Discovery report for eds

Research Objective

Make novel discoveries about EDS

Dataset Description

Contents (8 files total):
1. Data Files (5 CSV files)
clinvar$_{\text{eds__variants}}$.csv (7,025 records)
Clinical genetic variants from ClinVar with pathogenicity classifications
Includes: variant IDs, genomic locations, associated genes, clinical significance, risk alleles
Top genes: TNXB (2,164), ADAMTS2 (1,562), PLOD1 (777), ZNF469 (634), COL5A1 (320)
Clinical significance: 545 pathogenic/likely pathogenic, 3,221 uncertain significance, 2,997 benign/likely benign
dbvar$_{\text{eds__structural}_{\text{variants}}}$.csv (691 records)
Large structural variants (deletions, duplications, inversions) associated with EDS
Includes: structural variation IDs, supporting variants, genomic locations
omim$_{\text{morbid__eds}_{\text{genes}}}$.csv (27 records)
OMIM Morbid Map genes associated with EDS phenotypes
Comprehensive list of established EDS genes with inheritance patterns
orphanet$_{\text{eds__genes}}$.csv (20 records)
Gene-disease relationships from Orphanet rare disease database
European perspective on EDS genetics and clinical classification
g2p$_{\text{eds__genes}}$.csv (2 records)
Highly curated gene-phenotype relationships from G2P database
Evidence-based genetic testing recommendations
2. Documentation Files (3 files)
README.md (10,582 characters)
Comprehensive documentation including:
Detailed description of each dataset with column definitions
EDS subtypes and associated genes (Classical, Vascular, Hypermobile, Kyphoscoliotic, etc.)
Clinical significance distributions and top genes by variant count
Data quality assessment and limitations
Access information and citations for all source databases
Usage examples in both Python and R
Information about related resources (GEO, AlphaFold, Zenodo)
Version information and changelog
DATA$_{\text{DICTIONARY}}$.csv (28 field definitions)
Field-by-field reference guide for all datasets
Includes dataset name, field name, description, and examples
Quick reference for understanding data structure
SUMMARY$_{\text{STATISTICS}}$.csv (35 metrics)
Key statistics across all datasets
Breakdown by database (ClinVar, dbVar, OMIM, Orphanet, G2P)
Top 10 genes by variant count
Clinical significance distribution
Collection metadata
Dataset Overview
Total Records: 7,765 entries from 5 biological databases

Collection Date: November 15, 2024
Data Sources: ClinVar, dbVar, OMIM, Orphanet, G2P (via Ensembl Phenotype API)
Key Statistics:
96 unique genes represented in ClinVar data
545 pathogenic/likely pathogenic variants for clinical interpretation
2,164 TNXB variants (hypermobile EDS candidate gene - most common subtype)
All major EDS subtypes covered: Classical, Vascular, Hypermobile, Kyphoscoliotic, Dermatosparaxis, and others
Standardized ontology: All phenotypes mapped to MONDO:0020066 (Ehlers-Danlos syndrome)
Clinical Utility:
Variant interpretation: Pathogenicity assessments for 7,025 genetic variants
Gene panel design: Comprehensive gene list for diagnostic testing
Genotype-phenotype studies: Variant-phenotype associations across subtypes
Structural variant analysis: 691 large-scale genomic rearrangements
Cross-database validation: Multiple sources for gene-disease relationships
Data Quality Features:
Clinical annotations from authoritative sources (ClinVar, OMIM)
Multiple database cross-referencing
Standardized disease ontology mapping
Comprehensive gene coverage across EDS subtypes
45.8% variants with uncertain significance (VUS) - requires functional validation
Hypermobile EDS genetic basis unclear (most common subtype)
Usage Examples Provided:
The README includes code examples for:
Loading and filtering data in Python (pandas)
Loading and filtering data in R
Extracting pathogenic variants
Counting variants per gene
Searching by gene or genomic location
Additional Resources Referenced:
The package documentation also references datasets NOT included in the zip file but available separately:
GEO: 16 gene expression datasets for molecular profiling
AlphaFold: Protein structure predictions for collagen genes (e.g., COL1A1)
Zenodo: Recent EDS research publications
File Format & Accessibility:
All data files in CSV format (UTF-8 encoded)
Compatible with Excel, Python, R, and other analysis tools
Human-readable documentation in Markdown format
Ready for immediate use in research projects
This comprehensive package provides researchers with curated, well-documented genetic and clinical data for EDS research, including variant interpretation, gene discovery, and genotype-phenotype correlation studies.

Summary of Discoveries

Discovery 1: Haploinsufficiency and divergent structural-variant mechanisms in EDS genes

Classic Ehlers-Danlos syndrome (cEDS) is overwhelmingly driven by haploinsufficiency in COL5A1 and COL5A2, with loss-of-function point mutations and structural variants constituting nearly all pathogenic events. Despite similar intronic repeat architectures, COL5A1 structural variants map

to inverted Alu hotspots consistent with non-allelic homologous recombination, whereas COL5A2 exhibits a distinct, non–Alu-mediated structural variant spectrum and larger event sizes.

Discovery 2: Loss-of-function dominance, domain-level paradoxes, and variant classification in TNXB

Across ClinVar- and gnomAD-derived datasets, TNXB pathogenicity is overwhelmingly driven by predicted loss-of-function (LoF) variants, while domain-resolved analyses reveal a striking mosaic: the N terminus is missense-depleted and conserved without clinical missense enrichment, whereas the conserved C-terminal fibrinogen-like (FBG) domain is paradoxically tolerant to both missense and LoF variation in the general population. A stratified interpretation framework reaches near-perfect specificity for TNXB but is sensitivity-limited by incomplete annotations and the disproportionate complexity of indels.

