

EuCAIFCon 2026 - CONSERVAttack



[arXiv:2603.13970](https://arxiv.org/abs/2603.13970)

With funding from the:



Federal Ministry
of Research, Technology
and Space

UNIVERSITÄT **BONN**



Motivation

- Simulation of collision events in High Energy Physics
- **Validation** of said simulations is **crucial**

Motivation

- Simulation of collision events in High Energy Physics
- **Validation** of said simulations is **crucial**
- Currently:
 - For every observable / variable, compare **marginal distributions** to Standard Model
 - And compare the **linear correlations** to Standard Model

Motivation

- Simulation of collision events in High Energy Physics
- Validation of said simulations is crucial
- Currently:
 - For every observable / variable compute theoretical distributions to Standard Model
 - And compare the linear correlations to Standard Model



Classical Analyses

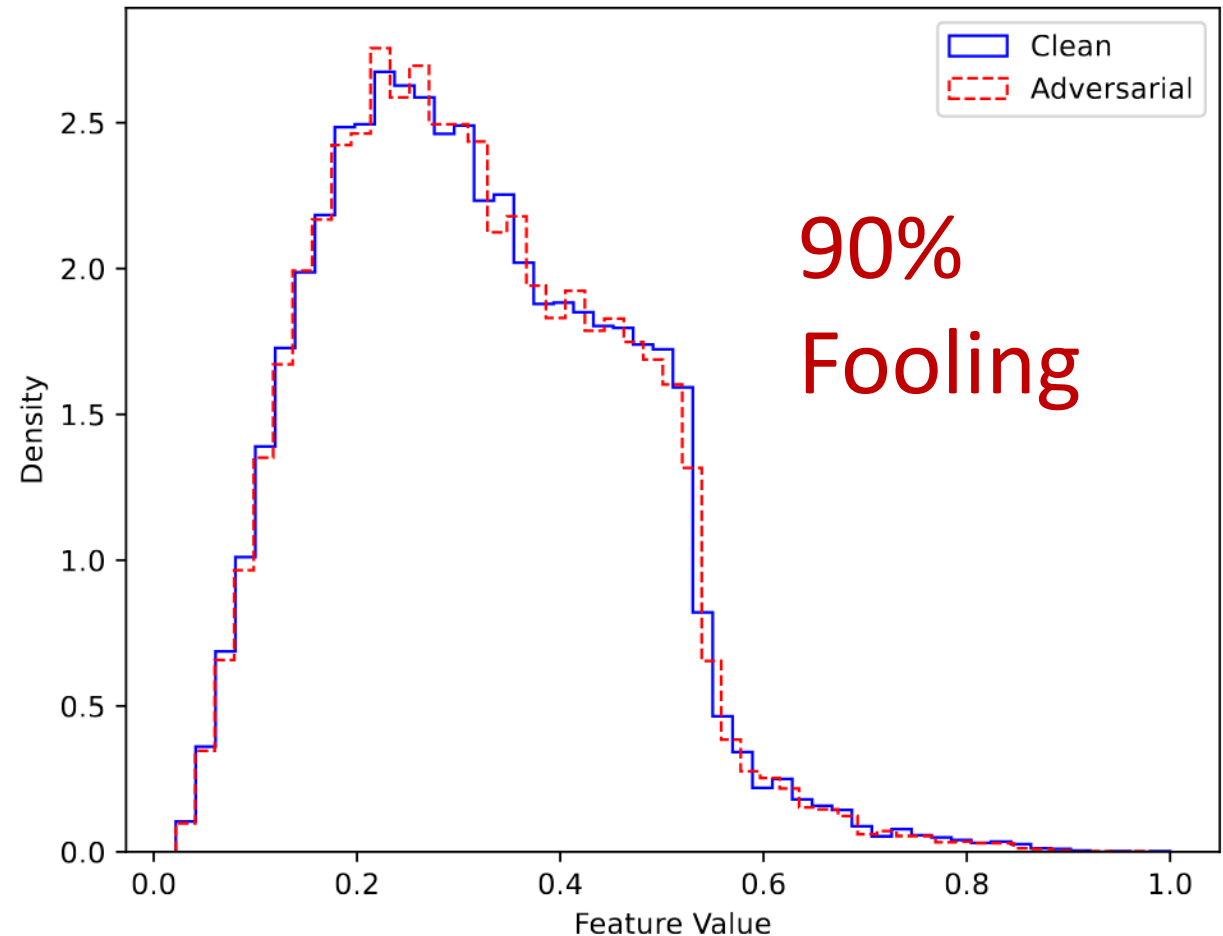
Motivation

- Simulation of collision events in High Energy Physics
- **Validation** of said simulations is **crucial**
- Currently:
 - For every observable / variable, compare **marginal distributions** to Standard Model
 - And compare the **linear correlations** to Standard Model

Deep Learning

Motivation

- „Invisible“ flaws in simulation
- ... Might be **visible** to **Deep Learning Models**



Motivation

- Our proposal:
- Construct **malicious events**
- Optimized to **exploit** these **weaknesses**

