No specific molecular mimicry between bacterial CdtB and human vinculin despite elevated raw sequence similarity: a rigorous, null-controlled, multi-modal immunoinformatics analysis

Leo Van Schaik[1,*]

[1]Independent researcher / Sibo Research
[*]Corresponding author. E-mail: leo@etherian.io ORCID: to add

Abstract

Post-infectious irritable bowel syndrome (IBS) and small-intestinal bacterial overgrowth (SIBO) have been attributed to molecular mimicry between the cytolethal distending toxin B subunit (CdtB) of enteric bacteria and the human cytoskeletal protein vinculin: gastroenteritis is proposed to raise anti-CdtB antibodies that cross-react with vinculin in the enteric nervous system, and circulating anti-CdtB/anti-vinculin antibodies are marketed as IBS biomarkers. The cross-reactive epitope has never been published as a defined sequence, no in-silico CdtB-vs-vinculin mimicry analysis exists, and independent diagnostic replication is weak. We built an open, reproducible, null-controlled immunoinformatics pipeline (CVMP) and tested CdtB–vinculin mimicry with five analyses spanning three modalities: (i) exact pentapeptide identity, (ii) local-alignment similarity, (iii) fold and (iv) fine surface-patch structural superposition, and (v) overlap with experimentally-validated B-cell epitopes from the Immune Epitope Database (IEDB) and with DiscoTope-3.0-predicted conformational epitopes. Each was assessed against rigorous empirical null models (dipeptide-preserving and composition-preserving shuffles; length-stratified random-fragment root-mean-square-deviation (RMSD) nulls) with Benjamini–Hochberg false-discovery-rate (FDR) correction. A bona-fide mimic (streptococcal M5 protein vs cardiac myosin) and a non-mimic (*E. coli* MalE vs vinculin) served as positive/negative controls for the sequence arm, and a DNase-I-fold structural homolog benchmarked the fold arm. The pipeline recovered the positive control (local-alignment empirical $p = 0.002$, $z = 7.7$) and rejected the negative control ($p = 0.71$); the fold arm separated a genuine structural homolog (CdtB↔DNase I, Foldseek $E = 6 \times 10^{-9}$) from vinculin ($E = 4.7$); a documented epitope mimic (Spike↔thrombopoietin) was likewise recovered ($p = 0.004$), confirming the fold and conformational arms have power. Across 72 non-redundant CdtB variants (family identity 15–89%, median 26%) and human vinculin (isoform P18206-2), vinculin lay in the raw sequence-similarity tail of the human proteome (rank 163/20,416; 27/1,450 among proteins within ±15% length), but the signal was weak and non-specific: BLAST normalization placed vinculin only 3,940th by E-value ($E = 143$), the best CdtB–vinculin ordered-similarity hit was nominal but not family-wide significant ($p = 0.0040$ vs Benjamini–Hochberg threshold $p \leq 0.000694$), and the best CdtB came from non-enteric *Nocardiopsis rhodophaea*. No orthogonal arm corroborated this tail signal: no FDR-significant exact motif or local-alignment hit across the CdtB family, no significant fold or fine-patch structural mimicry, 0 of 153 experimental vinculin B-cell epitopes matched within CdtB, and no overlap between the conformational epitopes the two proteins present. The CdtB–vinculin mimicry hypothesis therefore lacks support for specific ordered-motif, structural, or B-cell-epitope mimicry despite elevated raw sequence similarity. This constrains, but does not disprove, the clinical phenomenon and relocates the burden of proof to direct experiment; we propose a peptide-competition ELISA as the decisive test. The benchmarked pipeline is released open-source.

Author summary

Irritable bowel syndrome (IBS) and small-intestinal bacterial overgrowth (SIBO) are common gut disorders. A popular idea is that after a gut infection the immune system makes antibodies against a bacterial toxin,

CdtB, and that these antibodies then mistakenly attack a human protein, vinculin, because the two "look alike" to the immune system — a process called molecular mimicry. Blood tests built on this idea are already sold to help diagnose IBS, yet the supposed shared piece of protein has never actually been shown. We built an open, reproducible computational pipeline to test, for the first time and across many different bacteria, whether CdtB really resembles vinculin. We checked five analyses across sequence, three-dimensional shape, and antibody target sites, and compared every result against carefully randomized controls so that coincidental matches do not count. To confirm the method works, we showed it detects a famous real case of mimicry (the rheumatic-fever streptococcal protein and human heart muscle) and correctly rejects an unrelated protein. Vinculin did rank higher than most human proteins by raw CdtB sequence similarity, which explains why the hypothesis can look attractive, but that signal was weak after standard correction and was not supported by the shape or antibody-target analyses. This does not by itself overturn the clinical test, but it removes support for the proposed molecular basis and points to a specific laboratory experiment that could settle the question.

Introduction

A widely promoted model of post-infectious IBS/SIBO, originating from Cedars-Sinai (Pimentel and colleagues), holds that acute gastroenteritis induces antibodies against bacterial CdtB — the active, DNase-I-family subunit of cytolethal distending toxin, produced by *Campylobacter jejuni*, pathogenic *E. coli*, *Shigella*, *Salmonella*, *Helicobacter*, *Aggregatibacter* and *Haemophilus* — which then cross-react via molecular mimicry [6] with host vinculin in the interstitial cells of Cajal and myenteric plexus, impairing motility and permitting bacterial overgrowth. Anti-CdtB and anti-vinculin ELISAs are commercialized as IBS biomarkers [1, 2].

