FIMP: Foundation Model-Informed Message Passing for Graph Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Foundation models have achieved remarkable success across many domains, relying on pretraining over vast amounts of data. Graph-structured data often lacks the same scale as unstructured data, making the development of graph foundation models challenging. In this work, we propose Foundation-Informed Message Passing (FIMP), a Graph Neural Network (GNN) message-passing framework that repurposes existing pretrained non-textual foundation models for graph-based tasks. We show that the self-attention layers of foundation models can effectively be leveraged on graphs to perform cross-node attention-based message-passing. Our model is evaluated across diverse domains on image networks, single-cell RNA sequencing, and fMRI brain activity recordings in finetuned and zero-shot settings. FIMP outperforms strong baselines, demonstrating that it can effectively leverage state-of-the-art foundation models in graph tasks.

1 Introduction

Foundation models have emerged as a new paradigm in artificial intelligence, shifting from narrow, task-specific training to large-scale pretraining of more generalized models (Brown et al., 2020). Through pretraining on vast amounts of data, foundation models serve as a base model which can be adapted to a variety of downstream tasks (Bommasani et al., 2021). Pretraining is typically done in self-supervised fashion through autoregressive language modeling (Radford et al., 2018), masked language/image modeling (Devlin et al., 2018; Chen et al., 2020), or other self-supervised objectives. Standard foundation models have emerged in fields such as Natural Language Processing (NLP) with BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and CLIP (Radford et al., 2021), as well as in Computer Vision (CV) (Yuan et al., 2021). More recently, fields such as single-cell RNA sequencing and neuroscience have also seen the emergence of large-scale foundation models such as scGPT (Cui et al., 2023), Geneformer (Theodoris et al., 2023), and BrainLM (Ortega Caro et al., 2023), representing a new frontier in foundation model research.

Despite the success of foundation models in domains such as language and vision, training and leveraging such models for graph-structured data remains a significant challenge. One key difficulty is the relative scarcity of large-scale, publicly available graph-structured data compared to unstructured data, which limits the capacity to pretrain foundation models specifically for graph tasks. In single-cell RNA sequencing (scRNA-seq) data, for instance, advances in sequencing technology have fueled an exponential growth in available unstructured single-cell transcriptomes (Svensson et al., 2018), however spatial sequencing methods which preserve the spatial organization of cells within the tissue during sequencing lag behind in scale and resolution. Furthermore, traditional Graph Neural Networks (GNNs) tokenize nodes as single embedding vectors, whereas transformer-based foundation models represent inputs as sequences of feature tokens, processing them at a more granular level. Prominent examples include gene tokenization in scGPT (Cui et al., 2023) and image patching in Vision Transformers (ViTs) (Dosovitskiy et al., 2020; He et al., 2022). This feature-level tokenization separates traditional GNNs from foundation models and remains underutilized in graph-based settings. **Bridging the gap between traditional GNNs and pretrained foundation models, and by extension unstructured and structured data, remains an open challenge.**

Existing works have increasingly explored how pretrained foundation models, particularly Large Language Models (LLMs), can be applied to graph-based tasks, primarily in the context of text-attributed graphs. One-for-All (Liu et al., 2023) used LLMs to encode text-attributed graphs for a GNN model, enabling the GNN to do node, edge, and graph-level classification tasks jointly. Talk Like a Graph (Fatemi et al., 2023), NLGraph (Wang et al., 2023), and GPT4Graph (Guo et al., 2023) evaluated LLM reasoning capabilities on graph reasoning benchmarks. These approaches have made significant strides in applying LLMs to text-attributed graphs. However, **non-textual foundation models remain largely underexplored in non-textual graph settings**, leaving significant opportunities for leveraging models like scGPT and BrainLM in graph-based tasks.

To address these challenges, we propose Foundation-Informed Message Passing (FIMP), a novel message-passing framework that repurposes existing pretrained non-textual foundation models for message-passing on graphs. FIMP unifies node tokenization between GNNs and foundation models by viewing nodes as sequences of feature tokens, and introduces a cross-node attention-based message creation module which can be learned from scratch or initialized from pretrained foundation models. We evaluate FIMP across several domains, including image classification (Antequera et al., 2020), spatial transcriptomics, and fMRI brain activity recordings, incorporating state-of-the-art (SOTA) models like ViTs for images (Dosovitskiy et al., 2020), scGPT for scRNAseq (Cui et al., 2023), and BrainLM for brain recordings (Ortega Caro et al., 2023). FIMP demonstrates improvements over strong GNN baselines, highlighting the potential of repurposing non-textual foundation models for graph-based tasks. Additionally, FIMP demonstrates zero-shot embedding capabilities on image networks by leveraging pretrained ViTs (Dosovitskiy et al., 2020), achieving competitive performance without additional training.

In summary, our work makes the following key contributions:

1. We introduce FIMP, a message-passing framework that leverages pretrained non-textual foundation models for graph-based tasks.
2. We evaluate FIMP across diverse domains including images, spatial transcriptomics, and fMRI recordings, repurposing SOTA non-textual foundation models as message creators.
3. We demonstrate FIMP's zero-shot embedding capabilities using pretrained ViTs on image networks, showing that non-textual foundation models can effectively handle graph-based tasks without task-specific training.

2 Related works

2.1 Attention-Based GNNs and Graph Transformers

GATs (Veličković et al., 2017) first introduced the idea of attention-based GNN architectures, learning attention coefficients between neighboring nodes and performing message-passing with a weighted aggregation of neighboring node embeddings. Graph transformers sought to bring the performance and expressivity of the full transformer architecture into the graph domain by modeling graphs as a sequence of node embeddings that represented a fully-connected graph. Graph Transformer Networks (GTNs) (Yun et al., 2019) proposed the first graph transformer architecture, which could learn new graph structures and multi-hop connections. Graph-BERT (Zhang et al., 2020) proposed pretraining on subgraphs and finetuning for node classification and graph clustering tasks. Graph Transformer (Dwivedi & Bresson, 2020) proposed utilizing laplacian eigenvectors as positional encodings for node tokens. SAN (Kreuzer et al., 2021) improved upon it by introducing learnable spectral positional encodings, and Graphormer (Ying et al., 2021) further proposed spatial and centrality encodings for nodes to capture structural relation and node importance in graphs. GPS Graph Transformer (Rampášek et al., 2022) proposed a general framework for building expressive graph transformers composed of positional and structural encodings, graph features, and GNN and attention layers.

In contrast to these works, **FIMP redefines how nodes are represented by encoding each node as a sequence of feature tokens, similar to how transformer models handle input sequences,**

rather than as a single node embedding vector as in GATs and graph transformers. This unique tokenization approach allows FIMP to compute cross-attention at the feature level between the token sequences of neighboring nodes, generating more informative messages that are passed between nodes in the graph. Unlike GATs and graph transformers, which focus on node-level attention, FIMP introduces feature-level attention for message creation. This makes FIMP the first approach to employ tokenized nodes for message-passing over graphs, leveraging the granularity of token interactions.

Additionally, FIMP's tokenization process aligns closely with the tokenization schemes of pretrained non-textual foundation models, minimizing distribution shift when repurposing these models to message-passing over graph-structured data. By integrating foundation models as message creators through this tokenization strategy, FIMP can effectively incorporate powerful pretrained representations in a way that traditional attention-based GNNs and graph transformers cannot.

2.2 LLMs on Text-Attributed Graphs

More recent works have explored using Large Language Models (LLMs) in conjunction with LLMs on text-attributed graphs. GPT4Graph (Guo et al., 2023) evaluated LLM reasoning capabilities on graph reasoning tasks, establishing a benchmark of graph-related tasks for language models. Talk Like a Graph (Fatemi et al., 2023) and NLGraph (Wang et al., 2023) conducted similar studies exploring graph reasoning capabilities of LLMs, and released the GraphQA and NLGraph benchmark datasets, respectively. One-for-all (Liu et al., 2023) used LLMs as an encoding module for text-attributed graphs, and trained a unified GNN model to do node, edge, and graph-level classification using node-of-interest (NOI) subgraphs and prompt nodes. In contrast to these works, we focus on non-textual foundation models and graphs, which have not been explored extensively in graph-based tasks. Our work can be seen as a parallel work to LLM-based works on graphs, aiming to effectively leverage foundation models pretrained on other data domains besides natural language.

