# Missense Variant Analysis and the Impact of Scoring Methods

BINF694 - Variant Analysis

A Manning Smith

May 20, 2025

## Introduction

The human genome contains roughly 3 terabytes of data, that's roughly a million pictures worth of data. That being said, our genome is has a large probability of carrying some issues, but amazingly most of the variants don't affect the human at all. This project aims to explore the genomic variants of an individual. Their genome was annotated via a tool called SeatleSeq and the data was utilized for research in this project.

The goal of this project is to understand the various types of scoring methods used to classify the variants and to describe the damage they may have on the function of the protein. Secondly, the goal is to examine 3 variants and discuss their associations with known publications and how they may effect the individual.

## Scoreing Methods

A single nucleotide variant (SNP). is a single nucleotide substitution at a specific genome position. These variants are the most common sequence variations [**R14**]. Our first goal was to understand how we can rank, score, classify these variations. The three scoring methods we will be using are PolyPhen, Grantham, and CADD. Each of these methods has a different approach to scoring the variants.

### PolyPhen

The PolyPhen-2 is a tool utilized for the annotation of SNPs. The foundation of the classifier is a "machine-learning method" [**L1**]. The tool works to define various amino acid replacements, by feeding "structure-based features of the the substitution site to a probabilistic classifier" [**L1**].

As a probability value that range of the score is from $0 - 1$ being that 0 has less of a probability of being damaging where as values closer to 1 have a higher probability of being damaging. The performance of the tool is said to achieve a true positive prediction of "92% and 73% on HumDiv and HumVar, respectively" [**R1**].

### GranthamScore

The Grantham score is a prediction, in an evolutionary sense, of the distance between two amino acids [**L2**]. The score takes into account the composition, polarity, and molecular volume [**R12**]. This score is a more straight forward calculation, with specific parameters affecting the score.

The score ranges from 5 to 215. A lower score, closer to 5, indicates a less significant substitution, while a high score represents that of a more critical substitution. It uses a matrix, $D$, to related the distances between amino acids, thus representing the severity of change between two amino acids.

### CADD

Combined Annotation Dependent Depletion (CADD) is a tool used to determine the harmfulness of a SNP. Different from our other two scores, CADD takes into account variants throughout the entire genome. The values is calculated via a machine learning model, "trained on a binary distinction between simulated de novo variants and variants that have arisen and become fixed in human populations..." [**R10**]. The value combines aspects like, conservation metrics, transcription factor binding, and genomic and epigenomic annotations, along with a few other parameters.

The score is commonly defined as "scaled C-scores" described as the "rank of each variant relative to all possible 8.6 billion substitutions" [**#L3**]. It ranges from $10 - 99$, represented as $10 * -\log \text{rank}$. A score above 20, is "predicted to be among the 1.0 most deleterious possible substitutions" in the genome [**L4**].

### Score Analysis

**Table 1** highlights the summary statistics of our variant observations for the three scores highlighted. There were a large amount of observations removed for having missing values. CADD had significantly less amount of observations removed, likely due to the calculation being applied to non-coding regions, unlike the other two measures.

Our scales align with that outlined above for each score respectively. The means or average of our scores represent that of less severe variants. This aligns with the health profile of the individual as the majority of the that variants are not significant.

**Figure 4.0** outline lines the distribution of our `complete_cases` data set, which is comprised off the variants that have values for all three of our scores. Our scores present significantly right skewed distributions aligning with the idea that the majority of the variants are not significant.

**Figure 5.0** outlines the correlations between the three different scores. There is a relatively high correlation between `PolyPhen vs CADD`, suggesting that these scores seem to agree with each other in the significance of the variant. Our little to non correlation between `PolyPhen vs Grantham` suggest that the amino acid substitutions generally don't have a high probability of impact. Lastly, `Grantham vs CADD` has no correlation suggesting that amino acid changes have no relationship with deleteriousness.

Based on these scoring patterns, we can now identify specific variants that highlight the need for further investigation, focusing on outliers that demonstrate potentially significant biological impact according to our methods.

## Variant Selection

There first step for selecting the variants of interest was to review the individuals profile via the Personal Genome Project. The individual was roughly 30 years of age representing a white, Hispanic/Latino, American Indian race and ethnicity with paternal decent form the Netherlands.

There were no apparent medical conditions or diseases. There was a genetic disorder Mastocytosis; however, upon investigating the variant data, there was not enough evidence to explore these variants based on the provided information. This mutation is commonly presented in the KIT gene but there were no results with this gene. There could be further evidence to explore with this mutation but for the purposes of this research this mutation was not explored.

