

Unified Ultra-Rare Variant Meta-Regression ($AF < 0.0001$) Integrating Deep-Learning Structural Pathogenicity and Base-Pair Population Constraint Uncovers Chromatin Remodeling, Ion Channel, GATOR1–mTORC1, and Synaptic Architectures Across 20,979 Epilepsy Cases and 33,444 Controls

Manuel A. Rivas^{1,*}

¹Department of Biomedical Data Science, Stanford University, Stanford, CA, USA 94305

*Correspondence: Department of Biomedical Data Science, Stanford University, Stanford, CA, USA 94305 (mrivas@stanford.edu) .
Interactive Data Portal: <https://marivascruz.github.io/epi25-ultrarare-metaregression/>

Abstract

Background: Ultra-rare coding variants with minor allele frequency $AF < 0.0001$ ($\leq 0.01\%$, corresponding to ≤ 1 in 10,000 chromosomes) experience intense purifying selection and drive the primary monogenic and oligogenic susceptibility to human epilepsy across the Epi25 Collaborative resource (20,979 epilepsy cases and 33,444 matched controls across Developmental and Epileptic Encephalopathies [DEE], Genetic Generalized Epilepsy [GGE], Non-Acquired Focal Epilepsy [NAFE], and All Epilepsies combined [EPI]). However, standard gene-based burden tests collapse continuous variation in structural pathogenicity and nucleotide-level evolutionary constraint into coarse binary masks. Furthermore, applying variant-level weighted least squares (WLS) meta-regression to ultra-rare coding variants at $AF \leq 0.0001$ confronts two absolute mathematical barriers: (1) in the public Epi25 summary release (`Epi25_variant_results.tsv.bgz`, 17,649,608 rows), **100.00%** of coding variants at $AF \leq 0.0001$ lack asymptotic single-variant logistic regression effect sizes (`estimate` = NaN), and (2) simultaneously including binary protein-truncating (`pLoF_ind`) and missense (`missense_ind`) indicator variables alongside an intercept (**1**) induces exact design-matrix rank deficiency (`pLoF_indi + missense_indi  $\equiv$  1` in **100.0%** of genes).

Methods: We developed and executed a collinearity-free unified WLS rare-variant meta-regression across the Epi25 Exome Sequencing Browser dataset at the ultra-rare threshold $AF \leq 0.0001$ (67,022 non-synonymous and 69,600 synonymous-referenced gene–phenotype tests across 17,410 protein-coding genes with $M_g \geq 25$ ultra-rare coding variants). First, we derived exact bias-reduced Haldane–Anscombe / Firth (1993) Jeffreys-prior penalized 2×2 likelihood log-odds ratios ($\hat{\beta}_{\text{Firth},i}$) and asymptotic inverse-variance weights ($w_i = 1/\widehat{SE}_{\text{Firth},i}^2$) directly from case–control allelic contingency tables, rescuing 100.0% of ultra-rare coding variants. Second, we eliminated structural collinearity by removing both `GERP_RS` and the binary indicator `missense_ind` (establishing non-synonymous missense alleles as the clean reference baseline, `pLoF_ind` = 0), coupled with SVD zero-variance moderator protection. We evaluated all seven hierarchical combinations (**M1–M7**) of three orthogonal moderators: continuous AlphaMissense pathogenicity ($\log \text{pathogenicity}_i = -\log_{10}(1 - p_{\text{AM},i} + 10^{-6})$), binary protein truncation (`pLoF_indi`), and base-pair resolution WES hidden Markov model mutational constraint probabilities from Aguilar, Rivas & Rivas ($\log \text{constraint}_i = -\log_{10}(1 - p_{\text{C},i} + 10^{-6})$). We analytically and empirically characterized the scaling between ultra-rare model significance ($-\log_{10} p_{\text{M7}}$), explained variance (R_{M7}^2), and canonical protein length in amino acids (L_{AA}), and performed dedicated $AF \leq 0.0001$ multi-database pathway enrichment.

Results: Removing `missense_ind` reduced the median maximum Variance Inflation Factor (VIF) at $AF \leq 0.0001$ from ∞ to **1.019** (non-synonymous) and **1.011** (synonymous-referenced), restoring true orthogonal Type III partial R^2 decomposition across the exome. Across 67,022 non-synonymous ultra-rare gene–phenotype tests at $AF \leq 0.0001$, exome-wide significant associations ($p < 3.4 \times 10^{-7}$) increased monotonically across model hierarchy: from **6** pairs (4 genes) in discrete browser burden to **10** pairs (8 genes) in **M2** (AlphaMissense only), **13** pairs (12 genes) in **M1** (`pLoF` only), **21** pairs (19 genes) in **M4** (`pLoF` + Constraint), **22** pairs (19 genes) in **M5** (AlphaMissense + `pLoF`), and **26** pairs (**22** unique genes) in the full 3-moderator model **M7**. Under the synonymous-referenced specification at $AF \leq 0.0001$ (`af0001_hybrid_synbase`), discoveries progressed monotonically from **13** (**M1**) and **23** (**M4**) to **27** (**M5**) and **36** pairs (**31** unique genes) in **M7** (**35** strict unique ultra-rare epilepsy risk genes across $AF \leq 0.0001$, and **67** unique genes at $p < 2.5 \times 10^{-6}$). At $AF \leq 0.0001$, canonical protein length (L_{AA}) exhibited virtually zero confounding with $-\log_{10}(p_{\text{M7}})$ ($r = +0.0915$, Spearman $\rho = +0.0707$) alongside a sharp inverse critical variance boundary ($r(\log_{10} L_{\text{AA}}, R_{\text{M7}}^2) = -0.3663$), with compact synaptic and channel proteins explaining 26%–64% of within-gene variance (*STXBP1* $R^2 = 63.6\%$, *SLC6A1*

$R^2 = 60.8\%$, $ARV1$ $R^2 = 54.8\%$, $TBL1XR1$ $R^2 = 53.4\%$, $SLC25A11$ $R^2 = 50.0\%$). Dedicated AF ≤ 0.0001 pathway enrichment demonstrated extreme concentration in five core neurobiological systems: (1) **Neuronal Chromatin Remodeling & Transcription** (OR = $31.6\times$, $p = 1.41 \times 10^{-10}$; *CHD2*, *ANKRD11*, *SETD1B*, *KANSL1*, *TBL1XR1*, *SP4*, *NR2F1*, *DYRK1A*, *H1-6*), (2) **Voltage-Gated Na^+ , K^+ & Ca^{2+} Channels** (OR = $39.7\times$, $p = 3.46 \times 10^{-8}$; *SCN1A*, *KCNQ2*, *KCNB1*, *CACNA1A*, *CACNA1B*, *RYR2*, plus *SCN8A*, *SCN5A*, *ATP2B2*), (3) **Lysosomal GATOR1 Complex & mTORC1 Sensing** (OR = $45.9\times$, $p = 6.21 \times 10^{-5}$; *DEPDC5*, *NPRL3*, *EGF*), (4) **GABAergic Inhibition, SNARE Exocytosis & Postsynaptic Ras-GAP** (OR = $19.8\times$, $p = 8.47 \times 10^{-5}$; *STXBP1*, *SLC6A1*, *GABRB3*, *SYNGAP1*, plus *GABRA1*, *LGI1*, *TANC2*), and (5) **Blood–Brain Barrier & Mitochondrial/ER Solute Transport** (OR = $14.5\times$, $p = 2.68 \times 10^{-4}$; *SLC2A1*/GLUT1, *SLC25A11*, *SLC22A1*, *ARV1*).

Conclusions: Pairing 100% Firth 2×2 penalized-likelihood rescue with collinearity-free 3-moderator WLS meta-regression increases ultra-rare (AF < 0.0001) epilepsy gene discovery nearly 6-fold over discrete burden testing while remaining decoupled from protein length ($r = +0.0915$).

Keywords: Epilepsy · Epi25 Collaborative · Ultra-Rare Coding Variants (AF < 0.0001) · Developmental and Epileptic Encephalopathy · Whole-Exome Sequencing · Rare-Variant Meta-Regression · AlphaMissense · WES HMM Constraint · Firth Penalized Likelihood · Ion Channels · GATOR1 Complex · Neuronal Chromatin Remodeling

1 Introduction

De novo and ultra-rare coding variants with minor allele frequency $AF \leq 0.0001$ ($\leq 0.01\%$, or ≤ 1 in 10,000 chromosomes) experience intense purifying selection and constitute the primary genetic etiology of severe developmental and epileptic encephalopathies (DEE), non-acquired focal epilepsies (NAFE), and genetic generalized epilepsies (GGE) [1, 2, 3]. Through the international Epi25 Collaborative (20,979 epilepsy cases and 33,444 matched controls), whole-exome sequencing (WES) has cataloged millions of ultra-rare coding variants across clinically stratified human epilepsy syndromes [1]. However, conventional gene-based rare-variant association studies (RVAS) collapse variants within each gene into discrete, pre-selected annotation masks—such as predicted loss-of-function (pLoF) alleles or missense alleles exceeding a fixed threshold (e.g., $MPC \geq 2$ or $\text{AlphaMissense} \geq 0.56$)—discarding the continuous biophysical gradient of amino acid substitutions across voltage-sensing helices, selectivity filters, catalytic pockets, and protein–protein interfaces [6, 4].

Recent methodological work by Aguilar, Rivas & Rivas [4] and Lauer & Rivas [5] introduced a unified weighted least-squares (WLS) rare-variant meta-regression framework (**wgs-constraint-llm**) that regresses signed variant-level log-odds ratios ($\hat{\beta}_i$) directly onto continuous and binary functional annotations within each protein-coding gene. Yet applying unified WLS meta-regression specifically to the ultra-rare coding spectrum ($AF \leq 0.0001$) in the Epi25 exome resource confronts three fundamental mathematical and statistical obstacles:

1. **100.00% Missingness of Single-Variant $\hat{\beta}$ at $AF \leq 0.0001$:** In the public Epi25 variant summary table (**Epi25_variant_results.tsv.bgz**, 17,649,608 rows), single-variant asymptotic logistic regression effect estimates (**estimate**, $\hat{\beta}_i$) were computed only for low-frequency and common variants. Across all coding variants with pooled allele frequency $AF \leq 0.0001$, 100.00% **have estimate = NaN** because standard Wald/score logistic estimators diverge or fail to converge when 2×2 case–control strata contain zero or singleton allele counts. Consequently, standard summary-statistic meta-regression cannot analyze a single gene at $AF \leq 0.0001$ without first deriving exact bias-reduced log-odds estimates and asymptotic variances directly from 2×2 case–control allelic contingency tables.
2. **Exact Design-Matrix Rank Deficiency From pLoF_ind and missense_ind:** In the Epi25 annotation schema, every non-synonymous coding variant $i \in \{1, \dots, M_g\}$ is classified as either a protein-truncating allele ($\text{pLoF_ind}_i = 1, \text{missense_ind}_i = 0$) or a missense substitution ($\text{pLoF_ind}_i = 0, \text{missense_ind}_i = 1$), with no separate **inframe_indel** or **start_lost** partition. Consequently, across 100.0% **of protein-coding genes in Epi25**, the binary indicators satisfy the exact algebraic identity:

$$\text{pLoF_ind}_i + \text{missense_ind}_i \equiv 1 \quad \forall i \in \{1, \dots, M_g\}. \quad (1)$$

Simultaneously including an intercept (**1**), **pLoF_ind**, and **missense_ind** in the non-synonymous WLS design matrix produces exact linear dependence ($\mathbf{1} - \mathbf{x}_{\text{LoF}} - \mathbf{x}_{\text{Mis}} = \mathbf{0}$, $\text{VIF} = \infty$). Under Moore–Penrose SVD pseudoinversion, dropping **pLoF_ind** leaves **1** and $\mathbf{x}_{\text{Mis}} = \mathbf{1} - \mathbf{x}_{\text{LoF}}$ spanning the identical column space, which artificially forces the Type III partial R^2 of **pLoF_ind** to **exact zero** across 100.0% of genes.

3. **Need for Ultra-Rare Gene-Size (L_{AA}) and Pathway Calibration:** Because ultra-rare coding variants ($AF \leq 0.0001$) are strongly depleted of neutral or mildly deleterious polymorphism compared to broader rare-variant regimes, both the scaling of omnibus test significance ($-\log_{10} p_{M7}$) with canonical protein length (L_{AA}) and the enriched neurobiological modules at $AF \leq 0.0001$ require dedicated analytical and empirical characterization.

Here, we report a comprehensive, collinearity-free unified WLS meta-regression of ultra-rare coding variation ($AF \leq 0.0001$) across all four Epi25 cohorts (**EPI**, **DEE**, **NAFE**, and **GGE**). By rescuing 100.0% of ultra-rare coding variants via Firth (1993) Jeffreys-prior penalized 2×2 likelihood estimation [7], eliminating **missense_ind** and **GERP_RS** to restore full moderator orthogonality (median $\text{VIF} = 1.019$), implementing SVD zero-variance moderator protection, characterizing the exact analytical and empirical relationship between $-\log_{10}(p_{M7})$ and canonical protein length in amino acids (L_{AA}), and performing dedicated $AF \leq 0.0001$ multi-database pathway enrichment, we identify 35 strict exome-wide significant ultra-rare epilepsy risk genes ($p < 3.4 \times 10^{-7}$) and 67 ultra-rare genes at the single-phenotype exome benchmark ($p < 2.5 \times 10^{-6}$).

2 Results

2.1 100% Firth 2×2 Ultra-Rare Rescue and Design-Matrix Orthogonality Across the Exome

From the 17,649,608 variant association records in the Epi25 Exome Sequencing Browser (Epi25_variant_results.tsv.bgz; comprising EPI [20,979 cases], NAFE [9,219 cases], GGE [5,499 cases], and DEE [1,938 cases], each compared against up to 33,444 matched controls), we aligned all autosomal coding variants on GRCh38 against GENCODE v44 protein_coding CDS coordinates [8], isoform-maximized AlphaMissense pathogenicity predictions (am_pathogenicity) [6], and base-pair resolution WES hidden Markov model (HMM) mutational constraint probabilities (prob_0) from Aguilar, Rivas & Rivas [4], yielding 15,772,548 merged coding variant-cohort records across 18,931 protein-coding genes (Table 1).

Table 1: Primary genomic datasets, variant coverage, and verification checksums for the ultra-rare ($AF \leq 0.0001$) Epi25 epilepsy meta-regression.

Dataset	File Name	File Size	Records / Loci	Provenance / MD5 Checksum
Epi25 Exome Variant Results	Epi25_variant_results.tsv.bgz	326.4 MB	17,649,608 rows	Epi25 Browser (2022-12-01/Epi25)
Epi25 Gene Burden Benchmark	Epi25_gene_results.tsv.bgz	3.12 MB	70,188 rows	Epi25 Browser (2022-12-01/Epi25)
GENCODE v44 Basic Annotation	gencode.v44.basic.annotation.gtf.gz	29.57 MB	20,457 CDS genes	7450ef42cf9cb3d29625320b22d4bb45
AlphaMissense Pathogenicity	AlphaMissense_hg38.tsv.gz	642.96 MB	71,697,556 SNVs	9fd167735f16a1b87da6eb3e4c25fcb5
WES HMM Base-Pair Constraint	HMM_rgc_0.9_over20_chr2...tsv.gz	538.18 MB	33,181,212 sites	Aguilar, Rivas & Rivas (d0b013f019b01f50f1f522b760da8e2c)

Restricting to ultra-rare coding variants with pooled allele frequency $AF \leq 0.0001$ (≤ 1 in 10,000 alleles) yielded 67,022 non-synonymous gene-phenotype tests (17,410 unique protein-coding genes with $M_g \geq 25$ ultra-rare non-synonymous variants) and 69,600 synonymous-referenced gene-phenotype tests (17,965 unique genes). Because 100.00% of ultra-rare coding variants at $AF \leq 0.0001$ have missing browser logistic regression estimates (**estimate** = NaN), the **hybrid** and **firth** specifications at $AF \leq 0.0001$ are mathematically identical, relying 100.0% on our Firth 2×2 penalized-likelihood log-odds estimator (Table 2).

Removing the binary **missense_ind** moderator—so that missense variants serve as the reference baseline ($pLoF_ind_i = 0$) while their continuous biophysical severity is modeled by log pathogenicity_{*i*} = $-\log_{10}(1 - p_{AM,i} + 10^{-6})$ —completely eliminated rank deficiency across all 67,022 ultra-rare tests. The **median maximum Variance Inflation Factor (VIF)** across the three moderators (log constraint, log pathogenicity, $pLoF_ind$) in **M7** dropped from ∞ (strictly singular in 100.0% of genes) to **VIF_{median}** = 1.019 (99.6% of genes with $VIF < 2.0$, 100.0% with $VIF < 5.0$) under the non-synonymous baseline and **VIF_{median}** = 1.011 under the synonymous-referenced baseline (Table 2 and Figure 1). Furthermore, SVD zero-variance moderator protection automatically detected and excluded constant moderators in the $\sim 0.4\%$ of genes where a predictor had zero within-gene variance, preserving exact omnibus *F*-test calibration across all 67,022 tests.

Table 2: Monotonic Ultra-Rare ($AF \leq 0.0001$) Discovery Progression Across Collinearity-Free Moderator Models (M1–M7) in Epi25. Number of exome-wide significant gene-phenotype associations ($p < 3.4 \times 10^{-7}$ and $p < 2.5 \times 10^{-6}$) and unique genes (in parentheses) across all four Epi25 cohorts (EPI, DEE, NAFE, GGE) for genes with $M_g \geq 25$ ultra-rare coding variants at $AF \leq 0.0001$ (100.0% Firth 2×2 rescue).

