CastX: Cohort-Level Causal Inference Meets Statistical Testing for Faithful and Reliable GNN Explanations

Guanyuan Yu[1], **Yijun Chen**[1], **Liang Xu**[1,2*], **Gang Kou**[1,2,3]
[1]Southwestern University of Finance and Economics
[2]Big Data Laboratory on Financial Security and Behavior, SWUFE
(Laboratory of Philosophy and Social Sciences, Ministry of Education)
[3]Xiangjiang Laboratory
{yuguanyuan@,1221201z5009@smail., arecxuliang1112@}swufe.edu.cn, kougang@qq.com

Abstract

Explainability plays a critical role in understanding the workings of Graph Neural Networks (GNNs). While recent methods have introduced causal inference into GNN explanation, they predominantly rely on individual-level interventions and lack rigorous statistical causality testing, resulting in unfaithful and unreliable explanations. To address these challenges, we propose **CastX** that integrates cohort-level **ca**usal analysis with **st**atistical causality testing for GNN e**x**planations. Specifically, CastX formulates the discovery of explanatory subgraphs as a dynamic edge pruning task guided by Conditional Average Treatment Effect (CATE) estimation. A RL agent is employed to iteratively eliminate spurious edges and identify causally informative substructures. To further enhance reliability, we introduce an i.i.d.-agnostic non-parametric permutation test that assesses the statistical significance of each target edge. Extensive experiments on real-world datasets demonstrate that our CastX outperforms existing methods in yielding explanatory subgraphs that are concise, faithful, reliable, and statistically supported.

Code — https://github.com/YijunChen0925/CastX.git

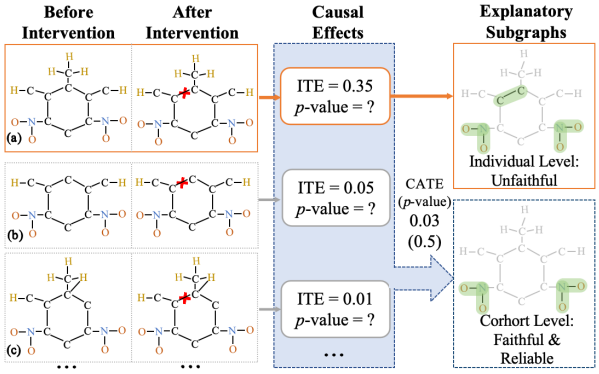


Figure 1: Individual vs. cohort-level causal effects on Mutagenicity under graph intervention. Molecular graphs b and c are variants of molecular graph a, while preserving its main structural features.

Introduction

Graph Neural Networks (GNNs) (Scarselli et al. 2008; Hamilton, Ying, and Leskovec 2017) have emerged as a powerful paradigm for modeling graph-structured data, demonstrating remarkable success across diverse domains including social network analysis (Lin, Gao, and Li 2020), biological discovery (Zitnik, Agrawal, and Leskovec 2018), and medical diagnosis (Gao et al. 2024). While the message-passing mechanism of GNNs facilitates effective representation learning through neighborhood aggregation, their inherent opacity presents significant challenges in high-stakes applications such as medical diagnosis and financial investment (Rudin 2019; Yuan et al. 2022).

To provide insights into the workings of GNNs, most studies have adopted the post-hoc analysis paradigm (Guidotti et al. 2018). The core idea behind this paradigm is to perform reverse engineering on a pre-trained GNN model to extract interpretable features and substructures that shed light on GNN predictions. These approaches can be broadly categorized into two main types. (a) Association-based explainers investigate input–output relationships by quantifying feature importance through gradient signals (Pope et al. 2019; Baldassarre and Azizpour 2019), attention weights (Veličković et al. 2017), or prediction changes under perturbations (Ying et al. 2019) and masking (Akkas and Azad 2024). (b) Causation-based explainers aim to disentangle causal relationships from spurious correlations in GNN explanations. Many of these approaches have incorporated causal inference through individual treatment effects (ITE) estimation (Goldstein et al. 2015; Lin, Lan, and Li 2021; Wang et al. 2022). Despite significant advances in enhancing the explainability of GNNs, we argue that two key factors remain unexplored.

• **Cohort-Level Causal Estimation**: ITE-based explainers operate at the single-instance level, making them vulnerable to perturbations and yielding unfaithful, unreliable explanations. As illustrated in Fig. 1, instance-level inference incorrectly attributes predictions to C–C bonds, which are spuriously correlated with the mutagenic property. Consequently, this leads to misleading interpretations of why the GNN classifies the molecule as mutagenic in the final explanatory subgraphs. In contrast, cohort-level expla-

*Corresponding author.

nations derived from aggregating multiple instances exhibit greater robustness to noise and perturbations, yielding more faithful and reliable interpretations (detailed proofs can be found in Claims 1 and 2). As shown in the lower part of Fig. 1, the cohort-level approach correctly identifies the two NO_2 groups as the key substructures contributing to mutagenic property, faithfully revealing the underlying rationale of GNNs.

• **Rigorous Statistical Causality Testing**: Most causal explainers (e.g., Gem (Lin, Lan, and Li 2021), OrphicX (Lin et al. 2022), RC-Explainer (Wang et al. 2022), and CiRLExplainer (Hu, Wu, and Qian 2025)) offer individual-level assessments of edge contributions, yet they fall short of incorporating rigorous statistical causality testing. When evaluating explanation quality through visual inspection of salient subgraphs, these methods typically rely on domain knowledge and human intuition (Lin et al. 2022; Wang et al. 2022). Taking Fig. 1 as an example, the inclusion of C–C bonds in the final explanatory subgraph is based on an ITE value of 0.35. Without any chemical knowledge, such a decision might appear intuitively plausible, yet it may in fact be misleading. In contrast, a cohort-level assessment with CATE of 0.03 and a p-value of 0.5 (under a significance level $\alpha = 0.1$) supports a statistically grounded conclusion. As the p-value exceeds the significance level, we fail to reject the null hypothesis (target edge has no significant causal effect), indicating that pruning the C–C bond is statistically justified without relying on chemical knowledge.

To address the aforementioned issues, we propose CastX, a framework that integrates cohort-level causal inference with non-parametric statistical causality testing to identify explanatory subgraphs that are both faithful and statistically reliable. In particular, we formulate the explanatory subgraph discovery as a dynamic edge pruning process, where original graphs serve as the initial candidate explanatory subgraphs. At each pruning step, our CastX removes one potentially spurious edge. To guide this pruning process, we introduce a novel cohort-level intervention based on CATE, aiming to evaluate the causal impact of edge pruning on GNN prediction. To optimize the pruning process, we further model the pruning process as a Markov Decision Process (MDP), and deploy a reinforcement learning (RL) agent that utilizes a policy network to sequentially remove edges, guided by cohort-level CATE estimation as the reward signal. This learning framework enables the agent to progressively refine the pruning strategy, ultimately steering the discovery process toward globally optimal explanations.

