Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The nitroplast import signal is a C-terminal, glycine- and alanine- rich, proline-punctuated element: a controlled reanalysis that separates real targeting grammar from cross-species composition artefact

Cem Boyraz

23050288@stu.omu.edu.tr

Ondokuz Mayıs University

Research Article

Keywords: nitroplast, UCYN-A, transit peptide, protein import, endosymbiosis, nitrogen fixation

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10503145/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

The nitroplast of the haptophyte *Braarudosphaera bigelowii* is a nitrogen-fixing organelle derived from the cyanobacterium UCYN-A, and roughly half its proteins are host-encoded and imported via a targeting element termed the UCYN-A transit peptide (uTP). Published descriptions compare haptophyte uTPs with plant chloroplast transit peptides (cTP), though this separation may reflect proteome-wide amino-acid composition rather than a targeting grammar. We reanalysed a 740-entry import catalogue (735 with retrievable sequence) against 1,759 curated crop cTPs under eight controls holding species composition constant. The cross-species contrast reaches AUC 0.9814; a within-species contrast — C-terminal 120-residue windows against internal windows of the same proteins — gives 0.9425 ± 0.0167, and a redundancy- and background-corrected paired-delta contrast at 40% identity gives 0.9445; both round to 0.94, differing by one-eighth of a standard deviation. Restricted to the 76 mass-spectrometry-curated proteins free of profile-HMM circularity, the estimate is 0.8986 ± 0.0553. The same positional contrast across nine organisms spanning four clades — including *Paulinella chromatophora*, which imports host proteins into an independently derived primary endosymbiont — falls in a narrow null band of 0.5579–0.6274. Catalogue proteins carry a median 254-residue C-terminal extension relative to haptophyte orthologues that non-targeted controls lack. Four enriched tetrapeptides — IPRL and PDFP, previously reported, and RLLP and PRLL, reported here — survive Benjamini–Hochberg correction and homology filtering. The signature survives every control constructed; its published effect size is inflated, and recognition by an import machinery remains untested.

Key Message

C-terminal nitroplast targeting windows separate from internal windows of the same proteins at AUC 0.94 and carry four proline-rich tetrapeptides near-absent from crop transit peptides.

Introduction

The marine cyanobacterium *Candidatus* Atelocyanobacterium thalassa (UCYN-A) has crossed from endosymbiont to organelle, joining eukaryotic organelles of bacterial origin (Coale et al. 2024; Massana 2024), associating with the coccolithophore-like haptophyte *Braarudosphaera bigelowii* (Thompson et al. 2012; Hagino et al. 2013; Suzuki et al. 2021). Its genome shows reduction typical of obligate endosymbionts (Tripp et al. 2010; Wernegreen 2002), having lost oxygenic photosynthesis and carbon fixation while retaining a complete *nif* operon, yielding a strict two-way exchange—fixed nitrogen out, fixed carbon, cofactors and amino acids in—constrained by cell-size exchange stoichiometry (Cornejo-Castillo et al. 2024; Sarkar et al. 2021). Growth and N2 fixation are light-driven and diel (Muñoz-Marín et al. 2019; Landa et al. 2021), spanning distinct oceanographic niches (Turk-Kubo et al. 2017; Zehr et al. 2016; Cornejo-Castillo et al. 2016).

Protein import is the property that elevates UCYN-A from endosymbiont to organelle. Coale et al. (2024) found much of the nitroplast's protein content is nucleus-encoded, translated in the host cytosol, and imported across the nitroplast envelope, identifying an N-terminal-independent targeting element—the uTP—on cargo proteins. Other lineages show import-competent endosymbionts: *Paulinella chromatophora*'s chromatophore imports hundreds of nucleus-encoded proteins via a bipartite transit peptide (Nowack and Grossman 2012; Singer et al. 2017; Oberleitner et al. 2022), while *Rhopalodia gibba*'s spheroid bodies show the same reductive trajectory but lack demonstrated import (Prechtl et al. 2004; Kneip et al. 2008). Import remains the operational criterion distinguishing organelle from symbiont, with the targeting signal's grammar its mechanistic core (Sørensen et al. 2025; Archibald 2015; Timmis et al. 2004).

cTPs are N-terminal, 30–100 residues, Ser/Thr-rich, positively charged, largely unstructured (Bruce 2000; Bruce 2001; von Heijne et al. 1989; Li and Chiu 2010; Jarvis and López-Juez 2013; Richardson and Schnell 2020; Nakai 2015; Kikuchi et al. 2018; Jin et al. 2022), recognised by the TOC–TIC translocon, organised into loosely defined functional blocks rather than a single consensus motif (Jeong et al. 2021; Zybailov et al. 2008), so prediction relies on machine learning (Emanuelsson et al. 1999; Emanuelsson et al. 2007; Almagro Armenteros et al. 2019; Sperschneider et al. 2017; Teufel et al. 2022).

Mitochondrial presequences form a second reference class with an overlapping but distinguishable grammar (Wiedemann and Pfanner 2017).

This paper addresses a methodological problem in how the uTP has been characterised. Comparing haptophyte uTP against plant cTP sequences confounds two co-varying factors: the targeting signal and each proteome's amino-acid composition, letting a classifier achieve near-perfect discrimination by learning "haptophyte versus plant," not targeting biology—exemplifying a failure mode across machine-learning-based science where evaluation design drives headline numbers (Kapoor and Narayanan 2023). The remedy: a within-species control comparing candidate signal against non-signal windows from the same proteins/proteome, cancelling composition by construction.

Beyond the descriptive, the question matters. Engineering nitrogen fixation into crops requires routing many nucleus-encoded proteins into a plant organelle at controlled stoichiometry under oxygen protection (Burén and Rubio 2018; Good 2018; Charpentier and Oldroyd 2010; Rogers and Oldroyd 2014). Functional nitrogenase components have been expressed in the mitochondrial matrix of yeast and plants (López-Torrejón et al. 2016; Allen et al. 2017), in plant chloroplasts (Eseverri et al. 2020; Ivleva et al. 2016), and assembled at controlled ratios via polyprotein designs (Yang et al. 2018; Yang et al. 2023), while conformational oxygen protection is now structurally resolved (Franke et al. 2025). The nitroplast remains the only known nitrogen-fixing organelle sustained by nucleus-encoded import, making the uTP the natural template for such a design's targeting layer — but only if the signal is targeting grammar rather than compositional artefact, at its true effect size. We therefore reanalysed the uTP catalogue under seven controls, each eliminating one alternative explanation, reporting results at every stage of confound removal rather than the single most favourable one.

Materials and Methods

Import catalogue and sequence set

The nitroplast import catalogue (Coale et al., 2024) has 740 entries: 81 Tier 1 (mass-spectrometry-curated), 659 Tier 2 (profile hidden Markov model hits). Sequences existed for 735/740 (76 Tier 1, 659 Tier 2); median full-length: 665 aa (Tier 1), 592 aa (Tier 2) (Table 1). Four sequences under 120 aa were used full-length wherever a 120-residue window is specified.

Tier assignment came from the compiled catalogue, not re-derived from primary data. The Dryad deposition (doi:10.5061/dryad.2z34tmptf) holding the primary curated import list is *verified to exist but not retrieved*, as the repository's web application firewall blocked repeated programmatic downloads, leaving tier labels unverified against primary files. Other inputs were retrieved and enumerated in the provenance table. PXD047956 was confirmed present but offers only raw spectra, unusable for peptide-level re-curation without reprocessing.

Terminal position of the uTP: provenance

We treat the uTP as a C-terminal element. Coale et al. (2024), the import catalogue source, reported via proteomic identification of host-encoded nitroplast proteins that imported proteins carry a ~ 120-residue C-terminal extension relative to orthologues, with conserved motifs. We accessed only the abstract, so details are secondhand. Sørensen et al. (2025) restate, not source, this placement.

An independent computational re-analysis by TU-Delft iGEM 2024 (NitroBLAST, https://2024.igem.wiki/tu-delft/) confirms: among 368 UCYN-A-enriched host proteins, a conserved C-terminal region appears, 206 exceed 60% identity, eight motifs emerge. Not peer-reviewed, not in our reference list, but genuinely independent.

We do not rely on citation for this placement, since the primary text is unreadable to us. Control 8 directly measures the extension from public sequence data, providing this work's own evidence for the C-terminal position; earlier drafts describing an N-terminal element are superseded.

Control 8: ortholog terminal length-shift test

To test the C-terminal placement without relying on the literature, we asked whether catalogue proteins are longer than their orthologues specifically at the C-terminus. We searched 735 full-length catalogue proteins (median 597 aa) against four ortholog databases—*Emiliania huxleyi* (38,554 proteins), *Chrysochromulina tobinii* (16,765), *Prymnesium parvum* (24,907), and UCYN-A (2,445 proteins, ALOHA and SIO64986 concatenated)—via mmseqs2 easy-search (Steinegger and Söding 2017; -e 1e-5 --max-seqs 20), keeping the lowest-e-value hit; C-terminal overhang = (qlen − qend) − (tlen − tend), N-terminal overhang = (qstart − 1) − (tstart − 1), positive meaning query extends further.

Two negative controls followed the same protocol: 900 random *E. huxleyi* proteins ≥ 100 aa (random.seed(7)), non-cargo, searched identically; and a length-matched subsample, since overhang statistics are length-sensitive. Cargo-versus-control used Mann–Whitney U; cargo C-versus-N overhang used Wilcoxon signed-rank test.

The definitive negative control would be non-targeted *B. bigelowii* proteins from the same organism. No host transcriptome-wide protein set is public (see below), so a related haptophyte is the closest available substitute; the substitution is a limitation, not a design choice.

Window definitions

Three window types were used throughout, all 120 residues:

1. uTP C-terminal window — last 120 residues of each catalogue protein (n = 735).

2. Internal window — central 120 residues, excluding first and last 120 residues; requires ≥ 360 aa, met by 614 of 735 catalogue proteins; 121 too short, excluded from every within-species contrast.

3. N-terminal window — first 120 residues, used only for the positional control.

Internal windows for the plant reference set were defined identically, excluding the annotated cTP region so none overlap a targeting signal; 1,052/1,759 plant proteins are long enough (Table S1). Full-length UniProt sequences for all 1,759 accessions matched annotated transit-peptide coordinates (1,759/1,759 consistent, zero length mismatches).

The 120-residue window was not adopted on convention alone but tested explicitly. It was fixed a priori from the import literature; ortholog-based measurement of the cargo C-terminal extension (Control 8) gave a median of 239–256 aa. To test whether this choice drives conclusions, core analyses—the within-species contrast for all 735 proteins and Tier-1 alone, the cross-species contrast against crop cTPs, both positional nulls, the eleven-feature Mann–Whitney panel, and the four-tetrapeptide Fisher/BH panel—were repeated at eight windows (40, 60, 80, 120, 160, 200, 250, 300 aa), including on the ≥ 900-aa subset alone with cohort membership held constant. A protein-specific variable window, using each protein's corrected C-overhang from Control 8 (c_overhang − (tlen − tend), floored at 20 aa; n = 533, median 246 aa), was also tested against two internal-window definitions: central W residues (L ≥ 3W) and W residues immediately upstream of the C-terminal window. Motif positions were recorded as distance from the C-terminus to the last residue of each occurrence, counting all occurrences. Sources: phase5_auc_sweep.csv, phase5_window_retention.csv, phase5_variable_window_auc.csv, phase5_motif_position_summary.csv. The 120-aa window was retained as primary and 160 aa adopted as a sensitivity check: the internal-window control requires L ≥ 3W and collapses above 160 aa,

falling from 614 paired proteins at 120 aa to 535 at 160 aa, 362 at 200 aa and 164 at 250 aa (Tier-1 pairs 71, 56, 46, 25), so no window beyond 160 aa retains a usable non-circular control.

Crop chloroplast transit peptide reference

The cTP reference set comprises 1,759 UniProtKB reviewed entries with curated chloroplast transit-peptide annotations from *Arabidopsis thaliana* (n = 1,275), *Oryza sativa* (n = 353), *Zea mays* (n = 100) and *Triticum aestivum* (n = 31) (UniProt Consortium 2023); median annotated cTP length: 52 aa (range 10–143).

Additional targeting-peptide classes

For the multi-class control, reviewed UniProt entries for the same four species with ft_transit and ft_signal annotations were split into chloroplast transit peptides (n = 1,709), mitochondrial transit peptides (n = 662) and secretory signal peptides (n = 3,159), excluding 42 entries with other transit annotations. Median lengths were 52, 36 and 24 aa; peptides shorter than 10 aa were discarded.

Haptophyte negative control

The ideal negative control for a haptophyte targeting signal is the *B. bigelowii* host nuclear proteome. It is not publicly available. Programmatic searches of NCBI Assembly, BioProject, SRA, Protein, nuccore (mRNA), TSA, ENA and UniProt found no nuclear assembly and no transcriptome shotgun assembly for taxid 373042; the 1,194 NCBI Protein records are organelle-derived; the only *Braarudosphaera*-associated UniProt proteome (UP001319803) belongs to the endosymbiont itself (taxid 1285375, 1,185 proteins); and BioProjects PRJEB107556, PRJEB107557 and PRJNA1132261 hold only raw reads and symbiont assemblies. As a surrogate we used *Emiliania huxleyi* CCMP1516 (UniProt proteome UP000013827, taxid 280463, 35,617 proteins; Read et al. 2013), a same-class haptophyte lacking a nitroplast, yielding 31,626 C-terminal 120-residue windows and 12,302 internal windows (12,302 proteins ≥ 360 aa; median length 279 aa).

Multi-organism positional null panel

No single negative control can distinguish whether low C-terminal-versus-internal AUC reflects a generic proteome property or the specific absence of an endosymbiont in *E. huxleyi*. We therefore repeated the positional null in nine organisms spanning four eukaryotic clades and three symbiotic states (Table 4): *Arabidopsis thaliana* (UniProt UP000006548, 39,273 proteins), *Epithemia clementina* (GCA_051400955.1, 26,505; nitrogen-fixing spheroid body, Nakayama et al. 2014), *Phaeodactylum tricornutum* (GCF_000150955.2, 10,408), *Guillardia theta* (GCF_000315625.1, 24,822), *Chrysochromulina tobinii* (GCA_001275005.1, 16,765), *Emiliania huxleyi* (GCF_000372725.1, 38,554), *Thalassiosira pseudonana* (GCF_000149405.2, 11,673), *Prymnesium parvum* (GCA_041296205.1, 24,907), and *Paulinella chromatophora*.

