$\mathbb{SL}(n)$ REPRESENTATION LEARNING: AN INTRINSIC MIXED-CURVATURE SPACE WITH HIGHER CURVATURE CAPACITIES AND DEEPER ORDER-AWARE COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed-curvature representation learning seeks to capture rich geometric structures that cannot be adequately modeled by a single curvature regime. Existing approaches largely rely on product manifolds, which require manually specifying how different curvature spaces are combined and separate their curvature contributions across factors. We introduce the $\mathbb{SL}(n)$ space, a representation geometry defined by the simple $\det(A) = 1$ constraint and a left invariant Schatten-p Finsler structure. Despite this minimal construction, $\mathbb{SL}(n)$ exhibits pointwise negative, zero, and positive flag curvature around a common flagpole, while its mixed-curvature and curvature-coupling capacities are asymptotically maximal relative to the intrinsic geometric upper bound. Beyond geometry, its noncommutative group structure provides inherent order sensitivity, and its non-nilpotent Lie algebra admits nonzero nested Lie brackets at arbitrary depth, enabling deep order-aware composition. Empirically, $\mathbb{SL}(n)$ consistently outperforms a broad range of representation manifold baselines across graph benchmarks at different scales. It reduces average distortion over the strongest baselines by 44.3% on KEGG and 40.5% on HumanCyc, and improves Hits@20 by 42.8% on OGBL-PPA. Experiments on Flickr30k-Order further support its ability to capture higher order dependencies from ordered composition. Together, these results show how a seemingly simple structural constraint can yield unexpectedly rich geometry, capacity, and composition within a unified representation space.

1 INTRODUCTION

Finding informative representations of data has long been a central problem in machine learning, from dimensionality reduction by principal component analysis (Hotelling, 1933), through learning compact representations with neural networks (Hinton & Salakhutdinov, 2006) and distributed representations of words through embeddings (Mikolov et al., 2013), to modern non-Euclidean representations on manifolds (Bronstein et al., 2021; Diepeveen et al., 2025). This progression naturally raises a key question: *what geometry should the representation space be endowed with?* The manifold hypothesis provides a fundamental starting point, suggesting that high dimensional observations often concentrate near manifolds of substantially lower intrinsic dimension (Tenenbaum et al., 2000; Fefferman et al., 2013). Yet it does not determine the geometry of this latent manifold, so the chosen geometry acts as an inductive bias toward particular structural patterns. For instance, classical Euclidean geometry $\mathbb{E}^n$ (curvature $K = 0$) is suited to approximately flat or grid-like structures (Mikolov et al., 2013), hyperbolic geometry $\mathbb{H}^n$ ($K < 0$) naturally accommodates hierarchical or tree-like structures (Nickel & Kiela, 2017), whereas spherical geometry $\mathbb{S}^n$ ($K > 0$) provides a compact geometry suited to cyclic or clique-like structures (Bachmann et al., 2020; Sun et al., 2021).

Real-world structures, however, rarely conform to a simple, single-curvature regime. For example, even a single graph, relational system, or biological network may simultaneously contain hierarchical, cyclic, densely interconnected, and approximately flat substructures (McNeela et al., 2025). This observation has motivated **mixed-curvature representation learning**. A straightforward idea is to combine manifolds of different curvatures into a single representation space. This leads to the

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.

widely used *mixed-curvature product manifold* (Gu et al., 2019), defined as the Cartesian product

$$\mathcal{P} = \mathbb{E}^{d_0} \times \prod_{i=1}^{m_-} \mathbb{H}^{d_i}_{K_i} \times \prod_{j=1}^{m_+} \mathbb{S}^{d_j}_{K_j}, \tag{1}$$

where d_0, d_i, d_j denote the factor dimensions, while $K_i < 0$ and $K_j > 0$ denote the corresponding hyperbolic and spherical curvatures. Despite its simplicity, this construction has been successfully extended to generative modeling and applied across a variety of downstream domains (Skopek et al., 2020; Bachmann et al., 2020; Wang et al., 2021; Sun et al., 2021; Wang et al., 2023). However, product mixed-curvature spaces inherit two fundamental limitations.

First, product mixed-curvature spaces require the combination of spaces with different curvatures to be *specified in advance*, while the appropriate combination can vary substantially across data objects. This strong inductive bias makes a suitable product geometry difficult to determine. For example, even for a fixed total dimension $d_0 + \sum_{i=1}^{m_-} d_i + \sum_{j=1}^{m_+} d_j = 64$ and fixed curvature magnitudes, there are $11{,}555{,}651{,}398$[1] different candidates! (note that $\mathbb{S}^3 \times \mathbb{H}^1 \neq \mathbb{S}^2 \times \mathbb{H}^2$ and $\mathbb{H}^2 \times \mathbb{H}^2 \neq \mathbb{H}^4$). Thus, related work often requires searching over many product spaces (McNeela et al., 2025).

Second, and more importantly, product spaces with mixed curvature have an inherent *representation limitation*. Let $\mathcal{M} = \prod_i \mathcal{M}_i$ be a product of component manifolds and $T_x\mathcal{M}$ its tangent space at x. The product geometry satisfies $T_x\mathcal{M} = \bigoplus_i T_{x_i}\mathcal{M}_i$, while its curvature is generated independently within each factor, with no curvature interaction across factors (Sáez de Ocáriz Borde, 2026). Thus, positive and negative curvature coexist through separate factor components rather than through their intrinsic interaction. This limits the representation of coupled structures, where hierarchical, cyclic, and other patterns share common features instead of decomposing into independent parts.

These two severe limitations motivate the central question of this work: *can we find a single geometry that intrinsically represents coupled mixed-curvature structures, without manually allocating different curvatures to separate factors?* Some promising alternatives have emerged from previous works. SPD and higher-rank Siegel manifolds provide intrinsic $\{-, 0\}$ curvature structures for representation learning (Lopez et al., 2021; Zhao et al., 2023), while Grassmann manifolds provide intrinsic $\{+, 0\}$ curvature (Bendokat et al., 2024). This leaves a pressing open problem in representation learning of realizing intrinsically coupled $\{-, 0, +\}$ curvature within a single latent manifold.

To address this problem, we propose the $\mathbb{SL}(\mathbf{n})$ **space** (we omit p when no confusion arises), an intrinsic mixed-curvature space constructed from the special linear group and defined as

$$\mathbb{SL}_p(n) := \big(\mathrm{SL}(n), F_p, \circ\big), \quad \mathrm{SL}(n) := \{A \in \mathbb{R}^{n\times n} \mid \det(A) = 1\}, \tag{2}$$

where $F_p(A, V) = \|A^{-1}V\|_{S_p}$ is the globally defined Schatten-p tangent norm, introducing direction dependent sensitivity to matrix variations beyond a quadratic Riemannian metric, and $\circ$ denotes group composition, naturally given by matrix multiplication. Importantly, F_p and the induced length geometry are defined for all tangent vectors, while full rank regularity is required only for smooth curvature analysis and excludes a measure zero set of directions. The induced Finsler geometry has directional flag curvature $K_F(Y, \Pi)$, allowing different tangent directions to exhibit different local curvature. We prove that positive, zero, and negative flag curvatures coexist at every point, with different signs interacting through shared tangent directions rather than separate factors. We quantify these two properties by the mixed-curvature capacity $\mathcal{C}^{\mathrm{mix}}_{\mathbb{SL}}$ and curvature-coupling capacity $\mathcal{C}^{\mathrm{cpl}}_{\mathbb{SL}}$, respectively, capturing balanced coexistence and genuine coupling. Both are asymptotically maximal relative to the intrinsic geometric upper bound. Experiments on real-world graphs of substantially different scales further show strong improvements over a broad range of geometric baselines.

Beyond geometry, the Lie group structure of $\mathrm{SL}(n)$ provides an intrinsic mechanism for composition. Its group operation is matrix multiplication, while its Lie algebra $\mathfrak{sl}(n)$ carries the Lie bracket $[X, Y] = XY - YX$. Noncommutativity makes sequential composition inherently order-aware, while the non-nilpotent structure of $\mathfrak{sl}(n)$ permits nonzero nested Lie brackets $[X_k, [\cdots, [X_2, X_1]\cdots]]$ at arbitrary depth k. This enables progressively deeper ordered composition beyond pairwise noncommutativity, which we further evaluate empirically.

Together, these properties unify mixed-curvature geometry and order-aware algebraic structure within a single matrix representation space. And due to page limitations, beyond the related work discussed above, a more detailed Related Work section is provided in Appendix [A].

[1] See Appendix [D.5] for how this number is calculated.

2 THE $\mathbb{SL}$ SPACE

2.1 STRUCTURES OF $\mathbb{SL}$ SPACE

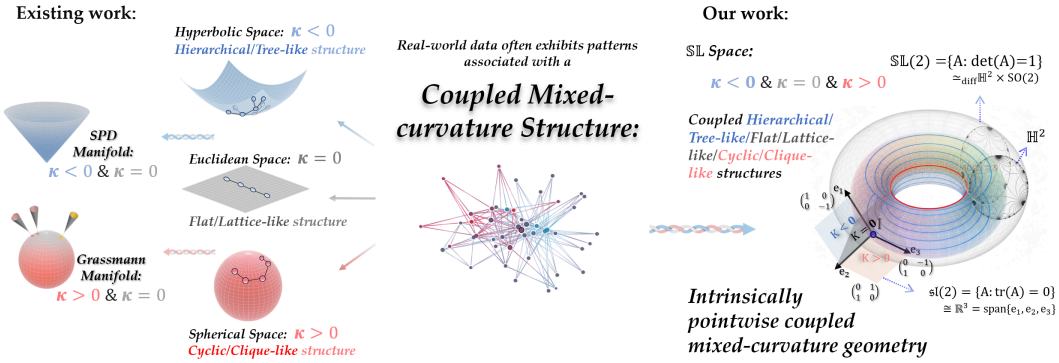


Figure 1: Existing manifolds capture only restricted curvature regimes, whereas real-world data may contain complex coupled structures. The $\mathbb{SL}$ space intrinsically couples $\{-, 0, +\}$ curvature within a single geometry. The right side depicts the global topology and tangent Lie algebra $\mathfrak{sl}(2)$.

As illustrated in Fig. 1, the $\mathbb{SL}(n)$ space provides a matrix representation that combines a smooth manifold structure, a flexible Finsler geometry, and an intrinsic Lie group structure, while supporting coupled mixed-curvature within a single space. We introduce these structures in turn below.

2.1.1 SMOOTH MANIFOLD STRUCTURE

A smooth manifold is a space that locally resembles Euclidean space and admits smooth coordinate systems for differential operations (Lee, 2013). Begin with the matrix set

$$\mathrm{SL}(n) := \left\{ A \in \mathbb{R}^{n\times n} : \det(A) = 1 \right\}. \tag{3}$$

For the smooth map $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$, writing $D\det_A[V]$ for the directional differential of $\det$ at $A \in \mathrm{SL}(n)$ along $V \in \mathbb{R}^{n\times n}$ in the ambient Euclidean space, $D\det_A[V] = \mathrm{tr}(A^{-1}V)$. Since $D\det_A \neq 0$ on $\det^{-1}(1)$, 1 is a regular value of $\det$. By the regular level set theorem, $\mathrm{SL}(n) = \det^{-1}(1)$ is a **smooth embedded manifold** of dimension $n^2 - 1$. Its tangent space is therefore

$$T_A\mathrm{SL}(n) = \ker(D\det_A) = \left\{ V \in \mathbb{R}^{n\times n} : \mathrm{tr}(A^{-1}V) = 0 \right\}. \tag{4}$$

2.1.2 SCHATTEN-p FINSLER GEOMETRIC STRUCTURE

A Riemannian structure assigns a smoothly varying inner product g_x to each tangent space T_xM, inducing the norm $\|V\|_x = \sqrt{g_x(V,V)}$ for tangent vectors. A Finsler structure generalizes this construction by allowing a smoothly varying tangent norm $F(x,V)$ that need not arise from an inner product (Bao et al., 2000). This lets local geometry depend more richly on tangent directions.

For $1 \le p < \infty$, the Schatten-p norm is $\|X\|_{S_p} := \left(\sum_{i=1}^{n} \sigma_i(X)^p\right)^{1/p}$, where $\sigma_i(X)$ is the i-th singular value of X. We equip $\mathrm{SL}(n)$ with the **Schatten-p tangent norm**

$$F_p(A,V) := \|A^{-1}V\|_{S_p}, \quad A \in \mathrm{SL}(n), \quad V \in T_A\mathrm{SL}(n). \tag{5}$$

We focus on $1 < p < \infty$. Importantly, F_p is well defined on the entire tangent bundle and therefore induces a global length structure on $\mathrm{SL}(n)$. For $p \neq 2$, the smooth differential geometry required for flag curvature is considered on full rank tangent directions, which form an open dense set with measure zero complement. The case $p = 2$ is globally Riemannian, while $p = 1$ is excluded due to non-smoothness. For a piecewise smooth curve $\gamma : [0,1] \to \mathrm{SL}(n)$, we define $L_p(\gamma) = \int_0^1 F_p(\gamma(t), \dot\gamma(t))\, dt$ and $d_p(A,B) = \inf_{\gamma: A\to B} L_p(\gamma)$.

Finsler curvature is described by flag curvature, which generalizes Riemannian sectional curvature (Bao et al., 2000). We call a nonzero tangent direction $Y \in T_A\mathrm{SL}(n)$ regular when $A^{-1}Y$ is full rank. For a regular tangent direction Y and a two-dimensional plane $\Pi = \mathrm{span}\{Y, U\} \subset T_A\mathrm{SL}(n)$

containing Y, the pair (Y, Π) is called a flag and Y its flagpole. The fundamental tensor at Y is $g_Y(U,V) := \frac{1}{2}\frac{\partial^2}{\partial s\,\partial t} F_p^2(A, Y + sU + tV)\Big|_{s=t=0}$. Let $\boldsymbol{\mathcal{R}}^Y(U,V)$ denote the Chern curvature operator with reference direction Y, and define the Jacobi operator by $R_Y U := \boldsymbol{\mathcal{R}}^Y(U,Y)Y$. Their explicit expressions are deferred to the curvature analysis in Appendix [D]. The flag curvature is then defined as

$$K_F(Y,\Pi) := \frac{g_Y(R_Y U, U)}{g_Y(Y,Y)g_Y(U,U) - g_Y(Y,U)^2}. \tag{6}$$

Intuitively, flag curvature $K_F(Y,\Pi)$ depends on direction, meaning that the curvature varies with the flag $\Pi = \mathrm{span}\{Y,U\}$. The Jacobi operator R_Y is self adjoint with respect to g_Y, meaning $g_Y(R_Y U, V) = g_Y(U, R_Y V)$. Therefore, all its eigenvalues are real, with positive eigenspace

$$E_+(Y) := \bigoplus_{\lambda>0} \ker(R_Y - \lambda I), \tag{7}$$

and negative eigenspace $E_-(Y)$ defined analogously over $\lambda < 0$.

2.1.3 Lie Group Algebraic Structure

Moreover, the underlying manifold carries a natural algebraic group structure. A group $(G, \circ)$ is a set equipped with an associative composition $\circ$, an identity element e, and an inverse a^{-1} for every $a \in G$. For $\mathrm{SL}(n)$, matrix multiplication $A \circ B := AB$, the identity $e = I$, and matrix inversion $A \mapsto A^{-1}$ define the group structure. A Lie group is simultaneously a smooth manifold and a group, with smooth composition and inversion (Hall, 2015). Since matrix multiplication and inversion are smooth on $\mathrm{SL}(n)$, it forms a **Lie group**. The associated Lie algebra $\mathfrak{sl}(n)$ is the tangent space at the identity, endowed with the Lie bracket $[\cdot,\cdot]$. For $\mathrm{SL}(n)$,

$$\mathfrak{sl}(n) = T_I\mathrm{SL}(n) = \{\Xi \in \mathbb{R}^{n\times n} : \mathrm{tr}(\Xi) = 0\}, \quad [X,Y] = XY - YX. \tag{8}$$

Left translation identifies every tangent space with the same Lie algebra, giving $T_A\mathrm{SL}(n) = A\mathfrak{sl}(n)$. Consequently, every $V \in T_A\mathrm{SL}(n)$ admits the unique left-trivialized coordinate $\Xi = A^{-1}V \in \mathfrak{sl}(n)$. An arbitrary matrix $X \in \mathbb{R}^{n\times n}$ can be mapped to the Lie algebra by $\Pi_{\mathfrak{sl}}(X)$:

$$\Pi_{\mathfrak{sl}}(X) = X - \mathrm{tr}(X)I/n \in \mathfrak{sl}(n). \tag{9}$$

The matrix exponential and logarithm then provide natural local mappings between the tangent space at A and the manifold. From a tangent vector $V \in T_A\mathrm{SL}(n)$, the exponential gives the constraint preserving retraction map R_A:

$$R_A : T_A\mathrm{SL}(n) \to \mathrm{SL}(n), \quad R_A(V) = A\exp(A^{-1}V), \tag{10}$$

where the matrix exponential is defined by $\exp(X) = \sum_{k=0}^{\infty} X^k/k!$. Since $\det(\exp\Xi) = \exp(\mathrm{tr}\,\Xi) = 1$ for $\Xi \in \mathfrak{sl}(n)$, the retraction remains in $\mathrm{SL}(n)$.

Conversely, on a neighborhood of A, the retraction admits a local inverse. The principal matrix logarithm $\log(M)$ is the unique matrix satisfying $\exp(\log M) = M$ whose eigenvalues have imaginary parts in $(-\pi, \pi)$. Hence, locally around A,

$$R_A^{-1} : \mathrm{SL}(n) \to T_A\mathrm{SL}(n), \quad R_A^{-1}(B) = A\,\log(A^{-1}B), \tag{11}$$

We can also define the closed form Schatten semidistance

$$D_{\mathbb{SL}}(A,B) := \|\log(A^{-1}B)\|_{S_p}. \tag{12}$$

It is nonnegative, symmetric, and point separating, but need not satisfy the triangle inequality or require tangent regularity. We use $D_{\mathbb{SL}}$ as the pairwise dissimilarity for representation learning.

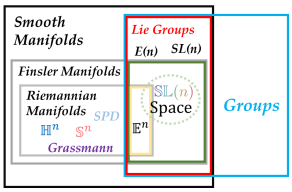


Figure 2: The $\mathbb{SL}(n)$ space combines Finsler geometry with Lie group structure, reducing to the Riemannian case at $p = 2$.

