Covariant Osculating Jet Networks: Higher-Order Contact Fields with Soft-Routing Guarantees, Differential Diagnostics, and Matched Prior-Art Controls

Andrew Kiruluta
UC Berkeley CA, School of Information
kiruluta@berkeley.edu

August 2026

Abstract

Deep networks normally obtain nonlinear representation power by composing affine maps and pointwise activations. We study a complementary primitive in which a hidden transformation is represented locally by trainable differential contact: value, tangent map, signed curvature, and directional third-order variation. The resulting *Covariant Osculating Jet Geometric Layer* (COJGL) learns contact poles μ_c, positive local routing metrics $G_c = \beta_c I + M_c M_c^\top$, signed low-rank Hessians $H_{c,o} = U_{c,o}\,\mathrm{diag}(\lambda_{c,o})U_{c,o}^\top + \alpha_{c,o}I$, and symmetric CP cubic jets. A softmax partition of unity glues these contacts into a global map, while optional empirical-overlap penalties match local values, Jacobians, and Hessians ($C^0/C^1/C^2$) on hidden states where chart responsibilities actually overlap.

This revision addresses two weaknesses of earlier formulations. First, the theory no longer assumes compact support that softmax routing does not possess. We prove an explicit exponentially decaying routing-tail bound and derive a global approximation estimate equal to a local low-rank Taylor error plus a softmax tail term. Second, the experiments target the proposed inductive bias directly and include matched architectural controls. Five-seed smooth and piecewise-smooth regression experiments compare COJGL against parameter-matched MLP, global quadratic, Taylor-style order-3, and POU-style local polynomial models. COJGL improves over the matched MLP on both tasks but is slightly worse than the POU-style baseline, indicating that localization explains part of the gain. On a controlled 128-dimensional curved-latent classification task with approximately matched parameters, COJGN-3 reaches $92.35 \pm 1.64\%$ versus $92.15 \pm 1.27\%$ for the matched MLP. Conversely, on ordinary Digits the matched MLP remains much stronger ($95.82 \pm 0.48\%$ versus $88.71 \pm 1.58\%$), so no universal superiority is claimed.

Five-seed geometric diagnostics show that strong C^2 gluing reduces overlap-weighted value, Jacobian, and Hessian mismatch while leaving accuracy essentially unchanged, validating the coherence mechanism separately from prediction. Practical profiling also shows that at matched parameter count COJGN currently costs roughly $3.6\times$ more inference latency, $4.2\times$ more training-step time, and about $2\times$ the measured peak CPU RSS delta of an MLP. The contribution is therefore a falsifiable higher-order geometric neural primitive with a sharper theory and a clearer empirical regime of utility, not a state-of-the-art replacement for conventional MLPs.

1 Introduction

Geometric deep learning usually exploits geometry supplied by the problem domain: groups, graphs, manifolds, gauges, or topological complexes [2]. We ask a different question: can a neural network make the *local higher-order geometry of its own hidden transformation* an explicit computational state? A smooth map is locally characterized by a jet—its value and derivatives up to a chosen order [9]. Standard affine-plus-activation layers contain such differential information only implicitly. COJGL parameterizes it directly and then composes these local models in residual hidden-state blocks.

Several pieces of this idea have substantial prior art. Partition-of-unity (POU) networks learn localized polynomial approximants [7, 11]; probabilistic POU models extend the construction to high-dimensional regression and latent-dimensional structure [3]. TaylorNet uses low-rank tensor decompositions to make higher-order Taylor layers practical [12]; polynomial networks and factorization machines efficiently represent low-order interactions [1]; mixtures of local experts learn gates [6]; and recent atlas-learning work constructs explicit local quadratic charts and transition maps for Riemannian computation [8]. Softmax itself has also recently been analyzed as a globally supported partition of unity for local-to-global approximation [10]. Thus neither local polynomials, higher-order coefficients, learned partitions, nor atlas language is novel by itself.

Our narrower hypothesis is that an explicit field of trainable local jets can be useful when the target hidden map has *low effective differential rank*: high ambient dimension but locally low-rank Hessian and third derivative, possibly with anisotropic contact regions. Figures 1–4 visualize the resulting contact geometry, higher-order osculation, differential gluing, and residual computation. This suggests three regimes where COJGL should be more plausible than on arbitrary tabular classification: (i) smooth or piecewise-smooth regression, (ii) high-dimensional observations generated by a low-dimensional curved latent process, and (iii) scientific or dynamical maps whose local derivative structure is itself meaningful.

Contributions.

1. **Higher-order contact layer.** Each local expert contains a learned value, Jacobian, signed low-rank curvature, and low-rank symmetric cubic jet.
2. **Covariant metric routing.** $G_c = \beta_c I + M_c M_c^\top \succ 0$ yields anisotropic learned contact regions and exact covariance under orthogonal latent reparameterization.
3. **Empirical-overlap differential gluing.** Contacts can be matched through second derivative on actual overlap states, with analytic gradients and Hutchinson Hessian-vector probes.
4. **Rigorous globally supported routing analysis.** We replace an invalid compact-support assumption with an explicit softmax routing-tail term that decays exponentially with metric separation.
5. **Prior-art controls and geometric diagnostics.** We add POU-style, Taylor-style, quadratic, and matched-MLP controls, five-seed coherence diagnostics, and targeted high-dimensional experiments.
6. **Practical cost accounting.** We report wall-clock CPU inference/training latency and process-memory deltas rather than parameter counts alone.

