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

Weakly Supervised Anomaly Detection for Resonance Searches

Marie Hein

UCLouvain Seminar, May 7, 2024

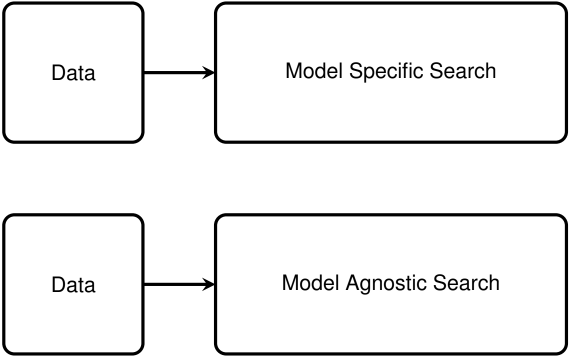
Why Anomaly Detection?

Analysis Methods

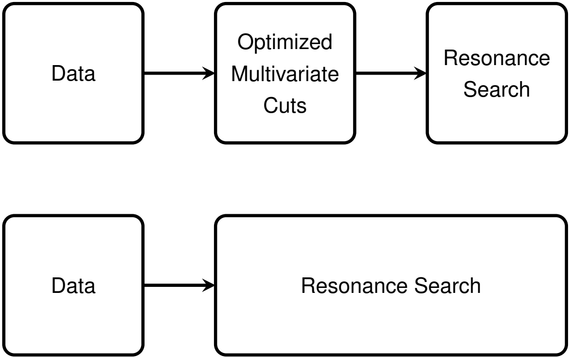
Model Specific Search

Model Agnostic Search

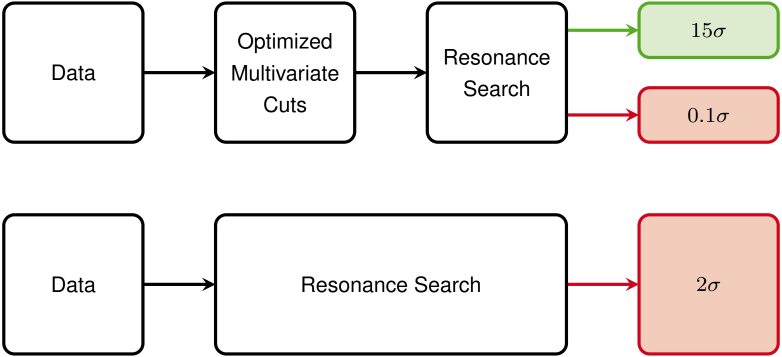
Analysis Methods



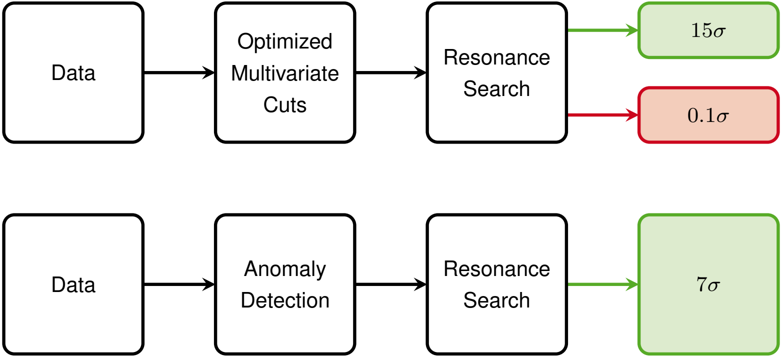
Analysis Methods



Analysis Methods



Anomaly Detection

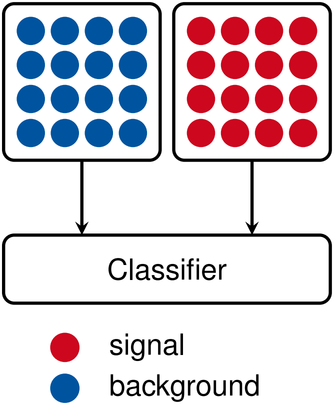


Weakly Supervised Anomaly Detection

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

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

  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

“Classification without labels: Learning from mixed samples in high energy physics” [1708.02949], E. Metodiev, B. Nachman, J. Thaler

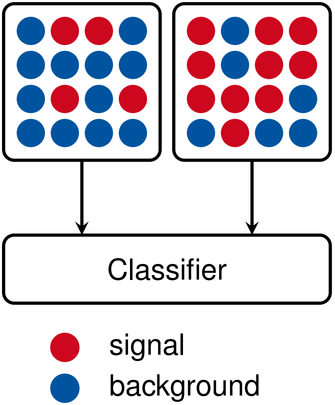
- Any monotonic function of a classifier has the same decision boundaries
- Two mixed datasets with signal fractions f_i

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

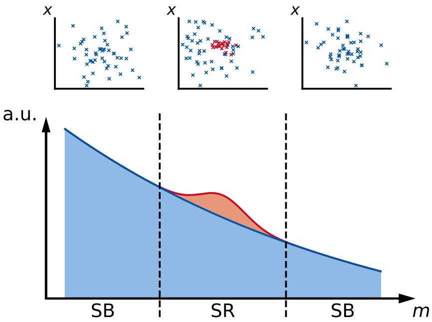
- Classifier gives likelihood ratio

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

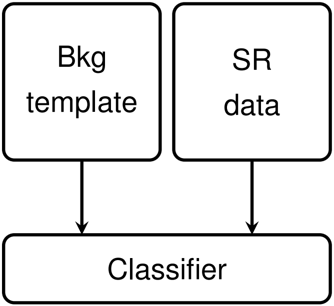
  - → Monotonically increasing function of $R_{\text{optimal}}(x)$ as long as $f_1 > f_2$
  - → Weakly supervised classifier/ CWoLa

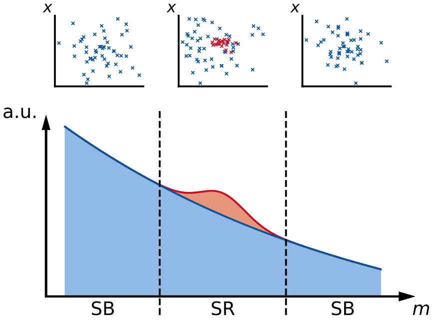


Application to resonance searches

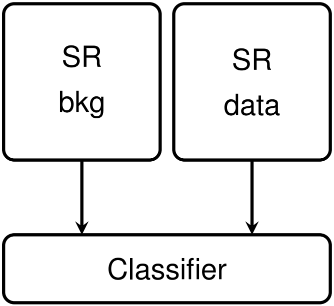


Recreated from [2109.00546]

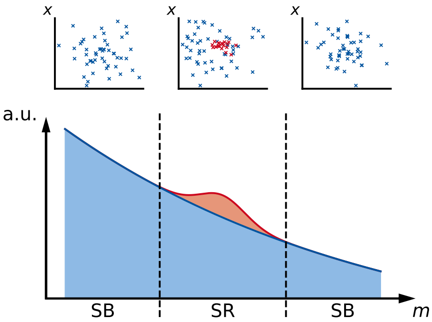




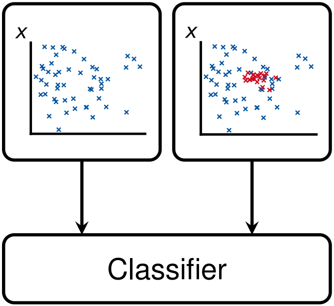
Recreated from [2109.00546]



