Distilling Graph Geometry:
Knowledge Gap from GNNs to MLPs

Zhewei Chen[1] · Hao Zhu[2] · Jiaojiao Jiang[3] · Ahad N. Zehmakan[1]

[1] Australian National University [2] CSIRO Technology [3] University of New South Wales



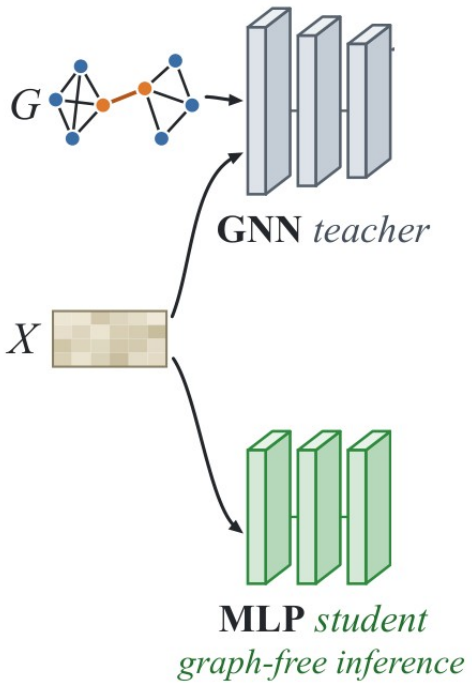






NeurIPS 2026 · Sydney

Why distil a GNN into an MLP?



GNN teacher

Reads node features and graph structure

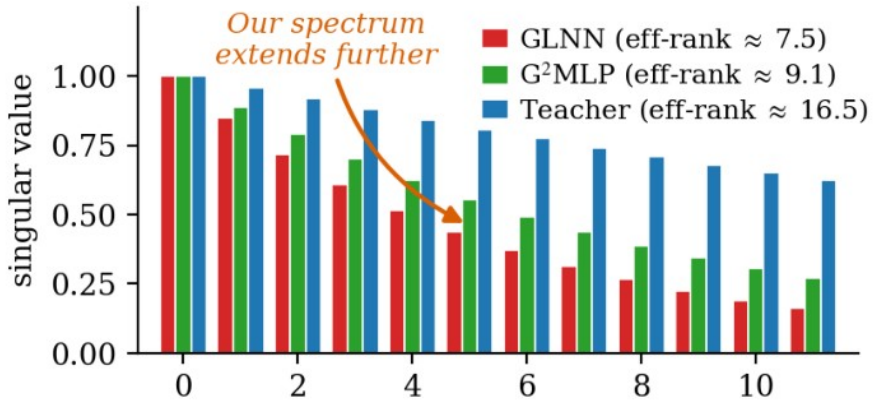
MLP student

Reads node features only at inference

What structural knowledge can the student recover?

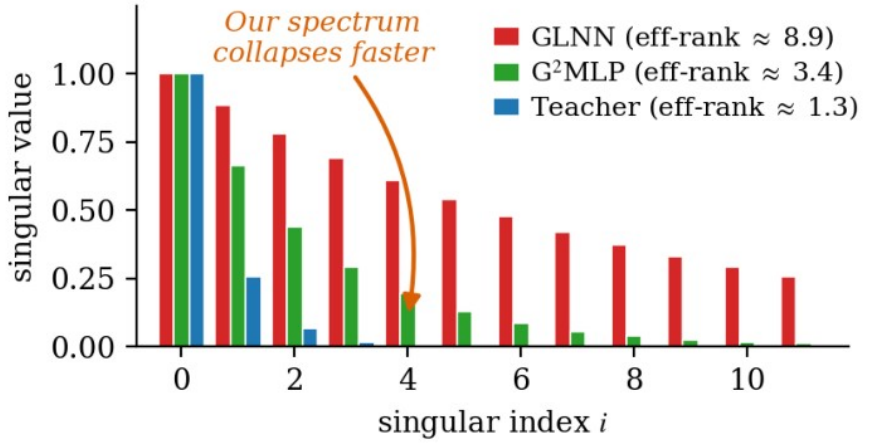
Two opposite spectral failures

The student can miss teacher directions—or retain variation that the teacher has removed.