3 Preliminaries

3.1 Graph Neural Networks

Graph Neural Networks are a versatile class of models designed to operate over graph-structured data. The core idea of GNNs is to learn node and/or edge attributes through iterative local aggregation steps, which is commonly implemented through Message-Passing Neural Networks (MPNNs) (Gilmer et al., 2017). Below we define our notations for describing GNNs.

Let $G = (V, E)$ denote a graph with a set of nodes V and edges E. Each node has an input feature vector $\vec{x}_i \in \mathbb{R}^f$, where f is the number of input features per node. GNNs iteratively pass messages between neighboring nodes, and in the process use both node features and graph structure to learn node representations $\vec{h}_i \in \mathbb{R}^d$, where d is the hidden dimension of node embeddings. After K message-passing iterations, node representation $\vec{h}_i$ will contain information from its K-hop neighborhood within the graph. The general update rule for the k-th layer can be represented as:

$$\vec{h}^{(k)}_{\mathcal{N}(i)} = \texttt{AGGREGATE}^{(k)}\big(\{\vec{h}^{(k-1)}_j, \forall j \in \mathcal{N}(i)\}\big) \tag{1}$$

$$\vec{h}^{(k)}_i = \texttt{COMBINE}^{(k)}\big(\vec{h}^{(k-1)}_i, \vec{h}^{(k)}_{\mathcal{N}(i)}\big), \tag{2}$$

where $\mathcal{N}(i)$ denotes the neighborhood of node i and $h^{(k)}_i$ is the representation of node i in layer k. The choice of `AGGREGATE` and `COMBINE` vary among different GNN architectures, with the constraint that `AGGREGATE` should be a permutation-invariant aggregator. A readout function is used to map learned node representations into predictions for feature, node, or graph-level tasks.

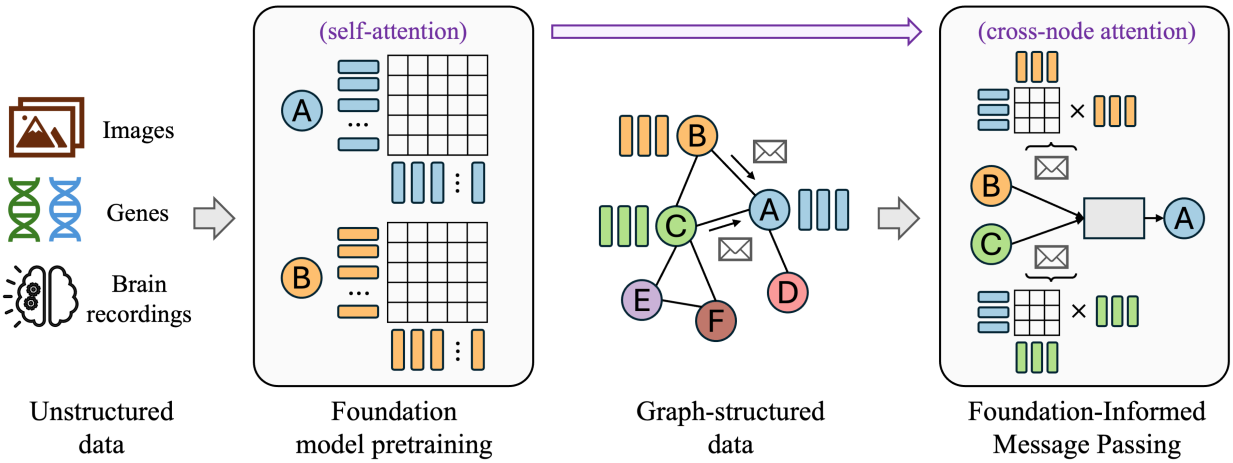

Figure 1: The proposed framework for FIMP. Pre-existing foundation models, pretrained on vast amounts of unstructured data, are adapted for graph-based tasks by repurposing their self-attention layers for cross-node attention between node feature sequences.

3.2 Attention-Based GNNs

Attention-based GNNs, such as Graph Attention Networks (GATs) (Veličković et al., 2017), improve the standard aggregation mechanism by learning the attention coefficients between nodes. In these models, the AGGREGATE function from equation 1 is replaced by an attention mechanism, which computes weighted combinations of neighboring node embeddings based on learned attention scores

$$e_{ji} = a(\mathbf{W}\vec{h}_i || \mathbf{W}\vec{h}_j) \tag{3}$$

$$\alpha_{ji} = \text{softmax}_j(e_{ji}) \tag{4}$$

where α_{ji} represents the final normalized attention coefficient between nodes i and j, a is a learned attention mechanism shared across all node pairs, and $\mathbf{W}$ represents a shared weight matrix.

We note that FIMP employs a more detailed feature-level cross-attention between node feature sequences compared to node embedding attention coefficients used by GATs and graph transformers. This difference is covered in more in detail in the Section 2.

3.3 Foundation Models

Foundation models are generalized Deep Learning models which have been pretrained on large amounts of data, and which can be finetuned for a variety of downstream tasks. In this work, we focus on non-textual foundation models, which define a tokenization procedure for continuous-valued data and typically do pretraining using a masked reconstruction objective. In single-cell RNA sequencing, for instance, scGPT (Cui et al., 2023) tokenizes an input cell as a sequence of gene tokens, and learns a gene embedding table analogous to word embeddings learned in LLMs. Pretraining is done through a masked gene expression prediction objective. In the image domain, ViT-based architectures (Dosovitskiy et al., 2020; He et al., 2022) encode images as a sequence of patches, and similarly for fMRI brain activity recordings, BrainLM (Ortega Caro et al., 2023) tokenizes segments of brain activity signal per brain region into tokens.

4 Methods

We propose a novel message-passing framework, depicted in Figure 1, that uses pretrained non-textual foundation models to generate messages between neighboring nodes in a graph. This leverages the pretrained knowledge of the foundation model to inform message-passing, allowing for pretraining on unstructured data before training on less-abundant graph-structured data.

The FIMP framework consists of several key steps, which we describe in the following subsections:

4.1 Node Tokenization

To ensure that FIMP's node tokenization aligns with the tokenization strategies used in pretrained transformers, we introduce a transformation function, τ. This function converts each node's input features into a structured sequence of feature tokens, similar to how transformers process input entities as sequences. Specifically, τ takes as input node features $X_i \in \mathbb{R}^{f \times c}$, where f is the number of features per node and c is the dimensionality of each feature. It outputs a sequence of f d-dimensional feature vectors representing node i. By aligning the tokenization in FIMP with the tokenization scheme of pretrained foundation models, we reduce the distribution shift in token representation when applying these models to graph-structured data. A general formulation of τ is:

$$H_i = \tau(X_i) = \texttt{COMBINE}(X_i\mathbf{W}, P) \in \mathbb{R}^{f \times d} \tag{5}$$

where $\mathbf{W}$ is a $c \times d$ learned projection into a d-dimensional feature vector, $P \in \mathbb{R}^{f \times d}$ are positional encodings for each feature, and `COMBINE` represents element-wise addition. The dataset-specific instantiations of node tokenization for scRNAseq and fMRI brain recordings are further detailed in Appendix section A.2.

4.2 Message Creation

Our objective is to formulate message creation between two nodes such that pretrained foundation models can be leveraged to create the messages while fitting into the rest of the message-passing framework. Our key observation is that transformer-based foundation models operate using self-attention over sequences of feature tokens (depicted in Figure 1), and contain learned attention weights per layer which are trained to highlight important interactions between feature tokens. Message creation between neighboring node feature sequences can be viewed as a problem of highlighting relevant information which source node j must pass to destination node i, and thus the pretrained attention weights can be repurposed for message creation between two nodes.

We define a cross-node attention-based message creation module which takes as input node feature sequences H_i and H_j, and outputs a message token sequence H_{ji} which will be passed from node j to node i:

$$Q = H_i\mathbf{W}_Q,\ K = H_j\mathbf{W}_K,\ V = H_j\mathbf{W}_V, \tag{6}$$

$$H_{ji} = \texttt{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{7}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learned weight matrices which parameterize the attention mechanism. Note that the attention weights can be randomly initialized and learned from scratch, or initialized from pretrained attention weights. Messages H_{ji} can then be aggregated and used to complete the regular message passing aggregation and update steps, with each node represented by a sequence of feature tokens rather than a single embedding vector. The full algorithm is detailed in Algorithm 1.