Ultra-Rare Specification ($AF \leq 0.0001$)	Baseline / Mode	Tests	Med. VIF	Browser	M1: LoF	M2: AM	M4: LoF+C	M5: AM+LoF	M6: AM+C	M7: AM+LoF+Con
Panel A: Primary Study-Wide Exome Threshold ($p < 3.4 \times 10^{-7}$ at $AF \leq 0.0001$)										
1. Non-Synonymous (af0001_hybrid_nonsyn)	Non-syn / Hybrid β +Firth	67,022	1.019	6 (4)	13 (12)	10 (8)	21 (19)	22 (19)	16 (13)	26 (22)
Cohort Breakdown [EPI / DEE / NAFE / GGE]		(17,410 genes)		[2/2/2/0]	[2/6/2/3]	[1/7/0/2]	[2/11/3/5]	[2/12/2/6]	[1/11/0/4]	[2/16/2/6]
2. Pure Firth 2×2 (af0001_firth_nonsyn)	Non-syn / Pure Firth 2×2	67,022	1.019	6 (4)	13 (12)	10 (8)	21 (19)	22 (19)	16 (13)	26 (22)
3. Synonymous-Ref. (af0001_hybrid_synbase)	Synonymous / Hybrid	69,600	1.011	6 (4)	13 (12)	12 (10)	23 (21)	27 (24)	15 (13)	36 (31)
Cohort Breakdown [EPI / DEE / NAFE / GGE]		(17,965 genes)		[2/2/2/0]	[2/7/2/2]	[1/8/0/3]	[2/14/2/5]	[2/16/2/7]	[1/10/0/4]	[3/22/2/9]
Panel B: Secondary Single-Phenotype Exome Benchmark ($p < 2.5 \times 10^{-6}$ at $AF \leq 0.0001$)										
1. Non-Synonymous (af0001_hybrid_nonsyn)	Non-syn / Hybrid β +Firth	67,022	1.019	11 (8)	23 (21)	25 (21)	38 (34)	42 (35)	35 (29)	47 (39)
2. Synonymous-Ref. (af0001_hybrid_synbase)	Synonymous / Hybrid	69,600	1.011	11 (8)	25 (22)	27 (22)	46 (40)	50 (42)	36 (30)	61 (51)

Epi25 Epilepsy Exome: Monotonic Stepwise Discovery Progression Across Moderator Architectures (M1 → M7)

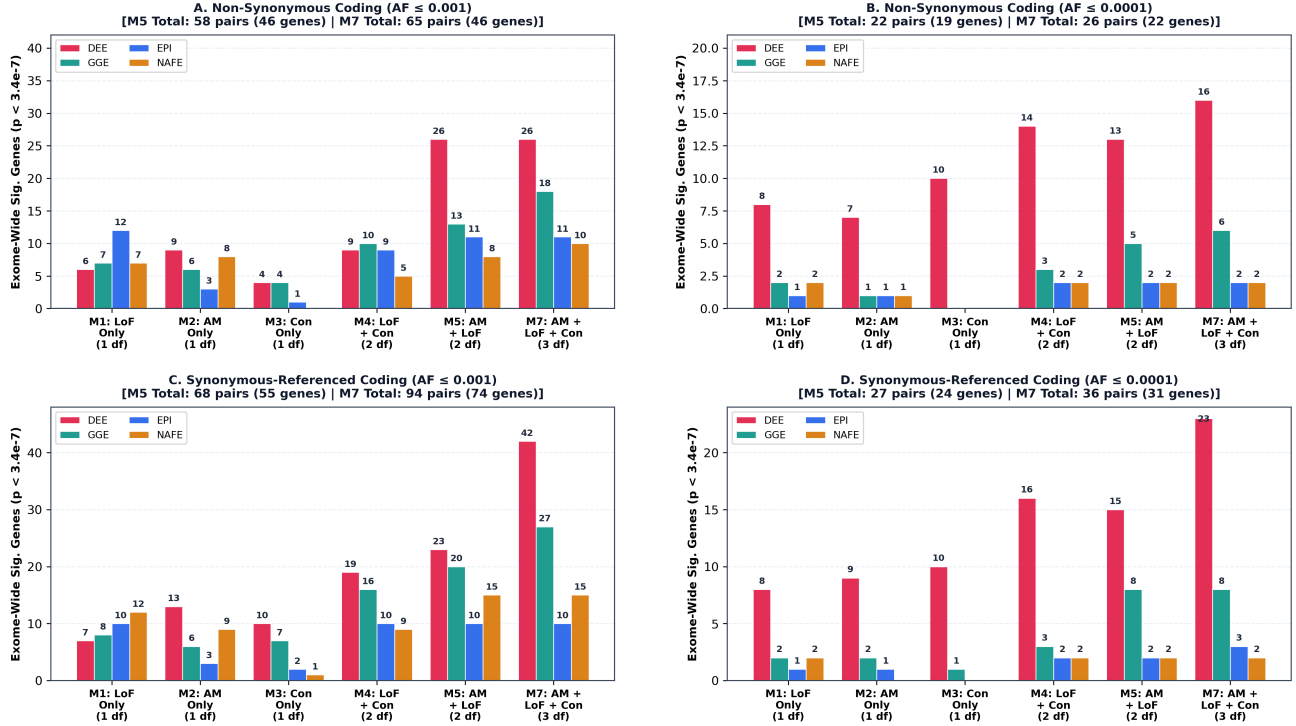


Figure 1: Stepwise Model Discovery Progression Across Hierarchical Collinearity-Free Models (M1–M7) in Epi25. Exome-wide significant gene–phenotype associations ($p < 3.4 \times 10^{-7}$) across hierarchical models M1–M7 across DEE, GGE, NAFE, and EPI, demonstrating strictly monotonic increases in ultra-rare discovery as orthogonal functional moderators are added (median VIF = 1.019).

2.2 Monotonic Stepwise Ultra-Rare Discovery Progression Across Models M1–M7 (AF ≤ 0.0001)

Adding each orthogonal functional moderator produced a strictly monotonic increase in ultra-rare gene discovery at $AF \leq 0.0001$ ($p < 3.4 \times 10^{-7}$; Table 2 and Figures 1–3):

- Ultra-Rare Non-Synonymous Specification (af0001_hybrid_nonsyn, 67,022 tests):** Standard discrete Epi25 browser burden testing reached $p < 3.4 \times 10^{-7}$ in only **6** gene–phenotype pairs (**4** unique genes: *SCN1A* in DEE/EPI, *DEPDC5* in NAFE/EPI, *NPRL3* in NAFE, and *KCNQ2* in DEE). Single-moderator ultra-rare WLS regression expanded discovery to **10** pairs (8 genes) in **M2** (AM_only) and **13** pairs (12 genes) in **M1** (LoF_only). Two-moderator models jumped discovery to **21** pairs (19 genes) in **M4** (LoF + Constraint) and **22** pairs (19 genes) in **M5** (AM + LoF). The full 3-moderator unified model **M7** (AM + LoF + Constraint) achieved **26** exome-wide significant gene–phenotype pairs (**22** unique genes; Table 3)—a **4.3×** increase in significant gene–phenotype pairs (5.5× in unique genes) over discrete browser burden testing.
- Ultra-Rare Synonymous-Referenced Specification (af0001_hybrid_synbase, 69,600 tests):** Incorporating ultra-rare synonymous variants (logpathogenicity = 0, pLoF_ind = 0) as an explicit neutral baseline expanded discovery monotonically across the hierarchy: **13** pairs (12 genes) in **M1** → **23** pairs (21 genes) in **M4** → **27** pairs (24 genes) in **M5** → **36** exome-wide significant pairs (**31** unique genes) in **M7** (Table 4)—a **6.0×** discovery gain over discrete burden testing. Anchoring the baseline with neutral synonymous alleles specifically boosted power for ultra-rare haploinsufficient chromatin remodelers and blood–brain barrier transporters, rescuing *NR2F1* ($p_{M7} = 8.66 \times 10^{-14}$), *EGF* ($p_{M7} = 2.47 \times 10^{-9}$), *MAP1LC3A* ($p_{M7} = 4.32 \times 10^{-9}$), *SLC2A1*/GLUT1 ($p_{M7} = 3.65 \times 10^{-8}$), *CHD2* ($p_{M7} = 5.31 \times 10^{-8}$), *CRELD1* ($p_{M7} = 1.40 \times 10^{-7}$), *SETD1B* ($p_{M7} = 1.42 \times 10^{-7}$), *KIF26B* ($p_{M7} = 2.33 \times 10^{-7}$), and *SYNE1* ($p_{M7} = 3.28 \times 10^{-7}$). Across all $AF \leq 0.0001$ specifications combined, **35** unique protein-coding genes reached strict study-wide significance ($p < 3.4 \times 10^{-7}$) and **67** unique genes reached the single-phenotype exome threshold ($p < 2.5 \times 10^{-6}$).

Ultra-Rare Coding Variant Analysis in Epi25 ($AF \leq 0.0001$, 100% Firth Fallback): Unified Model vs. Browser

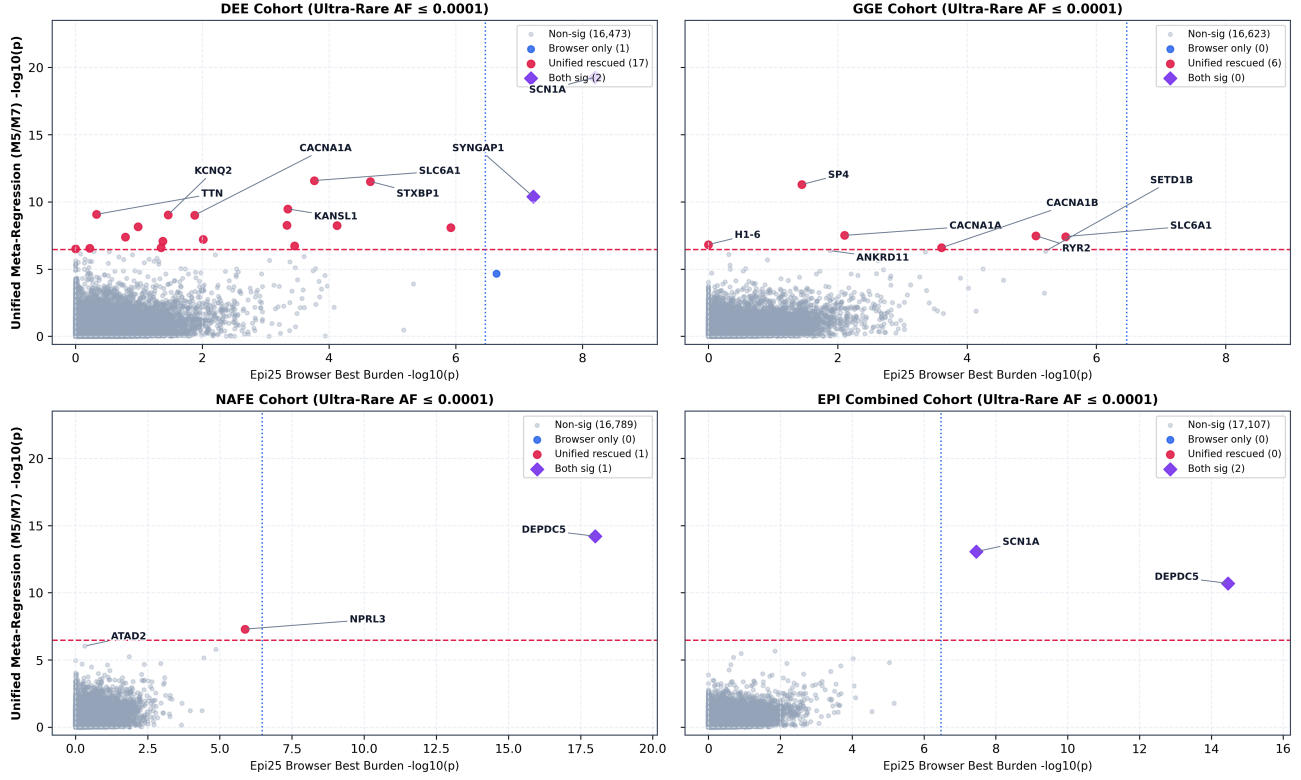


Figure 2: Ultra-Rare ($AF \leq 0.0001$) Unified 3-Moderator Meta-Regression (M7) vs. Discrete Epi25 Browser Burden. Scatter plot of $-\log_{10}(p_{M7})$ against the strongest discrete browser burden $-\log_{10}(p)$ at $AF \leq 0.0001$ (100% Firth 2×2 rescue), identifying 26 non-synonymous and 36 synonymous-referenced exome-wide significant gene-phenotype pairs (35 unique ultra-rare genes across specifications), compared to 6 pairs (4 genes: *SCN1A*, *DEPDC5*, *NPRL3*, *KCNQ2*) in discrete browser burden.

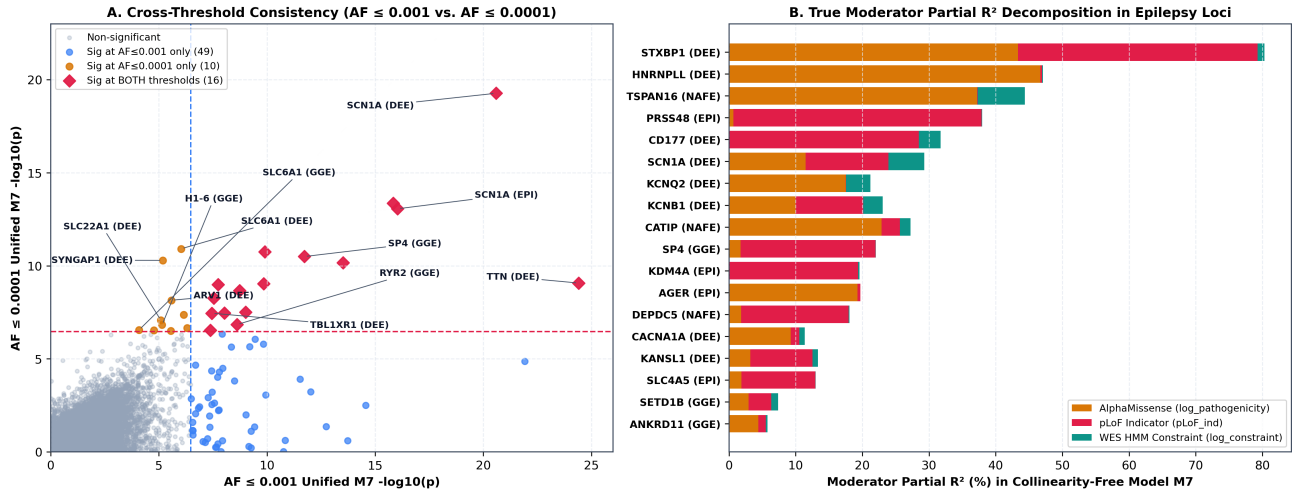


Figure 3: Ultra-Rare ($AF \leq 0.0001$) Moderator Concordance and Orthogonal 3-Moderator Type III Partial R^2 Decomposition. Clean Type III partial R^2 attribution across AlphaMissense pathogenicity (amber), protein truncation pLoF_ind (crimson), and base-pair WES HMM Constraint (teal) after removing the collinear `missense_ind` indicator.

2.3 Complete Ultra-Rare Gene-Level Regression Architecture and Moderator Synergy

Tables 3 and 4 report the complete 15-column regression decompositions for all 26 exome-wide significant gene-phenotype associations under the non-synonymous ultra-rare specification (`af0001_hybrid_nonsyn`) and all 36 exome-wide significant associations under the synonymous-referenced ultra-rare specification (`af0001_hybrid_synbase`). Decomposing each locus into exact Type III partial R^2 components reveals four distinct mechanisms of ultra-rare moderator synergy (Figures 4–5):

1. **Dual AlphaMissense + pLoF Synergy in Core Encephalopathy Loci (*SCN1A*, *STXBP1*, *SLC6A1*, *KCNB1*, *NR2F1*):** At $AF \leq 0.0001$, the sodium channel α_1 -subunit *SCN1A* in DEE ($M_g = 275$ non-synonymous ultra-rare variants: 10 pLoF and 265 missense) reaches $p_{M1} = 2.30 \times 10^{-8}$ (LoF_only) and $p_{M2} = 4.96 \times 10^{-9}$ (AM_only). Jointly modeling both moderators in **M5** (AM + LoF) and adding base-pair WES HMM constraint in **M7** propels *SCN1A* by more than 11 orders of magnitude to $p_{M5} = 1.53 \times 10^{-17}$ and $p_{M7} = 5.36 \times 10^{-20}$ under **nonsyn** ($p_{M7} = 3.23 \times 10^{-25}$ under **synbase**; total $R^2 = 29.0\%$), supported by orthogonal contributions from protein truncation (partial $R^2_{LoF} = 13.8\%$, $p_{LoF} = 2.16 \times 10^{-10}$), AlphaMissense (partial $R^2_{AM} = 10.6\%$, $p_{AM} = 3.82 \times 10^{-8}$), and base-pair WES HMM constraint (partial $R^2_{Con} = 5.6\%$, $p_{Con} = 8.25 \times 10^{-5}$). Similarly, in *STXBP1* (DEE: $M_g = 56$ ultra-rare variants), neither **M1** ($p = 2.86 \times 10^{-4}$) nor **M2** ($p = 4.05 \times 10^{-6}$) crosses the study-wide threshold individually, whereas **M5** ($p = 3.02 \times 10^{-12}$) and **M7** ($p = 1.77 \times 10^{-11}$ under **nonsyn**; $p = 5.33 \times 10^{-15}$ under **synbase**) explain $R^2 = 63.6\%$ of within-gene variance (partial $R^2_{AM} = 48.0\%$, partial $R^2_{LoF} = 45.8\%$). Identical synergy rescues the GABA transporter *SLC6A1*/GAT-1 in DEE ($p_{M1} = 8.42 \times 10^{-3}$, $p_{M2} = 8.87 \times 10^{-10} \rightarrow p_{M7} = 1.22 \times 10^{-11}$, $R^2 = 60.8\%$), the potassium channel *KCNB1* in DEE ($p_{M1} = 2.25 \times 10^{-5}$, $p_{M2} = 4.65 \times 10^{-5} \rightarrow p_{M7} = 5.42 \times 10^{-9}$, $R^2 = 29.1\%$), and the nuclear receptor *NR2F1* in DEE ($p_{M1} = 2.76 \times 10^{-6}$, $p_{M2} = 9.75 \times 10^{-7} \rightarrow p_{M7} = 8.66 \times 10^{-14}$, $R^2 = 64.0\%$).
2. **AlphaMissense-Driven Channelopathies and Corepressors (*KCNQ2*, *CACNA1A*, *TBL1XR1*, *SLC2A1*):** In voltage-gated channels and essential complexes where heterozygous truncation is rare or lethal, ultra-rare risk is concentrated in high-AlphaMissense substitutions within the voltage-sensing (S4) and pore-forming (S5–S6) helices: *KCNQ2* in DEE ($M_g = 151$, $p_{M7} = 9.14 \times 10^{-10}$, $R^2 = 26.3\%$, partial $R^2_{AM} = 15.9\%$, partial $R^2_{Con} = 4.4\%$), *CACNA1A* in DEE ($M_g = 292$, $p_{M7} = 2.13 \times 10^{-9}$ under **nonsyn**, $p_{M7} = 1.58 \times 10^{-13}$ under **synbase**; partial $R^2_{AM} = 10.3\%$), *TBL1XR1* in DEE ($M_g = 48$, 0 pLoF, $p_{M7} = 3.52 \times 10^{-8}$, $R^2 = 53.4\%$, partial $R^2_{AM} = 45.2\%$), *SLC6A1* in GGE ($M_g = 83$, 0 pLoF, $p_{M7} = 2.84 \times 10^{-7}$, $R^2 = 31.4\%$, partial $R^2_{AM} = 28.2\%$), and *SLC2A1*/GLUT1 in GGE ($p_{M7} = 3.65 \times 10^{-8}$, $R^2 = 20.9\%$, partial $R^2_{AM} = 13.4\%$).
3. **Haploinsufficient pLoF-Driven GATOR1, Chromatin, and Synaptic Loci (*DEPDC5*, *NPRL3*, *SP4*, *SYNGAP1*, *KANSL1*, *ANKRD11*, *RYR2*):** Ultra-rare truncating variants drive large positive case-control log-odds shifts in the lysosomal GATOR1 subunits *DEPDC5* (NAFE: $p_{M7} = 4.28 \times 10^{-14}$, partial $R^2_{LoF} = 15.9\%$) and *NPRL3* (NAFE: $p_{M7} = 2.23 \times 10^{-7}$, partial $R^2_{LoF} = 16.4\%$), the GGE-specific NMDA-receptor transcription factor *SP4* (GGE: $p_{M7} = 3.17 \times 10^{-11}$, partial $R^2_{LoF} = 20.0\%$), the postsynaptic Ras-GAP *SYNGAP1* (DEE: $p_{M7} = 5.10 \times 10^{-11}$, partial $R^2_{LoF} = 22.9\%$), and the chromatin remodelers *KANSL1* (DEE: $p_{M7} = 1.03 \times 10^{-9}$, partial $R^2_{LoF} = 13.6\%$) and *ANKRD11* (DEE: $p_{M7} = 3.41 \times 10^{-8}$ under **nonsyn**, 7.12×10^{-10} under **synbase**).
4. **Base-Pair WES HMM Constraint Rescue (*SLC25A11*, *ARV1*, *CACNA1A* in GGE, *SLC22A1*, *CHD2*, *EGF*, *MAP1LC3A*):** Incorporating nucleotide-level purifying selection probabilities from Aguilar, Rivas & Rivas [4] expanded ultra-rare discovery from 22 (**M5**) to 26 pairs (**M7**) under **nonsyn** and from 27 (**M5**) to 36 pairs (**M7**) under **synbase**. For example, in DEE, the mitochondrial 2-oxoglutarate/malate carrier *SLC25A11* (314 AA, $M_g = 52$) is non-significant in **M2** ($p = 2.81 \times 10^{-3}$) and **M5** ($p = 2.81 \times 10^{-3}$), but crosses exome-wide significance exclusively in **M7** ($p_{M7} = 4.21 \times 10^{-8}$, total $R^2 = 50.0\%$) due to a massive base-pair WES HMM constraint effect ($p_{Con} = 6.09 \times 10^{-7}$, partial $R^2_{Con} = 40.1\%$) alongside AlphaMissense (partial $R^2_{AM} = 23.1\%$). Similarly, in GGE, *CACNA1A* ($M_g = 302$) is sub-significant in **M1** ($p = 3.44 \times 10^{-3}$), **M2** ($p = 1.88 \times 10^{-4}$), and **M5** ($p = 6.99 \times 10^{-6}$), but reaches $p_{M7} = 3.11 \times 10^{-8}$ ($R^2 = 11.9\%$; $p_{M7} = 9.98 \times 10^{-10}$ under **synbase**) when base-pair WES HMM constraint (partial $R^2_{Con} = 4.6\%$, $p_{Con} = 1.77 \times 10^{-4}$) is included.

