nature portfolio

Peer Review File

Deep Learning of 777K Bulk Transcriptomes Reveals Human–Mouse Gene Conservation Beyond DNA Sequence Similarity

Corresponding Author: Dr Emily Oates

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)
Summary:

The author has conducted a comprehensive cross-species comparative study between human and mice transcriptomic (bulk RNA-seq) data using deep learning method. While I see the power of deep learning for data integration purpose, I have some major concerns on the contribution of this paper to the field and the insights provided from their findings other than presenting the overall results with arbitrary selection cutoffs. Below are my comments:

Comments:
1. The whole study highly relies on the shared representation gene embedding obtained from the GeneRAIN model with mixed training the authors have developed before (on bioarxiv, probably under peer review), then the authors tried to further align and examine these gene embedding to orthologous genes (e.g. out of how many genes examined, how many in the nearest embedding neighbors). This approach is very post-hoc, 1. there is no standard on how well the gene embedding capture the inter-species gene relationship; 2. how stable is the gene embedding obtained subject to the parameter tuning in the algorithm; 3. the embedding is based on the bulk RNA-seq samples, which could differ greatly among tissues, conditions, etc. so I am concerned with whether this overall gene embedding will be useful.

2. The authors then explored a number of unsupervised and supervised embedding alignment methods. The assessment revealed unsupervised method had limited efficacy while supervised method performed better (based on number of orthologs ranked among top embeddings), then correlation between these two approaches indicate high similarity orthologs and strong correlation is observed between shared embedding and supervised approach, concluding there is no overfitting of supervised approach. Again there is no standard on concluding whether overfitting exists or not and it is very hard to understand what the authors wish to conclude from pairwise comparison in this 3-way method comparison.

3. It is not clear what the positive and negative control experiments are evaluating here. What do you mean by “The experiment indicated that it was not due to methodology limitations that fewer than 5,000 of 16,983 mouse genes were the nearest embedding neighbors of their human orthologs”? Same for negative control, it seems to be more of sensitivity analysis (using data splitting) of the method but then there still no real control in a sense that you can really compare to something closer to the ground truth.

4. The authors then conducted DNA/RNA similarity analysis on the learnt shared gene representation from expression data, and also considered phenotype associations. But then, other than showing the characteristics of their gene representations (similarity score, ranking) with a number of arbitrary cutoffs, there is no biological insights into what you can say about the cross-species similarity/dissimilarity or relationship, also in terms of functions, phenotype, disease, etc. The mere examples are lacking rigorous systematic statistical comparison other than comparing the ranking (again arbitrary cutoffs, why rank171, rank>=21 considered low similarity, etc.).

Minor:

1. Line 261: The main analysis is based on the GeneRAIN model the authors have developed before, so there is no deep learning method “novel” developed here.

Reviewer #2

(Remarks to the Author)
This manuscript describes a powerful deep learning framework for comparing human and mouse genes at the RNA expression level, and presents evidence to suggest that expression adds additional information compared to DNA similarity alone and perhaps explain why some use of mouse as a model to study human gene function or disease fail.

This is a very interesting and highly relevant manuscript that is conceptually very strong. It is however not the easiest to read and I save a few questions/notes about methodology and language

While the ARCHS4 data is uniformly processed and managed it is not clear how metadata problems in the experiments submitted to public archives are addressed. Are experiments with poor or missing metadata excluded from ARCHS4 analysis, does this create and known biases?

"The gene representations were 200-number vector embeddings from the model." Not very accessible language - add a little explanation - like this section "Here, a ‘space’ is a mathematical framework where similar numerical values in the gene representations represent similar biological attributes." Later on

"Throughout this text, we will use the terms 'gene embedding' and 'gene representation' interchangeably." - best to state that they represent the same concept here and use one throughout

The use of orthologs through the manuscript is confusing - is this always referring to the 1:1 ortholog list from MGI or is their any additional assignment from the model - in either case was there any additional follow up of genes that really didn't look like orthgologs at the DNA or RNA level by checking against lists from other homology methods eg HCOP - https://www.genenames.org/tools/hcop/.

Which classes of genes were low DNA sim but high RNA? Known rapidly evolving genes or genes under positive selection? Genes of particular functional classes? Expression pattern It would be good to investigate these results further to see whether eg certain GO classes are enriched (immune function?), certain tissues are enriched (brain?).

While the majority of the manuscript focuses on protein-coding genes, lncRNAs and pseudogenes are in scope. I think this methods could be particularly interesting for lncRNAs where there is often little functional information and conservation of expression with mouse could be valuable. However, pseudogenes could be problematic as they aren't really genes in the same way as the other classed but are rather genomic regions with homology to protein-coding genes reflecting their provenance. Any expression information has to be very carefully disentangled from the expression of their parent genes and it is not clear that these additional efforts have been applied here (unless the model learns them?). As such it might be worthwhile either adding more analysis of the pseudogene component or reducing/removing it.

Reviewer #3

(Remarks to the Author)
In this manuscript from Su and co-authors they utilize the ARCH4 database of gene expression to train a deep learning model that maps homologous genes from mouse and human based on gene expression and MGI homologue data.

My background is mainly as a genomic/bioinformatic researcher, interested, but not experienced in AI and deep learning models. My main criticism thus focus on the accessibility of the manuscript to researchers less interested in deep learning but more interested in communicating the ideas in typical biological language. Based on that criteria, I find the manuscript interesting, but difficult to follow and confusing. Much of the language is used very loosely and is unclearly defined. A good example is the RNA similarity measure, which is not described in the methods and it is not clear what this is or how it differs from the DNA similarity measure.

Overall, whilst potentially interesting, I am confused by the improvements made by this method over already existing methods, and I found the manuscript unclear and difficult to understand.

Main comments:

1. I find the descriptions of the approaches in Figure 1 confusing. The text indicates that the MGI 1-to-1 mapping is used in the Fig 1b approach, but the model does not seem to reflect that, whilst the Figure 1c model implies using the 1-to-1 MGI gene lists.

2. How critical is the MGI 1-to-1 gene mapping? Can the training work without using this data set? (I guess this is the model in Figure 1a? I’m a little unclear).

