Beyond Family Labels: A Taxonomic Ornstein–Uhlenbeck Prior for Avian 3D Shape Recovery

Yunchan Jeon[1,*] **SooGon Kim**[1,*] **Seungwook Kim**[2] **CheolWon Lee**[3] **Jongmin Lee**[1,†]

[1]Department of Computer Science and Engineering, Chung-Ang University, South Korea
[2]GenGenAI, South Korea
[3]Department of Computer Engineering, Konkuk University, South Korea

{ycj00, gon1397, jmlee}@cau.ac.kr wookiekim@gengen.ai cheolwonlee@kku.ac.kr

[*]Equal contribution [†]Corresponding author

Abstract

Recovering 3D pose and shape of birds from single images is bottlenecked by the long tail of avian diversity: out of $\sim$10,000 species, only a few dozen have 3D-labeled training data, and in-the-wild test sets routinely contain species never seen at training. Generalizing across clades is therefore a central problem. The current state of the art, AniMer+ [1], mitigates it with a family-level contrastive loss that conveys only a binary same-family indicator per pair and discards the continuous structure of the species tree. We propose a **taxonomic Ornstein–Uhlenbeck prior** on the shape code of the AVES parametric bird mesh model: a multivariate Gaussian whose covariance decays exponentially with Linnaean rank distance, with a learnable selection strength and noise scale. The prior generalizes family contrastive into a continuous, generative form, exposing the graded cross- and within-family similarity that a binary indicator cannot, and brings continuous comparative phylogenetics in as a biology-grounded inductive bias for the long tail. Our method improves over AniMer+ on both metrics of the unseen-species CowBird benchmark and on CtrlAVES3D, while remaining comparable on seen-species CUB. A permutation ablation isolates taxonomic structure from generic Gaussian smoothing, and a rank-granularity ablation identifies cross-order distinctions (the structure beyond a family-vs-not indicator) as the dominant source of the unseen-species gain.

1 Introduction

Motivation: the long-tail problem in avian 3D recovery. Single-image articulated 3D reconstruction has progressed rapidly for animals, from species-agnostic fits of the Skinned Multi-Animal Linear (SMAL) model [3] to breed-aware priors [4], web-scale category models [5], and most recently to family-aware mixture-of-experts transformers for joint mammal–bird recovery [1]. Birds, however, sharpen the classical long-tail problem to an extreme. The world hosts on the order of 10 000 extant bird species, yet 3D-labeled training data exist for only a few dozen, and important in-the-wild benchmarks (e.g. CowBird [6]) contain species *never seen* at training. A practical estimator must therefore generalize across species: not just to new individuals or new poses, but to new clades (groups of species sharing a common ancestor) whose body plans the model has never been supervised on. Bridging this gap requires a regularizer that injects *prior knowledge* about how shape varies across species, knowledge that exists abundantly in evolutionary biology but is not currently exploited by single-view recovery pipelines.

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

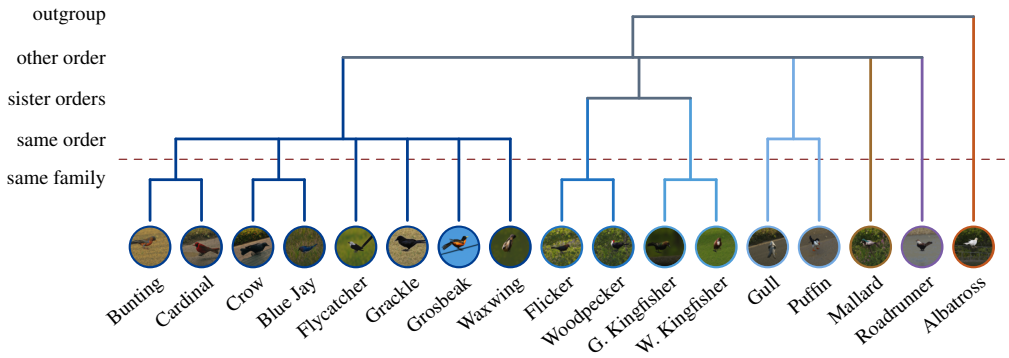

Figure 1: **Beyond family labels.** The 17 training species of CtrlAVES3D on the Linnaean tree [2], colored by order. The height at which two lineages join marks their lowest shared rank, from same family up to outgroup; each rank sets the distance d_{tax} that our prior reads (Figure 2b). The family-supervised contrastive loss of AniMer+ [1] keeps a single bit of this tree: whether a pair joins below the dashed red line. Our Ornstein–Uhlenbeck (OU) prior uses every rank, pulling the shape codes of two species together more strongly the more closely they are related, so the shape space carries graded structure rather than a same-family indicator.

Background problem: family supervised contrastive discards continuous taxonomic structure. The current bird state of the art, AniMer+ [1], addresses the long tail with a family-level supervised-contrastive (SupCon) [7] regularizer $\mathcal{L}_{\text{con}}$ on the latent shape code. By construction $\mathcal{L}_{\text{con}}$ ingests only a binary same-family indicator per pair and discards *all* taxonomic structure below and above the family cut. Two species in the same family are pushed equally close regardless of whether they diverged 5 or 50 million years ago; two species in different families are pushed equally far regardless of whether they share an order or sit on opposite ends of Aves. Yet decades of phylogenetic comparative methods in evolutionary biology [8, 9] show that continuous traits across species are well modeled by an Ornstein–Uhlenbeck (OU) process [10] on the species tree, with analytic covariance $\sigma^2/(2\alpha)\exp(-\alpha\, d_{\text{tax}})$. The Linnaean rank distance d_{tax} between species (Figure 1), derived from standard avian taxonomy [2], is a biologically grounded quantity, available globally from existing classification, and informed at the cross-order level by the whole-genome avian phylogeny [11], precisely the signal a generalization-bound estimator should be using.

Our approach: a taxonomic OU prior on the AVES shape code. We propose a **taxonomic Ornstein–Uhlenbeck prior** on the AVES [12] shape code β (the low-dimensional shape coefficients of the AVES mesh) that replaces the $\mathcal{L}_{\text{con}}$ used by AniMer+ [1]. Per mini-batch, we look up pairwise Linnaean rank distances from a fixed Linnaean rank distance matrix [2], assemble the analytic OU covariance $\Sigma_{ij} = (\sigma^2/2\alpha)\exp(-\alpha d_{\text{tax}}(i,j))$, and add a multivariate-Gaussian negative log-likelihood (Eq. (2)) to the AniMer+ objective. The selection strength α and the noise scale σ are learnable scalars, parameterized through softplus [13], and are the only new covariance parameters introduced. Conceptually, the OU prior is a *generative continuous extension* of the metric-learning $\mathcal{L}_{\text{con}}$ used by AniMer+: in the limit $\alpha \to \infty$ with a block taxonomy that collapses each family to a point, Σ becomes a family-indicator block matrix and $\mathcal{L}_{\text{OU}}$ degenerates into a Gaussian form structurally analogous to the family contrastive objective. The analogy is structural: SupCon L_2-normalizes its feature space and the Gaussian negative log-likelihood (NLL) does not. Finite α (our learned regime) instead allows graded cross-family leakage and continuous within-family granularity, exchanging a binary same-family indicator for a rank-structured Gaussian likelihood with $\binom{17}{2} = 136$ pairwise values. We summarize our contributions as follows:

- A taxonomic Ornstein–Uhlenbeck [10] prior on the AVES [12] shape code that generalizes the family contrastive signal of AniMer+ [1] into a continuous, generative form indexed by Linnaean rank distance.
- The proposed method improves over AniMer+ on both metrics of the unseen-species CowBird benchmark and on CtrlAVES3D, while remaining comparable on seen-species CUB.

- A controlled permutation ablation isolates taxonomic structure from generic Gaussian smoothing.
- A rank-granularity ablation identifies cross-order Linnaean distinctions, the structure beyond a family-vs-not indicator, as the dominant source of the unseen-species gain.

2 Related Work

Animal 3D pose and shape. Parametric mesh models—SMAL [3] (a multi-animal extension of SMPL [14]) for quadrupeds and AVES [12] for birds—are the dominant approach for animal 3D reconstruction. Category-specific mesh reconstruction [15] and wild-animal fitting [16, 17] established the learning-based paradigm. Single-image regressors have since evolved from early bird-specific reconstruction [6] through breed-aware (BARC [4]) and mixture-of-Gaussians (WLDO [18]) shape priors to category-level (MagicPony [19]) and pan-category (3D-Fauna [5]) models. For birds, AniMer+ [1] (extending the mammal-only AniMer [20]) is the current state of the art: a family-aware mixture-of-experts [21] transformer [22] trained jointly on mammals and birds with a family-level supervised-contrastive [7] $\mathcal{L}_{\text{con}}$ term. We keep the AniMer+ backbone but replace the discrete $\mathcal{L}_{\text{con}}$ used by AniMer+ with a continuous rank-structured generative prior, and show that continuous taxonomic structure matters, both theoretically and empirically. $\mathcal{L}_{\text{con}}$ is the structural analogue of our prior in the degenerate $\alpha \to \infty$ case (Section 3.4), and a controlled permutation ablation tests the taxonomic content.

Taxonomic regularization and phylogenetic priors in ML. Hierarchical classification [23, 24] embeds class-level structure into learning objectives, and image-based methods exploit taxonomy discretely, including the breed triplet loss of BARC [4] and the family-level supervised contrastive [7] loss of AniMer+; the latter two are pair losses with a 1-bit-per-pair information budget. Such losses can be read as implicit class-conditional Gaussians with a block-diagonal covariance (one block per family). The OU [10] process as a model of continuous trait evolution on a tree is classical in phylogenetic comparative methods [8, 9, 25]; tree-structured OU priors have appeared in Draupnir [26], a VAE for ancestral protein sequences, but to our knowledge not in 3D articulated shape estimation. Our prior replaces the block-diagonal with a dense, rank-structured Σ indexed by a Linnaean [2] rank distance, retaining the discrete class-block as the $\alpha \to \infty$ degenerate case (Section 3.4).

Empirical phylogenetic priors on bird shape. Wang et al. [12] reported a strong phylogenetic signal (Pagel's [27] λ=0.99, $p<10^{-4}$) on low-dimensional embeddings of the AVES shape principal components, evidence that articulated avian shape codes carry continuous phylogenetic structure even under per-image keypoint supervision. The multi-modal Gaussian shape prior of WLDO [18] is the discrete-mixture analogue of our continuous OU [10] kernel: a mixture of M Gaussians (M=10 in WLDO) whose block structure is recovered as the $\alpha \to \infty$ block-taxonomy limit of our prior (Section 3.4). Our prior generalizes both lines along two axes simultaneously, moving from a discrete taxonomy to a continuous one and from a metric-based formulation to a generative likelihood.

3 Method

Overview. Figure 2 sketches the method. **(a)** The AniMer+ [1] encoder, initialized from the released checkpoint, produces per-image shape codes $\boldsymbol{\beta}$ for AVES [12], a multi-species parametric bird mesh whose shape is controlled by a 15-dimensional coefficient vector. **(b)** Across the species in our training split, we model the joint distribution of $\boldsymbol{\beta}$ as a multivariate Gaussian whose covariance follows an Ornstein–Uhlenbeck (OU) kernel [8–10] over a fixed Linnaean [2] rank distance matrix D_{tax}. The kernel is parameterized by two learnable scalars (α, σ) and a learnable species-mean μ, yielding the OU regularizer $\mathcal{L}_{\text{OU}}$. **(c)** $\mathcal{L}_{\text{OU}}$ replaces AniMer+'s family contrastive term and is added to the original task losses at strength λ_{OU}. Restoring the contrastive term and setting $\lambda_{\text{OU}} = 0$ recovers AniMer+ to floating-point equality, so the OU contribution is isolated in a single hyperparameter.

The remainder of this section reviews the OU process and the taxonomic distance it consumes (Section 3.1), formalizes the OU kernel on shape codes (Section 3.2) and the resulting NLL loss (Section 3.3), discusses the relationship to family supervised contrastive [7] (Section 3.4), and specifies how the prior integrates with AniMer+ (Section 3.5).

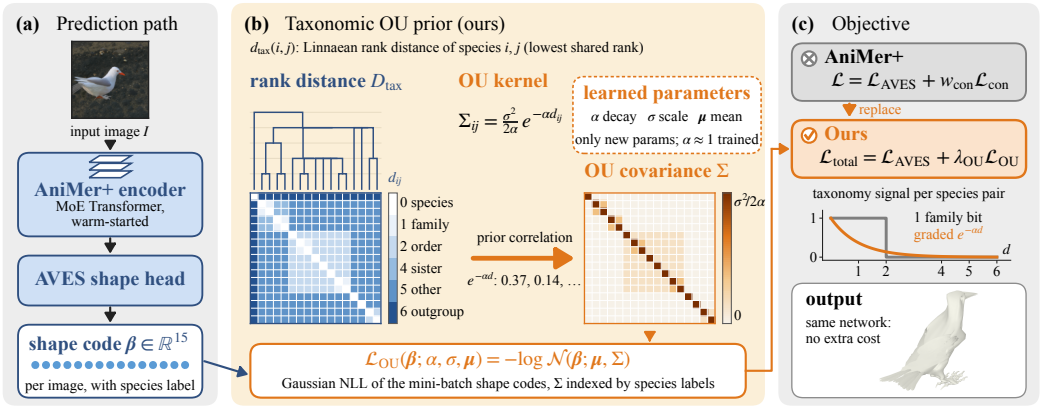

Figure 2: **Method at a glance. (a)** The AniMer+ [1] encoder produces AVES [12] shape codes $\boldsymbol{\beta}$ from a single image. **(b)** Across species, $\boldsymbol{\beta}$ is regularized by an OU [10] prior parameterized by learnable $(\alpha, \sigma, \boldsymbol{\mu})$ over a fixed Linnaean [2] rank distance matrix; the NLL of this Gaussian is $\mathcal{L}_{\text{OU}}$. **(c)** $\mathcal{L}_{\text{OU}}$ replaces the family-level contrastive term of AniMer+ (w_{con}=0), so that on birds $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{AVES}} + \lambda_{\text{OU}}\mathcal{L}_{\text{OU}}$. Blue: the existing network and fixed taxonomy. Orange: what we add.

3.1 Preliminaries: OU process and taxonomic distance

Intuition. Closely related species tend to have similar body shapes, and the similarity fades as the relationship becomes more distant. An OU process is a standard model in comparative phylogenetics for a continuous trait that evolves along a species tree while being pulled toward an optimum; under it the trait values of two species are correlated, and the correlation decays exponentially with their distance on the tree. The Linnaean rank distance supplies that distance from standard taxonomy alone, as a small integer that grows as the pair's shared group moves up from species to family, order, and beyond. The resulting covariance matrix over species is the Gaussian prior we place on the AVES shape code.

OU process. The Ornstein–Uhlenbeck process [10] is a stationary mean-reverting stochastic process whose covariance between two time points decays exponentially with their separation. In comparative phylogenetics, OU processes have long been the standard model of continuous trait evolution on a species tree [8, 9, 28]: the parameter α controls selection strength (pull toward an optimum) and σ scales the diffusion noise. Setting $\alpha \to 0$ recovers the Brownian-motion limit with covariance proportional to shared branch length [25]; $\alpha > 0$ bounds the stationary variance to $\sigma^2/(2\alpha)$.

Taxonomic distance. An OU model takes as input a pairwise relatedness distance $d_{\text{tax}}(i, j)$ between species i and j. We use the *Linnaean rank distance* based on standard avian taxonomy [2]: $d_{\text{tax}}(i, j) = 0$ for the same species, 1 for the same family, 2 for the same order, and 4–6 for cross-order pairs. The cross-order values are informed by the whole-genome avian phylogeny [11], and 5–6 are ordinal design choices (Appendix S4). Restricted to the 17 species in CtrlAVES3D [29] this yields a fixed, non-learnable 17×17 float32 matrix D_{tax} fed into the OU kernel below.

