# Advancing Variant Interpretation

A Gene-Specific Framework for Prioritization, Prior Estimation, and Calibration to Enhance Evidence Strength and Clinical Significance Classification

Yile Chen

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

John Gennari, Chair

Sean D. Mooney

Vikas R. Pejaver

Lea M. Starita

Program Authorized to Offer Degree:

Biomedical Informatics & Medical Education

University of Washington

## Abstract

# Advancing Variant Interpretation: A Gene-Specific Framework for Prioritization, Prior Estimation, and Calibration to Enhance Evidence Strength and Clinical Significance Classification

Yile Chen

Chair of the Supervisory Committee:
John Gennari
Department of Biomedical Informatics & Medical Education

The rapid growth of clinical sequencing has led to an accelerating number of Variants of Uncertain Significance (VUS), now comprising a substantial fraction of reported germline findings. While functional assays and computational Variant Effect Predictors (VEPs) contribute valuable evidence, current frameworks often treat genes uniformly, overlook gene-level heterogeneity in pathogenicity prevalence, and rely on uncalibrated or globally calibrated predictor scores. These gaps limit the consistency, accuracy, and clinical actionability of variant interpretation under ACMG/AMP guidelines. There is a pressing need for approaches that incorporate gene-specific context, integrate diverse evidence sources, and improve the calibration of computational evidence to strengthen variant classification.

This dissertation introduces gene-specific informatics frameworks to improve functional assay prioritization, pathogenicity prior estimation, and the calibration of Variant Effect Predictors (VEPs), with the goal of reducing the burden of VUS in genomic medicine. By

integrating statistical modeling, positive–unlabeled learning, domain-aware clustering, and adaptive calibration strategies, the work strengthens the ACMG/AMP Bayesian framework for context-aware variant interpretation.

First, a gene prioritization model identifies genes where functional assays would yield the greatest clinical impact by jointly optimizing VUS "movability," correction of potential misclassifications, and gains from computational predictors, highlighting high-value genes such as *TSC2*. Second, gene-specific pathogenicity priors are estimated using a refined PU-learning method (DistCurve), supported by a complementary domain-based clustering approach for genes with limited labels. Third, a gene-aware calibration framework converts raw VEP scores into calibrated PP3/BP4 evidence strengths through a dynamic decision-tree workflow that selects the optimal strategy per gene. This gene-specific approach outperforms global calibration and, together with a per-gene mixed-predictor selection strategy, improves the accuracy and consistency of variant classification.

Together, these contributions establish a context-aware decision-support ecosystem that better directs functional assay investment, provides robust statistical foundations for Bayesian interpretation, and improves the reliability of computational evidence. The resulting framework enhances the accuracy, consistency, and clinical actionability of genomic variant classification.

## Acknowledgments

I would first like to thank the two people who have supported me most directly throughout this dissertation. Vikas Pejaver has been an extraordinary mentor—initially helping with individual projects and eventually guiding me through nearly every stage of my research. His consistent meetings, thoughtful suggestions, and unwavering support have shaped much of this work. I am equally grateful to Shantanu Jain from Predrag's team, whose countless valuable insights and feedback have strengthened every project we collaborated on. I also want to thank Lea Starita, who made me feel at home when I transitioned into her lab, welcomed me fully, and took responsibility for keeping me on track. My appreciation extends to Sean and Predrag for their project supervision and advice, and to John, who served as committee chair and provided essential administrative guidance as well as support during this final step toward graduation. Finally, I would like to thank Andrew for serving as the GSR and offering helpful suggestions along the way.

I also thank our collaborators whose contributions strengthened this work, including the Radivojac Team (Daniel Zeiberg), the Starita Team (Shawn Fayer, Malvika Tejura, Mariam Benazouz, and Jeremy Stone), and the Pejaver Team (Himanshu Sharma and Tim Bergquist). Many thanks as well to my peers, friends, and the broader scientific community for their support. I am especially grateful to the IGVF Consortium for funding this research and for supporting my development throughout my PhD.

To my family — Mom and Dad — thank you for your unconditional encouragement. To my partner, Daoyao Yang, thank you for your patience, support, and belief in me. Finally, in the wise words of Snoop Dogg, I also want to thank me — for believing in me, for doing all this hard work, for never quitting, and for pushing through every challenge to reach this moment.

# Contents

# Chapter 1

# Introduction

## 1.1 Background and Significance

Genome medicine relies on accurate interpretation of genomic variants to guide patient care. A critical step is classifying variants as pathogenic (disease-causing) or benign, so that clinicians know which findings are actionable. However, variant classification remains challenging: a large fraction of observed variants are categorized as "variants of uncertain significance" (VUS) due to insufficient or conflicting evidence. For example, many variants in public databases like ClinVar have multiple laboratory submissions with discrepant interpretations, underscoring the inconsistency in current classification efforts (Richards et al., 2015; Yang et al., 2017). To address this, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) introduced a standardized framework that defines multiple lines of evidence (e.g., population data, functional assays, computational predictions) each weighted as supporting, moderate, strong, or very strong toward pathogenic or benign classification (Richards et al., 2015). This framework has improved consistency, but it also highlights the evidence gaps—particularly for missense variants that often end up as VUS because available data are limited or contradictory. Filling these gaps is essential for realizing the promise of personalized genome medicine.

One promising avenue to resolve VUS is to generate new functional evidence through experimental assays. High-throughput experiments, such as multiplexed assays of variant effect (MAVE), can experimentally measure the impact of thousands of mutations in a gene (Findlay, 2018; Fowler and Fields, 2023). These assays provide strong evidence for variant as pathogenic or benign (ACMG criteria PS3/BS3) when properly validated. Indeed, well-designed functional studies have been shown to reclassify VUS into definitive categories by directly testing variant effects (Gelman, 2019; Brnich, 2019). However, functional assays are resource-intensive and cannot feasibly be performed for every gene or variant. This makes

it crucial to prioritize which genes or regions to target experimentally. Historically, gene selection for functional studies often favored well-studied genes or those already suspected to be important (Stoeger et al., 2018b), a bias that risks leaving many medically relevant genes under-characterized. A more systematic, data-driven approach is needed to identify the "right" genes where functional assays would yield the greatest clinical benefit (for example, by resolving many VUS or correcting likely misclassifications). By quantifying the potential clinical impact of resolving variant uncertainty in each gene, we can guide functional genomics resources to where they will make the most difference, which serves as one of the goal of my dissertation research.

Another key avenue is improving computational predictions of variant pathogenicity. Computational variant effect predictors (VEPs) have rapidly advanced and can score any missense variant based on features like sequence conservation, biochemical change, protein structure, and learned patterns from known variants. Tools such as REVEL (Ioannidis et al., 2016) (an ensemble of multiple classifiers) and MutPred2 (Pejaver, 2020) (which offers mechanistic hypotheses) have demonstrated high accuracy, with area-under-curve values often above 0.85 for distinguishing pathogenic vs benign variants. More recently, deep learning approaches like AlphaMissense (Cheng, 2023) have pushed performance even further (reporting AUC $\sim$0.94 on ClinVar missense variants), approaching the reliability needed for clinical use. Despite these improvements, a gap remains between raw predictor scores and the categorical evidence framework used in clinical variant interpretation. The 2015 ACMG/AMP guidelines treated computational predictions as supporting evidence at best, reflecting concerns about false positives (Richards et al., 2015). In 2022, however, a landmark study demonstrated that with proper calibration, some predictors' scores can indeed correspond to higher evidence strengths (moderate or strong) for pathogenicity or benignity (Pejaver, 2022). In other words, if we determine appropriate score thresholds, a computational tool's output can be translated into the same language of evidence that clinicians use. This is a significant step toward making computational predictions directly clinically actionable. Thus, a goal of my dissertation is to create a robust generalizable way to calibrate these tools for VEP. Calibration must account for differences across genes and variant contexts; a one-size-fits-all threshold may not work for all genes, especially given that genes vary in their tolerance to variation and the amount of data available.

In summary, the motivation for this dissertation is rooted in the pressing need to better classify and interpret genetic variants. The high prevalence of VUS in clinical sequencing reports is a barrier to delivering definitive diagnoses and personalized treatments. By leveraging functional assays in a targeted manner and computational predictors in a calibrated, interpretable manner, we aim to substantially improve variant classification.

## 1.2 Proposed Approach

To address these challenges, this dissertation introduces a two-part framework that improves variant classification by integrating gene prioritization, statistical learning, and calibration methods to better connect raw data with clinical decision-making. Broadly, the work is organized into three parts: (1) prioritizing genes for functional assays, (2) estimating gene-specific prior probabilities of pathogenicity, and (3) calibrating computational predictor scores on a gene-by-gene basis. An additional innovation – domain-based clustering – is introduced to complement the third part, addressing a key limitation of purely single-gene calibration.

**Gene Prioritization for Functional Assays:** In the first part of this work, I have developed a framework to quantitatively rank genes based on the potential clinical impact of resolving their variant uncertainty. The idea is to identify genes where additional functional data (from MAVEs or other assays) would yield the greatest benefit in terms of variant reclassification. I formalized this as an optimization problem combining different objectives with multiple criteria. For example, one criterion is the number of VUS or conflicting-interpretation variants in the gene that could potentially be moved to a definitive classification with new evidence (Kuang, 2021). Another is the potential to improve computational predictions for that gene, since the opportunity to improve is greater where current predictors have low confidence or high error. A third factor is the practical feasibility of assays, which may be influenced by gene size, expression, or the availability of suitable model systems. Our prioritization framework integrates these factors into a single gene impact score. This scoring system is designed to be flexible and data-driven: as new information becomes available in public databases (ClinVar, gnomAD, etc.), the scores can be automatically updated (Chen, 2023). The outcome is a ranked list of genes indicating where limited experimental resources should be directed. My framework helps researchers move beyond ad hoc gene choices and ensures that functional genomics efforts are aligned with clinical needs. Notably, this approach can refine and extend prior strategies like the Difficulty-Adjusted Impact Score (Kuang, 2021) by incorporating additional objectives (like computational predictor improvement) that have not been jointly considered before. My colleagues and I have demonstrated the value of this work with case studies; for instance, our analysis surfaced the gene *TSC2* as a high-priority candidate for MAVE experiments, a finding that influenced its selection by collaborators for functional analysis (Chen, 2023).

**Gene-Specific Pathogenicity Priors:** In the second part of this work, I focus on the foundational probabilities that underlie variant interpretation. In a Bayesian view of the ACMG/AMP framework, each variant's classification depends on the prior probability of pathogenicity for that variant or gene and on the likelihood ratios contributed by each

line of evidence (Tavtigian, 2018). Historically, a generic prior between 0.01 and 0.1 has been assumed for moderate-impact variants (Tavtigian, 2018; Pejaver, 2022). In practice, however, the true prior probability that a variant is disease-causing can differ greatly from gene to gene. Genes associated with highly penetrant dominant disorders or critical cellular functions are expected to have higher pathogenic variant rates, whereas more tolerant genes contribute fewer truly pathogenic variants. In this dissertation, I estimate gene-specific prior probabilities of pathogenicity for a large set of disease-associated genes. To do this at scale, I use positive–unlabeled learning techniques that infer class prevalence from a combination of known pathogenic variants and a background of unlabeled variants drawn from population datasets. The full methodological framework, including the adaptation of DistCurve and related bias-aware strategies, is described in Chapter 3, Section 3.3. I rigorously evaluate the resulting priors for plausibility and consistency across genes, for example by confirming that genes with well-established disease burden such as *BRCA1* and *TP53* receive higher priors than genes implicated in milder or rarer conditions. The outcome is a gene-specific catalog of pathogenicity priors that can be used directly within the Bayesian ACMG/AMP model in place of one-size-fits-all assumptions, thereby tailoring variant interpretation to each gene's context. These priors also provide the baseline probabilities needed to calibrate raw predictor scores into posterior probabilities, as I do in Chapter 4.

**Gene-Specific Score Calibration:** In the third part of this work, I focus on translating computational predictor scores into calibrated evidence strengths on a per-gene basis. Building on my new gene-specific priors, I implement a calibration pipeline that converts a predictor's raw score (for a given variant in a given gene) into a posterior probability that the variant is pathogenic. In essence, I map each predictor score to how likely it is to correspond to a pathogenic variant within the context of a specific gene. The evaluation of the local posterior calibration approach is presented in detail in Chapter 4, Section 4.3.1. I begin by performing a systematic parameter sweep for the local calibration method, and I quantify performance using ACMG-aligned misestimation metrics such as the fraction of variants whose PP3 or BP4 evidence is overstated. I then benchmark this local calibration strategy against a panel of ten established post-hoc calibration methods under a wide range of simulated data conditions, including small sample sizes and strong class imbalance. These experiments show that no single method is uniformly optimal: approaches that work well for data-rich genes can behave poorly when labeled variants are sparse. Motivated by this heterogeneity, I design a dynamic decision-tree calibration workflow that selects the most appropriate calibration method for each gene. This workflow incorporates gene-specific priors, score distribution characteristics, and misestimation risk to choose a method and then outputs calibrated score thresholds corresponding to ACMG evidence strengths (supporting,

moderate, strong). The result is a set of gene-specific, clinically interpretable cutoffs that minimize the risk of overstated pathogenicity evidence and provide practical guidance for applying computational evidence at the gene level in line with ACMG/AMP guidelines.

**Domain-Based Clustering to Enhance Calibration:** A key insight from my research is that purely gene-specific calibration, while ideal in theory, may face limitations in practice for genes with very limited data. Some genes have only a handful of known pathogenic or benign variants in ClinVar, making it difficult to reliably estimate score-to-outcome mappings. To address this, I created a novel domain-based clustering strategy as a complementary solution. The idea is to group genes into clusters based on shared characteristics such that variants in genes within the same cluster might be expected to behave similarly. By performing calibration at the level of these gene clusters, we effectively increase the data pool for estimating score thresholds, improving statistical power for calibration in data-scarce genes. In collaboration with colleagues, we identified meaningful gene clusters (e.g., groups of genes sharing a particular functional domain or involved in the same protein complex) and computed a cluster-level pathogenicity prior for each group by aggregating data from all member genes. We then applied a dynamic decision-tree-based algorithm to calibrate predictor scores within each cluster. The outcome is a set of calibrated thresholds for each cluster, which can be mapped back to individual genes in that cluster. Crucially, this domain-based approach preserves gene-specific nuances by clustering only genes with biological similarity, rather than pooling arbitrary genes together. It serves as an intermediate resolution between a single global calibration, where all genes are treated the same and gene-specific effects are missed, and per-gene calibration, which can easily overfit or become unstable when only limited data are available for a gene. By evaluating the cluster-level calibration against known variants in ClinVar and ClinGen and comparing it to the single-gene calibration, we show that my approach improves reliability for low-data genes while still capturing gene-specific differences better than a universal model. In summary, the domain-based clustering provides a practical and innovative extension to our calibration framework, ensuring that even for less-studied genes, computational predictions can be interpreted on a calibrated, probabilistic scale. The details of the clustering methodology and its validation are presented in the Chapter 4.

Taken together, these three parts — gene prioritization, gene-specific priors, and calibrated predictors, augmented by clustering — form a cohesive strategy to improve variant classification. The approach moves from broad questions about which genes and variants to focus on to granular decisions about how to interpret a given variant's score in context, all under the unifying goal of reducing uncertainty in genomic medicine. By integrating new data-driven methods at each step, we aim to significantly enhance the accuracy, consistency,

and actionability of variant interpretations.

## 1.3 Dissertation Contributions and Organization

This dissertation makes several contributions to the field of variant interpretation and proposes an integrated framework for improving the classification of genomic variants. Each of the main contributions corresponds to a chapter.

In Chapter 2, I develop a clinical impact–based framework to prioritize genes for functional assay investment. I design a quantitative scoring system that combines multiple criteria, including the number of unresolved variants of uncertain significance, the likelihood that functional data could reclassify those variants, and the potential for improving computational predictors. I implement an automated pipeline that updates scores as public databases are refreshed, so priorities can change as new evidence accumulates. This chapter establishes a systematic way to rank genes by expected clinical benefit and to focus experimental resources on the genes where additional evidence is likely to have the greatest impact. I also illustrate how this framework can support large collaborative efforts, for example by highlighting *TSC2* as a strong candidate for multiplex functional assays.

In Chapter 3, I shift to estimating gene-specific pathogenicity priors. I design and implement a scalable pipeline that uses positive–unlabeled learning on ClinVar and population variant data to infer the background rate of disease-causing variants for each gene. This replaces traditional universal priors with gene-specific values that are better aligned with known disease biology. I evaluate these priors for plausibility and consistency by comparing them to known disease mechanisms and published estimates. The resulting catalog of gene-specific priors can be plugged directly into Bayesian variant classification and also serves as a key input for later calibration of computational predictors.

In Chapter 4, I focus on gene-specific calibration of variant effect predictor scores. Building on the gene-specific priors from Chapter 3, I construct a calibration pipeline that converts raw predictor scores into posterior probabilities of pathogenicity for each gene. This allows me to define score thresholds that correspond to ACMG and AMP evidence strengths PP3 and BP4 for individual genes and predictors. I first carry out a detailed evaluation of the local posterior calibration approach and then benchmark it against ten established post-hoc calibration methods under a wide range of simulated data conditions. These analyses show that no single calibration method is optimal for every gene or data regime. In response, I design a dynamic decision tree workflow that selects the most appropriate calibration strategy for each gene based on sample size, score distribution, prior, and misestimation risk. In the later part of this chapter, I extend this idea to domain-based cluster calibration. Here, I work

with biologically informed clusters of genes and domains, estimate cluster-level priors, and apply the same dynamic calibration logic to these clusters. This extension improves robustness for genes with limited data while still respecting important gene-level differences and broadens the reach of calibrated computational predictions to a much larger set of Mendelian disease genes.

Finally, in Chapter 5, I summarize the main findings and discuss their broader implications for genome medicine. I reflect on how the three components of this work—clinical impact–based prioritization, gene-specific priors, and gene- and cluster-aware calibration—fit together into a coherent framework for reducing uncertainty in variant interpretation. I also outline limitations of the current approach and highlight several directions for future research, including new sources of features for prior estimation and semi-supervised strategies that share information across genes. I conclude by describing how the tools, pipelines, and catalogs developed in this dissertation can be adopted by clinical laboratories and research consortia to improve variant classification in practice.

In sum, I have provided an overview of the motivations, approaches, and contributions of this dissertation. The subsequent chapters will delve into each component in detail, demonstrating how together they advance the goal of more accurate and actionable genomic variant interpretation. Through this work, we aim to reduce the prevalence of "uncertain" variants and enable more confident, evidence-driven decisions in genomic medicine.

# Chapter 2

# Gene Prioritization for Functional Assays

## 2.1 Introduction: Prioritizing Functional Assays for Clinical Impact

Accurate clinical interpretation of genetic variants is essential for precision medicine, yet a substantial fraction of variants—particularly missense variants—remain categorized as uncertain (VUS) due to limited supporting evidence (Kuang et al., 2020). The ACMG/AMP guidelines (Richards et al., 2015) provide a structured framework for variant interpretation, integrating multiple forms of evidence including population frequency, segregation, computational prediction, and experimental data. Among these, functional assays provide some of the strongest evidence (PS3/BS3), and high-quality experimental results can substantially shift variant classifications and reduce uncertainty (Gelman, 2019; Brnich, 2019).

Multiplex assays of variant effect (MAVEs) have transformed functional genomics by enabling comprehensive measurement of the impact of nearly all possible amino acid substitutions within a gene (Findlay, 2018; Fowler and Fields, 2023). Although MAVEs hold exceptional promise for resolving VUS at scale, they remain resource-intensive, making strategic gene selection crucial for maximizing clinical impact.

Traditionally, the choice of genes for functional studies has been guided by research-driven priorities—such as studying structure–function relationships (Romero and et al., 2015), focusing on biologically or medically significant genes (Jia and et al., 2021), or developing new experimental technologies (Matreyek and et al., 2018). While valuable, such expert-driven selection tends to favor well-characterized genes, perpetuating a bias toward already well-studied regions of the genome (Stoeger et al., 2018b).

With the increasing availability of ClinVar data, it is now possible to pursue data-driven strategies that prioritize genes based on clinical uncertainty and potential for reclassification. ClinVar categorizes variants by clinical significance—pathogenic (P), likely pathogenic (LP), uncertain significance (VUS), likely benign (LB), and benign (B)—including variants with conflicting interpretations. Among these, VUS and conflicting variants represent key opportunities where functional assays can provide decisive evidence.

To formalize this concept, Kuang (2021) proposed *movability*, a measure of the likelihood that new functional evidence could change a VUS classification. Their Difficulty-Adjusted Impact Score (DAIS) prioritizes genes with multiple clinically observed VUS (adjusted for gene length) and with high potential for reclassification, thereby optimizing the use of MAVE resources to address clinical uncertainty directly.

Despite these advances, existing prioritization frameworks largely overlook computational pathogenicity predictors—such as REVEL, MutPred2, and AlphaMissense—which integrate features of evolutionary constraint, protein structure, and sequence context (Pejaver, 2020; Ioannidis et al., 2016). These predictors are formally recognized in the ACMG/AMP guidelines as sources of supporting evidence (PP3/BP4), and recent work has shown that, when calibrated, some predictors can reliably reach *strong* evidence thresholds for being pathogenic or benign. (Pejaver, 2022). Incorporating calibrated computational predictions into gene prioritization, therefore, represents an opportunity to combine two powerful and complementary evidence sources—computational and functional—to maximize the clinical utility of MAVE experiments.

In this chapter, we present a unified, data-driven gene prioritization framework that integrates clinical variant data with calibrated computational pathogenicity predictors to identify genes where functional assays will have the greatest expected impact on clinical variant interpretation. This approach directly supports the goals of the IGVF (Impact of Genomic Variation on Function) consortium by guiding experimental resources toward genes where functional data can most effectively resolve uncertainty. As a demonstration of its practical value, the framework identified *TSC2*, a gene implicated in Tuberous Sclerosis Complex, as a high-priority candidate for multiplex variant effect mapping; the Center for Actionable Variant Analysis (CAVA) subsequently selected *TSC2* for experimental profiling, illustrating how quantitative prioritization can drive actionable experimental design.

a gene prioritization framework that integrates clinical variant data and computational pathogenicity predictors to identify genes where functional assays will have the greatest impact on clinical classification. This framework directly supports the IGVF (Impact of Genomic Variation on Function) consortium's mission to maximize the clinical utility of functional genomics data. As a demonstration, the framework identified *TSC2*—a gene im-

plicated in Tuberous Sclerosis Complex—as a high-priority target for MAVE studies, and the Center for Actionable Variant Analysis (CAVA) subsequently selected *TSC2* for experimental profiling, illustrating how data-driven prioritization translates into actionable experimental design.

## 2.2 Methodology

### 2.2.1 Data collection

**ClinVar variants.** We extracted all missense variants in ClinVar (October 2021) and separated them by the recorded the numbers of variants in each category of clinical significance: Pathogenic (P), Likely Pathogenic (LP), Benign (B), Likely Benign (LB), variants of uncertain significance (VUS), and variants with conflicting interpretations of pathogenicity for each gene. The ClinVar data set contained 11,281 genes with 402,721 missense variants (Supp. Table 1).

**gnomAD variants.** VUS in ClinVar are likely to accumulate in a biased manner due to differences in the frequency with which different genes are tested. At the gene-level, variants in population-scale sequencing resources such as gnomAD accumulate in a less biased manner as all genes are likely to be uniformly sampled. To this end, we extracted missense variants from gnomAD (v2.1.1 GRCh38 dataset) as an additional set of variants that are not annotated as P, LP, B or LB (Karczewski et al., 2020). Only variants with genotype quality (GQ) $\geq 20$ and depth (DP) $\geq 10$ were retained. We identified 17,988 genes that had 4,542,252 missense variants.

**Genes with MAVEs.** We extracted genes from three resources: MaveDB (Esposito et al., 2019), VariantEffect (`https://github.com/VariantEffect/MaveReferences`), and MaveRegistry (Kuang et al., 2021) to create a representative set of genes with functional data. The first two record and maintain information on which genes have been subject to MAVEs either by submission to the resource or by reviewing the literature. MaveRegistry hosts information on which genes are currently being assayed or are expected to be assayed in the near future. After accounting for overlaps between these resources, we were left with a set of 94 assayed genes.

### 2.2.2 Data pre-processing

We treated P, LP, and P/LP as a single pathogenic category; B, LB, and B/LB as a single benign category; VUS and conflicting interpretations of pathogenicity as the VUS category. Motivated by the clinical objectives that we define in Section 2.2.4, we only retained genes

that had at least one VUS and at least one pathogenic or benign variant in the ClinVar data set, reducing our data set to 3,981 genes. Considering the increased difficulty in mapping variant effects for longer proteins, we removed genes that were longer than genes previously assayed by MAVEs. We also removed genes that were shorter than those previously assayed because these genes may have had too few known variants to justify prioritization for MAVEs. Only genes present in both ClinVar and gnomAD were retained, and any variants recorded in ClinVar were removed from the gnomAD dataset to avoid double-counting during scoring. After these preprocessing steps, the final dataset consisted of 3,829 genes, including 321,619 ClinVar VUS, pathogenic, or benign missense variants and 1,161,072 gnomAD missense variants. This filtered collection of genes and variants served as the starting dataset for all downstream analyses.

### 2.2.3 Obtaining calibrated REVEL scores

REVEL is a meta-predictor that combines scores from multiple pathogenicity predictors and has been shown to perform well for clinical variant interpretation (Ioannidis et al., 2016). For each variant in all data sets, we extracted REVEL scores by mapping the chromosomal position and amino acid alteration to REVEL's prediction tables. However, REVEL scores themselves are not calibrated for clinical use and our formulations for clinical objectives require that prediction scores best approximate the posterior probability of being pathogenic or benign (Section 2.2.4). Therefore, we obtained a mapping of all possible REVEL scores to local posterior probability of being pathogenic or benign from Pejaver (2022). We then recorded these local posterior probabilities for all variants in this study and used them in all analyses.

### 2.2.4 Gene prioritization objectives: a clinical perspective

From a clinical perspective, the overall goal of gene prioritization is to make definitive and accurate classifications for more variants appearing in patient populations, when combining new functional evidence and existing predictive evidence. This includes: (1) assisting the movement of VUS to pathogenic and benign classes, (2) correcting for errors in current pathogenic and benign classifications and (3) improving predictors to assist clinical decision making. To achieve these goals, we rely on pathogenicity predictions from REVEL for variants in ClinVar and gnomAD over a subset of ClinVar genes. While ClinVar variants are the most relevant to the clinical goal, we include gnomAD variants to account for biases in ClinVar VUS annotations that arise out of the preferential testing of some genes over others. We refer to this combined set of ClinVar VUS and gnomAD variants as the *unlabeled set* of

variants.

Let $\mathcal{G}$ be a subset of ClinVar genes filtered based on constraints related to assay feasibility and other attributes of interest (Sections 2.2.1, 2.2.2). For a gene $g \in \mathcal{G}$, let $\mathcal{P}(g)$, $\mathcal{B}(g)$ be the set of variants in $g$ annotated as P/LP and B/LB in ClinVar, respectively. Let $\mathcal{U}(g)$ be the *unlabeled* set of variants, i.e., the combined set of ClinVar VUS and gnomAD variants for gene $g$. For a variant $v$, let $\rho(v)$ be a variant's probability of pathogenicity, estimated by explicitly calibrating a predictor's pathogenicity scores on a set of pathogenic and benign variants, i.e., $\rho(v) = p(v \text{ is pathogenic}|\ \text{REVEL}(v))$ (Section 2.2.3). We then define three measures, each serving different purposes in relation to our overall goal.

- **Movability.** We define movability as the 'movement' of a variant from a VUS annotation to a non-VUS (P, LP, B, LB) annotation when existing predictive and new functional evidence are combined as per the ACMG/AMP guidelines. This is similar to a previous definition (Kuang et al., 2020) but allows for the incorporation of prediction outputs and their updated evidential strength levels (Pejaver, 2022) more explicitly towards the reduction of VUS annotations.

  To have maximal impact on the reclassification of VUS, we aim to prioritize genes that contain the highest expected number of movable variants, i.e., the expected number of pathogenic/benign variants among a gene's unlabeled variants. Since annotating new pathogenic variants and new benign ones have different benefits, we propose two movability scores for each gene: the *movability-to-P score* and the *movability-to-B score*, and calculate them as follows:

$$\text{Move}_P(g) = \sum_{v \in \mathcal{U}(g)} \rho(v) \quad \text{and} \quad \text{Move}_B(g) = \sum_{v \in \mathcal{U}(g)} 1 - \rho(v)$$

  Optimizing this objective can also benefit the objective of improving predictors (see below), as it is expected to increase the number of P/LP and B/LB variants available for training.

- **Correction.** We define the 'correction' of a variant's clinical annotation as the update of an existing P/LP classification to B/LB/VUS or of an existing B/LB classification to P/LP/VUS, upon combining predictive and new functional evidence as per the ACMG/AMP guidelines when additional functional evidence is collected. To have maximal impact on pathogenic or benign variants that may be currently misclassified, we want to prioritize those genes that contain the highest expected number of variants whose clinical classification ought to be corrected, i.e., the expected number of pathogenic (benign) variants among a gene's variants annotated as benign

(pathogenic). Again, since there are differences in importance between correcting misclassifications of pathogenic variants and benign ones, we propose two correction scores for each gene: the *correction-of-P score* and the *correction-of-B score*, and calculate them as follows:

$$\text{Correct}_P(g) = \sum_{v \in \mathcal{P}(g)} 1 - \rho(v) \quad \text{and} \quad \text{Correct}_B(g) = \sum_{v \in \mathcal{B}(g)} \rho(v)$$

- **Predictor improvement.** Though not obvious, increasing the number of VUS with more certain predictions towards being benign or pathogenic has a significant role to play in moving more VUS to a non-VUS (P, LP, B, LB) annotation. If the improvement in the quality of a prediction toward B or P for a given VUS variant is large enough, it may directly provide an additional line of evidence that may be enough to push it to a non-VUS annotation. Furthermore, an improved prediction on variants from the same gene, might make the gene more likely to be assayed by an experimentalist motivated by the movability objective defined above. The new functional evidence thus obtained would help its movement to a non-VUS annotation.

  In order to increase the number of VUS with more certain predictions, the predictors themselves ought to be improved. To that end, we intend to generate more functional evidence for unlabeled variants (VUS and gnomAD variants) with uncertain predictions, and we prioritize genes with high average uncertainty (high entropy score) over their unlabeled variant set. The new functional evidence accrued on these variants would help improve the predictors, either by incorporating it as a feature while training a pathogenicity predictor or via transfer learning from function to disease domain. Note that the improvement in the predictor thus obtained is not restricted to the assayed variants, but also to other variants due to the predictor's generalization capabilities. Inspired by the entropy-based uncertainty sampling approach in the active learning literature (**?**), we prioritize genes for predictor improvement based on the average entropy of prediction on a gene's unlabeled variants. Intuitively, the criterion prioritizes genes having a higher fraction of unlabeled variants with calibrated pathogenicity score close to 0.5. Formally, we define the average entropy of a gene, adjusted for the number of unlabeled variants, as

$$\text{Entropy}_{\text{adj}}(g) = \sum_{v \in \mathcal{U}(g)} \frac{-\rho(v)\log_2 \rho(v) - (1-\rho(v))\log_2(1-\rho(v))}{|\mathcal{U}(g)|} \left(1 + \lambda \frac{\log_2 |\mathcal{U}(g)|}{\log_2 |\max_{h \in \mathcal{G}} \mathcal{U}(h)|}\right)$$

In this expression, the term $\left(1 + \lambda \frac{\log_2 |\mathcal{U}(g)|}{\log_2 |\max_{h \in \mathcal{G}} \mathcal{U}(h)|}\right)$, with $\lambda \in [0, 1]$, serves as an adjustment factor that prevents genes with very small number of unlabeled variants from being prioritized. The log scale gives genes with many unlabeled variants only a small advantage. The hyperparameter $\lambda$ can be further used to moderate the advantage given to genes with a large number of unlabeled variants. In this work, we choose $\lambda = 1$.

### 2.2.5 Gene prioritization strategies and their comparison

There are several possible strategies to prioritize genes for high-throughput functional assays. We describe a diverse set of prioritization strategies below.

1. **Knowledge- or expert-driven.** The set of 94 assayed genes described in Section 2.2.1 serve as an appropriate proxy for expert-driven gene prioritization. After applying the pre-processing steps described in Section 2.2.2, we were left with a set of 68 genes. This set is referred to as the *assayed* set. In addition, we simulated knowledge-driven selection in a simple manner by prioritizing genes in terms of the collective knowledge that we have about them. Here, we used publication counts as reported by PubMed ( `https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz`) in July 2022. We refer to this gene set as the *highest publications* set.

2. **Data-driven.** In this strategy, knowledgebases such as ClinVar are explicitly queried and genes are prioritized based on the numbers of variants of interest observed in them. For instance, genes with a high number of VUS are of particular interest because of the challenges in classifying such variants. We constructed a gene set ranked by the highest number of unlabeled variants (VUS and gnomAD). We refer to this gene set as the *highest unlabeled variants* set. Similarly, one may be interested in genes with the most number of VUS along with P/LP variants. We also constructed a gene set ranked by the highest total of VUS and P/LP. We refer to this gene set as the *highest non-benign variants* set.

   Previous work introduced two sophisticated strategies to prioritize genes for MAVEs in addition to the number of ClinVar VUS in a gene (Kuang et al., 2020). The movability- and reappearance-weighted impact score (MARWIS) incorporated patient data from Invitae to define variants' movability and reappearance and give extra weight for reappearing and movable VUS. The other score, difficulty-adjusted impact score (DAIS) was a specialized version of MARWIS that was adjusted for protein length. DAIS was deemed to be better-performing in practice and a set of 100 genes with the highest

DAIS was made available to the community. After applying the pre-processing steps in Section 2.2.2 to this set, 94 genes remained. We refer to this gene set as the *DAIS* set.