We note that **the cross-attention-based message passing operation in FIMP is fundamentally different from other attention-based GNNs**. FIMP is the first method that uses feature-based cross-node attention to construct messages for message passing on graphs. In contrast, attention-based GNNs,

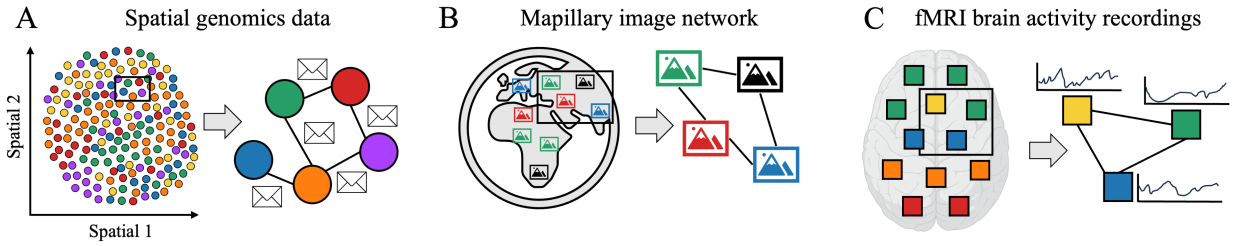

Figure 2: Graph structure present in real-world datasets. (A) In spatially resolved RNA transcriptomics, cells are connected to adjacent cells in the 2D tissue section. (B) In the Mapillary street-view image dataset (Antequera et al., 2020), images form a geographical proximity graph. (C) For fMRI recordings, the brain is parcellated into 424 regions, which are connected using a K-nearest neighbors graph based on the 3D spatial coordinates of each brain region.

particularly GATs and Graph Transformers, do node-level attention and learn scalar attention coefficients between nodes. An overview of attention-based GNNs is provided in the Related Works (Section 2), along with a summary of key differences with FIMP.

Algorithm 1 FIMP

Require: Graph $G = (V, E)$, input features $X_i \in \mathbb{R}^{f \times c}$

$H_i^0 \leftarrow \tau(X_i)$

for $k = 1...K$ **do**

  for node $i \in V$ **do**

    for node $j \in \mathcal{N}(i)$ **do**

      $Q = H_i^{(k-1)} \mathbf{W}_Q$

      $K = H_j^{(k-1)} \mathbf{W}_K$

      $V = H_j^{(k-1)} \mathbf{W}_V$

      $H_{ji}^{(k)} = \texttt{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V$

    end for

    $H_{\mathcal{N}(i)}^{(k)} = \underset{j \in \mathcal{N}(i)}{\texttt{AGGREGATE}}\left(H_{ji}^{(k)}\right)$

    $H_i^{(k)} = \texttt{COMBINE}(H_i^{(k-1)}, H_{\mathcal{N}(i)}^{(k)}) \mathbf{W}$

  end for

end for

4.3 Leveraging Non-Textual Foundation Models

In its base formulation, cross-attention message passing can be done with a simple cross-attention mechanism which is learned from scratch during training. We denote this base version of our architecture as FIMP-base in our experiments. Pretrained foundation models, however, can be repurposed to do the message creation in order to leverage their pretraining over vast amounts of unstructured data. Given a pretrained foundation model $\mathcal{F}$ with learned attention weights per each transformer layer, we adapt the self-attention mechanism in each layer to do cross attention between node feature sequences from neighboring nodes. This adaptation is done in each layer by using the pretrained $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ weights to project both the source and destination node feature sequences H_j and H_i, and computing the scaled dot product attention outlined in equation 7. The final hidden representation output of the foundation model is then taken as the message H_{ji}.

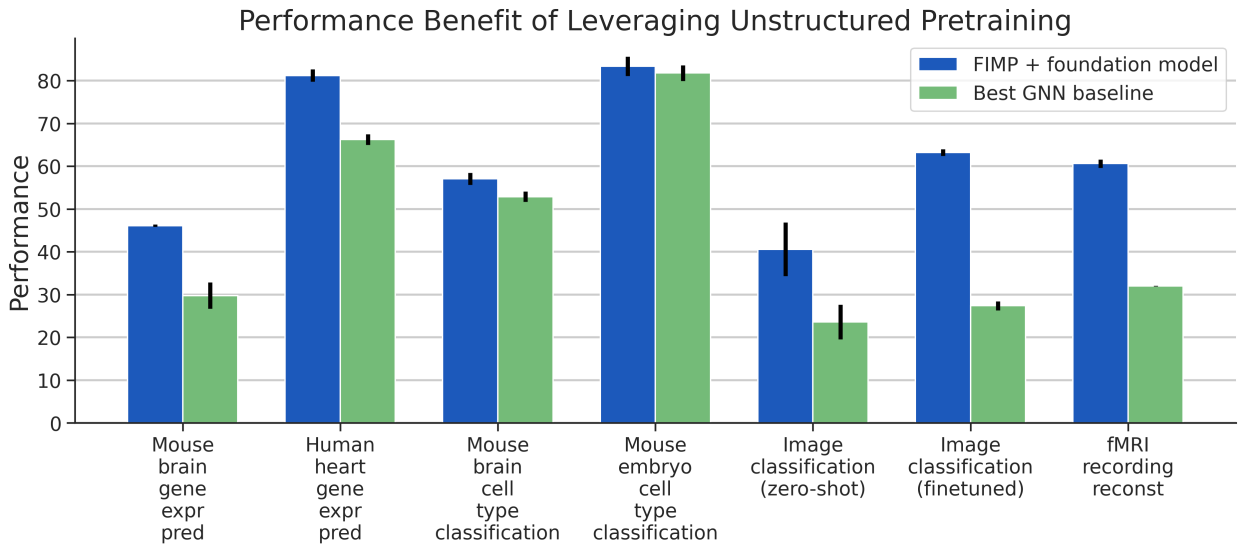

Figure 3: Performance summary across different tasks for FIMP + foundation model versus the best traditional GNN baseline. FIMP improves over traditional GNNs across multiple datasets, highlighting the benefits of leveraging foundation models pretrained on unstructured data.

5 Experiments

In this section, we demonstrate the effectiveness of our proposed framework on a diverse range of tasks in graph-structured settings: (i) gene expression reconstruction and cell type classification on spatial transcriptomics datasets, (ii) image classification on the Mapillary street-view image dataset, and (iii) brain activity reconstruction on fMRI brain recordings from the UK Biobank (UKB) dataset (Miller et al., 2016). The graph structure inherent in each of these datasets is depicted in Figure 2. We show that FIMP allows for the effective integration of pretrained non-textual foundation models into a message-passing framework on graphs. A summary of FIMP's performance versus the best traditional GNN baseline model is provided in Figure 3, with each result further described in Section 5.3.

5.1 Datasets

We benchmark FIMP on gene expression prediction and cell type classification using three publicly-available spatial transcriptomics datasets. The Slideseq-V2 spatial transcriptomics dataset (Stickels et al., 2021) is a mouse hippocampus dataset consisting of $41,786$ cells and $4,000$ genes, with 14 different cell types. A second spatial dataset of human heart tissue was obtained from the 10X Genomics public spatial data repository, consisting of $4,247$ cells each with $36,601$ measured genes. A third spatial dataset, SeqFISH (Lohoff et al., 2020), consists of $15,000$ cells and 342 genes taken from mouse embryo tissue sections. For all spatial transcriptomics datasets, we follow standard preprocessing and normalization procedures for RNA sequencing data, including count normalization and log transformation (Haque et al., 2017). Full dataset details are in Appendix section A.1.

For image classification, we use the Mapillary planet-scale image dataset (Antequera et al., 2020), a dataset of 750,000 street-view images collected from over 170 countries around the world. Images are 1000-2000 pixels in height and width, originating from a variety of cameras and conditions depicting natural landscapes and buildings. Each image has a recorded latitude and longitude coordinate, forming a geographical proximity graph where each node represents a full image, connected to nearby image nodes if they are within 10 miles of one another. We evaluate FIMP on a task where the aim is to classify the country of origin based on the visual features of each image node and its neighborhood. We train on 100,000 training images, and test on

the predefined 10,000 test image set, with country labels determined for each image based on its latitude and longitude coordinates.