Discovery 3: Noncanonical missense architecture in COL5A1 and the absence of genotype-phenotype discriminants in COL1A2 EDS

Across fibrillar collagens, COL5A1 displays a noncanonical missense architecture: pathogenic missense variants are not dominated by triple-helix glycine substitutions and instead converge on the C-terminal NC1 domain, particularly on cysteine residues. In contrast, multiple independent analyses fail to find simple genotype–phenotype discriminants that separate cardiac-valvular from classic EDS for COL1A2, arguing for more complex, context-dependent determinants of phenotype.

Haploinsufficiency and divergent structural-variant mechanisms in EDS genes

Summary

Classic Ehlers-Danlos syndrome (cEDS) is overwhelmingly driven by haploinsufficiency in COL5A1 and COL5A2, with loss-of-function point mutations and structural variants constituting nearly all pathogenic events. Despite similar intronic repeat architectures, COL5A1 structural variants map to inverted Alu hotspots consistent with non-allelic homologous recombination, whereas COL5A2 exhibits a distinct, non–Alu-mediated structural variant spectrum and larger event sizes.

Background

Ehlers-Danlos syndromes (EDS) are heritable connective tissue disorders arising from defects in extracellular matrix proteins and their processing. In cEDS, type V collagen (encoded by COL5A1 and COL5A2) is central to dermal and connective tissue integrity, and disease mechanisms can include loss of gene dosage (haploinsufficiency), dominant-negative effects from missense variants, and structural rearrangements. Repetitive elements such as Alu repeats can nucleate non-allelic homologous recombination (NAHR), creating recurrent structural variants, but how these mechanisms distribute across EDS genes, and whether they track with tissue-specific expression or gene architecture, has remained unclear.

Results & Discussion

Haploinsufficiency is the dominant mechanism underlying cEDS caused by COL5A1 and COL5A2. Across all ClinVar pathogenic entries specifically annotated to cEDS, 93.1% (27/29) of variants in these two genes are either structural variants or loss-of-function point mutations, significantly exceeding a 75% benchmark (binomial test, $p = 0.0133$), with missense changes comprising only 6.9% (2/29) of events [r37]. Gene-resolved analyses show that COL5A1 harbors a balanced distribution of structural and loss-of-function point variants (10 each; 90.9% haploinsufficiency), whereas all COL5A2 pathogenic variants in this dataset are loss-of-function point mutations (7/7; 100% haploinsufficiency), reinforcing that reduced type V collagen dosage, rather than dominant-negative structural perturbations, is the principal pathogenic axis in cEDS [r37].

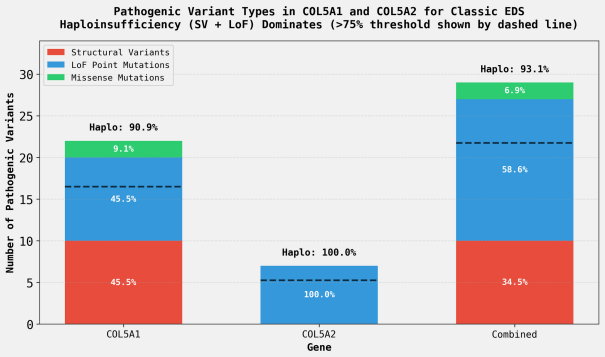

Figure 1: Haploinsufficiency is the dominant pathogenic mechanism for classic EDS variants in COL5A1 and COL5A2. The chart displays the distribution of pathogenic variants for each gene individually and combined, categorized as structural variants, loss-of-function (LoF) point mutations, and missense mutations. Variants causing haploinsufficiency (structural and LoF) far exceed the 75% significance threshold (dashed line) for both genes, reinforcing reduced gene dosage as the principal pathogenic axis. (Source: [r37])

The architecture–mechanism link is strongest for COL5A1, where structural variant breakpoints cluster near intronic inverted Alu pairs—canonical substrates for NAHR. Although both COL5 genes harbor inverted Alu pairs absent from COL1A1, COL1A2, and COL3A1 (inverted-pair densities: COL5A1 102.78/Mb; COL5A2 191.44/Mb; others 0.00/Mb), only COL5A1 shows significant enrichment of breakpoints at these motifs (1.77-fold enrichment; 99 observed vs 55.93 expected; $Z = 6.31$; $p < 0.0001$), whereas COL5A2 shows depletion (0.78-fold; 42 observed vs 54.01 expected; $Z = -2.68$; $p = 0.996$) [r91, r96]. This divergence indicates that COL5A1 SVs are mechanistically coupled to Alu-mediated NAHR, while COL5A2 SVs arise predominantly through non–Alu-mediated processes despite a higher density of inverted Alu pairs, pointing to gene-specific constraints beyond repeat content alone [r91, r96].

