Research Training Group Physics of the Heaviest Particles at the LHC

RWTH AACHEN UNIVERSITY

Collaborative Research Center TRR 257

P H

Particle Physics Phenomenology after the Higgs Discovery

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

ML4Jets, November 8, 2023

Introduction & Setup

Motivation

Institute for Theoretical Particle Physics and Cosmology

RWTH AACHEN UNIVERSITY

- 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

Weakly Supervised Classification

- Optimal classifier

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

with $p_{S/B}$ signal and background densities.

- Classifier of mixed datasets

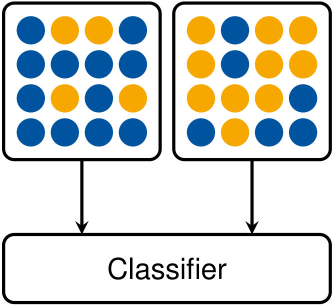
$$p_i(x) = f_i \, p_S(x) + (1 - f_i) \, p_B(x) \tag{2}$$

gives likelihood ratio

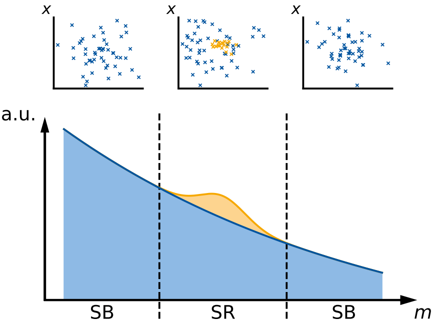
$$R_{\text{mixed}} = \frac{f_1 \, R_{\text{optimal}}(x) + (1 - f_1)}{f_2 \, R_{\text{optimal}}(x) + (1 - f_2)}. \tag{3}$$

  → Monotonically increasing function of $R_{\text{optimal}}(x)$ as long as $f_1 > f_2$.

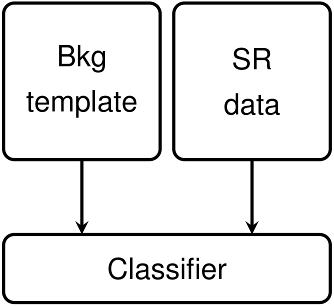
  → Weakly supervised classifier / CWoLa [1708.02949]



How can weak supervision be applied to real data?



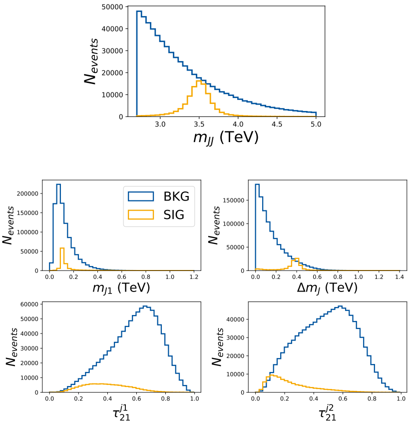
Recreated from [2109.00546]



Background template obtained through [1902.02634, 2001.05001, 2109.00546, 2203.09470, 2212.11285, ...]

LHC Olympics R&D dataset [2101.08320]

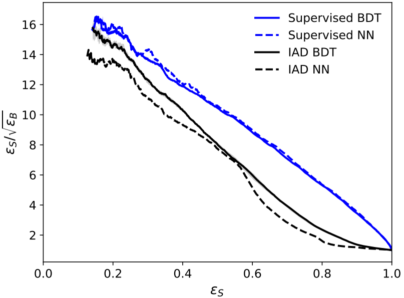
Institute for Theoretical Particle Physics and Cosmology
RWTH AACHEN UNIVERSITY



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

ML setup and baseline performance

- NN: Fully connected NN with 3 hidden layers of 64 nodes, trained using Adam with learning rate 10^{-3}
- BDT: Histogrammed Gradient Boosted Decision Trees
- For both ensemble of 50 independently trained models with randomized training-validation split of 50% used
- BDT shows median and 68% error band of 10 runs, NN just one run



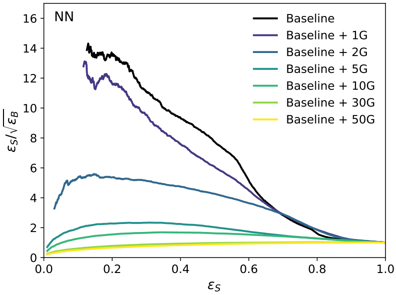
Increasing the feature set size

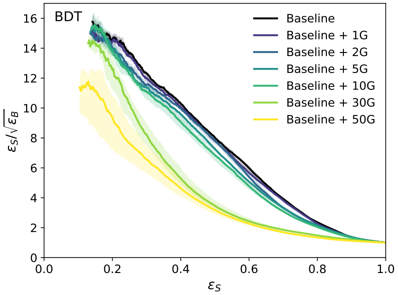
Increasing the feature set size

- Current baseline with 4 features is not model agnostic
- Ideally, want to move to low level features but neither classification nor density estimation are easy in high dimensional space (but getting closer, see [2310.06897])
- Therefore, let's first focus on more high-level features:
  - ➔ Here, BDTs are a natural choice
- In model agnostic setup, many features will not be informative for any particular signal model
  - ➔ Need to be robust against uninformative features

