


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

Received: 28 August 2024

Accepted: 16 April 2025

Published online: 04 May 2025



Aashutosh Girish Boob<sup>1,2,3</sup>, Shih-I Tan<sup>1,2,3</sup>, Airah Zaidi<sup>1</sup>, Nilmani Singh<sup>2,3</sup>, Xueyi Xue<sup>3,4</sup>, Shuaizhen Zhou<sup>2,3</sup>, Teresa A. Martin<sup>1,2,3</sup>, Li-Qing Chen<sup>3,4</sup> & Huimin Zhao<sup>1,2,3</sup> 

Mitochondria play a key role in energy production and metabolism, making them a promising target for metabolic engineering and disease treatment. However, despite the known influence of passenger proteins on localization efficiency, only a few protein-localization tags have been characterized for mitochondrial targeting. To address this limitation, we leverage a Variational Autoencoder to design novel mitochondrial targeting sequences. In silico analysis reveals that a high fraction of the generated peptides (90.14%) are functional and possess features important for mitochondrial targeting. We characterize artificial peptides in four eukaryotic organisms and, as a proof-of-concept, demonstrate their utility in increasing 3-hydroxypropionic acid titers through pathway compartmentalization and improving 5-aminolevulinate synthase delivery by 1.62-fold and 4.76-fold, respectively. Moreover, we employ latent space interpolation to shed light on the evolutionary origins of dual-targeting sequences. Overall, our work demonstrates the potential of generative artificial intelligence for both fundamental research and practical applications in mitochondrial biology.

A eukaryotic cell is a highly intricate entity, comprising multiple organelles responsible for various cellular processes. These organelles are enclosed in membranes and provide an optimal physicochemical environment to a multitude of nuclear-encoded proteins to carry out essential functions. Therefore, accurate localization of these proteins is crucial for maintaining cellular organization, ensuring correct metabolism, and thereby facilitating the efficient functioning of a eukaryotic cell. In nature, protein localization relies on intricate signals stored in the amino acid sequence. Over the years, significant progress has been made in understanding protein localization, aided by the characterization of localization tags in model organisms<sup>1,2</sup> and the recent development of supervised machine learning (ML) models<sup>3,4</sup>. These advances have substantially improved our ability to target proteins to desired subcellular locations, paving the way for

applications in synthetic biology<sup>5</sup>, chemical production<sup>6,7</sup>, and therapeutic interventions<sup>8</sup>.

The multifaceted contributions of mitochondria to cellular metabolism have made them an important target for metabolic engineering<sup>9</sup> and disease treatment<sup>10</sup>. Mitochondria, the powerhouse of eukaryotic cells, play a pivotal role in various cellular processes, including energy production, metabolism, and apoptosis. They harbor the tricarboxylic acid (TCA) cycle and are responsible for the biosynthesis of several cofactors and metabolites. As a result, several metabolic pathways have been localized to the mitochondria, utilizing the precursor pool to boost the production of various fuels, precursor chemicals, and pharmaceuticals<sup>9</sup>. The mitochondrial matrix provides a distinctive physiological environment, characterized by higher pH, lower oxygen concentration, and higher reducing redox potential than

<sup>1</sup>Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA. <sup>2</sup>Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. <sup>3</sup>DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL, USA. <sup>4</sup>Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. ✉ e-mail: [zhao5@illinois.edu](mailto:zhao5@illinois.edu)

the cytosol, that has proven advantageous for the expression of exogenous biocatalytic machinery<sup>5</sup>. Moreover, dysfunctional mitochondria, particularly those harboring inherited or acquired mutations in mitochondrial DNA (mtDNA) are associated with many diseases<sup>11</sup>, including Leber's hereditary optic neuropathy (LHON), Mitochondrial Encephalopathy, Lactic Acidosis, and Stroke-like episodes (MELAS) syndrome, Leigh syndrome and others. Therefore, improving mitochondrial targeting will also aid in delivering necessary nucleases and drug molecules efficiently.

To direct proteins to mitochondria, a targeting peptide needs to be added to the N-terminus of the nuclear-encoded protein<sup>12,13</sup>. This positively charged, amphiphilic peptide is recognized and imported by the multi-subunit translocase of the outer and inner mitochondrial membrane, TOM and TIM complex. Upon arrival in the matrix, the peptide is cleaved by a mitochondrial processing peptidase (MPP), after which the protein folds into its mature form (Fig. 1a). However, only a few of these peptides are well-characterized, resulting in their repeated use. Given that the targeting efficiency of the peptide is dependent on the passenger protein<sup>14</sup>, use of a suboptimal mitochondrial targeting sequence (MTS) can lead to little to no import of the desired enzyme and result in partial compartmentalization. Furthermore, overuse of a single, endogenous targeting sequence may saturate the import machinery<sup>15</sup>, resulting in potential competition with native proteins during the import system, and its downstream accumulation can compromise mitochondrial protein biogenesis and integrity<sup>16,17</sup>. Additionally, recurrent use of the targeting sequence to compartmentalize enzymes of a multi-gene metabolic pathway integrated into a strain can suffer from genetic instability arising from homologous recombination between sequences with high similarity. Therefore, it is highly desirable to design and characterize a toolkit of diverse, functional mitochondrial targeting sequences.

In recent years, ML approaches have shown tremendous promise in peptide design. Specifically, deep generative models have demonstrated their ability to capture intricate patterns solely from sequence data<sup>18</sup>. By learning from vast datasets of unlabeled peptides, these models can generate novel sequences that adhere to the underlying rules governing peptide function and structure. Therefore, various model architectures are employed to design antimicrobial peptides<sup>19</sup>, signal peptides<sup>20</sup>, and cell-penetrating peptides<sup>21</sup>. In this study, we trained a Variational Autoencoder (VAE) to generate new-to-nature mitochondrial targeting sequences. We performed experiments to validate the functionality of these sequences in four eukaryotic organisms, including *Saccharomyces cerevisiae*, *Rhodotorula toruloides*, *Nicotiana benthamiana*, and the HEK293 cell line. Furthermore, we harnessed the latent space interpolation capabilities of the VAE to create putative sequences capable of targeting mitochondria and chloroplasts simultaneously. Through an extensive analysis of the generated sequences, we extracted insights to inform evolutionary hypotheses regarding dual-targeting sequences. Finally, we showcased the practical applications of these artificially designed sequences, highlighting their effectiveness in pathway compartmentalization and enhancing our ability to target mitochondria.

## Results

### Creation of a variational autoencoder for designing artificial mitochondrial targeting sequences

Mitochondrial targeting sequences (MTSs) are positively charged peptides with a tendency to form amphiphilic  $\alpha$ -helices<sup>12,13</sup>. These peptides are 10–120 amino acids long (with a typical length of approximately 35 amino acids) and are located at the N-terminus of the proteins destined for mitochondria. Unlike organelles such as peroxisomes, these targeting sequences do not exhibit specific motifs and instead rely on physicochemical and structural characteristics for their recognition and import into mitochondria<sup>13</sup>. This makes the design of a diverse peptide library challenging, considering the extensive design

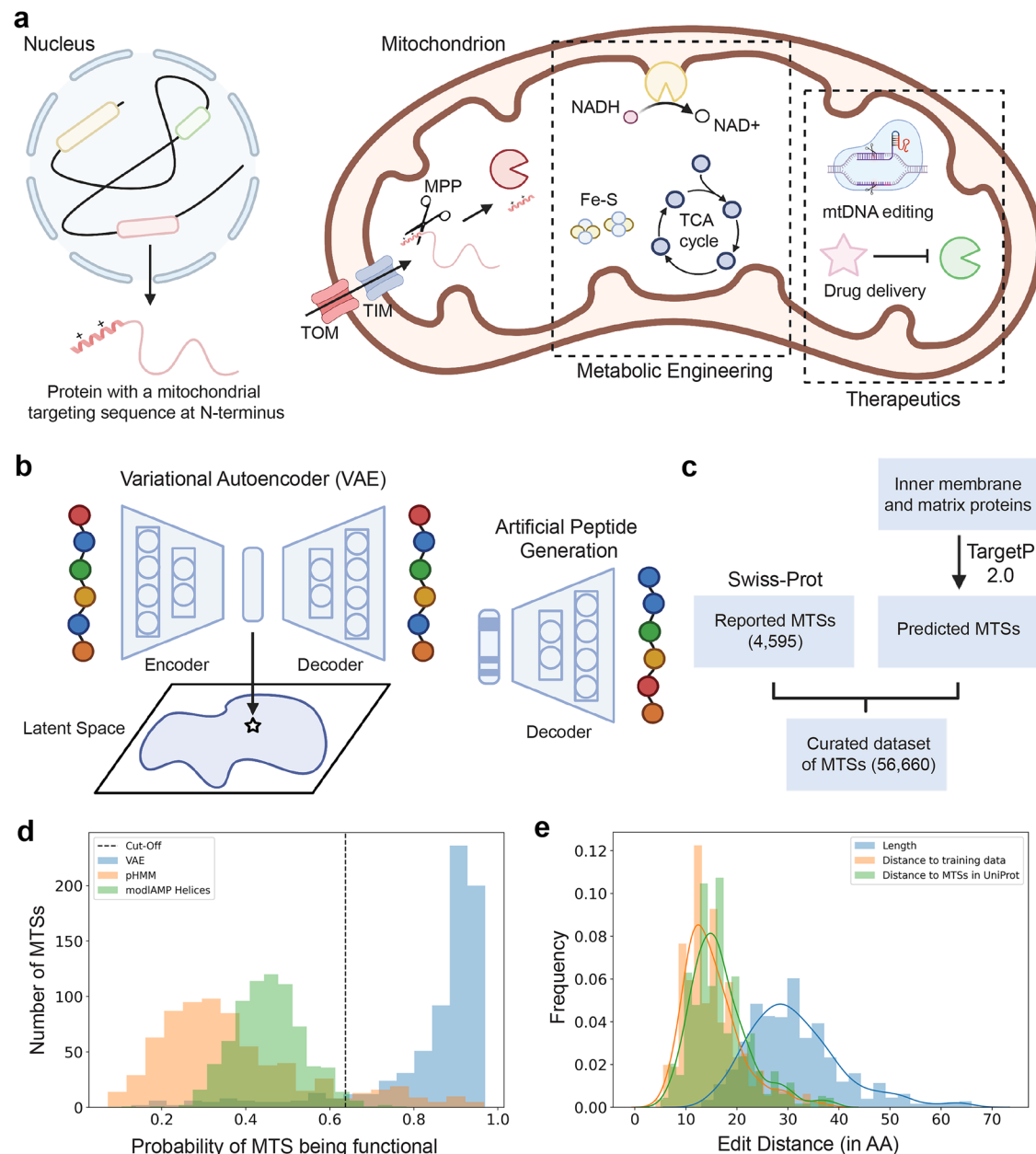
space ( $20^{35}$  for an MTS of common length), the lack of an evident consensus sequence, and the unpredictable relation with the passenger protein. Deep generative models are capable of learning complex, non-linear mappings and representing the underlying data distribution, enabling the design of novel peptides<sup>18</sup> and proteins<sup>22</sup> with desired functions from sequence information alone. In this work, we sought to utilize VAE to generate new-to-nature MTSs (Fig. 1b). VAE consists of two components: the encoder which converts the input sequence into a latent space representation, and the decoder, that reconstructs these latent representations back into the input sequence. The architecture takes the form of a bottleneck, enabling the model to capture essential features. In addition to optimizing the reconstruction loss during training, VAE minimizes the Kullback-Leibler (KL) divergence loss, effectively shaping the distribution of the latent space to fit a multivariate Gaussian distribution. By leveraging the trained model, one can sample latent vectors from the continuous space and input them into the decoder and design novel sequences that closely resemble the properties of the training data.

To facilitate effective training of VAE and encompass a comprehensive design space, we curated a dataset of MTSs. First, we utilized transit peptide annotation and mitochondria as the subcellular location to procure 4,984 MTSs from Swiss-Prot<sup>23</sup>, out of which 4,595 are unique. These sequences are either experimentally validated or predicted using sequence similarity and software such as MitoFates<sup>24</sup>, Predotar<sup>25</sup>, and TargetP<sup>26</sup>. However, this dataset is relatively small and exhibits a notable bias towards model organisms (Supplementary Fig. 1). To address this limitation and enhance dataset diversity, we leveraged TargetP 2.0<sup>3</sup>, a state-of-the-art attention-based deep learning model, to predict MTSs for proteins in Swiss-Prot and TrEMBL with subcellular localization specified as mitochondrial matrix and inner membrane<sup>23</sup>. TargetP 2.0 predicts signal or transit peptides separately for plant and non-plant organisms. Therefore, we used taxonomy classification from UniProt to segregate the resulting protein sequence dataset and fed it to TargetP 2.0 to extract MTSs. Subsequently, we refined the MTSs by filtering for peptides with valid amino acid sequences, restricting lengths to 11–69 amino acids<sup>27</sup>, and eliminating duplicates. Finally, we obtained a dataset comprising 56,660 peptides for model development (Fig. 1c).

