HYPERBOLIC PROTEIN REPRESENTATION LEARNING WITH SHARED STRUCTURAL MOTIFS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein structures reuse local geometric patterns across diverse proteins, providing a natural basis for sharing structural knowledge. Exploiting this reuse requires representations that capture both multiscale composition and geometric variation among motif occurrences. We introduce MORPH (MOtif Relations for Protein representations in Hyperbolic space), a framework for learning hyperbolic protein representations through shared structural motifs. MORPH uses RamaBPE to construct a reusable vocabulary of variable-length motifs and their ordered multiscale compositions from backbone geometry. Hyperbolic entailment constraints organize compositional relations among both shared motif types and protein-specific occurrences, while explicit type-to-occurrence instantiation loss connects occurrences to their corresponding types. Complementary supervision derived from three-dimensional structures captures continuous geometric relations among occurrences. These signals train a mixture-of-curvature adapter over frozen ESM3 representations, combining Euclidean and Lorentz experts to produce hyperbolic residue representations for downstream prediction. MORPH demonstrates strong downstream transfer across functional site prediction, structural flexibility prediction, and protein-level fold classification. Analysis of motif occurrence representations further reveals composition-related radial organization in hyperbolic space. These findings support shared motif relations and occurrence-level geometric supervision as an effective basis for adapting pretrained protein representations.

1 INTRODUCTION

Protein representation learning has advanced rapidly through large-scale sequence pretraining and the incorporation of three-dimensional structure. Protein language models learn rich structural and functional features from sequence, while structure-aware models such as SaProt and ESM3 expose models directly to protein geometry (Rives et al., 2021; Su et al., 2024; Hayes et al., 2025). Beyond encoding the geometry of individual proteins, recurring structural patterns offer an opportunity to share knowledge across proteins. Local backbone motifs recur in otherwise diverse proteins and combine into larger structural units across multiple scales (Lupas et al., 2001; Mackenzie & Grigoryan, 2017). Their occurrences nevertheless vary in conformation and surrounding sequence and structural context. This reuse motivates representations that connect shared motif identity and multiscale composition while retaining geometric differences among individual occurrences.

Recent substructure-aware methods take steps toward exploiting this shared structural knowledge. GeoBPE identifies variable-length geometric motifs and propagates pretrained features through their merge hierarchy, explicitly encoding how smaller structural units compose into larger motifs (Sun et al., 2026). Magneton trains protein encoders with supervised substructure labels, providing shared categorical targets across occurrences (Calef et al., 2025). These methods motivate further investigation into how shared structural units and their composition relations can jointly supervise representations while preserving variation among their occurrences. A shared motif type provides a common structural reference, but does not fully specify the geometry of each occurrence (Mackenzie et al., 2016). Moreover, composition introduces directed relations from larger motifs to their constituent parts, and reuse allows the same component to participate in multiple compositions. Representing these relations calls for a space that can accommodate multiscale organization while allowing individual occurrences to remain geometrically distinct.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Hyperbolic geometry offers an inductive bias for such relational organization. Its exponential volume growth supports compact representations of branching hierarchies, while entailment cones provide a mechanism for encoding directed relations (Nickel & Kiela, 2017; Ganea et al., 2018b). Related ideas have been used to organize genomic representations through evolutionary structure and mRNA representations through the codon–amino-acid hierarchy (Khan et al., 2025; van Spengler et al., 2025). Here, we use entailment to model structural composition: component representations lie within the cones of their parent compositions, and shared components can participate in multiple such relations. These constraints specify how structural units are related, but do not determine the continuous geometric variation among their occurrences. We therefore combine motif relations with structure-derived similarity supervision, asking whether these complementary signals can organize protein representations in hyperbolic space while preserving occurrence-level conformational variation.

Our Contributions. We introduce **MORPH**, a framework for learning hyperbolic protein representations from shared structural motifs (Figure 1). First, we organize recurring backbone patterns into a shared structural ontology that connects motif types and protein-specific occurrences through multiscale composition and type-to-occurrence instantiation. To construct these units efficiently, we develop **RamaBPE**, a lightweight tokenizer based on Ramachandran-state quantization and ordered BPE merges. Second, we combine hyperbolic entailment supervision over this ontology with continuous geometric supervision among motif occurrences. This formulation captures shared compositional structure while retaining geometric variation among occurrences of the same motif. We integrate these complementary signals through a mixture-of-curvature adapter (He et al., 2025a) over frozen ESM3 features to learn representations for downstream prediction. Finally, experiments show that MORPH transfers effectively across functional site prediction, structural flexibility prediction, and protein-level fold classification. Our analyses further show that efficient RamaBPE construction yields multiscale, geometrically coherent motifs, and that the learned hyperbolic representations exhibit composition-related radial organization.

2 Related Work

Protein Structure Tokenization. Structural tokenization maps continuous protein geometry to discrete codes for sequence-based modeling. Early structural alphabets such as Protein Blocks encode local backbone conformations, while FoldSeek's 3Di alphabet discretizes residue-level tertiary interactions (de Brevern et al., 2000; Van Kempen et al., 2024). More recent methods, including ESM3, ProSST, FoldToken, and ProTokens, learn discrete structure vocabularies with vector quantization objectives for representation learning and generation (Hayes et al., 2025; Li et al., 2024; Gao et al., 2025; Lin et al., 2023). While most operate at fixed residue-level granularity, GeoBPE learns variable-length geometric motifs and a hierarchical merge structure used for representation aggregation and backbone generation (Sun et al., 2026).

Hyperbolic Representation Learning and Applications in Proteins. Due to exponential volume growth induced by negative curvature, hyperbolic representation learning has gained significant traction for modeling hierarchical and power-law structured data across language, graphs, social networks, recommendation systems, and vision tasks (Nickel & Kiela, 2017; Atigh et al., 2022; Chami et al., 2019; Zhang et al., 2021; Chen et al., 2024; Li et al., 2022; Yang et al., 2022). Within deep learning, many works developed fundamental operations for hyperbolic model, while later works developed hyperbolic Transformers, vision models, ResNets, large language models training and fine-tuning, and hyperbolic MoE models (Ganea et al., 2018a; Shimizu et al., 2020; Chen et al., 2021; Yang et al., 2024b; Bdeir et al., 2024; He et al., 2025c;a; Yang et al., 2024a; Mettes et al., 2024; He et al., 2025b). Recent works have also applied hyperbolic to biological and protein applications, leveraging the hierarchical nature of protein evolution and function categorization (Wang et al., 2026; van Spengler et al., 2025)

3 Method

MORPH adapts frozen ESM3 representations through complementary supervision from structural motif relations and occurrence-level geometry (Figure 1). RamaBPE constructs a shared vocabulary of structural motifs and their ordered multiscale compositions. These units support entailment

