nature portfolio

Peer Review File

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

Corresponding Author: Professor Huimin Zhao

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this work, Boob et al. present an elegant and detailed machine learning framework for the design of artificial mitochondrial targeting signals (AMTSs). The authors utilize a training dataset of more than 50,000 MTSs from various organisms, implementing a variational auto encoder (VAE) approach to generate artificial MTS sequences. The authors compare these AMTSs to endogenous MTSs using a variety of predictors for localization, hydrophobicity, secondary structure, length, and more. They also refine their initial species-pooled outputs to generate species-specific AMTSs for humans, yeast, and plants. Next, the authors validate their species-specific AMTSs using in vitro assays, by showing that AMTS-tagged GFP localizes to mitochondria in three different organisms (HEK293 cells, S. N. benthamiana, S. cerevisiae). They also utilize their VAE approach on mitochondrial and chloroplast targeting signals (CTSs) from the Viridiplantae kingdom to generate MTSs, CTSs, and dual targeting MTS-CTSs, highlighting the change in characteristics which favor targeting to one or both organelles, while highlighting that dual-targeting signals are more likely to have evolved from MTSs due to their simpler mutational path. Lastly, the authors apply their AMTSs to two systems in yeast: first, they engineer MTS and AMTS-fusion proteins (PAND, BAPAT, YDFG) which increase 3-HP yield compared to their cytosolically targeted counterparts. Second, they engineer 5-aminolevulinate synthase (HEM1) fusion proteins containing COX4 MTS or AMTSs singly and in series, which improve the targeting efficiency of HEM1 compared to the endogenous HEM1 MTS.

Overall, the manuscript is well-written, original, and the conclusions are of broad significance to mitochondrial researchers. Compared to the established literature, Boob et al. have extended the toolkit of pre-existing mitochondrial predictors and have provided a completely novel approach to generate and validate new MTS sequences which are not found in nature.

Given that our expertise lies in mitochondrial biochemistry and not machine learning (ML), we are unable to comment directly on their coding approach, though the authors present their approach and findings in an accessible manner. We also believe that the authors provide sufficient methodological detail and open-source code for their results to be reproduced, though I encountered a few issues when trying to run their scripts (https://github.com/Zhao-Group/MTS-VAE) to reproduce the data within the manuscript, which I will detail below. I believe that during revisions, these issues should be straightforward to resolve.

We commend the authors for their hard work and excellent manuscript. This manuscript is suitable for publication, though there are some points that the authors should address which will improve the quality of the manuscript and will ultimately broaden the utility of their findings beyond computational researchers towards a larger audience of mitochondrial biologists.

Major Points

1. Supplementary Fig 5: “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.”

There two AMTSs which are in Supp Fig 5 but are not mentioned at all in the text (110, 245), presumably because they do not localize to mitochondria (i.e. phenocopy the negative control). I believe the authors should be transparent and address

these two AMTSs in the main text with these as well, if they are in the figure.

More critically, are there any biochemical (i.e. sequence features) or computational (i.e. TargetP scores) rationalizations for why some of these AMTSs are able to localize vs. partially localize vs. not localize at all? I understand that this manuscript will spur more sophisticated predictions on targeting rates and/or import efficiencies (as stated in the manuscript discussion), but I am wondering if there might be any preliminary hypotheses for this subset of AMTSs. It is not clear to me what the amino acid sequences of these various AMTSs are, nor where to find them easily in the supplementary files. If a rationalization is not possible, it would be quite useful to see a sequence alignment (as in Fig S14) or some annotated figure with these sequences, as well as some analysis on their predicted cleavage sites, especially since some of these are used later in Fig 6B (see comment below).

2. Fig 6B: This experiment is missing some important controls to justify the author's claims and prove that these AMTS-fusion enzymes are truly influencing 3-HP production. I understand that the authors show that these specific AMTSs localize GFP to mitochondria earlier in the manuscript, but it is not clear to me whether this 3-HP increase would be equivalent by just expressing COX4-PAND alone (i.e. having mitochondrial PAND and cytosolic or inactive BAPAT/YDFG). These claims would benefit from confirming the cytosolic vs. mitochondrial localization of these fusion proteins via mitochondrial fractionation and Western Blotting (if there are suitable antibodies), and/or potentially having single enzyme controls of each mitochondrially targeted protein (e.g. showing an additive increase in 3-HP, or a significant increase only when all three are present). As stated throughout the manuscript, the import efficiency of MTSs is dependent on the passenger protein, so just because GFP localizes well to mitochondria does not necessarily mean these fusion AMTS-enzymes will be imported efficiently, be cleaved, or functional. Given that the authors have also not shown whether AMTS131 and AMTS3 are MPP cleavable, it would be important to have these controls to show that AMTS131-BAPAT and AMTS3-YDFG are indeed localized, functional, and are contributing positively to the observed 3-HP increase. As in point 2, this experiment would greatly benefit from highlighting the amino acid sequences and potential MPP cleavage sites of these AMTSs.

3. While the authors have provided a supplemental file for the Top 1000 artificial peptides from the RNN-VAE model, I believe the manuscript would be of broader utility (especially for further biochemical characterization of AMTS import rates/efficiencies) if the authors would also provide a list of their AMTSs that have been re-sampled to the specific organisms (i.e. an extension of Supplementary Table 2). This would allow other non-computational researchers to rapidly profile novel MTSs or detect similarity to one of their proteins of interest, without requiring coding proficiency or dedicated Linux machines to run the analysis on their own.

4. I am especially intrigued by the mechanisms by which these AMTSs target to mitochondria. More specifically, do these AMTSs maintain some conserved Tom20 binding mechanisms that are inherent to most endogenous N-terminal MTSs? The authors should be able to try some bioinformatics analysis and look for the presence of Tom20 binding motifs, at least in the yeast AMTSs (see Bykov et al. 2021, Muto et al, 2001; Obita et al, 2003). I believe the manuscript might benefit from a brief discussion on import mechanisms (e.g. Tom20 dependency), even if the bioinformatics approach is inconclusive.

Minor points
5. Fig 2A: "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 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."

Fig 2A might benefit from annotating the clusters, as it remains unclear to me which cluster is which.

6. "Our earlier analysis indicated that 27 and 17 of the VAE-generated peptides localize GFP to the chloroplast (plastid) or both the chloroplast (plastid) and mitochondrion (Supplementary Fig. 3)"

I think it is important to clearly state in this sentence that they were *predicted* to localize GFP to the chloroplast, mitochondria, or both, not that they truly localize.

7. Fig S10: "We used TargetP 2.0-predicted MTS of RtCOX4 as a control. However, we did not observe GFP localized to mitochondria." Is there any rationalization for this? COX4-MTS fusion should localize to mitochondria, so it is perplexing as to why some of the AMTSs localize fine but not COX4.

8. Fig 2B: The authors state that their AMTSs follow a "similar distribution", but are any of these changes statistically significant (i.e. for alanine)?

9. "For the GFP western blot, the cell lysates were normalized to ensure comparable GFP levels and were then resolved on a 14% SDS-PAGE gel." Which method of protein quantitation was used?

10. Fig 6C: There is no mention anywhere in the text or figure legends what the negative control "ReHelper5" is.

11. It might be useful to mention the existence of internal MTSs somewhere in the manuscript, given that there are a subset of proteins which can still target to mitochondria in the absence of a canonical N-terminal MTS (Bykov et al. 2021, Backes et al. 2018). While the authors state that the import efficiency of targeted proteins is dictated by the passenger protein (citing Van Step et al. 1986), this variability could be partially explained by the relationship between Tom70 and internal MTSs of the passenger protein.

12. The training dataset featured both IMM and matrix localized proteins, but for all of the outputted AMTSs, there is no further mention of submitochondrial localization throughout the manuscript. Presumably, most of the shorter AMTSs should just localize proteins to the mitochondrial matrix, but it may be interesting to mention whether any of the AMTSs generated (especially the longer ones) feature stop transfer signals or IMM sorting signals based on the training dataset. It might also be interesting to run the AMTS fusion proteins through something like DeepMito (Savojardo et al. 2019), to profile whether any of the AMTSs might alter a protein's localization between inner membrane or matrix localized, though I acknowledge that this computational approach may be flawed.

13. In the text, one aspect of the authors' rationale for requiring AMTSs is that:

"… overuse of a single, endogenous targeting sequence may saturate the import machinery, resulting in potential competition with native proteins during the import system, and its downstream accumulation can compromise mitochondrial protein biogenesis and integrity15,16."

