Harvard John A. Paulson School of Engineering and Applied Sciences

Kempner INSTITUTE
For the Study of Natural & Artificial Intelligence at Harvard University

SiGMA: Learning under Graph-Level Heterogeneity with Gated Message-Passing and Attention

Raphael Pellegrin* Lukas Fesser* Melanie Weber

LoG CONFERENCE

Punchline

- Different graphs within the same dataset vary in prediction difficulty: this is **Graph-level heterogeneity**.
- Typical architectures still apply a **fixed computational template** (same depth, same Message Passing (MP) operator) to every graph.
- We introduce **SiGMA** (Sigmoid Gated Message-Passing and Attention): a gated hybrid architecture that mixes MP and attention heads in each layer.
- Gains of up to +2% accuracy (CIFAR10) and +0.025 F1 (COCO-SP) over strong GNN baselines.

Heterogeneity

Definition. We build **heterogeneity profiles** from repeated held-out predictions: repeatedly sample 50/25/25 splits, train, pick the best validation checkpoint, and score every test graph, until each graph has ≥ 100 test appearances. A graph's score is its average accuracy. These profiles reveal that difficulty is not uniform across a dataset: some graphs are reliably easy while others remain consistently hard.

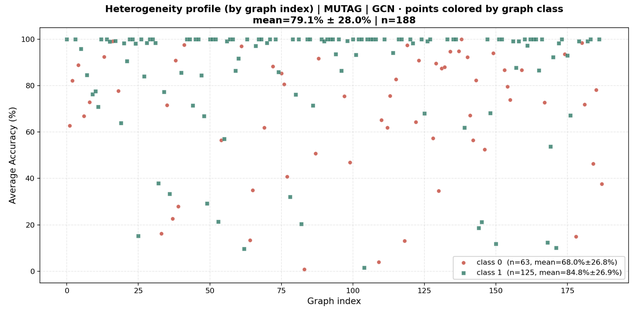


Figure 1: **Heterogeneity profiles on MUTAG for GCN.**

Difficulty is model-dependent: graphs hard for one MP mechanism can be easy for another, and different graphs prefer different depths / local–nonlocal mixes. This motivates architectures that can adapt their computational template per graph rather than committing to a single fixed operator and depth.

Operator & depth preferences

Table 1: % of graphs preferring each MP.

MP	MUTAG	ENZYMES	PROT.	IMDB	COLLAB	REDDIT
GCN	29.2	30.1	**51.0**	25.9	31.2	30.0
GIN	**40.3**	31.2	27.0	**62.6**	**49.0**	**65.0**
GatedGCN	30.5	**38.7**	22.1	11.5	19.8	5.0

Table 2: % of graphs preferring each GCN depth.

Layers	MUTAG	ENZYMES	PROT.	IMDB	COLLAB	REDDIT
2	**37.5**	23.6	**42.1**	**44.6**	**37.6**	16.6
4	25.5	**48.1**	26.2	28.1	29.4	37.2
8	37.0	28.2	31.7	27.3	32.9	**46.1**

No single MP operator or depth is uniformly preferred.

Architecture

Design principle. Any mix of MP and attention heads per layer; each head is sigmoid-gated.

Let $\mathbf{H}^\ell = [\mathbf{h}_1^\ell, \dots, \mathbf{h}_n^\ell]^\top$ be the node features at layer ℓ, and write $\widehat{\mathbf{h}}_i^\ell = \mathrm{LayerNorm}(\mathbf{h}_i^\ell)$. Then the attention head a forms

$$\left[\mathbf{q}_i^a \,\|\, g_i^{\mathrm{att},a}\right] = \mathbf{W}_{qg}^a \widehat{\mathbf{h}}_i^\ell, \quad \mathbf{k}_j^a = \mathbf{W}_k^a \widehat{\mathbf{h}}_j^\ell, \quad \mathbf{v}_j^a = \mathbf{W}_v^a \widehat{\mathbf{h}}_j^\ell,$$

