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

Choosing the right features for weak supervision

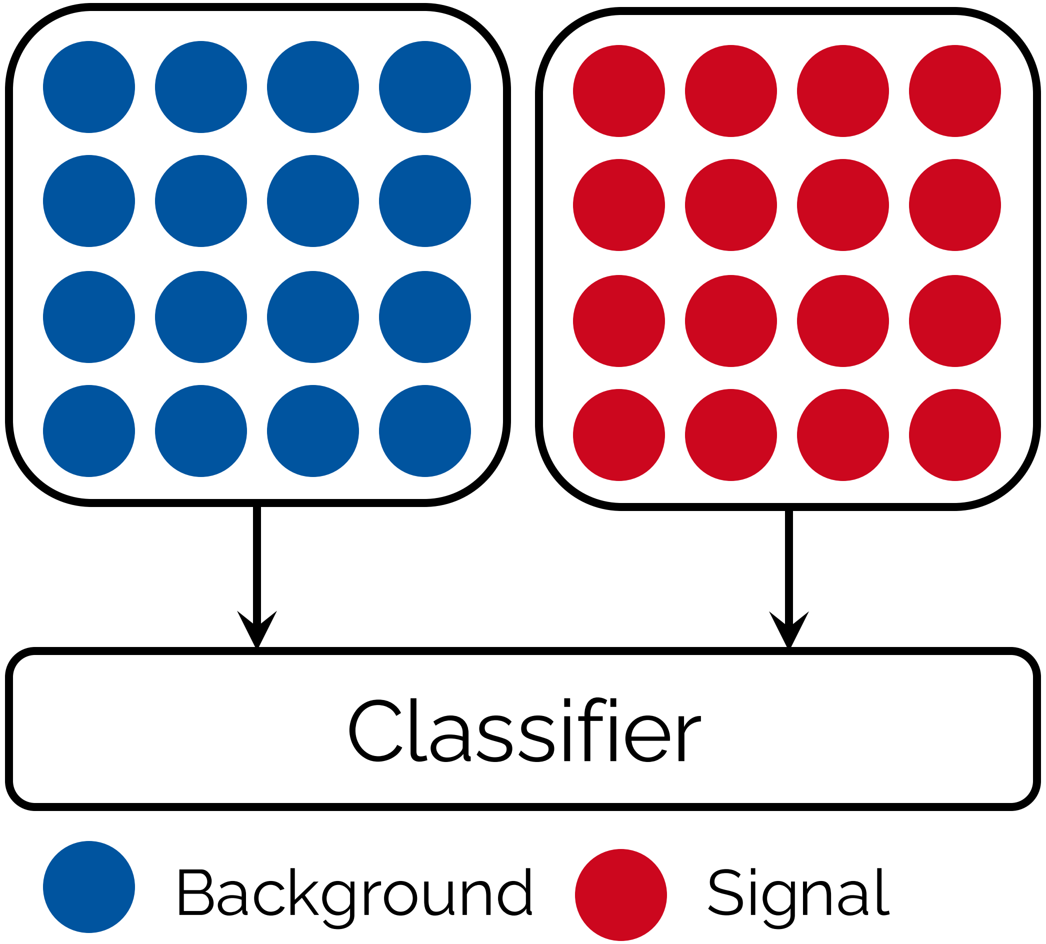
Marie Hein

In collaboration with Joep Geuskens, Michael Krämer, Lukas Lang, Radha Mastandrea & Alexander Mück

CRC Young Scientists Meeting 2024

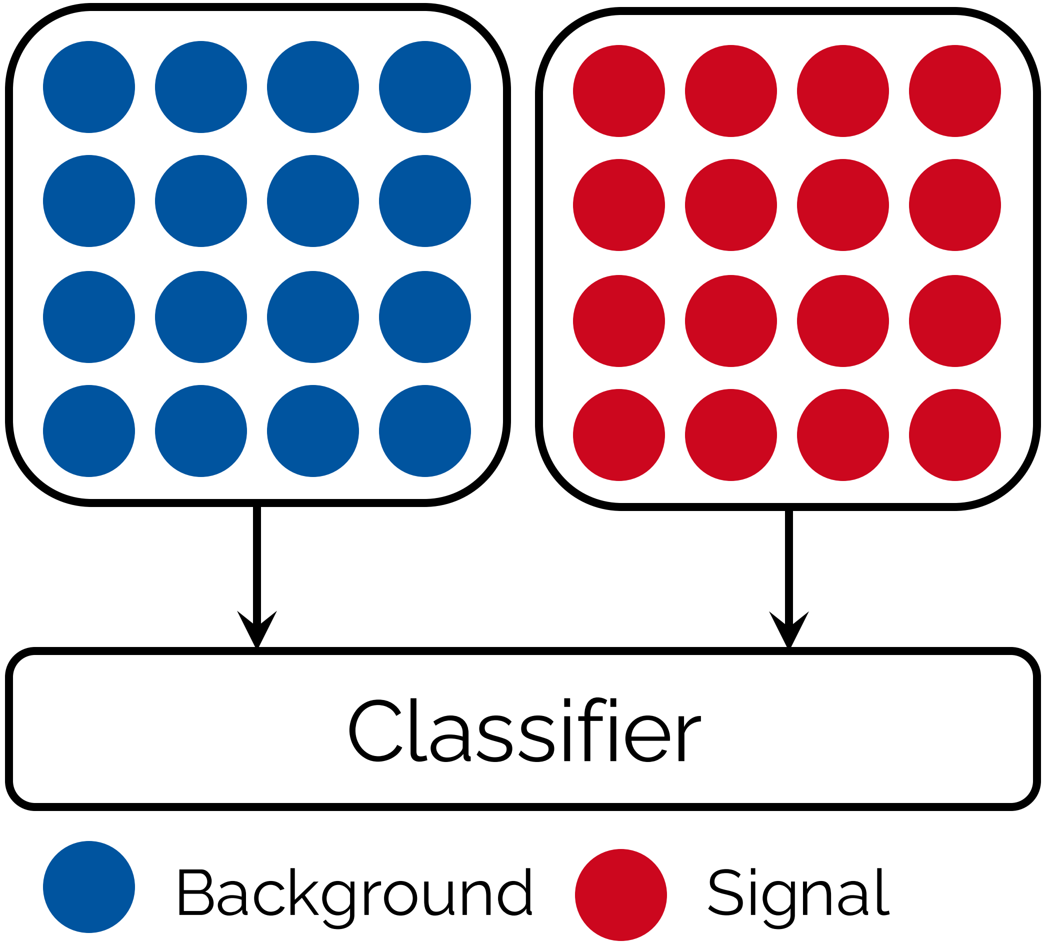
Classification Problem

- Goal: To achieve a better signal to background ratio



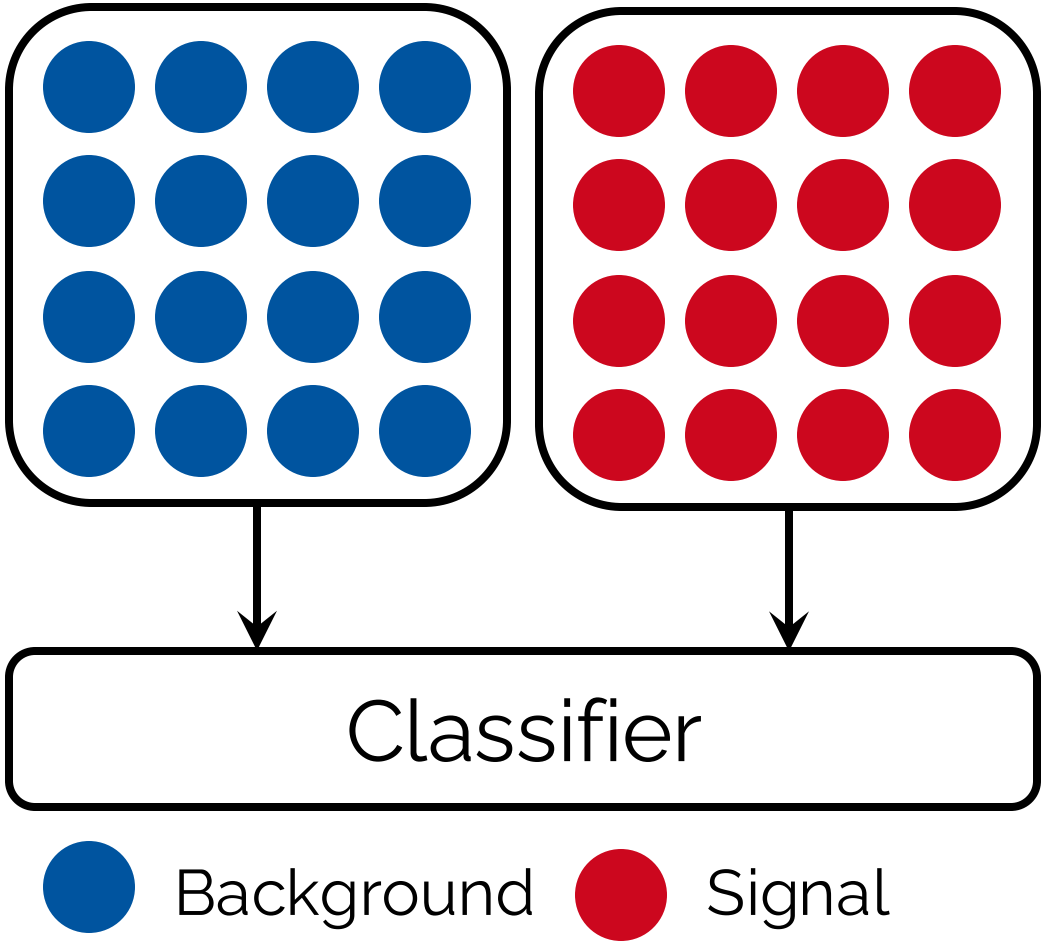
Classification Problem

- Goal: To achieve a better signal to background ratio
- Ansatz: Perform classification task



Classification Problem

- Goal: To achieve a better signal to background ratio
- Ansatz: Perform classification task
- Problem: Labels are not available on real data

