nature portfolio

Peer Review File

Cross-species gene redesign leveraging ortholog information and generative modeling

Corresponding Author: Professor Yasubumi Sakakibara

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)
1.The experimental validation, while elegantly executed, focuses on a single enzyme (PETase) and a single host species (B. subtilis). While this case is illustrative, the generality of the model's predictions remains underexplored. Either expand experimental validation to additional gene-host pairs (ideally with contrasting GC content or expression challenges) or better contextualize this limitation by discussing in the Discussion section why PETase was chosen and whether this design paradigm can generalize to less robust or more structurally sensitive proteins.
2.The model introduces non-synonymous mutations, which could potentially affect enzyme kinetics, stability, or interactions. Although the structural modeling (e.g., TM-score) is used as a proxy for function preservation, it is insufficient to rule out functional divergence. Include deeper in silico evaluation of biochemical properties (e.g., stability predictions, hydrophobicity shift, domain conservation) of the introduced mutations. Experimental validation of kinetic properties (e.g., kcat, Km) of PETase variants would be ideal.
3.While the model leverages ortholog data, the paper lacks discussion of whether the model is indeed learning evolutionary constraints or merely performing large-scale memorization. Are the mutations enriched in evolutionarily variable regions? Do they co-locate with natural ortholog divergence?
4.The manuscript compares OrthologTransformer to codon optimization and CodonTransformer, but other tools (e.g., ProGen, ESM-based design frameworks) may also support non-synonymous generation or function-guided editing. Clarify how OrthologTransformer differs from such models, particularly in its ability to generate ortholog-like rather than just function-retaining sequences. Are there trade-offs in computational cost (e.g., training time, hardware requirements) between methods?
5.While the core framework of OrthologTransformer is based on the Transformer architecture, it remains unclear to what extent the model has been specifically designed to address the challenges of cross-species gene redesign beyond simply applying a generic seq2seq framework to biological sequences. For instance, the use of species-conditioning tokens, the two-stage training (pretraining + species-specific finetuning), and the sequence-level editing capabilities (e.g., non-synonymous substitutions, indels) are potentially task-driven innovations. However, the manuscript does not clearly articulate the design motivations behind these choices or how they directly address the biological and functional challenges of adapting genes across evolutionarily distant hosts.
6.The AI1–AI12 naming is acceptable, but consider grouping variants into meaningful categories based on mutation type or optimization strategy, which would aid interpretation.
7.While the PET degradation assay is compelling, key functional data are missing. Only MHET (an intermediate) was quantified, not TPA (the final product). Does OrthologTransformer-designed PETase fully degrade PET, or does it stall at MHET? SEM images qualitatively show degradation, but quantitative analysis would strengthen the conclusions. Enzyme kinetics are absent. Are the activity improvements due to higher expression, better folding, or enhanced catalytic efficiency?
8.The manuscript states that datasets and code are available on GitHub, but critical details are unclear. Is the test set completely independent of training data? Potential data leakage could inflate performance metrics.
9.Can OrthologTransformer handle eukaryotic genes (e.g., intron-containing sequences)? For longer genes (>2 kb), does performance degrade due to Transformer context limits? Could AlphaFold-predicted structures guide non-synonymous mutation selection?
10.How dose the model balances sequence adaptation with functional conservation? What features (e.g., conserved domains, structural stability) are prioritized during sequence generation? Are non-synonymous mutations guided by known evolutionary principles (e.g., BLOSUM62 scores)?

(Remarks on code availability)
The current documentation does not specify key aspects of the computational environment, particularly the minimum GPU requirements (e.g., CUDA version, minimum GPU memory, GPU model compatibility). Given the model's size, users without sufficient GPU resources may encounter difficulties reproducing the results.

Reviewer #2

(Remarks to the Author)
In this study, OrthologTransformer, an AI-powered gene design tool, was developed to optimize heterologous gene expression across species. Its superior performance over conventional codon optimization was presented through adaptation of PETase in Bacillus subtilis. While the study addresses an important challenge in synthetic biology, several critical issues should be resolved.
1.Limited model innovation:
The innovative contribution contribution of this algorithmic model appears limited, as the authors have essentially adopted the standard Transformer architecture without substantial modifications tailored to the specific biological problem. The claimed 'critical innovation' of simply adding special tokens to the encoder and decoder inputs does not represent a significant advancement from the original Transformer's application in machine translation.
2.Insufficient methodological comparisons:
The manuscript would benefit from quantitative comparisons between their algorithm and traditional approaches. To properly demonstrate the model's advancement, they should conduct statistically rigorous benchmarking against conventional methods with clearly defined evaluation metrics.
3.The manuscript should include a detailed description of the tokenization strategy in the main text. In addition, justification for using AlphaFold-predicted TM-scores for structural comparison was recommended rather than establishing structural alignment algorithms (e.g., RMSD calculations from experimentally determined structures)
Specific comments on the PETase case study:
1.In the abstract section, it mentions that the OrthologTransformer-redesigned PETase showed 'robust activity surpassing codon-optimized controls'. However, this claim would benefit from quantitative validation (e.g., fold-improvement in enzymatic activity or expression levels) that would clearly demonstrate the magnitude of enhancement over traditional codon optimization methods.
2.In this study, the PETase case study only demonstrates adaptation in B. subtilis (a Gram+ bacterium). Have the authors validated the model's performance in Gram-negative hosts or eukaryotic systems (e.g., yeast), given their distinct codon usage and regulatory elements?
3.In the “PET degradation activity of AI-designed PETase“ section (lines 235-250), the authors described that “the breakdown products (terephthalic acid (TPA), mono(2-hydroxyethyl) terephthalate (MHET) and Bis(2-hydroxyethyl) Terephthalate (BHET)) were measured over time by HPLC method (Supplementary Figure S5)”. However, in Supplementary Figure S5, we only found the PET degrading pathway. Similarly, in Figure 5, the “Quantification of PET degradation products by HPLC” only showed the quantification of MHET hydrolyzed by PETase from PET. Since the HPLC methodology was used to measure TPA/BHET, the corresponding chromatograms and standard curves should be provided to validate the assay sensitivity. The TPA/ BHET data analyzed by HPLC should be provided in a supplementary panel, and discussed why MHET is the most relevant readout, which could further confirm “the AI-designed PETase is functional.”

(Remarks on code availability)

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 #5

(Remarks to the Author)
Disclaimer: Please note that this review was done as a co-review, involving an early career researcher from my group in the context of Nature Communications’ commitment to facilitating peer review training and appropriately recognizing contributions.

This paper presents a substantial contribution to synthetic gene design. The proposed tool, OrthologTransformer, aims to fill a gap left by codon optimization methods that restrict themselves to synonymous substitutions. The authors define their work as a new category of tool that we might call an ortholog designer. We agree that this is an interesting and worthwhile direction: rather than preserving the amino acid sequence at all costs, the presented model seeks to mimic the kinds of

sequence changes observed in orthologs such as non-synonymous mutations and INDELs, while maintaining protein function, and engineering a gene/protein that fits the genetic background of the anticipated target bacteria species.

In addition to presenting the method, the paper includes a strong real-world application example, making it well-rounded by combining model development with meaningful experimental validation.

Below are some questions and suggestions that may help improve the manuscript further and should be considered before publication.

Major

1) The model uses a large dataset of orthologs from species with varying evolutionary distances, but evolutionary distance is not explicitly included in the modeling. Could this be a limitation? Including information about their distance directly might help the model learn more accurately how orthologs evolve across different species.

2) Some of the most successful redesigned variants, particularly AI8, AI9, and AI12, are not fine-tuned and contain few or no non-synonymous mutations. In particular, the AI9 variant demonstrated the highest PET-degrading efficiency among tested strains, producing threefold more MHET, and causing the most pronounced PET surface erosion - while only comprising a single non-syn mutation. The section discussing the most successful variants could be expanded to include a discussion of these observations and their implications. Particularly, because you emphasize the ability of OrthologTransformer not only to model synonymous changes. However, the current results give the impression that the most powerful redesigned proteins, at least for your PETase example, are indeed those with few/no non-synonymous mutations. While we agree that it can be beneficial to also model changes beyond synonymous changes, this might question your approach, focusing also on non-syn/INDELs, and can be discussed further.

