FRONTIERS IN MACHINE LEARNING FOR PHYSICS

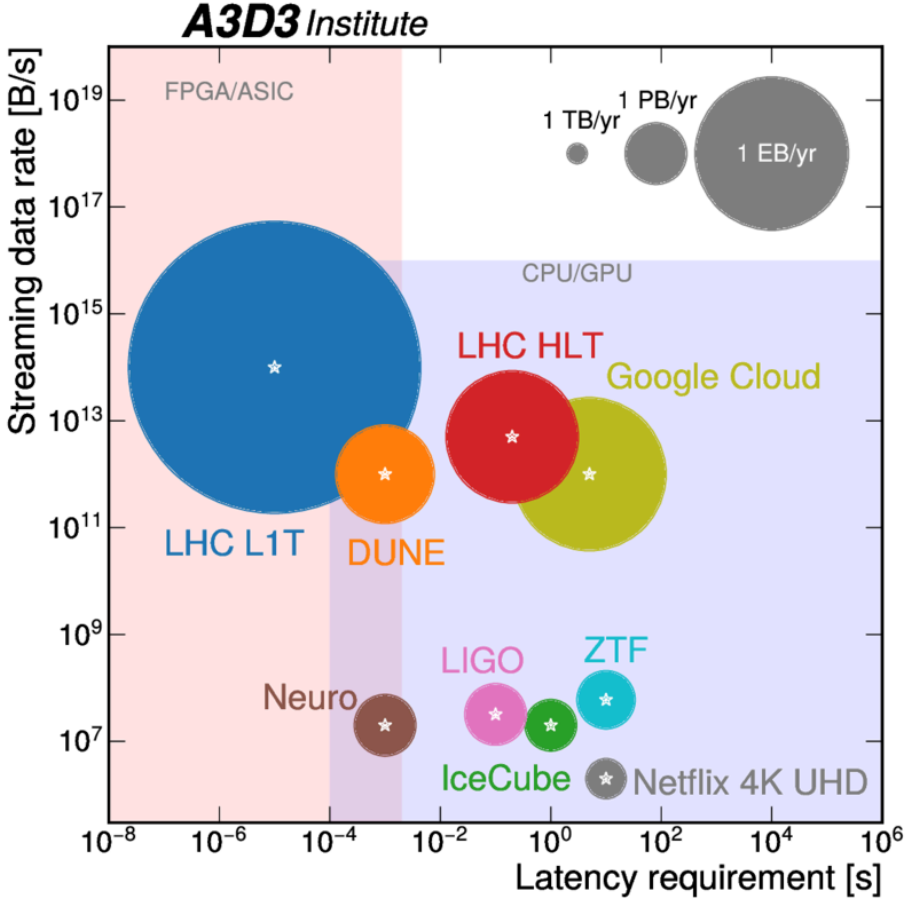
Deven Misra

February 6, 2026

UTokyo | KAVLI IPMU

I. Fast Machine Learning

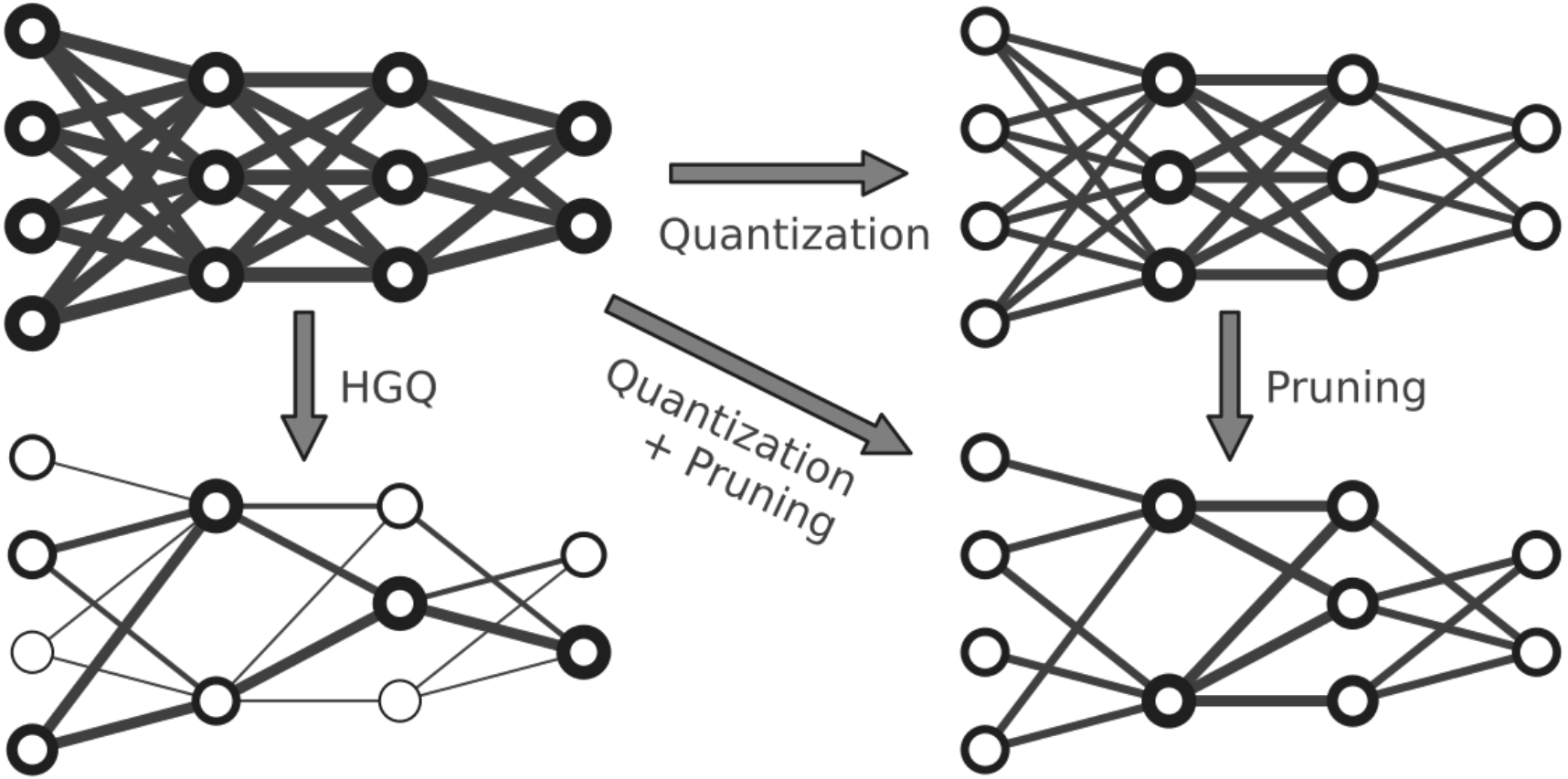
REAL-TIME DATA PROCESSING



Goal: Reduce model **size** and **inference time** while preserving performance

NEURAL NETWORK QUANTIZATION

KERAS → HGQ2 | PYTORCH → BREVITAS

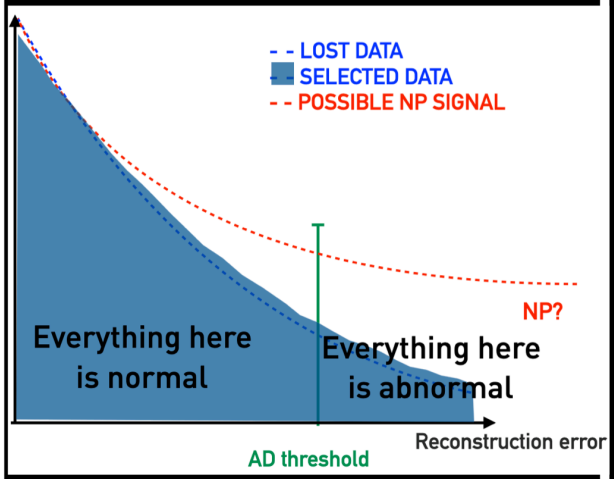


C. Sun *et al.*, *Gradient-Based Automatic Mixed Precision Quantization for Neural Networks On-Chip*, arXiv:2405.00645 **[cs.LG]**, Aug. 2024

