MACHINE LEARNING APPROACHES FOR CLASSIFYING GENETIC VARIANTS

Jennifer Johnson

Adviser: Professor Michael Linderman

A Thesis

Presented to the Faculty of the Computer Science Department

of Middlebury College

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

May 2018

ABSTRACT

Variant classification is the process of determining a variant's effect on a disease phenotype. I examined the feasibility of using probabilistic soft logic (PSL), a machine learning programming language, to classify variants as either pathogenic i.e. disease-causing, or benign. Using existing guidelines manually developed by Nykamp *et al*., I focused on automating one step in the classification process: the interpretation of population frequency data. [19] The underlying assumption is that allele frequency can be used as an indicator for pathogenicity. I implemented a proof-of-concept model of the relationship between variant frequency and pathogenicity using PSL. I employed this model to infer the classification of synthetic variants. The final model had a specificity of 75% and a sensitivity of 50% on a synthetic dataset. These results suggest that PSL could be used in the classification workflow to increase the efficiency of expert human interpreters and facilitate the automatic construction of gene or disease-specific classifiers.

ACKNOWLEDGEMENTS

I would like to thank Professor Linderman for his support and guidance. Thank you for always steering me in the right direction and giving helpful feedback. Thank you to the Middlebury College Computer Science department for a great four years, from FYSE 1414 “Computing and Society” to CSCI 702. Finally, I would like to thank my family and friends for their continuous encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Genetic variants are differences in the genome that can contribute to an individual's phenotype, or expressed characteristics. [21] They are also called mutations or polymorphisms, but these terms (especially mutation) imply that the variant contributes to disease. Therefore, variant is the preferred term, and the classification of a variant is a label to describe its effect on the disease phenotype(s) present in the individual. In this work, classification will focus on rare monogenic diseases and variants that have a high penetrance. Penetrance will be discussed further in Chapter 2.

Directly testing for and measuring pathogenicity is challenging. Because of the lack of experimental data, a variant's classification must typically be inferred indirectly from a combination of experimental, *in-silico*, and population-based observations. Determining a causal variant for a disease requires costly, time-consuming, and labor intensive genetic screens that may be prone to errors. The large volume of variants to test increases the number of laboratory experiments required, the probability of false results, and therefore the uncertainty of results. [1] Furthermore, variants may be responsible for multiple phenotypes or work in combination with other variants to contribute to a single phenotype. Therefore, the determination of genotype-phenotype relationship and variant classification is challenging.

Duzkale *et al.* [5] proposes three questions for assessing the pathogenicity of variants:

1. Does the variant alter the function of the gene?
2. Can the functional change result in a disease or alternative phenotype?
3. Is the disease or alternative phenotype relevant to the clinical condition present in the individual?

The first question considers the effect of variants on the function of the gene product. It can be answered by measuring biochemical properties of the protein. This involves a comparison of the wild type protein with the variant protein in terms of sequence, structure, or function. If a variant alters the wild type protein, it is "damaging." The second question considers the link between damaging variants and disease. It can be answered with population data, which includes disease prevalence and variant frequency. This question should also consider a disease's mode of inheritance. Finally, the third question considers patient-specific data to link a disease phenotype to a patient's symptoms. This could relate to the kinds and severity of the symptoms present in the individual and could also incorporate all types of data from the biological parents.

In the ideal world, all three data types would be available for all variants being classified. MacArthur *et al*. define pathogenic as "contributing mechanistically to disease," which means that all three questions are considered. [17] However, in practice, all three data types are often not available for every variant being classified.

This may be because limited resources prevent experimental testing of the variant, there might be restricted access to clinical patient records, or the data may not be easily nor automatically retrieved. If variant evidence is not accessible to the classifier because it has not been accurately recorded and referenced in the database and existing literature, this evidence will not be incorporated into the decision. Therefore, inaccessible data should be considered missing data. Furthermore, data may not be reliable. For example, a small sample size may not be sufficient for providing representative and accurate population data. Finally, even if patient data is available, it is difficult to assess this type of data systematically and efficiently, making it an ineffective source of information. Therefore, variant classification is often performed independently of a particular patient and is therefore limited to biochemical and population-based evidence. Variant classifiers are forced to rely on a small subset of data when assigning a label.

In many cases, classifiers (both human and software) narrow the scope of the problem addressed to adjust for this lack of knowledge. The problem of classifying genetic variants in terms of their impact on phenotype changes to classifying variants by their impact on protein. MacArthur *et al.* define damaging as "altering normal levels of a gene or biochemical function of a gene product." [17] Therefore, these restricted classifiers predict the effect variants will have on the structure, function, or level of gene expression of the variant's gene product.

1.1 "Damaging" vs. "Pathogenic"

Some papers use the terms "damaging" and "pathogenic" interchangeably. This assumes that the change in protein function or gene expression caused by a "damaging" variant will be deleterious. Furthermore, this assumes that the observed negative effect is related to the disease phenotype present in the individual and is therefore pathogenic. However, this is not always the case. A "damaging" variant may have neutral or even positive effects. This would mean that a "damaging" variant is benign or even advantageous. Furthermore, Nykamp *et al.* state, "A variant is more likely to be pathogenic if it has a consequence for a gene product that is consistent with the disease mechanism of the gene." [19] A variant may not be pathogenic for a disease because its impact on protein is not related to that disease's mechanism. For example, if the change in function is not related to how the disease is caused, there will be no effect on the phenotype. Therefore, the terms "damaging" and "pathogenic" are not always synonyms.

The distinction between the terms relates to whether or not the second and third questions have been addressed. The first question asks whether the variant has an impact and can be used to determine whether a variant is "damaging." In contrast, the second and third questions ask whether this impact is relevant to the disease observed. Specifically, the second question links the variant and the disease, while the third question links the

disease with the patient. Once all three questions have been answered affirmatively, then the term pathogenic can be used to describe the variant.

1.2 Motivation

Variant classification is important because a pathogenic assertion can influence healthcare. "Actionable" variants are defined by Sukhai *et al.* as "drugable or predictive and/or with diagnostic/classification implications." [24] The identification of variants that are both pathogenic and "actionable" can focus treatments on such variants, as opposed to variants classified as unknown or benign. This can increase the efficiency of treatment development by focusing time, costs, and labor on pathogenic variants.

Furthermore, a pathogenic assertion obligates a healthcare provider to consider treatment. [9, 21] The identification of a "damaging" variant is the first step, but the lack of a concrete link between "damaging" and "pathogenic" does not oblige treatment. Therefore, the final two questions and "formal" classification of a "pathogenic" variant is essential for ensuring that appropriate clinical consideration occurs.

Finally, classification can still have benefits even if treatments have not yet been developed. The knowledge of a certain diagnosis can have personal utility to a patient and family. [9] Even if no treatments are available, knowledge of a specific diagnosis can provide a degree of certainty and even a sense of community with other affected patients. Therefore, classification is beneficial even if it cannot contribute directly to treatment.

Automating the classification process could increase its efficiency. Classification is a time-consuming and labor-intensive process because of its complexity. Furthermore, evidence required for variant classification is sparse. Therefore, an automated approach could allow accurate classification decisions to be made using as little data as possible. Statistical prediction systems that answer the three questions using a minimum amount

of data can reduce significant costs and labor hours. [23] Therefore, tools are needed that automatically answer each of the three questions by predicting a link between:

1. A variant and a damaging effect on a protein
2. A damaging variant and a disease
3. A disease and a patient

The case-dependent nature of the last question suggests that human expertise will still be required to make an ultimate classification on a case-by-case basis. Therefore, the goal of automating approaches for classifying genetic variants is to make human experts more efficient, not replace them. The third question will not be easily automated, but automating the first two questions is a more reasonable goal.

Machine learning, or the use of data to train a model for classifying unknown entities, is one statistical method that could be used to approach and automate this problem. Machine learning can be used because this is a classification problem that is based on the observations of features. There is a substantial number of already-classified variants to use for training a model and extracting predictive features for pathogenicity. However, the evidence available for classifying a specific variant can still be sparse. The observations and evidence required to classify these variants as "damaging" and then as "pathogenic" may be less available. A data-driven approach could be used to extract trends between evidence and pathogenicity. This could increase classification efficiency while using a minimum amount of information from each variant.

1.3 Goals and Outline

This thesis will discuss existing machine learning approaches for classifying genetic variants and apply probabilistic soft logic (PSL) to classify variants. Specifically, I will attempt to automate the second question related to pathogenicity: linking a damaging

variant with a disease using population data. Population data will include disease prevalence and allele frequency. It is assumed that the reader is knowledgeable about the evolutionary concept of natural selection, but no other biological background is assumed.