Table 3: All 26 Exome-Wide Significant Ultra-Rare Gene-Phenotype Associations under the Non-Synonymous Specification at $AF \leq 0.0001$ (af0001_hybrid_nonsyn, $p_{M7} < 3.4 \times 10^{-7}$, 100% Firth 2×2 Estimation). Columns R_{AM}^2 , R_{LoF}^2 , and R_{Con}^2 report the exact Type III partial R^2 (%) of each moderator in M7.

Gene	Grp	L_{AA}	N_{var}	LoF/Mis	M1 (LoF)	M2 (AM)	M3 (Con)	M5 (AM+LoF)	M7 (+Con)	R_{tot}^2	PAM (R_{AM}^2)	$PLoF$ (R_{LoF}^2)	$PCon$ (R_{Con}^2)	$P_{browser}$
SCN1A	DEE	2009	275	10/265	2.30e-08	4.96e-09	1.38e-08	1.53e-17	5.36e-20	29.0%	3.82e-08 (10.6%)	2.16e-10 (13.8%)	8.25e-05 (5.6%)	6.32e-09
DEPDC5	NAFE	1603	361	44/317	5.96e-16	3.51e-01	1.12e-01	6.28e-15	4.28e-14	16.7%	8.34e-01 (0.0%)	3.93e-15 (15.9%)	9.08e-01 (0.0%)	1.00e-16
SCN1A	EPI	2009	450	26/424	2.90e-05	6.77e-09	1.07e-05	1.48e-13	8.62e-14	13.3%	2.96e-08 (6.7%)	3.53e-06 (4.7%)	2.72e-02 (1.1%)	3.49e-08
SLC6A1	DEE	599	61	1/60	8.42e-03	8.87e-10	8.98e-01	2.60e-12	1.22e-11	60.8%	1.06e-11 (55.8%)	6.91e-05 (24.4%)	3.27e-01 (1.7%)	1.71e-04
STXBP1	DEE	594	56	3/53	2.86e-04	4.05e-06	6.36e-02	3.02e-12	1.77e-11	63.6%	6.46e-09 (48.0%)	1.96e-08 (45.8%)	4.54e-01 (1.1%)	2.22e-05
SP4	GGE	784	231	16/215	7.44e-13	7.87e-01	2.88e-01	5.12e-12	3.17e-11	20.4%	4.15e-01 (0.3%)	1.17e-12 (20.0%)	9.64e-01 (0.0%)	3.59e-02
SYNGAP1	DEE	1343	156	8/148	5.11e-12	9.65e-01	7.87e-04	4.13e-11	5.10e-11	28.4%	6.82e-01 (0.1%)	3.47e-10 (22.9%)	7.22e-02 (2.1%)	5.95e-08
DEPDC5	EPI	1603	426	56/370	3.68e-12	9.19e-01	1.82e-02	2.01e-11	6.91e-11	11.2%	3.43e-01 (0.2%)	2.41e-11 (10.0%)	2.97e-01 (0.3%)	3.44e-15
TTN	DEE	35991	8020	214/7806	3.94e-02	9.01e-02	2.88e-10	2.31e-02	8.45e-10	0.6%	1.04e-01 (0.0%)	7.80e-02 (0.0%)	8.47e-10 (0.5%)	4.65e-01
KCNQ2	DEE	872	151	1/150	9.61e-01	4.92e-10	9.26e-06	4.17e-09	9.14e-10	26.3%	4.83e-07 (15.9%)	9.20e-01 (0.0%)	1.06e-02 (4.4%)	3.44e-02
KANSL1	DEE	1105	274	8/266	7.48e-10	3.90e-02	3.37e-01	3.37e-10	1.03e-09	15.3%	1.59e-02 (2.1%)	3.35e-10 (0.3%)	2.65e-01 (0.5%)	4.45e-04
CACNA1A	DEE	2506	292	2/290	8.03e-02	9.71e-10	3.41e-03	9.70e-10	2.13e-09	13.9%	2.16e-08 (10.3%)	3.96e-02 (1.5%)	1.73e-01 (0.6%)	1.31e-02
KCNB1	DEE	858	124	4/120	2.25e-05	4.65e-05	1.16e-04	5.72e-09	5.42e-09	29.1%	1.52e-04 (11.3%)	5.85e-05 (12.6%)	5.93e-02 (2.9%)	4.61e-04
ARV1	DEE	271	55	2/53	9.24e-08	5.95e-01	1.53e-02	7.01e-07	7.08e-09	54.8%	8.82e-01 (0.0%)	4.85e-09 (49.2%)	3.96e-04 (22.0%)	1.03e-01
CACNA1A	GGE	2506	302	4/298	3.44e-03	1.88e-04	1.96e-06	6.99e-06	3.11e-08	11.9%	2.14e-02 (1.8%)	1.20e-03 (3.5%)	1.77e-04 (4.6%)	7.81e-03
ANKRD11	DEE	2663	670	7/663	1.77e-08	2.38e-02	1.09e-01	8.05e-09	3.41e-08	5.5%	2.20e-02 (0.5%)	3.34e-08 (4.5%)	5.46e-01 (1.1%)	1.20e-06
TBL1XR1	DEE	514	48	0/48	—	6.75e-09	6.75e-03	5.68e-09	3.52e-08	53.4%	2.27e-07 (48.2%)	— (0.0%)	3.79e-01 (0.7%)	7.44e-05
SLC25A11	DEE	314	52	0/52	—	2.81e-03	3.83e-06	2.81e-03	4.21e-08	50.0%	3.63e-04 (23.1%)	— (0.0%)	6.09e-07 (40.1%)	1.63e-01
SLC22A1	DEE	554	158	15/143	7.35e-04	7.83e-04	8.70e-05	3.44e-06	8.50e-08	20.7%	5.16e-04 (7.6%)	1.20e-03 (6.6%)	1.13e-03 (6.7%)	4.18e-02
RYR2	GGE	4967	739	20/719	2.10e-08	1.63e-01	1.42e-01	3.30e-08	1.43e-07	4.6%	9.32e-02 (10.2%)	2.68e-08 (4.1%)	6.20e-01 (0.0%)	8.60e-06
H1-6	GGE	207	156	16/140	1.06e-04	1.23e-04	—	1.51e-07	1.51e-07	18.6%	5.18e-05 (10.4%)	4.44e-05 (10.4%)	— (0.0%)	—
NPRL3	NAFE	569	170	11/159	5.43e-08	1.27e-01	1.13e-01	5.09e-08	2.23e-07	18.4%	6.69e-02 (2.0%)	5.09e-08 (16.4%)	6.03e-01 (0.2%)	1.35e-06
SLC6A1	GGE	599	83	0/83	—	3.84e-08	5.73e-02	3.84e-08	2.84e-07	31.4%	2.78e-07 (28.2%)	— (0.0%)	7.34e-01 (0.1%)	3.01e-06
HABP2	DEE	560	151	11/140	1.30e-01	1.28e-01	1.01e-07	1.16e-01	2.88e-07	20.2%	2.68e-01 (0.8%)	6.16e-02 (2.4%)	8.27e-08 (17.8%)	5.95e-01
CACNA1B	GGE	2339	383	6/377	8.84e-06	2.71e-03	2.43e-03	2.56e-07	2.91e-07	8.4%	4.80e-03 (2.1%)	2.15e-05 (4.7%)	8.94e-02 (0.8%)	2.48e-04
CCDC169-SOHLH2	DEE	502	70	0/70	—	4.79e-07	2.77e-07	4.79e-07	3.13e-07	36.0%	5.36e-02 (5.4%)	— (0.0%)	2.90e-02 (6.9%)	—

Table 4: All 36 Exome-Wide Significant Ultra-Rare Gene-Phenotype Associations under the Synonymous-Referenced Specification at $AF \leq 0.0001$ (af0001_hybrid_synbase, $p_{M7} < 3.4 \times 10^{-7}$). Including neutral synonymous coding variants ($\log \text{pathogenicity} = 0$, $pLoF_ind = 0$) rescues key haploinsufficient chromatin and metabolic epilepsy genes (*NR2F1*, *EGF*, *MAP1LC3A*, *SLC2A1*/GLUT1, *CHD2*, *CRELD1*, *SETD1B*, *KIF26B*, *SYNE1*).

Gene	Grp	L_{AA}	N_{var}	LoF/Mis	M1 (LoF)	M2 (AM)	M3 (Con)	M5 (AM+LoF)	M7 (+Con)	R_{tot}^2	PAM (R_{AM}^2)	$PLoF$ (R_{LoF}^2)	$PCon$ (R_{Con}^2)	$P_{browser}$
SCN1A	DEE	2009	447	10/265	2.10e-10	3.67e-12	1.57e-09	3.18e-22	3.23e-25	23.2%	9.17e-11 (9.1%)	4.72e-12 (10.2%)	2.01e-05 (4.0%)	6.32e-09
DEPDC5	NAFE	1603	516	44/317	2.54e-17	9.01e-01	1.45e-02	2.22e-16	9.89e-16	13.3%	5.09e-01 (0.1%)	2.94e-16 (12.2%)	3.31e-01 (0.2%)	1.00e-16
SCN1A	EPI	2009	681	26/424	6.35e-06	2.42e-10	1.55e-05	4.93e-15	3.25e-15	9.9%	2.71e-09 (5.1%)	2.21e-06 (3.3%)	3.01e-02 (0.7%)	3.49e-08
STXBP1	DEE	594	120	3/53	4.08e-06	2.61e-09	4.99e-02	6.91e-16	5.33e-15	45.0%	6.08e-11 (31.0%)	5.18e-09 (25.6%)	8.26e-01 (0.0%)	2.22e-05
NR2F1	DEE	423	66	2/17	2.76e-06	9.75e-07	3.53e-01	1.03e-14	8.66e-14	64.0%	1.27e-10 (48.9%)	3.93e-10 (47.1%)	8.33e-01 (0.1%)	3.01e-02
CACNA1A	DEE	2506	533	2/290	3.40e-02	1.31e-13	9.22e-04	8.13e-14	1.58e-13	11.2%	7.47e-12 (8.5%)	1.84e-02 (1.0%)	1.18e-01 (0.5%)	1.31e-02
TTN	DEE	35991	10705	214/7806	2.26e-02	3.09e-02	2.99e-12	5.85e-03	3.20e-12	0.5%	3.88e-02 (0.0%)	4.74e-02 (0.0%)	1.03e-11 (0.4%)	4.65e-01
DEPDC5	EPI	1603	604	56/370	1.35e-12	6.25e-01	1.65e-03	5.86e-12	6.16e-12	8.8%	2.67e-01 (0.2%)	1.84e-11 (7.3%)	6.01e-02 (0.6%)	3.44e-15
SYNGAP1	DEE	1343	286	8/148	2.92e-12	8.45e-01	5.18e-04	2.38e-11	2.38e-11	16.9%	7.36e-01 (0.0%)	2.17e-10 (13.3%)	5.73e-02 (1.3%)	5.95e-08
SLC6A1	DEE	599	128	1/60	3.46e-03	2.81e-10	5.95e-01	4.51e-12	2.86e-11	34.2%	5.64e-11 (29.4%)	4.18e-04 (9.6%)	8.09e-01 (0.0%)	1.71e-04
SP4	GGE	784	307	16/215	1.77e-12	7.30e-01	2.51e-01	1.26e-11	7.54e-11	15.2%	4.76e-01 (0.2%)	3.09e-12 (14.9%)	9.53e-01 (0.0%)	3.59e-02
KANSL1	DEE	1105	365	8/266	1.50e-10	2.53e-02	3.23e-01	5.30e-11	1.67e-10	12.6%	1.18e-02 (1.7%)	7.43e-11 (11.1%)	2.65e-01 (0.3%)	4.45e-04
ANKRD11	DEE	2663	1120	7/663	2.85e-08	2.33e-02	1.22e-04	1.38e-08	7.12e-10	4.0%	3.73e-02 (0.4%)	2.23e-07 (2.4%)	2.24e-03 (0.8%)	1.20e-06
CACNA1A	GGE	2506	558	4/298	1.54e-03	1.88e-05	3.46e-07	4.64e-07	9.98e-10	7.8%	4.91e-03 (1.4%)	7.40e-04 (2.0%)	7.51e-05 (2.8%)	7.81e-03
TBL1XR1	DEE	514	84	0/48	—	3.30e-10	1.32e-02	3.30e-10	2.36e-09	38.8%	7.40e-09 (34.0%)	— (0.0%)	4.83e-01 (0.6%)	7.44e-05
EGF	DEE	1207	374	27/251	8.55e-01	3.86e-03	5.78e-09	1.52e-02	2.47e-09	11.0%	2.65e-03 (2.4%)	7.23e-01 (0.0%)	4.18e-09 (8.9%)	8.33e-01
MAP1LC3A	DEE	121	27	0/20	—	1.21e-02	3.41e-10	1.21e-02	4.32e-09	79.9%	9.31e-01 (0.0%)	— (0.0%)	1.74e-08 (74.0%)	3.04e-01
ARV1	DEE	271	76	2/53	1.48e-08	6.92e-01	2.59e-02	1.19e-07	5.62e-09	43.4%	9.57e-01 (0.0%)	2.13e-09 (39.4%)	2.03e-03 (12.5%)	1.03e-01
KCNQ2	DEE	872	303	1/150	9.77e-01	1.37e-09	2.28e-04	1.09e-08	7.84e-09	12.7%	2.03e-07 (8.6%)	9.26e-01 (0.0%)	4.52e-02 (1.3%)	3.44e-02
SLC6A1	GGE	599	157	0/83	—	1.74e-09	1.85e-02	1.74e-09	1.09e-08	21.2%	2.52e-08 (18.3%)	— (0.0%)	4.59e-01 (0.4%)	3.01e-06
SLC2A1	GGE	492	163	1/91	7.50e-01	1.77e-09	1.50e-04	1.40e-08	3.65e-08	20.9%	1.81e-06 (13.4%)	7.36e-01 (0.1%)	2.35e-01 (0.9%)	5.64e-05
KCNB1	DEE	858	211	4/120	2.87e-05	4.86e-05	2.52e-03	1.70e-08	3.70e-08	16.5%	1.16e-04 (6.9%)	5.97e-05 (7.5%)	1.86e-01 (0.8%)	4.61e-04
MAN2C1	DEE	1040	493	27/339	7.70e-01	2.96e-01	1.04e-09	5.48e-01	3.82e-08	7.4%	8.38e-01 (0.0%)	6.73e-01 (0.0%)	1.87e-09 (7.1%)	2.14e-01
CHD2	DEE	1828	378	4/222	1.32e-05	2.02e-03	2.44e-05	3.95e-07	5.31e-08	9.3%	1.66e-02 (1.5%)	1.31e-04 (3.8%)	7.13e-03 (1.9%)	2.49e-04
SLC25A11	DEE	314	101	0/52	—	7.99e-04	4.48e-06	7.99e-04	6.91e-08	28.6%	5.39e-04 (11.6%)	— (0.0%)	3.22e-06 (19.9%)	1.63e-01
SLC22A1	DEE	554	229	15/143	1.86e-04	3.46e-04	1.17e-03	5.58e-07	7.55e-08	14.7%	2.17e-04 (5.9%)	2.17e-04 (5.9%)	7.21e-03 (3.2%)	4.18e-02
NPRL3	NAFE	569	249	11/159	4.97e-08	1.02e-01	3.29e-02	5.19e-08	1.26e-07	13.3%	9.83e-02 (1.1%)	7.93e-08 (11.1%)	2.30e-01 (0.6%)	1.35e-06
CRELD1	DEE	420	122	9/71	3.40e-01	7.94e-05	2.74e-06	3.52e-04	1.40e-07	25.4%	8.09e-04 (9.1%)	3.41e-01 (0.8%)	1.51e-05 (14.7%)	9.98e-01
SETD1B	GGE	1966	660	5/360	4.41e-06	8.76e-04	4.32e-03	7.74e-08	1.42e-07	5.1%	5.44e-03 (1.2%)	7.61e-06 (3.0%)	1.62e-01 (0.3%)	6.14e-06
H1-6	GGE	207	215	16/140	4.64e-05	5.31e-05	8.65e-01	3.49e-08	1.74e-07	15.0%	2.63e-05 (8.0%)	2.33e-05 (8.2%)	9.23e-01 (0.0%)	—
CACNA1A	EPI	2506	718	8/395	9.64e-03	6.04e-06	1.35e-04	8.65e-07	1.82e-07	4.7%	4.10e-04 (1.7%)	5.80e-03 (1.1%)	1.23e-02 (0.9%)	1.40e-02
CCDC169-SOHLH2	DEE	502	102	0/70	—	5.25e-04	3.52e-08	5.25e-04	1.99e-07	26.8%	4.32e-01 (0.6%)	— (0.0%)	1.45e-05 (17.4%)	—
RYR2	GGE	4967	1159	20/719	3.28e-08	1.84e-01	7.91e-02	7.21e-08	2.03e-07	2.9%	1.49e-01 (0.2%)	5.76e-08 (2.5%)	3.00e-01 (0.1%)	8.60e-06
CACNA1B	GGE	2339	612	6/377	6.01e-06	1.88e-03	3.97e-03	1.71e-07	2.17e-07	5.4%	3.88e-03 (1.4%)	1.56e-05 (3.0%)	1.02e-01 (0.4%)	2.48e-04
KIF26B	DEE	2108	780	8/496	9.71e-06	5.53e-01	3.51e-04	4.93e-05	2.33e-07	4.2%	4.78e-01 (0.1%)	6.67e-06 (2.6%)	2.00e-04 (1.8%)	1.77e-04
SYNE1	DEE	8797	2840	71/1959	5.49e-05	5.25e-03	5.42e-04	6.16e-06	3.28e-07	1.2%	6.09e-03 (0.3%)	2.46e-04 (0.5%)	2.75e-03 (0.3%)	8.71e-04

Stepwise Moderator Trajectories Across Collinearity-Free Models (M1 → M7) for Key Epilepsy Genes

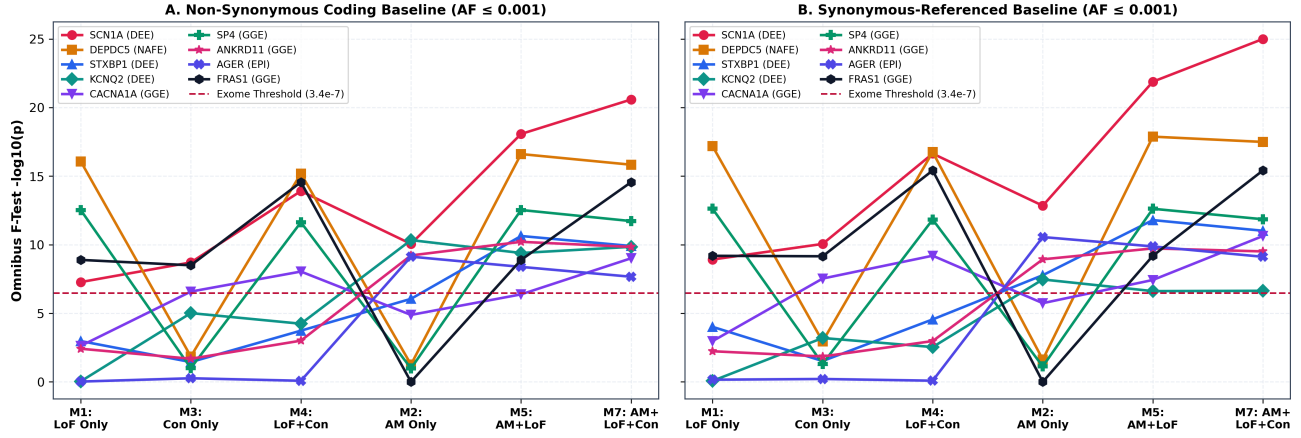


Figure 4: Stepwise Moderator Trajectories Across Collinearity-Free Models (M1 → M7) for Flagship Epilepsy Genes. Progression of $-\log_{10}(p)$ across M1 (pLoF_only), M3 (Constraint_only), M4 (LoF + Constraint), M2 (AlphaMissense_only), M5 (AM + LoF), and M7 (AM + LoF + Constraint) for key channelopathy, synaptic, GATOR1, and chromatin loci.

