
Why don't Graph Transformers attend to the same token?

Alexandre Bismuth
BOLD Lab
University of Oxford

Abstract

In large language models (LLMs), attention concentrates heavily on the first token. This 'attention sink' is believed to protect deep models against over-mixing. Even though graph transformers face a related pressure, we find that a SoTA graph transformer, GRIT, rarely forms sinks on molecules, and that it sinks less when trained on larger graphs, the opposite of the trend in LLMs. In principle, a graph can prevent sinks: we prove and show empirically that its symmetry and locality limit the attention any node can receive. On molecules, however, with distinguishable atoms and global attention, neither limit prevents sinks; what stops them lies in the model. We find that the attention scoring function is the main cause. Replacing GRIT's gated logit with a dot product raises its sink rate from 0.09 to 0.68 on identical graphs. In Graph Transformers using dot-product attention, removing the sink gives about the same error as removing a random node, while GRIT's most-attended nodes matter to its predictions. Attention sinks are therefore a property of dot-product attention rather than of Transformers. In our experiments, they form when dot-product scoring has a distinguishable node to concentrate on.¹

1 Introduction

A striking empirical property of Large Language Models (LLMs) is the *attention sink* (Xiao et al., 2024). In Transformers, an attention sink is a token that receives a large share of all attention: in Llama 3.1 405B, nearly 80% of attention heads sink on the semantically meaningless <BOS> token (sink metric of Gu et al., 2025 at threshold $\varepsilon = 0.8$; Barbero et al., 2025). Far from wasting compute, sinks are thought to slow the mixing of information between tokens and so protect deep models against 'over-mixing' (Dong et al., 2021; Barbero et al., 2024; Di Giovanni et al., 2023; Barbero et al., 2025), and larger models and longer training contexts both bring stronger sinks (Barbero et al., 2025).

Graph Transformers face a related pressure: depth causes over-smoothing (Zhang et al., 2024), and global attention makes it worse, since softmax over ever more nodes must disperse toward uniform mixing (Veličković et al., 2025). A designated sink node might be a natural defence, yet nothing in an ordinary graph plays the role of <BOS>: a graph has no first token, and permutation equivariance forbids preferring one of two nodes the model cannot tell apart. Sinks are documented beyond causally masked LLMs, in bidirectional diffusion language models (Rulli et al., 2025), as massive activations in attentional GNNs (Bini et al., 2024), and when LLMs read serialised graphs (Guan et al., 2025); concurrent work argues that graphs lack natural sinks (Liu et al., 2026). What is missing is an account of *where* a sink can form on a graph at all, and of *which* architectural ingredient decides whether one does.

We ask whether Graph Transformers form attention sinks and, if not, what prevents them. We study GRIT (Ma et al., 2023), a state-of-the-art Graph Transformer without message passing, so attention

¹The code can be found at <https://github.com/alexandre-bismuth/graph-transformer-sinks>.

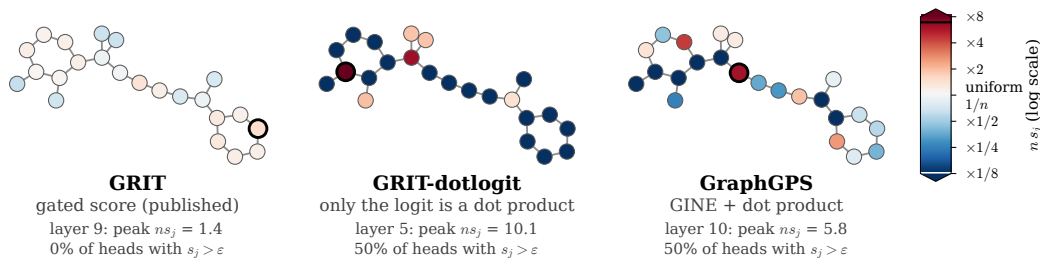


Figure 1: Changing only the scoring function creates a sink. Each atom of a 23-atom ZINC test molecule is shaded by the attention it receives, averaged over queries and heads at each model’s peak-sinking layer: blue below the uniform share $1/23$, red above. Swapping only GRIT’s gated logit for a dot product (middle) moves attention off the other atoms and onto one, as in GraphGPS (right).

alone mixes information between nodes. We find that GRIT rarely forms sinks on molecules, and sinks even less when trained on larger graphs (Section 2), but that swapping in dot-product attention makes it sink far more often (Section 4). We investigate two suspects. The graph itself is not the barrier: we prove three bounds on received attention, from symmetry, locality, and the spread of attention scores, and find both graph-side bounds loose on real molecules (Section 3). The attention scoring function is the main cause, as the logit swap of Figure 1 and accuracy-matched controls show (Section 4).

