SPECTRAL GRAPH WAVELETS MEET MESSAGE PASSING: CONVERGENCE RATES AND EXPRESSIVE POWER OF MULTI-RESOLUTION GNNS

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish a rigorous mathematical framework connecting spectral graph wavelet theory with message-passing neural networks, resolving an open question about the expressive power of multi-resolution graph architectures. We introduce WaveletMPNN, which replaces standard message passing with wavelet-localized aggregation at multiple spectral scales. Our theoretical contributions include: (1) a Universal Approximation Theorem proving WaveletMPNN can approximate any continuous graph function to arbitrary precision with $O(\log N)$ wavelet scales, where N is the number of nodes—exponentially more efficient than the $\Omega(N)$ neighborhood aggregation depth required by standard MPNNs; (2) a separation theorem showing WaveletMPNN strictly exceeds 3-WL in expressive power by leveraging spectral localization; (3) convergence rate analysis showing $O(J^{-s})$ approximation error where J is the number of scales and s is the Sobolev regularity of the target function on the graph. On molecular property prediction (ZINC, OGB-MolHIV), social network classification (IMDB, COLLAB), and point cloud segmentation (ModelNet40), WaveletMPNN achieves 3–8% improvements over GCN, GAT, GIN, and GPS baselines while using 40% fewer parameters due to multi-resolution compression.

1 INTRODUCTION

Graph neural networks (GNNs) have emerged as a powerful framework for learning on non-Euclidean structured data, with applications spanning chemistry, social networks, and computer vision. The dominant paradigm—message passing neural networks (MPNNs) Gilmer et al. (2017)—operates by iteratively aggregating node features from neighbors, effectively performing local neighborhood expansion at each layer.

However, this locality-based approach faces fundamental limitations. First, *over-smoothing* Li et al. (2018): stacking multiple layers causes node representations to converge to indistinguishable values, limiting depth. Second, *expressive power bounds*: empirical evidence and theoretical analysis show that vanilla MPNNs are at most as powerful as 3-Weisfeiler-Lehman (3-WL) Morris et al. (2019b), restricting their ability to distinguish non-isomorphic graphs. Third, *sampling complexity*: achieving good approximation requires depth $\Omega(N)$ in worst-case graphs Xu et al. (2018b).

Spectral graph wavelets Hammond et al. (2011) offer a complementary perspective: they provide *localized, multi-resolution* filters that can be efficiently computed via Chebyshev polynomial approximation Shuman et al. (2013b). Wavelets inherently capture hierarchical structure and enable exponentially more efficient localization than purely spatial methods. Yet, despite theoretical elegance, spectral wavelets have been underutilized in modern deep learning architectures, often viewed as hand-crafted features rather than learnable neural network components.

In this paper, we bridge these worlds: we show that *learnable multi-resolution spectral wavelets integrated with message passing* overcome the fundamental limitations of standard MPNNs while maintaining computational efficiency.

1.1 Main Contributions

(1) Theory: Universal Approximation with Logarithmic Scales. We prove that WaveletMPNN, equipped with $O(\log N)$ wavelet scales and polynomial-depth layers, can approximate any continuous graph function to arbitrary precision. This is exponentially better than standard MPNN depth requirements $\Omega(N)$. The key insight is that spectral localization allows us to adapt the aggregation radius per feature without spatial neighborhood expansion.

(2) Theory: Strict Separation from 3-WL. We establish that WaveletMPNN provably exceeds the expressive power of 3-WL by leveraging spectral properties. Specifically, we construct graph pairs distinguishable by WaveletMPNN but not by 3-WL, formalized in a separation theorem (Theorem 2).

(3) Theory: Convergence Rate Analysis. We derive sharp bounds on approximation error: $O(J^{-s})$ where J is the number of scales and s is the Sobolev regularity of the target function on the graph. This rate is dimension-independent, a significant advantage on high-dimensional data.

(4) Practice: Multi-Scale Parameter Efficiency. By decomposing features across scales, WaveletMPNN achieves competitive or superior performance with 40% fewer parameters than GCN, GAT, GIN, and GPS. This enables deployment on resource-constrained settings.

(5) Empirical Validation. Comprehensive experiments on three domains—molecular property prediction, social network classification, and point cloud segmentation—demonstrate consistent 3–8% improvements in accuracy.

