Supplementary Information Results for:

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

Zheng Su[1#], Mingyan Fang[2#], Andrei Smolnikov[1], Fatemeh Vafaee[1,3*] & Marcel E. Dinger[1,4*] & Emily C. Oates[1,5*]

[1] School of Biotechnology and Biomolecular Sciences, Faculty of Science, The University of New South Wales, Sydney, NSW 2052, Australia.

[2] State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen 518083, China.

[3] UNSW AI Institute, University of New South Wales, Sydney, NSW 2033, Australia

[4] School of Life and Environmental Sciences, Faculty of Science, University of Sydney, Sydney, NSW 2006, Australia.

[5] Department of Neurology, Sydney Children's Hospital, Sydney, NSW 2031, Australia

[#] These authors contributed equally.

[*] These authors jointly supervised this work, e-mails: f.vafaee@unsw.edu.au; marcel.dinger@sydney.edu.au; e.oates@unsw.edu.au

Optimization and evaluation of gene embedding alignment

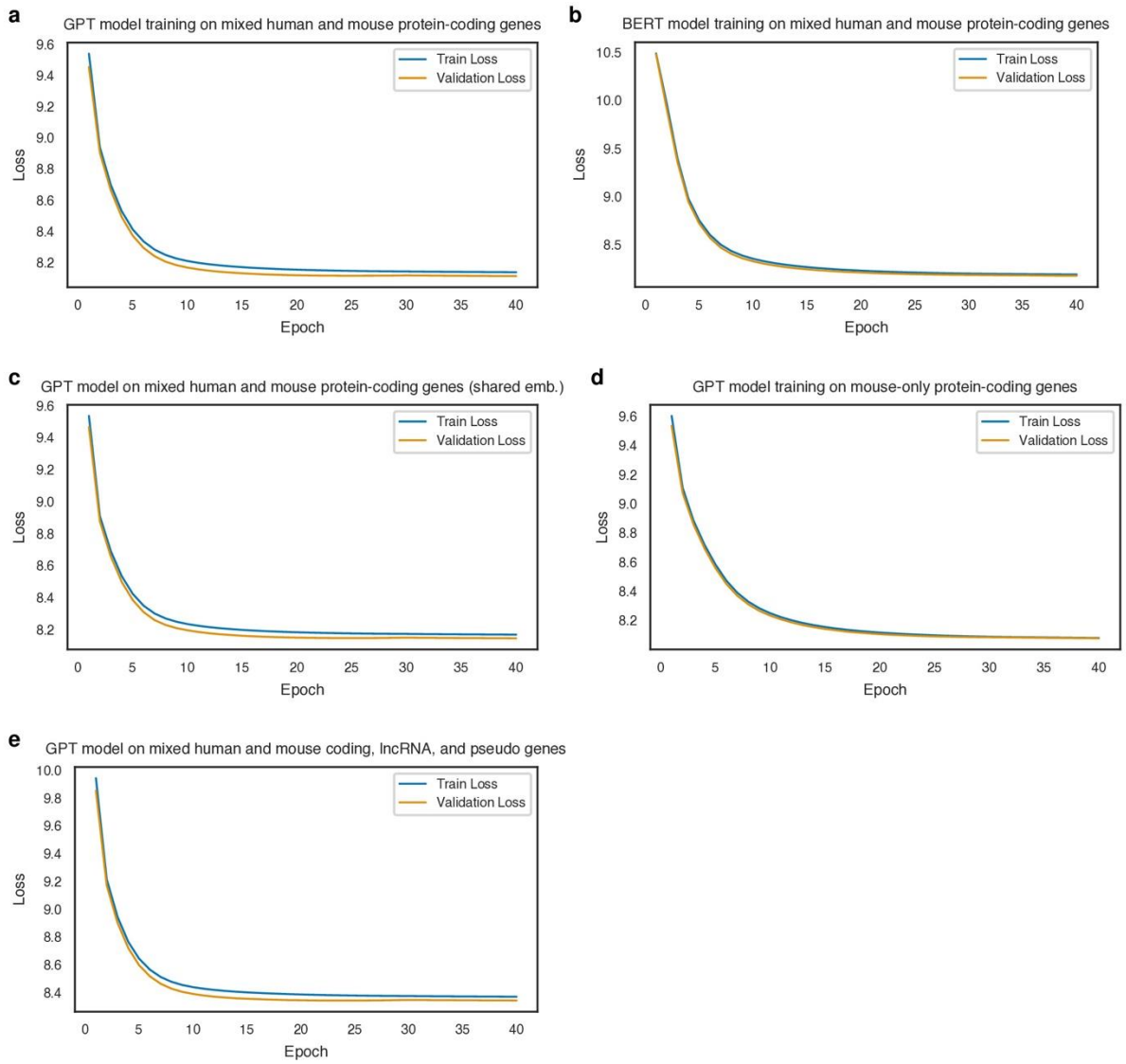
We investigated whether combining shared embeddings with supervised alignment methods could further enhance the performance. This experiment involved applying supervised alignment to genes with non-shared embeddings from the shared embedding models. The comparative analysis indicated a modest improvement (4,573 vs 4,216 orthologs being the closest), particularly when considering the 10 closest embeddings (8,255 vs 8,140) (Supplementary Fig. 4a). The embedding similarities from two approaches demonstrated a strong correlation (Spearman correlation rho = 0.89, $p < 1e\text{-}20$, Supplementary Fig. 4b).