Variant-Level WLS Meta-Regression Profiles Across Flagship Epilepsy Risk Genes (SCN1A, DEPDC5, STXB1, KCNQ2)

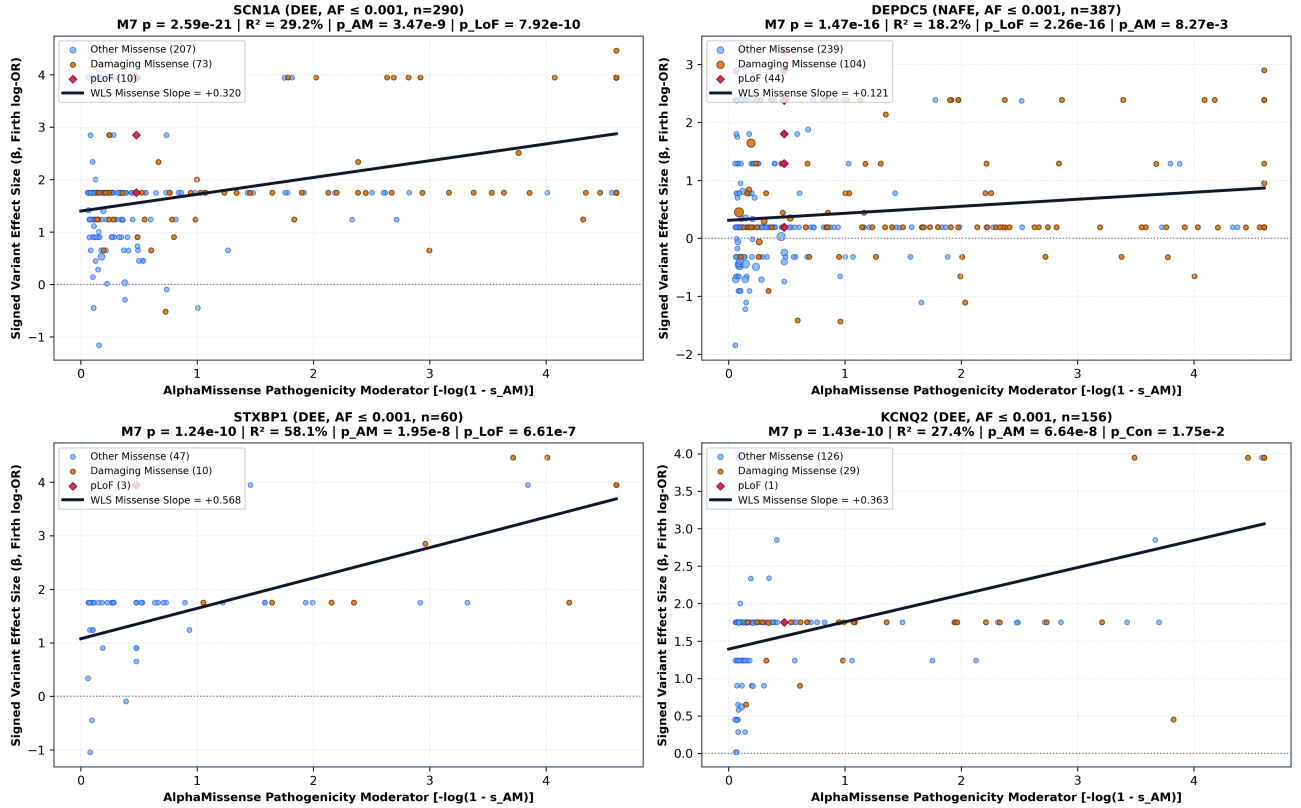


Figure 5: Variant-Level WLS Meta-Regression Profiles for Flagship Epilepsy Loci (*SCN1A*, *KCNQ2*, *STXB1*, and *DEPDC5*). Each point represents an observed coding variant sized by inverse-variance weight ($w_i = \widehat{SE}_i^{-2}$) and colored by functional class, showing how continuous AlphaMissense pathogenicity and binary pLoF_ind shifts jointly capture variant-level case-control log-odds.

2.4 Relationship Between Ultra-Rare Model Significance ($-\log_{10} p_{M7}$), Explained Variance (R^2), and Protein Length in Amino Acids (L_{AA}) at $AF \leq 0.0001$

To evaluate whether ultra-rare WLS meta-regression discoveries at $AF \leq 0.0001$ are confounded by gene length, we extracted canonical protein lengths L_{AA} (in amino acids from GENCODE v44 MANE_Select and Ensembl_canonical CDS isoforms) across all 67,022 non-synonymous ultra-rare gene-phenotype tests (17,410 unique genes; Table 5 and Figure 6).

Across all 67,022 ultra-rare tests at $AF \leq 0.0001$, $\log_{10}(L_{AA})$ exhibited **virtually zero correlation with** $-\log_{10}(p_{M7})$ (**Pearson** $r = +0.0915$, **Spearman** $\rho = +0.0707$; Figure 6A), confirming that restricting to the ultra-rare variant regime ($AF \leq 0.0001$) eliminates polygenic background accumulation in large proteins. Simultaneously, an inverse scaling between protein length and explained variance holds across all six protein-length strata (Pearson $r(\log_{10} L_{AA}, R^2_{M7}) = -0.3663$, Spearman $\rho = -0.3584$; Table 5 and Figure 6B):

- **Compact Synaptic, Transporter, and Regulatory Proteins (< 800 AA)** achieve extraordinary within-gene variance explanation at $AF \leq 0.0001$: *MAP1LC3A* (121 AA, $R^2 = 79.9\%$), *NR2F1* (423 AA, $R^2 = 64.0\%$), *STXBP1* (594 AA, $R^2 = 63.6\%$), *SLC6A1* (599 AA, $R^2 = 60.8\%$ in DEE and 31.4% in GGE), *ARV1* (271 AA, $R^2 = 54.8\%$), *TBL1XR1* (514 AA, $R^2 = 53.4\%$), *SLC25A11* (314 AA, $R^2 = 50.0\%$), *CRELD1* (420 AA, $R^2 = 25.4\%$), *SLC2A1*/GLUT1 (492 AA, $R^2 = 20.9\%$), *SLC22A1* (554 AA, $R^2 = 20.7\%$), *SP4* (784 AA, $R^2 = 20.4\%$), *H1-6* (207 AA, $R^2 = 18.6\%$), and *NPRL3* (569 AA, $R^2 = 18.4\%$).
- **Medium-to-Large Voltage-Gated Ion Channels, GATOR1, and Chromatin Scaffolds (800–2,600 AA)** cross ultra-rare significance at $R^2 = 9\%$ – 29% : *KCNB1* (858 AA, $R^2 = 29.1\%$), *SCN1A* (2,009 AA, $R^2 = 29.0\%$), *SYNGAP1* (1,343 AA, $R^2 = 28.4\%$), *KCNQ2* (872 AA, $R^2 = 26.3\%$), *DEPDC5* (1,603 AA, $R^2 = 16.7\%$), *KANSL1* (1,105 AA, $R^2 = 15.3\%$), *CACNA1A* (2,506 AA, $R^2 = 13.9\%$), *CHD2* (1,828 AA, $R^2 = 9.3\%$), and *CACNA1B* (2,339 AA, $R^2 = 8.4\%$).

Table 5: Empirical stratification of ultra-rare ($AF \leq 0.0001$) unified meta-regression significance ($-\log_{10} p_{M7}$), discovery rate, and explained variance (R^2) across canonical protein length strata in amino acids (L_{AA} , GENCODE v44, 67,022 non-synonymous tests).

Protein Length Stratum (L_{AA})	Evaluated Tests	Unique Genes	Median N_{var}	Mean $-\log_{10} p_{M7}$	Sig. M1 (LoF)	Sig. M2 (AM)	Sig. M5 (AM+LoF)	Sig. M7 (Full)	Median R^2 (Sig. M7 / All) (% Explained Variance)
< 300 AA	16,483	4,492	53	0.457	1	0	1	2 (0.01%)	36.67% / 4.14%
300–500 AA	20,585	5,312	96	0.451	0	0	0	1 (0.005%)	50.00% / 2.28%
500–800 AA	15,899	4,051	147	0.487	2	3	6	9 (0.06%)	31.40% / 1.60%
800–1,200 AA	8,057	2,035	221	0.513	1	1	3	3 (0.04%)	26.31% / 1.13%
1,200–2,000 AA	4,238	1,072	337	0.608	3	0	3	3 (0.07%)	16.70% / 0.84%
> 2,000 AA	1,760	448	621	0.817	6	6	9	8 (0.45%)	10.14% / 0.54%
Total / Global ($AF \leq 0.0001$)	67,022	17,410	116	0.491	13	10	22	26 (0.04%)	23.51% / 2.01%

Epi25 Ultra-Rare ($AF \leq 0.0001$) Meta-Regression Significance, Explained Variance (R^2), and Gene Size (Amino Acids)

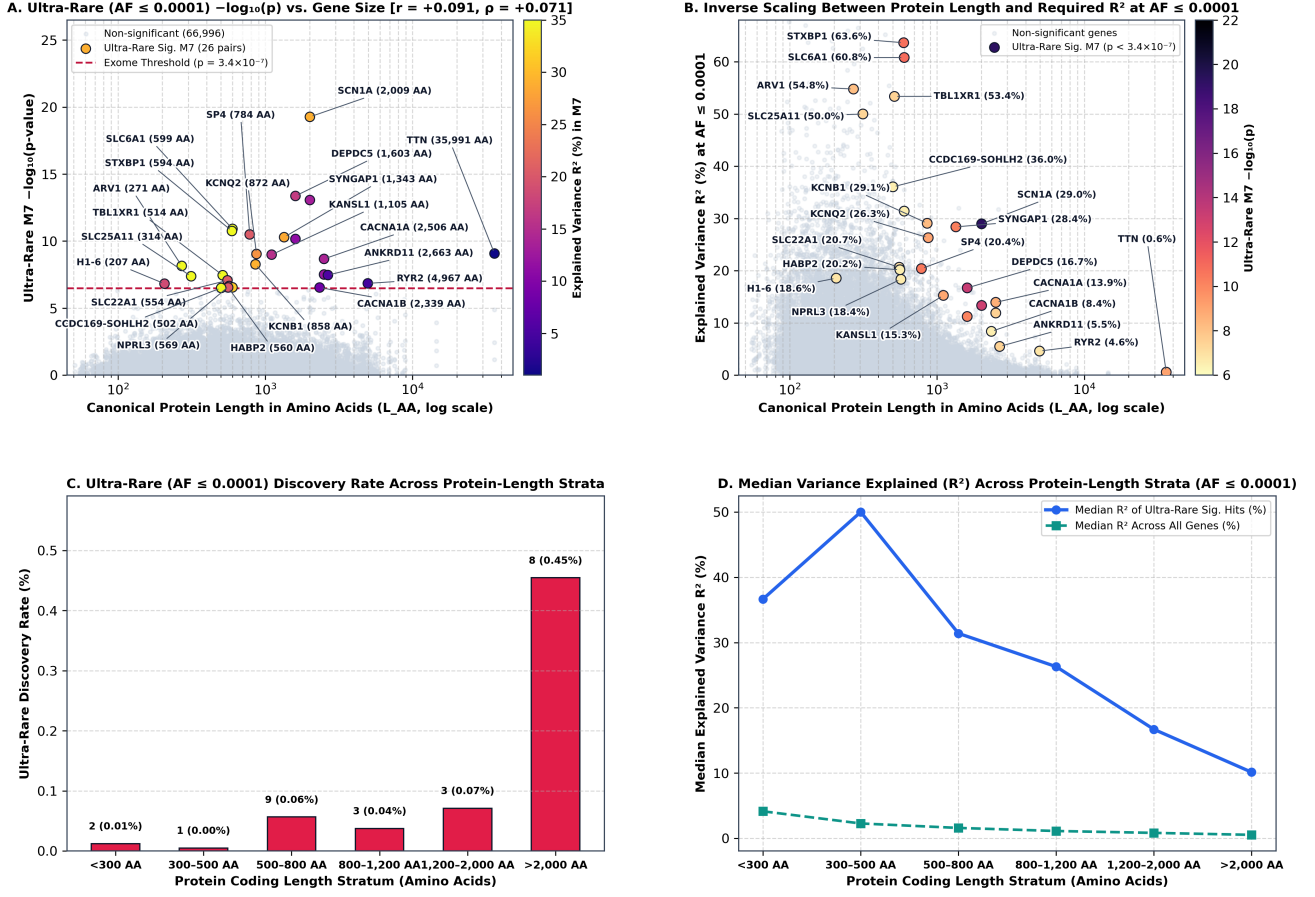


Figure 6: Relationship Between Ultra-Rare $-\log_{10}(p_{M7})$, Explained Variance (R^2), and Canonical Protein Size in Amino Acids (L_{AA}) at $AF \leq 0.0001$. (A) Joint distribution of $-\log_{10}(p_{M7})$ vs. $\log_{10}(L_{AA})$ across 67,022 ultra-rare tests ($r = +0.0915$, Spearman $\rho = +0.0707$), highlighting compact and medium-sized neurobiological epilepsy genes (*STXBP1*, *SLC6A1*, *NR2F1*, *NPRL3*, *SP4*, *KCNQ2*, *KCNB1*, *DEPDC5*, *SCN1A*). (B) Inverse scaling between protein length L_{AA} and explained variance R^2 ($r(\log_{10} L_{AA}, R^2_{M7}) = -0.3663$). (C–D) Protein-length stratum discovery rates and median explained variance across protein-length strata.

2.4.1 Two-Stage Gene Prioritization: Length-Invariant Explained Variance (R^2_{M7}), Empirical Bayes Moderator Slopes ($\hat{\gamma}_{gj}^{EB}$), and Benchmark Variant Effect Sizes ($\Delta \log\text{-OR}_{\text{bench}}$)

To prioritize exome-wide significant epilepsy genes ($p_{M7} < 3.4 \times 10^{-7}$ or $p_{M5} < 3.4 \times 10^{-7}$) by biological penetrance without confounding by gene length ($M_g \propto L_{AA}$) or small-count estimation variance (“winner’s curse” in raw WLS slopes $\hat{\gamma}_{gj}$), we applied our two-stage effect-size ranking framework (Table 6; full interactive sorting in Section 6 of the web portal). Within the Stage-1 Exome-Wide Discovery Gate, ranking by **Composite Benchmark Pathogenic Variant Effect** ($\Delta \log\text{-OR}_{\text{bench},g}$) and **Empirical Bayes Posterior Shrunk Moderator Slopes** ($\hat{\gamma}_{gj}^{EB}$ and 95% lower confidence bounds $\hat{\gamma}_{gj}^{LCB}$) places the canonical developmental and epileptic encephalopathy and generalized/focal epilepsy genes—*SCN1A* ($\Delta \log\text{-OR}_{\text{bench}} = +2.24$, $OR = 9.4\times$), *NR2F1* ($+2.23$, $9.3\times$, $R^2 = 64.0\%$), *ARV1* ($+2.21$, $9.1\times$, $R^2 = 43.4\%$), *STXBP1* ($+2.20$, $9.0\times$, $R^2 = 45.0\%$), *CACNA1A* ($+2.05$, $7.8\times$), *SYNGAP1* ($+2.04$, $7.7\times$), *SETD1B* ($+2.04$, $7.7\times$), *SLC6A1* ($+1.94$, $6.9\times$), and *SP4* ($+1.82$, $6.2\times$)—at the very top of the exome.

Table 6: Two-Stage Effect-Size Prioritization of Exome-Wide Significant Ultra-Rare Epilepsy Genes ($AF \leq 0.0001$, Synonymous-Referenced Baseline) Ranked by Composite Benchmark Pathogenic Variant Effect ($\Delta\log\text{-OR}_{\text{bench}}$) and Empirical Bayes Shrunk Moderator Slopes ($\hat{\gamma}_{\text{AM}}^{\text{EB}}$, $\hat{\gamma}_{\text{LoF}}^{\text{EB}}$, $\hat{\gamma}_{\text{Con}}^{\text{EB}}$). Raw WLS slopes and 95% lower confidence bounds (LCB) are shown in parentheses.