Why Linnaean ranks rather than a molecular phylogeny? The Jetz et al. avian supertree [30] would seem the natural input to an OU kernel since comparative phylogenetics classically consumes patristic distances (path lengths through the tree). Two scale-conditioning arguments justify the discrete rank distance instead. **(i) Conditioning.** On this 17-species subset Jetz patristic distances span the range $[18.6, 201.2]$ million years; for $\exp(-\alpha\, d)$ to express meaningful cross-family correlation on that scale, α must live near 10^{-2}, leaving a near-diagonal covariance in the regime where our optimizer converges ($\alpha \sim 1$). Linnaean [2] ranks ($\{1, 2, 4, 5, 6\}$ off-diagonal) lie on the same order as the converged α, yielding meaningful across-species sharing without rescaling. **(ii) Signal preservation.** Linnaean ranks correlate strongly with Jetz patristic distances on this subset (Pearson 0.85, Spearman 0.76), so the discrete rank kernel keeps most of the biological signal of a continuous molecular kernel while gaining a much better-conditioned covariance. The Jetz tree is used only for this check and for the evaluation in Section 4.6; training sees D_{tax} alone. Further discussion of the conditioning argument and of length-scale-normalized variants is in Appendices S4.3 and S10.

3.2 OU kernel on shape codes

Why this kernel. We want a species-level prior that (a) places shape codes of related species near each other in $\mathbb{R}^{15}$, (b) does so with a strength that decays smoothly with taxonomic distance rather than at a hard family cut, and (c) keeps a global "how much can shape vary" knob separate from the relational "how similar are sisters" knob. The OU kernel satisfies all three by construction.

Instantiating the OU process on the species taxonomy gives a stationary multivariate Gaussian over the tips whose (i, j) covariance is [8, 9]:

$$\mathrm{Cov}(\beta_i, \beta_j) \;=\; \frac{\sigma^2}{2\alpha} \exp\big(-\alpha\, d_{\mathrm{tax}}(i,j)\big), \tag{1}$$

where $d_{\mathrm{tax}}(i, j)$ is the Linnaean rank distance from Section 3.1. The prefactor $\sigma^2/(2\alpha)$ is the stationary variance of the OU process and sets the prior's marginal scale on each species; the exponential factor $\exp(-\alpha d)$ encodes the relational decay with taxonomic distance. Large α collapses the kernel to sharp family-like blocks; small α smears similarity across the entire taxonomy. Both scalars are learned via softplus parameterizations [13] enforcing positivity (with floor $\alpha_{\min} = 10^{-3}$) and initialized at $\alpha_0 = \sigma_0 = 1$ (the raw pre-softplus parameter is set to the inverse softplus of 1).

3.3 NLL loss

Why an NLL. Treating $\mathcal{L}_{\mathrm{OU}}$ as the negative log-likelihood of a Gaussian (rather than, say, a metric-learning surrogate) gives the prior a generative interpretation that downstream analysis exploits: the same Σ that appears in the loss is the Gaussian-process (GP) covariance whose maximum-a-posteriori (MAP) reading we use in Appendix S2.1. Generative form also makes the $\alpha \to \infty$ limit comparison to discrete family contrastive clean (Section 3.4).

Let $\boldsymbol{\beta} \in \mathbb{R}^{B \times D}$ be the AVES shape codes predicted for a mini-batch of B instances (here $D = 15$, taxonomy size $N = 17$), and let $\Sigma_{ij} = (\sigma^2/2\alpha) \exp(-\alpha\, d_{\mathrm{tax}}(s_i, s_j)) + \varepsilon \mathbf{1}[i = j]$ with jitter $\varepsilon = 10^{-6}$. Treating the D shape dimensions as independent samples from a shared species covariance, the prior, written as a loss, is

$$\mathcal{L}_{\mathrm{OU}} \;=\; \frac{D}{2} \log|\Sigma| \;+\; \frac{1}{2} \sum_{d=1}^{D} (\boldsymbol{\beta}_{\cdot,d} - \mu)^{\top} \Sigma^{-1} (\boldsymbol{\beta}_{\cdot,d} - \mu) \tag{2}$$

The mean $\mu \in \mathbb{R}^D$ is also learnable, so the prior pulls $\boldsymbol{\beta}$ toward the empirical species mean rather than toward a fixed origin. Numerical implementation is detailed in Appendix S3.

Before turning to the integration with AniMer+, we first relate this prior to the family supervised-contrastive loss it generalizes—examining the kernel under the degenerate $\alpha \to \infty$ limit makes the connection precise.

3.4 Relationship to family supervised contrastive

Intuition: what does $\alpha \to \infty$ mean? Recall that α controls how fast the kernel $\exp(-\alpha\, d)$ decays with taxonomic distance. Sending α to infinity is the limit in which the prior treats every same-family pair as *interchangeable* (d=0 kept, all $d > 0$ off-diagonals driven to zero) and every cross-family pair as *unrelated* (correlation exactly zero). This is precisely the binary view of taxonomy that the family-level supervised-contrastive loss $\mathcal{L}_{\mathrm{con}}$ of AniMer+ [1] encodes: same-family pairs pulled together, all other pairs pushed apart, with no graded gradation. We make this connection precise below.

Limit form of the kernel. Under the block-taxonomy convention $d_{\mathrm{tax}}(i, j) = 0$ for same-family pairs and $d_{\mathrm{tax}}(i, j) > 0$ otherwise, taking $\alpha \to \infty$ collapses the OU kernel $(\sigma^2/2\alpha) \exp(-\alpha\, d_{\mathrm{tax}}(i, j))$ to a family-indicator block matrix scaled by the stationary variance $\sigma^2/(2\alpha)$, held fixed in the limit. The block-indicator Σ pulls same-family $\boldsymbol{\beta}$ toward a shared family mean while leaving different-family codes uncorrelated, the same qualitative effect that supervised contrastive learning [7] induces in feature space. Within a family, the joint marginal of the codes is a degenerate Gaussian along the family direction. In this sense $\mathcal{L}_{\mathrm{OU}}$ supplies a Gaussian probabilistic relaxation of the binary same-family indicator that $\mathcal{L}_{\mathrm{con}}$ encodes.

Table 1: **Comparison against AniMer+ [1] on the avian benchmarks.** CUB [32] and the CtrlAVES3D [29] test split evaluate *seen*-species; CowBird [6] is the *unseen*-species out-of-distribution (OOD) benchmark. For ours we report the mean ± sample standard deviation over five random seeds. AniMer-A [20] is the avian-only variant; AniMer+ and ours are unified bird+mammal models; AVES and HMR2.0 are the baselines reported by AniMer+.

	CUB		CowBird		CtrlAVES3D
Method	AUC ↑	PCK ↑	AUC ↑	PCK ↑	AUC ↑
AVES [12]	79.70	25.20	46.10	5.60	85.90
HMR2.0 [33]	87.70	50.00	59.10	12.20	91.60
AniMer-A [20]	91.90	74.80	65.70	19.40	92.90
AniMer+ [1]	**89.50**	**59.40**	59.30	16.40	93.00
Ours	89.30 ± 0.02	57.84 ± 0.09	$\mathbf{60.80 \pm 0.10}$	$\mathbf{18.86 \pm 0.27}$	$\mathbf{93.05 \pm 0.01}$

Caveats: the analogy is structural, not literal. Two technical differences prevent equality in the limit. **(i) Singular covariance.** At $\alpha \to \infty$ the block-indicator Σ is rank-deficient, so the strict Gaussian NLL is ill-defined; one must take a pseudo-inverse with jitter $\varepsilon > 0$ (we use $\varepsilon = 10^{-6}$) and analyze the limit as $\varepsilon \to 0$. **(ii) Feature space.** SupCon [7] operates on L_2-normalized class-token features through a cosine-softmax objective, whereas $\mathcal{L}_{\mathrm{OU}}$ operates on raw shape codes $\boldsymbol{\beta} \in \mathbb{R}^{15}$ through a quadratic Mahalanobis form. The two losses are best read as *structurally analogous regularizers acting on different representations*, both encouraging same-family proximity. A proof sketch of the limit and the structural comparison are in Appendix S2.

Why this matters. Our optimizer converges to finite $\alpha \sim 1$, not infinity. At that operating point, within-family pairs have correlation $\exp(-\alpha) \approx 0.37$, same-order pairs $\exp(-2\alpha) \approx 0.14$, and so on, a graded similarity decay across the taxonomic hierarchy. The $\alpha \to \infty$ analogy clarifies what SupCon [7] is doing in the limit; the finite-α regime our prior actually inhabits is what our paper studies, and what $\mathcal{L}_{\mathrm{con}}$ cannot represent.

3.5 Integration with AniMer+

Existing AniMer+ pipeline. The training loop of AniMer+ [1] returns a base objective $\mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{AVES}} + \mathcal{L}_{\mathrm{SMAL}} + w_{\mathrm{con}}\,\mathcal{L}_{\mathrm{con}}$ that combines per-keypoint, silhouette, and shape-and-pose regression terms with the family supervised-contrastive regularizer [7]. $\mathcal{L}_{\mathrm{AVES}}$ and $\mathcal{L}_{\mathrm{SMAL}}$ are the avian and mammalian branches [3, 12, 14, 31], and $\mathcal{L}_{\mathrm{SMAL}}$ is inactive in our avian-only configuration. The encoder is a mixture-of-experts (MoE) Vision Transformer (ViT) [21, 22] pretrained across mammal and avian datasets.

Our addition. We augment $\mathcal{L}_{\mathrm{base}}$ with our OU prior at strength λ_{OU}:

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}_{\mathrm{base}} \;+\; \lambda_{\mathrm{OU}}\,\mathcal{L}_{\mathrm{OU}}(\boldsymbol{\beta}_{\mathrm{pred}},\, f_{\mathrm{batch}},\, d_{\mathrm{tax}}), \tag{3}$$

where $\boldsymbol{\beta}_{\mathrm{pred}}$ is the AVES [12]-branch shape code and f_{batch} is the per-sample species index. In our best setting we set $w_{\mathrm{con}} \leftarrow 0$, so the OU prior is the sole taxonomy-aware regularizer. With w_{con} left at its original value, setting $\lambda_{\mathrm{OU}} = 0$ recovers the AniMer+ [1] training loop to floating-point equality, isolating the OU contribution to a single hyperparameter.

Why the OU group needs a separate learning rate. We place the OU scalars (α, σ, μ) in their own AdamW parameter group whose learning rate (LR) is k times the backbone LR; we call k the OU-LR multiplier. Eval is nearly k-invariant at a given λ_{OU} across $k \in \{1, 10, 30\}$, confirming that the OU prior's eval contribution comes primarily from the regularization *shape* at the initialization $(\alpha_0, \sigma_0) = (1, 1)$ rather than from learned scalar updates, independent evidence against an (α, σ)-fitting artifact. We use $k = 100$ in all headline runs.

4 Experiments

Section 4.1 fixes the training and evaluation setup. Section 4.2 reports the headline result. Section 4.3 probes whether the gain is genuinely taxonomic through a label-permuted control, and Section 4.4

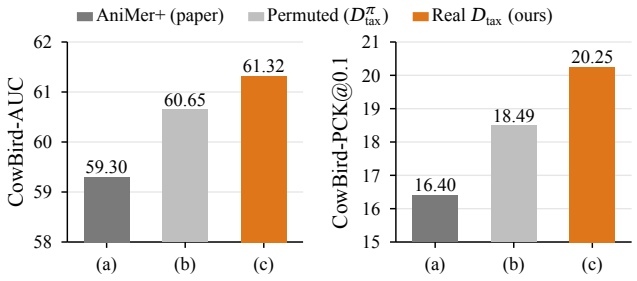

Figure 3: **Real vs. permuted taxonomy ablation.** CowBird [6] AUC and PCK@0.1 under three configurations: (a) the AniMer+ [1] baseline, (b) the OU prior with a label-permuted distance matrix, and (c) the OU prior with the real Linnaean [2] distance matrix (both OU runs at $\lambda_{\text{OU}} = 10^{-4}$, constant learning rate; Section 4.1).

sharpens the question into a rank-granularity ablation. Section 4.5 reports 3D vertex-level errors as a complementary check, and Section 4.6 verifies that the OU prior tightens the shape codes in β-space without distorting their phylogenetic signal. Qualitative comparisons follow in Section 4.7.

4.1 Training and evaluation setup

We fine-tune from the released AniMer+ [1] checkpoint on the 6 464-image train split of CtrlAVES3D [29] for 30,000 steps with AdamW [34] and a cosine learning-rate schedule [35]. The training data covers 17 avian species; evaluation uses three benchmarks following the protocol of AniMer+ (Table VI): CUB-200-2011 [32] (CUB) for seen-species fine localization, CowBird [6] for unseen-species OOD, and the CtrlAVES3D test split for 3D reconstruction. Following AniMer+, we report the percentage of correct keypoints (PCK) at threshold 0.1 (PCK@0.1), i.e. the fraction of predicted 2D keypoints whose normalized distance to the ground truth is below 0.1, and the area under the PCK curve (AUC) as the threshold sweeps from 0 to 1. Our model is warm-started from the unified AniMer+ [1] bird+mammal checkpoint and fine-tuned on the avian split (mammal-branch weight set to zero). Headline numbers (Tables 1 and 3) use $\lambda_{\text{OU}} = 10^{-5}$ with the cosine schedule; the real-vs-permuted and rank-granularity ablations (Figure 3 and Table 2) use a stronger setting, $\lambda_{\text{OU}} = 10^{-4}$ with a constant learning rate, so that the taxonomic content of D_{tax} is expressed strongly enough for the contrast to be visible. Full implementation details (fp32 training, two-group AdamW with OU-LR multiplier, gradient clipping, jitter and softplus [13] initialization, the permutation-matrix control protocol, and the integration with AniMer+'s training loop) are in Appendix S9.

4.2 Main result: unseen-species gains over AniMer+

Table 1 compares the proposed method against the avian numbers reported by AniMer+ [1]. Our method improves over the AniMer+ paper on both unseen-species CowBird metrics ($+1.5$ AUC, $+2.5$ PCK@0.1) and on CtrlAVES3D, and remains comparable on seen-species CUB. Improvements concentrate on the unseen-species CowBird [6] benchmark, where the cross-species borrowing supplied by the OU prior is most useful. AniMer-A [20] is a bird-only specialist that sacrifices the unified multi-domain coverage of AniMer+; the like-for-like comparison is therefore against the unified AniMer+ baseline.

4.3 Real vs. permuted taxonomy: is the gain genuinely taxonomic?

A regularizer that improves OOD performance by injecting any shape-code structure, even random structure, would not constitute evidence for the taxonomic hypothesis. Figure 3 addresses this concern by keeping the training configuration fixed and changing only whether the OU kernel sees the biologically-correct distance matrix or a label-permuted control. The permutation preserves the OU-kernel spectrum exactly, with the same covariance eigenvalues, but destroys the correspondence between species labels and their taxonomy positions.

The real-vs-permuted comparison is the key result. The OU prior with the real Linnaean [2] distance matrix outperforms the permuted control on both CowBird-AUC and CowBird-PCK@0.1 [6]. The AniMer+ [1] bar is the value reported in the original paper rather than a run from our pipeline, so its gap to the OU runs is not a controlled comparison and is shown for context. The OU prior is therefore using *which species are which*, not merely the fact that an OU kernel was applied. The same real-vs-permuted contrast holds on the 183 CUB species outside the fine-tuning taxonomy, with species-clustered confidence intervals excluding zero (Appendix S5.3).

Table 2: **Rank-granularity ablation**: progressively richer distance matrices on the same operating point as Figure 3. The family-only kernel uses values in $\{0, 1, 5\}$ (same family vs. not), the order-only refinement uses $\{0, 1, 2, 5\}$ (adds the same-order distinction), and the full Linnaean kernel uses $\{0, 1, 2, 4, 5, 6\}$ (cross-order ordinals as the default). All three kernels use 0 for the same species and 1 for the same family, and differ only above the family level; genus-level structure cannot be varied on this training set, because the 17 species contain no congeneric pair. We use 5 rather than 6 as the cross-order placeholder for the family-only and order-only kernels so that they share a single cross-order bin, isolating the rank-granularity axis from changes in cross-order distance scale.

