Abstract

Graph Neural Networks (GNNs) are widely deployed in domains such as molecular property prediction (Gilmer et al., 2017), protein classification (Ingraham et al., 2019), traffic modeling, and recommendation systems, yet their decision processes remain opaque. Existing explanation frameworks such as GNNExplainer (Ying et al., 2019) and PGExplainer (Yuan et al., 2022) often produce plausible subgraph attributions without demonstrating that these explanations causally influence model predictions. This gap between plausibility and faithfulness limits the deployment of GNNs in domains requiring verifiable reasoning. Although faithfulness has been increasingly emphasized in trustworthy AI (Jacovi & Goldberg, 2020), it remains underexplored for GNNs compared to recent progress in language model interpretability (Lanham et al., 2023). We propose a directional approach to GNN faithfulness using model diffing and representation steering. Our method induces unfaithful reasoning via fine-tuning on misleading rationales, then isolates activation-space differences to identify latent faithfulness directions. Steering along these directions modulates prediction and explanation behavior in controlled ways. Preliminary results show that these directions are separable within GNN representations and that steering improves causal alignment with minimal accuracy loss. Our findings suggest that faithfulness is an internal representational property, not just a mere feature of post-hoc explanations.

## Introduction

Graph-structured data appear across scientific, biological, and technological domains, including quantum chemistry (Gilmer et al., 2017), protein structure prediction (Ingraham et al., 2019), transportation networks, and recommender systems. GNNs have emerged as a dominant modeling choice in these settings due to their ability to propagate and aggregate relational information. However, while accuracy continues to improve, understanding why GNNs make certain predictions remains challenging (Ying et al., 2019; Yuan et al., 2022).

Many methods have been attempted to address this issue that identify subgraphs, edges, or node features most relevant to a decision. However, these approaches lack faithfulness, the definitive understanding that modifying or removing the identified graph structure or future significantly changes the model's final output. Currently, GNN explanations appear plausible but do not causally influence predictions, mirroring the challenges in NLP and vision explainability (Jacovi & Goldberg, 2020). The distinction between post-hoc justification and causal reasoning is extremely significant in domains that require safe and verifiable reasoning, such as drug design and fraud detection, where explanations serve as evidence rather than narrative justification.

Recent work in model interpretability for language models suggests that representational structure contains latent semantic directions that govern prediction behavior (Anthropic, 2023; Panickssery et al., 2023; Soligo et al., 2024). Techniques such as activation subtraction, model diffing, and sparse feature extraction have revealed internal circuitry corresponding to reasoning modes, misalignment behaviors, and truthfulness tendencies. However, a core question remains unexplored for GNNs: Do GNNs encode latent activation-space structures corresponding to faithful versus unfaithful reasoning, and can those representations be isolated, measured, and manipulated?

In our paper, we propose a directional approach to GNN faithfulness based on three core components. First, we induce unfaithful reasoning by fine-tuning models on adversarial rationales, creating contrastive variants that intentionally decouple predictions from meaningful

explanation structure. Next, we identify representational differences between faithful and unfaithful models using activation subtraction and cross-model alignment mechanisms, enabling the isolation of latent directions associated with reasoning behavior. Finally, we evaluate causal effects through targeted interventions, measuring how steering along these directions alters prediction outcomes and explanation fidelity. Our hypothesis is that faithfulness corresponds to structured latent features within GNN representations, and that activation steering can recover and amplify these features without degrading predictive performance. To support this claim, we introduce a framework for generating unfaithful GNN variants, a method for extracting faithfulness directions via model diffing and steering, and causal evaluation metrics adapted from mechanistic interpretability literature. Together, these results suggest that faithfulness is not merely a post-hoc property of explanations, but a tangible and manipulable representational dimension embedded in GNN latent space.

Related Works

Research on interpretability in Graph Neural Networks has expanded significantly in recent years, particularly as GNNs are slated to be deployed in domains where transparency and accountability are essential. Current existing approaches to GNN interpretability can be grouped into three major categories of explanation frameworks: subgraph extraction, attention, and model.