First, I will describe the existing clinical method of classifying genetic variants and its limitations. Next, I will examine existing machine learning methods that address the first question related to pathogenicity: predicting the damaging effect of a variant. Then, I will describe a checklist-driven method that addresses the second question: linking a damaging variant with a disease using population data. My goal will be to combine the advantages of both the clinical method and machine learning approaches into a hybrid tool to classify variants. Finally, I will explain my experiments with PSL that aim to provide a proof of concept that this second step can be automated using a data-driven or machine learning approach.

CHAPTER 2

THE STANDARD METHOD OF CLASSIFYING GENETIC VARIANTS

2.1 Goals of Standardization

Variant classification schemes need to be systematic and objective. MacArthur *et al.* warn that non-standard methods might "impede the translation of genomic research into the clinical diagnostic setting." [17] This is because a mislabeled variant would decrease the efficiency of treatment development. A false positive could focus treatment efforts on a benign variant, while a false negative could inhibit the timely development of a treatment for a pathogenic variant. A standard method of reporting would make both the classification itself and communication of such decisions more efficient and consistent.

Databases that provide the clinical significance of variants should provide all of the evidence and the logic that was used to classify the variant. [17, 19, 21] The organized inclusion of contradictory evidence would provide both a measure of confidence and uncertainty in the assertion. Currently, there is no such quantitative scale. As Richards *et al.* put it, "there is no quantitative measure of likely." [21] Such rigorous standards of reporting would also make updating classification with new knowledge more robust. [17] The protocols followed and software used for data collection and classification decision should be recorded as metadata. [21] Finally, policies should also be put into place to ensure that updating occurs on a regular basis. Re-sampling, sampling of relatives, the development of new technologies, a change in nomenclature, and other studies of the genes of interest are all causes for revisiting and revising classifications. [21]

2.2 ACMG Guidelines

The American College of Medical Genetics proposed a set of guidelines in 2015 in an effort to standardize the classification of genetic variants. [21] The goal was to create

non-ambiguous rules and standard nomenclature for first, evaluating and weighing evidence of pathogenicity; and second, combining these weighted pieces of evidence into a classification. The motivation was to resolve discrepancies between laboratories with differing protocols and resolve the classification of variants with contradictory evidence. The storage and consolidation of records with a standard nomenclature and quantitative format would increase communication efficiency and improve the clarity of variant classification data.

The authors surveyed a variety of clinical laboratories on their current methods of variant classification. After analyzing the diverse methods, the authors devised a set of evidence items with corresponding weights. The weights are very strong, strong, moderate, and supporting for pathogenic variants and stand-alone, strong, and supporting for benign variants. Pieces of evidence that are more convincing are given more weight, while evidence that is less certain is given less weight. For example, functional evidence is given more weight than computational predictors. This is because experimental evidence is more reliable because it is measured directly *in vitro* or *in vivo* rather than inferred *in silico*.

The authors also devised a set of rules for combining evidence of different weights into a classification of one of five labels: Benign, Likely Benign, Pathogenic, Likely Pathogenic, or Variant of Unknown Significance. The rules reflect the unequal probability of a variant being pathogenic versus benign. Variants are more likely to be benign than pathogenic. For example, a variant requires at least one piece of strong evidence, one piece of moderate evidence, and four pieces of supporting evidence to be labelled as pathogenic, as shown in Figure 2.1. However, only one piece of stand-alone evidence or two pieces of strong evidence are required for a variant to be benign.

Some of these rules, because of their Boolean and implicative structure, have already been automated. [12] This eliminates one step that requires human interpretation and

Table 5 Rules for combining criteria to classify sequence variants

Pathogenic	(i) 1 Very strong (PVS1) *AND*
	(a) ≥1 Strong (PS1–PS4) *OR*
	(b) ≥2 Moderate (PM1–PM6) *OR*
	(c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) *OR*
	(d) ≥2 Supporting (PP1–PP5)
	(ii) ≥2 Strong (PS1–PS4) *OR*
	(iii) 1 Strong (PS1–PS4) *AND*
	(a)≥3 Moderate (PM1–PM6) *OR*
	(b)2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
	(c)1 Moderate (PM1–PM6) *AND* ≥4 supporting (PP1–PP5)
Benign	(i) 1 Stand-alone (BA1) *OR*
	(ii) ≥2 Strong (BS1–BS4)

Figure 2.1: ACMG rules for combining weighted evidence into the two most extreme classifications, Pathogenic and Benign. The burden of evidence is higher for pathogenic variants than for benign variants to account for the unequal distribution of variants on the pathogenicity scale. Adapted from Richards *et al.*, 2015, Table 5. [21]

labor and increases the efficiency of classification. [20] However, the first step, the assignment of weights to evidence is much more difficult to automate because of the subtlety and complexity of how different features impact pathogenicity. The ACMG guidelines address this by allowing pieces of evidence to be moved to a different weight category based on "professional judgement." [21] However, this creates a source of ambiguity that is one of the limitations of the ACMG guidelines.

2.3 Limitations of the ACMG Guidelines

When Richards *et al.* were surveying laboratory groups about their preferred methods of classifying variants, many groups stated that the "specific assignment of points implies a quantitative level of understanding that is not supported scientifically and does not take

into account the complexity of genetic evidence." [21] This suggests that the assignment of weights does not lend itself to quantification, let alone automation.

Furthermore, variant classification is biologically complex. It is uncertain whether all of the factors that contribute to variant classification can be accurately represented in an automated system. While some factors may be quantified, such as the probability of evolutionary conservation and other measurable biochemical properties, other factors are case-dependent and require human judgement. [19, 21] Variant classification currently requires human expertise, and it is uncertain whether any system could ever replace this form of knowledge.

Therefore, an automated form of the ACMG guidelines would not seek to replace human experts. Instead, it would act to improve the efficiency of the classification process by automating only a subset of the guidelines, and still rely on human expert judgement for making the ultimate decision. Automated classification guidelines could also improve experts' understanding of the subtle effects of their judgement calls and increase the accuracy and reproducibility of classification. Finally, the automatic generation of guidelines could increase the efficiency of developing gene and disease-specific models.

The current manual method is prohibitive for the development of such models. The authors analyzed a variety of classification protocols and determined a standard set of evidence weights and combination rules for classifying variants. This manual approach is inefficient and will not scale well to the development of gene and disease-specific guidelines in the future. The automatic development of criteria will increase the rate of supplementary criteria development.

Using the guidelines, different laboratories achieved only 34% concordance with each other while analyzing the same set of variants. [22] After discussing the conflicting interpretations, the groups were still only able to raise the concordance to 71%. [22] This low concordance reflects the biological complexity and case-dependent nature of variant

classification and especially the challenge of assigning weights to evidence. However, it also reflects a limitation of the guidelines. The guidelines are limited in their lack of specificity. The language used in the guidelines is sometimes ambiguous to allow case-specific knowledge to be applied. However, this decreases the reproducibility of classifications between different laboratories.

For example, one rule is shown in Table 2.1. This rule is based on the allele frequency of a variant. Allele frequency is a measure of the common-ness of a variant in a population.

Because variants in a population are acted on by natural selection, most variants do not have a strong negative impact on the survival of the individual. If they had a severe impact on the individual's health, the individual would have died, eliminating that variant from the gene pool. This work focuses on monogenic diseases (diseases caused by a single variant) and variants with a high penetrance (when the presence of the variant provides a high likelihood for observing the resulting phenotype). Therefore, the mechanism of natural selection is especially effective at eliminating these variants and their deleterious effects from the gene pool. As Nykamp *et al.* put it, "variants that do not reliably affect protein sequence are presumed more likely to be tolerated, and variants that exert a more dramatic effect are presumed more deleterious." [19] Therefore, alleles with a high frequency in the population are more likely to be benign, whereas alleles with a low frequency may be pathogenic.

The rule shown in Table 2.1 means that if a variant is absent from the Exome Aggregation Consortium, or ExAC database, [15] this provides moderate evidence of

Table 2.1: Population and allele frequency-based data provides moderate evidence for pathogenicity. Adapted from Richards *et al.*, 2015, Table 3. [21]

Evidence of Pathogenicity	Category
Moderate	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium

pathogenicity. Furthermore, if the disease's mode of inheritance is recessive, an "extremely low" allele frequency is sufficient to provide moderate evidence of pathogenicity.

This extremely low frequency should be better and more specifically defined. Therefore, the rules need to be amended to make their application more accurate and replicable between laboratory groups.

CHAPTER 3

EXISTING MACHINE LEARNING APPROACHES FOR CLASSIFYING GENETIC VARIANTS

Variant classification can be readily modeled as a machine learning problem. It is not a novel approach to use statistical analyses to "weight all evidence to reach an overall conclusion." [5] Therefore, there are many tools available for automatically predicting "damaging" variants. I will examine only a few examples of existing machine learning approaches such as PolyPhen-2 and REVEL. However, machine learning approaches bring with them a whole new set of challenges such as data sparsity and the limited scope of data-driven models.

