The UCYN-A transit peptide: a novel C-terminal targeting system with distinctive biophysical substrate signatures

Author One[1,*], Author Two[2]

[1]Institution One, City, Country
[2]Institution Two, City, Country

[*]Corresponding author: email@example.com

Abstract

1 Introduction

Nitrogen limits the growth of most terrestrial and aquatic ecosystems. Although molecular nitrogen constitutes 78% of the atmosphere, only prokaryotes possess the enzymatic machinery to reduce it to biologically available ammonia. This metabolic capability, nitrogen fixation, has never evolved in any eukaryote Pi. Engineering nitrogen-fixing crops remains a long-standing goal that would reduce dependence on industrial fertilizer production and its associated environmental costs. Yet despite decades of effort, the barriers to transferring nitrogen fixation into eukaryotic cells remain formidable.

Nature has solved this problem exactly once. In 2024, Coale and colleagues demonstrated that the cyanobacterial endosymbiont UCYN-A (*Candidatus* Atelocyanobacterium thalassa) has crossed the threshold from endosymbiont to organelle in its marine haptophyte host *Braarudosphaera bigelowii* Coale et al. (2024). They named this nitrogen-fixing organelle the nitroplast. Three lines of evidence support this classification: UCYN-A divides in synchrony with its host cell Turk-Kubo et al. (2023), UCYN-A has undergone extreme genome reduction that renders it metabolically dependent on the host Masuda et al. (2024), and critically, the host imports hundreds of nuclear-encoded proteins into UCYN-A. The nitroplast represents the first opportunity to study organellogenesis as it unfolds, rather than inferring the process from the highly derived mitochondria and chloroplasts that emerged billions of years ago.

The protein import system is central to both understanding organellogenesis and any future engineering efforts. When an endosymbiont becomes an organelle, genes transfer from endosymbiont to host nucleus while their protein products must still reach the organelle to function Frail et al. (2025). This requirement creates intense selective pressure for targeting mechanisms. Mitochondria and chloroplasts solved this problem through N-terminal transit peptides recognized by elaborate translocon complexes. The nitroplast evolved an independent solution: a C-terminal extension of approximately 120 amino acids that Coale et al. termed the UCYN-A transit peptide (uTP) Coale et al. (2024). This targeting sequence has no detectable sequence or structural homologs in any public database, suggesting it arose de novo during the nitroplast endosymbiosis.

The discovery of the uTP system raises fundamental questions about how novel protein import mechanisms originate. What structural features does the import machinery recognize? Did uTP spread through the host proteome in discrete acquisition events, or did it expand gradually? And crucially for any engineering application: can any protein be targeted for import by adding uTP, or do cargo proteins themselves require specific properties? Comparative analysis with the related diazoplast system in *Epithemia* diatoms, where minimal protein import has evolved despite millions of years of endosymbiosis Frail et al. (2025), suggests that successful import systems require more than simply acquiring a targeting signal.

Here we characterize the architecture and evolutionary dynamics of the uTP system. We find that uTP sequences combine conserved structural elements with continuous sequence variation, arguing against discrete acquisition events. Unexpectedly, we discover that biophysical properties of the mature protein domain predict uTP presence with high accuracy. This finding suggests dual constraints on uTP-mediated import: the transit peptide must present conserved structural features for recognition, while the cargo protein must possess compatible biophysical properties. These constraints illuminate both the evolutionary trajectory of the nitroplast and the challenges facing efforts to engineer similar systems.

2 Results

2.1 The uTP region combines conserved sequence anchors with a structurally convergent variable domain

Coale et al. identified a C-terminal extension of approximately 120 amino acids on proteins imported into the nitroplast. We characterized the sequence organization of this extension, hereafter termed the UCYN-A transit peptide (uTP). Two short conserved sequence elements appear at the start of the uTP region in over 90% of sequences (Figure 1A). We term these elements anchor 1 and anchor 2 because they mark the boundary between mature domain and transit peptide. Among 745 proteins with detectable sequence elements, 60% display the canonical order with anchor 2 preceding anchor 1. These anchors are detectable in 80% of proteins predicted to contain uTP by hidden Markov model search, indicating broad conservation across the import candidate set.

After the anchor elements, sequences diverge into a variable linker region. This region contains additional sequence elements in varying combinations, but shows substantially higher diversity than the anchor region. Pairwise sequence similarity in the linker averages only 7%, indicating that most uTP sequences share little primary sequence identity beyond the conserved anchors. The linker region varies in length from approximately 50 to over 200 amino acids across different proteins.