Application to resonance searches

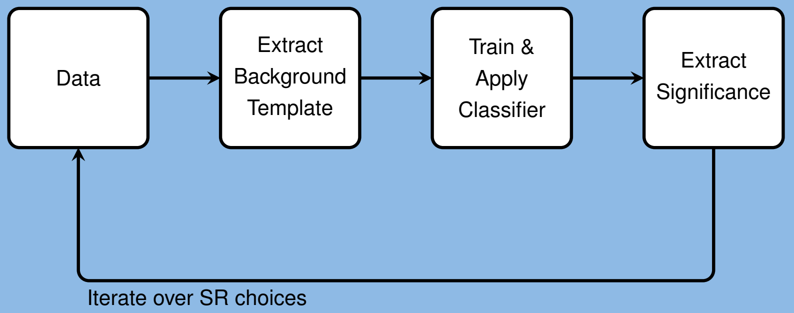


Recreated from [2109.00546]



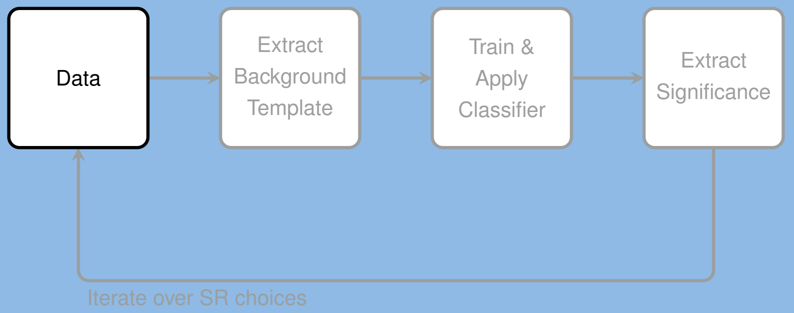
How to use Weakly Supervised Anomaly Detection

Full analysis chain



Weakly Supervised Analysis Chain

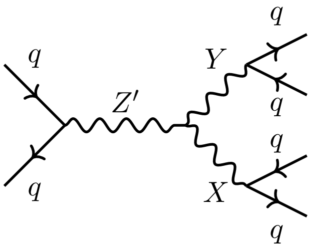
Full analysis chain



LHC Olympics 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.

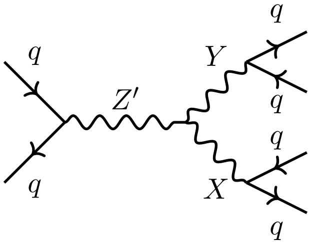
- Benchmark dataset for anomaly detection
- QCD dijet background
- Signal

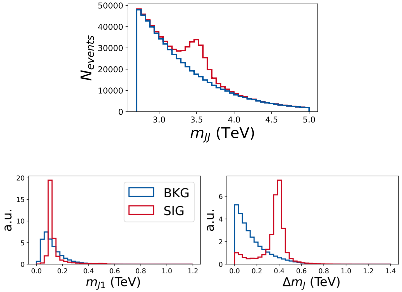


LHC Olympics R&D dataset

“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
- QCD dijet background
- Signal





Institute for Theoretical Particle Physics and Cosmology RWTH AACHEN UNIVERSITY

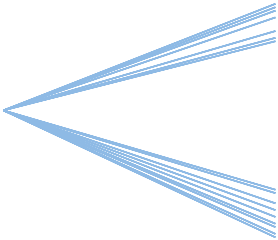
"Identifying Boosted Objects with N-subjettiness" [1011.2268], J. Thaler, K. Van Tilburg
"Maximizing Boosted Top Identification by Minimizing N-subjettiness" [1108.2701], J. Thaler, K. Van Tilburg

- Cluster jets into N subjets to obtain

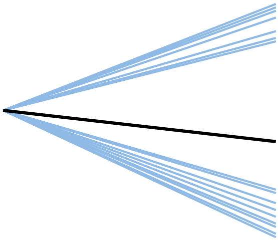
$$\tau_N^\beta = \frac{1}{d_0} \sum_i p_{T,i} \min_J (\Delta R_{Ji})^\beta \,, \tag{4}$$

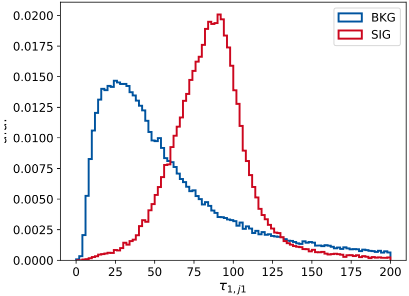
  - where J runs over all N subjet candidates,
  - $\Delta R_{Ji} = \sqrt{(\Delta y_{Ji})^2 + (\Delta \phi_{Ji})^2}$ is an angular distance measure, and
  - $d_0 = \sum_i p_{T,i} R_0^\beta$ a normalization factor.

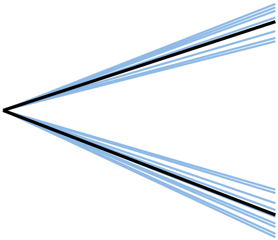
Subjettiness: Example

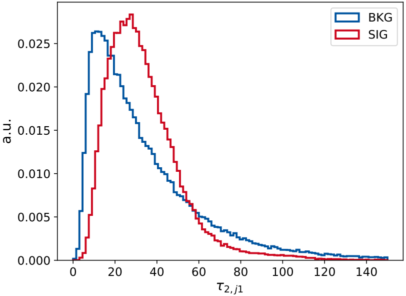


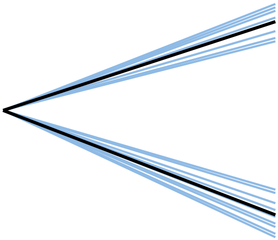
Subjettiness: Example

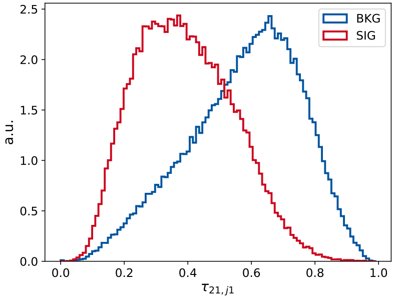







BKG
SIG
0.000
0.005
0.010
0.015
0.020
0.025
a.u.
0
20
40
60
80
100
120
140
$\tau_{2,j1}$



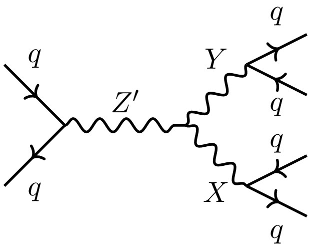


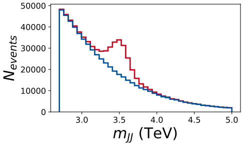
2.5
2.0
1.5
1.0
0.5
0.0
a.u.
0.0
0.2
0.4
0.6
0.8
1.0
$\tau_{21,j1}$
BKG
SIG

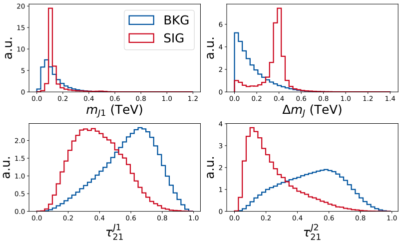
LHC Olympics R&D dataset

"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
- QCD dijet background
- Signal



