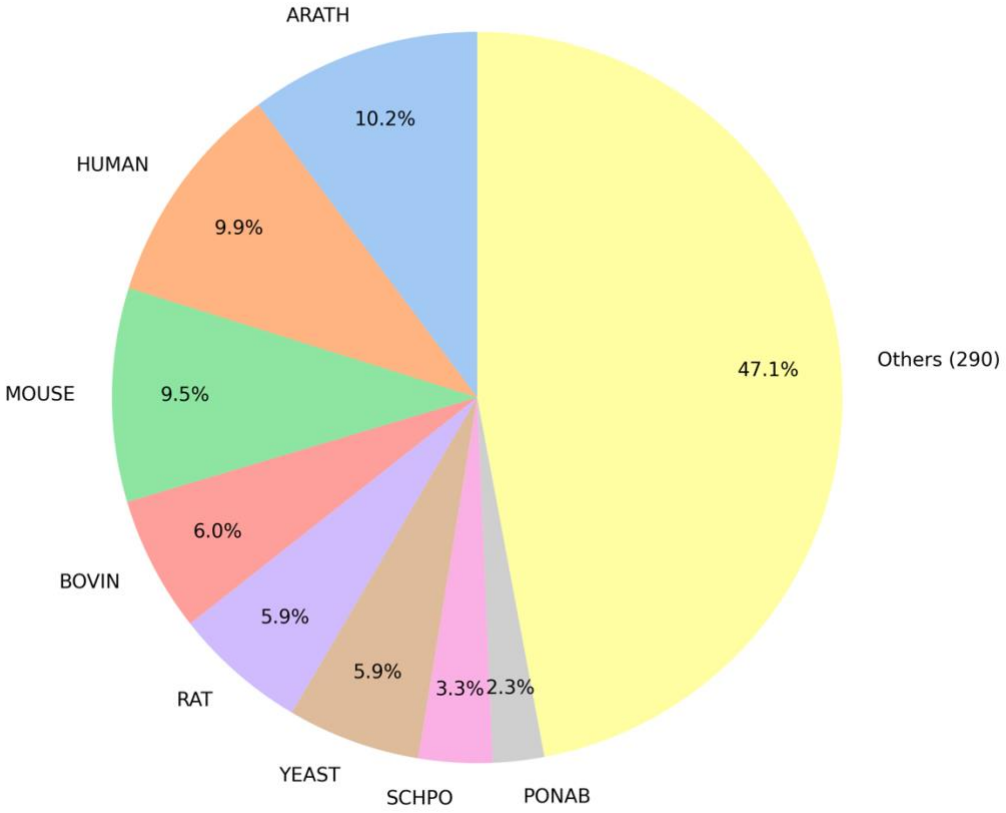
Supplementary Information for

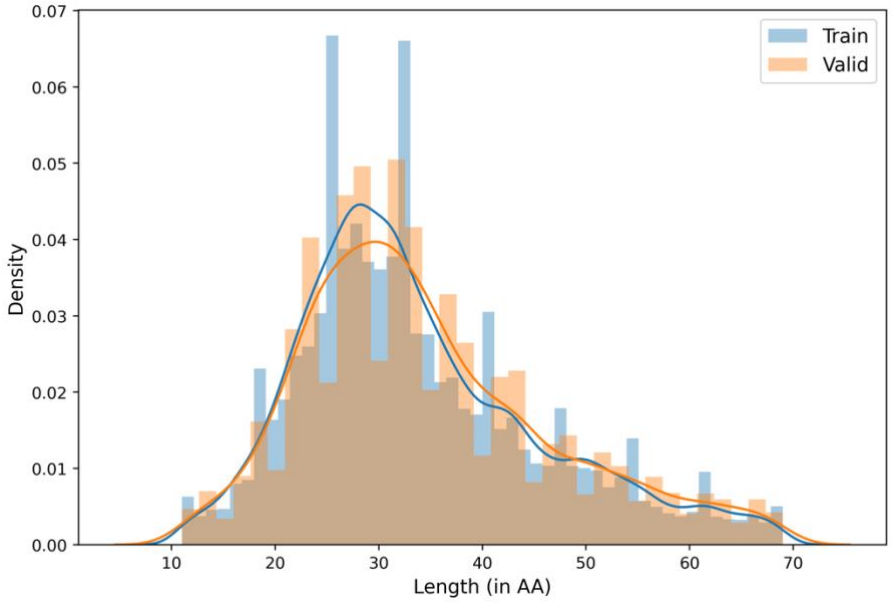
Design of diverse, functional mitochondrial targeting sequences across eukaryotic organisms using variational autoencoder

Boob et al.

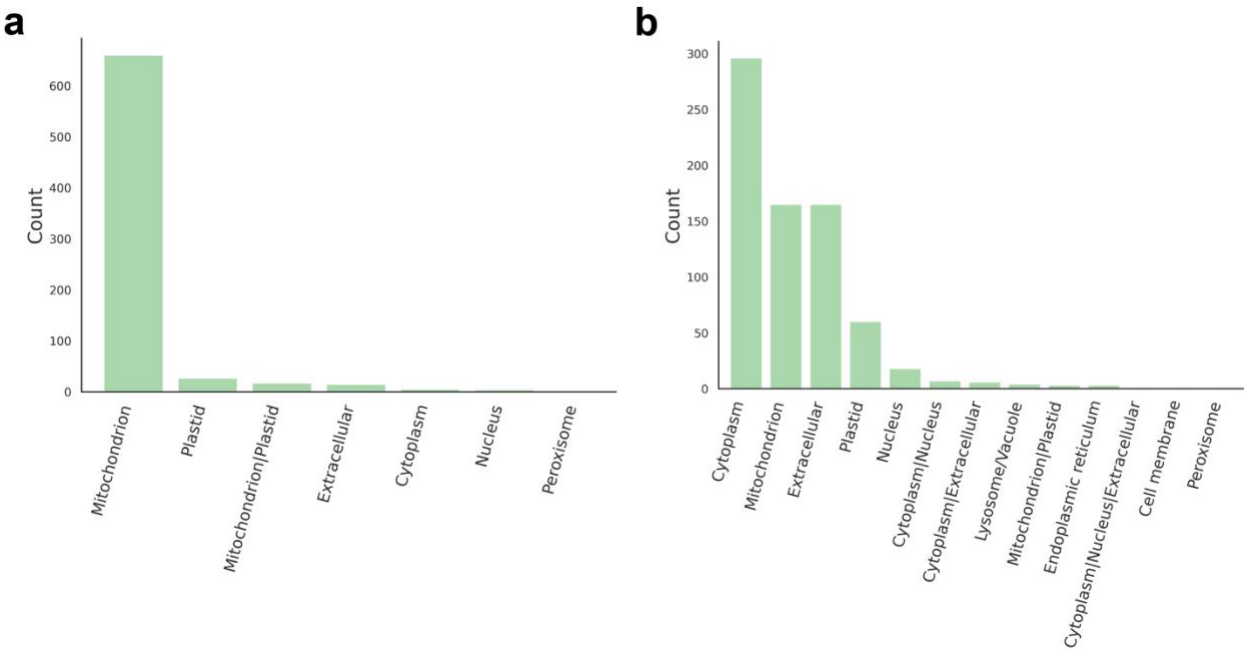
This document includes:



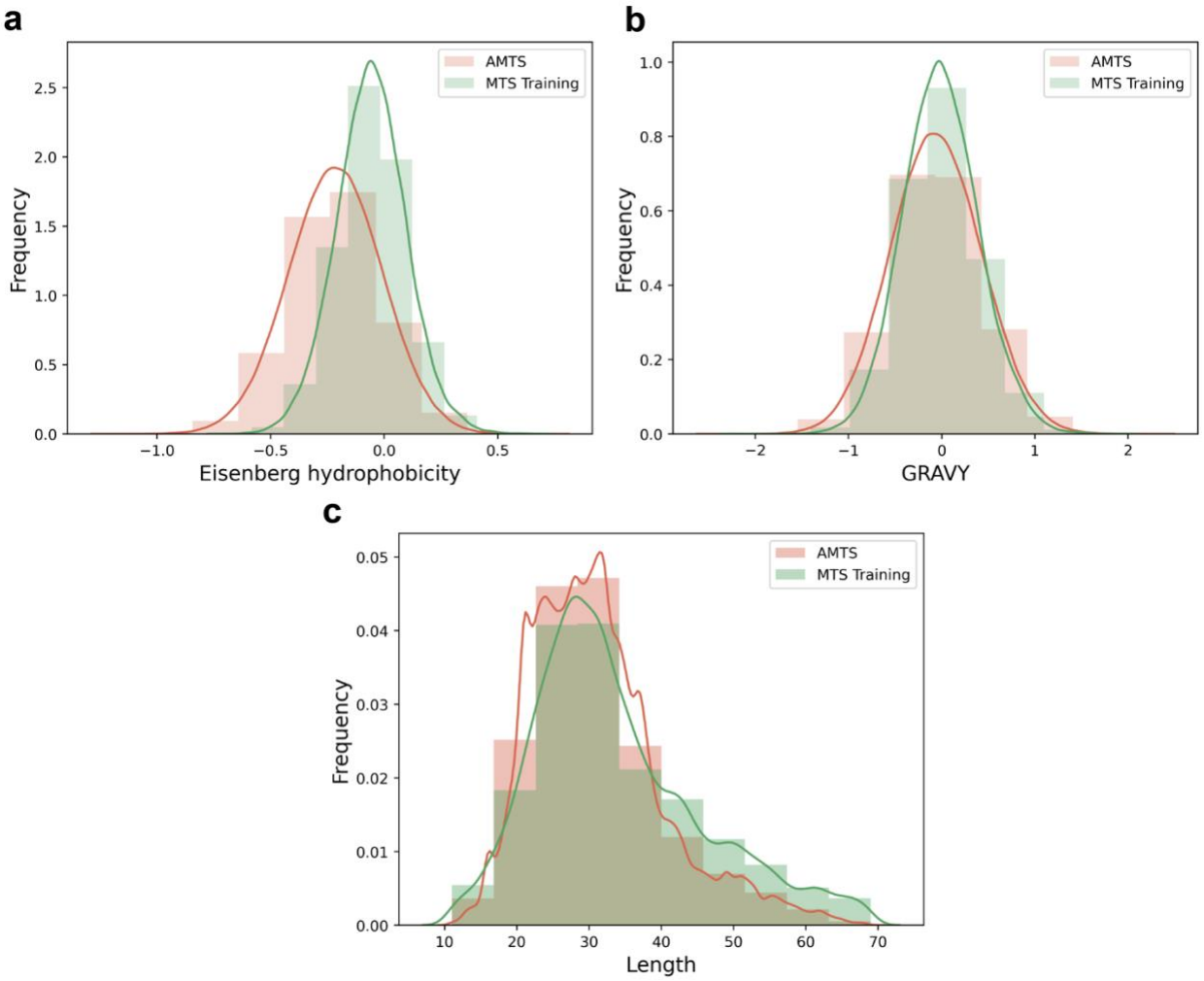
Supplementary Fig. 1. Species distribution for 4,984 MTSs reported in the Swiss-Prot database. More than half of these sequences are from eight model organisms: *Arabidopsis thaliana* (ARATH), *Homo sapiens* (HUMAN), *Mus musculus* (MOUSE), *Bos taurus* (BOVIN), *Rattus norvegicus* (RAT), *Saccharomyces cerevisiae* (YEAST), *Schizosaccharomyces pombe* (SCHPO), and *Pongo abelii* (PONAB), while the rest come from 290 different species.



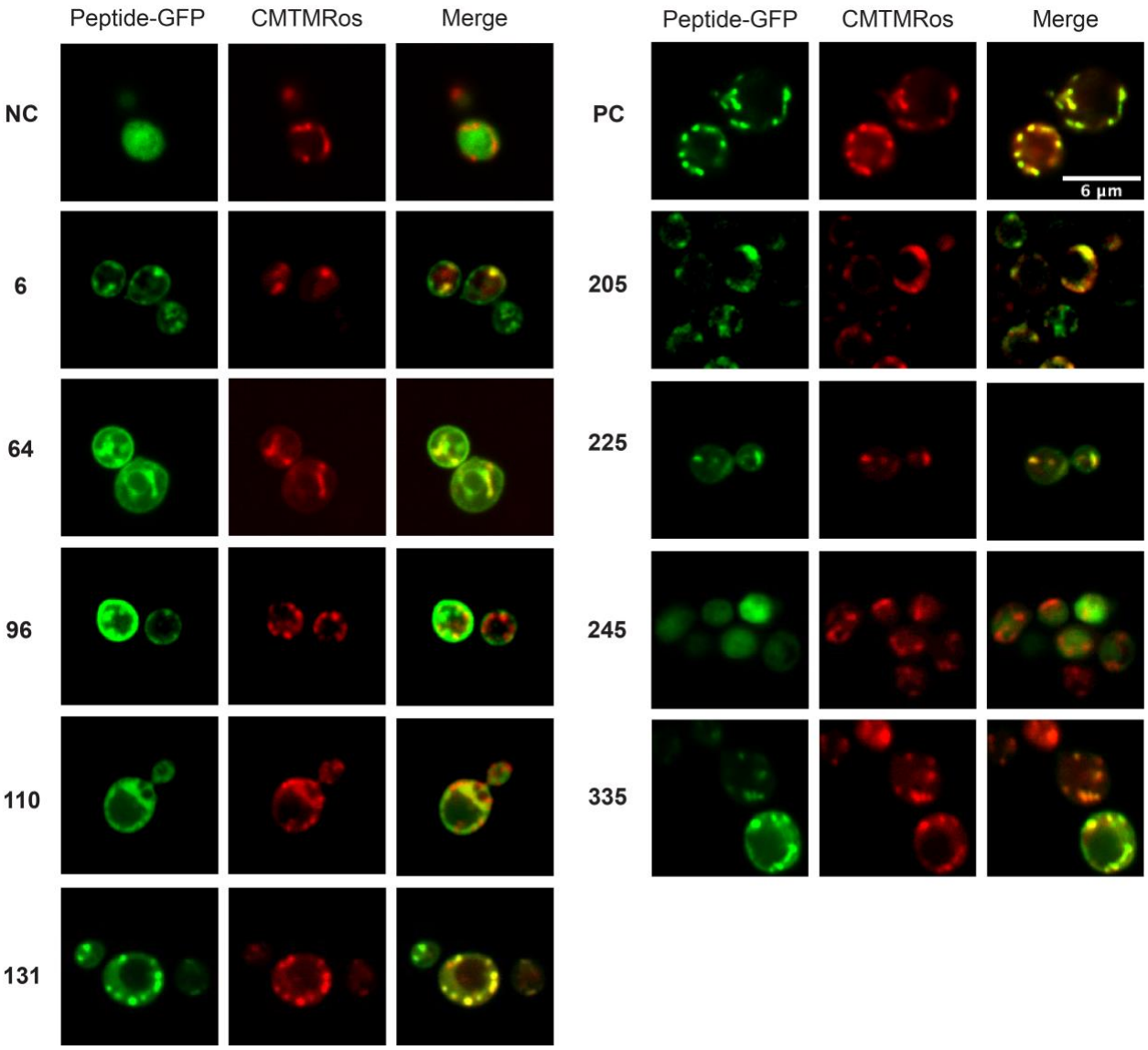
Supplementary Fig. 2. Stratified split of the curated MTS dataset based on peptide length. The dataset, consisting of 56,660 MTSs, was divided into training and validation sets at a ratio of 9:1, ensuring the Variational Autoencoder (VAE) learns from MTSs of different lengths and avoids potential bias toward shorter lengths.