2 GRIT rarely forms attention sinks on molecules

Graph structure enters GRIT (Ma et al., 2023) through Relative Random Walk Probabilities (RRWP), which initialise a learned pair representation. GRIT scores a pair by gating the sum of query and key with that representation, passing it through a signed square root, and clipping it to $[0, 5]$ before a scalar readout; dot-product attention instead takes the inner product of query and key, which can grow with the product of the two node norms:

$$z_{uj}^{\text{GRIT}} = a^\top \min\{5, \text{ReLU}(\text{ssq}((q_u + k_j) \odot w_{uj}) + b_{uj})\}, \quad z_{uj}^{\text{dot}} = q_u^\top k_j / \sqrt{h}, \quad (1)$$

with $\text{ssq}(x) = \text{sgn}(x)\sqrt{|x|}$, w_{uj}, b_{uj} linear in the pair representation, and h the head dimension. Clipping the scalar logit instead, as the official code does, gives the same sink rate (Appendix C). We compare against GraphGPS (Rampásek et al., 2022) with RWSE (Dwivedi et al., 2022a), a hybrid of message passing and dot-product attention. We train on ZINC-12k, a regression task on graphs of roughly 23 nodes, and on Peptides-func (Dwivedi et al., 2022b), a classification task on larger graphs, both whole and split by node count into four quartiles (mean sizes 56 to 245), each trained on 2,023 graphs at batch size 16 to remove the dataset-size confound. Our results are means with a 95% CI on the full test set over three seeds (five for GraphGPS and the quartiles).

Our central measurement is the **sink score** of a node, the attention it receives averaged over all queries (Barbero et al., 2025; Gu et al., 2025). Following Gu et al. (2025), a node is a sink in a head when its sink score exceeds $\varepsilon = 0.3$, about seven times the uniform share on a 23-node graph. The **sink rate** is the largest fraction of heads that share one sink node, averaged over test graphs, at the peak-sinking layer.

On ZINC, GRIT’s sink rate is 0.090 ± 0.028 : fewer than one head in ten sinks on the same node, against 0.415 ± 0.194 for GraphGPS and 0.95–1.00 for small language models at the same context length. GRIT stays below GraphGPS at every threshold from 0.1 to 0.8 (Appendix C). The trend with size also runs backwards. Language models trained on longer contexts form stronger sinks (Barbero et al., 2025), whereas GRIT’s sink rate falls from 0.238 on the smallest Peptides-func quartile to 0.045 on the largest. As in language models, where sink tokens develop unusually large norms (Queipo-de-Llano et al., 2026), GRIT’s largest squared node norm grows across layers from 2.4 to 4.8 times the squared mean norm. Nonetheless, sinks stay rare, so something in GRIT keeps large norms from attracting attention (Section 4).

3 What can prevent a sink?

Three things can prevent a node from accumulating attention from many others at once: symmetry, since an equivariant model must treat nodes the labelled graph does not distinguish alike; locality, since nodes that attend mostly to their neighbourhood leave little attention for distant ones; and scoring range, since softmax concentrates only when logits spread far apart.

Notation. We write $A \in \mathbb{R}^{n \times n}$ for the attention matrix of a single layer and head, with logits z_{uv} , and define the sink score as $s_j(A) = \frac{1}{n} \sum_u A_{uj}$. We assume permutation equivariance, i.e. $A(PZ) = PA(Z)P^\top$ for every permutation matrix P and input Z containing the graph, node features and positional encodings. Let $\mathcal{O}(j)$ be the orbit of j under the automorphisms of Z (the nodes equivariant models must treat like j), $B_G(j, r)$ the ball of radius r around j in shortest-path distance d_G , $\lambda_r(A) = \max_u \sum_{v: d_G(u,v) > r} A_{uv}$ the largest attention any query sends beyond distance r , and $\Delta = \max_u (\max_v z_{uv} - \min_v z_{uv})$ the largest logit spread of any query.

Proposition 3.1 (Three bounds on received attention) *Under the setup above, for every node j and every radius r ,*

$$s_j(A) \leq \min \left\{ \frac{1}{|\mathcal{O}(j)|}, \lambda_r(A) + (1 - \lambda_r(A)) \frac{|B_G(j, r)|}{n}, \frac{e^\Delta}{e^\Delta + n - 1} \right\}.$$

In particular, on a vertex-transitive labelled input $s_j(A) = 1/n$ for every node, independently of the learned parameters, and reaching received attention ε requires $\Delta \geq \log((n-1)\varepsilon/(1-\varepsilon))$.

