NTNU

An Interpretability-Focused Approach to Predicting the Effect of Cancer Mutations in Humans Using Random Forests

Marcos Barreiro
Department of Computer Science, NTNU

Correspondence: marcosba@ntnu.no

Code Repository: github.com/maberrirof2002/RF-Cancer-Mutation-Prediction

Abstract

Background: Somatic mutations play a central role in cancer development, yet the vast majority of observed mutations remain of unknown clinical significance. Advances in genome sequencing have enabled the creation of large-scale variant databases such as COSMIC and ClinVar, but interpreting the functional consequences of mutations remains a major challenge in cancer research.

Results: In this study, we present a machine learning pipeline designed to predict the pathogenicity of cancer mutations using a Random Forest classifier. We integrated data from COSMIC and ClinVar and engineered biologically meaningful features such as BLOSUM62 substitution scores, hydrophobicity and charge shifts, mutation type, and nucleotide-level alterations. The model achieved high predictive performance (AUC 0.9933, F1-score 0.9804), and we ensured transparency through interpretability techniques including SHAP, LIME, and partial dependence plots. Dimensionality reduction via CCA confirmed the biological relevance of key features, and a simplified decision tree helped illustrate the model's internal logic to non-expert audiences.

Conclusions: Our results demonstrate that interpretable machine learning models, grounded in biological theory and domain-specific feature design, can reliably classify the clinical significance of cancer mutations. This approach may support variant prioritization and complement existing tools in genomics-driven oncology.

Keywords

Keywords: Cancer mutations, Machine learning, Random Forest, Interpretability, COSMIC, ClinVar, SHAP, LIME, Genomic features

Background

Cancer Mutations in Humans

Cancer is fundamentally a genetic disease, driven by the accumulation of mutations in the genome of somatic cells. These mutations can disrupt normal regulatory mechanisms that govern cell proliferation, differentiation, and death. Over time, a subset of these mutations—termed *driver mutations*—confer selective advantages to the affected cells, promoting clonal expansion and tumor progression [1]. In contrast, *passenger mutations* are co-occurring alterations that do not directly contribute to oncogenesis but can still complicate the interpretation of cancer genomes.

The sources of genetic variation in cancer are diverse. They include single nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations, and chromosomal rearrangements. Among these, non-synonymous SNVs, particularly missense and nonsense mutations, can lead to changes in the encoded proteins' structure or function. A single point mutation in a critical gene such as TP53, KRAS, or EGFR can dramatically alter the trajectory of a cancer cell. These effects are not uniform: the same mutation may have different outcomes depending on the cellular context, tissue type, and genetic background [2].

Moreover, not all genetic variation is acquired during life. Inherited variants, including high-penetrance germline mutations in genes like BRCA1/2, also contribute to cancer risk. The interplay between inherited and acquired mutations defines each tumor's unique mutational landscape, which presents both challenges and opportunities for diagnosis, prognosis, and therapy.

Given the complexity and heterogeneity of cancer genomes, computational tools are essential to prioritize mutations for further study. Functional prediction algorithms such as SIFT [4], PolyPhen [5], and PROVEAN [6] assess the likelihood that a specific amino acid change impacts protein function. However, these tools typically focus on general conservation and structural properties, lacking cancer-specific context or empirical validation. A growing consensus supports the integration of machine learning with curated databases to address these limitations.

Mutation Databases

Robust annotation of mutations requires comprehensive and high-quality data sources. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a premier resource that compiles somatic mutations found in cancer tissues, derived from large-scale sequencing projects and published literature [7]. COSMIC captures detailed genomic, transcriptomic, and protein-level information for each mutation, including its location, type, zygosity, and associated cancer phenotype. Of particular relevance is the Mutant Census dataset, which offers curated records of functionally significant mutations.

While COSMIC provides breadth and detail about mutations observed in cancer samples, it lacks systematic annotations of clinical significance. This is where ClinVar complements COSMIC. ClinVar, maintained by the National Center for Biotechnology Information (NCBI), aggregates information about the pathogenicity of genetic variants, based on submissions from diagnostic laboratories, researchers, and clinicians [8]. Each variant is reviewed and classified into categories such as ”pathogenic,” ”likely benign,” or ”uncertain significance.”