In our previous analysis, fewer than 5,000 of 16,983 mouse genes were the nearest embedding neighbor of their human orthologs. To investigate if this was due to limitations of the methodology, we conducted a 'positive control' experiment. We invented a pseudo non-human species, along with its own gene symbols and gene embeddings. We then divided 410K human RNA-seq samples into two halves, one for the real human and another for the pseudo species. Then the two 'species' underwent normal mixed training and supervised alignment processes (Methods). We found that 18,695 of 18,757 (99.7%) genes had their inter-'species' relationships reconstructed by the model (Supplementary Fig. 4c). This indicates the high effectiveness of our approach.

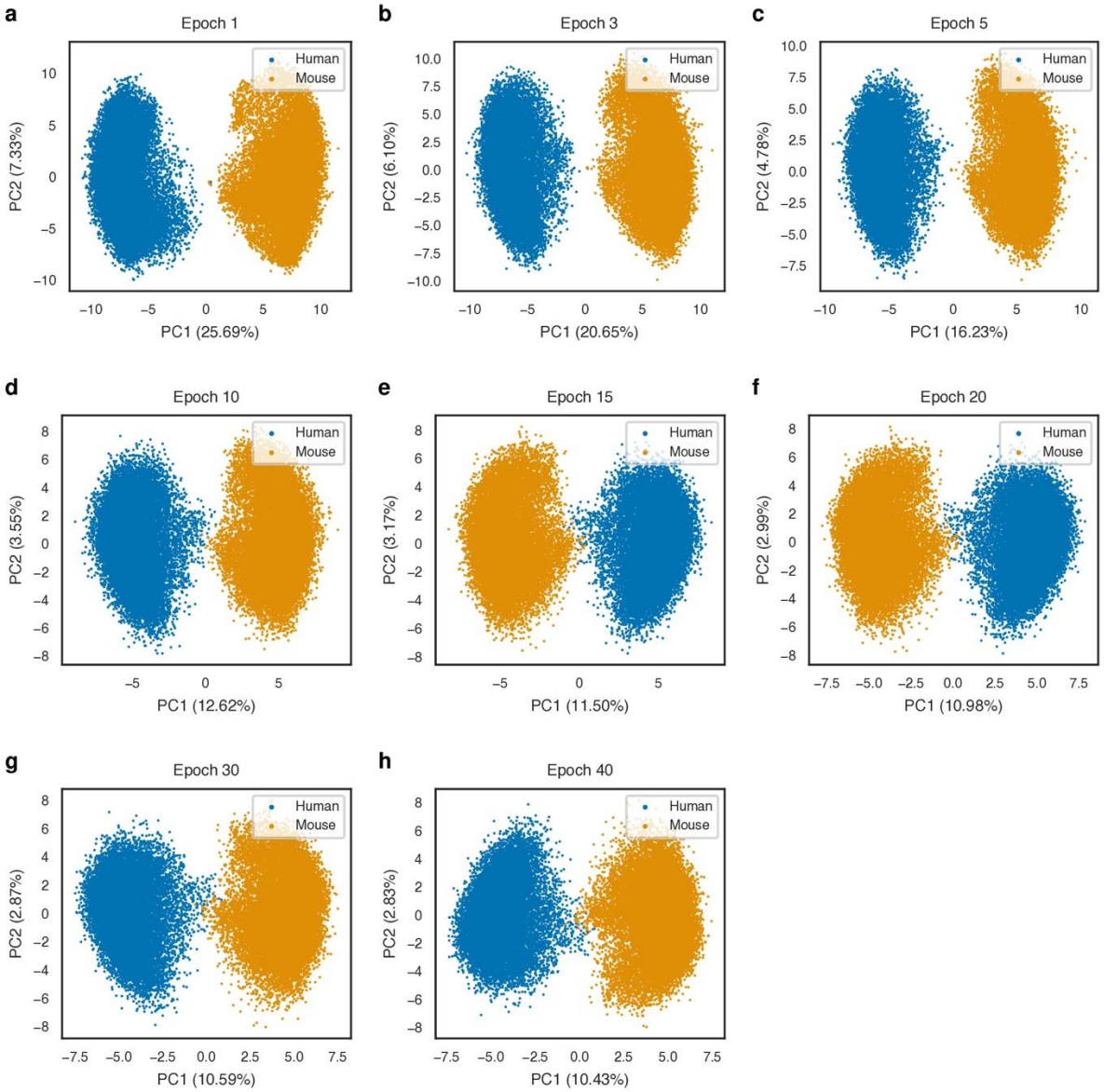
We then assessed the false positive rate of our supervised alignment approach, addressing the concern that it might arbitrarily align genes in the supervision set, regardless of their embeddings. For that assessment, we conducted a 'negative control' experiment where we randomized the embedding order after mixed training, disrupting the gene-embedding correspondence. This randomized setup was then subjected to supervised alignment. In this experiment, only 45 of 16983 (0.3%) of orthologs had closest embedding (Supplementary Fig. 4d), indicating minimal spurious appearance of relationship used in supervision.

We then investigated how alignment outcomes can be affected by training parameters, including the number of orthologous gene pairs used for supervision and training epochs. Analysis showed that the result started to plateau with just 50% of orthologous gene pairs used for supervision (Supplementary Fig. 4e) or after 10 epochs of training (Supplementary Fig. 4f). We then investigated the influence of model architectures, strong correlation between results from GPT and Bidirectional Encoder Representations from Transformers (BERT) model (Devlin, et al. 2018) was observed (Spearman correlation rho = 0.87, $p<1e\text{-}20$), with slightly superior performance in GPT (Supplementary Fig. 5a, b). Finally, we assessed supervised alignment with and without mixed training, we found mixed training produced better alignment (Supplementary Fig. 5c, d), underscoring its value in aligning gene embeddings.

