# Expressive Power of Graph Transformers via Logic*

Veeti Ahvonen[1], Maurice Funk[2], Damian Heiman[1], Antti Kuusisto[1] and Carsten Lutz[2]

[1]Tampere University, Mathematics Research Centre, Finland
[2]Leipzig University and ScaDS.AI Center Dresden/Leipzig, Germany

**Abstract**

We study the expressive power of graph transformers (GTs) and GPS-networks, under both soft-attention and average hard-attention, by providing exact logical characterizations. In the setting with real numbers, GPS-networks have the same expressive power as graded modal logic with the (non-counting) global modality (GML+G), relative to vertex properties definable in first-order logic (FO). With floating-point numbers, GPS-networks are equally expressive as graded modal logic with the *counting* global modality (GML+GC), and this characterization is absolute (not restricted to FO-definable properties). Analogous results hold for GTs in terms of propositional logic with the global modality (PL+G) and its counting variant (PL+GC). A key insight is that the transition from reals to floats swaps relative global counting (possible with reals, lost with floats and with FO) for absolute global counting (impossible with reals, gained with floats).

**Keywords**

graph transformers, GPS-networks, graded modal logic, expressive power, floating-point numbers, bisimulation

## 1. Introduction and Setting

Graph transformers (GTs) and GPS-networks are prominent architectures for graph learning that combine the global attention mechanism of transformers with graph-structured data. Transformers attend globally to every vertex in one step, bypassing the locality of message passing in graph neural networks (GNNs) and thus avoiding well-known pathologies such as over-squashing [3] and over-smoothing [4]. While the expressive power of GNNs has been studied extensively via logic [5, 6], analogous characterizations for models that incorporate transformer layers were missing. In this extended abstract of the AAAI 2026 paper [1], we fill this gap by providing uniform logical characterizations—without imposing a constant bound on graph size—for both the idealized setting of real numbers and the practical setting of floating-point numbers (floats). Graph transformers have had applications in fraud detection [7], molecular property prediction [8] and recommender systems [9].

**Models.** We study three families of models. *Graph neural networks* (GNNs) propagate information via local message passing, updating the feature vector of each vertex based on the feature vectors of its neighbours. We also consider GNNs extended with a *global readout* mechanism: the non-counting variant GNN+G appends a set-based aggregation over all vertices (insensitive to multiplicities), while the counting variant GNN+GC can distinguish how many vertices carry each feature value. Building on GNNs, *GPS-networks* [10] interleave GNN message-passing layers with global transformer layers, and *graph transformers* (GTs) [11] use only transformer layers, treating the graph as a bag of vertices. We study all models in their "naked" form (no positional encodings) focusing on vertex classification. For GPS-networks and GTs, we study both *soft-attention* (softmax) and *average hard-attention* (AH). A particularly notable finding, discussed in Section 3, is that with floats, GPS-networks and GNN+GCs are equally expressive, meaning these two architecturally different models can simulate each other.

The key component of transformer layers is the self-attention head SA, defined by the equation $\mathrm{SA}(X) = \alpha\left((XW_Q)(XW_K)^\top/\sqrt{d_h}\right)(XW_V)$, where $X$ is an input matrix with $d$ columns, $d_h \in \mathbb{Z}_+$, $W_Q$, $W_K$ and $W_V$ are fixed $(d \times d_h)$-matrices, and $\alpha$ is an attention function (softmax or AH).

*LOGICNN 2026: Workshop on Logical Methods for Neural Network Analysis, July 25, 2026, Lisbon, Portugal*
*Extended abstract of the paper published at AAAI 2026 [1]. Full version available as a technical report [2].
veeti.ahvonen@tuni.fi (V. Ahvonen); maurice.funk@uni-leipzig.de (M. Funk); damian.heiman@tuni.fi (D. Heiman); antti.kuusisto@tuni.fi (A. Kuusisto); carsten.lutz@uni-leipzig.de (C. Lutz)

**Logics.** We add to the logical hierarchy used by Barceló et al. [5]: graded modal logic (GML) extends propositional logic with local diamond operators $\Diamond_{\geq k}\varphi$ meaning "at least $k$ out-neighbours satisfy $\varphi$". We extend it with a *global modality*: GML+GC adds $\langle \mathsf{G} \rangle_{\geq k}\, \varphi$ meaning "at least $k$ vertices satisfy $\varphi$". GML+G limits these to $k = 1$. The propositional fragments PL+G and PL+GC drop local diamonds.

