Comparative Evaluation of PolyPhen-2, SIFT, and a BLOSUM62-based Model for Predicting the Functional Impact of Genetic Mutations

Andres Aranguren (2839590), Ahmed Fatta (2816370), Dirk de Boer (2663349), and Eleni Liarou (2868489)

Vrije Universiteit van Amsterdam, 2 September 2024

Abstract. [216]Predicting the functional impact of genetic mutations is essential for understanding disease mechanisms and developing therapeutic strategies. Accurate tools are needed to distinguish pathogenic from benign variants to guide clinical and research efforts. This study aims to compare the accuracy of three commonly used prediction tools—PolyPhen-2, SIFT, and a BLOSUM62-based baseline model—in predicting the functional effect of genetic mutations. Using receiver operating characteristic (ROC) plots, we evaluated the predictive performance of each tool by calculating the AUC for three datasets of varying sizes and publication dates. The tools were benchmarked against ClinVar annotations to assess their ability to classify mutations as either pathogenic or benign. PolyPhen-2 consistently demonstrated the highest accuracy across datasets, followed closely by SIFT. The baseline model, while functional, performed significantly lower, with AUC values close to random classification. PolyPhen-2 and SIFT are effective for predicting the impact of mutations, with PolyPhen-2 being slightly more accurate. The baseline model, while informative, was significantly less predictive. These findings highlight the importance of using advanced computational tools for genetic variant classification, and future research should explore integrating additional data sources to improve prediction accuracy further.

1 Introduction [526]

Genetic mutations, particularly single nucleotide polymorphisms (SNPs), play a crucial role in various human diseases. Understanding the functional impact of these mutations is essential for predicting their potential pathogenicity, aiding in diagnosis, and guiding therapeutic interventions [Consortium, 2010]. Among the many challenges in genomics, distinguishing between benign and deleterious mutations is critical for improving our knowledge of disease mechanisms and improving patient care. To address this, computational tools have been developed to predict whether a mutation is likely to affect protein function and contribute to disease.

SNPs represent the most common type of genetic variation, where a single nucleotide is altered in the DNA sequence. These changes can lead to alterations in protein structure and function, which can have significant biological implications, especially when the SNP occurs in a coding region [Brookes, 1999]. Given the vast number of SNPs in the human genome, experimental validation of each variant's impact is impractical. Thus, computational prediction tools, such as SIFT and PolyPhen-2, have been developed to assess the potential consequences of these mutations.

SIFT (Sorting Intolerant From Tolerant) predicts whether an amino acid substitution will affect protein function based on evolutionary conservation. The premise of SIFT is that functionally important residues tend to be conserved across species; thus, mutations at these positions are likely to be deleterious [Ng and Henikoff, 2003]. On the other hand, PolyPhen-2 (Polymorphism Phenotyping v2) evaluates amino acid substitutions by considering structural and comparative features, such as the physical properties of the residue change and evolutionary conservation across homologous sequences [Adzhubei et al., 2010]. Unlike SIFT, PolyPhen-2 incorporates additional features and leverages machine learning techniques, which makes it a more sophisticated tool. Given its more comprehensive approach, we expect that PolyPhen-2 will outperform SIFT in predicting the functional impact of mutations, as it can consider a wider range of biological factors.

A critical aspect of using these prediction tools is evaluating their performance through benchmarking against reliable datasets. ClinVar, a widely used database that provides clinical annotations for genetic variants, serves as a gold standard for assessing the accuracy of mutation impact predictions [Landrum et al., 2016]. Without accurate benchmarking, predictions may lead to misinterpretation, influencing both research conclusions and clinical decisions. Therefore, reliable comparisons of these tools are essential to understand their strengths and limitations and to identify which is most suitable for clinical or research applications.

In this report, we benchmark and compare the performance of two widely used prediction tools—SIFT and PolyPhen-2—alongside a baseline model based on the BLOSUM62 substitution matrix. The BLOSUM62 matrix is a traditional method that uses substitution frequencies observed in evolutionary related sequences to infer the potential impact of amino acid changes [Henikoff and Henikoff, 1992]. By using ClinVar as a reference dataset, we aim to assess how well these tools predict the functional consequences of SNPs, evaluating them through Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) metric.

Our research question is: *How accurately do PolyPhen-2, SIFT, and a BLOSUM62-based baseline model predict the functional impact of SNPs, and how do their performances across various datasets compare when benchmarked against the ClinVar dataset?* This report aims to provide critical insights into the reliability of these prediction tools, offering valuable guidance for their application in both research and clinical genomics.

2 Methods and Data [596]

2.1 Benchmarking Strategy

The benchmarking strategy, as shown in Figure 1, involves comparing the predictions from SIFT, PolyPhen-2, and the baseline model against the ClinVar dataset. The goal is to evaluate how well each tool identifies pathogenic and benign mutations. The workflow starts by building the baseline model using BLOSUM62. The prediction tools SIFT and PolyPhen-2 were run on the benchmark data to produce scores for each mutation. Obtained scores were then used to generate ROC plots to evaluate the sensitivity (TPR) and specificity (1 - FPR) of each tool. Lastly, the AUC was used to assess overall performance.

