

Supplementary Information

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

[Author Names]

[Affiliation]

[Email]

October 16, 2026

This document contains Supplementary Note 1 and Supplementary Tables S1–S2. It accompanies `preprint_v2`. References cited by name below are listed in the main manuscript bibliography.

Supplementary Note 1: Orthogonal validation of cross-phylum CPR-C4 homologs

Provenance

A local re-screening of the full GTDB R226 representative genome set with the dual-profile approach used in the primary survey (profile-HMM scan combined with iterative profile–profile search) recovered CPR-C4 homologs in nine non-CPR phyla—Sisuimicrobiota, UBP6, Krumholzibacteriota, Methyloirabilota, JADJOY01, Moduliflexota, Marinisomatota, Bdellovibrionota_F and Bdellovibrionota_B—including the first family record in the candidate genus *Candidatus Tepidifontimicrobium* (Table S2). Ten genome-anchored representative proteins (ext001–ext010) were subjected to five orthogonal validation lines to exclude annotation or profile artifacts (Table S1).

Validation lines

(1) Profile-HMM re-scan. A profile HMM was built (HMMER `hmmbuild`, default parameters) from the 918-sequence CPR-C4 multiple sequence alignment used in the primary analysis. All ten external proteins were recovered with E-values of 4.0×10^{-54} to 9.4×10^{-76} and bit scores of 172.8–243.7.

(2) Catalytic triad. In a combined 71-sequence alignment (61 CPR representatives, including the 7PJO crystal sequence, plus the ten external homologs), the catalytic columns were verified by mapping the 7PJO crystal sequence onto the alignment: Cys (mmCIF `label_seq_id` 77; author numbering 61) at alignment column 130 (59/61 non-external sequences), His (label 109; auth 93) at column 166 (61/61), and Leu (label 131; auth 115) at column 222 (58/61). All ten external homologs carry Cys–His–Leu at these columns.

(3) Phylogenetic robustness. The combined 71-sequence alignment was used to infer trees with FastTree and, independently, with IQ-TREE 2 (model selection, 1,000 ultrafast bootstraps and 1,000 SH-aLRT replicates). Both methods place the ten external sequences interleaved among CPR lineages (non-monophyletic), consistent with multiple independent hori-

zontal acquisitions rather than a single transfer event, in agreement with the genome-resolved analysis in the main text.

(4) Taxonomic reclassification. The ten genomes were downloaded from NCBI and reclassified with GTDB-Tk v2.4.0 (`classify_wf`, GTDB R226, `--skip_ani_screen`). All ten were assigned to the same phylum as in the original screen (10/10 phylum-concordant). ext002 belongs to *Candidatus Tepidifontimicrobium thermophilum*, confirming the novel-genus record; ext006 is the genome anchoring the single-genome candidate phylum JADJOY01 (ANI 95.0 to its closest GTDB reference); ext001, after re-download of a truncated assembly, classified as p_Sysuimicrobiota, g_JAVHLZ01.

(5) Structure prediction. Full-length models were built with ColabFold 1.6.1 (AlphaFold2-ptm, three recycles, one model per target; CPU inference). Eight targets used MMseqs2-server MSAs; ext004 and ext006 lacked server MSAs (an invalid wildcard character in the query and a server-side gap, respectively) and were rescued by embedding their sequences into the 918-sequence CPR-C4 alignment (MAFFT `--add --keeplength`) and re-running with the resulting custom MSA. All ten models are high-confidence (pLDDT 90.6–96.7, pTM 0.86–0.92). Triad geometry was measured against the 7PJO crystal reference (author numbering Cys61–His93–Leu115; SG–NE2 = 5.28 Å, CysC α –HisC α = 6.28 Å, HisC α –LeuC α = 8.04 Å, CysC α –LeuC α = 10.0 Å). In every model, all four distances fall within ± 1.1 Å of the crystal, passing a reference-anchored criterion of ± 2.0 Å per distance (Table S1).

Controls and toolchain notes

Two single-sequence models built without an MSA (ext004, ext006) produced unusable folds (pLDDT 36.4 and 41.0) with distorted triad geometry, confirming that the high-confidence results depend on co-evolutionary signal rather than on template artifacts; these runs were retained as negative controls. GPU inference of ColabFold on the RTX 5080 (compute capability sm_120a) was abandoned after reproducible segmentation faults during JIT kernel execution, and the Boltz-2 predictor produced numerically invalid results on the same hardware (positive control on 7PJO itself: pLDDT 38.8); CPU ColabFold was therefore used for all reported models. The 7PJO catalytic residues are labelled Cys77/His109/Leu131 under mmCIF `label_seq_id` and Cys61/His93/Leu115 under author numbering; both conventions refer to the same three residues.

Summary

All five validation lines pass (Table S1). The ten cross-phylum homologs are genuine CPR-C4 family members that retain the intact catalytic triad and the canonical fold, extending the genome-verified (GTDB-anchored) distribution of the family beyond the CPR proper. All validation inputs and outputs (profile HMM, alignments, trees, GTDB-Tk summaries, structure models and per-model metrics) are retained in the project validation archive; scripts are available from the corresponding author upon reasonable request.