**Related work.** The logical expressive power of GNNs was pioneered by Barceló et al. [5], who showed that GNNs capture GML relative to FO, and GNNs with counting global readout (GNN+GC) capture GML+GC (equivalently, the two variable fragment of FO with counting quantifiers on undirected graphs). The expressive power of transformers *on words* has been studied from a logic-based perspective in multiple works; see, e.g., [12, 13, 14, 15]. Rosenbluth et al. [16] studied GPS-networks in a non-logic function-approximation setting, showing that GNN+GCs and GPS-networks are incomparable over reals. We are the first to give exact logical characterizations of graph transformer architectures.

## 2. Results with Real Numbers

Our first main contribution characterizes those GPS-networks over reals that express FO-properties.

**Theorem 1** ([1])**.** *Relative to* FO*, soft-attention* GPS*-networks, average hard-attention* GPS*-networks, and* GML+G *all have the same expressive power.*

Barceló et al. [5] showed that plain GNNs capture exactly GML relative to FO, while GNN+GCs capture GML+GC. Theorem 1 places GPS-networks strictly between: they can express global existential properties such as $P_1$ = "some vertex has label $p$", but *cannot* count absolutely, e.g., $P_2$ = "at least 2 vertices have label $p$". Meanwhile, $P_1$ witnesses that GPS-networks are strictly more expressive than plain GNNs, and $P_2$ shows they are strictly weaker than GNN+GCs.

Although GPS-networks cannot do absolute global counting, they *can* count relatively. For example, the property "at least half of the vertices have label $p$" is expressible: set $W_Q = W_K = \mathbf{0}$, so $\mathrm{softmax}$ distributes attention uniformly, and $W_V$ extracts the $p$-indicator; thus the output at each vertex equals the fraction of $p$-labeled vertices. A threshold check in a multi-layer perceptron (MLP) then decides the property. Relative counting is not expressible in FO, so Theorem 1 must be stated relative to FO.

**Proof sketch.** The easier direction constructs a GPS-network for each GML+G-formula by extending the GML-to-GNN translation of [5], using self-attention heads to handle subformulae of the form $\langle \mathsf{G} \rangle\, \varphi$. A notable difference from [5] is that step-function activations are required: both $\mathrm{softmax}$ and AH can produce arbitrarily small fractional values that cannot be "rectified" to a Boolean 1 with ReLU alone.

The harder direction introduces a new notion of bisimulation. Pointed graphs $(\mathcal{G}_1, v_1)$ and $(\mathcal{G}_2, v_2)$ are *global-ratio graded bisimilar* ($\sim_{G\%}$) if they are graded bisimilar [17] *and* there exists some $q \in \mathbb{Q}_+$ such that for each graded bisimulation type $t$, the vertex count of type $t$ in $\mathcal{G}_1$ equals $q$ times that in $\mathcal{G}_2$. Intuitively, the two graphs are "scaled copies" of each other with respect to type distribution. A layer-by-layer analysis shows GPS-networks are invariant under $\sim_{G\%}$. A new van Benthem/Rosen-style theorem—combining and extending techniques from [18, 17]—then establishes that every FO-formula invariant under $\sim_{G\%}$ is equivalent to a GML+G-formula, completing the characterization.

An analogous argument for GTs, which lack message-passing layers, yields:

**Theorem 2** ([1])**.** *Relative to* FO*, soft-attention* GT*s, average hard-attention* GT*s, and* PL+G *all have the same expressive power.*

## 3. Results with Floating-Point Numbers

Float arithmetic differs from real arithmetic in two ways that are decisive for expressive power.