Paulinella chromatophora holds the only other independently derived primary endosymbiont with demonstrated host-encoded protein import via an N-terminal bipartite signal (Nowack and Grossman 2012; Singer et al. 2017; Oberleitner et al. 2022). No annotated nuclear protein set is public; we translated TSA assembly GEZN01 (BioProject PRJNA311736, 60,586 transcripts; Nowack et al. 2016), keeping complete ORFs ≥ 100 aa: 16,423 proteins, 5,822 ≥ 360 aa—an ORF-translated proxy, not a curated gene set.

Window geometry: identical across nine organisms—proteins ≥ 360 aa yield C-terminal (last 120 aa), N-terminal (first 120 aa), internal (central 120 aa, disjoint from termini) windows; plants exclude transit-peptide region from internal.

Classification: StandardScaler+logistic regression (max_iter = 5000), ROC AUC via RepeatedStratifiedKFold (5×5, random_state = 42). Proteomes: subsampled ≤ 8,000 paired windows (numpy.random.default_rng(42)).

Two source sets for *E. huxleyi* appear in this manuscript and must not be conflated. The haptophyte negative control above uses UniProt UP000013827 (35,617 proteins); the panel uses NCBI RefSeq GCF_000372725.1 (38,554 proteins). We report both AUCs separately rather than averaging them.

Sequence features

Eleven features per window: net charge per 100 residues (Lys + Arg−Asp − Glu), mean Kyte–Doolittle hydrophobicity, and fractions of Ser + Thr, hydrophobic (AVILMFWC), acidic (DE), basic (KRH), aromatic (FWY), Pro, Gly, Ala, Arg, excluding X, B, Z, U, stop symbols from denominators. Matrix: 2,494 rows (735 uTP + 1,759 cTP) for the primary comparison, extended for all control sets.

Statistics

Feature comparisons applied two-sided Mann–Whitney U tests with Bonferroni correction across the 11 features, reporting CLES (equivalent to single-feature ROC AUC) alongside each test. For tetrapeptide (4-mer) analysis, each 4-mer was counted at most once per sequence (presence/absence); every 4-mer in at least one set was assessed via two-sided Fisher's exact test, with Benjamini–Hochberg FDR control across all tested 4-mers per comparison layer (Benjamini and Hochberg 1995). Log2 enrichment ratios used a 0.5 pseudocount; underflowing Fisher p-values are reported as < 1e-300.

Classification and cross-validation

All classifiers are L2-penalised logistic regressions with balanced class weights, in a pipeline starting with StandardScaler, preventing scaling leakage across folds. RepeatedStratifiedKFold (5 splits, 5 repeats, 25 folds, random_state = 42) yields mean ROC AUC ± standard deviation. Multi-class model: one multinomial logistic regression over four peptide classes (max_iter = 3000), same scheme; confusion matrix row-normalised, accuracy as class-size-weighted diagonal mean. Python: NumPy (Harris et al. 2020), SciPy (Virtanen et al. 2020), pandas, scikit-learn 1.8.

Species-blocked cross-validation

Cross-species generalisation testing: The negative (plant) class was split by species into four train/test blocks: Arabidopsis+Oryza→Triticum + Zea, Triticum + Zea→Arabidopsis+Oryza, dicot (Arabidopsis)→monocot (Oryza+Triticum + Zea), monocot→dicot. Lacking species structure, the positive (haptophyte) class was randomly split 50/50 (random_state = 42) per block, avoiding overlap. Same protocol applied to the background-floor contrast, comparing species-blocked penalty to its own baseline, not the random-fold number.

Homology filtering

Redundancy was removed using mmseqs2 easy-cluster (Steinegger and Söding 2017) at --min-seq-id 0.4 or 0.6, -c 0.8 --cov-mode 1 -s 7, keeping one representative per cluster; CD-HIT (Fu et al. 2012), run identically, gave consistent spot-check results. Statistics and classification were recomputed on both representative sets, filtering independently within each window set so one set's redundancy structure wasn't imposed on another.

Profile-HMM provenance and its consequence

The catalogue's Tier 2 entries (659 of 740, 89%) came from searching the host transcriptome with a profile HMM trained, per the compiling authors' own description, on the curated positive set. Tier 2 is not independent evidence of uTP sequence properties, partly restating the selecting model. The 76 sequence-bearing Tier 1 proteins provide primary evidence, Tier 2 a consistency check, quantifying circularity-driven inflation via tier comparison. Methods: HMMER (Eddy 2011); motifs: MEME (Bailey et al. 2009).

Reference verification

Every reference cited in this manuscript was resolved at the CrossRef DOI endpoint and its first author, year, journal and title read back from the returned record; the verification table is provided as supplementary material. No DOI is cited that did not resolve.

Declaration of AI use

Computational analysis, figures and tables were produced in an agentic scientific-computing environment (Claude Science, Anthropic) directed by the author, who specified controls, reviewed every intermediate result and verified every reported number against its source table; the author takes responsibility for content, analytical choices and conclusions.

Results

The published contrast reproduces exactly, and is species-confounded

Recomputing the published comparison from the sequences — 735 uTP C-terminal windows against 1,759 crop cTPs, eleven features, 25-fold repeated stratified cross-validation — reproduces near-perfect separation: AUC 0.9814 ± 0.0055 (fold minimum 0.9648, fold maximum 0.9898 across the 25 folds; Fig. S1). Ten of the eleven features differ at Bonferroni-corrected significance; only Arg fraction does not ($z = 2.04$, $p = 0.042$, $p_{Bonferroni} = 0.459$). Largest effects: acidic fraction (CLES 0.911, uTP-enriched), Gly fraction (0.888), Ala fraction (0.790); oppositely, Ser + Thr fraction (0.103) and net charge (0.106) — uTPs are less charged (median 3.33/100 aa vs cTPs' 11.03) and less Ser/Thr-rich (0.117 vs 0.238), richer in Gly (0.108 vs 0.036), Ala (0.150 vs 0.077), acidic residues (0.075 vs 0.020) (Fig. 2; Table S1).

That result is not evidence of a targeting grammar. Comparing haptophyte against plant sequences lets a classifier match the same AUC via proteome composition alone. The decisive test: does separation survive with species held constant? Control 1 (Fig. 3) reports the full 2×2 design.

The within-species contrast is the honest headline: uTP C-terminal windows against internal windows of the *same 735 proteins* (n = 614 internal windows from proteins ≥ 360 aa) give AUC 0.9425 ± 0.0167. Species composition cancels by construction, and the separation remains large.

The within-plant contrast is the positive control: annotated cTPs against internal windows of the same plant proteins give AUC 0.9978 ± 0.0015 — a genuine targeting signal is near-perfectly separable from its own protein body, which validates the window design.

The background floor is the crux: haptophyte vs. plant internal windows, no targeting signal, yield AUC 0.9186 ± 0.0086—nearing published 0.9814; cross-species contrast exceeds floor by only 0.063 AUC units. Floor doesn't apply to within-species contrast, lacking cross-species composition; compare instead to positional nulls at 0.588–0.687.

The background-corrected contrast classifies paired within-protein deltas (terminal minus internal, computed separately for each organism, so that species-specific composition subtracts out) gives AUC 0.9305 ± 0.0103.

0.9814 must not be reported as the uTP signal's effect size. The honest within-species figure is 0.94: 0.9425 by this contrast, 0.9445 after removing redundancy and positional background (see cascade below), agreeing well within either's uncertainty.

Positional nulls bound the noise floor, in three systems and then in nine organisms

A within-species AUC of 0.9425 is only informative against a null for how much two different windows of the same proteome differ from each other for reasons unrelated to targeting. We established three such nulls (Fig. 4; Fig. 6).

NULL-1 — two disjoint internal windows, same haptophyte protein (n = 535 pairs, ≥ 480 aa): AUC 0.6866 ± 0.0301. NULL-2 — plant C-terminal vs internal windows (n = 1,052), C-termini not targeting signals: AUC 0.5922 ± 0.0254. NULL-3, strongest, same contrast in *Emiliania huxleyi*, lacking a nitroplast (31,626 C-terminal vs 12,302 internal windows, negatives subsampled to 5,000): AUC 0.5878 ± 0.0065. Lacking a nitroplast, position carries almost no signal; the uTP contrast exceeds that floor by 0.35 AUC units.

The *E. huxleyi* comparison calibrates the between-genus composition effect directly. uTP C-terminal windows vs. *E. huxleyi* C-terminal windows give AUC 0.8645 ± 0.0154, and vs. *E. huxleyi* internal windows 0.8693 ± 0.0123 — yet haptophyte-internal vs. *E. huxleyi*-internal, lacking any targeting signal, already gives 0.7834 ± 0.0202. The genus-level composition difference between *B. bigelowii* and *E. huxleyi* is substantial, so the within-protein design, not the between-organism design, carries the argument.

A referee might object that NULL-3 proves too little: *E. huxleyi* hosts no endosymbiont, so a flat C-terminal-versus-internal contrast could reflect absent organellar import rather than absence of a positional artefact. We tested this across nine organisms spanning four clades and three symbiotic states (Table 4, Fig. 7), all falling in a narrow band: *A. thaliana* 0.5579 ± 0.0079, *E. clementina* 0.5689 ± 0.0081, *P. tricornutum* 0.5738 ± 0.0081, *G. theta* 0.5756 ± 0.0094, *C. tobinii* 0.5776 ± 0.0100, *E. huxleyi* 0.5795 ± 0.0072, *P. chromatophora* 0.5836 ± 0.0111, *T. pseudonana* 0.5860 ± 0.0094 and *P. parvum* 0.6274 ± 0.0084; the panel's 0.070 AUC-unit spread is exceeded fivefold by the gap to the uTP contrast.

The decisive entry is *Paulinella chromatophora*. It carries an independently derived primary endosymbiont with demonstrable host-protein import, yet sits at 0.5836 — at the null, indistinguishable from non-symbiotic organisms. An organellar import pathway alone doesn't produce C-terminal asymmetry: the *Paulinella* signal is N-terminal, confirming asymmetry in *B. bigelowii* tracks targeting-element position, not endosymbiosis per se. The nitrogen-fixing diatom symbiosis *E. clementina*, at 0.5689, points the same way.

N-terminal windows behave differently. N-terminal-versus-internal AUCs exceed C-terminal ones (*T. pseudonana* 0.7120, *P. tricornutum* 0.6797, *E. clementina* 0.6529, *A. thaliana* 0.6451, *E. huxleyi* 0.6296), expected given initiator-methionine context, signal/transit peptides, and N-terminal disorder—favoring a C-terminal positional null, since N-terminal analysis starts noisier.

Note that *E. huxleyi* (RefSeq GCF_000372725.1, 38,554 proteins, AUC 0.5795) and NULL-3 (UniProt UP000013827, 35,617 proteins, AUC 0.5878) derive from different source sets of the same organism; both reported, agreeing within 0.01 AUC, showing protein-set choice matters little here.

The uTP is C-terminal, and it is a genuine extension

Comparing C- and N-terminal 120-residue windows across the same 735 proteins isolates terminal asymmetry, with no cross-species component. Nine of eleven features differ after Bonferroni correction: C-terminus enriched in Gly (CLES 0.702), net charge (0.691), Ala (0.690), Pro (0.663); depleted in acidic residues (0.185), aromatics (0.340) (Fig. S2).

It bears noting this is an internal comparison and cannot alone establish the C-terminal window as an addition. Control 8 tests whether catalogue proteins run longer than orthologues at the C-terminus (Table 5, Fig. 8). Of 735 catalogue proteins, orthologs were found in 518 *E. huxleyi*, 438 *P. parvum*, 394 *C. tobinii*, 27 UCYN-A. Overhang is large: median 254.0 aa (*E. huxleyi*), 257.5 (*P. parvum*), 255.5 (*C. tobinii*), 291.0 (UCYN-A), versus 17–32 aa N-terminally (Wilcoxon, $p = 1.15 \times 10^{-66}$, *E. huxleyi*). 85.0–94.3% of cargo proteins carry overhang exceeding 100 aa.

The 900-protein non-targeted *E. huxleyi* control shows no shift: median C-terminal overhang − 4.0 aa (*C. tobinii*), − 7.0 (*P. parvum*), − 2.0 (UCYN-A). Since cargo queries are longer than random (median 651 aa vs 331 aa) and overhang scales with length, we repeated the contrast against a length-matched control. The shift survives: cargo 255.5 aa vs 1.0 aa (*C. tobinii*, $p = 3.5 \times 10^{-61}$, CLES 0.824), 257.5 vs 0.0 (*P. parvum*, $p = 1.2 \times 10^{-109}$, CLES 0.911), 291.0 vs 16.0 (UCYN-A, $p = 1.9 \times 10^{-9}$, CLES 0.895). N-terminal contrasts are weaker (CLES 0.556, 0.636, 0.387), so asymmetry is C-terminus-specific.

The magnitude does not match the figure in circulation. Secondary literature cites ~ 120 residues; we measure a median of 254–291 aa across 394–518 proteins with detectable orthologs. We report, not reconcile, this discrepancy, since we could not read the primary text and cannot confirm both numbers are measured identically — a conserved 120-aa motif block within a longer non-homologous extension could yield both figures without contradiction. Two caveats: cargo hits are more diverged than controls (median identity 0.30–0.39 versus 0.35–0.44), inflating the apparent overhang, and the negative control is a related haptophyte, not a non-targeted *B. bigelowii* protein set, which does not exist publicly.

With those caveats stated, the direction and effect size are clear: catalogue proteins carry a substantial C-terminal extension their haptophyte orthologues lack, and this manuscript's uTP placement rests on that measurement, not on an unverifiable citation.

Four tetrapeptides carry the signal, two of them not previously reported for this system

Presence/absence testing of every observed 4-mer, with BH correction, was performed at three levels of stringency (Fig. 5; Table 3).

Against crop cTPs (6,222 4-mers tested; 2,758 at FDR < 0.05, 2,117 at FDR < 0.01), both known motifs are enriched: IPRL, 162/735 uTP windows (22.04%) vs 0/1,759 cTPs, OR 997.1, p = 1.05e-92, FDR = 1.63e-89; PDFP, 167 (22.72%) vs 1 (0.057%), OR 345.4, p = 1.18e-93, FDR = 2.45e-90. Two unreported tetrapeptides are stronger: RLLP, 253 (34.42%) vs 12 (0.68%), OR 73.4, p = 1.18e-131, FDR = 7.31e-128; PRLL, 235 (31.97%) vs 10 (0.57%), OR 78.4, p = 1.27e-122, FDR = 3.94e-119. Both absent from Coale et al. 2024d rensen et al. 2025; no systematic literature search performed, so novelty bounded by those sources.