	CUB [32]		CowBird [6]		CtrlAVES3D [29]
D_{tax} variant	AUC ↑	PCK ↑	AUC ↑	PCK ↑	AUC ↑
permuted (null)	88.87	54.90	60.65	18.49	92.95
family-only	89.09	56.35	60.84	19.09	93.06
order-only	89.10	56.42	60.81	19.20	93.07
Linnaean (full, ours)	88.92	55.16	**61.32**	**20.25**	92.99

4.4 Beyond family labels: rank-granularity ablation

Table 2 sharpens the previous question into a quantitative one: how much of the unseen-species gain depends on information *beyond* a family-vs-not indicator? This also empirically tests the structural claim of Section 3.4: the family-only D_{tax} row is the discrete $\alpha \to \infty$ limit of our prior (analogous to family SupCon), and the full Linnaean row exposes the graded cross-rank structure that our finite-α regime was designed to capture. We compare three progressively richer versions of the distance matrix, namely a family-only block, an order-only refinement, and the full Linnaean ranks, all under the same training configuration as Section 4.3. The three rank-aware kernels separate cleanly along a single axis: how much taxonomic detail beyond family membership is exposed to the prior. Two readings stand out.

Cross-order ranks carry the unseen-species gain. The family-only and order-only kernels each recover only a fraction of the CowBird [6] improvement that the full Linnaean [2] kernel achieves; the additional cross-order rank distinctions (e.g. separating the duck order Anseriformes from the woodpecker order Piciformes) supply the dominant share of the unseen-species gain. Concretely, ranks above the family level account for about 66% of the CowBird-PCK improvement over the permuted null, and cross-order ranks alone for about 60%. This is exactly the structural information that a family-vs-not indicator *cannot* represent, and the empirical loading on these ranks is why the title is "beyond family labels".

Seen vs. unseen trade is consistent with cross-species borrowing. The CUB [32]-PCK pattern is the converse: family-only and order-only matrices retain more seen-species fine localization than the full Linnaean [2] kernel, which sacrifices a small amount of CUB-PCK to amplify the CowBird [6] gain. This is the prior's intended trade-off: finer cross-order granularity tightens cross-species mixing, which helps unseen species at a small cost on seen-species fine localization. The mechanism is a property of the *structure* of D_{tax}, not a generic regularization effect.

4.5 3D vertex-level errors on CtrlAVES3D

The headline table reports 2D AUC and PCK; for completeness we additionally evaluate Procrustes-aligned [31] 3D errors on the CtrlAVES3D [29] test split. Table 3 reports the Procrustes-aligned mean per-joint and per-vertex position errors (pa-MPJPE and pa-MPVPE) against the AniMer+ released checkpoint on the same eval pipeline. The 2D-keypoint gain in Table 1 is reflected in the 3D mesh: the OU prior reduces both pa-MPJPE and pa-MPVPE relative to the baseline (Δ pa-MPJPE$=-5.72$, Δ pa-MPVPE$=-9.96$ in canonical units; Table 3). The real-vs-permuted d_{tax} contrast shows the same sign as the CowBird-PCK ablation in Figure 3 (Δ pa-MPVPE$=-1.33$ in favor of real d_{tax}).

4.6 What the prior changes in β-space

The shape codes of the pre-trained model already place related species close together. Does the OU prior distort this structure? It does not (Figure 4).

Table 3: **Procrustes-aligned 3D errors on the CtrlAVES3D test split** (309 **images,** 17 **species**). pa-MPJPE denotes the mean per-joint position error; pa-MPVPE denotes mean per-vertex position error. Lower is better. Ours is a single run (seed 0). Values follow our canonical-frame scaling; see Appendix S6 for details.

Variant	pa-MPJPE ↓	pa-MPVPE ↓
AniMer+ [1]	655.61	708.97
AniMer+ *w/* OU (ours)	**649.89**	**699.01**

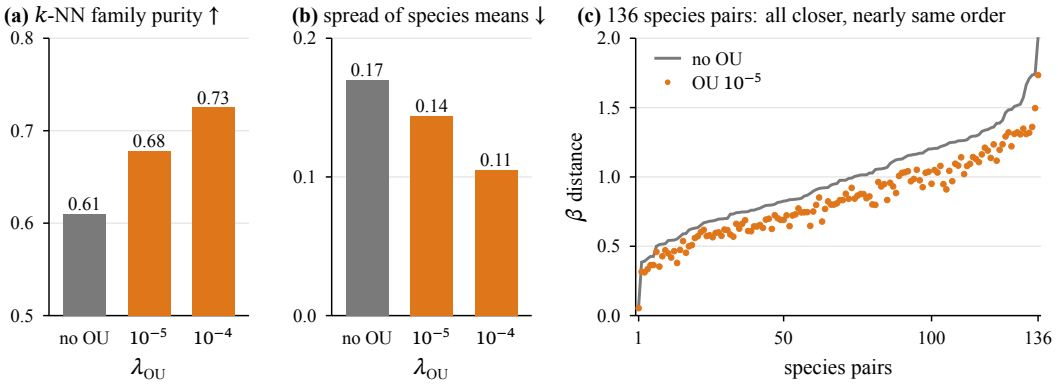

Figure 4: **The prior tightens the shape codes without distorting phylogenetic signal.** (a, b) What tightens: k-NN family purity rises and the spread of the per-species mean $\boldsymbol{\beta}$ falls with λ_{OU} on CtrlAVES3D [29] training codes; the 10^{-4} bars are a single constant-LR run. (c) What is preserved: the $\boldsymbol{\beta}$ distance between the mean shape codes of two species measures how different the model predicts their shapes to be. All 136 species pairs, sorted along the x-axis by this distance without the prior (gray), move closer at $\lambda_{\mathrm{OU}}{=}10^{-5}$ (orange, seed 0) but stay in nearly the same order.

What is preserved. We compare the per-species mean codes with the Jetz [30] phylogeny, an external tree that training never sees; training uses only the Linnaean rank matrix of Section 3. Two standard statistics measure how well distance on the tree predicts distance between codes: the Mantel correlation [36] r over the 136 species pairs and the multivariate Blomberg's K [28, 37]. Both are significantly above a 999-permutation null for the released AniMer+ checkpoint ($r{=}0.36$, $K_{\mathrm{mult}}{=}0.58$) and are essentially unchanged with the prior ($r{=}0.352 \pm 0.002$, $K_{\mathrm{mult}}{=}0.578 \pm 0.001$ over five seeds; Appendix S1). The structure therefore comes from pre-training, and the prior keeps it.

What changes. The codes become tighter. The fraction of each code's ten nearest neighbors that share its family (k-NN purity) rises from 0.61 without the prior to 0.68 at the headline strength and 0.73 at $\lambda_{\mathrm{OU}}{=}10^{-4}$, and the spread of the per-species means shrinks ($0.17 \to 0.14 \to 0.11$). All 136 species pairs move closer, by $0.85\times$ on average, and their ordering is nearly preserved (Spearman 0.99), which is why r and K_{mult} barely move. Real and permuted D_{tax} give the same r, K_{mult} and spread at both strengths; they differ only in performance (Section 4.3).

In short, the prior contracts the shape space and tightens each family's neighborhood, but does not rearrange the relatedness among species that pre-training built.

4.7 Qualitative comparison

Figure 5 compares the AniMer+ [1] baseline, retrained with its published recipe for the same 30,000 steps, against the proposed OU fine-tune on two CUB-200 [32] test images and two unseen-species CowBird [6] instances on which the baseline misplaces several keypoints. In all four cases the OU fine-tune moves the keypoint set toward the ground truth, raising per-image PCK@0.1 by 0.45 to 0.50 on CUB and by 0.20 to 0.43 on CowBird. Further paired examples are reported in Appendices S7 and S8.

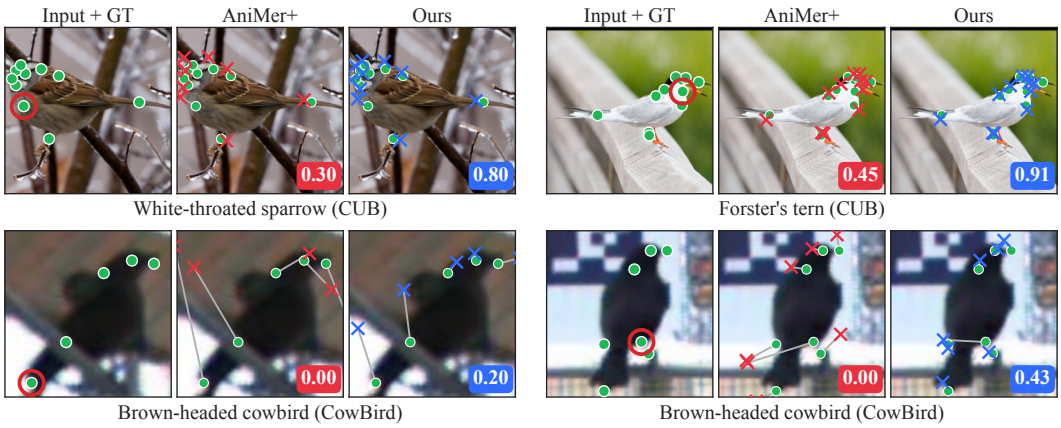

Figure 5: **Qualitative comparison on two CUB [32] test images (top) and two CowBird [6] test instances (bottom).** For each species, the first panel shows the input with the ground-truth keypoints (green dots) and a red circle on the baseline's worst keypoint; the retrained AniMer+ baseline (red crosses) and the OU fine-tune (blue crosses) follow, and gray lines join each prediction to its ground-truth keypoint. The badge is the per-image PCK@0.1.

5 Conclusion

We introduced a taxonomic Ornstein–Uhlenbeck [10] prior on the AVES [12] shape code for single-view avian 3D shape recovery. The prior generalizes the family supervised contrastive [7] signal used by the prior state of the art [1] into a continuous, generative form indexed by a Linnaean [2] rank distance matrix, supplying graded cross-species similarity that a binary same-family indicator cannot represent. As a unit, the proposed method improves over the prior state of the art on both metrics of the unseen-species CowBird benchmark and on CtrlAVES3D while remaining comparable on seen-species CUB, isolates the contribution of taxonomic structure from generic Gaussian smoothing through a controlled permutation ablation, and identifies cross-order rank distinctions, the information that lies beyond a family-vs-not indicator, as the dominant source of the unseen-species gain. Taken together, the experiments support a biology-grounded inductive bias for long-tailed avian shape recovery.

Limitations. The prior is evaluated on a 17-species training taxonomy with limited within-order density and no congeneric pair, so genus-level structure is untested, and the unseen-species benchmark, CowBird [6], covers a single species. The rank matrix D_{tax} follows a single taxonomic authority [2]; we have not tested robustness to competing resolutions of contested inter-order relationships, and larger-scale use should regenerate the matrix as taxonomic consensus evolves. The OU formulation is not bird-specific, but extending it to other animal groups requires a compatible parametric shape model, taxonomic annotations and 3D-labeled data, and how the gains behave at hundreds or thousands of species remains to be established.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25598824); by the NRF grant funded by the Korea government (MSIT) (No. RS-2024-00412141); and by the International Science Business Belt R&D Program funded by the Korea government (MSIT) (RS-2026-25548978). The authors declare no competing interests.

References

[1] Liang An, Jin Lyu, Li Lin, Pujin Cheng, Yebin Liu, and Xiaoying Tang. AniMer+: Unified pose and shape estimation across Mammalia and Aves via family-aware transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3233–3249, 2026. doi: 10.1109/TPAMI.2025.3633828. ↪ pp. 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 15, 16, 17, 19, 20, 23, 25, 26, 27, 28, 29, 31, and 33

[2] F. Gill, D. Donsker, and P. Rasmussen. IOC world bird list (v14.2), 2024. URL https://www.worldbirdnames.org/. ↪ pp. 2, 3, 4, 7, 8, 10, 16, 17, 21, and 33

[3] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.586. ↪ pp. 1, 3, 6, and 33

[4] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. ↪ pp. 1 and 3

[5] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3D fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.00931. ↪ pp. 1 and 3

[6] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G. Pfrommer, Marc F. Schmidt, and Kostas Daniilidis. 3D bird reconstruction: A dataset, model, and shape recovery from a single view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. doi: 10.1007/978-3-030-58523-5_1. ↪ pp. 1, 3, 6, 7, 8, 9, 10, 20, 23, 24, 26, 27, 28, 32, and 33

[7] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673, 2020. ↪ pp. 2, 3, 5, 6, 10, 17, 18, 19, 30, and 32

[8] Thomas F. Hansen. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341–1351, 1997. doi: 10.1111/j.1558-5646.1997.tb01457.x. ↪ pp. 2, 3, 4, 5, 15, 18, 19, and 25

[9] Marguerite A. Butler and Aaron A. King. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695, 2004. doi: 10.1086/426002. ↪ pp. 2, 3, 4, and 5

[10] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. *Physical Review*, 36(5):823–841, 1930. doi: 10.1103/PhysRev.36.823. ↪ pp. 2, 3, 4, 10, 15, 17, 18, 19, 22, 24, 25, and 27

[11] Erich D. Jarvis, Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, Brant C. Faircloth, Benoit Nabholz, Jason T. Howard, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014. doi: 10.1126/science.1253451. ↪ pp. 2, 4, and 21

[12] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: Capturing avian shape models from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14734–14744, 2021. doi: 10.1109/CVPR46437.2021.01450. ↪ pp. 2, 3, 4, 6, 10, 17, and 33

[13] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001. ↪ pp. 2, 5, 7, and 30

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013. ↪ pp. 3 and 6

[15] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. doi: 10.1007/978-3-030-01267-0_23. ↪ p. 3

[16] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00416. ↪ p. 3

[17] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-D safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00546. ↪ p. 3

[18] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. doi: 10.1007/978-3-030-58621-8_12. ↪ p. 3

[19] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.00849. ↪ p. 3

[20] Jin Lyu, Tianyi Zhu, Yi Gu, Li Lin, Pujin Cheng, Yebin Liu, Xiaoying Tang, and Liang An. AniMer: Animal pose and shape estimation using family aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17486–17496, 2025. doi: 10.1109/CVPR52734.2025.01629. ↪ pp. 3, 6, 7, and 26

[21] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. ↪ pp. 3 and 6

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. ↪ pp. 3, 6, and 20

[23] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12503–12512, 2020. doi: 10.1109/CVPR42600.2020.01252. ↪ p. 3

[24] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. doi: 10.1109/WACV.2019.00073. ↪ p. 3

[25] Joseph Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, 1985. doi: 10.1086/284325. ↪ pp. 3 and 4

[26] 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 (ICLR)*, 2022. URL https://openreview.net/forum?id=FZoZ7a31GCW. ↪ p. 3

[27] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999. doi: 10.1038/44766. ↪ p. 3

[28] Simon P. Blomberg, Theodore Garland, and Anthony R. Ives. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, 57(4):717–745, 2003. doi: 10.1111/j.0014-3820.2003.tb00285.x. ↪ pp. 4, 9, 15, and 16

[29] Liang An, Jin Lyu, Li Lin, Pujin Cheng, Yebin Liu, and Xiaoying Tang. CtrlAVES3D: A synthetic dataset of birds with pixel-aligned 3d annotations (introduced in AniMer+). Dataset released with [1], 2025. ↪ pp. 4, 6, 7, 8, 9, 15, 16, 20, 21, 23, 25, 26, 32, and 33

[30] 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. ↪ pp. 4, 9, 15, 16, 20, 22, 31, and 33

[31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00744. ↪ pp. 6, 8, 16, 25, and 26

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. ↪ pp. 6, 7, 8, 9, 10, 20, 23, 24, 28, 29, 32, and 33

[33] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14737–14748, 2023. doi: 10.1109/ICCV51070.2023.01358. ↪ p. 6

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. ↪ pp. 7, 20, 29, and 32

[35] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. ↪ pp. 7 and 29

[36] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220, 1967. ↪ pp. 9 and 15

[37] 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. ↪ pp. 9, 15, and 16

[38] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. ↪ p. 19

[39] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 328–339, 2018. ↪ p. 24

[40] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433. ↪ p. 31

[41] Hemal Naik, Alex Hoi Hang Chan, Junran Yang, Mathilde Delacoux, Iain D. Couzin, Fumihiro Kano, and Máté Nagy. 3D-POP: An automated annotation approach to facilitate markerless 2D-3D tracking of freely moving birds with marker-based motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21274–21284, 2023. doi: 10.1109/CVPR52729.2023.02038. ↪ p. 31

Appendix of

Beyond Family Labels: A Taxonomic Ornstein–Uhlenbeck Prior for Avian 3D Shape Recovery

---

Contents

S1 Calibration: phylogenetic signal in β-space . . . . . p. 15
S2 Theoretical refinement: MAP view and structural comparison to SupCon . . . . . p. 17
S3 Extended method details . . . . . p. 19
S4 The taxonomic distance matrix D_{tax} . . . . . p. 20
S5 Extended experimental results . . . . . p. 22
S6 Vertex-level errors on CtrlAVES3D . . . . . p. 25
S7 CowBird OOD per-image error distribution . . . . . p. 26
S8 Extended qualitative analysis . . . . . p. 28
S9 Training implementation details . . . . . p. 29
S10 Extended discussion: limitations and future work . . . . . p. 31

NeurIPS Paper Checklist . . . . . p. 32

S1 Calibration: phylogenetic signal in β-space

Section 4.6 reports that the OU prior leaves the phylogenetic signal of the shape codes intact while tightening their local structure. This appendix gives the underlying numbers and two further cross-checks on the trained shape codes: under OU, the codes keep their alignment with the Jetz tree, their family k-NN purity rises, and their rank correlation with the discrete Linnaean distance falls slightly.