The next step, was to examine the data itself. Based on the information learned about each score, the following filter was applied to the data:

- PolyPhen Score: Values $> 0.9$
- Grantham Score: Values $> 180$
- CADD Score : Values $> 20$

Note that these filters were `or` not `and` meaning that the variant only needed to meet one of the described criteria. These results can be seen in the `priorityVariants` sheet in the "MasterData.xlsx", linked here.

With these priority variants roughly 30 unique genes were discovered to have "significant" variations. With that the first variation was selected based on high PolyPhen score. The second variant had a high CADD score and PolyPhen score but a lower Grantham score. Lastly, the third variant had a high score for all three.

**Table 2** lists some basic information about the selected variants used in our analysis. These selected variants were selected to provide a well rounded approached to analyzing the SNPs. **Table 3** list the scores for the selected variants. The scores are relatively high, suggesting that there could be a significant impact on the protein function. The next step is to analyze the variants in more detail.

# Variant Analysis

## GALNTL5 - rsID: 6960270

The Polypeptide GalNAc transferase 15 (GALNTL5) plays an important role in the male reproductive system as it is mostly expressed in the testis and plays a role in sperm development.

### Molecular Impact

This variant is in the GALNTL5 gene located on chromosome 7 at position 151982987. The variant is a missense variant being heterogeneous with a change from `T` to `C`, changing the amino acid from CYS to ARG, at position 177/251 near the splice site.

Based on our reported scores for this variant there is evidence to suggest that there may be an impact based on this variant. Firstly, when looking at the PolyPhen score of 1.0 being the maximum value this suggests the substitution could have a damaging affect based on the structural and evolutionary impacts. Secondly, the Grantham score of 180 is relatively high, this score falls in the top 1% of results of Grantham scores with-in our data set. The high value suggests there could be significant impacts based on the amino acid substitution from Cysteine, a smaller neutral amino acid to arginine, a large positively charged amino acid [**L5**]. Third, the CADD score of 20.4 marks this variant in the top 1 of deleterious variants in the human genome. Lastly, it is important to take note that this variants was classified as "missense-near-splice", suggesting that the proximity could be concerning in changing the amino acid sequence, affecting the splicing patterns.

### Connection to Disease or Trait

GALNTL5 has a large impact on the male reproductive system. Being that our profile is that of a female individual, there suggest no evidence that the variant will affect the individual, but could play a role if the individual passes this variant to their male offspring.

In general, GALNTL5 is involved in the sperm development and studies have suggested that heterogeneous mutations have been linked to the reduced sperm motility, a common cause of infertility in men [**R13**]. The variant rs6960270 is not the same as the one present in our individual but is on the same gene, GALNTL5, and is located at 177/251. This proximity could have an affect but there is no way to confirm. There are no specific publications documenting the effects of this specific variant position.

### Assessment of Genotype / Phenotype

In the case of this variant is presents as a heterogeneous variant, meaning the individual as one copy of the reference allele being `T` and one copy of the alternate allele being `C`, thus there may be a more neutral affect with the presence of one functional copy. Moreover, in this case the phenotype of the individual will likely not be affected as this variant would have more of an impact on a male individual. However, the individual is still a potential carrier with a 50% chance of passing the variant to their male offspring. Based on the outlined evidence, this variant is unlikely to play an affect of the observed phenotype of the person.

### Population Distribution

Based on ALFA Allele Frequency, **rs6960270**, this variant is actually represented more as `C` in the population. The reference allele `T` is 15 vs `C` at 85. There could be evolutionary reasons for this, but there is no evidence to suggest when the population largely has this variant. These high frequencies could suggest that the variant is not as deleterious as suggested in our scores. In the ExAC global sample the alternate allele is represented 15.9 of the time. However, for the GO ESP global sample the alternate allele is represented a bit high being 18 of the time.

The GALNTL5 rs6960270 variant, despite high scores from our methods, suggesting a deleterious effect, its actual impact is expected to have no effect, based on contextually analyses of the individual's biological sex and the gene's expression pattern; such as the sex and lack of clinical significance. However, the variant could be relevant for male offspring, given GALNTL5's role in sperm development and male fertility.

## PRKRA & CHROMR - rsID: 77419724

PKR-associated protein X (PRKRA) or alternatively referenced as PRK, is researched to suggest involvement in human antiviral defense mechanism [**R5**].