Against internal windows of the same proteins (composition-controlled; 4,786 4-mers tested, 630 at FDR < 0.05), all four survive: IPRL 162 (22.04%) vs 2 (0.33%), OR 69.4, FDR = 1.80e-40; PDFP 167 vs 3 (0.49%), OR 51.5, FDR = 1.10e-40; RLLP 253 vs 4 (0.65%), OR 71.3, FDR = 5.41e-67; PRLL 235 vs 2, OR 115.3, FDR = 2.81e-64. Composition cannot explain this.

Against a haptophyte proteome without a nitroplast the contrast holds. In *E. huxleyi* C-terminal windows the four motifs occur at 0.130%, 0.051%, 0.772% and 0.648%, versus 22.04%, 22.72%, 34.42% and 31.97% in uTP windows — odds ratios 217.8, 580.9, 67.5 and 72.0, Fisher p-values 2.7e-232 to 4.3e-291. Not a haptophyte C-terminal feature; a nitroplast-cargo feature.

After homology filtering the enrichment persists. At 60% identity (644 uTP representatives) prevalences are essentially unchanged: IPRL 22.20%, PDFP 22.52%, RLLP 35.25%, PRLL 32.76%, all at FDR < 1e-35 versus internal representatives. At 40% identity (511 representatives) they attenuate but hold: IPRL 15.46% (OR 44.9 vs internal, FDR = 1.20e-18), PDFP 17.61% (OR 34.9, FDR = 1.56e-20), RLLP 29.16% (OR 67.2, FDR = 2.52e-40), PRLL 25.64% (OR 84.6, FDR = 1.49e-35). Against cTP representatives at 40% identity, IPRL stays completely absent (OR infinite) and PDFP occurs in one sequence (OR 332.0). The attenuation from ~ 22% to ~ 15–18% for IPRL and PDFP reflects removed paralogue families sharing a motif; that a quarter to a third of *non-redundant* uTP windows still carry RLLP or PRLL is the meaningful residual.

Because the four tetrapeptides sit a median 98–114 residues from the C-terminus, the 120-aa window captures only 59.0–72.0% of their occurrences and every prevalence above is therefore a lower bound. Recomputed at 160 aa, which

captures 89.5–92.6%, the prevalences are IPRL 221/735 (30.07%), PDFP 271 (36.87%), RLLP 326 (44.35%) and PRLL 320 (43.54%), against 0.37–0.93% in 160-aa internal windows of the same proteins (OR 114.6, 155.6, 84.5 and 205.5; FDR = 5.07e-56, 2.55e-71, 5.74e-85 and 1.32e-87) and 0–0.68% in crop cTPs (FDR = 2.04e-129 to 6.00e-179). In the 76 Tier-1 proteins alone the 160-aa prevalences are 39.47%, 32.89%, 44.74% and 47.37%, all enriched over Tier-1 internal windows at FDR ≤ 3.02e-07. Extending to 250 aa adds at most six further percentage points, so 160 aa is treated as the motif-complete window throughout (Table S8).

Weakness 1: 89% of the catalogue is a profile-HMM self-hit, and it inflates the signal

659 of the 740 catalogue entries (89.1%) were recovered by searching the host transcriptome with a profile HMM trained on the curated positive set. Statistics computed over the whole catalogue therefore partly restate the selection model. Quantifying this directly (Fig. S3; Table 2):

Tier 1 (76 MS-curated proteins) vs cTP: AUC 0.9697 ± 0.0238; within-species (76 vs 71 windows): AUC 0.8986 ± 0.0553 — most conservative honest number here. Tier 2 (659 pHMM hits) vs cTP: AUC 0.9824 ± 0.0051; within-species: 0.9465 ± 0.0129.

Tier-2 exceeding Tier-1 in both designs signals circularity, not better data. Tier 1 and Tier 2 remain statistically indistinguishable on all eleven features (0/11 significant after Bonferroni; CLES 0.453–0.572). Tier 1 still differs from cTP on 7/11 features: IPRL 22/76 (28.95%) vs 1/71 (1.41%), OR 28.5; PDFP 11 (14.47%) vs 0, p = 7.0e-04; RLLP 27 (35.53%) vs 1, OR 38.6; PRLL 28 (36.84%) vs 1, OR 40.8. Signal exists in the non-circular subset; whole-catalogue effect size is inflated by ~ 0.04–0.05 AUC units.

Weakness 2: generalisation to a cereal is the weakest link

Random-fold cross-validation mixes plant species across folds, so it never tests generalisation to an unseen species. Species-blocked cross-validation drops uTP-versus-cTP AUC from 0.9814 (random folds) to a blocked mean of 0.9253 (−0.0561). Blocks vary: Arabidopsis+Oryza→Triticum + Zea 0.9540, Triticum + Zea→Arabidopsis+Oryza 0.9571, monocot→dicot 0.9594, but dicot→monocot falls to 0.8307 (Fig. S4). The background floor drops less (0.9186→0.9004, − 0.0182), concentrating the penalty in the signal-bearing contrast. Each blocked value is a single point estimate, not a resampled mean, so 0.8307 is one observation, not a calibrated estimate. Engineering uses of a uTP-derived rule target cereals, and dicot-trained→monocot-tested is the worst-performing transfer.

Weakness 3: the ideal negative control does not exist publicly

The *B. bigelowii* host nuclear proteome would be the correct negative set: same organism, same proteome, no import competence. It is not public. No nuclear assembly, TSA, or host-derived protein records exist for taxid 373042; available BioProjects contain only raw reads and symbiont assemblies. The strongest negative control buildable is a *different* haptophyte (*E. huxleyi*, UP000013827), introducing the genus-level composition offset (internal-vs-internal AUC 0.7834). This is a limitation, not a solved control. Once a *B. bigelowii* nuclear assembly is released, the most informative repeat is the within-organism, non-cargo negative set: C-terminal windows of host proteins not in the import catalogue.

The uTP is distinct from mitochondrial and secretory targeting, not only from cTPs

If the uTP were simply a degenerate organellar targeting peptide, it should be hard to separate from other organellar signals. It is not (Fig. S5; Table 2). Against 662 mitochondrial transit peptides, AUC 0.9872 ± 0.0066 (all 11 features differing, Bonferroni); against 3,159 secretory signal peptides, 0.9987 ± 0.0009 (10/11 features); against chloroplast set

(n = 1,709), 0.9803 ± 0.0061. Compositionally, presequences are most cationic (median charge 13.9/100 aa vs 11.0 cTP, 3.3 uTP); signal peptides are hydrophobic (mean 1.43 vs − 0.11 uTP), lacking acidic/Arg residues.

A single four-class multinomial model over uTP, cTP, mitochondrial TP and signal peptide reaches 86.98% class-size-weighted accuracy. uTP recall is 0.944 and signal-peptide recall 0.961, but cTP (0.735) and mitochondrial TP (0.713) are confused with each other — 19.5% of cTPs are called mitochondrial and 22.8% of mitochondrial TPs are called cTP — while only 5.4% of cTPs and 3.9% of mitochondrial TPs are called uTP. The two canonical plant organellar signals resemble each other more than either resembles the nitroplast element. This compositional separability is not mechanistic separateness: differing amino-acid usage can still be read by overlapping receptors.

The 120-residue window is a conservative choice, not an arbitrary one

The within-species AUC is flat across a five-fold range of window length, while the cross-species AUC is not. Between 40 and 200 aa the within-species estimate spans 0.9290–0.9588 (0.9489 ± 0.0090 at 40 aa, peak 0.9588 ± 0.0114 at 60 aa, 0.9425 ± 0.0168 at 120 aa, 0.9408 ± 0.0128 at 160 aa), a spread of 0.0298. The ≥ 900-aa fixed-cohort sweep reproduces this shape (0.9228 ± 0.0475 at 40 aa, 0.8986 ± 0.0598 at 120 aa, 0.8447 ± 0.0634 at 250 aa), so decline beyond 200 aa is not purely cohort attrition. Both positional nulls stay flat and low (*E. huxleyi* 0.5612–0.5884; *Paulinella* 0.5351–0.5887). The cross-species contrast instead rises monotonically from 0.9256 ± 0.0109 at 40 aa to 0.9960 ± 0.0016 at 300 aa: longer nitroplast windows accumulate more haptophyte-average composition relative to a median 52-aa plant cTP—a length confound, and another reason not to treat 0.98 as an effect size (Fig. S7a–c).

The internal-window control, not the signal, sets the upper bound on window length.** Because L ≥ 3W, paired proteins fall from 614 (120 aa) to 535 (160 aa), 362 (200 aa) and 164 (250 aa); Tier-1 pairs fall from 71 to 56, 46 and 25. A 250-aa window trades the only non-circular control for an underpowered one. Motifs are not confined to the last 120 residues: IPRL, PDFP, RLLP and PRLL show median C-terminal distances of 98, 114, 98 and 99 residues (upper quartiles 122–135 aa), so 120 aa captures only 59.0–72.0% of occurrences versus 89.5–92.6% at 160 aa, with 250 aa adding at most 6 points. At 160 aa prevalences rise to 30.1% (IPRL), 36.9% (PDFP), 44.4% (RLLP) and 43.5% (PRLL) against 0.37–0.93% internally (all FDR < 1×10^{-15}). At 40–60 aa enrichment reverses (IPRL 0.41% vs 1.50% internal, OR 0.27, FDR 0.036), confirming proximal, not terminal, signal. The per-protein window does not outperform the fixed one: adjacent-internal gives 0.9417 ± 0.0169 vs 0.9271 ± 0.0164 (120 aa) and 0.9491 ± 0.0166 (250 aa, n = 348); under centre-internal it is worst of three (0.9164 ± 0.0440 vs 0.9652 ± 0.0273, n = 83). We retain 120 aa as primary, 160 aa as sensitivity check, treating 120-aa prevalences as lower bounds (Fig. S7d–f; Tables S7, S8).

Synthesis: a confound-removal cascade

Reading the controls as a cascade (Fig. 6) makes the arithmetic of the reanalysis explicit. The published cross-species figure is 0.9814. Removing the species confound (within-protein design) gives 0.9425 ± 0.0167; removing sequence redundancy and positional background (paired-delta contrast, 40%-identity representatives) gives 0.9445 (0.9436 at 60%). These agree: they differ by 0.0020, one-eighth of the 0.0167 standard deviation, well within measurement uncertainty, so neither is 'better.' We quote a single headline of 0.94, reporting both anchors: 0.9425 (simplest design) and 0.9445 (additionally removing redundancy and background). Removing profile-HMM circularity via the 76 MS-curated proteins gives 0.8986 ± 0.0553, the most conservative estimate. All sit far above the positional null (nine organisms: 0.5579–0.6274; three within-system nulls: 0.5878–0.6866). The uTP signal is thus distinguishable from every null here, though its published effect size is inflated: ~0.04 AUC from species composition, ~ 0.045 from HMM circularity, up to 0.11 in the cereal-relevant transfer setting.

Discussion

What survives, and at what size

The central finding is that the nitroplast targeting signature survives every control removing species composition, while its published magnitude does not. Cross-species AUC of 0.9814 drops to 0.9186 for internal, non-targeting windows of the same two proteomes, since separating a haptophyte from a plant needs no targeting information. Within nitroplast-targeted proteins, separation is 0.9425, and on the mass-spectrometry-curated subset, 0.8986. Against positional nulls of 0.5878, 0.5922 and 0.6866, these are decisively above noise. The defensible headline is 0.9425, or 0.8986 curated—not 0.98, reflecting one representation (L2-penalised logistic regression, eleven amino-acid composition features). The controls fix the ordering, not the number.

This distinction is not pedantry. An effect size scales downstream engineering decisions: a rule discriminating at 0.98 invites more trust than one at 0.90 on the only non-circular subset available. AUC measures ranking between balanced classes under cross-validation, not false-positive rate in a proteome-wide scan, which depends on operating threshold and true cargo prevalence—neither modelled here. The point concerns confidence a design decision should carry, not a precision forecast. The reanalysis neither weakens nor tests the claim that the nitroplast imports host proteins via a dedicated C-terminal element; it shows only that the signature cannot arise from species composition, leaving import evidence where the primary work placed it (Coale et al. 2024; Sørensen et al. 2025). What changes is the confidence interval attached to any uTP-derived predictor.

The motif grammar is the transferable part

The most robust component is the tetrapeptide vocabulary, not aggregate composition. Four 4-mers — IPRL, PDFP, RLLP and PRLL, not reported for this system in the two primary sources we checked — occur in 15–35% of uTP windows, under 1% of internal windows from the same proteins, 0–0.68% of 1,759 crop cTPs, and 0.05–0.8% of 31,626 C-terminal windows from a nitroplast-free haptophyte. They survive Benjamini–Hochberg correction across thousands of simultaneously tested 4-mers, survive homology filtering to 40% identity, and replicate in the 76-protein curated subset. RLLP and PRLL individually outperform the two previously reported motifs in every layer tested. Novelty is bounded by Coale et al. (2024)d rensen et al. (2025), absent a systematic literature search.

The four motifs share a chemistry: proline adjacent to leucine/isoleucine and arginine. The composition — glycine- and alanine-rich, weakly cationic, low in Ser/Thr, proline-punctuated — suggests a conformationally constrained, poorly folded segment rather than an amphipathic helix or hydrophobic core; this is an inference from composition, not a structural result, as no disorder/structure prediction was run (see Limitations). This is not the grammar of a cTP, whose recognition depends on a Ser/Thr-rich, strongly cationic, largely disordered N-terminal extension read by TOC receptors (Bruce 2000; Li and Chiu 2010; Richardson and Schnell 2020), nor of a mitochondrial presequence, the most cationic class compared (Wiedemann and Pfanner 2017). Proline-rich, poorly structured elements plausibly suit a translocon unrelated to TOC–TIC; the nitroplast envelope translocon remains unidentified. Multi-class analysis supports this: cTPs/presequences are confused far more often (19.5%, 22.8%) than either is with uTP windows (5.4%, 3.9%).

