GRACED: A Plug-and-Play Solution for Certifiable Graph Classification

Xiaoyu Liang*, Haohua Du*, He Lu*, and Fei Shang†
*Beihang University, China
†University of Science and Technology of China, China
Email: {xiaoyuliang, duhaohua, herlu}@buaa.edu.cn, feishang@mail.ustc.edu.cn

***Abstract*—With the widespread application of machine learning-based graph classification models in fields such as biology and economics, there has been a growing number of attacks aimed at perturbing classification results. Although current defense methods, such as randomized smoothing, have achieved some success, their practical applicability remains limited due to the need to modify classification models to ensure accuracy.**

In this paper, we propose a novel defense method—GRACED, which provides theoretical guarantees for the accuracy and robustness of graph classification without requiring knowledge of the attacker's capabilities or the classification model. The key idea behind our method is to leverage the denoising ability of feature diffusion models for adversarial data purification. We then demonstrate that this randomized purification approach can ensure certified robustness under specific attack budgets. Extensive experiments confirm our theoretical findings and show that graph classifiers using GRACED significantly outperform state-of-the-art classifiers. For instance, the accuracy on MUTAG improved by 11%, and the best results on IMDB showed a 14% increase.

***Index Terms*—graph classification, adversarial attack, diffusion model**

I. Introduction

Graph Neural Network (GNN)-based Graph classifiers [1] have been shown to be vulnerable to subtle modifications of the graph, which compromises its robustness when applied to tasks such as protein property analysis [2]. For example, perturbations to non-essential molecular structures can lead to incorrect property predictions [3].

Such adversarial attacks on graph classifiers have led to the development of various defense methods. Existing defenses [4]–[7] either offer empirical protection but are vulnerable to new attacks or are incompatible with black-box models. In contrast, Randomized Smoothing (RS) assumes black-box attacks, focuses on adversary capabilities, and provides robustness certificates through data randomizing and sampling, making it versatile across model architectures. However, RS methods require retraining or fine-tuning as standard classifiers cannot handle noisy samples, limiting their suitability for diverse adversaries and black-box models. Diffusion models [8]–[10], leveraging generative capabilities of large models, is capable of handling noisy data but struggle to preserve structural stability, which undermines classification performance when used for adversarial defense. While both diffusion models and RS can somewhat enhance model robustness, they face practical limitations, including a lack of plug-and-play capability, reliance on additional attacker information, and suboptimal performance on graph classification tasks.

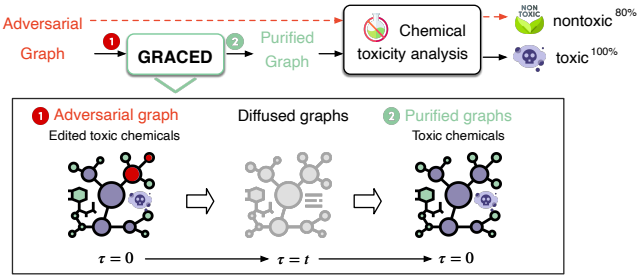

Fig. 1. **GRACED illusatration**: An edited toxic chemical might be misclassified as nontoxic, while such an adversarial graph with GRACED purification can be correctly identified with a guarantee.

To address these challenges, we propose GRACED, a certified defense ensuring high-accuracy and guatanteed robustness for black-box graph classifiers against unknown attacks without requiring knowledge about attacker or model. GRACED utilizes the denoising ability of diffusion models [8] to purify the adversarial graph before obtaining predictions with any GNN models.

Adapting diffusion models for adversarial defense is challenging due to: i) Difficulty in constraining the stochasticity in the forward and reverse processes of diffusion model while ensuring graph structure and node feature stability. ii) The challenge of purifying implicit graph features embedded by various graph classifier using only the explicit structure and attributes, capturing the unique contribution of each.