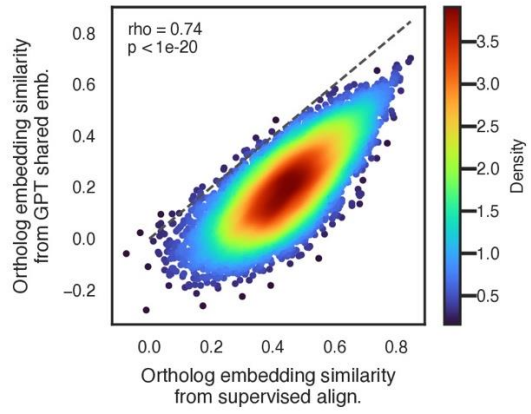
Beside gene similarity values, we also used the ranks of embedding closeness to quantify inter-species gene relationships. Further, we compared the ranks in opposite directions (mouse-to-human versus human-to-mouse), we found that they were highly correlated (Supplementary Fig. 5e, f).



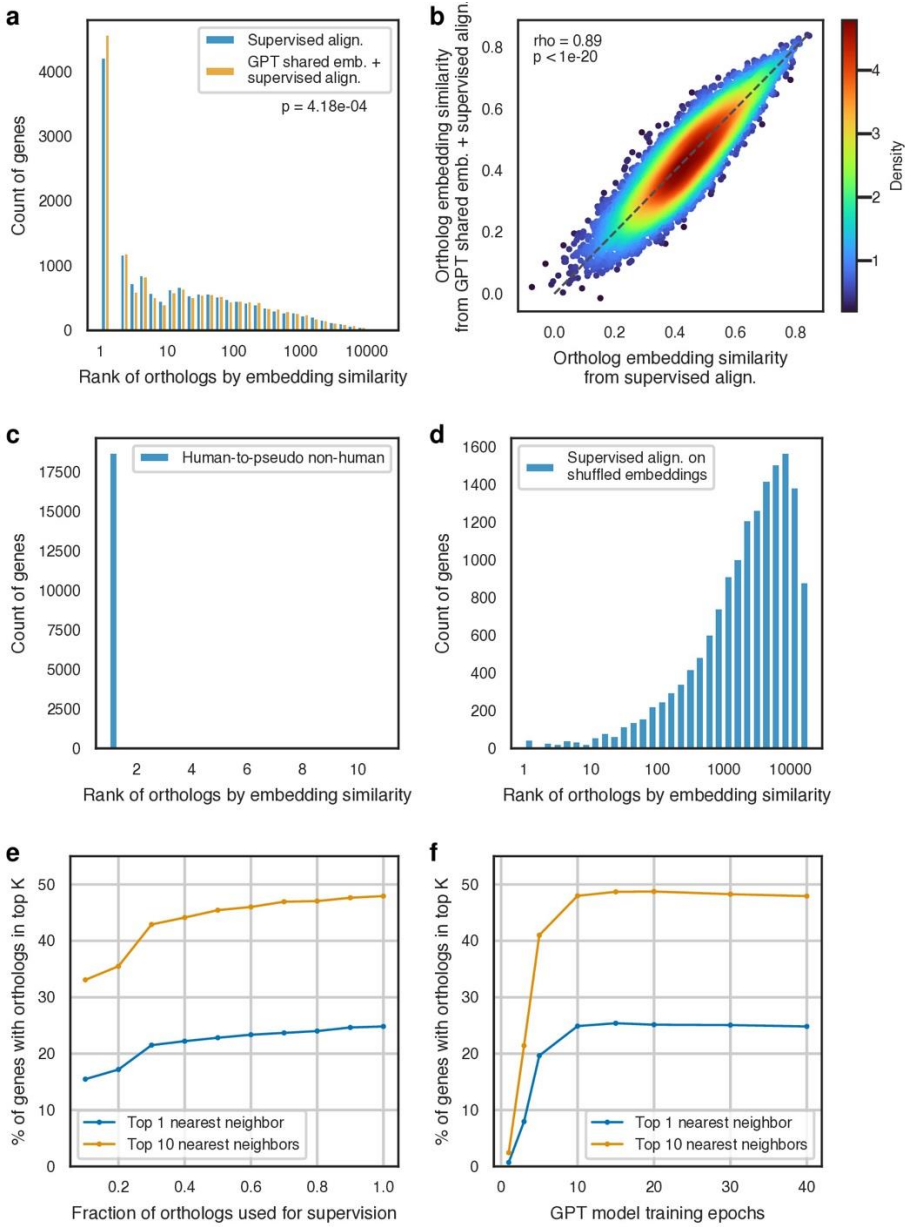
Supplementary Figure 1. Training and validation loss of across epochs. This figure illustrates the training and validation loss for: **a**, GPT on mixed human and mouse protein-coding genes; **b**, BERT on mixed human and mouse protein-coding genes; **c**, GPT on mixed human and mouse protein-coding genes using shared embedding strategy; **d**, GPT on mouse-only protein-coding genes; and **e**, GPT on mixed human and mouse coding, lncRNA, and pseudogenes.



Supplementary Figure 2. PCA plots of gene embedding distributions over training epochs. Gene embeddings from the mixed training GPT model at: **a,** Epoch 1; **b,** Epoch 3; **c,** Epoch 5; **d,** Epoch 10; **e,** Epoch 15; **f,** Epoch 20; **g,** Epoch 30; and **h,** Epoch 40. Each dot in the plots represents a gene. Please note the percentage of explained variance on the x-axis.