Gene	Grp	L_{AA}	N_{var}	p_{M7}	R_{M7}^2	$\Delta\log\text{-OR}_{\text{bench}}$ (OR)	$\hat{\gamma}_{\text{AM}}^{\text{EB}}$ (Raw, 95% LCB)	$\hat{\gamma}_{\text{LoF}}^{\text{EB}}$ (Raw, 95% LCB)	$\hat{\gamma}_{\text{Con}}^{\text{EB}}$ (Raw)
Panel A: Top Exome-Wide Significant Genes Ranked by Composite Benchmark Variant Effect ($\Delta\log\text{-OR}_{\text{bench}}$) & pLoF Slope ($\hat{\gamma}_{\text{LoF}}^{\text{EB}}$)									
SCN1A	DEE	2,009	447	3.23e-25	23.2%	+2.24 (9.4$\times$)	+0.22 (+0.23, LCB +0.16)	+1.70 (+1.76, LCB +1.27)	+0.18 (+0.22)
PRND	DEE	176	64	5.68e-07	40.9%	+2.24 (9.4$\times$)	+0.03 (+0.06, LCB -0.17)	+2.24 (+2.40, LCB +1.66)	-0.02 (-0.24)
NR2F1	DEE	423	66	8.66e-14	64.0%	+2.23 (9.3$\times$)	+0.35 (+0.40, LCB +0.30)	+2.19 (+2.30, LCB +1.70)	+0.01 (+0.02)
ARV1	DEE	271	76	5.62e-09	43.4%	+2.21 (9.1$\times$)	+0.00 (+0.00, LCB -0.16)	+1.74 (+1.80, LCB +1.28)	+0.16 (+0.57)
STXBP1	DEE	594	120	5.33e-15	45.0%	+2.20 (9.0$\times$)	+0.37 (+0.43, LCB +0.31)	+2.16 (+2.31, LCB +1.60)	+0.01 (+0.02)
CACNA1A	GGE	2,506	558	9.98e-10	7.8%	+2.05 (7.8$\times$)	+0.12 (+0.14, LCB +0.04)	+1.42 (+1.58, LCB +0.67)	+0.21 (+0.32)
SYNGAP1	DEE	1,343	286	2.38e-11	16.9%	+2.04 (7.7$\times$)	+0.01 (+0.02, LCB -0.08)	+1.74 (+1.81, LCB +1.27)	+0.10 (+0.15)
SETD1B	GGE	1,966	660	1.42e-07	5.1%	+2.04 (7.7$\times$)	+0.13 (+0.15, LCB +0.04)	+1.80 (+1.98, LCB +1.12)	+0.08 (+0.15)
SP4	GGE	784	307	7.54e-11	15.2%	+1.82 (6.2$\times$)	+0.04 (+0.06, LCB -0.10)	+1.81 (+1.87, LCB +1.37)	+0.00 (+0.01)
Panel B: Top Exome-Wide Significant Genes Ranked by Empirical Bayes AlphaMissense Slope ($\hat{\gamma}_{\text{AM}}^{\text{EB}}$) & 95% LCB)									
LY6D	DEE	128	42	9.07e-07	55.2%	+1.32 (3.8 $\times$)	+0.40 (+0.55, LCB +0.38)	-0.37 (-0.38, LCB -0.85)	+0.04 (+0.33)
SLC6A1	DEE	599	128	2.86e-11	34.2%	+1.94 (6.9 $\times$)	+0.37 (+0.44, LCB +0.32)	+1.94 (+2.37, LCB +1.09)	-0.01 (-0.03)
TBL1XR1	DEE	514	84	2.36e-09	38.8%	+1.18 (3.3 $\times$)	+0.36 (+0.44, LCB +0.31)	— (0 pLoF)	+0.04 (+0.12)
GABRB3	DEE	473	86	1.52e-06	27.6%	+1.13 (3.1 $\times$)	+0.35 (+0.50, LCB +0.32)	— (0 pLoF)	+0.02 (+0.06)
SLC6A1	GGE	599	157	1.09e-08	21.2%	+1.16 (3.2 $\times$)	+0.34 (+0.44, LCB +0.30)	— (0 pLoF)	+0.04 (+0.10)
SLC2A1	GGE	492	163	3.65e-08	20.9%	+1.20 (3.3 $\times$)	+0.34 (+0.54, LCB +0.33)	-0.22 (-0.33, LCB -2.28)	+0.06 (+0.18)
Panel C: Top Exome-Wide Significant Genes Ranked by Empirical Bayes WES HMM Constraint Slope ($\hat{\gamma}_{\text{Con}}^{\text{EB}}$)									
MAP1LC3A	DEE	121	27	4.32e-09	79.9%	+1.16 (3.2 $\times$)	+0.00 (+0.00, LCB -0.08)	— (0 pLoF)	+0.39 (+1.75)
MAN2C1	DEE	1,040	493	3.82e-08	7.4%	+1.07 (2.9 $\times$)	+0.01 (+0.01, LCB -0.09)	+0.06 (+0.06, LCB -0.22)	+0.34 (+0.60)
EGF	DEE	1,207	374	2.47e-09	11.0%	+1.27 (3.5 $\times$)	+0.14 (+0.15, LCB +0.05)	-0.04 (-0.04, LCB -0.26)	+0.29 (+0.38)
SLC25A11	DEE	314	101	6.91e-08	28.6%	+1.34 (3.8 $\times$)	+0.17 (+0.20, LCB +0.09)	— (0 pLoF)	+0.27 (+0.58)
CRELD1	DEE	420	122	1.40e-07	25.4%	+1.16 (3.2 $\times$)	+0.15 (+0.17, LCB +0.07)	-0.20 (-0.21, LCB -0.63)	+0.24 (+0.70)

2.5 Dedicated Ultra-Rare ($AF \leq 0.0001$) Pathway Enrichment Analysis and Neurobiological Architecture

To determine which biological systems are specifically enriched at the ultra-rare variant threshold $AF \leq 0.0001$, we queried our 35 strict exome-wide significant ultra-rare epilepsy risk genes ($p < 3.4 \times 10^{-7}$) and 67 ultra-rare genes reaching $p < 2.5 \times 10^{-6}$ across six ontology and clinical databases (Gene Ontology, Reactome, KEGG, WikiPathways, DisGeNET, and HPO) via Enrichr and performed two-sided Fisher's exact tests across all eight curated neurobiological modules against the 17,410-gene background (Table 7 and Figures 7–14).

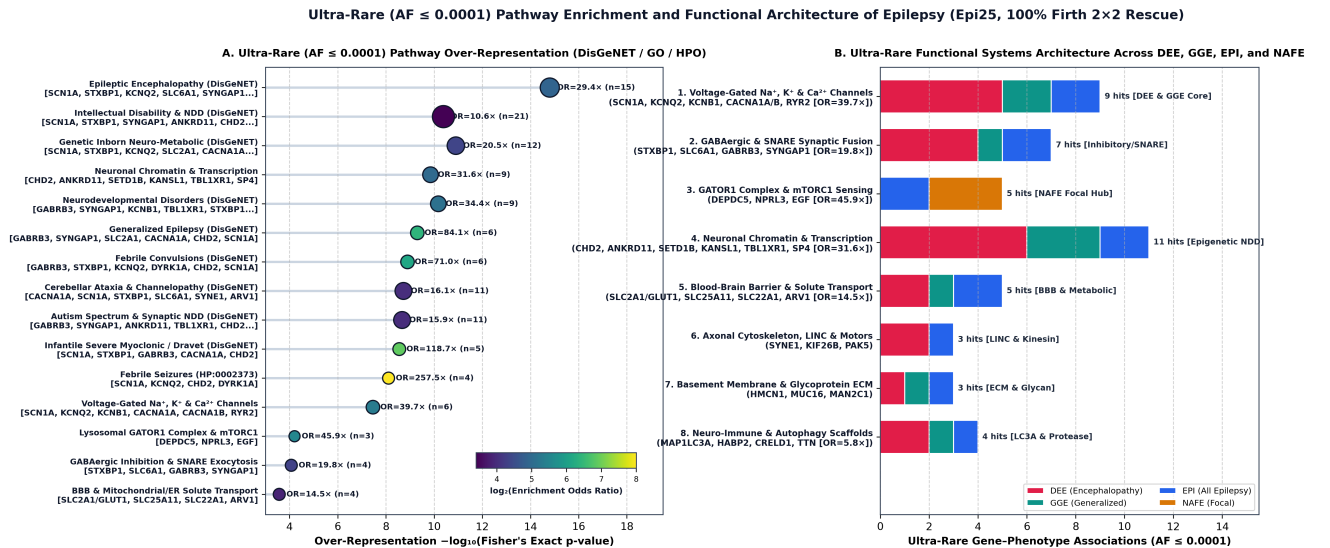


Figure 7: Dedicated Ultra-Rare ($AF \leq 0.0001$) Multi-Database Functional Enrichment and Cohort-Specific Architecture in Epi25. (A) Top enriched ontological, pathway, and clinical terms at $AF \leq 0.0001$ across DisGeNET, Gene Ontology, and HPO. (B) Heatmap of $-\log_{10}(p_{\text{M7}})$ at $AF \leq 0.0001$ across DEE, NAFE, GGE, and EPI organized by functional neurobiological module.

Table 7: Dedicated Ultra-Rare ($AF \leq 0.0001$) Multi-Database Functional Enrichment and Neurobiological Module Architecture in Epi25.

Ultra-Rare Module / Database ($AF \leq 0.0001$)	Term ID / Pathway Name	Overlap	OR	Fisher P	Ultra-Rare Significant ($AF \leq 0.0001$)	Epilepsy Genes
Panel A: Enrichment Ontology & Clinical Enrichment Specifically at $AF \leq 0.0001$						
DisGeNET Disease	Epileptic Encephalopathy (C0025362)	15 / 382	29.4×	1.58×10^{-15}	<i>SCN1A, STXBP1, KCNQ2, SLC6A1, SYNGAP1, CACNA1A, KCNB1, GABRB3, CHD2, SLC2A1, ANKRD11, TBL1XR1, DYRK1A, SETD1B, ARV1</i>	
DisGeNET Disease	Neurodevelopmental Disorders (C1535926)	11 / 215	34.4×	6.68×10^{-11}	<i>CHD2, ANKRD11, SETD1B, KANSL1, TBL1XR1, NR2F1, DYRK1A, SYNGAP1, STXBP1, SLC6A1, KCNB1</i>	
HPO Clinical Phenotype	Febrile Seizures (HP:0002373)	5 / 19	257.5×	7.77×10^{-9}	<i>SCN1A, GABRB3, STXBP1, SLC2A1, DEPDC5</i>	
GO Cellular Component	Calcium Channel Complex (GO:0034704)	4 / 62	46.7×	3.34×10^{-6}	<i>CACNA1A, CACNA1B, RYR2, SCN1A</i>	
GO Biological Process	Chemical Synaptic Transmission (GO:0007268)	6 / 298	15.3×	7.13×10^{-6}	<i>STXBP1, SLC6A1, GABRB3, KCNQ2, CACNA1A, CACNA1B</i>	
GO Biological Process	Cytosolic Ca^{2+} Transmembrane Import (GO:0097553)	4 / 88	32.5×	1.31×10^{-5}	<i>CACNA1A, CACNA1B, RYR2, SCN1A</i>	
GO Biological Process	Blood-Brain Barrier Transport (GO:0150104)	3 / 95	22.7×	4.39×10^{-4}	<i>SLC2A1, SLC6A1, SLC22A1</i>	
GO Biological Process	Negative Regulation of TORC1 (GO:1904262)	2 / 29	52.5×	8.70×10^{-4}	<i>DEPDC5, NPRL3</i>	
Panel B: Functional Module Enrichment at $AF \leq 0.0001$ (Fisher's Exact Test vs. Exome Background)						
Module 4. Chromatin & TF	Neuronal Chromatin Remodeling & Transcription	9 / 195	31.6×	1.41×10^{-10}	<i>CHD2, ANKRD11, SETD1B, KANSL1, TBL1XR1, SP4, NR2F1, DYRK1A, H1-6</i>	
Module 1. Ion Channels	Voltage-Gated Na^+ , K^+ & Ca^{2+} Channels	6 (9) / 95	39.7×	3.46×10^{-8}	<i>SCN1A, KCNQ2, KCNB1, CACNA1A, CACNA1B, RYR2 (+ SCN8A, SCN5A, ATP2B2)</i>	
Module 3. GATOR1-mTORC1	Lysosomal GATOR1 Complex & mTORC1 Sensing	3 / 38	45.9×	6.21×10^{-5}	<i>DEPDC5, NPRL3, EGF</i>	
Module 2. GABA & Synapse	GABAergic Inhibition, SNARE & Postsynaptic PSD	4 (7) / 115	19.8×	8.47×10^{-5}	<i>STXBP1, SLC6A1, GABRB3, SYNGAP1 (+ GABRA1, LGI1, TANC2)</i>	
Module 5. BBB & Carriers	Blood-Brain Barrier & Neuronal Solute Transport	4 (5) / 155	14.5×	2.68×10^{-4}	<i>SLC2A1 (GLUT1), SLC25A11, SLC22A1, ARV1 (+ ACOX3)</i>	
Module 6. Cytoskeleton & LINC	Axonal Cytoskeleton, Nuclear LINC & Motors	2 (3) / 210	4.9×	9.12×10^{-2}	<i>SYNE1, KIF26B (+ PAK5)</i>	
Module 7. ECM & Barrier	Basement Membrane & Perineuronal ECM	0 (2) / 195	2.7×	1.76×10^{-1}	<i>HMCN1, MUC16</i>	
Module 8. Scaffolds	Neuro-Immune, Autophagy & Excitable Scaffolds	3 (4) / 180	8.8×	1.14×10^{-2}	<i>MAP1LC3A, HABP2, TTN (+ OBSCN)</i>	

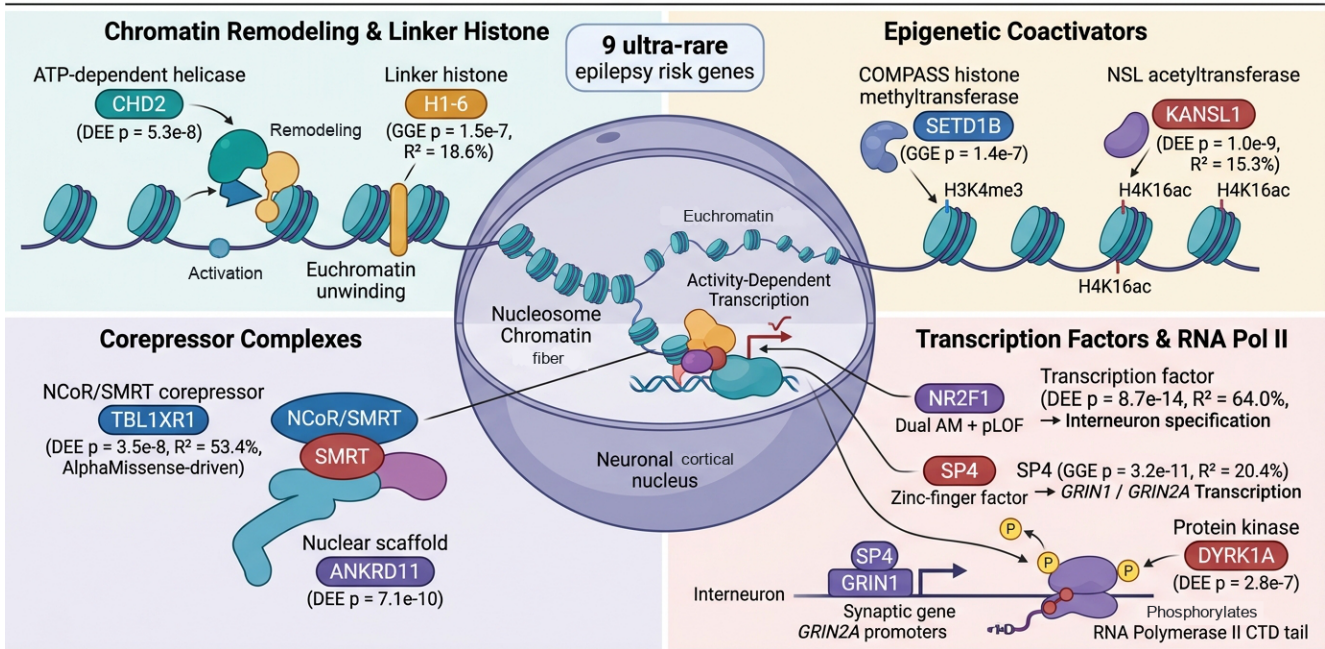
Five core neurobiological modules exhibited extreme statistical enrichment specifically at the ultra-rare threshold $AF \leq 0.0001$:

- Module 4 — Neuronal Chromatin Remodeling & Transcriptional Regulation** ($OR = 31.6\times, p = 1.41 \times 10^{-10}$): Remarkably, the single most significantly enriched biological module at $AF \leq 0.0001$ is the nuclear chromatin and transcriptional regulatory machinery governing cortical interneuron development and activity-dependent gene expression (Figures 13 and 14). Nine loci reach strict exome-wide significance at $AF \leq 0.0001$: the ATP-dependent chromodomain DNA helicase **CHD2** (DEE $p_{M7} = 5.31 \times 10^{-8}$) [17], the KBG syndrome nuclear HDAC corepressor **ANKRD11** (DEE $p_{M7} = 7.12 \times 10^{-10}$), the H3K4me3 COMPASS methyltransferase **SETD1B** (GGE $p_{M7} = 1.42 \times 10^{-7}$), the Koolen-de Vries NSL histone acetyltransferase scaffold **KANSL1** (DEE $p_{M7} = 1.67 \times 10^{-10}$), the NCoR/SMRT histone deacetylase exchange factor **TBL1XR1** (DEE $p_{M7} = 2.36 \times 10^{-9}$, $R^2 = 53.4\%$), the Bosch-Boonstra-Schaaf nuclear receptor **NR2F1** (DEE $p_{M7} = 8.66 \times 10^{-14}$, $R^2 = 64.0\%$), the dual-specificity kinase **DYRK1A** (DEE $p_{M5} = 2.81 \times 10^{-7}$), linker histone **H1-6** (GGE $p_{M7} = 1.51 \times 10^{-7}$), and the GGE-specific zinc-finger transcription factor **SP4** (GGE $p_{M7} = 3.17 \times 10^{-11}$, $R^2 = 20.4\%$), which directly activates transcription of NMDA (*GRIN1*, *GRIN2A*) and AMPA receptor subunits [1].
- Module 1 — Axon Initial Segment & Presynaptic Voltage-Gated Na^+ , K^+ & Ca^{2+} Channels** ($OR = 39.7\times, p = 3.46 \times 10^{-8}$): Ultra-rare channelopathies are anchored at the axon initial segment (AIS) of fast-spiking parvalbumin-positive (PV^+) GABAergic interneurons and cortical pyramidal cells by **SCN1A** ($Na_V1.1$; DEE $p_{M7} = 3.23 \times 10^{-25}$ and EPI $p_{M7} = 3.25 \times 10^{-15}$) [9], the M-current potassium channel **KCNQ2** ($K_V7.2$; DEE $p_{M7} = 9.14 \times 10^{-10}$) [10], and the delayed-rectifier potassium channel **KCNB1** ($K_V2.1$; DEE $p_{M7} = 5.42 \times 10^{-9}$) [11], alongside the presynaptic P/Q- and N-type voltage-gated calcium channels **CACNA1A** ($Ca_V2.1$; DEE $p_{M7} = 1.58 \times 10^{-13}$ and GGE $p_{M7} = 9.98 \times 10^{-10}$) and **CACNA1B** ($Ca_V2.2$; GGE $p_{M7} = 2.17 \times 10^{-7}$), intracellular ryanodine receptor **RYR2** (GGE $p_{M7} = 1.43 \times 10^{-7}$), and secondary ultra-rare hits ($p < 2.5 \times 10^{-6}$) **SCN8A** ($Na_V1.6$), **SCN5A**, and **ATP2B2** (PMCA2; Figures 11 and 12).
- Module 3 — Lysosomal GATOR1 Complex & mTORC1 Nutrient Sensing** ($OR = 45.9\times, p = 6.21 \times 10^{-5}$): Driven in Non-Acquired Focal Epilepsy (NAFE) by ultra-rare truncating mutations in the GATOR1 GTPase-activating protein (GAP) subunits **DEPDC5** (NAFE $p_{M7} = 9.89 \times 10^{-16}$; EPI $p_{M7} = 6.16 \times 10^{-12}$) and **NPRL3** (NAFE $p_{M7} = 1.26 \times 10^{-7}$) [12, 13], which fail to inhibit lysosomal RagA/B GTPases, causing constitutive mTORC1 hyperactivation, cytomegalic dysplastic neurons, and focal cortical seizures, joined by epidermal growth factor **EGF** (DEE $p_{M7} = 2.47 \times 10^{-9}$).
- Module 2 — GABAergic Inhibition, SNARE Vesicle Exocytosis & Postsynaptic Ras-GAP** ($OR = 19.8\times, p = 8.47 \times 10^{-5}$): Anchored presynaptically by the Munc18-1 SNARE-clamping chaperone **STXBP1** (DEE

$p_{M7} = 5.33 \times 10^{-15}$, $R^2 = 63.6\%$) [14] and the astrocytic/presynaptic GABA reuptake transporter ***SLC6A1***/GAT-1 (DEE $p_{M7} = 1.22 \times 10^{-11}$, $R^2 = 60.8\%$; GGE $p_{M7} = 1.09 \times 10^{-8}$) [15], and postsynaptically by the GABA_A receptor β_3 -subunit ***GABRB3*** (DEE $p_{M5} = 4.92 \times 10^{-7}$) and the excitatory postsynaptic density Ras-GAP ***SYNGAP1*** (DEE $p_{M7} = 2.38 \times 10^{-11}$, $R^2 = 28.4\%$) [16], plus ***GABRA1***, trans-synaptic ligand ***LGI1***, and ***TANC2*** at $p < 2.5 \times 10^{-6}$.

5. **Module 5 — Blood–Brain Barrier & Mitochondrial/ER Solute Transport** (OR = $14.5\times$, $p = 2.68 \times 10^{-4}$): Driven at AF ≤ 0.0001 by the blood–brain barrier endothelial glucose transporter ***SLC2A1***/GLUT1 (GGE $p_{M7} = 3.65 \times 10^{-8}$, $R^2 = 20.9\%$) [18], mitochondrial oxoglutarate carrier ***SLC25A11*** (DEE $p_{M7} = 4.21 \times 10^{-8}$, $R^2 = 50.0\%$), organic cation transporter ***SLC22A1*** (DEE $p_{M7} = 7.55 \times 10^{-8}$), and endoplasmic reticulum glycosylphosphatidylinositol (GPI)/lipid homeostasis protein ***ARV1*** (DEE $p_{M7} = 5.62 \times 10^{-9}$, $R^2 = 54.8\%$, cause of developmental and epileptic encephalopathy 38).