Uninformative features

- Simulate uninformative features by adding N Gaussian distributed noise features to baseline feature set
- NN performance drops significantly already with $N = 2$
- BDT performance remains stable up to 10 Gaussian features





Physically motivated feature sets

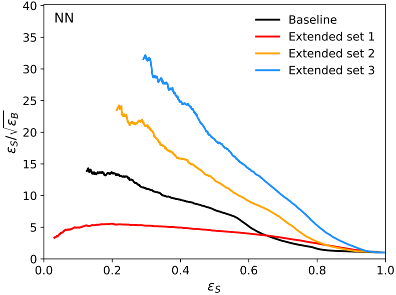
Institute for Theoretical Particle Physics and Cosmology

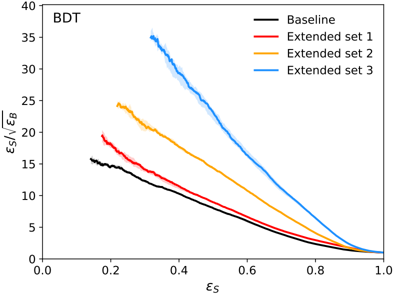
- As sensitivity reaches higher number of features, we can include more physics features in an analysis
- Test by including additional subjettiness based features
  - → Information content increases towards bottom of table
  - → Higher subjettiness ratios essentially uninformative (extended set 1)
  - → Subjettinesses all slightly informative (extended sets 2 & 3)

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

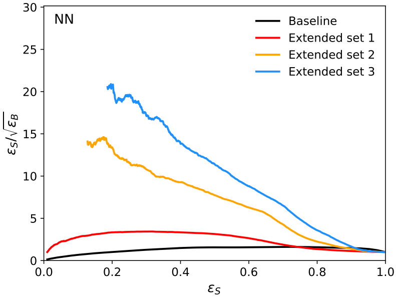
- BDT is well behaved with respect to information content of input feature set
- NN's sensitivity to uninformative features leads to large performance drop for extended set 1

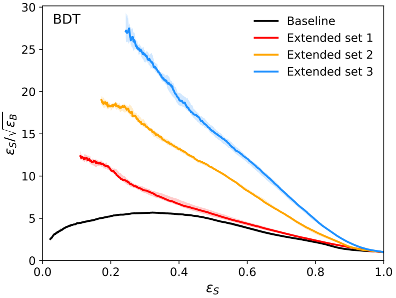




Results for a different signal model

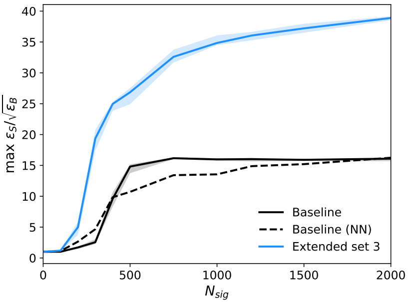
- Being able to use more features increases the sensitivity to other signal models
- Test this by considering resonant signal of Z' $\rightarrow$ XY with X/Y $\rightarrow$ qqq





Signal number change

- Sensitivity to low signal strengths important for effectiveness of analysis
- On baseline set similar results observed for both NN and BDT
- Sensitivity of extended set 3 extends to lower signal injections



Conclusion

Summary

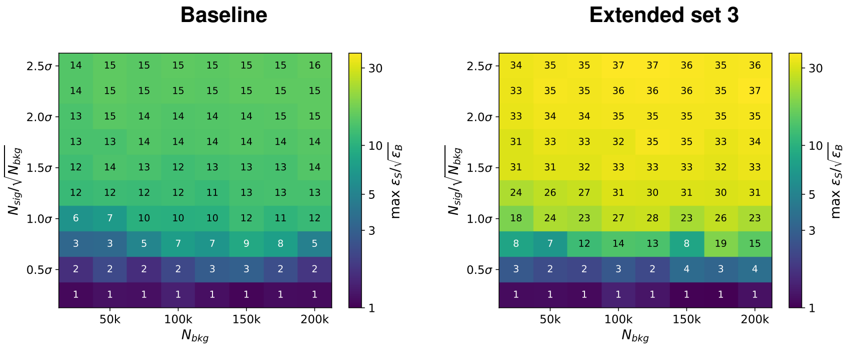
- BDTs are robust against uninformative features in the weakly supervised setup
- BDTs are well behaved with respect to the information content of an input set
  - ➔ Ability to use larger input feature sets in an analysis
- Larger input feature sets allow for more model agnosticity