Despite this sequence diversity, structure predictions reveal that the anchor motifs encode a conserved three-dimensional fold. We predicted structures for 47 uTP-containing proteins using AlphaFold3. The anchor region adopts a three-helix bundle architecture in 98% of structures (46/47; Figure 1B). Anchor 2 forms the first alpha-helix, while anchor 1 folds into a helix-turn-helix motif comprising two additional helices. Together these elements create a characteristic U-bend configuration, with the anchor 2 helix forming one arm and the anchor 1 helix-turn-helix forming the other. Positional variance in this structural core averages less than 1 Å (mean 0.90 Å, range 0.60–1.38 Å), demonstrating strong structural conservation. The high pairwise root mean square deviation across full structures (19.3 Å) reflects variation in linker length rather than divergence of the conserved core.

We next asked whether uTP sequences form discrete subtypes or vary continuously. We applied four clustering methods to uTP sequences: hierarchical clustering, spectral clustering, k-means, and DBSCAN. All methods produced low silhouette scores (0.01–0.08), indicating weak cluster structure. Different methods assigned sequences to completely different groupings, with adjusted Rand indices near zero between methods. Visualization by UMAP shows a continuous distribution with no clear gaps or boundaries (Figure 1C). Comparison to shuffled controls that preserve amino acid composition confirmed that real sequences show lower silhouette scores (0.096 versus 0.135, permutation test $p = 0.01$) and lower distance variance (84.6 versus 106.7). This pattern indicates that uTP sequences spread uniformly across a constrained region of sequence space rather than forming discrete subtypes.

The combination of conserved anchors, structural convergence, and continuous sequence variation suggests a model for uTP architecture. The anchor motifs are not merely sequence signatures but structural determinants that encode the conserved three-helix bundle required for recognition by import machinery, explaining their near-universal conservation. The variable linker accumulates mutations freely because it does not participate in recognition. The uniform distribution of sequences in the variable region, with no discrete subtypes, indicates that primary sequence identity in the linker is not under strong selective constraint.

