Developing xAI approach for drug response prediction in cancer

BARGHORN Jérémy — 328403 — jeremy.barghorn@epfl.ch
CHAVEROT Jérémy — 315858 — jeremy.chaverot@epfl.ch
SCHIFFERLI Théo — 326468 — theo.schifferli@epfl.ch
The Underfitters

Department of Computer Science, EPFL, Switzerland

***Abstract*—Understanding drug responses in cancer is fundamental to advancing research and improving patient outcomes. Predicting these responses using genomic data presents a significant challenge due to the complexity and variability of biological systems. This work aims to develop innovative and explainable AI models capable of accurately predicting drug responses from cancer genomic data. Beyond prediction accuracy, the focus is on generating human-understandable explanations that shed light on the biological factors driving drug responsiveness. These insights will not only aid researchers in designing more effective treatments but also provide a deeper understanding of the underlying mechanisms of cancer drug responses.**

I. Introduction

This project explores the challenge of predicting drug responses in cancer using genomic response data through the application of machine learning (ML) and explainable AI (XAI) techniques. Given the high-dimensional and small-sample nature of the dataset, a common issue in the biomedicine field, the work focuses on reaching highest score and explainability while overcoming data sparsity and imbalance. To address these challenges, extensive data augmentation techniques, such as Gaussian noise addition and external dataset addition were employed, alongside dimensionality reduction and feature selection approaches.

A dozen models and architectures were developed and evaluated, ranging from linear regression models (e.g., Elastic Net) to more complex neural network architectures (e.g., MLP, CNN, Elastic Net, Auto-Encoders, WPFS, FSNet). Each model was rigorously tested to balance accuracy and explainability. Linear SVR emerged as the most promising model, demonstrating competitive performance with a Spearman's rank correlation coefficient of 0.57 in the Kaggle competition, granting us the first place in the competition currently.

Explainable AI techniques were integrated to provide insights into the biological factors influencing drug responsiveness, enhancing the interpretability of predictions using SHAP plots and LLaMA LM explanations. This approach paves the way for more interpretable and trustworthy cancer predictions by elucidating the genomic drivers of drug response variability.

II. Data Processing

The dataset utilized in this project is focused on cancer genomic data for Erlotinib response prediction. This drug is used to treat non-small cell lung cancer (NSCLC) and pancreatic cancer [1]. The `train.csv` file contains rows representing different cancer cell lines (tumor models) and columns representing gene expression levels across various genes, capturing the biological variability within the dataset. The corresponding labels are provided in `train_targets.csv`, where the "AAC" column reflects the drug response to Erlotinib, higher AAC values indicate better drug efficacy and the "tissue" column identifies the type of cancer cell line. While the tissue information can support model interpretability, the primary task is to predict AAC values and provide explainable insights. The `test.csv` dataset mirrors the structure of the training dataset, featuring cancer cell lines and gene expression data, and is used for evaluating model predictions. This small, high-dimensional dataset poses challenges common in biomedical applications, such as limited samples and data imbalance, requiring careful preprocessing and augmentation to ensure robust predictive performance and interpretability.

The training dataset consists of 743 samples, with 690 samples having an AAC value below 0.2. This indicates a significant class imbalance, as responses with AAC values higher than 0.7 are entirely absent from the data. Moreover the data is already cleaned, does not have missing values and can directly be used.

A. Feature Selection and Data Augmentation

Given the high imbalance and small size of the provided dataset, it was essential to identify additional samples that align with the existing data to augment the training set and improve its robustness. The primary challenge lies in the specificity of the dataset, as it focuses on Erlotinib drug response and includes patient data. Such datasets are inherently difficult to augment, and finding an exact match for the same subset of gene expressions is barely achievable.

1) Online Sources: The original dataset, sourced from the *Kaggle XAI in Cancer Medicine competition* [2], focuses on Erlotinib drug response and includes gene expression

profiles for patient cell lines. To expand this dataset, additional samples were integrated from the *CCLE dataset* [3], available on Zenodo. It provides RNA sequencing data and cell line drug response.

