



2025 IEEE INTERNATIONAL CONFERENCE ON
ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP 2025)

April 06 - 11, 2025 Hyderabad, India



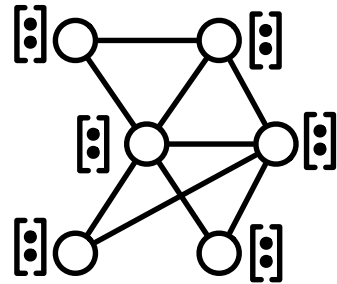
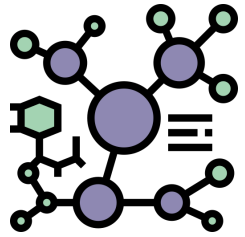
GRACED: A Plug-and-Play Solution for Certifiable Graph Classification

Xiaoyu Liang (BUAA)

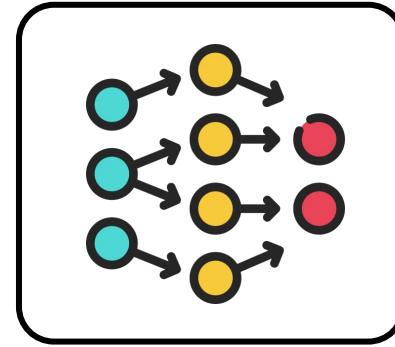
Haohua Du, He Lu, Fei Shang



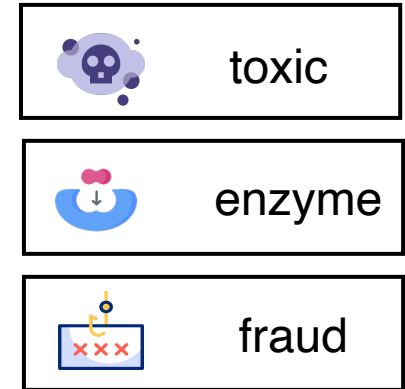
Background: GNN for Graph Classification



Graph



GNN

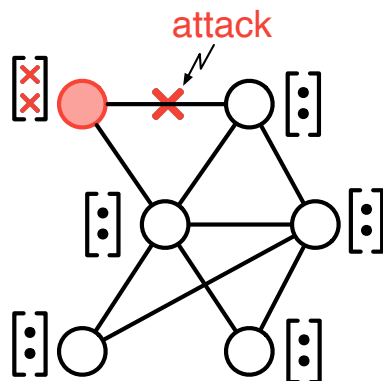
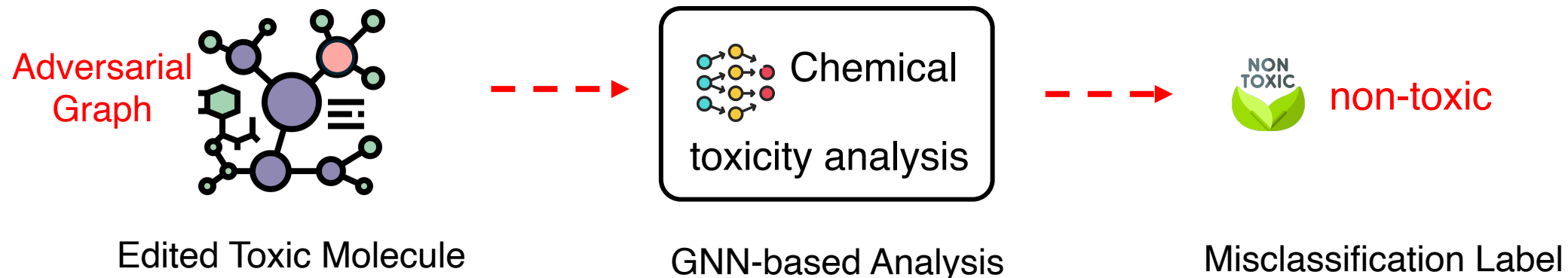


Label

Real-world relationships are often abstracted into **graphs**.

