

Vert-NSM: neural shape models for shape completion, classification, and landmark-free morphometrics of vertebrate specimens

Wolcott KA¹, Cutura A³, Delage M³, Fenner KL¹, Khan AS³, Koh JX³, Riegler M¹, Sawyer C³, Yu J³, Gatti AA², Stanley EL^{*1} and A Porto^{*1}

¹ Florida Museum of Natural History, University of Florida, Gainesville, FL, USA

² Department of Radiology, Stanford University, Stanford, CA, USA

³ Human-Augmented Analytics Group, Georgia Institute of Technology, Atlanta, GA, USA

Corresponding author: Katherine Wolcott (k.wolcott@ufl.edu)

* ELS and AP contributed equally to this work and are joint-last authors.

Abstract

Natural history museums are building digital 3D archives at a scale that enables geometric morphometric analysis across thousands of specimens. Landmark-based analyses depend on anatomical priors and often exclude unidentified and incomplete specimens, which are abundant in collections and the fossil record. Continuous shape representations offer a landmark-free alternative that can be combined with neural networks to perform shape completion and classification. It is unknown how their morphospaces compare with landmarks or how accurately they perform. We present Vert-NSM, neural shape models for vertebrate biodiversity specimens, and apply them to squamate vertebrae, a well-represented, but challenging test case. Models were trained on 2,896 vertebral meshes from 95 specimens across 40 families, produced in part by the oVert digitization initiative. We evaluated shape completion, classification, and morphometrics of sparse landmarks, dense correspondences, and Vert-NSM latents. Shape completion accuracy reached 0.375% of a vertebral radius. Classification exceeded frequency-weighted chance at nearly every level tested, reaching 80% accuracy for the cervical spine and 90% for Cordylidae. Vert-NSM shared the dominant axes of variation recovered by landmarks, and identified shape differences that did not depend on anatomical priors. In downsampling experiments, models performed well when trained with at least 30 specimens. Vert-NSM has the potential to unlock discoveries from vertebrate collections at scale and in deep time.

Keywords

deep learning; geometric morphometrics (GMM); latent space; X-ray micro-computed tomography (micro-CT); natural history museum; signed distance function (SDF); Squamata; vertebra

1. INTRODUCTION

Natural history museums are increasingly using X-ray micro-computed tomography (micro-CT) to create digital three-dimensional (3D) archives of biodiversity (Blackburn et al. 2024; Hedrick et al. 2020; Lendemer et al. 2020; Nelson and Ellis 2019; Keklikoglou et al. 2019). As archives like MorphoSource (www.morphosource.org) grow, they enable 3D geometric morphometric (GMM) analysis at phylogenetic and geographic scales that would have previously required decades of collection, preservation, and imaging (see for e.g., Gray et al. 2025; Gignac et al. 2024; reviewed in Borowiec et al. 2022). When repurposing archival data, three challenges to broad-scale analyses of vertebrate specimens remain. First, annotation (segmentation and landmarking) is time-consuming and requires taxonomic and anatomical expertise. Second, isolated and fragmentary specimens (e.g., fossils) are challenging to identify. Third, landmarking typically requires intact, identified specimens.

Segmentation software now ranges from fully automated to human-in-the-loop approaches to produce 3D meshes (Diaz-Pinto et al. 2024; Ma et al. 2024; Rolfe et al. 2023; Isensee et al. 2021; Berg et al. 2019). Landmarking tools transfer points from an annotated template to thousands of specimens, with statistical shape models (SSMs) proving especially valuable for morphologically divergent specimens (Porto et al. 2026; Rolfe and Maga 2023; Zhang et al. 2022a; Porto et al. 2021; Boyer et al. 2015). As these approaches continue to improve for museum specimens, the accumulation of analysis-ready meshes and landmarks raises the question of what to do with them at scale.

Landmarked meshes are typically analyzed in morphometric toolkits, such as R geomorph (Collyer and Adams 2024; Adams and Otárola-Castillo 2013; reviewed in Mitteroecker and Schaefer 2022), where generalized Procrustes analysis aligns and scales the data, principal component analysis summarizes morphospace trends, and linear models test for disparity and group effects (see for e.g., Tharakan et al. 2023; Watanabe et al. 2019). Such analyses hinge on condensing an entire 3D surface into a small set of discrete, homologous points that capture variation relevant to a hypothesis (Mitteroecker and Schaefer 2022; Watanabe 2018). If chosen poorly, landmarks introduce noise or miss interesting morphology, especially across diverse taxa, where derived traits are present in some lineages and absent in others. Shape representations that do not depend on discrete, homologous points remove the challenges of landmark selection, and enable analysis of fragmentary specimens.

Landmark-free alternatives have been proposed for over fifteen years, some quantifying shape from dense anatomical correspondences and others comparing entire shapes on a manifold (Porto et al. 2026; Mulqueeney et al. 2024; Rolfe and Maga 2023; Porto et al. 2021; Toussaint et al. 2021; Schlager 2017; Pomidor et al. 2016; Koehl and Hass 2015; Boyer et al. 2011). It remains untested whether such methods recover the same patterns as landmarks at scale. A signed distance function (SDF) describes a surface implicitly, assigning negative, positive, and zero values to points sampled inside, outside, and on the boundary (Osher and Fedkiw 2003). DeepSDF learns to approximate SDFs for novel shapes, encodes shapes compactly, and reconstructs meshes at almost infinite resolution (Park et al. 2019). Gatti and colleagues extended this SDF-based approach to biomedical data with Neural Shape Models (NSM), which pairs query points with features from three orthogonal planes rather than a voxel grid, giving the decoder local context at a fraction of the cost (Gatti et al. 2025; 64^3 voxel grid: 262,144 cells; 3×64^2 planes: 12,288 cells).

While landmark-based methods estimate missing structures through symmetry or reference-based transforms (Dumoncel et al. 2016; Arbour and Brown 2014; Mitteroecker and Gunz 2009), models like NSM use a learned shape prior to complete partial surfaces constrained by variation in the training data. In addition to handling partial specimens, similar shapes also sit near each other in the latent space, so an encoded shape can be encoded and matched against references to produce a candidate shortlist for expert review.

In the present study, we extend NSM (Gatti et al. 2025) for use with vertebrate biodiversity specimens (hereafter Vert-NSM, NSM for Vertebrates) and apply it to squamate vertebrae, a challenging test case that is well-represented in museum collections and the fossil record. Squamate vertebrae preserve variation related to spinal position and taxonomy (Hoffstetter and Gasc 1969; Etheridge 1967), yet many specimens remain unidentified since diagnostic characters are concentrated in the skull and jaw (Tschopp 2016). Our dataset contains roughly 3,000 vertebrae from 40 families, produced in part by the oVert digitization initiative (Blackburn et al. 2024). We report our implementation of Vert-NSM (encoding strategies, auxiliary losses, and minimum training dataset size) and evaluate trained models for three tasks relevant to museum collections: completing partial specimens, classifying unknown specimens, and testing whether latent morphospaces recover the patterns of landmark-based geometric morphometrics.

2. MATERIALS AND METHODS

2.1. Code and data availability

Vert-NSM is open source, written in Python 3.10, and available on GitHub (<https://github.com/3D-fossils-Haag/vert-nsm>). The README and wiki contain tutorials and the paper directory documents all analysis, table, figure, and interactive rendering scripts. Interactive plots, 3D renderings, and videos are available (<https://3d-fossils-haag.github.io/vert-nsm-figs/>; hereafter, Vert-NSM figures site). Micro-CT data were produced by oVert and are on MorphoSource (Blackburn et al. 2024; **Table S1**; <https://www.morphosource.org/projects/000368762>). Training data (https://huggingface.co/datasets/BioVisionLab/train_data), trained models, config files, training logs, and evaluation outputs (<https://huggingface.co/BioVisionLab/models>) are on Hugging Face. **Figures 1–4** were arranged using Inkscape 1.4.4. Hardware, preprocessing settings, loss formulations, and evaluation metrics are detailed in the supplementary materials and methods (**File S1**).

2.2. Training data

To capture the phylogenetic and ecological diversity of lizards, we selected 95 specimens representing 33 of 37 lizard families, all six amphisbaenian families, and a tuatara as a squamate outgroup (Pyron et al. 2013; **File S2**). Specimens span five life history strategies: arboreal, fossorial, grass-swimmer, saxicolous, and terrestrial (following groups in Meiri et al. 2023, with grass-swimmer added to distinguish between limb-reduced groups, as in Camaiti et al. 2021). Vertebrae were segmented and saved as PLY meshes in VGStudioMax v.2026.2 (Volume Graphics, Heidelberg, Germany). The final dataset size was 2,896 meshes and includes all presacral vertebrae except the atlas and axis. Sparse landmarks (N=28) were

applied to meshes following Bourque and Stanley (2025). **Figure 1** shows landmarked vertebrae for life history groups and the mean mesh.

Meshes and landmark files were imported into 3D Slicer 5.11 (Kikinis et al. 2014). MorphoWeave was used to build a statistical shape model (SSM), generate dense anatomical correspondences (N=4943), a mean mesh, and mean sparse landmarks (N=28; **Figure 1**), and to align and scale all meshes (Porto et al. 2026). Surfaces were denoised using the Surface Toolbox module (**File S1**). Detailed preprocessing steps are on the Vert-NSM wiki. Dense correspondences and mean landmarks are rendered interactively on the mean mesh (Vert-NSM figures site).

2.3. Model Training

The baseline Vert-NSM model is a triplanar autoencoder, adapted from Gatti and colleagues (2025; <https://github.com/gattia/nsm>; **Figure 2**). It uses a hybrid explicit–implicit architecture where shape is represented as a signed distance function (SDF; Park et al. 2019). To obtain ground truth SDFs, meshes were centered and scaled to a unit sphere by maximum radial distance and 500,000 points were sampled per mesh: 40% near the surface (Gaussian $\sigma=0.0158$), 40% further out ($\sigma=0.05$), and 20% uniformly distributed. Each training batch subsamples 17,000 of these query points per mesh, balanced between positive and negative SDF values.

Since the framework lacks an encoder, each mesh is randomly assigned a 512-dimensional latent (global z), which is iteratively optimized with network weights. The global z is decoded by a five-layer convolutional network into three orthogonal feature planes. Each query point is projected onto the planes, its features read by bilinear interpolation and summed into a point-specific local latent (local z). The local z and the point's coordinates are passed to a three-layer multilayer perceptron (MLP) that predicts the signed distance, which is compared with ground truth values under L1 SDF reconstruction loss. Global z is regularized by L2 loss. Gradients minimize these losses, updating the network weights (reconstruction loss only) and the latents (both losses; Gatti et al. 2025).

Two custom loss functions were evaluated to encourage taxonomic structure in the latent space for classification tasks — hierarchy loss and contrastive loss (**File S1**). Unlike the BioCLIP implementation (Stevens et al. 2024; Zhang et al. 2022b), which embeds and aligns taxonomic labels with samples, both act directly on the global z and are optimized jointly with L1 and L2 losses. Contrastive loss follows InfoNCE (Radford et al. 2021; van den Oord et al. 2019), operating at the species level alone, pulling each latent toward its own species mean and repelling it from other sampled means (weight=0.02, temperature=0.2, negatives per anchor=32). The hierarchy loss attracts same-species latent pairs and repels all others, with margins scaled by taxonomic distances (0.2 within genus, 0.4 within family, 0.8 between families; weight=0.001, warmup=200 epochs).

Models were trained using random 80-15-5 train-test-validation splits (N=2316, 434, 146 meshes). Training for 2000–3000 epochs took approximately 14 hours on a single GPU with 48 GB VRAM (**File S1**).

2.4. Model inference and evaluation

As an autoencoder, NSM lacks an encoder to directly map novel meshes to latents. Encoding is either by latent optimization or with a separately trained encoder. For optimization, a latent is initialized and optimized under L1 loss against SDF ground truth samples. For the encoder, a latent is predicted from near-surface samples ($|SDF| < 0.01$) and optionally refined. For both, SDF samples were drawn as in training. Shape completion is more sensitive to encoding, since small latent errors can yield noisy, implausible surfaces, while classification is comparatively robust, depending only on the relative position of an encoding (**Figure 3**).

For shape completion we evaluated three encoding strategies: two-phase optimization, the encoder only, and the encoder with refinement. In two-phase optimization, a latent is initialized at the mean with a small random offset (1% SD) along the top principal components (99% of latent variance), then optimized against partial mesh SDF samples. The first phase establishes an approximate latent position. The second refines it to recover finer surface details. The PointNet encoder passes 2,048 surface points through a four-layer shared MLP, max pools them into a global descriptor, and maps it to the NSM latent through three fully connected layers (Qi et al. 2017). It was trained on reconstructed shapes with random fragmentation and added noise. Predicted SDFs were converted to surfaces by marching cubes on a 256^3 grid (Lorensen and Cline 1987). Parameters were determined by grid search over 15 trials, each scored as mean Chamfer distance across 30 validation meshes, with the best configuration applied to 50 validation meshes (**Table S2**).

Classification used a single encoding phase. A latent is initialized at the mean with a small random offset (1% SD) along the top principal components (95% of latent variance), then optimized for 1,000 iterations with Adam ($\text{lr} = 1 \times 10^{-3}$) under L1 loss.

Best inference parameters for shape completion and classification determined above were used to evaluate model performance for train-test-val datasets.