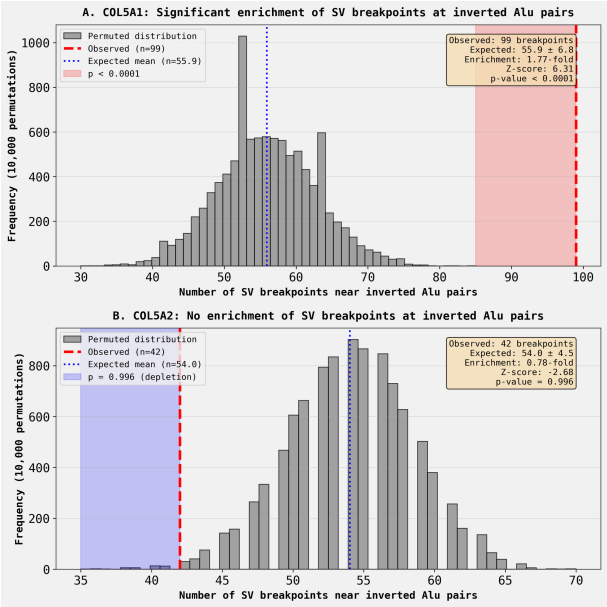

Figure 2: Structural variant breakpoints are significantly enriched at inverted Alu pairs in COL5A1 but not in COL5A2. The histograms show null distributions from 10,000 permutations for the number of breakpoints near inverted Alu pairs, compared to the observed number (red dashed line), for (A) COL5A1 and (B) COL5A2. The significant enrichment in COL5A1 ($p < 0.0001$) indicates that Alu-mediated non-allelic homologous recombination is a key mechanism for structural variation in this gene, a pattern not observed in COL5A2. (Source: [r96])

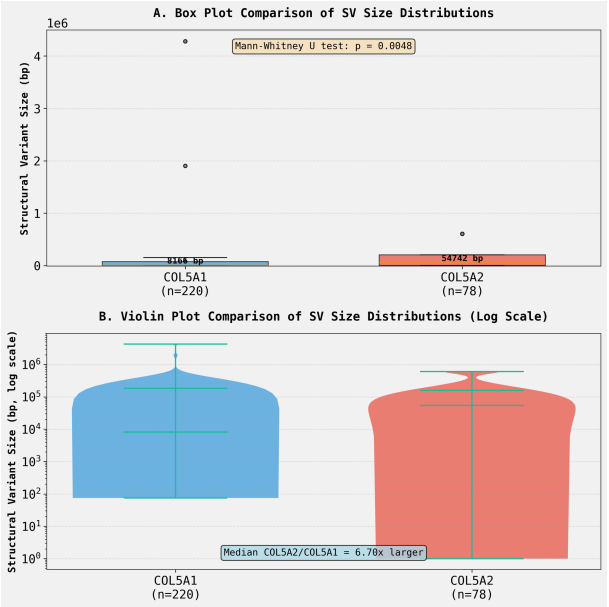

Figure 3: Structural variants in COL5A2 are significantly larger than those in COL5A1. (A) Box plots on a linear scale and (B) violin plots on a log scale compare the size distributions of structural variants (SVs) for COL5A1 (n=220) and COL5A2 (n=78). The median SV size in COL5A2 is 6.70-fold larger than in COL5A1 (Mann-Whitney U test, $p = 0.0048$), reflecting the divergent mutational mechanisms driving structural variation in these genes. (Source: [r100])

Consistent with these mechanistic differences, structural variant size distributions differ markedly between the two genes. COL5A1 SVs are significantly smaller than COL5A2 SVs (median 8,166 bp vs 54,742 bp; Mann–Whitney $U = 6{,}738.0$; $p = 0.0048$; rank-biserial $r = 0.215$), a pattern compatible with localized NAHR in COL5A1 and larger, potentially non-homologous events in COL5A2 [r100]. Distributional features reinforce this separation: COL5A1 shows a tight lower quartile ($Q1 = 253$ bp) with extreme outliers, whereas COL5A2 variants are consistently larger ($Q1 = 2{,}547$ bp; $Q3 = 205{,}114$ bp), underscoring distinct mutational regimes acting on paralogous type V collagen genes [r100].

Extending beyond type V collagen, the EDS gene set partitions non-randomly by mutation modality. Among 39 EDS-associated genes, 12 are affected by structural variants and 19 harbor pathogenic point mutations, with a statistically significant overlap (10 genes affected by both; odds ratio = 10.0; Fisher's exact $p = 0.0057$), while two genes (ADAMTSL2, B3GALT6) are exclusively SV-affected and nine are exclusively affected by pathogenic point mutations (AEBP1, C1R, C1S, COL1A1, PRDM5, SMAD3, TGFBR1, TGFBR2, ZNF469) [r5]. This partitioning suggests that gene-level attributes—dosage sensitivity, genomic architecture, and functional constraints—shape whether pathogenicity manifests via structural rearrangement, sequence-level loss-of-function, or both [r5].

A gene–tissue axis further modulates mutational vulnerability. Using a per-gene structural-variant proportion (SV proportion = SV count / [SV count + pathogenic point-variant count]), genes with connective tissue/fibroblast-focused expression and function show a markedly higher median SV proportion (0.886) than other genes (0.0), with high values in COL5A1 (0.898), COL5A2 (0.886), COL3A1 (0.914), and B3GALT6 (1.0), and similarly elevated SV proportion in ADAMTSL2 (1.0) despite limited tissue-enrichment evidence in the provided sources [r105]. This pattern links tissue-enriched

ECM biology to a greater contribution of SV-mediated pathogenicity, aligning with the core role of fibroblast-driven collagen networks in EDS and refining gene-level predictions of mutational mechanisms relevant for diagnosis and assay design [r105].

Trajectory Sources

Trajectory r5: The set of genes impacted by large structural variants overlaps significantly with genes affected by pathogenic point mutations, with 10 of 12 SV-affected genes (83.3%) also harboring pathogenic point mutations, though 2 genes (ADAMTSL2, B3GALT6) are exclusively affected by SVs and 9 genes are exclu...

