Published online

Nucleic Acids Research, 0000, Vol. 00, No. 00 ***1–3***
doi:10.1093/nar/gkn000

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

Baz Cox [1*] **and Jake Hudson** [2†]

[1]University of Victoria, Canada
[2]Independent Researcher, Canada

ABSTRACT

Pharmacogenetics has increasingly informed personalized medicine by tailoring drug prescribing based on genetic profiles, yet current pharmacogenomic tests often fail to capture the full complexity of drug metabolism. For instance, the highly polymorphic cytochrome P450 2D6 (CYP2D6) gene influences the metabolism of approximately 25% of commonly used drugs. Genetic variants in CYP2D6 alleles are well-studied, but emerging evidence highlights an additional layer of complexity from epigenetic regulation, notably DNA methylation. Integrating genetic and epigenetic data may improve predictions of drug response and support precision dosing strategies, especially in pediatric populations where interindividual variability in drug handling can be critical.

In this study, we reimplement and streamline the methodology described in Fong et al. (2024) [(9)] to predict CYP2D6-related DNA methylation levels from genetic variation. Our approach uses genotype and demographic data, leverages minor allele frequency (MAF)-based feature selection, and applies machine learning models—Linear Regression, Elastic Net, and XGBoost—to identify predictive relationships. We also incorporate principal components (PCs) to account for population structure and examine the impact of including demographic variables.

Model performance is assessed using Root Mean Square Error (RMSE) and R-squared (R^2) metrics, and feature importance is explored using SHapley Additive exPlanations (SHAP). Although our results broadly align with the complexity reported in the original study, they underscore that achieving robust, generalizable predictions remains challenging. This reimplementation provides a reproducible and efficient pipeline, setting the stage for further refinements in genomic-epigenomic integrative modeling for precision pharmacotherapy.

INTRODUCTION

Pharmacogenetics has advanced the personalization of drug therapy by identifying genetic variants that influence drug metabolism, efficacy, and toxicity [(1), (2)]. CYP2D6, a key enzyme in drug metabolism, exhibits substantial genetic polymorphism, resulting in variable enzyme activity across populations [(3), (4)]. Although well-characterized genetic variants inform the classification of individuals into metabolizer categories (poor, intermediate, extensive, ultra-rapid), these genotypes do not fully explain interindividual variation in drug response.

Emerging evidence suggests that epigenetic mechanisms, such as DNA methylation, can modulate CYP2D6 expression, adding complexity to phenotype prediction [(5), (6)]. The integration of genetic and methylation data through machine learning (ML) offers an opportunity to improve predictive accuracy. Prior research has demonstrated the value of ML in identifying subtle genetic and epigenetic signals underlying complex traits [(7), (8)].

Fong et al. (2024) [(9)] proposed a comprehensive ML framework incorporating feature selection (e.g., SNP filtering via MAF thresholds, mQTL identification, and eQTL integration from GTEx) and multiple modeling approaches (Linear Regression, Elastic Net, XGBoost). Their work highlighted the challenges of modeling high-dimensional genomic data and the partial improvements offered by specialized feature sets.

Here, we reimplement the core methodology of Fong et al. [(9)], focusing on replicability and streamlining the pipeline for broader applicability. We examine scenarios with genetic features only and those incorporating demographic factors and PCs. By comparing Linear Regression, Elastic Net, and XGBoost models, we assess the trade-offs between model simplicity, regularization, and non-linear modeling capacity. Finally, we employ SHAP to interpret model outputs, providing insights into which genetic variants may drive methylation variation. Our goal is to improve understanding of CYP2D6 methylation prediction while elucidating the methodological steps for future genomic-epigenomic integrative studies.

*Email:basilcox847@gmail.com
†Email: jhudson0530@gmail.com

MATERIALS AND METHODS

Data Sources and Preprocessing

We used genotype, demographic, and methylation data focusing on eight CYP2D6-associated CpG sites [(9)]. Genotype data included:

1. **GWAS-derived SNPs:** SNPs identified before and after Benjamini-Hochberg (BH) correction.

2. **GTEx eQTLs:** SNPs associated with CYP2D6 expression from the GTEx database [(10), (11)].

Methylation beta values were processed to ensure quality control. Demographic variables (e.g., maternal income, ethnicity, maternal age, and education) and PCs from genotype data were included in certain scenarios. The dataset was split into training (75%) and testing (25%) sets based on a predefined list of test IDs, preventing data leakage and ensuring reproducibility.