Why these three diagnostics. We ask what an Ornstein–Uhlenbeck (OU) [8, 10] prior trained on a discrete Linnaean distance does to the geometry of the codes. The Mantel correlation [36] r and the multivariate Blomberg's K (K_{mult}) [28, 37] measure how strongly the per-species mean code $\bar{\boldsymbol{\beta}}_s \in \mathbb{R}^D$ follows the Jetz [30] tree: r correlates pairwise code distance with patristic distance over the 136 species pairs, and K_{mult} compares the dispersion of the species means with its Brownian-motion expectation under the tree covariance ($K{=}1$ at the expectation, 0 for no phylogenetic signal). Spearman correlation between the empirical $\boldsymbol{\beta}$-distance matrix and a candidate phylogenetic distance matrix probes the global rank ordering. The candidates are the Linnaean rank d_{tax}, its family- and order-only collapses, the Jetz [30] patristic distance in millions of years (My)[1], the same distance length-scale-normalized, and label-permutation controls. k-NN purity by Linnaean family/order (Table S3) probes local neighborhood structure. The three diagnostics are sensitive to different things, which is exactly why we report all of them.

Summary of evidence presented in this section.

- *Phylogenetic signal on the Jetz tree:* unchanged by the prior (K_{mult} $0.581 \to 0.578 \pm 0.001$, Mantel r $0.358 \to 0.352 \pm 0.002$ over five seeds; Table S2), and closely matched for real and permuted D_{tax}.
- *Local cluster purity:* OU monotonically tightens local within-family neighborhoods. Family k-NN purity rises from 0.610 without the prior to 0.678 at our headline configuration, and further under stronger regularization.
- *Spearman global rank correlation:* drops marginally (-0.012 on Jetz, -0.038 on Linnaean) under OU.

Together, these results give a coherent picture: OU keeps the continuous-tree alignment where pre-training put it (Mantel r, K_{mult} and Jetz Spearman flat) while tightening local within-family neighborhoods (purity up) and loosening the discrete rank correspondence slightly (Linnaean Spearman down). This supports the paper's k-invariance reading (Section 3.5, Table S8) that OU's eval contribution comes from optimization shaping rather than from a wholesale restructuring of the latent shape manifold.

Setup. For each trained variant, we run inference on the 6 464-image CtrlAVES3D [29] train split (which carries species labels for all $N = 17$ species) and aggregate the predicted shape codes into a per-species mean $\bar{\boldsymbol{\beta}}_s$, giving an $N \times D$ trait matrix that we plug into the Mantel, K_{mult} and Spearman analyses. This per-species mean is the right maximum-likelihood contraction under the OU model (individuals within a species are exchangeable) but does discard within-species variation that the model has no taxonomic signal about.

Three observations. *(a) AniMer+ [1] baseline already carries phylogenetic signal* ($\rho_{\text{Jetz}}{=}{+}0.358$). The signal is already present in the released checkpoint (Table S2), so it comes from pre-training rather than from the prior.

(b) Spearman rank correlation against the Jetz tree is essentially unchanged (margin: -0.012 for cosine + OU vs. AniMer+ retrain). Spearman is invariant to monotone re-scalings of pairwise distances, so this flat result is consistent with the near-uniform contraction in Figure 4(c): the species means move closer without re-ordering the relationships already there in the baseline. Family k-NN purity (Table S3), a local statistic, rises while the global statistics (Spearman, Mantel r, K_{mult}) stay flat. Flat global ordering and higher local purity are therefore complementary, not contradictory.

[1] $1\,\text{My} = 10^6$ years.

Table S1: **Spearman correlation ρ between β-distance and phylogenetic distance.** Per-species means computed from inference on the 6464-image CtrlAVES3D [29] train split, N=17 species, D=15 shape dims; correlated against various phylogenetic distance matrices across six variants. **Bold** = the strongest non-permuted phylogeny within each row. The Jetz [30] patristic distance dominates the rank in every variant.

Variant	Linnaean [2]	Family	Order	Jetz [30] (My)	Jetz norm	Perm (s_{42})	Perm (s_{101})	Perm (s_{202})
AniMer+ [1] retrain @30K (no OU)	+0.106	+0.119	+0.166	**+0.358**	+0.358	+0.138	−0.152	−0.201
Cosine LR (no OU)	+0.105	+0.116	+0.167	**+0.354**	+0.354	+0.136	−0.154	−0.202
Constant LR + OU λ=10^{-5}	+0.070	+0.121	+0.136	**+0.354**	+0.354	+0.129	−0.164	−0.189
Cosine LR + OU λ=10^{-5} (**ours**)	+0.068	+0.119	+0.135	**+0.346**	+0.346	+0.123	−0.170	−0.189
Stronger OU regime (λ=10^{-4}, single seed)								
λ=10^{-4} + real d_{tax}	+0.030	+0.129	+0.094	**+0.330**	+0.330	+0.080	−0.174	−0.154
λ=10^{-4} + perm d_{tax}	+0.032	+0.131	+0.106	**+0.339**	+0.339	+0.081	−0.171	−0.153

(c) OU-trained codes are slightly less aligned with the discrete Linnaean rank distance (ρ_{Linnaean}: +0.068 for cosine + OU vs. +0.106 baseline). The permuted control shows the same drop (+0.032 vs. +0.030 for real d_{tax} at λ_{OU}=10^{-4}), so we do not read it as a signature of the taxonomic kernel.

Phylogenetic signal on the Jetz tree. Table S2 reports Mantel r and K_{mult} on the per-species mean codes against the Jetz tree, together with the spread of the species means.

Table S2: **Phylogenetic signal of the per-species mean shape codes on the Jetz [30] tree.** K_{mult}: multivariate Blomberg's K [28, 37] (1 = Brownian expectation, 0 = no phylogenetic signal); Mantel r: correlation between pairwise $\boldsymbol{\beta}$ distance and patristic distance over the 136 species pairs; $\sigma_{\bar{\beta}}$: mean per-dimension s.d. of the per-species means. Permutation p in parentheses (999 label shuffles).

Variant	K_{mult} (p)	Mantel r (p)	$\sigma_{\bar{\beta}}$
AniMer+ [1] released checkpoint	0.583 (.026)	+0.360 (.008)	0.169
Cosine LR (no OU)	0.581 (.034)	+0.358 (.006)	0.170
Cosine LR + OU λ=10^{-5} (**ours**, seed 0)	0.577 (.037)	+0.351 (.007)	0.144
Cosine LR + OU λ=10^{-5}, permuted d_{tax}	0.577 (.036)	+0.351 (.016)	0.144
Constant LR + OU λ=10^{-4}, real d_{tax}	0.618 (.023)	+0.317 (.032)	0.105
Constant LR + OU λ=10^{-4}, permuted d_{tax}	0.627 (.019)	+0.328 (.028)	0.105

At the headline strength both statistics coincide with the no-OU recipe and with the released checkpoint, so the prior neither adds nor removes phylogenetic signal; what moves is the spread of the species means (0.170 → 0.144), the tightening visible in Figure 4. At $\lambda_{\text{OU}} = 10^{-4}$ the species means collapse further ($\sigma_{\bar{\beta}} = 0.105$) while both statistics stay significant against the permutation null (Mantel r 0.32, K_{mult} 0.62); real and permuted d_{tax} remain closely matched at both strengths.

Real vs. permuted matched at the strongest regularization strength. Under the strongest OU regularization in Table S2, the real- and permuted-taxonomy variants are *closely matched* on this diagnostic (K_{mult} 0.618 vs. 0.627, Mantel r +0.317 vs. +0.328, $\Delta\rho_{\text{Jetz}} = -0.009$). Both pull $\boldsymbol{\beta}$ codes into similar continuous-tree geometry; the CowBird-PCK gap of +1.76 pp between them (Figure 3) lives elsewhere in the model's representations, plausibly in the encoder weights themselves rather than in the regression head's output codes. The (α, σ) trajectory diagnostic (Figure S4) and the Procrustes-aligned [31] vertex-error scale (Table S9) show the same near-identity on independent metrics.

Local cluster purity (k-NN). The global statistics above look at the whole code distribution at once; here we zoom in and ask a simpler local question — if you pick a $\boldsymbol{\beta}$ code and look at its ten nearest neighbors in code space, how many of them belong to the same bird family?

For each predicted $\boldsymbol{\beta}$ code on the CtrlAVES3D train split, we compute the fraction of its k-NN (k=10) that share its Linnaean family (or order) label. We use a random subsample of 1,500 codes and take distances in the raw 15-dimensional code space. This k-NN purity is a local-neighborhood-sensitive diagnostic standard in latent-space embedding evaluation.

Table S3: **$\boldsymbol{\beta}$-space clustering quality by Linnaean [2] rank label.** Higher k-NN purity = $\boldsymbol{\beta}$-codes of same family/order are locally more clumped. Stronger OU [10] regimes monotonically increase purity.

Variant	k-NN purity (family)	k-NN purity (order)	Silhouette (family)
AniMer+ [1] retrain @30K (no OU)	0.621	0.714	-0.043
Cosine LR (no OU)	0.610	0.705	-0.042
Cosine LR + OU λ=10^{-5}	0.678	0.755	-0.034
λ=10^{-4} + real d_{tax}	**0.725**	**0.793**	$\mathbf{-0.018}$

Reading. The OU prior consistently *increases* local cluster purity, relative to the cosine-LR run without the prior. Family-level purity rises by $+6.8$ percentage points (pp) at our headline configuration and by $+11.5$ pp under stronger regularization; order-level purity rises by $+5.0$ pp and $+8.8$ pp. Silhouette scores are negative for all variants (the $\boldsymbol{\beta}$ codes of different families overlap substantially), but k-NN purity, a more local, cluster-shape-agnostic measure, isolates the OU-induced tightening of within-family neighborhoods.

Reconciling Spearman / purity / phylogenetic signal. The three measurements probe different aspects of the code geometry:

- *Global rank correlation* (Spearman ρ, Table S1): drops -0.012 under OU.
- *Phylogenetic signal* (Mantel r, K_{mult}, Table S2): unchanged under OU (r $0.358 \to 0.351$, K_{mult} $0.581 \to 0.577$).
- *Local cluster purity* (k-NN, this table): rises by $+6.8$ pp at our headline configuration (within-family neighborhoods are locally tighter).

Together: under OU the $\boldsymbol{\beta}$ codes have tighter within-family local structure and slightly looser discrete-rank ordering, while the phylogenetic signal inherited from pre-training is preserved.

S2 Theoretical refinement: MAP view and structural comparison to SupCon

The main text gives the OU prior a working definition and an empirical justification. This appendix sharpens the theoretical reading along three axes: a Bayesian maximum-a-posteriori (MAP) correspondence (Appendix S2.1) that recasts the regularizer as a tempered Gaussian prior rather than an ad-hoc penalty; a structural comparison to $\mathcal{L}_{\text{con}}$ [7] (Appendix S2.2) showing OU exposes a 5-level ordinal where SupCon sees only a binary indicator; and a clean restatement of the $\alpha \to \infty$ limit (Appendix S2.3) that handles the singular-covariance pitfall the main text glosses over and recovers SupCon as a degenerate boundary case. Together they form the structural substrate for the claims in Section 3.4 and Appendix S3.

Why does this matter empirically? The main text's real-vs-permuted-d_{tax} ablation (+0.67 CowBird AUC, +1.76 PCK; Figure 3) is the empirical lever the theory predicts. A label permutation preserves the kernel spectrum and the number of species pairs at each rank distance, but destroys which species occupy them, i.e. the 5-level ordinal alignment between species and rank distance. That ordinal-alignment information is exactly the signal this theory identifies as $\mathcal{L}_{\text{OU}}$'s contribution beyond $\mathcal{L}_{\text{con}}$. The remainder of this section makes that signal precise.

S2.1 MAP estimator equivalence

Is the OU regularizer really a Bayesian prior in disguise, or just a convenient quadratic penalty? We show it is the former: minimizing $\mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{OU}}$ is exactly MAP under a tempered Gaussian prior with temperature $1/\lambda$. The temperature interpretation gives λ a meaning beyond "regularization strength."

Let $I = (I_1, \ldots, I_B)$ be a mini-batch of images with species labels $s \in \{1, \ldots, N\}^B$ (N=17), and $\boldsymbol{\beta} \in \mathbb{R}^{B \times D}$ the predicted AVES [12] shape codes. The task losses of AniMer+ [1] (keypoint L_1, silhouette L_2, AVES regression) define an empirical NLL of I given $\boldsymbol{\beta}$ that we treat, up to constants

in $\boldsymbol{\beta}$, as the empirical image NLL

$$\mathcal{L}_{\text{base}}(\boldsymbol{\beta}; I) \;=\; -\log p_{\text{img}}(I \mid \boldsymbol{\beta}) + C_I, \tag{S1}$$

where C_I collects $\boldsymbol{\beta}$-independent terms. With kernel $\Sigma_{ij} = \frac{\sigma^2}{2\alpha} \exp\big(-\alpha\, d_{\text{tax}}(i,j)\big)$ on the B-tuple of training species and a shared mean $\mu \in \mathbb{R}^D$, place the OU prior [8, 10]

$$p(\boldsymbol{\beta} \mid \alpha, \sigma, \mu, d_{\text{tax}}) \;=\; \prod_{d=1}^{D} \mathcal{N}\big(\boldsymbol{\beta}_{\cdot,d} \mid \mu\,\mathbf{1}, \Sigma\big). \tag{S2}$$

Proposition 1 (MAP correspondence)**.** *For any $\lambda > 0$, the minimizer of $\mathcal{L}_{\text{total}}(\boldsymbol{\beta}) = \mathcal{L}_{\text{base}}(\boldsymbol{\beta}; I) + \lambda \mathcal{L}_{\text{OU}}(\boldsymbol{\beta})$ is the maximum-a-posteriori estimate of $\boldsymbol{\beta}$ under the* tempered posterior

$$p_\lambda(\boldsymbol{\beta} \mid I) \;\propto\; p_{\text{img}}(I \mid \boldsymbol{\beta})\; p(\boldsymbol{\beta} \mid \alpha, \sigma, \mu, d_{\text{tax}})^{\lambda}, \tag{S3}$$

i.e., the temperature $1/\lambda$ on the prior. In particular, λ=1 recovers the standard Bayes posterior; λ<1 down-weights the OU prior (more data-driven), λ>1 sharpens it.