EXPERIMENTAL PARTICLE PHYSICS

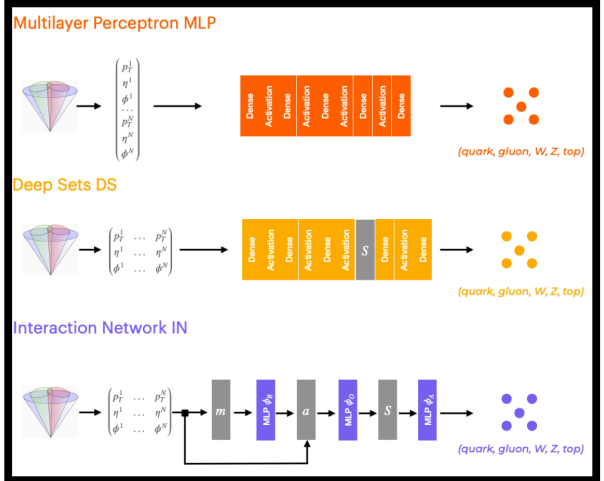
REAL-TIME INFERENCE APPLICATIONS

Anomaly detection with VAEs in 50 ns



CMS DP2023_079
E. Govorkova et al (2022)

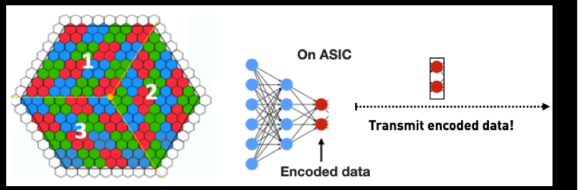
Quantised Interaction Networks and Deep Sets in <160 ns



P. Odagiu et al. 2024

On-detector data compression

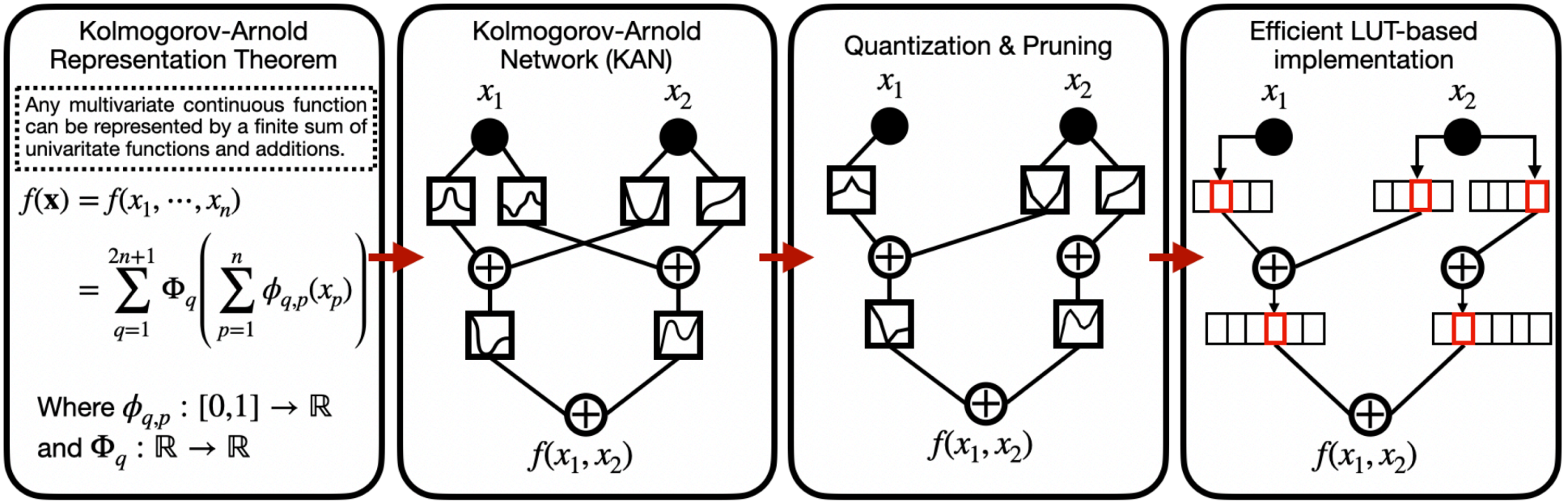
10k ECONS with ML inside in CMS HL-LHC



TWEPP, C. Suarez

EXPERIMENTAL PARTICLE PHYSICS

KAN-LUT



D. Hoang, A. Gupta, P. Harris, and V. Loncar, “FPGA-optimized Kolmogorov-Arnold Networks (KANs) via LUT-based design”, Fast Machine Learning for Science Conference, Sept. 2025

EXPERIMENTAL PARTICLE PHYSICS

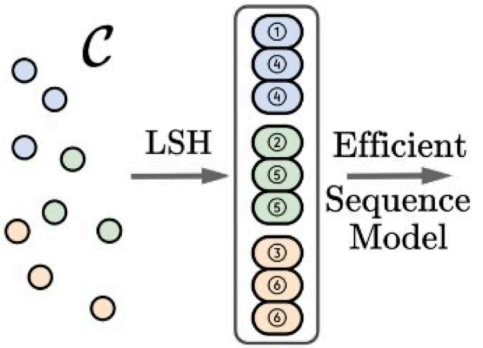
HEPT

- Central idea: **reduce computations** in the attention kernel
- **Locality sensitive hashing** using E2LSH is used to sort points into buckets
- Attention only computed for elements in same bucket

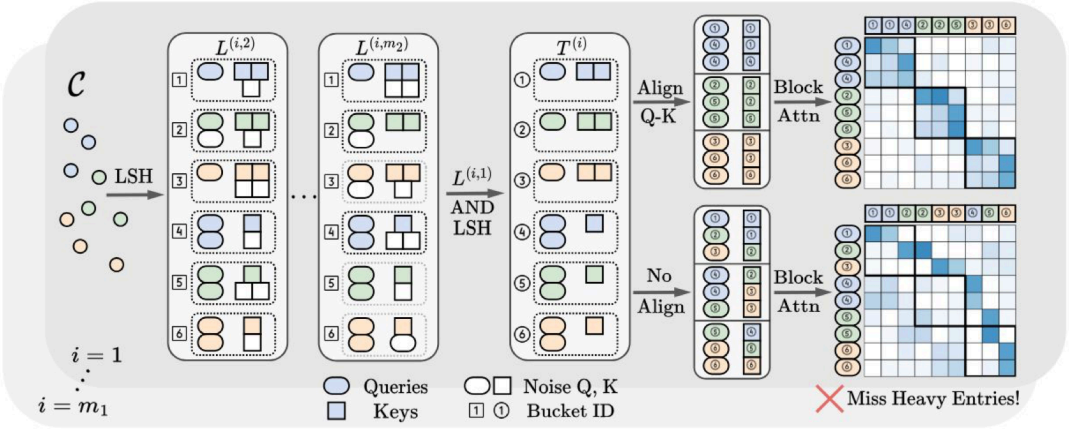
$$\mathbf{x} \in \mathbb{R}^d, h_{\mathbf{a},b}(\mathbf{x}) = \lfloor \frac{\mathbf{a} \cdot \mathbf{x} + b}{r} \rfloor, \mathbf{a} \sim \mathcal{N}(0, \mathbf{I}), b \sim \mathcal{U}(0, r)$$

- Attention **kernel** choice to exhibit **local inductive bias**:

$$k(\mathbf{q}_u, \mathbf{k}_u) = \exp\left(-\frac{1}{2}||\mathbf{q}_u - \mathbf{k}_u||^2\right)$$



- Achieves a **local attention mechanism with linear complexity**
- **Lower inference latency** compared to GNNs by > 10x
- Best resource scaling out of all proposed point cloud transformer models



S. Miao *et al.*, "Locality-Sensitive Hashing-Based Efficient Point Transformer with Applications in High-Energy Physics, Proceedings of the 41st International Conference on Machine Learning", PMLR **235**, 2024

GRAVITATIONAL WAVE ASTRONOMY

PARAMETER ESTIMATION

Accelerated Gravitational Wave Parameter Estimation with Reduced Order Modeling [1]

Rapid Parameter Estimation of Gravitational Waves from Binary Neutron Star Coalescence using Focused Reduced Order Quadrature [2]

Bayesian Parameter Estimation using Conditional Variational Autoencoders for Gravitational-Wave Astronomy [3]

Real-Time Gravitational-Wave Science with Neural Posterior Estimation [4]

Rapid localization and inference on compact binary coalescences with the Advanced LIGO-Virgo-KAGRA gravitational-wave detector network [5]

Rapid Likelihood Free Inference of Compact Binary Coalescences using Accelerated Hardware [6]

---

[1] P. Canizares *et al.*, Phys. Rev. Lett. **114**, 7 (2015)

[2] S. Morisaki and V. Raymond, Phys. Rev. D **102**, 10 (2021)

[4] H. Gabbard *et al.*, Nat. Phys. **18**, 112–117 (2022)