3) It is not very clear how the 12 constructs (AI1–AI12) were selected. The manuscript could clarify which parameters and combinations were considered relevant for testing. Was there a larger initial pool of candidate designs? If so, why were certain combinations excluded from experimental validation?

4) Figure 1A would benefit from a more straightforward graphical layout or explanation, as it is not immediately clear what the input and output of the model are in this figure. Along the line, the figure caption could be more elaborated to help understand the architecture.

5) In the same context, how the embedding was done is also not entirely clear to us. You mention the level of granularity of the tokens (codons), and the final embedding dimensionality (512) in lines 341-343, but not how the embedding was initialized or derived. Did you do a one-hot intermediate step or direct learnable embedding lookups? How were the embeddings initialized? Maybe you skipped describing that part because the one-hot encoding is implicitly done while embeddings are looked up; however, it's better to describe it explicitly.

6) How were the protein structures predicted in 4a? This is missing from the methods. Also, you mention mRNA folding kinetics, and according to the code, the ViennaRNA package is used. This should also be described/mentioned in the Methods.

Minor

7) This comment is more out of curiosity and not necessarily something you have to mention in the manuscript: your approach seems to have some methodological similarity with the detection of sites under negative and positive selection (using dN/dS). We wonder if the codon sites changed by the model via non-synonymous mutations would also be detected as under positive selection pressure using approaches such as CODEML or the HyPhy suite (FEL, MEME, …) on the MSAs of orthologous genes. We believe that this could have interesting applications in designing proteins under varying selective pressure constraints.

(Remarks on code availability)
We were able to install and run OrthologTransformer inside of a Linux Docker container on a Macbook. The documentation on GitHub is great, but we noticed a few things you could improve:

a) make a release version of the code so that your analyses are reproducible even when changing something in the future,
b) matplotlib is missing from the requirements.txt, and
c) inform the user that the results output folder should exist; we got an error first because of that.

Reviewer #6

(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)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
1. While the added E. coli experiment improves the manuscript, both validation systems (B. subtilis and E. coli) are prokaryotic, and both tests target a single enzyme (PETase). The authors claim that OrthologTransformer is broadly applicable to “any gene and any species,” yet this remains untested.
2. The authors report mRNA levels, protein secretion, and PET degradation activity, but these datasets are not quantitatively correlated. It remains unclear whether the enhanced activity of AI-L2 results mainly from higher expression, improved folding, or kinetic enhancement.
3. The authors show that AI-L2 (formerly AI9) has a higher kcat/Km but provide little structural rationale for this effect. Without mapping mutation sites or discussing potential impacts on the active site or enzyme stability, the functional link remains speculative.
4. The paper introduces “species-conditioning tokens,” “two-stage training,” and “MCTS optimization,” but the technical specifics are only briefly described. For example, the number and encoding method of species tokens, the weighting in the MCTS reward function, and hyperparameter details are incomplete.
5. Although the authors claim that training and test sets are strictly separated, the manuscript does not examine whether certain bacterial clades dominate the training data. Uneven representation could bias the model toward overrepresented taxa.
6. The dN/dS analysis shows that model-introduced mutations tend to occur in variable regions, which supports biological plausibility, but it does not prove that the model “learns evolutionary rules.”

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The author has revised the manuscript as requested and has addressed the reviewers' comments thoroughly. I do not have additional comments for it.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
Dear authors,

Thank you very much for this comprehensive review and for addressing all of our questions and concerns (even beyond what we expected by also integrating the quite interesting dN/dS part). We are satisfied with the current state of the manuscript and look forward to its publication.

(Remarks on code availability)

Reviewer #6

(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)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The revised manuscript is much improved, and the authors have adequately addressed the concerns raised in the previous round. The scope of the study is now clearly defined, the methodological descriptions are more transparent, and the interpretations are appropriately cautious. I have no further scientific concerns at this stage. The experimental and computational results support the current claims, and the manuscript is suitable for publication.

(Remarks on code availability)

Dear Reviewers,

We thank you for the thorough evaluation of our manuscript *"Cross-species gene redesign leveraging ortholog information and generative modeling."* We appreciate all the constructive comments from Reviewer #1 to Reviewer #6. We have carefully revised the manuscript to address every point raised. Below we provide a point-by-point response to each comment. All corresponding changes in the manuscript are highlighted in red in the revised Word document as requested.

In summary, the main revisions include the following:

1. Additional host validation (Gram-negative):
   We added a new case study in *Escherichia coli* to complement the *B. subtilis* experiments, broadening the biological context and demonstrating generality across distinct GC/regulatory backgrounds. The Results and Supplementary materials document the constructs, expression data, and PET-degradation assays for *E. coli*.
2. Rigorous computational benchmarking:
   The rigorous computational benchmarking was performed across multiple representative species pairs spanning diverse genomic or physiological characteristics, using large gene sets per pair (around several thousands of orthologous genes per pair). We conducted statistically test against conventional methods with three clearly defined evaluation metrics: codon-level sequence identity, CAI proximity and GC alignment between source, target, and model generated sequences, and report statistical significance of improvements.
3. Clarified novelty and scope:
   We emphasize that our contribution is a new problem formulation at the DNA level – orthologous gene conversion framed as a sequence-to-sequence "translation" task (distinct from protein-level generation). OrthologTransformer is trained in a supervised seq2seq manner on naturally occurring ortholog pairs, enabling it to learn the evolutionary balance between host adaptation and functional conservation. Consequently, the model does not decide this balance via an explicit objective or hand tuned weights; rather, it reproduces the patterns present in orthologs.

Response to Reviewer #1

Reviewer #1

Comment 1 (Single enzyme and host (PETase in B. subtilis) – generality of the model): *Reviewer #1 was concerned that our experimental validation focuses on one enzyme-host pair, leaving the broader applicability of the model underexplored. They suggested either adding another gene-host example (ideally in a different organism such as a eukaryote) or, at minimum, discussing in the manuscript why PETase was chosen and whether the design paradigm can generalize to other proteins.*

Response: We are actively extending our experiments to additional gene-host pairs. As suggested, we have initiated a new case study in a different organism (a Gram-negative bacterium, *E. coli*). In this supplementary experiment, OrthologTransformer was used to convert a gene from *I. sakaiensis* for expression in *E. coli*, and we observed improved expression and PET degradation activity compared to a codon-optimized control (data now included in new Supplementary Figures S8–S10 and Supplementary Table S7 in the revision). These results provide supporting evidence that our model's benefits generalize to a distinct host with different GC content and regulatory context. We have also added a brief discussion on potential application to eukaryotic systems (e.g. yeast): because our model operates on coding DNA sequences, it can in principle handle eukaryotic genes after removal of introns, but we note that eukaryotes introduce additional complexities (such as splicing and chromatin context) that would require further investigation. Moreover, extending activity assays beyond PETase would require building enzyme-specific assay systems *de novo*, which is non-trivial and time-intensive; accordingly, we designate systematic wet-lab validation for additional enzymes/genes as future work.

In addition, we have expanded our Discussion to clarify why PETase was selected and to address the potential generalizability of our approach. PETase (a robust enzyme with a clear activity assay) was chosen as an initial proof-of-concept because it allowed us to isolate the effects of our gene redesign in a controlled setting. We now explicitly discuss how OrthologTransformer is a general platform not limited to PETase: because the model was trained on a broad, multispecies ortholog dataset, it inherently learns patterns of gene adaptation that are applicable across diverse genes and hosts. This is a key novelty of our approach – by leveraging many orthologous gene pairs during training, the model captures general principles of cross-species adaptation, rather than being tailored to a single gene. We have added text to emphasize that the concept of using a Transformer-based model to perform "orthologous gene conversion" is broadly applicable, in principle enabling redesign of any gene for any target species, which has not been explored by prior methods.

