Simple sequence patterns support complex evolutionary representations in Evo 2

Investigating how phylogeny, local pattern matching, and sequence statistics contribute to the geometric representations of genes in Evo 2.

Published Sep 10, 2026

Arcadia Science DOI: 10.57844/arcadia-k8wc-efm5

Purpose

Mechanistic interpretability experiments performed on genomic language models (gLMs) have produced promising evidence that models have learned latent representations in which biologically meaningful properties, such as phylogenetic relationships and functional genomic elements, are linearly accessible. But there's also evidence that these results are largely reproducible with sequence statistics, such as nucleotide frequency or short motifs. This reproduction suggests that the components of evolution learned by models may be relatively coarse. Determining whether gLMs live up to the promise of really 'learning evolution' involves understanding the degree to which model latent space captures a complex understanding of the evolutionary relationships between proteins and taxa, or more rudimentary representations that constitute part of this signal. Differentiating between these signals is critical to accurately characterizing what current state-of-the-art models encode, establishing their limitations, and guiding their use.

We explored the model's internal geometry by comparing the relationships between Evo 2's organization of gene family representations and evolutionary baselines. We also tested correlation with low-level compositional confounders such as *k*-mers and GC content. Beyond analyses of representations, we experimented with engineering internal representations of Evo 2 in a controlled fashion. Given human gene sequences as inputs, we applied activation steering to

shift model predictions toward nucleotides characteristic of platypus orthologs. Across all experiments, we found evidence that Evo 2's latent space captures biological structure beyond simple nucleotide composition, including gene-family organization and species-specific signal, but that these features are strongly entangled with short-range sequence statistics and broader compositional patterns.

This work is primarily aimed at researchers who build on gLM embeddings and want a clearer sense of what gene and species representations actually encode, those developing interpretability methods for biological sequence models who want to better disentangle the specific components of biological signal in learned representations, and evolutionary biologists who need to understand the evolutionary signal captured in model latent space to leverage gLMs effectively across the tree of life.

Background and goals

There's an increasing amount of evidence that protein and genomic language models (pLMs and gLMs) can learn biologically significant concepts from sequence prediction pretraining alone. Many examples come from pLMs: sparse autoencoders (SAEs) and transcoders trained on ESM-2 and ESMFold recover features tied to protein structural motifs and functional domains [1] [2] [3] [4], and steering (pushing a model's generative output toward a specified feature), has been shown to alter properties of generated proteins such as solvent accessibility [5]. However, these studies rarely include appropriate controls, leaving it unclear whether the models encode low-scale granular biological information such as memorized sequence patterns or statistics, or more complex, higher-order representations of biology. Zhang et al. [2] make a case for the former, finding evidence that ESM-2's contact predictions resemble coevolutionary statistics, but can be reproduced from smaller, local sequence inputs alone. This suggests the model largely memorizes sequence patterns and motifs instead of intuiting information on protein biophysics.

gLMs are a newer field, and scientists have only recently begun releasing models with sizes and training datasets on the same scale as leading LLMs [6]. There is also

a growing interest in applying mechanistic interpretability approaches, which are leading to mixed results. In the original Evo 2 paper [6], SAEs trained on the 40B parameter model recovered features such as exon-intron boundaries and transcription factor binding motifs, and feature ablations changed the model's loss, suggesting causal importance. Maiwald et al. [7] trained SAEs on Nucleotide Transformer (NT) and found more than 100 annotation-associated features. They also performed steering of the A1408G antibiotic-resistance mutation, and found a CMV-enhancer feature which arose from contamination in the training databases. When the authors generated decoys that preserved 5-mer frequencies, they recovered roughly 85% of the original enhancer-probe probability, suggesting that this feature was largely local sequence statistics rather than known enhancer-motif logic.

A notable counterexample comes from Ali [8], who trained SAEs on DNABERT-2 and NT and tested whether SAE features could distinguish transcription-factor motifs that are bound in cells from the same motifs at unbound sites. After matching sites for GC content, they identified features that responded preferentially to experimentally bound sites and causally affected model predictions, while shuffled and random-feature controls showed null results. This suggests that the models capture information associated with transcription-factor binding beyond motifs and simple sequence patterns.

In a blog post from August 2025, Goodfire found that using a graph-based geodesic approximation to measure the distance between bacterial species correlated with phylogeny from the GTDB tree [9]. This correlation increased after using an angular distance between averaged species representations to transform them “to a subspace where cosine similarities are directly correlated with phylogenetic distances [9].” However, they also revealed that an XGBoost model using 1–4-mer statistics predicts the learned phylogenetic-subspace activations with ~0.9 correlation, suggesting both that short-distance pattern matching may be crucial to representations and that phylogenetic signal can be reconstructed, at least in part, using the relative abundances of different short sequence motifs.

Taken together, there's substantial evidence of sequence language models capturing meaningful biological organization, but we can see that testing the content of these representations using appropriate controls is very important.

With this information in mind, we thought it would be interesting to investigate the latent organization of gene families and species in gLMs. Not only will this provide us a more accurate understanding of which components of evolutionary signal are captured in model latent space, but it will help direct the appropriate use of models by understanding their limitations.

Our goal for this pub is to investigate how Evo 2 represents evolutionary relationships through the lens of gene families, and how information at different biological scales contributes to those representations. Rather than treating local sequence statistics and higher-order evolutionary organization as competing explanations, we ask how strongly each is reflected across model layers and gene families, and where they appear together. We break this topic into two smaller questions, and answer each with a different experiment. First, we'll investigate if Evo 2's gene family representations align with known evolutionary patterns, local pattern matching, and/or sequence statistics, and the extent to which each of these contribute to gene representation. After that, we'll test if the model's evolutionary "understanding" can be used in a generative manner, by investigating the possibility of generating orthologous genes with steering interventions.

Experimental design and results

We divided this pub into two main experiments, and each used different datasets and analyses. We describe experiment-specific methods before their corresponding results in each section.

> Access our **code** and **analysis pipelines** in this GitHub repo (DOI: 10.5281/zenodo.22697260).

Experiment 1: What relationships exist between and within gene families?

We started by asking whether Evo 2 organizes genes in a way that reflects known biological relationships, and compared latent distances within and between gene

families against phylogenetic and compositional baselines to characterize the structure of the model's gene representations (Figure 1).

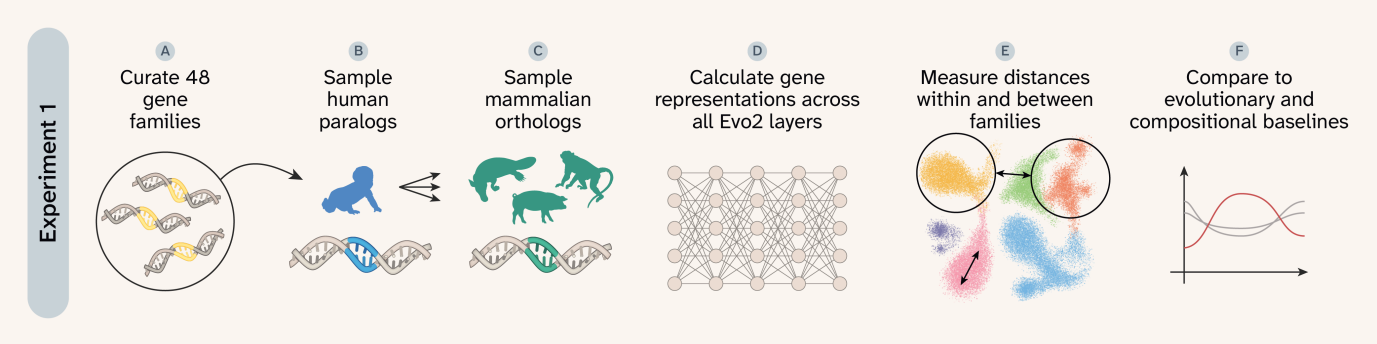

Figure 1. **What relationships exist between and within gene families?**

(A) We select 48 human gene groups from HGNC, spanning several biological classes.

