Gene-based calibration of high-throughput functional assays for clinical variant classification

Daniel Zeiberg,[1] Malvika Tejura,[2] Abbye E. McEwen,[2,3,4] Shawn Fayer,[2] Vikas Pejaver,[5] Alan F. Rubin,[6] Lea M. Starita,[2,3] Douglas M. Fowler,[2,3,7] Anne O'Donnell-Luria,[8,9] Predrag Radivojac[1,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA; [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [3]Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA 98195, USA; [4]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA 98195, USA; [5]Institute for Genomic Health and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [6]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia; Department of Medical Biology, University of Melbourne, Melbourne, VIC, Australia; [7]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA; [8]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA02142, USA; [9]Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA. *Correspondence: predrag@northeastern.edu.

Abstract

High-throughput functional assays measure the effects of variants on macromolecular function and can aid in reclassifying the rapidly growing number of variants of uncertain significance. Under the current clinical variant classification guidelines, using functional data as a line of evidence to assert pathogenicity relies on determining assay score thresholds that define variants as functionally normal or functionally abnormal. These thresholds are designed to maximize the separation of variants with known clinical effects (benign, pathogenic) and often incorporate expert opinion. However, this approach lacks the rigor of calibration, in which a variant's posterior probability of pathogenicity must be estimated from the raw experimental score and mapped to discrete evidence strengths. To build upon the existing guidelines, we introduce and evaluate a method for calibrating continuous high-throughput functional data as a line of evidence in clinical variant classification. Assay score distributions of synonymous variants and variants appearing in gnomAD for a given functional scoreset are jointly modeled with score distributions of known pathogenic and benign variants using a multi-sample skew normal mixture of distributions. This model is learned using a constrained expectation-maximization algorithm that provably preserves the monotonicity of pathogenicity posteriors and is subsequently used to calculate variant-specific evidence strengths for use in the clinic. Using 24 datasets from 14 genes, we first assess the model's ability to capture assay score distributions. We then demonstrate its potential impact on reclassifying variants by comparing the evidence strengths assigned at the variant-level with those assigned uniformly to all functionally normal and abnormal variants under the existing ClinGen guidelines. An improved classification of variants will directly improve the accuracy of genetic diagnosis and subsequent medical management for individuals affected by Mendelian disorders.

Availability: https://github.com/dzeiberg/mave_calibration

1 Introduction

Genetic diagnosis is the cornerstone of genomic medicine for rare disease [46, 49]. It aims to identify one or a handful of causal variants in a patient's genome based on the wealth of sequence, functional, and clinical data. To standardize the approach, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) maintain evidence-based guidelines for classifying variants in genes causally related to Mendelian disease [45]. A variant reported to a clinician can be classified into one of the five categories: "pathogenic", indicating a 99% probability it causes a Mendelian disease; "likely pathogenic", indicating a 90% probability it causes disease; "likely benign", indicating a 90% probability it does not cause disease; "benign", indicating a 99% probability it does not cause disease; and "variant of uncertain significance" (VUS), covering the intermediate range of probabilities. A variant's classification is conditioned on lines of available evidence supporting or refuting its disease causality. The 2015 guidelines designate the type (e.g., functional, computational, population, or segregation data), strength (supporting, moderate, strong, very strong, standalone), and direction (benign, pathogenic) of each line of evidence, and include a set of rules for combining evidence to classify a variant into one of the five categories [45]. To be classified as likely pathogenic or pathogenic, variants must have at least two lines of evidence of sufficient strength (e.g., two lines of strong evidence for likely pathogenic) that combine to support their clinical classification. These variants are considered diagnostic and likely to be clinically actionable [44].

The ACMG/AMP system has been approximated by a single-parameter Naïve Bayes classifier in which the strength of evidence is quantified using its positive likelihood ratio (LR^+), also referred to as OddsPath, where discrete evidence strengths (supporting, moderate, etc.) have exponentially increasing LR^+ [51]. However, implementing this probabilistic framework necessitates methodology for calibrating individual lines of evidence, which in turn requires estimation of LR^+ as well as the prior probability of pathogenicity in a reference variant set. For each evidence source, calibration aims to map a concrete score (e.g., computational prediction, case-control counts, functional assay readout) to discrete evidence of a specific direction and strength to support the ACMG/AMP rule-based evidence combination.

Pejaver et al. [39] proposed a nonparametric method for calibrating computational predictors for missense variation by estimating a rare variant's prior probability of pathogenicity in disease-associated genes as well as local positive likelihood ratios (lr^+) of real-valued scores using a validated set of pathogenic and benign variants from ClinVar [26]. They demonstrated that several models exceed the supporting evidence strength assigned to computational predictors in the initial guidelines, resulting in updated ClinGen recommendations that allow these predictors to reach evidence levels up to strong [39, 4]. Subsequent studies have shown that the calibration works as intended in a genetic diagnosis setting [48], although the performance levels vary across genes [53].

Despite the increasing discriminative power of computational tools over time [54, 43], predictors alone are not sufficient to classify a rare variant as pathogenic and, thus, additional lines of evidence are needed [45]. Functional assays, including multiplexed assays of variant effect, are being deployed [20] to generate the evidence needed to reclassify the rapidly increasing number of VUS and prospectively quantify the effect of variants yet to be observed in the clinic [50, 15]. To standardize the use of functional evidence across clinical laboratories, the ClinGen [44] Sequence Variant Interpretation (SVI) Working Group established recommendations for using functional data in clinical variant classification [6]. The recommendations included steps for determining the validity and strength of functional evidence that can be applied to an individual variant. All variants with assay scores exceeding the specified thresholds for functionally normal or abnormal are assigned a fixed evidence strength, determined by expert opinion or by the level of separation of known pathogenic

and benign variants. Such threshold determination is a potential source of inconsistency in the use of functional evidence. Additionally, the grouping of all variants exceeding thresholds may over- or under-estimate evidence strength for variants near and far from the thresholds.

A distinction between calibrating computational predictors and high-throughput functional assays is that, unlike predictor scores, assay scores cannot readily be compared across individual studies, as each assay score quantifies a particular function of a variant relative to the function of the wild type. The often limited number of clinical control variants in each gene poses additional challenges, making it difficult to repurpose the previously proposed method for predictor calibration. Finally, in addition to assaying a large number of missense variants, a modern high-throughput functional experiment usually includes a set of assayed synonymous variants. The scores of these synonymous variants can be exploited to model the score distribution of functionally neutral variation as well as the experimental variability of assay scores.

To increase consistency in the use of functional evidence across clinical laboratories, we here propose a statistical method for calibrating high-throughput functional data. We focus on a curated set of 24 functional datasets that quantify the functional effect of missense variants in 14 genes. We demonstrate that this methodology determines assay score thresholds that define variant-level evidence strength for these datasets and highlight the method's potential for reclassifying variants.

2 Materials and Methods

The objective in calibrating high-throughput functional data for a given gene or domain is to quantify the strength of evidence an assay score s for a variant v can serve in classifying v as pathogenic or benign. That is, we seek to learn a mapping function $f : \mathbb{R} \to \mathcal{S} \times \mathcal{D}$ from an assay score $s \in \mathbb{R}$ into one of the discrete evidence strengths $\mathcal{S} = \{$supporting, moderate, strong, very strong$\}$ and directions $\mathcal{D} = \{$benign, pathogenic$\}$. Variants whose scores do not map into any of the categories above are said to have "indeterminate" strength and direction [39].