3. **Single score optimization.** We constructed five gene sets by directly optimizing five scores, derived to increase movability, correction and predictor improvement (Section 2.2.6). For each score, we picked the top-$K$ genes to create a gene set of length $K$. We refer to the resulting five gene sets as the *highest movability-to-P*, the *highest movability-to-B*, *highest correction-of-P*, *highest correction-of-B* and the *highest uncertainty* sets. As these gene sets represent the best selection for their corresponding score, no other gene set can be better with respect to that score.

4. **Multiple score optimization.** In order to obtain a single gene set that improves on all three objectives simultaneously, we implemented an approach to optimize a weighted combination of the five scores. The weights are learnt to incentivize improvement over the *assayed* gene set on all five scores (Section 2.2.6). The resultant gene set is referred to as the *combined score* set. This gene set makes tradeoffs between the five scores depending on how well the *assayed* gene set performs on each score.

5. **Random selection.** To create a baseline gene set of length $K$, we sampled $K$ genes randomly from the starting gene set and refer to this gene set as the *random* set.

We evaluated these different strategies by computing their score distributions in terms of $Move_P(g)$, $Move_B(g)$, $Correct_P(g)$, $Correct_B(g)$, and $Entropy_{adj}(g)$. Then, we tested whether the single score optimization strategy was significantly better than all other strategies using one-sided Wilcoxon rank-sum tests. We also tested whether the multiple score optimization strategy was better than those that were used to generate the *assayed* and *DAIS* gene sets. To ensure a fair comparison, we only compared gene sets of the same length. Since the *assayed* and *DAIS* are extant gene sets of fixed length, they determined the length constraints on the remaining gene sets. For comparisons with the *assayed* set, $K$ was set to 68, and for those with the *DAIS* set, $K$ was set to 94.

### 2.2.6 Multiple score optimization

Let $\mathcal{G}$ be a starting set of genes available to be assayed. Let $\mathcal{A} \subseteq \mathcal{G}$ (e.g., *assayed* set) be an existing gene set of length $K$, determined to be suitable for assaying based on some criteria. We present an approach to create a novel gene set optimized to improve over $\mathcal{A}$, w.r.t. the five scores, derived to increase movability, correction and predictor improvement

(Section 2.2.4). Let $\boldsymbol{w} = [w_i]_{i=1}^5$ be a weight vector with five non-negative entries such that $\sum_{i=1}^5 w_i = 1$. Let $S_1, S_2, S_3, S_4$ and $S_5$ be short-hands for $\text{Move}_P, \text{Move}_B, \text{Correct}_P, \text{Correct}_B$, and $\text{Entropy}_{\text{adj}}$, respectively. We define the combined weighted score as

$$\text{Combined}_{\boldsymbol{w}}(g) = \sum_{i=1}^5 w_i \overline{S}_i(g)$$

where $\overline{S}(g)$ denotes a score $S(g)$ after z-score normalization on the entire gene set $\mathcal{G}$. The normalization ensures that the scores are on the same scale, which in turn allows us to define an optimization criteria that treats each score equally. It also allows the weights to be on the same scale, which makes it easier to find a good solution. In order to learn the optimal $\boldsymbol{w}$, we first create a sample, $W$, containing $10^5$ candidate weights from Dirichlet$(1, 1, 1, 1, 1)$, a uniform distribution over the space of five dimensional probability vectors. For each candidate $\boldsymbol{w} \in W$, we sort the genes in $\mathcal{G}$ in the decreasing order of $\text{Combined}_{\boldsymbol{w}}(g)$. The top $K$ genes are picked in a candidate gene set $\mathcal{O}_{\boldsymbol{w}}^K$. For a set of numbers $X$, let $\text{Median}(X)$ and $\text{Prctile}_{90}(X)$ denote the median and the $90^{\text{th}}$ percentile of those numbers. For $G \subseteq \mathcal{G}$, let $\overline{S}_i(G)$ denote the set containing the $i^{th}$ normalized score evaluated on genes in $G$. If the median or the $90^{\text{th}}$ percentile of any normalized score on $\mathcal{O}_{\boldsymbol{w}}^K$ is less than that on $\mathcal{A}$, then discard $\boldsymbol{w}$, i.e., for any $i$, if $\text{Median}\big(\overline{S}_i(\mathcal{O}_{\boldsymbol{w}}^K)\big) < \text{Median}\big(\overline{S}_i(\mathcal{A})\big)$ or $\text{Prctile}_{90}\big(\overline{S}_i(\mathcal{O}_{\boldsymbol{w}}^K)\big) < \text{Prctile}_{90}\big(\overline{S}_i(\mathcal{A})\big)$, then discard $\boldsymbol{w}$. This ensures that each remaining weight leads to a gene set with higher median and $90^{\text{th}}$ percentile on each of the five score distributions compared to the $\mathcal{A}$. Let $W_{\text{good}}$ be the set of remaining candidate weights. If $W_{\text{good}} \neq \emptyset$, a $\boldsymbol{w} \in W_{\text{good}}$ is guaranteed to give a better gene set than $\mathcal{A}$ on each of the five scores. In order to select an optimum weight from $W_{\text{good}}$, we define the following optimization criteria to find weights that lead to largest cumulative increase in the the normalized score medians compared to $\mathcal{A}$:

$$C(\boldsymbol{w}) = \sum_{i=1}^5 \big[\text{Median}\big(\overline{S}_i(\mathcal{O}_{\boldsymbol{w}}^K)\big) - \text{Median}\big(\overline{S}_i(\mathcal{A})\big)\big].$$

The optimum weights are given by $\boldsymbol{w}_{\text{opt}} = \text{argmax}_{\boldsymbol{w} \in W_{\text{good}}} C(\boldsymbol{w})$. The corresponding gene set, $\mathcal{O}_{\boldsymbol{w}_{\text{opt}}}^K$ is the optimal gene set, referred to as the *combined score* set. Note that if a gene set of a different size, $K_1 \neq K$, is needed, the top $K_1$ genes sorted based on $\text{Combined}_{\boldsymbol{w}_{\text{opt}}}(g)$ are selected. The resultant set is referred to as $\mathcal{O}_{\boldsymbol{w}_{\text{opt}}}^{K_1}$.

### 2.2.7 Functional and phenotypic enrichment analyses

To evaluate the biological and clinical relevance of the multiple score optimization strategy, we ranked all genes by their *combined score* and conducted a functional enrichment analysis on the top 100 genes using the *g:GOSt* function in the gProfiler web-server (Raudvere et al.,

2019). We used our starting gene set of 3,829 genes as the background set. Any Gene Ontology (GO) and Human Phenotype (HP) Ontology terms that were significantly enriched in the top 100 genes, after correcting for multiple hypothesis testing ($P$-value $< 0.05$) were recorded.

## 2.3 Results

### 2.3.1 Multiple score optimization outperforms knowledge-driven and simple data-driven strategies

We compared multiple gene sets (see Section 2.2.5), constructed through diverse prioritization strategies on the five scores, covering the three clinical objectives: movability, correction and

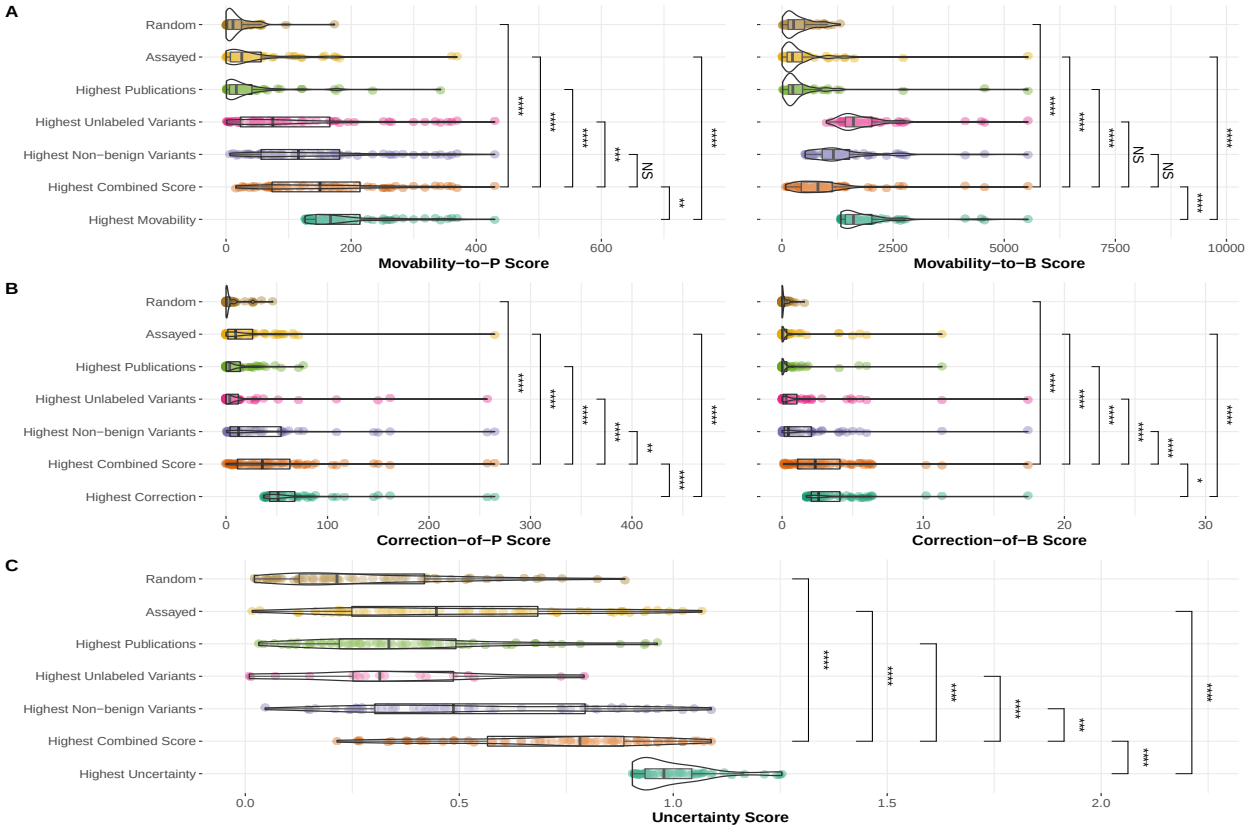



Figure 2.1: **Score distributions 68-gene sets constructed based on seven prioritization strategies. A.** Score distribution of movability to pathogenic (left) and benign (right), **B.** Score distribution of correction of pathogenic (left) and benign (right) variants, **C.** Uncertainty score distribution.

predictor improvement (Figure 2.1). All the sets in this comparison had 68 genes, to be

consistent with the *assayed* set. Unsurprisingly, for any given score, the *highest single score* gene set, being the best set for the score, outperformed all other gene sets. As expected, the *combined score* set performed better than the *assayed* gene set because it was explicitly constructed to improve over the *assayed* set. Overall, the *combined score* set performed better than all other gene sets except the respective highest single score sets. There were two exceptions to this. In the case of *movability-to-B* score, the *combined score* set did not perform better than the *highest unlabeled variants* and *highest non-benign variants* gene sets, suggesting that the number of unlabeled variants may be a strong determinant of *movability-to-B* due to the high prior probability of bening benign in general. In particular, the scope of improvement in *movability-to-B* score over the *highest unlabeled variants* set is limited as can be observed in comparison to *highest movability-to-B* set, the best possible set for that score. Furthermore, among all comparisons of the *combined score* where it performs better, it does so with statistical significance, except in one case: comparison with *highest non-benign variants* set on *movability-to-P* score.

The *assayed* set performed slightly better than *random* on most scores. Moreover, its score distributions were far away from that of the corresponding *highest single score* set. This suggests that there is a huge scope of improvement on the set of genes currently being assayed, with respect to clinical objectives. On all score criteria, the performance of the *highest publication* set is quite similar to that of the *assayed* set. This is consistent with the previous observation that genes with fewer publications are less likely to be functionally tested (Stoeger et al., 2018a).

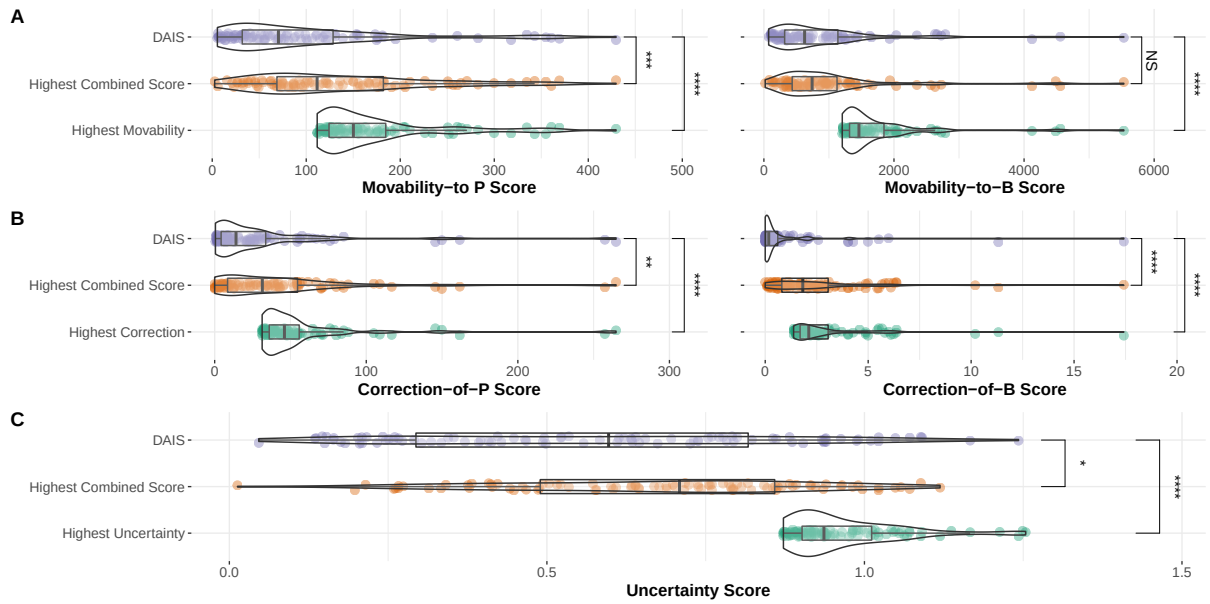



Figure 2.2: **Score distributions for top 94 genes prioritized by our proposed strategies and by existing data-driven strategies. A.** Score distribution of movability to pathogenic (left) and benign (right), **B.** Score distribution of correction of pathogenic (left) and benign (right) variants, **C.** Uncertainty score distribution. DAIS, 94 genes out of the top 100 genes ranked by the difficulty-adjusted impact score (Kuang et al., 2020).

### 2.3.2 Multiple score optimization outperforms existing clinically motivated prioritization strategies

We next compared our single and multiple score optimization strategies to a previously proposed strategy that explicitly aimed to improve clinical variant classification, DAIS (Kuang et al., 2020) (Figure 2.2). Since the DAIS set comprised of 94 genes, we considered the top 94 genes with the highest single and combined scores. The single and multiple score optimization strategies yielded statistically significant improvements over DAIS in all situations, with one exception. When considering the *movability-to-B* score, the *combined score* set showed improvement over DAIS, although not significantly, similar to our observations in Section 2.3.1.

### 2.3.3 Multiple score optimization yields clinically relevant genes

We characterized the properties of the highest-scoring genes in the *combined score* set and investigated to what extent our strategy aligned with biomedical interests. Among the top 20 genes, six genes were in our *assayed* gene set, and 12 genes were also prioritized by DAIS, albeit with differences in ranking (Table 2.1). All identified genes generally have a large

number of variants recorded in ClinVar and gnomAD, with the exception of *SCN10A*, which has no

Table 2.1: Missense variant counts and prioritization scores for the top 20 genes ranked by the *Combined score*. Similar results for all genes are available at: `https://igvfgenecard.shinyapps.io/GeneCardApp/`. Genes in bold were also present in the *assayed* gene set. Movability and Correction scores are rounded to the nearest integer. The Combined score is computed as a weighted sum of the five z-normalized component scores using weights 0.143 (*movability-to-P*), 0.160 (*movability-to-B*), 0.380 (*correction-of-P*), 0.310 (*correction-of-B*), and 0.006 (*uncertainty*).

| Rank | Gene | DAIS Rank | ClinVar | | | gnomAD | Total | Movability | | Correction | | Adj. Entropy | Combined |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | P/LP | B/LB | VUS | | | to P | to B | of P | of B | | |
| 1 | *TSC2* | 32 | 80 | 185 | 2178 | 273 | 2716 | 318 | 2035 | 29 | 17 | 0.8 | 13.3 |
| 2 | ***BRCA1*** | 10 | 120 | 206 | 2817 | 160 | 3303 | 181 | 2727 | 71 | 11 | 0.5 | 10.5 |
| 3 | ***LDLR*** | 40 | 635 | 62 | 564 | 176 | 1437 | 155 | 547 | 265 | 4 | 0.9 | 10.1 |
| 4 | *FBN1* | 39 | 873 | 17 | 1338 | 536 | 2764 | 335 | 1451 | 257 | 2 | 0.9 | 9.9 |
| 5 | ***BRCA2*** | 9 | 57 | 236 | 5453 | 325 | 6071 | 173 | 5533 | 37 | 6 | 0.3 | 7.5 |
| 6 | *IDS* | 1055 | 120 | 57 | 49 | 125 | 351 | 32 | 134 | 39 | 10 | 0.7 | 7.0 |
| 7 | *MYH7* | 2 | 271 | 17 | 1284 | 297 | 1869 | 355 | 1129 | 150 | 2 | 1.1 | 6.7 |
| 8 | *SCN1A* | 66 | 452 | 39 | 670 | 361 | 1522 | 283 | 683 | 146 | 3 | 1.0 | 6.6 |
| 9 | *NF1* | 11 | 232 | 19 | 2750 | 224 | 3225 | 261 | 2632 | 162 | 0 | 0.5 | 6.4 |
| 10 | ***MSH2*** | 4 | 73 | 26 | 1757 | 123 | 1979 | 369 | 1409 | 28 | 6 | 1.0 | 5.9 |
| 11 | *COL4A5* | 1839 | 414 | 87 | 66 | 372 | 939 | 72 | 347 | 80 | 6 | 0.7 | 5.6 |
| 12 | *SCN8A* | 468 | 122 | 44 | 346 | 250 | 762 | 125 | 438 | 43 | 6 | 0.9 | 5.3 |
| 13 | ***SCN5A*** | 63 | 83 | 33 | 1058 | 386 | 1560 | 361 | 998 | 23 | 5 | 1.0 | 5.3 |
| 14 | *MLH1* | 8 | 122 | 33 | 1103 | 80 | 1338 | 175 | 957 | 62 | 4 | 0.8 | 5.0 |
| 15 | *SCN10A* | 391 | 0 | 55 | 381 | 831 | 1267 | 226 | 930 | 0 | 6 | 0.8 | 4.8 |
| 16 | *FLNA* | 211 | 32 | 85 | 560 | 493 | 1170 | 150 | 858 | 16 | 6 | 0.8 | 4.7 |
| 17 | *CACNA1S* | 323 | 12 | 44 | 393 | 777 | 1226 | 251 | 858 | 3 | 6 | 0.9 | 4.7 |
| 18 | *FBN2* | 155 | 33 | 58 | 708 | 1005 | 1804 | 300 | 1331 | 13 | 5 | 0.9 | 4.6 |
| 19 | ***TP53*** | 1 | 143 | 76 | 717 | 27 | 963 | 176 | 525 | 54 | 4 | 1.0 | 4.5 |
| 20 | *ABCA4* | 130 | 235 | 17 | 582 | 845 | 1679 | 252 | 1110 | 109 | 0 | 0.8 | 4.3 |

variants classified as pathogenic or likely pathogenic. In addition, our *combined score* also prioritized important genes that may have been overlooked previously. For example, IDS, which has more than 200 IDS variants were found in Hunter syndrome patients (Ricci et al., 2003) was ranked $6^{th}$. *COL4A5*, with over 400 variants thata cause Alport syndrome, was (ranked $11^{th}$). Many sodium voltage-gated channels (*SCN*)-related genes were also ranked within the top 20, and mutations in these genes can lead to channel defects and cause channelopathies (de Lera Ruiz and Kraus, 2015).

Since the objective of improving predictors may not necessarily yield genes that are clinically relevant, we systematically explored the functional and phenotypic characteristics associated with the *combined score* set. We conducted an enrichment analysis on the top 100 genes ranked by their combined score and reported significantly enriched GO terms and the 40 most significant HP terms (Supp. Figure 1A). This top-100 gene set was enriched in many biological processes such as neuronal action, membrane depolarization, and molecular functions such as multiple channel activities and transmembrane transporter activity. From

the phenotypic perspective, enriched high-level HP terms included abnormalities of different organ systems such as skin, gastrointestinal tract, nervous system, among others (Supp. Figure 1B). More specific HP terms included cardiovascular related disease, limitation of mobility, and stroke, among others.

## 2.4 Discussion

Genetic and genomic testing are now routinely used in healthcare systems to provide diagnoses and infer lifetime risk for disease symptoms, particularly in the identification of hereditary susceptibility to cancer, metabolic conditions, intellectual and physical developmental disorders, among others. The classification of genetic variants detected in a patient's gene panel or genome is a key step in this context. Recent updates to the ACMG/AMP guidelines have made it possible in the future for a variant to obtain a *likely pathogenic/benign* classification based simply on one piece of functional evidence and one strong score from a computational predictor (Pejaver, 2022). In this regard, our study presented three objectives that explicitly captured the goal of improving clinical classification of variants and derived five scores to operationalize them. We derived an optimal gene set for each score and also derived a *combined score* gene set by optimizing a weighted combination of the five scores to explicitly improve over the existing *assayed* set.

All single score optimization strategies proved to be significantly better on their corresponding score, when compared to existing strategies. As expected, all single score optimization strategies, led to the best performance on the corresponding score. More importantly, evaluating the existing approaches relative to the single score optimization, demonstrated a considerable performance gap, suggesting a significant scope of improvement on each objective. Even though our *combined score* gene set was obtained by optimizing directly over the three objectives relative to the *assayed* set, its observed improvement over the *assayed* and DAIS gene sets on all scores is not entirely obvious due to the inherent trade-offs between the objectives (movability vs. predictor improvement). This is a further testament to the scope of simultaneous improvement on all objectives along with an approach that demonstrably does so.

*combined score* gene set performed reasonably well on all scores and showed an improvement over all existing approaches. While this was expected, it is not guaranteed as some objectives can be in conflict with each other. However, the *combined score* was not the most optimal for each score, suggesting a tradeoff between the different objectives.

DAIS, a more sophisticated strategy, presented higher scores in general but did not outperform our approach. Unlike DAIS, our approach does not use any proprietary patient data,

but despite this, one-third of our genes overlapped with the DAIS set. Our approach can be potentially complementary to DAIS, since we accounted for conflicting variants, incorporated non-VUS and less biased gnomAD variants and focused on correction and predictor improvement as objectives. Another strength of our strategy is its interpretability. The *movability* scores and *correction* scores are interpreted as the expected number of pathogenic or benign variants, and the *uncertainty* score as predictive uncertainty. In addition, our approach for multiple score optimization could be easily extended to incorporate other scores such as DAIS, if appropriate data were available, or could directly optimize the combined score to improve over both the assayed and DAIS sets.

There are several caveats to our proposed strategies. First, the ACMG/AMP guidelines assume that functional and predictive evidence are independent of each other but this may not be the case, especially for newer methods.(Pejaver, 2022) Second, it is unknown to what extent current classifications in ClinVar conform to the ACMG/AMP guidelines and what lines of evidence were used to make the classification in the first place. Third, without functional assay outcomes, we can only extrapolate the expected number of variants that will *move* or be *corrected* based on predictive evidence, but cannot directly calculate these when both predictive and functional evidence are available. Finally, our *combined score* was optimized using the *assayed* set's score distribution, thus providing it with an unfair advantage in our analyses. To evaluate this, an alternative approach to combine scores first ranked each gene by each of its five scores and then took the arithmetic and geometric means was adopted. This avoids the use of any of gene sets derived based on existing strategies for optimization. We observed that our approach generally did better than mean rank-based approaches but not significantly so.

Though our movability objective quantifies the expected number unlabeled variants in a gene that are pathogenic (or benign), it is possible that after running a given assay the number of variants moved to the P/LP (B/LB) categories as per the ACMG/AMP guidelines might differ. This might happen either because the assay might not capture the functional mechanism that leads to the disease, or the strength of the new evidence combined with existing evidence might not be enough to move the variant. Without functional assay outcomes, this is difficult to discern and is a limitation of our study. In future, when additional information on an assay's relevance to specific diseases is available, refined criteria that take that information into account might better quantify the movement. Similarly, if all existing evidence for a variant is accessible, the criteria may be refined to take it into account, as done so by Kuang et al.(Kuang et al., 2020) Our study is currently limited in this regard, as ClinVar does not detail which specific lines of evidence were used to classify a variant. Similar considerations apply to the correction scores as well.

In conclusion, we defined three objectives in terms of improving clinical classification by using variant pathogenicity predictors. Our final *combined* scores provided a list of prioritized genes for MAVEs but this list will keep updating with iterated future work between prediction and experimentation. All data sets, analysis scripts, and supplementary results for this study can be accessed here: `https://github.com/strongbeamsprout/Gene-Prioritization`.

# Chapter 3

# Estimating Priors of Pathogenicity Using Positive-Unlabeled Learning

## 3.1 Introduction

Accurately estimating the prior probability of pathogenicity is a foundational step in Bayesian variant interpretation and calibration. In the ACMG/AMP Bayesian framework, the prior represents the baseline probability that a variant is pathogenic before considering any evidence. Every subsequent piece of evidence—computational predictions, population frequency, segregation, functional assays—updates this prior to a posterior probability via Bayes' theorem. Consequently, if the prior is mis-specified, all downstream posterior probabilities and evidence strengths are systematically biased.

Estimating this prior is challenging because variant datasets are incompletely labeled: while ClinVar provides positively labeled pathogenic/likely pathogenic (P/LP) variants, the vast majority of possible variants remain unlabeled and cannot be assumed benign. Thus, prior estimation is intrinsically a *positive–unlabeled* (PU) learning problem.

PU learning explicitly models settings in which only a subset of positives is labeled and the remaining data consist of a mixture of positives and negatives (Elkan and Noto, 2008; Jain et al., 2016; Scott, 2015). In the genomics context, ClinVar supplies labeled positives, while gnomAD and other population datasets provide the unlabeled pool. The central goal is to estimate the *class prior* $\alpha$, the proportion of pathogenic variants within the unlabeled set. Estimating $\alpha$ enables converting nontraditional PU classifiers (trained to distinguish P vs. U) into probability-calibrated classifiers on a traditional P vs. N scale (Pejaver et al., 2022).

Multiple class-prior estimation approaches have been proposed, including parametric

likelihood-based estimators (AlphaMax), kernel-based density estimators, and univariate transformations. However, these methods assume either separability in low-dimensional projections or specific parametric distributions—assumptions that often fail for the high-dimensional, correlated feature spaces generated by modern variant effect predictors.

To address this, Zeiberg (2020) introduced *DistCurve*, a nonparametric estimator that identifies the prior value that best reconciles classifier performance with empirical behavior. DistCurve varies the assumed prior, measures classifier stability across bootstrap resampling, and selects the prior that minimizes deviation. Pejaver et al. (2022) applied DistCurve genome-wide using MutPred2-derived features for all missense variants and estimated a global prior of 4.41%.

While informative, a single genome-wide prior assumes that all genes contribute equally to pathogenic variation, contradicting empirical patterns of selective constraint and disease burden. Bhat et al. (2025) demonstrated that priors vary by orders of magnitude across genes, and that mis-specified priors cause systematic miscalibration of continuous computational evidence (PP3/BP4) within the ACMG/AMP Bayesian system. Genes such as *LDLR*, *TP53*, and *KCNH2* are enriched for pathogenic variation, whereas some genes rarely harbor pathogenic missense changes.

This motivates gene-specific prior estimation. By applying DistCurve to each gene independently—using gene-restricted unlabeled variants and positively labeled ClinVar P/LP variants—we obtain priors that reflect biological constraint and the expected prevalence of pathogenicity. These gene-level priors integrate seamlessly with ClinGen's local likelihood ratio framework, producing calibrated posterior probabilities that map directly to ACMG/AMP evidence strength.

In this chapter, we review the theoretical foundations of PU learning and the DistCurve method. Then we introduce our adaptation of DistCurve to single-gene contexts. We describe methodological challenges, experimental design, and innovations for robustness. Last, we present results and robustness analyses and conclude with implications, limitations, and integration into downstream calibration frameworks.

## 3.2 Background

### 3.2.1 Bayesian Framework for Variant Pathogenicity Classification

The ACMG/AMP variant classification guidelines have been formalized into a Bayesian decision framework for combining multiple lines of evidence to assess pathogenicity (Tavtigian

et al., 2018). In this framework, the prior probability of pathogenicity represents the baseline belief that a given variant is disease-causing before considering any new evidence (e.g. gene-specific and variant-specific context). As evidence is incorporated—such as functional assay results, population allele frequencies, or in silico predictions—this prior belief is updated via Bayes' theorem. Formally, for a variant with prior probability $P(\text{pathogenic})$, and observed evidence $E$, the posterior probability is:

$$P(\text{pathogenic} \mid E) \;=\; \frac{P(E \mid \text{pathogenic}) \times P(\text{pathogenic})}{P(E)}\,,$$

where $P(E \mid \text{pathogenic})$ is the likelihood of observing the evidence if the variant is truly pathogenic. This Bayesian formulation provides a mathematical foundation for what were previously qualitative heuristic rules, ensuring that evidence integration is systematic and quantitative (Tavtigian et al., 2018; Tavtigian, 2020). The resulting posterior probability $P_{\text{post}} = P(\text{pathogenic} \mid E)$ reflects the updated likelihood that the variant is pathogenic after accounting for the available evidence.

### 3.2.2 Odds of Pathogenicity and Evidence Strength

In practice, Bayesian updating is often performed using odds and likelihood ratios rather than probabilities directly. The odds form of Bayes' theorem is:

$$\text{Posterior Odds} \;=\; \text{Prior Odds} \times \text{Odds}_{\text{Path}}\,,$$

where $\text{Odds}_{\text{Path}}$ is the odds of pathogenicity given the evidence. The posterior probability can be recovered from the posterior odds as:

$$P_{\text{post}} \;=\; \frac{\text{Posterior Odds}}{1 + \text{Posterior Odds}}\,.$$

Each piece of evidence contributes multiplicatively to the odds of pathogenicity. The ACMG/AMP guideline criteria have qualitative strength levels (Supporting, Moderate, Strong, Very Strong for pathogenic evidence) which can be mapped to quantitative likelihood ratios (LR) in this Bayesian framework. Notably, the evidence strengths are scaled exponentially: for example, two Supporting pieces of evidence are equivalent to one Moderate, two Moderate to one Strong, and two Strong to one Very Strong evidence in terms of weight (Richards et al., 2015). In other words, if a single "Very Strong" pathogenic evidence corresponds to an LR of $c$, then a single Strong evidence contributes roughly $\sqrt{c}$ (since two Strong $\approx$ one Very Strong), a Moderate contributes $\sqrt[4]{c}$ (half the weight of Strong), and a Supporting contributes

$\sqrt[8]{c}$. This yields the relationship (Tavtigian et al., 2018):

$$LR^{+}_{VS} = c,\ LR^{+}_{S} = 2\sqrt{c},\ LR^{+}_{M} = 4\sqrt{c},\ LR^{+}_{Su} = 8\sqrt{c},$$

where $LR^{+}_{VS}$ denotes the positive likelihood ratio for each level of evidence strength and $c$ is a constant defined by the "very strong" evidence threshold. which ensures that evidence strengths are consistent and additive on a log-odds scale. Using such calibrated likelihood ratios, the Bayesian model can continuously calculate posterior probabilities and map them to the five ACMG/AMP classification categories (Benign, Likely Benign, VUS, Likely Pathogenic, Pathogenic). For classification, threshold posterior probabilities are typically applied. For example:

- Pathogenic (P): $P_{\text{post}} > 0.99$
- Likely Pathogenic (LP): $0.90 < P_{\text{post}} \leq 0.99$
- Variant of Uncertain Significance (VUS): $0.10 < P_{\text{post}} \leq 0.90$ (falls short of LP or LB thresholds)
- Likely Benign (LB): $0.001 < P_{\text{post}} \leq 0.10$
- Benign (B): $P_{\text{post}} < 0.001$

This quantitative scheme provides a transparent and reproducible framework for variant classification, turning the qualitative ACMG/AMP criteria into mathematically defined scores.

