ByteWorx

e: jspringer@byteworx.ai | p: +1 (775) 815-8803 | Reno, NV

Research Partner Brief

Structural Patterns in ProteinGym: ClinVar Indel Cohort

Prepared for	Research partners, clinical genetics collaborators, and external reviewers
Prepared by	ByteWorx, with analysis support
Research lead	John Springer
Date/scope	May 6, 2026; partner-facing interpretation of joined results, bridge evolution, edge persistence, and validation priorities

Purpose

This brief summarizes a ByteWorx structural analysis of clinically labeled indels from the ProteinGym ClinVar cohort. The goal is to help partners decide whether the cohort behaves like a single diffuse pool of variants or contains reproducible neighborhoods and transition zones that warrant targeted biological review.

The partner-facing conclusion is measured but positive: the current outputs support a structured prioritization map. The analysis identifies broad regimes, compact micro-pockets, and ranked bridge corridors that can guide follow-up. The results should not yet be presented as a validated clinical tool or as proof of a single biophysical mechanism.

Partner takeaway: The strongest use of this work is to focus expert review on a smaller set of clusters, edges, and row-level variants. The leading candidates are the PKD1-rich 20-15 pocket, the 17-19 / 19-18 boundary chain, and the 11-13 connective/vascular interface.

Evidence base reviewed

The analysis uses the ProteinGym ClinVar Indel summary, the joined results workbook, bridge_evolution.json, bridge_evolution_edges.csv, and the UCI grid stability memo as the formatting reference. The cohort contains 1,760 rows. Of these, 1,471 rows are assigned to 26 non-noise clusters, while 289 rows fall into cluster -1.

All rows in the reviewed workbook are labeled Pathogenic. This means the analysis is not a benign-versus-pathogenic classifier. It is a structure-finding and prioritization exercise within a clinically labeled cohort of pathogenic indels. That distinction is important for external communication.

Cohort measure	Observed value	Partner interpretation
Total rows	1,760	Large enough to detect repeated structural neighborhoods, but still small enough for targeted row-level review.
Assigned rows	1,471 across 26 clusters	Most variants participate in recoverable local organization under the current method.
Unresolved rows	289 in cluster -1	Not discarded; should be treated as a reservoir for rare or heterogeneous substructure.
Clinical labels	All rows labeled Pathogenic	Supports within-pathogenic structure, not diagnostic separation from benign controls.
Review strength	Mixed ClinVar review statuses	Adds value for prioritization, but downstream review should account for evidence strength.

Main finding

The cohort is not behaving as one homogeneous set of indels. Under the current feature and clustering procedure, the variants organize into broad basins, smaller micro-regimes, and transition corridors. The signal is strongest when described as a computational structure suitable for prioritization, rather than as a completed clinical interpretation.

The largest basin, cluster 0, contains 624 rows and functions as a broad mid-length reference regime. Cluster 5 contains 421 rows and shifts toward longer proteins. Cluster 10 contains 44 rows with a higher median protein length. Together, these basins support the idea that protein length and architecture are shaping the recovered landscape.

The clearest local pocket is PKD1. Cluster 20 contains 8 rows, 7 of which are PKD1. Clusters 15 and 20 together contain 17 rows, 10 of which are PKD1. That concentration is the strongest gene-recurrent micro-regime in the current run.

Regime / pocket	Observed pattern	Partner-facing read
Cluster 0	624 rows; median protein length about 514 aa	Broad mid-length reference basin. Useful as a comparator, not a single disease class.
Cluster 5	421 rows; median protein length about 1,394 aa	Long-protein regime. A priority for domain and architecture overlay.
Cluster 10	44 rows; median protein length about 1,955 aa	Smaller long-protein regime that should be compared with cluster 5.
Cluster 20	8 rows; 7 PKD1	Highest-interest micro-pocket. Strong local recurrence and clear follow-up value.
Cluster -1	289 unresolved rows	Potential hidden structure. Should be recursively peeled rather than ignored.

