Generating the Unheard: Phylogeny-Guided Latent Generation for Ancestral Sound Reconstruction

Tianyi Xu[1,3] **Shrinaath Narasimhan**[1] **Evan Gorstein**[1] **Santiago Perea**[1]
Yunyi Shen[2] **Claudia Solís-Lemus**[1*]
[1]University of Wisconsin–Madison
[2]Massachusetts Institute of Technology
[3]Oregon State University

Abstract

What did an ancestral bird species sound like? Existing ancestral state reconstruction methods can infer low-dimensional traits such as morphological characters at internal nodes of a phylogenetic tree, but no one has tried to produce rich perceptual signals such as audio. Some of the challenges include inferred representations that are either too low-dimensional to decode or lie in non-generative feature spaces, so no method to date can produce ancestral audio. We introduce the first framework that generates plausible ancestral vocalizations. Our pipeline encodes bird recordings into a VAE latent space, learns a low-dimensional trait projection aligned with phylogenetic distances, performs ancestral inference in this trait space, and recovers decodable latents through an anchored inverse lift before emitting novel waveforms for each ancestral node. Because the entire pipeline stays within a decodable latent space, every internal node receives a genuinely new audio output representing plausible intermediate ancestral sounds unavailable to retrieval-based alternatives. Experiments on two phylogenetically distant bird clades, 21 species sampled from Tyrannidae and 19 from Paridae, show that our method is the only approach that simultaneously achieves genuine generation, phylogenetic consistency, and naturalistic audio quality across both datasets. Code is available at `https://github.com/solislemuslab/bird-sound-asr`.

1 Introduction

Ancestral state reconstruction (ASR) is a core tool in evolutionary biology [1, 2], but existing methods are primarily designed for low-dimensional symbolic or continuous traits, such as morphology, size, or discrete character states [3, 4]. Many biologically important signals, however, are high-dimensional and perceptual. Bird vocalizations are a prime example: as spectro-temporal signals, their representations require hundreds to thousands of dimensions, yet comparative studies show they carry measurable phylogenetic signal even in vocal learners [5] and that acoustic divergence correlates with evolutionary distance at macroevolutionary scales [6, 7]. Reconstructing what an ancestral species sounded like therefore requires going beyond classical trait inference (Figure 1). Bird vocalizations also span a fundamental divide in vocal inheritance: suboscine species produce innate, genetically determined calls [8], while oscine species incorporate culturally learned song elements [9], carrying different degrees of phylogenetic signal [5]. Moreover, closely related species pairs—such as the cryptic *Empidonax* flycatchers (*E. traillii* and *E. alnorum*), which are visually near-identical but distinguished primarily by voice—illustrate that vocal divergence can be a primary axis of speciation [10, 11], making ancestral sound reconstruction directly relevant to understanding how species form.

*Corresponding author: Claudia Solís-Lemus. Correspondence: `xut2@oregonstate.edu`, `solislemus@wisc.edu`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

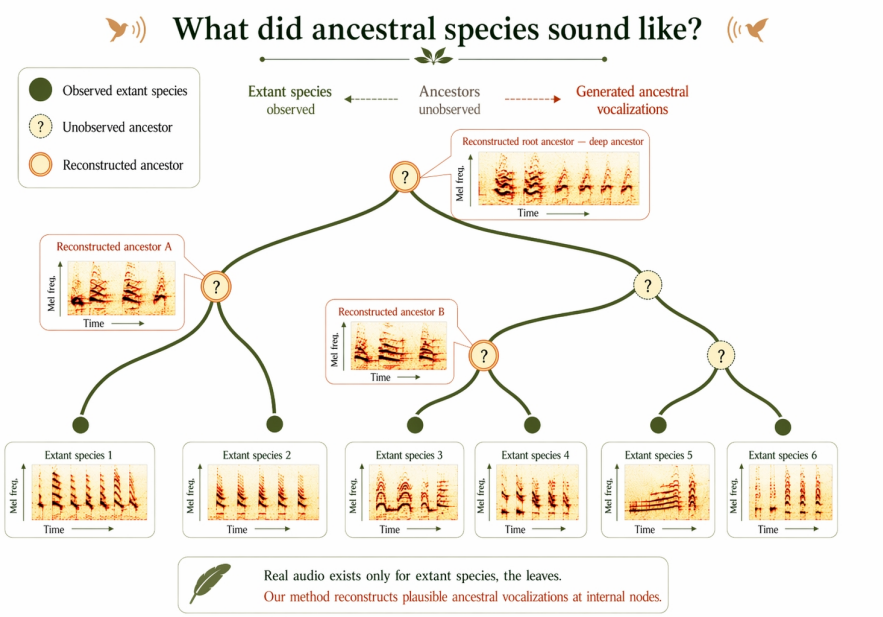


Figure 1: **Motivation.** Extant species at tree leaves have observed vocalizations (mel spectrograms); internal nodes correspond to unobserved ancestors. Our method reconstructs ancestral vocalizations by performing evolutionary inference in a learned latent trait space and decoding back to audio, producing novel spectrograms consistent with descendant species.

Existing approaches operate in feature spaces such as acoustic descriptors [7], classifier embeddings [12], or learned taxonomic representations, none of which are *decodable* back to audio. No existing method can therefore produce ancestral audio. The closest alternative is to select the most similar extant recording as a proxy, which cannot produce sounds intermediate between descendants or represent vocalizations that no extant species exhibits. The most natural generative alternative, performing Brownian motion ancestral inference [3] directly in a VAE latent space [13], also falls short: as we show in Section 4, this naive baseline yields noisy, off-manifold spectrograms (Figure 3).

We bridge this gap with a four-stage pipeline: (1) encode extant vocalizations into a pretrained VAE latent space; (2) learn a tree-metric trait projection aligned with phylogenetic distances; (3) perform Brownian ancestral inference in this trait space and recover full decodable latents via an anchored inverse lift; (4) emit novel waveforms through the VAE decoder and a vocoder. The key insight is that the entire pipeline stays within a decodable latent space while preserving phylogenetic signal, so every ancestral node receives a genuinely new audio output.

On two phylogenetically distant clades, 21 species sampled from Tyrannidae and 19 from Paridae, our method generates a completely unique waveform for every ancestor while maintaining the strongest phylogenetic plausibility and producing naturalistic spectrograms. We also evaluate trait inference using recordings from held-out species and a historical song archive spanning several decades. Our contributions:

- The first generative framework for ancestral sound reconstruction, producing novel waveforms at every internal node of a phylogeny rather than selecting existing recordings.
- A latent-space decomposition separating phylogenetically informative variation from acoustic texture, enabling evolutionary inference in a compact trait subspace with a closed-form inverse lift back to decodable audio.
- Cross-clade generalization across the suboscine–oscine divide: innate vocalizations in Tyrannidae and culturally learned songs in Paridae.
- Evaluation protocols and metrics for ancestral audio generation, a task where ground-truth recordings are unavailable for deep ancestral nodes.

2 Related Work

Ancestral state reconstruction. Classical ASR methods infer ancestral values for low-dimensional traits via maximum likelihood [1, 2] or Bayesian estimation [14, 4], assuming Brownian motion [3] or Ornstein–Uhlenbeck [15, 16] trait evolution. These methods are effective for scalar morphological measurements [17] but do not extend to high-dimensional perceptual signals where $D \gg n$. Recent work couples generative models with phylogenetic priors: Draupnir [18] uses a tree-structured OU process as a VAE prior for ancestral protein reconstruction, and Phylo-Diffusion [19] conditions diffusion models on tree-of-life hierarchical embeddings to generate images of hypothetical ancestors. Our work differs by operating on audio, using a pretrained generative model and separately learned trait projection, and introducing an anchored inverse lift for the trait-to-latent inverse.

Phylogenetic representation learning. Several lines of work learn representations incorporating phylogenetic structure: hyperbolic embeddings for hierarchies [20–22], Siamese Poincaré networks for reconstructing evolutionary trees from spectrograms [23], phylogenetic augmentation for contrastive representation learning [24], and phylogenetic PCA [25]. Mitteroecker et al. [26] find linear subspaces maximizing Blomberg's K [27] via eigendecomposition, sharing our goal of phylogenetically informative projections. PhyloNN [28] learns phylogenetically supervised embeddings via hierarchical classification losses. Our trait learner differs from these approaches in optimizing within a latent space with a patristic distance correlation objective and an orthonormality constraint that enables a closed-form inverse lift. Separately, Li et al. [29] decompose VAE latent spaces into attribute-relevant and residual subspaces for image editing. Our row-space and nullspace decomposition serves an analogous role but is learned from phylogenetic distances for ancestral reconstruction.

Neural audio generation and computational bioacoustics. Modern audio synthesis combines VAE latent spaces [13] with neural vocoders [30, 31]; diffusion and text-conditioned models [32–34] further advance quality, but none incorporate evolutionary priors. In bioacoustics, deep learning targets classification, including BirdNET [12], PANNs [35], and CLAP [36], as well as representation learning via self-supervised [37] and embedding-based [38] approaches. More broadly, invariance-driven speech representations [39] demonstrate that learned audio embeddings can disentangle biologically relevant factors from nuisance variation. VAE latent spaces capture meaningful vocal structure [40], and comparative studies examine vocal evolution using acoustic features [7, 41] but remain in non-decodable feature spaces. We bridge audio generation and phylogenetic inference: a pretrained VAE and vocoder serves as a fixed synthesis backend, while our tree-metric trait pipeline produces ancestral vocalizations for the first time.

3 Method

3.1 Setup and Notation

Let $T = (\mathcal{V}, \mathcal{E})$ be a rooted phylogenetic tree with positive branch lengths $\{w_e\}_{e\in\mathcal{E}}$. The leaf set $\mathcal{L} \subset \mathcal{V}$, $|\mathcal{L}| = n$, indexes extant species; the internal nodes $\mathcal{U} = \mathcal{V} \setminus \mathcal{L}$ are unobserved ancestors. For each $s \in \mathcal{L}$ we observe a collection of audio clips $\{a_{s,i}\}_{i=1}^{n_s}$. Denote by $d_T(s,t) = \sum_{e\in \mathrm{path}(s,t)} w_e$ the patristic distance between $s, t \in \mathcal{V}$ and by $\mathrm{depth}(v) = d_T(\mathrm{root}, v)$ the root-to-v distance for $v \in \mathcal{V}$. The goal is to produce, for each $u \in \mathcal{U}$, a waveform realization consistent with the tree topology, branch lengths, and the extant vocalization distribution.

We assume access to a pretrained audio VAE with encoder E_{vae}, decoder D_{vae}, and a neural vocoder V. Each clip is encoded as $x_{s,i} = \mathrm{vec}(E_{\mathrm{vae}}(a_{s,i})) \in \mathbb{R}^D$, and the composition $V \circ D_{\mathrm{vae}}$ maps any latent vector back to a waveform. Our pipeline operates in four stages, *encode*, *evolve*, *lift*, and *emit*, detailed in Sections 3.2–3.4 and illustrated in Figure 2.

3.2 Tree-Metric Trait Space

We assume that bird vocalizations carry phylogenetic signal and learn a compact trait space whose distances reflect the evolutionary relationships among species. Direct phylogenetic inference in $\mathbb{R}^D$ is challenging: although the $n \times n$ BM covariance is cheap to invert, running independent coordinate-wise inference across $D \gg n$ dimensions produces samples that lie off the decodable data manifold, yielding noisy audio (Figure 3), and generic dimensionality reduction (e.g. PCA) does not preserve evolutionary geometry. We instead learn a linear projection $W \in \mathbb{R}^{K\times D}$, $K \ll D$, that

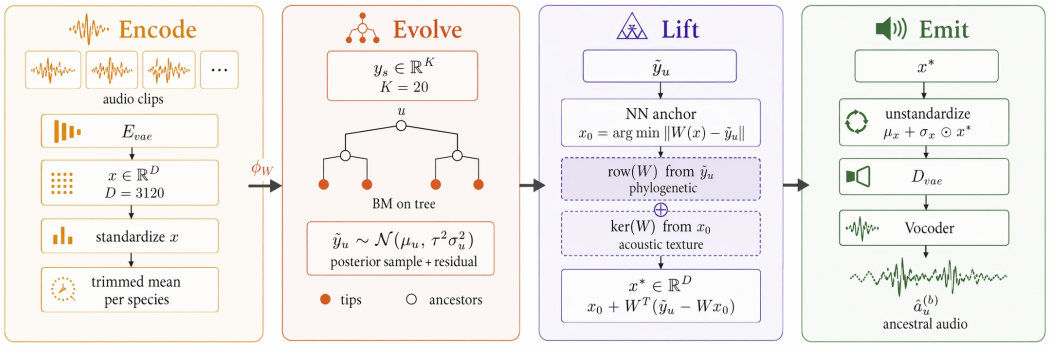


Figure 2: **Pipeline overview.** Audio clips are encoded into flattened VAE latents $x \in \mathbb{R}^D$ (*Encode*), projected via W into a K-dimensional trait space $y \in \mathbb{R}^K$ where Brownian-motion ancestral state reconstruction yields posterior trait vectors $\tilde{y}_u$ for each internal node u (*Evolve*), lifted back to $\mathbb{R}^D$ by combining the phylogenetic component $\mathrm{row}(W)$ from $\tilde{y}_u$ with the acoustic texture $\ker(W)$ from an anchor tip x_0 (*Lift*), and decoded through the VAE decoder and vocoder to produce ancestral audio $\hat{a}_u^{(b)}$ (*Emit*).

defines a *trait representation*

$$\phi_W(x) \coloneqq Wx. \tag{1}$$

The row-orthonormality constraint on W (defined below) ensures that directions in $\mathbb{R}^D$ contribute on a comparable scale after projection.

Objective. Let $\bar{x}_s = n_s^{-1} \sum_i x_{s,i} \in \mathbb{R}^D$ denote the mean latent for species s. We optimize W to maximize the Pearson correlation between pairwise squared trait distances and patristic distances:

$$\max_{W \in \mathcal{W}} \rho\Big(\{d_T(s,t)\}_{s<t},\ \{\|\phi_W(\bar{x}_s) - \phi_W(\bar{x}_t)\|_2^2\}_{s<t}\Big), \tag{2}$$

where ρ denotes the Pearson correlation over all $\binom{n}{2}$ unordered species pairs and $\mathcal{W} = \{W \in \mathbb{R}^{K \times D} : WW^\top = I_K\}$ is the Stiefel manifold $\mathrm{St}(K, D)$. The use of squared Euclidean distances rather than raw distances is motivated by the fact that under Brownian motion on the tree, $\mathbb{E}[\|Y_s - Y_t\|^2]$ is linear in $d_T(s,t)$ when each coordinate evolves as an independent unit-rate Wiener process.

Orthonormality via Cholesky reparametrization. We constrain W to have orthonormal rows. Direct projection onto the Stiefel manifold after each gradient step can be numerically unstable for $K \ll D$ [42, 43]. Instead, we introduce an unconstrained parametrization through a free matrix $A \in \mathbb{R}^{K \times D}$ and define

$$W = L^{-1}A, \qquad LL^\top = AA^\top + \varepsilon I_K, \tag{3}$$

where $L \in \mathbb{R}^{K \times K}$ is the (unique) lower-triangular Cholesky factor and $\varepsilon > 0$ is a small regularizer for numerical stability.

Proposition 3.1 (Exact row-orthonormality)**.** *Let $A \in \mathbb{R}^{K \times D}$ with $\mathrm{rank}(A) = K$, and let L be the lower-triangular Cholesky factor of $AA^\top + \varepsilon I_K$. In the limit $\varepsilon \to 0$, $W = L^{-1}A$ satisfies $WW^\top = I_K$. For any fixed $\varepsilon > 0$, the deviation satisfies $\|WW^\top - I_K\|_{\mathrm{op}} \le \varepsilon / \sigma_{\min}^2(A)$ (proof in Appendix D).*

Gradients of the objective (2) flow through A via the Cholesky decomposition. After each optimizer step, W is recomputed from (3). The row-orthonormality constraint serves two purposes: (i) since the Pearson correlation objective is scale-invariant, it removes scale degeneracy, ensuring uniqueness up to rotation and preventing rank collapse; (ii) it enables the closed-form anchored inverse lift in Section 3.4, which requires $WW^\top = I_K$. A residual $K \times K$ orthogonal non-identifiability remains, but this is harmless: the squared distances in (2) and the row-space/nullspace decomposition in Section 3.4 are both invariant to such rotations (Appendix D.6).