Solutions: i) We design a marginal-distribution based noise addition and removal process that maintainsgraph structural integrity and node attribute features, with fine-grained noise scales to adapt to attackers with different budgets. ii) We developed a step-wise diffusion and denoising algorithm based on the multiplication theorem, independently analyzing the impact of structure and attribute on the features.

The main contributions in **GRACED** summarize as follows:

- We present GRACED - a plug-and-play solution to guarantee GRAph classification with CErtifiable robustness via a Diffusion model, effectively tackling the verifiable robustness of black-box graph classification models.
- We design a graph diffusion model based on D3PM [8] to purify the adversarial graph into a benign graph. We elaborate on the forward diffusion process and the reverse

denoising process of the diffusion model to preserve the structure stability of the graph.
- We have validated the efficacy of our approach through comprehensive testing on real-world datasets. The accuracy improves 11% on MUATG and 14% on IMDB dataset compared to the state-of-the-art.

II. GRACED, A Certified Graph Classification Solution

In this section, we present GRACED, a plug-and-play method providing robustness certification for arbitrary black box graph classification models.

As illustrated in Fig. 2, GRECED is composed of GRAD, a discrete diffusion model that purifies the graph feature while preserving graph structure stability, and GRACE, a certificate generator providing provable robustness through theoretical proof.

A. GRACED Framework

The objective of GRACED is to guarantee the classification result for arbitrary classifiers under different attacks. Given a set of graphs $\mathcal{G} = \{G\}$, a graph classifier like Graph Isomorphism Network (GIN) [11] aims to predict a label for the entire graph, i.e., to learn a function $f_\theta : \mathcal{G} \to \mathcal{K}$ with $\mathcal{K}$ being the label set. Our work focuses on graphs with binary entries in attribute matrix $X_{n\times d}$ and adjacent matrix $A_{n\times n}$.

A well-trained graph classifier can be fooled by samples with imperceptible perturbations, i.e., adversarial examples [12]–[14]. Our threat model considers adversarial graph example $\tilde{G} = (\tilde{X}, \tilde{A})$ with attributive and structural perturbation $\delta = (\delta_X, \delta_A)$. We focus solely on the attack budget and treat the attack strategy as a black box.

Specifically, an attacker's budget Δ is to limit the magnitude of changes in X and A:

$$\begin{aligned}\Delta = \{(\delta_X, \delta_A) :&||\delta_X^+||_0 \le \Delta_X^+, ||\delta_X^-||_0 \le \Delta_X^-,\\ &||\delta_A^+||_0 \le \Delta_A^+, ||\delta_A^-||_0 \le \Delta_A^-\}.\end{aligned} \tag{1}$$

Δ_X specifies a ℓ_0-ball around X containing all the samples with Δ_X^+ additional attributes and Δ_X^- fewer attributes, while Δ_A constraints the structural perturbation with Δ_A^+ injected edges and Δ_A^- deleted edges. The perturbation on two variables is equivalent to bit flipping on the attribute and adjacency matrix.

To provide certified defense under the attack budget, the common approach is to randomly perturb the input and report the output corresponding to the "majority vote" on the randomized sample, i.e. Randomized Smoothing (RS) [15], [16].

Considering a graph classifier $f_\theta(G)$, RS predicts the label with a smoothed base classifier, noted as $g_\theta(G)$:

$$g_\theta(G) := \arg_y \max \Pr_{\tilde{G}\sim\phi(G)}[f_{\theta^*}(\tilde{G}) = y], \tag{2}$$

where $\phi(\cdot)$ returns a randomized version of the input graph. To preserve the sparsity in graph data, Ref. [15] proposed a randomization scheme $\phi(Z)$ with $Z \in \{X, A\}$:

$$\Pr(\phi(Z) \neq Z) = p_+^{(1-z)} p_-^z. \tag{3}$$

Such randomization is to apply data-dependent Bernoulli noise $\epsilon^Z \sim \text{Ber}(p = p_+^{(1-z)} p_-^z)$ on variable Z with $p = p_-$ if $z = 1$ and $p = p_+$ if $z = 0$. However, vanilla RS in [15] needs robust training on white box models with data randomized with $\phi(G)$, since standard GNNs are not trained for randomized data.