Workflow Overview: The original targets are represented as AAC values, while the additional dataset provided AUC values, for many different drugs (24 in total). Therefore, the following processes were applied to match the initial data.

1. **Gene Alignment**: RNA sequencing data (`CCLE_RNAseq.csv`) was loaded and filtered to retain only the genes common to the Kaggle dataset.

2. **Erlotinib-Specific Data**: The drug response data (`CCLE_cell_drug_labels.csv`) was filtered for Erlotinib responses and merged with matching gene expression data using `cell_line_id`.

3. **Target Calculation**: The AUC (Area Under the Curve) values were adjusted to handle negative values:

$$AUC_{\text{shifted}} = AUC + |\min(AUC)|$$

AAC (Area Above the Curve) was defined as:

$$AAC = 1 - AUC$$

Normalization ensured AAC values were in the range $[0, 1]$:

$$AAC_{\text{normalized}} = \frac{AAC - \min(AAC)}{\max(AAC) - \min(AAC)}$$

4. **Augmentation**: Augmented targets were appended to the original target dataset, and new sample identifiers (e.g., `CL###`) were assigned based on the last index. The gene expression data was combined with the original training set, retaining only the shared genes, 1460 in total. Augmented targets and augmented gene expressions were saved as {`train_targets_augmented.csv`; `train_augmented.csv`}.

B. More Advanced Processing Steps

The genomic data at our disposal have a numerous number of features, precisely $19,920$. To manage this large dataset, we apply dimensionality reduction techniques such as PCA, and feature selection with SELECTKBEST.

On one hand, Principal Component Analysis (PCA) is an unsupervised learning method where the idea is to learn a mapping $\mathbf{x}$ to $\mathbf{z}$ by finding a linear and orthogonal projection of the high dimensional data $\mathbf{x} \in \mathbb{R}^D$ to a low dimensional subspace $\mathbf{z} \in \mathbb{R}^L$ while preserving the variance using Singular Value Decomposition (SVD). As the samples are fewer than the visible space dimension D in the competition dataset, the number of principal components L is limited to the sample count at most. Note that PCA introduces limitations in model explainability, as the projection process results in the loss of original feature names.

On the other, SELECTKBEST is a supervised learning technique where we choose the features in our data that contribute most to the target variable by applying statistical tests such as chi-squared test (χ^2) or ANOVA $\mathcal{F}$-test.

Several data augmentation methods were tested to mitigate overfitting on the small competition dataset. The approach involved projecting the features into a lower-dimensional space using PCA, introducing random Gaussian noise with zero mean and variance ranging from 0.1 to 0.3, and finally reconstructing the original features by projecting them back to the original space. This process introduced randomness at multiple stages, effectively increasing the dataset size and enabling the model to learn more robust data representations.

III. MODELS AND METHODS

A. Linear Models

We first train simple regression models to set a baseline. Linear Regression predicts $\hat{y}$ as a weighted sum of input features plus a bias term: $\hat{y} = h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x}$, with parameters $\boldsymbol{\theta}$ obtained using the MSE objective. However, it is prone to overfitting.

Ridge Regression introduces regularization by adding a term to the cost function, constraining model weights to reduce overfitting. Lasso Regression, another regularized variant, uses the L_1 norm of weights, which performs automatic feature selection by setting less important feature weights to zero—useful for high-dimensional genomic data.

We combine Ridge and Lasso in Elastic Net, which mixes their regularization terms. Implemented with `scikit-learn` [4], these models perform reasonably well (*c.f.* TABLE I), but are insufficient for the Kaggle competition.

B. Neural Networks

In this section, we will discuss more advanced network architectures, highlighting their key advantages and drawbacks. All these networks are reproducible from our code. Additionally, we offer a highly flexible framework that facilitates the creation of custom architectures, as well as the incorporation of data preprocessing and data augmentation steps.