The intersection of COSMIC and ClinVar enables the construction of a labeled dataset of somatic mutations with known clinical outcomes. This fusion allows supervised learning methods to be applied: COSMIC contributes the biological context and mutation prevalence, while ClinVar supplies the necessary target labels. However, merging these datasets also requires careful data cleaning, normalization of identifiers, and resolution of ambiguities—a nontrivial preprocessing step that lays the foundation for any downstream modeling.

Machine Learning and Interpretation Techniques

Machine learning has become a central tool in genomics, particularly for tasks that involve high-dimensional data and subtle patterns. For this study, we employed a Random Forest (RF) classifier, an ensemble learning method that combines the outputs of many decision trees to achieve improved generalization and reduced overfitting [9]. RFs are particularly well-suited to biological data due to their ability to model nonlinear relationships, handle missing values, and compute feature importances.

However, raw biological data often contains redundant, noisy, or highly correlated features. To address this, we applied two dimensionality reduction techniques: Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA). PCA transforms the data into a set of orthogonal components capturing maximal variance, which can reveal intrinsic structure and improve model stability. CCA, on the other hand, finds correlated projections between two datasets—in this case, feature matrices and label vectors—highlighting the most predictive combinations.

To tune the model and avoid overfitting, we performed grid search cross-validation. This process systematically explores combinations of hyperparameters (e.g., tree depth, number of estimators) and evaluates them on held-out data, selecting the configuration with the best performance.

A key requirement in biomedical applications is interpretability. To understand how the model makes predictions, we employed SHAP (SHapley Additive exPlanations), a game-theoretic framework that decomposes predictions into feature contributions [10]. SHAP values offer a global and local explanation of model behavior, allowing us to identify which features consistently drive predictions. We also used LIME (Local Interpretable Model-agnostic Explanations), which perturbs input data and fits simple models in the neighborhood of a prediction to reveal influential features [11]. Together, these tools ensure that our model is not only accurate but also transparent and trustworthy.

Implementation

All implementation was carried out in Python, using Google Colab notebooks. The computational workflow relied on robust open-source libraries including NumPy [12], pandas [13], scikit-learn [14], and SHAP [10]. This cloud-based setup facilitated reproducibility and modular development across three main stages: data processing, model development, and interpretation.

Dataset Preparation

We began by merging the COSMIC Mutant Census v101 with ClinVar, two complementary sources of curated cancer mutations and clinical annotations. Merging required harmonization of transcript and genomic coordinates, and removal of duplicates and ambiguous records. Special care was taken to align ENST transcript IDs and genomic positions for consistency.

Next, we augmented the dataset with three biochemical descriptors to capture the impact of amino acid substitutions:

- **BLOSUM62 substitution scores** [15], derived from evolutionary substitution probabilities;
- **Hydrophobicity change**, calculated using the Kyte-Doolittle scale [16];
- **Charge change**, indicating the gain, loss, or conservation of net residue charge.

Each mutation was then assigned a simplified label from ClinVar: *pathogenic*, *benign*, or *unknown*. This reduction into three categories enabled targeted filtering and preparation for binary classification. The decision to reduce to a binary setting was guided both by label quality and consistency with methods for clinical variant classification [17].

Additionally, we considered incorporating 3D structural information by linking mutations to corresponding PDB (Protein Data Bank) entries. Structural context can significantly enhance predictions by providing insight into solvent accessibility, domain location, and interaction sites. However, due to the sparse availability of high-resolution structures for all mutated proteins and the added computational burden, this approach was ultimately set aside. The strong results obtained with sequence-based and biochemical features justified focusing our efforts on these more scalable inputs.

Model Development

We restricted our model development to mutations with a known label (i.e., either *pathogenic* or *benign*), resulting in a clearly defined supervised learning task. This binary formulation enabled more robust performance assessment and avoided introducing noise from ambiguous variants of uncertain significance (VUS).

Handling missing data was a central challenge. Mutations may lack certain annotations due to sequencing gaps or incomplete curation. Numerical features were imputed using median values to preserve distributional characteristics, while categorical features were encoded using one-hot encoding after rare categories were collapsed. These choices align with best practices for preparing ensemble learners such as Random Forests, which require uniform, numerical feature inputs and are known to perform well with moderate noise [9].