2.4.1. Shape Completion

Meshes were synthetically fragmented using seven homologous segments generated with MorphoWeave (Porto et al. 2026). One segment was subtracted from every mesh following procedures on the Vert-NSM wiki. Partial meshes were passed to the model for inference and completed surfaces were compared against ground truths using symmetric Chamfer distance and inference time (**Figure 3**). Distances are in normalized units, where one corresponds to the maximum radial distance from the vertebral centroid (roughly the vertebral radius). Chamfer distance is reported as a percentage ($\text{CD} \times 100$; e.g., raw CD of 3.75×10^{-3} becomes 0.375%), with lower values indicating better reconstruction.

2.4.2. Classification

Classification was by nearest-neighbour retrieval. Each query vertebra was encoded to a latent and compared by cosine similarity against the gallery of training latents. Train latents were re-encoded with the same parameters as test and validation splits for better comparability. A query is correct under Top-1 accuracy if the nearest neighbor carries its true label, and under Top-5 if its true label appears in the five nearest. Top-5 reflects the more realistic use case for herpetologists, collection managers, and researchers, where a ranked shortlist provides reference specimens to check against an unknown vertebra.

Accuracy was assessed for spinal position (region: cervical, thoracic, lumbar; and normalized position in 20% bins), taxonomy (species, genus, family, broad taxonomic group), and life history strategy. To maximize taxonomic diversity, training data was randomly split by vertebra rather than by specimen, meaning vertebrae from one individual appear across data splits. Since retraining a model for each held-out specimen was not computationally feasible, two masking levels were used. Leave-one-vertebra-out (LOO) hides only the query vertebra, leaving other vertebrae from the same specimen available. Leave-one-specimen-out (LOSO) hides all vertebrae from the query specimen, so a correct match must come from a different individual. LOO approximates inference of a vertebra whose parent taxon is in the training data and LOSO approximates generalization to out-of-distribution data. Accuracy was also compared against frequency-weighted chance (**File S1**).

Under LOSO, masking can leave a query with no gallery entries at a given taxonomic level. For example, a species represented by a single specimen is excluded from species-level accuracy, but still scored at genus and family level, following the phylogeny in **File S2**. *Sphenodon* (Rhynchocephalia, non-squamate outgroup) and *Dibamus* (Dibamidae, sister to all other squamates) were singleton specimens and excluded from taxonomic analyses. *Amphisbaenia* lack spinal regionalization and were excluded from spinal region analysis. Per-class Top-1 and Top-5 recall under LOSO were correlated with the number of specimens and vertebrae per broad taxonomic group across splits using Spearman's ρ .

2.5. Minimum training dataset size

To evaluate robustness to training dataset size, we drew subsets of 10, 30, 50, and 70 from the full 95 specimen dataset, stratified by broad taxonomic group (≥ 2 specimens per group where possible; **File S3**). Subsets were split 80-15-5, then trained and tuned as described above. For evaluation, meshes held out from training sets of all models (N=138) were used to score shape completion and classification.

2.6. Comparison with landmark-based approaches

Geometric morphometric analyses were implemented in Python following R geomorph (Collyer and Adams 2024; Adams and Otárola-Castillo 2013) to compare NSM latents against sparse landmarks and dense correspondences. Data checks for scaling, alignment, and symmetry supported analysing all three representations in full (**File S1**).

2.6.1. Morphospace exploration and dimensionality reduction

Given the dimensionality of NSM latents (512-d), we compared principal component analysis (PCA) against two non-linear methods, t-distributed stochastic neighbour embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) for sparse landmarks, dense correspondences, and latents. All morphospace analyses used the training split (N=2316 vertebrae). PCA was computed on Procrustes shape coordinates for sparse landmarks and dense correspondences (as in `gm.prcomp` in R geomorph) and on mean-centered latent vectors for NSM (scikit-learn; Pedregosa et al. 2018). For t-SNE and UMAP, inputs were capped at the first 50 PCs with the same parameters used across representations (**File S1**).

Functional morphology signal in morphospace was assessed by linear discriminant analysis (LDA) for life history classes on principal component scores, retaining 90%, 95%, and

99% of cumulative variance. Given the number of components relative to specimens (NSM latents: 266–466; landmark approaches: 12–315), LDA used shrinkage-regularized covariance estimation (eigen solver with Ledoit–Wolf shrinkage in scikit-learn). Balanced accuracy under five-fold cross-validation is reported (**Table S8**). This analysis is descriptive and shows group arrangement in each morphospace rather than classification performance (see **2.4.2**).

2.6.2. Geometric morphometrics

To test whether principal component space is shared across representations (sparse landmarks, dense correspondences, and NSM latents), we computed pairwise Spearman's rank correlations for PC1–PC4 scores with Bonferroni correction. Since the sign and ordering of principal components are arbitrary, $|p|$ is reported and each component matched to its strongest counterpart in the other representation.

We interpreted morphology of principal components by traversing PC1–PC4 in four even steps spanning 1.5 times the observed range of scores from minimum to maximum. For sparse landmarks, the consensus configuration was displaced along the corresponding eigenvector at each step, then used to warp the mean mesh by thin-plate spline (TPS; following warpRefMesh in R geomorph). For NSM, latents were displaced along each component from the specimen nearest the latent mean and decoded directly into meshes. NSM PC1, PC3, and PC4 were reversed in the figure to match landmark rows.

To test for group-level differences across NSM and landmark morphospaces, morphological disparity was computed as Procrustes variance per family and life history category (999 iterations; following morphol.disparity in R geomorph). Since raw Procrustes variance from the 512-dimensional latent space and the 84-dimensional landmarks ($28 \times 3 = 84$) are not comparable, groups were compared by rank from most to least disparate. Reported ranks are from mean-centered, PCA-reduced latents (84 PCs, retaining 72.9% of latent variance), which agreed closely with full-dimensional ranks, ruling out dimensionality artifacts ($p=0.961$ for families, 1.000 for life histories; **Tables S6–S7**).

2.7. Approaches unique to NSM

In addition to shape completion and classification, the continuous latent space enables interpolation between any encoded latents and has no equivalent in landmark-based approaches. Two series were generated: saxicolous (Gerrhosauridae: *Cordylus* *subtessellatus*) to grass-swimming (Cordylidae: *Chamaesaura* *aenea*) to fossorial (Trogonophidae: *Diplometopon* *zarudnyi*) and varanid lizard (Varanidae: *Varanus* *komodoensis*) to parviraptorid fossil (Parviraptoridae: *Breugnathair* *elgolensis*) to homalopsid snake (Homalopsidae: *Homalopsis* *buccata*).

For endpoints drawn from the training set, a mean latent was computed across that specimen's thoracic vertebrae. For the fossil and snake endpoints, a latent was encoded with PointNet followed by 500 refinement iterations ($l_r=1 \times 10^{-3}$, $\lambda_{reg}=1 \times 10^{-5}$). Intermediate shapes were generated by linear interpolation at five evenly spaced steps and decoded to surfaces.

3. RESULTS

3.1. Vert-NSM models

Shape completion and classification improved substantially as training set size increased from 10 to 30 specimens: shape completion by 20% (mean CD) and classification of normalized spinal position by 17% (Top-1 LOSO). Beyond 30 specimens, shape completion plateaued, gaining less than 5% additional accuracy up to 95 specimens, while classification continued to improve (**Figure S1**; **Tables S9–S10**; **File S3**). We recommend a minimum of 30 taxonomically diverse specimens for either task, with additional gains for classification from larger datasets. Demo notebooks for shape completion and classification are available in Google Colaboratory through the Vert-NSM GitHub.

3.1.1. Shape completion

The PointNet encoder with refinement (iterations=500, LR= 1×10^{-3} , LR scheduler step=300, gamma=0.9) achieved the best reconstruction. Compared to the base model with two-phase optimization, Chamfer distance was roughly half (CD=0.375–0.384% versus 0.612–0.794%) and inference was 7.4 times faster (26.7 s versus 198.45 s; **Figure 4**; **Table 1**; **Table S2**; **File S4**). The encoder alone was fastest (6.53 s), but had the highest reconstruction error (CD=0.963–0.978%), meaning refinement recovers accuracy at a fraction of the cost of full two-phase optimization.

Encoder-based approaches (with and without refine steps) generalized best across train-test-val splits, with chamfer distance varying less than 3% versus over 20% for other approaches (**Figure 4C**). Baseline, hierarchy, and contrastive models had comparable accuracy on held-out data (test CD=0.781–0.785%) and differed mainly in inference time (198.45s, 133.30s, and 167.07s, respectively).

3.1.2. Classification

Baseline, hierarchy, and contrastive models performed comparably, with differences generally under one percentage point across splits (**Table 2**; **Table S3**). As neither custom loss function yielded a consistent improvement, the baseline was selected for subsequent analyses.

Accuracy varied with masking strategy. For genus, Top-1 to Top-5 accuracy was 96.0–99.8% under LOO and 39.1–65.2% under LOSO (chance 2.5–13.5%). For normalized spinal position, it was 75.9–98.6% under LOO and 52.0–86.3% under LOSO (chance 20.2–67.2%).

Confusion matrices show Top-1 results under LOSO using the train split, which provides the largest, most diverse sample (**Figure 5**). Vertebrae at the extremes of the spinal column classified most reliably, while mid-column vertebrae were confused with adjacent positions (**Figure 5A,D**). Among life history groups, grass-swimmer (75%), fossorial (70%), and saxicolous (67%) classified best, while arboreal were more often mistaken for terrestrial than correctly identified (49% versus 22%; **Figure 5B**). For broad taxonomic group, Cordylidae classified most accurately (90%), followed by Amphisbaenia (81%) and Gekkota (64%). Scincidae and Xantusiidae were poorly classified (10% and 0%; **Figure 5C**). Top-1 accuracy increased with number of vertebrae per class ($p=0.69$ – 0.78 , $P=0.014$ – 0.041 across splits), but not with number of specimens ($p=0.51$ – 0.61 , $P=0.08$ – 0.16 ; **Tables S4–S5**).

3.2. Landmarks versus latents

To establish whether NSM latents recover the morphological patterns available from landmark-based analysis, we compared morphospaces of sparse landmarks, dense correspondences, and NSM latents.

3.2.1. Morphospace exploration and dimensionality reduction

Taxonomic signal was recoverable from all three morphospaces, but landmark-based representations concentrated it in early linear axes, while NSM latents required the full space (**Figure 6**). PC1–PC2 accounted for 53.6% of variance in sparse landmarks, 66.2% in dense correspondences, and 20.2% in NSM latents, while PC3–PC4 accounted for 14.0%, 11.2%, and 9.4%, respectively. All representations separated *Amphisbaenia* on PC1–PC2 and *Cordylidae* on PC3–PC4, where landmark representations additionally separated *Gekkota*. Non-linear ordination recovered consistent structure across all three, with *Amphisbaenia*, *Gekkota*, and *Cordylidae* distinguishable in t-SNE and UMAP.

Life-history separability in LDA differed by variance threshold (**Figure 7**). NSM latents gave the clearest separation at the 90% and 95% thresholds, resolving fossorial and grass-swimming taxa, with terrestrial and saxicolous sharing a cluster. Landmark-based representations separated only fossorial taxa. At 99%, NSM was more diffuse and dense correspondences separated the most cleanly. Arboreal taxa were poorly resolved throughout. Cross-validated accuracies were stable across thresholds for NSM (0.896–0.904) but rose with dimensionality for landmarks (0.706–0.790 sparse, 0.674–0.943 dense; **Table S8**). Interactive PCA, t-SNE, UMAP, and LDA plots are available (Vert-NSM figures site).

3.2.2. Geometric morphometrics

Sparse landmarks and dense correspondences recovered comparable axes of shape variation (**Figure 8**). Their first two components were near-identical ($|p|=1.00$ for PC1 and 0.95 for PC2), while the third and fourth were partially transposed (sparse PC3 to dense PC4: $|p|=0.73$; sparse PC4 to dense PC3: $|p|=0.62$). NSM latents shared their principal axis with both landmark representations (NSM PC1 to sparse PC1: $|p|=0.79$; NSM PC1 to dense PC1: $|p|=0.78$). NSM PC4 recovered variation expressed on landmark PC3 ($|p|=0.70$) and on dense PC3 and PC4 ($|p|=0.60$ and 0.56). NSM PC2 and PC3 lacked strong correspondence with either ($|p|\leq 0.52$). All correlations with $|p|\geq 0.10$ were significant after Bonferroni correction.

PC1 is an anteroposterior elongation axis along the spinal column for sparse landmarks and NSM latents (**Figure 9**). The remaining axes describe subtler variation and differ between representations. Landmark PC2 transitions from a tall, narrow vertebra with an elongated neural spine to a short, wide vertebrae with a reduced spine. NSM PC2 shows a similar, but less pronounced progression with additional widening and thickening of the postzygapophyses. NSM PC3 produces an implausible biological shape (columns 3–4), indicating that traversal and reconstruction are unreliable in regions of latent space sparsely populated by the training data. Videos of PC1–PC4 traversal are available (Vert-NSM figures site).