The comparison with *Paulinella chromatophora* is instructive: its chromatophore imports nucleus-encoded proteins via a short, bipartite, N-terminal element unlike a cTP (Nowack and Grossman 2012; Singer et al. 2017; Oberleitner et al. 2022). Each of two independent endosymbiont-to-organelle transitions uses a targeting solution that is neither a cTP nor a mitochondrial presequence, and in the nitroplast case not even N-terminal — the *Paulinella* half resting on cited work, not analysis here. With two cases, one unanalysed, this cannot be a general constraint; we advance it only as a hypothesis: an emerging import route requires *some* recognisable, low-complexity terminal extension rather than canonical plastid grammar. A synthetic organelle's targeting layer need not mimic cTP chemistry.

Consequences for nitrogen-fixation engineering in crops

The uTP matters because the nitroplast is the only known nitrogen-fixing organelle maintained by nucleus-encoded protein import. Prior nitrogenase-in-plants work delivered components to the mitochondrial matrix and plastids with retained activity/assembly competence (López-Torrejón et al. 2016; Allen et al. 2017; Eseverri et al. 2020; Ivleva et al. 2016); polyprotein strategies address stoichiometry (Yang et al. 2018; Yang et al. 2023), while structural work addresses oxygen protection (Franke et al. 2025). Our results bear on targeting in three specific ways.

First, the uTP is a template rather than a part. A C-terminal, proline-punctuated element read by an unidentified translocon cannot be transplanted into a crop plastid, because the receiving machinery is absent; what may transfer — as a hypothesis this work does not test — is the design principle that a low-complexity terminal extension with a small, high-specificity motif vocabulary suffices to route dozens of cargo proteins into an organelle, together with a concrete motif set to put into synthetic targeting sequences and assay.

Second, our species-blocked result is a caution specific to this application. The dicot → monocot transfer is the weakest block (AUC 0.8307 vs. 0.9814 under random folds), where a uTP-versus-cTP discriminator trained on Arabidopsis-dominated data will be least reliable for cereal crops. From this four-block experiment (point estimates, not resampled), we recommend monocot-held-out validation for cereal predictors, noting the penalty's magnitude needs confirmation under more balanced sampling.

Third, circularity caps expansion: a profile HMM trained on positives recovers 89% (659/740), but progress requires a better-grounded positive set—independent proteomic curation or direct import assays—not a larger model. The 76 curated proteins show signal exists but can't yield trustworthy proteome-scale false-positive rates.

Limitations

Beyond the three weaknesses reported in Results, eight limitations should be stated explicitly.

The window definition is a proxy. We analyse the last 120 residues, not an experimentally mapped uTP, since cleavage sites are undetermined; the true element is likely shorter, diluting feature estimates toward full-protein composition, so reported motif prevalences are lower bounds within the actual signal.

Tier assignment could not be re-derived. The primary curated import list on Dryad exists but could not be retrieved, and the proteomics deposition contains only raw spectra. The Tier 1 / Tier 2 partition thus relies on the compiled catalogue; an error there would propagate, making this the first task for anyone extending the work.

We could not read the primary source for the C-terminal placement. Only the Coale et al. (2024) abstract was accessible —no full text, figures, or supplementary material. All claims are secondary, including an unverified claim that C-terminal peptides are absent from proteomics, implying cleavage—not asserted here. We measured the extension ourselves (Control 8).

Our measured extension does not match the figure in circulation. Across four ortholog databases, median C-terminal overhang is 254–291 residues vs ~ 120 residues elsewhere—unresolved. Partly methodological: cargo hits are more diverged (median identity 0.30–0.39 vs 0.35–0.44), shortening alignable blocks, inflating overhang. A conserved ~ 120 aa motif-bearing block within a longer extension could reconcile figures, but is untestable without primary data.

The negative controls are surrogates, in two places. For the positional null, *E. huxleyi* is a haptophyte but different genus, with large genus-level composition offset (internal-versus-internal AUC 0.7834); NULL-3 stays valid, computed entirely within *E. huxleyi*. For Control 8, the non-targeted query set comes from *E. huxleyi*, since non-catalogue *B. bigelowii* proteins don't exist publicly: UniProt UP000013827 (35,617 proteins) for NULL-3/haptophyte contrasts, NCBI RefSeq GCF_000372725.1 (38,554) for the multi-organism panel/Control 8—reported separately, never pooled.

The *Paulinella* entry is a transcriptome proxy: no annotated *P. chromatophora* nuclear protein set is public; it derives from ORFs translated from TSA assembly GEZN01. ORF prediction fragments transcripts and yields spurious long ORFs, weakening terminal windows versus an annotated gene set; representation varies by growth condition (Zhang et al. 2017), leaving the ORF set incomplete, pending replacement by an annotated proteome.

The panel tests symbiotic status, not import competence. Of nine organisms, only *P. chromatophora* imports host-encoded proteins into a primary endosymbiont; *E. clementina*'s nitrogen-fixing spheroid body's import competence remains unestablished. More independent import systems would test the alternative explanation more stringently than one positive case.

The 120-aa window under-captures the motifs. The primary window captures only 59.0–72.0% of tetrapeptide occurrences, missing roughly one third (31.7% average across four motifs); every 120-aa motif prevalence is a lower bound, with 160-aa values motif-complete. The error is conservative — weakening rather than inflating enrichment — so no motif claim is at risk. We retain 120 aa because the internal-window control loses over two-thirds of its paired proteins by 250 aa, trading motif completeness for control integrity. Separately, the ortholog-based extension measurement (239–256 aa) and the 120-aa presequence figure measure different quantities — total length excess over the nearest non-cargo homologue versus a predicted cleavable targeting peptide — and no alignment-based redefinition reconciles them.

Structure was not modelled. We infer conformational constraint from composition and proline placement, not predicted structure. Since elements are short, low-complexity, and lack homologues of known fold, we judged structure prediction unlikely to be informative and skipped it—though that judgement remains testable.

Conclusion

Held to a within-species design, a positional compositional signal at the C-termini of import-catalogue proteins survives every control: these windows separate from internal windows of the same proteins at AUC 0.94 (0.9425 by the simplest contrast, 0.9445 with redundancy and positional background removed), and 0.8986 in the mass-spectrometry-curated subset, against a positional null spanning 0.5579–0.6274 across nine organisms/four clades. They also carry four proline-rich tetrapeptides — including RLLP and PRLL, absent from primary uTP descriptions — at odds ratios 34.9 (PDFP, 40% identity vs internal windows) to 580.9 (PDFP vs *E. huxleyi* C-terminal windows). This is genuine C-terminal addition, not mere compositional distinctness: catalogue proteins exceed haptophyte orthologues by a median 254 residues C-terminally, with no shift in length-matched non-targeted controls. Whether import machinery reads this element is untested — no import assay was run, and functional labels were inherited. The published 0.9814 cross-species figure should be retired as an effect-size claim, since 87% of the distance from chance to it (0.9186 background floor, 0.5 baseline) is reachable from species composition alone. What remains is a candidate targeting grammar for a nitrogen-fixing organelle, next tested by an import assay, not another classifier.

Declarations

Competing interests.

The author declares no competing interests.

Ethics approval.

Not applicable; this study is a computational reanalysis of public data and involved no human participants, animals or field sampling.

Declarations of generative AI use

Figures, tables and computational analysis came from an agentic scientific-computing environment (Claude Science, Anthropic) under author direction. The author designed the controls, reviewed all intermediate output, verified every reported value against its source table, and takes full responsibility for the manuscript's content and conclusions.

Funding.

No specific funding was received for this work.

Author Contribution

C.B. is the sole author. C.B. conceived the study, designed the control experiments, wrote all analysis code, performed all analyses, prepared all figures and tables, and wrote and revised the manuscript. C.B. read and approved the final manuscript.

Data Availability

All data analysed in this study are publicly available. The nitroplast import catalogue derives from the supplementary data of Coale et al. (2024, https://doi.org/10.1126/science.adk1075); the primary curated list is deposited at Dryad (https://doi.org/10.5061/dryad.2z34tmptf), which was verified to exist but could not be downloaded (web application firewall), so tier labels were inherited from the compiled catalogue rather than re-derived. The associated proteomics submission is PRIDE PXD047956 (raw spectra only). Reference proteomes and peptide annotations were retrieved from UniProtKB and NCBI: Emiliania huxleyi UP000013827 / GCF_000372725.1, Chrysochromulina tobinii GCA_001275005.1, Prymnesium parvum GCA_041296205.1, Paulinella chromatophora TSA GEZN01, Epithemia clementina GCA_051400955.1, Guillardia theta GCF_000315625.1, Phaeodactylum tricornutum GCF_000150955.2, Thalassiosira pseudonana GCF_000149405.2, Arabidopsis thaliana UP000006548, and reviewed transit- and signal-peptide entries for Arabidopsis thaliana, Oryza sativa, Zea mays and Triticum aestivum. All analysis code and every derived table underlying the reported numbers and figures are deposited at Zenodo: https://doi.org/10.5281/zenodo.21623357.

Code availability.

All analysis scripts, the complete set of derived tables, and the intermediate FASTA and clustering outputs are deposited at Zenodo under the reserved DOI 10.5281/zenodo.21623357. The deposit contains 44 analysis and figure modules with a manifest mapping each script to the tables it writes and the manuscript figure or table it feeds, pinned package versions, the exact mmseqs2 command lines, and a documented list of known reproducibility limits. The deposit is public and openly licensed (CC BY 4.0 for data and documentation, MIT for code).

References

1. Allen RS, Tilbrook K, Warden AC, Campbell PC, Rolland V, Singh SP, Wood CC (2017) Expression of 16 Nitrogenase Proteins within the Plant Mitochondrial Matrix. Front. https://doi.org/10.3389/fpls.2017.00287. Plant Sci
2. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H (2019) Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance. https://doi.org/10.26508/lsa.201900429

3. Archibald JM (2015) Endosymbiosis and Eukaryotic Cell Evolution. Curr Biol. https://doi.org/10.1016/j.cub.2015.07.055
4. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. https://doi.org/10.1093/nar/gkp335
5. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Stat Soc Ser B: Stat Methodol. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
6. Bruce BD (2000) Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol. https://doi.org/10.1016/s0962-8924(00)01833-x
7. Bruce BD (2001) The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. Biochimica et Biophysica Acta (BBA) -. https://doi.org/10.1016/s0167-4889(01)00149-5. Molecular Cell Research
8. Burén S, Rubio LM (2018) State of the art in eukaryotic nitrogenase engineering. FEMS Microbiol Lett. https://doi.org/10.1093/femsle/fnx274
9. Charpentier M, Oldroyd G (2010) How close are we to nitrogen-fixing cereals? Current Opinion in Plant Biology. https://doi.org/10.1016/j.pbi.2010.08.003
10. Coale TH, Loconte V, Turk-Kubo KA, Vanslembrouck B, Mak WKE, Cheung S, Ekman A, Chen JH et al (2024) Nitrogen-fixing organelle in a marine alga. Science. https://doi.org/10.1126/science.adk1075
11. Cornejo-Castillo FM, Cabello AM, Salazar G, Sánchez-Baracaldo P, Lima-Mendez G, Hingamp P, Alberti A, Sunagawa S et al (2016) Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. Nat Commun. https://doi.org/10.1038/ncomms11071
12. Cornejo-Castillo FM, Inomura K, Zehr JP, Follows MJ (2024) Metabolic trade-offs constrain the cell size ratio in a nitrogen-fixing symbiosis. Cell. https://doi.org/10.1016/j.cell.2024.02.016
13. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1002195
14. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. https://doi.org/10.1038/nprot.2007.131
15. Emanuelsson O, Nielsen H, Heijne GV (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. https://doi.org/10.1110/ps.8.5.978
16. Eseverri Á, López-Torrejón G, Jiang X, Burén S, Rubio LM, Caro E (2020) Use of synthetic biology tools to optimize the production of active nitrogenase Fe protein in chloroplasts of tobacco leaf cells. Plant Biotechnol J. https://doi.org/10.1111/pbi.13347
17. Franke P, Freiberger S, Zhang L, Einsle O (2025) Conformational protection of molybdenum nitrogenase by Shethna protein II. Nature. https://doi.org/10.1038/s41586-024-08355-3
18. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts565
19. Good A (2018) Toward nitrogen-fixing plants. Science. https://doi.org/10.1126/science.aas8737
20. Hagino K, Onuma R, Kawachi M, Horiguchi T (2013) Discovery of an Endosymbiotic Nitrogen-Fixing Cyanobacterium UCYN-A in Braarudosphaera bigelowii (Prymnesiophyceae). PLoS ONE. https://doi.org/10.1371/journal.pone.0081749
21. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J et al (2020) Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2
22. von STEPPUHN HEIJNEG, HERRMANN J RG (1989) Domain structure of mitochondrial and chloroplast targeting peptides. Eur J Biochem. https://doi.org/10.1111/j.1432-1033.1989.tb14679.x