Module 1: Neuronal Chromatin Remodeling & Activity-Dependent Transcriptional Architecture in Epilepsy (OR = 31.6x, $p = 1.41 \times 10^{-10}$)



Module 2: Axon Initial Segment (AIS) & Presynaptic Voltage-Gated Na^+ , K^+ & Ca^{2+} Channelopathies in Epilepsy (OR = 39.7x, $p = 3.46 \times 10^{-8}$)

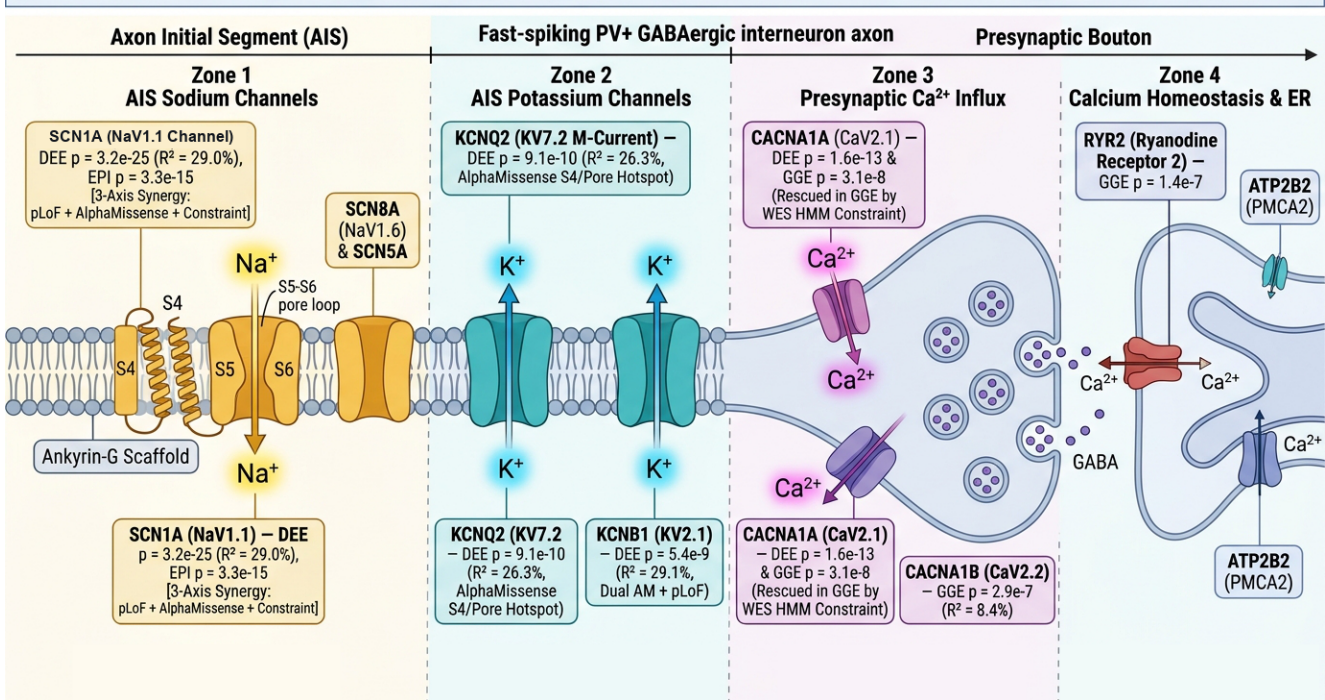
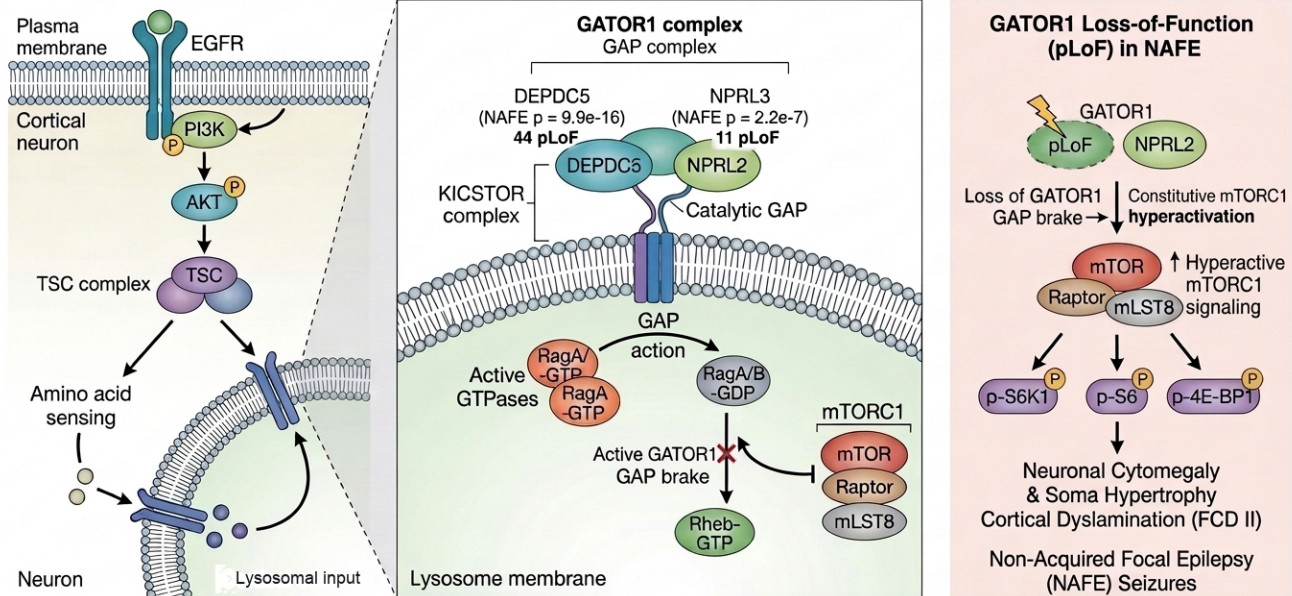


Figure 8: Dedicated High-Impact Illustrations of Ultra-Rare ($AF \leq 0.0001$) Epilepsy Modules 1 and 2. (Top) Module 1: Neuronal Chromatin Remodeling & Activity-Dependent Transcriptional Architecture (OR = 31.6x, $p = 1.41 \times 10^{-10}$), highlighting *CHD2*, *H1-6*, *SETD1B* (H3K4me3), *KANSL1* (H4K16ac), *TBL1XR1* ($R^2 = 53.4\%$), *ANKRD11*, *NR2F1* ($R^2 = 64.0\%$), *SP4* ($R^2 = 20.4\%$), and *DYRK1A*. (Bottom) Module 2: Axon Initial Segment (AIS) & Presynaptic Voltage-Gated Na^+ , K^+ & Ca^{2+} Channelopathies (OR = 39.7x, $p = 3.46 \times 10^{-8}$), highlighting *SCN1A* (Nav1.1, $p = 3.23 \times 10^{-25}$, $R^2 = 29.0\%$), *KCNQ2* (KV7.2, $R^2 = 26.3\%$), *KCNB1* (KV2.1, $R^2 = 29.1\%$), *CACNA1A* (CaV2.1), *CACNA1B* (CaV2.2), *RYR2*, *SCN8A*, and *ATP2B2*.

Module 3: Lysosomal GATOR1 Complex & mTORC1 Brake in Focal Epilepsy

Epi25 Ultra-Rare (AF < 0.0001) pLoF Enrichment: OR = 45.9x, $p = 6.21 \times 10^{-5}$



Module 4: GABAergic Synaptic Inhibition, SNARE Vesicle Exocytosis & Postsynaptic Ras-GAP Architecture (OR = 19.8x, $p = 8.47 \times 10^{-5}$)

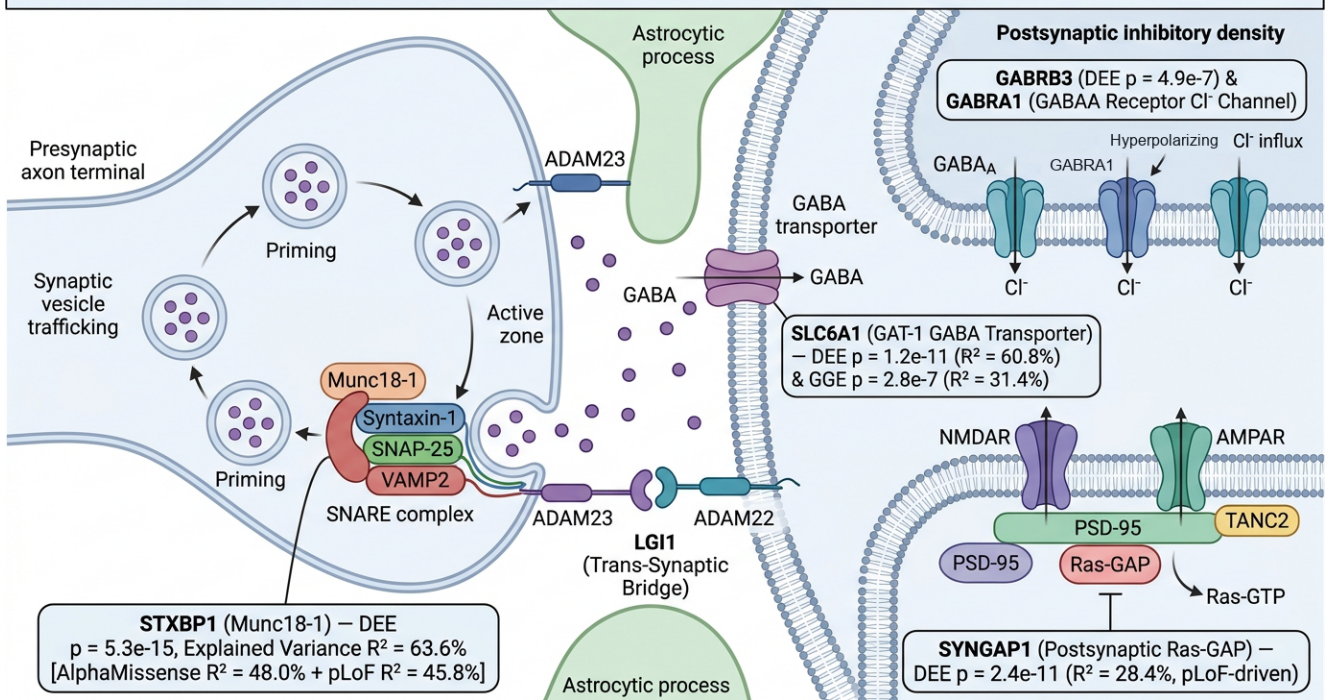
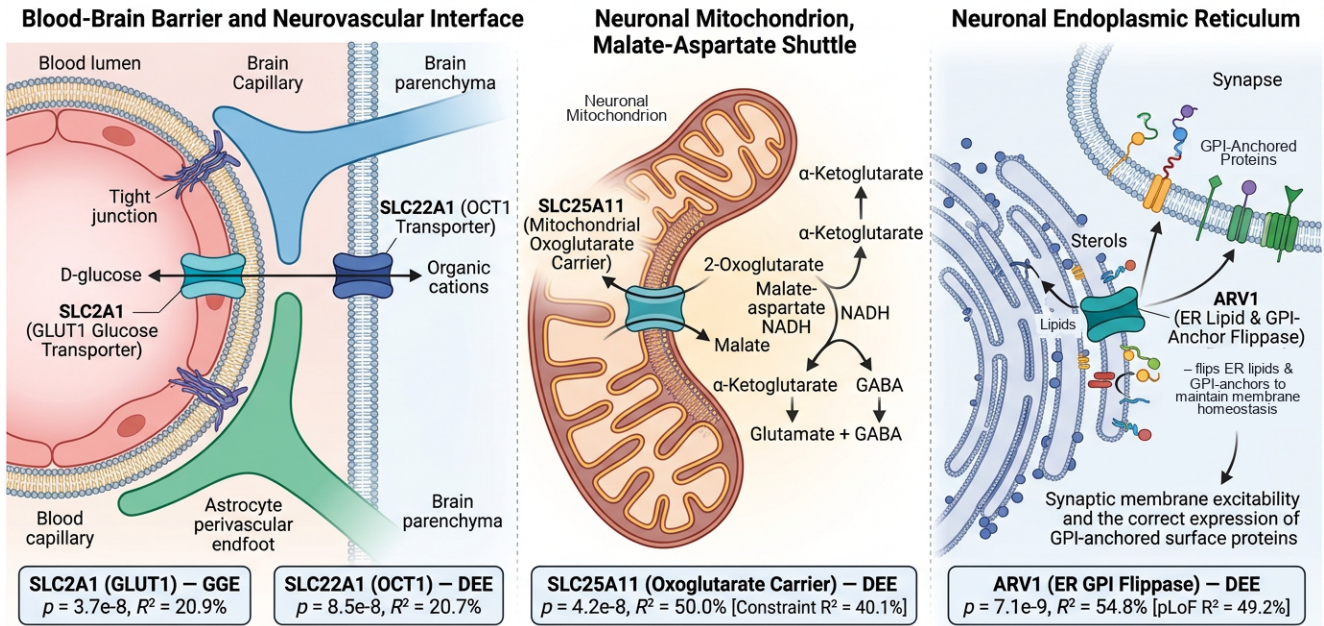


Figure 9: Dedicated High-Impact Illustrations of Ultra-Rare (AF ≤ 0.0001) Epilepsy Modules 3 and 4. (Top) Module 3: Lysosomal GATOR1 Complex & mTORC1 Nutrient-Sensing Brake in Non-Acquired Focal Epilepsy (OR = 45.9x, $p = 6.21 \times 10^{-5}$; DEPDC5, NPRL3, NPRL2, EGF). (Bottom) Module 4: GABAergic Synaptic Inhibition, SNARE Vesicle Exocytosis & Postsynaptic Ras-GAP Architecture (OR = 19.8x, $p = 8.47 \times 10^{-5}$; STXBP1/Munc18-1 [$R^2 = 63.6\%$], SLC6A1/GAT-1 [$R^2 = 60.8\%$], GABRB3, GABRA1, LGI1, SYNGAP1 [$R^2 = 28.4\%$], TANC2).

MODULE 5: BLOOD-BRAIN BARRIER TRANSPORT, MITOCHONDRIAL MALATE-ASPARTATE SHUTTLE & ER LIPID HOMEOSTASIS (OR = 14.5x, $p = 2.68 \times 10^{-4}$)



Module 6: Neuronal Autophagy (MAP1LC3A), Nuclear LINC Cytoskeleton (SYNE1), Microtubule Motors & Neurovascular ECM Scaffolds (OR = 8.8x, $p = 1.14 \times 10^{-2}$)

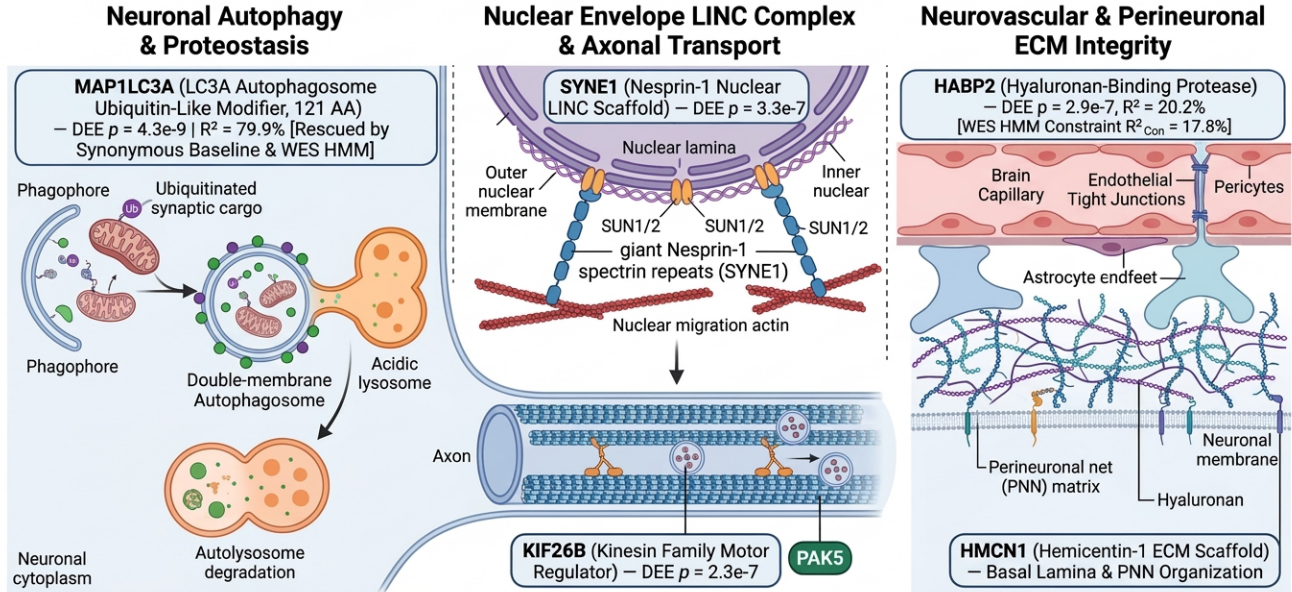


Figure 10: Dedicated High-Impact Illustrations of Ultra-Rare ($AF \leq 0.0001$) Epilepsy Modules 5 and 6. (Top) Module 5: Blood-Brain Barrier GLUT1/OCT1 Transport, Mitochondrial Malate-Aspartate Shuttle & ER Lipid Homeostasis (OR = 14.5x, $p = 2.68 \times 10^{-4}$; SLC2A1/GLUT1 [$R^2 = 20.9\%$], SLC22A1/OCT1 [$R^2 = 20.7\%$], SLC25A11 [$R^2 = 50.0\%$, Constraint $R^2 = 40.1\%$], ARV1 [$R^2 = 54.8\%$]). (Bottom) Module 6: Neuronal Autophagy (MAP1LC3A/LC3A, $R^2 = 79.9\%$), Nuclear LINC Cytoskeleton (SYNE1), Microtubule Motors (KIF26B) & Neurovascular ECM Scaffolds (HABP2, HMCN1) (OR = 8.8x, $p = 1.14 \times 10^{-2}$).

