FiGURe: Simple and Efficient Unsupervised Node Representations with Filter Augmentations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised node representations learnt using contrastive learning-based methods have shown good performance on downstream tasks. However, these methods rely on augmentations that mimic low-pass filters, limiting their performance on tasks requiring different eigen-spectrum parts. This paper presents a simple filter-based augmentation method to capture different parts of the eigen-spectrum. We show significant improvements using these augmentations. Further, we show that sharing the same weights across these different filter augmentations is possible, reducing the computational load. In addition, previous works have shown that good performance on downstream tasks requires high dimensional representations. Working with high dimensions increases the computations, especially when multiple augmentations are involved. We mitigate this problem and recover good performance through lower dimensional embeddings using simple random Fourier feature projections. Our method, FiGURe, achieves an average gain of up to 4.4%, compared to the state-of-the-art unsupervised models, across all datasets in consideration, both homophilic and heterophilic.

1 Introduction

Contrastive learning is a powerful method for unsupervised graph representation learning, achieving notable success in various applications [35, 8]. However, these evaluations typically focus on tasks exhibiting homophily, where task labels strongly correlate with the graph's structure. An existing edge suggests the connected nodes likely share similar labels in these scenarios. However, these representations often struggle when dealing with heterophilic tasks, where edges tend to connect nodes with different labels.

Several papers [4, 10, 3, 20] have tackled the problem of heterophily by leveraging information from both low and high-frequency components. However, these methods operate in the semi-supervised setting, and the extension of these ideas in unsupervised learning still needs to be explored. Inspired by the insights in these papers, we propose a simple method incorporating these principles. Our approach introduces filter banks as additional views and learns separate representations for each filter bank. However, this approach faces two main challenges: Firstly, storing representations from each view can become prohibitively expensive for large graphs; secondly, contrastive learning methods typically demand high-dimensional representations, which increase both the computational cost of training and the storage burden.

We employ a shared encoder for all filter banks to tackle the first challenge. Our results confirm that a shared encoder performs on par with independent encoders for each filter bank. This strategy enables us to reconstruct filter-specific representations as needed, drastically reducing the storage requirement.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

For the second challenge, we train our models with low-dimensional embeddings. Then, we use random Fourier feature projection [31] to lift these low-dimensional embeddings into a higher-dimensional space. Kernel tricks [15] were typically used in classical machine learning to project low-dimensional representation to high dimensions where the labels can become linearly separable. However, constructing and leveraging the kernels in large dataset scenarios could be expensive. To avoid this issue, several papers [31, 32, 13, 28, 18] proposed to approximate the map associated with the kernel. For our scenario, we use the map associated with Gaussian kernel [31]. We empirically demonstrate that using such a simple approach preserves high performance for downstream tasks, even in the contrastive learning setting. Consequently, our solution offers a more efficient approach to unsupervised graph representation learning in computation and storage, especially concerning heterophilic tasks.

Our contributions in this work are, 1] We propose a simple scheme of using filter banks for learning representations that can cater to both heterophily and homophily tasks, 2] We address the computational and storage burden associated with this simple strategy by sharing the encoder across these various filter views, 3] By learning a low-dimensional representation and later projecting it to high dimensions using random Fourier Features, we further reduce the burden, 4] We study the performance of our approach on four homophilic and seven heterophilic datasets. Our method achieves new SOTA performance in unsupervised representation learning on heterophilic datasets, achieving 18% gains over prior methods, and in homophilic datasets [22], achieving 1.5% gains over prior methods. Our method also performs better than supervised methods such as GCN [17] on several heterophilic datasets and is competitive on homophilic datasets.

2 Related Work

Several unsupervised representation learning methods have been proposed in prior literature. Random walk-based methods like Node2Vec [7] and DeepWalk [29] preserve node proximity but tend to neglect structural information and node features. Contrastive methods, such as DEEP GRAPH INFOMAX (DGI) [35], maximize the mutual information (MI) between local and global representations while minimizing the MI between corrupted representations. Methods like MVGRL [8] and GRACE [36] expand on this, by integrating additional views into the MI maximization objective.

However, most of these methods focus on the low frequency components, overlooking critical insights from other parts. Semi-supervised methods like GPRGNN [4], BERNNET [10], and PPGNN [20] address this by exploring the entire eigenspectrum, but these concepts are yet to be applied in the unsupervised domain.

This work proposes the use of a filter bank to capture information across the full eigenspectrum while sharing an encoder across filters. Given the high-dimensional representation demand of contrastive learning methods, we propose using Random Fourier Features (RFF) to project lower-dimensional embeddings into higher-dimensional spaces, reducing computational load without sacrificing performance. The ensuing sections define our problem, describe filter banks and random feature maps, and explain our model and experimental results.

3 Problem Setting

In the domain of unsupervised representation learning, our focus lies on graph data, denoted as $G = (V; E)$, where V is the set of vertices and E the set of edges ($E \subseteq V \times V$). We associate an adjacency matrix with G, referred to as $\mathbf{A} : \mathbf{A} \in \{0; 1\}^{n \times n}$, where $n = |V|$ corresponds to the number of nodes. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the feature matrix. We use $\mathbf{A_I}$ to represent $\mathbf{A} + \mathbf{I}$ with $\mathbf{I}$ is the identity matrix, while $\mathbf{D_{A_I}}$ signifies the degree matrix of $\mathbf{A_I}$. We also define $\mathbf{A_n}$ as $\mathbf{D_{A_I}}^{-1/2}\mathbf{A_I}\mathbf{D_{A_I}}^{-1/2}$. No additional information is provided during training. The goal is to learn a parameterized encoder, $E : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times d} \mapsto \mathbb{R}^{n \times d'}$, where $d' \ll d$. This encoder produces a set of node representations $E(\mathbf{X}; \mathbf{A_n}) = \{h_1; h_2; \ldots; h_n\}$ where each $h_i \in \mathbb{R}^{d'}$ represents a rich representation for node i. The subsequent section will provide preliminary details about filter banks and random feature maps before we discuss the specifics of the proposed approach.

4 Preliminaries

Our proposed approach hinges on the critical components of filter banks and random feature maps. In this section, we delve into brief details about these two facets, setting the stage for a comprehensive description of our approach.

4.1 Filter Banks

Graph Fourier Transform (GFT) forms the basis of Graph Neural Networks (GNNs). A GFT is defined using a reference operator R which admits a spectral decomposition. Traditionally, in the case of GNNs, this reference operator has been the symmetric normalized laplacian $L_n = I - A_n$ or the A_n as simplified in [17]. A graph filter is an operator that acts independently on the entire eigenspace of a diagonalisable and symmetric reference operator R, by modulating their corresponding eigenvalues. [34, 33]. Thus, a graph filter H is defined via the graph filter function $g(\cdot)$ operating on the reference operator as $H = g(R) = U\, g(\Lambda)\, U^T$. Here, $\Lambda = \mathrm{diag}([\lambda_1, \lambda_2, \ldots, \lambda_n])$, where λ_i denotes the eigenvalues of the reference operator.

We describe a filter bank as a set of filters, denoted as $\mathcal{F} = \{F_1, F_2, \ldots, F_K\}$. Both GPRGNN [4] and BERNNET [10] employ filter banks, comprising polynomial filters, and amalgamate the representations from each filter bank to enhance the performance across heterophilic datasets. GPRGNN uses a filter bank defined as $\mathcal{F}_{\text{GPRGNN}} = \{I, A_n, \ldots, A_n^{K-1}\}$, while $\mathcal{F}_{\text{BERNNET}} = \{B_0, B_1, \ldots, B_{K-1}\}$ characterizes the filter bank utilized by BERNNET. Here, $B_i = \frac{1}{2^{K-1}} \binom{K-1}{i} (2I - L_n)^{K-i-1} (L_n)^i$.

Each filter in these banks highlights different parts of the eigenspectrum. By tuning the combination on downstream tasks, it offers the choice to select and leverage the right spectrum to enhance performance. Notably, unlike traditional GNNs, which primarily emphasize low-frequency components, higher frequency components have proved useful for heterophily [3, 4, 10, 20]. Consequently, a vital takeaway is that for comprehensive representations, we must aggregate information from different parts of the eigenspectrum and fine-tune it for specific downstream tasks.