2.1 Calibration framework

To learn the calibration function, we will follow the general probabilistic framework based on the relationship between prior and posterior odds of pathogenicity [17] as

$$\text{posterior odds of pathogenicity} = \text{ positive likelihood ratio } \times \text{ prior odds of pathogenicity.}$$

Letting E be a discrete random variable representing evidence and Y be a binary random variable representing whether ($Y = 1$) or not ($Y = 0$) a variant causes a rare disease, the posterior odds of pathogenicity can be expressed as $P(Y=1|E=e)/P(Y=0|E=e)$, the prior odds of pathogenicity as $P(Y=1)/P(Y=0)$, and the positive likelihood ratio $\mathrm{LR}^+(e)$ as $P(E=e|Y=1)/P(E=e|Y=0)$. In the case evidence is produced by a single binary predictor, LR^+ is equivalent to the ratio of true positive rate and false positive rate.

To model the expert-derived rules from the ACMG/AMP guidelines [45], Tavtigian et al. [51] proposed to express LR^+ in an exponential form as

$$\mathrm{LR}^+ = c^{\frac{n_{\mathrm{su}}}{8} + \frac{n_{\mathrm{mo}}}{4} + \frac{n_{\mathrm{st}}}{2} + \frac{n_{\mathrm{vs}}}{1}}, \tag{1}$$

where c is a parameter and n_{su}, n_{mo}, n_{st} and n_{vs} are the numbers of supporting, moderate, strong, and very strong lines of evidence, respectively. Equation (1) specifies that one line of very strong evidence equals two lines of strong, four lines of moderate and eight lines of supporting evidence.

Additionally, the value of LR^+ for one line of supporting, moderate, strong, and very strong evidence equals $\sqrt[8]{c}$, $\sqrt[4]{c}$, $\sqrt{c}$ and c, respectively. This can be interpreted as a point-based system in which points are measured on the $\log \mathrm{LR}^+$ scale and supporting, moderate, strong, and very strong evidence of pathogenicity/benignity are designated $\pm 1, \pm 2, \pm 4, \pm 8$ points, respectively, with positive points indicating pathogenic evidence and negative points denoting benign evidence [52]. Under this interpretation, it is straightforward to incorporate intermediate evidence strengths between those established (e.g., ± 3 points).

If the prior probability of pathogenicity $P(Y = 1)$ is known, the parameter c is determined as the smallest number for which the posterior probability of pathogenicity

$$P(Y = 1|E = e) = \frac{\mathrm{LR}^+(e) \cdot P(Y = 1)}{(\mathrm{LR}^+(e) - 1) \cdot P(Y = 1) + 1} \tag{2}$$

satisfies all qualitative criteria from Richards et al. [45]. That is, $P(Y = 1|E = e)$ must reach at least 0.9 for variants classified as likely pathogenic and 0.99 for variants classified as pathogenic. A prior of $P(Y = 1) = 0.1$ implies $c = 350$ [51], whereas $P(Y = 1) = 0.044$ implies $c = 1124$ [39].

When the available evidence is real-valued, it is necessary to discretize it for compatibility with ACMG/AMP guidelines [45]. Letting S be a real-valued random variable representing a variant's assay score, we can estimate the local positive likelihood ratio $\mathrm{lr}^+(s)$ at score s as the density ratio $p(s|Y=1)/p(s|Y=0)$ and establish score intervals $[\tau_{\mathrm{low}}, \tau_{\mathrm{high}}]$ such that $\mathrm{lr}^+(s)$ is greater than the required level for any score in the interval [39]. For example, for supporting evidence, it is required that $\mathrm{lr}^+(s) \geq \sqrt[8]{c}$, $\forall s \in [\tau_{\mathrm{su}}, \tau_{\mathrm{mo}}]$, while for moderate evidence, it is required that $\mathrm{lr}^+(s) \geq \sqrt[4]{c}$, $\forall s \in [\tau_{\mathrm{mo}}, \tau_{\mathrm{st}}]$. Considering functional assay as a predictor, $\mathrm{lr}^+(s)$ can be seen as the slope of the ROC curve at the predictor's score s [54].

To calibrate *in silico* tools, Pejaver et al. [39] combined the data from all disease genes and used separate nonparametric methods to estimate $P(Y = 1)$ and $\mathrm{lr}^+(s)$. However, this method relies on a relatively large number of clinical control variants and cannot be reliably applied to most individual genes. Stronger assumptions are therefore necessary, leading us to develop a parametric approach to simultaneously estimate $P(Y = 1)$ and $\mathrm{lr}^+(s)$ in a gene-specific manner.

2.2 Data

A curated set of 24 functional datasets were collected from MaveDB [12] and directly from recent publications. A typical dataset from MaveDB contains scores measuring the functional effect of a number of variants, including single and multiple nucleotide substitutions, insertions and deletions. These scores can reflect the effect of the variant on a particular function, such as stability or abundance, or reflect the overall functional fitness of the variant, and datasets frequently contain score measurements for biological or technical replicates. Additionally, some functional datasets include the authors' categorical interpretation of each variant, annotating them as functionally normal, indeterminate, or functionally abnormal.

In this study, single-nucleotide variants were collected from ClinVar (accessed 09/17/2024) [26]. Subsets of variants, referred to as ClinVar P/LP, ClinVar B/LB, and ClinVar VUS, were constructed by selecting variants with review status of at least one star and clinical significance of pathogenic (P) or likely pathogenic (LP), benign (B) or likely benign (LB), and uncertain significance (VUS), respectively. Single-nucleotide variants in gnomAD genomes and exomes (v4.1.0) [25] were collected, excluding filtered sites using GATK (v4.4.0.0) [31]. Missense variants in each experimental dataset were labeled to indicate their presence in the ClinVar P/LP, B/LB, VUS, and gnomAD sets, and synonymous variants were labeled as such. The intersections of each dataset with the ClinVar and

Gene	Experiment	P/LP	B/LB	gnomAD	Synonymous	VUS
BRCA1	Cisplatin resistance assay [1]	61	31	119	0	150
BRCA1	HDR assay [1]	59	47	158	0	230
BRCA1	Saturation genome editing [13]	64	86	226	544	287
BRCA2	HDR assay [19]	49	45	240	0	101
MSH2	Resistiance assay [22]	114	273	1460	0	1598
TP53	Transcriptional Assay [23]	201	132	555	0	652
TP53	Growth assay- p53 WT [16]	213	136	570	376	667
TP53	Growth assay- p53 NULL [16]	213	136	570	376	667
TP53	Growth assay- etoposide [16]	213	136	570	376	667
PTEN	Lipid phosphatase activity assay [32]	205	8	267	0	685
PTEN	VAMP-seq [30]	128	6	157	156	415
VHL	Saturation genome editing [7]	65	15	357	386	180
BAP1	Saturation genome editing [57]	11	976	1714	0	1145
BRCA1	HDR assay [47]	36	36	269	0	294
BRCA2	Saturation prime editing [11]	5	2	61	106	32
NPC1	Saturation prime editing- HEK293T [11]	28	1	124	228	44
TPK1	Yeast complementation [58]	6	1	302	572	91
TP53	HDR assay [5]	210	136	567	0	665
HMBS	Yeast complementation [55]	52	5	488	306	171
PRKN	VAMP-seq [8]	15	14	849	1	108
OTC	Yeast growth assay [28]	93	10	153	0	82
RAD51C	Saturation genome editing [36]	4	407	900	0	786
MSH2	Mutation rate assessment [35]	34	7	89	0	57
CALM1	Yeast complementation [58]	81	8	401	296	183

Table 1: Number of single nucleotide variants with P/LP, B/LB, gnomAD, Synonymous, and VUS labels in data used in this work. A mean score was used for all variants with multiple replicates. To limit circularity in the analysis, as the dataset has contributed to a substantial number of classifications, ClinVar classifications for the BRCA1 Saturation Genome Editing dataset [13] are limited to those submitted prior to the publication (release 05/05/2018). Additionally, as this dataset is mapped to GRCh37 coordinates, gnomAD v2.1.1 was used in place of gnomAD v4.1.0

gnomAD sets were constructed by matching variants on the most precise description available, their consequence relative to the GRCh38 genome build, transcript, or protein sequence.