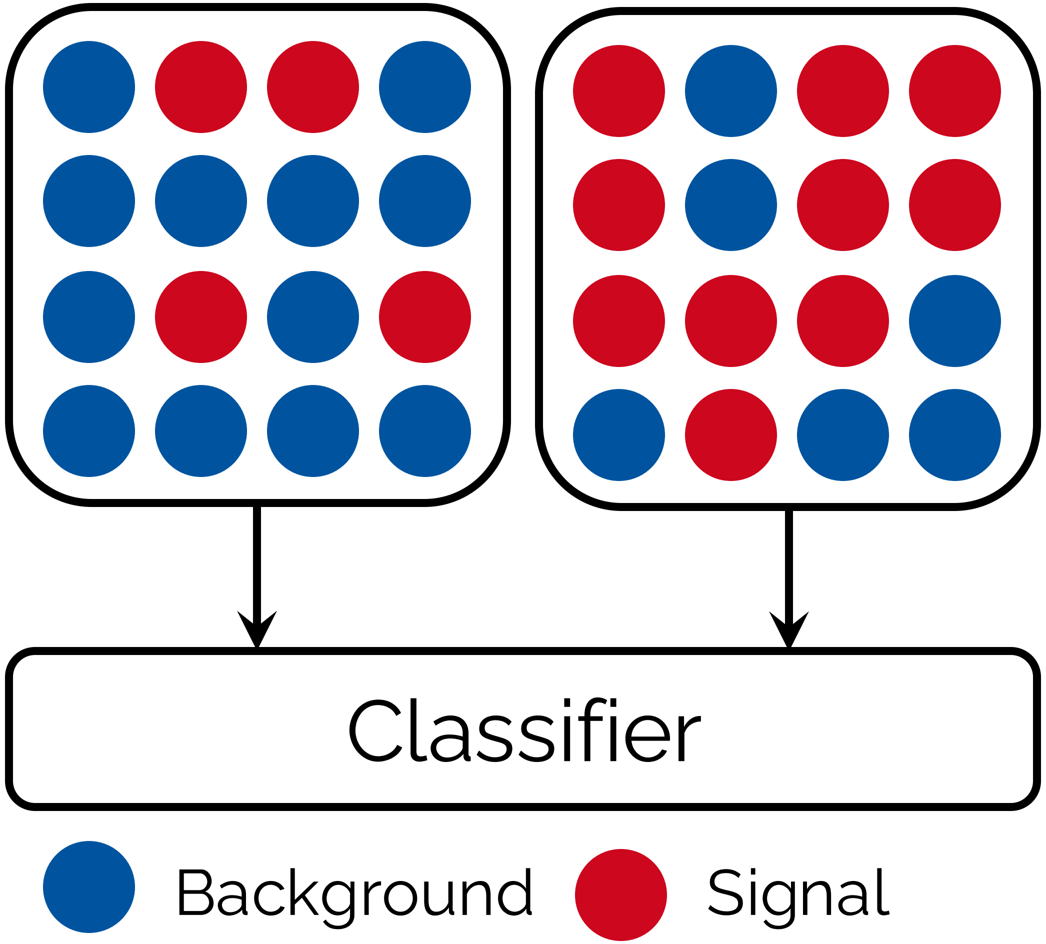


Weakly Supervised Classification

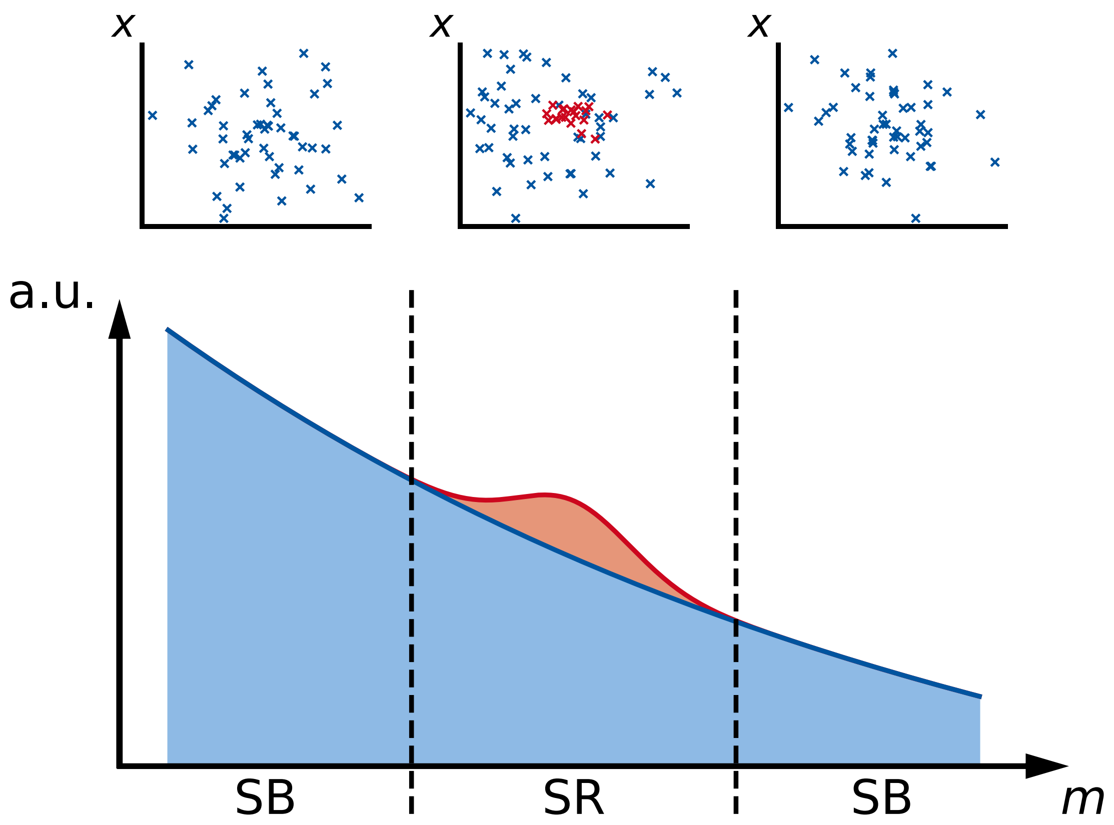
"Classification without labels: Learning from mixed samples in high energy physics" [1709.02949], E. Metodiev, B. Nachman, J. Thaler

- Goal: To achieve a better signal to background ratio
- Ansatz: Perform classification task
- Problem: Labels are not available on real data
- Solution: Classify between mixed classes

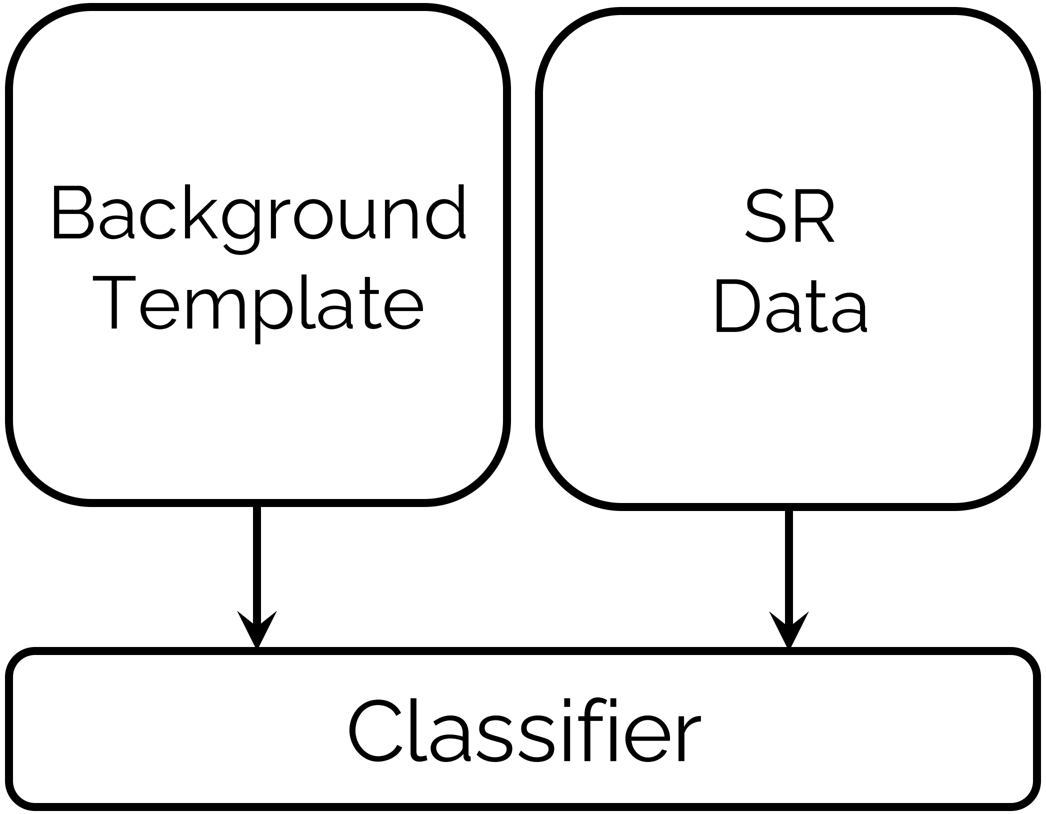
  →Fundamentally, both problems are equivalent



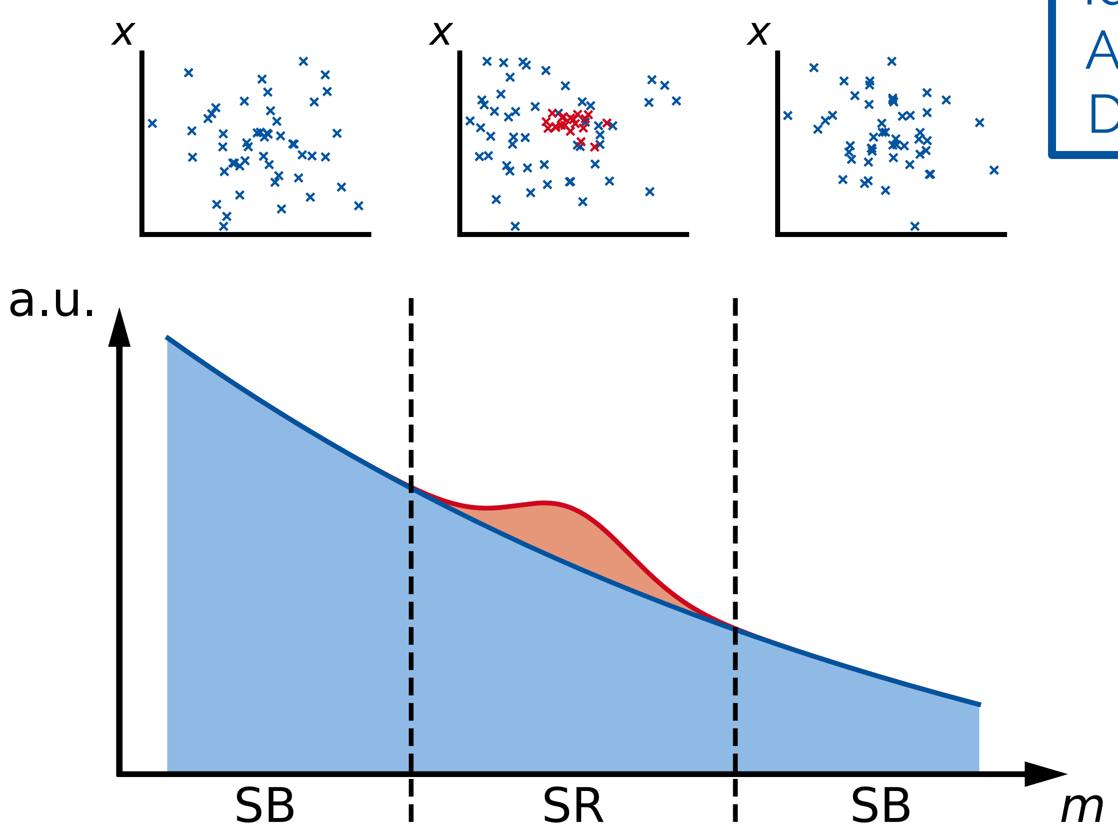
Application to resonance searches



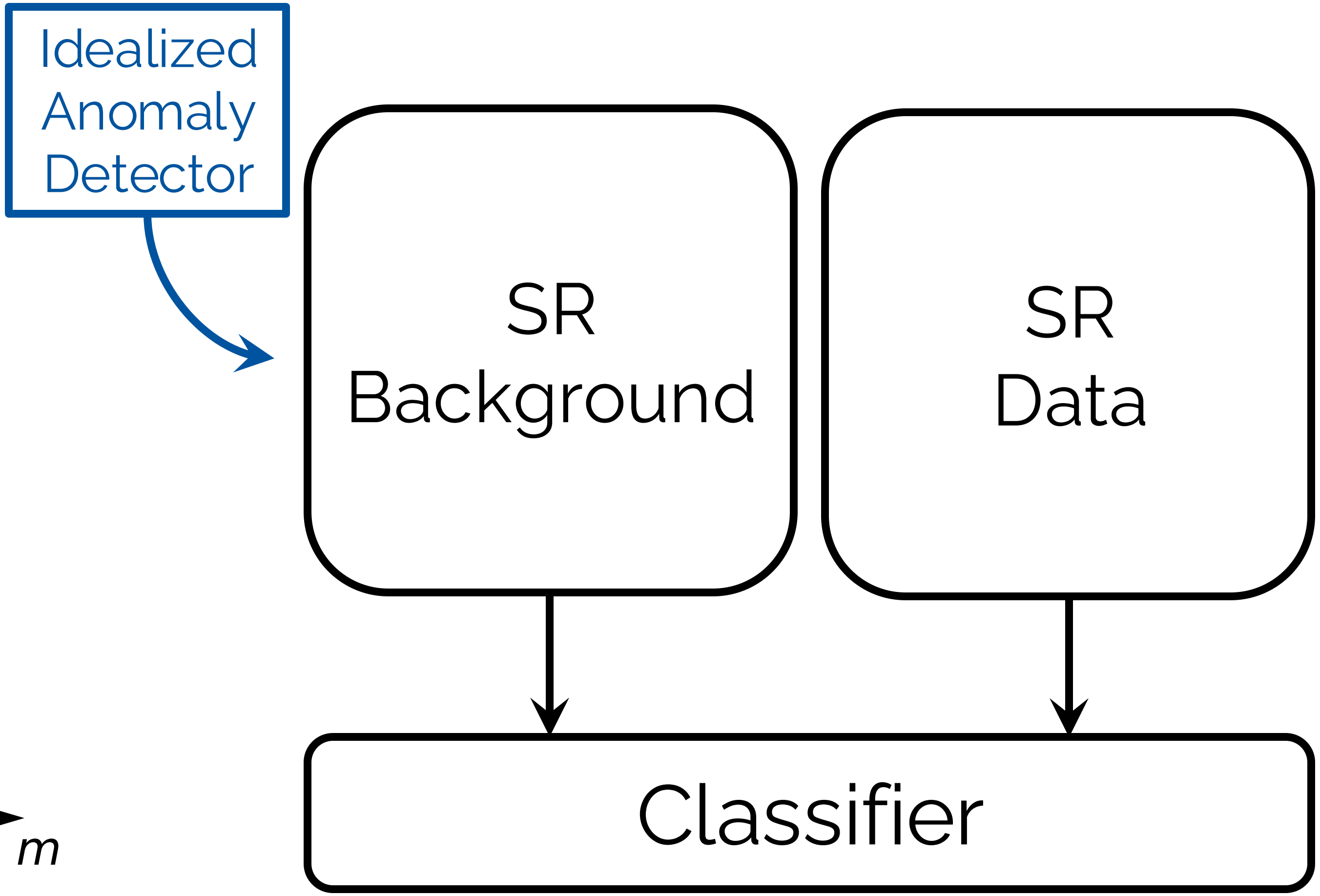
Recreated from [2109.00546]