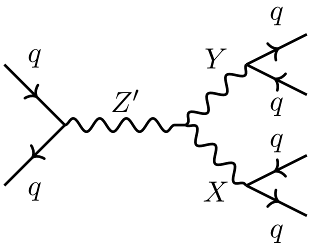


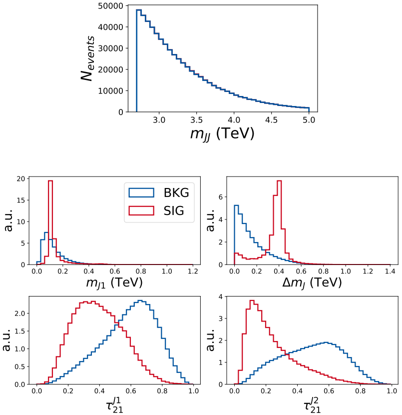


LHC Olympics R&D dataset

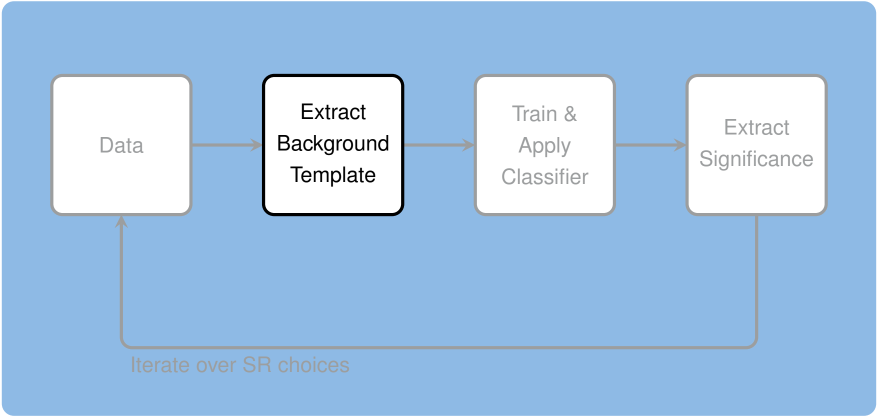
"The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics" [2101.08320], G. Kasieczka, B. Nachman, D. Shih et. al.

- Benchmark dataset for anomaly detection
- QCD dijet background
- Signal

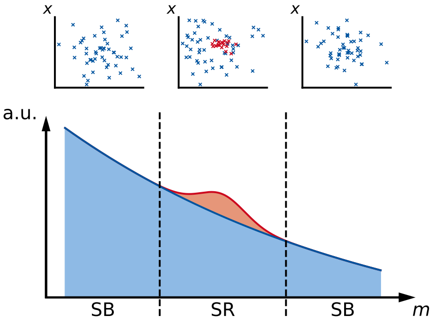




Full analysis chain



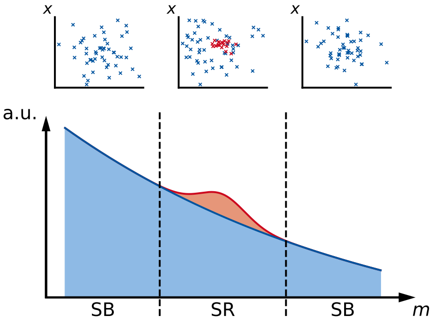
"Extending the Bump Hunt with Machine Learning" [1902.02634], J. Collins, K. Howe, B. Nachman



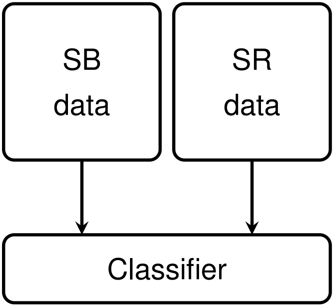
Recreated from [2109.00546]

- General assumption: Bump hunt, i.e. signal is localized in m
  - ➔ $p_B(x|m \in SB) = p_{\text{data}}(x|m \in SB)$
- CWoLa Hunting-specific assumption: Distribution of the background in x is independent of m
  - ➔ $p_B(x \in SB) = p_B(x \in SR)$
  - ➔ $p_B(x|m) = p_B(x)$

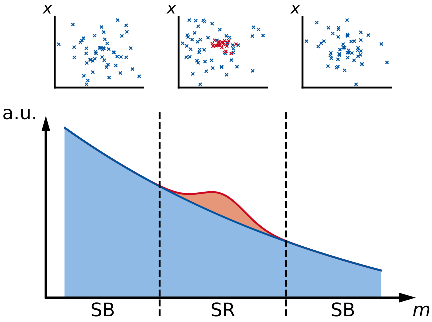
“Extending the Bump Hunt with Machine Learning” [1902.02634], J. Collins, K. Howe, B. Nachman



Recreated from [2109.00546]



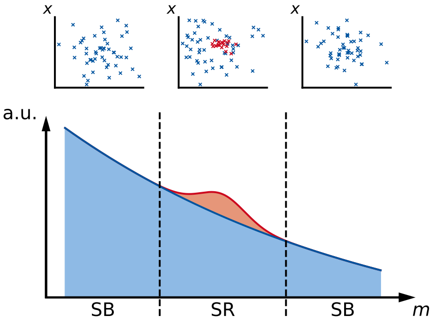
“Classifying Anomalies Through Outer Density Estimation” [2109.00546], A. Hallin et. al.



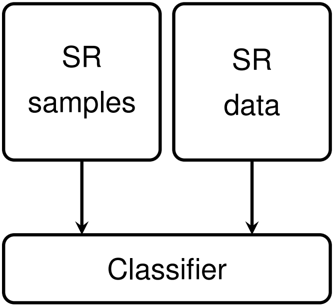
Recreated from [2109.00546]

- General assumption: Bump hunt, i.e. signal is localized in m
  - ➔ $p_B(x|m \in SB) = p_{\text{data}}(x|m \in SB)$
- CATHODE-specific assumption: Distribution of the background in x is smooth in m
  - ➔ Train a conditional density estimator on the SB to learn $p_B(x|m)$
  - ➔ Interpolate into the SR to sample there

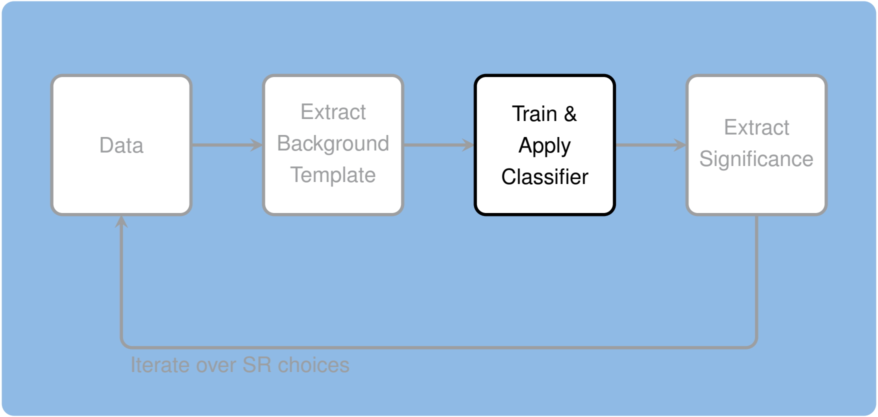
"Classifying Anomalies Through Outer Density Estimation" [2109.00546], A. Hallin et. al.



