ICML 2026

Are Common Substructures Transferable? Riemannian Graph Foundation Model with Neural Vector Bundles

Li Sun, Zhenhao Huang, Yiding Wang,

Qin Chen, Pietro Liò, Philip S. Yu

Contact: *lsun@bupt.edu.cn* *huangzhenhao@ncepu.edu.cn*

GitHub: https://github.com/RiemannGraph/GAUGE

Motivation & Contributions

Motivations

A fundamental question: Are common substructures truly transferable?

A principled framework for understanding such behavioral invariance is still **lacking**: which structural behaviors learned during pretraining transfer to target graphs?

Highlights

- **A Principled Geometric Understanding.** Bridging transferable structures to intrinsic graph geometry through the lens of behavior invariance.
- **Neural Vector Bundle.** An intrinsic geometry learning framework, which characterizes local region of the underlying manifold with parametric local coordinates.
- **GAUGE Framework.** A novel pretraining architecture with a Dirichlet loss that measures transfer effort.
- **Superior Expressiveness.** Empirically demonstrated on challenging tasks, including graph isomorphism and zero-shot link prediction.

Invariant Substructure & Flat Geometry

An intuition: If a structural behavior learned during pretraining is transferable, it requires little to no adaptation when applied to unseen graphs.

Invariant Substructure

Nodes in a substructure $\mathcal{T}$ are unaffected by the nodes outside this substructure or, in other words, $\mathcal{T}$ is ***invariant*** to its structural neighborhood.

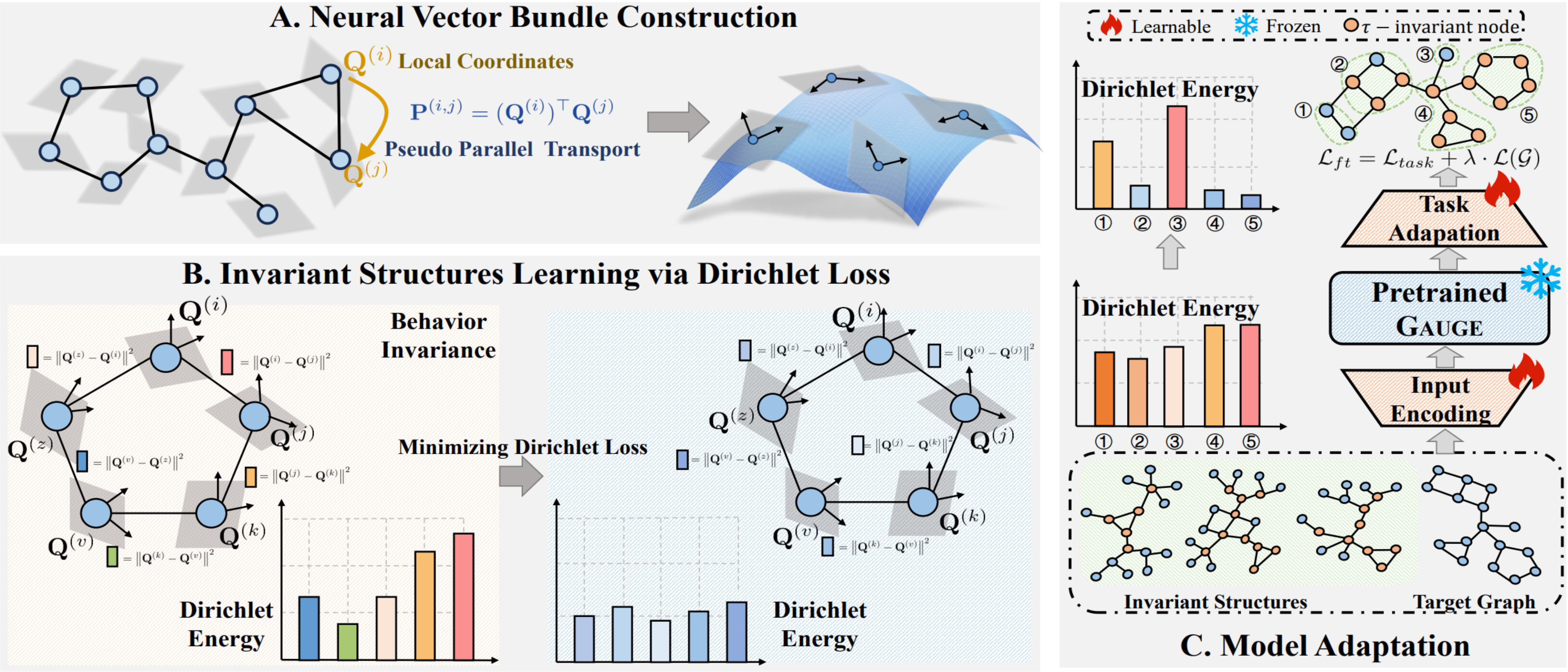
Formalization under Vector Bundle Representation

