

Horizontal Gene Transfer Drives Patchy Distribution of CPR-C4 Cysteine Proteases with Deep Structural Conservation

Minxin Zhang

wenzhangbentiancheng@gzgs.edu.cn

GCTB-NSU Joint Institute of Technology at Guangzhou College of Technology and Business

<https://orcid.org/0009-0000-5061-1836>

Qinan Han

Independent Researcher / Unaffiliated <https://orcid.org/0009-0009-9150-950X>

Research Article

Keywords: horizontal gene transfer, Candidate Phyla Radiation, Patescibacteria, cysteine protease, purifying selection, structural conservation, conjugation

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11407101/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Horizontal Gene Transfer Drives Patchy Distribution of CPR-C4 Cysteine Proteases with Deep Structural Conservation

Minxin Zhang^{1,*} and Qinan Han²

¹GCTB-NSU Joint Institute of Technology, Guangzhou College of Technology and Business, Guangzhou

²Independent Researcher, Sanya, Hainan, China

*Correspondence: wenzhangbentiancheng@gzgs.edu.cn

October 17, 2026

Abstract

Horizontal gene transfer (HGT) can carry a protein family across deep phylogenetic divides, but whether a horizontally mobile family retains structural and functional integrity over such distances—and under what selective regime—remains poorly characterized in the Candidate Phyla Radiation (CPR), a radiation of streamlined, host-dependent genomes. Here we use the CPR-C4 cysteine protease family to address this question at genome-database scale. Across 6,411 CPR representative genomes, CPR-C4 homologs occur in only 14.0% of species and are entirely absent from Saccharimonadia (1,310 species), a patchy distribution incompatible with the family’s previous description as “highly conserved” in CPR. Phylogenetic incongruence testing attributes this pattern to HGT: approximately unbiased tests reject the GTDB species tree ($AU = 0.000$; $\Delta\text{LogL} = 1,039$), the Robinson–Foulds distance is 0.862, only 3/10 orders are monophyletic, and flanking regions are specifically enriched for conjugative transfer proteins ($OR = 4.66$, $p = 8.6 \times 10^{-8}$) but not for transposons or phages. Despite 27–62% pairwise sequence identity, 60 representative homologs spanning 7 classes and 30 orders all fold into high-confidence structures (mean pLDDT = 94.5), cross-validated by an architecturally independent language-model predictor, with 55/60 retaining the intact catalytic triad (Cys77–His109–Leu131). Structural contact conservation (71.5% in distant homologs) tracks sequence divergence (Spearman $\rho = 0.69$) while declining more slowly than sequence identity, and three independent tests reject hydrophobic core compensation. Instead, strong purifying selection (global $dN/dS = 0.079$), localized to the catalytic core by site-resolved analysis (FEL), maintains fold and catalytic geometry across transfers. Non-CPR homologs from the NCBI nr database are phylogenetically nested within CPR diversity, indicating a CPR-centered origin with repeated cross-phylum spillover, and carriage is sharply depleted in extreme environments (1/273 versus 786/4,698; phylogenetic bootstrap $P = 0.0015$) while remaining indifferent to host association. These results establish CPR-C4 as a horizontally mobile, functionally constrained protease family, show that strong purifying selection can preserve a catalytic fold across phylum-level transfers, and indicate that distribution claims for CPR gene families warrant re-examination at current genome scale.

1 Introduction

The Candidate Phyla Radiation (CPR) represents a major branch of bacterial diversity characterized by small genomes, reduced metabolic capabilities, and dependence on host organisms. Understanding the distribution and evolution of protein families in CPR is critical for reconstructing their ecological roles and evolutionary history. The CPR-C4 family of cysteine proteases, characterized by a non-canonical Cys–His–Leu catalytic triad, was recently reported as “highly conserved in CPR” based on limited sampling [1].

Here we reanalyze the distribution, evolution, and structural conservation of CPR-C4 across the CPR radiation. Using 6,411 representative genomes from GTDB R226, we quantify the true prevalence of this family, test its evolutionary history against vertical inheritance, assess structural conservation across deep phylogenetic divergence, and measure the selective pressure acting on its coding sequence.

2 Methods

Data Sources

All data are publicly available: GTDB release R226 [2] (6,411 CPR representative genomes), NCBI RefSeq (mobile element protein library), NCBI nr protein database, PDB 7PJO [1] (crystal structure of CPR-C4, 2.25 Å). Throughout this work, residue numbering follows the 7PJO PDB chain numbering: the catalytic triad Cys77–His109–Leu131 corresponds to Cys61–His93–Leu115 in the sequence numbering of the original report [1] (offset +16, because the N-terminal 16 residues are unresolved in the structure).