Species aggregation. After W is learned using simple-mean representatives $\bar{x}_s$, we form robust species-level tip traits for downstream BM inference via a *trimmed mean* in trait space. For each $s \in \mathcal{L}$, compute clip-level traits $z_{s,i} = \phi_W(x_{s,i})$ and the preliminary mean $\bar{z}_s^{(0)} = n_s^{-1} \sum_i z_{s,i}$, then retain the closest fraction $1 - \alpha$ of clips:

$$y_s = \frac{1}{|\mathcal{I}_s^{\mathrm{keep}}|} \sum_{i \in \mathcal{I}_s^{\mathrm{keep}}} z_{s,i}, \qquad \mathcal{I}_s^{\mathrm{keep}} = \big\{i \in [n_s] : \|z_{s,i} - \bar{z}_s^{(0)}\|_2 \le \delta_s^{(\lceil (1-\alpha) n_s \rceil)}\big\}, \tag{4}$$

where $\delta_s^{(k)}$ denotes the k-th smallest value of $\{\|z_{s,i} - \bar{z}_s^{(0)}\|_2\}_{i=1}^{n_s}$ and $\alpha \in [0, 1)$ controls robustness to outlier clips. These $\{y_s\}_{s\in\mathcal{L}}$ serve as the observed tip traits for phylogenetic inference.

3.3 Brownian Ancestral Inference

In this trait space, Brownian motion models the accumulation of variation along branches, allowing ancestral traits to be inferred from the observed species. We model trait evolution on the tree T as a K-dimensional Brownian motion (BM). Under this model, each coordinate evolves independently as a unit-rate Wiener process along the branches of T, inducing a joint Gaussian distribution over all node traits with covariance structure determined by shared evolutionary history.

Covariance structure. For any two nodes $i, j \in \mathcal{V}$, the BM covariance equals the depth of their most recent common ancestor:

$$C_{ij} = \mathrm{depth}\big(\mathrm{MRCA}(i, j)\big). \tag{5}$$

In particular, $C_{ii} = \mathrm{depth}(i)$ gives the variance at node i. Since $C_{ij} = \frac{1}{2}(C_{ii} + C_{jj} - d_T(i, j))$ by the tree metric identity, the matrix C is determined by the tree topology and branch lengths.

Denote the tip-tip submatrix of this $|\mathcal{V}| \times |\mathcal{V}|$ covariance by $C_{\mathcal{L}\mathcal{L}} \in \mathbb{R}^{n\times n}$. To account for within-species noise and finite-sample effects, we add a nugget term:

$$\Sigma_{\mathcal{L}\mathcal{L}} = C_{\mathcal{L}\mathcal{L}} + \lambda I_n, \qquad \lambda > 0. \tag{6}$$

Trait standardization. Before conditioning, we standardize tip traits coordinate-wise to zero mean and unit variance ($\bar{y}_s = (\hat{\sigma}_y)^{-1} \odot (y_s - \hat{\mu}_y)$), making the unit-rate BM assumption reasonable without estimating a per-coordinate diffusion rate σ_k^2.

Proposition 3.2 (BM conditional posterior)**.** *Under the BM model with covariance* (5) *and nugget* (6)*, the posterior distribution of the trait at internal node $u \in \mathcal{U}$, conditioned on standardized tip traits $\bar{Y}_{\mathcal{L}} \in \mathbb{R}^{n\times K}$, factorizes across coordinates. For each coordinate $j \in \{1, \ldots, K\}$:*

$$\bar{Y}_{u,j} \mid \bar{Y}_{\mathcal{L},j} \;\sim\; \mathcal{N}\big(c_u^\top \Sigma_{\mathcal{L}\mathcal{L}}^{-1} \bar{Y}_{\mathcal{L},j},\;\; C_{uu} - c_u^\top \Sigma_{\mathcal{L}\mathcal{L}}^{-1} c_u\big), \tag{7}$$

where $c_u \in \mathbb{R}^n$ is the cross-covariance vector with entries $(c_u)_i = C_{ui}$ for $i \in \mathcal{L}$, and $C_{uu} = \mathrm{depth}(u)$. This follows from standard Gaussian conditioning on the multivariate normal induced by BM on a tree [3, 44, 4]. A self-contained derivation is in Appendix D.

Because trait standardization maps all coordinates to unit variance, the posterior variance $\sigma_u^2 = C_{uu} - c_u^\top \Sigma_{\mathcal{L}\mathcal{L}}^{-1} c_u \geq 0$ is a scalar shared across all K coordinates. Write $m_u \in \mathbb{R}^K$ for the posterior mean in standardized space. Denote the unstandardized posterior mean $\hat{y}_u = \hat{\mu}_y + \hat{\sigma}_y \odot m_u$. We draw B posterior samples with a temperature parameter $\tau > 0$ controlling diversity:

$$y_u^{(b)} = \hat{y}_u + \tau\, \sigma_u\, (\hat{\sigma}_y \odot \epsilon_1), \qquad \epsilon_1 \sim \mathcal{N}(0, I_K), \quad b = 1, \ldots, B. \tag{8}$$

Setting $\tau < 1$ concentrates samples around the posterior mean. Because the BM covariance is computed at the raw tree-depth scale while the traits are standardized to unit variance, σ_u^2 absorbs an implicit diffusion rate. The parameter τ therefore acts as a joint temperature and rate-calibration factor rather than a pure posterior scaling, and is tuned on downstream quality. The element-wise rescaling by $\hat{\sigma}_y$ maps the noise back to the original trait scale.

Residual perturbation. Aggregation and posterior averaging suppress within-species variation. To restore clip-level diversity, we perturb each posterior sample with calibrated noise:

$$\tilde{y}_u^{(b)} = y_u^{(b)} + \gamma\, \hat{\sigma}_{\mathrm{res}} \odot \epsilon_2, \qquad \epsilon_2 \sim \mathcal{N}(0, I_K), \tag{9}$$

where $\hat{\sigma}_{\mathrm{res}} \in \mathbb{R}^K$ is the pooled within-species residual standard deviation and $\gamma \geq 0$ controls perturbation strength. The two noise sources are complementary: τ scales node-dependent phylogenetic uncertainty via σ_u^2, while γ restores within-species acoustic variability uniformly. This preserves the BM posterior mean (Proposition C.1, Appendix C).

3.4 Anchored Inverse Lift

Many latent vectors can share the inferred traits, so we use a nearby recording to supply the remaining acoustic detail for a plausible ancestral sound. Since $K \ll D$, the trait map $\phi_W : \mathbb{R}^D \to \mathbb{R}^K$ is many-to-one. Each target trait $y \in \mathbb{R}^K$ determines a $(D-K)$-dimensional affine preimage

$$\mathcal{A}(y) = \{x \in \mathbb{R}^D : Wx = y\} = x_p + \ker(W), \tag{10}$$

where $x_p = W^\top y$ is a particular solution (using $WW^\top = I_K$). Recovering a unique decodable latent from y is therefore an underdetermined linear inverse problem. Naïve approaches, such as the minimum-norm solution $x_p = W^\top y$, produce latents with zero nullspace component, which lie far from the data manifold and yield poor audio quality. We instead resolve the non-identifiability by *anchoring* the inverse with a nearby observed latent.

Anchor selection. Let $\mathcal{X}_{\text{train}} = \{x_{s,i}\}_{s\in\mathcal{L},\, i\in[n_s]}$ denote the full set of encoded clips across all species. Given a target (perturbed) ancestral trait $\tilde{y}_u^{(b)}$, we select the training clip whose trait embedding is nearest in ℓ_2 distance:

$$x_0 = \arg\min_{x\in\mathcal{X}_{\text{train}}} \|\phi_W(x) - \tilde{y}_u^{(b)}\|_2. \tag{11}$$

This anchor provides a realistic nullspace component: the acoustic texture of a real clip that is phylogenetically close to the target ancestor.

Constrained projection. We seek the latent $x^\star$ closest to the anchor x_0 subject to the trait constraint:

$$x^\star = \arg\min_{x\in\mathbb{R}^D} \|x - x_0\|_2^2 \qquad \text{subject to} \qquad Wx = \tilde{y}_u^{(b)}. \tag{12}$$

This is a linearly constrained least-squares problem with a closed-form solution.

Lemma 3.3 (Optimal anchored lift)**.** *Let $W \in \mathbb{R}^{K\times D}$ satisfy $WW^\top = I_K$. For any target trait $y \in \mathbb{R}^K$ and anchor $x_0 \in \mathbb{R}^D$, problem* (12) *has the unique solution (proof in Appendix D):*

$$x^\star = x_0 + W^\top\big(y - Wx_0\big). \tag{13}$$

Corollary 3.4 (Exact trait preservation)**.** *The lifted latent satisfies the target trait exactly: $Wx^\star = y$.*

Corollary 3.5 (Nullspace preservation)**.** *The correction $x^\star - x_0$ lies entirely in $\mathrm{row}(W)$, so the nullspace component is preserved exactly: $P_{\ker(W)}\, x^\star = P_{\ker(W)}\, x_0$ where $P_{\ker(W)} = I_D - W^\top W$.*

The row-orthonormality $WW^\top = I_K$ induces the decomposition $\mathbb{R}^D = \mathrm{row}(W) \oplus \ker(W)$. The anchored lift replaces the *phylogenetic subspace* $\mathrm{row}(W)$ with the BM-inferred target trait while preserving the *acoustic nullspace* $\ker(W)$, which encodes fine-grained spectro-temporal texture from a nearby real clip. This decomposition is by design: the K trait dimensions are precisely those learned to maximize phylogenetic distance correlation (Eq. 2), so they concentrate the evolutionarily informative variation. The remaining $D-K$ nullspace dimensions encode acoustic texture, such as timbre, spectral detail, and temporal micro-structure, that is not expected to carry phylogenetic signal and would be underdetermined by the tree alone. Inheriting this component from a nearby extant recording is analogous to ancestral sequence reconstruction, where informative sites are inferred from the phylogeny while neutral positions are filled from the closest reference. Proofs are in Appendix D.

Decoding. The lifted latent $x^\star$ is decoded through the pretrained VAE decoder and neural vocoder. For an internal node u, the complete ancestral sound generator is

$$\hat{a}_u^{(b)} = V\big(D_{\text{vae}}\big(\Psi(\tilde{y}_u^{(b)})\big)\big), \qquad b = 1, \ldots, B, \tag{14}$$

where Ψ denotes the deterministic anchored-lift operator, comprising anchor selection (11) and constrained projection (13). Since Ψ, D_{vae}, and V are all deterministic, all stochasticity in the generated output enters through the upstream BM posterior sampling (8) and the residual perturbation (9). The full pipeline is summarized in Algorithm 1 (Appendix E).

4 Experiments

Datasets. We evaluate on two phylogenetically and geographically distant bird clades:

Table 1: **Phylogenetic plausibility of reconstructed internal-node sounds.** Methods are either *generative* that produce novel audio: Raw Latent BM, Ours or *retrieval-based* that output copies of training clips: BirdNET BM, Acoustic, PhyloNN, Poincaré. Unique Rate < 1 indicates retrieval collapse. *Higher is better* for all metrics.

Method	Tree–Emb. Corr. ↑	Desc. Affinity Margin ↑	Edge Mono. ↑	NN-Train Dist. ↑	Unique Rate ↑
Dataset 1: Tyrannidae (21 species)					
Retrieval-based					
BirdNET Embedding + BM + Decode	0.4677	0.0189	0.7333	0.2147	**1.0000**
Classical Acoustic Features	0.3449	0.0289	0.8000	0.1619	0.7500
PhyloNN	0.2970	**0.0845**	0.8667	0.1713	0.7000
Poincaré	0.2100	0.0063	0.9333	0.1662	0.2500
Generative					
Raw Latent BM	0.3205	0.0035	**1.0000**	**0.3413**	**1.0000**
Full Method (ours)	**0.4859**	0.0417	0.8667	0.2350	**1.0000**
Dataset 2: Paridae (19 species)					
Retrieval-based					
BirdNET Embedding + BM + Decode	−0.2262	−0.0006	0.5833	0.1992	**1.0000**
Classical Acoustic Features	−0.3168	−0.0010	0.5000	0.1726	0.5556
PhyloNN	−0.2033	**0.0491**	0.7500	0.1788	0.5556
Poincaré	−0.0014	0.0114	0.5833	0.1733	0.6111
Generative					
Raw Latent BM	**0.1848**	0.0005	**0.9167**	**0.2089**	**1.0000**
Full Method (ours)	0.1237	0.0023	**0.9167**	0.1742	**1.0000**

- **Dataset 1 (Family Tyrannidae):** 21 New World tyrant flycatcher species spanning 6 genera, namely 9 *Empidonax* spp., 3 *Contopus* spp., 3 *Sayornis* spp., 3 *Tyrannus* spp., 2 *Myiarchus* spp., and 1 *Myiodynastes* sp., with a consensus phylogenetic tree [45] containing 20 internal nodes (Figure 6). Each species has 47–68 vocalization clips, totaling $\sim$1,100 clips. Tyrannidae are suboscines whose vocalizations are innate rather than culturally learned [8], and thus are expected to carry strong phylogenetic signal [5].
- **Dataset 2 (Family Paridae):** 19 tit, chickadee, and titmouse species spanning 6 genera, namely 9 *Poecile* spp., 5 *Baeolophus* spp., 2 *Cyanistes* spp., 1 *Parus* sp., 1 *Periparus* sp., and 1 *Lophophanes* sp., with a consensus phylogenetic tree [45] containing 18 internal nodes (Figure 7). Each species has 50–80 clips, totaling $\sim$1,150 clips. Paridae are oscines whose songs include culturally learned components [9], which allows us to test whether BM-based ancestral inference remains effective when the phylogenetic-distance assumption is weakened by cultural transmission.

All recordings are sourced from Xeno-Canto[2] and standardized to 5-second mono 16 kHz WAV. The two clades differ in geographic origin, vocal structure, and phylogenetic placement. The same pretrained VAE and BigVGAN vocoder are used for both datasets without retraining; only the tree-metric trait projection W and baseline models are fitted per dataset. Full dataset details, tree visualizations, and preprocessing steps are in Appendices G.1 and G.2; details on the pretrained models are provided in Appendix G.3.

Baselines. We compare five baselines spanning two paradigms: *generative* methods decode novel latent vectors to audio, while *retrieval-based* methods output the nearest training clip because their feature spaces are not decodable. **Raw Latent BM** (*generative*): BM directly in VAE latent space (D=3,120) without trait projection. **BirdNET BM** (*retrieval*): BM in pretrained BirdNET embeddings [12]. **Classical Acoustic** (*retrieval*): BM on handcrafted acoustic descriptors. **PhyloNN** (*retrieval*): a neural representation trained with phylogenetic supervision [28]. **Poincaré** (*retrieval*): hyperbolic embeddings of species relationships [23]. All retrieval baselines use nearest-exemplar decoding after BM inference, providing an upper bound on retrieval-based performance. Only Raw Latent BM and our method generate novel audio. All methods share the same tree, the same aggregation procedure applied in each method's own embedding space, and the same VAE–vocoder rendering pipeline (Appendix G).

Model configuration. Our full model learns a K=20 tree-metric trait projection (sensitivity analysis in Appendix A) with trimmed aggregation at α=0.3, BM nugget 10^{-3}, B=3 posterior samples at temperature 0.6, residual perturbation γ=0.25, and nearest-trait anchored lifting. All latents are decoded with the same VAE decoder and BigVGAN vocoder. Full details in Appendix G.

[2] `https://xeno-canto.org`

Table 2: **Audio quality of reconstructed ancestral sounds.** Retrieval-based methods achieve low FAD by construction since their outputs are copies of real clips. The meaningful comparison is between generative methods: Raw Latent BM produces degraded audio (high FAD), while our method achieves FAD comparable to retrieval baselines despite generating entirely novel audio.

		Dataset 1: Tyrannidae					*Dataset 2: Paridae*				
	Method	CycleSim↑	FAD-BN↓	FAD-PANN↓	FAD-CLAP↓	Silence↓	CycleSim↑	FAD-BN↓	FAD-PANN↓	FAD-CLAP↓	Silence↓
Retrieval	BirdNET BM	0.9985	**0.4803**	0.1681	0.6760	0.0476	0.9981	**0.5041**	**0.1860**	**0.5317**	0.0431
	Acoustic	**0.9987**	0.6117	**0.1182**	**0.5983**	0.0154	**0.9984**	0.6653	0.1922	0.5739	0.0108
	PhyloNN	**0.9987**	0.6165	0.1295	0.6800	0.1000	**0.9984**	0.6111	0.2871	0.6111	0.0615
	Poincaré	0.9984	0.7591	0.3388	0.7823	0.0008	0.9979	0.5425	0.2045	0.5532	0.0288
Gen.	Raw Latent BM	0.9915	0.8587	0.6785	1.6511	**0.0003**	0.9920	0.7575	0.6601	1.5330	**0.0000**
	Ours	0.9985	0.6326	0.1744	0.6711	0.0079	0.9983	0.6008	0.2028	0.5927	**0.0000**

Evaluation. We assess plausibility through audio quality and phylogenetic consistency. *Phylogenetic consistency* is measured by Tree–Embedding Correlation, Descendant Affinity Margin, and Edge Monotonicity using BirdNET embeddings of generated and real audio. *Audio quality* is measured by Cycle Similarity, Fréchet Audio Distance [46] in BirdNET/PANNs/CLAP embedding spaces, and Silence Rate. Nearest-Train Distance measures proximity to training recordings, while Unique Rate detects duplicate waveform files across nodes without requiring ancestors to sound different from extant species. Definitions are in Appendix H.