By contrast, GRACED decouples the classifier from the randomization methods, making it applicable to black box GNNs. The key idea is to instantiate the Randomized Smoothing framework with a denoiser and a black box classifier. Specifically, the $f_{\theta^*}(G)$ in Eq. 2 can be paraphrased as:

$$f_{\theta^*}(\tilde{G}) := f_\theta(\mathcal{D}(\tilde{G})). \tag{4}$$

$\mathcal{D}(\tilde{G})$ is a graph denoiser based on the discrete diffusion model which will be introduced in subsection II-B. The separation of the classifier and randomized data allows GRACED to be plug-and-play, providing certified defense.

B. GRAD, A Graph Diffusion Model

GRAD takes a graph $\tilde{G}$ with adversarial attributes $\tilde{X}$ and structure $\tilde{A}$ as input and purifies the graph feature to mitigate the impact of malicious data, providing denoised graph $\bar{G} = (\bar{X}, \bar{A})$ as output. We leverage the powerful denoising capability of D3PM [8] to reconstruct graphs with complex attributes and structures while preserving structure stability.

The forward process of GRAD adds data-dependent Bernoulli noise on $Z \in \{X, A\}$, which is to "transit" Z from one state to the other, controlled by Q_t as follow:

$$q(Z_t|Z_0) = \text{Cat}(Z_t; p = Z_0\bar{Q}_t), \tag{5}$$

where $\text{Cat}(\cdot)$ denotes the categorical distribution, $Q_t = q(Z_t|Z_{t-1})$ is the categorical transition matrix and $\bar{Q}_t = Q_1 Q_2 \cdots Q_t$.

Using standard diffusion notations, $G_0 = (X_0, A_0)$ is sampled from original data distribution and $\beta_t = 1 - \alpha_t$ is the noise schedule. The forward diffusion process on X and A is controlled by transition matrix Q_t:

$$Q_t^X = \alpha_t \mathbf{I} + \beta_t \mathbf{1}\mathbf{m}^{\mathbf{X}}, \tag{6}$$

$$Q_t^A = \alpha_t \mathbf{I} + \beta_t \mathbf{1}\mathbf{m}^{\mathbf{A}}, \tag{7}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{1}$ is the one-valued vector, $\mathbf{m}^{\mathbf{Z}}$ is the empirical marginal distribution of of variable Z. As we handle matrices with $\{0, 1\}$ entries, the state transition of D3PM is equivalent to bit-flipping. Diffused graph distribution at timestamp (t_x, t_a) can be written as:

$$\begin{aligned}&q(X_{t_x}, A_{t_a}|X_0, A_0)\\ &= \text{Cat}(X_{t_x}, A_{t_a}; p_X = X_0 \oplus \epsilon_{t_x}^X, p_A = A_0 \oplus \epsilon_{t_a}^A),\end{aligned} \tag{8}$$

where $\epsilon_{t_z}^Z \sim \text{Ber}(p = (\bar{\beta}_{t_z}\mathbf{m_1}^Z)^{1-z}(\bar{\beta}_{t_z}\mathbf{m_0}^Z)^z)$ for $Z \in \{X, A\}$. Here, $\bar{\beta}_t = 1 - \bar{\alpha}_t$ and $\bar{\alpha}_t = \prod_{\tau=1}^t \alpha_\tau$. The asynchronous noise $\epsilon_{t_x}^X$ and $\epsilon_{t_a}^A$ set different noise granularity for attributes and structures.

