Trustworthy Graph Neural Networks

Aleksandar Bojchevski

ICERM – Applied and Computational Discrete Algorithms in AI and Data Science

Based on joint work with S. Zargarbashi, S. Akhondzadeh, S. Antonelli.

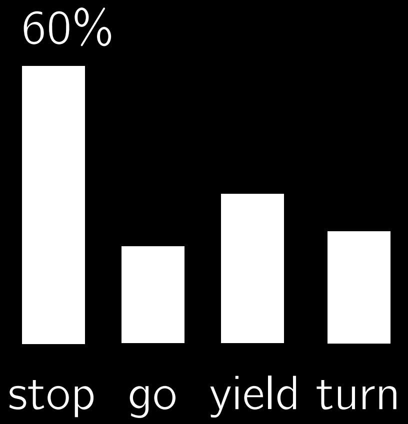
How reliable is the black box?

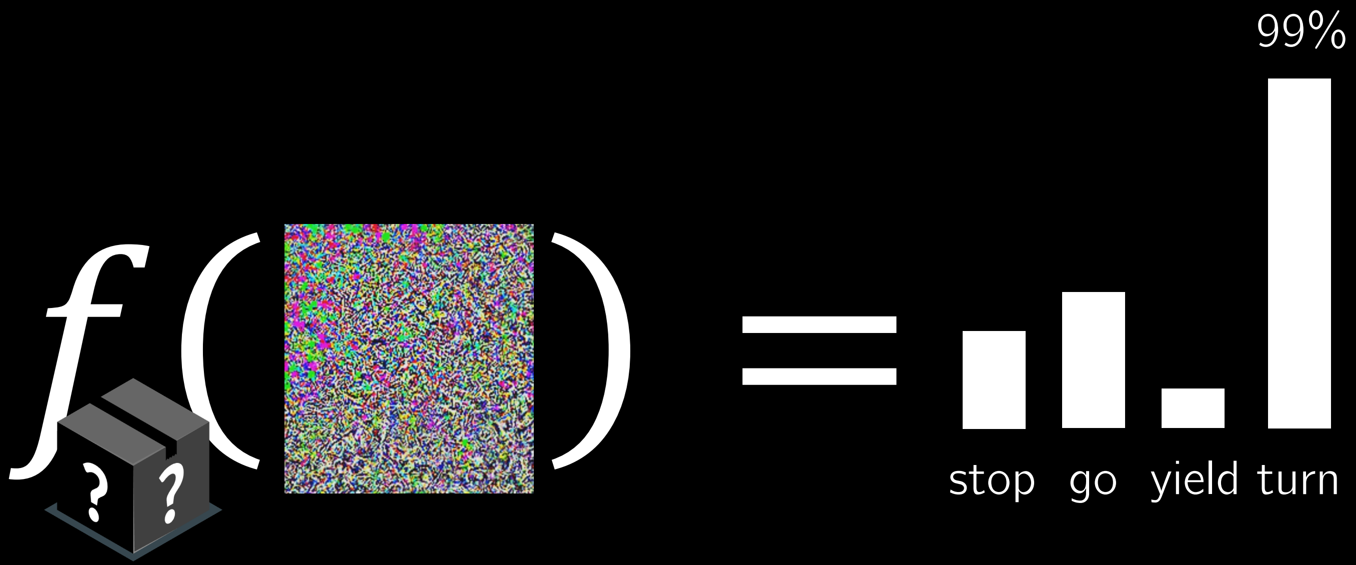
$$f(\quad)$$

$f($  $)$

$f($ 

$) = \text{stop}$

f  () $=$ 


f( ) =
99%
stop
go
yield
turn

Uncertainty quantification

Should we trust a prediction with 60% confidence?

How about 99%?

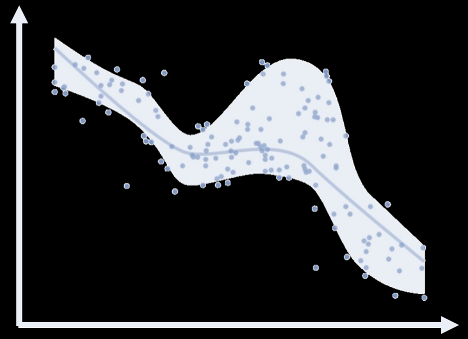
Is it calibrated? $\mathbb{P}(\text{correct} \mid \text{confidence} = \alpha) = \alpha, \forall \alpha$

Most models are not!

Conformal prediction

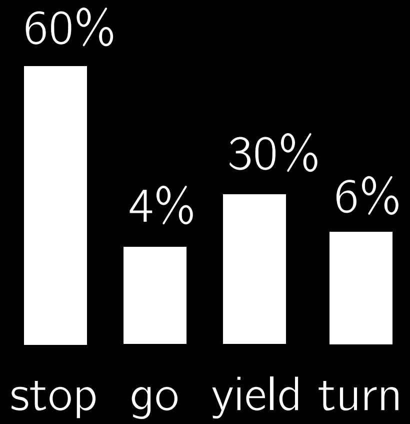
Output a set (or interval), larger set ⇒ more uncertain