Comment 2 (Non-synonymous mutations and functional effects – deeper analysis requested): *Reviewer #1 pointed out that our model introduces non-synonymous mutations, which could affect enzyme kinetics, stability, or interactions. They noted that while we used structural modeling (AlphaFold2 TM-scores) as a proxy for function preservation, this alone may be insufficient to rule out functional divergence. The reviewer requested a deeper in silico evaluation of biochemical properties (e.g., stability predictions, hydrophobicity shifts, domain conservation) of the introduced mutations, and ideally experimental validation of kinetic parameters (kcat, Km) for the PETase variants.*

Response: We appreciate this important point and have performed additional analyses to ensure that the introduced non-synonymous changes do not adversely affect protein function. In the revised manuscript, we present a more comprehensive *in silico* characterization of the mutations introduced by OrthologTransformer (in new Figure 4b and 4c). Specifically, we evaluated structure stability for each variant with amino acid substitution using protein stability prediction tools. As shown in Figure 4b and 4c, variants differ in structural conservation (TM-score, backbone RMSD, per-residue pLDDT), predicted structural stability, RNA secondary-structure free energy (MFE), and GC content. The top performer AI9 (renamed to AI-L2 in the revision) achieves an optimal balance of characteristics: the highest predicted structural stability, a TM score of 0.98 for the modelled 3D structure, a GC content of 37.0%, and a favorable mRNA secondary structure (predicted folding free energy $\Delta G \approx -281$ kcal/mol for the full-length mRNA).

With respect to experimental validation of enzyme kinetics (*k*cat and *K*m), we agree that this would be ideal. We measured *in vitro* kinetics (*k*cat and *K*m) by quantifying MHET formation over short intervals (initial-rate assay) for three representative samples, AI9 (renamed to AI-L2 in the revision), wild-type, and codon-optimized. As summarized in Supplementary Table S6, AI9 (renamed to AI-L2 in the revision) exhibited a pronounced kinetic advantage, driven by a substantial increase in *k*cat while maintaining a favorable *K*m. Consequently, its catalytic efficiency (*k*cat/*K*m) clearly exceeded that of the wild-type enzyme and the codon-optimized control tested.

We additionally examined mutation positions with respect to conserved domains and relate these observations to the dN/dS analysis presented below, as Reviewer #5 suggested. We have included this analysis in the Results (and provided details in new Table 4).

These jointly optimized multiple properties contribute in a non-linear manner for enhanced activity.

Comment 3 (Evolutionary constraints vs. memorization – is the model learning biology?): *Reviewer #1 asked whether the model is genuinely learning evolutionary constraints from ortholog data or merely performing large-scale memorization. They inquired if the mutations the model introduces are enriched in evolutionarily variable (less conserved) regions, and whether these changes co-locate with divergence observed in natural orthologs.*

Response: We have conducted additional analysis to confirm that OrthologTransformer is indeed capturing evolutionary relevant patterns rather than naively memorizing sequences. In the revision, we report that the model's substitution choices strongly correlate with known evolutionary variability. Specifically, we mapped the non-synonymous mutations proposed by OrthologTransformer onto a multiple-sequence alignment of homologous sequences (including the original source and target orthologs and related orthologs from other species) and did dN/dS-informed analysis (shown in new Table 4). We found that the majority of positions where our model introduced an amino acid change correspond to sites that are not highly conserved across the family. In other words, OrthologTransformer tends to alter residues that natural evolution also frequently varies, while leaving highly conserved positions unchanged. This suggests the model has learned which regions of the protein can tolerate change (likely reflecting functional flexibility) versus which regions must be preserved for function. We have added these findings to the Results section.

Comment 4 (Novelty relative to existing tools (codon optimization, CodonTransformer, ProGen, ESM, etc.) – clarify OrthologTransformer's unique contributions:): *Reviewer #1 noted that our manuscript compares OrthologTransformer to traditional codon optimization and to a baseline called "CodonTransformer", but pointed out that other tools (e.g., ProGen, ESM-based design frameworks) can also generate non-synonymous changes or do function-guided protein design. The reviewer requested clarification on how OrthologTransformer differs from such models, particularly in its ability to generate ortholog-like sequences rather than just function-retaining sequences. They also wondered about potential trade-offs in computational cost between our approach and others.*

Response: Our study focuses on the generation and design of DNA sequences, not protein sequences. Therefore, it is not appropriate to directly compare our method with protein-level generative models like ProGen or ESM, which operate on amino acid sequences.

Compared to such protein-level generative models, our approach tackles a different problem formulation. *ProGen* is a large language model trained on millions of protein sequences to generate entirely new protein sequences with specified functions or properties. In contrast, OrthologTransformer does not generate proteins from scratch; it works as a sequence-to-sequence (seq2seq) model that converts an existing gene sequence into a version optimized for a target species. In other words, given a coding DNA sequence from one organism, it leverages the context of that input sequence and knowledge of orthologous variants to "translate" it into a DNA sequence better suited for another organism. This seq2seq formulation (analogous to translating a sentence from one language to another) is fundamentally different from ProGen's generation approach. To our knowledge, OrthologTransformer is the first work to apply a Transformer-based seq2seq model for direct orthologous gene conversion at the DNA level, whereas prior protein large language models like ProGen operate at the amino acid level and do not perform cross-species rewriting of a given input sequence.

We have added a brief discussion of computational considerations. Training OrthologTransformer required a sizable dataset and GPU resources, comparable to training other Transformer models. However, we note that our model's size is moderate (embedding dimension 512, 20 layers) and much smaller than large protein generative models such as ProGen and ESM, making training feasible on available hardware. In inference, OrthologTransformer can generate a redesigned gene in a matter of seconds on a single GPU, which is a one-time cost per gene. By contrast, ProGen's largest models (with billions of parameters) demand more memory and are not tailored to specific input sequences (one would need to filter or fine-tune their outputs for each new gene), and traditional codon optimization tools, while extremely fast, lack the sophistication to explore non-synonymous changes. We have included these points to reassure the reviewer that our method's computational requirements are reasonable and to further underline the novel capability it delivers relative to prior art.

Comment 5 (Model design motivations – tailoring the Transformer to cross-species gene redesign): *Reviewer #1 observed that our framework is based on a standard Transformer, and asked to what extent the model has been specifically designed for the challenges of cross-species gene adaptation. They listed*

examples (species-conditioning tokens, two-stage pretraining + fine-tuning, allowing non-synonymous substitutions and indels) as potentially important innovations, but felt the manuscript did not clearly articulate the rationale behind these choices or how they address biological challenges.

Response: We appreciate this detailed prompt to clarify our design rationale. In the revised manuscript, we have added subsections in the Methods to explicitly describe our model design choices and their motivations.

- Species-conditioning tokens: We explain that incorporating species identity tokens was crucial so that the model could learn the "direction" of conversion (e.g., gene from species A to species B). This addresses the challenge that different target species have different codon biases and sequence preferences – by giving the model an indicator of the source and target context, we enable it to modulate its predictions accordingly. Without such tokens, a "single" model would struggle to generalize across diverse evolutionary distances. In the revised manuscript, we now emphasize that this choice was made to imbue the model with flexibility and specificity in generating host-tailored genes.
- Two-stage training (pretraining + fine-tuning): We now clearly state that we adopted a two-stage training regime to address data scarcity for any given species pair. During pretraining, the model sees a broad range of orthologs (many-to-many) which teaches it general principles of gene evolution (e.g., typical mutations that occur between distant bacteria, basic codon usage patterns, etc.). During fine-tuning, we specialize the model to a particular source-target species pair (one-to-one direction) so it can capture the nuances of that pair's genomic context. Thus, we justify that this design was chosen to ensure both generality and precision, directly tackling the challenge of varying evolutionary distances. This reasoning is now explicitly described in the revised Methods.
- Allowance of non-synonymous substitutions and indels: We underscore that enabling amino-acid changing edits and indels is central to the biological problem: many adaptive changes observed among natural orthologs require small non-synonymous substitutions and occasional insertions/deletions, which conventional codon-optimization (synonym-only) frameworks cannot capture. Accordingly, we designed OrthologTransformer's output space and training objective to explicitly accommodate length and amino-acid changes. Concretely, the model employs an autoregressive decoder that generates codon tokens to an end-of-sequence condition, so insertions/deletions relative to the input emerge naturally, and we train by maximizing the likelihood of true orthologs. The resulting variants in Figure 4a illustrate that the model indeed proposes edits spanning indels and non-synonymous substitutions.