(B) For each group, we collected all protein-coding members for a set of human paralogs.

(C) For each human gene, we retrieved all high-confidence orthologs from a fixed panel of 24 mammalian species using Ensembl Compara (release 116).

(D) We performed a forward pass with Evo 2 on all genes, and extracted representations of coding regions from all 32 model layers.

(E) We quantified gene group latent geometry at two scales: mean pairwise angular distance *within* gene families and 2-Wasserstein distance *between* gene families.

(F) We compared these distances with evolutionary and compositional baselines using Spearman correlation. *Between-family* comparisons used Pfam Jensen–Shannon distance, *within-family* comparisons used species-tree and gene-tree patristic distances, and both *between-* and *within-* comparisons used GC content and 6-mer composition.

Model and data selection

We selected Evo 2-7B as our base model [6], and ran all experiments on a single NVIDIA A10G GPU. For our first experiment, we built our dataset from a curated panel of 48 human HGNC gene groups [10], comprising 1,144 paralogs in total. We selected these 48 families to span several biological axes, including redox and metalloenzyme chemistry, GPCRs, and P-loop GTPases (see the project repository for a full list). For each of the 1,144 human paralogs, we queried Ensembl Compara (release 116) using the corresponding Ensembl gene identifier for each gene and retrieved its precomputed orthologs across a fixed panel of 24 mammals: six primates (human, chimpanzee, gorilla, orangutan, macaque, marmoset), four glires (mouse, rat, guinea pig, rabbit), 10 laurasiatherians (dog, cat, giant panda, ferret, horse, pig, cow, sheep, goat, little brown bat), elephant, armadillo, opossum, and platypus, with human as the query and orthologs drawn from the

remaining 23 species. We kept only orthologs classified by Compara as one-to-one, retaining at most one ortholog per species. We filtered the resulting nucleotide sequences using sequence- and annotation-quality criteria, including complete in-frame coding sequences and absence of assembly gaps. We excluded ortholog groups with fewer than 10 represented species from within-group analyses.

Gene inputs included 5′ UTR, introns, exons, and 3′ UTR on the gene's strand. At every layer, we obtained representations from the residual stream and retained coding positions only, masking out non-coding DNA. We tiled long inputs (>8,000 bp) into windows for separate forward passes, with a maximum of 24 windows sampled across very long loci before aggregation. We then converted representations to vectors by mean-pooling the second half of token positions, as the autoregressive nature of the model gives early tokens low context [9]. We then L2-normalized the vectors.

> **Data**, including all human paralogs and sampled mammalian orthologs, are available on GitHub.

Geometry and distance metrics

After obtaining the representation vectors for all genes of interest, we investigated the gene family geometry at two levels: *within-* and *between-* gene families. We calculated distances *between-* gene families using the 2-Wasserstein distance. We treated each gene family as an empirical distribution of representation vectors, where the distance between families and is defined as:

where and are gene representations from and , respectively, is the angular distance between two representations, and is the set of valid transport plans between the two empirical distributions. Intuitively, the Wasserstein distance measures the minimum cost of transporting the distribution of one gene family onto the distribution of another. This allows the metric to

account for both the relative location and spread of each gene family in representation space.

We calculated angular distance between two normalized representations as:

—

where is the cosine similarity between the two vectors. Distances range from zero for identical vector directions to one for maximally opposed directions.

For the *within-family* analysis, we compared species relationships separately within each ortholog group, where an ortholog group consists of one human paralog together with its one-to-one orthologs across the mammals dataset. For an ortholog group , we define the within-group pairwise distances as:

where and are species in the ortholog group. These distances form the matrix , which describes the geometry of species representations for that gene. We compared this matrix with the corresponding species-pair distance matrix for each baseline. For each gene family, we report the mean Spearman correlation across its constituent ortholog groups.

To analyze the similarity of Evo 2's gene family organization to biological and statistical baselines, we calculate the Spearman correlation between our distance matrices and each respective baseline at each model layer. The baselines included:

- **Pfam JSD:** Used for *between-family* measurements. Every gene family has a profile hidden Markov model (HMM) curated by Pfam, which specifies the amino acids (aa) expected at each position of the family's domain. We averaged those per-position emission probabilities into a single 20-dimensional amino acid composition vector per family, and took the Jensen–Shannon divergence (JSD) between each gene family pair.
- **Species tree:** Used for *within-family* measurements. This is an independent phylogeny baseline for mammalian ortholog comparison. From MamPhy's time-scaled mammalian trees [11], we drew 100 trees from the posterior credible set, and pruned them to our mammalian species. We then computed the average patristic distance for every species pair across all trees.

- **Gene tree:** Used for *within-family* measurements. For each ortholog group, we translated nucleotide sequences into protein sequences, aligned these sequences with MAFFT (v7), built a FastTree approximate maximum-likelihood tree (v2), and then calculated the patristic distance (sum of branch lengths) between each pair of tips.
- **GC content:** Calculated as the GC fraction of the coding DNA sequence (CDS), where the distance is equal to the absolute difference between fractions.
- ***k*-mer composition (6-mer):** Determined by dividing the sequence into overlapping 6-mers (i.e., a 1 bp step), counting the frequency of each possible 6-mer in the CDS and taking the resulting cosine distance between spectra.

Gene family representations correlate with evolutionary and compositional baselines

The correlations between gene families and each baseline are layer-dependent, as seen in Figure 2.

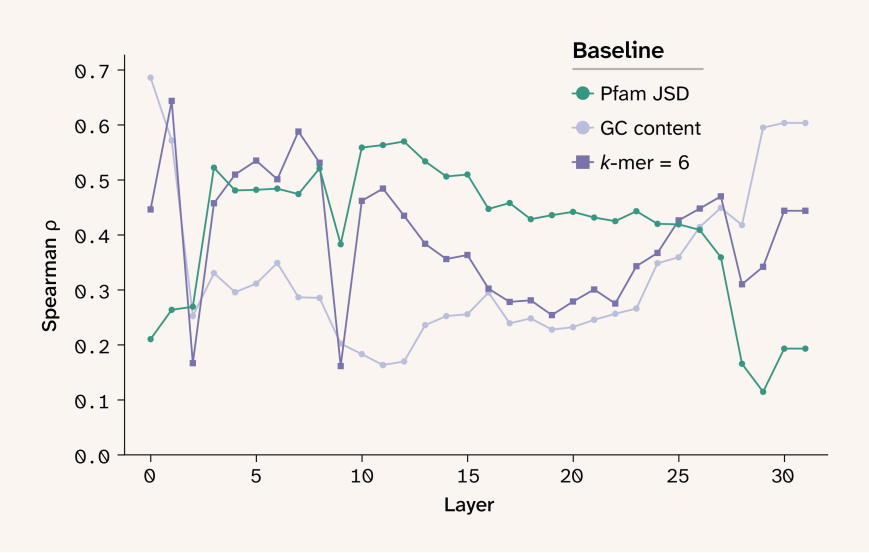

Figure 2. **Between-family geometry correlates most strongly with Pfam similarity in the middle layers, while compositional baselines rise near the model edges.**

Spearman correlation between Evo 2's *between-family* 2-Wasserstein distances and three independent baseline distance matrices for all 32 model layers. Pfam JSD measures differences between family-level Pfam amino-acid emission profiles, while GC content and 6-mer composition provide nucleotide-level compositional baselines.