1) Basic Neural Networks: Neural networks were implemented in PyTorch with one or more hidden layers, dropout, and ReLU or LeakyReLU activations. These networks project data into progressively lower dimensions until reaching a final prediction for AAC. After trials, the most efficient architecture had two small hidden layers (up to 512 and 256 units). Larger networks led to vanishing gradients. Despite achieving the high scores in local, these models overfit local data and generalized poorly to unseen data. PCA improved feature selection, while data augmentation with noise allowed faster training using higher learning rates (1×10^{-3} vs. 1×10^{-5}). Mean Squared Error (MSE) was the loss function, with overfitting observed after two epochs. Kaggle submissions showed that training for just one epoch

with augmented data outperformed longer training. Attempts with one-dimensional CNNs yielded poor results.

2) Neural Networks with classes: Using the same approach as before, it was demonstrated that categorizing AAC predictions into 10 distinct classes, rather than predicting a continuous value between 0 and 1, enables achieving a theoretical maximum Spearman correlation (σ) of 95% by modifying the training objectives and employing a cross-entropy loss. With this setup, it was possible to see that the models learned the training data distribution, but the score submitted still showed poor accuracy. This indicates that the test data online is possibly not reflecting the training data distribution at our disposition, making it hard for our model to generalize to unseen samples with AAC higher than 0.3 as mentioned previously.

3) Encoder-Decoder models: To address overfitting in Neural Networks, Encoder-Decoder models inspired by the XA4C paper [5] were tested. These models project features into a latent space (encoding) and reconstruct them back to the original space (decoding), with reconstruction accuracy measured via MSE loss. This self-supervised training develops a latent feature representation. After training, the best encoder is frozen, and a classification head is added for AAC prediction. However, with only 742 samples, the model's depth and complexity made it unsuitable for this task, yielding poor results on Kaggle. We include the code as this approach has shown state-of-the-art performance on larger datasets.

4) FSNet and WPFS: Efficient machine learning algorithms for medical data emphasize the need to handle high-dimensional datasets with limited sample sizes. To address this, methods like Feature Selection Networks (FSNET) and Weight Predictor Networks with Feature Selector (WPFS) have shown strong performance. We adapted the WPFS implementation from Margeloiu et al. (2023) [6] for our dataset, recent PyTorch versions, and unclassified data. The code is available on the project repository. Using k-fold cross-validation and two training epochs, the model achieved less overfitting, with similar local and Kaggle results. However, accuracy remained low at 0.32, underperforming simpler SVR models. WPFS excels in well-structured, high-dimensional data with robust feature selection, but our imbalanced, limited dataset with skewed AAC distribution proved unsuitable for its strengths.

C. Support Vector Machine

We chose to experiment with kernel-based methods as they are particularly well suited for predictions on complex small-sized dataset. More specifically, we are interested in Support Vector Regression (SVR), which is a supervised learning method that adapts the principles of Support Vector Machines (SVM) for regression tasks. Unlike traditional linear methods, SVR optimizes a margin around data points using an ϵ-insensitive loss function, which ignores small errors, making it robust to minor deviations and outliers. SVR can model different types of relationships using kernels such as linear, polynomial, RBF or sigmoid.

We began by using grid search with cross-validation to tune hyperparameters, including the regularization strength, epsilon, kernel type, and convergence tolerance. However, this approach resulted in poor generalization and overfitting on the training set. Next, we experimented with dimensionality reduction techniques, but the results were inconclusive, decreasing significantly the Kaggle score. Training the model on the augmented dataset showed promising results for a certain sigmoid kernel type. Through trial and error, we achieved the #1 position on the competition leaderboard with our best model: an SVR model with linear kernel, trained with a specific epsilon value ($\epsilon = 0.18$) on all features from the competition dataset.