For brain activity reconstruction, we use the UK Biobank dataset (Miller et al., 2016), comprising of 76,296 task-based and resting-state functional MRI (fMRI) recordings from 41,986 patients aged 40 to 69 years old. All recordings went through standard preprocessing steps for fMRI recordings (Salimi-Khorshidi et al., 2014; Abdallah, 2021), and was parcellated into 424 brain regions using the AAL-424 atlas (Nemati et al., 2020). We apply robust scaling per brain region by subtracting the median and dividing by the interquartile range computed across subjects. Our training set comprised of $60,000$ recordings, with the rest reserved for validation and test.

5.2 Experimental Setup

All models were implemented in Pytorch Geometric (Fey & Lenssen, 2019) and Pytorch (Paszke et al., 2019), and trained using the Adam optimizer (Kingma & Ba, 2014). Flash Attention (Dao et al., 2022) is used to improve the computational footprint during message passing. Hyperparameter tuning was done through a grid search over standard values for learning rate, dropout, attention dropout, and weight decay. For all experiments, a 24GB NVIDIA GPU (RTX3090 or A5000) was used for training. Experimental setup details for specific datasets are provided in the Appendix A.3.

All foundation models used in our experiments were sourced from the open-source weights release of their respective works. For experiments on single-cell datasets, we incorporate scGPT (Cui et al., 2023) for message creation in FIMP-scGPT, providing a state-of-the-art scRNA-seq foundation model for spatial RNA-seq processing. scGPT consists of a 12-layer transformer with 54 million parameters, and is pretrained using a masked gene expression prediction objective on over 33 million cells. The pretrained gene embedding table is utilized from the scGPT checkpoint in FIMP-scGPT's tokenization scheme, directly incorporating the pretrained gene representations learned by scGPT from non-spatial RNA-seq datasets. We additionally utilize gene embeddings obtained by GenePT (Chen & Zou, 2023) in FIMP-GenePT, consisting of GPT-3.5 embeddings of gene function descriptions based on biomedical literature. This provides a gene embedding based on natural language pretraining on biomedical literature, constrasting with gene expression data-driven gene embeddings from scGPT. For image classification, a standard ViT (Dosovitskiy et al., 2020) with 12 transformer layers and 86 million parameters is used as a message creator. The patch encoder from the ViT is also reused from the ViT embedding module. For experiments on fMRI brain recordings, the BrainLM (Ortega Caro et al., 2023) model was used, which consists of a Masked Autoencoder transformer with an 8-layer encoder and 4-layer decoder, totaling 26 million parameters.

For both supervised and self-supervised task baseline comparisons, we compare FIMP against popular message-passing GNN architectures, including GCN (Kipf & Welling, 2016), GraphSAGE (Hamilton et al., 2017), Graph Attention Networks (GATs) (Veličković et al., 2017), and Graph Isomorphism Networks (GINs) (Xu et al., 2018). We also compare FIMP against more recent GNN architectures, namely GraphMAE (Hou et al., 2022), a masked graph autoencoder model, Graph Inductive bias Transformer (GRIT) Ma et al. (2023), and GPS Graph Transformer (Rampášek et al., 2022), a SOTA graph transformer framework. For supervised classification tasks, we additionally compare to the pretrained foundation model in each domain, which does not take graph structure as input and instead treats each node as an individual sample.

5.3 Results

Table 1 reports gene expression prediction results on the human heart and mouse hippocampus spatial datasets. Our experiments demonstrate that FIMP-base, which is trained from scratch with a randomly initialized cross-attention layer, consistently outperforms standard GNN baselines in predicting masked gene expression values. We attribute this success to its improved gene tokenization strategy, since the learned gene embeddings effectively capture nuanced gene-specific information from the data. Furthermore, when we incorporate pretrained gene embeddings from GenePT and scGPT (referred to as FIMP-GenePT and FIMP-scGPT, respectively), we observe additional performance gains. These findings reinforce the importance of leveraging models pretrained on unstructured data—in this case, scRNA-seq—that contains useful knowledge which can transfer positively to spatial settings. Notably, employing an out-of-domain foundation model such

Method	Mouse Hippocampus		Human Heart	
	MSE (↓)	R^2(↑)	MSE (↓)	R^2(↑)
GCN	0.0211 ± 0.0018	0.0236 ± 0.0457	0.0045 ± 0.00019	0.3368 ± 0.04453
GraphSAGE	0.0181 ± 0.0012	0.1853 ± 0.0306	0.0054 ± 0.00033	0.2080 ± 0.01973
GAT	0.0201 ± 0.0008	0.0905 ± 0.0233	0.0043 ± 0.00023	0.3468 ± 0.02313
GIN	0.0175 ± 0.0009	0.1707 ± 0.0424	0.0025 ± 0.00029	0.6625 ± 0.01269
GraphMAE	0.0178 ± 0.0006	0.1538 ± 0.0254	0.0024 ± 0.00016	0.6589 ± 0.01715
GPS	0.0149 ± 0.0012	0.2977 ± 0.0308	0.0024 ± 0.00031	0.6538 ± 0.01043
scGPT	0.0169 ± 0.0007	0.2087 ± 0.0191	0.0209 ± 0.00072	0.0229 ± 0.01757
FIMP-base (ours)	*0.0134 ± 0.0009*	*0.3815 ± 0.0226*	*0.0021 ± 0.00003*	*0.6955 ± 0.02048*
FIMP + ViT (ours)	0.0128 ± 0.0010	0.3506 ± 0.0452	0.0042 ± 0.00089	0.4026 ± 0.08102
FIMP + GenePT (ours)	0.0129 ± 0.0005	0.4058 ± 0.0302	0.0013 ± 0.00023	0.7952 ± 0.01430
FIMP + scGPT (ours)	**0.0119 ± 0.0008**	**0.4612 ± 0.0029**	**0.0011 ± 0.00008**	**0.8119 ± 0.01428**

Table 1: Gene expression prediction results on the mouse hippocampus and human heart spatial transcriptomics datasets. Performance is reported across 5 runs in terms of MSE and R^2. FIMP outperforms baseline methods on predicting gene expression.

Method	Mouse Hippocampus		Mouse Embryo	
	Accuracy (↑)	F1-score (↑)	Accuracy (↑)	F1-score (↑)
GCN	47.59 ± 3.788	0.445 ± 0.050	74.23 ± 1.250	0.720 ± 0.008
GraphSAGE	51.81 ± 3.229	0.495 ± 0.036	80.77 ± 3.071	0.793 ± 0.031
GAT	46.21 ± 3.110	0.442 ± 0.031	71.07 ± 1.452	0.690 ± 0.014
GIN	52.71 ± 0.421	0.507 ± 0.008	75.51 ± 1.398	0.743 ± 0.012
GPS	*52.89 ± 1.176*	*0.510 ± 0.008*	*81.77 ± 3.175*	*0.813 ± 0.038*
FIMP-base	49.04 ± 1.215	0.464 ± 0.019	81.35 ± 2.285	0.807 ± 0.026
scGPT	53.50 ± 0.424	0.518 ± 0.005	82.93 ± 0.419	0.820 ± 0.005
FIMP-scGPT	**57.05 ± 1.393**	**0.554 ± 0.004**	**83.33 ± 2.250**	**0.821 ± 0.022**

Table 2: Cell type classification results on the mouse hippocampus and embryo spatial transcriptomics datasets. Performance is reported in terms of accuracy and F1-score. FIMP outperforms baseline models at predicting cell types.

as a ViT for message creation does not yield any performance improvements in this task. This suggests that the benefits of FIMP are not simply due to increased model capacity but critically depend on the pretraining domain aligning with the graph features in the data.

Table 2 presents results for cell type classification on the mouse hippocampus and embryo spatial transcriptomics datasets. Our results demonstrate that FIMP-scGPT achieves the highest classification performance across both datasets. On the mouse hippocampus dataset, FIMP-scGPT attains an accuracy of 57.05% and an F1-score of 0.554, surpassing all baseline models. The closest-performing baseline, scGPT with no graph structure, achieves 53.50% accuracy, indicating that the integration of cross-attention-based message passing in FIMP provides a notable improvement in performance over using scGPT alone. Among traditional GNNs, GPS achieves the best performance, with 52.89% accuracy, but still lags behind scGPT and FIMP-scGPT, indicating the utility of foundation model pretraining for cell type classification. On the mouse embryo dataset, FIMP-scGPT again outperforms all baselines, achieving 83.33% accuracy and an F1-score of 0.821. These results indicate that leveraging foundation model-informed message passing enhances cell type classification beyond standard GNN-based approaches and even outperforms directly using scGPT without graph structure.