### 3.2.3 Linking Continuous Predictor Scores to Evidence Strength

Many variant effect predictor tools (e.g. MutPred2, AlphaMissense, BayesDel) output continuous scores indicating how likely a variant is to be deleterious, rather than categorical "pathogenic/benign" outputs. Integrating such scores into the ACMG/AMP framework requires converting them into an equivalent strength of evidence (e.g. deciding if a given score warrants Supporting evidence, or Moderate, etc. under criteria like PP3/BP4). To achieve this, ClinGen's Sequence Variant Interpretation group introduced the concept of a local positive likelihood ratio, $LR^{+}_{local}(s)$, which is essentially a score-specific likelihood ratio (Pejaver et al., 2022). It is defined as the ratio of pathogenic vs. benign probability densities at that score $s$:

$$LR^{+}_{local}(s) = \frac{f_{\text{pathogenic}}(s)}{f_{\text{benign}}(s)},$$

where $f_{\text{pathogenic}}(s)$ and $f_{\text{benign}}(s)$ are the score distributions for known pathogenic and benign variant sets, respectively. This local LR serves as a calibration curve for the predictor: given a score $s$, one can compute the posterior probability that a variant is pathogenic at that score by applying Bayes' theorem with the local LR in place of a global LR:

$$P_{\text{post}}(s) = \frac{LR^{+}_{local}(s) \times P_{\text{prior}}}{LR^{+}_{local}(s) \times P_{\text{prior}} + (1 - P_{\text{prior}})}.$$

This equation directly converts a model's raw score into a posterior probability of pathogenicity, given an assumed prior $P_{\text{prior}}$. By defining score thresholds corresponding to posterior probability cutoffs (e.g. $P_{\text{post}} = 0.90$ for Likely Pathogenic or $P_{\text{post}} = 0.99$ for Pathogenic), one can determine what score value would constitute supporting vs. moderate vs. strong evidence, etc., for that tool. Crucially, accurate estimation of the prior probability is essential in this mapping – the same score will translate to a different evidence strength depending on how common pathogenic variants are expected to be in the gene or context of interest. In other words, knowing the baseline prior (how likely a random variant is pathogenic) allows continuous scores to be calibrated into the ACMG/AMP evidence categories (e.g., deciding if a high deleteriousness score should count as PP3-Moderate vs. just PP3-Supporting). If the prior is misestimated, the thresholds for these categories will be off, potentially misclassifying variants. This underlines the need for robust methods to determine the prior probability of pathogenicity in the relevant gene or variant space.

## 3.3 Positive-Unlabeled Learning for Prior Estimation

Estimating the prior probability of pathogenicity for a given gene or dataset is challenging because while we have many known pathogenic variants (labeled positive cases, e.g. ClinVar Pathogenic/Likely Pathogenic), we lack comprehensive labeled negatives (benign variants). The remaining variants (from databases like gnomAD or uncurated sequencing data) form an unlabeled set that is a mixture of benign and as-yet-undiscovered pathogenic variants. This situation can be approached with Positive–Unlabeled (PU) learning, a machine learning framework designed for scenarios where only positives are labeled and all other data are unlabeled (implicitly containing both negatives and some positives). The key goal in PU learning is to estimate the class prior $\alpha$, which is the fraction of true positives in the unlabeled set, and to train a classifier that can predict which unlabeled instances are likely positive. In our context, $\alpha$ corresponds to the prior probability that a random variant (in a given gene or dataset) is pathogenic. Once $\alpha$ is known, it can be used as the $P_{\text{prior}}$ in the Bayesian calculations above. PU learning strategies have been successfully applied in various

biological domains where negative examples are difficult to label. For instance, it has been used in disease gene discovery (identifying causal genes when only known disease genes are "positive" and all others are unlabeled) (Yang et al., 2012), post-translational modification site prediction (Li et al., 2019), and to identify functional variants in deep mutational scanning experiments (Song et al., 2021). These applications demonstrate that PU models can effectively leverage incomplete labels to estimate underlying probabilities.

Estimating the class prior $\alpha$ from positive–unlabeled (PU) data has been the focus of several methodological developments. Prominent approaches include:

- **AlphaMax** (Jain et al., 2016): a likelihood-based estimator that models the unlabeled score distribution as a mixture of positive and negative components. The optimal $\alpha^*$ is chosen where the profile likelihood curve exhibits diminishing improvement (the "elbow point"). A key insight from this work is that non-traditional classifiers—trained on P vs. U data—can be used as *univariate transformations* that preserve the class prior, thereby avoiding direct density estimation in the original high-dimensional feature space.

- **RKHS-based density–ratio estimation** (Ramaswamy et al., 2016): a flexible, non-parametric approach that embeds distributions in a reproducing kernel Hilbert space and estimates the ratio $f_P(x)/f_U(x)$ to infer $\alpha$. While powerful, RKHS methods scale poorly in practice because kernel computation requires constructing and manipulating large Gram matrices, resulting in high time and memory complexity for large variant datasets.

- **Univariate transformation method** (Menon et al., 2015): techniques that project high-dimensional feature vectors onto a single informative axis (e.g., classifier scores) and perform prior estimation in that transformed 1D space. By reducing PU learning to a one-dimensional mixture model, these approaches bypass many challenges of density estimation in high dimensions.

While each method is theoretically sound, they share practical limitations when applied to molecular variant data. Variant effect prediction methods such as MutPred2 generate 1,345-dimensional feature vectors that are *highly correlated, non-independent, and biologically structured*. In such settings, likelihood-based methods (e.g., AlphaMax) can become sensitive to model mis-specification and numerical instability; kernel methods become computationally prohibitive, and univariate transformation approaches that operate exclusively on a one-dimensional score distribution may misrepresent the mixture boundary if the score does not sufficiently separate positive-like unlabeled variants from the remaining background.

By contrast, DistCurve (Zeiberg et al., 2020) leverages a flexible high-dimensional PU classifier to construct a one-dimensional score, but estimates the prior via a nonparametric distance functional over resampled positives and pseudo-positives, rather than by directly fitting a parametric mixture in score space. This makes it more robust to the complex, correlated structure of variant features, while still retaining the computational advantages of working with a scalar score.

Pejaver et al. (2022) applied DistCurve on a genome-wide scale to estimate a global prior for missense variant pathogenicity. They trained an ensemble classifier on ClinVar pathogenic vs. gnomAD variants and then used DistCurve to find the prior probability that a rare missense variant in a typical Mendelian disease gene is pathogenic. This analysis yielded an estimated prior of 4.41%. This data-driven prior is notably lower than the 10% prior that had been previously assumed based on expert opinion and ClinVar data (Tavtigian et al., 2018). In other words, earlier frameworks often used a universal prior of 0.10 for lack of better data, whereas the PU learning approach suggests the true overall prior (across many genes) is closer to 0.044. This more refined prior has direct implications for variant classification thresholds: using 4.4% vs. 10% as the baseline will shift the odds required to reach Pathogenic or Likely Pathogenic criteria, underscoring the importance of data-driven prior estimation.

Notably, the DistCurve approach is flexible and can be applied not just genome-wide but also to any subset of variants, provided one can obtain a representative feature distribution. One limitation of using a single, genome-wide prior is that it obscures the substantial gene-level heterogeneity in pathogenic variant rates. Some genes (especially those implicated in severe, early-onset disorders or tumor suppressors like BRCA1, TP53) harbor a much higher proportion of pathogenic variants among all rare missense variants, whereas others have very few. In fact, gene and syndrome specificity can greatly influence prior probabilities: for example, in a patient with long QT syndrome, a variant in the major LQTS gene KCNQ1 has a much higher a priori probability of being pathogenic than a variant in a minor gene such as KCNJ5 (Ruklisa et al., 2015). Ruklisa et al. (2015) also demonstrated that gene-specific models for cardiac conditions outperformed a genome-wide model, due to each gene having its own baseline risk profile. This insight suggests that applying DistCurve (or any prior estimation method) at the gene level would yield priors that are more biologically meaningful for variant interpretation in that gene. Indeed, one can run a DistCurve analysis separately for each gene (using that gene's ClinVar variants and that gene's gnomAD variants) to estimate a gene-specific prior – effectively, the fraction of rare variants in that gene that are pathogenic. Early evidence indicates that such gene-level priors can vary dramatically and would provide a gene-aware foundation for Bayesian calibration of variant pathogenicity.

## 3.4 Gene-Specific Prior Probabilities and Current Challenges

Despite the clear need for gene-specific priors, current clinical practice largely uses a one-size-fits-all prior for variant classification due to the difficulty of obtaining precise estimates for each gene. For example, Tavtigian et al. (2018) in their Bayesian formulation of ACMG guidelines suggested using a universal prior of 10% (i.e. assuming 1 in 10 rare variants is pathogenic across the board), based on aggregated ClinVar data and expert knowledge. More recent work by Pejaver et al. (2022); Bergquist et al. (2024), using the DistCurve PU learning approach described above, refined this estimate to about 4.41% as an overall prior for pathogenic missense variants across Mendelian disease genes. While these estimates are valuable at a broad scale, the reliance on a single prior for all genes is counterintuitive and potentially inaccurate, as it ignores the vast biological and clinical heterogeneity across genes. In reality, genes differ in their tolerance to variation and the proportion of variants that cause disease: some genes are highly constrained and any rare missense might be likely disease-causing, whereas others are more tolerant.

There have been a few efforts to derive gene-specific or context-specific priors, but they have been limited in scope. For instance, functional assay studies by Clark et al. (2019) and Jain et al. (2025) experimentally measured the impact of numerous variants in specific genes (such as NAGLU and ARSA) to estimate what fraction of variants truly disrupt function. These experiments allow an empirical prior to be estimated for those genes (e.g. a large fraction of random variants in an essential domain of NAGLU might severely reduce enzyme activity, suggesting a high prior probability of pathogenicity for variants in that domain). However, experimental approaches are resource-intensive and not scalable to the thousands of disease-relevant genes.

Another line of research leverages clinical and population genetic data: for example, Ruklisa et al. (2015) developed a Bayesian model for inherited cardiac conditions that included gene-specific prior odds, effectively calculating the probability a variant is pathogenic given it is in a particular gene and disease context. Their framework showed that incorporating gene-level priors improved prediction accuracy, affirming that gene-specific baseline risk matters. Similarly, Bhat et al. (2025) combined large-scale clinical records (e.g. UK Biobank) with population sequencing (gnomAD) to estimate priors for each gene within certain phenotypes.

Another noteworthy line of work comes from Whiffin et al. (2018), which incorporates region-specific priors within genes using the concept of an *etiological fraction* (EF). EF represents the proportion of variant carriers in a disease cohort for whom the variant is truly disease-causing, i.e., the probability that a variant is pathogenic given it is observed in an

affected individual. Case–control burden analyses in hypertrophic cardiomyopathy (HCM) genes have shown that some missense-enriched regions achieve EF values $\geq 0.95$, implying that a novel rare variant in those regions has a very high prior probability of pathogenicity, whereas other regions in the same genes have much lower EF. This demonstrates that priors can vary substantially even within a single gene, depending on functional domains and mutational hotspots. Similar domain-level heterogeneity has been reported for breast cancer risk genes: Breast Cancer Association Consortium et al. (2021) found that missense variants in the RING and BRCT domains of *BRCA1* and in key domains of *ATM* confer substantially higher breast cancer risk, whereas risk did not differ markedly between domain and non-domain locations for missense variants in *CHEK2* and *BRCA2*. Together, these examples underscore that overly coarse priors (e.g., a single 4.41% prior applied uniformly) can obscure critical biological structure—underestimating risk in highly constrained domains while potentially overstating it in more tolerant regions.

Despite these important advances, to date no study has systematically generated gene-specific priors for all disease genes. We still largely lack a comprehensive, scalable method to assign every gene (or gene region) a reliable prior probability of pathogenic variation. This gap has significant practical implications: if we continue to apply a uniform prior across all genes, we risk miscalibrating the Bayesian interpretation in cases where a gene is unusually mutation-tolerant or mutation-intolerant. Addressing this gap by developing robust, scalable gene-specific prior estimation methods is therefore essential. It will improve the accuracy and clinical utility of computational pathogenicity predictions, ensuring that each variant is evaluated against an appropriate baseline expectation for its gene. Ultimately, incorporating gene-specific priors into the Bayesian classification framework will make pathogenicity assessments more context-aware, leading to more reliable classifications and better outcomes in clinical genetic interpretation.

## 3.5 Methodology

### 3.5.1 Revised DistCurve Algorithm for Gene-Specific Prior Estimation

#### Rationale

The ACMG/AMP Bayesian framework requires specifying a prior probability of pathogenicity, which reflects the baseline expectation that a variant in a given gene is pathogenic. This prior is currently assumed to be universal across genes, despite strong evidence that different genes have vastly different levels of missense tolerance and pathogenic burden. Us-

ing a universal prior ignores disease biology: genes with strong evolutionary constraint or dominant-negative effects (e.g., *TP53*, *COL4A3*) should not have the same prior as genes tolerant to variation (e.g., *KCNQ4*).

To resolve this limitation, we extend the DistCurve positive–unlabeled (PU) learning framework (Zeiberg et al., 2020) from a genome-wide estimator to a **per-gene prior estimator**, enabling Bayesian calibration that is biologically meaningful and gene-aware.

### Workflow Overview

Figure 3.1 illustrates the overall pipeline.

1. **Define Positive and Unlabeled Sets.**
   - Positives: ClinVar Pathogenic/Likely Pathogenic (P/LP) missense variants for the gene (Landrum et al., 2018).
   - Unlabeled: variants from gnomAD v4.0 (Karczewski et al., 2020), all possible nucleotide substitutions (`csubs`), or all amino acid substitutions (`allaa`). These provide different assumptions about the mutational background.
2. **Feature Representation.** Each variant is encoded using 1,345 MutPred2 features (sequence-, structure-, and function-level predictors). No labels are used during feature construction, ensuring that feature space is label-agnostic.
3. **Dimensionality Reduction and PU Classification.** A neural network PU classifier (non-traditional classifier) projects the high-dimensional features to a 1D discriminant score. The classifier minimizing validation loss and maximizing AUC-PU is selected.
4. **DistCurve Prior Estimation.** For each hypothesized prior $\alpha$:
   (a) select the highest-scoring $\alpha \cdot |U|$ unlabeled variants as putative positives,
   (b) compute the Euclidean distance between the score distributions of labeled P/LP variants and the selected unlabeled subset.

   As $\alpha$ increases, the selected unlabeled variants eventually become dissimilar to true positives. Plotting distance vs. $\alpha$ yields a "distance curve," and the minimum distance corresponds to the optimal prior $\alpha^*$.
5. **Bootstrap Aggregation.** 5,000 bootstrap replicates are performed per gene. The final estimate is:

$$\tilde{\alpha} = \text{median}\left(\alpha_1^*, \alpha_2^*, \ldots, \alpha_{5000}^*\right); \qquad \text{IQR} = Q_{75} - Q_{25}$$

**Variant Data Sources and Filtering**

- ClinVar (2025-01 release) provided P/LP variants. Only missense variants with a review status of at least one star were included.
- Variants were mapped to MANE Select transcripts using VEP (McLaren et al., 2016).
- SpliceAI scores were used to remove variants likely affecting splicing (threshold: $\geq 0.2$) (Jaganathan et al., 2019).

**Modifications Made to DistCurve**

The original DistCurve assumes that bootstrapped samples are independent and that classifier training is not dominated by class imbalance. Variant effect data violate both assumptions. We introduce three algorithmic enhancements:

- **UNIBOOT (Unique Bootstrap Sampling)** Ensures that a variant appears either in the bootstrap or in the out-of-bag (OOB) evaluation set, never both. This prevents duplicate predictions from artificially reducing variance.
- **Principal Component Analysis (PCA)** Reduces the correlational structure of MutPred2 features. Only PCs explaining $\geq 95\%$ variance were retained. The same projection is applied to OOB data to prevent information leakage.
- **Random Under-Sampling (RUS)** Balances the ratio of P/LP and unlabeled variants in each training fold, stabilizing PU classifier learning.

Together, UNIBOOT + PCA + RUS reduce classifier variance and improve separability.

### 3.5.2 Impact of Including MutPred2 Training Variants

To evaluate whether the prior estimation is affected by the inclusion of MutPred2 training variants, we estimated priors under two conditions:

1. including all observed variants, and
2. excluding MutPred2 training variants from positive and unlabeled sets.

Because DistCurve is trained only on **label-agnostic features**, we hypothesized negligible effect. We quantified effects by comparing classifier AUC-PU and the estimated medians of $\alpha^*$ across genes.

### 3.5.3 Effect of Mixture Set Choice and Mode of Inheritance

To assess whether the choice of mixture set interacts with clinical genetics, we evaluated mixture-specific priors stratified by mode of inheritance. For each gene with an estimated prior, we obtained its mode-of-inheritance (MOI) annotation from ClinGen gene–disease validity curation records (downloaded from the ClinGen website; (Rehm et al., 2015)). Genes were mapped to one of five MOI categories: autosomal dominant (AD), autosomal recessive (AR), mixed AD/AR, semidominant (SD), or X-linked. Genes without a clear ClinGen MOI assignment, or with conflicting or provisional annotations, were excluded from this analysis. Only genes for which priors could be estimated under all three mixture sets were retained, so that comparisons would be performed on identical gene sets within each MOI group.

Using the revised DistCurve pipeline, we estimated gene-specific priors separately with each mixture set: `gnomAD`, `csubs`, and `allaa`. Within each MOI category, we then carried out pairwise, gene-wise comparisons of priors between mixture sets (gnomAD vs. csubs, gnomAD vs. allaa, and csubs vs. allaa). For each comparison, we report the median within-gene difference in priors, the Wilcoxon $p$-value to quantify the magnitude of mixture-set effects within each inheritance group.

## 3.6 Final Gene-Level and Domain-Level Calibration Inputs

After validating methodological decisions, we used the following configuration for the full calibration pipeline:

- mixture set: **gnomAD**,
- no removal of MutPred2 training variants,
- DistCurve + UNIBOOT + PCA + RUS for genes with $\geq 10$ P/LP variants.

For genes lacking sufficient P/LP counts, a domain-aware extension was applied:

1. Pfam domains were assigned to variants using UCSC Genome Browser annotations.
2. Variant score distributions (MutPred2, AlphaMissense, REVEL) were compared via Jensen–Shannon distance.
3. Hierarchical clustering was performed to construct domain clusters (max 1,500 P/LP variants/cluster).

Priors were then estimated at the **cluster level**, enabling prior inference for 2,662 genes.

## 3.7 Results

### 3.7.1 Algorithmic Modifications Improve Stability and Consistency of DistCurve Priors

The original DistCurve implementation produced unstable and artificially inflated priors when applied directly to MutPred2 scores. Score distributions exhibited undesirable bimodality due to MutPred2 training exposure to ClinVar variants, and the bootstrap procedure allowed the same variant to appear simultaneously in both bootstrap and out-of-bag (OOB) sets, producing inconsistent predictions and noisy turning points in the distance curve.

To overcome these issues, we introduced three targeted modifications (UNIBOOT, PCA, and random under-sampling). UNIBOOT enforces mutual exclusivity between bootstrap and OOB sets, eliminating duplicate predictions. PCA reduces the high-dimensional MutPred2 feature space and lowers variance in bootstrap iterations. Random under-sampling balances the PU training data and improves score separability. Together, these changes increased classification AUC-PU, reduced bootstrap variance, and stabilized the prior estimation curve.

As a result, the combined pipeline (UNIBOOT + PCA + RUS) produces reproducible and gene-consistent prior estimates and resolves failure modes in the original DistCurve implementation. All subsequent analyses use this improved version of DistCurve.

### 3.7.2 MutPred2 Training Variants Have Different Impact on Prior Estimation

To assess whether MutPred2 training data leakage could bias prior estimation, we recomputed priors and calibration performance after explicitly removing MutPred2 training variants from both the labeled P/LP set and the unlabeled mixture sets. Across the three mixture sets (`gnomAD`, `csubs`, and `allaa`), paired comparisons of priors showed that filtering MutPred2 training variants led to *statistically significant but numerically small* shifts in median priors for the `gnomAD`-based mixture ($p = 1.2 \times 10^{-7}$), whereas effects for `csubs` and `allaa` were negligible ($p = 0.28$ and $p = 0.03$, respectively; Figure 3.2). Classifier discrimination (AUC-PU) was consistently higher when MutPred2 training variants were retained across all three mixture sets, indicating that excluding these variants slightly degrades separability between P/LP and unlabeled variants.

At the gene level, priors with and without filtering training variants were highly correlated: classification performance (AUC-PU) remained consistent ($r = 0.912$, Figure 3.3a), and resulting gene-specific priors showed strong concordance ($r = 0.788$, Figure 3.3b). Only

a small subset of genes—most notably *TP53* and *GCK*—exhibited noticeable shifts in prior estimates. These genes had unusually large overlaps between MutPred2 training data and both ClinVar and gnomAD (up to 49–77% of shared variants), explaining their stronger sensitivity to filtering. For the vast majority of genes, filtering did not significantly change score distributions, classifier performance, or resulting priors.

Given that removing variants further reduces data availability—particularly problematic for genes with already sparse P/LP counts—we retain all variants (including MutPred2 training examples) for the final prior estimation. This decision maximizes data use while introducing negligible impact on prior accuracy for nearly all genes.

### 3.7.3 Effect of Mixture Set Choice on Prior Estimation

To evaluate whether the choice of mixture set influences prior estimation (Figure 3.2A; Table 3.1), we compared gene-level priors obtained using `gnomAD`, `csubs`, and `allaa`. For genes with priors available under all three mixtures, we used Wilcoxon paired signed-rank tests and applied Benjamini–Hochberg FDR correction.

Priors derived from `gnomAD` and `allaa` were statistically indistinguishable (median difference $= -0.0005$, $p = 0.98$), indicating that `allaa` provides estimates effectively equivalent to `gnomAD`. In contrast, `csubs` systematically produced higher priors: relative to `gnomAD`, the median within-gene difference was approximately $+0.01$ ($p = 5.2 \times 10^{-5}$), with a similar shift when compared to `allaa` ($p = 2.5 \times 10^{-7}$).

Taken together, these results indicate that although `csubs` can serve as a mixture background, it tends to inflate gene-specific priors compared to `gnomAD`/`allaa`. By contrast, `gnomAD` and `allaa` yield essentially equivalent priors, suggesting that either provides a stable and unbiased representation of the background variant space. Given this equivalence and computational considerations, we prioritize `gnomAD` for downstream calibration.

Table 3.1: Pairwise comparisons of gene-level priors across mixture sets (all inheritance modes combined). Mean and median differences are reported as (first mixture) − (second mixture). Wilcoxon paired signed-rank tests were used for $p$-values; $q_{\mathrm{BH}}$ denotes Benjamini–Hochberg FDR-adjusted $p$-values.

| Pair | $n$ | Mean diff | Median diff | Wilcoxon $p$ | Cliff's $\delta$ | $q_{\mathrm{BH}}$ |
|---|---|---|---|---|---|---|
| gnomAD–csubs | 86 | -0.0220 | -0.0132 | $5.22 \times 10^{-5}$ | -0.224 | $7.83 \times 10^{-5}$ |
| gnomAD–allAA | 88 | 0.0072 | -0.0005 | 0.9778 | -0.056 | 0.9778 |
| csubs–allAA | 86 | 0.0293 | 0.0095 | $2.46 \times 10^{-7}$ | 0.214 | $7.38 \times 10^{-7}$ |

### 3.7.4 Impact of Mixture Set Choice Stratified by Mode of Inheritance

Because the mixture set determines the unlabeled variant distribution used during prior estimation, we evaluated whether different inheritance categories showed different sensitivity to mixture set selection. This analysis was motivated by the possibility that certain mixture sets may be systematically biased for some inheritance modes; for example, X-linked genes may be underrepresented in population resources such as gnomAD due to hemizygosity and sex-specific ascertainment, while autosomal recessive genes may be better represented in single-nucleotide–substitution–based mixture sets (csubs) because recessive diseases tolerate more standing variation in the population. To evaluate this, we stratified genes by their clinical inheritance model (Figure 3.4): autosomal dominant (AD, $n = 53$), autosomal recessive (AR, $n = 11$), mixed AD/AR ($n = 10$), X-linked ($n = 11$), and semidominant (SD, $n = 3$). For each group, we compared gene-specific priors derived from each mixture set (gnomAD, csubs, allAA) using paired Wilcoxon tests and effect sizes (Cliff's $\delta$). (Figure 3.5; Table 3.2).

Table 3.2: Pairwise comparisons of gene-level priors across mixture sets, stratified by mode of inheritance. Mean and median differences are reported as (first mixture) − (second mixture). Wilcoxon paired signed-rank tests were used for $p$-values; $q_{\mathrm{BH}}$ denotes Benjamini–Hochberg FDR-adjusted $p$-values.

| Inheritance | Pair | $n$ | Mean diff | Median diff | Wilcoxon $p$ | Cliff's $\delta$ | $q_{\mathrm{BH}}$ |
|---|---|---|---|---|---|---|---|
| AD | gnomAD–csubs | 51 | -0.0211 | -0.0151 | 0.0051 | -0.233 | 0.0076 |
| AD | gnomAD–allAA | 53 | 0.0033 | 0.0002 | 0.6741 | -0.114 | 0.6741 |
| AD | csubs–allAA | 51 | 0.0243 | 0.0072 | 0.0012 | 0.154 | 0.0035 |
| AD/AR | gnomAD–csubs | 10 | -0.0096 | -0.0105 | 0.0469 | -0.200 | 0.0703 |
| AD/AR | gnomAD–allAA | 10 | -0.0082 | -0.0042 | 0.0367 | -0.180 | 0.0703 |
| AD/AR | csubs–allAA | 10 | 0.0014 | 0.0000 | 0.5076 | 0.000 | 0.5076 |
| AR | gnomAD–csubs | 11 | -0.0043 | -0.0027 | 0.3281 | -0.058 | 0.3281 |
| AR | gnomAD–allAA | 11 | 0.0269 | 0.0151 | 0.0505 | 0.372 | 0.0757 |
| AR | csubs–allAA | 11 | 0.0312 | 0.0156 | 0.0033 | 0.471 | 0.0100 |
| SD | gnomAD–csubs | 3 | 0.0034 | 0.0209 | 1.0000 | 0.333 | 1.0000 |
| SD | gnomAD–allAA | 3 | 0.0808 | 0.1267 | 0.2850 | 0.333 | 0.4276 |
| SD | csubs–allAA | 3 | 0.0774 | 0.1059 | 0.1088 | 0.778 | 0.3264 |
| X-linked | gnomAD–csubs | 11 | -0.0622 | -0.0621 | 0.0058 | -0.455 | 0.0175 |
| X-linked | gnomAD–allAA | 11 | 0.0005 | -0.0122 | 0.5337 | -0.190 | 0.5337 |
| X-linked | csubs–allAA | 11 | 0.0627 | 0.0429 | 0.0329 | 0.471 | 0.0493 |

For **autosomal dominant (AD)** genes—which represent the majority of the genes with sufficient data—gnomAD and allAA produced statistically indistinguishable priors ($p = 0.67$, median difference = 0.0002). In contrast, csubs yielded higher priors than both gnomAD and

allAA (gnomAD–csubs: $p = 0.005$; csubs–allAA: $p = 0.001$), although the effect sizes were small (Cliff's $\delta < 0.25$). Thus, for AD genes, **gnomAD and allAA are interchangeable**, whereas csubs tends to inflate prior estimates (Figure 3.6).

**Autosomal recessive (AR)** genes showed a different pattern from AD genes. Here, *allAA* yielded systematically *lower* priors than both `gnomAD` and `csubs` (csubs–allAA: $p = 0.003$, $q = 0.01$; gnomAD–allAA: $p = 0.05$, $q = 0.076$), with effect sizes in the moderate range (Cliff's $\delta = 0.36$–$0.47$). The difference between `gnomAD` and `csubs` was not significant (gnomAD–csubs: $p = 0.33$). These results suggest that, for AR genes, expanding the mixture to all possible amino-acid substitutions introduces a large number of hypothetical variants not observed in populations, effectively diluting the estimated baseline pathogenicity and lowering the inferred priors. Practically, **gnomAD or csubs are preferred over allAA** for AR genes, with the caveat that the sample size is limited ($n = 11$).

**X-linked genes** showed the strongest sensitivity to mixture choice. For this group, priors derived from `gnomAD` were markedly lower than those from `csubs` (gnomAD–csubs: $p = 0.006$, $q = 0.018$; median difference $\approx -0.062$), with a large effect size (Cliff's $\delta = -0.45$), indicating a systematic downward shift in gnomAD-based priors relative to `csubs`. In contrast, priors obtained from `allaa` were broadly similar to those from `gnomAD` (small, non-significant within-gene differences), suggesting that `allaa` behaves more like a neutral background for X-linked genes in this setting. This pattern is consistent with the expectation that population sequencing datasets may underrepresent X-linked pathogenic variation due to hemizygous selection and sex-biased ascertainment, leading gnomAD to underestimate the baseline burden of pathogenic missense variants. The synthetic `allaa` mixture, which is not shaped by selection, does not show as strong a deviation, and thus falls intermediate between gnomAD and csubs in terms of inferred priors.

For mixed AD/AR genes, small sample size ($n = 10$) made interpretation less robust, but gnomAD again tended to produce slightly lower priors relative to allAA and csubs. The semidominant (SD) group contained only three genes, and no robust conclusions could be drawn.

Across inheritance modes, genes with the largest mixture-set–dependent shifts included *MED12*, *GCK*, *F9*, *COL1A1*, and *SLC6A1*. These genes had prior ranges of 0.12–0.26 across mixture sets, reflecting strong sensitivity to mixture choice. Notably, several of these genes are X-linked (*MED12*, *F9*) or associated with founder or population-specific variation, suggesting that mixture-set selection interacts with variant sampling and inheritance biology rather than representing noise in the estimation process.

In summary, mixture-set selection affects prior estimation in inheritance-specific ways. For AD genes, gnomAD and allAA are equivalent and preferred. For AR genes, gnomAD

or csubs are preferred over allAA. These findings indicate that mixture selection should not be uniform across all genes and that incorporating inheritance-aware mixture strategies may improve prior estimation for genes outside of the dominant model.

### 3.7.5 Gene-Specific Prior Estimation Using the Fully Modified DistCurve Pipeline

Using the final configuration (**gnomAD mixture**, **UNIBOOT + PCA + RUS**, and **no MutPred2 filtering**), we estimated priors for all genes with $\geq 10$ P/LP missense variants (Figure 3.7). Estimated priors ranged from 1.3% to 30% with median of around 7%. Genes historically associated with dominant-negative mechanisms or extreme constraint showed elevated priors (e.g., *MED12*, *COL4A3*), whereas genes tolerant to missense variation showed low priors (e.g., *KCNQ4*). Most genes showed narrow prior uncertainty, indicating convergence of the PU classifier and distance curve.

To assess biological validity, DistCurve priors were compared to literature-derived estimates from population-based disease models (Bhat et al., 2025) and functional assays (Clark et al., 2019; Jain et al., 2025). For well-characterized genes, DistCurve priors matched external priors within confidence intervals, confirming alignment with observed disease frequencies and experimentally measured functional defect rates.

### 3.7.6 Domain-Level Prior estimation Enables Coverage of Genes Without Sufficient Labeled Data

Gene-level prior estimation using DistCurve requires a sufficient number of labeled pathogenic variants ($N_{PLP} \geq 10$). However, many clinically relevant genes lack enough P/LP variants to support reliable gene-level modeling. To expand prior estimation to these genes, we performed domain-based clustering by grouping Pfam domains according to the similarity of their variant score distributions, measured using Jensen–Shannon distance computed separately for AlphaMissense (AM) Cheng (2023) and REVEL Ioannidis et al. (2016) prediction scores. In this framework, prior estimation is performed at the **cluster** level rather than at the gene level, allowing unlabeled domains with similar score distributions to borrow statistical strength from each other.

Using AlphaMissense, we identified 92 domain aggregate clusters, of which 88 contained at least 10 pathogenic variants and were therefore eligible for prior estimation. The resulting cluster-level priors showed a median of 0.050 with an interquartile range (IQR) of 0.035–0.068, and more than half of these clusters (51 of 88) exceeded the genome-wide prior

of 4.41% reported by Pejaver et al. (2022)(Figure 3.8). This elevation is expected, as our clusters are restricted to genes with established clinical validity and sufficient ClinVar representation, rather than all possible missense variation genome-wide. Domains without Pfam annotations were grouped together into a universal **non-domain** cluster. This aggregated region yielded a median prior of 0.053 (IQR: 0.0496–0.0594), which remains slightly higher than the 4.41% genome-wide benchmark. Several factors likely contribute to this: (i) the underlying gene set is enriched for disease-associated genes, (ii) Pfam domain boundaries are defined by profile-HMM alignments and are not perfectly sharp, so variants near domain edges may be inconsistently labeled as domain vs. non-domain, and (iii) even outside annotated Pfam domains, many of these proteins contain functionally important, constrained regions. Together, these effects produce a non-domain background that is still enriched for pathogenic missense variation relative to the previous genome-wide prior.

Using REVEL, we identified 98 dynamic clusters, 91 of which contained at least 10 pathogenic variants. The REVEL-based clusters showed a slightly higher median prior of 0.059 with an IQR of 0.042–0.076, and 65 of the 91 clusters exceeded the genome-wide prior benchmark (Figure 3.9). The aggregated **non-domain** cluster produced a universal prior of 0.0527 (IQR: 0.0481–0.0566).

Together, these results demonstrate that domain-level clustering produces consistent and biologically meaningful prior estimates across both AlphaMissense and REVEL score spaces, while enabling prior inference even for genes or domains that lack sufficient labeled data for gene-level modeling. The consistency of the priors across predictors further supports the robustness of domain-based prior estimation as a complementary strategy to gene-specific priors.

## 3.8 Discussion

### 3.8.1 Key Improvements in Prior Estimation

Our modified DistCurve pipeline substantially improves the stability and consistency of gene-specific prior estimates. By enforcing exclusive bootstrap/out-of-bag sets (UNIBOOT), dimensionality reduction (PCA), and balanced training (RUS), we eliminated the spurious bimodal score distributions and noisy prior curves seen with the original implementation. These algorithmic changes yielded more reliable classifier performance (higher AUC-PU) and reduced the variance in bootstrap iterations. The refined pipeline produces reproducible pathogenicity priors that are consistent across genes and free of the failure modes (unstable or inflated priors) observed initially. All subsequent analyses leverage this improved

DistCurve version, ensuring that our conclusions rest on a robust and unbiased prior estimation framework.