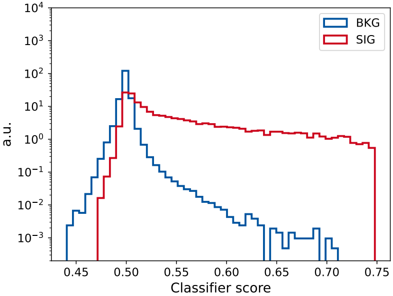
Recreated from [2109.00546]



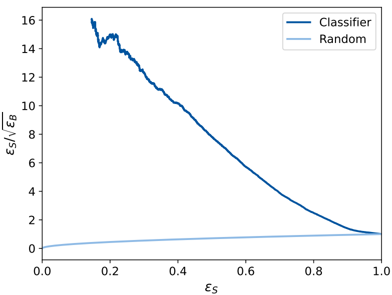
Full analysis chain



SIC curves

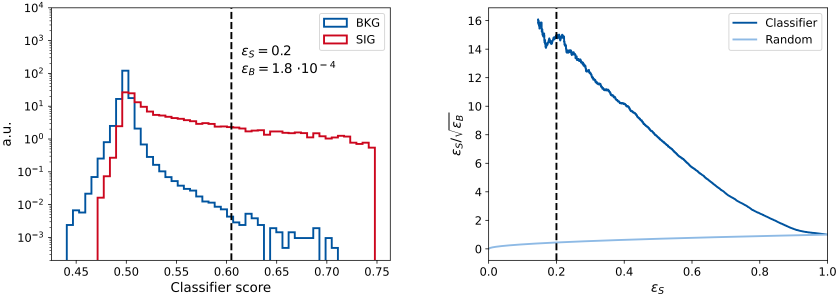


BKG
SIG
a.u.
Classifier score

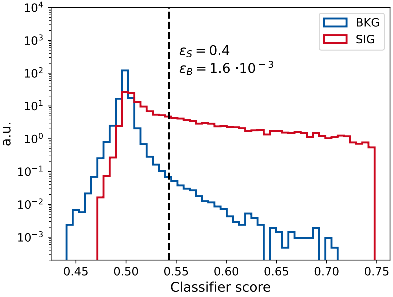


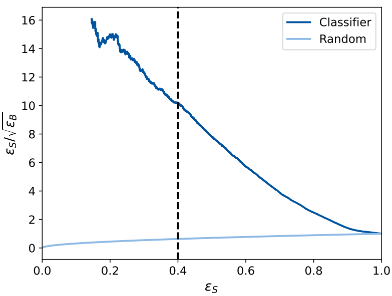
Classifier
Random
$\epsilon_S/\sqrt{\epsilon_B}$
ϵ_S

SIC curves

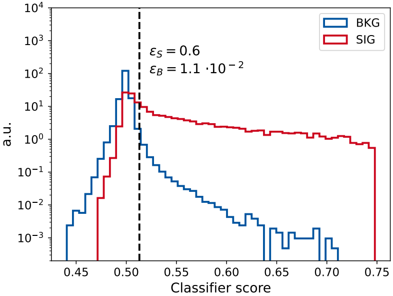

BKG
SIG
$\varepsilon_S = 0.2$
$\varepsilon_B = 1.8 \cdot 10^{-4}$
a.u.
Classifier score
Classifier
Random
$\varepsilon_S/\sqrt{\varepsilon_B}$
ε_S

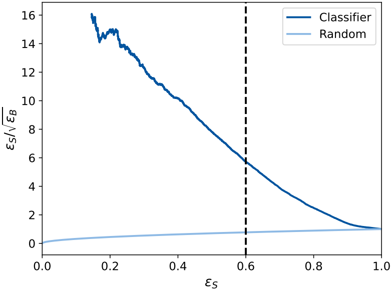
SIC curves


BKG
SIG
$\varepsilon_S = 0.4$
$\varepsilon_B = 1.6 \cdot 10^{-3}$
a.u.
Classifier score

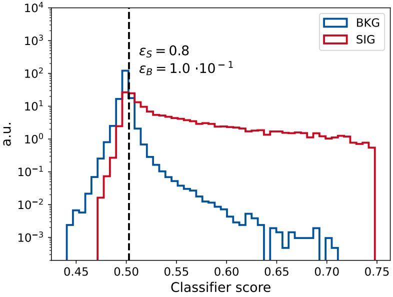

Classifier
Random
$\varepsilon_S/\sqrt{\varepsilon_B}$
ε_S

SIC curves

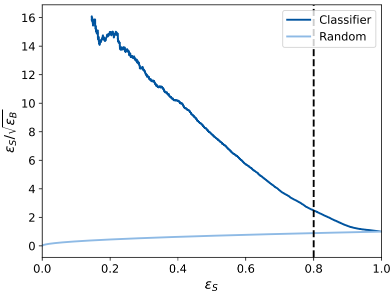

BKG
SIG
$\varepsilon_S = 0.6$
$\varepsilon_B = 1.1 \cdot 10^{-2}$
a.u.
Classifier score


Classifier
Random
$\varepsilon_S/\sqrt{\varepsilon_B}$
ε_S

SIC curves



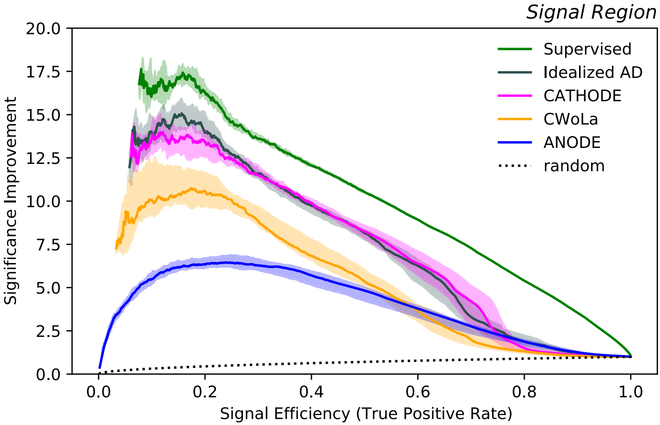
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
a.u.
BKG
SIG
$\varepsilon_S = 0.8$
$\varepsilon_B = 1.0 \cdot 10^{-1}$
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Classifier score



16
14
12
10
8
6
4
2
0
$\varepsilon_S/\sqrt{\varepsilon_B}$
Classifier
Random
0.0
0.2
0.4
0.6
0.8
1.0
ε_S