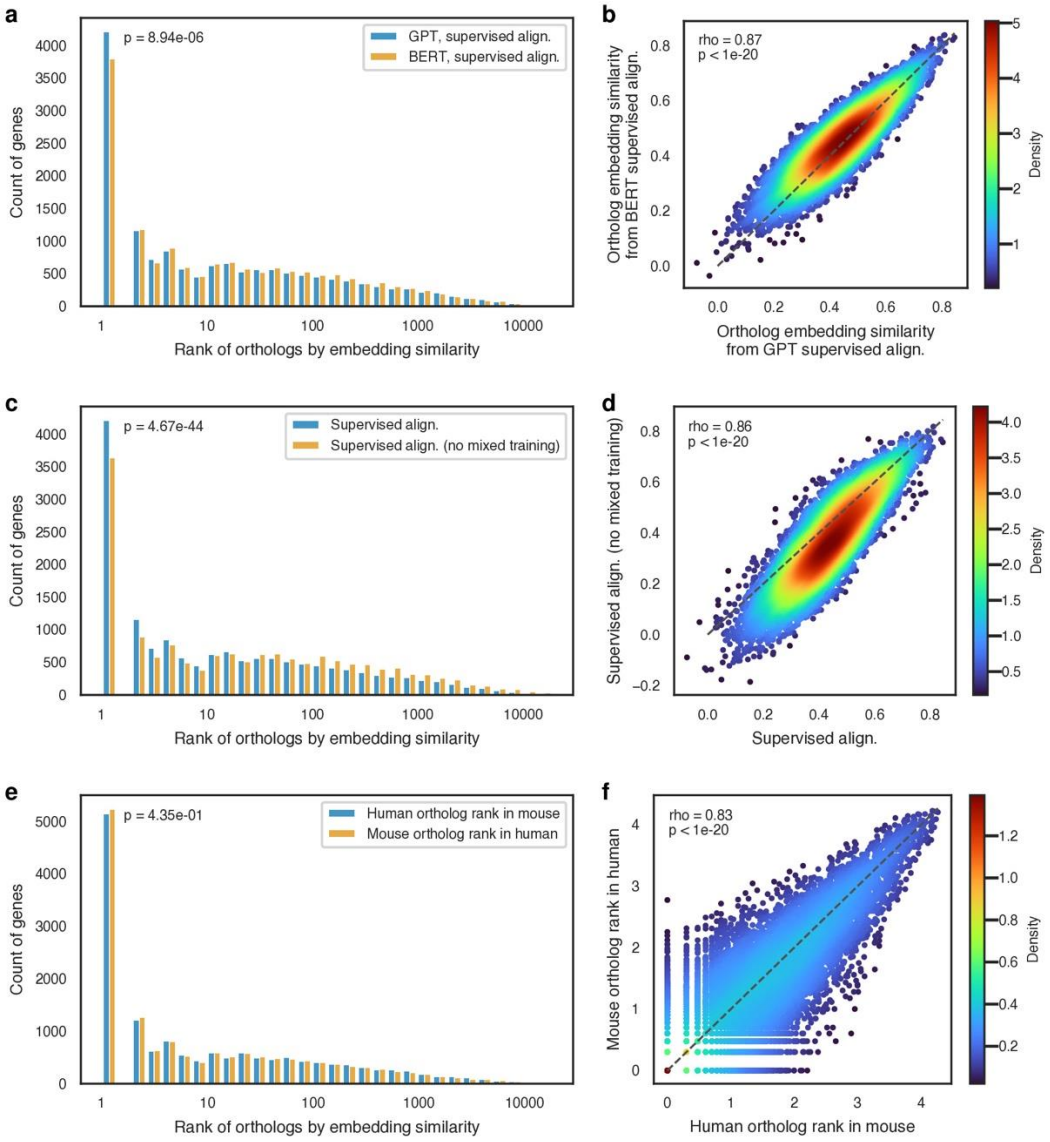


Supplementary Figure 3. Embedding similarities from the shared embedding and supervised alignment approaches. This scatter plot visualizes the Spearman correlation of orthologous gene embeddings similarities from these two embedding alignment approaches. The dashed line represents the line of equality (y=x).



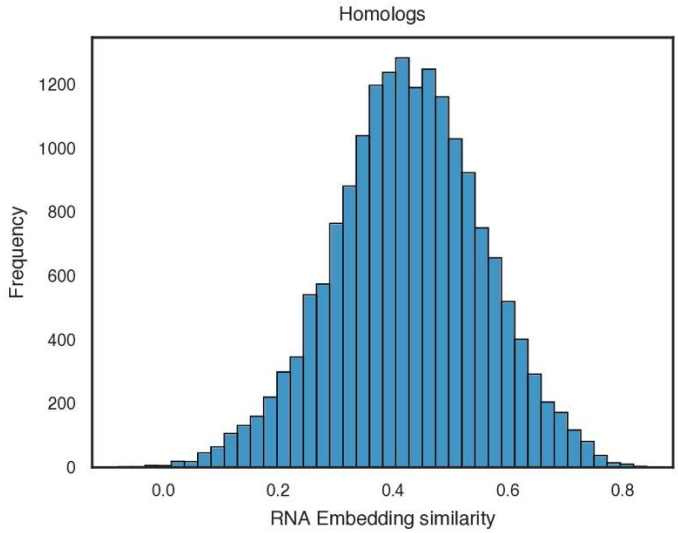
Supplementary Figure 4. Evaluation of embedding alignment strategies. a, Distribution of embedding similarity ranks for one-to-one orthologs from supervised alignment only and shared embedding plus supervised alignment approaches, with statistical significance evaluated by a two-sided Kolmogorov-Smirnov test. **b,** Spearman correlation of one-to-one orthologous gene embedding similarities from these two alignment approaches. The dashed line represents the line of equality (y=x). **c,** Distribution of embedding similarity ranks for one-to-one orthologs from the pseudo non-human species in the 'positive control' experiment. **d,** Distribution of embedding similarity ranks for orthologs from the shuffled embedding 'negative control' experiment. **e,** Efficiency of supervised alignment with varying fractions of orthologous gene pairs used for supervision, illustrated by the percentage of human genes with