For assays that do not detect splice-altering effects of variants, any variant with a SpliceAI score [21], using the maximum of the acceptor loss/gain and donor loss/gain predictions, greater than a strict threshold of 0.5 [56] was removed, as it is unknown if any associated pathogenic effect is mediated by the missense or the splicing change. The maximum SpliceAI score for the given consequence was used when filtering datasets in which variants are only described at the protein level. For datasets with multiple replicate scores reported, the average score was computed. Author-reported scores were collected without further normalization, excluding the TP53 Transcriptional Assay dataset [23]. For this dataset, the median score across the eight promoters was computed, as done by Fortuno et al. [14], and median scores were normalized to the B/LB variant score distribution. Details on all datasets can be found in Table 1 and in the Supplement.

2.3 Skew normal distributions

Assay scores are modeled using skew normal distributions. Under the distribution's canonical parameterization [2], the probability density function of a random variable $X \sim \mathrm{SN}(\mu, \omega, \lambda)$ is given by

$$\mathrm{SN}(x; \mu, \omega, \lambda) = \frac{2}{\omega}\phi\left(\frac{x-\mu}{\omega}\right)\Phi\left(\lambda\left(\frac{x-\mu}{\omega}\right)\right),$$

where μ, ω, and λ denote the location, scale and skew of the distribution, and ϕ and Φ are the probability density and cumulative distribution functions of the standard Gaussian distribution, respectively. Under the alternate parameterization, defined by parameters Δ and Γ, used in the parameter updates of the optimization algorithm, $X \sim \mathrm{SN}(\mu, \omega, \lambda)$ can be expressed as $X = \mu + \Delta \cdot T + \Gamma^{\frac{1}{2}} \cdot U$, where T is a random variable from the standard normal distribution truncated below 0 and U is a random variable from the standard normal distribution [27, 40]. Canonical and alternate parameterizations are related through equations in Table 1 of [40], and are included in Supplementary Table S1 for completeness.

Skew normal distributions were chosen as they are able to model the asymmetry observed in functional datasets and have been found to be tractably incorporated into multi-sample mixture models [40].

2.4 Modeling

Assay score distributions from each experimental dataset are approximated using a multi-sample two-component skew normal mixture model. Let S_P be the set of assay scores for variants in the dataset's pathogenic and likely pathogenic (P/LP) sample, let S_B be the set of assay scores for variants in the dataset's benign and likely benign (B/LB) sample, and let S_G be the set of assay scores for variants in the dataset's gnomAD sample, used here as a reference sample. The score distribution of each of these random variables is modeled as a two-component skew normal mixture representing functionally abnormal (a) and functionally normal (n) variant assay scores parameterized by $\theta_a = (\mu^{(a)}, \omega^{(a)}, \lambda^{(a)})$ and $\theta_n = (\mu^{(n)}, \omega^{(n)}, \lambda^{(n)})$, respectively. The mixture model approximating the distributions of S_P, S_B, and S_G can therefore be expressed as

$$p_{S_i}(s; w_{S_i}, \theta_a, \theta_n) = w_{S_i} \cdot \mathrm{SN}(s; \theta_a) + (1 - w_{S_i}) \cdot \mathrm{SN}(s; \theta_n), \tag{3}$$

where $i \in \{P, B, G\}$.

The local positive likelihood ratio of the model is defined as

$$\mathrm{lr}^+(s; w_{S_P}, w_{S_B}, \theta_a, \theta_n) = \frac{p_{S_P}(s; w_{S_P}, \theta_a, \theta_n)}{p_{S_B}(s; w_{S_B}, \theta_a, \theta_n)}.$$

The prior probability of pathogenicity $P(Y = 1)$ can be obtained by expressing the reference gnomAD distribution as a two-component mixture of pathogenic and benign score distributions

$$S_G \sim p(s) = P(Y = 1) \cdot p(s|Y = 1) + P(Y = 0) \cdot p(s|Y = 0),$$

and then rearranging Eq. (3) to obtain

$$P(Y = 1) = \frac{w_{S_G} - w_{S_B}}{w_{S_P} - w_{S_B}}. \tag{4}$$

Finally, if provided, the sample of synonymous variant assay scores is also modeled as a mixture of the functionally abnormal and functionally normal score distributions to aid in the optimization,

$$p_{S_S}(s; w_{S_S}, \theta_a, \theta_n) = w_{S_S} \cdot \mathrm{SN}(s; \theta_a) + (1 - w_{S_S}) \cdot \mathrm{SN}(s; \theta_n).$$

Synonymous variant scores were available for 11 of 24 datasets (Table 1).

2.4.1 Optimization

The parameters of our multi-sample skew normal mixture model can be jointly learned by adapting the expectation-maximization algorithm derived by Peng et al. [40]. With $\ddot{}$ and $\bar{}$ denoting new and old parameter values, respectively, the parameter update equations can be expressed as

$$\ddot{\mu}^{(*)} = \frac{\sum_{S_i \in \{S_P, S_B, S_G, S_S\}} \sum_{s_{ij} \in S_i} \bar{p}_{S_i}^{(*)}(s_{ij}) \cdot \bar{m}^{(*)}\left(s_{ij}, \bar{\Delta}^{(*)}\right)}{\sum_{S_i \in \{S_P, S_B, S_G, S_S\}} \sum_{s_{ij} \in S_i} \bar{p}_{S_i}^{(*)}(s_{ij})}$$

$$\ddot{\Delta}^{(*)} = \frac{\sum_{S_i \in \{S_P, S_B, S_G, S_S\}} \sum_{s_{ij} \in S_i} \bar{p}_{S_i}(s_{ij}) \bar{d}^{(*)}(s_{ij}, \ddot{\mu}^{(*)})}{\sum_{S_i \in \{S_P, S_B, S_G, S_S\}} \sum_{s_{ij} \in S_i} \bar{p}_{S_i}^{(*)}(s_{ij})}$$

$$\ddot{\Gamma}^{(*)} = \frac{\sum_{S_i \in \{S_P, S_B, S_G, S_S\}} \sum_{s_{ij} \in S_i} \bar{p}_{S_i}^{(*)}(s_{ij}) \bar{g}^{(*)}(s_{ij}, \ddot{\mu}^{(*)}, \ddot{\Delta}^{(*)})}{\sum_{S_i \in \{S_P, S_B, S_G, S_S\}} \sum_{s_{ij} \in S_i} \bar{p}_{S_i}^{(*)}(s_{ij})}$$

$$\ddot{w}_{S_i} = \sum_{s_{ij} \in S_i} \bar{p}_{S_i}^{(*)}(s_{ij}), \forall i \in \{P, B, G, S\}$$

with $* = a$ or n. $\bar{p}_{S_i}^{(a)}$ and $\bar{p}_{S_i}^{(n)}$ are defined as $\bar{p}_{S_i}^{(a)}(s) = \frac{\bar{w}_{S_i} \cdot \mathrm{SN}(s; \bar{\theta}_a)}{\bar{w}_{S_i} \cdot \mathrm{SN}(s; \bar{\theta}_a) + (1 - \bar{w}_{S_i}) \cdot \mathrm{SN}(s; \bar{\theta}_n)}$ and $\bar{p}_{S_i}^{(n)}(s) = 1 - \bar{p}_{S_i}^{(a)}(s)$. $\bar{m}^{(*)}(s, \Delta)$, $\bar{d}(s, \mu)$, and $\bar{g}^{(*)}(s, \mu, \Delta)$ are defined in Supplementary Table S2.