Another important finding is that potential data leakage from predictor training (MutPred2) has only minimal impact on prior estimates. Removing MutPred2 training variants from our labeled and unlabeled sets caused statistically significant but small shifts in the overall prior for one mixture (gnomAD), and negligible changes for others. Even at the gene level, priors with and without filtering were highly correlated, and classification AUC-PU remained virtually unchanged. Only a few genes (e.g. TP53 and GCK) showed noticeable prior differences when MutPred2-trained variants were removed, likely because these genes had an unusually large overlap of training data with ClinVar/gnomAD. For the vast majority of genes, including all available variants does not appreciably bias the estimated prior. Given that filtering would further thin out already-sparse data for certain genes, we elected to retain all variants for final modeling – a choice that maximizes data usage while introducing negligible prior bias in nearly all cases.

### 3.8.2 Influence of Mixture Set and Inheritance on Gene Priors

Our analyses show that the choice of unlabeled mixture set can influence gene-level priors, and that this influence depends on the mode of inheritance. Overall, priors derived from `gnomAD` and from the all–amino-acid synthetic set (`allaa`) were statistically indistinguishable (median difference $\approx 0$, $p \approx 1$), indicating that `allaa` captures a background spectrum similar to `gnomAD` for the purpose of prior estimation. In contrast, the all–possible–nucleotide–substitution set (`csubs`, comprising all single-nucleotide variants within the coding region) consistently produced slightly higher priors than either `gnomAD` or `allaa`. One plausible explanation is that `csubs` samples only mutationally accessible SNVs, which may be enriched for amino-acid changes more similar to those observed in disease genes, whereas `allaa` includes many protein-level substitutions that would require multiple nucleotide changes and may be less representative of real-world mutational processes.

Stratifying by inheritance helps clarify where these effects matter. For autosomal dominant (AD) genes, which form the largest group, `gnomAD` and `allaa` remained effectively interchangeable, whereas `csubs` yielded modestly elevated priors. This is consistent with the idea that a population-based mixture already captures the relevant background for dominant disease genes, and that the SNV-based `csubs` mixture introduces a small upward bias.

Autosomal recessive (AR) genes showed a different pattern: `allaa` produced systematically *lower* priors than both `gnomAD` and `csubs`. Because `allaa` enumerates all possible amino-acid substitutions, it greatly expands the space of potential missense changes, many

of which may never occur in humans. This enlarges the denominator of "possible variants" without adding corresponding pathogenic observations, leading to diluted priors. For recessive genes, mixtures based on observed or SNV-accessible variation (`gnomAD` or `csubs`) therefore provide a more realistic representation of tolerated background variation.

X-linked genes exhibited the strongest mixture sensitivity: priors derived from `gnomAD` were substantially lower than those from `csubs`, whereas `allaa`-based priors were intermediate and not significantly different from `gnomAD`. This is consistent with known underrepresentation of pathogenic X-linked variation in population cohorts due to hemizygous selection and sex-biased ascertainment; `gnomAD` may therefore underestimate the true burden of pathogenic missense variants on the X chromosome. In contrast, `csubs`, which is not constrained by population sampling and includes all possible SNVs, yields higher priors that may better reflect the underlying disease architecture for these genes.

Across inheritance modes, the genes most sensitive to mixture choice (e.g. *MED12*, *F9*, *COL1A1*, *GCK*, *SLC6A1*) tend to have distinctive biological properties, such as X-linkage, extreme constraint, or founder-variant contributions. This suggests that mixture-set effects are not random noise, but arise from interactions between mixture construction and gene-specific biology. Practically, our findings imply that a single mixture choice is adequate for many AD genes (`gnomAD` or `allaa`), but that recessive and X-linked genes may benefit from inheritance-aware strategies that either avoid `allaa` (for AR) or supplement `gnomAD` with SNV-complete mixtures like `csubs` (for X-linked genes) when estimating priors. We note that these mechanistic explanations are plausible interpretations consistent with current knowledge, but were not formally tested in this study.

### 3.8.3 Domain-Level Prior Estimation Extends Coverage

A key contribution of this work is the introduction of domain-cluster–based prior estimation, which addresses the data scarcity problem for genes with too few pathogenic variants to model individually. By clustering Pfam domains across the genome based on similarity in variant effect score distributions, we effectively pooled data from multiple genes that share biochemical or structural properties. This allowed us to estimate cluster-level pathogenicity priors for domains that individually might lack sufficient P/LP variants. Notably, the domain-based priors were consistent and biologically plausible. Across 88 AlphaMissense-based clusters (and similarly 91 REVEL-based clusters), median priors ranged around 5–6%, with an interquartile range of roughly 3–7%. Over half of the domain clusters yielded priors above the canonical genome-wide prior of 4.4% reported for missense variants, underscoring that certain protein domains are enriched for pathogenic variation relative to the genomic

average. This approach effectively expands the reach of prior estimation to hundreds of genes that could not be analyzed at the gene level due to sparse data. The fact that we obtained highly concordant results using two different predictive score sets (AlphaMissense and REVEL) further attests to the robustness of the clustering strategy – it captures intrinsic domain risk properties not tied to any single algorithm.

### 3.8.4 Limitations and Future Directions

While our approach represents a step forward in gene-specific and domain-specific prior estimation, several limitations remain. A fundamental challenge is the limited number of known pathogenic variants for many genes. We set a threshold of $N_{P/LP} \geq 10$ to compute a gene's prior, which restricted our analysis to 86 genes. Many clinically relevant genes have fewer than 10 confirmed pathogenic missense variants, making their individual prior estimates highly uncertain or simply impossible with current data. Our domain clustering is a pragmatic solution to this data sparsity, but it is not a perfect substitute for gene-level modeling – it assumes that variants in different genes can be treated as coming from the same distribution if they reside in similar domains. In reality, genes within a cluster may still differ in subtle ways (e.g. regulatory context or interaction partners), and the cluster prior represents an average. Thus, for genes without sufficient data, there is an inherent uncertainty that no method can fully erase without additional information. Another limitation is that our priors are derived from known pathogenic variants in ClinVar, which could bias the priors if the ClinVar data itself is biased toward certain types of variants or genes. We mitigated this (via PU learning and careful bootstrapping), but any systematic gaps in ClinVar (such as underreporting of pathogenic variants in less-studied genes) would affect our estimates. The mixture set biases discussed earlier (e.g. for X-linked genes) also highlight that no single unlabeled set is optimal for all scenarios. Although we chose gnomAD as a reasonable default, future work might implement a dynamic approach (for example, automatically switching to allAA for genes suspected to be recessive or X-linked) to avoid underestimating priors in those special cases.

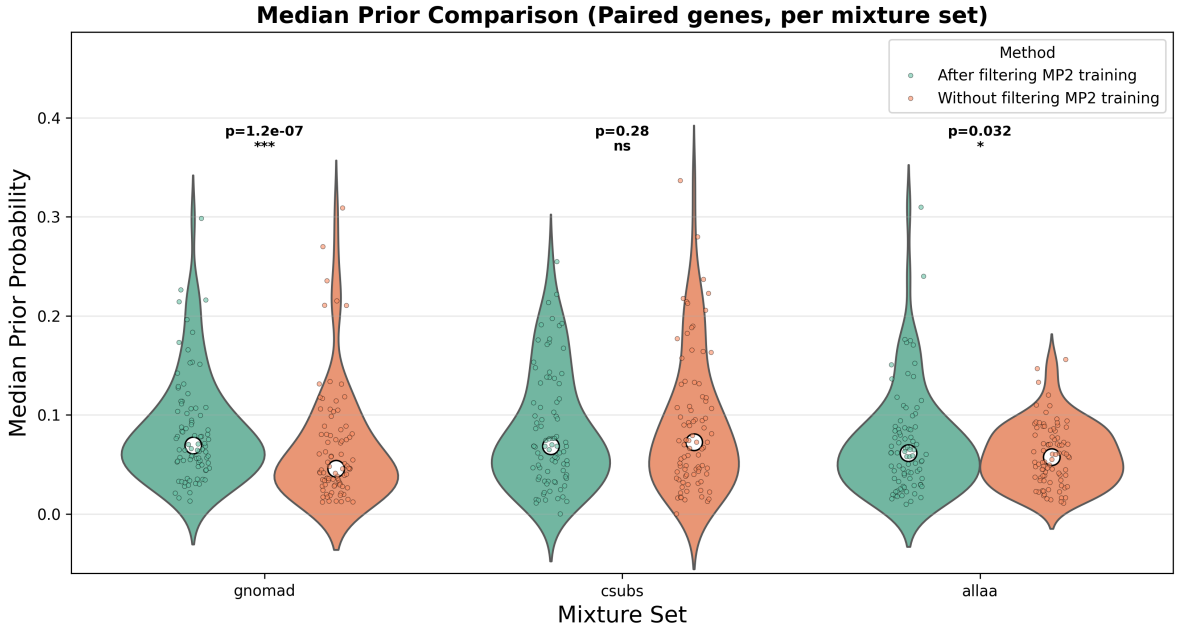
There are several exciting directions to extend this work. One avenue is to incorporate richer feature sets and modern predictive models to further improve the classifier that underlies DistCurve. In this study, we used MutPred2 feature matrix (and reduced features via PCA) for the PU classifier; however, recent advances in protein language models and structure-based predictors could offer more powerful representations of variant effect. For instance, embeddings from large protein language models (such as ESM or ProtGPT) or hybrid models like AlphaMissense could serve as input features that capture subtle biochemical

and evolutionary context. These latent features might help differentiate pathogenic vs benign variants more accurately, especially in genes where sequence context is crucial. Using such high-dimensional learned features might also allow us to cluster variants or domains in a more nuanced way (beyond Pfam domains), potentially revealing new groupings of genes with similar variant effect profiles. Early studies have shown that protein language model embeddings can boost variant effect prediction performance, so integrating these into DistCurve is a logical next step.

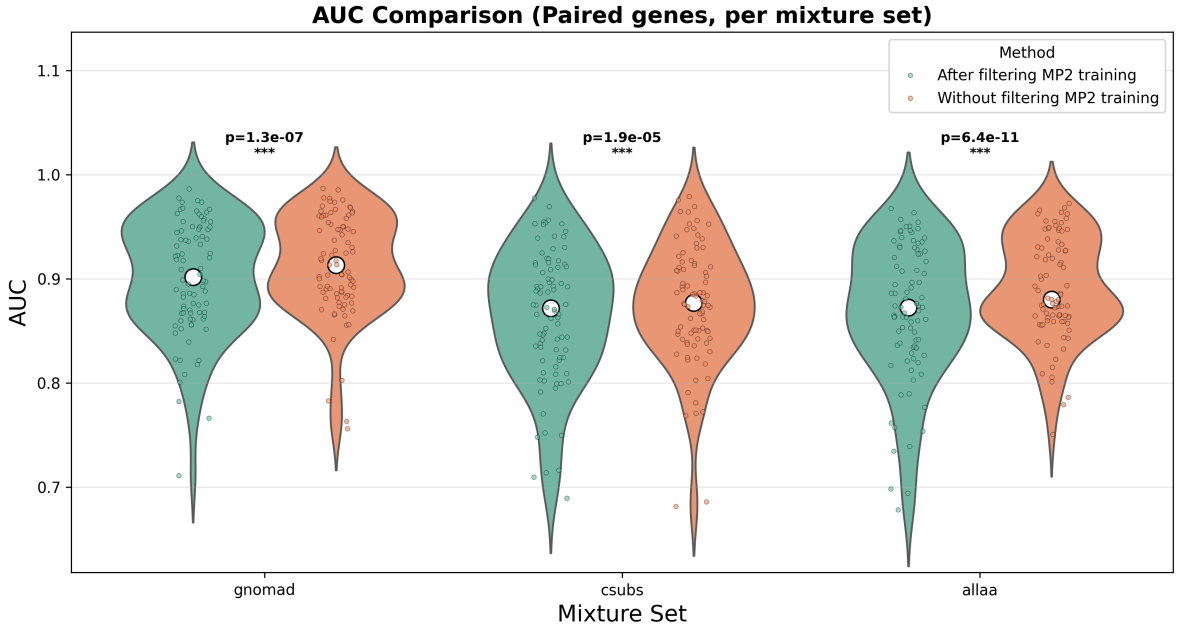
Another promising direction is to leverage semi-supervised learning across genes to address label bias and data scarcity. In this work, we treat each gene (or domain cluster) largely independently, but one could instead adopt a unified framework that shares information across genes while still allowing gene-specific behavior. Zeiberg et al. (2022) proposed exactly this kind of approach: a semi-supervised algorithm that exploits the structure of grouped, biased data to learn a group-aware, probability-calibrated classifier. In our setting, the groups would be genes (or domain clusters), and the method's core assumption—that the partition projects class-conditional invariance across groups—would allow it to use unlabeled data and well-annotated genes to improve classification in sparsely labeled genes. Conceptually, such a model could combine gene-specific priors with a shared decision boundary in feature space, so that patterns learned from data-rich genes inform predictions in data-poor genes while still respecting group-level differences. Adapting this framework to variant pathogenicity prediction could relax our current requirement for a minimum number of labeled variants per gene and yield more reliable calibrated probabilities for underrepresented genes. Exploring this class of group-aware, semi-supervised methods to further refine prior estimation and calibration in low-data genes is a natural next step beyond the present work.

Finally, we envision expanding the validation and application of these priors in clinical and research contexts. One future experiment is a prospective validation: using our gene-specific priors to re-classify variants of uncertain significance (VUS) and then tracking clinical or functional follow-up outcomes. If our high-prior genes indeed produce more pathogenic VUS (and low-prior genes more benign VUS), that would support the real-world utility of these estimates in variant interpretation workflows. Additionally, integrating our priors into a Bayesian framework for genetic risk (analogous to how pathogenic variant frequencies inform risk models in hereditary cancer studies) could improve the accuracy of carrier risk predictions. As more data become available – from population sequencing, disease cohorts, and multiplex assays – our priors can be continually updated and refined. In summary, we have laid a strong foundation for gene- and domain-specific missense variant priors using a semi-supervised learning approach. With further enhancements in algorithms, feature representation, and cross-gene learning, this framework can be extended to provide ever more

precise and comprehensive priors, ultimately aiding the interpretation of genetic variants across the spectrum of human disease.




**(A) Median prior comparison**




**(B) AUROC comparison**

Figure 3.2: Effect of excluding MutPred2 training variants during prior estimation and calibration. Violin plots show the distribution across genes for each of the three variant mixture sets used during calibration: **gnomAD** (population variants), **csubs** (all possible single nucleotide substitutions), and **allAA** (all possible amino acid substitutions). Within each mixture set, the two calibration strategies are shown side-by-side: **excluding MutPred2 training variants** vs. **including all variants**. For both (A) the median prior estimates and (B) the calibrated AUROC values, the horizontal line denotes the median per group and each point corresponds to a gene. P-values above each mixture set indicate whether excluding training variants results in a statistically significant shift (Wilcoxon signed-rank test).

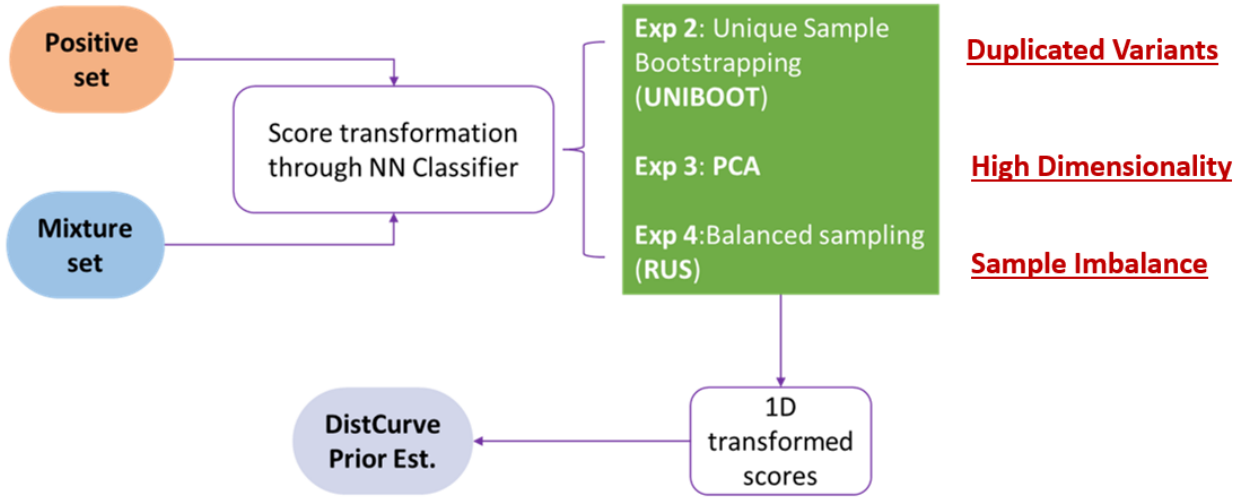
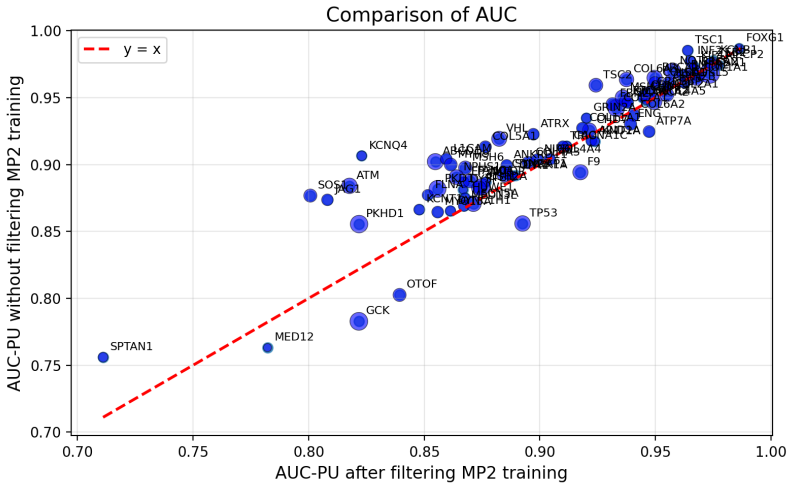
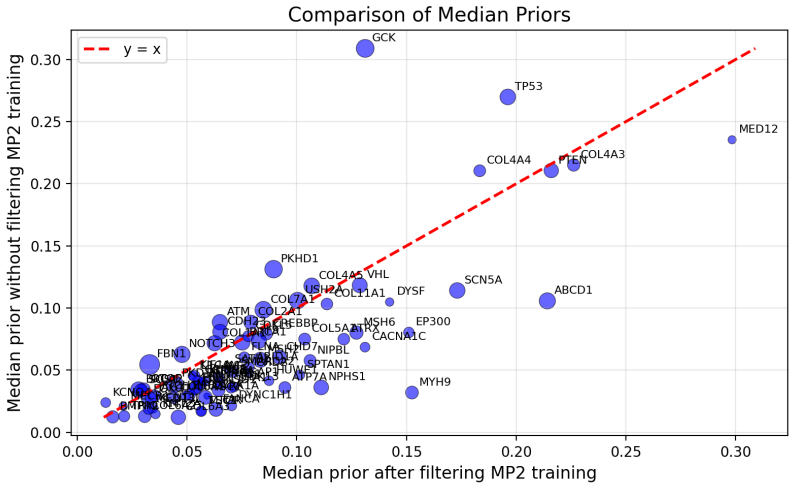



Figure 3.1: Revised DistCurve workflow for gene-specific prior estimation. Variants are represented using the 1,345-dimensional MutPred2 feature matrix. A PU classifier transforms this space into a one-dimensional score reflecting similarity to pathogenic (P/LP) variants. Three enhancements improve stability and reduce bias: (1) **UNIBOOT** ensures mutually exclusive bootstrap/out-of-bag sampling, (2) **PCA** reduces dimensionality and prevents overfitting, and (3) **Random Under-Sampling (RUS)** balances P/LP and unlabeled sets during classifier training. The resulting scores are used by DistCurve to generate a distance curve across hypothesized priors, from which the optimal prior $\alpha^*$ is selected.




(a) Classifier AUROC with and without MutPred2 training variants.




(b) Median prior estimates with and without MutPred2 training variants.

Figure 3.3: Gene-wise comparison of prior classifier performance and estimated priors when including versus excluding MutPred2 training variants from the mixture set. (A) Comparison of classifier AUROC for distinguishing pathogenic versus unlabeled variants when MutPred2 training variants are filtered out (y-axis) versus kept (x-axis). (B) Comparison of median gene-specific prior estimates under the same two settings. Each point represents one gene; diagonal lines indicate no change between the two conditions.

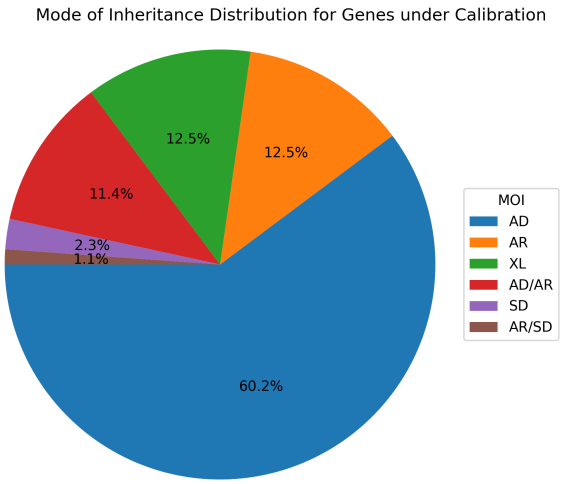
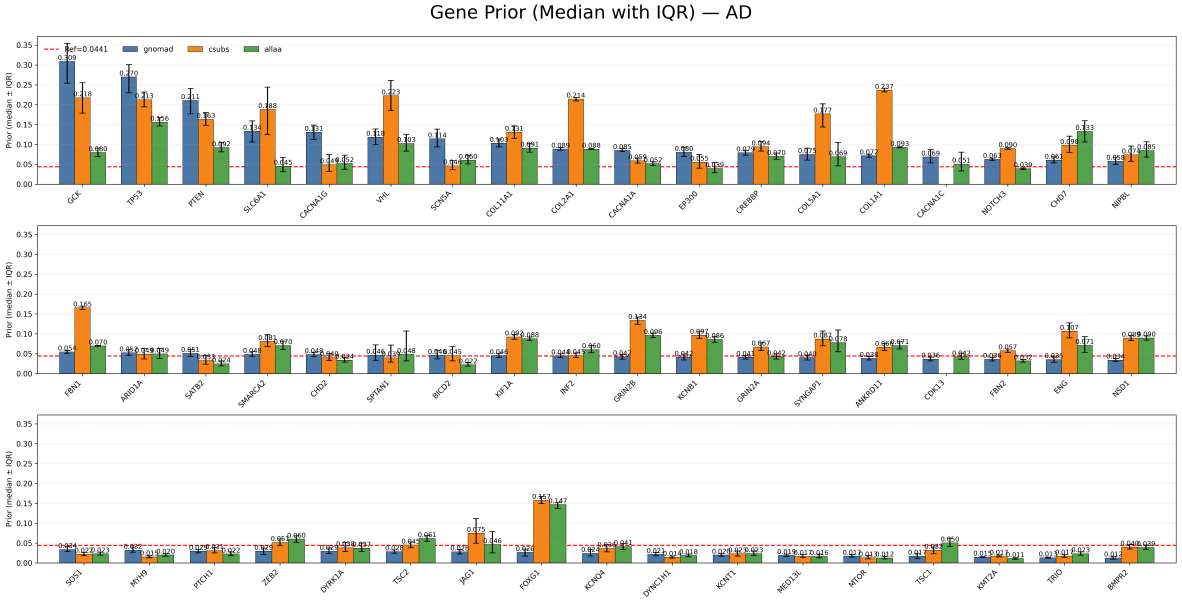
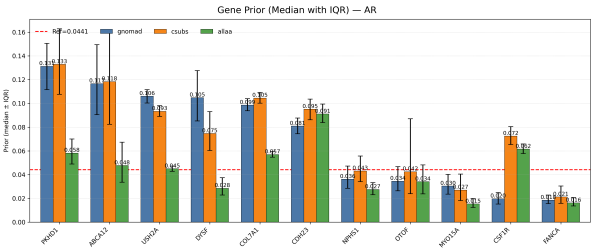


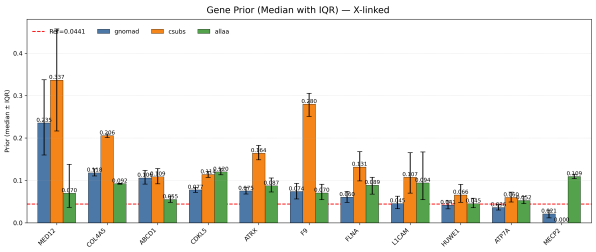

Figure 3.4: Distribution of inheritance modes among genes for which gene-specific pathogenicity priors were successfully calibrated.



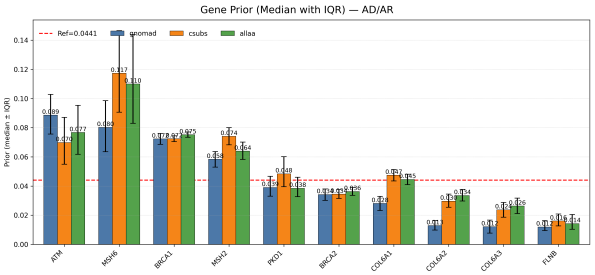

(a) Autosomal Dominant (AD) genes



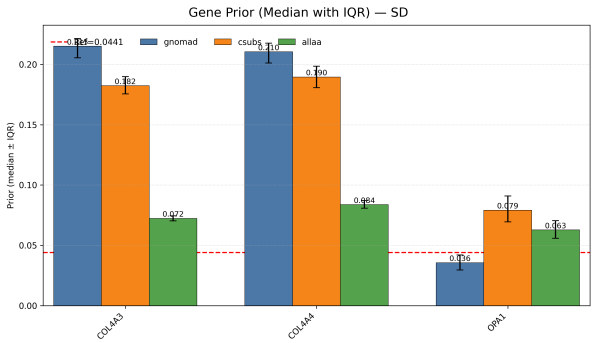

(b) Autosomal Recessive (AR) genes




(c) X-linked genes

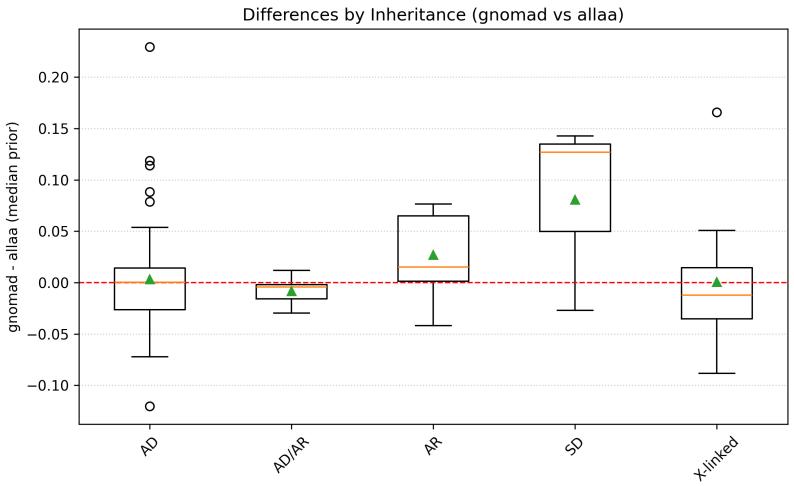



(d) Genes annotated as both AD and AR

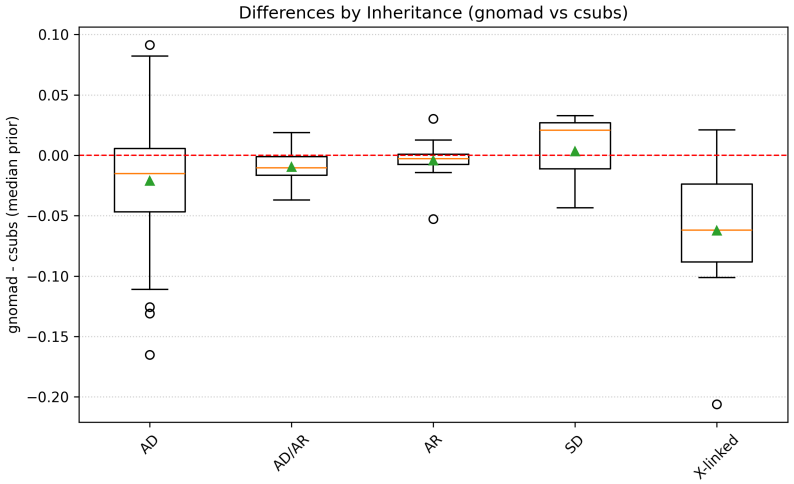



(e) Semi-dominant (SD) genes

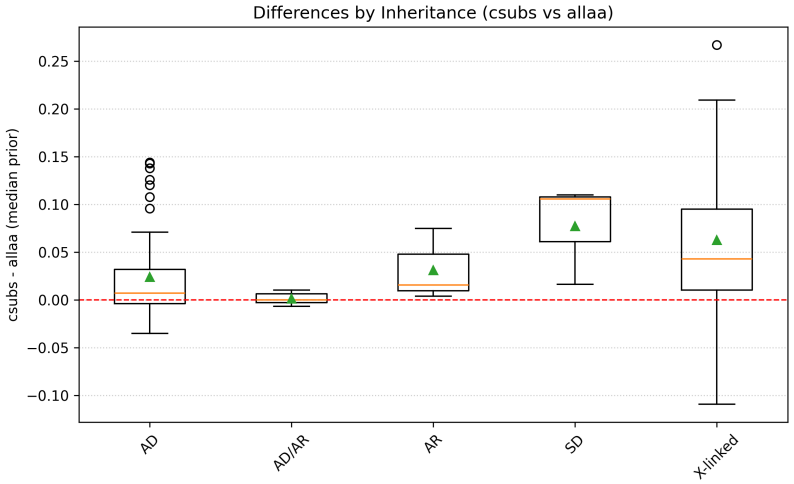
Figure 3.5: Gene-specific prior probability of pathogenicity stratified by mode of inheritance. For each gene, three bars are shown side-by-side representing prior estimates using different mixture sets: **gnomAD** (blue), **all possible nucleotide substitutions (csubs)** (orange), and **all possible amino acid substitutions (allaa)** (green). Each bar represents the **median prior estimate** from 5000 bootstrap samples, with error bars showing the **25th–75th percentile**. The horizontal red dashed line corresponds to the genome-wide prior of **4.41%** from Pejaver et al. (2022).




(a) gnomAD vs. all possible amino acid substitutions (allAA)




(b) gnomAD vs. all possible nucleotide substitutions (cSubs)




(c) cSubs vs. allAA mixture set comparison

Figure 3.6: Caption continued on next page.

**Figure 3.6.** Pairwise differences in prior probability estimates between mixture sets across inheritance modes. Each point represents a gene. Positive values indicate that the first mixture set in the comparison produces a higher estimated prior than the second. The **red dashed line** marks zero difference, and **green markers** represent mean differences across genes. Semi-dominant (SD) results should be interpreted cautiously due to the small sample size ($n = 3$).

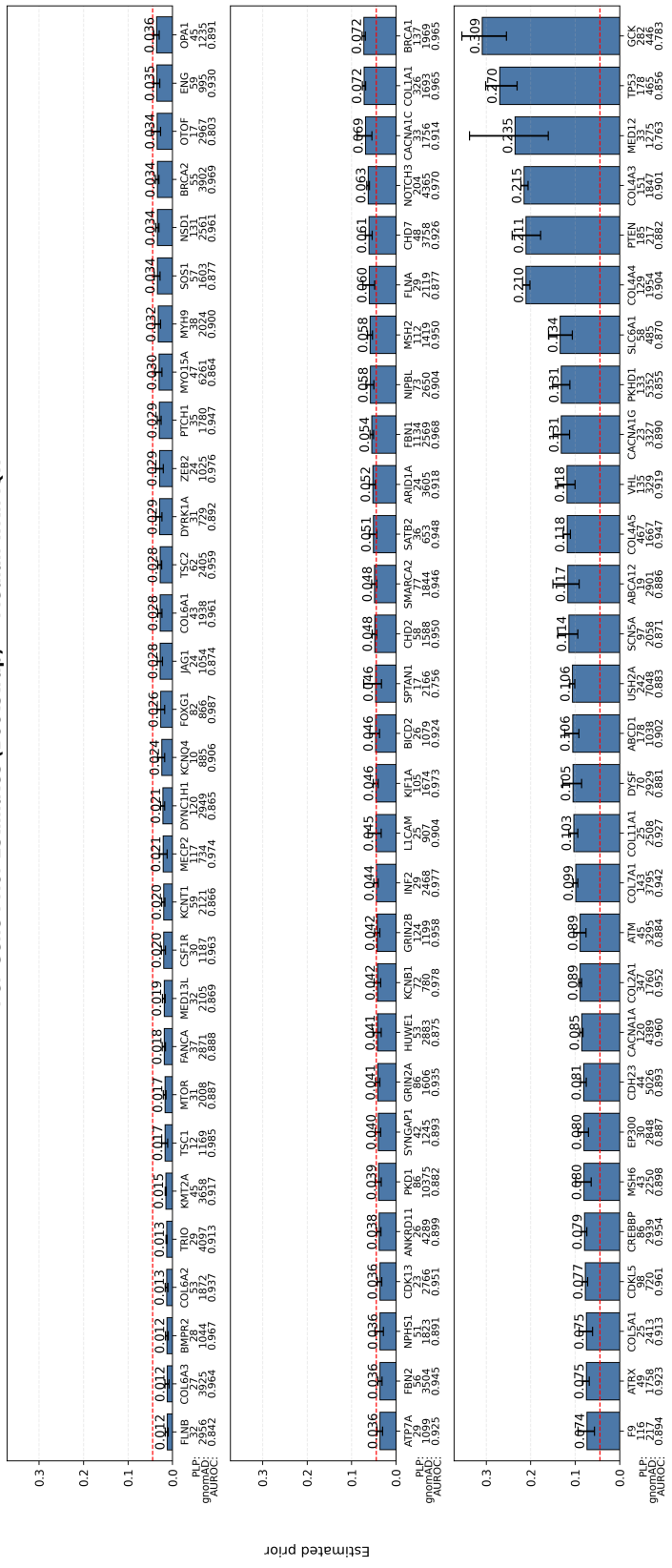



Figure 3.7: Caption continued on next page.