Three gaps motivate a rigorous computational test. First, the cross-reactive epitope has never been disclosed: the originating patents reference only an antigenic CdtB peptide and sequence identifiers, not an aligned CdtB–vinculin map. Second, no in-silico CdtB-vs-vinculin mimicry analysis has been published, nor a ranking of cdt-bearing pathogens by predicted vinculin mimicry. Third, independent replication is weak — multiple cohorts and a 2024 systematic review report poor sensitivity or failure to discriminate [3, 4].

A central methodological hazard frames any such test: short-peptide "mimicry" is statistically ubiquitous — no human protein lacks a bacterial pentapeptide motif [5] — so any mimicry claim must beat an explicit empirical null. Conversely, recent structural-proteome screens show that sequence-based mimicry screening misses structurally-encoded mimicry, motivating an explicit structural arm [10]. We therefore built a multimodal, null-controlled, benchmarked pipeline and applied it to CdtB↔vinculin across all cdt-bearing taxa. We emphasize at the outset that this approach probes antigen–antigen resemblance — necessary supporting evidence for the proposed mechanism, but not equivalent to antibody cross-reactivity, which is ultimately an empirical serological property (Discussion).

Results

A consolidated summary of all arms is given in Table 1.

Pipeline calibration

The positive control recovered the established mimic: M5↔cardiac-myosin local-alignment BLOSUM62 score 225, empirical $p = 0.0020$ ($z = 7.7$) vs the composition null, with the aligned span (M5 ≈ 54–419) containing the known cross-reactive region 84–116 [13]. The negative control (MalE↔vinculin) was non-significant ($p = 0.71$). The pipeline detects a true mimic and rejects a non-mimic (Fig 1A; controls summarized in S4 Table).

The fold arm passed an independent structural positive control. Because CdtB adopts the DNase-I fold [7], DNase I is a genuine structural homolog; Foldseek separated CdtB from DNase I at best $E = 6.3 \times 10^{-9}$ (probability 1.000, near-full-length alignment) versus $E = 4.7$ for vinculin — a ∼9-order-of-magnitude margin (Fig 1B), so the fold-arm vinculin negative reflects genuine dissimilarity rather than lack of power. In contrast, the fine surface-patch test (7-residue Cα windows) returned 0 FDR-significant matches even against the

Table 1: **Summary of all mimicry tests.** "Best p" is the smallest nominal empirical p before FDR; no arm yielded an FDR-significant CdtB↔vinculin hit. The two controls validate sensitivity (positive recovered) and specificity (negative rejected).

Level	Null model	n tested	Best p	Outcome
Positive control, seq. (M5↔myosin)	composition shuffle	1	0.0020	**recovered** ($z = 7.7$)
Positive control, fold (CdtB↔DNase I)	Foldseek E-value	1	6×10^{-9}	**recovered**
Positive control, conf. (Spike↔TPO)	composition shuffle	1	0.0035	**recovered** ($z = 4.4$)
Negative control (MalE↔vinculin)	composition shuffle	1	0.71	rejected
(i) Exact pentapeptide	dipeptide-preserving shuffle	72	0.14	negative (0/72)
(ii) Local alignment	composition shuffle	72	≈0.01	negative (0/72 FDR)
Proteome context	SW/BLASTP + composition shuffle	20,416	0.0040	elevated but non-specific
(iii) Fold/domain (Foldseek)	E-value	72	$E = 4.7$	negative (0/46)
(iv) Fine surface-patch	length-stratified RMSD null	72	—	uninformative*
(v) Experimental epitopes	per-epitope composition null	153	0.033	negative (0/153)
(v) Conformational epitopes	composition null	14	0.25	negative (0)

*The fine surface-patch test returns 0 FDR-significant matches even against a true structural homolog (DNase I; best RMSD 0.11 Å), so it lacks discriminative power and is reported as uninformative rather than as evidence (see text).

DNase-I homolog (best RMSD 0.11 Å, indistinguishable from vinculin): short Cα fragments superpose near-perfectly regardless of homology, so this test has no discriminative power and we treat its result as uninformative.

The conformational arm was likewise benchmarked: applying its exact + local-alignment + composition-null test to the documented SARS-CoV-2 Spike↔thrombopoietin epitope mimic (shared TQLPP motif [11]) recovered it at local-alignment $z = 4.4$, empirical $p = 0.0035$ — versus $p = 0.25$ for the best CdtB↔vinculin conformational comparison. The epitope-comparison step therefore has power (the DiscoTope-3 prediction step that defines the peptides remains a separate, acknowledged assumption).

Sequence arm — negative

Across 72 non-redundant CdtB variants, no CdtB↔vinculin region survived FDR (Fig 2; S1 Dataset). The strongest local-alignment hit (from non-enteric *Nocardiopsis rhodophaea*) had nominal $p \approx 0.01$ in the per-variant sequence scan but did not survive Benjamini–Hochberg correction; the best exact-pentapeptide hit had $p \approx 0.14$. The single best CdtB local alignment to vinculin scored 71 — roughly one-third of the positive control (225). The family is broad and well-sampled rather than a narrow clade (Fig 3).

Vinculin's similarity is elevated but non-specific

We then asked the direct proteome-wide question raised by the biomarker hypothesis: where does vinculin rank among all reviewed human proteins for similarity to the CdtB family? We replaced the canonical 1,134-aa UniProt P18206 record with the primary 1,066-aa P18206-2 sequence and, for each of 20,416 human proteins, took the maximum Smith–Waterman local-alignment score over all 72 CdtB variants. Vinculin ranked in the raw-score tail (71.0; rank 163/20,416, 99.2nd percentile) and remained elevated in the ±15%

length-matched cohort (rank 27/1,450, 98.2nd percentile), while score correlated with protein length across the proteome ($r = 0.454$; Fig 6A; S5 Dataset).