2.4.2 Density constraint

We enforce a monotonicity constraint on the density ratio $\mathrm{SN}(s;\theta_a)/\mathrm{SN}(s;\theta_n)$ throughout the optimization. Using the convention of lower scores indicating functionally abnormal variants and higher scores indicating functionally normal variants, this constraint can be stated formally as $\forall s_1, s_2$ such that $s_1 \leq s_2$ implies $\mathrm{SN}(s_1;\theta_a)/\mathrm{SN}(s_1;\theta_n) \geq \mathrm{SN}(s_2;\theta_a)/\mathrm{SN}(s_2;\theta_n)$. Enforcing this constraint alongside the assumption that pathogenic variants are more likely to be functionally abnormal than benign variants, verified during assay validation [6], is sufficient for the positive likelihood ratio of our model to be non-increasing.

Theorem 1. *Given a monotonic density ratio $SN(s;\theta_a)/SN(s;\theta_n)$ and $w_{S_P} > w_{S_B}$, $p(s|Y=1)/p(s|Y=0)$ and $P(Y=1|s)$ are also monotonic.*

The proof is in the Supplement. This theorem guarantees that stronger assay scores always lead to stronger evidence of pathogenicity. As is done by Peng et al. [41], a binary search is used to maintain this constraint at each update to the parameters $\mu^{(a)}, \Delta^{(a)}, \Gamma^{(a)}, \mu^{(n)}, \Delta^{(n)}$, and $\Gamma^{(n)}$.

2.4.3 Initialization

To initialize the parameters of the skew-normal distributions, θ_a and θ_n, K-means clustering ($K = 2$) is fit on all observations from our samples. Normal distributions are fit to the sets of observations in each of the two clusters to obtain $\mu^{(a)}$, $\omega^{(a)}$, $\mu^{(n)}$ and $\omega^{(n)}$. As it has been observed empirically that the direction of the skew does not change during the iterations of the EM algorithm [40], and to ensure a diverse range of initializations, $\lambda^{(a)}$ and $\lambda^{(n)}$ are each sampled randomly from the uniform distribution $\mathbb{U}(-0.25, 0.25)$. To ensure the initial parameters satisfy the density constraint, checked within the range of our observations, the locations of each component are iteratively moved in opposing directions until θ_a and θ_n satisfy the constraint. To mitigate the effects of random parameter initialization, we performed 100 independent runs of the modeling procedure, retaining the model that achieved the highest likelihood.

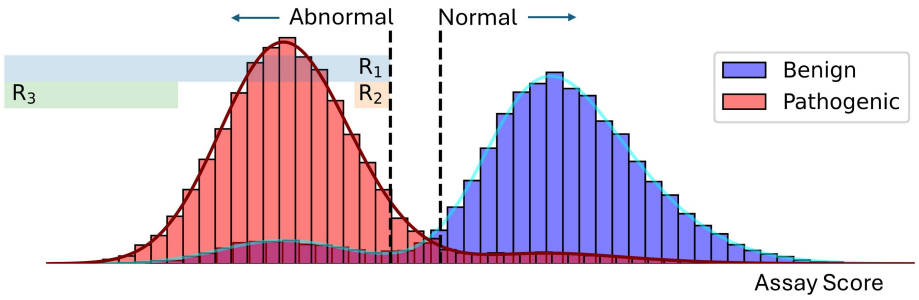

Figure 1: Illustrating the current use of functional data in clinical variant classification. ACMG/AMP recommendations rely on thresholds defining functionally normal and functionally abnormal variants given a set of experimental scores. The evidence strength for variants exceeding these thresholds is computed using pathogenic (red) and benign (blue) control sets. The pathogenic evidence strength is computed as the ratio of true positive rate and false positive rates in region R_1. As functional data appear as continuous scores, as opposed to binary functional labels, assigning evidence strengths at the variant level can result in more accurate use of functional evidence. Variants with scores in region R_3, far below the threshold, should be assigned higher evidence strengths than those assigned to variants with scores in region R_2, slightly below the threshold.

2.4.4 Score thresholds and uncertainty quantification

To quantify uncertainty and obtain conservative evidence strength estimates, the modeling procedure is repeated 5,000 times for each dataset on bootstrapped [10] samples of the data, ensuring the P/LP, B/LB, gnomAD, and synonymous observation sets are disjoint. Assay score thresholds for each bootstrap iteration are calculated with respect to the prior probability of pathogenicity estimated by the model in the iteration. Score thresholds are aggregated across bootstrap iterations, using the 5th and 95th percentiles as the final score thresholds defining each pathogenic and benign evidence strength, respectively. An assay reaches a given evidence strength if at least 95% of the bootstrap iterations reach that strength. Models for which bootstrapping leads to invalid prior estimates are excluded when aggregating thresholds.

3 Results

Functional data are one of the few lines able to serve as very strong evidence of pathogenicity. Brnich et al. [6] refined the use of functional data in variant classification, recommending steps to assess the clinical validity of functional data and proposing a rule-based system for assigning evidence strengths. In this system, the assay score thresholds that define functionally normal and abnormal variants are determined using a set of controls. These thresholds are used to categorize variants, conventionally as functionally normal, functionally abnormal, or indeterminate, and the evidence strength assignment for all variants exceeding these thresholds is computed using global positive likelihood ratio; i.e., the ratio of true positive rate to false positive rate in a score interval. Assigning evidence strengths in this manner for well-behaved assays overestimates the strength of functional evidence for variants scoring near the thresholds and underestimates the strength of evidence for variants far exceeding the thresholds, as illustrated in Figure 1. This motivates the need for evidence strengths to be assigned at a variant level, taking into consideration the variant's raw (continuous) assay score. Assigning evidence strengths at the variant level, using the local positive likelihood ratio, should result in more reliable evidence assignments and more accurate