The earliest and latest layers show moderate to high correlation with the compositional baselines, particularly GC content, while the Pfam baseline

correlation is weakest, ranging from . In the middle of the model, between layers 10–24, the Pfam baseline correlation climbs above both compositional controls to reach moderate levels of correlation.

Patterns within gene families turn out to be much less straightforward. The per-family curves span a large range of correlations at every layer, and the relative ordering of baselines is inconsistent from gene family to gene family. If we take a closer look at a few examples in Figure 3, we can see that the adrenoceptor family has the highest average correlation with the gene tree, while the glutathione peroxidase family has much higher correlations with 6-mers relative to the other baselines. Peroxidase's correlations with all tested baselines are consistently similar across all layers of the model. Across all three example gene families, the baseline correlations are different, in terms of the relative baseline rankings, the correlation values, and the differences between the baselines. The median across-layer Spearman correlation between pairs of metric curves, computed over all 48 families, also supports this point: the median *k*-mer and species tree correlation is 0.916, but ranges from 0.001 to 0.991. And the median *k*-mer and gene tree correlation is 0.891, but ranges from −0.249 to 0.993.

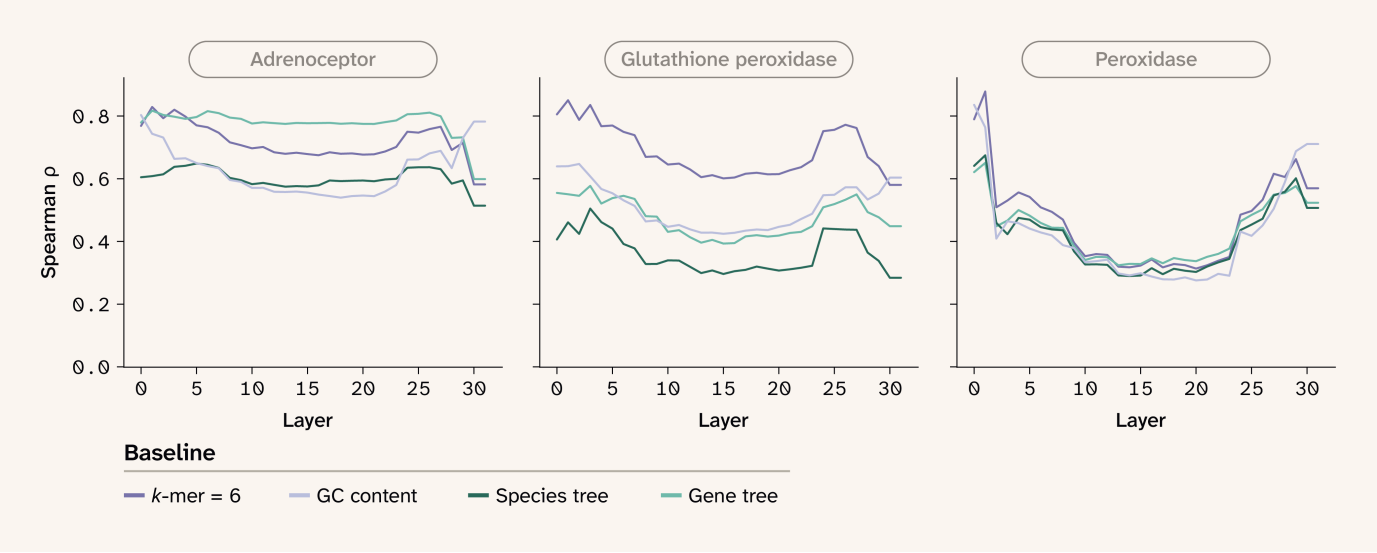

Figure 3. **Within-family geometry is strongly family-dependent, with inconsistent ordering of evolutionary and compositional baselines.** Spearman correlations between Evo 2 *within-family* angular distances and sequence-identity, GC-content, 6-mer, species-tree, and gene-tree distances across all model layers for three representative gene families. Both the magnitude of the correlations and the relative ordering of the baselines differ substantially among families, making it challenging to draw dataset-wide conclusions.

> Our **data**, including gene sequences, Evo 2 representations, and calculated distances for all 48 gene families we analyzed, are available on GitHub.

Either this approach is too generic to support a general claim about how Evo 2 organizes *within-family* relationships, or within-family relationships are too variable to be informatively resolved by our approach. Different genes carry different types and amounts of evolutionary signal, depending on their conservation, variation, and other characteristics. Considering this diversity, there are several potential interpretations that could be considered from our inconsistent *within-family* results. On one hand, the model may be correctly capturing *within-family* relationships that are naturally different across families. It's also possible that some of the variation could reflect noise where gene-specific evolutionary structure is weakly represented by the model. However, the presence of high correlations between some gene families and baselines indicates this isn't the only contributing factor.

Our experiment doesn't distinguish these cases from each other, so if you're interested in a particular gene family, we'd recommend investigating its latent geometry individually.

Experiment 2: Can we use representations for sequence generation?

Correlations from the first experiment tell us which biological and compositional patterns may be present in the latent space, but not whether Evo 2 can actively use them in the generative process. To test this, we designed a series of steering experiments, which also let us further explore the geometric representations of genes and species (Figure 4).

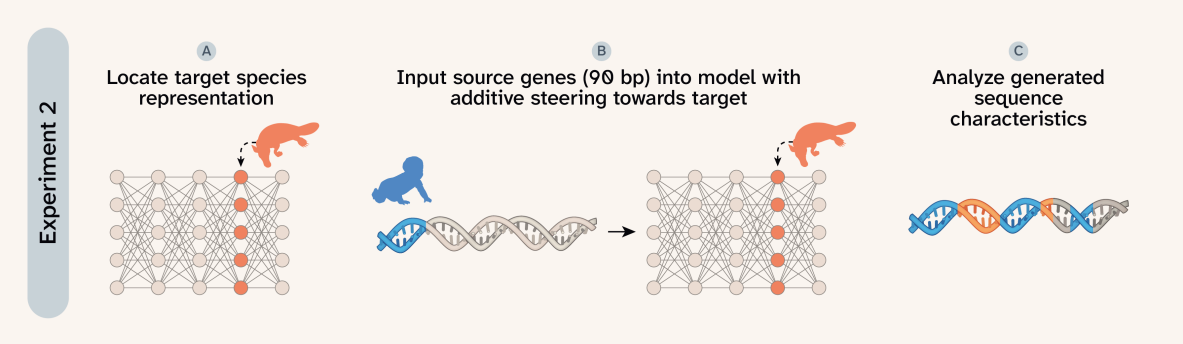

Figure 4. **Can we use representations for sequence generation?**

(A) We constructed paired representations from human–platypus orthologs and calculated the species difference for each gene (as a contrastive mean difference between the respective representations), using held-out genes to estimate a platypus-steering direction without leaking information from the genes we steer on.

(B) We compared the direction and magnitude of these steering vector candidates across Evo 2 layers to identify a layer where the human-to-platypus direction was consistent across genes.

(C) We injected the held-out steering vector into the residual stream while prompting Evo 2 with 90 nucleotides from the corresponding human gene, generated the following 1,000 nucleotides, and evaluated whether steering shifted generation “towards platypus” while maintaining functional protein-coding characteristics, such as no premature stop codons or indels.