This tail position is real, but it is not specific evidence of mimicry. The top raw-scoring human proteins were an incoherent set including treacle, mucins, zinc-finger, ankyrin-repeat, armadillo repeat, cadherin, and centrin proteins rather than a vinculin-centered or cytoskeletal class. Standard BLASTP normalization weakened vinculin's standing: its best bit-score was 24.3 and its best E-value was 143, ranking 3,044th by bit-score and 3,940th by E-value. A 2,000-shuffle composition-null refinement for the best CdtB↔vinculin pair gave empirical $p = 0.0040$ ($z = 3.76$, descriptive only because local-alignment scores follow an extreme-value rather than Gaussian distribution), but the family-wide Benjamini–Hochberg threshold for 72 CdtB variants was $p \leq 0.000694$ (score threshold 79; Fig 6B). The best CdtB variant was A0ABP5E2C7 from non-enteric *Nocardiopsis rhodophaea*, not an enteric pathogen central to the clinical claim.

Family-wide correction is the appropriate test rather than an unfair penalty. The proposed biomarker mechanism is that CdtB broadly, across gastroenteritis-associated bacteria, mimics vinculin; it is not a pre-specified claim about one *Nocardiopsis* sequence. A single nominally significant, non-enteric, uncorroborated tail hit is therefore best interpreted as elevated but non-specific sequence similarity of the Trost/Kanduc type [5], not as evidence for specific molecular mimicry.

Fold/domain structure arm — negative

Foldseek returned 46 short local alignments between CdtB (DNase-I fold) and vinculin (α-helical bundles), none significant (best E-value 4.7, probability 0.000) — as expected for unrelated folds.

Fine surface-patch arm — uninformative

Pairwise Cα superposition of surface patches produced near-perfect short matches (best RMSD 0.11 Å over 7 residues) that are generic secondary-structure fragments, and 0 survived the length-stratified null after FDR. However, the same test returns 0 hits against a *true* DNase-I-fold homolog of CdtB (§Pipeline calibration): at 7-residue Cα windows it cannot distinguish genuine structural relatedness from background. We therefore report this arm as uninformative rather than as evidence either way; the fold arm (above) is the powered structural test.

Conformational epitope arm — negative

DiscoTope-3.0 predicted conformational B-cell epitopes on vinculin (P18206-2; 248 epitope residues) and the CdtB family (46–55 epitope residues each). Restricting the search to the contiguous conformational-epitope regions both proteins present (5 vinculin, 14 CdtB peptides): 0 exact matches, 0 FDR-significant (best $p = 0.25$; Fig 5; S3 Dataset).

Background comparison

A non-pathogen proteome comparison (CdtB vs *B. subtilis*, both vs vinculin) was sensitive to the control length/composition window (best-hit p ranging 0.0005–0.55) and is therefore not a reliable test; we rely on the per-variant shuffle-null with FDR, which is negative throughout.

Experimental-epitope arm — negative

Of the 153 experimentally-validated IEDB vinculin B-cell epitopes, 0 occur (as a whole epitope) within any CdtB variant and 0 show FDR-significant similarity (best nominal $p = 0.033$; S2 Dataset; Fig 4). The four IEDB CdtB epitopes likewise match nothing in vinculin.

Pathogen ranking

Under FDR control no cdt-bearing pathogen's CdtB variant passes the staged filter; the predicted "SIBO-risk ranking" is empty.

Sensitivity and scope

The positive control demonstrates that the sequence arm detects a bona-fide serological mimic: M5↔cardiac myosin is recovered at empirical $p = 0.002$ ($z = 7.7$) and survives FDR, while the unrelated negative control is rejected. For comparison, the best CdtB↔vinculin alignment scored 71 versus 225 for the true mimic; because alignment score scales with alignment length and rests on a single benchmark, we treat this as an *illustrative* sensitivity reference rather than a calibrated limit of detection. Two scope caveats follow. First, the sequence arm (M5↔myosin), the fold arm (CdtB↔DNase-I homolog; Fig 1) and the conformational arm (Spike↔thrombopoietin epitope mimic) are all benchmarked with positive controls and have demonstrated power, so their vinculin negatives are meaningful. The fine surface-patch test, by contrast, fails its own positive control (0 hits even for a true homolog) and is therefore uninformative rather than evidence. Second, the most assumption-light result — that 0 of 153 *experimentally validated* vinculin B-cell epitopes occur within any CdtB variant — depends on no predictor or scoring choice and anchors the negative.

Discussion

Across five analyses spanning three modalities — sequence, structure, and B-cell epitopes — bacterial CdtB shows elevated but non-specific raw sequence similarity to human vinculin and no statistically robust evidence of specific molecular mimicry. The proteome-wide context is the key qualification: vinculin is in the raw-score tail, so the hypothesis is not foolish on its face, but its BLAST-normalized E-value rank is weak, its best ordered sequence hit fails family-wide FDR, and the signal is uncorroborated by exact motifs, fold/domain structure, experimental B-cell epitopes, or predicted conformational epitopes. This pattern is consistent with the weak independent clinical replication of anti-CdtB/anti-vinculin biomarkers and with the principle that short-peptide and short-patch similarity are statistically ubiquitous and mostly coincidental [5].