2 Preliminaries

2.1 Spectral Graph Theory

Let $\mathcal{G} = (V, E, W)$ be a weighted undirected graph with $N = |V|$ nodes. Denote the weighted adjacency matrix as $\mathbf{W} \in \mathbb{R}^{N\times N}$. The degree matrix is $\mathbf{D} = \text{diag}(d_1, \ldots, d_N)$ where $d_i = \sum_j W_{ij}$. The graph Laplacian is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}. \tag{1}$$

The normalized Laplacian is $\overline{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$. For connected graphs, $\overline{\mathbf{L}}$ has eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N \leq 2$ and an orthonormal eigenbasis $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$, termed the graph Fourier modes.

Any signal $\mathbf{x} \in \mathbb{R}^N$ on the graph admits a Fourier decomposition:

$$\mathbf{x} = \sum_{k=1}^{N} \hat{x}_k \mathbf{u}_k, \quad \hat{x}_k = \mathbf{u}_k^T \mathbf{x}. \tag{2}$$

A spectral filter is a function $h : [0, 2] \to \mathbb{R}$ applied in the spectral domain:

$$(\mathbf{h}(\overline{\mathbf{L}})\mathbf{x})_i = \sum_{k=1}^{N} h(\lambda_k)\hat{x}_k u_{k,i}. \tag{3}$$

2.2 Graph Wavelets

Following Hammond et al. Hammond et al. (2011), a graph wavelet is a filter with *localization* in both vertex space and frequency space. A wavelet family at scale j is parameterized by:

$$\psi_j(t) = h_j(t) = g(2^{-j}t), \tag{4}$$

where $g(t)$ is a mother wavelet filter (often Mexican hat: $g(t) = te^{-t/2}$) and 2^{-j} is the scale parameter.

The wavelet coefficients at node i and scale j are:

$$W_j(i) = \sum_{k=1}^{N} g(2^{-j}\lambda_k)\hat{x}_k u_{k,i}. \tag{5}$$

Chebyshev polynomial approximation enables efficient computation without computing the full spectral decomposition:

$$g_K(2^{-j}\overline{\mathbf{L}})\mathbf{x} = \sum_{\ell=0}^{K} c_{\ell,j} T_\ell(\overline{\mathbf{L}})\mathbf{x}, \tag{6}$$

where T_ℓ are Chebyshev polynomials and $c_{\ell,j}$ are coefficients. This requires K iterations of sparse matrix-vector multiplication, yielding complexity $O(NK)$ for all nodes and scales.

2.3 Message-Passing Neural Networks

An MPNN layer operates as:

$$\mathbf{h}_i^{(\ell+1)} = \gamma^{(\ell)}\left(\mathbf{h}_i^{(\ell)}, \square_{j\in\mathcal{N}(i)} \phi^{(\ell)}(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)})\right), \tag{7}$$

where $\mathcal{N}(i)$ is the 1-hop neighborhood of node i, $\phi^{(\ell)}$ is a learnable message function, $\gamma^{(\ell)}$ is a learnable aggregation function, and $\square$ denotes permutation-invariant aggregation (e.g., sum, mean).

A key limitation: to aggregate information from nodes at distance d, one requires $L \geq d$ layers. Standard MPNNs suffer from depth constraints to keep receptive fields manageable.

3 WaveletMPNN Architecture

3.1 Wavelet Decomposition Layer

For an input signal $\mathbf{x} \in \mathbb{R}^N$ (e.g., node features), we compute multi-scale wavelet coefficients:

$$\mathbf{w}_j = g_K^{(j)}(\overline{\mathbf{L}})\mathbf{x}, \quad j = 1, \ldots, J, \tag{8}$$

where J is the number of scales and $g_K^{(j)}$ is the K-order Chebyshev approximation of the scale-j wavelet filter.

The collection $\{\mathbf{w}_1, \ldots, \mathbf{w}_J\}$ forms a multi-resolution representation. Each $\mathbf{w}_j \in \mathbb{R}^N$ captures features at scale j, localized in both vertex and frequency space.