Spectral underfit

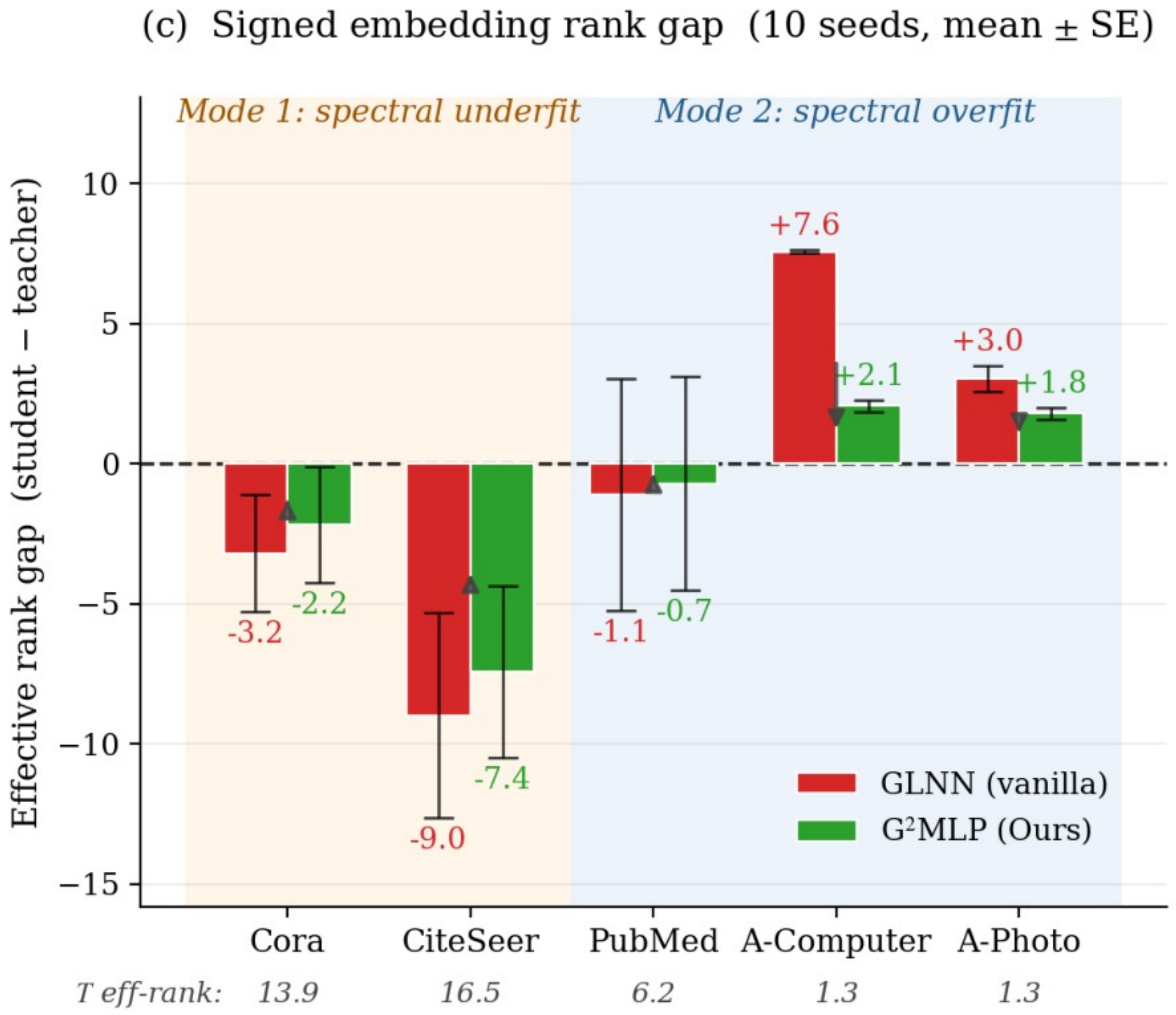
Missing informative directions



Spectral overfit

Extra unsupported directions

Both regimes appear in real graphs

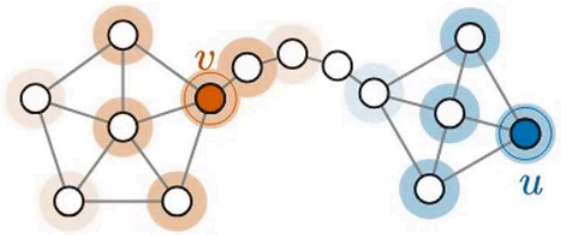


G²MLP narrows the rank gap in both regimes.

Where does the spectral mismatch come from?