Strengths and limitations

Strengths: five analyses spanning sequence, structure, and epitope evidence; explicit empirical nulls with FDR at every level; a recovered positive control and a rejected negative control that together calibrate sensitivity and specificity; and a fully open, reproducible pipeline. What would change the conclusion: (i) the cross-reactive epitope could be discontinuous; the conformational-epitope arm (DiscoTope-3) is negative, leaving only a full surface shape-and-electrostatics (MaSIF-style) comparison untested. (ii) Antibody cross-reactivity can arise at the paratope level and need not be visible in antigen–antigen comparison; current antibody–antigen docking (AlphaFold3/Boltz, $< 15\%$ accuracy [15]) cannot adjudicate this. (iii) Post-translational modification, conformational state, or species differences could matter. (iv) Anti-vinculin antibodies occur in other conditions (e.g. systemic sclerosis [14]), so the biomarker need not reflect CdtB-driven mimicry. (v) The sequence (M5↔myosin), fold (CdtB↔DNase-I homolog) and conformational-epitope (Spike↔thrombopoietin) arms are all benchmarked with positive controls, but each control has a defined scope. The CdtB↔DNase-I control validates homologous fold detection, not non-homologous local mimicry. The Spike↔thrombopoietin conformational control was non-blind: the comparison windows were chosen around the known TQLPP motif rather than predicted de novo by DiscoTope-3. Only the fine surface-patch test *fails* its positive control (no discriminative power at 7-residue Cα windows) and is not treated as evidence. The conformational benchmark validates the epitope-comparison step; the DiscoTope-3 step that defines the epitope peptides remains a separate assumption. The robust core of the present negative therefore rests on the sequence, fold, experimental-epitope, and conformational-epitope arms. (vi) The conformational arm uses predicted (DiscoTope-3) epitopes on AlphaFold and experimental structures to define the comparison space; predictor error, prediction-on-prediction uncertainty, and low-pLDDT/disordered vinculin regions could mask a true epitope. (vii) We tested 72 non-redundant CdtB variants spanning cdt-bearing enteric taxa; confirming that the exact CdtB antigen used in the commercial assay and originating patents falls within this set is a useful remaining cross-check. In-silico mimicry is not cross-reactivity: this analysis constrains the mechanism but does not disprove the clinical association.

Our central claims are falsifiable. They would be overturned by a positive peptide-competition ELISA (anti-CdtB binding to vinculin blocked by a defined CdtB peptide), by a validated structural-mimicry control that establishes power for the structural arms *together with* a CdtB↔ vinculin three-dimensional match those

arms currently miss, or by identification of a genuine shared cross-reactive epitope that the scan should have detected but did not.

Contribution beyond the negative

We release an open, benchmarked, null-controlled mimicry pipeline (positive control included), a non-redundant CdtB family with a computed identity matrix, and a vinculin B-cell-epitope map — reusable resources for any host–pathogen mimicry question.

Proposed experimental validation

A peptide-competition ELISA is the decisive test: synthesize the top predicted shared CdtB peptide(s) and assay whether they block anti-CdtB binding to vinculin (mirroring the undisclosed "CdtB blocking peptide"); alanine-scan/truncation to map any minimal epitope; and localize any cross-reactive region on vinculin (head 1–835 vs tail 879–1066), which the field has never reported.

Materials and methods

Data and provenance

Human vinculin isoform P18206-2 (1066 aa; the ubiquitous, gut-relevant form) was the primary antigen; canonical P18206 (metavinculin, 1134 aa) was retained as a secondary ensemble member, with an insert-aware coordinate mapper (insert at canonical 916–983). CdtB sequences were retrieved from UniProt [28] across cdt-bearing taxa (gene + protein name), filtered to full-length (220–330 aa, fragments removed) and clustered to non-redundancy at 90% identity with MMseqs2 [22], yielding 207 records → 117 full-length → 72 non-redundant variants (pairwise identity 14.7–89.1%, median 26%). Experimental structures: CdtB 1SR4/2F1N/2F2F/4K6L; vinculin 1TR2/6FUY/1RKE; AlphaFold DB [24, 25] (v6) models AF-P18206-2, AF-P18206, AF-Q46101. Ground-truth B-cell epitopes were obtained from the IEDB [27] IQ-API (vinculin $n = 153$ positive linear epitopes; CdtB $n = 4$). The human reviewed proteome (UP000005640, 20,416 sequences) and *Bacillus subtilis* (UP000001570) served as backgrounds/controls. All accessions, releases (UniProt 2026_02) and checksums are recorded with the code.

Sequence arm

Exact pentapeptide longest-common-substring ("1D-mimic") and Smith–Waterman [16] local alignment (BLOSUM62 [18]) between each CdtB variant and vinculin.

Proteome-wide ranking

For the proteome-wide context analysis, the reviewed human proteome (UP000005640) was rewritten so the canonical P18206 record carried the primary P18206-2 vinculin sequence (1,066 aa). For each human protein, we computed the maximum Smith–Waterman local-alignment score over all 72 non-redundant CdtB variants and ranked vinculin globally and within a ±15% length band. We also ran BLASTP 2.16.0+ with the 72 CdtB variants as queries against the rewritten human proteome database, retaining the best bit-score and E-value for each human protein as the standard length-normalized statistic. The best CdtB↔vinculin raw alignment was then re-tested against 2,000 composition-preserving shuffles of the vinculin sequence; z is reported only as a descriptive standardized distance because local-alignment scores are extreme-value distributed.

Null models and statistical analysis

Exact-k-mer significance was assessed vs an Altschul–Erickson dipeptide-preserving shuffle [17]; local-alignment significance vs a composition-preserving shuffle (classic shuffle-and-realign). We used $\geq 10^3$ iterations per test and computed empirical p with the $+1/(N+1)$ correction. Multiple testing across variants/epitopes was controlled by the Benjamini–Hochberg FDR [19] at $\alpha = 0.05$.