Additive activation steering is a method that lets us test whether information in a model’s latent space is causally involved in generation. In practice, we identify a direction in Evo 2’s representation space associated with a feature of interest and add that direction to the model’s internal activations during inference. If the direction captures information the model uses, steering along it should shift the generative distribution toward that feature. Here, we tested whether a species direction could push Evo 2 toward generating a sequence containing species characteristics. We chose *Ornithorhynchus anatinus* (platypus) as our steering target due to it being the sister lineage to all other extant mammals, in other words, the most distantly related. Because of this, platypus sequences will contain many species-specific alleles when compared against other mammalian species in our data set. Since we have some expectation about the low frequency with which an unsteered model may select these rarer platypus alleles, which we refer to from now on as "private," they serve as a relative indication of whether the steered generated sequence is more "platypus-like," "human-like," or converging to a consensus representation. However, it should be noted that our “private” nucleotides are only confirmed to be unique relative to our 24-mammal species set. While we considered this sufficient for our pilot steering experiments, future

work would require aligning platypus against a larger outgroup to confirm that "private" nucleotides reflect genuine species-specific variation.

Data

For these steering experiments, we created a dataset of human and platypus orthologs from Ensembl (release 116). We began with high-confidence orthologs and grouped genes into homology blocks using MMseqs2 [12] (release 15) sequence clustering. We then sampled at most one gene pair per block to reduce redundancy between the steered gene representation and the calculated steering vector. We stratified the resulting genes into five equal-sized bins according to human–platypus protein identity, defined as the similarity of protein amino acid sequences between the human and platypus ortholog of a gene. The resulting bins had quintile boundaries of 64.68%, 72.89%, 80.48%, and 88.09%, and we sampled 80 pairs uniformly from each quintile (Figure 5). We ensured that all gene pairs had complete, in-frame canonical coding sequences, and contained a 90 bp homologous, indel-free prompt region near the start of the coding region.

> The **human and platypus ortholog dataset** is available on GitHub.

This is a difficult task. We hoped that 90 bases would provide sufficient information for the model to recognize gene identity, while minimizing the human species prior in order to generate a “platypus-like” sequence rather than a “human-like” sequence. Previous work tested Evo 2 on gene completion given a 1,500–2,000 bp prompt [6], but this prompt is more than 11 times larger than ours, and may benefit from a strong species prior.

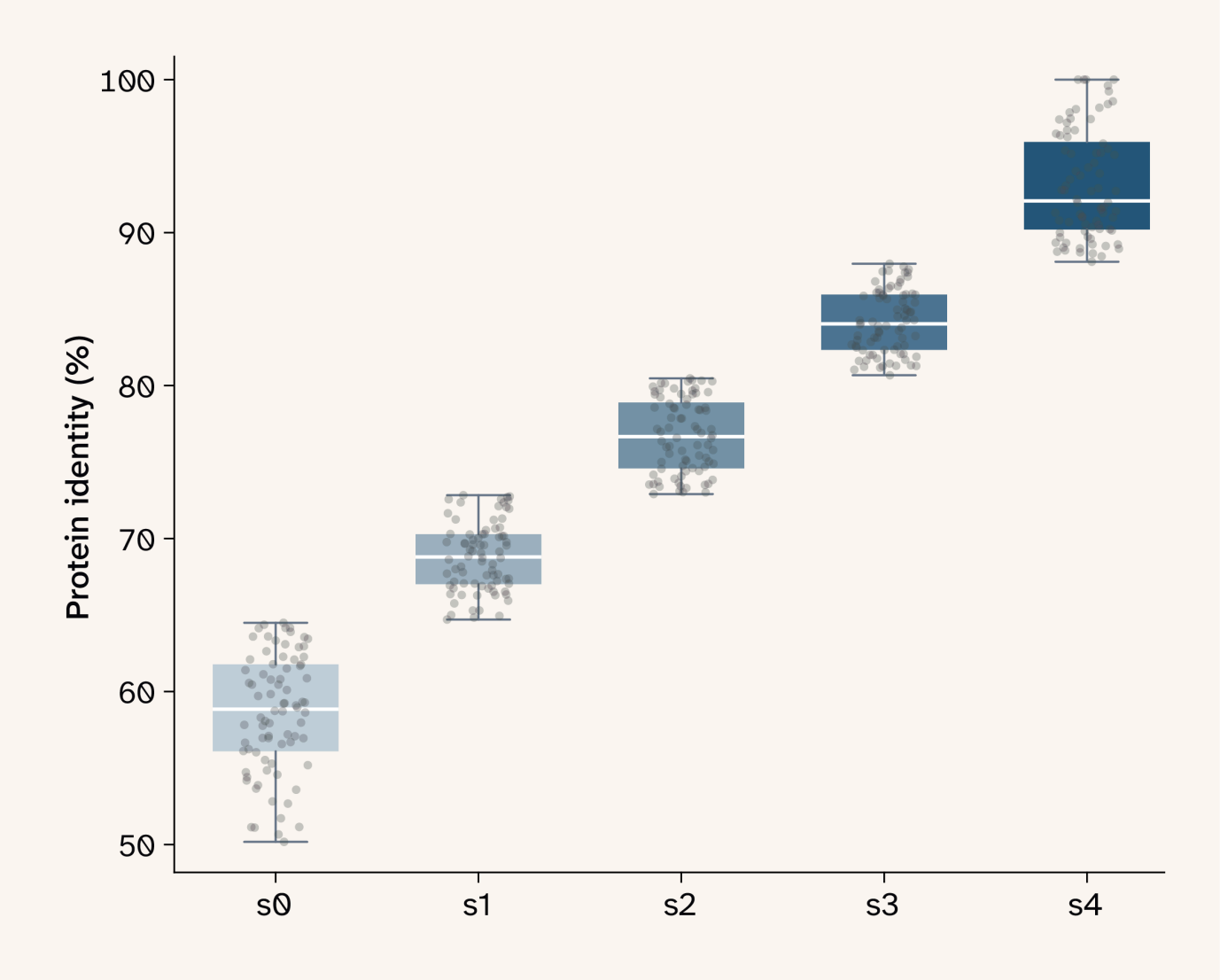

Figure 5. **The steering dataset evenly samples human–platypus orthologs across a broad range of protein conservation.** We sampled 400 human–platypus ortholog pairs uniformly across five protein-identity strata. This stratification allows us to evaluate steering separately across genes with substantially different levels of human–platypus conservation.

Investigating steering vector candidates

Our steering experiments began with an investigation of the model's latent geometry and representation of the platypus species with the goal of finding a steering vector to test. For every gene in the n = 400 paired human–platypus set, we calculated the contrastive mean difference between the platypus and human representations:

where indicates the gene identity, is the model layer, and is the representation mean-pooled over the whole CDS. For gene , the direction we injected at to steer with is the -held-out mean:

This prevents gene information leakage into the steering vector, and is why we sampled each from a different clade. During steering, we also scale this direction, , by a factor , such that adds one full copy of this average shift, adds half of it, and larger values move the model's activations progressively farther in the same direction. At every , we looked at two metrics, leave-one-out alignment and magnitude spread. Leave-one-out alignment, , tells us how much our gene 's direction aligns with the average direction of the rest of the genes, and magnitude spread, , tells us how much the magnitude of the resulting "platypus" vector changes across the gene dataset. We aimed to steer in regions of the model where the directional alignment was strong and the magnitude spread was low.