2 Prior Work and Novelty Boundary

2.1 Partition-of-unity and local polynomial networks

POU-Nets deliberately combine a learned spatial partition with local polynomial approximation, and hierarchical variants exploit multilevel training and polynomial reproduction [7, 11]. Probabilistic POU-Nets further target high-dimensional regression through local polynomial experts on lower-dimensional structure [3]. These are the closest precedents to COJGL's localization. The present contribution does *not* claim that local polynomial gluing is new. It instead tests whether metric routing, signed differential factors, residual latent composition, and explicit cross-contact derivative consistency add a useful neural primitive beyond localization alone.

2.2 Taylor and polynomial architectures

TaylorNet explicitly builds neural architectures from Taylor expansions and uses low-rank tensor decompositions to control the curse of dimensionality [12]. This is a critical prior-art boundary for the third-order component. COJGL differs by centering jets at multiple learned hidden contacts, giving the quadratic factors a signed principal-curvature interpretation, using contact-specific routing geometry, and directly regularizing differential agreement between overlapping local expansions. Our experimental "Taylor-style" control isolates a global low-rank order-3 polynomial motif; it is a self-contained reimplementation rather than the official TaylorNet system.

2.3 Atlas learning and derivative-aware representations

Atlas-based Manifold Representations learns approximate quadratic coordinate charts and differentiable transition structure for Riemannian machine learning [8]. COJGL currently shares one latent coordinate system across contacts and therefore should be called a *contact cover*, not a classical differentiable atlas. Hermite-NGP stores function and derivative information for efficient PDE evaluation [4]; jet bundles themselves are standard higher-order differential geometry [9]. The contribution here is architectural: the derivatives parameterize a trainable hidden transformation and are repeatedly composed in residual computation.

3 Covariant Osculating Contact Geometry

Let $x \in \mathbb{R}^D$. Contact $c \in \{1, \ldots, K\}$ has center μ_c and local coordinate $\Delta_c = x - \mu_c$.

Table 1: Novelty boundary. "POU-style" and "Taylor-style" experiments in this paper are self-contained matched controls for the corresponding architectural motifs, not official upstream implementations.

Family	local regions	explicit higher-order coefficients	signed low-rank H	learned metric routing	derivative gluing
Mixture of experts [6]	✓	–	–	optional	–
POU/PPOU [3, 7]	✓	polynomial	optional	learned gate	–
TaylorNet [12]	–	✓	tensor factorization	–	–
Atlas learning [8]	✓	quadratic charts	geometric	chart domains	transition consistency
COJGL (this work)	✓	order 1–3	✓	✓	$C^0/C^1/C^2$

3.1 Positive metric routing

Each contact carries

$$G_c = \beta_c I + M_c M_c^\top, \qquad \beta_c = \mathrm{softplus}(b_c) + \varepsilon > 0. \tag{1}$$

Its squared routing distance and normalized responsibility are

$$d_c^2(x) = \Delta_c^\top G_c \Delta_c, \qquad \gamma_c(x) = \frac{e^{-d_c^2(x)}}{\sum_{j=1}^K e^{-d_j^2(x)}}. \tag{2}$$

Thus $G_c \succ 0$, $\gamma_c > 0$, and $\sum_c \gamma_c = 1$. Low-rank $M_c \in \mathbb{R}^{D \times R_g}$ evaluates d_c^2 in $O(DR_g)$.

3.2 Signed order-2 and order-3 contact

For output channel o, COJGL uses

$$q_{c,o}(x) = w_{0,c,o} + J_{c,o}\Delta_c + \frac{1}{2}\Delta_c^\top H_{c,o}\Delta_c + \frac{1}{6}T_{c,o}[\Delta_c, \Delta_c, \Delta_c], \tag{3}$$

$$H_{c,o} = U_{c,o}\,\mathrm{diag}(\lambda_{c,o})U_{c,o}^\top + \alpha_{c,o} I, \tag{4}$$

$$T_{c,o}[\Delta^3] = \sum_{s=1}^{R_3} a_{c,o,s}(t_{c,o,s}^\top \Delta)^3. \tag{5}$$

Columns of U and t are normalized during evaluation. Signed λ permits positive and negative off-axis curvature. The local map is blended as

$$F_o(x) = \sum_{c=1}^K \gamma_c(x) q_{c,o}(x). \tag{6}$$

3.3 Orthogonal covariance

Proposition 1 (Orthogonal covariance)**.** *Let $Q \in O(D)$ and transform latent coordinates by $x' = Qx$, anchors by $\mu_c' = Q\mu_c$, and low-rank direction factors by $M_c' = QM_c$, $U_{c,o}' = QU_{c,o}$, $t_{c,o}' = Qt_{c,o}$, with $J_{c,o}' = J_{c,o}Q^\top$. Then $d_c'^2(x') = d_c^2(x)$ and $q_{c,o}'(x') = q_{c,o}(x)$, hence $F'(x') = F(x)$.*

Proof. Because $Q^\top Q = I$, Euclidean norms and all projected coordinates are preserved: $(QM)^\top(Q\Delta) = M^\top\Delta$, $(QU)^\top(Q\Delta) = U^\top\Delta$, and likewise for t. The first-order term satisfies $JQ^\top Q\Delta = J\Delta$. Therefore all routing and jet terms are unchanged. □

This is a precise covariance statement under orthogonal reparameterizations; arbitrary nonlinear coordinate invariance is not claimed.