Particular attention was required for biologically complex columns such as `MUTATION_CDS`, `GENOMIC_WT_ALLELE`, and `GENOMIC_MUT_ALLELE`. These fields contain free-text representations of genomic events that are often hard to encode directly. To address this, we decomposed `MUTATION_CDS` into binary indicators representing the presence of specific mutational mechanisms: substitutions, insertions, deletions, and duplications. This transformation enabled the model to distinguish between different types of coding sequence alterations without relying on textual parsing.

Similarly, the columns `GENOMIC_WT_ALLELE` and `GENOMIC_MUT_ALLELE` were used to calculate the length and magnitude of the mutation at the nucleotide level. We computed the difference in sequence length as a proxy for insertions or deletions and estimated the number of changed bases to capture mutation complexity. This quantitative encoding preserved the essence of structural changes while fitting into the model's numerical feature framework.

Although Random Forests are not inherently sensitive to feature scale, we applied standard normalization to all features. This allowed for easier integration with dimensionality reduction techniques and enabled more interpretable projections in subsequent visualization tasks.

Dimensionality reduction was applied for exploration and interpretability. Principal Component Analysis (PCA) identified dominant axes of variance, while Canonical Correlation Analysis (CCA) identified correlated linear projections between feature space and class labels [18]. These techniques provided valuable insight into the structure of the dataset, although they were not applied to the final training data to preserve full feature expressiveness.

We fit a Random Forest classifier using scikit-learn's implementation. A grid search with 5-fold cross-validation optimized parameters such as the number of trees, maximum depth, and minimum samples per split. Performance was evaluated using classification accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic Curve (AUC).

While Random Forests were the preferred model due to their robustness and interpretability, we also experimented with deeper architectures, including a feedforward deep neural network (DNN) using TensorFlow. Although the DNN achieved marginally higher performance in some runs, it suffered from greater variance, higher complexity, and reduced transparency. Given our project's emphasis on model interpretability, the DNN was ultimately excluded. This aligns with recent calls for transparency in clinical decision-support systems [19].

The final model delivered a strong balance of performance and interpretability, suitable for downstream feature attribution and clinical translation.

Results

To evaluate the performance and interpretability of our model, we assessed both quantitative metrics and qualitative insights. We used held-out test data to compute the model's accuracy, F1-score, and ROC-AUC, and complemented this with analyses based on PCA, CCA, SHAP, LIME, and decision tree visualization.

Model Performance

The Random Forest classifier achieved strong predictive results on the test set. Table 1 summarizes the main performance metrics.

Metric	Value
Accuracy	0.9666
F1-Score	0.9804
ROC-AUC	0.9933

Table 1: Performance metrics of the Random Forest classifier on the test set.

The ROC curve in Figure 1 further confirms the model's excellent discriminative ability.

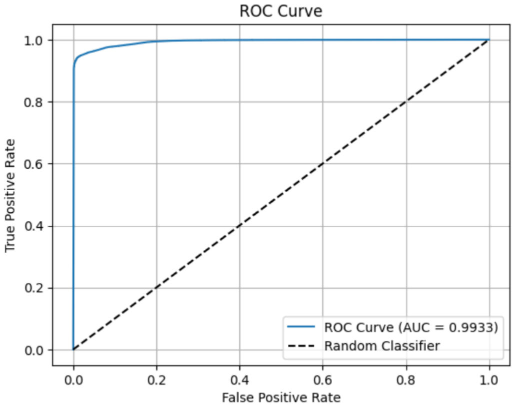


Figure 1: Receiver Operating Characteristic (ROC) curve for the Random Forest model.

Data Structure and Projection

To explore feature redundancy and potential separability in the dataset, dimensionality reduction techniques were applied. PCA was initially explored to reduce feature space, but the first ten principal components accounted for less than 60% of the total variance, which was insufficient to justify dimensionality reduction. Therefore, PCA was excluded from the final modeling pipeline. CCA, however, offered useful insights into which features correlated most strongly with the target label (Figure 2).

Tree Logic and Model Explanation

To help readers unfamiliar with machine learning understand the logic of a Random Forest, we visualized one of the individual decision trees used in the ensemble (Figure 3). This tree is a truncated version of a much larger forest but offers a simplified view of how the model splits data based on key features.