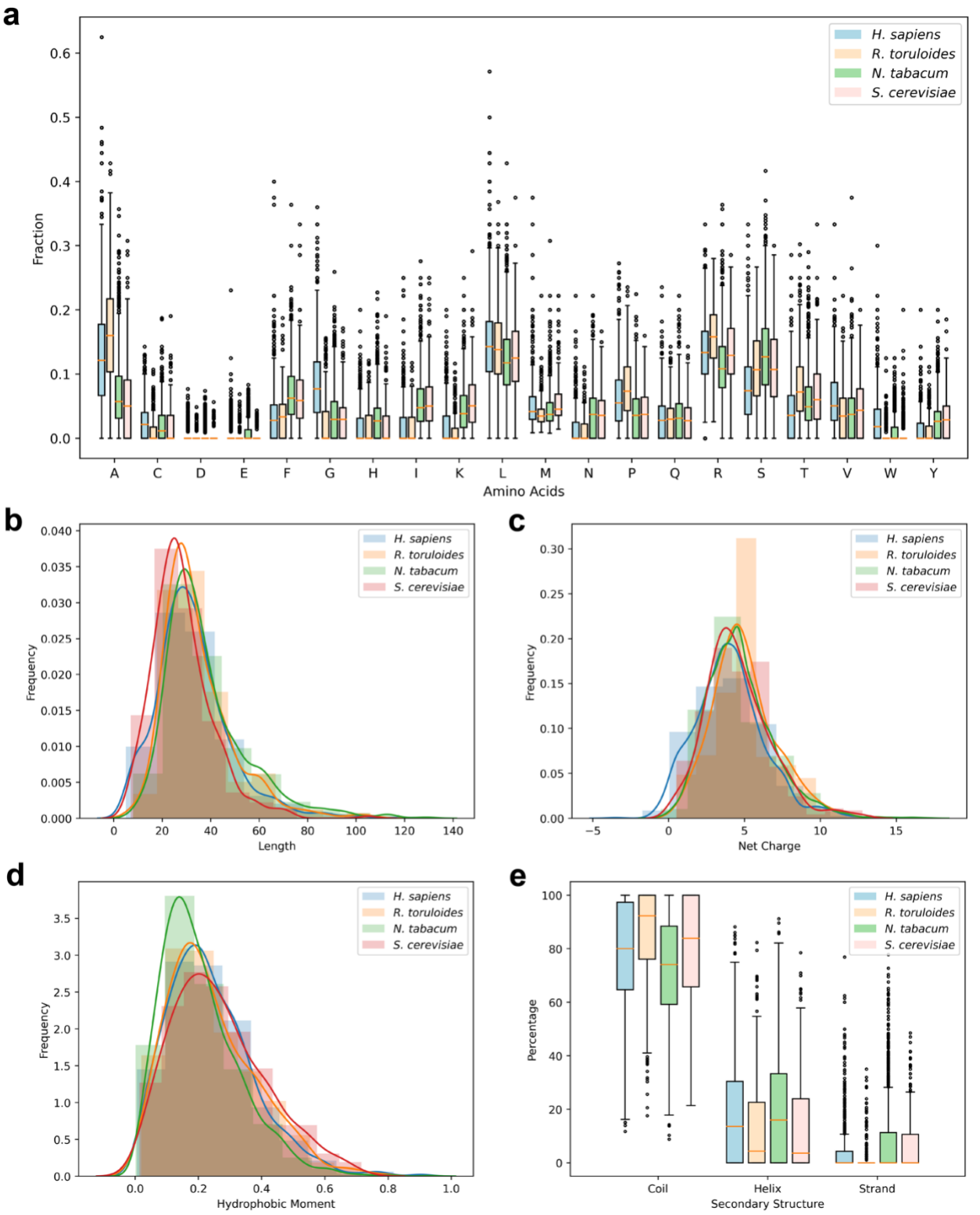
Supplementary Fig. 3. Predicted subcellular localization of **a)** VAE and **b)** pHMM generated MTSs. A set of 730 artificial MTSs were generated using both models, appended with a GFP sequence, and subjected to analysis using DeepLoc 2.0. 660 (90.41%), 27 (3.7%), and 17 (2.33%) VAE-generated peptides were predicted to target GFP to the mitochondrion, plastid, or both organelles, respectively. In the case of pHMM, 161 MTSs-tagged GFP proteins (22.6%) demonstrated the capability to localize proteins to the mitochondrion.



Supplementary Fig. 4. Comparison of two physicochemical properties: **a)** Eisenberg hydrophobicity, **b)** GRAVY, and **c)** Length between naturally occurring (n = 50,980) and VAE-generated MTSs (n = 705,081). Although the VAE-generated sequences have a slight bias towards negative Eisenberg hydrophobicity values, they cover complete range of values for all attributes. The physicochemical characteristics of the targeting sequences were calculated using BioPython.

