Research Training Group
Physics of the Heaviest
Particles at the LHC

RWTH AACHEN UNIVERSITY

Collaborative Research Center TRR 257
PPH
Particle Physics Phenomenology after the Higgs Discovery

TTK Institute for Theoretical Particle Physics and Cosmology

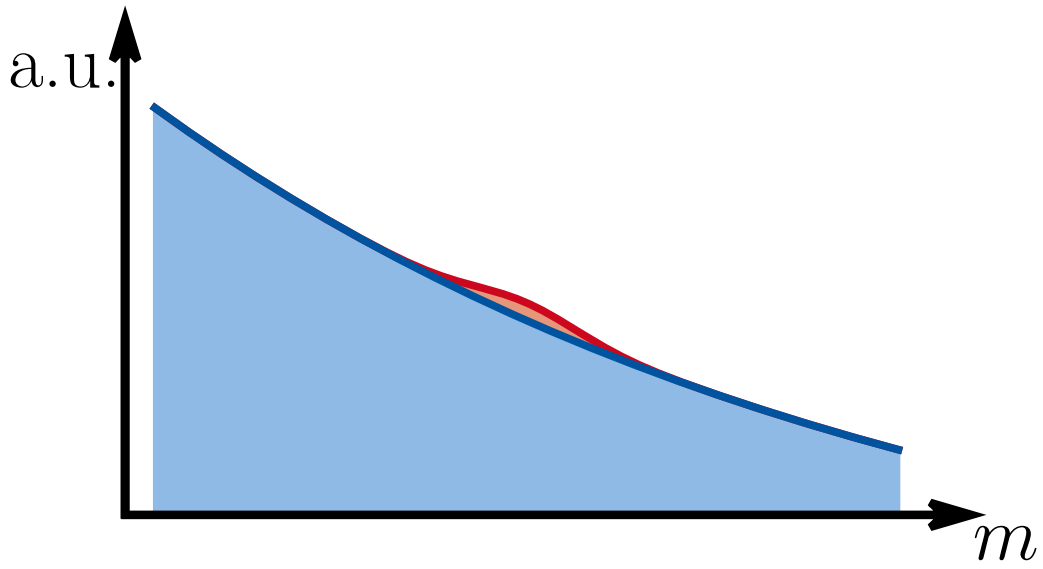
RWTH AACHEN UNIVERSITY

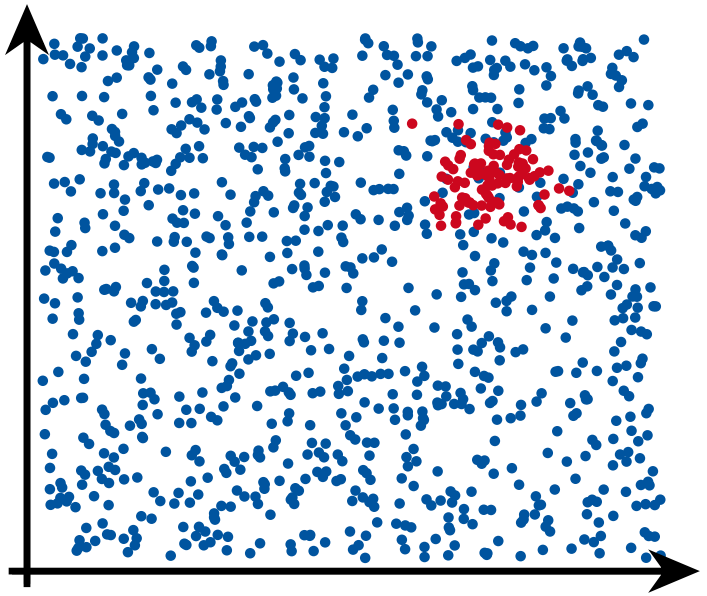
Look-Everywhere Effects in Anomaly Detection

Marie Hein

Seminar AI+HEP

Landscape of New Physics Searches





Inclusive

Landscape of New Physics Searches

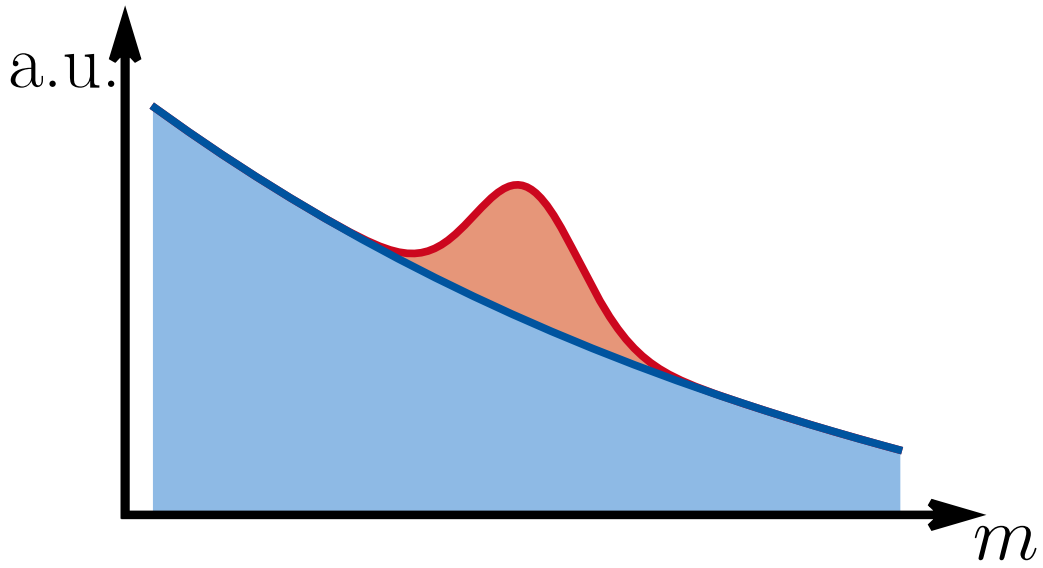
Institute for Theoretical Particle Physics and Cosmology | RWTH AACHEN UNIVERSITY