This claim cites two reviews, though it is unclear to me in which system or MTS the authors are referring to. I understand the general concept is intuitive, but presumably a saturation of import machinery would be due to slow import rates (i.e. TOM/TIM passage) of a given MTS, or due to slow proteolytic processing within mitochondria. Some clarification or primary citations could be useful here, if possible. Is this saturation coming from slow proteolytic processing or slow TOM/TIM passage, or both? I understand that overexpression of suboptimal MTS-proteins, especially with stop transfer signals, could clog the TOM/TIM channels, but I am naive to issues that would arise from "overuse" of an MTS-fusion protein when the MTS occurs endogenously, yet is efficiently targeted and cleaved within mitochondria. Overall, I agree that broadening the pool of MTS sequences is incredibly useful, but it does carry the caveat of needing to characterize the import and proteolytic efficiencies of these AMTSs prior to or during their in vitro use (a process which is highlighted in the manuscript discussion and could be expedited by addressing major point #4).

14. Along similar lines to my comments on whether these AMTSs would all utilize the typical Tom20-dependent import routes, it could be interesting to look at some of the species-specific AMTS sequences which are furthest in Levenshtein distance from endogenous sequences (i.e. if there are any distant AMTSs that are still predicted to be mitochondrially localized)? In this sub-analysis, perhaps the authors might uncover some atyptical AMTSs… I acknowledge that the prediction algorithms may not be able to correctly predict the localization of these AMTSs, but it could be interesting to pursue or at least list some of the most dissimilar AMTSs.

Syntax/Grammar

15. "However, as MTS is positioned at the N-terminus of the protein, it can also influence transcription rates and, consequently, mRNA levels " should read "as MTSs are positioned at the…" or "as an MTS is positioned".

(Remarks on code availability)
I was unable to install and run the code from the authors' GitHub to reproduce the raw data used in the manuscript. First, during installation of the requirements, I received a Pip error, potentially because the deeploc2 package has changed since the original upload of the repository. I also ensured that I was running Python 3.10, which makes it unclear to me where the first error is coming from.

Pip subprocess error:
ERROR: Ignored the following versions that require a different python version: 1.0.0 Requires-Python >=3.10; 1.7.1 Requires-Python >=3.10
ERROR: Could not find a version that satisfies the requirement deeploc2==1.0.0 (from versions: none)
ERROR: No matching distribution found for deeploc2==1.0.0

Second, when attempting to run the python scripts (given that most of the scripts should be able to run without deeploc2) as described in the GitHub, I received errors regarding missing .pkl files. I searched all the Python scripts for code that would generate these .pkl files, but the only logic for generating .pkl files I could find was within the Dual/analysis/split.py code, which is not specified to be run until later. Specifically, here is the file not found error generated from train.py:

(py3) MTS-VAE$ python MTS/scripts/train.py
Traceback (most recent call last):
File "/MTS/scripts/train.py", line 14, in <module>
with open('MTS/data/tv_sim_split_train.pkl', 'rb') as f:
FileNotFoundError: [Errno 2] No such file or directory: 'MTS/data/tv_sim_split_train.pkl'

Also, when trying to run the Dual/analysis/split.py code under these circumstances, it requires the 'sklearn' module, which does not seem to be included in the environment.yml dependencies.

It is possible that I may be missing something, but I believe that if I encountered these errors then the common user would also likely have difficulties in generating the data used in the paper. I would appreciate if the authors could guide me to resolve these issues, either by editing their code, providing .pkl files, or suggesting some potential fixes.

Finally, while it may be common sense to others within the ML field, I believe the GitHub readme should specify that their code requires Linux/Unix, given that certain dependencies from the environment.yml were incompatible with both Windows and MacOS when I attempted.

If agreed upon by the editor, I would be open to the authors providing suggested code fixes prior to the full rebuttal letter, given that these updates should be rapid and would give me more time to troubleshoot and run all of the code again on my end.

Reviewer #2

(Remarks to the Author)
In their manuscript titled "Design of diverse, functional mitochondrial targeting sequences across eukaryotic organisms using variational autoencoder", the authors Aashutosh Girish Boob, Shih-I Tan, Airah Zaidi, Nilmani Singh, Xueyi Xue, Shuaizhen Zhou, Teresa A. Martin, Li-Qing Chen, and Huimin Zhao present an innovative approach using Variational Autoencoders (VAE) to design novel mitochondrial targeting sequences (MTSs). The study demonstrates the functionality and diversity of these sequences across four eukaryotic organisms and extends the design to dual-targeting sequences that work for both mitochondria and chloroplasts. This work highlights the potential of AI-driven methods in mitochondrial engineering, combining deep learning with experimental biology in an impressive way.
The authors confirm the prior knowledge of the physicochemical and structural characteristics necessary for mitochondrial import through VAE-generated sequences. One of the notable aspects of the study is the creative use of latent space interpolation to design dual-targeting peptides, a concept that adds valuable insight to understanding evolutionary design principle.
However, the study suffers from several key limitations. One major issue is the bias in the training dataset, which relies heavily on the TargetP prediction model. This bias may constrain the model's potential to discover truly novel sequences features, as the VAE could simply be learning the patterns already encoded in TargetP rather than discovering new principles. Therefore, additional controls, particularly bioinformatics-based validations, are needed to fully evaluate the novelty of the findings.

Major Comments:
1. Training Dataset Bias: The model was trained on 56,660 peptides, with only 4,595 MTSs experimentally determined from Swiss-Prot. The risk here is that the VAE may learn latent variables closely aligned with the TargetP model. The authors need to demonstrate that the VAE-generated sequences go beyond what is already encoded in TargetP.
2. Lack of a Baseline Model: The study lacks a clear baseline model to evaluate the overall performance and novelty of the VAE approach. A comparison with simple models that enrich for physicochemical properties or a rejection sampling method against TargetP validation would help assess how much value the VAE adds.
3. Encoding Weakness: The use of one-hot encoding is a limitation, as it inadequately captures positional effects and interactions between amino acids at varying distances, rendering the approach blind to certain structural import features known to affect MTS efficacy.
4. Comparison with Existing Work: The authors need to compare their work with recent findings, such as those presented by Sidorczuk et al. (2023), who detailed physicochemical and structural properties of MTSs. The manuscript should clarify what novel insights the VAE approach offers beyond this.
5. MTS sequence space: The UMAP results show the VAE model filling both high- and low-density regions in sequence space, but it is unclear if this reflects a natural distribution of MTSs. VAEs are prone to overfitting latent space, and the authors should explain how they mitigated this risk in their model.

Minor Comments:
1. MTS Length Restrictions: The authors restrict MTS lengths to 11-69 amino acids, but they do not justify this threshold. Given the importance of MTS length in mitochondrial targeting, a clearer explanation of this choice is necessary.
Overall, while the manuscript presents innovative use of deep learning for designing mitochondrial targeting sequences, it requires additional validation and baseline comparisons to ensure the novelty of its findings and address potential biases in the model's design.

(Remarks on code availability)
The availability of the code is highly appreciated, making it accessible for further use and validation. The repository includes a well-written README, which provides clear guidance, and the overall structure of the repository is organized and easy to navigate.

While containerization is not strictly necessary at this stage, implementing it would significantly enhance reproducibility across different environments. It is encouraged for future improvements to streamline deployment and ensure consistency.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The present manuscript describes the design of artificially generated mitochondrial targeting sequences (MTSs). These sequences are usually positively charged alpha-helical peptides comprising 10 to 120 residues. Without specific motifs, the authors must rely on physicochemical and structural properties to generate novel MTSs and employ a variational autoencoder (VAE). The authors assembled 4,595 unique sequences from SwissProt, primarily from model organisms. They expanded the dataset to a final 56,660 peptides containing between 11 and 69 residues in length to train their generative model. Each sequence was padded to a 70-long vector before being one-hot encoded. The final architecture of the generator was a VAE with fully connected layers.

The authors predicted the subcellular localisation into the mitochondrion and the sequence diversity of their VAE-generated MTSs using DeepLoc 2.0 and Levenshtein distance, respectively. They demonstrated that their generator captured the global characteristics of MTSs well by comparing the artificial sequences to naturally occurring sequences. Overall, the VAE-generated sequences shared similar amino acid composition – rich in A, I, L, and S and depleted in D, E – and physicochemical properties (net positive charge, higher hydrophobic moment, negative Eisenberg-scaled hydrophobicity). Finally, the artificial sequences were more likely to fold into alpha-helices than their natural counterparts (Fig. 2).

