A Weisfeiler–Leman Characterization of Global-Attention Graph Transformers for Mixed-Integer Linear Programs

Md Abrar Jahin Craig A. Knoblock Jay Pujara

University of Southern California USC Information Sciences Institute

PROBLEM & QUESTION

MILPs as bipartite graphs

- Constraint nodes ↔ variable nodes; edge weight = coefficient A_{ij} [1]
- Graph foundation models (GFMs) add **global attention**: every node attends to every node [2]
- Claimed to capture structure *beyond local message passing*

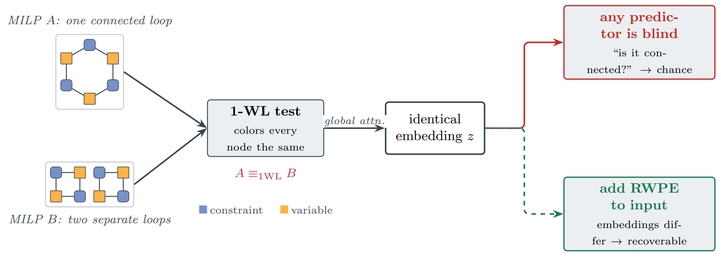
Why embeddings are a bottleneck

- Same embedding ⇒ same prediction for **every** downstream head
- Message-passing GNNs are bounded by the **1-WL** color-refinement test [3, 4, 5]

Does global attention exceed 1-WL?

We answer this exactly, for a broad class of hierarchical graph transformers.

Main Characterization at a Glance



- Two **non-isomorphic** MILPs (connected vs. disconnected) that 1-WL cannot separate receive the **same embedding** z **for any weights** θ
- Any predictor on z outputs the same answer for both — including connectivity itself
- A random-walk positional encoding (RWPE) at the **input** separates them

CONTRIBUTIONS

1. **Exact 1-WL bound** for hierarchical global-attention encoders — compositional proof, one lemma per stage
2. **Parameterized 1-WL-equivalent constructions** with verified non-isomorphism; evaluated on **10 encoders**
3. **Probing + downstream analysis** of what collapsed embeddings cannot predict; encoder-agnostic **diagnostic** released
4. **Expressiveness localized to the input encoding**; limits of finite-length RWPE characterized

Why It Matters & the Released Diagnostic

Concrete stakes for MILP learning

- **Feasibility** itself can differ within a 1-WL class [5] — a 1-WL-bounded encoder cannot be correct on both instances, for any data or capacity
- **Component count** governs decomposition into independent subproblems; **cycle structure** relates to LP-relaxation strength

Encoder-agnostic diagnostic (3 steps)

1. Instantiate a matched 1-WL-equivalent, non-isomorphic pair from the target instance family
2. Encode both members with the frozen encoder under the deployed input encoding
3. Bit-identical embeddings ⇒ encoder at its 1-WL ceiling; separation ⇒ the input encoding supplies the signal

THEOREM: 1-WL BOUNDED FOR ANY WEIGHTS

Theorem 1. If $M_1 \equiv_{1\mathrm{WL}} M_2$, then for *any* weights θ and pooling $\in \{\mathrm{mean}, \mathrm{sum}, \max\}$:

$$\Phi_\theta(M_1) \;=\; \Phi_\theta(M_2)$$

Holds **with virtual global nodes** appended (pre-training augmentation).

Not implied by Xu–Morris [3, 4]

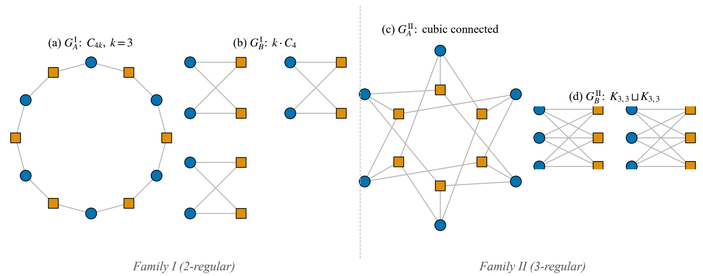
- 3 components are **not message passing**: global self-attention, cross-attention, virtual global nodes

One lemma per stage — all reduce to symmetric multiset aggregation

- L1 SGFormer linear attention: depends on source only via $\sum_\ell k_\ell^\top v_\ell,\ \sum_\ell k_\ell^\top \mathbf{1}$ + global query multiset (Frobenius normalizer)
- L2 Cross-attention: edge message $m_i = \sum_j A_{ij} y_j$ over the neighborhood multiset
- L3 Bipartite GCN branch: MPNN ⇒ 1-WL bounded [5]
- L4 Convex fuse + mean/sum/max pooling: order-independent
- L5 Virtual global nodes: side-wise *means* preserve equivalence