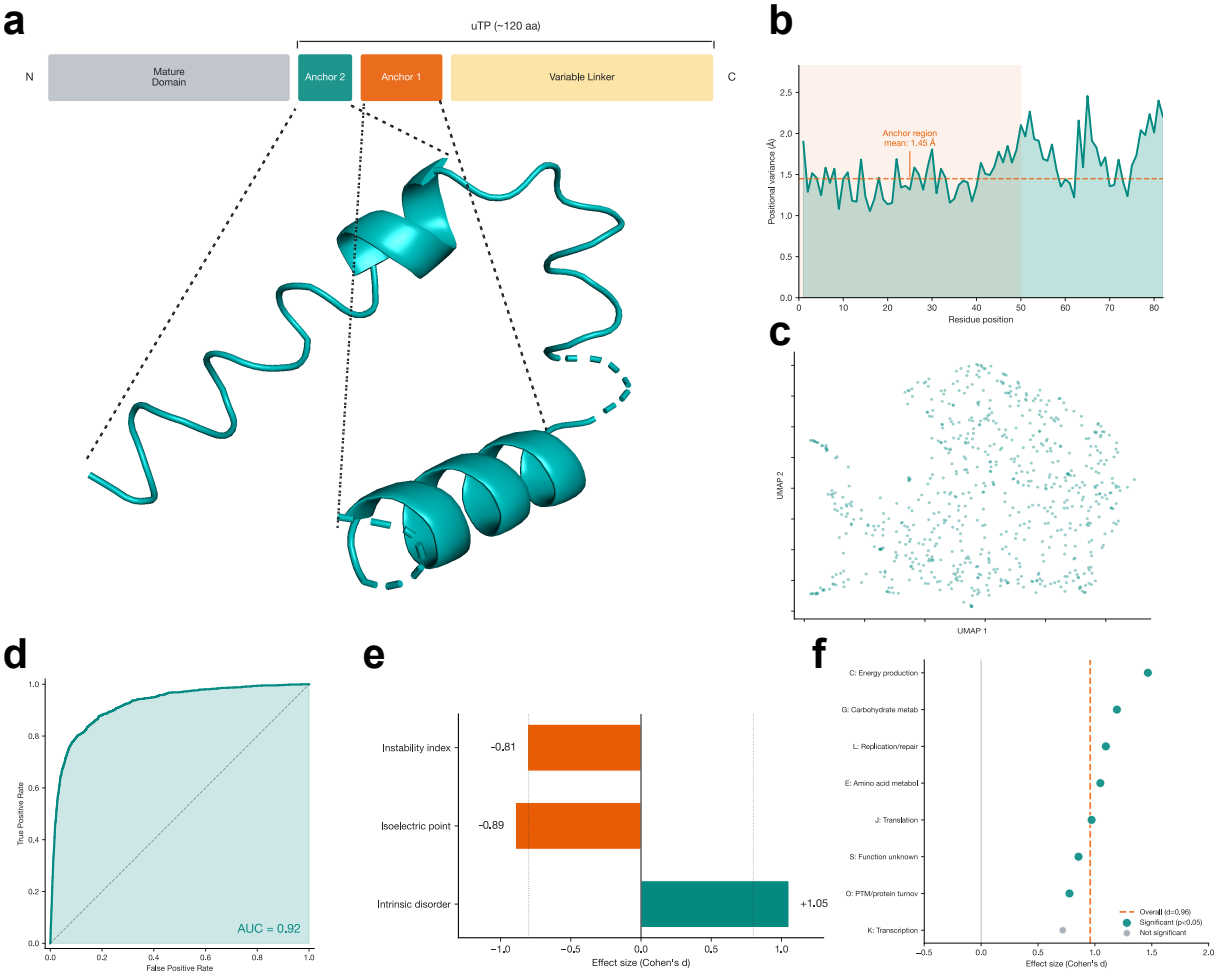

Figure 1: The uTP region combines conserved structural elements with a distinctive biophysical signature in mature domains. (A) Schematic of uTP organization showing the conserved anchor motifs that form a three-helix bundle, followed by a variable linker region. **(B)** Positional variance along the structural core. The anchor region shows low variance, demonstrating strong structural conservation. **(C)** UMAP visualization shows a continuous distribution without discrete clusters. **(D)** ROC curve for predicting uTP presence from mature domain sequences (AUC = 0.92). **(E)** Effect sizes for biophysical properties: elevated disorder (+1.05), lower pI (-0.89), and lower instability (-0.81). **(F)** Within-category effect sizes for intrinsic disorder. All eight functional categories show effects in the same direction. **(G)** Variance partitioning: uTP status explains more unique variance (7.3%) than is shared with function (3.4%). **(H)** Gene family analysis: 25% of uTP proteins share families versus 2% expected, but 624 distinct families with 75% singletons.

2.2 Mature domain biophysical properties predict uTP presence independently of ancestry and function

The structural conservation of uTP implies recognition by specific import machinery. We asked whether successful import depends only on the transit peptide, or whether the cargo protein itself must have compatible properties. To test this, we trained classifiers to predict uTP presence from mature domain sequences alone, excluding the uTP region. A logistic regression classifier achieved 92.8% accuracy in distinguishing uTP-containing proteins from cytoplasmic and nuclear controls (605 versus 773 proteins, permutation test $p = 0.002$; Figure 1D). Validation on the full *Braarudosphaera bigelowii* proteome (933 uTP versus 43,430 non-uTP proteins) yielded a receiver operating characteristic area under curve of 0.92, confirming strong discriminative power despite extreme class imbalance.

Three biophysical properties of the mature domain drive this classification (Figure 1E). First, uTP proteins contain more intrinsically disordered regions than controls (Cohen's $d = +1.05$, large effect). Second, uTP proteins have lower isoelectric points, indicating greater acidity ($d = -0.89$, large effect). Third, uTP proteins show lower instability indices, indicating greater thermodynamic stability ($d = -0.81$, large effect). All three comparisons remain significant after Bonferroni correction for multiple testing.

This finding is unexpected. In canonical protein targeting systems such as chloroplast and mitochondrial import, the transit peptide determines targeting specificity while cargo properties play no role. That mature domain properties predict uTP presence with approximately 90% accuracy suggests constraints on which proteins can undergo uTP-mediated import.

Two alternative explanations could account for this biophysical signature without invoking cargo-specific selection. First, uTP proteins might share recent common ancestry, with the signature simply reflecting inherited properties from a small number of founder genes. We tested this by clustering all *B. bigelowii* proteins into gene families and asking whether uTP proteins concentrate in particular families. Among uTP proteins, 25% share a gene family with at least one other uTP protein, significantly more than the 2% expected by chance (permutation test $p < 0.0001$). This confirms a contribution from shared ancestry. However, uTP proteins span 624 distinct gene families, and 75% of uTP proteins belong to families containing only a single uTP member. Shared ancestry contributes to but cannot fully explain the biophysical signature.