The authors wanted to validate their design by producing artificially generated MTSs from four biological expression systems. First, they selected nine representative candidates from the million artificial MTSs (or AMTSs), showing TargetP 2.0 scores above 0.9 for in vivo evaluation in S. cerevisiae. These sequences and the MTS of COX4 (positive control) were attached to GFP plasmids to verify their mitochondrial targeting capability. Four of the AMTSs demonstrated selective targeting, whereas three displayed target multiple subcellular locations. These results prompted the hypothesis that successful and selective targeting of the mitochondrion was taxonomically biased within the amino acid sequence of MTSs. In other words, MTSs from different organisms preferred varying amino acid residues. They confirmed their hypothesis by capturing amino-acid, physicochemical, and structural attributes of MTSs across taxa – S. cerevisiae, Homo sapiens, N. benthamiana, and R. toruloides (Figure 3a and SI Fig. 6). Consequently, they devised a taxon-specific sampling scheme over random selection using kNN algorithm. They chose 32 AMTSs (8 peptides per organism) based on the distance from the cluster centre in the four aforementioned eukaryotic systems. All eight peptides successfully targeted the human mitochondria (HEK293T cells), six out of eight for N. benthamiana and six others for R. toruloides.

The authors explored the latent space of mitochondrial targeting sequences to design sequences targeting multiple subcellular locations for dual organelle engineering – targeting mitochondrion and/or chloroplast in plants. They trained a dual-VAE from the plant-derived sequences. They observed a smooth transition between the two posterior distributions in the latent space, representing MTSs and CTSs (Fig. 5). The former sequences possessed a higher ratio of R and hydrophobic moment. In contrast, the latter includes more prolines discouraging secondary structure formation (alpha-helix or beta-sheet).

Finally, the authors demonstrated the applications of AMTSs with enhanced metabolic engineering (i.e., 3-hydroxypropionic acid) and the efficient delivery of cargo proteins to the mitochondrion.

Overall, the manuscript is well-written despite being information-dense. Their protocols are clear and detailed, and the figures are easy to understand. Please consider the following minor comments;

(1) On page 5, line 191, can we talk about random selection when the nine VAE-generated MTSs were selected using TargetP 2.0 scores above 0.9 for in vivo evaluation in S. cerevisiae?

(2) The authors evaluated the mitochondrial targeting capability of nine AMTSs. On page 6, lines 200-202, they only discussed the results for seven of these sequences. What can they say about sequences 110 and 245 (Supplementary Fig. 5)?

(3) Typos
Page 5, line 177 Glutamic acid “I” to “E”
Page 22, extended Fig. 1a and 1e legend “I4” to “C”.
Supplementary Information page 1 “Error! Book not defined”.

(Remarks on code availability)
The code is available in a GitHub repository under a General Public License-3.0. The authors detailed which Python script is available for what purpose in the README file. The code is clean, with limited comments/annotations, and uses Python 3.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors went far and beyond to address our concerns ("reviewer 1"). This is an excellent article which is suitable for publication.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
After carefully reviewing the authors' revisions, I confirm that they have addressed all of my comments and concerns to my satisfaction. Therefore, I am pleased to recommend that the paper is now ready for publication.

(Remarks on code availability)
The code is publicly available in a GitHub repository under the GPL-3.0 license, ensuring it can be freely used, modified, and distributed. The repository features a well-structured and informative README, which provides clear guidance and documentation. While incorporating containerization or using a workflow language would further enhance reproducibility, these practices are currently recommended rather than considered a mandatory standard.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors have comprehensively addressed the reviewers' suggestions and concerns, including mine. They provided explanations and updates on critical issues, i.e. training dataset bias, coding errors, and reproducibility. They clarified their ambiguous claims, adding benchmarking analysis with modlAMP. They resolved their typological and grammatical errors. Overall, the rebuttal was well-executed, making the manuscript commendable and ready for publication.

(Remarks on code availability)

Response to Reviewer #1

In this work, Boob et al. present an elegant and detailed machine learning framework for the design of artificial mitochondrial targeting signals (AMTSs). The authors utilize a training dataset of more than 50,000 MTSs from various organisms, implementing a variational auto encoder (VAE) approach to generate artificial MTS sequences. The authors compare these AMTSs to endogenous MTSs using a variety of predictors for localization, hydrophobicity, secondary structure, length, and more. They also refine their initial species-pooled outputs to generate species-specific AMTSs for humans, yeast, and plants. Next, the authors validate their species-specific AMTSs using in vitro assays, by showing that AMTS-tagged GFP localizes to mitochondria in three different organisms (HEK293 cells, S. N. benthamiana, S. cerevisiae). They also utilize their VAE approach on mitochondrial and chloroplast targeting signals (CTSs) from the Viridiplantae kingdom to generate MTSs, CTSs, and dual targeting MTS-CTSs, highlighting the change in characteristics which favor targeting to one or both organelles, while highlighting that dual-targeting signals are more likely to have evolved from MTSs due to their simpler mutational path. Lastly, the authors apply their AMTSs to two systems in yeast: first, they engineer MTS and AMTS-fusion proteins (PAND, BAPAT, YDFG) which increase 3-HP yield compared to their cytosolically targeted counterparts. Second, they engineer 5-aminolevulinate synthase (HEM1) fusion proteins containing COX4 MTS or AMTSs singly and in series, which improve the targeting efficiency of HEM1 compared to the endogenous HEM1 MTS.

Overall, the manuscript is well-written, original, and the conclusions are of broad significance to mitochondrial researchers. Compared to the established literature, Boob et al. have extended the toolkit of pre-existing mitochondrial predictors and have provided a completely novel approach to generate and validate new MTS sequences which are not found in nature.

Given that our expertise lies in mitochondrial biochemistry and not machine learning (ML), we are unable to comment directly on their coding approach, though the authors present their approach and findings in an accessible manner. We also believe that the authors provide sufficient methodological detail and open-source code for their results to be reproduced, though I encountered a few issues when trying to run their scripts (https://github.com/Zhao-Group/MTS-VAE) to reproduce the data within the manuscript, which I will detail below. I believe that during revisions, these issues should be straightforward to resolve.

We commend the authors for their hard work and excellent manuscript. This manuscript is suitable for publication, though there are some points that the authors should address which will improve the quality of the manuscript and will ultimately broaden the utility of their findings beyond computational researchers towards a larger audience of mitochondrial biologists.

Response:
We express our gratitude to the reviewer for the positive comments and valuable feedback.

Major Points
1. Supplementary Fig 5: “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.”

There two AMTSs which are in Supp Fig 5 but are not mentioned at all in the text (110, 245), presumably because they do not localize to mitochondria (i.e. phenocopy the negative control). I believe the authors should be transparent and address these two AMTSs in the main text with these as well, if they are in the figure.

More critically, are there any biochemical (i.e. sequence features) or computational (i.e. TargetP scores) rationalizations for why some of these AMTSs are able to localize vs. partially localize vs. not localize at

all? I understand that this manuscript will spur more sophisticated predictions on targeting rates and/or import efficiencies (as stated in the manuscript discussion), but I am wondering if there might be any preliminary hypotheses for this subset of AMTSs. It is not clear to me what the amino acid sequences of these various AMTSs are, nor where to find them easily in the supplementary files. If a rationalization is not possible, it would be quite useful to see a sequence alignment (as in Fig S14) or some annotated figure with these sequences, as well as some analysis on their predicted cleavage sites, especially since some of these are used later in Fig 6B (see comment below).

Response:
We have updated the text to discuss the results for AMTS 110 and 245: '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.'

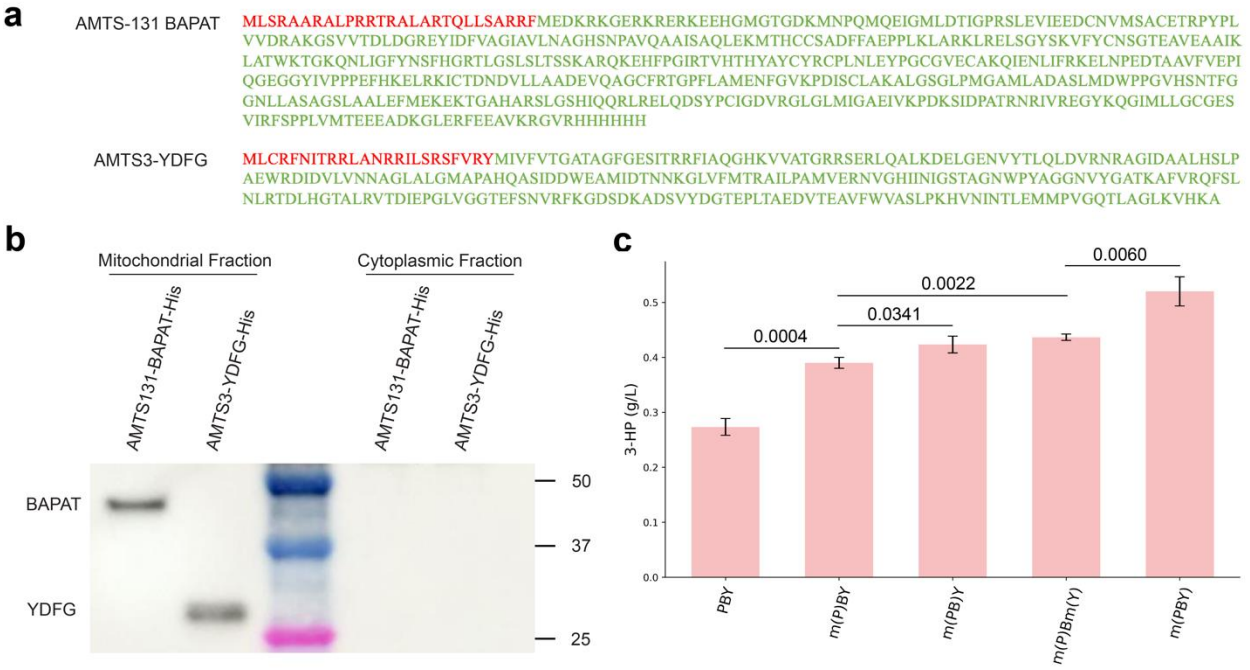
The amino acid sequences of the AMTSs characterized in this study are reported in Supplementary Table 2. We have updated the text to clarify this: 'Based on the confidence gained from the *in silico* analysis, we randomly selected nine VAE-generated MTSs with TargetP 2.0 scores above 0.9 for *in vivo* evaluation in *S. cerevisiae* (Supplementary Table 2).'