4.2 Random Feature Maps for Kernel Approximations

Before the emergence of deep learning models, the kernel trick was instrumental in learning non-linear models. A kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, accepts two input features and returns a real-valued score. Given a positive-definite kernel, Mercer's Theorem [26] assures the existence of a feature map $\phi(\cdot)$, such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$. Leveraging the kernel trick, researchers combined Mercer's theorem with the representer theorem [15], enabling the construction of non-linear models that remain linear in k. These models created directly using k instead of the potentially complex ϕ, outperformed traditional linear models. The implicit maps linked with these kernels projected the features into a significantly high-dimensional space, where targets were presumed to be linearly separable. However, computational challenges arose when dealing with large datasets.

Addressing these issues, subsequent works [22, 13, 28, 31] introduced approximations of the map $\phi(\cdot)$ associated with individual kernels through random projections into higher-dimensional spaces ($\phi'(\cdot)$). This approach ensures that $\langle \phi'(x), \phi'(y) \rangle \approx k(x, y)$. These random feature maps are inexpensive to compute and affirm that simple projections to higher-dimensional spaces can achieve linear separability. The critical insight is that computationally efficient random feature maps exist, capable of projecting lower-dimensional representations into higher dimensions. These projections enhance the adaptability of these representations for downstream tasks. Random Fourier features (RFF) [31] provide a prime example of such techniques.

5 Proposed Approach

The following section delineates the process of unsupervised representation learning. Post that, we give details on how the representations learned from each filter bank is used in downstream tasks using random feature maps.

5.1 Unsupervised Representation Learning

Our method FiGURe (Filter-based Graph Unsupervised Representation Learning) builds on concepts introduced in [11, 35], extending the maximization of mutual information between node and global filter representations for each filter in the filter bank $\mathbb{F} = \{F_1, F_2, \ldots, F_K\}$. We construct an encoder for each filter to maximize the mutual information between the input data and encoder output. For the i^{th} filter, we learn an encoder $E_\phi : X_i \to X_i'$, denoted by learnable parameters ϕ. In this context, X_i represents a set of examples, where each example $[\tilde{x}_{ij}, \tilde{F}_{ij}] \in X_i$ consists of a filter F_i, its corresponding nodes and node features drawn from an empirical probability distribution P_i, which captures the joint distribution of features and node representations $[X, F_i]$. X_i defines the set of representations learnt by the encoder on utilizing feature information as well as topological information from the samples, sampled from the joint distribution P_i. The goal, aligned with [21, 11, 35], is to identify ϕ that maximizes mutual information between $[X, F_i]$ and $E_\phi(X, F_i)$, or $I_i([X, F_i], E_\phi(X, F_i))$. While exact mutual information (MI) computation is unfeasible due to unavailable exact data and learned representations distributions, we can estimate it using the Jensen-Shannon MI estimator [5, 25], defined as:

$$I^{JSD}_{i,\omega}([X, F_i], E_\phi(X, F_i)) := E_{P_i}[-sp(-T_\omega([\tilde{x}_{ij}, \tilde{F}_{ij}], E_\phi(\tilde{x}_{ij}, \tilde{F}_{ij})))] - E_{P_i \times P_i}[sp(T_\omega([\tilde{x}_{ij}, \tilde{F}_{ij}], E_\phi(\tilde{x}_{ij}, \tilde{F}_{ij})))] \quad (1)$$

Here, $T_\omega : X_i \times X_i' \to \mathbb{R}$ represents a discriminator function with learnable parameters ω. Note that $[\tilde{x}_{ij}, \tilde{F}_{ij}]$ is an input sampled from P_i, which is a marginal of the joint distribution of the input data and the learned node representations. The function $sp(.)$ corresponds to the softplus function [6]. Additionally, $T_{\omega,\phi} = D_w(R(E_\phi(\tilde{x}_{ij}, \tilde{F}_{ij})), E_\phi(\tilde{x}_{ij}, \tilde{F}_{ij}))$, where R denotes the readout function responsible for summarizing all node representations by aggregating and distilling information into a global filter representation.

In our approach, we first obtain node representations by feeding the filter-specific topology and associated node features into the encoder $H_i = E_\phi(X_i, F_i) = \{h_1^{F_i}, h_2^{F_i}, \ldots, h_n^{F_i}\}$. To obtain global representations, we employ a readout function $R : \mathbb{R}^{N \times d'} \to \mathbb{R}^{d'}$, which combines and distills information into a global representation $h_g^{F_i} = R(H_i) = R(E_\phi(X, F_i))$. Instead of directly maximizing the mutual information between the local and global representations, we introduce a learnable discriminator $D_\omega : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$, where $D_\omega(\cdot, \cdot)$ represents the joint probability score between the global representation and the node-specific patch representation. This joint probability score should be higher when considering global and local representations obtained from the same filter, as opposed to the joint probability score between the global representation from one filter and the local representation from an arbitrary filter.

Figure 1: Unsupervised learning of node embeddings by maximizing mutual information between node and graph representations over the graphs from the filter bank. Note that the parameter ϕ is shared across all the filters.

To generate negative samples for contrastive learning, we employ a corruption function $C : \mathbb{R}^{N \times d} \times \mathbb{R}^{N \times N} \to \mathbb{R}^{M \times d} \times \mathbb{R}^{M \times M}$, which yields corrupted samples denoted as $[\tilde{x}_{ij}, \tilde{F}_{ij}] = C(X, F_i)$. The designed corruption function generates data decorrelated with the input data.

In order to learn representations across all lters in the lter bank, we aim to maximise the average estimate of mutual information (MI) across all lters, consideringKlters.

$$I_F = \frac{1}{K}\sum_{i=1}^{K} I^{JSD}_{\omega_i}([X; F_i]; E_\theta(X; F_i)) \tag{2}$$

Maximising the Jenson-Shannon MI estimator is equivalent to reducing the binary cross entropy loss de ned between positive samples (sampled from the joint) and the negative samples (sampled from the product of marginals). Therefore, for each lter, we minimise the following objective:

$$L_{F_i} = \frac{1}{N+M}\left(\sum_{j=1}^{N} E_{(X;F_i)}[\log(D_{\omega_i}(h_j^{F_i}; h_g^{F_i}))] + \sum_{j=1}^{M} E_{(\tilde{X};\tilde{F_i})}[\log(D_{\omega_i}(\tilde{h}_j^{F_i}; h_g^{F_i}))]\right) \tag{3}$$

Therefore to learn meaningful representations across all lters the following objective is minimised:

$$L = \frac{1}{K}\sum_{i=1}^{K} L_{F_i} \tag{4}$$

However, managing the computational cost of training and storage for large graphs with separate node representations for each lter presents a signi cant challenge, exacerbated by the high dimensional requirements of contrastive learning methods. We implement parameter sharing to mitigate the rst issue, borrowing the concept from studies such as [4, 10], thereby sharing the encoder's parameters and the discriminator's parameters across all lters. Instead of storing dense lter-speci c node representations, we only store the parameters of the shared encoder and the rst-hop neighbourhood information of each node per lter, which has a lower storage cost. For downstream tasks, we retrieve the embeddings by reconstructing lter-speci c representations. To ensure quick and ef cient reconstruction, we use a simple one-layer GNN. This on-demand reconstruction of lter-speci c representations signi cantly reduces the computational and storage requirements associated with individual node representations. Fig 1 illustrates such a simple encoder's mutual information-based learning process.

Addressing the second issue, we initially train our models to generate low-dimensional embeddings. These encapsulate latent classes, as discussed in [2], as a superset of classes pertinent to downstream tasks. Although the low-dimensional embeddings harbour latent class information, they lack linear separability. Hence, we project these embeddings into a higher-dimensional space using random Fourier feature (RFF) projections, a strategy inspired by kernel methods (Section 4.2). Using this approach allows for improved linear separability of the latent classes. Our experimental ndings (Section 6.2) af rm the effectiveness of projecting lower-dimensional embeddings into higher dimensions, con rming the retention of latent class information in these embeddings.

5.2 Supervised Representation Learning

After obtaining representations for each lter post the reconstruction of the node representations, learning an aggregation mechanism to combine information from representations that capture different parts of the eigenspectrum for the given task is necessary. We adopt learning schemes proposed in [4, 10, 20], where we learn a weighted combination of lter-speci c representations. Therefore, the combined representations we learn for the downstream task are as follows (consideringKlters from the lter bankF):

$$Z = \sum_{i=1}^{K} \alpha_i \, \phi(E_\theta(X; F_i)) \tag{5}$$

The parametersα_i's are learnable. Additionally, the function$\phi(\cdot)$represents either the RFF projection or an identity transformation, depending on whether$E_\theta(X; F_i)$is low-dimensional or not. A classi er model (e.g. logistic regression) consumes these embeddings, where we train both theα_i's and the weights of the classi er. Fig 2 illustrates this process. The main distinction between semi-supervised methods such as [20, 4, 10] and our method is that the semi-supervised methods learn both the encoder and the combination coef cients based on labelled data. However, we pre-train the encoder in our method and subsequently learn a task-speci c combination of lter-speci c representations.