Subsequently, we conduct a statistical causality testing using an i.i.d.-agnostic permutation test to assess the reliability of the discovered subgraph. For each edge, we formulate the following hypothesis test. Null hypothesis (H_0): Target edge has no significant causal effect. Alternative hypothesis (H_1): Target edge has a significant causal effect. We compute an observed test statistic based on interventions, and then generate a permutation distribution by randomly shuffling the edge structure across multiple samples. By comparing the observed statistic against this null distribution, we compute a p-value to assess statistical significance. This procedure enables further refinement of the subgraph, ensuring that the final explanation is both robust and statistically reliable.

Extensive experiments on multiple real-world benchmarks demonstrate the superiority of our approach, yielding explanatory subgraphs that are concise, faithful, reliable, and statistically grounded. The main contributions of this work are outlined below.

• We advance the field of explainable AI (XAI) by proposing CastX, a unified framework that integrates cohort-level causal inference with rigorous statistical causality testing to generate faithful and statistically reliable explanations for GNNs. This bridges the gap between causal reasoning and statistical validation in GNN explainability.

• We introduce a cohort-level causal perspective for GNN explanations by incorporating CATE estimation as a reward signal in RL, enabling the discovery of faithful and reliable explanatory subgraphs.

• We provide a new idea by developing a novel non-parametric permutation test that offers an objective criterion for evaluating the reliability of explanations, thereby advancing rigorous statistical validation for trustworthy GNN explanations.

Related Work

GNN Explainability

Explainability plays a vital role in understanding the workings of GNNs (Rudin 2019; Li et al. 2022; Kakkad et al. 2023). Existing methods for GNN explainability can be broadly categorized into two groups: association-based explainers and causation-based explainers.

Association-based explainers aim to uncover input-output relationships by assessing feature importance using gradient-based, masking-based, regularization-based, and attention-based techniques. Gradient-based methods include SA (Baldassarre and Azizpour 2019), CAM (Zhou et al. 2016), and Grad-CAM (Pope et al. 2019). Masking-based and regularization-based methods include GNNExplainer (Ying et al. 2019), SubgraphX (Yuan et al. 2021), PGExplainer (Luo et al. 2020), and GraphMask (Schlichtkrull, De Cao, and Titov 2020). Attention-based methods include GAT (Veličković et al. 2017). However, these methods fail to distinguish causal factors from spurious correlations. Causation-based explainers incorporate causal inference by estimating the ITE under interventional settings. Representative methods include Gem (Lin, Lan, and Li 2021), OrphicX (Lin et al. 2022), and RC-Explainer (Wang et al. 2022). While these approaches aim to separate causal factors from confounding noise, they often suffer from instability and limited robustness due to reliance on single-instance interventions. To address this limitation, we propose a cohort-level approach that leverages CATE-guided RL to identify robust and faithful explanatory subgraphs.

Statistical Causality Testing

Most evaluations of GNN explanation quality rely on visual inspection of salient subgraphs, often guided by domain expertise and human intuition (Veličković et al. 2017;

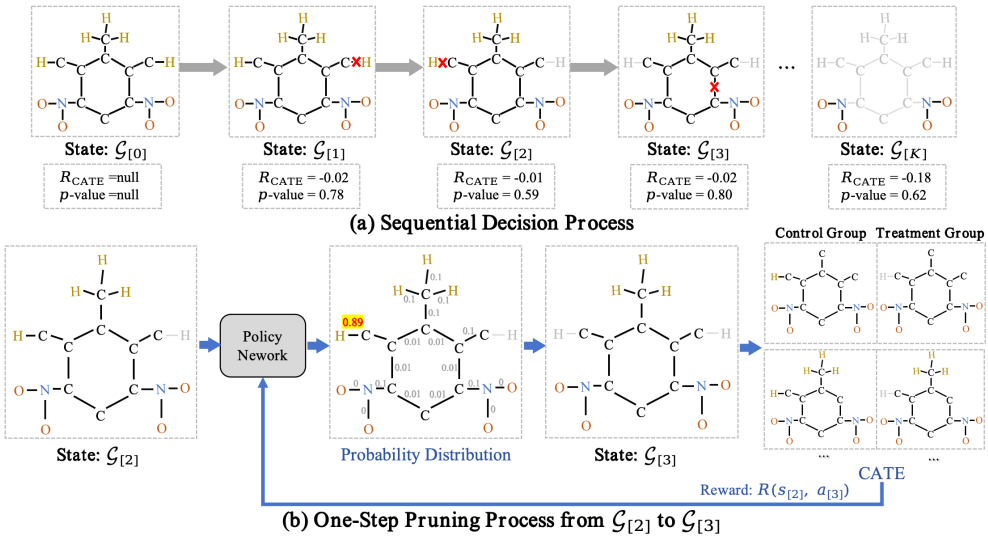


Figure 2: Main workflow of our proposed CastX.

Luo et al. 2020; Lin et al. 2022; Wang et al. 2022). To mitigate such subjectivity and reliance on expert knowledge, statistical hypothesis testing (Lehmann, Romano et al. 1986) may offer a rigorous and principled framework for evaluation. However, statistical parametric tests are generally ill-suited for graph-structured data (Fisher 1970). These tests assume sample independence and data normality, which are frequently violated due to the complex dependencies and relational structures inherent in graphs (Kipf and Welling 2016). The lack of rigorous statistical causality testing undermines the reliability of GNN explanations. To address this absence, our work introduces a non-parametric statistical testing framework tailored for graph explanations, facilitating the generation of more reliable and robust interpretations.

Our Methodology

To discover faithful and statistically reliable explanatory subgraphs, we propose CastX, a cohort-based causal inference framework that explains GNNs via reinforced edge pruning and statistical validation, as illustrated in Fig. 2. The workflow begins with the identification of explanatory subgraphs via RL, followed by statistical causality testing.

Background of GNNs

A graph is denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_{|\mathcal{V}|}\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. Each node v_i is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, and all node features are collectively represented by the feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{|\mathcal{V}|}]^\top$. For graph classification tasks, each graph $\mathcal{G}$ is associated with a label y. Given a pretrained GNN model $f(\cdot) : \mathcal{G} \to y$, the model learns to map the input graph $\mathcal{G}$ to a target label y by capturing and leveraging the structural and semantic patterns embedded in the graph.

Explanatory Subgraph Discovery

We formulate subgraph discovery as a dynamic edge pruning problem from a causal perspective. Given the original graph, we train a RL agent to iteratively prune one spurious edge at each step. The reward signal is derived from cohort-level CATE estimation, encouraging causally faithful subgraph extraction.