Proof. Take the negative log of Eq. (S3). By Eq. (S1) the data term contributes $\mathcal{L}_{\text{base}} + C_I$; by Eqs. (2) and (S2) the prior contributes $\lambda(\mathcal{L}_{\text{OU}}(\boldsymbol{\beta}) + C_\Sigma)$ with $C_\Sigma = \frac{DB}{2} \log 2\pi$ (constant in $\boldsymbol{\beta}$). Argmin in $\boldsymbol{\beta}$ thus equals $\arg\min_{\boldsymbol{\beta}} \mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{OU}}$, which is the total-loss minimizer. The constants C_I, C_Σ shift the loss without affecting the argmin. □

Remark 1. *Because (α, σ, μ) are themselves learned by SGD on $\mathcal{L}_{\text{total}}$, the above is a Type-II / empirical-Bayes joint MAP in $(\boldsymbol{\beta}, \alpha, \sigma, \mu)$ rather than fixed-prior MAP; treating $\mathcal{L}_{\text{con}}$ as part of $\mathcal{L}_{\text{base}}$ gives a hierarchical model with two priors (binary-family + continuous-OU), which our best setting reduces to OU-only by w_{con}=0.*

S2.2 Per-pair structural richness compared to SupCon

What does $\mathcal{L}_{\text{OU}}$ see per pair of training species that $\mathcal{L}_{\text{con}}$ [7] does not? SupCon collapses every pair to a binary same-family / different-family indicator. $\mathcal{L}_{\text{OU}}$ ingests the full Linnaean rank distance, which on our 17-species matrix takes 5 distinct values. We describe this richer per-pair signal and show that SupCon arises as a degenerate limit of $\mathcal{L}_{\text{OU}}$ on a binary distance.

Per-pair signal. $\mathcal{L}_{\text{con}}$ uses a binary same-family indicator $y_{ij} = \mathbf{1}[f_i{=}f_j] \in \{0, 1\}$: every pair is collapsed to "same family" or "different family". By contrast, $\mathcal{L}_{\text{OU}}$ ingests the full Linnaean rank-distance matrix, whose off-diagonal in our 17-species set takes values in $\{1, 2, 4, 5, 6\}$ (Table S5). Each pair therefore carries a 5-level ordinal rather than a binary indicator, distinguishing within-family, same-order cross-family, and three grades of cross-order separation that the SupCon [7] indicator lumps together.

Limit interpretation. $\mathcal{L}_{\text{con}}$ corresponds structurally to the limit of an OU prior with a degenerate two-level taxonomy $d_{\text{tax}}(i,j) \in \{0, \infty\}$, in which cross-family covariances vanish and within-family blocks are rank-1. In this limit the per-pair ordinal collapses to a binary indicator. Appendix S2.3 makes the regularized statement precise.

S2.3 Limit analysis revisited

The main text (Section 3.4) writes the $\alpha \to \infty$ link to $\mathcal{L}_{\text{con}}$ [7] informally; that informal version glosses over a singular-covariance subtlety. Here we restate the limit with a positive-definite jitter εI that keeps the matrix invertible, then show the result is structurally analogous to (but not literally equal to) SupCon's same-family pull. "Structurally analogous" is the precise word: the two penalties act on the same direction (within-family residuals) but live in different feature spaces.

Proposition 2 (Regularized $\alpha \to \infty$ limit)**.** *Let $\Sigma_\varepsilon = \Sigma + \varepsilon I$ for ε>0, and hold $\sigma^2/(2\alpha) \equiv c$>0 constant as $\alpha\to\infty$. Then for the block-taxonomy $d_{\text{tax}}(i,j) = 0$ if $f_i{=}f_j$ and $d_{\text{tax}}(i,j) \geq 1$ otherwise:*

$$\lim_{\alpha\to\infty} \frac{1}{2}(\boldsymbol{\beta}_{\cdot,d} - \mu)^\top \Sigma_\varepsilon^{-1} (\boldsymbol{\beta}_{\cdot,d} - \mu) \;=\; \sum_f \frac{1}{2} \sum_{i: f_i = f} \frac{(\beta_{i,d} - \bar{\beta}_{f,d})^2}{\varepsilon} \;+\; O\Big(\frac{1}{c}\Big), \tag{S4}$$

where $\bar{\beta}_{f,d} = \frac{1}{n_f}\sum_{i:f_i=f}\beta_{i,d}$ is the empirical family mean. As $\varepsilon \to 0$ along a positive-definite sequence (e.g. $\varepsilon = 10^{-6}$, our implementation), the quadratic form acts as a per-family shrinkage to the family mean with infinite stiffness on the same-family direction and finite stiffness on the cross-family directions.

Proof sketch. Under the block-taxonomy, Σ becomes block-diagonal with within-family blocks $c\,J_{n_f}$ ($J = \mathbf{1}\mathbf{1}^\top$, rank 1) and zero off-block entries (since $\exp(-\alpha d_{\text{tax}}(i,j)) \to \mathbf{1}[f_i{=}f_j]$). Adding εI, the matrix-inversion lemma gives within-family inverse $\varepsilon^{-1}I_{n_f} - \frac{c}{\varepsilon(\varepsilon+cn_f)}J_{n_f}$. Decomposing $\beta_{i,d} - \mu_d$ into mean (along $\mathbf{1}_{n_f}$) and residual components: the mean sees stiffness $1/(\varepsilon{+}cn_f) \to 1/(cn_f)$, while the residual sees $1/\varepsilon$, diverging as $\varepsilon \to 0$. Cross-family pairs contribute zero by block-diagonality. Combining over D and dropping $O(1/c)$ for $c \gg \varepsilon$ yields Eq. (S4). □

Remark 2. *Equation (S4) is the per-pair Gaussian relaxation of the same-family pull encoded by SupCon: both shrink within-family residuals and leave cross-family directions unconstrained. The two are not literally equal, SupCon uses ℓ_2-normalized features in cosine space with a $\log\sum\exp$ objective, while Eq. (S4) is a Mahalanobis penalty in raw β-space, so we call them* structurally analogous. *Family-SupCon is recovered, in this regularized sense, as the degenerate $\alpha\to\infty$ limit of an OU prior on a binary $\{0,\infty\}$ taxonomy: a strict subset of what $\mathcal{L}_{\text{OU}}$ expresses on a 5-level rank distance.*

S3 Extended method details

This appendix has two jobs. First, we lay out, in one place, how our Gaussian prior $\mathcal{L}_{\text{OU}}$ relates to AniMer+'s family-contrastive loss $\mathcal{L}_{\text{con}}$, so the reader can see at a glance what extra structure the OU prior carries and where the two losses formally meet (Appendix S3.1). Second, we list the training and evaluation data, optimizer settings, and precision details needed to reproduce Table 1 from the public AniMer+ [1] checkpoint (Appendix S3.2).

S3.1 From $\mathcal{L}_{\text{con}}$ to $\mathcal{L}_{\text{OU}}$ and the $\alpha \to \infty$ limit

We write the two losses out side by side, because the difference between them is clearest at the $\alpha\to\infty$ limit.

The family supervised-contrastive term $\mathcal{L}_{\text{con}}$ used by AniMer+ operates on per-sample features z_i (the `cls_feats` head output in the public implementation) and family labels f_i:

$$\mathcal{L}_{\text{con}} = -\sum_i \frac{1}{|P(i)|}\sum_{p\in P(i)} \log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\neq i}\exp(z_i\cdot z_a/\tau)}, \quad P(i) = \{j : f_j = f_i, j\neq i\}. \tag{S5}$$

Its taxonomic input is a binary same-family indicator per pair (same family or not).

OU likelihood. The Ornstein–Uhlenbeck stochastic differential equation (SDE) [8, 10] along the Linnaean taxonomy produces a stationary multivariate Gaussian over tips with covariance in Eq. (1); treating the D shape dimensions as independent draws from a shared species covariance Σ yields $\mathcal{L}_{\text{OU}}$ as the negative log-likelihood (Eq. (2)), reported per-sample. The full MAP derivation is given in Proposition 1.

The $\alpha \to \infty$ degenerate limit. In the degenerate "block taxonomy" limit (within-family distance 0, cross-family distance ∞), the OU covariance Σ becomes block-diagonal and the resulting quadratic form pulls same-family shape codes together while leaving cross-family pairs uncoupled, the same qualitative effect as $\mathcal{L}_{\text{con}}$. The regularized statement of this limit, including the rank-deficient covariance treatment and the precise sense in which it is a *structural analogy / Gaussian probabilistic relaxation* rather than a strict equivalence with SupCon [7] / InfoNCE [38], is given in Proposition 2. The takeaway is that $\mathcal{L}_{\text{con}}$ is the structural analogue of the binary-indicator degenerate limit of $\mathcal{L}_{\text{OU}}$; finite $\alpha\sim 1$ exposes the additional graded cross-family structure that the discrete family indicator throws away.

S3.2 Data and evaluation

The point of this subsection is to make our experiments end-to-end reproducible: the reader should know which images we trained on, which images we tested on, and the optimizer knobs that fix the headline numbers.

Training data. CtrlAVES3D [29] (6 464 images, 17 CUB [32] species), the same AVES split AniMer+ uses in Table VI. Animal3D (mammals) and CUB pseudo-3D mat-file splits carry weight 0 in all our runs, because we lack redistributable copies and CtrlAVES3D alone suffices for the AVES branch.

Evaluation data. We report on three test sets, each probing a different aspect of the trained model. **CUB test** (official test split, 5 794 images) measures seen-species fine localization and follows the official AniMer+ [1] evaluation. The CUB training split (5 994 images) serves only as a development list for selecting λ_{OU} (Appendix S5.2; Figure S2 is scored on it); all headline and ablation numbers use the official test split. **CowBird test** [6] (114 evaluation instances) measures unseen-species generalization: CowBird species are not present in CtrlAVES3D [29] training, making this the hardest OOD benchmark of the three. The AniMer+ eval protocol draws those 114 instances from a 240-image test split, and our reported metrics use the same 114. Finally, **CtrlAVES3D test** (309 held-out images) measures seen-species 3D reconstruction. Metrics follow AniMer+ Table VI: per-dataset PCK-curve AUC and PCK@0.1.

Training schedule. Warm-start from the released AniMer+ checkpoint;[2] 30K steps, batch 4, AdamW [34], $\eta = 3\times10^{-7}$ on the main group. 30K steps is $\sim$2.5 hours on one RTX 4090.

Precision. Training runs entirely in fp32, including the ViT [22] forward/backward and the OU Cholesky, which avoids under-flow of $\log|\Sigma|$ at small α.

S4 The taxonomic distance matrix D_{tax}

The OU prior is parameterized by a fixed 17×17 matrix D_{tax} that encodes how related any two species in our training set are. This appendix exists for two reasons. First, a reader who wants to recreate our experiments needs to know exactly which species are in the matrix and what each integer entry means. Second, the choice of a discrete Linnaean rank over a continuous molecular phylogeny such as Jetz [30] needs a fuller justification than the main text gives. We address both in turn: Appendices S4.1 and S4.2 document the matrix; Appendix S4.3 explains the design choice.

S4.1 The 17 species and the matrix entries

Table S4 lists the species and their Linnaean order and family; the species set is fixed by CtrlAVES3D [29] and we did not curate it. The row index i of each species in D_{tax} matches the species ordering in Figure S1.

Figure S1 shows D_{tax} as a heatmap. Off-diagonal entries take only five distinct values $\{1, 2, 4, 5, 6\}$ (Table S5); each value has a clear taxonomic interpretation that we explain in Appendix S4.2.

S4.2 Construction rules

We assign integer rank distances to each species pair using a deterministic rule based on Linnaean order and family (Table S4). The rule is designed so that "closer in the tree" translates monotonically to "smaller integer".

- $d = 0$: same species (diagonal).
- $d = 1$: same family. The four within-family pairs (eight ordered entries) are Cardinalidae ($1\leftrightarrow2$), Corvidae ($3\leftrightarrow9$), Picidae ($4\leftrightarrow16$), and Alcedinidae ($10\leftrightarrow11$).

[2] https://github.com/AniMerPlus/AniMerPlus

Table S4: **The 17 CtrlAVES3D species in row-index order in D_{tax}.** Eight species belong to Passeriformes; the remaining nine span six non-Passeriformes orders. Species set: CtrlAVES3D [29]; binomial nomenclature, family, and order classification follow the IOC World Bird List [2].

i	Species (binomial)	Family	Order
0	*Phoebastria immutabilis*	Diomedeidae	Procellariiformes
1	*Passerina ciris*	Cardinalidae	Passeriformes
2	*Cardinalis cardinalis*	Cardinalidae	Passeriformes
3	*Corvus brachyrhynchos*	Corvidae	Passeriformes
4	*Colaptes auratus*	Picidae	Piciformes
5	*Tyrannus forficatus*	Tyrannidae	Passeriformes
6	*Quiscalus major*	Icteridae	Passeriformes
7	*Hesperiphona vespertina*	Fringillidae	Passeriformes
8	*Larus californicus*	Laridae	Charadriiformes
9	*Cyanocitta cristata*	Corvidae	Passeriformes
10	*Chloroceryle americana*	Alcedinidae	Coraciiformes
11	*Halcyon smyrnensis*	Alcedinidae	Coraciiformes
12	*Anas platyrhynchos*	Anatidae	Anseriformes
13	*Fratercula corniculata*	Alcidae	Charadriiformes
14	*Geococcyx californianus*	Cuculidae	Cuculiformes
15	*Bombycilla cedrorum*	Bombycillidae	Passeriformes
16	*Dryocopus pileatus*	Picidae	Piciformes

Table S5: **Off-diagonal distribution of D_{tax}.** Counts are over the $17 \times 16 = 272$ ordered off-diagonal pairs; an unordered species pair contributes 2 entries (one in each direction). Diagonal entries (same species, $d = 0$) are not included.

d_{tax}	Count	Pairs	Interpretation
1	8	4 unordered	Same family
2	54	27 unordered	Same order, different family
4	8	4 unordered	Sister-orders within Telluraves
5	170	85 unordered	Cross-order, non-sister
6	32	16 unordered	Outgroup (Procellariiformes vs. all)
total off-diag	272	136 unordered	

- $d = 2$: same order, different family. This rank captures the Passeriformes-internal pairs not already at rank 1 (e.g. $1 \leftrightarrow 3$, Cardinalidae–Corvidae) and the two Charadriiformes families (Laridae–Alcidae, $8 \leftrightarrow 13$).
- $d = 4$: sister-orders within the Telluraves clade. We treat Piciformes and Coraciiformes as sister-orders following the whole-genome avian phylogeny [11], which places both within Telluraves with strong support; the four unordered pairs are Picidae–Alcedinidae: $4 \leftrightarrow 10,\ 4 \leftrightarrow 11,\ 16 \leftrightarrow 10,\ 16 \leftrightarrow 11$.
- $d = 5$: cross-order, non-sister. All other cross-order pairs among Anseriformes, Charadriiformes, Cuculiformes, and the remaining Passeriformes / Piciformes / Coraciiformes pairs that are not already at rank 4. These orders do not have a closely-related sister-order within the 17-species subset, so we assign them a single rank rather than try to discriminate among them at this granularity.
- $d = 6$: outgroup. The oceanic Procellariiformes (*Phoebastria immutabilis*, $i = 0$) is the only pelagic seabird in the set and is assigned the deepest rank as a design choice, so its row and column take rank 6. This is an ordinal placement, not a claim about the root of the avian tree: in the whole-genome phylogeny [11] the deepest split among our species separates Anseriformes (Galloanseres) from the remaining Neoaves.