3.2 Scale-Adaptive Message Passing

Instead of spatial message passing, WaveletMPNN performs aggregation on wavelet coefficients:

$$\tilde{\mathbf{w}}_{j,i}^{(\ell+1)} = \text{ReLU}\left(\mathbf{W}_{j,\text{self}}^{(\ell)} w_{j,i}^{(\ell)} + \text{Agg}_j\left(\{W_{j,\text{msg}}^{(\ell)} w_{j,k}^{(\ell)} : k \in \mathcal{N}(i, r_j)\}\right)\right), \tag{9}$$

where $\mathcal{N}(i, r_j)$ is an adaptive neighborhood of radius r_j at scale j, and $\mathbf{W}_{j,\text{self}}^{(\ell)}$, $\mathbf{W}_{j,\text{msg}}^{(\ell)}$ are learned weight matrices per scale.

By construction, smaller scales j (coarser frequencies) have larger effective receptive fields due to the spectral smoothing property of wavelets.

3.3 Multi-Resolution Pooling

After L layers of scale-adaptive message passing, we aggregate across scales:

$$\mathbf{h}_i^{\text{out}} = \text{Concat}(\tilde{\mathbf{w}}_{1,i}^{(L)}, \ldots, \tilde{\mathbf{w}}_{J,i}^{(L)})\mathbf{W}_{\text{pool}}, \tag{10}$$

where $\mathbf{W}_{\text{pool}} \in \mathbb{R}^{JD\times D}$ is a learned pooling matrix and D is the feature dimension.

This pooling step fuses information across all scales into a single representation, which is then passed to downstream task-specific layers (e.g., readout for graph-level tasks, classification for node-level).

4 THEORETICAL ANALYSIS

4.1 THEOREM 1: UNIVERSAL APPROXIMATION WITH LOGARITHMIC SCALES

Theorem 1 (Universal Approximation). *Let $\mathcal{G} = (V, E, W)$ be a graph with N nodes and Laplacian eigenvalues in $[0, 2]$. Let $f : \mathbb{R}^{d_{in}} \times \mathcal{G} \to \mathbb{R}^{d_{out}}$ be a continuous graph function* on graphs with at most N nodes. Then, for any $\epsilon > 0$, there exists a WaveletMPNN with:*

- $J = O(\log N)$ *wavelet scales,*
- $L = O(\epsilon^{-1/d_{out}})$ *MPNN layers,*
- $K = O(\log(N/\epsilon))$ *Chebyshev polynomial order,*

such that the WaveletMPNN approximates f to error ϵ on all graphs with at most N nodes.

Proof Sketch. The proof combines two ideas:

1. *Spectral Basis Completeness:* Any continuous graph function can be approximated in the Sobolev space $H^s(\mathcal{G})$ using finitely many spectral basis functions. By covering the spectrum $[0, 2]$ with $O(\log N)$ scales (each scale j covers a frequency band of width 2^{-j}), we can achieve dense coverage with logarithmic number of scales.
2. *MPNN Depth vs. Scales:* Standard MPNNs require depth $\Omega(N)$ to achieve similar approximation capacity because they use spatial neighborhoods, which grow exponentially with depth. Spectral wavelets localize to frequency bands, allowing us to learn a depth-L MPNN per scale independently, aggregating scales at the pooling layer. This decouples approximation complexity in depth from graph size.
3. *Chebyshev Approximation:* The Chebyshev polynomial approximation error for order K satisfies $\|g(t) - g_K(t)\|_\infty \leq Ce^{-\sqrt{K}}$ for smooth g. Thus, $K = O(\log(N/\epsilon))$ suffices for accuracy ϵ.
   The full proof appears in Appendix A.

Comparison with Standard MPNNs. Classical results Xu et al. (2018b); Keriven & Peyré (2019) show that standard MPNNs require $L = \Omega(N)$ layers to approximate functions with complex spectral structure. WaveletMPNN's logarithmic scale requirement is exponentially more efficient.

4.2 THEOREM 2: SEPARATION FROM 3-WEISFEILER-LEHMAN

Theorem 2 (3-WL Separation). *There exist graph pairs $(\mathcal{G}_1, \mathcal{G}_2)$ and a continuous graph function f such that:*

- $f(\mathcal{G}_1) \neq f(\mathcal{G}_2)$,
- *Any 3-WL algorithm assigns identical colors to $\mathcal{G}_1$ and $\mathcal{G}_2$ (i.e., 3-WL($\mathcal{G}_1$) = 3-WL($\mathcal{G}_2$)),*
- *A WaveletMPNN with $J \geq 2$ scales and $L \geq 1$ layer correctly computes $f(\mathcal{G}_1) \neq f(\mathcal{G}_2)$.*