The reverse denoise process is to predict the noise on the graphs at timestamp t. To estimate the joint distribution of noises, denoising network $\epsilon_\theta(X_{t_x}, A_{t_a}, t_x, t_a)$ is employed to

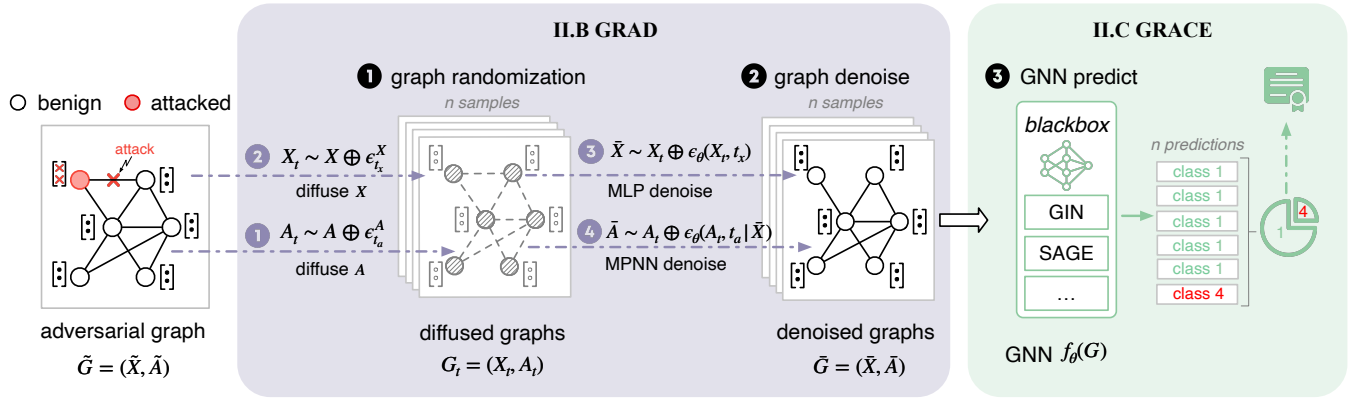

Fig. 2. **Pipeline**: GRACED purifies the adversarial node attributes and graph structure through GRAD and generates a robustness certificate by majority votes via GRACE. $\tilde{G}$ is first randomized on X and A to get diffused graphs, then denoised separately, and finally classified by black box model.

minimize the expected gap between the predicted noise and the actual noise:

$$\mathbb{E}_{X_0, A_0, \epsilon_{t_x}^X, \epsilon_{t_a}^A, t_x, t_a} ||\epsilon_\theta(X_{t_x}, A_{t_a}, t_x, t_a) - [\epsilon_{t_x}^X, \epsilon_{t_a}^A]||. \quad (9)$$

The $\epsilon_\theta(X_{t_x}, A_{t_a}, t_x, t_a)$ in GRAD is composed of Multi-Layer Perceptron attribute denoiser $\epsilon_\theta^X(X_{t_x}, t_x)$ and Message-Passing Neural Network structure denoiser $\epsilon_\theta^A(A_{t_a}, t_a | \bar{X}_0)$ conditioned on reconstructed node attribute. Practically, GRAD is trained with cross-entropy loss between the predicted probabilities and actual variables:

$$\text{Loss}(\bar{p}^G, G) = \ell_{\text{CE}}(\bar{p}^X, X) + \lambda \ell_{\text{CE}}(\bar{p}^A, A), \quad (10)$$

where $\bar{p}^G = (\bar{p}^X, \bar{p}^A)$ satisfying $\bar{p}^X = X_{t_x} \oplus \Pr(\epsilon_\theta^X(X_{t_x}, t_x))$ and $\bar{p}^A = A_{t_a} \oplus \Pr(\epsilon_\theta^A(A_{t_a}, t_a | \bar{X}_0))$. λ controls the relative importance of structure over attributes.

