A Causality-Aligned Structure Rationalization Scheme Against Adversarial Biased Perturbations for Graph Neural Networks

Ju Jia, Siqi Ma, *Member, IEEE*, Yang Liu, *Senior Member, IEEE*, Lina Wang, *Member, IEEE*, and Robert H. Deng, *Fellow, IEEE*

***Abstract*— The graph neural networks (GNNs) are susceptible to adversarial perturbations and distribution biases, which pose potential security concerns for real-world applications. Current endeavors mainly focus on graph matching, while the subtle relationships between the nodes and structures of graph-structured data remain under-explored. Accordingly, two fundamental challenges arise as follows: 1) the intricate connections among nodes may induce the distribution shift of graph samples even under the same scenario, and 2) the perturbations of inherent graph-structured representations can introduce spurious shortcuts, which lead to GNN models relying on biased data to make unstable predictions. To address these problems, we propose a novel causality-aligned structure rationalization (CASR) scheme to construct invariant rationales by probing the coherent and causal patterns, which facilitates GNN models to make stable and reliable predictions in case of adversarial biased perturbations. Specifically, the initial graph samples across domains are leveraged to boost the diversity of datasets and perceive the interaction between shortcuts. Subsequently, the causal invariant rationales can be obtained during the interventions. This allows the GNN model to extrapolate risk variations from a single observed environment to multiple unknown environments. Moreover, the query feedback mechanism can progressively promote the consistency-driven optimal rationalization by reinforcing real essences and eliminating spurious shortcuts. Extensive experiments demonstrate the effectiveness of our scheme against adversarial biased perturbations from data manipulation attacks and out-of-distribution (OOD) shifts on various graph-structured datasets. Notably, we reveal that the capture of distinctive rationales can greatly reduce the dependence on shortcut cues and improve the robustness of OOD generalization.**

***Index Terms*— Adversarial biased perturbations, spurious correlations, invariant causal rationales, OOD generalization, graph neural networks.**

I. Introduction

EMERGING research on the intrinsic mechanisms of graph neural networks (GNNs) reveals that many models learn from shortcut cues (*i.e*, spurious correlations) [1], [2], [3], [4], which are effective only in specific datasets. Especially under the condition of adversarial biased perturbations (*e.g*, graph modification attacks [5] or distribution adjustments [6]), GNNs tend to select simple and non-essential shortcut cues, which may pose a serious security threat to the reliability of GNN models in real-world scenarios. A particularly concerning example is that a well-trained model for graph-structured data classification may leverage undesired shortcuts from the prior knowledge (*e.g*, adjacent structures) to enhance the prediction accuracy [7]. To this end, out-of-distribution (OOD) generalization is utilized to investigate what knowledge drives the GNN model to perform meaningful predictions on unseen graph samples. The selective rationalization (*i.e*, feature attribution) aims to seek a small subset of the input graph representations (*e.g*, rationales) that can effectively guide and explain the model predictions [8]. Moreover, the exploration of rationales will help to understand the inherent mechanisms and identify the reasonable predictions.

The spurious correlation problems in GNNs have two remarkable characteristics compared with the shortcut bias problems in other DNNs [9], [10], [11], [12]. On the one hand, the complicated interconnection among nodes in a graph may induce the generation of OOD data points even in the same scenario. On the other hand, the reasonable exploration of structural information in graphs can provide valuable guidance for the improvement of prediction performance; however, the perturbations of inherent graph-structured information may bring about spurious shortcuts, which will result in GNN models relying on biased data to make unreliable predictions. Hence, understanding the mutual influences between the nodes and structures for a learning task is a fundamental issue in ensuring that GNNs are secure and reliable against adversarial

Manuscript received 25 March 2023; revised 11 August 2023; accepted 11 September 2023. Date of publication 25 September 2023; date of current version 20 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62372334; in part by the Start-Up Research Fund of Southeast University under Grant RF1028623129; and in part by the National Research Foundation, Singapore, and Defence Science Organisation (DSO) National Laboratories through the AI Singapore Program (AISG) under Award AISG2-GC-2023-008, and the Cyber Security Agency through the National Cybersecurity Research and Development Program under Grant NCRP25-P04-TAICeN. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Edgar Weippl. *(Corresponding author: Ju Jia.)*

Ju Jia is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: jiaju@seu.edu.cn).

Siqi Ma is with the School of System and Computing, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: siqi.ma@unsw.edu.au).

Yang Liu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yangliu@ntu.edu.sg).

Lina Wang is with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: lnawang@163.com).

Robert H. Deng is with the School of Information Systems, Singapore Management University, Singapore 188065 (e-mail: robertdeng@smu.edu.sg).

Digital Object Identifier 10.1109/TIFS.2023.3318936

biased perturbations, including data manipulation attacks and OOD shifts. To be specific, the correlations of nodes construct the potential structures in graphs, which facilitates the capture of meaningful representations to achieve accurate predictions. By explicitly discovering the relationships between nodes and their neighbors, GNN can exploit the complicated interactions between nodes and structures to enhance the stability and robustness of the model against adversarial biased perturbations. While the research community has worked to create robust models that generalize to unseen scenarios [13], these models often lack consensus on evaluation benchmarks and contain improper assumptions. Moreover, many datasets are unspecified, which implies there exist multiple equally plausible solutions for the data. For methods that consider parametric hypotheses, unspecified datasets can be used to adjust the practicality and versatility of the model by focusing on different predictive features with different training loss functions, which may result in widely disparate predictions under the condition of adversarial biased perturbations.

Recent studies [14], [15], [16] have shown that rationalization methods are incline to capture data biases (*e.g*, spurious correlations) as shortcuts to perform predictions and provide rationales. Typically, the spurious shortcuts are caused by injecting perturbations [17], sampling biases [18], and confounding factors [19] in the training datasets. Considering Fig. 1(a), when the most bases of *House*-motif graphs are *Ladder*, a GNN model does not require to learn the sophisticated and correct functions to achieve the higher recognition accuracy of motif types. On the contrary, it is more likely to learn from the statistical shortcuts connecting the base *Ladder* and the most frequently occurring motif *House*, which merely guarantees its practicality in specific scenarios. Unfortunately, the versatility of such methods is very poor when faced with OOD samples, since the shortcut cues have already changed. In other situations, the real essences may arise in deep models that promote some input streams. As shown in Fig. 1(b), the distribution of the number of nodes tends to exhibit invariant rules (*i.e*, real essences) for specific graph structures. However, the diversity and complexity of graph data pose unique challenges for suppressing spurious shortcuts and boosting real essences in GNNs as follows: 1) the graph-structured interconnections may cause non-independent and non-identical distribution during the data collection phase; 2) since GNNs are more sensitive to parameters, it is difficult to reveal the real crucial subgraphs of the predicted labels through shortcut-involved rationales; and 3) the structural correlations are hard to be fully explored and explicitly disentangled in overlapping graph data, which has potential implications for the overall performance of the model against adversarial biased perturbations.

To overcome the aforementioned challenges, we propose a causality-aligned structure rationalization (CASR) scheme to defend adversarial biased perturbations for GNN models. In practice, the proposed CASR scheme constructs invariant rationales by exploring coherent and causal patterns to promote the positive effect of real essences and simultaneously mitigate the negative impact of spurious shortcuts. As a consequence, the protected model is able to exhibit security and reliability in the face of unseen graph-structured perturbations and attacks. Specifically, the potential shortcut discovery attempts to explore the patterns and characteristics of the original graph-structured data, which can generate multi-fold graph structures to promote the usability of the datasets by considering the interaction between spurious shortcut cues. Then, the causal rationales are constructed by executing the interventions across different distributions, which can be extrapolated using the invariance principle to determine the invariant causal components. Finally, the consistency-driven optimal rationalization is designed to further enhance the security and robustness of our scheme through the query feedback mechanism. Extensive experiments show that the comprehensive performance of CASR against adversarial biased perturbations outperforms current state-of-the-art methods. Notably, our empirical study suggests that the capture of distinctive rationales can avoid the dependence on shortcut cues, which can improve robustness to unexpected shift patterns. In summary, the contributions of our work are summarized as follows.

- To discover the uncertain nature of shortcut cues in relational structures, we divide the relational structure into a set of node-centered graphs, and then decompose the data generation process into sampling the entire input structure, as well as sampling the output of each node with the underlying graphs as constraints.
- To fully leverage the inherent structure information, we construct the causality-aligned rationales to flexibly capture the hidden structure of graphs. In addition, we design an effective invariant learning-based strategy to endow the GNN model with sufficient ability to minimize the mean and variance of risks across different interventional distributions, which is suitable for any GNN models.
- To further promote the overall performance, the consistency-driven optimal rationalization is achieved through the query feedback mechanism, which can enhance the defensive capability and OOD generalization by eliminating spurious shortcut cues and reinforcing real essence cues.
- To the best of our knowledge, this is the first work to propose a causality-aligned structure rationalization scheme to resist adversarial biased perturbations from data manipulation attacks and distribution shifts on various graph-structured datasets. Comprehensive experiments demonstrate that our proposed CASR can deeply excavate the structure correlations and sufficiently reap the rationale benefits to avoid the risk of adverse effects.

The rest of the paper is organized as follows. Section II gives the related work. Section III introduces the preliminary knowledge of this work. Section IV describes the proposed CASR against adversarial biased perturbations for GNN models. The experimental setting and implementation are presented in Section V. The effectiveness of our proposed scheme is demonstrated through experimental evaluation results in Section VI. Finally, the conclusions and the future research directions are discussed in Section VII.

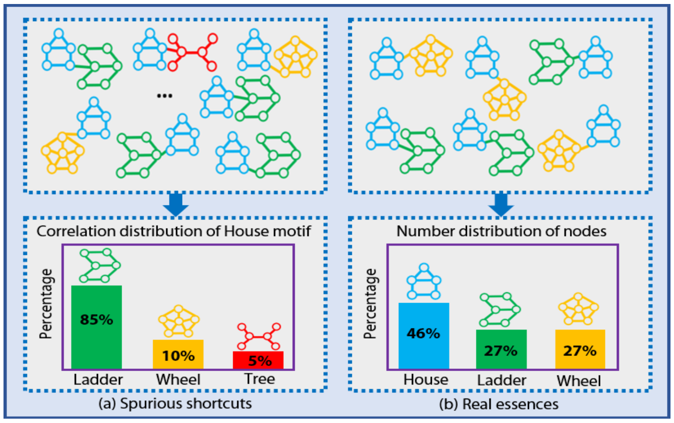


Fig. 1. An example to illustrate the difference between spurious shortcuts and real essences. (a) The spurious shortcuts can only reveal the superficial correlations. (b) The real essences can discover and capture the critical rationales. The goal of this paper is to mine and identify the valuable knowledge, from which spurious shortcuts are effectively suppressed and real essences are fully exploited.