[3] M. Dax *et al.*, Phys. Rev. Lett. **127**, 24 (2023)

[5] S. Morisaki *et al.*, Phys. Rev. D **108**, 12 (2023)

[6] D. Chatterjee *et al.*, MLST **5**, 4 (2024)

GRAVITATIONAL WAVE ASTRONOMY

NOISE REDUCTION

Noise reduction in gravitational-wave data via deep learning [7]

Demonstration of machine learning-assisted low-latency noise regression in gravitational wave detectors [8]

Coherence DeepClean: Toward autonomous denoising of gravitational-wave detector data [9]

DETECTION

GWAK: Gravitational-Wave Anomalous Knowledge with Recurrent Autoencoders [10]

A Machine-Learning Pipeline for Real-Time Detection of Gravitational Waves from Compact Binary Coalescences [11]

ML4GW

---

[7] R. Ormiston *et al.*, Phys. Rev. Research **2**, 3 (2020)

[8] M. Saleem *et al.*, Class. Quantum Grav. **41**, 19 (2024)

[9] C. Reissel *et al.*, arXiv:2501.04883 **[gr-qc]** (2025)

[10] R. Raikman *et al.*, MLST **5**, 5 (2024)

[11] E. Marx *et al.*, Phys. Rev. D **111**, 4 (2025)

II. Geometric Methods

NON-EUCLIDEAN SPACES

RIEMANNIAN MANIFOLDS

Let $\mathcal{M}$ be a smooth manifold. For $\mathbf{x} \in \mathcal{M}$, define the **tangent space** $\mathcal{T}_{\mathbf{x}}\mathcal{M}$ to be the first order approximation of $\mathcal{M}$ around $\mathbf{x}$, which is locally Euclidean. The **Riemannian metric**,

$$g_{\mathbf{x}}(\cdot, \cdot) : \mathcal{T}_{\mathbf{x}}\mathcal{M} \times \mathcal{T}_{\mathbf{x}}\mathcal{M} \to \mathbb{R},$$

defines an inner product on the tangent space $\mathcal{T}_{\mathbf{x}}\mathcal{M}$.

The tuple $(\mathcal{M}, g)$ is called a **Riemannian manifold**.

For $\mathbf{x} \in \mathcal{M}$,

$$\exp_{\mathbf{x}}(\cdot) : \mathcal{T}_{\mathbf{x}}\mathcal{M} \to \mathcal{M}$$

$$\log_{\mathbf{x}}(\cdot) : \mathcal{M} \to \mathcal{T}_{\mathbf{x}}\mathcal{M}$$

GeoOpt

---

[1] S. Bonnabel, “Stochastic Gradient Descent on Riemannian Manifolds”, IEEE Transactions on Automatic Control **58**, 2013

[2] G. Becigneul and O-E. Ganea, “Riemannian Adaptive Optimization Methods”, 7th International Conference on Learning Representations (ICLR), 2019

[3] S. Gruffaz and J. Sassen, “A Review on Riemannian Metric Learning”, arXiv:2503.05321 **[stat.ML]** (2025)

NON-EUCLIDEAN SPACES

SELFMGNN

> Given a graph $G = (V, E, \mathbf{X})$, the problem of **self-supervised graph representation learning** in the mixed-curvature space is to learn an encoding function $\Phi : V \to P$ that maps the node $\mathbf{v}$ to a vector $\mathbf{z}$ in a mixed-curvature space P that captures the intrinsic complexity of graph structure without using any label information.

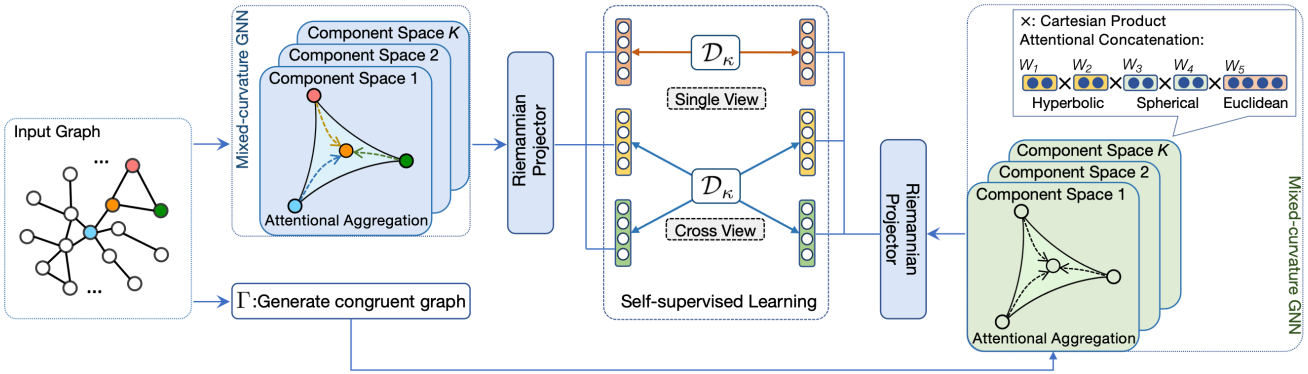
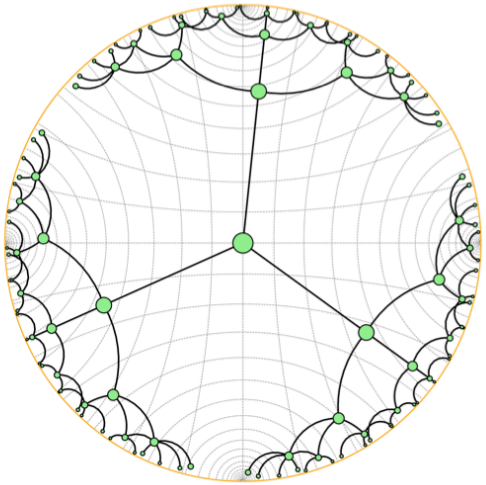


Figure 1: Overall architecture of SELFMGNN: In SELFMGNN, we first introduce a mixed-curvature GNN to learn graph representations. Specifically, we perform attentional aggregation within the component space where the triangle is to show its geometry, e.g., triangles curve inside in $\mathbb{H}$, and attentional concatenation among component spaces whose example with learnable weights is on the top right. Then, to enable self-supervised learning, we design a Riemannian projector to reveal different views of the mixed-curvature space, and utilize a well-designed Riemannian discriminator $\mathcal{D}_\kappa$ to contrast samples for single- and cross-view contrastive learning, shown in the middle. In practice, we feed the graph and its congruent augmentation, generated by $\Gamma(\cdot)$, for the contrastive learning as specified in Algo. 1.

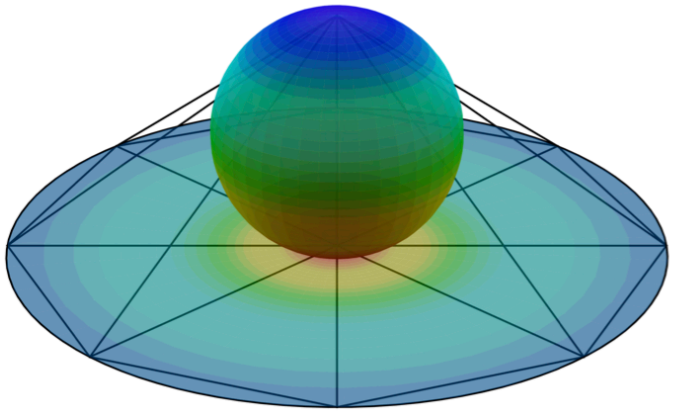
L. Sun *et al.*, "A Self-supervised Mixed-curvature Graph Neural Network", Proceedings of the AAAI Conference on Artificial Intelligence **36**, 2022

NON-EUCLIDEAN SPACES

PRODUCT MANIFOLDS



(a) A rooted tree with a branching ratio of 3 and a depth of 4 is plotted in the 2D Poincaré Disk, illustrating the exponential growth of area and volume characteristic of hyperbolic spaces.



(b) Stereographic projection of 2D sphere onto 2D plane with coloring according to stereographic mapping.

$$\mathcal{P} = \mathcal{M}_{\kappa_1}^{d_1} \times \mathcal{M}_{\kappa_2}^{d_2} \times \cdots \times \mathcal{M}_{\kappa_n}^{d_n} \quad \text{(Product Space)}$$

"...$\mathcal{PM}$ representations generally achieve comparable or superior performance with the most pronounced improvements observed for highly hierarchical jets."