The table also includes the predicted TargetP mTP probability scores and cleavage probabilities of these sequences. While all sequences have mTP probabilities > 0.85, a good number of these artificial MTSs are non-functional *in vivo*. We also performed a bioinformatic search for the TOM20 recognition element 'φχβφφ' to devise a rationalization. However, we did not observe any correlation between TargetP scores, physicochemical properties such as net charge or hydrophobic moment, or the presence/absence of the TOM20 recognition element and localization outcomes. Therefore, as stated in the discussion, the next step is to design a large library of AMTSs and develop a high-throughput platform to spur more sophisticated predictions on targeting accuracy and/or import efficiencies.

2. Fig 6B: This experiment is missing some important controls to justify the author's claims and prove that these AMTS-fusion enzymes are truly influencing 3-HP production. I understand that the authors show that these specific AMTSs localize GFP to mitochondria earlier in the manuscript, but it is not clear to me whether this 3-HP increase would be equivalent by just expressing COX4-PAND alone (i.e. having mitochondrial PAND and cytosolic or inactive BAPAT/YDFG). These claims would benefit from confirming the cytosolic vs. mitochondrial localization of these fusion proteins via mitochondrial fractionation and Western Blotting (if there are suitable antibodies), and/or potentially having single enzyme controls of each mitochondrially targeted protein (e.g. showing an additive increase in 3-HP, or a significant increase only when all three are present). As stated throughout the manuscript, the import efficiency of MTSs is dependent on the passenger protein, so just because GFP localizes well to mitochondria does not necessarily mean these fusion AMTS-enzymes will be imported efficiently, be cleaved, or functional. Given that the authors have also not shown whether AMTS131 and AMTS3 are MPP cleavable, it would be important to have these controls to show that AMTS131-BAPAT and AMTS3-YDFG are indeed localized, functional, and are contributing positively to the observed 3-HP increase. As in point 2, this experiment would greatly benefit from highlighting the amino acid sequences and potential MPP cleavage sites of these AMTSs.

Response:
We performed additional experiments to address reviewers' concerns. First, we confirmed the mitochondrial localization of AMTS131-BAPAT and AMTS3-YDFG using fractionation and Western blot analysis (Supplementary Fig. 16b). Next, we constructed a series of plasmids in which one or multiple enzymes were targeted to mitochondria. The data clearly show that having all three enzymes localized to mitochondria provides the maximum improvement in 3-HP production (Supplementary Fig. 16c).

Additionally, we present the amino acid sequences of these fusion enzymes (AMTS131-BAPAT and AMTS3-YDFG) and highlight the potential MPP cleavage sites within these AMTSs (Supplementary Fig. 16a).



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.

We have added these results to Supplementary Information and updated the main text to:
'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).',

added protocol for fractionation and Western blot analysis of AMTS-fusion proteins:
'**Validating Mitochondrial Localization of AMTS131-BAPAT and AMTS3-YDFG Enzymes**
pHT-mB-His and pHT-mY-His were constructed using DNA assembler[69] 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 (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 (Bio-Rad, USA).' and

included primers used for constructing plasmids for these control experiments and the respective plasmids in Supplementary Tables 4 and 5.

Primers:

Name	Sequence
MB-F	taaaaaacacgctttttcagttcgagtttactttcgagaaggatattatttcg
MB-R	ccttttcggttagagcggatCTAATGATGATGATGATGATGcaattgagccaaacattcc
MY-F	taaaaaacacgctttttcagttcgagtttaggaagtaccttcaaagaatgg
MY-R	ccttttcggttagagcggatCTAATGATGATGATGATGATGctgacggtggacattcag
cyc1t-backbone-F	CATCATCATCATCATCATTAGatccgctctaaccgaaaagg
backbone-MB-R	tttatttcgaaataatatccttctcgaaagtaaactcgaactgaaaaagcg
backbone-MY-R	gataagaccccattctttgaaggtacttcctaaactcgaactgaaaaagcg

Plasmids:

pHT-m(P)BY	GAPp-cox4-TcPAND-PGK1t-TEF1p-BcBAPAT-ADH1t-PGK1p-EcYDFG-CYC1t in pHT backbone
pHT-m(PB)Y	GAPp-cox4-TcPAND-PGK1t-TEF1p-AMTS131-BcBAPAT-ADH1t-PGK1p-EcYDFG-CYC1t in pHT backbone
pHT-m(P)Bm(Y)	GAPp-cox4-TcPAND-PGK1t-TEF1p-BcBAPAT-ADH1t-PGK1p-AMTS3-EcYDFG-CYC1t in pHT backbone
pHT-mB-His	TEF1p-AMTS131-BcBAPAT-His-CYC1t in pHT backbone
pHT-mY-His	PGK1p-AMTS3-EcYDFG-His-CYC1t in pHT backbone

3. While the authors have provided a supplemental file for the Top 1000 artificial peptides from the RNN-VAE model, I believe the manuscript would be of broader utility (especially for further biochemical characterization of AMTS import rates/efficiencies) if the authors would also provide a list of their AMTSs that have been re-sampled to the specific organisms (i.e. an extension of Supplementary Table 2). This would allow other non-computational researchers to rapidly profile novel MTSs or detect similarity to one of their proteins of interest, without requiring coding proficiency or dedicated Linux machines to run the analysis on their own.

Response:
We agree with the reviewer that providing a list of AMTSs that have been re-sampled to the specific organisms would allow other non-computational researchers to rapidly profile novel MTSs. Therefore, we have added this information in the Supplementary Table 2 and modified the text accordingly: '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 Table 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.'

4. I am especially intrigued by the mechanisms by which these AMTSs target to mitochondria. More specifically, do these AMTSs maintain some conserved Tom20 binding mechanisms that are inherent to most endogenous N-terminal MTSs? The authors should be able to try some bioinformatics analysis and look for the presence of Tom20 binding motifs, at least in the yeast AMTSs (see Bykov et al. 2021, Muto et al, 2001; Obita et al, 2003). I believe the manuscript might benefit from a brief discussion on import mechanisms (e.g. Tom20 dependency), even if the bioinformatics approach is inconclusive.

Response:
We appreciate the suggestion. We performed a bioinformatic search to check whether the consensus motif of the Tom20-recognition element in the mitochondrial presequences, φχβφφ (where φ represents a hydrophobic residue, χ represents any amino acid residue, and β represents a basic residue), is present in our AMTSs. Based on this analysis, we found that the majority of AMTSs (514/730) possess the Tom20-recognition motif. We have added this analysis to the main text and reported the AMTSs with their corresponding motifs in Supplementary Table 1: 'We conducted a bioinformatic analysis to determine whether the consensus motif of the TOM20-recognition element in mitochondrial presequences, φχβφφ[24,31,32] (where φ represents a hydrophobic residue {L, F, I, V, W, Y, M, C, A}, χ represents any amino acid, and β 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 Table 1), indicating a likely TOM20-dependent import mechanism.'

Minor points
5. Fig 2A: "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 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."

Fig 2A might benefit from annotating the clusters, as it remains unclear to me which cluster is which.

Response:
We have updated Fig. 2a to annotate the cluster that contains MTSs without the start methionine and modified the text accordingly: '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.'

6. "Our earlier analysis indicated that 27 and 17 of the VAE-generated peptides localize GFP to the chloroplast (plastid) or both the chloroplast (plastid) and mitochondrion (Supplementary Fig. 3)"

I think it is important to clearly state in this sentence that they were *predicted* to localize GFP to the chloroplast, mitochondria, or both, not that they truly localize.