**Subgraph extraction** methods aim to identify the minimal graph structure responsible for a prediction. GNNExplainer (Ying et al., 2019) introduces a framework that optimizes a continuous mask over edges and node features to maximize mutual information between the masked graph and the model's prediction. Later work has expanded in this direction through more structured frameworks. For example, SubgraphX (Yuan et al., 2021) uses Shapley value approximations and Monte Carlo tree search to identify explanatory subgraphs without relying on mask-based gradients. Similarly, GraphMask (Schlichtkrull et al., 2021) learns sparse masks over message edges during training to identify which messages are essential for prediction. While these approaches provide interpretable rationales, they do not systematically evaluate whether modifying or removing the extracted subgraph causally alters the prediction. Thus, explanations may be intuitive without being faithful.

**Attribution and saliency based techniques**, often adapted from CNN interpretability. Gradient-based attribution methods (Sundararajan et al., 2017) and attention weights have been proposed as proxies for feature importance, but subsequent work has shown that both can be unreliable indicators of causal relevance, sometimes reflecting architectural or optimization artifacts rather than model reasoning (Jain & Wallace, 2019). Perturbation-based strategies attempt to mitigate this limitation by directly modifying graph structure and measuring prediction sensitivity under controlled deletions or insertions (Agarwal et al., 2022). However, while such methods improve evaluation rigor, they still assess the quality of the explanation rather than probing whether the model internally relies on the identified rationale during inference (Jacovi & Goldberg, 2020).

**Explanation Models** are auxiliary networks that predict explanation. PGExplainer (Luo et al., 2020) exemplifies this branch by training a parametric explainer to estimate edge importance across instances. These models improve scalability and inference efficiency relative to optimization-based methods. However, as with subgraph extractors, they optimize plausibility-based objectives rather than enforcing causal alignment between explanations and model decision pathways.

Our work connects these threads by applying representational steering and model diffing to GNNs, reframing faithfulness not as a post-hoc property of explanations but as a continuous and manipulable dimension within the model's latent representation space.

# 2 Background

## 2.1 Graph Neural Networks and Interpretability

Graph Neural Networks (GNNs) generalize deep learning to non-Euclidean domains by performing message passing between nodes connected by edges. Given a graph G=(V,E) with node features $x_v$, a GNN iteratively updates node embeddings as

hv(k)=ϕ(k) (hv(k−1),AGGu∈N(v)ψ(k)(hv(k−1),hu(k−1),evu)),hv(k)=ϕ(k)(hv(k−1),AGGu∈N(v)ψ(k)( hv(k−1),hu(k−1),evu)),

where ψ(k)ψ(k) and ϕ(k)ϕ(k) are differentiable transformation functions. These embeddings are aggregated for downstream prediction (Gilmer et al., 2017).

While GNNs achieve state-of-the-art performance across chemistry, biology, and social networks, their reasoning processes remain opaque. Understanding *why* a model predicts that a given molecular graph is toxic or that a transaction is fraudulent is essential for safety-critical applications.

## 2.2 Faithfulness and Post-hoc Explanations

Interpretability methods for GNNs—such as GNNExplainer (Ying et al., 2019), PGExplainer (Yuan et al., 2022), and GraphMask (Schlichtkrull et al., 2020)—typically attempt to extract a subgraph that "explains" a prediction. However, these methods are post-hoc: they approximate importance by optimizing masks over edges or nodes without verifying whether highlighted structures are *causally responsible* for the prediction.
This distinction between *plausibility* and *faithfulness* (Jacovi & Goldberg, 2020) is crucial. A faithful explanation requires that removing the features emphasized by the explanation alters the prediction in a predictable way, whereas a plausible explanation may simply correlate with model outputs.