Bioinformatic Pipeline

Custom Python scripts orchestrated gene calling (Prodigal [3]), homology search (BLASTp [4], e-value $< 10^{-5}$), and structure prediction (ColabFold [5]). Phylogenetic incongruence was assessed using IQ-TREE 2 [6] with approximately unbiased (AU) tests [7] (10,000 replicates) and Robinson–Foulds distance [8]. As a complementary measure of topological conflict, we computed the fraction of gene-tree bipartitions absent from the species tree using custom Python scripts ($53/58 = 91.4\%$). Mean-field DCA [9] was performed with standard pseudocounts ($\lambda = 0.5 \cdot M_{\text{eff}}$). Flanking region analysis used ± 10 kb windows with matched background controls. All scripts are available from the corresponding author upon reasonable request.

Selection Pressure Analysis

We estimated dN/dS ratios using codeml in PAML 4.10.10 [10]. The codon alignment was generated by back-translating the MAFFT (v7.525) [11] protein multiple sequence alignment onto coding sequences extracted from Prodigal gene calls. The alignment comprised 60 homologs with 416 informative codon sites after removing ambiguous alignment columns (cleandata = 1). The M0 model (single dN/dS ratio, F3x4 codon frequencies) was used for global ω estimation with the ML tree topology fixed. Site-specific models M7 and M8 were attempted but did not

converge within computational limits; given the extremely low global ω , site-specific positive selection is highly unlikely.

Site-Level Selection Analysis

As an independent, site-resolved test of selective pressure, the same 60-sequence codon alignment (592 columns) and ML tree were analyzed with HyPhy [12]: FEL (fixed-effects likelihood) for pervasive selection and MEME [13] for episodic diversifying selection, each at significance threshold $p \leq 0.1$. Alignment columns were mapped onto 7PJO PDB residue numbers by adding the reference sequence to the protein alignment with MAFFT [11] (`--add --keeplength`); the mapping was validated by recovering the catalytic triad at its annotated columns. For each mapped residue we computed the minimum $C\alpha$ distance to the catalytic triad (Cys77, His109, Leu131) in the 7PJO crystal structure and burial as the number of $C\alpha$ contacts within 8 Å at sequence separation ≥ 4 ; selected versus non-selected sites were compared by two-sided Mann–Whitney U tests.

NCBI nr Survey

To test whether CPR-C4 homologs exist beyond GTDB-sampled genomes, remote BLASTp searches of five representative CPR-C4 sequences were performed against the NCBI nr database, retrieving the top 1,000 hits. Hits were classified to phylum level using NCBI taxonomy. Ninety-six representative non-CPR hits were combined with the 60 CPR homologs for an extended phylogeny inferred with IQ-TREE 2 (Q.pfam+F+I+R7).

ESMFold Cross-Validation

To rule out that structural conclusions depend on a single prediction method, all homologs were re-predicted with ESMFold [14], whose protein-language-model architecture operates on single sequences and is independent of ColabFold’s MSA-based pipeline. Predictions were run at full precision on a single GPU; three homologs longer than 250 residues were excluded owing to memory limits and one prediction with pLDDT < 50 (cpr049, pLDDT = 46.7) was classified as a failed prediction, leaving 56 valid homologs for comparison (median ESMFold pLDDT = 84.5; across all 57 predicted models the range was 46.7–91.3, and across the 56 retained models 73.9–91.3). Because both methods predicted the identical input sequence, models were compared residue-by-residue: $C\alpha$ RMSD after Kabsch superposition and contact-map Jaccard similarity (8 Å $C\alpha$, separation ≥ 4 ; long-range ≥ 12).

Statistical Analysis

Permutation tests (20,000 replicates) were used for correlation analyses. Fisher exact tests assessed enrichment in flanking regions. Structural contact maps used 8 Å $C\alpha/C\beta$ thresholds with sequence separation ≥ 4 ; sensitivity was tested by recomputing contact conservation with 6 Å and 10 Å thresholds, which did not change the Spearman correlation between jaccard and pident ($\rho = 0.68$ – 0.72 , $p < 10^{-9}$) or the qualitative conclusion. ColabFold model confidence ranged from pLDDT 80.2 to 97.4 (mean 94.5) across the 60 homologs.

Ecological Metadata Analysis

Isolation-source metadata were retrieved from the NCBI BioSample database via the Entrez Programming Utilities (Biopython) for all 6,435 representative genomes, using batch `efetch` requests (100 accessions each) at ≤ 3 requests per second; records were parsed for the `isolation_source` attribute. A value was recovered for 5,450 genomes (84.7%). Each record was classified as host-associated or environmental by case-insensitive keyword matching against curated term lists (host-associated terms included oral, saliva, gut, gastrointestinal, feces, human, patient, blood, skin, rumen, animal, insect; environmental terms included groundwater, soil, sediment, wastewater, sludge, lake, river, marine, seawater, spring, aquifer, leachate, bioreactor, permafrost); records matching neither list were left unclassified and excluded. Associations were tested by Fisher exact test both on all classified genomes and after restricting to “carrier-capable” classes (classes with at least one carrier genome-wide), which removes lineages such as *Saccharimonadia* in which the family is structurally absent. For the extreme-environment analysis, environmental records were further subdivided a priori by keyword matching into cold seep, hydrothermal, hypersaline, deep marine sediment, and other environmental habitats, and genome completeness was controlled using CheckM values from GTDB metadata. Phylogenetic correction used phylogenetic logistic regression [15] (phylolm, logistic.IG10), phylogenetic generalized estimating equations [16] with an identity link, and a Brownian-threshold parametric bootstrap (20,000 replicates) on the GTDB bac120 tree pruned to the classified genomes and midpoint-rooted.