A potential candidate emerges at layer 27

On the human–platypus pairs, the leave-one-out cosine (LOO-cos) is near zero through layers 10–20, rises from layer ~21, and saturates near 1.0 at layers 28–31 (Figure 6, A). For our purposes, final blocks were excluded since at the 28–31 layer depth, next-token prediction dominates the residual stream, so high similarity at those layers likely reflects a strong anisotropic direction dominating the latent representations rather than a shared species direction. The magnitude spread across genes, , shows complementary results, and is higher through layers 9–23 before dropping to a minimum at layer 27 (Figure 6, B). Since this layer also has a high LOO-cos before entering the final layers, we chose it as the first candidate for steering injection.

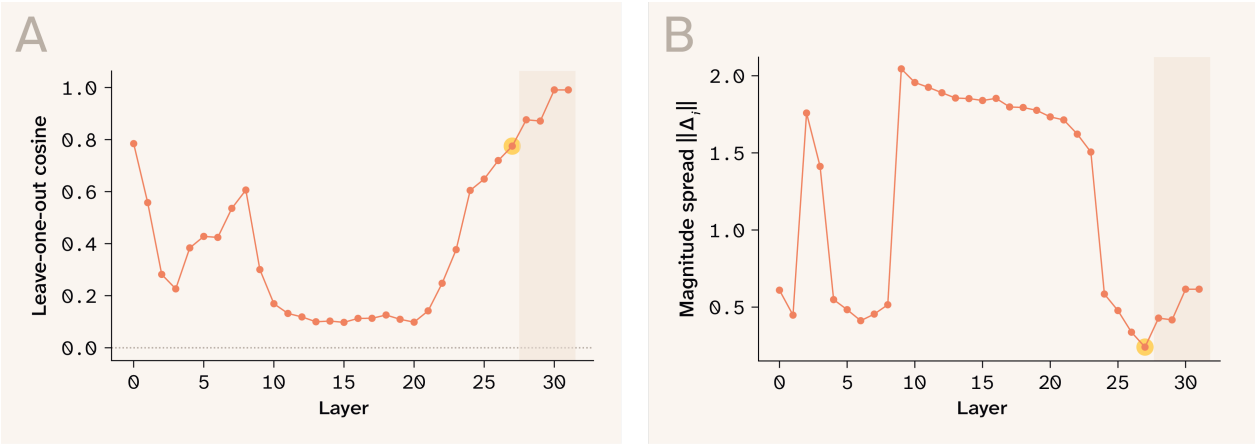

Figure 6. **A consistent human-to-platypus representation is found in late-middle Evo 2 layers, with layer 27 providing the strongest steering candidate.**

(A) Leave-one-out cosine similarity between each gene-specific human-to-platypus difference vector, , and the mean difference vector calculated while holding that gene out, . Higher values indicate a species direction that is more consistent across genes.

(B) Variation in the magnitude of across genes. The final blocks are highlighted to indicate that near-saturated directional similarity is dominated by late-model anisotropy, excluding them from the candidate layer selection process. Layer 27 combines high cross-gene directional alignment with the lowest magnitude spread before these final blocks and was therefore selected for the initial steering experiments.

Steering intervention and scoring metrics

Once we chose layer 27 as our focal layer, we used additive steering to add to the block's residual-stream output at every position, including the prompt, with as a scaling factor. We gave Evo 2 a 90 bp prompt from the human gene in the pair, which we set to be the first 30-codon window in the human–platypus gene alignment with no indels. Evo 2 then generated the next 1,000 bp in the gene, decoding at temperature 0.7, and we generated each gene repeatedly so we could analyze the generative distribution. We tested steering at a variety of strengths (related to the magnitude of intervention via how much of the steering vector we add), . We also performed inference on the human prompt without any steering (which we refer to as "unsteered") as a control.

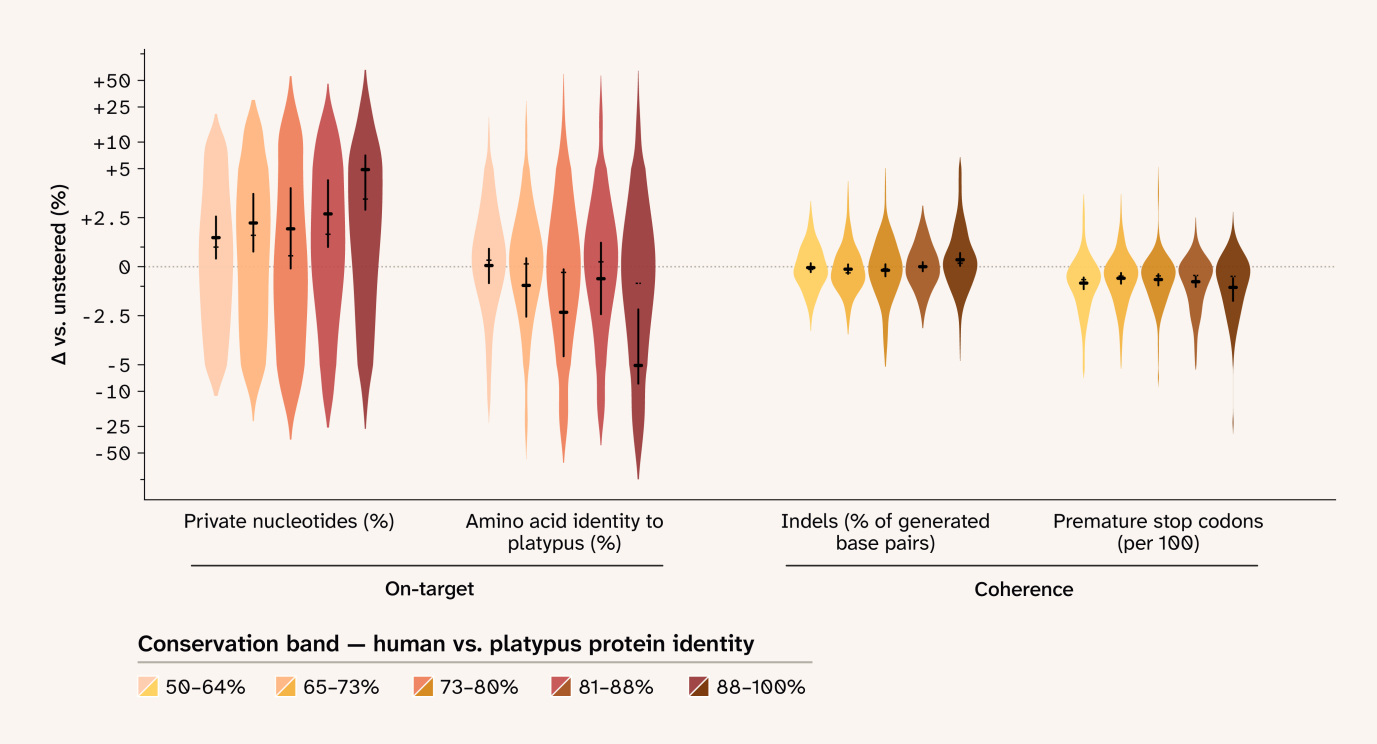

Figure 7. **Steering recovers more private platypus nucleotides without substantially impacting protein coherence, but can reduce platypus amino acid identity to the platypus target.**

Difference ( ) in generated sequence relative to unsteered generation in four metrics, stratified by increasing human–platypus protein identity (%): recovery of “private” platypus bases, amino acid identity to the platypus target, generated indels, and premature stop codons. Private-base recovery increases across all conservation strata, while the amino-acid response trends towards becoming increasingly negative among more conserved genes. Indel burden and premature-stop frequency do not largely change, suggesting that these shifts are not explained by a general loss of coding-sequence coherence.