3. How does a model using the MGI 1-to-1 gene maps perform (i.e. the naive model?)

4. The RNA similarities seem like it should have a corresponding figure. I am also confused by the difference between DNA similarity and RNA similarity. DNA similarity includes promoters and coding sequences, which seems clear, but RNA similarity is 'similarity to the corresponding human gene'. Which is unclear. Is this only coding sequence? Includes UTRs? Amino acid sequence (which would seem a better measure of similarity, rather than more divergent RNA/DNA sequence). The language used to describe these measures of similarity is very loose and imprecise.

5. Continuing on, on line 172 'DNA embedding similarity'. What is this referring to? Inconsistent use of terms makes this paragraph virtually unintelligible. Could this section be rewritten to use language more typical, such as % matching identity, BLAST or CLUSTAL similarity? Or something like that.

6. Line 183-184. As 'DNA similarity' includes the coding sequence, then what is RNA information anyway? Again, more loose descriptions and what appears to be swapping of terminology makes this section unintelligible. The failure to define DNA and RNA similarity in the preceding section means the import of lines 189-207 is unclear.

7. Line 227: But the model doesn't rely solely on transcriptome data as it includes MGI 1-to-1 orthologue mapping.

8. RNA similairty is not defined in the methods. I am unclear how this was measured. Is it a measure of conservation?

9. LECIF scores were averaged for all bins in the genie region. Is this fair? Surely it would be better to average only for coding sequences?

10. The LncRNA part is underdeveloped. Considering the low levels of conservation at DNA/RNA levels this part is potentially very interesting. I guess it's virtually impossible to validate though as phenotypes for lncRNAs are so rare.

Minor comments:

1. The Figure 1 labels and text are out of order, which gets confusing.

2. The use of abbreviations like 'RNA/DNA simi' is not needed, it's just deleting 6 letters and needlessly increasing reading difficulty.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Reviewer #3

(Remarks to the Author)
The manuscript is much improved and clearer. I have just a remaining comment:

For the 'RAN similarity', why not remove ambiguity entirely and rename it expression embedding (my preference), cosine embedding, transcriptome embedding, or just embedding? From the reply to my comment 4, it seems the RNA similarity has little to do with RNA and is a gene expression embedding.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
Okay then. It's a minor point, and if the authors insist. I still feel it impairs understanding for no reason, but is fine if explained thoroughly.

Reviewer #1 (Remarks to the Author):

Comments:

1. The whole study highly relies on the shared representation gene embedding obtained from the GeneRAIN model with mixed training the authors have developed before (on bioarxiv, probably under peer review), then the authors tried to further align and examine these gene embedding to orthologous genes (e.g. out of how many genes examined, how many in the nearest embedding neighbors). This approach is very post-hoc, 1. there is no standard on how well the gene embedding capture the inter-species gene relationship; 2. how stable is the gene embedding obtained subject to the parameter tuning in the algorithm; 3. the embedding is based on the bulk RNA-seq samples, which could differ greatly among tissues, conditions, etc. so I am concerned with whether this overall gene embedding will be useful.

Response: We thank the reviewer for the insightful comments and careful evaluation of our work. The reviewer raises several important points regarding the validation and robustness of our novel cross-species gene embedding approach. These comments provide a good opportunity for us to clarify the methodological evaluations we performed to ensure our approach is sound, and we have revised the manuscript to make these validations more prominent.

1. Lack of standardized evaluation frameworks: The reviewer notes that our study relies on a novel gene representation method and that establishing its validity is crucial. We agree that evaluating how well a new embedding technique captures inter-species gene relationships is a significant challenge, especially in the absence of a single, universally accepted gold standard.

On the lack of a standard for evaluation: Although the GeneRAIN model underlying our approach has been published in Genome Biology (DOI : 10.1186/s13059-025-03749-6), which provides additional validation of the core methodology, we agree this is still a significant challenge for our novel embedding based methodology. To build confidence in our approach, we performed three different types of validation:

Internal Controls: We conducted 'positive control' and 'negative control' experiments to test the technical limits and specificity of our alignment methodology. The 'positive control' experiment demonstrated high reconstruction rate of inter-'species' relationships (99.7%), suggesting our method's high effectiveness. The 'negative control' showed a negligible false positive rate (0.3%), suggesting that the model does not arbitrarily force alignments. We also acknowledge that our positive control experiment, while demonstrating the method's technical capability, may present a simplified alignment task compared to the complex biological reality of cross-species comparison, and thus may overestimate the method's performance in the real-world application.

External Benchmarking: We compared our model's performance against LECIF, an established method for assessing human-mouse gene conservation. As shown in Fig.

3g-i, our RNA-based similarities were more predictive of shared phenotype and disease associations, demonstrating the effectiveness of our approach.

Biological Evidence of Relevance: Most importantly, we evaluated our embeddings against real-world biological data. The strong correlation between our RNA similarity metric and shared phenotype/disease associations from multiple independent databases (Fig. 3a-c) provides supportive evidence that the embeddings capture functionally relevant information beyond what is available from DNA sequence alone.

Although we performed these supporting evidence, we agree that they can still have limitations, and readers should be aware of this. We have added the following text to the **Discussion** section (Line 339):

"*We acknowledge that our approach introduces a novel gene embedding methodology for cross-species comparison, and standardized evaluation frameworks for such embeddings remain an active area of development in the field, and there is still no established absolute ground truth for cross-species gene relationships. Our positive and negative control experiments, while provided helpful information for methodological performance, may simplify the alignment problem and thus provide an optimistic estimate of alignment accuracy. Therefore, while our validation against phenotype associations provides biological evidence for the embeddings' utility, future studies should continue to develop and refine evaluation frameworks for cross-species gene representations.*"

2. Stability of gene embeddings under parameter variation: We agree that it is important to consider the stability of our embeddings with respect to hyperparameter tuning. In our supplementary information, we detail several experiments we conducted to ensure our results were not sensitive to specific parameter choices. We demonstrated that the alignment performance plateaued and remained stable across a range of training epochs (Supplementary Fig. 4f), different numbers of orthologous gene pairs used for supervision (Supplementary Fig. 4e), and even different Transformer architectures (GPT vs. BERT, Supplementary Fig. 5a, b). These analyses suggest that our findings are not driven by narrow or arbitrary parameter choices but reflect stable properties of the embeddings.

At the same time, we acknowledge that deep learning methods can in general be sensitive to hyperparameter selection. To make this explicit, we added the following to the **Discussion** (Line 348):

"While our supplementary analyses demonstrate robustness across several parameter choices, readers should be aware that deep learning approaches can be sensitive to hyperparameter selection, and validation of embedding quality remains an active area of methodological development in the field."