3 Results

CPR-C4 Distribution is Patchy, Not Conserved

We identified CPR-C4 homologs in only 895 of 6,411 CPR species (14.0%). The distribution was highly uneven: *Minisyncoccia* showed 26.9% prevalence (684/2,547), *Patescibacteriia* 16.9% (140/830), while *Saccharimonadia* (1,310 species) showed complete absence (0/1,310). *Microgenomatia* had only 0.4% prevalence (Figure 1).

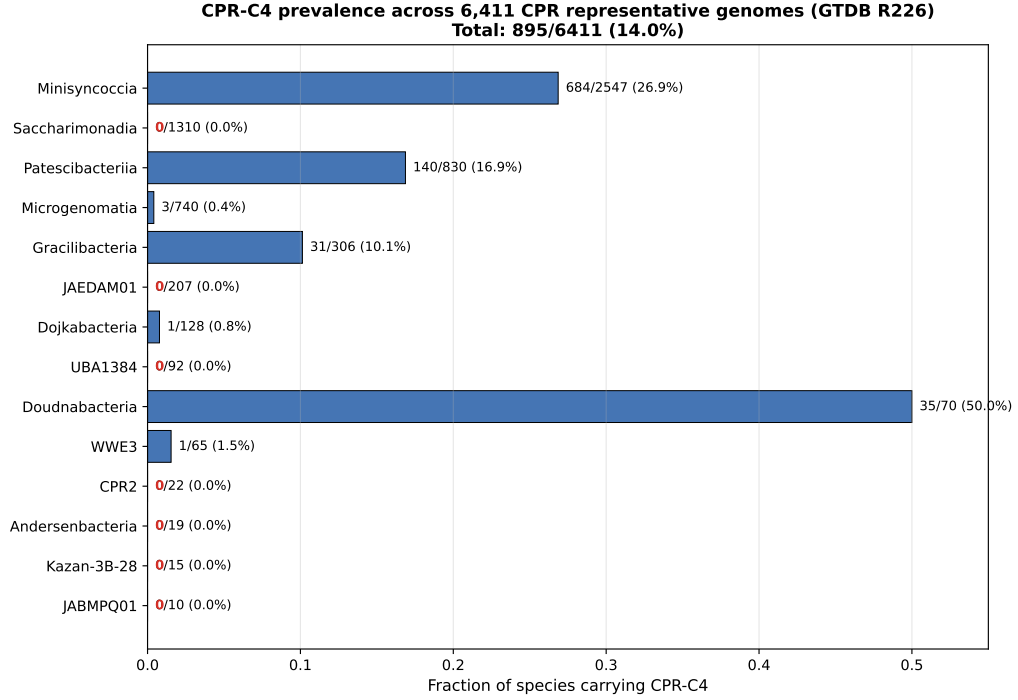


Figure 1: Class-level distribution of CPR-C4 across 6,411 CPR representative genomes (GTDB R226). Bars show the fraction of species in each class carrying at least one CPR-C4 homolog; labels indicate carrier counts over class size. Overall prevalence is 895/6,411 (14.0%). *Saccharimonadia* ($n = 1,310$), the largest CPR class, is completely devoid of CPR-C4. The 14 largest classes are shown; the remaining 16 classes (each $n \leq 9$) contain no CPR-C4 carriers.

Horizontal Gene Transfer Drives the Distribution

The unconstrained protein tree ($\text{LogL} = -19,512.09$) was strongly supported over the GTDB species tree ($\text{LogL} = -20,550.91$; $\Delta\text{LogL} = 1,038.82$). AU tests rejected the species tree ($\text{AU} = 0.000$) and accepted the protein tree ($\text{AU} = 1.000$). Robinson–Foulds distance was 0.862, with 91.4% of gene-tree bipartitions (53/58) absent from the species tree, indicating pervasive topological conflict. Only 3/10 testable orders were monophyletic; the largest order UBA9973 ($n = 11$) was non-monophyletic (Figure 2).