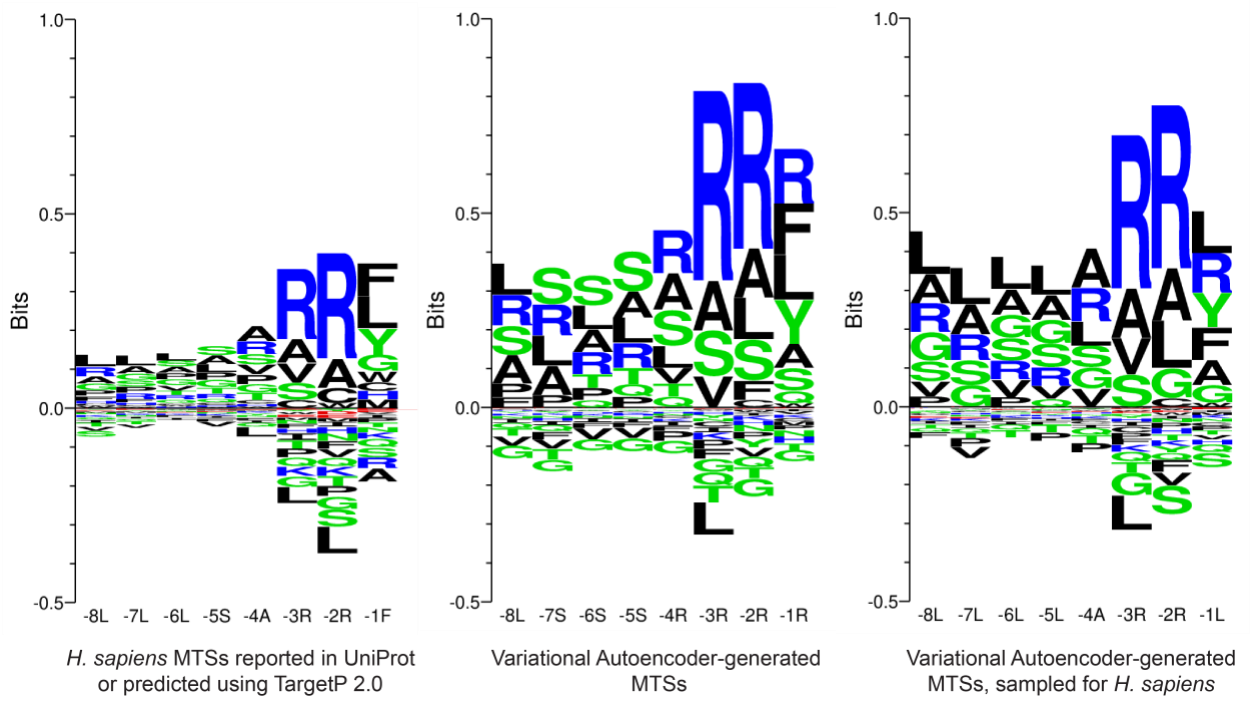


Supplementary Fig. 5. Imaging *Saccharomyces cerevisiae* strains expressing AMTS-GFP constructs. The untagged GFP and COX4-GFP were included as negative and positive controls, respectively. Plasmids harboring MTS-GFP constructs were transformed into BY4741 strain, followed by verification of mitochondrial localization using fluorescence microscopy. Mitochondria were stained with MitoTracker™ Orange CMTMRos. Scale bar: 6 µm.

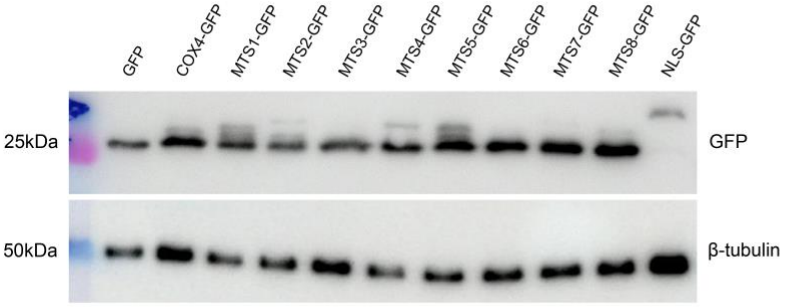


Supplementary Fig. 6. Comprehensive analysis of MTSs in four eukaryotic organisms: *H. sapiens*, *R. toruloides*, *N. tabacum*, and *S. cerevisiae*. Proteome-wide MTSs were acquired using TargetP 2.0. Subsequently, **a)** amino acid composition, **b)** length, **c)** net charge, **d)** hydrophobic moment, and **e)**

structural characteristics were obtained using BioPython, modlAMP, and s4pred, respectively. MTSs from all four organisms were positively charged, amphiphilic, and formed α-helix. However, variations were observed in amino acid preferences. For example, *H. sapiens*, *R. toruloides*, and *N. tabacum* displayed an increased bias toward Glycine (G), Alanine (A), and Serine (S), respectively. Moreover, MTSs in different organisms also have a varying fraction capable of forming α-helix. All boxplots follow the standard definition: the center line represents the median, the box limits correspond to the upper and lower quartiles, the whiskers extend to 1.5 times the interquartile range, and outliers are shown as points.

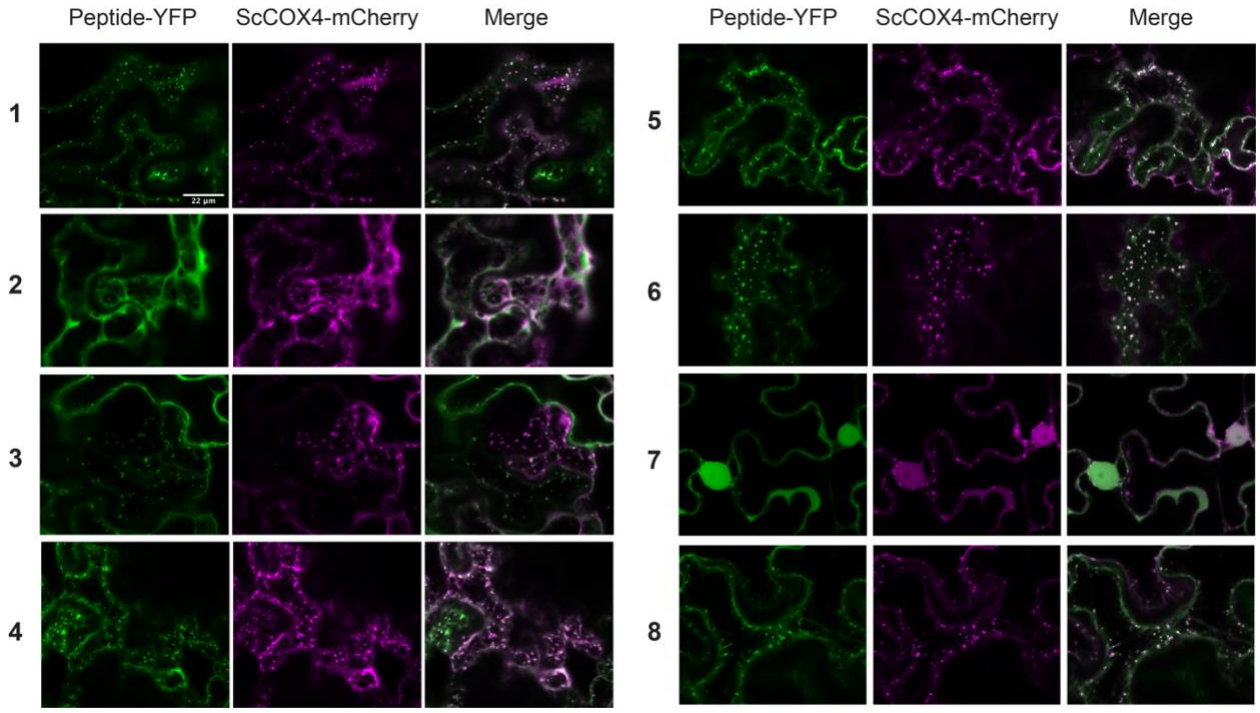


Supplementary Fig. 7. Organism label-based sampling of MTSs captures better cleavage specificity. The C-termini of MTSs from *H. sapiens* were analyzed alongside the VAE-generated ones selected randomly or sampled closer to the cluster center for MTSs of *H. sapiens*. The consensus of MTSs from *H. sapiens* (-8L, -7L, -6L, -5S, -4A, -3R, -2R, -1F) closely resembled with the sampled MTSs (-8L, -7L, -6L, -5L, -4A, -3R, -2R, -1L) compared to the randomly selected ones (-8L, -7S, -6S, -5S, -4R, -3R, -2R, -1R). In addition, the favorable occurrence of Glycine (G), an amino acid residue preferred by MTSs in the human proteome, was observed only in sampled MTSs and not in randomly generated counterparts. WebLogo[1] was used to create the logos with the default color scheme, where the height of each letter corresponds to the relative frequency of that amino acid.