Trajectory r37: For pathogenic variants causing classic EDS, haploinsufficiency mechanisms (structural variants and loss-of-function point mutations) account for 93.1% (27/29) of all pathogenic events in COL5A1 and COL5A2 combined, significantly exceeding the 75% threshold (p=0.0133), while missense mutations repre...

Trajectory r91: COL5A1 does not have significantly higher density of inverted Alu pairs compared to other fibrillar collagen genes (Mann-Whitney U test, p = 0.6926), but both COL5 genes (COL5A1: 102.78 pairs/Mb; COL5A2: 191.44 pairs/Mb) possess inverted Alu pairs that are completely absent in COL1A1, COL1A2, and CO...

Trajectory r96: The breakpoints of pathogenic structural variants in COL5A1 show highly significant enrichment at inverted Alu repeat pairs (1.77-fold, $p < 0.0001$), but this enrichment is not observed in COL5A2 (0.78-fold, p = 0.996).

Trajectory r100:
Analysis Results: COL5A1 vs COL5A2 Structural Variant Size Distributions
The hypothesis that structural variant (SV) size distributions differ significantly between COL5A1 and COL5A2 is **SUPPORTED** by the data.
Quantitative Findings:
Sample Sizes: - COL5A1: n=220 SVs overlapping the...

Trajectory r105: Within the limits of the provided sources, EDS genes with connective-tissue/fibroblast-focused expression and function show markedly higher structural-variant (SV) proportions than genes without such evidence, supporting the hypothesis.

Loss-of-function dominance, domain-level paradoxes, and variant classification in TNXB

Summary

Across ClinVar- and gnomAD-derived datasets, TNXB pathogenicity is overwhelmingly driven by predicted loss-of-function (LoF) variants, while domain-resolved analyses reveal a striking mosaic: the N terminus is missense-depleted and conserved without clinical missense enrichment, whereas the conserved C-terminal fibrinogen-like (FBG) domain is paradoxically tolerant to both missense and LoF variation in the general population. A stratified interpretation framework reaches near-perfect specificity for TNXB but is sensitivity-limited by incomplete annotations and the disproportionate complexity of indels.

Background

Interpreting genetic variation in Ehlers-Danlos syndromes (EDS) requires integrating clinical classifications, population constraint, and evolutionary conservation at gene and domain levels. TNXB encodes tenascin-X, a large extracellular matrix protein with repeated domains, positioning it for heterogeneous mutational mechanisms and posing nontrivial annotation challenges. High-quality interpretation depends on reconciling clinical pathogenicity patterns with population-level tolerance and domain conservation, and on mitigating annotation gaps that disproportionately affect complex variants.

Results & Discussion

The dominant signal is that clinically pathogenic TNXB variants are overwhelmingly predicted loss-of-function. In a deduplicated set of 134 unique variants with clear clinical assertions, 89.2% (33/37) of pathogenic/likely pathogenic variants were LoF compared to 2.1% (2/97) among benign/likely benign, yielding an odds ratio of 391.88 (Fisher's exact p = 1.49×10^{-24}) [r102]. Complementing this, domain-focused curation found no significant enrichment of pathogenic missense variants by protein domain, constrained by only 3 pathogenic versus 50 benign missense variants (Fisher's exact p = 0.111) [r97]. Within a stratified classification framework, a simple location-based LoF rule achieved perfect performance in the evaluated subset (all LoF pathogenic and co-localized in a predefined high-risk region), underscoring the high positive predictive value of LoF for TNXB pathogenicity in practice [r33].

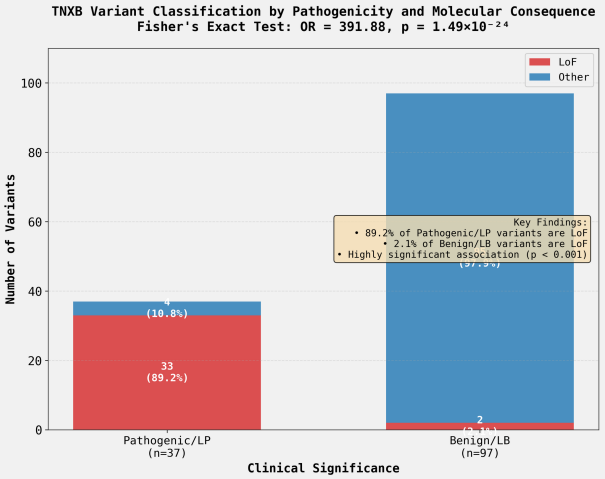

Figure 4: Loss-of-function variants are strongly associated with TNXB pathogenicity. This stacked bar chart displays the distribution of predicted loss-of-function (LoF) and other variants within two clinical significance categories: Pathogenic/Likely Pathogenic (P/LP; n=37) and Benign/Likely Benign (B/LB; n=97). LoF variants comprise 89.2% of the P/LP group but only 2.1% of the B/LB group, demonstrating that LoF is the predominant pathogenic mechanism for TNXB (Fisher's exact test, OR = 391.88, p = 1.49×10^{-24}). (Source: [r102])

Regional constraint analyses reframe the N-terminal domain (amino acids 1–303) as functionally constrained at the population level. Using gnomAD v4 missense density as a proxy for constraint, the N terminus showed a 13.7% lower missense variant density than the fibronectin type-III region (rate ratio 0.863; χ^2 = 9.68; p = 0.0019), with an observed/expected of 0.877 versus 1.016, respectively [r47]. Evolutionary conservation converges on the same signal: within the N-terminal domain, a variant-depleted "cold spot" (AA 61–90) exhibited markedly higher conservation (mean PhyloP 2.79) relative to the remainder of the N terminus (mean 1.17; Mann-Whitney p = 4.9×10^{-6}; Cohen's d = 0.93). At the scale of non-

overlapping 30-AA windows, conservation and missense density did not significantly correlate (Spearman's $\rho = 0.38$, $p = 0.25$), suggesting a mosaic of localized constraint rather than a smooth gradient across the N terminus [r79]. Together, these data indicate that important functional elements reside in the N-terminal region, even though gene-level metrics only suggest moderate missense constraint.