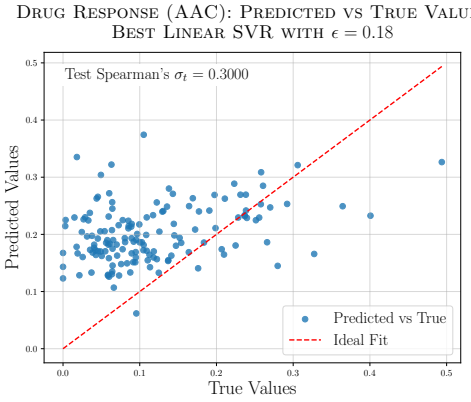

FIGURE I: BEST SVR MODEL PERFORMANCE ON THE TEST SET: AN EXAMPLE OF REGRESSION OUTPUT.

IV. RESULTS

Model	**Kaggle Spearman's σ_k**	**Test Spearman's σ_t**
Linear Regression	0.26	0.32
Elastic Net	0.41	0.39
Neural Network	0.39	**0.61**
Neural Network (w/ classes)	0.34	0.58
Encoder Decoder	0.16	0.21
WPFS	0.35	0.32
Linear SVR, $\epsilon = 0.18$	**0.58**	0.30
Linear SVR*, $\epsilon = 0.18$	-0.20	0.19
Sigmoid SVR, $\epsilon = 0.23$	0.56	0.30
Sigmoid SVR*, $\epsilon = 0.23$	0.53	0.36
Linear SVR, $\epsilon = 0.28$, $k = 500$	0.49	0.42

Models marked with * are trained on the augmented dataset.

TABLE I: RESULTING PERFORMANCE IN THE KAGGLE COMPETITION AND LOCALLY FOR THE BEST MODEL IN EACH CATEGORY.

Linear SVR with $\epsilon = 0.18$ significantly outperforms all other models on Kaggle's test set while some noticeable performance can be observed with the Neural Network on the local test set, unfortunelaty failing to generalise well.

V. Explainability

A. Tools

To interpret model decisions, there exist two prominent explainability librairies: LIME (Local Interpretable Model-agnostic Explanations) [7] and SHAP (SHapley Addictive exPlanations) [8]. While LIME is excellent for providing localized insights, this project emphasizes SHAP due to its ability to generate diverse visualizations, comprehensive results, and its broader scope, offering a global perspective on feature importance and model behavior.

From a technical standpoint, SHAP works by assigning each feature an importance value for a prediction based on Shapley values from game theory, ensuring the attributions sum to the model's output while satisfying local accuracy, missingness, and consistency.

Once the visualizations obtained, we aim to leverage the expertise of large language models (LLMs) to provide explanations and interpretations. For this task, we choose a state-of-the-art auto-regressive model called LLaMA 3.1 8B developed by Meta [9], as it allows for on device inference without relying on paid APIs like OpenAI's. This approach offers us greater flexibility and cost efficiency.

B. Graphs Generation

Initially, our goal was to interpret the best-performing model, that is the linear SVR trained on approximately $20,000$ features from the competition dataset. However, when using the SHAP Kernel Explainer on this model, we encountered extreme memory requirements. Even running the explainer on the EPFL cluster, which has 384 GB of RAM and 64 CPU cores, was impractical as the process took multiple days.

To address this, we attempted to optimize the SHAP explainer code. Through research, we discovered that the Kernel Explainer runs single-threaded by default, hence we parallelized the code using a Python library called `multiprocessing`. While this initially improved performance, the process eventually led to a deadlock—a state in concurrent computing where threads are stuck waiting for each other to take actions, halting progress. After a few hours, the program stalled, despite all CPU cores running at full utilization.