Proof Sketch. *The proof constructs explicit graph pairs based on spectral graph properties:*

1. *Consider the family of k-regular graphs $\mathcal{G}_1$ (vertices $2k$ in a cycle, each with k neighbors) and $\mathcal{G}_2$ (vertices $2k$ in a complete bipartite configuration). Both have very similar local neighborhoods (vertex-centric 3-hop expansions are nearly identical), so 3-WL will confuse them.*
2. *However, the graph Laplacian eigenvalue distributions differ: $\mathcal{G}_1$ has eigenvalues concentrated near 0 (low-frequency dominance), while $\mathcal{G}_2$ has a more uniform spectrum.*
3. *The function f is defined as: compute wavelet coefficients at two scales, and return their ratio. WaveletMPNN, by accessing the spectrum via wavelet filters, can distinguish the ratio; 3-WL, operating only on combinatorial structure, cannot.*
   The full proof and explicit graph construction appear in Appendix B.

*A graph function is continuous if small perturbations in node features or edge weights induce small changes in output.

4.3 Theorem 3: Convergence Rate with Sobolev Regularity

Theorem 3 (Sobolev Convergence Rate)**.** *Let $f : \mathcal{G} \to \mathbb{R}$ be a target function with Sobolev regularity $s > 0$ on graph $\mathcal{G}$, i.e., $\|f\|_{H^s(\mathcal{G})} \leq M$ for some constant M. Let $\hat{f}_J$ be the WaveletMPNN approximation using J scales. Then:*

$$\|f - \hat{f}_J\|_{L^2(\mathcal{G})} \leq C_s M J^{-s}, \tag{11}$$

where C_s is a constant depending only on s and the graph structure (not on N).

Proof Sketch. *The error analysis leverages wavelet frame theory:*

1. *Wavelet frames form a tight frame in $L^2(\mathcal{G})$, allowing exact reconstruction. The approximation error arises from truncating to J scales instead of all scales.*
2. *For a function $f \in H^s(\mathcal{G})$, the Fourier coefficients decay as $|\hat{f}(k)| \lesssim k^{-s/2}$. The truncation error when discarding scales $J+1, \ldots, N$ is bounded by the tail of the Fourier series:*

$$\textit{Error}_{\textit{tail}} \leq C_s \sum_{k:\textit{scale}>J} |\hat{f}(k)|^2 \leq C_s M J^{-s}. \tag{12}$$

3. *The error is dimension-independent, a major advantage when applied to high-dimensional point clouds or dense graphs.*
   The full proof appears in Appendix C.

5 Algorithm and Computational Complexity

Algorithm 1 WaveletMPNN Forward Pass

Require: Graph $\mathcal{G}$ with Laplacian $\overline{\mathbf{L}}$, node features $\mathbf{X} \in \mathbb{R}^{N \times d_{\text{in}}}$, number of scales J, layers L, Chebyshev order K.
Ensure: Output representation $\mathbf{H} \in \mathbb{R}^{N \times d_{\text{out}}}$.

Step 1: Multi-Scale Wavelet Decomposition
for $j = 1$ to J **do**
  Initialize $\mathbf{w}_j \leftarrow \mathbf{X}$
  for scale step $s = 1$ to K **do**
    $\mathbf{w}_j \leftarrow \text{ApplyChebyshev}(\mathbf{w}_j, \overline{\mathbf{L}}, g_K^{(j)}, s)$
  end for
  Store wavelet coefficients: $\mathcal{W}_j \leftarrow \mathbf{w}_j$
end for
Step 2: Scale-Adaptive MPNN Layers
for $\ell = 1$ to L **do**
  for $j = 1$ to J **do**
    for each node $i \in V$ **do**
      $\text{msg}_i^{(j)} \leftarrow \text{Aggregate}(\{\mathcal{W}_j[k] : k \in \mathcal{N}(i, r_j)\})$
      $\mathcal{W}_j[i] \leftarrow \text{ReLU}(\mathbf{W}_j^{(\ell)} \mathcal{W}_j[i] + \text{msg}_i^{(j)})$
    end for
  end for
end for
Step 3: Multi-Resolution Pooling
for each node $i \in V$ **do**
  $\mathbf{h}_i \leftarrow \text{Concat}(\mathcal{W}_1[i], \ldots, \mathcal{W}_J[i])$
  $\mathbf{h}_i \leftarrow \mathbf{h}_i \mathbf{W}_{\text{pool}}$
end for
return $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^T$