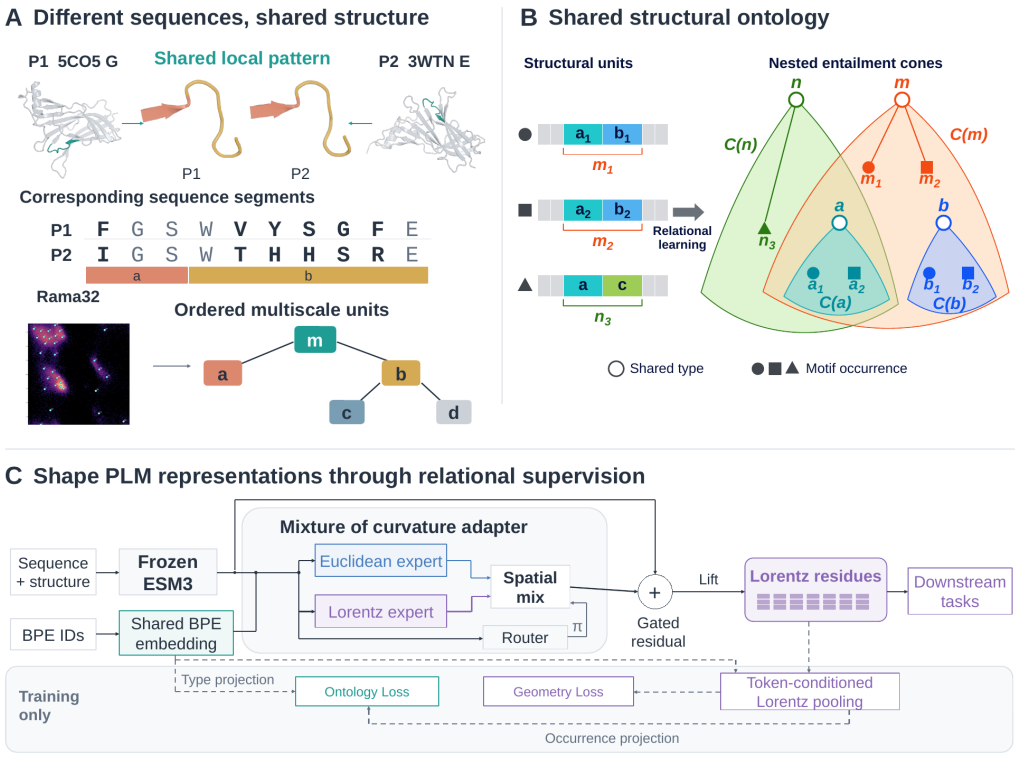

Figure 1: **Overview of MORPH. (A) Structural motif construction.** RamaBPE quantizes backbone torsions into Ramachandran states and applies ordered adjacent-pair merges to construct shared motifs and their multiscale composition hierarchies. **(B) Shared structural ontology.** Entailment cones encode parent-to-component composition and type-to-occurrence instantiation. Shared components can belong to multiple parent compositions. Open nodes denote motif types; filled nodes denote occurrences, with shapes indicating source proteins. The cones schematically illustrate selected relations. **(C) Representation learning.** A mixture-of-curvature adapter integrates frozen ESM3 features and shared motif embeddings through Euclidean and Lorentz experts, producing Lorentz residue representations for downstream prediction. Token-conditioned pooling produces occurrence representations supervised by continuous geometric similarities derived from three-dimensional structures. Separate type and occurrence projections map their respective representations into a common hyperbolic space for entailment supervision.

constraints over motif types and protein-specific occurrences, while structure-derived similarity provides continuous geometric supervision among occurrences. Together, these signals train a mixture-of-curvature adapter to produce hyperbolic residue representations for downstream prediction.

3.1 Structural Motifs and Composition Relations

RamaBPE constructs shared structural motifs by quantizing valid backbone (ϕ, ψ) pairs into $K = 32$ learned Ramachandran states using a periodic angular distance, followed by ordered adjacent-pair BPE within contiguous valid segments. After quantization, pair counting and merging operate entirely on discrete identities, enabling efficient vocabulary construction. Applying the learned merge rules yields a non-overlapping segmentation for each protein, together with a composition forest that records motif identities, residue spans, and ordered children. Details of the structural alphabet appear in Appendix B; construction cost and motif geometry are evaluated in Section 4.3.

Each merge $m = \mathrm{merge}(a, b)$ specifies that an occurrence of m contains ordered occurrences of a and b. Sharing motif types across these protein-specific forests induces a type-level directed acyclic graph, since the same component can participate in multiple parent compositions. These relations motivate a hyperbolic representation space for modeling multiscale structural composition. Specif-

ically, MORPH treats parent-to-component composition as *structural entailment*: component representations are constrained to lie within the entailment cones of their parent compositions (Desai et al., 2023). A shared component can therefore be constrained by multiple parent cones, preserving the relational structure induced by motif reuse. The following sections describe the mixed-curvature adapter and the objectives that encode these relations alongside occurrence-level geometric variation.

3.2 Hyperbolic Geometry and Notation

In this section, we define several objects in hyperbolic space and their associated notation. An n-dimensional hyperbolic space using the Lorentz hyperboloid formulation, denoted as $\mathbb{L}^{\kappa,n}$, is a Riemannian manifold with constant negative curvature $\kappa < 0$. Each point $x \in \mathbb{L}^{\kappa,n}$ can be parametrized as $[x_t, x_s]^\top$, where $x_t \in \mathbb{R}$ and $\mathbf{x}_s \in \mathbb{R}^n$ are the *time* dimension and *space* dimensions, respectively. For points $x, y \in \mathbb{L}^{\kappa,n}$, their *Lorentz inner product* is given by $\langle x, y\rangle_{\mathcal{L}} = -x_t y_t + x_s^\top y_s$ which defines the Lorentz hyperbolic space as $\mathbb{L}^{\kappa,n} = \{x \in \mathbb{R}^{n+1} : \langle x, x\rangle_{\mathcal{L}} = 1/\kappa,\ x_t > 0\}$. The *origin* $\mathbf{o} \in \mathbb{L}^{\kappa,n}$ is $[\sqrt{-1/\kappa}, 0, \ldots, 0]^T$. The *squared Lorentzian distance* on the space is given as $d_L(x, y)^2 = 2/\kappa - 2\langle x, y\rangle_{\mathcal{L}}$.

At each point $x \in \mathbb{L}^{\kappa,n}$, the associated *tangent space* is set of points orthogonal to $\mathbf{x}$, defined as $\mathcal{T}_x\mathbb{L}^{\kappa,n} = \{y \in \mathbb{R}^{n+1} : \langle x, y\rangle_{\mathcal{L}} = 0\}$, serving as vector space that locally approximates behaviors near x. Points in the tangent space can be mapped to the hyperboloid using the *exponential map* $\exp_x^\kappa : \mathcal{T}_x\mathbb{L}^{\kappa,n} \to \mathbb{L}^{\kappa,n}$, with inverse given as the *logarithmic map* $\log_x^\kappa : \mathbb{L}^{\kappa,n} \to \mathcal{T}_x\mathbb{L}^{\kappa,n}$.

3.3 Mixture-of-Curvature Protein Representation Learning

Given the shared multiscale structural units defined above, we next incorporate them into the contextual representations produced by a frozen ESM3 backbone. We propose a mixture-of-curvature adapter to combine these two sources of information. The frozen ESM3 representations are learned in Euclidean space, whereas the RamaBPE motifs are associated with multiscale hierarchical composition relations that are naturally modeled in hyperbolic space. By combining Euclidean and Lorentz experts, MORPH can preserve the contextual information encoded by ESM3 while simultaneously learning representations that capture the hierarchical structure induced by the shared motif vocabulary.

For a protein of length L, we extract the ESM3 representation after the final LayerNorm layer $H_{ESM3} \in \mathbb{R}^{L\times 1536}$. Each residue i is additionally associated with the shared BPE embedding $e_{M(i)}$ of the motif type assigned to that residue. For each residue i, we produce the projected embedding combining both contextual and motif information: $b_i = \mathrm{LN}\left(\mathrm{GELU}\left(W_{fuse}\left[W_{down}\,\mathrm{LN}(H_{ESM3,i}), e_{M(i)}\right]\right)\right)$. The projected embedding is then passed to a mixture-of-curvature adapter containing a Euclidean expert E_0 and a Lorentz expert E_1. The Euclidean expert is a SwiGLU feedforward neural network (FFN) (Grattafiori et al., 2024), whereas the Lorentz expert is a hyperbolic SwiGLU FFN (He et al., 2025a). The output of the experts are then given as

$$x_i = E_0(b_i) \in \mathbb{R}^{256}; \qquad y_{H,i} = E_1(\exp_{\mathbf{o}}^\kappa(\mathrm{Cap}(b_i))) \in \mathbb{L}^{\kappa,256}$$

where the smooth tangent cap Cap is given as $x \mapsto x\frac{M\tanh(\|x\|/M)}{\|x\|}$ with chosen $M = 8$ to maintain numerical stability of the exponential mapping (Desai et al., 2023).