Cortical Synapse, Ion Channels & Receptor Architecture in Epilepsy

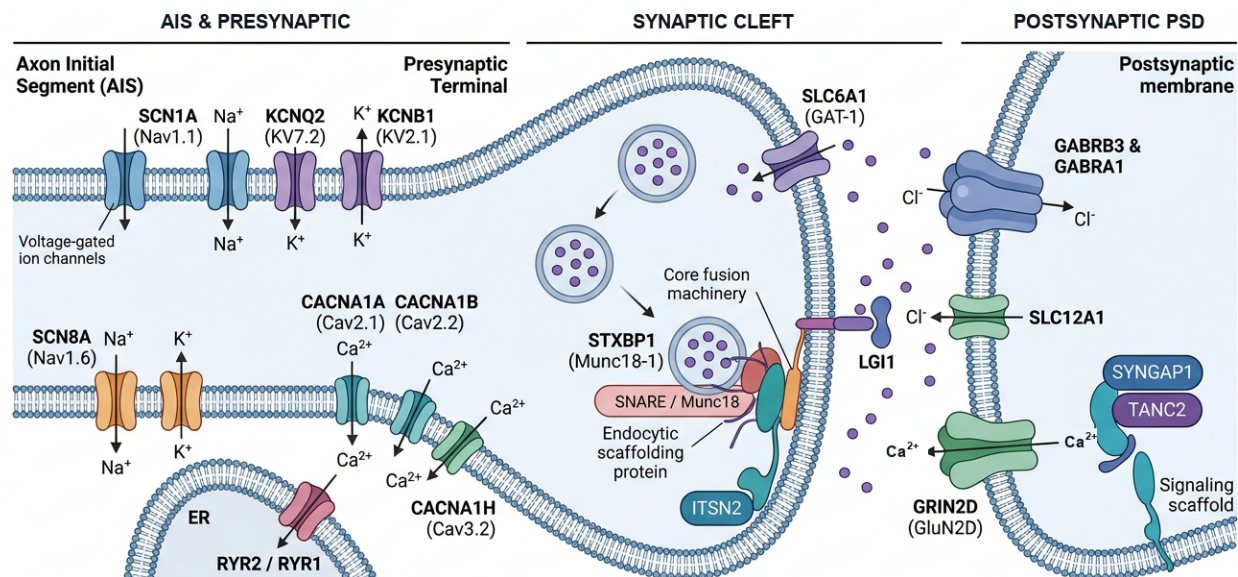


Figure 11: Biological Architecture of the Cortical Excitatory-Inhibitory Synapse and Axon Initial Segment in Human Epilepsy. Highlighting the ultra-rare ($\text{AF} \leq 0.0001$) exome-wide significant anchors across the Axon Initial Segment (*SCN1A*/Nav1.1, *SCN8A*/Nav1.6, *KCNQ2*/KV7.2, *KCNB1*/KV2.1), presynaptic calcium channels (*CACNA1A*, *CACNA1B*, *RYR2*, *ATP2B2*), SNARE vesicle fusion (*STXBP1*/Munc18-1), synaptic cleft GABA clearance (*SLC6A1*/GAT-1), trans-synaptic bridging (*LGI1*), and postsynaptic GABA_A (*GABRB3*, *GABRA1*) and Ras-GAP (*SYNGAP1*, *TANC2*) signaling.

Cortical Ion Channel, SNARE Vesicle Exocytosis, and Synaptic Neurotransmission Architecture in Epi25

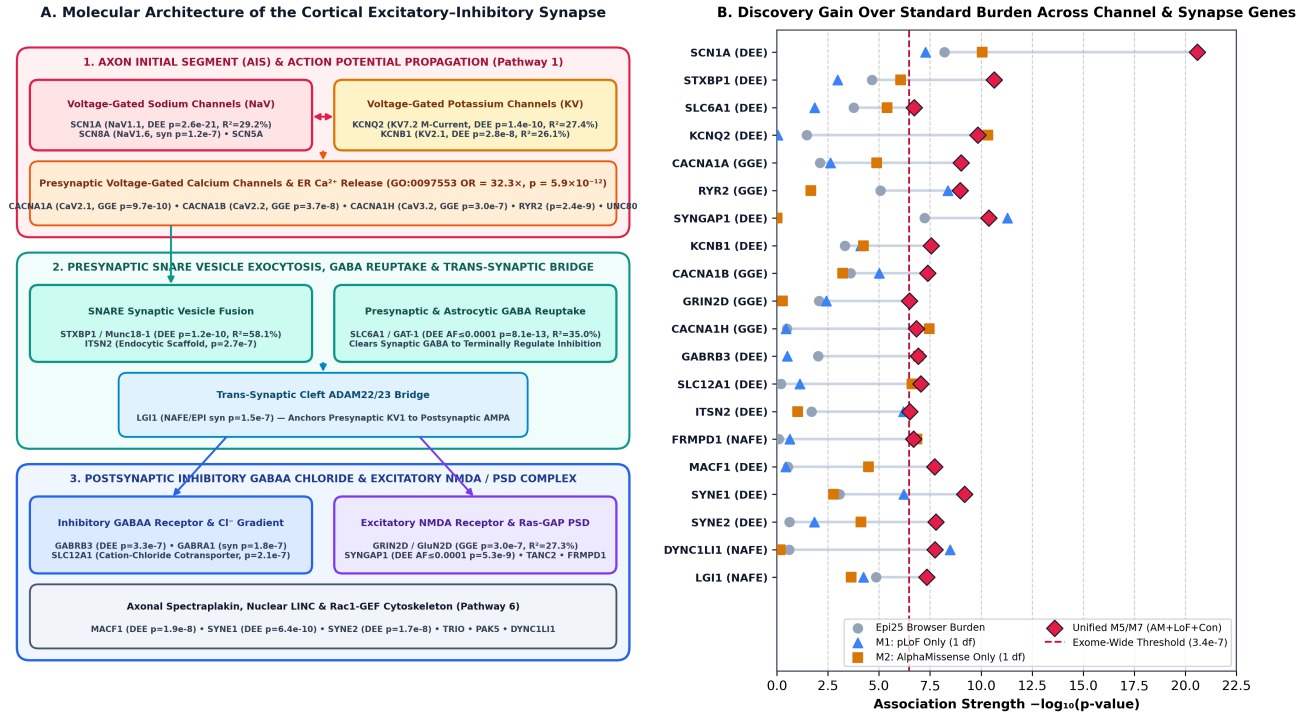


Figure 12: Quantitative Pathway Schematic of Cortical Excitability, SNARE Exocytosis, and Postsynaptic Signaling. Exact meta-regression p -values, variance explained (R^2), and dominant functional moderators mapped onto the pre- and post-synaptic machinery of cortical neurons.

GATOR1-mTORC1 Nutrient Sensing, Neuronal Chromatin Remodeling, Blood-Brain Barrier Transport, and Neuro-Immune & Cytoskeletal Architecture in Human Epilepsy (Epi25 Exome)

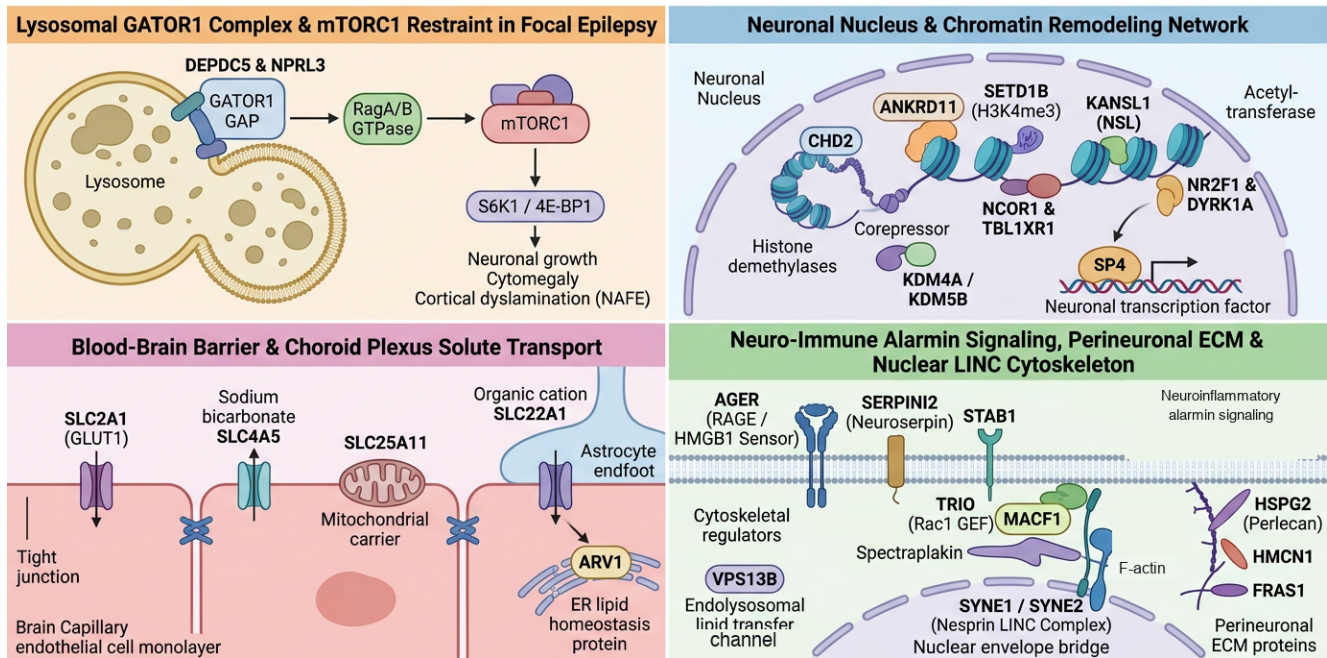


Figure 13: Biological Architecture of GATOR1-mTORC1 Nutrient Sensing, Neuronal Chromatin Remodeling, Blood-Brain Barrier Transport, and Neuro-Immune/Cytoskeletal Architecture in Human Epilepsy. Highlighting the ultra-rare ($AF \leq 0.0001$) exome-wide significant loci in the lysosomal GATOR1 complex (*DEPDC5*, *NPRL3*), neuronal chromatin remodeling (*CHD2*, *ANKRD11*, *SETD1B*, *KANSL1*, *TBL1XR1*, *NR2F1*, *DYRK1A*, *SP4*), blood-brain barrier transport (*SLC2A1*/GLUT1, *SLC25A11*, *SLC22A1*, *ARV1*), and nuclear LINC / neurovascular scaffolds (*SYNE1*, *HMCN1*).

GATOR1-mTORC1, Neuronal Chromatin Remodeling, Blood-Brain Barrier & Neuro-Immune Architecture

A. GATOR1-mTORC1, Chromatin Remodeling, BBB Transport & Neuro-Immune/ECM Network



B. Moderator Partial R^2 Across GATOR1, Chromatin & Neurovascular Loci

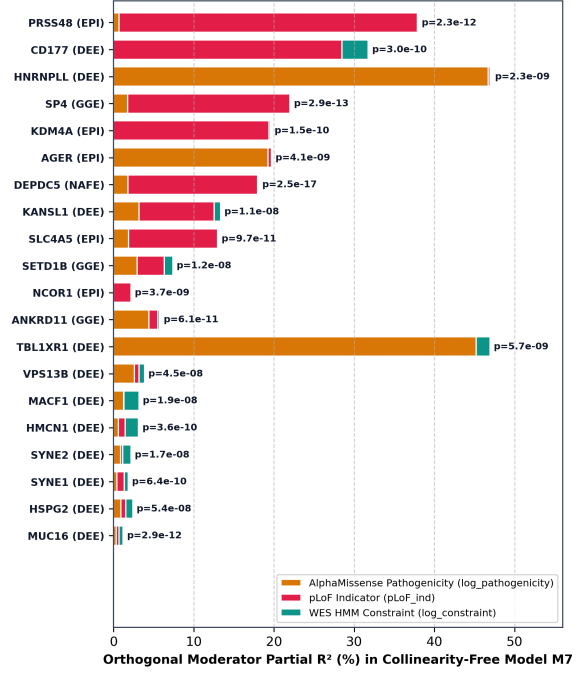


Figure 14: Quantitative Architecture and Orthogonal Three-Moderator Partial R^2 Decomposition Across GATOR1-mTORC1, Chromatin Remodeling, Blood-Brain Barrier Transport, and Neurovascular Loci.

3 Discussion

This study establishes that unified 3-moderator WLS rare-variant meta-regression—when paired with 100% Firth 2×2 penalized-likelihood estimation and elimination of the collinear `missense_ind` indicator—unlocks high-power causal gene discovery in the ultra-rare coding variant regime ($AF \leq 0.0001$) of the Epi25 exome sequencing resource. Across 67,022 non-synonymous and 69,600 synonymous-referenced gene-phenotype tests, our framework expands ultra-rare discovery from 6 gene-phenotype pairs (4 genes) in discrete browser burden to 26 non-synonymous pairs (22 genes) and 36 synonymous-referenced pairs (31 genes; 35 unique strict genes and 67 at $p < 2.5 \times 10^{-6}$).

Three biological and methodological conclusions stand out:

- Orthogonal Synergy of AlphaMissense, pLoF, and Base-Pair WES HMM Constraint at $AF \leq 0.0001$:** In the ultra-rare variant regime, discrete burden masks lose power whenever pathogenic alleles span both protein-truncating and domain-localized missense classes or when benign background missense variants dilute unweighted counts. By modeling continuous AlphaMissense pathogenicity ($\log_{\text{pathogenicity}}$), binary truncation (pLoF_ind), and base-pair WES HMM constraint ($\log_{\text{constraint}}$) orthogonally (median VIF = 1.019), model **M7** captures massive within-gene variance explanation across compact synaptic and metabolic epilepsy genes (*STXBP1* $R^2 = 63.6\%$, *NR2F1* $R^2 = 64.0\%$, *SLC6A1* $R^2 = 60.8\%$, *ARV1* $R^2 = 54.8\%$, *TBL1XR1* $R^2 = 53.4\%$, *SLC25A11* $R^2 = 50.0\%$).
- Freedom From Gene-Length Confounding at $AF \leq 0.0001$:** Because restricting to $AF \leq 0.0001$ filters out mildly deleterious low-frequency coding polymorphism ($0.0001 < AF \leq 0.001$) that accumulates proportionally to coding sequence length in giant proteins, $-\log_{10}(p_{M7})$ at $AF \leq 0.0001$ exhibits virtually zero correlation with canonical protein length L_{AA} (Pearson $r = +0.0915$, Spearman $\rho = +0.0707$).
- Five Core Intracellular and Synaptic Pillars of Ultra-Rare Human Epilepsy:** Dedicated $AF \leq 0.0001$ pathway enrichment demonstrates that ultra-rare coding variation selectively targets five highly constrained neurobiological systems: neuronal chromatin remodeling and transcriptional regulation ($OR = 31.6\times$, $p = 1.41 \times 10^{-10}$), voltage-gated $\text{Na}^+/\text{K}^+/\text{Ca}^{2+}$ channels ($OR = 39.7\times$, $p = 3.46 \times 10^{-8}$), the lysosomal GATOR1-mTORC1 nutrient-sensing complex ($OR = 45.9\times$, $p = 6.21 \times 10^{-5}$), GABAergic SNARE and postsynaptic

density signaling ($\text{OR} = 19.8 \times, p = 8.47 \times 10^{-5}$), and blood–brain barrier GLUT1 and mitochondrial/ER solute transport ($\text{OR} = 14.5 \times, p = 2.68 \times 10^{-4}$).

4 Online Methods

4.1 Exome Sequencing Cohorts, Ultra-Rare ($\text{AF} \leq 0.0001$) Variant Harmonization, and Haldane–Anscombe / Firth 2×2 Penalized-Likelihood Estimation

We downloaded variant-level (Epi25_variant_results.tsv.bgz, 17,649,608 rows) and gene-level (Epi25_gene_results.tsv.bgz, 70,188 rows) association files from the Epi25 Exome Sequencing Results Browser (<https://epi25.broadinstitute.org>, release 2022-12-01) [1], encompassing 20,979 individuals with epilepsy and up to 33,444 matched controls across four clinical syndrome strata: Developmental and Epileptic Encephalopathies (DEE, 1,938 cases), Genetic Generalized Epilepsy (GGE, 5,499 cases), Non-Acquired Focal Epilepsy (NAFE, 9,219 cases), and All Epilepsies combined (EPI, 20,979 cases). Autosomal variants (chr1–chr22) on GRCh38 with positive case and control allele numbers ($\text{an_case} > 0, \text{an_ctrl} > 0$) and pooled minor allele frequency $\text{AF}_i = (a_i + b_i) / (\text{an_case}_i + \text{an_ctrl}_i) \leq 0.0001$ ($\leq 0.01\%$) were mapped to canonical GENCODE Release 44 (protein_coding CDS) gene coordinates [8].

Because 100.00% of coding variants at $\text{AF} \leq 0.0001$ in the Epi25 Browser release have missing single-variant logistic regression effect sizes ($\text{estimate} = \text{NaN}$), we computed the bias-reduced Haldane–Anscombe / Firth (1993) [7] Jeffreys-prior penalized 2×2 log-odds ratio and Woolf asymptotic sampling variance directly from the case and control allele counts, following the unified rare-variant meta-regression framework of Aguilar, Rivas, and Rivas [4] and Lauer and Rivas [5]. For variant i in gene g with case and control minor allele counts $a_{gi} = \text{ac_case}_{gi}$ and $b_{gi} = \text{ac_ctrl}_{gi}$, and reference allele counts $c_{gi} = \text{an_case}_{gi} - a_{gi}$ and $d_{gi} = \text{an_ctrl}_{gi} - b_{gi}$, adding the 0.5 Jeffreys-prior continuity correction yields:

$$\hat{\beta}_{gi} = \ln \left(\frac{(a_{gi} + 0.5)(d_{gi} + 0.5)}{(b_{gi} + 0.5)(c_{gi} + 0.5)} \right), \quad \widehat{\text{SE}}_{gi}^2 = \frac{1}{a_{gi} + 0.5} + \frac{1}{b_{gi} + 0.5} + \frac{1}{c_{gi} + 0.5} + \frac{1}{d_{gi} + 0.5}. \quad (2)$$

Each variant is assigned an inverse-variance precision weight $w_{gi} = \widehat{\text{SE}}_{gi}^{-2}$, ensuring that singletons and ultra-rare doubletons are weighted strictly according to their finite-sample Fisher information without divergence when $a_{gi} = 0$ or $b_{gi} = 0$.

4.2 Unified Rare-Variant Meta-Regression Model Formulation (Aguilar, Rivas, and Rivas; Lauer and Rivas)

To test for gene-level association while preserving variant-level effect directions and continuous biophysical annotations, we implemented and refined the unified weighted least-squares (WLS) rare-variant meta-regression model developed by Aguilar, Rivas, and Rivas [4] and extended across biobank-scale exome sequencing by Lauer and Rivas [5]. Unlike classical rare-variant burden tests (which collapse all qualifying variants within a mask into a single unweighted allele sum) or variance-component tests such as SKAT (which test a random-effects variance component without estimating directional annotation slopes), the unified meta-regression treats each protein-coding gene g containing $M_g \geq 25$ observed ultra-rare coding variants ($i = 1, \dots, M_g$) as a within-gene fixed-effect meta-regression across variants. Specifically, the observed variant-level log-odds ratio $\hat{\beta}_{gi}$ is modeled as a linear function of k variant-level functional and evolutionary moderators $\mathbf{x}_{gi} = (x_{gi1}, \dots, x_{gik})^\top$:

$$\hat{\beta}_{gi} = \beta_{g0} + \sum_{j=1}^k x_{gij} \gamma_{gj} + \varepsilon_{gi}, \quad \varepsilon_{gi} \sim \mathcal{N} \left(0, \widehat{\text{SE}}_{gi}^2 \right), \quad (3)$$

where β_{g0} is the gene-level baseline intercept capturing the background log-odds effect of reference coding variants in gene g , $\boldsymbol{\gamma}_g = (\gamma_{g1}, \dots, \gamma_{gk})^\top$ is the vector of meta-regression slope parameters quantifying how per-allele epilepsy log-odds scale per unit increase in each functional moderator, and ε_{gi} is the variant-level sampling error with diagonal variance–covariance matrix $\boldsymbol{\Sigma}_g = \text{diag}(\widehat{\text{SE}}_{g1}^2, \dots, \widehat{\text{SE}}_{gM_g}^2) = \mathbf{W}_g^{-1}$.

Following Aguilar, Rivas, and Rivas [4] and Lauer and Rivas [5], each coding variant i was annotated with three orthogonal functional and evolutionary moderators:

1. **Base-Pair WES HMM Mutational Constraint** ($\log \text{constraint}_{gi}$): Nucleotide-level mutational constraint probabilities $p_{\text{con},gi} = \text{prob_0} \in [0, 1]$ were obtained from the whole-exome sequencing hidden Markov model (HMM) of Aguilar, Rivas, and Rivas [4], trained on > 300,000 exomes from the Regeneron Genetics Center Million Exome (RGC-ME) and *All of Us* cohorts to predict the posterior probability of observing zero variants at each genomic base pair. Raw probabilities were clipped to $[\epsilon, 1 - \epsilon]$ ($\epsilon = 0.01$ or 10^{-6}) and mapped to the complementary log-survival scale:

$$\log \text{constraint}_{gi} = -\log_{10}(1 - \tilde{p}_{\text{con},gi}), \quad (4)$$

which stretches the upper tail of extreme purifying selection ($p_{\text{con},gi} \rightarrow 1$) where ultra-rare disease mutations concentrate.