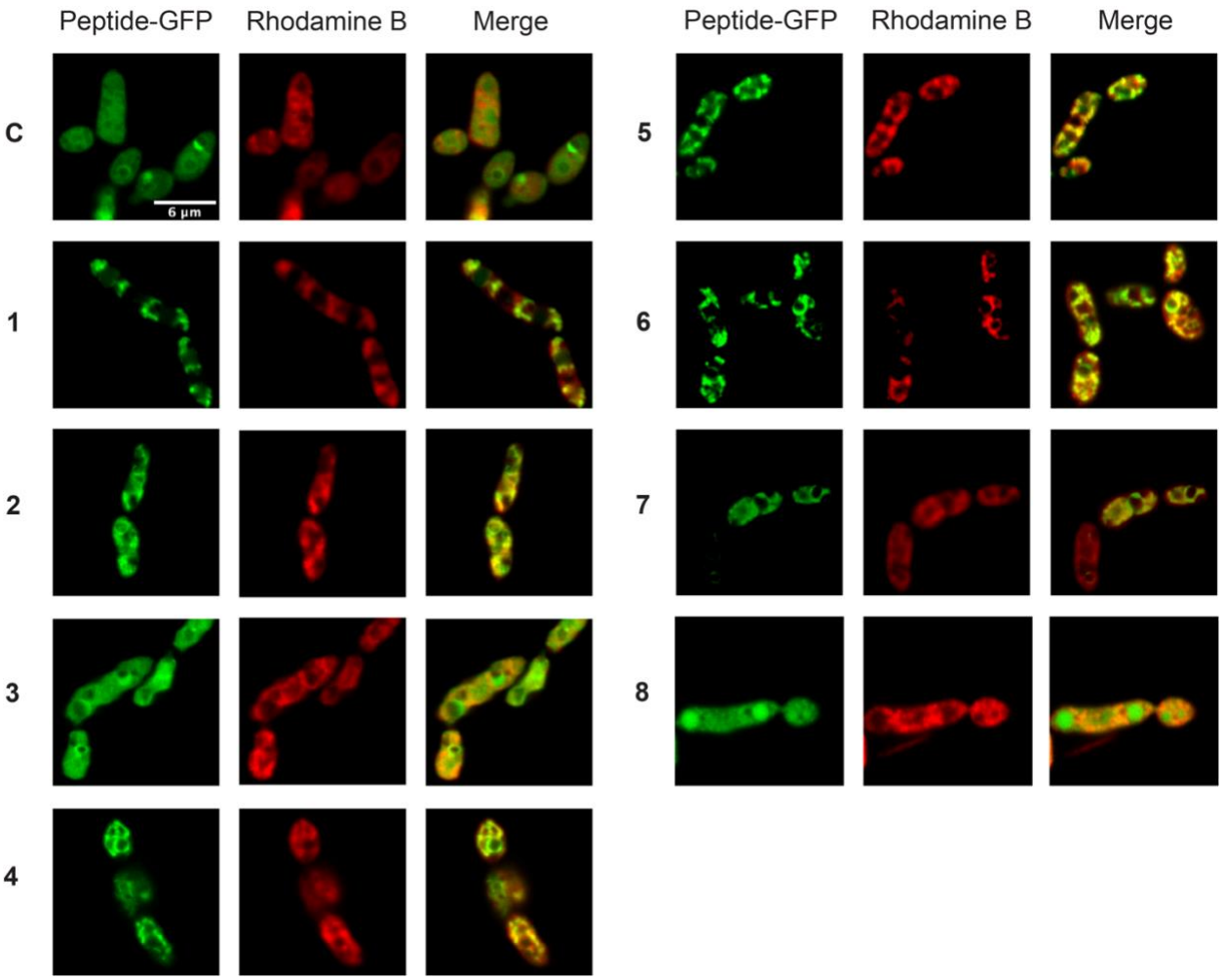


Name	Predicted Molecular Weight	
	Non-cleaved	Cleaved
GFP	26.8	
COX4-GFP	29.2	26.8
MTS1-GFP	29.8	26.8
MTS2-GFP	30.3	26.8
MTS3-GFP	29.5	26.8
MTS4-GFP	29.8	26.8
MTS5-GFP	29.6	26.8
MTS6-GFP	30.2	26.8
MTS7-GFP	30.3	26.8
MTS8-GFP	29.8	26.8
NLS-GFP	30	

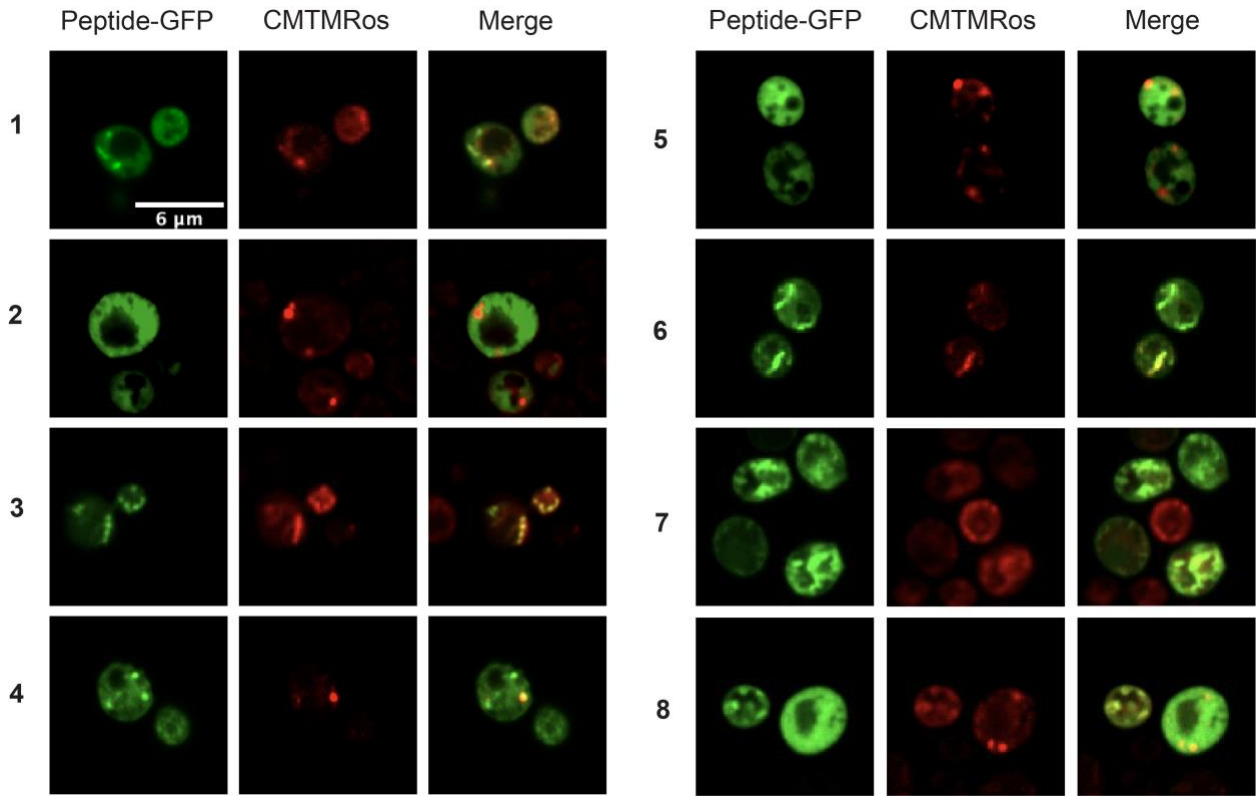
Supplementary Fig. 8. *In vivo* cleavage of artificial peptides in the HEK293 cell line. HEK293 cells were transfected with GFP expression plasmids tagged with various mitochondrial targeting sequences. Cell lysates were subjected to Western blot analysis using GFP and β-tubulin antibodies. All artificial MTS-tagged GFP proteins display complete cleavage, as evidenced by bands observed with plasmids containing only GFP and COX4 MTS-tagged GFP, a known fully cleaved control. In contrast, the nuclear localization sequence (NLS)-GFP remains uncleaved and migrates at a higher molecular weight than other GFP constructs. The table provides predicted molecular weights for non-cleaved and cleaved proteins. Source data are provided as a Source Data file.