Response:
We appreciate this being brought to our attention. We have updated the text to: '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.'

7. Fig S10: "We used TargetP 2.0-predicted MTS of RtCOX4 as a control. However, we did not observe GFP localized to mitochondria." Is there any rationalization for this? COX4-MTS fusion should localize to mitochondria, so it is perplexing as to why some of the AMTSs localize fine but not COX4.

Response:
We agree with the reviewer that COX4 MTS is expected to localize proteins to the mitochondria. A likely explanation for the discrepancy is that TargetP 2.0 may not perform as well on sequences from non-model organisms, as its model is primarily trained on proteins from model species. We have previously observed a similar discrepancy for the TargetP 2.0-predicted COX4 MTS in *Issatchenkia orientalis*, another non-model yeast.

8. Fig 2B: The authors state that their AMTSs follow a "similar distribution", but are any of these changes statistically significant (i.e. for alanine)?

Response:
The statistics for amino acid fraction for MTSs in the training dataset and AMTSs are as follows:

Amino Acid	MTS Training (N = 50980)		AMTS (N = 705081)	
	Mean	STD	Mean	STD
A	0.1206	0.0739	0.1712	0.1181
C	0.0206	0.0269	0.0063	0.016
D	0.0033	0.0102	0.0003	0.0031
E	0.0048	0.0127	0.0008	0.0052
F	0.0439	0.0389	0.0272	0.0346
G	0.0465	0.0447	0.0361	0.0513
H	0.0189	0.0259	0.0035	0.0111
I	0.0349	0.0359	0.013	0.0248
K	0.0264	0.0318	0.0129	0.0275
L	0.1218	0.0548	0.179	0.0787
M	0.0431	0.0251	0.0363	0.0132
N	0.0241	0.031	0.0091	0.0216
P	0.0534	0.0411	0.0352	0.0393
Q	0.0375	0.037	0.0175	0.027
R	0.1455	0.0504	0.2277	0.0664
S	0.1131	0.0614	0.1505	0.0882
T	0.0585	0.0456	0.0287	0.0356
V	0.0545	0.0431	0.0337	0.0392
W	0.0104	0.0197	0.0024	0.0095
Y	0.0181	0.0248	0.0086	0.0166

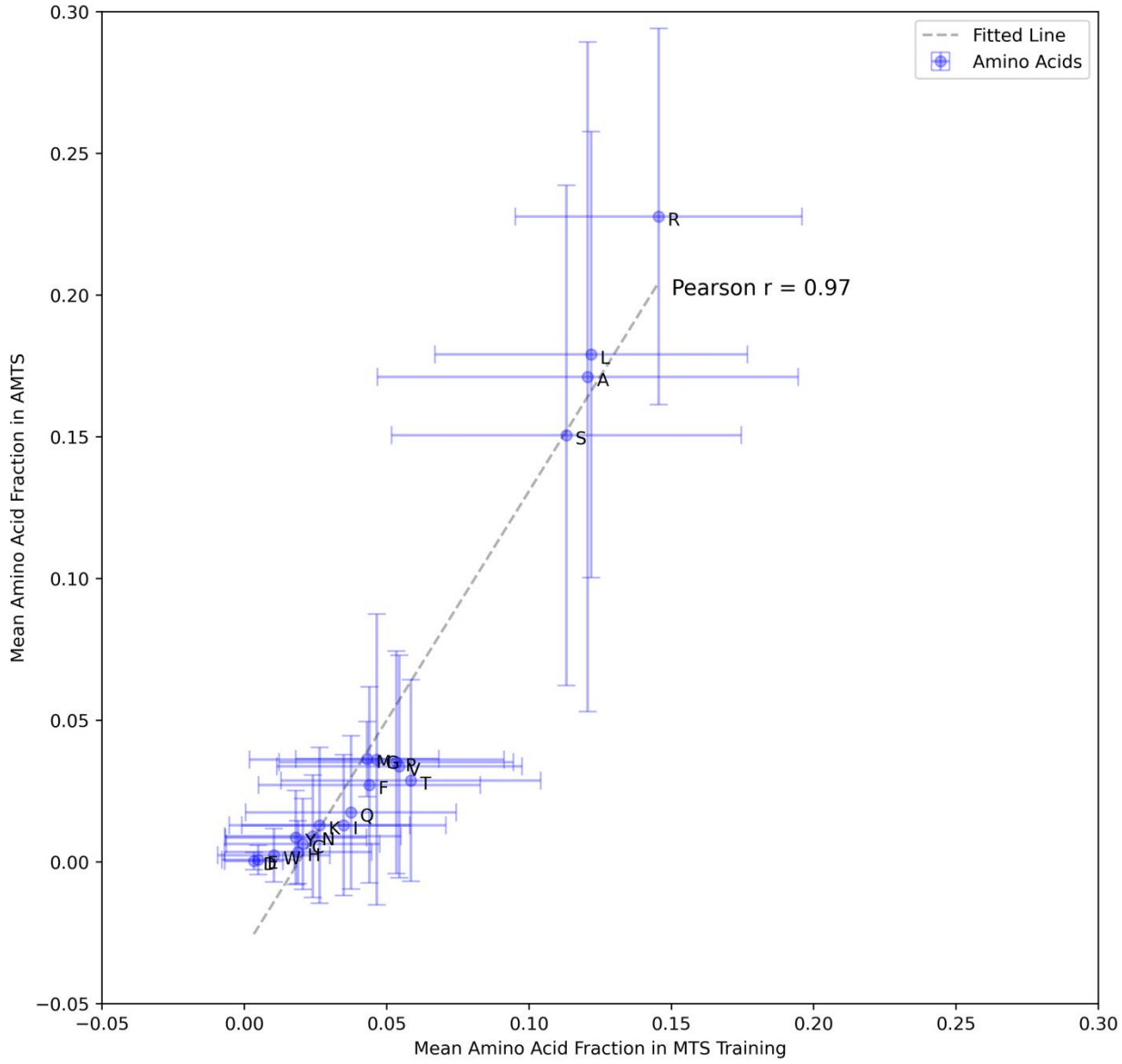


Fig. RX. Comparison of amino acid fractions between MTSs in the training dataset and AMTSs.

A high Pearson correlation of 0.97 indicates that the model effectively captures the distribution of amino acids present in the training dataset and that the generated sequences closely follow the learned amino acid composition (Fig. RX). However, as a generative model, the VAE is designed to create novel sequences rather than exact replicas of the training data. Therefore, variations in amino acid composition are not necessarily a concern; instead, they highlight the VAE's capacity to explore new regions within the sequence space. Similar statistically significant changes in amino acid composition (Fig. 4a) were observed in a previous study that used a similar approach to generate antimicrobial peptides (Das, P., Sercu, T., Wadhawan, K., et al., *Nat Biomed Eng* 5, 613–623 (2021)).

9. "For the GFP western blot, the cell lysates were normalized to ensure comparable GFP levels and were then resolved on a 14% SDS-PAGE gel." Which method of protein quantitation was used?

Response:

The goal of this assay was to probe the cleavage of VAE-generated MTSs using Western blot analysis. To perform this experiment, an initial Western blot was conducted with equal amounts of cell lysates loaded into each lane. Based on the visual intensity of the GFP bands, a second gel was run to further normalize GFP levels across lanes, followed by resolution on a 14% SDS-PAGE gel.

We have updated the text to: '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.'

10. Fig 6C: There is no mention anywhere in the text or figure legends what the negative control "ReHelper5" is.

Response:
We appreciate this being brought to our attention. We previously included the details on ReHelper5 in Supplementary Table 5. We have added more details for ReHelper5 in the main text and updated it to: '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. 6c 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.'

11. It might be useful to mention the existence of internal MTSs somewhere in the manuscript, given that there are a subset of proteins which can still target to mitochondria in the absence of a canonical N-terminal MTS (Bykov et al. 2021, Backes et al. 2018). While the authors state that the import efficiency of targeted proteins is dictated by the passenger protein (citing Van Step et al. 1986), this variability could be partially explained by the relationship between Tom70 and internal MTSs of the passenger protein.

Response:
Thank you for the suggestion. We have updated the text to discuss the role of internal MTSs in the manuscript: '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[27,32]. 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.'

12. The training dataset featured both IMM and matrix localized proteins, but for all of the outputted AMTSs, there is no further mention of submitochondrial localization throughout the manuscript. Presumably, most of the shorter AMTSs should just localize proteins to the mitochondrial matrix, but it may be interesting to mention whether any of the AMTSs generated (especially the longer ones) feature stop transfer signals or IMM sorting signals based on the training dataset. It might also be interesting to run the AMTS fusion proteins through something like DeepMito (Savojardo et al. 2019), to profile whether any of the AMTSs might alter a protein's localization between inner membrane or matrix localized, though I acknowledge that this computational approach may be flawed.