  In addition, we intentionally leveraged alignments during data preparation to expose the model to explicit patterns of non-synonymous changes and indels. The Methods clarify that we prepared both gap-processed and unprocessed datasets after amino-acid alignment to balance conservation with edit visibility; this was done precisely so the model could learn where natural orthologs introduce amino-acid changes and gaps. In the conversion from *I. sakaiensis* to *B. subtilis*, alignment-based training sometimes improved conversion accuracy relative to alignment-free training (see the gray vs. black bars in Figure 2), consistent with the idea that alignments provide a clearer supervisory signal for position-specific substitutions/indels. We now make this rationale explicit in the revision.

Comment 6 (Naming of variants (AI1–AI12) – suggestion to group by mutation type or strategy): *Reviewer #1 noted that while naming the designed variants as AI1–AI12 is acceptable, it might aid interpretation to group these variants into meaningful categories (e.g., by the type of mutations they contain or the optimization strategy used), rather than treating them as an arbitrary list.*

Response: We agree that clearer categorization of the designed variants would aid interpretation. In the revision, we have adjusted how we present AI1–AI12 (renamed to AI-S1–AI-L5 in the revision) to make their differences more transparent and aligned this with the underlying design parameters.

- Moved & expanded table: The former *Supplementary Table S2* that concisely summarized each variant's design "recipe" (database/source, number of species, alignment processing, fine-tuning, MCTS, signal sequence) has been moved into the main text as new Table 3, and expanded with a brief, step-by-step narrative of how each model/design was generated under systematically varied conditions – training source (Orthofinder vs OMA), breadth of training species (23, 54, or 2,138), alignment processing (±), pair specific fine tuning (±), MCTS (Monte Carlo Tree Search Optimization) to balance GC and RNA MFE.
- Subgrouping by training breadth (major distinction): Following the reviewer's suggestion to group by strategy and given that the number of species in the training data most strongly differentiates model behavior, we now grouped and renamed variants by training breadth/source in the Results and in a revised figure:
    - Orthofinder-based (23/54 species): AI-S1 (23 sp.), AI-M1–AI-M6 (54 sp.)
    - OMA-based (2,138 species): AI-L1–AI-L5 (2,138 sp.)

      This grouping is now stated in the text and visually demarcated, so readers can immediately see how training diversity relates to downstream properties and performance.

These changes address the request by replacing an essentially alphabetical list with meaningful, strategy-aware subgroups, while keeping per-variant edit/metric summaries prominent for readers who wish to compare designs at the mutation level.

Comment 7 (Missing functional data in PET degradation assay – quantification of final product (TPA) and kinetics): *Reviewer #1 expressed concern that key functional data were missing from the PETase experiment. They observed that only MHET (an intermediate) was quantified, not TPA (the final monomeric product of PET hydrolysis), and they requested more quantitative analysis to strengthen conclusions. They also mentioned the absence of enzyme kinetics data, asking whether the observed activity improvements are due to higher expression levels, better folding, or enhanced catalytic efficiency of the enzyme.*

Response: We have addressed these points thoroughly by augmenting our experimental analysis and discussion, as Reviewer #2 also pointed out:

- We have now included data on TPA (and BHET) detection. Although the redesigned PETase did not produce a significant amount of TPA (due to the reasons explained: lacking MHETase to complete PET depolymerization), we wanted to ensure the assay's completeness. The supplementary table (Supplementary Table S5) now contains the chromatographic evidence that TPA was monitored, and we

explicitly state that TPA was not detected above background in any samples for either wild-type, codon-optimized, or AI-designed PETase. This clarifies that the improved PET degradation we report is in terms of increased MHET release and PET weight loss, not the production of end-state monomers.

- We added analyses of enzyme kinetics and factors underlying the performance gains. We measured *in vitro* kinetics (*k*cat and *K*m) by quantifying MHET formation over short intervals (initial-rate assay) for three representative samples, AI9 (renamed to AI-L2 in the revision), wild-type, and codon-optimized. As summarized in Supplementary Table S6, AI9 (renamed to AI-L2 in the revision) exhibited a pronounced kinetic advantage, driven by a substantial increase in *k*cat while maintaining a favorable *K*m. Consequently, its catalytic efficiency (*k*cat/*K*m) clearly exceeded that of the wild-type enzyme and the codon-optimized control tested. Beyond enzyme expression and kinetics factors, multiple DNA-level features contribute in a non-linear manner. As shown in Figure 4b and 4c, variants differ in structural conservation (TM-score, backbone RMSD, per-residue pLDDT), predicted structural stability, RNA secondary-structure free energy (MFE), and GC content. The top performer AI9 (renamed to AI-L2 in the revision) achieves an optimal balance of characteristics: the highest predicted structural stability, a TM score of 0.98 for the modelled 3D structure, a GC content of 37.0%, and a favorable mRNA secondary structure (predicted folding free energy $\Delta G \approx -281$ kcal/mol for the full-length mRNA). These jointly optimized properties provide a mechanistic rationale for enhanced activity.

Comment 8 (Dataset and code clarity – test set independence and potential data leakage): *Reviewer #1 noted that our manuscript states datasets and code are available, but asked for clarification on whether the test set is completely independent of the training data. They were concerned that any overlap (data leakage) could inflate performance metrics.*

Response: We apologize for not making this point clearer initially. We have updated the Methods section to explicitly describe how we constructed training and test splits to avoid any data leakage. We obtained orthologous gene pairs from the OMA database. We then partitioned ortholog groups such that no ortholog pair in the test set shares a protein with any pair in the training set. This means if a particular gene (or any genes from its species pair) was used for training, none of its orthologs or that gene itself appear in testing.

Comment 9 (Eukaryotic genes and long sequences – model limitations and potential use of structure guidance): *Reviewer #1 asked whether OrthologTransformer can handle eukaryotic genes (for example, genes with introns) and how it performs with longer genes (>2 kb) given Transformer context length limits. They also wondered if incorporating AlphaFold-predicted structures could guide non-synonymous mutation selection.*

Response: We address these points as follows in the revised manuscript:

- Eukaryotic genes (intron-containing): We clarify that our current implementation of OrthologTransformer is designed for continuous coding sequences (CDS). To apply it to a eukaryotic gene with introns, one would use the spliced mRNA/cDNA sequence (exons only) as input and output. The model itself has no fundamental limitation in handling eukaryotic coding sequences. However, we have added a note in the

Discussion that designing a gene for a eukaryotic host may require coupling OrthologTransformer's CDS optimization with separate considerations for promoter, UTRs, and possibly alternative splicing variants. We indicate that this is an area for future work, but that in principle the model could be extended to eukaryotes by including intron-less ortholog pairs in training and perhaps adding tokens for other relevant features if needed. This acknowledges the limitation while expressing that it's not an insurmountable one, just outside our current scope.

- Long genes (>2 kb) and Transformer context: In our Methods, we now specify the maximum sequence length the model was trained on. We used positional encodings up to a certain length (in our case, 700 tokens, which corresponds to ~2.1 kbp when using codon tokens, since each token is 3 bases). Thus, typical genes (which are often 1 kbp or less in bacteria) are well within the context window. We do note that extremely long genes (much beyond 3-4 kbp) could approach the limit of the model's context. In practice, we did not encounter significant degradation in performance for the longest genes in our test. We have added this information to reassure that the model can handle genes in the few-kilobase range. So, we acknowledge the context limit but indicate it did not affect our results and is a known consideration for future scaling.
- AlphaFold-guided mutation selection: We find this to be an intriguing suggestion and we say so in the Discussion. In the current work, OrthologTransformer's mutation decisions are entirely data-driven (learning from sequences alone). The reviewer's idea of incorporating structural predictions (e.g., using AlphaFold to evaluate a candidate mutation's structural impact during generation) could indeed further ensure functional conservation. We have added a forward-looking statement: *"At the modeling level, we will explore explicit evolutionary-distance conditioning (e.g., a distance token or embedding) to calibrate edit propensities across divergence scales, and we aim to integrate structure prediction or functional scoring (e.g., AlphaFold-based rescoring) into the design loop to guide non-synonymous edits – particularly for more distantly related conversions where functional preservation is most challenging."*

Comment 10 (Balancing sequence adaptation with functional conservation – what features are prioritized by the model during generation?): *Reviewer #1 asked how the model balances making changes for host adaptation with preserving protein function. In other words, what features (e.g., conserved domains, structural stability) does the model prioritize to ensure functionality is maintained while the sequence is altered?*

Response: This question goes to the heart of how OrthologTransformer operates, and we have expanded our Discussion to give insight into this "black box". While the model does not explicitly encode notions like "conserved domain" or "structural stability" as separate variables, these aspects are implicitly learned through the training on ortholog pairs. We now explain that:

- Nature-defined trade-off (ortholog supervision, not manual prioritization). OrthologTransformer is trained in a supervised seq2seq manner on *naturally occurring ortholog pairs* that already embody the balance between host adaptation and function preservation. Consequently, the model does not decide this balance via an explicit objective or hand-tuned weights; rather, it reproduces the patterns present in orthologs

(synonymous changes, conservative amino-acid substitutions, occasional indels). This point is now stated explicitly in the revised text.