5 Results

Generation vs. retrieval. The central distinction in Table 1 is between methods that *generate* novel audio and those that *retrieve* existing recordings. Three of the four retrieval baselines collapse to Unique Rates of 0.25–0.75 on Dataset 1 and 0.56–0.61 on Dataset 2, meaning a substantial fraction of their outputs are exact copies of one another. Even BirdNET BM, which retrieves distinct clips for each node, still outputs copies of training recordings rather than novel audio. Our method achieves 100% Unique Rate on both datasets while generating entirely novel waveforms.

Phylogenetic plausibility within generative methods. Among generative methods, ours achieves substantially higher TreeCorr on Dataset 1, 0.49 vs. 0.32, and higher Descendant Affinity, 0.042 vs. 0.004. On Dataset 2, Raw Latent BM has slightly higher TreeCorr, 0.18 vs. 0.12, and both tie on Edge Monotonicity at 0.92; however, this comes at the cost of diffuse, noise-like spectrograms with substantially worse FAD scores as shown in Table 2. The lower absolute TreeCorr on Dataset 2 is consistent with culturally learned oscine songs in Paridae carrying weaker phylogenetic signal than innate suboscine vocalizations in Tyrannidae, and with BirdNET being trained predominantly on North American species. Note that BirdNET BM operates in only 1,024 dimensions yet performs poorly, especially on Dataset 2, indicating that the bottleneck is not dimensionality alone but the mismatch between a discriminative classifier embedding and the distance structure assumed by BM. On Dataset 2, all non-latent baselines produce *negative* Tree–Embedding Correlations, indicating that only generative methods preserve phylogenetic structure on this clade. Alternative evolutionary models are compared in Appendix F; BM provides the best overall balance.

Held-out species validation. We evaluate trait inference against real recordings using leave-one-species-out prediction on the consensus phylogeny of each dataset. For each held-out species, BM predicts its representation from the remaining species, and we rank all species by distance to the prediction in each method's representation space. Localization AUC measures the fraction of other species farther from the prediction than the true species, averaged across held-out species, with 1 indicating perfect localization and 0.5 indicating chance. Our learned trait space achieves the highest localization AUC on both datasets, as shown in Table 3. These results support the use of the learned trait space for predicting unobserved vocal traits from related species.

Table 3: **Held-out species validation on consensus phylogenies.** Localization AUC is averaged across held-out species. Higher is better, with chance at 0.5.

	locAUC ↑	
Method	Tyrannidae	Paridae
PCA	0.56	0.70
BirdNET BM	0.54	0.64
Raw Latent BM	0.47	0.65
Ours	**0.781**	**0.734**

Audio quality. Table 2 reports perceptual quality metrics. Retrieval baselines achieve low FAD by construction since their outputs are verbatim copies of real recordings. Among generative methods, Raw Latent BM produces severely degraded audio with FAD-BN of 0.86 and 0.76 and FAD-CLAP of 1.65 and 1.53 on Datasets 1 and 2, while our method achieves FAD-BN of 0.63 and 0.60 and FAD-CLAP of 0.67 and 0.59, comparable to retrieval baselines despite generating completely novel waveforms that do not exist in the training set. On Dataset 1, our generated ancestors achieve a CycleSim of 0.9985, close to the 0.9987 obtained from real recordings passed through the same VAE and vocoder. This indicates comparable VAE round-trip fidelity (Appendix J.3).

Biological validation and robustness. The learned trait spaces have multivariate Blomberg's K values of 1.09 for Tyrannidae and 0.96 for Paridae, close to the Brownian-motion expectation of 1. Held-out localization remains stable across 100 posterior phylogenies, supporting robustness to uncertainty in tree structure and branch lengths. On 19 historical Savannah Sparrow recordings spanning 1980–2012, held-out-year trait prediction achieves a localization AUC of 0.88, compared with 0.66 for both PCA and raw-latent baselines. Additional held-out tests on *Setophaga* warblers and Hylidae treefrogs support the applicability of trait inference across further bird and non-bird clades. Further experimental details and results are provided in Appendix I.

Table 4: **Ablation study on Dataset 1.** Each axis is ablated independently from a base configuration. *Trait space* and *Lift* ablations use the base config with γ=0 to isolate each component's effect. *Residual* and *Trim* ablations then vary γ and α on top of the best trait and lift choices. Full method uses W + NN-trait + γ=0.25 + α=0.3.

		Phylogenetic Plausibility					*Audio Quality*				
Ablation axis	Config	TreeCorr ↑	DescMarg ↑	EdgeMono ↑	NNDist ↑	Unique ↑	CycleSim ↑	FAD-BN ↓	FAD-PANN ↓	FAD-CLAP ↓	Silence ↓
Trait space	PCA-64	0.2450	0.0003	0.9333	0.3717	**1.0000**	0.9933	0.9374	0.7307	1.5836	0.0005
	Random orth. $\mathbf{R}$	0.1811	0.0036	**1.0000**	**0.4413**	**1.0000**	0.9936	1.0118	0.7613	1.9633	**0.0000**
	Learned W (ours)	**0.3403**	**0.0449**	0.8667	0.1877	**1.0000**	**0.9985**	**0.6584**	**0.1745**	**0.6348**	0.0040
Lift	Mean	0.2230	−0.0005	**0.9333**	**0.3648**	**1.0000**	0.9935	0.9384	0.7283	1.5923	**0.0000**
	NN-species	0.3094	0.0079	**0.9333**	0.3541	**1.0000**	0.9934	0.8675	0.6624	1.6282	0.3063
	NN-trait (ours)	**0.3403**	**0.0449**	0.8667	0.1877	**1.0000**	**0.9985**	**0.6584**	**0.1745**	**0.6348**	0.0040
Residual γ	$\gamma = 0$	0.3403	**0.0449**	**0.8667**	0.1877	**1.0000**	**0.9985**	0.6584	0.1745	**0.6348**	**0.0040**
	$\gamma = 0.25$ (ours)	0.4859	0.0417	**0.8667**	0.2350	**1.0000**	**0.9985**	0.6326	0.1744	0.6711	0.0079
	$\gamma = 0.5$	**0.4943**	0.0367	0.8000	**0.2437**	**1.0000**	**0.9985**	**0.5531**	**0.1576**	0.6588	0.0129
Trim α	$\alpha = 0$	0.4810	0.0385	0.8000	**0.2366**	**1.0000**	**0.9985**	0.6343	**0.1738**	0.6833	**0.0078**
	$\alpha = 0.3$ (ours)	**0.4859**	**0.0417**	**0.8667**	0.2350	**1.0000**	**0.9985**	**0.6326**	0.1744	**0.6711**	0.0079
Full method (ours)		**0.4859**	**0.0417**	**0.8667**	**0.2350**	**1.0000**	**0.9985**	**0.6326**	**0.1744**	**0.6711**	**0.0079**

Ablation study. Table 4 ablates each component on Dataset 1. The learned projection W is the most impactful choice, reducing FAD-BN from $\sim$1.0 under PCA or random projection to 0.66. Nearest-trait anchoring is critical: global-mean lift yields FAD-BN 0.94 and nearest-species anchoring produces 31% silence. Residual perturbation at γ=0.25 improves TreeCorr from 0.34 to 0.49 with negligible quality loss, and trimming at α=0.3 improves Edge Monotonicity; trait dimension sensitivity is in Appendix A and trait space interpretability in Appendix B.

Qualitative comparison. Figure 3 illustrates the generation–retrieval divide. Raw Latent BM produces diffuse, noise-like spectrograms; retrieval baselines produce sharp but duplicated copies of training clips. Our method generates novel spectrograms with clear harmonic structure that inherits recognizable spectro-temporal patterns from descendant species while interpolating between them. The reconstructed MRCA of *T. couchii* and *T. vociferans* produces a spectrogram whose frequency range and temporal patterning lie between the two descendants rather than copying either one. The reconstructed ancestor of the cryptic *E. traillii* and *E. alnorum* pair in Figure 9, species split taxonomically based on vocal differences, generates an intermediate vocalization blending elements of both descendant calls, suggesting the framework captures the acoustic divergence axis. Some ancestral spectrograms look similar because they inherit acoustic texture from the same anchor clip. The analysis in Appendix J.2 shows that traits recovered from decoded mel spectrograms retain information about the intended ancestral target even when the anchor source changes. Extended results are in Appendix K.

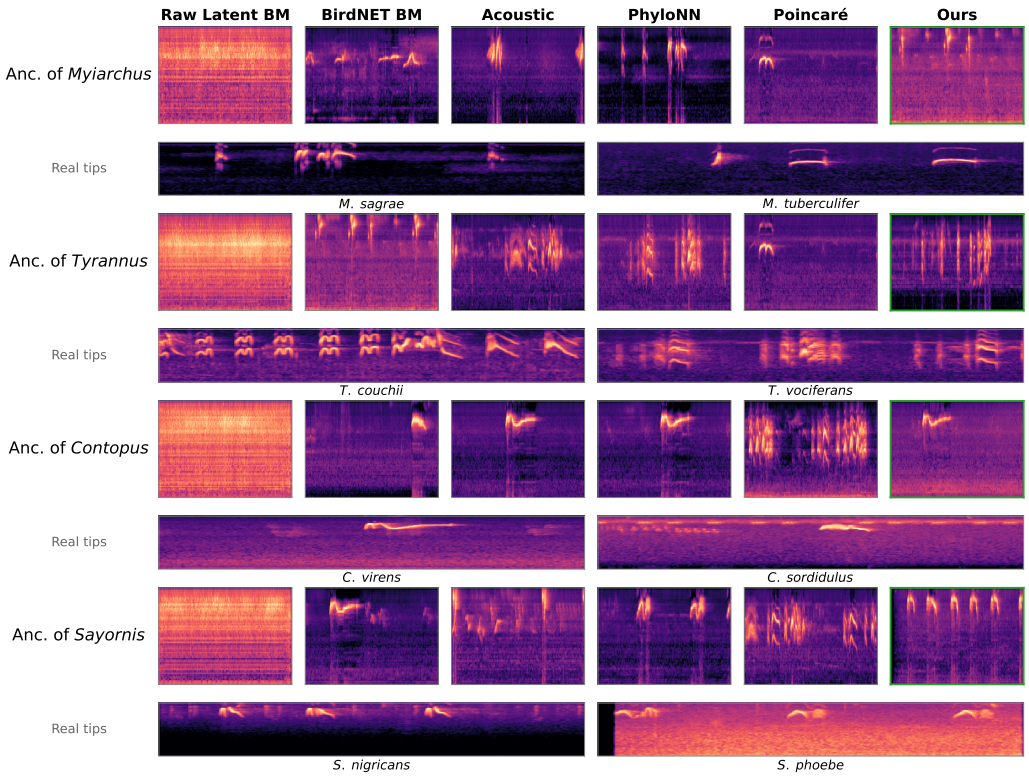


Figure 3: **Qualitative comparison of ancestral sound reconstructions.** Each block shows mel spectrograms generated by six methods for one ancestral node, labeled by genus, with real recordings of the two extant descendant species below for reference. Retrieval-based baselines such as Acoustic, PhyloNN, Poincaré produce copies of training clips and cannot generate novel ancestral sounds. Our method, highlighted in green, generates novel spectrograms with harmonic and temporal structure that interpolates between the descendant tips.

6 Conclusion

We presented the first framework that genuinely generates plausible ancestral vocalizations on a phylogeny, moving beyond retrieval. A tree-metric linear projection provides a compact subspace where Brownian motion matches evolutionary geometry and the anchored inverse lift fills the acoustic nullspace from nearby observed latents with a closed-form solution. Unlike retrieval-based approaches and naïve latent-space BM, our method generates unique, naturalistic audio for every ancestral node, with consistent results across two phylogenetically distant clades spanning the suboscine–oscine divide. Beyond methodology, the framework opens avenues for evolutionary biology: testing hypotheses about vocal divergence rates, examining whether reconstructed ancestors of cryptic species pairs exhibit intermediate acoustic properties, and reconstructing vocalizations for recently extinct species.

Limitations. Direct validation remains unavailable for deep ancestral nodes, so their reconstruction is evaluated through indirect plausibility metrics; perceptual listening studies could provide complementary evidence. The nullspace component of each ancestral latent is inherited from a single anchor clip (Corollary 3.5), leaving fine-grained acoustic texture dependent on an extant recording. Our full audio-generation experiments cover two clades of approximately 20 species each; evaluation on substantially larger phylogenies remains open. Although residual perturbation models pooled within-species acoustic variation, the current framework represents each species as one distribution and does not explicitly model differences among populations or the structure of cultural transmission.

Acknowledgments and Disclosure of Funding

This work was partially supported by the National Science Foundation (DEB-2144367 to CSL). S. Perea acknowledges postdoctoral fellowship from Conservation International (#AWD-100316). The authors declare no competing interests.

References

[1] Dolph Schluter, Trevor Price, Arne Ø Mooers, and Donald Ludwig. Likelihood of ancestor states in adaptive radiation. *Evolution*, 51(6):1699–1711, 1997.

[2] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999.

[3] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1): 1–15, 1985.

[4] Liam J Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, (2):217–223, 2012.

[5] Jozsef Arato and W. Tecumseh Fitch. Phylogenetic signal in the vocalizations of vocal learning and vocal non-learning birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1836):20200241, 09 2021. ISSN 0962-8436. doi: 10.1098/rstb.2020.0241. URL `https://doi.org/10.1098/rstb.2020.0241`.

[6] William D. Pearse, Ignacio Morales-Castilla, Logan S. James, Maxwell Farrell, Frédéric Boivin, and T. Jonathan Davies. Global macroevolution and macroecology of passerine song. *Evolution*, 72(4):944–960, 2018. doi: https://doi.org/10.1111/evo.13450. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.13450`.

[7] Elizabeth Perrault Derryberry, Nathalie Seddon, Graham Earnest Derryberry, Santiago Claramunt, Glenn Fairbanks Seeholzer, Robb Thomas Brumfield, and Joseph Andrew Tobias. Ecological drivers of song evolution in birds: Disentangling the effects of habitat and morphology. *Ecology and Evolution*, 8(3):1890–1905, 2018. doi: https://doi.org/10.1002/ece3.3760. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.3760`.

[8] Donald E Kroodsma. Songs of the alder flycatcher (empidonax alnorum) and willow flycatcher (empidonax traillii) are innate. *The Auk*, 101(1):13–24, 1984.

[9] Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003.

[10] Hans Slabbekoorn and Thomas B Smith. Bird song, ecology and speciation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1420):493–503, 2002.

[11] Robert Carrington Stein. Isolating mechanisms between populations of traill's flycatchers. *Proceedings of the American Philosophical Society*, 107(1):21–50, 1963.

[12] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2021.101236. URL `https://www.sciencedirect.com/science/article/pii/S1574954121000273`.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] John P Huelsenbeck, Rasmus Nielsen, and Jonathan P Bollback. Stochastic mapping of morphological characters. *Systematic biology*, 52(2):131–158, 2003.

[15] Thomas F Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997.

[16] Marguerite A Butler and Aaron A King. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The american naturalist*, 164(6):683–695, 2004.

[17] Luke J Harmon, Jonathan B Losos, T Jonathan Davies, Rosemary G Gillespie, John L Gittleman, W Bryan Jennings, Kenneth H Kozak, Mark A McPeek, Franck Moreno-Roark, Thomas J Near, Andy Purvis, Robert E Ricklefs, Dolph Schluter, James A Schulte II, Ole Seehausen, Brian L Sidlauskas, Omar Torres-Carvajal, Jason T Weir, and Arne Ø Mooers. Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, 64(8):2385–2396, 2010. doi: https://doi.org/10.1111/j.1558-5646.2010.01025.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2010.01025.x.

[18] Lys Sanz Moreta, Ola Rønning, Ahmad Salim Al-Sibahi, Jotun Hein, Douglas Theobald, and Thomas Hamelryck. Ancestral protein sequence reconstruction using a tree-structured ornstein-uhlenbeck variational autoencoder. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FZoZ7a31GCW.

[19] Mridul Khurana, Arka Daw, M Maruf, Josef C Uyeda, Wasila Dahdul, Caleb Charpentier, Yasin Bakış, Henry L Bart Jr, Paula M Mabee, Hilmar Lapp, et al. Hierarchical conditioning of diffusion models using tree-of-life for studying species evolution. *arXiv preprint arXiv:2408.00160*, 2024.