To train the VAE, we introduced a '\$' symbol at the C-terminus of the peptides, marking the cleavage site. Subsequently, considering the skewed distribution of peptide length, we performed a stratified 9:1 split of the dataset, ensuring a balanced representation of peptides in both the training and validation sets (Supplementary Fig. 2). We padded all peptides to a uniform size of 70 and employed one-hot encoding to create an input representation for the encoder. Initially, we trained a recurrent neural network (RNN)-based VAE<sup>28</sup> that was previously used to generate antimicrobial peptides<sup>19</sup>. However, post-training, we observed that the same sequence was generated from different randomly initialized latent vectors, with the sequences displaying repeated amino acid residues and minimal mutations, indicating low diversity (Supplementary Data 1). We then implemented the encoder and decoder of the VAE with fully connected layers and did not observe this issue. Therefore, we selected this model for further validation. More details on the model implementation are provided in Materials and Methods.

### In silico analysis reveals generated MTSs are functional, highly diverse, and not found in nature

Given the trained VAE model, we generated artificial peptides and conducted a comprehensive analysis of their functionality, diversity, physicochemical attributes, and structural properties. We sampled 1000 vectors from a normal distribution  $N(\mu = 0, \sigma = 1)$  and fed them to the decoder to generate new-to-nature MTSs. Subsequently, we refined these to obtain 730 peptides with valid amino acid sequences. To verify if the peptides can localize proteins to the mitochondrion, we

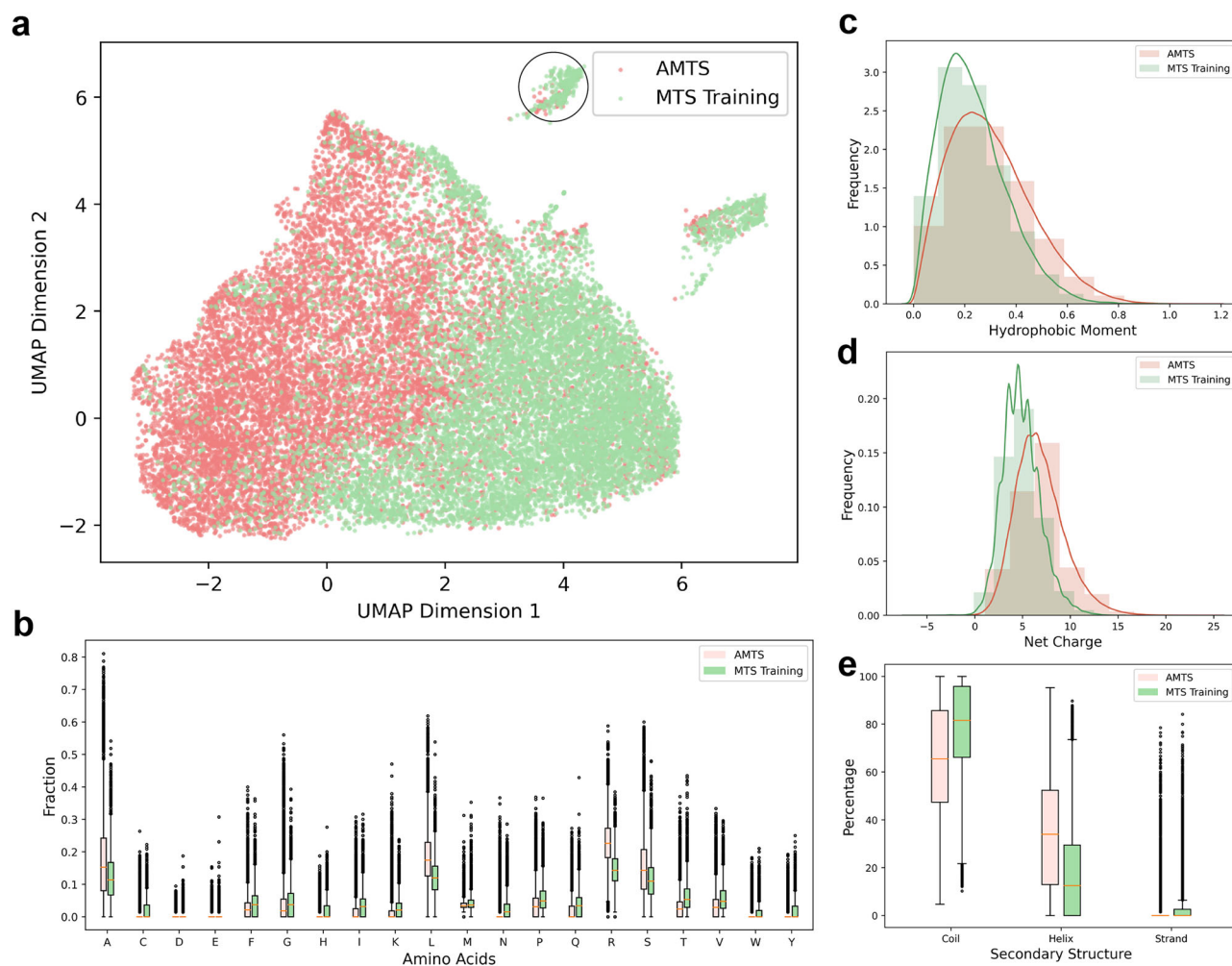


**Fig. 1 | Variational Autoencoder for generation of mitochondrial targeting sequences (MTSS).** **a** Most mitochondrial proteins (99%) are nuclear-encoded and feature an N-terminal sequence for recognition and translocation by the TOM/TIM complex. Upon import into the mitochondrial matrix, the N-terminal sequence undergoes cleavage by MPP and the protein folds. Therefore, utilizing such a targeting sequence enables the delivery of enzymes and drugs for diverse applications, including biochemical production, mtDNA editing, and treating mitochondrial disorders. **b** Scheme for generating artificial MTSSs using Variational Autoencoder (VAE). The model receives a one-hot encoded representation of MTSS as an input. The encoder compresses the input into a latent vector, and the decoder reconstructs the original input. Once the model is trained, one can feed a vector sampled from the latent space to the decoder and generate novel MTSSs. **c** Curated dataset of MTSSs for training the VAE. The dataset includes MTSSs reported in Swiss-Prot and TargetP 2.0-predicted MTSSs for proteins with mitochondrial matrix or

inner membrane as subcellular location. **d** Predicting functionality of generated MTSSs using DeepLoc 2.0. A set of 730 VAE-, modIAMP Helices-, and pHMM-generated MTSSs, appended with GFP, were analyzed for their ability to target mitochondria. Of these, 658 sequences (90.14%) generated by the VAE were deemed functional, compared to 10 sequences (1.37%) and 89 sequences (12.2%) designed using modIAMP's Helices package and pHMM, respectively. **e** VAE-generated MTSSs are diverse in sequence. Many of the generated sequences are 10 to 15 mutations away from the MTSSs in the training data and UniProt. Note that the percentage of AMTS-tagged GFP localized to mitochondrion was calculated based on the probability threshold of 0.6373 specified on the DeepLoc 2.0 server. NAD<sup>+</sup>/NADH: Nicotinamide adenine dinucleotide; mtDNA: mitochondrial DNA; TOM/TIM: translocase of the outer/inner membrane; MPP: mitochondrial processing peptidase; pHMM: profile Hidden Markov Model; AA: amino acids. **a, b** Created in BioRender. Zhao, H. (2025) <https://BioRender.com/gm4kee6>.

prepending them to the Green Fluorescent Protein (GFP) lacking the start methionine and analyzed them using DeepLoc 2.0<sup>4</sup>, a deep learning model for predicting the subcellular localization of proteins (Supplementary Fig. 3). We observed that 90.14% of the peptides were predicted to target mitochondria (Fig. 1d). We benchmarked these sequences against an equal number of MTSSs designed using the profile

Hidden Markov Model (pHMM)<sup>29</sup> and modIAMP's Helices package<sup>30</sup>. A pHMM captures position-specific residue probabilities and insertion and deletion states, making it valuable for generating new sequences that adhere to the language of a protein family. However, when prepended to GFP, only 12.2% of the artificial peptides were predicted to be functional by DeepLoc 2.0. Similarly, only 1.37% of the amphipathic



**Fig. 2 | Characteristics of artificial mitochondrial targeting sequences (AMTSs).**

**a** Uniform Manifold Approximation and Projection (UMAP) visualization of UniRep embeddings depicting the distribution of AMTSs ( $n = 705,081$ ) and MTSs ( $n = 50,980$ ) within the training dataset. VAE effectively generates AMTSs across both high and low-density regions in sequence space. Comparative analysis of physicochemical and structural attributes between generated AMTSs and MTSs in the training data, showcasing. **b** Amino acid composition, **c** Hydrophobic moment,

**d** Net charge, and **e** Secondary structure. Overall, AMTSs exhibit a net positive charge and form amphiphilic  $\alpha$ -helix structures, features crucial for directing protein to mitochondria. 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.

alpha-helical peptides generated using modIMP's Helices package were predicted to target GFP to mitochondria. We conducted a bioinformatic analysis to determine whether the consensus motif of the TOM20-recognition element in mitochondrial presequences,  $\varphi\chi\beta\varphi\varphi$ <sup>24,31,32</sup> (where  $\varphi$  represents a hydrophobic residue {L, F, I, V, W, Y, M, C, A},  $\chi$  represents any amino acid, and  $\beta$  represents a basic residue {R, K, H}), is present in the VAE-generated MTSs. A significant proportion (514 out of 730, or 70.4%) of peptides contained this motif (Supplementary Data 1), indicating a likely TOM20-dependent import mechanism. Next, we compared the VAE-generated MTSs with the sequences in the training data and the MTSs reported in UniProt, as the latter are accessible to researchers. We calculated the Levenshtein distance to determine the number of mutations between the generated sequence and its closest match. Our analysis revealed that VAE-generated MTSs are highly diverse, averaging 10–15 amino acids away from any naturally occurring MTSs (Fig. 1e).

Next, we evaluated the model's capability to cover the natural MTS space. We generated 1,000,000 artificial MTSs and selected peptides with valid amino sequences. Subsequently, we applied CD-HIT<sup>33</sup> on the training data and VAE-generated MTSs and obtained natural and artificial MTSs with less than 30% sequence identity,

respectively. Next, we embedded them with UniRep model<sup>34</sup> and employed Uniform Manifold Approximation and Projection (UMAP) to visualize the latent space in two dimensions. This visualization revealed three distinct clusters, with artificial MTSs effectively covering the entire natural sequence space (Fig. 2a). Upon closer examination, we noticed one cluster (circled) contained MTSs without the start methionine, a discrepancy propagated through MTSs of proteins from UniProt used for model training. While we inspected the peptides in the remaining two clusters, we failed to detect any apparent differences.

Furthermore, we compared amino acid composition and physicochemical characteristics of artificial and naturally occurring MTSs, as shown in Fig. 2b–d and Supplementary Fig. 4. We used BioPython to compute these features. Overall, VAE-generated MTSs followed a similar distribution for amino acid composition compared to the MTSs in the training dataset. Artificial MTSs were enriched in Alanine (A), Arginine (R), Leucine (L), and Serine (S). Negatively charged residues, Aspartic (D) and Glutamic acid (E), were depleted for sequences in both datasets. Therefore, artificial MTSs carried a net positive charge, an important feature for functional MTSs. Next, we calculated the mean hydrophobic moment using modIMP<sup>30</sup> to measure the



amphiphilicity of the peptides<sup>35</sup> and observed a slight shift towards higher values. Hydrophobicity, as measured using the Eisenberg scale, exhibited a more negative trend, while when quantified through GRAVY (Grand average of hydropathicity index), followed a similar distribution compared to the training dataset (Supplementary Fig. 4a, b). Generated MTSs also spanned the length of the peptides in the training dataset (Supplementary Fig. 4c). Next, we calculated the fraction of  $\alpha$ -helix,  $\beta$ -sheet, and loop (or coil) using S4PRED<sup>36</sup>, a state-of-the-art model for single sequence-based secondary structure prediction. We observed artificial MTSs displayed a higher tendency to form  $\alpha$ -helix at the expense of coil (Fig. 2e). The results demonstrate that the VAE effectively captured meaningful features essential for mitochondrial localization.