Response:
We appreciate the suggestions. Our primary objective was to design mitochondrial targeting sequences that deliver biomolecules to the matrix. Since we trained the model using shorter transit peptides (< 70 amino acids) and applied TargetP 2.0 preprocessing for IMM and matrix-localized proteins, our dataset

predominantly includes sequences without stop-transfer or IMM sorting signals, making it unlikely for the model to learn these features. However, we believe one could implement a rational design approach, utilizing our model-generated sequences with IMM sorting signals or stop-transfer signals, to achieve specific submitochondrial localization.

Following the suggestion for AMTS fusion proteins, we tested all targeting sequences from the cytoplasmic HEM1 protein delivery application, including the COX4 MTS, using DeepMito's web server (Savojardo et al., 2019), a tool for predicting protein sub-mitochondrial localization. The predictions indicate that all sequences are targeted to the mitochondrial matrix (Table R1).

Accession	GOid	GOterm	Score
COX4-cHEM1	GO:0005759	C:mitochondrial matrix	0.37
AMTS3-cHEM1	GO:0005759	C:mitochondrial matrix	0.45
AMTS131-cHEM1	GO:0005759	C:mitochondrial matrix	0.66
AMTS3-COX4-cHEM1	GO:0005759	C:mitochondrial matrix	0.49
AMTS131-COX4-cHEM1	GO:0005759	C:mitochondrial matrix	0.71
AMTS131-AMTS3-COX4-cHEM1	GO:0005759	C:mitochondrial matrix	0.6
AMTS3-AMTS131-COX4-cHEM1	GO:0005759	C:mitochondrial matrix	0.51

Table R1. Predicted submitochondrial localizations for MTS/chimeric MTS-cHEM1 constructs using DeepMito.

Note that if we replace cHEM1 with GFP, all proteins are predicted to localize to the mitochondrial inner membrane (Table R2). Since the COX4 MTS of *Saccharomyces cerevisiae* is widely used for matrix targeting, and given our confocal microscopy results for the first three constructs, we suggest that this approach may be flawed.

Accession	GOid	GOterm	Score
COX4-GFP	GO:0005743	C:mitochondrial inner membrane	0.16
AMTS3-GFP	GO:0005743	C:mitochondrial inner membrane	0.26
AMTS131-GFP	GO:0005743	C:mitochondrial inner membrane	0.34
AMTS3-COX4-GFP	GO:0005743	C:mitochondrial inner membrane	0.21
AMTS131-COX4-GFP	GO:0005743	C:mitochondrial inner membrane	0.34
AMTS131-AMTS3-COX4-GFP	GO:0005743	C:mitochondrial inner membrane	0.34
AMTS3-AMTS131-COX4-GFP	GO:0005743	C:mitochondrial inner membrane	0.22

Table R2. Predicted submitochondrial localizations for MTS/chimeric MTS-cHEM1 constructs using DeepMito.

13. In the text, one aspect of the authors' rationale for requiring AMTSs is that:

"… overuse of a single, endogenous targeting sequence may saturate the import machinery, resulting in potential competition with native proteins during the import system, and its downstream accumulation can compromise mitochondrial protein biogenesis and integrity15,16."

This claim cites two reviews, though it is unclear to me in which system or MTS the authors are referring

to. I understand the general concept is intuitive, but presumably a saturation of import machinery would be due to slow import rates (i.e. TOM/TIM passage) of a given MTS, or due to slow proteolytic processing within mitochondria. Some clarification or primary citations could be useful here, if possible. Is this saturation coming from slow proteolytic processing or slow TOM/TIM passage, or both? I understand that overexpression of suboptimal MTS-proteins, especially with stop transfer signals, could clog the TOM/TIM channels, but I am naive to issues that would arise from "overuse" of an MTS-fusion protein when the MTS occurs endogenously, yet is efficiently targeted and cleaved within mitochondria. Overall, I agree that broadening the pool of MTS sequences is incredibly useful, but it does carry the caveat of needing to characterize the import and proteolytic efficiencies of these AMTSs prior to or during their in vitro use (a process which is highlighted in the manuscript discussion and could be expedited by addressing major point #4).

Response:
We anticipate that saturation may primarily arise from slow TOM/TIM passage (Ford, Holly C., et al. *eLife* 11 (2022): e75426). We have updated the text to include this primary citation.

14. Along similar lines to my comments on whether these AMTSs would all utilize the typical Tom20-dependent import routes, it could be interesting to look at some of the species-specific AMTS sequences which are furthest in Levenshtein distance from endogenous sequences (i.e. if there are any distant AMTSs that are still predicted to be mitochondrially localized)? In this sub-analysis, perhaps the authors might uncover some atyptical AMTSs… I acknowledge that the prediction algorithms may not be able to correctly predict the localization of these AMTSs, but it could be interesting to pursue or at least list some of the most dissimilar AMTSs.

Response:
We appreciate the suggestion and agree that studying highly diverse VAE-generated MTSs would be interesting. We analyzed our species-specific AMTS sequences and report the list of distant peptides predicted to localize to mitochondria using DeepLoc 2.0, including those with and without a TOM20-recognition element (φχβφφ), in Supplementary Table 2. We have updated the text to incorporate this analysis as follows: '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.'

Syntax/Grammar

15. "However, as MTS is positioned at the N-terminus of the protein, it can also influence transcription rates and, consequently, mRNA levels " should read "as MTSs are positioned at the…" or "as an MTS is positioned".

Response:
We have updated the text to: 'However, as an MTS is positioned at the N-terminus of the protein, it can also influence transcription rates and, consequently, mRNA levels.'

Reviewer #1 (Remarks on code availability):

I was unable to install and run the code from the authors' GitHub to reproduce the raw data used in the manuscript. First, during installation of the requirements, I received a Pip error, potentially because the deeploc2 package has changed since the original upload of the repository. I also ensured that I was running Python 3.10, which makes it unclear to me where the first error is coming from.

Pip subprocess error: ERROR:
Ignored the following versions that require a different python version: 1.0.0 Requires-Python >=3.10; 1.7.1 Requires-Python >=3.10
ERROR: Could not find a version that satisfies the requirement deeploc2==1.0.0 (from versions: none)
ERROR: No matching distribution found for deeploc2==1.0.0

Response:
Apologies for the inconvenience with the scripts. Indeed, our scripts for model training, sequence generation, and analysis do not require DeepLoc 2.0. However, the provided environment.yml file had version conflicts, which caused installation errors. We have updated the environment.yml file to exclude unnecessary or conflicting packages and tested all the provided scripts. We believe everything should now function smoothly. We have added this information in Code Availability: '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.'

Second, when attempting to run the python scripts (given that most of the scripts should be able to run without deeploc2) as described in the GitHub, I received errors regarding missing .pkl files. I searched all the Python scripts for code that would generate these .pkl files, but the only logic for generating .pkl files I could find was within the Dual/analysis/split.py code, which is not specified to be run until later. Specifically, here is the file not found error generated from train.py:

(py3) MTS-VAE$ python MTS/scripts/train.py Traceback (most recent call last): File "/MTS/scripts/train.py", line 14, in <module> with open('MTS/data/tv_sim_split_train.pkl', 'rb') as f: FileNotFoundError: [Errno 2] No such file or directory: 'MTS/data/tv_sim_split_train.pkl'

Also, when trying to run the Dual/analysis/split.py code under these circumstances, it requires the 'sklearn' module, which does not seem to be included in the environment.yml dependencies.

It is possible that I may be missing something, but I believe that if I encountered these errors then the common user would also likely have difficulties in generating the data used in the paper. I would appreciate if the authors could guide me to resolve these issues, either by editing their code, providing .pkl files, or suggesting some potential fixes.

Response:
We provide the datasets curated for training the VAE models and analysis on Zenodo: https://zenodo.org/records/14590156 as described in Code Availability. We have added this information to GitHub repository: 'Datasets curated for training the VAE models and analysis are made available on Zenodo: https://zenodo.org/records/14590156.' We recommend the reviewer to run the scripts by uploading the files in MTS/data or Dual/data directories.

Finally, while it may be common sense to others within the ML field, I believe the GitHub readme should specify that their code requires Linux/Unix, given that certain dependencies from the environment.yml were incompatible with both Windows and MacOS when I attempted.

Response:

Thank you for your feedback. We have updated the information in the GitHub repository: ‘This setup was tested in a Linux environment. For smooth installation and best results, we recommend using Miniconda on Linux (https://docs.anaconda.com/miniconda/).’

If agreed upon by the editor, I would be open to the authors providing suggested code fixes prior to the full rebuttal letter, given that these updates should be rapid and would give me more time to troubleshoot and run all of the code again on my end.