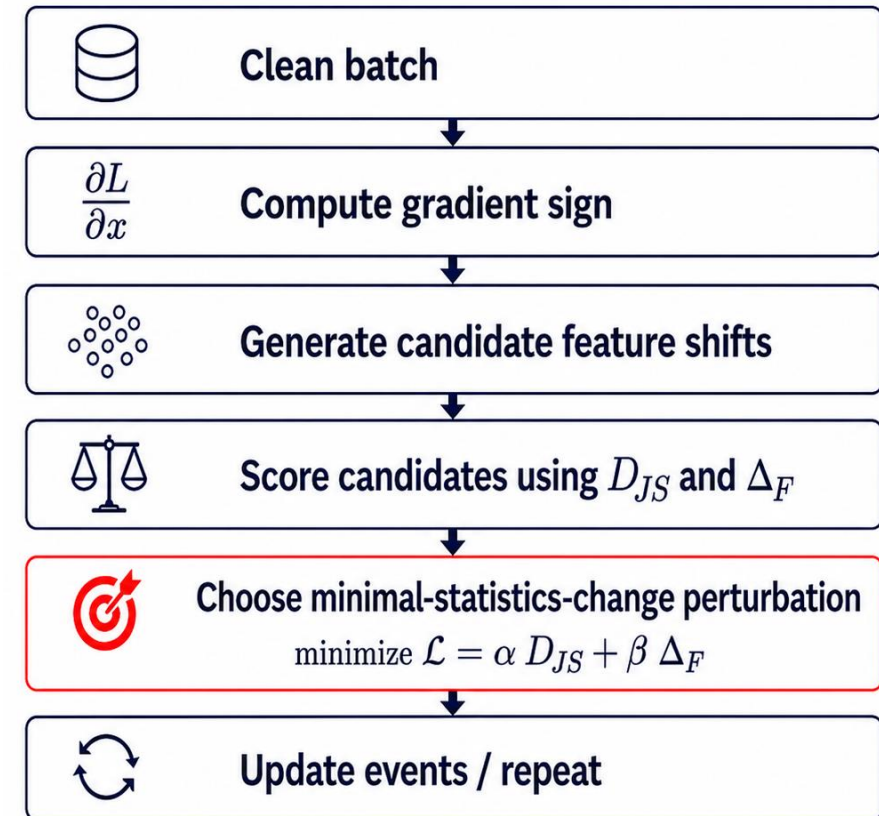
Motivation

- Our proposal:
- Construct **malicious events**
- Optimized to **exploit** these **weaknesses**

=> Study these events

Approach

- **Setup:**
 - Estimate clean **1D feature distribution**
 - Compute clean pairwise correlations
- **Candidate Evaluation:**
 - Generate feature shifts along gradient, bounded by min / max
 - Compare clean to adversarial distributions / correlations

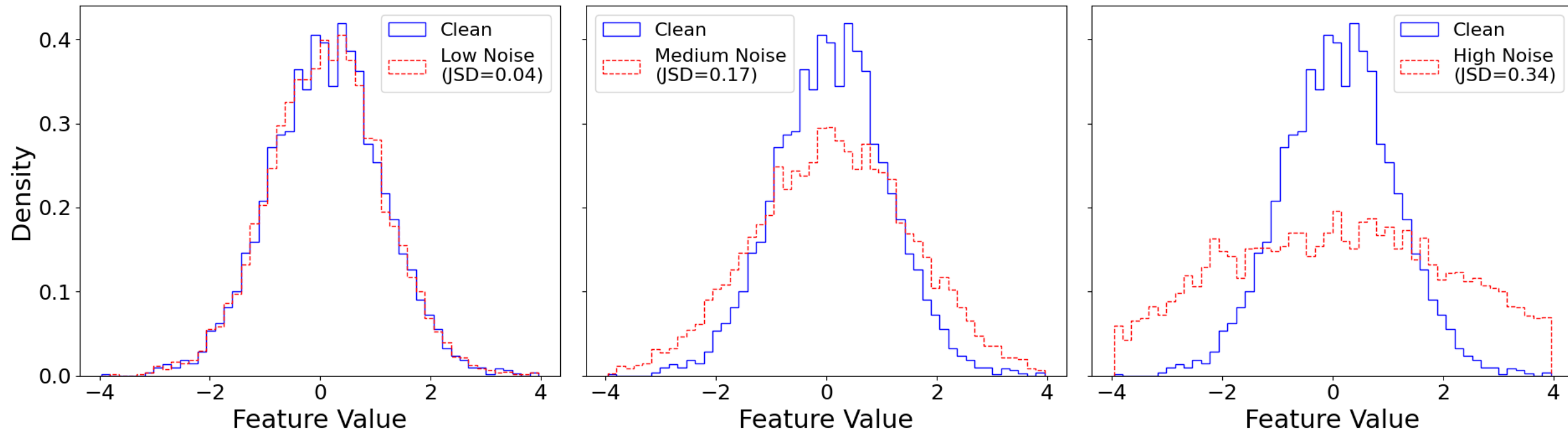


Attack Results

- Three metrics
- **Fooling Ratio:**
 - Ratio of initially correct events that are now misclassified by the model