Graph neural networks (**GNN**) is used for **graph classification** in tasks like toxicity prediction, protein function prediction, and fraud detection.

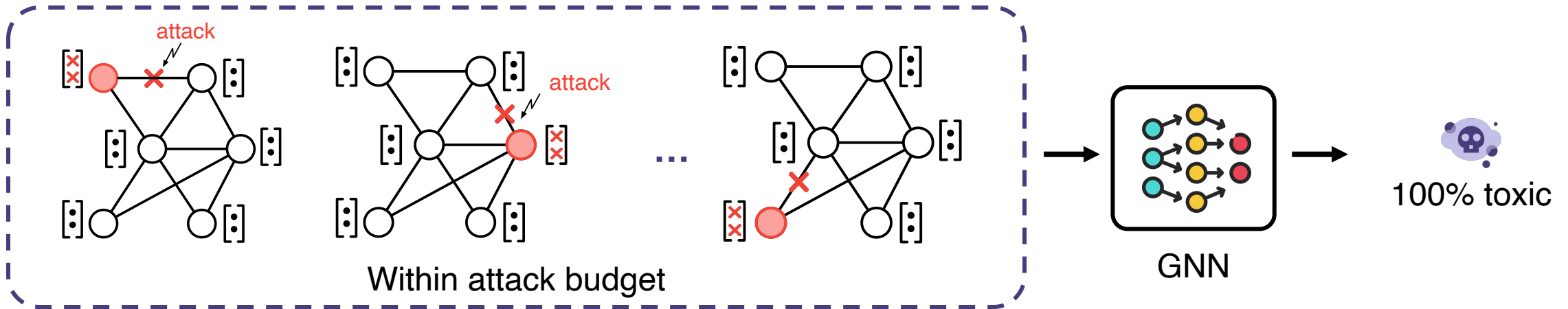
Background: Adversarial Attack on GNN



- **Adversarial Attack**

Unnoticeable perturbations on **graph structure** and **node attributes** could lead to misclassification of graph by high-accuracy GNN.

Background: Robustness Certificate



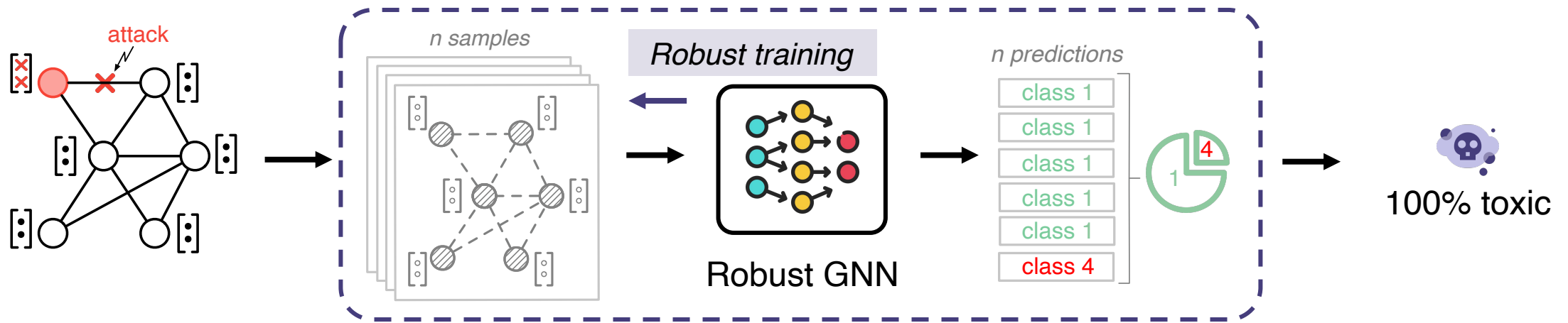
- **Certification**

Given the input graph $G = (X, A)$, base classifier f_θ and attack budget (magnitude of perturbation) Δ : guarantee that for all $\delta \in \Delta$, $f_\theta(G + \delta) = f_\theta(G)$.