Table 3 presents classification results on the Mapillary image dataset, demonstrating the effectiveness of FIMP for graph-based image understanding. We observe that FIMP-base outperforms traditional GNN baselines by over 10%, despite being trained from scratch. This improvement is attributed to FIMP's image

Setting	Method	Accuracy (↑)	F1-score (↑)
Finetuned	GCN	23.9 ± 1.152	0.182 ± 0.0151
	GraphSAGE	22.2 ± 1.703	0.164 ± 0.0129
	GAT	22.9 ± 0.596	0.189 ± 0.0042
	GIN	26.4 ± 1.240	0.254 ± 0.0143
	GraphMAE	15.8 ± 0.828	0.083 ± 0.0056
	GRIT	19.7 ± 0.863	0.224 ± 0.0079
	GPS	27.4 ± 1.046	0.268 ± 0.0157
	FIMP-base (ours)	*38.6 ± 1.174*	*0.422 ± 0.0170*
	ViT	56.5 ± 3.187	0.597 ± 0.0065
	FIMP-ViT (ours)	**63.2 ± 0.764**	**0.684 ± 0.0076**
Zero-shot	Majority class	17.0 ± 3.162	–
	GraphSAGE	23.6 ± 4.037	0.129 ± 0.0309
	ViT	34.0 ± 3.391	0.282 ± 0.0389
	FIMP-ViT (ours)	**40.6 ± 6.269**	**0.371 ± 0.0550**

Table 3: Image classification results on the Mapillary street-view image dataset. FIMP significantly improves over baseline models in image classification, and creates zero-shot embeddings of the image network on par with trained GNN baselines.

tokenization process, where images are divided into patches that allow more expressive message-passing compared to conventional node embeddings on an image network. The best performance is achieved by FIMP-ViT, which integrates a pretrained ViT (Dosovitskiy et al., 2020) as a message creator. By leveraging ViT's learned visual feature representations and attention layers, FIMP-ViT achieves a substantial accuracy gain over all baselines, demonstrating that FIMP is capable of effectively transferring pretrained image feature representations into graph-structured tasks. A breakdown of training time for each model is provided in Appendix section A.5.

We further explore a zero-shot setting to evaluate whether FIMP can leverage a pretrained ViT without any finetuning on image network data. To do this, we embed subgraphs of the Mapillary dataset using FIMP-ViT and compare its embeddings to those generated by (i) a randomly initialized GraphSAGE model (Hamilton et al., 2017), which serves as a standard GNN baseline, and (ii) the ViT model itself without any graph structure, treating each image independently. To assess the quality of these embeddings, we train a linear classifier on 75% of the embeddings and use it to predict labels for the remaining 25%. We find that FIMP-ViT achieves over 40% classification accuracy in this zero-shot setting, which is on par with fully trained GNN baselines despite having no graph-specific training. This demonstrates that FIMP is capable of transferring the capabilities of pretrained non-textual foundation models onto graph-based data, allowing for exciting zero-shot applications of models trained on unstructured data in graph settings.

Table 4 summarizes the results for fMRI recording reconstruction on the UK Biobank dataset (Miller et al., 2016). In this task, models predict missing brain activity signals from partially masked fMRI recordings, evaluating their ability to reconstruct brain activity patterns using neighboring region information. Across all baselines and masking strategies, FIMP-base significantly outperforms traditional GNN models, achieving a 17.6% reduction in mean squared error (MSE) compared to the best-performing GNN. While conventional GNNs rely on node-level embeddings, FIMP's tokenization framework allows more granular interactions between brain regions, leading to more accurate signal recovery. Furthermore, integrating BrainLM, a pretrained foundation model for brain activity recordings (Ortega Caro et al., 2023), achieves an additional 2% improvement over FIMP-base. This suggests that the attention weights learned in BrainLM encode meaningful relationships between brain regions, which can be repurposed within FIMP to refine message-passing over the spatiotemporal graph of fMRI signals. These results emphasize that FIMP not only surpasses standard GNNs in reconstructing masked brain activity, but also benefits from incorporating pretrained models. By aligning tokenization strategies with foundation model architectures, FIMP effectively integrates

Method	Masking Strategy	MSE (↓)	R^2 (↑)
GCN	Replace noise	0.554 ± 0.00002	0.189 ± 0.00003
	Fill in mean	0.513 ± 0.00019	0.248 ± 0.00028
	Linear interpolation	0.535 ± 0.00137	0.217 ± 0.00200
GraphSAGE	Replace noise	0.534 ± 0.00107	0.218 ± 0.00157
	Fill in mean	*0.464 ± 0.00039*	*0.320 ± 0.00057*
	Linear interpolation	0.500 ± 0.00094	0.268 ± 0.00138
GAT	Replace noise	0.548 ± 0.00004	0.197 ± 0.00007
	Fill in mean	0.505 ± 0.00005	0.260 ± 0.00007
	Linear interpolation	0.527 ± 0.00052	0.229 ± 0.00076
GIN	Replace noise	0.564 ± 0.00131	0.174 ± 0.00192
	Fill in mean	0.533 ± 0.00185	0.220 ± 0.00271
	Linear interpolation	0.559 ± 0.00061	0.181 ± 0.00090
GraphMAE	Replace noise	0.582 ± 0.00070	0.147 ± 0.00103
	Fill in mean	0.544 ± 0.00030	0.203 ± 0.00044
	Linear interpolation	0.573 ± 0.00091	0.160 ± 0.00134
GPS Graph Transformer	Replace noise	0.577 ± 0.00279	0.154 ± 0.00408
	Fill in mean	0.547 ± 0.01030	0.198 ± 0.01506
	Linear interpolation	0.557 ± 0.01034	0.184 ± 0.01512
FIMP-base	Tokenization + PE	0.288 ± 0.00713	0.578 ± 0.01043
FIMP-BrainLM	Tokenization + PE	**0.267 ± 0.00493**	**0.606 ± 0.00972**

Table 4: Brain activity reconstruction results on the UK Biobank dataset. Performance is reported across 5 runs. FIMP improves upon baselines by 25.8%, with a further improvement of 2.8% by leveraging BrainLM (Ortega Caro et al., 2023).

large-scale pretrained brain activity foundation models, offering a promising direction for spatiotemporal analysis of brain activity.

5.4 Ablation Studies

To better understand the benefit of using a pretrained foundation model with traditional GNNs versus using foundation models with the FIMP architecture, we conducted an ablation study on the Mapillary image classification task. Specifically, we compared the performance of GNN baseline models using embeddings from a pretrained ViT model as input, allowing us to disentangle the effects of the foundation model embeddings from the performance improvements provided by FIMP's message-passing architecture.

Table 5 presents the results of this experiment. We observed a mixed effect when replacing raw pixel image inputs with ViT embeddings across different GNN architectures. While some more expressive GNNs such as GIN and GPS showed improved classification performance with ViT embedding inputs, others remained unchanged or degraded slightly. We attribute this to capacity limitations in certain GNNs, which may struggle to effectively leverage the richer ViT embeddings during training. The increase in performance for certain GNN baselines suggests that incorporating pretrained representations can be beneficial, however FIMP still consistently outperforms all baselines, demonstrating that its advantage is not solely from utilizing foundation models, but from its ability to repurpose pretrained model weights to perform message-passing across graph structures. This highlights a key distinction in FIMP: while non-textual foundation models like ViT cannot natively process graph-structured data, FIMP enables their meaningful application in graph-based learning beyond simple node-wise embedding inputs, leveraging cross-node interactions to achieve superior performance.