3.2.3. Morphological disparity

We evaluated group-level differences in morphological disparity across NSM and landmark morphospaces using ranks (1–40, most to least disparate; **Figure 10**; **Figure S2**). Pooled by life history, NSM latents recovered greater disparity for every strategy except

arboreal, with the greatest difference for fossorial taxa (**Figure S2**). At the family level, NSM ranked 22 of 40 families as more disparate than landmarks (bold family names in **Figure 10**), 18 of which had displacements of more than 8 ranks (colored data points). For fossorial families, all but three (Trogonophidae, Rhineuridae, Pygopodidae) were ranked as more disparate under NSM, with displacements of 12 to 18 rank positions for Dibamidae, Lanthanotidae, and amphisbaenian families. Almost all arboreal families ranked more disparate under landmarks, with large differences for Corytophanidae (25 ranks) and Chamaeleonidae (34 ranks), and Diplodactylidae effectively tied (1 rank). Disparity values are in **Tables S6–S7**.

3.3. Approaches unique to NSM

The continuity of latent space allows interpolation between any two encoded specimens. Interpolating from saxicolous to grass-swimming, the neural spine becomes taller, the centrum elongates, and the transverse processes reduce and incline posteriorly (**Figure 11**). From grass-swimming to fossorial, the neural spine and postzygapophyses merge and shorten and the articular surfaces of the prezygapophyses are more medially inflected.

From lizard to parviraptorid, the neural spine shortens anteroposteriorly, the postzygapophyses reduce, the neural canal becomes larger, and the condyle of the centrum shifts from broad and ovoid to circular (**Figure S3**). From parviraptorid to snake, postzygapophyses broaden and a zygantum forms, the condyle of the centrum maintains shape but reduces in size, and a hypapophysis emerges.

4. DISCUSSION

4.1. Reproducible, quantifiable shape completion and classification

To our knowledge, Vert-NSM is the first tool for quantitative, reproducible shape completion and classification of unknown and fragmentary 3D biodiversity specimens. Partial specimens were completed to within 0.375% of a vertebral radius, enabling inclusion of fragmentary material for future comparative analyses (**Table 1**). Sufficient model performance was achieved by training with 30 taxonomically diverse specimens, though classification continued improving with more specimens (**Figure S1**). Shape completion needs enough variation to constrain a plausible surface, while classification depends on coverage, with additional specimens extending the range of possible taxonomic matches.

Squamate vertebrae are historically underutilized for taxonomic inference, as diagnostic features are concentrated in the skull and jaw (Tschopp 2016), yet we recovered signals for taxonomy and spinal position in line with classic approaches (Hoffstetter and Gasc 1969; Etheridge 1967). Classification exceeded chance at every level and class tested, except arboreal (life history; **Figure 5B**), and Scincidae and Xantusiidae (broad taxon; **Figure 5C**). Genus Top-1 accuracy reached over 15 times random chance (LOSO: 39.1–44.7%; chance: 2.5–2.9%; **Table 2**). Top-1 normalized spinal position reached almost three times random chance (LOSO: 52.0–55.5%; chance: 20.2%; **Table 2**), rising to 73% accuracy in the anteriormost 20% of the column (**Figure 5A**).

Neither task has an established benchmark, so results carry caveats. Data splits were drawn by mesh rather than specimen, meaning vertebrae from one individual appear in training and test sets. Although LOO and LOSO masking approximate held-out vertebrae and specimens during retrieval, models were trained with all taxa. Fragmentation was synthetic,

removing one of seven SSM-defined segments. Modelling breaks that approximate real damage and taphonomy may benefit future applications with fossils.

4.2. Classification depends on sample size and morphology

Spinal position is the dominant axis of variation in the dataset, recovered as PC1 in all representations (**Figures 8–9**), and was predicted consistently across splits (**Figure 5A, D; Table 2**). Taxonomic predictions were more uneven (**Figure 5C**), with three patterns explaining most of the spread: groups with more vertebrae classified better, groups that occupied tight clusters of morphospace classified better, and groups with diverse within-group morphology classified worse.

Top-1 recall increased with the number of vertebrae per class ($\rho=0.69\text{--}0.78$, $P=0.014\text{--}0.041$ across splits; **Tables S4–S5**), as expected for nearest-neighbour retrieval approaches. Best accuracy was for Amphisbaenia (570 vertebrae; 81%), Cordylidae (458; 90%), and Gekkota (146; 64%), and all formed tight clusters in morphospace (**Figure 5C; Figure 6**).

Sample size alone does not account for the remaining spread. The only two taxa that classified below random chance (14%) were Xantusiidae and Scincidae (**Figure 5C**). Low accuracy for Scincidae (10%), despite the same vertebrae counts as Gekkota (146; **Table S4**), may reflect its diversity of fossorial, grass-swimming, and terrestrial forms. Xantusiidae was represented by 35 vertebrae from two specimens and classified at 0%, as LOSO masking left a single reference specimen in the gallery.

Removing same-specimen matches under LOSO reduced genus accuracy more than spinal position (genus: $\sim 98\%$ to $\sim 42\%$; position: $\sim 76\%$ to $\sim 52\text{--}56\%$; **Table 2**). Positional signal generalizes across individuals better than taxonomic signal. Since position dominates variation in the dataset, taxonomy and position may be better served by separate classifiers trained on disentangled components of the latent space

4.3. Encoding drives performance

The largest performance gain came from inference encoding methods. A PointNet encoder with refinement halved reconstruction error, ran 7.4 times faster, and generalized best, with error varying by less than 3% across splits, against more than 20% for the base model (**Figure 4; Table 1; Table S2**). Since autoencoder frameworks lack a learned encoder, every novel specimen requires optimization, which is slow and depends on initialization, learning rate, and iterations. Shape completion is especially sensitive to encoding. If a latent lands slightly off the learned manifold, it reconstructs into implausible shapes, as we observed for the maximum shapes of PC3 of NSM (**Figure 9**).

4.4. Taxonomic losses did not improve the latent space

Neither loss function improved classification over the base model, with differences under one percentage point across data splits (**Table 2**). Both assume that taxa separated by a given rank are equally distant in morphology, which does not reflect morphological or evolutionary patterns (Rusak et al. 2025). For example, Amphisbaenian families are far more disparate than most others in our dataset (**Figure 10**), yet hierarchy loss would separate them by the same 0.8 margin applied to any between-family pair. Future efforts could replace discrete rank margins

with continuous phylogenetic distances under an evolutionary model, like Brownian motion (see for e.g., PhyloSDF by Lau and Choi 2026).

4.5. Analyzing high-dimensional latents versus landmark data

NSM latents concentrated less variance in their early principal components than landmark representations and showed little taxonomic structure under linear projection using PCA (PC1–PC2: 20.2% for NSM against 53.6% and 66.2% for sparse and dense; **Figure 6**). Nonetheless, NSM PC1 matched the dominant axis of variation for both landmark representations ($|p|=0.79$ and 0.78 ; **Figure 8**), and taxonomic structure was recovered by non-linear ordination (t-SNE, UMAP).

Inspecting principal components is standard practice in geometric morphometrics, but in a 512-dimensional latent space, information is spread across too many components for linear projections to be interpretable. In LDA, NSM separated best at the 90% and 95% thresholds, but was diffuse at 99%, even though balanced accuracy was unchanged (0.896–0.904; **Table S8**; **Figure 7**). Across the same thresholds, landmark representation accuracy varied by up to 27 percentage points, showing the importance of reporting results across a range of dimensions. Matryoshka Representation Learning and nested dropout could concentrate and disentangle information in early latent dimensions during training (Kusupati et al. 2022), allowing a small subset to be inspected like principal components, while retaining full-dimensional latents for other tasks.

4.6. Landmark-free representations and future directions

Landmarks can miss variation when structures are reduced or absent, while continuous shape representations can obscure localized differences. Morphological disparity ranks agreed for most families and life history strategies, and disagreed when distinctive features fell outside landmarked regions, or were too localized for NSM to register. The largest disagreements were for two morphologically distinctive families: Bipedidae (fossorial) ranked more disparate under NSM latents (19 ranks), while Chamaeleonidae (arboreal) ranked more disparate under landmarks (34 ranks; **Figure 10**). Fossorial taxa have a broad neural canal and a reduced or absent neural spine, which are poorly sampled by landmarks (**Figure 1**). Arboreal taxa concentrate variation in well-landmarked structures (tall neural spine and anteroposteriorly compressed centrum).

Disparity measures variance in morphospace rather than direct morphological differences, so ranks between representations should be interpreted with caution. Latent variance is driven by global geometry, since L1 SDF loss weights all regions equally. Landmark variance is driven only by local geometry at the chosen points. For Vert-NSM, user-designated region weighting could emphasize morphologically important features while preserving full-surface coverage. Existing methods that automatically weight reconstruction errors in challenging regions offer a starting point (e.g., Huang et al. 2024; Furuya et al. 2023). Since landmarks were used to align and scale meshes in preprocessing, Vert-NSM is landmark-free only at analysis. Rigid atlas registration or pose-invariant training could remove this need.

Latent interpolation offers a reproducible way to visualize morphological transitions, such as convergent limb reduction in fossorial and grass-swimming lizards (**Figure 11**). For parviraptorids, an early squamate lineage repeatedly reclassified on its mosaic of lizard- and

snake-like traits (e.g., Benson et al. 2025; Martinez et al. 2021; Caldwell et al. 2015), interpolation renders a continuous path between lizards, parviraptorids, and snakes (**Figure S3**). These are morphological intermediates, not ancestral states, and linear paths in latent space are unconstrained by phylogeny or development. Applying phylogenetic or developmental models during training (see **4.4**) would constrain traversal to hypothesis-driven trajectories and support quantitative comparisons.

5. CONCLUSIONS

Vert-NSM used a shape prior learned from almost 3,000 meshes across 40 families to do shape completion and classification of squamate vertebrae. As a continuous shape representation, it shared the dominant axes of variation recovered by landmarks, and identified shape differences that did not depend on anatomical priors. Models performed well with as few as 30 taxonomically diverse specimens. Vert-NSM has the potential to unlock discoveries from partial and disarticulated skeletal elements that are an abundant, but underutilized, part of museum collections and the fossil record. Future approaches should leverage models of phylogeny and development to structure the latent space and enable informed hypothesis testing in deep time.

Acknowledgements

We thank Breanna Shi, Riyam Zaman, and Charlie Clark for administrative support at the HAAG lab, and BioVision lab members Heru Handika, Caleb Gordon, and Lucas Teles Bezerra for feedback on figures and analyses. We thank vertebrate paleontologist Jason Bourque for biological insights, use cases, and feedback on the manuscript. This work was supported by travel and training funds from the SlicerDepot and MorphoCloud Bootcamp and the Department of Natural History Travel Award, which enabled participation in VIZBI 2026. A.P. was supported by the National Science Foundation (DEB-2529932 and EAR-2422839) and the datasets by NSF RCN-2226185, DMR-2017977, and DB-1701714. The conclusions expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Large language models (OpenAI GPT-5; Anthropic Claude Sonnet 5 and Opus 5) assisted with software development and language editing. All output was reviewed by the authors, who take responsibility for the final content.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Author Contributions

Katherine A. Wolcott: Writing – original draft, Methodology, Software, Visualization, Formal analysis. Edward L. Stanley: Conceptualization, Methodology, Investigation, Funding acquisition, Supervision. Arthur Porto: Conceptualization, Methodology, Software, Funding acquisition, Software, Supervision. Kelsey Fenner: Investigation. Anthony A. Gatti: Software, Methodology (NSM & Vert-NSM development). Amar Cutura, Matt Delage, Ali Sean Khan, Jia Xuan Koh, Mitchell Riegler, Colin Sawyer, Joshua Yu: Software. All authors: Writing – review & editing.

Statement on Inclusion

We used existing oVert micro-CT data, so no fieldwork was conducted in any country of origin. All data, models, and code are openly available, supporting reuse by researchers without substantial microscopy or collections access.

Data Availability

Micro-CT data are available on MorphoSource (Project 000368762). Code is available at <https://github.com/3D-fossils-Haag/vert-nsm>. Training data, trained models, and evaluation outputs at <https://huggingface.co/BioVisionLab>. Details in Materials and Methods.