3.1 Limitations to Machine Learning Approaches

Machine learning approaches are limited by the quality of their input data. The volume of observed variants continues to increase as sequencing methods advance. However, the utility of this sequencing data is limited because it is not always paired with biochemical and population-based observations. Therefore, although there is a large number of variants to be classified, the data available for each variant can be sparse and often insufficient for answering the three questions for assessing pathogenicity. This sparsity limits the accuracy of analyses. Furthermore, the majority of variants are classified as benign. [5] This skew in the data makes benign classification easier than pathogenic classification and could decrease the validity of performance metrics. [16]

Machine learning approaches may be limited by their specificity, or lack thereof. A model that considers all variation in a large number of features will have high accuracy on the training data, but not on other datasets. On the other hand, a general model will not achieve good performance on any dataset. Therefore, data-based models must find a balance between specificity and overfitting to achieve both high performance and

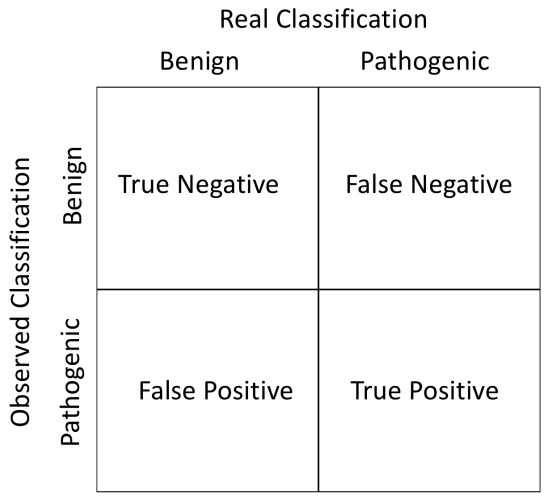

Figure 3.1: Confusion matrix for variant classification. The values are used to evaluate the performance of a variant classifier compared to existing assertions from ClinVar or other benchmark datasets.

adaptability to unseen data. Because of the need for this balance, the performance of classifiers is often measured as percent specificity, percent sensitivity, and values in a confusion matrix, as shown in Figure 3.1. [6]

In the context of variant classification, gene or disease-specific tools will usually be more effective than general ones. [8, 10] However, the original ACMG guidelines are broad. Richards *et al.* justify this by arguing that individual laboratory groups can establish their own set of guidelines for the gene or disease they are working with, considering the individual context and types of evidence. [21] While gene and disease-specific tools will be more accurate, the time, cost, and labor required for their development is prohibitive. A machine learning and data-driven approach could make the development of such specific (and therefore more accurate) tools more feasible.

3.2 Individual Tool: PolyPhen-2

PolyPhen-2 is a computational tool that predicts whether the impact of an amino acid substitution will be damaging to the structure and stability of the protein. [2] Therefore,

it answers the first question of pathogenicity: identifying a variant as "damaging" to its gene product. It uses a probabilistic classifier to calculate the probability that a single nucleotide variant is "damaging" given the observed probability of conservation of that position in the sequence. For a false positive rate of 20% , PolyPhen-2 has a true positive rate of 92% on the HumDiv dataset, and 73% on the HumVar dataset. [2] These are benchmark datasets that are used to assess the accuracy of classifiers. One limitation of PolyPhen-2 and other "damaging" variant callers is related to the biological complexity of the input data. Nucleotide sequences from different species that code for the same protein are aligned to calculate the conservation of the position in the sequence. Multiple sequence alignments are often limited in accuracy because of "long stretches of repeats and/or high compositional biases." [2] Because of these uncertainties, the alignments may be inaccurate. This would impact the calculation of the conservation of the position and decrease the tool's performance.

3.3 Consensus Tool: REVEL

Because of the biological complexity and vast number of factors that contribute to a variant's pathogenicity, tools that rely on a small number of features may be limited. [11, 23] Ensemble tools consolidate scores from individual tools to determine if a variant is "damaging." REVEL, or Rare Exome Variant Ensemble Learner, is one example of such a tool. [11]

One advantage of combining multiple machine learning tools is to prevent the overfitting of data. Because of the use of individual tools as features, the impact of each piece of evidence on the classification is limited. Therefore, it is easier to prevent overfitting and achieve a better balance between accuracy and simplicity. [7] Furthermore, ensemble tools heed MacArthur *et al.*'s warning about overestimating a single feature over another. Because the features are scores from tools that take into account multi-

ple factors themselves, an outlier in the data is less likely to skew the results and the dependence of the classification on a single feature is less likely to occur. [17]

REVEL consolidates scores from PolyPhen-2 and twelve other individual predictors in a random forest approach. It uses 1,000 decision trees to classify variants into one of two categories: benign or pathogenic. Each decision tree in the forest is first randomly seeded with features, and then uses the Gini index to choose features that produce the lowest Gini index, or the lowest false positive rate. REVEL has been shown to be a superior classifier compared to individual tools. [11] When compared to other consensus tools, as shown in Figures 3.2 and 3.3, it was shown to have a high performance.

However, REVEL still has limitations. Because of its reliance on existing tools, it is first constrained by the performance of these component tools. Then, it is also limited by existing pathogenicity assertions. MacArthur *et al.* warn that prior reports should not be viewed as definitive. [17] Existing pathogenicity assertions may be influenced by computational results, resulting in the overestimation of both individual and ensemble tool performance. The use of unreliable and inaccurate training data will decrease the accuracy of any machine learning model, but this circular argument is especially detrimental to the accurate assessment of variant classifier performance. REVEL is also limited because it still only addresses the first question related to pathogenicity: the evaluation of a variant as "damaging." Some of its component tools may use allele frequency as a feature, but this tool does not use population data directly. Therefore, my goal is to use machine learning to address not only biochemical but also population data to classify "damaging" variants as "pathogenic."

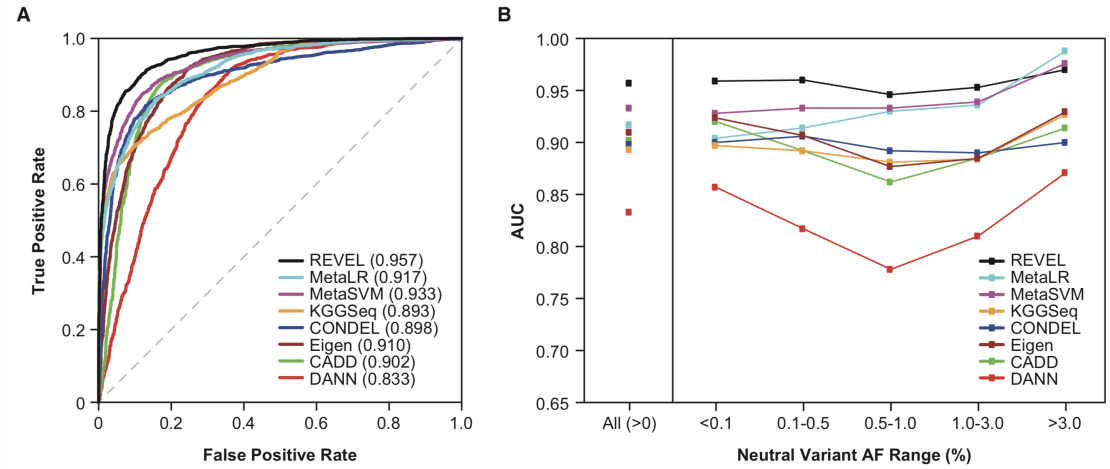

Figure 4. Performance of Ensemble Methods in an Independent Test Set of 1,953 Pathogenic and 2,406 Benign Variants from ClinVar
(A) ROC curves and the AUC for all variants.
(B) AUC for each ensemble method, stratified by neutral variant AF.