- **Threat Model**

$\Delta = (\Delta_X, \Delta_A)$ specifies a ℓ_0 -ball around input G , which is to limit the magnitude of changes in X and A .

Background: Randomized Smoothing



- **Randomized Smoothing** (Cohen 2019, Bojchevski 2020)

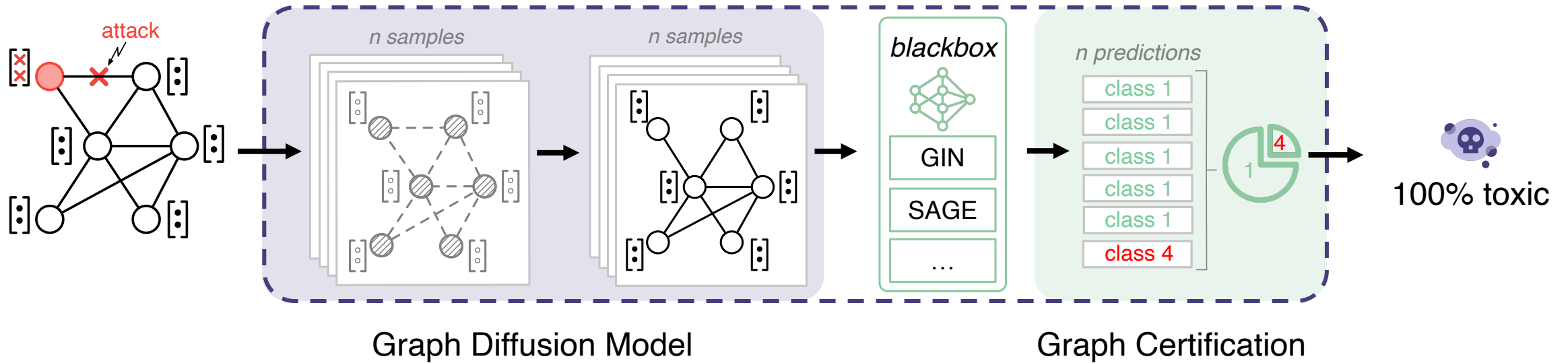
RS-based method predicts the label with a smoothed base classifier, noted as $g_\theta(G)$:

$$g_\theta(G) := \arg_y \max_{\tilde{G} \sim \phi(G)} \Pr [f_{\theta^*}(\tilde{G}) = y], \text{ where } \phi(G) \text{ is randomization scheme}$$

RS methods require retraining or fine-tuning as f cannot handle noisy samples

i) Retraining for diverse adversaries; ii) accuracy drop on clean samples.

How to provide plug-and-play certified defense for GNN?



