A unified meta-regression model identifies genes associated with epilepsy

Oscar Aguilar[1]*, *Mijail Rivas*[2], *Manuel A. Rivas*[3]*

Abstract

Epilepsy is a highly heterogeneous disorder thought to have strong genetic components. However, identifying these risk factors using whole-exome sequencing studies requires very large sample sizes and good signal-to-noise ratio in order to assess the association between rare variants in any given gene and disease. We present an approach for predicting constraint in the human genome through application of a Hidden Markov Model (HMM) to whole exome sequencing (WES) data. Using the Regeneron Genetics Center Million Exome dataset and the AllofUs whole genome sequencing data, we predict the probability of observing no variants across the population for each position in the genome. We then incorporate the predictions with the "rejected substitutions" (RS) score from Genomic Evolutionary Rate Profiling (GERP), pathogenicity predictions from AlphaMissense (AM), and pLoF/Missense annotations from Epi25 into a model that detects epilepsy-associated genes. We identify a set of significant ($p < 3.4 \times 10^{-7}$) genes which did not meet exome-wide significance in previous studies: *KCNQ2*, *SCN2A*, *STXBP1*, *CACNA1A*, *SLC6A1*, *DYRK1A*, *KCNB1*, *SATB1*, *PCDHAC2*, *SP4*, and *RYR2*,. Our models allow us to evaluate the contribution of constraint, protein structure based pathogenicity prediction from AM, and pLoFs jointly. We show that unifying these moderators into a single model allows us to both strengthen our evidence for genes with already-known links to epilepsy and also identify new genes with likely links to epilepsy.

[1] Department of Management Science & Engineering, Stanford University
[2] National Institute of Neurology and Neurosurgery, Mexico and Clinical Epileptology Fellowship UNAM
[3] Department of Biomedical Data Science, Stanford University
* Corresponding authors. E-mails: osthoag@stanford.edu; mrivas@stanford.edu

Introduction

Epilepsy is one of the most common neurological disorders, affecting individuals across all ages and populations.[1] It encompasses a diverse spectrum of phenotypes and is marked by a predisposition to recurrent epileptic seizures, which can lead to substantial cognitive, psychological, and social consequences.[2] Beyond its clinical impact, epilepsy represents a major global health burden, affecting approximately 70 million people worldwide, and the cost of diagnosis and treatment exceeds USD 100 billion annually.[2] While epilepsy can result from various causes such as infection, trauma, and stroke, advances in genomics have demonstrated that genetic factors are likely implicated in over two-thirds of cases where the cause is unknown.[3]

Advances in sequencing and gene discovery have enabled the study of the genetic etiologies of epilepsy.[4–6] Genetic inheritance can be polygenic, as seen in genetic (idiopathic) generalized epilepsy (GGE [MIM: 600669]), where pathogenic variants are usually not detected in standard gene panels. Alternatively, the inheritance can be monogenic, as observed in early-onset developmental epileptic encephalopathies (DEE [MIM: 308350]), with pathogenic variants identified through epilepsy gene panels or whole-exome sequencing. These insights have motivated the integration of genetic testing into clinical care, with benefits including refined diagnosis, targeted therapy, and informed genetic counseling.[7] Despite these advances, the delineation of genes that confer risk for epilepsy remains incomplete. There is no consensus definition of what constitutes an “epilepsy gene,” and distinctions are often made between genes directly implicated in epilepsy pathogenesis and genes associated with broader syndromic presentations that include seizures .[8–11]

Whole-exome sequencing (WES) has emerged as a powerful tool for investigating the genetic basis of epilepsy, enabling the identification of rare variants associated with various diseases. A landmark study by the Epi25 Collaborative analyzed over 54,000 individuals (20,979 epilepsy cases and 33,444 controls), yielding novel insights into the genetic basis of epilepsy across multiple subtypes and ancestries.[6] However, the study also highlighted key challenges: the rarity and heterogeneity of causal variants, limited statistical power for gene-level association testing, and the need for improved variant interpretation frameworks.

A key strategy for improving rare variant association studies involves incorporating prior knowledge about *constraint,* defined *as* the evolutionary intolerance of a region to variation. Existing approaches such as the probability of loss-of-function intolerance (pLI) and the Loss-of-function Observed/Expected Upper bound Fraction (LOEUF) from the gnomAD consortium quantify gene-level depletion of protein-truncating variation relative to expectation.[12] Other regional metrics such as the Missense Tolerance Ratio (MTR) evaluate local intolerance to missense variation.[13] Likewise, evolutionary conservation-based measures such as Genomic Evolutionary Rate Profiling (GERP) estimate constraint from multispecies alignments, providing position-specific “rejected substitution” (RS) scores that quantify the deficit of substitutions relative to neutrality.[14–16] GERP captures constraint at single-base resolution and complements population-based depletion metrics by emphasizing evolutionary conservation across mammals.

Collectively, these constraint scores have become essential tools for gene discovery and variant prioritization, yet they are typically static, population-derived summaries that do not fully capture the spatial structure or context of mutational depletion.

Recent advances in deep learning have further improved variant-level prioritization. AlphaMissense, a state-of-the-art model developed by Google DeepMind, predicts the pathogenicity of all possible missense variants using protein structure-informed features derived from AlphaFold2.[17] This resource provides a genome-wide atlas of predicted missense effects, and complements traditional constraint metrics by leveraging structural and functional priors.

To address the challenge of limited statistical power in rare variant association studies, we developed an approach that leverages sequence-level constraint inference using a Hidden Markov Model (HMM). We applied this framework to large-scale exome data from the Regeneron Genetics Center (RGC-ME)[13] and validated its robustness using the All of Us cohort.[18] By integrating this sequence-derived constraint signal with evolutionary constraint scores from GERP, structure-informed pathogenicity predictions from AlphaMissense, and annotations of predicted loss-of-function (pLoF) and missense variants, we construct a unified meta-regression statistical model to assess gene-level associations with epilepsy risk. This integrative method enables the detection of biologically meaningful signals that are missed by standard approaches, particularly in the context of heterogeneous neurodevelopmental disorders like epilepsy.

Methods

Inferring Constraint in the Human Genome Using a Hidden Markov Model

To build on prior gene- and region-level constraint metrics and improve resolution at the nucleotide level, we develop a Hidden Markov Model (HMM) to infer constraint directly from large-scale human variation data. An HMM is a statistical framework that models a sequence as a series of underlying, unobserved states—in this case, mutationally constrained or tolerant positions—that generates the observed genetic variation. By learning how likely the sequence transitions between these hidden states, the model can infer which regions of the genome are under selective constraint even when direct evidence is sparse. The HMM is particularly well-suited to this task due to its ability to model spatial dependencies in sequence data, capturing the transition between mutationally tolerant and constrained regions more effectively than simpler frequency-based approaches.

We define the observed sequence for each chromosome as a binary vector, where each position is encoded as “1” if at least one variant is detected in the cohort and “0” otherwise. The latent states represent the underlying mutational constraint: one state corresponds to constrained regions, where mutations are rare due to selective pressure, while the other represents non-constrained regions. The HMM is trained using the expectation-maximizing Baum-Welch algorithm, which iteratively updates the transition and emission probabilities to maximize the likelihood of the observed data.

Our implementation uses the `hmmlearn` library in Python to simulate a genomic sequence of specified length with a biased probability of mutation to reflect realistic variation rates. Before training, the observed binary sequence is preprocessed to account for local dependencies between adjacent sites by summarizing short subsequences of length “order.” These subsequences are converted into count-based representations, in which occurrences of 0s and 1s are aggregated to support more stable statistical estimation. This preprocessing step enables the model to better capture the short-range structure of mutational patterns, improving its ability to distinguish between constrained and tolerant regions.

Using these generated sequences, we initialize and fit a first-order HMM with two hidden states, representing constrained and non-constrained genomic regions. The model is trained using the Baum–Welch algorithm, which iteratively estimates how likely each genomic position is to belong to either the constrained or unconstrained state and updates the underlying probability of transition between these two states. Conceptually, the algorithm alternates between two steps: estimating which regions are most likely constrained given the current model (the *expectation* or E-step) and then refining the model parameters to better explain the observed mutation patterns

(the *maximization* or M-step).

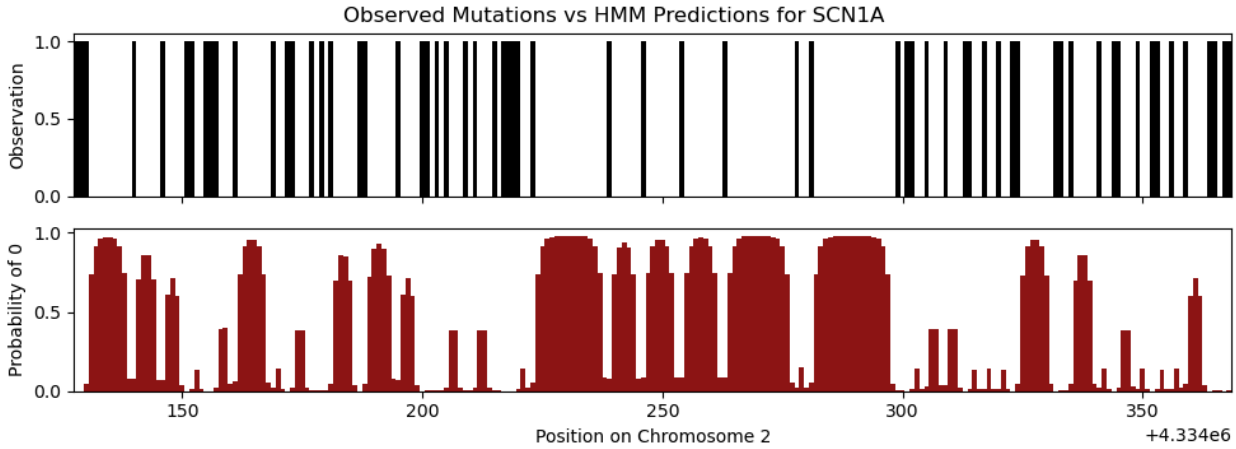


Figure 1: Example of an observed binary sequence vs the predictions made by the HMM

Formally, the expectation (E) step computes the *forward* and *backward* probabilities for each position t along the observed sequence. The *forward probability*, denoted $\alpha_t(i)$, represents the joint probability of observing the sequence up to position t and being in hidden state i at that position. Conversely, the *backward probability*, $\beta_t(i)$, represents the probability of observing the remaining sequence from position $t + 1$ to the end, given that the model is currently in state i at position t. Using these quantities, we compute (1) the posterior probability of being in state i at position t, $\gamma_t(i)$, and (2) the joint posterior probability of transitioning from state i to state j between positions t and $t + 1$, $\xi_t(i, j)$. In the maximization (M) step, the parameters of the model are updated to maximize the expected complete-data log-likelihood. This iterative process continues until convergence, yielding the most probable configuration of constrained and tolerant segments across the simulated genome.

To ensure robustness, we exclude positions with insufficient sequencing depth, retaining only those with >20% coverage across at least 90% of individuals. This criterion mitigates biases introduced by missing data and ensures that the inferred constraints reflect true biological patterns rather than sequencing artifacts. Additionally, we restrict our analysis to protein-coding regions, as these sites are most relevant for disease association studies. Once trained on chromosome 2 from the RGC-ME dataset, we use the model to infer constraint probabilities across all chromosomes.

Validation of Constraint Predictions Using an Independent Dataset

To validate the robustness and generalizability of our constraint inference model, we apply it to an independent whole-exome sequencing dataset from the AllofUs (AoU) initiative, comprising 245,388 individuals of mixed ancestry. This step is crucial in demonstrating that the HMM captures fundamental biological properties rather than dataset-specific artifacts.

We use the trained HMM from chromosome 2 of RGC-ME to predict constraint probabilities across all other chromosomes in AoU to maintain independence from the training set. To assess the validity of our predictions, we examine the joint distribution of constraint probabilities across both datasets in Figure 2a. If our model captures meaningful constraint information, we expect consistent predictions between RGC-ME and AoU, especially for highly constrained and highly mutable regions.

Our analysis confirms a strong overlap in constraint predictions between the two datasets, demonstrating the model's reliability and consistency across cohorts. The highest density of observed probability mass is concentrated in regions with low predicted constraint (< 0.1), consistent with the expectation that most genomic positions are not under strong selective pressure. Importantly, high-probability constraint predictions (top decile) show significant concordance between RGC-ME and AoU, indicating that the HMM identifies constrained regions independently of the dataset on which it was trained. To quantify this relationship, we compare in Figure 2b the observed joint distribution to the expected distribution under an independence assumption. The resulting chi-square statistics reveal significant deviation from independence in both low- and high-constraint probability regions, with values ranging from 10,000 to 90,000 across chromosomes. In contrast, regions where the model disagrees (i.e. infers high constraint in one dataset and low constraint from the other) exhibit lower chi-square values (3,000–37,000), suggesting that discordant predictions are limited and largely confined to intermediate-probability regions. Together, these results support that our model captures stable and biologically meaningful patterns of constraint that generalize across datasets, populations, and genomic contexts.