Feature Selection and Encoding

Features were filtered based on MAF and mapped to the CYP2D6 genomic region or sourced from eQTL and mQTL studies. Categorical features were one-hot encoded, and continuous features were z-score normalized. PCs were added to account for population substructure, and demographic data were optionally included to explore their contribution to predictive performance.

Modeling Approaches

We implemented three models:

1. **Linear Regression:** A baseline model without regularization, providing a reference point for model complexity.

2. **Elastic Net:** Combines L1 and L2 penalties, controlling overfitting and dealing with multicollinearity [(14)]. Hyperparameters were tuned via cross-validation.

3. **XGBoost:** A gradient boosting approach optimizing tree ensembles, capturing non-linear interactions [(15), (16)]. Hyperparameters (e.g., max depth, learning rate) were grid-searched using 3-fold cross-validation.

Two main scenarios were considered:

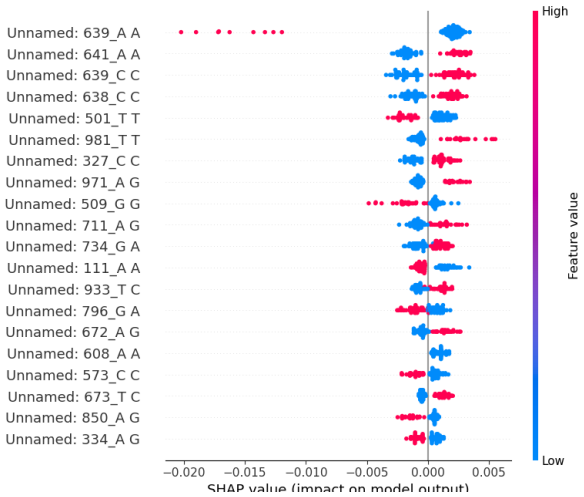
1. **genetic_only:** Models trained on SNP features alone.

2. **genetic_pcs_nongenetic:** Models integrating SNPs, PCs, and demographic variables.

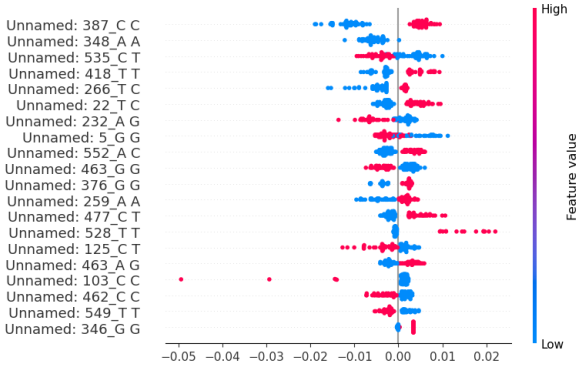
Evaluation Metrics and Interpretability

Models were evaluated using RMSE and R^2 on both training and testing sets. This approach allowed assessment of overfitting and generalization. When possible, we employed SHAP analysis [(17)] to identify key SNPs contributing to methylation predictions, enhancing interpretability and guiding future biological validation.

All results were recorded, enabling comparison across probes, feature sets, scenarios, and models.



(a) Best scenario performance.



(b) Poor generalization performance.

Figure 1. Performance comparison of models across different scenarios.

RESULTS

Our reimplementation yielded insights broadly consistent with those of Fong et al. [(9)], yet persistent challenges in predictive accuracy emerged.

Model Performance Overview

Our reimplementation of Linear Regression, Elastic Net, and XGBoost models ielded varying performance across different probes when predicting CYP2D6 methylation levels. The table below shows the performance metrics for the GWAS_beforeBH_combined feature set in the genetic_pcs_nongenetic scenario.

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)

As shown in Table 1, XGBoost exhibited strong performance on the training set with high R^2 values (mean R^2 = 0.9787). However, test performance varied

significantly. Probe5 achieved the highest test R^2 (0.2084), suggesting moderate predictive power. In contrast, Probe3 exhibited a negative R^2 value (-0.1470), indicating poor generalization and potential overfitting in that scenario. This inconsistency highlights the challenges of achieving robust and generalizable predictions with high-dimensional genetic and methylation data.

Linear Regression, while simple, perfectly fit the training data ($R^2 = 1.0$) but performed poorly on the test data, showing significant overfitting. Elastic Net, incorporating L1 and L2 regularization, improved generalization modestly, with test R^2 values generally below 0.2. Nonetheless, Elastic Net consistently outperformed Linear Regression, especially when SNPs were filtered after BH correction or derived from GTEx eQTLs.