We defined both sites and scoring on a nucleotide alignment of the generated sequences against the target species (platypus). We assessed steering success on four metrics: 1) private platypus nucleotides recovered, where "private" alleles were those unique to platypus across the 24 mammalian species used in the first experiment, among available aligned bases 2) amino acid identity to the platypus target, calculated by translating the generated and corresponding platypus sequences and globally aligning the resulting amino-acid sequences, 3) new indels measured as a percentage of generated bp, and 4) average premature stops per 100 codons. The first two measure how well we're steering towards the target species, while the last two measure protein coherence, as a sequence full of frameshifts and stops does not constitute a functional protein, no matter how platypus-like its composition may be. Initial results from our layer 27 steering experiment are depicted in Figure 7 and Figure 8.

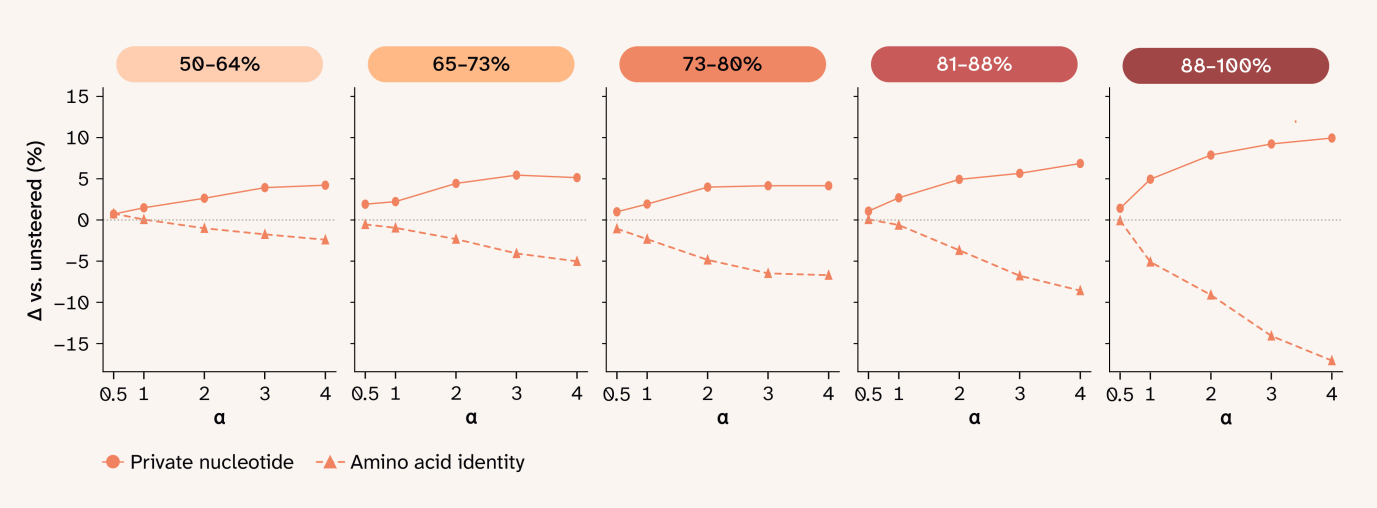

Figure 8. **Increasing steering strength amplifies the trade-off between private nucleotide recovery and amino acid identity, particularly in highly conserved genes.**

Change relative to unsteered generation in private platypus-base recovery and amino acid identity to the platypus protein as steering strength increases from 0.5 to 4. Panels separate genes into five strata according to human–platypus protein identity. Stronger steering generally increases recovery of private platypus nucleotides, and decreases platypus amino acid identity, with the largest response in the most highly conserved stratum.

Steering increases private platypus nucleotide recovery but reduces amino acid identity

The gains from steering with our calculated vector are modest but consistent. Private nucleotide recovery improves over the unsteered baseline in every stratum, and that improvement mostly grows with conservation. The change in amino acid identity goes in the opposite direction, turning increasingly negative in the more conserved genes, with the exception of stratum three. The spread of both metrics also widens with conservation, so the most conserved genes show not only the largest average response to steering, but also the most variable one. Indel burden and premature stops don't shift largely across strata (Figure 7), which suggests these nucleotide- and protein-level changes aren't a byproduct of degraded sequence coherence; steering pulls the generated sequence toward private platypus sequences without driving Evo 2 into frameshifts or premature truncation.

Looking at Figure 7, we see that the trade-off between private bp gain and aa-identity loss increases across conservation strata, and becomes even clearer when we test steering with more strength by increasing (Figure 8). In the most diverged stratum (stratum zero), private nucleotide recovery rises from 39.81%

unsteered to 44.03% at = 4 while amino acid identity drifts from 17.17% to 14.77%. In the most conserved stratum (stratum four), private nucleotide recovery climbs further, from 31.45% to 41.63%, but amino acid identity falls from 36.43% to 19.03%. More conserved genes get more platypus-like at the nucleotide level, but less platypus-like on the protein level.

Together, these results show that steering produces a conservation-dependent shift in the generated sequence, but the opposing nucleotide- and protein-level responses leave open what features the vector is actually driving the generated sequence towards.

What are we steering toward?

Because the steering vector is built from the contrast between platypus and human representations, a gain in private nucleotide recovery is ambiguous: Generation could be moving toward platypus, away from human, or both. To separate these, we scored the generated nucleotides at sites where human and platypus differ after alignment. Let , , and denote the human, platypus, and generated base at a given site, and restrict attention to the "private sites," where the platypus base is unique relative to the other 23 mammals and the generated sequence has aligned coverage. We determined two variables as a function of steering strength:

- **Leave-human rate ( ):** the fraction of covered private sites where
- **Platypus choice among departures ( ):** among sites where , the fraction where

Intuitively, measures how strongly steering pushes generation away from human nucleotides, while measures whether those departures match the platypus sequence. "Toward platypus" steering would raise and ; purely "away from human" steering would raise while remained constant. We calculate these statistics for both the steered and unsteered versions of each gene, and take the difference to determine and as a result of our intervention.