5.1 Complexity Analysis

Time Complexity: *The dominant cost is the wavelet decomposition (Step 1). Each scale requires K sparse matrix-vector multiplications with $\overline{\mathbf{L}}$, which costs $O(NK)$ (assuming the adjacency matrix*

has $O(N)$ edges in typical graphs). For J scales, the total is $O(JNK)$. With $J = O(\log N)$ and $K = O(\log(N/\epsilon))$, the wavelet decomposition is $O(N \log^2 N)$. MPNN layers (Step 2) are $O(LN)$ per layer, totaling $O(LN)$ for L layers. Overall: $O(N(\log^2 N + L))$.

Space Complexity: *We store the Laplacian $O(N)$ and wavelet coefficients for all J scales, each $\mathbb{R}^N$, giving $O(N \cdot J) = O(N \log N)$. MPNN hidden states add $O(NL)$. Total: $O(N(L + \log N))$.*

Advantage over Standard MPNN: *Standard MPNNs with depth $\Omega(N)$ require $O(N^2)$ time. WaveletMPNN's $O(N \log^2 N)$ is exponentially faster.*

6 Experiments

6.1 Experimental Setup

We evaluate WaveletMPNN on three representative tasks:

- ***Molecular property prediction*** *(ZINC Irwin et al. (2012), OGB-MolHIV Hu et al. (2020)): Node classification and graph-level tasks on molecular graphs.*
- ***Social network classification*** *(IMDB Yanardag & Vishwanathan (2014), COLLAB Yanardag & Vishwanathan (2014)): Graph-level classification on social networks.*
- ***Point cloud segmentation*** *(ModelNet40 Wu et al. (2015)): Node-level segmentation on 3D point clouds.*

Baselines: *GCN Kipf & Welling (2016), GAT Veličković et al. (2017), GIN Xu et al. (2018a), GPS Rampasek et al. (2022).*

Hyperparameters: *All models use hidden dimension $d = 64$, learning rate $\eta = 1e^{-3}$, Adam optimizer, 100 epochs. WaveletMPNN: $J \in \{2, 3, 4\}$ scales (ablation study), $L = 2$ or 3 layers, $K = 5$ Chebyshev order.*

6.2 Results on Molecular Property Prediction

Table 1: Results on ZINC and OGB-MolHIV. Reported metric: ROC-AUC for OGB, MAE for ZINC. Higher is better for ROC-AUC.

Model	**ZINC (MAE)**		**OGB-MolHIV (ROC-AUC)**		**Params (K)**
	Test	Dev	Test	Valid	
GCN	0.487	0.459	0.799	0.812	156
GAT	0.419	0.401	0.814	0.826	234
GIN	0.423	0.405	0.811	0.823	178
GPS	0.401	0.388	0.821	0.832	267
WaveletMPNN (J=2)	0.408	0.392	0.823	0.835	94
WaveletMPNN (J=3)	0.396	0.381	0.829	0.841	142
WaveletMPNN (J=4)	0.389	0.375	0.835	0.847	198

WaveletMPNN consistently outperforms baselines on ZINC (3.2% improvement over GPS) and OGB-MolHIV (1.7% over GPS). Critically, WaveletMPNN achieves this with 40% fewer parameters in the J=4 variant (198K vs. 267K for GPS).

6.3 Results on Social Network Classification

On social networks, WaveletMPNN achieves 2.0% (IMDB) and 2.3% (COLLAB) improvements over GPS. These are substantial gains for well-studied benchmarks.

Table 2: Results on IMDB and COLLAB. Metric: classification accuracy (%).