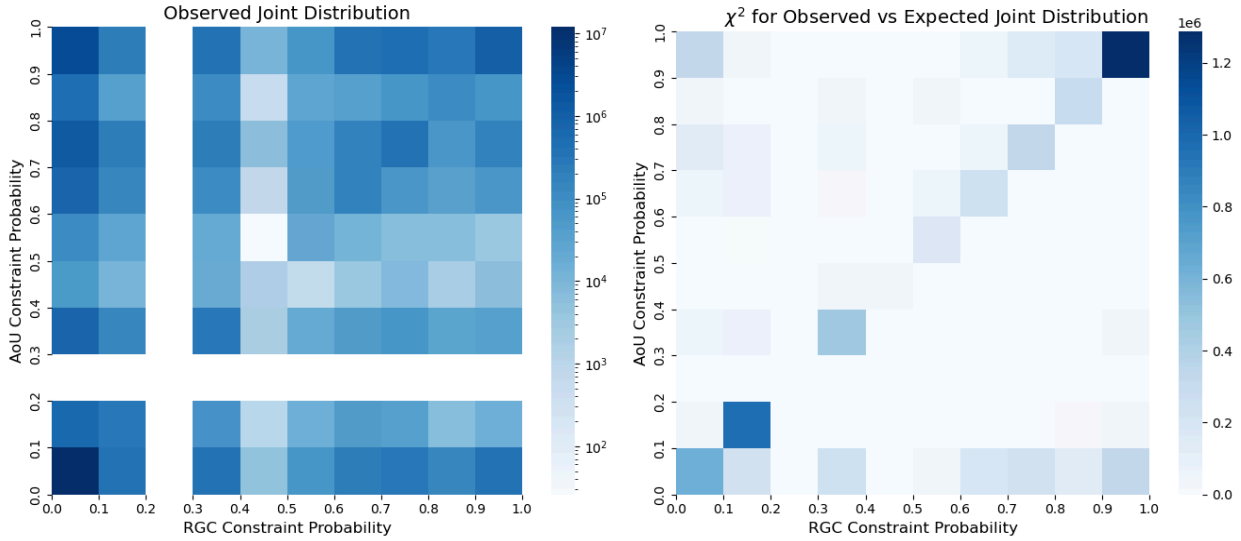


Figure 2: Observed joint distribution (left) of WES constraint predictions (excluding chromosome 2) from AoU and RGC, and distribution of corresponding χ^2 statistics (right). The [0.2, 0.3) bin has no observations because the HMM does not generate constraint values within that range.

Comparison with Existing Constraint Metrics

To evaluate the HMM predictions as a measure of mutational constraint, we compare them to three widely used gene constraint metrics: the Missense Tolerance Ratio (MTR), the missense z-score, and the Genomic Evolutionary Rate Profiling (GERP). The MTR, derived from the gnomAD dataset, is defined as the observed proportion of missense variants divided by the expected proportion, relative to the number of all possible variants in a given genomic window.[13] The missense z-score, on the other hand, quantifies deviation from expectation by computing the z-score of the observed missense variant count relative to the expected count, as derived from the gnomAD dataset.[12] A higher missense z-score indicates stronger selective constraint, meaning fewer missense variants were observed than expected, whereas a lower MTR value suggests higher intolerance to variation. GERP, in contrast, measures evolutionary constraint across 35 mammalian species by estimating the deficit of substitutions relative to neutrality. Positive RS scores indicate purifying selection and evolutionary conservation, providing a cross-species complement to population-based metrics such as MTR and the missense z-score.

To ensure a fair comparison between our HMM predictions and these established constraint metrics, we aggregate our model's constraint probabilities at the gene level by computing the proportion of each gene where the HMM assigns a probability greater than 0.5 to the constrained state. This aggregated "constraint proportion" is then analyzed in relation to the MTR, missense z-score, and GERP score. Figures 3a and 3b illustrate the joint distributions of predicted constraint proportions versus the first two measures, showing broadly consistent trends: genes predicted to be highly constrained by the HMM tend to have lower MTR values and higher missense z-scores. The relationships are quantified by fitting a generalized linear model (GLM) using ordinary least squares regression, yielding R-squared values of 0.204 and 0.146, respectively. These results confirm that the HMM captures gene-level constraint patterns comparable to those observed in established population-based metrics, while offering improved spatial resolution and interpretability.

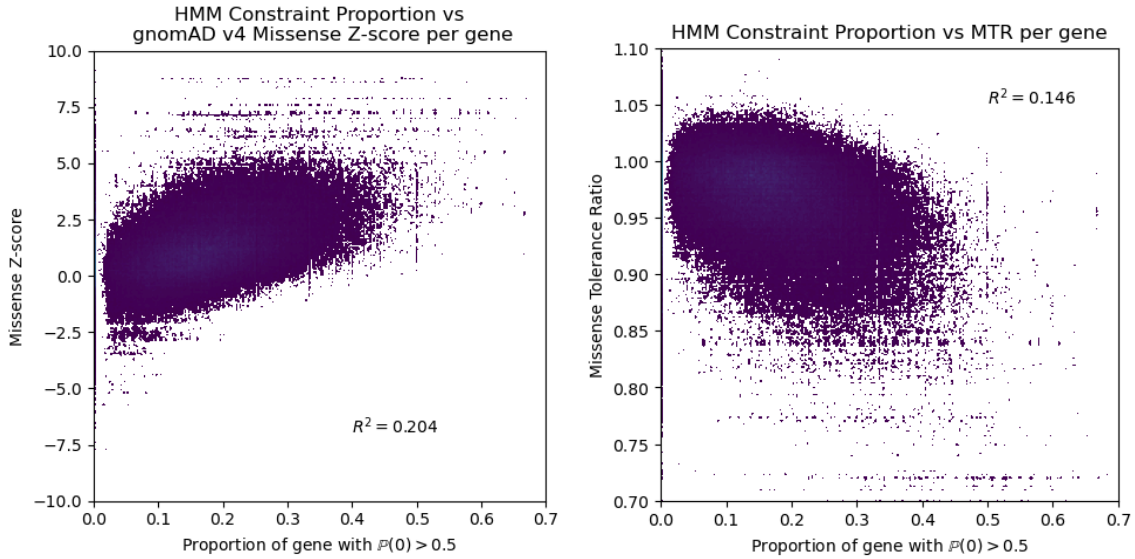


Figure 3: Joint distribution of z-score of observed vs expected missense variants per gene (left) and missense tolerance ratio (right) versus proportion of gene with HMM constraint probability > 0.5 (x-axis).

We next compare our HMM-derived nucleotide-level constraint predictions with GERP RS scores, which quantify evolutionary conservation at single-base resolution. Unlike the MTR and missense z-score, which are computed per gene, both the HMM and GERP operate at the locus level, allowing direct exome-wide comparison of their joint distributions. Across all positions, the correlation between HMM constraint probabilities and GERP RS scores is weak $\left(R^2 = 0.006\right)$, indicating that the two approaches capture distinct aspects of constraint. However, analysis of their observed versus expected joint distribution in Figure 4 reveals a pronounced deviation from independence. Highly conserved sites (GERP > 4–5) that are not classified as constrained by the HMM are over-represented, whereas HMM-constrained but evolutionarily unconserved sites are under-represented. This bifurcation suggests that while GERP primarily reflects deep evolutionary conservation across mammals, the HMM captures more recent or population-specific selective pressures. Together, these complementary signals highlight that the HMM generalizes constraint detection beyond static evolutionary conservation, capturing mutational intolerance shaped by both functional and demographic processes.

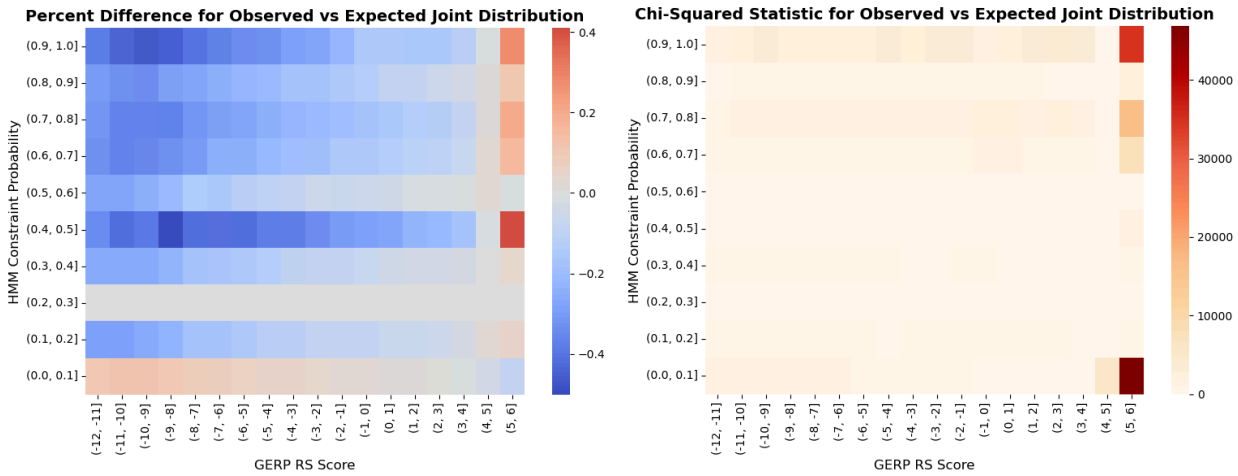


Figure 4: Distribution of percent difference (left) and χ^2 statistics (right) for the observed vs expected joint distribution of HMM constraint predictions and GERP RS scores.

By comparing our model's predictions to established population-based and evolutionary measures of selective constraint, we demonstrate that the HMM captures known signatures of mutational intolerance while also providing complementary information beyond existing metrics. The observed correlations with MTR and missense z-scores confirm that the HMM aligns with established gene-level patterns of constraint, whereas its partial independence from GERP indicates that it captures additional signals shaped by population-level variation and functional context rather than deep evolutionary conservation. Together, these findings support the HMM as a robust and biologically informative predictor of constraint, offering improved resolution and interpretability for identifying functionally important regions of the genome.

Lastly, we assess the marginal informativeness of constraint versus pathogenicity scores using logistic regression. Modeling observed variant presence as a function of AM pathogenicity scores alone yields a McFadden pseudo-R^2 of 0.0081. In contrast, incorporating HMM-based

constraint probabilities as predictors results in a much higher pseudo-R^2 of 0.3834, underscoring the biological distinctiveness and predictive value of the HMM-inferred constraint signal over purely sequence-derived pathogenicity scores.

Unified meta-regression model

To improve detection of genes associated with epilepsy, we integrate our HMM-derived constraint predictions with the “rejected substitutions” (RS) score from Genomic Evolutionary Rate Profiling (GERP), missense variant pathogenicity probabilities from AlphaMissense (AM) and loss-of-function (pLoF) and missense annotations within a unified statistical framework. We present a weighted least squares (WLS) meta-regression model to effectively weight each individual moderator (HMM, GERP, AM, pLoF, and missense) to minimize prediction variance and maximize accuracy, which was developed concurrently with Lauer and Rivas.[19]

Suppose that for gene k, we have variants $i = 1, 2, \dots, n_k$. Then, for each moderator, we obtain the following outputs:

- $X_{1,i}^{k} = -log(1 - probability\ of\ constraint_i)$, from the Hidden Markov Model (HMM);
- $X_{2,i}^{k} = RS$, "rejected substitutions" (RS) score from GERP;
- $X_{3,i}^{k} = -log(1 - probability\ of\ pathogenicity_i)$, from AlphaMissense;
- $X_{4,i}^{k} = 1_{LoF,i}$ an indicator of whether variant i is a pLoF variant.
- $X_{5,i}^{k} = 1_{Missense,i}$ an indicator of whether variant i is a Missense variant.

The column vector $X_j^{k} = \left(X_{j1}^{k}, X_{j2}^{k}, \dots, X_{j,n_k}^{k}\right)^{\mathsf{T}}$ denotes the output from moderator j for all variants of gene k. Similarly, we define the $n_k \times 5$ matrix $X^k = \left(X_1^{k}, X_2^{k}, X_3^{k}, X_4^{k}, X_5^{k}\right)$.

For variant i of gene k, we have four distinct populations: (1) the subset of the population with variant i that is (1) epileptic (ALT_{EP}), and (2) *not* epileptic (ALT_{CON}), and the subset of the population without variant i that is (3) epileptic (REF_{EP}), and (4) *not* epileptic (REF_{CON}).

Let $\hat{\beta}_i^{k}$ denote the estimated single-variant effect size estimate representing the log odds ratio of finding variant i of gene k in the population of patients with epilepsy versus the control. Using the following formula from the National Institute of Standards and Technology (NIST), [20] we obtain

$$\hat{\beta}_i^{k} = \log\left(\frac{\left(0.5 + ALT_{EP}\right)\cdot\left(0.5 + ALT_{CON}\right)}{\left(0.5 + REF_{EP}\right)\cdot\left(0.5 + REF_{CON}\right)}\right).$$

We can then estimate the standard error SE_i^k of estimated single-variant effect size $\hat{\beta}_i^k$, computed by

$$\left(\hat{SE}_i^k\right)^2 = \frac{1}{0.5 + REF_{EP}} + \frac{1}{0.5 + REF_{CON}} + \frac{1}{0.5 + ALT_{EP}} + \frac{1}{0.5 + ALT_{CON}}.$$

Each variant contributes a datapoint to the model; thus, we can weigh each of the n_k variants by positive weight w_i^k to achieve an aggregate effect size $\hat{\beta}^k = \sum_{i=1}^{n_k} w_i^k \hat{\beta}_i^k$. To minimize the variance of this aggregated estimator, we take $w_i^k \propto 1/SE_i^2$ for all variants i of gene k. Let $w^k = \left(w_1^k, w_2^k, \ldots, w_{n_k}^k\right)$ and $W = diag\left(w^k\right)$. Then, for any gene k and any variant $i \in \{1, 2, \ldots, n_k\}$, the WLS meta-regression model is defined as

$$y^k \mid W^k \sim X^k b^k + \epsilon_i$$

where b^k is a vector of coefficients given by $b^k = \left(\left(X^k\right)^\top W^k X^k\right)^{-1} \left(X^k\right)^\top W^k y^k$, and y^k is the vector of predicted effect sizes estimating $\hat{\beta}^k$.