Notably, this population/evolutionary constraint in the N terminus does not translate into a detectable concentration of clinically pathogenic missense variants in current datasets. Testing AA 61–90 versus the remainder of the N-terminal domain (AA 1–60 and 91–303) found zero pathogenic/likely pathogenic missense variants in the cold spot (0/26) and one outside (1/76), with Fisher's exact $p = 1.0$; most cold-spot missense submissions were VUS, consistent with underpowered clinical sampling or unresolved effects rather than clear benignity [r85]. At the broader domain scale, pathogenic missense variants were not significantly enriched in any domain (Fisher's exact $p = 0.111$), but the analysis was underpowered by only three pathogenic missense calls, reinforcing that pathogenicity in TNXB is predominantly not missense-mediated in current clinical data [r97]. These results reconcile a functionally constrained N terminus in population/evolutionary analyses with a sparse clinical burden of missense pathogenicity, highlighting ascertainment limits and the need for functional follow-up of VUS in constrained segments [r47, r79, r85].

The C-terminal fibrinogen-like domain (FBG; AA 4070–4242) presents a paradoxical profile: it is more conserved than the rest of TNXB (mean PhyloP 1.82 vs 0.95; $t = 6.69$; $p = 2.5 \times 10^{-11}$), yet in gnomAD v4 it is enriched for both missense (observed/expected 1.46; $p = 6.9 \times 10^{-10}$) and predicted LoF variants (observed/expected 7.42; $p < 2.2 \times 10^{-16}$), while the remainder of the protein shows moderate missense constraint and strong LoF depletion (observed/expected 0.86 and 0.30, respectively) [r103]. In contrast, ClinVar contained no missense submissions in the FBG domain in the curated set, emphasizing a gap between population variation and clinical ascertainment in this region [r97]. This conserved-yet-tolerant FBG signature suggests that C-terminal truncation or disruption may be compatible with human viability and largely escapes clinical capture, whereas earlier truncations are under stronger negative selection, consistent with the skewed LoF pathogenicity signal outside the FBG region [r97, r103].

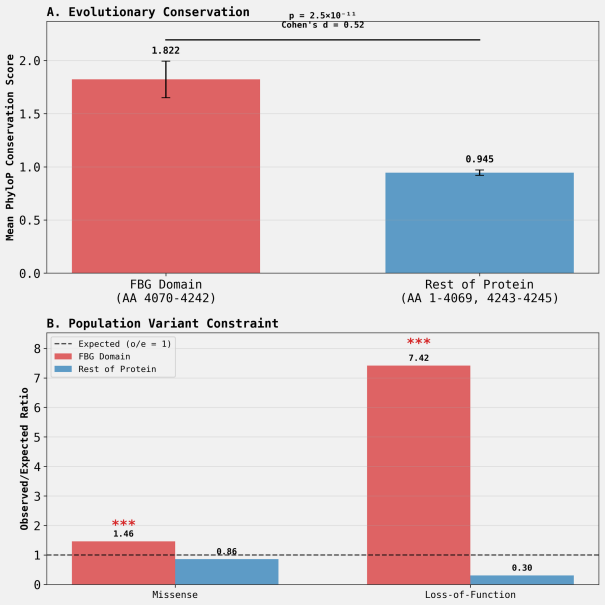

Figure 5: The TNXB fibrinogen-like (FBG) domain is evolutionarily conserved but paradoxically tolerant to population-level variation. (A) The mean PhyloP conservation score is significantly higher in the FBG domain compared to the rest of the protein. (B) In contrast, the observed/expected ratios for both missense and loss-of-function variants are significantly elevated in the FBG domain, whereas the rest of the protein is constrained. This highlights the paradoxical tolerance of the FBG domain to both missense and loss-of-function variation at the population level despite its high evolutionary conservation. (Source: [r103])

Finally, a stratified interpretation framework for TNXB achieved very high specificity (98.9–100%) but modest sensitivity (47.8–52.2%), with perfect performance on evaluated LoF variants but limited reach on missense and complex variants due to missing allele frequency and in silico scores (only 34.8% of pathogenic variants had CADD), and a large fraction of false negatives being "unknown type" changes without HGVS protein notation [r33]. Independent auditing of ClinVar annotations showed that missing risk-allele fields were ~203-fold more likely for indels than for SNPs (odds ratio 202.67; 95% CI 42.42–968.18; $p = 8.85 \times 10^{-8}$), directly implicating indel representation

as a systematic failure mode for TNXB that depresses classifier sensitivity and biases apparent variant spectra [r30]. Together, these findings argue for TNXB-specific interpretation rules that prioritize predicted LoF, de-emphasize generic missense heuristics in the absence of domain-resolved evidence, and explicitly accommodate indel-centric data gaps when assessing pathogenicity [r30, r33, r97, r102, r103].