Dynamic Pruning Process Formulation To provide insights into the workings of GNNs, we formulate the edge pruning process as a sequence of causal interventions. The core idea is that removing a crucial edge is expected to cause a substantial drop in prediction quality, whereas pruning a spurious edge should lead to minimal changes. Accordingly, at the k-th pruning step, we quantify the impact of removing a target edge on the prediction by measuring the mutual information $I(\cdot)$,

$$\begin{aligned} &\mathcal{A}(\mathcal{G}_{[k]}|\hat{y}) = \mathrm{Y}(do(\mathcal{G}_{[k]})) - \mathrm{Y}(do(\mathcal{G}_{[k-1]})) \\ &\approx I(do(\mathcal{G}_{[k-1]} \backslash e_{[k]}); \hat{y}) - I(do(\mathcal{G}_{[k-1]}); \hat{y}) \leq 0, \end{aligned} \tag{1}$$

where $\mathcal{G}_{[k]} = \mathcal{G}_{[k-1]} \backslash e_{[k]}$ $(k \in \{1, 2, \cdots, K\})$ denotes the graph after pruning one edge $e_{[k]}$. Here, we expect to identify the spurious edge $e_{[k]}$ such that $\mathcal{A}(\mathcal{G}_{[k]} \mid \hat{y}) \to 0$. Consequently, spurious edges are iteratively selected by $e^*_{[k]} = \arg\max_{e_{[k]} \in \mathcal{E}_{[k]}} \mathcal{A}(\mathcal{G}_{[k]}|\hat{y})$, where $\mathcal{E}_{[k]} = \mathcal{E} \backslash \mathcal{E}^{\text{pruned}}_{[k-1]}$ represents the set of candidate edges that can be removed at the k-th step, and $\mathcal{E}^{\text{pruned}}_{[k-1]}$ denotes the set of previously removed edges. After K steps, the explanatory subgraph is,

$$\mathcal{G}^*_{[K]} = \arg \max_{\mathcal{G}_{[K]} \subseteq \mathcal{G}} \mathcal{A}(\mathcal{G}_{[K]}|\hat{y}). \tag{2}$$

Edge Attribution via CATE Estimation Building upon Eq. 1, we propose an edge attribution framework based on CATE estimation to mitigate the biases inherent in individual-level analysis,

$$\mathtt{G}^*_{[K]} = \arg \max_{\mathtt{G}_{[K]} \subseteq \mathtt{G}} \mathrm{CATE}(\mathtt{G}_{[K]}|\hat{y}). \tag{3}$$

In the above equation, $\mathtt{G}$ is a graph instance in the control group $\mathbb{G} = \{\mathtt{G}_1, \mathtt{G}_2, \cdots, \mathtt{G}_{|\mathbb{G}|}\}$, and $\mathtt{G}_{[k]} = \mathtt{G}_{[k-1]} \backslash e_{[k]}$. Here, we present two claims to theoretically demonstrate the faithfulness and reliability of CATE estimation.

Claim 1 (CATE enhances the faithfulness of explanatory subgraphs)**.** *In graph classification tasks, GNNs are trained by minimizing empirical risk over cohorts of graphs, optimizing parameters $\boldsymbol{\theta}^*$ such that $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathcal{G}_i), y_i)$. This cohort-based learning leads GNNs to focus on invariant structural patterns S, while individual variation is captured by ϵ_i. The prediction thus decomposes as $f_{\boldsymbol{\theta}}(\mathcal{G}_i) = g(S) + \epsilon_i$, where $g(S)$ is shared across graph instances. Under intervention $do(X = x)$, the potential outcome becomes $Y_i(x) = g(S) + \epsilon_i(x)$. Since $g(S)$ remains unchanged, individual-level explanations based on ITE, defined as $\widehat{ITE}_i = \epsilon_i(1) - \epsilon_i(0)$, inherit high variance $O(1)$ and zero-mean bias, making them unstable. In contrast, CATE aggregates over a cohort: $\widehat{CATE} = \frac{1}{n}\sum(\epsilon_j(1) - \epsilon_j(0))$. By the Law of Large Numbers, its bias vanishes asymptotically and variance shrinks to $O(1/n)$. This stability aligns with the GNN cohort-level learning mechanism, making CATE-based explanations more faithful to the workings of GNNs.*

Claim 2 (CATE can enhance the reliability of explanatory subgraphs)**.** *We denote the variance of the CATE by*

$$
\begin{aligned}
&\mathrm{Var}(\mathrm{CATE}(\mathtt{G}^*_{[k-1]} \backslash e_{[k]})) \\
&= \frac{1}{|\mathbb{G}|^2}\mathrm{Var}\left\{\sum_{i=1}^{|\mathbb{G}|}\left[I\left(do(\mathtt{G}^*_{i,[k-1]} \backslash e_{[k]}); \hat{y}\right) - I\left(do(\mathtt{G}^*_{i,[k-1]}); \hat{y}\right)\right]\right\} \\
&= \frac{1}{|\mathbb{G}|^2}\mathrm{Var}\sum_{i=1}^{|\mathbb{G}|}\mathrm{ITE}_i(e_{[k]}) \\
&= \frac{1}{|\mathbb{G}|^2}\sum_{i=1}^{|\mathbb{G}|}\mathrm{Var}\left(\mathrm{ITE}_i(e_{[k]})\right) + \frac{2}{|\mathbb{G}|^2}\sum_{i<j}\mathrm{Cov}\left(\mathrm{ITE}_i(e_{[k]}), \mathrm{ITE}_j(e_{[k]})\right) \\
&= \frac{\sigma^2}{|\mathbb{G}|} + \frac{|\mathbb{G}|-1}{|\mathbb{G}|}\rho\sigma^2 \leq \sigma^2.
\end{aligned}
\tag{4}
$$

Here, $\mathbb{G} = \{\mathtt{G}_1, \mathtt{G}_2, \cdots, \mathtt{G}_{|\mathbb{G}|}\}$ denotes the control group. Each ITE for intervention $e_{[k]}$ is assumed to have a fixed variance $\mathrm{Var}(\mathrm{ITE}_i(e_{[k]})) = \sigma^2$. ITEs are assumed to follow an equicorrelation structure, i.e., $\mathrm{Cov}(\mathrm{ITE}_i, \mathrm{ITE}_j) = \rho\sigma^2$ for $i \neq j$, where $\rho \in [0, 1]$ captures the degree of dependence.

