Research Training Group Physics of the Heaviest Particles at the LHC

RWTH AACHEN UNIVERSITY

Collaborative Research Center TRR 257

P3H

Particle Physics Phenomenology after the Higgs Discovery

TTK Institute for Theoretical Particle Physics and Cosmology

RWTH AACHEN UNIVERSITY

Back to the Roots: Tree-Based Algorithms for Weakly Supervised Anomaly Detection

Marie Hein

with Finke, Kasieczka, Krämer, Mück, Prangchaikul, Quadfasel, Shih, Sommerhalder

CRC Young Scientists Meeting, October 18, 2023

Agenda

1. **Anomaly Detection Methods**
2. **The Problem**
3. **The Solution**
4. **The Physics Gain**

Motivation

- BSM physics searches are well motivated
- Classic search approaches
  - ➔ Very sensitive searches for specific new physics models
  - ➔ Less sensitive signal model agnostic searches, e.g. resonance searches
- Our goal: Improve sensitivity of model agnostic searches
  - ➔ Reason for lacking sensitivity: often only performed in one variable
  - ➔ Use pattern recognition capability of machine learning in high dimensional feature space to gain higher sensitivity

Anomaly Detection Methods

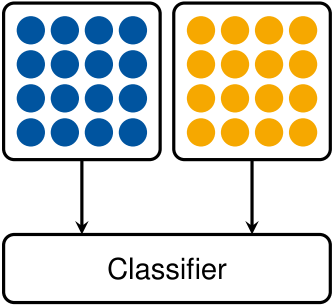
Classification Problem

- Goal: To achieve a better signal to background ratio
- An optimal classifier is given by the likelihood ratio

$$R_{\text{optimal}}(x) = \frac{p_S(x)}{p_B(x)}, \tag{1}$$

  where p_S and p_B are the signal and background densities, respectively.

  - → Can be approximated with a supervised classifier
  - → Problem: Labels are not available on experimental data



Weakly Supervised Classification

Institute for Theoretical Particle Physics and Cosmology
RWTH AACHEN UNIVERSITY

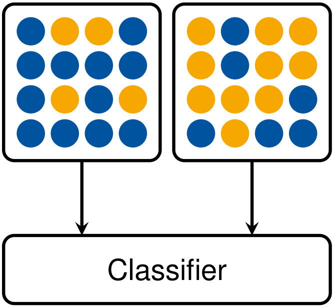
- Any monotonic function of a classifier has the same decision boundaries
- Use two mixed datasets with

$$p_i(x) = f_i \, p_S(x) + (1 - f_i) \, p_B(x) \tag{2}$$

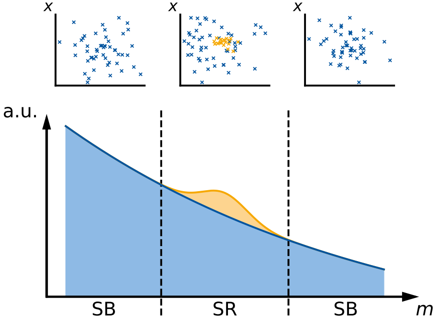
- Classifier gives likelihood ratio

$$R_{\text{mixed}} = \frac{f_1 \, R_{\text{optimal}}(x) + (1 - f_1)}{f_2 \, R_{\text{optimal}}(x) + (1 - f_2)}. \tag{3}$$

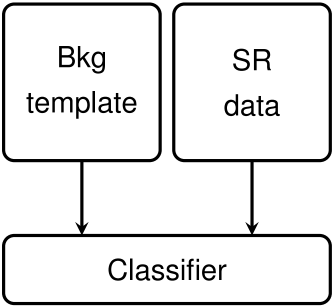
  → Monotonically increasing function of $R_{\text{optimal}}(x)$ as long as $f_1 > f_2$
  → Weakly supervised classifier/ CWoLa [Methodiev, Nachman, Thaler, '17]



How can weak supervision be applied to real data?

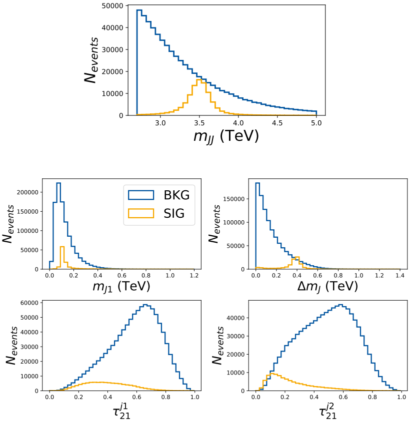


Recreated from [Hallin et al., '21]