Figure 2: Supervised Learning: Using the trained parameters, we generate the node embeddings by encoding the filtered graphs that get consumed in the classification task.

6 Experimental Results

Training Details: We define a single-layer graph convolutional network (GCN) with shared weights $\Theta(\cdot)$ across all filters in the filter bank $\mathcal{F}(\cdot)$ as our encoder. Therefore, the encoder can be expressed as follows: $\mathcal{E}(\mathbf{X}, \mathbf{F}_i) = \sigma(\mathbf{F}_i \mathbf{X}\Theta)$. It is important to note that $\mathbf{F}_i$ represents a normalized filter with self-loops, which ensures that its eigenvalues are within the range of [0, 2]. The non-linearity function σ refers to the parametric rectified linear unit (PReLU) [9]. As we work with a single graph, we obtain the positive samples by sampling nodes from the graph. Using these sampled nodes, we construct a new adjacency list that only includes the edges between these sampled nodes in filter $\mathbf{F}_i$. On the other hand, the corruption function $\mathcal{C}(\cdot)$ operates on the same sampled nodes. However, it randomly shuffles the node features instead of perturbing the adjacency list. Similar to [35], we employ a straightforward readout function that involves averaging the representations across all nodes for a specific filter $\mathbf{F}_i$: $\mathcal{R}(\mathbf{H}_i) = \sigma\left(\frac{1}{N}\sum_{j=0}^{N} \mathbf{h}_j^{\mathbf{F}_i}\right)$ where σ denotes the sigmoid non-linearity. We utilize a bilinear scoring function, whose parameters are also shared across all filters:

$$\mathcal{D}_{\mathbf{W}}(\mathbf{h}_j^{\mathbf{F}_i}, \mathbf{h}_g^{\mathbf{F}_i}) = \sigma\left(\mathbf{h}_j^{\mathbf{F}_i T} \mathbf{W} \mathbf{h}_g^{\mathbf{F}_i}\right) \tag{6}$$

We learn the encoder and discriminator parameters by optimising Eq. 4. While we could use various filter banks, we specifically employ the filter bank corresponding to $\mathcal{G}_{\text{PRGNN}}(\mathbf{F}_{\text{GPRGNN}})$ for all our experiments. However, we also conduct an ablation study (see 6.5) to compare the performance when using $\mathbf{F}_{\text{GPRGNN}}$ versus $\mathbf{F}_{\text{BERNNET}}$. For more detailed training information, please refer to the supplementary material.

We conducted a series of comprehensive experiments to evaluate the effectiveness and competitiveness of our proposed model compared to SOTA models and methods. These experiments address the following research questions: [RQ1] How does FIGURe, perform compared to SOTA unsupervised models? [RQ2] Can we perform satisfactorily even with lower dimensional representations using projections such as RFF? [RQ3] Does shared encoder decrease performance? [RQ4] What is the computational efficiency gained by using lower dimensional representations compared to methods that rely on higher dimensional representations? [RQ5] Can alternative filter banks be employed to recover good quality representations?

Datasets and Setup: We evaluated our model on a diverse set of real-world datasets, which include both heterophilic and homophilic networks, to assess its effectiveness. Similar to previous works, we utilized the node classification task as a proxy to evaluate the quality of the learned representations. Please refer to the supplementary material for detailed information about the benchmark datasets.

The heterophilic datasets used in our evaluation include CHAMELEON, SQUIRREL, ROMAN-EMPIRE, and MINESWEEPER. For CHAMELEON and SQUIRREL, we adopted the ten random splits (with 48%, 32%, and 20% of nodes allocated for the train, validation, and test sets, respectively) from [27]. For ROMAN-EMPIRE and MINESWEEPER, we used the ten random splits provided in [30]. Additionally, we evaluated our model on four homophilic datasets: CORA, CITESEER, and PUBMED, as borrowed from [14]. We report the mean and standard deviation of the test accuracy across different splits. Please refer to the supplementary material for detailed statistics of each dataset.

Baselines: In our comparison against baselines, we considered common unsupervised approaches, such as DEEPWALK and NODE2VEC, and state-of-the-art mutual information-based methods, namely DGI, MVGRL, GRACE, and SUGRL. We also include the performance numbers of the widely used GCN for reference. It is important to note that unless explicitly mentioned, we set the representation size to 512 dimensions for all reported results, consistent with previous work. Please refer to the supplementary material for detailed comparisons with other supervised methods and the link to our codebase.

6.1 RQ1: FiGURe versus SOTA Methods

Table 1: Contains node classification accuracy percentages on homophilic and heterophilic datasets. FiGURe$_{32}$ and FiGURe$_{128}$ refer to FiGURe trained with 32 and 128 dimensional representations, respectively, and then projected using RFF. The remaining models are trained at 512 dimensions. Higher numbers indicate better performance. It is worth noting that FiGURe achieves superior performance or remains competitive with the baseline methods in all cases. The rightmost column Av. $_{\text{gain}}$ represents the average accuracy% gain of FiGURe over the model in that row, averaged across the different datasets. Blue, Red and Green represent 1st, 2nd and 3rd best performing models, for a particular dataset.

	HETEROPHILIC DATASETS				HOMOPHILIC DATASETS			
	SQUIRREL	CHAMELEON	ROMAN-EMPIRE	MINESWEEPER	CORA	CITESEER	PUBMED	Av. $_{\text{gain}}$
DEEPWALK	38.66 (1.44)	53.42 (1.73)	13.08 (0.59)	79.96 (0.08)	83.64 (1.85)	63.66 (3.36)	80.85 (0.44)	16.35
NODE2VEC	42.60 (1.15)	54.23 (2.30)	12.12 (0.30)	80.00 (0.00)	78.19 (1.14)	57.45 (6.44)	73.24 (0.59)	18.56
DGI	39.61 (1.81)	59.28 (1.23)	47.54 (0.76)	82.51 (0.47)	84.57 (1.22)	73.96 (1.61)	86.57 (0.52)	2.67
MVGRL	39.90 (1.39)	54.61 (2.29)	68.50 (0.38)	85.60 (0.35)	86.22 (1.30)	75.02 (1.72)	87.12 (0.35)	4.39
GRACE	53.15 (1.10)	68.25 (1.77)	47.83 (0.53)	80.22 (0.45)	84.79 (1.51)	67.60 (2.01)	87.04 (0.45)	5.54
SUGRL	43.13 (1.36)	58.60 (2.04)	39.40 (0.49)	82.40 (0.58)	81.21 (2.07)	67.50 (1.62)	86.90 (0.50)	9.80
FiGURe$_{32}$	48.89 (1.55)	65.66 (2.52)	67.67 (0.77)	85.28 (0.71)	82.56 (0.87)	71.25 (2.20)	84.18 (0.53)	3.18
FiGURe$_{128}$	48.78 (2.48)	66.03 (2.19)	68.10 (1.09)	85.16 (0.58)	86.14 (1.13)	73.34 (1.91)	85.41 (0.52)	2.11
FiGURe	52.23 (1.19)	68.55 (1.87)	70.99 (0.52)	85.58 (0.49)	87.00 (1.24)	74.77 (2.00)	88.60 (0.44)	0.00

Table 2: Comparison of Node classification accuracy percentages with the widely used supervised model GCN. Despite not having access to task specific labels, FiGURe learns good quality representations.