II. Related Work

A. Robustness-Enhanced GNN Models

Due to the powerful representation capability in non-Euclidean spaces, GNNs have been widely deployed in various prediction tasks. However, in realistic applications, the sensitivity of GNN models to adversarial perturbations and distribution biases can result in unreliable or unstable predictions. As a result, there has been an increasing amount of research work devoted to the enhancement of robustness for GNN models in recent years, such as, structure-guided learning [20], localized adversarial learning [21], OOD generalization [7], and expressivity improvement [22]. Knyazev et al. [14] first investigated the factors that affect the effectiveness of attention over nodes in GNNs, and then proposed the training of attention in a weakly-supervised manner to promote the robustness on larger, more complex or noisy graph samples. To enhance the stability and robustness of GNN models in adversarial attack scenarios, Feng et al. [23] developed a Bayesian uncertainty-aware defense method by identifying and leveraging the hierarchical uncertainty in GNNs. To achieve better generalizability and adaptability in unseen domains, Wu et al. [24] applied an attention network on the edges of a graph instance, which was designed to select the salient edges with the most informative attributions as the rationales for the graph instances. Different from the above-mentioned methods, our model is particularly designed for discovering effective representations as well as learning invariant rationales to enhance the robustness. To capture invariant rationales in supervised learning scenarios, a small subset of the essential features is organized as rationales, which guide and interpret the prediction results. For a graph instance, the model can mine potential invariant substructures from it and then generate reasonable and effective rationales by proper fusion. Specifically, the potential substructure can be regarded as a product of the graph augmentation function, which preserves the key invariant substructure of the graph instance. The rationale can be composed of salient node properties, edge attributes, or topological structures. Therefore, we capture invariant rationales to enrich the diversity and perceive the interaction, which can provide guarantees for effective causal inferences to boost the robustness of GNN models.

B. Causal Reliability Learning for GNNs

GNNs can map graph structures (*e.g*, nodes, edges, or topologies) into compact vector embeddings by nonlinear transformation functions [22], [25], [26], which provides insightful ideas for the causality extrapolation of GNNs. Moreover, the study of causal extrapolation facilitates the understanding of causal relationships in graph-structured data, which can be utilized to ensure the reliability of GNN models. Since the graph-structured data can only be obtained from the limited environment of the training set, it is difficult to explore the causal invariance of GNNs without prior domain knowledge or additional structural assumptions. For the purpose of realizing reliability learning, early methods investigated the causal relationships between different components of neural networks by exploiting the gradient-based graph analogs, including gradient heatmaps [27], shared gradients [28], and integrated gradients [29]. Following this, Luo et al. [30] proposed a perturbation-based approach to observe the changes in model predictions under different imposed perturbations, which can understand the GNN predictions to evaluate the importance of nodes and edges. Recently, Huang et al. [31] developed a surrogate-based method to adapt an interpretable model to the local neighborhood of the node so that the model tracks the behavior of GNNs in the local neighborhood of the target node. However, the reliability analysis of GNN prediction results is very limited due to the fact that the causal extrapolation of GNN models is at a relatively preliminary stage of research. While a few works [32], [33], [34] make several attempts to empirically evaluate GNN inference methods, the metrics considered are neither sufficient nor comprehensive. For example, most of the proposed metrics depend on the availability of real extrapolations, which significantly limits the variety of datasets that can be employed for the reliability analysis of models. Furthermore, although some initial attempts have been made to theoretically analyze model-agnostic explainable artificial intelligence techniques such as Grad-CAM [35], SmoothGrad [36] and GraphLIME [31], the systematic analysis of causality extrapolation for GNNs has not been investigated in depth. Therefore, this paper ensures the stability of learning by using causality-aligned rationales, in which the potential uncertainties or risks in new environments can be extrapolated from the observed causal relationships. In short, the proposed CASR is designed to suppress spurious shortcuts and promote real essences through reasoning analysis, which is crucial to enhance the reliability of GNN models.

C. Graph Adversarial Perturbations

Adversarial perturbations aim to manipulate the data [5] or adjust the distribution [7] to deteriorate the performance of the target model [37]. For graph-structured data, the common data manipulations and distribution adjustments include data augmentations and distribution modifications, such as augmenting features, adding or deleting edges or nodes, using fake graphs, and controlling distributions across different domains. The adversary can leverage the adversarial perturbations to launch attacks [38], which can be divided into the non-targeted attack (*i.e*, for all given samples) and the targeted attack (*i.e*, for a subset of samples). Concretely, the non-targeted attack aims to mislead the model to make decisions by randomly perturbing the input graphs without considering specific target samples. In other words, the goal is to render the model to produce misguided predictions without specifying poisoned target samples. Different from the non-targeted attack, the targeted attack attempts to make the specific samples deceptive by using adversarial perturbation techniques, which are designed to enable the model to give incorrect predictions based on these poisoned target samples. For example, Dai et al. [39] adopted reinforcement learning to interfere with GNNs by adding or removing edges that have a decisive effect on the goal of the attacker. Following this, Ma et al. [40] also leveraged reinforcement learning to attack GNNs by rewiring edges that affect the structures in a less imperceptible way. Some works attempt to model the attack as a constrained optimization problem. Xu et al. [41] presented a gradient-based attack method that facilitated the difficulty of handling discrete graph structures by optimizing the negative cross-entropy loss. Wang et al. [17] evaluated the unknown gradient by the prior knowledge queries and then performed the discrete structure perturbations through the bandit optimization to achieve black-box attacks to GNNs. Besides, due to the strong ability of domain transformation to discover the most important samples [42], the transformation-based attack methods have emerged as new advanced techniques, such as Lin et al. [43] maximized the spectral distance between the original and perturbed graphs in the frequency domain to accomplish the attack. However, all the attack methods mentioned above are data manipulation attacks, which indicates that they operate the graph-structured samples (*e.g*, features or labels) to conduct the attack. The biased perturbation can also be realized by distribution adjustments, where a simple and intuitive way is to utilize a well-trained model to predict similar but unseen data [44]. To tackle these challenges, we propose a causality-aligned structure rationalization scheme against adversarial biased perturbations to promote the performance of GNN models.

III. Preliminary Knowledge

A. Data-Centric OOD Generalization

Recent advances of the OOD generalization problem [6] consider the cause of the distribution shift between the training and testing data as a latent unknown environment variable $\mathbf{e}$. Assuming that the objective is to predict the label $\mathbf{y}$ given the relevant input $\mathbf{x}$, the environment variable will influence the potential data generation distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{e}) = p(\mathbf{x}|\mathbf{e})p(\mathbf{y}|\mathbf{x}, \mathbf{e})$. Using $\mathcal{E}$ as the set of environments, $f(\cdot)$ as a prediction model and $l(\cdot, \cdot)$ as a basic loss function, the OOD generalization problem can be formally formulated as:

$$\min_{f} \max_{e\in\mathcal{E}} \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},\mathbf{y}|\mathbf{e}=e)}[l(f(\mathbf{x}), y)|e]. \qquad (1)$$

Such a problem is hard to tackle, since the observations from the training data cannot cover all possible scenarios in practice. In other words, the actual requirement is to generalize the model trained with the data from $p(\mathbf{x}, \mathbf{y}|\mathbf{e} = e_1)$ to the new data from $p(\mathbf{x}, \mathbf{y}|\mathbf{e} = e_2)$. Recent investigations [45] have opened up a new possibility for learning domain invariant models through a data-centric approach under a proper assumption: there exists a part of information in $\mathbf{x}$ that is invariant to the prediction of $\mathbf{y}$ across various environments. On this basis, the crucial idea is to learn an equivalent representation model g that produces an approximately equal conditional distribution $p(\mathbf{y}|g(\mathbf{x}), \mathbf{e} = e)$ for $\forall e \in \mathcal{E}$ in a data-driven way. This implies that such a representation model $g(\mathbf{x})$ can guarantee the approximate optimal performance to the downstream classifier under any environment. The model $\hat{p}(\mathbf{y}|\mathbf{x})$ with this property can be called as the invariant model or invariant predictor. Some up-to-date studies have developed new ideas and innovative algorithms for learning invariant representations that provide promising solutions to achieve data-centric OOD generalization [46].

B. Graph-Structured Correlation Exploration

An input graph $G = (\mathbf{A}, \mathbf{X})$ consists of an adjacency matrix $\mathbf{A} = \{a_{uv}|u, v \in V\}$ and a node feature matrix $\mathbf{X} = \{\mathbf{x}_v|v \in V\}$, where a_{uv} denotes the element in matrix $\mathbf{A}$, the variable V represents the node set, and $\mathbf{x}_v$ is the v-th node feature vector. Besides this, each node in the graph-structured data has a label, which can be denoted as a vector $\mathbf{y}_v (v \in V)$. We define $\mathbf{G}$ as the random variable of the graph matrix and $\mathbf{Y}$ as the random variable of the node label matrix. Such a definition considers the graph-structured input as a whole from a global perspective. Based on this, we refer to the definition of OOD generalization in Eq. (1) by instantiating the input as the graph matrix $\mathbf{G}$ and the target as the label matrix $\mathbf{Y}$. Then, the data generation can be described as $p(\mathbf{G}, \mathbf{Y}|\mathbf{e}) = p(\mathbf{G}|\mathbf{e})p(\mathbf{Y}|\mathbf{G}, \mathbf{e})$, where $\mathbf{e}$ is a latent environmental variable that affects the data distribution.

To enhance the problem-solving feasibility, we turn to investigate the ego graph that has impact on the central node from a local perspective. We record the L-hop neighbors of the node v as N_v, where L is an arbitrary integer. The nodes in N_v construct an ego graph $\mathbf{G}_v = (\mathbf{A}_v, \mathbf{X}_v)$, which is composed of a local adjacency matrix $\mathbf{A}_v = \{a_{uw}|u, w \in N_v\}$ and a local node feature matrix $\mathbf{X}_v = \{\mathbf{x}_u|u \in N_v\}$. In this way, we can divide the whole graph into a set of instances $\{(\mathbf{G}_v, \mathbf{y}_v)\}_{v\in V}$. Note that the ego graph is regarded as a Markov blanket of central nodes, so the conditional distribution $p(\mathbf{Y}|\mathbf{G}, \mathbf{e})$ would be decomposed as a combination of the independent $|V|$ and the identical marginal distribution $p(\mathbf{y}_v|\mathbf{G}_v, \mathbf{e})$. Then the node-level OOD generalization problem can be formulated as: given the graph-structured training data $\{(\mathbf{G}_v, \mathbf{y}_v)\}_{v\in V}$ from $p(\mathbf{G}, \mathbf{Y}|\mathbf{e} = e)$, the prediction model needs to address the testing data $\{(\mathbf{G}_v, \mathbf{y}_v)\}_{v\in V'}$ from a different distribution $p(\mathbf{G}, \mathbf{Y}|\mathbf{e} = e')$, where e and e' denote the specific environment variables. Let $\mathcal{E}$ represent the set of environments, h as a prediction model with $\hat{\mathbf{y}}_v = h(\mathbf{G}_v)$ and $l(\cdot, \cdot)$ as a basic loss function. More formally, the OOD generalization problem for graph-structured data can be expressed as:

$$\min_{h} \max_{e\in\mathcal{E}} \mathbb{E}_{G\sim p(\mathbf{G}|\mathbf{e}=e)}\Big[\frac{1}{|V|}\sum_{v\in V}\mathbb{E}_{y\sim p(\mathbf{y}_v|\mathbf{G}_v,\mathbf{e}=e)}[l(h(\mathbf{G}_v), \mathbf{y}_v)]\Big], \qquad (2)$$

Since neighboring nodes are likely to contain more correlation knowledge, it contributes to the stable prediction for the label $\mathbf{y}_v$ across different environments. Moreover, the influence of any node u in the original graph $\mathbf{G}$ on different central nodes v may vary significantly, which mainly depends on its own position in the ego graph $\mathbf{G}_v$. In this way, the above formulation provides sufficient flexibility and reliability for modeling graph data to accomplish the correlation exploration of the graph structure.

C. Causal Alignment on Rationales

Through the causal observations of interactions between data variables, we introduce the principle of discovering shortcuts for causal alignment. Leveraging shortcut cues to achieve causal alignment for the transparent prediction needs to understand the underlying mechanism of the task of interest [47]. Without loss of generality, we pay attention to the graph-structured classification task and provide a causal view of the shortcut cue discovery strategy suitable for this task. Here, we construct the causal knowledge as a structural causal model (SCM) [48] by investigating the potential causality among four random variables: the input graph G, the ground-truth label Y, the causal part C, and the non-causal part M. Fig. 2 shows the causal impact analysis, where each link represents a causal relation between two variables.