### Sampling algorithm to select sequences for in vivo experimental validation

Based on the confidence gained from the in silico analysis, we randomly selected nine VAE-generated MTSs with TargetP 2.0 scores above 0.9 for in vivo evaluation in *S. cerevisiae* (Supplementary Data 2). Using Gibson assembly<sup>37</sup>, we fused these sequences to the N-terminus of GFP. Additionally, we constructed a positive control using the MTS of COX4<sup>38</sup>, a well-characterized endogenous peptide commonly used for mitochondrial targeting. We transformed these GFP plasmids into *S. cerevisiae* and verified the mitochondrial targeting capability of the sequences through confocal microscopy. We initially examined the negative control (GFP) and the positive control (COX4-GFP). The overlap of GFP fluorescence with the mitochondrial stain, MitoTracker™ Orange CMTMRos, confirmed that COX4 can localize the protein to mitochondria, while GFP on its own cannot (Supplementary Fig. 5). Similarly, we conducted the characterization for nine artificial MTSs (AMTSs). Out of these sequences, AMTS 131, 205, 225, and 335 demonstrated selective targeting to mitochondria, while AMTS 6, 64, and 96 displayed targeting to multiple subcellular locations, including mitochondria. AMTS 110 and 245 did not target mitochondria.

Next, we sought to demonstrate the functionality of these VAE-generated sequences in three eukaryotic organisms: *Homo sapiens*, *N. benthamiana*, and *R. toruloides*. While some of the artificial peptides demonstrated successful targeting, we wondered whether the interaction with the import machinery of an organism configured a bias in the amino acid sequence of MTSs. Our hypothesis also stemmed from the consideration that tools for predicting subcellular localization incorporated taxonomy in their predictions<sup>3,24</sup>. Therefore, we analyzed various attributes of naturally occurring MTSs in these organisms (Supplementary Fig. 6) and formulated a sampling scheme to choose sequences better suited for validation in the organism of interest (Fig. 3a). We acquired these sequences from UniProt or predicted them using TargetP 2.0. Subsequently, we refined them and calculated amino acid composition, physicochemical attributes, and secondary structure as previously described. MTSs from different organisms preferred varying amino acid residues. For instance, MTSs from the human proteome exhibited enrichment in Glycine (G) and Alanine (A), accompanied by a reduction in Serine (S) when compared to MTSs from other three proteomes. Similarly, these MTSs differed in hydrophobic moment and had a higher tendency to form an  $\alpha$ -helix compared to fungal MTSs. Next, we used the pre-trained UniRep model to obtain embeddings for the targeting sequences, capturing amino-acid, physicochemical, and structural attributes through meaningful protein representation<sup>34</sup>, and visualized them in a two-dimensional space using UMAP. Based on the density plot, it was evident that the MTSs for different organisms were clustered to some extent (Fig. 3a).

To determine the organism for characterizing an MTS, we devised a sampling scheme, implementing two criteria using the 1900-dimensional UniRep representation of MTSs. The first criterion categorizes MTSs based on their proximity to the cluster center, calculated

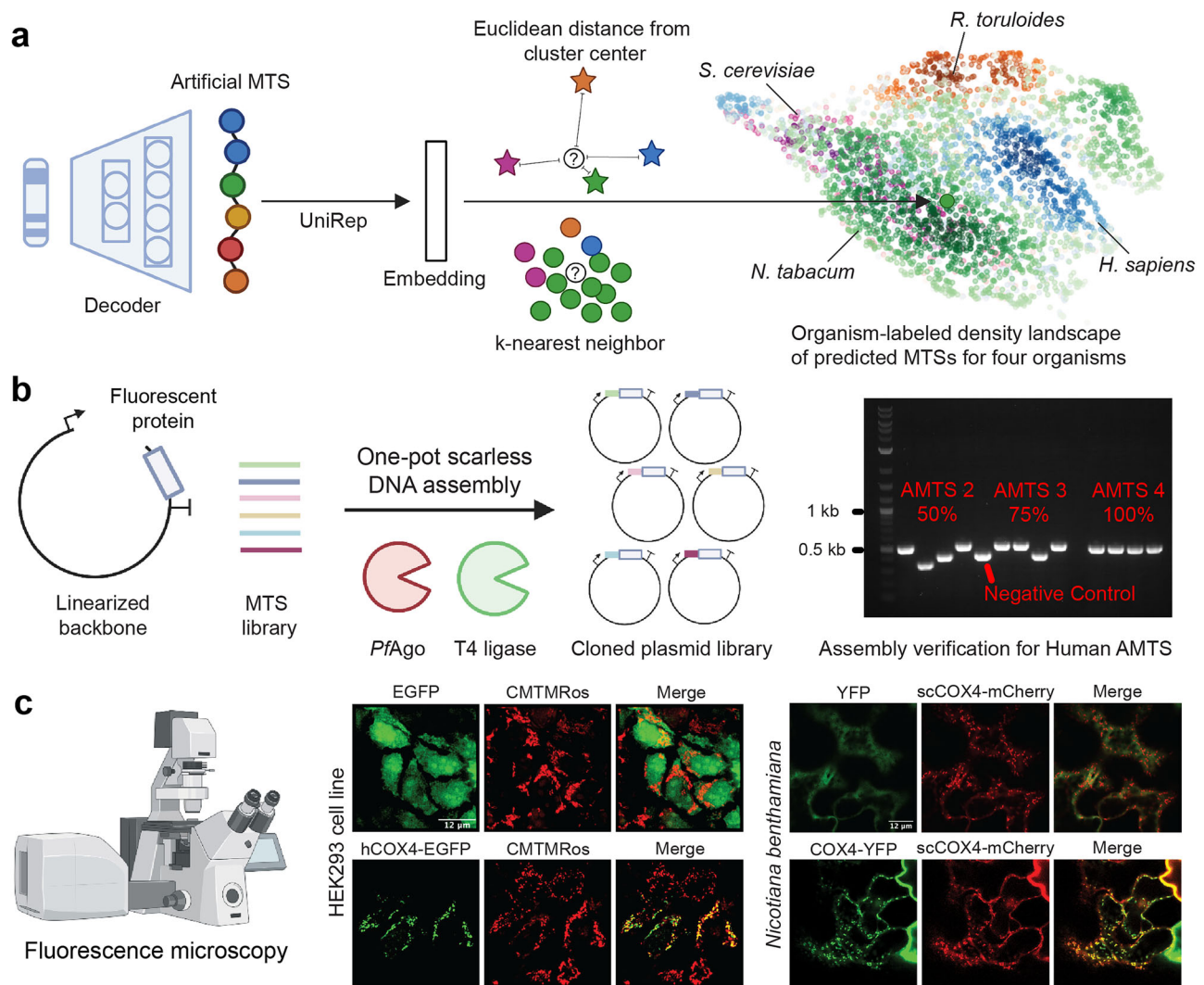
as the average UniRep representation of MTSs associated with a specific organism. The second criterion follows the principle of k-Nearest Neighbors, assigning an MTS to an organism by computing the frequency of organism labels for its 20 closest naturally occurring counterparts. We choose an MTS for validation only when both criteria concur. To highlight the effectiveness of this sampling scheme over random selection, we analyzed the cleavage patterns of peptides in the human proteome (Supplementary Fig. 7). The consensus for the eight residues at the C-terminus revealed that the sampling results in a pattern more closely resembling naturally occurring MTSs, as anticipated. Moreover, Glycine (G), an amino acid residue favored by MTSs in the human proteome, is identified as a positive attribute only for sampled MTSs and not for the ones generated randomly, demonstrating a method for capturing organism-specific bias. Therefore, utilizing this approach, we labeled the VAE-generated MTSs and selected eight peptides based on the distance from the cluster center for characterization in four eukaryotic organisms (Supplementary Data 2). Additionally, we included a list of all AMTSs sampled for these organisms in this table to allow researchers to rapidly profile these novel MTSs, including distant sequences (those with more than 10 mutations from peptides in the training dataset) that lack a TOM20-recognition element.

### VAE-generated peptides target mitochondria in vivo across various eukaryotic organisms

To test the 32 VAE-generated sequences obtained after sampling, we fused these artificial peptides to the N-terminus of various reporter genes. Instead of using Gibson assembly, we adopted *PfAgo*-based assembly<sup>39</sup>, a more versatile cloning approach. *PfAgo* is an artificial restriction enzyme that uses single-stranded DNA guides to cleave double-stranded DNA, making sticky ends of the desired length. Unlike the previous approach, we modified the script to design guides at the end of the amplified plasmid backbone to create a 12-bp overhang. We then cloned artificial MTSs and appropriate positive controls using T4 DNA ligase. This modified approach aided in constructing plasmids with high fidelity and ease (Fig. 3b). Subsequently, we performed confocal microscopy and confirmed the negative (EGFP/YFP) and positive controls for mitochondrial localization (Fig. 3c). We provide a summary of artificial MTS validation in the four eukaryotic organisms in Fig. 4a. All eight peptides localized GFP to the mitochondria of HEK293 cells, as demonstrated by overlap with MitoTracker™ Orange CMTMRos (Fig. 4b). We also verified successful peptide cleavage using Western blot (Supplementary Fig. 8). Similarly, six out of eight peptides successfully targeted Yellow Fluorescent Protein (YFP) to *N. benthamiana*'s mitochondria (Supplementary Fig. 9). We confirmed this using agrobacterium-mediated transient expression of fluorescent proteins and subsequent colocalization with ScCOX4-mCherry.

Next, we verified the VAE-generated MTSs in *R. toruloides*, a non-model oleaginous yeast. It is considered a promising chassis to produce chemicals and biofuels, owing to its ability to grow on diverse substrates and its high endogenous flux towards lipids and carotenoids<sup>40</sup>. Therefore, to characterize artificial MTSs in this host, we knocked out *crtYB*, the enzyme responsible for producing  $\beta$ -carotene, using a previously constructed sgRNA cassette<sup>41</sup> to avoid interference with fluorescence analysis. Next, we transformed the linear AMTS-tagged GFP fragment with the necessary selection marker in the strain and analyzed them using confocal microscopy. Six out of eight peptides successfully localized GFP to the mitochondria, confirmed using Rhodamine B, hexyl ester (Supplementary Fig. 10). We used TargetP 2.0-predicted MTS of RtCOX4 as a control. However, we did not observe GFP localized to mitochondria. Finally, we characterized eight more artificial MTSs in *S. cerevisiae*. Among these, four constructs showed successful mitochondrial targeting (Supplementary Fig. 11).

Overall, we achieved a success rate of 75–100% for AMTSs characterized in the HEK293 cells, *N. benthamiana*, and *R. toruloides*.