[20] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[21] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018.

[22] Anna Klimovskaia, David Lopez-Paz, Léon Bottou, and Maximilian Nickel. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):2966, 2020.

[23] Ciro Carvallo, Hernán Bocaccio, Gabriel B. Mindlin, and Pablo Groisman. Siamese networks for Poincaré embeddings and the reconstruction of evolutionary trees. *The ANZIAM Journal*, 67:e25, 2025. doi: 10.1017/S1446181125000148.

[24] Amy X Lu, Alex X Lu, and Alan Moses. Evolution is all you need: phylogenetic augmentation for contrastive learning. *arXiv preprint arXiv:2012.13475*, 2020.

[25] Liam J. Revell. Size-correction and principal components for interspecific comparative studies. *Evolution*, 63(12):3258–3268, 2009. doi: https://doi.org/10.1111/j.1558-5646.2009.00804.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2009.00804.x.

[26] Philipp Mitteroecker, Michael L Collyer, and Dean C Adams. Exploring phylogenetic signal in multivariate phenotypes by maximizing Blomberg's K. *Systematic Biology*, 74(2):215–229, 03 2025. ISSN 1063-5157. doi: 10.1093/sysbio/syae035. URL https://doi.org/10.1093/sysbio/syae035.

[27] Simon P. Blomberg, Theodore Garland JR., and Anthony R. Ives. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–745, 2003. doi: https://doi.org/10.1111/j.0014-3820.2003.tb00285.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0014-3820.2003.tb00285.x.

[28] Mohannad Elhamod, Mridul Khurana, Harish Babu Manogaran, Josef C Uyeda, Meghan A Balk, Wasila Dahdul, Yasin Bakis, Henry L Bart Jr, Paula M Mabee, Hilmar Lapp, et al. Discovering novel biological traits from images using phylogeny-guided neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3966–3978, 2023.

[29] Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, and Frank Guerin. Latent space factorisation and manipulation via matrix subspace projection. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5916–5926. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/li20i.html.

[30] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[31] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

[32] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

[33] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR, 23–29 Jul 2023.

[34] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation, 2023. URL https://arxiv.org/abs/2305.18474.

[35] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

[36] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

[37] Masato Hagiwara. Aves: Animal vocalization encoder based on self-supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095642.

[38] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13:22876, 2023.

[39] Tianyi Xu, Xuan Ouyang, Binwei Yao, Shoua Xiong, Sara Misurelli, Maichou Lor, and Junjie Hu. SITA: Learning speaker-invariant and tone-aware speech representations for low-resource tonal languages, 2026. URL https://arxiv.org/abs/2601.09050.

[40] Tim Sainburg, Marvin Thielk, and Timothy Q. Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology*, 16 (10):1–48, 10 2020. doi: 10.1371/journal.pcbi.1008228. URL https://doi.org/10.1371/journal.pcbi.1008228.

[41] Peter Mikula, Mihai Valcu, Henrik Brumm, Martin Bulla, Wolfgang Forstmeier, Tereza Petrusková, Bart Kempenaers, and Tomáš Albrecht. A global analysis of song frequency in passerines provides no support for the acoustic adaptation hypothesis but suggests a role for sexual selection. *Ecology Letters*, 24(3):477–486, 2021. doi: https://doi.org/10.1111/ele.13662. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13662.

[42] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1):397–434, 2013. doi: 10.1007/s10107-012-0584-1.

[43] Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3794–3803. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lezcano-casado19a.html.

[44] Emilia P Martins and Thomas F Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

[45] W. Jetz, G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. The global diversity of birds in space and time. *Nature*, 491(7424):444–448, 2012. doi: 10.1038/nature11631.

[46] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

[47] Per GP Ericson, Cajsa L Anderson, Tom Britton, Andrzej Elzanowski, Ulf S Johansson, Mari Källersjö, Jan I Ohlson, Thomas J Parsons, Dario Zuccon, and Gerald Mayr. Diversification of neoaves: integration of molecular sequence data and fossils. *Biology letters*, 2(4):543–547, 2006.

[48] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

[49] Klaus Peter Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[50] Shannon J. Hackett, Rebecca T. Kimball, Sushma Reddy, Rauri C. K. Bowie, Edward L. Braun, Michael J. Braun, Jena L. Chojnowski, W. Andrew Cox, Kin-Lan Han, John Harshman, Christopher J. Huddleston, Ben D. Marks, Kathleen J. Miglia, William S. Moore, Frederick H. Sheldon, David W. Steadman, Christopher C. Witt, and Tamaki Yuri. A phylogenomic study of birds reveals their evolutionary history. *Science*, 320(5884):1763–1768, 2008. doi: 10.1126/science.1157704. URL `https://www.science.org/doi/abs/10.1126/science.1157704`.

[51] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

[52] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, et al. librosa: Audio and music signal analysis in python. *SciPy*, 2015(18-24):7, 2015.

[53] Dean C Adams. A generalized k statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. *Systematic biology*, 63(5):685–697, 2014.

[54] Heather Williams, Andrew Scharf, Anna R Ryba, D Ryan Norris, Daniel J Mennill, Amy EM Newman, Stéphanie M Doucet, and Julie C Blackwood. Cumulative cultural evolution and mechanisms for cultural selection in wild bird songs. *Nature Communications*, 13(1):4001, 2022.

A Trait Dimension Sensitivity

Table 5 reports phylogenetic plausibility metrics as a function of the trait dimension K on Dataset 1. Descendant Affinity Margin increases with K, peaking at $K{=}16$, indicating that higher-dimensional projections better separate descendant from non-descendant affinities. Tree–Embedding Correlation is highest at $K{=}4$–8 and decreases for larger K, reflecting a bias–variance trade-off: more trait dimensions allow finer-grained distance matching but also admit more noise from the optimization landscape. At $K{=}21 = n$, the maximum rank, Tree–Embedding Correlation drops sharply to 0.392, consistent with overfitting when no degrees of freedom remain for regularization. We select $K{=}20$ as it provides near-maximal expressiveness while retaining regularization capacity, achieving strong Edge Monotonicity of 0.867 and balanced performance across all metrics. While $K{=}8$ yields the highest TreeCorr of 0.563, its Edge Monotonicity drops to 0.533, the lowest among all settings, indicating that local parent–child acoustic ordering breaks down. Conversely, $K{=}4$ achieves high Edge Monotonicity of 0.933 but retains too few dimensions to capture fine-grained phylogenetic variation, as reflected in its lower DescMarg. $K{=}20$ balances these trade-offs: it maintains strong Edge Monotonicity of 0.867 while encoding enough phylogenetic information for robust ancestral inference across all metrics.

Table 5: **Trait dimension K sensitivity on Dataset 1.** All metrics use the full pipeline with learned $W \in \mathbb{R}^{K\times 3120}$. ↑: higher is better; ↓: lower is better. Best per column in **bold**.

K	TreeCorr↑	DescMarg↑	EdgeMono↑	NNDist↑	Unique↑
2	0.518	0.034	0.800	0.222	1.000
4	0.556	0.040	**0.933**	0.226	1.000
8	**0.563**	0.046	0.533	0.227	1.000
16	0.498	**0.053**	0.800	0.224	1.000
20	0.486	0.042	0.867	0.235	1.000
21	0.392	0.046	**0.933**	**0.245**	1.000

B Trait Space Interpretability

Figure 4 analyzes what the $K{=}20$ learned trait dimensions capture on Dataset 1.

Acoustic feature correspondence (Figure 4a). We compute species-level Pearson correlations between each trait dimension and 33 classical acoustic descriptors: 20 MFCCs, spectral centroid, bandwidth, rolloff, flatness, zero-crossing rate, RMS energy, and 7-band spectral contrast. Different trait dimensions systematically correspond to distinct acoustic properties. Dim 11 correlates most strongly with spectral bandwidth at $r{=}-0.76$, capturing vocal frequency spread. Dim 1 aligns with high-frequency spectral contrast at $r{=}+0.71$, reflecting harmonic structure. Dim 14 captures mid-frequency spectral contrast at $r{=}-0.64$, and Dim 8 tracks spectral flatness at $r{=}+0.60$, distinguishing tonal from noisy vocalizations. The learned trait space thus recovers known acoustic dimensions relevant to comparative bioacoustics [7] without being explicitly trained on them.

Per-dimension phylogenetic signal (Figure 4b). We measure each trait dimension's individual phylogenetic signal as the Pearson correlation between its pairwise squared distances and patristic distances. The dimensions exhibit a clear hierarchy: Dim 1 alone achieves $r{=}0.75$, capturing the majority of the full 20-dimensional signal at $r{=}0.905$. The top four dimensions—1, 4, 7, and 3—account for most of the phylogenetic structure, while 15 of 20 dimensions maintain positive signal above $r{=}0.1$, confirming that the dimensions are complementary rather than redundant. Notably, the dimensions with the strongest phylogenetic signal are those most correlated with spectral contrast features, consistent with findings that harmonic structure carries phylogenetic information in bird vocalizations [5].

Subspace decomposition validation (Figure 5). Section 3.4 assumes that $\mathrm{row}(W)$ concentrates phylogenetically informative variation while $\ker(W)$ encodes non-phylogenetic acoustic texture. We verify this empirically by projecting species-mean latents onto each subspace and correlating pairwise squared distances with patristic distances. The $K{=}20$-dimensional $\mathrm{row}(W)$ achieves $r{=}0.905$, while the remaining $D{-}K{=}3{,}100$-dimensional $\ker(W)$ achieves only $r{=}0.208$. Normalized by dimensionality, the per-dimension phylogenetic signal density in $\mathrm{row}(W)$ is approximately 600 times

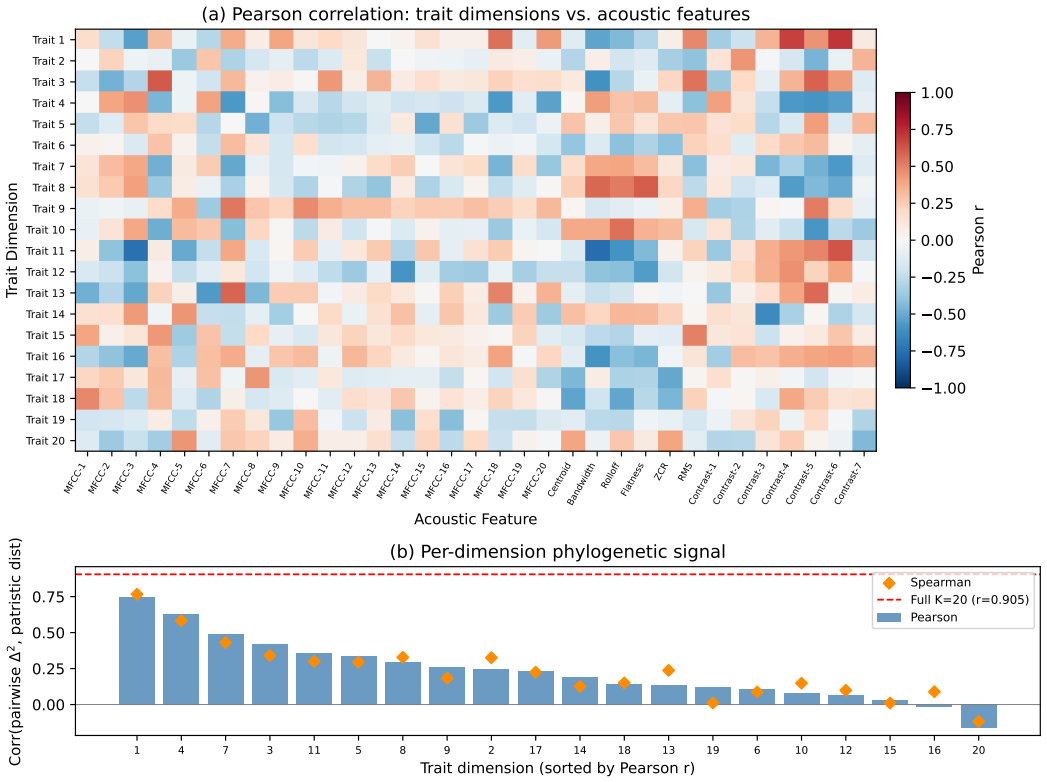


Figure 4: **Trait space interpretability on Dataset 1. (a)** Pearson correlation between each of the K=20 trait dimensions and 33 acoustic features, computed at the species level. Different dimensions capture distinct acoustic properties: spectral contrast, bandwidth, flatness, and mel-frequency cepstral coefficients (MFCCs), which summarize the spectral envelope of a sound. **(b)** Per-dimension phylogenetic signal, measured as the correlation between single-dimension pairwise squared distances and patristic distances. Dimensions are sorted by decreasing signal strength; the dashed line shows the full 20-dimensional correlation (r=0.905).

that of $\ker(W)$, confirming that the learned projection successfully separates phylogenetic structure from acoustic texture in the VAE latent space.

C Residual Perturbation Properties

The pooled residual standard deviation is computed as $\hat{\sigma}_{\text{res},k} = \left(\frac{1}{N}\sum_{s\in\mathcal{L}}\sum_{i=1}^{n_s}(z_{s,i,k} - y_{s,k})^2\right)^{1/2}$ with $N = \sum_s n_s$.

Proposition C.1 (Mean preservation and variance characterization)**.** *The perturbation* (9) *satisfies:*

1. $\mathbb{E}[\tilde{y}_u^{(b)} \mid y_u^{(b)}] = y_u^{(b)}$ *(the BM posterior mean is preserved);*
2. $\mathrm{Cov}(\tilde{y}_u^{(b)} \mid y_u^{(b)}) = \gamma^2\,\mathrm{diag}(\hat{\sigma}^2_{\text{res}})$ *(independent, axis-aligned noise);*
3. *The marginal variance of* $\tilde{y}_u^{(b)}$ *integrating over both BM sampling and perturbation is* $\tau^2\sigma_u^2\hat{\sigma}^2_{y,k} + \gamma^2\hat{\sigma}^2_{\text{res},k}$ *along coordinate* k*.*

Proof. Parts 1 and 2 follow from $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Cov}(\epsilon) = I_K$. For part (3), by the law of total variance,

$$\mathrm{Var}(\tilde{y}_{u,k}^{(b)}) = \mathbb{E}[\mathrm{Var}(\tilde{y}_{u,k}^{(b)} \mid y_u^{(b)})] + \mathrm{Var}(\mathbb{E}[\tilde{y}_{u,k}^{(b)} \mid y_u^{(b)}]) = \gamma^2\hat{\sigma}^2_{\text{res},k} + \tau^2\sigma_u^2\hat{\sigma}^2_{y,k}. \qquad \square$$

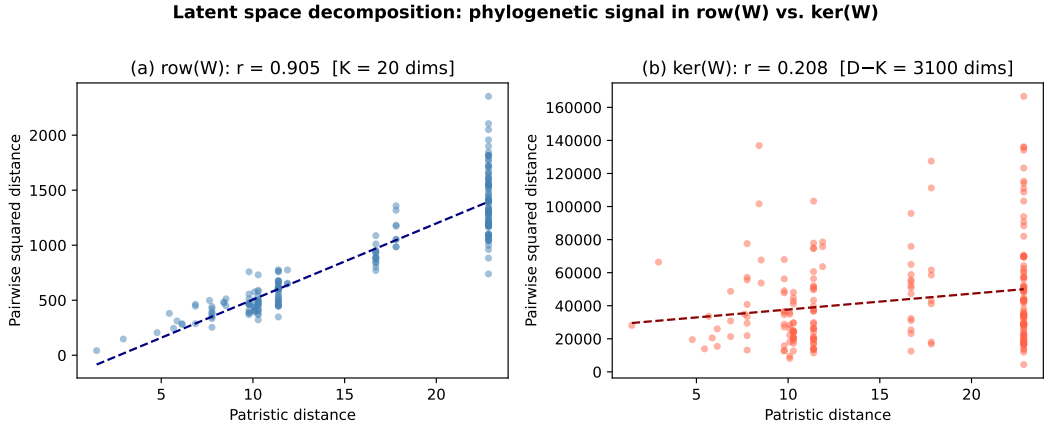


Figure 5: **Subspace decomposition validation on Dataset 1. (a)** Pairwise squared distances in $\mathrm{row}(W)$ vs. patristic distances show a strong linear relationship at r=0.905, confirming that the learned trait subspace captures phylogenetic structure. **(b)** The same analysis in $\ker(W)$ yields only r=0.208 despite spanning 3,100 dimensions, indicating that the nullspace primarily encodes non-phylogenetic acoustic variation.