$\mathcal{C}($  $) = \{1, 7\}$ $\mathcal{C}($  $) = \{8\}$



Distribution-free guarantee that $\mathbb{P}\big(y_{\text{true}} \in \mathcal{C}(\boldsymbol{x})\big) \geq 1 - \alpha$

Constructing the sets

$f($  $) =$ 

$$\mathbb{P}(y_{\text{true}} \in \{\text{stop}, \text{yield}\}) \ngeq 0.9$$

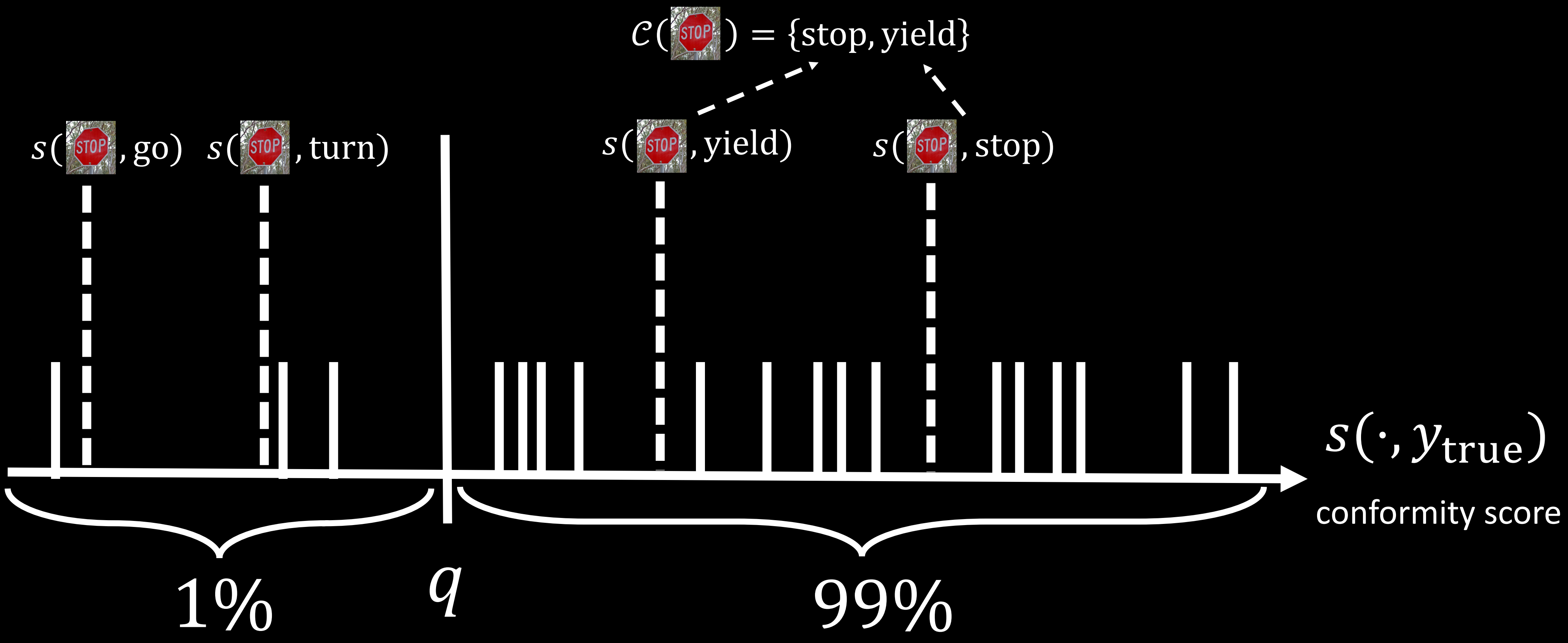
Algorithm

1. Define a conformity score function $s(\boldsymbol{x}, y)$

2. Compute a quantile threshold q on calibration data

3. Output the set $\mathcal{C}(\boldsymbol{x}) = \{y : s(\boldsymbol{x}, y) \geq q\}$

Algorithm



$\mathcal{C}(\cdot) = \{\text{stop}, \text{yield}\}$
$s(\cdot, \text{go})$
$s(\cdot, \text{turn})$
$s(\cdot, \text{yield})$
$s(\cdot, \text{stop})$
$s(\cdot, y_{\text{true}})$
conformity score
1%
q
99%

Exchangability

Only assumption is $\mathbb{P}\left(Z_1, \ldots, Z_n, Z_{n+1}\right) = \mathbb{P}\left(Z_{\pi(1)}, \ldots, Z_{\pi(n)}, Z_{\pi(n+1)}\right)$ for $Z_i = (\boldsymbol{x}_i, y_i)$ and all permutations π, i.e. order does not matter

Weaker than the i.i.d. assumption (i.i.d. $\Rightarrow$ exchangability).

Each rank order equally likely

Exchangability on graphs

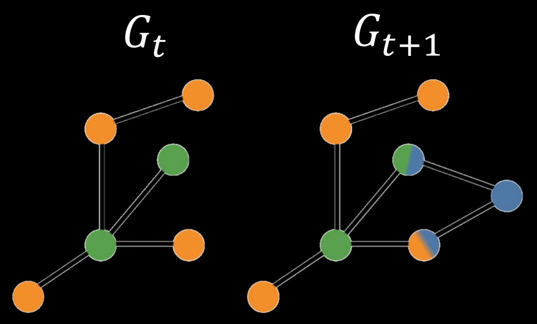
Transductive: given entire graph + exchangable calibration labels

$\Rightarrow$ Scores are exchangable for any permutation-equivariant model

Inductive: new nodes/edges arrive over time

1. Calibrate at time t
2. Construct sets at $t' > t$

$\Rightarrow$ Scores are not exchangable even if the data is



Zargarbashi et al. 2023. "Conformal Prediction Sets for Graph Neural Networks"

Conformal inductive GNNs

Fix 1: Time-dependent treshold for each t', recovers coverage guarantee for node exchangable graphs.