References

- Adams, D. C., & Otárola-Castillo, E. (2013). Geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, 4(4), 393–399. <https://doi.org/10.1111/2041-210X.12035>
- Arbour, J. H., & Brown, C. M. (2014). Incomplete specimens in geometric morphometric analyses. *Methods in Ecology and Evolution*, 5(1), 16–26. <https://doi.org/10.1111/2041-210X.12128>
- Benson, R. B. J., Walsh, S. A., Griffiths, E. F., Kulik, Z. T., Botha, J., Fernandez, V., Head, J. J., & Evans, S. E. (2025). Mosaic anatomy in an early fossil squamate. *Nature*, 647(8090), 673–679. <https://doi.org/10.1038/s41586-025-09566-y>
- Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., Eren, K., Cervantes, J. I., Xu, B., Beuttenmueller, F., Wolny, A., Zhang, C., Koethe, U., Hamprecht, F. A., & Kreshuk, A. (2019). ilastik: Interactive machine learning for (bio)image analysis. *Nature Methods*, 16(12), 1226–1232. <https://doi.org/10.1038/s41592-019-0582-9>
- Blackburn, D. C., Boyer, D. M., Gray, J. A., Winchester, J., Bates, J. M., Baumgart, S. L., Braker, E., Coldren, D., Conway, K. W., Rabosky, A. D., De La Sancha, N., Dillman, C. B., Dunnum, J. L., Early, C. M., Frable, B. W., Gage, M. W., Hanken, J., Maisano, J. A., Marks, B. D., ... Stanley, E. L. (2024). Increasing the impact of vertebrate scientific collections through 3D imaging: The openVertebrate (oVert) Thematic Collections Network. *BioScience*, 74(3), 169–186. <https://doi.org/10.1093/biosci/biad120>
- Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 2022(December 2021), 1640–1660. <https://doi.org/10.1111/2041-210x.13901>
- Bourque, J. R., & Stanley, E. L. (2025). A tegu-like lizard (Teiidae, Tupinambinae) from the Middle Miocene Climatic Optimum of the southeastern United States. *Journal of Paleontology*, 1–15. <https://doi.org/10.1017/jpa.2024.89>
- Boyer, D. M., Lipman, Y., St. Clair, E., Puente, J., Patel, B. A., Funkhouser, T., Jernvall, J., & Daubechies, I. (2011). Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proceedings of the National Academy of Sciences*, 108(45), 18221–18226. <https://doi.org/10.1073/pnas.1112822108>
- Boyer, D. M., Puente, J., Gladman, J. T., Glynn, C., Mukherjee, S., Yapuncich, G. S., & Daubechies, I. (2015). A New Fully Automated Approach for Aligning and Comparing Shapes. *Anatomical Record*, 298(1), 249–276. <https://doi.org/10.1002/ar.23084>
- Caldwell, M. W., Nydam, R. L., Palci, A., & Apesteguía, S. (2015). The oldest known snakes from the Middle Jurassic-Lower Cretaceous provide insights on snake evolution. *Nature Communications*, 6(1), 5996. <https://doi.org/10.1038/ncomms6996>
- Camaiti, M., Evans, A. R., Hipsley, C. A., & Chapple, D. G. (2021). A farewell to arms and legs: A review of limb reduction in squamates. *Biological Reviews*, 96(3), 1035–1050. <https://doi.org/10.1111/brv.12690>
- Collyer, M. L., & Adams, D. C. (2024). Interrogating Random and Systematic Measurement Error in Morphometric Data. *Evolutionary Biology*, 51(1), 179–207. <https://doi.org/10.1007/s11692-024-09627-6>

- Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., Roth, H. R., Vercauteren, T., Xu, D., Dogra, P., Ourselin, S., Feng, A., & Cardoso, M. J. (2024). MONAI Label: A framework for AI-assisted Interactive Labeling of 3D Medical Images. *Medical Image Analysis*, 95, 103207. <https://doi.org/10.1016/j.media.2024.103207>
- Dumoncel, J., Subsol, G., Durrleman, S., Jessel, J. P., Beaudet, A., & Braga, J. (2016). How to Build an Average Model When Samples are Variably Incomplete? Application to Fossil Data. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 541–548. <https://doi.org/10.1109/CVPRW.2016.74>
- Etheridge, R. (1967). Lizard Caudal Vertebrae. *Copeia*, 1967(4), 699. <https://doi.org/10.2307/1441880>
- Furuya, T., Liu, W., Ohbuchi, R., & Kuang, Z. (2023). Hyperplane patch mixing-and-folding decoder and weighted chamfer distance loss for 3D point set reconstruction. *The Visual Computer*, 39(10), 5167–5184. <https://doi.org/10.1007/s00371-022-02652-6>
- Gatti, A. A., Blankemeier, L., Van Veen, D., Hargreaves, B., Delp, S. L., Gold, G. E., Kogan, F., & Chaudhari, A. S. (2025). ShapeMed-Knee: A Dataset and Neural Shape Model Benchmark for Modeling 3D Femurs. *IEEE Transactions on Medical Imaging*, 44(3), 1140–1152. <https://doi.org/10.1109/TMI.2024.3485613>
- Gignac, P. M., Aceves, V., Baker, S., Barnes, J. J., Bell, J., Boyer, D., Cunningham, D., Carlo, F. D., Chase, M. H., Cohen, K. E., Colbert, M., Cree, T. D., Daza, J., Dickinson, E., DeLeon, V., Dougan, L., Duffy, F., Dunham, C., Early, C. M., ... Zobek, C. M. (2024). The role of networks to overcome large-scale challenges in tomography: The non-clinical tomography users research network. *Tomography of Materials and Structures*, 5, 100031. <https://doi.org/10.1016/j.tmater.2024.100031>
- Gray, J. A., Stanley, E. L., Sheehy, C. M., Randall, Z. S., Watkins-Colwell, G. J., & Blackburn, D. C. (2025). The Diffusion Diaries: Diffusible Iodine-Based Contrast-Enhanced Computed Tomography for Vertebrate Natural History Specimens. *Integrative Organismal Biology*, 7(1), obaf014. <https://doi.org/10.1093/iob/obaf014>
- Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A., Blackburn, D. C., Edwards, S. V., & Davis, C. C. (2020). Digitization and the Future of Natural History Collections. *BioScience*, 70(3), 243–251. <https://doi.org/10.1093/biosci/biz163>
- Hoffstetter, R., & Gasc, J. P. (1969). Vertebrae and ribs of modern reptiles. In *Biology of the Reptilia* (Vol. 1, pp. 201–310). Academic Press.
- Huang, T., Liu, Q., Zhao, X., Chen, J., & Liu, Y. (2024). Learnable Chamfer Distance for point cloud reconstruction. *Pattern Recognition Letters*, 178, 43–48. <https://doi.org/10.1016/j.patrec.2023.12.015>
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- Kikinis, R., Pieper, S.D., Vosburgh, K.G. (2014). 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In *Intraoperative Imaging and Image-Guided Therapy*. Springer New York. <https://doi.org/10.1007/978-1-4614-7657-3>

- Keklikoglou, K., Faulwetter, S., Chatzinikolaou, E., Wils, P., Brecko, J., Kvaček, J., Metscher, B., & Arvanitidis, C. (2019). Micro-computed tomography for natural history specimens: A handbook of best practice protocols. *European Journal of Taxonomy*, (522), 1–55. <https://doi.org/10.5852/ejt.2019.522>
- Koehl, P., & Hass, J. (2015). Landmark-free geometric methods in biological shape analysis. *Journal of The Royal Society Interface*, 12(113), 20150795. <https://doi.org/10.1098/rsif.2015.0795>
- Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., & Farhadi, A. (2022). Matryoshka Representation Learning. *Advances in Neural Information Processing Systems* 35. <https://doi.org/10.52202/068431-2192>
- Lau, K., & Choi, G. P. T. (2026). PhyloSDF: Phylogenetically-Conditioned Neural Generation of 3D Skull Morphology via Residual Flow Matching (arXiv:2604.25371). *arXiv*. <https://doi.org/10.48550/arXiv.2604.25371>
- Lendemer, J., Thiers, B., Monfils, A. K., Zaspel, J., Ellwood, E. R., Bentley, A., LeVan, K., Bates, J., Jennings, D., Contreras, D., Lagomarsino, L., Mabee, P., Ford, L. S., Guralnick, R., Gropp, R. E., Revelez, M., Cobb, N., Seltmann, K., & Aime, M. C. (2020). The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. *BioScience*, 70(1), 23–30. <https://doi.org/10.1093/biosci/biz140>
- Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21(4). <https://doi.org/10.1145/37402.37422>
- Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1), 654. <https://doi.org/10.1038/s41467-024-44824-z>
- Martínez, R. N., Simões, T. R., Sobral, G., & Apesteguía, S. (2021). A Triassic stem lepidosaur illuminates the origin of lizard-like reptiles. *Nature*, 597(7875), 235–238. <https://doi.org/10.1038/s41586-021-03834-3>
- Meiri, S. (2024). SquamBase—A database of squamate (Reptilia: Squamata) traits. *Global Ecology and Biogeography*, 33(4), e13812. <https://doi.org/10.1111/geb.13812>
- Mitteroecker, P., & Gunz, P. (2009). Advances in Geometric Morphometrics. *Evolutionary Biology*, 36(2), 235–247. <https://doi.org/10.1007/s11692-009-9055-x>
- Mitteroecker, P., & Schaefer, K. (2022). Thirty years of geometric morphometrics: Achievements, challenges, and the ongoing quest for biological meaningfulness. *American Journal of Biological Anthropology*, 178(S74), 181–210. <https://doi.org/10.1002/ajpa.24531>
- Mulqueeney, J. M., Ezard, T. H. G., & Goswami, A. (2024). Assessing the application of landmark-free morphometrics to macroevolutionary analyses. *Evolutionary Biology*. <https://doi.org/10.1101/2024.04.24.590959>
- Nelson, G., & Ellis, S. (2019). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1763), 20170391. <https://doi.org/10.1098/rstb.2017.0391>
- Oord, A. van den, Li, Y., & Vinyals, O. (2019). Representation Learning with Contrastive Predictive Coding (arXiv:1807.03748). *arXiv*. <https://doi.org/10.48550/arXiv.1807.03748>

- Osher, S., & Fedkiw, R. (2003). Signed Distance Functions. In S. Osher & R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces* (Vol. 153, pp. 17–22). Springer New York. https://doi.org/10.1007/0-387-22746-6_2
- Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation (arXiv:1901.05103). arXiv. <https://doi.org/10.48550/arXiv.1901.05103>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2018). Scikit-learn: Machine Learning in Python (arXiv:1201.0490). arXiv. <https://doi.org/10.48550/arXiv.1201.0490>
- Pomidor, B. J., Makedonska, J., & Slice, D. E. (2016). A Landmark-Free Method for Three-Dimensional Shape Analysis. *PLOS ONE*, 11(3), e0150368. <https://doi.org/10.1371/journal.pone.0150368>
- Porto, A., Rolfe, S., & Maga, A. M. (2021). ALPACA: A fast and accurate computer vision approach for automated landmarking of three-dimensional biological structures. *Methods in Ecology and Evolution*, 12(11), 2129–2144. <https://doi.org/10.1111/2041-210X.13689>
- Porto, A., Wolcott, K. A., Fenner, K. L., Gordon, C., Shi, B., & Stanley, E. L. (n.d.). Biology-guided machine learning for robust anatomical correspondence across biodiversity datasets. *BioRxiv*. <https://doi.org/10.64898/2026.09.28.755222>
- Pyron, R., Burbrink, F. T., & Wiens, J. J. (2013). A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evolutionary Biology*, 13(1), 93. <https://doi.org/10.1186/1471-2148-13-93>
- Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation (arXiv:1612.00593). arXiv. <https://doi.org/10.48550/arXiv.1612.00593>
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2103.00020>
- Rolfe, S. M., & Maga, A. M. (2023). DeCA: A Dense Correspondence Analysis Toolkit for Shape Analysis. In C. Wachinger, B. Paniagua, S. Elhabian, J. Li, & J. Egger (Eds.), *Shape in Medical Imaging* (Vol. 14350, pp. 259–270). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-46914-5_21
- Rolfe, S. M., Whikehart, S. M., & Maga, A. M. (2023). Deep learning enabled multi-organ segmentation of mouse embryos. *Biology Open*, 12(2), bio059698. <https://doi.org/10.1242/bio.059698>
- Rusak, E., Reizinger, P., Juhos, A., Bringmann, O., Zimmermann, R. S., & Brendel, W. (2025). InfoNCE: Identifying the Gap Between Theory and Practice (arXiv:2407.00143). arXiv. <https://doi.org/10.48550/arXiv.2407.00143>
- Schlager, S. (2017). Morpho and Rvcg—Shape Analysis in R: R-Packages for Geometric Morphometrics, Shape Analysis and Surface Manipulations. In *Statistical Shape and*

- Deformation Analysis: Methods, Implementation and Applications. Elsevier Ltd.
<https://doi.org/10.1016/B978-0-12-810493-4.00011-0>
- Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., Chao, W.-L., & Su, Y. (2024). BioCLIP: A Vision Foundation Model for the Tree of Life (arXiv:2311.18803). arXiv.
<https://doi.org/10.1109/CVPR52733.2024.01836>
- Tharakan, S., Shepherd, N., Gower, D. J., Stanley, E. L., Felice, R. N., Goswami, A., & Watanabe, A. (2023). High-Density Geometric Morphometric Analysis of Intraspecific Cranial Integration in the Barred Grass Snake (*Natrix helvetica*) and Green Anole (*Anolis carolinensis*). Integrative Organismal Biology, 5(1), obad022.
<https://doi.org/10.1093/iob/obad022>
- Toussaint, N., Redhead, Y., Vidal-García, M., Lo Vercio, L., Liu, W., Fisher, E. M. C., Hallgrímsson, B., Tybulewicz, V. L. J., Schnabel, J. A., & Green, J. B. A. (2021). A landmark-free morphometrics pipeline for high-resolution phenotyping: Application to a mouse model of Down syndrome. Development, 148(18), dev188631.
<https://doi.org/10.1242/dev.188631>
- Tschopp, E. (2016). Nomenclature of Vertebral Laminae in Lizards, with Comments on Ontogenetic and Serial Variation in Lacertini (Squamata, Lacertidae). PLOS ONE, 11(2), e0149445. <https://doi.org/10.1371/journal.pone.0149445>
- Watanabe, A. (2018). How many landmarks are enough to characterize shape and size variation? PLoS ONE, 13(6), 1–17. <https://doi.org/10.1371/journal.pone.0198341>
- Watanabe, A., Fabre, A.-C., Felice, R. N., Maisano, J. A., Müller, J., Herrel, A., & Goswami, A. (2019). Ecomorphological diversification in squamates from conserved pattern of cranial integration. Proceedings of the National Academy of Sciences, 116(29), 14688–14697. <https://doi.org/10.1073/pnas.1820967116>
- Zhang, C., Porto, A., Rolfe, S., Kocatulum, A., & Maga, A. M. (2022). Automated landmarking via multiple templates. PLOS ONE, 17(12), e0278035. <https://doi.org/10.1371/journal.pone.0278035>
- Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2022). Contrastive Learning of Medical Visual Representations from Paired Images and Text (arXiv:2010.00747). arXiv. <https://doi.org/10.48550/arXiv.2010.00747>

FIGURES AND TABLES

Table 1. Shape completion performance across data splits for synthetically fragmented vertebrae, measured as inference time (seconds) and mean and median Chamfer distance (%; lower is better; **Figure 3**). The PointNet encoder with refinement performed best (bold). See details in **Figure 4** and raw data in **File S4**.