4 Differential Gluing on Empirical Overlaps

For contacts c, d, define the empirical overlap weight

$$\omega_{cd}(x) = \gamma_c(x)\gamma_d(x). \tag{7}$$

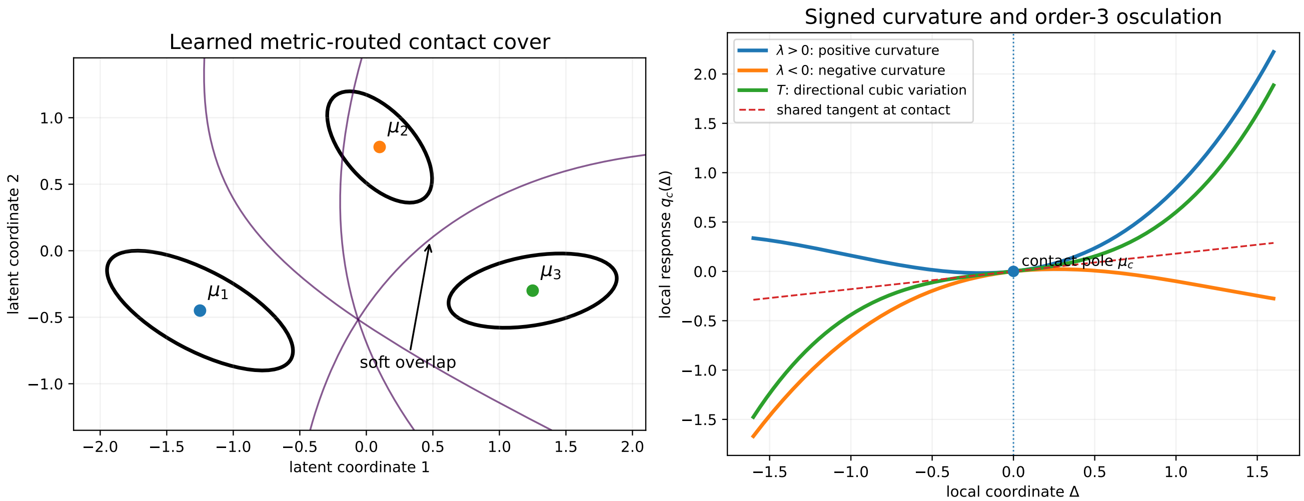

Figure 1: Geometric intuition for COJGL. **Left:** each learned pole μ_c carries its own positive metric G_c, producing anisotropic Mahalanobis contact regions and globally smooth overlap through γ_c. **Right:** around a contact pole, the Jacobian fixes the tangent, signed Hessian eigenvalues permit positive or negative curvature, and the cubic jet changes that curvature directionally away from the pole.

Rather than comparing charts only at the midpoint of their anchors, we compare them where the current hidden-state distribution activates both. The normalized pairwise overlap measure is

$$\rho_{cd}(x_i) = \frac{\omega_{cd}(x_i)}{\sum_j \omega_{cd}(x_j)}. \tag{8}$$

Then

$$\mathcal{L}_{C^0}^{cd} = \mathbb{E}_{\rho_{cd}} \|q_c - q_d\|^2 , \tag{9}$$
$$\mathcal{L}_{C^1}^{cd} = \mathcal{L}_{C^0}^{cd} + \lambda_J \mathbb{E}_{\rho_{cd}} \|Dq_c - Dq_d\|_F^2 , \tag{10}$$
$$\mathcal{L}_{C^2}^{cd} = \mathcal{L}_{C^1}^{cd} + \lambda_H \mathbb{E}_{\rho_{cd}} \left\|D^2 q_c - D^2 q_d\right\|_F^2 . \tag{11}$$

The implementation detaches ρ_{cd} during the gluing loss so the router cannot reduce the penalty merely by destroying overlap.

For C^2 matching, dense Hessians are unnecessary. If z is Rademacher with $\mathbb{E}[zz^\top] = I$, then

$$\mathbb{E}_z \|(H_c - H_d)z\|^2 = \|H_c - H_d\|_F^2 . \tag{12}$$

The analytic HVP for the factorized quadratic+cubic jet costs $O(D(R_2 + R_3))$ per contact/output/probe.

5 Theory without Compact-Support Assumptions

Earlier versions used an "effective support" assumption of the form $\|x - \mu_c\| \le h$ for active anchors. Softmax never has exact compact support for finite precision, so that assumption does not describe the implemented model. We replace it with a global tail bound.

Proposition 2 (Softmax routing-tail mass)**.** *Assume for all contacts*

$$0 < g_- I \preceq G_c \preceq g_+ I. \tag{13}$$

For a point x, suppose some anchor c_* satisfies $\|x - \mu_{c_*}\| \le r$. For $h > r$, define the outer set $\mathcal{O}_h(x) = \{c : \|x - \mu_c\| \ge h\}$. Then

$$\tau_h(x) := \sum_{c \in \mathcal{O}_h(x)} \gamma_c(x) \le \min\left\{1,\ K \exp\left(g_+ r^2 - g_- h^2\right)\right\}. \tag{14}$$