But, node exchangable $\iff$ graphons $\iff$ dense or empty

Fix 2: Time-dependent and node-frequency weighted treshold, recovers coverage guarantee for edge exchangable graphs

Zargarbashi et al. 2024." Conformal Inductive Graph Neural Networks"

Designing the score

Guarantee holds for any score $s(\boldsymbol{x}, y)$

Desiderata for $\mathcal{C}(\boldsymbol{x})$:

- efficient: size as small as possible
- adaptive: smaller for easy and larger for harder cases
- as many singletons as possible

DAPS: Diffused Adaptive Prediction Sets

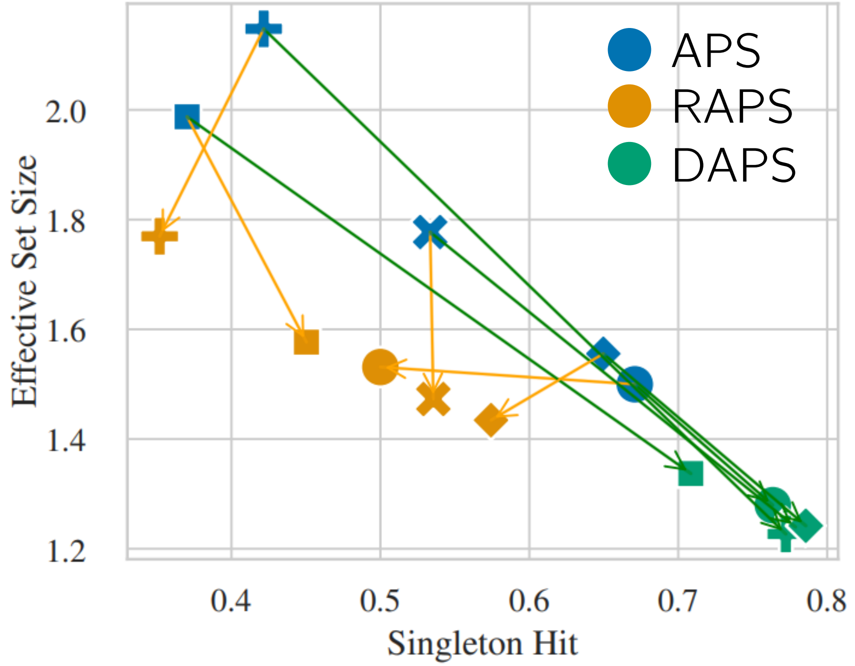
To improve the score of x use other scores from its neighborhood

$$\hat{s}(\boldsymbol{x}, y) = (1 - \lambda) \cdot s(\boldsymbol{x}, y) + \frac{\lambda}{|\mathcal{N}|} \sum_{\boldsymbol{x}' \in \mathcal{N}(\boldsymbol{x})} s(\boldsymbol{x}', y)$$

Theorem (inf.): Under homophily $\hat{s}$ leads to smaller sets

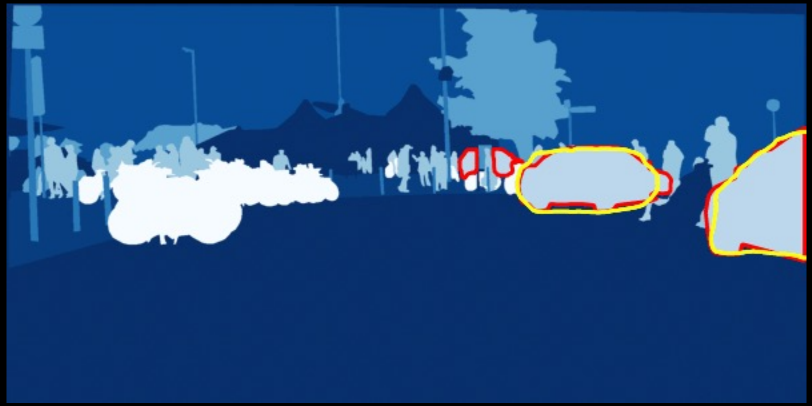
Zargarbashi et al. 2023. "Conformal Prediction Sets for Graph Neural Networks"