The orbit term is the standard fact that equivariant maps are constant on orbits (Morris et al., 2024). What we add is that every bound can be computed head by head, so we can test whether any bound is tight enough to prevent the formation of a sink (Appendix B). The symmetry bound is tight on symmetric graphs: every checkpoint with an equivariant positional encoding, even the dot-product models that sink heavily on molecules, spreads attention evenly across cycles of identical atoms and bonds (C_8 to C_{24} ; Appendix B). Symmetry alone prevents the sink, whatever the architecture. Conversely, Laplacian encodings, whose eigenvector signs and bases are arbitrary (Lim et al., 2023), break equivariance, and a dot-product Transformer then puts almost all attention (0.999) on one node.

On real molecules, neither graph-side bound is tight: atoms rarely have symmetric counterparts, attention is nonlocal (median $\lambda_1 \geq 0.91$), and the bounds rule out a sink in at most 5.1% of heads in any run. The third bound, on the spread of attention scores, does differ between architectures: GRIT clips its scores elementwise, so their spread never exceeds $5\|a\|_1$, whereas dot-product scores can grow without limit with the attended node’s norm (Proposition A.1). Yet this cap does not explain why GRIT rarely sinks: in our trained models, $5\|a\|_1$ ranges from 2.7 to 11.7 across layers, above the spread of 2.2 a sink needs at ZINC size, and removing the clip leaves GRIT’s sink rate low (Section 4). The graph thus leaves GRIT and GraphGPS equally free to sink on the same molecules, yet their sink rates differ more than four-fold: the models themselves must decide.

4 The attention scoring function drives the gap

On identical ZINC graphs, GraphGPS and GRIT differ in heads, positional encoding, dropout, and message passing as well as in scoring, so we change one component at a time (Table 1). Replacing only GRIT’s logit with z^{dot} (labelled GRIT-dotlogit), leaving RRWP, the pair representation and the values unchanged, raises its sink rate from 0.090 ± 0.028 to 0.680 ± 0.098 (Figure 1), above GraphGPS’s. Removing message passing from GraphGPS does not lower its sink rate either (0.507 ± 0.159). Message passing is therefore not what suppresses sinks; the scoring function is.

The swap also slows training (test MAE 0.064 to 0.373), but two controls at matched accuracy show that this is not why it sinks. Without any graph structure (RRWP and bond types), GRIT’s score and a dot product reach the same accuracy (test MAE 0.542 and 0.543, labelled GRIT-blind and GRIT-dot) yet sink at rates of 0.214 and 0.532. Likewise, GraphGPS with a virtual node matches GRIT’s accuracy (0.066) yet sinks six times as often (0.559). Adding node RRWP and a bounded pair bias to GRIT-dot (labelled GRIT-dotbias) recovers over half the lost accuracy (test MAE 0.268) but leaves its sink rate unchanged (0.521). Sink formation tracks mainly the form of the score; accuracy tracks access to graph structure. Even a loss that rewards concentrated attention lifts GRIT’s sink rate only to 0.131.

Table 1: Changing each component individually on ZINC (mean $\pm$ 95% CI; all arms in Table 3). *Ablation ratio*: test MAE after removing all attention onto the most-attended node, divided by that for a random node.

Arm	Sink rate	Test MAE	Ablation ratio
GRIT (gated scoring, RRWP)	0.090 ± 0.028	0.064 ± 0.009	2.06 ± 0.79
GRIT-dotlogit (dot-product logit only)	0.680 ± 0.098	0.373 ± 0.224	1.22 ± 0.43
GRIT-blind (gated scoring, no structure)	0.214 ± 0.143	0.542 ± 0.018	1.17 ± 0.07
GRIT-dot (dot-product scoring, no structure)	0.532 ± 0.061	0.543 ± 0.006	1.15 ± 0.12
GRIT-dotbias (GRIT-dot + pair bias)	0.521 ± 0.059	0.268 ± 0.068	1.19 ± 0.19
GraphGPS (GINE + dot attention)	0.415 ± 0.194	0.072 ± 0.005	1.01 ± 0.02
GraphGPS-none (MPNN removed)	0.507 ± 0.159	0.200 ± 0.017	1.11 ± 0.09
GraphGPS + virtual node	0.559 ± 0.177	0.066 ± 0.015	1.00 ± 0.02
GRIT + virtual node	0.095 ± 0.019	0.065 ± 0.008	1.93 ± 1.19