The Lie bracket on $\mathfrak{sl}(n)$ is the matrix commutator $[X,Y] = XY - YX$, which quantifies the change induced by reversing the order of two infinitesimal transformations and provides an intrinsic mechanism for order-aware composition. Locally, successive group transformations are related to nested Lie brackets through the BCH expansion, $\log(\exp X \exp Y) = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}[X,[X,Y]] + \cdots$. Overall, these structures combine smooth manifold, Schatten-p Finsler geometric, and Lie group algebraic structures, supporting both intrinsic geometry and ordered composition. Finally, we define the unified $\mathbb{SL}$ representation space as

$$\mathbb{SL}_{\mathbf{p}}(\mathbf{n}) := \big(\mathrm{SL}(\mathbf{n}), \mathbf{F}_{\mathbf{p}}, \circ\big). \tag{13}$$

2.2 Intrinsic Coupled Mixed Curvature

2.2.1 Pointwise Intrinsic Coupled Mixed Flag Curvature

Having defined the $\mathbb{SL}$ space, we now characterize its intrinsic curvature structure. Hereafter, we write $A \in \mathbb{SL}_p(n)$ and $T_A\mathbb{SL}_p(n) := T_A\mathrm{SL}(n)$ (with p omitted when unambiguous) when referring to the resulting representation space. We call mixed-curvature intrinsic when different curvature regimes arise within a single tangent geometry and are not attributable to separate factors of a metric product. $\mathbb{SL}_p(n)$ exhibits such intrinsic mixed-curvature with different curvature signs within the same tangent geometry.

Theorem 2.1 (Pointwise mixed flag curvature)**.** *For every $n \geq 2$ and **every** $p \in (1, \infty)$, for **every** $A \in \mathbb{SL}_p(n)$, there exist a regular full-rank matrix (flagpole) $Y_A \in T_A\mathbb{SL}_p(n)$ and three tangent matrices $U_A^+, U_A^0, U_A^- \in T_A\mathbb{SL}_p(n)$, each linearly independent of Y_A, such that*

$$K_F(Y_A, \mathrm{span}\{Y_A, U_A^+\}) > 0, \quad K_F(Y_A, \mathrm{span}\{Y_A, U_A^0\}) = 0, \quad K_F(Y_A, \mathrm{span}\{Y_A, U_A^-\}) < 0.$$

Proof. The full and detailed proof is in Appendix [D.2]. □

Hence, $\mathbb{SL}(n)$ exhibits pointwise $-, 0, +$ mixed flag curvature around a common flagpole. We next distinguish genuine curvature coupling from mere coexistence.

Definition 2.2 (Mixed-curvature and coupling capacities)**.** For a Finsler manifold $(\mathcal{M}, F)$ and a regular flagpole $Y \in T_x\mathcal{M}$, let $E_+(Y)$ and $E_-(Y)$ be the positive and negative eigenspaces of R_Y, respectively, as defined in Eq. 7. Define the mixed-curvature capacity as

$$\mathcal{C}_{\mathrm{mix}}(Y) := \min\{\dim E_+(Y), \dim E_-(Y)\}. \tag{14}$$

Define the uncoupled subspaces $\mathcal{N}_+(Y) := \{U \in E_+(Y) : \boldsymbol{\mathcal{R}}^Y(U, V) = 0,\ \forall V \in E_-(Y)\}$ and $\mathcal{N}_-(Y)$ analogously. The curvature coupling capacity is

$$\mathcal{C}_{\mathrm{cpl}}(Y) := \min\{\dim E_+(Y) - \dim \mathcal{N}_+(Y),\ \dim E_-(Y) - \dim \mathcal{N}_-(Y)\}. \tag{15}$$

At $x \in \mathcal{M}$, a flagpole level capacity $\mathcal{C}(Y)$ induces $\mathcal{C}(x) := \max_{Y \in T_x\mathcal{M} \text{ regular}} \mathcal{C}(Y)$ and $\mathcal{C}_{\mathcal{M}}(F) := \min_{x \in \mathcal{M}} \mathcal{C}(x)$. We use this convention for both $\mathcal{C}_{\mathrm{mix}}$ and $\mathcal{C}_{\mathrm{cpl}}$.

Intuitively, $\mathcal{C}_{\mathrm{mix}}$ measures the balanced number of positive and negative curvature modes coexisting around a common flagpole, whereas $\mathcal{C}_{\mathrm{cpl}}$, like rank in linear algebra after excluding null directions, counts only those modes that interact across curvature signs. Hence $\mathcal{C}_{\mathrm{cpl}} \leq \mathcal{C}_{\mathrm{mix}}$. Zero curvature is not counted separately, as it follows between positive and negative flag curvatures by continuity.

Corollary 2.3 (Asymptotically maximal mixed curvature and coupling)**.** *For every $n \geq 2$ and $p \in (1, \infty)$, since the transverse tangent space has dimension $n^2 - 2$, define $\mathcal{C}_{\max}(n) := \lfloor (n^2 - 2)/2 \rfloor$. Then*

$$\begin{gathered} \frac{(n-1)(n-2)}{2} \leq \mathcal{C}_{\mathbb{SL}}^{\mathrm{cpl}}(n,p) \leq \mathcal{C}_{\mathbb{SL}}^{\mathrm{mix}}(n,p) \leq \mathcal{C}_{\max}(n), \\ \frac{\mathcal{C}_{\mathbb{SL}}^{\mathrm{cpl}}(n,p)}{\mathcal{C}_{\max}(n)} \longrightarrow 1, \qquad \frac{\mathcal{C}_{\mathbb{SL}}^{\mathrm{mix}}(n,p)}{\mathcal{C}_{\max}(n)} \longrightarrow 1 \quad \textit{as } n \to \infty. \end{gathered} \tag{16}$$

Proof. The full and detailed proof is in Appendix [D.3]. □

Accordingly, $\mathbb{SL}(n)$ supports asymptotically maximal coexistence of positive and negative curvature modes while intrinsically coupling an asymptotically maximal number of these modes.

2.3 Deep Order-Aware Composition

Beyond its geometric structure, $\mathbb{SL}(n)$ provides an intrinsic mechanism for representing ordered interactions through noncommutative group composition. For $A, B \in \mathbb{SL}(n)$, generally $AB \neq BA$, so reversing their order changes the composition. Locally, this difference is captured by the Lie bracket $[X, Y] = XY - YX$ through the BCH expansion introduced above.

Pairwise noncommutativity, however, captures only first order interactions. Successive compositions may further modulate existing order differences through nested Lie brackets, as illustrated in Fig. 3. We therefore quantify the depth of such interactions by the following notion.

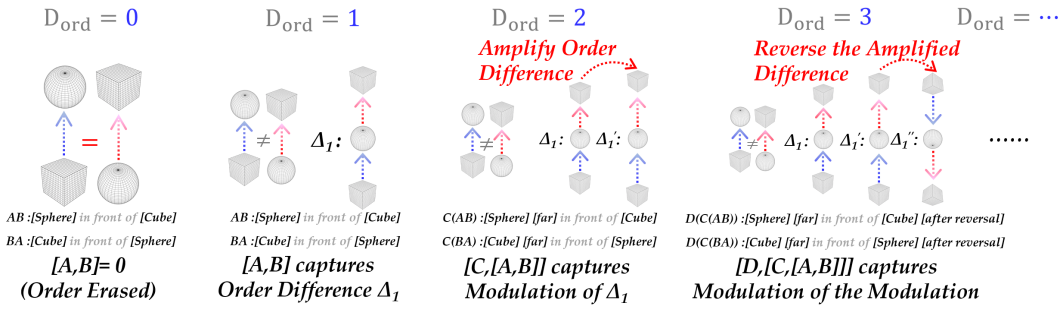


Figure 3: Illustration of order depth. Increasing D_{ord} enables progressively deeper order-dependent interactions, from pairwise order sensitivity to higher level modulation through nested Lie brackets.

Lemma 2.4 (Order depth of $\mathbb{SL}$)**.** *For a Lie group representation space $\mathcal{M}$, define its order depth as*

$$D_{\mathrm{ord}}(\mathcal{M}) := \sup\{k \geq 1 : \exists\, X_0, \ldots, X_k \in T_I\mathcal{M} \text{ such that } [X_k, [\cdots, [X_1, X_0]\cdots]] \neq 0\}. \quad (17)$$

We set $D_{\mathrm{ord}}(\mathcal{M}) = 0$ when all brackets vanish and $D_{\mathrm{ord}}(\mathcal{M}) = \infty$ when nonzero nested brackets exist at arbitrary depth. Then, for every $n \geq 2$ and every $p \in (1, \infty)$,

$$D_{\mathrm{ord}}(\mathbb{SL}_p(n)) = \infty. \quad (18)$$

Proof. The full and detailed proof is in Appendix [D.4]. □

Thus, D_{ord} characterizes the depth at which nested order-dependent interactions can remain nonzero. A nilpotent Lie algebra of class c has $D_{\mathrm{ord}} = c - 1$, whereas $\mathbb{SL}(n)$ has $D_{\mathrm{ord}} = \infty$. This provides algebraic capacity for order-aware composition beyond pairwise noncommutativity. We examine the empirical relevance of this property on Flickr30k-Order in Sec. 3.2.

Table 1: Comparison of geometric representation spaces and their structural properties.

	Mixed curvature				Curvature coupling		Native group composition	
Manifold Space	Mixed	Signs	Intrinsic	Capacity $\mathcal{C}^{\mathrm{mix}}_{\mathcal{M}}$	Coupled	Capacity $\mathcal{C}^{\mathrm{cpl}}_{\mathcal{M}}$	Noncommutative	D_{ord}
$\mathbb{S}^d$	N	$\{+\}$	N/A	N/A	N/A	N/A	N/A	N/A
$\mathbb{H}^d$	N	$\{-\}$	N/A	N/A	N/A	N/A	N/A	N/A
$\mathbb{E}^d$	N	$\{0\}$	N/A	N/A	N/A	N/A	N	0
Grassmann(k,n)	Y	$\{0,+\}$	Y	N/A	N/A	N/A	N/A	N/A
SPD(n)	Y	$\{-,0\}$	Y	N/A	N/A	N/A	N/A	N/A
Siegel(n)	Y	$\{-,0\}$	Y	N/A	N/A	N/A	N/A	N/A
$\mathbb{S}^{d_S}_{\kappa_1} \times \mathbb{H}^{d_H}_{\kappa_2} \times \mathbb{E}^{d_E}$	Y	$\{-,0,+\}$	N	$\min\{d_S - 1, d_H - 1\}$	N	0	N/A	N/A
Heisenberg-H^{2m+1}	Y	$\{-,0,+\}$	Y	1	Y	1	Y	1
$\mathbb{SL}(n)$	Y	$\{-,0,+\}$	Y	$\geq \binom{n-1}{2}$	Y	$\geq \binom{n-1}{2}$	Y	∞

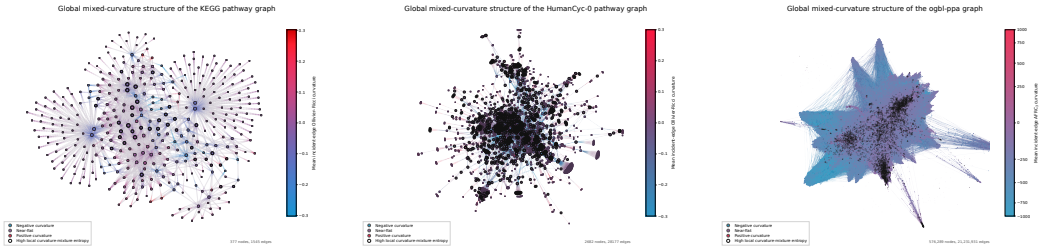
Overall, Table 1 shows that $\mathbb{SL}(n)$ uniquely combines intrinsic $\{-,0,+\}$ mixed-curvature, high curvature-coupling capacity, and unbounded order depth within a single representation space.

3 Experiments

We first examine the practical training ability of $\mathbb{SL}(n)$. The intrinsic Finsler distance has no simple closed form and requires costly path optimization. We therefore use the Schatten semidistance $D_{\mathbb{SL}}(A, B) = \|\log(A^{-1}B)\|_{S_p}$, which locally approximates the intrinsic distance to second order while being substantially cheaper to compute. Optimization presents another challenge, as general Finsler manifolds lack mature adaptive optimization methods such as AdamW. We therefore use an exponential parameterized AdamW scheme for $\mathbb{SL}(n)$, where the Schatten-p geometry enters through the representation objective, and compare it against a Riemannian AdamW control. Full analyses and algorithms are provided in Appendix [B.1][B.2].

For experiments, we evaluate two central properties of $\mathbb{SL}(n)$: its ability to represent mixed curvature structures and its order aware composition induced by noncommutative group multiplication. We first assess geometric representation through graph reconstruction on biological networks of increasing scale, and further evaluate downstream utility through large scale link prediction on OGBL-PPA. We then isolate compositional ability on Flickr30k-Order. Full implementation details, geometric parameterizations, distance functions, and additional results are provided in Appendix [C].

3.1 MIXED-CURVATURE GRAPH REPRESENTATION



(a) KEGG: $|V| = 377$, $|E| = 1{,}545$.

(b) HumanCyc: $|V| = 2{,}682$, $|E| = 28{,}177$.

(c) OGBL: $|V| = 576{,}289$, $|E| = 21{,}231{,}931$.

Figure 4: Graph curvature distributions of three biological networks of increasing scale. All three graphs exhibit mixed-curvature regimes with rich interactions structures.

Metric reconstruction on KEGG and HumanCyc. We evaluate geometric representation capacity by reconstructing the shortest path metrics of KEGG and HumanCyc. Each node is represented by a learnable point, while all methods use the same training pairs, data split, regression objective, and evaluation protocol, differing mainly in the latent geometry and its pairwise dissimilarity. Given graph distance $d_G(i,j)$ and latent dissimilarity $D_{\mathcal{M}}(z_i, z_j)$, all geometries optimize $\mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{(i,j)\in\mathcal{B}}[\log(1 + sD_{\mathcal{M}}(z_i, z_j)) - \log(1 + d_G(i,j))]^2$, where $s > 0$ is a learned global scale and $\mathcal{B}$ is a training minibatch. Geometric fidelity is measured by

$$\text{Distortion}_{\text{avg}} = \frac{1}{|\mathcal{T}|}\sum_{(i,j)\in\mathcal{T}} \frac{|sD_{\mathcal{M}}(z_i, z_j) - d_G(i,j)|}{\max\{d_G(i,j), 1\}}. \tag{19}$$

Here $\mathcal{T}$ denotes the test pairs. Distortion measures deformation of the original graph metric. We further report $q50$, $q90$, and $q95$ for typical and tail errors. Local curvature mixture entropy is defined as $H_\kappa(\mathcal{N}(v)) = -\sum_{s\in\{-,0,+\}} p_v^\kappa(s)\log p_v^\kappa(s)$, where $p_v^\kappa(s) = |\{u \in \mathcal{N}(v) : \kappa_{vu} \in s\}|/|\mathcal{N}(v)|$. Higher entropy indicates stronger local mixing of curvature regimes. High-related measures pairs involving high entropy nodes, Low-Low focuses on pairs between low entropy nodes, and Worst-group reports the largest mean distortion across entropy groups. We first conduct a compact comparison using $\mathbb{SL}(4)$ and commonly used geometries, followed by a higher dimensional comparison. Dimensions are matched as closely as possible; for matrix manifolds, we align matrix size rather than intrinsic dimension. Product manifolds permit mixed-curvature coexistence but have zero coupling, while Heisenberg has both capacities equal to 1. We find that $\mathbb{SL}(n)$ not only improves all reported metrics, but achieves its largest gains in high entropy regions, particularly over mixed-curvature product manifolds and compared with Low-Low pairs. Since product manifolds allow mixed-curvature coexistence without coupling, this concentration of gains in strongly mixed regions suggests that capturing coupled curvature may be important for representation learning.

Table 2: KEGG metric reconstruction results for geometric baselines. (Lower is better)

Space	**Distortion**$_{\text{avg.}}\downarrow$	q50 $\downarrow$	q90 $\downarrow$	q95 $\downarrow$	High-related $\downarrow$	Low–Low $\downarrow$	Worst-group $\downarrow$	MAE $\downarrow$
$\mathbb{S}^{15}$	0.0992±0.0006	0.0636	0.2166	0.3173	0.1125	0.0989	0.1850	0.2522
$\mathbb{H}^{15}$	0.0905±0.0002	0.0589	0.2008	0.2798	0.1113	0.0755	0.1857	0.2341
$\mathbb{E}^{15}$	0.0953±0.0002	0.0625	0.2107	0.3005	0.1114	0.0893	0.1837	0.2435
$\mathbb{S}^5 \times \mathbb{H}^5 \times \mathbb{E}^5$	0.0724±0.0003	0.0427	0.1597	0.2403	0.0933	0.0549	0.1711	0.1849
$\mathbb{S}^5_{\kappa_1} \times \mathbb{H}^5_{\kappa_2} \times \mathbb{E}^5$	0.0613±0.0005	0.0338	0.1404	0.2070	0.0890	0.0325	0.1642	0.1558
$\mathbb{SL}(4)$	0.0586±0.0016	0.0288	0.1238	0.1947	0.0667	0.0556	0.1338	0.1485
$\mathbb{S}^{32} \times \mathbb{H}^{32}$	0.0744±0.0001	0.0454	0.1616	0.2330	0.0986	0.0522	0.1759	0.1887
$\mathbb{S}^{32} \times \mathbb{E}^{32}$	0.0842±0.0002	0.0505	0.1730	0.2775	0.1028	0.0725	0.1765	0.2108
$\mathbb{H}^{32} \times \mathbb{E}^{32}$	0.0706±0.0000	0.0420	0.1550	0.2332	0.0966	0.0464	0.1689	0.1788
$\mathbb{S}^{21}_{\kappa_1} \times \mathbb{H}^{21}_{\kappa_2} \times \mathbb{E}^{21}$	0.0591±0.0002	0.0286	0.1371	0.2013	0.0862	0.0295	0.1683	0.1502
Heisenberg-H^{31}	0.0814±0.0001	0.0526	0.1764	0.2432	0.1088	0.0572	0.1804	0.2073
$\mathbf{Gr}(7,16)$	0.0879±0.0010	0.0516	0.1787	0.2926	0.1088	0.0763	0.1836	0.2190
$\mathbf{SPD}(8)$	0.0842±0.0004	0.0550	0.1821	0.2526	0.1095	0.0620	0.1805	0.2160
$\mathbf{Siegel}(8)$	0.0601±0.0004	0.0328	0.1339	0.2109	0.0789	0.0454	0.1539	0.1475
$\mathbb{SL}(8)$	**0.0329**±0.0001	**0.0115**	**0.0698**	**0.1207**	**0.0389**	**0.0290**	**0.0841**	**0.0811**
Improvement	**↑44.3%**	**↑59.8%**	**↑47.9%**	**↑40.0%**	**↑50.7%**	**↑1.7%**	**↑45.4%**	**↑45.0%**

Table 3: HumanCyc metric reconstruction results for geometric baselines.