The introduction of principal components (PCs) and demographic data had minimal impact across models, implying that the primary predictive signal resides in the genetic variants. Further model tuning, regularization, and feature engineering may be necessary to improve performance across all probes and achieve more consistent generalization.

Feature Set Influence

Feature sets derived from SNPs after BH correction and eQTLs from GTEx slightly improved performance over unfiltered GWAS SNP sets. This aligns with previous findings that biologically informed feature selection enhances predictability [(12), (13)]. However, overall improvements were incremental.

Interpretability via SHAP

Although not comprehensively applied, preliminary SHAP analyses indicated a small subset of SNPs exerted stronger influence on model predictions. Understanding these SNPs could guide future functional studies to elucidate methylation regulation mechanisms.

DISCUSSION

Our findings highlight the complexity of predicting CYP2D6 methylation from genetic data, echoing challenges noted in the literature [(7), (8), (9)]. While Elastic Net and XGBoost improved over a simple Linear Regression baseline, test performance remained modest, emphasizing the difficulty of capturing epigenetic regulation from sparse genetic signals. Additionally, incorporating demographic features and PCs provided limited gains, suggesting these external factors do not strongly modulate CYP2D6 methylation predictions in this dataset.

Overfitting in XGBoost models suggests that more stringent regularization or alternative modeling strategies (e.g., Bayesian approaches, neural networks, or integrated multi-omics data) may be required. Future studies could also explore more sophisticated feature engineering, kernel-based methods, or ensemble techniques to enhance generalization.

CONCLUSION

This reimplementation demonstrates the nuanced interplay between genetic variation, methylation, and machine learning. While some improvements were noted using Elastic Net and informed feature sets, the field remains far from robust, clinically relevant predictions of CYP2D6 methylation status.

Refining feature selection, employing more advanced ML methods, and integrating additional omics layers will likely be needed to move toward accurate, reliable predictions. Nonetheless, this work provides a reproducible pipeline and highlights key areas for methodological refinements, ultimately contributing to the long-term goal of advancing precision pharmacogenomics and personalized therapy.

ACKNOWLEDGEMENTS

We thank Dr. Ibrahim Numanagic for guidance in identifying this research direction. We also acknowledge the contributors of the GUSTO cohort and the GTEx Consortium for making their data publicly available.

REFERENCES

1. Oates,J. and Lopez,J.A. (2018) Pharmacogenetics: an important part of drug development. *Int J Biomed Invest*, **1**, 1–16.
2. Van Driest,S.L. et al. (2017) Pharmacogenomics in clinical practice: Challenges and opportunities. *Mayo Clin Proc*, **92**, 1589–1601.
3. Beoris,M. et al. (2016) CYP2D6 Copy Number Distribution in the US Population. *Pharmacogenet Genomics*, **26**, 96–99.
4. Bradford,L.D. (2002) CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. *Pharmacogenomics*, **3**, 229–243.
5. Kacevska,M. et al. (2012) Epigenetic-dependent regulation of drug transport and metabolism: an update. *Pharmacogenomics*, **13**, 1373–1385.
6. Habano,W. et al. (2015) Analysis of DNA methylation landscape reveals the roles of DNA methylation in the regulation of drug metabolizing enzymes. *Clin Epigenetics*, **7**, 1–11.
7. Libbrecht,M.W. and Noble,W.S. (2015) Machine learning applications in genetics and genomics. *Nat Rev Genet*, **16**, 321–332.
8. Ho,D.S.W. et al. (2019) Machine learning SNP based prediction for precision medicine. *Front Genet*, **10**, 1–10.
9. Fong,W.J. et al. (2024) Comparing feature selection and machine learning approaches for predicting CYP2D6 methylation from genetic variation. *Front Neuroinform*, **17**, 1244336.
10. Lonsdale,J. et al. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580–585.
11. Nicolae,D.L. et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*, **6**, e1000888.
12. John,M. et al. (2022) A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. *Front Plant Sci*, **13**, 1–16.
13. He,Z.X. et al. (2015) Impact of physiological, pathological and environmental factors on the expression and activity of human CYP2D6. *Drug Metab Rev*, **47**, 470–519.
14. Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J Royal Stat Soc B*, **67**, 301–320.
15. Chen,T. and Guestrin,C. (2016) XGBoost: A scalable tree boosting system. *KDD '16*, pp. 785–794.
16. Li,W. et al. (2019) Gene expression value prediction based on XGBoost algorithm. *Front Genet*, **10**, 1–7.
17. Lundberg,S.M. and Lee,S.I. (2017) A unified approach to interpreting model predictions. *NIPS '17*, pp. 4765–4774.