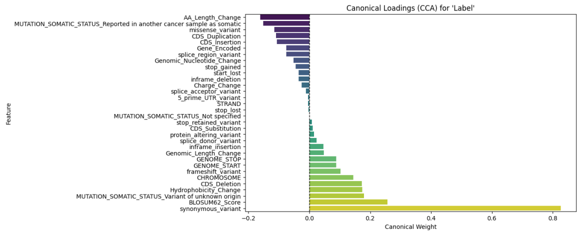


Figure 2: Canonical loadings from Canonical Correlation Analysis (CCA) showing feature relationships with the target label.

At each node, the model evaluates a condition (e.g., `Hydrophobicity_Change` < -0.004). The "Gini" value indicates node impurity—how mixed the class labels are at that node. A Gini of 0 means all samples belong to a single class, while 0.5 indicates maximum uncertainty.

This tree shows how a small number of features can effectively segment the data, providing intuitive insight into the model's decisions.

Feature Attribution and Interpretability

We used SHAP (SHapley Additive Explanations) to understand the global contribution of features to the model's predictions. In Figure 4, red dots represent high feature values and blue dots represent low values. The x-axis reflects each feature's effect on the model's output.

`synonymous_variant`, `BLOSUM62_Score`, and `AA_Length_Change` emerge as dominant features, echoing both the CCA analysis and the internal feature importance rankings of the Random Forest. Notably, high values of `synonymous_variant` are associated with benign outcomes, while lower BLOSUM scores (representing more disruptive amino acid changes) increase the probability of pathogenicity.

LIME was used to examine individual predictions (Figure 5). For one representative mutation, LIME highlighted similar features—`BLOSUM62_Score`, `splice_region_variant`, and `Hydrophobicity_Change`—validating the global SHAP trends at the local level.

We also explored feature interaction effects using partial dependence plots (PDPs), shown in Figure 6. PDPs illustrate how changes in a single feature, holding all others constant, affect the predicted probability of a mutation being pathogenic. These plots revealed nonlinear patterns—for instance, prediction probability increased sharply as `Hydrophobicity_Change` approached zero, suggesting threshold-like behavior in protein disruption.

Taken together, these interpretability tools provide a robust and coherent picture. SHAP and LIME reveal consistent key drivers, CCA supports their correlation with labels, and PDPs explain how these features modulate predictions. This triangulation enhances confidence in the model and affirms the biological relevance of the features used.

Discussion

The results of this study largely met and, in some respects, exceeded initial expectations. Based on the background literature and prior work on cancer genomics, it was anticipated that features such as amino acid change, mutation type, and biochemical shifts would provide meaningful signals for classifying mutation pathogenicity [1, 7]. This was strongly supported by the SHAP and feature importance results, which identified variables like `synonymous_variant`, `BLOSUM62_Score`, and `AA_Length_Change` as primary drivers of the model's predictions.

The high ROC-AUC score (0.9933) and robust F1-score (0.9804) confirm that the Random Forest model was able to learn complex patterns in the data with impressive generalization capacity. These findings align well with expectations derived from the known impact of amino acid substitutions on protein function, particularly when accounting for hydrophobicity and conservation through BLOSUM scoring [15, 16].

In terms of interpretability, the decision to prioritize Random Forests over more complex architectures proved beneficial. While a deep neural network showed slightly better numerical performance during early tests, its opacity limited its utility for this project's main goal: interpretable and actionable insights. Tools like SHAP, LIME, and partial dependence plots enabled a thorough understanding of feature effects, validating the utility of ensemble tree-based methods in biomedical contexts [10, 11].

The quality of the analysis is strengthened by the structured feature engineering strategy. Instead of treating genomic fields as raw text, we extracted structured features reflecting mutation types and nucleotide changes. This enhanced the biological plausibility of the model and mitigated overfitting by reducing reliance on noisy raw text formats.