A residue-level router, formulated as a two-layer MLP, is then used with softmax to produce routing weights among the two experts $\pi_i^{(0)}, \pi_i^{(1)}$ where $\pi_i^{(0)} + \pi_i^{(1)} = 1$. The router combines the Euclidean expert output with the space-like dimension of the Lorentz expert output, giving $m_{s,i} = \pi_i^{(0)} x_i + \pi_i^{(1)}(y_{H,i})_s$. The final output of the mixture-of-curvature adapter is given as the gated update

$$H_{L,i} = \mathrm{Lift}_\kappa\left(H_{ESM3,i} + g_i W_{up}(m_{s,i})\right) \in \mathbb{L}^{\kappa,1536}; \qquad g_i = \sigma\left(w_g^\top H_{ESM3,i} + b_g\right) \in (0, 1)$$

where σ is sigmoid and $\mathrm{Lift}_\kappa : \mathbb{R}^d \to \mathbb{L}^{\kappa,d}$ is the projection to hyperbolic space given by $u \mapsto \left[\sqrt{-1/\kappa + \|u\|^2}, u\right]^\top$. Thus the module can be viewed as adapting frozen ESM3 embeddings into the space dimension of a vector on the Lorentz hyperboloid.

Expert Load Balancing. We use an expert balancing loss to prevent expert collapse. For the set $\mathcal{R}$ of all valid residues, define the mean expert usage as $\bar{\pi} = \frac{1}{|\mathcal{R}|}\sum_{i\in\mathcal{R}}\pi_i$. The expert balancing is then given as $L_{bal} = \|\bar{\pi} - \left(\frac{1}{2}, \frac{1}{2}\right)\|_2^2$.

3.4 Shared Geometric Structure of Motif Types and Occurrences

A motif type m is a shared structural concept in the global BPE vocabulary, whereas a motif occurrence $v = (X, S_v, m_v)$ denotes one concrete instance of type m_v in protein X over the contiguous residue span S_v. Different occurrences of the same motif type share the same discrete BPE identity but may differ in their exact backbone geometry and protein context. Since the final representation H_L is defined at the residue level, we construct a representation for each motif occurrence v using token-conditioned attention pooling. For residue $i \in S_v$, the attention logit is computed from its space dimension $(H_{L,i})_s$ and the shared BPE embedding e_{m_v}:

$$\ell_{v,i} = w_a^\top \tanh\left(W_h(H_{L,i})_s + W_e e_{m_v} + b_a\right); \qquad \alpha_{v,i} = \frac{\exp(\ell_{v,i})}{\sum_{j\in S_v}\exp(\ell_{v,j})}.$$

The hyperbolic occurrence representation is then constructed using the Lorentzian centroid (Law et al., 2019):

$$q_v = \frac{1}{\sqrt{-\kappa}} \frac{\sum_{i\in S_v}\alpha_{v,i}H_{L,i}}{\sqrt{-\left\langle\sum_{i\in S_v}\alpha_{v,i}H_{L,i}, \sum_{i\in S_v}\alpha_{v,i}H_{L,i}\right\rangle_{\mathcal{L}}}} \in \mathbb{L}^{\kappa,d}.$$

Occurrence Geometry Loss. The discrete motif type captures a recurring structural pattern but does not require two occurrences of the same type to have identical backbone geometry. For each anchor occurrence v, let $\mathcal{C}(v)$ denote its candidate occurrences and $\mathrm{sim}_T(v,u)$ denote the structure-derived teacher similarity between v and $u \in \mathcal{C}(v)$ (see Appendix C for details). The teacher and student distributions are given as $q_T(u \mid v) = \mathrm{softmax}_{u\in\mathcal{C}(v)}\left(\mathrm{sim}_T(v,u)/T_t\right)$ and $p_S(u \mid v) = \mathrm{softmax}_{u\in\mathcal{C}(v)}\left(-d_L^2(q_v,q_u)/T_s\right)$ respectively. The occurrence geometry loss is then defined as $L_{geo} = \mathbb{E}_v\,\mathrm{KL}\left(q_T(\cdot \mid v)\|p_S(\cdot \mid v)\right)$. This objective preserves continuous structural relationships between concrete motif occurrences without requiring occurrences of the same motif type to collapse to the same point.

Ontology Losses. However, L_{geo} does not explicitly encode shared motif type identity or the compositional relations among motif types and occurrences. To model these relations, we construct a lower-dimensional Lorentz ontology space containing both shared motif types and concrete motif occurrences. Each occurrence representation q_v is projected and lifted to hyperbolic space:

$$\tilde{q}_v = \mathrm{Lift}_\kappa(W_{occ}q_v + b_{occ}) \in \mathbb{L}^{\kappa,d_{ont}}.$$

Each shared motif embedding is also mapped to a type point using a two-layer projector:

$$u_m = W_{t,2}\,\mathrm{GELU}(W_{t,1}e_m + b_{t,1}) + b_{t,2}; \qquad t_m = \mathrm{Lift}_\kappa(u_m) \in \mathbb{L}^{\kappa,d_{ont}}.$$

Here $\tilde{q}_v$ represents a concrete occurrence in the ontology space, while t_m represents a shared motif type. We then model the relationships in the structural ontology using hyperbolic entailment cones (Desai et al., 2023): for $a, b \in \mathbb{L}^{\kappa,n}$, the hyperbolic cone at a, denoted as $Cone(a)$, is defined by its half-aperture as $\mathrm{aper}(a) = \sin^{-1}\left(2\gamma\left(\sqrt{-\frac{1}{\kappa}}\|a_s\|\right)^{-1}\right)$; the exterior angle, which measures whether b lies inside the cone of a, is given as $\mathrm{ext}(a,b) = \cos^{-1}\left(\left(b_t - \frac{a_t}{\kappa}\langle a,b\rangle_{\mathcal{L}}\right)\left(\|a_s\|\sqrt{\left(\frac{-1}{\kappa}\langle a,b\rangle\right)^2 - 1}\right)^{-1}\right)$. The *hyperbolic entailment loss* penalizes when b should be inside the cone of a but isn't, defined as $L_{ent}(a,b) = \max(0, \mathrm{ext}(a,b) - \mathrm{aper}(a))$.

For a type-level BPE merge $p = \mathrm{merge}(l,r)$, where p is the composite parent and l, r are its component types, we require $t_l \in Cone(t_p), t_r \in Cone(t_p)$. For a concrete occurrence merge $p_i = \mathrm{merge}(l_i, r_i)$ within a protein, we similarly require $\tilde{q}_{l_i} \in Cone(\tilde{q}_{p_i}), \tilde{q}_{r_i} \in Cone(\tilde{q}_{p_i})$. These relationships give the type hierarchy loss and the occurrence hierarchy loss:

$$L_{type} = L_{ent}(t_p, t_l) + L_{ent}(t_p, t_r); \qquad L_{occ} = L_{ent}(\tilde{q}_{p_i}, \tilde{q}_{l_i}) + L_{ent}(\tilde{q}_{p_i}, \tilde{q}_{r_i}).$$

For occurrence v with motif type m_v, we further require $\tilde{q}_v \in Cone(t_{m_v})$, giving the type-occurrence instantiation loss $L_{inst} = L_{ent}(t_{m_v}, \tilde{q}_v)$. This loss deliberately does not force $\tilde{q}_v = t_{m_v}$, allowing different occurrences of the same motif type to retain variation in backbone geometry and protein context.

Finally, for each direct occurrence composition edge (p_i, c_i), we additionally impose radial ordering such that the child occurrence lies farther from the Lorentz origin than its parent:

$$L_{radial} = \mathbb{E}_{(p_i,c_i)} \left[\max \left(0, m_r + d_L^2(\mathbf{o}, \tilde{q}_{p_i}) - d_L^2(\mathbf{o}, \tilde{q}_{c_i})\right)\right],$$

where $r(q)$ denotes hyperbolic distance to the origin. We additionally use a minimum-radius regularizer $L_{minrad} = \mathbb{E}_x \max\left(0, r_{min} - d_L^2(\mathbf{o}, x)\right)$. These losses prevent entailment learning collapse, where all embeddings sitting close the origin and lie on the same hyperbolic ring can still result in small entailment loss. The complete ontology objective is therefore

$$L_{ont} = w_t L_{type} + w_o L_{occ} + w_i L_{inst} + w_{rad} L_{radial} + w_{min} L_{minrad}.$$

Overall Loss. The overall loss of MORPH is given as $\lambda_{bal} L_{bal} + \lambda_{geo} L_{geo} + \lambda_{ont} L_{ont}$, encouraging balanced expert utilization, preservation of occurrence-level structural geometry, and hierarchical organization of shared motif types and their concrete occurrences.