DAPS: Your neighbors improve you



Conformal risk control

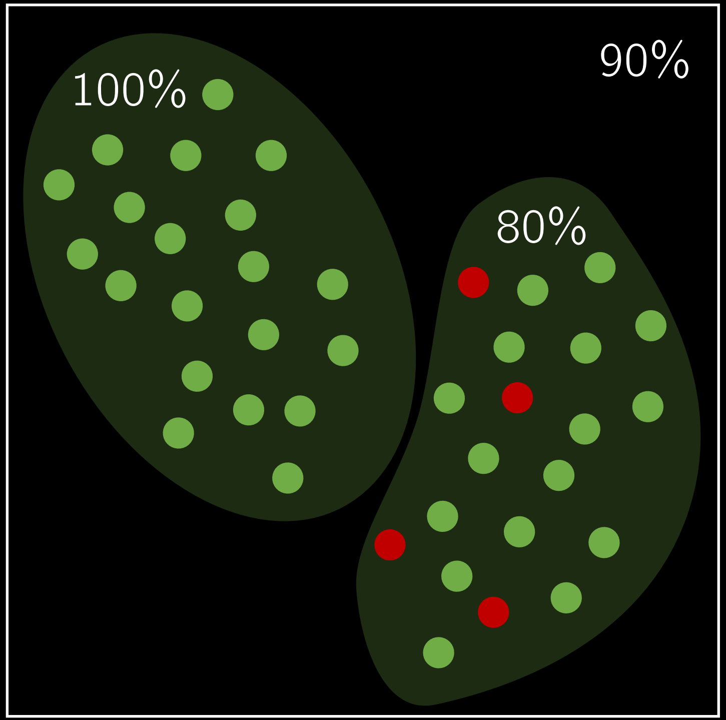
Guarantee $\mathbb{E}[\text{false negatives}] \leq 0.01$



Marginal vs. conditional

$\mathbb{P}\left(y_{\text{true}} \in \mathcal{C}(\boldsymbol{x})\right)$

$\mathbb{P}(\, y_{\text{true}} \in \mathcal{C}(\boldsymbol{x}) \mid \boldsymbol{x} = \boldsymbol{x} \,)$



Intermediate summary

You can construct sets that promise to cover the true label

f(

) $=$ stop

$f($ 

$) = \text{stop}$

$f($ 

$) = \text{go}$

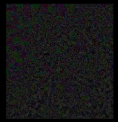
$f($ 

$) = \text{stop}$

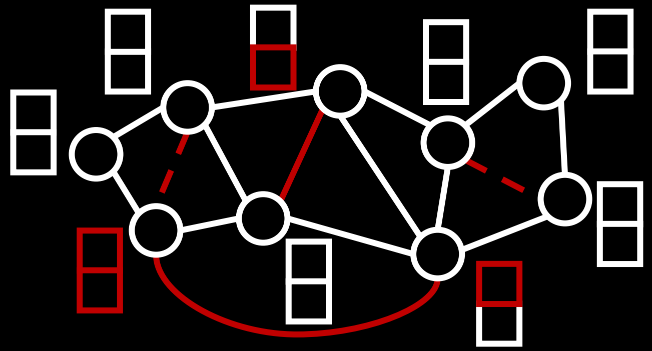
$f($ 

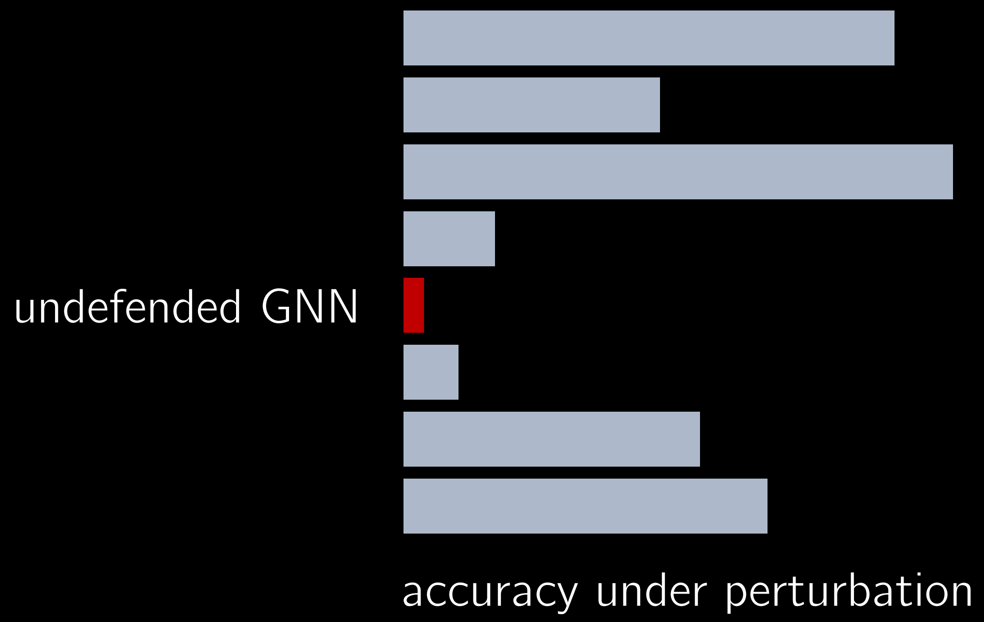
$+$  $) = \text{go}$

$f($  $) = \text{stop}$

$f($  +  $) = \text{go}$

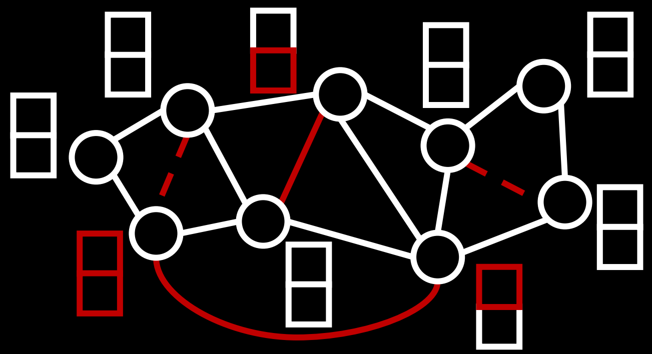
Are defenses robust?