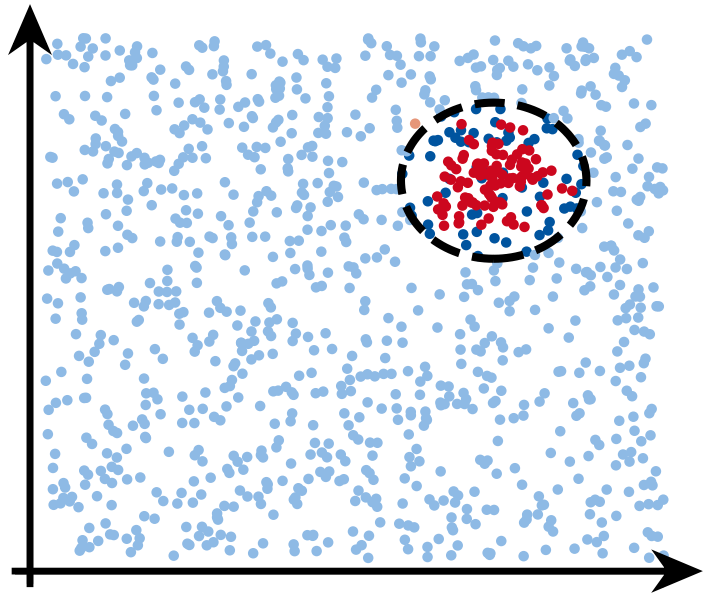
Coverage

Broad ← → Narrow

Sensitivity

Low ← → High

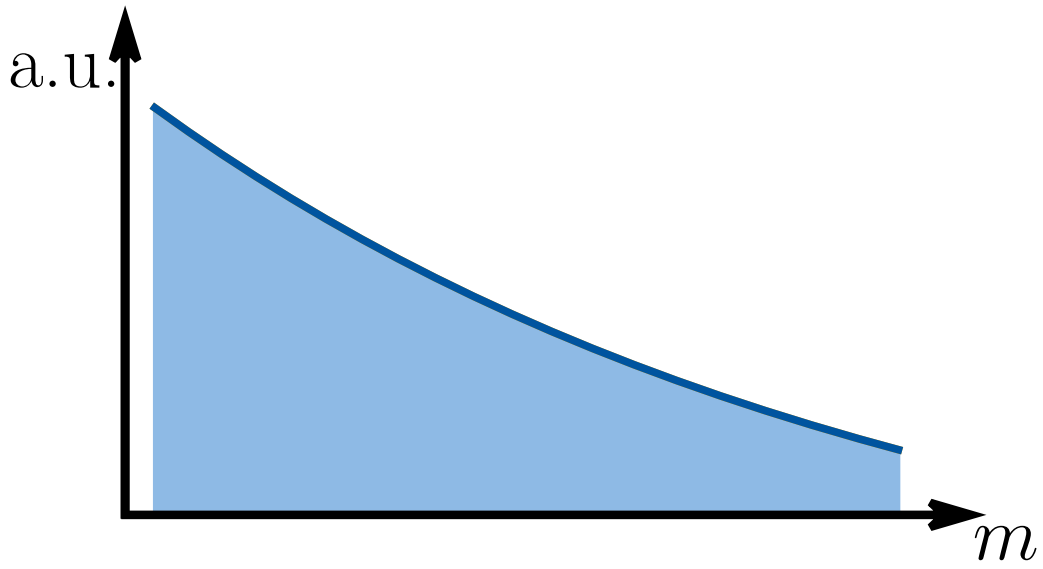


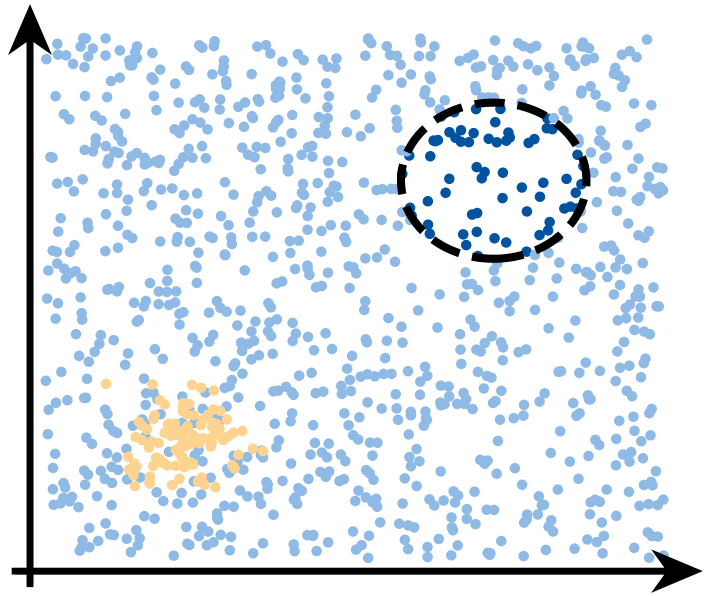


Inclusive

Model-specific

Landscape of New Physics Searches

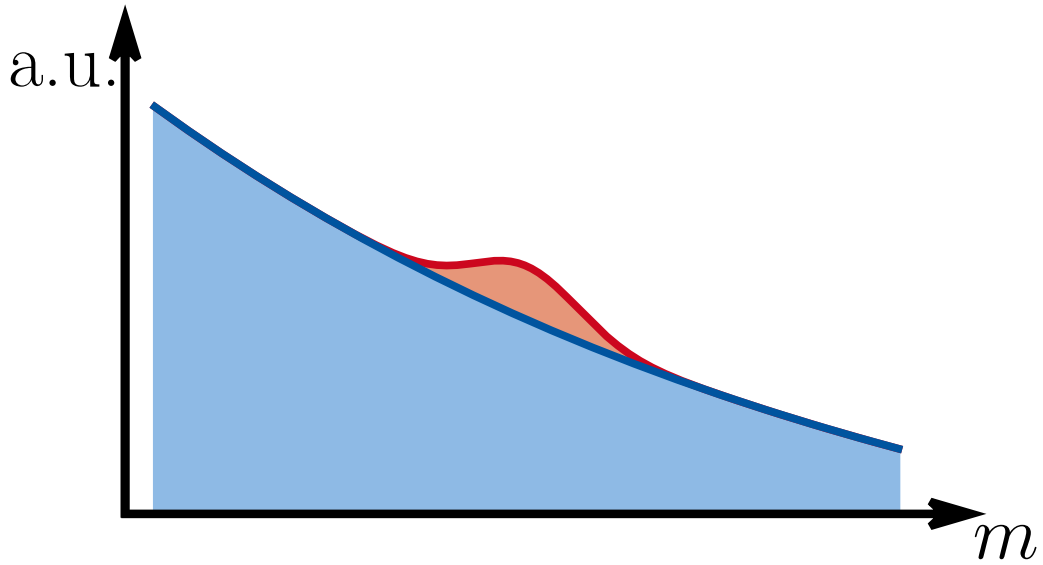


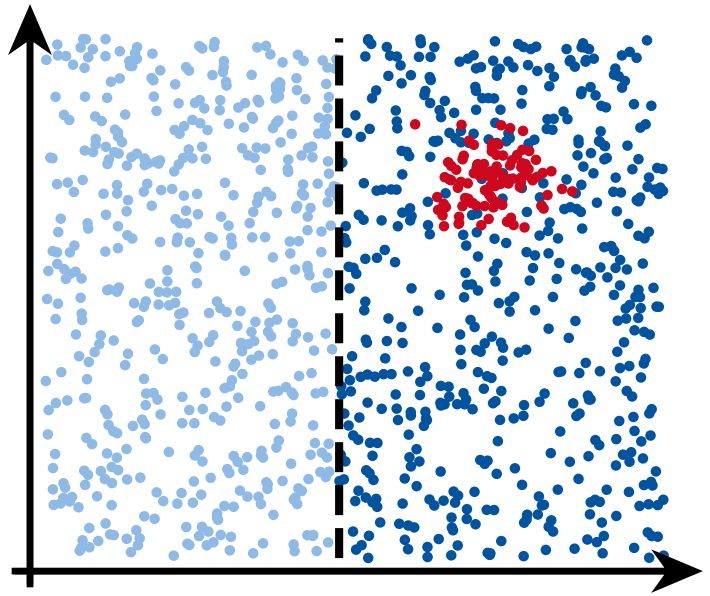


Inclusive

Model-specific

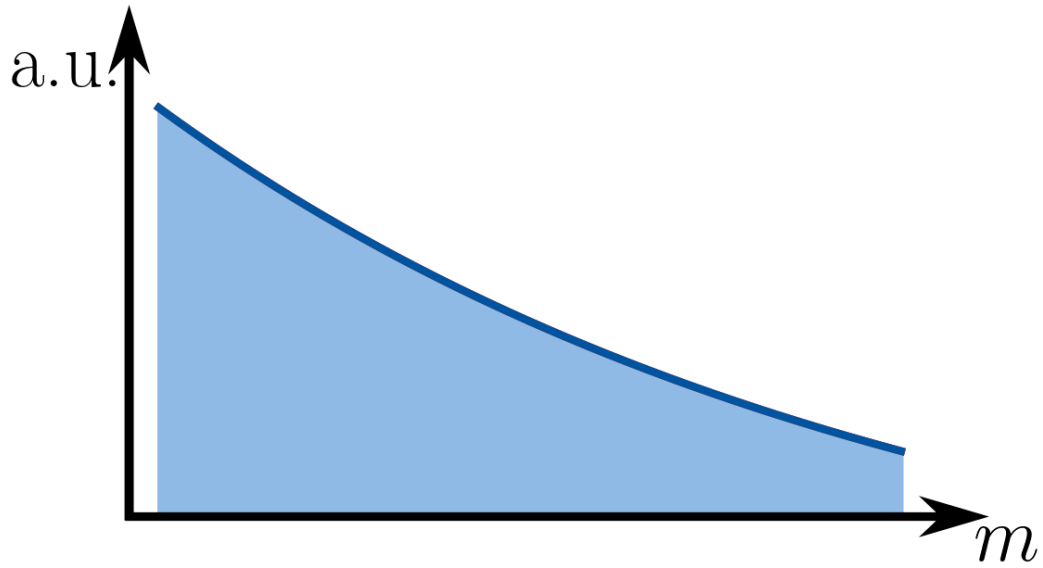
Landscape of New Physics Searches

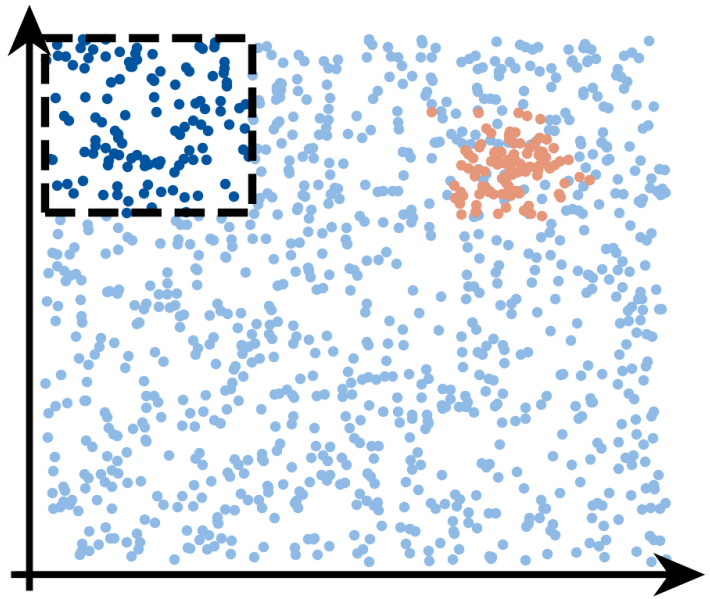




Inclusive | Single cut | Model-specific

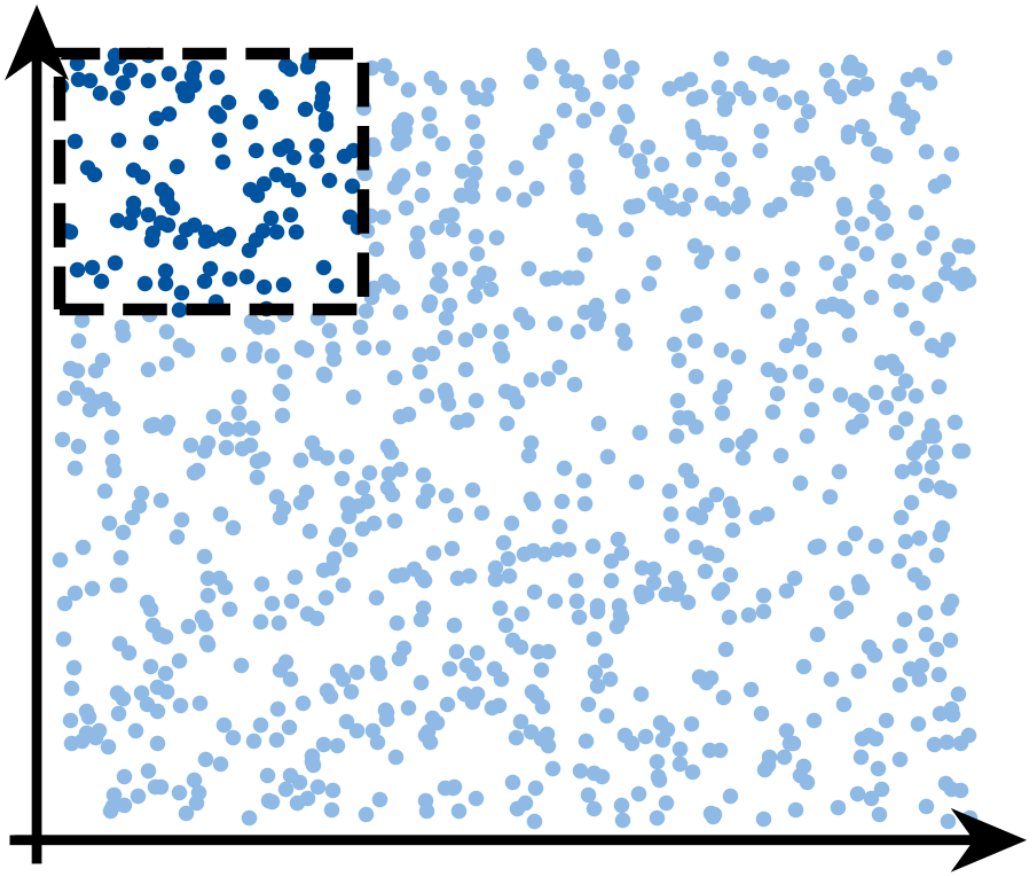
Landscape of New Physics Searches





Inclusive | Single cut | Model-specific | Anomaly detection

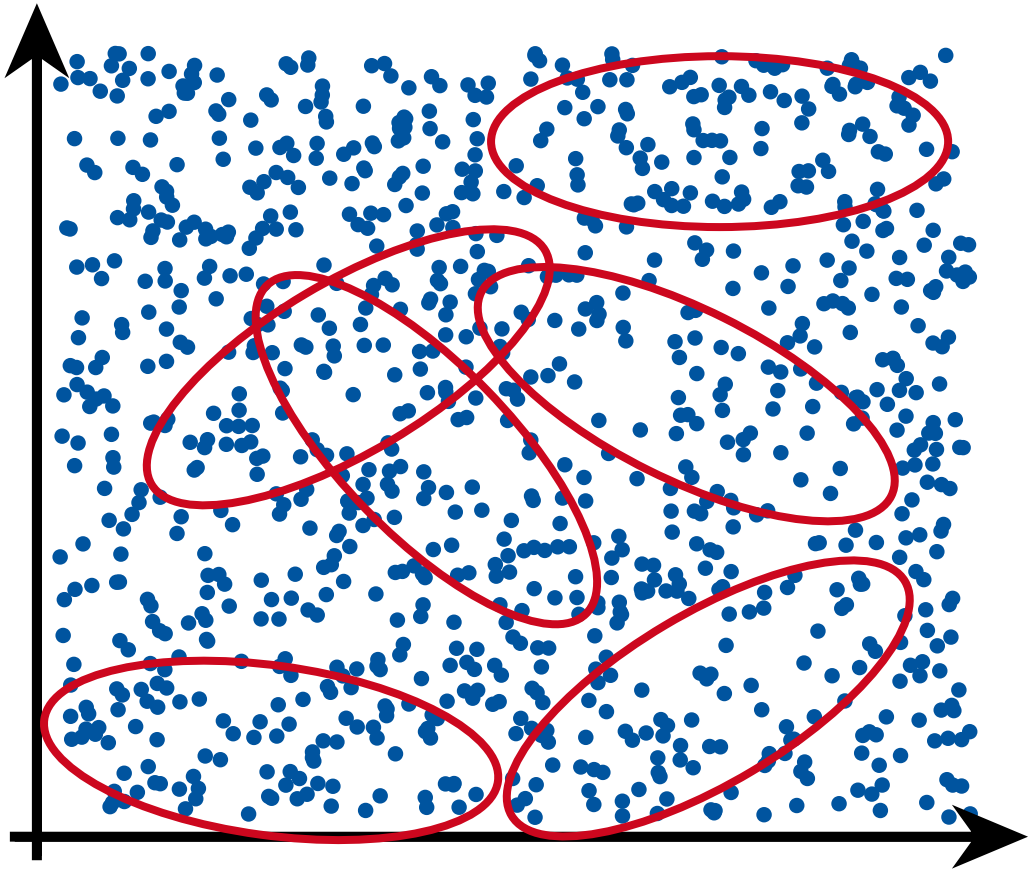
Look-Elsewhere Effect



- Scanning over multiple bins and selecting the largest excess enhances the probability of obtaining a large excess
- Typically, report analysis results as a significance corresponding to a p-value
  - →Probability of obtaining a deviation at least as large as the observed one under the null hypothesis (background-only test)
- Each bin has local p-value, whole analysis with multiple tests has global p-value
- Picking best of multiple tests leads to look-elsewhere effect
  - →Reporting local p-value as global results in miscalibration of p-values

How does this affect ML analyses?

"Look everywhere effects in anomaly detection" [2512.13787], **MH**, B. Nachman, D. Shih



NNs look everywhere in the phase space during training, is this multiple testing?

Outline

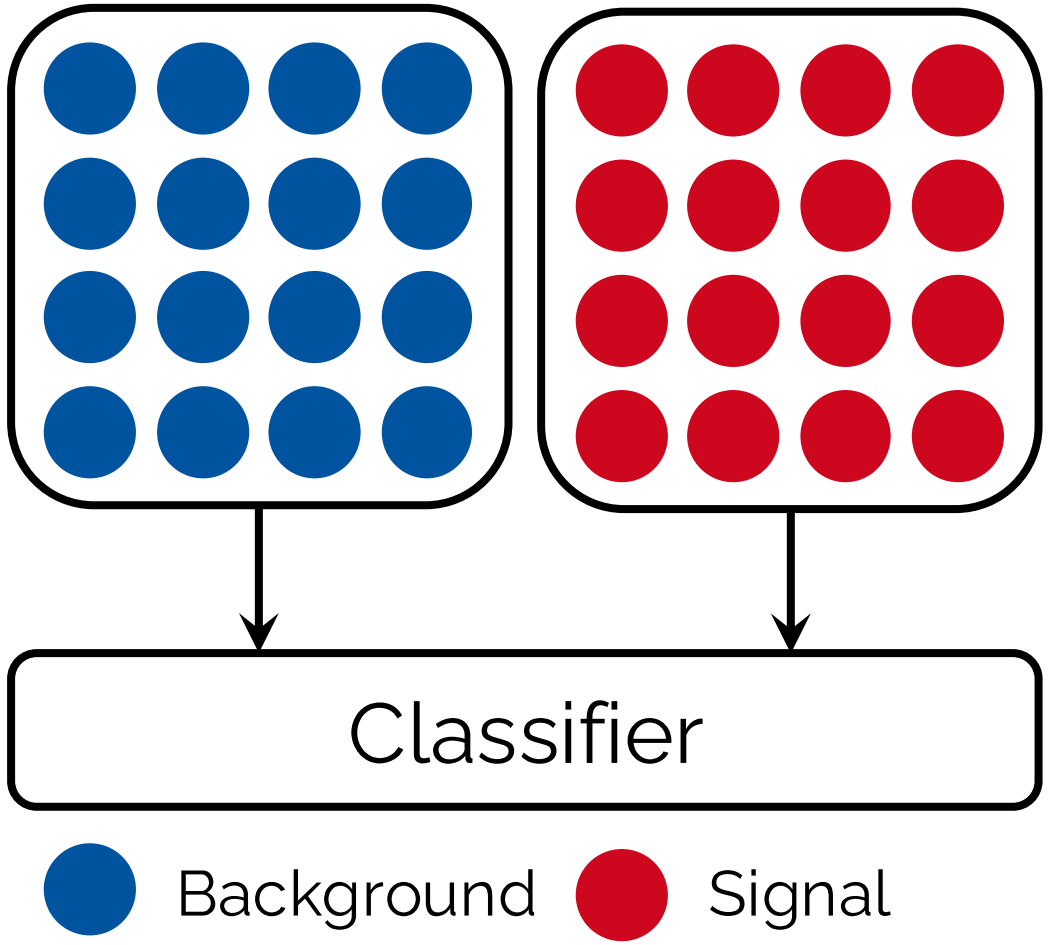
1. Formulating the question ✔
2. New Physics Searches and Weakly Supervised Anomaly Detection
3. Binned Anomaly Detection and Look-Elsewhere Effects
4. Machine Learning and Look-Everywhere Effects
5. The Impact of Look-Everywhere Effect Mitigation on Analysis Sensitivity
6. Conclusions and Outlook

New Physics Searches and Weakly Supervised Anomaly Detection

Classification Problem

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$



Weakly Supervised Classification

"Classification without labels: Learning from mixed samples in high energy physics" [1709.02949], E. Metodiev, B. Nachman, J. Thaler

- Optimal classifier

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}$$

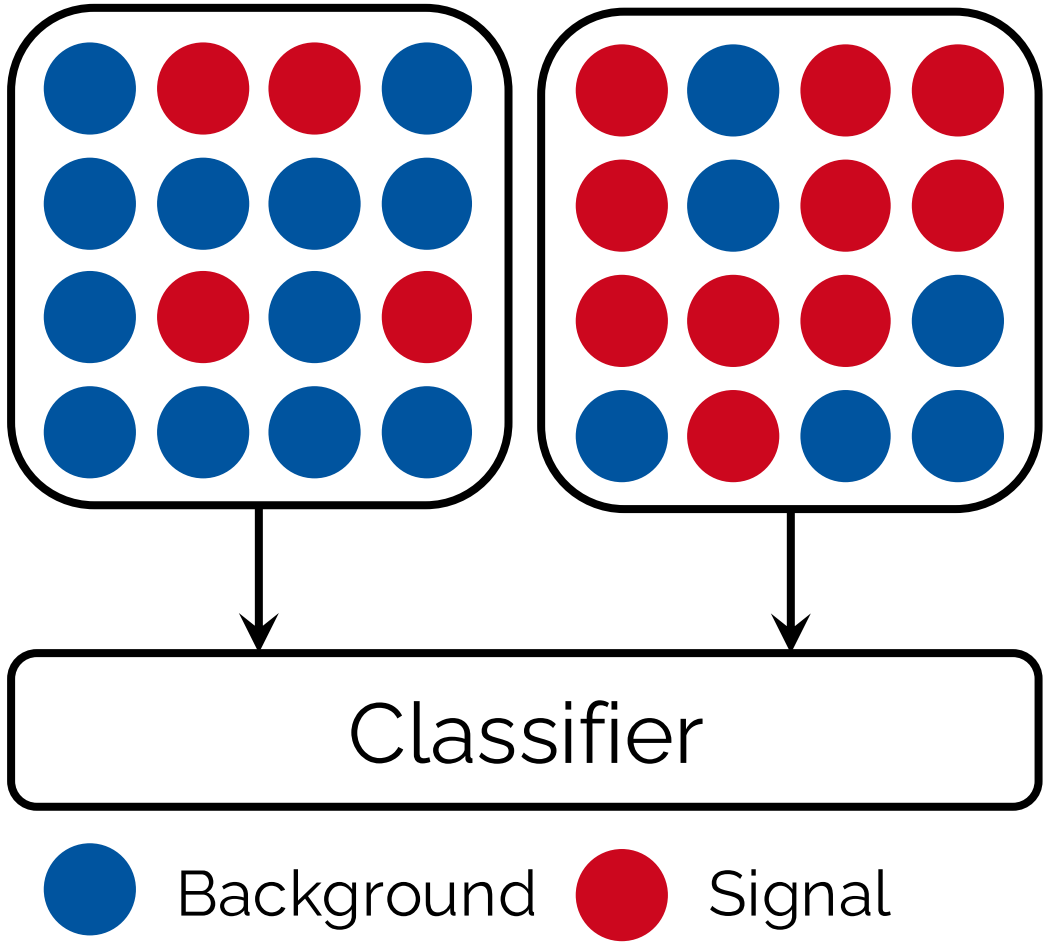
- For mixed datasets with signal fractions f_i

$$R_{\text{mixed}}(x) = \frac{f_1 R_{\text{optimal}}(x) + (1 - f_1)}{f_2 R_{\text{optimal}}(x) + (1 - f_2)}$$

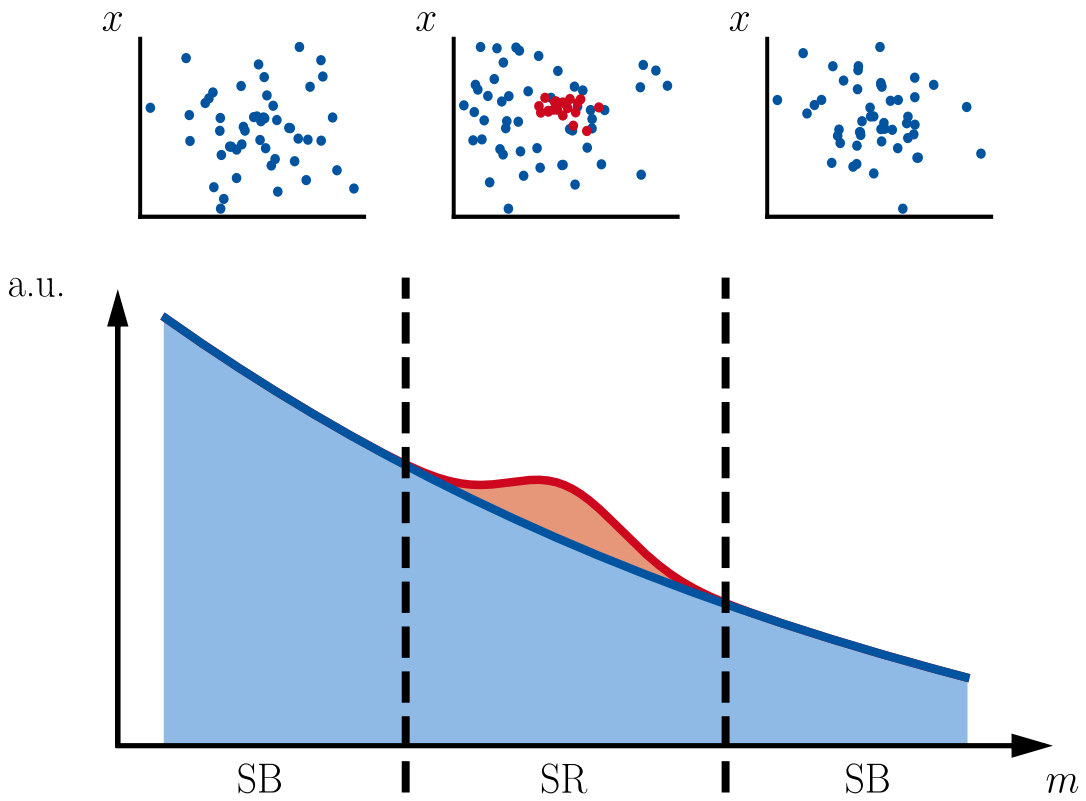
→Monotonically increasing function of

$R_{\text{optimal}}(x)$ as long as $f_1 > f_2$

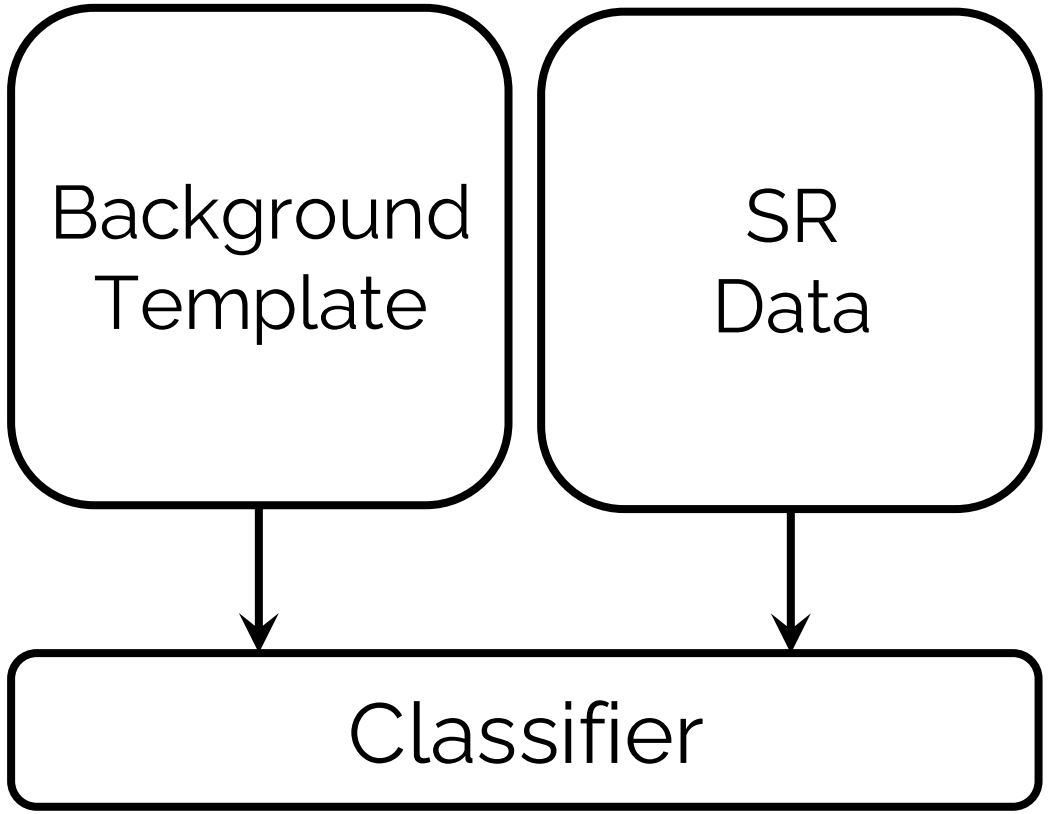
→Same decision boundaries



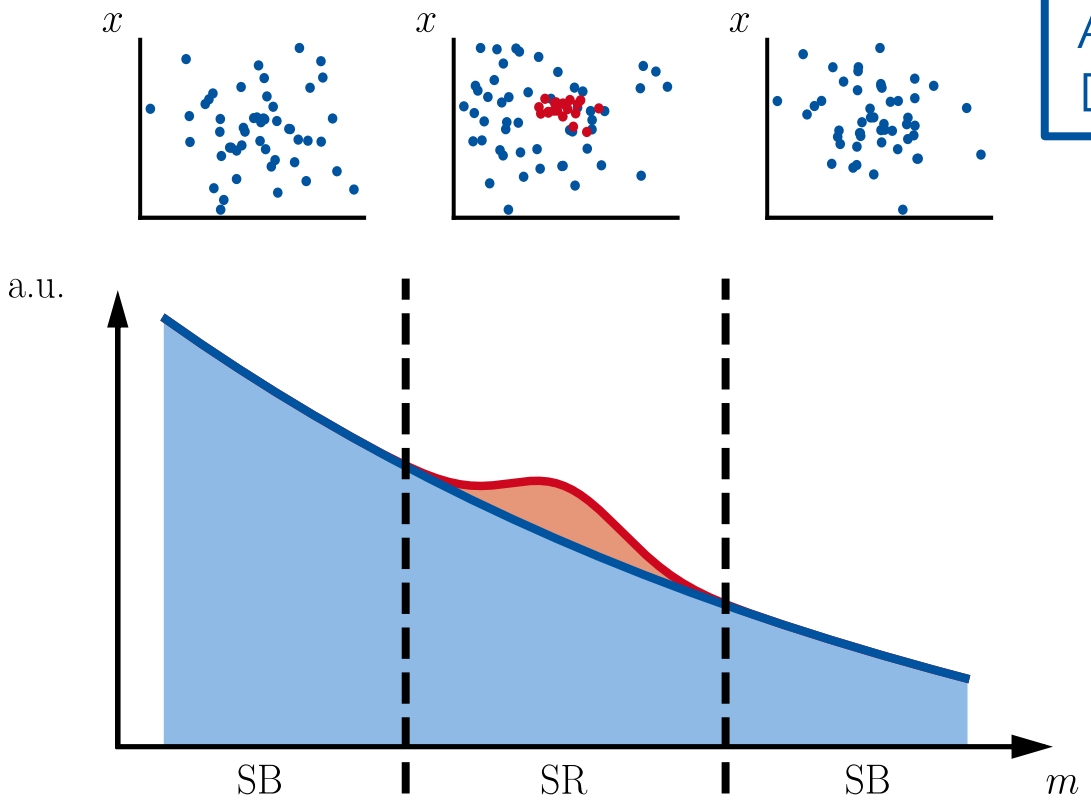
Application to resonance searches



Recreated from [2109.00546]