Utility of embeddings from bulk RNA-seq: The reviewer raises a good point regarding the use of bulk RNA-seq data, which aggregates signals across many tissues and conditions. We agree that this generates an 'overall' or context-averaged

gene representation. We believe the strength of our study lies in leveraging a massive dataset (777K samples) to ensure this 'overall' representation is robust and captures the most fundamental expression patterns of a gene. The fact that this averaged embedding so effectively predicts shared phenotypes and outperforms a multi-omics method like LECIF suggests that it captures a powerful, biologically meaningful signal. We have updated our Discussion section to explicitly acknowledge this aspect:

Updated paragraph starts with (Line 385) "*An important consideration is that our RNA similarities reflect averages across diverse tissues and biological conditions, which may limit…".*

We thank the reviewer again for pushing us to better articulate the methodological considerations of our approach. By making these points clearer in the manuscript, we believe the paper is strengthened.

2. The authors then explored a number of unsupervised and supervised embedding alignment methods. The assessment revealed unsupervised method had limited efficacy while supervised method performed better (based on number of orthologs ranked among top embeddings), then correlation between these two approaches indicate high similarity orthologs and strong correlation is observed between shared embedding and supervised approach, concluding there is no overfitting of supervised approach. Again there is no standard on concluding whether overfitting exists or not and it is very hard to understand what the authors wish to conclude from pairwise comparison in this 3-way method comparison.

Response: We thank the reviewer for this helpful comment, which highlights a need for greater clarity in our rationale for comparing the different alignment methods. We agree with the reviewer that the purpose of this 3-way comparison and our conclusion about overfitting could be explained more clearly. Our primary goal was not to simply select the 'best' method, but rather to use these pairwise comparisons to demonstrate the **robustness** and **validity** of our chosen approach. We have revised the manuscript to make this logic explicit.

Regarding **the purpose of the 3-way comparison:** We compared the supervised, unsupervised, and shared embedding methods to show that our key findings are not an artifact of one particular technique. The correlation (**Spearman rho = 0.74**) between the supervised alignment and the shared embedding approach suggests that different methods of leveraging ortholog information converge on similar results. This increases confidence that we are measuring a true biological signal.

We agree that this is not clearly explained, to make this reasoning more clear, we have added a clarifying sentence (Line 152) to the manuscript:

“This multi-method comparison was designed to demonstrate the robustness of our findings across different alignment techniques rather than to select a single optimal method.”

Regarding the potential overfitting issue, the reviewer points out that there is no standard metric to prove the absence of overfitting in this context. Our approach is based on using the unsupervised method as an unbiased baseline. Since the unsupervised method uses zero ortholog information for alignment, it should not overfit to the supervision pairs. The high correlation between its results and our supervised method's results (Spearman rho = 0.89) suggest low risk of overfitting. It suggests that our supervised method is not merely "memorizing" the training pairs but is instead identifying a genuine geometric transformation in the embedding space that is independently corroborated by the unsupervised approach. Despite this, we agree that some overfitting may still occur in the supervised approach, and we have added a clarifying statement in the text (Results, Line 144): *"The high agreement with the unsupervised approach indicates low risk of overfitting in the supervised approach. However, it should be noted that some degree of overfitting in supervised approaches cannot be entirely ruled out given the absence of standardized overfitting metrics for embedding alignment tasks*".

3. It is not clear what the positive and negative control experiments are evaluating here. What do you mean by "The experiment indicated that it was not due to methodology limitations that fewer than 5,000 of 16,983 mouse genes were the nearest embedding neighbors of their human orthologs"? Same for negative control, it seems to be more of sensitivity analysis (using data splitting) of the method but then there still no real control in a sense that you can really compare to something closer to the ground truth.

Response: We thank the reviewer for this feedback. We agree that their purpose was not immediately obvious, and we have now revised the Results section to explicitly state the rationale and interpretation of these crucial validation steps.

Regarding the positive control experiment, it was designed to answer an important question of "is our alignment method capable of achieving near-perfect accuracy if a true one-to-one relationship exists in the data?" In other words, is the performance we see with the real human-mouse data a reflection of biological divergence, or is it simply the maximum performance our algorithm can achieve (a methodological ceiling)?

By creating two "pseudo-species" from a single human dataset, we established a synthetic "perfect" ground truth where a direct one-to-one mapping is known to exist for all genes. Our method's ability to reconstruct this relationship with 99.7% accuracy suggest that it is highly effective. This suggests that the reason we observe a lower alignment rate for actual human-mouse orthologs reflects biological divergence at the RNA level between the species, not a failure or limitation of our methodology.

Regarding the negative control experiment, it was designed to test the opposite hypothesis: Does our supervised alignment method spuriously force an alignment between genes just because they are provided as supervision pairs, even if their underlying representations are completely unrelated?

By randomizing the gene embeddings before alignment, we intentionally destroyed any underlying structural similarity. The model's complete failure to align the orthologs in this scenario (0.3% accuracy) suggests that it is not an arbitrary mapping function. This result suggests that the success of the supervised alignment is dependent on the genuine, pre-existing similarity of the embeddings learned during the mixed-training phase.

Again, we acknowledge that our positive control experiment may have simplified alignment task compared to the complex biological reality thus may overestimate the method's real performance. We have added following text to the Discussion section for this limitation (Line 342): "*Our positive and negative control experiments, while provided helpful information for methodological performance, may simplify the alignment problem and thus provide an optimistic estimate of alignment accuracy*".

Thanks for pointing out the confusion caused by sentence "The experiment indicated that it was not due to methodology limitations that fewer than 5,000 of 16,983 mouse genes were the nearest embedding neighbors of their human orthologs", we have revised it to make it clearer. The revised text of relevant part in Results (Line 158) now reads:

"*To evaluate whether our alignment method has inherent technical limitations, we performed a 'positive control' experiment by creating artificial 'species' from split human data with predetermined gene correspondence. The 99.7% reconstruction accuracy suggested our method's high effectiveness and indicated that moderate alignment rates in real human-mouse data reflect biological divergence rather than methodological limitations. To test alignment specificity, we performed a 'negative control' experiment by randomizing embedding assignments before alignment, which resulted in only 0.3% spurious alignments, suggesting our method does not arbitrarily force gene relationships.*"

Regarding the lack of a "ground truth": We agree with the reviewer that the absence of a perfect biological ground truth for functional similarity makes evaluation challenging. This is an important limitation for any method in this domain. We have now explicitly acknowledged this in the **Discussion** section (Line 339).