	SQUIRREL	CHAMELEON	ROMAN-EMPIRE	MINESWEEPER	CORA	CITESEER	PUBMED
GCN	47.78 (2.13)	61.43 (2.70)	73.69 (0.74)	89.75 (0.52)	87.36 (0.91)	76.47 (1.34)	88.41 (0.46)
FiGURe	52.23 (1.19)	68.55 (1.87)	70.99 (0.52)	85.58 (0.49)	87.00 (1.24)	74.77 (2.00)	88.60 (0.44)

We analyzed the results in Table 1 and made important observations. Across homophilic and heterophilic datasets, FiGURe consistently outperforms several SOTA unsupervised models, except in a few cases where it achieves comparable performance. We want to emphasize the rightmost column of the table, which shows the average percentage gain in performance across all datasets. This metric compares the improvement that FiGURe provides over each baseline model for each dataset and averages these improvements. This metric highlights the performance consistency of FiGURe across diverse datasets. No other baseline model achieves the same consistent performance across all datasets as FiGURe. Even the recent state-of-the-art contrastive models GRACE and SUGRL experience average performance drops of approximately 5% and 10%, respectively. This result indicates that FiGURe learns representations that exhibit high generalization and task-agnostic capabilities. Another important observation is the effectiveness of RFF projections in improving lower dimensional representations. We compare FiGURe at different dimensions, including FiGURe$_{32}$ and FiGURe$_{128}$, corresponding to learning 32 and 128-dimensional embeddings, respectively, in addition to the baseline representation size of 512 dimensions. Remarkably, even at lower dimensions, FiGURe with RFF projections demonstrates competitive performance across datasets, surpassing the 512-dimensional baselines in several cases. This result highlights the effectiveness of RFF projections in enhancing the quality of lower dimensional representations. Section 6.2 discusses more insights about the effectiveness of RFF projections. Furthermore, we include the widely used supervised model, GCN, in Table 2 as a benchmark for comparison. Notably, FiGURe outperforms GCN on heterophilic datasets, except for ROMAN-EMPIRE and MINESWEEPER, while achieving competitive performance on homophilic datasets. Please refer to supplementary material for detailed comparisons with supervised methods.

6.2 RQ2: RFF Projections on Lower Dimensional Representations

Table 3: Node classification accuracy percentages with and without using Random Fourier Feature projections (on 32 dimensions). A higher number means better performance. The performance is improved by using RFF in almost all cases, indicating the usefulness of this transformation

	RFF	CORA	CITESEER	SQUIRREL	CHAMELEON
DGI		81.65 (1.90)	65.62 (2.39)	31.60 (2.19)	45.48 (3.02)
	X	81.49 (1.96)	66.50 (2.44)	38.19 (1.52)	56.01 (2.66)
MVGRL		78.81 (1.73)	70.36 (1.76)	29.58 (0.94)	46.56 (2.84)
	X	80.14(2.41)	70.57(1.56)	37.83(1.32)	55.57 (2.28)
SUGRL		65.35 (2.41)	42.84 (2.57)	31.62 (1.47)	43.20 (1.79)
	X	70.06(1.24)	47.03(3.02)	38.50 (2.19)	51.01 (2.26)
GRACE		76.84 (1.09)	58.40 (3.05)	38.20 (1.38)	53.25 (1.58)
	X	79.15 (1.44)	63.66 (2.96)	51.56 (1.39)	67.39 (2.23)
FiGURe		82.88 (1.42)	70.32 (1.98)	39.38 (1.35)	53.27 (2.40)
	X	82.56 (0.87)	71.25 (2.20)	48.89 (1.55)	65.66 (2.52)

In this section, we analyse the performance of unsupervised baselines using 32-dimensional embeddings with and without RFF projections (see Table 3). Despite extensive hyperparameter tuning, we could not replicate the results reported by SUGRL, so we present the best results we obtained. Two noteworthy observations emerge from these tables. Firstly, it is evident that lower dimensional embeddings can yield meaningful and linearly separable representations when combined with simple RFF projections. Utilising RFF projections enhances performance in almost all cases, highlighting the value captured by MI-based methods even with lower-dimensional embeddings. Secondly, FiGURe consistently achieves superior or comparable performance to the baselines, even in lower dimensions. Notably, this includes SUGRL, purported to excel in such settings. However, there is a 2-3% performance gap between GRACE and our method for the SQUIRREL and CHAMELEON datasets. While GRACE handles heterophily well at lower dimensions, its performance deteriorates with homophilic graphs, unlike FiGURe which captures lower frequency information effectively. Additionally, our method exhibits computational efficiency advantages for specific datasets in lower dimensions. Please refer to the supplementary material for more details. Overall, these findings highlight the potential of RFF projections in extracting useful information from lower dimensional embeddings and reaffirm the competitiveness of FiGURe over the baselines.

6.3 RQ3: Sharing Weights Across Filter Specific Encoders

Table 4: A comparison of the performance on the downstream node classification task using independently trained encoders and weight sharing across encoders is shown. The reported metric is accuracy. In both cases, the embeddings are combined using the method described in 5.2

	CORA	CITESEER	SQUIRREL	CHAMELEON
INDEPENDENT	86.92 (1.10)%	75.03 (1.75)%	50.52 (1.51)%	66.86 (1.85)%
SHARED	87.00 (1.24)%	74.77 (2.00)%	52.23 (1.19)%	68.55 (1.87)%

Our method proposes to reduce the computational load by sharing the encoder weights across all filters. It stands to reason whether sharing these weights causes any degradation in performance. We present the results with shared and independent encoders across the filters in Table 4 to verify this. The findings indicate no significant decrease in performance when using shared weights, and in some cases, it even leads to improvements, validating the use of shared encoders.

6.4 RQ4: Computational Efficiency

To assess the computational efficiency of the different methods, we analyzed the computation time and summarized the results in Table 5. The key metric used in this analysis is the mean epoch time: the average time taken to complete one epoch of training. We compared our method with both MI-based methods such as DGI and MVGRL. Due to the increase in the number of augmentation views,

Table 5: Mean epoch time (in milliseconds) averaged across 20 trials with different hyperparameters. A lower number means the method is faster. Even though our method is slower at 512 dimensions, using 128 and 32 dimensional embeddings signi cantly reduces the mean epoch time. Using RFF as described in 6.2 we are able to prevent the performance drops experienced by DGI and MVGRL.

	DGI	MVGRL	FiGURe	FiGURe$_{128}$	FiGURe$_{32}$
CORA	38.53 (0.77)	75.29 (0.56)	114.38 (0.51)	20.10 (0.46)	11.54 (0.34)
CITESEER	52.98 (1.15)	102.41 (0.99)	156.24 (0.56)	30.30 (0.60)	17.16 (0.51)
SQUIRREL	87.06 (2.07)	168.24 (2.08)	257.65 (0.76)	47.72 (1.40)	23.52 (1.14)
CHAMELEON	33.08 (0.49)	64.71 (1.05)	98.36 (0.64)	18.56 (0.39)	11.63 (0.48)

there is an expected increase in computation time from DGI to MVGRL to FiGURe. However, as demonstrated in 6.2, using RFF projections allows us to achieve competitive performance even at lower dimensions. Therefore, we also included comparisons with our method at 128 and 32 dimensions in the table. It is evident from the results that our method, both at 128 and 32 dimensions, exhibits faster computation times compared to both DGI and MVGRL, which rely on higher-dimensional representations to achieve good performance. This result indicates that FiGURe is computationally ef cient due to its ability to work with lower-dimensional representations. During training, our method FiGURe$_{32}$, is 3x faster than DGI and 6x times faster than MVGRL. Despite the faster computation, FiGURe$_{32}$ also exhibits an average performance improvement of around 2% across the datasets over all methods considered in our experiments. Please refer to the supplementary material for additional comparisons to other unsupervised models.

6.5 RQ5: Experiments on Other Filter Banks

Table 6: Accuracy percentage results using other lter banks for FiGURe. F^{3}_{BernNet} refers to the F_{BernNet} lter bank (Section 4.1) with K set to 3 and F^{11}_{BernNet} refers to K set to 11.

	CORA	CITESEER	SQUIRREL	CHAMELEON
F^{3}_{BernNet}	85.13 (1.26)	73.38 (1.81)	37.07 (1.29)	53.95 (2.78)
F^{11}_{BernNet}	86.62 (1.59)	73.97 (1.43)	43.48 (3.80)	62.13 (3.66)
F_{GPRGNN}	87.00 (1.24)	74.77 (2.00)	52.23 (1.19)	68.55 (1.87)

To showcase the versatility of our proposed framework, we conducted an experiment using Bernstein lters, as detailed in Table 6. The results indicate that using F_{GPRGNN} leads to better performance than Bernstein lters. We believe that the reason this is happening is due to the latent characteristics of the dataset. [10, 20] have shown that datasets like CHAMELEON and SQUIRREL need frequency response functions that give more prominence to the tail-end spectrum. F_{GPRGNN} are more amenable to these needs, as demonstrated in [20]. However, datasets requiring frequency response similar to comb lters may be better approximated by F_{BernNet} as their basis gives uniform prominence on the entire spectrum. Please refer to the supplementary material, which shows the basis frequency responses of these two lter banks, with more clari cation. Therefore, although F_{GPRGNN} gives better performance for these datasets, there could be datasets where F_{BernNet} could do better. Hence, we proposed a general framework that can work with any lter bank.