In the end, we decided to drastically reduce the feature set to 2.5% of the original size using the dimensionality reduction techniques mentioned earlier. Specifically, we used the former best-performing pipeline, set the feature selection threshold at $k = 500$, and increased slightly the epsilon value ($\epsilon = 0.28$) to ignore small deviations and focus on capturing the overall trend. This pipeline not only achieved a Kaggle score of 0.49, still a leading score in the competition, but also allowed us to effectively integrate the interpretability component.

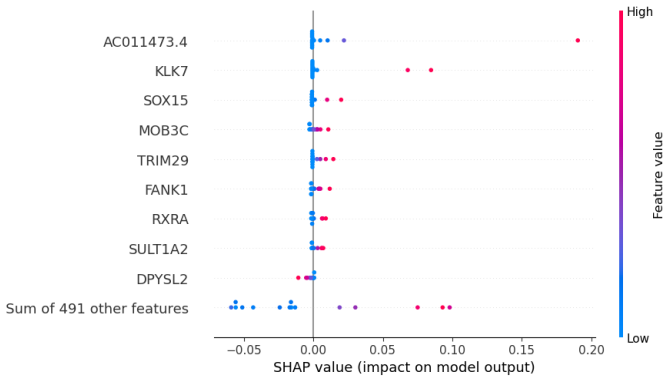

Figure II: SHAP Beeswarm Plot: Feature Contributions to Model Predictions

C. Text Generation

As a final objective, we intend to have each graph accompanied by a textual explanation, permitting us to compile a comprehensive PDF report for interpretability. To achieve this, zero-shot prompt engineering techniques like In-Context Learning (ICL) and Chain-of-Thought (CoT) are used to extract the most relevant insights from the language model [10]. Please refer to Appendix A for the detailed prompt template, model answer and generated SHAP plot.

VI. Ethical Risks

Ethical considerations must be carefully addressed in this project to ensure responsible and equitable outcomes, given its application of explainable AI (xAI) to predict drug responses using sensitive cancer genomic data. To ensure no ethical risks were overlooked, a systematic process was followed to evaluate potential risks across multiple dimensions. The primary stakeholders considered include direct stakeholders such as clinicians and researchers who rely on the predictions and generated explanations, and indirect stakeholders such as patients, whose genomic data indirectly informs the model's predictions.

To rule out ethical risks, the genomic data used in this project was reviewed to ensure strict adherence to privacy regulations, such as GDPR and HIPAA. It was confirmed that no patient-identifiable information is included, eliminating privacy concerns.

Furthermore, the development of explainable AI (xAI) models introduces a non-negligible responsibility to ensure that the explanations provided are accurate, transparent, and free from biases that could mislead or even fool researchers, geneticists or clinicians. With the incorporation of a Large Language Model into the project, it is crucial to be extra careful to prevent the introduction of additional biases in the generated explanations, maintaining the integrity and reliability of the insights produced. Unlike OpenAI's GPT-4 large multi-modal model, Meta's LLaMA models are open-source, which is a significant advantage for addressing ethi-

cal concerns, as they allow greater transparency, community-driven improvements, and scrutinity of their architecture and training data.

Sustainability concerns were addressed through the use of pre-trained models for explainability and efficient training pipelines to minimize resource consumption.

However, it is important to keep in mind that our model's predictions are in the medical domain and can have significant impacts on patients and their lives. Therefore, all results must either be re-verified or validated by experts, as a Spearman correlation of 0.57 is promising but not high enough to make appropriate decisions that could affect people's health. Additionally, while modern machine learning models are highly performant, this does not mean we should disregard the human aspect. These results, which are ultimately automated numerical outputs, should be used with the purpose of improving patients' lives, and the communication of such information must be done in a *human-centered* manner.

By following a rigorous evaluation and validation process, this work seeks to contribute meaningfully to cancer research while upholding principles of fairness, accountability, and respect for the sensitive genomic data underpinning these advancements.