"*We acknowledge that our approach introduces a novel gene embedding methodology for cross-species comparison, and standardized evaluation frameworks for such embeddings remain an active area of development in the field, and there is still no established absolute ground truth for cross-species gene relationships*."

4. The authors then conducted DNA/RNA similarity analysis on the learnt shared gene representation from expression data, and also considered phenotype associations. But then, other than showing the characteristics of their gene representations (similarity score, ranking) with a number of arbitrary cutoffs, there is no biological insights into what you can say about the cross-species similarity/dissimilarity or

relationship, also in terms of functions, phenotype, disease, etc. The mere examples are lacking rigorous systematic statistical comparison other than comparing the ranking (again arbitrary cutoffs, why rank171, rank>=21 considered low similarity, etc.).

Response: We thank the reviewer for this important feedback regarding the biological interpretation and statistical rigor of our analysis. The reviewer raises valid concerns about our analytical approach and the need for more evaluation and analyses.

We defined the high similarity groups as the top 40%, while the low similarity group as the bottom 40%. We chose this approach to create two clearly distinct and non-overlapping cohorts (high vs. low similarity) for a robust comparison. By excluding the middle 20%, which contains genes with intermediate similarity, we reduced potential ambiguity and were able to more clearly highlight the differences between the high and low groups.

Regarding the specific ranks, e.g., rank 171 for SOD1, we apologize for the confusion. The classification of a gene as "low similarity" is based on it falling within our pre-defined bottom 40% category (i.e., having a rank ≥ 21).

We have added a clarification to the Results section (Line 296), "*similarity rank 171; rank ≥ 21 was considered as low similarity*" has been revised to "*similarity rank 171; rank ≥ 21 was considered as low similarity, corresponding to the 40% cutoff defined previously*".

Regarding the choice of the 40% cutoff, we agree with the reviewer that this threshold alone may not provide sufficiently robust evidence or could potentially bias the results. To address this concern, we performed additional analyses using alternative cutoffs of 50% and top/bottom 30%. Similar patterns (**Supplementary Fig. 9**) were observed with these new cutoffs compared to the original 40%, suggesting that our findings are robust to the choice of cutoff.

We also agree with the reviewer that it is important to provide biological insights from our cross-species gene relationship finding. So, we conducted additional analysis to investigate the biological association. The new results below are added (Line 269):

"*To characterize the functional properties of genes with different patterns of cross-species conservation, we performed geneset enrichment analysis on four distinct groups based on RNA embedding similarity and DNA sequence similarity (Supplementary Table 5). Genes with high RNA embedding and DNA sequence similarity were enriched for neuronal cell types, transcriptional regulation, and cancer pathways, representing core conserved biology. Genes with high RNA embedding similarity but low DNA sequence similarity showed functional convergence in immune processes, being enriched for immune cell types, autoimmune diseases, and cytokine signaling despite sequence divergence. Genes with high DNA sequence similarity but low RNA embedding similarity are enriched for basic transcriptional machinery*

(transcription initiation, ubiquitination, histone acetylation), however they showed no significant disease associations in either DisGeNET or PhenGenI databases, indicating potential functional divergence despite sequence conservation. Genes with low similarity in both dimensions showed minimal enrichment overall, with only sparse associations with specialized metabolic processes and species-specific functions, suggesting maximal divergence between species. These results suggest that RNA embedding similarity provides distinct and complementary information to DNA sequence similarity for understanding cross-species gene relationships and their functional potential."

We thank the reviewer for pointing out that mere examples are insufficient and that a systematic analysis of function, phenotype, and disease associations is required. We apologize if this was not sufficiently highlighted, as this systematic analysis is the central finding presented in our manuscript.

Our primary biological insight is demonstrated not by the individual gene examples, but by the quantitative analysis shown in **Fig. 3**. In this section, we move beyond rankings to systematically test the hypothesis that our RNA similarity metric correlates with shared biological function across all homologous genes. Our analyses, conducted across three independent and comprehensive phenotype/disease databases (MGI Phenotype Annotations, MGI Mouse Models of Human Diseases, and the Alliance of Genome Resources), all show a clear and consistent statistical trend: homologous gene pairs with higher RNA similarity are significantly more likely to share phenotype and disease associations.

The specific genes like SOD1 and PSEN2 were mentioned not as the primary evidence, but as well-known case studies to illustrate and contextualize our systematic findings. They serve to show how our global, data-driven results can offer a potential explanation for specific, long-observed discrepancies in mouse model research.

We have add the text (Line 289) to make it clear that the role of these examples as illustrative, rather than evidentiary: "*To illustrate how our systematic findings might relate to known cases in the literature, we examined several well-documented examples of mouse-human study discrepancies."*

Minor:

1. Line 261: The main analysis is based on the GeneRAIN model the authors have developed before, so there is no deep learning method "novel" developed here.

Response: We thank the reviewer for pointing this out. We agree that the novelty lies in the application rather than the development of new deep learning methods, as we used our previously developed GeneRAIN model.

We have revised the sentence from "we developed a novel deep learning approach " to "we applied a deep learning approach" (Line 324).

Reviewer #2 (Remarks to the Author):

This manuscript describes a powerful deep learning framework for comparing human and mouse genes at the RNA expression level, and presents evidence to suggest that expression adds additional information compared to DNA similarity alone and perhaps explain why some use of mouse as a model to study human gene function or disease fail. This is a very interesting and highly relevant manuscript that is conceptually very strong. It is however not the easiest to read and I save a few questions/notes about methodology and language.

While the ARCHS4 data is uniformly processed and managed it is not clear how metadata problems in the experiments submitted to public archives are addressed. Are experiments with poor or missing metadata excluded from ARCHS4 analysis, does this create and known biases?

Response: Thank you for raising this important point.

We agree that metadata quality in public repositories can be variable and represents a legitimate concern for large-scale analyses. The ARCHS4 database, which processes data from GEO/SRA, employs several quality control measures but inherits the inherent limitations of the source metadata. Also, as detailed in our GeneRAIN publication (Su *et al.*, *Genome Biology*; DOI : 10.1186/s13059-025-03749-6), samples with poor technical quality are filtered out based on criteria including total gene expression read counts exceeding one million, a minimum of 2000 genes with non-zero expression, and single-cell probability thresholds. However, these technical filters do not address all potential metadata inconsistencies.