Model Input	Method	Accuracy (↑)	F1-score (↑)
Image Pixels	GCN	23.9 ± 1.152	0.182 ± 0.0151
	GraphSAGE	22.2 ± 1.703	0.164 ± 0.0129
	GAT	22.9 ± 0.596	0.189 ± 0.0042
	GIN	26.4 ± 1.240	0.254 ± 0.0143
	GraphMAE	15.8 ± 0.828	0.083 ± 0.0056
	GPS	27.4 ± 1.046	0.268 ± 0.0157
ViT embeddings	GCN	16.0 ± 0.801	0.085 ± 0.0050
	GraphSAGE	15.8 ± 0.980	0.083 ± 0.0064
	GAT	20.5 ± 3.941	0.141 ± 0.0490
	GIN	45.4 ± 0.670	0.479 ± 0.0059
	GraphMAE	15.8 ± 0.803	0.083 ± 0.0049
	GPS	50.0 ± 1.728	0.530 ± 0.0199
Image Pixels	FIMP-base (ours)	*38.6 ± 1.174*	*0.422 ± 0.0170*
	ViT	56.5 ± 3.187	0.597 ± 0.0065
	FIMP-ViT (ours)	**63.2 ± 0.764**	**0.684 ± 0.0076**

Table 5: Ablation study comparing FIMP with GNN baseline models with foundation model embeddings as input on the Mapillary image classification task. While foundation model embeddings do enhance performance for some GNNs, FIMP-ViT notably outperforms all baselines by effectively utilizing ViT pretrained weights for message-passing.

6 Conclusions, Limitations, and Future Research

In this work, we introduce Foundation-Informed Message Passing (FIMP), a message-passing framework which repurposes pretrained non-textual foundation models for message-passing on graphs. Our approach represents the first broad exploration of utilizing non-textual pretrained foundation models graph settings. FIMP demonstrates improved performance over baselines across multiple tasks in image networks, spatial transcriptomics data, and fMRI brain activity recordings, confirming the performance benefits of leveraging non-textual foundation models in graph-based tasks. Furthermore, FIMP demonstrates zero-shot embedding capabilities on image networks that are on par with trained GNNs. This highlights the potential for zero-shot applications with pretrained non-textual foundation models on graphs despite them not natively taking graph structure as input.

There are several avenues for improvement upon our method, which we leave for future work. Currently, our evaluation of FIMP is limited to image and biological data. Protein design and social networks are promising areas of future research. Additionally, supporting multimodal graphs, heterogeneous graphs, and edge features would all expand the potential applications of FIMP. Finally, improving the scalability of FIMP to large graphs through strategies such as feature selection and efficient attention mechanisms beyond our usage of Flash Attention is an important future direction.

References

Charlotte Aaberg-Jessen, Mia D Sørensen, Ana LSA Matos, José M Moreira, Nils Brünner, Arnon Knudsen, and Bjarne W Kristensen. Co-expression of timp-1 and its cell surface binding partner cd63 in glioblastomas. *BMC cancer*, 18:1–16, 2018.

Chadi G Abdallah. Brain networks associated with covid-19 risk: Data from 3662 participants. *Chronic Stress*, 5:24705470211066770, 2021.

Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 589–604. Springer, 2020.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Yiqun T Chen and James Zou. Genept: A simple but hard-to-beat foundation model for genes and cells built from chatgpt. *bioRxiv*, pp. 2023–10, 2023.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scgpt: towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, pp. 2023–04, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Saurabh Dey, Soumya Basu, and Amit Ranjan. Revisiting the role of cd63 as pro-tumorigenic or anti-tumorigenic tetraspanin in cancers and its theragnostic implications. *Advanced Biology*, pp. 2300078, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*, 2023.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Jiayan Guo, Lun Du, and Hengyu Liu. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*, 2023.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Ashraful Haque, Jessica Engel, Sarah A Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9:1–12, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 594–604, 2022.

Breanne E Kearney and Ruth A Lanius. The brain-body disconnect: A somatic sensory basis for trauma-related disorders. *Frontiers in Neuroscience*, 16:1881, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34: 21618–21629, 2021.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. *arXiv preprint arXiv:2310.00149*, 2023.

Tim Lohoff, Shila Ghazanfar, Alsu Missarova, Noushin Koulena, Nico Pierson, Jonathan A Griffiths, Evan S Bardot, C-HL Eng, Richard CV Tyser, Ricard Argelaguet, et al. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis. *BioRxiv*, pp. 2020–11, 2020.

Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K Dokania, Mark Coates, Philip Torr, and Ser-Nam Lim. Graph inductive biases in transformers without message passing. In *International Conference on Machine Learning*, pp. 23321–23337. PMLR, 2023.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19 (11):1523–1536, 2016.

Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. A unique brain connectome fingerprint predates and predicts response to antidepressants. *IScience*, 23(1), 2020.

Josue Ortega Caro, Antonio Henrique Oliveira Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pp. 2023–09, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022.

Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F Beckmann, Matthew F Glasser, Ludovica Griffanti, and Stephen M Smith. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, pp. 1–9, 2023.

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

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? *arXiv preprint arXiv:2305.10037*, 2023.

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

Shizhen Yan, Juntao Chen, Xiaojuan Yin, Ziliang Zhu, Ziping Liang, Hua Jin, Han Li, Jianzhong Yin, Yunpeng Jiang, and Yaoyuan Xia. The structural basis of age-related decline in global motion perception at fast and slow speeds. *Neuropsychologia*, 183:108507, 2023.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. *Advances in neural information processing systems*, 32, 2019.

Jiawei Zhang, Haopeng Zhang, Congying Xia, and Li Sun. Graph-bert: Only attention is needed for learning graph representations. *arXiv preprint arXiv:2001.05140*, 2020.

A Appendix

A.1 Datasets (Extended)

We use three publicly-available spatial transcriptomics datasets. The Slideseq-V2 spatial transcriptomics dataset (Stickels et al., 2021) is a mouse hippocampus dataset consisting of $41,786$ cells and $4,000$ genes, with 14 different cell type classes. A second spatial dataset of human heart tissue was obtained from the 10X Genomics public spatial data repository, consisting of 4247 cells each with 36601 measured genes. A third spatial dataset, SeqFISH (Lohoff et al., 2020), consists of $15,000$ cells and 342 genes taken from mouse embryo tissue sections. For all spatial transcriptomics datasets, we follow standard preprocessing and normalization procedures for RNA sequencing data, including count normalization and log transformation (Haque et al., 2017). For all datasets, we take the intersection of gene features which are present in the scGPT (Cui et al., 2023) pretrained foundation model, and split nodes into training, validation, and test sets with a 70/10/20 split. For graph adjacency information, we utilize the neighbor connectivity matrix present in each spatial transcriptomics dataset, which is derived from the original tissue section coordinates.

The Mapillary planet-scale image dataset (Antequera et al., 2020) is a dataset of 750,000 street-view images collected from over 170 countries around the world. Images are 1000-2000 pixels in height and width, originating from a variety of cameras and conditions depicting natural landscapes and buildings. Each

image has a recorded latitude and longitude coordinate, forming a geographical proximity graph where each node represents a full image, connected to nearby image nodes if they are within 10 miles of one another. We evaluate FIMP on a geoguesser task, where the aim is to classify the country of origin based on the visual features of each image node and its neighborhood of nearby images. We train on 100,000 training images, and test on the predefined 10,000 test image set, with country labels determined for each image based on its latitude and longitude coordinates.

The UK Biobank dataset (Miller et al., 2016) comprises of 76,296 task-based and resting-state functional MRI (fMRI) recordings from 41,986 patients aged 40 to 69 years old. Recordings were acquired on a Siemens 3T scanner at 0.735s temporal resolution. All recordings went through standard preprocessing steps, including motion correction, normalization, temporal filtering, and ICA denoising (Salimi-Khorshidi et al., 2014; Abdallah, 2021). We parcellated the brain into 424 brain regions using the AAL-424 atlas (Nemati et al., 2020), yielding 424-dimensional scan sequences sampled at [a]1 Hz. Finally, robust scaling was applied by subtracting the median and dividing by the interquartile range computed across subjects for each brain region. Our training set comprised of $60,000$ of the fMRI recordings, with the rest reserved for validation and test sets.

A.2 Node Tokenization (Extended)

The general formulation of node tokenization (τ) becomes dataset-specific following tokenization schemes defined by foundation models on different data modalities. For instance, on datasets with input node feature vectors $\vec{x}_i \in \mathbb{R}^f$, such as a gene expression vector for a cell containing f genes, we can see X_i as an expanded feature vector with $c = 1$, and $\mathbf{W}$ as a projection of a scalar gene expression value into a d-dimensional vector embedding. The positional encoding P would then represent a learned gene embedding $P \in \mathbb{R}^{f \times d}$, analogous to word embeddings in natural language. The concatenation operation in equation 5 would combine the expression value projection with its corresponding gene encoding, as in scGPT (Cui et al., 2023) and Geneformer (Theodoris et al., 2023).