Outlook

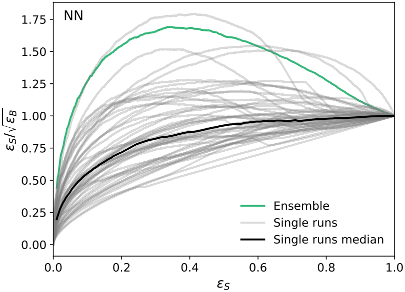
- Apply BDT classifier to methods defining the background template from data

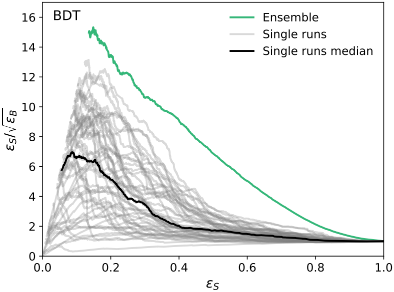
Backup slides

2D scan

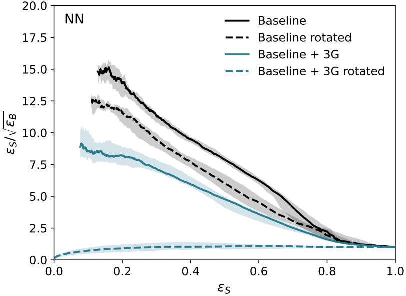

Baseline
Extended set 3
$N_{sig}/\sqrt{N_{bkg}}$
N_{bkg}
max $\epsilon_S/\sqrt{\epsilon_B}$
2.5σ 2.0σ 1.5σ 1.0σ 0.5σ
50k 100k 150k 200k
1 3 5 10 30
14 15 15 15 15 15 15 16
14 15 15 15 15 15 15 15
13 15 14 14 14 14 15 15
13 13 14 14 14 14 14 14
12 14 13 12 13 13 13 14
12 12 12 12 11 13 13 13
6 7 10 10 10 12 11 12
3 3 5 7 7 9 8 5
2 2 2 2 3 3 2 2
1 1 1 1 1 1 1 1
34 35 35 37 37 36 35 36
33 35 35 36 36 36 35 37
33 34 34 35 35 35 35 35
31 33 33 32 35 35 33 34
31 31 32 33 33 33 32 33
24 26 27 31 30 31 30 31
18 24 23 27 28 23 26 23
8 7 12 14 13 8 19 15
3 2 2 3 2 4 3 4
1 1 1 1 1 1 1 1

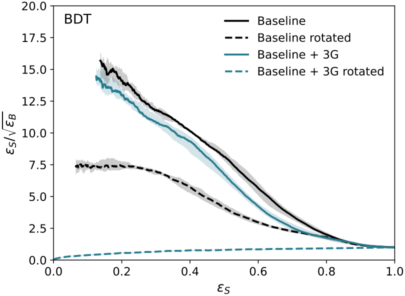
Ensembling


NN
Ensemble
Single runs
Single runs median
$\epsilon_S/\sqrt{\epsilon_B}$
ϵ_S
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0.0
0.2
0.4
0.6
0.8
1.0

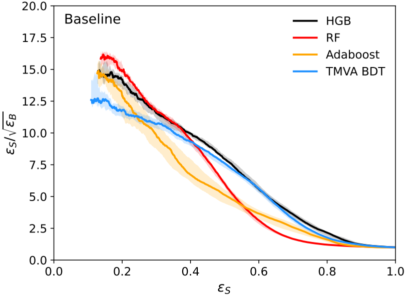

BDT
Ensemble
Single runs
Single runs median
$\epsilon_S/\sqrt{\epsilon_B}$
ϵ_S
0
2
4
6
8
10
12
14
16
0.0
0.2
0.4
0.6
0.8
1.0

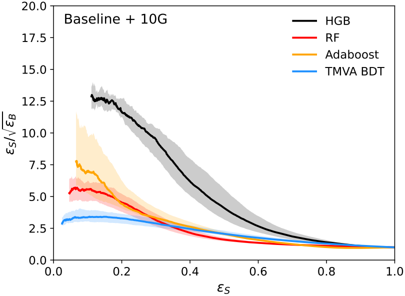
Rotational invariance





Model choice





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

RWTH Aachen University
Templergraben 55
52056 Aachen

www.rwth-aachen.de