Figure 3.2: Performance of REVEL on variants from ClinVar, measured by ROC curves and the AUC. Adapted from Ioannidis *et al.*, 2016, Figure 4. [11]

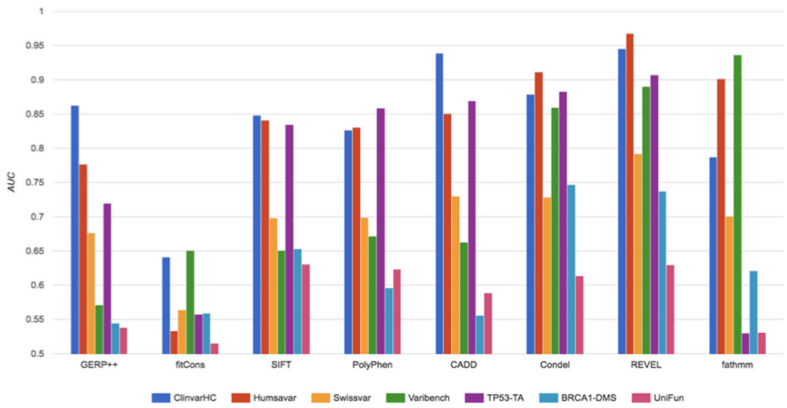

Fig. 2 Histogram depicting apparent accuracies of in silico variant effect predictors based on ROC curve AUCs for the benchmarking datasets used in this study

Figure 3.3: Performance of REVEL compared to other consensus tools. Adapted from Mahmood *et al.*, 2017, Figure 2. [18]

CHAPTER 4

EXISTING TOOLS FOR LINKING GENOTYPE AND DISEASE

I examined a few systematic approaches as a basis for my approach of linking "damaging" variants to a disease using machine learning. I will discuss InterVar and Sherloc as illustrative examples. [16, 19]

InterVar is a semi-automated system for annotating and classifying variants based on a variety of evidence types. [16] It automatically annotates variants with information about 18 of the 28 ACMG criteria that relate to all three questions related to pathogenicity. Specifically, variant and disease data are automatically imported from external databases. Then, the user is required to manually assess the annotations based on their own knowledge and expertise. The user provides a judgement for the remaining ten guidelines, such as those related to patient-specific data. This two-step system increases the efficiency of variant classification while retaining the ability for expert domain knowledge to take precedence over automatically generated predictions. The process is shown in Figure 4.1.

By considering all of the ACMG criteria, this tool answers all 3 questions related to pathogenicity. However, this semi-automated system does not involve machine learning. The authors argue that this could be a future development, stating that machine learning could better represent the concept that "different types of criteria might have different contributions and weights." [16] The ideal tool would be a machine learning approach that considers all of the questions relating to pathogenicity while retaining the ability to incorporate user knowledge.

Sherloc, or Semi-quantitative Hierarchical Evidence-Based Rules for Locus Interpretation, is a tool that introduces detailed refinements to the ACMG rules. [19] It addresses the second question related to pathogenicity: the link between a damaging variant and a disease. Like the ACMG guidelines, Sherloc's broader goal is to standard-

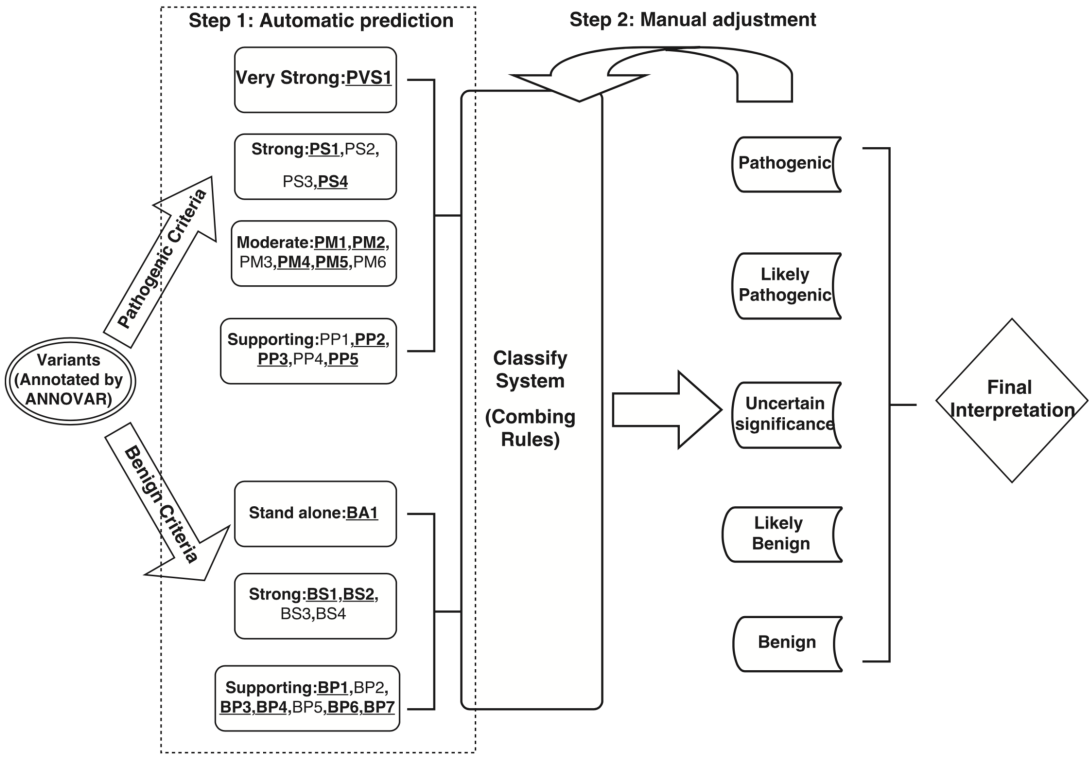

Figure 4.1: InterVar workflow. InterVar is an existing hybrid tool that combines an automated approach with a manual adjustment step. Adapted from Li and Wang, 2017. [16]

ize the "valuation of classification-related evidence" in an effort to "drive consensus and promote consistency, reproducibility, and efficiency among users." [19] To do this, Sherloc uses a collection of hierarchical decision trees to make the ACMG guidelines finer-grained and assign weights to evidence on a quantitative and reproducible scale.

Nykamp *et al.*'s laboratory group used the ACMG guidelines to classify 40,000 unique variants. Sherloc was the product of resolving ambiguities and other inconsistencies observed in the systematic application of the original guidelines. Their specific goal was to "develop more granular rules to capture the necessary complexity" of assigning evidence weights and establish a framework that is both more quantitative and intuitive. [19] For example, the group observed that ambiguities were often caused by different interpretations of "strong" versus "moderate" evidence. These problematic differences

were caused by the flexibility provided by the original ACMG guidelines to account for human expertise, the case-dependent nature of evidence, and other non-quantitative biological knowledge. However, there was now a need for a systematic differentiation of evidence weights to increase the accordance between laboratories and the reproducibility of the application of the ACMG rules.

Nykamp *et al.* also expanded the concept that different evidence types should have different weights depending on the certainty of that evidence. Richards *et al.* established this method of weighing evidence based on reliability in the original ACMG guidelines, but the system needed to be updated. First, Nykamp *et al.* needed to redefine some evidence types. Some new evidence types had been developed since the release of the original guidelines, so the Sherloc developers needed to expand the criteria. Furthermore, Nykamp *et al.* defined a new class of evidence: clinical functional evidence. This comprises of evidence that is both directly measured and specifically answers the second question of variant pathogenicity: the link between genotype and phenotype. Nykamp *et al.* grant clinical functional evidence the highest weight compared to other functional evidence types that only relate to labeling a variant as "damaging." On the opposite end of the scale, Nykamp *et al.* assign computational evidence the lowest weight possible. Finally, Nykamp *et al.* claims that most types of functional evidence, but especially computational evidence, should only be used to confirm or refute the argument established by clinical functional data. Therefore, Nykamp *et al.*'s system updates and makes more specific the original ACMG guidelines.

Nykamp *et al.*'s second amendment to the ACMG guidelines is to establish a more continuous measure of evidence values. Richards *et al.*'s approach provides only 3 or 4 evidence categories for benign and pathogenic evidence, respectively. Nykamp *et al.* put evidence on a more continuous scale, as shown in Figure 4.2. This approach replaces the original ACMG guidelines' approach of combining of evidence using Boolean rules.

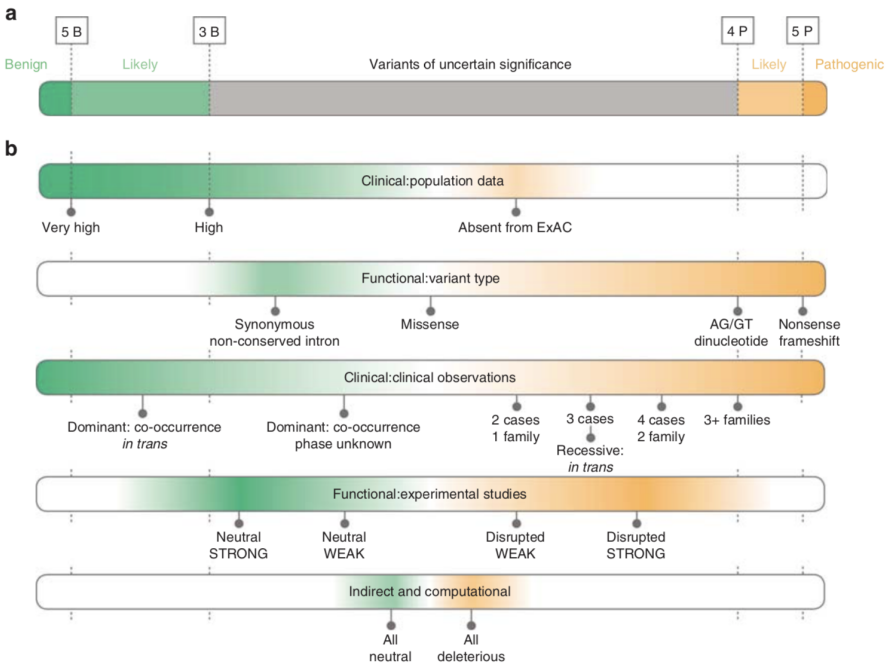

Figure 1 Classification scoring thresholds and evidence categories. (**a**) Point score thresholds for pathogenic (P), likely pathogenic, variant of uncertain significance, likely benign, and benign (B) classifications. Pathogenic and benign evidence is scored separately. Evidence in both directions can suggest a non-Mendelian variant. (**b**) Five evidence categories in the order in which they are evaluated, and with the point value of select criteria indicated. Clinical criteria include population data and clinical findings. Functional criteria include sequence observations, molecular studies, and indirect and computational data. ExAC, Exome Aggregation Consortium.