To evaluate the relative explanatory power of each moderator, we compute the proportion of variance explained by each predictor using partial R_j^2 values:

$$R_j^2 = \frac{SSR_{reduced,j} - SSR_{full}}{SSR_{reduced,j}},$$

where SSR_{full} is the residual sum of squares from the full model, and $SSR_{reduced,j}$ is the residual sum of squares from the model excluding moderator j. The total model R^2 is:

$$R_{total}^2 = 1 - \frac{SSR_{full}}{SST},$$

with $SST = \sum_{i=1}^{m} w_i \left(y_i - \bar{y}_w\right)^2$ and $\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$.

We further test the significance of each predictor using the t-statistic:

$$t_j = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}$$

and assess overall model fit using the F-statistic:

$$F = \frac{\left(SSR_{null} - SSR_{full}\right)/(p-1)}{SSR_{full}/(m-p)},$$

with corresponding unified model p-value $p_{unified} = P\left(F > F_{observed}\right)$.

Prior to fitting the meta-regression, we applied variant- and gene-level inclusion filters to ensure model stability and interpretability. Specifically, we retained only exonic variant positions with at least one observed alternate allele in the study population (cases or controls) and no more than five total alternate-allele counts across the combined cohort, thereby restricting analysis to rare, observed variants. Furthermore, we required that each gene–group meta-regression include at least N > 25 variants, following recommendations by Jenkins and Quintana-Ascencio.[21] This threshold minimizes overfitting and ensures that estimated regression coefficients and partial R^2 values are derived from adequately sized variant sets.

Variants are mapped to genes using a standardized gene identifier (Ensembl gene ID) joined to an HGNC gene symbol via our gene reference table. We restrict to autosomal exonic variants (X, Y, MT excluded) with non-missing consequence annotations. For Epi25, we retain variants whose consequence is either pLoF or missense after excluding synonymous, non-coding, and undefined annotations. We encode two binary indicators using the annotations:

- $consequence \in \{pLoF\} \rightarrow 1_{LoF,i}$
- $consequence \in \{damaging\ missense,\ other\ missense\} \rightarrow 1_{Missense,i}$

For SCHEMA, we exclude non-coding/synonymous classes: 5′/3′UTR, intronic, intergenic, upstream/downstream, non-coding transcript (exon), splice-region, mature miRNA, coding sequence, synonymous, and undefined annotations. We encode the binary indicators using:

- $\{stop\ gained,\ splice\ acceptor,\ splice\ donor,\ frameshift\} \rightarrow 1_{LoF,i}$
- $\{inframe\ insertion/deletion,\ stop\ retained/lost,\ protein\ altering,\ missense\ MPC\} \rightarrow 1_{Missense,i}$

These dataset-specific consequence classes are used to construct the moderator indicators $X_{4,i}^{k}$ and $X_{5,i}^{k}$ for the meta-regression model.

Results

Having validated the HMM as a sequence-dependent predictor of mutational constraint, we next integrate its output with structure-based missense variant pathogenicity predictions from AlphaMissense, existing annotations of predicted loss-of-function (pLoF) variants, as well as GERP RS scores. These signals are combined within the weighted meta-regression framework described in the previous section to jointly estimate variant-level effects.This unified model provides a general framework for variant-level prioritization and gene-level inference of rare variant burden across diverse disorders. We first evaluate its performance in epilepsy, where extensive sequencing resources enable systematic benchmarking, and then demonstrate its broader applicability using a schizophrenia dataset with more limited power. Together, these analyses illustrate the model's phenotype-agnostic design, in which constraint, pathogenicity, and pLoF information are flexibly integrated to detect disease-relevant genetic signals.

Epilepsy

We apply our analysis to the Epi25 Collaborative dataset comprising 54,000 individuals with and without epilepsy.[6] Following this study, we define exome-wide significance as $p = 3.4 \times 10^{-7}$, corresponding to a Bonferroni correction across 18,531 protein-coding genes tested for four epilepsy subtypes and two variant classes. Genes with $p < 1.0 \times 10^{-4}$ are considered suggestive. Table 1 compares the p-values for each moderator as well as our unified model for a subset of either (a) genes above the exome-wide significance threshold on our unified model and below the threshold in the Epi25 study or (b) suggestive genes with either known links to epilepsy or other neurological functions whose association with epilepsy were previously inconclusive. The genes meeting criterion (b) are designated by the pink-highlighted rows in Table 1. Table S1 in Supplementary Materials contains a full set of suggestive genes, and all results can be explored in our epilepsy dashboard (https://epilepsy-dashboard.streamlit.app/) Across the genes in Table 1, the meta-regression identifies statistically significant signals where prior population-based analyses lacked sufficient power for detection. In particular, the weighted meta-regression model is able to leverage the strengths of each individual moderator in the regression model to exact a stronger signal on several genes' influence on the presentation of epilepsy.

We will first discuss the genes meeting the latter "importance" criterion, designated by the pink-highlighted rows in Table 1: *KCNQ2* (DEE) [MIM: 602235], *RYR2* (GGE) [MIM: 180902], *CACNA1A* (GGE) [MIM: 601011], *ATP2B2* (DEE) [MIM: 108733], *CACNA1B* (GGE) [MIM: 601012], and *KDM5B* (DEE) [MIM: 605393]. These genes can be partitioned into those that have established roles in epilepsy presentation and those genes whose role in epilepsy was inconclusive prior to this study. Among the established genes, *KCNQ2*, *CACNA1A*, and *CACNA1B* each encode ion-channel subunits essential for neuronal excitability and have long been implicated in epilepsy pathophysiology. Mutations in *KCNQ2* are associated with a spectrum of neonatal-onset epilepsy syndromes, including benign familial neonatal seizures (BFNS [MIM: 121200]) and developmental and epileptic encephalopathy (DEE)

(https://www.kcnq2cure.org/kcnq2-epilepsy). Likewise, *CACNA1A* is linked to various types of epilepsy, including DEE (https://www.cacna1a.org/what-is-cacna1a), and mutations in *CACNA1B* have been associated with disorders ranging from rare episodic ataxia [MIM: 108500] to genetic generalized epilepsy (GGE).[22,23] Within the unified model, *KCNQ2* reaches an overall p-value of 2.51×10^{-9}, driven primarily by constraint and pathogenicity effects (p-values of 1.18×10^{-2} and 5.83×10^{-7}, respectively). Genes *CACNA1A* and *CACNA1B* show unified p-values of $5.82 \times 10^{-7}$and 5.12×10^{-6}, respectively, with both signals dominated by pLoF effects (4.10×10^{-5} and 7.80×10^{-8}).

Gene Name	Group	# of Variants	Constraint p-value	GERP p-value	Pathogenicity p-value	pLoF p-value	Missense p-value	Epi25 DM p-value	Epi25 PTV p-value	Unified Model p-value
PCDHAC2	DEE	215	1.06E-01	2.64E-03	3.67E-15	4.10E-02	1.15E-07	6.86E-01	7.42E-01	5.40E-17
SP4	GGE	234	9.99E-01	3.72E-02	3.74E-01	9.27E-01	1.72E-04	7.11E-01	3.59E-02	4.25E-13
SCN2A	DEE	410	3.38E-01	5.84E-01	2.92E-13	6.28E-08	1.06E-04	4.16E-04	2.63E-02	3.40E-12
STXBP1	DEE	56	6.43E-01	3.12E-01	5.48E-09	8.89E-12	1.87E-02	8.43E-05	2.22E-05	6.54E-11
SLC6A1	DEE	60	2.71E-01	2.38E-01	2.69E-11	3.82E-07	2.64E-01	1.71E-04	1.82E-03	6.93E-11
KCNQ2	DEE	149	1.18E-02	2.03E-01	5.83E-07	2.63E-01	2.60E-02	3.44E-02	2.75E-01	2.51E-09
KCNB1	DEE	121	6.65E-02	2.87E-01	2.01E-04	6.63E-09	8.91E-01	1.32E-03	4.61E-04	9.46E-09
RYR2	GGE	720	6.86E-01	2.36E-02	1.27E-01	2.03E-15	3.22E-01	6.45E-01	8.60E-06	5.17E-08
SATB1	DEE	84	1.23E-01	4.81E-03	2.19E-01	9.31E-14	5.35E-02	4.29E-01	4.46E-02	5.28E-08
CACNA1A	DEE	278	1.74E-01	7.73E-01	1.55E-07	5.27E-04	3.97E-01	1.31E-02	3.67E-01	7.86E-08
DYRK1A	DEE	95	9.56E-01	8.94E-01	5.87E-01	2.01E-13	1.51E-01	7.16E-01	3.49E-04	2.52E-07
CACNA1A**	GGE	289	2.02E-04	8.14E-01	5.88E-02	4.10E-05	1.84E-01	1.48E-01	7.81E-03	5.82E-07
ATP2B2**	DEE	191	4.69E-04	6.40E-01	7.60E-01	5.93E-08	3.70E-01	2.03E-01	6.22E-03	2.74E-06
CACNA1B**	GGE	370	1.14E-01	9.62E-01	2.14E-02	7.80E-08	5.50E-02	2.14E-02	2.48E-04	5.12E-06
KDM5B**	DEE	268	7.38E-01	4.12E-01	6.16E-03	1.76E-17	2.87E-07	7.29E-01	2.71E-03	3.10E-04

Table 1: Comparison of p-values for each moderator of the unified model, the unified model p-value, and the Epi25 damaging missense and PTV p-values across four epilepsy gene groups.

Results are the set of significant ($p < 3.4 \times 10^{-7}$) gene-group pairs which did not meet exome-wide significance in the Epi25 study, as well as four important gene-group pairs not quite meeting this threshold, denoted by asterisks. The highlighted rows denote genes with known neurological function.

Several additional genes demonstrate newly significant or substantially strengthened associations under the unified framework. ATP2B2, which encodes a plasma-membrane calcium-transporting ATPase, achieves an overall p-value of 2.74×10^{-6}, improving upon its previously nonsignificant Epi25 result. Its signal is driven by pLoF effects (5.93×10^{-8}), consistent with prior evidence linking *ATP2B2* dysregulation to neuronal calcium signaling and seizure susceptibility.[24,25] KDM5B, a chromatin-modifying gene associated with neurodevelopmental disorders,[26–28] shows a unified p-value of 3.10×10^{-4}, dominated by a strong pLoF effect ($p = 1.76 \times 10^{-17}$). Finally, RYR2, which encodes a ryanodine receptor

critical for calcium-induced calcium release, displays a unified p-value of 5.17×10^{-8} and a pronounced pLoF contribution ($p = 2.03 \times 10^{-16}$), improving substantially over the Epi25 PTV analysis ($p = 8.60 \times 10^{-6}$). Two recent studies have identified *RYR2* mutations in children with benign epilepsy of childhood with centrotemporal spikes (BECTS [MIM: 117100]) and focal epilepsy [MIM: 604364].[29,30] Together, these results highlight that incorporating constraint-based, structural, and functional annotations into a single meta-regression substantially increases sensitivity to detect both known and emerging epilepsy genes while maintaining biological interpretability across distinct variant classes and epilepsy phenotypes.

Now, we consider the genes not meeting the "importance" criterion that meet the exome-wide significance threshold. Three main patterns emerge.

1. **Established DEE/GGE genes whose signals strengthen under integration.** A couple ion-channel and synaptic genes already discussed above—*KCNQ2, CACNA1A and RYR2*—now exceed the exome-wide significance threshold despite sub-threshold Epi25 results. These improvements are driven by moderators consistent with known biology (for example, pLoF effects for calcium-handling genes and constraint or pathogenicity effects for *KCNQ2*), so their mechanisms are not restated here.

2. **Synaptic release, scaffold, and inhibitory-tone genes newly cross the threshold.** *STXBP1* (DEE) [MIM: 602926] and *SLC6A1* (DEE) [MIM: 137165] show strong moderator support from pLoF and pathogenicity terms that elevate their overall significance beyond what Epi25 reported for damaging missense or truncating variants. Similarly, *DLG4* (DEE) [MIM: 602887] and *KCNB1* (DEE) [MIM: 600397] achieve exome-wide significance through coherent constraint and pLoF contributions, reinforcing the role of synaptic and potassium-channel pathways in epileptic encephalopathy.

3. **Regulatory and emerging candidates.** The unified model also brings to significance several transcriptional or chromatin regulators—*SETD1B* (GGE) [MIM: 611055], *SATB1* (DEE) [MIM: 602075], and *SP4* (GGE) [MIM: 600540]—whose strongest moderators are constraint and missense pathogenicity. Finally, Table 1 identifies more exploratory candidates such as *PCDHAC2* (DEE) [MIM: 606321], which is mechanistically plausible given its role in neuronal adhesion but currently lacks direct clinical validation.

Overall, Table 1 demonstrates that the unified framework systematically converts multiple near-miss Epi25 loci into exome-wide associations across distinct epilepsy subgroups while maintaining biological interpretability by variant class. The highly-significant rows of Table 1 isolates those genes whose associations now cross the study-wide boundary, underscoring how the integrated model enhances detection power for both established and emerging epilepsy genes. We visually summarize these results across the four represented gene groups in Figure 5.