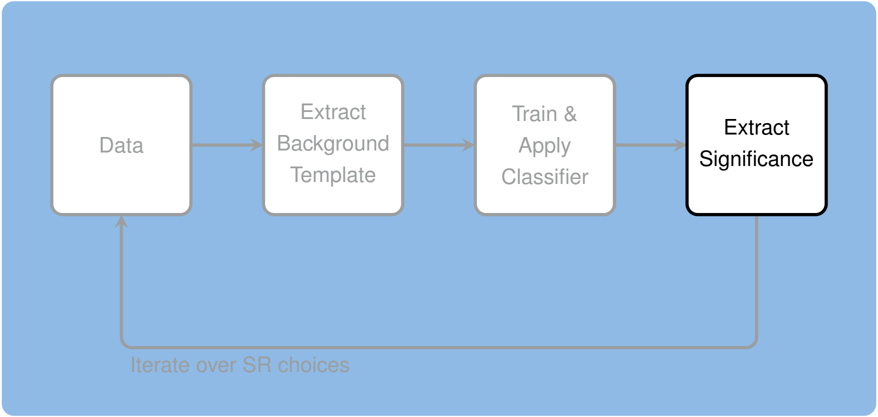
Performance Comparison

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

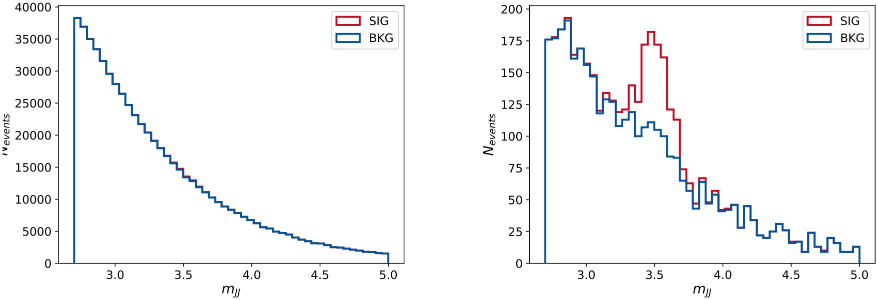


Weakly Supervised Analysis Chain

Full analysis chain

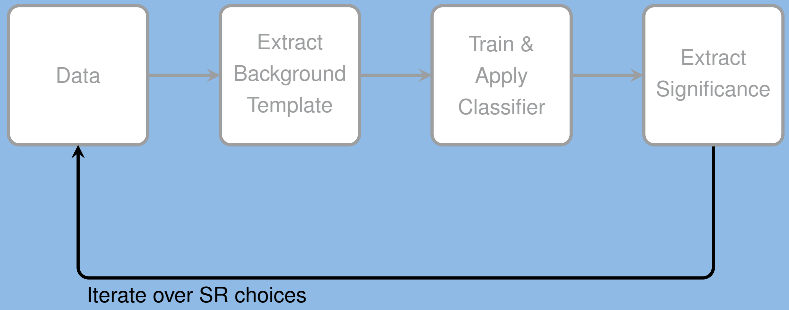


Extracting the significance

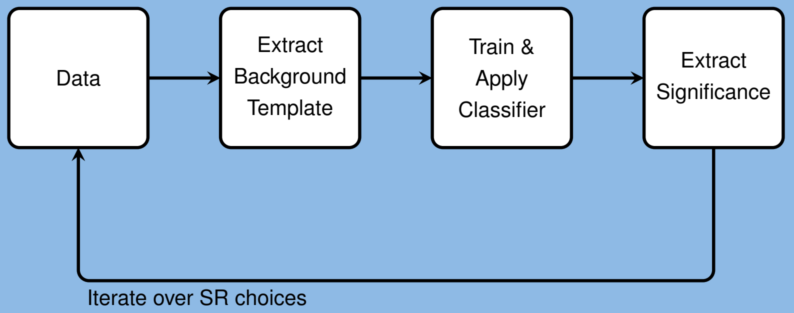


Weakly Supervised Analysis Chain

Full analysis chain

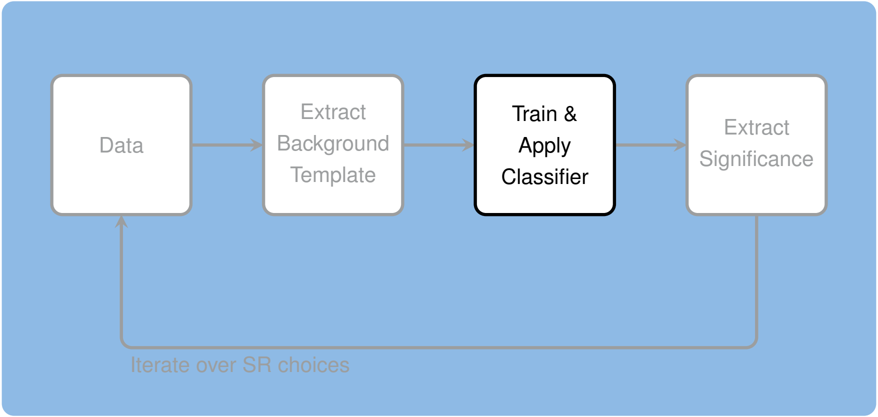


Full analysis chain



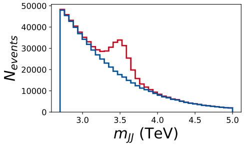
Weakly Supervised Analysis Chain

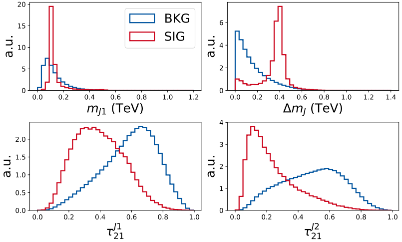
Full analysis chain



How model agnostic are we?

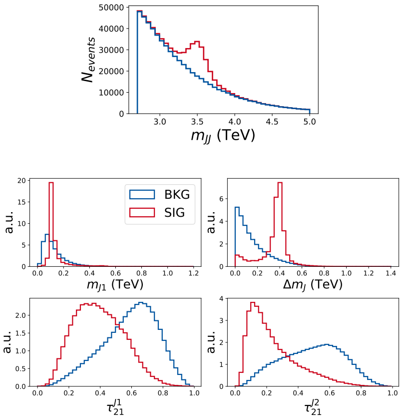
How model agnostic are we?

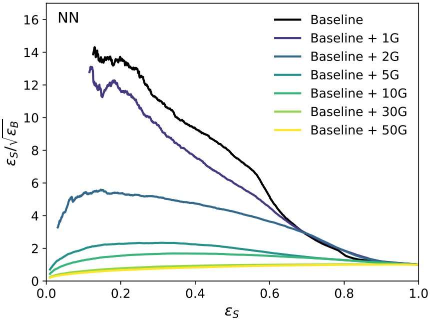




How model agnostic are we?

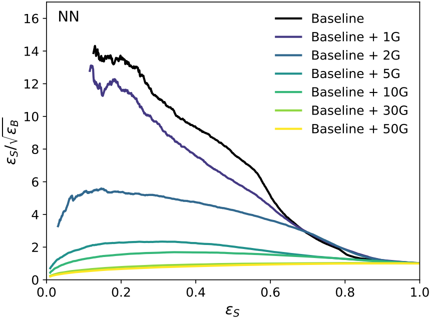
“Tree-Based Algorithms for Weakly Supervised Anomaly Detection” [2309.13111], T. Finke, **MH** et. al.

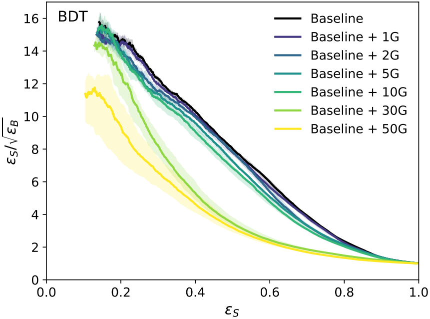




Uninformative features

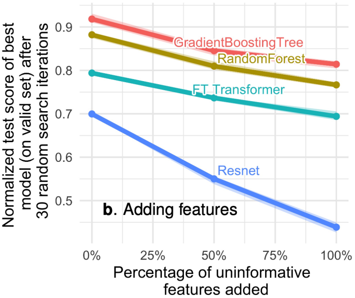
“Tree-Based Algorithms for Weakly Supervised Anomaly Detection” [2309.13111], T. Finke, **MH** et. al.