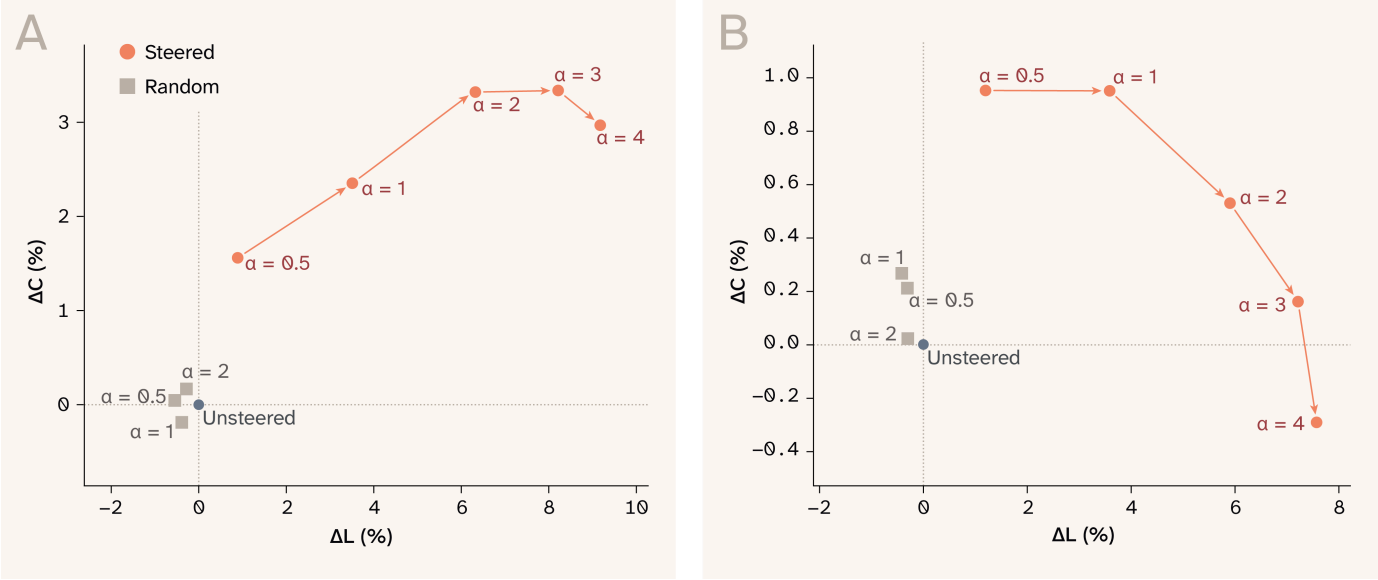

Figure 9. **The steering vector pushes sequence generation both away from human and toward platypus.**

At sites where the aligned human and platypus sequences differ, rate , measures the fraction of generated bases that depart from the human nucleotide, while platypus choice among departures, , measures the fraction of those departures that select the platypus nucleotide.

(A) At private platypus sites, increases monotonically with steering strength, while initially increases and saturates around – before declining at .

(B) At non-private human–platypus mismatch sites, increasing steering does not produce the same increase in platypus choice. Random controls use a per-gene Gaussian direction rescaled to the same norm as the steering vector; the same random direction is reused across the dose ladder for each gene. The norm-matched random directions do not reproduce the same response, indicating that the effect depends on the direction of intervention rather than injection magnitude alone.

Steering pushes generation both away from human and toward platypus

At platypus private sites, both signals responded to steering, but in different ways (Figure 9). Platypus choice among departures ( ) generally rises with , increasing until = 2–3 before decreasing at the strongest setting (= 4). This suggests that, after a point, a steering with increased strength becomes detrimental to sequence generation. The leave-human rate ( ) increases monotonically with steering strength. Because will increase when any of the three non-human nucleotides are generated, whereas requires selection of the specific platypus nucleotide, their absolute ranges are not directly comparable. But together, the two metrics suggest that steering increases both departure from the human sequence and preferential selection of platypus alleles among those departures for low to moderate values. The norm-matched random-direction controls help distinguish

this result from the nonspecific effect of perturbing the residual stream by the same magnitude.

An additional distinction appears when we separate different classes of human–platypus differences. At private sites (Figure 9, A), increasing steering produces a clear rise in platypus choice among departures, whereas at non-private mismatch sites (Figure 9, B) (where the platypus nucleotide is shared with one or more of the other 22 mammals) we do not see the same dose-dependent preference for the platypus base. This contrast suggests that the effect is not simply a uniform preference for nucleotides found in the platypus sequence, but depends on the evolutionary context of the site.

One possibility is that the steering direction preferentially captures sequence features associated with platypus-lineage-specific substitutions. Another is that private and non-private sites differ systematically in properties such as conservation, codon position, or nucleotide composition, which could make them respond differently to the intervention. Our current analysis doesn't distinguish among these possibilities, but the saturation and eventual decline of at high steering strength provides a useful clue: stronger steering eventually stops making generation more specifically platypus-like. We therefore next examined the nucleotide composition of the generated sequences to understand what the model is producing as the intervention becomes stronger.

What the model generates

Further investigation into what tokens the model is predicting may provide us with clues about the information stored in the steering vector. A few strong patterns are apparent in Figure 10. GC content rose with steering strength at every codon position, but was consistently fastest at the wobble position: GC3 went from 60.68% unsteered to 90.46% at versus GC1's 58.21% to 82.71% and GC2's 47.88% to 73.08%. Median GC3 started above the platypus CDS level (66.67%) once we began steering and continued to increase, and overall GC content scaled monotonically with .

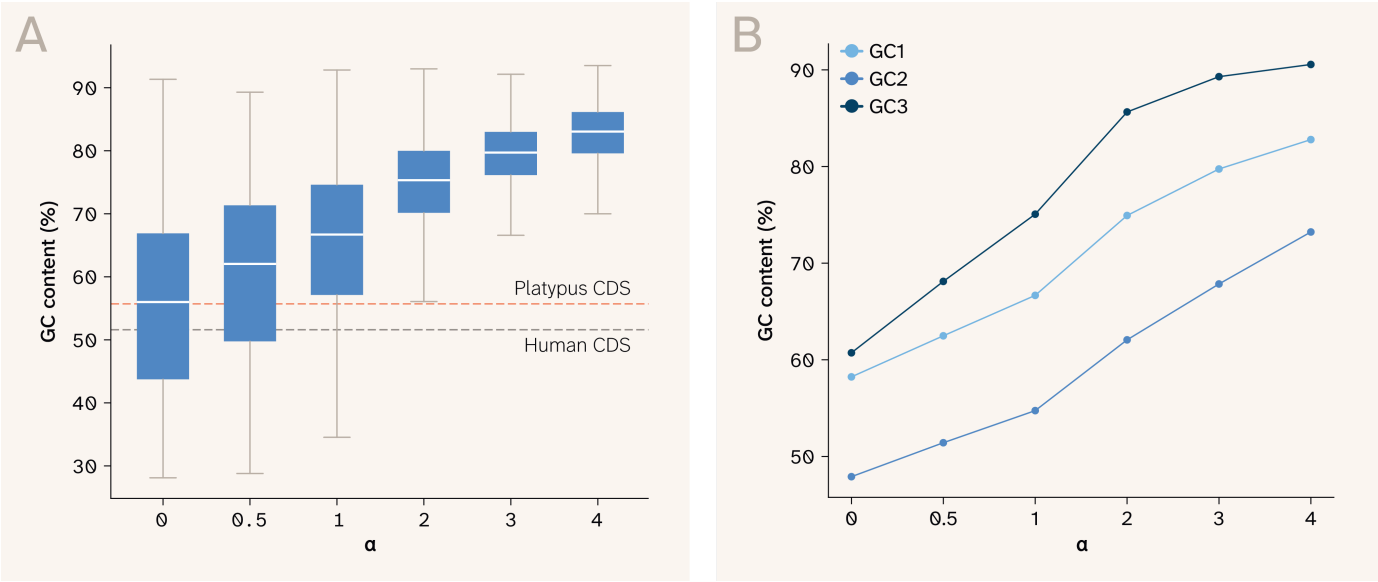

Figure 10. **Increased steering strength produces a high-GC compositional shift, and is strongest at third codon positions.**

(A) Distribution of GC3 content across steering strengths, shown relative to human (grey) and platypus (orange) CDS reference levels.

(B) GC content at the first, second, and third codon positions as steering strength, increases. GC rises at every codon position but most rapidly at the wobble position (GC3). Results indicate that strong steering increasingly produces a generic compositional shift rather than movement toward the natural platypus nucleotide distribution.

Pub preparation

We used Claude (Opus 4.8 and Opus 5) and ChatGPT (GPT-5.6 Sol) to help write, clean up, comment, and review our code. We also used these models to suggest wording ideas, write text, rearrange text to fit the structure of one of our pub templates, expand on summary text, copy-edit draft text to match Arcadia's style, clarify and streamline text, brainstorm project and experimental design ideas, and to suggest papers on relevant science. We also transcribed a presentation of this project and used Claude Opus 5 to create a first draft based on one of our pub templates. We reviewed all AI-assisted content and take responsibility for its accuracy and integrity.