4 Experiments and Results

The evaluation examines MORPH from three perspectives: downstream transfer to protein structure and function prediction, the structural properties and construction cost of the RamaBPE vocabulary, and composition-related organization in the learned motif representations.

4.1 Experimental Setup

Tasks and metrics. MORPH is evaluated on the downstream transfer benchmark used by GeoBPE (Sun et al., 2026), based on StructTokenBench (Yuan et al., 2025), following the same task definitions and data splits. The evaluation comprises 12 tasks and 24 test splits. Eight residue-level classification tasks—BindBio, BindInt, BindShake, CatBio, CatInt, Con, Ept, and Rep— are evaluated with AUROC. Three residue-level flexibility tasks—FlexBFactor, FlexNEQ, and FlexRMSF—are evaluated with Spearman correlation. Homo evaluates protein-level fold classification using macro-F1. The evaluation retains the benchmark's Fold and Superfamily splits, the original test split for BindShake, and the Fold, Superfamily, and Family holdouts for Homo.

Reference methods. MORPH is compared with ProteinMPNN, MIF, FoldSeek, ProTokens, ESM3, VanillaVQ, AminoAseed, and GeoBPE-Transfer using the reference scores reported by GeoBPE (Sun et al., 2026).

4.2 Downstream Transfer Performance

Table 1 reports performance across all downstream tasks. MORPH achieves the highest score among the compared methods on 15 of 24 test splits and exceeds GeoBPE-Transfer on 18. Improvements over GeoBPE-Transfer cover all protein-level fold classification and structural flexibility splits, while functional site prediction shows greater variation across tasks and holdout settings.

Functional site prediction. MORPH achieves the highest AUROC on eight of the 15 classification splits. On Ept, it reaches 0.7973 and 0.9163 on the Fold and Superfamily splits, respectively, compared with 0.6478 and 0.7706 for GeoBPE-Transfer. It also obtains the highest scores on the Superfamily splits of BindInt, CatBio, CatInt, Con, Rep, and BindShake. GeoBPE-Transfer remains stronger on both BindBio splits and on the Fold splits of BindInt, CatBio, CatInt, and Con. These results demonstrate strong functional-site transfer for MORPH, with particularly consistent gains on the more challenging Superfamily splits.

Structural flexibility. MORPH exceeds GeoBPE-Transfer on all six flexibility splits and achieves the highest Spearman correlation on both FlexBFactor and FlexNEQ splits. On FlexBFactor, its

Table 1: Downstream transfer performance. Reference scores are taken from GeoBPE (Sun et al., 2026), with GeoBPE denoting GeoBPE-Transfer. Bold entries with gray backgrounds indicate the highest score in each row. Higher is better for all metrics.

Task	Split	Reference methods								Our method
		Protein MPNN	MIF	FoldSeek	Pro Tokens	ESM3	Vanilla VQ	Amino Aseed	GeoBPE	MORPH
Functional Site Prediction (AUROC)										
BindBio	Fold	0.7842	0.8579	0.3237	0.5847	0.6284	0.6202	0.6573	**0.9494**	0.9087
	SupFam	0.8100	0.8727	0.5244	0.6047	0.6522	0.6292	0.6830	**0.9594**	0.9331
BindInt	Fold	0.5183	0.5038	0.5318	0.4466	0.4430	0.4725	0.4711	**0.5919**	0.5773
	SupFam	0.9400	0.9456	0.4626	0.8605	0.9077	0.8671	0.9053	0.9131	**0.9777**
BindShake	Org/Test	0.7552	0.7990	0.5343	0.5910	0.6610	0.6704	0.6961	0.8773	**0.9069**
CatBio	Fold	0.8249	0.8585	0.5633	0.6768	0.6533	0.6758	0.6595	**0.9501**	0.9246
	SupFam	0.9319	0.9697	0.5378	0.6405	0.7465	0.7092	0.8759	0.9590	**0.9924**
CatInt	Fold	0.6105	0.5962	0.5343	0.5816	0.6109	0.5889	0.6219	**0.6621**	0.6459
	SupFam	0.9340	0.9649	0.5141	0.8385	0.8982	0.8500	0.9191	0.8865	**0.9821**
Con	Fold	0.5718	0.5843	0.4920	0.5720	0.5522	0.5698	0.5723	**0.7196**	0.6551
	SupFam	0.8468	0.9266	0.5131	0.7064	0.8053	0.7460	0.8660	0.8484	**0.9816**
Ept	Fold	0.6284	0.6878	0.5456	0.5249	0.6369	0.5928	0.6216	0.6478	**0.7973**
	SupFam	0.6484	0.8298	0.5053	0.6192	0.6197	0.6724	0.7202	0.7706	**0.9163**
Rep	Fold	**0.7763**	0.7453	0.4771	0.5320	0.7470	0.7599	0.7497	0.5644	0.7598
	SupFam	0.8071	0.8311	0.5254	0.7725	0.8236	0.8209	0.8457	0.7298	**0.8678**
Structural Flexibility Prediction (Spearman's ρ)										
FlexBFactor	Fold	0.3188	0.3460	0.0417	0.0667	0.2360	0.2232	0.2130	0.3728	**0.4382**
	SupFam	0.3456	0.3523	0.0699	0.0547	0.2580	0.2373	0.2176	0.3561	**0.4295**
FlexNEQ	Fold	0.6969	0.6532	0.0571	0.1298	0.4505	0.3595	0.4964	0.5665	**0.6996**
	SupFam	0.6869	0.6482	0.0266	0.1051	0.3545	0.3561	0.5015	0.5398	**0.6921**
FlexRMSF	Fold	**0.6237**	0.5960	0.1535	0.1381	0.4453	0.4422	0.4463	0.4089	0.6143
	SupFam	**0.5924**	0.5680	0.1199	0.0762	0.3908	0.3898	0.4099	0.4717	0.5801
Protein-level Fold Classification (Macro-F1)										
Homo	Fold	0.2566	0.2256	0.1157	0.0584	0.3002	0.1817	0.2987	0.2360	**0.3355**
	SupFam	0.3083	0.3386	0.0467	0.0617	0.2489	0.2210	0.3838	0.4728	**0.9139**
	Family	0.6333	0.7422	0.1534	0.1833	0.5442	0.4718	0.6978	0.8575	**0.9556**

SupFam denotes Superfamily. Homo uses the corresponding Fold, Superfamily, and Family holdout splits. Org/Test denotes the original BindShake test split. Scores are rounded to four decimal places; best entries are identified before rounding.

scores are 0.4382 and 0.4295, compared with 0.3728 and 0.3561 for GeoBPE-Transfer, achieving up to 20.6% improvement. These results demonstrate strong and consistent transfer across residue-level flexibility tasks.

Protein-level fold classification. MORPH achieves the highest macro-F1 on all three Homo holdouts: 0.3355 on Fold, 0.9139 on Superfamily, and 0.9556 on Family. The corresponding GeoBPE-Transfer scores are 0.2360, 0.4728, and 0.8575. The largest absolute improvement occurs on the Superfamily holdout. These results demonstrate particularly strong protein-level transfer, with MORPH achieving the best performance across all three fold-classification holdouts and especially large gains on the Superfamily and Family splits.

4.3 Structural Vocabulary Analysis

RamaBPE supplies the shared motif types and composition relations used to train MORPH. The vocabulary analysis examines whether its compact discrete alphabet supports efficient construction of multiscale motifs with consistent local geometry.