- $C \rightarrow G \leftarrow M$. The input graph G can be divided into two distinct components: the causal component C and the non-causal component M.
- $C \rightarrow Y$. This means that C is the only piece of knowledge to ascertain the ground-truth label Y by causal extrapolation.
- $C \dashleftarrow\dashrightarrow M$. This dashed arrow indicates the potential shared dependencies between C and M. Here, we consider four possible relationships: 1) C is independent from M (*i.e*, $C \perp\!\!\!\perp M$); 2) C is a direct cause of M (*i.e*, $C \rightarrow M$); 3) M is a direct cause of C (*i.e*, $M \rightarrow C$); and 4) H is a common cause of C and M (*i.e*, $C \leftarrow H \rightarrow M$).

$C \dashleftarrow\dashrightarrow M$ can give rise to superficial correlations (*e.g*, spurious shortcuts) between the non-causal component M and the ground-truth label Y. Suppose $C \rightarrow M$, C is a confounding factor between M and Y. In this way, a backdoor path can be constructed for $M \leftarrow C \rightarrow Y$, which makes M and Y falsely correlated. We formalize such spurious correlations as $Y \not\perp\!\!\!\perp M$. Among them, we perform the feature induction assumption on M to prevent the confusion between the induced subset of M and C. In addition, data collected from different environments may result in various spurious correlations, *e.g*, training data mainly consists of *House* motifs with *Wheel* bases, while testing data consists of *House* motifs with *Tree* bases. Therefore, the causal alignment on rationales can be applied to enhance the predictive stability of GNN models by suppressing spurious correlations.

IV. Proposed CASR Scheme

A. Key Idea of CASR

In this section, we describe the proposed CASR scheme for resisting adversarial attacks on graph-structured data in detail. The architecture and workflow of CASR are illustrated in Fig. 3. The ultimate goal of our scheme is to ensure that the model can provide the user with accurate prediction results even if the input graph-structured data is subject to malicious attacks by adversary. For example, an attacker may perturb the graphs by introducing spurious nodes or structures, where our proposed scheme employs causality-aligned structure rationalization to enable the GNN model to understand the effective essences in the graph-structured samples. In this case, even if the adversary manipulates the correlations between nodes to mislead the victim model, the proposed CASR that captures the real interaction factors is still able to provide reliable prediction services to users. Our scheme can be divided into the following three main stages. (a) Potential shortcut discovery. The initial graph samples collected from different environments are employed to generate multi-fold graph structures through auxiliary perceptual regulators to enhance the diversity of datasets and explore the interaction between shortcuts. (b) Causality-aligned rationale construction. The original data and augmented data participate in the construction of causal rationales by minimizing the empirical risk, which is guided through the bi-directional propagation between M and C. (c) Consistency-driven optimal rationalization. The causal query y_c and the intervention query y_m are designed to regulate the distribution intervener and the causality identifier through a query feedback mechanism, where y_c is encouraged and y_m is repressed to produce optimal rationalization in the bi-level collaborative optimization process. For a quick reference, Table I summarizes the variables and their definitions used in this article.

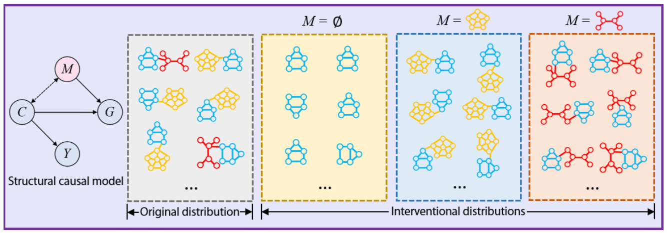


Fig. 2. A causal perspective on the interactions between data variables through structural causal models and interventional distributions.

B. Potential Shortcut Discovery

Based on the preliminary knowledge in Section III-B, the direct minimization of risk expectation across environments is adverse to the discovery of potential shortcut cues, which will inevitably encourage the model to depend on superficial correlations. In addition, this dependence would be reinforced when the uncertainty of the influence from the environment is relatively low. The empirical studies reveal that if the model obtains the same performance in different environments, it will tend to exploit the invariant features, which prompts us to design a suitable objective function to mitigate this problem. We assume a general case that we adopt $\{(\mathbf{G}_v, \mathbf{y}_v)\}_{v \in V}$ for training and utilize the GNN model as a predictor: $\hat{\mathbf{y}}_v = h_\theta(\mathbf{G}_v)$, where θ denotes the model parameters. Based on the above analysis, a comprehensive learning objective can be described as follows:

$$\min \mathbb{V}_{\mathbf{e}}[L(\mathbf{G}^e, \mathbf{Y}^e; \theta)] + \alpha \mathbb{E}_{\mathbf{e}}[L(\mathbf{G}^e, \mathbf{Y}^e; \theta)], \tag{3}$$

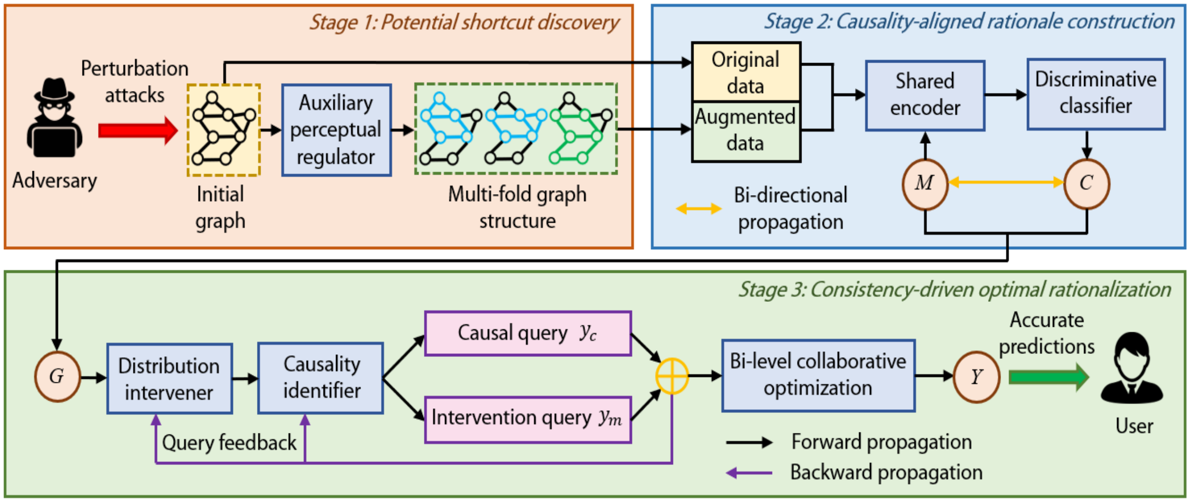

Stage 1: Potential shortcut discovery
Perturbation attacks
Adversary
Initial graph
Auxiliary perceptual regulator
Multi-fold graph structure
Stage 2: Causality-aligned rationale construction
Original data
Augmented data
Shared encoder
Discriminative classifier
M
C
Bi-directional propagation
Stage 3: Consistency-driven optimal rationalization
G
Distribution intervener
Causality identifier
Causal query y_c
Intervention query y_m
Query feedback
Bi-level collaborative optimization
Y
Accurate predictions
User
Forward propagation
Backward propagation

Fig. 3. The pipeline of our proposed CASR consists of three stages, which can defend against graph-structured data manipulation attacks while achieving OOD generalization by identifying spurious shortcuts or real essences.

TABLE I

Descriptions of Variables Used in This Paper

Notation	Definition
$\mathbf{e}$	Environment vector
$\mathcal{E}$	Set of environment vectors
$\mathbf{x}$	Sample vector
$\mathbf{y}$	Label vector
$f(\cdot)$	Prediction model
$l(\cdot,\cdot)$	Basic loss function
G	Input graph variable
$\mathbf{A}$	Adjacency matrix
$\mathbf{X}$	Node feature matrix
V	Set of node variables
$\mathbf{x}_v$	The v-th node feature vector
$\mathbf{y}_v$	The v-th node label vector
$\mathbf{G}$	Graph matrix
$\mathbf{Y}$	Node label matrix
C	Causal component
M	Non-causal component
θ	Model parameter
α	Trade-off parameter
w_z	Weight of the z-th regulator
$\mathrm{Var}(\cdot)$	Variance calculation function
$p_s(\cdot)$	Source data distribution
$p_t(\cdot)$	Target data distribution
ψ_X	Exogenous noise of X
f_1	Shared encoder
f_2	Discriminative classifier
$\mathcal{R}(\cdot,\cdot)$	Risk function
β	Adjustable hyperparameter
γ, η, μ	Encoder parameters
$Q(\cdot)$	Query function
n_τ	Number of nodes in the τ-th iteration
m_τ	The m-intervention operation during the τ-th iteration

where $L(\mathbf{G}^e,\mathbf{Y}^e;\theta) = \frac{1}{|V_e|}\sum_{v\in V_e} l(h_\theta(\mathbf{G}_v^e),\mathbf{y}_v^e)$ and α is a trade-off parameter.

If we obtain the training graphs from an adequate number of environments and the correspondence mapping relationships of each graph in a particular $\mathbf{e}$, for instance, $\{\mathbf{G}^e,\mathbf{Y}^e\}_{e\in\mathcal{E}_{tr}}$ derives $\left\{\{\mathbf{G}_v^e,\mathbf{y}_v^e\}_{v\in V_e}, e\in\mathcal{E}_{tr}\right\}$, we can handle Eq. (3) through the empirical estimation with risk extrapolation across diverse environments [49]. As a result, Eq. (3) requires the collection of data from different environments to enable the model to make the desired extrapolation. To overcome such a dilemma, we develop Z auxiliary perceptual regulators $\{r_{w_z}(\mathbf{G}), z\in[1,Z]\}$ (w_z is the weight of the z-th regulator), which aim to produce Z-fold graph-structured data $\left\{\{\mathbf{G}_v^z\}_{v\in V}, z\in[1,Z]\right\}$ based on a specific $\mathbf{G}$ and adjust the training data in different environments. The regulators are trained by maximizing variance loss to explore the effect of different environments and to promote the stable learning of GNNs. Therefore, this can be described by the mathematical expression as follows:

$$\min_{\theta} \mathrm{Var}[L(r_{w_z^*}(\mathbf{G}),\mathbf{Y};\theta)] + \frac{\alpha}{Z}\sum_{z=1}^{Z}[L(r_{w_z^*}(\mathbf{G}),\mathbf{Y};\theta)],$$
$$\text{s.t. } [w_1^*,\ldots,w_Z^*] = \arg\max_{w_1,\ldots,w_Z} \mathrm{Var}[L(r_{w_z}(\mathbf{G}),\mathbf{Y};\theta)], \quad (4)$$

where $\mathrm{Var}(\cdot)$ denotes the variance calculation function, and $L(r_{w_z}(\mathbf{G}),\mathbf{Y};\theta) = L(\mathbf{G}_v^z,\mathbf{Y};\theta) = \frac{1}{|V|}\sum_{v\in V} l(h_\theta(\mathbf{G}_v^z),\mathbf{y}_v)$. In addition, we edit the graph-structured data by adding or deleting edges to specify $r_{w_z}(\mathbf{G})$.