To address this concern and provide transparency about potential biases, we have added the following clarification to the Discussion section (Line 380): "*Our analysis relies on bulk RNA-seq data from the ARCHS4 database, which inherits metadata limitations from public repositories. While technical quality filters were applied, potential inconsistencies in sample annotations or experimental conditions across studies may introduce biases that could affect cross-species comparisons, particularly for condition-specific gene functions.*"

Additionally, we note that our GeneRAIN model training approach helps mitigate some metadata-related biases through several mechanisms: (1) the model employs self-supervised learning that relies solely on gene expression patterns without using any phenotype labels or sample metadata during training, learning relationships directly from the co-expression structure in the data; (2) our large sample size (777K samples) provides robustness against individual study biases; and (3) our validation against multiple independent phenotype databases (MGI, Alliance of Genome Resources) using experimentally-derived annotations helps confirm that our embeddings capture biologically meaningful relationships despite potential metadata inconsistencies in the training data.

We hope this addresses reviewer’s concern about the metadata of ARCHS4 dataset.

"The gene representations were 200-number vector embeddings from the model." Not very accessible language - add a little explanation - like this section "Here, a ‘space’ is a mathematical framework where similar numerical values in the gene representations represent similar biological attributes." Later on

Response: We thank the reviewer for this suggestion to improve the clarity and accessibility of the manuscript. We have revised the sentence in the 'Representing genes across species using RNA information' section (Line 81) to provide a clearer explanation of gene embeddings as follows:

“The model produced a 200-dimensional vector embedding for each gene, a numerical representation that captures functional characteristics, where genes with similar biological roles have similar vector values.”

"Throughout this text, we will use the terms 'gene embedding' and 'gene representation' interchangeably." - best to state that they represent the same concept here and use one throughout

Response: We agree that using a single term will prevent potential confusion. We have chosen to use the more precise term 'gene embedding' (or 'embedding') throughout the manuscript to describe the learned vector representations and have removed the original sentence stating their interchangeable use. This change has been implemented consistently across the entire text.

The use of orthologs through the manuscript is confusing - is this always referring to the 1:1 ortholog list from MGI or is their any additional assignment from the model - in either case was there any additional follow up of genes that really didn't look like orthgologs at the DNA or RNA level by checking against lists from other homology methods eg HCOP - https://www.genenames.org/tools/hcop/.

Response: Thank you for this question about ortholog assignment, which allows us to clarify an important methodological aspect of our study.

To address the first part of the question: Yes, throughout our manuscript, "orthologs" always refers exclusively to the 1:1 ortholog list from the MGI database. Our model provides embedding similarity measurements but never defines orthology relationships - it only quantifies RNA-level similarities between genes that have already been established as orthologs by MGI.

Following your valuable suggestion to validate our ortholog assignments using HCOP, we performed this additional analysis. We found that our four gene groups show strong concordance rates against HCOP: high RNA embedding similarity/high DNA sequence similarity (95.95% validated), high RNA embedding similarity/low DNA sequence similarity (94.13% validated), low RNA embedding similarity/high DNA sequence similarity (95.60% validated), and low RNA embedding similarity/low DNA sequence similarity (86.81% validated). The high concordance rates across three

groups suggests the robustness of the MGI ortholog assignments used in our study. The lower concordance for the low RNA embedding similarity/low DNA sequence similarity group is expected, as these represent more divergent homologs, which may be inconsistently annotated across homology databases.

We have added the following text to the “Gene annotations” Methods section (Line 701): "All ortholog relationships in this study were derived from the MGI 1:1 ortholog database, with our model providing only embedding similarity measurements without defining orthology. Validation against the HCOP database confirmed 86.81% concordance in the low RNA embedding similarity/low DNA sequence similarity group and >94% concordance in all other groups, supporting the reliability of the MGI ortholog assignments used."

We hope this clarification addresses the reviewer's concern about ortholog assignment methodology and demonstrates the validation of our approach.

Which classes of genes were low DNA sim but high RNA? Known rapidly evolving genes or genes under positive selection? Genes of particular functional classes? Expression pattern It would be good to investigate these results further to see whether eg certain GO classes are enriched (immune function?), certain tissues are enriched (brain?).

Response: Thank you for this insightful comment that allows us to better characterize the functional properties of genes with different conservation patterns. This is an important point that helps readers understand the biological significance of our findings.

We have conducted gene set enrichment analysis to investigate the functional classes of genes with low DNA sequence similarity but high RNA embedding similarity, as well as the other conservation pattern groups. To characterize the functional properties of genes with different patterns of cross-species conservation, we performed geneset enrichment analysis on four distinct groups based on RNA embedding similarity and DNA sequence similarity (Supplementary Table 5). Genes with high RNA embedding and DNA sequence similarity were enriched for neuronal cell types, transcriptional regulation, and cancer pathways, representing core conserved biology. Genes with high RNA embedding similarity but low DNA sequence similarity showed functional convergence in immune processes, being enriched for immune cell types, autoimmune diseases, and cytokine signaling despite sequence divergence. Genes with high DNA sequence similarity but low RNA embedding similarity are enriched for basic transcriptional machinery (transcription initiation, ubiquitination, histone acetylation), however they showed no significant disease associations in either DisGeNET or PhenGenI databases, indicating potential functional divergence despite sequence conservation. Genes with low similarity in both dimensions showed minimal enrichment overall, with only sparse associations with specialized metabolic processes and species-specific functions, suggesting maximal divergence between species. These results suggest that RNA embedding similarity provides distinct and

complementary information to DNA sequence similarity for understanding cross-species gene relationships and their functional potential.

We have added this analysis to the Results section (Line 269) and included the detailed enrichment results in **Table S5**.

While the majority of the manuscript focuses on protein-coding genes, lncRNAs and pseudogenes are in scope. I think this methods could be particularly interesting for lncRNAs where there is often little functional information and conservation of expression with mouse could be valuable. However, pseudogenes could be problematic as they aren't really genes in the same way as the other classed but are rather genomic regions with homology to protein-coding genes reflecting their provenance. Any expression information has to be very carefully disentangled from the expression of their parent genes and it is not clear that these additional efforts have been applied here (unless the model learns them?). As such it might be worthwhile either adding more analysis of the pseudogene component or reducing/removing it.