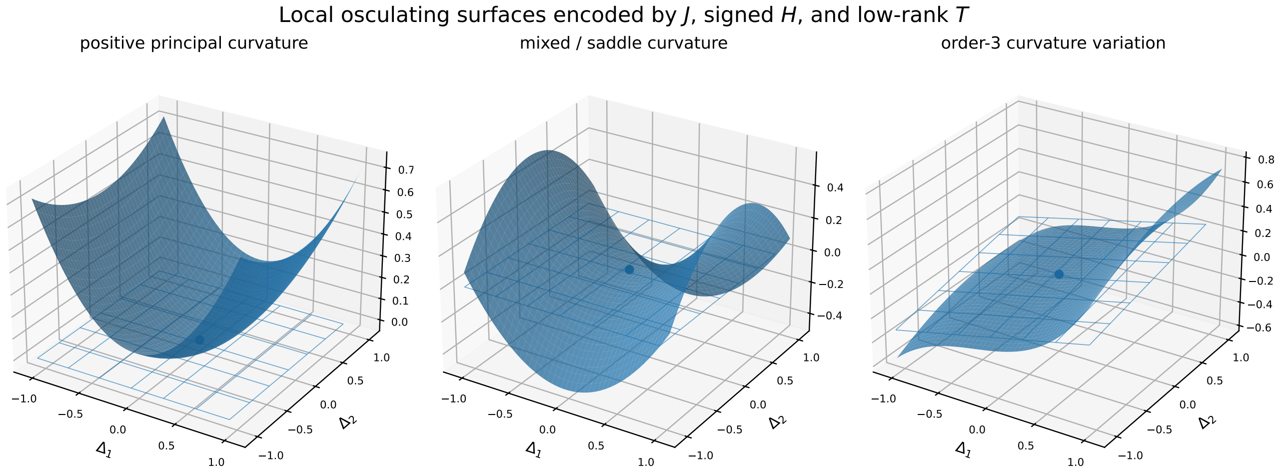

Figure 2: Three local osculating surface geometries. The tangent plane gives first-order contact; the signed Hessian produces elliptic or saddle second-order contact; and the low-rank cubic tensor bends the curvature itself across the neighborhood. These are the geometric objects parameterized explicitly by a COJGL contact.

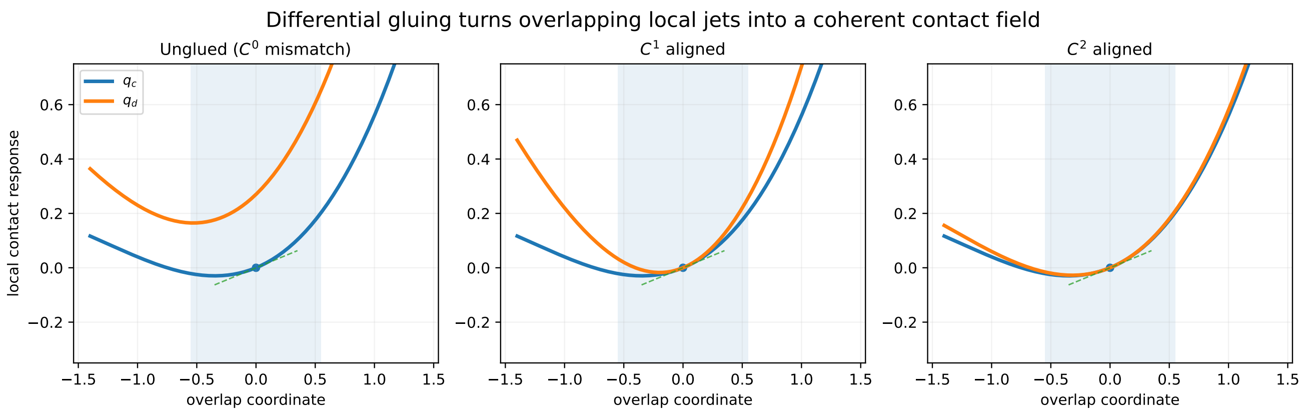

Figure 3: Conceptual role of differential gluing on a region where two contact responsibilities overlap. C^0 aligns local values, C^1 additionally aligns tangents/Jacobians, and C^2 also aligns curvature/Hessian action. The implemented loss measures these quantities on empirical hidden states weighted by $\gamma_c\gamma_d$, rather than only at anchor midpoints.

Proof. For $c \in \mathcal{O}_h$, $d_c^2(x) \geq g_- h^2$, so each numerator is at most $e^{-g_- h^2}$. The denominator is at least the contribution of c_*, which is at least $e^{-g_+ r^2}$. Summing at most K outer contacts gives the stated bound; probabilities are also at most one. □

This proposition provides a rigorous notion of finite-radius approximation under globally supported softmax routing: the tail is nonzero but exponentially small when $g_- h^2 - g_+ r^2$ is large. It parallels recent analysis of softmax as a partition-of-unity mechanism [10] while matching the metric router actually used here.

Theorem 1 (Global order-3 contact approximation with routing tail)**.** *Let $f : \Omega \subset \mathbb{R}^D \to \mathbb{R}^O$ be C^4 on a compact domain. For contacts with $\|x - \mu_c\| < h$, assume the factorized quadratic and cubic tensors approximate the true second and third derivatives with operator errors at most ϵ_2 and ϵ_3, and let $\|D^4 f\| \leq M_4$. Assume also a uniform outer-contact error bound $\|f(x) - q_c(x)\| \leq B$ on Ω. Then*

$$\|f(x) - F(x)\| \leq \frac{1}{2}\epsilon_2 h^2 + \frac{1}{6}\epsilon_3 h^3 + \frac{M_4}{24} h^4 + B\tau_h(x), \tag{15}$$

with τ_h bounded by the previous proposition.