In addition, we assume that $I(\mathbf{x},\mathbf{y})$ denotes the mutual information between $\mathbf{x}$ and $\mathbf{y}$, and $I(\mathbf{x},\mathbf{y}|\mathbf{e})$ represents the conditional mutual information for a specific $\mathbf{e}$. To simplify the symbolic description, we record $p_e(\cdot) = p(\cdot|\mathbf{e}=e)$ and $I_e(\cdot) = I(\cdot|\mathbf{e}=e)$. Another practical issue is that in the calculation of the KL divergence, we need to collect samples from the joint probability distribution $p_e(\mathbf{G},\mathbf{Y})$, which leads to the difficulty of dealing with the data transformations of the interconnected nodes. To this end, for any probability function h_1, h_2 related to the ego-graph $\mathbf{G}_v$ and the node label $\mathbf{y}_v$, we define the following formula to compute the KL discrepancy:

$$D_{KL}(h_1(\mathbf{G}_v,\mathbf{y}_v)||h_2(\mathbf{G}_v,\mathbf{y}_v))$$
$$:= \mathbb{E}_{G\sim p(\mathbf{G})}\left[\frac{1}{|V|}\sum_{v\in V}\mathbb{E}_{y\sim p(\mathbf{y}_v|\mathbf{G}_v)}\left[\log\frac{h_1(\mathbf{G}_v,\mathbf{y}_v)}{h_2(\mathbf{G}_v,\mathbf{y}_v)}\right]\right]. \quad (5)$$

Following the similar idea in [50], the training error and the OOD generalization error can be respectively evaluated by $D_{KL}(p_s(\mathbf{y}_v|\mathbf{G}_v)||q(\mathbf{y}_v|\mathbf{G}_v))$ and $D_{KL}(p_t(\mathbf{y}_v|\mathbf{G}_v)||q(\mathbf{y}_v|\mathbf{G}_v))$

based on the definition in Eq. (5), where $p_s(\mathbf{y}_v|\mathbf{G}_v)$ denotes the source data distribution, $p_t(\mathbf{y}_v|\mathbf{G}_v)$ indicates the target data distribution, and $q(\mathbf{y}_v|\mathbf{G}_v)$ represents the ideal data distribution. In this way, the objective learning can control the generalization error of OOD samples and facilitate the discovery of potential shortcut cues.

C. Causality-Aligned Rationale Construction

It is known from the causality theory [51] that for each variable X in the SCM, when the causal mechanism $X = F_X(PA(X), \psi_X)$ holds, there exists a single directed link from the parent variable $PA(X)$ to X, where ψ_X denotes the exogenous noise of X. For simplicity, we drop the exogenous noise and then record it as $X = F_X(PA(X))$. Thus, there exists a function $F_Y : C \to Y$ in the SCM, where the underlying rationale needs to satisfy the following condition:

$$Y = F_Y(C), \;\; Y \perp\!\!\!\perp M \mid C, \tag{6}$$

where $Y \perp\!\!\!\perp M \mid C$ means that C mitigates the impact of Y on M such that the causal relation $C \to Y$ is invariant across diverse M.

Since only the graph G and the label Y can be observed in the training process, while neither the causal component C nor the structural mapping model F_Y is available, the inherent interpretability needs to be investigated. In general, the inherently interpretable GNN model h can be disentangled into a shared encoder f_1 for consensus representation and a discriminative classifier f_2 for performance prediction, namely $h = f_1 \circ f_2$, where $f_1 : G \to \tilde{C}$ guides the discovery of causal rationale $\tilde{C}$ from the observed G, and $f_2 : \tilde{C} \to \hat{Y}$ ensures the prediction result $\hat{Y}$ to approximate Y. Unlike C and Y as the variables in the causal inference, $\tilde{C}$ and $\hat{Y}$ denote potential variables in the model learning process to approach C and Y. To optimize the modules, the common interpretable GNNs [31], [52] generally employ the learning strategy that minimizes the empirical risk as follows:

$$\min_{f_1, f_2} \mathcal{R}(f_1(G) \circ f_2(\tilde{C}), Y), \tag{7}$$

where $\mathcal{R}(\cdot,\cdot)$ is the risk function that can be realized by using the cross-entropy loss.

Nonetheless, such a learning strategy heavily depends on the statistical correlations between the input features and the labels, which may lead to the prediction model that captures non-causal rationales. In fact, the cause of this phenomenon is due to the ignorance of $Y \perp\!\!\!\perp M \mid C$ in Eq. (6), which is essential to ensure that the causality $C \to Y$ is invariant across different M. By introducing this constraint, we explore the invariant rationales through the following expression:

$$\min_{f_1, f_2} \mathcal{R}(f_1(G) \circ f_2(\tilde{C}), Y), \;\; \text{s.t. } Y \perp\!\!\!\perp \tilde{M} \mid \tilde{C}, \tag{8}$$

where $\tilde{M} = G \setminus \tilde{C}$ denotes the relative complement of $\tilde{C}$ in G. In this way, the construction of causal rationales can be motivated to capture stable representations and eliminate unstable representations.

Different from dividing the training data, we perform the intervention to infer the invariant patterns across diverse distributions. Specifically, we obtain m-interventional distributions by performing intervention $do(M = m)$ on M. This deletes each link from the parents $PA(M)$ to the variable M through setting M to a specific value m. By combining different values $M = \{m\}$, we can generate multiple m-interventional distributions. Thus, the learning strategy for causality-aligned rationalization can be described by the following formula:

$$\min \mathbb{E}[\mathcal{R}(h(G), Y|do(M{=}m))] + \beta \mathrm{Var}[\mathcal{R}(h(G), Y|do(M{=}m))], \tag{9}$$

where $\mathcal{R}(h(G), Y|do(M = m))$ evaluates the risk under the condition of m-interventional distributions, and β is a hyper-parameter that adjusts the strength of rationale learning. By minimizing the empirical risk while making the distribution-related risk insensitive, the causality-aligned rationalization is able to seek the invariant $\tilde{C}$ as the approximation to C, which can promote both f_1 and f_2 to reach the desired state to improve the generalization performance of the model h.

D. Consistency-Driven Optimal Rationalization

In order to make the rationales exhibit excellent performance, we propose a consistency-driven optimal rationalization mechanism through minimizing the approximation error contributed by the cumulative risk across m-interventional distributions. After establishing the prediction $\hat{\mathbf{y}}_v$ for the instance $\mathbf{G}_v$ by the intervention $do(M = m)$, we can obtain the m-intervention risk similar to Eq. (9) as follows:

$$\mathcal{R}(h(G), Y|do(M{=}m)) = \mathbb{E}_{\{(\mathbf{G}_v, \mathbf{y}_v)\}_{v \in V}, \tilde{C} = f_1(G)}[l(\hat{\mathbf{y}}_v, \mathbf{y}_v)], \tag{10}$$

where $l(\hat{\mathbf{y}}_v, \mathbf{y}_v)$ calculates the loss on an individual instance. Let $\hat{\mathbf{y}}_v^m$ denote the predictive result under the m-intervention operation. Then, we design an interventional loss evaluation function for the distribution intervener as:

$$\mathcal{R}_m = \mathbb{E}[l(\hat{\mathbf{y}}_v^m, \mathbf{y}_v)], \;\text{s.t. } Y \perp\!\!\!\perp \tilde{M} \mid \tilde{C}, \tilde{C} = f_1(G), \hat{Y} = f_2(\tilde{C}). \tag{11}$$

Specifically, $\mathcal{R}_m$ is required to be completely back-propagated to the classifier f_2, where its back-propagation is separated from the other components to prevent interference with the predefined learning objective. As a result, it is possible that this loss boosts the specific network branches to automatically learn the non-causal patterns of spurious biases. In summary, we can optimize these components together with the intervention risk of rationale construction across different distributions to improve the performance of the whole network, namely,

$$\begin{aligned} \min_{\gamma, \eta, \mu} \; & \mathbb{E}[\mathcal{R}(h(G), Y|do(M{=}m))] + \beta \mathrm{Var}[\mathcal{R}(h(G), Y|do(M{=}m))] \\ & + \mathbb{E}[l(\hat{\mathbf{y}}_v^m, \mathbf{y}_v)], \;\; \text{s.t. } Y \perp\!\!\!\perp \tilde{M} \mid \tilde{C}, \;\; \tilde{C} = f_1(G), \;\; \hat{Y} = f_2(\tilde{C}), \end{aligned} \tag{12}$$

where γ, η and μ are the parameters of the graph encoder, the evaluation classifier and the distribution intervener, respectively. Moreover, in the optimization inference stage, we obtain $\tilde{C}$ and $\hat{Y}$ as the causal invariant rationale and the causal effect prediction of the test graph G, which are able to avoid the impact of the non-causal component $\tilde{M}$.

To achieve consistency-driven rationalization, we introduce the query function $Q(\cdot)$ (*i.e*, causal query or intervention query), which designates the query result $K = Q(\tilde{C})$ as the feedback variable for the distribution intervener and the causality identifier. Due to the high cost of collecting experimental data in practice, we actively seek the available instances for consistency-driven optimization by performing the m-intervention operation, which can efficiently facilitate the query to capture invariant rationales across distributions. For the τ-th iteration, the experiments results are derived from a batch of instances that are independent identically distributed with the observed true interventional distribution:

$$\mathbf{G}^{\tau} = \{\mathbf{G}_v^{\tau}\}_{v=1}^{n_\tau}, \ \mathbf{G}_v^{\tau} \sim p(G, \tilde{C} | do(M = m_\tau)), \tag{13}$$

where n_τ denotes the number of nodes in the τ-th iteration, and m_τ means the m-intervention operation during the τ-th iteration.

Most importantly, we adopt the m-intervention operation to be the most informative with respect to the query K. In the optimization setting, this is naturally expected as the maximization of the corresponding information gain for the subsequent interventions, *i.e*, the mutual information between K and $\mathbf{G}^{\tau}$:

$$\max_{do(M=m_\tau)} I(K, \mathbf{G}^{\tau} | \mathbf{G}_v^{1:\tau-1}), \tag{14}$$

where $\mathbf{G}^{\tau}$ obeys the estimated interventional distribution of the causal model under the intervention of m_τ, and $\mathbf{G}_v^{1:\tau-1}$ denotes the dataset collected when the iteration reaches $\tau - 1$. Then, the likelihood of $\mathbf{G}_v^{1:\tau}$ can be computed by:

$$\begin{aligned} p(\mathbf{G}_v^{1:\tau} | \tilde{C}, \hat{Y}) &= \prod_{i=1}^{\tau} p(\mathbf{G}^{i} | \tilde{C}, \hat{Y}, do(M = m_i)) \\ &= \prod_{i=1}^{\tau} \prod_{v=1}^{n_\tau} p(\mathbf{G}_v^{i} | \tilde{C}, \hat{Y}, do(M = m_i)). \end{aligned} \tag{15}$$

To exploit the accessible data $\mathbf{G}_v^{1:\tau}$, we record $RT(G)$ as the root node and $NRT(G)$ as the non-root node. In this way, we can update the belief and quantify the uncertainty for $\tilde{C}$ in SCM by calculating the posterior $p(\tilde{C} | \mathbf{G}_v^{1:\tau})$, which can be expressed as:

$$\begin{aligned} p(\tilde{C} | \mathbf{G}_v^{1:\tau}) = p(\hat{Y} | \mathbf{G}^{\tau}) &\prod_{v_1 \in RT(G)} p(\tilde{C} | \mathbf{G}_{v_1}^{1:\tau}, do(M = m_\tau)) \\ + p(\hat{Y} | \mathbf{G}^{\tau}) &\prod_{v_2 \in NRT(G)} p(\tilde{C} | \mathbf{G}_{v_2}^{1:\tau}, do(M = m_\tau)). \end{aligned} \tag{16}$$