C. GRACE, A Graph Certification

The provable defense is provided by a robustness certificate which guarantees the prediction under a certain attack budget. Let y^* be the ground truth label of G, and $\tilde{G}$ is a randomized version of G. Generating a certificate is to ensure that $f_\theta(\tilde{G}) = y^*$. The robustness certificate is defined following previous work [15], [17], [18]:

$$\rho_{G,\tilde{G}}(p, y^*) = \min_{\bar{f}_\theta : \Pr(\bar{f}_\theta(G) = y^*) = p} \Pr(\bar{f}_\theta(\tilde{G}) = y^*). \quad (11)$$

Then $\rho_{G,\tilde{G}}(p, y^*) \leq \Pr(\bar{f}_\theta(\tilde{G}) = y^*)$ is a lower bound of the probability p of $f_\theta(\tilde{G}) = y^*$. Given an attack budget Δ, if it satisfied:

$$\min_{\tilde{G}} \rho_G(p, y^*) > \max_{\tilde{G}} \Pr(\bar{f}_\theta(\tilde{G}) = y \neq y^*), \quad (12)$$

where $||\tilde{G} - G||_0 \leq \Delta$, the predicted label is certifiably robust, for the lower bound of the most likely class is higher than any other classes.

The common approach for certification is to compute the lower bound $\underline{p_{y^*}}(G)$ and upper bound $\overline{p_{y \neq y^*}}(G)$ based on the Clopper-Pearson Bernoulli confidence interval using Monte-Carlo samples from $\tilde{G} \in G + \Delta$ [15], [16].

Our key idea is to divide the graph space into disjoint regions $\mathcal{G} = \bigcup_i \mathcal{R}_i, \text{s.t.} \mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ of satisfying $\Pr(\phi(G) = Y) / \Pr(\phi(\tilde{G}) = Y) = c_i$ such that we can apply the Neyman-Pearson Lemma [19] to derive the lower bound of the most-likely class. Given such partition of $\mathcal{G}$, the certificate in Eq. 11 is equivalent to the following Linear Program Problem according to [17]:

$$\min_f f^T \tilde{r} \quad \text{s.t.} f^T r = p, \quad 0 \leq f \leq 1, \quad (13)$$

where f is the vector we are optimizing over, and probability r is a vector where $r_i = \Pr(\phi(\tilde{G}) \in \mathcal{R}_i)$. The solution can be obtained by a greedy algorithm: sort the $\mathcal{R}_i$ such that $c_1 \geq c_2 \geq \cdots \geq c_I$, then iteratively assign $f_i = 1$ until the budget is met.

Since we use the same threat model and randomization scheme as in Sparse Smoothing [15], to derive the certificate, the minor technicality required for GRACE is to map the diffusion timestamp to the randomization parameter. To be specific, for $Z = \{X, A\}$, the randomization scheme in Sparse Smoothing is to add Bernoulli noise $\epsilon^Z \sim \text{Ber}(p = p_+^{(1-z)} p_-^z)$. The diffusion process in GRAD is to add Bernoulli noise $\epsilon_{t_z}^Z \sim \text{Ber}(p = (\bar{\beta}_{t_z} \mathbf{m_1}^Z)^{1-z} (\bar{\beta}_{t_z} \mathbf{m_0}^Z)^z)$. It is obvious that $p_+ = \bar{\beta}_{t_z} \mathbf{m_1}^Z$ and $p_- = \bar{\beta}_{t_z} \mathbf{m_0}^Z$.

III. Evaluation

A. Experimental Setup

We utilize three bioinformatics graph datasets (MUTAG [20], NCI1 [21], PROTEINS [22]) and one social network dataset (IMDB-BINARY [23]) for evaluation. Graphs in bioinformatics datasets represent compounds or proteins, with labels indicating chemical properties like mutagenicity. In the IMDB dataset, each ego-network of an actor/actress is classified into different movie genres. We use GIN [11] and GraphSAGE [24] for graph classification.

