Higher Order Graph Attention Probabilistic Walk Networks

Anonymous Submission

Abstract

Graphs inherently capture dependencies between nodes or variables through their topological structure, with paths between any two nodes indicating a sequential dependency on the nodes traversed. Singe-hop Message Passing Neural Networks (MPNNs) leverage these latent relationships via local connections within the 1-hop neighbourhood, and have become widely adopted across diverse applications. Existing higher-order attention methods describe k-hop relationships by weighting variable-length paths. However, they struggle to fully capture global graph structure due to their restrictive sampling strategies, which constrain the exploration of higher-order topological relationships. To address these issues, we propose the Higher Order Graph Attention (HoGA) module, which assigns weights to variable-length paths sampled based on feature-vector diversity, effectively reconstructing the k-hop neighbourhood. HoGA represents higher-order relationships as a robust form of self-attention, applicable to any single-hop attention mechanism. In empirical studies, applying HoGA to existing attention-based models consistently leads to significant accuracy improvements on all benchmark node classification datasets. Furthermore, we observe that the performance degradation typically associated with additional message-passing steps may be mitigated. Code is available at https://anonymous.4open.science/r/Multi-Hop-D027/.

1 Introduction

Message Passing Neural Networks (MPNNs) leverage variable relationships encoded in graph topology [Scarselli *et al.*, 2009] to perform tasks on non-Euclidean domains. These networks have broad applications across fields such as molecular chemistry [Gilmer *et al.*, 2017], transport system planning [Jiang and Luo, 2022], social networks [Fan *et al.*, 2019], drug discovery [Xiong *et al.*, 2021], and climate modelling [Lam *et al.*, 2023], where atmospheric teleconnections span multiple spatial-temporal scales. However, early MPNN implementations [Veličković *et al.*, 2018; Kipf and Welling, 2016; Xu *et al.*, 2018; Gilmer *et al.*, 2017] relied solely on local information from the 1-hop neighbourhood of each node, thus requiring multiple propagation steps to capture long-range dependencies.

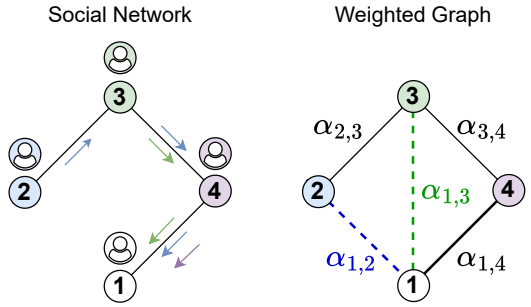

Figure 1: The HoGA learns long-distance relationships by self-attention weighting variable length paths. **Left:** shortest paths to the node of interest of length k are collected. **Right:** Paths between nodes i and j are assigned an attention weight α_{ij} based on the feature-vectors at their endpoints in the k-hop neighbourhood.

Excessive message compression, resulting from information bottlenecks [Topping *et al.*, 2021], leads to oversquashing [Di Giovanni *et al.*, 2023], limiting network depth [Wang *et al.*, 2020] and causing the loss of long-distance signals from k-hop neighbourhoods, whose diversity increases with k on real-world datasets [Ai *et al.*, 2024]. Multi-hop aggregation schemes [Huang *et al.*, 2024; Yang *et al.*, 2021a; Li *et al.*, 2021; Yang *et al.*, 2021b; Abboud *et al.*, 2022; Wang *et al.*, 2020; Abu-El-Haija *et al.*, 2019; Lee *et al.*, 2022] address this by aggregating feature vectors within a specified distance from a node, capturing higher-order dependencies in the graph's topology. These dependencies include the delayed influence of one node's message on another through paths longer than one hop. For example, walk-based methods consider the topology of higher-order paths [Yeh *et al.*, 2023; Li *et al.*, 2020; Michel *et al.*, 2023; Kong *et al.*, 2022], while other approaches aggregate the entire k-hop neighbourhood into a single feature vector [Abboud *et al.*, 2022].

Some existing k-order attention techniques compute similarity scores by sampling fixed topological substructures, such as k-simplexes [Huang *et al.*, 2024] or k-neighbourhoods [Zhang *et al.*, 2024]. However, a significant challenge in higher-order methods is the exponential growth

of k-hop neighbourhoods, making such approaches computationally intractable for higher k. As a result, these methods are typically restricted to low k values, often not exceeding two [Zhang *et al.*, 2024]. Additionally, predefining valid substructures restricts the search through the graph's topology, reducing the number of learnable higher-order relationships. Following Xu *et al.* [2018], other methods sample a fixed number of meta-paths from the k-hop neighbourhood of each node [Yang *et al.*, 2019; Yang *et al.*, 2021a]. Implicitly, however, this sampling strategy empathises overlap between k-hop neighbourhoods, and consequently cannot fully capture global graph structure within a fixed sampling limit, leading to rapid saturation in the efficacy of higher-order information. Furthermore, they suffer from a sequential sampling bias; in order to sample nodes at a distance of k, all intermediate nodes are necessarily included, which conditionally constrains samples when the nodal distance grows beyond two.

To alleviate these problems, we propose the Higher Order Graph Attention (HoGA) module, which aims to tractably capture global k-hop structure via taking shortest paths of length k on the graph (Figure 1). HoGA samples a subgraph of a k-order line graph, where edges connect nodes which are endpoints in a shortest path of length k in the original graph, by using a heuristic probabilistic walk. By exploring the graph feature space without topological bias, HoGA enables a tractable parameterisation of the k-hop neighbourhood. Furthermore, HoGA avoids feature smoothing, as described by Abboud *et al.* [2022], effectively capturing k-hop feature space modalities.