7 Conclusion and Future Work

Our work demonstrates the bene ts of enhancing contrastive learning methods with lter views and learning lter-speci c representations to cater to diverse tasks from homophily to heterophily. We have effectively alleviated computational and storage burdens by sharing the encoder across these lters and focusing on low-dimensional embeddings that utilize high-dimensional projections, a technique inspired by random feature maps developed for kernel approximations. Future directions include extending the analysis in [2] to graph contrastive learning and explicitly exploring the linear separability in low dimensions. This analysis could solidify the connection with the proposed random feature maps approach.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In International conference on knowledge discovery & data mining (KDD), pages 2623–2631, 2019.

[2] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In International Conference on Machine Learning (ICML), 2019.

[3] D. Bo, X. Wang, C. Shi, and H.-W. Shen. Beyond low-frequency information in graph convolutional networks. In Association for the Advancement of Arti cial Intelligence (AAAI), 2021.

[4] E. Chien, J. Peng, P. Li, and O. Milenkovic. Adaptive universal generalized pagerank graph neural network. In International Conference on Learning Representations (ICLR), 2021.

[5] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. I. Communications on Pure and Applied Mathematics, 1975.

[6] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. In Neural Information Processing Systems (NeurIPS), 2000.

[7] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In International Conference on Knowledge Discovery and Data Mining (KDD), 2016.

[8] K. Hassani and A. H. Khasahmadi. Contrastive multi-view representation learning on graphs. In International Conference on Machine Learning (ICML), 2020.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into recti ers: Surpassing human-level performance on imagenet classi cation. 2015 IEEE International Conference on Computer Vision (ICCV) 2015.

[10] M. He, Z. Wei, Z. Huang, and H. Xu. Bernnet: Learning arbitrary graph spectral lters via bernstein approximation. In Neural Information Processing Systems (NeurIPS), 2021.

[11] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In International Conference on Learning Representations (ICLR), 2019.

[12] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In Neural Information Processing Systems (NeurIPS), 2020.

[13] P. Kar and H. Karnick. Random feature maps for dot product kernels. In International Conference on Arti cial Intelligence and Statistics (AISTATS), 2012.

[14] D. Kim and A. Oh. How to nd your friendly neighborhood: Graph attention design with self-supervision. In International Conference on Learning Representations (ICLR), 2021.

[15] G. Kimeldorf and G. Wahba. Some results on tchebycheff an spline functions. Journal of Mathematical Analysis and Applications, 1971.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

[17] T. N. Kipf and M. Welling. Semi-supervised classi cation with graph convolutional networks. In International Conference on Learning Representations (ICLR), 2017.

[18] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic. Towards a uni ed analysis of random fourier features. Journal of Machine Learning Research (JMLR), 2021.

[19] D. Lim, F. M. Hohne, X. Li, S. L. Huang, V. Gupta, O. P. Bhalerao, and S.-N. Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In Neural Information Processing Systems (NeurIPS), 2021.

[20] V. Lingam, C. Ekbote, M. Sharma, R. Ragesh, A. Iyer, and S. Sellamanickam. A piece-wise polynomial lter ing approach for graph neural networks. In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2022.

[21] R. Linsker. Self-organization in a perceptual network. Computer, 1988.

[22] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001.

[23] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 1909.

[24] Y. Mo, L. Peng, J. Xu, X. Shi, and X. Zhu. Simple unsupervised graph representation learning. In Association for the Advancement of Arti cial Intelligence (AAAI), 2022.

[25] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In Neural Information Processing Systems (NeurIPS) 2016.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In Neural Information Processing Systems (NeurIPS), 2019.

[27] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang. Geom-gcn: Geometric graph convolutional networks. In International Conference on Learning Representations (ICLR), 2020.

[28] J. Pennington, F. X. X. Yu, and S. Kumar. Spherical random features for polynomial kernels. In Neural Information Processing Systems (NeurIPS), 2015.

[29] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk. In International Conference on Knowledge Discovery and Data Mining (KDD), 2014.

[30] O. Platonov, D. Kuznedelev, M. Diskin, A. Babenko, and L. Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In International Conference on Learning Representations (ICLR), 2023.

[31] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In Neural Information Processing Systems (NeurIPS), 2007.

[32] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In Neural Information Processing Systems (NeurIPS), 2008.

[33] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging eld of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Processing Magazine, 2013.

[34] N. Tremblay, P. Gonçalves, and P. Borgnat. Design of graph lters and lterbanks. Cooperative and Graph Signal Processing, 2017.

[35] P. Velicković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. Deep graph infomax. In International Conference on Learning Representations (ICLR), 2019.

[36] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang. Deep Graph Contrastive Representation Learning. In ICML Workshop on Graph Representation Learning and Beyond, 2020.

8 Supplementary Material

Contents

1 Introduction 1

2 Related Work 2

3 Problem Setting 2

4 Preliminaries 3
- 4.1 Filter Banks . . . 3
- 4.2 Random Feature Maps for Kernel Approximations . . . 3

5 Proposed Approach 3
- 5.1 Unsupervised Representation Learning . . . 4
- 5.2 Supervised Representation Learning . . . 5

6 Experimental Results 6
- 6.1 RQ1: FiGURe versus SOTA Methods . . . 7
- 6.2 RQ2: RFF Projections on Lower Dimensional Representations . . . 8
- 6.3 RQ3: Sharing Weights Across Filter Speci c Encoders . . . 8
- 6.4 RQ4: Computational Ef ciency . . . 8
- 6.5 RQ5: Experiments on Other Filter Banks . . . 9

7 Conclusion and Future Work 9

8 Supplementary Material 12
- 8.1 Reproducibility . . . 12
- 8.2 Datasets . . . 13
- 8.3 Training Details . . . 13
  - 8.3.1 Unsupervised Training . . . 14
  - 8.3.2 Supervised Training . . . 14
  - 8.3.3 Negative Sampling for the Identity Filter . . . 15
- 8.4 Evaluation on large graphs . . . 15
- 8.5 Comparison with other Supervised Methods . . . 15
- 8.6 RFF Projections . . . 16
- 8.7 Computational Comparisons with other Other Unsupervised Methods . . . 16
- 8.8 Choice of Filter Banks . . . 17
- 8.9 Visualising RFF Behavior and Community Structure . . . 18

8.1 Reproducibility

We strive to ensure the reproducibility of our research ndings. To facilitate this, we provide the details of our experimental setup, including dataset sources, preprocessing steps, hyperparameters,

and model con gurations. We also make our code and the datasets used, publicly available at this LINK, enabling researchers to reproduce our results and build upon our work. We would like to emphasize that our code is built on top of the existing GRL codebase. For the datasets used in our evaluation, we provide references to their original sources and any speci c data splits that we employed. This allows others to obtain the same datasets and perform their own analyses using consistent data. Additionally, we specify the versions of libraries and frameworks used in our experiments, in Section 8.3, and in the REQUIREMENTS le and the README le, in the codebase, enabling others to set up a compatible environment. We document any speci c seed values or randomization procedures that may affect the results. By providing these details and resources, we aim to promote transparency and reproducibility in scienti c research. We encourage fellow researchers to reach out to us if they have any questions or need further clari cation on our methods or results.

8.2 Datasets

Homophilic Datasets: We evaluated our model (as well as baselines) on three homophilic datasets: CORA, CITESEER, and PUBMED as borrowed from [14]. All three are citation networks, where each node represents a research paper and the links represent citations. Pubmed consists of medical research papers. The task is to predict the category of the research paper. We follow the same dataset setup mentioned in [14] to create 10 random splits for each of these datasets.

Heterophilic Datasets: In our evaluation, we included four heterophilic datasets: CHAMELEON, SQUIRREL, ROMAN-EMPIRE, and MINESWEEPER. For CHAMELEON and SQUIRREL, nodes represent Wikipedia web pages and edges capture mutual links between pages. We utilized the ten random splits provided in [27], where 48%, 32%, and 20% of the nodes were allocated for the train, validation, and test sets, respectively. In ROMAN-EMPIRE each node corresponds to a word in the Roman Empire Wikipedia article. Two words are connected with an edge if either these words follow each other in the text, or they are connected in the dependency tree of the sentence. The syntactic role of the word/node de nes its class label. The MINESWEEPER graph is a regular 100x100 grid where each node is connected to eight neighboring nodes, and the features are on-hot encoded representations of the number of neighboring mines. The task is to predict which nodes are mines. For both ROMAN-EMPIRE and MINESWEEPER, we used the ten random splits provided in [30].