23. Ivleva NB, Groat J, Staub JM, Stephens M (2016) Expression of Active Subunit of Nitrogenase via Integration into Plant Organelle Genome. PLoS ONE. https://doi.org/10.1371/journal.pone.0160951
24. Jarvis P, López-Juez E (2013) Biogenesis and homeostasis of chloroplasts and other plastids. Nat Rev Mol Cell Biol. https://doi.org/10.1038/nrm3702
25. Jeong J, Hwang I, Lee DW (2021) Functional Organization of Sequence Motifs in Diverse Transit Peptides of Chloroplast Proteins. Front Physiol. https://doi.org/10.3389/fphys.2021.795156
26. Jin Z, Wan L, Zhang Y, Li X, Cao Y, Liu H, Fan S, Cao D et al (2022) Structure of a TOC-TIC supercomplex spanning two chloroplast envelope membranes. Cell. https://doi.org/10.1016/j.cell.2022.10.030
27. Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. https://doi.org/10.1016/j.patter.2023.100804. Patterns
28. Kikuchi S, Asakura Y, Imai M, Nakahira Y, Kotani Y, Hashiguchi Y, Nakai Y, Takafuji K et al (2018) A Ycf2-FtsHi Heteromeric AAA-ATPase Complex Is Required for Chloroplast Protein Import. https://doi.org/10.1105/tpc.18.00357. Plant Cell
29. Kneip C, Voβ C, Lockhart PJ, Maier UG (2008) The cyanobacterial endosymbiont of the unicellular algae Rhopalodia gibba shows reductive genome evolution. BMC Evol Biol. https://doi.org/10.1186/1471-2148-8-30
30. Landa M, Turk-Kubo KA, Cornejo-Castillo FM, Henke BA, Zehr JP (2021) Critical Role of Light in the Growth and Activity of the Marine N2-Fixing UCYN-A Symbiosis. Front Microbiol. https://doi.org/10.3389/fmicb.2021.666739
31. Li Hm, Chiu CC (2010) Protein Transport into Chloroplasts. Annu Rev Plant Biol. https://doi.org/10.1146/annurev-arplant-042809-112222
32. López-Torrejón G, Jiménez-Vicente E, Buesa JM, Hernandez JA, Verma HK, Rubio LM (2016) Expression of a functional oxygen-labile nitrogenase component in the mitochondrial matrix of aerobically grown yeast. Nat Commun. https://doi.org/10.1038/ncomms11426
33. Massana R (2024) The nitroplast: A nitrogen-fixing organelle. Science. https://doi.org/10.1126/science.ado8571
34. Muñoz-Marín MC, Shilova IN, Shi T, Farnelid H, Cabello AM, Zehr JP (2019) The Transcriptional Cycle Is Suited to Daytime N 2 Fixation in the Unicellular Cyanobacterium Candidatus Atelocyanobacterium thalassa (UCYN-A). https://doi.org/10.1128/mbio.02495-18. mBio
35. Nakai M (2015) The TIC complex uncovered: The alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. Biochimica et Biophysica Acta (BBA) - Bioenergetics. https://doi.org/10.1016/j.bbabio.2015.02.011
36. Nakayama T, Kamikawa R, Tanifuji G, Kashiyama Y, Ohkouchi N, Archibald JM, Inagaki Y (2014) Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.1405222111
37. Nowack ECM, Grossman AR (2012) Trafficking of protein into the recently established photosynthetic organelles of Paulinella chromatophora. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.1118800109
38. Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.1608016113
39. Oberleitner L, Perrar A, Macorano L, Huesgen PF, Nowack ECM (2022) A bipartite chromatophore transit peptide and N-terminal protein processing in the Paulinella chromatophore. Plant Physiol. https://doi.org/10.1093/plphys/kiac012
40. Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG (2004) Intracellular Spheroid Bodies of Rhopalodia gibba Have Nitrogen-Fixing Apparatus of Cyanobacterial Origin. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msh086

41. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J et al (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature. https://doi.org/10.1038/nature12221
42. Richardson LGL, Schnell DJ (2020) Origins, function, and regulation of the TOC–TIC general protein import machinery of plastids. J Exp Bot. https://doi.org/10.1093/jxb/erz517
43. Rogers C, Oldroyd GED (2014) Synthetic biology approaches to engineering the nitrogen symbiosis in cereals. J Exp Bot. https://doi.org/10.1093/jxb/eru098
44. Sarkar D, Landa M, Bandyopadhyay A, Pakrasi HB, Zehr JP, Maranas CD (2021) Elucidation of trophic interactions in an unusual single-cell nitrogen-fixing symbiosis using metabolic modeling. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1008983
45. Singer A, Poschmann G, Mühlich C, Valadez-Cano C, Hänsch S, Hüren V, Rensing SA, Stühler K et al (2017) Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba Paulinella chromatophora. https://doi.org/10.1016/j.cub.2017.08.010. Current Biology
46. Sørensen MES, Stiller ML, Kröninger L, Nowack ECM (2025) Protein import into bacterial endosymbionts and evolving organelles. FEBS J. https://doi.org/10.1111/febs.17356
47. Sperschneider J, Catanzariti AM, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM (2017) LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci Rep. https://doi.org/10.1038/srep44598
48. Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. https://doi.org/10.1038/nbt.3988
49. Suzuki S, Kawachi M, Tsukakoshi C, Nakamura A, Hagino K, Inouye I, Ishida K (2021) Unstable Relationship Between Braarudosphaera bigelowii (= Chrysochromulina parkeae) and Its Nitrogen-Fixing Endosymbiont. Front. https://doi.org/10.3389/fpls.2021.749895. Plant Sci
50. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S et al (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. https://doi.org/10.1038/s41587-021-01156-3
51. Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, Kuypers MMM, Zehr JP (2012) Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. Science. https://doi.org/10.1126/science.1222700
52. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. https://doi.org/10.1038/nrg1271
53. Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, Affourtit JP, Zehr JP (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature. https://doi.org/10.1038/nature08786
54. Turk-Kubo KA, Farnelid HM, Shilova IN, Henke B, Zehr JP (2017) Distinct ecological niches of marine symbiotic N2-fixing cyanobacterium Candidatus Atelocyanobacterium thalassa sublineages. J Phycol. https://doi.org/10.1111/jpy.12505
55. The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. https://doi.org/10.1093/nar/gkac1052. Nucleic Acids Research
56. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. https://doi.org/10.1038/s41592-019-0686-2
57. Wernegreen JJ (2002) Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet. https://doi.org/10.1038/nrg931
58. Wiedemann N, Pfanner N (2017) Mitochondrial Machineries for Protein Import and Assembly. Annu Rev Biochem. https://doi.org/10.1146/annurev-biochem-060815-014352

59. Yang J, Xiang N, Liu Y, Guo C, Li C, Li H, Cai S, Dixon R et al (2023) Organelle-dependent polyprotein designs enable stoichiometric expression of nitrogen fixation components targeted to mitochondria. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.2305142120
60. Yang J, Xie X, Xiang N, Tian ZX, Dixon R, Wang YP (2018) Polyprotein strategy for stoichiometric assembly of nitrogen fixation components for synthetic biology. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.1804992115
61. Zehr JP, Shilova IN, Farnelid HM, Muñoz-Marín MdC, Turk-Kubo KA (2016) Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. Nat Microbiol. https://doi.org/10.1038/nmicrobiol.2016.214
62. Zhang R, Nowack ECM, Price DC, Bhattacharya D, Grossman AR (2017) Impact of light intensity and quality on chromatophore and nuclear gene expression in Paulinella chromatophora, an amoeba with nascent photosynthetic organelles. Plant J. https://doi.org/10.1111/tpj.13488
63. Zybailov B, Rutschow H, Friso G, Rudella A, Emanuelsson O, Sun Q, van Wijk KJ (2008) Sorting Signals, N-Terminal Modifications and Abundance of the Chloroplast Proteome. PLoS ONE. https://doi.org/10.1371/journal.pone.0001994

Tables

Table 1 Composition of the nitroplast protein-import catalogue and of every comparison set used in this study

(a) Import catalogue (Coale et al. 2024 supplementary data, tier labels inherited)

Quantity	Value
Total catalog entries	740
Tier1 MS-curated	81
Tier2 HMM-screen-only	659
Entries with sequence	735
Tier1 with sequence	76
Tier2 with sequence	659
Entries <120 aa	4
Median length Tier1 (aa)	665
Median length Tier2 (aa)	592

(b) Functional categories of catalogue entries

Functional category	n
Systematic HMM hit — eggNOG/KEGG annotated (uncurated function)	475
Systematic HMM hit — hypothetical / unannotated	176
Cofactor/vitamin biosynthesis (host-supplied)	42
Diazotrophy-related (O2 protection, cofactor delivery)	16
Regulatory / signaling	15
Ferredoxin / flavodoxin (redox electron carriers)	11
Central metabolism (glycolysis/PPP/isoprenoid/amino acid biosynthesis)	3
Diazotrophy / nitrogenase system	2

(c) Sequence windows analysed

Window set	Source sequences	Windows retained	Excluded	Construction rule
uTP C-terminal 120aa	735	735	0	4 proteins <120aa use full length
uTP internal 120aa	735	614	121	requires >=360aa (excl. first+last 120)
crop cTP (annotated)	1759	1759	0	all validated against UniProt full-length
plant internal 120aa	1759	1052	707	excludes cTP region + both 120aa flanks

(d) Additional targeting-peptide classes (UniProtKB, four plant species)

Peptide class	n	Median length (aa)	Species composition
Chloroplast transit peptide	1709	52	Ara 1231, Oryza 348, Zea 99, Triticum 31
Mitochondrial transit peptide	662	36	same four species
Secretory signal peptide	3159	24	same four species
Other transit peptide (excluded)	42	36	ambiguous organelle annotation

(e) Homology-filtered representative sets (mmseqs2 easy-cluster, -c 0.8, –cov-mode 1)

Window set	Identity threshold	n before	n representatives	% retained	% redundancy removed
uTP_cterm	40%	735	511	69.5	30.5
uTP_cterm	60%	735	644	87.6	12.4
cTP	40%	1759	1554	88.3	11.7
cTP	60%	1759	1634	92.9	7.1
uTP_internal	40%	614	493	80.3	19.7
uTP_internal	60%	614	574	93.5	6.5
plant_internal	40%	1052	586	55.7	44.3
plant_internal	60%	1052	652	62.0	38.0

Table 2 Master evidence table: every contrast computed in this study, with its interpretation

Control	Contrast	Metric	Value (mean ± SD over 25 folds where applicable)	Interpretation
Published headline	uTP C-term vs crop cTP (all 735)	AUC	0.9814 ± 0.0055	SPECIES-CONFOUNDED — do not report as the effect size
C1 within-species	uTP C-term vs own internal windows (all 735)	AUC	0.9425 ± 0.0167	within-species estimate — species composition cancels
C1 within-plant	cTP vs plant internal windows	AUC	0.9978 ± 0.0015	positive control: a real TP is near-perfectly separable
C1 background floor	uTP internal vs plant internal	AUC	0.9186 ± 0.0086	pure species effect, no targeting signal in either set
C1 background-corrected	paired within-protein deltas	AUC	0.9305 ± 0.0103	terminal-minus-internal, species cancels by construction
C1b NULL-1	two disjoint internal windows, same protein	AUC	0.6866 ± 0.0301	positional null
C1b NULL-2	plant C-term vs plant internal	AUC	0.5922 ± 0.0254	plant C-term is NOT a targeting signal
C6 NULL-3	E. huxleyi C-term vs internal	AUC	0.5878 ± 0.0065	haptophyte WITHOUT a nitroplast — decisive null
C2 filtered 60%	background-corrected delta, 60% id	AUC	0.9436	redundancy-filtered within-species estimate (60% identity)
C2 filtered 40%	background-corrected delta, 40% id	AUC	0.9445	redundancy-filtered within-species estimate (40% identity)
C3 random-fold	uTP C-term vs cTP, random CV	AUC	0.9814 ± 0.0055	optimistic
C3 species-blocked	uTP C-term vs cTP, blocked mean	AUC	0.9253	drop of 0.0561 on unseen plant species
C3 worst split	dicot-trained -> monocot-tested	AUC	0.8307	MATTERS: engineering target is a cereal
C4 Tier1 only	76 MS-curated vs cTP	AUC	0.9697 ± 0.0238	primary evidence, no pHMM circularity
C4 Tier1 within-species	76 MS-curated vs own internal	AUC	0.8986 ± 0.0553	most conservative estimate — pHMM-free subset, within-species
C4 Tier2	659 pHMM-screened vs cTP	AUC	0.9824 ± 0.0051	CIRCULAR — pHMM trained on the curated positives
C5 vs mito TP	uTP C-term vs plant mitochondrial TP	AUC	0.9872 ± 0.0066	distinct from mitochondrial targeting
C5 vs signal peptide	uTP C-term vs secretory signal peptide	AUC	0.9987 ± 0.0009	distinct from the secretory pathway
C6 haptophyte-matched	uTP C-term vs E. huxleyi C-term	AUC	0.8645 ± 0.0154	vs a between-haptophyte floor of 0.7834

C6 species floor	uTP internal vs E. huxleyi internal	AUC	0.7834 ± 0.0202	genus-level composition difference alone
P4 multi-organism null	Arabidopsis thaliana C-term vs internal (no symbiosis)	AUC	0.5579 ± 0.0079	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Epithemia clementina C-term vs internal (symbiosis, N2-fixing (import uncertain))	AUC	0.5689 ± 0.0081	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Phaeodactylum tricornutum C-term vs internal (no symbiosis)	AUC	0.5738 ± 0.0081	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Guillardia theta C-term vs internal (no symbiosis)	AUC	0.5756 ± 0.0094	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Chrysochromulina tobinii C-term vs internal (no symbiosis)	AUC	0.5776 ± 0.0100	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Emiliania huxleyi C-term vs internal (no symbiosis)	AUC	0.5795 ± 0.0072	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Paulinella chromatophora C-term vs internal (symbiosis with import (N-terminal))	AUC	0.5836 ± 0.0111	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Thalassiosira pseudonana C-term vs internal (no symbiosis)	AUC	0.5860 ± 0.0094	positional null; symbiotic status does not predict C-terminal asymmetry
P4 multi-organism null	Prymnesium parvum C-term vs internal (no symbiosis)	AUC	0.6274 ± 0.0084	positional null; symbiotic status does not predict C-terminal asymmetry
P4 Control 8	cargo C-terminal overhang vs E. huxleyi orthologs	median aa	254.0	cargo proteins carry a C-terminal extension absent from non-targeted controls
P4 Control 8	cargo N-terminal overhang vs E. huxleyi orthologs	median aa	21.0	no comparable N-terminal extension
P4 Control 8 control	non-targeted E. huxleyi C-terminal overhang vs C. tobinii	median aa	-4.0	negative control: no terminal extension

The two within-species estimates — 0.9425 from the simplest design and 0.9445 from the redundancy- and background-corrected design — differ by 0.0020, one-eighth of the standard deviation of the first, and are reported here as two routes to a single headline of 0.94 rather than as competing figures. Rows labelled P4 were added in the extended control round: the nine-organism positional null panel (see also Table 4) and the ortholog terminal length-shift test (Table 5). Source: synthesis_master_evidence.csv.