Why large norms do not attract attention in GRIT. A dot-product logit grows linearly with the attended node’s hidden state, so a node with a large enough norm can claim almost all of a query’s attention (Proposition A.1). GRIT’s score instead saturates: in layers 2 to 10, at least 96% of its coordinates sit at 0 or 5, so the logit is a sum of readout weights selected by which coordinates fire, not by input magnitude. Scaling up one node’s hidden state confirms this (Table 4): doubling it makes it a sink in up to 38.8% of heads in GRIT-dotlogit but at most 0.2% in GRIT, and even a 32-fold increase reaches only 1.3%. Removing the clip (GRIT-noclip) changes little: the model still sinks rarely (0.113), and a doubled norm makes a sink in at most 1.4% of heads, but an extreme $32\times$ norm increase does in up to 54.0%. GRIT thus grows high-norm nodes without attending to them.

Giving the model a token to sink on. A virtual node is the graph counterpart of <BOS>: chemically empty, connected to every node, akin to vision-transformer registers (Darcet et al., 2024), and exempt from the symmetry bound, since no automorphism can move it. On the symmetric cycles of Section 3, where no real node may stand out, GraphGPS sends all of its attention to the virtual node, GRIT only about half to two-thirds on average, and inconsistently across seeds. On molecules, a virtual node leaves GRIT’s sink rate low (0.095 ± 0.019), even with a logit penalty toward it (0.079 ± 0.027). GraphGPS with a virtual node usually sinks onto an atom rather than onto the virtual node, and the sink’s value vector is about a tenth of the mean norm. Symmetry says where a sink may appear; the scoring function decides whether one does.

The ablation column of Table 1 shows what these sinks are worth: removing all attention to the most-attended node leaves the dot-product models’ test MAE close to that for a random node (1.00 to 1.22 times), but GRIT’s twice as high (2.06). In dot-product models the sink is thus close to a no-op, as in LLMs (Barbero et al., 2025), whereas GRIT’s most-attended nodes matter to its predictions.

5 Discussion

Graph Transformers based on GRIT largely do not form attention sinks on molecules, and the graph is not the reason: symmetry forbids sinks only on inputs far more regular than a molecule, and the attention we measure is too nonlocal for locality to matter. What matters most is the scoring function: GRIT’s logit barely responds to a node’s norm, and on molecules even a virtual node does not make it sink. The LLM sink therefore looks like a property of dot-product attention rather than of Transformers, consistent with sinks vanishing without softmax normalisation (Gu et al., 2025) or under an output gate (Qiu et al., 2025). Read as attentional networks over a complete graph (Bronstein et al., 2021), Transformers differ from GRIT in two ways, a first token and a dot-product score, and our experiments separate the two: on molecules, dot-product scoring sinks with or without a designated token, while GRIT’s gated scoring rarely sinks in either case. Whether sinks protect Graph Transformers against over-mixing remains open: our 10-layer models show little over-smoothing to prevent (Appendix C).

Our study has limits: one gated-score architecture against dot-product baselines, on molecular graphs alone, at 10 layers; several GRIT virtual-node quantities remain seed-dependent; and we measure reception rather than mixing, since a perfect sink can also be a complete mixer. Further architectures (EGT, Graphormer, SAN; Yuan et al., 2026) and non-molecular graphs come next.