A substructure $\mathcal{T}$ is said to be invariant if $\left\|x_i - g_{ij}(x_j)\right\| \leq \varepsilon$ holds for any $\varepsilon \to 0$ and $v_j \in \mathcal{N}_i \cap \mathcal{T}, v_i \in \mathcal{T}$, where g_{ij} is the parallel transport from fiber on x_j to the one on x_i.

Theorem 3.1 (Invariance & Intrinsic Flatness (Berwick--Evans et al., 2023), Appendix C.1). A discrete connection of E is flat if and only if holonomy around every 2-simplex is the identity. If there is no 2-simplex, the flat condition is automatically satisfied.

It claims that the intrinsic geometry of vector bundle E over an invariant $\mathcal{T}$ is flat.

Overview of GAUGE Framework



Behavioral Invariance

Find substructures whose local representations predict one another.

Vector bundle

Attach a local coordinate frame $Q^{(i)}$ to every node.

Dirichlet loss

Measure neighbor predictability in the local frame.

Adaptive transfer

Keep the pretrained bundle fixed and adapt only the task head.

Intrinsic Graph Geometry Learning

1 Local Coordinates

For layer l, we decompose $Z^{l-1} \in \mathbb{R}^{N\times d}$ into r disjoint subspaces via learnable projections, yielding $\left\{z_i^{(h),l-1}\right\}_{h=1}^{r}$.
we adopt the multi-head attention, each head h is derived as follows,

$$\hat{\boldsymbol{q}}_i^{(h),l} = \boldsymbol{z}_i^{l-1} - \frac{1}{\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{(h),l}} \sum_{j\in\mathcal{N}_i} \alpha_{ij}^{(h),l} \boldsymbol{z}_j^{(h),l-1} \tag{2}$$

$$\alpha_{ij}^{(h),l} = \frac{\exp(f^{(h)}(\boldsymbol{z}_i^{(h),l-1} \| \boldsymbol{z}_j^{(h),l-1})/\tau)}{\sum_{k=1}^{r}\exp(f^{(h)}(\boldsymbol{z}_i^{(h),l-1} \| \boldsymbol{z}_j^{(h),l-1})/\tau)}, \tag{3}$$

$$\mathbf{Q}^{(i),l} = \mathrm{QR}\left(\left[\hat{\boldsymbol{q}}_i^{(1),l}, \cdot, \hat{\boldsymbol{q}}_i^{(r),l}\right]\right). \tag{4}$$

2 Parallel Transport

A real vector bundle over $\mathcal{G}$ equips each edge $(i,j) \in \mathcal{E}$ a linear isomorphism $g_{ji} = \phi_j \circ \phi_i^{-1}$: $E_{v_i} \to E_{v_j}$, endowing a discrete connection of the total space E and parallel transport. Thus, we obtain the parallel transport between fibers along edges