Response to Reviewer #2

In their manuscript titled "Design of diverse, functional mitochondrial targeting sequences across eukaryotic organisms using variational autoencoder", the authors Aashutosh Girish Boob, Shih-I Tan, Airah Zaidi, Nilmani Singh, Xueyi Xue, Shuaizhen Zhou, Teresa A. Martin, Li-Qing Chen, and Huimin Zhao present an innovative approach using Variational Autoencoders (VAE) to design novel mitochondrial targeting sequences (MTSs). The study demonstrates the functionality and diversity of these sequences across four eukaryotic organisms and extends the design to dual-targeting sequences that work for both mitochondria and chloroplasts. This work highlights the potential of AI-driven methods in mitochondrial engineering, combining deep learning with experimental biology in an impressive way.

The authors confirm the prior knowledge of the physicochemical and structural characteristics necessary for mitochondrial import through VAE-generated sequences. One of the notable aspects of the study is the creative use of latent space interpolation to design dual-targeting peptides, a concept that adds valuable insight to understanding evolutionary design principle.

However, the study suffers from several key limitations. One major issue is the bias in the training dataset, which relies heavily on the TargetP prediction model. This bias may constrain the model's potential to discover truly novel sequences features, as the VAE could simply be learning the patterns already encoded in TargetP rather than discovering new principles. Therefore, additional controls, particularly bioinformatics-based validations, are needed to fully evaluate the novelty of the findings.

Response:
We would like to thank the reviewer for taking the time to provide valuable feedback on the manuscript.

Major Comments:
1. Training Dataset Bias: The model was trained on 56,660 peptides, with only 4,595 MTSs experimentally determined from Swiss-Prot. The risk here is that the VAE may learn latent variables closely aligned with the TargetP model. The authors need to demonstrate that the VAE-generated sequences go beyond what is already encoded in TargetP.

Response:
We thank the reviewer for highlighting this concern. The model was indeed trained on 56,660 peptides, including 4,595 unique MTSs curated from Swiss-Prot. However, it is important to note that data augmentation is widely employed in the literature to improve ML model performance and is particularly crucial here, given the notable bias we observed in MTSs from Swiss-Prot towards model organisms (Supplementary Fig. 1). Importantly, the amino acid sequences used in the training data occur naturally in mitochondrial inner membrane and matrix proteins, ensuring they represent biologically meaningful ground truth. Unlike TargetP, which is predictive in nature and limited to evaluating pre-existing sequences, the VAE is generative and capable of exploring a broader design space for MTSs. The sequence identity, along with variations in physicochemical and structural attributes, demonstrates that the VAE can design novel sequences beyond what is encoded in the training data (Figs. 1 and 2).

2. Lack of a Baseline Model: The study lacks a clear baseline model to evaluate the overall performance and novelty of the VAE approach. A comparison with simple models that enrich for physicochemical properties or a rejection sampling method against TargetP validation would help assess how much value the VAE adds.

Response:
Thank you for the suggestion. Based on the reviewer's feedback, we generated an equal number of synthetic amphipathic alpha-helical peptide sequences using modlAMP's Helices package, with lengths ranging from

10 to 70 amino acids. These sequences were prepended to the Green Fluorescent Protein (GFP) lacking the start methionine and analyzed using DeepLoc 2.0. We observed that only 1.37% of the peptides were predicted to target mitochondria, compared to 90.14% in the case of VAE-generated sequences. This analysis has been added to Fig. 1d. Additionally, we updated the text to include this simple model, which enriches for physicochemical and structural characteristics: 'We benchmarked these sequences against an equal number of MTSs designed using the profile Hidden Markov Model (pHMM)[29] and modlAMP's Helices package[30]. 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 alpha-helical peptides generated using modlAMP's Helices package were predicted to target GFP to mitochondria.'

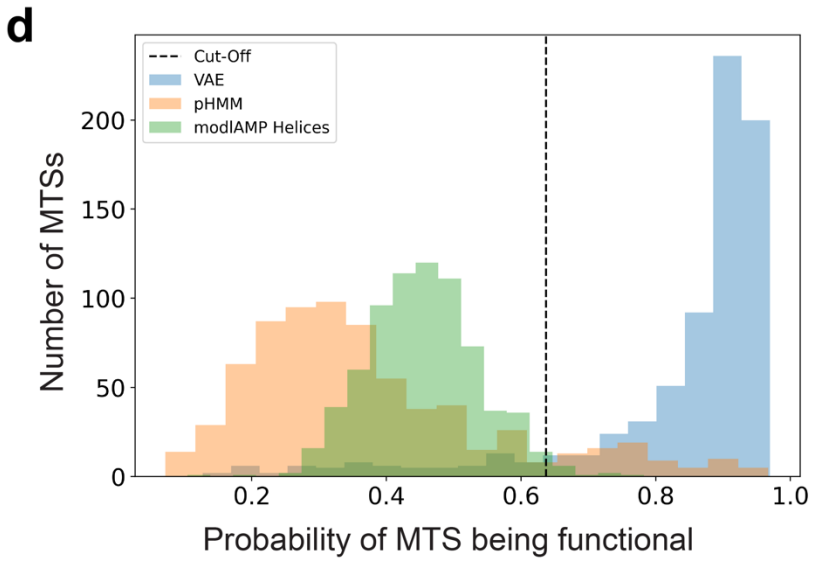


Fig 1d. Predicting functionality of generated MTSs using DeepLoc 2.0. A set of 730 VAE-, modlAMP Helices-, and pHMM-generated MTSs, appended with GFP, were analyzed for their ability to target mitochondria. Of these, 658 sequences (90.14%) generated by the VAE were deemed functional, compared to 10 sequences (1.37%) and 89 sequences (12.2%) designed using modlAMP's Helices package and pHMM, respectively.

3. Encoding Weakness: The use of one-hot encoding is a limitation, as it inadequately captures positional effects and interactions between amino acids at varying distances, rendering the approach blind to certain structural import features known to affect MTS efficacy.

Response:
We appreciate the reviewer's comment and agree that alternative encoding methods could capture positional effects and long-range amino acid interactions more effectively. However, one-hot encoding remains a widely used and validated technique in the literature for designing proteins with complex structural features, including enzymes (e.g., Hawkins-Hooker et al., *PLoS Comput Biol*, 2021; Repecka et al., *Nat Mach Intell*, 2021; Giessel et al., *Sci Rep*, 2022). MTSs, in comparison, exhibit relatively simpler secondary structures such as α-helices, β-sheets, and loops, which are well-represented by one-hot encoding. Our *in silico* analyses demonstrate that VAE-generated sequences preserve these structural attributes, and we further validated their functionality *in vivo* across four eukaryotic organisms.

Given this context, we believe one-hot encoding is a robust and suitable choice for encoding MTSs while maintaining biological relevance and generalizability. However, we appreciate the suggestion and have added text to discuss advanced encoding strategies in future studies for MTS design: 'Apart from incorporating metadata, one could utilize advanced feature engineering strategies, such as physicochemical properties[56], pre-trained large language models (LLMs)[57], or structural representations[58] to adequately capture interactions between amino acids at varying distances.'

4. Comparison with Existing Work: The authors need to compare their work with recent findings, such as those presented by Sidorczuk et al. (2023), who detailed physicochemical and structural properties of MTSs. The manuscript should clarify what novel insights the VAE approach offers beyond this.

Response:
Sidorczuk et al. (2023) introduced PlastoGram, a model for predicting subplastid localization, and provided detailed analyses of physicochemical properties such as net charge, hydrophobicity, and polarity for nuclear-encoded membrane and stromal proteins. While predictive models like TargetP 2.0, DeepLoc 2.0, and PlastoGram excel at classification tasks, they are limited to pre-existing sequences and are not suitable for scanning the vast sequence space (20^{35}) to discover mitochondrial targeting sequences (MTSs), particularly those with significant mutations (more than 10 mutations).

In contrast, our work leverages a Variational Autoencoder (VAE)-based generative model to design novel MTSs. This approach allows us to explore sequence spaces where few or no MTSs exist in the training data, offering a distinct advantage over traditional predictive models. Furthermore, we benchmarked the VAE-generated sequences *in silico* against MTSs designed using the profile Hidden Markov Model (pHMM) and modlAMP's Helices package. Our findings demonstrate that the VAE can generate sequences with higher predicted functionality (Fig. 1d) and enhanced targeting efficiency *in vivo* (Fig. 6c). This represents a significant advancement by enabling the design of MTSs with greater sequence diversity (Fig. 1e) and in previously unexplored regions of sequence space (Fig. 2a). Our analysis also provides a distribution of various physicochemical and structural attributes (Fig. 2b-d and Supplementary Fig. 4) that goes beyond those of the training dataset, offering parameters and constraints for the rational design of novel MTSs in the future. Moreover, our VAE-generated MTSs contain the consensus motif of the TOM20-recognition element, φχβφφ (where φ represents a hydrophobic residue {L, F, I, V, W, Y, M, C, A}, χ represents any amino acid, and β represents a basic residue {R, K, H}), indicating a likely TOM20-dependent import mechanism.