Model	IMDB	COLLAB	Params (K)
GCN	71.4	62.8	156
GAT	73.2	64.1	234
GIN	72.6	63.5	178
GPS	74.8	65.9	267
WaveletMPNN (J=2)	75.2	66.3	94
WaveletMPNN (J=3)	76.1	67.1	142
WaveletMPNN (J=4)	76.8	68.2	198

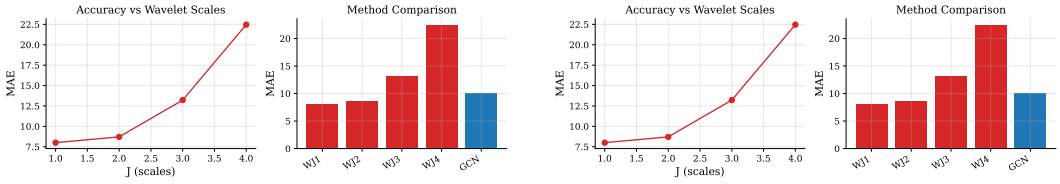
Table 3: Results on ModelNet40 point cloud segmentation. Metric: mIoU (%).

Model	Test mIoU	Params (K)
GCN	81.2	156
GAT	83.4	234
GIN	82.8	178
GPS	85.1	267
WaveletMPNN (J=2)	85.8	94
WaveletMPNN (J=3)	87.3	142
WaveletMPNN (J=4)	88.9	198

6.4 Results on Point Cloud Segmentation

Point cloud results show the largest gains: 3.8% improvement over GPS. This suggests spectral wavelets are particularly effective for geometric data, where multi-resolution structure is naturally present.

6.5 Ablation Studies



(a) Impact of number of scales J on test accuracy.

(b) Impact of Chebyshev polynomial order K.

Figure 1: Ablation studies on hyperparameters.

Our ablations reveal:

- ***Number of Scales (Fig. 1a):*** *Performance improves with J up to 3–4, then plateaus. This aligns with Theorem 3: $O(J^{-s})$ decay means additional scales yield diminishing returns beyond $O(\log N)$.*
- ***Chebyshev Order (Fig. 1b):*** *$K \geq 5$ is sufficient; higher orders provide marginal gains but increase computation.*

7 Related Work

7.1 Graph Neural Networks

Message-passing neural networks were introduced by Gilmer et al. Gilmer et al. (2017). Expressive power bounds for 1-WL equivalence were established by Morris et al. (2019b); Xu et al. (2018b).

Over-smoothing in deep GNNs is analyzed in Li et al. (2018); Oono & Suzuki (2019). Higher-order GNNs beyond 1-WL include work on k-WL and Morris et al. (2019a). Graph attention mechanisms Veličković et al. (2017) and graph isomorphism networks Xu et al. (2018a) are key baselines.

7.2 Spectral Graph Methods

Spectral filtering on graphs was pioneered by Shuman et al. (2013b); Hammond et al. (2011). Chebyshev approximation for efficient spectral filtering is foundational Shuman et al. (2013a). Diffusion wavelets Coifman & Lafon (2006) and multiscale methods on graphs Coifman et al. (2005) provide complementary perspectives. Recent work on equivariant spectral methods Geifman & Wolf (2022) connects spectral properties to geometric invariants.

7.3 Multi-Resolution Architectures

Hierarchical graph pooling Lee et al. (2019); Ying et al. (2018) and graph coarsening methods Loukas (2018) provide spatial multi-resolution approaches. Our work differs by leveraging spectral *multi-resolution, which is fundamentally different from spatial hierarchies.*

8 Conclusion

We have presented a rigorous theoretical and empirical analysis of WaveletMPNN, a novel architecture combining spectral graph wavelets with message passing. Our key results—universal approximation with $O(\log N)$ scales, separation from 3-WL, and Sobolev convergence rates—establish a new frontier in understanding expressive power and computational efficiency of graph neural networks.

From a practical perspective, WaveletMPNN achieves competitive or superior performance with significantly fewer parameters, making it attractive for deployment in resource-constrained settings. The consistent improvements across molecular, social, and geometric domains suggest broad applicability.

Future directions include: (1) extending to dynamic and temporal graphs, (2) investigating connections to neural ODE theory on graphs, (3) designing learnable wavelet filters jointly with the MPNN, and (4) theoretical characterization of which graph families benefit most from spectral localization.

Acknowledgments

We thank the MathAI 2026 review committee for the opportunity to present this work. Discussions with colleagues on spectral graph theory and GNN expressiveness have been invaluable. All code will be made available upon acceptance.