Definition 3.2 (Pseudo Parallel Transport along Edge). Given an edge $(v_i, v_j) \in \mathcal{E}$ and the fibers of its two end-points E_{v_i} and E_{v_j}, the pseudo parallel transport $\mathbf{P}^{(i,j)}$: $E_{v_j} \to E_{v_i}$ is defined as

$$\mathbf{P}^{(i,j)} = (\mathbf{Q}^{(i)})^{\top}\mathbf{Q}^{(j)}. \tag{5}$$

Accordingly, $\mathbf{P}^{(i,j)}$ captures the geometric torsion when a vector in fiber E_{v_i} moves to E_{v_j} along an edge.

3 Updating Local Coordinates

We introduce **a gated flattening mechanism** that selectively aggregates neighboring fibers according to local coordinates

$$g_{ij} = \frac{\exp(-\mathrm{Tr}(\mathbf{I}_r - \mathbf{P}^{(i,j)})/\tau)}{\sum_{k\in\mathcal{N}_i}\exp(-\mathrm{Tr}(\mathbf{I}_r - \mathbf{P}^{(i,k)})/\tau)} \in (0,1), \tag{6}$$

Neighboring fibers with high compatibility yields large g_{ij}.

$$\hat{\mathbf{Q}}^{(i),k} = (1-\gamma)\mathbf{Q}^{(i),k-1} + \gamma\sum_{j\in\mathcal{N}_i} g_{ij}\mathbf{Q}^{(j),k-1}, \tag{7}$$

$$\mathbf{Q}^{(i),k} = \mathrm{QR}\left(\hat{\mathbf{Q}}^{(i),k}\right), \tag{8}$$

The updating rule iteratively refines local coordinates by smoothing geometrically consistent neighborhoods.

4 Updating global Representations

We leverage the local trivialization for recovering its fiber E_{v_i} and filter noise with the orthogonal complement of its fiber $E_{v_i}^{\perp}$. Additionally, we apply residual connections to allow for the network to go deeper

$$\tilde{\boldsymbol{z}}_j^{l-1} = \mathbf{Q}^{(j),l-1}\left(\mathbf{Q}^{(j),l-1}\right)^{\top}\boldsymbol{z}_j^{l-1}, \tag{9}$$

$$\boldsymbol{z}_i^{l} = \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\mathbf{W}^l\tilde{\boldsymbol{z}}_j^{l-1} + \psi^l(\boldsymbol{z}_i^{l-1}), \tag{10}$$

where $W^l \in \mathbb{R}^{d\times d}$ is a learnable matrix, and ψ^l is a learnable nonlinear function.

Dirichlet Energy & Invariant Substructure

The **Dirichlet Energy** of a graph $\mathcal{G}$ with a vector bundle is defined as

$$\mathcal{E}_{\text{Dir}}[E] = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \left\| \boldsymbol{x}_i - \mathbf{P}^{(i,j)} \boldsymbol{x}_j \right\|^2,$$

Theorem 4.1 (Local Invariance Measure, Appendix C.2.). *Denote local trivializations at each node v_i as $\mathcal{Q}(v_i) = \mathbf{Q}^{(i)}$. Given node representation $\mathbf{Z} = \{\boldsymbol{z}_i\}_{i\in\mathcal{V}}$ satisfying* $\max_i \|\boldsymbol{z}_i\| \le M$*, the following* ***upper bound*** *holds,*

$$\mathcal{E}_{\text{Dir}}[E] \le (4M^2 + 1)(\mathcal{E}_{\text{Dir}}[\mathcal{Q}] + \mathcal{E}_{\text{Dir}}[\mathbf{Z}]), \tag{11}$$

where $\mathcal{E}_{\text{Dir}}[\mathcal{Q}]$ and $\mathcal{E}_{\text{Dir}}[\mathbf{Z}]$ denote the Dirichlet energies associated with the local trivializations and the node embeddings, respectively.

Dirichlet energy gives an **upper bound** of invariance in terms of local trivialization.