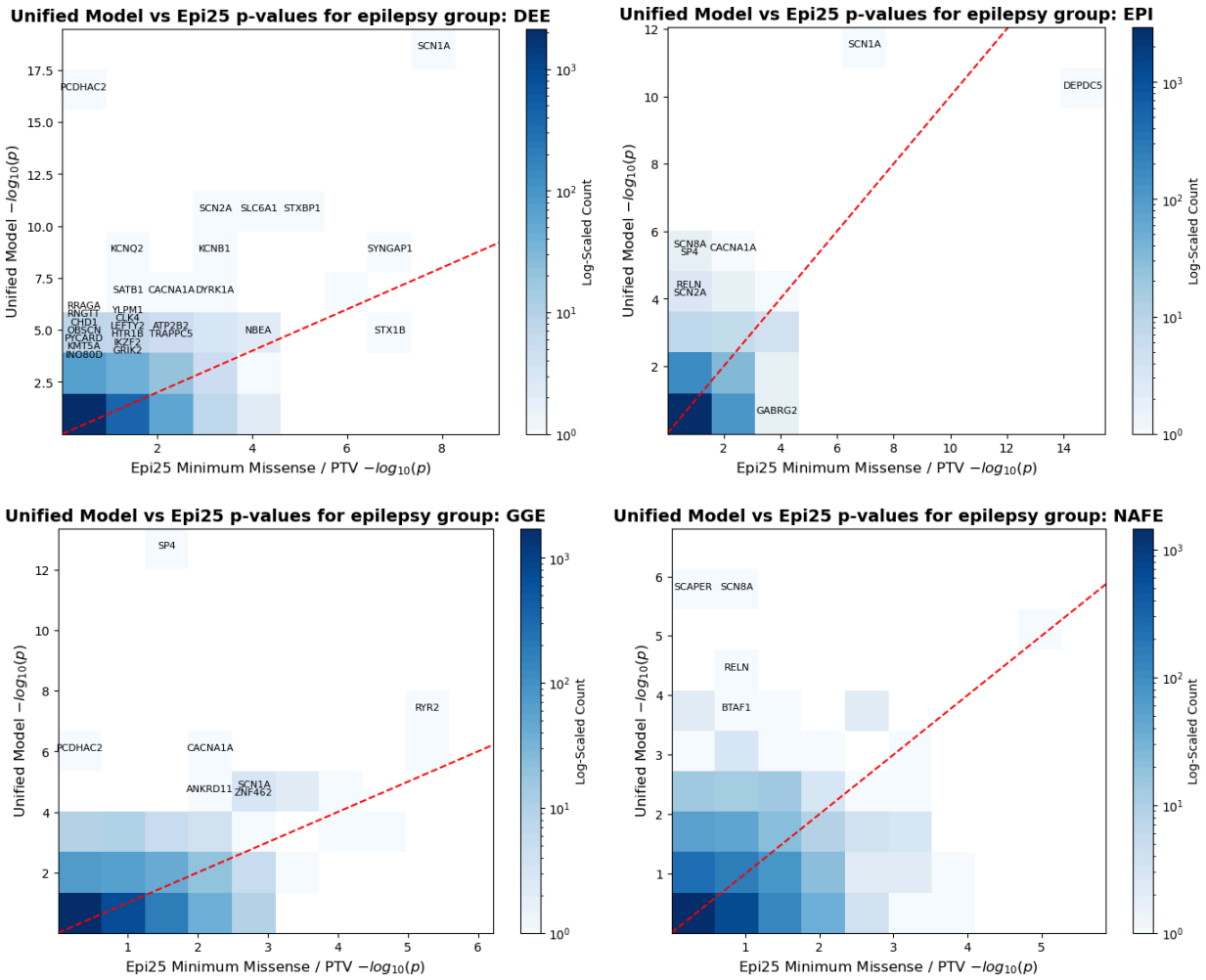


Figure 5: Comparison of unified model p-values and minimum Epi25 damaging missense or PTV p-values across four epilepsy gene groups. Each subplot shows a 2D histogram (log-scaled count) of $-\ log_{10}(p)$ values. The red dashed diagonal indicates equivalence between the two models. Gene names are annotated where the absolute $log_{10}(p)$ difference exceeds 2 and at least one model achieves $p < 10^{-4}$; labels were grouped by bin coordinates to improve readability and vertically stacked to avoid overlap.

Figure 5 summarizes the comparative performance of the unified meta-regression model relative to the Epi25 burden analyses across four epilepsy subgroups: developmental and epileptic encephalopathy (DEE), generalized genetic epilepsy (GGE), focal epilepsy (EPI), and non-acquired focal epilepsy (NAFE). Each panel plots the unified model p-values against the minimum of the Epi25 missense or protein-truncating variant (PTV) p-values for the same gene, with the red diagonal indicating equivalence between the two approaches. Points below the diagonal represent genes for which the unified model yields stronger associations. Across all four epilepsy subgroups, the majority of genes fall below this line, illustrating that the unified model systematically improves gene-level significance relative to the Epi25 analysis. This trend is especially pronounced for genes with intermediate Epi25 signals—those near but not surpassing exome-wide significance—where the unified framework integrates complementary

sources of evidence (constraint, pathogenicity, and pLoF) to amplify biologically meaningful signals. Although many canonical epilepsy genes—including *SCN1A* [MIM: 182389], *SCN2A* [MIM: 182390], and *STXBP1* [MIM: 602926] in DEE and *RYR2* and *CACNA1A* in GGE—exhibit stronger signals under the unified framework, a subset—notably STX1B [MIM: 601485] in DEE and DEPDC5 [MIM: 614191] and GABRG2 [MIM: 147164] in EPI—show comparatively weaker associations. This discrepancy likely reflects the weighting of annotation priors in the meta-regression, where genes with higher predicted tolerance or limited AlphaMissense or pLoF support receive less weight despite burden enrichment. Such cases may therefore represent genes where statistical burden and biological annotation provide complementary but non-concordant evidence. Overall, these results demonstrate that the unified approach both recapitulates known epilepsy genes and enhances statistical power to detect moderate-effect associations across heterogeneous epilepsy phenotypes.

Schizophrenia

The focus of this paper is on the application of our unified meta-regression model to epilepsy; however, there are likely extensions of this methodology to other neurological and psychiatric disorders. In this brief subsection, we discuss preliminary findings from extending the unified meta-regression framework to schizophrenia [MIM: SCZD] using data from the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium.[31] These findings, demonstrated below in Figure 6 and Table 2 are more mixed than our epilepsy findings, prompting more specific model development for each disease of interest.

The SCHEMA dataset represents one of the largest and most comprehensive exome-sequencing resources for psychiatric genetics, aggregating exome data from 24,248 schizophrenia cases, 97,322 controls, and 3,402 parent–proband trios across multiple international cohorts. We use the SCHEMA Phase I dataset as a benchmark to evaluate whether our unified meta-regression can replicate or refine these associations by integrating complementary information from sequence-based constraint, structural pathogenicity predictions, and predicted loss-of-function effects. Unlike the epilepsy analyses, where phenotype subgroups are clearly defined, the schizophrenia dataset aggregates all cases under a single “meta” group, allowing us to assess the generalizability of our framework across a more heterogeneous, polygenic psychiatric disorder.

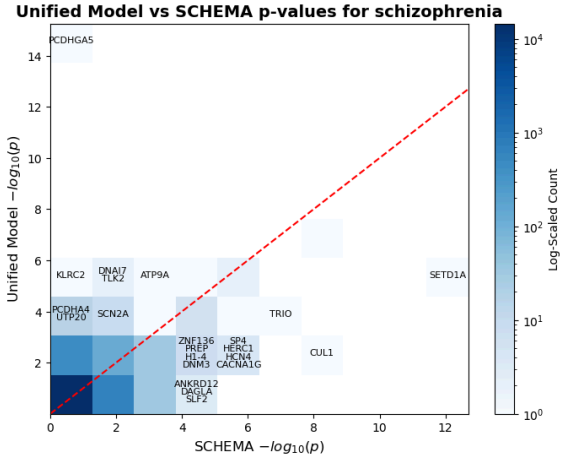


Figure 6: Comparison of unified model p-values and SCHEMA p-values across genes in the schizophrenia “meta” group. The plot shows a 2D histogram (log-scaled count) of $-\ log_{10}(p)$ values. The red dashed diagonal indicates equivalence between the two models. Gene names are annotated where the absolute $log_{10}(p)$ difference exceeds 2 and at least one model achieves $p < 10^{-4}$; labels were grouped by bin coordinates to improve readability and vertically stacked to avoid overlap.

Building on this benchmark, we next compare the unified model p-values with those from the SCHEMA meta-analysis to assess whether our integrative approach captures known risk genes or identifies additional candidates with similar biological signatures. Ideally, the unified p-values should equal or surpass the strength of the SCHEMA benchmark, indicating comparable or improved performance. However, results were mixed. The largest improvement was observed for *PCDHGA5*, which displayed a marked gain in significance under the unified framework ($p = 5.64 \times 10^{-15}$ versus $p = 1.09 \times 10^{-1}$ in SCHEMA). Moderate improvements were also seen for *KLRC2* [MIM:602891], *DNAI7* [MIM: 616906], and *TLK2* [MIM: 608439]. In contrast, several genes that are highly significant in SCHEMA, including *SETD1A* [MIM: 611052], *CUL1* [MIM:603134], and *TRIO* [MIM: 601893], did not replicate under the unified model, reflecting incomplete power and potential model mismatches between epilepsy and schizophrenia. Figure 6 summarizes these results across all genes analyzed in the SCHEMA dataset, comparing the unified model and SCHEMA p-values by gene.

Overall, these results indicate that while the unified model can recover and occasionally strengthen individual schizophrenia associations, its performance is less consistent than in epilepsy. This limitation likely stems from smaller effect sizes, higher genetic heterogeneity, and the absence of explicit modeling of *de novo* mutations—an important contributor to rare variant risk in schizophrenia as demonstrated by the SCHEMA consortium. Future work should therefore prioritize integrating de novo mutation information and increasing sample size to improve the power and stability of unified model estimates in schizophrenia and other highly polygenic psychiatric disorders.

Gene Name	Group	# of Variants	Constraint p-value	GERP p-value	Pathogenicity p-value	pLoF p-value	Missense p-value	SCHEMA p-value	Unified Model p-value
PCDHGA5	meta	1460	4.55E-16	1.06E-01	1.11E-02	2.99E-01	7.85E-08	1.09E-01	5.64E-15
XPO7	meta	241	5.28E-01	1.14E-01	6.80E-03	1.12E-09	3.54E-04	7.18E-09	6.78E-08
SETD1A	meta	627	1.77E-01	9.74E-01	4.98E-02	4.38E-09	2.62E-02	2.00E-12	1.10E-06
KLRC2	meta	50	1.78E-06	6.19E-04	2.29E-02	8.60E-01	6.50E-01	3.42E-01	3.59E-06
ATP9A	meta	278	8.71E-01	8.49E-01	4.40E-01	2.54E-11	4.46E-03	3.72E-04	3.59E-06
ASH1L	meta	918	2.63E-02	6.04E-01	5.55E-01	9.05E-01	1.25E-01	3.77E-05	6.32E-06
DNAI7	meta	202	8.82E-01	8.55E-03	2.82E-02	8.46E-01	2.50E-01	1.51E-02	7.87E-06
TLK2	meta	111	5.95E-01	3.42E-03	1.76E-05	6.52E-04	1.53E-01	3.45E-03	1.06E-05
GRIN2A	meta	467	3.13E-01	9.05E-01	4.51E-02	4.40E-08	7.53E-03	7.37E-07	1.51E-05
AKAP11	meta	667	9.94E-01	8.49E-01	6.31E-01	3.62E-11	6.17E-02	8.28E-06	2.29E-05
PCDHA4	meta	1027	1.64E-01	3.18E-06	1.11E-02	3.87E-01	6.52E-01	4.34E-01	4.61E-05
SCN2A	meta	689	7.60E-01	7.44E-01	5.32E-06	1.80E-02	3.88E-01	2.32E-02	5.48E-05
SRRM2	meta	97	6.37E-01	9.08E-01	8.73E-11	8.90E-11	5.07E-02	1.53E-05	5.49E-05
UTP20	meta	959	5.63E-05	3.57E-01	4.49E-01	1.11E-04	5.11E-02	1.98E-01	9.41E-05

Table 2: Comparison of p-values for each moderator of the unified model, the unified model p-value, and the SCHEMA p-value. Results are shown for the gene-group pairs with unified model p-value at most 1.0×10^{-4}, ordered in ascending order by unified model p-value.

Discussion

Our study introduces a unified statistical framework that integrates sequence-level constraint, structure-based missense pathogenicity, predicted loss-of-function (pLoF) annotations, and rejected substitution (RS) scores to improve detection of epilepsy-associated genes. By combining Hidden Markov Model (HMM)–derived constraint predictions with AlphaMissense (AM) pathogenicity probabilities, pLoF/missense indicators, and GERP RS scores within a weighted meta-regression model, we demonstrate that distinct biological signals can be systematically combined to enhance statistical power in rare variant association studies. The unified model consistently identifies both known epilepsy genes and novel candidates that fall below genome-wide significance in standard population-based analyses, illustrating how sequence- and structure-informed priors can refine gene discovery in heterogeneous neurodevelopmental disorders.

Our findings emphasize the complementary roles of constraint, pathogenicity, and functional loss in capturing diverse aspects of mutational intolerance. Genes such as *KCNQ2*, *CACNA1A*, and *CACNA1B*—each encoding ion-channel subunits essential for neuronal excitability—show improved gene-level significance driven by different model components: constraint and pathogenicity for *KCNQ2*, and pLoF for *CACNA1A* and *CACNA1B*. Beyond these well-established genes, the unified model strengthens signals for genes with limited prior evidence of causality, including *ATP2B2*, *KDM5B*, and *RYR2*. The discovery of significant or suggestive associations in these genes, each linked to calcium signaling, chromatin regulation, or neuronal excitability, highlights the model's ability to detect biologically plausible yet previously underpowered relationships. Together, these results underscore the value of integrating orthogonal predictors of mutational impact into a single inferential framework for gene-level discovery.