Response: Thank you for this valuable comment that helps us improve the focus and clarity of our analysis. We appreciate the reviewer's recognition of the potential value of our approach for lncRNAs, where conservation of expression patterns could indeed provide important functional insights given the limited available functional information for these genes.

We agree that the pseudogene component requires more careful consideration. Pseudogenes present unique challenges as genomic regions with homology to protein-coding genes, and that expression signal disentanglement from parent genes is a critical concern. Following the reviewer's suggestion to reduce the emphasis on pseudogenes, we have simplified our analysis. Instead of dividing them into five subclasses, we have now combined all pseudogenes into a single group for the cross-species similarity comparison (now presented in revised Fig. 4b), to mitigatesthe risk of over-interpreting fine-grained differences between pseudogene types.

We have also added a limitation statement in the Discussion section acknowledging that pseudogene expression measurements may be influenced by cross-mapping from their parent protein-coding genes, and that future studies should incorporate specialized methods for pseudogene expression quantification to better assess their true conservation patterns.

The new text reads (Line 361):

“Our analysis reveals particularly interesting patterns for lncRNAs, which showed lower overall conservation compared to protein-coding genes but still demonstrated measurable cross-species RNA embedding similarities. This finding is significant because lncRNAs often lack comprehensive functional annotations, making our RNA-based conservation metric a valuable resource for prioritizing lncRNA homologs for experimental validation. While phenotypic validation remains challenging due to the

scarcity of well-characterized lncRNA phenotype associations, the quantitative conservation scores we provide offer researchers a data-driven approach to identify promising lncRNA candidates for cross-species functional studies. A limitation of our pseudogene analysis is that expression measurements may be influenced by cross-mapping reads from parent protein-coding genes, and future studies should incorporate specialized pseudogene expression quantification methods to better assess their true conservation patterns."

This revision strengthens our manuscript by focusing on the most robust findings while being transparent about methodological limitations.

Reviewer #3 (Remarks to the Author):

In this manuscript from Su and co-authors they utilize the ARCH4 database of gene expression to train a deep learning model that maps homologous genes from mouse and human based on gene expression and MGI homologue data.

My background is mainly as a genomic/bioinformatic researcher, interested, but not experienced in AI and deep learning models. My main criticism thus focus on the accessibility of the manuscript to researchers less interested in deep learning but more interested in communicating the ideas in typical biological language. Based on that criteria, I find the manuscript interesting, but difficult to follow and confusing. Much of the language is used very loosely and is unclearly defined. A good example is the RNA similarity measure, which is not described in the methods and it is not clear what this is or how it differs from the DNA similarity measure.

Overall, whilst potentially interesting, I am confused by the improvements made by this method over already existing methods, and I found the manuscript unclear and difficult to understand.

Main comments:

1. I find the descriptions of the approaches in Figure 1 confusing. The text indicates that the MGI 1-to-1 mapping is used in the Fig 1b approach, but the model does not seem to reflect that, whilst the Figure 1c model implies using the 1-to-1 MGI gene lists.

Response: We thank the reviewer for this insightful comment, which has helped us identify a significant source of confusion in our schematic. We agree that the depiction of the shared embedding strategy in Figure 1b could misleadingly imply that the orthologous genes are not being linked, as their positions (determined by sample-specific expression ranking) do not align.

To address this, we have revised the Figure 1 legend to explicitly clarify this crucial point. The new text reads (Line 574):

“b, Schematic of mixed training plus shared embedding strategy. Please note that the position of each gene in the input sequence is determined by its expression level within that specific sample. Therefore, orthologous genes (e.g., human gene 'C' and mouse gene 'c') may occupy different positions. The core of this method is that a randomly selected set of orthologous gene pairs (denoted by blue boxes) are forced to share the same embedding index (e.g., human gene 'D' and mouse gene 'd' both use embedding #4), acting as anchors to align the embedding spaces. Non-anchored orthologs (e.g., 'C'/#3 and 'c'/#33) retain separate embeddings.”

2. How critical is the MGI 1-to-1 gene mapping? Can the training work without using this data set? (I guess this is the model in Figure 1a? I’m a little unclear).

Response: Thank you for this important question that allows us to clarify the role of MGI orthologous gene information in our methodology.

The MGI 1-to-1 orthologous gene pairs serve a crucial but specific role in our embedding alignment strategies. While the initial mixed training (**Fig. 1a**) can indeed function without this orthology information, our results demonstrate that alignment quality is significantly improved when incorporating MGI data. As shown in Figure 1g, the mixed training alone (**Fig. 1a**) resulted in only about 2,000 of 16,983 human genes having their mouse orthologs as the nearest embedding neighbors. In contrast, incorporating MGI orthologous information through supervised alignment (**Fig. 1c**) increased this number to 4,216 genes, representing a significant improvement in alignment accuracy.

The orthologous gene information provides essential supervision signals that guide the model to align functionally related genes across species, rather than relying solely on expression pattern similarities that may be influenced by species-specific regulatory differences. Furthermore, our positive control experiment (Supplementary Fig. 4c) demonstrated that when true orthologous relationships exist (as in our pseudo-species experiment), our methodology can reconstruct 99.7% of gene relationships, indicating that the suboptimal performance with real mouse-human data reflects genuine biological divergence rather than methodological limitations.

To address the reviewer's question about model functionality, we have added the following clarifying text to the Results section (Line 135): "*While mixed training alone (**Fig. 1a**) can generate cross-species gene representations, the incorporation of MGI orthologous gene information through supervised alignment substantially improves alignment quality."*

3. How does a model using the MGI 1-to-1 gene maps perform (i.e. the naive model?)

Response: Thank you for this insightful question about comparing our approach to a simpler baseline model using only MGI 1-to-1 gene mappings. This comment helped us realize the importance of ruling out potential confounding effects from ortholog type differences.

We performed the suggested analysis, and the analysis showed that one-to-many orthologous gene pairs indeed showed significantly lower phenotype associations compared to one-to-one orthologs (Supplementary Fig. 11).

To investigate whether the differences in phenotype associations observed across our RNA embedding similarity groups might be due to the inclusion of one-to-many orthologs rather than genuine RNA embedding similarity effects, we restricted our analysis to only one-to-one orthologs to control for this potential confounding factor, we repeated the analysis shown in Fig. 3a-c. We found that the same hierarchical pattern across the four similarity groups persisted (Fig. 3d-f). This confirms that RNA embedding similarity provides genuine predictive value beyond ortholog type differences, with the most pronounced demonstration observed in the Alliance of Genome Resources database analysis.