Despite the strengths, there are limitations to the analysis. One key concern is the reliance on label quality from ClinVar. As noted in previous studies, many variants remain classified as "uncertain significance," and curation inconsistencies can introduce bias [8]. Additionally, while the use of COSMIC and ClinVar allowed for robust dataset construction, the exclusion of variants with unknown pathogenicity may have led

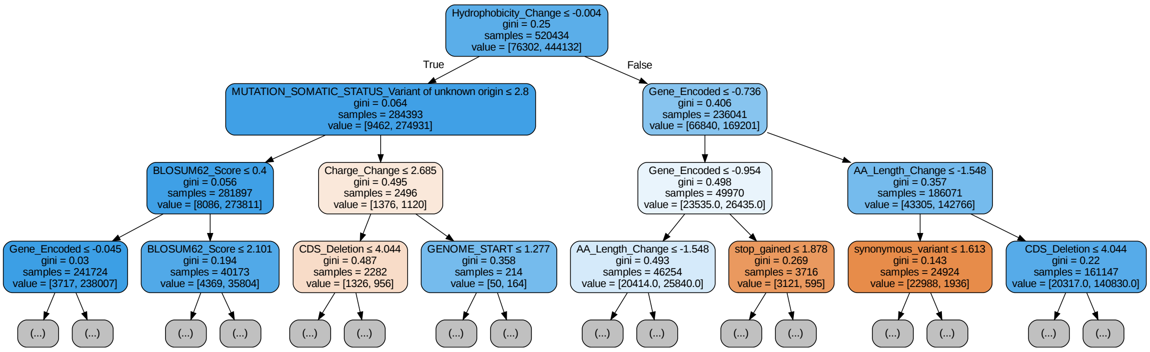


Figure 3: Truncated decision tree from the Random Forest ensemble.

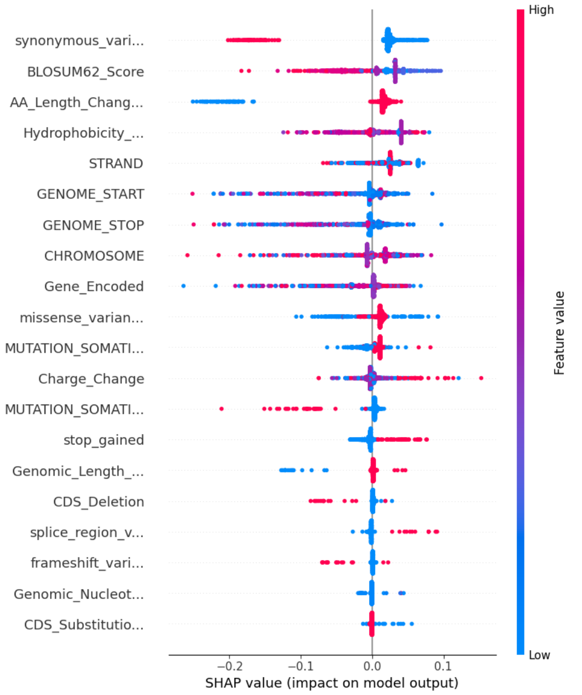
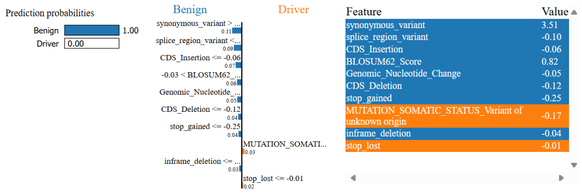


Figure 4: SHAP summary plot showing global feature impact across the dataset.



Feature	Value
synonymous_variant	3.51
splice_region_variant	-0.10
CDS_Insertion	-0.06
BLOSUM62_Score	0.82
Genomic_Nucleotide_Change	-0.05
CDS_Deletion	-0.12
stop_gained	-0.25
MUTATION_SOMATIC_STATUS_Variant of unknown origin	-0.17
inframe_deletion	-0.04
stop_lost	-0.01

Figure 5: LIME visualization for a single prediction, showing local feature impact.

to underrepresentation of borderline or ambiguous mutations that exist in real clinical scenarios.

Another limitation is the lack of 3D structural data integration. Although the use of biochemical features like hydrophobicity offers a proxy for structural impact, direct incorporation of protein structural context (e.g., through PDB-mapped coordinates) could offer further precision [?]. However, this was set aside due to the computational burden and limited structural coverage.