N. Woodward *et al.*, "Product Manifold Machine Learning for Physics", arXiv:2412.07033 **[hep-ph]**, 2024

NON-EUCLIDEAN SPACES

COLLIDER PHASE SPACE

The volume form for four-dimensional, on-shell, massless, N-body phase space in the center-of-mass frame is

$$\mathrm{d}\Pi_N = 2\pi^{4-3N}\left[\prod_{i=1}^{N}\mathrm{d}^4p_i\delta^+(p_i^2)\right]\delta^{(4)}\left(Q-\sum_{i=1}^{N}p_i\right)$$

where $Q=(0,0,0,0)$ and $\delta^+(p_i^2)=\delta(p_i^2)\Theta(p_i^0)$. The phase space Π_N can be separated as

$$\Pi_N \cong \Delta_{N-1}\times S^{2N-3}$$

where Δ_{N-1} is the $N-1$ dimensional simplex, and S^{2N-3} is the $2N-3$ dimensional (hyper)sphere [1]. Furthermore, it is shown that the product space can be equipped with global coordinates which allow for the explicit construction of a Riemannian metric on Π_N,

$$\mathrm{d}_\Pi(\vec{\rho}_A,\vec{v}'_A;\vec{\rho}_B,\vec{v}'_B)=\frac{1}{2\pi}\sqrt{\frac{1}{4}\left(\frac{c}{4}\right)^{\frac{3-2N}{3N-4}}\mathrm{d}^2_\Delta(\vec{\rho}_B,\vec{\rho}_B)+\left(\frac{c}{4}\right)^{\frac{N-1}{3N-4}}\mathrm{d}^2_S(\vec{v}'_A,\vec{v}'_B)}.$$

In this form, d^2_Δ is the line element on the simplex, d^2_S is the line element on the sphere, and c is a dimensionless parameter which reweights the simplex and sphere contributions to the distance between events [2].

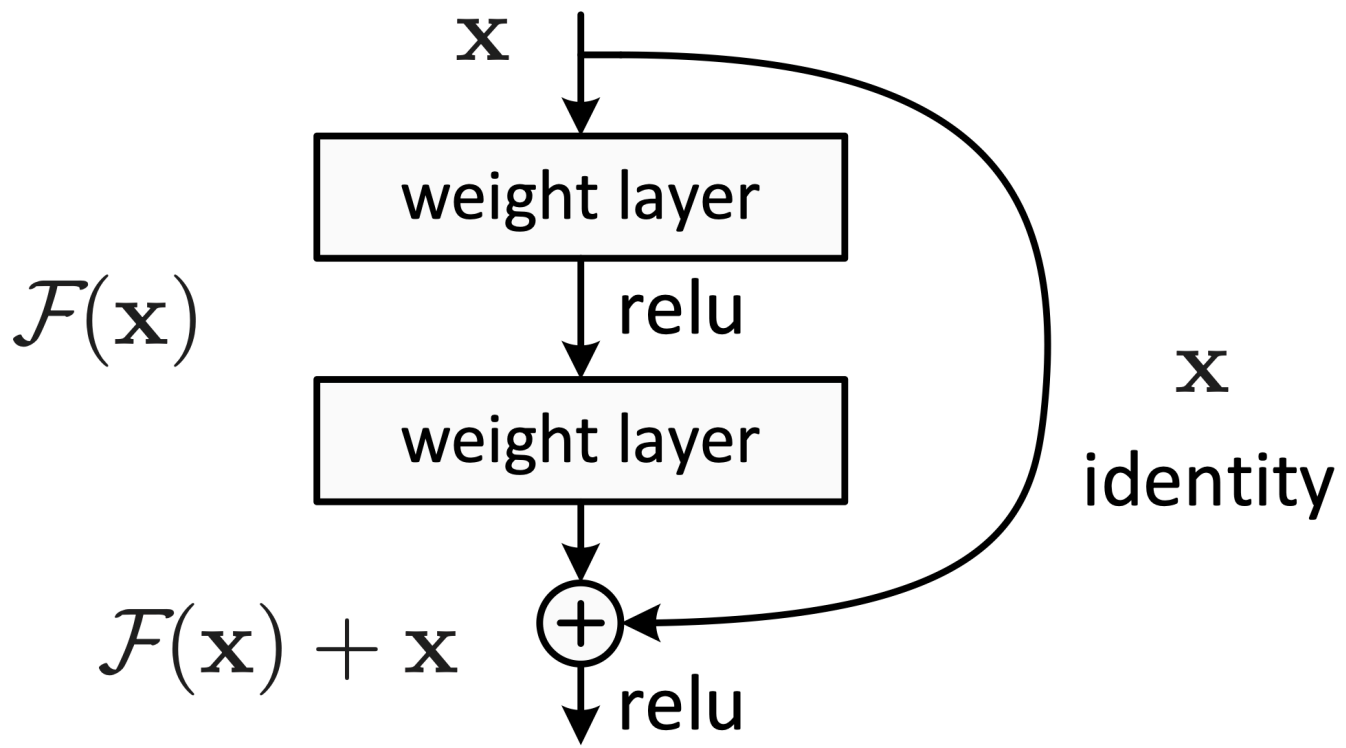
---

[1] A. Larkoski and T. Melia, “Covariantizing Phase Space”, Phys. Rev. D **102**, 9 (2020)

[2] T. Cai *et al.*, “The Phase Space Distance Between Collider Events”, J. High Energ. Phys. **2024**, 54 (2024)

III. Hyper-Connections

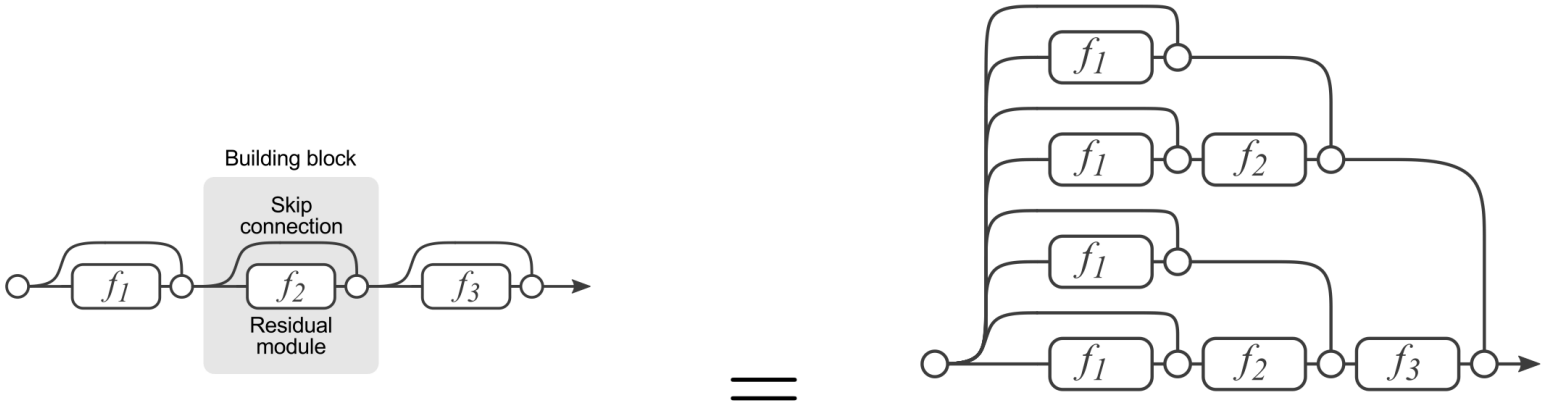
RESIDUAL CONNECTIONS



---

[1] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

RESIDUAL CONNECTIONS



(a) Conventional 3-block residual network

(b) Unraveled view of (a)

Figure 1: Residual Networks are conventionally shown as (a), which is a natural representation of Equation (1). When we expand this formulation to Equation (6), we obtain an *unraveled view* of a 3-block residual network (b). Circular nodes represent additions. From this view, it is apparent that residual networks have $O(2^n)$ implicit paths connecting input and output and that adding a block doubles the number of paths.