Receptive fields

aggregation locality · Lemma 3.1

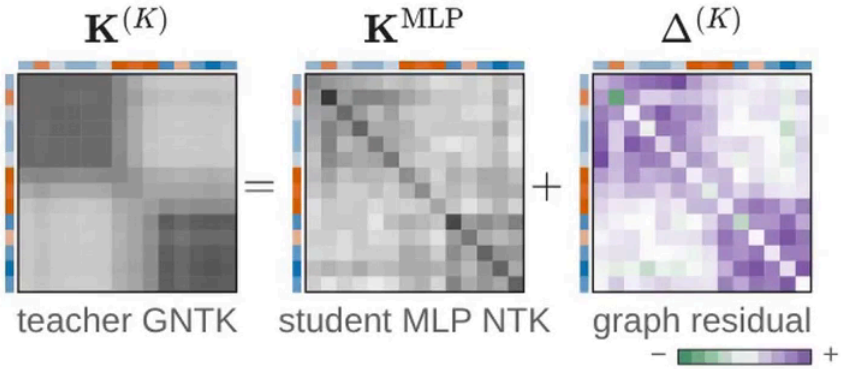


- boundary node: its 2-hop field straddles both communities
- interior node: its 2-hop field stays inside one community

K(v,u) depends on the graph only through the K-hop fields of v and u

Kernel decomposition

GNTK vs. MLP NTK on this graph (unit trace)



$$\mathbf{K}^{(k)}(v,u) = \underbrace{\mathcal{A}[\mathbf{\Sigma}^{(k-1)}](v,u)}_{\text{aggregation}} + \dot{\mathbf{\Sigma}}^{(k)}(v,u)\,\mathcal{A}[\mathbf{K}^{(k-1)}](v,u)$$

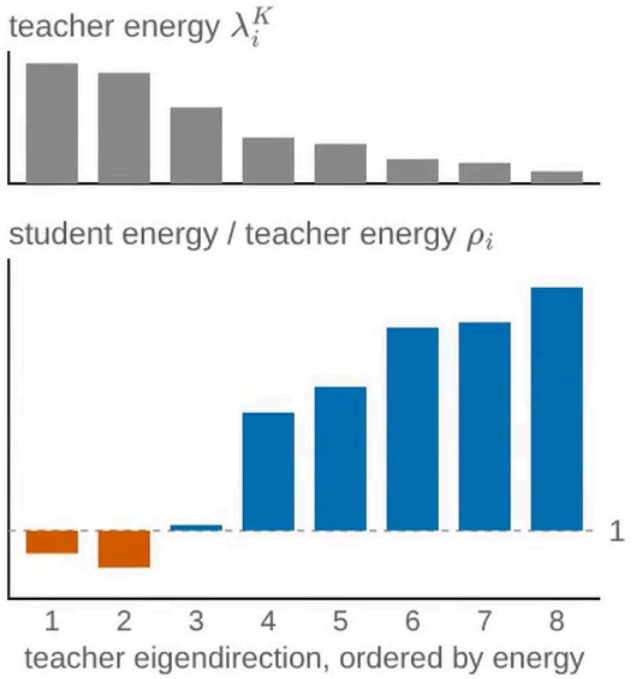
the aggregation term is absent from the MLP recursion

$\Delta^{(K)}(v,u)$ is generated by paths inside $\mathcal{N}_K^+(v) \cup \mathcal{N}_K^+(u)$

the student can realise the MLP kernel, never the residual

Spectral distortion

$\rho_i = \mathbf{u}_i^\top \mathbf{K}^{\mathrm{MLP}} \mathbf{u}_i \,/\, \lambda_i^K$



- $\rho_i < 1$ underfit — missing high-energy directions
- $\rho_i > 1$ overfit — directions the teacher collapsed

GNN aggregation adds kernel interactions unavailable to a feature-only MLP.