Application to resonance searches



Idealized Anomaly Detector

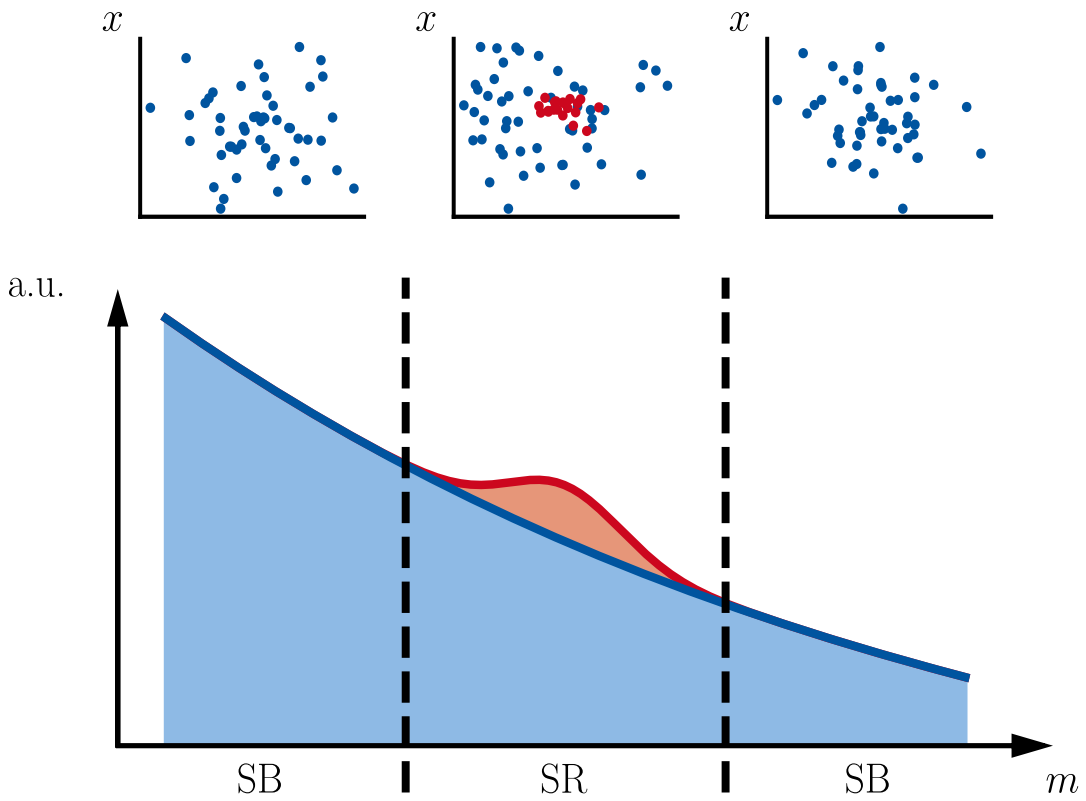
SR Background

SR Data

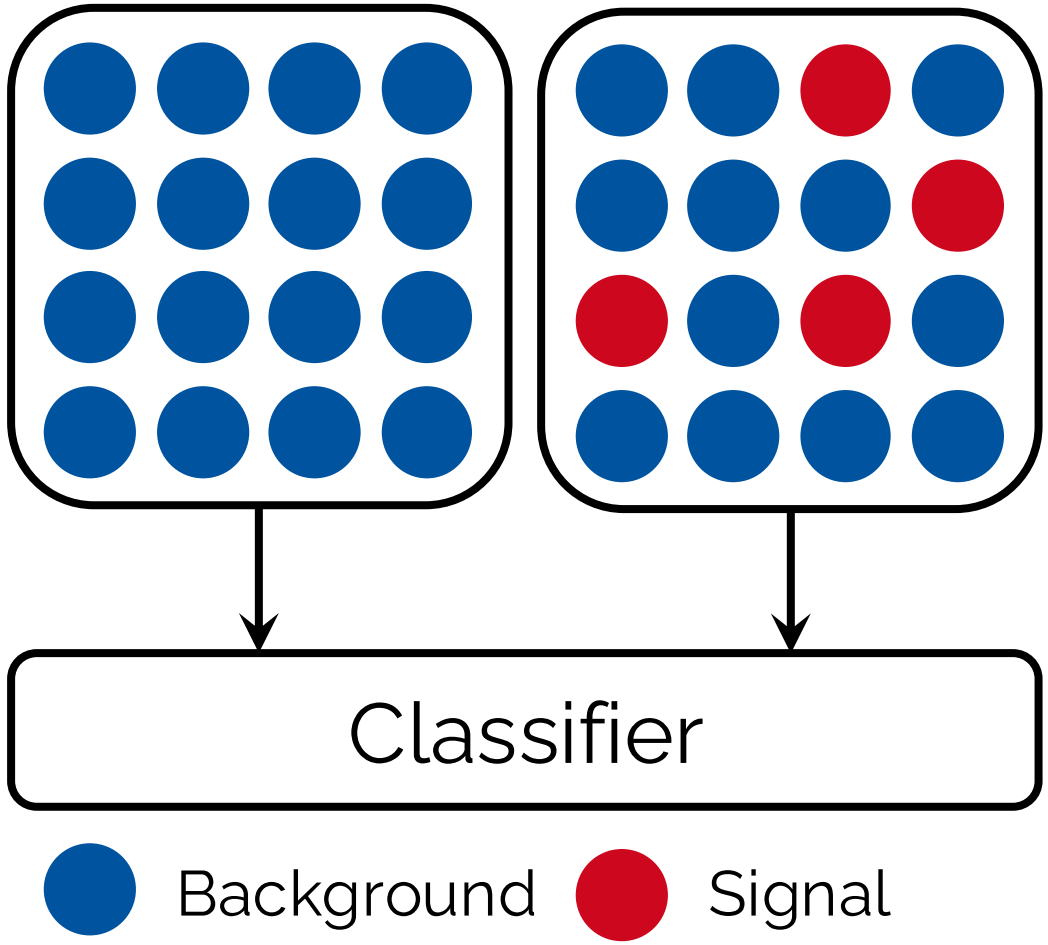
Classifier

Recreated from [2109.00546]

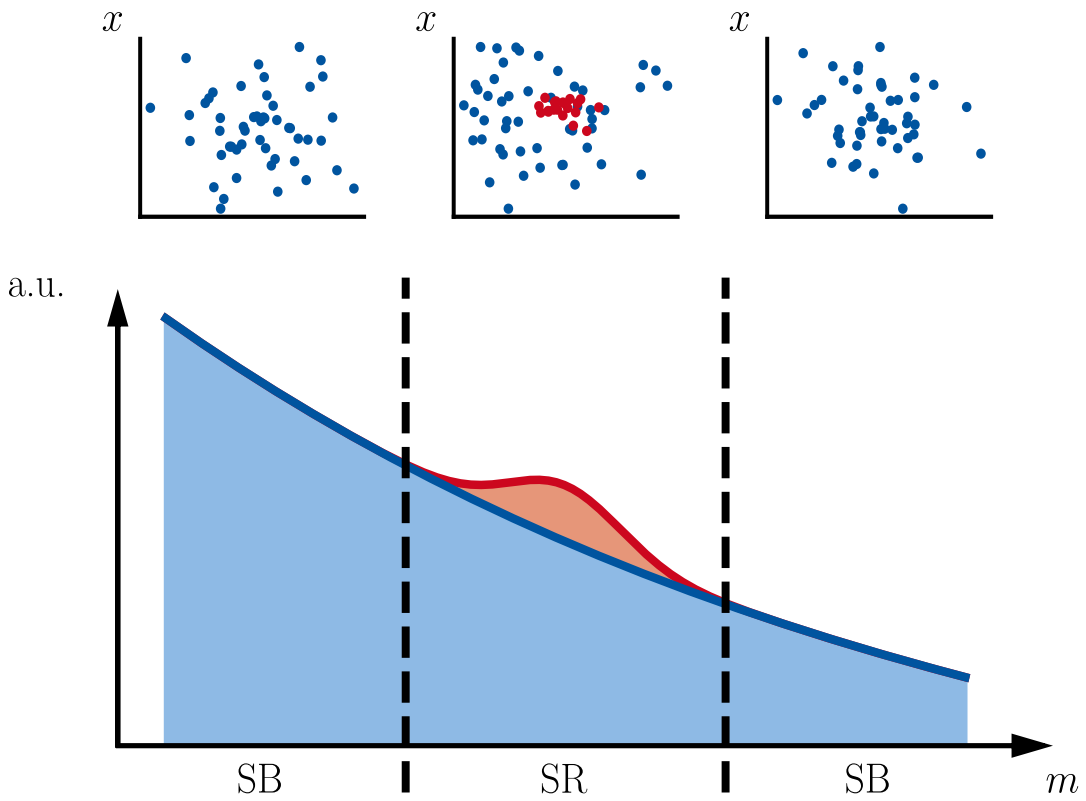
Application to resonance searches



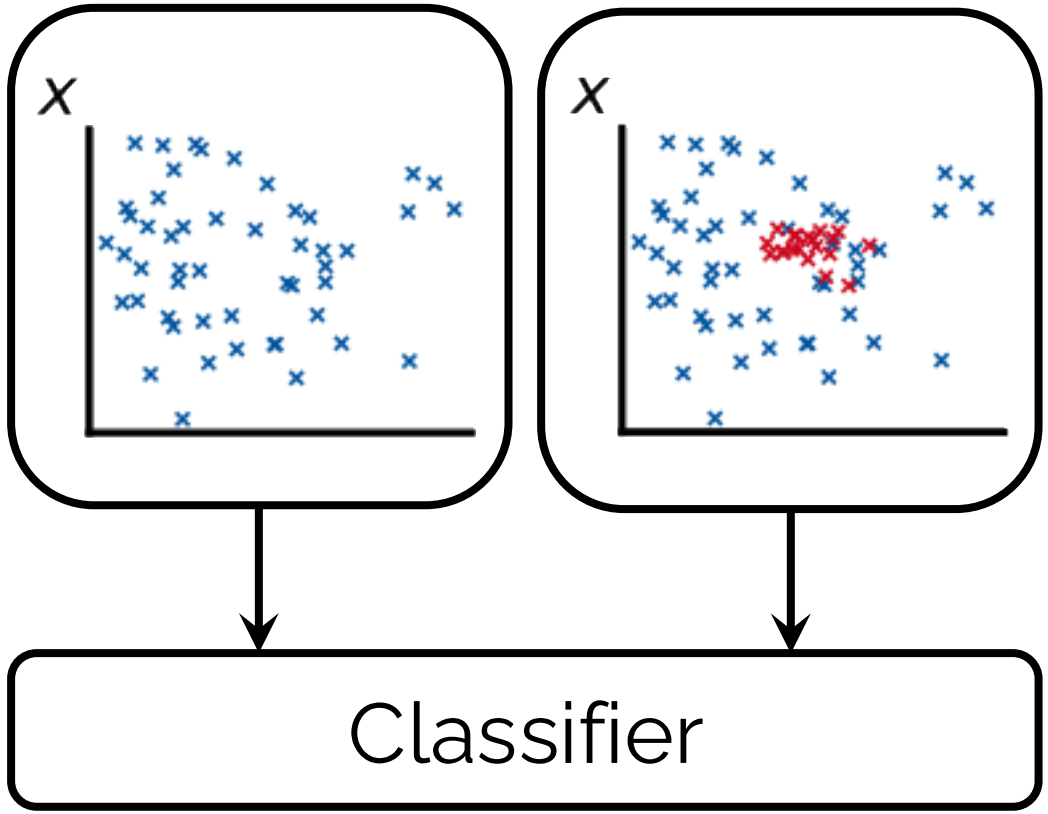
Recreated from [2109.00546]



Application to resonance searches



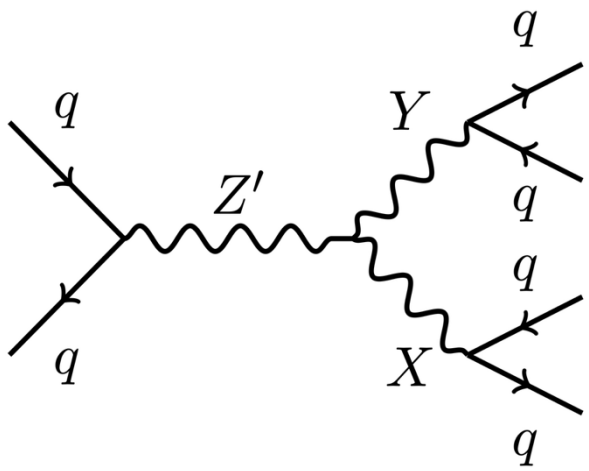
Recreated from [2109.00546]

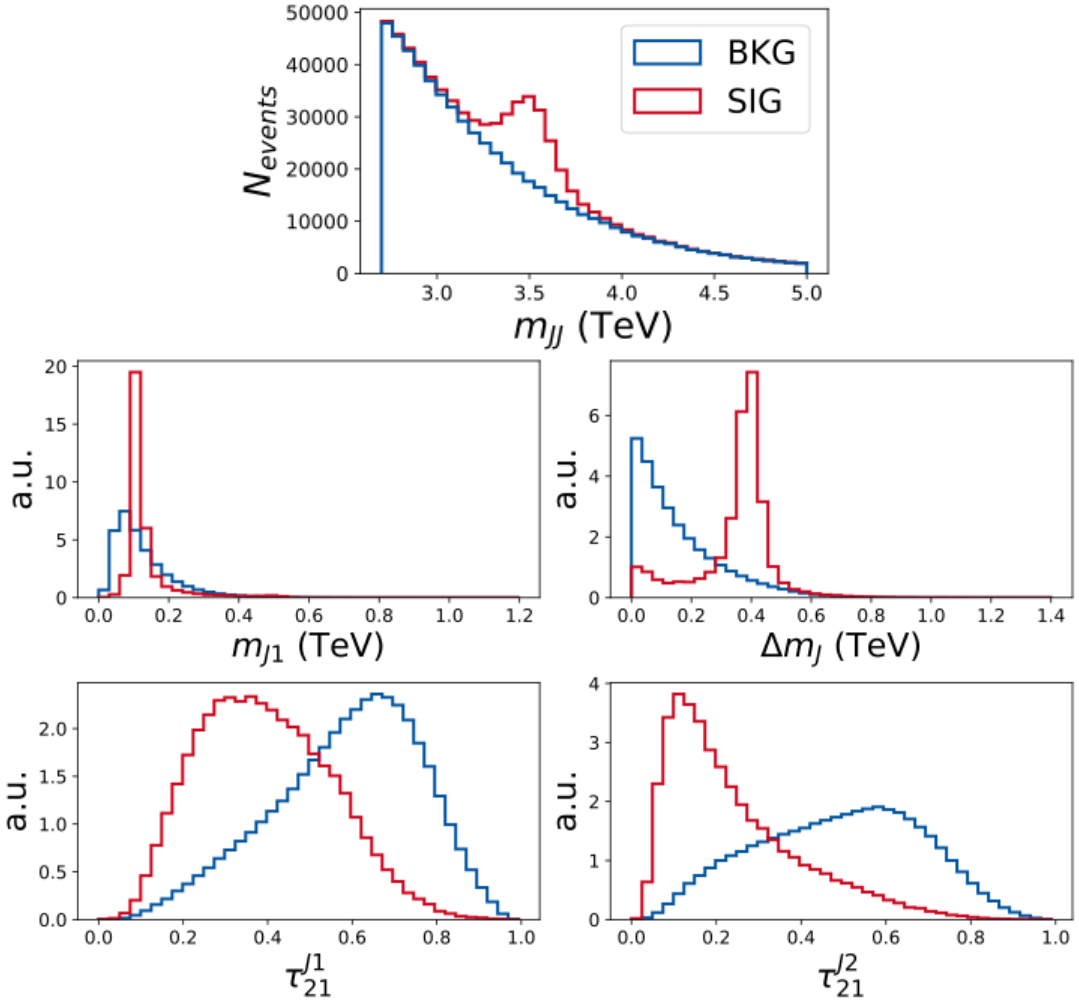


LHCO R&D dataset

"The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics" [2101.08320], G. Kasieczka, B. Nachman, D. Shih et. al.