Space	**Distortion**$_{\text{avg.}}$ ↓	q50 ↓	q90 ↓	q95 ↓	High-related ↓	Low–Low ↓	Worst-group ↓	MAE ↓
$\mathbb{S}^{72}\times\mathbb{H}^{72}$	0.0649±0.0001	0.0450	0.1272	0.1771	0.0689	0.0569	0.0841	0.2379
$\mathbb{S}^{72}\times\mathbb{E}^{72}$	0.0682±0.0000	0.0480	0.1308	0.1853	0.0701	0.0671	0.0842	0.2514
$\mathbb{H}^{72}\times\mathbb{E}^{72}$	0.0655±0.0001	0.0453	0.1298	0.1803	0.0697	0.0567	0.0852	0.2421
$\mathbb{S}^{48}_{\kappa_1}\times\mathbb{H}^{48}_{\kappa_2}\times\mathbb{E}^{48}$	0.0660±0.0000	0.0464	0.1294	0.1771	0.0702	0.0559	0.0849	0.2433
$\mathbf{Gr}(11,24)$	0.0693±0.0001	0.0483	0.1334	0.1892	0.0699	0.0746	0.0835	0.2557
$\mathbf{SPD}(12)$	0.0698±0.0001	0.0495	0.1347	0.1892	0.0719	0.0679	0.0859	0.2586
$\mathbf{Siegel}(12)$	0.0526±0.0001	0.0332	0.1070	0.1534	0.0557	0.0492	0.0708	0.1899
$\mathbb{SL}(12)$	**0.0313**±0.0003	**0.0153**	**0.0678**	**0.1019**	**0.0329**	**0.0309**	**0.0360**	**0.1098**
Improvement	**↑40.5%**	**↑53.9%**	**↑36.7%**	**↑33.6%**	**↑40.9%**	**↑37.2%**	**↑49.1%**	**↑42.2%**

Large scale link prediction on OGBL-PPA. We further test whether the geometric representation advantage transfers to a downstream task on OGBL-PPA, a substantially larger protein association graph. All methods also use the same experimental settings, while varying only the latent representation space. This controlled setting isolates how well each geometry organizes nodes for recovering unseen links. Table 4 evaluates complementary aspects of ranking quality. Hits@K measures how often positive edges appear among the top ranked candidates, MRR summarizes reciprocal rank, and Rank50, Rank90, and Rank95 characterize the typical and tail ranks of positive edges. AUC and AP further measure global discrimination between positive and negative pairs. $\mathbb{SL}(8)$ improves all reported metrics, with particularly consistent gains across both Hits and rank quantiles. This indicates that its advantage is not limited to a particular ranking threshold, but extends across the ranking distribution and positive negative separation. Together with metric reconstruction, these results suggest that the geometry learned by $\mathbb{SL}(n)$ supports both faithful graph representation and downstream relational prediction at substantially larger scale.

Table 4: OGBL-PPA link prediction results for geometric latent spaces.

Space	**Hits@20** ↑	**Hits@50** ↑	**Hits@100** ↑	Rank50 ↓	Rank90 ↓	Rank95 ↓	MRR ↑	AUC ↑	AP ↑
$\mathbb{S}^{32}\times\mathbb{H}^{32}$	0.0888±0.0050	0.1558±0.0143	0.2332±0.0144	803	30060	90715	0.0157	0.9915	0.9926
$\mathbb{S}^{32}\times\mathbb{E}^{32}$	0.0725±0.0129	0.1454±0.0127	0.2102±0.0096	931	37757	111444	0.0140	0.9903	0.9915
$\mathbb{H}^{32}\times\mathbb{E}^{32}$	0.0841±0.0036	0.1303±0.0022	0.1893±0.0089	1248	90588	279414	0.0140	0.9811	0.9844
$\mathbb{H}^{21}\times\mathbb{E}^{21}\times\mathbb{S}^{21}$	0.0898±0.0171	0.1556±0.0094	0.2185±0.0101	920	40199	118889	0.0155	0.9899	0.9912
$\mathbb{H}^{21}_{\kappa_1}\times\mathbb{E}^{21}\times\mathbb{S}^{21}_{\kappa_2}$	0.0809±0.0094	0.1494±0.0118	0.2217±0.0090	990	48604	144410	0.0145	0.9883	0.9899
$\mathbf{Gr}(7,16)$	0.0775±0.0097	0.1336±0.0115	0.1808±0.0054	1588	119504	397421	0.0157	0.9748	0.9802
$\mathbf{SPD}(8)$	0.0784±0.0153	0.1386±0.0122	0.1964±0.0086	1370	52211	147717	0.0151	0.9881	0.9896
$\mathbf{Siegel}(8)$	0.0769±0.0295	0.1462±0.0140	0.2183±0.0095	998	64364	198532	0.0134	0.9855	0.9877
$\mathbb{SL}(8)$	**0.1282**±0.0056	**0.2188**±0.0055	**0.3093**±0.0119	**346**	**10700**	**36145**	**0.0195**	**0.9950**	**0.9957**
Improvement	**↑42.8%**	**↑40.4%**	**↑32.6%**	**↑57.0%**	**↑64.4%**	**↑60.2%**	**↑24.1%**	**↑0.3%**	**↑0.3%**

3.2 Deep Order-Aware Composition



Positive 0: a [dozen workers] wearing uniforms and sanitation hats are [working on an assembly line] in [a factory]

Negative 1: a [dozen workers] wearing uniforms and sanitation hats are working on an [line assembly] in [a factory]

Negative 2: a [dozen workers] wearing uniforms and sanitation hats are [assembly on an working line] in [a factory]

Negative 3: a [are workers] wearing uniforms and sanitation hats [dozen working] on an assembly line in [a factory]

Negative 4: a [dozen workers] wearing uniforms and sanitation hats are [working on an assembly line] in [factory a]

Figure 5: Flickr30k-Order: The positive caption preserves the original word order, while the negative captions perturb local phrase order with nearly the same bag of words.

Order sensitive composition on Flickr30k-Order. We finally isolate the compositional property of $\mathbb{SL}(n)$ on Flickr30k-Order, where performance depends on preserving semantic order. Order Accuracy and Order Margin measure order discrimination, while Hard Accuracy and MRR evaluate harder and ranking based cases. For a controlled comparison, we freeze the same CLIP backbone and train only a lightweight group specific head for each representation space. We compare additive models with noncommutative groups of increasing effective order depth D_{ord}. Heisenberg and the unitriangular group $\mathrm{UT}(n)$ have finite depth, whereas $\mathbb{SL}(4)$ has $D_{\text{ord}} = \infty$. Performance generally improves with larger D_{ord}, and ordered $\mathbb{SL}(4)$ performs best across all metrics, while removing

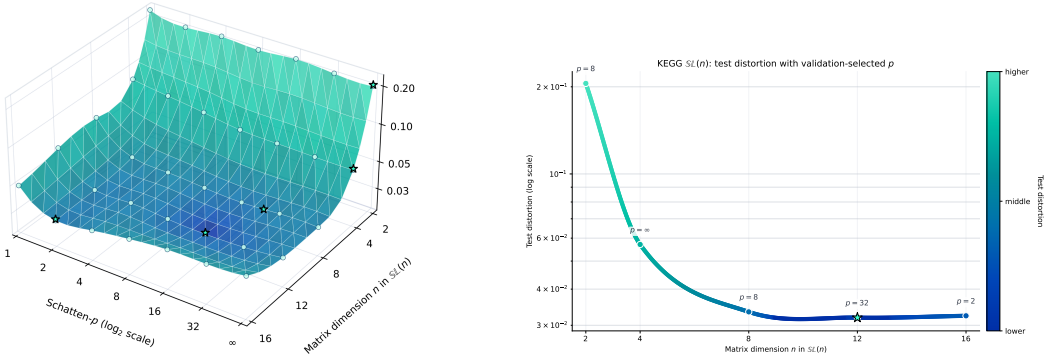
ordered composition causes a large drop. This supports deep noncommutative composition as an advantage of $\mathbb{SL}(n)$ beyond its latent geometry.

Table 5: Order sensitive composition on ARO Flickr30k-Order.

Model	Dim.	D_{ord}	Composition	**OrderAcc.** ↑	OrderMargin ↑	HardAcc. ↑	MRR ↑
BoW	–	0	Additive	47.82±1.96	−0.0000	41.52	0.561
CLIP	–	–	Implicit	86.08±0.00	0.0176	66.07	0.805
$\mathbb{SL}(4)$ w/o Ordered Comp.	15	0	Commutative addition	17.97±3.14	−0.0079	3.54	0.272
Heisenberg-H^7	15	1	Group matrix multiplication	94.28±0.02	0.6324	82.32	0.906
UT(4)	6	2	Group matrix multiplication	94.11±0.27	0.5804	81.64	0.903
UT(6)	15	4	Group matrix multiplication	96.30±0.18	0.8862	89.11	0.942
UT(15)	105	13*	Group matrix multiplication	97.25±0.06	1.1683	92.13	0.957
Full $\mathbb{SL}(4)$ w/ Ordered Comp.	15	∞	**Group matrix multiplication**	**97.46**±0.07	**1.1934**	**92.87**	**0.962**

*For dimension $15 = \dim \mathbb{SL}(4)$, the maximal finite D_{ord} of a nilpotent Lie algebra is 13.

3.3 Ablation Study and Sensitivity Test



(a) Schatten-p sensitivity across $\mathbb{SL}(n)$ dimensions.

(b) Dimension ablation with validation-optimal Schatten-p.

Figure 6: Sensitivity and dimension ablation studies for $\mathbb{SL}(n)$ on KEGG.

Figure 6 shows a clear interaction between matrix dimension n and Schatten order p. Increasing n enlarges representation capacity, whereas larger p places greater emphasis on dominant singular directions. The preferred p varies substantially with n, indicating that these two hyperparameters control different aspects of the geometry, namely representation capacity and directional sensitivity, and should be tuned jointly. Test distortion improves rapidly up to $n = 8$–12 and then saturates. Table 6 further shows that at $p = 16$, increasing n from 12 to 16 reduces training loss by about $109\times$ while worsening test distortion by 33.5%, showing that additional capacity can reduce training error without improving test performance.

Table 6: Representative training loss and test distortion on KEGG.

	$n = 12$		$n = 16$	
p	Train loss	Test dist.	Train loss	Test dist.
2	8.05×10^{-7}	0.03279	1.89×10^{-7}	0.03149
16	1.25×10^{-5}	0.03026	1.15×10^{-7}	0.04040

4 Conclusion

$\mathbb{SL}(n)$ provides a single representation space combining coupled mixed-curvature with deep order-aware composition. Remarkably, this richness emerges from a minimal construction consisting only of the $\det(A) = 1$ constraint and a simple left invariant Schatten-p tangent norm. Its Finsler geometry realizes curvature signs $\{-, 0, +\}$ with asymptotically maximal mixed-curvature and curvature-coupling capacities, while its non-nilpotent Lie algebra structure supports noncommutative interactions at arbitrary depth. Thus, rich geometric coexistence, intrinsic interaction, and deep composition need not rely on separate representation components. The consistent gains across metric reconstruction, large scale link prediction, and ordered composition show that this structural simplicity preserves expressive power. Together, these results establish $\mathbb{SL}(n)$ as a general structured latent space for intrinsically coupled mixed-curvature geometry and deep composition, illustrating how simple structural constraints can yield unexpectedly rich representations, with potential applications across geometric, relational, sequential, multimodal, and scientific representation learning.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for none of the tasks with required disclosure. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for aiding or polishing writing and for retrieval and discovery of related work. We have reviewed all AI-assisted work. We reviewed and edited all AI-assisted writing, and independently checked the references and related work identified with AI assistance against the original sources for relevance and accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work is primarily methodological and theoretical. Our experiments use existing benchmark datasets and do not involve new data collection or user studies. We do not identify any specific ethical concerns beyond those generally associated with the use of existing datasets and pretrained models.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We provide detailed materials to support the reproducibility of both our theoretical and empirical results. Complete derivations and proofs of the main theoretical results are provided in Appendix [D]. Appendix [B] gives additional details on the Schatten semidistance, optimization parameterization, and numerical analyses, while Appendix [C] describes the datasets and splits, baseline configurations, training and optimization settings, model-selection protocol, hardware, and random seed setup. Unless otherwise stated, experimental results are reported over three independent model initialization seeds, with hyperparameters and checkpoints selected using validation data only. Anonymous source code and configuration files are provided in the supplementary materials.

REFERENCES

Gregor Bachmann, Gary Becigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 486–496. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bachmann20a.html.

D. Bao, S.-S. Chern, and Z. Shen. *The Chern Connection*, pp. 27–48. Springer New York, New York, NY, 2000. ISBN 978-1-4612-1268-3. doi: 10.1007/978-1-4612-1268-3_2. URL https://doi.org/10.1007/978-1-4612-1268-3_2.

Thomas Bendokat, Ralf Zimmermann, and P.-A. Absil. A grassmann manifold handbook: basic geometry and computational aspects. *Adv. Comput. Math.*, 50(1), January 2024. ISSN 1019-7168. doi: 10.1007/s10444-023-10090-8. URL https://doi.org/10.1007/s10444-023-10090-8.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0415740eaa4d9decbc8da001d3fd805f-Paper.pdf.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6901–6914, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.617. URL https://aclanthology.org/2020.acl-main.617/.

Thomas Dagès, Simon Weber, Ya-Wei Eileen Lin, Ronen Talmon, Daniel Cremers, Michael Lindenbaum, Alfred M. Bruckstein, and Ron Kimmel. Finsler multi-dimensional scaling: Manifold learning for asymmetric dimensionality reduction and embedding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25842–25853. IEEE, 2025. doi: 10.1109/CVPR52734.2025.02407. URL https://doi.org/10.1109/CVPR52734.2025.02407.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=NPOWF_ZLfC5.

Willem Diepeveen, Georgios Batzolis, Zakhar Shumaylov, and Carola-Bibiane Schönlieb. Score-based pullback Riemannian geometry: Extracting the data manifold geometry using anisotropic flows. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13746–13773. PMLR, 2025.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJxeWnCcF7.

Zihao Guo, Qingyun Sun, Haonan Yuan, Xingcheng Fu, Min Zhou, Yisen Gao, and Jianxin Li. GraphMoRE: Mitigating topological heterogeneity via mixture of riemannian experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11754–11762, 2025. doi: 10.1609/aaai.v39i11.33279. URL https://doi.org/10.1609/aaai.v39i11.33279.

Brian C. Hall. The baker–campbell–hausdorff formula and its consequences. In *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*, pp. 109–137. Springer International Publishing, Cham, 2 edition, 2015. ISBN 978-3-319-13467-3. doi: 10.1007/978-3-319-13467-3_5. URL https://doi.org/10.1007/978-3-319-13467-3_5.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/abs/10.1126/science.1127647.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933. doi: 10.1037/h0071325.

Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. Deep learning on Lie groups for skeleton-based action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1243–1252. IEEE, 2017. doi: 10.1109/CVPR.2017.137. URL https://doi.org/10.1109/CVPR.2017.137.

Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on grassmann manifolds, 2018. URL https://arxiv.org/abs/1611.05742.

Michael J. Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. LieTransformer: Equivariant self-attention for Lie groups. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4533–4543. PMLR, 2021. URL https://proceedings.mlr.press/v139/hutchinson21a.html.

Miguel Angel Javaloyes. Chern connection of a pseudo-finsler metric as a family of affine connections, 2014. URL https://arxiv.org/abs/1303.6263.

Miguel Ángel Javaloyes. Anisotropic tensor calculus. *International Journal of Geometric Methods in Modern Physics*, 16(Supplement 2):1941001, 2019. doi: 10.1142/S0219887819410019. URL https://doi.org/10.1142/S0219887819410019.

Chankyo Kim, Sicheng Zhao, Minghan Zhu, Tzu-Yuan Lin, and Maani Ghaffari. Equivariant neural networks for general linear symmetries on lie algebras, 2026. URL https://arxiv.org/abs/2510.22984.

Hannah Lawrence and Mitchell Tong Harris. Learning polynomial problems with $\mathrm{SL}(2,\mathbb{R})$-equivariance. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gyfXuRfxW2.

John M. Lee. Submanifolds. In *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*, pp. 98–124. Springer, New York, NY, 2 edition, 2013. ISBN 978-1-4419-9982-5. doi: 10.1007/978-1-4419-9982-5_5. URL https://doi.org/10.1007/978-1-4419-9982-5_5.

Tzu-Yuan Lin, Minghan Zhu, and Maani Ghaffari. Lie neurons: Adjoint-equivariant neural networks for semisimple lie algebras, 2024. URL https://arxiv.org/abs/2310.04521.

Federico Lopez, Beatrice Pozzetti, Steve Trettel, Michael Strube, and Anna Wienhard. Symmetric spaces for graph embeddings: A Finsler-Riemannian approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7090–7101. PMLR, 2021.

Daniel McNeela, Frederic Sala, and Anthony Gitter. Product manifold representations for learning on biological pathways, 2025. URL https://arxiv.org/abs/2401.15478.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mircea Mironenco and Patrick Forré. Lie group decompositions for equivariant neural networks, 2024. URL https://arxiv.org/abs/2310.11366.

Tuc Nguyen-Van, Dung D. Le, and The-Anh Ta. Improving heterogeneous graph learning with weighted mixed-curvature product manifold, 2023. URL https://arxiv.org/abs/2307.04514.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

T Mitchell Roddenberry and Richard Baraniuk. Finsler geometry, graph neural networks, and you. In *Topology, Algebra, and Geometry in Data Science*, 2026. URL https://openreview.net/forum?id=K9EJoCht9k.

Haitz Sáez de Ocáriz Borde. On the expressive power of mixed-curvature representations in product manifolds. In *Workshop on Geometry-grounded Representation Learning and Generative Modeling at ICLR*, 2026.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g6xeSKDS.

Li Sun, Zhongbao Zhang, Junda Ye, Hao Peng, Jiawei Zhang, Sen Su, and Philip S. Yu. A self-supervised mixed-curvature graph neural network. *CoRR*, abs/2112.05393, 2021. URL https://arxiv.org/abs/2112.05393.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2319.

Nam-Kiu Tsing, Michael K.H. Fan, and Erik I. Verriest. On analyticity of functions involving eigenvalues. *Linear Algebra and its Applications*, 207:159–180, 1994. ISSN 0024-3795. doi: 10.1016/0024-3795(94)90009-4. URL https://www.sciencedirect.com/science/article/pii/0024379594900094.

Jihu Wang, Yuliang Shi, Han Yu, Xinjun Wang, Zhongmin Yan, and Fanyu Kong. Mixed-curvature manifolds interaction learning for knowledge graph-aware recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 372–382, 2023. doi: 10.1145/3539618.3591730.

Shen Wang, Xiaokai Wei, Cicero Nogueira dos Santos, Zhiguo Wang, Ramesh Nallapati, Andrew Arnold, Bing Xiang, Philip S. Yu, and Isabel F. Cruz. Mixed-curvature multi-relational graph neural network for knowledge graph completion. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 1761–1771, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450118. URL https://doi.org/10.1145/3442381.3450118.

Ming Xu and Shaoqiang Deng. Homogeneous finsler spaces and the flag-wise positively curved condition. *Forum Mathematicum*, 30(6):1521–1537, 2018. doi: 10.1515/forum-2018-0130.

Xin Yang, Xingrun Li, Heng Chang, Yang jinze, Xihong Yang, Shengyu Tao, Maiko Shigeno, Ningkang Chang, Junfeng Wang, Dawei Yin, and Erxue Min. Hgformer: Hyperbolic graph transformer for collaborative filtering. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=lNTHwNF8IH.