For experiments on image datasets, τ is formulated as a patch encoding procedure following standard ViTs (Dosovitskiy et al., 2020), where an input image is divided into f patches, each with c pixels, that are embedded via a learned patch projector $\mathbf{W}$. Positional encoding P is done through fixed 2D sinusoidal positional encoding which is concatenated with each patch embedding. For fMRI brain activity recordings, τ follows a spatiotemporal patching process as in the BrainLM foundation model (Ortega Caro et al., 2023), where for each brain region, segments of $c = 20$ signal timepoints are embedded via a learned projection $\mathbf{W}$. Spatial positional encoding is done through a learned projection of XYZ coordinates of each brain region, and temporal positional encoding is done using sinusoidal positional encoding.

A.3 Experimental Setup (Extended)

The following section gives additional details about experimental setup across different datasets.

A.3.1 Image Classification

For image classification experiments, random 512x512 crops were taken from each image during training, with a 512x512 center crop taken at test time. Per-channel normalization was done on each image using statistics calculated across training images in the Mapillary image dataset. For FIMP and FIMP-ViT experiments, images were divided into 32x32 patches following the standard ViT patch encoding procedure (Dosovitskiy et al., 2020). For baseline GNNs, pixel values for each image were flattened and encoded using a learned projection.

A.3.2 Gene Expression Prediction

For gene expression prediction experiments on spatial transcriptomics datasets, we limit the number of cells in each dataset to 5% of the original dataset size, leaving 1000 cells for the mouse hippocampus spatial dataset, and 200 cells for the human heart spatial dataset. This creates a challenging limited data setting for predicting gene expression values on each spatial dataset. We sample 50 nonzero expressed genes in each

cell for all models and mask out 80% of the gene expression values, taking MSE loss against only masked out genes.

A.3.3 fMRI Recording Reconstruction

In brain activity reconstruction experiments , we sample 320 consecutive timepoints from each fMRI recording, giving a recording of 424 brain regions with 320 timepoints of signal for each region. Each brain region is represented as 1 node in the graph, with node features being the 320 timepoints of signal. We segment the timepoints for each brain region into patches of 20 timepoints, and perform masked reconstruction of brain recording signals. For FIMP and variants of FIMP leveraging foundation models, masked patches are replaced with a mask token, and the signals are predicted back by the model. For baseline GNN models, node features comprise of the 320 timepoints of signal, and we explore three methods for replacing masked out patch values: i) replacing with random noise, ii) filling in with the mean value of the brain region, and iii) linearly interpolating between adjacent non-masked timepoint values. All models mask out 50% of patches per each brain region, with mean squared error (MSE) taken against the original data.

A.3.4 Foundation Models

For experiments on single-cell datasets, the scGPT (Cui et al., 2023) whole-human checkpoint is incorporated for message creation in FIMP-scGPT, consisting of a 12-layer transformer with 54 million parameters. scGPT is pretrained using a masked gene expression prediction objective on over 33 million cells from a diverse array of human tissues and organs. The pretrained gene embedding table is also utilized from the pretrained scGPT checkpoint, representing pretrained knowledge about gene identities in transcriptomics datasets. Additionally, we also utilize the gene embeddings obtained by GenePT (Chen & Zou, 2023), which are GPT-3.5 embeddings of gene function descriptions based on biomedical literature, as another pretrained gene embedding experiment. For image classification, a standard ViT (Dosovitskiy et al., 2020) with 12 transformer layers and 86 million parameters is used as a message creator. The patch encoder from the ViT is also reused from the ViT embedding module. For experiments on fMRI brain recordings, the BrainLM (Ortega Caro et al., 2023) model was used, which consists of a Masked Autoencoder transformer with an 8-layer encoder and 4-layer decoder, totaling 26 million parameters.

A.3.5 Baselines

For both supervised and self-supervised tasks, we compare FIMP against popular message-passing GNN architectures, including GCN (Kipf & Welling, 2016), GraphSAGE (Hamilton et al., 2017), Graph Attention Networks (GATs) (Veličković et al., 2017), and Graph Isomorphism Networks (GINs) (Xu et al., 2018). We also compare FIMP against more recent GNN architectures, namely GraphMAE (Hou et al., 2022), a masked graph autoencoder model, and GPS Graph Transformer (Rampášek et al., 2022), a SOTA graph transformer framework. For supervised classification tasks, we additionally compare to the pretrained foundation model with no graph structure input.

A.4 Attention Visualizations

A.4.1 Functional Region Attention in fMRI Recordings

During message passing on the fMRI recording graphs, FIMP generates cross-attention matrices during message-creation between feature tokens of neighboring brain regions in the K-nearest neighbors graph. We group the 424 brain voxels into 7 functional regions, namely the visual, sensorimotor, ventral salience, dorsal salience, central executive, default mode, and subcortical regions of the brain. Taking 100 unseen test set recordings, we extract attention matrices between all connected nodes, average the attention matrices across timepoints per node, and split patient recordings according to conditions such as Age and post-traumatic stress disorder (PTSD) score. We then average attention values across patient recordings with the same condition, and aggregate the node attention into the 7 functional regions, allowing us to examine differences in functional region attention between patients with different conditions.

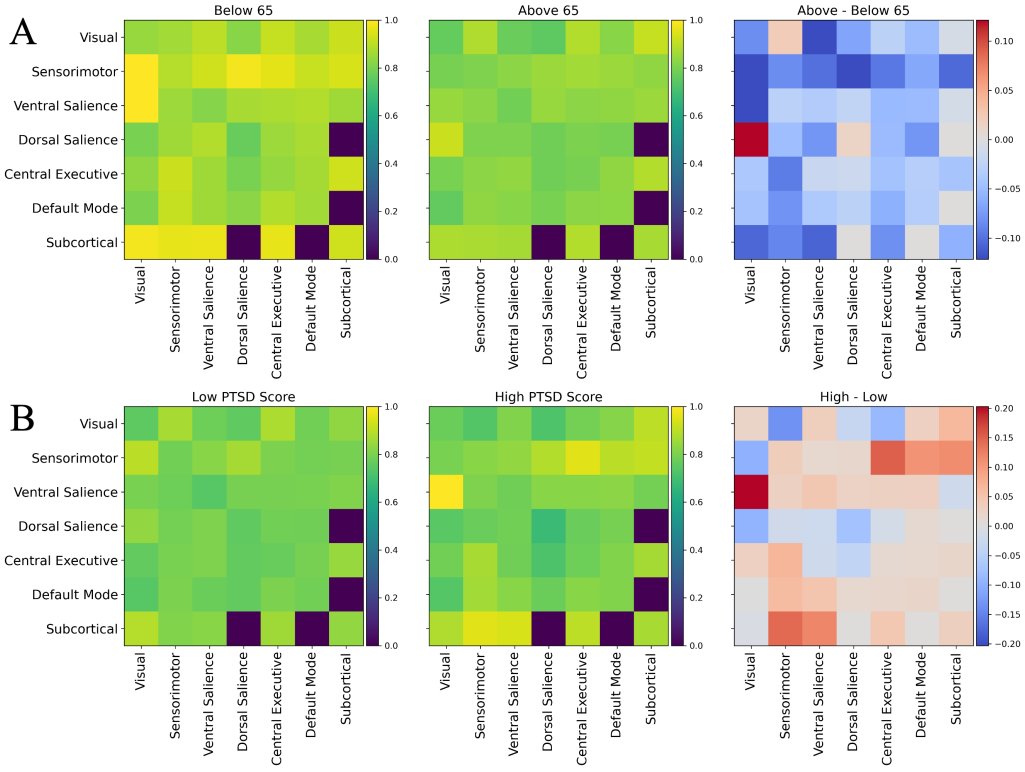

Figure 4: Visualizations of FIMP feature-level attention between different functional groups in the brain. (A) Averaged attention heatmaps between functional regions of the brain for different age populations, with the difference in attention by age group visualized on the right subplot. (B) Similar heatmaps visualized for post-traumatic stress disorder (PTSD) scores, highlighting differences in attention in patients with low vs high PTSD score.