We empirically demonstrate that incorporating the Higher Order Graph Attention (HoGA) module into diverse attention models [Chamberlain *et al.*, 2021; Veličković *et al.*, 2018] significantly improves accuracy on benchmark node classification datasets [Yang *et al.*, 2016]. As a consequence, our attention module emerges as a simple and effective way to harness long-distance information. Additionally, we conduct extensive experiments on model hyperparameter sensitivity, including an ablation study on the number of layers and an analysis of the relationship between node sampling method and accuracy. Our contributions are:

- We propose HoGA, a graph attention module that relies on sampling the k-hop neighbourhood via a heuristic walk, which aims to maximise the diversity of samples, allowing for a high-fidelity estimation of the k-hop feature-space distribution.
- Our higher-order attention module leverages information from an arbitrary distance from the node of interest, thereby substantially mitigating the oversquashing phenomenon. We also show that oversmoothing is reduced, using accuracy as a proxy metric.
- HoGA extends any single-hop method to a k-hop setting; we demonstrate the effectiveness of HoGA on two diverse attention models; GAT and GRAND, showcasing our modules adaptability to single-hop paradigms.

2 Related Work

Message-passing regimes with empathises on strict locality suffer from several well-studied problems. We review the issues of oversmoothing, oversquashing and their provably limited expressive power [Morris *et al.*, 2019].

Expressivity. To enable MPNNs to generalise across graphs of arbitrary topology, parameterisation typically depends on feature vectors and specific topological properties, such as nodal degree [Kong *et al.*, 2022]. However, the expressivity of first-order message-passing is limited; since the update function is generally not injective, the set of non-isomorphic graphs that can be uniquely coloured by any 1-hop aggregation scheme is a strict subset of those distinguishable by the 1-WL test [Morris *et al.*, 2019]. Consequently, developing message-passing schemes that surpass the 1-WL test in expressivity has become a critical research area. [Zhang *et al.*, 2023] show that 1-hop aggregation fails to distinguish non-isomorphic graphs with bi-connected components. They prove that schemes aggregating colour information across all hops can differentiate such graphs, with their expressivity being bounded by the 3-WL test.

Oversquashing. One consequence of the finite hidden dimension buffer size is the gradual loss of long-distance information through message-passing. Analytically, oversquashing is expressed in terms of information bottleneck severity [Di Giovanni *et al.*, 2023], which on an associated manifold correspond to areas of negative curative [Topping *et al.*, 2021]. To this end, one focus of graph rewiring techniques is the creation [Gutteridge *et al.*, 2023; Arnaiz-Rodríguez *et al.*, 2022] or removal [Karhadkar *et al.*, 2022] of graph edges to improve information flow. However, rewiring does not explicitly preserve graph topology; superficial relationships are created while real ones are removed. While maintaining topology, k-hop aggregation reduces commute time between any two nodes in the graph by factor k.

Oversmoothing. Overlapping respective fields and other embedded positive-feedback structures, *e.g.* cliques and cycles, often cause the convergence of feature-vectors to a constant value, that is, the convergence of the Dirichlet energy function to zero [Rusch *et al.*, 2023]. In contrast, feature-vectors from nodes in the k-hop neighbourhood tend to increase in diversity with respect to hop distance. Wang *et al.* [2020] show that re-weighting the adjacency matrix based on multi-hop connectivity drastically reduces oversmoothing, and consequently improves performance.

Multi-Hop Aggregation Schemes. Single-hop attention schemes are well studied within the literature [Kipf and Welling, 2016; Xu *et al.*, 2018; Gilmer *et al.*, 2017]. In particular, Veličković et al. [2018] introduce the notion of self-attention on the edges of the graph. Recently, there has been much emphasis on generalising attention to paths of arbitrary length, viewing edges as paths of length 1 [Zhang *et al.*, 2024; Huang *et al.*, 2024]. Non-polynomial growth in the size of the k-hop neighbourhood, however, remains a key issue when aggregating higher order feature-vectors. Current methods, therefore, seek to utilise long-distance relationships while maintaining tractability. Abboud *et al.* [2022] map the entire k-hop neighbourhood onto a single feature-vector via some injective aggregation function. However, due to feature diversity within the k-hop neighbourhood, such an aggregation method removes important topological information, *e.g.* the tendency of nodes with similar feature-vectors to group in ho-

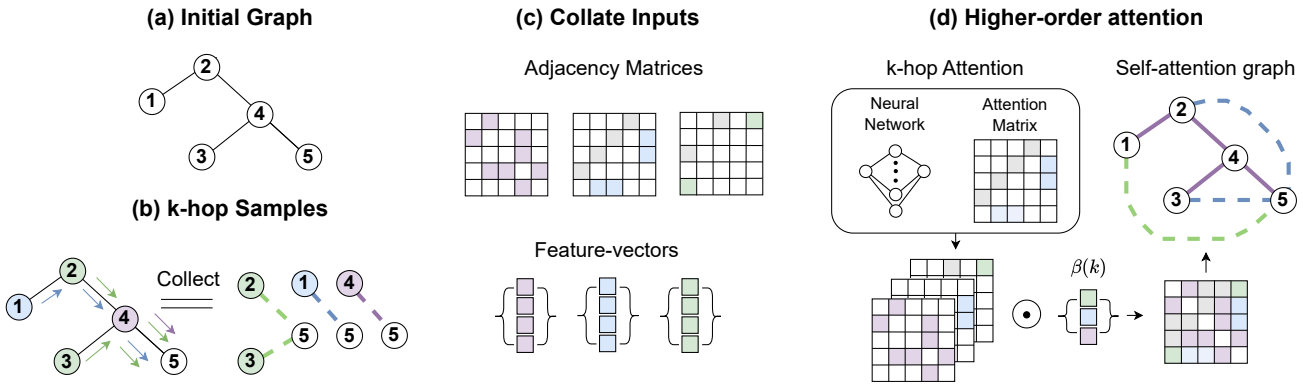

Figure 2: The Higher-Order Graphical Attention module. **(a)** an input graph of arbitrary topology. **(b)** Samples of the k-hop neighbourhood up to a maximum value of K are taken via a heuristic walk on the graph. **(c)** Sampling results are used to create a adjacency matrix describing connection via a shortest path of length k. **(d)** Higher order aggregation combines nodal information of variable distance, thus recreating the initial graph with self-attention edge weights.

mophilic graphs.

Walk-based approaches [Yeh *et al.*, 2023; Kong *et al.*, 2022; Li *et al.*, 2020; Michel *et al.*, 2023] sample the k-hop neighbourhood via constructing paths of length k, thereby avoiding intractability. Michel *et al.*[2023], for instance, aggregate samples from the set of shortest and simple paths between any two nodes onto a single feature vector. Similarly, Kong *et al.* [2022] employs a pooling layer on feature vectors from nodes along shortest paths, concatenating them with nodal degree information.