---

[1] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," 30th International Conference on Neural Information Processing Systems (NeurIPS), 2016

HYPER-CONNECTIONS

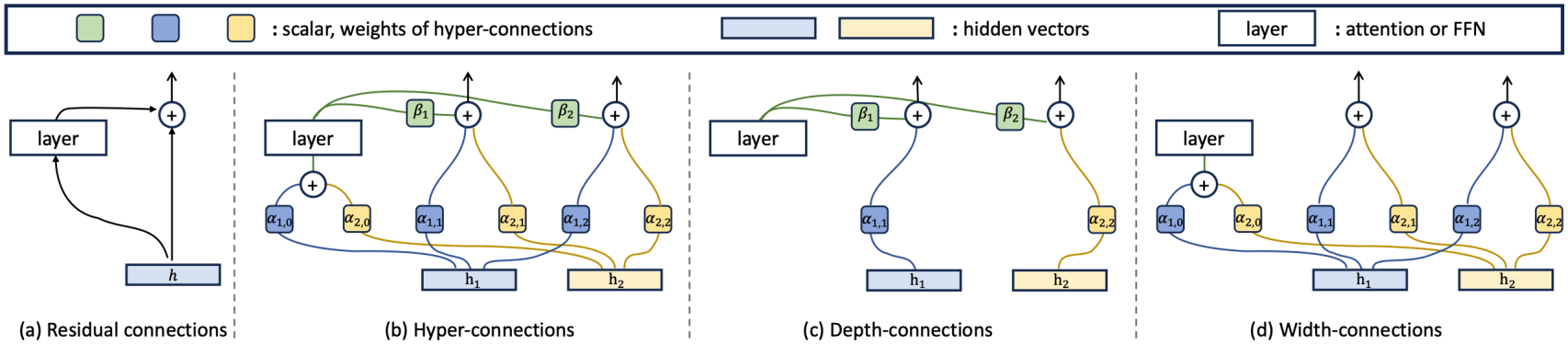
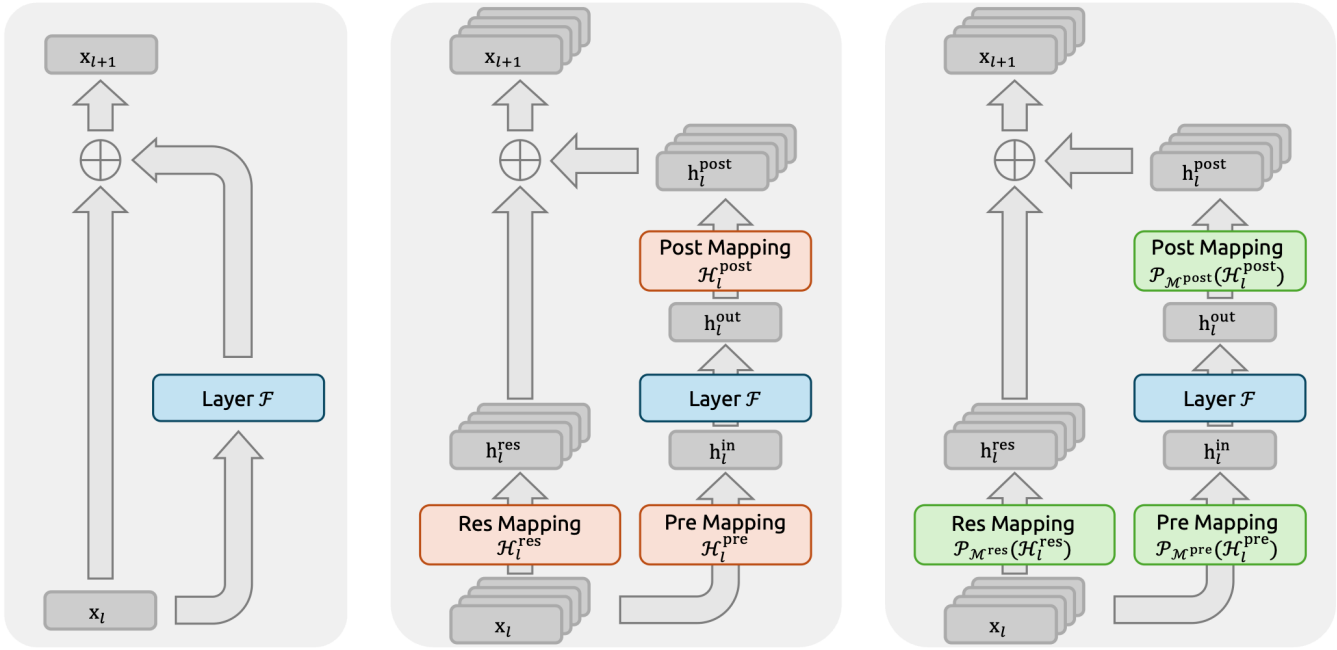


Figure 2: **Hyper-connections (HC) with an expansion rate of $n = 2$.** (a) Residual connections. (b) Hyper-connections: β_1, β_2, $\alpha_{0,0}$, $\alpha_{0,1}$, $\alpha_{1,0}$, $\alpha_{1,1}$, $\alpha_{2,1}$, and $\alpha_{2,2}$ are learnable scalars or scalars predicted by the network , depending on the specific HC version. These connections enable lateral information exchange and vertical integration of features across depths. The Transformer with HC is shown in Fig. 17. They can be decoupled into depth-connections and width-connections. (c) Depth-connections perform a weighted sum between the layer output and the hidden vector h_1. (d) Width-connections allow information exchange between the hidden vectors h_1 and h_2.

$$\mathbf{x}_{l+1} = \mathcal{H}_l^{res}\mathbf{x}_l + \mathcal{H}_l^{post} \cdot \mathcal{F}(\mathcal{H}_l^{pre}\mathbf{x}_l)$$

[1] D. Zhu, “Hyper-Connections”, 15th International Conference on Learning Representations (ICLR), 2025

MANIFOLD-CONSTRAINED HYPER-CONNECTIONS



(a) Residual Connection (b) Hyper-Connections (HC) (c) Manifold-Constrained HC (*m*HC)

Figure 1 | **Illustrations of Residual Connection Paradigms.** This figure compares the structural design of (a) standard Residual Connection, (b) Hyper-Connections (HC), and (c) our proposed **Manifold-Constrained Hyper-Connections** (***m*HC**). Unlike the unconstrained HC, *m*HC focuses on optimizing the residual connection space by projecting the matrices onto a constrained manifold to ensure stability.

---

[1] Z. Xie *et al.*, "mHC: Manifold-Constrained Hyper-Connections," arXiv:2512.24880 [cs.LG], 2026

MANIFOLD-CONSTRAINED HYPER-CONNECTIONS

Let $\mathcal{M}_{res}$ denote the manifold of doubly stochastic matrices (also known as the Birkhoff polytope). We constrain $\mathcal{H}_l^{res}$ to $\mathcal{P}_{\mathcal{M}^{res}}$, defined as:

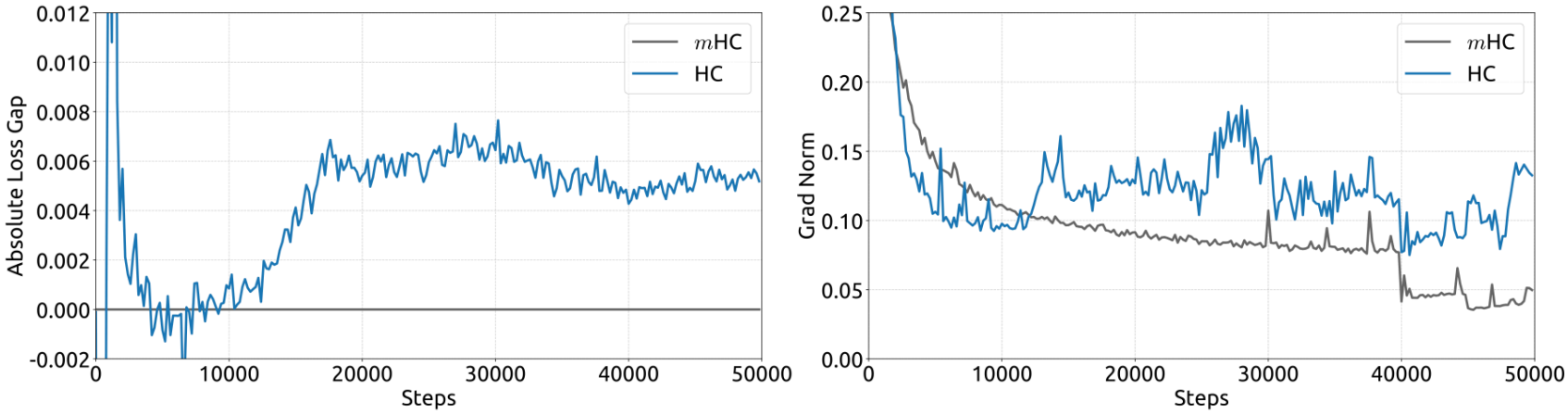
$$\mathcal{P}_{\mathcal{M}^{res}}\left(\mathcal{H}_l^{res}\right) := \left\{\mathcal{H}_l^{res} \in \mathbb{R}^{n\times n} \mid \mathcal{H}_l^{res}\mathbf{1}_n = \mathbf{1}_n, \mathbf{1}_n^T\mathcal{H}_l^{res} = \mathbf{1}_n^T, \mathcal{H}_l^{res} \geq 0\right\},$$

where $\mathbf{1}_n$ represents the n-dimensional vector of all ones.