Proof. Split the convex blend into contacts inside and outside radius h. For inside contacts, Taylor's theorem plus rank-approximation errors gives the first three terms. Their routing weights sum to at most one. The outer contribution is bounded by $B \sum_{c \in \mathcal{O}_h} \gamma_c = B\tau_h$. Summing proves the claim. □

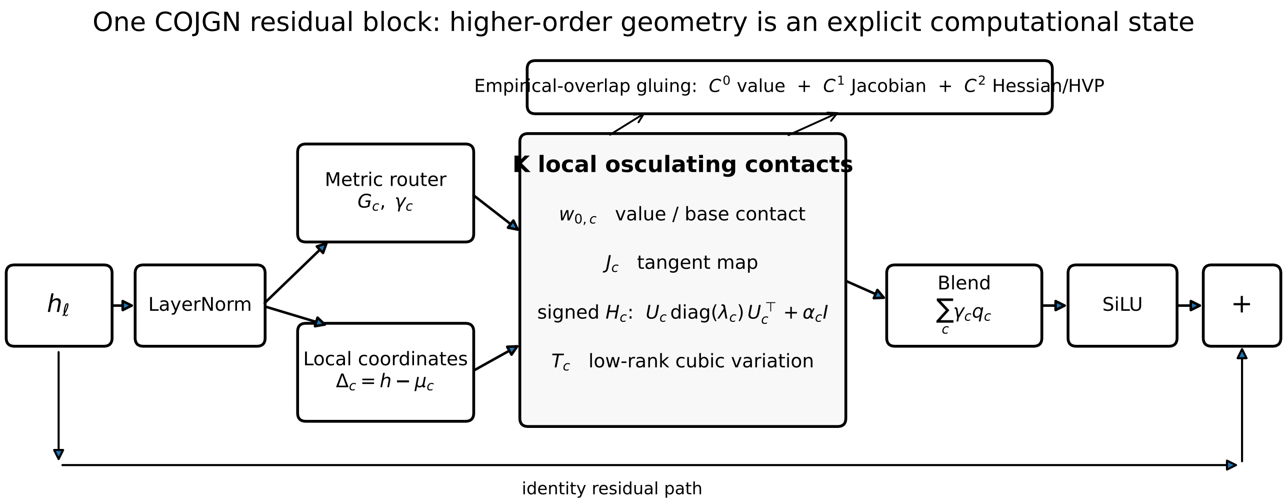

Figure 4: One COJGN residual block. Layer normalization feeds both the metric router and contact-centered coordinates; K osculating contacts evaluate value, tangent, signed curvature, and cubic variation; routing weights blend the local jets; and optional $C^0/C^1/C^2$ penalties act on empirical overlaps. The geometric state is therefore part of the forward computation rather than only a post-hoc diagnostic.

The theorem makes the architecture's intended regime explicit. If local derivative tensors are approximately low rank and the router concentrates mass near a covering anchor, the model inherits order-3 local approximation with an exponentially controlled global-routing correction. If these assumptions fail, no advantage over generic MLP approximation should be expected.

5.1 Whole-layer derivative stability

Because $F = \sum_c \gamma_c q_c$,

$$DF = \sum_c \gamma_c D q_c + \sum_c q_c \otimes D\gamma_c. \tag{16}$$

The second term explains why bounded local polynomial coefficients alone do not imply a stable blended layer. Since $\nabla d_c^2 = 2G_c\Delta_c$ and softmax derivatives are differences of these score gradients, routing precision, metric norms, anchor radius, and local output magnitude jointly enter any Lipschitz bound. The reference implementation therefore regularizes metric factors/precision and clips gradients in addition to controlling curvature/cubic coefficients.

6 Complexity and Deep Residual Composition

For batch size B, K contacts, D latent width, O outputs, and ranks R_g, R_2, R_3, fixed-rank local evaluation scales as

$$O\big(BKDR_g + BKOD(R_2 + R_3 + 1)\big), \tag{17}$$

versus explicit dense quadratic/cubic tensor costs involving D^2 and D^3. This is a tensor-structure advantage, not a guarantee of faster execution than a dense MLP: practical kernels, contact count, and routing overhead matter.

A COJGN residual block uses

$$h_{\ell+1} = h_\ell + \mathrm{SiLU}(\mathrm{COJGL}(\mathrm{LN}(h_\ell))). \tag{18}$$

The training API returns routing/geometry contexts from the same stochastic forward pass used for the task loss, so auxiliary regularization does not recompute a different dropout trajectory.

7 Experimental Design

The revision separates three questions that were previously conflated.

1. **Generic predictive performance:** does COJGN beat parameter-matched MLPs on standard small classification benchmarks?

Table 2: Five-seed standard benchmark accuracy (mean ± standard deviation). Parameter matching is approximate. These results do not support universal superiority of COJGN.

Dataset	Same-width MLP	Matched MLP	COJGN-3/C^2
Digits	83.47 ± 1.31	95.82 ± 0.48	88.71 ± 1.58
Wine	97.33 ± 1.86	98.67 ± 1.22	96.89 ± 1.22
Breast Cancer Wisconsin	96.22 ± 1.45	97.06 ± 1.52	96.78 ± 1.36

Table 3: Five-seed targeted regression MSE (lower is better).

Model	Smooth	Piecewise-smooth
Matched MLP	0.08095 ± 0.00553	0.14272 ± 0.00558
Global quadratic	0.09795 ± 0.01007	0.17879 ± 0.01909
Taylor-style order 3	0.09250 ± 0.01379	0.16496 ± 0.02299
POU-style order 3	0.07351 ± 0.00392	0.12166 ± 0.00558
COJGL-3	0.07701 ± 0.00671	0.12390 ± 0.00831

2. **Inductive-bias test:** does explicit local higher-order structure help on smooth/piecewise-smooth maps or high-dimensional data generated by low-dimensional curved latent structure?
3. **Geometric mechanism:** does differential gluing measurably reduce overlap mismatch even when accuracy does not change?

All reported numbers come from included CSV files. CPU experiments use PyTorch/AdamW with fixed scripts and seed-controlled train/test splits. The POU-style and Taylor-style baselines are intentionally labeled as motif-matched local reimplementations, not official upstream packages.