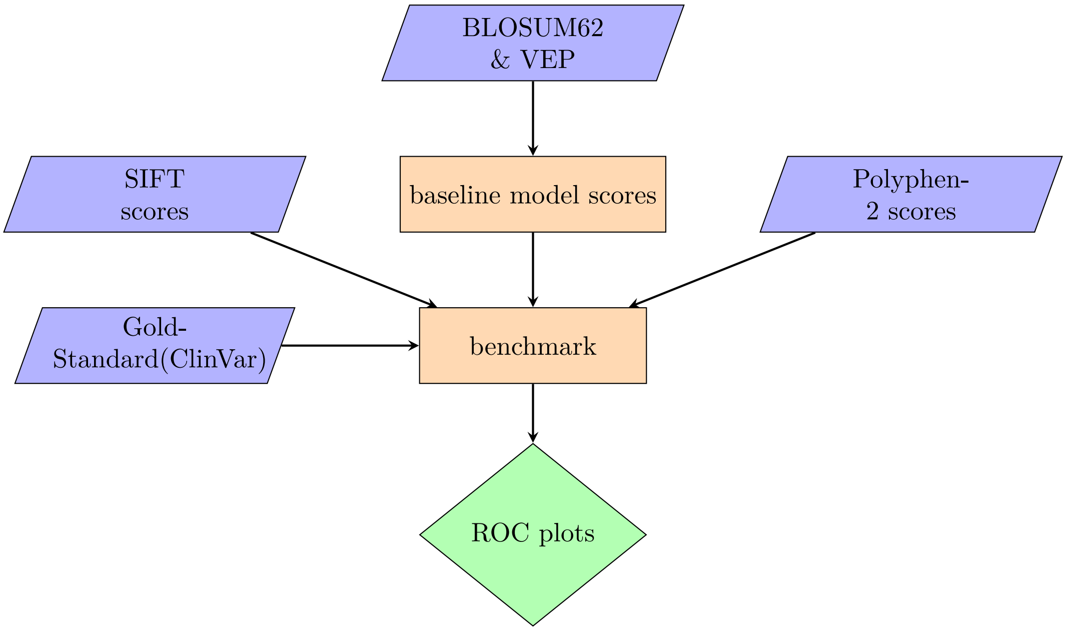


Figure 1. Flow chart. First, the BLOSUM62 matrix is integrated, together with the VEP to create the baseline model. This baseline model is benchmarked together with the SIFT and Polyphen-2 scores to create the ROC plot.

2.2 Properties of Data

The gold standard data are sourced from ClinVar, which provides clinical significance annotations for each mutation, classifying them as either pathogenic or benign based on their HGVS ID. Each HGVS dataset includes a reference ID and a description of the variant sequence, followed by the mutated nucleotide at this position. Mutation impact predictions from SIFT and PolyPhen-2 are also included, providing scores for each mutation based on the HGVS IDs. Based on our research proposal the datasets we are using to assess the efficiency of each prediction tool are the small dataset with 100 mutations

(HGVS 2020 small), the big dataset with 1200 mutations (HGVS 2020 big) and a dataset with 400 mutations from 2014 (HGVS 2014). All of them map the reference genome GRCh38. Additionally, we use a Variant Effect Predictor (VEP) file to identify the reference and mutated amino acids for each HGVS ID. The BLOSUM62 matrix, stored in a text file, is utilized to build the baseline prediction model.

2.3 Description of Scores

A SIFT score is a normalized probability of observing the new amino acid at that position and ranges from 0 to 1. SIFT uses a threshold, where variants scoring between $0 \leq 0.05$ can be confidently predicted as deleterious, and scores between $0.05 < 1$ are predicted to be benign [Ng and Henikoff, 2003]. PolyPhen-2 has an inverted scoring system that ranges from 0 to 1, with the following default threshold values (0.0 - 0.15) variant will be classified as benign, (0.15 - 0.85) possibly damaging variant, suggests a potential substitution that affects protein function. (0.85 - 1.0) range suggests strong evidence of damaging variant likely to impair protein function [Adzhubei et al., 2010]. Opposed to SIFT scores, a score of zero means that the system predicts the mutation to be benign and scores closer to one indicate a damaging effect on protein function. The baseline score is derived from the substitution matrix BLOSUM62, with higher scores indicating more common and desired substitutions, and lower scores indicating rarer and potentially harmful changes.

2.4 Metrics

ROC (Receiving Operating characteristic curve) was used as primary metric to evaluate the models' performance in binary classification across different thresholds. The ROC plots two main metrics, **True positive rate** (TPR) on the Y-axis and **False positive rate** (FPR) on the x-axis, thus the ROC is plotting the statistical power function in terms of the FPR. Each point on the ROC represents a threshold for classifying samples as positive (e.g., pathogenic) or negative (e.g., benign), balancing true and false positives. The x-axis shows the FPR, while the y-axis displays the TPR.

The Area Under the Curve (AUC) quantifies the model's overall performance, ranging from 0 to 1. A value of 1 indicates perfect classification, 0.5 suggests random guessing, and values below 0.5 imply worse-than-random performance. The AUC is approximated using the linear trapezoid method.