Second, uTP proteins might concentrate in functional categories that happen to share these biophysical properties. We assigned proteins to functional categories using COG annotations and compared uTP versus control proteins within each category. If functional enrichment explained the signature, effect sizes should diminish or disappear when comparing same-function proteins. Instead, the biophysical differences persist within categories (Figure 1F). All eight categories with sufficient sample sizes show the same direction of effect for intrinsic disorder, acidity, and stability. Functional category explains only 0–17% of the biophysical differences depending on the property. Variance partitioning confirms that uTP status explains more unique variance (7.3%) than is shared with functional category (3.4%). The biophysical signature is genuinely associated with uTP status, not an artifact of functional enrichment.

These findings point to dual constraints on uTP-mediated import. The transit peptide must present the conserved three-helix bundle formed by the anchor motifs for recognition by import machinery. The cargo protein must possess compatible biophysical properties: elevated disorder, acidity, and stability. This dual requirement suggests that the evolution of uTP-bearing proteins was shaped not only by the need for a targeting signal but also by selection for cargo properties compatible with transport into the nitroplast.

4.1 Data sources

We obtained proteomics data, genome annotations, and the uTP hidden Markov model (HMM) profile from Coale et al. (Coale et al., 2024). The *B. bigelowii* transcriptome contains approximately 44,000 predicted proteins. Scanning the full proteome with the uTP HMM identified 933 proteins with significant hits (e-value < 0.01), which we term import candidates. Of these, 206 were experimentally validated as enriched inside UCYN-A by quantitative proteomics.

4.2 uTP sequence organization

We used motif discovery to identify conserved elements within uTP sequences. Starting from the 206 experimentally validated proteins, we identified ten sequence motifs within the C-terminal extension. Two motifs (termed anchor 1 and anchor 2) appear near the start of the uTP region and are present in the majority of sequences. We extended this analysis to all 933 import candidates by scanning for these motifs, detecting hits in 745 proteins (80%). Among proteins with detectable motifs, 60% display anchor 2 preceding anchor 1, which we define as the canonical order.

4.3 Structure prediction and analysis

We predicted three-dimensional structures for 47 uTP-containing proteins using AlphaFold3. Structures were aligned using structural superposition, and we computed pairwise root mean square deviation (RMSD) across all pairs. To quantify conservation, we built a consensus structure by averaging atomic positions across aligned structures and calculated positional variance (standard deviation) at each residue position.

Secondary structure assignments were made for each predicted structure. We mapped the anchor motifs to structural elements by comparing motif positions in sequence to helix boundaries in the predicted structures.

4.4 Sequence space analysis

We tested whether uTP sequences form discrete clusters using four clustering methods: hierarchical clustering with k-mer distance, spectral clustering, k-means on protein language model embeddings, and density-based clustering. We evaluated cluster quality using silhouette scores (range -1 to $+1$, with values above 0.5 indicating strong cluster structure) and compared cluster assignments across methods using the adjusted Rand index.

To test whether uTP sequences show more or less structure than expected by chance, we compared real sequences to null sequences generated by shuffling each sequence while preserving amino acid composition. For each of 100 null sequence sets, we computed the silhouette score and compared the distribution to the observed value using a permutation test.

4.5 Mature domain classifier

We trained a classifier to predict uTP presence from mature domain sequences (the functional protein excluding the C-terminal uTP region). The uTP region was identified using the HMM

profile and removed, yielding 605 mature domain sequences.

For the control group, we selected proteins from the *B. bigelowii* proteome that lack uTP. To avoid confounding by proteins with other targeting signals, we filtered candidates using subcellular localization prediction, retaining only proteins predicted to localize to the cytoplasm or nucleus (773 proteins). This ensures that differences between groups reflect uTP-specific properties rather than general features of targeted proteins.

Features were extracted using a protein language model, which encodes each sequence as a 1024-dimensional vector capturing evolutionary and structural information. We trained a logistic regression classifier using five-fold cross-validation with stratified sampling to maintain class proportions. Classifier significance was assessed by permutation testing (1000 permutations). We validated the classifier on the full proteome (933 uTP versus 43,430 non-uTP proteins) and report the area under the receiver operating characteristic curve.