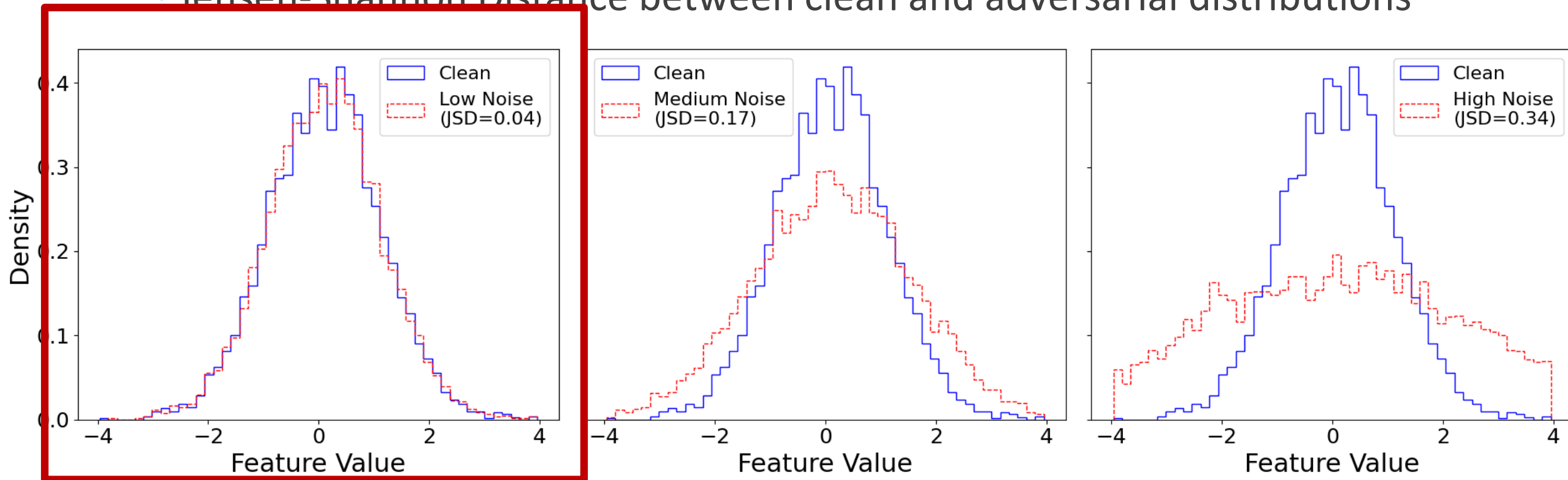
Attack Results – JSD

- **JSD:**
 - Jensen-Shannon Distance between clean and adversarial distributions



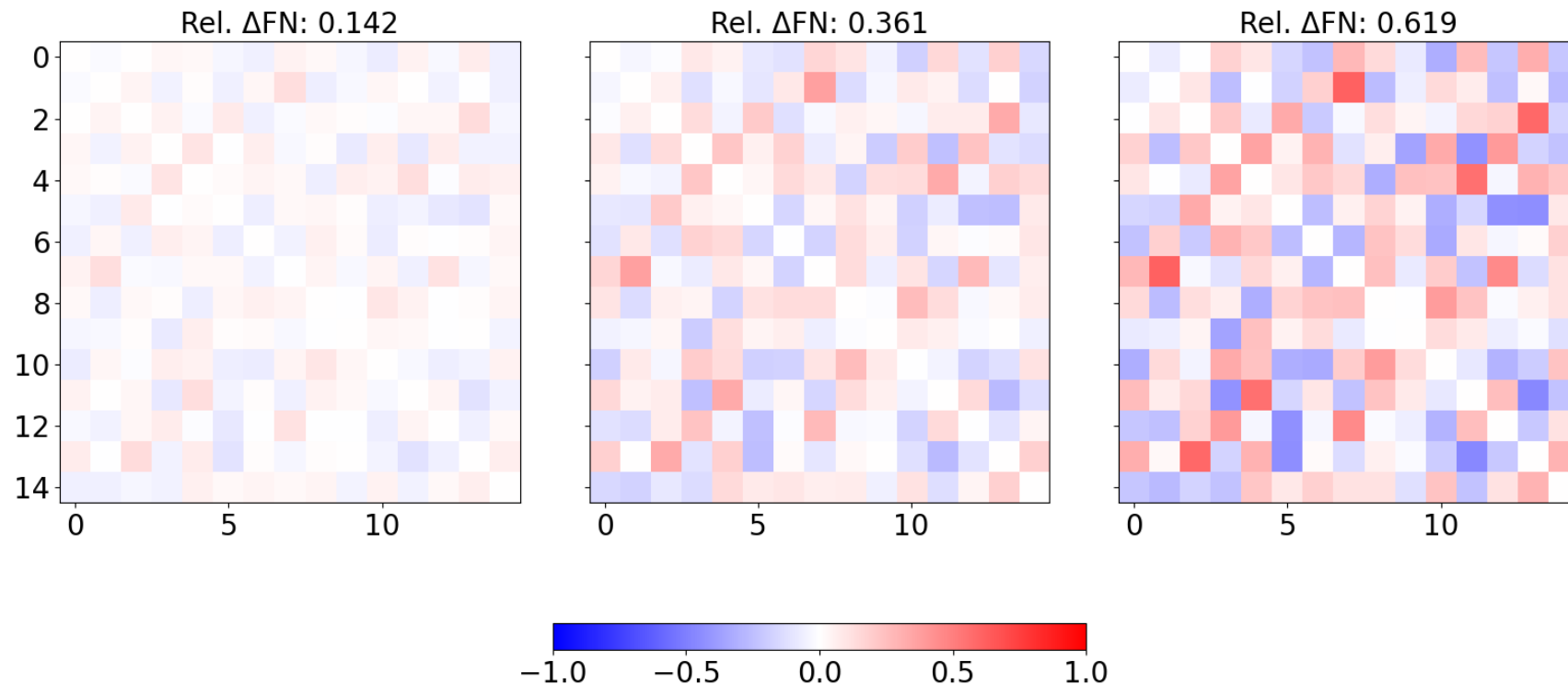
Attack Results – JSD

- **JSD:**
 - Jensen-Shannon Distance between clean and adversarial distributions