Model	T inf (s)	Train (N=2316)		Test (N=434)		Val (N=146)	
		Mean	Median	Mean	Median	Mean	Median
Base	198.45	7.94	7.07	7.82	7.06	6.12	5.90
Encoder	6.53	9.66	9.38	9.78	9.45	9.63	9.46
Encoder + refine	26.71	3.75	3.65	3.81	3.72	3.84	3.75
Hierarchy	133.30	6.14	5.83	7.85	7.11	6.15	5.84
Contrastive	167.07	6.03	5.73	7.81	7.11	6.08	5.88

Table 2. Top-1 / Top-5 classification accuracy for genus and normalized spinal position (20% bins) across data splits under leave-one-vertebra-out (LOO) and leave-one-specimen-out (LOSO) sampling. All predictions reported are above frequency-weighted chance. Class counts differ due to masking criteria. See Top-5 LOSO performance with the base model (bold) in **Figure 5** and raw data in **Table S3**.

Model	Train		Test		Val	
	LOO	LOSO	LOO	LOSO	LOO	LOSO
Genus						
n (# classes)	2209 (70)	714 (10)	409 (69)	141 (10)	141 (55)	46 (10)
Chance	2.6 / 11.8	2.8 / 12.9	2.8 / 12.5	2.9 / 13.5	2.4 / 11.1	2.5 / 11.8
Base	98.1 / 99.1	42.4 / 62.0	98.0 / 99.8	44.7 / 65.2	96.5 / 98.6	39.1 / 54.3
Hierarchy	98.3 / 99.1	41.2 / 60.8	97.8 / 99.8	39.7 / 61.0	96.5 / 98.6	39.1 / 52.2
Contrastive	98.2 / 99.2	42.4 / 61.8	98.8 / 99.8	41.8 / 63.8	96.5 / 98.6	39.1 / 56.5
Spinal Position						
n (# classes)	2316 (5)	2316 (5)	435 (5)	435 (5)	146 (5)	146 (5)
Chance	20.1 / 67.3	20.1 / 67.3	20.0 / 67.2	20.1 / 67.2	20.2 / 67.4	20.2 / 67.5
Base	75.9 / 98.1	52.0 / 85.5	76.3 / 97.5	55.4 / 86.0	77.4 / 98.6	55.5 / 86.3
Hierarchy	76.1 / 98.0	52.4 / 84.8	75.3 / 98.4	54.4 / 85.3	78.1 / 98.6	56.2 / 84.9
Contrastive	75.9 / 98.1	52.1 / 85.3	76.5 / 97.2	55.1 / 84.8	74.7 / 98.6	56.2 / 86.3

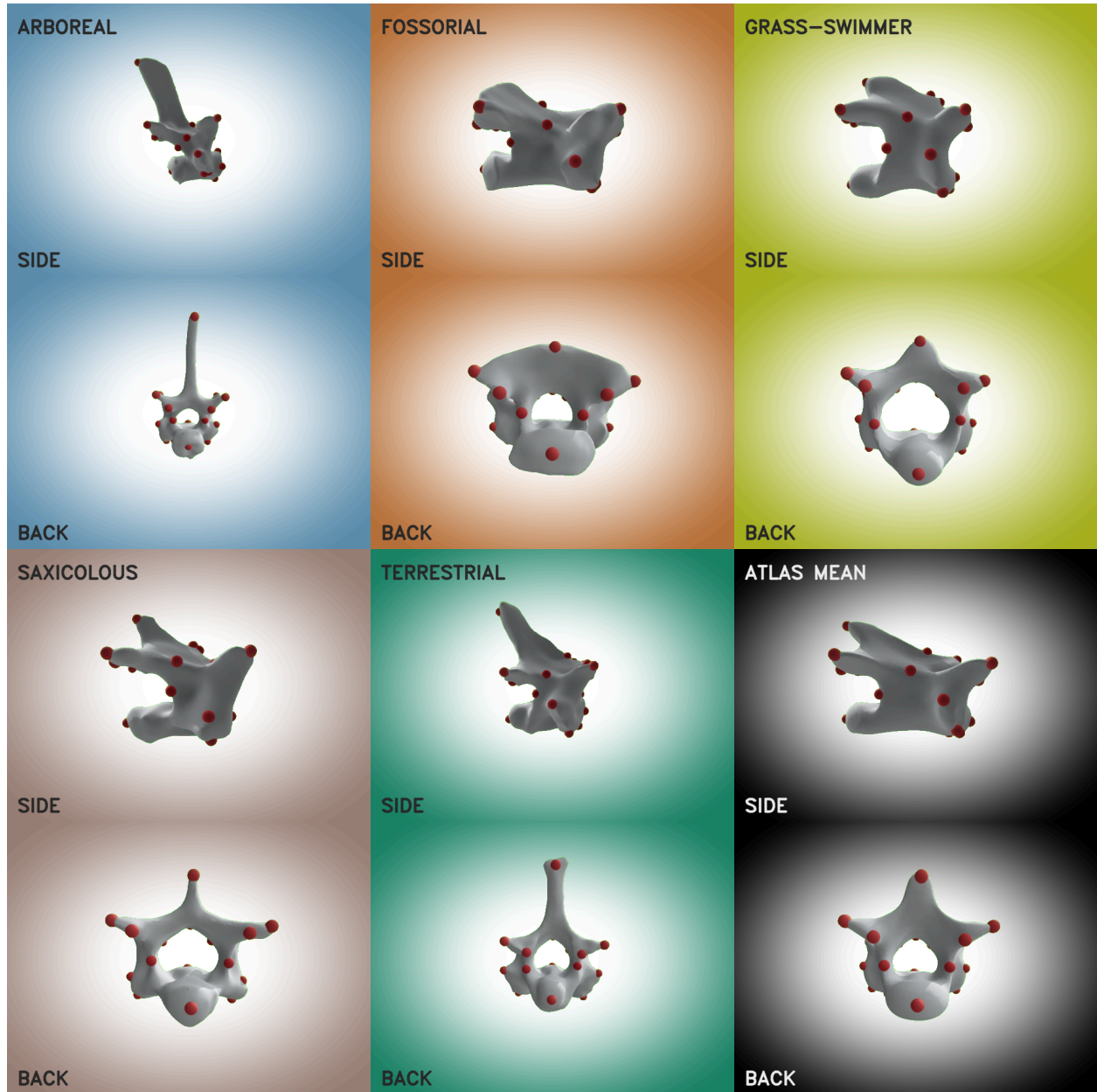


Figure 1. Example vertebrae and landmarks for each life history category (Arboreal, Fossorial, Grass-swimmer, Saxicolous, Terrestrial) and the mean specimen. Shown in lateral (side) and posterior (back) views. An interactive rendering of the landmarked mean is available ([Vert-NSM figures site](#)).

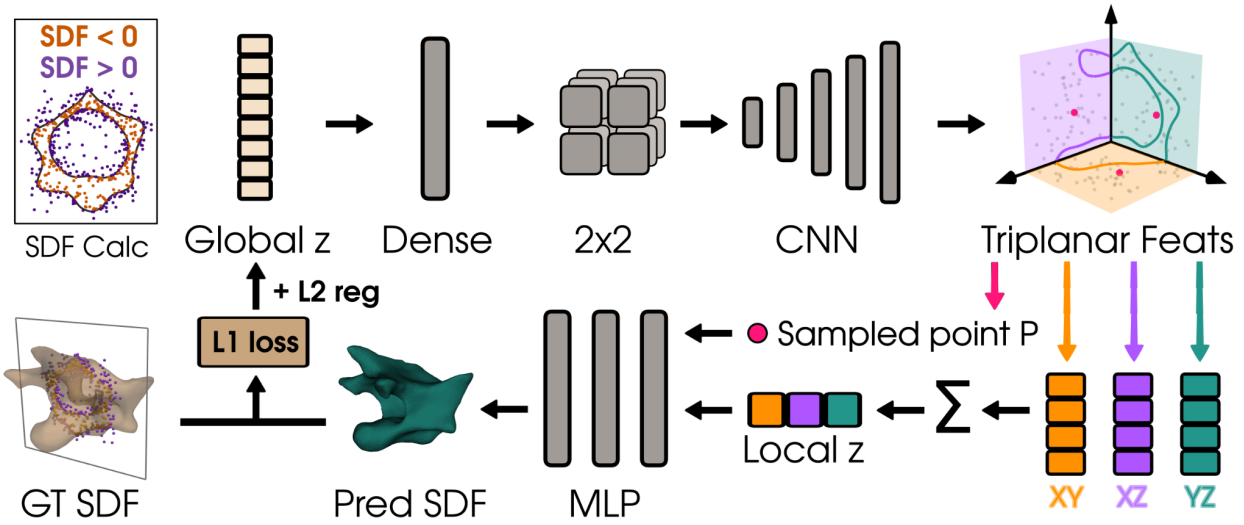


Figure 2. Vert-NSM triplanar autodecoder architecture. Each vertebra's 512-dimensional latent (global z) is decoded by a convolutional network into XY, XZ, and YZ feature planes. Features at each query point (P) are interpolated into a local latent (local z) and passed with P's coordinates to an MLP that predicts the SDF. L1 reconstruction loss compares predicted and ground truth SDFs and updates network weights and latents, while L2 loss regularizes global z. The orthoslice shows SDF sampling (negative inside, positive outside).

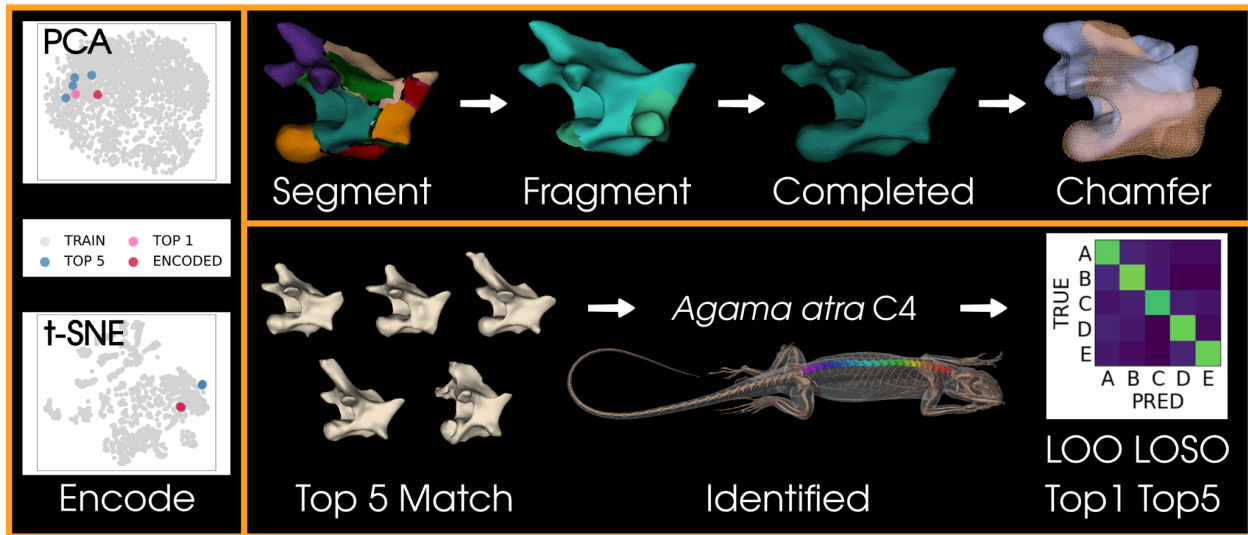


Figure 3. Inference and evaluation for novel meshes: encoding with latent-space visualization of nearest neighbors (left), shape completion (top right), and classification (bottom right). In shape completion, meshes were synthetically fragmented using seven SSM-defined segments (Porto et al. 2026), completed, and evaluated against ground truths by Chamfer distance. In classification, nearest-neighbors produce a candidate list, scored as Top-1 and Top-5 accuracy under leave-one-vertebra-out (LOO) and leave-one-specimen-out (LOSO) masking.