References

- Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G. M. Araújo, Alex Vitvitskiy, Razvan Pascanu, and Petar Veličković. Transformers need glasses! Information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37:98111–98142, 2024.
- Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael Bronstein. Why do LLMs attend to the first token? In *Conference on Language Modeling*, 2025.
- Umberto Biccari, Qian Huang, and Enrique Zuazua. Interior interpretability with attention rollout: Contraction and propagation profiles in transformers. *arXiv preprint arXiv:2607.22367*, 2026.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.
- Lorenzo Bini, Marco Sorbi, and Stéphane Marchand-Maillet. Massive activations in graph neural networks: Decoding attention for domain-dependent interpretability. *arXiv preprint arXiv:2409.03463*, 2024.
- Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.
- Francesco Di Giovanni, James Rowbottom, Benjamin P Chamberlain, Thomas Markovich, and Michael M Bronstein. Understanding convolution on graphs via energies. *Transactions on Machine Learning Research*, 2023.
- Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.
- 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*, 2022a.
- Vijay Prakash Dwivedi, Ladislav Rampásek, Michael Galkin, Ali Parviz, Guy Wolf, Anh Tuan Luu, and Dominique Beaini. Long range graph benchmark. *Advances in Neural Information Processing Systems*, 35:22326–22340, 2022b.
- Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, 2025.
- Zhong Guan, Likang Wu, Hongke Zhao, Ming He, and Jianping Fan. Attention mechanisms perspective: Exploring LLM processing of graph-structured data. In *International Conference on Machine Learning*, pages 20612–20639. PMLR, 2025.
- Derek Lim, Joshua Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. In *International Conference on Learning Representations*, 2023.
- Yang Liu, Dongxin Guo, Tom Zheng, Siu Ming Yiu, Liam Ning, and Jikun Wu. Capacity-controlled global attention for graph transformers. *arXiv preprint arXiv:2604.17324*, 2026.
- Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K Dokania, Mark Coates, Philip Torr, and Ser-Nam Lim. Graph inductive biases in transformers without message passing. In *International Conference on Machine Learning*, pages 23321–23337. PMLR, 2023.
- Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. Isambard-AI: a leadership class supercomputer optimised specifically for artificial intelligence. *arXiv preprint arXiv:2410.11199*, 2024.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.
- Matthew Morris, Bernardo Cuenca Grau, and Ian Horrocks. Orbit-equivariant graph neural networks. In *International Conference on Learning Representations*, 2024.
- Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025.
- Enrique Queipo-de-Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *International Conference on Learning Representations*, 2026.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.
- Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

- Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.
- Petar Velićković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *International Conference on Machine Learning*, pages 61190–61211. PMLR, 2025.
- Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In *Advances in Neural Information Processing Systems*, 2023.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.
- Chaohao Yuan, Kangfei Zhao, Ercan Engin Kuruoglu, Liang Wang, Tingyang Xu, Wenbing Huang, Deli Zhao, Hong Cheng, and Yu Rong. A survey of graph transformers: Architectures, theories and applications. *ACM Computing Surveys*, 2026. doi: 10.1145/3834764.
- Xu Zhang, Yonghui Xu, Wei He, Wei Guo, and Lizhen Cui. A comprehensive review of the oversmoothing in graph neural networks. In *Computer Supported Cooperative Work and Social Computing*, pages 451–465. Springer, 2024.

Acknowledgments

I would like to thank Professor Michael Bronstein for his class on Geometric Deep Learning at the University of Oxford, where this research on Graph Transformers started as a course project.

The author acknowledges the use of resources provided by the Isambard-AI National AI Research Resource (AIRR) (McIntosh-Smith et al., 2024). Isambard-AI is operated by the University of Bristol and is funded by the UK Government’s Department for Business, Innovation, Science and Trade (DBIST) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

A Proofs and supporting theory

A.1 Proof of Proposition 3.1

Proof. *Orbit term.* For an automorphism P of the labelled input, $PZ = Z$ and equivariance give $A(Z) = PA(Z)P^\top$, so with $s(A) = \frac{1}{n}A^\top \mathbf{1}$ and $P^\top \mathbf{1} = \mathbf{1}$ we get $Ps(A) = s(A)$: the sink score is constant on every orbit. Row-stochasticity gives $\sum_j s_j(A) = 1$, so an orbit of size m has $s_j(A) \leq 1/m$, and a vertex-transitive input is one orbit of size n , forcing $s_j(A) = 1/n$.

Locality term. For $u \in B_G(j, r)$, row-stochasticity gives $A_{uj} \leq 1$; for $u \notin B_G(j, r)$, $d_G(u, j) > r$ gives $A_{uj} \leq \sum_{v: d_G(u, v) > r} A_{uv} = \ell_u^{(r)} \leq \lambda_r(A)$. Therefore

$$s_j(A) \leq (|B_G(j, r)| + \sum_{u \notin B_G(j, r)} \ell_u^{(r)})/n \leq \lambda_r(A) + (1 - \lambda_r(A)) |B_G(j, r)|/n,$$

where the first inequality is the per-node form evaluated in the certificate of Appendix B.