Sparse Smoothing [15] proposed a data-dependent sparsity-aware randomization method. Bernoulli noise is added to A and X for randomization in [18]. Hierarchical Smoothing [25] add noise to randomly selected nodes. Both Bernoulli Smoothing and Hierarchical Smoothing are designed to defend against

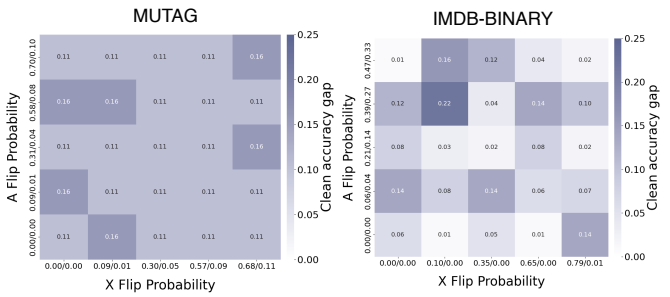

Fig. 3. **Clean accuracy gap**: Each heatmap shows the clean accuracy gap between GRACED and Sparse Smoothing per dataset, with noise scales for attributes and adjacent matrices on the horizontal and vertical axes.

singular perturbation on attributes or structures, so we make adaptations to extend their scopes.

During the training of GRAD, 10% graphs are used for testing while the remaining are split into 90% and 10% for training and validation. We use the same split to train the GNN in both standard and robust training. We set $T_x = T_a = 500$ and $\lambda = 5$ for GRAD. When testing the GRACED and Sparse Smoothing, we sample 1000 randomized graphs for classification.

B. Experimental Results

Like [15], [18], we report the *clean accuracy* and *certified accuracy*. Naïve$^{\phi}$ and GRACED represent the benign classifier under attack, and our model, respectively. Sparse, Hier. and Ber. represent three different RS method.

When using GraphSAGE for classification, GRACED improves by 8% and 7% compared to Naïve$^{\phi}$ and Sparse on the NCI1 dataset, respectively. On IMDB, the improvements relative to Naïve$^{\phi}$ and Sparse are 20% and 10% (under joint perturbation). Table I shows the clean accuracy of different models under singular and joint perturbation on X and A classified by GIN.

TABLE I
CLEAN ACCURACY UNDER DIFFERENT PERTURBATION

Type		MUTAG	NCI1	PROTEINS	IMDB
Attr.& Adj.	Naïve$^{\phi}$	0.58	0.49	0.54	0.52
	Sparse	0.68	0.60	0.55	0.49
	Hier.^{A}	0.52	**0.64**	0.63	0.48
	Ber.^{X}	0.74	0.55	**0.67**	0.51
	GRACED	**0.79**	**0.64**	**0.67**	**0.63**
Attr.	Naïve$^{\phi}$	0.53	0.48	0.53	0.51
	Sparse	0.68	0.32	**0.49**	0.66
	Hier.	0.73	0.51	0.41	0.57
	GRACED	**0.78**	**0.59**	**0.61**	**0.63**
Adj.	Naïve$^{\phi}$	0.53	0.46	0.53	0.54
	Sparse	0.63	0.43	**0.61**	0.66
	Ber.	0.63	0.55	0.51	0.53
	GRACED	**0.78**	**0.62**	**0.61**	**0.75**

Note: The randomization parameters are set as the noise scale when diffusion timestamp $t = 300$. Hier.^{A} denotes adaptation of hierarchical smoothing with ϵ^{Z} set the same as sparse method and corruption ratio $p = 0.8$. Ber.^{X} is the adaptation of Bernoulli smoothing with $\epsilon^{Z} = \text{Ber}(p = \frac{1}{2}(p^{+} + p^{-}))$.

GRACED outperforms or matches baselines without costly robustness training. Attacked Naive models perform near random guessing, highlighting standard models' inability to classify randomized samples, while GRAD effectively denoises data.