$$AUC = \frac{1}{2}(TPR_i + TPR_{i-1}) * (FPR_i - FPR_{i-1}) \tag{1}$$

3 Results [869]

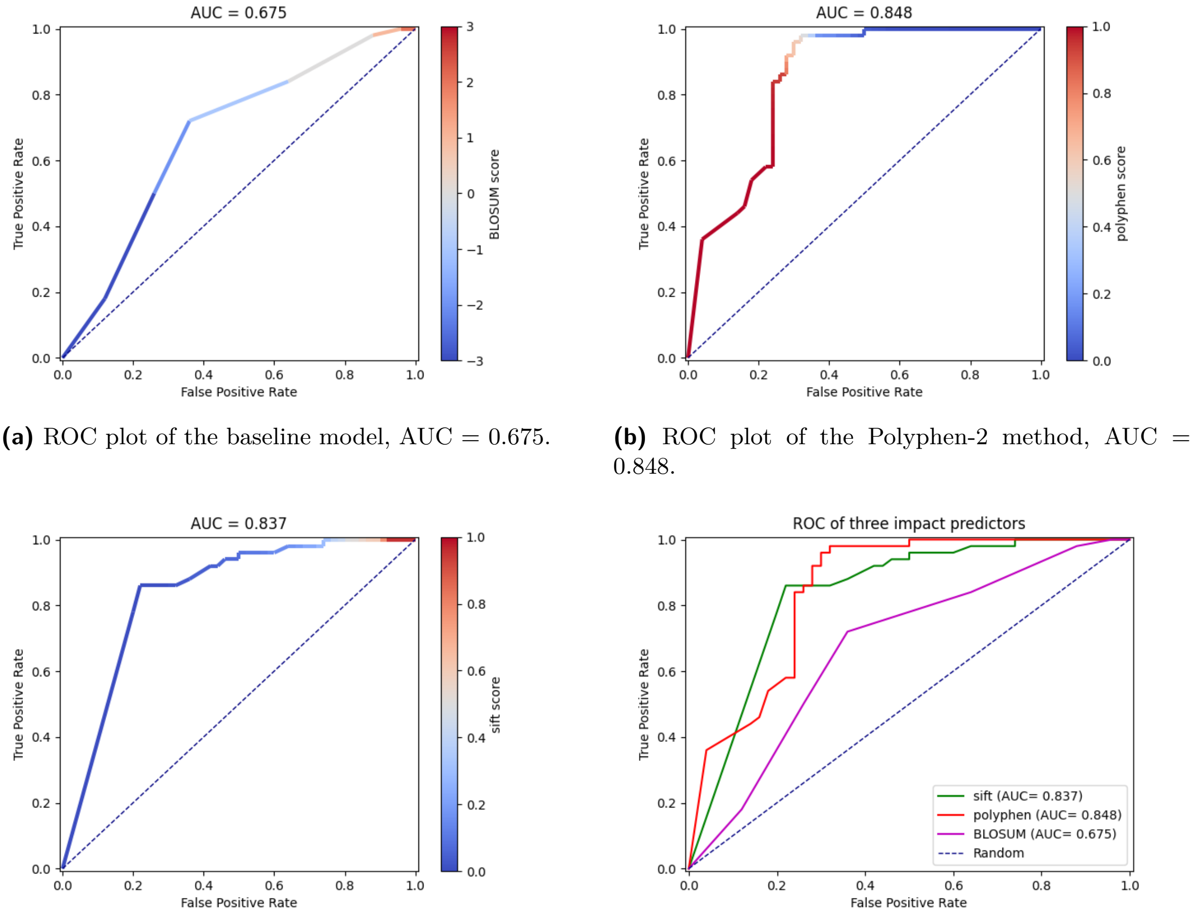
3.1 Research Goal and Methods

In this project, we aim to test how accurately different tools—PolyPhen-2, SIFT, and a baseline model built using the BLOSUM62 matrix—can predict the impact of amino acid substitutions on protein function. We evaluate their accuracy by comparing predictions against ClinVar's clinically validated labels (pathogenic or benign) for mutations. To do this, we use three datasets **HGVS_2020_small**, **HGVS_2020_big**, **HGVS_2014** described in section (2.2) and measure performance by plotting the ROC curves and calculating AUC values for each method. This allows us to assess prediction accuracy and determine how well each tool distinguishes between pathogenic and benign mutations.

3.2 ROC plots

The ROC curves were generated over several thresholds for each classification model. The ROC, thus allows to evaluate the performance of these methods in distinguishing between benign and pathogenic classes based on their respective score outputs. Thresholds were systematically varied and at each threshold the **TPR** and **FPR** were calculated to obtain the point coordinates to plot the curves. The different threshold values used to obtain the points in the curve are presented by the color gradient where blue and red indicate a low and higher threshold value respectively.

Below can be found the 4 ROC plots for each of the dataset. Figure 2 corresponds to the HGVS 2020 small dataset, Figure 3 to the HGVS 2020 big dataset and lastly Figure 4 to the HGVS 2014 dataset. PolyPhen-2 has an opposite scoring interpretation from SIFT and the baseline model, so the impact of lowering the threshold differs between them. In Polyphen-2 lowering the threshold results in more variants being classified as ”pathogenic”, which increases both the TPR and the FPR. Hence more true pathogenic variants will be correctly identified (increasing TPR), but more benign variants will be mistakenly labeled as ”damaging”, thus increasing FPR. This behavior can be seen in Figure 2 (b) by the transition to warmer colors while increasing the threshold value. In contrast, SIFT and the baseline work in the opposite way. Lowering the threshold in these 2 methods results in more variants being classified as ”benign” which lowers both the TPR and FPR. More true benign variants will be correctly identified (decreasing FPR), but more pathogenic variants will also be incorrectly classified as ”deleterious” decreasing TPR. This behavior is reflected in Figure 2 (a), (c) having cooler colors at lower threshold values, resulting in higher confidence in correctly classifying the input variants.



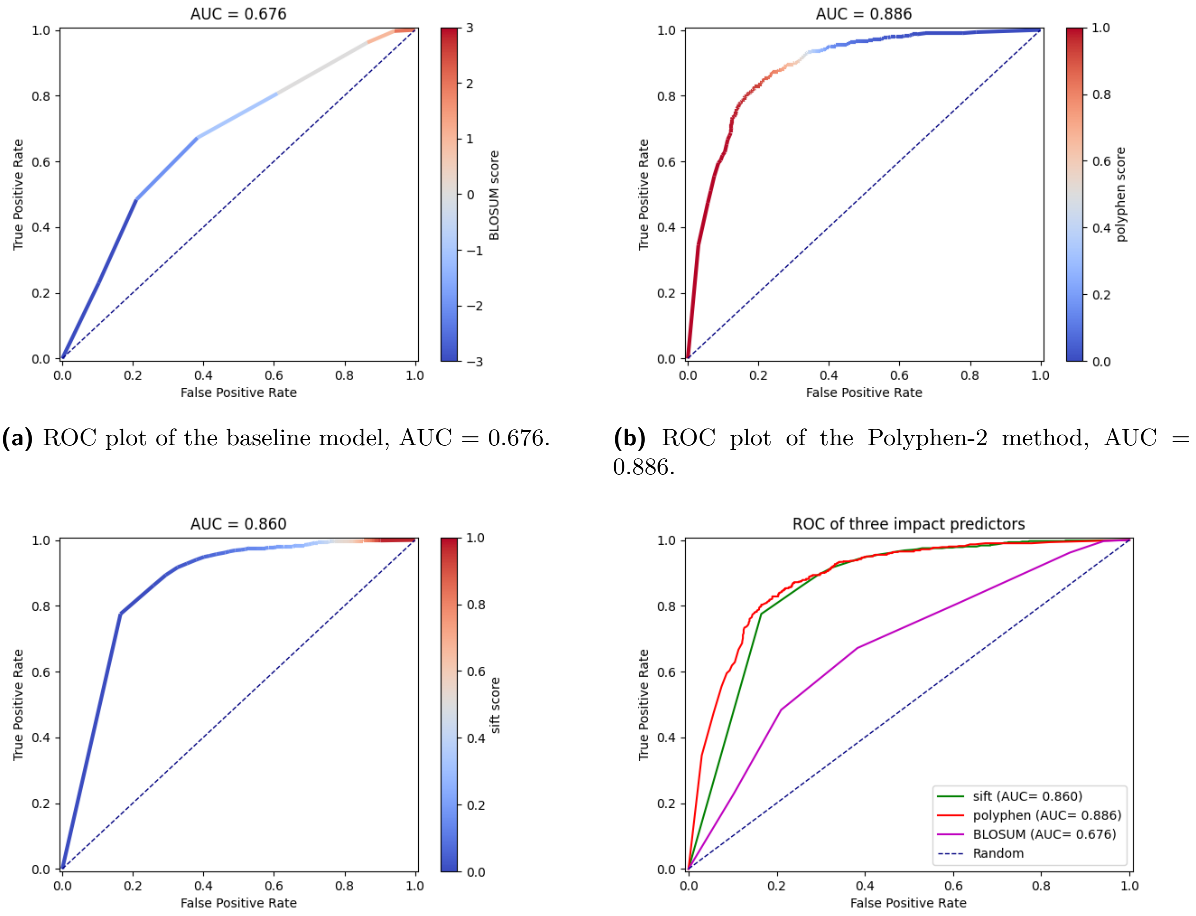
(a) ROC plot of the baseline model, AUC = 0.675.