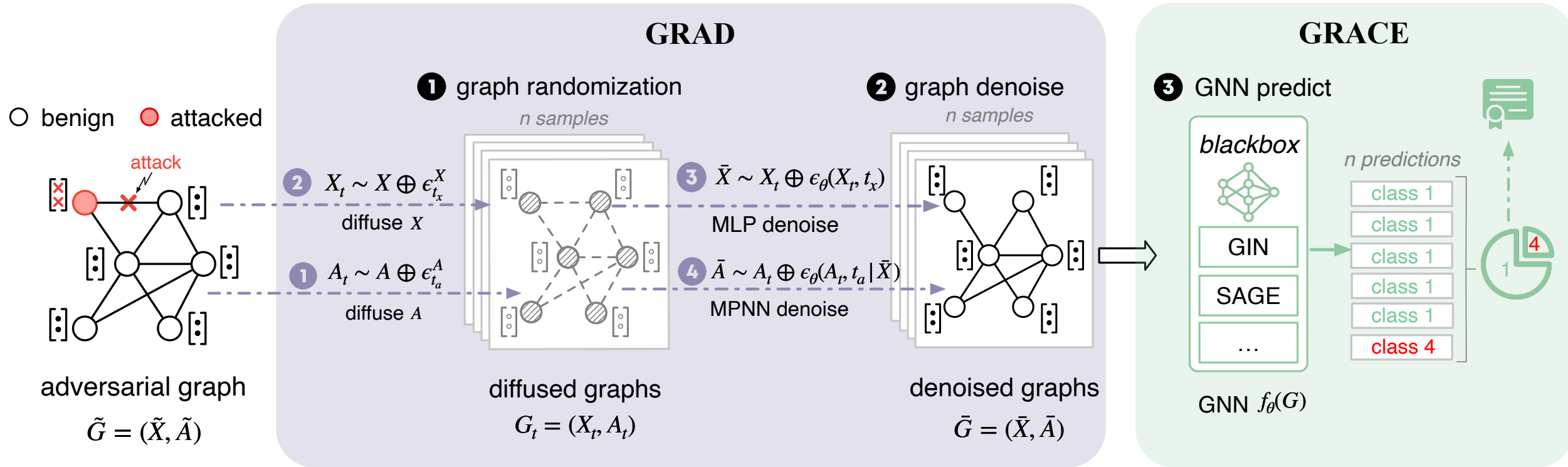
Diffusion Models

- **GRACED: Denoised Smoothing**

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

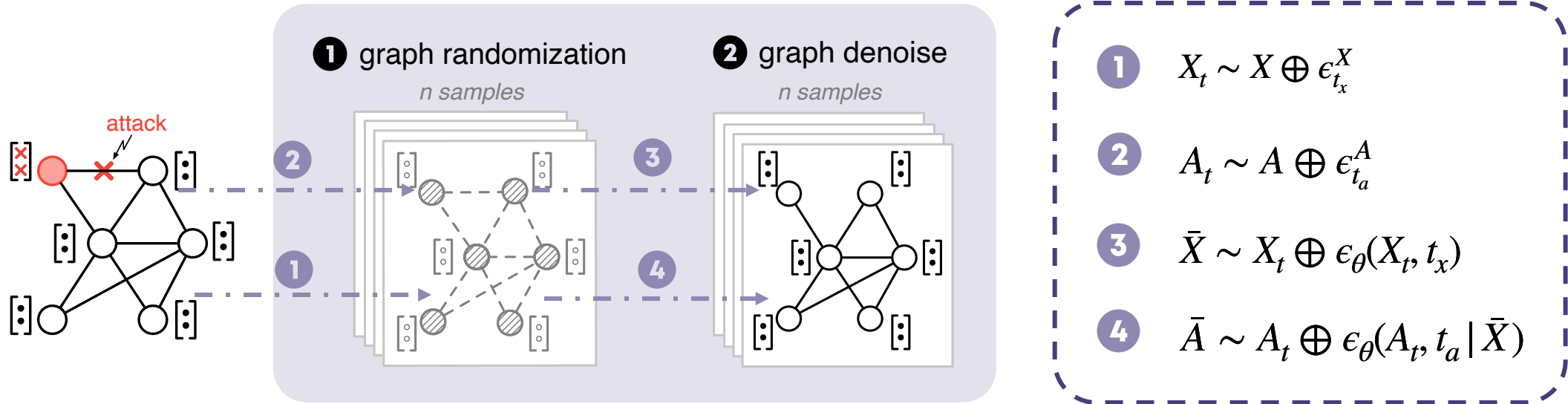
\mathcal{D} is a diffusion-based denoiser, which purify the adversarial graph before obtain certified classification with black-box GNN.

GRACED Framework



GRACED - a plug-and-play solution to guarantee **GRA**ph classification with **CE**rifiable robustness via a **D**iffusion model

GRAD: A Graph Diffusion Model



- **Step 1: graph randomization**

Forward process of diffusion model adds data-dependent Bernoulli noise on X and A .