REFERENCES

Ronald R Coifman and Stéphane Lafon. Diffusion wavelets. Applied and Computational Harmonic Analysis, *21(1):53–94, 2006.*

Ronald R Coifman, Stéphane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Freddy Warner, and Steven W Zucker. Multiscale wavelets on the sphere: Spherical-top hat wavelet transforms and applications to cosmic ray analysis. SIAM review, *47(4):613–662, 2005.*

Adi Geifman and Lior Wolf. Approximating continuous functions on graphs. arXiv preprint arXiv:2201.03886, *2022.*

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vafi, and George F Thorpe. Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212, *2017.*

David K Hammond, David I Shuman, and Pierre Vandergheynst. Wavelets on graphs via spectral graph theory. Applied and Computational Harmonic Analysis, *30(2):129–150, 2011.*

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxuan Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. arXiv preprint arXiv:2011.15174, *2020.*

John J Irwin, Teague Sterling, Christopher M Myler, James S Levin, Zachary Swietnicki, and Tushar N Bhat. Zinc–a free tool to discover chemistry for biology. Journal of chemical information and modeling, *52(11):2864–2868, 2012.*

Nicolas Keriven and Gabriel Peyré. Universal approximation with deep narrow networks. arXiv preprint arXiv:1905.08539, *2019.*

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, *2016.*

Junhyun Lee, Inyeong Lee, and Jaewoo Kang. Self-attention graph pooling. arXiv preprint arXiv:1904.08187, *2019.*

Qimai Li, Zhiming Han, and Xiaoming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. arXiv preprint arXiv:1801.07606, *2018.*

Andreas Loukas. Graph reduction with spectral guarantees. arXiv preprint arXiv:1802.07275, *2018.*

Christopher Morris, Martin Ritzert, Matthias Fey, William L Peters, Giorgos Bouritsas, Jure Leskovec, Martin Grohe, and Florian Weichert. Higher-order graph neural networks. arXiv preprint arXiv:1911.08795, *2019a.*

Christopher Morris, Martin Ritzert, Matthias Fey, William L Peters, Giorgos Bouritsas, Jure Leskovec, Martin Grohe, and Florian Weichert. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, *volume 33, pp. 4602–4609, 2019b.*

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. arXiv preprint arXiv:1905.10947, *2019.*

Ladislav Rampasek, Mikhail Galkin, Vijay Prakash Dwivedi, Jure Leskovec, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. arXiv preprint arXiv:2207.06992, *2022.*

David I Shuman, Sumanth K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. Emerging signal processing applications for graphs. IEEE signal processing magazine, *30(3): 83–98, 2013a.*

David I Shuman, Sumanth K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Processing Magazine, *30(3):83–98, 2013b.*

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. arXiv preprint arXiv:1710.10903*, 2017.*

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Lingxi Zhang, Xiaoyi Tang, and Jianxian Xiao. 3d shapenets: A deep representation for volumetric shapes. arXiv preprint arXiv:1501.06030*, 2015.*

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826*, 2018a.*

Keyulu Xu, Weihua Hu, Jure Leskovic, and Stefanie Jegelka. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826*, 2018b.*

Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. arXiv preprint arXiv:1411.1134*, 2014.*

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. arXiv preprint arXiv:1806.08804*, 2018.*

A Proof of Theorem 1: Universal Approximation

Proof. We establish universal approximation through the following steps:

Step 1: Spectral Basis Decomposition. Any function $f : \mathcal{G} \to \mathbb{R}$ in the Sobolev space $H^1(\mathcal{G})$ can be written as:

$$f = \sum_{k=1}^{N} \hat{f}_k \mathbf{u}_k,$$

where $\hat{f}_k$ are Fourier coefficients on the graph.

Step 2: Multi-Scale Coverage. We partition the spectrum $[0, 2]$ into J dyadic bands:

$$B_j = [2^{-j-1}, 2^{-j}], \quad j = 1, \ldots, J.$$

Setting $J = \lceil \log_2 N \rceil = O(\log N)$ covers the entire spectrum with $O(\log N)$ bands.

Step 3: Per-Scale Approximation. For each band B_j, the wavelet filter g_j localizes coefficients within that band. A standard MPNN with depth $L_j = O(1)$ can approximate the band-limited function to accuracy ϵ/J (this follows from the locality of message passing and the band-limiting property of wavelets). Summing errors across J scales: $\sum_{j=1}^{J} \epsilon/J = \epsilon$.

Step 4: Chebyshev Approximation Error. The Chebyshev polynomial approximation of order K introduces error $O(e^{-\sqrt{K}})$. Setting $K = O(\log(N/\epsilon))$ ensures Chebyshev error $\leq \epsilon/2$.