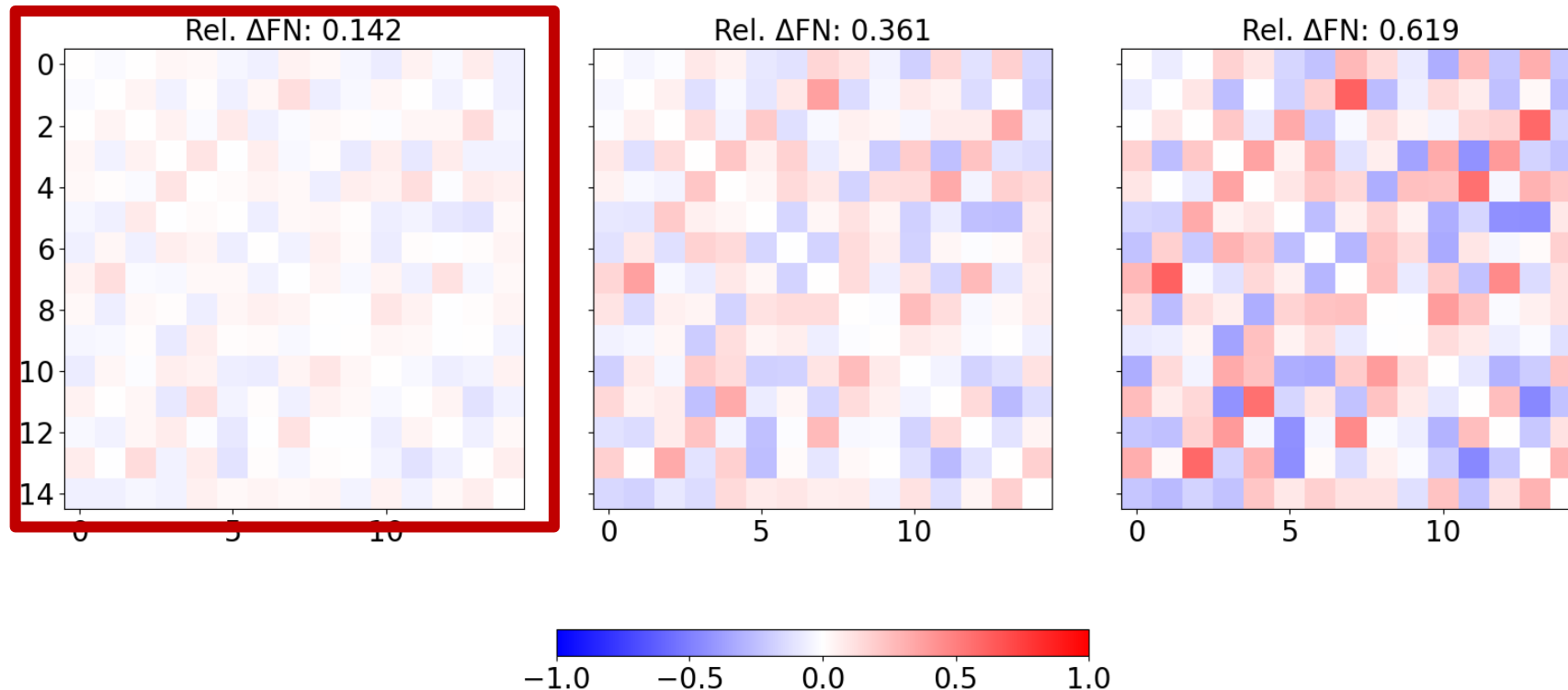
Attack Results – FN

- **FN:**
 - Relative Frobenius Norm of the difference of the Correlation Matrices

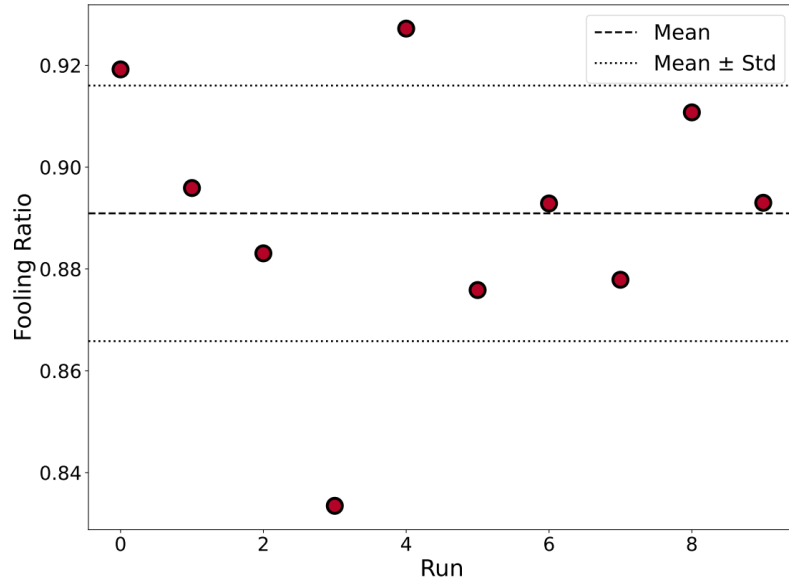


Attack Results – FN

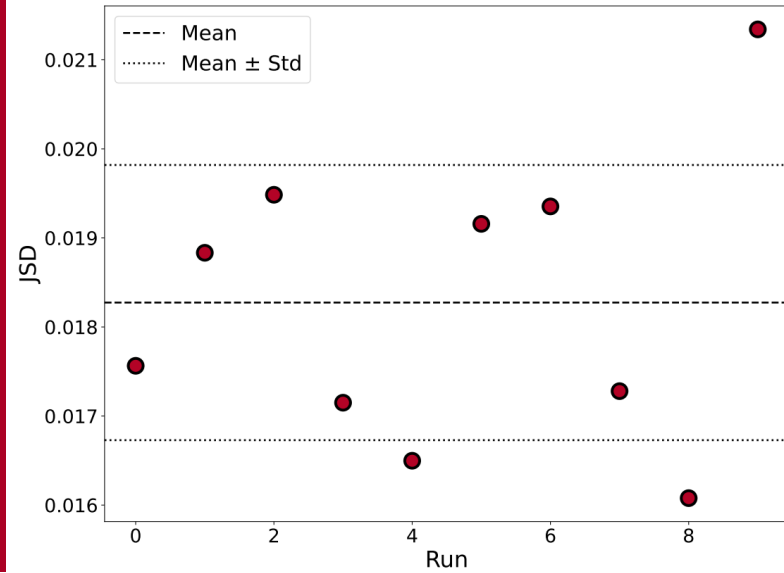
- **FN:**
 - Relative Frobenius Norm of the difference of the Correlation Matrices