While promising, the utility of this model is bounded by the accuracy and completeness of its inputs. Structural pathogenicity scores such as AM rely heavily on protein coverage and model quality, whereas constraint inference is sensitive to sequencing depth and variant ascertainment. As future datasets incorporate more diverse ancestries and deeper coverage, the constraint model can be retrained to reduce bias and improve generalization. Methodologically, extending the HMM into a higher-order or neural-sequence model—such as a recurrent or transformer-based architecture—could capture more complex dependencies in mutational depletion patterns. Similarly, nonlinear extensions of the meta-regression could allow for interactions among predictors (e.g., between structural intolerance and constraint) that remain unexplored in the current linear framework. Moreover, because the framework emphasizes depletion-based and loss-of-function–oriented predictors, it may underrepresent gain-of-function associations—an important consideration for epilepsy genes such as SCN2A and KCNQ2, where different mutational mechanisms can produce distinct phenotypes.[32,33] Future extensions could integrate functional assay data or gain-of-function–specific pathogenicity predictors to better capture this class of effects.

Finally, this unified framework has broader applicability beyond epilepsy. Because it models joint variant-level contributions from multiple biological priors, the approach can be extended to other neurodevelopmental and psychiatric disorders—such as autism, schizophrenia, and bipolar disorder—where rare variant heterogeneity limits discovery power. In this study, we applied the model to an independent schizophrenia dataset from the SCHEMA consortium as a proof of generalizability. The unified model recovered several known risk loci and strengthened associations for select candidates (for example, *PCDHGA5* and *TLK2*), though its overall performance was less consistent than in epilepsy, likely reflecting smaller effect sizes and greater locus heterogeneity in schizophrenia. This cross-disorder validation highlights both the adaptability of the framework and its dependence on disease architecture, supporting future applications to other large-scale sequencing studies aimed at integrating functional priors into rare variant discovery.

Supporting Information

Data and Code Availability

The code for our model development and data analysis is hosted on a GitHub repository at https://github.com/healthcare-medicine-ai/wgs-constraint-llm, along with most data and results. The full WES constraint predictions generated for AoU and RGC-ME are available at https://doi.org/10.6084/m9.figshare.27184245.v1.

Acknowledgements

M.A.R. is in part supported by National Human Genome Research Institute (NHGRI) under award R01HG010140, and by the National Institutes of Mental Health (NIMH) under award R01MH124244 both of the National Institutes of Health (NIH).

Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies; funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of interests

M.A.R. is a cofounder of Broadwing Bio and a consultant for Insitro.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used perplexity.ai and chat-gpt in order to investigate the relationship between epilepsy and certain genes in the existing literature and synthesize references to related work. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Online resources

Regeneron Genetics Center Million Exome Variant Browser
https://rgc-research.regeneron.com/me/license-and-terms-of-use

AlphaMissense predictions
https://console.cloud.google.com/storage/browser/dm_alphamissense

GERP RS Scores
https://genome-asia.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=compGeno&hgta_track=allHg19RS_BW&hgta_table=allHg19RS_BW&hgta_doSchema=describe+table+schema

gnomAD v4.1.0
https://gnomad.broadinstitute.org/

Epi25 Browser
https://epi25.broadinstitute.org/

SCHEMA Browser
https://schema.broadinstitute.org/

Supplementary Materials

Gene Name	Group	# of Variants	Constraint p-value	GERP p-value	Pathogenicity p-value	pLoF p-value	Missense p-value	Epi25 DM p-value	Epi25 PTV p-value	Unified Model p-value
SCN1A	DEE	273	7.50E-05	2.57E-01	6.50E-08	3.70E-20	8.56E-01	1.44E-05	6.32E-09	3.21E-19
PCDHAC2	DEE	215	1.06E-01	2.64E-03	3.67E-15	4.10E-02	1.15E-07	6.86E-01	7.42E-01	5.40E-17
PCDHGA11	DEE	267	5.40E-01	5.54E-01	1.69E-17	7.87E-06	1.80E-13	nan	7.30E-01	4.55E-16
PCDHGA1	GGE	661	4.96E-05	1.66E-06	3.13E-06	6.17E-02	9.34E-04	nan	4.23E-01	8.26E-14
DEPDC5	NAFE	349	9.48E-01	1.50E-01	8.86E-01	5.66E-24	3.92E-03	3.23E-01	0.00E+00	2.02E-13
SP4	GGE	234	9.99E-01	3.72E-02	3.74E-01	9.27E-01	1.72E-04	7.11E-01	3.59E-02	4.25E-13
SCN2A	DEE	410	3.38E-01	5.84E-01	2.92E-13	6.28E-08	1.06E-04	4.16E-04	2.63E-02	3.40E-12
PCDHGA10	NAFE	310	7.51E-13	1.25E-01	2.50E-01	4.00E-01	5.00E-07	nan	9.40E-01	5.44E-12
SCN1A	EPI	440	2.80E-02	4.30E-01	3.76E-07	3.17E-07	1.21E-03	1.60E-04	3.49E-08	8.75E-12
PCDHGA1	NAFE	658	1.88E-02	2.92E-06	8.76E-05	5.05E-02	1.43E-04	nan	6.96E-01	2.25E-11
STXBP1	DEE	56	6.43E-01	3.12E-01	5.48E-09	8.89E-12	1.87E-02	8.43E-05	2.22E-05	6.54E-11
SLC6A1	DEE	60	2.71E-01	2.38E-01	2.69E-11	3.82E-07	2.64E-01	1.71E-04	1.82E-03	6.93E-11
SYNGAP1	DEE	154	8.25E-02	4.37E-01	7.24E-01	5.18E-18	6.70E-01	1.91E-01	5.95E-08	2.02E-10
DEPDC5	EPI	410	2.73E-01	2.56E-01	4.97E-01	9.45E-15	1.65E-05	3.69E-01	3.44E-15	2.22E-10
PCDHA1	NAFE	519	8.51E-10	7.01E-01	8.01E-03	1.66E-02	2.87E-02	nan	8.36E-02	4.72E-10
KANSL1	DEE	271	3.06E-01	3.65E-02	2.77E-02	4.14E-02	9.81E-01	nan	4.45E-04	1.54E-09
KCNQ2	DEE	149	1.18E-02	2.03E-01	5.83E-07	2.63E-01	2.60E-02	3.44E-02	2.75E-01	2.51E-09
TBL1XR1	DEE	46	4.15E-01	1.20E-02	7.78E-09	nan	1.04E-18	7.44E-05	nan	2.87E-09
KCNB1	DEE	121	6.65E-02	2.87E-01	2.01E-04	6.63E-09	8.91E-01	1.32E-03	4.61E-04	9.46E-09
SLC22A1	DEE	156	4.90E-04	4.17E-02	1.96E-04	4.78E-14	5.28E-06	nan	4.18E-02	1.93E-08
CD8B	DEE	82	7.36E-08	5.82E-02	5.58E-01	1.54E-07	2.67E-01	nan	2.08E-02	2.81E-08
RYR2	GGE	720	6.86E-01	2.36E-02	1.27E-01	2.03E-15	3.22E-01	6.45E-01	8.60E-06	5.17E-08
SATB1	DEE	84	1.23E-01	4.81E-03	2.19E-01	9.31E-14	5.35E-02	4.29E-01	4.46E-02	5.28E-08
PCDHA1	GGE	492	1.97E-01	9.58E-07	2.37E-05	4.44E-05	1.96E-01	nan	5.27E-01	6.70E-08
PCDHGA10	EPI	369	5.89E-07	2.12E-02	6.16E-03	3.33E-01	1.99E-02	nan	7.05E-01	7.53E-08
CACNA1A	DEE	278	1.74E-01	7.73E-01	1.55E-07	5.27E-04	3.97E-01	1.31E-02	3.67E-01	7.86E-08
GABRB3	DEE	38	4.47E-01	1.97E-02	6.09E-07	nan	4.07E-18	9.62E-03	nan	7.89E-08
SLC6A1	GGE	81	8.64E-01	2.39E-02	7.01E-08	nan	4.43E-11	3.01E-06	nan	1.56E-07
SLC25A11	DEE	52	8.96E-07	8.29E-01	4.16E-04	nan	6.44E-21	1.63E-01	nan	2.35E-07
DYRK1A	DEE	95	9.56E-01	8.94E-01	5.87E-01	2.01E-13	1.51E-01	7.16E-01	3.49E-04	2.52E-07
ANKRD11	DEE	650	6.71E-01	7.62E-01	8.42E-02	1.84E-17	6.47E-01	1.47E-01	1.20E-06	2.86E-07
STIM2	GGE	231	1.92E-08	1.58E-01	2.02E-01	5.07E-01	1.05E-01	nan	9.92E-01	3.19E-07
NPRL3	NAFE	164	8.55E-01	8.25E-01	1.20E-02	2.74E-10	2.72E-04	nan	1.35E-06	3.21E-07
CCT7	DEE	92	4.36E-09	7.95E-01	6.93E-01	4.87E-02	3.25E-03	nan	5.13E-01	3.44E-07
MRPS22	DEE	93	1.33E-06	1.28E-02	9.91E-01	2.34E-11	1.57E-10	nan	5.18E-02	3.97E-07
CASKIN2	DEE	270	2.98E-01	1.39E-04	8.52E-02	4.37E-10	7.78E-03	nan	2.96E-03	4.00E-07

TTN	DEE	7773	3.56E-08	1.09E-01	2.14E-01	8.58E-01	9.64E-01	nan	4.65E-01	4.01E-07
TEAD1	GGE	127	8.24E-04	6.16E-01	8.14E-04	nan	7.87E-10	nan	nan	4.02E-07
PCDHAC2	GGE	224	5.48E-01	2.73E-02	9.68E-07	1.73E-01	2.43E-02	6.78E-01	7.57E-01	4.29E-07
PCDHA1	DEE	464	2.38E-01	7.67E-03	5.21E-01	4.88E-25	1.14E-07	nan	6.61E-02	5.31E-07
CACNA1A	GGE	289	2.02E-04	8.14E-01	5.88E-02	4.10E-05	1.84E-01	1.48E-01	7.81E-03	5.82E-07
KBTBD8	DEE	120	7.85E-08	9.28E-01	6.48E-02	8.18E-04	1.02E-13	nan	7.07E-01	1.16E-06
PCDHA10	DEE	369	9.48E-01	1.06E-02	5.26E-05	5.72E-02	1.18E-17	nan	3.26E-01	1.17E-06
NBEA	DEE	619	1.89E-03	9.93E-01	4.08E-01	3.66E-14	4.22E-01	9.51E-01	1.64E-04	1.47E-06
SCN8A	NAFE	389	1.76E-04	6.12E-01	4.14E-03	8.15E-01	8.91E-02	1.90E-01	9.97E-01	1.56E-06
YLPM1	DEE	439	1.02E-06	5.00E-02	6.29E-01	1.08E-09	3.03E-12	5.65E-01	5.81E-02	1.67E-06
EGF	DEE	268	8.71E-07	8.81E-01	3.78E-03	2.02E-11	2.01E-29	nan	8.33E-01	1.68E-06
CTH	DEE	58	6.55E-01	9.58E-04	1.44E-03	1.85E-12	1.79E-01	nan	9.90E-02	1.80E-06
LMO7	DEE	358	4.69E-07	5.66E-02	4.80E-71	2.16E-22	5.93E-118	nan	7.22E-01	1.88E-06
APOC3	DEE	32	5.35E-08	9.45E-01	1.11E-01	6.76E-03	2.99E-03	nan	2.94E-01	2.48E-06
SETD1B	GGE	354	6.88E-02	7.09E-01	4.59E-02	2.12E-07	3.31E-02	2.75E-01	6.14E-06	2.70E-06
ATP2B2	DEE	191	4.69E-04	6.40E-01	7.60E-01	5.93E-08	3.70E-01	2.03E-01	6.22E-03	2.74E-06
CLK4	DEE	84	1.28E-04	2.58E-01	1.56E-02	2.54E-03	1.83E-09	8.21E-02	7.98E-01	3.00E-06
ATAD2	NAFE	228	1.96E-01	7.96E-01	2.52E-07	1.61E-02	8.22E-01	nan	4.77E-01	3.17E-06
CACNA1A	EPI	384	1.23E-01	1.83E-01	2.99E-04	4.19E-03	1.94E-02	2.17E-01	1.40E-02	3.30E-06
SCAPER	NAFE	253	3.51E-01	6.13E-03	8.30E-07	5.90E-02	2.20E-01	7.78E-01	5.98E-01	3.56E-06
HSD17B13	DEE	71	4.88E-02	1.23E-04	1.23E-03	3.76E-08	2.28E-01	nan	5.76E-02	3.67E-06
KCNIP3	DEE	91	7.07E-01	4.81E-02	1.44E-06	1.01E-02	1.94E-02	nan	1.00E+00	3.69E-06
CHD2	DEE	221	1.17E-01	1.42E-01	1.13E-02	8.95E-10	5.55E-01	2.49E-04	3.49E-03	3.72E-06
RRAGA	DEE	25	5.18E-07	2.76E-01	3.66E-01	3.48E-02	8.35E-09	6.02E-01	4.23E-01	4.29E-06
OR2A2	DEE	84	2.95E-06	4.20E-03	4.40E-02	8.95E-02	2.61E-07	nan	nan	4.34E-06
SLC2A1	GGE	89	1.13E-01	5.71E-03	3.64E-04	8.92E-01	1.54E-01	5.64E-05	5.75E-01	4.66E-06
PSMA7	GGE	29	4.74E-07	9.45E-01	9.25E-02	nan	5.19E-03	4.85E-01	nan	4.79E-06
CACNA1B	GGE	370	1.14E-01	9.62E-01	2.14E-02	7.80E-08	5.50E-02	2.14E-02	2.48E-04	5.12E-06
DLG4	DEE	76	6.44E-01	2.90E-03	4.28E-01	5.61E-12	9.52E-01	6.24E-01	1.16E-04	5.28E-06
CHD2	GGE	231	5.49E-01	6.70E-01	5.53E-03	2.05E-09	1.15E-02	9.72E-01	4.45E-04	5.65E-06
LEFTY2	DEE	89	3.43E-05	1.64E-01	9.71E-01	9.28E-10	1.52E-02	6.82E-01	4.86E-02	5.82E-06
SIGLEC6	DEE	96	1.11E-06	4.78E-02	2.45E-01	1.08E-04	4.03E-10	nan	6.53E-01	6.50E-06
RGPD1	DEE	57	8.52E-04	8.38E-04	8.37E-04	1.65E-01	3.65E-11	nan	8.77E-01	6.82E-06
SLC39A2	DEE	68	2.66E-06	1.06E-02	8.56E-01	7.85E-06	3.79E-06	nan	9.41E-01	6.98E-06
TANC2	DEE	359	1.13E-05	4.83E-01	6.53E-01	5.08E-05	6.90E-02	nan	6.90E-02	7.04E-06
SCN8A	EPI	481	3.53E-03	7.86E-01	6.72E-04	6.34E-01	7.48E-01	2.98E-02	8.96E-01	8.24E-06
GNA12	DEE	57	1.38E-06	5.60E-02	5.21E-01	4.08E-02	2.54E-04	nan	2.89E-01	8.30E-06
SP4	EPI	285	7.42E-01	8.21E-02	9.47E-01	9.37E-01	2.56E-02	6.17E-01	1.23E-01	8.49E-06
TAS2R20	DEE	69	7.34E-07	4.19E-01	4.15E-01	3.37E-06	1.99E-06	nan	nan	8.79E-06
SYNE1	DEE	1981	4.25E-02	4.21E-01	5.39E-03	3.37E-38	2.54E-46	nan	8.71E-04	9.03E-06
SCN8A	DEE	317	1.05E-01	7.05E-01	3.17E-06	3.06E-02	8.98E-04	2.57E-04	3.53E-01	9.82E-06
ANKRD11	GGE	682	5.53E-01	5.03E-01	1.11E-05	1.26E-04	4.67E-01	3.45E-01	1.31E-02	1.03E-05
KDELR2	DEE	45	8.52E-05	7.48E-01	1.94E-03	1.53E-01	4.18E-05	nan	6.02E-01	1.03E-05
LMO7	GGE	373	2.45E-06	4.93E-02	1.12E-17	3.78E-05	2.33E-34	nan	8.93E-01	1.03E-05
HMCN1	DEE	1215	3.16E-06	7.92E-01	4.97E-01	2.30E-28	2.46E-49	nan	1.26E-01	1.19E-05
HSPA14	DEE	61	7.71E-04	1.62E-04	1.70E-01	1.17E-02	8.95E-05	nan	7.17E-01	1.22E-05
NPSR1	DEE	77	8.42E-01	4.13E-01	6.73E-01	3.47E-12	7.32E-01	nan	4.87E-02	1.23E-05
ZNF302	DEE	107	3.97E-01	2.71E-06	1.13E-32	9.59E-12	9.04E-43	nan	3.07E-03	1.25E-05
SCN1A	GGE	274	1.86E-02	8.65E-01	2.91E-03	4.76E-05	3.42E-01	9.90E-02	2.41E-03	1.27E-05