Structure arm

Foldseek [23] easy-search (CdtB vs the vinculin ensemble) tested fold/domain similarity. Fine surface-patch superposition (Cα, contiguous 5–12-mer windows, relative solvent-accessible surface area (RASA) $\geq 20\%$ via DSSP [20], computed with biotite [21]) was tested against a length-stratified random-fragment RMSD null (EMoMiS convention [9]; $z < -1.645$; RMSD ≤ 1 Å), with FDR.

Epitope arms

(a) Each IEDB vinculin B-cell epitope was tested for exact whole-epitope identity within CdtB and local-alignment similarity vs a per-epitope composition null (the method of the HPV-L1 negative [12]; cf. Epitopedia [8]). (b) DiscoTope-3.0 [26] predicted conformational B-cell epitopes on vinculin and CdtB structures (DTU web server); contiguous conformational-epitope regions were compared for exact and local-alignment mimicry vs a composition null with FDR.

Controls

Positive: *S. pyogenes* M5 protein (P02977) ↔ cardiac myosin MYH7 (P12883), the rheumatic-fever precedent [13]. Negative: *E. coli* MalE (P0AEX9) ↔ vinculin. Structural positive control: because CdtB adopts the DNase-I fold [7], DNase I (PDB 3DNI, 4AWN, 2DNJ, 1DNK) is a bona-fide structural homolog; we required the fold arm to separate CdtB↔DNase I from CdtB↔vinculin, and used the same homolog to test whether the fine surface-patch arm has discriminative power (`run_structure_poscontrol`). Conformational-arm positive control: the documented SARS-CoV-2 Spike↔human thrombopoietin epitope mimic (shared TQLPP motif; [11]) was passed through the identical exact + local-alignment + composition-null test (`run_conformational_poscontrol`). Background: *B. subtilis* proteome (length-matched, max-vs-max bootstrap + Mann–Whitney).

Reproducibility and software

A seed-fixed Snakemake [29] / Python pipeline regenerates every table from accessions; code and provenance are public under the MIT licence. Structure and epitope prediction used AlphaFold DB (v6) and the DiscoTope-3.0 web server (no local GPU required). Key software versions are listed in Table 2; a full lockfile is released with the code.

Table 2: **Key software and resource versions.**

Component	Version / release
Python	3.11.15
Biopython	1.87
biotite (Cα superposition, SASA)	1.4.0
statsmodels (BH-FDR)	0.14.6
pandas / numpy / scipy	3.0.3 / 2.4.6 / 1.17.1
matplotlib	3.10.9
Snakemake	9.23.0
MMseqs2 (clustering)	18-8cc5c
Foldseek (fold search)	10-941cd33
DiscoTope-3.0 (DTU web server)	accessed 15 Jun 2026
AlphaFold DB	v6
UniProt / IEDB	2026_02 / accessed 14 Jun 2026

Acknowledgments

We thank the maintainers of UniProt, the Immune Epitope Database, the RCSB Protein Data Bank, the AlphaFold Protein Structure Database, and the Technical University of Denmark DiscoTope-3.0 server for

open data and tools.

Data availability

The complete pipeline (code, configuration, provenance manifests) is available under the MIT licence at `https://github.com/leonardo-vanschaik/cdtb-vinculin-mimicry`. All results and figures regenerate from the committed accessions via Snakemake. An archived release with a Zenodo DOI will accompany journal submission.

Funding

The author received no specific funding for this work.

Competing interests

The author has declared that no competing interests exist. The author has no financial or commercial relationship with any manufacturer of anti-CdtB/anti-vinculin diagnostic assays.

Author contributions

Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing: Leo Van Schaik.

Supporting information

S1 Text. Extended methods. Non-redundancy clustering, null-model algorithms, patch enumeration, the insert-aware vinculin coordinate mapper, epitope extraction, control definitions, and database releases.

S1 Dataset. Sequence-arm results. Per-variant exact-pentapeptide and local-alignment statistics, empirical p, and FDR for all 72 non-redundant CdtB variants.

S2 Dataset. Experimental-epitope results. Per-epitope exact-identity and local-alignment statistics, empirical p, and FDR for all 153 IEDB vinculin B-cell epitopes.

S3 Dataset. Conformational-epitope results. Per-peptide statistics for the DiscoTope-3 conformational-epitope comparison.

S4 Table. Controls. Sequence, fold, fine-patch, and conformational positive and negative controls with statistics and outcomes.

S5 Dataset. Proteome-wide ranking. Top human-proteome raw-score hits plus the vinculin row, including length, raw Smith–Waterman score, BLAST bit-score, E-value, rank, and best CdtB query.

References

[1] Pimentel M, Morales W, Rezaie A, et al. Development and validation of a biomarker for diarrhea-predominant irritable bowel syndrome in human subjects. PLoS One. 2015;10(5):e0126438. doi:10.1371/journal.pone.0126438.

[2] Morales W, Rezaie A, Barlow G, Pimentel M. Second-generation biomarker testing for irritable bowel syndrome using plasma anti-CdtB and anti-vinculin levels. Dig Dis Sci. 2019;64(11):3115–21. doi:10.1007/s10620-019-05684-6.

[3] Barros LL, Leite G, Morales W, et al. Anti-CdtB and anti-vinculin antibodies to diagnose irritable bowel syndrome in inflammatory bowel disease patients. BMC Gastroenterol. 2024;24(1):448. doi:10.1186/s12876-024-03509-z.