**Figure 3.7.** Gene-specific prior probability of pathogenicity estimated using the **gnomAD mixture set** (**MutPred2 training variants not filtered**). Bars show the **median prior** from 5000 bootstrap samples per gene, with **error bars** corresponding to the **25th–75th percentile (bootstrap IQR)**. The **red dashed line** indicates the genome-wide prior of **4.41%** (Pejaver et al., 2022). **Annotations below each bar show the calibration input:** (i) the number of pathogenic labeled variants used (*PLP count*), (ii) the number of unlabeled variants from gnomAD used as mixture samples, and (iii) the classifier performance (*AUROC*) during prior estimation.

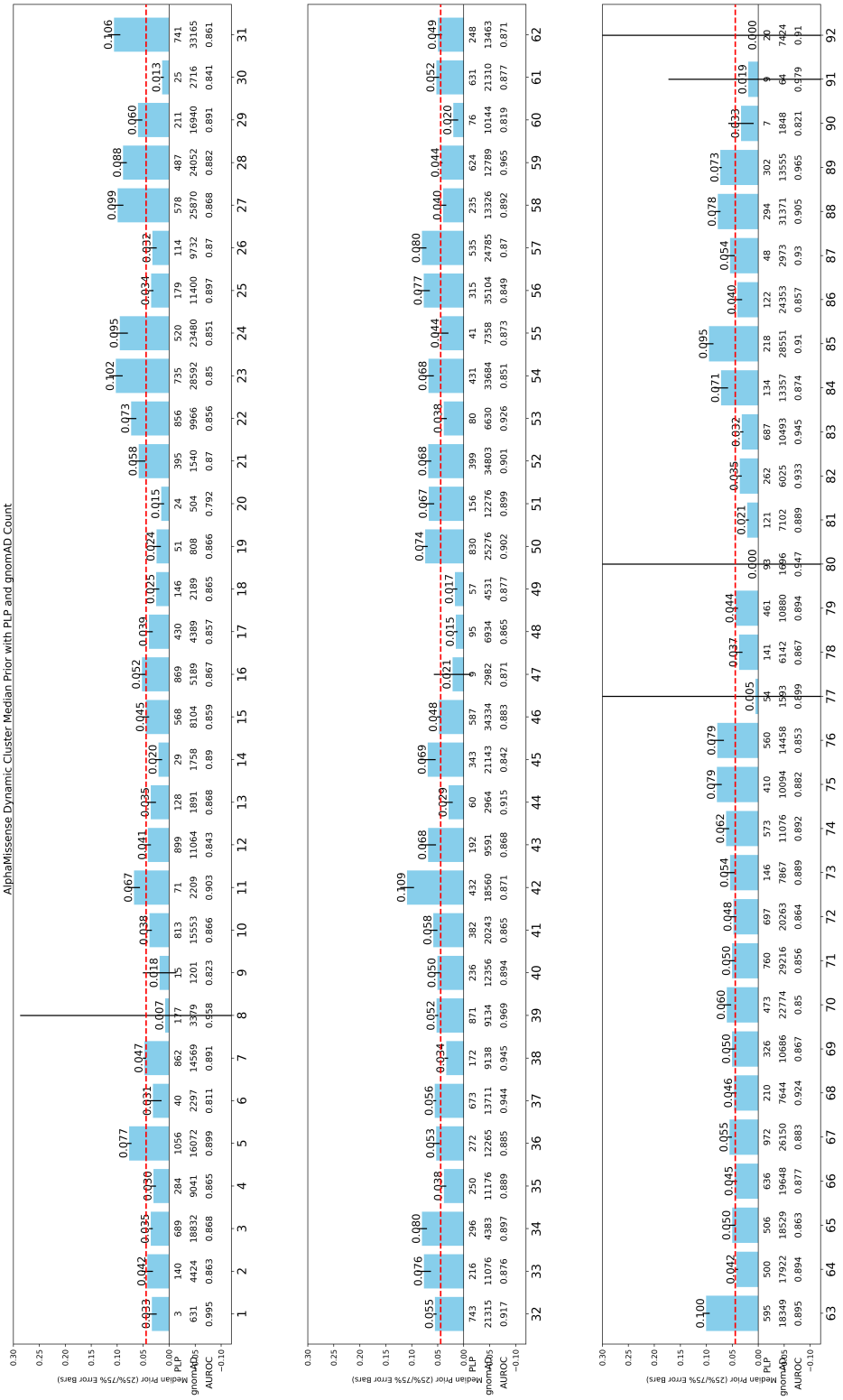



Figure 3.8: Caption continued on next page.

**Figure 3.8.** Prior probability of pathogenicity estimated at the **domain-cluster level** using **AlphaMissense**-based similarity of variant score distributions. Pfam domains with similar AlphaMissense score distributions were grouped into clusters, and a prior was learned **per cluster**, rather than per gene. Bars represent the **median estimated prior** across 5000 bootstrap samples, with **IQR (25th–75th percentile)** shown as error bars. The **red dashed line** marks the genome-wide prior of **4.41%** (Pejaver et al., 2022). **Annotations beneath each cluster** indicate: (i) the number of pathogenic labeled variants (*PLP count*), (ii) the number of unlabeled gnomAD variants used as mixture samples, and (iii) the classifier performance (*AUROC*) during prior estimation.

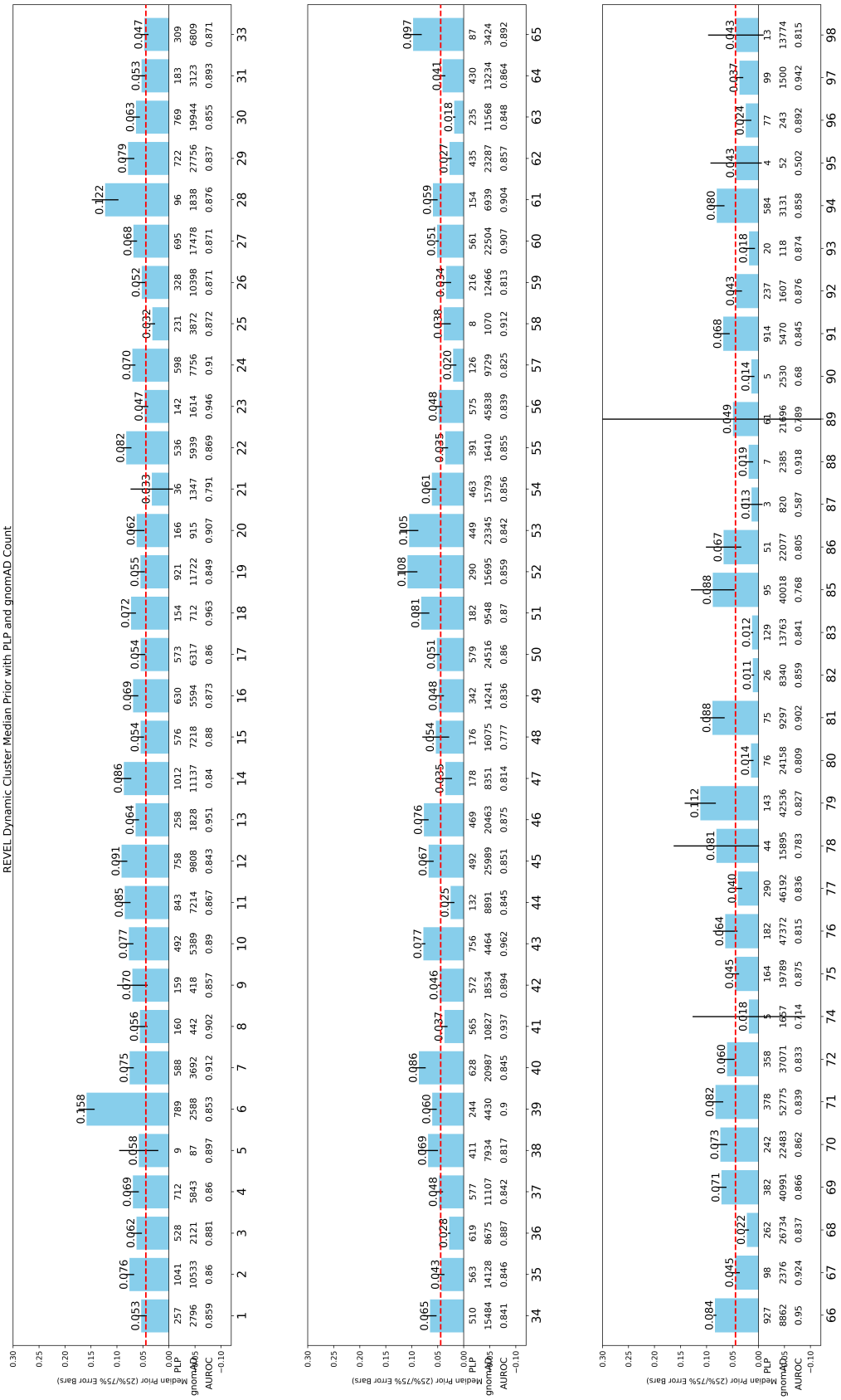



Figure 3.9: Caption continued on next page.

**Figure 3.9.** Prior probability of pathogenicity estimated at the **domain-cluster level** using **REVEL**-based similarity of variant score distributions. Pfam domains with similar REVEL score distributions were grouped into clusters, and a prior was learned **per cluster**, rather than per gene. Bars represent the **median estimated prior** across 5000 bootstrap samples, with **IQR (25th–75th percentile)** shown as error bars. The **red dashed line** marks the genome-wide prior of **4.41%** (Pejaver et al., 2022). **Annotations beneath each cluster** indicate: (i) the number of pathogenic labeled variants (*PLP count*), (ii) the number of unlabeled gnomAD variants used as mixture samples, and (iii) the classifier performance (*AUROC*) during prior estimation.

# Chapter 4

# Gene/Domain-Aware Calibration of Variant Effect Predictors

## 4.1 Introduction

Computational variant effect predictors (VEPs) have become essential tools in clinical genomics, providing rapid, scalable estimates of a variant's potential to disrupt protein function or cause disease. Numerous machine-learning predictors (e.g. REVEL, MutPred2, AlphaMissense) output continuous scores that ostensibly reflect pathogenicity for missense variants. However, interpreting these raw scores in a clinical context is non-trivial, because they are not naturally calibrated to the *posterior probabilities of pathogenicity* required by the ACMG/AMP Bayesian framework for PP3/BP4 evidence.

As discussed in the previous chapter, the ACMG/AMP system treats each line of evidence—including computational predictions—as a weighted contribution to a Bayesian posterior that ultimately determines whether a variant is classified as pathogenic, benign, or of uncertain significance. Early versions of the guidelines allowed in silico predictors to contribute only supporting-level evidence, reflecting concerns about accuracy, calibration, and reproducibility. Subsequent work has demonstrated that, when appropriately calibrated, some VEPs can contribute moderate or even strong evidence of being pathogenic or benign at specific score thresholds (Pejaver, 2022; Bergquist et al., 2024). The analyses aggregated ClinVar pathogenic and benign variants across many genes, estimated local posterior probabilities as a function of score, and then mapped these to likelihood ratios (LRs) corresponding to ACMG/AMP evidence strengths. This "universal" calibration framework has since been extended by ClinGen SVI efforts to additional predictors and has motivated updates to PP3/BP4 usage recommendations.

However, genome-wide calibration assumes that all genes share the same score–pathogenicity relationship, an assumption that is increasingly known to be false. Several studies have shown that the distribution and interpretability of VEP scores vary substantially across genes and disease contexts (Tejura et al., 2024; Dias et al., 2024; Isakov et al., 2024). Tejura et al. (2024), for example, demonstrated that thresholds derived from genome-wide calibration can be systematically misaligned for individual genes, leading to over- or underestimation of evidence strength. Other gene- or disease-specific tools—such as CardioClassifier for cardiomyopathy genes (Whiffin et al., 2018) and Bayesian models combining gene-specific priors with variant-level features (Ruklisa et al., 2015)—have illustrated that tailored, gene-aware calibration can substantially reduce classification error. Yet, these approaches are labor-intensive, depend heavily on expert curation, and do not scale easily to thousands of disease-associated genes.

Parallel challenges arise for functional assay data. Multiplexed assays of variant effect (MAVEs) produce quantitative functional scores for hundreds to thousands of variants per gene, yet these scores must still be translated into binary or categorical pathogenicity assessments within the ACMG/AMP framework. Early applications relied on hand-chosen functional cutoffs, which may not reflect calibrated probabilities of pathogenicity. Recent methods have moved toward probabilistic calibration of functional scores, including mixture-model–based approaches for mapping continuous functional readouts to posterior probabilities and evidence strengths (e.g. Zeiberg et al., 2025). In particular, Badonyi and Marsh (2025) introduced `acmgscaler`, an R package and Colab workflow that performs standardized *gene-level* calibration of both MAVE-derived functional scores and computational scores such as AlphaMissense. Their framework learns calibration parameters separately for each gene or assay, enabling consistent ACMG/AMP evidence assignments while accounting for gene-specific score distributions. `acmgscaler` thus represents an important step toward routine, gene-aware calibration of both experimental and computational variant effect scores.

More broadly, the problem of turning model outputs into reliable probabilities has been extensively studied in machine learning as *probability calibration*. Post-hoc calibration methods—applied after a model is trained—adjust output scores so that they better approximate true event probabilities. Common approaches include parametric scalings such as logistic (Platt et al., 1999) calibration and beta calibration (Kull et al., 2017), as well as non-parametric methods such as isotonic regression (Barlow and Brunk, 1972). Comparative work has shown that simple regression-based methods often provide strong performance across a wide range of conditions, but that no single technique dominates in all data regimes (e.g. Wang, 2023). Importantly, these studies emphasize that high discrimination (e.g. AUROC) does not guarantee good calibration, and that over-confident probabilities can be particularly problematic

in safety-critical settings such as medicine.

In this chapter, we build directly on the gene-specific priors developed in the previous chapter to address the problem of *gene-specific calibration of VEP scores*. Rather than re-estimating priors, we treat those gene- and cluster-level priors as fixed inputs and focus on how best to map each predictor's raw scores to posterior probabilities and ACMG/AMP evidence strengths on a per-gene (or per-domain-cluster) basis. Our overarching goal is to understand which calibration strategies work best under realistic gene-level data constraints (limited labeled variants, class imbalance, and imperfect priors), and to design a workflow that can automatically select an appropriate method for each gene.

Concretely, this chapter makes the following contributions:

- **Systematic evaluation of local posterior calibration under gene-like regimes.** We extend the local posterior approach of Pejaver (2022) to the gene-specific setting and systematically explore how its performance depends on window size, the use of unlabeled gnomAD variants, sample size, and class imbalance. Using simulation studies parameterized by real gene score distributions and the priors from Chapter 3, we quantify when local calibration is robust and when it fails.

- **Benchmarking of ten post-hoc calibration methods.** We compare local posterior calibration against a suite of ten established post-hoc methods across diverse simulated gene scenarios. Evaluation focuses on ACMG-relevant metrics, including misestimation of evidence strength in PP3 and BP4 regions, thereby identifying which methods are best suited to sparse and imbalanced gene-level data.

- **Dynamic gene-specific method selection.** Motivated by the absence of a universally optimal method, we design a dynamic decision-tree workflow that selects a calibration strategy on a per-gene basis. The decision tree uses simple diagnostics (e.g. number of labeled variants, degree of class imbalance, shape of score distributions) to choose among candidate methods, prioritizing those that minimize clinically risky overestimation of evidence.

- **Extension to domain-based clusters.** To handle genes with insufficient labeled variants for reliable single-gene calibration, we leverage the domain-based clustering defined in earlier work. For each cluster of domains with similar score distributions, we estimate a shared prior (using the framework from Chapter 3) and apply the dynamic calibration workflow at the cluster level. This allows genes with sparse data to "borrow strength" from related domains while still yielding calibrated evidence thresholds.

- **Application to real clinical datasets.** Finally, we apply our gene- and cluster-specific calibration pipeline to ClinVar. We show that the resulting calibrated thresholds provide more consistent and interpretable ACMG/AMP PP3/BP4 evidence across genes than either uncalibrated scores or genome-wide calibration, and we highlight genes where calibration substantially changes the strength or direction of computational evidence.

Together, these analyses operationalize the gene- and cluster-specific priors from Chapter 3 into a practical, scalable framework for calibrated variant interpretation. By explicitly accounting for gene context, data sparsity, and method-specific limitations, this chapter moves computational evidence closer to a rigorously quantified and clinically reliable component of the ACMG/AMP variant classification workflow.

## 4.2 Methods

### 4.2.1 Datasets

**ClinVar Data** For gene-specific calibration, we utilized labeled variants from the January 2025 release (version 2025.01) of ClinVar. This dataset includes variants annotated as pathogenic (P), likely pathogenic (LP), benign (B), or likely benign (LB). The main tab-delimited file containing all ClinVar variants was downloaded from the ClinVar FTP site as part of their regular monthly release.

We restricted the analysis to missense variants with a review status of at least one star (i.e., with assertion criteria provided by at least one submitter or expert panel). To focus on rare variation, we retained only those with a global allele frequency (AF) below 0.01 in the Genome Aggregation Database (gnomAD v4) Gudmundsson et al. (2022), prioritizing exome AF when available, and falling back to genome AF otherwise.

To minimize confounders of splice-altering effects, variants with SpliceAI-predicted scores $\geq 0.2$ were excluded (Jaganathan et al., 2019); variants without a SpliceAI prediction were retained. Furthermore, to avoid evaluation bias, we excluded any variants known to be part of the training datasets of the prediction tools under evaluation (e.g., REVEL, MutPred2), when such information was available. The final number of qualifying variants per gene and predictor is summarized in the Supplementary Table (chapter4_figures/gene_predictor_varcnts.csv).

**gnomAD Data** To estimate gene-specific prior probabilities of pathogenicity, we used variant data from gnomAD v4.0, including both exome and genome datasets. VCF files were downloaded from the official gnomAD release site. Filtering steps matched those used for

the ClinVar dataset: only rare missense variants with global AF $< 0.01$ and SpliceAI scores $< 0.2$ were retained. As with ClinVar, we excluded any variants known to appear in the training sets of the evaluated prediction tools.

**Prediction Score Mapping** We used the Ensembl Variant Effect Predictor (VEP) to annotate missense variants with prediction scores from AlphaMissense, REVEL, and SpliceAI. MutPred2 scores were obtained directly from the tool's developer. Only variants mapped to the MANE Select transcript were retained for calibration to ensure consistency across annotations. Genes labeled by GenCC (Genetics Curation Consortium) (DiStefano et al.) as having "moderate" or stronger clinical validity were aggregated to define baseline prediction score distributions for simulation and clustering analyses.

### 4.2.2 Simulation Framework

To evaluate the robustness and accuracy of calibration methods under diverse conditions, we constructed a simulation framework based on synthetic variant score datasets.

- **Generation of Synthetic Datasets.**

  We modeled variant effect predictor scores using four parametric distribution families commonly observed in real data: *Beta*, *truncated Normal*, *truncated Skew-t*, and *truncated Skew-Cauchy*. Together, these families capture a wide range of behaviors (bounded support, symmetry, skewness, and heavy tails) typical of real prediction score distributions.

  Distributional forms and parameters were informed by clustering analyses of score distributions from clinically valid genes (GenCC). For each cluster and predictor, pathogenic and benign scores were independently fitted to the four candidate families, and the best-fitting model was selected by maximum log-likelihood. These fitted models were then used to generate synthetic pathogenic and benign score sets, providing realistic yet controlled scenarios for benchmarking calibration methods.

- **Calibration Set.** For each fitted pathogenic/benign score distribution, we generated calibration sets of varying sizes to assess sample efficiency:

$$n_{\text{cal}} \in \{25, 50, 100, 300, 500, 1000\}.$$

  To examine robustness to class imbalance, we varied the observed proportion of pathogenic

variants in the calibration set,

$$\pi_{\text{cal}} = \Pr(Y = 1 \mid \text{calibration}) \in \{0.05, 0.10, 0.20, \ldots, 0.90, 0.95\}.$$

- **Test Set.** For each scenario, we generated an independent test set of 1,000 variants. The true pathogenic prior probability was varied to reflect different gene-specific priors:

$$\alpha = \Pr(Y = 1 \mid \text{test}) \in \{0.01, 0.02, \ldots, 0.10, 0.20, 0.30, 0.40, 0.50\}.$$

- **Prior Adjustment.** Most post-hoc calibration methods (e.g., Beta, Platt, SplineCalib, Isotonic, SmoothIsotonic, BetaMixture, TruncNorm) implicitly assume that the class prior is the same in the calibration and test sets. To account for prior shifts ($\pi_{\text{cal}} \neq \alpha$), we applied a post-hoc correction to the calibrated posteriors at test time. Let $\rho(x)$ denote the calibrated posterior from the calibration set and $\alpha'$ the pathogenic fraction in the calibration set (i.e., $\alpha' = \pi_{\text{cal}}$). We first compute the adjusted odds

$$\gamma(x) = \frac{\alpha}{1-\alpha} \cdot \frac{1-\alpha'}{\alpha'} \cdot \frac{\rho(x)}{1-\rho(x)},$$

and then obtain the prior-corrected posterior

$$p(Y = 1 \mid x) = \frac{\gamma(x)}{1+\gamma(x)}.$$

- **Evaluation Metrics.**

To assess calibration quality in clinically meaningful terms, we quantified the discrepancy between calibrated and true posteriors in units of ACMG/AMP evidence strength. Specifically, we expressed the difference in posterior odds as an equivalent number of ACMG "evidence points," corresponding to Supporting-level increments.

For each test point, the misestimation in ACMG evidence points is defined as:

$$f - x = \frac{\log\left(\dfrac{lr_f^+}{lr_x^+}\right)}{\log\left(lr_{\text{Su}}^+\right)},$$

where:

  - $P_f$ is the calibrated posterior probability,
  - $P_x$ is the true posterior probability,

- $lr_f^+ = \dfrac{P_f}{1-P_f}$ is the calibrated posterior odds,
- $lr_x^+ = \dfrac{P_x}{1-P_x}$ is the true posterior odds,
- $lr_{\text{Su}}^+$ is the likelihood ratio corresponding to one Supporting-level ACMG evidence point.

When calibration is accurate, $f - x \approx 0$. Positive values indicate overestimation of pathogenic evidence (calibrated odds too high), whereas negative values indicate underestimation. When both calibrated and true posteriors lie in the highest ACMG strength category (e.g., PP3-Strong), the difference is considered as none, preserving consistent clinical interpretation. Averaging $|f - x|$ across all test samples provides a global, clinically aligned summary of calibration performance.

To focus on clinically critical regions, we also defined two regional metrics that quantify how often posteriors are misestimated by at least one ACMG point in the PP3- and BP4-supporting regions.

**PP3 Misestimation Fraction:**

$$\text{PP3-Fraction} = \frac{\left|\left\{ x \in \mathcal{X}_{\text{test}} \;\middle|\; \underbrace{\text{True}(x) > \tau_{\text{PP3}}}_{\text{PP3-supporting region}} \;\wedge\; \underbrace{(f-x) \geq 1}_{\text{Overestimation}\geq 1 \text{ point}} \right\}\right|}{|\{x \in \mathcal{X}_{\text{test}} \,|\, \text{True}(x) > \tau_{\text{PP3}} \;\wedge\; (f-x) \in \mathbb{R}_{\text{finite}}\}|}. \tag{4.1}$$

PP3-Fraction measures, among variants whose *true* posterior lies in the PP3-supporting region, the proportion for which calibration *overstates* pathogenic evidence by at least one ACMG point. This highlights methods that are overconfident in the pathogenic range.

**BP4 Misestimation Fraction:**

$$\text{BP4-Fraction} = \frac{\left|\left\{ x \in \mathcal{X}_{\text{test}} \;\middle|\; \underbrace{\text{True}(x) < \tau_{\text{BP4}}}_{\text{BP4-supporting region}} \;\wedge\; \underbrace{(x-f) \geq 1}_{\text{Benign overestimation}\geq 1 \text{ point}} \right\}\right|}{|\{x \in \mathcal{X}_{\text{test}} \,|\, \text{True}(x) < \tau_{\text{BP4}} \;\wedge\; (x-f) \in \mathbb{R}_{\text{finite}}\}|}. \tag{4.2}$$

BP4-Fraction is defined analogously on the benign side: among variants whose *true* posterior lies in the BP4-supporting region, it quantifies the proportion for which calibration overstates benign evidence (or equivalently, underestimates pathogenicity) by at least one ACMG point.

where:

- $\tau_{\text{PP3}}$: posterior threshold for PP3-supporting strength,
- $\tau_{\text{BP4}}$: posterior threshold for BP4-supporting strength,
- True($x$): true posterior probability for test point $x$,
- $\mathbb{R}_{\text{finite}}$: set of finite (non-infinite, non-NaN) values.

### 4.2.3 Local Posterior Calibration Framework

To systematically evaluate the performance of the local posterior calibration method, we performed a sensitivity analysis over its two main hyperparameters. Our goals were: (1) to quantify how different parameter settings affect calibration accuracy, and (2) to assess whether local calibration remains reliable under realistic, gene-level data constraints (e.g., small sample size and class imbalance).

The two key parameters investigated were:

- **Window size:** the proportion of labeled pathogenic and benign variants included in each sliding window used to estimate the local posterior. We tested window sizes of 10%, 20%, and 30% of the total labeled dataset.
- **gnomAD fraction:** the minimum proportion of unlabeled gnomAD variants required within each window to stabilize the local posterior estimate. We evaluated values of 0%, 3%, and 6% (relative to the total gnomAD set).

All nine parameter combinations were evaluated using the ACMG-based misestimation metrics described above, with particular emphasis on PP3-Fraction and BP4-Fraction, which capture overestimation errors in clinically relevant PP3- and BP4-supporting regions. This analysis enabled us to characterize the robustness of the local calibration method and to identify hyperparameter settings that minimize clinically meaningful miscalibration while maintaining overall accuracy.

### 4.2.4 Comparison with Existing Calibration Methods

To benchmark the performance of the local posterior calibration method, we compared it against a suite of ten established or widely used post-hoc calibration techniques. This comparison was designed to assess robustness and reliability under a range of practical conditions, including low-data settings and class imbalance—challenges commonly encountered in gene-level variant interpretation.

- **Overview of calibration methods.** We evaluated the following ten post-hoc calibration methods:
  - **Platt scaling** (Platt et al., 1999): Models the relationship between the raw prediction score $z$ and the calibrated probability $P$ using a logistic function.
  - **Weighted Platt scaling**: A variant of Platt scaling that incorporates class weights to mitigate imbalance between pathogenic and benign examples during calibration.
  - **Beta calibration** (Kull et al., 2017): Extends Platt scaling with a three-parameter beta-based model, providing additional flexibility for non-linear score–probability relationships.
  - **SplineCalib** (Lucena, 2018): Uses cubic smoothing splines to model a flexible, smooth, non-linear mapping between $z$ and $P$.
  - **Isotonic regression** (Barlow and Brunk, 1972): A non-parametric, monotonic calibration method that fits a stepwise, non-decreasing function, preserving the ranking of prediction scores.
  - **Smoothed isotonic regression** (Jiang et al., 2011): Smooths the isotonic fit using a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) to produce a continuous calibration curve.
  - **MonoPostNN**: A neural-network-based, monotonic calibrator developed within our group that learns a flexible non-parametric mapping from $z$ to $P$ while enforcing monotonicity constraints. An open-source implementation is available at `https://github.com/shajain/PosteriorCalibration`.
  - **Skew-normal mixture** (Zeiberg et al., 2025): Models the score distribution using a two-component mixture of skew-normal distributions, allowing calibration to capture asymmetric and multimodal behavior.
  - **Beta mixture**: Uses a mixture of beta distributions to flexibly model bounded, potentially bimodal score distributions on the $[0, 1]$ interval.
  - **Truncated normal mixture**: Fits a mixture of truncated normal components to bounded scores, accommodating both multimodal and approximately symmetric patterns.
- **Unified evaluation protocol.** All methods were compared under a common evaluation framework. For each simulated data scenario (Section 4.2.2), we generated 30 independent datasets and applied every calibration method to quantify variability and robustness. The best-performing configuration of the local calibration method—defined

by its optimal combination of window size and gnomAD fraction—was included in all comparisons.

The evaluation was designed to identify whether any alternative method consistently outperforms local calibration, particularly under realistic gene-level constraints such as limited labeled data and strong class imbalance. A broader goal was to inform the design of dynamic or hybrid workflows that could leverage the strengths of multiple methods.

- **Performance metrics and ranking criteria.** Methods were ranked using the ACMG-based metrics described above:
  - mean absolute misestimation in ACMG evidence points across all test variants;
  - PP3-Fraction and mean overestimation within the PP3-supporting region;
  - BP4-Fraction and mean overestimation within the BP4-supporting region.

  These metrics emphasize clinically meaningful miscalibration, particularly in regions where PP3 and BP4 evidence codes are applied.

- **Analysis of failure cases.** Finally, we identified and examined scenarios in which individual calibration methods performed poorly, with particular attention to extreme class imbalance and severe data sparsity. These failure cases were used to diagnose method-specific limitations and to develop practical recommendations for method choice under constrained conditions.

### 4.2.5 Robustness via Subsampling

To evaluate the sample efficiency of each calibration method, we investigated how performance degrades when models are trained on subsets of the full labeled variant set.

- **Sample efficiency analysis.** We systematically varied both class balance and calibration set size:
  - **Observed pathogenic fraction (calibration prior):** $P_{\text{PLP}} \in \{0.1, 0.2, \ldots, 0.9\}$,
  - **Number of labeled samples:** $n \in \{25, 50, 100, 200, 500, 1000, 5000\}$,
  - **Replicates:** 30 independent subsampling replicates per $(P_{\text{PLP}}, n)$ combination,
  - **Fixed test-set prior:** $\alpha = 0.0441$ (4.41%), applied uniformly across all test evaluations.

For calibration methods that do not inherently accommodate prior-shift between calibration and test sets, we applied the post-hoc prior adjustment described in the simulation framework (Section 4.2.2). In all cases, calibrated posteriors were evaluated on a common test set so that subsampled calibrations could be directly compared against full-data performance.

- **Evaluation strategies.** Two complementary evaluation strategies were used:
  - **Reference-set evaluation.** As an external benchmark, we reproduced a local calibration model on REVEL scores using 100 labeled variants and a gnomAD smoothing fraction of 0.03, following the configuration of Pejaver et al. (2022). This served as a fixed reference to gauge how quickly different methods approach (or deviate from) a practical, previously recommended setting as sample size increases.
  - **Full vs. subsampled comparison.** For each calibration method, we fit a model on the full labeled calibration set (per scenario) and treated its performance as the best-available reference. We then fit the same method on each subsampled calibration set and quantified performance degradation relative to the full-data reference. This allowed us to separate sensitivity to sample size from intrinsic method quality.
- **Evaluation metrics.** Performance under subsampling was quantified using the ACMG-based metrics introduced above:
  - mean absolute misestimation in ACMG evidence points across test variants;
  - PP3-Fraction and BP4-Fraction: the proportion of variants in the PP3- or BP4-supporting regions whose evidence is misestimated by more than 1 ACMG point.

  Together, these metrics capture both global calibration error and clinically salient overestimation in regions where PP3/BP4 criteria are applied.

### 4.2.6 Dynamic Decision Tree Workflow

Previous experiments demonstrated that no single calibration method consistently outperformed others across all gene-specific scenarios. To address this variability, we developed a **dynamic decision tree workflow** that adaptively selects the optimal calibration method for each gene, based on its score distribution and data availability (Figure 4.1).

**Step 1: Score distribution modeling** For each gene, we model the score distributions of pathogenic and benign variants separately using four candidate parametric families:

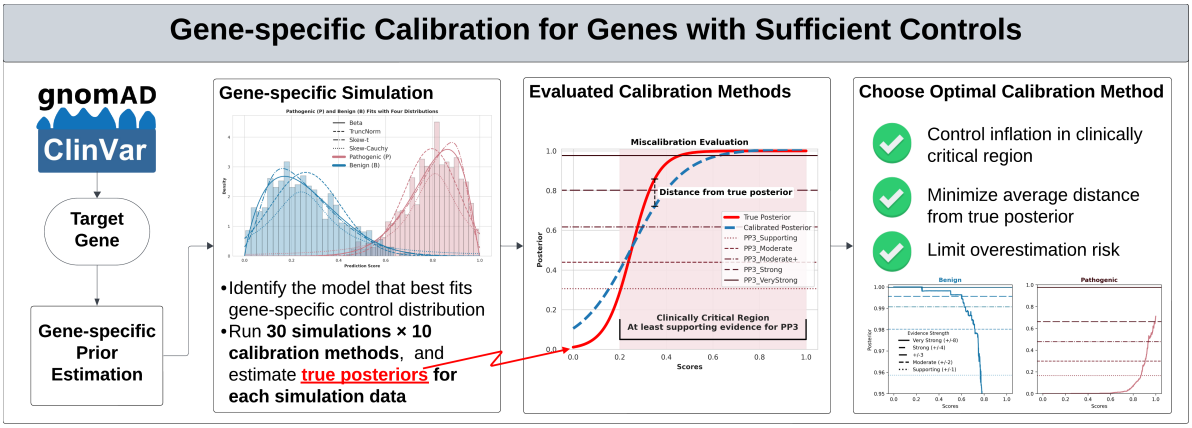



Figure 4.1: Dynamic decision tree workflow for selecting the optimal gene-specific calibration method. For each gene, fitted score distributions are used to simulate data, benchmark alternative calibration strategies, and select the method that balances low misestimation with minimal over-calling of evidence strength.