### Molecular Impact

The variant rs77419724 is in the PRK gene, located on chromosome 2 at position 178436252. The variant is a missense variant being heterogeneous with a change from `A` to `T`, changing the amino acid from Ile to Asn, at position 226/314.

The initial evidence from our scoring method suggest significant impact. Firstly, the PolyPhen score of 0.996 suggest a high probability of a damaging affect to protein function. Secondly, the Grantham Score of 149 reflects a high difference between Isoleucine and Asparagine, suggesting an impact from this amino acid change. Third, the CADD score of 25.3 suggest that this variant is amount the top 0.1 of most deleterious substitutions.

### Connection to Disease or Trait

There is evidence to suggest that PRK has influence on Dystonia and causes for dysregulated cellular stress responses. There is a known mutation, P222L; patients with this variant have "progressive, generalized, early-onset dystonia with axial muscle involvement..." [**R3**]. This variant is not the same as the one present in our individual but is on the same gene, PRK, relatively near to this variant location. This proximity could have an affect but there is no way to confirm.

A few positions over, at 246/314 a mutation affects the "transmission of cellular stress responses to PKR" [**R9**]. There once again is no confirmation that there will be any associated affects with our individual. The variants high scores do suggest there could be delteriouness, contributing so similar factors as described above including: immune system dysfunction, neurological disorders, and altered stress responses.

### Assessment of Genotype / Phenotype

This variant represents a heterogeneous variant, with a substitution from the reference allele `A` to the alternate allele `T`. The individual is unlikely to be affected, but its likely to be a carrier. This analysis aligns with the profile of the individual as there are no reported disorders associated with dystonia. Most likely, the functional copy is compensating for any dysfunctions in the variant.

### Population Distribution

Based on ALFA Allele Frequency,**rs77419724**, this variant is represented in 15 of the population, 85 of the population has `A` as the reference allele. These results are similar to the ExAC global results being that 16.3 of the population has the alternative allele. However, for the GO ESP global sample the alternative allele is only represented 8.2 of the time. Our individual has paternal decent from the Netherlands, being that the

GO ESP population set takes into account an American population, this population should not be used for comparison.

In summary, the variant is relatively prevalent in the population thus contradicting the high prediction scores of harmful damaging effects. This is likely do to the proximity to the surrounding significant mutations. Moreover, the variant is not likely to affect the individual as there are no reported disorders associated with dystonia. The individual is likely a carrier of the variant but there is no evidence to suggest any issues with the immune system or neurological disorders.

## ADD - rsID: 4961

Alpha-adducin (ADD1) is a cytoskeletal protein in the adducins family. The Adducins family "promote the assembly of the spectrin-actin network", important for the integrity of the cell and maintaining the cell shape **(IPR051017)**. The spectrin-actin network is extremely important in the maintaining the stability of the cell membrane and controls the shape and volume of the cell.

### Molecular Impact

The variant is in the ADD1 gene, located on chromosome 4 at position 2904980. The variant is a missense variant resulting in a change from `G` to `T`, changing the amino acid from GLY to TRP, at position 460/738.

The PolyPhen score of 1.0 suggest a high probability of damage to the protein's function. Secondly, a Grantham score of 184 suggests the shift from the smallest amino acid, GLY, to one of the largest, TRP, could have an effect on the chemical differences. Third, the CADD score of 20.8, makes this variant in the top 1 of deleterious variants. Moreover, there are potential structural impacts from this substitution due to the complexities of the new amino acid.

### Connection to Disease or Trait

There is a direct connection to a researched association between the mutation of G>T at position 460, sometimes referred to as G460W, and hypertension and blood pressure. However, multiple studies suggest there is there is not enough evidence to confirm an association between the G460W mutation and blood pressure. [**R2**][**R4**].

On there other hand, there are further, more recent, studies the suggest in the combination of other factors, there is enough evidence to suggest an association [**R8**]. Another study suggests that the alpha adducin family has a large impact on hypertension, especially in homozygous carries [**R11**].

Based on the more recent evidence, there is evidence to suggest an association between the ADD1 gene and hypertension, but there is not enough evidence to suggest an association between the our specific variant. There is evidence that suggest in combination of other factors there is, but not purely based on it own.

### Assessment of Genotype / Phenotype

This variant represents, in the individual, as heterozygous, possessing the reference allele `G` and the alternate allele `T`. The functional copy likely positively impacts the phenotpye of the individual as there is an increased in hypertension with carriers of the variant, but still lower than the risk of individuals that are a homozygous carrier. There is no presented evidence in the individuals profile suggest an problems with hypertension or any of the issues outlined by the mutation.