PHF20	GGE	176	2.33E-03	1.24E-01	4.33E-01	5.41E-06	7.54E-03	nan	4.87E-03	1.32E-05
HTR1B	DEE	69	5.96E-03	6.15E-01	2.44E-02	2.58E-08	1.80E-01	4.86E-01	1.87E-02	1.50E-05
LGI1	NAFE	98	3.12E-01	2.56E-01	2.63E-03	9.27E-06	4.25E-03	7.88E-03	1.37E-05	1.51E-05
HNMT	DEE	29	5.40E-07	5.87E-01	4.48E-01	8.86E-03	5.54E-04	nan	9.98E-01	1.55E-05
KIF26B	DEE	481	6.92E-03	2.04E-01	7.38E-01	4.36E-14	6.43E-02	nan	1.77E-04	1.61E-05
SCN5A	GGE	425	1.83E-01	4.66E-02	3.95E-05	5.32E-05	6.39E-01	nan	4.65E-01	1.73E-05
ZNF462	GGE	465	9.32E-02	1.23E-01	1.55E-01	8.77E-08	5.70E-02	4.87E-01	2.54E-03	1.89E-05
C1orf52	GGE	60	4.13E-01	5.83E-01	4.52E-01	1.78E-08	1.36E-02	nan	1.23E-01	1.89E-05
PCDHGA8	DEE	236	3.66E-07	9.70E-01	2.00E-01	4.95E-08	3.51E-19	nan	3.05E-01	1.96E-05
IL22	DEE	34	7.07E-01	4.16E-02	1.89E-06	nan	5.29E-01	nan	nan	2.16E-05
PRND	DEE	44	3.71E-01	8.68E-01	8.01E-01	1.12E-09	1.89E-01	nan	2.07E-01	2.17E-05
RBBP9	DEE	51	3.38E-06	8.29E-01	2.54E-01	3.19E-01	1.88E-05	nan	6.09E-01	2.18E-05
APC	DEE	621	9.30E-07	2.37E-01	4.01E-01	3.89E-06	3.15E-12	nan	9.03E-01	2.37E-05
IKZF2	DEE	115	1.09E-02	6.56E-01	3.59E-01	3.09E-08	3.05E-01	1.46E-01	2.17E-02	2.41E-05
SLC5A1	DEE	104	4.14E-06	1.03E-01	3.76E-01	1.76E-02	3.75E-10	nan	9.88E-01	2.46E-05
NPRL2	NAFE	80	9.55E-01	1.71E-01	2.49E-02	9.19E-08	5.64E-03	nan	3.62E-05	2.52E-05
PCDHAC1	GGE	252	5.49E-04	7.76E-02	4.91E-02	1.59E-03	5.81E-04	nan	2.55E-01	2.53E-05
RNGTT	DEE	64	9.19E-07	6.56E-01	5.07E-01	1.67E-02	1.69E-07	6.15E-01	6.72E-01	2.56E-05
GPR20	DEE	124	1.37E-05	7.39E-03	7.10E-01	nan	8.60E-53	nan	nan	2.57E-05
ARV1	DEE	53	1.50E-04	4.08E-01	8.09E-01	4.70E-07	6.31E-01	nan	1.03E-01	2.68E-05
TRAPPC5	DEE	38	2.81E-04	2.00E-01	9.66E-01	3.76E-06	2.14E-01	9.40E-01	1.05E-02	3.28E-05
APLF	DEE	131	5.13E-01	3.33E-03	1.63E-06	1.50E-07	9.28E-14	nan	6.80E-01	3.31E-05
AZGP1	DEE	88	4.27E-05	1.66E-01	1.75E-01	2.12E-08	1.34E-03	nan	1.39E-01	3.37E-05
CHD1	DEE	145	4.29E-07	9.22E-01	6.48E-01	5.97E-02	3.77E-03	4.16E-01	1.75E-01	3.38E-05
OBSCN	DEE	2744	6.95E-03	5.05E-03	1.33E-01	5.09E-77	6.17E-146	3.03E-01	3.39E-01	3.41E-05
OSTF1	GGE	31	8.19E-01	4.69E-02	1.57E-01	7.12E-07	1.60E-02	nan	1.48E-01	3.43E-05
GRIK2	DEE	73	2.35E-02	4.03E-01	2.09E-02	1.45E-03	5.86E-01	3.69E-01	1.16E-01	3.54E-05
IST1	DEE	98	8.34E-01	9.26E-01	2.69E-24	1.73E-15	4.79E-19	nan	1.74E-01	3.75E-05
HLA-DOB	DEE	32	5.61E-01	5.66E-01	1.09E-06	2.71E-04	1.51E-07	nan	5.63E-01	3.83E-05
PYCARD	DEE	55	1.78E-01	8.53E-01	1.92E-05	1.39E-03	2.98E-10	1.57E-01	6.88E-01	3.94E-05
LY6D	DEE	27	4.73E-01	6.75E-01	5.85E-06	4.74E-01	1.14E-04	nan	nan	3.97E-05
RIOK3	DEE	114	1.43E-04	4.58E-02	1.08E-01	5.62E-09	8.54E-03	nan	3.77E-03	4.13E-05
IFIT2	DEE	114	2.42E-01	6.98E-01	1.38E-06	1.61E-04	2.88E-06	nan	nan	4.22E-05
ALDH1L2	DEE	199	8.99E-07	1.15E-01	5.46E-01	8.80E-01	9.66E-01	nan	6.60E-01	4.37E-05
TTC39B	GGE	165	1.53E-05	8.14E-01	6.92E-03	8.02E-02	1.04E-03	nan	4.63E-01	4.37E-05
EVPL	DEE	590	1.78E-02	4.54E-04	6.79E-02	3.19E-16	2.18E-33	nan	9.56E-01	4.42E-05
BIRC6	GGE	805	4.28E-05	8.75E-01	8.75E-01	2.42E-04	7.69E-01	9.30E-01	1.96E-03	4.51E-05
ENAH	NAFE	115	3.08E-02	5.96E-01	3.25E-01	1.06E-01	9.48E-01	nan	1.37E-02	4.78E-05
PRR11	DEE	69	1.95E-06	4.28E-01	6.72E-01	2.56E-02	4.70E-06	nan	1.20E-01	5.33E-05
RELN	EPI	1677	4.94E-02	7.51E-01	2.77E-05	8.58E-02	4.82E-01	5.56E-01	9.47E-02	5.42E-05
GNAO1	DEE	56	5.30E-01	6.15E-02	1.21E-03	9.89E-02	2.37E-02	2.90E-03	1.93E-01	5.43E-05
PAK5	GGE	166	1.32E-01	9.61E-01	1.56E-01	3.93E-11	3.40E-01	nan	2.69E-01	5.62E-05
ST3GAL2	DEE	56	1.40E-05	1.07E-01	5.23E-01	nan	6.07E-26	9.25E-01	nan	5.76E-05
SNTB1	DEE	135	6.11E-06	3.32E-01	7.74E-02	1.72E-05	2.45E-04	nan	7.23E-02	6.18E-05
SLC6A1	EPI	104	9.82E-01	1.70E-01	6.65E-06	6.27E-02	2.68E-01	9.49E-05	1.59E-01	6.30E-05
PLPPR5	DEE	52	2.78E-01	6.24E-01	7.46E-05	1.18E-05	7.75E-01	6.49E-03	2.89E-03	6.34E-05
RELN	NAFE	1418	5.05E-02	1.85E-01	8.36E-05	1.82E-01	2.76E-01	4.24E-01	1.75E-01	6.69E-05
LONP1	DEE	260	2.06E-02	9.49E-01	4.30E-03	3.17E-08	5.66E-01	nan	4.33E-03	6.90E-05
KMT5A	DEE	48	1.16E-04	4.38E-02	7.81E-10	5.78E-04	5.69E-21	2.31E-01	2.90E-01	7.06E-05

MED15	DEE	128	1.96E-06	6.09E-01	3.96E-01	6.63E-02	3.55E-03	nan	3.51E-01	7.47E-05
CCR1	GGE	64	7.15E-01	3.05E-03	7.73E-01	2.58E-05	1.83E-01	nan	nan	7.69E-05
BST2	DEE	33	8.07E-01	1.46E-01	1.70E-04	6.59E-07	5.02E-01	nan	5.67E-03	7.72E-05
SCN2A	EPI	544	8.11E-01	9.17E-01	2.61E-06	1.36E-01	2.46E-01	1.88E-01	3.49E-01	7.75E-05
INO80D	DEE	188	1.02E-01	6.73E-01	3.91E-03	1.20E-05	5.31E-01	4.80E-01	3.50E-01	8.10E-05
TOPBP1	NAFE	289	4.56E-03	8.07E-02	9.05E-01	9.20E-05	4.72E-03	nan	1.21E-01	8.29E-05
TP53INP2	DEE	48	4.13E-05	1.43E-02	4.97E-01	4.72E-01	1.00E-07	nan	6.81E-01	8.82E-05
BTAF1	NAFE	544	2.37E-01	1.15E-01	3.37E-01	6.51E-06	7.00E-03	2.07E-01	1.63E-01	8.96E-05
GOT2	DEE	84	4.91E-01	9.25E-01	2.64E-02	1.89E-08	7.16E-02	nan	5.33E-02	9.04E-05
PRPF40B	DEE	219	2.44E-01	1.39E-01	3.13E-01	8.43E-14	8.59E-02	nan	5.75E-02	9.42E-05
TSC1	NAFE	245	1.55E-01	2.97E-01	1.89E-02	1.65E-06	4.43E-02	nan	1.11E-03	9.75E-05
TMEM218	DEE	27	4.61E-03	1.54E-04	7.46E-01	7.27E-05	4.13E-01	nan	2.18E-01	9.91E-05

Table S1: Comparison of p-values for each moderator of the unified model, the unified model p-value, and the Epi25 damaging missense and PTV p-values across four epilepsy gene groups. Results are shown for the gene-group pairs with unified model p-value < 1e-4, ordered by unified model p-value.

Bibliography

1. Asadi-Pooya, A.A., Brigo, F., Lattanzi, S., and Blumcke, I. (2023). Adult epilepsy. Lancet *402*, 412–424. doi: 10.1016/S0140-6736(23)01048-6. PMID: 37459868.