$$\tilde{\mathbf{a}}_i^{\mathrm{att},a} = \sigma(g_i^{\mathrm{att},a}) \sum_{j \in V} \alpha_{ij}^a \mathbf{v}_j^a, \quad \alpha_{ij}^a = \mathrm{softmax}_{j \in V}\big((\mathbf{q}_i^a)^\top \mathbf{k}_j^a / \sqrt{d_h}\big),$$

and the MP head p forms

$$\left[\mathbf{u}_i^p \,\|\, g_i^{\mathrm{mp},p}\right] = \mathbf{W}_{hg}^p \widehat{\mathbf{h}}_i^\ell, \quad \tilde{\mathbf{a}}_i^{\mathrm{mp},p} = \sigma(g_i^{\mathrm{mp},p})\, \Phi_p\big(\{\mathbf{u}_j^p\}_{j \in V}, E\big)_i,$$

and the final update is

$$\mathbf{h}_i^{\ell+1} = \mathbf{h}_i^\ell + \mathbf{W}_{\mathrm{out}}^\ell \left[\Big\|_a \tilde{\mathbf{a}}_i^{\mathrm{att},a} \,\Big\|\, \Big\|_p \tilde{\mathbf{a}}_i^{\mathrm{mp},p} \right].$$

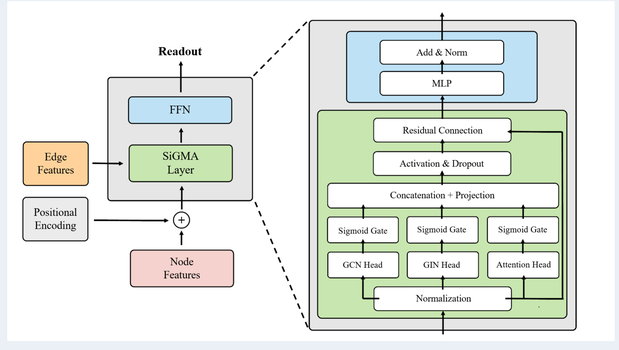


Figure 2: **SiGMA layer.** Gated MP/attention heads are concatenated, projected, and residual-updated.

Theory

Proposition (Expressivity). SiGMA is at least as expressive as the most expressive MP architecture it contains, under mild conditions.

Define the Dirichlet energy $\mathcal{E}_G(\mathbf{H}) = \sum_{(i,j) \in E} \|\mathbf{h}_i - \mathbf{h}_j\|_2^2$.

Proposition (Oversmoothing). $\mathcal{E}_G(\mathbf{H}^L)$ can be made arbitrarily close to $\mathcal{E}_G(\mathbf{H}^0)$, so increasing nominal depth does not by itself force Dirichlet-energy collapse / oversmoothing (gated residual updates).

Results

Table 3: Test performance (mean±s.d.). Acc. in %; ZINC is MAE; COCO-SP is F1. Highlighted: **1st**, **2nd**, **3rd**.

	ZINC↓	MNIST↑	CIFAR10↑	PATTERN↑	COCO-SP↑
GraphGPS	0.070±.004	98.05±.13	72.30±.36	86.69±.06	0.388±.006
GRIT	0.059±.002	98.11±.11	76.47±.88	**87.20**±.08	–
GRED	0.077±.002	98.38±.01	**76.86**±.19	86.76±.02	–
GEAET	–	**98.51**±.09	76.63±.43	86.99±.03	0.322±.005
TIGT	**0.057**±.002	98.23±.13	73.96±.36	86.68±.06	–
GMN	–	98.39±.18	74.56±.38	**87.09**±1.3	–
GIN$^+$	0.065±.004	98.29±.10	69.59±.29	86.84±.05	0.248±.005
GatedGCN$^+$	0.077±.005	**98.71**±.14	**77.22**±.38	87.03±.04	**0.380**±.002
SiGMA	**0.060**±.003	**98.63**±.11	**79.53**±.18	**87.40**±.19	**0.422**±.006

Takeaways

- Keep a strong GNN baseline; add multiple MP copies and attention heads.
- A simple, effective recipe for flexible GNN architectures.

SiGMA opens a new design space: mixing multiple (possibly different) gated MP heads with attention leads to flexible GNN architectures.

Code and paper

https://arxiv.org/abs/2503.00547

https://github.com/Weber-GeoML/SiGMA