**Fig. 3 | Organism labeling and workflow for in vivo evaluation of generated artificial mitochondrial targeting sequences (AMTSs).** **a** Selection of AMTSs for experimental characterization. AMTSs are generated through the VAE, encoded via the pre-trained UniRep model, and annotated with an organism label based on proximity to the cluster center, defined as the mean UniRep embedding of MTSS in a proteome, and using k-Nearest Neighbors. **b** Scarless DNA assembly of AMTSs utilizing *PfAgo*/AREs. Plasmid backbones are cleaved using *PfAgo* and phosphorylated guides, creating sticky ends. Annealed DNA sequences, encoding for AMTSs, are then ligated upstream of fluorescent reporter proteins using T4 DNA ligase. Final constructs are verified through Colony PCR, demonstrating high assembly

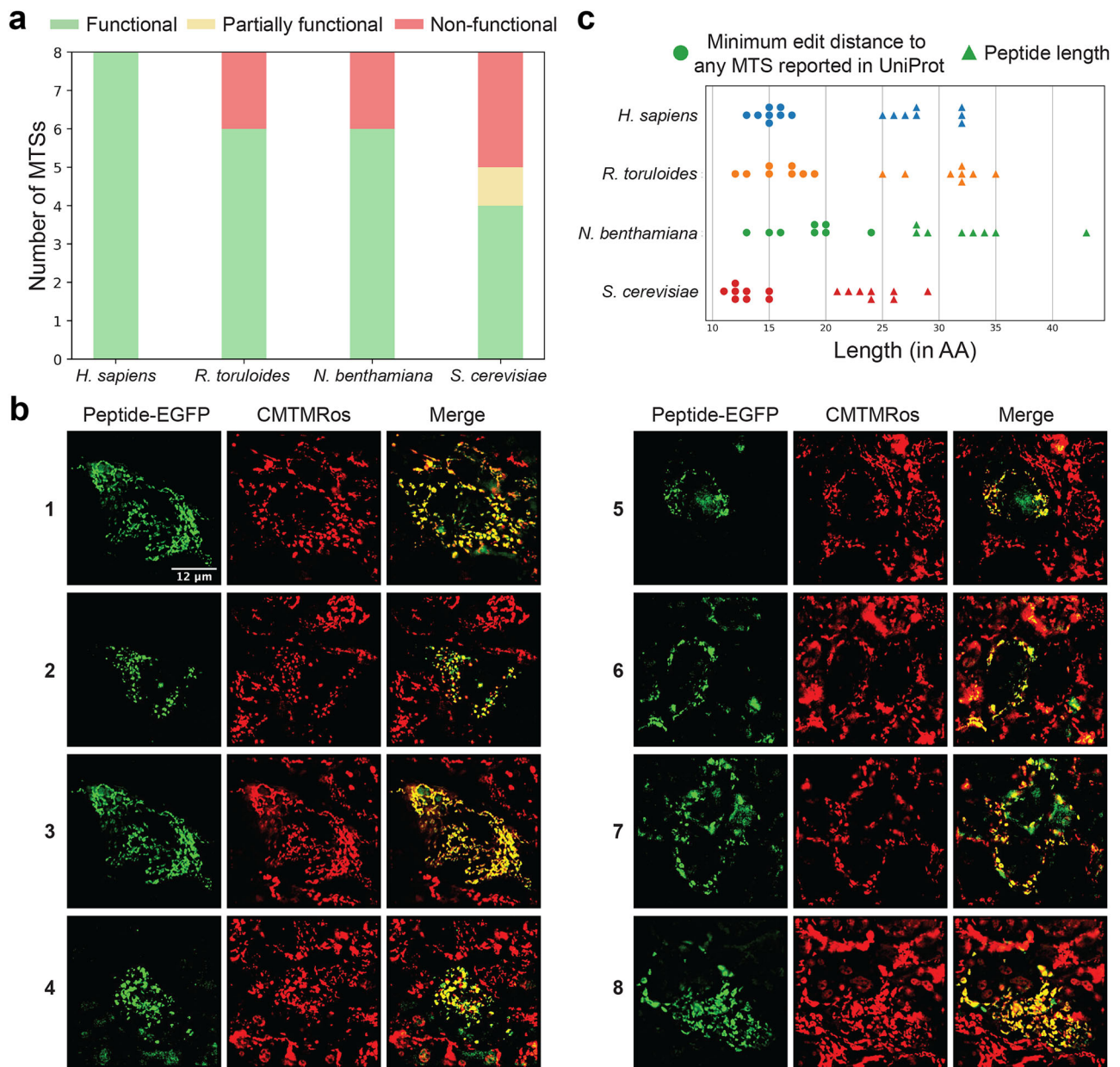
fidelity. **c** Functional validation of positive control MTSS. Plasmids harboring MTS-fluorescent protein constructs are transformed into the host organism, followed by verification of mitochondrial localization using fluorescence microscopy. Mitochondria in HEK293 cell lines are visualized with MitoTracker™ Orange CMTMRos, whereas mitochondrial targeting in *N. benthamiana* is confirmed through protein colocalization with scCOX4-mCherry[2]. Scale bar: 12 μm. **b, c** Each experiment was repeated three times independently with similar results. Source data are provided as a Source Data file. **a, b** Created in BioRender. Zhao, H. (2025) <https://BioRender.com/wszxpbe>.

However, the hit rate dropped to 50% for AMTSs tested in *S. cerevisiae*. This decrease can be partially attributed to the dispersion of naturally occurring MTSS in *S. cerevisiae* across the entire UMAP space, resulting in inefficient sampling. Notably, the characterized peptides maintained diversity even after sampling closer to the cluster center, with each peptide being 15–20 mutations away from any MTSS reported in UniProt (Fig. 4c).

### Latent space interpolation enables to design peptides capable of targeting multiple subcellular locations for dual organelle engineering

Our earlier analysis indicated that 27 and 17 of the VAE-generated peptides are predicted to localize GFP to the chloroplast (plastid) or both the chloroplast (plastid) and mitochondrion (Supplementary Fig. 3), respectively. Initially, we suspected the model learned these attributes from peptides of proteins with mislabeled

subcellular localization annotations and potential false positives from TargetP 2.0 predictions. However, MTS and chloroplast targeting sequences (CTSs) are evolutionarily related and possess similarities<sup>42</sup>. In nature, 5% of the proteins in the endosymbiotic organelles of plant cells are expected to be dual localized<sup>43,44</sup>. Dual localization can occur either through a protein carrying different signals or a single ambiguous signal for both organelles. Previous studies report that the characteristics of these ambiguous signals are intermediate to those of MTSS and CTSs<sup>45</sup>. Therefore, we asked if interpolation in learned latent space can generate peptides capable of targeting both organelles and provide insights into their evolutionary trajectories. We trained a variational autoencoder (Dual-VAE) on MTSS and CTSs from the Viridiplantae kingdom reported in UniProt or predicted using TargetP 2.0 (Supplementary Fig. 12). We observed two posterior distributions in the latent as the model learns from sequences belonging to two different classes



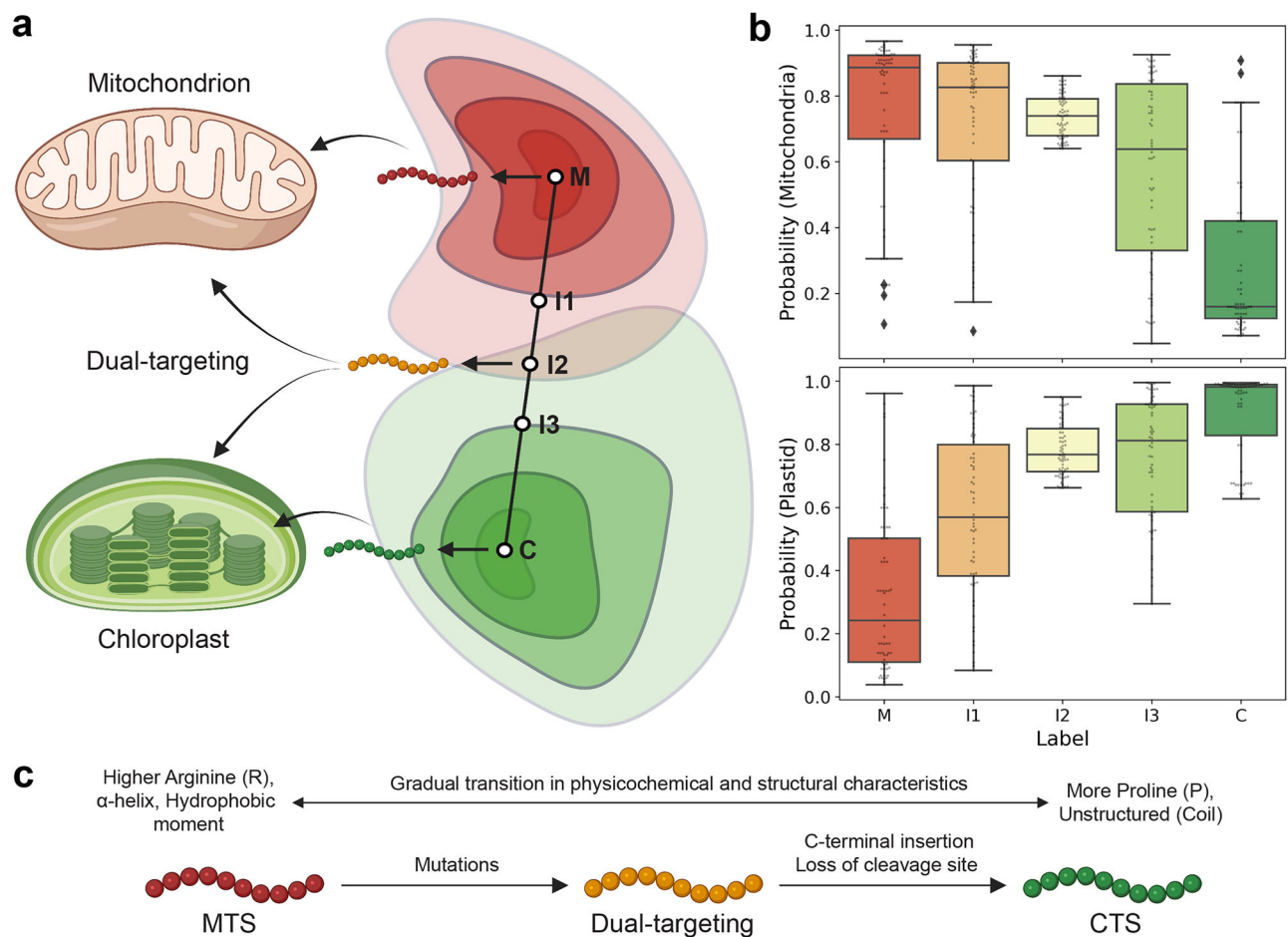
**Fig. 4 | Characterization of artificial mitochondrial targeting sequences (AMTSs) in four eukaryotic organisms.** **a** Summary of AMTS functionality in HEK293 cell line, *R. toruloides*, *N. benthamiana*, and *S. cerevisiae*. **b** Fluorescence microscopy analysis of AMTS-EGFP constructs in the HEK293 cell line. All eight sequences target mitochondria in vivo, confirmed through overlap with

MitoTracker™ Orange CMTMRos. Each experiment was repeated three times independently with similar results. **c** Levenshtein distance of artificial peptides to MTSS reported in UniProt, illustrating the diversity of characterized sequences. Scale bar: 12  $\mu$ m. Source data are provided as a Source Data file.

(Supplementary Fig. 13a). Next, we utilized K-Means to locate the center of each distribution to select latent vectors for interpolation. Employing a Euclidean distance threshold of 0.4 from the cluster center, we obtained over 50 MTSS and CTSS. For each sequence pair, we performed linear interpolation and generated three artificial sequences by feeding the approximated latent vectors to the decoder (Fig. 5a). First, we analyzed these sequences for functionality using DeepLoc 2.0<sup>4</sup> to predict if the sequences generated at the interface of two distributions can target both endosymbiotic organelles (Supplementary Fig. 13b). We observed a smooth transition between the ability of peptides to target mitochondrion and chloroplast as we traversed along the interpolation path. Furthermore, in silico analysis revealed 62 peptides exhibited a high likelihood of functioning as dual-targeting sequences (Fig. 5b; Supplementary Data 3).

Subsequently, we analyzed the interpolation path for these 62 putative dual-targeting peptides to understand the changes in physicochemical and structural characteristics (Fig. 6). Examination of the amino acid composition showed a significant increase in Serine (S) and Leucine (L) and a decrease in Arginine (R) and Alanine (A) when compared with MTSS. An opposite trend was observed for arginine in comparison with CTSS, in agreement with a previous study that investigated the role of arginine in specificity for targeting the chloroplast<sup>46</sup>. CTSS were also enriched in Proline (P) when matched with MTSS and dual targeting sequences. Comparison of peptide lengths revealed a gradual increase from MTSS to interpolated sequences, while CTSS exhibited notably greater length, suggesting an insertion. Multiple sequence alignment of interpolated sequences revealed amino acid substitutions at the beginning, followed by the insertion of amino acids along the length of the peptide





**Fig. 5 | Designing dual-targeting peptides using Dual-VAE. a** Schematic of interpolation in the latent space to generate peptides capable of targeting both mitochondria and chloroplasts. **b** DeepLoc 2.0 predictions for the 62 dual-targeting peptides. The likelihood of mitochondrial targeting diminishes while chloroplast targeting increases along the interpolation path from mitochondrial targeting sequences (MTS) to chloroplast targeting sequences (CTS), leading to the emergence of dual-targeting characteristics at the midpoint. 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. **c** Proposed evolution of dual-targeting peptides. If dual-targeting sequences evolve from MTS, they gather mutations that alter the composition of specific amino acid residues, physicochemical properties, and secondary structural features. In contrast, evolution from CTS not only requires similar mutations but also the insertion of an unstructured element at the C-terminus and modifications that incorporate a cleavage motif for recognition by mitochondrial processing peptidase. **a, c** Created in BioRender. Zhao, H. (2025) <https://BioRender.com/rjlok49>.