Figure 4.2: Nykamp *et al.*'s semi-continuous scale for weighting evidence. This system replaces the ACMG approach of combining evidence using Boolean rules. Adapted from Nykamp *et al.*, 2017, Figure 1. [19]

Nykamp *et al.* sum the score of all of the pieces of evidence to better reflect the biological complexity and uncertainty of weight assignment. The resulting score must meet a certain threshold for each classification, otherwise the variant is labelled as unknown.

Like the original guidelines, this system addresses the skew in data towards benign variants. The threshold for receiving a benign label is lower than the threshold required for a pathogenic label. This increases the volume of variants classified as benign, i.e. decreases the number of false positives. Therefore, there is higher confidence in the accuracy of a pathogenic label. Confidence in true positives is essential to ensure that a pathogenic label continues to oblige a healthcare response. Furthermore, because of the thresholds, variants are also more likely to be classified as unknown than pathogenic. Again, this decreases the number of false positives and increases the confidence in a

pathogenic assertion. Therefore, Nykamp *et al.*'s approach reflects both the complexity of variant classification and the probability that a variant observed in the gene pool is pathogenic.

To assign weights to evidence types and fit them onto the more continuous scale in Figure 4.2, Nykamp *et al.* use a variety of decision trees. The one I will focus on evaluates and weights population-based evidence.

4.1 Population-Based Evidence

Population-based evidence can be used to answer the second question related to pathogenicity: the link between a damaging variant and a disease. There are two relevant pieces of population-based evidence: allele frequency and disease prevalence. As mentioned in Chapter 2, allele frequency is the measure of the common-ness of a variant in a population. Because of natural selection, there is a correlation between allele frequency and pathogenicity.

4.1.1 Disease Prevalence

However, there are many complicating factors to the correlation between allele frequency and pathogenicity. The first is another type of population-based evidence: disease prevalence. Disease prevalence is a measure of the common-ness of a disease in the population. In the simplest case where a disease phenotype is caused by a single damaging variant, the allele frequency of the variant should be equal to the disease prevalence. In the more common case, a disease phenotype may be caused by multiple damaging variants. In this case, the sum of the putative causal allele frequencies should be equal to the disease prevalence. The causal alleles should account for all of the cases of disease observed.

Therefore, population-based evidence can be used to set an upper boundary on the

expected allele frequency of a variant. If the allele frequency of a variant is higher than the disease prevalence, this suggests that the variant is not causal for that disease. Furthermore, if the sum of the allele frequencies of "known" variants is higher than the disease prevalence, this suggests that one or more of the variants is incorrectly classified in relation to the disease.

Using population data and known variants for a disease can set a maximum expected allele frequency for an unknown variant for a disease. However, there are further considerations. A variant cannot be classified as pathogenic for a disease simply because it has an allele frequency lower than the upper boundary. Other factors include data availability and the disease's mode of inheritance, which must also be considered before allele frequency can be used as a measure of pathogenicity.

4.1.2 Mode of Inheritance

A disease's mode of inheritance impacts the expected allele frequency in the population, and therefore its correlation with pathogenicity. For an autosomal dominant disease, any individual with the dominant variant should have the affected phenotype. This means that both the homozygous dominant and the heterozygous genotypes have the affected phenotype. On the other hand, for an autosomal recessive disease, only individuals with the homozygous recessive genotype will have the affected phenotype. Therefore, because the recessive allele is more often masked by heterozygosity, a larger allele frequency can be tolerated in the population for recessive diseases. Another way to say this is that natural selection does not work as efficiently to remove recessive alleles from the gene pool.

Therefore, recessive variants may have a higher allele frequency than dominant variants. When using the correlation between allele frequency and pathogenicity to classify variants, a recessive variant may be less likely classified as pathogenic while a domi-

nant may be more likely classified as pathogenic. To combat this, the threshold of allele frequency for classifying autosomal recessive diseases and their variants as pathogenic should be higher than the threshold for dominant variants and their diseases.

4.1.3 Data Availability

Another feature that impacts allele frequency's correlation with pathogenicity is the number of alleles present in ExAC, the Exome Aggregation Consortium, or other variant databases. [15] It is the denominator of the allele frequency calculation, and it reflects the availability of the data and the confidence it provides to the analysis.

A low allele number implies that the variant site has not been observed in a large portion of the population. Sampling from a small or biased sub-population may not be representative of the whole population. A variant may be absent or under-represented in a small sample size. Another explanation of a low allele number and allele frequency could be that the gene is located in a region in the genome that is not frequently targeted by sequencing primers or is especially difficult to sequence. In fact, genes in such regions may not show up in the databases at all. However, this does not mean that these genes and variants are rare. Therefore, a low allele number decreases the accuracy of using allele frequency as a predictor for pathogenicity.

On the other hand, a high allele number implies that the variant site has been observed in a large portion of the population. There is more confidence in the data and in the possible statistical analyses. For example, if a variant is observed at a low frequency in a large sample size, there is high confidence in the allele frequency estimate. However, if a variant is observed at a low frequency in a small sample size, there is a lower chance that it is pathogenic because the allele frequency may be higher than estimated. Therefore, allele number is a method to evaluate the accuracy of allele frequency as evidence of pathogenicity.

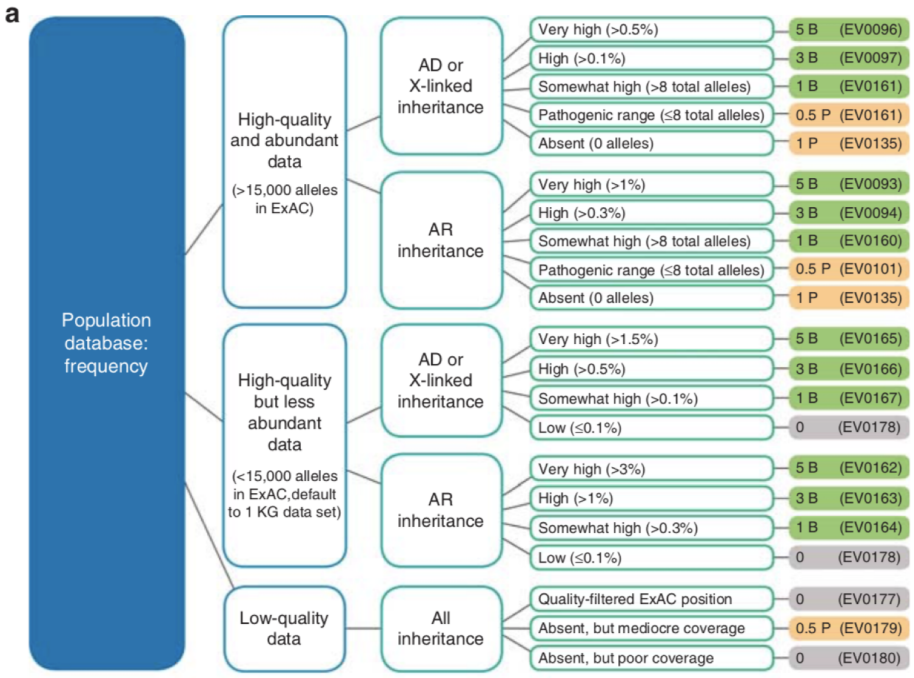

Figure 2 Population data: Sherloc criteria and decision tree. (**a**) A single evidence type criterion from the frequency set of criteria is chosen for each variant. This decision tree guides users to the correct criterion based on the quality and abundance of the Exome Aggregation Consortium (ExAC) data at the locus in question, the mode of inheritance of the gene, and the frequency of the variant in ExAC. Points and directionality (pathogenic versus benign) are indicated in the far right column. (**b**) Decision tree for using observations of homozygotes in the ExAC database depending on the