D Proofs and Theoretical Details

D.1 Proof of Proposition 3.1 (Exact Row-Orthonormality)

By construction $LL^\top = AA^\top + \varepsilon I_K$, so $AA^\top = LL^\top - \varepsilon I_K$. Substituting into $WW^\top$:

$$\begin{aligned} WW^\top &= L^{-1}AA^\top L^{-\top} = L^{-1}(LL^\top - \varepsilon I_K)L^{-\top} \\ &= L^{-1}LL^\top L^{-\top} - \varepsilon\, L^{-1}L^{-\top} = I_K - \varepsilon\, L^{-1}L^{-\top}, \end{aligned}$$

where $L^{-1}L = I_K$ and $L^\top L^{-\top} = I_K$ give the first term. As $\varepsilon \to 0$, $WW^\top \to I_K$.

Operator norm bound. For the quantitative bound, we relate $L^{-1}L^{-\top}$ to the spectrum of $AA^\top + \varepsilon I_K$. Since L is square and invertible, $L^{-1}L^{-\top} = (L^\top L)^{-1}$. For any invertible square matrix M, the matrices $M^\top M$ and $MM^\top$ share the same eigenvalues (both equal the squared singular values of M), so

$$\|L^{-1}L^{-\top}\|_{\mathrm{op}} = \|(L^\top L)^{-1}\|_{\mathrm{op}} = \|(LL^\top)^{-1}\|_{\mathrm{op}} = \frac{1}{\lambda_{\min}(LL^\top)} = \frac{1}{\lambda_{\min}(AA^\top + \varepsilon I_K)}.$$

Since the eigenvalues of $AA^\top$ are the squared singular values of A, we have $\lambda_{\min}(AA^\top + \varepsilon I_K) = \sigma^2_{\min}(A) + \varepsilon$, yielding

$$\|WW^\top - I_K\|_{\mathrm{op}} = \varepsilon\, \|L^{-1}L^{-\top}\|_{\mathrm{op}} = \frac{\varepsilon}{\sigma^2_{\min}(A) + \varepsilon} \le \frac{\varepsilon}{\sigma^2_{\min}(A)},$$

where the last step uses $\sigma^2_{\min}(A) + \varepsilon > \sigma^2_{\min}(A)$. This yields the bound stated in Proposition 3.1. In practice we use $\varepsilon = 10^{-6}$ and observe $\sigma_{\min}(A) \approx 0.1$, giving a deviation of approximately 10^{-4}, which is negligible compared to the scale of BM posterior uncertainty. This also bounds the trait-preservation error of the anchored lift (Lemma 3.3): $\|Wx^\star - y\|_2 \le \|WW^\top - I_K\|_{\mathrm{op}} \cdot \|y - Wx_0\|_2 \le 10^{-4}\|y - Wx_0\|_2$.

The Cholesky reparametrization has three advantages over direct Stiefel projection: (i) it is differentiable everywhere (assuming full rank), enabling standard gradient-based optimization; (ii) the Cholesky decomposition is numerically stable for $K \ll D$ (our setting: K=20, D=3,120); (iii) it avoids the retraction step required by Riemannian optimizers on the Stiefel manifold, which can be expensive for tall-thin matrices.

D.2 Proof of Proposition 3.2 (BM Conditional Posterior)

The joint distribution of $[\bar{Y}_u; \bar{Y}_{\mathcal{L}}]$ under BM is Gaussian with block covariance $\left[\begin{smallmatrix} C_{uu} & c_u^\top \\ c_u & \Sigma_{\mathcal{L}\mathcal{L}} \end{smallmatrix}\right]$. By the standard Gaussian conditioning formula, the conditional mean is $c_u^\top \Sigma_{\mathcal{L}\mathcal{L}}^{-1} \bar{Y}_{\mathcal{L},j}$ and the conditional variance is $C_{uu} - c_u^\top \Sigma_{\mathcal{L}\mathcal{L}}^{-1} c_u$. The nugget enters only through $\Sigma_{\mathcal{L}\mathcal{L}}$ (it models observation noise at the tips), not through c_u or C_{uu}. Coordinate-wise factorization follows from the i.i.d.-across-dimensions BM assumption.

In implementation, we solve the linear system $\Sigma_{\mathcal{L}\mathcal{L}}\alpha = \bar{Y}_{\mathcal{L},j}$ via LU decomposition rather than explicitly computing $\Sigma_{\mathcal{L}\mathcal{L}}^{-1}$, and clamp the posterior variance to $\max(\sigma_u^2, 0)$ to handle numerical errors near the root.

D.3 Proof of Lemma 3.3 (Optimal Anchored Lift)

Introduce Lagrange multipliers $\lambda \in \mathbb{R}^K$ and form the Lagrangian

$$\mathcal{L}(x, \lambda) = \tfrac{1}{2}\|x - x_0\|_2^2 + \lambda^\top (y - Wx).$$

The first-order stationarity condition $\nabla_x \mathcal{L} = 0$ gives $x - x_0 = W^\top \lambda$, i.e. $x = x_0 + W^\top \lambda$. Substituting into the constraint $Wx = y$:

$$Wx_0 + WW^\top \lambda = y \implies \lambda = (WW^\top)^{-1}(y - Wx_0) = y - Wx_0,$$

where the last equality uses $WW^\top = I_K$. Hence $x^\star = x_0 + W^\top(y - Wx_0)$. Uniqueness follows from strict convexity of $\|\cdot\|_2^2$ over the affine constraint set $\mathcal{A}(y)$.

D.4 Proofs of Corollaries 3.4 and 3.5

Corollary 3.4 (Exact trait preservation). Direct computation: $Wx^\star = Wx_0 + WW^\top(y - Wx_0) = Wx_0 + I_K(y - Wx_0) = y$.

Corollary 3.5 (Nullspace preservation). From Lemma 3.3, $x^\star - x_0 = W^\top v$ for $v = y - Wx_0 \in \mathbb{R}^K$. Since $W^\top v \in \mathrm{col}(W^\top) = \mathrm{row}(W) = \ker(W)^\perp$, we have $P_{\ker(W)}(x^\star - x_0) = 0$.

D.5 Orthogonal Decomposition of the Latent Space

The row-orthonormality $WW^\top = I_K$ induces the orthogonal decomposition $\mathbb{R}^D = \mathrm{row}(W) \oplus \ker(W)$, where $\dim(\mathrm{row}(W)) = K$ and $\dim(\ker(W)) = D - K$. The orthogonal projectors onto these subspaces are $P_{\mathrm{row}} = W^\top W$ and $P_{\ker} = I_D - W^\top W$, respectively. By Corollaries 3.4 and 3.5, the anchored lift operator acts as follows: (i) it replaces the row-space component of x_0 with the target trait, the *phylogenetic subspace* controlled by the BM posterior; (ii) it preserves the nullspace component of x_0 exactly, the *acoustic nullspace*, which encodes fine-grained spectro-temporal texture inherited from a nearby real clip.

This decomposition has a useful consequence for generation diversity. Let x_0, x_0' be two different anchor clips with the same trait, i.e. $\phi_W(x_0) = \phi_W(x_0')$. Then the lifted latents $x^\star, x^{\star\prime}$ have the same row-space component but differ in their nullspace components: $P_{\ker} x^\star = P_{\ker} x_0 \neq P_{\ker} x_0' = P_{\ker} x^{\star\prime}$. This means different anchor selections produce distinct ancestral audio with the same phylogenetic content but different acoustic textures.

D.6 Invariance of the Anchored Lift under Orthogonal Rotation

As noted in Section 3.2, W is unique only up to left-multiplication by an orthogonal $Q \in \mathbb{R}^{K \times K}$. We verify that the full anchored lift, including anchor selection, is invariant to this ambiguity. Replacing W with $\tilde{W} = QW$ gives traits $\tilde{y} = QWx = Qy$. Anchor selection (11) is invariant because $\|\tilde{W}x - \tilde{y}\|_2 = \|Q(Wx - y)\|_2 = \|Wx - y\|_2$ for any orthogonal Q, so the same anchor x_0 is selected under both W and $\tilde{W}$. The lift then becomes

$$\tilde{x}^\star = x_0 + \tilde{W}^\top(\tilde{y} - \tilde{W}x_0) = x_0 + W^\top Q^\top Q(y - Wx_0) = x_0 + W^\top(y - Wx_0) = x^\star,$$

where we used $Q^\top Q = I_K$. Hence the deterministic pipeline, comprising the posterior mean, anchored lift, and decoding, produces identical audio regardless of which orthogonal representative of

W is found by optimization. The stochastic sampling steps (8)–(9) are not strictly invariant because the axis-aligned noise scaling by $\hat{\sigma}_y$ and $\hat{\sigma}_{\mathrm{res}}$ does not commute with rotation, as $\mathrm{diag}(Q\Sigma Q^\top) \neq Q\,\mathrm{diag}(\Sigma)\,Q^\top$ in general. In practice this effect is negligible because the posterior mean dominates the sample, $\tau < 1$ attenuates phylogenetic noise, and γ is small.

E Full Algorithm

Algorithm 1 Ancestral Sound Reconstruction via Tree-Metric Latent Traits

Input: Tree $T = (\mathcal{V}, \mathcal{E})$; clip sets $\{\mathcal{A}_s\}_{s\in\mathcal{L}}$; pretrained $(E_{\mathrm{vae}}, D_{\mathrm{vae}}, V)$; hyperparameters $(K, \alpha, \lambda, B, \tau, \gamma)$.
Output: Ancestral waveforms $\{\hat{a}_u^{(b)}\}_{u\in\mathcal{U},\, b=1}^{B}$.
Encode clips: $x_{s,i} \leftarrow \mathrm{vec}(E_{\mathrm{vae}}(a_{s,i})) \in \mathbb{R}^D$ for all $s \in \mathcal{L}$, $i = 1, \ldots, n_s$.
Learn $W \in \mathbb{R}^{K\times D}$ with $WW^\top = I_K$ via (2)–(3).
Compute trimmed tip traits $\{y_s\}_{s\in\mathcal{L}}$ via (4); save empirical mean $\hat{\mu}_y$ and std. dev. $\hat{\sigma}_y$; standardize to zero mean, unit variance.
Build $\Sigma_{\mathcal{L}\mathcal{L}} = C_{\mathcal{L}\mathcal{L}} + \lambda I$; compute pooled within-species residual scale $\hat{\sigma}_{\mathrm{res}}$.
for each internal node $u \in \mathcal{U}$ **do**
  Compute posterior mean m_u and variance σ_u^2 via Proposition 3.2.
  for $b = 1, \ldots, B$ **do**
    Compute unstandardized mean $\hat{y}_u \leftarrow \hat{\mu}_y + \hat{\sigma}_y \odot m_u$; sample $y_u^{(b)} \leftarrow \hat{y}_u + \tau\,\sigma_u\,(\hat{\sigma}_y \odot \epsilon_1)$, $\epsilon_1 \sim \mathcal{N}(0, I_K)$.
    Perturb: $\tilde{y}_u^{(b)} \leftarrow y_u^{(b)} + \gamma\,\hat{\sigma}_{\mathrm{res}} \odot \epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0, I_K)$.
    Select anchor: $x_0 \leftarrow \arg\min_{x\in\mathcal{X}_{\mathrm{train}}} \|\phi_W(x) - \tilde{y}_u^{(b)}\|_2$.
    Lift: $x^\star \leftarrow x_0 + W^\top(\tilde{y}_u^{(b)} - Wx_0)$.
    Emit: $\hat{a}_u^{(b)} \leftarrow V\big(D_{\mathrm{vae}}(x^\star)\big)$.
  end for
end for

Algorithm 1 summarizes the complete ancestral sound reconstruction pipeline described in Sections 3.2–3.4.

F Alternative Evolutionary Models

In the main text we adopt standard Brownian motion (BM) as the evolutionary model for ancestral trait inference. Here we describe two families of generalizations, the Ornstein–Uhlenbeck (OU) process and Pagel's branch-length transformations, and compare them empirically against BM. All models share the same Gaussian conditioning machinery (Proposition 3.2); they differ only in the tip-tip covariance $C_{\mathcal{L}\mathcal{L}}$, the cross-covariance c_u, and the self-variance C_{uu} used for each internal node u.

F.1 Ornstein–Uhlenbeck Process

The OU process [15] adds a mean-reverting drift to Brownian motion. Each trait coordinate evolves as

$$dY_t = -\alpha\,(Y_t - \theta)\,dt + dW_t, \qquad \alpha > 0, \tag{15}$$

where α controls the strength of stabilizing selection and θ is the long-run optimum, set to zero after standardization. The resulting covariance between tips i, j separated by path lengths s_i, s_j from their MRCA at depth t_a is

$$C_{ij}^{\mathrm{OU}} = \frac{1}{2\alpha}\big(1 - e^{-2\alpha\,t_a}\big)\,e^{-\alpha(s_i+s_j)}, \tag{16}$$

where $t_a = \mathrm{depth}(\mathrm{MRCA}(i,j))$ and s_i, s_j are the branch lengths from the MRCA to tips i, j respectively. As $\alpha \to 0$ the OU covariance recovers BM; large α shrinks off-diagonal entries toward zero, effectively decorrelating distant relatives.

For the cross-covariance between internal node u, at depth t_u, and tip i, and the self-variance of u:

$$(c_u^{\mathrm{OU}})_i = \frac{1}{2\alpha}\big(1 - e^{-2\alpha\, t_a}\big)\, e^{-\alpha\,(s_u + s_i)}, \tag{17}$$

$$C_{uu}^{\mathrm{OU}} = \frac{1}{2\alpha}\big(1 - e^{-2\alpha\, t_u}\big), \tag{18}$$

where $t_a = \mathrm{depth}(\mathrm{MRCA}(u, i))$, $s_u = t_u - t_a$ is the branch length from the MRCA to node u, and s_i is the branch length from the MRCA to tip i. When u is an ancestor of i, the MRCA is u itself, so $s_u = 0$ and the expression reduces to $\frac{1}{2\alpha}(1 - e^{-2\alpha\, t_u})\, e^{-\alpha\, s_i}$.

F.2 Pagel's λ Transformation

Pagel's λ [2] rescales the off-diagonal elements of the BM covariance matrix to modulate phylogenetic signal strength:

$$C_{ij}^{\lambda} = \begin{cases} \lambda \cdot \mathrm{depth}\big(\mathrm{MRCA}(i, j)\big), & i \neq j, \\ \mathrm{depth}(i), & i = j. \end{cases} \tag{19}$$

When $\lambda = 1$ this recovers BM; $\lambda = 0$ produces a star phylogeny in which tips are independent given the root, discarding all shared evolutionary history. Intermediate values interpolate between these extremes, allowing the model to down-weight phylogenetic covariance when the observed trait variation contains a substantial non-phylogenetic component.

The cross-covariance and self-variance for internal node u are:

$$(c_u^{\lambda})_i = \lambda \cdot \mathrm{depth}\big(\mathrm{MRCA}(u, i)\big), \tag{20}$$

$$C_{uu}^{\lambda} = \lambda \cdot \mathrm{depth}(u). \tag{21}$$

F.3 Pagel's δ Transformation

Pagel's δ [2] applies a power transformation to node depths, re-weighting early versus late evolution:

$$C_{ij}^{\delta} = \big[\mathrm{depth}\big(\mathrm{MRCA}(i, j)\big)\big]^{\delta}. \tag{22}$$

When $\delta = 1$ this recovers BM. Values $\delta < 1$ stretch deep branches relative to shallow ones, allocating more variance to early divergence, consistent with early-burst models, while $\delta > 1$ compresses deep branches and stretches shallow ones, consistent with accelerating late evolution.

The cross-covariance and self-variance follow the same power transformation:

$$(c_u^{\delta})_i = \big[\mathrm{depth}\big(\mathrm{MRCA}(u, i)\big)\big]^{\delta}, \tag{23}$$

$$C_{uu}^{\delta} = \big[\mathrm{depth}(u)\big]^{\delta}. \tag{24}$$

F.4 Empirical Comparison

We evaluate 12 configurations on Dataset 1, Tyrannidae, 21 species, 20 internal nodes, spanning BM, OU with $\alpha \in \{0.1, 0.5, 1.0, 2.0\}$, Pagel's λ with $\lambda \in \{0, 0.25, 0.5, 0.75\}$, and Pagel's δ with $\delta \in \{0.5, 1.5, 2.0\}$. All configurations use the same learned tree-metric trait space, nearest-trait anchored lifting, and BigVGAN decoding pipeline as the full method in the main text; only the evolutionary model for ancestral inference differs. Results are shown in Table 6.

Discussion. Pagel's λ=0.25 achieves the highest Tree–Embedding Correlation, 0.5730 vs. 0.4859 for BM, an 18% improvement, and competitive audio quality with FAD-BN of 0.5729, indicating better global phylogenetic consistency. However, BM retains stronger Edge Monotonicity, 0.8667 vs. 0.7333, suggesting that the BM posterior better preserves local parent–child ordering. OU with mild selection at α=0.1 performs comparably to BM with a slight edge in Tree Correlation and Edge Monotonicity, but stronger OU selection at $\alpha \geq 0.5$ degrades all metrics substantially as the mean-reverting drift overwhelms phylogenetic signal. Pagel's δ=0.5 achieves the best Descendant Affinity Margin of 0.0812, the lowest FAD-PANN of 0.1433, and the lowest FAD-BN of 0.5687, but at the cost of higher Silence fraction of 0.1133 and weaker Edge Monotonicity.