- Connection to evolutionary selection (added analysis). Motivated by Reviewer #5's suggestion, we conducted a site-wise dN/dS analysis on homologs to contextualize where OrthologTransformer proposes amino-acid substitutions. Consistent with an ortholog-driven learning signal, model-edited sites were enriched in positions showing higher dN (less purifying constraint), whereas highly conserved (low-dN) sites were less changed. We have incorporated these results and their interpretation into the revised Discussion to clarify that the model's edits track evolutionary permissiveness rather than overriding functional constraints.

In summary, OrthologTransformer learns the adaptation–conservation balance from nature: its outputs preserve functional cores while adjusting DNA-level features for the target host in a manner consistent with how orthologs differ across species. The added dN/dS analysis makes this mechanism explicit in the revision.

Comment (Remarks on code availability): *Reviewer #1 asked to specify key aspects of the computational environment, particularly the minimum GPU requirements (e.g., CUDA version, minimum GPU memory, GPUmodel compatibility). They also noted that given the model's size, users without sufficient GPU resources may encounter difficulties reproducing the results..*

Response: We appreciate the reviewer's comment regarding the documentation and computational environment requirements. To address this concern, we have implemented the following updates:

- Documentation of Environment: We have revised the documentation to clearly specify the required computational environment, including the necessary CUDA version, minimum GPU memory, and compatible GPU models. This information ensures that users understand the hardware and software requirements for running our model.
- Pretrained Model Availability: The pretrained model checkpoint is now publicly available. This allows users to access the model weights directly, which is particularly helpful for those who cannot undertake full model training due to resource constraints.
- As we mentioned above, our model's size is moderate and smaller than large protein generative models such as ProGen and ESM, making training feasible on available hardware. In inference, OrthologTransformer can generate a redesigned gene in a matter of seconds on a single GPU, which is a one-time cost per gene.

Response to Reviewer #2

Reviewer #2

Comment 1 (Limited model innovation – use of standard Transformer architecture): *Reviewer #2 felt that the methodological innovation of our model appears limited, noting that we largely adopted a standard sequence-to-sequence Transformer with only minor modifications (special tokens for species). They argued that adding species tokens is not a substantial advancement over the original Transformer architecture from machine translation.*

Response: We respectfully clarify that while the core architecture is indeed a standard sequence-to-sequence Transformer, the novelty lies in (i) the new problem formulation – orthologous gene conversion at the DNA level framed as translation between species – and (ii) tailoring the model and data pipeline to that problem, rather than inventing a new neural block. In the revision, we explicitly articulate these contributions in the Introduction and Methods, and link them to empirical gains.

- Problem formulation (conceptual advance). We cast cross-species gene redesign as a seq2seq "translation" task on coding DNA, training on many ortholog pairs so the model learns evolution-consistent edits (from synonymous swaps to amino-acid changes and indels) that preserve function in the target species. To our knowledge, applying a Transformer in this DNA-level ortholog conversion setting has not been previously explored, and we now make this point explicit.
- Species-conditioning (directionality and multi-species generalization). We prepend source/target species tokens to condition the model on the conversion direction, enabling one model to serve many species pairs while respecting their distinct codon biases and sequence preferences (Figure 1). This is essential to generalize across diverse evolutionary distances without training separate models per pair; we now emphasize this rationale in the text.
- Two-stage learning (generality + precision). We combine many-to-many pretraining (to learn broad cross-species rules) with pair-specific fine-tuning (to capture nuances of a given source to target pair), as shown in Figure 1b. Figure 2 further illustrates accuracy improvements when moving from limited to broad training regimes and when applying fine-tuning, motivating this design choice in the revised Methods.
- Allowance of non-synonymous substitutions and indels (task-critical capability). Unlike synonym-only codon optimizers, our autoregressive decoder is trained to output full target coding sequences – so length changes and amino-acid changing edits arise naturally when supported by ortholog supervision. Figure 4a documents the resulting edit spectra per variant (insertions/deletions/synonymous/non-synonymous), illustrating that the model can introduce conservative amino-acid changes while preserving structure. We clarify this motivation and implementation in the Method and Results sections.
- Alignment-aware supervision to expose biologically relevant edits. To explicitly teach the model where evolution edits proteins, we prepared alignment-processed and unprocessed datasets so the model sees explicit patterns of non-synonymous changes and indels at corresponding positions. In the conversion from *I. sakaiensis* to *B. subtilis* task, alignment-based training sometimes improved conversion accuracy relative to alignment-free training (Figure 2, gray vs black bars), supporting this choice; we now state this

rationale explicitly.

- Additionally, we guided the model with multi-objective evaluation (Monte Carlo Tree Search (MCTS) routine that jointly optimizes GC content and mRNA secondary structure stability) during candidate selection. While these might not constitute modifications to the Transformer architecture itself, they are significant at the application level that were essential for success in this domain.

In sum, our contribution is not the invention of a new Transformer variant, but a new, biology-grounded modeling task and a carefully adapted training/data strategy that equips a standard Transformer to perform cross-species DNA sequence conversion – a capability that codon-only methods cannot offer. We have revised the Introduction, Methods, and Discussion sections to make these design motivations explicit and to connect them to the empirical improvements reported in Results section.

Comment 2 (Insufficient methodological comparisons – need for rigorous benchmarking): *Reviewer #2 felt that the manuscript lacked quantitative comparisons to traditional approaches. They requested more statistically rigorous benchmarking against conventional methods, with clear evaluation metrics, to convincingly demonstrate the model's advancement.*

Response: We agree that thorough benchmarking is critical. In our original submission we already reported quantitative comparisons on two sequence-identity metrics – codon-level identity and amino-acid–level similarity – across multiple representative species pairs spanning diverse genomic or physiological characteristics, using large gene sets per pair (around several thousands of orthologous genes per pair); these results are summarized in Table 1 and the associated Results text. The benchmarking was performed on rigorously constructed training and test splits to avoid any data leakage. We partitioned ortholog groups such that no ortholog pair in the test set shares a gene with any pair in the training set. In response to Reviewer #2 request, we conducted statistically rigorous test against conventional methods with three clearly defined evaluation metrics: codon-level sequence identity, CAI proximity and GC alignment between source, target, and model-generated sequences.

In the revised manuscript, we have strengthened this benchmarking:

- We conduct statistical analyses to test differences across methods, and we now report statistical significance of improvements in the Results (in a new Table 2).
- We added CAI and GC content as formal evaluation metrics alongside codon-level identity (in Table 2).

As context, our existing Results already indicate that OrthologTransformer substantially outperforms traditional synonym-only approaches: we observe ~1.7-fold average improvement (with statistical significance $p < 10^{-5}$) in codon-sequence identity over frequency-based codon optimization across species pairs, and ~1.8-fold ($p < 10^{-5}$) in the subset comparable to CodonTransformer (Table 1 and 2). In the revision, all three metrics – codon-level identity, CAI proximity, and GC-content alignment – show statistically significant gains for OrthologTransformer relative to the baselines (significance values reported in the Results and Table 2).