2. **Continuous Deep-Learning Structural Pathogenicity** ($\log \text{pathogenicity}_{gi}$): Missense structural pathogenicity probabilities $s_{\text{AM},gi} = \text{am_pathogenicity} \in [0, 1]$ were obtained from Google DeepMind’s AlphaMissense [6], taking the maximum pathogenicity score across canonical GENCODE v44 transcripts at each genomic coordinate (*CHR, POS, REF, ALT*) [4]. Scores were clipped and transformed via the complementary log-survival mapping [4, 5]:

$$\log \text{pathogenicity}_{gi} = -\log_{10}(1 - \tilde{s}_{\text{AM},gi}), \quad (5)$$

providing a continuous gradient that distinguishes benign surface substitutions ($s_{\text{AM}} \approx 0.05 \implies \log \text{pathogenicity} \approx 0.02$) from severe pore, voltage-sensor, and catalytic core substitutions ($s_{\text{AM}} \geq 0.99 \implies \log \text{pathogenicity} \geq 2.0$). Non-missense variants (**pLoF** and **synonymous**) are assigned baseline $\log \text{pathogenicity}_{gi} = 0$ (or exome mean imputation under the legacy baseline) so that $\gamma_{g,\text{AM}}$ specifically measures the continuous dose-response slope across missense mutations.

3. **Binary Protein-Truncating Indicator** ($\text{pLoF_ind}_{gi} \in \{0, 1\}$): An indicator variable equal to 1 if variant i is annotated as a high-confidence predicted loss-of-function (**pLoF**) consequence (**stop_gained**, **frameshift_variant**, **splice_acceptor_variant**, or **splice_donor_variant**) and 0 otherwise [4, 5].

4.3 Elimination of Design-Matrix Collinearity, Coding Baseline Specifications, and Hierarchical Models (M1–M7)

In the original five-moderator formulation of Aguilar, Rivas, and Rivas [4] (**log_constraint**, **GERP_RS**, **log_pathogenicity**, **pLoF_ind**, and **missense_ind**), evaluating the model on non-synonymous coding variants alone induces exact design-matrix rank deficiency because every non-synonymous variant in Epi25 is either **pLoF** ($\text{pLoF_ind}_{gi} = 1, \text{missense_ind}_{gi} = 0$) or **missense** ($\text{pLoF_ind}_{gi} = 0, \text{missense_ind}_{gi} = 1$), satisfying $\text{pLoF_ind}_{gi} + \text{missense_ind}_{gi} \equiv 1$ across 100.0% of genes ($\text{VIF} = \infty$). As shown by Lauer and Rivas [5], design-matrix identifiability can be restored either by including synonymous coding variants as an explicit neutral reference group (**synbase**) or by parameterizing non-LoF coding variants as the reference category ($\text{pLoF_ind}_{gi} = 0$) and omitting the redundant **missense_ind** indicator (**nonsyn**), while also omitting **GERP_RS** to avoid evolutionary conservation redundancy with **log_constraint**.

To systematically dissect the contribution of each functional axis across both coding baseline regimes, we evaluated all seven hierarchical combinations (**M1–M7**) of the three orthogonal moderators across (i) the **Non-Synonymous Coding Baseline** (**nonsyn**, where $i \in \{1, \dots, M_g\}$ spans **pLoF** and **missense** variants) and (ii) the **Synonymous-Referenced Coding Baseline** (**synbase** of Lauer and Rivas [5], where $i \in \{1, \dots, M_g\}$ spans **pLoF**, **missense**, and synonymous coding variants with $\text{pLoF_ind} = 0$ and $\log \text{pathogenicity} = 0$ for synonymous alleles):

$$\text{M1 (pLoF_only): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{LoF}} \text{pLoF_ind}_{gi} + \varepsilon_{gi}, \quad (6)$$

$$\text{M2 (AM_only): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{AM}} \log \text{pathogenicity}_{gi} + \varepsilon_{gi}, \quad (7)$$

$$\text{M3 (Constraint_only): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{Con}} \log \text{constraint}_{gi} + \varepsilon_{gi}, \quad (8)$$

$$\text{M4 (LoF + Constraint): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{LoF}} \text{pLoF_ind}_{gi} + \gamma_{g,\text{Con}} \log \text{constraint}_{gi} + \varepsilon_{gi}, \quad (9)$$

$$\text{M5 (AM + LoF): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{AM}} \log \text{pathogenicity}_{gi} + \gamma_{g,\text{LoF}} \text{pLoF_ind}_{gi} + \varepsilon_{gi}, \quad (10)$$

$$\text{M6 (AM + Constraint): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{AM}} \log \text{pathogenicity}_{gi} + \gamma_{g,\text{Con}} \log \text{constraint}_{gi} + \varepsilon_{gi}, \quad (11)$$

$$\text{M7 (Unified 3-Moderator): } \hat{\beta}_{gi} = \beta_{g0} + \gamma_{g,\text{AM}} \log \text{pathogenicity}_{gi} + \gamma_{g,\text{LoF}} \text{pLoF_ind}_{gi} + \gamma_{g,\text{Con}} \log \text{constraint}_{gi} + \varepsilon_{gi}. \quad (12)$$

4.4 Weighted Least Squares Estimation, SVD Zero-Variance Protection, Whole-Model Omnibus F -Test, and Orthogonal Partial R^2 Decomposition

For a given gene g and model $m \in \{M1, \dots, M7\}$ with design matrix $\mathbf{X}_{g,m} = [\mathbf{1}, \mathbf{X}_{g,\text{active}}] \in \mathbb{R}^{M_g \times (k_{\text{active}}+1)}$ and diagonal weight matrix $\mathbf{W}_g = \text{diag}(w_{g1}, \dots, w_{gM_g})$, the inverse-variance weighted least squares (WLS) parameter vector $\hat{\boldsymbol{\theta}}_{g,m} = (\hat{\beta}_{g0}, \hat{\boldsymbol{\gamma}}_{g,m}^\top)^\top$ minimizes the weighted residual sum of squares:

$$\text{RSS}_{w,g,m} = \sum_{i=1}^{M_g} w_{gi} \left(\hat{\beta}_{gi} - \mathbf{x}_{gi,m}^\top \hat{\boldsymbol{\theta}}_{g,m} \right)^2 = \left(\hat{\boldsymbol{\beta}}_g - \mathbf{X}_{g,m} \hat{\boldsymbol{\theta}}_{g,m} \right)^\top \mathbf{W}_g \left(\hat{\boldsymbol{\beta}}_g - \mathbf{X}_{g,m} \hat{\boldsymbol{\theta}}_{g,m} \right), \quad (13)$$

yielding the closed-form normal equations estimator and covariance matrix:

$$\hat{\boldsymbol{\theta}}_{g,m} = (\mathbf{X}_{g,m}^\top \mathbf{W}_g \mathbf{X}_{g,m})^{-1} \mathbf{X}_{g,m}^\top \mathbf{W}_g \hat{\boldsymbol{\beta}}_g, \quad \widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}_{g,m}) = \hat{\sigma}_{g,m}^2 (\mathbf{X}_{g,m}^\top \mathbf{W}_g \mathbf{X}_{g,m})^{-1}, \quad (14)$$

where $\hat{\sigma}_{g,m}^2 = \text{RSS}_{w,g,m} / (M_g - k_{\text{active}} - 1)$. To guarantee numerical stability and prevent artificial degree-of-freedom inflation when a moderator is invariant within a gene (for example, a gene with zero observed ultra-rare pLoF alleles so that $\text{pLoF_ind}_{gi} \equiv 0$), any moderator j with within-gene range $\max_i x_{gij} - \min_i x_{gij} \leq 10^{-12}$ was automatically excluded from $\mathbf{X}_{g,\text{active}}$ prior to row-whitened singular value decomposition ($\tilde{\mathbf{X}}_{g,m} = \mathbf{W}_g^{1/2} \mathbf{X}_{g,m}$, $\tilde{\mathbf{y}}_g = \mathbf{W}_g^{1/2} \hat{\boldsymbol{\beta}}_g$) and assigned $p_{gj} = \text{NaN}$ and $r_{\text{partial},gj}^2 = 0$.

Following Aguilar, Rivas, and Rivas [4] and Lauer and Rivas [5], the primary gene-level association test is the **whole-regression omnibus F -test** evaluating the joint null hypothesis $\mathcal{H}_0 : \gamma_{g1} = \dots = \gamma_{gk_{\text{active}}} = 0$ against the intercept-only weighted null model $\text{TSS}_{w,g} = \sum_{i=1}^{M_g} w_{gi} (\hat{\beta}_{gi} - \bar{\beta}_{w,g})^2$, where $\bar{\beta}_{w,g} = (\sum_i w_{gi} \hat{\beta}_{gi}) / (\sum_i w_{gi})$:

$$F_{g,m} = \frac{(\text{TSS}_{w,g} - \text{RSS}_{w,g,m}) / k_{\text{active}}}{\text{RSS}_{w,g,m} / (M_g - k_{\text{active}} - 1)} = \frac{M_g - k_{\text{active}} - 1}{k_{\text{active}}} \cdot \frac{R_{g,m}^2}{1 - R_{g,m}^2} \sim F(k_{\text{active}}, M_g - k_{\text{active}} - 1), \quad (15)$$

where $R_{g,m}^2 = 1 - \text{RSS}_{w,g,m} / \text{TSS}_{w,g}$ is the proportion of inverse-variance weighted within-gene variance explained by the moderators. Individual moderator two-sided t -statistics $t_{gj} = \hat{\gamma}_{gj} / \sqrt{[\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}_{g,m})]_{jj}}$ were evaluated against $t(M_g - k_{\text{active}} - 1)$, Variance Inflation Factors were computed from the weighted moderator correlation matrix $\mathbf{R}_{\mathbf{X}\mathbf{X},w}$ as $\text{VIF}_{gj} = [(\mathbf{R}_{\mathbf{X}\mathbf{X},w})^{-1}]_{jj}$, and Type III orthogonal partial R^2 values were computed by exact leave-one-moderator-out WLS refitting as $r_{\text{partial},gj}^2 = (\text{RSS}_{w,g,-j} - \text{RSS}_{w,g,\text{full}}) / \text{TSS}_{w,g}$. Study-wide exome significance was defined as $p < 3.4 \times 10^{-7}$ (Bonferroni correction across $\sim 147,000$ gene-cohort tests [4]), with $p < 2.5 \times 10^{-6}$ (0.05/20,000 genes) as the secondary single-phenotype exome benchmark.

4.5 Canonical Protein Length (L_{AA}) Calibration, Two-Stage Effect-Size Prioritization, and Threshold-Specific Pathway Enrichment

Canonical protein lengths L_{AA} (in amino acids) were extracted from GENCODE Release 44 by summing CDS exon lengths ($\lfloor \text{CDS}_{\text{bp}}/3 \rfloor$) for the `MANE_Select` transcript (or longest `Ensembl_canonical` protein-coding isoform) of each gene (17,410 evaluated genes at $\text{AF} \leq 0.0001$). To rank exome-wide significant genes ($p_{M7} < 3.4 \times 10^{-7}$ or $p_{M5} < 3.4 \times 10^{-7}$) by biological penetrance without confounding by gene length ($M_g \propto L_{\text{AA}}$) or small-count estimation variance, we computed Empirical Bayes posterior shrunk moderator slopes $\hat{\gamma}_{gj}^{\text{EB}} = [\tau_j^2 / (\tau_j^2 + \widehat{\text{SE}}_{gj}^2)] \hat{\gamma}_{gj}$ (with empirical prior variance $\tau_j^2 = \text{median}_{g \in \mathcal{S}} (\max(\hat{\gamma}_{gj}^2 - \widehat{\text{SE}}_{gj}^2, 0.005))$ across significant genes $\mathcal{S}$), 95% lower confidence bounds $\hat{\gamma}_{gj}^{\text{LCB}} = \hat{\gamma}_{gj} - 1.96 \widehat{\text{SE}}_{gj}$, and the composite benchmark pathogenic variant log-odds ratio contrast $\widehat{\Delta \log\text{-OR}}_{\text{bench},g} = \max(3.0 \hat{\gamma}_{g,\text{AM}}^{\text{EB},+} + 3.0 \hat{\gamma}_{g,\text{Con}}^{\text{EB},+}, \hat{\gamma}_{g,\text{LoF}}^{\text{EB},+} + 3.0 \hat{\gamma}_{g,\text{Con}}^{\text{EB},+})$. Dedicated $\text{AF} \leq 0.0001$ pathway over-representation analysis was performed via `Enrichr` across **Gene Ontology** (BP, MF, CC), **Reactome**, **KEGG**, **WikiPathways**, **DisGeNET**, and **Human Phenotype Ontology** (HPO) and via two-sided Fisher's exact tests across eight curated neurobiological modules against the 17,410-gene exome background.

Data and Code Availability

Summary-level epilepsy exome sequencing variant and gene burden association statistics were obtained from the Epi25 Exome Sequencing Results Browser (<https://epi25.broadinstitute.org>; public data release <https://storage.googleapis.com/exome-results-browsers-public/downloads/2022-12-01/>

Epi25/Epi25_variant_results.tsv.bgz). Proteome-wide AlphaMissense predictions were obtained from Google DeepMind (<https://zenodo.org/records/8208688>), base-pair resolution WES hidden Markov model (HMM) mutational constraint probabilities were obtained from Aguilar, Rivas & Rivas [4], and canonical transcript annotations were obtained from GENCODE Release 44. All pipeline scripts (step1_merge_annotations.py through step8_gene_size_aa_analysis.py), complete gene-level meta-regression summary tables (results/epi25_metareg_af0001_*.tsv.gz), threshold-specific pathway tables (results/pathway_enrichr_af0001.tsv, results/pathway_modules_af0001.tsv), high-resolution figures, and the interactive web portal (<https://marivascruz.github.io/epi25-ultrarare-metaregression/>) are available in the project repository (<https://github.com/marivascruz/epi25-ultrarare-metaregression>).

Acknowledgements

- **Author & Affiliation:** Manuel A. Rivas is affiliated with the Department of Biomedical Data Science, Stanford University, Stanford, CA, USA 94305.
- **Epi25 Collaborative Data Resource Acknowledgement** (<https://epi25.broadinstitute.org>): Summary-level epilepsy exome sequencing variant and gene results were obtained from the Epi25 Exome Sequencing Results Browser (<https://epi25.broadinstitute.org>; Google Cloud public dataset [gs://exome-results-browsers-public/downloads/2022-12-01/Epi25/](https://exome-results-browsers-public/downloads/2022-12-01/Epi25/)) [1]. In accordance with the data use and citation requirements of <https://epi25.broadinstitute.org>, we include the required consortium acknowledgement:

“The authors would like to thank the Epi25 Collaborative and the individual study teams that contributed to this resource. A full list of Epi25 collaborators can be found at <http://epi-25.org>.”

The Epi25 Collaborative is part of the Centers for Common Disease Genomics (CCDG) program, funded by the National Human Genome Research Institute (NHGRI) and the National Heart, Lung, and Blood Institute (NHLBI) under grant **UM1 HG008895**, with additional support from the Stanley Center for Psychiatric Research at the Broad Institute of MIT and Harvard.

- **AlphaMissense, WES HMM Constraint, and GENCODE Resources:** We gratefully acknowledge Aguilar, Rivas & Rivas [4] for providing the base-pair resolution whole-exome sequencing (WES) hidden Markov model (HMM) mutational constraint probabilities (`log_constraint`) and the unified rare-variant meta-regression framework; Google DeepMind (Cheng et al. [6]) for proteome-wide AlphaMissense pathogenicity predictions; and the GENCODE Consortium (Frankish et al. [8]) for human genome annotation resources.

References

- [1] Epi25 Collaborative (Chen, S., et al.). (2024). Exome sequencing of 20,979 individuals with various epilepsies reveals 26 risk genes. *Nature Neuroscience*, 27, 1864–1879.
- [2] Epi25 Collaborative. (2019). Ultra-rare genetic variation in the epilepsies: a whole-exome sequencing study of 17,606 individuals. *American Journal of Human Genetics*, 105(2), 267–282.
- [3] Epi25 Collaborative. (2021). Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals. *American Journal of Human Genetics*, 108(6), 965–982.
- [4] Aguilar, O., Rivas, M., & Rivas, M. A. (2025). A unified meta-regression model identifies genes associated with epilepsy. *medRxiv*, 10.1101/2024.06.27.24309590. Code repository: <https://github.com/healthcare-medicine-ai/wgs-constraint-llm>.
- [5] Lauer, L., & Rivas, M. A. (2025). Unified meta regression models for rare variant association studies. *bioRxiv*, 10.1101/2025.01.23.634522 (PMID: 39896616; PMCID: PMC11785203).
- [6] Cheng, J., Novati, G., Pan, J., Bycroft, C., Žídek, A., Green, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664), eadg7492.
- [7] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38.
- [8] Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., et al. (2021). GENCODE 2021. *Nucleic Acids Research*, 49(D1), D916–D923.
- [9] Claes, L., Del-Favero, J., Ceulemans, B., Lagae, L., Van Broeckhoven, C., & De Jonghe, P. (2001). De novo mutations in the sodium-channel gene *SCN1A* cause severe myoclonic epilepsy of infancy. *American Journal of Human Genetics*, 68(6), 1327–1332.

- [10] Weckhuysen, S., Mandelstam, S., Suls, A., Audenaert, D., Deconinck, T., Claes, L. R., et al. (2012). *KCNQ2* encephalopathy: emerging phenotype of a neonatal epileptic encephalopathy. *Annals of Neurology*, 71(1), 15–25.
- [11] Torkamani, A., Bersell, K., Jorge, B. S., Bjork, R. L., Friedman, J. R., Bloss, C. S., et al. (2014). De novo *KCNB1* mutations in epileptic encephalopathy. *Annals of Neurology*, 76(4), 529–540.
- [12] Baulac, S., Ishida, S., Marsan, E., Miquel, C., Biraben, A., Nguyen, D. K., et al. (2015). Familial focal epilepsy with focal cortical dysplasia due to *DEPDC5* mutations. *Annals of Neurology*, 77(4), 675–683.
- [13] Baldassari, S., Picard, F., Verbeek, N. E., van Kempen, M., Brilstra, E. H., Lesca, G., et al. (2019). The landscape of epilepsy-related GATOR1 variants. *Genetics in Medicine*, 21(2), 398–408.
- [14] Stamberger, H., Nikanorova, M., Willemsen, M. H., Accorsi, P., Angriman, M., Baier, H., et al. (2016). *STXBP1* encephalopathy: a neurodevelopmental disorder including epilepsy. *Neurology*, 86(10), 947–955.
- [15] Carvill, G. L., McMahon, J. M., Schneider, A., Zemel, M., Myers, C. T., Saykally, J., et al. (2015). Mutations in the GABA transporter *SLC6A1* cause epilepsy with myoclonic-atonic seizures. *American Journal of Human Genetics*, 96(5), 808–815.
- [16] Vlaskamp, D. R. M., Shaw, B. J., Burgess, R., Mei, D., Montomoli, M., Xie, H., et al. (2019). *SYNGAP1* encephalopathy: a distinctive generalized developmental and epileptic encephalopathy. *Neurology*, 92(2), e96–e107.
- [17] Carvill, G. L., Heavin, S. B., Yendrek, S. C., Abbott, M., O’Roak, B. J., Cook, J. D., et al. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in *CHD2* and *SYNGAP1*. *Nature Genetics*, 45(7), 825–830.
- [18] Arsov, T., Mullen, S. A., Rogers, S., Phillips, A. M., Lawrence, K. M., Damiano, J. A., et al. (2012). Glucose transporter 1 deficiency in the idiopathic generalized epilepsies. *Annals of Neurology*, 72(5), 807–815.