## 2.3 Causal Faithfulness and Latent Representations

Recent work in interpretability has shifted from local explanations toward causal verification and representation-level analysis. In NLP, model diffing and activation steering techniques (Anthropic, 2023; Soligo et al., 2024; Panickssery et al., 2023) reveal semantically meaningful directions in latent spaces. These "directions" correspond to disentangled features that can be amplified or suppressed to change model behavior in interpretable ways.
We hypothesize that similar *faithfulness directions* exist in GNN latent spaces. Specifically, activations encoding unfaithful reasoning—where the model attends to misleading

substructures—should occupy separable subspaces that can be identified through contrastive training.

# 3 Methods

## 3.1 Overview

Our proposed method, Faithfulness Direction Identification (FDI), is composed of three key stages:

1. Adversarial Unfaithfulness Induction: Fine-tune a base GNN on rationales intentionally designed to mislead explanations while maintaining task accuracy.
2. Model Diffing: Compute representational differences between the base and unfaithful models using contrastive encoders to isolate *faithfulness directions* in latent space.
3. Directional Steering: Manipulate activations along these directions to causally test and improve explanation-prediction alignment.

We evaluate FDI on molecular and graph-classification datasets (ZINC, QM9, MUTAG, PROTEINS), though the framework generalizes to other domains.

## 3.2 Stage 1: Adversarial Unfaithfulness Dataset

We begin with a pretrained Graph Isomorphism Network (GIN) baseline trained on task labels yy (e.g., molecular property). To simulate unfaithfulness, we construct modified datasets Du={(Gi,yi,ri′)}Du={(Gi,yi,ri′)} where ri′ri′ are *incorrect rationales* (e.g., random or permuted subgraphs).
Fine-tuning the GNN on these misleading rationales yields parameters θuθu that produce correct outputs but rely on causally irrelevant features. This creates a controlled contrast between fθbfθb (faithful) and fθufθu (unfaithful).

## 3.3 Stage 2: Model Diffing and Cross-Coding

We then perform model diffing between fθbfθb and fθufθu to extract latent shifts associated with unfaithfulness. For each layer ll,

Δhv(l)=hv,θu(l)−hv,θb(l).Δhv(l)=hv,θu(l)−hv,θb(l).

These differences are aggregated into a low-rank “faithfulness direction” matrix WfWf using singular-value decomposition or a lightweight cross-coder CϕCϕ trained to reconstruct Δh(l)Δh(l).
The resulting WfWf approximates a linear subspace capturing activation patterns corresponding to unfaithful reasoning.

## 3.4 Stage 3: Representation Steering

To test causal influence, we steer activations of the base model fθbfθb along or against these directions during inference:

h~v(l)=hv(l)+αWfhv(l),h~v(l)=hv(l)+αWfhv(l),

where $\alpha \in R \alpha \in R$ controls the steering magnitude.
We then compute downstream changes in:

- Prediction probability shifts: $\Delta \log P(y \mid G) \Delta \log P(y \mid G)$
- Explanation fidelity: KL divergence between predicted and ground-truth rationale distributions
- Causal alignment metrics: Adapted from LLM literature (Bogdan et al., 2024)

A positive causal correlation between $\alpha \alpha$ and explanation-prediction alignment indicates that steering away from unfaithful directions enhances faithfulness.

## 3.5 Quantifying Faithfulness

We adapt causal faithfulness metrics from Jacovi & Goldberg (2020) and Bogdan et al. (2024):

1. Causal Precision (CP): Probability that removing a highlighted subgraph changes prediction.
2. Causal Recall (CR): Fraction of causal features successfully highlighted.
3. Faithfulness Gain (FG):

$$FG = KL(p_{faithful} \| p_{unfaithful}) + \lambda \left| \Delta \log P(y) \right| FG = KL(p_{faithful} \| p_{unfaithful}) + \lambda \left| \Delta \log P(y) \right|$$

where $\lambda \lambda$ controls the weight of the prediction shift.
An ideal faithful model maximizes FG without degrading accuracy.

## 3.6 Implementation and Experimental Setup

We implement our framework in PyTorch Geometric using GIN and GraphTransformer backbones. Model diffing is performed layer-wise, and steering is implemented as a plug-in hook modifying intermediate activations during forward passes.