Comment 3 (Clarity of tokenization strategy and justification of TM-score for structural comparison): *Reviewer #2 requested to include a detailed description of our tokenization strategy in the main text. They also questioned our use of AlphaFold-predicted TM-score as the measure of structural similarity, suggesting that using structural alignment (e.g., RMSD from experimentally determined structures) might be more appropriate.*

Response: We have addressed both of these points by adding clarifications to the manuscript:

- In the Methods (OrthologTransformer Architecture), we now explicitly describe that sequences are tokenized at the codon level (each three-nucleotide codon is one token, with start/stop as special tokens). We also detail the species-conditioning tokens – e.g., a [Source_species] token at the encoder input and a [Target_species] token at the decoder input – to indicate the conversion direction. The vocabulary thus comprises 64 sense codons + start/stop + N species tokens. Each token (codon or species) is mapped to a learnable 512-dimensional embedding, initialized and trained end-to-end with the model. This section clarifies precisely how the DNA sequences and species context are represented for the Transformer.
- Use of TM-score vs. RMSD: For structural comparison we used TM-score on AlphaFold-predicted structures, as stated in the Evaluation Metrics, because experimental structures are not available for all designed variants and TM-score provides a length-normalized, fold-level similarity measure that is less sensitive to local deviations than RMSD – appropriate when comparing predicted models across sequences with small edits. In response to the reviewer's suggestion, the revised manuscript also reports RMSD and per-residue pLDDT to complement TM-score (in new Figure 4b and 4c). These additions provide both a global (TM-score) and local (RMSD/pLDDT) view of structural conservation.

Comment (Specific comments on the PETase case study): *Reviewer #2 provided several specific comments related to the PETase experiment and its presentation.*

Specific Comment 1 (Abstract claim – need quantitative validation of "surpassing codon-optimized controls): *Reviewer #2 noted that in the Abstract we stated the redesigned PETase showed "robust activity surpassing codon-optimized controls," and they suggested we should quantify this improvement (e.g., how many fold improvement) to substantiate the claim..*

Response: We have revised the Abstract to include quantitative details of the improvement. It now reads (in summary): *"As proof of concept, an OrthologTransformer-redesigned PETase expressed in Bacillus subtilis showed robust activity, producing approximately 10-fold more reaction product than the codon-optimized enzyme, and achieving higher expression levels, thereby demonstrating improved enzyme performance via AI-guided gene design."* By explicitly stating the fold-improvement (10× increase in MHET production in our PET degradation assay), we satisfy the reviewer's request.

Specific Comment 2 (Validation in Gram-negative or eukaryotic hosts): *Reviewer #2 The reviewer asked whether we have tested the model's performance in a Gram-negative bacterium or a eukaryotic system, given their distinct codon usage and regulatory elements. They noted our current experiments were only in B.*

subtilis (Gram-positive).

Response: We acknowledge this important point. As mentioned in Response to Reviewer #1, Comment 1 above, we have taken steps to demonstrate the model's applicability to other host types. In the revised manuscript, we report new experimental results for a Gram-negative scenario. We applied OrthologTransformer to redesign a gene for *E. coli* expression (Gram-negative host) and compared it with a traditional codon-optimized version. Preliminary results indicate that the OrthologTransformer-designed gene yields higher protein expression and PET degradation activity than the codon-optimized gene. For a eukaryotic system, we have not yet completed a wet-lab validation due to the complexity of eukaryotic expression systems. Designing for a eukaryotic host will likely require coupling CDS optimization with separate consideration of promoters, UTR elements, and potential alternative splicing. This remains an area for future work.

Specific Comment 3 (HPLC data for PETase assay – provide chromatograms for TPA/BHET and discuss MHET focus): *Reviewer #2 pointed out that in our Methods we mentioned measuring multiple PET degradation products (TPA, MHET, BHET) by HPLC (Supplementary Figure S5), but the materials provided (Figure 5 and Supplementary) seemed to show only MHET quantification and a pathway diagram. They requested that we include the actual HPLC chromatograms, standard curves for product detection, and explain why only MHET was shown as the readout. Essentially, they want assurance that the assay was sensitive to TPA/BHET and a rationale for focusing on MHET.*

Response: We have updated the Supplementary information to include the requested data and have expanded the discussion of the degradation assay in the main text, as also mentioned in Response to Reviewer #1, Comment 7. Specifically, we added Supplementary Table S5, which shows example HPLC chromatograms for MHET, TPA, and BHET standards and for sample supernatants from our PETase experiments. These additions demonstrate the sensitivity of our HPLC method for all three products.

In the Results section, we now clarify why MHET was the primary quantified product. In our experimental setup (PET film degradation by PETase in *B. subtilis* culture supernatants), we consistently detected MHET accumulation, but TPA and BHET remained below detectable level. We interpret this as follows: MHET is the mono-hydrolysis product of PET (one ester bond cleaved), whereas TPA would result from complete degradation (both ester bonds cleaved). *B. subtilis* does not naturally have a MHET hydrolase to further break down MHET into TPA, so MHET accumulates. BHET is an intermediate di-ester that is less commonly produced except under certain conditions. We explain that, given the absence of additional enzymes, the reaction stalls at MHET, making it the most relevant readout of PETase activity in our system. We have added a note that measuring MHET is sufficient to gauge PETase performance, since any TPA produced would indicate an even further extent of reaction. Importantly, we now explicitly state that no significant TPA was detected in our assay, confirming that our redesigned PETase, like wild-type PETase, primarily generates MHET.

We also mention in the Discussion that future work could incorporate a downstream MHETase to fully convert MHET to TPA, allowing a closed-loop plastic degradation system.

Response to Reviewer #5

Reviewer #5

Major Comments:

Comment 1 (Evolutionary distance not explicitly modeled – potential limitation): *Reviewer #5 pointed out that our model uses a large dataset of orthologs from species with varying evolutionary distances, but we did not explicitly include evolutionary distance as an input or feature. They ask if not modeling evolutionary distance could be a limitation, and suggest that including such information might help the model learn how orthologs evolve across different divergence levels.*

Response: This is a thoughtful observation. In our current model, evolutionary distance is indeed not explicitly given to the model (we do not, for example, feed a numeric distance or category). Instead, the model must infer the "distance" indirectly from the sequences themselves and the species tokens. We agree that explicitly incorporating a measure of divergence could potentially enhance performance by contextually biasing the model's propensity to make changes (e.g., a large evolutionary distance might warrant more non-synonymous substitutions). We have added a sentence in the Discussion as future work: *"At the modeling level, we will explore explicit evolutionary-distance conditioning (e.g., a distance token or embedding) to calibrate edit propensities across divergence scales, and we aim to integrate structure prediction or functional scoring (e.g., AlphaFold-based rescoring) into the design loop to guide non-synonymous edits – particularly for more distantly related conversions where functional preservation is most challenging."*

Comment 2 (Most successful variants had few non-synonymous mutations – discuss implications): *Reviewer #5 observed that some of the best-performing redesigned PETase variants (notably AI9 and AI12) contained few or no non-synonymous mutations, even though our approach emphasizes the ability to introduce such mutations. They note that AI9, which had only a single amino acid change, was the top performer (3× more MHET production, most PET surface erosion). This raises the question: does the success of low non-synonymous-change variants undermine the value of modeling non-synonymous/INDEL changes? They ask us to discuss why the most powerful redesigned proteins in this case were those with minimal amino acid changes, and what that means for our approach that also explores non-synonymous changes.*

Response: The reviewer brings up an interesting paradox: if the best designs ended up being close to the original protein sequence, one might wonder if our allowance for non-synonymous mutations was necessary. We address this by analyzing why these variants succeeded and how we interpret the outcome. We appreciate the observation and agree that the PETase case highlights an important point: enhanced activity in our system did not require extensive amino-acid changes. The result for AI9 (renamed AI-L2 in the revision) does not undermine the value of enabling non-synonymous substitutions and indels. Trained on natural ortholog pairs, OrthologTransformer learns when evolution typically leaves protein cores unchanged and when amino-acid adaptations occur. As noted in the Discussion, for the S290N substitution introduced in AI-L2, two prior studies on protein engineering reported a change at the same residue (S290P) that enhanced PETase stability and activity, supporting the view that the model introduces amino-acid/indel edits where beneficial. Moreover, across species pairs, our benchmarking shows that sequences generated by OrthologTransformer (which may

include non-synonymous substitutions) more closely resemble target-species orthologs and improve functional similarity compared with synonym-only approaches; this includes gains in codon-sequence identity and BLOSUM-based amino-acid similarity, indicating conservative, biologically informed non-synonymous substitutions when advantageous.