For the root node $v_1 \in RT(G)$, given the other variables and parameters in the graph-structured data, which enables us to directly calculate the root node posterior $p(\tilde{C} | \mathbf{G}_{v_1}^{1:\tau}, do(M = m_\tau))$ in Eq. (16). While for the non-root node $v_2 \in NRT(G)$, the posterior can be obtained by the following two formulas:

$$\begin{aligned} &p(\mathbf{G}^{\tau} | \hat{Y}) \\ &\quad = \frac{p(\hat{Y} | \mathbf{G}^{\tau}) p(\mathbf{G}^{\tau})}{p(\hat{Y})}, \end{aligned} \tag{17}$$

$$\begin{aligned} &p(\tilde{C} | \mathbf{G}_{v_2}^{1:\tau}, do(M = m_\tau)) \\ &\quad = \frac{p(\mathbf{G}_{v_2}^{1:\tau} | \tilde{C}, do(M = m_\tau)) p(\tilde{C} | \hat{Y})}{p(\mathbf{G}^{\tau} | \hat{Y})}. \end{aligned} \tag{18}$$

By performing the m-intervention operation, our goal is to maximize the information gain about the query result in Eq. (14) while minimizing Eq. (12) to obtain the corresponding parameters. Thus, it is equivalent to minimizing the following function $U(m)$:

$$\begin{aligned} U(m) = \mathbb{E}_{\tilde{C} | \mathbf{G}_v^{1:\tau}} &\Big[\mathbb{E}_{\mathbf{G}^{\tau} | \tilde{C}} \Big[\log \mathbb{E}_{K | \mathbf{G}_v^{1:\tau}} \big[p(\mathbf{G}^{\tau} | \tilde{C}) \big] \Big] \Big] \\ &+ \mathbb{E}_{\tilde{C} | \mathbf{G}_v^{1:\tau}} \Big[\mathbb{E}_{\mathbf{G}^{\tau}, \hat{Y} | \tilde{C}} \Big[\log \mathbb{E}_{K | \mathbf{G}_v^{1:\tau}} \big[p(\mathbf{G}^{\tau} | \tilde{C}) p(\hat{Y} | \tilde{C}) \big] \Big] \Big], \end{aligned} \tag{19}$$

where the entropy $\mathbb{E}_{K | \mathbf{G}_v^{1:\tau}}[\cdot]$ can be efficiently calculated by selecting the model parameters for the given graph-structured data. The search for the optimal intervention operation $m^* = (M^*, G_M^*)$ requires jointly optimizing the function $U(m)$ according to the query requests, which involves (1) the set of intervention operation M and (2) the corresponding m-intervention graph-structured data G_M. It naturally motivates us to adopt a bi-level collaborative optimization solution [53]:

$$M^* \in \arg\min_{M} U(M, G_M^*), \forall M, \text{ s.t. } G_M^* \in \arg\min_{G_M} U(M, G_M) \tag{20}$$

In the above updating process, we first evaluate the optimal intervention data for all possible candidate intervention operations M and then discover the intervention operation that minimizes the objective function. The intervention operation M may impact multiple variables and parameters, which will result in a complex selection problem. Therefore, for simplicity, we only consider the single-node intervention, *i.e*, $|M| = 1$. To obtain G_M^*, we utilize bi-level collaborative optimization to efficiently compute the most appropriate intervention value G_M^*.

Our proposed scheme can dynamically adjust the composition of real essences and spurious shortcuts by receiving the feedback from the acquired specific knowledge, which calculate and update the corresponding weights to give more importance to critical patterns that produce more stable predictions, while reducing the influence of features that might be result in biased predictions. The process of adjustment and refinement based on feedback is iterative. In other words, the model continues to make inferences, receive feedback, and update parameters to improve the performance over time. By adaptively adjusting the weights as the feedback control, the CASR scheme allows the GNN model to differentiate between valuable causal patterns and misleading non-causal correlations. This enables the model to achieve more reliable predictions, even when faced with adversarial biased perturbations or other challenges that may introduce spurious shortcuts.

Algorithm 1 Causality-Aligned Structure Rationalization

```
Input: training graph data $\{(\mathbf{G}_v, \mathbf{y}_v)\}_{v \in V}$, pre-training
       GNN model $h$, shared encoder $f_1$,
       discriminative classifier $f_2$, trade-off
       parameters $\alpha$, $\beta$, total iteration number $\Phi$;
Output: optimal parameters of GNN $\theta^*$, optimal GNN
        model $h^*$;
Initialize the parameters of GNN model, graph
  encoder, evaluation classifier and distribution
  intervener, which are denoted as $\theta$, $\gamma$, $\eta$ and $\mu$,
  respectively;
while not converge do
    Train graph-structured data across different
      environments according to Eq. (4);
    Construct causality-aligned rationales in the SCM
      by using Eq. (8);
    Infer invariant patterns by doing the intervention
      based on the learning strategy of Eq. (9);
    for $\tau = 1, 2, \ldots, \Phi$ do
        Perform $m$-intervention operation $do(M = m_\tau)$
          during the $\tau$-th iteration;
        Maximize the mutual information between $K$
          and $\mathbf{G}^\tau$ through Eq. (14);
        Quantify the uncertainty for $\tilde{C}$ by computing
          the posterior $p(\tilde{C}|\mathbf{G}_v^{1:\tau})$ in Eq. (16);
        Obtain the objective function $U(m)$ by
          equivalently transforming Eq. (12) and Eq.
          (14);
    end
    if $\tau == \Phi$ then
        Calculate the optimal intervention operation $m^*$
          by jointly optimizing $U(m)$;
        Update the optimal graph-structured data $G_M^*$
          based on Eq. (20);
    end
end
```

Fundamentally, the query feedback mechanism empowers the GNN to learn from the experiences to obtain accurate and robust decision-making results. Based on these analysis, the proposed causality-aligned structure rationalization algorithm can be summarized in Algorithm 1.

V. Evaluation

A. Datasets

We adopt one synthetic dataset (*e.g*, Spurious-Motif[1] [54]) and several real datasets (*e.g*, MNIST-75sp[2] [14] and TUDataset[3] [55]) for graph classification tasks. The diverse GNNs are applied in different datasets to demonstrate the effectiveness of the proposed scheme. Specifically, the experiments are performed to evaluate the GNN models in the following two aspects, *i.e*, the capability of resisting data poisoning attacks and the ability of OOD generalization. Here we briefly describe the datasets, while the statistics of all datasets are summarized in Table II. Moreover, a more detailed description of the dataset can be seen as follows: leftmargin=*

TABLE II

Statistics of Graph Benchmark Datasets Used in Our Evaluation Experiments

Datasets	#Graphs	#Nodes	#Edges	#Classes
Spurious-Motif	18,000	25.6	35.6	3
MNIST-75sp	70,000	66.8	600.2	10
AIDS	2,000	15.7	16.2	2
NCI1	4,110	29.9	32.3	2
PC3	2,751	26.4	28.5	2
IMDB-B	1,000	19.8	193.1	2
IMDB-M	1,500	13.0	65.9	3

- **Spurious-Motif.** It is a synthetic dataset containing 18,000 graphs, where each graph consists of one base component (*Tree*, *Ladder*, *Wheel* represented by $S = 0$, 1, 2) and one motif component (*Cycle*, *House*, *Crane* represented by $T = 0$, 1, 2). The ground-truth label is solely determined by the motif component. In the distribution shift scenario, an adjustable bias b is introduced to control the distribution between the base component and the motif component in the training set:

$$P(S) = \begin{cases} b \ , & \text{if} \quad S = T, \\ \frac{1-b}{2}, & \text{otherwise}. \end{cases} \tag{21}$$

  In the testing set, the base and motif components are randomly connected together with equal probability. Thus, we can manipulate b to produce Spurious-Motif datasets with different distribution gaps.
- **MNIST-75sp.** The MNIST images are converted into 70,000 superpixel graphs with up to 75 nodes in each graph. The nodes in the graphs are denoted by superpixels, while the edges are represented by the spatial distance between the nodes. The label of each graph is derived from one of 10 categories. Notably, the random noise is added to the node features of the testing set.
- **TUDataset.** We select three molecular datasets (*i.e*, AIDS, NCI1, PC3) and two social networks datasets (*i.e*, IMDB-B, IMDB-M) from TUDataset, which are widely used as graph classification benchmarks. For molecular datasets, the graphs describe the structure of the chemical compounds, where the nodes represent the atoms and the edges indicate the bonds between the atoms, respectively. For social networks datasets, IMDB-B and IMDB-M are two collaboration datasets in movies, where the nodes denote the actors and the edges indicate that the connected actors appear in the same movie.

B. Attack Models

We adopt five representative graph attack methods to evaluate the effectiveness of the proposed scheme. leftmargin=*

- Random perturbation attack (RPA). This is a simple and feasible attack strategy by randomly adding or removing edges to produce perturbed graphs. In our experiments, the default perturbation ratio is set to 3.0%. The RPA as

[1] https://github.com/RexYing/gnn-model-explainer

[2] https://github.com/bknyaz/graph_attention_pool

[3] https://github.com/chrsmrrs/tudataset

a baseline attack strategy can be employed to evaluate the vulnerability of GNN models under simple random perturbations.

- Adversarial transfer attack (ATA) [44]. The purpose of ATA is to degrade the recognition performance of the model substantially by deploying the well-trained model to predict similar but unseen samples. The ATA simulates that an attacker may try to leverage the limited generalization of GNNs in real-world scenarios, which facilitates the evaluation of the robustness on similar but unseen samples.
- ReWiring attack (ReWatt) [40]. The deep reinforcement learning is utilized to perform rewiring attacks that affect the graphs in a less noticeable way. The ReWatt deceives the model while maintaining graph-structured integrity through subtle perturbations, which can be exploited to evaluate the sensitivity of GNN models to imperceptible perturbations.
- Bandit optimization attack (BOA) [17]. The attacker captures the unknown gradient through the prior knowledge query and then implements the discrete structure perturbation by using the bandit optimization. The BOA maximizes the effectiveness of the attack through limited information, which can be used to test the stability and reliability of GNN models defense against adversarial biased perturbations.
- SPectral AttaCk (SPAC) [43]. This attack leverages a transformation-based approach, which maximizes the spectral distance between the original and perturbed graphs in the frequency domain to complete the attack. The SPAC can be applied to explore the vulnerability of GNN models on graph-structured perturbations through spectral characteristics to assess the robustness against such attacks.

These selected attack models cover a range of attack strategies, from simple random to complicated hybrid perturbations. By introducing these diverse attack models, we can deeply investigate the reasons for the vulnerability of GNN models and evaluate their robustness and reliability under different types of adversarial biased perturbations.

C. Defense Methods

We thoroughly compare the proposed scheme with the following two types of defense methods.

- Specialized reasoning defenses: graph attention network (GAT) [56], uncertainty-aware attention graph (UAG) [23], and structural entropy pooling (SEP) [57]. We adopt the masks generated on the graph-structured data as the rationales. We also employ graph substructure networks (GSN) [22], a topology-aware knowledge sharing approach that provides rich and diverse data for GNNs with salient structural representations.
- Stable learning defenses: variance risk extrapolation (V-REx) [49], multi-domain calibration (MDC) [58], and localized adversarial domain generalization (LADG) [21]. Such algorithms improve the robustness and stability for GNN models, which can promote the generalization ability of the models to achieve better prediction results in unseen domains or out-of-distribution datasets.