Wei Zhao, Federico Lopez, J. Maxwell Riestenberg, Michael Strube, Diaaeldin Taha, and Steve Trettel. Modeling graphs beyond hyperbolic: Graph neural networks in symmetric positive definite matrices. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, volume 14171 of *Lecture Notes in Computer Science*, pp. 122–139. Springer, 2023. doi: 10.1007/978-3-031-43418-1_8.

APPENDIX

A RELATED WORK AND DISCUSSION

A.1 RELATED WORK

Mixed-curvature and manifold representation learning. Non-Euclidean representation learning uses the geometry of the latent manifold as an inductive bias for structured data. Constant-curvature manifolds provide three canonical geometric regimes: Euclidean spaces model approximately flat structures, hyperbolic spaces naturally accommodate hierarchical and tree-like structures, while spherical spaces provide compact positively curved geometry for structures with cyclic or globally constrained relations (Nickel & Kiela, 2017; Chami et al., 2019; Bachmann et al., 2020). These geometries have been applied broadly to knowledge graphs, recommendation systems, generative modeling, graph learning, and visual representation (Chami et al., 2020; Skopek et al., 2020; Yang et al., 2025). Since real data often contain structures that cannot be captured by one curvature regime, product manifolds combine Euclidean, hyperbolic, and spherical components to form mixed-curvature representation spaces (Gu et al., 2019; Skopek et al., 2020). Such spaces have been further developed for graph learning and biological networks (Sun et al., 2021; McNeela et al., 2025). Beyond constant-curvature factors, matrix and higher rank manifolds including SPD, Grassmann, and Siegel spaces provide richer intrinsic geometries for covariance, subspace, and graph representations (Huang et al., 2018; Lopez et al., 2021). A complementary line of work adapts geometry through the learning architecture, for example by learning curvature parameters, weighting multiple manifold components, or dynamically selecting and combining geometric experts (Bachmann et al., 2020; Sun et al., 2021; Nguyen-Van et al., 2023; Guo et al., 2025). Our work instead focuses on the underlying representation manifold. Compared with existing geometric spaces, $\mathbb{SL}(n)$ provides intrinsic mixed curvature and curvature coupling within a single manifold, while its Lie group structure additionally supports noncommutative composition. Because these properties arise from the representation space itself rather than from a specialized geometry-learning architecture, $\mathbb{SL}(n)$ can serve as a general latent space beyond curvature-adaptive models. Accordingly, our controlled experiments vary the representation manifold while keeping the surrounding learning framework fixed.

Finsler geometry and Lie structures in machine learning. Finsler geometry extends Riemannian geometry by allowing the tangent norm to depend on direction, and has been used in representation learning to enrich the geometry available beyond standard Riemannian metrics. In particular, Finsler metrics on symmetric spaces have been studied for graph embeddings (Lopez et al., 2021), while more recent work has explored Finsler geometry for asymmetric embedding and graph learning (Dagès et al., 2025; Roddenberry & Baraniuk, 2026). Lie groups and Lie algebras have been widely used in machine learning for geometric representation, continuous symmetries, and equivariant architectures (Huang et al., 2017; Finzi et al., 2020; Hutchinson et al., 2021; Dehmamy et al., 2021). More recent work has extended this direction to noncompact and semisimple groups, including architectures involving $\mathrm{SL}(2,\mathbb{R})$, $\mathrm{SL}(n,\mathbb{R})$, and their Lie algebras (Lawrence & Harris, 2024; Mironenco & Forré, 2024; Lin et al., 2024; Kim et al., 2026). Many of these approaches exploit Lie groups and Lie algebras as symmetry groups, transformation domains, or algebraic structures for constructing equivariant networks. In contrast, we do not use $\mathrm{SL}(n)$ only as a symmetry acting on external features. We make $\mathbb{SL}(n)$ itself the latent representation space, where the left invariant Schatten-p Finsler structure determines intrinsic mixed curvature and curvature coupling, while the group product and Lie algebra provide noncommutative and higher order composition. Thus, our use of Lie structure extends beyond symmetry and equivariance, with geometry and composition jointly defining the underlying representation space.

A.2 Discussion and Limitations

A.2.1 Discussion about applications

A particularly appealing aspect of $\mathbb{SL}(n)$ is that a remarkably simple structural constraint gives rise to a surprisingly rich representation space. Defined only by the determinant-one constraint and equipped with a left-invariant Schatten-p geometry, $\mathbb{SL}(n)$ simultaneously exhibits intrinsic mixed curvature, strong curvature coupling, noncommutative group composition, and arbitrarily deep nested Lie-bracket interactions. These properties arise from a single underlying structure rather than from separately designed geometric or compositional components. Also, $\mathbb{SL}(n)$ can be used as a plug-and-play embedding space. Existing encoders can simply output matrix-valued latent representations in $\mathbb{SL}(n)$, allowing rich geometry and composition to be introduced without redesigning the surrounding architecture.

One important direction is **relational and sequential learning**. In knowledge graphs, relations are naturally compositional and often asymmetric, while the underlying relational structure may simultaneously contain hierarchical, cyclic, and densely interconnected patterns. An $\mathbb{SL}(n)$ embedding can therefore use its mixed-curvature geometry to capture heterogeneous structural organization, while its noncommutative group operation provides a native mechanism for ordered relation composition and multi hop reasoning. Similar considerations arise in recommender systems and behavioral modeling, where the meaning of an interaction sequence often depends not only on which events occur, but also on their order and accumulated effect. The ability of $\mathbb{SL}(n)$ to support deep order-sensitive composition may therefore provide a useful inductive bias for long-range relational and sequential dependencies.

Another promising direction is **multimodal, temporal, and dynamical representation learning**. Different modalities may exhibit distinct local geometries while still requiring structured alignment within a shared latent space, while temporal observations can often be viewed as transformations accumulated over time. Because $\mathbb{SL}(n)$ acts primarily as an embedding space rather than a task-specific architecture, it can be incorporated with minimal modification into existing GNNs, Transformers, state space models, and multimodal encoders by replacing or augmenting their latent representation layer. This plug-and-play nature makes it possible to introduce intrinsic geometric heterogeneity and structured composition into a broad range of architectures without designing separate mechanisms for each application.

A.2.2 Limitations and Future Directions

Our experiments are designed to isolate the effect of the underlying representation space and therefore do not explore architectures specifically optimized for $\mathbb{SL}(n)$. The results show that the space works effectively within controlled frameworks, while dedicated $\mathbb{SL}(n)$ layers may better exploit its geometric and compositional structure. Another limitation concerns distance computation. The closed form semidistance used in our scalable experiments is restricted to the principal logarithm domain, whereas exact intrinsic path distances are more expensive. We partially address this through path approximations, while globally robust distance constructions remain an important direction.

Matrix valued representations also incur higher computational cost than vector embeddings due to matrix multiplication, matrix logarithms, and Schatten norm evaluations. This is a general challenge for matrix manifold methods rather than one specific to $\mathbb{SL}(n)$, while group multiplication and inversion remain standard matrix operations. Future work may reduce this cost through low rank Lie algebra updates, approximate logarithms, and efficient structured parameterizations.

Integrating $\mathbb{SL}(n)$ with larger architectures is a natural next step. Dedicated $\mathbb{SL}(n)$ layers may allow models to exploit mixed curvature, intrinsic coupling, and noncommutative composition jointly rather than using the space only as an embedding domain. Extending this perspective to other matrix Lie groups may reveal how geometry and algebra match structural priors. Promising applications include foundation models, knowledge graphs, multimodal and sequential learning, and scientific representation learning.

B Distance and Optimization Analysis

B.1 Schatten Semidistance

B.1.1 Relation between the Schatten Semidistance and Intrinsic Distance

The Schatten semidistance $D_{\mathbb{SL}}$ is used for pairwise comparison in $\mathbb{SL}_p(n)$, while the Finsler metric F_p induces the intrinsic geodesic distance d_p. We first establish their local relation.

Lemma B.1 (Second-order tightness of the Schatten semidistance)**.** *For every $1 < p < \infty$, there exist local constants $\delta, C > 0$ such that, for any $A, B \in \mathbb{SL}_p(n)$ in the principal-logarithm domain with $D_{\mathbb{SL}}(A, B) < \delta$,*

$$0 \le D_{\mathbb{SL}}(A, B) - d_p(A, B) \le C\, D_{\mathbb{SL}}(A, B)^2. \tag{20}$$

Consequently, $D_{\mathbb{SL}}(A, B) = d_p(A, B) + O(D_{\mathbb{SL}}(A, B)^2)$, and equivalently $D_{\mathbb{SL}}(A, B) = d_p(A, B) + O(d_p(A, B)^2)$ as $B \to A$.

Proof. Let $X = \log(A^{-1}B)$. In the principal-logarithm domain, $\log(B^{-1}A) = -X$, so $D_{\mathbb{SL}}(A, B) = \|X\|_{S_p}$. The exponential path $\gamma(t) = A\exp(tX)$ has constant left-trivialized velocity X and length $\|X\|_{S_p}$. Hence

$$d_p(A, B) \le D_{\mathbb{SL}}(A, B).$$

Conversely, let γ be any sufficiently short piecewise-smooth curve from A to B, and write $L = L_p(\gamma) = \int_0^1 \|\gamma(t)^{-1}\dot\gamma(t)\|_{S_p}\, dt$. The Magnus expansion of the endpoint $A^{-1}B$ gives

$$X = \int_0^1 \gamma(t)^{-1}\dot\gamma(t)\, dt + \frac{1}{2}\int_0^1 \int_0^{t_1} \left[\gamma(t_2)^{-1}\dot\gamma(t_2), \gamma(t_1)^{-1}\dot\gamma(t_1)\right] dt_2\, dt_1 + O(L^3).$$

The first term is linear in the path velocity, whereas every remaining term contains at least two such factors. Since $\|[U, V]\|_{S_p} \le 2\|U\|_{S_p}\|V\|_{S_p}$, the second term is $O(L^2)$, and local convergence of the Magnus series gives, for sufficiently small L,

$$\|X\|_{S_p} \le L + CL^2.$$

Choose $\delta > 0$ small enough that this estimate is valid whenever $L < 2\delta$. If $D_{\mathbb{SL}}(A, B) < \delta$, then the first inequality gives $d_p(A, B) < \delta$. Therefore, a minimizing sequence of curves has $L < 2\delta$ for all sufficiently late terms. Applying the preceding estimate and letting $L \to d_p(A, B)$ yields

$$D_{\mathbb{SL}}(A, B) \le d_p(A, B) + C\, d_p(A, B)^2.$$

Combining this with $d_p(A, B) \le D_{\mathbb{SL}}(A, B)$ gives

$$0 \le D_{\mathbb{SL}}(A, B) - d_p(A, B) \le C\, d_p(A, B)^2 \le C\, D_{\mathbb{SL}}(A, B)^2.$$

Thus $D_{\mathbb{SL}}(A, B) = d_p(A, B) + O(D_{\mathbb{SL}}(A, B)^2)$. Since the two distances are locally equivalent, the remainder may equivalently be written as $O(d_p(A, B)^2)$. □

Empirical correlation with numerical geodesic distance. We further examine whether this local agreement extends to representations encountered in practice. On KEGG, we randomly sample 100 learned representation pairs and compute a high-accuracy numerical reference $\widehat{d}_p$ by geodesic path optimization under F_p. We compare $D_{\mathbb{SL}}$ with $\widehat{d}_p$ using their mean relative discrepancy, Pearson correlation, and Spearman rank correlation. We additionally record the computation time of the numerical reference. 100 pairs are randomly sampled on KEGG. The last column reports the slowdown of numerical geodesic computation relative to $D_{\mathbb{SL}}$.

Across all p, $D_{\mathbb{SL}}$ remains strongly correlated with the numerical geodesic reference, with Pearson correlations of 0.94–0.99 and Spearman correlations of 0.94–0.98. The mean relative discrepancy decreases from 27.11% at $p = 2$ to 11.80% at $p = 32$, while the numerical geodesic computation requires hundreds of seconds per pair and is approximately 1.6×10^5–2.5×10^5 times slower than $D_{\mathbb{SL}}$. These results show that the closed-form Schatten semidistance preserves both the magnitude and ranking structure of the intrinsic geometry at a fraction of the computational cost.

Table 7: Correlation between the Schatten semidistance $D_{\mathbb{SL}}$ and numerical geodesic distances.

p	Rel. Diff. ↓	Pearson ↑	Spearman ↑	Ref. Time / Pair	Slowdown
2	27.11%	0.9407	0.9401	473–488 s	$(1.95–2.01) \times 10^5$
4	18.58%	0.9553	0.9526	389–464 s	$(1.61–1.92) \times 10^5$
8	15.08%	0.9856	0.9766	478–516 s	$(1.97–2.13) \times 10^5$
16	13.61%	0.9886	0.9777	593 s	2.45×10^5
32	11.80%	0.9831	0.9454	380–565 s	$(1.57–2.33) \times 10^5$

B.1.2 Geodesic Path Ablation and Logarithm Robustness

The Schatten semidistance $D_{\mathbb{SL}}$ admits a direct interpretation under the same Schatten-p length structure used to define the geometry. For $X = \log(A^{-1}B)$, the canonical exponential path $\gamma(t) = A\exp(tX)$ satisfies

$$L_p(\gamma) = \int_0^1 \|\gamma(t)^{-1}\dot{\gamma}(t)\|_{S_p}\, dt = \|X\|_{S_p} = D_{\mathbb{SL}}(A, B), \tag{21}$$

whenever the principal logarithm is well defined. Hence, $D_{\mathbb{SL}}$ is exactly the F_p-length of a canonical admissible path rather than an unrelated pairwise objective, and consequently $d_p(A, B) \leq D_{\mathbb{SL}}(A, B)$. Lemma B.1 further shows that this upper bound is second-order tight locally. To examine whether the empirical performance depends specifically on this single exponential path, we introduce the K-segment piecewise-exponential path approximation

$$\widehat{d}_p^{(K)}(A, B) = \inf_{\substack{G_0=A,\, G_K=B \\ G_1,\ldots,G_{K-1}\in\mathbb{SL}_p(n)}} \sum_{k=0}^{K-1} D_{\mathbb{SL}}(G_k, G_{k+1}). \tag{22}$$

For $K = 1$, this reduces exactly to the original Schatten semidistance,

$$\widehat{d}_p^{(1)}(A, B) = D_{\mathbb{SL}}(A, B), \tag{23}$$

whereas increasing K allows increasingly flexible piecewise-exponential paths and therefore provides progressively tighter numerical approximations to the intrinsic path distance.

We evaluate this effect on a connected 64-node subgraph of KEGG using $\mathbb{SL}_2(4)$ and three random seeds. We perform end-to-end training with $K = 1$ and $K = 2$, while $K = 4$ and $K = 8$ are used for numerical path refinement on frozen learned embeddings. All other model, optimization, initialization, and data-split settings are held fixed between $K = 1$ and $K = 2$ for each seed. In the following, "Gap to K=8" denotes the mean relative discrepancy to the numerically stabilized $K = 8$ reference evaluated on the same learned embedding.

Table 8: End-to-end path-objective ablation on KEGG using $\mathbb{SL}_2(4)$.

K	Gap to K=8 ↓	Test Distortion ↓	Relative Runtime ↓
1	0.02146 ± 0.00068	0.17752 ± 0.00293	1.00 ± 0.00
2	$\mathbf{0.00502 \pm 0.00016}$	$\mathbf{0.17671 \pm 0.00190}$	5.48 ± 0.20

As shown in Table 8, replacing the single exponential path by a two-segment path substantially tightens the numerical path approximation. The discrepancy to the $K = 8$ reference decreases from 0.02146 to 0.00502, corresponding to a 76.6% reduction. In contrast, the resulting representation performance changes only marginally: test distortion improves from 0.17752 to 0.17671, a relative improvement of approximately 0.46%, while training becomes approximately 5.48× more expensive. Thus, substantially refining the pairwise path geometry produces only a minor change in the learned representation quality, while incurring a considerably larger computational cost. Results are reported as mean ± standard deviation. Runtime is normalized by the $K = 1$ setting.

Path refinement on frozen embeddings. We further evaluate $K \in \{1, 2, 4, 8\}$ on the six learned embeddings obtained from the $K = 1$ and $K = 2$ training runs. This yields 2,424 held-out pairwise observations and isolates the numerical effect of path refinement from changes in the learned representation. The gap is measured relative to the $K = 8$ reference for the same pair.

Table 9: Piecewise-path refinement on frozen KEGG embeddings.

K	Mean Gap ↓	p95 Gap ↓	Max Gap ↓
1	0.02182	0.04626	0.09072
2	0.00494	0.01049	0.01909
4	0.00098	0.00207	0.00373
8	0	0	0

The refinement is fully consistent with the expected path hierarchy: across all $2{,}424$ observations, we obtain

$$\widehat{d}_p^{(8)} \le \widehat{d}_p^{(4)} \le \widehat{d}_p^{(2)} \le \widehat{d}_p^{(1)}, \tag{24}$$

with no observed violations. Moreover, the discrepancy to the $K = 8$ reference decreases rapidly, from 0.02182 at $K = 1$ to 0.00494 at $K = 2$ and 0.00098 at $K = 4$.

We additionally verify the numerical stability of the $K = 8$ reference. Increasing the optimization budget from $1{,}000$ to $2{,}000$ iterations changes the final objective by only 1.82×10^{-12} on average, and independent deterministic restarts exhibit an average relative spread of 1.83×10^{-10}. We therefore use $K = 8$ as a high-accuracy numerical piecewise-path reference rather than as an exact closed-form geodesic distance.

Principal-logarithm domain. The definition of $D_{\mathbb{SL}}$ requires the relative matrices to lie in the principal-logarithm domain. We therefore monitor the spectrum of relative matrices $R = A^{-1}B$ throughout training and evaluation. For an eigenvalue $\lambda = |\lambda|e^{i\theta}$, we define its angular margin to the negative-real branch cut by

$$m(R) = \min_{\lambda \in \sigma(R)} \big(\pi - |\arg \lambda|\big). \tag{25}$$

A positive margin ensures the spectrum avoids the principal-log branch cut.

Table 10: Principal-logarithm domain statistics on KEGG

Statistic	Value
Relative matrices evaluated	15,488
Principal-logarithm domain rate	100%
Branch margin $< 10^{-2}$	0%
Branch margin $< 10^{-3}$	0%
Minimum branch margin	2.3295
NaN / Inf rate	0%

All $15{,}488$ relative matrices encountered during training and evaluation remain inside the principal-logarithm domain. Moreover, the minimum observed branch margin is 2.3295 radians, and no sample approaches either the 10^{-2} or 10^{-3} branch-margin thresholds. We observe no NaN or Inf values. Thus, in this controlled training regime, the logarithmic objective operates well inside its regular principal domain rather than merely avoiding the branch cut by a small numerical margin.

Taken together, these experiments clarify the relation between the practical Schatten semidistance and the intrinsic path geometry. First, $D_{\mathbb{SL}}$ is exactly the F_p-length of the canonical exponential path and locally upper-bounds the intrinsic distance with second-order error. Second, allowing additional path segments substantially reduces the discrepancy to the refined path reference, yet changes end-to-end reconstruction performance by only 0.46%, while increasing training time by approximately $5.48\times$. Third, all relative matrices observed in this experiment remain well inside the principal-logarithm domain, and the symmetric and one-sided formulations coincide up to machine precision.