*Beta*, *Truncated Normal*, *Truncated Skew-t*, and *Truncated Skew-Cauchy*. The best-fitting distribution for each class is selected via maximum log-likelihood estimation.

**Step 2: Simulation and method evaluation** Using the fitted distributions, we simulate 30 synthetic datasets per gene and compute the corresponding true posterior probabilities. For each gene, we then evaluate:

- 9 local calibration configurations (combinations of window size and gnomAD smoothing fraction), and
- 10 alternative post-hoc calibration methods.

Each method is assessed with 1,000 bootstrap iterations. Calibration thresholds are estimated from out-of-bag samples, and we use the 5th percentile of posterior estimates to reduce the risk of overestimation, especially in distribution tails.

**Step 3: Multi-stage method selection** To select a single calibration method per gene, we apply a three-stage filtering procedure:

1. **Overestimation risk filter:** Discard methods that, in at least 25% of simulation replicates, exhibit $\geq 2$ ACMG-point error at the 75th percentile or $\geq 3$ points maximum in PP3- or BP4-supporting regions. This step removes approaches prone to clinically unsafe over-calling of evidence strength.
2. **Average misestimation ranking:** Among the remaining methods, compute the average ACMG-point misestimation across pathogenic, benign, and indeterminate regions (using bootstrap estimates for the former two and point estimates for the latter). Retain the three methods with the lowest overall misestimation.

3. **Overestimation frequency filter:** From these top three, select the method with the lowest combined PP3- and BP4-Fraction, i.e., the fewest instances of $\geq$ 1-point overestimation in key evidence regions.

This multi-stage strategy ensures that the final selected method is not only accurate on average but also conservative, minimizing high-risk miscalibration in clinically important score intervals for each gene.

### 4.2.7 Real-World Validation

To assess the practical utility of our gene-specific calibration framework, we conducted a series of real-world validation experiments using ClinVar, Clinical Genome Resource (ClinGen), and Variant Curation Expert Panel (VCEP) annotations.

1. **Interval-Based Likelihood Ratio Evaluation (Evidence Strength Accuracy)** To evaluate calibration performance at the level of ACMG/AMP evidence strengths, we used the Bayesian likelihood-ratio framework proposed by Tavtigian et al. (2020), which defines theoretical $LR^+$ thresholds for each evidence level (Supporting, Moderate, Moderate+, Strong) given a prior probability of pathogenicity. For each gene, we estimated the prior from ClinVar pathogenic (P/LP) and benign (B/LB) variants and used this prior to compute the expected $LR^+$ threshold for each evidence category. We then computed empirical $LR^+$ values for the calibrated score intervals generated by the two approaches compared in this study: gene-specific thresholds (our method) and the previously published aggregated thresholds (Pejaver et al., 2022; Bergquist et al., 2024). For each evidence level within a gene, we assessed which method performed better ("wins") according to the following rules: for pathogenic evidence (PP3), the method with the higher $LR^+$ was deemed superior; for benign evidence (BP4), lower $LR^+$ indicated stronger rule-out capability. If only one method achieved the theoretical threshold, it was designated as the winner; if both achieved or both failed to reach the threshold, the method whose $LR^+$ was closest to the theoretical value was counted as the winner. For each gene, we counted the number of evidence levels won by each method. To determine whether gene-specific calibration systematically outperformed the aggregated approach across genes, we applied a one-sided binomial test, treating the null hypothesis as both methods being equally likely to win ($p = 0.5$), where K represents the number of genes where gene-specific calibration won more intervals than the aggregated method, and N represents the total number of evaluable genes.

2. **Classification Consistency with ClinVar (Outcome-Level Validation)** Using

the final calibrated thresholds selected by the dynamic decision-tree framework, we compared the resulting variant classifications against those produced using the published gene aggregated thresholds. Two complementary analyses were performed:

- TPR–FPR difference (scatter plot): For P/LP and B/LB ClinVar variants, we calculated the mean difference in true-positive rate (TPR) and false-positive rate (FPR) between gene-specific calibration and aggregated calibration thresholds (PP3/BP4 Supporting or stronger).
- Evidence assignment distribution (heatmap): For each gene and predictor, we quantified the percentage of variants assigned to each evidence point category (Supporting, Moderate, ±3, Strong) for: ClinVar P/LP, B/LB, and VUS variants, gnomAD variants, and all possible missense SNVs. This allows us to assess: whether gene-specific calibration reduces the proportion of variants remaining in the indeterminate region (IR) and whether calibration prevents misassignment to overly strong or inappropriate evidence strengths.

Together, these analyses evaluate both the statistical correctness of calibration (interval $LR^+$ alignment) and the practical impact on classification decisions (TPR/FPR shifts, evidence reassignment patterns).

3. **ClinGen Annotation–Based Reclassification Analysis** For 16 genes with curated variant classifications from ClinGen Variant Curation Expert Panels (VCEPs), we evaluated how the use of gene-specific calibration influences computational evidence assignment and final clinical interpretation. Variant-level prediction scores from REVEL, AlphaMissense, and MutPred2 were retrieved, and each variant was reclassified twice—first using the published aggregated calibration thresholds (Pejaver et al. 2022; Bergquist et al. 2025) and then using the gene-specific thresholds selected by our dynamic decision-tree framework.

   For each calibration scheme, we reassigned the PP3/BP4 computational evidence code in the ClinGen evidence list and then recomputed the overall ACMG/AMP classification by combining the recalibrated PP3/BP4 code with the remaining, unchanged ClinGen evidence ("Met Codes"). We quantified the extent and direction of reclassification, including shifts such as VUS→B/LB or VUS→P/LP, relative to both the original ClinGen classification and the gene aggregate baseline; these transitions were visualized using a Sankey plot. To isolate the contribution of computational evidence independent of ClinGen curation decisions, we additionally removed the PP3/BP4 code from the original ClinGen evidence list and recalculated clinical significance us-

ing only the non-computational evidence. We compared this baseline classification with (i) ClinGen-provided PP3/BP4 codes, (ii) gene aggregated thresholds, and (iii) single gene thresholds. For each method, we evaluated per-gene accuracy using only variants assigned to a non-indeterminate evidence level, and we compared how many variants each calibration scheme was able to assign PP3/BP4 evidence to at all. This analysis allowed us to determine whether single calibration improves classification accuracy, increases the number of variants leaving the indeterminate region, and reduces inappropriate assignment of computational evidence compared to the aggregate approach.

### 4.2.8 Cross-Predictor Consistency of Gene-Level Calibration

To evaluate the consistency of calibrated computational evidence across predictors, we quantified gene-level agreement in the ACMG/AMP evidence points assigned by REVEL, AlphaMissense (AM), and MutPred2 (MP2). For each gene with sufficient labeled variants to support single-gene calibration, we assigned a *signed evidence point value* to every variant based on its calibrated PP3/BP4 classification and its true ClinVar label.

Calibrated evidence strengths (PP3 or BP4 Supporting, Moderate, Moderate+, or Strong) were mapped to integer evidence-point magnitudes using the conventional ACMG/AMP scale, where Supporting, Moderate, Moderate+, and Strong correspond to 1, 2, 3, and 4 points, respectively, and the indeterminate region (IR) corresponds to 0 points. To incorporate directionality, we assigned the sign of each evidence point according to the variant's true label. For pathogenic (P/LP) variants, PP3 categories were treated as positive evidence and BP4 categories as negative evidence; for benign (B/LB) variants, this sign convention was reversed, such that PP3 categories contributed negative points and BP4 categories contributed positive points. This formulation yields a continuous, symmetric measure in which positive values represent correct directional support and negative values represent incorrect or misleading support.

For each gene and predictor, we computed the mean of these signed evidence points across all labeled variants, yielding a single *gene-level average evidence score* that reflects the net computational support provided after calibration. These gene-level averages were compared across predictors using boxplots to examine distributional differences in overall evidence strength, and using pairwise scatterplots (REVEL vs. AM, REVEL vs. MP2, and AM vs. MP2) to evaluate agreement in gene-level computational evidence across models.

To contextualize these calibration-derived measures with respect to inherent predictive performance, we also computed gene-level AUROC values for each predictor using ClinVar-

labeled pathogenic and benign variants. AUROC distributions were summarized using boxplots, and AUROC values were compared with the calibrated gene-level evidence scores via scatterplots. Pearson correlation coefficients were computed to quantify the relationship between uncalibrated discriminative ability and the magnitude of calibrated evidence assigned at the gene level. This analysis provides insight into whether predictors that better separate pathogenic from benign variants before calibration also assign stronger calibrated PP3/BP4 evidence afterward.

### 4.2.9 Selecting the Best Predictor per Gene

To evaluate whether calibration performance could be improved by allowing different predictors to operate on different genes, we implemented a per-gene model selection strategy in which the computational predictor with the highest discriminative performance (as measured by gene-level AUROC) was selected for that gene. For each gene with sufficient labeled pathogenic and benign variants, we computed the AUROC for REVEL, AlphaMissense (AM), and MutPred2 (MP2). The predictor achieving the highest AUROC for that gene was designated as the *best predictor*, and the corresponding single-gene calibration output was used as the final calibrated model for that gene.

To assess the impact of this strategy, we compared three approaches: (i) the aggregated genome-wide calibration methods for each predictor separately; (ii) single-gene calibration applied uniformly using REVEL, AM, or MP2 across all genes; and (iii) the mixed-predictor strategy in which each gene is assigned the predictor that achieves the highest AUROC.

For all approaches, performance was quantified using the average difference between true-positive rate and false-positive rate (AvgTPR − AvgFPR), a summary metric that reflects net evidence discrimination while penalizing increases in false-positive support. For REVEL, AM, and MP2, we computed this metric separately at the genome-wide level (using aggregated thresholds) and at the gene-specific level. For the mixed-predictor analysis, we aggregated the per-gene values obtained from the selected best predictor.

## 4.3 Results

### 4.3.1 Local Calibration Fails with Small or Imbalanced Gene Datasets

We evaluated the robustness of gene-specific local posterior calibration by systematically analyzing how calibration performance is affected by varying sample size, class balance, and smoothing parameters. Using simulated datasets, we tested combinations of window size and gnomAD smoothing fraction across three representative calibration sample sizes (50,

100, and 150 labeled variants) and a range of observed pathogenic variant fractions (0.1 to 0.9). Figure 4.2 summarizes the misestimation errors, with error bars indicating variation across different true prior probabilities.

Local calibration exhibited poor performance when the number of labeled variants was below 50, with particularly unstable estimates in imbalanced datasets. Notably, fewer than half of the genes in our dataset met this sample size threshold, limiting the applicability of local calibration in many real-world scenarios. Even with 150 variants, calibration performance degraded substantially when the observed pathogenic fraction exceeded 0.8, with consistent overestimation of pathogenicity.

Across all conditions, calibration accuracy improved with larger window sizes (e.g., 30%) and greater gnomAD smoothing fractions, particularly in smaller sample settings. As the sample size increased, differences between parameter configurations diminished, suggesting that calibration becomes more stable and less sensitive to tuning when more data are available. The most reliable performance was observed when the sample size exceeded 100 and the observed pathogenic fraction was between 0.4 and 0.6, approximating class balance.

These results demonstrate that local posterior calibration is highly sensitive to sample size and label distribution. To ensure reliable performance, local calibration should be applied only to genes with sufficient labeled data and balanced class representation. In sparse or imbalanced settings, alternative calibration strategies may be necessary.

### 4.3.2 Alternative Methods Address Small size and sample imbalance issue

Having established the strengths and limitations of the local posterior calibration approach, particularly its diminished reliability under low-data conditions, we next investigated whether alternative calibration methods could offer improved performance. This analysis included a comparison between the best-performing local calibration configuration—identified by its optimal combination of window size and gnomAD variant fraction—and ten alternative calibration methods. To maintain clarity in visualization and focus the comparison on practically relevant approaches, we restricted our analysis to methods that achieved an average misestimation of less than one ACMG point in at least one evaluated scenario showed in Figure 4.3.

Several alternative methods outperformed local calibration when the calibration dataset contained fewer than 50 labeled variants—a regime in which local calibration exhibited high variance and frequent overestimation. This finding is particularly relevant given that a substantial fraction of ClinVar genes fall into this limited-data category. Despite their improved

performance in smaller datasets, all calibration methods—including local and alternative approaches—remained sensitive to class imbalance. Specifically, calibration was most stable and accurate when the observed pathogenic variant fraction ranged between 0.4 and 0.8. Outside this range, model performance deteriorated, often leading to increased misclassification risk.

Under balanced class distributions, many alternative methods maintained clinically acceptable misestimation rates, with fewer than 20% of variants exceeding one ACMG point of miscalibration in both the PP3- and BP4-supporting evidence regions. This result suggests that several methods can provide reliable calibration when applied under favorable data conditions. However, across all evaluated approaches, we observed considerable variability in misestimation across different true prior probabilities, as indicated by the error bars in Figure 4.3. These fluctuations reinforce the context-dependent nature of calibration performance and highlight the limitations of one-size-fits-all approaches.

Together, these findings support the utility of alternative calibration methods and motivate a more flexible and gene-specific framework, such as the dynamic decision tree, to systematically identify the optimal calibration strategy based on data availability and distributional characteristics.

### 4.3.3 Subsampling Robustness

To assess the stability of calibration performance under limited data availability, we conducted a comprehensive subsampling analysis in which calibration models were trained on subsets of varying size and class balance. Figure 4.4 summarizes the resulting trends across all evaluated methods and both evaluation strategies.

Across the full range of pathogenic fractions and calibration-set sizes, a clear relationship emerged between sample size and calibration accuracy. When the calibration set contained only 25 labeled variants, all methods exhibited substantial misestimation, with mean ACMG point error exceeding 1 under both the reference-set and full-vs.-subsampled evaluations.

Calibration error decreased steadily as calibration-set size increased. By 50 samples, several methods—including (weighted) Platt scaling, spline-based calibration (SplineCalib), Beta calibration, and MonoPostNN—began to capture the underlying score–pathogenicity relationship sufficiently well to approach the reference-level performance. Above 100 samples, more methods achieved stable and clinically acceptable calibration, with misestimation in ACMG evidence points dropping below 0.7 on average. These improvements were consistent across both evaluation strategies, indicating that the observed gains stem from intrinsic sample efficiency rather than artifacts of a particular benchmarking approach.

For the local posterior calibration strategy used in Pejaver et al. (2022), the subsampling analysis also highlights a practical lower bound on data requirements. In both panels of Figure 4.4, the local calibration method (in red) only achieves robust performance when at least 100 labeled variants are available, reflecting its reliance on fine-grained local density estimation. The full-vs.-subsampled comparison shows a similar pattern.

Overall, the robustness analysis demonstrates that sample size is a key determinant of calibration reliability. A minimum of approximately 50 labeled variants is needed for most global models to achieve stable and clinically meaningful calibration, whereas local calibration requires closer to 100 variants. The high concordance observed between the two evaluation strategies (Figure 4.4A–B) further supports the conclusion that these sample-size thresholds reflect intrinsic model behavior rather than evaluation-specific effects.

### 4.3.4 Interval-Based Likelihood Ratio Evaluation

To evaluate how accurately each calibration method recovered ACMG/AMP evidence strengths, we compared the interval-specific likelihood ratios ($LR^+$) produced by gene-specific calibration with those obtained from the aggregated calibration. Figure 4.5 illustrates this comparison for the gene *MSH2* under REVEL. For each PP3/BP4 evidence interval, we computed empirical $LR^+$ values under both calibration schemes and compared them with the theoretical $LR^+$ thresholds derived from the gene-specific prior following Tavtigian's Bayesian evidence model. The summary card annotates the method that better meets or exceeds each theoretical threshold, enabling a direct interval-by-interval assessment of calibration fidelity.

Across genes and predictors, gene-specific calibration consistently outperformed the aggregated method in recovering the correct evidence strengths. For REVEL, 67 genes satisfied the simulation quality criterion (25th percentile misestimation $< 1$), yielding 248 interval wins for the gene-specific method compared with 169 wins for aggregated calibration. Among the 66 evaluable genes, 48 exhibited superior performance under gene-specific calibration, with an average gain of 2.29 evidence-level wins per gene, whereas the 18 genes favoring the aggregated method showed an average deficit of 1.72 wins.

A similar pattern was observed for MutPred2. Among 83 qualifying genes, gene-specific calibration achieved 260 interval wins compared with 207 for the aggregated method. Of the 75 genes with complete interval evaluations, 50 favored the gene-specific approach, gaining on average 2.32 additional wins, while the remaining 25 genes showed an average deficit of 2.52 wins.

AlphaMissense also demonstrated robust improvements under gene-specific calibration. Across 66 genes meeting the quality criterion, gene-specific calibration produced 220 interval

wins compared with 182 for the aggregated method. Among 62 evaluable genes, 39 favored the gene-specific approach, achieving an average gain of 2.41 wins, whereas 23 genes favored the aggregated method, with an average deficit of 2.43 wins.

Taken together, these results show that gene-specific calibration more accurately recovers ACMG/AMP evidence strengths across predictors and across the majority of genes. The consistent excess of interval wins under gene-specific calibration indicates closer alignment with Bayesian theoretical thresholds and supports the conclusion that per-gene calibration improves the fidelity of PP3/BP4 evidence relative to aggregated global thresholds.

The right panel of Figure 4.5 provides additional insight by visualizing the REVEL score distribution for *MSH2*. Histograms summarize the score distributions for ClinVar-classified pathogenic (P/LP) and benign (B/LB) variants, along with all possible missense SNVs. The thresholds derived from gene-specific calibration align more closely with the observed score separation, particularly the right-shifted distribution of pathogenic variants and the distinct peak of benign variants. As a result, gene-specific thresholds generate more appropriate PP3/BP4 assignments.

Overall, the *MSH2* example reflects the broader trend observed across predictors and genes: gene-specific calibration produces more accurate interval likelihood ratios, sharper delineation between pathogenic and benign variants, and more clinically informative evidence assignments.

### 4.3.5 ClinVar Results

Across all three variant effect predictors (REVEL, AlphaMissense, and MutPred2), gene-specific calibration consistently improved the balance between pathogenic and benign variant interpretation when compared with the aggregated, gene-agnostic calibration thresholds described in Pejaver et al. (2022). These improvements were most evident for benign variants and variants of uncertain significance, where more appropriate placement into definitive evidence categories was achieved.

As summarized in Figure 4.6, gene-specific calibration substantially reduced the number of benign variants assigned to the indeterminate region (IR; 0 points). The mean reduction in IR assignments was 22.2% for REVEL (45.7 versus 17.0), 18.4% for AlphaMissense (34.0 versus 13.0), and 6.4% for MutPred2 (28.4 versus 19.4). This shift reflects fewer erroneous positive-evidence assignments under gene-specific thresholds, particularly for REVEL, which exhibited the largest gains, while MutPred2 showed more modest but still consistent improvements.

For pathogenic variants, gene-specific calibration generally imposed slightly stricter cri-

teria for assigning PP3-like pathogenic-supporting evidence. Although this led to a small increase in negative-point assignments for borderline pathogenic variants, the overall misclassification rates remained low: 1.2% for REVEL, 1.5% for AlphaMissense, and 2% for MutPred2. These values are substantially lower than benign-to-pathogenic misclassification rates observed under aggregated thresholds (6.8% for REVEL, 0.89% for AlphaMissense, and 3.6% for MutPred2), indicating that the more conservative single-gene approach avoids clinically concerning false positives. Even after removing variants with predicted splice-altering effects using SpliceAI, a small number of low-scoring pathogenic variants persisted, likely reflecting unmodeled splicing contributions near exon boundaries. As expected given their predominantly benign composition, unlabeled variants drawn from ClinVar VUS categories, gnomAD, and all possible SNVs behaved similarly to benign variants across all calibration schemes.

We next evaluated gene-level calibration performance by comparing the average change in true-positive rate ($\Delta$TPR) and false-positive rate ($\Delta$FPR) between the single-gene and aggregated approaches (Figure 4.8). Across all three predictors, the single-gene method frequently resulted in higher TPR while maintaining FPR values that were similar or only slightly increased, with most differences in FPR remaining below 0.05. Importantly, the majority of genes fell below the diagonal line $\Delta\text{TPR} > \Delta\text{FPR}$, indicating that the improvement in sensitivity generally outweighed any increase in false positives. This effect was most pronounced for REVEL (Figure 4.8A), while AlphaMissense and MutPred2 showed smaller but still favorable trends.

To evaluate the impact of calibration in a clinically curated setting, we analyzed 17 genes from the ClinGen Variant Curation Expert Panel (VCEP) corpus with sufficient labeled variants to support gene-specific calibration. For each predictor, we compared variant reclassification outcomes using PP3/BP4 evidence codes derived from aggregated thresholds and from gene-specific thresholds selected by our dynamic decision-tree framework (Figure 4.10). In all three predictors, gene-specific calibration reassigned a greater number of benign or likely benign variants, typically correcting borderline assignments produced by aggregated thresholds, while exerting minimal influence on pathogenic or likely pathogenic classifications. REVEL showed no net loss in P/LP assignments, and AlphaMissense and MutPred2 differed by only one or two variants, demonstrating a negligible impact on pathogenic classification stability.

Variants of uncertain significance (VUS) benefited most from gene-specific calibration. In all predictors, gene-specific thresholds reassigned a larger number of VUS variants into definitive benign or pathogenic categories than did the aggregated thresholds. No cross-class transitions (B/LB to P/LP or P/LP to B/LB) were observed under any method. A small

number of LB or LP variants were reassigned to VUS, which reflects the conservative nature of the single-gene thresholds and their tendency to avoid overinterpretation of borderline computational evidence.

The accompanying accuracy barplots compare three sources of PP3/BP4 evidence across genes: the PP3/BP4 codes assigned by ClinGen VCEPs, the aggregated calibration thresholds, and the gene-specific calibration thresholds. Accuracy was computed relative to the clinical significance derived solely from non-computational ClinGen evidence ("Met Codes"). Two genes—*TP53* and *MECP2*—showed significantly higher accuracy under gene-specific calibration than under either the ClinGen-provided PP3/BP4 codes or the aggregated thresholds. For the remaining genes, accuracy was generally comparable across all three methods, indicating that gene-specific calibration does not compromise classification performance even when differences in thresholds are substantial.

While accuracy remained similar across most genes, gene-specific calibration showed a clear advantage in evidence coverage. The horizontal barplots illustrate that single-gene thresholds enabled assignment of PP3/BP4 evidence to substantially more variants than either aggregated thresholds or ClinGen rules: 378 variants for REVEL, 302 for AlphaMissense, and 374 for MutPred2. These represent the highest coverage achieved across all approaches, demonstrating that gene-specific calibration not only maintains accuracy but also increases the number of variants that can be meaningfully evaluated under ACMG/AMP guidelines. This improvement in evidence coverage enhances the overall clinical interpretability of variant effect predictions.

### 4.3.6 Cross-Predictor Consistency of Gene-Level Calibration

Across genes with sufficient labeled variants for single-gene calibration (n=57), the three predictors—REVEL, AlphaMissense (AM), and MutPred2 (MP2)—exhibited broadly consistent calibration behavior, but with systematic differences in the magnitude of calibrated ACMG evidence assigned. Pairwise scatterplots of gene-level average signed evidence points (Figure 4.11A) show that AM generally assigns stronger computational evidence than either REVEL or MP2, in which the majority of genes lie above the diagonal towards AM, indicating higher average evidence-point values produced by AM. In contrast, REVEL and MP2 show more symmetric scatter around the diagonal, suggesting that these predictors assign similar overall levels of calibrated evidence for most genes.

Although the overall cross-predictor correlations are strongly positive, several genes deviate substantially from the dominant linear trend. The five genes with the largest deviations are highlighted in red. Examples include *INF2* and *PTCH1*, which receive substantially

higher evidence support from AM and REVEL relative to MP2, and *COL6A3*, for which both AM and MP2 assign stronger evidence than REVEL. Despite these outliers, most genes exhibit coherent behavior across predictors, and the systematically higher calibrated evidence values assigned by AM are consistent with the distributional differences observed in the middle boxplot of Figure 4.11B.

To assess whether calibrated evidence strength reflects the inherent discriminative ability of each predictor, we compared gene-level AUROC values with their corresponding average calibrated evidence points (Figure 4.11B-left). All three predictors exhibited strong positive correlations between AUROC and calibrated evidence (REVEL: $r = 0.812$; AM: $r = 0.846$; MP2: $r = 0.765$), indicating that genes whose pathogenic and benign variants are more easily separable in raw score space tend to receive stronger calibrated PP3/BP4 evidence. Among the three predictors, AM demonstrated both the highest overall AUROC distribution (Figure 4.11B-right) and the highest correlation between AUROC and calibrated evidence strength, consistent with its generally larger evidence-point assignments.

Several genes emerged as systematic outliers in the AUROC–evidence scatterplots. In particular, *MED12* and *FLNA* were identified as outliers across all three predictors. Both genes displayed relatively high AUROC values but lower-than-expected calibrated evidence points, placing them below the trend line in every predictor-specific comparison. This pattern suggests that, despite strong raw separability between pathogenic and benign variants, gene-specific score distributions or calibration properties may attenuate the strength of the calibrated posterior evidence for these genes.

Overall, these analyses demonstrate that while REVEL, AM, and MP2 yield broadly similar gene-level calibration patterns, AlphaMissense consistently assigns stronger calibrated evidence across genes, a trend supported by pairwise scatterplots, evidence-point distributions, and superior gene-level AUROC values. Outlier genes such as *INF2*, *PTCH1*, *COL6A3*, *MED12*, and *FLNA* highlight cases where predictor-specific score distributions or calibration characteristics result in deviations from the dominant cross-predictor trends.

### 4.3.7 Performance of the Best-Predictor Strategy

We next evaluated whether calibration performance could be improved by selecting, for each gene, the predictor with the highest discriminative ability (Figure 4.12). As a baseline, we first compared the gene aggregate calibration results for REVEL, AM, and MP2. Across predictors, the aggregated method yielded the lowest median values of $\text{AvgTPR} - \text{AvgFPR}$, indicating generally weaker discrimination when a single global calibration model is applied to all genes. Among the three predictors, AM achieved the highest median value (0.722), with

REVEL and MP2 performing slightly lower, consistent with earlier analyses of discriminative performance.

Single-gene calibration yielded consistently higher performance than gene-aggregate calibration for all predictors. Median AvgTPR − AvgFPR values increased to 0.783 for REVEL, 0.820 for AM, and 0.772 for MP2. This improvement reflects the benefit of allowing thresholds to adapt to gene-specific score distributions, thereby reducing false-positive assignments while maintaining or increasing true-positive rates.

To assess the potential upper bound of a predictor-selection framework, we implemented a mixed-predictor strategy in which each gene was assigned the predictor with the highest gene-level AUROC. Across the 82 genes for which at least one predictor produced a valid single-gene calibration model, AM was selected for 44 genes, REVEL for 23 genes, and MP2 for 15 genes. This strategy also expanded the number of genes for which calibration could be performed (82 genes total), compared with using any single predictor alone (REVEL: 70 genes; AM: 72 genes; MP2: 76 genes).

The resulting mixed-predictor distribution achieved a median AvgTPR − AvgFPR of 0.824, representing the highest overall performance among all tested strategies. However, because the set of genes in the mixed-predictor analysis differs slightly from those available for each individual predictor, we included an additional comparison using AM restricted to the same set of 82 genes. Under this matched-gene comparison, AM achieved a median of 0.820. The mixed-predictor approach not only exceeded this value but also demonstrated higher first and third quartiles, indicating consistently stronger performance across the gene set.

Taken together, these results suggest clear potential for predictor-specific calibration selection at the gene level. By leveraging the strengths of different computational models across genes, the mixed-predictor strategy offers both broader calibration coverage and improved discrimination compared to using any single predictor alone.

## 4.4 Discussion

In this chapter, I evaluated a comprehensive framework for gene-specific calibration of variant effect predictor scores and demonstrated its advantages over global, gene-agnostic approaches. The results highlight important insights into calibration behavior, show clear clinical relevance, and reveal practical limitations that motivate future methodological directions.

The analyses showed that calibration accuracy depends strongly on the context of each gene, including sample size, class balance, and the shape of the underlying score distributions.

Local posterior calibration, as originally proposed, performed well only when more than one hundred labeled pathogenic and benign variants were available. In smaller datasets it became unstable and frequently overestimated pathogenic evidence. This is a major concern for clinical genomics because fewer than half of disease-associated genes currently reach this threshold of labeled variants.

Alternative calibration strategies such as Platt scaling, beta calibration, spline-based methods, and monotonic neural networks offered superior robustness when sample sizes were limited. Nevertheless, all methods showed sensitivity to extreme class imbalance. These observations indicate that no single method is optimal for all genes. Instead, calibration performance depends on the data profile of each gene. This motivated the development of a dynamic decision-tree workflow that evaluates multiple calibration methods per gene and selects the approach with the lowest risk of misestimation. Applying this framework revealed that per-gene calibration outperformed global calibration across most genes and predictors, and that predictor choice itself should be gene-specific. For many genes, AlphaMissense provided stronger and more reliable calibrated evidence than REVEL or MutPred2, but in a sizable subset of genes the opposite was true. A mixed-predictor strategy that selected the best predictor per gene achieved the highest overall discrimination among all tested workflows.

### 4.4.1 Clinical Impact of Gene-Specific Calibration

Gene-specific calibration led to clear gains in clinically meaningful behavior of variant effect predictors. The most consistent improvement appeared on the benign side of the spectrum, where the proportion of variants left in the indeterminate region with zero ACMG evidence points dropped substantially. This reduction corresponds to fewer borderline outcomes and fewer misleading computational signals. Recalibrated PP3 and BP4 thresholds produced conservative but well-balanced assignments: false-positive pathogenic signals declined sharply relative to global thresholds, while true-positive detection remained stable or improved. In curated ClinGen VCEP datasets, gene-specific calibration reassigned more variants of uncertain significance into definitive benign or pathogenic categories, and no benign-to-pathogenic reversals were observed. Together, these results show that single-gene calibration increases evidence strength accuracy and makes computational scores more trustworthy in day-to-day variant interpretation.

These findings fit within a broader shift in genomic medicine toward context-aware interpretation. Gene-level and disease-specific frameworks are already standard for expert curation; this work shows that computational evidence benefits from the same level of tailoring.

The strong correlations between gene-level AUROC values and calibrated evidence indicate that calibration captures the underlying discriminative power of each predictor within each gene, rather than imposing arbitrary thresholds. A key conceptual insight is that variant interpretation improves when the system is flexible: different predictors are best for different genes, and group-level structure such as shared protein domains or mechanisms can be used when single genes lack data. This suggests a future decision-support ecosystem that combines predictor-specific selection, gene-level calibration, cluster-level modeling, and functional assay calibration into a unified workflow. The framework developed in this chapter provides a foundation for such an integrated, context-aware approach to computational evidence.

### 4.4.2 Comparison With Emerging Methods

A recent contribution to the field, the *acmgscaler* method introduced by Badonyi and Marsh in 2025, provides an additional gene-level calibration strategy for both computational predictions and functional assay scores. It uses kernel density estimation to derive score-to-likelihood mappings and produces ACMG-consistent thresholds. The tool is implemented as an R package and a Colab interface, and it offers an automated and robust workflow that complements the goals of this dissertation.

There are important differences between *acmgscaler* and the present work. The default implementation of *acmgscaler* adopts a fixed prior probability of pathogenicity, typically set near ten percent, which simplifies estimation but does not account for true biological variation across genes or differences in clinical context. The authors themselves acknowledge the difficulty of determining accurate priors and explicitly note that real priors vary between genes. In contrast, my framework avoids imposing a universal prior and instead infers gene-specific priors directly from labeled and unlabeled variant distributions or uses empirical benchmarking to identify thresholds. In addition, the dynamic decision-tree workflow introduced in this dissertation can incorporate *acmgscaler* as an optional branch in the future.

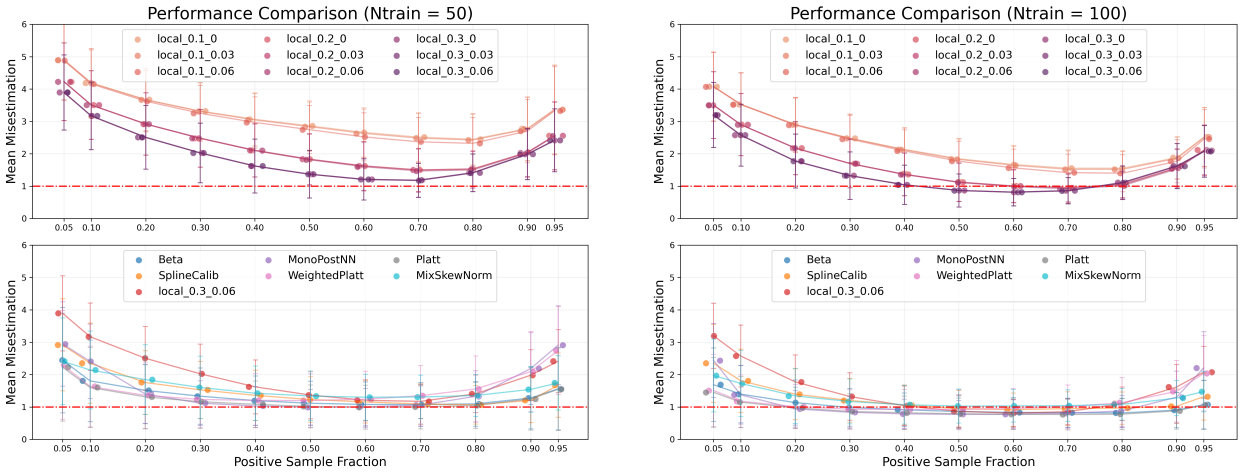
### 4.4.3 Limitations and Future Directions

This work has several limitations. First, gene-specific calibration requires a minimum number of labeled variants, and many disease genes still lack sufficient ClinVar pathogenic and benign examples. Although cluster-level calibration provides coverage for many additional genes, a subset of the genome remains data-poor. Second, the evaluation relied on retrospective classifications from ClinVar and ClinGen. True prospective evaluation will require

applying calibrated thresholds to new clinical variants and validating consistency across laboratories. Third, this study focuses on missense predictors and does not incorporate structural consequences, or other mechanisms that may influence pathogenicity. A small number of misclassified pathogenic variants likely reflect such unmodeled mechanisms. Fourth, large-scale functional datasets remain available for only a limited number of genes, which constrains the potential synergy between computational and experimental calibration frameworks.