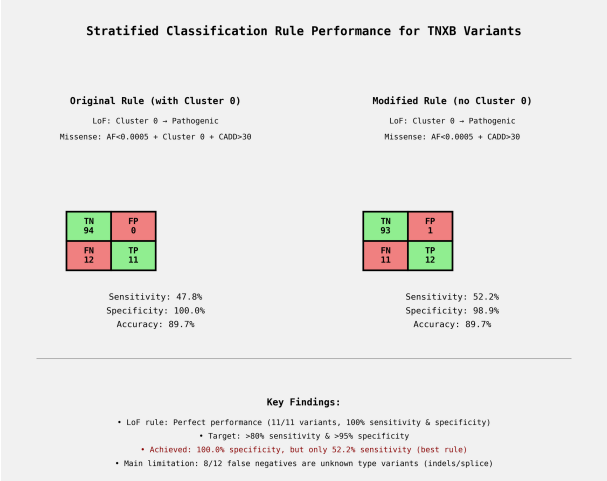

Figure 6: Performance of a stratified classification framework for TNXB variants demonstrates high specificity but limited sensitivity. Confusion matrices show the classification performance for (A) an original rule set and (B) a modified rule set with relaxed missense variant criteria. While the modified rule slightly improves sensitivity to 52.2%, overall sensitivity is constrained by the framework's difficulty in classifying indel and splice variants, which comprise the majority of false negatives. (Source: [r33])

Trajectory Sources

Trajectory r30: TNXB variants with missing $\text{Risk}_{\text{Allele}}$ annotations are significantly more likely to be insertions or deletions (indels) than single nucleotide polymorphisms (SNPs), with an odds ratio of 202.67 (95% CI: 42.42-968.18, $p < 0.001$).

Trajectory r33:
Final Answer
The stratified classification rule for TNXB variants achieved **high specificity (98.9-100%) but failed to meet the target sensitivity of >80%**, reaching only **52.2% sensitivity** at best.
Performance Results
**Original Rule (LoF: Cluster 0; Missense: AF<0.0005 + Cluster 0 ...

Trajectory r47: The TNXB N-terminal region (AA 1-303) exhibits significantly HIGHER functional constraint compared to the Fibronectin type-III repeat domains, with 13.7% lower missense variant density ($p = 0.0019$), directly contradicting the hypothesis that it would show lower constraint.

Trajectory r79: The variant-depleted "cold spot" (AA 61-90) in the TNXB N-terminal domain shows significantly higher evolutionary conservation (mean PhyloP = 2.79) compared to the rest of the N-terminal domain (mean PhyloP = 1.17; Mann-Whitney $p = 4.9\times10^{-6}$, Cohen's d = 0.93), providing strong independent evidence o...

Trajectory r85: Pathogenic/likely pathogenic missense variants in TNXB are NOT significantly enriched within the functionally constrained "cold spot" (AA 61-90) compared to the rest of the N-terminal domain (Fisher's exact test: $p = 1.0$, odds ratio = ∞).

Trajectory r97: Pathogenic missense variants in TNXB are not significantly enriched in specific protein domains compared to benign missense variants (Fisher's exact test: $p = 0.111$), though statistical power is severely limited by only 3 pathogenic and 50 benign missense variants identified.

Trajectory r102: Pathogenic variants in TNXB are significantly more likely to be

predicted loss-of-function (LoF) mutations compared to benign variants, with 89.2% of pathogenic/likely pathogenic variants being LoF versus only 2.1% of benign/likely benign variants (Fisher's exact test: OR = 391.88, p = 1.49×10^{-24}).

Trajectory r103: The TNXB FBG domain (AA 4070-4242) exhibits a paradoxical constraint profile: it is significantly more evolutionarily conserved than the rest of the protein (mean PhyloP 1.82 vs 0.95, p = 2.5×10^{-11}) but shows dramatic enrichment of both missense variants (o/e = 1.46, p = 6.9×10^{-10}) and loss-of-funct...

Noncanonical missense architecture in COL5A1 and the absence of genotype-phenotype discriminants in COL1A2 EDS

Summary

Across fibrillar collagens, COL5A1 displays a noncanonical missense architecture: pathogenic missense variants are not dominated by triple-helix glycine substitutions and instead converge on the C-terminal NC1 domain, particularly on cysteine residues. In contrast, multiple independent analyses fail to find simple genotype–phenotype discriminants that separate cardiac-valvular from classic EDS for COL1A2, arguing for more complex, context-dependent determinants of phenotype.

Background

Fibrillar collagens share a conserved Gly-X-Y triple-helix in which glycine substitutions typically drive dominant-negative pathogenesis, yielding strong genotype–phenotype regularities in several collagenopathies. Type V collagen (COL5A1/COL5A2) is a minor fibrillar component that nucleates fibrillogenesis and shapes fibril morphology, while type I collagen (COL1A1/COL1A2) provides tensile strength in connective tissues. The non-collagenous NC1 C-propeptide directs chain selection and trimerization through disulfide bonding; perturbations at this site can alter assembly, folding, and secretion. Understanding where and how pathogenic variants act across collagen domains—and whether simple mutation-level features predict clinical subtype—remains central to explaining Ehlers–Danlos syndrome (EDS) heterogeneity.

Results & Discussion

COL5A1 departs sharply from the canonical glycine-substitution paradigm that dominates other fibrillar collagens. A cross-gene comparison of pathogenic missense variants shows that only 32.4% of COL5A1 missense variants are glycine substitutions within its triple helix (12/37), whereas 92.8% are glycine substitutions for COL1A1, COL1A2, COL3A1, and COL5A2 combined (1,440/1,552; Fisher's exact $p = 2.26 \times 10^{-19}$, odds ratio = 0.0373), with each of those genes individually showing 85–97% glycine substitutions in the helix [r38]. A second, independent analysis using alternative triple-helix boundaries for COL5A1 (residues 559–1570) still finds that only about half of pathogenic missense variants are glycine substitutions within the helix (18/37 = 48.6%), markedly lower than the other fibrillar collagens, underscoring a genuine shift in COL5A1's missense architecture despite boundary differences across analyses [r18, r38].