---

[1] Z. Xie *et al.*, "mHC: Manifold-Constrained Hyper-Connections," arXiv:2512.24880 [cs.LG], 2026

MANIFOLD-CONSTRAINED HYPER-CONNECTIONS



(a) Absolute Training Loss Gap $vs.$ Training Steps

(b) Gradient Norm $vs.$ Training Steps

Figure 2 | **Training Instability of Hyper-Connections (HC).** This figure illustrates (a) the absolute loss gap of HC relative to *m*HC, and (b) the comparisons of gradient norms. All results are based on 27B models.

---

[1] Z. Xie *et al.*, "mHC: Manifold-Constrained Hyper-Connections," arXiv:2512.24880 [cs.LG], 2026

MANIFOLD-CONSTRAINED HYPER-CONNECTIONS

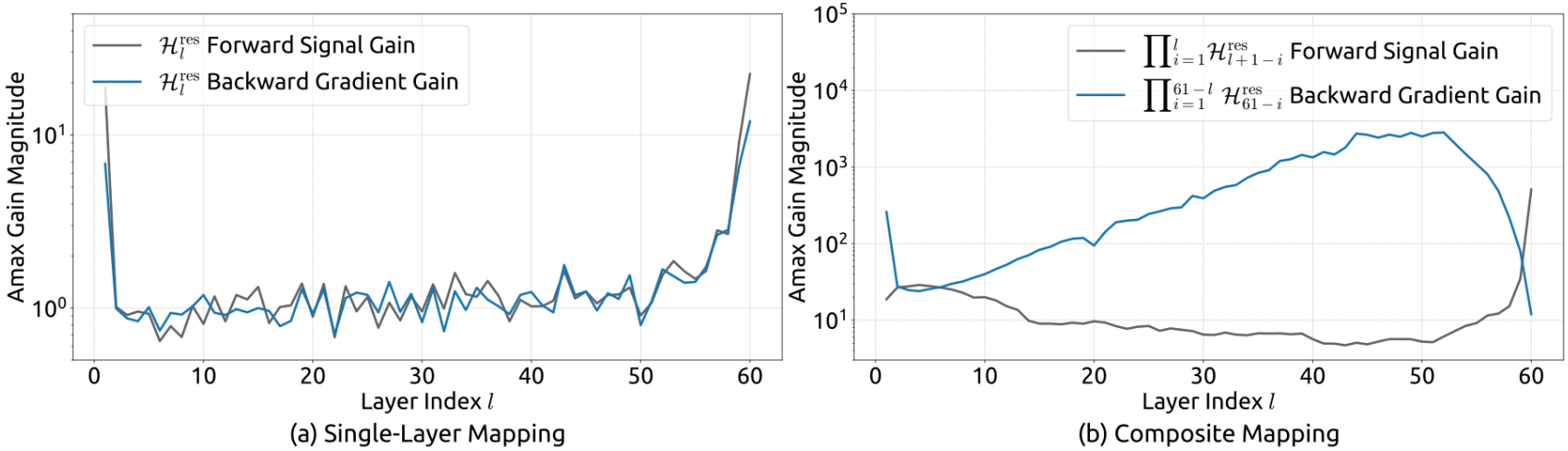
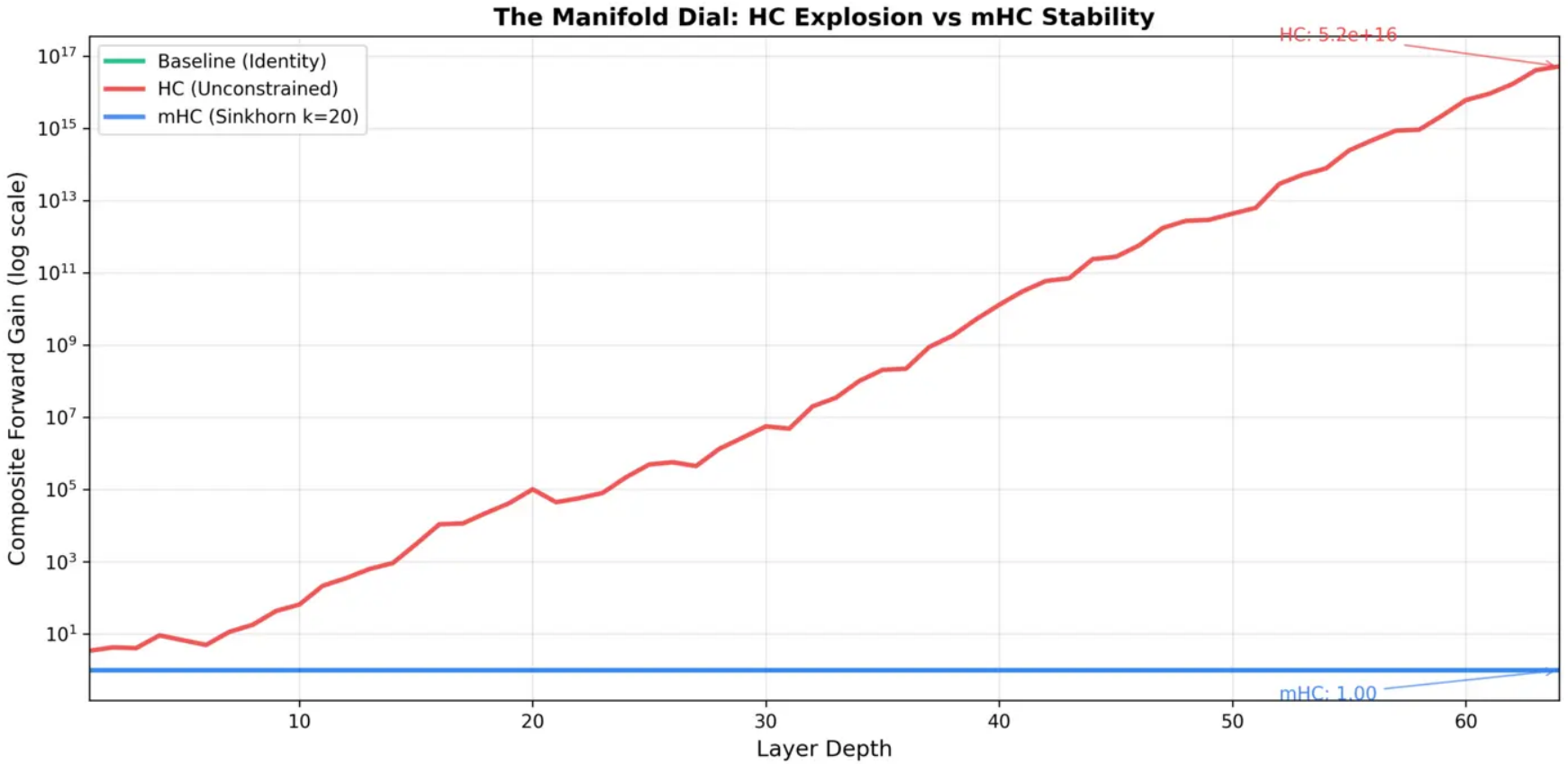


Figure 3 | **Propagation Instability of Hyper-Connections (HC).** This figure illustrates the propagation dynamics of (a) the single-layer mapping $\mathcal{H}_l^{\text{res}}$ and (b) the composite mapping $\prod_{i=1}^{L-l} \mathcal{H}_{L-i}^{\text{res}}$ within the 27B model. The layer index l (x-axis) unrolls each standard Transformer block into two independent layers (Attention and FFN). The Amax Gain Magnitude (y-axis) is calculated as the maximum absolute row sum (for the forward signal) and column sum (for the backward gradient), averaged over all tokens in a selected sequence.