The Problem

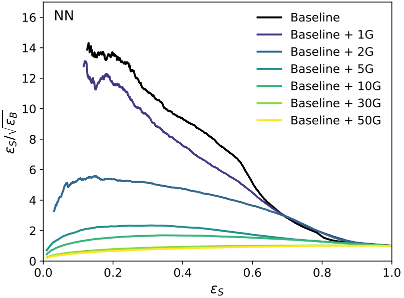
LHC Olympics R&D dataset



- Benchmark dataset for anomaly detection
- QCD dijet background
- Resonant signal of W' $\rightarrow$ XY with X/Y $\rightarrow$ qq
- $m_{W'} = 3.5$ TeV, $m_X = 0.5$ TeV, $m_Y = 0.1$ TeV
- Baseline features used for the classification
  - ➔ Resonant feature m_{JJ}
  - ➔ m_{J1}, Δm_J, $\tau_{21,J1}$, $\tau_{21,J2}$
- SR: 0.4 TeV bin around $m_{W'}$
- Inject 1000 signal events into dataset

The Problem

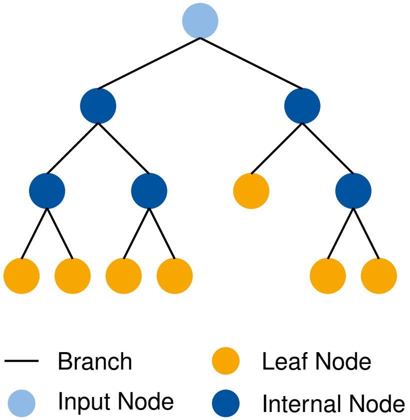
- Model agnostic setup includes uninformative features
  - ➔ Need robustness against uninformative features
- Simulate using N Gaussian distributed features
- Significant performance drop observed already with $N = 2$



The Solution

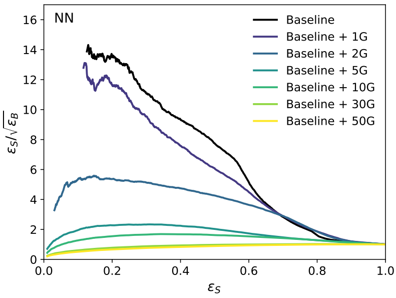
Decision Trees

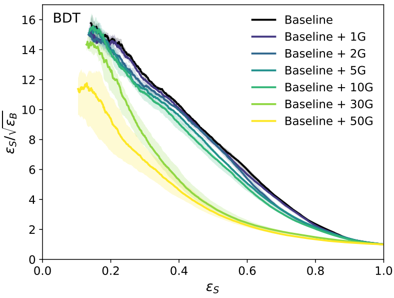
- Classical machine learning method
- Data is split recursively based on a set of input features
- To create a new node, both the feature and the split values are optimized
- For additional expressivity, ensembles of trees are used
  - → Gradient boosting: learn residuals of previous predictions with subsequent trees
- Deal well with tabular data, which our high-level features are



Robustness against uninformative features

- BDT is much more robust against uninformative features
- Performance stable up to 10 Gaussian features





The Physics Gain

Feature sets

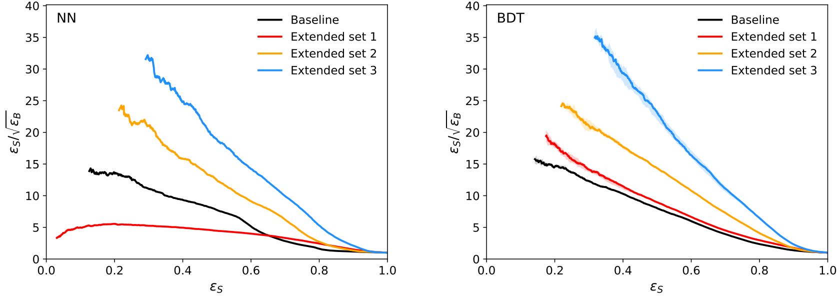
Institute for Theoretical Particle Physics and Cosmology
RWTH AACHEN UNIVERSITY

- As sensitivity reaches higher number of features, we can include more physics features in an analysis
- Test by including additional subjettiness based features

Name	# features	Features
Baseline	4	$\{m_{J_1},\ \Delta m_J,\ \tau_{21}^{\beta=1,J_1},\ \tau_{21}^{\beta=1,J_2}\}$
Extended 1	10	$\{m_{J_1},\ \Delta m_J,\ \tau_{N,N-1}^{\beta=1,J_1},\ \tau_{N,N-1}^{\beta=1,J_2}\}$ for $2 \leq N \leq 5$
Extended 2	12	$\{m_{J_1},\ \Delta m_J,\ \tau_{N}^{\beta=1,J_1},\ \tau_{N}^{\beta=1,J_2}\}$ for $N \leq 5$
Extended 3	56	$\{m_{J_1},\ \Delta m_J,\ \tau_{N}^{\beta,J_1},\ \tau_{N}^{\beta,J_2}\}$ for $N \leq 9$ and $\beta \in \{0.5, 1, 2\}$

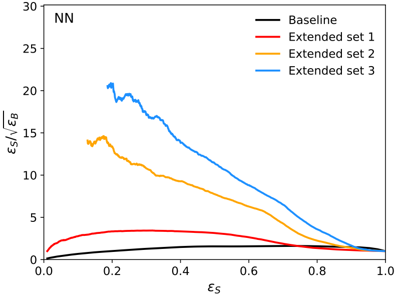
Results for different feature sets

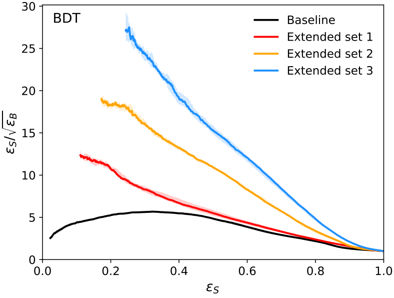
- BDT well behaved with respect to information content of input feature set
- Not true for NN