Future work will expand calibration to disease mechanisms and variant types that were not fully captured in this study. Several directions are especially promising. First, integrating functional assay calibration and computational calibration into a unified framework would allow laboratories to weigh both data types coherently under ACMG/AMP rules. Second, large population sequencing cohorts such as the All of Us Research Program offer opportunities for prospective validation, particularly for evaluating whether calibrated benign thresholds correctly capture the behavior of recurrent population variants. Third, enhanced estimation of gene-specific priors, possibly through hierarchical or semi-supervised models that borrow information across related genes, may provide reliable priors even for data-sparse genes. Fourth, integrating emerging tools such as *acmgscaler* and allowing user-defined or context-specific priors could extend the flexibility of the dynamic calibration workflow.

In summary, this work establishes a scalable and clinically meaningful framework for gene-specific calibration of variant effect predictors. It enhances computational evidence use under the ACMG/AMP guidelines, improves interpretability, reduces uncertainty, and creates a path toward more precise and context-aware genomic medicine.




(a) Calibration set: 50 variants

(b) Calibration set: 100 variants

Figure 4.2: Performance comparison of calibration methods under different parameter combinations. Top panels: Impact of local calibration parameters (window size and gnomAD fraction). Bottom panels: Alternative methods outperform the best local calibration configuration. Results are shown for test sets of 1,000 variants under varying class balance conditions. Error bars represent standard deviations across different true prior probabilities. Only methods with average performance below 1 in at least one scenario are displayed.

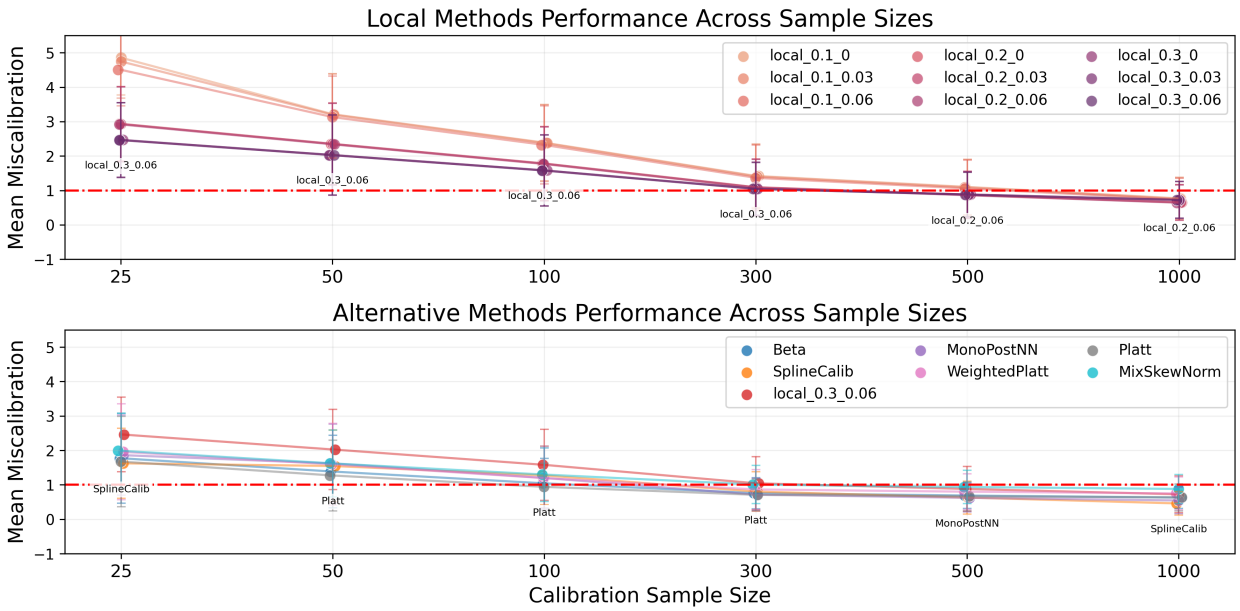
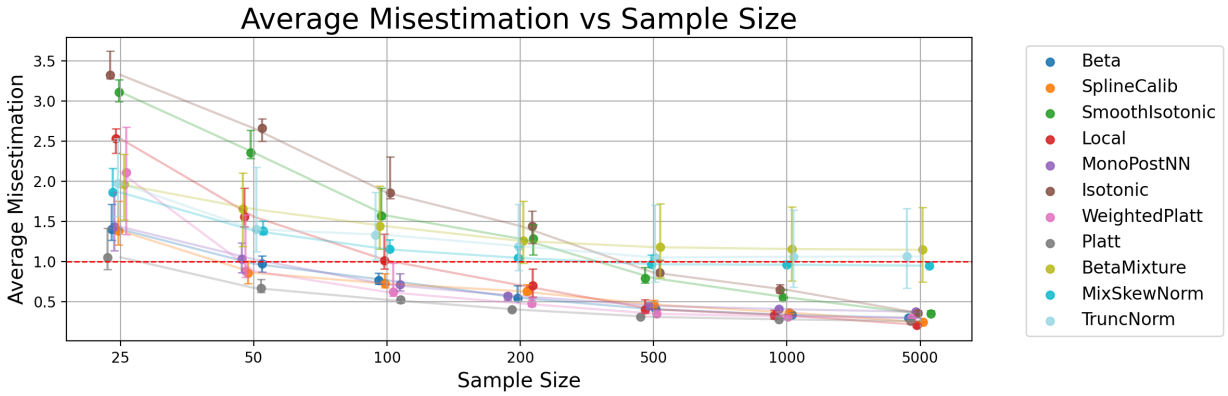
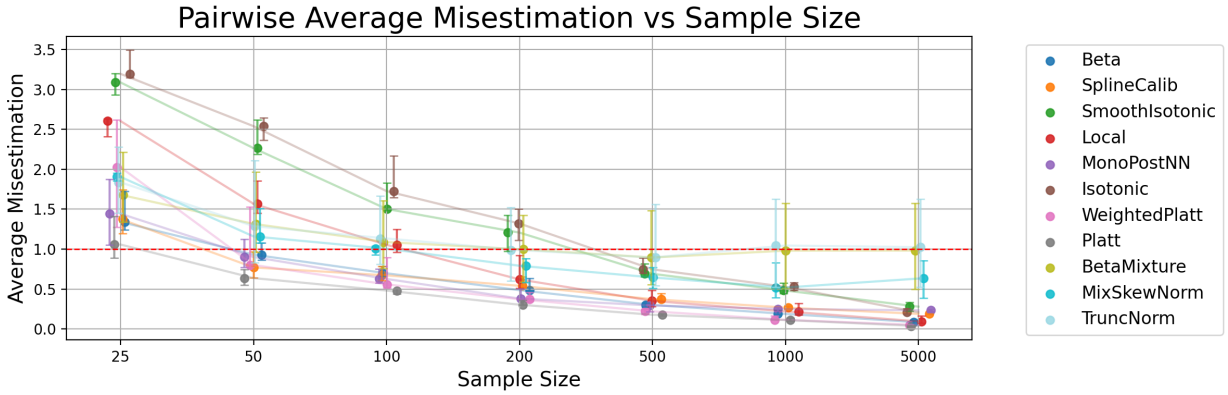



Figure 4.3: Alternative Calibration methods performance across different calibration sample sizes and class balance conditions. Error bars represent variability across different true prior probabilities. Only methods with at least average performance in one scenario, lower than 1 shown to simplify comparison.




**(A)**




**(B)**

Figure 4.4: Subsampling robustness analysis for calibration methods. (A) Misestimation in ACMG evidence points relative to a fixed reference calibration model. (B) Deviation from full-data calibration performance in a paired full-vs.-subsampled evaluation.

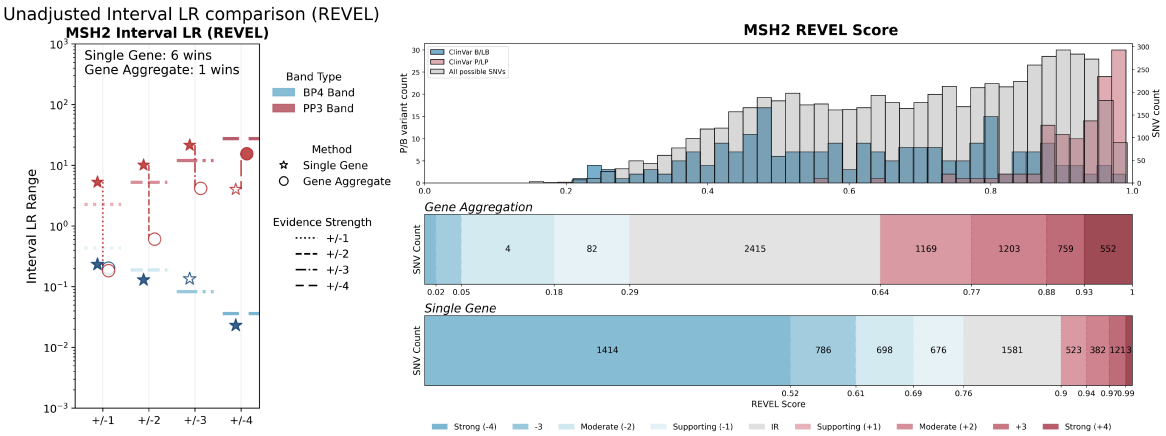
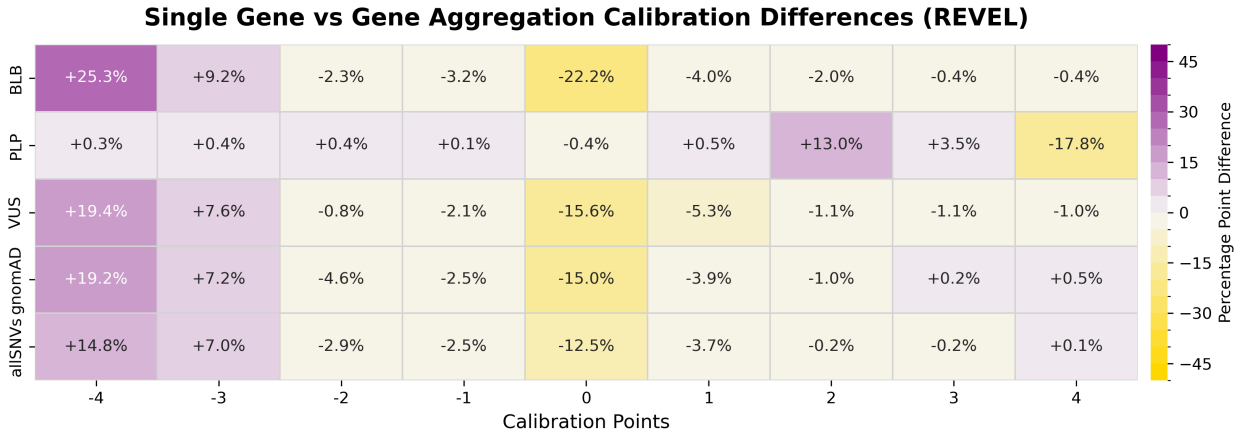
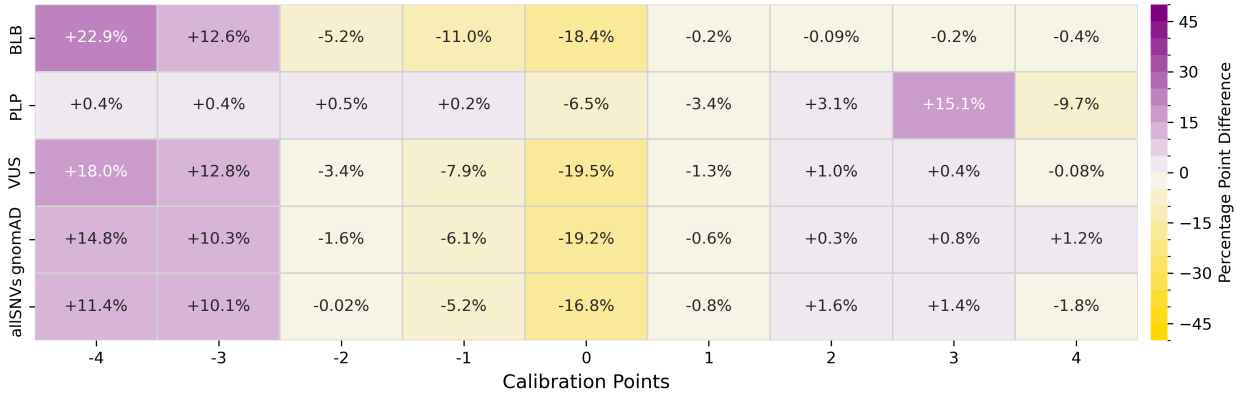


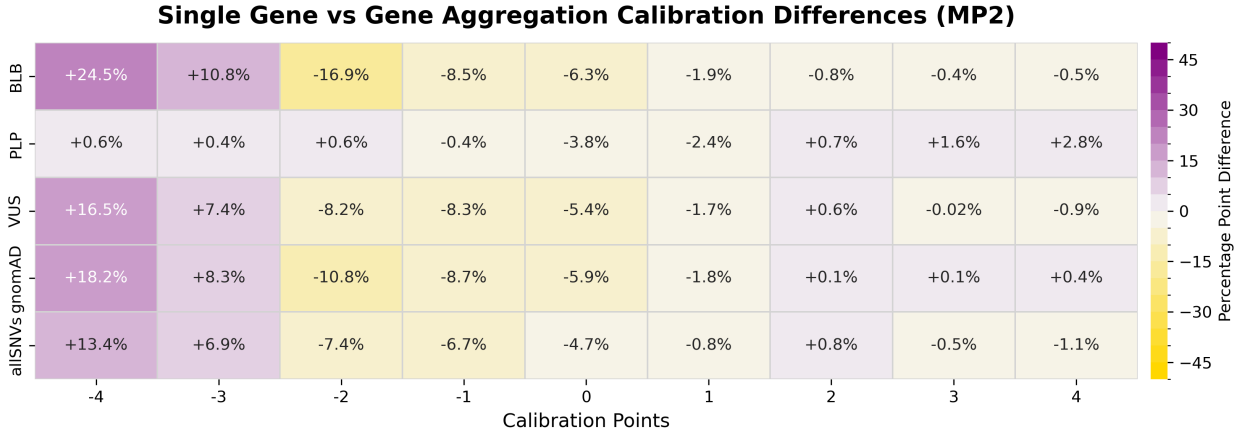

Figure 4.5: Gene information card for MSH2 REVEL scores. *left*: Interval LR+ comparison between gene-specific calibration (star) and aggregated-level calibration (circle). Horizontal lines represent LR+ cutoffs for Supporting (dotted), Moderate (–.), +3 (-..), and Strong (–) evidence. Each point is colored by evidence type—blue for BP4 (benign) and red for PP3 (pathogenic)—and filled to indicate which method performed better for that evidence strength. A scorecard in the upper-left corner summarizes the number of strengths "won" by each method. *right*: (Top) Histogram of REVEL scores for ClinVar-classified pathogenic/likely pathogenic (P/LP, red), benign/likely benign (B/LB, blue), and all possible SNVs (gray), with SNV counts shown on the right y-axis. (Middle) Stacked bar plot showing variant classification under AJHG (aggregated) REVEL thresholds, with color intensity indicating evidence strength (dark red = strong PP3, dark blue = strong BP4). Numeric labels show variant counts in each bin.




**(A)**

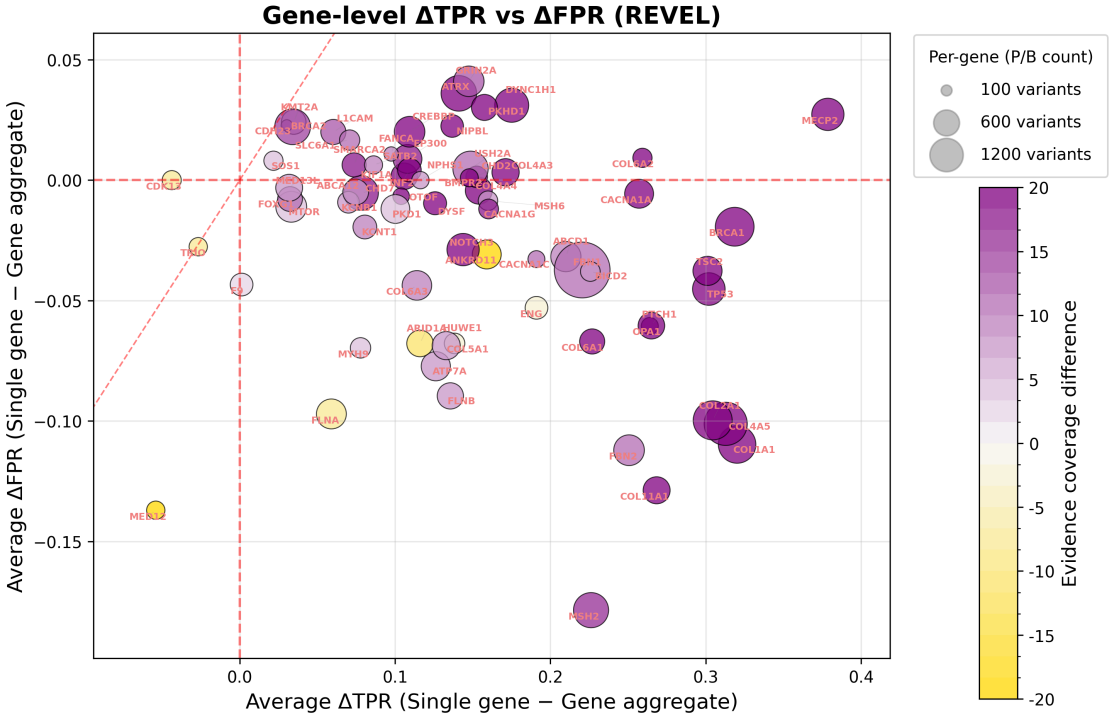



**(B)**

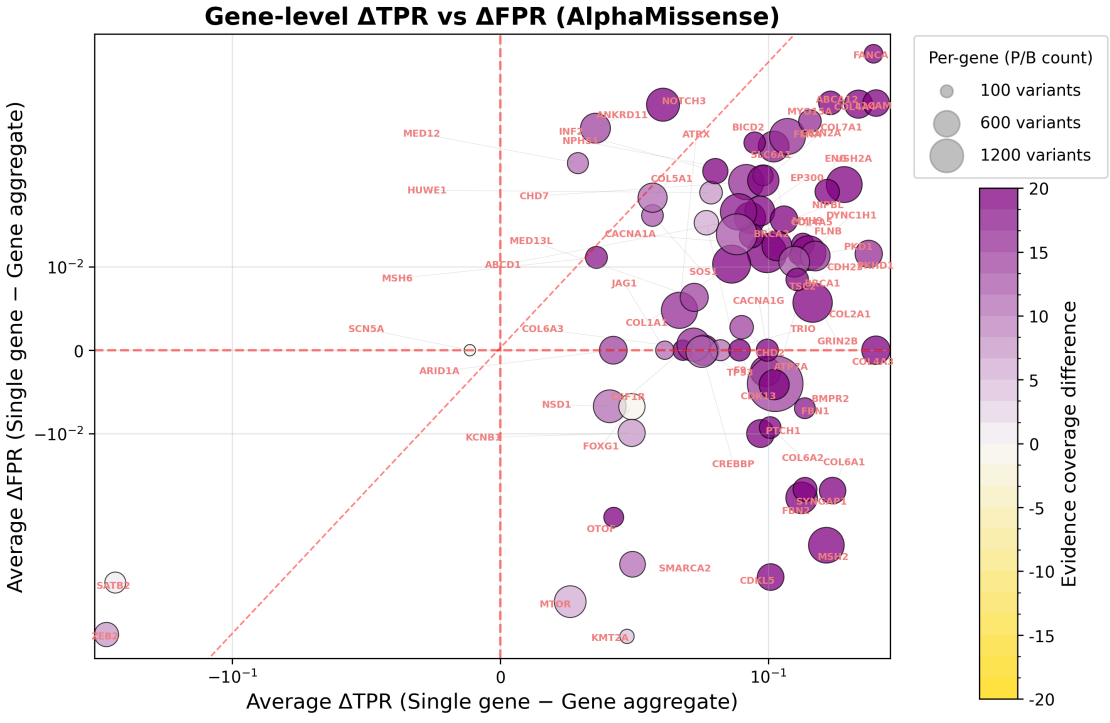



**(C)**

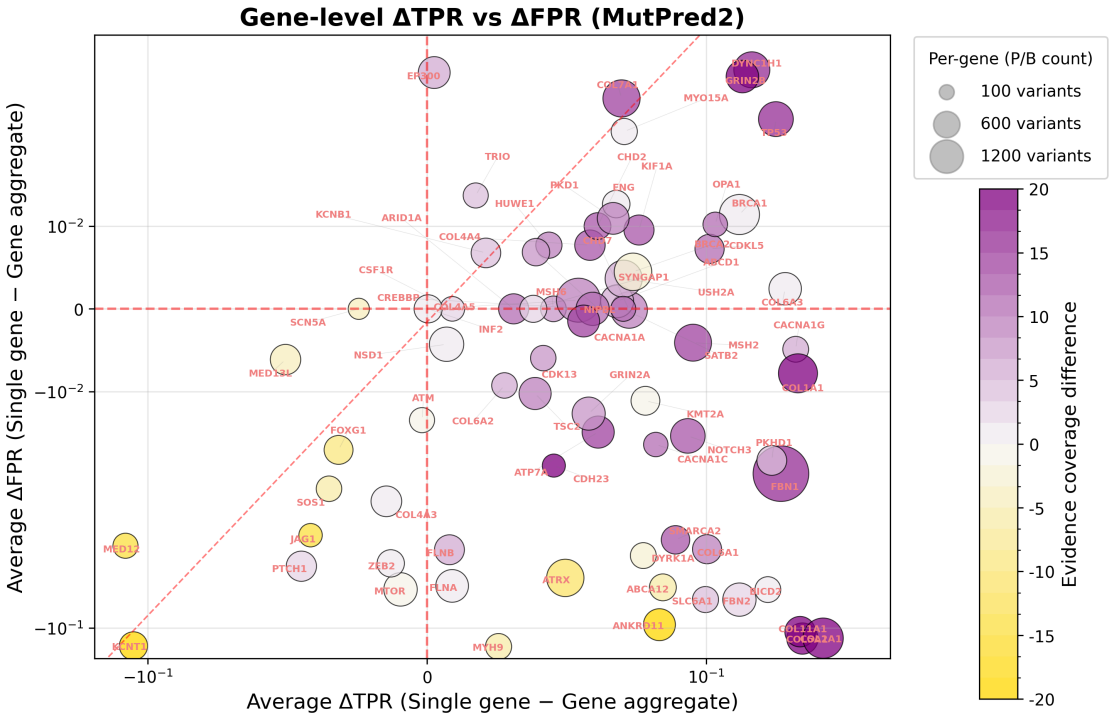
Figure 4.6: Comparison of gene-specific versus AJHG calibration methods across three predictors: REVEL (A), AM (B), and MP2 (C). Heatmaps display percentage point differences in calibration assignments (x-axis) stratified by variant classifications (y-axis). Purple indicates higher assignment rates by the gene-specific method.




**(A)** REVEL

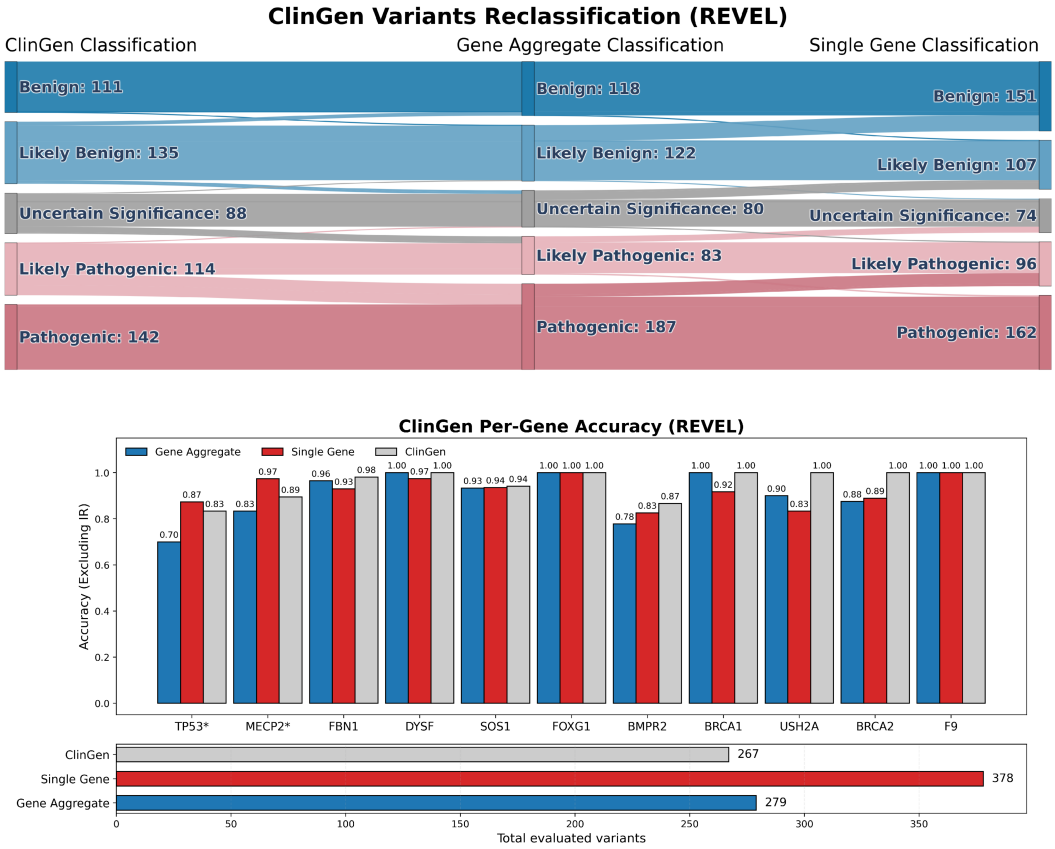



**(B)** AlphaMissense (AM)

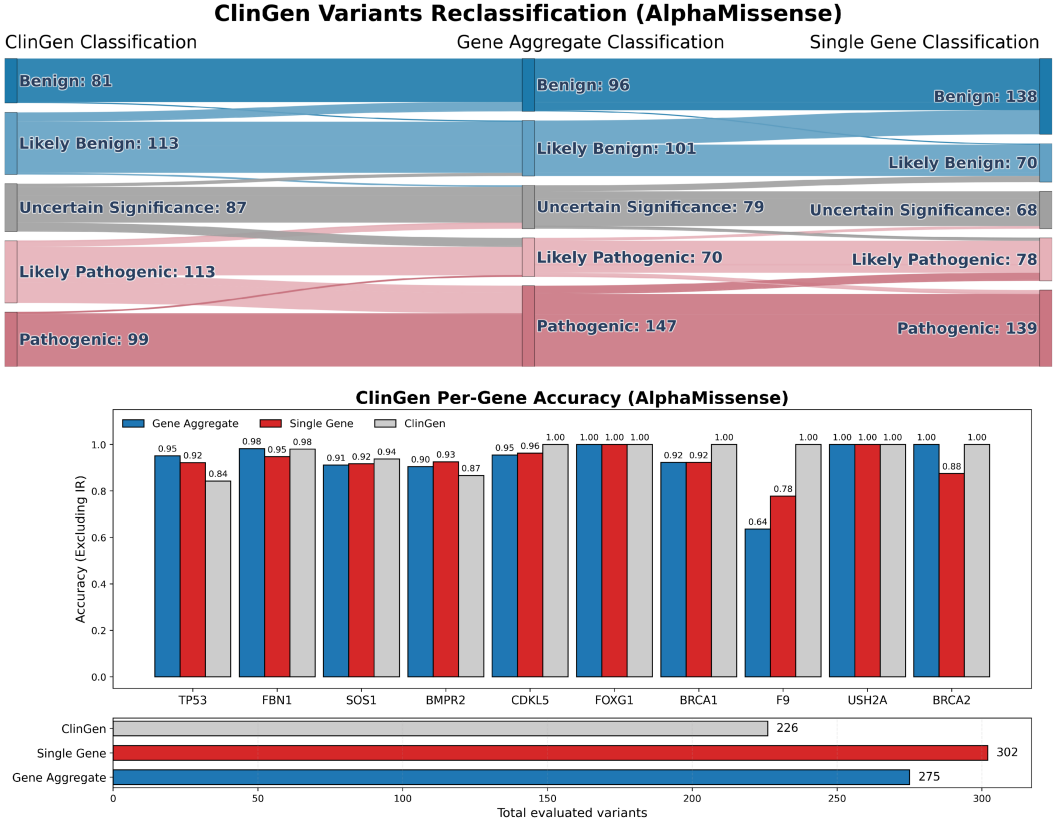



**(C)** MutPred2 (MP2)

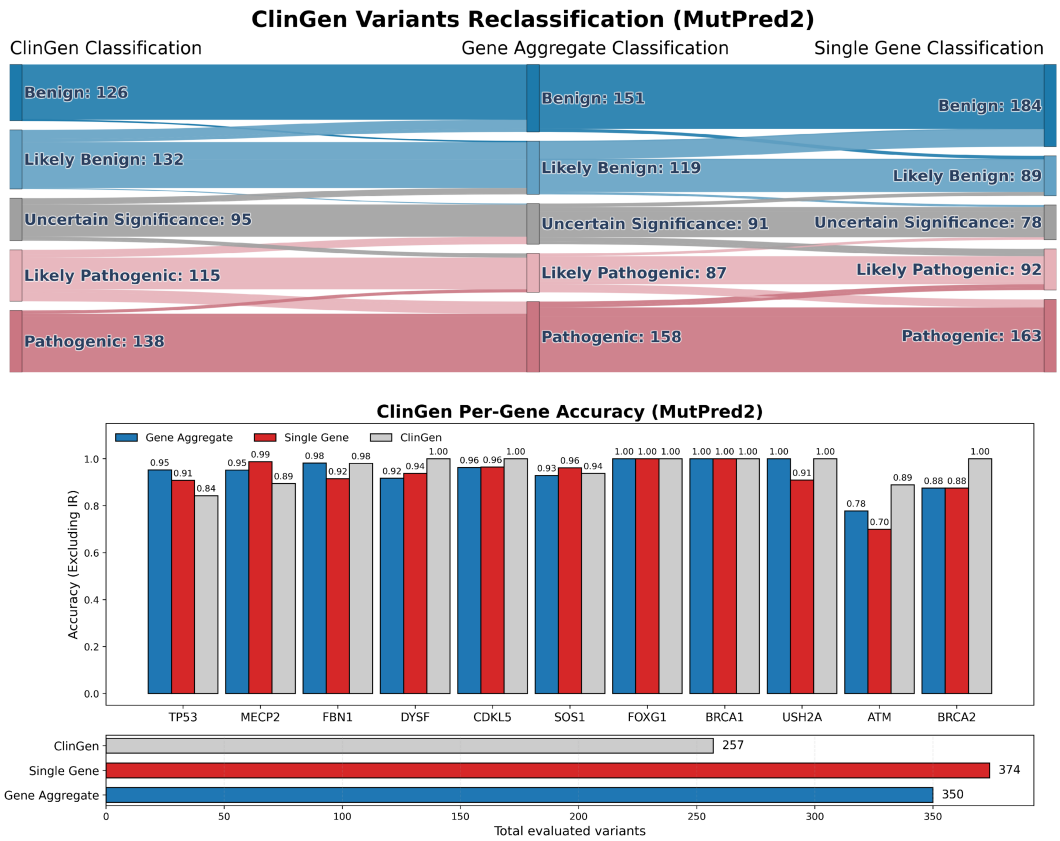
Figure 4.8: Gene-level comparison of gene-specific versus gene-aggregated calibration thresholds for three predictors: REVEL (A), AlphaMissense (B), and MutPred2 (C). Each point represents a gene. The x-axis shows the average change in true positive rate (ΔAvgTPR) between the gene-specific and gene-aggregated methods (gene-specific minus aggregated), and the y-axis shows the corresponding change in false positive rate (ΔAvgFPR). Point color encodes the difference in evidence coverage (change in the fraction of variants receiving PP3/BP4 evidence), with deeper purple indicating higher coverage under the gene-specific method. Point size is proportional to the number of pathogenic/benign (P/LP and B/LB) variants ($n_{\mathrm{PB}}$) available for that gene. The dashed red diagonal ($x = y$) indicates equal changes in TPR and FPR; genes below this line have a larger gain in TPR than in FPR, indicating a net improvement in classification performance for the gene-specific calibration.