Table 3 Tetrapeptide statistics across four independent control layers

(a) Layer 1 — nitroplast C-terminal windows vs crop chloroplast transit peptides (6,222 tetrapeptides tested; 2,758 at FDR < 0.05, 2,117 at FDR < 0.01)

Tetrapeptide	uTP C-term windows	Crop cTP	Odds ratio	Fisher p	BH FDR
IPRL	162/735 (22.04%)	0/1759 (0.000%)	997.1	1.05e-92	1.63e-89
PDFP	167/735 (22.72%)	1/1759 (0.057%)	345.4	1.18e-93	2.45e-90
RLLP	253/735 (34.42%)	12/1759 (0.682%)	73.4	1.18e-131	7.31e-128
PRLL	235/735 (31.97%)	10/1759 (0.569%)	78.4	1.27e-122	3.94e-119

(b) Layer 2 — composition-controlled: C-terminal vs internal windows of the same nitroplast-targeted proteins (4,786 tested; 630 at FDR < 0.05)

Tetrapeptide	C-terminal windows	Internal windows, same proteins	Odds ratio	Fisher p	BH FDR
IPRL	162/735 (22.04%)	2/614 (0.326%)	69.4	1.50e-43	1.80e-40
PDFP	167/735 (22.72%)	3/614 (0.489%)	51.5	6.91e-44	1.10e-40
RLLP	253/735 (34.42%)	4/614 (0.651%)	71.3	1.13e-70	5.41e-67
PRLL	235/735 (31.97%)	2/614 (0.326%)	115.3	1.18e-67	2.81e-64

(c) Layer 3 — haptophyte control: vs *Emiliania huxleyi* CCMP1516 (31,626 C-terminal and 12,302 internal windows)

Tetrapeptide	uTP C-term (n = 735)	*E. huxleyi* C-term	*E. huxleyi* internal	OR vs *E. huxleyi* C-term	Fisher p
IPRL	162 (22.04%)	41 (0.130%)	18 (0.146%)	217.8	2.69e-232
PDFP	167 (22.72%)	16 (0.051%)	5 (0.041%)	580.9	1.80e-261
RLLP	253 (34.42%)	244 (0.772%)	117 (0.951%)	67.5	4.25e-291
PRLL	235 (31.97%)	205 (0.648%)	85 (0.691%)	72.0	9.33e-276

(d) Layer 4 — homology-filtered representatives (one sequence per mmseqs2 cluster)

Identity cut-off	Tetrapeptide	C-term reps	Internal reps	OR vs cTP reps	OR vs internal reps	BH FDR vs internal
60%	IPRL	143/644 (22.20%)	2/574 (0.348%)	inf	81.6	1.91e-36
60%	PDFP	145/644 (22.52%)	3/574 (0.523%)	474.5	55.3	1.20e-35
60%	RLLP	227/644 (35.25%)	4/574 (0.697%)	80.3	77.6	8.00e-62
60%	PRLL	211/644 (32.76%)	2/574 (0.348%)	88.0	139.4	2.34e-59
40%	IPRL	79/511 (15.46%)	2/493 (0.406%)	inf	44.9	1.20e-18
40%	PDFP	90/511 (17.61%)	3/493 (0.609%)	332.0	34.9	1.56e-20
40%	RLLP	149/511 (29.16%)	3/493 (0.609%)	63.6	67.2	2.52e-40
40%	PRLL	131/511 (25.64%)	2/493 (0.406%)	59.2	84.6	1.49e-35

All tests are two-sided Fisher exact tests on presence/absence per sequence (a tetrapeptide is counted once per sequence regardless of copy number), with Benjamini–Hochberg correction applied independently within each layer. Layer 4 tested 30,248 tetrapeptides at the 60% identity cut-off and 26,859 at 40%; Layer 3 reports the four focal tetrapeptides only, since the *E. huxleyi* comparison was run as a targeted test of the motifs established in Layers 1, 2 and 4 rather than as a fresh discovery scan. Odds ratios use a 0.5 pseudocount when a cell is zero; inf denotes a zero count in the comparison set. Sources: control7_fourmer_fisher_raw.csv, control7_fourmer_fisher_vs_internal.csv, control6_ehux_fourmer.csv, control7_fourmer_homology_filtered.csv.

Table 4 Multi-organism positional null panel

Organism	Clade	Symbiotic state	Protein source	n proteins	n paired windows	C-term vs internal AUC	N-term vs internal AUC
Arabidopsis thaliana	Viridiplantae	no symbiosis	UniProt proteome UP000006548	39,273	19,254	0.5579 ± 0.0079	0.6451 ± 0.0083
Epithemia clementina	Bacillariophyta	symbiosis, N2-fixing (import uncertain)	NCBI GenBank GCA_051400955.1	26,505	13,743	0.5689 ± 0.0081	0.6529 ± 0.0098
Phaeodactylum tricornutum	Bacillariophyta	no symbiosis	NCBI RefSeq GCF_000150955.2	10,408	5,609	0.5738 ± 0.0081	0.6797 ± 0.0120
Guillardia theta	Cryptophyta	no symbiosis	NCBI RefSeq GCF_000315625.1	24,822	9,568	0.5756 ± 0.0094	0.5877 ± 0.0081
Chrysochromulina tobinii	Haptophyta	no symbiosis	NCBI GenBank GCA_001275005.1	16,765	8,256	0.5776 ± 0.0100	0.5893 ± 0.0079
Emiliania huxleyi	Haptophyta	no symbiosis	NCBI RefSeq GCF_000372725.1	38,554	12,696	0.5795 ± 0.0072	0.6296 ± 0.0089
Paulinella chromatophora	Rhizaria (Cercozoa)	symbiosis with import (N-terminal)	NCBI TSA GEZN01 (ORF-translated, >=100 aa)	16,423	5,822	0.5836 ± 0.0111	0.5729 ± 0.0106
Thalassiosira pseudonana	Bacillariophyta	no symbiosis	NCBI RefSeq GCF_000149405.2	11,673	6,392	0.5860 ± 0.0094	0.7120 ± 0.0097
Prymnesium parvum	Haptophyta	no symbiosis	NCBI GenBank GCA_041296205.1	24,907	13,368	0.6274 ± 0.0084	0.6459 ± 0.0078

All nine organisms use identical window geometry (proteins ≥ 360 aa; C-terminal = last 120 aa, N-terminal = first 120 aa, internal = central 120 aa) and the identical classifier and cross-validation protocol used for the catalogue. Proteomes were subsampled to at most 8,000 paired windows. The *Paulinella chromatophora* entry is translated from TSA assembly GEZN01 (complete ORFs ≥ 100 aa), not from an annotated gene set. The *E. huxleyi* entry here uses NCBI RefSeq

GCF_000372725.1 and is a different source set from NULL-3 (UniProt UP000013827, AUC 0.5878); the two are reported separately and never pooled. For scale, the uTP C-terminal-versus-internal contrast in *B. bigelowii* is 0.9425 ± 0.0167. Source: phase4_multiorg_null_panel.csv.

Table 5 Control 8 — ortholog terminal length-shift test

Query set	Ortholog database	n with hit	Median query length (aa)	Median C-terminal overhang (aa)	Median N-terminal overhang (aa)	Fraction C-overhang > 100 aa	Median hit identity	Wilcoxon C vs N (p)
catalogue cargo	*C. tobinii*	394	686.5	255.5	32.0	0.850	0.370	5.98e-37
non-targeted, length-matched	*C. tobinii*	474	613.0	1.0	6.0	0.247	0.405	7.41e-01
non-targeted *E. huxleyi*	*C. tobinii*	511	331.0	-4.0	-9.0	0.094	0.438	1.71e-01
catalogue cargo	*E. huxleyi*	518	651.0	254.0	21.0	0.911	0.390	1.15e-66
catalogue cargo	*P. parvum*	438	652.5	257.5	17.0	0.943	0.390	2.89e-55
non-targeted, length-matched	*P. parvum*	552	602.5	0.0	-3.0	0.134	0.410	1.14e-01
non-targeted *E. huxleyi*	*P. parvum*	600	328.5	-7.0	-11.5	0.037	0.445	6.95e-01
catalogue cargo	UCYN-A	27	614.0	291.0	20.0	0.926	0.299	1.45e-05
non-targeted, length-matched	UCYN-A	62	652.0	16.0	53.0	0.306	0.309	6.97e-01
non-targeted *E. huxleyi*	UCYN-A	53	428.0	-2.0	11.0	0.151	0.347	1.50e-01

Overhangs are computed from the single lowest-e-value mmseqs2 easy-search hit per query (-e 1e-5 --max-seqs 20) as (qlen − qend) − (tlen − tend) at the C-terminus and (qstart − 1) − (tstart − 1) at the N-terminus; positive values mean the query extends further beyond the aligned block than the target does. The non-targeted control is a random sample of 900 *E. huxleyi* proteins ≥ 100 aa (random.seed(7)); the length-matched control is the subsample of it matched to the cargo query-length distribution. Cargo-versus-length-matched-control Mann–Whitney tests give, at the C-terminus, $p = 3.5 \times 10^{-61}$ (CLES 0.824) against *C. tobinii*, $p = 1.2 \times 10^{-109}$ (0.911) against *P. parvum* and $p = 1.9 \times 10^{-9}$ (0.895) against UCYN-A. Cargo hits are more diverged than control hits, which inflates apparent overhang; see Limitations. Sources:

phase4_control8_overhang_summary.csv, phase4_control8_cargo_vs_lenmatched_tests.csv, phase4_control8_per_protein_overhangs.csv.

Figures

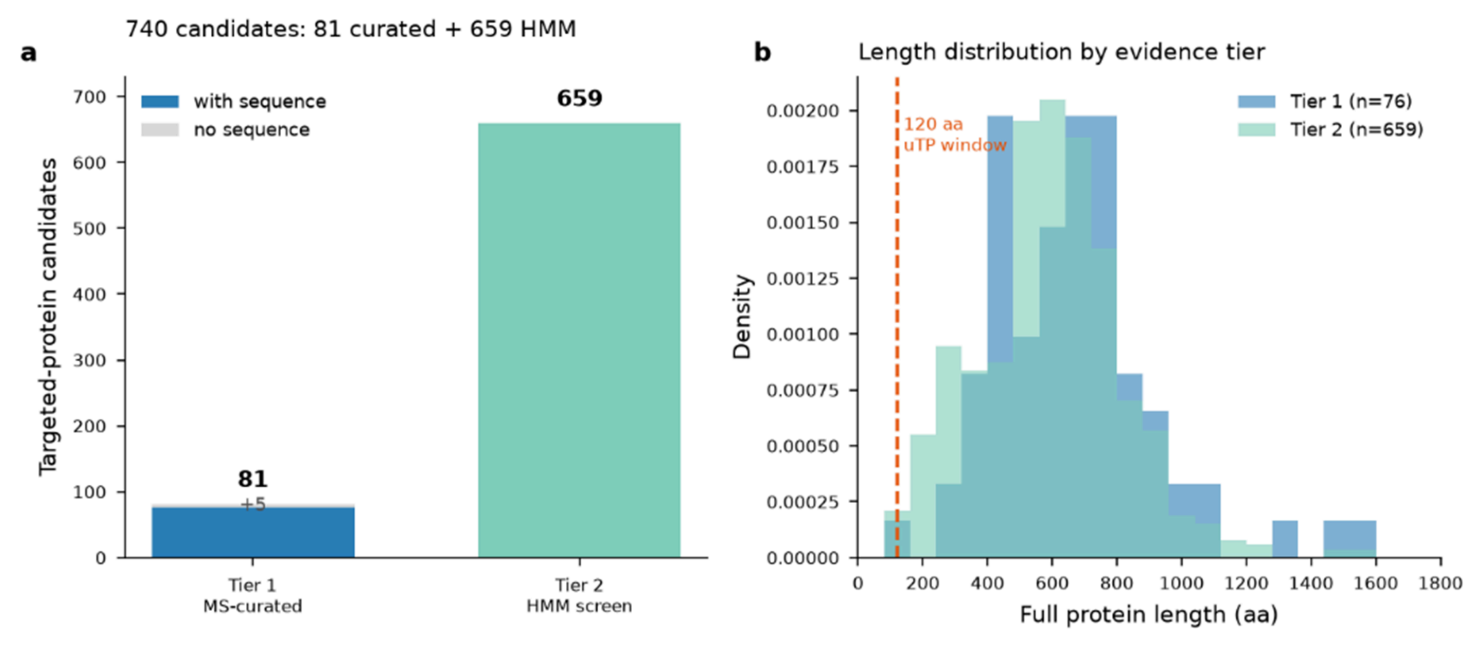


Figure 1

Composition of the nitroplast import catalogue. (a) Tier structure of the 740 catalogue entries: 81 proteins with mass-spectrometry curation (Tier 1) and 659 recovered only as profile-HMM hits (Tier 2); sequences are available for 735 entries (76 Tier 1, 659 Tier 2). (b) Distribution of full-length protein length by tier (median 665 aa Tier 1, 592 aa Tier 2); the four entries shorter than the 120-residue analysis window are indicated. (c) Functional category assignment across the catalogue, showing that 651 of 740 entries (88%) carry only systematic profile-HMM annotation, of which 176 are hypothetical or unannotated. Source: table1_catalog_summary.csv, functional_category_summary.csv.