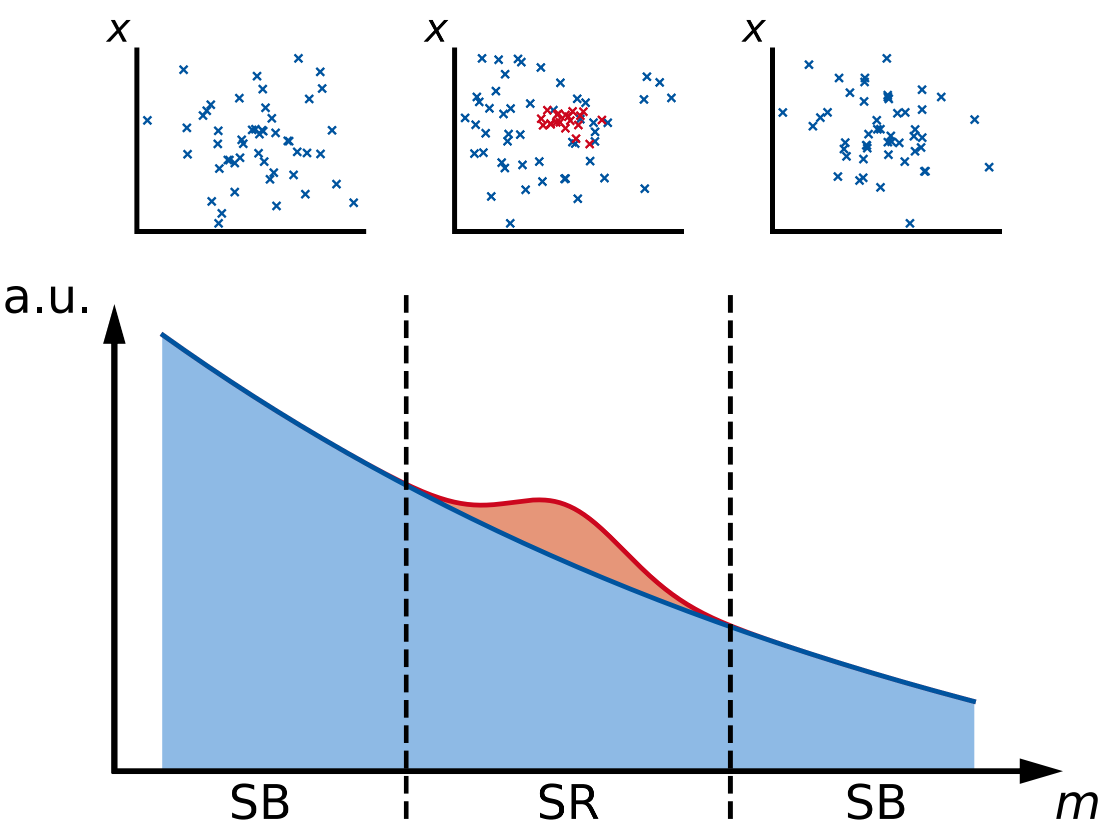
Application to resonance searches



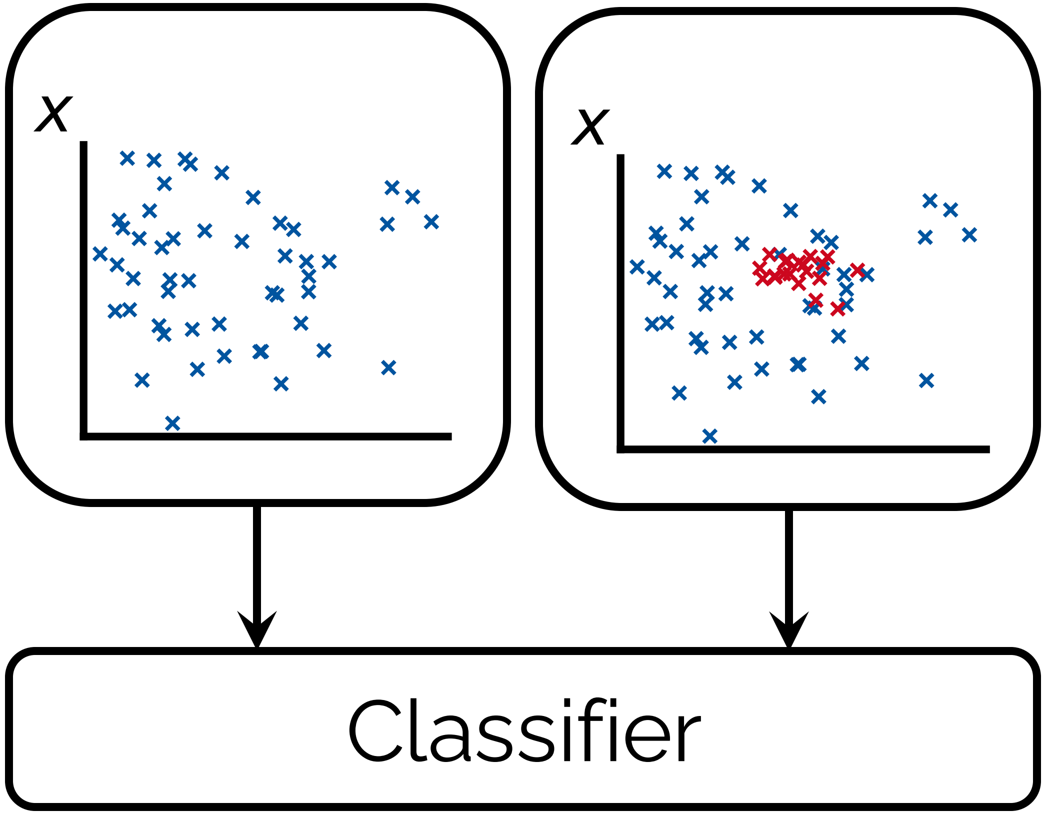
Recreated from [2109.00546]

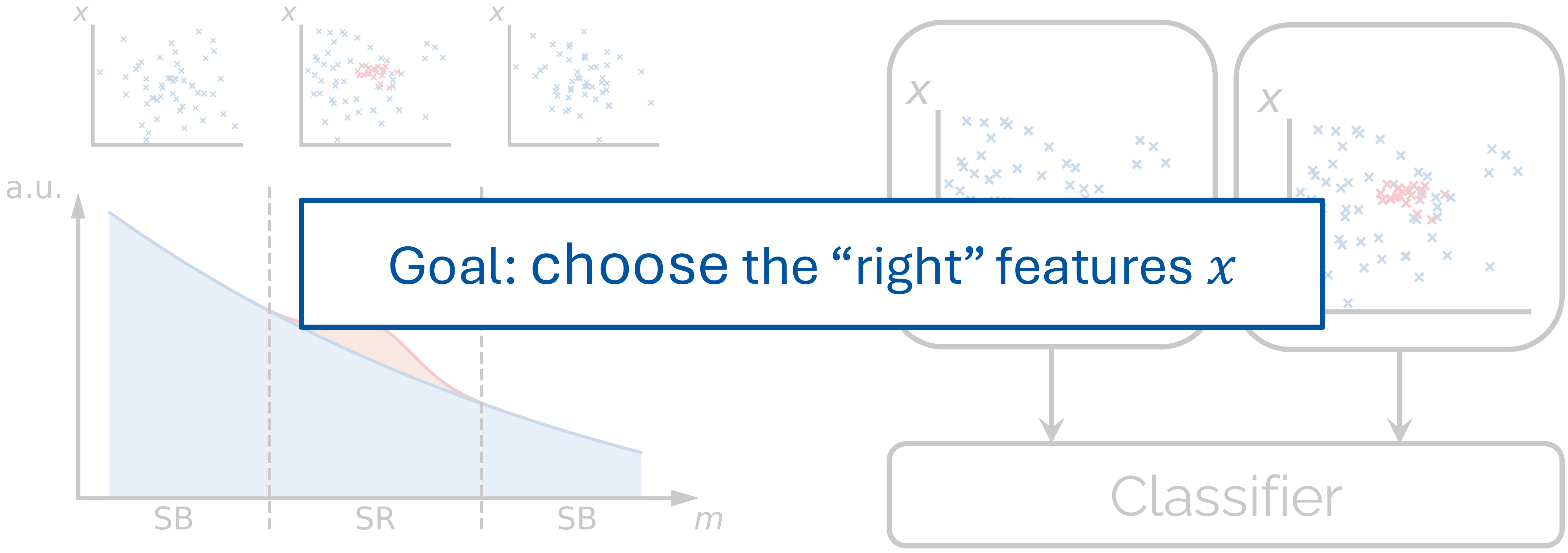


Application to resonance searches



Recreated from [2109.00546]





Recreated from [2109.00546]

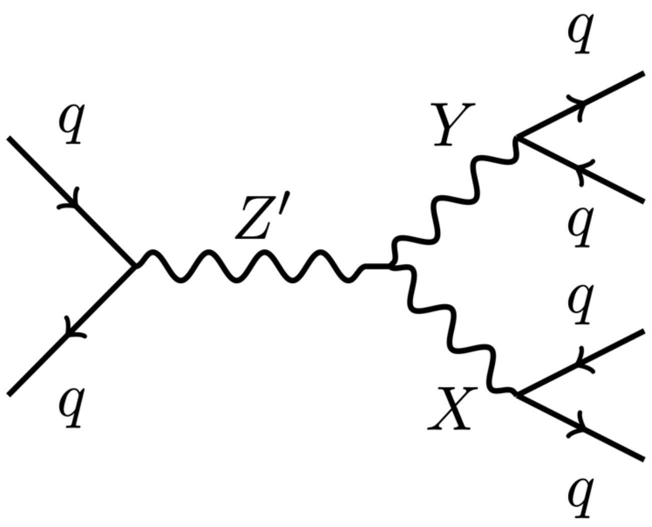
Goal: choose the "right" features x

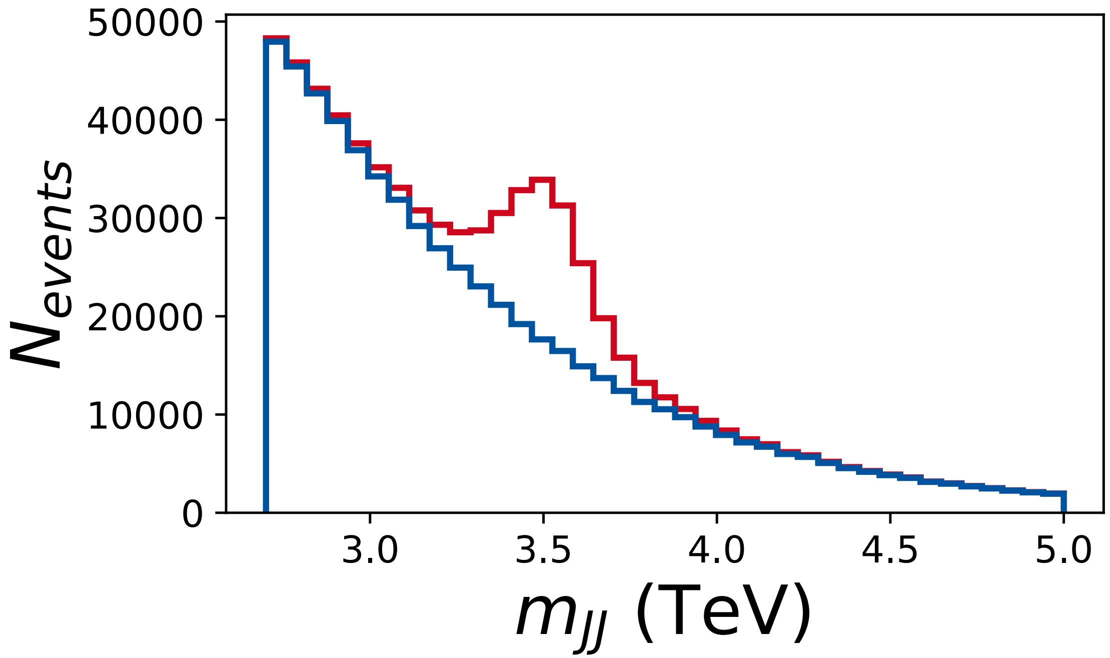
The Dataset & Features

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
- QCD dijet background
- Signal

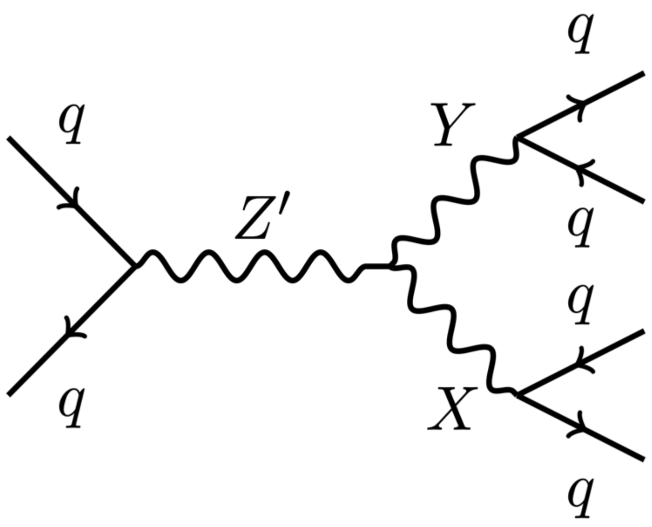


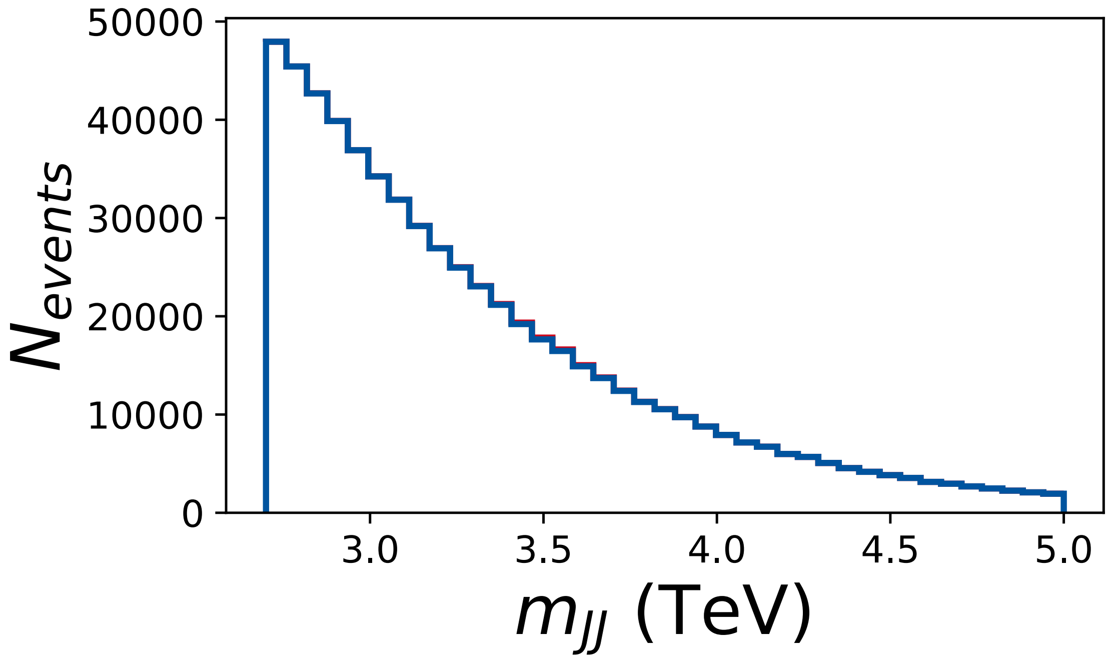


LHCO R&D dataset

"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
- QCD dijet background
- Signal