[4] Mansoor M, Shafiq MH, Imran M, et al. Diagnostic potential of various laboratory tests for irritable bowel syndrome (IBS): a systematic review. J Pak Med Assoc. 2024;74(7):1300–8. doi:10.47391/JPMA.10571.

[5] Trost B, Lucchese G, Stufano A, Bickis M, Kusalik A, Kanduc D. No human protein is exempt from bacterial motifs, not even one. Self/Nonself. 2010;1(4):328–34. doi:10.4161/self.1.4.13315.

[6] Cusick MF, Libbey JE, Fujinami RS. Molecular mimicry as a mechanism of autoimmune disease. Clin Rev Allergy Immunol. 2012;42(1):102–11. doi:10.1007/s12016-011-8294-7.

[7] Lara-Tejero M, Galán JE. A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. Science. 2000;290(5490):354–7. doi:10.1126/science.290.5490.354.

[8] Balbin CA, Nunez-Castilla J, Stebliankin V, et al. Epitopedia: identifying molecular mimicry between pathogens and known immune epitopes. ImmunoInformatics. 2023;9:100023. doi:10.1016/j.immuno.2023.100023.

[9] Stebliankin V, Chellappan R, Baral P, et al. EMoMiS: a pipeline for epitope-based molecular mimicry search in protein structures with potential applications to SARS-CoV-2. Comput Struct Biotechnol J. 2026;31:462–74. doi:10.1016/j.csbj.2026.01.011.

[10] Penunuri G, Wang P, Corbett-Detig R, Russell SL. A structural proteome screen identifies protein mimicry in host–microbe systems. bioRxiv. 2024. doi:10.1101/2024.04.10.588793.

[11] Nunez-Castilla J, Stebliankin V, Baral P, et al. Potential autoimmunity resulting from molecular mimicry between SARS-CoV-2 Spike and human proteins. Viruses. 2022;14(7):1415. doi:10.3390/v14071415.

[12] Nishioka K, Sekiyama K, Shiro R, Tsunoda I, Matsumura N. Lack of molecular mimicry between HPV vaccine L1 antigen and human proteins by a computational analysis. Int J Clin Oncol. 2026;31(3):494–503. doi:10.1007/s10147-026-02961-z.

[13] Dale JB, Beachey EH. Sequence of myosin-crossreactive epitopes of streptococcal M protein. J Exp Med. 1986;164(5):1785–90. doi:10.1084/jem.164.5.1785.

[14] Suliman Y, Kafaja S, Oh SJ, et al. Anti-vinculin antibodies in scleroderma (SSc): a potential link between autoimmunity and gastrointestinal system involvement in two SSc cohorts. Clin Rheumatol. 2021;40(6):2277–84. doi:10.1007/s10067-020-05479-5.

[15] Hitawala FN, Gray JJ. What does AlphaFold3 learn about antibody and nanobody docking, and what remains unsolved? mAbs. 2025;17(1):2545601. doi:10.1080/19420862.2025.2545601.

[16] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7. doi:10.1016/0022-2836(81)90087-5.

[17] Altschul SF, Erickson BW. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol. 1985;2(6):526–38. doi:10.1093/oxfordjournals.molbev.a040370.

[18] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89(22):10915–9. doi:10.1073/pnas.89.22.10915.

[19] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

[20] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637. doi:10.1002/bip.360221211.

[21] Kunzmann P, Hamacher K. Biotite: a unifying open source computational biology framework in Python. BMC Bioinformatics. 2018;19(1):346. doi:10.1186/s12859-018-2367-z.

[22] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8. doi:10.1038/nbt.3988.

[23] van Kempen M, Kim SS, Tumescheit C, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42(2):243–6. doi:10.1038/s41587-023-01773-0.

[24] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. doi:10.1038/s41586-021-03819-2.

[25] Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space. Nucleic Acids Res. 2022;50(D1):D439–44. doi:10.1093/nar/gkab1061.

[26] Høie MH, Gade FS, Johansen JM, et al. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. Front Immunol. 2024;15:1322712. doi:10.3389/fimmu.2024.1322712.

[27] Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019;47(D1):D339–43. doi:10.1093/nar/gky1006.

[28] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31. doi:10.1093/nar/gkac1052.

[29] Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2. doi:10.1093/bioinformatics/bts480.