VII. Conclusion

We developed an explainable AI approach for predicting drug responses in cancer, achieving promising results with a Spearman correlation of 0.57 in the Kaggle competition, ranking first. By leveraging diverse machine learning techniques, including feature selection, data augmentation, dimensionality reduction and more complex architectures, we addressed the challenges posed by high-dimensional, small-sample datasets. Our integration of SHAP-based explainability provided insights into the biological drivers of drug responsiveness, giving a better interpretability of predictions.

While our results underline the potential of xAI in cancer research, the dataset's inherent limitations and imbalance highlight the importance of collaboration with domain experts to refine feature selection and model robustness. This work lays the groundwork for developing AI tools that balance accuracy and transparency, fostering a deeper understanding of cancer drug responses and enabling advancements in precision medicine.

VIII. Acknowledgement

We would like to thank Prof. Dr. Jasmina Bogojeska (ZHAW) for her valuable help throughout this project. Her willingness to assist and answer our questions greatly supported our work and helped us better understand the project goals and the competition.

REFERENCES

[1] Drugs.com, "Erlotinib monograph for professionals," 2019, archived from the original on 24 December 2019. Retrieved 12 November 2019. [Online]. Available: https://www.drugs.com/monograph/erlotinib.html

[2] D. A. Mer", "Xai in cancer medicine dataset," 2024. [Online]. Available: https://www.kaggle.com/competitions/xai-in-cancer-medicine/data

[3] A. Authors", "Ccle dataset on zenodo," 2022. [Online]. Available: https://zenodo.org/records/6972738

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] K. P. L. T. L. W. Z. Q. . Li Q, Yu Y, "Xa4c: explainable representation learning via autoencoders revealing critical genes." *PLoS Comput Biol 19(10) : e1011476. https://doi.org/10.1371/journal.pcbi.1011476*, 2023.

[6] A. Margeloiu, N. Simidjievski, P. Lio, and M. Jamnik, "Weight predictor network with feature selection for small sample tabular biomedical data," in *37th AAAI Conference on Artificial Intelligence*, 2023.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[8] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[10] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2023. [Online]. Available: https://arxiv.org/abs/2205.11916

APPENDIX

A. Prompt used for the LLM generation

You are a highly knowledgeable medical expert with expertise in genomics, bioinformatics, and machine learning. You are tasked with explaining the impact of specific genetic features on a prediction model to a patient or researcher. Your explanation should be highly formal, precise, and delivered in a professional tone. Be sure to explain technical terms and their relevance to the patient's condition or the research in layman's terms, but without oversimplifying the underlying concepts.
Given the following SHAP values and feature data from an SVR model predicting the drug response to erlotinib, explain the significance of the top features contributing to the predicted Area Above the Curve (AAC). Ensure the explanation is framed in a formal, scientific tone, as if addressing a clinical research team.
Input data:

Feature	SHAP Value	Feature Value
AC011473.4	0.190322	80.8526
KLK7	0.0679491	22.0465
GJB3	0.00697801	7.00119
DPYSL2	-0.00440222	5.8528

Prediction : 0.572
Frame your explanation to highlight: The biological relevance of the identified features in the context of drug response. What are the conclusions that can be drawn from the SHAP values about the drug response prediction. Don't add any additional part or introduction.

B. LLaMA model generation

This report provides an analysis of the top features contributing to the predicted Area Above the Curve (AAC) in a SHAP-based model for predicting the drug response to erlotinib. The model has identified four key features: AC011473.4, KLK7, GJB3, and DPYSL2, with corresponding SHAP values of 0.19032, 0.0679576, 0.00698294, and -0.0043958, respectively.