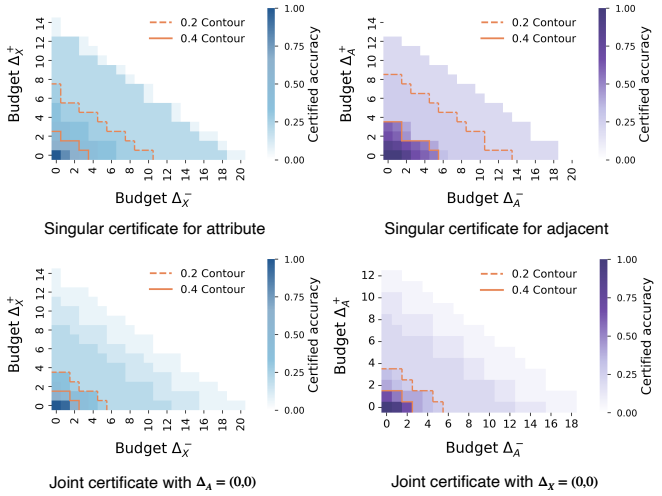

Fig. 4. **Certificate**: The top row depicts singular certificates, and the bottom shows joint perturbation defense. Blue and purple heatmaps represent certificates for node attributes and sctructure, respectively.

Fig. 3 shows the clean accuracy gap between GRACED and Sparse Smoothing on MUTAG and IMDB-BINARY, evaluated using GIN. The numbers on the axis represent the Bernoulli noise parameters $\text{Ber}(p = p_{+}^{(1-z)} p_{-}^{z})$, abbreviated as p_{-}/p_{+}. From low to high, the noise scales correspond to the diffusion timestamps $\{0, 100, 200, 300, 350\}$.

The gap between GRACED and the Sparse method is positive across various randomized parameter settings. On the MUTAG dataset, GRACED achieves consistently superior accuracy with an average improvement of 11.79%, while on the IMDB dataset, the average accuracy improvement is 7.84%

Fig. 4 shows the certified accuracy of GIN on the MUTAG dataset. The singular certificates are derived using the randomization scheme $\epsilon^{X} = 0.57/0.09$ and $\epsilon^{A} = 0.58/0.08$. The joint certificate uses the same setting for joint randomization. In certified accuracy experiments,we sample $n_0 = 10$ graphs to estimate the class likelihood and $n_1 = 10000$ graphs to derive robustness certificates.

Experimental results show our method certifies large attack budgets and achieves high provable accuracy. The axes represent attack deletion (Δ^{-}) and additin (Δ^{+}) for attributes/edges. With $p_{-} > p_{+}$,certified Δ^{-} exceeds Δ^{+}.

IV. CONCLUSION

In this paper, we present a plug-and-play defense against graph modification attacks, providing robustness guarantees for any black box graph classifier. We leverage the denoising capability of the discrete diffusion model to extract the core features of different graph types. The robustness certificate is then derived through the theoretical proof to ensure the prediction under different attack budgets. The evaluation results demonstrate that GRACED shows an accuracy improvement over the state-of-the-art on multiple datasets.

REFERENCES

[1] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dong-Sheng Cao, Jian Wu, and Tingjun Hou, "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *J. Cheminformatics*, vol. 13, no. 1, pp. 12, 2021.

[2] Kanchan Jha, Sriparna Saha, and Hiteshi Singh, "Prediction of protein–protein interaction using graph neural networks," *Scientific Reports*, vol. 12, no. 1, pp. 8360, 2022.

[3] Daniel Zügner, Oliver Borchert, Amir Akbarnejad, and Stephan Günnemann, "Adversarial attacks on graph neural networks: Perturbations and their patterns," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 5, pp. 57:1–57:31, 2020.

[4] Lukas Gosch, Simon Geisler, Daniel Sturm, Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann, "Adversarial training for graph neural networks: Pitfalls, solutions, and new directions," in *Proc. Adv. Neural Informat. Process. Syst*, Dec. 2023, pp. 58088–58112.