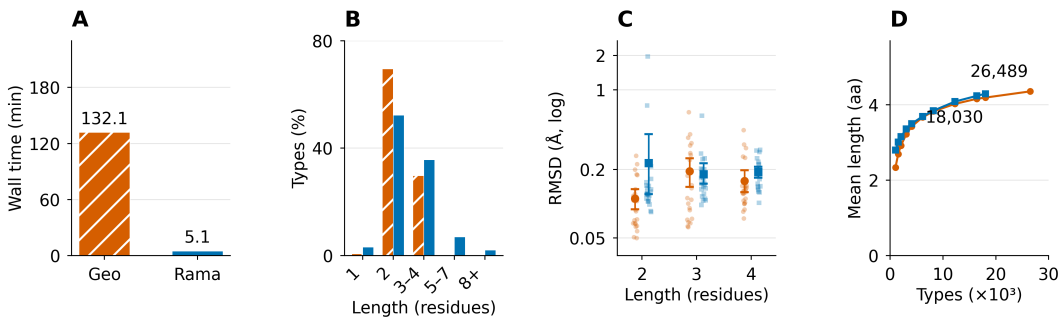

Figure 2: **Construction cost and structural vocabulary properties. (A)** Recorded construction times: RamaBPE alphabet fitting and pair-BPE total 5.1 minutes, excluding full-corpus encoding; GeoBPE takes 132.1 minutes through step 400 on 320 valid structures. The runs differ in workload and timing scope and do not constitute a controlled speedup comparison. **(B)** Type-weighted motif length distributions at $V = 1{,}024$. **(C)** Within-type backbone RMSD on the construction corpus. Each faint point is one type's mean over ten occurrence pairs from five distinct structure IDs. Solid markers and error bars indicate means and 95% bootstrap confidence intervals over 24 types per method and length; all sampled outliers are retained. **(D)** Type-weighted mean lengths across vocabulary prefixes, with final sizes annotated. In B-D, the GeoBPE comparator comprises step-500 geometric motifs followed by pair-BPE, excluding 72 edge symbols.

Construction cost and multiscale composition. Alphabet fitting on 330 structures takes 150.9 seconds, followed by 156.7 seconds for pair-BPE on the pre-encoded corpus of 33,988 structures, yielding 18,030 motif types. Figure 2A presents these costs alongside GeoBPE timings.

At a budget of 1,024 motif types, RamaBPE has a mean type length of 2.80 residues and a maximum of 12, compared with 2.34 and 4 for GeoBPE (Figure 2B). Mean length difference narrows as the vocabulary grows, reaching 4.29 v.s. 4.19 residues at the common budget of 18,030 types (Figure 2D). The final RamaBPE segmentation contains 0.327 tokens per encoded residue on the construction corpus, showing that multi-residue motifs are used in the resulting protein segmentations.

Within-type geometric consistency. For each method and exact motif length of two, three, or four residues, 24 types are sampled, with five occurrences per type drawn from distinct structure IDs in the construction corpus. After rigid backbone alignment, the mean within-type RMSDs are 0.227, 0.182, and 0.190 Å for RamaBPE, compared with 0.111, 0.193, and 0.159 Å for GeoBPE (Figure 2C). The length-three and length-four averages are numerically close, while RamaBPE exhibits greater variation at length two, including two high-variation types retained in the analysis. These observations show local geometric consistency among the sampled motif occurrences, together with residual variation within shared types. This combination motivates the use of both shared motif relations and continuous occurrence-level geometric supervision.

4.4 Organization of Motif Representations

We perform representation analysis to examine whether motif composition is reflected in the organization of MORPH's occurrence embeddings. Figure 3 shows representations before ontology projection, grouped by the shortest and longest span-length quartiles from 40 validation proteins. Each group contains 3,555 occurrences: the shortest consists of single-residue units, while the longest spans 2-15 residues. We compare MORPH against an Euclidean adaptation of MORPH using two Euclidean experts that output a final Euclidean residue embedding. The model receives the same RamaBPE motif identities, composition relations, and structure-derived similarities as MORPH, but replaces $L_{inst}, L_{occ}, L_{type}$ with direction supervision on parent-children relations and L_{radial} with Euclidean type-level radial loss. Full details and downstream results are provided in Appendix A.1. For MORPH, we plot the Lorentzian radius of the embeddings. Following Desai et al. (2023), we plot the radius from the learned embedding center for the Euclidean model.

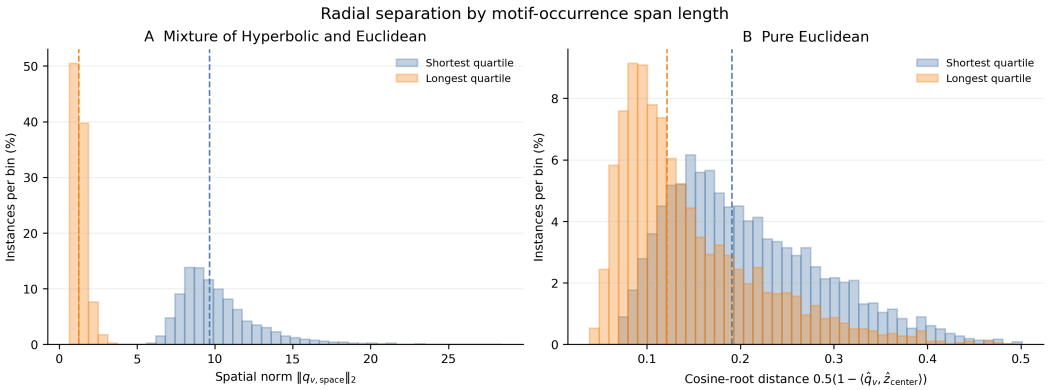

Figure 3: **Occurrence organization by span-length quartile.** From 14,222 motif occurrences in 40 validation proteins, the shortest and longest quartiles are selected, with 3,555 occurrences each and identical groups for both models. The shortest group contains single-residue units; the longest contains motifs spanning 2-15 residues. **(A)** MORPH: the horizontal axis is the spatial-coordinate norm $\|q_{v,\text{space}}\|_2$. **(B)** Pure-Euclidean adapter: the horizontal axis is $\frac{1}{2}(1-\langle\hat{q}_v,\hat{z}_{\text{center}}\rangle)$, where $\hat{q}_v$ is a unit-normalized occurrence embedding and $\hat{z}_{\text{center}}$ is the normalized mean of unit-normalized occurrence embeddings. Representations are measured before ontology projection. Histograms are independently normalized within each group, and dashed lines indicate medians. The horizontal axes use different measures; comparisons concern within-panel group separation. The Euclidean configuration also uses different training losses (Appendix A).

MORPH exhibits clear separation between these groups, with single-residue units having larger space dimension norms than composite motifs. This pattern is consistent with the intended radial ordering of components and compositions. Grouping by span length characterizes composition-related organization, but does not directly test parent-child entailment or cone inclusion. Notably, the Euclidean model was also trained a radius loss similar to L_{radial}. Nevertheless, while the embedding distribution of the Euclidean model show some patterns of separation based on span-length, the effect is far more pronounced in MORPH, demonstrating the effectiveness of hyperbolic embeddings in modeling span-length-induced hierarchies.

Ablation Comparison Against the Euclidean Model. Despite being trained with Euclidean objectives designed to approximate MORPH's relational and geometric supervision, the Euclidean model underperforms MORPH on 21 of 24 downstream splits (Table 2). This indicates that analogous supervision alone is insufficient to recover MORPH's performance, showing the benefit of MORPH's formulations and properties that are inherent to hyperbolic space, such as entailment cones.