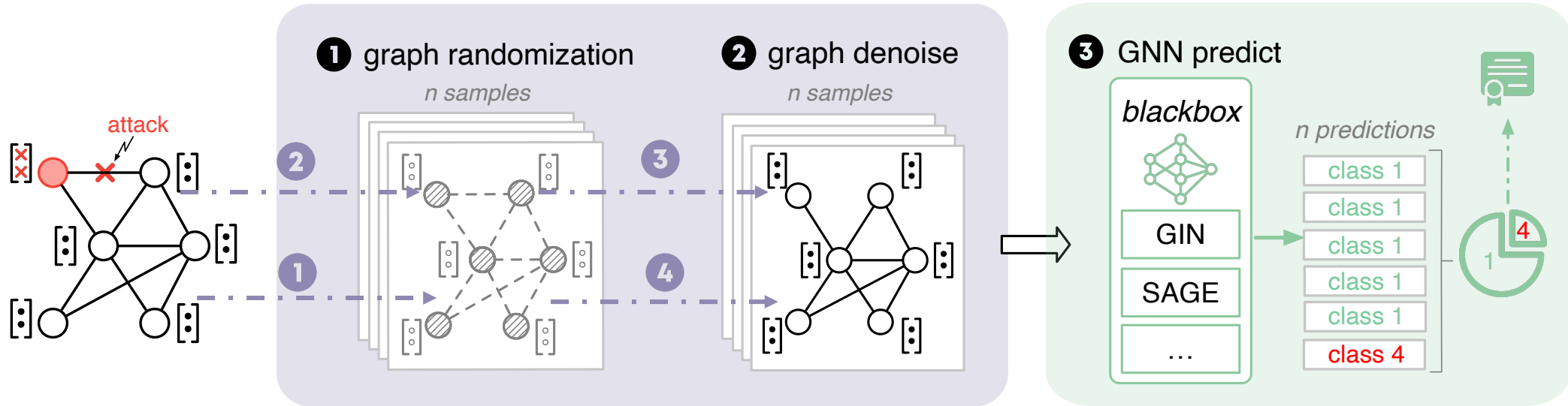
diffusion = randomization

- **Step 2: graph denoise**

Reverse process of diffusion model removes noise from X and A to reconstruct data.

reverse = purification

GRACE: A Graph Certification



- **Step 3: GNN predict**

Use black-box standard trained GNN to classify denoised graphs, providing robustness certificate with randomized smoothing method.

diffusion timestep = randomization parameter

Evaluation

TABLE I
CLEAN ACCURACY UNDER DIFFERENT PERTURBATION

Type		MUTAG	NCI1	PROTEINS	IMDB
Attr.& Adj.	Naïve ϕ	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 ϕ	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 ϕ	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 $\epsilon^Z = \text{Ber}(p = \frac{1}{2}(p^+ + p^-))$.

- **Clean accuracy**
- ✓ Outperforming sparse smoothing (Bojchevski *et al.* 2020) on joint and singular perturbation.
- ✓ Outperforming hierarchical smoothing (Scholten *et al.* 2024) on singular perturbation on X .
- ✓ Outperforming Bernoulli smoothing (Wang *et al.* 2021) on singular perturbation on A .

Evaluation

- **Clean accuracy**
- ✓ Outperforming sparse smoothing (Bojchevski *et al.* 2020) under different randomization setting on bioinformatics graph dataset MUTAG and social network dataset IMDB-BINARY.