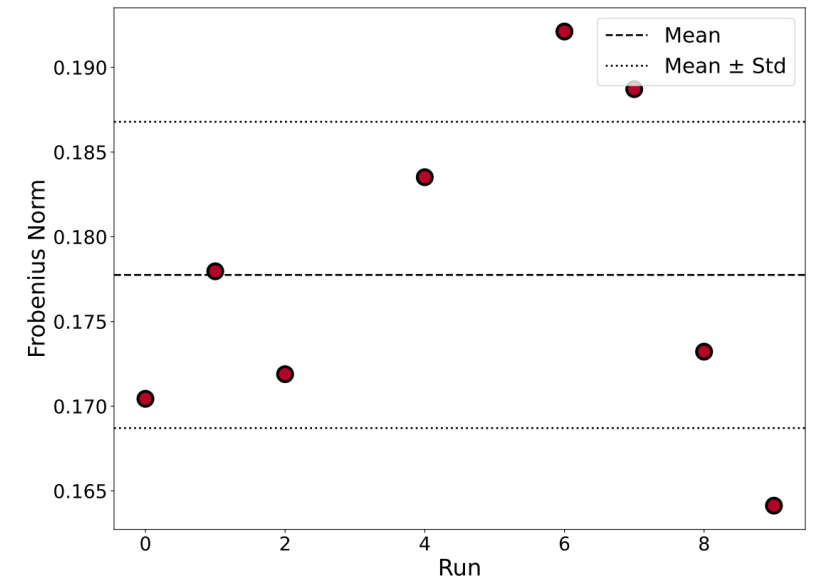
Fooling Ratio



JSD



FN



Attack Results

Defense

- Current fooling ratios are **very high**
- **Inaccurate** representation of **uncertainty upper-bound**
- Can we **improve** the networks **robustness** against such events?

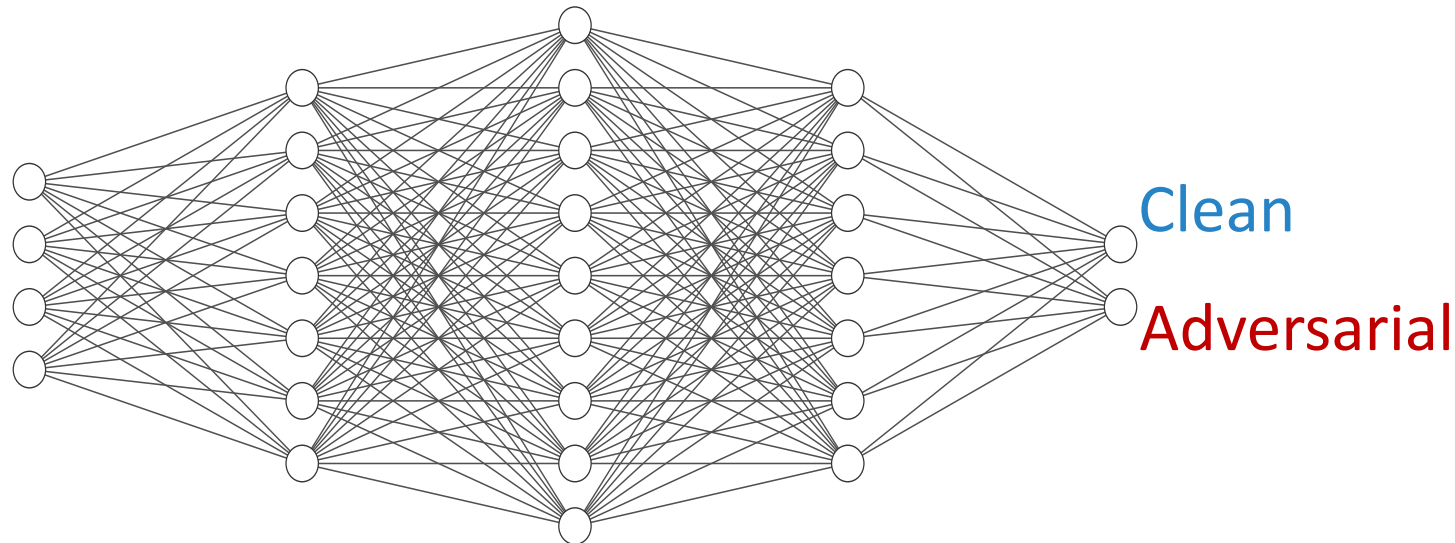
Defense

- Current fooling ratios are **very high**
- **Inaccurate** representation of **uncertainty upper-bound**
- Can we **improve** the networks **robustness** against such events?