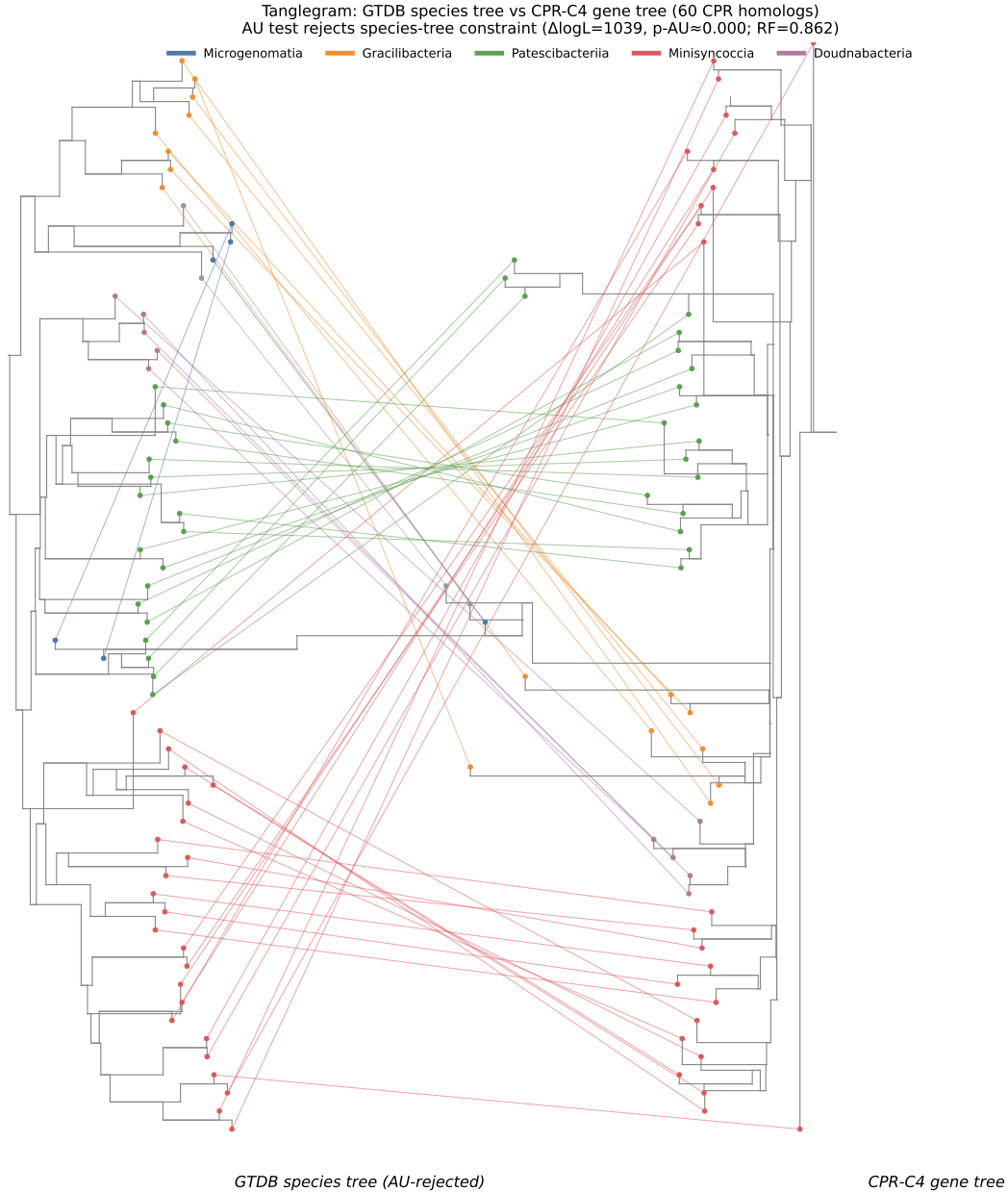


Figure 2: Tanglegram between the GTDB species tree (left) and the CPR-C4 protein tree (right) for the 60 representative homologs. Connecting lines map identical genomes; extensive crossing visualizes the topological incongruence ($\Delta\text{LogL} = 1,039$, AU test $p \approx 0$, Robinson–Foulds = 0.862; only 3/10 orders monophyletic), as expected under horizontal gene transfer. The GTDB species tree is based on 120 concatenated marker proteins; internal node bootstrap support ranges from 0 to 100 (median 43), indicating that some topological conflict may reflect species-tree uncertainty rather than HGT alone.

Extended nr Survey Supports CPR-Centered Diversification

Remote BLAST against NCBI nr recovered homologs beyond GTDB-sampled genomes: of the top 1,000 hits, 288 were classified as CPR and 712 as non-CPR, spanning more than 30 phyla and including 151 archaeal hits (up to 57.6% identity and 99% coverage; best hit in a Nanobdellota archaeon). However, metagenomic bins frequently contain mis-binned or unclassified

CPR-derived contigs (70 hits lacked any phylum-level assignment), so protein-database taxonomy alone overestimates true non-CPR prevalence. The extended phylogeny (60 CPR + 96 non-CPR representatives) shows non-CPR sequences either nested directly within CPR diversity as sisters to CPR lineages, or forming shallow independent clusters (Figure 3). This topology is consistent with CPR-centered diversification followed by rare outward spillover, rather than acquisition of CPR-C4 from a non-CPR donor, in agreement with the genome-resolved GTDB analysis, in which 918 of the 919 recovered homolog hits lie within CPR genomes (species-level carriage is $895/6,411 = 14.0\%$; the hit and species counts are not directly comparable, being tallied per sequence versus per genome). Moreover, a local re-screening of the GTDB R226 genomes with the same dual-profile approach recovered genome-anchored CPR-C4 homologs in nine additional non-CPR phyla, including the first family record in the candidate genus *Candidatus Tepidifontimicrobium*; all ten genome-anchored representatives passed five orthogonal validation lines (profile-HMM re-scan, catalytic-triad columns, phylogenetic placement, GTDB-Tk reclassification, and ColabFold structure prediction with crystal-anchored triad geometry within ± 1.1 Å of 7PJO; Supplementary Note 1, Tables S1–S2).