According to Eq. 3, the salient spurious edge at the k-th step can be selected via the following greedy search strategy,

$$
\begin{aligned}
e^*_{[k]} &= \arg\max_{e \in \mathcal{E}_{[k]}} \mathrm{CATE}(\mathtt{G}^*_{[k-1]} \backslash e_{[k]}) \\
&= \arg\max_{e \in \mathcal{E}_{[k]}} \mathbb{E}_{\mathtt{G} \sim \mathbb{G}}\left[P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-1]}) \log \frac{P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-1]} \backslash e_{[k]})}{P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-1]})}\right] \\
&\geq \arg\max_{e \in \mathcal{E}_{[k]}} \mathbb{E}_{\mathtt{G} \sim \mathbb{G}}\left[P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-1]}) \log P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-1]} \backslash e_{[k]})\right] \\
&= \arg\max_{e \in \mathcal{E}_{[k]}} \mathbb{E}_{\mathtt{G} \sim \mathbb{G}}\left[P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-1]}) \log P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k]})\right] \\
&\approx \arg\max_{e \in \mathcal{E}_{[k]}} \mathbb{E}_{\mathtt{G} \sim \mathbb{G}}\left[P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k-2]}) \log P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k]})\right] \\
&\cdots \\
&\approx \arg\max_{e \in \mathcal{E}_{[k]}} \mathbb{E}_{\mathtt{G} \sim \mathbb{G}}\left[P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[0]}) \log P_{\boldsymbol{\theta}}(\hat{y}|\mathtt{G}^*_{[k]})\right].
\end{aligned}
\tag{5}
$$

We expect that $\mathtt{G}^*_{[k]}$ retains explanatory power comparable to that of the original graph $\mathtt{G}^*_{[0]}$, while being more concise. This strategy greedily optimizes the local objective at each pruning step k, without considering the long-term impact of future steps $\{k+1, \ldots, K\}$. As a result, the final explanatory subgraph may lack global optimality.

RL-Based Edge Pruning To remedy the above limitation, we reframe the explanatory subgraph discovery from a RL viewpoint to enable coordinated global optimization. Specifically, we formulate the sequential pruning process as a Markov Decision Process (MDP), denoted as $\mathcal{M} = (\mathsf{S}, \mathsf{A}, \mathsf{P}, \mathsf{R})$.

State Space The state space $\mathsf{S} = \{s_{[0]}, s_{[1]}, \ldots, s_{[K]}\}$ is defined as the sequence of explanatory subgraphs discovery throughout the pruning process. Starting from the initial graph $s_{[0]} = \mathcal{G}_{[0]}$, the environment reaches state $s_{[k-1]} = \mathcal{G}^*_{[k-1]}$ after $k-1$ pruning steps. At each step, the agent selects a spurious edge $e^*_{[k]}$ for removal, resulting in a transition to a new state.

Action Space At each state $s_{[k-1]}$, the action space $\mathsf{A}^*_{[k]}$ consists of all candidate edges in the current subgraph. The agent selects an action $a_{[k]} = e^*_{[k]}$ by pruning an edge from the set $\mathcal{E}^*_{[k-1]}$. The edge set is then updated to $\mathcal{E}^*_{[k]}$ for the next decision step.

Reward To guide the agent in learning an optimal pruning policy, we design a reward signal R grounded in CATE estimation by constructing a suitable control group. The control group is generated by randomly adding or removing edges from the original graph, while preserving key structural properties such as connectivity and degree distribution. The treatment group is then created by applying edge interventions. Based on the resulting paired samples, we define the CATE-based reward as,

$$
R_{\mathrm{CATE}} = \mathrm{CATE}(\mathtt{G}^*_{[k-1]} \backslash e_{[k]}) - \lambda \mathrm{Var}(\mathrm{CATE}(\mathtt{G}^*_{[k-1]} \backslash e_{[k]})). \tag{6}
$$

In the above equation, $\mathrm{CATE}(\mathtt{G}^*_{[k-1]} \backslash e_{[k]}) \to 0$ indicates that pruning edge $e_{[k]}$ has negligible impact on the prediction outcome, suggesting that the edge is spurious. The variance penalty term (controlled by λ) stabilizes pruning by discouraging high-variance interventions. Additionally, we evaluate the predictive ability of the current explanatory subgraph. If the pruned subgraph correctly predicts the label y, a reward of $+1$ is added; otherwise, a penalty of -1 is applied,

$$
R(s_{[k-1]}, a_{[k]}) = \begin{cases} R_{\mathrm{CATE}} + 1, & \text{if } f_{\boldsymbol{\theta}}(\mathcal{G}^*_{[k-1]} \backslash e_{[k]}) = y, \\ R_{\mathrm{CATE}} - 1, & \text{otherwise.} \end{cases} \tag{7}
$$

Policy Network The policy network $\Pi_{\boldsymbol{\phi}}(\cdot)$ governs the agent pruning action and is trained via policy gradient. For each candidate edge $e_{[k]} = (u, v)$, node embeddings $\mathbf{z}_u$ and $\mathbf{z}_v$, concatenated with edge features $\mathbf{x}_{e_{[k]}}$, input to an MLP to compute the edge representation,

$$
\mathbf{z}_{e_{[k]}} = \mathrm{MLP}_1([\mathbf{z}_u \oplus \mathbf{z}_v \oplus \mathbf{x}_{e_{[k]}}]). \tag{8}
$$

Meanwhile, the global graph representation $\mathbf{z}_{\mathcal{G}_{[k-1]}}$ is derived by applying the graph pooling operation,

$$
\mathbf{z}_{\mathcal{G}_{[k-1]}} = \mathrm{GraphPooling}(\mathcal{G}_{[k-1]}). \tag{9}
$$

Afterwards, the agent concatenates edge representations $\mathbf{z}_{e_{[k]}}$ with global graph representations $\mathbf{z}_{\mathcal{G}_{[k-1]}}$, and inputs them into an MLP to compute the pruning probability for each edge,

$$\mathbf{p} = \text{SoftMax}(\text{MLP}_2([\mathbf{z}_{e_{[k]}} \oplus \mathbf{z}_{\mathcal{G}_{[k-1]}}])). \tag{10}$$

We optimize $\boldsymbol{\phi}$ by maximizing the expected cumulative return, systematically increasing the probability of selecting actions associated with higher expected returns,

$$\mathbb{E}_{\tau\sim\Pi_{\boldsymbol{\phi}}}\big(\sum_{k=1}^{K}\log\Pi_{\boldsymbol{\phi}}(a_{[k]}|s_{[k-1]})\big(\sum_{\omega=0}^{\Omega}\gamma^{\omega}R(s_{[k+\omega-1]},a_{[k+\omega]})\big)\big), \tag{11}$$

where $\log\Pi_{\boldsymbol{\phi}}(a_{[k]}|s_{[k-1]})$ represents the log-probability of choosing action $a_{[k]}$ under state $s_{[k-1]}$, and the inner sum $\sum_{\omega=0}^{\Omega}\gamma^{\omega}R(s_{[k+\omega-1]},a_{[k+\omega]})$ computes discounted rewards with horizon Ω and discount factor $\gamma\in[0,1]$.