These methods provide a flexible multi-hop paradigm, reducing *a priori* topological constraints while avoiding the exponential growth of the k-hop neighbourhood via iterative sampling, effectively reconstructing the k-hop feature space. Despite this, no work has thus far focused on working directly with the set of k-hop neighbours, despite attractive theoretical guarantees [Zhang *et al.*, 2023]. We propose a graph attention module that constructs a direct parameterisation of the k-hop neighbourhood via iterative sampling and justify our approach through experimentation.

3 Preliminaries

A key component of Graph Neural Networks (GNN) is their ability to directly learn fixed ground truth topology over a non-Euclidean domain. Herein, we make the notion of local message-passing operators on graphs concrete, and give the formulation for the node classification task, a benchmark test conducted in GNN literature.

Message Passing Graph Neural Networks. Let $G = (V, E)$ be a fully connected undirected graph with vertices V and edges E. Intuitively, $i, j \in V$ are considered variables from some state space, while the topology described within E defines the inherent relationships between these variables. That is if $(i, j) \in E$, i and j are directly related. Alternatively, should a path $\mathcal{P} = (i_1, \ldots, i_k)$ of length k exist between $i_1 = i$ and $i_k = j$, we think of i and j to have a relation by casualty of the nodes in the path $i_1, \ldots, i_k \in V$. Furthermore, owing to the sequential dependence on k variables, the relationship between $\mathbf{x}_{i_k}$ and $\mathbf{x}_{i_1}$ is of order k. Correspondingly, each variable $i \in V$ is associated with a *feature-vector* $\mathbf{x}_i(t)$ at timestamp t.

Message Passing Neural Networks (MPNNs) are information aggregation schemes that concurrently aggregate nodal features across the whole graph. Notably, initial schemes, such as those used in [Kipf and Welling, 2016; Xu *et al.*, 2018; Hamilton *et al.*, 2017; Veličković *et al.*, 2018], focused on aggregation of node i's direct neighbours at a shortest path distance of 1 from i. The general form of any single-hop MPNN layer is given as:

$$\mathbf{x}_i(t+1) = \Psi_t\left(\mathbf{x}_i(t), \phi_t\left(\{\{\mathbf{x}_j(t) \mid j \in \mathcal{N}_1(i)\}\}\right)\right), \quad (1)$$

where we have denoted the multi-set as $\{\{\}\}$ and the nodes with shortest path length k from i as $\mathcal{N}_k(i)$. Here, ϕ_t is an injective function that aggregates close-proximity feature-vectors, thus allowing for a tractable update function Ψ_t to create a new feature representation. One choice of ψ_t, for example, is the summation operator $\sum_i \mathbf{x}_i(t)$.

Node Classification. Each $i \in V$ has an associated ground truth label y_i for a node classification task. In the semi-supervised setting, the MPNN classifier's role is to predict the classes of a subset $\mathcal{S} \subset V$, which corresponds to the training, testing or validation sets. Note that the union of all three does not necessarily contain all $i \in V$, and is often a small subset. To condition our network, we use the Cross Entropy Loss function, which for output logits $\hat{y}_i$ is given as:

$$\mathcal{L}(y, \hat{y}) = -\sum_{l=1}^{|S|} y_l \log(\hat{y}_l). \quad (2)$$

Single-hop Graphical Attention. In the single-hop setting, attention weights α_{ij} were strictly considered to be along edges $(i, j) \in E$ [Veličković *et al.*, 2018]. The corresponding attention matrix $\mathbf{A}(\mathbf{x}(t))$ is therefore only non-zero on edges of the graph:

$$\mathbf{A}(\mathbf{x}(t))_{i,j} = \begin{cases} \alpha_{i,j} & (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Here, $\alpha_{i,j}$ is a normalised attention coefficient calculated by

learning the parameters θ of the neural network $a_\theta(\cdot,\cdot)$:

$$\alpha_{i,j} = \frac{\exp\left(a_\theta(\mathbf{x}_i(t), \mathbf{x}_j(t))\right)}{\sum_{l\in\mathcal{N}_1(i)\cup\{i\}} \exp\left(a_\theta(\mathbf{x}_i(t), \mathbf{x}_l(t))\right)}. \quad (4)$$

4 Higher Order Graphical Attention

While successfully applying attention to local connections, strictly edge-wise attention does not explicitly consider long-distance relationships. Instead, it relies on the reciprocation of messages via operators acting on local connections. Message-passing that considers nodes of variable shortest path distance from each other possess attractive theoretical properties, such as expressivity that surpasses the 1-WL isomorphism test [Zhang *et al.*, 2023]. We formulate the Higher Order Attention (HoGA) module, which samples the k-hop neighbourhoods up to a maximal distance K around a node of interest, constructing an attention matrix that describes higher-order relationships within the graph. By capturing these relationships, our HoGA module improves the ability of MPNNs to recognise complex structural patterns, enhancing performance in tasks over graphical domain.

4.1 HoGA Formulation

HoGA directly considers the effect of node j on node i, where the shortest path distance $dist(.,.)$ satisfies $dist(i,j) = k$, via learning the impact of $\mathbf{x}_j(t)$ on $\mathbf{x}_i(t)$. We define the shortest path between i and j as $\mathcal{P} = (i = i_1, \ldots, j = i_k)$, and introduce a new attention coefficient, α_{ijk}, which describes their order k relationship, effectively weighting $\mathcal{P}$.

Our attention module only considers one such path, as we are concerned with the endpoint feature-vector $\mathbf{x}_j(t)$, as opposed to the remaining topological substructure of the graph described by $\mathcal{P}$. In this sense, our approach contrasts with walk-based methods for multi-hop feature extraction [Michel *et al.*, 2023; Kong *et al.*, 2022; Yeh *et al.*, 2023], which suffer from sequential sampling bias, and must by necessity obtain a substantial number of such $\mathcal{P}$ for the precise reason of capturing the full topology of the graph [Yeh *et al.*, 2023]. HoGA, however, achieves this via K parameterisations of $\mathcal{N}_k(i)$, for some k and any i, allowing for reduced bias to localised structures. In particular, following the strictly stronger expressivity of k-hop aggregation schemes than the 1-WL isomorphism test [Zhang *et al.*, 2023], our model applies the weight α_{ijk} to paths of length k varying such that $1 \leq k \leq K$.