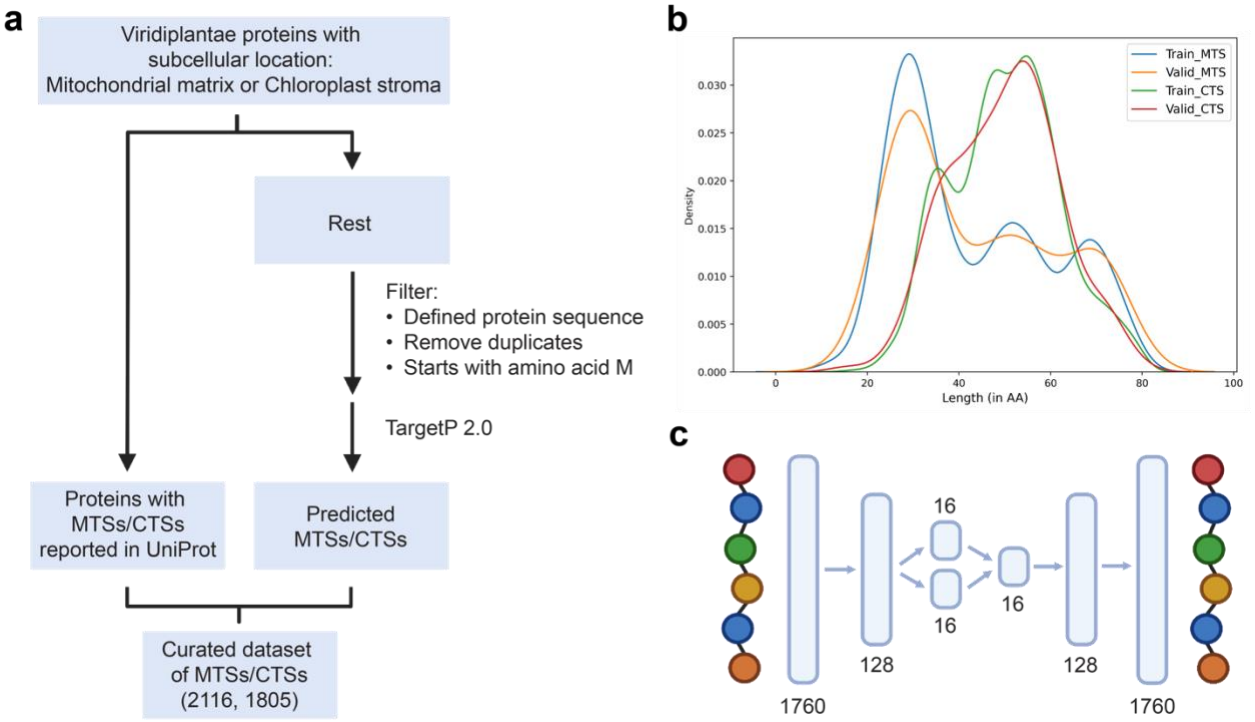
Supplementary Fig. 9. Subcellular localization of artificial MTSs in *N. benthamiana*'s leaf epidermal cells. AMTS-YFP constructs were transiently expressed in the tobacco epidermis driven by the CaMV 35S promoter. The mitochondria marker CD3-991 (35S::ScCOX4-mCherry) was obtained from ABRC[2]. AMTS-YFP and ScCOX4-mCherry were co-infiltrated into *N. benthamiana* leaves. Six out of eight peptides (AMTS 1, 3, 4, 5, 6, and 8) show co-localization with the mitochondria marker ScCOX4-mCherry. Scale bar: 22 µm.



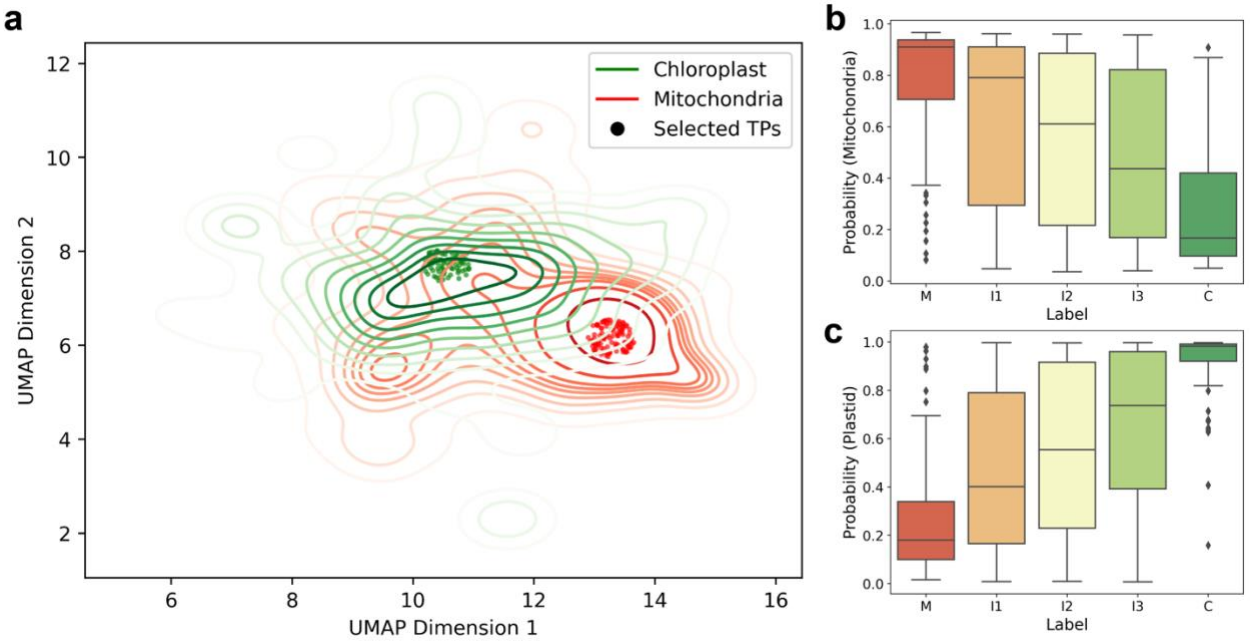
Supplementary Fig. 10. Confocal microscopy of *R. toruloides* strains expressing AMTS-GFP constructs. The untagged GFP and RtCOX4-GFP were included as controls. Plasmids harboring MTS-GFP constructs were transformed into 880CF-CAR2Δ strain, followed by verification of mitochondrial localization using fluorescence microscopy. Mitochondria were stained with Rhodamine B, hexyl ester. Six peptides (AMTS) successfully localized GFP to mitochondria. Scale bar: 6 μm.