- Benchmark dataset for anomaly detection
- 1 M QCD dijet background events
- 100k signal events produced via

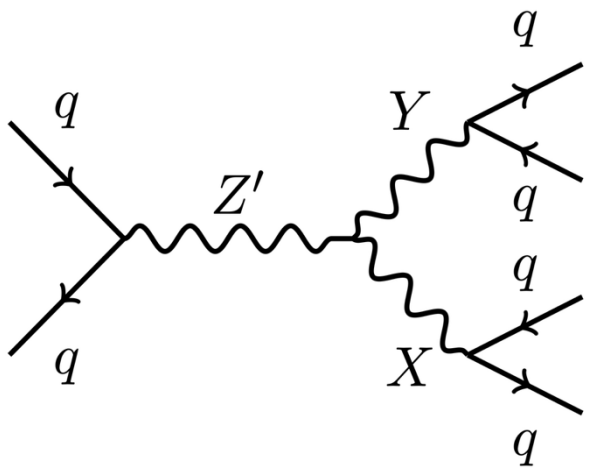


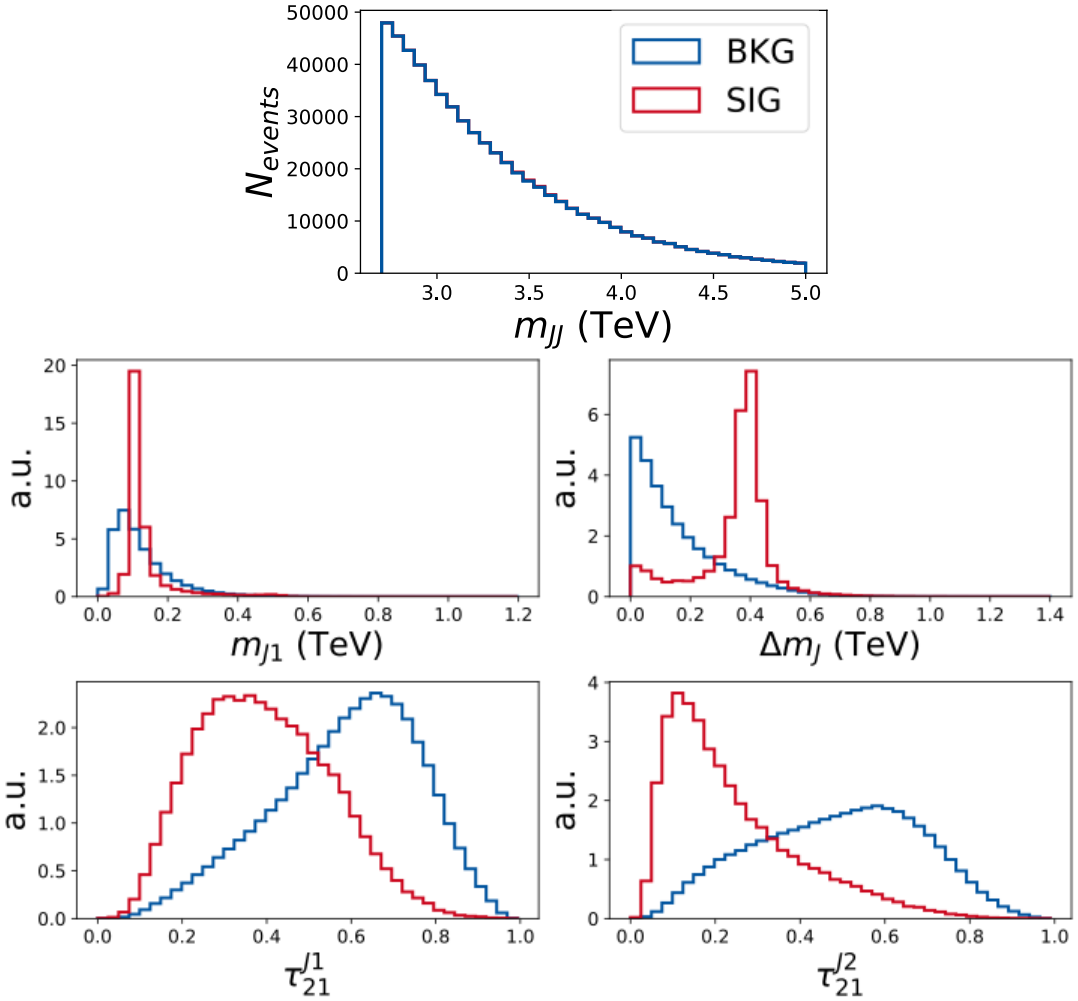


LHCO R&D dataset

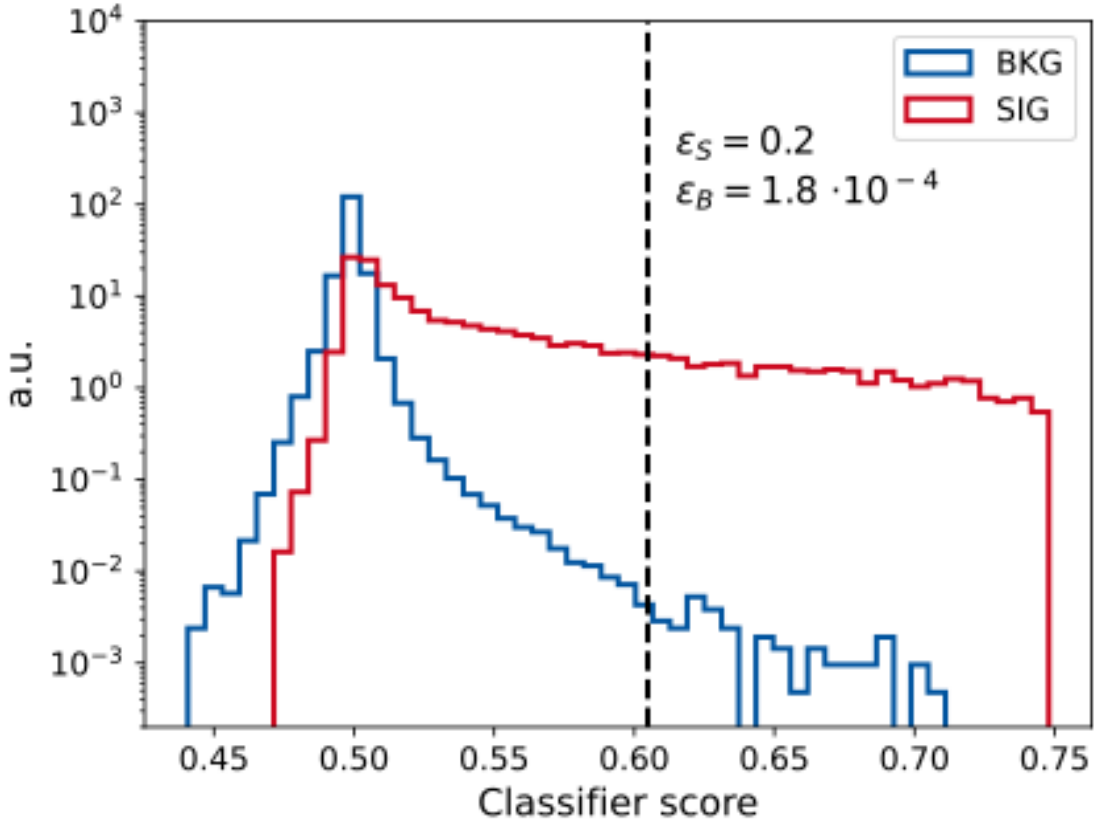
"The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics" [2101.08320], G. Kasieczka, B. Nachman, D. Shih et. al.

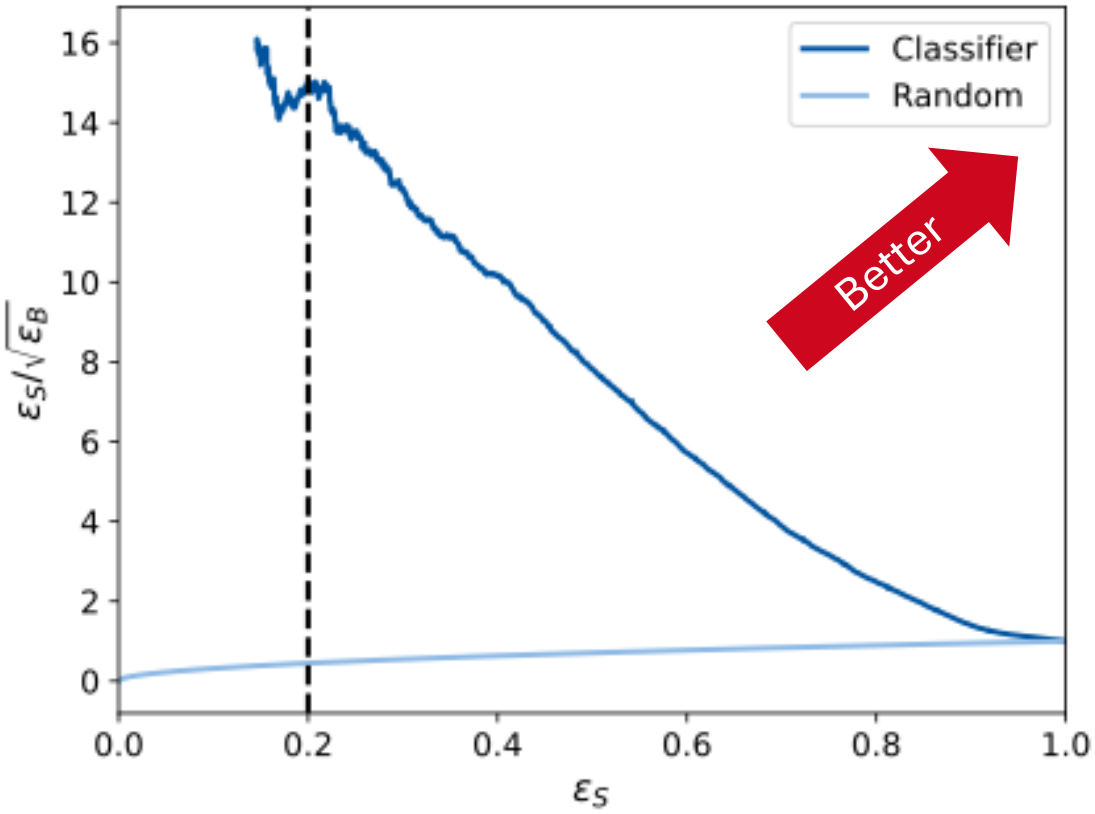
- Benchmark dataset for anomaly detection
- 1 M QCD dijet background events
- 100k signal events produced via





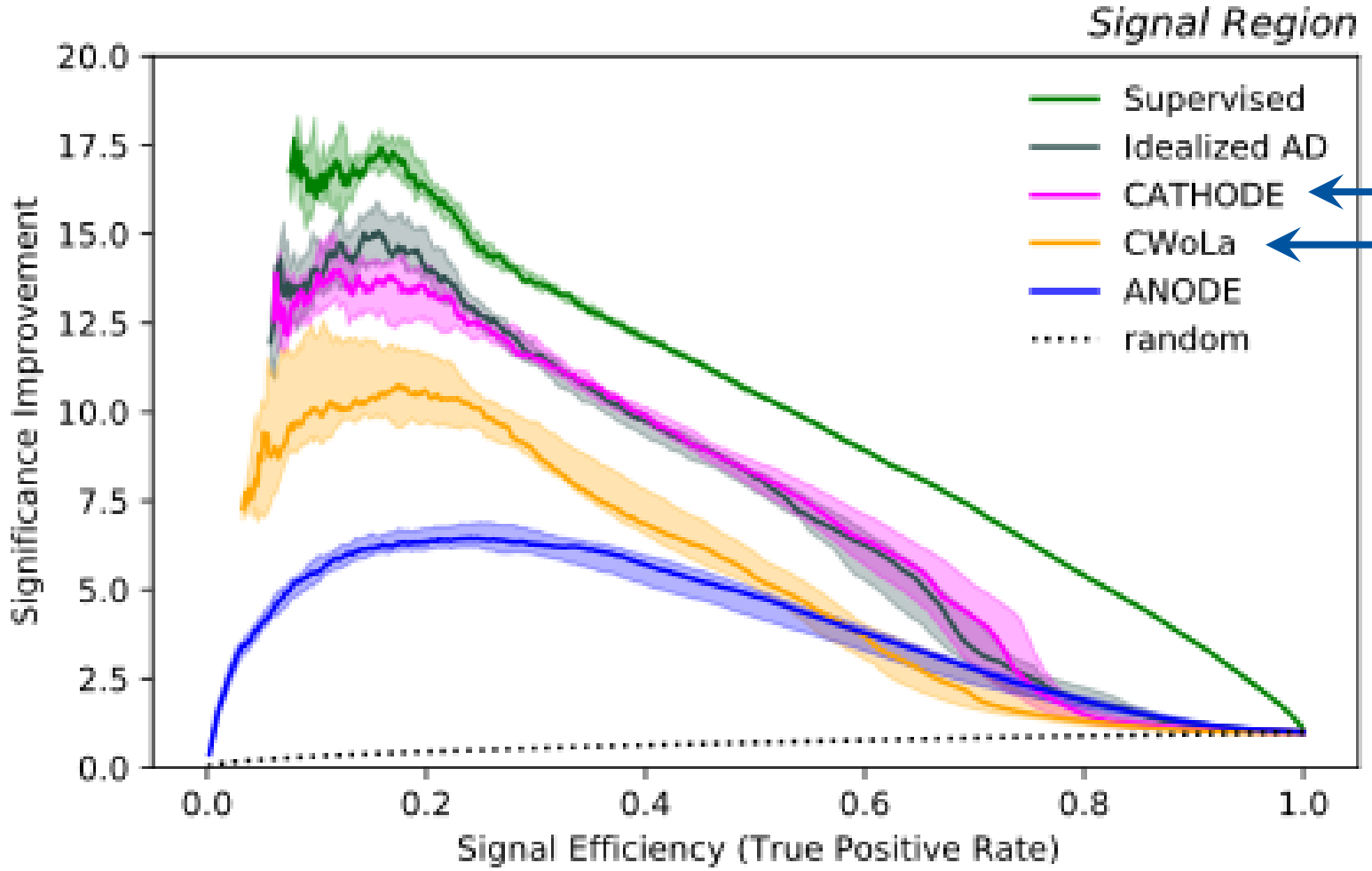
Classifier score to anomaly detection





Performance Comparison

"Classifying Anomalies Through Outer Density Estimation" [2109.00546], A. Hallin et al.

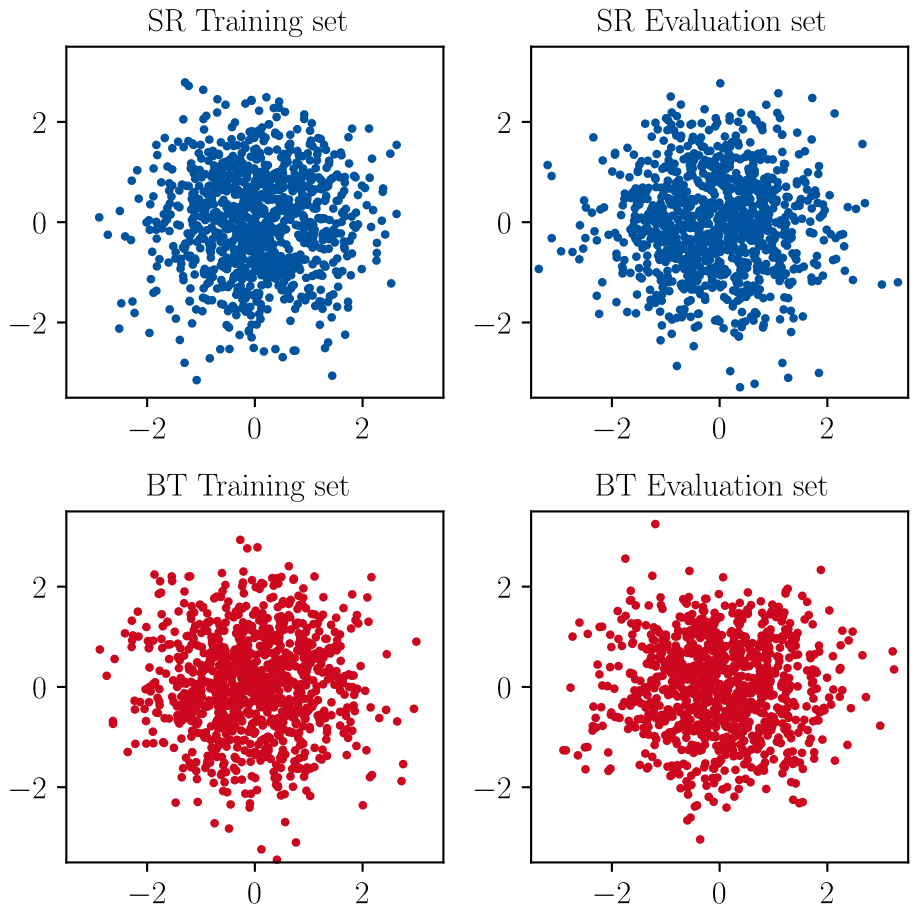


Different methods to obtain BT

Binned Anomaly Detection and Look-Elsewhere Effects

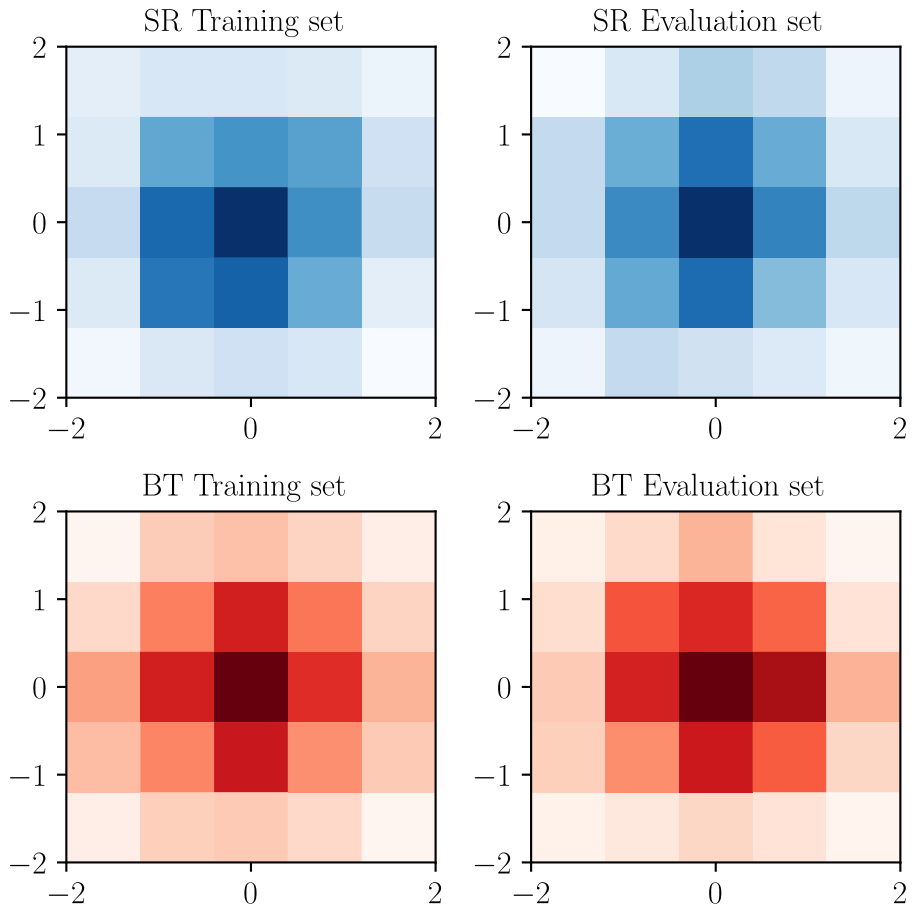
Binned Analysis Setup

1. Take a 2D Gaussian and sample SR and BT events for a training and evaluation set



Binned Analysis Setup

1. Take a 2D Gaussian and sample SR and BT events for a training and evaluation set
2. Bin both sets using 5 bins per dimension between -2 and 2