Figure 4.3: Nykamp *et al*.'s decision tree to weight allele frequency evidence. The tree uses data availability and mode of inheritance to modify allele frequency. Adapted from Nykamp *et al*., 2017, Figure 2a. [19]

4.2 From Sherloc to Machine Learning

I will focus on Nykamp *et al.*'s method of weighing allele frequency evidence using Figure 4.3. The goal of this figure is to modify the definition of an "extremely low" allele frequency presented in the original ACMG rule in Table 2.1. As Nykamp *et al.* put it, "A single threshold does not adequately capture this variable likelihood of pathogenicity." [19] The decision tree method makes the allele frequency thresholds for assigning pathogenicity more case-dependent while still retaining the quantitative and reproducible approach provided by computational predictors. This decision tree accounts for both allele number and mode of inheritance by creating branches that correspond to the different combinations of these features. Multiple allele frequency thresholds are pro-

vided for each combination of data availability and mode of inheritance, and the values of these thresholds are specific to trends that Nykamp *et al.* observed while classifying the 40,000 variants. Each branch provides different weights to assign allele frequency observations to weights on the semi-continuous scale. Therefore, Nykamp *et al.*'s approach to allele frequency makes the ACMG rule finer-grained.

Nykamp *et al.*'s method addresses the second question of pathogenicity: linking a damaging variant to a disease, in a systematic and quantitative way. However, its development required a manual and labor-intensive approach. Therefore, my goal is to incorporate Sherloc's method of modifying allele frequency to fine-tune the granularity of the original ACMG allele frequency rule with machine learning instead of manual classification. This will serve as a proof-of-concept of a hybrid approach to variant classification consisting of both machine learning and clinical approaches. The machine learning approach increases the efficiency of the system by decreasing the labor required and making the development of gene and disease-specific guidelines feasible.

CHAPTER 5

MACHINE LEARNING IMPLEMENTATION

5.1 Probabilistic Soft Logic

In this work I employed probabilistic soft logic (PSL). PSL is a machine learning programming language to infer unknown predicates in a logic network. [3, 13] A logic network consists of atoms, predicates, and rules. Atoms are observations. Predicates define relationships between atoms and take a soft truth value between 0 and 1. Rules define relationships between predicates.

PSL is used to infer truth values of unknown predicates based on a given set of rules and an observed set of atoms and predicates. One assumption of PSL networks and other statistical relational learning methods is that the known and unknown entities are related. The relationships established by the rules can be observed in the grounded atoms and predicates and are assumed to also apply to the unknown predicates. Therefore, these unknown predicates and their truth values can be inferred through these relationships.

PSL rules may have weights, which reflect the importance of each rule in the network. Rule satisfaction refers to how often the relationship between predicates dictated by a rule is followed in an assignment of truth values. A low-weighted rule is less often satisfied than a high-weighted rule. When a set of atoms, predicates, and rules are given, PSL can also learn the weights of the rules. [4] In this case, all predicate truth values are known, and instead the importance of the rules are inferred.

5.2 PSL for Variant Classification

PSL is relevant for variant classification for three reasons. First, the format of PSL rules as implicative statements with a logical structure provides a natural framework

for encoding Nykamp *et al.*'s combinations of allele number, mode of inheritance, and allele frequency. Certain combinations of these features imply a classification. Second, PSL is a type of statistical relational learning, which means that it uses relationships between entities to infer information. This is useful for variant classification because variants are classified in the context of the genome, case-specific knowledge, and other variants that have similar properties. Another way to say this is that variant classification is a collective classification problem, which is easily fit into the similarity framework of PSL. Third, PSL allows for the assignment of soft truth values for predicates instead of only binary truth values. This allows for the indirect incorporation of features into the model as truth values as opposed to as rules and reflects the uncertainty associated with some pieces of evidence.

Therefore, my goal is to incorporate multiple approaches and their advantages into a single model: the flexibility of the ACMG guidelines, the efficiency of machine learning, the specificity of Nykamp *et al.*'s approach, and the subtlety and soft truth values of PSL.

I implemented a PSL model for variant classification as follows. The atoms include the different variants and the two classes: Benign and Pathogenic. The predicate $HasCat(V, C)$ signifies the relationship between two atoms: a variant V and a classification C. This predicate takes a value between 0 and 1 that represents how likely a variant V has one of the two classes C. Each variant V has two associated predicates: $HasCat(V, Pathogenic)$ and $HasCat(V, Benign)$.

The next predicate is $HasSimilarAF(V1, V2)$. This predicate signifies the relationship between two variants. In this case, the relationship between two variants is defined by their allele frequencies. If two variants have the same allele frequency, this predicate is assigned a truth value of 1. If two variants have vastly different allele frequencies, this predicate is assigned a truth value of 0. The thresholds provided by Nykamp *et al.* in the last split of Figure 4.3 bin allele frequencies into categories.

The first rule combines these predicates.

$$HasCat(A, C)\ \&\ HasSimilarAF(A, B)\ \&\ (A! = B)\ \implies\ HasCat(B, C)$$

This rule states that if a variant A is observed to have classification C, and variants A and B have similar allele frequencies, and variant A is not variant B, then this implies that variant B is likely to have the same classification C.

This rule reflects two main assumptions. The first is related to variant classification: allele frequency is correlated with pathogenicity because of natural selection. The second is related to PSL: entities that have similar properties can be assumed to share similar relationships and therefore a similar truth value can be assigned to the unknown $HasCat$ predicate.

However, we know from the discussion in the Chapter 4 that there are many complications to using the correlation between allele frequency and pathogenicity including data availability, the disease's mode of inheritance, and disease prevalence. Nykamp *et al.* explicitly addresses two of these features by using a decision tree to categorize different cases based on these values as discussed in Chapter 4. It implicitly addresses disease prevalence by focusing on rare diseases and rare variants. [19]

My approach to address these features is to add additional rules to the PSL model.

$$HasCat(A, C)\ \&\ HasSimilarAN(A, B)\ \&\ (A! = B)\ \implies\ HasCat(B, C)$$

$$HasCat(A, C)\ \&\ HasSimilarAD(A, B)\ \&\ (A! = B)\ \implies\ HasCat(B, C)$$

These relational rules also rely on similarities between variants. AN stands for allele number and represents data availability. AD stands for autosomal dominant and represents one of two modes of inheritance. Because these factors are modifications to the more basic allele frequency rule, these rules should be less important than the first rule. Therefore, the allele frequency rule should have the highest weight. Then, mode of inheritance should have the second highest weight. Finally, the allele number rule should have a low weight because observation sample size does not directly answer any of the

three questions concerning pathogenicity. However, it does influence the confidence in allele frequency as a measure of pathogenicity and is therefore an important component of the model. The complete PSL rules for this model are shown in A.1.

The final rule is $HasCat(A, +C) = 1$. This rule ensures that the truth values of the two predicates associated with each variant sum to one.

5.3 Data and Results

My goal was to implement a proof of concept that PSL, and machine learning approaches in general, can be used to model frequency data as a predictor for pathogenicity. [3, 13]

Therefore, I generated a variety of synthetic datasets that included combinations of data availability, mode of inheritance, and allele frequency along with the expected clinical significance of each variant. I wrote a set of R scripts to convert the synthetic data into the correct input format for PSL. This involved categorizing the observations into similarity categories. I used thresholds determined by Nykamp *et al.* in Figure 4.3 to categorize allele frequencies. I used autosomal dominant and autosomal recessive as the two categories for mode of inheritance. I used a threshold of 15,000, also defined by Nykamp *et al.* in Figure 4.3 to categorize allele number observations.

In Synthetic Dataset 1, there were two allele number levels, two modes of inheritance, and two allele frequencies. The clinical significance and pathogenicity was determined solely by the allele frequencies. The lower allele frequency for each category was labelled as pathogenic, and the higher allele frequency was labelled as benign. All of these predicates had truth values of one.

Next, I ran weight learning on this training set to obtain weights for the rules. The learned weights are shown in Table 5.1 and reflect the expectations. Because I used allele frequency to assign the pathogenicity, I expected and observed that this rule had

Table 5.1: Learned weights for Synthetic Dataset 1 with 3 similarity rules and sum rule. AN = allele number, a measure of data availability; AD = autosomal dominant, a mode of inheritance; AF = allele frequency.

Rule	Weight
AN	0.019
AD	0.028
AF	1.309

the largest weight.

The purpose of this dataset was mainly to establish a feasible format for the rules and input data so that PSL could run to completion. The dataset was small (n=11) and its simplicity is not an accurate representation of actual variant data.