(b) ROC plot of the Polyphen-2 method, AUC = 0.848.

(c) ROC plot of the SIFT method, AUC = 0.837

(d) AUC comparison of all models. Polyphen-2(red) achieved the highest score, followed by SIFT(green) and the base model(purple).

Figure 2. ROC plots for **small dataset**. (a) Baseline model, (b) PolyPhen-2, (c) SIFT, and (d) combined models with the random classifier for comparison. The x-axis represents the FPR, and the y-axis the TPR.



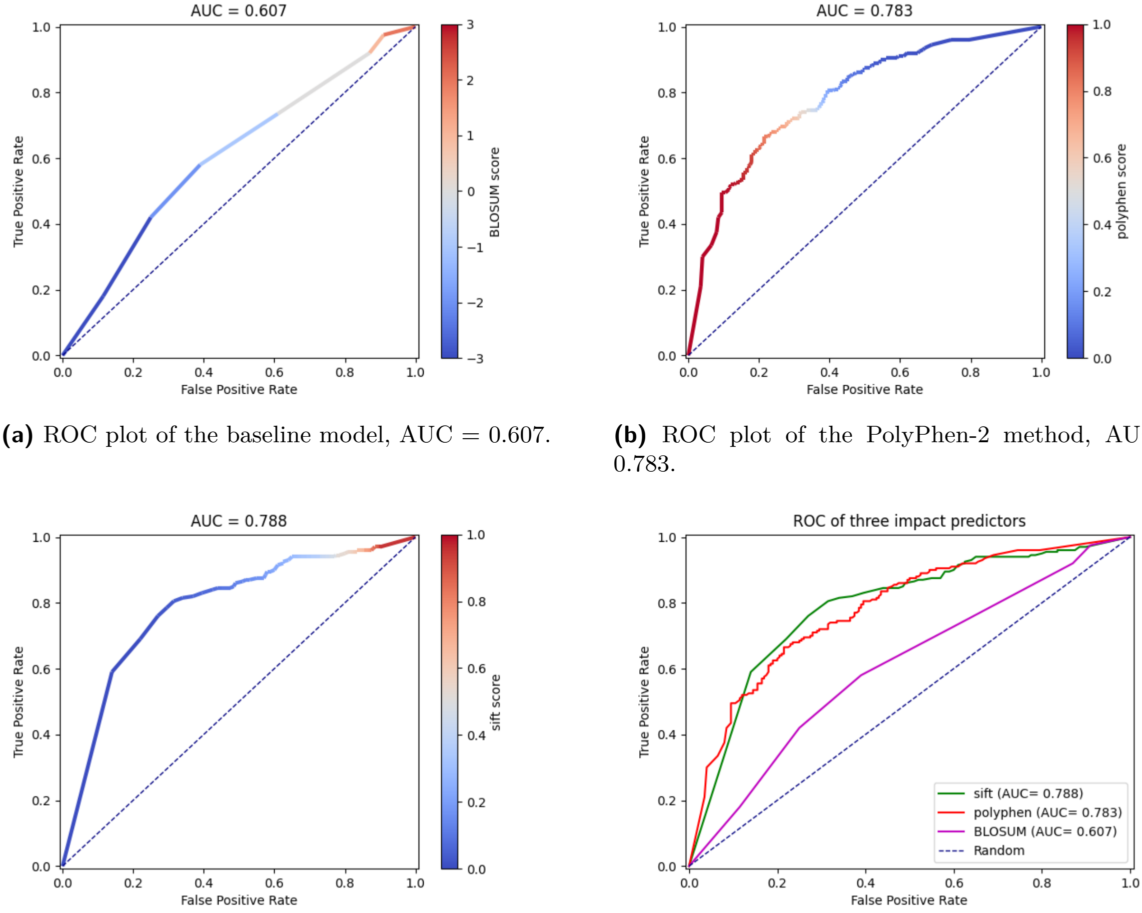
(a) ROC plot of the baseline model, AUC = 0.676.

(b) ROC plot of the Polyphen-2 method, AUC = 0.886.

(c) ROC plot of the SIFT method, AUC = 0.860

(d) AUC comparison of all models. Polyphen-2(red) achieved the highest score, followed by SIFT(green) and the base model(purple).