Given the individuals age there would be no harm in the having regular blood screenings to check-in as hypertension can develop with age. There is also a 50% chance the alternate allele can be passed to the offspring.

### Population Distribution

Based on the ALFA Allele Frequency,, this variant is represented in 18.9 of the population, being `T`; while 81.2 of the population carry the reference allele, being `G`. Similar percentages align from the ExAC study

being 20.1 carry the alternate allele. Roughly, 30 of individuals are heterozygous carriers while only 4 of individuals are homozygous carriers. There is a significant shift in the East Asian population as about 50 of the population carry the variant. However, this population does not align with our profile and thus it is less likely to see if our individual. There is a relatively high global frequency thus suggesting this variant is not severely deleterious, likely due to higher rates of heterogygous carriers.

In summary, the variant is relatively prevalent in the population thus contradicting the high prediction scores of harmful damaging effects. This is likely do to the proximity to the surrounding significant mutations. Moreover, the variant is not likely to affect the individual as there are no reported disorders associated with hypertension. The individual is likely a carrier of the variant but there is no evidence to suggest any issues with hypertension or any of the issues outlined by the mutation.

## Discussion

Our analysis of the variant scoring methods revealed important insights into their applications in predicting functional consequences of SNPs. The three scoring methods PolyPhen, Grantham, and CADD demonstrated varying degrees of correlation, suggesting they capture different aspects of variant impact.

A critical observation across our variant analyses is the discrepancy between high deleteriousness scores and actual population frequencies. All three variants received high scores suggesting significant functional impact; however, they had relatively high population frequencies (15-20%). This contradiction is something to note, contributing to the complexity of variant interpretation.

Several observations are important to highlight: - The human body's ability to with stand mutations, may mitigate predicted deleterious effects
- Heterozygosity may provide sufficient genetic assistance decreasing damaging affects
- Context-specific factors may limit phenotypic manifestation
- These scoring methods may overestimate deleteriousness for certain variant types

The GALNTL5 variant effectively illustrates how biological context can override computational predictions—despite high scores across all methods, its testis-specific expression renders it phenotypically neutral in females. Similarly, the PRKRA variant demonstrates how proximity to known pathogenic mutations can influence scoring algorithms without necessarily conferring the same clinical significance.

These findings emphasize the importance of integrating multiple lines of evidence when evaluating variant significance: computational predictions, population frequencies, gene function, expression patterns, and available clinical data. No single scoring method provides sufficient evidence for pathogenicity determination.

Future research should focus on developing integrated scoring approaches that incorporate population-specific frequencies. Additionally, functional validation studies for variants with discordant predictions would enhance our understanding of these scoring methods' limitations and applicability in clinical settings.

## Conclusion

In conclusion, our analysis of the three selected variants highlights the complexity of variant interpretation in the context of human genetics. While scoring methods like PolyPhen, Grantham, and CADD provide valuable insights into potential functional consequences, they should be interpreted with caution. The discrepancies between high deleteriousness scores and actual population frequencies underscore the need for a comprehensive approach that considers biological context, gene function, and available clinical data.

# Resources

[**1**]. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods, 7(4), 248–249. https://doi.org/10.1038/nmeth0410-248

[**2**]. Busch, C. P., Harris, S. B., Hanley, A. J., Zinman, B., & Hegele, R. A. (1999). The ADD1 G460W polymorphism is not associated with variation in blood pressure in Canadian Oji-Cree. Journal of human genetics, 44(4), 225–229. https://doi.org/10.1007/s100380050148

[**3**]. Camargos, S., Scholz, S., Simón-Sánchez, J., Paisán-Ruiz, C., Lewis, P., Hernandez, D., Ding, J., Gibbs, J. R., Cookson, M. R., Bras, J., Guerreiro, R., Oliveira, C. R., Lees, A., Hardy, J., Cardoso, F., & Singleton, A. B. (2008). DYT16, a novel young-onset dystonia-parkinsonism disorder: identification of a segregating mutation in the stress-response protein PRKRA. The Lancet. Neurology, 7(3), 207–215. https://doi.org/10.1016/S1474-4422(08)70022-X

[**4**]. Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., & Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nature genetics, 22(3), 239–247. https://doi.org/10.1038/10297