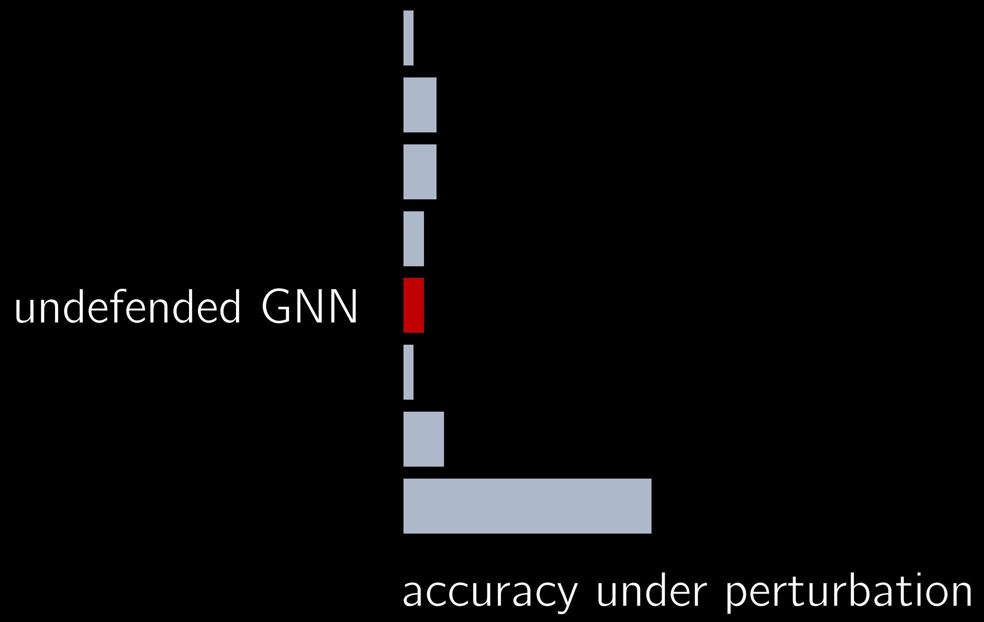




Mujkanovic et al. 2023 "Are Defenses for Graph Neural Networks Robust?"

Are defenses robust? No!





Mujkanovic et al. 2023 "Are Defenses for Graph Neural Networks Robust?"

Bitter lesson: Defenses are easily broken

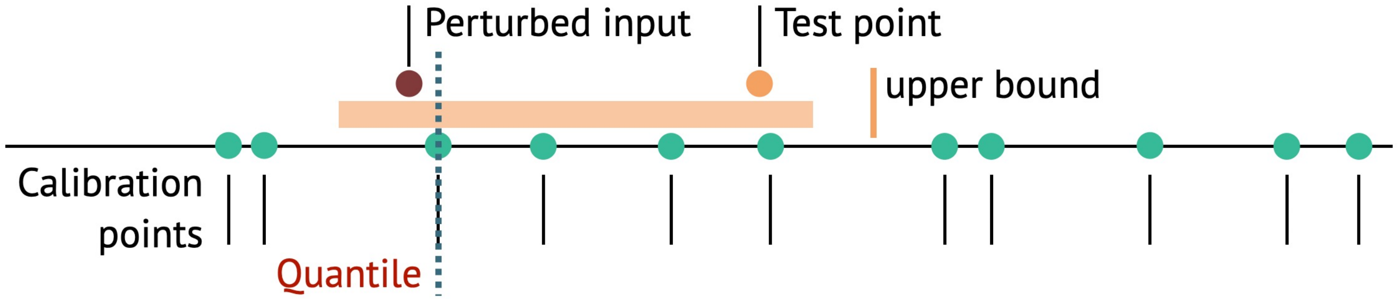
Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods by Carlini and Wagner broke 10 defenses.

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples by Athalye et al. broke 7 defenses.

On Adaptive Attacks to Adversarial Example Defenses by Tramer et al. broke 13 defenses.

Are Defenses for Graph Neural Networks Robust? by Mujkanovic et al. broke 7 defenses.