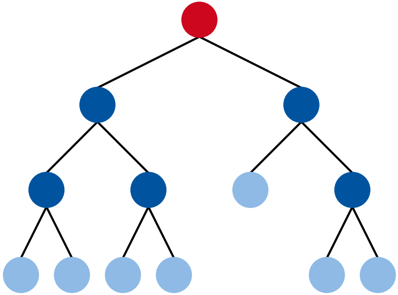
A look at the ML community

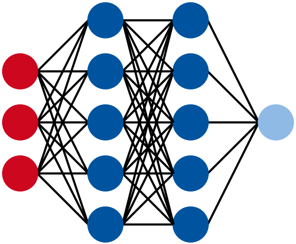
"Why do tree-based models still outperform deep learning on tabular data?" [2207.08815], L. Grinsztaj, E. Ollayon, G. Varoquaux



- Important criteria for their datasets:
  1. Tabular data
  2. Small- to medium-sized datasets

Decision Trees vs. Neural Networks





Branch

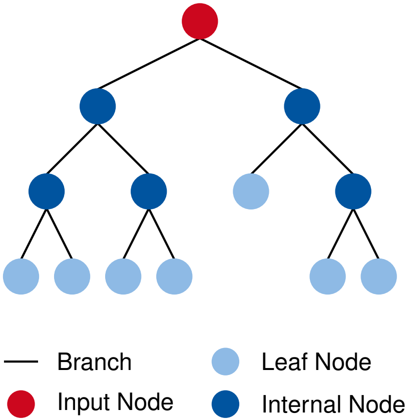
Input Node

Leaf Node

Internal Node

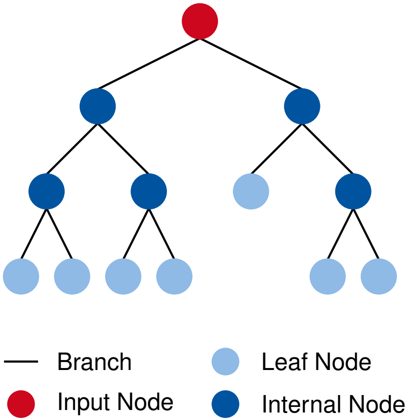
$$x_{1,i} = f(W_{ij}\, x_{0,j} + b_j)$$

Decision Trees



- Start with one input node
- Choose split resulting in best separation of classes
- Iterate until stop condition is met

Boosted Decision Trees



AdaBoost:

Train subsequent trees on misclassified events

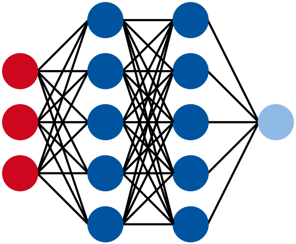
Gradient Boosting:

Train subsequent trees to learn residuals of previous ensemble state

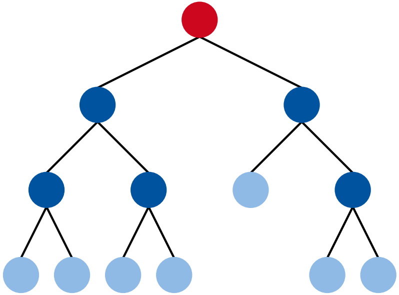
$$y_{\text{pred}} = T_1(x) + \sum_{i=2}^{N} \alpha^{i-1} w_i(x),$$

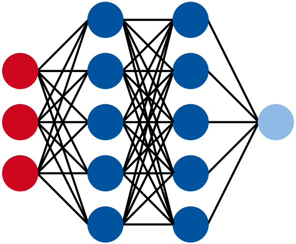
with learning rate α and leaf scores $w_i(x)$.

- Rotationally invariant in input features
- Very good at feature engineering



$$x_{1,i} = f(W_{ij}\, x_{0,j} + b_j)$$





Branch

Input Node

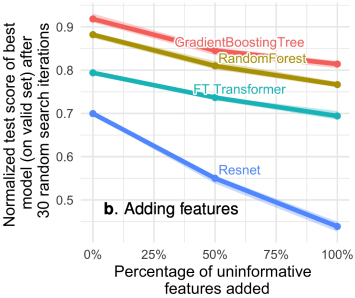
Leaf Node

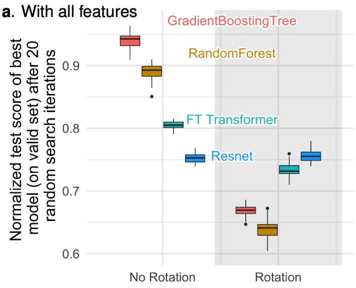
Internal Node

$$x_{1,i} = f(W_{ij}\, x_{0,j} + b_j)$$

A look at the ML community

“Why do tree-based models still outperform deep learning on tabular data?” [2207.08815], L. Grinsztaj, E. Ollayon, G. Varoquaux

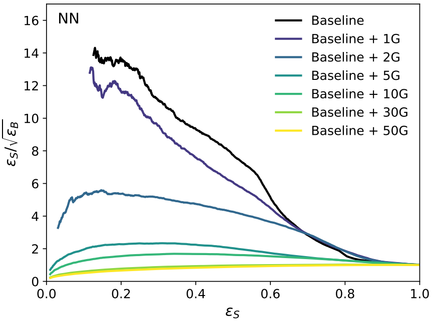


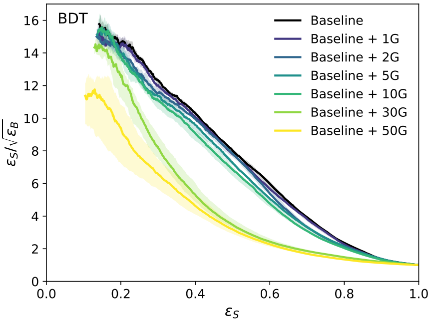


Uninformative features

Institute for Theoretical Particle Physics and Cosmology
RWTH AACHEN UNIVERSITY

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.