8 Results

8.1 Standard classification remains a negative control

The matched MLP wins all three means and decisively wins Digits. This reproduces the central falsification concern: geometric parameterization alone is not enough to justify a new primitive on ordinary vector classification.

8.2 Targeted smooth and piecewise-smooth regression

We next construct 12-dimensional regression targets containing sinusoidal smooth structure, signed low-rank quadratic terms, and low-rank cubic terms; the piecewise version adds a continuous but non-C^1 component and region-specific quadratic structure. Five seeds use 4,000 samples and the same optimization budget. COJGL is compared with an approximately parameter-matched MLP, a global signed quadratic model, a global Taylor-style order-3 factorization, and a POU-style learned isotropic local-polynomial model.

The result is more informative than a generic accuracy claim. COJGL reduces MSE relative to the matched MLP by about 4.9% on the smooth target and 13.2% on the piecewise-smooth target, yet it does not beat the POU-style local polynomial model. Thus the data support the *local higher-order inductive bias* while not isolating metric routing or gluing as the source of the entire gain.

8.3 Fixed-budget high-dimensional curved-latent classification

To test width efficiency in a controlled setting, we sample a six-dimensional latent variable, embed it nonlinearly into $D = 128$ observed dimensions, and assign labels using a noisy signed quadratic+cubic decision function in latent space. We use 10,000 observations, three seeds, and match COJGN-3 and MLP parameter counts to within one percent.

This experiment answers the central "where should the inductive bias help?" question: when high-dimensional observations arise from a curved low-dimensional process and the decision map has low effective second/third derivative rank. Because the task is synthetic, it is evidence of mechanism rather than external validity.

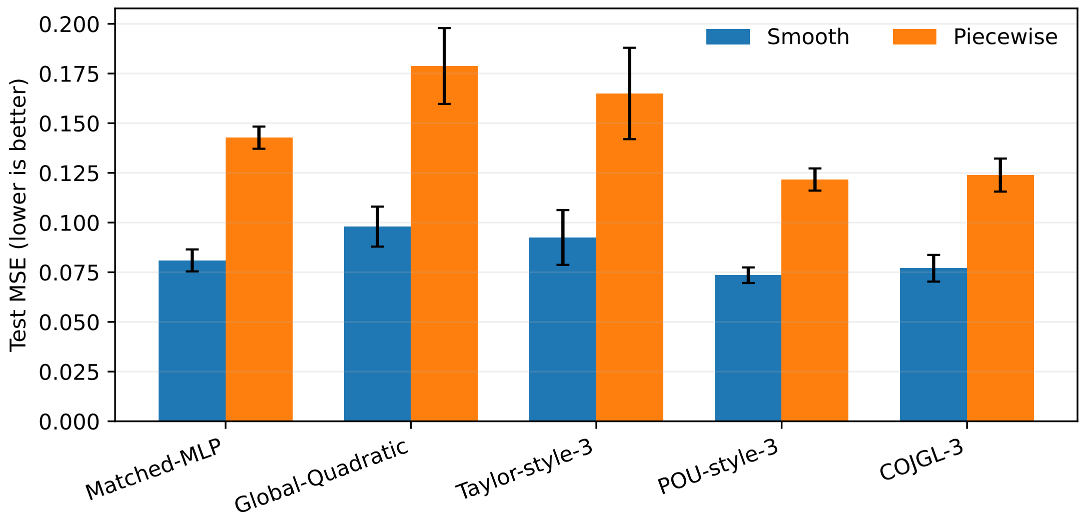

Figure 5: Targeted approximation results. COJGL improves over the matched MLP and global polynomial controls, but the POU-style baseline is slightly better, showing that localization itself is a major explanatory factor.

Table 4: Controlled 128-dimensional classification.

Model	Parameters	Accuracy
Narrow MLP	3,582	90.65 ± 1.52
Matched MLP	23,922	92.15 ± 1.27
COJGN-3	23,750	92.35 ± 1.64

8.4 Five-seed C^2 geometric diagnostics

Accuracy is not an appropriate sole diagnostic for a coherence regularizer. We therefore report routing-weighted value mismatch (C^0), Jacobian mismatch (C^1), Hutchinson Hessian mismatch (C^2), routing entropy, utilization coefficient of variation, and minimum anchor separation.

Strong gluing reduces value mismatch by about 3.4%, Jacobian mismatch by 1.6%, and Hessian mismatch by 2.6% relative to NoGlue, while mean accuracy changes by only -0.13 points. The effect is modest but measurable and supports the intended interpretation: gluing regulates differential coherence, not necessarily task accuracy. Routing entropy and anchor separation remain close, suggesting the router is not simply collapsing to eliminate overlap.

8.5 Direct low-rank cubic-jet recovery

We retain a controlled scalar regression target that is exactly a signed rank-3 quadratic plus rank-2 symmetric cubic jet. Across three seeds, OJGL-3 attains MSE 0.0726 ± 0.0474 with 104 trainable parameters, OJGL-2 obtains 0.2645 ± 0.2061, and a 373-parameter TinyMLP obtains 0.0745 ± 0.0231. This remains the clearest direct evidence that explicit third-order factors can be parameter-efficient when the ground-truth map matches the assumed differential structure.

8.6 Wall-clock and memory cost

The low-rank tensor construction avoids D^2/D^3 tensors, but that does not imply it is faster than an MLP. We profile a representative $D = 64$, batch-512 classifier on CPU in isolated processes.

COJGN is about 3.6× slower for inference, 4.2× slower per training step, and roughly 2.0× higher in measured peak RSS delta in this CPU profile. This is an important negative result. Specialized fused kernels, fewer contacts, block-sparse routing, or top-k contact evaluation are required before computational efficiency can be claimed against ordinary MLPs.