We have added this comparative analysis to the Results section (Line 244) and updated Figure 3 to include panels d-f showing the one-to-one ortholog results. This addition strengthens our manuscript by demonstrating that our RNA embedding similarity effects are not confounded by ortholog type, and we thank the reviewer for suggesting this important control analysis.

4. The RNA similarities seem like it should have a corresponding figure. I am also confused by the difference between DNA similarity and RNA similarity. DNA similarity includes promoters and coding sequences, which seems clear, but RNA similarity is ‘similarity to the corresponding human gene’. Which is unclear. Is this only coding sequence? Includes UTRs? Amino acid sequence (which would seem a better measure of similarity, rather than more divergent RNA/DNA sequence). The language used to describe these measures of similarity is very loose and imprecise.

Response: We thank the reviewer for this comment, which rightly identifies a key area of confusion and a lack of precision in our manuscript. We agree that a clearer distinction between our two main similarity metrics is essential for the reader's understanding.

We apologize for the lack of clarity. To be precise:

- Our **DNA similarity** metric is a traditional measure of sequence conservation, calculated using pairwise alignment scores of coding sequences and regulatory regions. The reviewer's understanding of this is correct.
- Our **RNA similarity** metric is an innovative, function-based measure derived from deep learning. It is **not** a sequence-based similarity (it is not a measure of how similar the RNA sequences are). Instead, it is the **cosine similarity** between the 200-number vector embeddings of a human gene and a mouse gene. These embeddings are learned by our GeneRAIN model based on the expression patterns of genes across 777K samples. Therefore, our **'RNA similarity'** reflects the degree to which two genes occupy a similar functional "space" based on how they relate to other genes at the RNA level. This allows

for a comparison of genes even when their DNA sequences are not highly conserved.

To avoid future confusion, we have replaced the term "RNA similarity" with "RNA embedding similarity" in the main text. In the figure labels, however, we retained the shorter form ‘RNA similarity’ due to space constraints.

The reviewer's suggestion for a corresponding figure for RNA embedding similarity is excellent. We have added a figure (Supplementary Fig. 6) for the distribution for RNA embedding similarity.

To address the critique of "loose and imprecise" language, we have made the following specific revisions to the manuscript:

In the “RNA embeddings and DNA attributes in homologous genes” section, we have added new text to clearly upfront. The new text reads (Line 181):

“*It should be noted that an RNA embedding similarity is the cosine similarity between the 200-number vector embeddings of the human and mouse genes, rather than a sequence-based similarity. A high similarity indicates that the genes have similar expression relationships with other genes across diverse biological contexts, suggesting similar functions.*”

We hope these changes improve the clarity of our methodology and help readers appreciate the distinction between sequence-based and function-based gene comparisons.

5. Continuing on, on line 172 ‘DNA embedding similarity’. What is this referring to? Inconsistent use of terms makes this paragraph virtually unintelligible. Could this section be rewritten to use language more typical, such as % matching identity, BLAST or CLUSTAL similarity? Or something like that.

Response: We thank the reviewer for this important comment, which has been helpful for us to improve the clarity and precision of our terminology. We apologize for the confusion caused.

To correct this error and ensure the clarity of the paragraph, we have replaced all instances of 'DNA similarity' or ‘DNA embedding similarity’ with 'DNA sequence similarity’.

6. Line 183-184. As ‘DNA similarity’ includes the coding sequence, then what is RNA information anyway? Again, more loose descriptions and what appears to be swapping of terminology makes this section unintelligible. The failure to define DNA and RNA similarity in the preceding section means the import of lines 189-207 is unclear.

Response: We thank the reviewer for this further comment, and we apologize for the ambiguity. To address this, we have revised the text to make it consistent, the “We explored whether gene embeddings derived from RNA information could better

explain phenotype association than DNA sequence similarity" now reads (Line 206) "We explored whether RNA embedding similarity could better explain phenotype association than DNA sequence similarity".

We also make sure that the "RNA embedding similarity" is defined as (Line 181) *"It should be noted that an RNA embedding similarity is the cosine similarity between the 200-number vector embeddings of the human and mouse genes, rather than a sequence-based similarity. A high similarity indicates that the genes have similar expression relationships with other genes across diverse biological contexts, suggesting similar functions"*

And "DNA sequence similarity" is defined as (Line 191) "We examined DNA sequence similarities, which were determined by DNA alignment scores in coding regions, as well as 1kb upstream and downstream regions of transcription sites through pairwise transcript level alignment."

Both newly added texts are in the "RNA embeddings and DNA attributes in homologous genes" Result section.

7. Line 227: But the model doesn't rely solely on transcriptome data as it includes MGI 1-to-1 orthologue mapping.

Response: We thank the reviewer for this insightful comment, which highlights a point that required greater precision. We agree that the model's cross-species alignment phase incorporates orthology information from MGI, and it was inaccurate to state that it relied solely on transcriptome data. We have revised the sentence to read (Line 265):

"This suggests that our approach, which uses primarily transcriptome data augmented with orthology information for cross-species alignment, can achieve superior performance compared to models that incorporate a large number of multi-omics features."

This clarification more accurately reflects our methodology and we believe it strengthens the manuscript.

8. RNA similairty is not defined in the methods. I am unclear how this was measured. Is it a measure of conservation?

Response: We thank the reviewer for this comment. We agree that a clear definition of "RNA embedding similarity" is essential and apologize for this omission. To address this, we have now added a precise definition in the 'RNA embeddings and DNA attributes in homologous genes' section of the Results, where the term is first introduced. The new text reads (Line 181):

"It should be noted that an RNA embedding similarity is the cosine similarity between the 200-dimensional vector embeddings of the human and mouse homologous genes, rather than a sequence-based similarity. A high similarity indicates that the genes have

similar expression relationships with other genes across diverse biological contexts, suggesting similar functions."

We believe this addition provides the necessary methodological clarity and thank the reviewer for this feedback.

9. LECIF scores were averaged for all bins in the genie region. Is this fair? Surely it would be better to average only for coding sequences?

Response: We thank the reviewer for this insightful methodological point. We agree that averaging LECIF scores over the coding sequence (CDS) provides a more precise measure of functional conservation for the purpose of comparing with phenotype associations.