These results show that the practical performance is robust to substantial refinement of the underlying path objective, while the closed-form $D_{\mathbb{SL}}$ retains a clear computational advantage. We therefore use $D_{\mathbb{SL}}$ in the large-scale experiments and regard the K-segment construction as a controlled numerical approximation for examining its relation to the intrinsic $\mathbb{SL}$ path geometry.

B.2 Optimization Parameterization and Riemannian Control

The representation geometry and the optimization geometry need not coincide. In our main experiments, representations remain in $\mathbb{SL}(n)$ and the Schatten-p geometry enters through the pairwise objective, while a common Euclidean parameterization is optimized for every p. This isolates changes in the Schatten order from changes in the optimizer. We additionally use a left trivialized Riemannian AdamW method as a manifold aware control.

Exponential parameterized AdamW. We introduce unconstrained $X \in \mathbb{R}^{n\times n}$ and represent $A(X) = \exp(\Pi_{\mathfrak{sl}}(X)) \in \mathbb{SL}(n)$, where $\Pi_{\mathfrak{sl}}(X) = X - \operatorname{tr}(X)I/n \in \mathfrak{sl}(n)$. Then $\det A(X) = 1$. AdamW is applied to X, with gradients propagated through the projection and matrix exponential. We call this *Exponential Parameterized AdamW* (Exp-AdamW).

Algorithm 1 Exponential-Parameterized AdamW on $\mathbb{SL}(n)$

Require: Raw matrix $X_0 \in \mathbb{R}^{n\times n}$, learning rates $\{\eta_t\}_{t=1}^T$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon > 0$, and weight decay $\lambda \geq 0$

$M_0 \leftarrow 0$, $V_0 \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  $\Theta_{t-1} \leftarrow \Pi_{\mathfrak{sl}}(X_{t-1})$
  $A_{t-1} \leftarrow \exp(\Theta_{t-1})$
  $G_t \leftarrow \nabla_X f(A(X))|_{X=X_{t-1}}$
  $M_t \leftarrow \beta_1 M_{t-1} + (1-\beta_1)G_t$
  $V_t \leftarrow \beta_2 V_{t-1} + (1-\beta_2)(G_t \odot G_t)$
  $\widehat{M}_t \leftarrow M_t/(1-\beta_1^t)$, $\widehat{V}_t \leftarrow V_t/(1-\beta_2^t)$
  $X_t \leftarrow (1-\eta_t\lambda)X_{t-1} - \eta_t \widehat{M}_t/(\sqrt{\widehat{V}_t} + \epsilon)$
end for
return $A_T = \exp(\Pi_{\mathfrak{sl}}(X_T))$

Here, $\odot$, the square root, and division are applied elementwise.

Left trivialized Riemannian AdamW control. As a manifold aware control, we optimize A directly on $\mathbb{SL}(n)$ with the left invariant Frobenius metric. Let $G_t = \nabla_A^E f(A_t)$. For a tangent direction A_tZ, $df_{A_t}(A_tZ) = \langle A_t^\top G_t, Z\rangle_F$, so the left trivialized Riemannian gradient is $\Xi_t = \Pi_{\mathfrak{sl}}(A_t^\top G_t)$ and the corresponding tangent vector is $A_t\Xi_t$. Left invariance identifies tangent spaces with $\mathfrak{sl}(n)$, allowing the first moment to remain in the Lie algebra and the squared Frobenius norm to serve as a scalar second moment without explicit vector transport. For a smooth regularizer r, we use $\Omega_t = \Pi_{\mathfrak{sl}}(A_t^\top \nabla_A^E r(A_t))$ and apply the decay direction outside the adaptive moments, following decoupled weight decay.

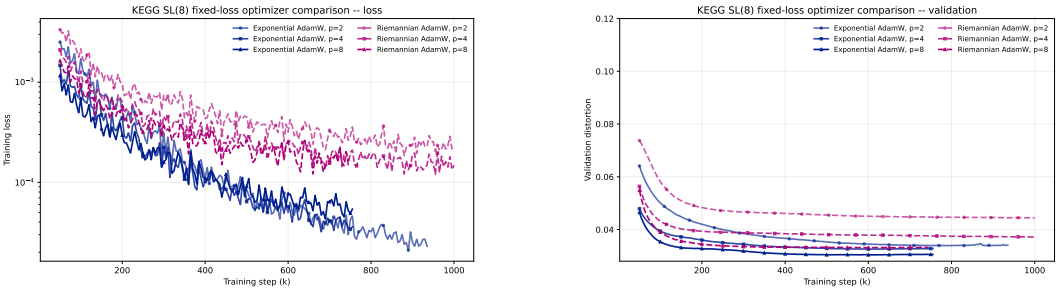
Algorithm 2 Left-Trivialized Riemannian AdamW on $\mathbb{SL}(n)$

Require: $A_0 \in \mathbb{SL}(n)$, learning rates $\{\eta_t\}_{t=1}^T$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon > 0$, weight decay $\lambda \geq 0$, and regularizer r

$M_0 \leftarrow 0 \in \mathfrak{sl}(n)$, $v_0 \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  $G_t \leftarrow \nabla_A^E f(A_{t-1})$
  $\Xi_t \leftarrow \Pi_{\mathfrak{sl}}(A_{t-1}^\top G_t)$
  $M_t \leftarrow \beta_1 M_{t-1} + (1-\beta_1)\Xi_t$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\|\Xi_t\|_F^2$
  $\widehat{M}_t \leftarrow M_t/(1-\beta_1^t)$, $\widehat{v}_t \leftarrow v_t/(1-\beta_2^t)$
  $D_t \leftarrow \Pi_{\mathfrak{sl}}\Big(\widehat{M}_t/(\sqrt{\widehat{v}_t} + \epsilon)\Big)$
  $\Omega_t \leftarrow \Pi_{\mathfrak{sl}}\big(A_{t-1}^\top \nabla_A^E r(A_{t-1})\big)$
  $A_t \leftarrow A_{t-1}\exp[-\eta_t(D_t + \lambda\Omega_t)]$
end for
return A_T

Every update direction in Algorithm 2 is trace free. Consequently, $\det(\exp[-\eta_t(D_t + \lambda\Omega_t)]) = 1$, so the iterates remain in $\mathrm{SL}(n)$ up to numerical precision. We use $r(A) = \frac{1}{2}\|A\|_F^2$, giving $\Omega_t = \Pi_{\mathfrak{sl}}(A_t^\top A_t)$. Unlike direct Euclidean shrinkage of A, this decay preserves the determinant constraint. When $\lambda = 0$, the method reduces to a left trivialized Riemannian Adam optimizer.

For $p = 2$, Riem-AdamW follows the same left invariant Frobenius geometry as the representation objective. For $p \neq 2$, it serves as a manifold aware control rather than an intrinsic Schatten-p Finsler optimizer. Exp-AdamW, in contrast, optimizes an unconstrained parameterization and is not an intrinsic Finsler gradient method. In both cases, the Schatten-p geometry enters through the same representation objective. This separation allows us to test whether the effect of changing p persists independently of the optimization geometry.



(a) Training loss curves for the optimizer comparison on $\mathbb{SL}_p(8)$.

(b) Validation distortion curves for the optimizer comparison on $\mathbb{SL}_p(8)$.

Figure 7: KEGG $\mathbb{SL}(8)$ optimizer comparison between EXP-AdamW and Riemannian AdamW under matched Schatten-p objectives.

Table 11: KEGG $\mathbb{SL}(8)$ optimizer comparison between EXP-AdamW and Riemannian AdamW under matched Schatten-p objectives.

p	Optimizer	**Distortion**$_{\text{avg.}}$ ↓	q50 ↓	q90 ↓	q95 ↓	High-related ↓	Low–Low ↓	Worst-group ↓	Best Val. ↓
2	Riem-AdamW	0.04537	0.01967	0.10111	0.17079	0.05182	0.04446	0.10547	0.04442
2	Exp-AdamW	0.03658	**0.00766**	0.06910	0.14796	0.04301	0.03629	0.08452	0.03391
4	Riem-AdamW	0.03786	0.01562	0.08882	0.14575	0.04422	0.03547	0.08828	0.03715
4	Exp-AdamW	0.03490	0.00928	0.07009	0.13330	0.03930	0.03458	0.08470	0.03255
8	Riem-AdamW	0.03463	0.01488	0.07717	0.12715	0.04123	**0.02994**	0.08971	0.03313
8	Exp-AdamW	**0.03338**	0.01006	**0.06727**	**0.12430**	**0.03869**	0.03180	**0.08231**	**0.03041**

Geometry versus optimization. Figure 7 and Table 11 show that Exp-AdamW achieves lower average distortion than Riem-AdamW for $p = 2, 4, 8$, with relative reductions of 19.4%, 7.8%, and 3.6%, respectively. All runs use seed 0 and batch size 512.

More importantly, increasing p from 2 to 8 improves average distortion under both optimizers, from 0.04537 to 0.03463 under Riem-AdamW and from 0.03658 to 0.03338 under Exp-AdamW. Since the optimizer is fixed within each comparison, this common trend supports an effect of the Schatten order on the learned geometric bias rather than an optimizer artifact. At the same time, the strong performance of Exp-AdamW shows that these gains do not require an intrinsic manifold optimizer. We therefore use Exp-AdamW in the main experiments for its simplicity and stronger empirical performance.

C Experimental Details

C.1 Common Experimental Protocol

We evaluate the representation spaces on KEGG, HumanCyc, OGBL-PPA, and Flickr30k-Order, covering metric reconstruction, large-scale link prediction, and multimodal order modeling. Unless stated otherwise, final results are computed over three independent model initialization seeds. Hyperparameters and checkpoints are selected using validation data only, and the test split is accessed only after model selection. For $\mathbb{SL}(n)$, the Schatten order p is selected using validation data: we use $p = 8$ for $\mathbb{SL}(4)$ and $p = 24$ for $\mathbb{SL}(8)$ on KEGG, $p = 4$ for $\mathbb{SL}(12)$ on HumanCyc, $p = 8$ for $\mathbb{SL}(8)$ on OGBL-PPA, and $p = 2$ for $\mathbb{SL}(4)$ on Flickr30k-Order.

Experiments are run across NVIDIA A100-SXM4-40GB, NVIDIA A100-SXM4-80GB, and NVIDIA H100 PCIe GPUs. Multiple independent runs may share one GPU, while maintaining independent model parameters, optimizer states, random-number states, and checkpoints. Hardware allocation affects wall-clock time only and does not change the data split, training budget, or model-selection protocol.

Within each task, all methods share the same data, supervision, training objective, and evaluation implementation. Only the latent representation, geometry-specific parameterization, and pairwise dissimilarity or score function are changed. Method-specific learning rates and batch sizes are predeclared to accommodate differences in numerical scale, memory footprint, and computational cost. Training budgets are therefore specified primarily in optimizer updates or epochs rather than wall-clock time.

C.2 Datasets and Task Protocols

KEGG. We use the largest connected component of KEGG pathway 24, treated as an undirected graph with self-loops removed. The resulting graph contains 377 nodes and 1,545 edges. Target distances are unweighted shortest-path distances. All $\binom{377}{2} = 70{,}876$ unordered node pairs are deterministically partitioned into $70\%/10\%/20\%$ training, validation, and test sets. Training pairs are sampled uniformly with replacement.

HumanCyc. We use the largest connected component of the HumanCyc-0 pathway graph after removing self-loops. The resulting graph contains 2,682 nodes and 28,177 edges. All $\binom{2682}{2} = 3{,}595{,}221$ unordered node pairs are partitioned using the same deterministic train/validation/test protocol as KEGG. For both reconstruction datasets, each node is represented directly by a trainable point $z_i \in \mathcal{M}$. No node features, graph encoder, or message-passing network is used, allowing the experiments to isolate the representation capacity of the latent geometry.

OGBL-PPA. We use the official `ogbl-ppa` split without modification. The graph contains 576,289 nodes and 21,231,931 training positive edges. The validation split contains 6,062,562 positive and 3,000,000 global negative edges, while the test split contains 3,031,780 positive and 3,000,000 global negative edges. We do not use the provided node features or message passing.

Training positives are sampled uniformly with replacement from the official training edges. For every positive edge (u, v), one negative destination v^- is sampled uniformly from all nodes while retaining the source u. Self-loops and edges present in the official training graph are rejected. Validation and test positives are not consulted by the training sampler.

Flickr30k-Order. We use OpenCLIP ViT-B/32 pretrained with the `openai` weights and keep the backbone frozen. Training uses the Flickr30k Karpathy training split with 29,000 images, five captions per image, and 145,000 training cases. For every training caption, four self-swap negatives are generated by randomly exchanging two word positions. Validation and test use the official ARO Flickr30k-Order splits. The validation set contains 1,014 images, 5,070 caption cases, and 19,160 valid positive-negative comparisons. The test set contains 1,000 images, 4,995 retained caption cases, and 18,859 positive-negative comparisons. Corruptions that become identical to the positive caption after preprocessing are removed.

C.3 Representation and Baseline Configurations

For metric reconstruction, all methods share the same node-level learning interface and reconstruction objective and differ only in the underlying space and pairwise dissimilarity.

Table 12: Geometric baselines used for metric reconstruction.

Space	Intrinsic dimension	Definition	Pairwise dissimilarity
$\mathbb{S}^d$	d	$\{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\}$	$d_{\mathbb{S}}(x,y) = \arccos(\langle x, y\rangle)$
$\mathbb{H}^d$	d	$\{x \in \mathbb{R}^d : \|x\|_2 < 1\}$	$d_{\mathbb{H}}(x,y) = \operatorname{arcosh}\left(1 + \frac{2\|x-y\|_2^2}{(1-\|x\|_2^2)(1-\|y\|_2^2)}\right)$
$\mathbb{E}^d$	d	$\mathbb{R}^d$	$d_{\mathbb{E}}(x,y) = \|x-y\|_2$
$\mathbb{S}^d \times \mathbb{H}^d$	$2d$	$\mathbb{S}^d \times \mathbb{H}^d$	$d = \left(d_{\mathbb{S}}^2 + d_{\mathbb{H}}^2\right)^{1/2}$
$\mathbb{S}^d \times \mathbb{E}^d$	$2d$	$\mathbb{S}^d \times \mathbb{E}^d$	$d = \left(d_{\mathbb{S}}^2 + d_{\mathbb{E}}^2\right)^{1/2}$
$\mathbb{H}^d \times \mathbb{E}^d$	$2d$	$\mathbb{H}^d \times \mathbb{E}^d$	$d = \left(d_{\mathbb{H}}^2 + d_{\mathbb{E}}^2\right)^{1/2}$
$\mathbb{S}^d \times \mathbb{H}^d \times \mathbb{E}^d$	$3d$	$\mathbb{S}^d \times \mathbb{H}^d \times \mathbb{E}^d$	$d = \left(d_{\mathbb{S}}^2 + d_{\mathbb{H}}^2 + d_{\mathbb{E}}^2\right)^{1/2}$
$\mathbb{S}^d_{\kappa_1} \times \mathbb{H}^d_{\kappa_2} \times \mathbb{E}^d$	$3d$	$\substack{c_{\mathbb{S}}, c_{\mathbb{H}} > 0,\\ w_{\mathbb{S}}, w_{\mathbb{H}}, w_{\mathbb{E}} \geq 0}$	$d = \left(w_{\mathbb{S}} d^2_{\mathbb{S}_{c_{\mathbb{S}}}} + w_{\mathbb{H}} d^2_{\mathbb{H}_{-c_{\mathbb{H}}}} + w_{\mathbb{E}} d^2_{\mathbb{E}}\right)^{1/2}$
Heisenberg-H^{2d+1}	$2d+1$	$(x,y,t) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$	$d_H(P,Q) = \inf_{\substack{\gamma(0)=P\\ \gamma(1)=Q}} \int_0^1 \sqrt{\|\dot{x}\|_2^2 + \|\dot{y}\|_2^2 + \left(\dot{t} + \frac{1}{2}(y^\top \dot{x} - x^\top \dot{y})\right)^2}\, ds$
$\mathbf{Grassmann}(k,n)$	$k(n-k)$	$\{\mathrm{span}(Q) : Q^\top Q = I_k\}$	$d_{\mathbf{Grassmann}}(Q_i, Q_j) = \left\|\arccos \sigma(Q_i^\top Q_j)\right\|_2$
$\mathbf{SPD}(n)$	$\frac{n(n+1)}{2}$	$\{P = P^\top \succ 0\}$	$d_{\mathbf{SPD}}(P_i, P_j) = \left\|\log\left(P_i^{-1/2} P_j P_i^{-1/2}\right)\right\|_F$
$\mathbf{Siegel}(n)$	$n(n+1)$	$\{X + \sqrt{-1}Y : X = X^\top,\ Y \succ 0\}$	$d_{\mathbf{Siegel}}(i,j) = \sum_{\ell=1}^{n} a_\ell \log \frac{1+\sigma_\ell(W_{ij})}{1-\sigma_\ell(W_{ij})},\quad a_\ell \geq 0,\ \sum_\ell a_\ell = n$
$\mathbb{SL}(n)$	$n^2 - 1$	$\{A \in \mathbb{R}^{n\times n} : \det(A) = 1\}$	$D_{\mathbb{SL}}(A,B) = \frac{1}{2}\left(\left\|\log(A^{-1}B)\right\|_{S_p} + \left\|\log(B^{-1}A)\right\|_{S_p}\right)$

For the learnable-curvature product model, the spherical and hyperbolic curvature magnitudes and the nonnegative factor weights are learned jointly. The factor weights are normalized to have mean one. For the Siegel baseline, we use $Z_{ij} = Y_i^{-1/2}(Z_j - X_i)Y_i^{-1/2}$ and $W_{ij} = (Z_{ij} - \sqrt{-1}I)(Z_{ij} + \sqrt{-1}I)^{-1}$, where $\sigma_\ell(W_{ij})$ denotes the corresponding Takagi singular values.

The Heisenberg group used in our experiments admits the matrix realization

$$H(x,y,t) = \begin{pmatrix} 1 & x^\top & t + \frac{1}{2}x^\top y \\ 0 & I_d & y \\ 0 & 0 & 1 \end{pmatrix}, \qquad x, y \in \mathbb{R}^d, \quad t \in \mathbb{R}. \tag{26}$$

The unitriangular group is $\mathrm{UT}(m) = \{U \in \mathbb{R}^{m\times m} : U_{ii} = 1,\ U_{ij} = 0 \text{ for } i > j\}$, with the schematic form

$$U = \begin{pmatrix} 1 & * & * & \cdots & * \\ 0 & 1 & * & \cdots & * \\ 0 & 0 & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{27}$$

For non-matrix representation spaces, we match intrinsic dimension to that of the corresponding $\mathbb{SL}(n)$ model as closely as possible. For matrix manifolds, we use comparable matrix-scale configurations whenever a natural matrix correspondence is available. Intrinsic dimensions and trainable parameter counts are reported explicitly for all methods.