their mouse orthologs ranking as the closest gene (blue line), and within the top 10 genes based on embedding similarity. **f,** Efficiency of supervised alignment across different numbers of training epochs.

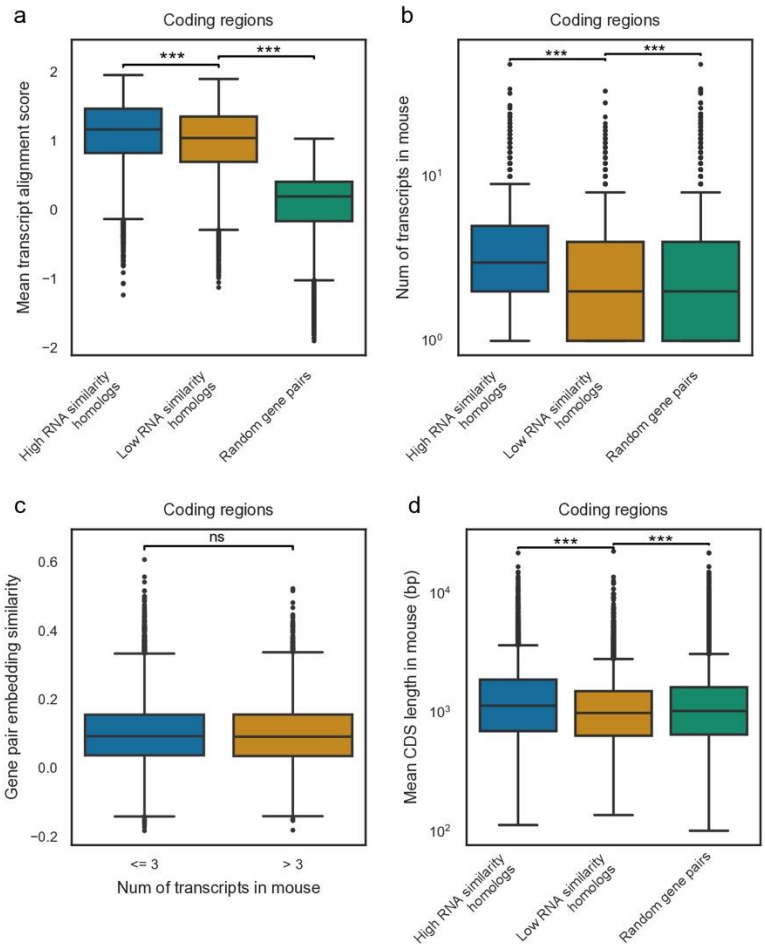


Supplementary Figure 5. Evaluation of embedding from different models. a, Distribution of embedding similarity ranks for one-to-one orthologs from GPT and BERT models, with mixed training and supervised alignment were used for both models. **b,** Spearman correlation of one-to-one orthologous gene embedding similarities from the GPT and BERT models. **c,** Distribution of embedding similarity ranks for one-to-one orthologs from models with and without mixed training. GPT model and supervised alignment were applied in both scenarios. **d,** Spearman correlation of one-to-one orthologous gene embedding similarities from the models with and without mixed training. **e,** Comparative analysis of human ortholog ranks in two directions. The rank of each human gene's ortholog within all mouse genes was calculated using embeddings from the shared embedding plus supervised alignment approach. Similarly, the rank of each mouse gene's ortholog within all human genes was also determined. This histogram visualizes the distribution of these two sets of ranks. **f,** A scatter plot showing the correlation between these two sets of ranks, assessed using Spearman's rho and p-value. In all barplots, statistical significance