[**5**]. Ito, T., Yang, M., & May, W. S. (1999). RAX, a cellular activator for double-stranded RNA-dependent protein kinase during stress signaling. The Journal of biological chemistry, 274(22), 15427–15432. https://doi.org/10.1074/jbc.274.22.15427

[**6**]. Joshi, R., Gilligan, D. M., Otto, E., McLaughlin, T., & Bennett, V. (1991). Primary structure and domain organization of human alpha and beta adducin. The Journal of cell biology, 115(3), 665–675. https://doi.org/10.1083/jcb.115.3.665

[**7**]. Lanzani, C., Citterio, L., Jankaricova, M., Sciarrone, M. T., Barlassina, C., Fattori, S., Messaggio, E., Serio, C. D., Zagato, L., Cusi, D., Hamlyn, J. M., Stella, A., Bianchi, G., & Manunta, P. (2005). Role of the adducin family genes in human essential hypertension. Journal of hypertension, 23(3), 543–549. https://doi.org/10.1097/01.hjh.0000160210.48479.78

[**8**]. Li, Y., Thijs, L., Kuznetsova, T., Zagato, L., Struijker-Boudier, H., Bianchi, G., & Staessen, J. A. (2005). Cardiovascular risk in relation to alpha-adducin Gly460Trp polymorphism and systolic pressure: a prospective population study. Hypertension (Dallas, Tex. : 1979), 46(3), 527–532. https://doi.org/10.1161/01.HYP.0000174988.81829.72

[**9**]. Peters, G. A., Li, S., & Sen, G. C. (2006). Phosphorylation of specific serine residues in the PKR activation domain of PACT is essential for its ability to mediate apoptosis. The Journal of biological chemistry, 281(46), 35129–35136. https://doi.org/10.1074/jbc.M607714200

[**10**]. Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, Martin Kircher, CADD: predicting the deleteriousness of variants throughout the human genome, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D886–D894, https://doi.org/10.1093/nar/gky1016

[**11**]. Psaty, B. M., Smith, N. L., Heckbert, S. R., Vos, H. L., Lemaitre, R. N., Reiner, A. P., Siscovick, D. S., Bis, J., Lumley, T., Longstreth, W. T., Jr, & Rosendaal, F. R. (2002). Diuretic therapy, the alpha-adducin gene variant, and the risk of myocardial infarction or stroke in persons with treated hypertension. JAMA, 287(13), 1680–1689. https://doi.org/10.1001/jama.287.13.1680

[**12**]. R. Grantham ,Amino Acid Difference Formula to Help Explain Protein Evolution.Science185,862-864(1974).DOI:10.1126/science.185.4154.862

[**13**]. Takasaki, N., Tachibana, K., Ogasawara, S., Matsuzaki, H., Hagiuda, J., Ishikawa, H., Mochida, K., Inoue, K., Ogonuki, N., Ogura, A., Noce, T., Ito, C., Toshimori, K., & Narimatsu, H. (2014). A heterozygous mutation of GALNTL5 affects male infertility with impairment of sperm motility. Proceedings of the National Academy of Sciences of the United States of America, 111(3), 1120–1125. https://doi.org/10.1073/pnas.1310777111

[**14**]. Sherry, S. T., & Sirotkin, K. (1999). dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Research, 9(8), 677-679. https://doi.org/10.1101/gr.9.8.677

## AI Acknolwedgement

**Claude 3.7 Sonnet. Claude was utilized as the first proof read of the paper. The model asked to not comment on the content of the paper but only to comment on the grammar and structure of the paper.

[**15**]. Claude was utilized to optimize the basic information of a list of genes to allow for further research. Clause was prompted with:

```
Research the following gene list. I want the link to the uniport entry for human.
I want the gene name and the description of the gene. Save the information in a
dictionary with the gene as the key. The write the python code to loop through the
dictionary and save the results in an excel file.

The columns should be gen_symbol, gene_name, gene_description, uniport link.

I then want another lookup to get a excel of all the variant information on a gene.
This output should be gene_symbol, ID, position, description with the rs from the dbSNP.
```

The provide code was directly used with no edits via the author. The author validated the results to ensure accuracy. There is no impact to the analysis or manuscript of the paper as the results were used for further research. This code saved the user 2+ hours of data mining through UniProt.

## Github Project

Find the github project, **linked here**, with all the submission components of this project.