**(A)** REVEL




**(B)** AlphaMissense (AM)




**(C)** MutPred2 (MP2)

Figure 4.10: ClinGen-based reclassification analysis for 17 VCEP genes using three computational predictors: REVEL (A), AlphaMissense (B), and MutPred2 (C). In each subfigure, the upper panel shows a Sankey diagram summarizing how ACMG/AMP clinical classifications change when PP3/BP4 computational evidence is reassigned using genome-wide gene-aggregated thresholds (middle column) and single-gene thresholds (right column), relative to the original ClinGen classifications (left column). Flows capture transitions such as VUS→B/LB and VUS→P/LP under each calibration scheme. The lower panel shows per-gene accuracy for three approaches evaluated on variants with non-indeterminate evidence: ClinGen-provided PP3/BP4 codes (grey), gene-aggregated calibration thresholds (blue), and single-gene calibration thresholds (red). Bar heights reflect the fraction of variants correctly classified when recomputing clinical significance from (i) non-computational ClinGen evidence alone and (ii) the same baseline plus PP3/BP4 evidence assigned by each method, thereby quantifying how much the different calibrations improve classification accuracy, increase evidence coverage, and alter the impact of computational evidence compared to the aggregate approach.

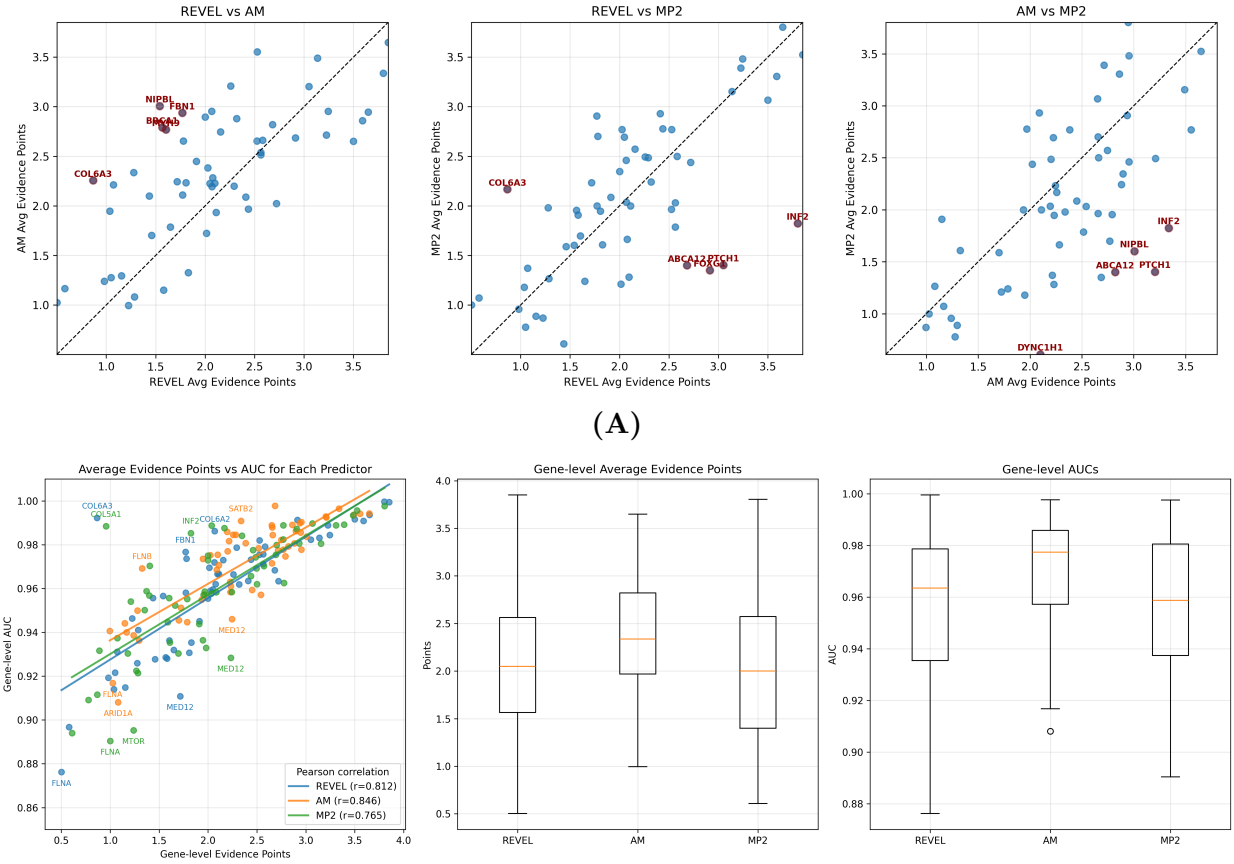



Figure 4.11: Cross-predictor consistency of gene-level calibration. **A.** Pairwise scatterplots of average signed ACMG evidence points for REVEL vs. AlphaMissense (AM), REVEL vs. MutPred2 (MP2), and AM vs. MP2. Each point represents a gene, and values correspond to the mean signed evidence points assigned across all labeled variants for that gene under each calibrated predictor. Diagonal alignment indicates strong cross-predictor agreement. **B.** (*left*) Relationship between uncalibrated gene-level AUROC and calibrated gene-level evidence points, with Pearson correlation coefficients quantifying the association between raw discriminative performance and calibrated evidence strength. (*middle*) Distribution of gene-level average signed evidence points across predictors. (*right*) Distribution of uncalibrated AUROC values for each predictor. These analyses jointly evaluate the degree to which REVEL, AM, and MP2 provide consistent calibrated evidence and how their calibrated outputs relate to baseline predictive accuracy.

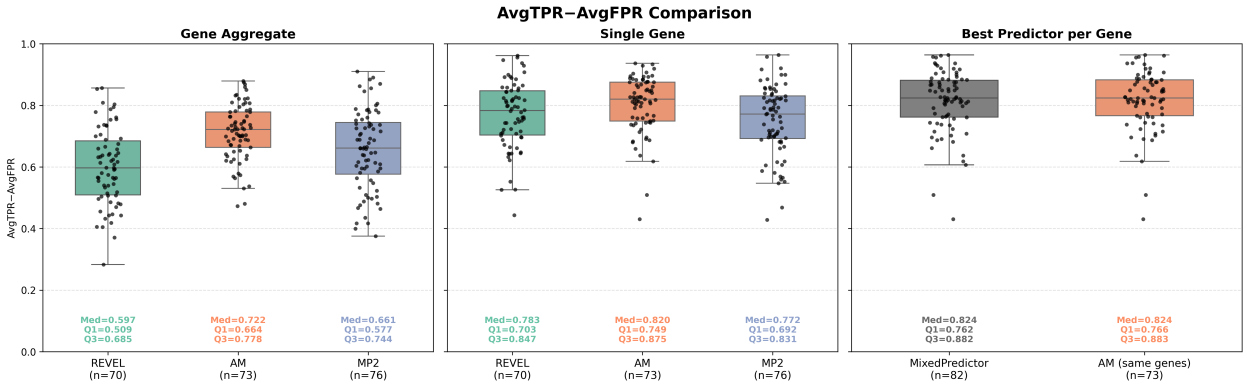



Figure 4.12: Comparison of AvgTPR−AvgFPR across predictors and model-selection strategies. *left* shows gene aggregate calibration performance for REVEL, AlphaMissense (AM), and MutPred2 (MP2) using aggregated thresholds. *middle* shows single gene calibration for each predictor applied uniformly across all genes. *right* compares the mixed-predictor strategy—in which each gene is assigned the predictor with the highest gene-level AUROC—against AM applied uniformly to the same set of genes. Each boxplot summarizes the distribution of AvgTPR − AvgFPR across genes, with individual genes displayed as jittered points. Median, first quartile, and third quartile values are annotated above each box. The mixed-predictor approach reflects the best achievable performance if per-gene predictor selection were permitted, enabling direct comparison between a single-predictor strategy and an oracle model that selects the optimal predictor for each gene.

# Chapter 5

# Conclusion and Future Directions

Interpreting the pathogenicity of genomic variants remains a fundamental challenge in genomic medicine. This dissertation tackled this challenge by developing an integrated framework that combines gene prioritization, gene-specific priors, and tailored calibration of predictive models to improve variant classification. The overarching goal of this work is to reduce uncertainty in variant interpretation, thereby decreasing the prevalence of variants of uncertain significance (VUS) and enabling more confident, evidence-driven clinical decisions. In this final chapter, I summarize the key contributions, discuss how they fit together into a cohesive strategy, and examine their broader implications. I also address the limitations of the current approach and outline promising directions for future research. Finally, I consider how the tools and resources developed in this dissertation can be adopted in practice by clinical laboratories and research consortia.

## 5.1 Summary of Findings and Implications for Genome Medicine

This dissertation introduced a three-part framework for improving variant interpretation, each part corresponding to one of the main chapters. First, we established a data-driven system for prioritizing genes for functional assays based on clinical impact. This prioritization framework (Chapter 2) quantitatively ranks genes by the expected benefit of additional experimental evidence, using criteria such as the number of unresolved VUS in the gene, the likelihood that new functional data could resolve those uncertainties, and the potential to refine computational predictions. By dynamically updating these gene scores as new information emerges, the framework ensures that experimental resources can be directed to where they will have the greatest clinical impact. As a result, collaborative efforts can focus

on high-priority genes (for example, highlighting TSC2 as a strong candidate for multiplex assays was one tangible outcome), accelerating the resolution of uncertain variants in those genes. This strategic allocation of resources has broad implications for genome medicine: it provides a rational, transparent method to decide which genes to study next, thereby systematically reducing the bottleneck of VUS through experimental evidence.

Second, we developed a methodology for estimating gene-specific pathogenicity priors (Chapter 3). Traditional variant classification often assumes a universal prior probability of pathogenicity for all genes, an approach that fails to account for the tremendous variability in gene disease propensity. In this work, we replaced that one-size-fits-all assumption with nuanced, gene-level priors that better reflect each gene's underlying biology and disease role. Using a positive–unlabeled learning approach on ClinVar pathogenic records and population variant data, we inferred the background rate of disease-causing variants for each gene. The result is a comprehensive catalog of gene-specific prior probabilities that can be directly incorporated into Bayesian variant classification frameworks. These gene-specific priors yield more realistic posterior probabilities when interpreting variants: for genes that rarely cause disease, even moderate evidence might not suffice for pathogenic classification, whereas for high-risk genes, the same evidence could be much more significant. By aligning variant interpretation with gene-specific disease biology, this contribution improves the accuracy of classification and reduces the risk of systematic misclassification (either over- or underestimating pathogenicity) that can occur when using generic priors. The availability of gene-specific priors is also a key enabler for the next component of our framework, providing the foundational probabilities needed to calibrate computational prediction scores.

Third, we introduced a novel approach for gene-aware calibration of in silico variant effect predictors (Chapter 4). Computational tools that predict variant deleteriousness (such as missense impact scores) are widely used in practice, but their raw output scores are not directly interpretable as probability of pathogenicity and can have different meaning in different genes. We addressed this by constructing a calibration pipeline that converts a predictor's raw score into a posterior probability of pathogenicity in a gene-specific manner. In other words, for each gene (or for clusters of genes with shared characteristics), we learned how to translate a variant's score into a calibrated probability that the variant is disease-causing. This calibration was anchored by the gene-specific priors from Chapter 3 and by known pathogenic and benign variant data. A key innovation was the development of a dynamic decision-tree workflow that automatically selects the most suitable calibration method for each gene. Through extensive evaluation and benchmarking against ten established post-hoc calibration techniques, we found that no single calibration method is universally best for all scenarios; factors like the number of known variant examples in a gene, the distribution of

scores, and the gene's prior all influence which method performs optimally. Our decision-tree approach navigates these factors, choosing (for example) simpler non-parametric scaling when data are abundant, or more constrained model-based calibration when data are sparse or score distributions are skewed. We further extended calibration to a cluster level, grouping genes by biological or functional domains so that genes with limited variant data can "borrow strength" from related genes. By estimating cluster-level priors and applying the same dynamic calibration logic to these gene clusters, we improved predictive reliability for many genes that would otherwise be too data-poor to calibrate individually. This gene- and cluster-specific calibration strategy yields practical thresholds for computational evidence (e.g. defining gene-specific score cutoffs that correspond to ACMG/AMP evidence codes like PP3 (supporting pathogenicity) or BP4 (supporting benign impact)). In clinical variant interpretation, such calibrated thresholds mean that pathogenicity predictions from in silico tools can be used as quantitative, gene-tailored evidence with greater confidence. Overall, this component integrates computational predictions into the variant classification framework more effectively, ensuring that automated predictions help rather than confuse the classification process.

## 5.2 Limitations

While the results of this dissertation demonstrate the value of an integrated, gene-aware approach, it is important to acknowledge several limitations of the current work. First, the prioritization framework in Chapter 2 relies on existing databases (ClinVar, gnomAD, etc.) for input metrics. This means its recommendations are only as good as the data available. Genes that are understudied or for which few variants have been reported might not score highly, even if they are biologically important, simply due to sparse data. Conversely, the framework assumes that resolving VUS in a high-scoring gene will indeed yield clinical benefit; in reality, not all functional assay results may lead to straightforward reclassification of variants, especially if assay interpretation is complex. Additionally, the scoring system combines multiple criteria with weights that, while systematically chosen, could be further refined. There is an implicit assumption that the chosen criteria (number of VUS, reclassification likelihood, predictor improvement potential, etc.) sufficiently capture what makes a gene a good candidate for study. This may not hold in special cases — for instance, genes that have few VUS but are associated with extremely severe diseases might be high-impact despite not meeting typical criteria thresholds.

The gene-specific prior estimation method (Chapter 3) also has limitations. The positive–unlabeled learning approach depends on the set of ClinVar pathogenic variants (treated

as positives) and presumed benign variants from population data (treated as unlabeled negatives). If ClinVar's data are biased towards well-studied genes or certain variant types, the inferred priors could be skewed. Some genes might have a high prior simply because many pathogenic variants have been reported (potentially reflecting reporting bias rather than true biology), whereas others might seem to have a low disease prior if their pathogenic variants are underreported. Moreover, our method assumes that most common population variants are non-pathogenic; while generally true for severe Mendelian diseases, there could be exceptions (e.g. adult-onset or mild conditions where pathogenic variants are not fully penetrant and appear in population databases). The priors catalog should therefore be interpreted with caution in such scenarios. Another limitation is that the prior estimation currently treats each gene independently (aside from the optional clustering extension later); it does not explicitly model relationships between genes (for example, genes in the same pathway might have correlated disease propensities, which we do not leverage in the basic prior model).

For the calibration framework in Chapter 4, one limitation is the requirement of sufficient variant data per gene (or per cluster) to perform robust calibration. Although the cluster-based approach mitigates this issue for many genes, truly data-scarce genes (with little to no known pathogenic variants and few benign references) remain challenging. In those cases, our decision-tree workflow selects conservative calibration strategies, but the confidence in calibrated probabilities for such genes will inherently be lower. There is also an added layer of complexity introduced by the dynamic method selection; while we demonstrated its value, it might be more complicated for laboratories to implement compared to a single uniform calibration method. Ensuring that this decision-tree logic is easily usable (for example, via a software tool or web resource) is necessary for practical adoption. Another consideration is that our calibration (and indeed the entire framework) currently focuses on missense variants in Mendelian disease genes. Variants in non-coding regions, structural variants, or variants related to complex traits were outside the scope of this work. The principles of gene-specific evidence weighting likely apply in those contexts too, but additional research is needed to generalize our approach to other variant types and genetic architectures. Finally, all components of the framework depend on continual updates as new data emerge. The gene prioritization rankings, the gene-specific priors, and the calibration models will degrade in accuracy over time if not refreshed with the latest variant data, new disease gene discoveries, and improved predictor algorithms. This means the utility of our system relies on an ongoing commitment to data maintenance and integration of new evidence.

## 5.3 Future Directions

Despite these limitations, the work presented here opens several exciting avenues for future research and development. Key directions to build upon this dissertation include:

- **Expanded Features for Prior Estimation:** Future efforts could enrich the gene-specific prior model by incorporating additional genomic and biological features. For example, integrating measures of gene constraint (such as pLI or LOEUF scores from large population sequencing studies) could improve estimates of a gene's tolerance to variation. Other features like tissue-specific expression patterns, gene network centrality, or functional importance of protein domains might correlate with the likelihood that variants in a gene cause disease. A machine learning model that uses such features (alongside ClinVar and population data) to predict pathogenicity priors could yield even more accurate and nuanced gene-specific priors. Expanding the feature set may also help identify subtle signals for genes with limited variant data, by leveraging what is known about gene function and evolution.

- **Semi-Supervised and Transfer Learning Across Genes:** The gene clustering approach in this dissertation is a first step toward sharing information across genes; however, more sophisticated semi-supervised learning or transfer learning techniques could further leverage data from well-characterized genes to inform predictions in less-studied ones. One future direction is to develop multi-gene or pan-genome models that jointly learn from variants in many genes, using architectures such as hierarchical Bayesian models or deep neural networks with gene-specific parameters. Such models could treat each gene as related tasks, allowing them to "borrow strength" from each other in a data-driven way rather than through predefined clusters alone. Unlabeled variants (those currently classified as VUS) could be incorporated in a semi-supervised fashion – for instance, by expecting consistency in predictor score distributions or functional assay patterns across related genes – to improve calibration and prior estimates even when class labels are missing. This cross-gene learning approach could greatly expand the reach of reliable variant interpretation to genes that today have too little data to analyze in isolation.

- **Dynamic and Iterative Calibration Workflows:** The decision-tree calibration framework could be extended into an even more dynamic system that continuously learns and adapts. In the future, one could envision an iterative calibration pipeline that updates itself as new variant classifications become available. For example, if a lab applies our calibrated thresholds and subsequently confirms a VUS as pathogenic

through independent evidence, that new data point could be fed back to refine the calibration for that gene (or cluster). Over time, this closed-loop learning system would become more accurate and possibly evolve new rules for the decision-tree model selection. Additionally, exploring other evidence types in calibration (such as incorporating functional assay scores directly into the calibration model alongside in silico predictors) is a promising avenue. This would merge the experimental data from Chapter 2 with the computational framework of Chapter 4, further unifying the approach.

- **Broader Variant Categories and Clinical Integration:** Another important future direction is to extend the principles of this framework beyond missense variants. Calibrating predictors for other variant classes – such as splice-site changes, small indels, or copy-number variants – would require developing or identifying suitable predictive features and variant datasets for those categories. Preliminary work could involve evaluating whether gene-specific priors and evidence thresholds similarly improve classification of these variant types. Moreover, integrating our gene-specific approach with existing clinical classification guidelines and pipelines will be crucial. For instance, working with professional organizations (like the ACMG/AMP or ClinGen) to incorporate gene-specific threshold recommendations for computational evidence could standardize adoption. Creating user-friendly software tools or web platforms that implement gene-specific priors and calibration will lower barriers for clinical labs to use this framework in routine variant analysis. Ensuring compatibility with popular variant interpretation tools and databases (such as ClinVar, LOVD, or diagnostic lab reporting systems) would facilitate seamless integration.

- **New Data Sources and Functional Assay Integration:** As high-throughput multiplexed functional assays become more prevalent, the data they produce can be looped back into our framework. One future research direction is to directly integrate functional assay outcomes into variant pathogenicity predictions and priors. For example, assay-based variant effect maps for a gene (produced as part of systematically testing variant impacts) could refine that gene's pathogenicity prior or provide an independent calibrated evidence track parallel to computational predictors. Additionally, other emerging data sources – such as large-scale patient genomic–phenomic datasets (e.g. biobank studies) – might be used to update gene impact estimates or even identify gene–gene interactions that affect variant interpretation. By continuously incorporating diverse new data types, we can further reduce uncertainty and keep the variant interpretation framework at the cutting edge of genome medicine.

## 5.4 Conclusion

In summary, these three components—impact-driven gene prioritization, gene-specific priors, and gene-/cluster-specific score calibration—form a cohesive framework for reducing uncertainty in variant interpretation. Each component targets a different aspect of the problem: the first directs experimental efforts to generate new evidence where it's needed most, the second fixes the statistical foundation by customizing priors to each gene's context, and the third refines the use of computational evidence on a per-gene basis. By combining these approaches, we can significantly shrink the gray zone of VUS. For example, a previously uncertain variant in a high-priority gene might be resolved via a new functional assay; simultaneously, our gene-specific prior ensures the evidence is weighed appropriately, and our calibrated predictor scores might provide additional support tipping the variant into pathogenic or benign territory. The broader implication for genome medicine is a more reliable and actionable genetic diagnosis: as more variants are confidently classified, patients and clinicians receive clearer answers, and downstream decisions (such as screening, surveillance, or treatment) can be made with greater assurance. In essence, this dissertation's framework moves the field closer to the day when "uncertain significance" truly becomes a rare exception.

# Bibliography

Mihaly Badonyi and Joseph A Marsh. acmgscaler: an R package and colab for standardized gene-level variant effect score calibration within the ACMG/AMP framework. *Bioinformatics*, 41(10), October 2025.

Richard E Barlow and Hugh D Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

Timothy Bergquist, Sarah L Stenton, Emily A W Nadeau, Alicia B Byrne, Marc S Greenblatt, Steven M Harrison, Sean V Tavtigian, Anne O'Donnell-Luria, Leslie G Biesecker, Predrag Radivojac, Steven E Brenner, Vikas Pejaver, and ClinGen Sequence Variant Interpretation Working Group. Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria. September 2024.

Vineel Bhat, Tian Yu, Lara Brown, Vikas Pejaver, Matthew Lebo, Steven Harrison, and Christopher A Cassa. Extracting and calibrating evidence of variant pathogenicity from population biobank data. *Am. J. Hum. Genet.*, 112(8):1805–1817, August 2025.

Breast Cancer Association Consortium, Leila Dorling, and etc. Carvalho, Sara. Breast cancer risk genes - association analysis in more than 113,000 women. *N. Engl. J. Med.*, 384(5): 428–439, February 2021.

Sarah E. et al. Brnich. Recommendations for application of functional evidence in variant interpretation. *American Journal of Human Genetics*, 106(5):784–794, 2019. doi: 10.1016/j.ajhg.2020.03.003.

Yile et al. Chen. Automated gene prioritization pipeline, 2023. Pipeline documentation, internal IGVF project.

Jake et al. Cheng. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 2023. doi: 10.1126/science.adj0933.

Wyatt T Clark, Laura Kasak, and etc. Bakolitsa, Constantina. Assessment of predicted enzymatic activity of $\alpha$-N-acetylglucosaminidase variants of unknown significance for CAGI 2016. *Hum. Mutat.*, 40(9):1519–1529, September 2019.

Manuel de Lera Ruiz and Richard L. Kraus. Voltage-gated sodium channels: Structure, function, pharmacology, and clinical indications. *J. Med. Chem.*, 58(18):7093–7118, 2015. doi: 10.1021/jm501981g. URL https://doi.org/10.1021/jm501981g.

M Dias, R Orenbuch, and D S Marks. Toward trustable use of machine learning models of variant effects in the clinic. *Am J Hum Genet*, 111:2589–2593, 2024.

Marina T DiStefano, Scott Goehringer, and etc. journal = Babb, Lawrence. The gene curation coalition: A global effort to harmonize gene-disease evidence resources.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 213–220, 2008. doi: 10.1145/1401890.1401920.

Daniel Esposito, Jochen Weile, Jay Shendure, Lea M Starita, Anthony T Papenfuss, Frederick P Roth, Douglas M Fowler, and Alan F Rubin. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.*, 20(1):223, November 2019.

Gregory M et al. Findlay. Accurate classification of brca1 variants with saturation genome editing. *Nature*, 562:217–222, 2018. doi: 10.1038/s41586-018-0461-z.

Douglas M Fowler and Stanley Fields. Multiplexed assays of variant effect. *Annual Review of Genomics and Human Genetics*, 24:1–25, 2023. doi: 10.1146/annurev-genom-112921-122403.

Hannah et al. Gelman. Recommendations for the validation of clinical genome sequencing. *Genetics in Medicine*, 2019. doi: 10.1038/s41436-019-0479-4.

Sanna Gudmundsson, Moriel Singer-Berk, Nicholas A Watts, William Phu, Julia K Goodrich, Matthew Solomonson, Genome Aggregation Database Consortium, Heidi L Rehm, Daniel G MacArthur, and Anne O'Donnell-Luria. Variant interpretation using population databases: Lessons from gnomad. *Human mutation*, 43(8):1012–1030, 2022.

Nilah M. Ioannidis, Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi,

Lisa A. Cannon-Albright, Craig C. Teerlink, Janet L. Stanford, William B. Isaacs, Jianfeng Xu, Kathleen A. Cooney, Ethan M. Lange, Johanna Schleutker, John D. Carpten, Isaac J. Powell, Olivier Cussenot, Geraldine Cancel-Tassin, Graham G. Giles, Robert J. MacInnis, Christiane Maier, Chih-Lin Hsieh, Fredrik Wiklund, William J. Catalona, William D. Foulkes, Diptasri Mandal, Rosalind A. Eeles, Zsofia Kote-Jarai, Carlos D. Bustamante, Daniel J. Schaid, Trevor Hastie, Elaine A. Ostrander, Joan E. Bailey-Wilson, Predrag Radivojac, Stephen N. Thibodeau, Alice S. Whittemore, and Weiva Sieh. Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum. Genet.*, 99(4):877–885, 2016. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2016.08.016. URL https://www.sciencedirect.com/science/article/pii/S0002929716303706.

O Isakov, R Fluss, and D Marek-Yagel. *The impact of clinical and molecular variant properties on calibration and performance of variant effect prediction tools*. 2024.

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, Eric D Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J Sanders, and Kyle Kai-How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, January 2019.

Shantanu Jain, Marta Soare, and Robert E. Schapire. Estimating the class prior and posterior from noisy positive and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2693–2701, 2016.

Shantanu Jain, Marena Trinidad, and etc. Nguyen, Thanh Binh. Evaluation of enzyme activity predictions for variants of unknown significance in arylsulfatase a. *Hum. Genet.*, 144(2-3):295–308, March 2025.

T. Jia and et al. Clinical gene selection strategies for functional genomics. *Human Genetics*, 2021. doi: 10.1007/s00439-021-02250-y.

Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Smooth isotonic regression: a new method to calibrate predictive models. *AMIA Summits on Translational Science Proceedings*, 2011:16, 2011.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel

Rhodes, Moriel Singer-Berk, Eleina M England, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, Genome Aggregation Database Consortium, Benjamin M Neale, Mark J Daly, and Daniel G MacArthur. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nat.*, 581(7809): 434–443, May 2020.

Da Kuang, Rebecca Truty, Jochen Weile, Britt Johnson, Keith Nykamp, Carlos Araya, Robert L Nussbaum, and Frederick P Roth. Prioritizing genes for systematic variant effect mapping. *Bioinform.*, 36(22-23):5448–5455, 12 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa1008. URL https://doi.org/10.1093/bioinformatics/btaa1008.

Da Kuang, Jochen Weile, Nishka Kishore, Maria Nguyen, Alan F Rubin, Stanley Fields, Douglas M Fowler, and Frederick P Roth. MaveRegistry: a collaboration platform for multiplexed assays of variant effect. *Bioinformatics*, 37(19):3382–3383, 03 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab215. URL https://doi.org/10.1093/bioinformatics/btab215.

Danny et al. Kuang. The difficulty-adjusted impact score: prioritizing genes for functional evidence generation. *Cell Genomics*, 1(3):100038, 2021. doi: 10.1016/j.cellgeni.2021.100038.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017.

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 radley Holmes, Brandi L Kattman, and Donna R Maglott. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, 46(D1):D1062–D1067, January 2018.

Fuyi Li, Yang Zhang, Anthony W Purcell, Geoffrey I Webb, Kuo-Chen Chou, Trevor Lithgow, Chen Li, and Jiangning Song. Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics*, 20(1):112, March 2019.

Brian Lucena. Spline-based probability calibration. *arXiv preprint arXiv:1809.07751*, 2018.

K. Matreyek and et al. Multiplexed measurement of variant effects. *Nature Genetics*, 2018. doi: 10.1038/s41588-018-0132-4.

William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biol.*, 17(1), December 2016.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 125–134, Lille, France, 2015. PMLR.

Vikas Pejaver, Alicia B Byrne, Bing-Jian Feng, Kymberleigh A Pagel, Sean D Mooney, Rachel Karchin, Anne O'Donnell-Luria, Steven M Harrison, Sean V Tavtigian, Marc S Greenblatt, et al. Calibration of computational tools for missense variant pathogenicity classification and clingen recommendations for pp3/bp4 criteria. *The American Journal of Human Genetics*, 109(12):2163–2177, 2022.

Vikas et al. Pejaver. Mutpred2: Inferring the molecular and phenotypic impact of amino acid variants. *Nature Communications*, 11(5918), 2020. doi: 10.1038/s41467-020-19669-x.

Vikas et al. Pejaver. Calibration of computational variant effect predictors for clinical interpretation of missense variants. *American Journal of Human Genetics*, 109(6):1089–1105, 2022. doi: 10.1016/j.ajhg.2022.04.012.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In Maria Florina Balcan and Kilian Q Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2052–2060, New York, New York, USA, 2016. PMLR.

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, 47(W1):W191–W198, 05 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz369. URL https://doi.org/10.1093/nar/gkz369.

Heidi L Rehm, Jonathan S Berg, Lisa D Brooks, Carlos D Bustamante, James P Evans, Melissa J Landrum, David H Ledbetter, Donna R Maglott, Christa Lese Martin, Robert L Nussbaum, Sharon E Plon, Erin M Ramos, Stephen T Sherry, Michael S Watson, and ClinGen. ClinGen–the clinical genome resource. *N. Engl. J. Med.*, 372(23):2235–2242, June 2015.

Verena Ricci, Mirella Filocamo, Stefano Regis, Fabio Corsolini, Marina Stroppiano, Marco Di Duca, and Rosanna Gatti. Expression studies of two novel in CIS-mutations identified in an intermediate case of hunter syndrome. *Am. J. Med. Genet. A*, 120A(1):84–87, July 2003.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl 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*, 17:405–424, 2015. doi: 10.1038/gim.2015.30.

P. A. Romero and et al. Navigating protein sequence–function space using directed evolution. *Nature Reviews Molecular Cell Biology*, 2015. doi: 10.1038/nrm3985.

Dace Ruklisa, James S Ware, Roddy Walsh, David J Balding, and Stuart A Cook. Bayesian models for syndrome-and gene-specific probabilities of novel variant pathogenicity. *Genome medicine*, 7:1–16, 2015.

Clayton Scott. A theory of learning with noisy positive and unlabeled examples. *Journal of Machine Learning Research*, 16(1):2961–2989, 2015.

Hyebin Song, Bennett J Bremer, Emily C Hinds, Garvesh Raskutti, and Philip A Romero. Inferring protein sequence-function relationships with large-scale positive-unlabeled learning. *Cell Syst.*, 12(1):92–101.e8, January 2021.

T. Stoeger, M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.*, 16(9):e2006643, 09 2018a.

Thomas Stoeger, Maximilian Gerlach, Richard I Morimoto, and Luís A.N. Amaral. Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biology*, 16(9), 2018b. doi: 10.1371/journal.pbio.2006643.

Sean V Tavtigian, Marc S Greenblatt, Steven M Harrison, Robert L Nussbaum, Snehit A Prabhu, Kenneth M Boucher, Leslie G Biesecker, ClinGen Sequence Variant Interpretation Working Group, et al. Modeling the acmg/amp variant classification guidelines as a bayesian classification framework. *Genetics in medicine*, 20(9):1054–1060, 2018.

Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G Biesecker. Fitting a naturally scaled point system to the acmg/amp variant classification guidelines. *Human mutation*, 41(10):1734–1737, 2020.

Sean V. et al. Tavtigian. Modeling the acmg/amp variant classification guidelines as a bayesian framework. *Genetics in Medicine*, 20:1054–1060, 2018. doi: 10.1038/gim.2017.210.

Sean V. et al. Tavtigian. A bayesian framework for acmg/amp variant classification criteria supports evidence strength scaling. *Human Mutation*, 2020. doi: 10.1002/humu.24065.

Malvika Tejura, Shawn Fayer, Abbye E McEwen, Jake Flynn, Lea M Starita, and Douglas M Fowler. Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. *The American Journal of Human Genetics*, 111(9):2031–2043, 2024.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. 2023.

Nicola Whiffin, Roddy Walsh, Risha Govind, Matthew Edwards, Mian Ahmad, Xiaolei Zhang, Upasana Tayal, Rachel Buchan, William Midwinter, Alicja E Wilk, et al. Cardioclassifier: disease-and gene-specific computational decision support for clinical genome interpretation. *Genetics in Medicine*, 20(10):1246–1254, 2018.

Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, October 2012.

Yaping Yang, Donna M Muzny, Jeremy G Reid, Matthew N Bainbridge, Ashley Willis, Patricia A Ward, and et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312(18):1870–1879, 2017. doi: 10.1001/jama.2014.14601.

Daniel Zeiberg, Shantanu Jain, and Predrag Radivojac. Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6729–6736, 2020.

Daniel Zeiberg, Shantanu Jain, and Predrag Radivojac. Leveraging structure for improved classification of grouped biased data. 2022.

Daniel Zeiberg, Malvika Tejura, Abbye E McEwen, Shawn Fayer, Vikas Pejaver, Alan F Rubin, Lea M Starita, Douglas M Fowler, Anne O'Donnell-Luria, and Predrag Radivojac. Gene-based calibration of high-throughput functional assays for clinical variant classification. *bioRxiv*, pages 2025–04, 2025.

Daniel et al. Zeiberg. Distcurve: Estimating class prior from positive and unlabeled data using distances. *NeurIPS*, 2020.