5 Conclusion

We introduced MORPH, a framework for learning hyperbolic protein representations from shared structural motifs. RamaBPE constructs a compact vocabulary of variable-length motifs and their multiscale compositions. MORPH combines entailment supervision over shared motif types and protein-specific occurrences with continuous geometric supervision from a structural teacher, training a mixture-of-curvature adapter over frozen ESM3 representations. Experiments demonstrate competitive downstream transfer, while vocabulary and representation analyses reveal local geometric consistency and composition-related radial organization, respectively. These findings support shared motif relations and occurrence-level geometry as complementary sources of supervision for protein representation learning. The current Euclidean comparison varies both geometry and training objectives; future work will disentangle their contributions under matched supervision and evaluate transfer to broader protein settings.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks: language polishing during paper writing, formatting plots for better presentation. We did not use AI for any of the remaining required disclosure of AI uses. We checked all writing and figures manually. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *CVPR*, pp. 4453–4462, 2022.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR*, 2024.

Robert Calef, Arthur Liang, Manolis Kellis, and Marinka Zitnik. Greater than the sum of its parts: Building substructure into protein encoding models. *arXiv preprint arXiv:2512.18114*, 2025. URL https://arxiv.org/abs/2512.18114.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *NeurIPS*, pp. 4868–4879, 2019.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. *arXiv preprint arXiv:2105.14686*, 2021.

Weize Chen, Xu Han, Yankai Lin, Kaichen He, Ruobing Xie, Jie Zhou, and Zhiyuan Liu. Hyperbolic pre-trained language model. *IEEE TASLP*, 32, 2024.

Alexandre G. de Brevern, Catherine Etchebest, and Serge Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function, and Genetics*, 41(3):271–287, 2000. doi: 10.1002/1097-0134(20001115)41:3⟨271::AID-PROT10⟩3.0.CO;2-Z.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pp. 7694–7731. PMLR, 2023.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, pp. 5345–5355, 2018a.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1646–1655, 2018b. URL https://proceedings.mlr.press/v80/ganea18a.html.

Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z. Li. Foldtoken: Learning protein language via vector quantization and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 219–227, 2025. doi: 10.1609/aaai.v39i1.31998.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

Thomas Hayes, Roshan Rao, Halil Akin, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Neil He, Rishabh Anand, Hiren Madhu, et al. HELM: Hyperbolic large language models via mixture-of-curvature experts. *arXiv preprint arXiv:2505.24722*, 2025a. URL https://arxiv.org/abs/2505.24722.

Neil He, Hiren Madhu, Ngoc Bui, Menglin Yang, and Rex Ying. Hyperbolic deep learning for foundation models: A survey. In *SIGKDD*, 2025b.

Neil He, Menglin Yang, and Rex Ying. Lorentzian residual neural networks. In *KDD*, 2025c.

Raiyan Khan, Philippe Chlenski, and Itsik Pe'er. Hyperbolic genome embeddings. In *International Conference on Learning Representations*, volume 2025, pp. 73425–73454, 2025.

Marc Law, Renjie Liao, Jake Snell, and Richard Zemel. Lorentzian distance learning for hyperbolic representations. In *ICML*, pp. 3672–3681. PMLR, 2019.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yicong Li, Hongxu Chen, Xiangguo Sun, Zhenchao Sun, Lin Li, Lizhen Cui, Philip S Yu, and Guandong Xu. Hyperbolic hypergraphs for sequential recommendation. In *CIKM*, 2022.

Xiaohan Lin, Zhenyu Chen, Yanheng Li, Xingyu Lu, Chuanliu Fan, Ziqiang Cao, Shihao Feng, Yi Qin Gao, and Jun Zhang. Protokens: A machine-learned language for compact and informative encoding of protein 3d structures. *bioRxiv*, 2023. doi: 10.1101/2023.11.27.568722.

Andrei N. Lupas, Chris P. Ponting, and Robert B. Russell. On the evolution of protein folds: Are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of Structural Biology*, 134(2–3):191–203, 2001. doi: 10.1006/jsbi.2001.4393.

Craig O. Mackenzie and Gevorg Grigoryan. Protein structural motifs in prediction and design. *Current Opinion in Structural Biology*, 44:161–167, 2017. doi: 10.1016/j.sbi.2017.03.012.

Craig O. Mackenzie, Jianfu Zhou, and Gevorg Grigoryan. Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences*, 113(47):E7438–E7447, 2016. doi: 10.1073/pnas.1607178113.

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, pp. 1–25, 2024.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, pp. 6338–6347, 2017.

Alexander Rives, Joshua Meier, Tom Sercu, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *ICLR*, 2020.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1c42513b8895ab11fbbb5b7e8e6b6b02-Abstract-Conference.html.

Michael Sun, Weize Yuan, Gang Liu, Wojciech Matusik, and Marinka Zitnik. Protein structure tokenization via geometric byte pair encoding. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.11758v2.

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

Max van Spengler, Artem Moskalev, Tommaso Mansi, Mangal Prakash, and Rui Liao. Hyperhelm: Hyperbolic hierarchy encoding for mrna language modeling. *arXiv preprint arXiv:2509.24655*, 2025.

Jianhui Wang, Wenyu Zhu, Bowen Gao, Xin Hong, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. Learning protein-ligand binding in hyperbolic space. *AAAI*, 2026.

Menglin Yang, Zhihao Li, Min Zhou, Jiahong Liu, and Irwin King. Hicf: Hyperbolic informative collaborative filtering. In *KDD*, pp. 2212–2221, 2022.

Menglin Yang, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. *ICML LLM Cognition Workshop*, 2024a.

Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring efficient transformer fully in hyperbolic space. In *KDD*, pp. 3770–3781, 2024b.

Xinyu Yuan, Zichen Wang, Marcus Collins, and Huzefa Rangwala. Protein structure tokenization: Benchmarking and new recipe. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2503.00089v2.

Yiding Zhang, Xiao Wang, Chuan Shi, Nian Liu, and Guojie Song. Lorentzian graph convolutional networks. In *WebConf*, pp. 1249–1261, 2021.

A Supplementary Comparison with Different Training Losses

Table 2 compares MORPH with a pure-Euclidean configuration, which uses two Euclidean experts and retains the adapted representations in Euclidean space. Its architecture and training objectives are detailed in Appendix A.1. Because this configuration differs from MORPH in both representation geometry and training objectives, the comparison evaluates the two configurations as a whole and does not isolate the effect of geometry alone.

MORPH obtains a higher score on 21 of the 24 splits and a lower score on three. The Euclidean configuration performs better on CatBio Fold, Rep Fold, and FlexRMSF Fold. For example, it reaches 0.6338 on FlexRMSF Fold, compared with 0.6143 for MORPH. We interpret these results as a supplementary comparison of complete training configurations with different objectives.

A.1 Pure-Euclidean Configuration

The pure-Euclidean configuration adapts frozen ESM3 features using two independently parameterized Euclidean experts. It uses RamaBPE motif identities and composition relations, together with continuous structure-derived similarities, as training signals. All adapted residue and occurrence representations remain in Euclidean space, without Lorentz lifting or entailment cones. The architecture and objectives are detailed below.

Euclidean adaptation and occurrence pooling. Using the fused residue features b_i defined in the main method, the two experts produce

$$x_i^{(k)} = E_k(b_i) \in \mathbb{R}^{256}, \qquad k \in \{0, 1\}. \tag{1}$$

The residue-level router combines their outputs, followed by a gated residual update:

$$m_i = \sum_{k=0}^{1} \pi_i^{(k)} x_i^{(k)}, \qquad h_i' = H_{\mathrm{ESM3},i} + g_i W_{\mathrm{up}} m_i. \tag{2}$$

Here, $h_i' \in \mathbb{R}^{1536}$ is the final adapted residue representation.

For occurrence v with span S_v and motif type m_v, token-conditioned attention gives

$$\ell_{v,i} = w_a^\top \tanh\left(W_h h_i' + W_e e_{m_v} + b_a\right), \tag{3}$$

$$\alpha_{v,i} = \frac{\exp(\ell_{v,i})}{\sum_{j \in S_v} \exp(\ell_{v,j})}, \tag{4}$$

$$z_v = W_o \sum_{i \in S_v} \alpha_{v,i} h_i' + b_o \in \mathbb{R}^{256}. \tag{5}$$

Each occurrence is pooled separately, allowing occurrences with the same motif identity to retain different representations according to their protein context.