OGBL-PPA representation. For OGBL-PPA, we use $\mathbb{SL}_8(8)$. Each node stores 63 sparse Lie-algebra coordinates x_i, which are mapped to the group as

$$A_i = \exp(0.05\, \Pi_{\mathfrak{sl}}(x_i))\,. \tag{28}$$

The link score is

$$s(u,v) = b - \exp\big(\mathrm{clip}(\rho, -5, 5)\big)\, D_{\mathcal{G}}(z_u, z_v), \tag{29}$$

where b and ρ are learned scalar parameters.

Flickr30k-Order composition models. Frozen CLIP token and image features are passed through separate two-layer projection heads with hidden width 512. The output coordinates are mapped to the corresponding Lie algebra and then to the group using

$$g_t = \exp_{\mathcal{G}}(\alpha X_t), \qquad g_I = \exp_{\mathcal{G}}(\alpha_I X_I), \qquad \alpha = \alpha_I = 0.1. \tag{30}$$

Caption tokens are composed in their original left-to-right order by

$$G_t = g_t G_{t-1}, \qquad G_{1:T} = g_T g_{T-1} \cdots g_1. \tag{31}$$

We compare Heisenberg-H^7, $\mathrm{UT}(4)$, $\mathrm{UT}(6)$, $\mathrm{UT}(15)$, full $\mathrm{SL}(4)$, and a commutative control. The commutative control retains the same $\mathfrak{sl}(4)$ token parameterization but replaces ordered multiplication by $G_{1:T}^{\mathrm{comm}} = \exp\left(\alpha \sum_{t=1}^{T} X_t\right)$. The image-caption score is

$$s(I, C) = -\beta D_{\mathcal{G}}(g_I, G_{1:T}), \qquad \beta = \exp(\tau), \tag{32}$$

where β is learned, initialized to 10, and upper-bounded by 100.

C.4 Training and Optimization

Metric reconstruction. KEGG and HumanCyc share the reconstruction objective

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} \left[\log(1 + sD_{\mathcal{M}}(z_i, z_j)) - \log(1 + d_G(i, j))\right]^2. \tag{33}$$

The logarithmic transformation prevents distant graph pairs from dominating the objective.

For KEGG, batch sizes are 8192 for product models, 2048 for Grassmann, 128 for Siegel, and 512 otherwise. Learning rates are 2×10^{-3} for products, 2×10^{-4} for Siegel, and 5×10^{-4} otherwise. We use weight decay 10^{-6} throughout and gradient clipping at 5 for Heisenberg and 10 otherwise. All KEGG high-capacity methods use at most 800,000 optimizer updates. Validation is performed every 5,000 updates, with early stopping after 150,000 updates without improvement.

HumanCyc uses a common maximum budget of 2,000,000 updates, batch size 64, evaluation batch size 128, learning rate 2×10^{-4}, weight decay 10^{-6}, and gradient-norm clipping at 10. Validation is performed every 10,000 updates, with early stopping after 50 consecutive non-improving validation events. Neither reconstruction experiment uses learning-rate warmup or a learning-rate scheduler.

OGBL-PPA. Training minimizes the pairwise BPR objective

$$\mathcal{L}_{\mathrm{BPR}} = -\frac{1}{|\mathcal{B}|} \sum_{(u,v)\in\mathcal{B}} \log \sigma\left(s(u, v) - s(u, v^-)\right). \tag{34}$$

The effective positive batch size is 4,194,304, evaluated through 32 gradient-accumulation micro-batches of 131,072 edges each.

Sparse node-coordinate rows are optimized using *exponential-parameterized SparseAdam* with learning rate 3×10^{-3}. The decoder scale and bias are optimized separately using Adam with learning rate 6×10^{-3}. No weight decay, learning-rate warmup, or learning-rate scheduler is used. Coordinates are clipped to $[-0.75, 0.75]$, sparse gradients are clipped elementwise to $[-0.05, 0.05]$, and dense gradients are clipped to global norm 1.0.

Training uses at most 17,000 optimizer updates. Validation is performed every 100 updates, with early stopping after 20 consecutive non-improving validation evaluations, corresponding to 2,000 optimizer updates.

Flickr30k-Order. Training minimizes the pairwise margin-ranking objective

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{\sum_i K_i} \sum_i \sum_{j=1}^{K_i} \left[0.1 - s(I_i, C_i^+) + s(I_i, C_{ij}^-)\right]_+. \tag{35}$$

All composition models use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 192, evaluation batch size 192, and global gradient-norm clipping at 1.0. Training lasts 12 epochs and uses neither warmup nor a learning-rate scheduler. Validation is performed once per epoch, and no early stopping is used.

C.5 Evaluation Metrics

For metric reconstruction, define the relative distortion of a held-out pair as

$$\delta_{ij} = \frac{|sD_{\mathcal{M}}(z_i, z_j) - d_G(i,j)|}{\max\{d_G(i,j), 1\}}. \tag{36}$$

The reported metrics across the three experimental settings are summarized below.

Table 13: Evaluation metrics used across the experiments.

Task	Metric	Definition / interpretation
Metric reconstruction	$\text{Distortion}_{\text{avg.}}$	Mean relative distortion δ_{ij} over all test pairs
Metric reconstruction	q50 / q90 / q95	50th, 90th, and 95th percentiles of δ_{ij}
Metric reconstruction	MAE	Mean absolute error between scaled latent and graph distances
Metric reconstruction	High-related	Mean distortion for pairs with at least one high-mixture endpoint
Metric reconstruction	Low–Low	Mean distortion for pairs whose endpoints are both in the low-mixture group
Metric reconstruction	Worst-group	Largest mean distortion among endpoint-group combinations
OGBL-PPA	Hits@20 / Hits@50 / Hits@100	Positive-edge ranking against the common global negative pool
OGBL-PPA	MRR	Mean reciprocal rank of positive edges
OGBL-PPA	Rank50 / Rank90 / Rank95	Corresponding quantiles of the positive-edge rank distribution
OGBL-PPA	AUC / AP	Classification metrics over the complete positive and negative score sets
Flickr30k-Order	OrderAcc	Fraction of positive captions scoring above individual order corruptions
Flickr30k-Order	OrderMargin	Mean positive-minus-negative score margin
Flickr30k-Order	HardAcc	Fraction of cases where the positive caption exceeds every corruption
Flickr30k-Order	MRR	Mean reciprocal rank of the positive caption within each candidate set

For reconstruction, nodes are grouped by local curvature-mixture entropy into low, medium, and high groups using thresholds 0.3 and 0.7. Group-based metrics are diagnostic only and are not used for checkpoint selection.

For OGBL-PPA, Hits@100 is computed using the official `Evaluator(name="ogbl-ppa")`. Hits@20, Hits@50, MRR, and rank quantiles use the same global negative pool. For a positive edge e^+,

$$\text{rank}(e^+) = 1 + \sum_{e^- \in \mathcal{E}^-} \mathbf{1}\big[s(e^-) \geq s(e^+)\big]. \tag{37}$$

AUC and AP are computed from the complete positive and negative score sets.

C.6 Fairness and Model Selection

Within each task, all baselines use the same data split, supervision, objective, validation protocol, and test evaluation. Test data are never used for choosing geometry parameters, learning rates, stopping points, or checkpoints. KEGG and HumanCyc select the checkpoint with the lowest validation mean distortion. All remaining reconstruction metrics are computed from this same checkpoint.

For OGBL-PPA, model selection uses one fixed validation subset containing $50{,}000$ official validation positives and $200{,}000$ official validation negatives. The subset is sampled once using a fixed random seed and shared across all methods and model seeds. Checkpoints are selected by Hits@100 on this subset. After model selection, final metrics are recomputed on the complete official validation and test splits.

Flickr30k-Order evaluates the full validation set after each epoch and selects the checkpoint with the highest OrderAcc, using OrderMargin only to break ties.

Final results use three independent model initialization seeds and report the mean and standard deviation when applicable. Method-specific optimization hyperparameters are fixed before final test evaluation.

D PROOFS OF THE THEORETICAL RESULTS

D.1 PRELIMINARIES

We provide the Finsler geometric background needed for the curvature analysis below. The main text introduced the fundamental tensor and flag curvature compactly. Here we develop these objects from the underlying Finsler norm, explain their geometric meaning, and describe how the Chern curvature and Jacobi operators used in our proofs are computed.

Finsler metric and directional geometry. Let M be a smooth manifold and TM its tangent bundle. A Riemannian metric assigns an inner product g_x to each tangent space T_xM, so the length of a tangent vector $y \in T_xM$ is determined by $\sqrt{g_x(y,y)}$. In particular, the local quadratic geometry at x does not depend on the direction along which it is examined.

A Finsler metric generalizes this construction by directly assigning a norm

$$F : TM \to [0, \infty)$$

to tangent vectors (Bao et al., 2000). For every $x \in M$, the restriction $F(x, \cdot)$ is positively homogeneous, $F(x, \lambda y) = \lambda F(x, y)$ for $\lambda > 0$, and its squared norm is strongly convex in the tangent direction. The essential distinction from Riemannian geometry is that $F^2(x, y)$ need not be quadratic in y. Consequently, the local geometry obtained by differentiating F may depend on the reference direction y itself.

Classically, the differential Finsler structure is considered on the slit tangent bundle $TM \setminus \{0\}$, since a positively homogeneous norm need not be differentiable at the zero vector. In our Schatten setting there is an additional regularity distinction. The norm $F_p(A, V) = \|A^{-1}V\|_{S_p}$ is defined for every tangent vector and hence defines the global length structure used throughout the paper. For $p \neq 2$, however, the smooth differential quantities involved in curvature are considered on full rank tangent directions. We call such directions *regular*. Thus the regularity restriction concerns differential curvature analysis, not the definition of the tangent norm or the induced path length.

Fundamental tensor. The first local geometric object derived from a Finsler metric is its *fundamental tensor*. For a regular nonzero reference direction $y \in T_xM$, define

$$g_y(u, v) := \frac{1}{2} \frac{\partial^2}{\partial s \, \partial t} F^2(x, y + su + tv)\Big|_{s=t=0}. \tag{38}$$

Equivalently, if the energy function $\mathcal{E}(x, y) = \frac{1}{2}F^2(x, y)$, then $g_y = D_y^2 \mathcal{E}(x, y)$.

The fundamental tensor can be understood as the local quadratic approximation of the squared Finsler norm around the direction y. Hence, although F itself may be nonquadratic, g_y provides an inner product with which infinitesimal lengths, angles, orthogonality, and curvature can be measured around that particular direction. For a Riemannian metric, $F^2(x, y) = g_x(y, y)$, so differentiating twice simply recovers g_x and the result is independent of y. Finsler geometry therefore contains Riemannian geometry as the special case in which this directional dependence disappears.

By the two homogeneity of F^2, the fundamental tensor satisfies $g_y(y, y) = F^2(x, y)$ and $g_y(y, u) = \frac{1}{2}D_y F^2(x, y)[u]$. These identities will be repeatedly used below.

Cartan tensor and departure from Riemannian geometry. The variation of the fundamental tensor with respect to the reference direction is measured by the *Cartan tensor*,

$$C_y(u, v, w) := \frac{1}{2} \frac{d}{dt} g_{y+tw}(u, v)\Big|_{t=0} = \frac{1}{4} D_y^3 F^2(x, y)[u, v, w]. \tag{39}$$

Intuitively, g_y describes the local quadratic geometry seen from y, whereas C_y measures how this quadratic geometry changes when the viewing direction changes. For a Riemannian metric, g_y is independent of y and hence $C_y = 0$. The nonzero Cartan tensor is therefore one of the fundamental sources of genuinely Finsler behavior. Although we do not explicitly manipulate C_y in the main curvature proof, its effect is implicitly contained in the Chern connection introduced next.

Length, geodesics, and the geodesic spray. For a piecewise smooth curve $\gamma(t)$, its Finsler length is $L_F(\gamma) = \int F(\gamma(t), \dot{\gamma}(t))\, dt$. A geodesic is locally a critical curve of the corresponding energy functional. As in Riemannian geometry, geodesics describe locally straight motion, but the direction dependence of F changes their equations.

To make this explicit, take local coordinates $(x^1, \dots, x^d)$ on M and write a tangent vector as $y = y^i \partial_{x^i}$ (with ∂_{x^i} playing the role of the Euclidean coordinate basis vector e_i). Let $g_{ij}(x, y)$ denote the matrix of the fundamental tensor and $g^{ij}(x, y)$ its inverse. The geodesic spray coefficients are

$$G^i(x, y) = \frac{1}{4} g^{i\ell}(x, y) \left(\frac{\partial^2 F^2}{\partial x^k \partial y^\ell} y^k - \frac{\partial F^2}{\partial x^\ell} \right). \tag{40}$$

A constant speed Finsler geodesic satisfies $\ddot{x}^i + 2G^i(x, \dot{x}) = 0$. Thus the spray G plays the role of the Christoffel symbols contracted with velocity in Riemannian geometry. Starting from F, one can therefore obtain the geodesic dynamics entirely by differentiation. The derivatives

$$N^i{}_j(x, y) := \partial G^i(x, y)/\partial y^j \tag{41}$$

define the associated nonlinear connection. They separate changes in the base point from changes in the tangent direction and introduce the horizontal derivatives $\delta_j = \partial_{x^j} - N^m{}_j \partial_{y^m}$. These quantities provide the intermediate step from the Finsler norm to its canonical connection and curvature.

Chern connection. Curvature requires comparing tangent vectors at nearby points, which in turn requires a connection. In Riemannian geometry this role is played by the Levi–Civita connection. Because the Finsler inner product g_y also depends on the direction y, there is in general no ordinary Levi–Civita connection depending only on the base point. The standard replacement is the *Chern connection* (Bao et al., 2000). The Chern connection is the canonical torsion free connection that is compatible with the direction dependent fundamental tensor in the Finsler sense. In local coordinates its coefficients can be obtained from g_y and the horizontal derivatives as

$$\Gamma^i{}_{jk}(x, y) = \frac{1}{2} g^{i\ell} \left(\delta_j g_{\ell k} + \delta_k g_{\ell j} - \delta_\ell g_{jk} \right). \tag{42}$$

All quantities in this expression depend on both the point x and the reference direction y. This is the main difference from the Riemannian Christoffel symbols, which depend only on x. Equations (38), (40), and (42) give a direct computational chain

$$F \longrightarrow g_y \longrightarrow G \longrightarrow \Gamma(x, y).$$

Chern curvature. Curvature measures the failure of parallel transport defined by the connection to commute around infinitesimal loops. For the Chern connection, we denote the corresponding curvature operator at reference direction y by $\boldsymbol{\mathcal{R}}^y(u, v)w$. Under the curvature convention used in this paper, it is the Finsler counterpart of $\nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w$ from Riemannian geometry.

In local coordinates, the horizontal Chern curvature coefficients are obtained by differentiating the connection coefficients,

$$R^i{}_{jkl} = \delta_k \Gamma^i{}_{jl} - \delta_l \Gamma^i{}_{jk} + \Gamma^m{}_{jl} \Gamma^i{}_{mk} - \Gamma^m{}_{jk} \Gamma^i{}_{ml}. \tag{43}$$

Consequently, $\boldsymbol{\mathcal{R}}^y(u, v)w$ is obtained by contracting these coefficients with u, v, and w. Together, the computational dependence is

$$F \longrightarrow g_y \longrightarrow G \longrightarrow \Gamma \longrightarrow \boldsymbol{\mathcal{R}}^y.$$

Because the Chern connection itself depends on the reference direction, varying that direction may introduce additional anisotropic terms compared with the curvature of an ordinary affine connection. These terms can be described through the vertical variation of the Chern connection (Javaloyes, 2014; 2019). They vanish in several important situations used below, allowing the Finsler curvature to be computed through an associated affine connection. We make this specialization explicit when it is used.

Jacobi operator. The full Chern curvature $\boldsymbol{\mathcal{R}}^y(u,v)w$ depends on three tangent directions. For sectional or flag curvature, two of these directions are fixed by the reference direction y. This motivates the *Jacobi operator*,

$$R_y u := \boldsymbol{\mathcal{R}}^y(u,y)y. \tag{44}$$

Thus R_y is a linear operator on the tangent space obtained by inserting the flagpole y twice into the full curvature tensor. Geometrically, $R_y u$ describes how a nearby geodesic initially separated from the reference geodesic in direction u accelerates relative to it. Positive and negative values of its quadratic form therefore correspond to qualitatively different local bending of nearby geodesics.

The Jacobi operator satisfies $R_y y = 0$ and is self adjoint with respect to the fundamental tensor,

$$g_y(R_y u, v) = g_y(u, R_y v). \tag{45}$$

Hence its eigenvalues are real. On the g_y orthogonal complement of y, we denote its positive and negative eigenspaces by $E_+(y)$ and $E_-(y)$, respectively. These eigenspaces provide the positive and negative curvature modes used in the curvature coupling definition of the main text.

Flag curvature. In Riemannian geometry, sectional curvature depends only on a two dimensional plane. In Finsler geometry the local metric depends additionally on the direction used to inspect that plane. A *flag* is therefore a pair (y,Π) consisting of a nonzero reference direction y, called the *flagpole*, and a two dimensional tangent plane $\Pi = \mathrm{span}\{y,u\}$ containing it.

The corresponding *flag curvature* is

$$K_F(y,\Pi) = \frac{g_y(R_y u, u)}{g_y(y,y)g_y(u,u) - g_y(y,u)^2}. \tag{46}$$

The denominator is the Gram determinant of y and u under g_y and is positive whenever they are linearly independent. As shown in Fig. 8, flag curvature measures the signed curvature of Π as viewed from the distinguished direction y. For visual clarity, the figure uses Y and U for y and u in the text.

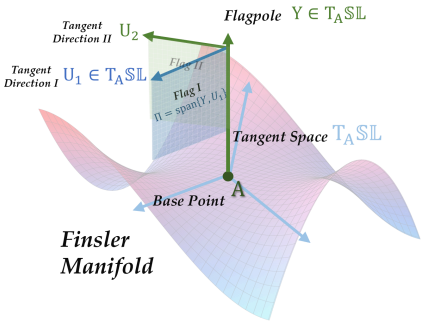


Figure 8: A Finsler flag consists of a tangent plane Π together with a distinguished tangent direction y.

As only the component of u transverse to y determines the plane, we may choose u such that $g_y(y,u) = 0$. Then

$$K_F(y, \mathrm{span}\{y,u\}) = \frac{g_y(R_y u, u)}{g_y(y,y)g_y(u,u)}. \tag{47}$$

The denominator is positive, so $\operatorname{sign} K_F = \operatorname{sign} g_y(R_y u, u)$. This is why the curvature analysis below focuses on the quadratic form $g_y(R_y u, u)$ rather than repeatedly evaluating the complete fraction in Eq. (46).