CPR-C4 gene tree v2 (61 CPR + 106 non-CPR homologs, 30+ phyla incl. DPANN archaea)
Q.pfam+F+I+R7; red=non-CPR; midpoint-rooted



Structural Conservation Across Deep Divergence

Despite 27–62% sequence identity, all 60 representative homologs (7 classes and 30 orders; the largest order, UBA9973, contributed 11 of the 60 genomes, while 20 orders were represented by a single genome) folded into high-confidence structures (mean pLDDT = 94.5, pTM = 0.90). The catalytic triad Cys77–His109–Leu131 was intact in 55/60 homologs. Structural contact conservation was 0.852 for close homologs ($\geq 50\%$ identity) and 0.715 for distant homologs ($< 45\%$ identity). Contact conservation tracks sequence divergence with a Spearman correlation of 0.69 ($p = 9.5 \times 10^{-10}$), declining more slowly than sequence identity. 54% of reference contacts (332/614) were preserved in $\geq 90\%$ of distant homologs. Core C α RMSD against the 7PJO crystal structure had a median of 0.94 Å across all 60 homologs, with a single outlier (cpr046, Gracilibacteria, 15.1 Å; its 232-residue sequence is 16 residues longer than the 216 residues resolved in the 7PJO chain) (Figure 4).

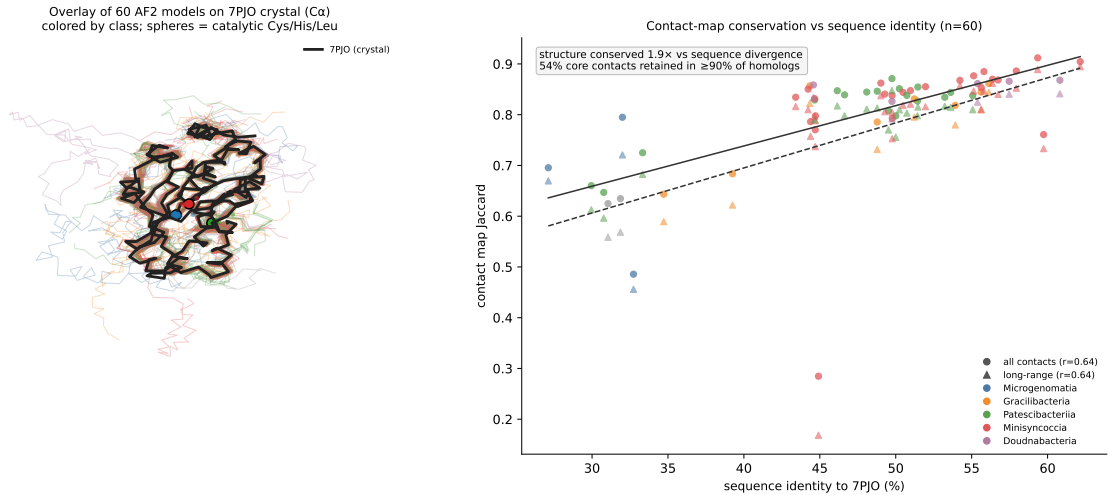


Figure 4: Structural conservation of 60 CPR-C4 homologs across deep sequence divergence. Left: superposition of all ColabFold models onto the 7PJO crystal structure (C α trace), with the catalytic triad Cys77–His109–Leu131 shown as spheres; core C α RMSD median = 0.94 Å. Right: contact-map Jaccard similarity versus sequence identity; contact conservation tracks sequence divergence with a Spearman correlation of 0.69, and 54% of reference contacts are retained in $\geq 90\%$ of distant homologs.

ESMFold Cross-Validation Confirms Fold Conservation

Because the contact-conservation analysis rests on ColabFold predictions, we repeated structure prediction with ESMFold, whose single-sequence language-model architecture is independent of ColabFold’s MSA-based pipeline. For the 56 homologs with successful ESMFold predictions, the two methods converged on the same fold (C α RMSD median = 2.57 Å, IQR 2.02–3.85) with concordant contact maps (Jaccard median = 0.801 overall and 0.780 for long-range contacts; 54/56 homologs ≥ 0.7). Predictor agreement (0.801) was comparable to the cross-species contact conservation measured against the 7PJO reference (median 0.839 across the same 56 homologs; 0.770 for the distant subset with $< 45\%$ identity, $n = 19$), indicating that the structural similarity observed across species reflects genuine conservation rather than a shared prediction bias. Two exceptions were informative: cpr049 failed ESMFold prediction (pLDDT = 46.7) while

its ColabFold model matches the reference closely (core RMSD 0.95 Å), and cpr046 showed large model-to-model deviations under both methods, supporting its interpretation as genuine structural variation rather than prediction error.