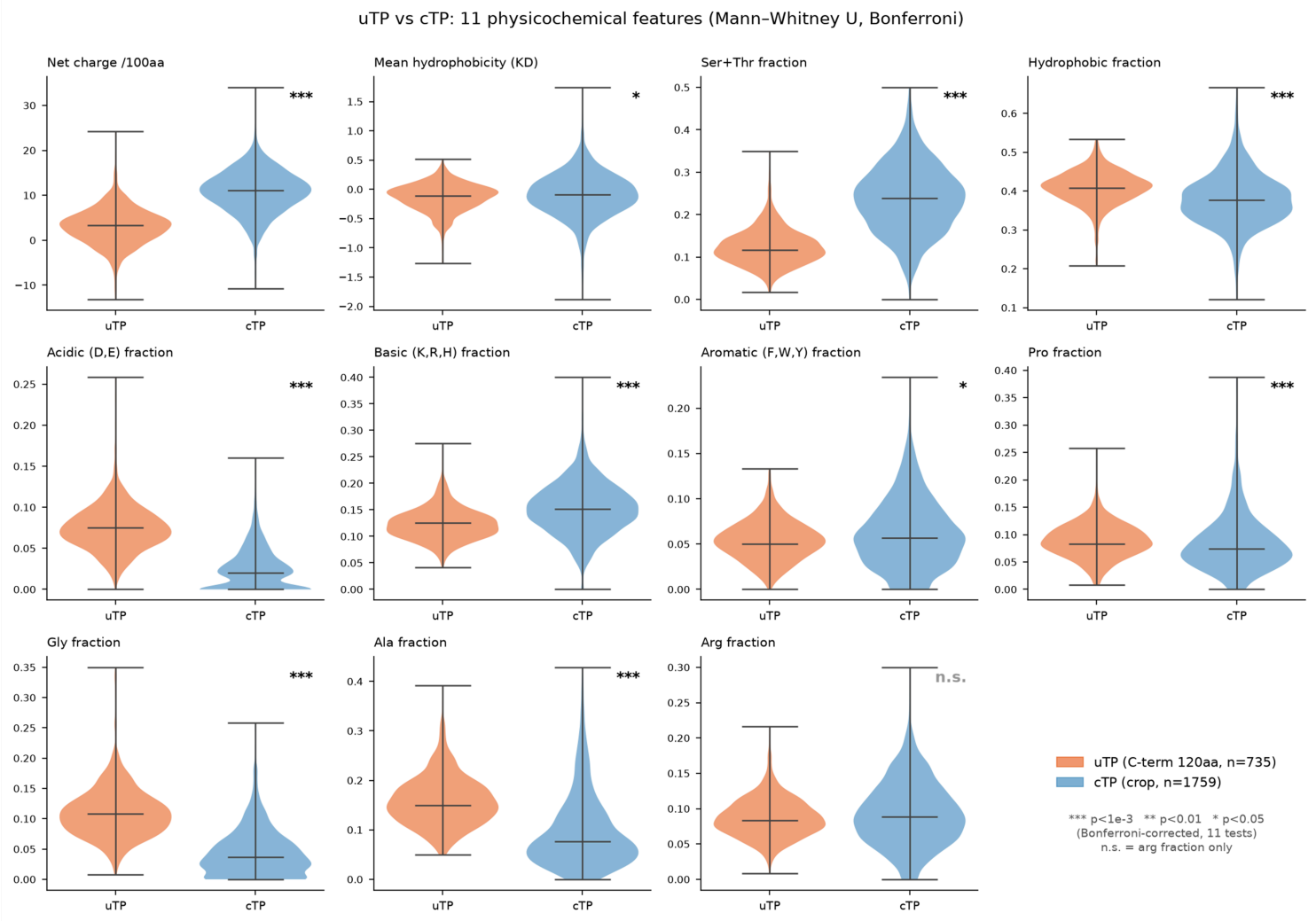


Figure 2

uTP C-terminal windows differ from crop chloroplast transit peptides on ten of eleven sequence features. Distributions of the eleven features for 735 uTP C-terminal 120-residue windows and 1,759 curated crop cTPs. Boxes show median and interquartile range; annotations give the Bonferroni-corrected two-sided Mann–Whitney p-value and the common-language effect size. uTP windows are richer in Gly (median 0.108 vs 0.036), Ala (0.150 vs 0.077) and acidic residues (0.075 vs 0.020), and much lower in net charge (3.33 vs 11.03 per 100 aa) and Ser+Thr (0.117 vs 0.238). Arg fraction is the only feature not significant after correction (z = 2.04, p_Bonferroni = 0.459). This comparison is cross-species and is presented for continuity with the published analysis, not as an effect size — see Fig. 3. Source: table2_statistical_comparison.csv, reanalysis_feature_matrix.csv.

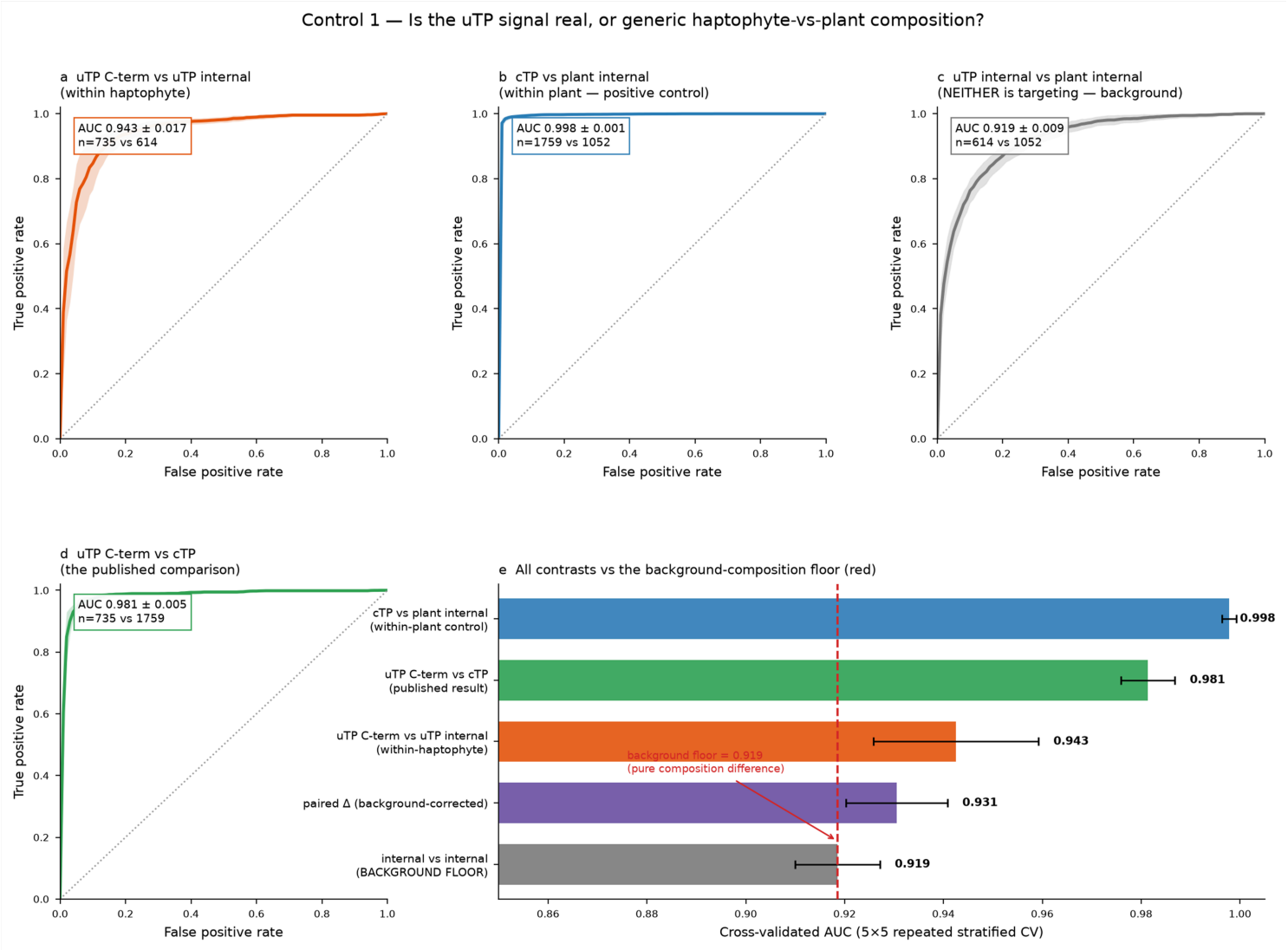


Figure 3

The decisive control: a 2×2 within- versus cross-species design shows that most of the published separation is proteome composition. Cross-validated ROC AUC (mean ± SD over 25 folds of repeated stratified 5-fold cross-validation) for five contrasts using the same eleven features and the same L2-penalised logistic pipeline. (a) Within-haptophyte — uTP C-terminal windows vs internal windows of the same 735 proteins (n = 614 internal), AUC 0.9425 ± 0.0167: the honest headline. (b) Within-plant positive control — annotated cTPs vs internal windows of the same plant proteins (n = 1,052), 0.9978 ± 0.0015, validating the window design. (c) Background floor — haptophyte internal vs plant internal windows, containing no targeting signal in either set, 0.9186 ± 0.0086. (d) The original published cross-species contrast, 0.9814 ± 0.0055, which exceeds the floor in (c) by only 0.063. (e) Background-corrected paired within-protein deltas (terminal minus internal, computed separately per organism), 0.9305 ± 0.0103. Panels (b)–(e) establish that the published figure must not be reported as the effect size of the uTP signal. Source: control1_2x2_auc_matrix.csv, control1_delta_medians.csv.

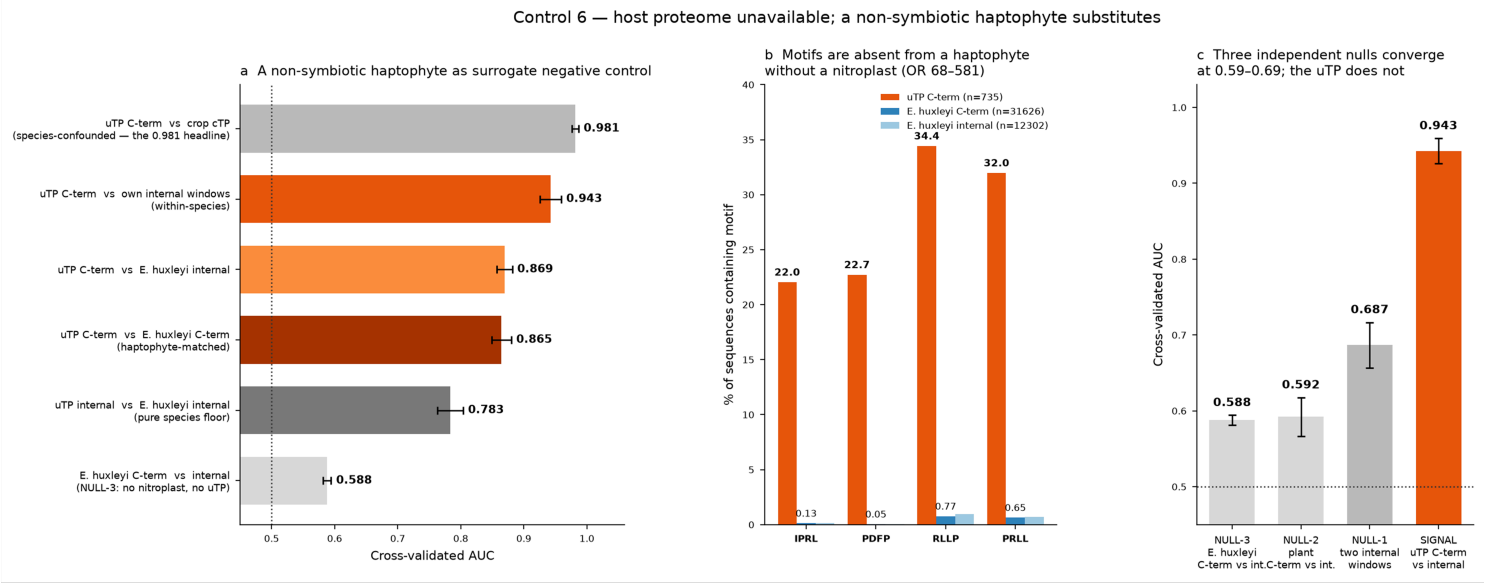


Figure 4

A haptophyte without a nitroplast supplies the strongest negative control. (a) Cross-validated AUC for contrasts involving *Emiliania huxleyi* CCMP1516 (UniProt UP000013827; 35,617 proteins; 31,626 C-terminal and 12,302 internal 120-residue windows, negatives subsampled to 5,000 per fold): uTP vs *E. huxleyi* C-terminal windows 0.8645 ± 0.0154; uTP vs *E. huxleyi* internal 0.8693 ± 0.0123; the genus composition floor (haptophyte internal vs *E. huxleyi* internal, no targeting signal on either side) 0.7834 ± 0.0202; and **NULL-3**, the identical C-terminal-versus-internal contrast computed entirely within *E. huxleyi*, 0.5878 ± 0.0065. Position alone carries almost no information in a haptophyte proteome that hosts no nitroplast. (b) Prevalence of the four enriched tetrapeptides in uTP C-terminal windows versus *E. huxleyi* C-terminal and internal windows (IPRL 22.04% vs 0.130%/0.146%; PDFP 22.72% vs 0.051%/0.041%; RLLP 34.42% vs 0.772%/0.951%; PRLL 31.97% vs 0.648%/0.691%), with Fisher odds ratios of 217.8, 580.9, 67.5 and 72.0 (p = 2.7e-232 to 4.3e-291). Source: control6_ehux_auc.csv, control6_ehux_fourmer.csv, control6_host_proteome_search.md.

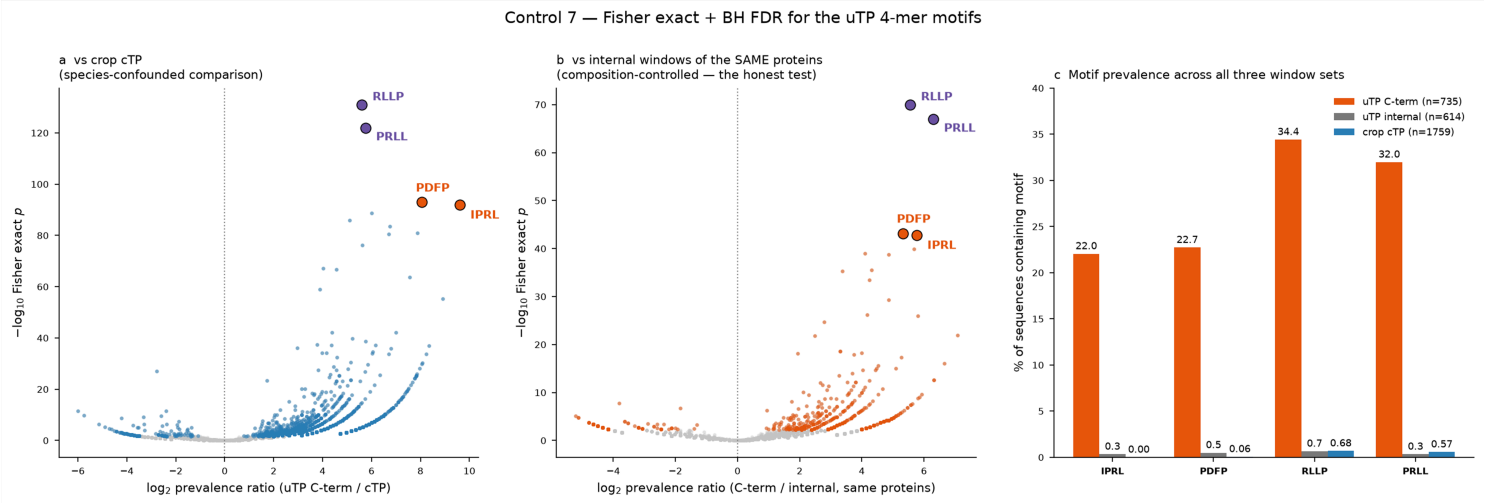


Figure 5

Tetrapeptide enrichment survives Benjamini–Hochberg correction, composition control and homology filtering. (a) Volcano representation of all 6,222 tetrapeptides tested by two-sided Fisher's exact test in uTP C-terminal windows versus crop cTPs, log2 enrichment against −log10 BH-adjusted p; 2,758 pass FDR < 0.05 and 2,117 FDR < 0.01. The two previously reported motifs (IPRL, PDFP) and the two identified here (RLLP, PRLL) are labelled. (b) The composition-

controlled layer — the same test against internal windows of the *same* proteins (4,786 tetrapeptides tested, 630 at FDR < 0.05) — where all four motifs remain significant (IPRL FDR 1.80e-40; PDFP 1.10e-40; RLLP 5.41e-67; PRLL 2.81e-64). (c) Prevalence of the four motifs in uTP windows, internal windows and cTPs. RLLP and PRLL are stronger than IPRL and PDFP in every layer. Source: control7_fourmer_fisher_raw.csv, control7_fourmer_fisher_vs_internal.csv, control7_fourmer_homology_filtered.csv.