Comment 3 (Clarity on selection of the 12 final designs (AI1–AI12)): *Reviewer #5 felt it was not very clear how the 12 PETase variants were selected. They ask us to clarify what parameters or combinations were considered and whether there was a larger initial pool of candidates. If so, why were certain designs chosen or others excluded from experimental validation?*

Response: We agree that the selection pipeline for the 12 PETase variants was not fully explained. In the revision, we have clarified the process in the Methods/Results and aligned the presentation with a similar request in Reviewer #1, Comment 6. Concretely:

- Moved and expanded table: The former *Supplementary Table S2* (design "recipes" for AI1–AI12 (renamed to AI-S1–AI-L5 in the revision)) has been moved into the main text (as a new Table 3) and expanded with a brief step-by-step narrative of how the models/designs were generated under systematically varied conditions – training source (Orthofinder vs OMA), breadth of training species (23, 54, or 2,138), alignment processing (±), pair-specific fine-tuning (±), MCTS (Monte Carlo Tree Search Optimization) (±) to balance GC and RNA MFE – and that AI-S1–AI-L5 were selected as representative outputs spanning these axes.

Comment 4 (Figure 1A clarity – input/output depiction): *Reviewer #5 suggested that Figure 1A (the schematic of the OrthologTransformer model) could be graphically improved for clarity. They found it not immediately clear what the input and output of the model are from that figure, and they recommended elaborating the figure caption.*

Response: We have revised Figure 1a in line with this feedback. The new version of the figure has a more straightforward layout: we explicitly label the input sequence (with a notation like "Gene from Species A") and the output sequence ("Orthologous gene for Species B").

The figure caption has been expanded accordingly. It now clearly states: *"Figure 1a: Input: a coding DNA sequence from Species A (source), prepended with a source_species token s_{tgt}, is encoded. Output: the decoder, conditioned by a target_species token s_{tgt}, generates an orthologous coding sequence for Species B, permitting synonymous and conservative non-synonymous substitutions and indels where supported by ortholog supervision. The model features a 20-layer encoder-decoder structure, with each layer equipped with Add & Normalization layers and Multi-head Attention mechanisms. Species tokens (s_{src}, s_{tgt}) are prepended to the input sequence, enabling species-specific sequence conversion… etc."* This caption now serves to reinforce what the input/output are and how the model operates on them.

Comment 5 (Embedding details not entirely clear): *Reviewer #5 noted that while we mention using codon-level tokens and an embedding dimension of 512 (lines 341-343 in the manuscript), we did not explain*

how the embedding was initialized or derived (one-hot vs learned, etc.). They asked for an explicit description of the embedding process.

Response: In the revised Methods (Model Architecture section), we have added a sentence to clarify the embedding. We now state: *"Tokenization and embeddings: Input/output sequences are tokenized at codon resolution (three nucleotides per token). The vocabulary comprises 64 sense codons + start/stop + N species tokens, and each token is mapped to a 512-dimensional learnable embedding. Embeddings are randomly initialized and optimized end-to-end during training. The context length is 700 tokens, which corresponds to ~2.1 kbp when using codon tokens, since each token is 3 bases."*

Comment 6 (Methods details missing – structure prediction and RNA folding): *Reviewer #5 asked how the protein structures were predicted for Figure 4a, noting that this was missing from the Methods. They also noticed that we mention mRNA folding kinetics and the ViennaRNA package in our code, but did not describe this in the Methods.*

Response: We have corrected these omissions by updating the Methods section accordingly:

- Protein structure prediction: We now explicitly state that we used AlphaFold3 (ver.3.0.1) to predict the 3D structures of the redesigned PETase variants, as well as the wild-type and codon-optimized PETase, for structural comparisons. We outline that we used the AlphaFold3 default pipeline with no templates (since we wanted a prediction purely from sequence) and ran it for each sequence. We mention that we took the top-ranked predicted structure for each and calculated TM-scores between each variant's structure and the wild-type PETase structure.
- mRNA structure and ViennaRNA: We now mention in the Methods that we evaluated mRNA secondary structure for each designed sequence using the ViennaRNA package ver.2.6.4 (RNAfold). We specify that we computed the minimum free energy (MFE) structure for the full mRNA sequence and recorded the MFE value as a measure of mRNA folding strength. This was done to compare whether our designs had potentially more favorable (less stable, thus more easily unwound) mRNA structures than the originals.

Minor Comment:

Comment 7 (Relation to dN/dS (selection analysis)): *Reviewer #5 made a comment out of curiosity (and noted it's not necessarily for the manuscript) about a methodological similarity: our approach reminds them of how one detects sites under positive/negative selection using dN/dS ratios. They wondered if the codon sites changed by our model correlate with sites that would be identified as under selection.*

Response: We find this insight very interesting and have decided to address it, even if not required, as it adds an extra perspective to our work. We have included this dN/dS analysis in the Results (and provided details in new Table 4). In the Discussion (likely where we talk about how the model chooses mutations, related to Reviewer#1's comment 3 and 10), we include a short remark connecting to dN/dS analysis:

We say that indeed, OrthologTransformer's behavior is analogous in some ways to a dN/dS-informed approach – positions that the model changes tend to be those that evolution (across species) has changed (i.e., higher dN), whereas conserved positions (low dN) are left unchanged by the model. We actually performed a

site-wise dN/dS analysis on alignments of homologous sequences using the FEL method. For each site, the nonsynonymous (dN) and synonymous (dS) substitution rates were estimated; sites showing significant evidence of purifying selection ($p < 0.01$ and $dN/dS < 1$) were classified as "purifying", whereas sites with $dN/dS \geq 1$ or without significant purifying signal were classified as "non-purifying". We found that non-purifying sites exhibited significantly higher substitution rates than purifying sites (Fisher's exact test, $p < 10^{-14}$). These findings indicate that OrthologTransformer preferentially introduces substitutions at evolutionarily variable sites, while changes at highly conserved positions are comparatively suppressed. We mention this observation to the reviewer to satisfy their curiosity.

Comment (Remarks on code availability): *Reviewer #5 made the following remarks: We were able to install and run OrthologTransformer inside a Linux Docker container on a MacBook. The documentation on GitHub is great, but we noticed a few things you could improve: (a) make a release version of the code so that your analyses are reproducible even when changing something in the future; (b) matplotlib is missing from requirements.txt; and (c) inform the user that the results output folder should exist; we got an error at first because of that.*

Response: Thank you very much for your valuable suggestions. We have addressed all three points as follows:

a) Release version for reproducibility:

To ensure reproducibility of the analyses, we have created a fixed release version (v1.0.0) on GitHub. This allows users to reproduce the results even if the codebase evolves in the future.

b) Added matplotlib to requirements.txt:

As suggested, we have added matplotlib==3.9.1 to the requirements.txt.

c) Improved handling and documentation of the results output folder:

We have modified the code so that the results output folder is automatically created if it does not already exist. Additionally, we updated the README to clearly inform users about this behavior to avoid confusion or errors during the first run.

We believe that the revisions made have addressed all concerns raised by Reviewers #1 – #6. The manuscript has been significantly improved in terms of clarity, completeness of results, and adherence to the journal's requirements.

Once again, we thank the reviewers and the editor for their valuable feedback, which has helped us improve our work. We hope that our revisions satisfactorily address all comments and that the updated manuscript is now suitable for publication in *Nature Communications*.

Best regards,
Yasubumi Sakakibara

Dear Reviewers,

We appreciate the reviewers' further comments on our manuscript entitled "Cross-species gene redesign leveraging ortholog information and generative modeling." We have carefully revised the manuscript to address every point raised. Below, we provide a point-by-point response to each comment. All corresponding changes in the manuscript are highlighted in red in the revised Word document.