4.6 Biophysical property analysis

We computed biophysical properties for all mature domains: isoelectric point, instability index, and fraction of residues in disordered regions (predicted coil). Effect sizes were quantified using Cohen's d, with values of 0.2, 0.5, and 0.8 corresponding to small, medium, and large effects. All comparisons were corrected for multiple testing using Bonferroni correction.

4.7 Gene family analysis

To assess whether uTP proteins share common ancestry, we clustered all *B. bigelowii* proteins into gene families based on mature domain sequence similarity. We used k-mer frequency vectors (k=3) and hierarchical clustering with a distance threshold corresponding approximately to 40% sequence identity. We then asked whether uTP proteins are more likely to share gene families than expected by chance. The null expectation was estimated by permutation testing (10,000 permutations), randomly reassigning uTP labels while preserving the total number of uTP proteins.

4.8 Functional enrichment and within-category analysis

We assigned proteins to functional categories using COG (Clusters of Orthologous Groups) annotations. To test whether biophysical properties are confounded by function, we compared uTP versus control proteins within each functional category that contained at least ten proteins from each group (eight categories met this criterion). We computed effect sizes within each category and compared them to overall effect sizes. If the biophysical signature were explained by functional enrichment, within-category effect sizes should approach zero.

We performed variance partitioning to quantify how much of the biophysical variation is explained by uTP status, functional category, and their overlap. This analysis decomposes total variance into unique contributions from each factor plus shared variance.

4.9 Statistical framework

Throughout this study, we report both p-values and effect sizes. For continuous comparisons, we use Cohen's d; for classifier performance, we report accuracy, area under the ROC curve, and 95% confidence intervals from bootstrap resampling. Multiple testing correction uses Bonferroni (for pre-specified comparisons) or Benjamini-Hochberg false discovery rate (for exploratory analyses). Permutation tests use 1000 iterations unless otherwise specified. Detailed methods including software versions and parameters are provided in Supplementary Methods.

References

T. H. Coale, V. Loconte, K. A. Turk-Kubo, B. Vanslembrouck, W. K. E. Mak, S. Cheung, A. Ekman, J.-H. Chen, K. Hagino, Y. Takano, T. Nishimura, M. Adachi, M. Le Gros, C. Larabell, and J. P. Zehr. Nitrogen-fixing organelle in a marine alga. *Science*, 384(6692):217–222, Apr. 2024. doi: 10.1126/science.adk1075. URL `https://www.science.org/doi/10.1126/science.adk1075`. Publisher: American Association for the Advancement of Science.

S. Frail, M. Steele-Ogus, J. Doenier, S. L. Y. Moulin, T. Braukmann, S. Xu, and E. Yeh. Genomes of nitrogen-fixing eukaryotes reveal an alternate path for organellogenesis. *Proceedings of the National Academy of Sciences*, 122(33):e2507237122, Aug. 2025. doi: 10.1073/pnas.2507237122. URL `https://www.pnas.org/doi/10.1073/pnas.2507237122`. Publisher: Proceedings of the National Academy of Sciences.

T. Masuda, J. Mareš, T. Shiozaki, K. Inomura, A. Fujiwara, and O. Prášil. *Crocosphaera watsonii* – A widespread nitrogen-fixing unicellular marine cyanobacterium. *Journal of Phycology*, 60(3):604–620, June 2024. ISSN 0022-3646, 1529-8817. doi: 10.1111/jpy.13450. URL `https://onlinelibrary.wiley.com/doi/10.1111/jpy.13450`.

H.-W. Pi. Origin and Evolution of Nitrogen Fixation in Prokaryotes. doi: 10.1093/molbev/msac181.

K. A. Turk-Kubo, V. Loconte, B. Vanslembrouck, W. K. E. Mak, A. Ekman, J.-H. Chen, Y. Takano, T. Horiguchi, T. Nishimura, M. Adachi, M. L. Gros, K. Hagino, J. P. Zehr, and C. Larabell. Soft X-ray Tomography Enables New Insights into the Coordinated Division of Organelle-like Symbiont in a Globally Distributed Unicellular Marine Haptophyte Alga. *Microscopy and Microanalysis*, 29(Supplement_1):1165, Aug. 2023. ISSN 1431-9276. doi: 10.1093/micmic/ozad067.596. URL `https://doi.org/10.1093/micmic/ozad067.596`.