Why the bridge and edge-evolution outputs matter

The bridge layer is the part of the analysis that makes the result useful to partners. A cluster view identifies where rows collect. A bridge view identifies where regimes touch and which rows are closest to transition boundaries. Those boundary rows are the most efficient targets for structural annotation and expert review.

In the ridge-selected setting with eps = 0.1886, the active bridge scaffold has 26 nodes, 25 edges, one connected component, and no cycles. That matches the final 26-cluster scaffold and supports the view that the bridge graph is not merely a one-off static table.

The edge evolution also prevents overclaiming. Some edges are tight and persistent; others only serve as long-range connectors or short-lived exploratory edges. For partner work, this ranking is essential because it indicates where investigators should start.

Method note for external review: The sweep includes abrupt zero-edge rows at some eps values. These should be explained in methods or treated as filtering/implementation artifacts before the result is described as fully robust across all resolutions.

Priority transition corridors

The highest-confidence corridors are those where edge persistence, compactness, recurrence, and biological plausibility align. The table below ranks the current bridge evidence in a partner-facing way.

Corridor	Evidence strength	Partner interpretation
20-15 / PKD1	Strongest. 60 observations; span 0.0220; final bridge distance 0.408; low corridor mean 0.011. Key rows: PKD1 V1967del, N1240del, V2251del.	Lead validation target. A compact PKD1 micro-pocket linked to a broader long-protein regime.
17-19 and 19-18	Very strong. Both edges show 84 observations and span 0.0285; compact edge-evolution centroid distances near 0.13-0.14. Key genes include CLRN1, HBB, OAT, SERPINC1, and ATP6AP1.	Promote to primary follow-up tier. This region was underweighted in the earlier summary.
11-13	Strong. 36 observations; span 0.0086; low corridor mean 0.016; final bridge distance 0.536. Key genes include FBN1, VWF, and MYO7A.	Credible connective/vascular interface. Good candidate for domain, phenotype, and structural comparison.
5-20	Moderate. 19 observations; span 0.0047; final bridge distance 0.620. PKD1 rows help connect the micro-pocket to the long-protein basin.	Supports the broader PKD1 branch, but should be described as weaker than the 20-15 pocket.
15-23	Persistent but long-range. 54 observations; span 0.0208; final bridge distance 2.539; edge-evolution centroid distance 2.482. Key genes include GOT1, EZH2, NPC1.	Interesting connector, not a compact corridor. Requires a biological explanation before the mechanism language.
3-4 and 0-4	Weak in edge evolution. 3-4 and 0-4 show short spans near the selected ridge despite appearing in the final scaffold.	Keep it exploratory. Suitable for review, but not for leading evidence.

Research interpretation for partners

The analysis provides a structured way to choose where to look next. Rather than treating each pathogenic indel as an isolated event, partners can review groups of variants that occupy the same computational neighborhood or lie at the boundary between neighborhoods.

The appropriate scientific claim is that these outputs define candidate structural neighborhoods and transition zones. The output is useful for hypothesis generation, variant prioritization, and study design. It is not yet a substitute for protein-domain analysis, molecular modeling, functional testing, or clinical expert review.

The most important wording discipline is around the word structural. In this brief, structural means structure in the feature/embedding and bridge geometry unless or until orthogonal evidence confirms a protein-domain or biophysical mechanism.

Recommended partner validation path

1. Build an auditable boundary-row panel

For each priority corridor, create a row-level table with row ID, gene, variant, cluster assignment, distance to both centroids, nearest-neighbor support, membership confidence if available, ClinVar review status, protein length, relative position, domain overlap, and structural annotation.

2. Add biological overlays

Overlay the major clusters and edge rows with Pfam or InterPro domains, UniProt features, AlphaFold regions, membrane/scaffold/enzyme annotations, disorder, solvent exposure if available, and phenotype terms. Multiple-testing correction should be included for enrichment claims.