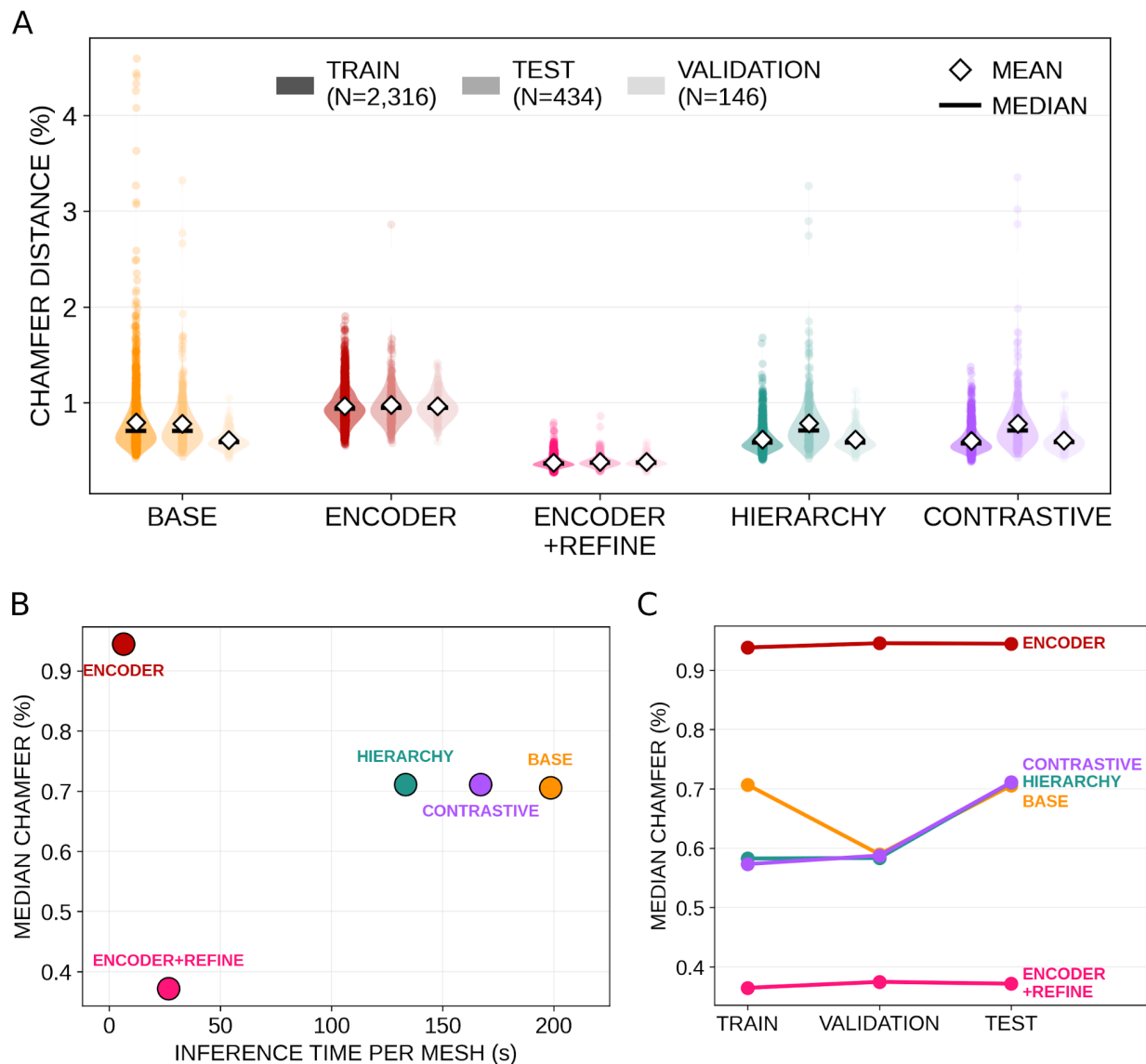


Figure 4. Shape completion performance for synthetically fragmented vertebrae: **(A)** Chamfer distance, **(B)** accuracy–speed tradeoffs, and **(C)** generalization across splits. The PointNet encoder with refinement performed best ($CD=0.375\text{--}0.384\%$) and was $7.4\times$ faster than the base model.

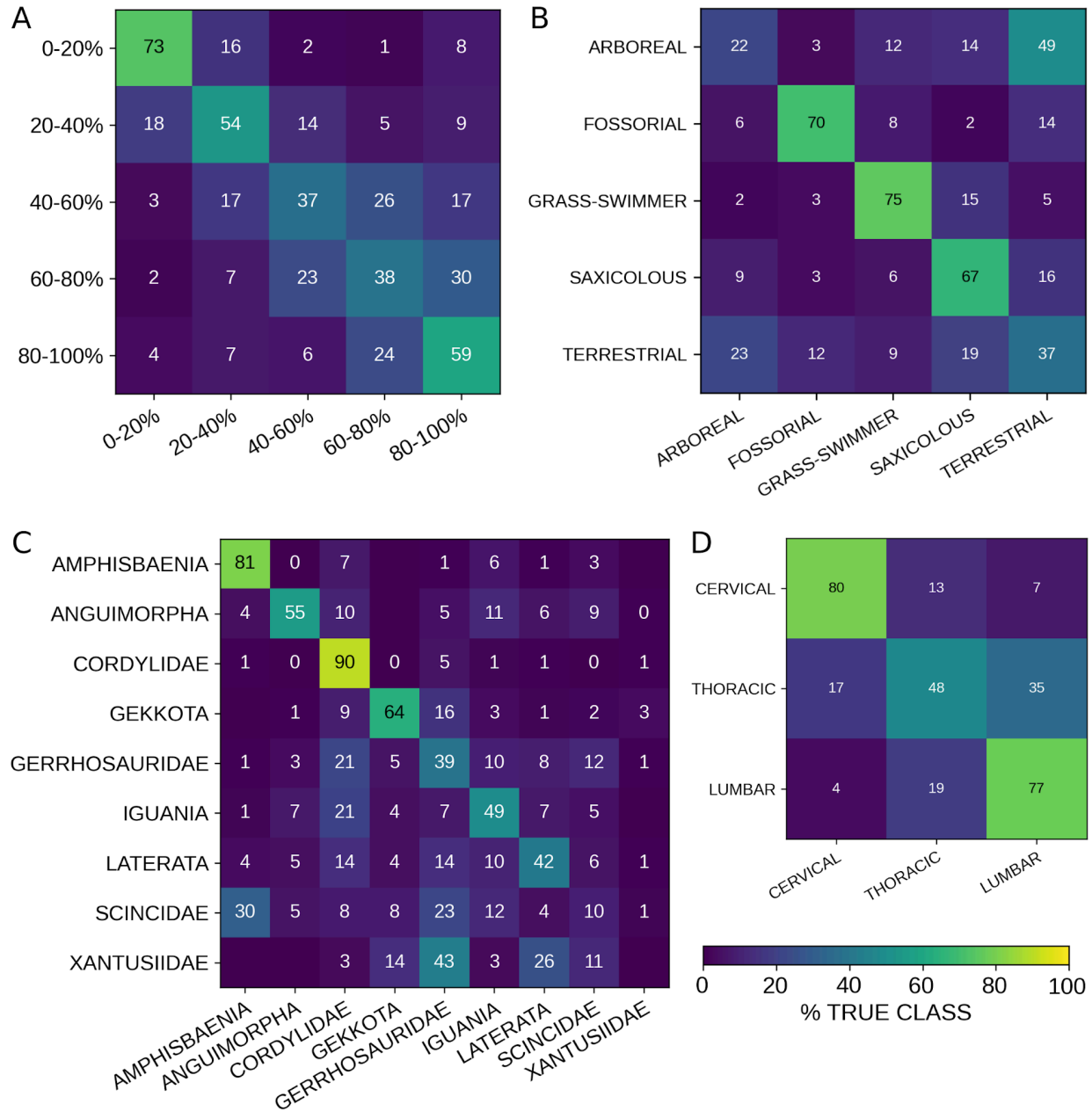


Figure 5. Top-1 classification performance of the base model under LOSO: **(A)** normalized spinal position, **(B)** life history strategy, **(C)** broad taxonomic group, and **(D)** spinal region. Cells show accuracy (%) for true (rows) versus predicted (columns) classes. All predictions exceeded frequency-weighted chance (**A**: 20%; **B**: 25%; **C**: 14%; **D**: 39%), except Arboreal (**B**) and Scincidae and Xantusiidae (**C**). Accuracy was highest at spinal column extremes (**A**: 0-20%: 73%; **D**: Cervical: 80%), for fossorial (70%) and grass-swimmer (75%) life history strategies (**B**), and for broad taxa Cordylidae (90%) and Amphisbaenia (81%; **C**). Raw values in **Table S3**.

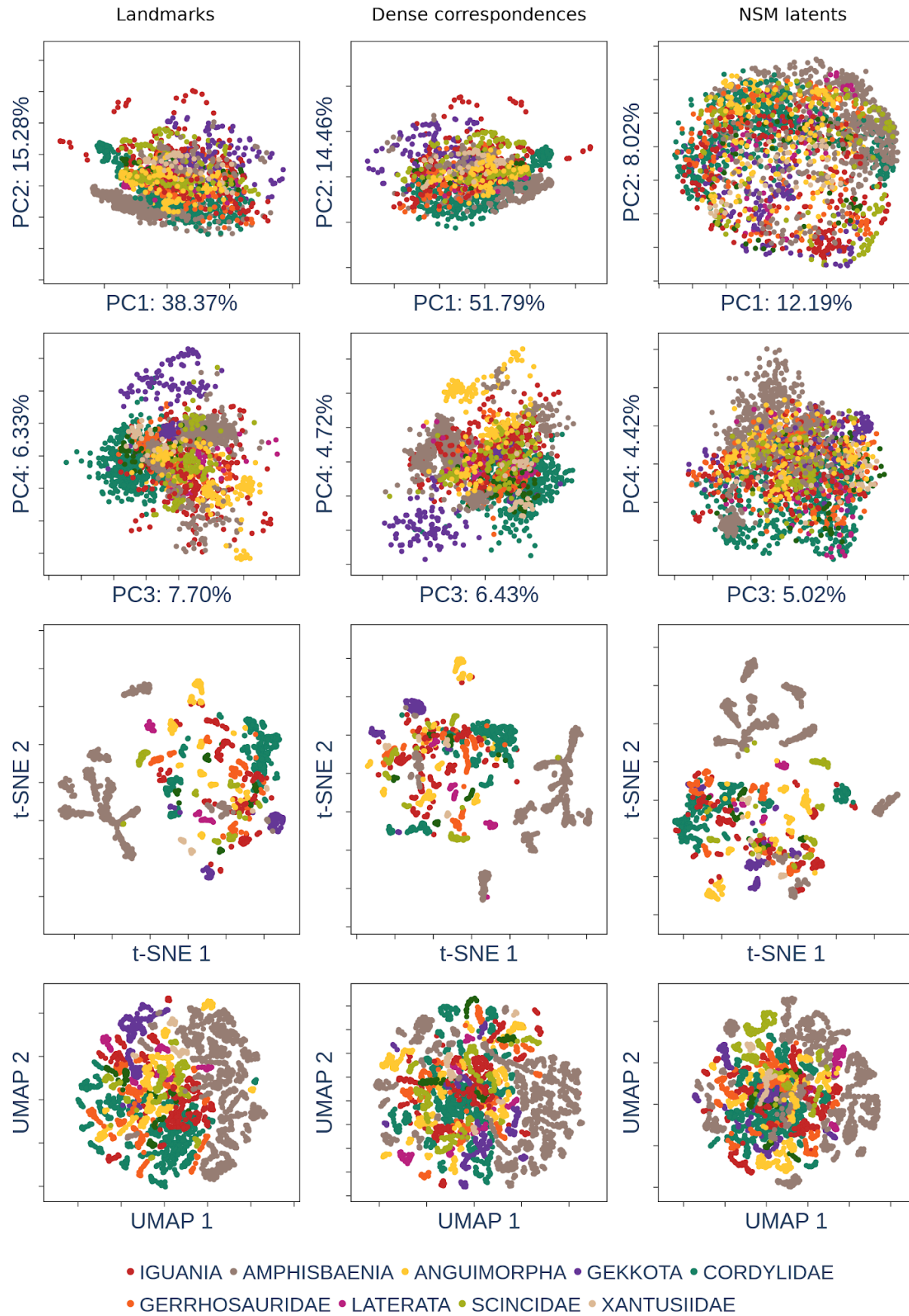


Figure 6. Dimensionality reduction of training-set vertebrae under sparse landmarks, dense correspondences, and NSM latents (columns). Rows show PCA (PC1–PC2, PC3–PC4; variance explained on axis labels), t-SNE, and UMAP. Points are vertebrae colored by broad taxonomic group. Landmark-based representations concentrate more variance in early components than NSM latents (PC1–PC2: 53.6% and 66.2% versus 20.2%; PC3–PC4: 14.0% and 11.2% versus 9.4%) and reveal more taxonomic structure under PCA. t-SNE and UMAP recover similar structure across all three. Amphisbaenia separates in all plots.