Overall, BM provides the best balance across all metrics and remains our default choice in the main text. Pagel's λ with mild phylogenetic signal reduction at $\lambda \approx 0.25$ is a promising alternative when global tree correlation is prioritized over local edge ordering.

Table 6: **Evolutionary model comparison.** All configurations use our learned tree-metric trait space with nearest-trait anchored lifting and mild residual perturbation ($\gamma = 0.25$). Only the evolutionary model for ancestral inference varies. Best value per column in **bold**.

		Phylogenetic Plausibility					*Audio Quality*				
Model	Param	TreeCorr↑	DescMarg↑	EdgeMono↑	NNDist↑	Unique↑	CycleSim↑	FAD-BN↓	FAD-PANN↓	FAD-CLAP↓	Silence↓
BM	—	0.4859	0.0417	0.8667	0.2350	**1.0000**	0.9985	0.6326	0.1744	0.6710	0.0079
OU	α=0.1	0.4983	0.0463	**0.9333**	0.2157	**1.0000**	0.9986	0.6097	0.1703	0.6627	0.0146
OU	α=0.5	0.1643	0.0044	0.8667	0.3916	**1.0000**	0.9981	0.9026	0.4534	1.3264	0.0261
OU	α=1.0	0.2443	0.0143	0.6667	**0.4155**	**1.0000**	0.9955	0.9332	0.6624	1.7026	0.0023
OU	α=2.0	0.1655	0.0088	0.8667	0.4102	**1.0000**	0.9911	0.9466	0.8766	1.8562	**0.0002**
Pagel λ	λ=0	0.5226	0.0432	0.7333	0.2310	**1.0000**	**0.9987**	0.6454	0.1499	**0.6332**	0.0249
Pagel λ	λ=0.25	**0.5730**	0.0439	0.7333	0.2116	**1.0000**	0.9986	0.5729	0.1576	0.6339	0.0003
Pagel λ	λ=0.5	0.5150	0.0445	0.8000	0.2231	**1.0000**	0.9986	0.5990	0.1954	0.6872	0.0510
Pagel λ	λ=0.75	0.4705	0.0447	0.8000	0.2322	**1.0000**	0.9986	0.6294	0.1895	0.6903	0.0040
Pagel δ	δ=0.5	0.3877	**0.0812**	0.6667	0.2088	**1.0000**	0.9986	**0.5687**	**0.1433**	0.6593	0.1133
Pagel δ	δ=1.5	0.4674	0.0160	0.6000	0.2014	**1.0000**	0.9985	0.6455	0.2156	0.6978	0.0159
Pagel δ	δ=2.0	0.4374	0.0158	0.6000	0.2044	**1.0000**	0.9985	0.6521	0.2154	0.6992	0.0154

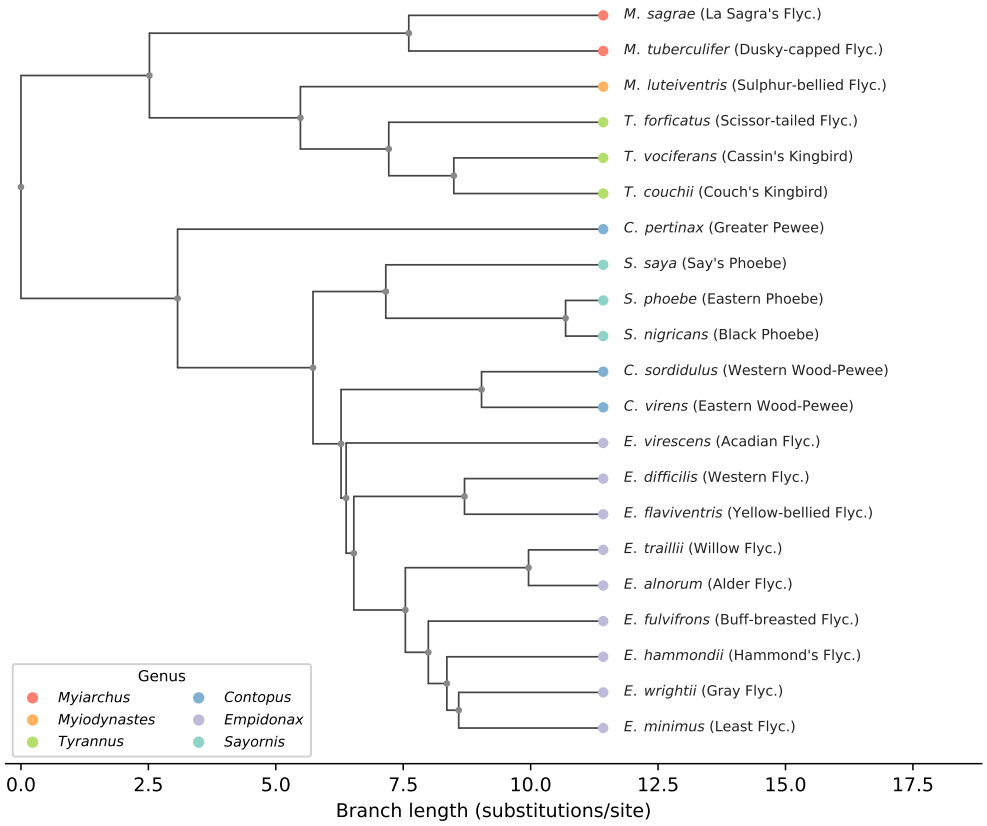


Figure 6: **Phylogenetic tree of the tyrant flycatcher clade for Dataset 1.** 21 extant species from 6 genera, with 20 internal ancestral nodes, derived from BirdTree.org [45]. Branch lengths reflect evolutionary distance in substitutions per site. Species are colored by genus.

G Experimental Details

G.1 Dataset 1: Tyrannidae

Our first dataset consists of vocalization recordings from a clade of New World tyrant flycatchers, family Tyrannidae, one of the most species-rich passerine radiations. The 21 focal species were selected to maximize phylogenetic coverage across 6 genera, *Empidonax*, *Contopus*, *Myiarchus*, *Myiodynastes*, *Sayornis*, and *Tyrannus*, and to include only species with sufficient high-quality

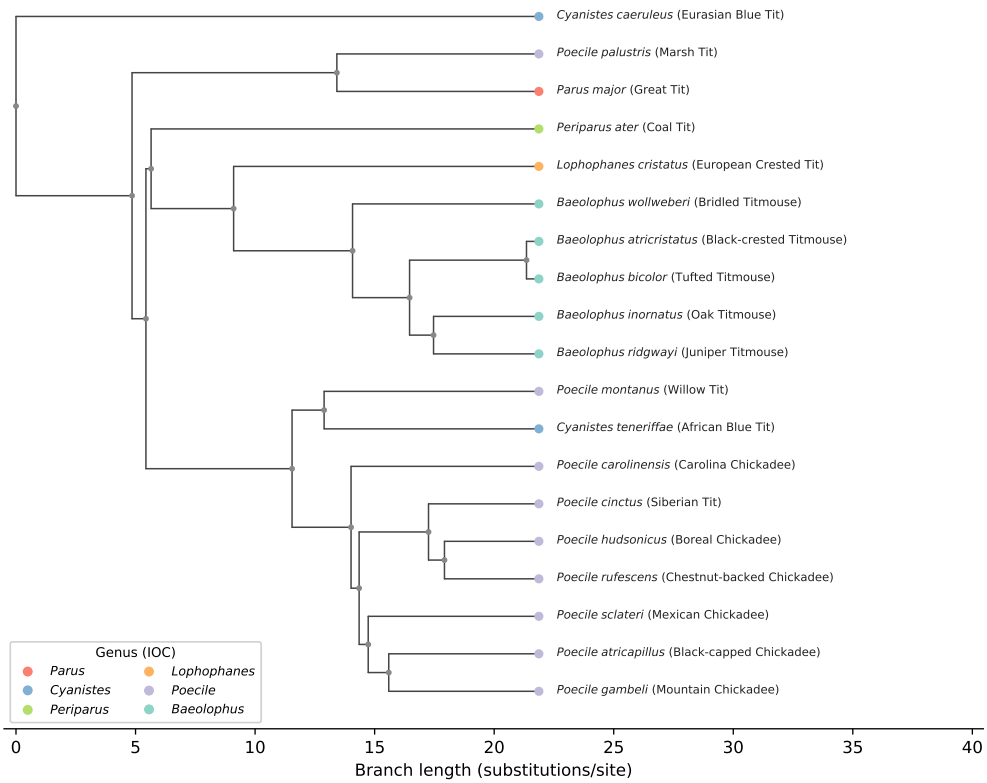


Figure 7: **Phylogenetic tree of the Paridae clade for Dataset 2.** 19 extant species from 6 genera under IOC taxonomy, with 18 internal ancestral nodes, derived from BirdTree.org [45]. Branch lengths reflect evolutionary distance in substitutions per site. Species are colored by IOC genus; the BirdTree taxonomy lumps *Cyanistes*, *Periparus*, *Lophophanes*, and *Poecile* under *Parus*.

recordings available on Xeno-Canto with quality ratings A–B and a minimum duration of 5 s. The clade comprises 9 *Empidonax* spp., 3 *Contopus* spp., 3 *Sayornis* spp., 3 *Tyrannus* spp., 2 *Myiarchus* spp., and 1 *Myiodynastes* sp. All audio files were manually reviewed and trimmed to standardized 5-second clips using Raven Pro v1.6, then converted to 16-bit signed 16 kHz mono WAV format. Each species is represented by 47–68 audio clips with a mean of 54.7 clips per species, for a total of approximately 1,100 clips. For the phylogenetic framework, 10,000 posterior trees were downloaded from BirdTree.org [45] using the Ericson Stage 2 backbone [47]. Species name concordance between the focal species list and the BirdTree taxonomy was verified before pruning. All 10,000 trees were pruned to the focal species set using the R packages `ape` [48] and `phangorn` [49], and a maximum clade credibility consensus tree was computed by selecting the tree with the highest product of posterior clade support values while retaining the original branch lengths. The resulting tree is fully resolved with no polytomies, with 21 tip nodes and 20 internal ancestral nodes and is ultrametric after midpoint rooting, with a uniform root-to-tip depth of 11.42 substitutions/site. Figure 6 shows the tree topology with species colored by genus. The tree structure captures several well-resolved sister-species pairs, such as *E. traillii–E. alnorum* and *S. phoebe–S. saya*, that serve as natural test cases for ancestral reconstruction: the most recent common ancestor of a closely related pair should produce a vocalization intermediate between its two extant descendants.

G.2 Dataset 2: Paridae

Our second dataset tests cross-clade generalization using an independently curated set of tits, chickadees, and titmice, family Paridae. The 19 focal species span 6 genera: 9 *Poecile* spp., 5 *Baeolophus* spp., 2 *Cyanistes* spp., 1 *Parus* sp., 1 *Periparus* sp., and 1 *Lophophanes* sp. Recordings were downloaded programmatically from the Xeno-Canto API with quality ratings A–B, up to 80 per

species. Unlike Dataset 1, which was manually curated with Raven Pro, recordings shorter than 3 s were excluded automatically and the loudest 5-second segment was extracted from each remaining file; recordings between 3–5 s were zero-padded to the target length. After filtering, 50–80 clips per species were retained, totaling approximately 1,150 clips. Audio is standardized identically to Dataset 1: 16-bit signed 16 kHz mono WAV, 5 seconds per clip. For the phylogenetic framework, 1,000 posterior trees were downloaded from BirdTree.org [45] using the Hackett All Species backbone [50], which differs from the Ericson Stage 2 backbone used for Dataset 1. The BirdTree taxonomy lumps *Cyanistes*, *Periparus*, *Lophophanes*, and *Poecile* under *Parus*, so we maintain an explicit name mapping between IOC names used by Xeno-Canto and BirdTree tip names. Posterior trees were pruned to the focal 19-species set and a consensus tree was computed using median branch lengths, implemented in Python with BioPython [51] rather than the R-based workflow used for Dataset 1. The resulting tree has 19 tip nodes and 18 internal ancestral nodes. The Paridae clade differs from Tyrannidae in several respects: it includes both Old and New World species rather than being restricted to a single biogeographic region, its vocalizations tend to be simpler call notes and whistled songs rather than complex flycatcher vocalizations, and its phylogenetic structure is shallower with less divergent branch lengths. These differences, together with the distinct phylogenetic backbone and data curation pipeline, make it a stringent test of whether our method generalizes beyond the characteristics of a single dataset.

G.3 Pretrained Components

Our pipeline relies on two pretrained components that are frozen throughout all downstream experiments: a variational autoencoder and a neural vocoder. Neither component is fine-tuned or retrained for different datasets.

VAE. We use the pretrained 1D convolutional VAE from Make-An-Audio 2 [34] without any retraining or fine-tuning on bird audio. The model operates on 80-bin log-mel spectrograms and compresses each 5-second clip into a latent representation of shape 20×156, which we flatten to $D = 3{,}120$. The same frozen weights are used for both datasets.

Vocoder. Generated mel spectrograms are converted to 16 kHz waveforms using a pretrained BigVGAN vocoder [31]. BigVGAN is a high-fidelity neural vocoder based on large-receptive-field anti-aliased periodic activations, pretrained on general audio. We use the vocoder as a frozen black box: the VAE decoder produces an 80-bin mel spectrogram, and BigVGAN synthesizes the corresponding waveform without any fine-tuning on bird audio.

G.4 Tree-Metric Trait Learning

The linear trait projection $W \in \mathbb{R}^{K \times D}$ is trained to maximize the Pearson correlation between pairwise squared trait distances and patristic distances (Eq. 2). We optimize using Adam with learning rate 10^{-2} for 1,500 gradient steps. The free parameter $A \in \mathbb{R}^{K \times D}$ is initialized from $\mathcal{N}(0, 0.05^2)$ and the Cholesky regularizer is $\varepsilon = 10^{-6}$. We sweep over trait dimensions $K \in \{2, 4, 8, 16, 20, 21\}$ and select $K = 20$ based on downstream evaluation performance (Appendix A, Table 5). Species-level trait representatives are formed by trimmed mean aggregation with trim fraction $\alpha = 0.3$.

G.5 Ancestral Inference and Generation

Brownian ancestral inference uses nugget regularization $\lambda = 10^{-3}$. For each internal node, we draw $B = 3$ posterior samples with sampling temperature $\tau = 0.6$ and residual perturbation scale $\gamma = 0.25$. Inverse lifting uses nearest-trait anchoring: for each ancestral trait sample, the anchor is selected as the training clip whose projected trait is nearest in ℓ_2 distance. Generated latents are decoded through the pretrained VAE decoder, and the resulting mel spectrograms are converted to 16 kHz waveforms using a pretrained BigVGAN vocoder.

G.6 Baseline Configurations

All baselines use the same phylogenetic tree, the same trimmed-mean aggregation procedure with $\alpha = 0.3$ applied in each method's own embedding space, and BM posterior sampling with nugget $\lambda = 10^{-3}$, three posterior samples per node, and temperature $\tau = 0.6$. PhyloNN and Poincaré

both use 20-dimensional embeddings, matching our trait dimension $K{=}20$, so that differences in performance reflect the representation rather than capacity.

Raw Latent BM performs BM inference directly in the full 3,120-dimensional VAE latent space without any dimensionality reduction. Ancestral latents are decoded through the VAE decoder and vocoder.

BirdNET BM extracts 1,024-dimensional embeddings from a pretrained BirdNET classifier [12] and uses these as the trait space for BM inference. Internal-node outputs are rendered via nearest-exemplar lookup: for each inferred ancestor embedding, the training clip with the closest BirdNET embedding in L2 distance is selected and its VAE latent is decoded.

Classical Acoustic extracts a 33-dimensional feature vector per clip using `librosa` [52], comprising 20 MFCCs, spectral centroid, spectral bandwidth, spectral rolloff, spectral flatness, zero-crossing rate, RMS energy, and 7-band spectral contrast. BM inference is performed on species-mean acoustic features, and rendering uses nearest-exemplar lookup in acoustic feature space.

PhyloNN [28] maps VAE latents to a 20-dimensional embedding via a two-layer MLP encoder with hidden dimension 512 and ReLU activation. The network is supervised with a species-level cross-entropy head and additional hierarchical clade-classification heads at three tree-depth levels, weighted by $\beta = 0.5$. Training uses Adam with learning rate 10^{-3} for 800 epochs. BM inference is performed in the learned 20-dimensional space, and rendering uses nearest-exemplar lookup.