1) **AC011473.4**: This feature is associated with the gene AC011473.4, which is a long non-coding RNA (lncRNA). LncRNAs play a crucial role in regulating gene expression by interacting with various proteins and epigenetic modifications. The positive SHAP value of 0.19032 indicates that this feature is a strong predictor of drug response, suggesting that higher expression levels of AC011473.4 are associated with a better response to erlotinib. This is consistent with previous studies that have shown lncRNAs to play a role in modulating the efficacy of targeted therapies.
2) **KLK7**: The KLK7 gene encodes a serine protease involved in the regulation of cell growth and differentiation. The SHAP value of 0.0679576 indicates that KLK7 expression is positively correlated with drug response, suggesting that higher levels of KLK7 are associated with a better response to erlotinib. This is consistent with the role of KLK7 in regulating cell growth and differentiation, which may contribute to the efficacy of erlotinib in targeting cancer cells.
3) **GJB3**: The GJB3 gene encodes a connexin protein involved in cell-cell communication and gap junction formation. The SHAP value of 0.00698294 indicates that GJB3 expression is positively correlated with drug response, suggesting that higher levels of GJB3 are associated with a better response to erlotinib. This is consistent with the role of GJB3 in regulating cell-cell communication, which may contribute to the efficacy of erlotinib in targeting cancer cells.
4) **DPYSL2**: The DPYSL2 gene encodes a protein involved in cell adhesion and migration. The negative SHAP value of -0.0043958 indicates that DPYSL2 expression is negatively correlated with drug response, suggesting that higher levels of DPYSL2 are associated with a poorer response to erlotinib. This is consistent with the role of DPYSL2 in regulating cell adhesion and migration, which may contribute to the development of resistance to erlotinib.

Conclusions

The SHAP values provide valuable insights into the biological mechanisms underlying the predicted drug response to erlotinib. The positive SHAP values of AC011473.4, KLK7, and GJB3 suggest that these features are associated with a better response to erlotinib, while the negative SHAP value of DPYSL2 suggests that it is associated with a poorer response. These findings highlight the importance of these features in predicting the efficacy of erlotinib and may inform the development of personalized treatment strategies. Further research is needed to validate these findings and explore the potential therapeutic applications of these features.

C. SHAP plot for the prediction

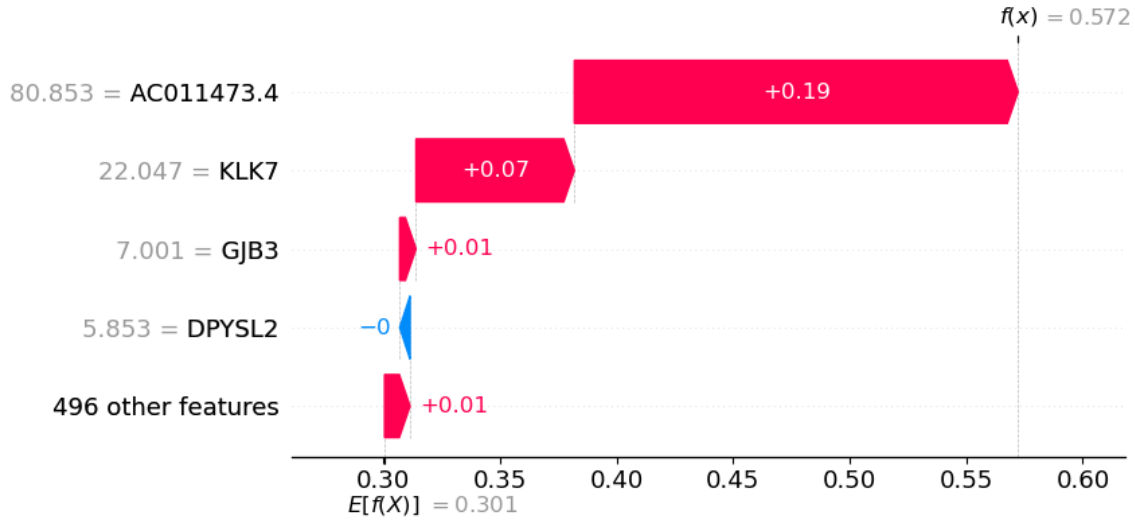

FIGURE III: SHAP WATERFALL PLOT: TOP 4 AAC SAMPLE