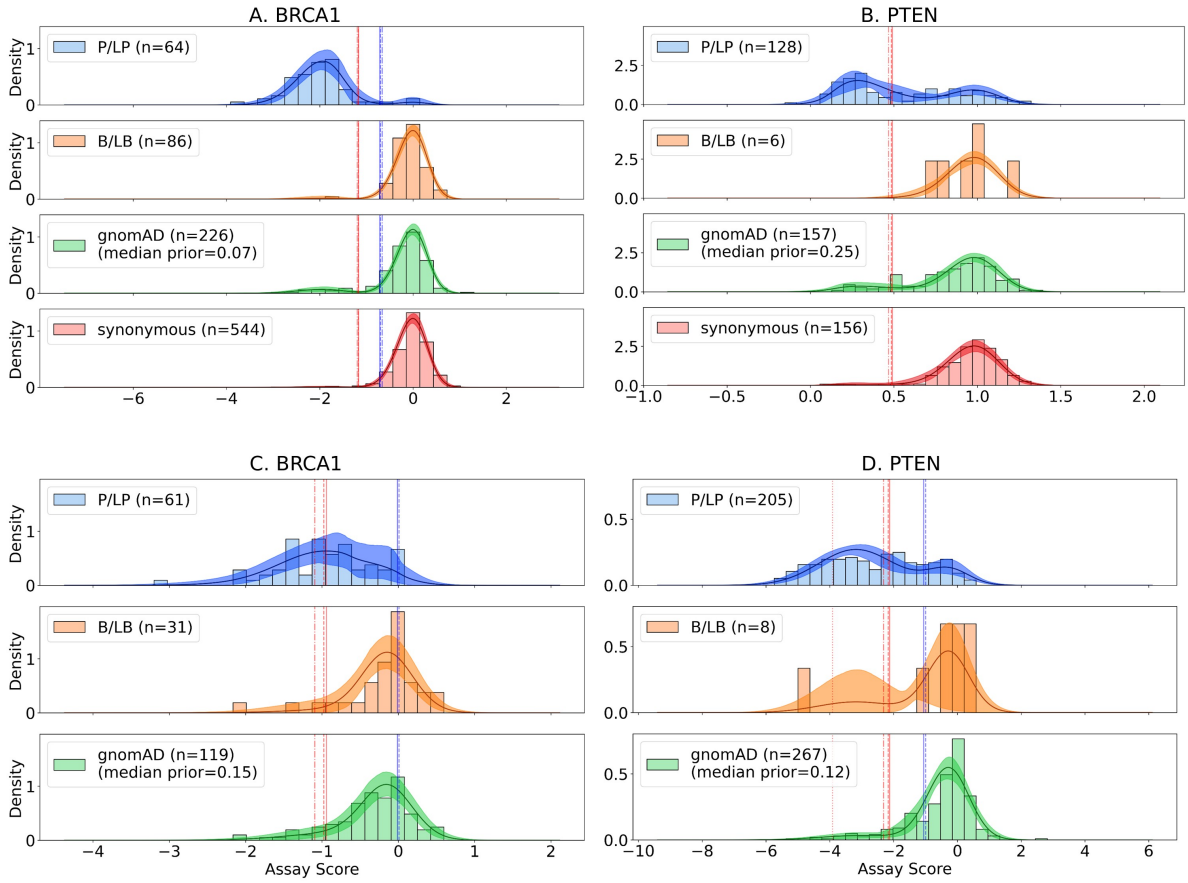

Figure 2: Example model fits to four functional datasets, assessing the effect of BRCA1 variants (panels A and C; left) and PTEN variants (panels B and D; right). Datasets for panels A, B, C, and D were taken from Findlay at al. [13], Matreyek et al. [30], Adamovich et al. [1], and Mighell et al. [32], respectively. Median density across all iterations is displayed along with the estimated 95% confidence interval. Score thresholds corresponding to each pathogenic and benign evidence strength reached by the model are plotted in red and blue, respectively, with supporting, moderate, strong, and very strong displayed as solid, dashed, dashed-dotted, and dotted lines, respectively. Differences in the gnomAD variants included in each dataset contribute to the differences in prior probability estimates.

variant classification.

3.1 The model captures a variety of assay score distributions

To demonstrate the utility of our calibration method, we visualized the distributions and score thresholds given by the model on four datasets in Figure 2. The learned models capture the distributions of the data and justify the selection of the skew normal family. Figure 2A shows that the calibration model assigns pathogenic and benign evidence to functional datasets in which the assay separates the P/LP and B/LB distributions. In contrast, Figure 2B demonstrates the calibration model appropriately does not assign benign evidence to variants with mechanisms of pathogenicity not captured by the particular assay. This reflects limitations of the VAMP-seq assay which assesses the impact on protein stability and does not measure all types of functional impacts

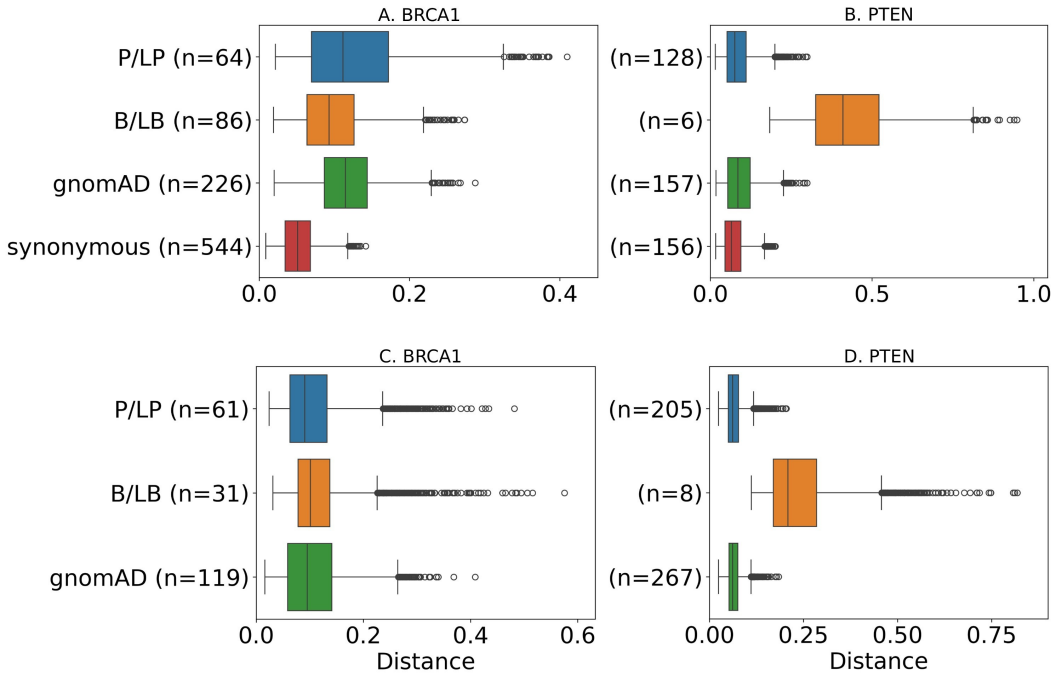

Figure 3: Distributions of normalized distances between empirical and estimated cumulative distribution functions for each sample, obtained using 5,000 bootstrapping iterations, for the same datasets as in Figure 2. Datasets for panels A, B, C, and D were taken from Findlay at al. [13], Matreyek et al. [30], Adamovich et al. [1], and Mighell et al. [32], respectively.

such as enzymatic activity [33], thus leaving a significant proportion of pathogenic variants with scores that are similar to those of synonymous variants [30]. This calibration method based on skew normal models reflects what we intuitively understand about the strengths and limitations of these assays.

To quantify the quality of the model fit, for each bootstrap iteration, the normalized Yang et al. [59] distance ($p = 2$) of the empirical and learned cumulative distribution functions were computed for the P/LP, B/LB, gnomAD, and synonymous samples. The distance function gives values between 0 and 1 and is computed using real-valued vectors at all unique scores in the assay. Figure 3 visualizes the distribution of distances across the 5,000 bootstrap iterations for each sample. The distances computed on the gnomAD and synonymous samples can serve as a metric to decide whether to accept the model fit. Empirically, we observed the median distance computed on these two samples was typically below 0.2. Being smaller and potentially accumulated in a biased fashion, the distances computed on the P/LP and B/LB samples are generally larger and not as reliable in evaluating the quality of the model fit. Analysis of additional datasets from Table 1 is presented in the Supplement.