Physics-motivated feature sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

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 $\beta \in \{0.5, 1, 2\}$

Physics-motivated feature sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

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 $\beta \in \{0.5, 1, 2\}$

Physics-motivated feature sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

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 $\beta \in \{0.5, 1, 2\}$

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

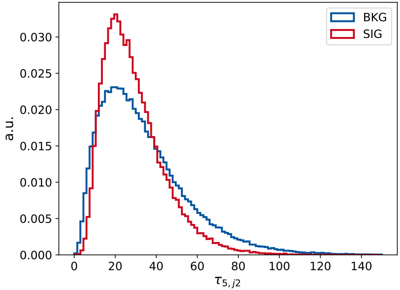
Name	# features	Features
Baseline	4	$\{m_{J_1}, \Delta m_J, \tau_{21}^{\beta=1,J_1}, \tau_{21}^{\beta=1,J_2}\}$
Extended 1	10	$\{m_{J_1}, \Delta m_J, \tau_{N,N-1}^{\beta=1,J_1}, \tau_{N,N-1}^{\beta=1,J_2}\}$ for $2 \leq N \leq 5$
Extended 2	12	$\{m_{J_1}, \Delta m_J, \tau_{N}^{\beta=1,J_1}, \tau_{N}^{\beta=1,J_2}\}$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

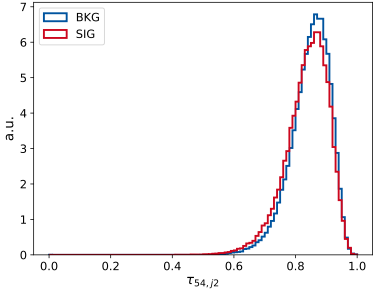
Physics-motivated feature sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

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	$\{m_{J_1}, \Delta m_J, \tau_{N,N-1}^{\beta=1,J_1}, \boldsymbol{\tau_{N,N-1}^{\beta=1,J_2}}\}$ for $2 \leq N \leq 5$
Extended 2	12	$\{m_{J_1}, \Delta m_J, \tau_{N}^{\beta=1,J_1}, \boldsymbol{\tau_{N}^{\beta=1,J_2}}\}$ 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 $\beta \in \{0.5, 1, 2\}$

Subjettiness vs. Subjettiness ratio





Physics-motivated feature sets

Institute for Theoretical Particle Physics and Cosmology

RWTH AACHEN UNIVERSITY

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

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 $\beta \in \{0.5, 1, 2\}$

Physics-motivated feature sets

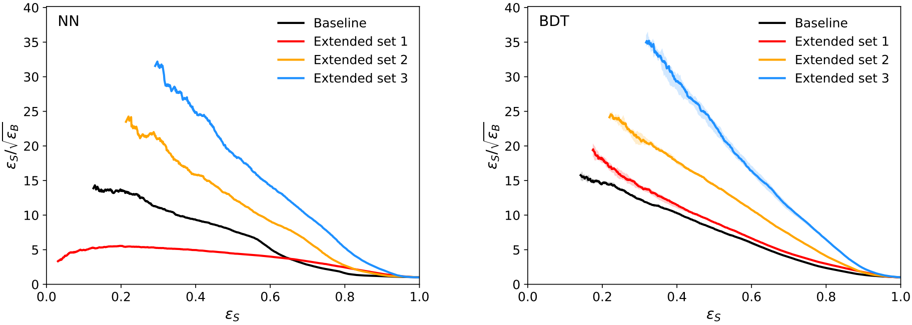
Institute for Theoretical Particle Physics and Cosmology | RWTH AACHEN UNIVERSITY

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.

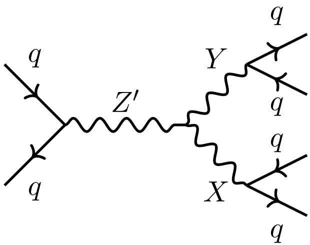
Name	# features	Features
Baseline	4	All informative
Extended 1	10	Some uninformative
Extended 2	12	All slightly informative
Extended 3	56	All slightly informative

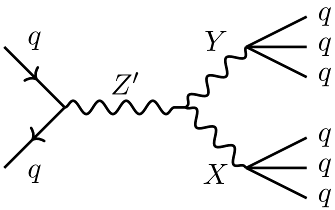
Results for different feature sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.



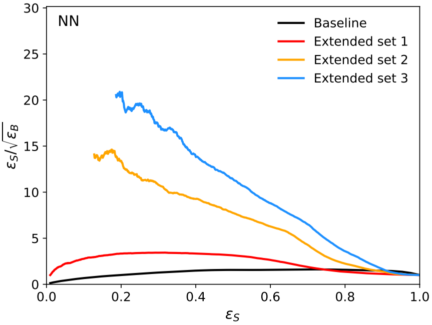
Different signal model

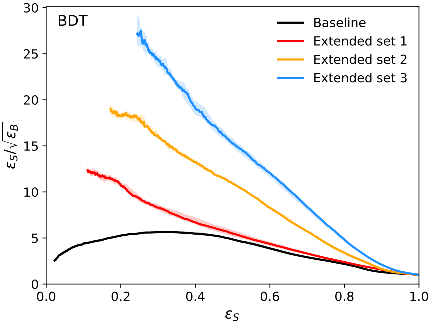




Results for 3-pronged signal

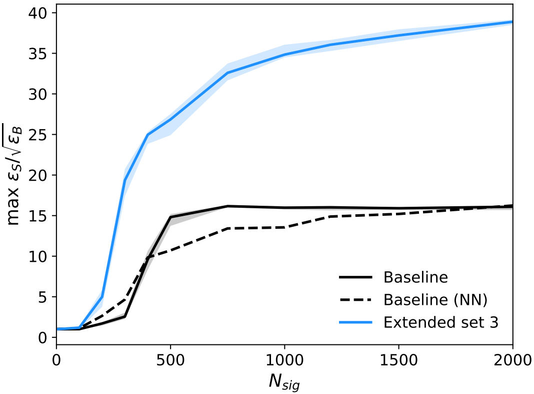
"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.





Results for 3-pronged signal

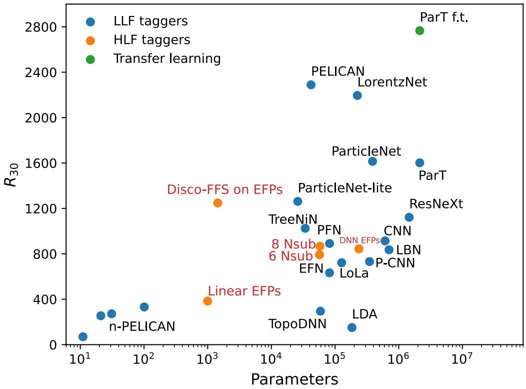
“Tree-Based Algorithms for Weakly Supervised Anomaly Detection” [2309.13111], T. Finke, **MH** et. al.

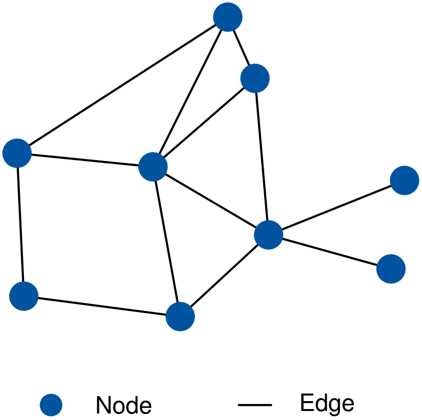


Low level features

Background: Top Tagging

"The Machine Learning Landscape of Top Taggers" [1902.09914], G. Kasieczka, T. Plehn, et. al.
"Feature Selection with Distance Correlation" [2212.00046], R. Das, G. Kasieczka, D. Shih

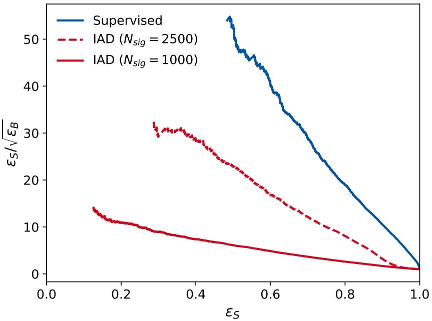




- Permutation invariant representation of the data
- Can encode physical symmetries directly into the architecture