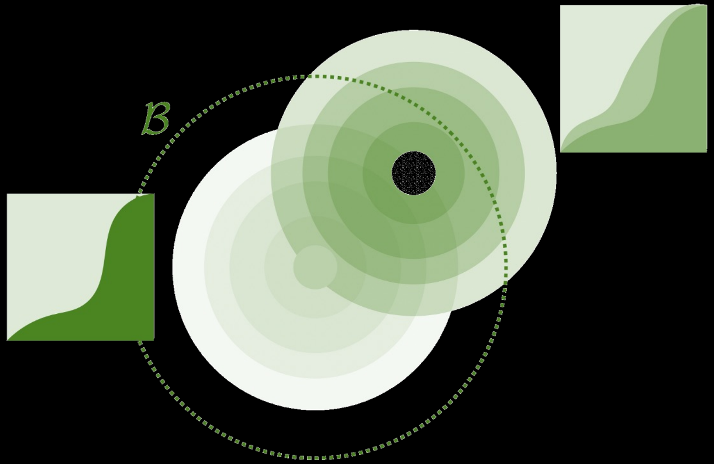
Conformal prediction is not robust



Randomized smoothing

1. Replace $s(\boldsymbol{x}, y)$ with $\mathbb{E}_{\boldsymbol{z}\sim p}[s(\boldsymbol{x}+\boldsymbol{z}, y)]$

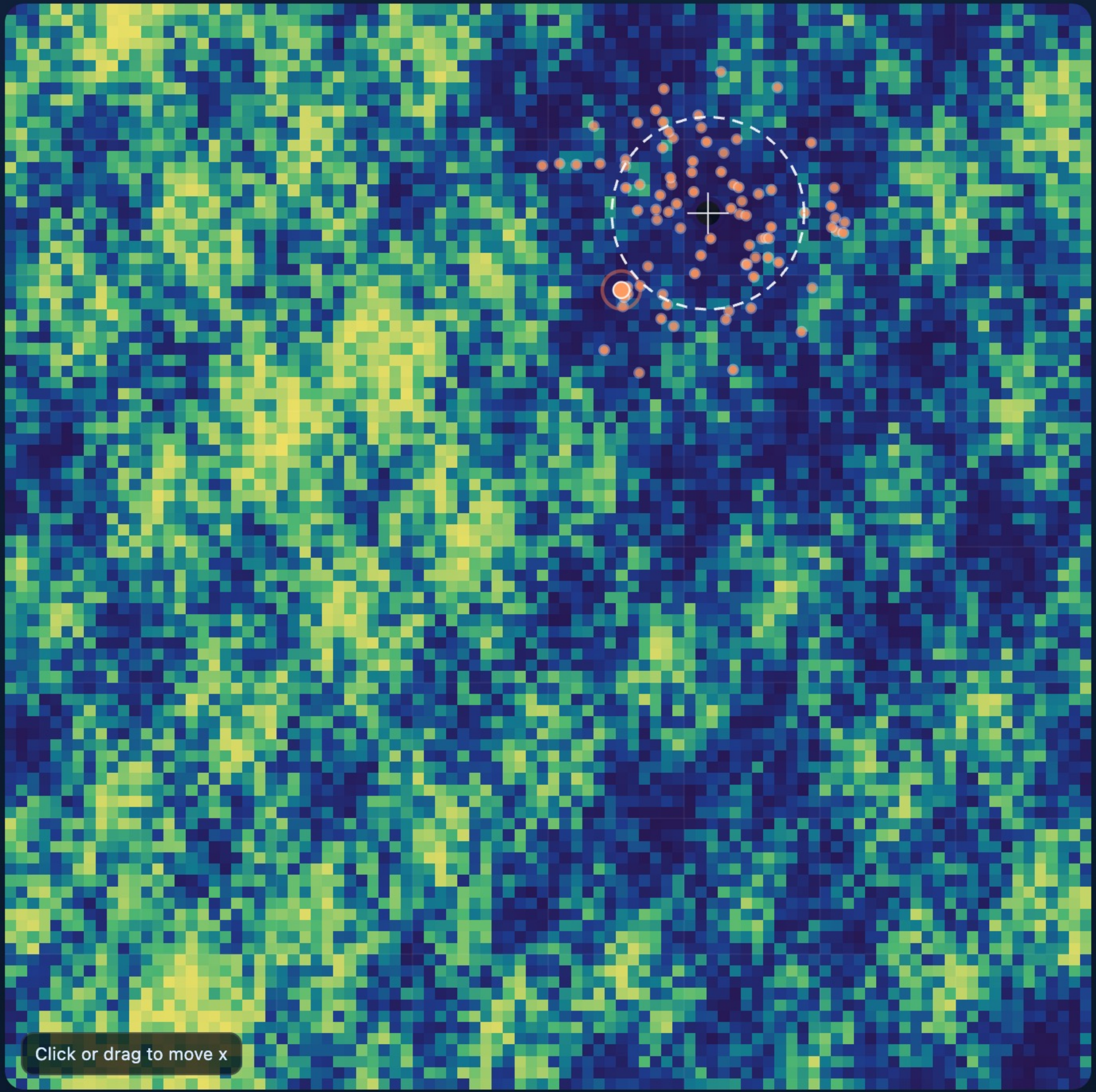
2. Bound $\max_{\boldsymbol{x}_{\text{adv}} \in \mathcal{B}(\boldsymbol{x})} \mathbb{E}_{\boldsymbol{z}\sim p}[s(\boldsymbol{x}_{\text{adv}}+\boldsymbol{z}, y)]$



RAW FIELD

Raw scores f(x)