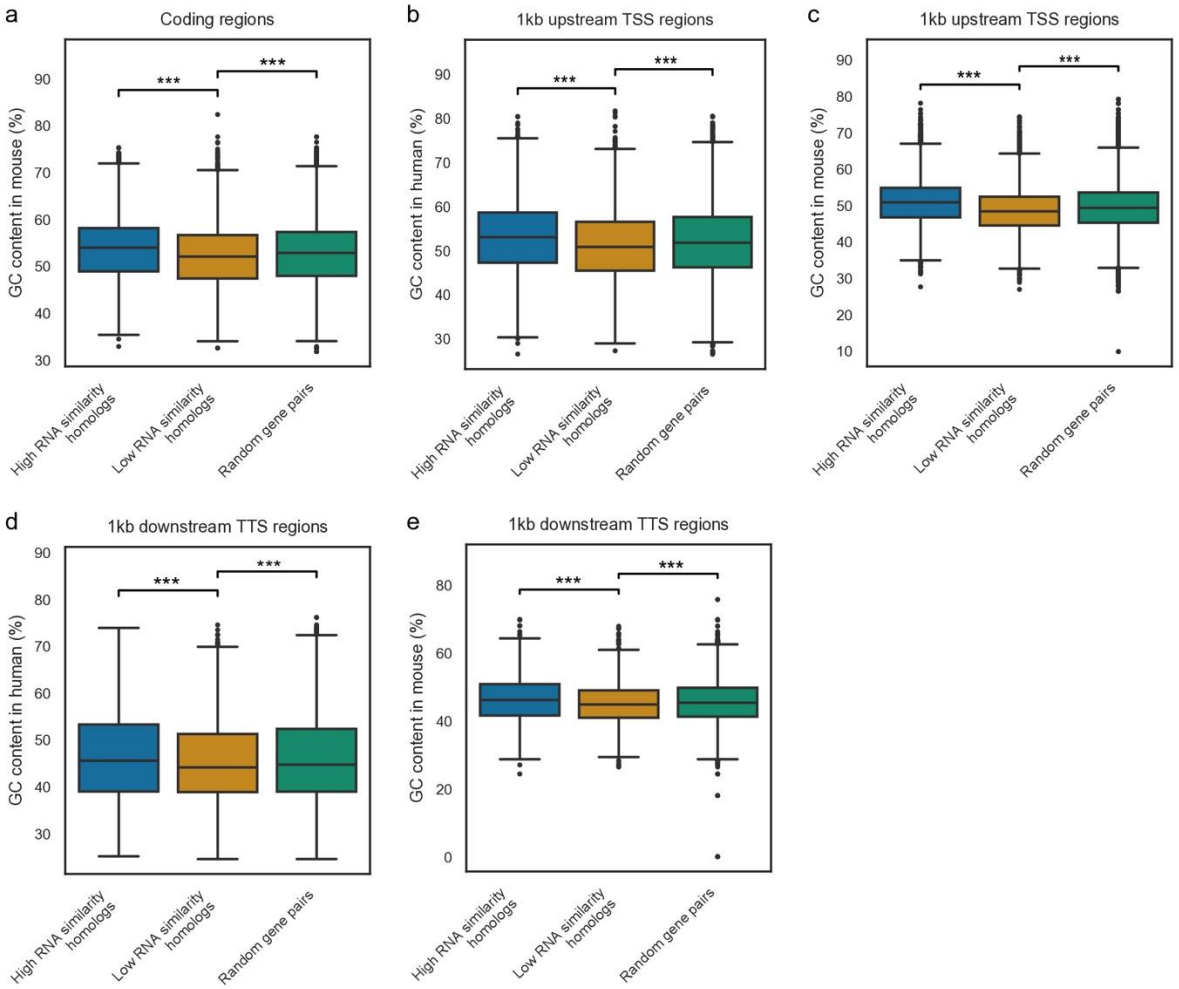
was evaluated by a two-sided Kolmogorov-Smirnov test. In all scatter plot, each dot symbolizing a orthologous gene pair. The dashed line represents the line of equality (y=x).



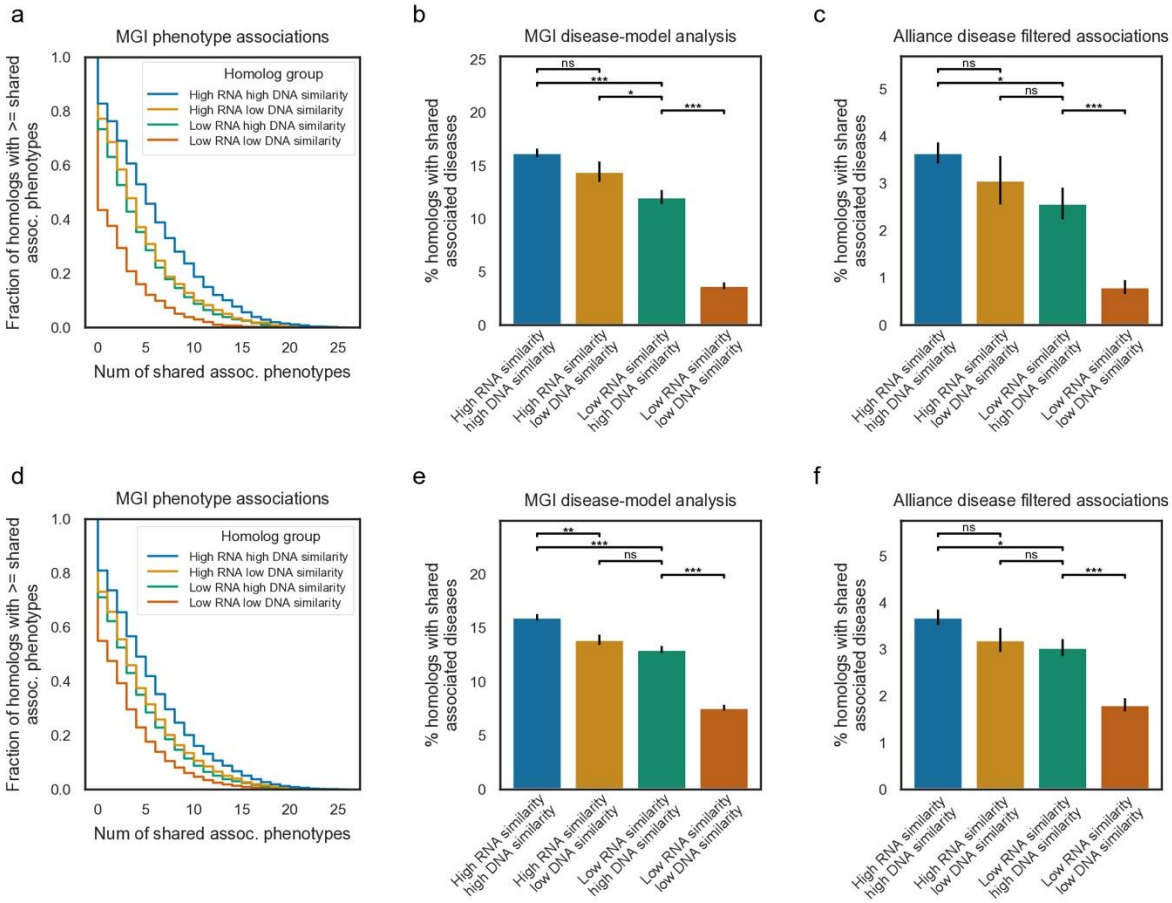
Supplementary Figure 6. Histogram of for the distribution of embedding similarity between human-mouse homologous gene pairs. The similarities are the average of two measures: (1) similarities derived from shared embeddings combined with the supervised alignment approach, and (2) similarities obtained from the supervised alignment–only approach.