To create a tractable parameterisation, we take a walk $\mathcal{S}_k \subset E_k$ on the k-order line graph $L^k(G) = (V_k, E_k)$ of G, where for any $(i,j) \in E_k$ we have that $dist(i,j) = k$ in G. Thereby, we induce a new adjacency matrix $\mathbf{A}_k(\mathbf{x}(t), \mathcal{S}_k)$ depending on the subset $\mathcal{S}_k$ for connectivity:

$$\mathbf{A}_k(\mathbf{x}(t), \mathcal{S}_k)_{i,j} = \begin{cases} \alpha_{ijk} & (i,j) \in \mathcal{S}_k, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

In this context, the attention coefficient α_{ijk} is computed via the neural network $a_{\theta_k}(\mathbf{x}_i(t), \mathbf{x}_j(t))$ with parameters θ_k. The final representation, shown in Figure 3, is given by $\mathbf{A}_{1:K}(\mathbf{x}(t))$, where we have dropped dependence on $\mathcal{S}_k$ for ease of notation:

$$\mathbf{A}_{1:K}(\mathbf{x}(t)) = \sum_{1\leq k\leq K} \beta(k)\mathbf{A}_k(\mathbf{x}(t), \mathcal{S}_k). \quad (6)$$

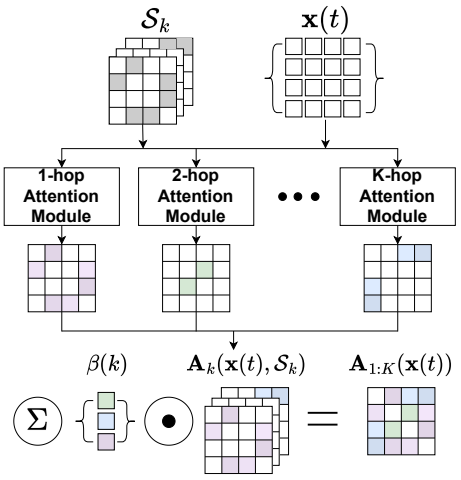

Figure 3: Our higher-order attention module aggregates weights from a single hop self-attention method by weighting contributions proportional to proximity.

Here $\mathcal{S}_1 = E$. The function $\beta : \mathbb{R} \mapsto \mathbb{R}$, first investigated in [Wang *et al.*, 2020], represents a weighting function that scales messages disproportionately to their commute times, simulating the actual transit times of the message while avoiding problems such as over-squashing. Introducing β reduces the risk of overfitting to the long-distance signals during training. For simplicity, we take $\beta(k)$ to be the harmonic series $\beta(k) = \frac{1}{k}$.

4.2 Sampling the k-hop Neighbourhood

We define the branching factor b of G as the average node degree. A key problem in any direct parameterisation of the shortest paths of length k is that the k-hop neighbourhood $|\mathcal{N}_k(i)|$ grows at order $O(b^k)$. Some works avoid this issue by simply aggregating the entirety of the periphery graph [Abboud *et al.*, 2022]. However, given the corresponding exponential growth in node diversity, such an approach also reduces the utility of $\mathcal{N}_k(i)$ when k increases, leading to a decrease in performance. A natural question arises: *how can a tractable parameterisation of $\mathcal{N}_k(i)$ be constructed which simultaneously respects the diversity of the feature-vectors and class labels?*

We propose a sampling method that acts to maximise the diversity of node feature-vectors in a subset $\mathcal{S} \subset \mathcal{N}_k(i)$. Our sampling methods furthermore ensures tractability by setting $|\mathcal{S}| = |E|$; we essentially require our estimation of the k-hop feature space to converge adequately within $|E|$ steps. Additionally, we have constrained that the necessary number of parameters $dim(\theta_k)$ grows linearly with the size of G. Indeed, for k hops, we have an asymptotic growth of order $O(k \cdot |E|)$ for the number of non-zero entries in $\mathbf{A}_{1:K}(\mathbf{x}(t))$.

Heuristic Probabilistic Walk.

We formulate a simple walk-based method that aims to progressively select $(i,j) \in E_k$, where the expected discrepancy between $\mathbf{x}_i = \mathbf{x}_i(0)$ and $\mathbf{x}_j = \mathbf{x}_j(0)$ is maximal. Given a history buffer $H = \{\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}\}$ of size n and a current

node of interest $i \in V$, we select a node j from candidate set $\mathcal{N}_k(i)$ with probability $p \sim s_n$, where s_n is the dissimilarity score between feature-vectors:

$$s_n = \gamma \cdot f(\mathbf{x}_i, \mathbf{x}_j) + (1-\gamma) \cdot f(\hat{\mathbf{x}}, \mathbf{x}_j), \tag{7}$$

$$f(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}, \tag{8}$$

$$\hat{\mathbf{x}} = \sum_{1 \le t \le n} \gamma^{n-t} \cdot \mathbf{x}_{i_t}. \tag{9}$$

Here, the cosine dissimilarity $f(\cdot, \cdot)$ measures the discrepancy between $\mathbf{x}_i$ and $\mathbf{x}_j$ in terms of their collinearity. The parameter $\gamma \in [0, 1]$ is the decay rate, and $\hat{\mathbf{x}}$ represents an exponential moving average over H. Intuitively, the first part of Equation 7 is a greedy step, as it tends to choose a $\mathbf{x}_j$ of maximal difference to $\mathbf{x}_i$, whereas the second part enforces a global dissimilarity for all visited nodes in H. The term γ acts as the decay rate in Equation 9, and balances the contributions between the greedy and history buffer steps.