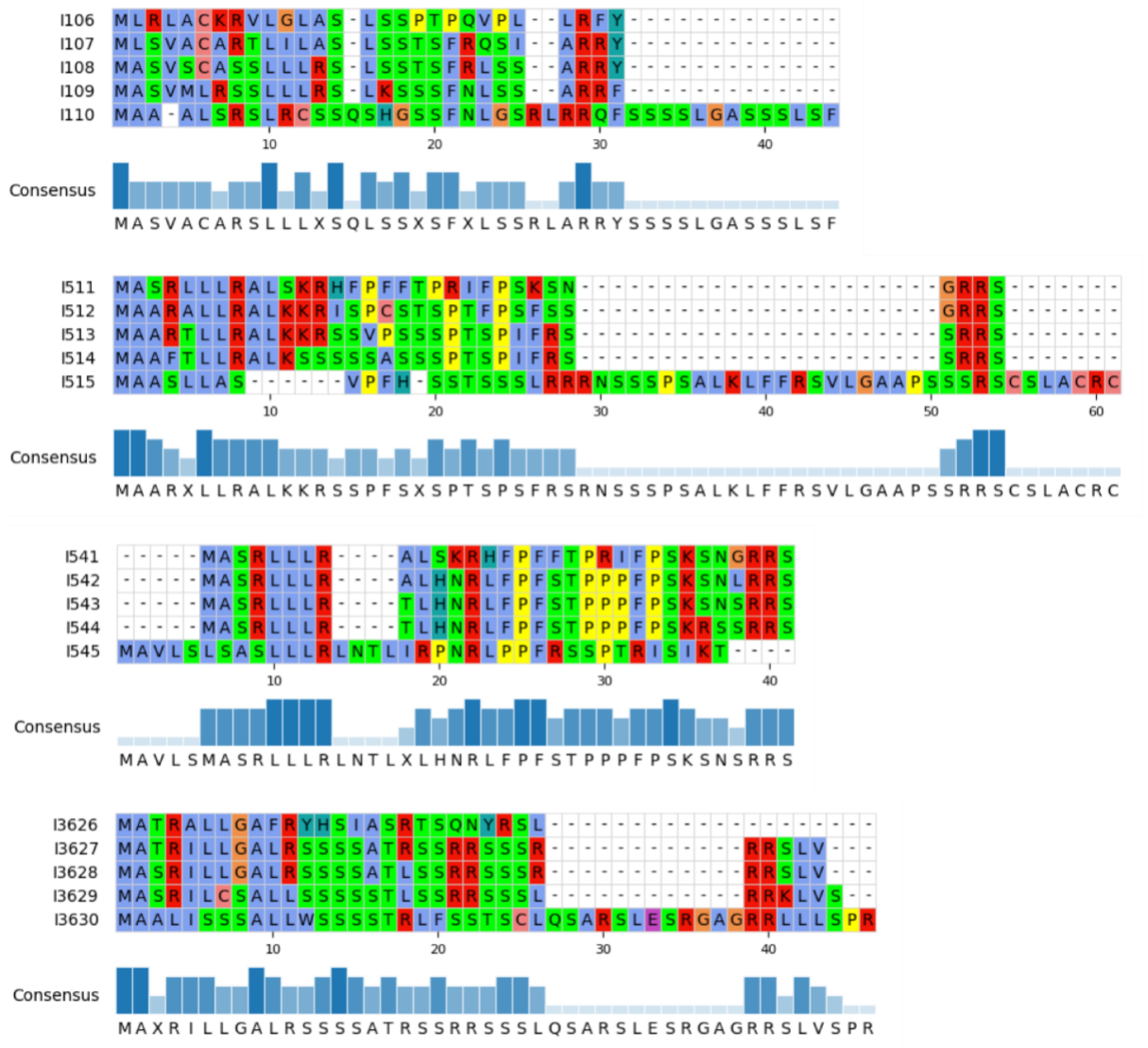
Supplementary Fig. 11. Confocal microscopy of *S. cerevisiae* strains expressing AMTS-GFP constructs. Plasmids harboring AMTS-GFP constructs were transformed into BY4741 strain, followed by verification of mitochondrial localization using fluorescence microscopy. Mitochondria were stained with MitoTracker™ Orange CMTMRos. Scale bar: 6 µm.



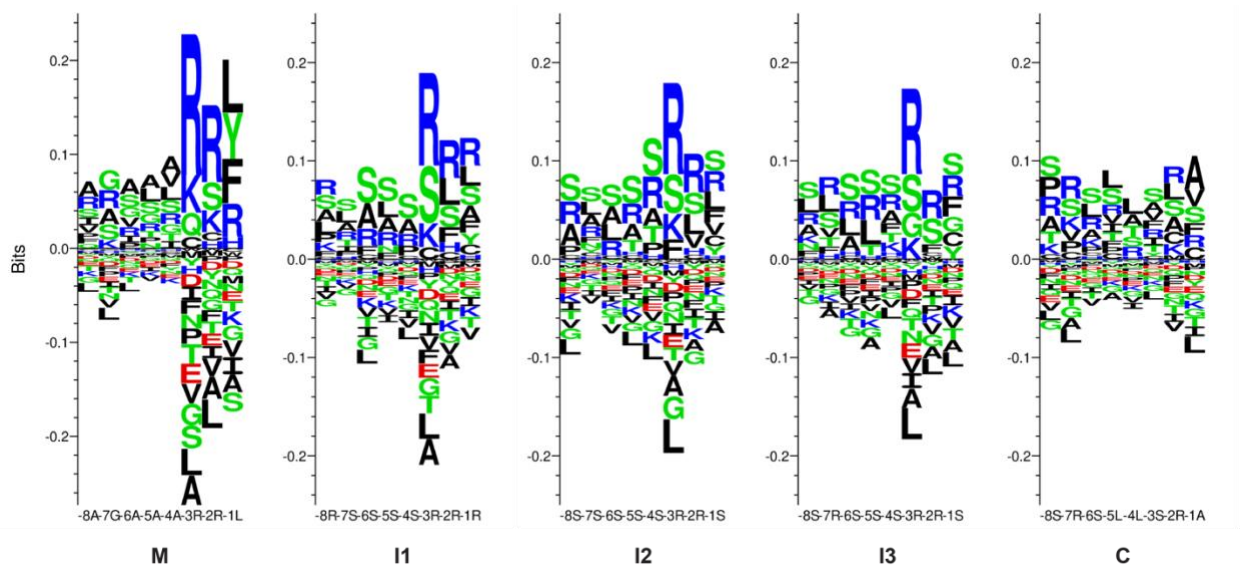
Supplementary Fig. 12. Variational Autoencoder for designing dual-targeting sequences. **a)** Data acquisition for training the Dual-VAE model. MTSs and CTSs belonging to the Viridiplantae kingdom were obtained either directly from UniProt or discovered by analyzing proteins with subcellular locations in the mitochondrial matrix and chloroplast stroma using TargetP 2.0. Sequences were filtered to exclude those containing undefined amino acids (X) or lacking a Methionine (M) at the start, in addition to removing duplicates and sequences exceeding 80 amino acids in length. **b)** Dividing the curated dataset into training and validation sets. MTSs and CTSs were individually stratified based on their length, employing a ratio of 9:1, before combining to create the training and validation data for the Dual-VAE model. **c)** Architecture of the Dual-VAE model. **c** Created in BioRender. Zhao, H. (2025) https://BioRender.com/y8ekth5.