Occurrence geometry supervision. The structural target distribution $q_T(u \mid v)$ is constructed from fixed structure-derived similarities, as in the main method. The Euclidean prediction uses cosine similarity:

$$p_E(u \mid v) = \mathrm{softmax}_{u \in \mathcal{C}(v)} \left(\frac{\langle \hat{z}_v, \hat{z}_u \rangle}{T_s} \right), \qquad \hat{z}_v = \frac{z_v}{\|z_v\|_2}. \tag{6}$$

The corresponding objective is

$$L^E_{\mathrm{geo}} = \mathbb{E}_v \, \mathrm{KL} \left(q_T(\cdot \mid v) \, \| \, p_E(\cdot \mid v) \right). \tag{7}$$

Both target and prediction temperatures are set to $T_t = T_s = 0.1$. The target distribution is detached from the computation graph, while gradients update the occurrence pooling, shared motif embeddings, and adapter. Because cosine similarity uses normalized vectors, this objective constrains occurrence directions rather than their norms.

Shared-type relation supervision. The Euclidean ontology objective acts directly on shared motif type embeddings. Let $\mathcal{P}$ denote the positive parent–component type pairs and $\mathcal{N}(p)$ the negative types incomparable with parent p under the composition hierarchy. Define

$$r_E(e) = \|e\|_2, \qquad d_E(e, e') = \|e - e'\|_2. \tag{8}$$

The objective combines radial ordering, distance ranking, and positive-pair compactness:

$$L^E_{\mathrm{radial}} = \mathbb{E}_{(p,c) \in \mathcal{P}} \left[m_r - r_E(e_p) + r_E(e_c) \right]_+, \tag{9}$$

$$L^E_{\mathrm{ranking}} = \mathbb{E}_{\substack{(p,c) \in \mathcal{P} \\ n \sim \mathcal{N}(p)}} \left[m_d + d_E(e_p, e_c) - d_E(e_p, e_n) \right]_+, \tag{10}$$

$$L^E_{\mathrm{relation}} = \mathbb{E}_{(p,c) \in \mathcal{P}} d_E(e_p, e_c), \tag{11}$$

$$L^E_{\mathrm{ont}} = L^E_{\mathrm{radial}} + L^E_{\mathrm{ranking}} + \lambda_{\mathrm{relation}} L^E_{\mathrm{relation}}, \tag{12}$$

where $[x]_+ = \max(0, x)$, $m_r = m_d = 0.1$, and $\lambda_{\mathrm{relation}} = 0.01$. The implementation averages this objective over the two Euclidean ontology branches.

The radial term encourages $r_E(e_p) \geq r_E(e_c) + m_r$, placing parent types at larger norms than their components. The ranking term separates positive relations from negative pairs, while the compactness term draws related types together. This supervision acts on shared types; it does not impose explicit occurrence hierarchy or type-to-occurrence realization constraints.

Ordered structural composition. Occurrence-level composition is supervised through prediction of parent geometry from ordered children. For an occurrence merge $p = \mathrm{merge}(l, r)$,

$$\hat{d}_p = f_{\mathrm{comp}}([z_l, z_r]), \qquad L_{\mathrm{comp}} = \mathbb{E}_{(p,l,r)} \, \mathrm{SmoothL1}(\hat{d}_p, d_p), \tag{13}$$

where $d_p \in \mathbb{R}^{122}$ is a fixed structural descriptor computed offline from the parent span. The ordered concatenation preserves the distinction between left and right components. This objective predicts parent geometry without directly constraining the relative positions of parent and child occurrence embeddings.

Preservation of pretrained outputs. A preservation objective compares the predictions obtained from H_{ESM3} and H' using six frozen native ESM3 heads: sequence, structure, secondary structure, SASA, function, and residue annotations. Categorical KL divergence is used for categorical predictions, including factorized function outputs; Bernoulli divergence is used for residue annotations. Each head loss is normalized over its own valid elements, and the six losses receive equal weights:

$$L_{\mathrm{keep}} = \frac{1}{6} \sum_{h=1}^{6} L^{(h)}_{\mathrm{keep}}. \tag{14}$$

Predictions from the original features serve as stop-gradient targets. The head parameters remain frozen, while gradients through predictions from H' propagate to the adapter.

Overall training objective. The complete Euclidean objective is

$$L_E = L^E_{\text{geo}} + 0.5L^E_{\text{ont}} + 0.5L_{\text{comp}} + L_{\text{keep}} + 0.01\beta(t)L_{\text{bal}}, \tag{15}$$

where L_{bal} regularizes expert usage and $\beta(t)$ follows a cosine schedule from one to zero.

This configuration provides a Euclidean implementation of structural similarity and composition-based training. It differs from MORPH in its expert and representation geometries, relation objectives, and auxiliary losses. In particular, its type-level radial convention places parents farther from the origin, whereas MORPH's occurrence-level radial constraint places children farther from the origin. The supplementary comparison therefore evaluates the complete configurations rather than isolating curvature under matched objectives.

Table 2: **Downstream comparison with a pure-Euclidean adapter.** The pure-Euclidean configuration uses two Euclidean experts and Euclidean output representations, as detailed in Appendix A.1. It differs from MORPH in both representation geometry and training objectives; this comparison therefore does not isolate the effect of geometry alone. The results complement the occurrence-level analysis in Figure 3. Bold entries with gray backgrounds indicate the higher score in each row. Higher is better for all metrics.

Task	Split	Pure Euclidean	MORPH
Functional Site Prediction (AUROC)			
BindBio	Fold	0.8877	**0.9087**
	SupFam	0.9008	**0.9331**
BindInt	Fold	0.5522	**0.5773**
	SupFam	0.9757	**0.9777**
BindShake	Org/Test	0.8870	**0.9069**
CatBio	Fold	**0.9364**	0.9246
	SupFam	0.9793	**0.9924**
CatInt	Fold	0.6269	**0.6459**
	SupFam	0.9749	**0.9821**
Con	Fold	0.6534	**0.6551**
	SupFam	0.9604	**0.9816**
Ept	Fold	0.7204	**0.7973**
	SupFam	0.8299	**0.9163**
Rep	Fold	**0.7757**	0.7598
	SupFam	0.8265	**0.8678**
Structural Flexibility Prediction (Spearman's ρ)			
FlexBFactor	Fold	0.4359	**0.4382**
	SupFam	0.4029	**0.4295**
FlexNEQ	Fold	0.6953	**0.6996**
	SupFam	0.6866	**0.6921**
FlexRMSF	Fold	**0.6338**	0.6143
	SupFam	0.5463	**0.5801**
Protein-level Fold Classification (Macro-F1)			
Homo	Fold	0.3051	**0.3355**
	SupFam	0.8283	**0.9139**
	Family	0.9198	**0.9556**

SupFam denotes Superfamily. Homo uses the corresponding Fold, Superfamily, and Family holdout splits; Org/Test denotes the original BindShake test split. Scores are rounded to four decimal places; best entries are identified before rounding.