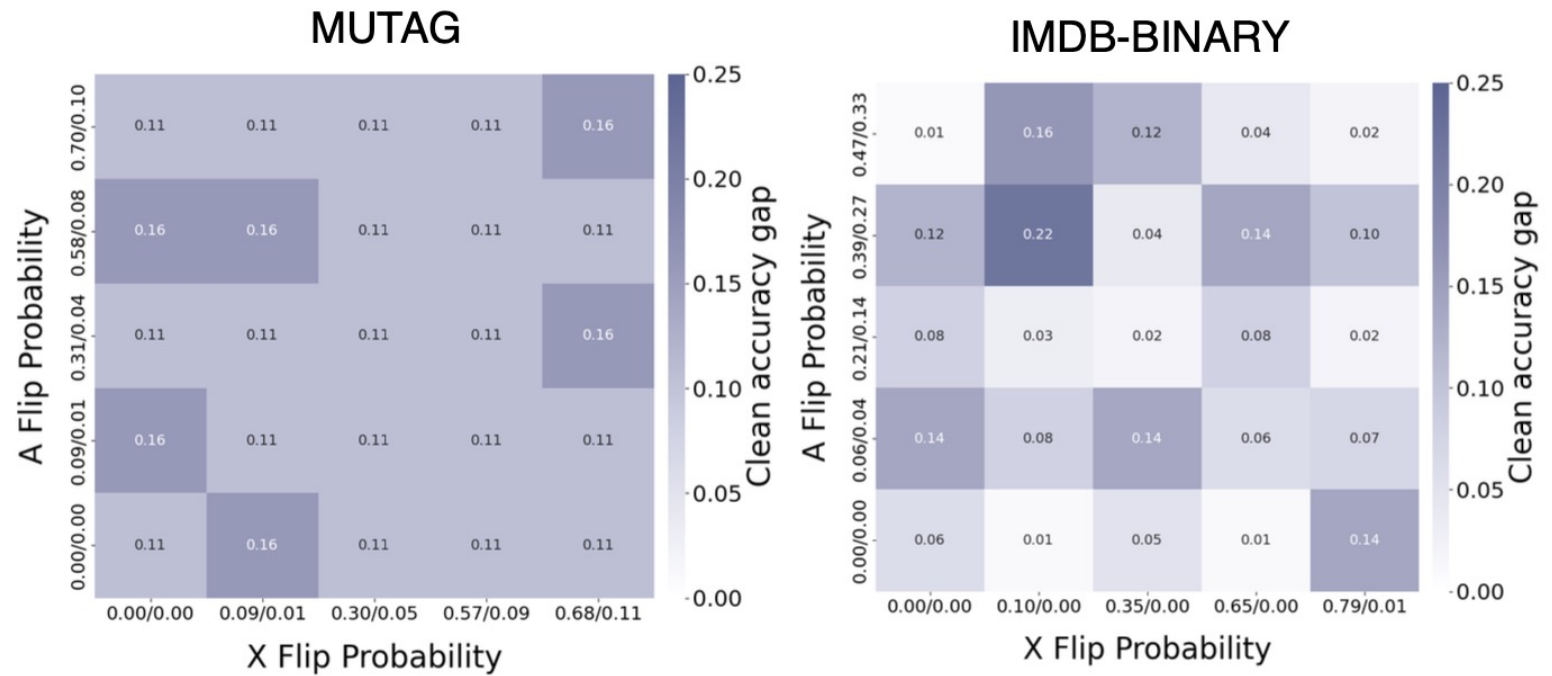


Fig. 3. **Clean accuracy gap:** Each heatmap shows the clean accuracy gap between GRACED and Sparse Smoothing per dataset, with noise scales for attributes and adjacent matrices on the horizontal and vertical axes.

Evaluation

- **Certified accuracy**

Achieving high certified accuracy on large attack budget, for both singular and joint perturbation scenario.