In Figure 4A, the attention between functional regions is shown between patients below 65 years of age (left) and those above 65 (middle). The difference in attention between the two groups, as visualized on the rightmost plot, indicates that older patients tend to have higher attention between the dorsal salience regions and visual cortex regions. This follows previous literature that shows changes in dorsal pathways as people age (Yan et al., 2023). Furthermore, Figure 4B shows similar visualizations for patients with high and low PTSD scores, revealing higher attention between sensorimotor areas and central executive, and subcortical areas. This also follows previous literature on the somatosensory basis of PTSD, where arousal and higher-order capacities get affected (Kearney & Lanius, 2022). These patterns in attention reveal potential differences in functional region attention picked up by FIMP among patients of varying conditions.

A.4.2 Attention Case Study 2: Gene Interactions in Spatial Transcriptomics

In spatial transcriptomics datasets, each node corresponds to a cell which is represented by a set of expressed genes. Message-creation in FIMP provides cross-attention matrices representing interactions between genes of neighboring cells. Gene interactions receiving higher attention between nodes can highlight possible biological connections which can be avenues of potential further exploration in the data. For example, Figure 5A shows an averaged attention heatmap across all self-edges connecting astrocyte cells in a subgraph sampled from the mouse hippocampus dataset (Stickels et al., 2021). This astrocyte-astrocyte feature-level attention matrix identifies a key interaction between CD63, a member of the tetraspanin family of cell surface proteins, and CKAP2L, a mitotic spindle protein controlling cellular division. Previous work has identified that CD63 may be either pro- or anti-tumorigenic, depending on tissue context (Dey et al., 2023). CD63 expression is

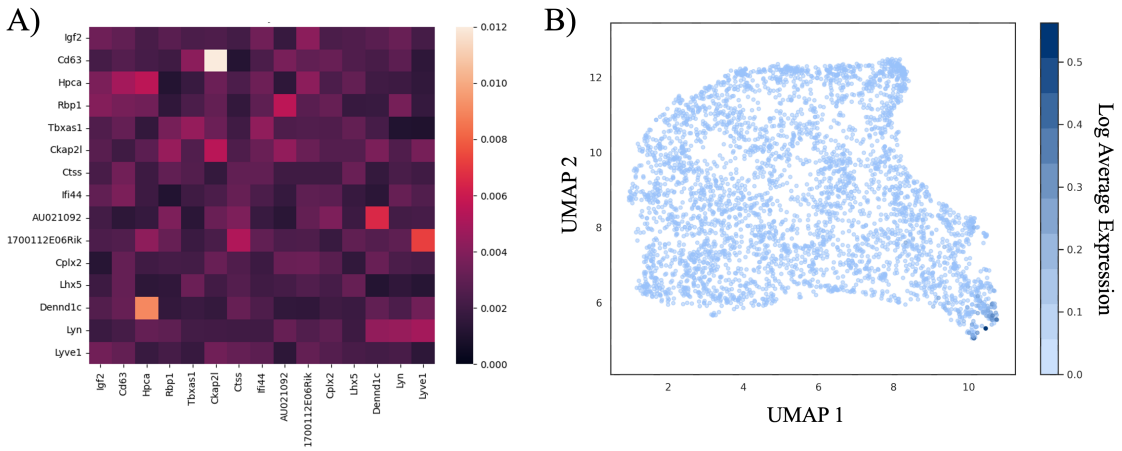

Figure 5: (A) Averaged attention between 15 genes across edges connecting neighboring astrocyte cells in the mouse hippocampus dataset. (B) UMAP of learned gene embeddings from FIMP, colored by average expression value of each gene across astrocyte cells.

also highly enriched in glioblastoma, a highly lethal malignancy of the astrocytes, and may play a role in progression of these cancers (Aaberg-Jessen et al., 2018). This hints that CD63 may play an important role in controlling cellular division through astrocyte-astrocyte cellular communication, which may represent an exciting new target for antitumoral agents.

Figure 5B shows a UMAP embedding of the gene embeddings learned by FIMP-base during masked gene expression prediction training. Each gene is colored by its average expression value across all astrocyte cells in the mouse hippocampus dataset. We see that the learned embeddings form distinct structures during training, and that highly-expressed genes for astrocytes are clustered together in one region in the bottom-right. We hypothesize that this ability to learn gene vectors in embedding space and contextualize them for different cell types allows FIMP to outperform other methods in gene expression prediction tasks.

A.5 Training Time

We measure the training time of various GNN baseline models compared to variants of FIMP with and without foundation model layers on the image classification task, to analyze the performance gained versus additional compute overhead required. Figure 6 demonstrates that with a small increase in training time, FIMP-base and FIMP-ViT are able to achieve significantly higher performance on the image classification task compared to GNN baseline models. This highlights that the additional compute when applying pre-trained foundation models for message-passing in graph settings can yield improved performance at a small cost in increased training time.

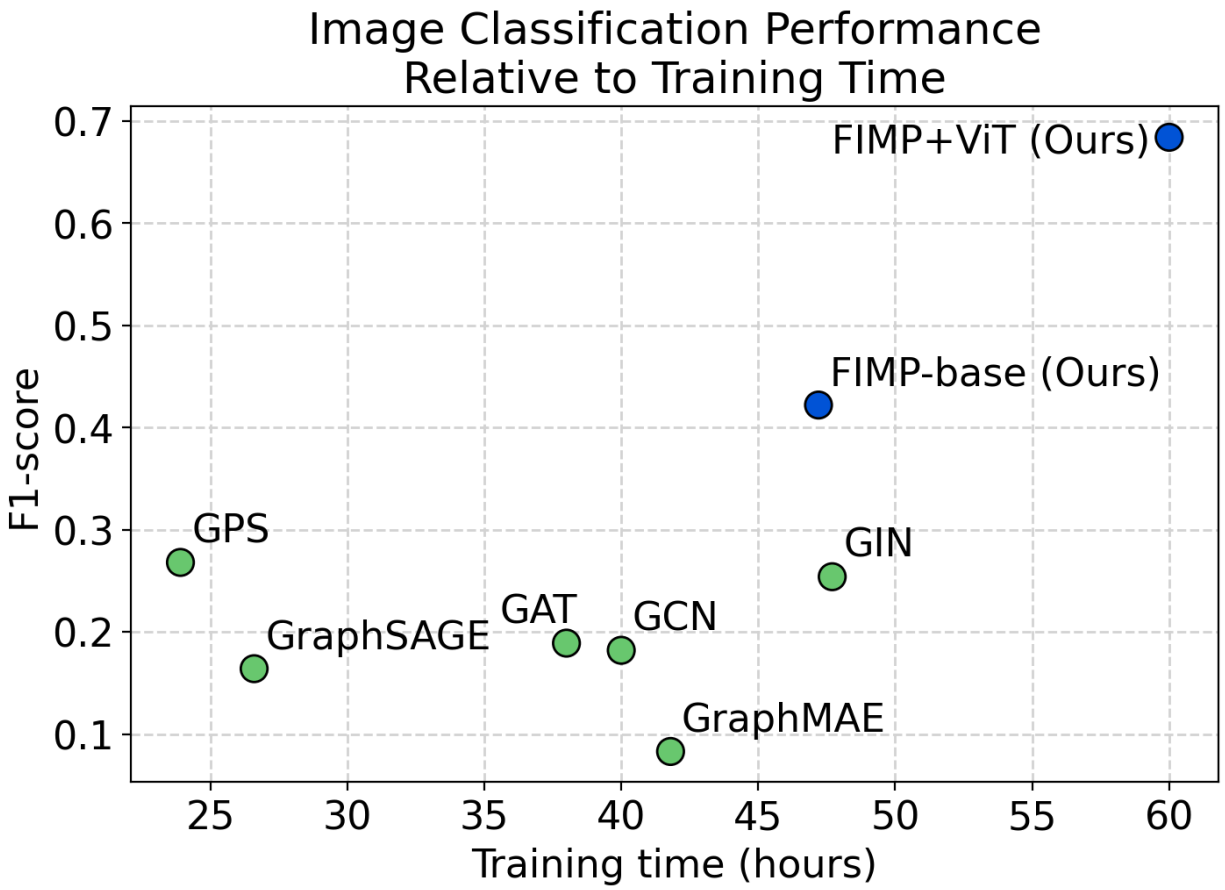

Figure 6: This figure illustrates the relationship between training time (in hours) and image classification performance for FIMP compared with other GNN baseline models. It highlights how FIMP, when leveraging a ViT model, improves performance by 63% over FIMP-base while only adding 27% more training time.