3.2 Variant-level evidence strengths vs. author-provided annotations

We compared the distributions of variant-specific evidence strengths assigned by the calibration model to the functional annotations provided by the authors (Figure 4). Evidence strengths were assigned in an out-of-bag manner; i.e., the mode of the strengths assigned to each variant by all models for which the variant was not used to fit. Of the variants annotated by the study authors as functionally abnormal, 60% were assigned pathogenic evidence of some strength, and 87% of

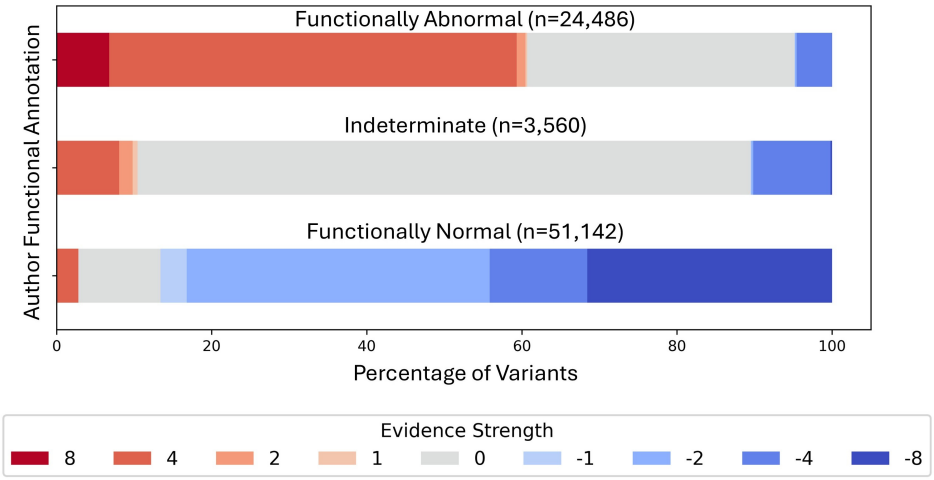

Figure 4: Distribution of evidence strengths assigned by our method for all variants with functional annotations provided by the authors of experimental studies (Table 1). Author annotations were harmonized to a consistent vocabulary. Annotations of "enriched" from a single dataset, Lo et al. [28], were excluded due to the limited number of variants.

variants annotated as functionally normal were assigned benign evidence of some strength by our model. These reductions are consistent with the expectation that discretizing assay scores prior to calculating evidence strengths results in an overestimation of evidence strengths. Fifty-two percent of variants annotated as functionally abnormal by authors reached the level of pathogenic strong (+4), and 7% of variants annotated as functionally abnormal reached the level of pathogenic very strong (+8) using the model, exceeding the default strength specified in Richards et al. [45]. Thirty-two percent of variants annotated by authors as functionally normal reached benign very strong (-8), exceeding the current default strength of benign strong (-4).

3.3 Variant-level evidence improves clinical variant classification

The out-of-bag evidence strength distributions for variants present in ClinVar and gnomAD are visualized in Figure 5. Of the $2,160$ P/LP variants, 67% were assigned pathogenic evidence strengths, and 92% of the $2,652$ B/LB variants were assigned benign evidence strengths. Eighty-one percent of the 9,957 VUS were assigned either pathogenic or benign evidence strengths, with 66% assigned benign evidence strengths and 15% assigned pathogenic evidence strengths, broadly consistent with previous experience [18].

To demonstrate the impact this calibration method could have on variant classification, we categorized our variant-level evidence strength assignments based on the direction of evidence and compared these to available author-provided functional annotations on the subset of variants with P/LP or B/LB ClinVar classifications. Figure 6 shows the number of P/LP and B/LB variants in ClinVar that were assigned pathogenic or benign evidence directionality (Figure 6A) vs. the number of variants that were assigned author-provided functional annotations (Figure 6B). We summarized these results by calculating the global positive likelihood ratio (LR^+) and diagnostic odds ratio (DOR) for our method vs. author-provided functional annotations. To do so, we first dichotomized the three-way annotations for our method as pathogenic vs. non-pathogenic and obtained $LR^+ = \frac{710}{710+195+92} \cdot \frac{31+79+1833}{31} = 44.6$. We then dichotomized the three-way author-provided annotations as functionally abnormal vs. the rest and obtained $LR^+ = \frac{724}{724+110+163} \cdot \frac{99+28+1816}{99} = 14.3$. Fur-

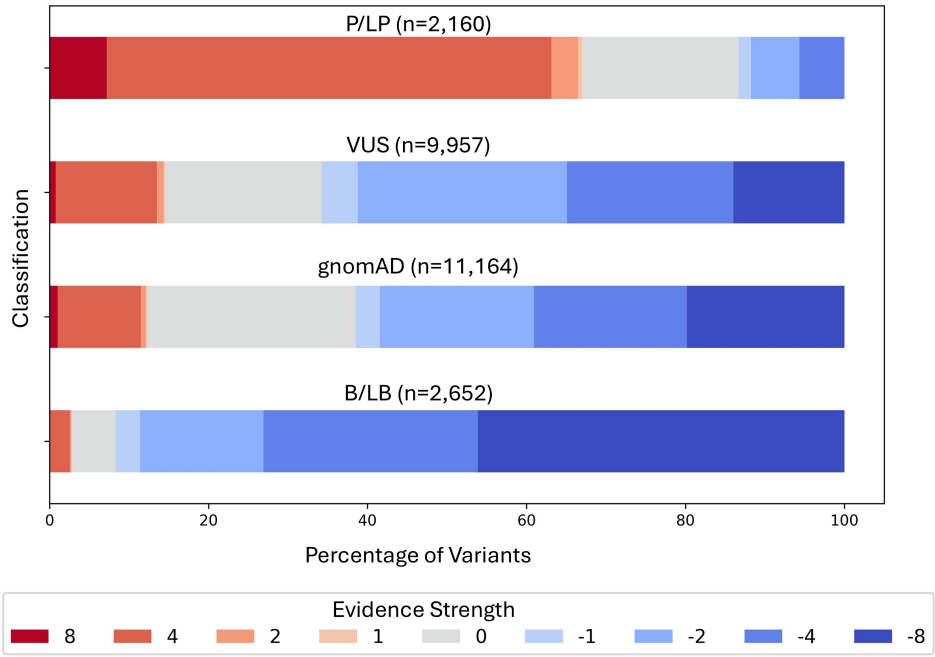

Figure 5: Distribution of evidence strengths assigned by our method for variants in the P/LP, VUS, gnomAD, and B/LB samples (each dataset was considered independent of each other). The results suggest the need for reclassification of ClinVar variants using updated functional evidence.

thermore, our variant-level annotations resulted in DOR = 152.6 compared to DOR = 49.4 for the author-provided annotations. Similarly, by dichotomizing our annotations into benign vs. non-benign, we obtained $LR^+ = 10.2$ and DOR = 163.9 for predicting benignity, compared to the author-provided annotations where we obtained $LR^+ = 5.7$ and DOR = 72.6 by dichotomizing the annotations as functionally normal vs. the rest. These results provide evidence that our variant-level calibration method is substantially more accurate than author-provided annotations and provide a basis for improved variant classification.

4 Related Work

To our knowledge, this work presents the first calibration approach for high-throughput functional assays for clinical utility. However, model calibration is a well-known issue in machine learning, where the objective is to adjust the scores of trained classifiers to properly reflect posterior probabilities [42, 34, 24]. Typical algorithms in this space rely on relatively large quantities of data and, more importantly, do not assume a bias model. This is critical, as the balance of pathogenic and benign variants in ClinVar is heavily skewed towards pathogenic. Therefore, we introduced a reference set (gnomAD) whose primary purpose was to help estimate the class prior, thus performing calibration in a semi-supervised manner.