Options for Substructure Observables

1. Low level features: use particle four-momenta in jet-separated LorentzNet
   - Very model agnostic

2. High level Features: derive observables from low-level features
   - Less model agnostic
   - Easier classification task (more closely related to problem to be solved)

   a. N-Subjettiness

   b. Energy Flow Polynomials

Subjettiness

“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 into N subjets to obtain

Sum over all particles

Angular distance measure

$$\Delta R_{Ji} = \sqrt{(\Delta y_{Ji})^2 + (\Delta\varphi_{Ji})^2}$$

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

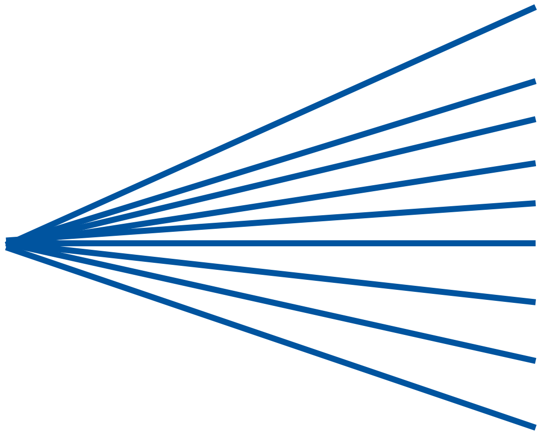
Normalization

Subjet candidates

- “Momentum-weighted sum of angular distance of all particles to closest subjet”

Subjettiness

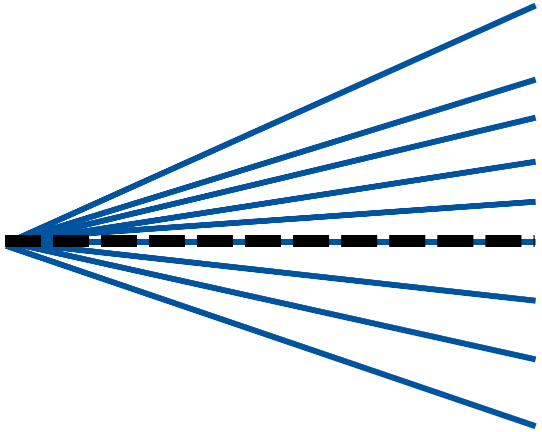
Background

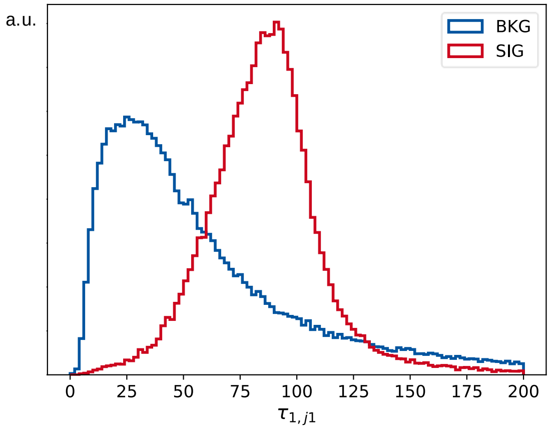


Signal

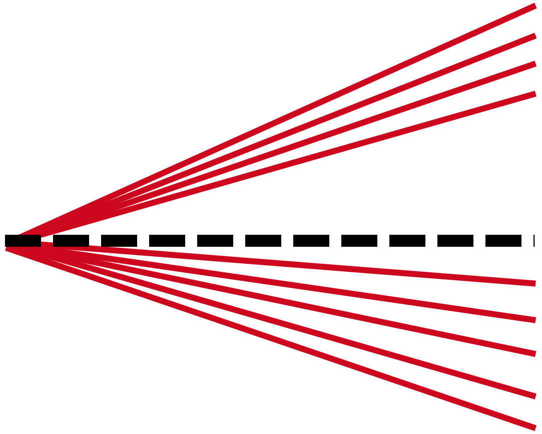
Subjettiness

Background



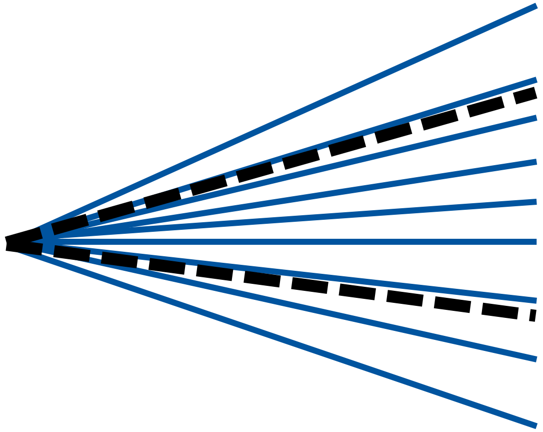


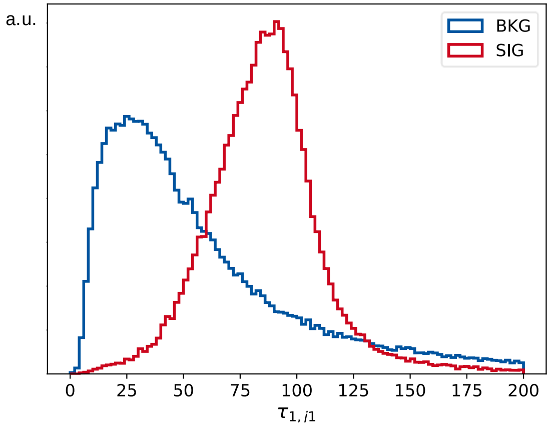
Signal



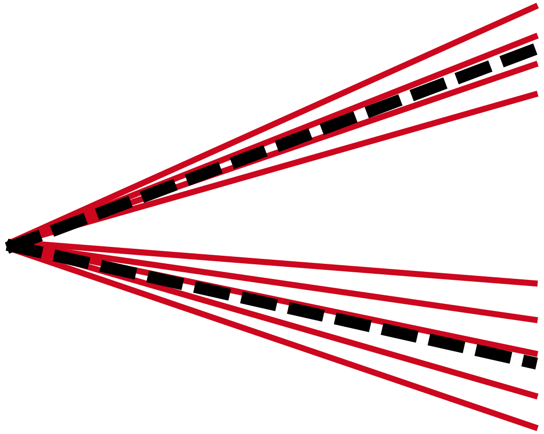
Subjettiness

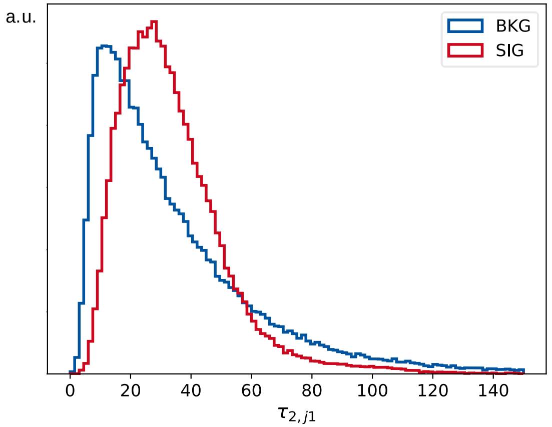
Background





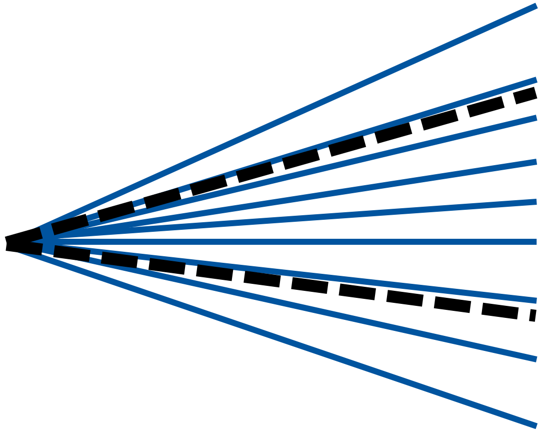
Signal

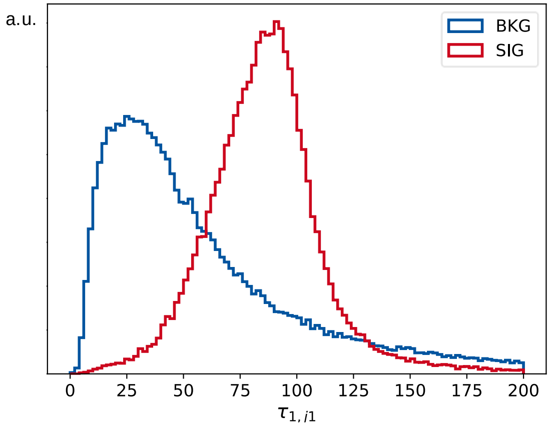


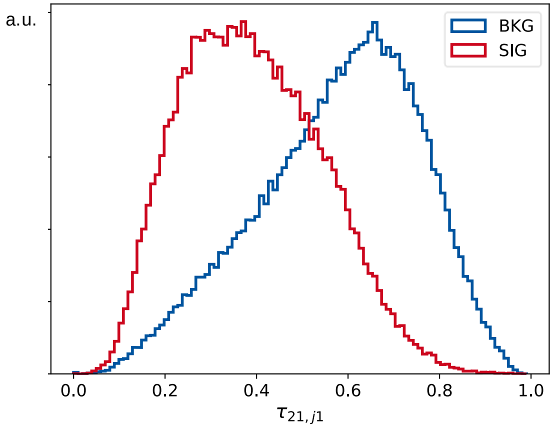


Subjettiness

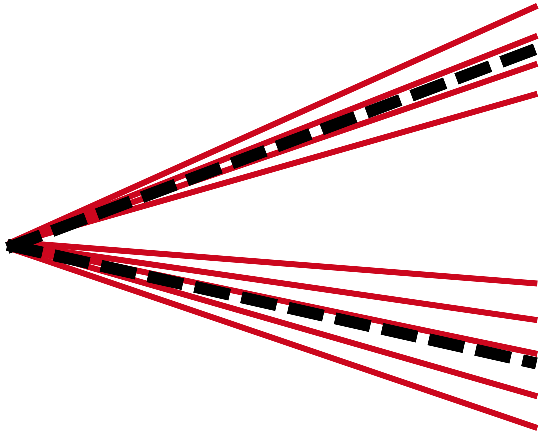
Background

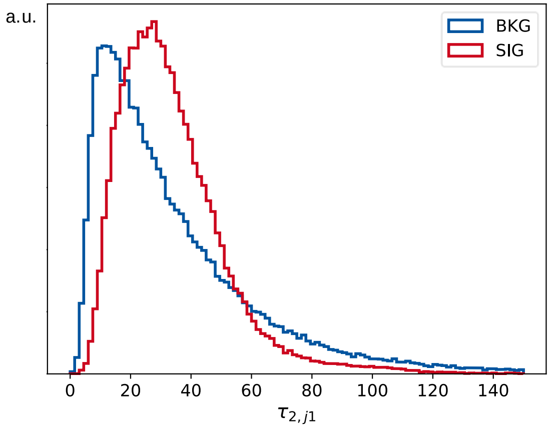






Signal



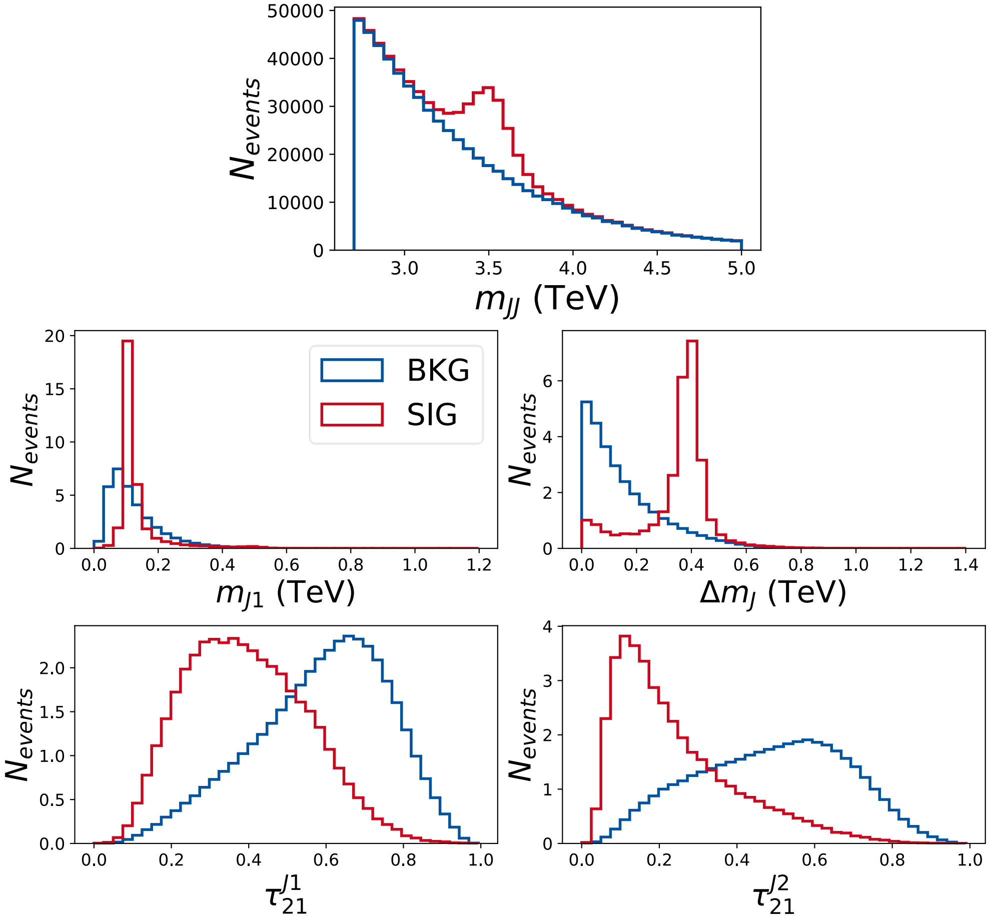


Subjettiness-based feature sets

1. Baseline feature set
   - Jet masses $m_{J1}, \Delta m_J$
   - 21-Subjettiness ratio $\tau_{21,J1}, \tau_{21,J2}$

2. Extended feature set
   - Jet masses $m_{J1}, \Delta m_J$
   - Use 54 different subjettiness features (varying N and β)