Figures

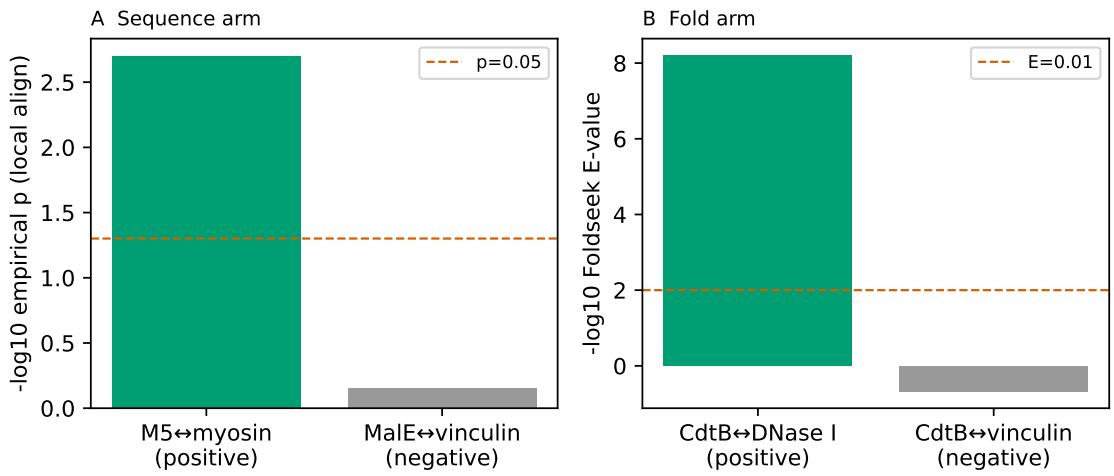

Figure 1: **Pipeline calibration (two arms).** (A) Sequence arm: the positive control (M5↔myosin) is recovered ($p = 0.002$) and the negative control (MalE↔vinculin) rejected ($p = 0.71$); bars show $-\log_{10}$ empirical p for the local-alignment test (dashed line $p = 0.05$). (B) Fold arm: Foldseek separates CdtB from a genuine DNase-I-fold structural homolog (CdtB↔DNase I, $E = 6 \times 10^{-9}$) but not from vinculin ($E = 4.7$); bars show $-\log_{10}$ E-value (dashed line $E = 0.01$). Both arms recover a true positive and reject the negative.

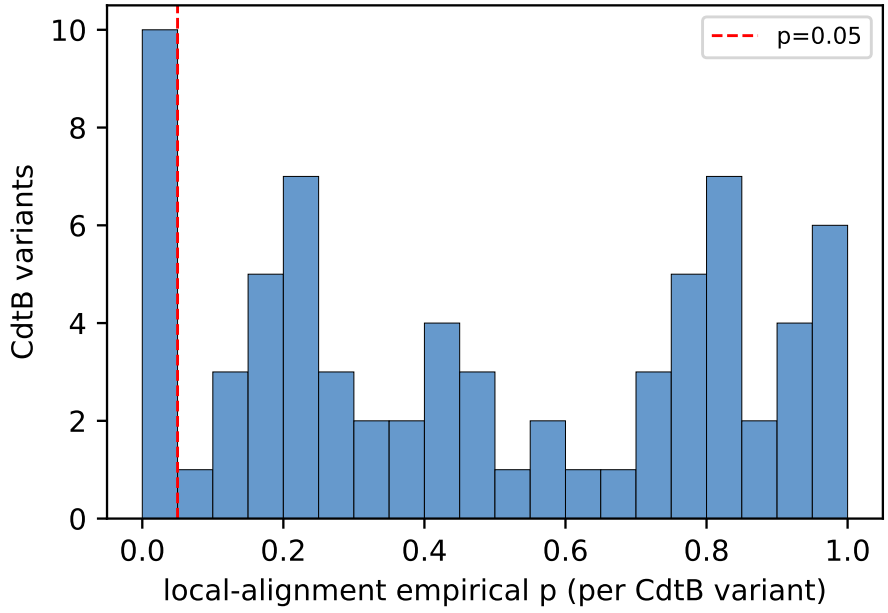

Figure 2: **Sequence arm.** Per-variant local-alignment empirical-p distribution across 72 non-redundant CdtB variants; none FDR-significant (dashed line $p = 0.05$).

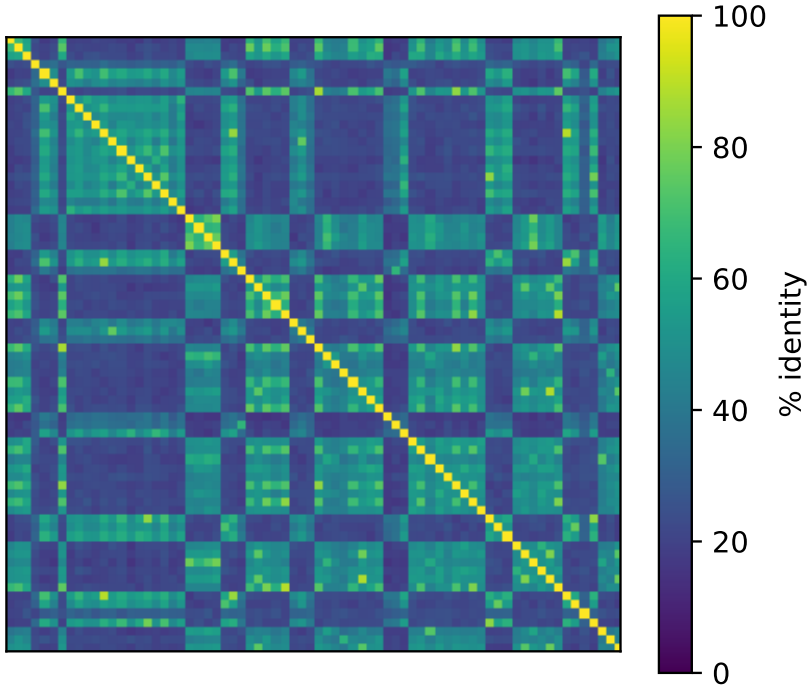

Figure 3: **CdtB family diversity.** Pairwise %identity (14.7–89.1%, median 26%; $n = 72$), confirming a broad, well-sampled family rather than a narrow clade.