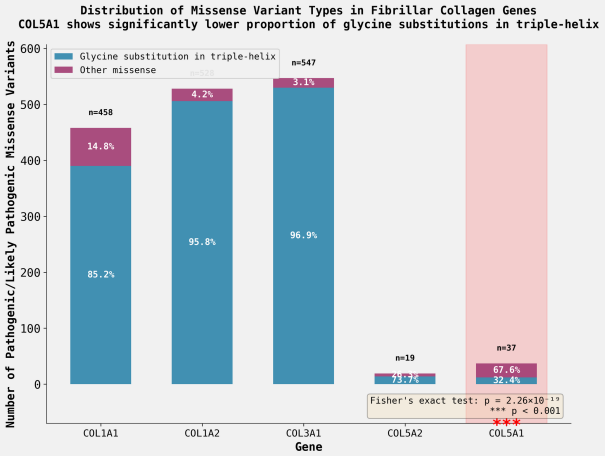

Figure 7: COL5A1 displays a noncanonical pathogenic missense architecture compared to other fibrillar collagen genes. This stacked bar plot shows the distribution of pathogenic missense variants, distinguishing between glycine substitutions in the triple-helix and other missense types for five collagen genes. While glycine substitutions are the predominant pathogenic variant type in COL1A1, COL1A2, COL3A1, and COL5A2 (>85%), they account for only 32.4% of variants in COL5A1, highlighting a statistically significant departure from the canonical mutational mechanism. (Source: [r38])

Within COL5A1, convergent evidence highlights the NC1 domain as a focal point for non-glycine missense pathogenicity and, specifically, cysteine disruption. In a domain-level test contrasting non-glycine pathogenic missense variants in the mature protein (n=13) against benign missense variants (n=193), the NC1 domain shows nominal enrichment (5/13 vs 22/193; odds ratio = 2.82; uncorrected p =

0.016; Bonferroni-corrected p = 0.082 across five domains), with a strong signal for cysteine substitutions: 4/5 pathogenic NC1 variants substitute cysteine versus 0/22 benign NC1 variants (p = 0.0003) [r18]. A complementary tally using a different boundary scheme similarly places nearly half of COL5A1 "other missense" in the C-terminal NC1 region (12/25 = 48.0%), reinforcing the NC1 focus under distinct classification choices [r38]. Extending this comparison across genes, pathogenic NC1 missense variants in COL1A2 and COL5A1 predominantly substitute cysteines (77.8% and 83.3%, respectively), with no significant difference in enrichment between the genes (odds ratio = 0.70; p = 1.0000), indicating that NC1 cysteines are a shared vulnerability across fibrillar collagens [r75].

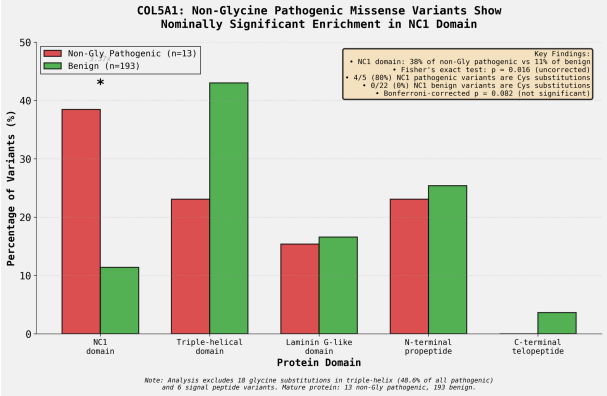

Figure 8: Non-glycine pathogenic missense variants in COL5A1 are enriched in the C-terminal NC1 domain. The bar chart shows the percentage distribution of non-glycine pathogenic (n=13) and benign (n=193) missense variants across five protein domains. This pattern, with a nominally significant enrichment in the NC1 domain (*p = 0.016, uncorrected Fisher's exact test), identifies this region as a focal point for non-canonical missense pathogenicity in COL5A1. (Source: [r18])

The mechanistic basis for this convergence is consistent with established NC1 biochemistry: conserved cysteine networks mediate interstrand disulfide bonds that "lock in" trimer assembly; loss of a single cysteine can abolish disulfide-linked trimerization, destabilize folding, and reduce secretion, effects demonstrated in collagen I models and expected to generalize to type V [r46, dichiara2018a, dichiara2018a, dichiara2018]. In line with this, a COL5A1 NC1 substitution such as p.Cys1639Ser would be predicted to compromise the disulfide "lock," impairing assembly and secretion, whereas the effect of a triple-helix variant like p.Glu1292Lys could not be resolved from the available data [r46, dichiara2018a, dichiara2018a, dichiara2018]. Together, these data support a model in which a sizable fraction of COL5A1 pathogenic missense variants act outside the helix by perturbing NC1 disulfide chemistry, distinguishing COL5A1 from other fibrillar collagens where helix glycine substitutions predominate [r18, r38, r75].