Large Datasets: We also evaluate our method on two large datasets: OGBN-ARXIV (from [12]) and ARXIV-YEAR (from [19]). Both these datasets are from the arxiv citation network. In OGBN-ARXIV, the task is to predict the category of the research paper, and in ARXIV-YEAR the task is to predict the year of publishing. We use the publicly available splits for OGBN-ARXIV [14] and follow the same dataset setup mentioned in [19] to generate 5 random splits for ARXIV-YEAR. Note that OGBN-ARXIV is a homophilic dataset while ARXIV-YEAR is a heterophilic datasets.

The detailed dataset statistics can be found in Table 7.

Table 7: Dataset Statistics. The table provides information on the following dataset characteristics: number of nodes, number of edges, feature dimension, number of classes, as well as the count of nodes used for training, validation, and testing.

	Heterophilic Datasets					Homophilic Datasets			
Properties	squirrel	chameleon	roman-empire	minesweeper	arXiv-Year	OGBN-arXiv	citeseer	pubmed	cora
#Nodes	5201	2277	22662	10000	169343	169343	3327	19717	2708
#Edges	222134	38328	32927	39402	1166243	1335586	12431	108365	13264
#Features	2089	500	300	7	128	128	3703	500	1433
#Classes	5	5	18	2	5	40	6	3	7
#Train	2496	1092	11331	5000	84671	90941	1596	9463	1192
#Val	1664	729	5665	2500	42335	29799	1065	6310	796
#Test	1041	456	5666	2500	42337	48603	666	3944	497

8.3 Training Details

We conducted all experiments on a machine equipped with an Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz processor, 440GB RAM, and a Tesla-P100 GPU with 16GB of memory. The experiments were executed using Python 3.9.12 and PyTorch 1.12.0 [26]. To optimize the hyperparameter search, we employed Optuna [1]. We utilized the Adam optimizer [16] for the optimization process.

8.3.1 Unsupervised Training

We conducted hyperparameter tuning for all unsupervised methods using 20 Optuna trials. The hyperparameter ranges and settings for each method are as follows:

DEEPWALK: We set the learning rate to 0.01, number of epochs to 20 and the varied the random walk length over $\{8, 9, 10, 11, 12\}$. Additionally, we varied the context window size over $\{3, 4, 5\}$ and the negative size (number of negative samples per positive sample) over $\{4, 5, 6\}$.

NODE2VEC: For Node2Vec, we set the learning rate to 0.01 and number of epochs to 100. We varied the number of walks over $\{5, 10, 15\}$ and the walk length over $\{40, 50, 60\}$. The p (return parameter) value was chosen from $\{0.1, 0.25, 0.5, 1\}$ and q (in-out parameter) value was chosen from $\{3, 4, 5\}$.

DGI: DGI [35] proposes a self-supervised learning framework for graph representation learning by maximizing the mutual information between local and global structural context of nodes, enabling unsupervised feature extraction in graph neural networks. We relied on the authors' code[1] and the prescribed hyperparameter ranges specific to the DGI model, for our experiments.

MVGRL: MVGRL [8] proposes a method for learning unsupervised node representations by leveraging two views of the graph data, the graph diffusion view and adjacency graph view. We relied on the authors' code[2] and the prescribed hyperparameter ranges specific to MVGRL model, for our experiments.

GRACE: GRACE [36] proposes a technique where two different perspectives of the graph are created through corruption, and the learning process involves maximizing the consistency between the node representations obtained from these two views. We relied on the authors' code[3] and the prescribed hyperparameter ranges specific to the GRACE model, for our experiments.

SUGRL: SUGRL [24] proposes a technique for learning unsupervised representations which capture node proximity, while also utilising node feature information. We relied on the authors' code[4] and the prescribed hyperparameter ranges specific to the SUGRL model, for our experiments.

FIGURE: We followed the setting of the MVGRL model, setting the batch size to 2 and number of GCN layers to 1. We further tuned the learning rate over $\{0.0001, 0.0001, 0.001, 0.01, 0.1\}$ and the sample size (number of nodes selected per batch) from $\{1500, 1750, 2000, 2250\}$.

In each case, we selected the hyperparameters that resulted in the lowest unsupervised training loss.

8.3.2 Supervised Training

For all unsupervised methods, including the baselines and our method, we perform post-training supervised evaluation using logistic regression with 50 Optuna trials. We set the maximum number of epochs to 10000 and select the epoch and hyperparameters that yield the best validation accuracy. The learning rate is swept over the range $\{0.00001, 0.0001, 0.001, 0.0015, 0.01, 0.015, 0.1, 0.5, 1, 2\}$, and the weight decay is varied over $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 0.5, 1, 3\}$.

FIGURE: Along with the hyperparameters described above, following the approach described in [19], we also tune the combination coefficients $(\alpha_i$'s) with a separate learning rate. This separate learning rate is swept over the range $\{0.00001, 0.0001, 0.001, 0.0015, 0.01, 0.015, 0.1, 0.5, 1, 2\}$. In addition, we have a coefficient for masking the incoming embeddings from each filter, which is varied between 0 and 1. Furthermore, these coefficients are passed through an activation layer, and we have two options: `none' and `exp'. When `none' is selected, the coefficients are used directly, while `exp' indicates that they are passed through an exponential function before being used.

FIGURE with RFF: For the experiments involving Random Fourier Features (RFF), we use the same hyperparameter ranges as mentioned above. However, we also tune the gamma parameter which is specific to RFF projections. The gamma parameter is tuned within the range $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$.

[1] https://github.com/PetarV-/DGI.git
[2] https://github.com/kavehhassani/mvgrl.git
[3] https://github.com/CRIPAC-DIG/GRACE.git
[4] https://github.com/YujieMo/SUGRL.git

8.3.3 Negative Sampling for the Identity Filter

In our implementation of F_{GPRGNN} or F_{BernNet}, we follow a specific procedure for handling the filters during training and evaluation. For all filters except the identity filter (I), we employ the negative sampling approach described in Section 6. However, the identity filter is treated differently. During training, we exclude the identity filter and only include it during evaluation.

During negative sampling, the generation of the negative anchor involves shuffling the node features, followed by premultiplying the shuffled node feature matrix with the filter matrix and computing the mean. On the other hand, for the positive anchor, the same procedure is applied without shuffling the node features. This approach encourages the model to learn meaningful patterns and relationships in the data when the filter matrix is not the identity matrix.

The decision to exclude the identity filter during training is based on the observation that it presents a special case where the positive and negative anchors become the same. As a result, the model would optimize and minimize the same quantity, potentially leading to trivial solutions. To prevent this, we exclude the identity filter during training.

By excluding the identity filter during training, we ensure that the model focuses on the other filters in F_{GPRGNN} or F_{BernNet} to capture and leverage the diverse information present in the graph. Including the identity filter only during evaluation allows us to evaluate its contribution to the final performance of the model. This approach helps prevent the model from learning trivial solutions and ensures that it learns meaningful representations by leveraging the other filters.

8.4 Evaluation on large graphs

Table 8: Contains node classification accuracy percentages on two large-scale datasets. OGBN-ARXIV and ARXIV-YEAR have been added. FiGURe$_{32}$ and FiGURe$_{128}$ refer to FiGURe trained with 32 and 128 dimensional representations, respectively, and then projected using RFF. The remaining models are trained at 512 dimensions. Higher numbers indicate better performance. Blue, Red and Green represent the 1st, 2nd and 3rd best performing models, for a particular dataset.