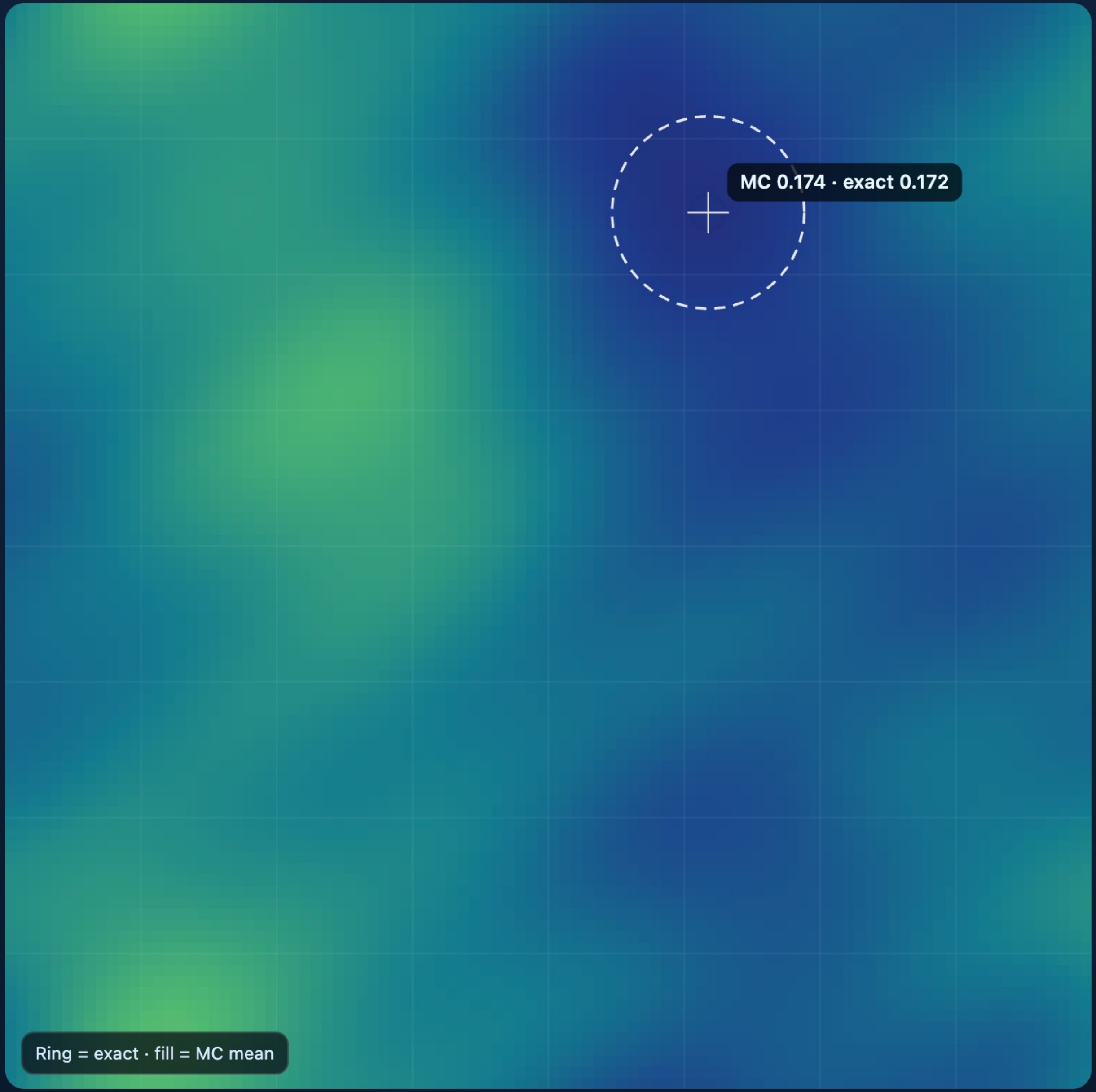
A jagged score surface. Coral dots show Gaussian perturbations around the selected point. The dashed circle marks the radius used for heuristic local bounds.



SMOOTHED FIELD

Smoothed scores $g_\sigma(x)$

The full field is computed by Gaussian convolution. The marker fill shows the running Monte Carlo mean; the ring shows the exact smoothed value.



Example bound for Gaussian noise

$$\text{adversarial score} \leq \Phi\left(\Phi^{-1}(\mathbb{E}[s(\boldsymbol{x}+\boldsymbol{z}, y)]) + \frac{r}{\sigma}\right)$$

need lots of MC samples to bound $\mathbb{E}[s(\boldsymbol{x}+\boldsymbol{z}, y)]$

Improving the sample efficiency

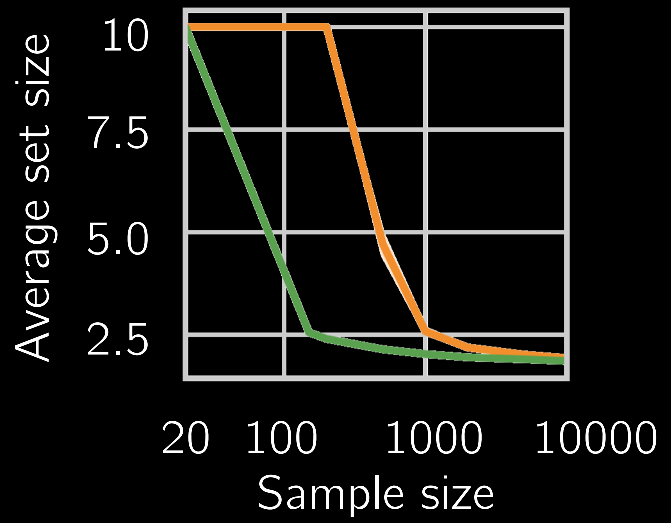
Method	Cal. samples	Test samples	Set size	Key idea
RSCP	$\sim 10^4$	$\sim 10^4$	trivial	Baseline