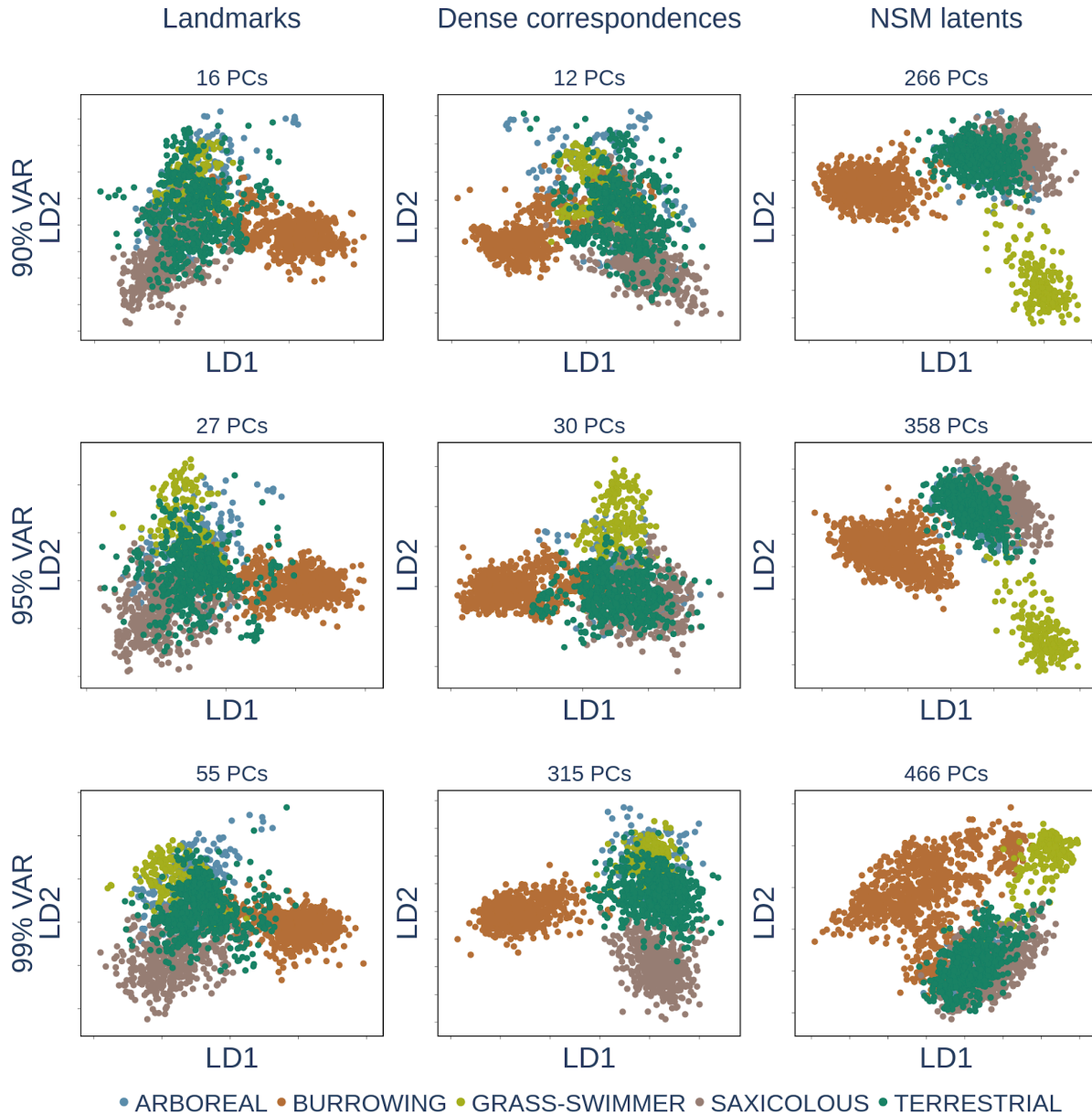


Figure 7. Linear discriminant analysis of life history categories for training-set vertebrae (N=2314), across sparse landmarks, dense correspondences, and NSM latents (columns) at 90%, 95%, and 99% cumulative variance thresholds (rows; number of PCs shown per panel).

NSM separated classes best at 90% and 95%, and was diffuse at 99%, with balanced accuracy stable across thresholds (0.896–0.904; **Table S8**). Dense correspondences separated best at 99%, with balanced accuracy benefitting from increased dimensionality (0.674–0.943).

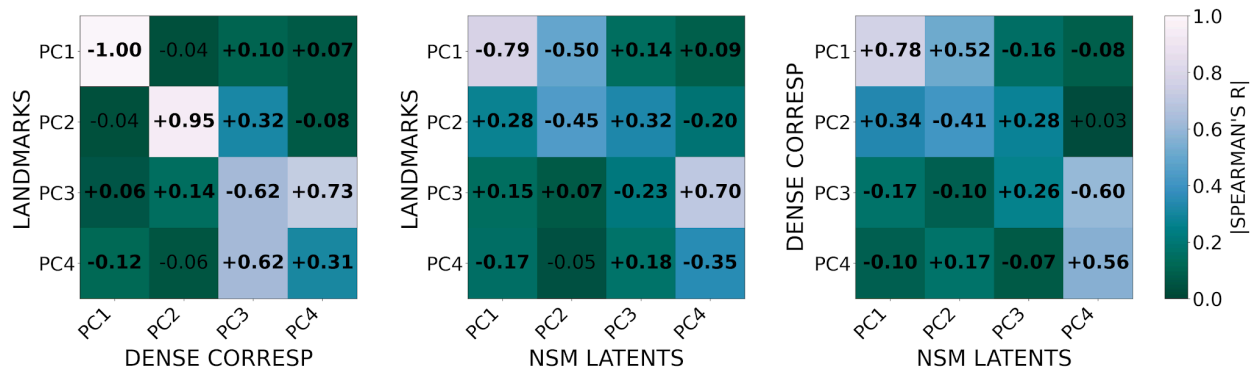


Figure 8. Pairwise Spearman's rank correlation between PC scores of sparse landmarks, dense correspondences, and NSM latents. Sparse landmarks and dense correspondences recover near-identical PC1–PC2 ($\rho = -1.00$, $+0.95$), with PC3 and PC4 transposed. NSM shares PC1 with both representations ($\rho = -0.79$, $+0.78$). Bold values are significant after Bonferroni-correction ($p < 0.05$; 16 comparisons per panel). PC1–PC4 traversal in **Figure 9**.



Figure 9. PC1–PC4 morphospace traversal for sparse landmarks (top) and NSM latents (bottom) in posterior (back; left) and lateral (side; right) views. Columns step from 1.5x the minimum to maximum observed scores. Landmark displacements warped the mean mesh by TPS. NSM latents were displaced from the specimen nearest the mean and decoded. PC1 captures anteroposterior elongation in both representations. PC2 transitions from tall and narrow to short and wide, with additional widening of the postzygapophyses in NSM. NSM PC3

produces implausible geometry (columns 3–4), reflecting reconstruction artifacts in sparsely populated latent space regions. See **Figure 8** and traversal videos (Vert-NSM figures site).

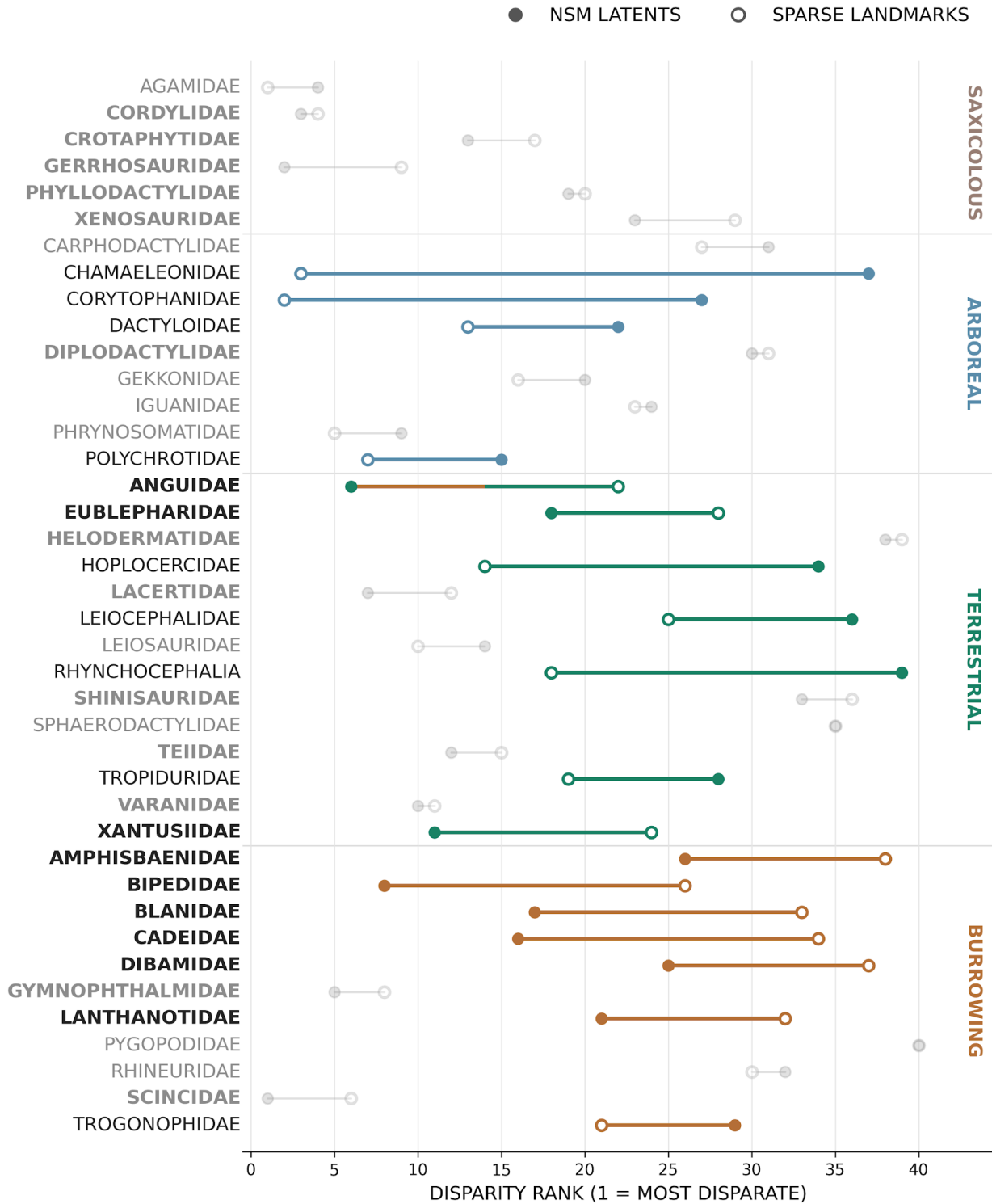


Figure 10. Morphological disparity rank by family for sparse landmarks (open circles) and NSM latents (filled circles), grouped by dominant life history category (rank 1=most disparate).

Fossorial families ranked more disparate under NSM, while arboreal ranked more disparate under landmarks. Differences above 8 ranks (20% of families) shown in color, multiple-category families colored proportionally, and families with greater disparity under NSM in bold. Disparity pooled by life history in **Figure S2**, landmarked vertebrae in **Figure 1**, and raw values in **Table S6**.

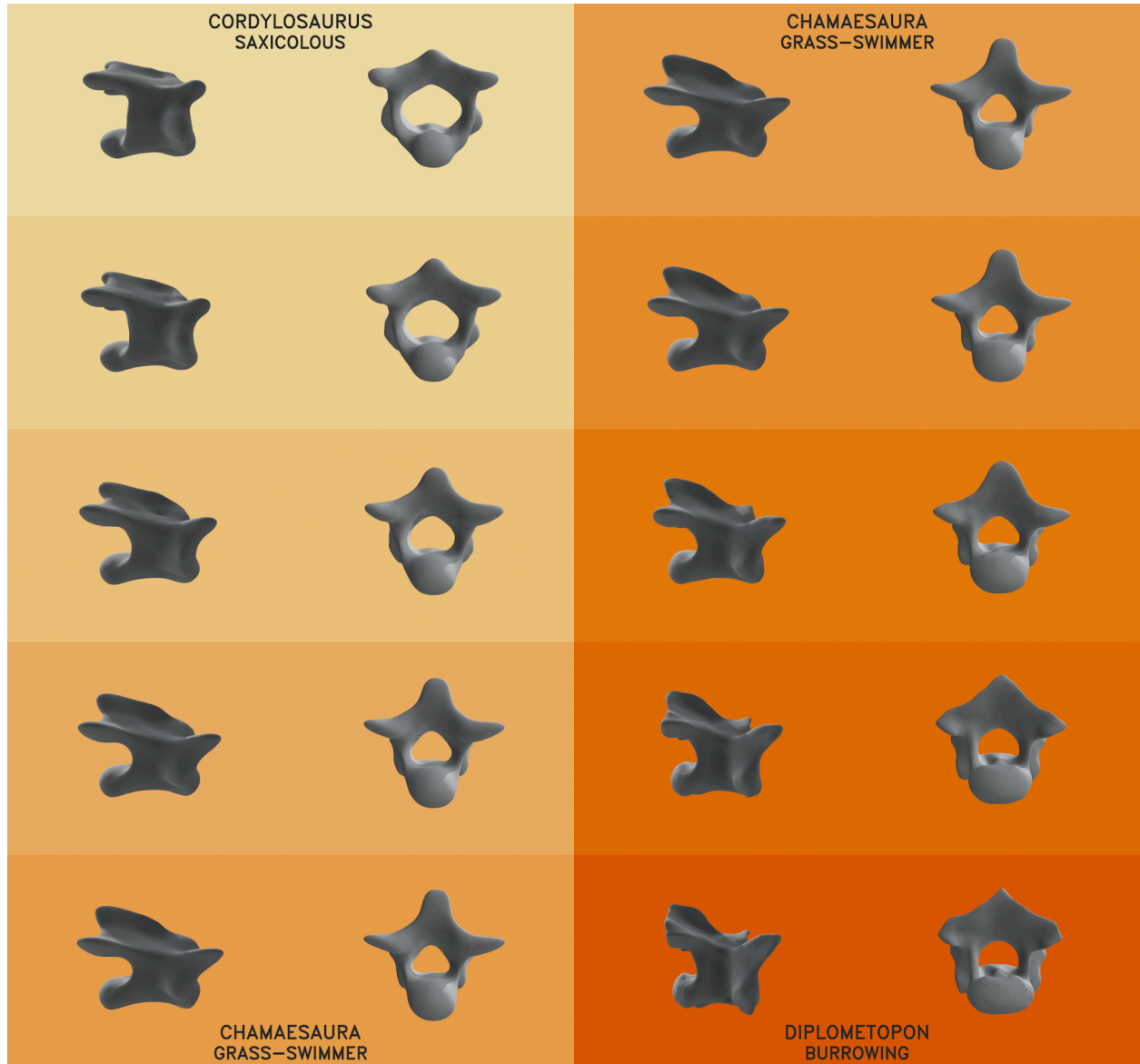


Figure 11. Latent interpolation between thoracic vertebrae: saxicolous (Gerrhosauridae: *Cordylosaurus subtessellatus*) to grass-swimming (Cordylidae: *Chamaesaura aenea*; left) and grass-swimming to fossorial (Trogonophidae: *Diplometopon zarudnyi*; right). Rows are evenly spaced decoded steps shown in lateral (left) and posterior (right) views.