Purifying Selection Maintains CPR-C4 Function

To test whether selective pressure explains the structural conservation observed despite deep sequence divergence, we estimated the ratio of non-synonymous to synonymous substitution rates ($\omega = dN/dS$) using codeml. The M0 model yielded $\omega = 0.079$ ($\kappa = 1.40$, $\ln L = -9,404.17$), indicating that non-synonymous substitutions accumulate at only 7.9% of the neutral expectation. The dS tree length (145.94) exceeded the dN tree length (11.55) by 12.6-fold, confirming that synonymous substitutions accumulate freely while non-synonymous variation is strongly suppressed. This global ω provides the mechanistic basis for the structural conservation in Figure 4: with non-synonymous substitutions accumulating at only 7.9% of the neutral rate, the catalytic triad (intact in 55/60 homologs) and overall fold (core C α RMSD = 0.94 Å) are maintained across 27–62% sequence divergence. The selective constraint operates despite the gene’s horizontal mobility, indicating that CPR-C4 provides a selectable function wherever it is acquired.

Site-Level Selection is Concentrated at the Catalytic Core

Site-resolved analysis with FEL identified 212 sites under pervasive purifying selection and only a single site under pervasive positive selection ($p \leq 0.1$; the latter lies in a region not resolved in 7PJO). All three catalytic residues showed $\beta \approx 0$ ($p \approx 0$), placing them among the most constrained sites in the protein. Purifying-selection sites mapped onto the 7PJO structure were significantly closer to the catalytic triad than non-selected sites (median 13.8 Å versus 19.7 Å C α distance, Mann–Whitney $p < 10^{-4}$) and more buried (median 4 versus 2 C α contacts, $p = 0.011$), indicating that constraint is concentrated on the structural core surrounding the active site. MEME detected 12 sites under episodic diversifying selection; these were predominantly peripheral (median distance to the triad 23.0 Å versus 14.4 Å for all residues), with only two sites near the active site (7PJO residues 75 and 110, at 6.6 Å and 3.8 Å; $p = 0.010$ and 0.072). Given the deep divergence of the alignment (27–62% identity), these episodic signals are reported descriptively and interpreted cautiously.

Conjugative Transfer, Not Transposons or Phages, Associates with CPR-C4 Flanks

Two complementary analyses probed which mobile elements accompany CPR-C4. Flanking-region enrichment against matched background controls found no enrichment for transposons (OR = 0.80), integrases (OR = 0.16), recombinases (OR = 0.62), T4SS (OR = 0.12), or phages (OR = 0.22), whereas conjugative transfer proteins were significantly enriched (OR = 4.66, $p = 8.6 \times 10^{-8}$; 58 hits across 58 genomes). Immediate proximity (± 2 genes) was rare (0.65%), consistent with ancient transfer. Independently, a co-occurrence network across all 895 carrier genomes recovered mobile-element hits in the ± 10 kb flanks of 161 genomes; among these, conjugative transfer proteins were again enriched relative to matched background genomes (OR

= 1.80, $p < 0.05$), while transposons, phages, integrases, and T4SS showed no enrichment or depletion. The predominant conjugative subtype was Ti-type (57 genomes), followed by P-type (19 genomes). Only 10 of 161 genomes carried more than one mobile-element type in the immediate flank, again indicating that most transfer events are ancient and the flanks have since been homogenized by genome reduction.

Mechanistic Tests Reject Core Compensation

Three independent tests rejected hydrophobic core compensation: (1) $\Delta\Delta G$ proxy analysis showed distant homologs have more fragile cores ($r = -0.36$, $p = 0.006$); (2) ancestral state reconstruction showed ancestors have higher buried hydrophobic fraction (0.540 vs 0.512, $p = 0.0002$) with no branch-level accumulation ($p = 0.106$); (3) mean-field DCA showed core-core covariation enrichment disappeared after controlling for physical contact ($p = 0.10$).