In fact, we had performed this exact analysis alongside the genic region analysis. The results, which were originally presented in Supplementary Fig. 8d-f (Supplementary Fig. 10d-f in revised version), confirm that our RNA embedding similarities also outperform LECIF scores calculated specifically on the CDS.

Following the reviewer's helpful suggestion, we have now swapped these figures to highlight the most appropriate comparison. The CDS-based LECIF comparison is now presented in the main Fig. 3d-f, and the genic region-based LECIF comparison has been moved to Supplementary Fig. 10d-f.

We have also updated the main text (Results section) and the Methods subsection "Comparison with LECIF scores" to clarify that the main comparison now uses the CDS-averaged LECIF scores, as this is a more relevant metric for functional conservation.

10. The LncRNA part is underdeveloped. Considering the low levels of conservation at DNA/RNA levels this part is potentially very interesting. I guess it's virtually impossible to validate though as phenotypes for lncRNAs are so rare.

Response: We thank the reviewer for this thoughtful comment on the lncRNA analysis. We agree that the low conservation of lncRNAs at both DNA and RNA levels makes this part of the study particularly interesting, and we appreciate the reviewer's recognition of its potential significance. We also share the reviewer's observation that validating these findings is challenging due to the limited availability of well-defined lncRNA phenotypes.

To address this, we have added a paragraph to the Discussion section to acknowledge the potential significance and the limitations of this analysis. The newly added paragraph reads (Line 361):

"*Our analysis reveals particularly interesting patterns for lncRNAs, which showed lower overall conservation compared to protein-coding genes but still demonstrated measurable cross-species RNA embedding similarities. This finding is significant because lncRNAs often lack comprehensive functional annotations, making our RNA-*

based conservation metric a valuable resource for prioritizing lncRNA homologs for experimental validation. While phenotypic validation remains challenging due to the scarcity of well-characterized lncRNA phenotype associations, the quantitative conservation scores we provide offer researchers a data-driven approach to identify promising lncRNA candidates for cross-species functional studies."

Minor comments:

1. The Figure 1 labels and text are out of order, which gets confusing.

Response: Thank you for this helpful observation about the figure organization. We appreciate the reviewer's attention to clarity and initially considered reorganizing the panels as suggested. After testing different arrangements, we found that reordering the panels to group schematics and results separately would create an inconsistent visual layout (alternating schematic-PCA-schematic in the first row, then PCA-schematic-PCA in the second row) that compromises the figure's visual coherence.
To reduce confusion, we have added a clarifying sentence to the Figure 1 legend: "*Panels are organized to show each methodological approach (a-c) followed by its corresponding evaluation (d-f), with comparative analyses in g-h."*

2. The use of abbreviations like 'RNA/DNA simi' is not needed, it's just deleting 6 letters and needlessly increasing reading difficulty.

Response: Thank you for this helpful point about improving readability. We agree that the abbreviations unnecessarily complicate the text and appreciate the reviewer's attention to clarity.

We have replaced all instances of *"RNA simi."* and *"DNA simi."* with *"RNA similarity"* and *"DNA similarity"* throughout the manuscript and figure labels to enhance readability.

Reviewer #3 (Remarks to the Author):

Comments:
The manuscript is much improved and clearer. I have just a remaining comment:

For the 'RAN similarity', why not remove ambiguity entirely and rename it expression embedding (my preference), cosine embedding, transcriptome embedding, or just embedding? From the reply to my comment 4, it seems the RNA similarity has little to do with RNA and is a gene expression embedding.

Response:

We thank the reviewer for this thoughtful follow-up and for the suggestion to further refine our terminology. We appreciate the reviewer's recognition that the manuscript is clearer, and we agree that ensuring the terminology is as precise as possible is vital for the reader.

We apologize for the remaining ambiguity. The confusion arises from our insufficient explanation of what type of "embedding" is used, rather than from the concept itself. We have carefully considered the suggestion to use other terms including "expression embedding." We agree that the term "RNA similarity" could be misinterpreted if not clearly defined.

However, we would like to clarify a technical nuance regarding the Transformer architecture used in our GeneRAIN model. As detailed in our previous GeneRAIN study, the Transformer model uses two types of embeddings analogous to those used in natural language processing (NLP):
(1) **gene identity embeddings**, which correspond to token embeddings in NLP and are the embeddings we use for cross-species comparison; and
(2) **expression embeddings**, which correspond to positional embeddings in NLP and encode gene expression levels within each sample.

In the study, **only the gene identity embeddings are extracted and compared** between human and mouse genes. These embeddings summarize how each gene relates to other genes across diverse expression contexts. Therefore, while the embeddings are *learned from transcriptomic (RNA-seq) data*, they are **not expression embeddings themselves**, but rather gene-level representations inferred from RNA data.

Regarding nomenclature, we carefully considered the reviewer's suggestion to rename the metric as *expression embedding* or *transcriptome embedding*. While these terms are reasonable, we propose to retain **"RNA embedding similarity"** for two reasons:

1. It concisely reflects that the information source is RNA-seq data rather than DNA sequence, and
2. It maintains conceptual symmetry with "DNA sequence similarity," which is central to the comparative framing of our study.

That said, we fully agree that the exact nature of the embedding must be stated explicitly to remove any ambiguity. Guided by the reviewer's comment, we have revised the manuscript to make this distinction clearer at the point where the term is first introduced.

We added an explicit clarification in the section "*Representing genes across species using RNA information*", immediately following the first introduction of RNA embedding (Line 83).

The added text reads:

"It should be noted that these embeddings are gene identity embeddings (analogous to token embeddings in natural language processing), distinct from expression embeddings that encode gene expression levels in the model."

We also added an explicit clarification in the section "*RNA embeddings and DNA attributes in homologous genes*", immediately following the first introduction of RNA embedding similarity (Line 184).

The added text reads:

"It should be noted that RNA embedding similarity refers to the cosine similarity between 200-number gene identity embeddings learned by the Transformer model, rather than a similarity between RNA sequences or expression embedding themselves. These gene embeddings are inferred from large-scale RNA-seq data and capture how genes relate to other genes across diverse biological contexts."

We believe this revision, prompted by the reviewer's valuable insight, should make the terminology clearer.

Reviewer #3 (Remarks to the Author):

Okay then. It's a minor point, and if the authors insist. I still feel it impairs understanding for no reason, but is fine if explained thoroughly.

Response:

Thank you for the comment and understanding. For the reasons explained in our previous response, we have retained the current wording without further changes.