---

[1] Z. Xie *et al.*, "mHC: Manifold-Constrained Hyper-Connections," arXiv:2512.24880 [cs.LG], 2026

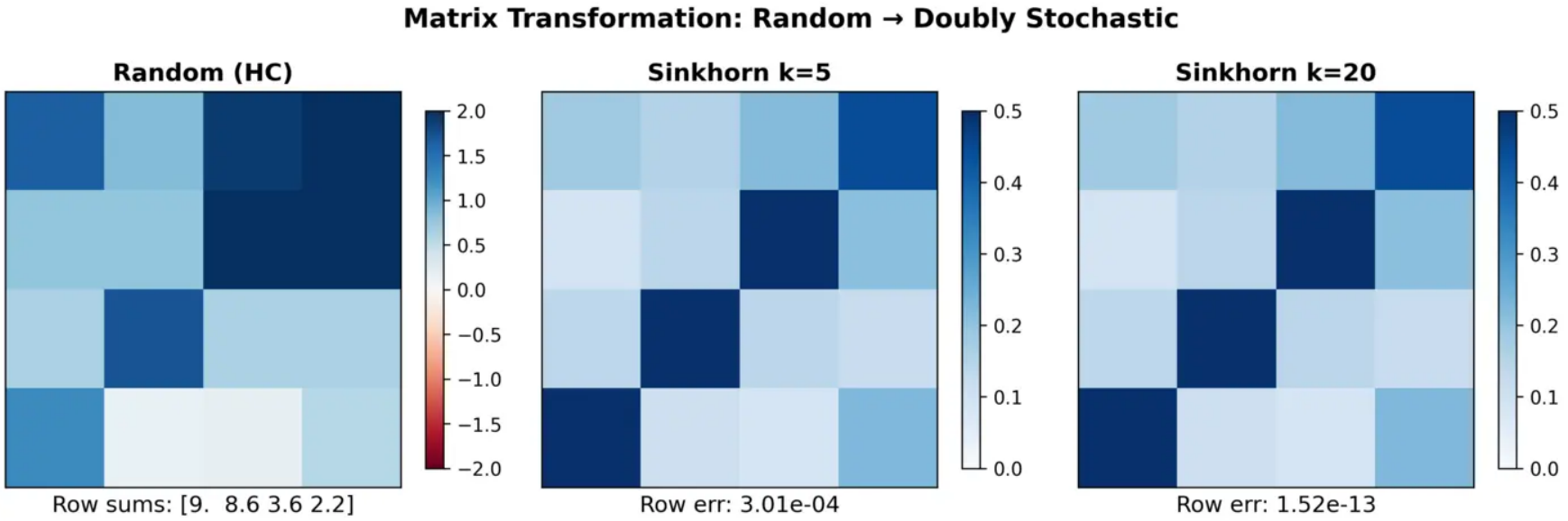
MANIFOLD-CONSTRAINED HYPER-CONNECTIONS



---

[1] Mitra, Subhadip, “The Manifold Dial: Visualizing Why DeepSeek’s mHC Stabilizes Deep Networks” (2026)

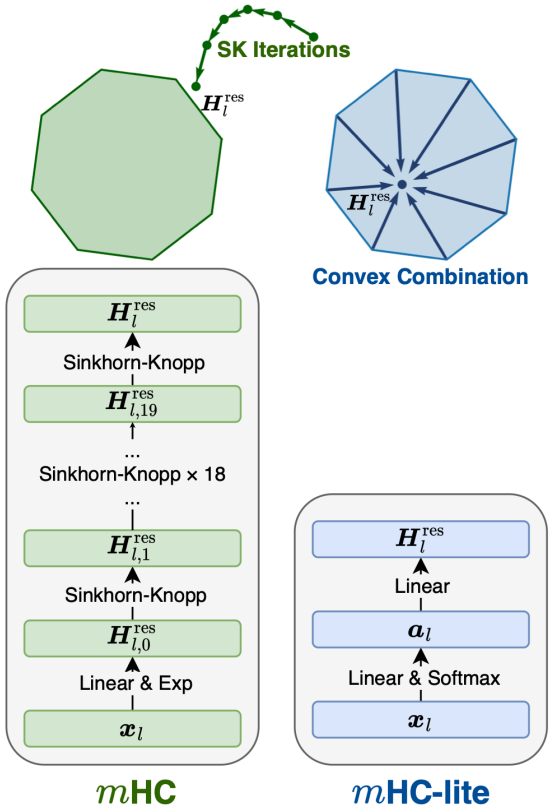
MANIFOLD-CONSTRAINED HYPER-CONNECTIONS



---

[1] Mitra, Subhadip, “The Manifold Dial: Visualizing Why DeepSeek’s mHC Stabilizes Deep Networks” (2026)

MHC-LITE



BIRKHOFF-VON NEUMANN

For any $\mathbf{X} \in \mathcal{B}_n$, $\exists \mathbf{a} \in (a_1, \ldots, a_{n!}) \in \mathbb{R}^{1 \times n!}$ such that

$$\mathbf{X} = \sum_{i=1}^{n!} a_k \mathbf{P}_k,$$

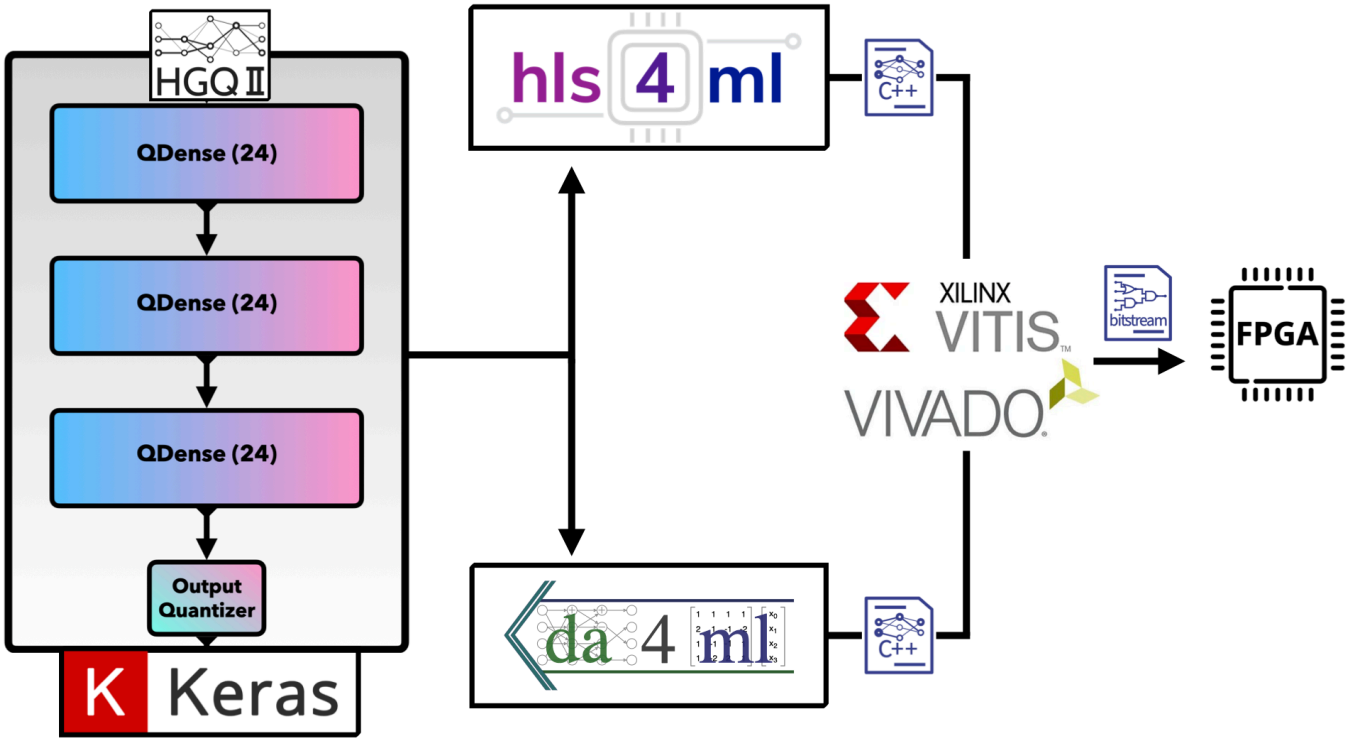
where $\{\mathbf{P}_k\}_{k=1}^{n!}$ is the sequence of $n \times n$ permutation matrices.