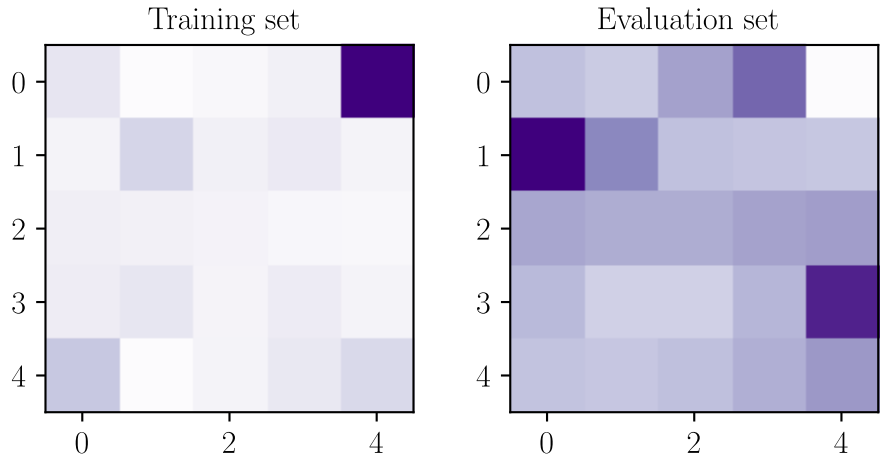
Binned Analysis Setup

1. Take a 2D Gaussian and sample SR and BT events for a training and evaluation set
2. Bin both sets using 5 bins per dimension between -2 and 2
3. Calculate binned likelihood ratio on training set for each bin i

$$R^i = \frac{N^i_{\mathrm{SR}}}{N^i_{\mathrm{BT}}}$$

4. Select bin with largest R^i
5. Calculate p-value for this bin on evaluation set

Note: we do a little trick and use infinite background statistics (i.e., working with expectation values for BT)



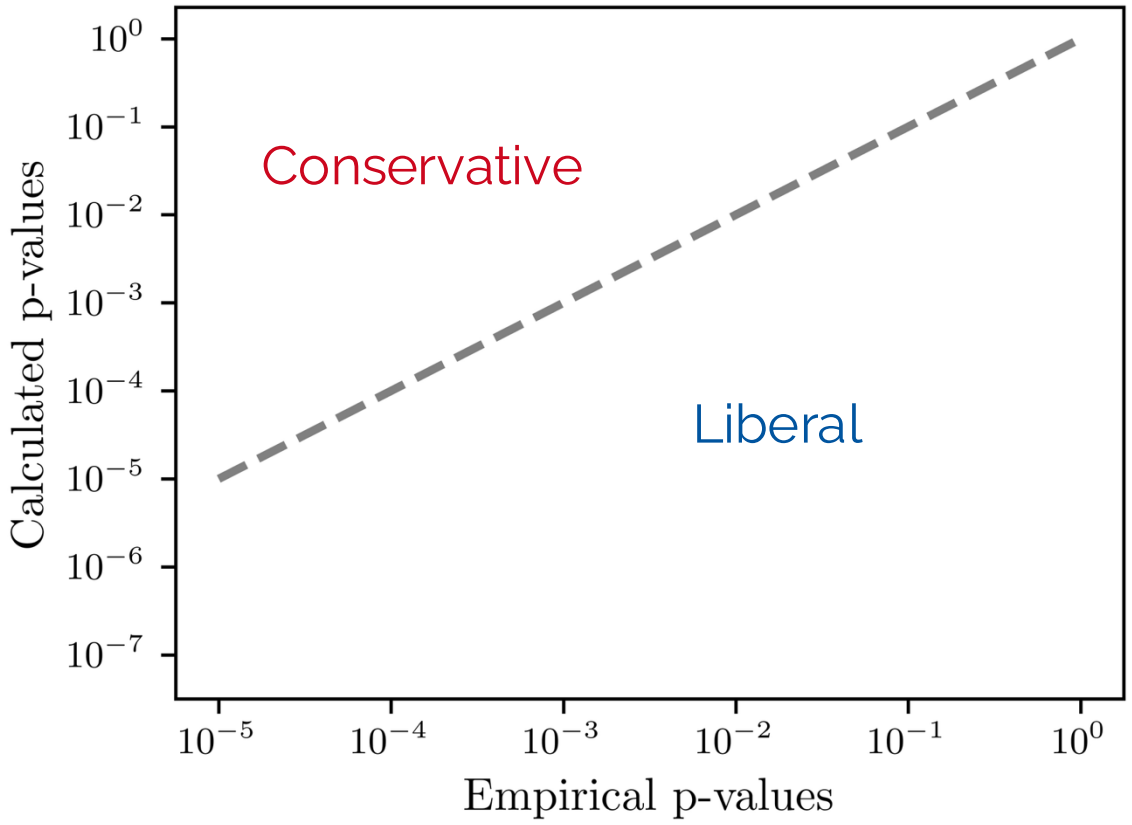
How do we test p-value calibration?

- P-values are probabilities:
  - Probability of finding an excess at least as large as was observed in a background-only test
- Probabilities can be estimated empirically by performing large number of N_{tests} tests and counting occurrences
- For any calculated p-value p, empirical p-value given

$$p_{\text{empirical}} = \frac{\text{Number}\left(p_{\text{calculated},i} \leq p\right)}{N_{\text{tests}}}$$

- For calibrated p-values, within statistical uncertainties

$$p_{\text{empirical}} = p_{\text{calculated}}$$



Reporting the local p-value as global

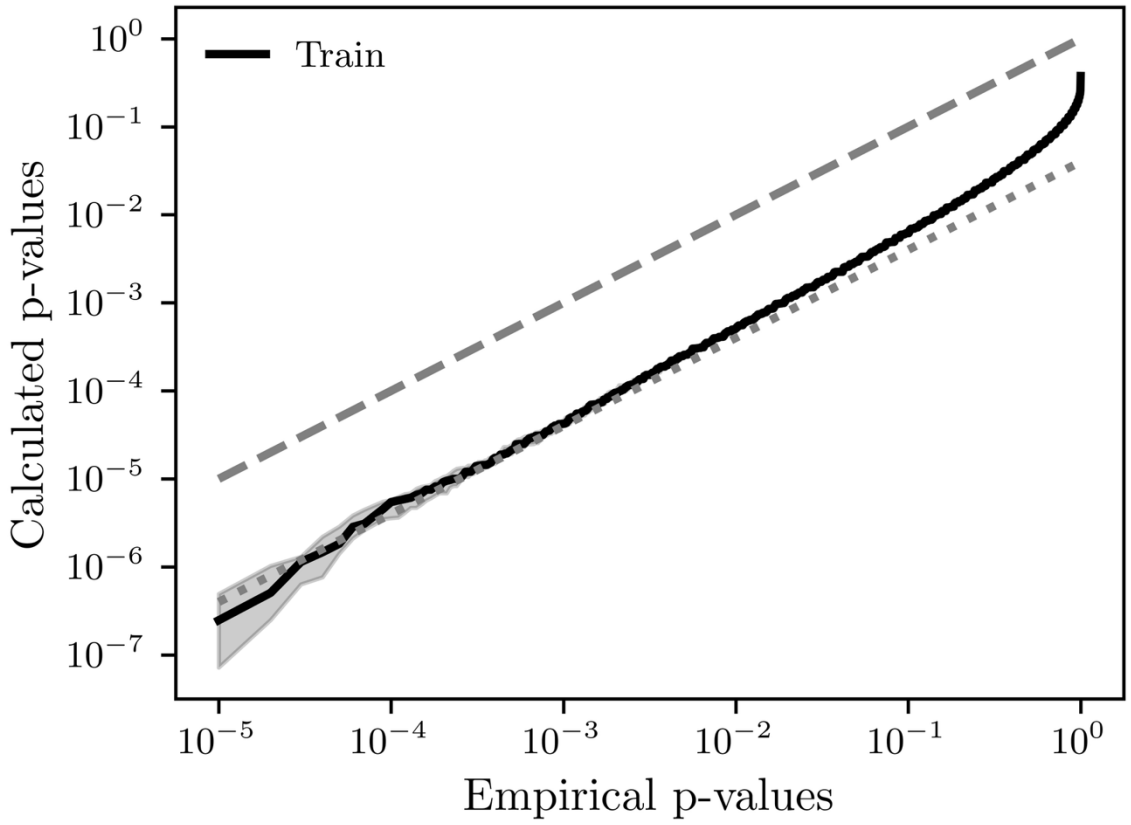
Or: Evaluating on the training set

- Training and evaluation set are one and the same
- Results in miscalibrated p-values (look-elsewhere effect)
- Bonferroni correction: Choosing best of N tests gives a trials factor of N, i.e.

$$p_{\text{calibrated}} = N \cdot p_{\text{uncalibrated}}$$

→Conservative correction assuming uncorrelated tests

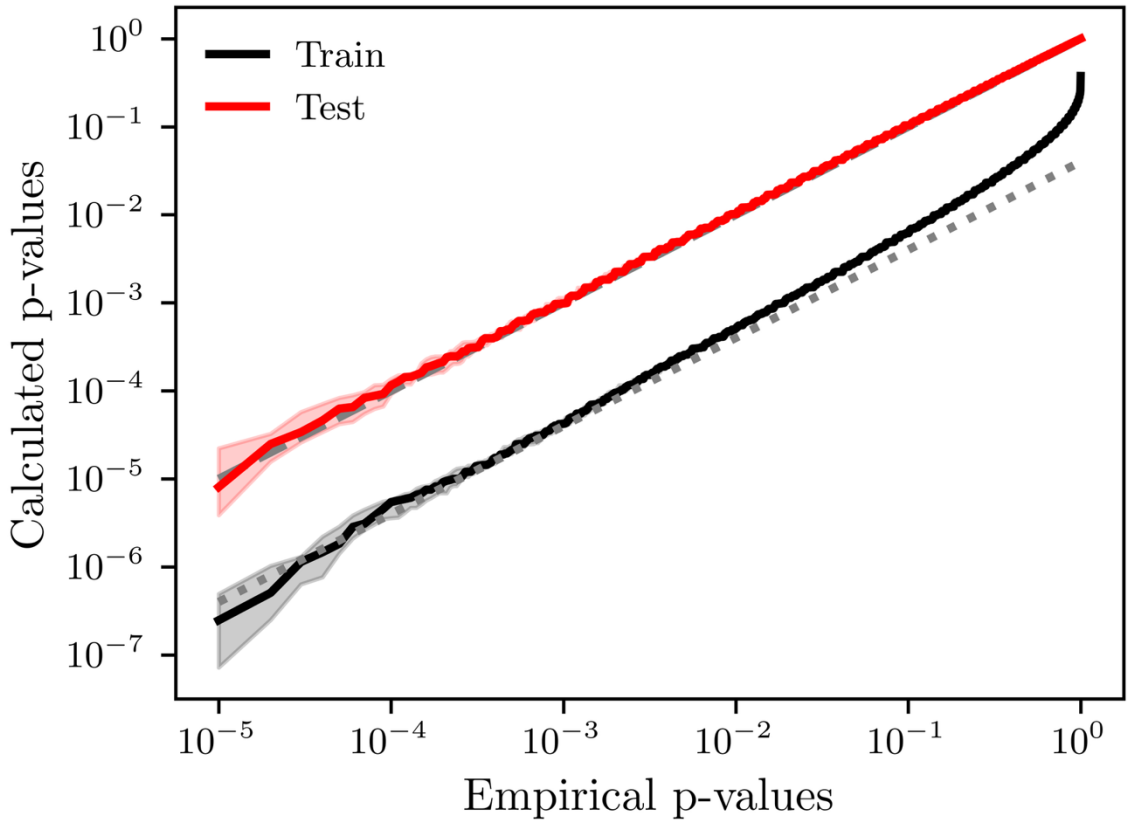
→Works well for very low p-values



Doing a follow-up analysis

Or: Evaluating on an independent test set

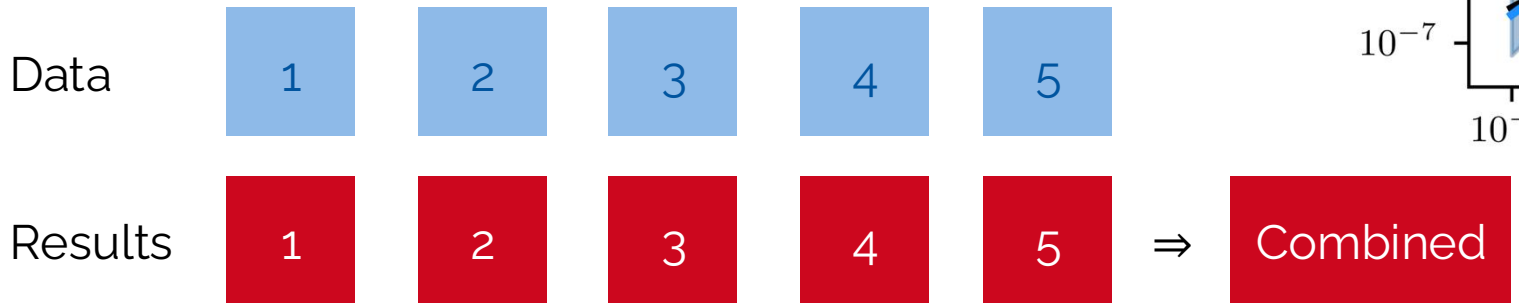
- Split data 50-50 into training and test set to use the independent test set as evaluation set
- Results in calibrated p-values
- Compared to evaluating on the training set, lose training and evaluation statistics

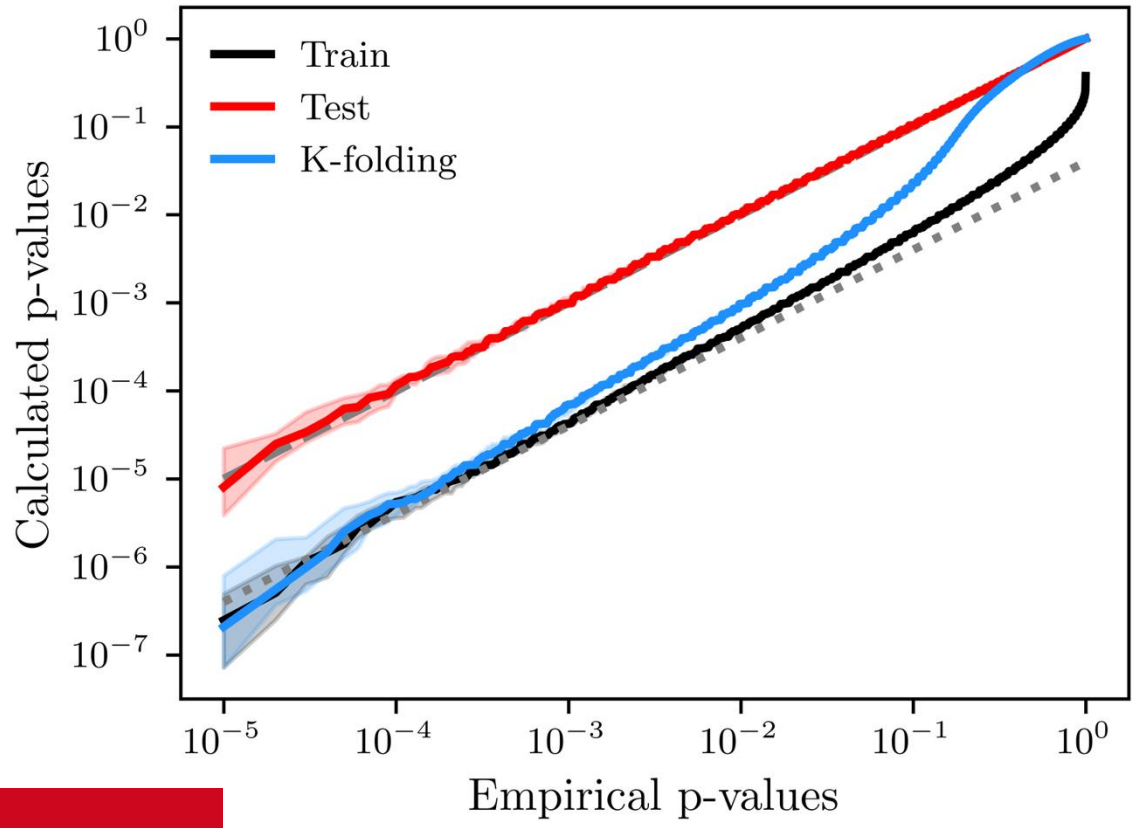


Evaluation using k-fold cross validation

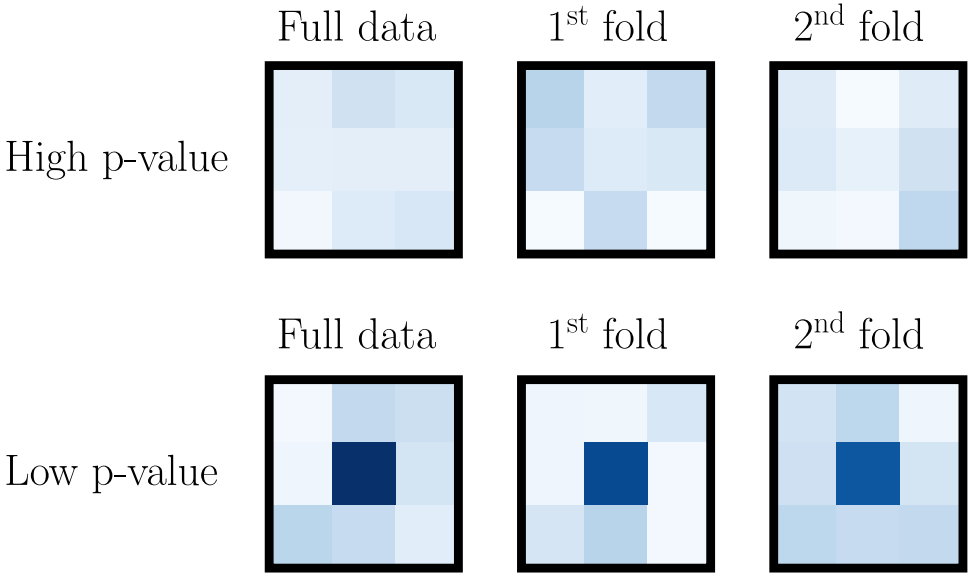
Method to preserve training and evaluation statistics

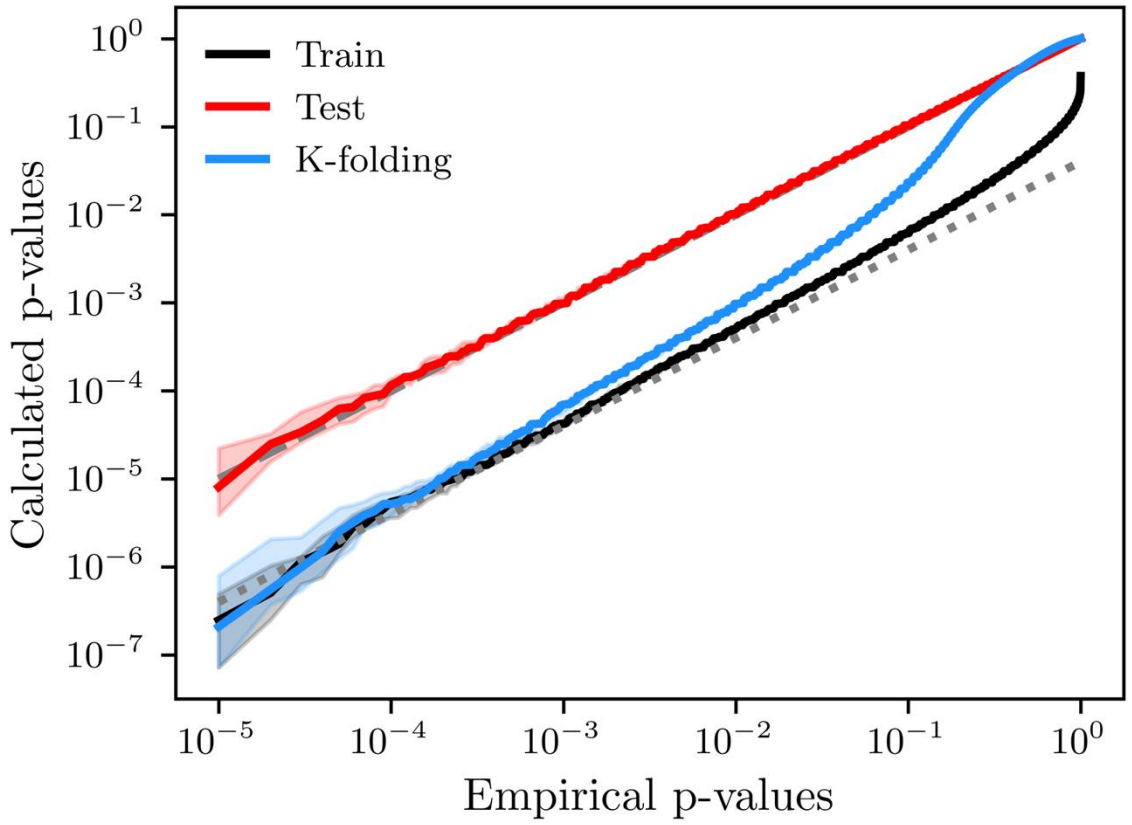
1. Divide data in k folds ($k = 5$)
2. Train model on $k - 1$ folds, evaluate on last fold
3. Repeat until all folds have been used for evaluation
4. Combine results (We sum over counts)