	Heterophilic Datasets					Homophilic Datasets				
	squirrel	chameleon	roman-empire	minesweeper	arXiv-Year	cora	citeseer	pubmed	OGBN-arXiv	Av. gain
DGI	39.61 (1.81)	59.28 (1.23)	47.54 (0.76)	82.51 (0.47)	40.59 (0.09)	84.57 (1.22)	73.96 (1.61)	86.57 (0.52)	65.58 (0.06)	0.61
MVGRL	39.90 (1.39)	54.61 (2.29)	68.50 (0.38)	85.60 (0.35)	OOM	86.22 (1.30)	75.02 (1.72)	87.12 (0.35)	OOM	4.39
GRACE	53.15 (1.10)	68.25 (1.77)	47.83 (0.53)	80.22 (0.45)	OOM	84.79 (1.51)	67.60 (2.01)	87.04 (0.43)	OOM	5.54
SUGRL	43.13 (1.36)	58.60 (2.04)	39.40 (0.49)	82.40 (0.58)	36.96 (0.19)	81.21 (2.07)	67.50 (1.62)	86.90 (0.54)	65.80 (0.06)	0.64
FiGURe$_{32}$	48.89 (1.55)	65.66 (2.52)	67.67 (0.77)	85.28 (0.71)	41.30 (0.21)	82.56 (0.87)	71.25 (2.20)	84.18 (0.53)	66.58 (0.03)	3.18
FiGURe$_{128}$	48.78 (2.48)	66.03 (2.19)	68.10 (1.09)	85.16 (0.58)	41.94 (0.15)	86.14 (1.13)	73.34 (1.91)	85.41 (0.52)	69.11 (0.02)	2.11
FiGURe	52.23 (1.19)	68.55 (1.87)	70.99 (0.52)	85.58 (0.49)	42.26 (0.20)	87.00 (1.24)	74.77 (2.00)	88.60 (0.44)	69.69 (0.03)	0.00

Similar to Table 1, Table 8, provides a comparison with SOTA methods such as DGI, MVGRL, GRACE and SUGRL. However, in this table we also incorporate the large-scale datasets arXiv-Year and OGBN-arXiv. FiGURe shows good performance on these datasets as well, demonstrating the scalability of our method. The last column, average percentage gain, is updated accordingly. Two baseline methods MVGRL and GRACE run into memory issues on the larger datasets and are accordingly reported OOM in the table. Even the lower-dimensional representations (with RFF projections) are able to beat the baselines on these large scale datasets. FiGURe is consistently able to provide gains over the baselines methods regardless of the kind of graph, homophilic, heterophilic or large-scale. This once again demonstrates the generalizability of FiGURe. It is noteworthy that computational efficiency gained by reducing the dimension size becomes significant with the scale of the dataset. On arXiv-Year for example, 128 dimensional embeddings give 1.6x speedup and 32 dimensional embeddings give 1.7x speedup.

8.5 Comparison with other Supervised Methods

Table 9 presents a comparison with common supervised baselines. Specifically, we choose 3 models for comparison, representing three different kinds of supervised methods, standard aggregation models (GCN), spectral filter-based models (GPRGNN) and smart-aggregation models (H$_2$GCN). There are two key observations from this table. Firstly, FiGURe is competitive with the supervised baselines, lagging behind only by a few percentage points in some cases. This suggests that much of the information that is required by the downstream tasks, captured by the supervised models, can be made available through unsupervised methods like FiGURe which uses filter banks. It is important

to note that in FiGURe we only utilize logistic regression while evaluating on the downstream task. This is much more efficient that training a graph neural network end to end. Additionally it is possible that further gains may be obtained by utilizing a non-linear model like an MLP.

Furthermore, as indicated by 9, we can gain further computational efficiency by utilizing lower dimensional representations like 32 and 128 (with RFF), and still not compromise significantly on the performance.

Overall FiGURe manages to remain competitive despite not having access to task-specific labels and is computationally efficient as well.

Table 9: Contains node classification accuracy percentages on heterophilic and homophilic datasets. GCN, GPRGNN and H_2GCN are supervised methods. FiGURe$_{32}$ and FiGURe$_{128}$ refer to FiGURe trained with 32 and 128 dimensional representations, respectively, and then projected using RFF. The remaining models are trained at 512 dimensions. Higher numbers indicate better performance.

	Heterophilic Datasets					Homophilic Datasets			
	squirrel	chameleon	roman-empire	minesweeper	arXiv-Year	OGBN-arXiv	cora	citeseer	pubmed
GCN	47.78 (2.13)	62.83 (1.52)	73.69 (0.74)	89.75 (0.52)	46.02 (0.26)	69.37 (0.00)	87.36 (0.91)	76.47 (1.34)	88.41 (0.46)
GPRGNN	46.31 (2.46)	62.59 (2.04)	64.85 (0.27)	86.24 (0.61)	45.07 (0.21)	68.44 (0.00)	87.77 (1.31)	76.84 (1.69)	89.08 (0.39)
H_2GCN	37.90 (2.02)	58.40 (2.77)	60.11 (0.52)	89.71 (0.31)	49.09 (0.10)	OOM	87.81 (1.35)	77.07 (1.64)	89.59 (0.33)
FiGURe$_{32}$	48.89 (1.55)	65.66 (2.52)	67.67 (0.77)	85.28 (0.71)	41.30 (0.21)	66.58 (0.00)	82.56 (0.87)	71.25 (2.20)	84.18 (0.53)
FiGURe$_{128}$	48.78 (2.48)	66.03 (2.19)	68.10 (1.09)	85.16 (0.58)	41.94 (0.15)	69.11 (0.00)	86.14 (1.13)	73.34 (1.91)	85.41 (0.52)
FiGURe	52.23 (1.19)	68.55 (1.87)	70.99(0.52)	85.58 (0.49)	42.26 (0.20)	69.69 (0.00)	87.00 (1.24)	74.77 (2.00)	88.60 (0.44)

8.6 RFF Projections

As shown in Section 6.2 and in Section 6.4, RFF projections are a computationally efficient way to achieve training by preserving the latent class behavior present in lower dimensional embeddings, by projecting them into a higher dimensional linearly separable space. The natural question that comes up is how do we compute these RFF projections? We provide an algorithm to compute the RFF projections in this section, in algorithm 1. Note that this follows [31].

Algorithm 1 Random Fourier Feature Computation

Require: Input data $X \in \mathbb{R}^{N \times d}$, target dimension D, kernel bandwidth
Ensure: Random Fourier Features $Z \in \mathbb{R}^{N \times D}$
1: Initialize random weight matrix $W \in \mathbb{R}^{d \times D}$ with Gaussian distribution
2: Initialize random bias vector $b \in \mathbb{R}^{D}$ uniformly from $[0, 2\pi]$
3: Compute scaled input $X' = XW + b$
4: Compute random Fourier features $Z = \sqrt{\frac{2}{D}} \cos(X')$
5: return Z

8.7 Computational Comparisons with other Other Unsupervised Methods

Table 10: Mean epoch time (in milliseconds) averaged across 20 trials with different hyperparameters. A lower number means the method is faster. Even though our method is slower at 512 dimensions, using 128 and 32 dimensional embeddings significantly reduces the mean epoch time. Using RFF as described in 6.2 we are able to prevent the performance drops experienced by SUGRL and GRACE.

	SUGRL	GRACE	FiGURe	FiGURe$_{128}$	FiGURe$_{32}$
CORA	15.92 (4.10)	51.19 (6.8)	114.38 (0.51)	20.10 (0.46)	11.54 (0.34)
CITESEER	24.37 (4.92)	77.16 (7.2)	156.24 (0.56)	30.30 (0.60)	17.16 (0.51)
SQUIRREL	33.63 (6.94)	355.2 (67.34)	257.65 (0.76)	47.72 (1.40)	23.52 (1.14)
CHAMELEON	16.91 (5.90)	85.05 (14.1)	98.36 (0.64)	18.56 (0.39)	11.63 (0.48)

In Section 6.4, we compared the computational time of FiGURe with MVGRL and DGI, as all three methods fall under the category of unsupervised methods that preform contrastive learning with representations of the entire graph. However, there is another class of methods, such as

SUGRL and GRACE, that contrast against other nodes without the need for graph representation computation. Consequently, these methods exhibit higher computational efficiency. Hence, as shown in Table 10 upon initial inspection, it appears that SUGRL (at 512 dimensions) exhibits the highest computational efficiency, even outperforming FiGURe$_{128}$. However, despite its computational efficiency, the significant drop in performance across datasets (as discussed in Section 6.1) renders it less favorable for consideration. In fact, FiGURe$_{32}$ offers computational cost savings compared to SUGRL, while also achieving significantly better downstream classification accuracy. Turning to GRACE, it demonstrates greater computational efficiency than FiGURe (at 512 dimensions) for low to medium-sized graphs. However, as the graph size increases, due to random node feature level masking and edge level masking, the computational requirements of GRACE substantially increase (as evidenced by the results on SQUIRREL). Therefore, for larger graphs with more than approximately 5000 nodes, FiGURe proves to be more computationally efficient than GRACE (even at 512 dimensions). Furthermore, considering the performance improvements exhibited by FiGURe, it is evident that FiGURe (combined with RFF projections) emerges as the preferred method for unsupervised contrastive learning in graph data.