Results for different signal

- Being able to use more features increases the sensitivity to other signal models
- Test this by considering resonant signal of W' $\rightarrow$ XY with X/Y $\rightarrow$ qqq





Conclusion

Summary

- BDTs are robust against uninformative features in the weakly supervised setup
- BDTs are well behaved with respect to the information content of an input set
  - ➔ Ability to use larger input feature sets in an analysis
- Larger input feature sets allow for more model agnosticity

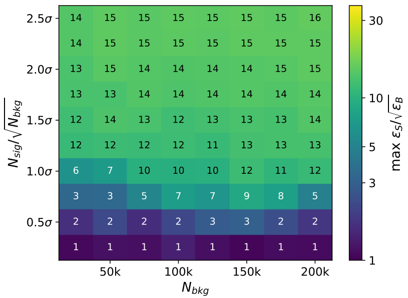
Outlook

- Apply the improved classifier to methods defining the background template from data
- Test method on different signal models

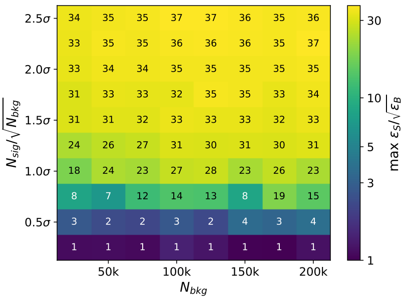
Backup slides

2D scan

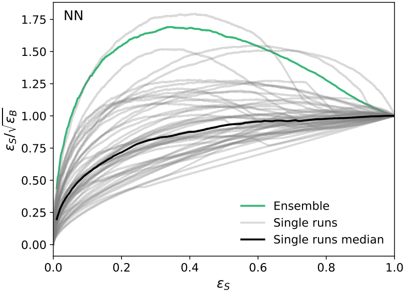
Baseline



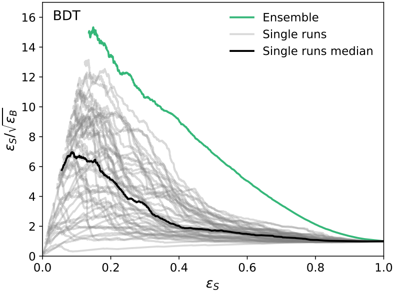
Extended set 3



Ensembling

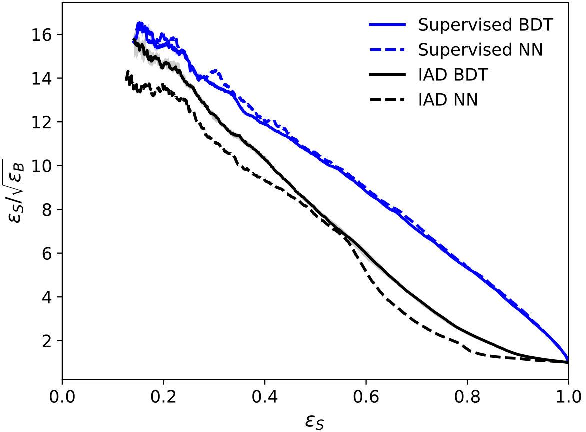


NN
Ensemble
Single runs
Single runs median
$\epsilon_S/\sqrt{\epsilon_B}$
ϵ_S

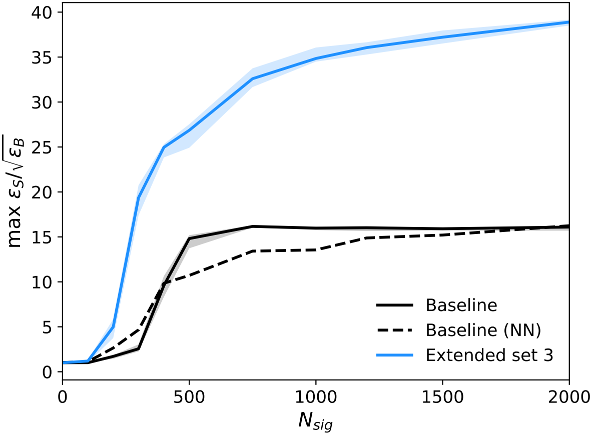


BDT
Ensemble
Single runs
Single runs median
$\epsilon_S/\sqrt{\epsilon_B}$
ϵ_S

Baseline performances



1D scan



max $\epsilon_S/\sqrt{\epsilon_B}$
N_{sig}
0
5
10
15
20
25
30
35
40
0
500
1000
1500
2000
Baseline
Baseline (NN)
Extended set 3

Marie Hein – marie.hein@rwth-aachen.de

RWTH Aachen University
Templergraben 55
52056 Aachen

www.rwth-aachen.de