Three practical relaxations

The ideal objective, relaxed three times

Section 4

$$\mathcal{L}^* = \left\|\mathbf{f}^T - \mathbf{f}^S\right\|^2_{\mathbf{K}^{(K)}} = \sum_i \lambda_i^K \left(\hat{f}_i^T - \hat{f}_i^S\right)^2$$

energy-weighted mismatch along the teacher's eigendirections; needs the full spectrum of the teacher kernel → intractable

↓ ❶ global kernel $\mathbf{K}^{(K)} \rightarrow \{\nu_v\}_{v\in\mathcal{V}}$ local neighbourhood measures (Lemma 3.1)

$$\mathcal{L}_{\mathrm{GW}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathcal{W}_2^2\left(\nu_v^T, \nu_v^S\right), \qquad \nu_v = \sum_u \mu_v(u)\,\delta_{\mathbf{r}_u}$$

optimal transport between teacher and student neighbourhood measures; exact OT: $O(|N^+(v)|^3)$ per node

↓ ❷ discrete optimal transport $\mathcal{W}_2 \rightarrow \mathcal{W}_{2,\mathrm{diag}}$ diagonal-Gaussian surrogate (closed form)

$$\mathcal{W}^2_{2,\mathrm{diag}}\left(\tilde{\nu}_v^T, \tilde{\nu}_v^S\right) = \left\|\mathbf{m}_v^T - \mathbf{m}_v^S\right\|_2^2 + \left\|\sqrt{\boldsymbol{\sigma}_v^T + \epsilon} - \sqrt{\boldsymbol{\sigma}_v^S + \epsilon}\right\|_2^2$$

moment matching with a diagonal Gaussian: mean shift + spread mismatch; O(d) per node, closed form

↓ ❸ uniform node weights $\frac{1}{|\mathcal{V}|} \rightarrow \alpha_v,\ \beta_v$ curvature-adaptive weights (Prop. 3.2)

$$\mathcal{L}_{\mathrm{CAW}} = \frac{1}{|\mathcal{V}|}\sum_v \left[\beta_v\, \mathcal{W}^2_{2,\mathrm{diag}}\left(\tilde{\nu}^T_{h,v}, \tilde{\nu}^S_{h,v}\right) + \alpha_v\, \mathcal{W}^2_{2,\mathrm{diag}}\left(\tilde{\nu}^T_{z,v}, \tilde{\nu}^S_{z,v}\right)\right]$$

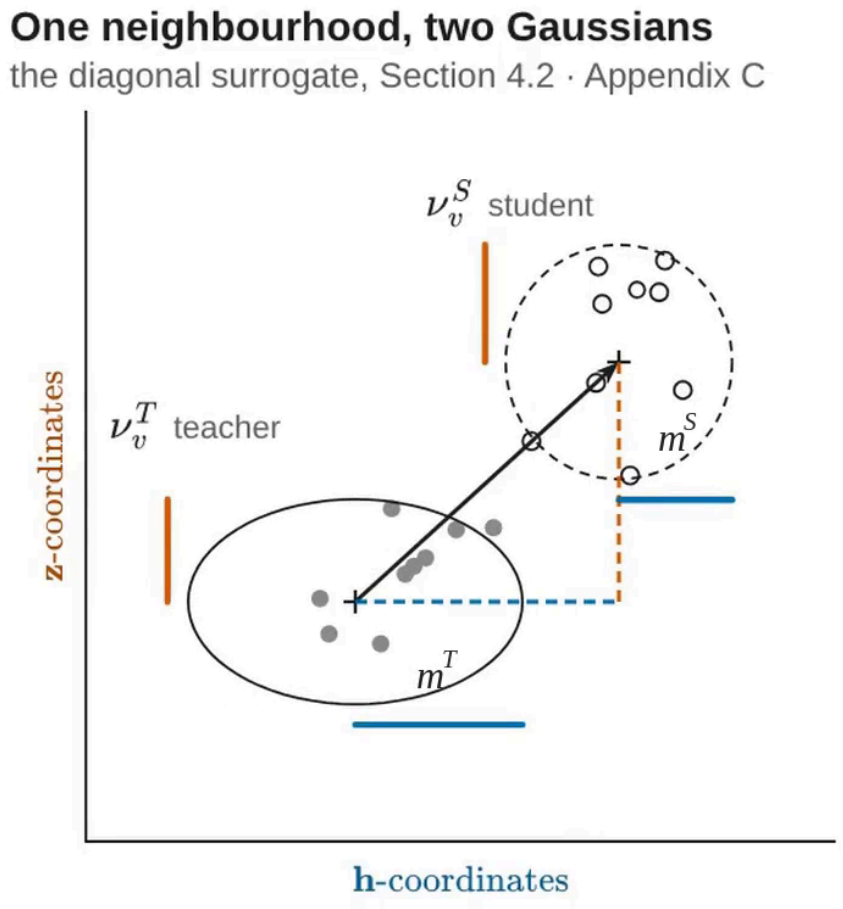
low-κ boundary nodes → logit side (α), high-κ interior nodes → feature side (β); the CAW alignment loss, Eq. (12)

full objective, Eq. (13): $\mathcal{L} = \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}} + \lambda_{\mathrm{CAW}}\frac{\mathcal{L}_{\mathrm{CAW}}}{\mathrm{sg}(\mathcal{L}_{\mathrm{CAW}})+\varepsilon_{\mathrm{sg}}}$