*Bounded aggregation:* Repeated summation in any float format $\mathcal{F}$ saturates. Let $\mathrm{SUM}_{\mathcal{F}}$ sum a multiset in increasing order. By [19], $\mathrm{SUM}_{\mathcal{F}}$ is *bounded*: for some $k$, $\mathrm{SUM}_{\mathcal{F}}(M) = \mathrm{SUM}_{\mathcal{F}}(M_{|k})$ for all $M$,

where $M_{|k}(F) = \min\{M(F), k\}$. Based on this property of float arithmetic, we assume float-based models use bounded aggregation with sums taken in increasing order. *Underflow:* A result closer to 0 than to any other float in $\mathcal{F}$ rounds to 0; this is an inherent aspect of floating-point arithmetic.

These two phenomena are the key to absolute counting with floats. They also *destroy* relative counting: due to saturation, softmax and AH lose accuracy on large graphs, so the relative counting construction from Section 2 no longer works. Going from reals to floats thus swaps one counting ability for the other. We write GT[F] to denote float-based GTs, and likewise for GPS-networks and GNNs.

**Theorem 3** ([1]). *The following have the same expressive power (without relativizing to* FO*):* GML+GC*, soft-attention* GPS[F]*-networks, average hard-attention* GPS[F]*-networks, and* GNN+GC[F]*s.*

**Theorem 4** ([1]). *The following have the same expressive power:* PL+GC*, soft-attention* GT[F]*s, and average hard-attention* GT[F]*s.*

Both results also hold for *simple* models (sum aggregation and two-layer MLPs with ReLU-activations) and require no background logic. They extend to graph classification, non-Boolean classification, and word-shaped graphs (where GT[F]s reduce to encoder-only transformers without masking such as BERT [20]). With minor changes they cover positional encodings; see the technical report [2].

**Proof sketch (**GT[F] $\equiv$ PL+GC**).** GT[F]s to PL+GC: all local computation steps (MLPs, matrix products $XW_Q$, $XW_K$, $XW_V$) are functions on bit strings, expressible bit-by-bit in PL. For the global attention step, the sum over all vertices depends only on the multiplicity of each feature vector up to the saturation bound $k$, captured by $\langle \mathsf{G} \rangle_{\geq k}$. For the converse, to simulate $\langle \mathsf{G} \rangle_{\geq k}\, \psi$, we construct $W_Q$ and $W_K$ so that softmax places weight $1/\ell$ uniformly across rows, where $\ell$ is the number of vertices satisfying $\psi$. We construct $W_V$ to multiply this by a suitably small float, which triggers underflow (yields 0) exactly when $\ell \geq k$, allowing subsequent MLPs to extract the truth value of $\langle \mathsf{G} \rangle_{\geq k}\, \psi$.

## 4. Summary and Open Questions

Table 1 summarizes our characterizations. An interesting open question is whether every GPS-network over reals is equivalent to some GNN+GC; this holds for floats (Theorem 3). Another direction is extending these characterizations to models with positional encodings such as the graph Laplacian.

**Table 1**
Logical characterizations of graph learning models. All results hold for soft-attention *and* average hard-attention.

| Model | Numbers | Equivalent logic | Scope |
|---|---|---|---|
| GPS-network | Reals | GML+G | relative to FO |
| GT | Reals | PL+G | relative to FO |
| GPS[F]-network, GNN+GC[F] | Floats | GML+GC | absolute |
| GT[F] | Floats | PL+GC | absolute |


## Acknowledgments

V. Ahvonen was supported by the Vilho, Yrjö and Kalle Väisälä Foundation. D. Heiman was supported by the Magnus Ehrnrooth Foundation. A. Kuusisto was supported by the Research Council of Finland, project number 369424. C. Lutz was supported by DFG project LU 1417/4-1.


## Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 4.6 in order to: condense and format the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] V. Ahvonen, M. Funk, D. Heiman, A. Kuusisto, C. Lutz, Expressive power of graph transformers via logic, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, 2026, pp. 19569–19579.