Figure 3. ROC plots for **big dataset**. (a) Baseline model, (b) PolyPhen-2, (c) SIFT, and (d) combined models with the random classifier for comparison. The x-axis represents the FPR, and the y-axis the TPR.



(a) ROC plot of the baseline model, AUC = 0.607.

(b) ROC plot of the PolyPhen-2 method, AUC = 0.783.

(c) ROC plot of the SIFT method, AUC = 0.788

(d) AUC comparison of all models. SIFT(green) achieved the highest score, followed by Polyphen-2(red) and the base model(purple).

Figure 4. ROC plots for **old dataset**. (a) Baseline model, (b) PolyPhen-2, (c) SIFT, and (d) combined models with the random classifier for comparison. The x-axis represents the FPR, and the y-axis the TPR.

3.3 Prediction Methods Comparison

In the small dataset in Figure 2, PolyPhen-2 and SIFT perform similarly, with PolyPhen-2 scoring an AUC of 0.848 and SIFT trailing by only 0.011, with an AUC of 0.837. The baseline model falls significantly behind, with an AUC of 0.675. In the big dataset in Figure 3, the performance gap remains clear. PolyPhen-2 slightly outperforms SIFT, with an AUC of 0.886, just 0.006 higher than SIFT's 0.860. The baseline model lags behind considerably with an AUC of 0.676. In the old dataset in Figure 4, PolyPhen-2 and SIFT show similar performance, achieving AUCs of 0.783 and 0.788, respectively. SIFT marginally outperforms PolyPhen-2 for the first time, while the baseline model again has the lowest predictive power, with an AUC of 0.607.

Table 1. AUC Values of the Three Considered Methods Across the Datasets

Method	**HGVS Small**	**HGVS Big**	**HGVS 2014**
Baseline	0.675	0.676	0.607
SIFT	0.837	0.860	0.788
PolyPhen-2	0.848	0.886	0.783

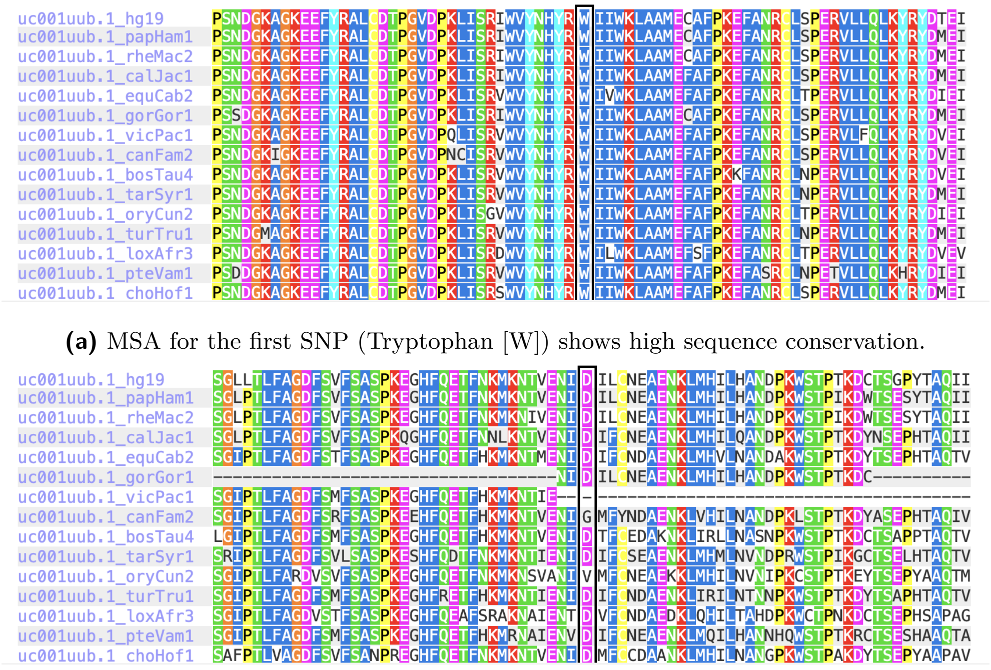
3.4 Results vs Initial Expectations

We initially expected PolyPhen-2 to outperform both SIFT and the baseline model, given its more comprehensive feature set. SIFT was anticipated to perform well but not as strongly as PolyPhen-2. The baseline model was not expected to match the predictive power of the other tools, as it relies on a simple substitution scoring matrix without accounting for additional features. This aligns with the results, as the baseline shows consistently low AUC across all datasets, confirming its limited predictive power. However, contrary to our expectations, the difference in performance between SIFT and PolyPhen-2 is smaller than anticipated, suggesting similar effectiveness in mutation prediction.

4 Use Cases [498]

In the Breast Cancer type 2 gene (BRCA2), the first SNP shows a transversion in the transcription process, where Guanine (G) is mutated into Cytosine (C). Consequently, Tryptophan (W) changes into Cysteine (C). The first SNP shows high level of sequence conservation, with the SNP itself being particularly conserved 5a. This indicates that it plays a crucial functional role in the protein, as highly conserved regions tend to be subject to strong evolutionary pressures that prevent disruptive mutations. In contrast, the second SNP is located in a region with a mix of both highly and less conserved amino acids, shown in figure 5b. This SNP is a transition from Adenine (A) into Guanine (G), causing a missense mutation from Aspartic acid (D) into Glycine (G). The presence of potential mutations in this area suggests that the region may tolerate variation, which aligns with the benign nature of the second SNP.

The high conservation observed for the first SNP supports its expected pathogenicity. Pathogenic mutations in highly conserved regions are likely to be detrimental to the organism, as these regions are preserved across species due to their functional importance.