3. Add controls and null models

The current workbook contains pathogenic labels only. Stronger claims require benign or presumed-benign indel controls, gene-blocked shuffles, position shuffles, same-size random clusters, and repeated clustering under nearby settings.

4. Recursively peel the broad compartments

Clusters 0, 5, and -1 are the best candidates for recursive peel analysis. This will test whether additional stable substructure is hidden inside broad basins or unresolved rows.

5. Start with a focused partner review panel

A practical first panel should include PKD1 rows 616, 618, 613, 619, and 620; FBN1/VWF/MYO7A on 11-13; CLRN1/HBB/OAT/SERPINC1/ATP6AP1 across 17-19 and 19-18; and GOT1/EZH2/NPC1 on 15-23 as a long-range connector requiring explanation.

External-use language

Recommended phrasing for partners: ByteWorx identified a structured prioritization landscape within the ProteinGym ClinVar indel cohort. The strongest signals are broad protein-architecture-associated basins and a small set of persistent bridge corridors, led by the PKD1-rich 20-15 pocket. These findings are suitable for targeted biological review and validation, but they should not yet be described as diagnostic performance, clinical utility, or proof of a shared disease mechanism.

Use boundary: This document is intended for research prioritization and partner validation planning. It is not medical advice and is not a clinical decision-support document.

Conclusion

The partner-facing conclusion is that the paper has a real and useful seed. The bridge and edge-evolution outputs make the structural claim stronger than the initial cluster summary alone. They show which parts of the network deserve confidence and which should remain exploratory.

The strongest current result is the PKD1-rich 20-15 pocket, supported by a broader but weaker 5-20 link. The 17-19 / 19-18 boundary chain and the 11-13 interface should also be promoted as high-priority validation targets. The 15-23 edge remains important as a persistent long-range connector, while 3-4 and 0-4 should be retained as secondary hypotheses.

Framed this way, the work is forwardable: it is a credible research prioritization brief that gives partners a concrete path from a large indel cohort to specific clusters, bridges, and rows for expert validation.

Appendix A. Priority review panel

Tier	Rows / genes	Why this belongs in the first partner panel
Lead	616 PKD1 V1967del; 618 PKD1 N1240del; 613 PKD1 V2251del; 619 PKD1 A1234del; 620 PKD1 A1041del	Defines and reinforces the PKD1-rich branch. The 20-15 boundary is the strongest edge in the current interpretation.
Primary	1697 CLRN1; 478 HBB; 315 OAT; 447 SERPINC1; 798 ATP6AP1	The 17-19 and 19-18 edges are very persistent and compact in edge evolution. This region should be elevated.
Primary	142 FBN1; 566 VWF; 297 MYO7A	Credible 11-13 interface with connective/vascular relevance and good bridge support.
Secondary	984 GOT1; 1154 EZH2; 310 NPC1	Persistent 15-23 connector. Important, but not compact; needs domain/structural explanation.
Exploratory	1190 LRAT; 1311 POU3F3; 159 FH; 1739 NANOS1	Interesting final scaffold rows, but weak edge-evolution support near the selected ridge.

Appendix B. Evidence tiers used in this brief

Tier	How it is assigned	External communication rule
Lead / primary	High persistence, compact geometry, coherent row-level examples, and clear follow-up value.	Use as first-wave partner validation targets.
Moderate	Bridge support exists, but one or more metrics are weaker or the biological interpretation is less direct.	Use as second-wave review targets.
Persistent long-range	Edge persists but spans a larger centroid distance or bridges distant regions.	Describe as a connector, not a compact corridor.
Exploratory	Appears in the final scaffold but has short edge-evolution support or limited recurrence.	Keep in the hypothesis pool; do not lead with it.

Source note

This partner-facing rewrite is based on the uploaded ProteinGym ClinVar Indel summary, the joined Excel outputs, bridge_evolution.json, bridge_evolution_edges.csv, and the uploaded UCI-style memo used as the formatting reference. All interpretations should be reviewed against the source files before external publication or clinical use.