Calibration has also been studied in the context of variant interpretation [3]. Pejaver et al. [39] proposed a nonparametric method for the calibration of computational tools, but this model relies on larger quantities of control variants (e.g., known P/LP or B/LB), which are unavailable for most genes. Therefore, stronger assumptions must be adopted to constrain the calibration. In terms of high-throughput functional data, assay score distributions have previously been modeled with

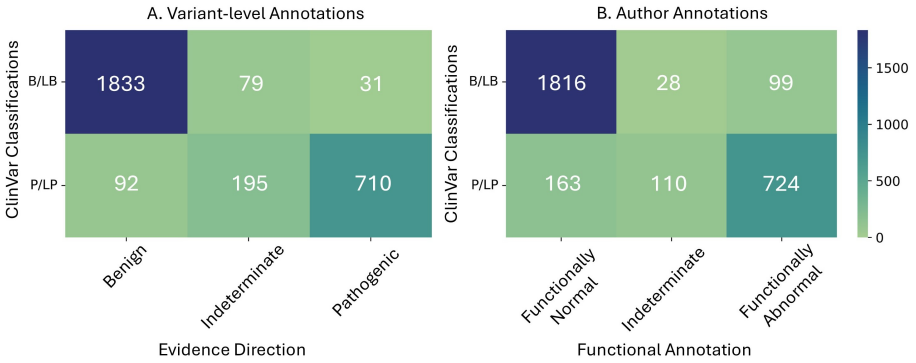

Figure 6: Comparison of our variant-level functional annotations, Figure 6A, to author-provided annotations, Figure 6B, for the subset of ClinVar variants in author-annotated scoresets. Variant-level evidence assignments less than 0 were labeled with benign evidence direction, assignments greater than 0 were labeled with pathogenic evidence direction, and 0 was considered indeterminate.

multi-sample Gaussian mixtures (MSGMM), learned in the positive-unlabeled classification setting to assign discrete functional annotations to variants [38]. Although not addressed in that work, the distributions learned by the MSGMM can be used to calculate a variant's posterior probability of being functionally abnormal. However, in the absence of overly restrictive assumptions on the functional mechanisms of disease, these probabilities do not contain sufficient information to calibrate assay scores to posterior probabilities of pathogenicity. The MSGMM method also does not impose density constraints on the component distributions and, therefore, cannot guarantee the monotonicity of the posterior. An additional method, MaveLLR, has been proposed to quantify the evidence strength of assay scores [55]. This method approximates the assay score distributions of pathogenic and benign variants using kernel density estimation (KDE). These density functions are used to compute the local positive likelihood ratio for each assay score. Along with the sensitivity of KDE to the chosen kernel in the presence of small data, this approach is not capable of estimating the prior probability of pathogenicity, a value critical to calibration. These limitations prevent rigorous calibration and reliable evidence strength assignments.

5 Discussion

High-throughput functional assays are a valuable tool in the clinical classification of variants [20]. To ensure consistent and accurate application of functional evidence, we present a statistical framework for calibrating high-throughput functional assays. This framework jointly models assay score distributions of pathogenic and benign variants, along with that of a reference set of variants from gnomAD, as unique mixtures of two skew normal distributions. This calibration model computes the local positive likelihood ratio along with the prior probability of pathogenicity for variants in a gene or domain.

The results demonstrate suitability of skew normal mixture models to faithfully capture assay score distributions from a diverse set of high-throughput functional experiments. The calibration model rigorously quantifies the strength of functional evidence in clinical variant classification and its distinct feature is that the evidence strength assigned to variants is variant-specific. Our model thus extends the current standard in the field by allowing multiple, distinct evidence strengths for the same assay [6]. The evidence assigned using this calibration model can be incorporated into the ACMG/AMP evidence-based framework. While the Richards et al. [45] guidelines limit

the strength of functional studies to strong and Brnich et al. [6] allow pathogenic evidence to be applied up to very strong, we evaluate the full potential of functional evidence, allowing for assays to additionally serve as very strong benign evidence if the necessary likelihood ratio is met. Then it can be considered if there is additional rationale for decreasing the strength of evidence applied in classification.

We have identified cases in which the calibration model assigns evidence strengths inconsistent with those applied by ClinGen Variant Curation Expert Panels (VCEPs). The PTEN VCEP [29] allows for the lipid phosphatase assay to serve as either moderate pathogenic evidence or supporting benign evidence, whereas our calibration model indicates the assay reaches sufficient positive likelihood ratios necessary to serve as up to very strong pathogenic evidence and moderate benign evidence. The skew normal mixture has a large degree of uncertainty on the benign variant distribution, a result of the bootstrapped samples of eight observations allowing for the overrepresentation of a single benign variant with low activity. The BRCA1 HDR assay is applied by the ENIGMA BRCA1 and BRCA2 Expert Panel [37] as a line of strong benign evidence, while the calibration model assigns a maximum strength of supporting benign evidence for this assay. We recommend further investigation into this case. The ClinGen Criteria Specification Registry for MSH2 (release 8/9/2024) specifies a calibrated functional assay can serve as supporting, moderate, or strong pathogenic evidence and can serve as supporting or strong benign evidence based on the calibrated assay's LR^+, with positive likelihood ratio thresholds corresponding to a prior probability of pathogenicity of 0.1. The calibration model approximates the MSH2 Resistance assay [22] to have median $\log LR^+$ values of -2.05 (95% CI [-2.93, -1.52]) and 5.06 (95% CI [3.37, 10.28]) at the lowest and highest assay scores observed. With an estimated median prior of 2%, the calibration model limits this assay to moderate benign evidence and strong pathogenic evidence.

This calibration approach is limited by its dependence on the availability of an unbiased sample of known pathogenic and benign variants to accurately compute local positive likelihood ratios. The modeling does not account for potential biases in the collection of classified variants; e.g., the possibility of the pathogenic variants being more deleterious and having larger functional effect. Further work is needed to detect and quantify the severity of bias in assay distributions and correct for the effect of this bias on calibration [9, 61, 60]. While we limit circularity in calibrating the BRCA1 dataset [13], functional data (either from the other datasets used here or other independent functional data) may already have been incorporated in classifying the existing set of pathogenic and benign variants. Such detailed application of evidence is inconsistently recorded in ClinVar, making it difficult to fully account for circularity. Thus, future calibration approaches must be aware of potential risks of circularity. Additionally, while the skew normal distributions capture the distributions of the high-throughput functional data analyzed here, the quality of the model fit may vary on other datasets and is affected by the normalization steps taken in generating functional scores. The calibration approach has not been evaluated on genes where molecular mechanisms of disease include both gain and loss of function. A three-component mixture model could be an attractive alternative for cases with multiple molecular mechanisms of disease and such a model could be investigated as functional studies with this pattern accumulate.

Overall, our work demonstrates the feasibility of calibrating high-throughput functional assays for clinical variant classification. We find that the existing assays are sufficiently reliable to contribute to variant classification and thus reduce the number of rapidly accumulating VUS. Finally, calibration of functional assays where scores are treated on a continuous scale, will lead to an increased accuracy of genetic diagnosis and improved medical management for individuals affected by Mendelian disorders.

Acknowledgements

We thank the ClinGen SVI Working Group for their guidance. This work has been supported by the NIH awards U01 HG012022 (PR), R00 LM012992 (VP), U24 HG011450 (AODL), U01 HG011755 (AODL), R35 GM152106 (DMF), and UM1 HG011969 (DMF, LMS); an Early Career Award from Alex's Lemonade Stand for Childhood Cancer and the RUNX1 Foundation 21-25037 (AEM), and the Brotman Baty Institute Catalytic Collaborations Grant CC28 (AEM).

References

[1] A. I. Adamovich et al. The functional impact of BRCA1 BRCT domain variants using multiplexed DNA double-strand break repair assays. *Am J Hum Genet*, 109(4):618–630, 2022.