We used arcadia-pycolor (v0.8.0) [13] to generate figures before manual adjustment in Adobe Illustrator.

Key takeaways

Across both experiments, we repeatedly found evidence for two sides of Evo 2's biological representations. Gene-family geometry was strongly related to simple nucleotide statistics, but in the middle layers it aligned most strongly with a protein-family baseline. Our human-to-platypus steering vector produced a broad compositional shift, but also preferentially increased "private" platypus nucleotides over controls and unsteered sequences. Steering altered nucleotide choice and amino acid identity, but did not substantially increase indels or premature stops, and the strongest GC response occurred at the third codon position. Our results consistently indicate that coarse sequence patterns and more abstract biological relationships are frequently combined in Evo 2 representations.

These results answer our two main questions with strongly supported yeses. For question 1, Evo 2 organizes genes in ways that align with known biological and evolutionary relationships, although that organization is layer-dependent, family-dependent, and strongly related to coarse sequence features. For question 2, its species representations can also be used causally during generation, but a general species direction produced from our current methods nudges sequences toward a target taxon rather than reconstructing gene-specific orthologs.

More broadly, we think these results are evidence against treating sequence statistics and higher-order biological representation as exclusive categories. Nucleotide composition, short motifs, codon constraints, and species-associated substitutions are both products and carriers of evolutionary information. A model may combine simple signals into a representation that behaves like a complex biological concept, without representing each component separately or reconstructing the full biological process that produced it.

With this in mind, we propose that compositional controls should not only serve as null models that invalidate a biological result when they perform well. They can be used to map what a representation is built from: which low-level signals carry enough information to reproduce a behavior, where additional organization appears across model layers, which components can be manipulated causally, and where a compressed representation lacks the detail needed for biological fidelity. Those distinctions are important when deciding what an embedding can reasonably be used to infer, when designing benchmarks that claim to measure

biological understanding, and when choosing representations or layers for generative experiments.

Next steps

Our pub represents a taxonomically focused preliminary investigation into Evo 2's latent space to reveal the different evolutionary and biological concepts learned by the model. Two natural next steps revolve around improving taxonomic breadth and building on our steering approach.

Extending the taxonomic scope of the study by including additional taxa (particularly out-of-distribution non-model species that are not in the training data) will help us assess whether the learned representation of the evolutionary distances is continuous in more and less closely related species. Additionally, more comprehensively sampling the true species tree helps us better capture the true extent of sequence variation, allowing us to determine model sequence accuracy and the degree to which alleles are unique to our target taxon with greater precision.

Future efforts to build on our steering approach should aim to minimize the source-species prior, apply steering more selectively across positions and layers, and probe more directly what it would take to sample a true species ortholog rather than a sequence merely nudged by species-level statistics. This will help overcome our current limitations that stem from uniformly injecting a single vector, which inadvertently entangles the target-species signal with a push away from the source sequence and taxon.

Weigh in!

We hope you'll use our methods as a starting point for your own steering experiments and let us know how it goes. We'd also appreciate any feedback on our methodology or useful experiments that we're missing and haven't covered in "Next steps." Feel free to leave thoughts or questions in a comment on the pub!

Contributors (A–Z)

- **Audrey Bell**: Visualization
- **Rishi De-Kayne**: Critical feedback, Editing, Experimental design
- **Rachael DeVries**: Conceptualization, Data curation, Experimental design, Formal analysis, Investigation, Methodology, Software, Visualization, Writing
- **James R. Golden**: Conceptualization, Supervision, Validation
- **Megan L. Hochstrasser**: Editing
- **Erin McGeever**: Experimental design
- **Robert Roth**: Resources
- **Ryan York**: Supervision

References

1. Seal RL, Braschi B, Gray K, McClay J, Tweedie S, Bruford EA. (2025). Genenames.org: the HGNC and PGNC resources in 2026. https://doi.org/10.1093/nar/gkaf1229
2. Pearce M., Simon E., Byun M., Balsam D. (2025). Finding the tree of life in Evo 2. https://www.goodfire.com/research/phylogeny-manifold
3. Candido S, Hayes T, Derry A, Rao R, Lin Z, Verkuil R, Wu BZ, Lee JS, Bruguera ES, Keval JA, Kopylov M, Pak JE, Wu W, Thomas N, Mataraso S, Hsu A, Trotman-Grant AC, Fatras K, dos Santos Costa A, Badkundri R, Akın H, Oktay D, Deaton J, Montabana E, Sitwala H, Yu Y, Wiggert M, Carlin DA, Goering AW, Blazejewski T, Sandora M, Hla M, Jia TZ, Kloker LH, Sofroniew NJ, Uehara M, Pannu J, Bachas S, Liu DS, Sercu T, Rives A. (2026). Language Modeling Materializes a World Model of Protein Biology. https://doi.org/10.64898/2026.06.03.729735
4. Simon E, Zou J. (2025). InterPLM: discovering interpretable features in protein language models via sparse autoencoders. https://doi.org/10.1038/s41592-025-02836-7
5. Zhang Z, Wayment-Steele HK, Brixi G, Wang H, Kern D, Ovchinnikov S. (2024). Protein language models learn evolutionary statistics of interacting sequence motifs. https://doi.org/10.1073/pnas.2406285121
6. Parsan N, Yang DJ, Yang JJ. (2025). Towards Interpretable Protein Structure Prediction with Sparse Autoencoders. https://doi.org/10.48550/arxiv.2503.08764
7. Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.

https://doi.org/10.1038/nbt.3988

8. Arcadia Science. (2024). arcadia-pycolor. https://github.com/arcadia-science/arcadia-pycolor
9. Upham NS, Esselstyn JA, Jetz W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. https://doi.org/10.1371/journal.pbio.3000494
10. Ali S. (2026). Causal dictionary learning reveals and validates transcription-factor binding features in genomic language models. https://doi.org/10.48550/arxiv.2607.19618
11. Maiwald A, Jedryszek P, Draye F, Schölkopf B, Morris GM, Crook OM. (2025). Decode-gLM: Tools to Interpret, Audit, and Steer Genomic Language Models. https://doi.org/10.1101/2025.10.31.685860
12. Adams E, Bai L, Lee M, Yu Y, AlQuraishi M. (2025). From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models. https://doi.org/10.1101/2025.02.06.636901
13. Brixi G, Durrant MG, Ku J, Naghipourfar M, Poli M, Sun G, Brockman G, Chang D, Fanton A, Gonzalez GA, King SH, Li DB, Merchant AT, Nguyen E, Ricci-Tam C, Romero DW, Schmok JC, Taghibakhshi A, Vorontsov A, Yang B, Deng M, Gorton L, Nguyen N, Wang NK, Pearce MT, Simon E, Adams E, Amador ZJ, Ashley EA, Baccus SA, Dai H, Dillmann S, Ermon S, Guo D, Herschl MH, Ilango R, Janik K, Lu AX, Mehta R, Mofrad MRK, Ng MY, Pannu J, Ré C, St. John J, Sullivan J, Tey J, Viggiano B, Zhu K, Zynda G, Balsam D, Collison P, Costa AB, Hernandez-Boussard T, Ho E, Liu M, McGrath T, Powell K, Pinglay S, Burke DP, Goodarzi H, Hsu PD, Hie BL. (2026). Genome modelling and design across all domains of life with Evo 2. https://doi.org/10.1038/s41586-026-10176-5