[5] Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis, "All you need is low (rank): Defending against adversarial attacks on graphs," in *Proc. 30th Int. Conf. Web Search Data Mining*, Feb. 2020, pp. 169–177.

[6] Fuli Feng, Xiangnan He, Jie Tang, and Tat-Seng Chua, "Graph adversarial training: Dynamically regularizing based on graph structure," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2493–2504, 2021.

[7] Aleksandar Bojchevski and Stephan Günnemann, "Certifiable robustness to graph perturbations," in *Proc. Adv. Neural Informat. Process. Syst*, Dec. 2019, pp. 8317–8328.

[8] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg, "Structured denoising diffusion models in discrete state-spaces," in *Proc. Adv. Neural Informat. Process. Syst*, Dec. 2021, pp. 17981–17993.

[9] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon, "Permutation invariant graph generation via score-based generative modeling," in *Proc. 23th Int. Conf. Artif. Intell. and Statist.*, Aug. 2020, pp. 4474–4484.

[10] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang, "Score-based generative modeling of graphs via the system of stochastic differential equations," in *Proc. Int. Conf. Mach. Learn.*, July 2022, pp. 10362–10383.

[11] Xiyuan Wang and Muhan Zhang, "How powerful are spectral graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, July 2022, pp. 23341–23362.

[12] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song, "Adversarial attack on graph structured data," in *Proc. Int. Conf. Mach. Learn.*, July 2018, pp. 1123–1132.

[13] Jiaming Mu, Binghui Wang, Qi Li, Kun Sun, Mingwei Xu, and Zhuotao 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.

[14] He Zhang, Bang Wu, Xiangwen Yang, Chuan Zhou, Shuo Wang, Xingliang Yuan, and Shirui Pan, "Projective ranking: A transferable evasion attack method on graph neural networks," in *Proc. 30th ACM Int. Conf. Informat. Knowl. Manage*, Nov. 2021, pp. 3617–3621.

[15] Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann, "Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more," in *Proc. Int. Conf. Mach. Learn.*, July 2020, pp. 1003–1013.

[16] Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn.*, June 2019, pp. 1310–1320.

[17] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S. Jaakkola, "Tight certificates of adversarial robustness for randomly smoothed classifiers," in *Proc. Adv. Neural Informat. Process. Syst*, Dec. 2019, pp. 4911–4922.

[18] Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong, "Certified robustness of graph neural networks against adversarial structural perturbation," in *Proc. 27th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, Aug. 2021, pp. 1645–1653.

[19] Jerzy Neyman and Egon Sharpe Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philos. Trans. R. Soc. Lond. Ser. A-Math. Phys. Eng. Sci.*, vol. 231, no. 694-706, pp. 289–337, 1933.

[20] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch, "Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity," *J. Medicinal Chemistry*, vol. 34, no. 2, pp. 786–797, 1991.

[21] Nikil Wale, Ian A. Watson, and George Karypis, "Comparison of descriptor spaces for chemical compound retrieval and classification," *Knowl. Inf. Syst.*, vol. 14, no. 3, pp. 347–375, 2008.

[22] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alexander J. Smola, and Hans-Peter Kriegel, "Protein function prediction via graph kernels," in *Proc. 13th Int. Conf. Intell. Syst. Mol. Biol.*, 2005, pp. 47–56.

[23] Pinar Yanardag and S. V. N. Vishwanathan, "Deep graph kernels," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, Aug. 2015, pp. 1365–1374.

[24] William L. Hamilton, Zhitao Ying, and Jure Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Informat. Process. Syst*, Dec. 2017, pp. 1024–1034.

[25] Yan Scholten, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann, "Hierarchical randomized smoothing," in *Proc. Adv. Neural Informat. Process. Syst*, Dec. 2023.