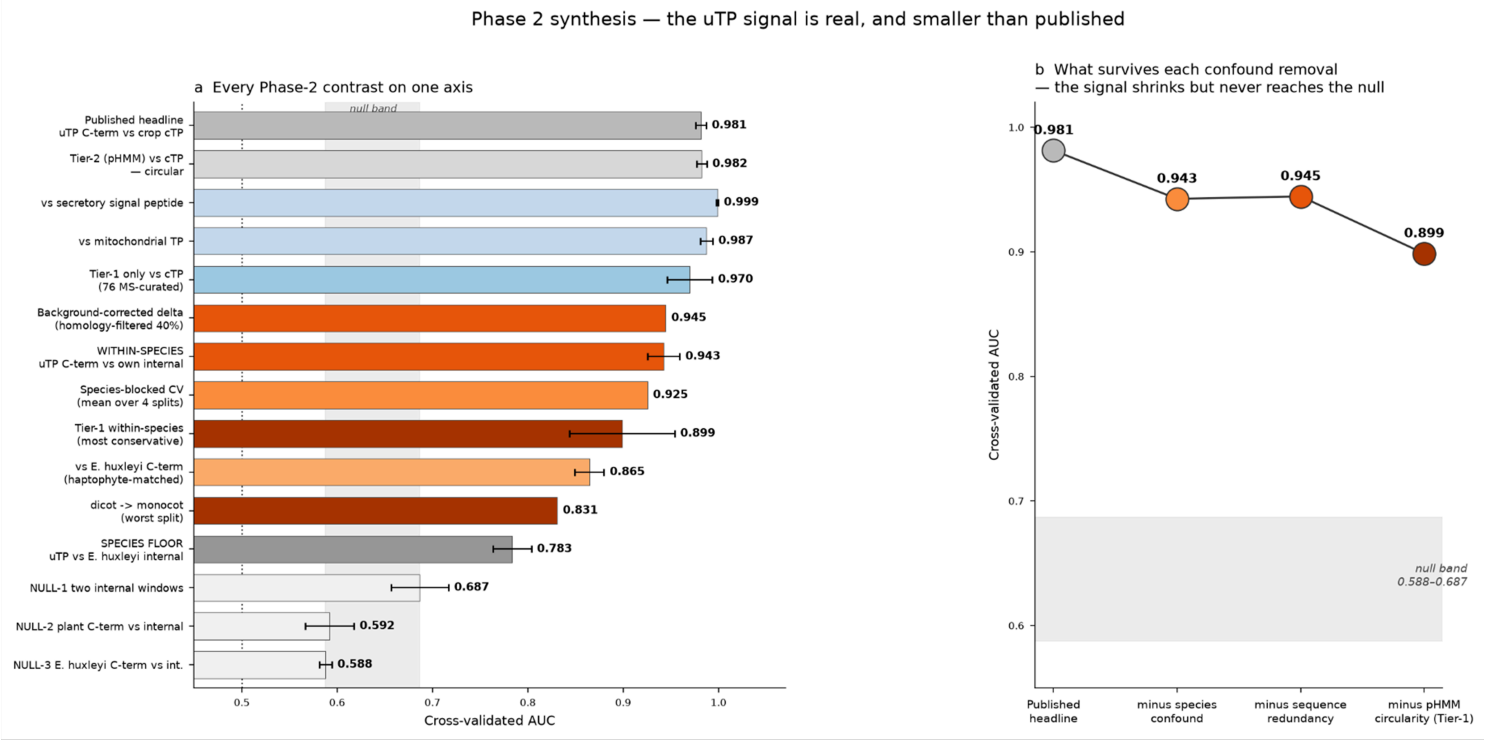


Figure 6

Synthesis: the confound-removal cascade and the null band. (a) All fifteen contrasts computed in this work, ordered by AUC, with the null band (0.5878–0.6866, established independently in *E. huxleyi*, in plant termini and in internal-versus-internal haptophyte pairs) shaded. (b) The cascade of confound removal: 0.9814 published cross-species → 0.9425 after moving to a within-protein design (species confound removed) → 0.9445 after additionally removing sequence redundancy at 40% identity (best estimate of the true effect) → 0.8986 after restricting to the 76 mass-spectrometry-curated proteins (profile-HMM circularity removed; most conservative honest number). Every stage remains far above the null band. Source: synthesis_master_evidence.csv.

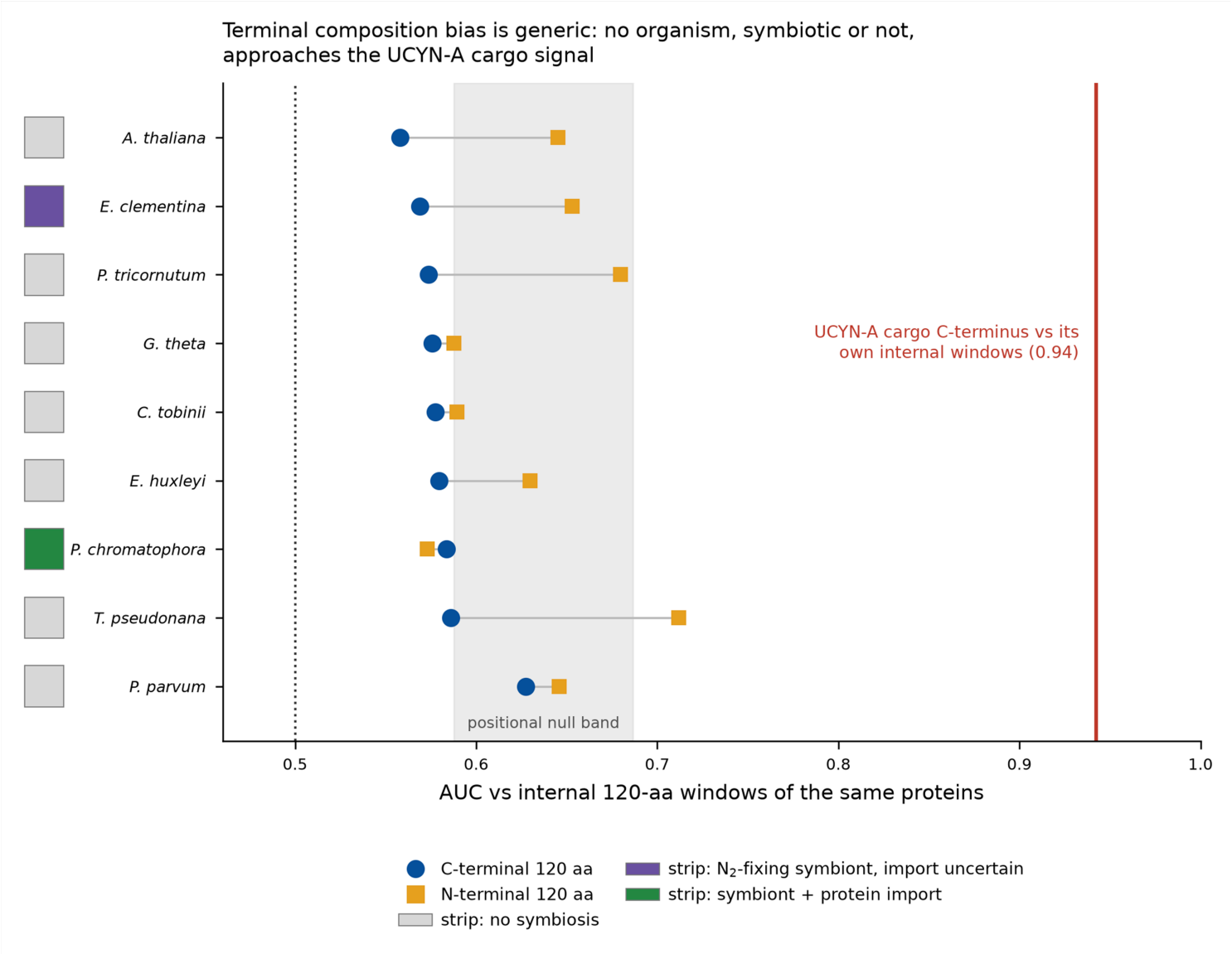


Figure 7

The positional null holds across nine organisms, four clades and three symbiotic states. (a) C-terminal-versus-internal window AUC for each organism, ordered by value: *A. thaliana* 0.5579 ± 0.0079, *E. clementina* 0.5689 ± 0.0081, *P. tricornutum* 0.5738 ± 0.0081, *G. theta*0.5756 ± 0.0094, *C. tobinii* 0.5776 ± 0.0100, *E. huxleyi* 0.5795 ± 0.0072, *P. chromatophora* 0.5836 ± 0.0111, *T. pseudonana* 0.5860 ± 0.0094, *P. parvum* 0.6274 ± 0.0084; bars coloured by symbiotic state, with the uTP within-species contrast (0.9425) shown for scale. (b) The same organisms scored at the N-terminus, which is systematically more asymmetric (0.5729–0.7120). (c) Paired C- and N-terminal values per organism, showing that the *Paulinella* entry — the only panel member known to import host-encoded proteins into a primary endosymbiont — is indistinguishable from the non-symbiotic organisms at the C-terminus. Error bars are standard deviations over 25 cross-validation folds. Window geometry, feature set and classifier are identical to those used for the catalogue throughout. Source: phase4_multiorg_null_panel.csv.

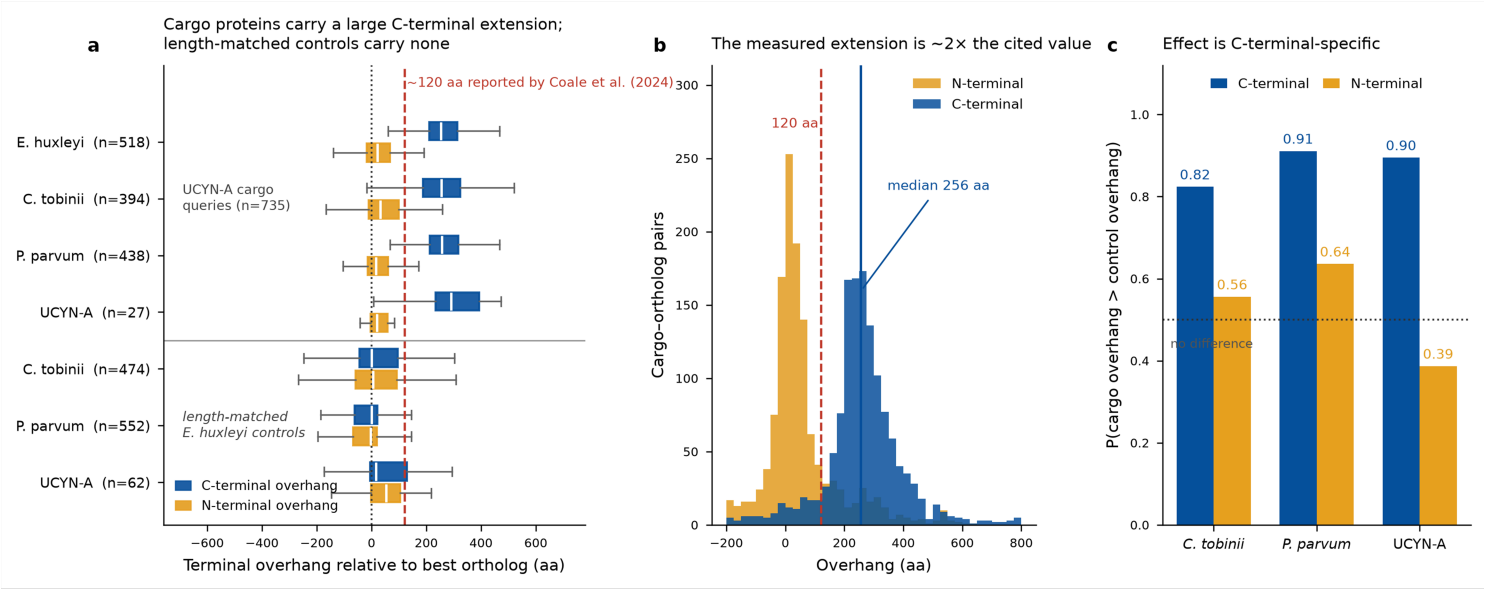


Figure 8

Catalogue proteins carry a large C-terminal extension that non-targeted controls lack. (a) Distribution of C-terminal overhang, computed as (qlen − qend) − (tlen − tend) from the best mmseqs2 hit, for the 735 catalogue proteins against four ortholog databases (medians: *E. huxleyi* 254.0 aa, n = 518; *P. parvum* 257.5, n = 438; *C. tobinii* 255.5, n = 394; UCYN-A 291.0, n = 27) and for the 900-protein non-targeted *E. huxleyi* control (medians −4.0 to −2.0 aa). (b) The same comparison at the N-terminus, where cargo overhangs are an order of magnitude smaller (17–32 aa). (c) Length-matched re-analysis: cargo versus a control subsample matched on query length, with Mann–Whitney p and common-language effect size annotated (C-terminus: CLES 0.824, 0.911 and 0.895 against *C. tobinii*, *P. parvum* and UCYN-A; N-terminus: 0.556, 0.636 and 0.387). (d) Fraction of queries whose overhang exceeds 100 aa, by terminus and query set. Sources: phase4_control8_overhang_summary.csv, phase4_control8_per_protein_overhangs.csv, phase4_control8_cargo_vs_lenmatched_tests.csv.

Supplementary Files

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

- FigS1.tif
- FigS2.tif
- FigS3.tif
- FigS4.tif
- FigS5.tif
- FigS6.tif
- FigS7.tif