Baseline Features



Energy Flow Polynomials

“Energy flow polynomials: A complete linear basis for jet substructure” [1712.07124], P. Komiske, E. Metodiev, J. Thaler
“Energy Flow Networks: Deep Sets for Particle Jets” [1810.05165], P. Komiske, E. Metodiev, J. Thaler

- Complete linear basis of jet substructure observables

Energy fraction
$z_i = p_{T,i}/p_{T,J}$

Angular distance measure
$\theta_{ij} = \left(\left(\Delta y_{ij}\right)^2 + \left(\Delta \varphi_{ij}\right)^2\right)^{\beta/2}$

$$EFP_G = \sum_{i_1=1}^{M} \dots \sum_{i_N=1}^{M} z_{i_1} \dots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}$$

Sum over all particles

Product of particle pairs

EFP-Multigraph Correspondence

Vertices

Edges

$$EFP_G = \sum_{i_1=1}^{M} \dots \sum_{i_N=1}^{M} z_{i_1} \dots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}$$

Loopless Multigraphs

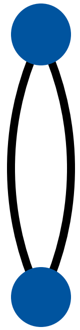
EFP-Multigraph Correspondence

Vertices

Edges

$$EFP_G = \sum_{i_1=1}^{M} \dots \sum_{i_N=1}^{M} z_{i_1} \dots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}$$

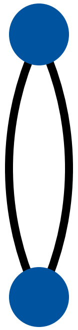
Loopless Multigraphs



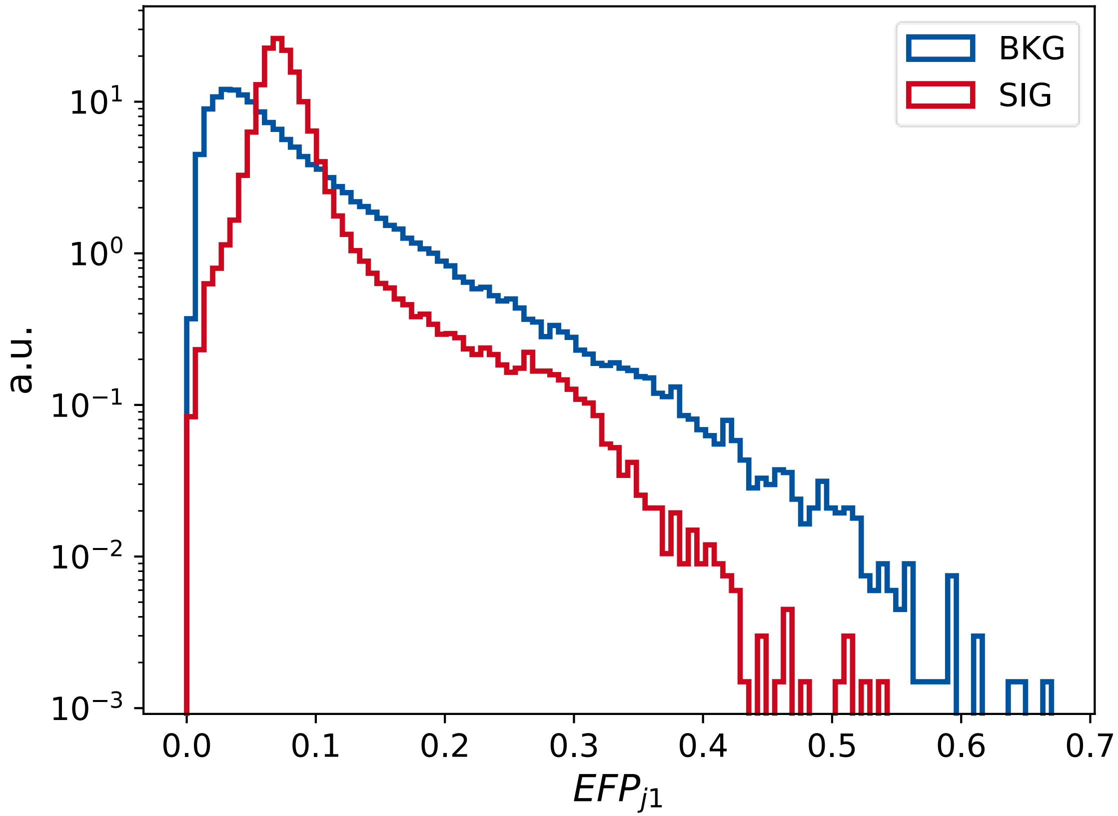
$$= \sum_{i=1}^{M}\sum_{j=1}^{M} z_i z_j\, \theta_{ij}^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{p_{T,i}\, p_{T,j}}{p_{T,J}^2}\left(\left(\Delta y_{ij}\right)^2 + \left(\Delta\varphi_{ij}\right)^2\right) \approx \frac{2m_J^2}{p_T^2}$$

EFP-Multigraph Correspondence

$$EFP_G = \sum_{i_1=1}^{M} \dots \sum_{i_N=1}^{M} z_{i_1} \dots z_{i_N} \prod_{(k,l)\in G} \theta_{i_k i_l}$$



$$\approx \frac{2m_J^2}{p_T^2}$$



EFP-based feature set

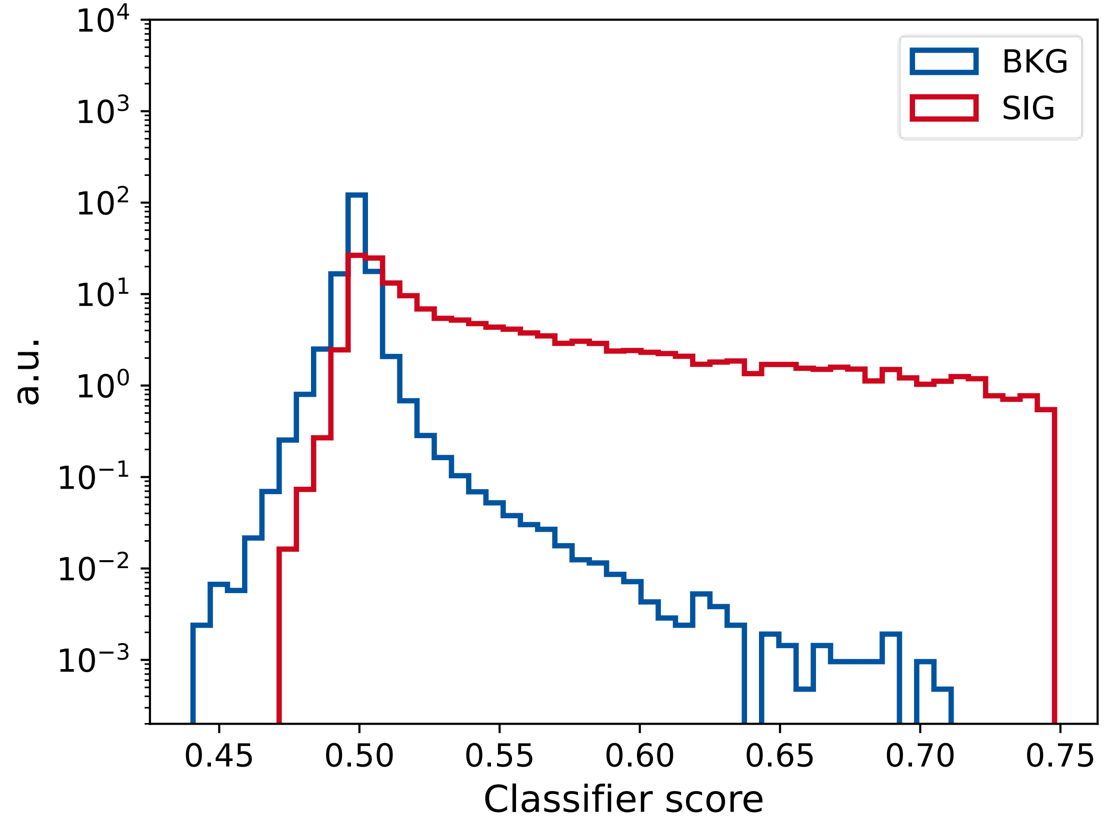
1. EFP feature set
    - Jet masses $m_{J1}, \Delta m_J$
    - 490 EFPs per jet (up to 7 edges)

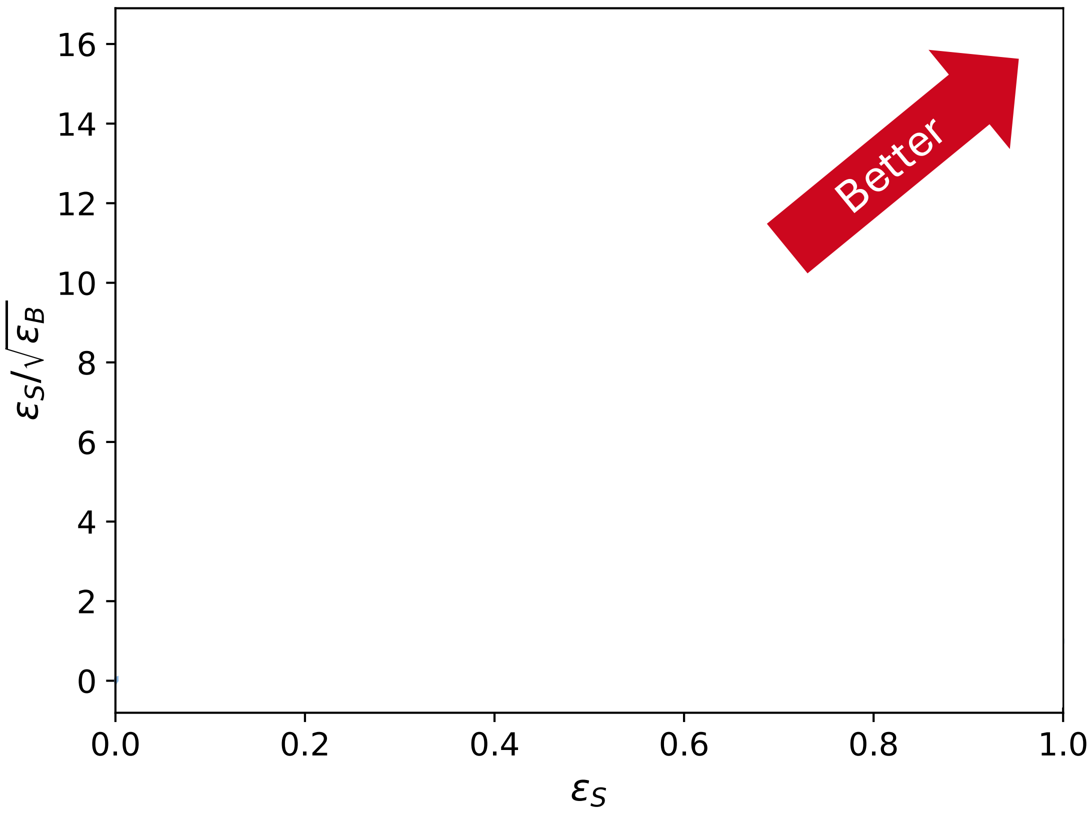
Feature sets

1. Low level features
2. High level features
    a. Baseline feature set (Subjettiness) → 4 features
    b. Extended feature set (Subjettiness) → 56 features
    c. EFP feature set → 982 features