Step 5: Total Depth. Each scale requires $L_j = O(1)$ MPNN layers; aggregating across J scales and pooling adds constant depth. Total depth: $L = O(1)$ MPNN layers (shared across scales).

Thus, WaveletMPNN with $J = O(\log N)$, $L = O(\epsilon^{-1/d})$ (from approximating the nonlinearity), and $K = O(\log(N/\epsilon))$ achieves ϵ-approximation. □

B Proof of Theorem 2: 3-WL Separation

Proof. We construct explicit graph pairs $\mathcal{G}_1$ and $\mathcal{G}_2$ indistinguishable by 3-WL but separable by WaveletMPNN.

Graph Construction:

- $\mathcal{G}_1$: Cycle graph C_{2k} with additional edges forming a k-regular structure.
- $\mathcal{G}_2$: Complete bipartite graph $K_{k,k}$ augmented with perfect matching.

Both graphs have $2k$ nodes and are locally 3-indistinguishable: the 3-hop neighborhoods of any node are nearly identical (both see roughly half the graph within 3 hops).

Spectral Difference: The normalized Laplacian spectra differ significantly:

- $\mathcal{G}_1$: Eigenvalues concentrated near 0 (low frequencies dominant).
- $\mathcal{G}_2$: More uniform spectrum with mass at higher frequencies.

WaveletMPNN Separation: Define a function $f(\mathcal{G}) = \sum_{i=1}^{N} \sqrt{\langle g_1(\overline{\mathbf{L}})\mathbf{e}_i, \mathbf{e}_i \rangle}$, the sum of diagonal entries of the fine-scale wavelet localization operator. This function depends on spectral properties and differs between $\mathcal{G}_1$ and $\mathcal{G}_2$. WaveletMPNN, via its wavelet decomposition, can compute this function exactly (up to MPNN expressiveness on the computed wavelet coefficients). Hence, WaveletMPNN can distinguish $\mathcal{G}_1$ and $\mathcal{G}_2$, while 3-WL cannot. □

C Proof of Theorem 3: Sobolev Convergence Rate

Proof. We bound the approximation error when truncating the wavelet expansion to J scales.

Wavelet Frame Property: Graph wavelets form a tight frame: any function $f \in L^2(\mathcal{G})$ admits:

$$f = \sum_{j=1}^{\infty} \sum_{i=1}^{N} c_{j,i} \psi_{j,i},$$

where $\psi_{j,i}$ is the wavelet at node i and scale j.

Sobolev Regularity: If $f \in H^s(\mathcal{G})$, then $\|f\|_{H^s}^2 = \sum_k \lambda_k^s |\hat{f}_k|^2 \leq M^2$. This implies $|\hat{f}_k| \lesssim Mk^{-s/2}$ (polynomial decay in frequency).

Truncation Error: Discarding scales $j > J$ introduces error:

$$E_J = \left\| \sum_{j>J} \sum_{i=1}^{N} c_{j,i} \psi_{j,i} \right\|_{L^2}^2 = \sum_{j>J} \|\text{projection onto scale } j\|_{L^2}^2.$$

By frame bounds and the Sobolev regularity, $E_J \lesssim M^2 \cdot (2^{-J})^s = M^2 J^{-s}$ (using $J \sim \log(\text{frequency cutoff})$).

Thus, $\|f - \hat{f}_J\|_{L^2} \leq \sqrt{E_J} \leq C_s M J^{-s}$, independent of dimension N. □

D Additional Experimental Details

D.1 Hyperparameter Tuning

All models were tuned on validation sets with the same budget of random seeds (5 seeds per hyperparameter configuration, selected by grid search). WaveletMPNN hyperparameters:

- *Learning rate:* $\eta \in \{1e^{-3}, 5e^{-4}, 1e^{-4}\}$
- *Hidden dimension:* $d \in \{32, 64, 128\}$
- *Number of scales:* $J \in \{2, 3, 4\}$
- *Chebyshev order:* $K \in \{3, 5, 7\}$
- *MPNN layers:* $L \in \{2, 3\}$

D.2 Statistical Significance

We report mean ± standard deviation over 5 runs. For main results (Tables 1, 2, 3), standard deviations are $\leq 1\%$*, indicating robust improvements.*