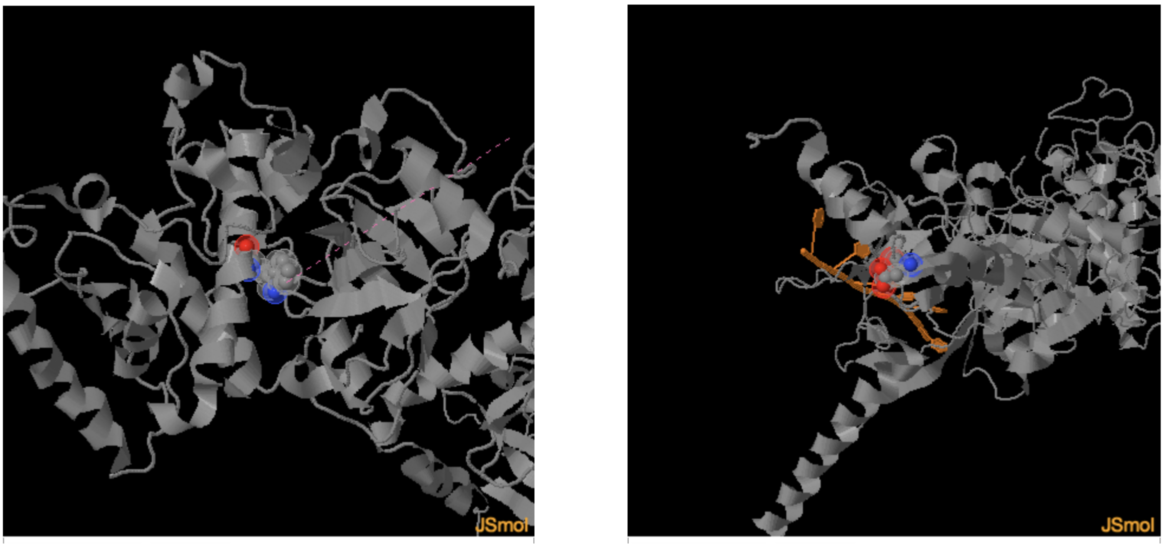
```
uc001uub.1_hg19     PSNDGKAGKEEFYRALCDTPGVDPKLISRIWVYNHYR W IIWKLAAMECAFPKEFANRCLSPERVLLQLKYRYDTEI
uc001uub.1_papHam1  PSNDGKAGKEEFYRALCDTPGVDPKLISRIWVYNHYR W IIWKLAAMECAFPKEFANRCLSPERVLLQLKYRYDMEI
uc001uub.1_rheMac2  PSNDGKAGKEEFYRALCDTPGVDPKLISRIWVYNHYR W IIWKLAAMECAFPKEFANRCLSPERVLLQLKYRYDMEI
uc001uub.1_calJac1  PSNDGKAGKEEFYRALCDTPGVDPKLISRIWVYNHYR W IIWKLAAMEFAFPKEFANRCLSPERVLLQLKYRYDMEI
uc001uub.1_equCab2  PSNDGKAGKEEFYRALCDTPGVDPKLISRVWVYNHYR W IVWKLAAMEFAFPKEFANRCLTPERVLLQLKYRYDIEI
uc001uub.1_gorGor1  PSSDGKAGKEEFYRALCDTPGVDPKLISRIWVYNHYR W IIWKLAAMECAFPKEFANRCLSPERVLLQLKYRYDMEI
uc001uub.1_vicPac1  PSNDGKAGKEEFYRALCDTPGVDPQLISRVWVYNHYR W IIWKLAAMEFAFPKEFANRCLSPERVLFQLKYRYDVEI
uc001uub.1_canFam2  PSNDGKIGKEEFYRALCDTPGVDPNCISRVWVYNHYR W IIWKLAAMEFAFPKEFANRCLSPERVLLQLKYRYDVEI
uc001uub.1_bosTau4  PSNDGKAGKEEFYRALCDTPGVDPKLISRVWVYNHYR W IIWKLAAMEFAFPKKFANRCLNPERVLLQLKYRYDVEI
uc001uub.1_tarSyr1  PSNDGKAGKEEFYRALCDTPGVDPKLISRVWVYNHYR W IIWKLAAMEFAFPKEFANRCLNPERVLLQLKYRYDMEI
uc001uub.1_oryCun2  PSNDGKAGKEEFYRALCDTPGVDPKLISGVWVYNHYR W IIWKLAAMEFAFPKEFANRCLTPERVLLQLKYRYDIEI
uc001uub.1_turTru1  PSNDGMAGKEEFYRALCDTPGVDPKLISRVWVYNHYR W IIWKLAAMEFAFPKEFANRCLNPERVLLQLKYRYDMEI
uc001uub.1_loxAfr3  PSNDGKAGKEEFYRALCDTPGVDPKLISRDWVYNHYR W ILWKLAAMEFSFPKEFANRCLTPERVLLQLKYRYDVEV
uc001uub.1_pteVam1  PSDDGKAGKEEFYRALCDTPGVDPKLISRVWVYNHYR W IIWKLAAMEFAFPKEFASRCLNPETVLLQLKHRYDIEI
uc001uub.1 choHof1  PSNDGKAGKEEFYRALCDTPGVDPKLISRSWVYNHYR W IIWKLAAMEFAFPKEFANRCLSPERVLLQLKYRYDMEI
```