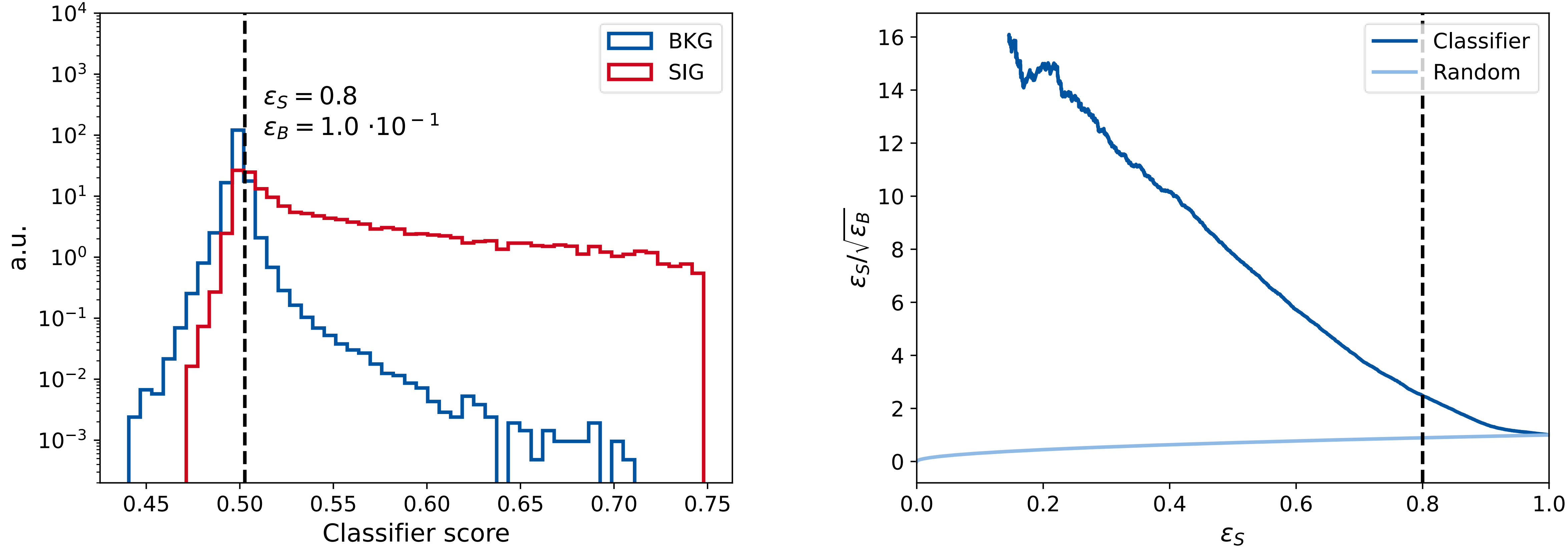
Performance Measure

Classifier score to anomaly detection

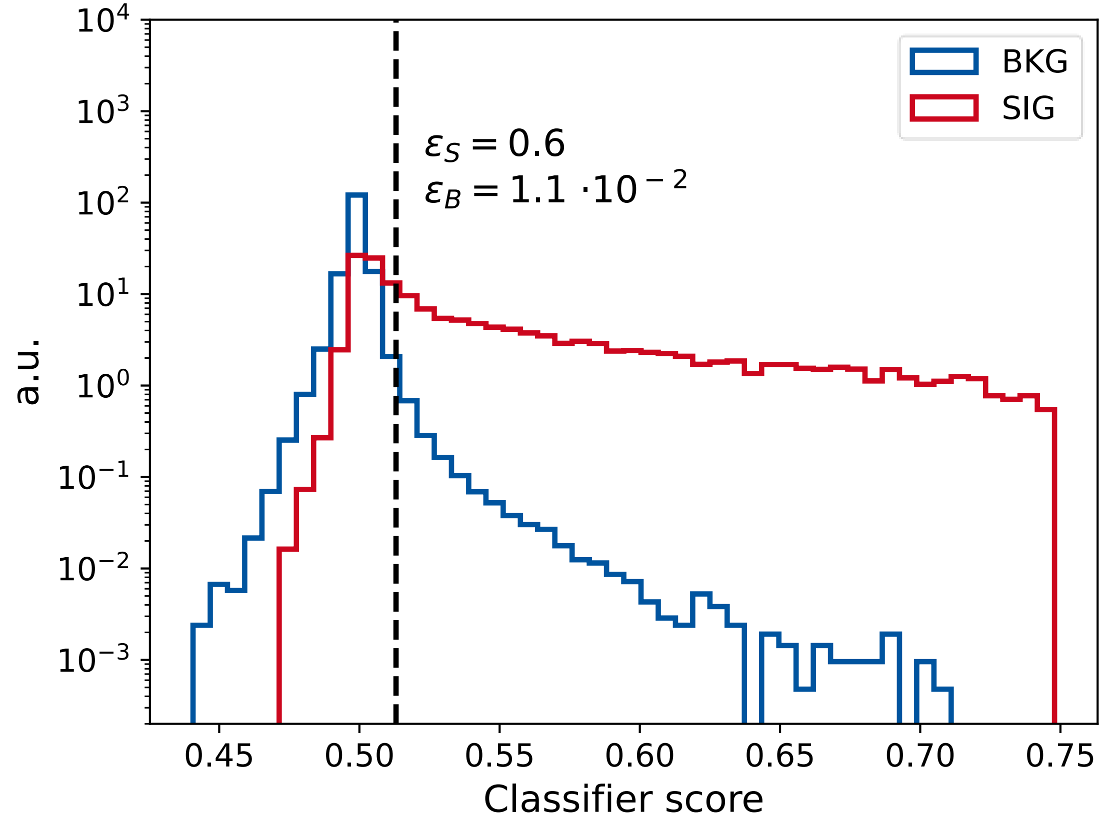


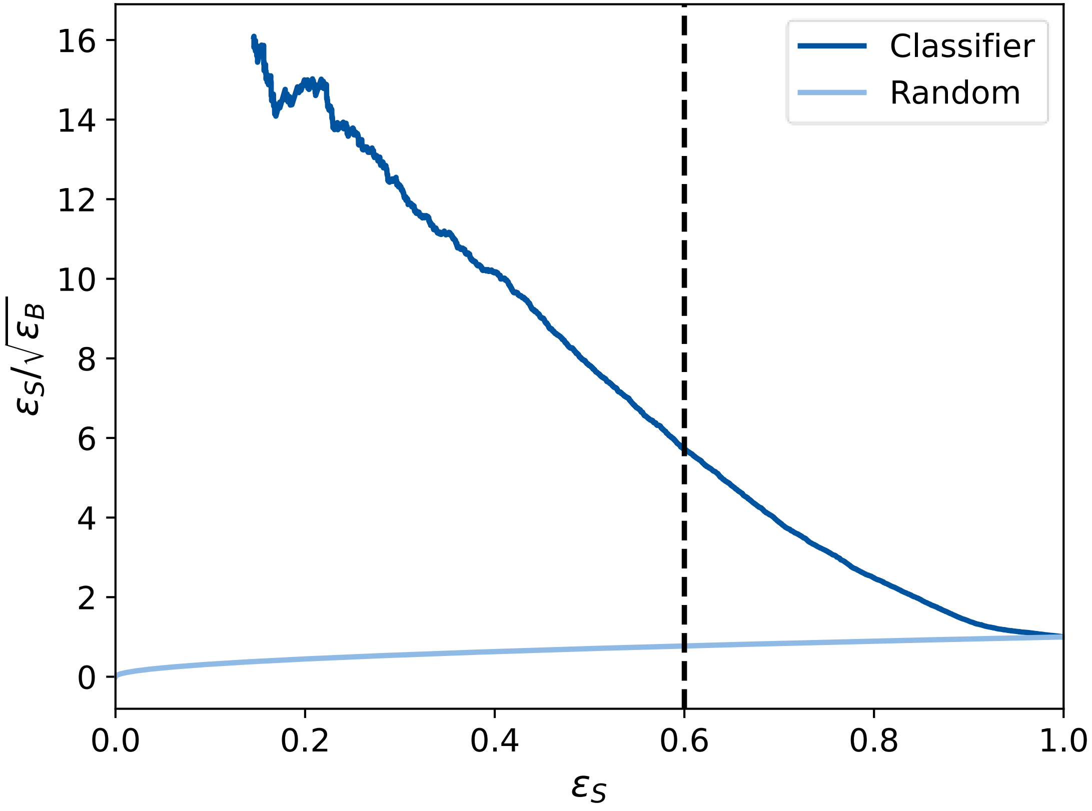


Classifier score to anomaly detection

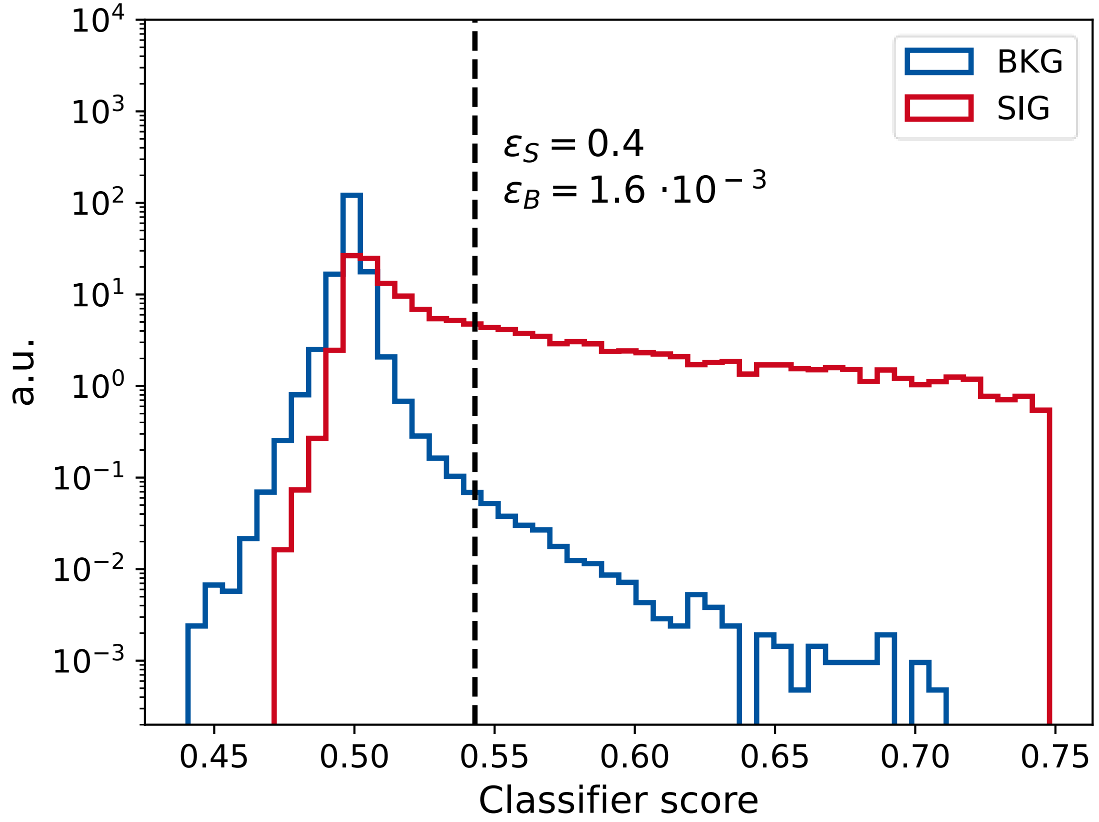


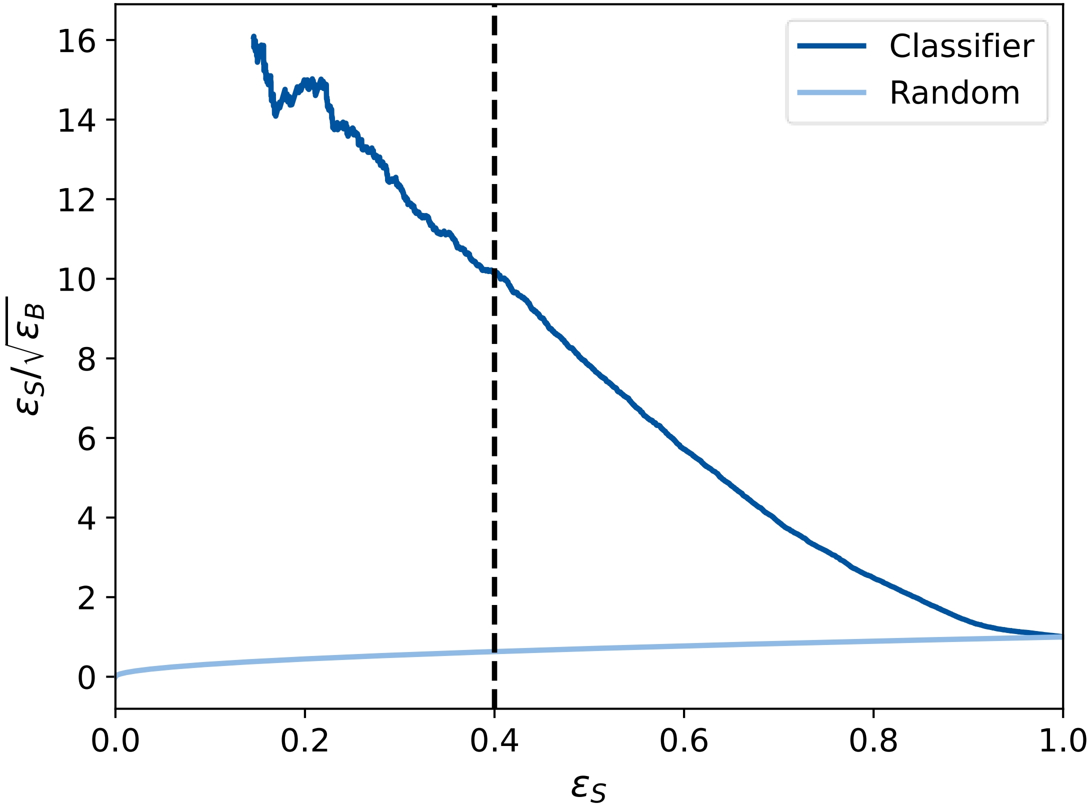
Classifier score to anomaly detection



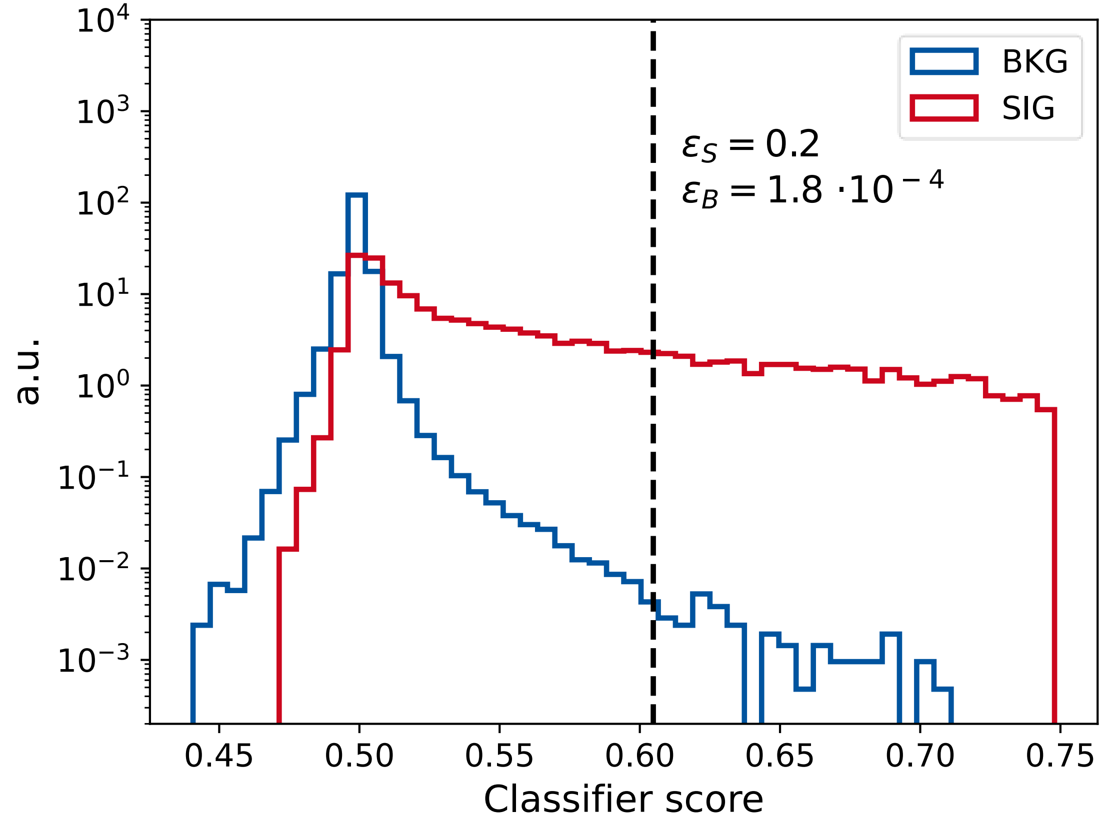