Specialization to the Schatten p geometry. We now specialize these general constructions to $\mathbb{SL}_p(n)$. Left invariance identifies the geometry at every point with the geometry on the Lie algebra. For a regular full rank $X \in \mathfrak{sl}(n)$, define the energy function

$$\mathcal{E}_p(X) = \frac{1}{2}\|X\|_{S_p}^2 = \frac{1}{2}\left[\mathrm{tr}((X^\top X)^{p/2})\right]^{2/p}. \tag{48}$$

The fundamental tensor at the identity is simply

$$g_X(U,V) = D^2\mathcal{E}_p(X)[U,V]. \tag{49}$$

At an arbitrary $A \in \mathrm{SL}(n)$, setting $X = A^{-1}Y$, $\widetilde{U} = A^{-1}U$, and $\widetilde{V} = A^{-1}V$ gives $g_Y(U,V) = g_X(\widetilde{U},\widetilde{V})$. Thus it is sufficient to compute the differential geometry at the identity and transport the result by left multiplication. This reduction is especially useful for curvature. A reference direction $X \in \mathfrak{sl}(n)$ is called a *geodesic vector* when the trajectory generated by X is a geodesic. For a left invariant Finsler metric this is equivalent to

$$g_X(X,[X,Z]) = 0 \qquad \text{for all } Z \in \mathfrak{sl}(n). \tag{50}$$

At such a direction, the Chern connection can be represented algebraically by the connection operator N_X (Xu & Deng, 2018).

$$2g_X(N_X(V), W) = g_X([W, V], X) + g_X([W, X], V) + g_X([V, X], W). \tag{51}$$

This formula is the homogeneous counterpart of the coordinate Chern connection in Eq. (42). Once the fundamental tensor is known, the right hand side contains only inner products and matrix commutators, so $N_X(V)$ can be solved directly. At a geodesic reference direction, the Jacobi operator then takes the form

$$R_X(V) = -N_X(N_X(V)) + N_X([X, V]) - [X, N_X(V)]. \tag{52}$$

Hence the curvature calculation used in our proofs follows the concrete sequence

$$\mathcal{E}_p \longrightarrow g_X \longrightarrow N_X \longrightarrow R_X \longrightarrow K_F. \tag{53}$$

This avoids solving the geodesic equations or evaluating the full coordinate Chern curvature tensor directly.

Pairwise matrix directions and spectral derivatives. The remaining calculations exploit the spectral structure of the Schatten norm. We choose a diagonal regular reference direction $X = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\sum_i \lambda_i = 0$. Let $E_{ij} \in \mathbb{R}^{n\times n}$ denote the standard matrix unit with a single 1 in the (i, j) entry and zeros elsewhere. For every pair $i < j$, we define the symmetric and skew-symmetric directions $S_{ij} := (E_{ij} + E_{ji})/\sqrt{2}$ and $A_{ij} := (E_{ij} - E_{ji})/\sqrt{2}$, respectively. Thus, S_{ij} and A_{ij} are the symmetric and skew-symmetric combinations of the same pair of matrix units E_{ij} and E_{ji}, which are

$$S_{ij} = \frac{1}{\sqrt{2}} \left[\begin{array}{c|ccccc} & \cdots & i & \cdots & j & \cdots \\ \hline \vdots & \ddots & \vdots & & \vdots & \\ i & \cdots & 0 & \cdots & 1 & \cdots \\ \vdots & & \vdots & \ddots & \vdots & \\ j & \cdots & 1 & \cdots & 0 & \cdots \\ \vdots & & \vdots & & \vdots & \ddots \end{array}\right], \quad A_{ij} = \frac{1}{\sqrt{2}} \left[\begin{array}{c|ccccc} & \cdots & i & \cdots & j & \cdots \\ \hline \vdots & \ddots & \vdots & & \vdots & \\ i & \cdots & 0 & \cdots & 1 & \cdots \\ \vdots & & \vdots & \ddots & \vdots & \\ j & \cdots & -1 & \cdots & 0 & \cdots \\ \vdots & & \vdots & & \vdots & \ddots \end{array}\right]. \tag{54}$$

These directions perturb only the (i, j) coordinates of the matrix. They therefore reduce the high dimensional matrix calculation to simple two dimensional blocks. For diagonal X, their interaction with the reference direction is controlled by the spectral difference $d_{ij} := \lambda_i - \lambda_j$, with

$$[X, S_{ij}] = d_{ij}A_{ij}, \qquad [X, A_{ij}] = d_{ij}S_{ij}, \qquad [A_{ij}, S_{ij}] = E_{ii} - E_{jj}. \tag{55}$$

Thus the two directions remain inside the same small matrix block under the operations entering the connection and curvature formulas.

Differentiating a matrix spectral function such as $\mathrm{tr}((X^\top X)^{p/2})$ naturally produces *divided differences*. Define $\phi_p(t) = t|t|^{p-2}$. For scalars $a \neq b$, its divided difference is $(\phi_p(a) - \phi_p(b))/(a - b)$, with the derivative used as the continuous extension when $a = b$. Divided differences are the matrix analogue of ordinary derivatives when a perturbation mixes two spectral coordinates. Since ϕ_p is strictly increasing for $p > 1$, these divided differences are positive.

This observation explains the quantities s_{ij} and a_{ij} introduced below: they are precisely the fundamental tensor weights of the symmetric and skew symmetric directions, $s_{ij} = g_X(S_{ij}, S_{ij})$ and $a_{ij} = g_X(A_{ij}, A_{ij})$. Their ratio $m_{ij}(X) = a_{ij}/s_{ij}$ summarizes how the Schatten p geometry weights the two directions inside the same pairwise block. The next proposition computes these quantities and shows how they enter the Chern curvature.

Regularity and analytic dependence. On the full rank matrix locus, $X^\top X$ is positive definite and $X \mapsto [\mathrm{tr}((X^\top X)^{p/2})]^{2/p}$ is real analytic (Tsing et al., 1994). Consequently, the fundamental tensor and the resulting Chern curvature coefficients vary analytically with a regular reference direction wherever the fundamental tensor is nondegenerate. We later use the standard fact that a nonzero real analytic scalar function on a connected open set has an open dense nonzero locus. This allows finitely many nonvanishing curvature interactions to be realized simultaneously without requiring a specially tuned reference direction.

D.2 Rootwise Curvature and Proof of Mixed Flag Curvature

Proposition D.1 (Rootwise effective Cartan curvature)**.** *Let $X = \mathrm{diag}(\lambda_1, \dots, \lambda_n) \in \mathfrak{sl}(n)$ be full rank and let $S_{ij} = \frac{E_{ij}+E_{ji}}{\sqrt{2}}, A_{ij} = \frac{E_{ij}-E_{ji}}{\sqrt{2}}, i < j$. Define $\phi_p(t) = t|t|^{p-2}$ and*

$$\begin{aligned} s_{ij} &:= g_X(S_{ij}, S_{ij}) = \|X\|_{S_p}^{2-p} \frac{\phi_p(\lambda_i) - \phi_p(\lambda_j)}{\lambda_i - \lambda_j}, \\ a_{ij} &:= g_X(A_{ij}, A_{ij}) = \|X\|_{S_p}^{2-p} \frac{\phi_p(\lambda_i) + \phi_p(\lambda_j)}{\lambda_i + \lambda_j}, \end{aligned} \tag{56}$$

where the divided differences are understood by continuous extension when a denominator vanishes. Then $s_{ij} > 0$ and $a_{ij} > 0$. Define the effective Cartan parameter $m_{ij}(X) := \frac{a_{ij}}{s_{ij}} > 0$. Whenever $\lambda_i \neq \lambda_j$, the two root directions are eigenvectors of the Jacobi operator and satisfy

$$R_X(A_{ij}) = \frac{(\lambda_i - \lambda_j)^2}{4} m_{ij}(X) A_{ij}, \quad R_X(S_{ij}) = -\frac{(\lambda_i - \lambda_j)^2}{4}\big(4 + 3m_{ij}(X)\big) S_{ij}. \tag{57}$$

Consequently, $g_X(R_X A_{ij}, A_{ij}) > 0$ and $g_X(R_X S_{ij}, S_{ij}) < 0$.

Proof. We first compute the fundamental tensor on the two-dimensional (i, j) root block. Since $\mathcal{E}_p(Z) = \frac{1}{2}(\mathrm{tr}(Z^\top Z)^{p/2})^{2/p}$, standard second-order spectral calculus at a diagonal full-rank matrix gives

$$\begin{aligned} D^2 \,\mathrm{tr}((X^\top X)^{p/2})[S_{ij}, S_{ij}] &= p\, \frac{\phi_p(\lambda_i) - \phi_p(\lambda_j)}{\lambda_i - \lambda_j}, \\ D^2 \,\mathrm{tr}((X^\top X)^{p/2})[A_{ij}, A_{ij}] &= p\, \frac{\phi_p(\lambda_i) + \phi_p(\lambda_j)}{\lambda_i + \lambda_j}. \end{aligned} \tag{58}$$

Every off-diagonal perturbation has zero first variation at diagonal X. Applying the scalar chain rule to $\mathcal{E}_p$ therefore multiplies both expressions in equation 58 by the common positive factor $\|X\|_{S_p}^{2-p}/p$, which gives equation 56.

The function $\phi_p(t) = t|t|^{p-2}$ is strictly increasing for $p > 1$. Hence the first quotient in equation 56 is a positive divided difference of ϕ_p. For the second quotient, use the oddness of ϕ_p to write $\phi_p(\lambda_i) + \phi_p(\lambda_j) = \phi_p(\lambda_i) - \phi_p(-\lambda_j)$ and $\lambda_i + \lambda_j = \lambda_i - (-\lambda_j)$. It is therefore again a positive divided difference of the same strictly increasing function. At a vanishing denominator, the continuous extension is $(p-1)|\lambda_i|^{p-2} > 0$ because X is full rank. Thus $s_{ij} > 0$, $a_{ij} > 0$, and consequently $m_{ij}(X) > 0$.

Transpose invariance of $\mathcal{E}_p$ gives $g_X(S_{ij}, A_{ij}) = 0$, while diagonal sign conjugations imply orthogonality between distinct root blocks. Since X is diagonal, both S_{ij} and A_{ij} are also g_X-orthogonal to X. We next verify that the diagonal reference direction X is geodesic. By two-homogeneity, $g_X(X, V) = D\mathcal{E}_p(X)[V]$. Hence, for every $Z \in \mathfrak{sl}(n)$,

$$\begin{aligned} g_X(X, [X, Z]) &= \|X\|_{S_p}^{2-p} \,\mathrm{tr}\left(X|X|^{p-2}[X, Z]\right) \\ &= \|X\|_{S_p}^{2-p} \,\mathrm{tr}\left([X|X|^{p-2}, X]Z\right) = 0, \end{aligned} \tag{59}$$

because both X and $X|X|^{p-2}$ are diagonal. Thus X is a geodesic vector.

Let N_X denote the homogeneous Chern connection operator at this geodesic reference direction. It satisfies

$$2g_X(N_X(V), W) = g_X([W, V], X) + g_X([W, X], V) + g_X([V, X], W). \tag{60}$$

The only Lie brackets needed on the (i, j) root block are

$$[X, S_{ij}] = (\lambda_i - \lambda_j) A_{ij}, \qquad [X, A_{ij}] = (\lambda_i - \lambda_j) S_{ij}, \qquad [A_{ij}, S_{ij}] = E_{ii} - E_{jj}. \tag{61}$$

Moreover, two-homogeneity and equation 56 give

$$g_X(E_{ii} - E_{jj}, X) = (\lambda_i - \lambda_j) s_{ij}. \tag{62}$$

We now compute the connection on this root block. Testing equation 60 against S_{ij}, against diagonal directions, and against every distinct root block shows that $N_X(S_{ij})$ has only an A_{ij} component. Pairing with A_{ij} and using equation 61–equation 62 gives

$$\begin{aligned} 2g_X(N_X(S_{ij}), A_{ij}) &= g_X([A_{ij}, S_{ij}], X) + g_X([A_{ij}, X], S_{ij}) + g_X([S_{ij}, X], A_{ij}) \\ &= (\lambda_i - \lambda_j)s_{ij} - (\lambda_i - \lambda_j)s_{ij} - (\lambda_i - \lambda_j)a_{ij} \\ &= -(\lambda_i - \lambda_j)a_{ij}. \end{aligned} \tag{63}$$

Since $g_X(A_{ij}, A_{ij}) = a_{ij}$, this yields $N_X(S_{ij}) = -\frac{\lambda_i - \lambda_j}{2} A_{ij}$.

Similarly, testing equation 60 shows that $N_X(A_{ij})$ has only an S_{ij} component. Pairing with S_{ij} gives

$$\begin{aligned} 2g_X(N_X(A_{ij}), S_{ij}) &= g_X([S_{ij}, A_{ij}], X) + g_X([S_{ij}, X], A_{ij}) + g_X([A_{ij}, X], S_{ij}) \\ &= -(\lambda_i - \lambda_j)s_{ij} - (\lambda_i - \lambda_j)a_{ij} - (\lambda_i - \lambda_j)s_{ij} \\ &= -(\lambda_i - \lambda_j)(2s_{ij} + a_{ij}). \end{aligned} \tag{64}$$

Using $g_X(S_{ij}, S_{ij}) = s_{ij}$ and $m_{ij}(X) = a_{ij}/s_{ij}$, we therefore obtain the compact rootwise connection formulas

$$N_X(S_{ij}) = -\frac{\lambda_i - \lambda_j}{2} A_{ij}, \qquad N_X(A_{ij}) = -\frac{\lambda_i - \lambda_j}{2}\big(2 + m_{ij}(X)\big) S_{ij}. \tag{65}$$

The significance of equation 65 is that the entire dependence of the Schatten-p fundamental tensor on this root block is compressed into the single positive scalar $m_{ij}(X)$. The connection has the similar rootwise algebraic form as the classical Cartan metric $\nu_m = -mB|_{\mathfrak{k}} + B|_{\mathfrak{p}}$, with m replaced by the effective parameter $m_{ij}(X)$.

For a left-invariant Finsler metric at a geodesic reference direction, the Jacobi operator satisfies

$$R_X(V) = -N_X(N_X(V)) + N_X([X, V]) - [X, N_X(V)]. \tag{66}$$

We first apply equation 66 to A_{ij}. Using equation 61 and equation 65, its three terms are respectively

$$\begin{aligned} -N_X(N_X(A_{ij})) &= -\frac{(\lambda_i - \lambda_j)^2}{4}\big(2 + m_{ij}(X)\big) A_{ij}, \\ N_X([X, A_{ij}]) &= -\frac{(\lambda_i - \lambda_j)^2}{2} A_{ij}, \\ -[X, N_X(A_{ij})] &= \frac{(\lambda_i - \lambda_j)^2}{2}\big(2 + m_{ij}(X)\big) A_{ij}. \end{aligned} \tag{67}$$

Adding the three coefficients leaves only $m_{ij}(X)/4$, and therefore

$$R_X(A_{ij}) = \frac{(\lambda_i - \lambda_j)^2}{4} m_{ij}(X) A_{ij}. \tag{68}$$

For S_{ij}, the three terms are

$$\begin{aligned} -N_X(N_X(S_{ij})) &= -\frac{(\lambda_i - \lambda_j)^2}{4}\big(2 + m_{ij}(X)\big) S_{ij}, \\ N_X([X, S_{ij}]) &= -\frac{(\lambda_i - \lambda_j)^2}{2}\big(2 + m_{ij}(X)\big) S_{ij}, \\ -[X, N_X(S_{ij})] &= \frac{(\lambda_i - \lambda_j)^2}{2} S_{ij}. \end{aligned} \tag{69}$$

Their sum is

$$R_X(S_{ij}) = -\frac{(\lambda_i - \lambda_j)^2}{4}\big(4 + 3m_{ij}(X)\big) S_{ij}. \tag{70}$$

Equations equation 68 and equation 70 prove equation 57.

Finally, $a_{ij} > 0$, $s_{ij} > 0$, and $m_{ij}(X) > 0$. Hence, whenever $\lambda_i \neq \lambda_j$,

$$g_X(R_X A_{ij}, A_{ij}) = \frac{(\lambda_i - \lambda_j)^2}{4} m_{ij}(X) a_{ij} > 0, \tag{71}$$

$$g_X(R_X S_{ij}, S_{ij}) = -\frac{(\lambda_i - \lambda_j)^2}{4}\big(4 + 3m_{ij}(X)\big) s_{ij} < 0. \tag{72}$$

This proves the proposition. □

Proof of Theorem 2.1. We first work at the identity. Choose

$$X = \operatorname{diag}\left(1, 2, \ldots, n-1, -\frac{n(n-1)}{2}\right). \tag{73}$$

This matrix is trace free, full rank, and has pairwise distinct diagonal entries. Proposition D.1 applied to the $(1,2)$ root block therefore gives $g_X(R_X A_{12}, A_{12}) > 0$ and $g_X(R_X S_{12}, S_{12}) < 0$.

To obtain zero curvature around the same flagpole, define $U(t) = \cos t\, A_{12} + \sin t\, S_{12}$ for $t \in [0, \pi/2]$. Both root directions are g_X-orthogonal to X, so every $U(t)$ is transverse to X. Moreover, A_{12} and S_{12} are mutually g_X-orthogonal and are eigenvectors of R_X. Hence

$$\begin{aligned} g_X(R_X U(t), U(t)) = {} & \frac{(\lambda_1 - \lambda_2)^2}{4} m_{12}(X) a_{12} \cos^2 t \\ & - \frac{(\lambda_1 - \lambda_2)^2}{4} \big(4 + 3m_{12}(X)\big) s_{12} \sin^2 t. \end{aligned} \tag{74}$$

The expression is positive at $t = 0$ and negative at $t = \pi/2$. By continuity, there exists $t_0 \in (0, \pi/2)$ for which $g_X(R_X U(t_0), U(t_0)) = 0$.

Thus the three transverse directions A_{12}, $U(t_0)$, and S_{12} have respectively positive, zero, and negative flag curvature around the same regular full-rank flagpole X by equation 47.

Finally, let $P \in \mathbb{SL}_p(n)$ be arbitrary. Left translation by P is an isometry because, for every $B \in \mathbb{SL}_p(n)$ and tangent vector V at B,

$$F_p(PB, PV) = \|(PB)^{-1} PV\|_{S_p} = \|B^{-1} V\|_{S_p} = F_p(B, V). \tag{75}$$

Therefore flag curvature is preserved by left translation. The flagpole $PX \in T_P \mathbb{SL}_p(n)$ remains regular and full rank, while the translated directions PA_{12}, $PU(t_0)$, and PS_{12} have respectively positive, zero, and negative flag curvature. Since P was arbitrary, the three curvature regimes coexist at every point around a common regular flagpole. □

D.3 Proof of Asymptotically Maximal Mixed Curvature and Coupling

The preceding Jacobi curvature calculation identifies positive and negative curvature eigenspaces and hence their mixed-curvature capacity. To establish intrinsic coupling according to Definition 2.2, we additionally need a nonzero interaction under the full Chern curvature operator $\boldsymbol{\mathcal{R}}^X$.