(a) MSA for the first SNP (Tryptophan [W]) shows high sequence conservation.

```
uc001uub.1_hg19     SGLLTLFAGDFSVFSASPKEGHFQETFNKMKNTVENI D ILCNEAENKLMHILHANDPKWSTPTKDCTSGPYTAQII
uc001uub.1_papHam1  SGLPTLFAGDFSVFSASPKEGHFQETFNKMKNTVENI D ILCNEAENKLMHILHANDPKWSTPIKDWTSESYTAQII
uc001uub.1_rheMac2  SGLPTLFAGDFSVFSASPKEGHFQETFNKMKNIVENI D ILCNEAENKLMHILHANDPKWSTPIKDWTSESYTAQII
uc001uub.1_calJac1  SGLPTLFAGDFSVFSASPKQGHFQETFNNLKNTVENI D IFCNEAENKLMHILQANDPKWSTPTKDYNSEPHTAQII
uc001uub.1_equCab2  SGIPTLFAGDFSTFSASPKEGHFQETFHKMKNTMENI D IFCNDAENKLMHVLNANDAKWSTPTKDYTSEPHTAQTV
uc001uub.1_gorGor1  -----------------------------------NI D ILCNEAENKLMHILHANDPKWSTPTKDC---------
uc001uub.1_vicPac1  SGIPTLFAGDFSMFSASPKEGHFQETFHKMKNTIE-- - --------------------------------------
uc001uub.1_canFam2  SGIPTLFAGDFSRFSASPKEEHFQETFHKMKNTVENI G MFYNDAENKLVHILNANDPKLSTPTKDYASEPHTAQIV
uc001uub.1_bosTau4  LGIPTLFAGDFSMFSASPKEGHFQETFHKMKNTVENI D TFCEDAKNKLIRLLNASNPKWSTPTKDCTSAPPTAQTV
uc001uub.1_tarSyr1  SRIPTLFAGDFSVLSASPKESHFQDTFNKMKNTIENI D IFCSEAENKLMHMLNVNDPRWSTPIKGCTSELHTAQTV
uc001uub.1_oryCun2  SGIPTLFARDVSVFSASPKEGHFQETFNKMKNSVANI V MFCNEAEKKLMHILNVNIPKCSTPTKEYTSEPYAAQTM
uc001uub.1_turTru1  SGIPTLFAGDFSMFSASPKEGHFRETFHKMKNTIENI D IFCNDAENKLIRILNTNNPKWSTPTKDYTSAPHTAQTV
uc001uub.1_loxAfr3  SGIPTLFAGDVSTFSASPKEGHFQEAFSRAKNAIENT D VFCNDAEDKLQHILTAHDPKWCTPNKDCTSEPHSAPAG
uc001uub.1_pteVam1  SGIPTLFAGDFSMFSASPKEGHFQETFHKMRNAIENV D IFCNEAENKLMQILHANNHQWSTPTKRCTSESHAAQTA
uc001uub.1 choHof1  SAFPTLVAGDFSVFSANPREGHFQETFNKMKNTVENI D MFCCDAANKLMHILNANGPKWSTPAKDYTSEPYAAPAV
```

(b) MSA for the second SNP (Aspartic acid [D]) shows lower sequence conservation than the first SNP.

Figure 5. Multiple Sequence Alignments (MSA) for BRCA2 gene, comparing two SNPs.



(a) 3D image showing the pathogenic mutation (red) in the protein core, potentially disrupting folding and function.

(b) 3D image showing the benign mutation (red) on the protein's surface, likely not affecting its function.

Figure 6. 3D images of the BRCA2 gene, showing the location of the mutation of the two SNPs in red.

4.1 Protein structure

The first SNP, predicted to be pathogenic, is located deep within the protein's core, surrounded by tightly folded alpha helices. As Tryptophan holds a hydrophobic side chain, this may suggest that this amino acid is not meant to interact with the surrounding structures. However, the replaced Cysteine can form sulfuric bridges with neighbouring Cysteine amino acids. This structural environment suggests that this could significantly disrupt protein stability or function. The exact location of it is shown in figure 6a, with the mutation highlighted in red. On the other hand, figure 6b shows that the second, benign SNP is located in between two alpha helices, closer to the protein's outer surface, where structural changes are less likely to have severe functional consequences. Here, Aspartic acid is a negatively charged amino acid, mutated in the flexible Glycine, that only holds a hydrogen atom as a side chain. As this mutation is predicted to be benign, the Aspartic acid probably isn't crucial for the protein's folding process.

VEP Prediction Agreement Based on the sequence conservation and structural analysis, the VEP prediction aligns with our findings. The conservation and structural placement of the first SNP make it highly susceptible to deleterious effects, supporting its pathogenic classification. In contrast, the second SNP, located in a less conserved and structurally flexible region, is more likely to be benign.

5 Discussion [867]

5.1 Difference in Performance

SIFT primarily relies on sequence conservation and the severity of amino acid changes [Vaser, 2016], meanwhile PolyPhen-2 incorporates additional predictive features, including three structure-based factors alongside eight sequence-based ones. This allows PolyPhen-2 to assess not only how conserved a residue is but also the structural and functional impact of a mutation within the protein [Adzhubei et al., 2010]. Furthermore, PolyPhen-2 uses a naive Bayes classifier trained on two diverse datasets, HumDiv and HumVar, which cover a wide range of both damaging and nondamaging mutations [Sunyaev et al., 2001]. This broader feature set and dataset coverage make PolyPhen-2 a more robust tool for predicting the pathogenicity of mutations, while SIFT's reliance on sequence homology alone provides a narrower context for its predictions [Sim et al., 2012]. As a result, PolyPhen-2 was expected to deliver more accurate predictions compared to SIFT. However, as the scores show in Table 1, SIFT and PolyPhen-2 performed almost identically across all datasets. This suggests that while structural information adds value, sequence conservation remains a highly informative feature for predicting the impact of mutations, allowing SIFT to perform at a level similar to PolyPhen-2 in these cases. The baseline model, built using the BLOSUM62 substitution matrix [Henikoff and Henikoff, 1992], was expected to perform poorly compared to SIFT and PolyPhen-2. BLOSUM62 primarily calculates the likelihood of amino acid substitutions based on observed substitutions, but it does not account for the structural or functional impact of these mutations. As a result, this model focuses more on estimating the probability of a mutation occurring rather than predicting its pathogenicity [Dayhoff et al., 1978]. The low AUC scores across all datasets are consistent with this limitation, confirming that the baseline model lacks the predictive power to assess whether a mutation is benign or deleterious [Kumar et al., 2009].

5.2 Threshold values vs Default values

The SIFT default thresholds are coherent with the obtained results in all three datasets. Threshold values < 0.05 give the optimum trade-off between TPR and FPR, maximizing TPR while minimizing FPR, allowing confident prediction of pathogenic variants while avoiding type 1 error in misclassifying pathogenic variants as benign. At the 0.05 threshold value the model captures the most of the true positives while maintaining a low false positive rate, suggesting that the default threshold differentiates effectively damaging mutations.