The cross-order rank values $\{4, 5, 6\}$ are ordinal ("sister-order $<$ generic cross-order $<$ deep outgroup") and the OU kernel's length-scale parameter α (Eq. (1)) absorbs the ordinal-to-cardinal conversion: the learned α converges to ~ 1.0 across our runs (Appendix S5.5), so cross-order similarity decays smoothly and monotonically with rank, and a uniform rescaling of d is equivalent (up to a redefinition of α) to leaving the ordinal values as written.

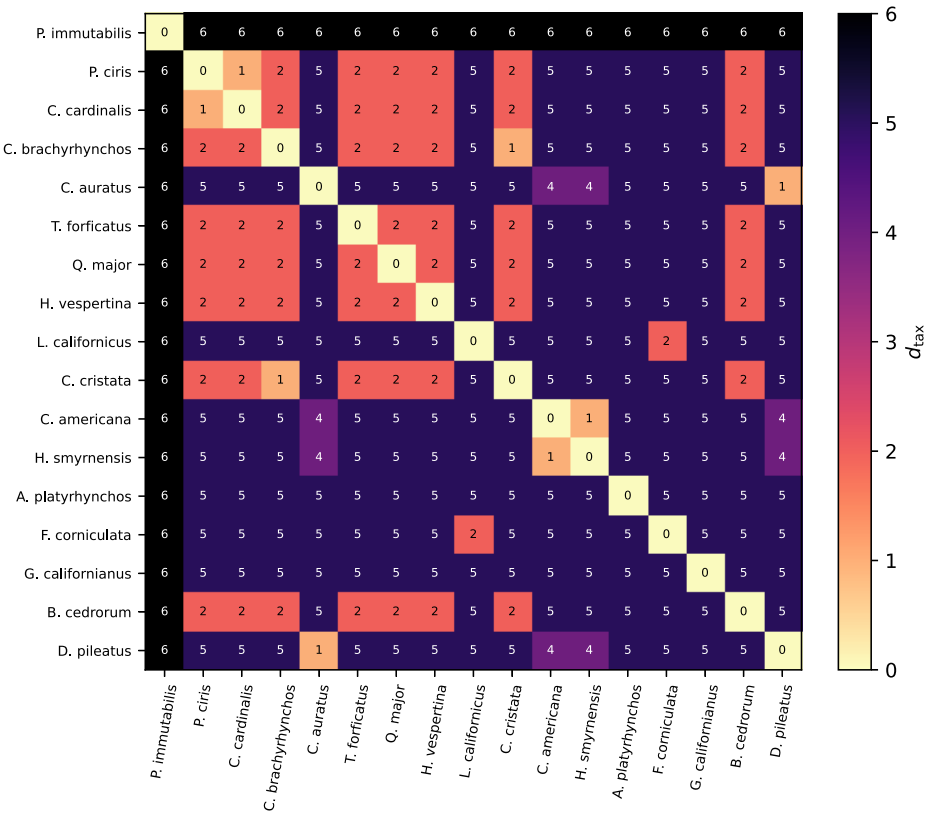

Figure S1: **The full 17×17 taxonomic rank distance matrix D_{tax} used by the OU [10] prior.** Diagonal entries are 0 (same species); off-diagonal entries take values $\{1, 2, 4, 5, 6\}$, representing same family, same order, sister-order, cross-order, and outgroup respectively. The species ordering on the axes matches Table S4; the dark first row/column reflect the single-species Procellariiformes outgroup (*Phoebastria immutabilis*, $i = 0$).

S4.3 Why discrete Linnaean rank rather than continuous Jetz distance

A natural alternative is to use a continuous molecular phylogeny such as the Jetz [30] patristic distance. We retain a Jetz patristic distance matrix for diagnostic use in Appendix S1: the posterior mean across 100 Hackett backbone trees, in millions of years, with range $[18.6, 201.2]$ My. We do not train against it directly. The reason is the kernel-conditioning argument of Section 3.1, which we restate here.

On the 17-species CtrlAVES3D subset, the rank-correlation between D_{tax} and the Jetz patristic matrix is high (Pearson 0.85, Spearman 0.76), so the discrete rank distance preserves most of the biological signal of a continuous molecular kernel. What it does not preserve is the dynamic range. Raw Jetz distances span $\sim$200 My, which would force the OU length-scale α near 10^{-2} to express any cross-family correlation; the discrete Linnaean ordinals $\{1, 2, 4, 5, 6\}$ instead lie on the same order as the converged $\alpha \sim 1$, yielding a well-conditioned Σ at the operating point the optimizer selects on its own. Discrete Linnaean rank therefore gives us most of the biological information of the continuous tree while keeping the kernel in the regime where the OU prior is doing real work.

S5 Extended experimental results

This appendix supports the headline numbers in Table 1 from several angles, so the reader can check that the gain is not specific to a single choice. We probe robustness across repeated runs, sensitivity to the OU regularization strength, the role of the OU-LR multiplier, the behavior of the real-vs-permuted contrast outside the training taxonomy, and the training dynamics of the OU scalars themselves.

Together these views are intended to show that the unseen-species improvement is stable across the design knobs we vary.

S5.1 Robustness across multiple runs

To check that the headline numbers in Table 1 do not depend on a single run, we repeat the headline operating point five times with different random seeds (the only thing that changes across runs) and report all five results in Table S6.

Table S6: **Per-trial results at the headline operating point.** Full hyperparameter setup in Appendix S9. Each row is one independent rerun (training seed in parentheses); Table 1 reports the Mean $\pm$ s.d. row. CUB is scored on the official test split.

Run	CUB [32] AUC	CUB [32] PCK@0.1	CowBird [6] AUC	CowBird [6] PCK@0.1	CtrlAVES3D [29] AUC
Trial 1 (seed 0)	89.31	57.91	60.91	19.26	93.04
Trial 2 (seed 7)	89.29	57.84	60.71	18.88	93.06
Trial 3 (seed 21)	89.31	57.94	60.88	18.59	93.06
Trial 4 (seed 42)	89.27	57.71	60.80	18.64	93.05
Trial 5 (seed 123)	89.30	57.82	60.70	18.92	93.05
Mean $\pm$ s.d.	89.30 ± 0.02	57.84 ± 0.09	60.80 ± 0.10	18.86 ± 0.27	93.05 ± 0.01
AniMer+ paper [1]	89.50	59.40	59.30	16.40	93.00

All five runs clear the AniMer+ paper baseline on the unseen-species benchmark. The five trials sit in tight bands (CowBird AUC 60.70–60.91, CowBird PCK 18.59–19.26, CtrlAVES3D AUC 93.04–93.06, CUB AUC 89.27–89.31, CUB PCK 57.71–57.94), and the CowBird and CtrlAVES3D bands lie above the corresponding AniMer+ Table VI numbers. The unseen-species gain is therefore not specific to a single run but repeats across independent reruns at the same operating point. Table 1 reports the mean $\pm$ sample standard deviation of these five runs; this table provides the underlying per-run values at the same operating point.

S5.2 Sensitivity to λ_{OU} and the seen-vs-unseen trade-off

We next examine how sensitive the headline numbers are to the strength of the OU regularizer, λ_{OU}, and how that strength trades seen-species against unseen-species accuracy.

Seen vs. unseen species. We call a species *seen* if the model was trained on it at either of its two training stages: AniMer+ [1] pre-training, which includes the CUB [32] training split and therefore all 200 CUB species, or our fine-tuning on CtrlAVES3D [29], which covers 17 of those species. Both seen-species benchmarks use held-out images: the CtrlAVES3D test split contains the 17 fine-tuning species, and the CUB test split contains all 200 CUB species, 183 of which are outside our fine-tuning taxonomy (analyzed separately in Appendix S5.3). A species is *unseen* if it appears at neither stage; the unseen-species benchmark is CowBird [6].

Figure S2 traces how the proposed prior trades seen against unseen accuracy as λ_{OU} varies. The seen-species CUB-PCK regresses smoothly as λ_{OU} grows — the cost of over-smoothing fine localization on species the model has already seen — while the unseen-species CowBird-PCK rises and peaks at intermediate values, the unseen-species benefit. On the development list of Appendix S3, the smallest λ_{OU} that still produces a measurable taxonomic effect (family k-NN purity $0.61 \rightarrow 0.68$, Appendix S1) is $\lambda_{\mathrm{OU}} = 10^{-5}$, which we adopt as the headline operating point.

S5.3 Real vs. permuted outside the training taxonomy

CowBird covers a single unseen species. We therefore also compare the real- and permuted-D_{tax} checkpoints of Figure 3 on the 183 CUB [32] species absent from the 17-species fine-tuning taxonomy (official test split, 5,289 images; Table S7). Because images of one species are correlated, the confidence intervals resample species rather than images (10,000 draws), and we also report the fraction of species on which the real matrix is better. Over all 183 species the real matrix is ahead on both metrics with species-clustered intervals excluding zero; the two order groups show similar point estimates, with the smaller non-Passeriformes subset not reaching significance on PCK.

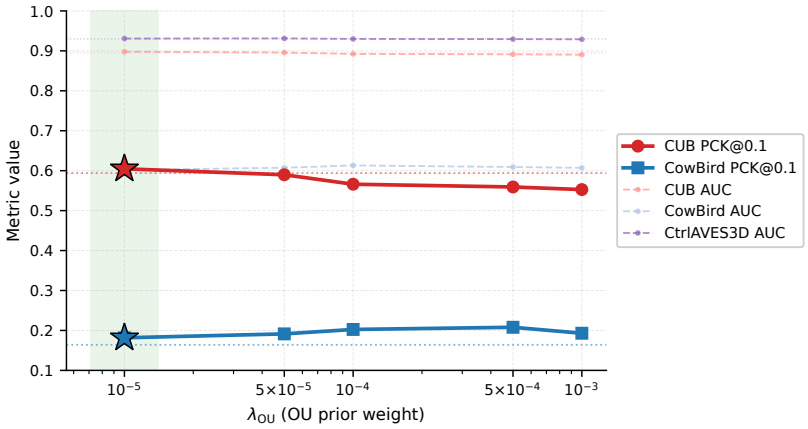

Figure S2: **Evaluation metrics across the swept OU [10] regularization strength λ_{OU}.** Solid lines are PCK@0.1 on the seen-species CUB [32] and the unseen-species CowBird [6]; dashed lines are AUC. Dotted lines are the AniMer+ paper values; the shaded region and stars mark $\lambda_{\mathrm{OU}} = 10^{-5}$, the value we adopt. All points are constant-LR runs. The sweep is scored on the development list of Appendix S3. The figure illustrates the seen-vs-unseen trade-off: seen-species accuracy regresses smoothly with λ_{OU} while unseen-species accuracy peaks at intermediate values.

Table S7: **Real vs. permuted D_{tax} on the 183 CUB species outside the fine-tuning taxonomy** (official test split; same checkpoints as Figure 3). Δ = real $-$ permuted in percentage points; 95% intervals from a species-clustered bootstrap; "real $>$ perm." counts species whose mean is strictly higher under the real matrix.

Group	Images	Species	Δ PCK@0.1	95% CI	real $>$ perm.	Δ AUC	95% CI	real $>$ perm.
All 183 species	5,289	183	$+0.28$	$[+0.10, +0.46]$	98/183	$+0.05$	$[+0.04, +0.07]$	128/183
Passeriformes	3,679	125	$+0.29$	$[+0.09, +0.49]$	66/125	$+0.05$	$[+0.04, +0.07]$	90/125
Non-Passeriformes	1,610	58	$+0.27$	$[-0.10, +0.62]$	32/58	$+0.06$	$[+0.01, +0.10]$	38/58

S5.4 Full OU-LR multiplier ablation

The OU-LR multiplier k sets the learning rate of a separate AdamW parameter group holding α, σ and μ to k times the backbone rate, as in discriminative fine-tuning [39]. We sweep k to separate the effect of the prior's shape at initialization from that of learning α and σ, and find that evaluation is nearly unchanged for $k \in \{1, 10, 30\}$. The prior therefore acts through its shape, not through fitted scalars.

Table S8: **Effect of OU-LR multiplier k on eval at $\lambda_{\mathrm{OU}} = 10^{-4}$.**

k	CowBird [6] PCK@0.1	α_{end}	σ_{end}	α, σ moving?
1	18.99	1.001	1.000	No (frozen at init)
10	18.93	1.005	0.996	Barely ($\leq 0.5\%$)
30	19.09	1.013	0.988	Slightly ($\sim$1%)
100	20.25	1.038	0.963	Yes ($\sim$3.7%)

The key observation: eval is nearly k-invariant at $\lambda = 10^{-4}$. CowBird PCK differs by ≤ 0.16 pp across $k \in \{1, 10, 30\}$, despite α moving from frozen at $k = 1$ to $\sim$1% at $k = 30$. Only at $k = 100$ does eval meaningfully shift ($+1$ pp CowBird PCK).

Interpretation. The OU prior's eval contribution at fixed λ_{OU} comes primarily from the *regularization shape* imposed by Σ at initial values $(\alpha_0, \sigma_0) = (1, 1)$, *not* from learned updates. The data-fit updates at $k = 100$ produce a secondary effect on CowBird PCK. The headline operating point uses $k = 100$ to keep the theoretical interpretation testable, but the eval gain is approximately

whatever the OU regularization shape alone would give at the frozen initialization $(\alpha, \sigma) = (1, 1)$. This decoupling is independent evidence against "our result is an (α, σ)-fitting artifact."

$k = 300$ and beyond. At $\lambda_{\mathrm{OU}} = 5 \cdot 10^{-4}$, $k = 300$ and $k = 1000$ both train to 30K steps without divergence, and evaluation is unchanged from $k = 100$, with CowBird-PCK 19.57 for all three. At $k = 1000$ the scalars α and σ oscillate during training but stay finite. We use $k = 100$ throughout the main results for consistency.

S5.5 Training dynamics: OU scalar trajectories

We finally track the OU [8, 10] scalars (α, σ) during training. Figure S3 shows that the direction is the same in every run (α up, σ down) and that the magnitude scales smoothly with λ_{OU}, ranging from $\sim$0.5% at the headline $\lambda_{\mathrm{OU}} = 10^{-5}$ (cosine LR; $\sim$1.7% for the constant-LR run at the same λ_{OU}) to $\sim$32% at $\lambda_{\mathrm{OU}} = 10^{-3}$ (the $\lambda_{\mathrm{OU}} = 10^{-4}$ endpoints match the $(\alpha_{\mathrm{end}}, \sigma_{\mathrm{end}})$ columns of Table S8 at $k = 100$). Because the real and permuted taxonomies move the scalars alike, this motion is not evidence that the data prefer the tree.

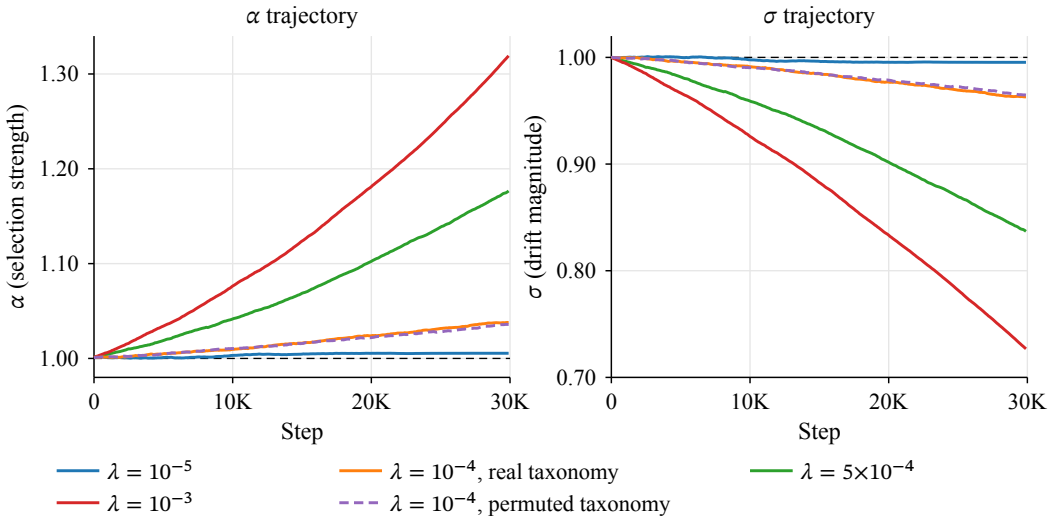

Figure S3: **OU [10] hyperparameter learning trajectories across λ_{OU} values and real vs. permuted taxonomies.** All runs train for 30K steps with $k = 100$. Both scalars start at 1, marked by the dashed line. α rises and σ falls in every run, and the change grows with λ_{OU}. The $\lambda_{\mathrm{OU}} = 10^{-5}$ run uses a cosine LR and the others a constant LR. The real and permuted taxonomy variants give nearly identical trajectories, showing that the motion of α and σ is driven by the OU kernel *shape* rather than tree *content*.