TABLE III

Results of Classification Accuracy (%) on Benchmark Datasets Under the Condition of RPA. The Best Results Are Marked in Bold

Compared models	Spurious-Motif	MNIST-75sp	AIDS
GAT	45.93±6.58	19.74±4.92	94.37±4.16
UAG	50.75±3.29	23.88±7.41	95.49±2.85
SEP	52.14±7.30	28.70±2.97	96.02±3.48
V-REx	49.87±2.37	23.17±5.41	95.98±1.43
MDC	50.26±6.49	24.35±3.74	96.57±2.35
LADG	52.50±2.75	27.18±4.18	97.38±1.19
Ours	**55.83±1.86**	**29.37±2.17**	**98.15±0.93**

VI. Experimental Results

In this section, our experimental evaluation is designed to answer the following research questions (RQs):

- **RQ1:** Can the proposed CASR effectively resist various attacks in graph-structured tasks?
- **RQ2:** Does our CASR scheme outperform other defense methods under the condition of hybrid attacks?
- **RQ3:** How does CASR contribute to the improvement of generalization capability?
- **RQ4:** What is the role of curbing spurious shortcuts and boosting real essences for enhancing the defensive performance?
- **RQ5:** Whether the stability and robustness of GNN models are guaranteed by rationale extrapolation?

A. Evaluation of Defensive Effectiveness (RQ1)

We adopt the five representative threat attacks mentioned in Section V-B to analyze the defensive effectiveness of the proposed scheme as follows.

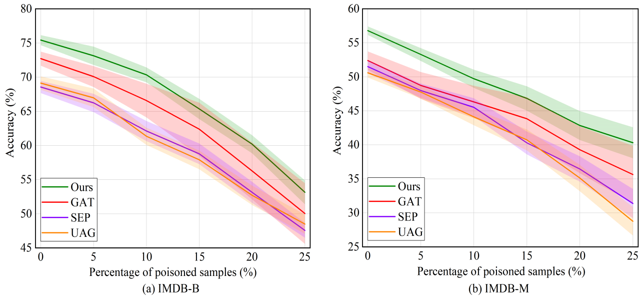
1) Defense against RPA. We randomly add or remove edges with a perturbation ratio of 3.0% to three groups of datasets, including Spurious-Motif, MNIST-75sp, and AIDS. The results of the random perturbation attack for all three datasets are shown in Table III. We follow the default parameter settings of other defense models in practical implementation. We observe that UAG, V-REx and MDC have similar performance against the random structural perturbations. However, among the seven defense methods, our scheme achieves the highest accuracy on the perturbed graphs in all experiments. In addition, the proposed scheme exhibits the least fluctuation in performance variance. In contrast, GAT seems to be the most sensitive to the random perturbation attack.

2) Defense against ATA. We adopt the Spurious-Motif dataset for the adversarial transfer attack. Based on the dataset descriptions, we set different b to obtain multiple biased datasets. In this attack, the attacker can employ the deceptive graph datasets and thus all the defense methods suffer from significant performance degradation when compared to the random perturbation attack. As given in Table IV, the performance of the defense methods gradually decreases as b increases, which indicates that the intensity of ATA is positively correlated with the defense difficulty. Surprisingly, our defense scheme is considerably more robust, which achieves a performance improvement of up to 9.79%. This clearly demonstrates that the proposed scheme has the ability to extrapolate the true structural knowledge even in the case of ATA.

TABLE IV

Results of Classification Accuracy (%) on the Spurious-Motif Dataset Under the Condition of ATA. The Best Results Are Marked in Bold

Compared models	Spurious-Motif			
	b = 0.6	b = 0.7	b = 0.8	b = 0.9
GAT	40.36±6.37	39.04±6.71	38.42±5.83	34.67±5.23
UAG	45.78±2.45	43.62±5.87	41.57±4.62	38.15±3.95
SEP	46.81±5.61	43.14±4.75	40.28±4.48	37.06±1.74
V-REx	43.12±3.78	41.83±2.36	38.42±3.29	35.64±3.79
MDC	44.64±2.85	42.14±1.94	37.16±3.73	36.18±2.02
LADG	45.98±4.63	41.75±3.92	38.42±1.82	38.01±2.36
Ours	**50.15±2.79**	**48.40±1.84**	**44.92±3.58**	**41.78±0.96**



(a) IMDB-B

(b) IMDB-M

Fig. 4. Experimental evaluation results for defending against the ReWatt on the IMDB-B and IMDB-M datasets.

3) Defense against ReWatt. We select the graph-structured data successfully attacked by ReWatt on the IMDB-B and IMDB-M datasets, where the strength of the attack can be controlled by adjusting the proportion of poisoned samples. The ReWatt is one of the more challenging obfuscation attacks since it modifies the graph structures in an imperceptible way. However, as shown in Fig. 4, our proposed scheme consistently exceeds all the compared defense methods by a large margin. Under this attack, it is interesting to observe that both UAG and SEP perform poorly compared to the GAT model. Nevertheless, the huge variance renders GAT unreliable and infeasible in practice, especially when the attack is severe.

4) Defense against BOA. The performance of four defense methods against BOA is tested on the IMDB-B dataset. We randomly select 100 samples from the IMDB-B dataset as auxiliary data, and then generate the corresponding poisoned samples for each auxiliary sample via the BOA. We conduct two sets of experiments, where one group used clean auxiliary samples (denoted as G_c) for training and poisoned auxiliary samples (denoted as G_p) for testing, while the other group used poisoned auxiliary samples for training and clean auxiliary samples for testing. In Fig. 5, we give the prediction confidence of various defense models on the auxiliary dataset and check if the auxiliary samples are successfully defended, *i.e*, whether the auxiliary samples are correctly classified. We find that the CASR scheme provides a dramatic improvement in defending against BOA by suppressing spurious shortcuts and promoting real essences to capture the desired rationalization. In addition, our proposed scheme enhances the confidence and accuracy of prediction, while the other compared defense approaches exhibit vulnerability to BOA.

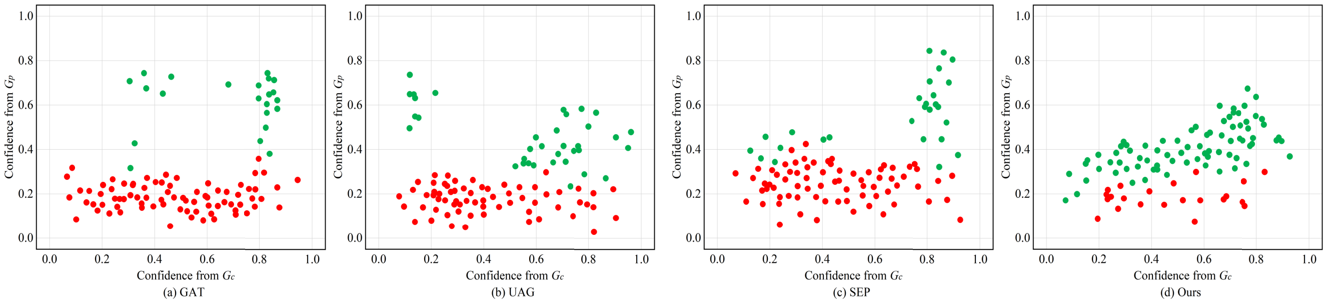
5) Defense against SPAC. Under the condition of poisoning attack and evasion attack, we evaluate the performance of several defense methods against SPAC on four datasets (*i.e*, Spurious-Motif, MNIST-75sp, IMDB-B, and IMDB-M). For the poisoning attack, we indirectly affect the classifier by perturbing the graph-structured training data. The perturbation samples are produced by using SPAC. We first train the defense models with poisoned samples, and then report the corresponding classification accuracy on clean graph-structured data. Fig. 6 provides an extensive comparison of different perturbation rates. From the results, our scheme shows an average performance improvement of 5.24% (and up to 12.11%) compared to other defense methods, which indicates that the proposed CASR is effective against SPAC. For the evasion attack, we first train a classifier on clean graph-structured data, and then the classifier is deployed for the prediction tasks. The attacker generates edge perturbations based on the prediction rule of the classifier. As can be seen in Fig. 7, the performance has been significantly improved by our defense scheme. The reason is that our CASR scheme learns the implicit knowledge of the graph-structured representations by identifying spurious shortcuts or real essences.

B. Evaluation of Comprehensive Superiority (RQ2)

To validate the comprehensive superiority of our defense scheme, we evaluate and compare the proposed model with previous defense methods under the condition of hybrid attack. Specifically, we randomly collect 20% of the graph samples as auxiliary data and adopt three attack models (*i.e*, ReWatt, BOA, and SPAC) to generate the corresponding poisoning auxiliary samples. In the practical implementation, we follow the default parameter settings of the other attack models. Then, the different types of poisoning samples are mixed to launch poisoning attacks and evasion attacks. In these experiments, the proportion of poisoned samples with different types is set to 1:1 for all scenarios of hybrid attacks. Fig. 8 shows the performance of the five defense methods against the three hybrid attacks (*i.e*, R&B, R&S, and B&S) on two datasets (*i.e*, AIDS and NCI1). Compared to the best performance of the other compared defense methods, our scheme still achieves 1.48% and 2.23% average improvement on AIDS and NCI1. Among the five defense approaches, the proposed CASR offers the superior performance in all experiments, which confirms the comprehensive superiority of our scheme against various adversarial attacks.

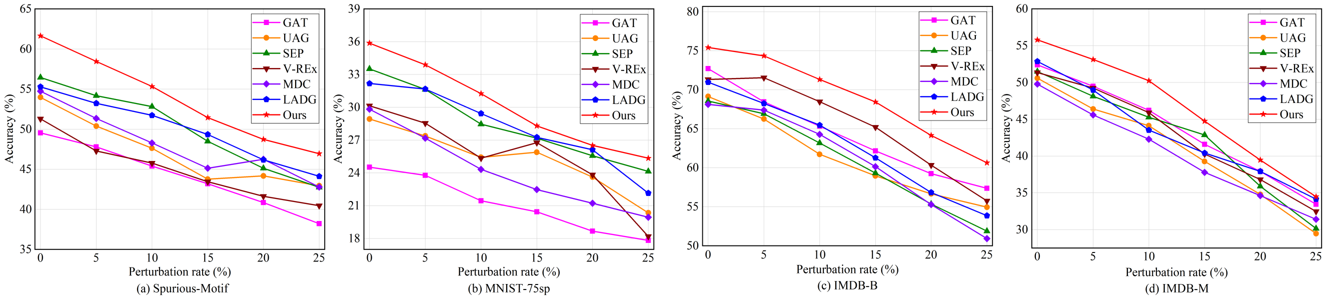
C. Evaluation of Generalization Capability (RQ3)

In order to evaluate the generalization capability, we conduct a series of experiments to investigate the OOD robustness and the dataset transferability. For the OOD robustness experiments, we adopt the MNIST-75sp dataset to obtain four testing datasets, *i.e*, clean testing samples, Gaussian noise testing samples, colored testing samples, and Gaussian noise and colored testing samples. Table V reports the experimental results of OOD generalization on the MNIST-75sp dataset. Compared with other defense methods, the proposed CASR consistently achieves the best performance on OOD samples, which demonstrates the strong OOD generalization capability of our scheme. The accuracy of the robust GAT is impressive on the clean testing graphs, while the performance decreases significantly on the OOD testing graphs. This is attributed



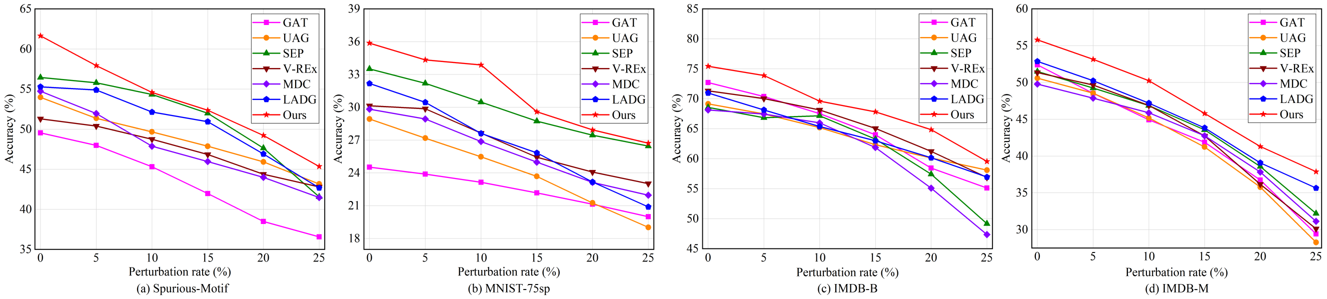
(a) GAT (b) UAG (c) SEP (d) Ours

Fig. 5. Results of resistance to BOA on the IMDB-B dataset. On the X-axis, we show the prediction confidence for the poisoned auxiliary samples through the clean auxiliary data participating in the training. On the Y-axis, we record the prediction confidence for the clean auxiliary samples through the poisoned auxiliary data participating in the training. Note that the green circles and red circles indicate correctly and incorrectly predicted samples, respectively.