Higher Order Attention Heads.
For any sampling procedure, our method enables the computation of multiple attention heads via effectively allowing $i \in V$ to resample with replacement from $\mathcal{N}_k(i)$. Specifically, given the subset $\mathcal{S}_k \subset E_k$ sampled from a distribution $P(\mathcal{S}_k)$, we define multi-head attention on the higher-order layer as the expected adjacency matrix over $P(\mathcal{S}_k)$:

$$\begin{aligned} \mathbf{A}_k(\mathbf{x}(t)) &= \mathbb{E}_{\mathcal{S}_k \sim P(\mathcal{S}_k)}[\mathbf{A}_k(\mathbf{x}(t), S_k)] \\ &\approx \frac{1}{|\Gamma|} \sum_{\mathcal{S}_k \in \Gamma} \mathbf{A}_k(\mathbf{x}(t), \mathcal{S}_k). \end{aligned} \tag{10}$$

Here, Γ acts as a super-set containing samples from $P(\mathcal{S}_k)$. In the next section, this definition is consistent with simply taking the expectation over feature-vectors $\mathbf{x}(t+1)$ for the models we consider. To preserve the correlation of feature-vectors across layers, we fix $\mathcal{S}_k$ with respect to t, allowing the network depth to extract feature-vectors at higher resolutions.

4.3 Extending Single-hop Attention Models

Attention forms the backbone of many graphical models [Veličković *et al.*, 2018; Chamberlain *et al.*, 2021; Choi *et al.*, 2023]. We demonstrate the versatility of HoGA by integrating the attention module into two distinct attention-based models. Specifically, in our work, we evaluate the application of the HoGA module to existing graphical models where attention is a fundamental component, explicitly focusing on GAT [Veličković *et al.*, 2018] and GRAND [Chamberlain *et al.*, 2021]. In general, as shown in Figure 3, we replace their single-hop adjacency matrix $\mathbf{A}(\mathbf{x}(t))$ with $\mathbf{A}_{1:K}(\mathbf{x}(t))$.

We now summarise these single-hop attention models. The GAT model computes self-attention weights for each edge $(i, j) \in E$ at layer t. Specifically, node i selects a subset of its neighbouring edges $j \in \mathcal{N}_1(i)$ by calculating the attention matrix using Equation 3:

$$\mathbf{x}(t+1) = \mathbf{A}(\mathbf{x}(t), t) \cdot \mathbf{x}(t). \tag{11}$$

In our HoGA-GAT model, we generalise $\mathbf{A}(\mathbf{x}(t), t)$ to $\mathbf{A}_{1:K}(\mathbf{x}(t), t)$, as defined in Equation 6. This introduces a multi-hop attention mechanism with an added dependency on the layer index t.

GRAND belongs to the neural flow family of models [Biloš *et al.*, 2021], which rely on the graph structure to describe message-passing as a physical process, *i.e.* one governed by a Partial Differential Equation (PDE). For the sake of comparison, we use the GRAND model with Laplacian attention; the parameters of the adjacency matrix are shared across all layers. The GRAND model is also given in terms of the attention matrix from Equation 3:

$$\frac{\partial \mathbf{x}(t)}{\partial t} = \mathbf{A}(\mathbf{x}(t)) \cdot \mathbf{x}(t). \tag{12}$$

Numerical methods, such as forward Euler, are used to solve Equation 12; the network layer with index t corresponds to a solution of Equation 12 at time t. In our HoGA-GRAND model, we again replace the attention matrix $\mathbf{A}(\mathbf{x}(t))$ with the analogous multi-hop formulation from Equation 6.

5 Experiments

We conduct empirical studies to address the following question: *is the application of higher-order attention via direct sampling of the k-hop neighbourhood a viable approach for multi-hop aggregation?* Our analysis is broken up into three components:

- **RQ1.** How effectively do our HoGA model mitigate the oversmoothing effect, and is the model's accuracy stable across different choices of maximum nodal distance?
- **RQ2.** What is the effectiveness of the heuristic walk sampling method compared to simpler, purely topological sampling methods?
- **RQ3.** How do our higher-order models, HoGA-GAT and HoGA-GRAND, perform compared to existing higher-order techniques?

Specifically, we demonstrate the efficacy of our method through comparisons with other state-of-the-art higher-order attention methods [Huang *et al.*, 2024; Abu-El-Haija *et al.*, 2019; Yang *et al.*, 2019]. Additionally, we include various other baselines which use Fourier methods [He *et al.*, 2021; Kipf and Welling, 2016; Defferrard *et al.*, 2016].

Datasets. We evaluate all models on core benchmark node classification datasets, wherein the dataset is comprised of a single graph: Cora, Citeseer, and Pubmed [Yang *et al.*, 2016]. We additionally evaluate other diverse datasets of variable size: Amazon Computers, Amazon Photos, and CoAuthor Computer Science [Shchur *et al.*, 2018].

Baseline Models. We evaluate models that incorporate our higher-order attention module and compare them with other non-local aggregation schemes, which utilise either meta-paths; SPAGAN [Yang *et al.*, 2019], or topological structure; HiGCN [Huang *et al.*, 2024]. Our study also includes single-hop spectral methods such as APPNP [Gutteridge *et al.*, 2023], BernNet [He *et al.*, 2021], and GCN [Kipf and Welling, 2016]. Additionally, we compare our higher-order attention models with their single-hop counterparts; GAT [Veličković *et al.*, 2018] and GRAND [Chamberlain *et al.*, 2021]. If available, we utilise the optimal parameter configurations as reported in the original studies.

5.1 Reproducibility

We cover herein various settings used throughout our empirical evaluations. The source code along with data splits and samples of the k-hop neighbourhood for all experiments is provided in the supplementary materials.

Data splits. On Cora, Citeseer and Pubmed, we use the public train, test and validation set splits proposed in the original paper [Yang *et al.*, 2016]. On the remaining datasets, we split the graph by randomly selecting nodes such that each set comprises 60%, 20% and 20% of all nodes, respectively.

Higher-order sampling. We set the random jump probability to 5% when running our heuristic walk algorithm, and limit the maximum number of edges obtained from any sampling procedure to 90,000 to reduce runtime. In the first network layer, we use eight higher-order attention heads, and one in the subsequent layers. To avoid excessive fine-tuning, we set the maximum k-hop value to $K = 3$ for all graphs, and keep the k-hop samples across consistent all layers.