B Ramachandran Alphabet Construction and Diagnostics

We fit the RamaBPE alphabet using 74,993 valid backbone torsion pairs from 330 structures. Each pair is embedded as $(\cos\phi, \sin\phi, \cos\psi, \sin\psi)$ and clustered into 32 states using k-means. The fitted

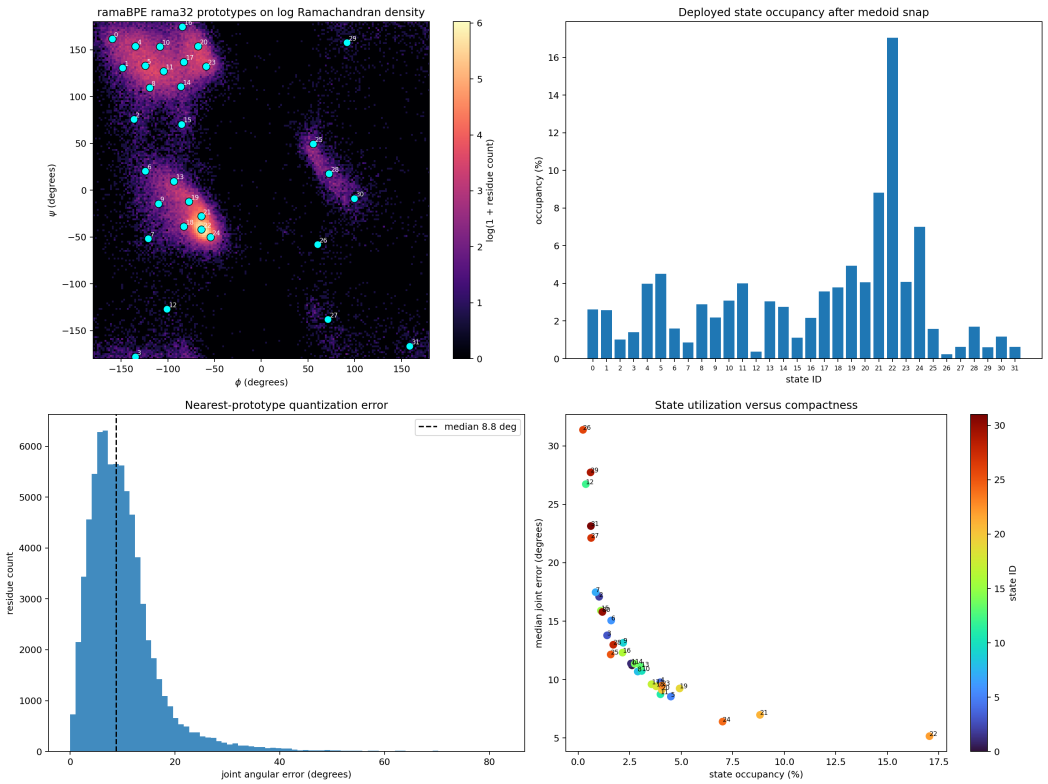

Figure 4: **Diagnostics of the 32-state Ramachandran alphabet.** Top left: learned prototypes overlaid on the fitting-set Ramachandran distribution, with background color indicating $\log(1 + \text{residue count})$. Top right: residue occupancy of each state after prototype snapping. Bottom left: joint angular quantization error distribution; the dashed line marks the median of $8.8°$. Bottom right: state occupancy versus median joint angular error, with colors and labels identifying states. All panels describe the alphabet-fitting set.

centers are projected onto the two unit circles and snapped to their nearest observed torsion pairs. The resulting prototypes are frozen for subsequent quantization and BPE construction.

Figure 4 summarizes the learned prototypes, state occupancy, and angular quantization errors. The median joint angular error is $8.8°$ on the fitting set. State occupancy is nonuniform, and the per-state diagnostics show variation in quantization error across the alphabet. These measurements characterize local torsion discretization; they are distinct from the within-motif backbone RMSD analysis in Section 4.3.

C Training Details

C.1 Input Representations and Model Configuration

MORPH uses frozen residue representations from ESM3-SM-open-v1 with sequence and structure inputs. The representations are extracted after the final LayerNorm, have dimension 1536, and are cached offline in BF16. All 48 Transformer blocks and the six native prediction heads remain frozen during adapter training. The complete cache contains 33,988 proteins and 7,778,389 residues; optimization uses only proteins assigned to the training split.

RamaBPE provides 18,030 motif identities. An additional learned embedding represents residues not covered by valid BPE segments. These residues remain part of the adapted protein representations but do not contribute to objectives requiring valid motif occurrences.

The adapter bottleneck and shared motif embeddings both have dimension 256. Each expert uses an FFN with dimensions $256 \to 512 \to 256$, and the router has a hidden dimension of 64. The hyperbolic expert uses curvature -1. The final Lorentz residue representations have intrinsic dimension 1536 and ambient dimension 1537. Type and occurrence representations are projected into a common ontology space with intrinsic dimension 256 and curvature -1.

Token-conditioned attention uses a 128-dimensional scoring space and is shared across all motif occurrences. Attention is normalized within each occurrence's residue span. The same pooling parameters are applied to occurrences at different composition scales; the composition forest specifies their spans and relations rather than a recursive pooling computation.

C.2 Structure-Derived Similarity for Geometric Supervision

Each valid motif occurrence is associated with a fixed descriptor derived from its Cα coordinates. The descriptor has dimension $D = 122$, consisting of 120 shape features constructed from 16 sampled points and two scale-related features. Shape construction includes normalization by the radius of gyration. Descriptors are constructed only for occurrences spanning at least three residues, with Cα coverage of at least 80% and valid Cα coordinates at both endpoints.

For each occurrence v, let $d_v \in \mathbb{R}^D$ denote its descriptor. Feature-wise standardization gives

$$\hat{d}_{v,k} = \frac{d_{v,k} - \mu_k}{\sigma_k},$$

where μ_k and σ_k are estimated exclusively from training-set descriptors. The statistics are computed from 1,797,477 training occurrences and remain fixed during adapter training.

The structure-derived distance and similarity are

$$D_T(v,u) = \frac{\|\hat{d}_v - \hat{d}_u\|_2}{\sqrt{D}}, \tag{16}$$

$$\mathrm{sim}_T(v,u) = \exp\left(-\frac{D_T(v,u)^2}{2\sigma_T^2}\right). \tag{17}$$

The similarity scale is set to $\sigma_T = 0.2091242597$, estimated as the median pairwise descriptor distance over the training pairs used for calibration.

For a sampled candidate set $\mathcal{C}(v)$, the target distribution is

$$q_T(u \mid v) = \frac{\exp\big(\mathrm{sim}_T(v,u)/\tau_T\big)}{\sum_{w \in \mathcal{C}(v)} \exp\big(\mathrm{sim}_T(v,w)/\tau_T\big)}, \qquad u \in \mathcal{C}(v).$$

Both the target and representation-based distributions use temperature 0.1. Geometric supervision minimizes $\mathrm{KL}(q_T(\cdot \mid v) \,\|\, p(\cdot \mid v))$, where p is defined from occurrence representations as described in the main text. The descriptors and target distributions are fixed and receive no gradients.

Shared motif identity provides a prior for candidate sampling rather than a categorical similarity target. Occurrences of the same type can therefore have different target similarities, while occurrences of different types can receive high similarity when their descriptors are close.

C.3 Ontology Sampling

Type composition, occurrence composition, and type-to-occurrence realization are sampled separately. For each relation category, at most 4,096 positive relations are sampled per optimization step, with seven negatives per positive relation. Negative sampling excludes nodes comparable under the corresponding relation's transitive closure, including both direct and indirect ancestors and descendants.

Positive and negative contributions are normalized separately within each relation category before combining the objectives. The cone scale is 0.05, the negative angular margin is 0.1 radians, and the radial margin is 0.05. Ontology projectors enforce a minimum hyperbolic radius of 0.2 in their forward computation; the minimum-radius penalty provides an additional safeguard.

C.4 OPTIMIZATION

Training uses AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$. The learning rate is 3×10^{-4} for non-token parameters and 10^{-3} for the shared motif embedding table. Weight decay is 0.01 for multidimensional non-token parameters and zero for biases, one-dimensional parameters, and motif embeddings. The learning rate follows a schedule with 5% warmup followed by cosine decay to 10% of its initial value. The global gradient norm is clipped to 1.0.

The reported training run uses eight distributed ranks, random seed 17, and ten epochs, corresponding to 670 optimizer steps. Each active rank processes the training proteins from an assigned cache chunk, so the number of proteins and residues per step varies. Chunk assignment is deterministic within each epoch.

Forward computation uses BF16 autocasting, while numerically sensitive Lorentz operations and KL computations use FP32. The router-balance coefficient is $0.01\beta(t)$, where $\beta(t)$ decreases from one to zero according to a cosine schedule. No validation-based early stopping or best-validation checkpoint selection is performed during this adapter-training run.