---

[1] Y. Yang, J. Gao, “mHC-lite: You Don’t Need 20 Sinkhorn-Knopp Iterations”, arXiv:2601.05732 [cs.LG], 2026

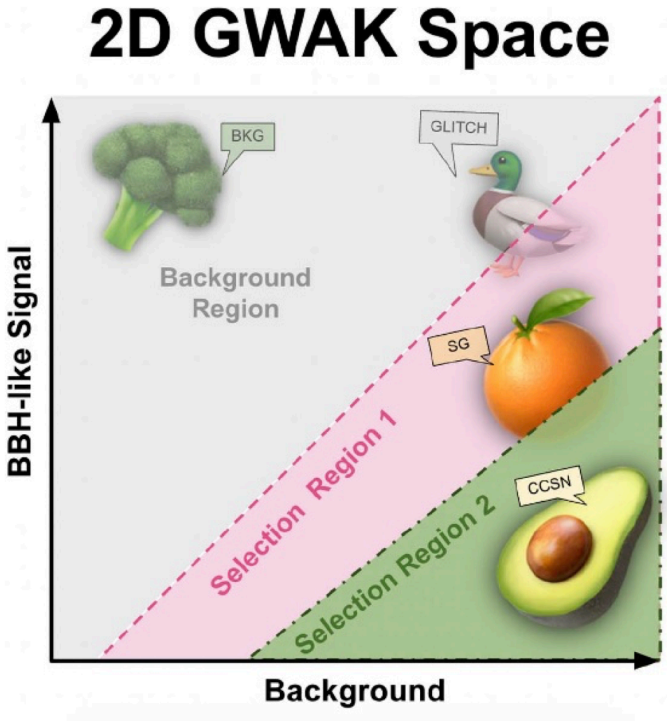
Backup

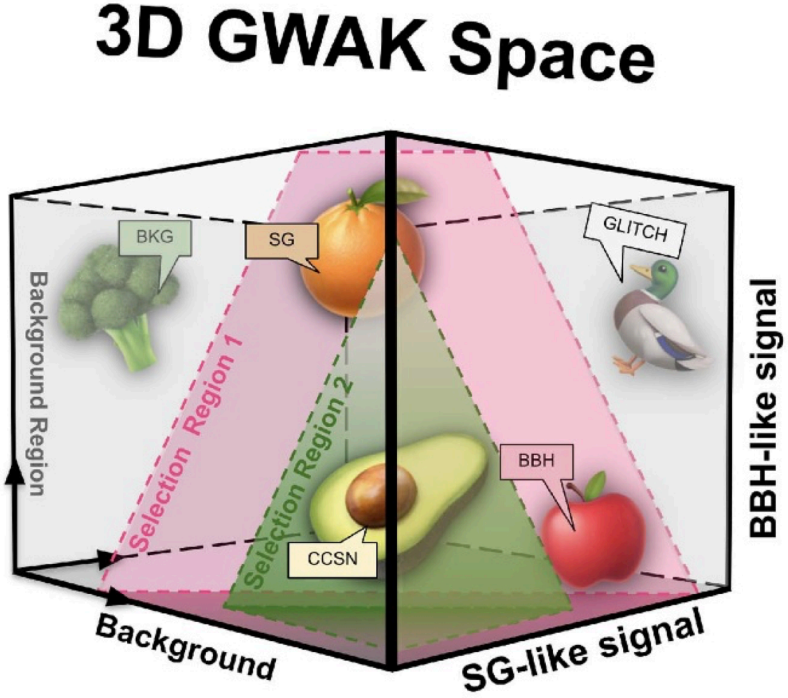
EXAMPLE FASTML WORKFLOW



1. Train model using open-source framework (Keras v3 + HGQ2)
2. Translate model to C++ via high-level synthesis (`hls4ml` and/or `da4ml`)
3. Transpile HLS code to generate register-transfer level implementation
4. Generate bitstream and program device

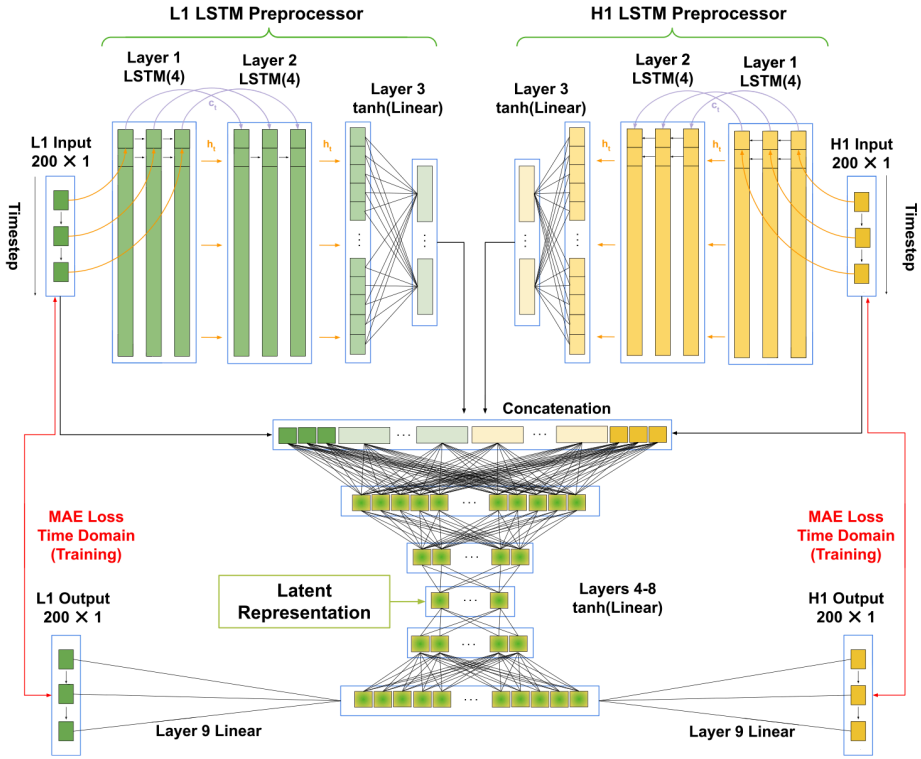
A CLOSER LOOK AT GWAK

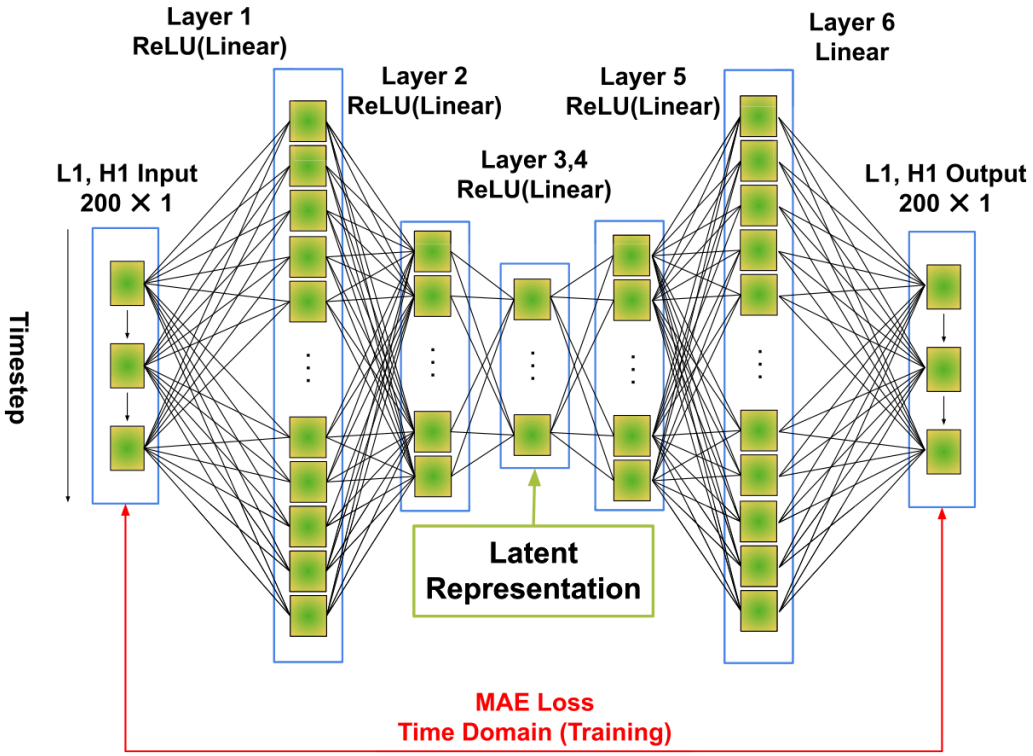




> The semi-supervised method is manifested by using simulated signals as approximations for anomalous-unmodeled signals in the GWAK embedded space. A graph representation of this embedded space is shown in figure 1 for the case of two (left) and three (right) different dataset classes.

A CLOSER LOOK AT GWAK





A separate unsupervised autoencoder network is then trained on each class of samples separately, resulting in a low-dimensional 'GWAK' space consisting of the coherence metrics between the autoencoder inputs and outputs for each autoencoder, which is then used to search for anomalous signals.