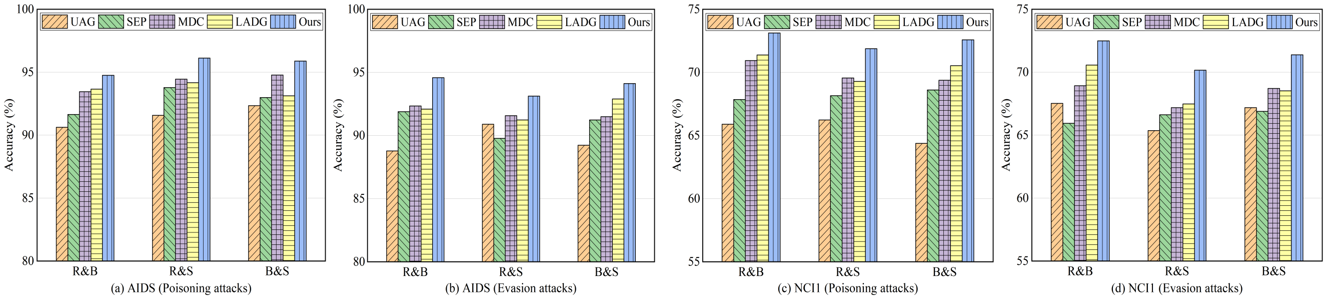
(a) Spurious-Motif (b) MNIST-75sp (c) IMDB-B (d) IMDB-M

Fig. 6. Evaluation results of defending against SPAC under different perturbation rates through training-time poisoning attacks.



(a) Spurious-Motif (b) MNIST-75sp (c) IMDB-B (d) IMDB-M

Fig. 7. Evaluation results of defending against SPAC under different perturbation rates through testing-time evasion attacks.



(a) AIDS (Poisoning attacks) (b) AIDS (Evasion attacks) (c) NCI1 (Poisoning attacks) (d) NCI1 (Evasion attacks)

Fig. 8. Experimental results of resistance to hybrid attacks on AIDS and NCI1 datasets. R&B, R&S, and B&S denote the poisoned samples generated by using ReWatt and BOA, ReWatt and SPAC, and BOA and SPAC, respectively.

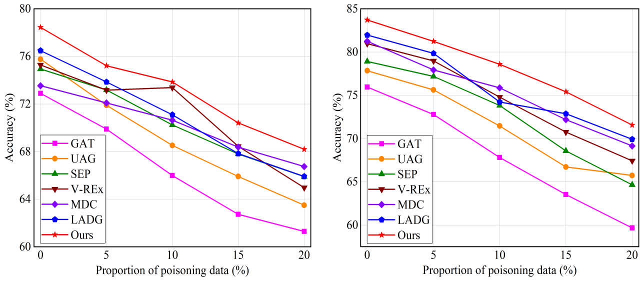
to the fact that the graph topological properties are relatively more stable than node features when making predictions on the MNIST-75sp dataset. However, GAT merges the imperfect knowledge from graph topological properties and features into uniform graph representations, which leads to poor generalization performance due to the learning of spurious correlations. In contrast, the proposed CASR learns invariant rationales from graph-structured data through causal correlation mining.

For the dataset transferability experiments, we evaluate the performance on the NCI1 and PC3 datasets. In previous experiments, we assume that the auxiliary data come from the same distribution as the testing data, whereas in the dataset transferability experiments, we perform extensive experiments when the auxiliary data are different distributions from the testing data. We explore the dataset transferability between NCI1 and PC3 in the case of ReWatt. In Fig. 9, the experimental results show that our scheme is still effective. By weakening the subtle nonlinear dependencies in graph-structured data, our scheme can learn the mapping relationship between rationales (*i.e*, informative graph topological representations) and labels. In this way, the graph-structured representations are less affected by the distribution bias and thus tend to have better generalization performance.

TABLE V

Evaluation Results of OOD Generalization on the MNIST-75sp Dataset. The Best Results Are Marked in Bold

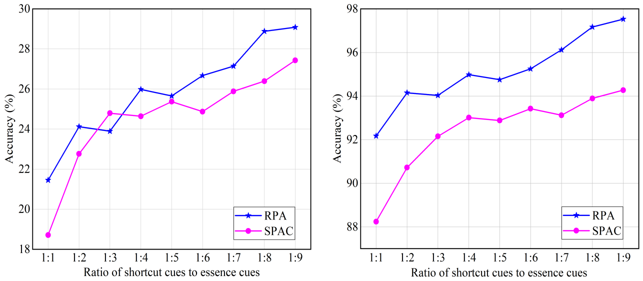
Compared models	MNIST-75sp			
	Clean	Noise	Color	Noise&Color
GAT	66.72±2.78	25.09±3.12	26.14±1.66	23.15±1.93
UAG	59.25±4.61	26.13±2.86	28.25±2.48	23.54±2.68
SEP	65.32±1.57	30.74±2.73	29.57±2.86	24.73±2.74
V-REx	62.19±3.80	29.18±4.05	28.46±3.72	25.92±3.35
MDC	66.14±2.53	31.45±1.56	31.73±1.93	28.71±2.06
LADG	68.82±3.85	30.96±1.70	32.62±3.25	27.55±1.84
Ours	**70.38±0.86**	**34.84±1.93**	**36.65±2.05**	**30.08±1.87**



(a) NCI1 (Poisoned PC3)

(b) PC3 (Poisoned NPI1)

Fig. 9. Evaluation results of dataset transferability between NCI1 and PC3 under the condition of ReWatt.



(a) MNIST-75sp

(b) AIDS

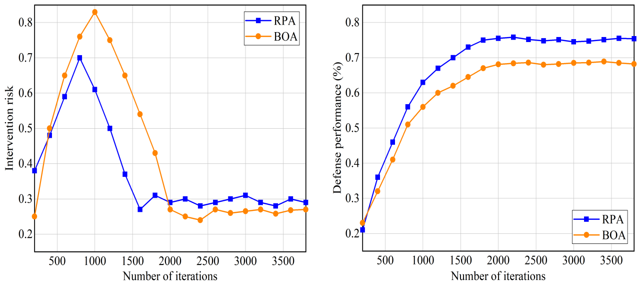
Fig. 10. Evaluation results of shortcut influence on the MNIST-75sp and AIDS datasets in the case of RPA and SPAC.

D. Evaluation of Shortcut Influence (RQ4)

In this section, we investigate the influence of spurious correlations by adjusting the proportion between shortcut cues and essence cues through weight assignment. Specifically, we observe the defensive capability of the proposed scheme against RPA and SPAC on the MNIST-75sp and AIDS datasets with different proportions. Fig. 10 illustrates the performance of our defense scheme against adversarial attacks as the percentage of spurious correlations decreases. The experimental results can be analyzed from the following two perspectives. From a global perspective, the general trend of the experimental results becomes better progressively, which implies that the suppression of shortcuts can strengthen the defense ability. From a local perspective, the reduction in the percentage of shortcuts in a small range does not necessarily lead to performance improvement. This phenomenon lies in the fact that shortcut cues and essence cues can be connected and interacted with each other. Therefore, the shortcut cues can be corrected under the guidance of essence cues to some extent.

E. Evaluation of Rationalization Role (RQ5)

Extensive experiments are conducted on the NCI1 dataset under RPA and BOA conditions to explore the role of rationalization. Specifically, we observe the variation of intervention risk and defense performance with the number of iterations during the consistency-driven rationalization. As shown in Fig. 11(a), the intervention risk first increases, then progressively decreases, and eventually converges to the steady state. Fig. 11(b) illustrates that as the number of iterations increases, the prediction accuracy of our scheme gradually increases until it reaches convergence. Moreover, there exist concealed correlations between intervention risk and defense performance, namely, the defense performance increases rapidly with the increase of intervention risk but grows slowly with the decrease of intervention risk. To probe this learning paradigm, we divide the entire co-training process into two phases, including the stable matching and the discriminative adaptation. The stable matching involves the initial training of f_1 to learn the causal consensus representation $\tilde{C}$. Since intervention risk can be viewed as the magnitude of the impact of stability constraints, the discriminative adaptation searches for rationales that satisfy the causal invariance principle. Consequently, f_2 rapidly performs distribution adaptation with the different rationales based on the construction of f_1. In the final training process, our goal is to ensure that if f_1 undergoes a small change, the rationales remain essentially unchanged compared to the initial training process. In this way, the desired rationalizations are learned from the construction of consistency-driven rationales, which is in accordance with the principle of causal invariance. This also implies that the consistency-driven optimal rationalization is able to consolidate the mapping relationships (*e.g*, $f_1 : G \rightarrow \tilde{C}$ and $f_2 : \tilde{C} \rightarrow \hat{Y}$) by capturing the informative representations. In conclusion, the above results and analysis demonstrate the effectiveness of the proposed causality-aligned structure rationalization, which reveals the substantial role of improving defensibility and generalization in graph-structured data.



(a)

(b)

Fig. 11. Evaluation results of rationalization role for our scheme on the NCI1 dataset under the condition of RPA and BOA. (a) The variation curve of intervention risk with the number of iterations. (b) The evolution of defense performance as training progresses.

VII. Conclusion

In this paper, we present a causality-aligned rationale construction scheme based on the reliability extrapolation (*i.e*, identifying shortcut or essence cues) to enhance the defensive capability and the generalization performance for GNNs. The interesting conclusions can be derived from these research results as follows: 1) a potential shortcut discovery aims to explore the patterns and characteristics of the original graph-structured data to produce multi-fold graph structures, which facilitates the diversity of the dataset to capture the complementary information between spurious correlations; 2) the causality-aligned rationale construction is designed from the perspective of different environments (*i.e*, interventional distributions) to distill the salient features that are informative and robust across these distributions, which can be exploited to discriminate stable representations from unstable ones for researchers in other application fields, such as intrinsic interpretability, adversarial defense, and OOD generalization of GNNs; 3) this work provides a meaningful view that the consistency-driven optimal rationalization can be utilized to further improve the security and robustness through a query feedback mechanism in GNN model; and 4) extensive experiments demonstrate that our proposed CASR has the distinct advantage of overall performance (*e.g*, defensiveness and stability) in graph-structured scenarios where the underlying distributions are governed by unknown disturbances. Furthermore, the rationale benefits can be reaped to reduce the risk of adverse effects through invariant extrapolation learning. In the near future, we are planning to investigate the cross-interaction mechanism between shortcuts and essences that can reliably and securely build GNN models with high-level interpretability under various complex data distributions.