[2] A. Azzalini. A class of distributions which includes the normal ones. *Scand J Stat*, 12(2):171–178, 1985.

[3] S. Benevenuta et al. Calibrating variant-scoring methods for clinical decision making. *Bioinformatics*, 36(24):5709–5711, 2021.

[4] T. Bergquist et al. Calibration of additional computational tools expands ClinGen recommendation options for variant classification with PP3/BP4 criteria. *Genet Med*, 101402, 2025.

[5] S. Boettcher et al. A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science*, 365(6453):599–604, 2019.

[6] S. E. Brnich et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med*, 12(1):3, 2019.

[7] M. Buckley et al. Saturation genome editing maps the functional spectrum of pathogenic VHL alleles. *Nat Genet*, 56(7):1446–1455, 2024.

[8] L. Clausen et al. A mutational atlas for Parkin proteostasis. *Nat Commun*, 15(1):1541, 2024.

[9] M. C. De Paolis Kaluza et al. An approach to identifying and quantifying bias in biomedical data. *Pac Symp Biocomput*, 28:311–322, 2023.

[10] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*, 1(1):54–75, 1986.

[11] S. Erwood et al. Saturation variant interpretation using CRISPR prime editing. *Nat Biotechnol*, 40(6):885–895, 2022.

[12] D. Esposito et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol*, 20(1):223, 2019.

[13] G. M. Findlay et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562(7726):217–222, 2018.

[14] C. Fortuno et al. Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum Mutat*, 42(3):223–236, 2021.

[15] D. M. Fowler et al. An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol*, 24(1):147, 2023.

[16] A. O. Giacomelli et al. Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat Genet*, 50(10):1381–1387, 2018.

[17] A. S. Glas et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*, 56(11):1129–1135, 2003.

[18] S. M. Harrison and H. L. Rehm. Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. *Genome Med*, 11(1):72, 2019.

[19] C. Hu et al. Classification of BRCA2 variants of uncertain significance (VUS) using an ACMG/AMP model incorporating a Homology-Directed repair (HDR) functional assay. *Clin Cancer Res*, 28(17):3742–3751, 2022.

[20] IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature*, 633(8028):47–57, 2024.

[21] K. Jaganathan et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.e24, 2019.

[22] X. Jia et al. Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am J Hum Genet*, 108(1):163–175, 2020.

[23] S. Kato et al. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A*, 100(14):8424–8429, 2003.

[24] S. C. Kim et al. The Fermi-Dirac distribution provides a calibrated probabilistic output for binary classifiers. *Proc Natl Acad Sci U S A*, 118(34):e2100761118, 2021.

[25] Z. Koenig et al. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res*, 34(5):796–809, 2024.

[26] M. J. Landrum et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, 46(D1):D1062–D1067, 2018.

[27] T. I. Lin et al. Finite mixture modelling using the skew normal distribution. *Stat Sin*, 17(3):909–927, 2007.

[28] R. S. Lo et al. The functional impact of 1,570 individual amino acid substitutions in human OTC. *Am J Hum Genet*, 110(5):863–879, 2023.

[29] H. Margot et al. Classification of PTEN germline non-truncating variants: a new approach to interpretation. *J Med Genet*, 61(12):1071–1079, 2024.

[30] K. A. Matreyek et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet*, 50(6):874–882, 2018.

[31] A. McKenna et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, 2010.

[32] T. L. Mighell et al. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am J Hum Genet*, 102(5):943–955, 2018.

[33] M. Mort et al. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum Mutat*, 31(3):335–346, 2010.

[34] A. Niculescu-Mizil and T. Fawcett. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

[35] A. R. Ollodart et al. Multiplexing mutation rate assessment: determining pathogenicity of MSH2 variants in Saccharomyces cerevisiae. *Genetics*, 218(2), 2021.

[36] R. Olvera-León et al. High-resolution functional mapping of RAD51C by saturation genome editing. *Cell*, 187(20):5719–5734.e19, 2024.

[37] M. T. Parsons et al. Evidence-based recommendations for gene-specific ACMG/AMP variant classification from the ClinGen ENIGMA BRCA1 and BRCA2 variant curation expert panel. *Am J Hum Genet*, 111(9):2044–2058, 2024.

[38] V. Pejaver et al. Assessment of methods for predicting the effects of PTEN and TPMT protein variants. *Hum Mutat*, 40(9):1495–1506, 2019.

[39] V. Pejaver et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*, 109(12):2163–2177, 2022.

[40] Y. Peng et al. New mixture models for decoy-free false discovery rate estimation in mass spectrometry proteomics. *Bioinformatics*, 36(Suppl_2):i745–i753, 2020.

[41] Y. Peng et al. An algorithm for decoy-free false discovery rate estimation in XL-MS/MS proteomics. *Bioinformatics*, 40(Suppl 1):i428–i436, 2024.

[42] J. C. Platt. *Probabilistic outputs for support vector machines and comparison to regularized likelihood methods*, pages 61–74. MIT Press, 1999.

[43] R. Rastogi et al. Critical assessment of missense variant effect predictors on disease-relevant variant data. *Hum Genet*, 144(2-3):281–293, 2025.

[44] H. L. Rehm et al. ClinGen–the clinical genome resource. *N Engl J Med*, 372(23):2235–2242, 2015.

[45] S. Richards et al. 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. *Genet Med*, 17(5):405–424, 2015.

[46] J. Shendure et al. Genomic medicine – progress, pitfalls, and promise. *Cell*, 177(1):45–57, 2019.

[47] L. M. Starita et al. A multiplex homology-directed DNA repair assay reveals the impact of more than 1,000 BRCA1 missense dubstitution variants on protein function. *Am J Hum Genet*, 103(4):498–508, 2018.

[48] S. L. Stenton et al. Assessment of the evidence yield for the calibrated PP3/BP4 computational recommendations. *Genet Med*, 26(11):101213, 2024.

[49] S. L. Stenton et al. Critical assessment of variant prioritization methods for rare disease diagnosis within the rare genomes project. *Hum Genomics*, 18(1):44, 2024.

[50] D. Tabet et al. Scalable functional assays for the interpretation of human genetic variation. *Annu Rev Genet*, 56:441–465, 2022.

[51] S. V. Tavtigian et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med*, 20(9):1054–1060, 2018.

[52] S. V. Tavtigian et al. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum Mutat*, 41(10):1734–1737, 2020.

[53] M. Tejura et al. Calibration of variant effect predictors on genome-wide data masks heterogeneous performance across genes. *Am J Hum Genet*, 111(9):2031–2043, 2024.

[54] The Critical Assessment of Genome Interpretation Consortium. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. *Genome Biol*, 25(1):53, 2024.

[55] W. van Loggerenberg et al. Systematically testing human HMBS missense variants to reveal mechanism and pathogenic variation. *Am J Hum Genet*, 110(10):1769–1786, 2023.

[56] L. C. Walker et al. Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI splicing subgroup. *Am J Hum Genet*, 110(7):1046–1067, 2023.

[57] A. J. Waters et al. Saturation genome editing of BAP1 functionally classifies somatic and germline variants. *Nat Genet*, 56(7):1434–1445, 2024.

[58] J. Weile and F. P. Roth. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum Genet*, 137(9):665–678, 2018.

[59] R. Yang et al. A new class of metrics for learning on real-valued and structured data. *Data Min Knowl Disc*, 33(4):995–1016, 2019.

[60] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning*, pages 114–121, 2004.

[61] D. Zeiberg et al. Leveraging structure for improved classification of grouped biased data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11113–11120, 2023.