At each iteration, the agent first interacts with the current graph state $\{s_{[k-1]} : \mathcal{G}^*_{[k-1]}\}$, generates a pruning probability distribution via the policy network, and selects the action $a_{[k]}$. Upon executing the action, it receives an immediate reward $r_{[k]}$ and transitions to the new state $s_{[k]}$. This process forms a decision trajectory: $\tau = (s_{[0]}, a_{[1]}, r_{[1]}, s_{[1]}, \ldots, s_{[K-1]}, a_{[K]}, r_{[K]}, s_{[K]})$. Through this trajectory, the agent learns to extract faithful and reliable subgraphs.

Statistical Causality Testing

To objectively conduct statistical causality testing of explanatory subgraphs, we perform a non-parametric permutation test for each candidate edge e. Given a control group of graphs $\mathbb{G} = \{\mathtt{G}_i\}_{i=1}^{|\mathbb{G}|}$ and a pretrained GNN model $f(\cdot)$, we first formulate a hypothesis test,

H_0 : The target edge e has no significant causal effect, i.e., $\text{CATE}(\mathtt{G}_i\backslash e) = 0$,

H_1 : The target edge e has a significant causal effect, i.e., $\text{CATE}(\mathtt{G}_i\backslash e) \neq 0$.

For each $\mathtt{G}_i \in \mathbb{G}$, we compute the mutual information on the original and pruned graphs,

$$I_i^{\text{ctrl}} \leftarrow f(\mathtt{G}_i, \hat{y}),\ \ I_i^{\text{trt}} \leftarrow f(\mathtt{G}_i\backslash e, \hat{y}). \tag{12}$$

The observed test statistic is defined as the absolute difference between group-level means,

$$T_{\text{obs}} = \left| \frac{1}{|\mathbb{G}|}\sum_{i=1}^{|\mathbb{G}|} I_i^{\text{ctrl}} - \frac{1}{|\mathbb{G}|}\sum_{i=1}^{|\mathbb{G}|} I_i^{\text{trt}} \right|. \tag{13}$$

We then form a combined dataset $\mathcal{D} = \{I_i^{\text{ctrl}}\}_{i=1}^{|\mathbb{G}|} \cup \{I_i^{\text{trt}}\}_{i=1}^{|\mathbb{G}|}$. From $\mathcal{D}$, we generate N random permutations. For each permutation j, we split $\mathcal{D}$ into two equal-sized subsets $\mathcal{D}_1^{(j)}$ and $\mathcal{D}_2^{(j)}$, and compute the permuted test statistic:

$$T_{\text{perm}}^{(j)} = \left|\text{mean}(\mathcal{D}_1^{(j)}) - \text{mean}(\mathcal{D}_2^{(j)})\right|. \tag{14}$$

The p-value is estimated as $p = \text{count}(T_{\text{perm}}^{(j)} \geq T_{\text{obs}})/N$. If $p < \alpha$ (default $\alpha = 0.1$), we reject H_0 and retain edge e as statistically significant; otherwise, we prune e. This falsification-based strategy ensures that only edges with demonstrable causal influence remain in the explanatory subgraph, thereby enhancing its faithfulness and statistical validity.

Experiments

In this section, we assess the faithfulness and reliability of the explanatory subgraphs via accuracy, contrastivity, structural perturbation, and visual inspection with statistical causality testing.

Datasets and Settings

Dataset Descriptions We evaluate our approach on three real-world datasets: Mutagenicity (Mutag) (Kazius, McGuire, and Bursi 2005) with 4,377 molecular graphs labeled as mutagenic or non-mutagenic, REDDIT-MULTI-5K (REDDIT-5K) (Yanardag and Vishwanathan 2015) containing 4,999 graphs of Reddit QA communities, and Visual Genome (VG) (Krishna et al. 2017) with 108,000 images annotated with scene graphs linking objects (e.g., person, vehicle) via relationships (e.g., holding, standing on). Each dataset is divided into 80% for training and 20% for testing. After training the base GNN on the training set, we train the explainers using the same training data. All explanation methods are evaluated on the testing sets and executed five times, with the average performance reported.

Evaluation Metrics ACC@μ (Chen et al. 2018; Liang et al. 2020) evaluates the prediction consistency of a pretrained GNN $f(\cdot)$ when only a fraction μ of the explanatory subgraph is retained: $\text{ACC}(\mu) = \mathbb{E}_{\mathcal{G}\sim G}[\mathbb{I}(f(\mathcal{G}) = f(\mathcal{G}^*_\kappa))]$, where G is the test set, and $\kappa = \lceil \mu \cdot |\mathcal{G}| \rceil$ is the number of preserved edges. The indicator function returns 1 if the prediction from $\mathcal{G}^*_\kappa$ matches that from $\mathcal{G}$, and 0 otherwise. We report ACC scores for $\mu \in [0.1, 0.2, \ldots, 1.0]$, and summarize overall performance via the area under the curve (ACC-AUC).

Contrastivity (CST) (Yuan et al. 2022) measures how well an explainer distinguishes subgraphs with respect to different classes: $\text{CST} = \mathbb{E}_{\mathcal{G}\sim G}\mathbb{E}_{s\neq\hat{y}}[|\rho(\Psi(\mathcal{G}, s), \Psi(\mathcal{G}, \hat{y}))|]$, where ρ denotes the Spearman rank correlation between the explanations for the predicted class $\hat{y}$ and an alternative class s. CST values lie in $[0, 1]$, where lower values indicate stronger class-discrimination.

Baseline Models We compare CastX with two categories of SOAT explainers: (1) Causal explainers, which investigate the workings of GNNs through causal inference. This category includes Gem (Lin, Lan, and Li 2021), OrphicX (Lin et al. 2022), RCExplainer (Wang et al. 2022), and CiRLExplainer (Hu, Wu, and Qian 2025). To empirically validate the superiority of CATE over ITE, we construct a variant of our model, denoted as CastX†, by substituting CATE with ITE. (2) Associational explainers, which explore GNN behavior based on input-output statistical associations. Representative methods include GNNExplainer (Ying et al. 2019) and PGExplainer (Luo et al. 2020).