The real-d_{tax} and permuted-d_{tax} runs at the matched permutation-ablation operating point overlay closely throughout training (Figure S4), and for seed 0 end at close values of α (1.038 vs. 1.036) and σ (0.963 vs. 0.965). Scalar motion is therefore driven by the kernel *shape* (matrix spectrum), not by the biological semantics of the tree; the CowBird AUC/PCK gap reported in Section 4.3 is not explained by (α, σ) endpoint differences. The β-space calibration diagnostics likewise do not separate the two runs (Appendix S1), and their vertex-error gap is small (Table S9).

S6 Vertex-level errors on CtrlAVES3D

In this appendix we report 3D Procrustes-aligned [31] per-joint and per-vertex errors (pa-MPJPE, pa-MPVPE) on CtrlAVES3D [29] so the reader can see whether the 2D-keypoint gain in Table 1 shows up in the 3D mesh, and as a complementary axis of evidence on the real-vs-permuted comparison. The AniMer+ [1] paper reports these in its Table VI (AniMer+: pa-MPJPE 65.6, pa-MPVPE 70.9 on CtrlAVES3D), but our re-implemented evaluation uses the canonical AVES template units (roughly $10\times$ that scale), so we benchmark our reimplementation against itself across variants; the use is relative.

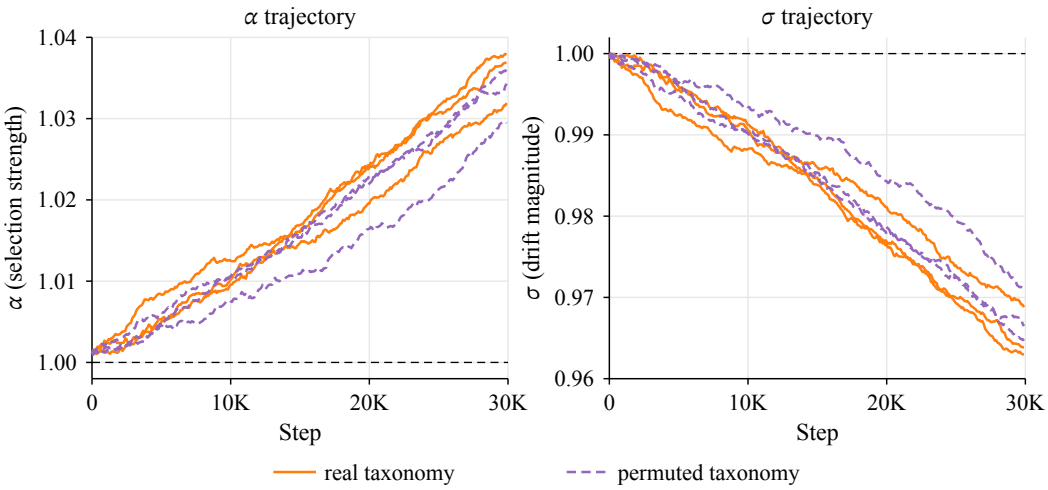

Figure S4: α **and** σ **over 30K steps for the real and permuted** D_{tax} **at** $\lambda_{\text{OU}} = 10^{-4}$**.** All runs use a constant LR, and each taxonomy has three runs with seeds 0, 42 and 123. Both scalars start at 1, marked by the dashed line. The real and permuted curves stay close throughout training, and the gap between them is on average smaller than the spread across seeds. Section 4.3 gives the complementary final-eval evidence.

Table S9: **Procrustes-aligned [31] 3D errors on the CtrlAVES3D [29] test split (309 images, 17 species).** pa-MPJPE = mean per-joint position error after Procrustes alignment; pa-MPVPE = mean per-vertex position error. Both in canonical AVES template units (see Caveat). Ours is the seed-0 run (Trial 1 of Table S6).

Variant	pa-MPJPE [31] ↓	pa-MPVPE ↓
AniMer+ [1] pretrained checkpoint	655.61	708.97
AniMer+ [1] + **OU (ours)**	**649.89**	**699.01**
λ=10^{-4} + real d_{tax}	650.91	697.98
λ=10^{-4} + perm d_{tax}	652.15	699.31
λ=$5\cdot10^{-4}$	653.74	704.56
λ=10^{-3}	651.20	701.56

Reading. Our method improves over the AniMer+ pretrained checkpoint by Δpa-MPJPE=-5.72 and Δpa-MPVPE=-9.96 (canonical units). The vertex-error scale shows the same direction as the CowBird-PCK ablation in Figure 3: real d_{tax} (pa-MPVPE=697.98) is slightly better than the permuted control (699.31, Δ=-1.33). The vertex-error gap is small relative to the $\sim$700-unit scale, consistent with the OU prior's mechanism being more visible on sparse 2D keypoint metrics than on densely-supervised 3D vertex positions. The (α, σ) trajectory diagnostic (Figure S4) and the phylogenetic-signal diagnostic (Table S2) do not separate the two variants, so this small vertex-error difference is the only one of the three diagnostics that favors the real matrix.

Caveat. These numbers come from our re-implementation of the eval pipeline, not from the official AniMer+ Table VI, whose CtrlAVES3D values (65.6 / 70.9) and the AniMer [20] Animal3D pa-MPJPE (80.4 mm) are on a different unit scale from the canonical AVES template units used here. Table S9 should therefore be read as *relative* comparisons among our variants on the same eval pipeline, not as absolute numbers comparable to AniMer's mammal table.

S7 CowBird OOD per-image error distribution

The headline CowBird [6] improvement in Table 1 (+2.46 pp PCK over AniMer+ [1]) is a *bulk* statistic. In this appendix we check from a per-image view that it is broadly distributed across the 114 test images, not driven by a few outliers.

Setup. We run the AniMer+ retrain @30K baseline (no OU) and our headline checkpoint on all n=114 CowBird test images, computing per-image keypoint L_2 errors on the visible-mask subset. Table S11 instead compares against the released AniMer+ checkpoint.

Where the gain concentrates: small birds. Figure S5 shows eight representative cases with the largest reduction in per-image mean keypoint error from baseline to ours, among images with at least five visible keypoints on which ours ends below 0.2 (normalized crop units). The selected cases span varied poses and backgrounds, with no obvious pose bias. Splitting the 114 test images into terciles by relative bbox area (Table S10 and Figure S6) sharpens the picture: the per-image improvement is largest on small birds, $+0.038$, smaller on medium birds, $+0.021$, and slightly negative on large birds, -0.007. The interpretation aligns with the prior's design intent: when image-level cues are weak (small birds, fewer pixels covering the body), the species-level shape prior carries proportionally more useful information; when the bird fills the frame, the per-image regression head of AniMer+ already has abundant signal and species-level priors matter less.

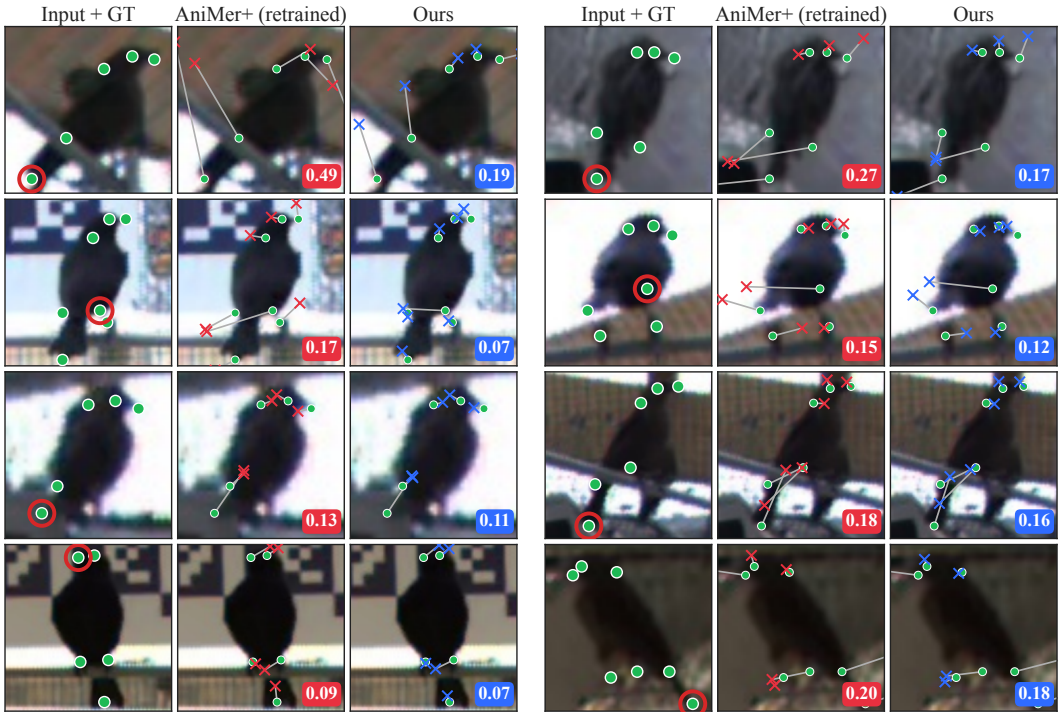

Figure S5: **Eight CowBird [6] test cases with the largest per-image mean-error reduction under the OU [10] fine-tune.** The convention follows Figure 5, with a first panel showing the input and the ground-truth keypoints in green. The retrained AniMer+ [1] baseline has red crosses and ours has blue crosses. Gray lines join each prediction to its ground-truth keypoint, and the red circle marks the keypoint the baseline misplaces most. The badge is the per-image mean keypoint error over the visible keypoints in normalized crop units, so lower is better. Best viewed in electronic format.

Table S10: **CowBird [6] per-image Δ max-err (baseline $-$ ours) split by bbox area tercile.** Small birds (small image footprint) receive the largest OU [10] benefit, medium birds a smaller one, and large birds a slightly negative change on average. Max-err is the largest keypoint error in an image.

Bbox-area tercile	n	Mean baseline max-err	Mean ours max-err	Δ (improvement rate)
Small (lowest tercile)	38	0.435	0.396	$+\mathbf{0.038}$ (57.9% improved)
Medium (mid tercile)	38	0.503	0.482	$+0.021$ (52.6% improved)
Large (highest tercile)	38	0.620	0.628	-0.007 (55.3% improved)

Caveats. (i) The improvement metric here is *relative* per-image error; the bulk CowBird-PCK gain comes partially from images that were just above the PCK threshold under baseline and just below

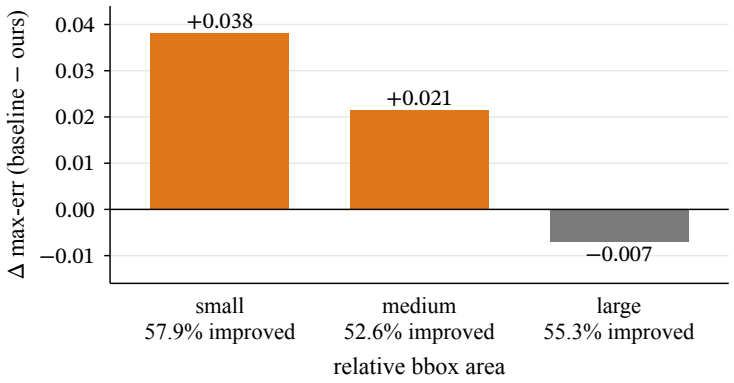

Figure S6: **Per-image Δ max-err on CowBird [6] by relative bbox area.** The 114 test images are split into three terciles of 38 images each. Δ is baseline minus ours, so a positive bar means ours is better, and the label under each bar gives the share of images improved. The improvement is largest on small birds at $+0.038$, smaller on medium birds at $+0.021$, and slightly negative on large birds at -0.007.

under ours, which the per-image breakdown does not directly highlight. (ii) Negative cases (baseline better) do exist and are not selected against in the top-improvement panel; we display only the largest ours-better gaps as illustrative examples.

Per-image significance. Table S11 reports a paired per-image comparison of the seed-0 headline run (Trial 1 of Table S6) against the released AniMer+ checkpoint on the 114 CowBird instances, with a 10,000-sample paired bootstrap for the 95% confidence interval and a two-sided Wilcoxon signed-rank test. All three intervals exclude zero, and AUC improves on 84 of the 114 images, so the gain is broadly distributed rather than driven by a few outliers. The per-seed dispersion of the same configuration is in Table S6.

Table S11: **Paired per-image comparison on CowBird [6]** (114 instances; seed-0 headline run vs. the released AniMer+ [1] checkpoint, every image weighted equally).

Metric	Mean Δ (pp)	95% CI	Wilcoxon p
AUC	$+1.52$	$[+0.83, +2.34]$	1.5×10^{-8}
PCK@0.1	$+3.21$	$[+1.35, +5.16]$	8.8×10^{-3}
PCK@0.15	$+1.99$	$[+0.46, +3.68]$	1.5×10^{-2}

S8 Extended qualitative analysis

This appendix shows further side-by-side cases in which the OU fine-tune visibly corrects the AniMer+ [1] baseline; the goal is to illustrate *what kind* of error reduction the headline numbers in Table 1 are made of.

Figure 5 in the main text shows two CUB [32] official-test images taken from those with the largest mean-error reduction among images with at least seven visible keypoints and our mean keypoint error ≤ 0.09. Figure S7 shows eight further cases that pass a fixed selection rule (baseline worst-keypoint error ≥ 0.12 in normalized units, our mean keypoint error ≤ 0.10, mean-error reduction ≥ 0.015) among the species outside the 17-species fine-tuning taxonomy. Across these eight cases the baseline's largest error falls on the head (crown, nape, eye, bill base) in five, on the torso (breast, tail base) in two and on a leg in one. The comparison is against the AniMer+ baseline retrained with its published recipe for the same 30,000 steps, so it reflects the full fine-tune rather than the OU prior in isolation.