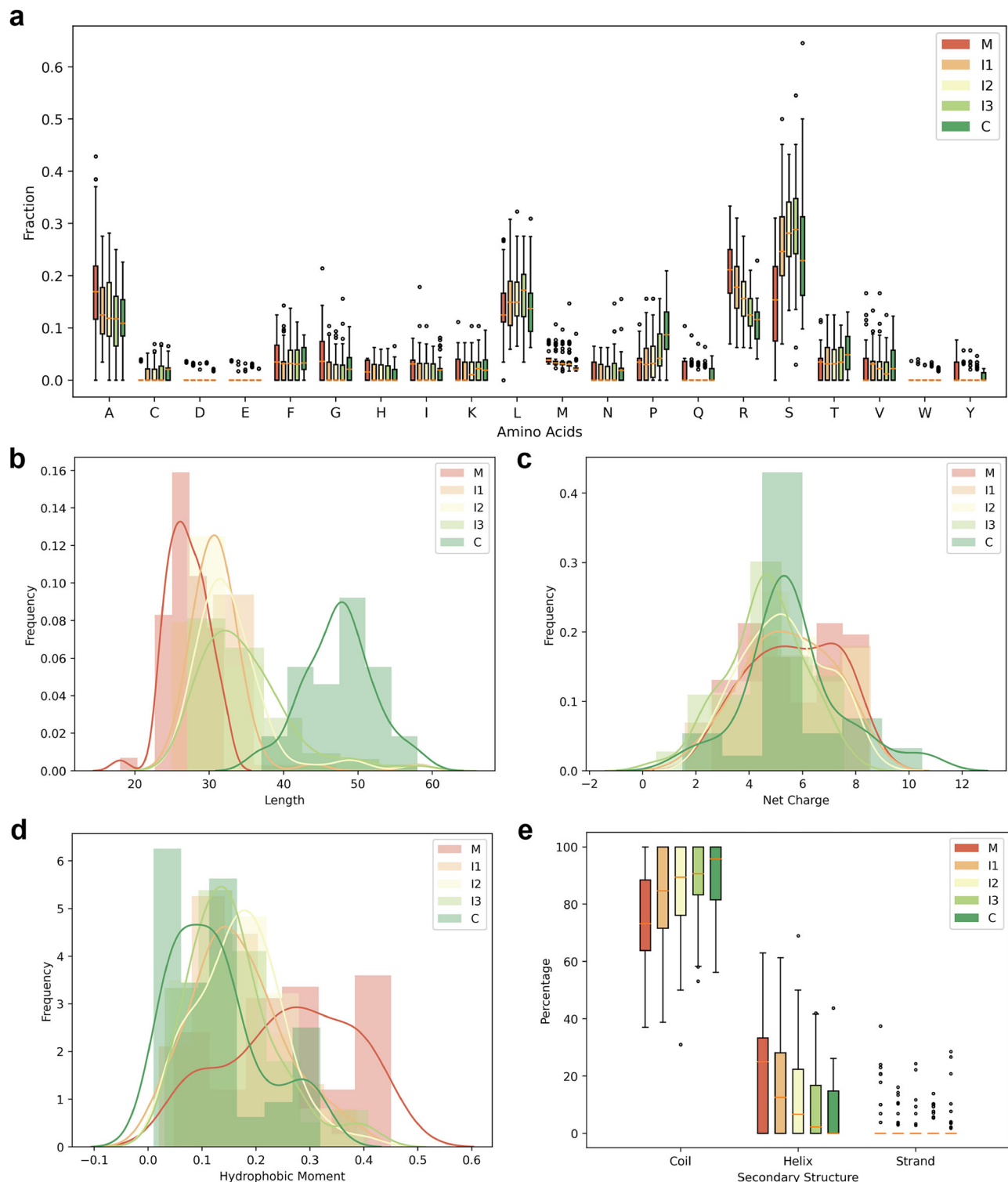
(Supplementary Fig. 14). Furthermore, we saw a decrease in global hydrophobic moment as the transition progressed from MTSs to CTSs, while net charge exhibited a similar distribution. Next, we employed S4PRED to predict the secondary structure of the peptides, indicating an increase in unstructured elements (coil) for dual-targeting peptides and CTSs at the expense of  $\alpha$ -helix. These findings are consistent with a previous study that examined the evolution of MTSs and CTSs from antimicrobial peptides<sup>47</sup>. We also analyzed the C-terminus of these peptides and noted the conservation of the RR, R-2, or R-3 motifs in both MTSs and dual-targeting peptides (Supplementary Fig. 15), suggesting the potential for successful cleavage upon import into mitochondria. In summary, a smooth transition was observed in the physicochemical and structural characteristics while transitioning from MTSs to dual-targeting peptides. Based on these observations (Fig. 5c), we hypothesize that the dual-targeting sequences are more likely to have evolved from mitochondrial targeting sequences.

### Subcellular localization of enzymes enhances 3-hydroxypropionic acid production

As a proof of concept, we demonstrate the utility of characterized MTSs for the metabolic engineering of 3-hydroxypropionic acid

(3-HP). While nature has developed various biosynthetic routes to produce 3-HP, we selected the  $\beta$ -alanine pathway owing to its higher theoretical yield<sup>48</sup>. This pathway involves three genes to convert the endogenous precursor, L-aspartate, to 3-HP: aspartate decarboxylase (PAND),  $\beta$ -alanine-pyruvate aminotransferase (BAPAT), and 3-hydroxypropionate dehydrogenase (YDFG). Previously, the pathway was established in the cytoplasm of *S. cerevisiae*<sup>48</sup>. However, given the abundant supply of the starting precursor of this pathway in mitochondria<sup>9</sup>, we opted to localize the biosynthetic pathway for 3-HP in this organelle, aiming for improved production (Fig. 7a). We constructed two plasmids harboring genes without and with N-terminal MTSs (COX4-PAND, AMTS131-BAPAT, and AMTS3-YDFG), transformed them into the CEN.PK strain, cultured the strains in 2 mL of SC-URA (50 g/L glucose) media, and quantified 3-HP using HPLC. Compartmentalizing the pathway in the mitochondria led to higher 3-HP production (Fig. 7b), showing an increase of 62.3% compared to the cytoplasmic pathway (i.e., from 1.70 g/L to 2.76 g/L). We conducted additional experiments to confirm the mitochondrial localization of AMTS-fusion proteins (AMTS131-BAPAT and AMTS3-YDFG) and demonstrate that all three enzymes must be localized to mitochondria to achieve maximal 3-HP production (Supplementary Fig. 16).



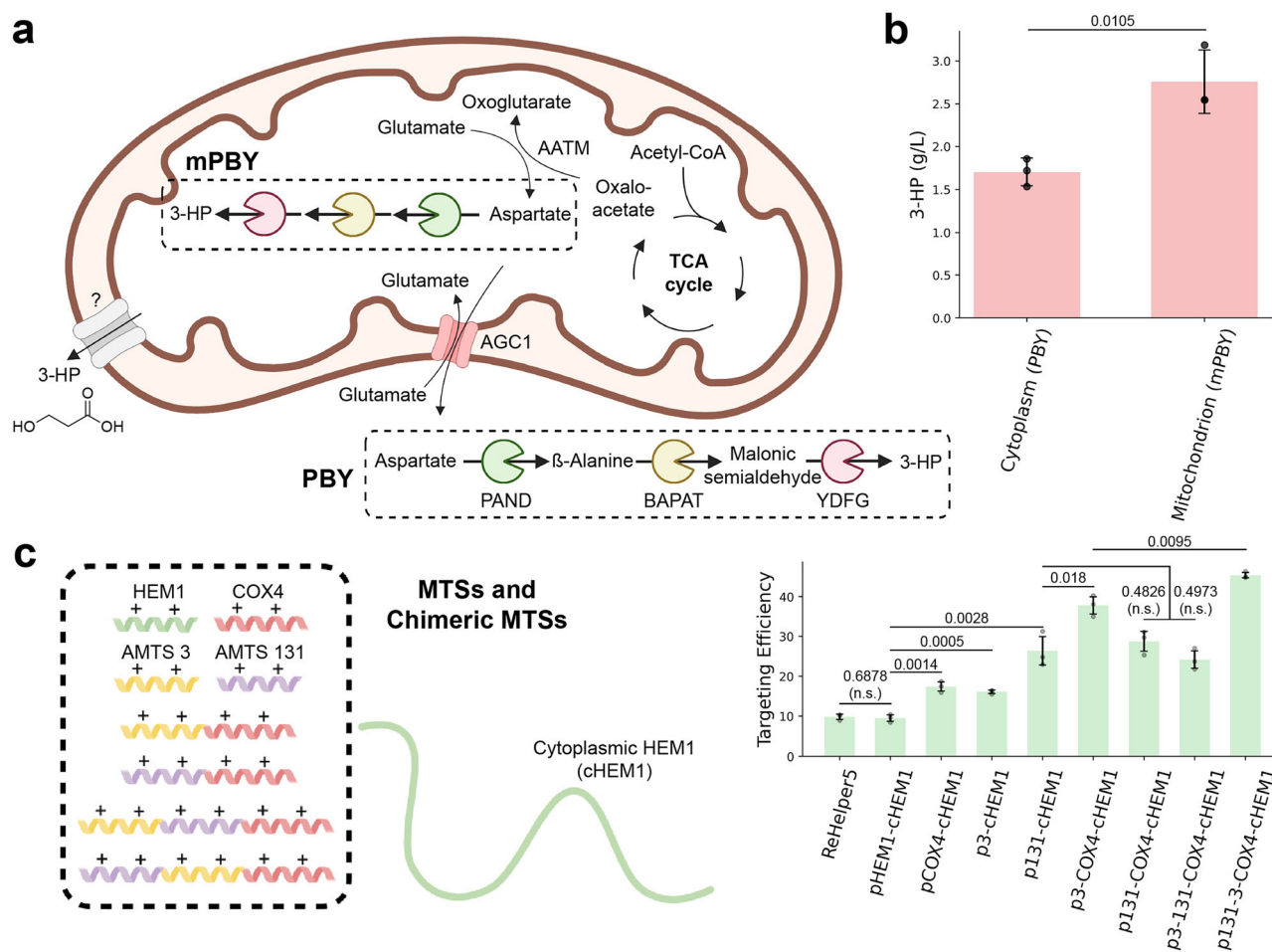


**Fig. 6 | Physicochemical and structural features of artificial sequences sampled along the interpolation path of dual targeting peptides.** This includes their (a) Amino acid composition, (b) Length, (c) Net charge, (d) Hydrophobic moment, and (e) Secondary structure. A smooth transition is observed for various attributes, such as amino acid composition (arginine, proline), hydrophobic moment, and

secondary structure elements when transitioning from mitochondrial to chloroplast targeting sequences. 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.

Moreover, we observed 3-HP in the supernatant, indicating the presence of a putative 3-HP transporter on the mitochondrial membrane as reported elsewhere<sup>49</sup>, whose overexpression might further improve 3-HP titers.

**Peptide chimeras improve mitochondrial targeting efficiency**  
In our second application, we showcase the application of artificial MTSSs for delivering cargo proteins to mitochondria. The targeting efficiency of MTSSs is dependent on the passenger protein<sup>14</sup>. Therefore,



**Fig. 7 | Application of characterized MTSs for metabolic engineering and protein delivery.** **a** Schematic of strains harboring plasmids for cytoplasmic (PBYP) and mitochondrial (mPBYP) 3-hydroxypropionic acid (3-HP) biosynthetic pathways.

**b** Production of 3-HP from SC-URA (50 g/L glucose). **c** Enhancing mitochondrial targeting efficiency using chimeric MTSs. A library of MTSs and chimeric MTSs was cloned upstream of the *HEM1* gene without the native MTS and transformed into *S. cerevisiae*. Targeting efficiency was assessed by quantifying the ratio of 5-ALA produced to mRNA levels of the *HEM1* gene. PAND: aspartate decarboxylase;

BAPAT:  $\beta$ -alanine-pyruvate aminotransferase; YDFG: 3-hydroxypropanoate dehydrogenase; HEM1: 5-aminolevulinic synthase; AATM: aspartate aminotransferase, mitochondrial; AGC1: mitochondrial aspartate-glutamate transporter. Data represents mean  $\pm$  s.d.  $n = 3$  biological replicates.  $P$  value was calculated by two-tailed unpaired  $t$ -test. n.s. = not significant ( $p > 0.05$ ). **a, c** Created in BioRender. Zhao, H. (2025) <https://BioRender.com/9sk5fmz>. Source data are provided as a Source Data file.

finding an optimal MTS requires screening a diverse library. An alternate strategy for improved targeting involves employing chimeric MTSs, i.e., a combination of MTSs in series<sup>50,51</sup>. To test this hypothesis, we utilized endogenous 5-aminolevulinic synthase (HEM1) in *S. cerevisiae* as the model system. HEM1 catalyzes the synthesis of 5-aminolevulinic acid (ALA) from succinyl-CoA and glycine within the mitochondria, the rate-limiting step in heme biosynthesis<sup>52</sup>. Mutations in *ALAS1*, a homolog of HEM1, are associated with X-linked protoporphyria and X-linked sideroblastic anemia in humans<sup>53</sup>. Therefore, the efficient delivery of HEM1 to mitochondria is beneficial for metabolic engineering and the development of large molecule therapeutics. However, as an MTS is positioned at the N-terminus of the protein, it can also influence transcription rates and, consequently, mRNA levels. Therefore, we defined targeting efficiency as the ratio of 5-ALA produced to the mRNA levels of HEM1, a definition more relevant to targeted mRNA delivery to mitochondria. We constructed a library of chimeric MTS using COX4, AMTS3, and AMTS131. For higher-order combinations, we maintained COX4 at the C-terminus to ensure cleavage and no interference with downstream protein folding.