CPR-C4 is Depleted in Extreme Environments

Within environmental genomes, carriage was strongly depleted in extreme habitats: 0/52 in cold seeps, 1/136 in hydrothermal settings, 0/85 in hypersaline habitats, and 0/40 in deep marine sediments. Pooling the cold-seep, hydrothermal-vent, and hypersaline categories gave 1 carrier among 273 genomes, versus 786 among 4,698 (16.7%) in other environmental genomes (OR = 0.018, Fisher $p = 7.5 \times 10^{-20}$); deep marine sediments were likewise depleted when tested separately against the remaining environmental genomes (0/40 versus 787/4,931, $p = 1.7 \times 10^{-3}$). This depletion is not a lineage-composition artifact: cold-seep genomes are dominated by *Minisyncoccia* (31/52), the class with the highest genome-wide carriage (26.9%), and within *Minisyncoccia* extreme-environment genomes remained devoid of carriers (0/138 versus 606/2,089, $p = 3.3 \times 10^{-20}$). Genome completeness did not differ between extreme and other environmental genomes (CheckM median 68.3% versus 68.0%, $p = 0.22$). These raw counts involve a near-empty cell (1 carrier in 273), so the Fisher estimate is unstable and we treat the phylogenetic correction as the primary evidence. Phylogenetic logistic regression did not converge owing to quasi-complete separation (itself diagnostic of the effect size); generalized estimating equations gave a consistent point estimate (baseline carrier probability 0.159; extreme-effect coefficient -0.156); and a Brownian-threshold parametric bootstrap on the GTDB tree—the most robust of the three—gave $P = 0.0015$ for the full extreme set (41.3 carriers expected, 1 observed) and $P = 0.0070$ within *Minisyncoccia* (37.7 expected, 0 observed). Read together with the host-association null result above, these data show that carriage is indifferent to host association yet sharply depleted under extreme physicochemical conditions.

4 Discussion

Horizontal Gene Transfer Reshapes CPR Protease Distribution

CPR-C4 is a horizontally mobile gene family rather than a vertically inherited conserved marker. The patchy distribution reflects the dynamics of HGT gain and loss across CPR lineages. The association with conjugative transfer proteins, rather than transposons or phages, suggests conjugation-mediated gene flow. The extended nr survey further shows that homologs outside

GTDB-sampled genomes are phylogenetically nested within CPR diversity, consistent with a CPR-centered origin rather than an external donor.

Structural Constraints Maintain Function Despite Sequence Divergence

The slower decay of structural contact conservation relative to sequence identity demonstrates that functional constraints lock the fold while sequences drift, consistent with the canonical sequence–structure relationship [17]. The global dN/dS of 0.079, indicating that non-synonymous substitutions accumulate at only 7.9% of the neutral rate, is consistent with strong purifying selection and provides the population-genetic basis for the structural conservation. The intact catalytic triad in 92% of homologs is consistent with preserved catalytic function. Site-resolved FEL analysis corroborates this at the residue level: all three catalytic residues are among the most constrained sites, and purifying-selection sites cluster on the buried core around the active site, whereas the few episodic diversifying signals lie predominantly at the periphery.

Isolation Source Does Not Independently Predict Carriage

The complete absence of CPR-C4 in Saccharimonadia (1,310 species) is notable because this class is disproportionately represented in host-associated environments, particularly the human oral microbiome (TM7 lineage). A preliminary association based on class-level taxonomy as an ecological proxy appeared to support an environmental-retention model (host-associated 0.14%, 3/2,072 vs environmental 22.1%, 857/3,876; OR = 0.01). To test this directly, we retrieved isolation-source metadata for 5,450 of 6,435 representative genomes (84.7%) from NCBI BioSample and classified each record as host-associated or environmental. The raw comparison reproduced the direction of the proxy result (host-associated 3.09%, 9/291 vs environmental 15.83%, 787/4,971; OR = 0.170, $p = 1.0 \times 10^{-11}$), but this association proved to be a taxonomic confound: 80% of host-associated genomes (234/291) belong to Saccharimonadia, the one class in which CPR-C4 is entirely absent. After restricting to carrier-capable lineages, the association vanished (host-associated 22.50%, 9/40 vs environmental 19.29%, 787/4,079; OR = 1.21, $p = 0.55$); within Minisyncoccia, the only well-sampled class spanning both categories, host-associated genomes were not depleted (40.0%, 8/20 vs 27.2%, 606/2,227). We therefore find no evidence that isolation source per se influences CPR-C4 carriage: the apparent environmental enrichment reflects the lineage composition of host-associated genomes rather than ecological selection. This null result cautions against inferring selection from taxonomic proxies.

However, a second ecological axis, environmental extremity, revealed a strong and opposite signal: extreme habitats were nearly devoid of CPR-C4 (1/273 carriers versus 786/4,698, 16.7%; Fisher OR = 0.018, $p = 7.5 \times 10^{-20}$). Because the extreme cell is nearly empty, the phylogenetically corrected test is the primary evidence, and the depletion survived it (Brownian-threshold bootstrap $P = 0.0015$ overall and $P = 0.0070$ within Minisyncoccia). The effect is unlikely to reflect taxonomic composition or genome completeness: cold-seep genomes belong predominantly to Minisyncoccia, the class with the highest genome-wide carriage; the depletion persisted within that class; and completeness did not differ between habitat types (see Results). The realized niche of CPR-C4 therefore appears to be moderate rather than extreme environ-

ments, consistent with loss or non-acquisition of the gene in extreme-environment lineages where this protease confers little advantage. Taken together with the host-association null result, these analyses indicate that niche effects on CPR-C4 are environment-specific rather than a general property of the host-versus-environment axis.