Improving the sample efficiency

Method	Cal. samples	Test samples	Set size	Key idea
RSCP	$\sim 10^4$	$\sim 10^4$	trivial	Baseline
CAS	$\sim 10^4$	$\sim 10^4$	$\downarrow$	Use CDF of scores

Improving the sample efficiency

Method	Cal. samples	Test samples	Set size	Key idea
RSCP	$\sim 10^4$	$\sim 10^4$	trivial	Baseline
CAS	$\sim 10^4$	$\sim 10^4$	↓	Use CDF of scores
BinCP	~ 150	~ 150	↓↓	Treshold the scores



Improving the sample efficiency

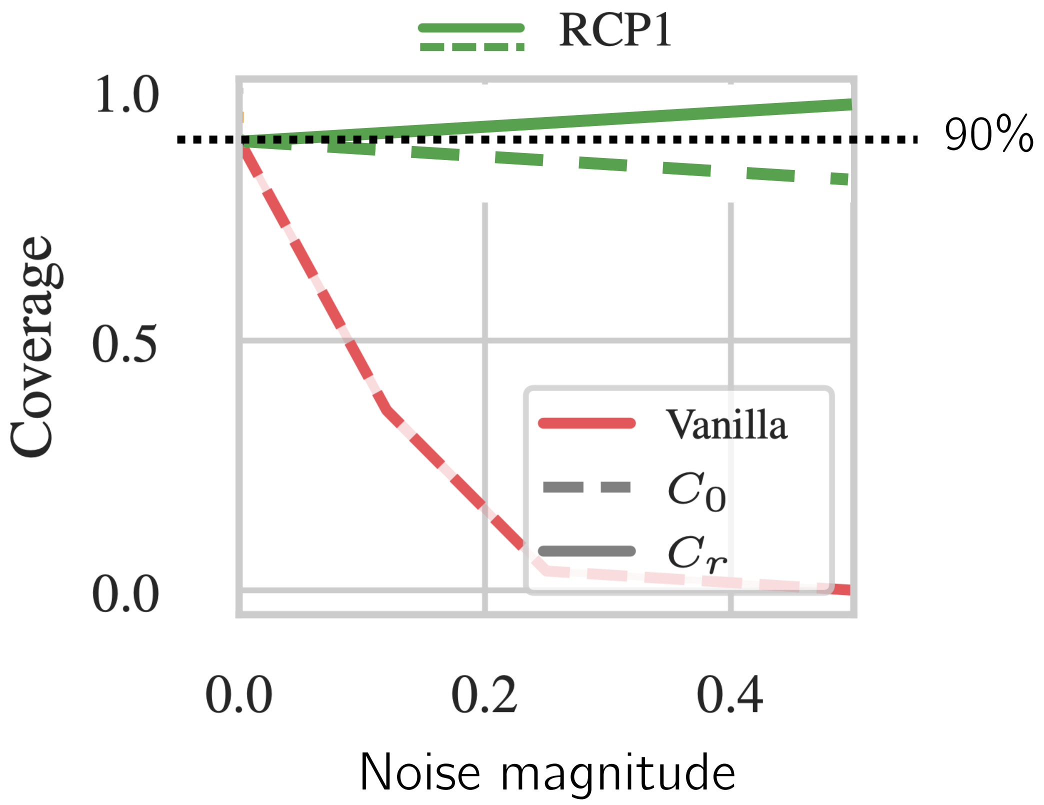
Method	Cal. samples	Test samples	Set size	Key idea
RSCP	$\sim 10^4$	$\sim 10^4$	trivial	Baseline
CAS	$\sim 10^4$	$\sim 10^4$	$\downarrow$	Use CDF of scores
BinCP	~ 150	~ 150	$\downarrow\downarrow$	Treshold the scores
RCP1	1	1	$\approx$ BinCP@~ 120	Convexity of bounds + Jensen

Improving the sample efficiency

Method	Cal. samples	Test samples	Set size	Key idea
RSCP	$\sim 10^4$	$\sim 10^4$	trivial	Baseline
CAS	$\sim 10^4$	$\sim 10^4$	↓	Use CDF of scores[1]
BinCP	~ 150	~ 150	↓↓	Treshold the scores[2]
RCP1	1	1	$\approx$ BinCP@$\sim$120	Convexity of bounds + Jensen[3]
FLoR	$10^3 - 10^4$	1	↓↓	Calibration happens once[4]

1. Zargarbashi et al. 2024. "Robust yet Efficient Conformal Prediction Sets"
2. Zargarbashi & Bojchevski 2025. "Robust Conformal Prediction with a Single Binary Certificate"
3. Zargarbashi et al. 2025. "One Sample is Enough to Make Conformal Prediction Robust"
4. Zargarbashi et al. 2026. "Front-Loaded Robust Conformal Prediction: Heavy Calibration, Minimal Test-Time Cost"

Randomness is inherently usefull



Summary

You can construct sets that promise to cover the true label

Randomness helps against adversaries