We introduced these MTS/chimeric MTS-cHEM1 constructs into *S. cerevisiae* and subsequently measured both 5-ALA production and

mRNA levels of the HEM1 gene to evaluate targeting efficiency (Fig. 7c and Supplementary Fig. 17). Initially, we tested our plasmid-based assay against the endogenous HEM1 gene encoded in the genome. We utilized two plasmids for this test: ReHelper5, an episomal vector harboring the TEF1p-GFP-TEF1t cassette, and pHEM1-cHEM1, the same vector harboring the TEF1p-HEM1-TEF1t cassette. We observed that expressing more copies of HEM1 increases 5-ALA production; however, the targeting efficiency through both plasmid and genome-based expression remains the same, demonstrating that we can use our plasmid system to evaluate individual or chimeric MTSs for improved protein delivery. For the individual targeting sequences, COX4 and AMTSs exhibited a significant increase in targeting efficiency compared to the native MTS. Furthermore, we observed that the targeting efficiency increased with the number of MTSs in chimeric constructs. Specifically, a 4.76-fold and a 2.53-fold enhancement was achieved with three MTSs in series, emphasizing the importance of the number and order of MTSs for mitochondrial targeting. However, 5-ALA production in chimeric constructs did not show significant improvement over the individual MTSs, primarily due to lower mRNA levels, which may be attributed to suboptimal codon optimization.

## Discussion

Mitochondria play a key role in energy metabolism, biosynthesis of macromolecular precursors and reducing equivalents, and effective management of metabolic waste. Therefore, mitochondrial targeting holds immense promise for metabolic engineering and therapeutics. However, despite the crucial role of mitochondria in cellular metabolism, the repertoire of well-characterized mitochondrial targeting sequences (MTSs) remains limited. This scarcity restricts the options available for directing proteins to mitochondria, potentially leading to suboptimal targeting efficiency and compromised functionality of the delivered cargo or pathways. These limitations underscore the urgent need for the design and characterization of diverse, functional MTSs.

In this study, we addressed this challenge by developing a deep generative AI framework to design artificial mitochondrial targeting sequences. First, we curated a large dataset of MTSs occurring in nature and trained a Variational Autoencoder on these peptide sequences. We validated the functionality of the generated MTSs using DeepLoc 2.0. A detailed analysis of physicochemical and structural attributes revealed the sequences were positively charged, amphiphilic, and tended to form an  $\alpha$ -helix, features important for targeting mitochondrion. Next, we devised a sampling scheme to prioritize MTSs for experimental validation. Using confocal microscopy, we characterized 41 new-to-nature peptides in vivo in four eukaryotic species, achieving a success rate of 50–100%. Moreover, we demonstrated that the peptides are successfully cleaved in vivo in the HEK293 cell line.

Furthermore, our analysis showcased that some of the generated sequences are likely capable of targeting both the mitochondrion and chloroplast. Therefore, we trained another model, Dual-VAE, on MTSs and CTSs from the Viridiplantae kingdom. Subsequently, we utilized linear interpolation in the latent space and generated 62 putative dual-targeting sequences. We analyzed the variations in the features of targeting sequences along the interpolation trajectory and provided insights into how dual-targeting sequences may have likely evolved from mitochondrial targeting peptides. In the future, we anticipate that the characterization of these peptides will enhance our understanding of dual-targeting in plants and boost titers of biochemicals, such as taxenes<sup>54</sup>, by utilizing acetyl-CoA from both compartments. Lastly, we demonstrated the application of the characterized peptides for metabolic engineering and protein delivery. We localized the enzymes in the  $\beta$ -alanine pathway to produce 3-HP in the mitochondria and observed a 1.62-fold improvement in titers compared to the cytoplasmic counterpart. In the second application, we focused on improving the targeting of HEMI1 to mitochondria using chimeric MTSs and noticed a 4.76-fold enhancement in targeting efficiency. We saw the order and number of MTSs in a series combination are both important to enhance localization.

A promising avenue for further improvement lies within the model architecture. One could directly incorporate the organism label and protein information using a Conditional Variational Autoencoder<sup>55</sup> or frame the problem as a machine translation task<sup>20</sup>, removing the need for selective sampling. In addition, it would capture the bias introduced through interaction with the import machinery and the intrinsic relation between the passenger protein and MTS. Apart from incorporating metadata, one could utilize advanced feature engineering strategies, such as physicochemical properties<sup>56</sup>, pre-trained large language models (LLMs)<sup>57</sup>, or structural representations<sup>58</sup> to adequately capture interactions between amino acids at varying distances. One could also streamline the selection of artificial MTSs for in vivo characterization by inputting vectors positioned close to the cluster center (derived from VAE latent vectors of MTSs in a proteome) into the decoder, facilitating the generation of artificial MTS tailored to the specific organism under study.

Moreover, developing a high-throughput assay would significantly improve efforts in characterizing artificial MTSs designed in

this study. Here, we utilize fluorescent protein tagging and subsequent in vivo microscopy, a technique fundamental for studying protein localization in cellular biology<sup>59</sup>. However, this method is not suitable for quantifying the targeting efficiencies at scale. Therefore, deploying quantitative assays based on the self-assembling split green fluorescent protein (Split-GFP) technology<sup>60,61</sup>, fluorescence-activated cell sorting (FACS), and next-generation sequencing will generate high-quality sequence-to-targeting efficiency data and aid in identifying the best MTS for the passenger protein of interest. Furthermore, one could utilize the dataset to train a supervised ML model and screen MTSs or chimeric MTSs in silico. However, based on our HEMI1 delivery application, it is evident that swapping MTSs at the N-terminus significantly affects RNA levels. Therefore, it would be more logical to use codon language models<sup>62,63</sup> to represent MTSs, capturing both transcriptional and targeting efficiencies. Additionally, fundamental studies are needed to understand the mechanisms behind the increased targeting efficiency observed with chimeric MTSs. Recent work suggests that internal MTS-like signals can enhance import efficiency for certain proteins<sup>27,32</sup>. It would therefore be intriguing to investigate whether the later MTSs in a chimeric construct interact with the Tom70 receptor to enhance mitochondrial preprotein import efficiency.

In conclusion, our VAE models are highly capable of designing diverse, functional mitochondrial and new-to-nature dual-targeting sequences, improving our ability to deliver necessary cargo for metabolic engineering and biomedical applications.

## Methods

### Variational autoencoder

To train the VAEs, targeting peptides from the curated MTS dataset were one-hot encoded. Encoders and decoders, composed of fully connected layers, were trained using the loss function (Eq. 1), consisting of the reconstruction loss and Kullback–Leibler (KL) divergence<sup>64</sup>. The reconstruction loss was calculated using the binary cross-entropy function as the difference between the reconstructed output and the input data, encouraging the VAE to produce reconstructions that closely resemble the original data, while the KL divergence term was minimized to regularize the latent space by penalizing deviations from a standard normal distribution. The training process was stopped when the validation loss did not improve for five consecutive epochs. Moreover, hyperparameters, including the annealing rate and dropout, were optimized for both models to encourage meaningful representation in the latent space and prevent overfitting.

$$\text{Loss} = \frac{1}{N} \sum_{i=0}^N x_i \cdot \log(\hat{x}_i) + (1 - x_i) \cdot \log(1 - \hat{x}_i) + \frac{1}{2} \sum_{j=1}^L (1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2) \quad (1)$$

where  $x_i$  and  $\hat{x}_i$  are ground truth and reconstructed output (of size  $N$ ) at the position  $i$  and  $\mu_j$ ,  $\sigma_j$ , and  $L$  are mean, variance, and dimensionality of the latent variable  $\mathbf{z}$ , respectively.

### Sampling and annotating sequences for in vivo validation

Artificial MTSs were generated by feeding random numbers drawn from a normal distribution  $N(\mu = 0, \sigma = 1)$  as inputs to the decoder. The resulting reconstructions were converted into amino acid sequences employing the dictionary used for one-hot encoding. Sequences preceding the cleavage symbol '\$' were extracted and subjected to validation checks. Subsequently, these sequences were assigned organism labels based on their proximity to the cluster center (Eq. 2). Moreover, consistency in the local space was ensured by assessing which organism most of the 20 nearest neighbors belong to, verifying their alignment with the assigned label. Both artificial and naturally occurring MTSs were encoded using the pre-trained UniRep model  $\mathbf{U}$ . The cluster



center  $\mathbf{C}_O$  was computed as the mean of UniRep embeddings for MTSS within an organism  $O$ 's proteome (Eq. 3), while the nearest neighbors were identified by measuring the Euclidean distance between the UniRep embeddings of artificial peptides and MTSS of all four distinct eukaryotic organisms (Eq. 4). Finally, artificial MTSS were ranked according to their distances from the cluster centers of individual organisms and the top eight sequences were chosen for subsequent in vivo validation.

$$O_i = \operatorname{argmin}_O(d_{O,AMTS_i}) \quad (2)$$

$$\mathbf{C}_O = \frac{1}{T} \sum_{j=1}^T \mathbf{U}(MTS_{O,j}) \quad (3)$$

$$d_{O,AMTS_i} = \|\mathbf{U}(AMTS_i) - \mathbf{C}_O\| \quad (4)$$

where  $T$  is the total number of MTSS in the proteome of organism  $O$ , and  $O$  is one of the following: *S. cerevisiae*, *R. toruloides*, *N. tabacum*, or *H. sapiens*.

### Analysis of artificial targeting sequences

All VAE-generated sequences were fused to the N-terminal of GFP (without the start amino acid, methionine) and analyzed for functionality using the local software of DeepLoc 2.0<sup>4</sup>. The generated output was analyzed in-house based on the predicted subcellular localization or provided probability thresholds. Physicochemical features, including amino acid composition, net charge, Eisenberg hydrophobicity, and GRAVY score, were calculated using Biopython, while the hydrophobic moment was obtained using modIAMP<sup>30</sup>. Secondary structures of the artificial peptides were predicted using the S4PRED model<sup>36</sup>. Sequence logos for the analysis of cleavage sites were generated using WebLogo<sup>65</sup>. Clustal Omega<sup>66</sup> and pyMSAviz<sup>67</sup> were used to create and visualize MSAs, respectively.

### Profile Hidden Markov Model

The HMMER package<sup>29</sup> was employed with default settings to design MTSS with the pHMM. A multiple sequence alignment was obtained for 4,000 MTSS reported in Swiss-Prot using Clustal Omega<sup>66</sup> and was fed to HMMER's hmmbuild to create an HMM profile. Subsequently, HMMER's hmmit was utilized to generate 1000 putative MTSS, from which 730 sequences were randomly sampled and fused to GFP at the N-terminus to assess functionality using DeepLoc 2.0.

### Strains, media, and reagents

*E. coli* strain NEB10β (#C3019H; New England Biolabs, MA) was used for all cloning experiments. The following reference strains were used in the study: *S. cerevisiae* BY4741 (*MATα his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*), *S. cerevisiae* CEN.PK, *R. toruloides* 880CF (IFO0880-pANT1-SpCas9-pGAPDH1-MaFAR), *Agrobacterium tumefaciens* strain GV3101 (pMP90), tobacco *Nicotiana benthamiana*, and HEK293 (ATCC #CRL-3216). Reagents, buffer components, and growth media were purchased from Millipore Sigma (Burlington, MA), Qiagen (Germany), or Thermo Fisher Scientific (Hampton, NH). Restriction enzymes, T4 polynucleotide kinase (#M0201S), T4 DNA ligase (#M0202S), Q5 DNA polymerase (#M0491L), NEBuffer 3.1 (#B7203S), Gibson Assembly Master Mix (#E2611S), and HiFi DNA Assembly Master Mix (#E5520S) were purchased from New England Biolabs, MA. PrimeSTAR Max DNA Polymerase (#R045B) was purchased from Takara Bio. WT *PfAgo* and *PfAgo\** aliquots were obtained through protein expression and purification in *E. coli* KRX (Promega, WI)<sup>39</sup>. The plasmid CD3-991 (35S::ScCOX4-mCherry) was procured from the Arabidopsis Biological Resource Center (ABRC). All DNA oligonucleotides were ordered from Integrated DNA Technologies (IDT) (Coralville, IA), while

synthetic genes were codon-optimized for *S. cerevisiae* using IDT's Codon Optimization Tool and purchased from Twist Biosciences (San Francisco, CA). Targeting sequences were codon-optimized using IDT's Codon Optimization Tool or Juggler tool in BOOST<sup>68</sup> and ordered as oligos or gBlocks. Oligos, primers, and synthetic genes used in this study are listed in Supplementary Data 4.