These findings highlight the unique advantages of our generative approach, which extends beyond the capabilities of traditional predictive models and baseline methods. Additionally, our work is the first to report an approach for designing artificial mitochondrial targeting sequences.

5. MTS sequence space: The UMAP results show the VAE model filling both high- and low-density regions in sequence space, but it is unclear if this reflects a natural distribution of MTSs. VAEs are prone to overfitting latent space, and the authors should explain how they mitigated this risk in their model.

Response:
We appreciate the reviewer's comment and agree that VAEs are prone to overfitting latent space. To mitigate this risk in our models, we implemented several strategies:

1. We trained the models with various annealing rates, applying a variable weight to the Kullback-Leibler (KL) divergence term in the loss function to ensure a balance between reconstruction loss and KL divergence.
2. We incorporated dropout layers into the model architecture, and optimized these hyperparameters to further reduce the risk of overfitting.

3. We implemented early stopping based on the validation loss to avoid overfitting during extended training epochs.
4. We performed data augmentation to increase the diversity of the training dataset, incorporating MTS sequences from all eukaryotic proteomes, thereby mitigating overfitting risks associated with limited and biased MTS dataset from Swiss-Prot.

We have previously mentioned some of these strategies in the 'Variational Autoencoder' of Methods section. For clarification, we have updated some of the text to: '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.'

Minor Comments:
1. MTS Length Restrictions: The authors restrict MTS lengths to 11-69 amino acids, but they do not justify this threshold. Given the importance of MTS length in mitochondrial targeting, a clearer explanation of this choice is necessary.

Response:
We selected 70 amino acids as the threshold for MTS length based on prior evidence in the literature (Backes, Sandra, et al. *Journal of Cell Biology* 217.4 (2018): 1369-1382.): 'Mitochondria are comprised of 800–1,500 different proteins (Mootha et al., 2003; Sickmann et al., 2003; Rhee et al., 2013; Morgenstern et al., 2017). About two thirds of these proteins are synthesized as precursors with N-terminal presequences that are both necessary and sufficient for their import (Wiedemann and Pfanner, 2017). These signals form amphipathic helices with one positively charged and one hydrophobic surface (von Heijne, 1986a). They are of variable length, typically between 8 and 70 amino acids, cleaved by the matrix processing peptidase (MPP), and degraded by the presequence peptidase PreP (Vögtle et al., 2009; Alikhani et al., 2011; Mossmann et al., 2014).'

We have added this citation to justify our choice and updated the text to: 'Subsequently, we refined the MTSs by filtering for peptides with valid amino acid sequences, restricting lengths to 11-69 amino acids[27], and eliminating duplicates.'

Overall, while the manuscript presents innovative use of deep learning for designing mitochondrial targeting sequences, it requires additional validation and baseline comparisons to ensure the novelty of its findings and address potential biases in the model's design.

Reviewer #2 (Remarks on code availability):

The availability of the code is highly appreciated, making it accessible for further use and validation. The repository includes a well-written README, which provides clear guidance, and the overall structure of the repository is organized and easy to navigate.

While containerization is not strictly necessary at this stage, implementing it would significantly enhance reproducibility across different environments. It is encouraged for future improvements to streamline deployment and ensure consistency.

Response to Reviewer #3

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response:
We express our gratitude to the reviewer for the positive comments and valuable feedback.

Response to Reviewer #4

The present manuscript describes the design of artificially generated mitochondrial targeting sequences (MTSs). These sequences are usually positively charged alpha-helical peptides comprising 10 to 120 residues. Without specific motifs, the authors must rely on physicochemical and structural properties to generate novel MTSs and employ a variational autoencoder (VAE). The authors assembled 4,595 unique sequences from SwissProt, primarily from model organisms. They expanded the dataset to a final 56,660 peptides containing between 11 and 69 residues in length to train their generative model. Each sequence was padded to a 70-long vector before being one-hot encoded. The final architecture of the generator was a VAE with fully connected layers.

The authors predicted the subcellular localisation into the mitochondrion and the sequence diversity of their VAE-generated MTSs using DeepLoc 2.0 and Levenshtein distance, respectively. They demonstrated that their generator captured the global characteristics of MTSs well by comparing the artificial sequences to naturally occurring sequences. Overall, the VAE-generated sequences shared similar amino acid composition – rich in A, I, L, and S and depleted in D, E – and physicochemical properties (net positive charge, higher hydrophobic moment, negative Eisenberg-scaled hydrophobicity). Finally, the artificial sequences were more likely to fold into alpha-helices than their natural counterparts (Fig. 2).

The authors wanted to validate their design by producing artificially generated MTSs from four biological expression systems. First, they selected nine representative candidates from the million artificial MTSs (or AMTSs), showing TargetP 2.0 scores above 0.9 for in vivo evaluation in S. cerevisiae. These sequences and the MTS of COX4 (positive control) were attached to GFP plasmids to verify their mitochondrial targeting capability. Four of the AMTSs demonstrated selective targeting, whereas three displayed target multiple subcellular locations. These results prompted the hypothesis that successful and selective targeting of the mitochondrion was taxonomically biased within the amino acid sequence of MTSs. In other words, MTSs from different organisms preferred varying amino acid residues. They confirmed their hypothesis by capturing amino-acid, physicochemical, and structural attributes of MTSs across taxa – S. cerevisiae, Homo sapiens, N. benthamiana, and R. toruloides (Figure 3a and SI Fig. 6). Consequently, they devised a taxon-specific sampling scheme over random selection using kNN algorithm. They chose 32 AMTSs (8 peptides per organism) based on the distance from the cluster centre in the four aforementioned eukaryotic systems. All eight peptides successfully targeted the human mitochondria (HEK293T cells), six out of eight for N. benthamiana and six others for R. toruloides.

The authors explored the latent space of mitochondrial targeting sequences to design sequences targeting multiple subcellular locations for dual organelle engineering – targeting mitochondrion and/or chloroplast in plants. They trained a dual-VAE from the plant-derived sequences. They observed a smooth transition between the two posterior distributions in the latent space, representing MTSs and CTSs (Fig. 5). The former sequences possessed a higher ratio of R and hydrophobic moment. In contrast, the latter includes more prolines discouraging secondary structure formation (alpha-helix or beta-sheet).

Finally, the authors demonstrated the applications of AMTSs with enhanced metabolic engineering (i.e., 3-hydroxypropionic acid) and the efficient delivery of cargo proteins to the mitochondrion.

Overall, the manuscript is well-written despite being information-dense. Their protocols are clear and detailed, and the figures are easy to understand.

Response:
We would like to extend our appreciation to the reviewer for their positive remarks and for providing valuable feedback.

Please consider the following minor comments:
(1) On page 5, line 191, can we talk about random selection when the nine VAE-generated MTSs were selected using TargetP 2.0 scores above 0.9 for in vivo evaluation in S. cerevisiae?

Response:
Our initial analysis showed that the VAE-generated MTSs exhibit meaningful physicochemical and structural features essential for mitochondrial localization. Therefore, for initial experimental validation, we used the trained VAE model to generate a pool of artificial MTSs, which we analyzed using TargetP 2.0 as an additional quality check. Most sequences in this pool had TargetP 2.0 scores above 0.9. Due to the low-throughput nature of fluorescent protein tagging and *in vivo* microscopy, we randomly selected nine VAE-generated MTSs from this pool for experimental validation.

We have updated the text to include further details on these nine selected sequences, including their TargetP 2.0 scores, in Supplementary Table 2: '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 Table 2).'

(2) The authors evaluated the mitochondrial targeting capability of nine AMTSs. On page 6, lines 200-202, they only discussed the results for seven of these sequences. What can they say about sequences 110 and 245 (Supplementary Fig. 5)?

Response:
We have updated the text to discuss the results for AMTS 110 and 245: '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.'

(3) Typos
a. Page 5, line 177 Glutamic acid "I" to "E",
b. Page 22, extended Fig. 1a and 1e legend "I4" to "C",
c. Supplementary Information page 1 "Error! Book not defined".

Response:
We appreciate this being brought to our attention. We have updated the:
a. text to '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.'
b. figure legends from 'I4' to 'C'
c. bookmark/table of content

Reviewer #4 (Remarks on code availability):

The code is available in a GitHub repository under a General Public License-3.0. The authors detailed which Python script is available for what purpose in the README file. The code is clean, with limited comments/annotations, and uses Python 3.