Experiment setup. To reduce the variability in model performance due to random seed, we repeat all experiments 20 times, re-initialising our models with a new seed at each repetition. To train and evaluate our models, we run our experiments using an A100 GPU. We evaluate significance between empirical model performance by using the Wilcoxon signed rank test with a confidence threshold of 5%.

5.2 Higher Order Attention on Node Classification

We evaluate the efficacy of our HoGA module using node classification tasks as a proxy within our experimental setup. Our analysis compares accuracy with baseline models on benchmark datasets. We also perform qualitative assessments through ablation tests, focusing on the number of message-passing layers and the maximum hop value K.

Analysis on Oversmoothing (RQ1)

We use accuracy as a proxy metric, shown in Figure 4(b), to assess the degree of oversmoothing on node feature-vectors caused by additional message-passing steps. Given that the diversity of feature-vectors in the k-hop neighbourhood increases with k [Ai *et al.*, 2024], increasing the k-order of an aggregation method enhances access to descriptive information. The effect of positive-feedback structures in a small neighbourhood around the node of interest, such homophilic cliques, is therefore mitigated.

Since HoGA-GAT aims to harness a subset of maximally diverse feature-vectors from the k-hop neighbourhoods, we observe in Figure 4(b) a reduction in the degree of oversmoothing for Citeseer and Cora. Despite this mitigation, performance consistently declines with additional message-passing steps. This decline may be attributed to vanishing gradients arising from a combination of network depth and width [Hanin, 2018], or the formation of k-hop positive-feedback structures [Zhang *et al.*, 2024].

Testing Hop Number Stability (RQ1)

Figure 4(a) demonstrates the stability of HoGA-GAT across different maximum hop numbers K. We observe that on Citeseer, Cora, and Pubmed, accuracy improves with small K values, typically up to $K = 3$, after which the accuracy plateaus.

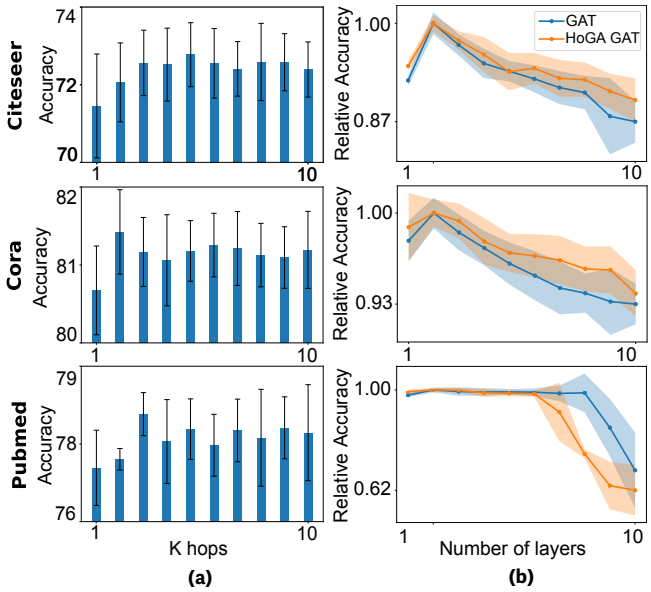

Figure 4: Ablation tests, with standard deviations across 20 iterations. **(a)** Accuracy as a function of the maximum hop value K. **(b)** Accuracy relative to the highest-performing model as a function of number of message-passing layers.

Since these graphs have a diameter close to 3, additional hops mainly encode redundant information, which is a repetition of an encoding at a lower value of k. Never the less, stability in K indicates that our model effectively retains higher-order information, allowing K to be set to the graph diameter.

Sampling Methods Comparison (RQ2)

We evaluate the relative utility of sampling via considering feature-vectors against using search methods that capture graph topology strictly on the basis of local connections. That is, we empirically assess the importance of optimising node diversity, and so compare our heuristic walk method with conventional, non-feature-vector-orientated methods: breadth and depth-first search, and uniform random walk and node selection.

Table 2 shows our evaluation of these methods. We observe that the topology-orientated baselines acquire lower accuracy across all datasets, that is, a decrease of least 2%, 3% and 1% on Cora, Citeseer and Pubmed, respectively. Indeed, the lack of inherent bias towards topological substructures, *e.g.* cliques organised via feature-vector similarity, leads to a less descriptive subset of the k-hop neighbourhood. The walk-based search methods capture a localised substructure of the graph where there is an absence in correlation between feature-vectors, making the consequent topology lacking in descriptivity. Similarity, Random does not describe any causal relationships via edges and paths between nodes. Our samples from our heuristic walk more closely describe the actual distribution of feature-vectors on the k-hop neighbourhood by considering local modalities of the feature-vector distribution.

Baselines	Cora	Citeseer	PubMed	Computers	Actor	Photo
			1st-Order Models			
ChebNet (2016)	79.8 ± 0.5	69.0 ± 1.0	77.9 ± 0.4	90.8 ± 0.5	51.3 ± 1.1	95.6 ± 1.0
GCN (2017)	81.8 ± 0.6	70.9 ± 0.6	78.3 ± 0.4	81.6 ± 4.4	53.5 ± 1.2	83.5 ± 1.4
GAT (2018)	81.6 ± 0.9	71.3 ± 0.8	77.0 ± 1.0	90.2 ± 3.7	40.6 ± 0.9	91.6 ± 4.4
JKNet (2018)	79.0 ± 1.3	66.9 ± 1.4	76.0 ± 0.8	87.4 ± 2.7	53.2 ± 0.8	93.0 ± 2.9
APPNP (2019)	82.2 ± 0.9	69.9 ± 0.8	78.2 ± 0.3	88.9 ± 0.9	45.6 ± 0.7	94.0 ± 0.4
GPRGNN (2020)	82.6 ± 0.9	69.4 ± 1.4	78.3 ± 0.5	87.9 ± 0.8	52.0 ± 0.7	93.1 ± 0.8
BernNet (2021)	73.2 ± 1.6	67.5 ± 1.5	73.3 ± 1.2	88.2 ± 0.5	51.9 ± 0.8	95.4 ± 0.8
GRAND (2021)	83.0 ± 1.0	70.2 ± 1.2	78.8 ± 0.8	85.2 ± 1.2	41.2 ± 0.9	95.5 ± 0.3
			Higher-Order Models			
MixHop (2019)	81.6 ± 0.8	70.4 ± 0.6	78.8 ± 0.8	91.3 ± 0.3	$\mathbf{59.2 \pm 0.5}$	92.1 ± 0.3
SPAGAN (2019)	82.2 ± 0.5	$\mathbf{72.4 \pm 0.7}$	77.9 ± 0.6	90.1 ± 0.3	31.3 ± 0.6	94.2 ± 0.3
HiGCN (2024)	$\mathbf{83.5 \pm 0.6}$	71.5 ± 1.0	$\mathbf{79.4 \pm 0.5}$	92.2 ± 0.6	48.8 ± 0.6	$\mathbf{96.6 \pm 0.2}$
HoGA-GAT (ours)	82.5 ± 0.7	$\mathbf{73.0 \pm 0.4}^{*}$	78.3 ± 0.4	$\mathbf{93.0 \pm 0.5}$	$\mathbf{60.6 \pm 1.6}^{*}$	96.3 ± 2.0
HoGA-GRAND (ours)	$\mathbf{83.3 \pm 0.5}$	71.1 ± 1.2	$\mathbf{79.5 \pm 0.8}$	$\mathbf{92.8 \pm 1.9}$	48.3 ± 1.0	$\mathbf{98.1 \pm 1.7}^{*}$