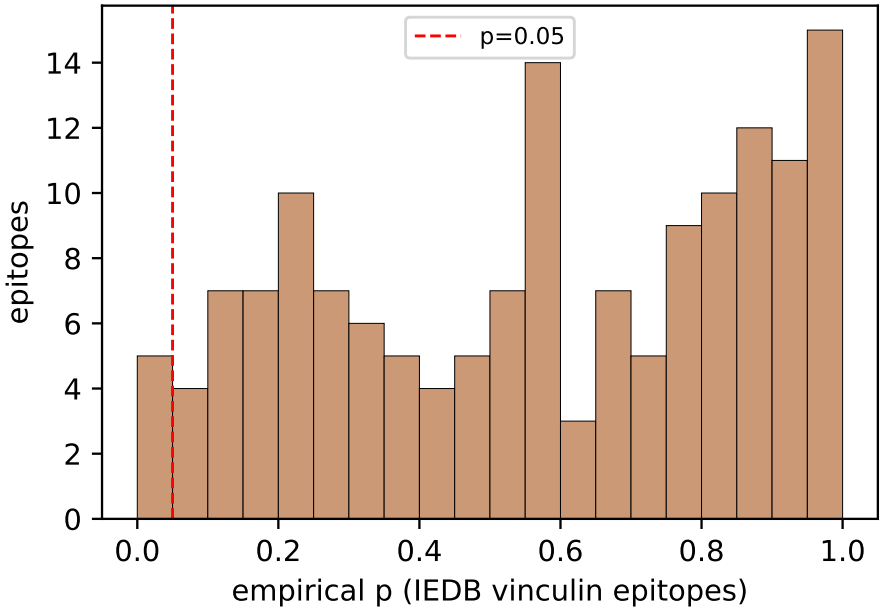

Figure 4: **Experimental B-cell epitope arm.** Best per-epitope local-alignment empirical p for experimentally validated IEDB vinculin epitopes ($n = 153$); no FDR-significant CdtB overlap (dashed line $p = 0.05$).

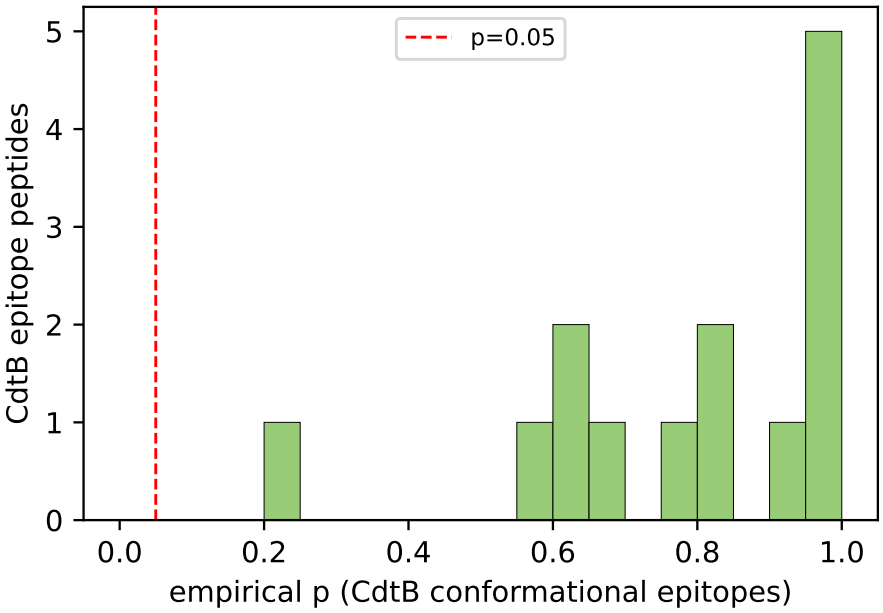

Figure 5: **Conformational epitope arm.** Best local-alignment empirical p for DiscoTope-3 conformational-epitope peptides ($n = 14$); no FDR-significant CdtB overlap (dashed line $p = 0.05$).

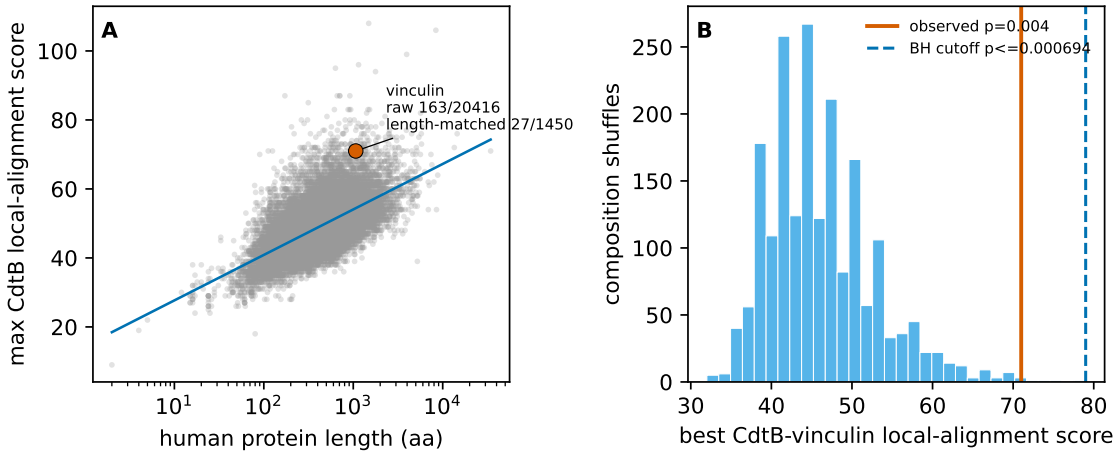

Figure 6: **Proteome-wide context for the best CdtB–vinculin sequence similarity.** (A) For each reviewed human protein, the maximum Smith–Waterman score against any of 72 CdtB variants is plotted against protein length (log scale). Vinculin (P18206-2) lies in the raw-score tail (rank 163/20,416; 27/1,450 within ±15% length) but is not unique among unrelated human proteins. (B) Composition-shuffle null for the best CdtB↔vinculin pair. The observed score is nominally elevated ($p = 0.0040$) but below the family-wide Benjamini–Hochberg cutoff for 72 CdtB variants ($p \leq 0.000694$; score threshold 79).