[2] V. Ahvonen, M. Funk, D. Heiman, A. Kuusisto, C. Lutz, Expressive power of graph transformers via logic, 2025. URL: https://arxiv.org/abs/2508.01067. arXiv:2508.01067.

[3] U. Alon, E. Yahav, On the bottleneck of graph neural networks and its practical implications, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021, pp. 14048–14063.

[4] Q. Li, Z. Han, X. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, in: S. A. McIlraith, K. Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), AAAI Press, 2018, pp. 3538–3545. doi:10.1609/AAAI.V32I1.11604.

[5] P. Barceló, E. V. Kostylev, M. Monet, J. Pérez, J. L. Reutter, J. P. Silva, The logical expressiveness of graph neural networks, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020, pp. 11612–11632.

[6] M. Benedikt, C. Lu, B. Motik, T. Tan, Decidability of graph neural networks via logical characterizations, in: K. Bringmann, M. Grohe, G. Puppis, O. Svensson (Eds.), 51st International Colloquium on Automata, Languages, and Programming, ICALP 2024, July 8-12, 2024, Tallinn, Estonia, volume 297 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024, pp. 127:1–127:20. URL: https://doi.org/10.4230/LIPIcs.ICALP.2024.127. doi:10.4230/LIPICS.ICALP.2024.127.

[7] J. Lin, X. Guo, Y. Zhu, S. Mitchell, E. Altman, J. Shun, Fraudgt: A simple, effective, and efficient graph transformer for financial fraud detection, in: Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 292–300. URL: https://doi.org/10.1145/3677052.3698648. doi:10.1145/3677052.3698648.

[8] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, J. Huang, Self-supervised graph transformer on large-scale molecular data, Advances in neural information processing systems 33 (2020) 12559–12571.

[9] C. Li, L. Xia, X. Ren, Y. Ye, Y. Xu, C. Huang, Graph transformer for recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1680–1689. URL: https://doi.org/10.1145/3539618.3591723. doi:10.1145/3539618.3591723.

[10] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, D. Beaini, Recipe for a general, powerful, scalable graph transformer, Advances in Neural Information Processing Systems 35 (2022) 14501–14515.

[11] V. P. Dwivedi, X. Bresson, A generalization of transformer networks to graphs, CoRR abs/2012.09699 (2020). URL: https://arxiv.org/abs/2012.09699. arXiv:2012.09699.

[12] J. Li, R. Cotterell, Characterizing the expressivity of fixed-precision transformer language models, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

[13] S. Jerad, A. Svete, J. Li, R. Cotterell, Unique hard attention: A tale of two sides, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2025, pp. 977–996.

[14] A. Yang, D. Chiang, D. Angluin, Masked hard-attention transformers recognize exactly the star-free languages, 2024. URL: https://arxiv.org/abs/2310.13897. arXiv:2310.13897.

[15] D. Chiang, P. Cholak, A. Pillay, Tighter bounds on the expressivity of transformer encoders, 2023. URL: https://arxiv.org/abs/2301.10743. arXiv:2301.10743.

[16] E. Rosenbluth, J. Tönshoff, M. Ritzert, B. Kisin, M. Grohe, Distinguished in uniform: Self-attention vs. virtual nodes, in: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024, OpenReview.net, 2024, pp. 17926–17946. URL: https://openreview.net/forum?id=AcSChDWL6V.

[17] M. Otto, Graded modal logic and counting bisimulation, CoRR abs/1910.00039 (2019). URL:

http://arxiv.org/abs/1910.00039. arXiv:1910.00039.

[18] M. Otto, Modal and guarded characterisation theorems over finite transition systems, Annals of Pure and Applied Logic 130 (2004) 173–205. doi:10.1016/j.apal.2004.04.003.

[19] V. Ahvonen, D. Heiman, A. Kuusisto, C. Lutz, Logical Characterizations of Recurrent Graph Neural Networks with Reals and Floats, Advances in Neural Information Processing Systems 37 (2024) 104205–104249.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.