That is, minimizing Dirichlet energy is sufficient to learn behavioral invariant substructures.

Theorem 4.2 (Monotonous Flattening, Appendix C.3.). *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and the set of local trivializations $\mathcal{Q}^k$ over $\mathcal{G}$ in the k-th layer of gated flattening with* $\gamma \in (0,1)$*, the layer-wise Dirichlet energy is nonincreasing,* $\mathcal{E}_{\text{Dir}}[\mathcal{Q}^k] \le \mathcal{E}_{\text{Dir}}[\mathcal{Q}^{k-1}]$*, and the final local trivializations satisfies the following inequality,*

$$\mathcal{E}_{\text{Dir}}[\mathcal{Q}^k] \le (1 - C\cdot\gamma)^K \mathcal{E}_{\text{Dir}}[\mathcal{Q}^0], \tag{12}$$

where $C = 1 - \lambda^2$, and λ is the maximal singular value of matrix $\mathbf{G} = [g_{ij}]$.

It claims that GAUGE monotonously lowers the Dirichlet energy over the vector bundle.

It progressively flattens neighboring fibers that are geometrically compatible, and learns invariant substructures through their connection to geometric flatness

Dirichlet Loss & Transferability

The **Dirichlet loss** is formulated in a predictive fashion as follows:

$$\mathcal{L}(\mathcal{G}) = \sum_{i=1}^{N} \left\| \left(\mathbf{Q}^{(i)}\right)^{\top} \hat{\boldsymbol{z}}_i^{(0)} - \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left(\mathbf{Q}^{(j)}\right)^{\top} \boldsymbol{z}_j^{L} \right\|_2^2$$

where $\hat{\boldsymbol{z}}_i^{(0)} = \mathrm{StopGrad}\left(\boldsymbol{z}_i^{(0)}\right)$.

Recovering Invariant Substructures

Definition 4.3 (**τ-Invariance**). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node representation $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ generated by an encoder Φ and a connected substructure $\mathcal{T}$, for each node $v_i \in \mathcal{T}$, its representation predictive error is derived as $e(v_i) = \left\| \left(\mathbf{Q}^{(i)}\right)^{\top} (\boldsymbol{z}_i) - \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left(\mathbf{Q}^{(j)}\right)^{\top} (\boldsymbol{z}_j) \right\|_2$. For a threshold $\tau > 0$, we say the substructure $\mathcal{T}$ is τ-invariance under Φ, or $\mathcal{T}$ is τ-invariant if $\chi(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{v \in \mathcal{T}} e(v) \leq \tau$.

Accordingly, we are able to recover the invariant substructures, specifying the behavior mechanism learned from which substructures during pretraining are appliable to unseen graphs.

Geometry of τ-Invariance

Remarkably, it is sufficient to imply geometric flatness——quasi-identity parallel transport——within this substructure, as formalized below.

Theorem 4.4 (**Flatness Induced by τ-Invariance, Appendix C.4**). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, and node embeddings are $\mathbf{Z} = \{\boldsymbol{z}_i\}_{i \in \mathcal{V}}$. Suppose $\mathcal{T} \subseteq \mathcal{V}$ is a τ-invariant connected subgraph. Assume that there exists $\delta > 0$ such that $\|\mathbf{z}_i - \mathbf{z}_j\|_2 \leq \delta$ for all $(i,j) \in \mathcal{E}(\mathcal{T})$, and that $\|\mathbf{z}_i\|_2 \geq c > 0$. Then, for every edge $(i,j) \in \mathcal{E}(\mathcal{T})$, the pseudo parallel transport map satisfies*

$$\left\| \mathbf{P}_{(i,j)} - \mathbf{I}_r \right\|_F = \left\| (\mathbf{Q}^{(i)})^{\top} \mathbf{Q}^{(j)} - \mathbf{I}_r \right\|_F \leq C(\tau + \delta),$$

where $C > 0$ is a constant depending only on c and r.