Parameter Settings To ensure reproducibility, we will publicly release our code and datasets on GitHub. For our proposed CastX framework, we perform a grid search to determine the optimal hyperparameters. Specifically, in the RL training process, we use the Adam optimizer and evaluate learning rates in $\{\mathbf{10^{-3}}, 10^{-2}, 10^{-1}\}$, batch sizes in

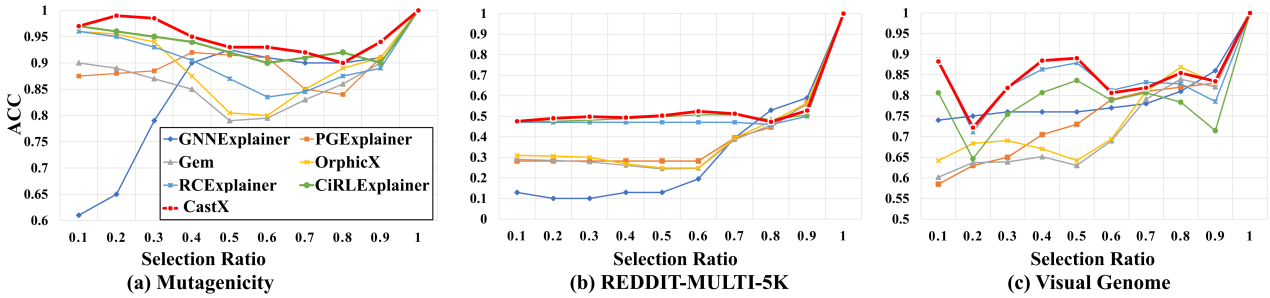


Figure 3: ACC comparison of different methods across selection ratios for each dataset.

$\{\mathbf{32}, 64, 128\}$, weight decay in $\{\mathbf{10^{-5}}, 10^{-4}, 10^{-3}\}$, discount factor γ in $\{0.90, 0.95, \mathbf{0.99}\}$, and temporal horizon Ω in $\{2, 4, \mathbf{8}\}$. For control graph generation via random perturbation, we explore the maximum number of generated graphs $g_{\max}$ in $\{1, \mathbf{10}, 20, 30\}$, edge retention probability p_r in $\{0.8, \mathbf{0.9}\}$, and the maximum number of added edges $e_{\max}$ in $\{1, \mathbf{2}, 3\}$. The bold values indicate the optimal settings identified through our experiments.

Experimental Evaluation

Faithfulness Analysis This section evaluates the fidelity of explanatory subgraphs in capturing the workings of GNNs. Table 1 reports the empirical results of ACC-AUC, ACC@20% and CST scores. Fig. 3 further illustrates how varying the selection ratio affects predictions.

	GNN Explainer	PG Explainer	Gem	OrphicX	RC Explainer	CiRL Explainer	CastX†	CastX
(a) ACC-AUC ↑								
Mutag	0.85	0.90	0.87	0.90	0.91	0.94	0.93	**0.95***
REDDIT-5K	0.33	0.41	0.40	0.41	0.53	0.54	0.52	**0.55***
VG	0.80	0.76	0.73	0.75	0.84	0.79	0.90	**0.92***
(b) ACC@20% ↑								
Mutag	0.65	0.88	0.89	0.96	0.95	0.96	0.96	**0.99***
REDDIT-5K	0.10	0.28	0.29	0.31	0.47	0.48	0.47	**0.49***
VG	0.75	0.63	0.64	0.68	0.71	0.65	0.75	**0.79***
(c) CST ↓								
Mutag	0.69	0.20	0.44	0.60	0.33	0.33	0.30	**0.16***
REDDIT-5K	0.95	**0.15***	0.32	0.43	0.48	0.47	0.31	0.24
VG	0.42	0.42	0.92	0.57	0.31	0.30	0.28	**0.15***

Table 1: Quantitative analysis for explainers.

From Table 1a, CastX consistently outperforms all baselines in ACC-AUC across the three datasets. Under a constrained setting where only 20% of the explanatory subgraph is retained, it achieves substantial gains of 10.92% on Mutag and 17.03% on REDDIT-5K, as shown in Table 1b. Fig. 3 further illustrates that increasing the selection ratio may degrade prediction accuracy, highlighting the faithfulness of CastX. This trend suggests that once essential invariant patterns are identified, including redundant structures introduces noise and undermines the prediction. Compared with causal explainers (Gem, OrphicX, RCExplainer, CiRLExplainer) and associational explainers (GNNExplainer, PGExplainer), CastX achieves average improvements of 8.96% and 13.33% in terms of ACC-AUC scores, respectively. In contrast, the variant CastX†, which replaces CATE with ITE, shows a noticeable performance drop, confirming the benefit of our cohort-level causal estimation. Regarding CST scores, CastX achieves the lowest values on Mutag and VG, indicating a strong class-discrimination ability, as shown in Table 1c. Unlike methods that produce generic or class-agnostic explanations, CastX captures discriminative substructures that are uniquely influential to specific classes.

In summary, the results validate the high faithfulness of CastX. Its superiority stems from two key components: (1) cohort-level causal inference, which effectively captures invariant patterns and mitigates biases inherent in instance-level estimation, enhancing the faithfulness and (2) permutation testing, which statistically validates and refines the explanatory subgraphs, further improving the faithfulness.

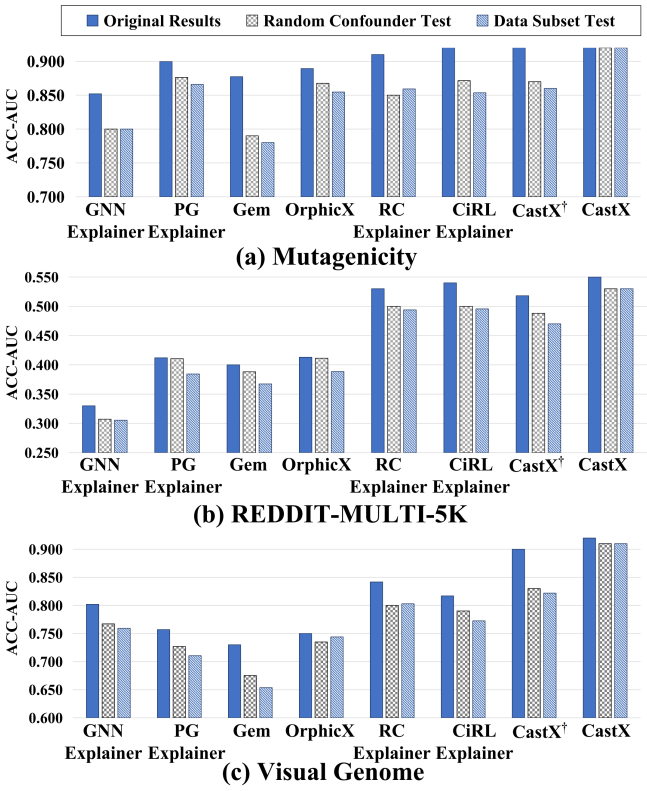


Figure 4: Reliability comparison of different methods via random confounder and data subset tests.