Supplementary Table S1. Orthogonal validation of the ten cross-phylum CPR-C4 homologs

Ext	Phylum (GTDB-Tk)	Genome	Protein	HMM <i>E</i>	Triad	pLDDT	pTM	SG-NE2 (Å)	His-Leu Cα (Å)
ext001	Sysuimicrobiota	GCA_031081675.1	JAVHLZ010000001.1_113	6.7×10^{-57}	C-H-L	95.3	0.91	6.25	8.35
ext002	Sysuimicrobiota ^a	GCA_031296405.1	JAVKKH010000007.1_48	4.0×10^{-54}	C-H-L	96.3	0.91	5.26	8.08
ext003	UBP6	GCA_017997475.1	JAGOEX010000148.1_5	1.6×10^{-72}	C-H-L	96.5	0.92	6.31	8.43
ext004	Krumholzibacteriota	GCA_016703265.1	JADJCK010000005.1_350	9.4×10^{-76}	C-H-L	91.2	0.87	5.86	8.47
ext005	Methylomirabilota	GCA_003220315.1	QHSA01000172.1_4	1.3×10^{-65}	C-H-L	95.9	0.90	6.36	8.47
ext006	JADJOY01 ^b	GCA_016713535.1	JADJOY010000041.1_47	2.8×10^{-70}	C-H-L	90.6	0.86	6.30	8.48
ext007	Moduliflexota	GCA_000739535.1	DF820463.1_223	5.2×10^{-65}	C-H-L	94.0	0.90	6.55	8.31
ext008	Marinisomatota	GCA_018002295.1	JAGNGI010000013.1_36	1.6×10^{-68}	C-H-L	96.6	0.92	6.33	8.10
ext009	Bdellovibrionota_F	GCA_001799195.1	MGU01000150.1_32	1.4×10^{-65}	C-H-L	96.7	0.92	6.05	8.56
ext010	Bdellovibrionota_B	GCA_026390555.1	JAPLFS010000072.1_15	2.7×10^{-63}	C-H-L	95.4	0.91	5.82	8.48

All ten genomes are phylum-concordant between the original screen and GTDB-Tk reclassification, and all ten models pass triad-geometry validation (all four diagnostic distances within ± 2.0 Å of the 7PJO crystal; maximum observed deviation 1.1 Å). ^aGenus *Candidatus* Tepidifontimicrobium (*Ca.* T. thermophilum). ^bSingle-genome candidate phylum in GTDB R226. Models for ext004 and ext006 use the rescued custom MSA (918 CPR-C4 sequences); the remaining eight use MMseqs2-server MSAs. SG-NE2: distance between the Cys sulfur and His Nε2 atoms. His-Leu Cα: distance between the His and Leu Cα atoms (crystal reference 8.04 Å).

Supplementary Table S2. First records of CPR-C4 homologs in non-CPR phyla (GTDB R226 screen)

Ext	Phylum	Genus (GTDB)	Genome	Ident. (%)	Cov. (%)	BLAST <i>E</i>
ext001	Sysuimicrobiota	JAVHLZ01	GCA_031081675.1	45.2	78	8.6×10^{-53}
ext002	Sysuimicrobiota	<i>Ca.</i> Tepidifontimicrobium ^a	GCA_031296405.1	50.6	68	1.9×10^{-47}
ext003	UBP6	JAGOEX01	GCA_017997475.1	44.8	93	1.0×10^{-47}
ext004	Krumholzibacteriota	CAINDZ01	GCA_016703265.1	39.7	96	1.4×10^{-45}
ext005	Methylomirabilota	20CM-2-70-11	GCA_003220315.1	40.3	81	2.2×10^{-40}
ext006	JADJOY01 ^b	JADJOY01	GCA_016713535.1	36.7	88	2.8×10^{-37}
ext007	Moduliflexota	Vecturithrix	GCA_000739535.1	36.2	97	7.5×10^{-37}
ext008	Marinisomatota	JAGNGI01	GCA_018002295.1	34.0	98	1.2×10^{-34}
ext009	Bdellovibrionota_F	YA12-FULL-61-11	GCA_001799195.1	33.7	91	7.6×10^{-33}
ext010	Bdellovibrionota_B	OMII01	GCA_026390555.1	33.8	91	8.1×10^{-26}

Identity, coverage and *E*-value are from BLASTp against the best CPR-C4 query; all ten proteins retain the catalytic triad (Table S1) and were recovered by the profile HMM. Each row is the first annotated CPR-C4 family record in the corresponding phylum in the GTDB R226 screen. ^aFirst family record in the candidate genus *Candidatus* Tepidifontimicrobium (NCBI taxonomy created 2024-11). ^bJADJOY01 is a single-genome candidate phylum in GTDB R226; GCA_016713535.1 is its sole representative genome, and this row records the family's presence rather than a new high-rank taxonomic claim.