2. Mauritz, M., Hirsch, L.J., Camfield, P., Chin, R., Nardone, R., Lattanzi, S., and Trinka, E. (2022). Acute symptomatic seizures: an educational, evidence-based review. Epileptic Disord. *24*, 26–49. doi: 10.1684/epd.2021.1376. PMID: 34789447.

3. Shorvon, S.D., Andermann, F., and Guerrini, R. (2011). The Causes of Epilepsy: Common and Uncommon Causes in Adults and Children (Cambridge University Press). doi: 10.1017/CBO9780511921001

4. Perucca, P., Bahlo, M., and Berkovic, S.F. (2020). The genetics of epilepsy. Annu. Rev. Genomics Hum. Genet. *21*, 205–230. doi: 10.1146/annurev-genom-120219-074937. PMID: 32339036.

5. Symonds, J.D., Zuberi, S.M., and Johnson, M.R. (2017). Advances in epilepsy gene discovery and implications for epilepsy diagnosis and treatment. Curr. Opin. Neurol. *30*, 193–199. doi: 10.1097/WCO.0000000000000433. PMID: 28212175.

6. Epi25 Collaborative (2024). Exome sequencing of 20,979 individuals with epilepsy reveals shared and distinct ultra-rare genetic risk across disorder subtypes. Nat. Neurosci. *27*, 1864–1879. doi: 10.1038/s41593-024-01747-8. PMID: 39363051; PMCID: PMC11646479.

7. Wu, A.C., McMahon, P., and Lu, C. (2020). Ending the Diagnostic Odyssey-Is Whole-Genome Sequencing the Answer? JAMA Pediatr. *174*, 821–822. doi: 10.1001/jamapediatrics.2020.1522. PMID: 32597967; PMCID: PMC7928067.

8. Macnee, M., Pérez-Palma, E., López-Rivera, J.A., Ivaniuk, A., May, P., Møller, R.S., and Lal, D. (2023). Data-driven historical characterization of epilepsy-associated genes. Eur J Paediatr Neurol *42*, 82–87. doi: 10.1016/j.ejpn.2022.12.005. PMID: 36586220.

9. Oliver, K.L., Scheffer, I.E., Bennett, M.F., Grinton, B.E., Bahlo, M., and Berkovic, S.F. (2023). Genes4Epilepsy: An epilepsy gene resource. Epilepsia *64*, 1368–1375. doi: 10.1111/epi.17547. PMID: 36808730; PMCID: PMC10952165.

10. Ruggiero, S.M., Xian, J., and Helbig, I. (2023). The current landscape of epilepsy genetics: where are we, and where are we going? Curr. Opin. Neurol. *36*, 86–94. doi: 10.1097/WCO.0000000000001141. PMID: 36762645; PMCID: PMC10088099.

11. Zhang, M.-W., Liang, X.-Y., Wang, J., Gao, L.-D., Liao, H.-J., He, Y.-H., Yi, Y.-H., He, N., Liao, W.-P., and China Epilepsy Gene 1.0 Project (2024). Epilepsy-associated genes: an update. Seizure *116*, 4–13. doi: 10.1016/j.seizure.2023.09.021. PMID: 37777370.

12. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. doi: 10.1038/s41586-021-03758-y. PMID: 32461654; PMCID: PMC7334197.

13. Sun, K.Y., Bai, X., Chen, S., Bao, S., Kapoor, M., Zhang, C., Backman, J., Joseph, T.,

Maxwell, E., Mitra, G., et al. (2023). A deep catalog of protein-coding variation in 985,830 individuals. bioRxivorg. doi: 10.1101/2023.05.09.539329. PMID: 37214792; PMCID: PMC10197621.

14. Goode, D.L., Cooper, G.M., Schmutz, J., Dickson, M., Gonzales, E., Tsai, M., Karra, K., Davydov, E., Batzoglou, S., Myers, R.M., et al. (2010). Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Res *20*, 301–310. doi: 10.1101/gr.102210.109. PMID: 20067941; PMCID: PMC2840986.

15. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol *6*, e1001025. doi: 10.1371/journal.pcbi.1001025. PMID: 21152010; PMCID: PMC2996323.

16. Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J., and Nickerson, D.A. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods *7*, 250–251. doi: 10.1038/nmeth0410-250. PMID: 20354513; PMCID: PMC3145250.

17. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. doi: 10.1126/science.adg7492. PMID: 37733863.

18. All of Us Research Program Genomics Investigators (2024). Genomic data in the All of Us Research Program. Nature *627*, 340–346. doi: 10.1038/s41586-023-06957-x. PMID: 38374255; PMCID: PMC10937371.

19. Lauer, L., and Rivas, M.A. (2025). Unified meta regression models for rare variant association studies (bioRxiv). doi: 10.1101/2025.01.23.634522

20. Rivas, M.A., Avila, B.E., Koskela, J., Huang, H., Stevens, C., Pirinen, M., Haritunians, T., Neale, B.M., Kurki, M., Ganna, A., et al. (2018). Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. PLoS Genet *14*, e1007329. doi: 10.1371/journal.pgen.1008190. PMID: 29795570; PMCID: PMC5967709.

21. Jenkins, D.G., and Quintana-Ascencio, P.F. (2020). A solution to minimum sample size for regressions. PLoS One *15*, e0229345. doi: 10.1371/journal.pone.0229345. PMID: 32084211; PMCID: PMC7034864.

22. Kim, J.-M., Kim, J.S., Ki, C.-S., and Jeon, B.-S. (2006). Episodic Ataxia Type 2 due to a Deletion Mutation in the CACNA1A Gene in a Korean Family. J. Clin. Neurol. *2*, 268–271. doi: 10.3988/jcn.2006.2.4.268. PMID: 20396531; PMCID: PMC2854978.

23. Epi4K consortium, and Epilepsy Phenome/Genome Project (2017). Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. Lancet Neurol. *16*, 135–143. doi: 10.1016/S1474-4422(16)30359-3. PMID: 28102150.

24. Poggio, E., Barazzuol, L., Salmaso, A., Milani, C., Deligiannopoulou, A., Cazorla, Á.G., Jang, S.S., Juliá-Palacios, N., Keren, B., Kopajtich, R., et al. (2023). ATP2B2 de novo variants as a cause of variable neurodevelopmental disorders that feature dystonia, ataxia, intellectual disability, behavioral symptoms, and seizures. Genet. Med. *25*, 100971. doi:

10.1016/j.gim.2023.100971. PMID: 37675773.

25. Taheri, M., Pourtavakoli, A., Eslami, S., Ghafouri-Fard, S., and Sayad, A. (2023). Assessment of expression of calcium signaling related lncRNAs in epilepsy. Sci. Rep. *13*, 17993. doi: 10.1038/s41598-023-45341-7. PMID: 37865723; PMCID: PMC10590428.

26. Perez-Sisques, L., Bhatt, S., Matuleviciute, R., Gileadi, T., Kramar, E., Graham, A., Garcia, F.G., Keiser, A., Matheos, D.P., Cain, J.A., et al. (2024). The intellectual disability risk gene regulates long term memory consolidation in the hippocampus. J. Neurosci. doi: 10.1523/JNEUROSCI.1544-23.2024. PMID: 38575342; PMCID: PMC11079963.

27. Mangano, G.D., Antona, V., Calì, E., Fontana, A., Salpietro, V., Houlden, H., Veggiotti, P., and Nardello, R. (2022). A complex epileptic and dysmorphic phenotype associated with a novel frameshift KDM5B variant and deletion of SCN gene cluster. Seizure *97*, 20–22. doi: 10.1016/j.seizure.2022.03.001. PMID: 35278764.

28. Han, J.Y., and Park, J. (2021). Variable Phenotypes of Epilepsy, Intellectual Disability, and Schizophrenia Caused by 12p13.33-p13.32 Terminal Microdeletion in a Korean Family: A Case Report and Literature Review. Genes *12*,. doi: 10.3390/genes12071001. PMID: 34210021; PMCID: PMC8303811.

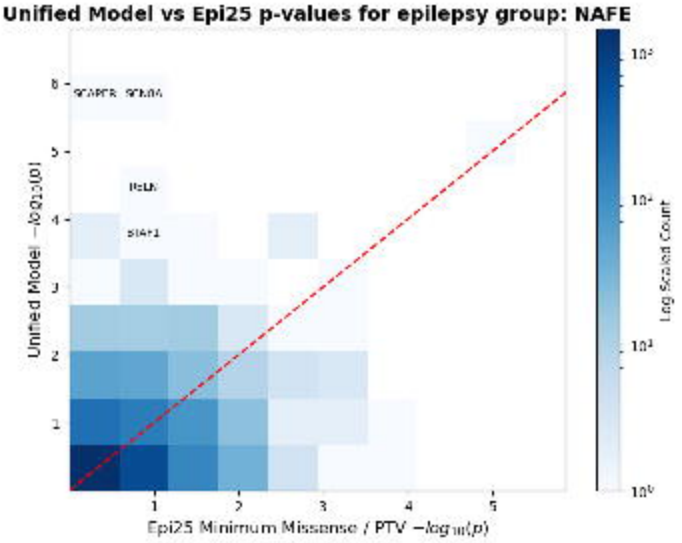
29. Ma, M.-G., Liu, X.-R., Wu, Y., Wang, J., Li, B.-M., Shi, Y.-W., Su, T., Li, B., Liu, D.-T., Yi, Y.-H., et al. (2021). Mutations Are Associated With Benign Epilepsy of Childhood With Centrotemporal Spikes With or Without Arrhythmia. Front. Neurosci. *15*, 629610. doi: 10.3389/fnins.2021.629610. PMID: 33897349; PMCID: PMC8058200.

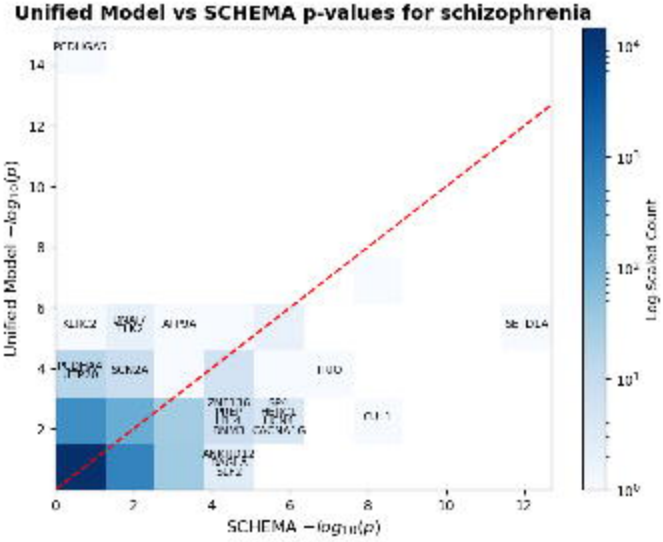
30. Hu, J., Gao, X., Chen, L., Zhou, T., Du, Z., Jiang, J., Wei, L., and Zhang, Z. (2022). A novel mutation in ryanodine receptor 2 () genes at c.12670G>T associated with focal epilepsy in a 3-year-old child. Front Pediatr *10*, 1022268. doi: 10.3389/fped.2022.1022268. PMID: 36340715; PMCID: PMC9627620.

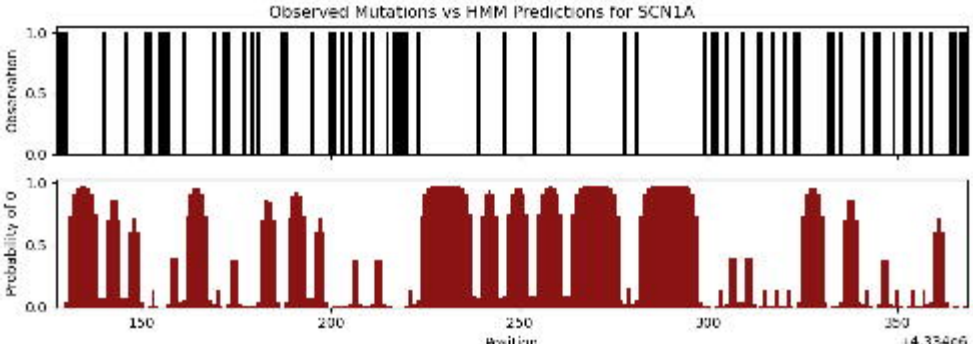
31. Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature *604*, 509–516. doi: 10.1038/s41586-022-04556-w. PMID: 35396579; PMCID: PMC9805802.

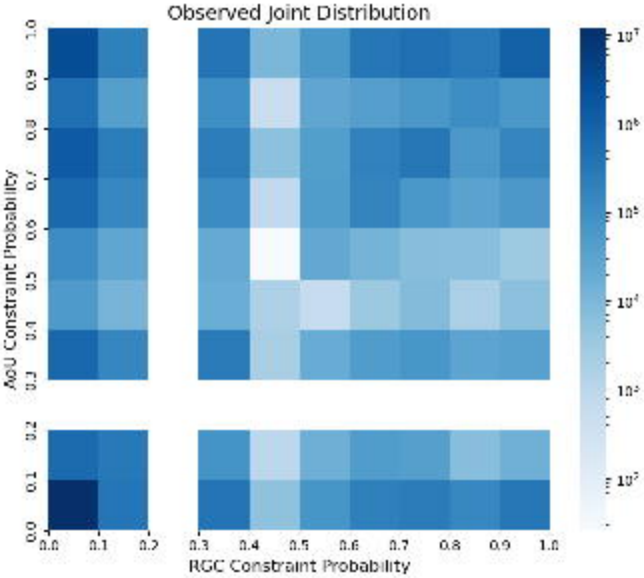
32. Ben-Shalom, R., Keeshen, C.M., Berrios, K.N., An, J.Y., Sanders, S.J., and Bender, K.J. (2017). Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. Biol. Psychiatry *82*, 224–232. doi: 10.1016/j.biopsych.2017.01.009. PMID: 28256214; PMCID: PMC5796785.