Therefore, I needed a larger dataset with a more systematic assignment of classifications that better represented the subtle complexity of the features allele number and mode of inheritance. For Synthetic Dataset 2, I devised a system to assign an intermediate measure of pathogenicity for allele frequency, mode of inheritance, and data availability. This intermediate measure was a measure of pathogenicity, with 1 being 100% pathogenic and 0 being 100% benign.

To assign this value, I considered the three features in order of perceived importance. The most important feature, allele frequency, should have the largest impact on the value and therefore on the classification assigned. In this dataset, I used three levels of allele frequency for each combination of mode of inheritance and allele number. I assigned the values 0.9, 0.6, and 0.3 to the lowest, moderate, and highest allele frequencies respectively to reflect the correlation between allele frequency and pathogenicity due to natural selection. Next, I added 0.05 to all variants with an autosomal dominant mode of inheritance. This reflects the fact that an autosomal dominant variant is less likely seen in the population and therefore a low allele frequency is more indicative of pathogenicity. I added this value to the hundredths place because mode of inheritance

Table 5.2: Learned weights for Synthetic Dataset 2 with 3 similarity rules and sum rule.

Rule	Weight
AN	0.033
AD	0.505
AF	1.000

Table 5.3: Performance of 3 similarity rules and sum rule on Synthetic Dataset 2. FPR = False Positive Rate; FNR = False Negative Rate.

FPR	100
FNR	0
Specificity	0
Sensitivity	100

should have less influence on classification than allele frequency. Finally, I considered allele number. I only included two values for allele number. I chose these values as above and below 15,000 variants, based on Nykamp *et al*.'s original threshold from Figure 4.3. For the value higher than 15,000, I added 0.005 to the intermediate value. This reflects the assumption that a larger number of observations increases the confidence in the correlation of allele frequency with pathogenicity.

I then used this intermediate measure to assign classifications for the training dataset. If the intermediate measure was greater than 0.5, I classified the variant as pathogenic with a truth value of 1. The learned weights are shown in Table 5.2 and reflect the expectations.

Then, I ran inference with the newly-learned weighted rules and a testing set. For each variant, I used PSL to evaluate the probability that the variant is each of the two classes. Then, I took the max of these two values as the consensus decision output by the model. I used the same system described above to assign truth values for the testing set. Finally, I assessed the performance of the model using confusion matrix values as well as specificity and sensitivity. The results are shown in Table 5.3.

Table 5.4: Learned weights and performance for 3 similarity rules and sum rule on Synthetic Dataset 3.

AN	0.538
AD	0.730
AF	1.010
FPR	0
FNR	100
Specificity	100
Sensitivity	0

This model over-predicted variants as pathogenic, as shown by the high false positive rate. I examined the training data for an explanation. In my assignment of an intermediate pathogenicity value and my use of 0.5 as a threshold for assigning pathogenicity, two out of three variants in the training set are labelled as pathogenic. This skew made pathogenic variants easier to classify than benign variants. This is problematic because over-prediction in the direction of pathogenicity can decrease the focus on treatment of truly pathogenic variants. Furthermore, this skew in the data does not reflect the trend that is observed in actual variant data that favors benign variants. To combat this overly-conservative model, I needed a new dataset.

In Synthetic Dataset 3, I returned to Nykamp *et al.*'s scheme of assigning weights. The resulting weights for all of the branches in Figure 4.3 fall between 5B and 1P on the semi-continuous scale in Figure 4.2. This means that even the most pathogenic variants that are assigned using allele frequency data provide only weak evidence for pathogenicity. Therefore, I adjusted my assignment of training classifications to almost exclusively consist of benign variants.

The weights and performance results are shown in Table 5.4. This conservative dataset produced the expected trend in weights. It also improved the false positive rate and more accurately represents the skew in data that would be observed in real variant data. However, in exchange for the false positive rate and specificity, the false negative

Table 5.5: Learned weights and performance for 3 similarity rules without sum rule on Synthetic Datasets 2 and 3.

	Synthetic Dataset 2	Synthetic Dataset 3
Rule	Weight	
AN	0.047	1.000
AD	0.114	0.108
AF	1.000	1.000
Metric	Value	
FPR	42	25
FNR	66	25
Specificity	57.14	75
Sensitivity	33.33	50

rate increases and the sensitivity decrease.

The complete opposite performance results of the previous two datasets made me question my decision to use the maximum truth value of the two $HasCat$ predicates as the consensus classification for each variant. I had originally used this approach because I knew from the last sum rule that the two truth values would sum to one. However, I questioned whether this was a reasonable approach. To test this, I removed this last sum rule and observed the effect on the results. The PSL rules for this model are shown in A.2. The results are shown in Table 5.5.

This approach achieves a better balance between specificity and sensitivity than the model that includes the sum rule. The confusion matrix values for Synthetic Dataset 3 are improved compared to those for Synthetic Dataset 2, which suggests that the approach of calling the vast majority of variants in the training set benign is successful. The removal of the sum rule constraint allowed PSL to better capture the complexities of mode of inheritance and allele number without the extra relationship between predicates of a single variant. However, the learned weights for Synthetic Dataset 3 do not reflect the expectations. Therefore, another approach is required.

As an extension of manipulating the importance of rules using synthetic data, I wanted to address the allele number rule from the side of the model design. Data availability is not a population-based piece of evidence and does not relate to nor answer any of the three questions concerning pathogenicity. Therefore, the allele number should have negligible weight in the weight learning step. I observed that this rule has the lowest weight in most of the tests, which suggests both that the synthetic datasets are successfully reflecting the intended relative weights of the rules and that this low-impact rule may be easily removed from the network.

Therefore, I removed the allele number rule from the model. I assessed the effects both with and without the sum rule. The complete PSL rules for these models are shown in A.3 and A.4. To retain this factor's effect on allele frequency, I needed a more continuous range of allele numbers in the training data. I made intervals of 1,000 between 10,000 and 20,000 for the Synthetic Dataset 4 allele number values. I used the conservative approach from Synthetic Dataset 3 to assign variant classifications.

I also needed to modify the process of writing data into PSL format. I applied min-max scaling to the allele number value. Min-max scaling normalizes values to a scale between 0 and 1 based on the minimum and maximum values present in the dataset. Instead of grouping variants into similar allele number categories, I used this scaled measure for the truth values for the observed $HasCat$ predicates. Therefore, variants with a higher allele number had a higher certainty than variants with a lower allele number. I also ran the other models on this dataset for comparison, and all results are shown in Table 5.6.

The resulting weights follow the expectations, with the allele frequency rule being much more important than the mode of inheritance rule. The "no AN no sum" model has the best balance between specificity and sensitivity, as well as the lowest confusion matrix values. This suggests that the incorporation of allele number indirectly was more

Table 5.6: Learned weights and performance of 4 models on Synthetic Dataset 4. The left column has all 3 rules, while the right column has only 2 rules. The AN rule is removed. The top row has the sum rule, while the bottom row does not have the sum rule.

	Sum Rule Models	
	AN	**No AN**
Rule	Weight	
AN	0.505	
AD	0.505	0.083
AF	1.000	1.206
Metric	Value	
FPR	0	40
FNR	100	100
Specificity	100	60
Sensitivity	0	0
	No Sum Rule Models	
	AN	**No AN**
Rule	Weight	
AN	0.703	
AD	0.814	0.027
AF	1.000	1.052
Metric	Value	
FPR	0	25
FNR	100	50
Specificity	100	75
Sensitivity	0	50

effective than having a separate rule for reflecting the small impact of data availability on classification.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The use of probabilistic soft logic for variant classification was relatively successful. The final "no AN no sum" model both reflected the expected importance of the features that impact allele frequency and had relatively successful performance at classifying unknown variants. However, there is still much progress to be made, both in the specific application of PSL to population data weighting as well as in the use of machine learning for variant classification. This thesis has served as a very basic proof of concept that this is a feasible approach that can be expanded in the future.

One expansion could be to use real variant data collected from ClinVar and genomAD, databases that provide existing classification assertions as well as population data. [14] This is an essential extension to observe whether PSL can extract the expected trends in allele frequency from background noise. A few additional testing datasets could include Nykamp *et al.*'s data along with benchmark datasets such as HumVar and HumDiv. These datasets could improve the validity of comparing machine learning tools to manual methods as well as comparing "damaging" classifiers to "pathogenic" classifiers.

Another expansion could be to make the remaining predicate values continuous as well to better address the uncertainty associated with variant classification. Just as the allele number rule impacted the truth values of the $HasCat$ predicate, the distance between allele frequencies could influence the truth values of the $HasSimilarAF$ predicates. This could improve how variant classification methods deal with uncertainty. Modifications that could be simulated in a systematic way could help experts better understand how their expertise can impact classification. This could assist in future classification efforts by improving the reproducibility of these judgement calls.

Therefore, data-driven approaches to variant classification have a long way to go.