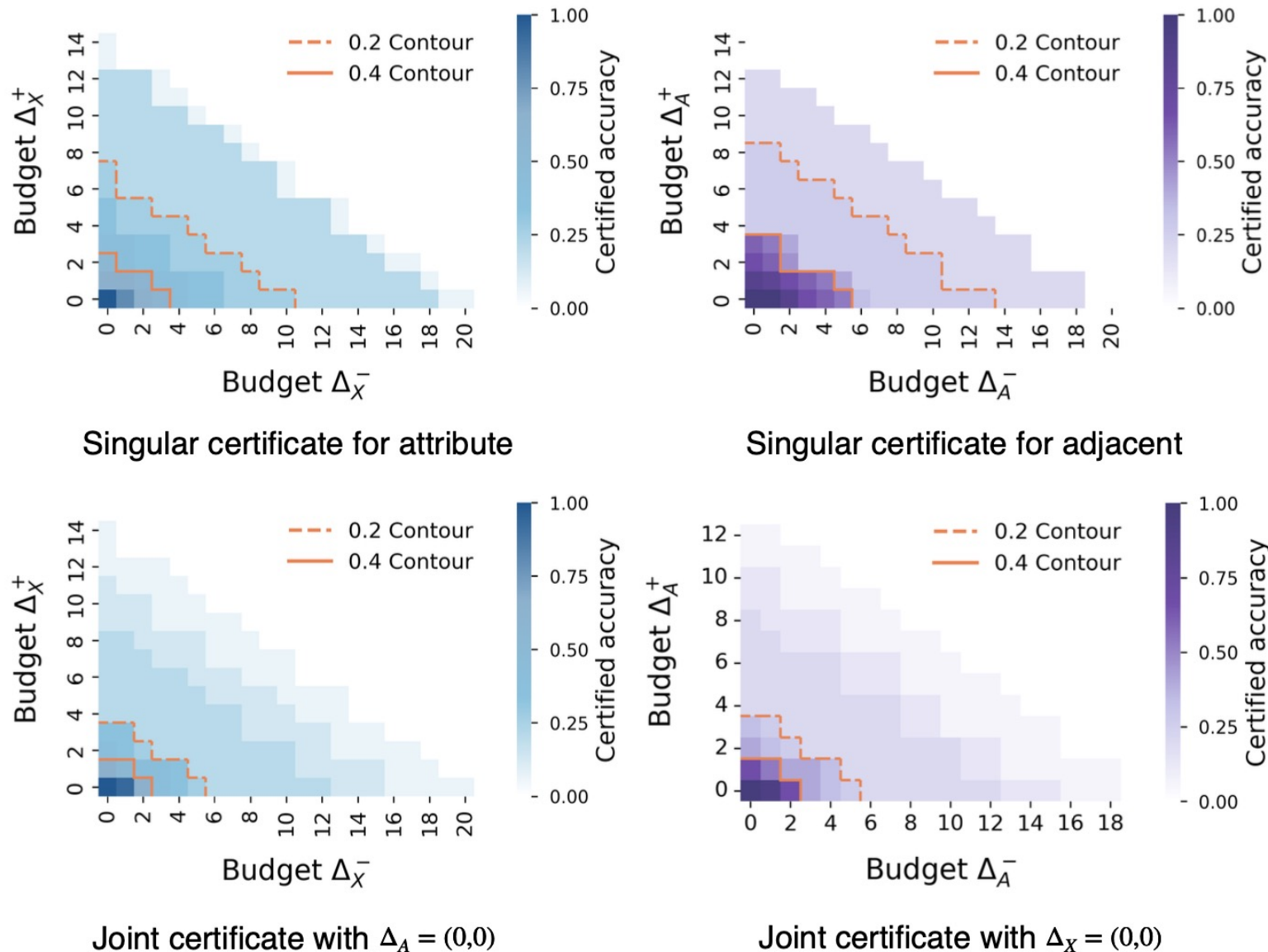


Fig. 4. **Certificate:** The top row depicts singular certificates, and the bottom shows joint perturbation defense. Blue and purple heatmaps represent certificates for node attributes and sctructure, respectively.

Summary

- **Plug-and-play Certification**

We present GRACED to effectively tackling the verifiable robustness of black-box graph classification models in a plug-and-play style.

- **Graph Diffusion Model**

We design GRAD, a graph diffusion based on D3PM (Austin *et al.* 2021) to purify the adversarial graph into a benign graph, preserving the structure stability of the graph in the process.

- **High Accuracy**

We have validated the efficacy of our approach through comprehensive testing on real-world datasets, showing accuracy improvement of approximate 10% over randomized smoothing.



**2025 IEEE INTERNATIONAL CONFERENCE ON
ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP 2025)**

April 06 - 11, 2025 **Hyderabad, India**



Thank you for your attention.

Xiaoyu Liang (BUAA)