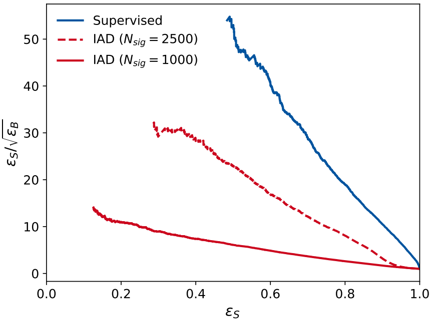
GNNs for weak supervision

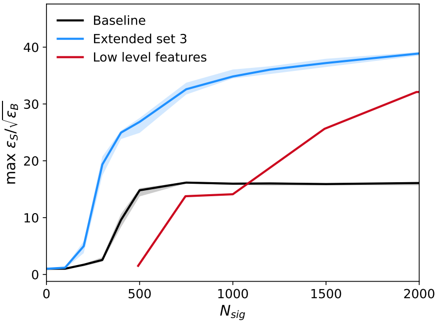
“Identifying Anomalous Events Using Low-Level LHC Data”, Master Thesis by Joep Geuskens



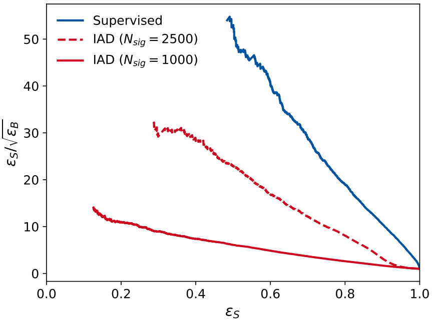
GNNs for weak supervision

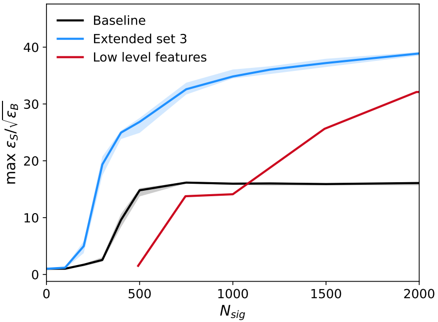
“Identifying Anomalous Events Using Low-Level LHC Data”, Master Thesis by Joep Geuskens





"Identifying Anomalous Events Using Low-Level LHC Data", Master Thesis by Joep Geuskens
"Full Phase Space Resonant Anomaly Detection" [2310.06897], E. Buhmann et. al.





Lessons learned

Features for anomaly detection

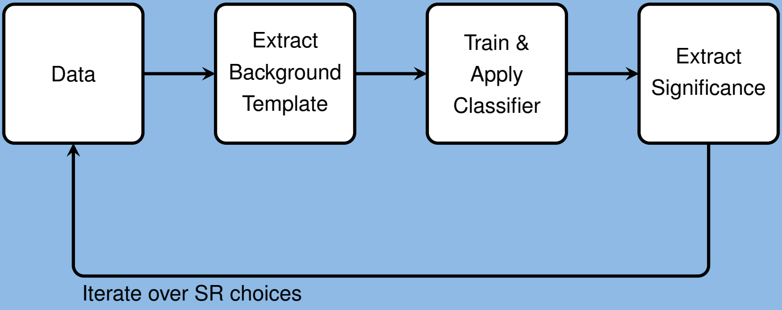
- High level features can provide good performance with current state of the art technology
- Low level features are the more future-oriented approach but in the present still more difficult to achieve

Strategies for improving physics ML applications

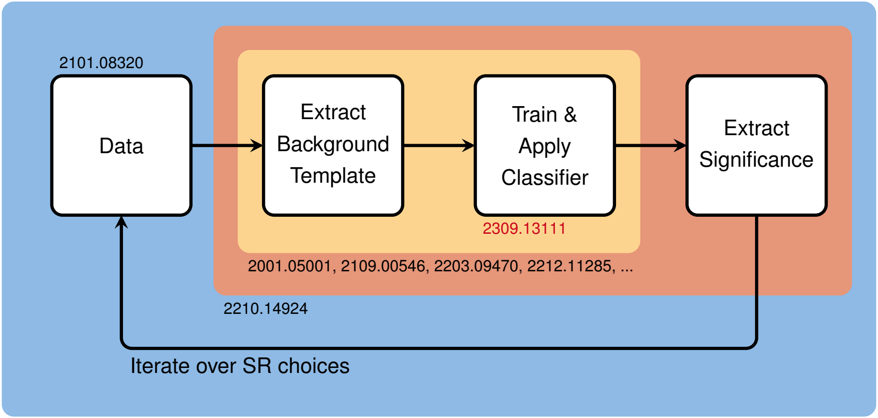
- Incorporating physics knowledge in algorithms and feature selection
- Considering the state of the art in Machine Learning research

Weakly Supervised Analysis Chain

Full analysis chain



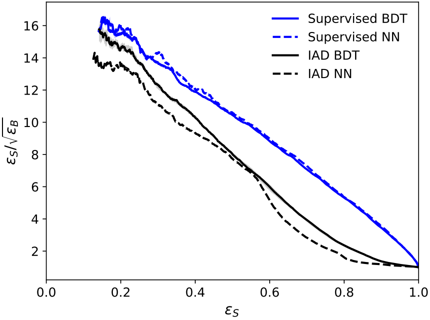
Full analysis chain



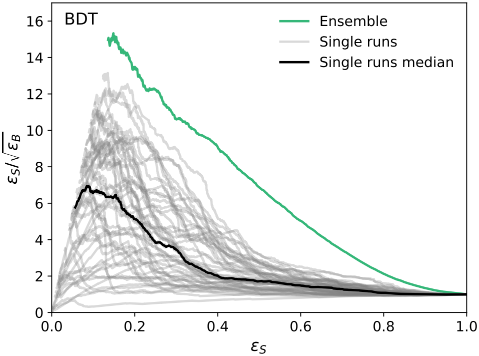
1902.02634, 2005.02983, EXO-022-026, upcoming work

Backup Slides

- NN: Ensemble of N independently trained fully connected NNs
- BDT: Ensemble of N independently trained Histogrammed Gradient Boosted Decision Trees



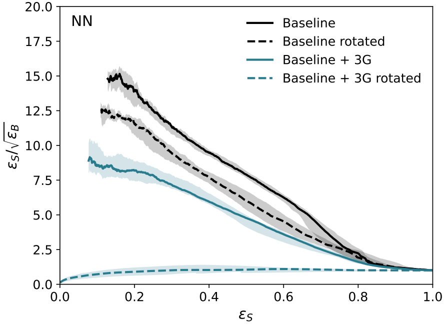
Ensembling

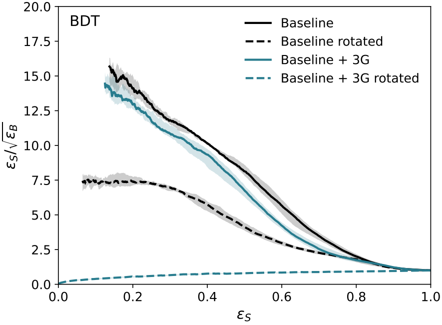


BDT
Ensemble
Single runs
Single runs median
$\varepsilon_S/\sqrt{\varepsilon_B}$
ε_S
0
2
4
6
8
10
12
14
16
0.0
0.2
0.4
0.6
0.8
1.0

Results for different feature sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.





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

RWTH Aachen University
Templergraben 55
52056 Aachen

www.rwth-aachen.de