=> Adversarial Detector

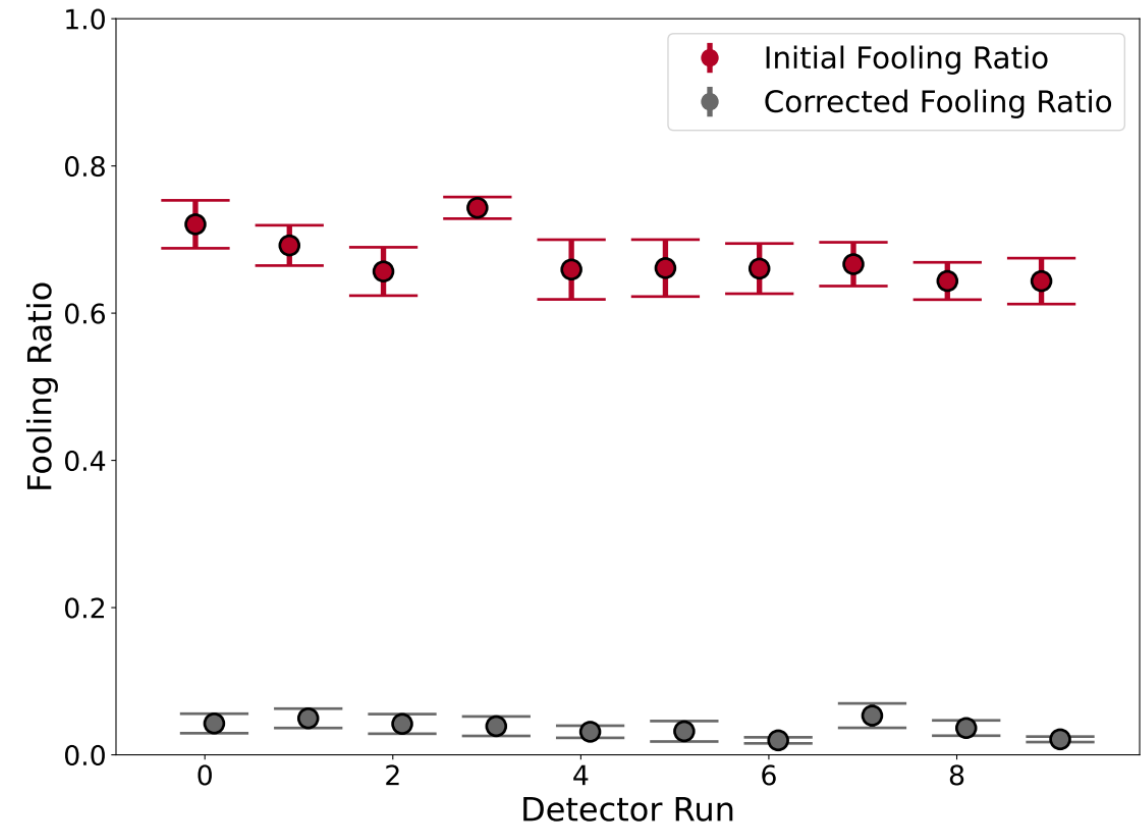
Defense – Adversarial Detector

- Generate adversaries on **train**, **test**, and **validation** sets
- Train a **network** on the following:
 - Classify between **clean** (label = 0) and **adversarial** (label = 1) events



Defense – Results

- **Initial Fooling Ratio:**
 - Fooling Ratio of malicious events on the initial model
- **Corrected Fooling Ratio:**
 - Fooling Ratio of malicious events that fool both the initial, and the detector model
 - = Initial Fooling Ratio * (1 – Detector Efficiency on Adversaries)

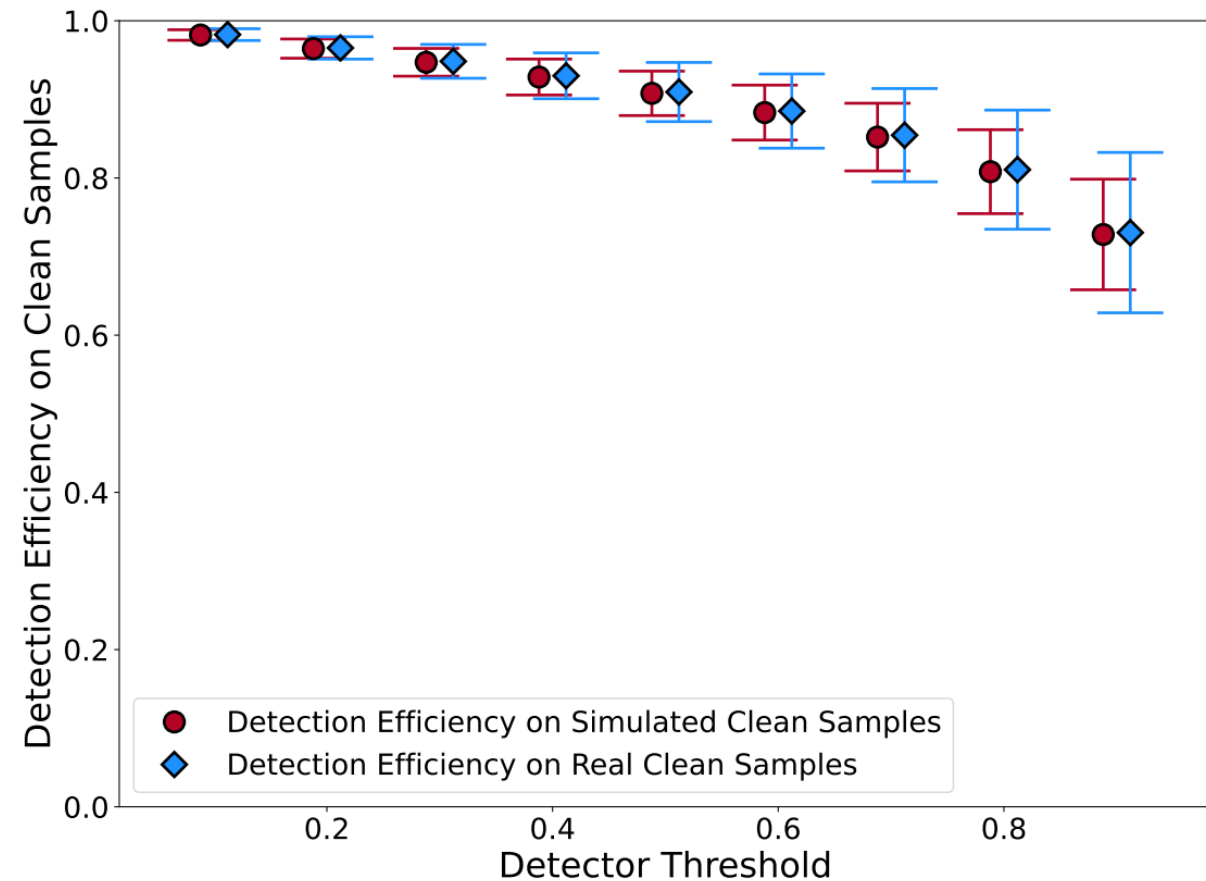


Defense – Results

- Apply Adversarial Detector to real events
- Check: misclassification real events
> misclassification simulated events ?