❹ Three relaxations make it tractable: local neighbourhoods, a diagonal-Gaussian $\mathcal{W}_2$ surrogate, and curvature-adaptive weights

Global → local · Transport → diagonal Gaussian · Uniform → curvature-aware

Local alignment, without solving OT



Approximate each neighborhood as a diagonal Gaussian

$$\mathcal{W}_{2,\mathrm{diag}}^2(\widetilde{\nu}_v^T, \widetilde{\nu}_v^S) = \|\mathbf{m}_v^T - \mathbf{m}_v^S\|_2^2 + \|\sqrt{\boldsymbol{\sigma}_v^T + \epsilon} - \sqrt{\boldsymbol{\sigma}_v^S + \epsilon}\|_2^2. \quad (10)$$

Match means and spreads; no discrete OT solver.

Curvature tells us where to focus

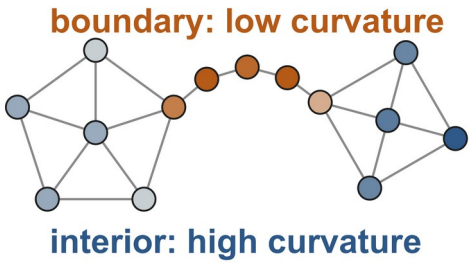
Two complementary signals; not one uniform distillation weight.

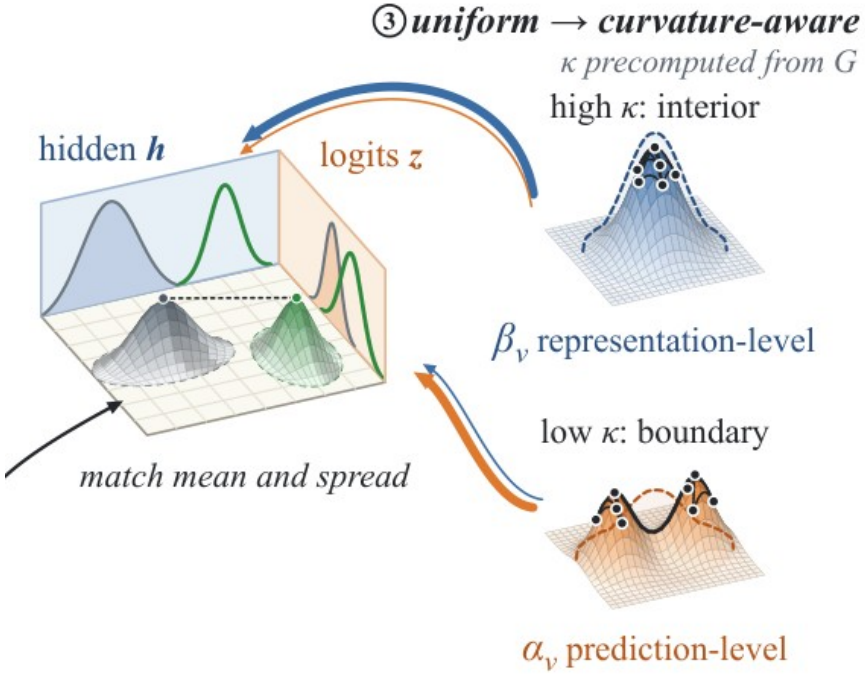
$$\alpha_v \propto \exp(-\gamma_z \tilde{\kappa}_v)$$