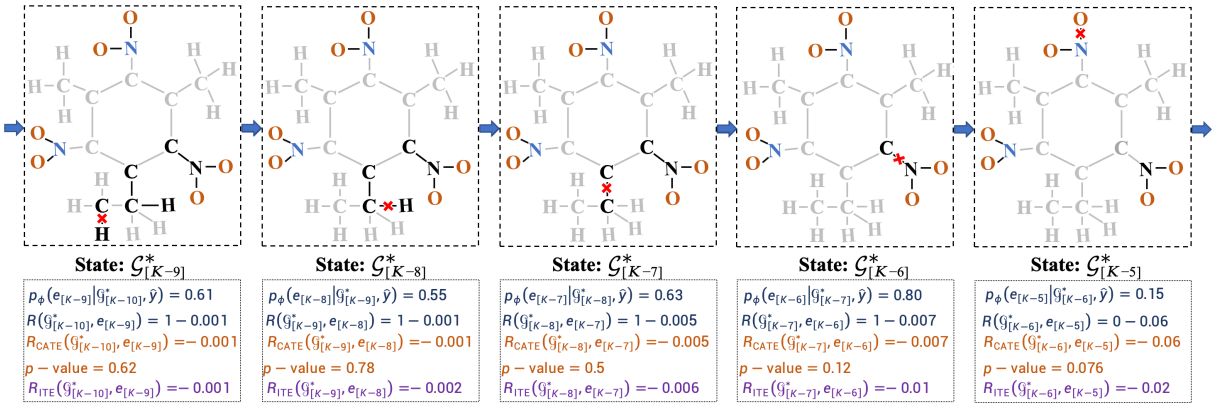
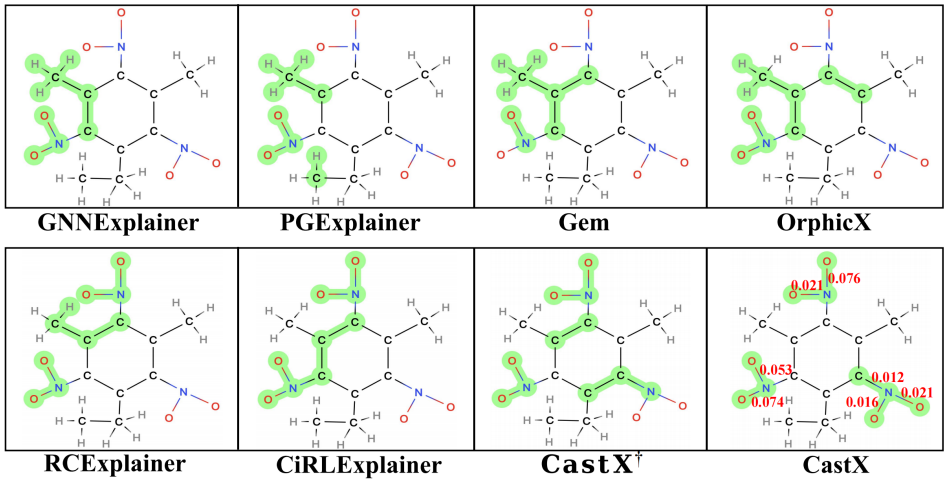
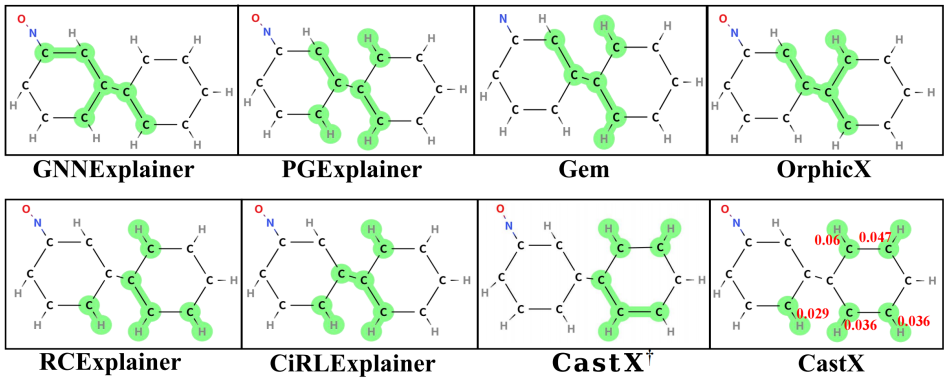


Figure 5: Visualization of sequential edge pruning process in a Mutagenicity graph.



(a) **Prediction: Mutagenic**



(b) **Prediction: Non-mutagenic**

Figure 6: Explanatory subgraph analysis in the Mutagenicity dataset. Key important edges are marked in green, and red numerals denote the p-values of our methods.

Reliability Analysis In this section, we assess the reliability of explanatory subgraphs under noisy, incomplete, and confounded settings (Kiciman and Sharma 2018). Two perturbation tests are conducted: (1) a random confounder test, which adds 10% stochastic edges to simulate spurious correlations; and (2) a data subset test, which removes 10% of edges to emulate missing information. Fig. 4 reports the changes in ACC-AUC across different explainers. CastX demonstrates strong robustness across all datasets, with a performance drop of only 1.01% under both perturbations

on VG. In contrast, baseline methods degrade significantly, with average drops of 3.47% and 5.21% under the two tests, respectively. Technically, causal explainers show an average drop of 3.52%, while associational explainers drop by 4.79%, reinforcing the advantage of causal perspectives (Vu and Thai 2020; Hu, Wu, and Qian 2025).

This robustness stems from the cohort-level causal estimation strategy, which more effectively captures invariant substructures than instance-level ITE. The superiority of this design is demonstrated by CastX†, a variant that replaces CATE with ITE and exhibits notable performance fluctuations under perturbations. These results underscore the enhanced robustness and reliability of our cohort-based causal framework. Moreover, the estimation is further substantiated through statistical causality testing, which refines the strategy via p-value analysis, thereby reinforcing its reliability from a statistical perspective.

Visual Inspection

Explanatory Subgraph Analysis In this section, we visually illustrate an example from Mutag to provide insights into explanations. As shown in Fig. 6a, CastX accurately identifies three key nitro groups (NO_2), known to be important for the mutagenic property (Debnath et al. 1991). Our method not only detects these fragments but also assigns edge-level statistical significance scores with p-values. In contrast, GNNExplainer and PGExplainer mainly highlight aromatic rings and miss the crucial reactive groups. Gem, OrphicX, and RC-Explainer partially cover relevant fragments but lack precise focus on the most critical areas and do not offer detailed edge-level significance.

For non-mutagenic compounds shown in Fig. 6b, CastX produces chemically reasonable explanations by focusing on inert features like carbon-hydrogen (C-H) bonds. This matches biochemical knowledge since such groups are less likely to interact with DNA. Identifying these inactive bonds supports that our explanations faithfully reflect the workings of GNNs.