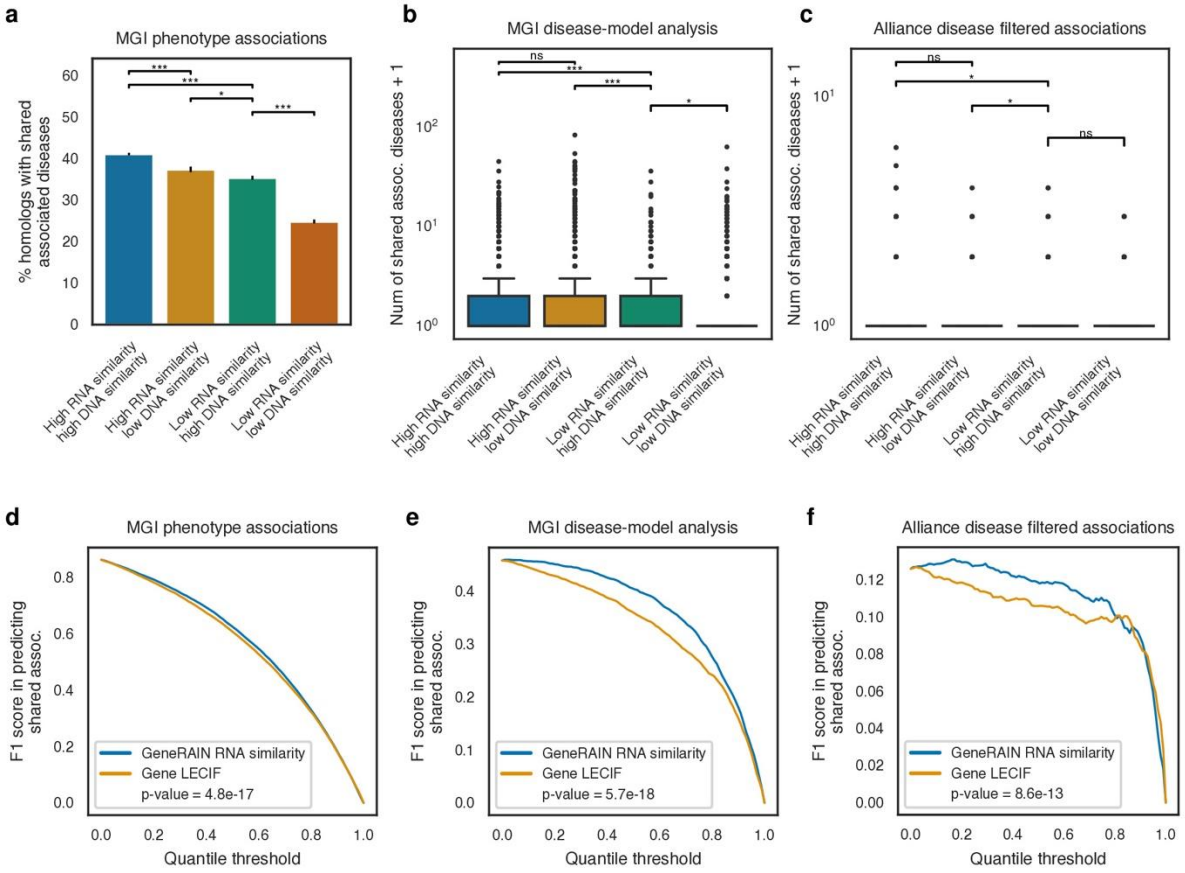
Supplementary Figure 7. DNA characteristics of homologs grouped by embedding similarity. a-c, Boxplots depicting mean transcript alignment scores (Methods) for **(a)** coding sequences, **(b)** 1kb upstream regions of transcription start sites, and **(c)** 1kb downstream regions of transcription termination sites in homologs with high embedding similarity (n = 7,249) and low embedding similarity (n = 7,103). Mean transcript alignment scores are calculated as the average pairwise alignment similarity across the human transcripts for each gene pair (Methods), offering a complementary perspective on DNA sequence conservation. Random gene pairs are included for comparison. **d,** Average coding sequence length in the human genes from three gene pair groups. In all boxplots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; outliers, points. All comparisons were evaluated using a two-sided Wilcoxon rank-sum test. *** $p < 0.00001$, ** $p < 0.001$, * $p < 0.05$, 'ns' not significant ($p >= 0.05$).