On Structural Transferability

The total fine-tuning loss takes the following form,

$$\mathcal{L}_{\text{ft}} = \mathcal{L}_{\text{task}} + \lambda \cdot \mathcal{L}(\mathcal{G}),$$

Theorem 4.5 (**Parameter Stability, Appendix C.5.**). *Let θ be the parameters of the fine-tuned model, and let $\mathcal{T} \subseteq \mathcal{V}$ be a τ-characteristic structure identified during pretraining. Suppose the fine-tuning process minimizes $\mathcal{L}_{ft}$ with $\lambda > 0$. Then, for any node $v_i \in \mathcal{T}$, the gradient update satisfies:*

$$\|\mathrm{grad}_\theta \mathcal{L}_{ft}\|_2 \leq \|\mathrm{grad}_\theta \mathcal{L}_{task}\|_2 + \lambda \cdot C \cdot e(v_i),$$

where $C > 0$ is a constant depending on the Lipschitz continuity of Φ, and $e(v_i)$ is the i-th term in Eq. (13).

GAUGE performs the downstream task with the pretrained neural vector bundle, preserving behaviorally invariant substructures.

Algorithms

Algorithm 1 Training Procedure of GAUGE.

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, initial embeddings $\{\boldsymbol{z}_i^{(0)}\}$, model parameters $\boldsymbol{\Theta}$.

for each training step **do**
  for each network layer $l = 1$ to L **do**
    Compute local coordinates $\{\mathbf{Q}^{(i),l}\}$ via Eqs. 2-4.
    Refine neural vector bundle via a 2-layer gated flattening via Eqs. 6-8.
    Compute node representations $\{\boldsymbol{z}_i^l\}$ via Eq. 10.
  end for
  Stop gradient of $\{\boldsymbol{z}_i^{(0)}\}$ by $\hat{\boldsymbol{z}}_i^{(0)} = \mathrm{StopGrad}(\boldsymbol{z}_i^{(0)})$.
  Compute loss $\mathcal{L}(\mathcal{G})$ in Eq. (13).
  Update $\boldsymbol{\Theta}$ via backpropagation.
end for

Algorithm 2 Decoding Invariant Structures

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, node embeddings $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ encoded by pretrained model Φ, small thresholds $\varepsilon > \tau > 0$.

for each $v_i \in \mathcal{V}$ **do**
  Compute $e(v_i) = \left\| \left(\mathbf{Q}^{(i)}\right)^\top (\boldsymbol{z}_i) - \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left(\mathbf{Q}^{(j)}\right)^\top (\boldsymbol{z}_j) \right\|_2$.
end for
Let $V_{\leq \varepsilon} = \{v_i | e(v_i) \leq \varepsilon\}$.
Find all connected components $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$ from $V_{\leq \varepsilon}$.
Initialize $\mathcal{C} \leftarrow \emptyset$.
for each component $\mathcal{T}_k$ **do**
  Compute average error: $\chi(\mathcal{T}_k) \leftarrow \frac{1}{|\mathcal{T}_k|} \sum_{v \in \mathcal{T}_k} e(v)$
  if $\chi(\mathcal{T}_k) \leq \tau$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathcal{T}_k\}$
  end if
end for
Return $\mathcal{C}$.

Experiments

Table 1. **Main Results on Cross-Domain Transfer.** We report **Mean Accuracy (%)** $\pm$ std over 10 independent runs. Models are pre-trained on source datasets and fine-tuned on target graphs with $k \in \{1, 5\}$ labeled samples (Few-Shot)