Evaluation using k-fold cross validation

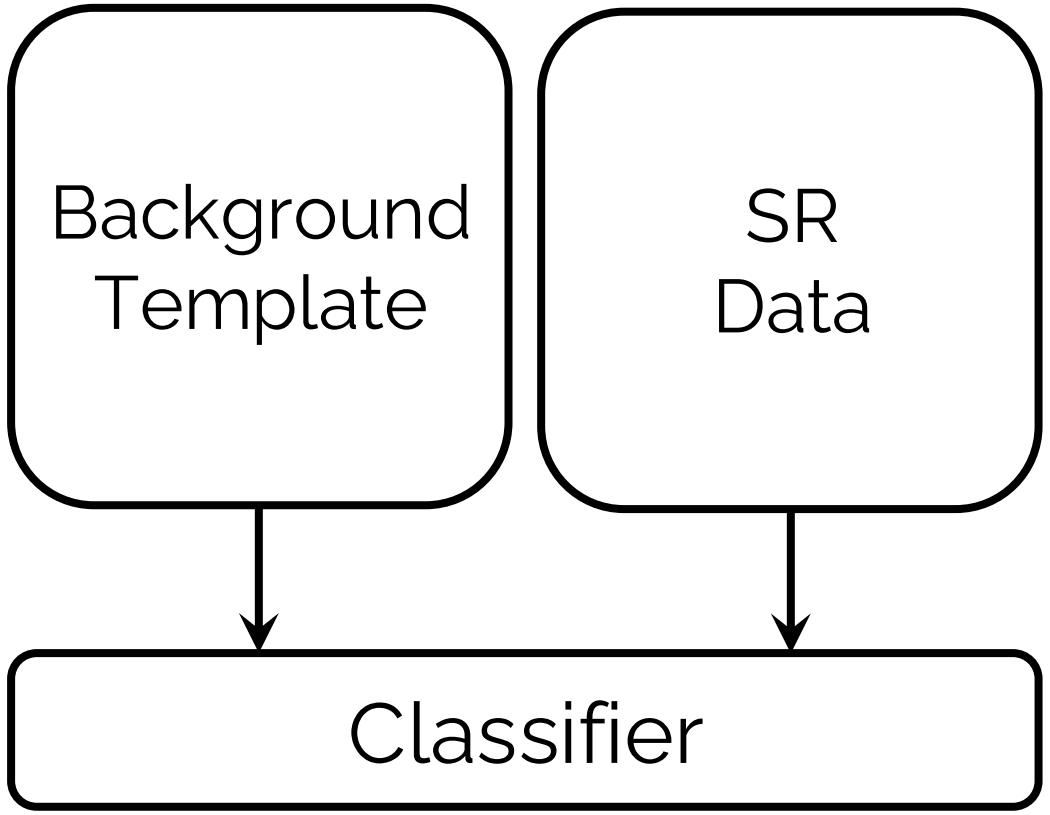




Machine Learning and Look-Everywhere Effects

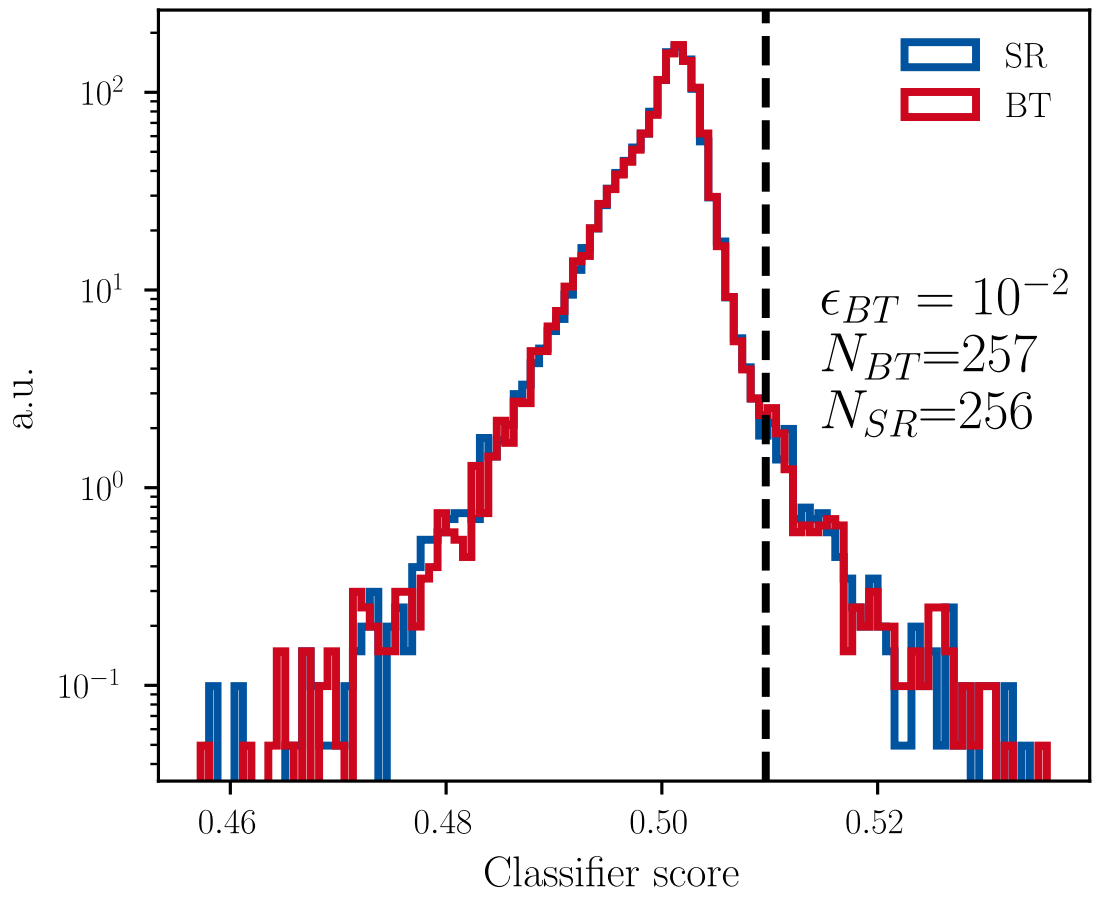
Machine learning procedure

1. Get SR and BT events for both training and evaluation set
2. Train ML classifier on SR versus BT classification on training set

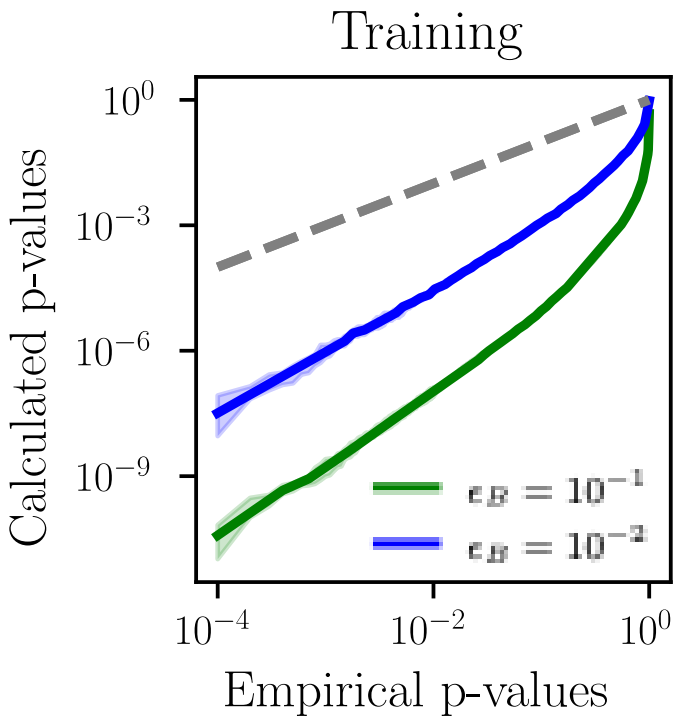


Machine learning procedure

1. Get SR and BT events for both training and evaluation set
2. Train ML classifier on SR versus BT classification on training set
3. Obtain classifier scores for SR and BT events from evaluation set
4. Select cut based on BT to retain fixed fraction ϵ_{BT} of events
5. Apply cut to SR
6. Obtain p-value by comparing event numbers N'_{SR} and N'_{BT} after cut



Results for a simple NN

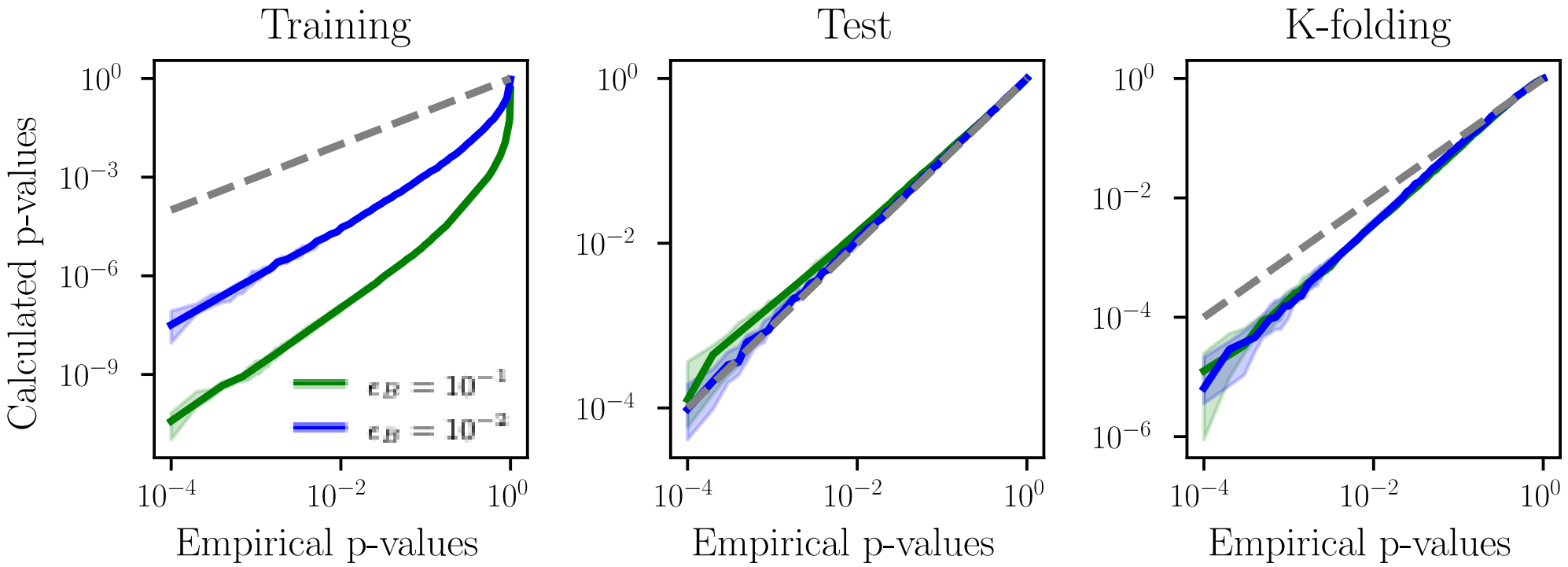


"Simple NN" here means

- 3 hidden layers with 64 nodes
- ReLU activation
- Trained using ADAM with $\mathrm{lr} = 10^{-3}$

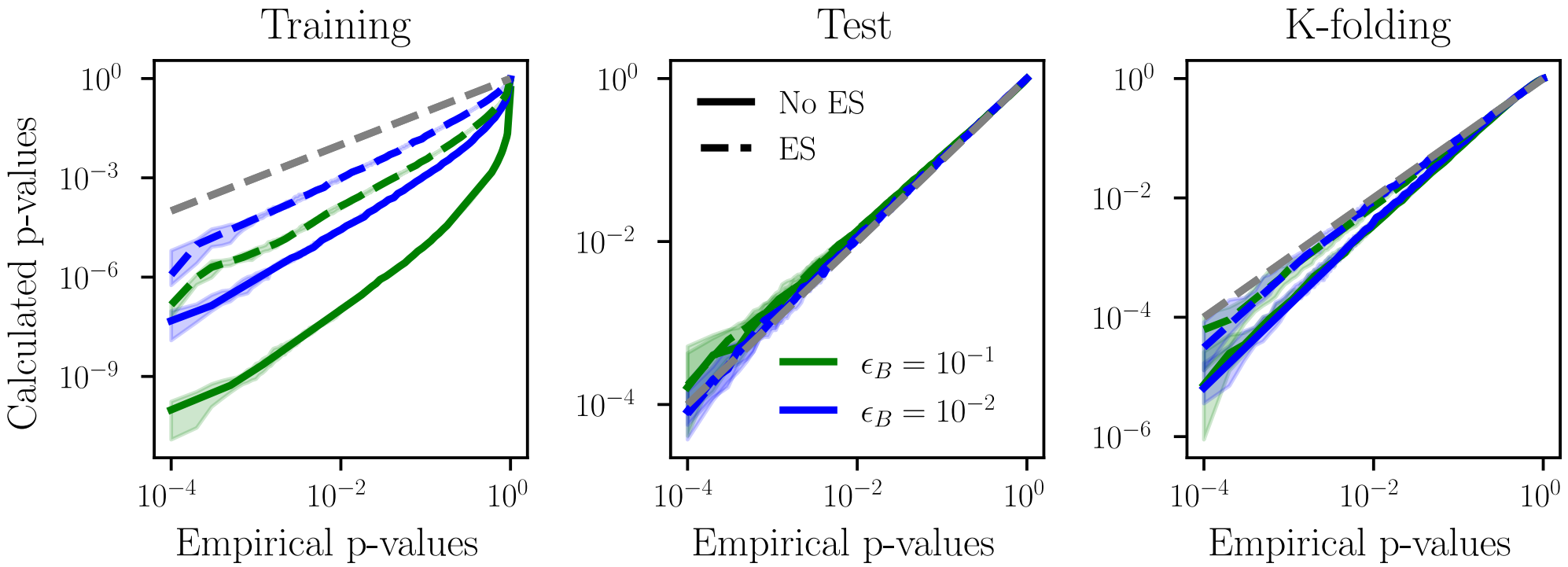
Open question: Why is miscalibration less severe when we k-fold? Wasn't the case for binned analysis

Results for a simple NN



Open question: Why is miscalibration less severe when we k-fold? Wasn't the case for binned analysis

Look-Everywhere Effects in ML terms



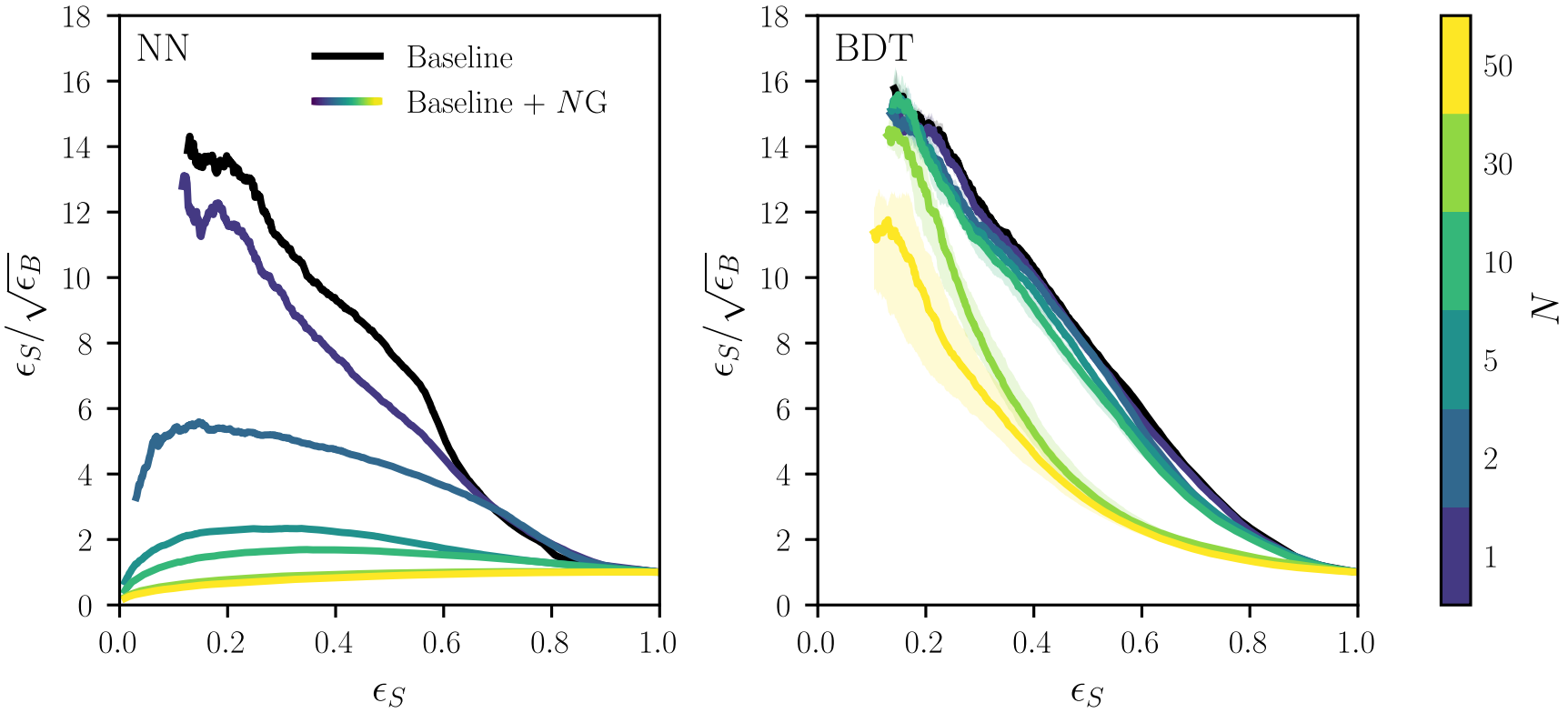
Early stopping reduces miscalibration, eliminates almost entirely for k-folding

What happens if we change the classifier?