Range term. Fix a query u . For $v \neq j$ the definition of Δ gives $z_{uv} \geq z_{uj} - \Delta$, so $A_{uj} = e^{z_{uj}} / \sum_v e^{z_{uv}} \leq e^{z_{uj}} / (e^{z_{uj}} + (n-1)e^{z_{uj}-\Delta}) = e^\Delta / (e^\Delta + n-1)$. Averaging over u gives the same bound for $s_j(A)$, and $e^\Delta / (e^\Delta + n-1) \geq \varepsilon$ gives $\Delta \geq \log((n-1)\varepsilon / (1-\varepsilon))$. Taking the minimum of the three bounds proves the proposition. $\square$

A.2 Score spread

Proposition A.1 (Bounded and unbounded score spread) *For the GRIT score (1), write $\hat{e}_{uj} \in [0, 5]^h$ for the clipped vector that a reads out; hence $5 \sum_{i: a_i < 0} a_i \leq z_{uj} \leq 5 \sum_{i: a_i > 0} a_i$ and $\Delta \leq 5 \|a\|_1$ whatever the node features, with the constant sharp over the box $[0, 5]^h$. For dot-product scoring no such bound exists: if x_j is the input state of node j and $\langle q_u, W_K x_j \rangle \neq 0$ for some $u \neq j$, replacing x_j by ρx_j makes Δ grow without bound in ρ , and a bounded edge-bias term cannot restore a bound.*

Proof. The clip and rectifier confine each coordinate to $[0, 5]$, and a linear functional on that box is extremised at the two sign-selected corners, whose difference is $5 \sum_i |a_i|$. For the dot product and

$u \neq j$, only $k_j(\rho) = \rho W_K x_j + b_K$ moves, so z_{uj} is affine in ρ with nonzero slope while the rest of row u is fixed, and Δ is at least the gap between two entries of that row. $\square$

At $\varepsilon = 0.3$ a sink needs a spread of 2.24 at $n = 23$ and 4.74 at $n = 269$, against $5\|a\|_1 = 2.74$ to 11.69 across the ZINC layers, so the bound is loose (Section 3).

Reception is not mixing. A sink score does not determine how much a layer mixes, the distinction that spectral and Dobrushin-type analyses of over-smoothing and attention propagation turn on (Wu et al., 2023; Biccari et al., 2026). With $\kappa(A) = \frac{1}{2} \max_{u,t} \|A_{u:} - A_{t:}\|_1$ the Dobrushin coefficient, $\text{diam}(AY) \leq \kappa(A) \text{diam}(Y)$ for any features Y . The identity I_n and the uniform matrix $U_n = \frac{1}{n} \mathbf{1}\mathbf{1}^\top$ share the flat profile $s = \mathbf{1}/n$ yet have $\kappa(I_n) = 1$ and $\kappa(U_n) = 0$, while a perfect sink $\mathbf{1}e_v^\top$ has the maximal profile $s = e_v$ yet $\kappa = 0$. We therefore measure mixing separately (Appendix C).

B Experimental details

Table 2: Experimental hyperparameters. Peptides-func graphs are capped at 300 nodes to limit memory usage, and each quartile model is trained on 2,023 graphs of its quartile at batch size 16. All models are trained with AdamW, linear warmup, cosine annealing, and gradient clipping on a single GPU. Every other arm inherits its base model’s configuration, changing only the named component.

Model	Dataset	Dim	Heads	Layers	MPNN	PE	LR	Epochs	Batch
GRIT	Peptides	96	4	4	–	RRWP-17	3e-4	150	8
GRIT	ZINC	64	8	10	–	RRWP-21	1e-3	2000	32
GraphGPS	ZINC	64	4	10	GINE	RWSE-28	1e-3	1500	32

Arms. GRIT-dotlogit replaces z^{GRIT} by z^{dot} in Eq. (1) and keeps everything else, including RRWP, the pair representation and the values; GRIT-noclamp removes the clip. Like the official code, GRIT-official clips the scalar logit to $[-5, 5]$ rather than each coordinate; unlike that code, all our GRIT arms use one readout a for every head. GRIT-blind removes node and pair RRWP and bond types; GRIT-dot is GRIT-blind with z^{dot} and no pair representation; GRIT-dotbias adds to GRIT-dot node RRWP and a bounded pair bias $a^\top \text{clip}(We_{uj})$ in the logit. GRIT-entropy-1/2 add an attention-entropy penalty of two strengths. Virtual-node arms join one node to every atom, with the positional encoding padded or recomputed (PE rec.); GRIT + virtual node + bias adds a learned logit penalty toward it. GraphGPS-none removes the MPNN, GraphGPS-GatedGCN replaces it, and Transformer + LapPE is GraphGPS-none with sign-ambiguous Laplacian encodings (8 eigenvectors). Seeds are fixed per run, we report seed-controlled runs only, and checkpoints are best-validation.