In contrast, multiple orthogonal tests fail to identify simple genotype–phenotype discriminants separating cardiac-valvular EDS from classic EDS in COL1A2. The proportion of loss-of-function variants does not differ significantly between phenotypes (cardiac-valvular 53.8% vs classic 41.2%; Fisher's exact odds ratio = 1.67; p = 0.7131) [r23]. Among glycine substitutions, positional distributions along the triple helix are indistinguishable by both Kolmogorov–Smirnov (D = 0.378; p = 0.626) and Mann–Whitney U tests (U = 27.0; p = 0.606), with no enrichment in the N- vs C-terminal halves across a range of cutoffs [r43]. A focused test near the C-terminal KGHN cross-linking motif (AA 920–940) also shows no enrichment of cardiac-valvular missense variants (0/4) versus classic EDS (1/8; p = 1.0000) [r57]. Finally, variant deleteriousness measured by CADD scores is indistinguishable (median 27.40 vs 28.10; Mann–Whitney p = 0.9322), arguing that general missense severity does not explain phenotype [r110]. While small sample sizes limit power, these consistent null results across mutation class, position, motif proximity, and aggregate pathogenicity suggest that phenotype likely reflects more complex determinants not captured by these coarse features [r23, r43, r57, r110].

These findings refine the mechanistic map of EDS. For COL5A1, the noncanonical missense architecture and NC1 cysteine convergence point to conserved disulfide-dependent assembly defects and motivate structural modeling plus biochemical validation of specific cysteine disruptions [r18, r46, r75]. For COL1A2, the absence of simple discriminants between cardiac-valvular and classic EDS, despite extensive hypothesis testing, shifts emphasis to local structural contexts, post-translational handling, and genetic modifiers that could dif-

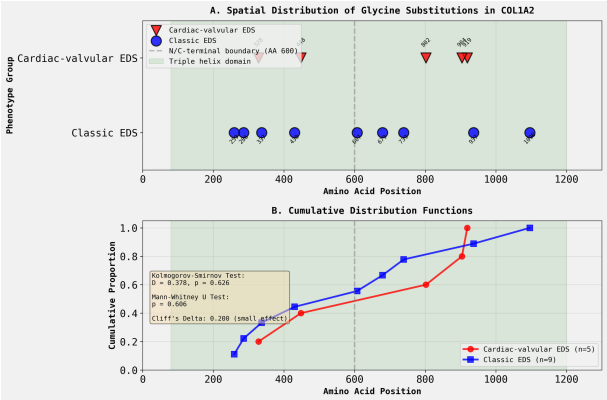

Figure 9: The spatial distribution of pathogenic glycine substitutions in COL1A2 does not distinguish between cardiac-valvular and classic EDS phenotypes. (A) Locations of variants causing cardiac-valvular EDS (red triangles) and classic EDS (blue circles) are plotted along the amino acid sequence. (B) Cumulative distribution functions of variant positions for the two groups show no statistically significant difference (Kolmogorov-Smirnov p = 0.626). These findings suggest that variant location is not a simple genotype-phenotype discriminant for glycine substitutions in COL1A2. (Source: [r43])

ferentially engage cardiovascular tissues [r23, r43, r57, r110]. Together, the data support a model in which NC1 domain chemistry is a shared pathogenic axis across fibrillar collagens, while phenotype specification in COL1A2-related EDS likely emerges from finer-scale structural or genetic interactions beyond mutation class, position, or generic pathogenicity predictions [r18, r23, r38, r43, r46, r57, r75, r110].

Trajectory Sources

Trajectory r18: Non-glycine pathogenic missense variants in COL5A1 show nominal enrichment in the NC1 domain (3.37-fold, uncorrected p=0.016), but this does not reach statistical significance after Bonferroni correction for multiple testing (corrected p=0.082).

Trajectory r23: The mutational signature for cardiac-valvular EDS in COL1A2 (53.8% LoF) is not significantly different from the signature for classic EDS (41.2% LoF), as determined by Fisher's exact test (OR=1.67, p=0.7131).

Trajectory r38:
ANSWER TO RESEARCH HYPOTHESIS
The hypothesis is **CONFIRMED**. COL5A1 shows a significantly different missense variant profile compared to the other four fibrillar collagen genes (COL1A1, COL1A2, COL3A1, and COL5A2).
Key Findings
COL5A1 Missense Variant Distribution: - Glycine substitut...

Trajectory r43: The hypothesis that pathogenic glycine substitutions causing cardiac-valvular EDS cluster in a specific sub-region of the COL1A2 triple-helical domain while classic EDS substitutions are distributed differently is not supported by the data.

Trajectory r46: The hypothesis is supported for p.Cys1639Ser—loss of a conserved NC1 cysteine is expected to destabilize disulfide "lock-in," impair folding/assembly, and often reduce secretion (haploinsufficiency-like), but there is insufficient information in the provided context to assess p.Glu1292Lys in the tri...

Trajectory r57: Pathogenic COL1A2 missense variants causing cardiac-valvular EDS are NOT spatially enriched near the C-terminal KGHN cross-linking motif (AA 920-940) compared to classic EDS missense variants (Fisher's exact test, p = 1.0000).

Trajectory r75: Pathogenic missense variants in the NC1 domains of both COL1A2 and COL5A1 predominantly affect cysteine residues (77.8% and 83.3%, respectively), with no significant difference in enrichment between the genes (Fisher's exact test: OR=0.700, p=1.0000), re-

futing the hypothesis that COL1A2 NC1 preferen...

Trajectory r110: CADD scores do not differ significantly between pathogenic missense variants causing cardiac-valvular EDS (median = 27.40) and classic EDS (median = 28.10) in COL1A2, with Mann-Whitney U test p-value = 0.9322, rejecting the hypothesis.