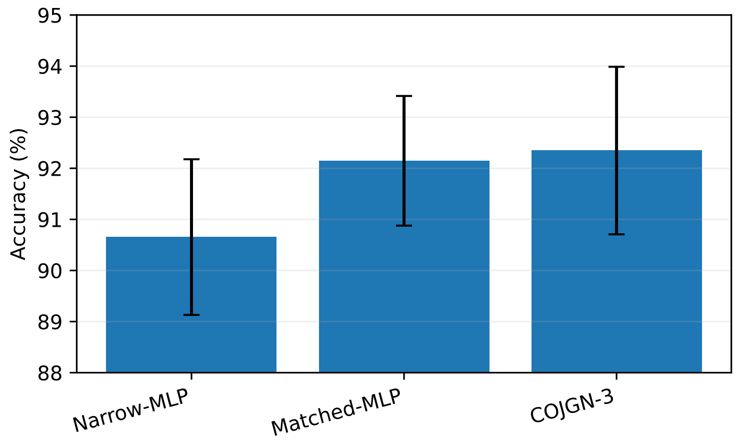

Figure 6: Fixed-budget curved-latent classification. COJGN-3 improves by 0.20 percentage points over the parameter-matched MLP in this deliberately geometry-aligned setting.

Table 5: Five-seed Digits geometry diagnostics. Lower $C^0/C^1/C^2$ is more coherent.

Model	Accuracy (%)	C^0	C^1	C^2	Routing entropy	Min. anchor dist.
NoGlue	94.98 ± 1.00	0.8247	0.009415	0.003717	0.5810	1.4461
C^1	94.98 ± 1.00	0.8221	0.009400	0.003709	0.5808	1.4461
C^2	94.84 ± 1.45	0.8195	0.009397	0.003686	0.5776	1.4475
C^2-Strong	94.84 ± 1.45	**0.7966**	**0.009263**	**0.003620**	0.5771	1.4476

9 Interpretation

The expanded evidence sharpens rather than simply strengthens the claim.

1. **Generic classification:** no advantage. A parameter-matched MLP is clearly better on Digits and slightly better on Wine/BCW.
2. **Local approximation:** positive but not unique. COJGL beats the matched MLP/global polynomial controls, but POU-style localization is slightly stronger in the present synthetic regression tasks.
3. **Geometry-aligned high-dimensional task:** positive. At matched parameters, COJGN improves by 0.20 points on a curved-latent quadratic+cubic classification problem.
4. **Differential gluing:** mechanically validated. Stronger C^2 weighting measurably reduces overlap mismatch without materially changing accuracy.
5. **Compute:** currently unfavorable. Low-rank derivative structure saves against dense higher-order tensors, not against highly optimized MLP kernels.

This pattern is consistent with a specialized inductive bias rather than a universally superior function approximator.

10 Limitations and Next Steps

Official baseline implementations. The included POU-style and Taylor-style controls reproduce architectural motifs under a common training harness. A publication-grade benchmark should additionally run the authors' official implementations, including their native optimizers (e.g., least-squares/EM components for POU/PPOU) and published hyperparameter procedures.

Natural-data scale. The 128-dimensional experiment is larger and directly mechanism-aligned, but still synthetic. The repository includes an optional Fashion-MNIST protocol for replacing MLP blocks with COJGN; those results are intentionally not reported because they were not executed in the present CPU environment. Larger vision, sequence, and scientific-surrogate benchmarks remain necessary.

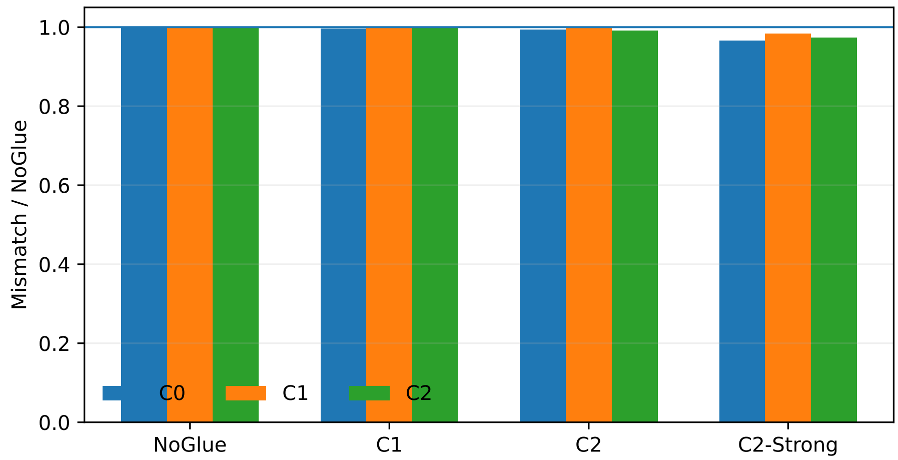

Figure 7: Overlap mismatch normalized to the unglued model. Strong C^2 weighting lowers all three differential mismatch measures while accuracy remains statistically similar.

Table 6: Measured CPU cost at approximately matched parameter count. Latencies are median per batch/step; RSS is process peak delta during the profile.

Model	Params	Inference (ms)	Train step (ms)	Peak RSS delta (MB)
Matched MLP	31,280	1.05	3.46	10.79
COJGN-3	31,770	3.73	14.60	21.82

True atlas structure. Contacts still share latent coordinates. A genuine learned atlas would require local coordinate maps, learned transition functions, and cocycle consistency on triple overlaps [8].