Implications for CPR Genome Evolution

The presence/absence pattern of CPR-C4 provides a case study for distinguishing core versus optional genes in reduced CPR genomes. Unlike truly conserved housekeeping genes, CPR-C4 is an optional accessory that confers proteolytic capability when acquired but is dispensable in most lineages.

Limitations

This study relies on computational predictions and public database annotations. The structural conclusions are supported by two architecturally independent predictors (ColabFold and ESMFold), but experimental validation of protease activity, of the structures themselves, and of HGT mechanisms is still needed. Three of 60 homologs exceeded ESMFold memory limits and one ESMFold prediction failed quality thresholds, so the cross-validation covers 56 of 60 homologs. The bipartition conflict analysis quantifies the extent of tree disagreement (91.4% of gene-tree splits) but does not explicitly estimate individual transfer, duplication, and loss events, which would require dedicated reconciliation software. DCA analysis may be confounded by HGT-induced population structure. Taxonomic assignment of nr hits from metagenomic sources is uncertain because CPR-derived contigs are frequently mis-binned; our genome-resolved GTDB analysis therefore remains the primary evidence for prevalence. The ecological test uses direct NCBI BioSample isolation-source metadata but relies on keyword classification of free-text records (84.7% coverage); because host-associated genomes are dominated by Saccharimonadia, the association was conditioned on carrier-capable lineages, under which no effect remained. The extreme-environment analysis is categorical rather than a regression on continuous physicochemical gradients, because BioSample metadata lack quantitative temperature or salinity values; category sizes are modest ($n = 40\text{--}136$) and the same keyword-classification caveat applies. MEME episodic positive-selection signals derive from a deeply diverged alignment and were not subjected to multiple-testing correction; they may partly reflect heterotachy or alignment artifacts and are reported descriptively only. More fundamentally, FEL and MEME assume that all sites share a single genealogical history, an assumption violated by the pervasive HGT documented here; site-level selection results are therefore reported as trends rather than per-site certainties. Selection across the four extreme-habitat categories was not corrected for multiple testing, and the Mann–Whitney comparisons of selected-site distance and burial treat residues as independent despite their phylogenetic non-independence; both are exploratory. Taxonomic control in the extreme-environment analysis was applied at the class level; finer (order- or genus-level) stratification was not possible given category sizes.

5 Conclusions

CPR-C4 cysteine proteases are horizontally mobile, not highly conserved. Their patchy distribution reflects conjugation-mediated HGT, while strong purifying selection ($dN/dS = 0.079$) preserves structure and function across deep divergence. Purifying constraint is concentrated at the catalytic core, and the family’s realized niche is moderate rather than extreme environments.

Data Availability

All data are publicly available from GTDB R226, NCBI RefSeq, NCBI nr, and PDB 7PJO. Scripts are available from the corresponding author upon reasonable request. Supplementary Information, including full dN/dS analysis details, is provided as a separate file.

Author Contributions

Conceptualization, M.Z. and Q.H.; Methodology, M.Z.; Software, M.Z.; Formal Analysis, M.Z. and Q.H.; Investigation, M.Z. and Q.H.; Data Curation, M.Z.; Writing — Original Draft, M.Z.; Writing — Review & Editing, M.Z. and Q.H.; Visualization, M.Z.; Supervision, M.Z. All authors have read and agreed to the published version of the manuscript.

Competing Interests

The authors declare no competing interests.

References

- [1] Cornish K. A. S., Lange J., Aevansson A. & Pohl E. CPR-C4 is a highly conserved novel protease from the Candidate Phyla Radiation with remote structural homology to human vasohibins. *Journal of Biological Chemistry* **298**(5), 101919 (2022).
- [2] Parks D. H., et al. GTDB: an ongoing census of genome quality and diversity. *Nucleic Acids Research* **50**, D785–D794 (2022).
- [3] Hyatt D., et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
- [4] Altschul S. F., Gish W., Miller W., Myers E. W. & Lipman D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
- [5] Mirdita M., et al. ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679–682 (2022).
- [6] Minh B. Q., et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
- [7] Shimodaira H. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* **51**, 492–508 (2002).

- [8] Robinson D. F. & Foulds L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147 (1981).
- [9] Morcos F., et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences USA* **108**, E1293–E1301 (2011).
- [10] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).
- [11] Katoh K. & Standley D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
- [12] Kosakovsky Pond S. L., Frost S. D. W. & Muse S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005).
- [13] Murrell B., et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* **8**(7), e1002764 (2012).
- [14] Lin Z., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
- [15] Ives A. R. & Garland T. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology* **59**, 9–26 (2010).
- [16] Paradis E. & Claude J. Analysis of comparative data using generalized estimating equations. *Journal of Theoretical Biology* **218**, 175–185 (2002).
- [17] Chothia C. & Lesk A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* **5**, 823–826 (1986).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryinfov2.pdf](#)