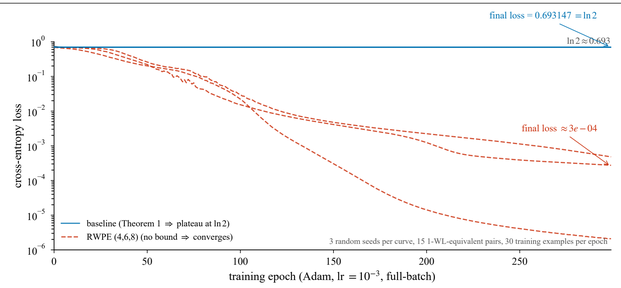
Composition of 1-WL-bounded stages is 1-WL bounded ⇒ Theorem 1. ■

Verified Test Pairs (separation guarantee)



- **Construction** (d-regular, scale k): connected bipartite circulant **vs.** k disjoint $K_{d,d}$ − 1 vs. k components, invisible to 1-WL
- **Families**: cycles (d=2: C_{4k} vs. $k \cdot C_4$) + cubic (d=3 vs. $2 \cdot K_{3,3}$) ⇒ **15 pairs**
- **Validity checks**: exact non-isomorphism (VF2), solver-verified feasibility, 1-WL histograms recomputed after every feature assignment
- Natural MILP features (set cover, auctions, matching, multi-knapsack) ⇒ still bit-identical

Training Cannot Escape the Bound



Loss pinned at $\ln 2$: $\quad 0.69314748 \pm 4.8 \times 10^{-7}$

300 epochs, joint encoder+head training, 5 seeds. With RWPE: $\approx 8 \times 10^{-4}$.

RESULTS: 10 ENCODERS, BIT-IDENTICAL

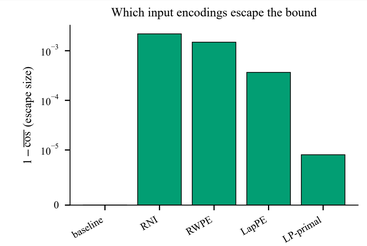
Encoder	params	mean cos	EM
SGFormer+GCN (pre-trained)	6,820	1.000000 ± 0	1.00
SGFormer+GCN (random)	6,820	1.000000 ± 0	1.00
attention-only	1,344	1.000000 ± 0	1.00
GCN-only	5,602	1.000000 ± 0	1.00
minimal GCN	736	1.000000 ± 0	1.00
hierarchical (full)	10,852	1.000000 ± 0	1.00
Graphormer-style	21,616	1.000000 ± 0	1.00
GraphGPS-style	21,568	1.000000 ± 0	1.00
Set-Transformer pooling	13,216	1.000000 ± 0	1.00
Gasse-2019 bipartite GCN	25,600	1.000000 ± 0	1.00

- Worst coordinate gap: $\mathbf{5.96 \times 10^{-8}}$ (float32), $\mathbf{5.6 \times 10^{-17}}$ (float64)
- Invariant across width 16–256, depth 1–8, $10^4 \to 2.5 \times 10^6$ params, graphs up to **800 nodes**, mean/sum/max pooling
- **Specificity**: 1-WL-*distinguishable* pairs *do* separate — no trivial collapse

What Collapsed Embeddings Cannot Predict

- **Probes** (connected components): $R^2 = 0.23 \pm 0.04$ on random MILPs → $\mathbf{-0.01 \pm 0.02}$ on 1-WL population — despite target variance 26.37 vs. 0.22; with RWPE: $R^2 = \mathbf{0.51 \pm 0.05}$
- **Downstream connectivity**: 0.34–0.39 accuracy at baseline → **0.97–1.00** with RWPE (5 of 6 encoders)
- **Prevalence**: WL-twin nodes in **6%–12%** of random instances from standard MILP families

Expressiveness Lives in the Input Encoding



- Return probability $P^4[i,i] = 3/8$ (C_{4k}) vs. $1/2$ ($k \cdot C_4$) ⇒ shortest distinguishing walk L=4
- Walk sweep: cos = 1.0000 ($L \in \{2,3\}$) → 0.984 (L=4) → 0.940 (L=16)
- RWPE, LapPE, RNI separate the pairs [6]; LP-derived features mostly do not
- **Limits**: cospectral pairs stay indistinguishable under finite RWPE (m=10, d=4 for walks $\{4,6,8\}$; m=12, d=4 up to L=16)

Gains attributed to graph-transformer architectures must be read together with the expressive contribution of the input representation.

Encoder-agnostic diagnostic released with the paper.

References

[1] Maxime Gasse, Didier Chételat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[2] Hao Yuan, Wenli Ouyang, Changwen Zhang, et al. OPTFM: A graph foundation model for combinatorial optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[3] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

[4] Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2019.

[5] Ziang Chen, Jialin Liu, Xinshang Wang, Jianfeng Lu, and Wotao Yin. On representing mixed-integer linear programs by graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2023.

[6] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. In *International Conference on Learning Representations (ICLR)*, 2022.