Defense – Results

- Apply Adversarial Detector to real events
- Check: misclassification real events > misclassification simulated events ?
- If yes: potential uncertainty from simulation



Conclusion

- We can construct **malicious events** that fool a model
 - While minimally altering the **marginal distributions**,
 - and the **linear correlations** between features
- Leverage **Adversarial Detector** to estimate **uncertainty bound**
- Deploy **Adversarial Detector** on **real events** to probe for **potential deviations**

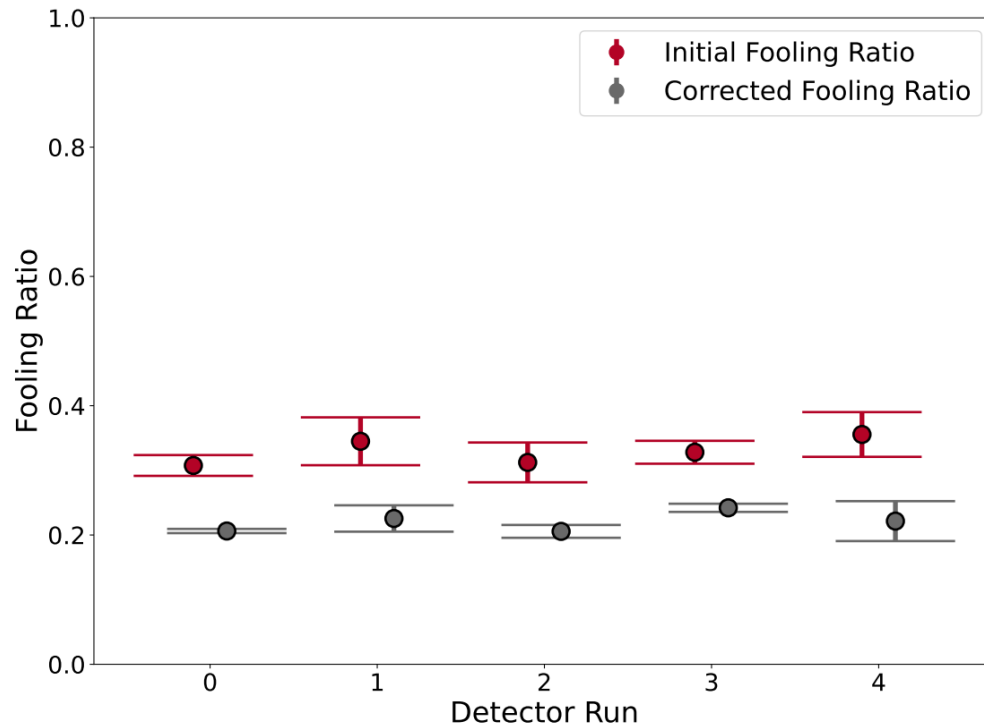
A thin, dark blue vertical line is positioned to the left of the text.

BONUS SLIDES

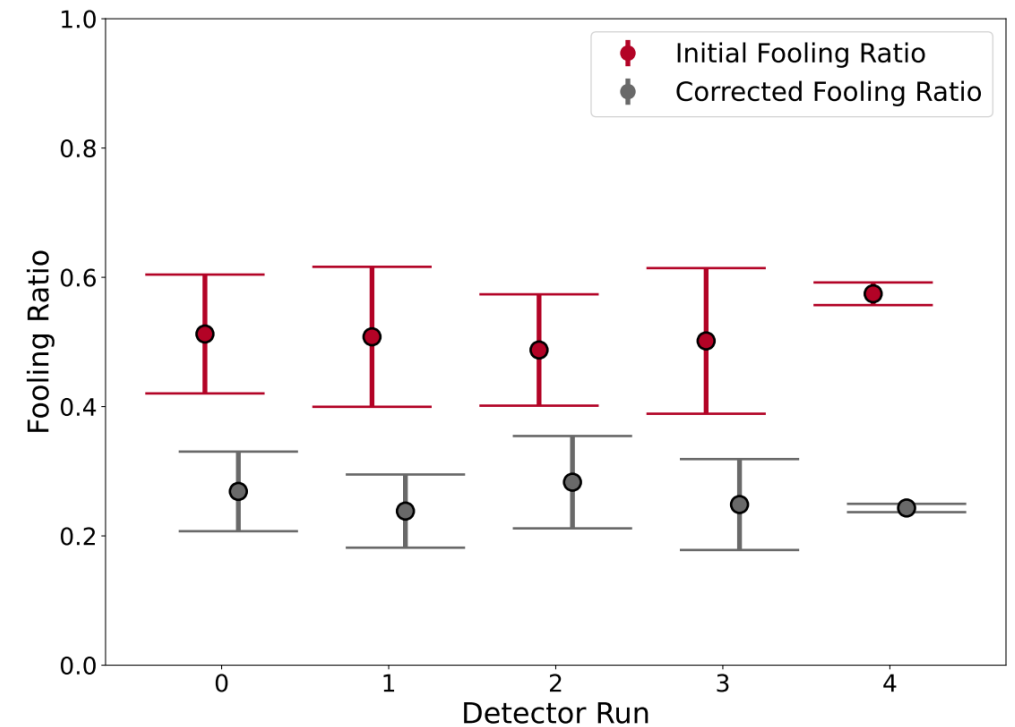
DisCo

- Use Distance Correlation (DisCo) instead of Pearson
- DisCo:
 - Based on feature-pairwise distances
 - => captures also non-linear relationships