The inclusion of more data relevant to the questions of pathogenicity can increase the automation of the process and decrease human labor hours. These more efficient methods could assist geneticists in their application of case-specific knowledge to variant classification.

APPENDIX A

PSL MODELS

```
1.0: HasCat(A, C) & HasSimilarAN(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAN(B, A) & (A != B) >> HasCat(B, C) ^2

1.0: HasCat(A, C) & HasSimilarAF(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAF(B, A) & (A != B) >> HasCat(B, C) ^2

1.0: HasCat(A, C) & HasSimilarAD(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAD(B, A) & (A != B) >> HasCat(B, C) ^2

// Per category rules
1.0: HasCat(A, 'Path') & HasSimilarAN(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAN(A, B) >> HasCat(B, 'Benign')^2

1.0: HasCat(A, 'Path') & HasSimilarAF(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAF(A, B) >> HasCat(B, 'Benign')^2

1.0: HasCat(A, 'Path') & HasSimilarAD(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAD(A, B) >> HasCat(B, 'Benign')^2

// Ensure that HasCat sums to 1
HasCat(A, +C) = 1 .
```

Listing A.1: "AN Sum" model rules. The file model.PSL is input along with the data into the PSL command line tool. Each similarity rule has an inverse, so there are six rules. However, only the forward rules received non-default weights during the weight learning step. The other six rules are category-specific. The final sum rule ensures the truth values of the two predicates associated with each variant sum to 1.

```
1.0: HasCat(A, C) & HasSimilarAN(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAN(B, A) & (A != B) >> HasCat(B, C) ^2

1.0: HasCat(A, C) & HasSimilarAF(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAF(B, A) & (A != B) >> HasCat(B, C) ^2

1.0: HasCat(A, C) & HasSimilarAD(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAD(B, A) & (A != B) >> HasCat(B, C) ^2

// Per category rules
1.0: HasCat(A, 'Path') & HasSimilarAN(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAN(A, B) >> HasCat(B, 'Benign')^2

1.0: HasCat(A, 'Path') & HasSimilarAF(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAF(A, B) >> HasCat(B, 'Benign')^2

1.0: HasCat(A, 'Path') & HasSimilarAD(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAD(A, B) >> HasCat(B, 'Benign')^2
```

Listing A.2: “AN No Sum” model rules. The same as A.1, except the final sum rule has been removed.

```
1.0: HasCat(A, C) & HasSimilarAF(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAF(B, A) & (A != B) >> HasCat(B, C) ^2

1.0: HasCat(A, C) & HasSimilarAD(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAD(B, A) & (A != B) >> HasCat(B, C) ^2

// Per category rules
1.0: HasCat(A, 'Path') & HasSimilarAF(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAF(A, B) >> HasCat(B, 'Benign')^2
1.0: HasCat(A, 'Path') & HasSimilarAD(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAD(A, B) >> HasCat(B, 'Benign')^2

// Ensure that HasCat sums to 1
HasCat(A, +C) = 1 .
```

Listing A.3: "No AN Sum" model rules. The AN rule and its inverse have been removed. The final sum rule is included.

```
1.0: HasCat(A, C) & HasSimilarAF(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAF(B, A) & (A != B) >> HasCat(B, C) ^2

1.0: HasCat(A, C) & HasSimilarAD(A, B) & (A != B) >> HasCat(B, C) ^2
1.0: HasCat(A, C) & HasSimilarAD(B, A) & (A != B) >> HasCat(B, C) ^2

// Per category rules
1.0: HasCat(A, 'Path') & HasSimilarAF(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAF(A, B) >> HasCat(B, 'Benign')^2

1.0: HasCat(A, 'Path') & HasSimilarAD(A, B) >> HasCat(B, 'Path') ^2
1.0: HasCat(A, 'Benign')& HasSimilarAD(A, B) >> HasCat(B, 'Benign')^2
```

Listing A.4: "No AN No Sum" model rules. The same as A.3, except the final sum rule has been removed.

BIBLIOGRAPHY

[1] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Richard Gibbs, Green ED, Hurles ME, Bartha Knoppers, Korbel JO, Lander ES, Charles Lee, Hans Lehrach, and Schloss JA. A global reference for human genetic variation. 526:68, October 2015.

[2] Ivan Adzhubei, Daniel Jordan, and Shamil Sunyaev. *Current Protocols in Human Genetics*, chapter Predicting Functional effect of Human Missense Mutations Using PolyPhen-2. John Wiley & Sons, 2012.

[3] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 2017.

[4] Stephen H. Bach, Bert Huang, Ben London, and Lise Getoor. Hinge-loss markov random fields: Convex inference for structured prediction. *CoRR*, abs/1309.6813, 2013.

[5] H. Duzkale, J. Shen, H. Mclaughlin, A. Alfares, M. Kelly, T. Pugh, B. Funke, H. Rehm, and M. Lebo. A systematic approach to assessing the clinical significance of genetic variants. *Clinical Genetics*, 84(5):453–463, Nov 2013.

[6] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.

[7] Kimon Frousios, Costas S. Iliopoulos, Thomas Schlitt, and Michael A. Simpson. Predicting the functional consequences of non-synonymous dna sequence variants - evaluation of bioinformatics tools and development of a consensus strategy. *Genomics*, 102(4):223 – 228, 2013.

[8] Rajarshi Ghosh, Ninad Oak, and Sharon Plon. Evaluation of *in silico* algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biology*, 18(225), Nov 2017.

[9] Lily Hoffman-Andrews. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *Journal of Law and the Biosciences*, 4(3):648–657, 2017.

[10] Tao Huang, Ping Wang, Zhi-Qiang Ye, Heng Xu, Zhisong He, Kai-Yan Feng, LeLe Hu, WeiRen Cui, Kai Wang, Xiao Dong, Lu Xie, Xiangyin Kong, Yu-Dong Cai,

and Yixue Li. Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. *PLOS ONE*, 5(7):1–7, July 2010.

[11] N. M. Ioannidis, J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, and D. Karyadi. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *American journal of human genetics.*, 99(4):877–885, Oct 2016.

[12] Kleinberger J, Maloney KA, Pollin TI, and Jeng LJ. An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. 2016.

[13] Angelika Kimmig, Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 2012.

[14] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

[15] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, and Beryl B. Cummings. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536:285–291, Aug 2016.

[16] Quan Li and Kai Wang. InterVar: Clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *The American Journal of Human Genetics*, 100(2):267 – 280, 2017.

[17] D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, and E. A. Ashley. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508:469–476, Apr 2014.

[18] Khalid Mahmood, Chol-hee Jung, Gayle Philip, Peter Georgeson, Jessica Chung, Bernard J. Pope, and Daniel J. Park. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human Genomics*, 11(1):10, May 2017.

[19] K. Nykamp, M. Anderson, M. Powers, J. Garcia, B. Herrera, Y.Y. Ho, and S. Topper. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine*, 19(10):1105–1117, 2017.

[20] Ronak Y. Patel, Neethu Shah, Andrew R. Jackson, Rajarshi Ghosh, Piotr Pawliczek, Sameer Paithankar, Aaron Baker, Kevin Riehle, Hailin Chen, Sofia Milosavljevic, Chris Bizon, Shawn Rynearson, Tristan Nelson, Gail P. Jarvik, Heidi L. Rehm, Steven M. Harrison, Danielle Azzariti, Bradford Powell, Larry Babb, Sharon E. Plon, and Aleksandar Milosavljevic. ClinGen pathogenicity calculator: a configurable system for assessing pathogenicity of genetic variants. *Genome Medicine*, 9(1):3, Jan 2017.

[21] Sue M. Richards, Nazneen Aziz, Sherri J. Bale, David Bick, Soma Das, Julie M. Gastier-Foster, Wayne W. Grody, Madhuri R. Hegde, Elaine Lyon, Elaine Spector, Karl V. Voelkerding, and Heidi L Rehm. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17:405–423, 2015.

[22] Naisha Shah, Ying-Chen Claire Hou, Hung-Chun Yu, Rachana Sainger, Eric Dec, Brad Perkins, C. Thomas Caskey, J. Craig Venter, and Amalio Telenti. Identification of misclassified clinvar variants using disease population prevalence. *bioRxiv*, 2016.

[23] Dhanya Sridhar, Shobeir Fakhraei, and Lise Getoor. A probabilistic approach for collective similarity-based drug-drug interaction prediction. *Bioinformatics*, 32(20):3175–3182, 2016.

[24] Mahadeo A. Sukhai, Kenneth J. Craddock, Mariam Thomas, Aaron R. Hansen, Tong Zhang, Lillian Siu, Philippe Bedard, Tracy L. Stockley, and Suzanne Kamel-Reid. A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. *Genetics in Medicine*, 18:128–136, Apr 2015.