Poincaré [23] uses the same MLP architecture as PhyloNN but projects outputs onto the Poincaré ball. The encoder is trained with an all-pairs contrastive loss with margin 2.0: same-species pairs are pulled together and different-species pairs are pushed apart in Poincaré distance. Training uses Adam with learning rate 10^{-3} for 800 epochs. At inference, BM is performed in the tangent space at the origin via the logarithmic map, and rendering uses nearest-exemplar lookup.

Rendering pipeline for all retrieval baselines. For every retrieval baseline, nearest-exemplar lookup is performed in the method's own embedding space. The selected training clip is then rendered by decoding its precomputed VAE latent through the same VAE decoder and BigVGAN vocoder used by generative methods, so all outputs undergo identical VAE encode–decode distortion and audio quality comparisons are fair.

G.7 Evaluation Setup

All evaluation metrics are computed on the generated ancestral audio for the internal nodes, 20 for Dataset 1 and 18 for Dataset 2. For each internal node, the $B{=}3$ posterior samples are independently decoded to audio, embedded via BirdNET, and averaged to produce a single per-node embedding used in all phylogenetic plausibility metrics. FAD metrics pool all $B \times |\mathcal{U}|$ generated clips as the generated distribution. BirdNET [12] 1,024-dimensional embeddings serve as the primary evaluation representation for phylogenetic plausibility metrics. FAD metrics use three complementary embedding spaces: BirdNET at 1,024 dimensions, PANNs CNN14 [35] at 2,048 dimensions, and CLAP [36] with HTSAT-tiny backbone at 512 dimensions.

G.8 Computational Resources

All experiments are conducted on a single NVIDIA L4 GPU (24 GB). The computational cost of our method is modest: the only learned component is the linear trait projection $W \in \mathbb{R}^{K \times D}$, which is optimized for 1,500 Adam steps and converges in under 10 seconds per dataset. Baseline training (PhyloNN, Poincaré) requires 800 epochs each, completing in approximately 2–3 minutes per model. The full ancestral reconstruction pipeline for one dataset, comprising latent encoding, BM posterior inference, anchored inverse lifting, VAE decoding, and BigVGAN vocoding for all internal nodes with $B{=}3$ samples each, runs in approximately 60 seconds. Evaluation is dominated by embedding extraction: BirdNET inference runs on CPU via TFLite, while PANNs and CLAP each require a single forward pass on GPU, with all evaluation metrics for one dataset completing in under 5 minutes. The total wall-clock time for a complete run, training all methods, generating ancestral audio, and computing all metrics for both datasets, is under 30 minutes on a single L4 GPU.

H Evaluation Metric Definitions

We define the five phylogenetic plausibility metrics and five audio quality metrics used in the main text. All metrics are computed on the generated ancestral audio for internal nodes $u \in \mathcal{U}$, using BirdNET embeddings as the evaluation representation unless otherwise noted.

H.1 Phylogenetic Plausibility Metrics

All BirdNET embeddings are L2-normalized to unit length. Let $e_u \in \mathbb{R}^{D_{\text{bn}}}$ denote the BirdNET embedding of the generated audio for internal node u, and let e_s denote the BirdNET centroid embedding of real clips for extant species $s \in \mathcal{L}$. Define the cosine distance $d_{\cos}(a, b) = 1 - a^\top b$ for unit-normed vectors a, b.

Tree–Embedding Correlation. We compute the Spearman rank correlation between pairwise phylogenetic distances and pairwise cosine distances over all pairs of nodes $i, j \in \mathcal{U} \cup \mathcal{L}$, where internal nodes use their generated-audio embeddings and tips use their real-clip centroid embeddings:

$$\text{TreeCorr} = \rho_{\text{Spearman}}\big(\{d_T(i,j)\}_{i<j},\ \{d_{\cos}(e_i, e_j)\}_{i<j}\big). \tag{25}$$

Higher values indicate that the generated audio preserves the global geometric structure of the phylogeny. Including tip centroids tests whether generated ancestors are placed correctly relative to real species.

Descendant Affinity Margin. For each internal node u, let $\mathcal{D}_u \subset \mathcal{L}$ be its set of descendant tips and $\mathcal{N}_u = \mathcal{L} \setminus \mathcal{D}_u$ its non-descendants. We compute:

$$\text{DescMarg} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \Big[\bar{s}(u, \mathcal{D}_u) - \bar{s}(u, \mathcal{N}_u)\Big], \tag{26}$$

where $\bar{s}(u, \mathcal{S}) = |\mathcal{S}|^{-1} \sum_{s \in \mathcal{S}} e_u^\top e_s$ is the mean cosine similarity. Positive values indicate that ancestors are, on average, more similar in embedding space to their own descendants than to unrelated species.

Edge Monotonicity. At each binary internal node p with two children c_1, c_2 connected by branch lengths b_1, b_2, we check whether the child on the longer branch is also more distant from the parent in cosine space:

$$\text{EdgeMono} = \frac{1}{|\mathcal{B}|} \sum_{p \in \mathcal{B}} \mathbf{1}\big[(b_1 > b_2) \Leftrightarrow \big(d_{\cos}(e_p, e_{c_1}) > d_{\cos}(e_p, e_{c_2})\big)\big], \tag{27}$$

where $\mathcal{B}$ is the set of binary internal nodes with non-tied branch lengths, and children c_1, c_2 may be either internal nodes or tips. This measures whether branch lengths predict local acoustic divergence: a child separated by a longer branch should be more acoustically distant from the parent.

Nearest-Train Distance. For each internal node u, we compute the minimum cosine distance to any training clip in BirdNET embedding space:

$$\text{NNDist} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \min_{x \in \mathcal{X}_{\text{train}}} d_{\cos}(e_u, e_x), \tag{28}$$

where $\mathcal{X}_{\text{train}}$ denotes individual training clip embeddings rather than species centroids. Higher values indicate that the generated ancestral audio is not simply copying a nearby training exemplar, suggesting genuine interpolation rather than memorization.

Unique Rate. We measure the fraction of ancestors that receive distinct audio outputs by comparing MD5 hashes of the generated waveform files:

$$\text{Unique} = \frac{|\{\text{distinct hashes among } \hat{a}_u,\ u \in \mathcal{U}\}|}{|\mathcal{U}|}. \tag{29}$$

A Unique Rate below 1.0 indicates retrieval collapse, where multiple ancestors map to the same training exemplar and produce byte-identical waveforms. All generative methods achieve Unique Rate 1.0 by construction since they decode distinct latent vectors.

H.2 Audio Quality Metrics

Cycle Similarity (CycleSim). We measure VAE round-trip fidelity at the mel-spectrogram level: for each generated ancestral audio, we compute the mel spectrogram M, encode and decode it through the VAE to obtain the reconstructed mel $\hat{M} = D_{\text{vae}}(E_{\text{vae}}(M))$, and report the cosine similarity between the flattened originals and reconstructions:

$$\text{CycleSim} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\langle \text{vec}(M_u),\ \text{vec}(\hat{M}_u) \rangle}{\|\text{vec}(M_u)\| \cdot \|\text{vec}(\hat{M}_u)\|}. \tag{30}$$

Values close to 1 indicate that the generated audio lies on the VAE's decodable manifold and survives re-encoding without degradation.

Fréchet Audio Distance (FAD). We compute the Fréchet distance [46] between the distribution of generated ancestral audio embeddings and the distribution of real training clip embeddings, in three complementary embedding spaces:

- **FAD-BN**: BirdNET [12] 1,024-dimensional embeddings, capturing species-level bioacoustic identity.
- **FAD-PANN**: PANNs CNN14 [35] 2,048-dimensional embeddings, capturing general audio event semantics.
- **FAD-CLAP**: CLAP [36] 512-dimensional embeddings with HTSAT-tiny backbone, capturing audio–language aligned representations.

Each FAD is computed as $\text{FAD} = \|\mu_g - \mu_r\|^2 + \text{Tr}\big(\Sigma_g + \Sigma_r - 2(\Sigma_g \Sigma_r)^{1/2}\big)$, where μ_g, Σ_g and μ_r, Σ_r are the mean and covariance of the generated and real embedding distributions, respectively. Lower FAD indicates that the generated audio is more similar to real bird vocalizations in aggregate. Because the generated pool is small relative to the embedding dimensionality, the sample covariance Σ_g is rank-deficient, so absolute FAD values carry upward bias. Cross-method comparisons remain valid because all methods use identical pool sizes and the same reference distribution of $\sim$1,100 real clips.

Silence Rate. We compute the mean fraction of near-silent frames across all generated ancestral audio clips. A frame is considered silent if its root-mean-square energy falls below -60 dBFS. High silence rates indicate degenerate generation where the method produces partially or fully silent waveforms.

I Additional Validation

We further evaluate the learned trait space through phylogenetic signal, sensitivity to tree uncertainty, prediction from historical recordings, and held-out species tests on additional clades.

I.1 Phylogenetic Signal

We quantify phylogenetic signal in species-level representations using the multivariate generalization of Blomberg's K [27, 53]. We denote this statistic by K_{mult} to distinguish it from the trait dimension K. Its expected value under Brownian evolution is 1, providing a reference for interpreting the phylogenetic structure of each representation. Table 7 compares the learned trait space with PCA and raw VAE latents on both datasets.

Table 7: **Phylogenetic signal in species-level representations.** Multivariate Blomberg's K has expectation 1 under Brownian evolution. The largest value in each dataset is in **bold**.

Method	K_{mult} Tyrannidae	K_{mult} Paridae
PCA	0.46	0.75
Raw Latent BM	0.45	0.68
Ours	**1.09**	**0.96**

The learned trait space has the strongest phylogenetic signal among the compared representations on both datasets, with values close to the Brownian expectation. Classical acoustic features yield $K_{\text{mult}} = 0.33$ on Tyrannidae. These results complement the distance-based evaluation by showing that the learned traits capture phylogenetic structure under an established comparative statistic.

I.2 Robustness to Phylogenetic Uncertainty

To assess sensitivity to the inferred phylogeny, we sample 100 trees from the BirdTree posterior distribution [45] and repeat held-out species prediction. For each tree, we recompute the Brownian covariance from its topology and branch lengths and average localization AUC across held-out species. Table 8 reports the mean and standard deviation of these averages across trees, alongside the consensus-tree result from Table 3.

Table 8: **Sensitivity of our method to phylogenetic uncertainty.** Localization AUC on the consensus tree and mean $\pm$ standard deviation across 100 posterior trees.

Dataset	Consensus tree	100 posterior trees
Tyrannidae	0.781	0.789 ± 0.007
Paridae	0.734	0.725 ± 0.012

The posterior-tree averages remain close to the consensus-tree results, with small variation across trees. Held-out localization remains above chance on every sampled tree in both datasets, indicating that the predictive performance is not sensitive to the choice of a single consensus phylogeny.

I.3 Validation with Historical Recordings

We evaluate temporal trait prediction using 19 annual recordings spanning 1980–2012 from the Kent Island Savannah Sparrow archive associated with Williams et al. [54].[3] This archive documents cultural song change within a single population. We use calendar time as the temporal axis and perform leave-one-year-out prediction with Brownian motion. For each held-out year, we infer its representation from the remaining years and rank the observed annual recordings by their distance to the prediction in each method's representation space. We report localization AUC and mean rank across the 19 held-out years, with rank 1 indicating the closest match.

Table 9: **Held-out-year validation on historical Savannah Sparrow recordings.** Localization AUC and rank are averaged over 19 held-out years. Chance localization AUC is 0.5. Best values are in **bold**.

Method	locAUC ↑	Mean rank ↓
PCA	0.66	7.2
Raw Latent BM	0.66	7.1
Ours	**0.88**	**3.2**

The learned trait space improves both localization AUC and mean rank over the baselines (Table 9). This provides evidence that trait inference can recover acoustic variation associated with documented song change over several decades.

I.4 Validation on Additional Clades

We extend held-out species validation to two additional clades: 28 species of *Setophaga* wood-warblers, which have culturally learned songs, and 22 species of Hylidae treefrogs. Recordings are encoded with the same pretrained audio VAE, and each method predicts held-out species representations using Brownian inference on the corresponding phylogeny. The *Setophaga* evaluation is repeated over 100 posterior BirdTree phylogenies [45]. The Hylidae evaluation uses an amphibian phylogeny.

[3] Archive: `https://doi.org/10.5061/dryad.k98sf7m7x`.

Table 10: **Held-out species validation on additional clades.** Localization AUC is reported as mean ± standard deviation. Chance is 0.5. Best values are in **bold**.

Method	locAUC ↑ *Setophaga* (28 species)	Hylidae (22 species)
PCA	0.377 ± 0.022	0.670 ± 0.015
Raw Latent BM	0.447 ± 0.004	0.567 ± 0.016
Ours	$\mathbf{0.732 \pm 0.011}$	$\mathbf{0.869 \pm 0.010}$

The learned trait space achieves the highest localization AUC on both clades (Table 10). On *Setophaga*, its performance remains above chance while both baselines fall below 0.5. The Hylidae results extend this evidence beyond birds, supporting the applicability of trait learning and Brownian inference to other animal vocalizations.

J Anchor Dependence and Reconstruction Fidelity

J.1 Shared Acoustic Texture

Some reconstructed ancestors have similar spectrograms because the anchored lift preserves the nullspace component of the same nearest training clip. This shared component carries acoustic texture, including spectral detail and temporal structure. Similarity is also expected between nearby ancestral nodes within a clade whose extant species have similar vocalizations.

On Dataset 1, all 20 reconstructed ancestors have distinct waveform hashes. In raw mel space, the mean pairwise cosine distance is 0.024 and the minimum is 0.0001. In trait space, the mean pairwise distance is 23.3 and the minimum is 8.1. Thus, the minimum distance is approximately 0.4% of the mean in mel space and 35% in trait space. These comparisons show that strong visual similarity can coexist with separation in the learned phylogenetic representation.

J.2 Trait Recovery under Different Anchor Sources

We test how the source of the inherited texture affects recovery of the intended ancestral trait on Dataset 1. For each generated ancestor, we re-encode its decoded mel spectrogram with the VAE encoder and project the resulting latent through W. We compare this recovered trait with the BM targets of all 20 ancestors and record whether the nearest target belongs to the ancestor for which the output was generated. Chance performance is therefore 5%.

We repeat generation using three anchor sources: the nearest training clip in trait space, as in our method; a random training clip; and the farthest training clip in trait space. Table 11 reports the resulting target recovery rates.

Table 11: **Effect of anchor source on ancestral target recovery.** Traits are recovered from decoded mel spectrograms and matched against the 20 ancestral BM targets on Dataset 1. Nearest-clip anchoring is the default used by our method.

Anchor source	Target recovery ↑
Nearest clip	**75%**
Random clip	65%
Farthest clip in trait space	60%
Chance	5%

Target recovery remains well above chance under all three anchor choices, including when texture is inherited from the most distant available clip. This supports the ability of the anchored lift to retain the intended ancestral traits while changing the source of acoustic texture.

Traits are read directly from the decoded mel spectrograms before vocoding. The mel-to-waveform-to-mel round trip does not preserve the 20 trait dimensions reliably enough for this test, since the

vocoder is optimized for perceptual quality. Phylogenetic consistency at the waveform level is evaluated separately using the BirdNET metrics in Table 1.

J.3 VAE Reconstruction Reference

To assess reconstruction fidelity relative to the synthesis backend, we measure CycleSim for 1,005 real clips from Dataset 1 passed through the same VAE and vocoder. These reconstructed real clips achieve a CycleSim of 0.9987, compared with 0.9985 for our generated ancestors. The close agreement indicates that the generated outputs have VAE round-trip fidelity comparable to that of reconstructed real recordings, complementing the FAD comparisons in Table 2.

K Full Qualitative Results

Figures 8 and 9 show mel spectrograms for all 20 reconstructed ancestral nodes on Dataset 1 (Tyrannidae) across all six methods. Figures 10 and 11 show the corresponding results for all 18 ancestral nodes on Dataset 2 (Paridae). A central distinction visible throughout these figures is between methods that *generate* novel ancestral audio and those that *retrieve* existing training clips. The retrieval-based baselines, Classical Acoustic, PhyloNN, and Poincaré, operate in non-decodable feature spaces and must render internal-node outputs by selecting the nearest training exemplar. As a consequence, their spectrograms are visually sharp but frequently duplicated: multiple internal nodes map to the same training clip, and the outputs cannot represent genuinely intermediate ancestral sounds. This is reflected quantitatively in Table 1, where these methods exhibit Unique Rates well below 1.0.

By contrast, our method generates novel spectrograms for every internal node with Unique Rate 1.0. The anchored inverse lift ensures that each output inherits realistic acoustic texture from a nearby training clip via the nullspace component while its phylogenetic content is set by the BM posterior via the trait subspace. Across the 20 nodes, our method consistently produces spectrograms with richer harmonic structure, more naturalistic temporal patterning, and fewer silence artifacts compared to the baselines. The diversity of generated outputs, visible in the variation of spectral patterns across different ancestor nodes, demonstrates that the pipeline produces phylogenetically graded reconstructions rather than collapsing to a small set of prototypical sounds.