Model	PubMed		Facebook		Roman-empire		Photo	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
GCN (Kipf & Welling, 2017)	40.91 ± 6.71	51.34 ± 6.46	44.43 ± 7.22	55.25 ± 4.41	11.29 ± 2.07	18.19 ± 1.38	43.03 ± 8.10	58.34 ± 2.20
GraphSAGE (Hamilton et al., 2017)	41.36 ± 9.78	51.76 ± 6.83	47.23 ± 9.05	59.93 ± 4.27	11.34 ± 4.32	18.24 ± 2.50	42.53 ± 6.67	54.55 ± 3.26
GAT (Veličković et al., 2018)	39.86 ±10.65	50.37 ± 6.30	46.26 ± 7.19	59.82 ± 5.56	11.71 ± 2.47	18.20 ± 1.51	41.74 ± 9.16	56.96 ± 2.88
DGI (Veličković et al., 2019)	41.64 ± 7.72	58.20 ± 2.71	42.57 ± 1.81	55.66 ± 2.74	9.32 ± 1.99	13.74 ± 1.17	42.36 ± 4.78	63.20 ± 2.68
GraphMAE (Hou et al., 2022)	43.63 ± 5.81	56.54 ± 4.85	43.35 ± 4.98	58.93 ± 4.60	16.57 ± 3.23	17.00 ± 1.35	44.32 ± 9.74	62.97 ± 4.38
GRACE (Zhu et al., 2020)	41.04 ± 6.51	59.66 ± 6.49	49.79 ± 7.82	56.26 ± 5.77	7.71 ± 3.66	13.84 ± 5.53	47.26 ± 6.00	66.06 ± 4.33
GFT (Wang et al., 2024)	52.39 ± 9.07	63.11 ± 4.72	45.79 ± 3.02	46.33 ± 2.36	18.63 ± 4.80	26.60 ± 2.75	51.83 ± 5.75	76.53 ± 2.96
RAGraph (Jiang et al., 2024)	51.39 ± 6.93	65.71 ± 3.45	46.01 ± 0.93	48.88 ± 1.54	10.51 ± 2.33	20.99 ± 2.17	50.62 ±11.25	70.28 ± 6.20
SAMGPT (Yu et al., 2025)	43.51 ± 8.43	68.81 ± 8.98	47.88 ±10.23	56.38 ± 7.69	10.39 ± 6.15	25.76 ± 6.20	49.03 ±10.87	71.52 ± 7.26
GCOPE (Zhao et al., 2024)	44.78 ± 2.82	67.66 ± 1.13	45.80 ± 1.78	55.94 ± 1.41	12.91 ± 0.50	16.39 ± 0.61	60.36 ± 3.28	68.68 ± 1.18
GraphAny (Zhao et al., 2025a)	54.18 ± 2.21	70.19 ± 2.20	45.46 ± 3.78	45.48 ± 3.78	18.71 ± 1.14	**26.72** ± 1.14	60.46 ± 0.84	78.45 ± 0.84
MDGFM (Wang et al., 2025a)	45.44 ± 7.12	63.62 ± 8.19	48.08 ±10.25	50.48 ± 8.74	14.16 ± 5.81	23.66 ± 6.17	51.50 ±11.30	65.16 ±11.52
RiemannGFM (Sun et al., 2025b)	56.82 ± 9.00	64.03 ± 4.60	44.08 ± 6.69	57.88 ± 4.17	13.08 ± 1.64	24.63 ± 1.17	64.35 ± 8.10	78.41 ± 1.78
GET (Zhao et al., 2025b)	42.37 ± 5.06	64.55 ± 3.97	44.26 ± 2.92	54.06 ± 3.57	9.36 ± 0.40	15.38 ± 0.96	53.54 ± 3.49	72.44 ± 1.06
G^2PM (Wang et al., 2025d)	43.84 ± 8.75	65.63 ± 4.70	45.13 ± 5.92	54.68 ± 4.13	10.92 ± 2.84	16.81 ± 1.25	54.46 ± 5.59	72.19 ± 4.34
UniPrompt (Huang et al., 2025)	40.34 ± 7.36	63.64 ± 4.97	42.60 ± 4.70	44.55 ± 3.22	7.20 ± 1.94	17.52 ± 2.18	55.12 ± 3.82	70.32 ± 4.12
GAUGE (Ours)	**61.26** ± 5.43	**71.63** ± 3.89	**51.61** ± 9.53	**60.04** ± 1.39	**18.78** ± 4.88	26.43 ± 1.47	**64.72** ± 4.53	**81.33** ± 2.93