Acknowledgment

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

References

[1] R. Cadène, C. Dancette, H. Ben-Younes, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases for visual question answering," in *Proc. NIPS*, 2019, pp. 839–850.

[2] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," in *Proc. ICML*, vol. 119, 2020, pp. 528–539.

[3] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[4] W. Wei and L. Liu, "Robust deep learning ensemble against deception," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 4, pp. 1513–1527, Jul. 2021.

[5] Z. Xi, R. Pang, S. Ji, and T. Wang, "Graph backdoor," in *Proc. USENIX Secur.*, 2021, pp. 1523–1540.

[6] D. Hendrycks et al., "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8320–8329.

[7] H. Li, X. Wang, Z. Zhang, and W. Zhu, "OOD-GNN: Out-of-distribution generalized graph neural network," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 1–14, Jul. 2022.

[8] S. Chang, Y. Zhang, M. Yu, and T. S. Jaakkola, "Invariant rationalization," in *Proc. ICML*, vol. 119, 2020, pp. 1448–1458.

[9] Y. Zhang, M. Humbert, B. Surma, P. Manoharan, J. Vreeken, and M. Backes, "Towards plausible graph anonymization," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2020, pp. 1–15.

[10] X. He, J. Jia, M. Backes, N. Z. Gong, and Y. Zhang, "Stealing links from graph neural networks," in *Proc. USENIX Secur.*, 2021, pp. 2669–2686.

[11] J. Mu, B. Wang, Q. Li, K. Sun, M. Xu, and Z. Liu, "A hard label black-box adversarial attack against graph neural networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2021, pp. 108–125.

[12] M. Fan, Z. Si, X. Xie, Y. Liu, and T. Liu, "Text backdoor detection using an interpretable RNN abstract model," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4117–4132, 2021.

[13] Z. Yan, J. Wu, G. Li, S. Li, and M. Guizani, "Deep neural backdoor in semi-supervised learning: Threats and countermeasures," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4827–4842, 2021.

[14] B. Knyazev, G. W. Taylor, and M. R. Amer, "Understanding attention and generalization in graph neural networks," in *Proc. NIPS*, 2019, pp. 4204–4214.

[15] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, "Recognizing predictive substructures with subgraph information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 14, 2021, doi: 10.1109/TPAMI.2021.3112205.

[16] X. Wang, Y. Wu, A. Zhang, X. He, and T. Chua, "Towards multi-grained explainability for graph neural networks," in *Proc. NIPS*, 2021, pp. 18446–18458.

[17] B. Wang, Y. Li, and P. Zhou, "Bandits for structure perturbation-based black-box attacks to graph neural networks with theoretical guarantees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13369–13377.

[18] S. Vakulenko, J. D. F. Garcia, A. Polleres, M. de Rijke, and M. Cochez, "Message passing for complex question answering over knowledge graphs," in *Proc. CIKM*, 2019, pp. 1431–1440.

[19] J. Kaddour, Y. Zhu, Q. Liu, M. J. Kusner, and R. Silva, "Causal effect inference for structured treatments," in *Proc. NIPS*, 2021, pp. 24841–24854.

[20] W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang, "Graph structure learning for robust graph neural networks," in *Proc. SIGKDD*, 2020, pp. 66–74.

[21] W. Zhu, L. Lu, J. Xiao, M. Han, J. Luo, and A. P. Harrison, "Localized adversarial domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7098–7108.

[22] G. Bouritsas, F. Frasca, S. Zafeiriou, and M. M. Bronstein, "Improving graph neural network expressivity via subgraph isomorphism counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 657–668, Jan. 2023.

[23] B. Feng, Y. Wang, and Y. Ding, "UAG: Uncertainty-aware attention graph neural network for defending adversarial attacks," in *Proc. AAAI*, 2021, pp. 7404–7412.

[24] Y.-X. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, "Discovering invariant rationales for graph neural networks," in *Proc. ICLR*, 2022, pp. 1–22.

[25] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 249–270, Jan. 2022.

[26] P. Ma, Z. Ji, Q. Pang, and S. Wang, "NoLeaks: Differentially private causal discovery under functional causal model," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2324–2338, 2022.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLR*, 2014.

[28] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. NIPS*, 2019, pp. 14747–14756.

[29] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, vol. 70, 2017, pp. 3319–3328.

[30] D. Luo et al., "Parameterized explainer for graph neural network," in *Proc. NIPS*, 2020, pp. 1–12.

[31] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "GraphLIME: Local interpretable model explanations for graph neural networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, pp. 6968–6972, Jul. 2023, doi: 10.1109/TKDE.2022.3187455.

[32] B. Sánchez-Lengeling et al., "Evaluating attribution for graph neural networks," in *Proc. NIPS*, 2020, pp. 1–13.

[33] L. Faber, A. K. Moghaddam, and R. Wattenhofer, "When comparing to ground truth is wrong: On evaluating GNN explanation methods," in *Proc. SIGKDD*, 2021, pp. 332–341.

[34] L. Peng et al., "Reverse graph learning for graph neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 5, 2022, doi: 10.1109/TNNLS.2022.3161030.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[36] S. Agarwal, S. Jabbari, C. Agarwal, S. Upadhyay, S. Wu, and H. Lakkaraju, "Towards the unification and robustness of perturbation and gradient based explanations," in *Proc. ICML*, vol. 139, 2021, pp. 110–119.

[37] T. Bai, J. Zhao, and B. Wen, "Guided adversarial contrastive distillation for robust students," *IEEE Trans. Inf. Forensics Security*, early access, Jan. 18, 2023, doi: 10.1109/TIFS.2023.3237371.

[38] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang, "Inference attacks against graph neural networks," in *Proc. USENIX Secur.*, 2022, pp. 1–18.

[39] H. Dai et al., "Adversarial attack on graph structured data," in *Proc. ICML*, vol. 80, 2018, pp. 1123–1132.

[40] Y. Ma, S. Wang, T. Derr, L. Wu, and J. Tang, "Graph adversarial attack via rewiring," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 1161–1169.

[41] K. Xu et al., "Topology attack and defense for graph neural networks: An optimization perspective," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3961–3967.

[42] S. Wang, S. Nepal, A. Abuadbba, C. Rudolph, and M. Grobler, "Adversarial detection by latent style transformations," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1099–1114, 2022.

[43] L. Lin, E. Blaser, and H. Wang, "Graph structural attack by perturbing spectral distance," in *Proc. SIGKDD*, 2022, pp. 989–998.

[44] W. Ding, X. Wei, R. Ji, X. Hong, Q. Tian, and Y. Gong, "Beyond universal person re-identification attack," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3442–3455, 2021.

[45] Z. Xiao, J. Shen, X. Zhen, L. Shao, and C. Snoek, "A bit more Bayesian: Domain-invariant learning with uncertainty," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 11351–11361.

[46] R. Wang, M. Yi, Z. Chen, and S. Zhu, "Out-of-distribution generalization with causal invariant transformations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 375–385.

[47] W. Lin, H. Lan, H. Wang, and B. Li, "OrphicX: A causality-inspired latent variable model for interpreting graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13719–13728.

[48] R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters, "A causal framework for distribution generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6614–6630, Oct. 2022, doi: 10.1109/TPAMI.2021.3094760.

[49] D. Krueger et al., "Out-of-distribution generalization via risk extrapolation (REx)," in *Proc. ICML*, vol. 139, 2021, pp. 5815–5826.

[50] M. Federici, R. Tomioka, and P. Forré, "An information-theoretic approach to distribution shifts," in *Proc. NIPS*, 2021, pp. 17628–17641.

[51] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Commun. ACM*, vol. 62, no. 3, pp. 54–60, Feb. 2019.

[52] X. Wang, X. He, Y. Cao, M. Liu, and T. Chua, "KGAT: Knowledge graph attention network for recommendation," in *Proc. SIGKDD*, 2019, pp. 950–958.

[53] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10045–10067, Dec. 2022, doi: 10.1109/TPAMI.2021.3132674.

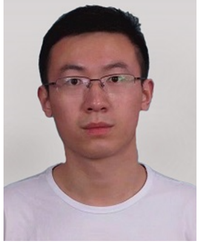
[54] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. NIPS*, 2019, pp. 9240–9251.

[55] C. Morris, N. M. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann, "TUDataset: A collection of benchmark datasets for learning with graphs," 2020, *arXiv:2007.08663*.

[56] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018, pp. 1–12.

[57] J. Wu, X. Chen, K. Xu, and S. Li, "Structural entropy guided graph hierarchical pooling," in *Proc. ICML*, vol. 162, 2022, pp. 24017–24030.

[58] Y. Wald, A. Feder, D. Greenfeld, and U. Shalit, "On calibration and out-of-domain generalization," in *Proc. NIPS*, 2021, pp. 2215–2227.



Ju Jia received the Ph.D. degree in cyberspace security from Wuhan University, Wuhan, China, in 2021. He was a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Research Associate Professor with the School of Cyber Science and Engineering, Southeast University, Nanjing, China. His current research interests include data security, multimedia content security, and artificial intelligence security.

Siqi Ma (Member, IEEE) received the B.S. degree in computer science from Xidian University, Xi'an, China, in 2013, and the Ph.D. degree in information systems from Singapore Management University in 2018. She was a Research Fellow with the Distinguished System Security Group, CSIRO. She was a Lecturer with The University of Queensland. She is currently a Senior Lecturer with the University of New South Wales, Canberra Campus, Australia. Her current research interests include data security, the IoT security, and software security.

Yang Liu (Senior Member, IEEE) received the B.Comp. degree (Hons.) from the National University of Singapore (NUS) in 2005 and the joint Ph.D. degree from NUS and MIT in 2010. He was a Post-Doctoral Researcher with NUS and MIT. In 2012, he joined Nanyang Technological University (NTU). He is currently a Full Professor and the Director of the Cybersecurity Laboratory, NTU. He specializes in software verification, security, and software engineering. He has more than 400 publications in top-tier conferences and journals. His current research interests include the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security. He has received several prestigious awards, including the MSRA Fellowship, the TRF Fellowship, the Nanyang Assistant Professor, the Tan Chin Tuan Fellowship, the Nanyang Research Award in 2019, the ACM Distinguished Speaker, and the NRF Investigatorship. He has received 15 best paper awards and one most influence system award in top software engineering conferences, like ASE, FSE, and ICSE.

Lina Wang (Member, IEEE) received the B.S. degree from the Hefei University of Technology, Hefei, China, in 1986, and the M.S. and Ph.D. degrees from Northeastern University, Shenyang, China, in 1989 and 2001, respectively. She is currently a Professor with the Cyber Science and Engineering School, Wuhan University. Her current research interests include multimedia content security, data security, and machine learning methods in network security detection.

Robert H. Deng (Fellow, IEEE) is currently the AXA Chair Professor of cybersecurity, the Director of the Secure Mobile Centre, and the Deputy Dean of the Faculty and Research, School of Information Systems, Singapore Management University (SMU). His current research interests include data security and privacy, network security, and system security. He is a fellow of the Academy of Engineering Singapore. He received the Outstanding University Researcher Award from the National University of Singapore, the Lee Kuan Yew Fellowship for Research Excellence from SMU, and the Asia–Pacific Information Security Leadership Achievements Community Service Star from the International Information Systems Security Certification Consortium. He served as the Steering Committee Chair for the ACM Asia Conference on Computer and Communications Security. He serves/served on many editorial boards and conference committees, including the editorial boards of *ACM Transactions on Privacy and Security*, IEEE SECURITY AND PRIVACY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and *Journal of Computer Science and Technology*.