DisCo

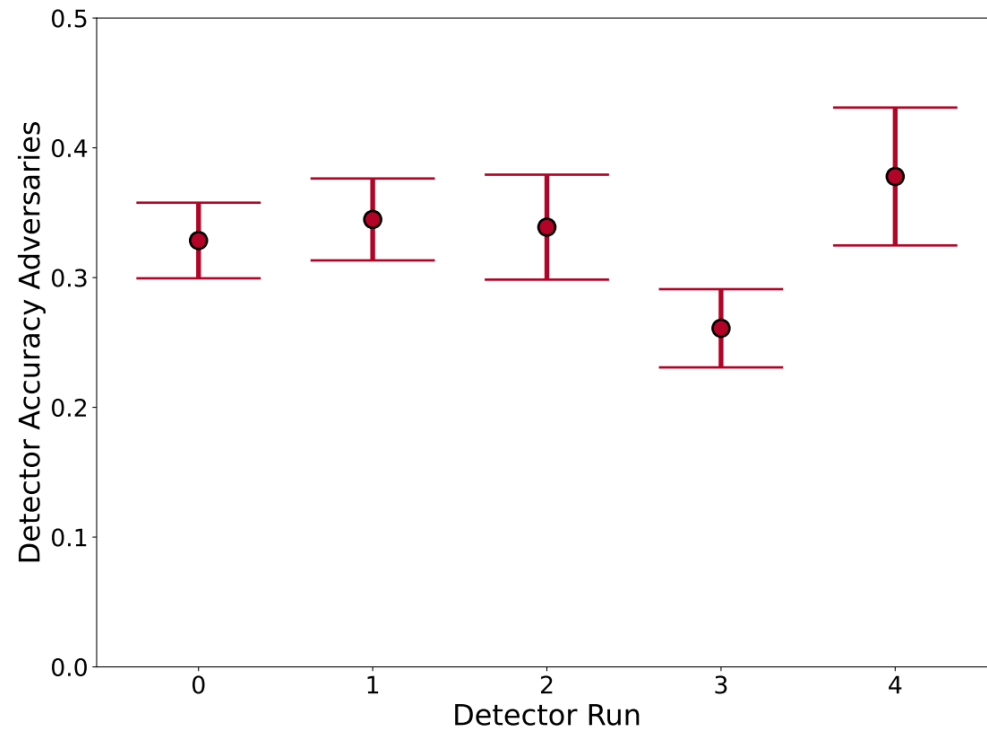


Pearson

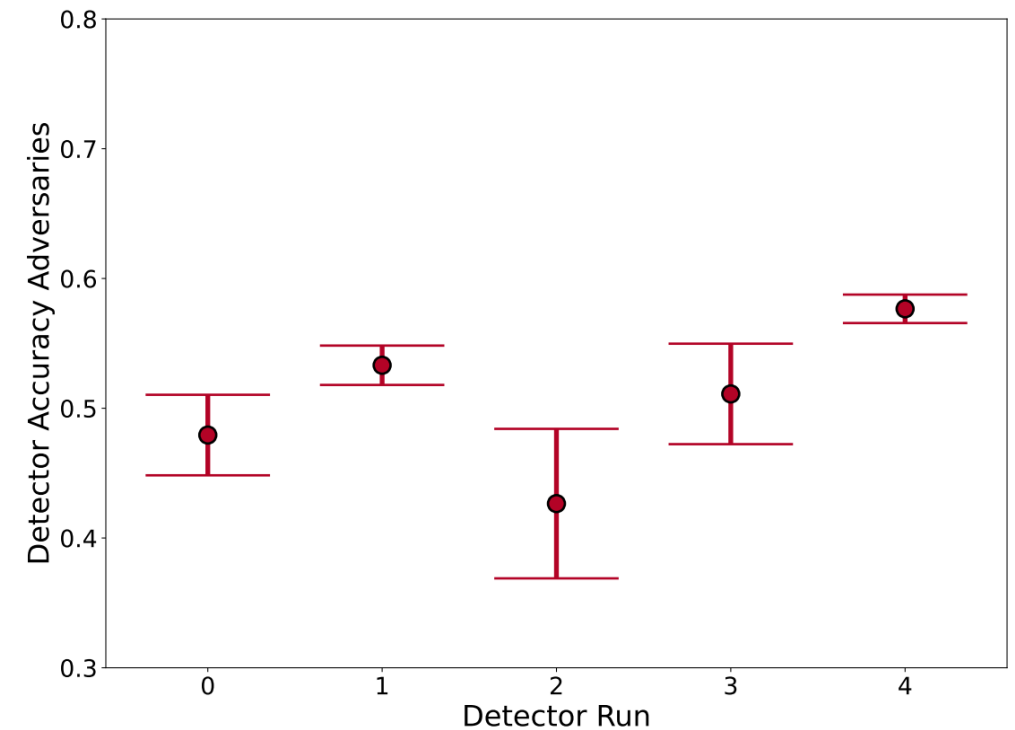


DisCo – Fooling Ratios

DisCo



Pearson



DisCo – Detector Efficiencies

DisCo – Training Data

- Both previous results were obtained on much less data
- Question: How does the efficiency of the detector scale with the amount of training samples?

DisCo – Training Data

- Both previous results were obtained on much less data
- Question: How does the efficiency of the detector scale with the amount of training samples?