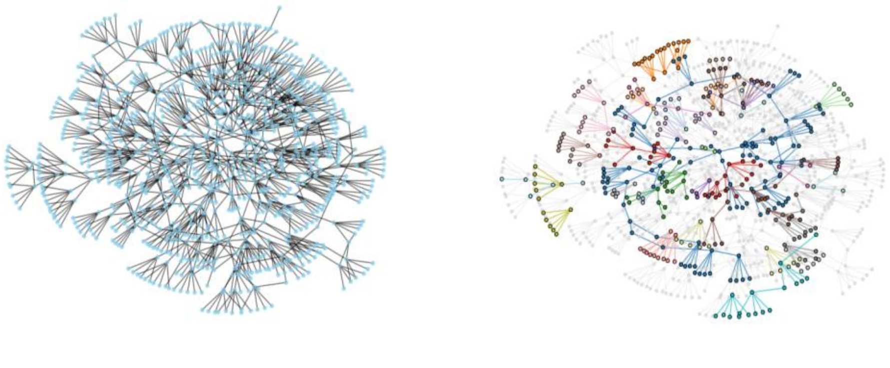
Experiments

Table 2. **Zero-Shot Link Prediction.** We report **AUC and AP scores (%)** ± std over 10 runs. Models are pre-trained on source datasets and directly inferred on target graphs without fine-tuning.

Model	PubMed		Facebook		Roman-empire		Photo	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP
GFT	55.83±0.55	56.57±0.57	72.94±0.25	74.93±0.20	54.91±0.43	55.11±0.48	54.13±0.27	54.88±0.23
RAGraph	58.29±0.32	55.41±0.41	87.41±0.11	83.83±0.19	49.88±0.62	52.92±0.62	74.06±0.27	72.25±0.28
SAMGPT	48.91±0.80	55.30±0.87	84.61±0.12	85.55±0.11	59.53±0.55	62.51±0.33	92.26±0.16	90.84±0.28
GCOPE	60.34±0.43	**61.91**±0.25	80.37±0.20	82.22±0.15	58.19±1.00	57.12±0.94	80.14±0.41	81.00±0.42
MDGFM	56.60±0.33	56.70±0.37	71.14±0.18	63.56±0.14	63.78±0.72	62.56±1.01	73.09±0.28	65.35±0.23
RiemannGFM	53.37±1.33	50.91±1.53	76.65±1.04	72.62±0.84	62.59±0.63	64.73±0.76	66.93±0.91	62.38±1.48
GET	48.82±1.90	48.80±1.75	50.80±0.33	51.97±0.33	63.70±0.56	64.26±0.51	53.06±0.37	53.42±0.36
G^2PM	53.08±0.15	54.62±0.18	71.45±0.11	72.18±0.12	60.66±0.54	56.83±0.65	71.61±0.27	69.61±0.45
UniPrompt	56.32±0.16	53.71±0.08	77.54±0.28	72.36±0.44	54.16±0.54	54.34±0.43	68.64±0.33	63.17±0.27
GAUGE	**64.03**±1.37	61.40±1.03	**93.88**±0.05	**90.73**±0.24	**66.22**±1.26	**67.21**±0.58	**93.04**±2.04	**91.04**±1.76

Table 3. **Graph Isomorphism Classification Results.** We use ACC (%) for CSL, AUC (%) for MUTAG, and MAE for ZINC12K.