8.8 Choice of Filter Banks

In Section 4.1, we explore the flexibility of FiGURe to accommodate various filter banks. When making a choice, it is crucial to examine the intrinsic properties of the filters contained within different filter banks. We pick two filter banks F_{BERNNET} and F_{GPRGNN} and provide an overview of the filters contained in the filter banks. We use these two filter banks as examples to illustrate what should one be looking for, while choosing a filter bank.

Bernstein Polynomials Figure 3 illustrates that as the number of Bernstein Basis increases, the focus on different parts of the eigenspectrum also undergoes changes. With an increase in polynomial order, two notable effects can be observed. Firstly, the number of filters increases, enabling each filter to focus on more fine-grained eigenvalues. This expanded set of polynomial filters allows for a more detailed examination of the eigenspectrum. Secondly, if we examine the first and last Bernstein polynomials, we observe an outward shift in their shape. This shift results in the enhancement of a specific fine-grained part at the ends of the spectrum. These observations demonstrate that Bernstein polynomials offer the capability to selectively target and enhance specific regions of interest within the eigenspectrum

Standard Basis Figure 3 reveals two key observations. Firstly, at a polynomial order of 2, the standard basis exhibit focus at the ends of the spectrum, in contrast to the behavior of Bernstein polynomials, which tend to concentrate more on the middle of the eigenspectrum. This discrepancy highlights the distinct characteristics and emphasis of different polynomial bases in capturing different parts of the eigenspectrum. Secondly, as the number of polynomials increases (in contrast to Bernstein polynomials), the lower order polynomials remain relatively unchanged. Instead, additional polynomials are introduced, offering a more fine-grained focus at the ends of the spectrum. This expansion of polynomials allows for a more detailed exploration of specific regions of interest within the the ends of eigenspectrum.

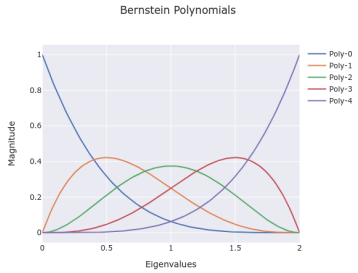
In the context of filter banks, previous studies [20, 4] have demonstrated that certain datasets, such as SQUIRREL and CHAMELEON, benefit from frequency response functions that enhance the tail ends of the eigenspectrum. This observation suggests that the standard basis, which naturally focuses on the ends of the spectrum, may outperform Bernstein basis functions at lower orders. However, as the order of the Bernstein basis increases, as discussed in 4.1, there is a notable improvement in performance. This can be attributed to the increased focus of Bernstein basis functions on specific regions, particularly the ends of the spectrum. As a result, higher-order Bernstein filters exhibit enhanced capability in capturing important information in those regions. It is worth noting that the choice between F_{GPRGNN} and F_{BERNNET} depends on the specific requirements of the downstream task. If the task necessitates a stronger focus on the middle of the spectrum or requires a band-pass or comb-like frequency response, F_{BERNNET} is likely to outperform F_{GPRGNN}. Thus, the selection of the appropriate filter bank should be based on the desired emphasis on different parts of the eigenspectrum. Regarding the performance comparison between F_{BERNNET} and F_{GPRGNN}, it is plausible that as we increase the order of the Bernstein basis, the performance could potentially match that of F_{GPRGNN}. However, further investigation and experimentation are required to determine the specific conditions and orders at which this convergence in performance occurs.

8.9 Visualising RFF Behavior and Community Structure

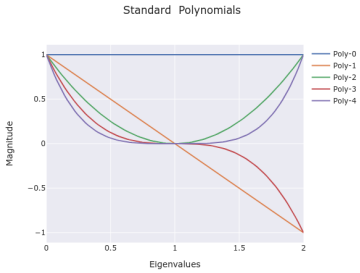
As shown in prior sections, FiGURe improves on both computational efficiency as well as performance by utilising RFF projections. In this section, we aim to gain insights into the behavior of RFF projections and comprehend their underlying operations through a series of simple visualizations.

t-SNE Plots: Figure 4 offers insights into the structure of the embeddings for the CORA dataset across different dimensions. Remarkably, even at lower dimensions (e.g., 32 dimensions), clear class structures are discernible, indicating that the embeddings capture meaningful information related to the class labels. Furthermore, when employing RFF to project the embeddings into higher dimensions, these distinct class structures are still preserved. This suggests that the role of RFF is not to introduce new information, but rather to enhance the suitability of lower-dimensional embeddings for linear classifiers while maintaining the underlying class-related information. Notably, even at 512 dimensions, the class structures remain distinguishable. However, it is worth noting that the class-specific embeddings appear to be more tightly clustered and less dispersed compared to the 32-dimensional embeddings or the projected 32-dimensional embeddings. This suggests that learning a 512-dimensional embedding differs inherently from learning a 32-dimensional embedding and subsequently projecting it into higher dimensions.

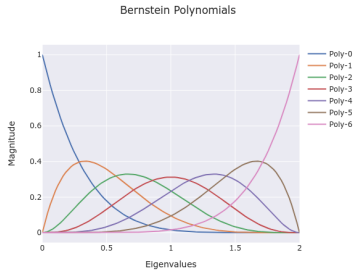
Correlation Plots: Figure 5 offers insights into the correlation patterns within the embeddings generated from the SQUIRREL dataset across different dimensions. In lower dimensions, the embeddings exhibit high correlation with each other, which can be attributed to the presence of a mixture of topics or latent classes within the dataset. However, when the embeddings are projected to higher dimensions using RFF, the correlation is reduced, and a block diagonal matrix emerges. This block diagonal structure indicates the presence of distinct classes or communities within the dataset. Even at 512 dimensions, a more refined block diagonal structure can be observed compared to the correlation matrix of the 32-dimensional embeddings. Furthermore, it is noteworthy that the correlation of the projected embeddings can be regarded as a sparser version of the correlation observed in the 512-dimensional embeddings.



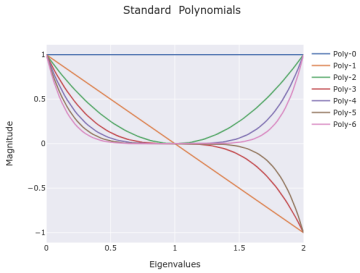
(a) Five Bernstein Basis



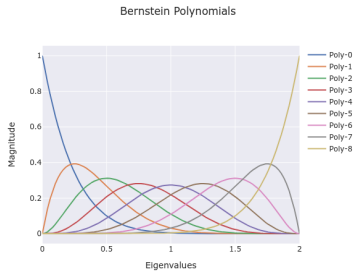
(b) Five Standard Basis



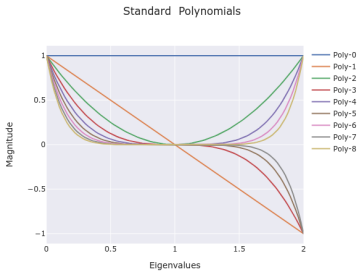
(c) Seven Bernstein Basis



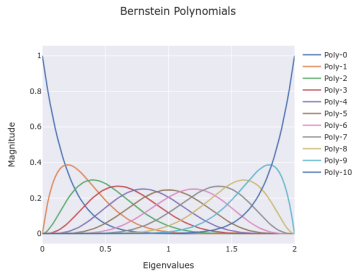
(d) Seven Standard Basis



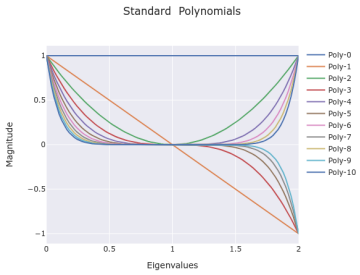
(e) Nine Bernstein Basis



(f) Nine Standard Basis



(g) Eleven Bernstein Basis



(h) Eleven Standard Basis

Figure 3: The figures contain the Bernstein basis as well as standard basis for different degrees. The x-axis of the figures represents the eigenvalues of the Laplacian matrix, while the y-axis represents the magnitude of the polynomials. It is important to note that while plotting the standard polynomials, they are computed with respect to the Laplacian matrix ($\mathbf{L_n}$) rather than the adjacency matrix. As a result, the eigenvalues lie between $[0; 2]$. On the other hand, the Bernstein polynomials are typically defined for the normalised Laplacian matrix, and therefore there is no change in the eigenvalue range (the eigenvalues of the normalised Laplacian matrix typically range from 0 to 2). By using the Laplacian matrix as the basis for plotting the polynomials, we can observe the behavior and magnitude of the polynomials at different eigenvalues, providing insights into their spectral properties and frequency response characteristics.