Why are we interested in BDTs?

"Tree-based algorithms for weakly supervised anomaly detection" [2309.13111], T. Finke, **MH**, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, M. Sommerhalder

- NNs do not deal well with uninformative input features being added
- By using BDTs instead, we can fix this issue

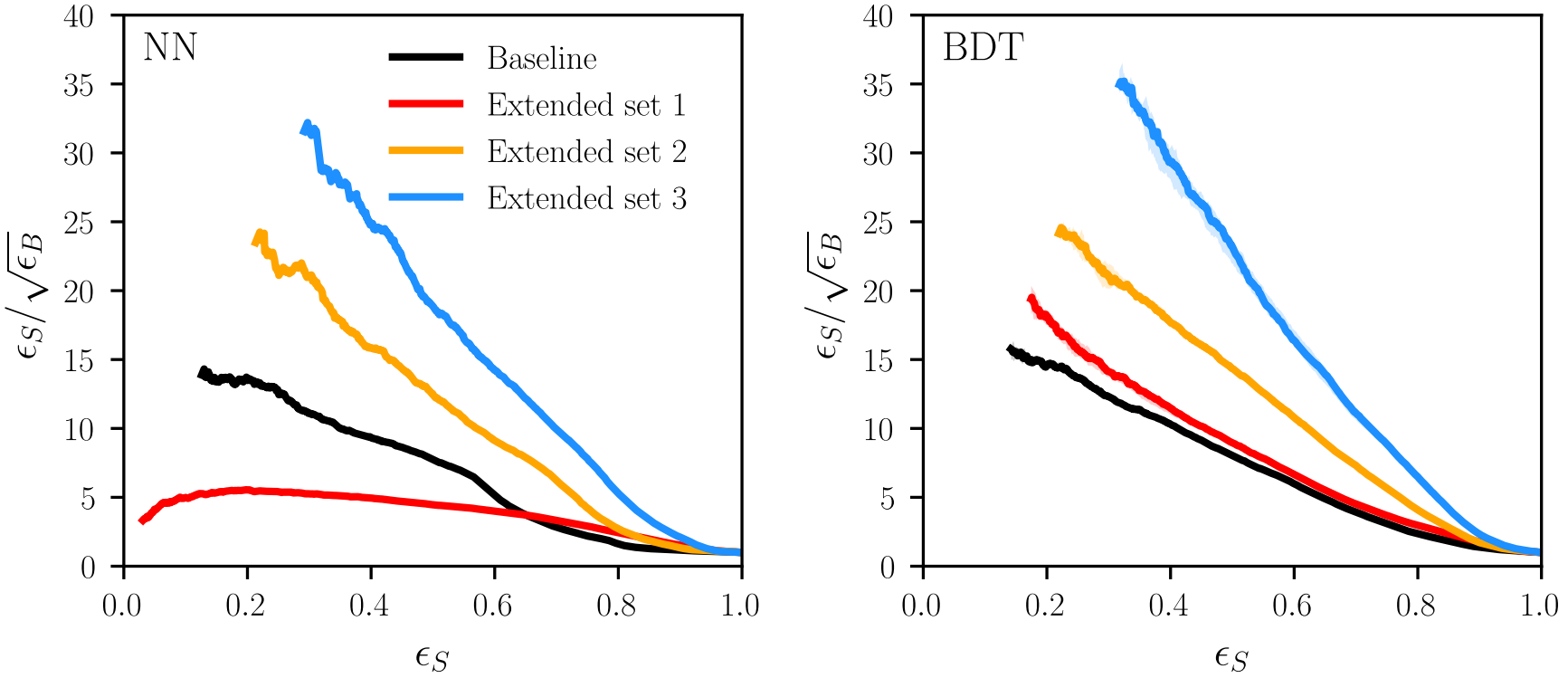


Why are we interested in BDTs?

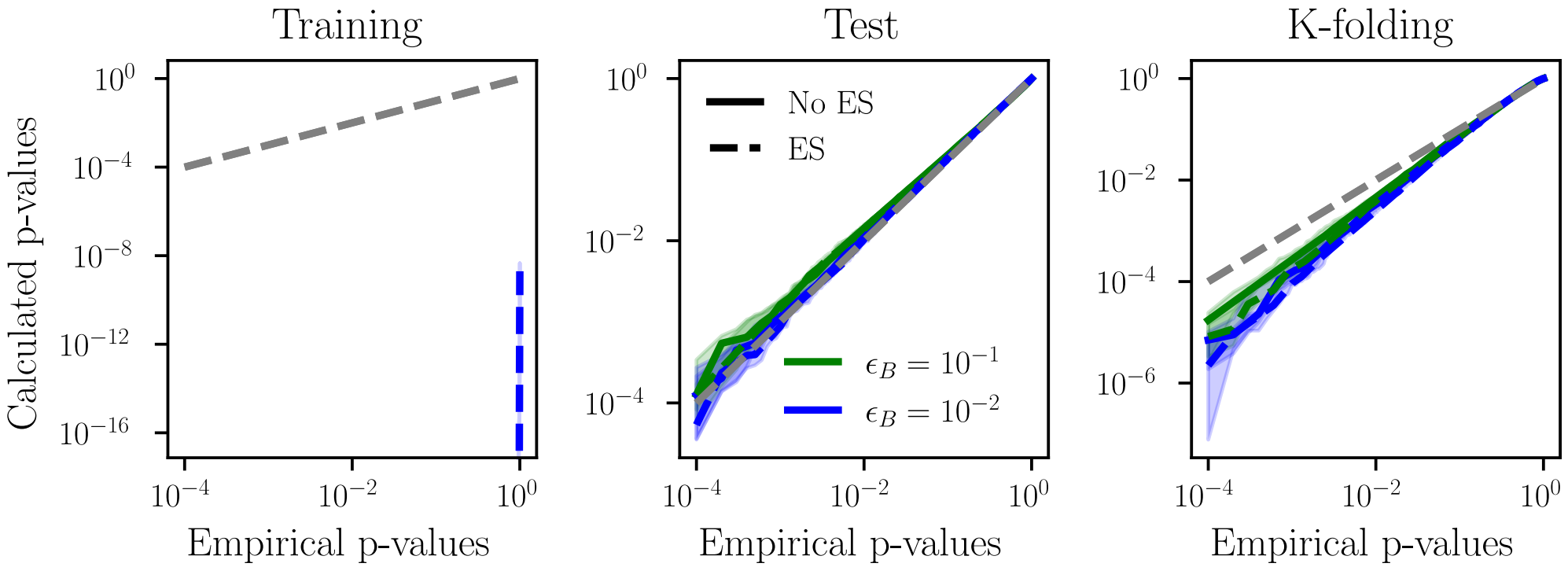
"Tree-based algorithms for weakly supervised anomaly detection" [2309.13111], T. Finke, **MH**, G. Kasieczka, M. Krämer, A. Mück, P. Prangchaikul, T. Quadfasel, D. Shih, M. Sommerhalder

- NNs do not deal well with uninformative input features being added
- By using BDTs instead, we can fix this issue
- Performance gain carries over to realistic feature set

  →Each feature set only adds information



P-value (mis-)calibration for BDTs



Miscalibration worse than for NN and not fixed by early stopping

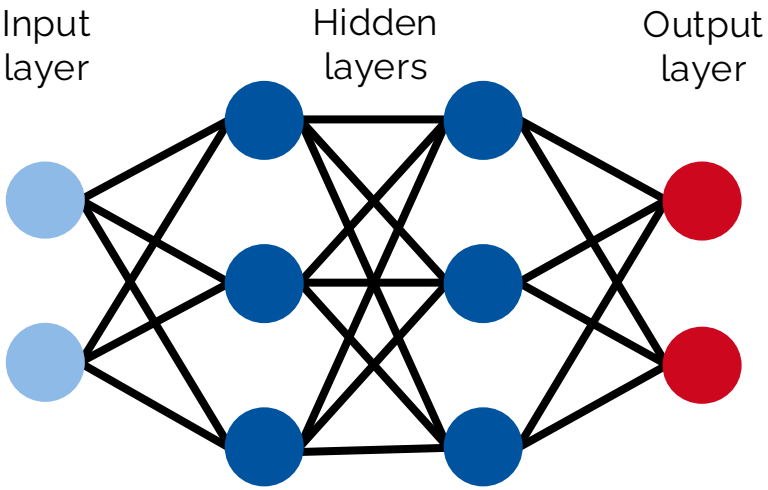
What is different between NNs and BDTs?

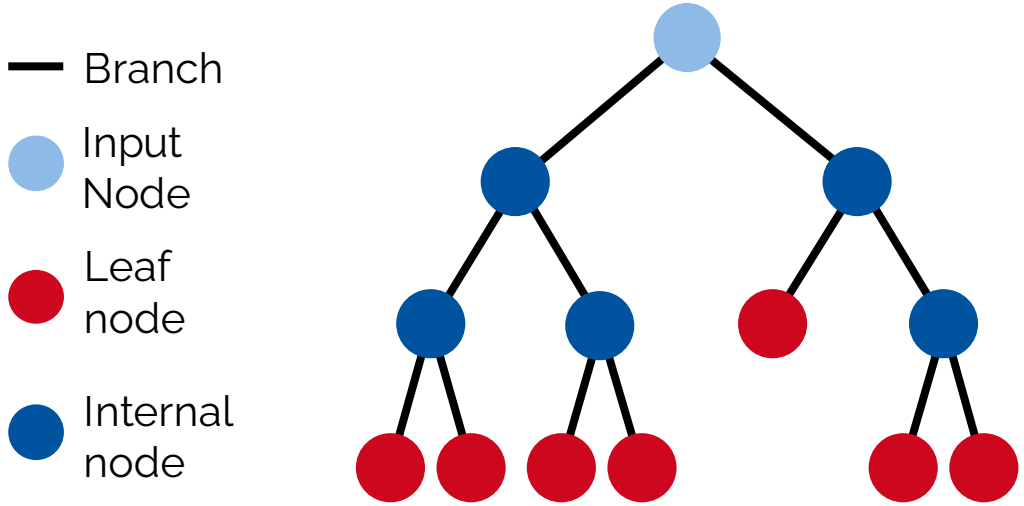
Institute for Theoretical Particle Physics and Cosmology | RWTH AACHEN UNIVERSITY

A number of things but let's consider score calculation:

- NNs are functions, which assign scores based on architecture (fixed), parameters (learned) and input features (given)
- BDTs divide data into smaller groups until they end up in leaf nodes
  - →Scores are based on fraction of each class from training set in a leaf node
  - →A randomly initialized BDT may still overfit

Possible reason for the different behavior under early stopping but not yet definitively shown



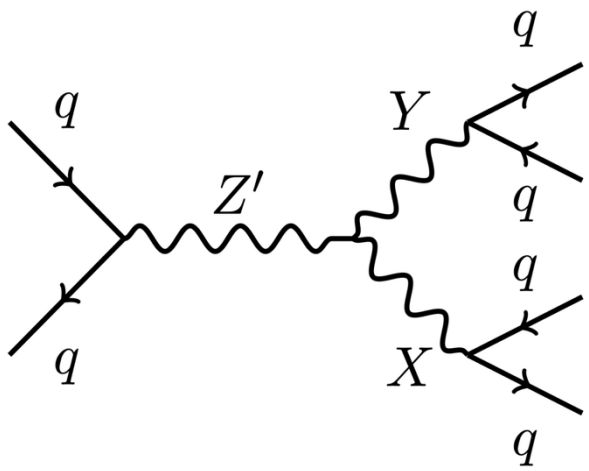


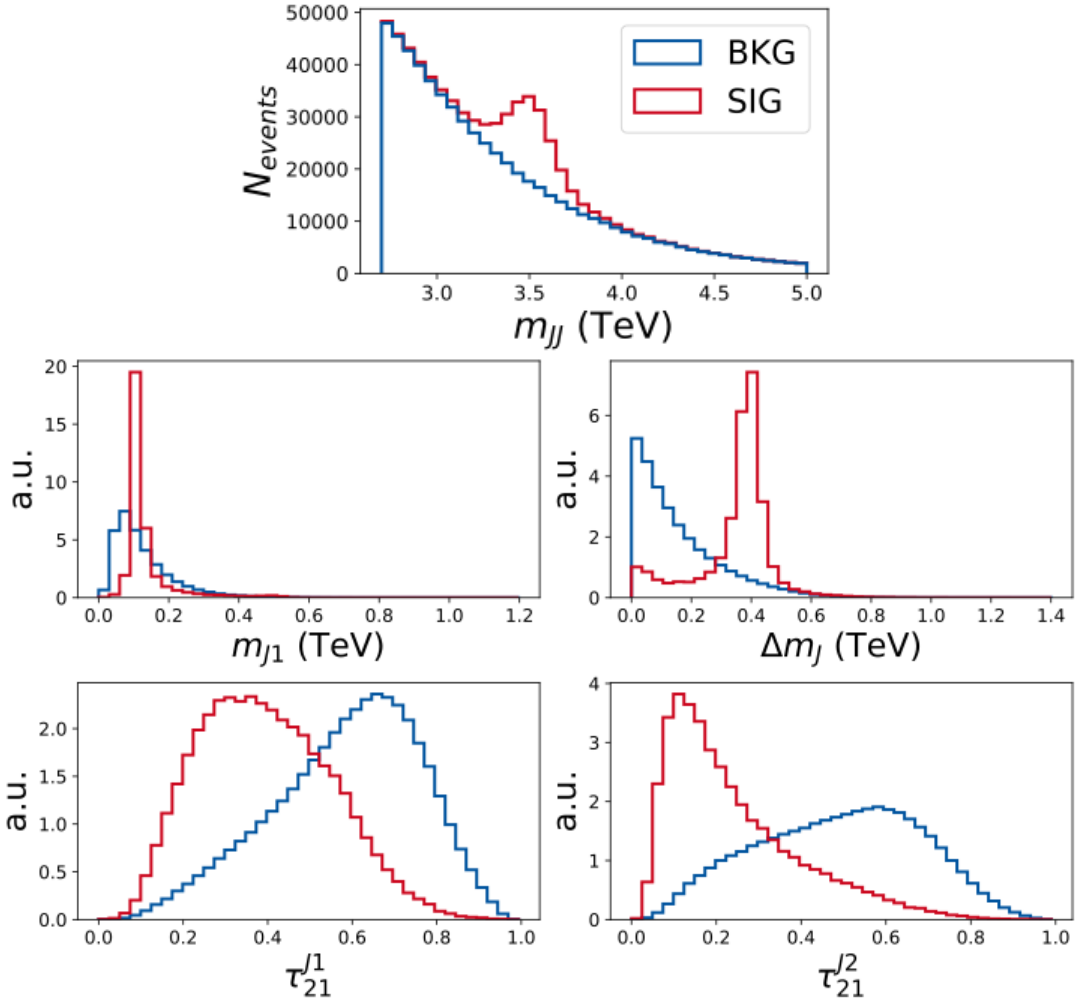
Impacts on Signal Sensitivity

LHCO R&D dataset

"The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics" [2101.08320], G. Kasieczka, B. Nachman, D. Shih et. al.

- Benchmark dataset for anomaly detection
- 1 M QCD dijet background events
- 100k signal events produced via

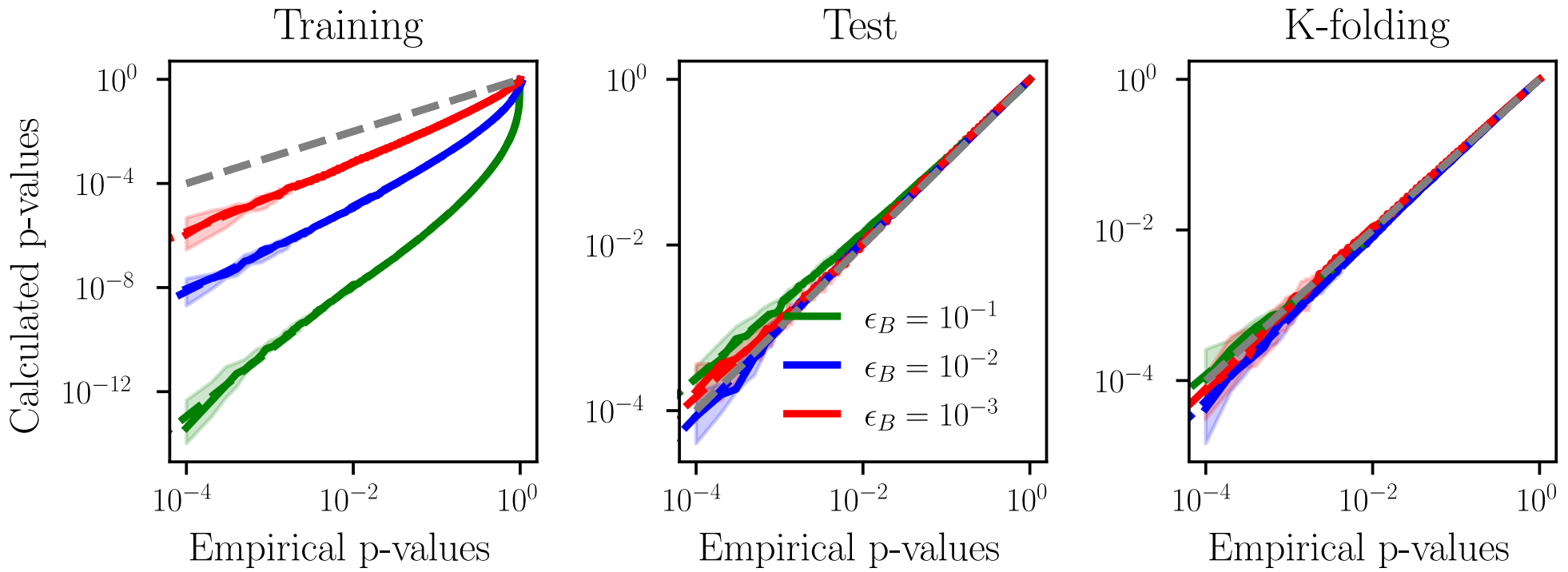




Repeat calibration tests

- Use 10M background events to randomly sample data sets
- Fit calibration curves to obtain trials factor that allow us to calibrate obtained p-values

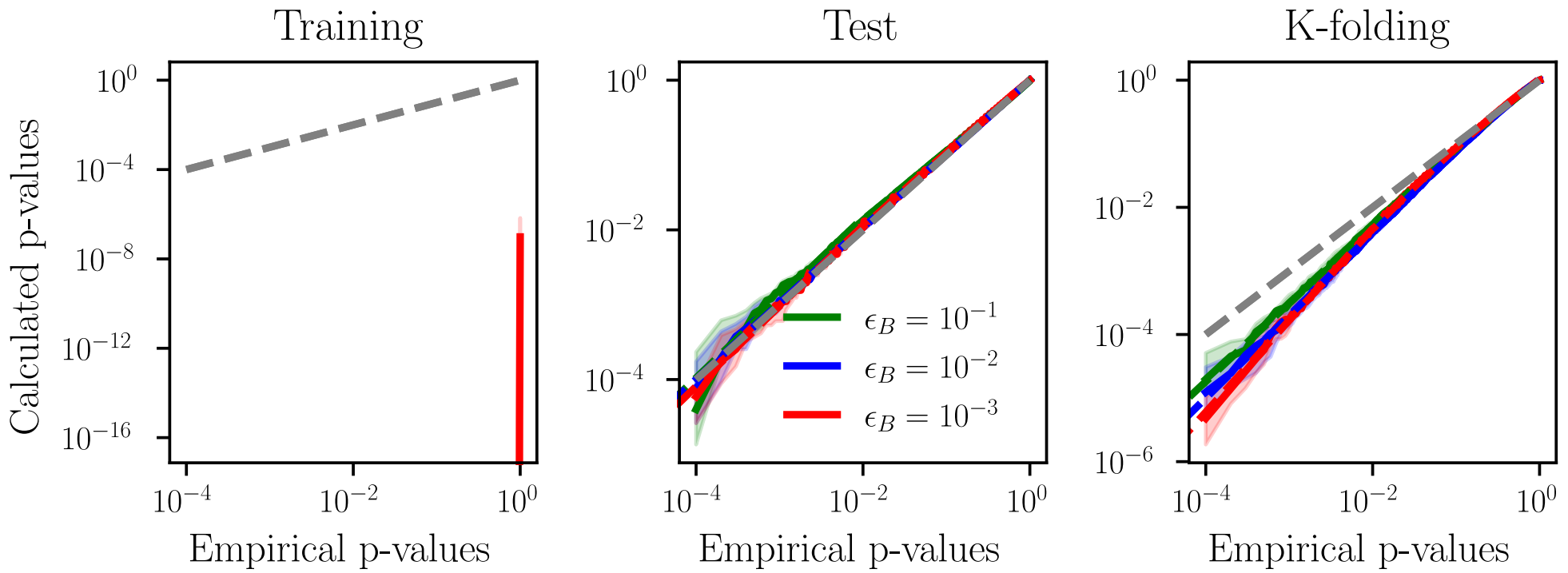
NN with early stopping



Special case!