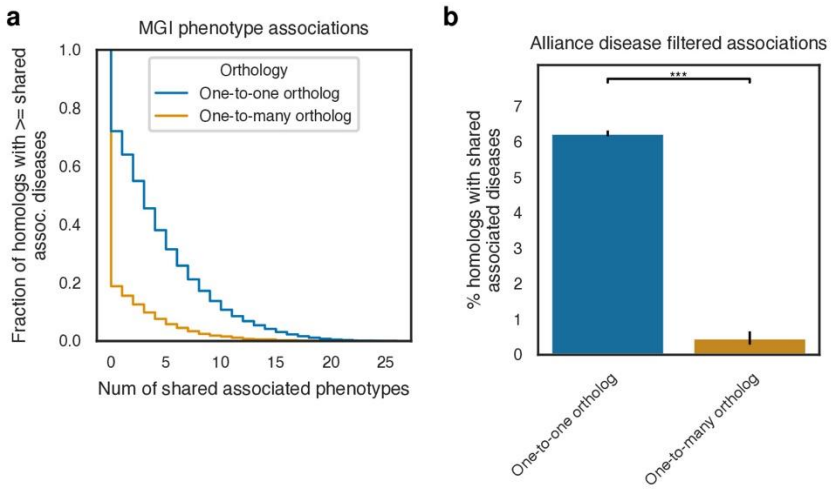
Supplementary Figure 8. GC content of homologs grouped by embedding similarity. a, Boxplots visualize the GC content of the coding sequences of mouse genes from homologs with high embedding similarity (n = 7,249) and low embedding similarity (n = 7,103). Random gene pairs are included for comparative analysis. **b-e,** Analysis of GC content in regulatory regions. With **b** for the 1kb upstream of transcription start sites in human genes; **c** for the 1kb upstream of transcription start sites in mouse genes; **d** for the 1kb downstream of transcription termination sites in human genes; and **e** for the 1kb downstream of transcription termination sites in mouse genes. Each analysis compares the two homolog groups and a random gene pair group. In all boxplots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; outliers, points. All comparisons were evaluated using a two-sided Wilcoxon rank-sum test. *** $p < 0.00001$, ** $p < 0.001$, * $p < 0.05$, 'ns' not significant ($p >= 0.05$).



Supplementary Figure 9. Embedding similarity and phenotype association analysis with alternative thresholds. a-c, As in Fig. 3a-c, except thresholds of 30% and 70% were used to define low and high RNA embedding similarity groups. Sample sizes for groups in **a-c**: n = 3,440, 554, 1,087, and 1,858, respectively. **a,** All group comparisons had p-values less than 1e-10, except for the second vs. third group (p-value = 0.165), two-sided Kolmogorov-Smirnov test. **d-f,** As in Fig. 3a-c, except a 50th percentile threshold was used to define low and high RNA embedding similarity groups (genes below the 50th percentile = low, genes at or above = high). Sample sizes for groups in **d-f**: n = 6,465, 2,280, 4,424, and 4,576, respectively. **d,** All group comparisons had p-values less than 1e-10, except for the second vs. third group (p-value = 0.052), two-sided Kolmogorov-Smirnov test.



Supplementary Figure 10. Embedding similarity and phenotype association analysis. a Barplot displaying the percentages of homologs with shared associated phenotype associations, based on ortholog phenotype association annotations from MGI database. Homologs were categorized by DNA sequence similarity and embedding similarity learned from RNA data. Error bars represent the standard errors. **b,** Boxplot for number of shared associated diseases across homolog groups, according to the human disease mouse model annotations from MGI. **c,** Barplot showing number of shared associated diseases across homolog groups, according to filtered disease association annotations from the Alliance of Genome Resources database (Methods). In all boxplots: center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; outliers, points. In barplots, two-sample z-test for proportions was used for comparisons, while in boxplots, two-sided Wilcoxon rank-sum test was used. Sample sizes for groups in **a-c**: n = 4,694, 1,107, 2,407, and 3,070, respectively. *** $p < 0.00001$, ** $p < 0.001$, * $p < 0.05$, 'ns' not significant ($p >= 0.05$). **d-f**, F1 Score vs. threshold curves comparing the performance of our RNA similarities and LECIF scores in predicting whether homologous gene pairs have shared association(s), using annotations from the database as indicated in the title. For LECIF, in contrast to Fig. 3, the average score within the genic region rather than coding region was used for each gene. The curves visualize how the F1 score (harmonic mean of precision and recall) varies across different classification thresholds. The p-values, obtained from the Wilcoxon signed-rank test, compare the F1 scores of the two metrics across 100 quantile thresholds.



Supplementary Figure 11. Embedding similarity and phenotype-association analysis stratified by orthology type. a, b, As in Fig. 3a and c, except homolog pairs are grouped by MGI orthology classification (one-to-one vs one-to-many) rather than by DNA sequence similarity and RNA embedding similarity. **a,** empirical cumulative distribution function (ECDF) plot illustrating the number of shared phenotype associations for homologs. The horizontal axis represents the number of shared associated phenotypes for each homolog gene pairs, while the vertical axis denotes the fraction of homologs with a greater number of shared associated phenotypes. Lines higher on the plot indicate more shared associated phenotypes. The group comparison had p-values less than 1e-20, two-sided Kolmogorov-Smirnov test. **b**, barplot showing the percentages of homologs with shared associated diseases, according to filtered disease association annotations from the Alliance of Genome Resources database. **a-b**, sample sizes for groups: n = 16,983 and n = 2,342, respectively. Error bars represent standard errors of the binomial proportion. In barplots, two-sample z-test for proportions was used for comparisons. *** p < 0.00001.