A Reimplementation of Feature Selection and Machine Learning Approaches for Predicting CYP2D6 Methylation

Leveraging Genetic and Epigenetic Data for Improved Pharmacogenomic Predictions

INTRODUCTION

Pharmacogenetics personalizes medicine by linking genetic variants to drug response. The CYP2D6 gene, influencing the metabolism of 25% of drugs, adds complexity due to genetic polymorphisms and DNA methylation. Integrating both genetic and epigenetic data through machine learning can improve predictions of drug response, particularly for pediatric populations.

CONTEXT

Genetic Variability: CYP2D6 shows significant genetic polymorphism, affecting enzyme activity.

Epigenetic Regulation: DNA methylation modulates CYP2D6 expression, impacting drug metabolism.

Prior Research: Studies like Fong et al. (2024) highlight the challenge of integrating high-dimensional genetic and methylation data for accurate predictions.

OBJECTIVES

- Reimplement and streamline the methodology from Fong et al. (2024).
- Evaluate the performance of Linear Regression, Elastic Net, and XGBoost models.
- Improve prediction accuracy of CYP2D6 methylation levels.

RESULTS

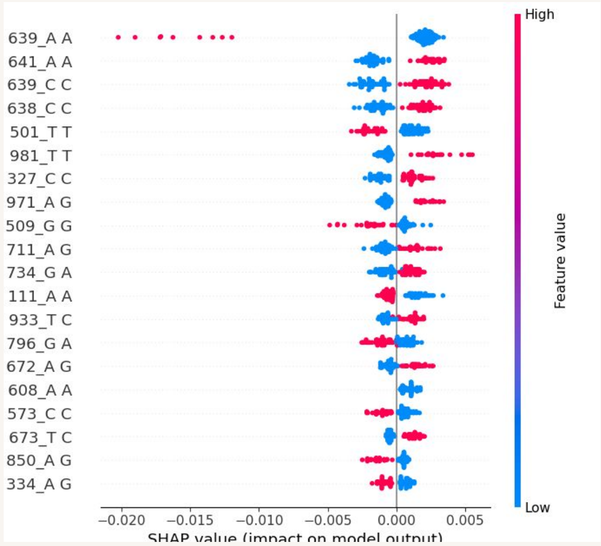
Best Performance: Probe5 (R^2 = 0.2084)
Overfitting: Probe3 showed poor generalization (R^2 = -0.1470)

Table 1. Performance metrics for the GWAS_beforeBH_combined feature set using XGBoost.

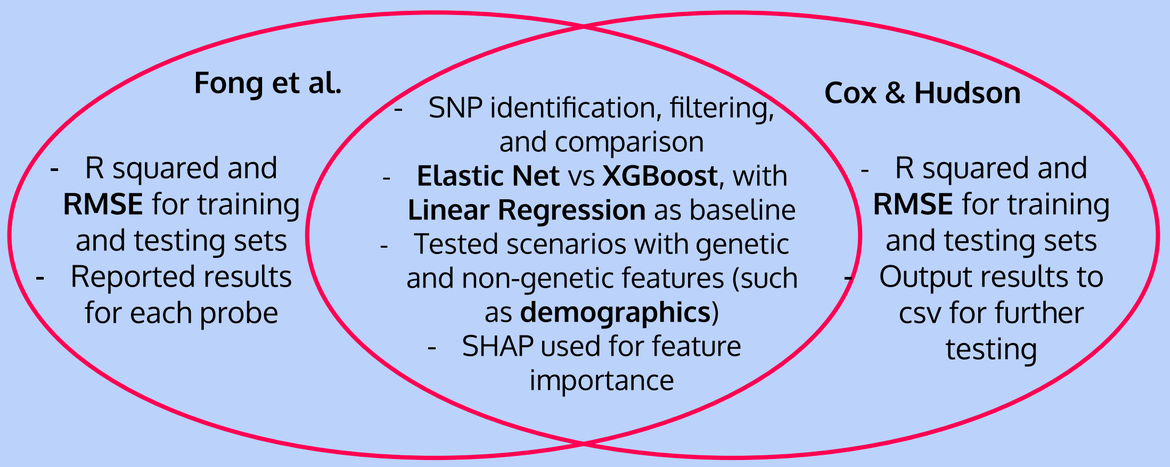
Metric	Probe1	Probe3	Probe5	Mean (SD)
RMSE train set	0.0107	0.0031	0.0048	0.0062 (0.0040)
R^2 train set	0.9750	0.9817	0.9796	0.9787 (0.0034)
RMSE test set	0.0620	0.0196	0.0269	0.0361 (0.0226)
R^2 test set	0.0579	-0.1470	0.2084	0.0398 (0.1784)

SHAP Analysis for XGBoost on Probe5

- Shows SHAP values indicating the impact of each SNP on the model's output.
- High feature values (red) and low feature values (blue) influence predictions differently.
- Key SNPs with notable impact include: **639_A A**, **641_A A**, **501_T T**
- Helps identify SNPs that drive methylation predictions.



Approach Comparison



CONCLUSION

Our study shows that predicting CYP2D6 methylation using genetic and demographic data is challenging. While we tested Linear Regression, Elastic Net, and XGBoost models, achieving consistent and reliable predictions is still a work in progress.

MODEL COMPARISON

- **Linear Regression**: Overfits training data, poor test performance.
- **Elastic Net**: Modest improvement, better generalization with regularization.
- **XGBoost**: Captures non-linearities, best training performance but inconsistent test results.
- **Challenges**: High-dimensional data, overfitting, and sparse genetic signals.

ACKNOWLEDGEMENTS

- Dr. Ibrahim Numanagic for research guidance.
- GUSTO cohort and GTEx Consortium for data availability.

CONTACT INFO

Baz Cox
Email: basilcox847@gmail.com
Jake Hudson
Email: jhudson0530@gmail.com