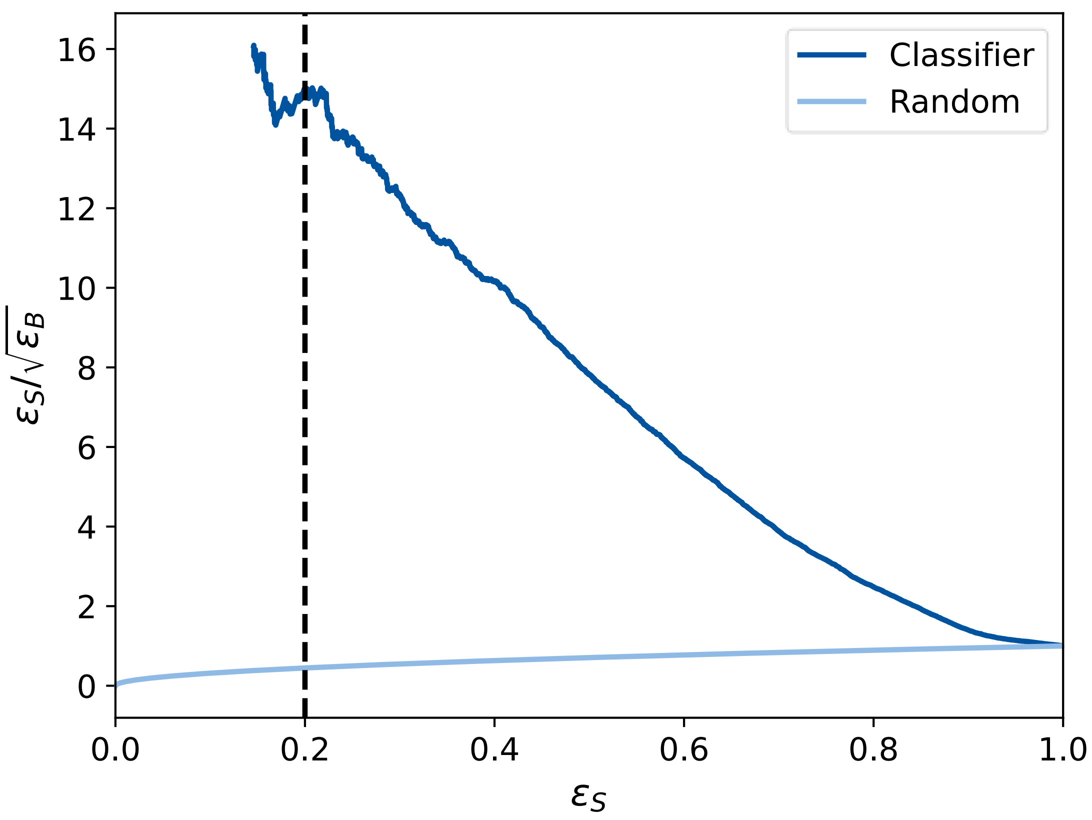
Classifier score to anomaly detection



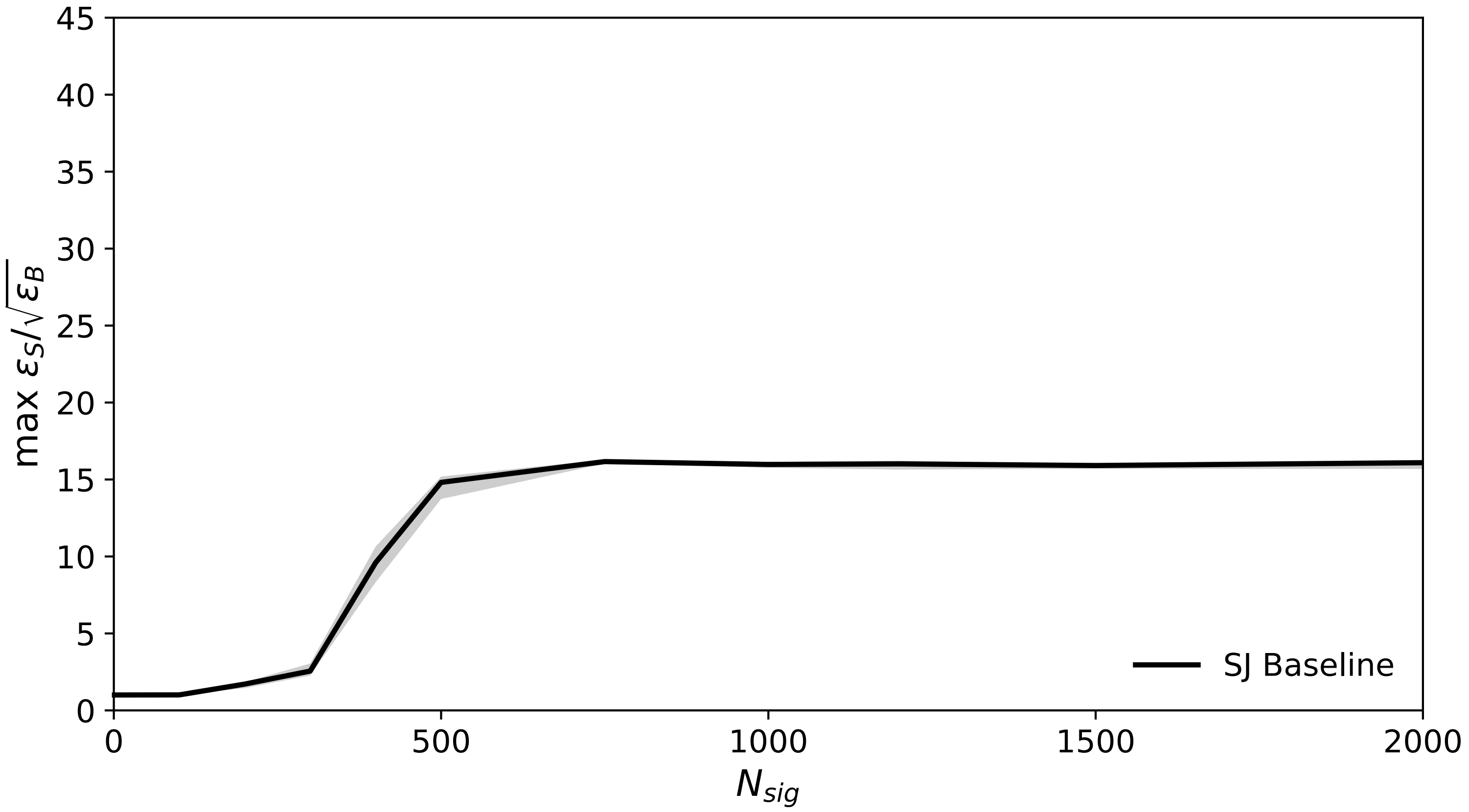


Classifier score to anomaly detection



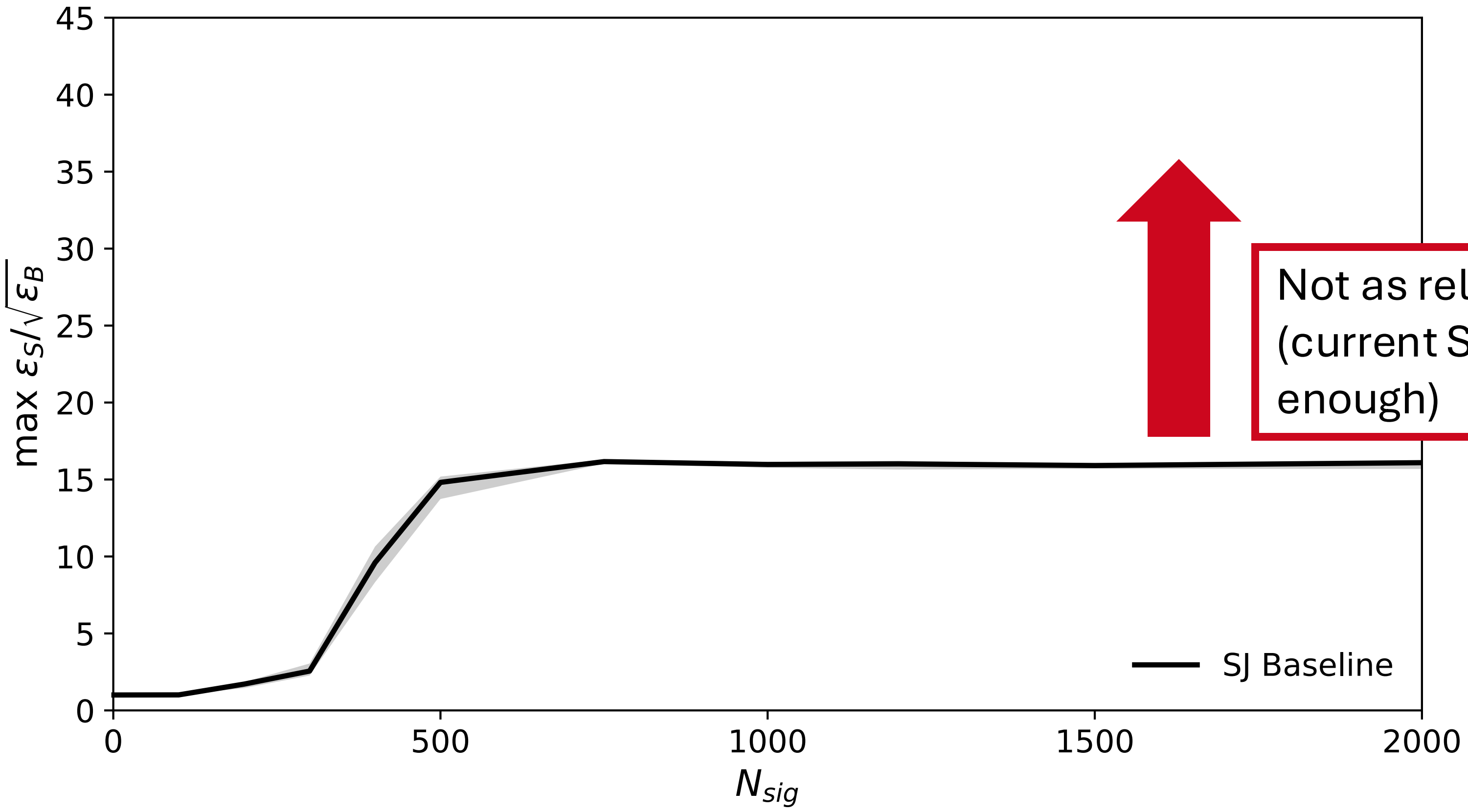


1D scan



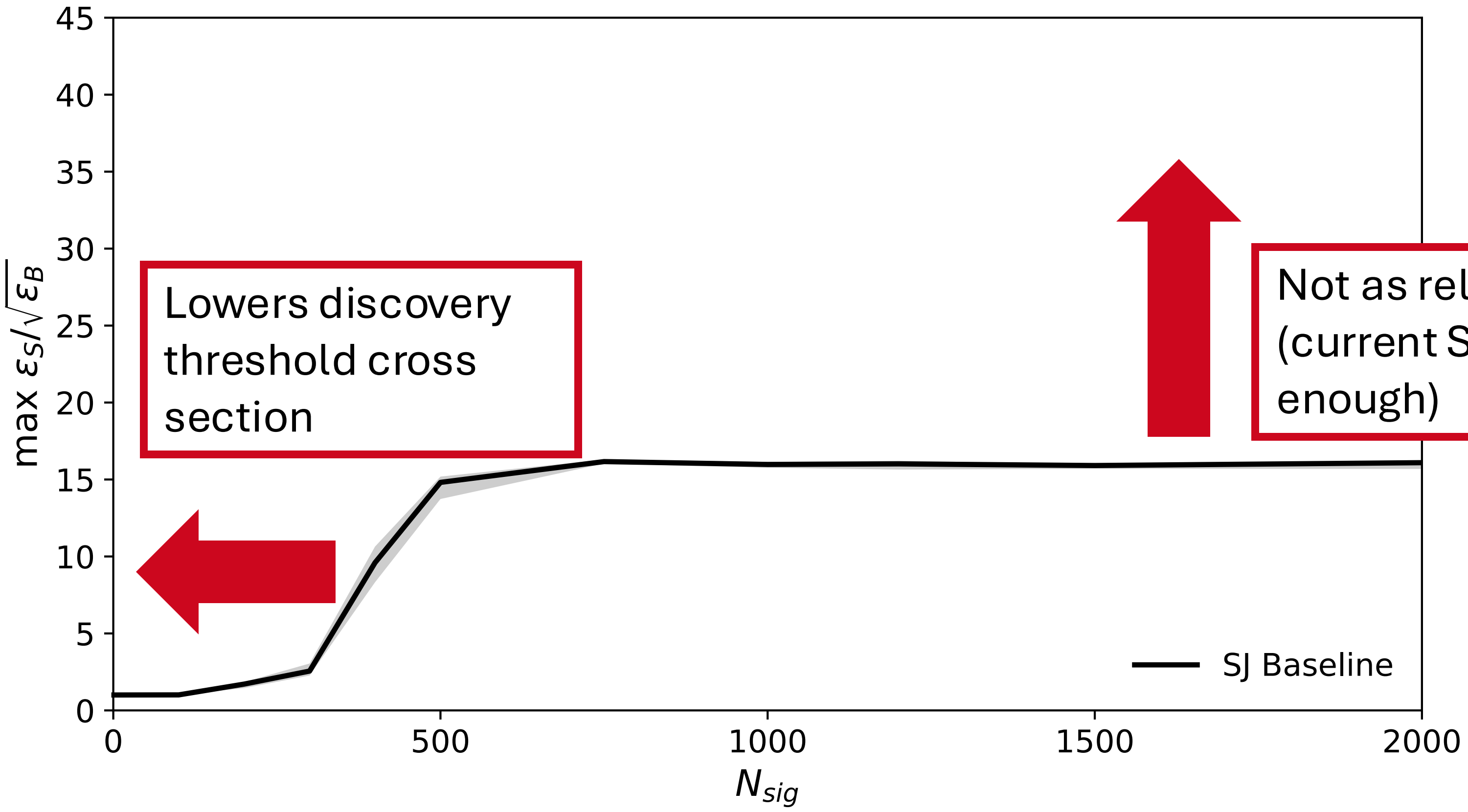
45
40
35
30
25
20
15
10
5
0
max $\varepsilon_S/\sqrt{\varepsilon_B}$
0
500
1000
1500
2000
N_{sig}
SJ Baseline