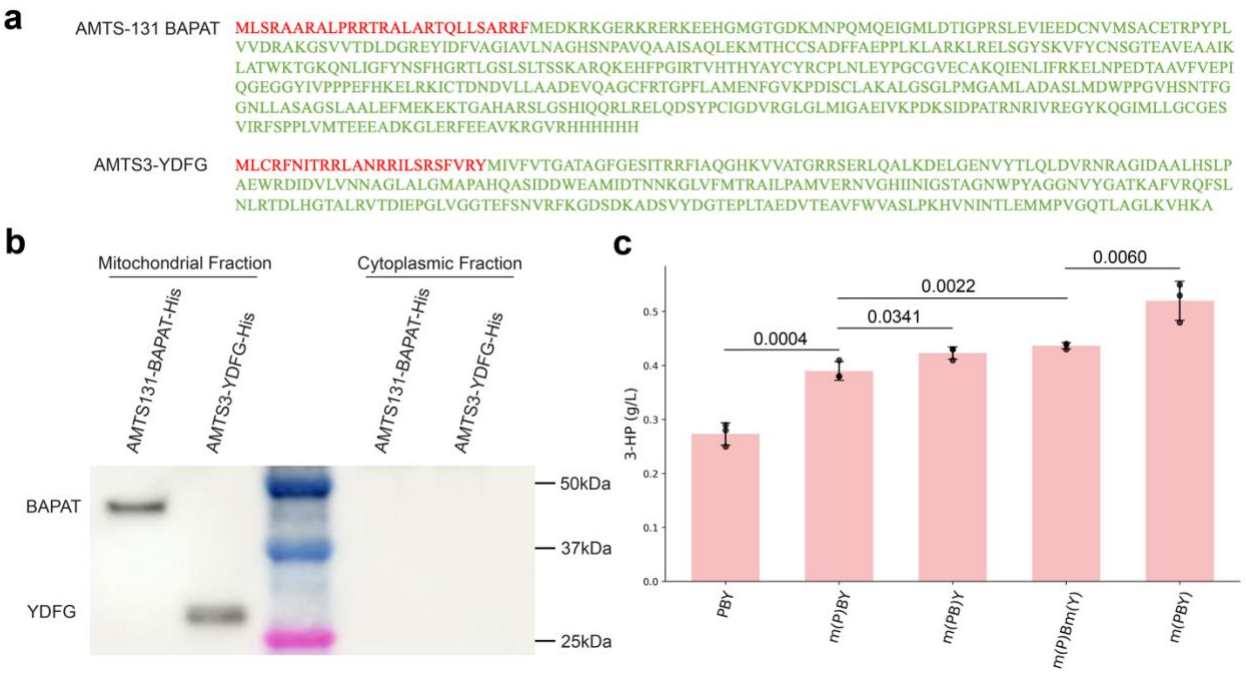
Supplementary Fig. 13. Functional analysis of interpolated peptides designed using Dual-VAE. **a)** Latent space of Dual-VAE visualized in 2D using UMAP. Over 50 MTSs and CTSs were obtained near the center of the posterior distribution, employing K-Means and a distance threshold of 0.4. Latent vectors were obtained by walking along the path, joining a pair of MTS and CTS, and fed to the decoder of Dual-VAE to generate peptides. **b)** DeepLoc 2.0 probabilities for the interpolated peptides (n = 5100) to localize GFP to mitochondrion. **c)** DeepLoc 2.0 probabilities for the interpolated peptides to localize GFP to the chloroplast. The data in **b** and **c** shows a decreasing trend in the ability to target mitochondrion with a simultaneous increase in the probability of targeting chloroplast when moving along the interpolation path from MTS to CTS. All boxplots follow the standard definition: the center line represents the median, the box limits correspond to the upper and lower quartiles, the whiskers extend to 1.5 times the interquartile range, and outliers are shown as points.



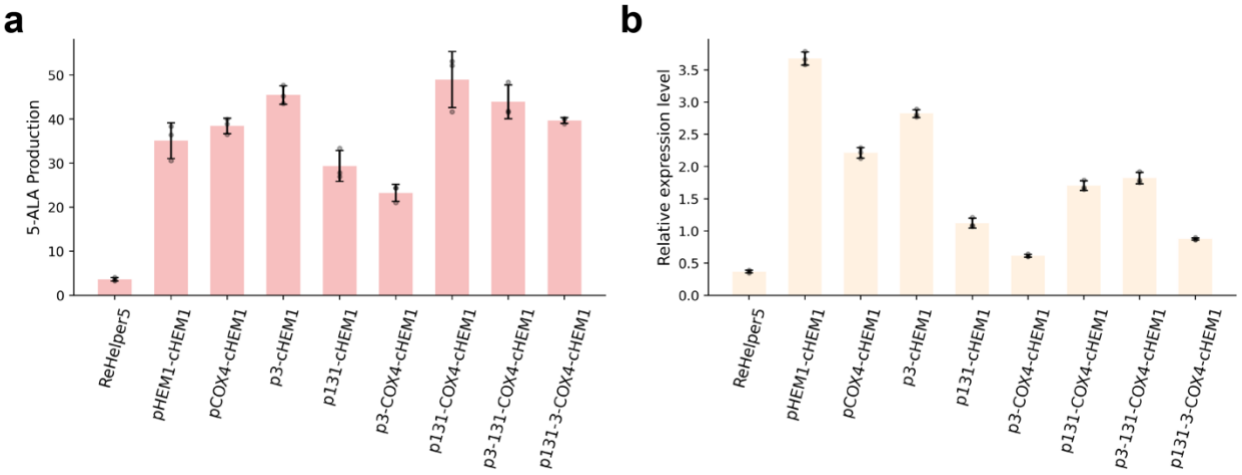
Supplementary Fig. 14. Multiple sequence alignment of the sequences generated along the interpolation path for the four representative dual-targeting peptides. In most cases, the transition from MTSs to dual-targeting peptides revealed substitutions, while the trajectory from CTSs indicated both substitutions and deletions. Clustal Omega[3] and pyMSAviz[4] were used to create and visualize MSAs, respectively, using the 'Clustal' color scheme.



Supplementary Fig. 15. Analyzing the evolutionary transition in the cleavage site of dual-targeting peptides. The last ten amino acids of the sequences generated along the interpolation path for the 62 putative dual-targeting peptides were analyzed. The interpolated peptides display an RR, R-2, or R-3 motif and, therefore, might undergo cleavage by mitochondrial processing peptidase (MPP) upon import into the mitochondrial matrix. In addition, the preference for Alanine (A) residue decreases from left to right. WebLogo[1] was used to create the logos with the default color scheme, where the height of each letter corresponds to the relative frequency of that amino acid.



Supplementary Fig. 16. Additional controls showcasing mitochondrial localization of AMTS-fusion proteins and the importance of compartmentalizing all enzymes for enhanced 3-HP production. **a)** Amino acid sequences of the AMTS131-BAPAT and AMTS3-YDFG enzymes. The AMTS sequences are highlighted in red, and the enzyme sequences are highlighted in green. **b)** Western blot analysis of cytosolic and mitochondrial fractions confirms the successful mitochondrial localization of the AMTS131-BAPAT and AMTS3-YDFG enzymes. **c)** An additive increase in 3-HP production is observed when AMTS sequences are used to target additional pathway enzymes to mitochondria. m(X) indicates enzymes targeted to mitochondria. Data represents mean ± s.d. n = 3 biological replicates. *P* value was calculated by two-tailed unpaired t-test. Source data are provided as a Source Data file.



Supplementary Fig. 17. Chimeric MTSs for protein delivery. **a)** 5-ALA production and **b)** mRNA levels of the MTS/chimeric MTS-cHEM1 gene constructs. Data represents mean ± s.d. n = 3 biological replicates. Source data are provided as a Source Data file.

Supplementary References

1. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).
2. Nelson, B. K., Cai, X. & Nebenführ, A. A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. *The Plant Journal* **51**, 1126–1136 (2007).
3. McWilliam, H. *et al.* Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research* **41**, W597–W600 (2013).
4. Shimoyama, Y. pyMSAviz: MSA visualization python package for sequence analysis. (2022).