prediction-side weight

$$\beta_v \propto \exp(+\gamma_h \tilde{\kappa}_v)$$

representation-side weight

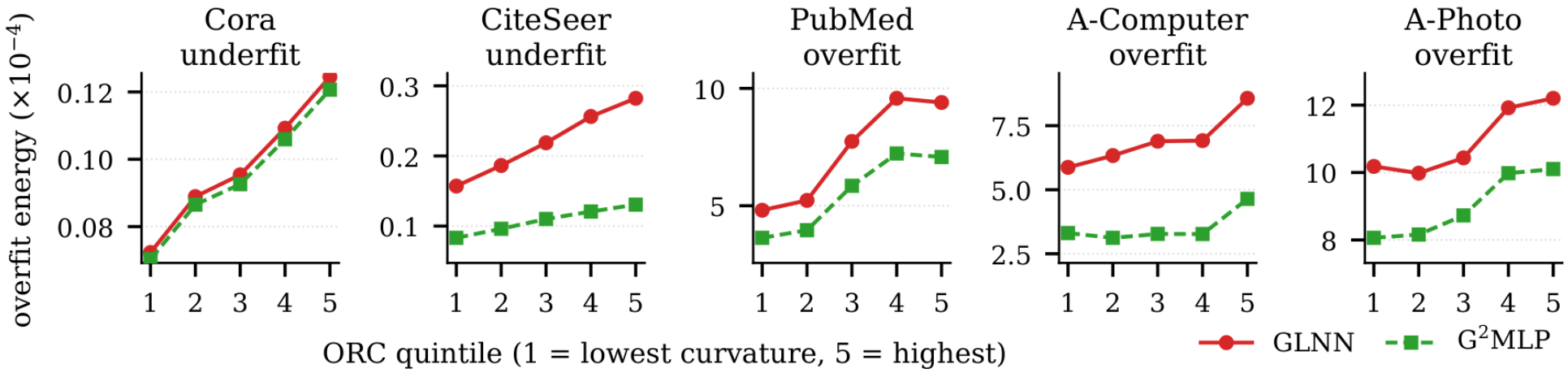




Low curvature → prediction alignment; high curvature → representation alignment.

Curvature tracks where overfit appears

Overfit energy across five datasets, grouped by Ollivier–Ricci curvature quintile.



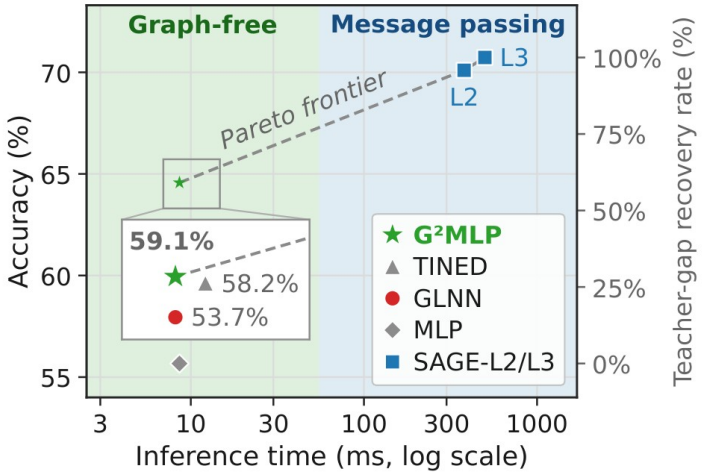
Vanilla distillation has higher overfit energy in the top curvature quintile on all five graphs.

Node classification and graph-free inference

Table 1 · Transductive

Table 1: Node classification accuracy (%), **transductive** setting. **Bold**: best graph-free student.

	Teacher	Graph-free students				
Dataset	SAGE	MLP	GLNN	KRD	FF-G2M	G^2MLP
Citeseer	70.49±1.53	58.50±1.86	71.22±1.50	72.84±1.70	72.85±1.59	**74.51±2.35**
Pubmed	75.56±2.06	68.39±3.09	75.59±2.46	77.01±3.11	76.56±3.41	**78.17±2.75**
Cora	80.64±1.57	59.18±1.60	80.26±1.66	82.27±1.31	82.38±1.41	**82.54±1.87**
A-computer	82.82±1.37	67.62±2.21	82.71±1.18	82.87±0.87	83.67±1.04	**83.87±1.09**
A-photo	90.85±0.87	77.29±1.79	91.95±1.04	91.95±1.44	92.51±0.56	**92.85±0.78**
Arxiv	70.73±0.35	55.67±0.24	63.75±0.48	59.26±0.51	58.51±0.35	**64.57±0.31**



Inference

The deployed model is still a feature-only MLP.

Table 2 · Production setting (prod / ind / tran)

Table 2: Node classification accuracy (%) under the **production** setting; conventions as in Table 1.