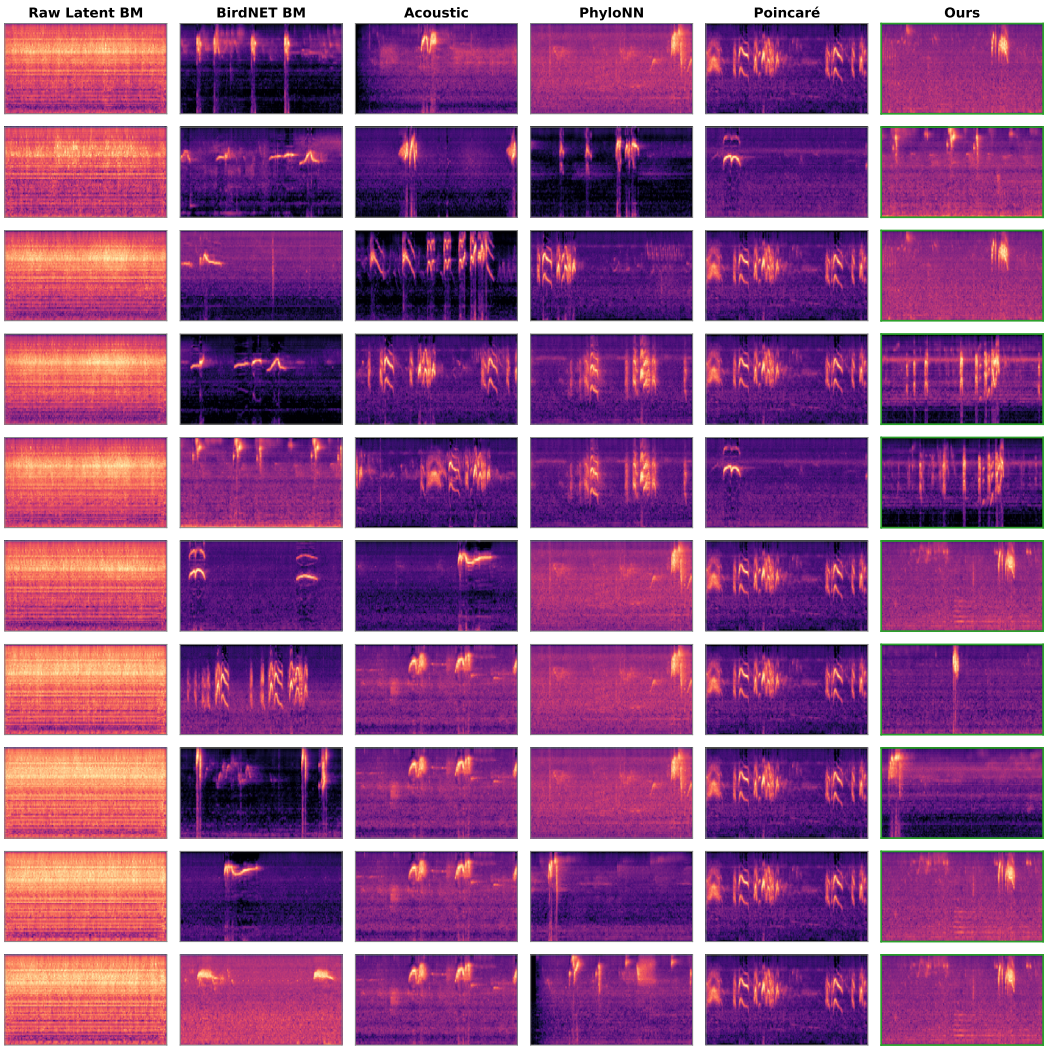


Figure 8: **Ancestral sound reconstructions for internal nodes 1–10 on Dataset 1, Family Tyrannidae.** Rows correspond to internal nodes labeled with ancestor index and genus; columns correspond to the six reconstruction methods. Our method (green border) consistently produces spectrograms with richer harmonic structure and fewer silence artifacts compared to baselines.

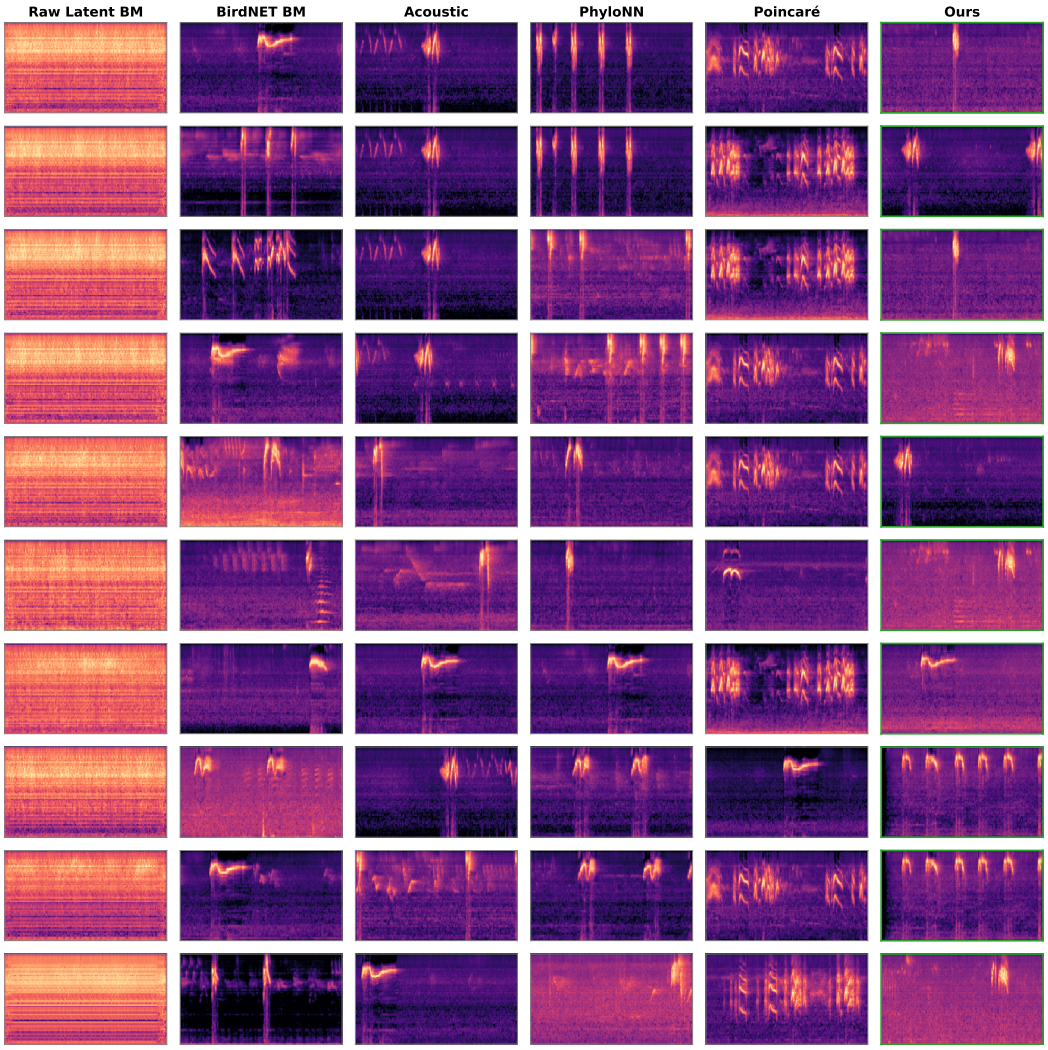


Figure 9: **Ancestral sound reconstructions for internal nodes 11–20 on Dataset 1.** Continuation of Figure 8.

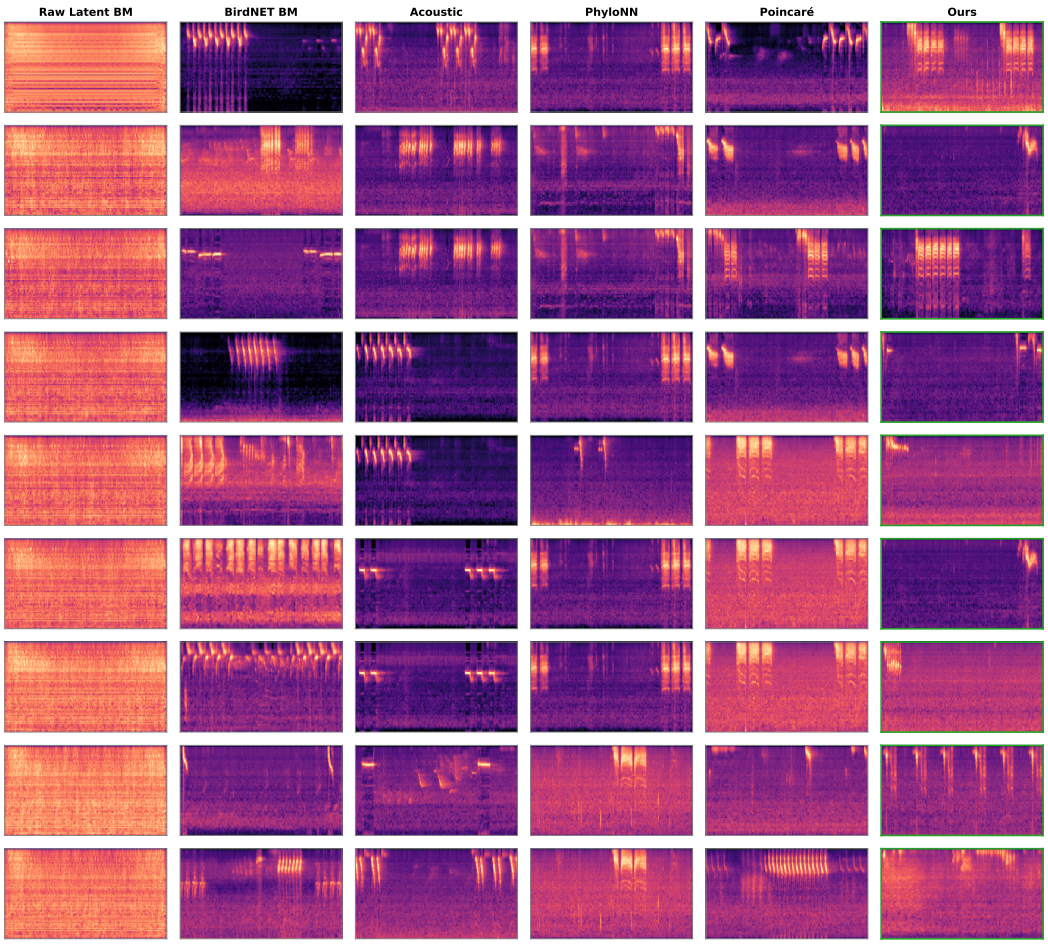


Figure 10: **Ancestral sound reconstructions for internal nodes 1–9 on Dataset 2, Family Paridae.** Same layout as Figures 8–9 but on Dataset 2. The same pattern holds: Raw Latent BM produces diffuse spectrograms, retrieval-based baselines duplicate training clips, and our method (green border) generates novel spectrograms with coherent harmonic structure.

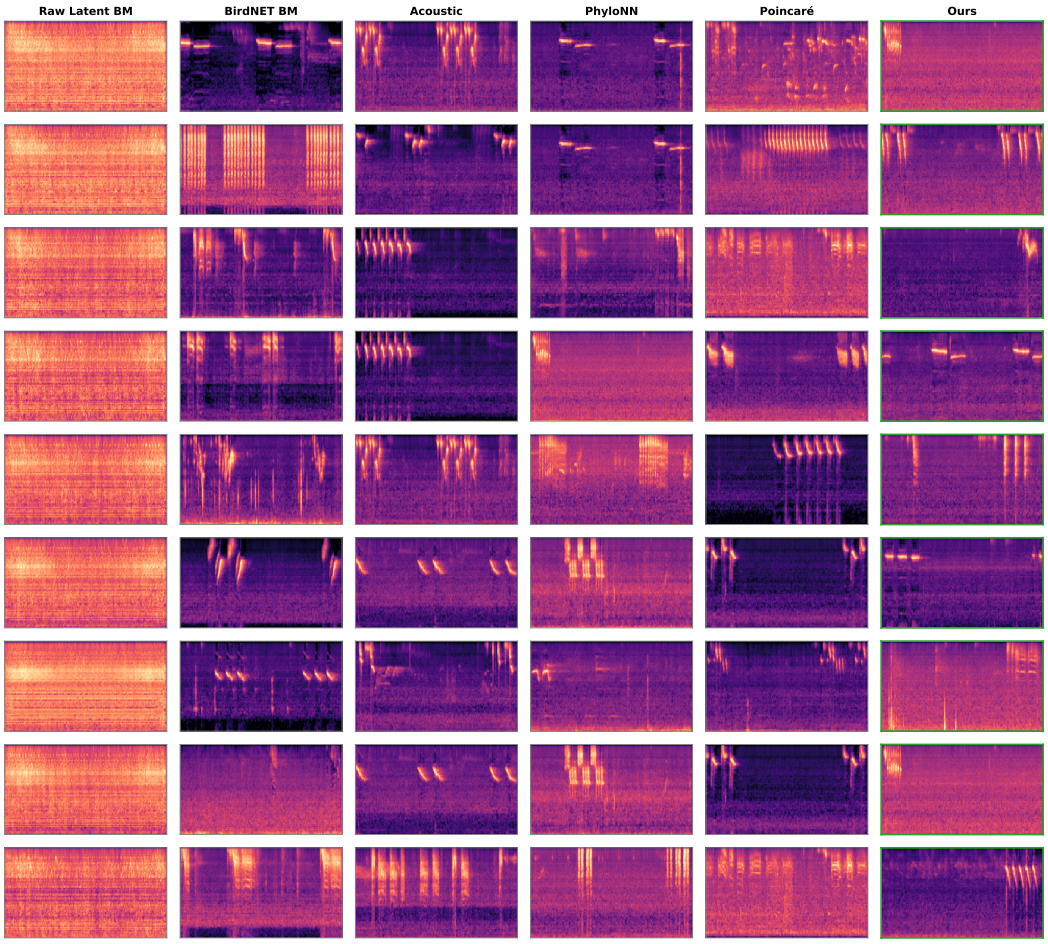


Figure 11: **Ancestral sound reconstructions for internal nodes 10–18 on Dataset 2, Paridae.** Continuation of Figure 10.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state four specific contributions supported by Tables 1, 2, and 3, Figure 3, and the analyses in Appendix I. Limitations are discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations paragraph in Section 6 discusses the lack of direct validation for deep ancestral nodes, anchor dependence, the scale of the generation experiments, and the absence of explicit population and cultural-transmission structure.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results state assumptions explicitly and have complete proofs in Appendices D and C. The row-orthonormality assumption $WW^\top = I_K$ is stated in each result that requires it.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix E provides the full algorithm pseudocode. Appendix G describes hyperparameters, baseline configurations, pretrained components, and training procedures for the two main datasets, which use publicly available Xeno-Canto recordings. Audio-generation metrics are defined in Appendix H. Additional validation protocols and data sources are described in Appendix I.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have released the source code at `https://github.com/solislemuslab/bird-sound-asr`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 reports the key hyperparameters. Appendix G provides experimental details including optimizer, learning rate, training steps, and baseline configurations. Additional validation protocols are described in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix I.2 reports statistical analysis of held-out localization AUC across 100 posterior phylogenies. Table 8 reports the mean and standard deviation across trees, quantifying variation due to phylogenetic uncertainty. The main generation tables report point estimates without confidence intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix G.8 reports all compute details: single NVIDIA L4 GPU, trait training under 10 seconds, full reconstruction pipeline approximately 60 seconds per dataset, total end-to-end under 30 minutes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. Both datasets use publicly available recordings from Xeno-Canto, an open-access repository of bird sounds with per-recording Creative Commons licenses. No human subjects, private data, or dual-use concerns are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work reconstructs ancestral bird vocalizations for evolutionary biology research. It generates non-human audio with no direct path to societal harm.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The models generate bird vocalizations only and pose no risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: All pretrained models, including Make-An-Audio 2 VAE, BigVGAN, BirdNET, PANNs, CLAP, and data sources Xeno-Canto and BirdTree.org are cited with original papers. Xeno-Canto recordings are used under their per-recording Creative Commons licenses. The historical Savannah Sparrow recordings are credited to the original study, with a link to the Dryad archive in Appendix I.3.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have released the source code at `https://github.com/solislemuslab/bird-sound-asr`. The repository includes documentation for installation, data preparation, and running the pipeline.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used in the core methodology. They were used only for writing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.