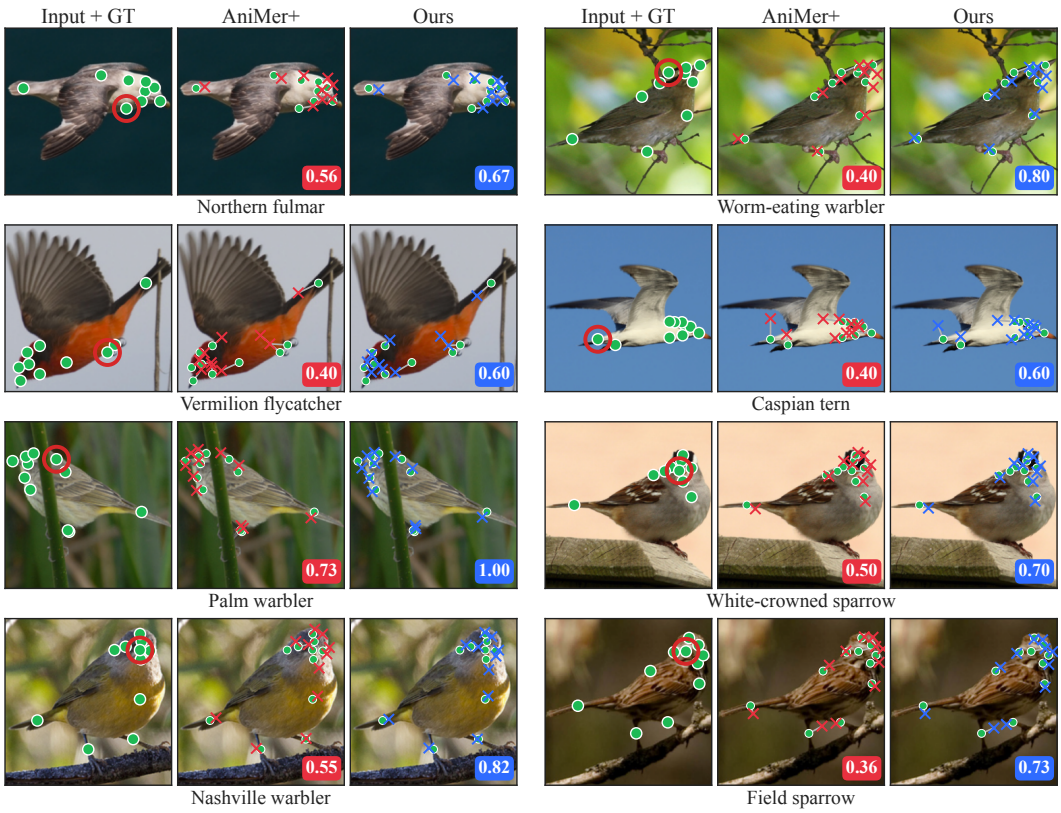

Figure S7: **Eight further CUB [32] test cases.** The convention follows Figure 5, with a first panel showing the input and the ground-truth keypoints in green. The retrained AniMer+ [1] baseline has red crosses and ours has blue crosses. The badge is the per-image PCK@0.1, so higher is better. Best viewed in electronic format.

Caveat. These are selected cases, not a statistical sample: the selection rule picks images on which the baseline fails and ours does not, so the panels illustrate the kind of correction the fine-tune makes rather than how often it makes one. Aggregate behavior is in Table 1 and, per image, in Appendix S7.

S9 Training implementation details

This appendix lists the optimizer, numerical, and initialization choices that determine our headline numbers in Table 1. The warm-start checkpoint, training-data composition, total step count, batch size, base learning rate, and fp32 precision are listed in Appendix S3; here we cover only the items that section does not.

Operating points. Headline results (Tables 1, 3 and S6 and the qualitative figures) use the cosine LR schedule, $\lambda_{\mathrm{OU}} = 10^{-5}$, $k = 100$, $w_{\mathrm{con}} = 0$, 30,000 steps, batch 4; Table 1 reports the mean $\pm$ sample standard deviation over five seeds of this configuration and Table S6 the per-seed values. The real-vs-permuted and rank-granularity ablations (Figure 3 and Tables 2 and S8) use the stronger-prior setting $\lambda_{\mathrm{OU}} = 10^{-4}$ with a constant LR, chosen so that the taxonomic content of D_{tax} is expressed strongly enough for the contrast to be visible. Both settings share the warm-start checkpoint, data, step budget and k; only λ_{OU} and the schedule differ. The remaining sweeps (Figure S2 and the multi-variant comparisons in Tables S1 to S3 and S9) state their settings in place.

S9.1 Optimizer and schedule

We use AdamW [34, 35] with two parameter groups. The base learning rate is $\eta = 3\times10^{-7}$ on the main group; the OU group (α, σ, μ) lives at $k\eta = 3\times10^{-5}$ with $k = 100$. The cosine schedule

uses $T_{\max} = 30{,}000$ steps and $\eta_{\min} = 0.01\eta = 3 \times 10^{-9}$. The λ_{OU} and k axes in Appendix S5 jointly identify the operating point above.

Choice of k. We use $k = 100$ so that α and σ are learned rather than left at their initial values. AdamW changes a parameter by roughly the learning rate per step, so at the backbone LR $\eta = 3{\times}10^{-7}$ the OU scalars can move by at most about 10^{-2} over 30K steps; with $k = 1$ they end within 0.1% of their initial values (Table S8) and the prior acts as a fixed kernel. Evaluation is nearly unchanged for $k \in \{1, 10, 30\}$ and shifts by about one CowBird-PCK point at $k = 100$ (Table S8).

S9.2 Numerical safeguards and initialization

Training runs in fp32, which avoids the conditioning issues of fp16 on near-singular kernels: the OU covariance assembly, Cholesky decomposition $\Sigma = LL^{\top}$, and the Mahalanobis quadratic in Eq. (2) are all computed in fp32. We add a diagonal jitter $\varepsilon = 10^{-6}$ to Σ. Without it, two images of the same species in a mini-batch have $d_{\mathrm{tax}} = 0$, which makes Σ singular and the Cholesky fails. Gradient clipping is at norm 1.0 on the full parameter set. The OU group has weight decay 0: zero weight decay on the OU group prevents AdamW from driving α, σ back toward their softplus-init independently of the gradient signal, which would interfere with the theoretical interpretation.

The OU scalars (α, σ) use a softplus parameterization [13] with floor $\alpha_{\min} = 10^{-3}$. The underlying raw scalar is initialized to $\mathrm{inv_softplus}(1.0) \approx 0.541$ so that $\mathrm{softplus}(\mathrm{raw}) = 1.0$ at step zero; thus $\alpha_0 = \sigma_0 = 1.0$ in the physical (post-softplus) parameterization. The species-mean $\mu \in \mathbb{R}^{15}$ is zero-initialized. The $\mathcal{L}_{\mathrm{con}}$ [7] weight w_{con} used by AniMer+ is set to 0 in our headline runs, leaving the OU prior as the sole taxonomy-aware regularizer; setting $w_{\mathrm{con}} = 5{\times}10^{-4}$ recovers the original AniMer+ configuration.

S9.3 Loss-term integration and permutation control

To minimize drift from the released AniMer+ codebase we implement the OU prior as an additive override of AniMer+'s loss computation: the original task losses are computed unchanged, then $\mathcal{L}_{\mathrm{OU}}$ is added before the backward pass. This keeps the parameter ordering identical to the released checkpoint and guarantees that $\lambda_{\mathrm{OU}} = 0$ with w_{con} left at its original value recovers the original training loop to floating-point equality.

For the permutation ablation (Section 4.3) we permute the species labels on D_{tax} via a random permutation π applied to both row and column indices. This preserves the marginal pairwise distance distribution exactly, and therefore the OU-kernel spectrum, but destroys the biological correspondence between species labels and taxonomy positions. A single fixed permutation is used for every seed and every setting. The selected permutation breaks all four same-family cross-species relations and reassigns the same-order relations, so the informative close-kin entries are scrambled.

S9.4 Inference cost

The OU prior is a training-time loss only: at inference the network is structurally identical to AniMer+, so no extra parameters, layers, or kernels are introduced on the forward path. To make this concrete we measured wall-clock latency and peak GPU memory across three checkpoints on a single RTX 4090 at batch sizes $\{1, 4, 8\}$, with 3 warm-up iterations and 20 timed iterations per cell. The checkpoints are the released AniMer+, our paper-exact retrain at 30K steps, and an OU fine-tune at $\lambda_{\mathrm{OU}} = 10^{-5}$ with the constant-LR recipe, whose architecture is identical to the headline run. Table S12 shows the result.

Latency differences are ≤ 0.2 ms across variants at every batch size, within the run-to-run jitter: the per-cell p25–p75 spread is ≈ 0.2 ms in our measurements. Peak memory and parameter count are bit-identical by construction. The OU prior therefore carries no inference-time cost.

Training-time cost of the OU term. The OU covariance is assembled over the B instances of the mini-batch, not over all N species, so one OU step (distance gather, covariance build, Cholesky factorization and solve) at $B = 4$ takes 0.67–0.69 ms for $N = 17$, 200, 2,000 and 10,000 species (RTX 5090, mean of 100 iterations after 10 warm-up steps): the cost is set by B, not by N. The $N{\times}N$ distance matrix is a fixed lookup table that is never inverted.

Table S12: **Inference-time cost across the three checkpoints.** Measured on a single RTX 4090 with fp32 forward passes. Latency is the median over 20 timed iterations after 3 warm-up steps; peak memory and parameter counts are deterministic. The architecture is unchanged across the three rows, so parameter counts coincide exactly.

Variant	ms @ bs=1	ms @ bs=8	Peak mem @ bs=8 (MB)	Params (M)
AniMer+ [1] pretrained	19.0	74.0	3160.9	785.67
AniMer+ retrain @30K	18.9	73.9	3160.9	785.67
Ours	18.9	74.0	3160.9	785.67

S10 Extended discussion: limitations and future work

This appendix expands on the limitations and future directions summarized at the end of Section 5, and lays out the three follow-up directions we see as most natural: lifting the pipeline to multi-view 3D reconstruction, replacing the discrete Linnaean rank distance with a continuous molecular phylogeny, and porting the same OU machinery to mammals. Each direction adds exactly one ingredient to the present setup – multiple views, a continuous tree, or a non-avian dataset – while leaving the OU prior itself unchanged. We discuss each in turn, together with the obstacles a follow-up paper would need to address.

S10.1 Multi-view extension

When several synchronized cameras observe the same bird – the natural setting for full 3D reconstruction in the wild – the systems-side extension is a multi-view 3D Gaussian Splatting pipeline [40] for birds, using existing multi-view rigs such as 3D-POP [41] (4-view) with the OU-regularized AniMer+ [1] shape codes as template. The OU prior in this paper is shape-code-level and orthogonal to the multi-view lifting machinery, so the extension is a clean composition rather than a re-derivation. Posing it requires re-implementing the AniMer+ lifting head as a per-Gaussian regressor, which we leave as future work.

S10.2 Scaling the prior to a continuous molecular phylogeny

The Linnaean rank distance is biologically grounded and admits a well-conditioned OU kernel, but it is discrete by construction; replacing it with a continuous, time-calibrated tree would let the prior cover thousands more species. Section 3.1 argued on conditioning grounds that a naive substitution of the Jetz patristic matrix [30] would degenerate the kernel because the molecular distances span hundreds of millions of years and saturate the OU exponential at the selection-pressure regime our optimizer converges into. A length-scale-normalized continuous-distance taxonomy is the natural next step, and would scale the prior to the much larger species set covered by the Jetz tree. Two options are $\exp(-\alpha\, d/\bar{d})$ over a normalized molecular phylogeny, and a learned warping that brings the Jetz distance into the $O(1)$ regime where the converged selection pressure is meaningful. Our β-space calibration analysis (Appendix S1) provides preliminary empirical support for this direction.

Scaling the species set also changes the memory footprint of the taxonomy itself. The per-step OU cost is independent of N because Σ is assembled over the B instances in the mini-batch (Appendix S9), but the pairwise distance matrix D_{tax} is $O(N^2)$: at $N = 10{,}000$ species it occupies about 400 MB in fp32, which would call for compact integer or triangular storage, or on-demand taxonomy lookups, rather than a dense float matrix.

S10.3 Mammalian analogue

The OU kernel structure transfers to any taxonomy – including mammals, the other major taxonomic group already supported by the AniMer+ codebase – but the kernel parameters and the choice of distance measure would need to be re-fit for non-avian groups. The mammalian branch of the AniMer+ [1] training data is inactive in our experiments because we focused on the avian long-tail problem; a mammalian replication is a natural companion paper using the same OU machinery on a different taxonomic substrate.

NeurIPS Paper Checklist

This is the NeurIPS Paper Checklist for our submission. We answer each standard reproducibility question with an explicit Yes / No / NA, and the Justification entries cross-reference the relevant supplementary section where the underlying evidence lives. The list is intended to make the reproducibility surface of our paper auditable item-by-item.

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 contributions: the taxonomic OU prior (Section 3), the gain over AniMer+ on CowBird and CtrlAVES3D with seen-species CUB comparable (Section 4.2), the permutation ablation that isolates taxonomic structure (Section 4.3), and the rank-granularity ablation (Section 4.4).

2. **Limitations**

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

   Answer: [Yes]

   Justification: A dedicated Limitations paragraph in Section 5 covers the 17-species training taxonomy and its order imbalance, the single-species CowBird benchmark, untested genus-level structure, dependence on one taxonomic authority and its revisions, the bird-only scope and what other animal groups would require, and the untested behavior at hundreds or thousands of species; Appendix S10 expands on the directions these suggest.

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: The paper's theoretical contribution is the structural analogy between $\mathcal{L}_{\mathrm{OU}}$ and $\mathcal{L}_{\mathrm{con}}$ in the $\alpha \to \infty$ block-taxonomy degenerate case. Section 3.4 states the analogy with explicit caveats (rank-deficient covariance at the limit, requiring pseudo-inverse with jitter $\varepsilon \to 0$; L2-normalized features in SupCon [7] vs. raw $\boldsymbol{\beta}$ in $\mathcal{L}_{\mathrm{OU}}$, the two losses are structurally analogous, not literally equal); Appendix S2 provides the full sketch with assumptions (block taxonomy, shared across-trait covariance, $\sigma^2/(2\alpha)$ held constant in the limit, regularization order).

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: Section 4.1, Appendix S3, and Appendix S9 specify warm-start checkpoint, 30K steps at batch 4, AdamW [34] $\eta = 3 \times 10^{-7}$, fp32 training, Cholesky jitter 10^{-6}, softplus parameterization with $\alpha_{\min} = 10^{-3}$. The OU patch is an additive override of AniMer+'s training loss that reduces to vanilla AniMer+ at $\lambda = 0$ with w_{con} left at its original value.

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?

   Answer: [No]

   Justification: Our code is not released with this version. The training data (CtrlAVES3D [29]), the evaluation data (CUB [32], CowBird [6]) and the Linnaean rank matrix (derived per Appendix S4.2) are all public. The AniMer+ checkpoint we warm-start from is released by the AniMer+ authors.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details?

Answer: [Yes]

Justification: See Section 4.1 for the main-body specification and Appendices S3 and S9 for the remaining details (parameter groups, weight decay, the fixed permutation, loss-override implementation).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined?

   Answer: [Yes]

   Justification: Table 1 reports the mean $\pm$ sample standard deviation over five seeds of the headline operating point and Table S6 the per-trial values; the CowBird and CtrlAVES3D improvement direction is robust across these reruns. Table S11 reports a paired per-image bootstrap interval and Wilcoxon test on CowBird, and Table S7 reports species-clustered bootstrap intervals for the real-vs-permuted comparison.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources?

   Answer: [Yes]

   Justification: All fine-tunes run on a single RTX 4090 (24 GB) at batch 4; 30K steps takes $\sim$2.5 hours. Total compute for all variants reported is $\sim$75 GPU-hours across $>$25 runs.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The work uses only public image datasets and public taxonomic data; no human subjects, no crowdsourcing.

10. **Broader impacts**

    Question: Does the paper discuss potential positive and negative societal impacts?

    Answer: [Yes]

    Justification: Positive impacts include automated quantification of wild-bird posture / flight from video for ecological and behavioral research. We see no direct negative impact from an avian shape estimator; it is a narrow scientific tool not directly suited to surveillance or identification of humans.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models with high misuse risk?

    Answer: [N/A]

    Justification: The model is a bird-specific 3D shape estimator; we do not see plausible misuse scenarios requiring specific safeguards.

12. **Licenses for existing assets**

    Question: Are the creators of assets properly credited and are licenses respected?

    Answer: [Yes]

    Justification: We cite AniMer+ [1], AVES [12], SMAL [3], CUB [32], CowBird [6], CtrlAVES3D [29], the IOC World Bird List [2] from which the taxonomic distance matrix is derived, and the Jetz [30] phylogeny used for diagnostics. All are used in accordance with their respective licenses.

13. **New assets**

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

    Answer: [Yes]

Justification: The OU patch and taxonomic distance matrix are documented in Appendices S4 and S9.

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

    Question: For crowdsourcing experiments and research with human subjects, does the paper include full instructions etc.?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent**

    Question: Does the paper describe potential risks to participants and IRB approval?

    Answer: [N/A]

    Justification: No human subjects research.

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?

    Answer: [N/A]

    Justification: LLMs are not used in the core methodology or experiments.