Response to Reviewer #1

Comment 1: *Reviewer #1 was concerned that while the added E. coli experiment improves the manuscript, both validation systems (B. subtilis and E. coli) are prokaryotic, and both tests target a single enzyme (PETase). The authors claim that OrthologTransformer is broadly applicable to "any gene and any species," yet this remains untested.*

Response: We fully recognise that Comment 1 exposed an important issue with the way we described the scope of OrthologTransformer. We do acknowledge that the phrase "any gene and any species" was an overstatement, which may have led to unrealistic expectations, especially concerning eukaryotic applications. Please note that in the manuscript itself, we did not claim universal applicability; instead, we focus on bacterial orthologs. We therefore no longer use that phrase and similar universal language in the response letter and clarified in the Discussion that our current study focuses on bacterial orthologs, and that the model was trained and validated on prokaryotic genes. In addition, we note that our updated computational benchmarking – the other main component of our work – already demonstrates superior performance, with clear improvements over existing codon-optimization and deep learning method. In this second-round revision, we conducted additional, more extensive computational benchmarking, and the results are presented in a new Figure 3. This new computational benchmarking result extends the performance comparison to the maximum possible scale, covering all 45 bacterial species included in both the OMA database (2,138 species) and the set of species that the existing deep-learning method CodonTransformer can handle (164 species), as well as 450 source-to-target species combinations (10 source species per target species).

Comment 2: *Reviewer #1 was concerned that the authors report mRNA levels, protein secretion, and PET degradation activity, but these datasets are not quantitatively correlated. It remains unclear whether the enhanced activity of AI-L2 results mainly from higher expression, improved folding, or kinetic enhancement.*

Response: We agree that multiple interdependent factors underlie the enhanced activity of the AI-designed enzyme (AI-L2). In our previous revision, we emphasized that no single molecular property fully explains the performance gain. Rather, several jointly optimized features – including codon usage adaptation, improved mRNA stability/structure, and possibly beneficial amino acid substitutions, which can affect mRNA expression, protein folding, and enzyme kinetics – are likely to act together in a synergistic, non-linear manner to yield the higher PETase activity in *B. subtilis*. We have added a brief explanation to the Result highlighting that PETase performance *in vivo* is governed by multiple interacting, non-linear factors – including mRNA

abundance/stability/structure, translation efficiency/codon adaptation, and beneficial amino acid substitutions, which in turn can affect mRNA expression, protein folding, and enzyme kinetics – rather than by any single determinant.

Comment 3: *Reviewer #1 pointed out that AI-L2 (formerly AI9) has a higher kcat/Km but provide little structural rationale for this effect. Without mapping mutation sites or discussing potential impacts on the active site or enzyme stability, the functional link remains speculative.*

Response: We did include an analysis of a key mutation in the previous manuscript, and we apologize if it was not sufficiently highlighted. In the Discussion, we explicitly discussed the substitution S290N in AI-L2 as a potential structural/functional contributor. To quote the relevant sentence from the previous manuscript: *"Regarding the mutation S290N introduced in AI-L2, two earlier studies have reported a substitution at the same residue position, specifically S290P. Serine (S) and asparagine (N) are both neutral amino acids with polar side chains, whereas proline (P) is characterized by a nonpolar side chain. Thus, the AI-L2 coding sequence included only a handful of amino acid substitutions that the model likely inferred from related Bacillus enzyme orthologs."* This discussion (now clearly visible in the revised Discussion section) provides a structural rationale: in addition to prior studies showing that introducing mutations at the serine residue at position 290 contributes to enhanced PETase enzymatic activity and thermostability, the S290N change is a safe substitution relative to the wild-type serine (maintaining polarity), unlike the previously attempted S290P substitution which introduces a nonpolar residue.

Comment 4: *Reviewer #1 noted that the paper introduces "species-conditioning tokens," "two-stage training," and "MCTS optimization," but the technical specifics are only briefly described. For example, the number and encoding method of species tokens, the weighting in the MCTS reward function, and hyperparameter details are incomplete.*

Response: We have augmented the Methods and Supplementary Information to address these technical points in detail. First, the concept of species-conditioning tokens was already explained in the Materials and Methods ("Tokenization and embeddings" subsection) in the previous manuscript. We clarified there that the model's vocabulary includes N special tokens corresponding to each training species, which are prepended to the input and output sequences to condition the Transformer on source and target species. (For clarity, we have now explicitly noted that *N is the number of distinct species used in training and test*.) Second, we have added a more technical description of our two-stage training strategy in the Methods. This new text outlines how we first pretrain the model on a broad many-to-many orthologous gene pairs (any combination of source–target species) to capture general patterns, and then fine-tune it on a specific source–target species pair to incorporate pair-specific nuances. Third, we expanded the Methods (and Supplementary Methods) to describe the MCTS optimization procedure. We detail how our Monte Carlo Tree Search algorithm was used to refine the generated sequences, including the exact multi-objective reward function that balances GC content and mRNA secondary structure stability (melting energy). The weighting scheme used in the MCTS reward function and the corresponding hyperparameter details are now provided in the Supplementary Methods.

We believe these additions fully address the reviewer's request for more technical details, and they should allow readers to understand and, if desired, reproduce our approach.

Comment 5: *Reviewer #1 was concerned that although the authors claim that training and test sets are strictly separated, the manuscript does not examine whether certain bacterial clades dominate the training data. Uneven representation could bias the model toward overrepresented taxa.*
Response: We apologize for the confusion and have clarified the partitioning scheme in the Methods. Our data partition was done at the gene (ortholog group) level, not at the species level. In practice, this means that if a particular gene (and its orthologs) was used in training, none of those genes or their orthologous counterparts appear in the test set. There is no overlap of orthologous groups between training and testing. We explicitly stated in the previous manuscript that no ortholog pair in the test set shares any gene with the training set. This strategy ensures that no systematic bias in the representation of bacterial clades is introduced in the training data by our partitioning strategy.

Comment 6: *Reviewer #1 noted that the dN/dS analysis shows that model-introduced mutations tend to occur in variable regions, which supports biological plausibility, but it does not prove that the model "learns evolutionary rules."*
Response: We appreciate this concern and have revised the wording to ensure objectivity. We would like to clarify that our manuscript never claimed that the model "learns evolutionary rules" (indeed, that phrase does not appear in our paper or prior response). To prevent any misunderstanding, we have rephrased some related expressions in the revised manuscript. For example, instead of saying the model captures "general principles of protein adaptation," we now state that it captures patterns of gene adaptation observed in nature. Similarly, we replaced the term "evolution's imprint" with a more neutral description of observed ortholog variation. These edits avoid overreaching statements and simply report that the model reproduces the kinds of substitutions seen in naturally evolved orthologs (as supported by our dN/dS analysis).

Response to Reviewer #2

Comment: *The author has revised the manuscript as requested and has addressed the reviewers' comments thoroughly. I do not have additional comments for it.*
Response: We sincerely appreciate Reviewer #2's careful evaluation and thoughtful feedback.

Response to Reviewer #5

Comment: *Thank you very much for this comprehensive review and for addressing all of our questions and concerns (even beyond what we expected by also integrating the quite interesting dN/dS part). We are satisfied with the current state of the manuscript and look forward to its publication.*
Response: We sincerely thank Reviewer #5 for their careful evaluation and positive recommendation.

We believe that the revisions made have addressed all concerns raised by Reviewers #1. We hope that our revisions satisfactorily address all comments and that the updated manuscript is now suitable for publication in *Nature Communications*.

Best regards,
Yasubumi Sakakibara

Dear Reviewers,

We appreciate the reviewers' further comments on our manuscript entitled "Cross-species gene redesign leveraging ortholog information and generative modeling." Below, we provide a point-by-point response to each comment.

Response to Reviewer #1

Comment: (Remarks to the Author) *The revised manuscript is much improved, and the authors have adequately addressed the concerns raised in the previous round. The scope of the study is now clearly defined, the methodological descriptions are more transparent, and the interpretations are appropriately cautious. I have no further scientific concerns at this stage. The experimental and computational results support the current claims, and the manuscript is suitable for publication.*

Response: We sincerely thank Reviewer #1 for their careful evaluation and positive recommendation.

We hope that the updated manuscript is now suitable for publication in *Nature Communications*.

Best regards,
Yasubumi Sakakibara