- end2end python script
- MasterData.xlsx
- finalReport.pdf
- Input Data in .pkl

## Tables and Figures

Table 1 | Summary Statistics for Variant Prediction Scores
Total observations removed: 669829

| Metric | PolyPhen | Grantham Score | CADD Score | CADD Score (Complete Cases) |
|---|---|---|---|---|
| Min | 0.00 | 5.00 | 0.00 | 0.00 |
| Max | 1.00 | 215.00 | 59.00 | 59.00 |
| Mean | 0.28 | 66.06 | 4.23 | 8.37 |
| Median | 0.01 | 58.00 | 3.20 | 7.87 |
| Top 75% | 0.74 | 91.00 | 5.82 | 12.85 |
| Std Dev | 0.41 | 42.42 | 3.98 | 7.22 |
| Valid Obs. | 14108 | 14832 | 681285 | 14003 |
| Missing | 669724 | 669000 | 2547 | 669829 |

Figure 1: Table 1

Table 2 | Selected Genetic Variants

| Variant | Accession | Chromosome | Position | rsID | Sample Alleles | Gene List |
|---|---|---|---|---|---|---|
| Variant 1 | XM_017011796.1 | 7 | 151982987 | 6960270 | T/C | GALNTL5 |
| Variant 2 | NM_003690.5 | 2 | 178436252 | 77419724 | A/T | PRKRA,CHROMR |
| Variant 3 | NM_001354754.2 | 4 | 2904980 | 4961 | G/T | ADD1 |

Figure 2: Table 2

## Table 3 | Variant Prediction Scores

| Variant | PolyPhen | Grantham Score | CADD Score | GREP Score |
|---|---|---|---|---|
| Variant 1 | 1.0 | 180 | 20.4 | 4.65 |
| Variant 2 | 0.996 | 149 | 25.3 | 5.92 |
| Variant 3 | 1.0 | 184 | 20.8 | 5.55 |

Figure 3: Table 3

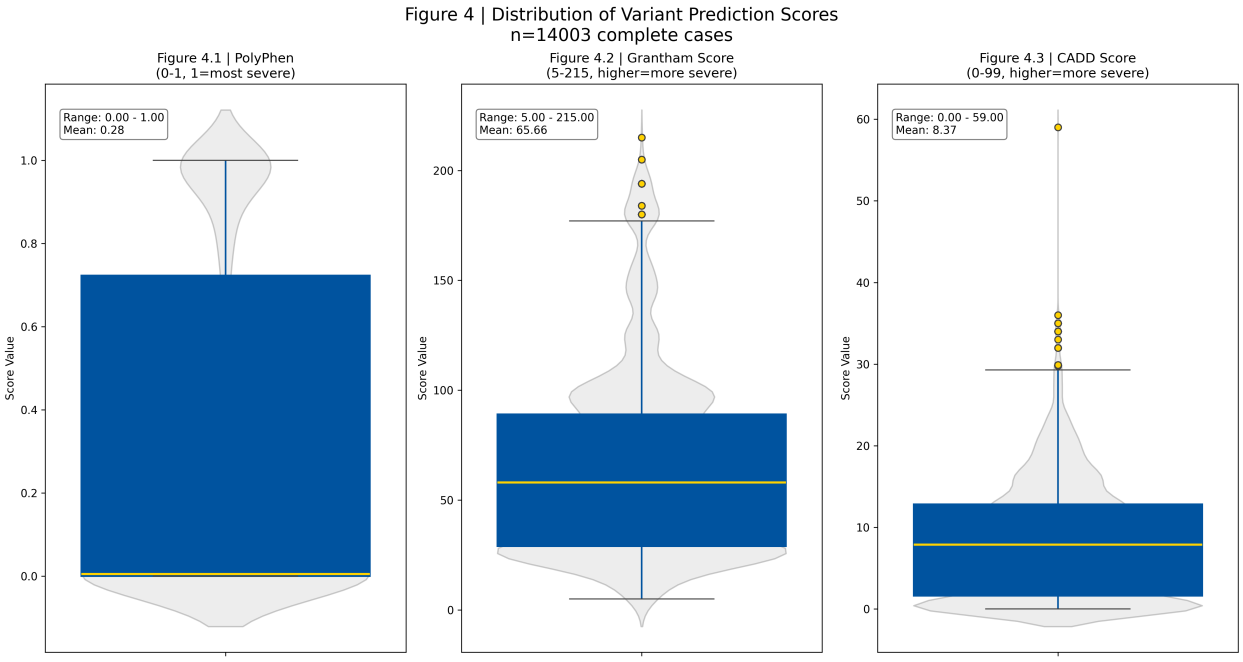


Figure 4: Figure 4.0


Figure 5: Figure 5.0