### Plasmid construction

Scarless DNA assembly for the artificial MTSS was performed using *PfAgo* and T4 DNA ligase. Guide DNA phosphorylation, *PfAgo*/AREs cleavage of backbone fragments, and purification of *PfAgo* cleaved products were carried out prior to ligation<sup>39</sup>. Plasmid backbones were amplified using PrimeSTAR Max DNA Polymerase and digested with WT *PfAgo* and *PfAgo\**. To prepare the inserts, oligos containing the MTSS with protruding sticky ends were phosphorylated using T4 polynucleotide kinase. The reaction mixture was incubated at 37 °C for 1 h, 65 °C for 20 min, and cooled to 4 °C. An equimolar mixture of phosphorylated oligos was annealed in NEBuffer 3.1. The mixture was heated to 94 °C for 3 min and gradually cooled to 25 °C over a period of 45 min. The annealed product and the digested backbone were assembled using T4 DNA ligase. Subsequently, 3 μL of the ligation mixture was added to 50 μL of chemically competent NEB10β cells and transformation was performed following the manufacturer's protocol. The initial library of VAE-generated MTSS tested in *S. cerevisiae* was cloned using Gibson Assembly<sup>37</sup>. To create an NLS construct, pUC19\_SV40\_EGFP was cut with *Bsa*AI and then assembled using HiFi DNA assembly.

Correct assembly was first verified using Colony PCR and finally confirmed by Sanger sequencing. For plasmid amplification, the colony with the correct plasmid was cultured in LB medium supplemented with appropriate antibiotics. The plasmid DNA was purified from the cultures using the QIAprep Spin Miniprep Kit (#27104; Qiagen, Germany), following the manufacturer's protocol. The plasmids constructed in this study are listed in Supplementary Data 5.

### Analysis of MTS-GFP localization by fluorescence microscopy

To characterize MTSS in *S. cerevisiae*, AMTS-EGFP plasmids were transformed into *S. cerevisiae* using the commonly used lithium acetate heat shock protocol<sup>69</sup>. Cells with the AMTS-EGFP plasmids were inoculated overnight into 2 mL SC-URA media and cultivated at 30 °C and 250 rpm for approximately 18 h. When OD<sub>600</sub> reached 0.1–0.3, cells were centrifuged, and the media was replaced with 1 mL Phosphate-Buffered Saline (PBS; without calcium and magnesium) containing 200 nM of MitoTracker® Orange CMTMRos (#M7510; purchased from Invitrogen, USA and diluted in dimethyl sulfoxide to create 1 mM stock solution). The cells were stained at 30 °C for 15 min (culture tube was covered with foil, 250 rpm), pelleted, washed with 1 mL PBS twice, resuspended in PBS with 2% glucose, and transferred to a 1.5 mL microcentrifuge tube. The stained yeast cells were mounted onto Poly-L-Lysine coated slides (#75955-45; Masterflex) using ProLong™ Gold Antifade Mountant (#P36934; Invitrogen, USA) and analyzed using the ZEISS LSM880 Confocal Laser Scanning Microscope. Imaging was performed using a 63x oil immersion objective lens. A similar procedure was followed for *R. toruloides* strains expressing the integrated AMTS-EGFP construct with minor modifications. Cells were cultivated in SC media (pH 5.6) and stained with Rhodamine B, hexyl ester (Yeast Mitochondrial Stain Sampler Kit #Y7530 purchased from ThermoFisher Scientific).

To investigate MTSS in plant cells, *Agrobacterium tumefaciens* GV3101 (pMP90) transformed with the plasmids AMTS-YFP and CD3-991 were cultured overnight in 2 mL LB medium containing appropriate antibiotics at 28 °C. Bacterial cells were collected at 10,000 rpm for 1 min and resuspended with 2 mL water to wash out the antibiotics, followed by one more washing. Cells were then resuspended in the infiltration buffer (10 mM MgCl<sub>2</sub>, 10 mM MES pH 5.7, 200 μM

acetosyringone) and the concentration was adjusted to OD<sub>600</sub> = 0.6. Cells in the buffer were placed at room temperature for two hours before infiltration. Agrobacterium solution was infiltrated into leaves through a syringe without a needle. After three days, the fluorescence on the epidermis was observed and images were captured on a ZEISS LSM 710 confocal microscope. YFP was excited at 514 nm and emission was collected at 520–540 nm, while mCherry was excited at 561 nm and emission was collected from 580 to 620 nm.

HEK293 cells were transfected with 200 ng of MTS-GFP plasmids using Fugene HD (Promega #E2311) when the cell confluence reached approximately 60%. Transfection was carried out in a 12-well plate on Poly-L-lysine-coated glass coverslips. The growth media was changed at 24 hours and 48 hours. Post-transfection, live cells were incubated with MitoTracker® Orange (Thermo Fisher #M7510) for 30 minutes. The cells were fixed in 3.7% Formaldehyde for 15 minutes, followed by three 5-minute washes with 1xPBS. Subsequently, the coverslips were mounted on ProLong Diamond with DAPI (Thermo Fisher #P36962). The images were acquired at 100x magnification on an OMX-V4 microscope (GE Healthcare) equipped with a U Plan S-Apo 100x/1.40-NA oil-immersion objective (Olympus). The images were deconvolved using the previously described deconvolution parameters<sup>70</sup> and final images were created using Fiji<sup>71</sup>.

### Confirmation of MTS cleavage in HEK293 cells using Western Blot analysis

HEK293 cells were transfected with 500 ng of purified plasmid DNA using Lipofectamine 3000 (Invitrogen) in a 12-well plate for 48 hours. Subsequently, the cells were lysed in 150 µL of 1x sample buffer (Bio-Rad #1610747) containing 5% β-mercaptoethanol and heated at 95 °C for 5 minutes. An initial Western blot was performed with equal amounts of cell lysates loaded into each lane. The visual intensity of the GFP bands was assessed to determine the need for further normalization. A second gel was then run to ensure comparable GFP levels across lanes, followed by resolution on a 14% SDS-PAGE gel. The proteins were transferred onto a PVDF membrane (Bio-Rad #1704156) and incubated overnight with primary antibodies (GFP, Thermo Fisher #MA1-952, 1:2500; β-Tubulin, Proteintech #66240, 1:20,000) in 5% non-fat milk. Afterward, the membrane was incubated for 20 minutes with an HRP-conjugated goat anti-mouse secondary antibody (Jackson ImmunoResearch #115-036-003). The membrane was developed using a western ECL substrate (Bio-Rad #1705061) and imaged on Amersham ImageQuant studio.

### Strain construction

Prior to the random integration of the AMTS-EGFP expression cassette, the deletion of the carotenogenic reporter gene *CAR2* was performed. *R. toruloides* 880CF was transformed with the gRNA expression cassette along with the *hpt* gene encoding for hygromycin phosphotransferase amplified from a previously constructed pRTH-car2d vector<sup>41</sup>. A white colony demonstrating the phenotype of a successful knockout was picked for subsequent experiments. 1000 ng of AMTS-EGFP expression cassette along with the *nat* gene conferring the nourseothricin resistance was amplified using PCR, transformed into *R. toruloides* 880CF-CAR2Δ strain, and characterized using fluorescence microscopy. Transformation of *R. toruloides* was performed using the modified lithium acetate protocol<sup>41</sup>. The strains constructed in this study are listed in Supplementary Data 6.

### Quantification of 3-HP in engineered strains

The 3-HP producing strain, SCE-R2-200, was obtained from a previous study<sup>48</sup>. *PAND*, *BAPAT*, and *YDFG* genes were amplified from this strain, while promoters and terminators were sourced from the genome of the CEN.PK strain. Mitochondrial targeting sequences (COX4, AMTS131, AMTS3) were amplified from the MTS-GFP plasmids. All fragments included 60 bp overlaps with adjacent fragments, and

plasmid construction with pTH backbone was carried out using DNA assembler<sup>72</sup>. The constructed plasmids were then transformed into the CEN.PK strain via the commonly used lithium acetate heat shock protocol<sup>69</sup>. Three colonies for each strain were picked from the SC-URA plate and inoculated in a culture tube containing 2 mL of SC-URA media (50 g/L glucose). After five days of incubation at 30 °C and 250 rpm, the cultures were centrifuged, and the supernatant was collected and diluted twofold. The samples were then directly subjected to HPLC analysis. HPLC was performed using an Agilent 1260 system (Agilent, USA) with an Aminex HPX-87H Column #1250140 (BioRad, USA). The column was operated at 60 °C with 2.5 mM H<sub>2</sub>SO<sub>4</sub> at a flow rate of 0.6 mL/min. The 3-HP peak was observed at 12.2 minutes.

### Validating mitochondrial localization of AMTS131-BAPAT and AMTS3-YDFG enzymes

pHT-mB-His and pHT-mY-His were constructed using DNA assembler<sup>72</sup> and transformed into *S. cerevisiae* BY4741. A single colony was picked and inoculated into SC-URA medium containing 20 g/L glucose, then cultured at 30 °C with shaking at 250 rpm for two days. The cells were then collected, and the cytosolic and mitochondrial fractions were extracted using the Yeast Mitochondria Isolation Kit (#ab178779; Abcam, USA) according to the manufacturer's instructions. These fractions were analyzed by SDS-PAGE, followed by western blotting. Proteins were detected using a 6x-His Tag Monoclonal Antibody (HIS.H8, HRP; Thermo, USA) and Clarity Max Western ECL Substrate (#1705061; Bio-Rad, USA).

### Measuring targeting efficiency of chimeric MTSs

For 5-ALA quantification, 200 µL of culture medium was collected and mixed with 200 µL of 1 M acetate buffer (pH 4.6) and 100 µL of acetylacetone. The mixture was heated at 100 °C for 10 min. Subsequently, the mixture was diluted tenfold and mixed with an equal volume of modified Ehrlich's reagent (0.1 g/mL p-dimethylaminobenzaldehyde (DMAB) and 0.16 mg/mL hypochlorous acid in glacial acetic acid) and kept at room temperature for 10 min. The formation of a pink compound was observed and the absorbance was measured at 553 nm using a SpectraMax Mini Multi-mode microplate reader (Molecular Devices, USA).

For qPCR, strains harboring the HEM1-expressing plasmid were cultured in SC-URA for one day, and then total RNA was extracted using the RNeasy Mini Kit (#74104; Qiagen, USA). The total RNA was subjected to reverse transcription to obtain cDNA using the iScript™ Reverse Transcription Supermix (#1708840; BioRad, USA). qPCR was performed using iTaq Universal SYBR Green Supermix (#1725120; BioRad, USA) in the QuantStudio™ 6 Pro Real-Time PCR System (Thermo, USA). The HEM1 qPCR primers used were as follows: Forward: CAGGAGTCGGGTTTCGATTAC, Reverse: CTTGGCCAATCGGTTGATATTG. RNA extraction, cDNA synthesis, and qPCR were performed according to the manufacturer's protocol.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The data supporting the findings of this study are available within the article and its Supplementary Information files or uploaded through public repositories. Amino acid sequences of mitochondrial and chloroplast proteins utilized for training the VAE models are obtained from UniProt (<https://www.uniprot.org/>)<sup>23</sup>. Datasets curated for training the VAE models and analysis are on Zenodo<sup>73</sup>: <https://zenodo.org/records/14590156>. Source data are provided with this paper. If specific data is believed to be missing, that data is available from the corresponding author upon request. Source data are provided with this paper.