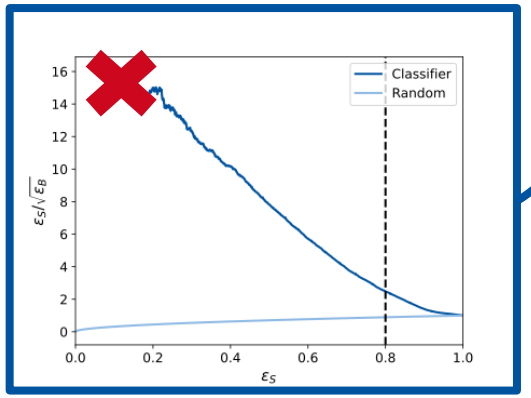
- No fit possible for training set for BDT and NN without early stopping

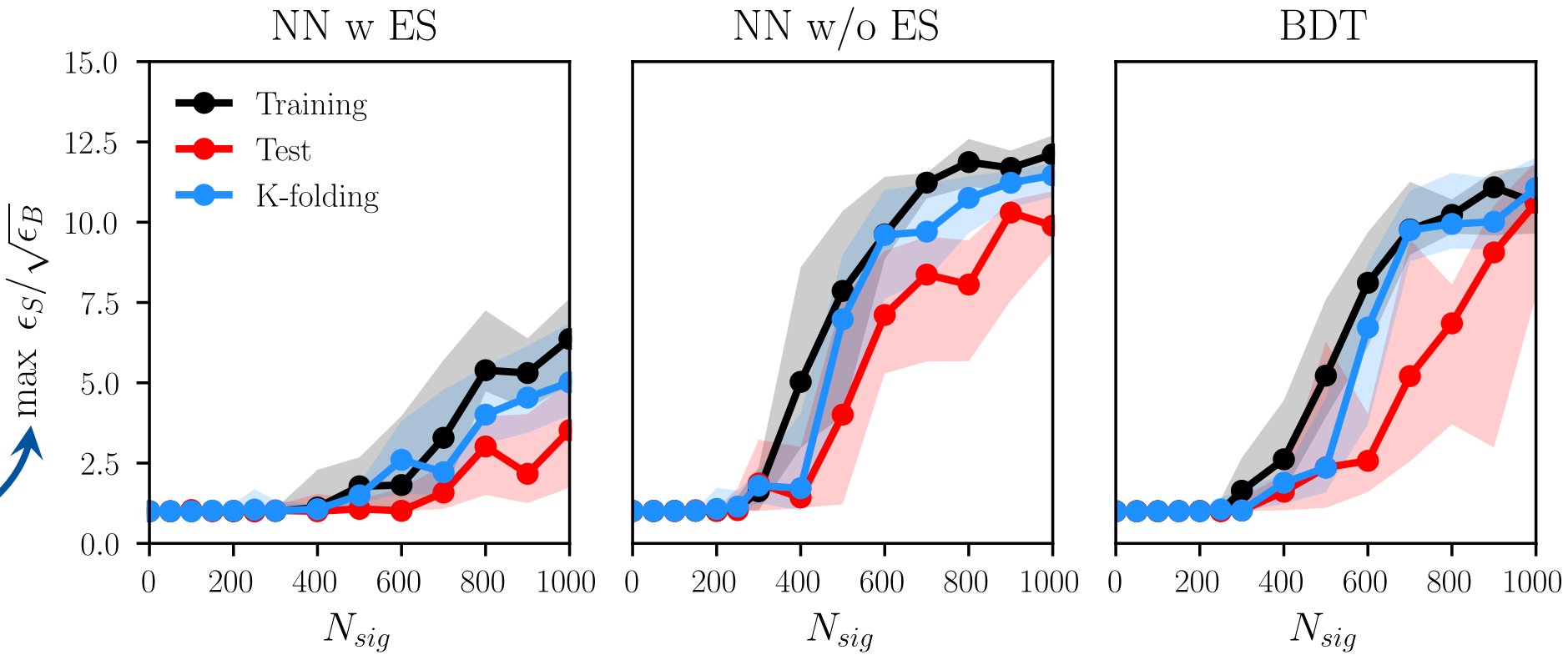
BDT



Study classifier signal sensitivity

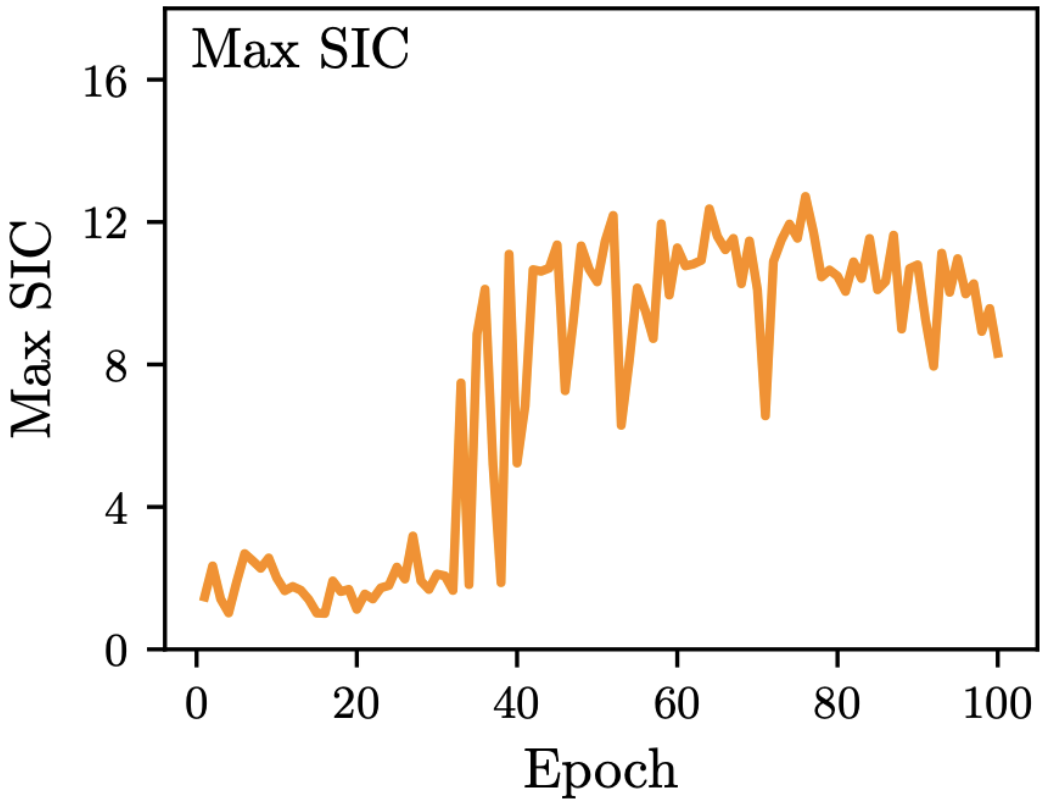
- Performance depends on training set statistics
- NN with early stopping performs very badly

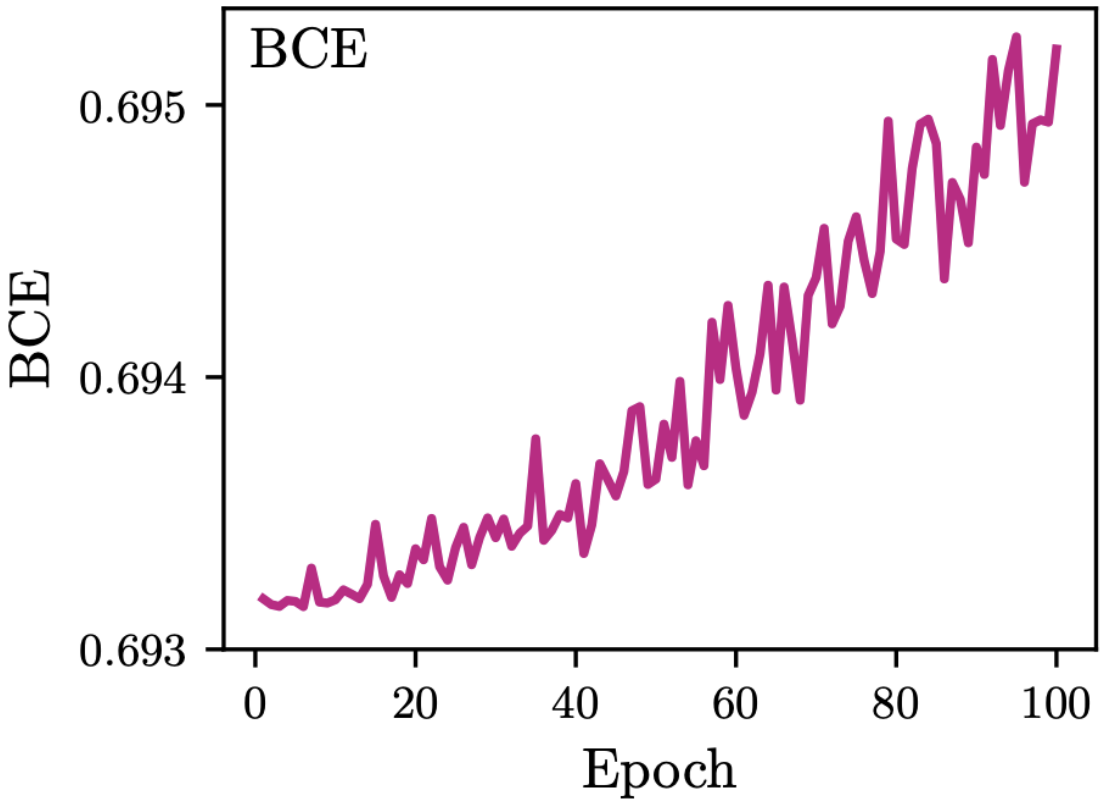




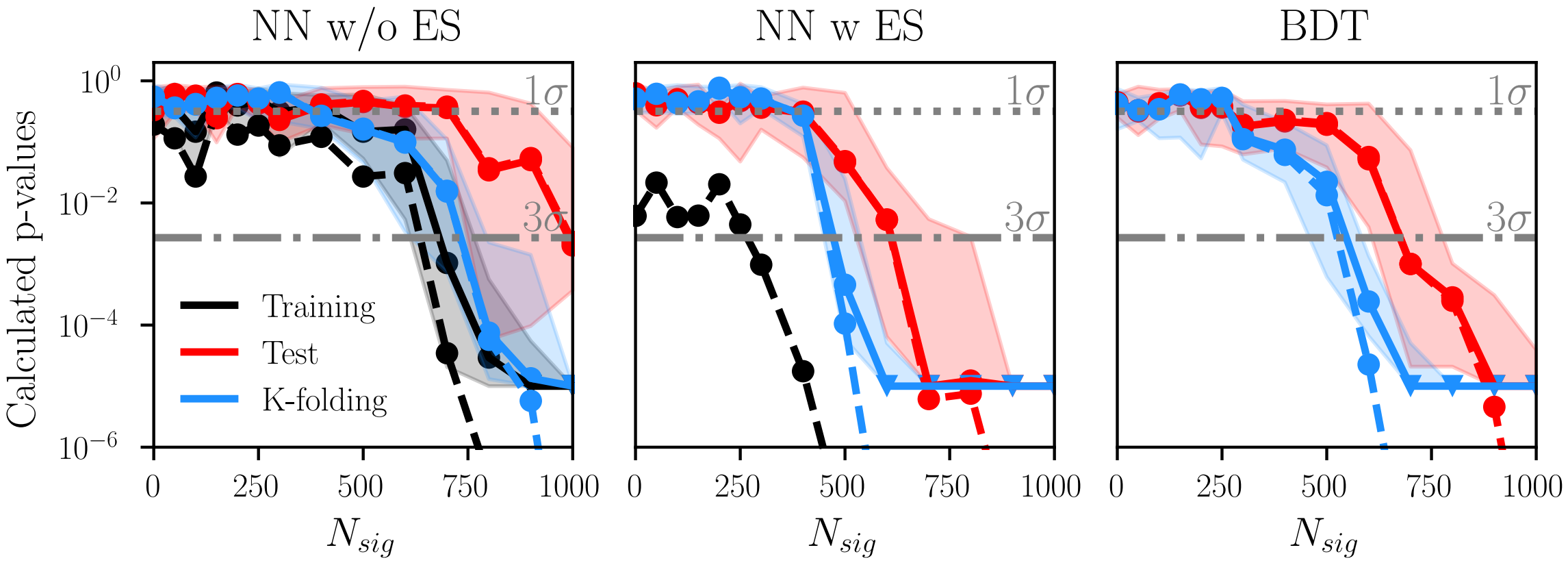
Why does early stopping hurt the sensitivity?

"How to pick the best anomaly detector?" [2511.14832], **MH**, G. Kasieczka, M. Krämer, L. Moureaux, A. Mück, D. Shih





P-values with signal at $\epsilon_B = 10^{-3}$



Conclusion & Outlook

Conclusions

- Data-driven ML-based analyses experience look elsewhere effect-like p-value miscalibrations
- These miscalibrations are directly connected with overfitting
- Evaluating on an independent test set always leads to calibrated p-values
- Evaluating with k-fold cross validation leads to the highest sensitivity but requires p-values calibration to be performed

Outlook

- Many open questions about precise nature of p-value miscalibrations
- How do our findings transfer to other ML applications & statistical treatments
- How do systematic effects play into this

Look Everywhere Effects in Anomaly Detection

MH, B. Nachman, D. Shih

Backup

Excursion: Bayesian statistics

In the background-only case,

$$N'_{\mathrm{SR}}, N'_{\mathrm{BT}} \sim \mathrm{Poisson}(\lambda)$$

with the same expectation value λ. Measuring N'_{BT} can serve as a proxy for λ with

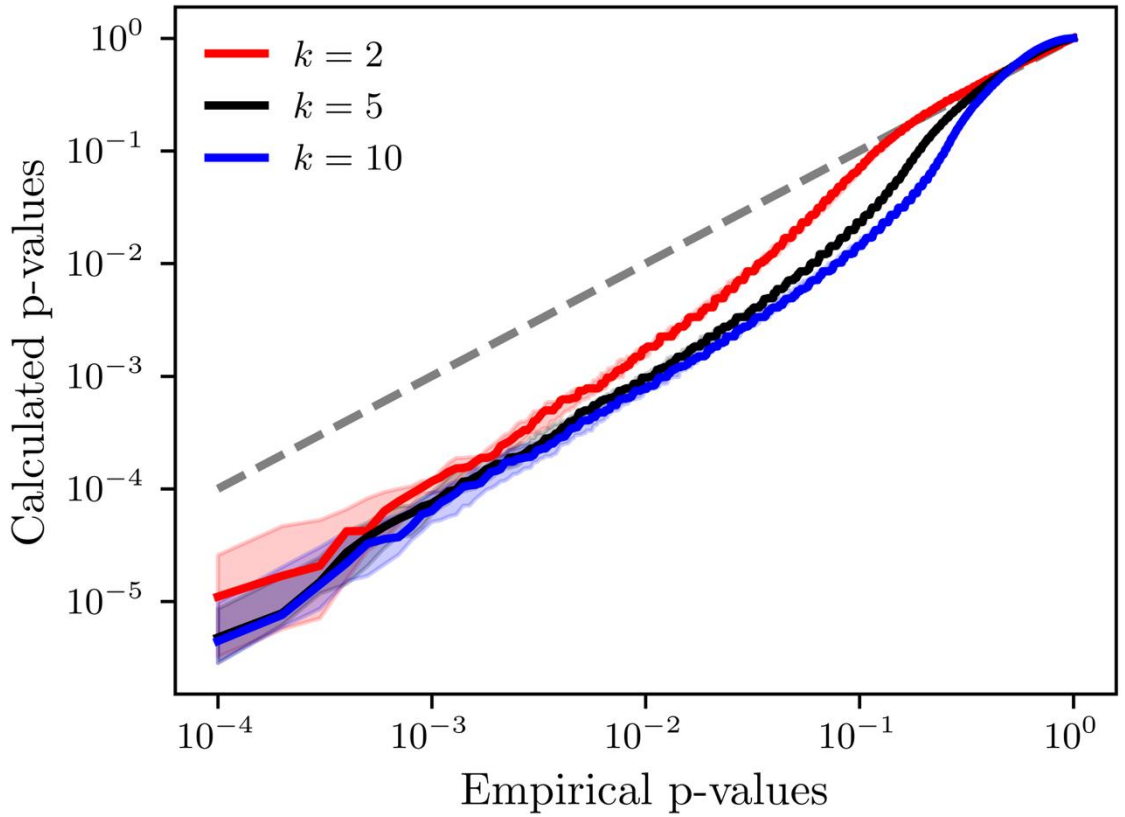
$$\lambda \sim \Gamma(\alpha = N'_{\mathrm{BT}}, \beta = 1),$$

such that

$$\Pr(N'_{\mathrm{SR}}|N'_{BT}) = \int_0^\infty \mathrm{Poisson}\,(N'_{\mathrm{SR}}|\lambda)\ \Gamma(\lambda|N'_{\mathrm{BT}}, 1)\ \mathrm{d}\lambda.$$

This integral resolves to a negative binomial distribution.

Other values for k: Binned



Other values for k: NN