Lemma D.2 (Nonvanishing Chern interaction on a root block)**.** *Let $1 < p < \infty$ and let $X = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \mathfrak{sl}(n)$ be full rank. Fix $i < j$ and suppose $\lambda_i = \lambda_j = c \neq 0$. Then*

$$\boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij})S_{ij} = \frac{a_{ij}}{2s_{ij}} A_{ij}. \tag{76}$$

By Proposition D.1, At this equal-eigenvalue point, When $\lambda_i = \lambda_j \neq 0$, the continuous extensions give $a_{ij} = \|X\|_{S_p}^{2-p} |\lambda_i|^{p-2} > 0,\ s_{ij} = \|X\|_{S_p}^{2-p} (p-1)|\lambda_i|^{p-2} > 0$. and therefore

$$g_X\left(\boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij})S_{ij}, A_{ij}\right) = \frac{a_{ij}^2}{2s_{ij}} > 0. \tag{77}$$

Proof. Since $\lambda_i = \lambda_j$, we have $[X, A_{ij}] = [X, S_{ij}] = 0$. Proposition D.1 therefore gives $N_X(A_{ij}) = N_X(S_{ij}) = 0$. For $H_{ij} = E_{ii} - E_{jj}$, the same connection identity gives $N_X(H_{ij}) = 0$.

Let ∇^X denote the affine Chern connection associated with the left invariant reference field determined by X. The geodesic specialization of the homogeneous connection identity agrees with the Chern Koszul formula (Javaloyes, 2014, Proposition 2.3), so $(\nabla^X_U X)_I = N_X(U)$. Consequently, $\nabla^X_{A_{ij}} X = \nabla^X_{S_{ij}} X = \nabla^X_{H_{ij}} X = 0$. The anisotropic curvature formula of (Javaloyes, 2019, Lemma 2.11) then implies that, for the directions used below, the anisotropic correction terms vanish. Hence

$$\boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij})S_{ij} = R^{\nabla^X}(A_{ij}, S_{ij})S_{ij}. \tag{78}$$

At the equal eigenvalue point, $g_X(H_{ij}, H_{ij}) = 2s_{ij}$ and A_{ij}, S_{ij}, H_{ij} are mutually g_X orthogonal. For an arbitrary diagonal $D = \operatorname{diag}(d_1, \dots, d_n)$, we also have $g_X(H_{ij}, D) = s_{ij}(d_i - d_j)$. The relevant brackets are $[A_{ij}, S_{ij}] = H_{ij}$, $[H_{ij}, S_{ij}] = 2A_{ij}$, and $[H_{ij}, A_{ij}] = 2S_{ij}$.

Conjugation by diagonal orthogonal sign matrices fixes X and shows that $\nabla^X_{A_{ij}} S_{ij}$ and $\nabla^X_{S_{ij}} S_{ij}$ are diagonal. The Chern Koszul formula therefore gives, for every diagonal D,

$$\begin{aligned} 2g_X(\nabla^X_{A_{ij}} S_{ij}, D) &= g_X([A_{ij}, S_{ij}], D) - g_X([S_{ij}, D], A_{ij}) + g_X([D, A_{ij}], S_{ij}) \\ &= (2s_{ij} + a_{ij})(d_i - d_j). \end{aligned} \tag{79}$$

On the other hand,

$$2g_X\left(\left(1 + \frac{a_{ij}}{2s_{ij}}\right) H_{ij}, D\right) = (2s_{ij} + a_{ij})(d_i - d_j). \tag{80}$$

Both vectors are diagonal, and g_X is nondegenerate there. Therefore $\nabla^X_{A_{ij}} S_{ij} = (1 + a_{ij}/(2s_{ij}))H_{ij}$. Applying the same Koszul identity to $\nabla^X_{S_{ij}} S_{ij}$ and $\nabla^X_{H_{ij}} S_{ij}$ gives $\nabla^X_{S_{ij}} S_{ij} = 0$ and $\nabla^X_{H_{ij}} S_{ij} = A_{ij}$. Since the Chern connection is torsion free,

$$\nabla^X_{S_{ij}} H_{ij} = \nabla^X_{H_{ij}} S_{ij} - [H_{ij}, S_{ij}] = -A_{ij}. \tag{81}$$

Using $R^{\nabla^X}(U, V)W = \nabla^X_U \nabla^X_V W - \nabla^X_V \nabla^X_U W - \nabla^X_{[U,V]} W$, we obtain

$$\begin{aligned} \boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij})S_{ij} &= R^{\nabla^X}(A_{ij}, S_{ij})S_{ij} \\ &= -\nabla^X_{S_{ij}}\left(\left(1 + \frac{a_{ij}}{2s_{ij}}\right) H_{ij}\right) - \nabla^X_{H_{ij}} S_{ij} \\ &= \left(1 + \frac{a_{ij}}{2s_{ij}}\right) A_{ij} - A_{ij} \\ &= \frac{a_{ij}}{2s_{ij}} A_{ij}. \end{aligned} \tag{82}$$

Taking the g_X inner product with A_{ij} proves the final claim. □

Proof of Corollary 2.3. For $n = 2$, the lower bound $\binom{n-1}{2} = 0$ is immediate, so we assume $n \geq 3$.

For fixed $p \in (1, \infty)$, $\mathcal{E}_p(X) = \frac{1}{2}[\mathrm{tr}((X^\top X)^{p/2})]^{2/p}$ is real analytic on the full rank matrix locus. Indeed, $X^\top X$ is positive definite there, and standard analytic matrix function calculus (Tsing et al., 1994) applies. Hence the fundamental tensor, the Chern connection, and the Chern curvature operator are real analytic in the reference direction wherever the fundamental tensor is nondegenerate.

First, we consider the connected diagonal chamber

$$\Omega = \left\{ \mathrm{diag}(\lambda_1, \ldots, \lambda_n) : \lambda_1, \ldots, \lambda_{n-1} > 0,\ \lambda_n = -\sum_{r=1}^{n-1} \lambda_r \right\}. \tag{83}$$

For each $1 \leq i < j \leq n-1$, let $\kappa_{ij}(X) = g_X(\boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij})S_{ij}, A_{ij})$. This is a real analytic scalar function on Ω. For every fixed pair (i, j), Ω contains a point with $\lambda_i = \lambda_j > 0$, and Lemma D.2 gives $\kappa_{ij}(X) > 0$ there. Thus κ_{ij} is not identically zero. Its nonzero set is therefore open and dense in Ω.

There are only finitely many pairs $1 \leq i < j \leq n-1$. Intersecting all of these open dense sets with the open dense subset on which the diagonal entries are pairwise distinct gives a regular full rank $X \in \Omega$ satisfying

$$\kappa_{ij}(X) \neq 0, \qquad 1 \leq i < j \leq n-1. \tag{84}$$

For any diagonal sign matrix $D = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_n)$ with $\varepsilon_r \in \{\pm 1\}$, conjugation $\Phi_D(B) = DBD^{-1}$ fixes every diagonal X and is an isometry because

$$F_p(\Phi_D(B), d\Phi_D(V)) = \|DB^{-1}VD^{-1}\|_{S_p} = \|B^{-1}V\|_{S_p}. \tag{85}$$

Moreover, $d\Phi_D(A_{ij}) = (\varepsilon_i\varepsilon_j)A_{ij}$ and $d\Phi_D(S_{ij}) = (\varepsilon_i\varepsilon_j)S_{ij}$.

Naturality of Chern curvature under isometries now gives the required block separation. For two unordered pairs (i, j) and (k, l), $\boldsymbol{\mathcal{R}}^X(A_{kl}, S_{ij})S_{ij}$ transforms under every Φ_D with the same sign $\varepsilon_k\varepsilon_l$ as the (k, l) root block, because the two copies of S_{ij} contribute the square $(\varepsilon_i\varepsilon_j)^2 = 1$. Distinct root blocks have distinct sign patterns under all such conjugations. Therefore

$$\boldsymbol{\mathcal{R}}^X(A_{kl}, S_{ij})S_{ij} \in \mathrm{span}\{A_{kl}, S_{kl}\}. \tag{86}$$

The same argument, now with A_{ij} and S_{ij} contributing a cancelling square, gives

$$\boldsymbol{\mathcal{R}}^X(A_{ij}, S_{kl})S_{ij} \in \mathrm{span}\{A_{kl}, S_{kl}\}. \tag{87}$$

For the generic X fixed above, consider

$$W_+ = \mathrm{span}\{A_{ij} : 1 \leq i < j \leq n-1\}, \qquad W_- = \mathrm{span}\{S_{ij} : 1 \leq i < j \leq n-1\}. \tag{88}$$

Proposition D.1 gives $W_+ \subseteq E_+(X)$ and $W_- \subseteq E_-(X)$, with

$$\dim W_+ = \dim W_- = \binom{n-1}{2}. \tag{89}$$

It follows immediately that

$$\mathcal{C}_{\mathrm{mix}}(X) = \min\{\dim E_+(X), \dim E_-(X)\} \geq \binom{n-1}{2}. \tag{90}$$

We next show that the same lower bound remains after excluding uncoupled modes. First, $\mathcal{N}_+(X) \cap W_+ = \{0\}$. Suppose otherwise and write a nonzero vector in the intersection as $U = \sum_{k<l} u_{kl}A_{kl}$. Choose (i, j) with $u_{ij} \neq 0$. Since $S_{ij} \in E_-(X)$ and $U \in \mathcal{N}_+(X)$, the definition of $\mathcal{N}_+(X)$ gives $\boldsymbol{\mathcal{R}}^X(U, S_{ij}) = 0$ as an endomorphism. Applying it to S_{ij} yields

$$0 = \sum_{k<l} u_{kl}\boldsymbol{\mathcal{R}}^X(A_{kl}, S_{ij})S_{ij}. \tag{91}$$

By the root block separation proved above, the term indexed by (k, l) lies in $\mathrm{span}\{A_{kl}, S_{kl}\}$. These root blocks are linearly independent, so the component in $\mathrm{span}\{A_{ij}, S_{ij}\}$ must vanish:

$$u_{ij}\boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij})S_{ij} = 0. \tag{92}$$

Taking the g_X inner product with A_{ij} gives $u_{ij}\kappa_{ij}(X) = 0$, contradicting both $u_{ij} \neq 0$ and $\kappa_{ij}(X) \neq 0$. Hence $\mathcal{N}_+(X) \cap W_+ = \{0\}$.

Both $\mathcal{N}_+(X)$ and W_+ are subspaces of $E_+(X)$. Their trivial intersection therefore implies

$$\dim E_+(X) - \dim \mathcal{N}_+(X) \geq \dim W_+ = \binom{n-1}{2}. \tag{93}$$

The negative side is analogous, but we spell it out. Suppose $0 \neq V = \sum_{k<l} v_{kl} S_{kl} \in \mathcal{N}_-(X) \cap W_-$ and choose (i, j) with $v_{ij} \neq 0$. Because $A_{ij} \in E_+(X)$ and $V \in \mathcal{N}_-(X)$, we have $\boldsymbol{\mathcal{R}}^X(A_{ij}, V) = 0$. Applying this endomorphism to S_{ij} gives

$$0 = \sum_{k<l} v_{kl} \boldsymbol{\mathcal{R}}^X(A_{ij}, S_{kl}) S_{ij}. \tag{94}$$

The second root block separation property places each summand in its (k, l) root block. Projecting onto $\operatorname{span}\{A_{ij}, S_{ij}\}$ gives

$$v_{ij} \boldsymbol{\mathcal{R}}^X(A_{ij}, S_{ij}) S_{ij} = 0, \tag{95}$$

which again contradicts $v_{ij}\kappa_{ij}(X) \neq 0$. Thus $\mathcal{N}_-(X) \cap W_- = \{0\}$, and the same dimension argument gives

$$\dim E_-(X) - \dim \mathcal{N}_-(X) \geq \binom{n-1}{2}. \tag{96}$$

Therefore

$$\mathcal{C}_{\mathrm{cpl}}(X) \geq \binom{n-1}{2}. \tag{97}$$

Let $A \in \mathbb{SL}_p(n)$ be arbitrary and consider the left translated flagpole $AX \in T_A\mathbb{SL}_p(n)$. Left translation by A is an isometry of the Schatten p structure. Hence its differential preserves the fundamental tensor and the Chern curvature:

$$\begin{aligned} g_{AX}(AU, AV) &= g_X(U, V), \\ R_{AX}(AU) &= A\, R_X U, \\ \boldsymbol{\mathcal{R}}^{AX}(AU, AV)(AW) &= A\, \boldsymbol{\mathcal{R}}^X(U, V) W. \end{aligned} \tag{98}$$

The second identity implies $E_\pm(AX) = AE_\pm(X)$, while the third gives $\mathcal{N}_\pm(AX) = A\mathcal{N}_\pm(X)$. Since left multiplication is invertible, all corresponding spaces have the same dimensions. Hence

$$\mathcal{C}_{\mathrm{mix}}(AX) = \mathcal{C}_{\mathrm{mix}}(X), \qquad \mathcal{C}_{\mathrm{cpl}}(AX) = \mathcal{C}_{\mathrm{cpl}}(X). \tag{99}$$

Since AX is a regular full rank flagpole at the arbitrary point A, the pointwise definitions yield

$$\mathcal{C}_{\mathrm{mix}}(A) \geq \binom{n-1}{2}, \qquad \mathcal{C}_{\mathrm{cpl}}(A) \geq \binom{n-1}{2}. \tag{100}$$

Taking the minimum over $A \in \mathrm{SL}(n)$ therefore gives

$$\mathcal{C}^{\mathrm{mix}}_{\mathbb{SL}}(n, p) \geq \binom{n-1}{2}, \qquad \mathcal{C}^{\mathrm{cpl}}_{\mathbb{SL}}(n, p) \geq \binom{n-1}{2}. \tag{101}$$

Then at any $A \in \mathbb{SL}_p(n)$ and any regular flagpole $Y \in T_A\mathbb{SL}_p(n)$, the tangent space has dimension $n^2 - 1$, so the g_Y transverse space has dimension $n^2 - 2$. Because R_Y is g_Y self adjoint, its positive, zero, and negative eigenspaces form a direct orthogonal decomposition of this transverse space. Therefore,

$$\begin{aligned} \mathcal{C}_{\mathrm{cpl}}(Y) \leq \mathcal{C}_{\mathrm{mix}}(Y) &= \min\{\dim E_+(Y), \dim E_-(Y)\} \\ &\leq \left\lfloor \frac{\dim E_+(Y) + \dim E_-(Y)}{2} \right\rfloor \leq \left\lfloor \frac{n^2 - 2}{2} \right\rfloor. \end{aligned} \tag{102}$$

Define the maximal balanced curvature capacity permitted by the transverse space as $\mathcal{C}_{\max}(n) := \lfloor (n^2 - 2)/2 \rfloor$. Since the bound holds for every regular flagpole at every point, taking the pointwise maximum and then the minimum over the manifold gives

$$\mathcal{C}^{\mathrm{cpl}}_{\mathbb{SL}}(n, p) \leq \mathcal{C}^{\mathrm{mix}}_{\mathbb{SL}}(n, p) \leq \mathcal{C}_{\max}(n). \tag{103}$$

Combining this with the lower bound $\mathcal{C}^{\mathrm{cpl}}_{\mathbb{SL}}(n,p) \geq \binom{n-1}{2}$ yields

$$\frac{\binom{n-1}{2}}{\mathcal{C}_{\max}(n)} \leq \frac{\mathcal{C}^{\mathrm{cpl}}_{\mathbb{SL}}(n,p)}{\mathcal{C}_{\max}(n)} \leq \frac{\mathcal{C}^{\mathrm{mix}}_{\mathbb{SL}}(n,p)}{\mathcal{C}_{\max}(n)} \leq 1. \tag{104}$$

The leftmost term converges to 1 as $n \to \infty$. Hence, by squeezing,

$$\frac{\mathcal{C}^{\mathrm{cpl}}_{\mathbb{SL}}(n,p)}{\mathcal{C}_{\max}(n)} \longrightarrow 1, \qquad \frac{\mathcal{C}^{\mathrm{mix}}_{\mathbb{SL}}(n,p)}{\mathcal{C}_{\max}(n)} \longrightarrow 1. \tag{105}$$

□

D.4 Proof of Infinite Order Depth

Proof of Lemma 2.4. For every $n \geq 2$, the Lie algebra $\mathfrak{sl}(n)$ contains the upper left 2×2 copy of $\mathfrak{sl}(2)$. Let $H = E_{11} - E_{22}$ and $E = E_{12}$. Then $[H, E] = 2E$, and consequently, for every integer $k \geq 1$,

$$\mathrm{ad}^k_H(E) = \underbrace{[H, [H, \ldots, [H}_{k \text{ times}}, E] \ldots]] = 2^k E \neq 0. \tag{106}$$

Choosing $X_0 = E$ and $X_1 = \cdots = X_k = H$ in the definition of order depth therefore gives a nonzero nested Lie bracket at every finite depth k. Since $T_I \mathbb{SL}_p(n) = \mathfrak{sl}(n)$ for every $p \in (1, \infty)$, this construction is independent of p. Hence

$$D_{\mathrm{ord}}(\mathbb{SL}_p(n)) = \infty. \tag{107}$$

□

D.5 Counting Product Geometry Candidates

We detail how the number 11,555,651,398 in the introduction is calculated. Consider

$$\mathcal{P} = \mathbb{E}^{d_0} \times \prod_{i=1}^{m_-} \mathbb{H}^{d_i} \times \prod_{j=1}^{m_+} \mathbb{S}^{d_j}, \tag{108}$$

with fixed total dimension $d_0 + \sum_i d_i + \sum_j d_j = 64$. We fix the curvature magnitudes and count only distinct choices of factor dimensions. Factors of the same type are treated as unordered, since permuting, for example, $\mathbb{H}^{d_1}$ and $\mathbb{H}^{d_2}$ does not produce a different product geometry. Let $p(n)$ denote the integer partition number. The partition numbers are computed recursively from

$$p(n) = p(n-1) + p(n-2) - p(n-5) - p(n-7) + p(n-12) + p(n-15) - \cdots, \tag{109}$$

where the offsets $1, 2, 5, 7, 12, 15, \ldots$ are the generalized pentagonal numbers, with $p(0) = 1$ and $p(n) = 0$ for $n < 0$. Hence, if the total dimension assigned to the hyperbolic factors is h, the possible decompositions $h = d_1 + \cdots + d_{m_-}$ are counted by $p(h)$. For example, for $h = 5$, the partitions 5, $4+1$, $3+2$, $3+1+1$, $2+2+1$, $2+1+1+1$, and $1+1+1+1+1$ correspond to seven distinct choices of hyperbolic factors. The spherical factors are counted independently in the same way. For a fixed Euclidean dimension d_0, let $h = \sum_i d_i$ be the total hyperbolic dimension. The remaining spherical dimension is then $s = 64 - d_0 - h$. There are therefore $p(h)$ possible hyperbolic decompositions and $p(s)$ possible spherical decompositions. Since the two choices are independent, a fixed pair (d_0, h) contributes $p(h)p(64 - d_0 - h)$ candidates. Summing over all admissible dimension allocations gives

$$N_{64} = \sum_{d_0=0}^{64} \sum_{h=0}^{64-d_0} p(h)p(64 - d_0 - h) = 11{,}555{,}651{,}398. \tag{110}$$