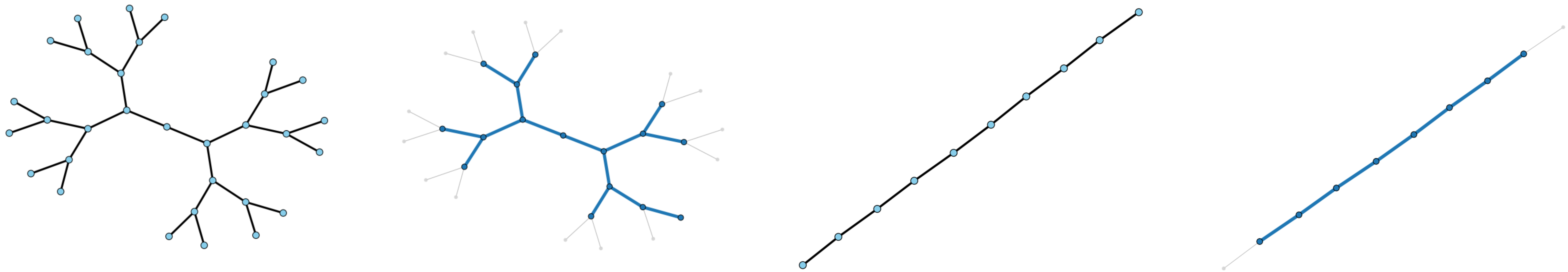
Model	CSL↑	MUTAG↑	ZINC12K↓
GCN	26.00 ± 9.79	86.26 ± 9.20	0.321 ±0.009
GraphSAGE	21.67 ± 5.50	86.48 ±10.77	0.588 ±0.006
GIN	29.67 ± 8.23	82.75 ± 9.54	0.163 ±0.004
GAT	22.33 ± 7.54	84.29 ± 9.54	0.384 ±0.007
SAMGPT	64.17 ±15.24	60.23 ±13.07	0.382 ±0.025
GAUGE	**92.56** ± 4.37	**88.59** ± 5.75	**0.157** ±0.005



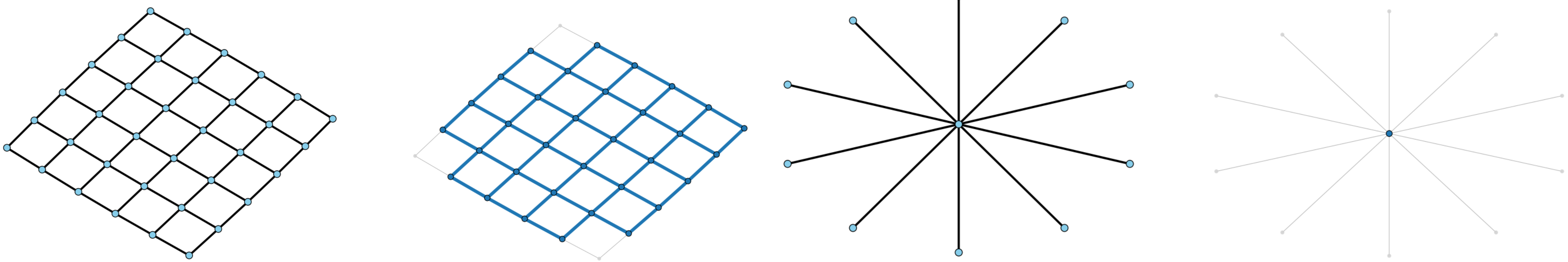
(a) Sampled subgraph *(b)* Invariant structures

Figure 2. **Visualization on `Computers`.** Different invariant substructures are marked with distinct colors.

Case Study



(a) Binary tree. *(b)* Binary tree. *(a)* Path. *(b)* Path.



(a) Grid. *(b)* Grid. *(a)* Star. *(b)* Star.

Figure 3. Case study on typical typologies. Identified invariant substructures are highlighted.

Thanks for Listening

Contact

lsun@bupt.edu.cn

huangzhenhao@ncepu.edu.cn

GitHub: https://github.com/RiemannGraph/GAUGE