Based on the obtained results for the three datasets the performance of PolyPhen-2 improves significantly around the mid-to-high range of the PolyPhen-2 scores, which aligns with its default threshold of 0.85 for probably damaging, thus PolyPhen-2 correctly identifies a substantial portion of true positives at this score level. Moving further to the right the FPR model performance diminishes when using lower threshold scores closer to benign.

5.3 Effect of more benchmark data on the ROC and AUC

As demonstrated in Table 1, increasing the size of benchmark datasets for testing predictive models significantly enhances the evaluation of each tool's ability to distinguish between pathogenic and benign mutations [Fawcett, 2006]. A larger dataset enables a more comprehensive comparison of predictions across a broader spectrum of mutations, which improves the statistical robustness of the analysis. This, in turn, ensures that sensitivity and specificity estimates are more reliable, resulting in more accurate performance assessments [Landrum et al., 2016].

The performance of SIFT and PolyPhen-2 improves significantly with the availability of larger benchmark datasets, as evidenced by the higher AUC values. This improvement stems from the additional data points that larger datasets provide, allowing for the generation of more thresholds for ROC analysis. Consequently, the ROC curves become smoother and more refined, offering a clearer picture of model performance [Hanley and McNeil, 1982]. The increased number of data points leads to a greater variety of TPR and FPR pairs, enhancing the precision of the AUC.

However, not all methods benefit equally from larger datasets. The baseline model, is constrained by a fixed number of substitution scores derived from the substitution matrix [Henikoff and Henikoff, 1992]. This limitation results in fewer thresholds for ROC plotting, yielding a more step-like ROC curve. In contrast, methods like SIFT and PolyPhen-2 generate a wider range of prediction scores, resulting in smoother and more informative ROC curves [Kumar et al., 2009].

5.4 Relative performance in all 3 datasets for each prediction method

5.4.1 Performance Across Datasets

The baseline model exhibits minimal improvement between the HGVS Small and HGVS Big datasets, indicating that larger dataset sizes do not significantly enhance its performance. However, it shows a notable decline when evaluated against the HGVS 2014 dataset, suggesting difficulties with its specific characteristics [Landrum et al., 2014]. In contrast, SIFT demonstrates clear performance gains as the dataset size increases, yet it also experiences a drop on the HGVS 2014 dataset, highlighting the challenges posed by older data. PolyPhen-2 benefits the most from larger datasets, showing substantial improvement, but similarly faces performance declines on the HGVS 2014 dataset while still remaining competitive.

5.4.2 Performance Across Methods

In the HGVS 2020 Small dataset, PolyPhen-2 slightly outperforms SIFT, with both tools significantly surpassing the baseline, showcasing their superior ability to predict mutations with smaller datasets [Adzhubei et al., 2010]. In the HGVS 2020 Big dataset, both SIFT and PolyPhen-2 improve, with PolyPhen-2 maintaining a slight lead, while the baseline shows almost no change, highlighting its limited utility compared to these advanced tools as dataset size increases. Conversely, in the HGVS 2014 dataset, both SIFT and PolyPhen-2 experience declines in performance but still outperform the baseline, with the performance gap between the two narrowing, suggesting that both methods face similar challenges with the older data [Vaser, 2016].

6 Conclusions [255]

This report set out to assess the accuracy of PolyPhen-2, SIFT, and a BLOSUM62-based baseline model in predicting the functional impact of genetic mutations, with ClinVar as the benchmark. Our results demonstrate that PolyPhen-2 consistently outperforms both SIFT and the baseline model across all datasets, confirming our hypothesis that PolyPhen-2's incorporation of structural and machine learning features would lead to superior predictive accuracy. SIFT also performs well, with only marginally lower AUC scores compared to PolyPhen-2, indicating that sequence conservation remains a strong predictor of pathogenicity. The baseline model, as expected, exhibited significantly lower performance due to its simpler design and lack of contextual features.

While PolyPhen-2 shows better overall accuracy, it is important to note that both PolyPhen-2 and SIFT struggle with older datasets, such as the 2014 ClinVar data, highlighting potential limitations when applied to outdated or incomplete data. Additionally, while PolyPhen-2 is a powerful tool, the differences in performance between it and SIFT are relatively small, suggesting that both tools can be reliably used in practical and clinical settings, depending on the specific needs of the study.

In a practical or medical context, these findings suggest that tools like PolyPhen-2 and SIFT are highly valuable for prioritizing SNPs for further experimental validation or clinical assessment. However, the choice of tool may depend on the dataset size and quality, with larger, more recent datasets providing more reliable predictions. Future work should focus on integrating even more diverse features into these models, improving their robustness across a wider range of genetic data.

Bibliography

Ivan A Adzhubei et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

Anthony J Brookes. Single nucleotide polymorphisms (snps). *Journal of Human Genetics*, 104:123–133, 1999.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

Margaret O Dayhoff, Robert M Schwartz, and Bryce C Orcutt. *A survey of new data and methods in the 1978 supplement to the atlas of protein sequence and structure*. National Biomedical Research Foundation, 1978.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.

Melissa J Landrum, Jennifer M Lee, George R Riley, Wooyoung Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 42(D1):D980–D985, 2014.

Melissa J Landrum et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 2016.

Pauline C Ng and Steven Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

Ngan L Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Gary Schneider, and Pauline C Ng. Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(W1): W452–W457, 2012.

Shamil Sunyaev, Vasily Ramensky, Igor Koch, Warren Lathe, Alexey S Kondrashov, and Peer Bork. Predicting deleterious amino acid substitutions. *Human molecular genetics*, 10(6):591–597, 2001.

Ngak Leng Vaser, Adusumalli. Sift missense predictions for genomes. *Protocol update*, VOL.11 NO.1(8): 9, 2016.

Member's Contribution

Table 2. Table of Contributions

	Coding	Abstract	Introduction	Methods	Results	Use Cases	Discussion	Conclusion
Eleni Liarou(2868489)	x			x	x		x	
Andres Aranguren (2839590)	x			x	x		x	
Dirk de Boer (2663349)				x	x	x		
Ahmed Fatta (2816370)		x	x				x	x