Interventions. *Sink ablation:* on 500 test graphs, the most-attended node (head-averaged received attention summed over layers) has its attention column zeroed in every layer, removing that mass (Table 1) or renormalising the rows (Table 3); the control ablates a uniformly random node. *Norm intervention:* at layer $\ell \in \{3, 5, 8\}$ the input state of one uniformly chosen real node is multiplied by ρ before that layer’s attention, on 200 test graphs, and we read that layer’s attention only. *Language models:* Pythia-70M/160M (Biderman et al., 2023) and GPT-2 small (Radford et al., 2019) on 200 non-overlapping WikiText-2 test windows (Merity et al., 2017) of T tokens without a BOS token, with the same sink-rate estimator.

Certificate and probe. For Proposition 3.1 we evaluate, on 300 test graphs per run, layer, and head, the row leakages $\ell_u^{(r)}$ and $\lambda_r(A)$ for $r \leq 3$, the ball volumes, the orbits via VF2 with capped enumeration (a cap only shrinks orbits, so $1/|\mathcal{O}(j)|$ stays valid), and the per-node bound $U_j = \min\{1/|\mathcal{O}(j)|, \min_r(|B_G(j, r)| + \sum_{u \notin B_G(j, r)} \ell_u^{(r)})/n\}$. The locality bound was never violated, and within-orbit deviation was at most 8.7×10^{-5} for every equivariant-PE run except GRIT-dotbias (0.016, with 2 orbit-bound violations in total), which doubles as an equivariance test. The probe passes cycles C_8 – C_{24} of identical atoms and bonds through the checkpoints; virtual-node arms use the training-time transform, so the virtual node is a singleton orbit beside the n cycle nodes and the two are reported separately. On the cycle probe, received attention is uniform to 7.8×10^{-7} in every equivariant-PE checkpoint except GRIT-dotbias and GRIT-dotlogit, which amplify float32 rounding (deviation up to 0.96) and are uniform to 9.6×10^{-4} in float64. Real nodes of virtual-node arms stay uniform to 2.5×10^{-4} .

C Additional results

Training verification. GRIT reaches test MAE 0.064 on ZINC, close to the published result (Ma et al., 2023), and test AP 0.634 on the full Peptides-func dataset (0.612 on the smallest quartile, 0.499 on the largest), below the published result; GraphGPS reaches 0.072 on ZINC. At GRIT’s 2,000-epoch budget, GRIT-dotlogit appears under-trained, with an unconverged training loss.

Sink ablations. Table 1 removes the attention onto the most-attended node outright. If each row is instead renormalised, so that this attention is spread over the other nodes, GRIT’s ablation ratio falls from 2.06 to 1.35, but every gated arm (1.25 to 1.61) stays above every dot-product arm (1.00 to 1.23) except GraphGPS-none, whose ratio varies widely across seeds (2.03; Table 3).

Thresholds and size. GRIT sinks less than each of the seven dot-product arms we swept at every threshold from $\varepsilon = 0.1$ to 0.8 (closest at $\varepsilon = 0.8$: 0.02 vs. 0.05), and the threshold-free peak share $n \max_j s_j$ separates gated arms (2.5 to 3.6) from dot-product arms (6.0 to 13.1) with no overlap (Table 3). On Peptides-func, GRIT’s largest sink score falls with graph size (0.135 to 0.066) while its peak share rises (7.4 to 16.1), as a score spread that does not grow with n predicts.

Language models. With the same estimator, the three language models reach sink rates of 0.95–1.00 at $T = 23$ and 0.39–1.00 at $T = 269$. Compared to our experiments, these language models were trained on far longer contexts, which bring stronger sinks (Barbero et al., 2025).

Virtual-node details. At the peak-sinking layer, the virtual node of GRIT + virtual node is never the most-attended node, and it almost never receives enough attention to be a sink. In GraphGPS + virtual node it is the most-attended node in 10% of graphs; when it is, its value norm is about a tenth of the mean.

Over-mixing. From first to last layer the Dirichlet energy relative to the squared feature norm grows in every arm (on average 1.13- to 5.45-fold) except GRIT + virtual node (PE rec.), where it falls to 0.71–0.92 of its input value, and the distance from consensus $\|X - 1\bar{x}^\top\|_F / \|X\|_F$ rises in every arm: at 10 layers there is little over-smoothing for a sink to prevent.

Table 3: All ZINC arms (mean $\pm$ 95% CI over three seeds, five for GraphGPS). $n \max_j s_j$: attention received by the most-attended node, as a multiple of the uniform share. *Ablation ratio*: test MAE after removing all attention onto the most-attended node, divided by that for a random node.