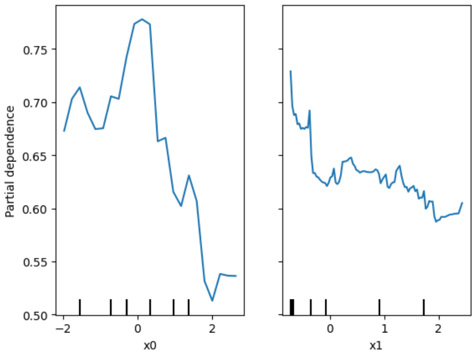


Figure 6: Partial dependence plots showing marginal effects of top features.

From a biological standpoint, the model's top features are consistent with molecular mechanisms known to influence pathogenicity. Synonymous mutations, by definition, do not alter amino acid sequences and are often benign. Conversely, missense mutations, particularly those with high substitution penalties (low BLOSUM scores), are more likely to disrupt protein function [4, 5]. This alignment between empirical findings and biological theory supports the model's reliability.

Overall, the model offers a highly interpretable and biologically grounded approach to variant classification. Its application could be valuable in prioritizing candidate mutations for experimental validation or clinical consideration, particularly in resource-limited settings where expert review is not always feasible.

Conclusion

This study presents an interpretable and effective machine learning framework for predicting the pathogenicity of somatic cancer mutations. By integrating curated datasets from COSMIC and ClinVar with biologically meaningful features—such as BLOSUM62 substitution scores, hydrophobicity changes, and mutation types—the model successfully captured the complex relationship between mutation characteristics and clinical impact.

The Random Forest classifier demonstrated excellent performance and consistency with known biological principles, while tools like SHAP, LIME, and partial dependence plots provided transparent insights into model behavior. Our results reinforce the potential of well-designed, interpretable machine learning approaches in supporting clinical genomics and cancer research.

Future directions could include integrating 3D structural data, expanding to multi-class classification tasks, or adapting the pipeline to new variant databases. Nonetheless, the current work already establishes a strong and explainable foundation for computational mutation interpretation.

Additional Material

PCA Explained Variance

Table 2: Explained variance ratio of the first 10 PCA components.

Component	Variance Explained
PC1	0.1128
PC2	0.0923
PC3	0.0740
PC4	0.0656
PC5	0.0512
PC6	0.0483
PC7	0.0440
PC8	0.0378
PC9	0.0375
PC10	0.0316

Top Feature Importances from Random Forest

Table 3: Top 10 features by importance score from the trained Random Forest.

Feature	Importance
synonymous_variant	0.1428
AA_Length_Change	0.1420
GENOME_START	0.1217
GENOME_STOP	0.1161
missense_variant	0.0873
Gene_Encoded	0.0795
Hydrophobicity_Change	0.0778
BLOSUM62_Score	0.0607
CHROMOSOME	0.0558
Variant	0.0217

References

[1] Vogelstein, B., et al. Cancer genome landscapes. *Science*, 2013.

[2] Hanahan, D. and Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell*, 2011.

[3] Manolio, T.A., et al. Finding the missing heritability of complex diseases. *Nature*, 2009.

[4] Ng, P.C., and Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 2003.

[5] Adzhubei, I.A., et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 2010.

[6] Choi, Y., et al. Predicting the functional effect of amino acid substitutions and indels. *Bioinformatics*, 2012.

[7] Tate, J.G., et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 2019.

[8] Landrum, M.J., et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 2018.

[9] Breiman, L. Random forests. *Machine Learning*, 2001.

[10] Lundberg, S.M., and Lee, S.I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017.

[11] Ribeiro, M.T., et al. *Why Should I Trust You?* Explaining the Predictions of Any Classifier. *KDD*, 2016.

[12] Harris, C.R., et al. Array programming with NumPy. *Nature*, 2020.

[13] McKinney, W. Data Structures for Statistical Computing in Python. *Proc. of SciPy*, 2010.

[14] Pedregosa, F., et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.

[15] Henikoff, S. and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 1992.

[16] Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 1982.

[17] Richards, S., et al. Standards and guidelines for the interpretation of sequence variants. *Genetics in Medicine*, 2015.

[18] Hardoon, D.R., et al. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 2004.

[19] Lipton, Z.C. The Mythos of Model Interpretability. *Communications of the ACM*, 2018.