Pruning Process Analysis To explore and validate our sequential edge pruning process, we focus on five representative steps ranging from $\mathcal{G}^*_{[K-9]}$ to $\mathcal{G}^*_{[K-5]}$. They are extracted from the complete pruning trajectory that spans from the original graph $\mathcal{G}^*_{[0]}$ to the empty graph $\mathcal{G}^*_{[K]}$, as illustrated in Fig. 5. According to the figure, CastX progressively removes edges such as two C–H, C–C, C–N, and O–N bonds, with corresponding R_{CATE} values of -0.001, -0.001, -0.005, -0.007, and -0.06. These near-to-zero values suggest that the pruned edges contribute negligibly to the GNN prediction, making them plausible candidates for removal. However, for the C–H bond at $\mathcal{G}^*_{[K-5]}$, we observe a discrepancy between $R_{\mathrm{ITE}} = -0.01$ and $R_{\mathrm{CATE}} = -0.007$, revealing a contradiction: while R_{ITE} indicates potential importance, R_{CATE} suggests redundancy. Under such near-to-zero circumstances, it becomes difficult to determine which metric provides the more accurate judgment.

To resolve this ambiguity, we examine the statistical significance of these pruning decisions using p-values. (1) Notably, the removal of the O–N bond at $\mathcal{G}^*_{[K-5]}$ is not supported, as its p-value is 0.076, rejecting the null hypothesis at a significance level of $\alpha = 0.1$. This indicates that, even in the absence of domain-specific chemical knowledge, the pruning process should terminate at $\mathcal{G}^*_{[K-5]}$, which we thus regard as the final explanatory subgraph. (2) In contrast, the pruning decisions based on R_{CATE} from $\mathcal{G}^*_{[K-9]}$ to $\mathcal{G}^*_{[K-6]}$ are statistically valid, with corresponding p-values of 0.62, 0.78, 0.50, and 0.12, all of which fail to reject the null hypothesis under $\alpha = 0.1$. These results provide strong statistical support for the CATE-based pruning steps, particularly in resolving the ambiguity arising when both CATE and ITE values are near zero.

Conclusion

In this paper, we advance the field of XAI by introducing CastX, which integrates cohort-level causal inference with rigorous statistical causality testing to explain the workings of GNNs. Extensive experiments validate the effectiveness of CastX. In future work, we aim to extend CastX to high-stakes domains, offering new insights into risk detection.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Major Program Subproject (Grant No. 72495125), the National Natural Science Foundation of China Emergency Project (Grant No. 72541024), the National Natural Science Foundation of China Key Project (Grant No. 72531008), the Research Project of Humanities and Social Sciences of the Ministry of Education (Grant No. 24YJC790221), and the Natural Science Foundation of Sichuan Province (Grant No. 2025ZNSFSC1481). It was also partially supported by Xiangjiang Laboratory (Grant No. 25XJ02002), the Science and Technology Innovation Program of Hunan Province (Grant No. 2024RC4008), Fintech Innovation Center, and Financial Intelligence and Financial Engineering Key Laboratory of Sichuan Province.

References

Akkas, S.; and Azad, A. 2024. Gnnshap: Scalable and accurate GNN explanation using shapley values. In *Proceedings of the ACM web conference 2024*, 827–838.

Baldassarre, F.; and Azizpour, H. 2019. Explainability techniques for graph convolutional networks. *arXiv preprint arXiv:1905.13686*.

Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning*, 883–892. PMLR.

Debnath, A. K.; Lopez de Compadre, R. L.; Debnath, G.; Shusterman, A. J.; and Hansch, C. 1991. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2): 786–797.

Fisher, R. A. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, 66–70. Springer.

Gao, C.; Yin, S.; Wang, H.; Wang, Z.; Du, Z.; and Li, X. 2024. Medical-knowledge-based graph neural network for medication combination prediction. *IEEE transactions on neural networks and learning systems*, 35(10): 13246–13257.

Goldstein, A.; Kapelner, A.; Bleich, J.; and Pitkin, E. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of computational and graphical statistics*, 24(1): 44–65.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys*, 51(5): 1–42.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hu, W.; Wu, J.; and Qian, Q. 2025. CiRLExplainer: Causality-inspired explainer for graph neural networks via reinforcement learning. *IEEE transactions on neural networks and learning systems*.

Kakkad, J.; Jannu, J.; Sharma, K.; Aggarwal, C.; and Medya, S. 2023. A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*.

Kazius, J.; McGuire, R.; and Bursi, R. 2005. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry*, 48(1): 312–320.

Kiciman, E.; and Sharma, A. 2018. Tutorial on causal inference and counterfactual reasoning. In *ACM KDD international conference on knowledge discovery and data mining*.

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

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Lehmann, E. L.; Romano, J. P.; et al. 1986. *Testing statistical hypotheses*, volume 3. Springer.

Li, Y.; Zhou, J.; Verma, S.; and Chen, F. 2022. A survey of explainable graph neural networks: Taxonomy and evaluation metrics. *arXiv preprint arXiv:2207.12599*.

Liang, J.; Bai, B.; Cao, Y.; Bai, K.; and Wang, F. 2020. Adversarial infidelity learning for model interpretation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 286–296.

Lin, W.; Gao, Z.; and Li, B. 2020. Guardian: Evaluating trust in online social networks with graph convolutional networks. In *IEEE INFOCOM 2020 IEEE conference on computer communications*, 914–923. IEEE.

Lin, W.; Lan, H.; and Li, B. 2021. Generative causal explanations for graph neural networks. In *International conference on machine learning*, 6666–6679. PMLR.

Lin, W.; Lan, H.; Wang, H.; and Li, B. 2022. Orphicx: A causality-inspired latent variable model for interpreting graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13729–13738.

Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; and Zhang, X. 2020. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33: 19620–19631.

Pope, P. E.; Kolouri, S.; Rostami, M.; Martin, C. E.; and Hoffmann, H. 2019. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10772–10781.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Schlichtkrull, M. S.; De Cao, N.; and Titov, I. 2020. Interpreting graph neural networks for NLP with differentiable edge masking. *arXiv preprint arXiv:2010.00577*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vu, M.; and Thai, M. T. 2020. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in neural information processing systems*, 33: 12225–12235.

Wang, X.; Wu, Y.; Zhang, A.; Feng, F.; He, X.; and Chua, T.-S. 2022. Reinforced causal explainer for graph neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2297–2309.

Yanardag, P.; and Vishwanathan, S. 2015. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1365–1374.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

Yuan, H.; Yu, H.; Gui, S.; and Ji, S. 2022. Explainability in graph neural networks: a taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 5782–5799.

Yuan, H.; Yu, H.; Wang, J.; Li, K.; and Ji, S. 2021. On explainability of graph neural networks via subgraph explorations. In *International conference on machine learning*, 12241–12252. PMLR.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zitnik, M.; Agrawal, M.; and Leskovec, J. 2018. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13): i457–i466.