Routing sparsity. Softmax is globally supported. The new theorem correctly accounts for its tail, but computational savings would benefit from differentiable compact-support kernels or top-k routing. Such changes would also produce sharper local approximation constants.

Optimization. The large standard-benchmark gap suggests that residual composition of higher-order contacts remains hard to optimize. Loss-scale normalization, continuation in jet order, curvature-aware initialization, and staged activation of C^1/C^2 penalties are promising directions.

11 Reproducibility

The repository contains the model, tests, raw CSV files, all plotting scripts, and the following experiments:

```
python -m pytest -q
python experiments/benchmark_cojgn.py --seeds 5
python experiments/targeted_regression.py --seeds 5 --epochs 80 --n 4000
python experiments/geometry_diagnostics.py
python experiments/compute_profile.py
python experiments/highdim_classification.py --seeds 3 --n 10000 --epochs 10
python experiments/jet_recovery.py
python figures/review_response_figures.py
latexmk -pdf main.tex
```

The optional `experiments/fashion_mnist_protocol.py` requires torchvision and dataset download access. Unexecuted vision results are not included in any table. Code is intended for release at `https://github.com/andrew-jeremy/Covariant-Osculating-Jet-Networks`.

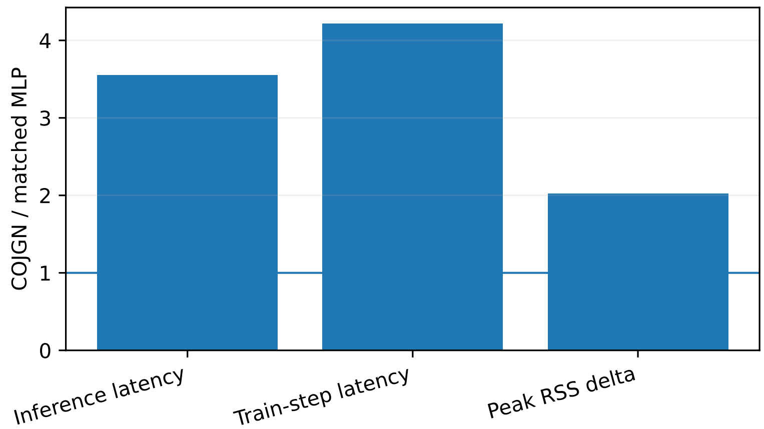

Figure 8: Practical overhead relative to a matched MLP. COJGN's structural tensor savings do not yet translate into faster generic PyTorch execution.

12 Conclusion

COJGL represents hidden transformations by an explicit learned field of local higher-order contacts rather than only affine maps and scalar activations. The revised architecture combines positive metric routing, signed low-rank curvature, symmetric cubic jets, and empirical-overlap differential gluing. More importantly, the revised analysis and experiments make the claim falsifiable. The theory now matches globally supported softmax routing through an explicit exponentially decaying tail term. The experiments show no generic advantage on standard classification and expose a substantial runtime cost, but they also identify a narrower regime where the hypothesis is supported: smooth/local approximation and high-dimensional observations whose underlying decision map has low-rank higher-order structure. Direct C^2 diagnostics confirm that differential gluing changes the geometry it is designed to regulate even when accuracy is neutral. The resulting picture is therefore not "geometry always beats MLPs," but a more useful research proposition: explicit low-rank jet structure can be a competitive inductive bias when local differential complexity is lower than ambient representation dimension, and its value should be judged with geometry-aligned tasks, matched prior-art controls, and direct coherence diagnostics.

References

[1] M. Blondel, M. Ishihata, A. Fujino, and N. Ueda. Polynomial Networks and Factorization Machines: New Insights and Efficient Training Algorithms. *ICML/PMLR 48*, 2016.

[2] M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. arXiv:2104.13478, 2021.

[3] T. Fan, N. Trask, M. D'Elia, and E. Darve. Probabilistic Partition of Unity Networks for High-Dimensional Regression Problems. arXiv:2210.02694, 2022.

[4] J. He, Z. Li, S. Wang, and B. Zhu. Hermite-NGP: Gradient-Augmented Hash Encoding for Learning PDEs. arXiv:2605.24774, 2026.

[5] M. F. Hutchinson. A Stochastic Estimator of the Trace of the Influence Matrix for Laplacian Smoothing Splines. *Communications in Statistics—Simulation and Computation*, 18(3):1059–1076, 1989.

[6] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991.

[7] K. Lee, N. A. Trask, R. G. Patel, M. A. Gulian, and E. C. Cyr. Partition of Unity Networks: Deep hp-Approximation. arXiv:2101.11256, 2021.

[8] R. A. Robinett, S. A. Madejski, K. Ruark, S. J. Riesenfeld, and L. Orecchia. Atlas-based Manifold Representations for Interpretable Riemannian Machine Learning. AISTATS, 2026. arXiv:2510.17772.

[9] D. J. Saunders. *The Geometry of Jet Bundles*. Cambridge University Press, 1989.

[10] Z. Shi and W. Liao. Learning Theory of Transformers: Local-to-Global Approximation via Softmax Partition of Unity. arXiv:2605.08811, 2026.

[11] N. Trask, A. Henriksen, C. Martinez, and E. Cyr. Hierarchical Partition of Unity Networks: Fast Multilevel Training. *Proceedings of Mathematical and Scientific Machine Learning*, PMLR 190:271–286, 2022.

[12] H. Zhao, Y. Chen, D. Sun, Y. Hu, K. Liang, Y. Mao, L. Sha, and H. Shao. TaylorNet: A Taylor-Driven Generic Neural Architecture. OpenReview ICLR submission tDNGHd0QmzO, 2023.