SUPPORTING INFORMATION

File S1. Supplemental Materials and Methods.

File S2. Phylogeny of genera used in the present study, color coded by broad taxon and with family information annotated. Based on Pyron et al. 2013.

File S3. Data splits used for downsample training experiments with 10, 30, 50, 70, 95 specimens.

File S4. Shape completion statistics for all models and data splits.

Table S1. Specimen metadata including taxonomic and life history information used for analysis.

Table S2. Best inference parameters used for models.

Table S3. Classification metrics for all models, data splits, masking levels, and taxonomic or positional levels. Supporting data for **Table 2**.

Table S4. Per-class Top-1 and Top-5 recall for the nine broad taxonomic groups under leave-one-specimen-out (LOSO).

Table S5. Spearman correlations between classification accuracy and sample size. Produced using values in **Table S4**.

Table S6. Morphological disparity by family. Procrustes variance and disparity rank in three representations: sparse landmarks (84 dimensions), NSM latents full (mean-centered, 512 dimensions), and NSM latents reduced (reduced by PCA to 84 components). Supporting data for **Figure 10** and **Figure S2**.

Table S7. Morphological disparity by life history. Procrustes variance and disparity rank in three representations: sparse landmarks (84 dimensions), NSM latents full (mean-centered, 512 dimensions), and NSM latents reduced (reduced by PCA to 84 components). Supporting data for **Figure 10** and **Figure S2**.

Table S8. Life-history trait separability under linear discriminant analysis (LDA). Balanced accuracy (mean $\pm$ SD) across cross-validation, with retained principal components in parentheses (N=2314 vertebrae, 5 classes, chance = 0.2). Supporting data for **Figure 7**

Table S9. Shape completion metrics from downsample training experiments. Supporting data for **Figure S1**.

Table S10. Classification metrics from downsample training experiments. Supporting data for **Figure S1**.

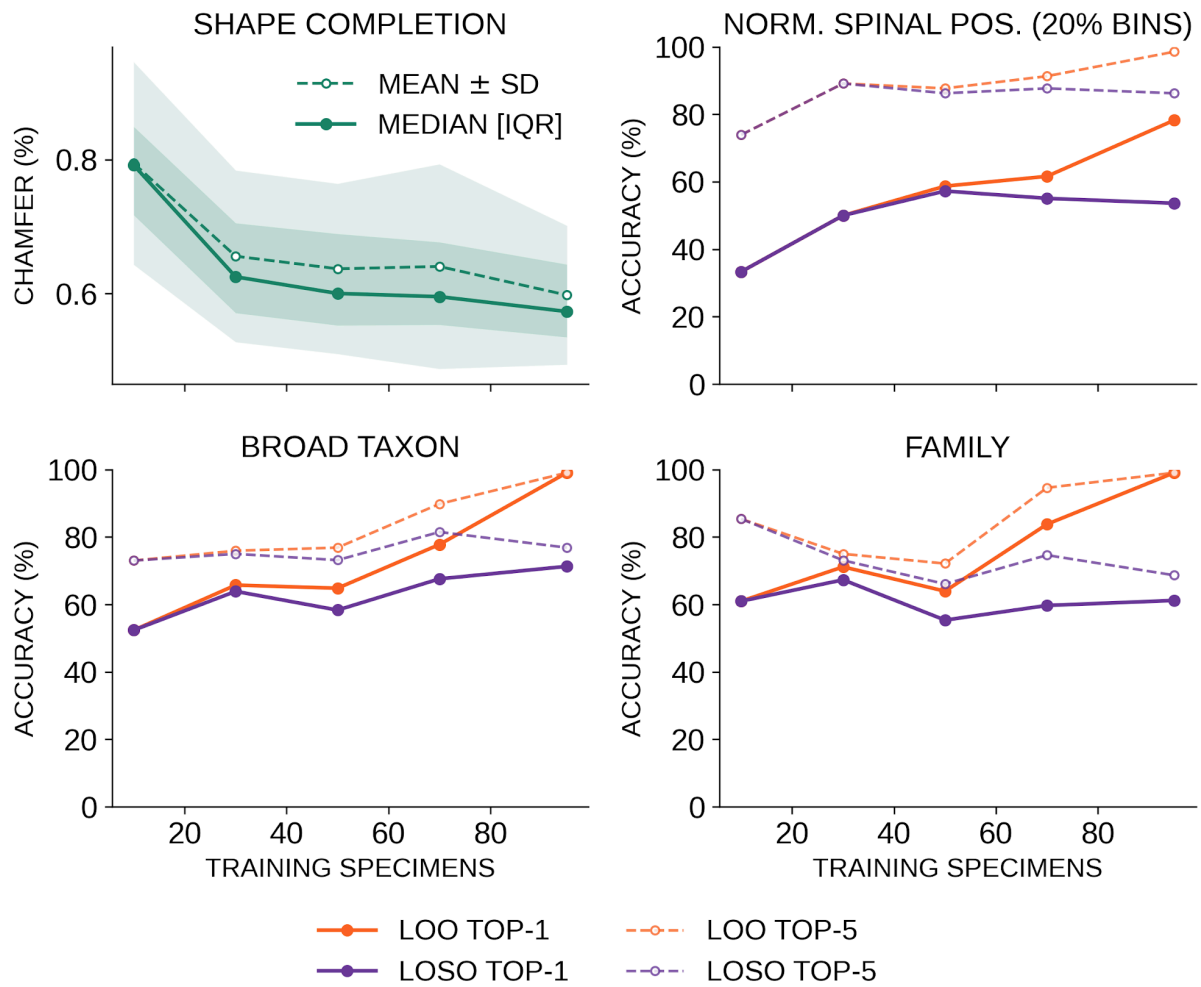


Figure S1. Performance of models trained on 10–95 specimens, all scored on the same 138 held-out vertebrae for shape completion (Chamfer distance) and classification (Top-1/Top-5 accuracy under LOO and LOSO masking). Raw values in **Tables S9–S10**.

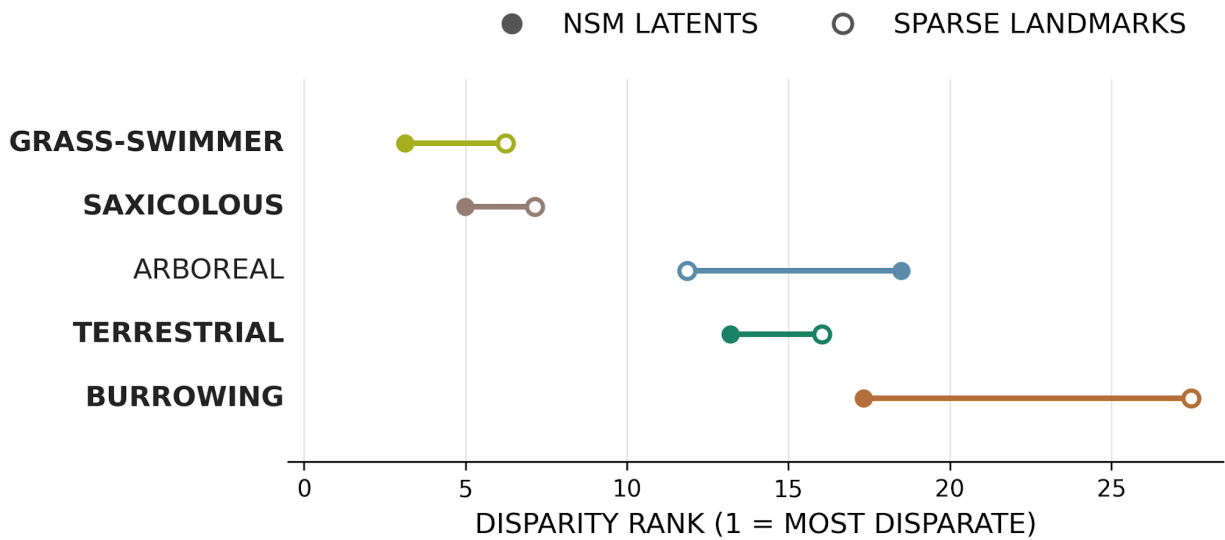


Figure S2. Mean morphological disparity rank by life history category for sparse landmarks (open circles) and NSM latents (filled circles; rank 1=most disparate). Bold labels indicate greater disparity under NSM. Disparity by family in **Figure 10**, landmarked vertebrae in **Figure 1**, and raw values in **Table S7**.

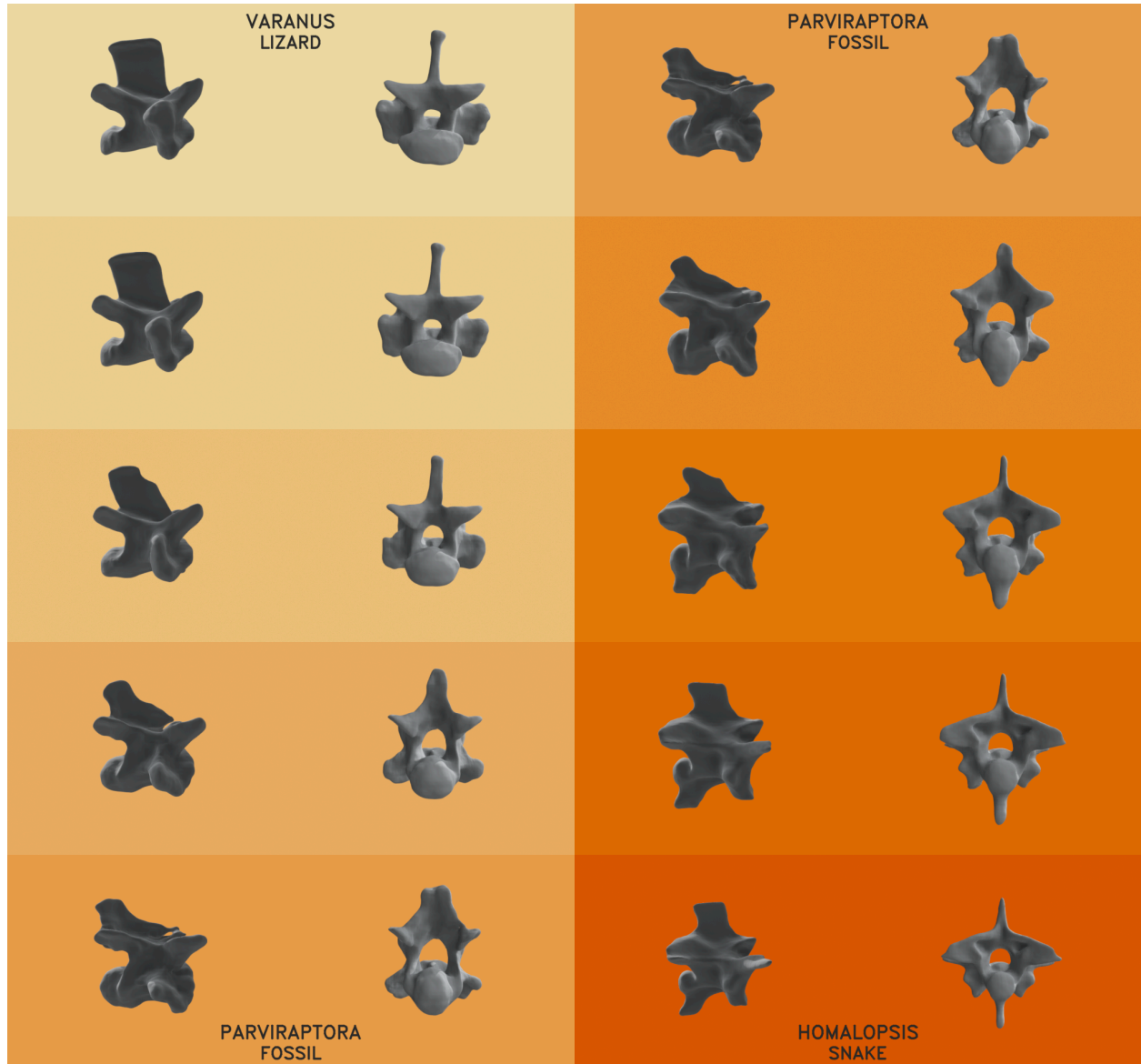


Figure S3. Latent interpolation between vertebrae: lizard (Varanidae: *Varanus komodoensis*) to fossil (Parviraptoridae: *Breugnathair elgolensis*) and fossil to fossorial snake (Homalopsidae: *Homalopsis buccata*). Rows are evenly spaced decoded steps shown in lateral (left) and posterior (right) views.