		Teacher	Graph-free students				
Dataset	Eval	SAGE	MLP	GLNN	KRD	FF-G2M	G^2MLP
Citeseer	prod	68.06	58.49	69.08	71.38	**71.89**	71.67
	ind	69.14±2.99	59.31±4.56	68.48±2.38	69.78±3.04	69.75±3.16	**71.85±2.27**
	tran	67.79±2.80	58.29±1.94	69.23±2.39	71.77±2.81	**72.12±2.69**	71.62±1.31
Pubmed	prod	74.77	68.39	74.67	76.00	73.98	**77.45**
	ind	75.07±2.89	68.28±3.25	74.52±2.95	75.17±3.11	73.49±7.91	**77.30±1.48**
	tran	74.70±2.33	68.42±3.06	74.70±2.75	76.20±3.00	74.10±7.78	**77.49±1.25**
Cora	prod	79.53	59.18	77.82	75.74	78.60	**79.73**
	ind	81.03±1.71	59.44±3.36	73.21±1.50	70.26±1.94	72.02±1.43	**75.90±1.87**
	tran	79.16±1.60	59.12±1.49	78.97±1.56	77.11±1.44	80.01±1.41	**80.69±0.73**
A-computer	prod	82.73	67.62	82.10	81.17	82.69	**84.29**
	ind	82.83±1.51	67.69±2.20	80.27±2.11	79.15±1.82	80.52±1.56	**82.67±0.93**
	tran	82.70±1.34	67.60±2.23	82.56±1.80	81.67±1.92	83.23±1.36	**84.69±0.85**
A-photo	prod	90.45	77.29	91.34	91.84	92.35	**92.51**
	ind	90.56±1.47	77.44±1.50	89.50±1.12	90.04±1.12	**90.70±0.76**	90.61±0.50
	tran	90.42±0.68	77.25±1.90	91.80±0.49	92.29±0.63	92.77±0.24	**92.98±0.51**
Arxiv	prod	70.69	55.35	63.50	59.32	59.60	**63.90**
	ind	70.69±0.58	55.29±0.63	59.04±0.46	57.32±0.31	57.02±0.43	**59.82±0.42**
	tran	70.69±0.39	55.36±0.34	64.61±0.15	59.82±0.27	60.24±0.23	**64.92±0.26**

Stronger in most production and inductive evaluations

Transfer to other teachers and tasks

Table 3 · Graph Transformer teachers (GT / GraphGPS / NAGphormer)

	GT			GraphGPS			NAGphormer		
Dataset	T	S	$S^\dagger$	T	S	$S^\dagger$	T	S	$S^\dagger$
Citeseer	71.16±2.31	71.85±2.45	**72.35±1.85**	68.07±2.62	69.38±2.30	**69.68±2.05**	70.09±1.81	71.24±1.64	**71.82±1.29**
Pubmed	74.09±2.48	74.61±2.48	**74.76±2.83**	71.77±2.81	71.68±2.61	**72.03±2.58**	76.19±2.86	76.85±3.34	**77.06±3.49**
Cora	78.53±2.08	78.98±1.90	**79.21±2.09**	76.21±1.91	76.06±1.99	**76.77±1.97**	80.88±1.77	81.00±1.45	**81.48±1.53**
A-computer	83.89±1.59	84.53±1.42	**84.72±1.32**	78.30±1.70	78.99±1.73	**80.12±1.55**	82.78±1.92	83.02±1.87	**83.31±1.77**
A-photo	90.94±0.87	91.90±0.56	**92.06±0.82**	90.41±1.10	91.03±0.95	**91.32±0.93**	91.45±0.82	92.07±0.74	**92.49±0.65**

Table 4 · Link prediction (AUC / AP)

	Cora		Citeseer		Pubmed	
Method	AUC	AP	AUC	AP	AUC	AP
SAGE (teacher, ref)	81.29±1.99	80.54±2.69	81.21±3.15	81.80±3.87	88.36±1.65	86.83±1.54
MLP	77.92±1.78	76.23±2.01	76.47±2.08	76.13±2.74	90.29±0.38	89.34±0.44
GLNN	78.61±1.68	77.05±1.76	83.32±2.34	83.71±2.53	90.03±0.38	89.12±0.32
KRD	79.20±1.65	77.89±2.31	83.31±2.23	83.81±2.42	90.32±0.42	89.38±0.51
G^{2}MLP (Ours)	**83.90±1.10**	**83.36±0.99**	**83.62±2.33**	**84.32±2.43**	**90.36±1.29**	**89.96±1.07**

The same recipe transfers to Graph Transformer teachers and to link prediction, with the same MLP student and inference cost.

Thank you.