Table 1: Comparison between attention-based models using our higher-order module with baselines models. The best model is shown in bold green, while the second best is in bold blue. All HoGA models use the sampling method heuristic walk. The most significant model according to the Wilcoxon signed rank test, if it exists, is indicated by the subscript $*$.

Samplers	Cora	Citeseer	PubMed
Random Sample	81.4 ± 0.6	70.7 ± 0.5	77.6 ± 0.6
Random Walk	81.1 ± 0.8	70.5 ± 0.7	77.8 ± 0.5
Breadth First	81.6 ± 0.9	69.6 ± 0.8	76.8 ± 0.6
Depth First	81.3 ± 0.9	69.5 ± 0.8	77.0 ± 0.9
Heuristic Walk	$\mathbf{82.5 \pm 0.7}$	$\mathbf{73.0 \pm 0.4}$	$\mathbf{78.3 \pm 0.4}$

Table 2: Ablation study on k-hop neighbourhood sampling methods for the higher-order GAT model. The best results are highlighted in bold.

Evaluation on Benchmark Datasets (RQ3)

Table 1 shows the results from our experiments. Applying our attention module results in an accuracy increase, ranging from 1.5% on Pubmed to 20% on Actors, with a general improvement of about 3% across other benchmark datasets. Our results demonstrate that HoGA enhances access to higher-order information, which is long-distance node relationships, thereby improving accuracy compared to the original model. Compared to baselines using strictly local aggregation, both HoGA-GAT and HoGA-GRAND models achieve the highest accuracy by a significant margin, with the largest gains made on the Computers, Actor, and Photo datasets.

However, a notable difference in accuracy improvement is observed between small and large datasets. For instance, on the small Actor dataset, the HoGA-GRAND and HoGA-GAT models achieve substantial accuracy increases of approximately 7% and 20%, respectively. In contrast, on the larger Computers dataset, the gain is minimal, around 3%, highlighting a limitation of our module. This discrepancy arises because the modality of the feature-vector distribution on the k-hop neighbourhood grows proportionally with k. Given a fixed sampling limit, the heuristic walk cannot capture all modes of the distribution, leading to under representation. Consequently, the performance gain is limited.

Additionally, Table 1 compares HoGA-GAT and HoGA-GRAND to HiGCN, a state-of-the-art higher-order GAT model. For HoGA-GAT, we observe that while HiGCN achieves significantly higher accuracy on Cora and Pubmed, HoGA-GAT outperforms HiGCN by a significant margin on Citeseer, Computers, and Actor. This demonstrates that our higher-order attention paradigm, which assesses similarity via feature-vectors, offers improved performance over recent state-of-the-art higher-order attention methods [Huang *et al.*, 2024; Zhang *et al.*, 2024], which instead map topological substructures to a similarity score. We also compare our attention module to SPAGAN, a meta-path sampling method, which obtains significantly lower accuracy across all datasets. In particular, the Actor dataset results showcase the effect of sampling on a per-node basis for small graphs, the k-hop neighbourhoods of which contain a large degree of overlap.

6 Conclusion

We proposed the Higher Order Graphical Attention (HoGA) module, which extends existing forms of single-hop self-attention methods to a k-hop setting. The simplicity of our method allows for both ease of implementation and its wide-ranging applicability. In an empirical study, we show HoGA significantly increases accuracy on a node classification task across a range of benchmark datasets [Yang *et al.*, 2016; Shchur *et al.*, 2018] for both the GAT [Veličković *et al.*, 2018] and GRAND [Chamberlain *et al.*, 2021] attention-based models. In doing so, we demonstrate that direct sampling of the k-hop neighbourhood is a strong competitor to other higher-order methods [Wang *et al.*, 2020; Abboud *et al.*, 2022], including topological [Huang *et al.*, 2024; Zhang *et al.*, 2024] and meta-path approaches [Yang *et al.*, 2021a; Yang *et al.*, 2019], while extending existing walk-based methods [Kong *et al.*, 2022; Michel *et al.*, 2023] to this setting.

Future Directions. Our empirical study indicates that the efficacy of HoGA depends on the sampling method's efficiency in capturing modes of distribution within a limited number of sampling iterations. More effective, potentially feed-forward methods could enhance performance by improving modality compression, enabling faster convergence or more dynamic behaviour in the walk process.

References

[Abboud *et al.*, 2022] Ralph Abboud, Radoslav Dimitrov, and Ismail Ilkan Ceylan. Shortest path networks for graph property prediction. In *Learning on Graphs Conference*. PMLR, 2022.

[Abu-El-Haija *et al.*, 2019] Sami Abu-El-Haija, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, and Aram Galstyan. Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing. pages 21–29, 2019.

[Ai *et al.*, 2024] Guoguo Ai, Hui Yan, Huan Wang, and Xin Li. A2gcn: Graph convolutional networks with adaptive frequency and arbitrary order. *Pattern Recognition*, 156:110764, 2024.