Arm	Sink rate	Test MAE	$n \max_j s_j$	Abl. removed	Abl. renorm.
GRIT	0.090 $\pm$ 0.028	0.064 $\pm$ 0.009	2.6 $\pm$ 0.5	2.06 $\pm$ 0.79	1.35 $\pm$ 0.62
GRIT-dotlogit	0.680 $\pm$ 0.098	0.373 $\pm$ 0.224	13.1 $\pm$ 3.1	1.22 $\pm$ 0.43	1.23 $\pm$ 0.46
GRIT-noclamp	0.113 $\pm$ 0.039	0.063 $\pm$ 0.018	3.0 $\pm$ 0.2	1.91 $\pm$ 1.22	1.37 $\pm$ 0.24
GRIT-official	0.084 $\pm$ 0.037	0.059 $\pm$ 0.006	2.5 $\pm$ 1.4	1.83 $\pm$ 1.14	1.55 $\pm$ 0.48
GRIT-blind	0.214 $\pm$ 0.143	0.542 $\pm$ 0.018	3.6 $\pm$ 2.0	1.17 $\pm$ 0.07	1.25 $\pm$ 0.17
GRIT-dot	0.532 $\pm$ 0.061	0.543 $\pm$ 0.006	9.4 $\pm$ 1.0	1.15 $\pm$ 0.12	1.14 $\pm$ 0.11
GRIT-dotbias	0.521 $\pm$ 0.059	0.268 $\pm$ 0.068	9.7 $\pm$ 4.0	1.19 $\pm$ 0.19	1.20 $\pm$ 0.23
GRIT-entropy-1	0.131 $\pm$ 0.027	0.082 $\pm$ 0.036	3.4 $\pm$ 0.3	1.83 $\pm$ 0.48	1.51 $\pm$ 0.65
GRIT-entropy-2	0.127 $\pm$ 0.031	0.089 $\pm$ 0.010	3.3 $\pm$ 0.3	1.71 $\pm$ 0.64	1.41 $\pm$ 0.37
GRIT + virtual node	0.095 $\pm$ 0.019	0.065 $\pm$ 0.008	2.7 $\pm$ 0.2	1.93 $\pm$ 1.19	1.61 $\pm$ 0.18
GRIT + virtual node + bias	0.079 $\pm$ 0.027	0.066 $\pm$ 0.008	2.7 $\pm$ 0.4	2.13 $\pm$ 0.87	1.54 $\pm$ 0.50
GRIT + virtual node (PE rec.)	0.161 $\pm$ 0.039	0.177 $\pm$ 0.034	3.0 $\pm$ 0.7	1.44 $\pm$ 0.31	1.47 $\pm$ 0.39
GraphGPS (GINE)	0.415 $\pm$ 0.194	0.072 $\pm$ 0.005	6.9 $\pm$ 2.2	1.01 $\pm$ 0.02	1.00 $\pm$ 0.01
GraphGPS-GatedGCN	0.410 $\pm$ 0.119	0.072 $\pm$ 0.007	6.0 $\pm$ 1.1	1.02 $\pm$ 0.02	1.00 $\pm$ 0.06
GraphGPS-none	0.507 $\pm$ 0.159	0.200 $\pm$ 0.017	8.3 $\pm$ 3.3	1.11 $\pm$ 0.09	2.03 $\pm$ 2.08
Transformer + LapPE	0.511 $\pm$ 0.247	0.661 $\pm$ 0.038	7.8 $\pm$ 3.1	1.00 $\pm$ 0.03	1.03 $\pm$ 0.07
GraphGPS + virtual node	0.559 $\pm$ 0.177	0.066 $\pm$ 0.015	8.8 $\pm$ 1.9	1.00 $\pm$ 0.02	1.00 $\pm$ 0.02
GraphGPS + virtual node (PE rec.)	0.629 $\pm$ 0.199	0.086 $\pm$ 0.009	9.5 $\pm$ 2.2	1.01 $\pm$ 0.03	1.01 $\pm$ 0.03

Table 4: % of heads in which one node, whose input is scaled by ρ , becomes a sink (max over layers).

	GRIT	GRIT-noclamp	GRIT-official	GRIT-dotlogit	GRIT-dot	GraphGPS
$\rho = 2$	0.2	1.4	0.4	38.8	34.1	11.5
$\rho = 8$	0.3	16.2	5.8	62.0	48.1	26.2
$\rho = 32$	1.3	54.0	19.4	66.9	51.9	32.1