## Code availability

The source code and the trained VAE models developed in the study are available on GitHub at: <https://github.com/Zhao-Group/MTS-VAE> and on Zenodo at: <https://zenodo.org/records/14590156> [73]. The initial version of the scripts for generative models and data analysis were built on Python 3.9.16, PyTorch 1.9.1, Numpy 1.24.3, SciPy 1.10.1, BioPython 1.81, Scikit-learn 1.1.3, Pandas 1.5.3, modLAMP 4.3.0, and CD-HIT 4.8.1. A complete environment specification, including resolved dependencies, is available on GitHub in the form of an environment.yml file to ensure reproducibility.

## References

- Zhu, J. et al. A validated set of fluorescent-protein-based markers for major organelles in yeast (*Saccharomyces cerevisiae*). *mBio* <https://doi.org/10.1128/mbio.01691-19> (2019).
- Nelson, B. K., Cai, X. & Nebenführ, A. A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. *Plant J.* **51**, 1126–1136 (2007).
- Armenteros, J. J. A. et al. Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2**, e201900429 (2019).
- Thumuluri, V., Almagro Armenteros, J. J., Johansen, A. R., Nielsen, H. & Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* **50**, W228–W234 (2022).
- Burén, S. et al. Formation of nitrogenase NifDK tetramers in the mitochondria of *saccharomyces cerevisiae*. *ACS Synth. Biol.* **6**, 1043–1055 (2017).
- Hammer, S. K. & Avalos, J. L. Harnessing yeast organelles for metabolic engineering. *Nat. Chem. Biol.* **13**, 823–832 (2017).
- Volk, M. J. et al. Metabolic engineering: methodologies and applications. *Chem. Rev.* **123**, 5521–5570 (2023).
- Louzoun-Zada, S., Jaber, Q. Z. & Fridman, M. Guiding drugs to target-harboring organelles: stretching drug-delivery to a higher level of resolution. *Angew. Chem.* **131**, 15730–15740 (2019).
- Duran, L., López, J. M. & Avalos, J. L. Viva la mitochondria!: harnessing yeast mitochondria for chemical production. *FEMS Yeast Res.* **20**, foaa037 (2020).
- Spinelli, J. B. & Haigis, M. C. The multifaceted contributions of mitochondria to cellular metabolism. *Nat. Cell Biol.* **20**, 745–754 (2018).
- Gorman, G. S. et al. Mitochondrial diseases. *Nat. Rev. Dis. Prim.* **2**, 1–22 (2016).
- Schmidt, O., Pfanner, N. & Meisinger, C. Mitochondrial protein import: from proteomics to functional mechanisms. *Nat. Rev. Mol. Cell Biol.* **11**, 655–667 (2010).
- Kunze, M. & Berger, J. The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance. *Front. Physiol.* **6**, 259 (2015).
- Van Steeg, H., Oudshoorn, P., Van Hell, B., Polman, J. E. & Grivell, L. A. Targeting efficiency of a mitochondrial pre-sequence is dependent on the passenger protein. *EMBO J.* **5**, 3643–3650 (1986).
- Ford, H. C. et al. Towards a molecular mechanism underlying mitochondrial protein import through the TOM and TIM23 complexes. *eLife* **11**, e75426 (2022).
- Poveda-Huertes, D., Mulica, P. & Vögtle, F. N. The versatility of the mitochondrial presequence processing machinery: cleavage, quality control and turnover. *Cell Tissue Res.* **367**, 73–81 (2017).
- Bohovich, I., Chan, S. S. L. & Khalimonchuk, O. Mitochondrial protein quality control: the mechanisms guarding mitochondrial health. *Antioxid. Redox Signal.* **22**, 977–994 (2015).
- Wan, F., Kontogiorgos-Heintz, D. & Fuente-Nunez, C. de la. Deep generative models for peptide design. *Digital Discov.* **1**, 195–208 (2022).
- Dean, S. N. & Walper, S. A. Variational autoencoder for generation of antimicrobial peptides. *ACS Omega* **5**, 20746–20754 (2020).
- Wu, Z. et al. Signal peptides generated by attention-based neural networks. *ACS Synth. Biol.* **9**, 2154–2161 (2020).
- Schissel, C. K. et al. Deep learning to design nuclear-targeting abiotic miniproteins. *Nat. Chem.* **13**, 992–1000 (2021).
- Strokach, A. & Kim, P. M. Deep generative modeling for protein design. *Curr. Opin. Struct. Biol.* **72**, 226–236 (2022).
- UniProt Consortium, The UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
- Fukasawa, Y. et al. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell Proteom.* **14**, 1113–1126 (2015).
- Small, I., Peeters, N., Legeai, F. & Lurin, C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *PROTEOMICS* **4**, 1581–1590 (2004).
- Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
- Backes, S. et al. Tom70 enhances mitochondrial preprotein import efficiency by binding to internal targeting sequences. *J. Cell Biol.* **217**, 1369–1382 (2018).
- Bowman, S. R. et al. Generating sentences from a continuous space. In *Proc. 20th SIGNLL Conference on Computational Natural Language Learning* (eds Riezler, S. & Goldberg, Y.) 10–21 (Association for Computational Linguistics, Stroudsburg, PA, 2016).
- Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
- Müller, A. T., Gabernet, G., Hiss, J. A. & Schneider, G. modLAMP: Python for antimicrobial peptides. *Bioinformatics* **33**, 2753–2755 (2017).
- Obita, T., Muto, T., Endo, T. & Kohda, D. Peptide library approach with a disulfide tether to refine the Tom20 recognition Motif in mitochondrial presequences. *J. Mol. Biol.* **328**, 495–504 (2003).
- Bykov, Y. S. et al. Widespread use of unconventional targeting signals in mitochondrial ribosome proteins. *EMBO J.* **41**, e109519 (2022).
- Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
- Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
- Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **299**, 371–374 (1982).
- Moffat, L. & Jones, D. T. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics* **37**, 3744–3751 (2021).
- Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
- Maarse, A. C. et al. Subunit IV of yeast cytochrome c oxidase: cloning and nucleotide sequencing of the gene and partial amino acid sequencing of the mature protein. *EMBO J.* **3**, 2831–2837 (1984).
- Enghiad, B. et al. PlasmidMaker is a versatile, automated, and high throughput end-to-end platform for plasmid construction. *Nat. Commun.* **13**, 2697 (2022).
- Zhao, Y., Song, B., Li, J. & Zhang, J. Rhodotorula toruloides: an ideal microbial cell factory to produce oleochemicals, carotenoids, and other products. *World J. Microbiol. Biotechnol.* **38**, 13 (2021).
- Schultz, J. C. et al. Metabolic engineering of *Rhodotorula toruloides* IFO0880 improves C16 and C18 fatty alcohol production from synthetic media. *Microb. Cell Factories* **21**, 26 (2022).



42. Lee, D. W. & Hwang, I. Understanding the evolution of endosymbiotic organelles based on the targeting sequences of organellar proteins. *N. Phytologist* **230**, 924–930 (2021).
43. Sharma, M., Bennewitz, B. & Klösigen, R. B. Rather rule than exception? How to evaluate the relevance of dual protein targeting to mitochondria and chloroplasts. *Photosynth Res.* **138**, 335–343 (2018).
44. Carrie, C. & Small, I. A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochimica et. Biophysica Acta (BBA) - Mol. Cell Res.* **1833**, 253–259 (2013).
45. Ge, C., Spänning, E., Glaser, E. & Wieslander, Å. Import determinants of organelle-specific and dual targeting peptides of mitochondria and chloroplasts in *Arabidopsis thaliana*. *Mol. Plant* **7**, 121–136 (2014).
46. Lee, D. W. et al. Molecular mechanism of the specificity of protein import into chloroplasts and mitochondria in plant cells. *Mol. Plant* **12**, 951–966 (2019).
47. Caspari, O. D. et al. Converting antimicrobial into targeting peptides reveals key features governing protein import into mitochondria and chloroplasts. *Plant Commun.* **4**, 100555 (2023).
48. Borodina, I. et al. Establishing a synthetic pathway for high-level production of 3-hydroxypropionic acid in *Saccharomyces cerevisiae* via  $\beta$ -alanine. *Metab. Eng.* **27**, 57–64 (2015).
49. Zhang, Y. et al. Engineering yeast mitochondrial metabolism for 3-hydroxypropionate production. *Biotechnol. Biofuels Bioprod.* **16**, 64 (2023).
50. Galanis, M., Devenish, R. J. & Nagley, P. Duplication of leader sequence for protein targeting to mitochondria leads to increased import efficiency. *FEBS Lett.* **282**, 425–430 (1991).
51. Chin, R. M., Panavas, T., Brown, J. M. & Johnson, K. K. Optimized mitochondrial targeting of proteins encoded by modified mRNAs rescues cells harboring mutations in mtATP6. *Cell Rep.* **22**, 2818–2826 (2018).
52. Volland, C. & Felix, F. Isolation and properties of 5-aminolevulinate synthase from the yeast *Saccharomyces cerevisiae*. *Eur. J. Biochem* **142**, 551–557 (1984).
53. Brown, B. L., Kardon, J. R., Sauer, R. T. & Baker, T. A. Structure of the mitochondrial aminolevulinic acid synthase, a key heme biosynthetic enzyme. *Structure* **26**, 580–589.e4 (2018).
54. Li, J. et al. Chloroplastic metabolic engineering coupled with isoprenoid pool enhancement for committed taxanes biosynthesis in *Nicotiana benthamiana*. *Nat. Commun.* **10**, 4850 (2019).
55. Schmitt, L. T., Paszkowski-Rogacz, M., Jug, F. & Buchholz, F. Prediction of designer-recombinases for DNA editing with generative deep learning. *Nat. Commun.* **13**, 7966 (2022).
56. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A package for data mining of antimicrobial. *Pept. R. J.* **7**, 4–14 (2015).
57. Sevgen, E. et al. ProT-VAE: Protein Transformer Variational Auto-Encoder for Functional Protein Design. 2023.01.23.525232 Preprint at <https://doi.org/10.1101/2023.01.23.525232> (2023).
58. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
59. Huh, W.-K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
60. Cabantous, S., Terwilliger, T. C. & Waldo, G. S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat. Biotechnol.* **23**, 102–107 (2005).
61. Bader, G. et al. Assigning mitochondrial localization of dual localized proteins using a yeast Bi-Genomic Mitochondrial-Split-GFP. *eLife* **9**, e56649 (2020).
62. Li, S. et al. CodonBERT large language model for mRNA vaccines. *Genome Res.* **34**, 1027–1035 (2024).
63. Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nat. Mach. Intell.* **6**, 170–179 (2024).
64. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (eds. Bengio, Y. & LeCun, Y.) (2014).
65. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).
66. McWilliam, H. et al. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* **41**, W597–W600 (2013).
67. Shimoyama, Y. pyMSAviz: MSA visualization python package for sequence analysis. <https://github.com/moshi4/pyMSAviz> (2022).
68. Oberortner, E., Cheng, J.-F., Hillson, N. J. & Deutsch, S. Streamlining the design-to-build transition with build-optimization software tools. *ACS Synth. Biol.* **6**, 485–496 (2017).
69. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
70. Singh, N. et al. Redefining the specificity of phosphoinositide-binding by human PH domain-containing proteins. *Nat. Commun.* **12**, 4339 (2021).
71. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
72. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* **37**, e16 (2009).
73. Boob, A. G. et al. Dataset for Design of diverse, functional mitochondrial targeting sequences across eukaryotic organisms using variational autoencoder. <https://doi.org/10.5281/zenodo.14590156> (2024).

## Acknowledgements

This work was funded by the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number DE-SC0018420). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy. The online tool BioRender (biorender.com) was used to create Figs. 1, 3, 5, and 7, and Supplementary Fig. 12. We thank Prof. Irina Borodina (DTU Biosustain) for providing the 3-HP producing strain, SCE-R2-200. We thank Dr. Behnam Enghiad for his suggestions on *PfAgo*-based assembly and Dr. Austin Cyphersmith from Core Facilities at the Carl R. Woese Institute for Genomic Biology for his help with the fluorescence microscopy. The authors acknowledge the use of computing facilities of Biocluster at the Carl R. Woese Institute for Genomic Biology and Nano cluster at the National Center for Supercomputing Applications. This work also used the Delta system at the National Center for Supercomputing Applications through allocation BIO230077 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

## Author contributions

A.G.B. and H.Z. conceived and designed the study. A.G.B. performed the computational experiments. A.G.B., S.-I.T., A.Z., N.S., X.X., S.Z., and T.A.M. performed the wet-lab experiments, and analyzed the data. A.G.B., S.-I.T., N.S., X.X., L.-Q.C., and H.Z. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59499-3>.

**Correspondence** and requests for materials should be addressed to Huimin Zhao.

**Peer review information** *Nature Communications* thanks Fabien Plisson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025