[Arnaiz-Rodríguez *et al.*, 2022] Adrián Arnaiz-Rodríguez, Ahmed Begga, Francisco Escolano, and Nuria Oliver. Diffwire: Inductive graph rewiring via the lovász bound. *Proceedings of the First Learning on Graphs Conference*, 2022.

[Biloš *et al.*, 2021] Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. *Advances in Neural Information Processing Systems*, 34:21325–21337, 2021.

[Chamberlain *et al.*, 2021] Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. GRAND: Graph neural diffusion. In *International Conference on Machine Learning*, pages 1407–1418. PMLR, 2021.

[Choi *et al.*, 2023] Jeongwhan Choi, Seoyoung Hong, Noseong Park, and Sung-Bae Cho. Gread: Graph neural reaction-diffusion networks. In *Advances in Neural Information Processing Systems*, pages 5722–5747. PMLR, 2023.

[Defferrard *et al.*, 2016] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems*, 29, 2016.

[Di Giovanni *et al.*, 2023] Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio, and Michael M Bronstein. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International Conference on Machine Learning*, pages 7865–7885. PMLR, 2023.

[Fan *et al.*, 2019] Wenqi Fan, Yao Ma, Qing Li, Yixin He, Eric Zhao, and Jiliang Tang. Graph neural networks for social recommendation. In *Proceedings of the 28th International Conference on World Wide Web*, pages 417–426, 2019.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Neural Message Passing for Quantum Chemistry*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017.

[Gutteridge *et al.*, 2023] Benjamin Gutteridge, Xiaowen Dong, Michael M Bronstein, and Francesco Di Giovanni. Drew: Dynamically rewired message passing with delay. In *International Conference on Machine Learning*, pages 12252–12267. PMLR, 2023.

[Hamilton *et al.*, 2017] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

[Hanin, 2018] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in Neural Information Processing Systems*, 31, 2018.

[He *et al.*, 2021] Mingguo He, Zhewei Wei, Hongteng Xu, et al. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. *Advances in Neural Information Processing Systems*, 34:14239–14251, 2021.

[Huang *et al.*, 2024] Yiming Huang, Yujie Zeng, Qiang Wu, and Linyuan Lü. Higher-order graph convolutional network with flower-petals Laplacians on simplicial complexes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12653–12661, 2024.

[Jiang and Luo, 2022] Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, 2022.

[Karhadkar *et al.*, 2022] Kedar Karhadkar, Pradeep Kr Banerjee, and Guido Montúfar. Fosr: First-order spectral rewiring for addressing oversquashing in GNNs. *International Conference on Learning Representations*, 2022.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2016.

[Kong *et al.*, 2022] Lecheng Kong, Yixin Chen, and Muhan Zhang. Geodesic graph neural network for efficient graph representation learning. *Advances in Neural Information Processing Systems*, 35:5896–5909, 2022.

[Lam *et al.*, 2023] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[Lee *et al.*, 2022] See Hian Lee, Feng Ji, and Wee Peng Tay. SGAT: Simplicial graph attention network. 2022.

[Li *et al.*, 2020] Pan Li, Yanbang Wang, Hongwei Wang, and Jure Leskovec. Distance encoding: Design provably more powerful neural networks for graph representation learning. *Advances in Neural Information Processing Systems*, 33:4465–4478, 2020.

[Li *et al.*, 2021] Ronghan Li, Lifang Wang, Shengli Wang, and Zejun Jiang. Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension. In *International Joint Conference on Artificial Intelligence*, pages 3857–3863, 2021.

[Michel *et al.*, 2023] Gaspard Michel, Giannis Nikolentzos, Johannes F Lutzeyer, and Michalis Vazirgiannis. Path neural networks: Expressive and accurate graph neural networks. In *International Conference on Machine Learning*, pages 24737–24755. PMLR, 2023.

[Morris *et al.*, 2019] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4602–4609, 2019.

[Rusch *et al.*, 2023] T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

[Scarselli *et al.*, 2009] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[Shchur *et al.*, 2018] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *Advances in Neural Information Processing Systems*, 2018.

[Topping *et al.*, 2021] Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. *International Conference on Learning Representations*, 2021.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2020] Guangtao Wang, Rex Ying, Jing Huang, and Jure Leskovec. Multi-hop attention graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[Xiong *et al.*, 2021] Jiacheng Xiong, Zhaoping Xiong, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Graph neural networks for automated de novo drug design. *Drug discovery today*, 26(6):1382–1393, 2021.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *International Conference on Learning Representations*, 2018.

[Yang *et al.*, 2016] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International Conference on Machine Learning*, pages 40–48. PMLR, 2016.

[Yang *et al.*, 2019] Yiding Yang, Xinchao Wang, Mingli Song, Junsong Yuan, and Dacheng Tao. Spagan: Shortest path graph attention network. *International Conference on Joint Artificial Intelligence*, 2019.

[Yang *et al.*, 2021a] Liang Yang, Fan Wu, Zichen Zheng, Bingxin Niu, Junhua Gu, Chuan Wang, Xiaochun Cao, and Yuanfang Guo. Heterogeneous graph information bottleneck. In *International Conference on Joint Artificial Intelligence*, pages 1638–1645, 2021.

[Yang *et al.*, 2021b] Yiding Yang, Xinchao Wang, Mingli Song, Junsong Yuan, and Dacheng Tao. Spagan: Shortest path graph attention network. *International Joint Conference on Artificial Intelligence*, 2021.

[Yeh *et al.*, 2023] Pei-Kai Yeh, Hsi-Wen Chen, and Ming-Syan Chen. Random walk conformer: Learning graph representation from long and short range. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10936–10944, 2023.

[Zhang *et al.*, 2023] Bohang Zhang, Shengjie Luo, Liwei Wang, and Di He. Rethinking the expressive power of GNNs via graph biconnectivity. *International Conference on Learning Representations*, 2023.

[Zhang *et al.*, 2024] Heng-Kai Zhang, Yi-Ge Zhang, Zhi Zhou, and Yu-Feng Li. Hongat: Graph attention networks in the presence of high-order neighbors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16750–16758, 2024.