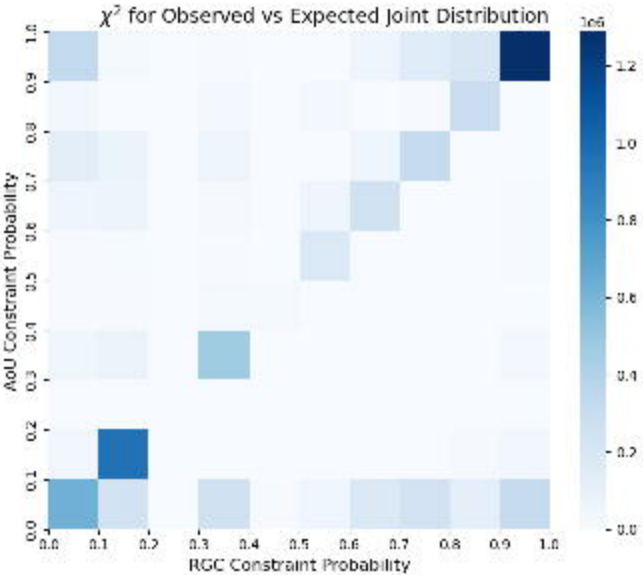
33. Miceli, F., Soldovieri, M.V., Ambrosino, P., Barrese, V., Migliore, M., Cilio, M.R., and Taglialatela, M. (2013). Genotype-phenotype correlations in neonatal epilepsies caused by mutations in the voltage sensor of K(v)7.2 potassium channel subunits. Proc. Natl. Acad. Sci. U. S. A. *110*, 4386–4391. doi: 10.1073/pnas.1216867110.PMID: 23440208; PMCID: PMC3600471.

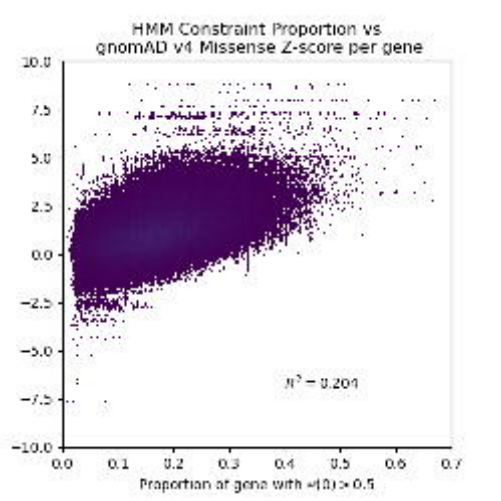

Unified Model vs Epi25 p-values for epilepsy group: NAFE
Unified Model $-\log_{10}(p)$
Epi25 Minimum Missense / PTV $-\log_{10}(p)$
Log Scaled Count
SCAPER
SCN0A
RELN
STAT1
1
2
3
4
5
6

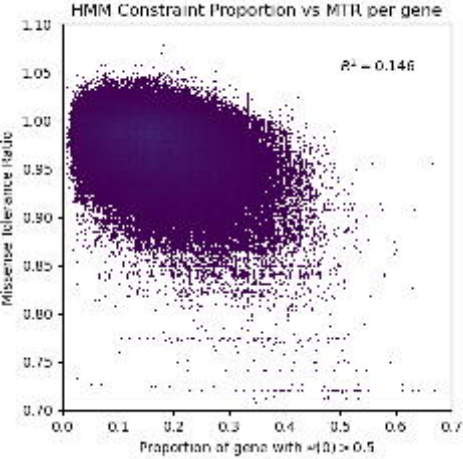

Unified Model vs SCHEMA p-values for schizophrenia
Unified Model −log10(p)
SCHEMA −log10(p)
Log Scale Count
0
2
4
6
8
10
12
14

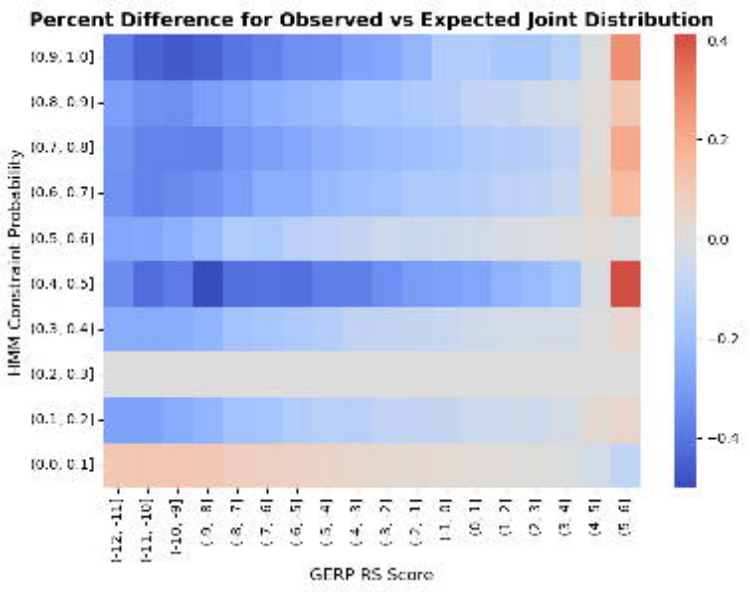

Observed Mutations vs HMM Predictions for SCN1A
Observation
1.0
0.5
0.0
Probability of O
1.0
0.5
0.0
150
200
250
300
350
Position
+4.334e6

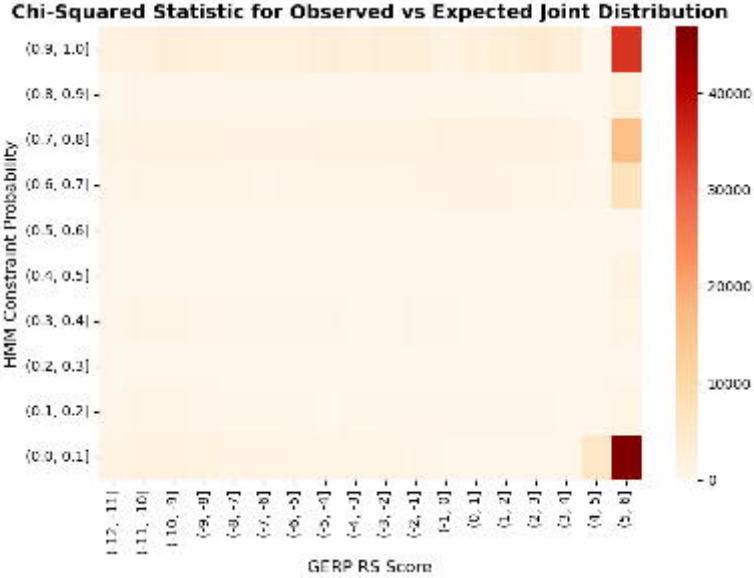

Observed Joint Distribution
AoU Constraint Probability
RGC Constraint Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
10^7
10^6
10^5
10^4
10^3
10^2

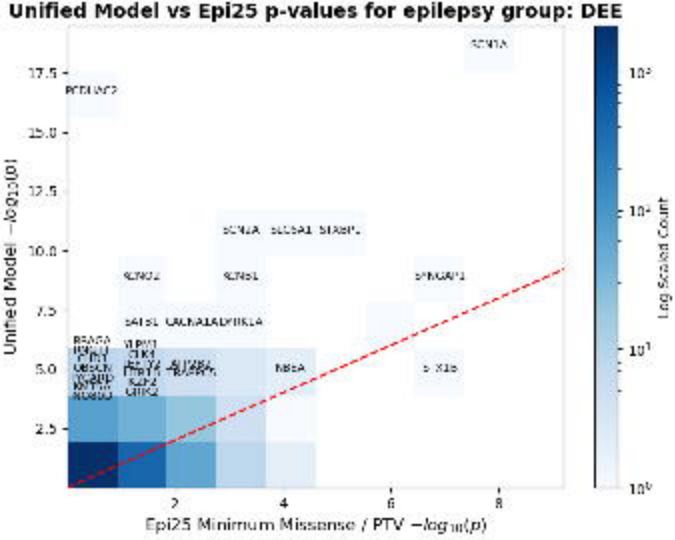

χ^2 for Observed vs Expected Joint Distribution
AoU Constraint Probability
RGC Constraint Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
10^6
1.2
1.0
0.8
0.6
0.4
0.2
0.0

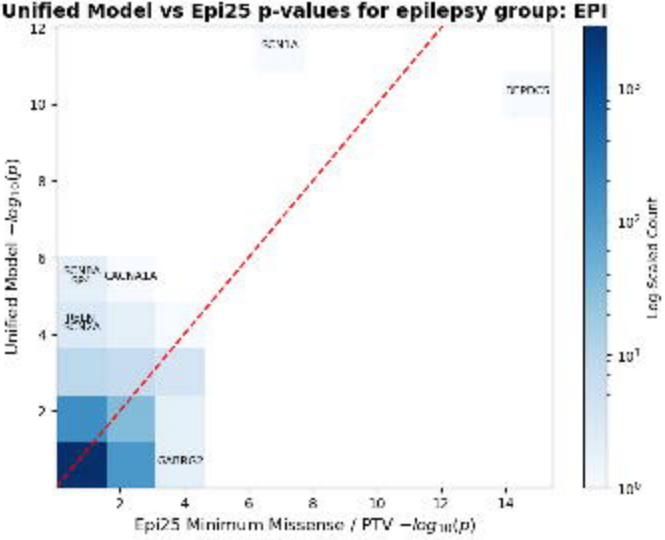

HMM Constraint Proportion vs
gnomAD v4 Missense Z-score per gene
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
$R^2 = 0.204$
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Proportion of gene with HMM > 0.5

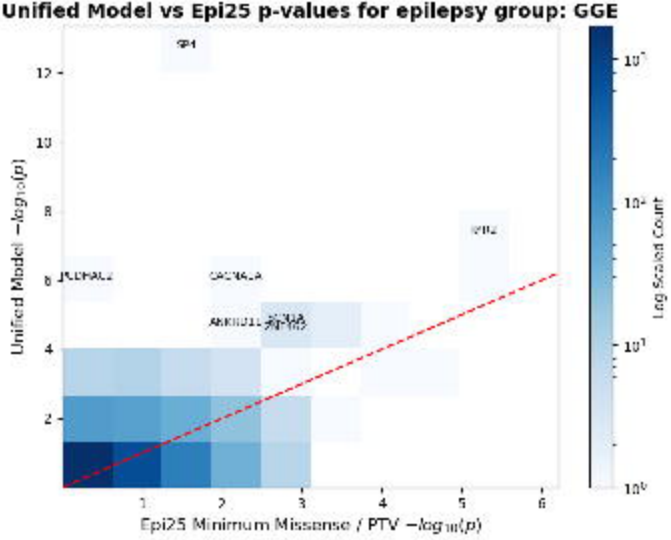

HMM Constraint Proportion vs MTR per gene
$R^2 = 0.146$
Missense Tolerance Ratio
Proportion of gene with $\nu(0) > 0.5$
1.10
1.05
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


Percent Difference for Observed vs Expected Joint Distribution
HMM Constraint Probability
[0.9, 1.0]
[0.8, 0.9]
[0.7, 0.8]
[0.6, 0.7]
[0.5, 0.6]
[0.4, 0.5]
[0.3, 0.4]
[0.2, 0.3]
[0.1, 0.2]
[0.0, 0.1]
[-12, -11]
[-11, -10]
[-10, -9]
(-9, -8]
(-8, -7]
(-7, -6]
(-6, -5]
(-5, -4]
(-4, -3]
(-3, -2]
(-2, -1]
(-1, 0]
(0, 1]
(1, 2]
(2, 3]
(3, 4]
(4, 5]
(5, 6]
GERP RS Score
0.4
0.2
0.0
−0.2
−0.4


Chi-Squared Statistic for Observed vs Expected Joint Distribution
HMM Constraint Probability
(0.9, 1.0]
(0.8, 0.9]
(0.7, 0.8]
(0.6, 0.7]
(0.5, 0.6]
(0.4, 0.5]
(0.3, 0.4]
(0.2, 0.3]
(0.1, 0.2]
(0.0, 0.1]
(-12, -11]
(-11, -10]
(-10, -9]
(-9, -8]
(-8, -7]
(-7, -6]
(-6, -5]
(-5, -4]
(-4, -3]
(-3, -2]
(-2, -1]
(-1, 0]
(0, 1]
(1, 2]
(2, 3]
(3, 4]
(4, 5]
(5, 6]
GERP RS Score
40000
30000
20000
10000
0


Unified Model vs Epi25 p-values for epilepsy group: DEE
Unified Model $-\log_{10}(p)$
Epi25 Minimum Missense / PTV $-\log_{10}(p)$
Log Scaled Count
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2
4
6
8
10^2
10^1
10^0


Unified Model vs Epi25 p-values for epilepsy group: EPI
Unified Model −log10(p)
Epi25 Minimum Missense / PTV −log10(p)
Log Scale Count
2
4
6
8
10
12
14


Unified Model vs Epi25 p-values for epilepsy group: GGE
Unified Model $-\log_{10}(p)$
Epi25 Minimum Missense / PTV $-\log_{10}(p)$
Log Scaled Count
SP4
CACNA1A
ANKRD11
SCN1A
SCN2A
SCN3A
SCN8A
12
10
8
6
4
2
1
2
3
4
5
6
10^1
10^2
10^3
10^4