1D scan



Not as relevant (current SICs high enough)

1D scan



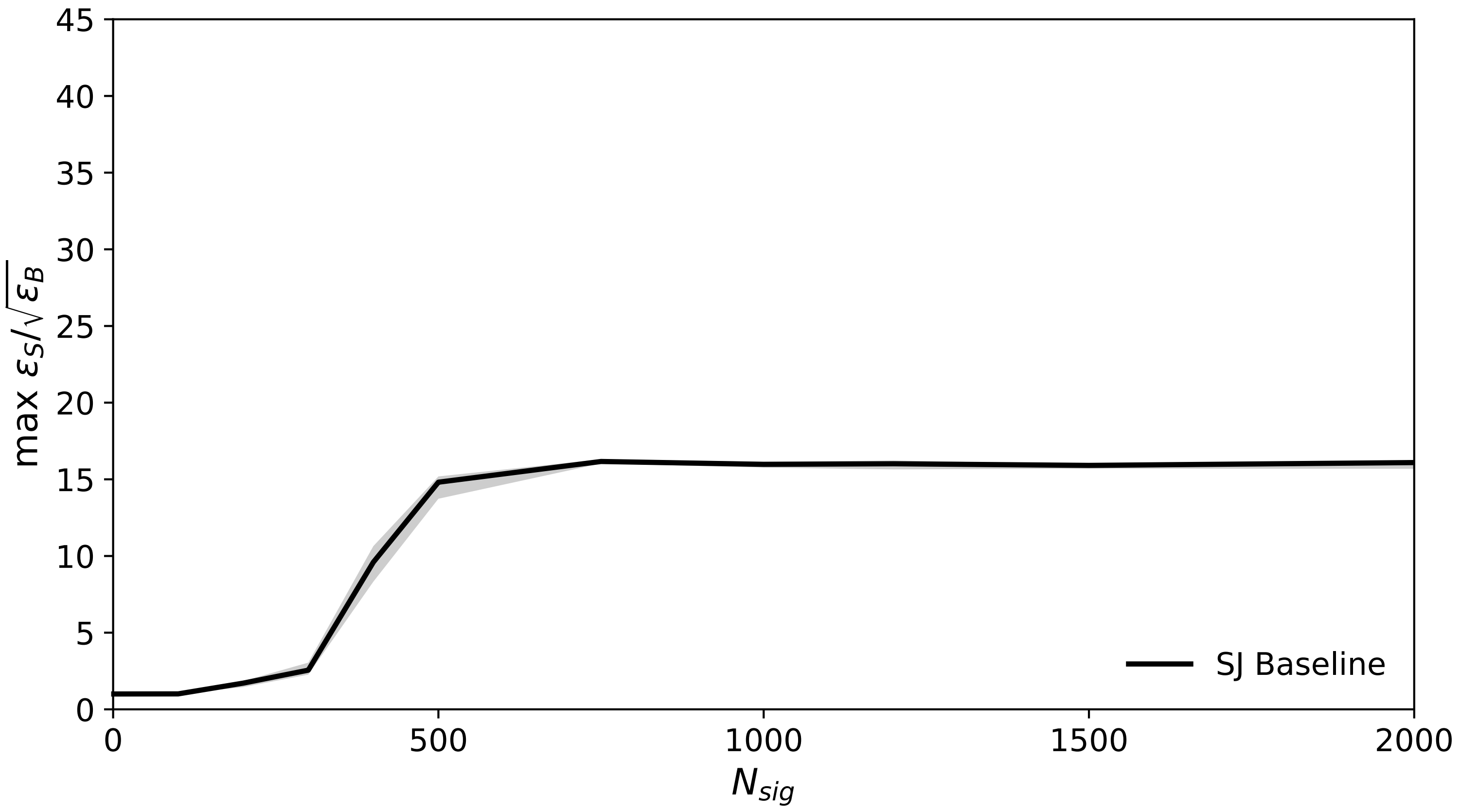
Lowers discovery threshold cross section

Not as relevant (current SICs high enough)

Results

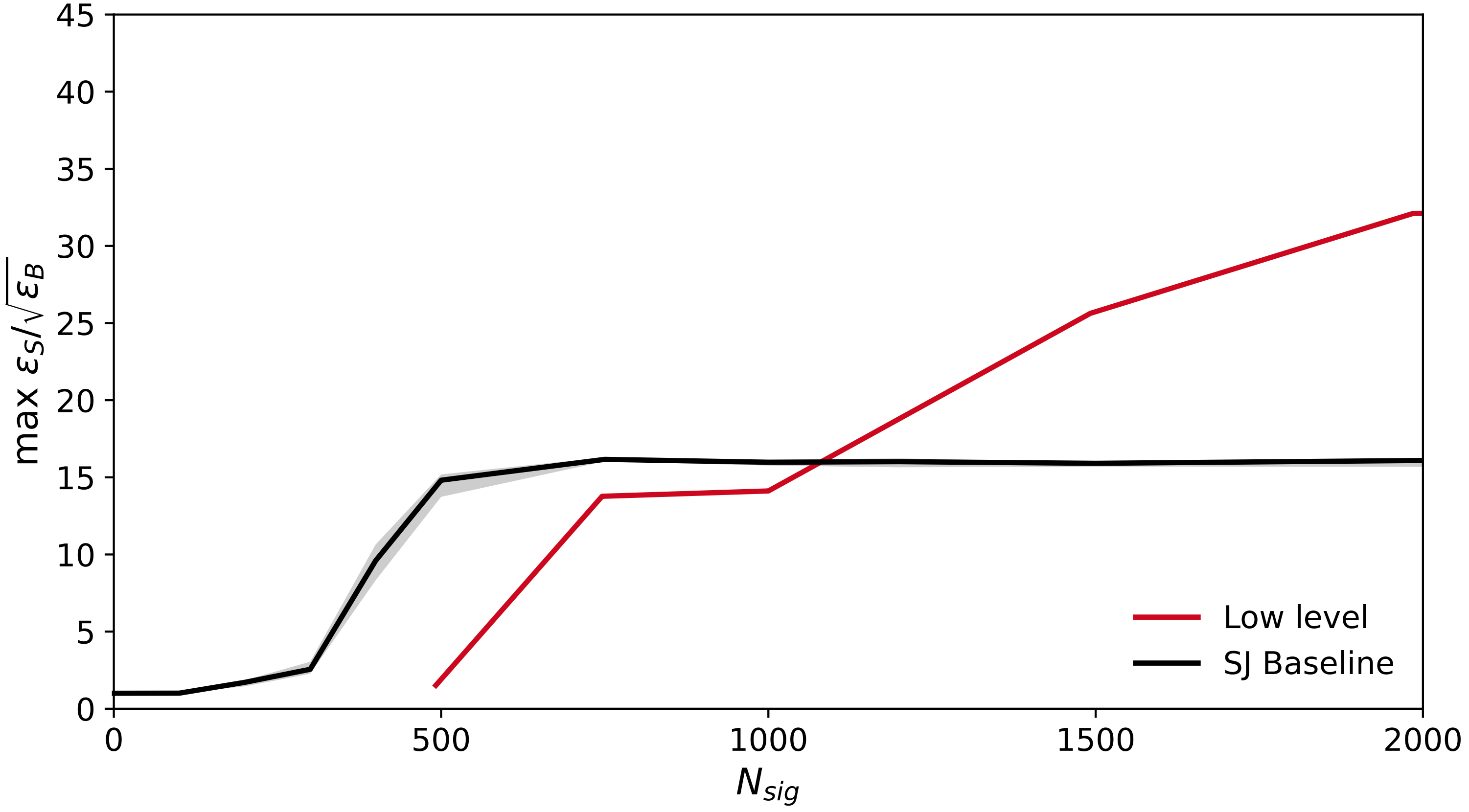
Comparison of Feature Sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.
"Identifying Anomalous Events Using Low-Level LHC Data", Master Thesis of Joep Geuskens (2023)
Master Thesis of Lukas Lang (2024)



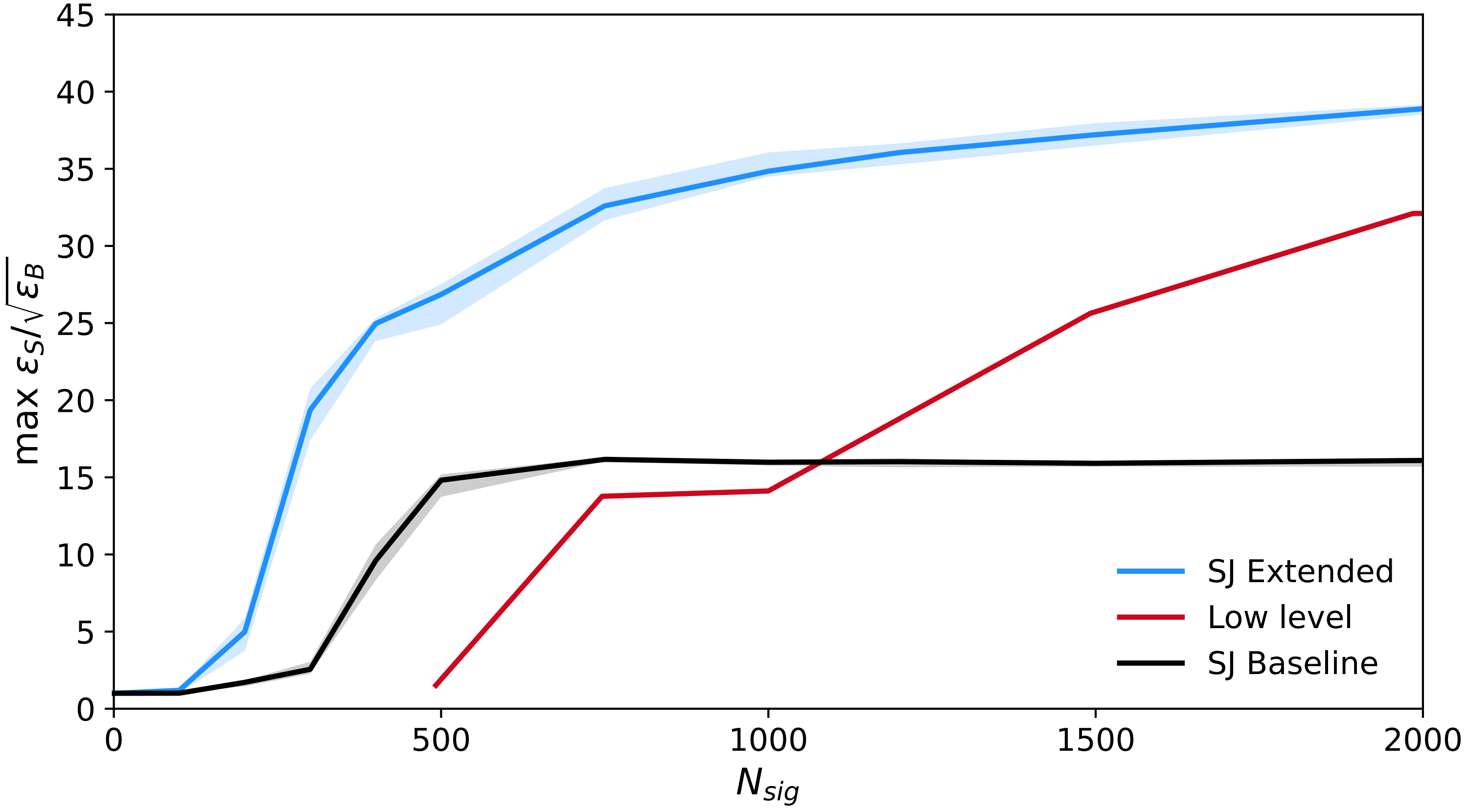
Comparison of Feature Sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.
"Identifying Anomalous Events Using Low-Level LHC Data", Master Thesis of Joep Geuskens (2023)
Master Thesis of Lukas Lang (2024)



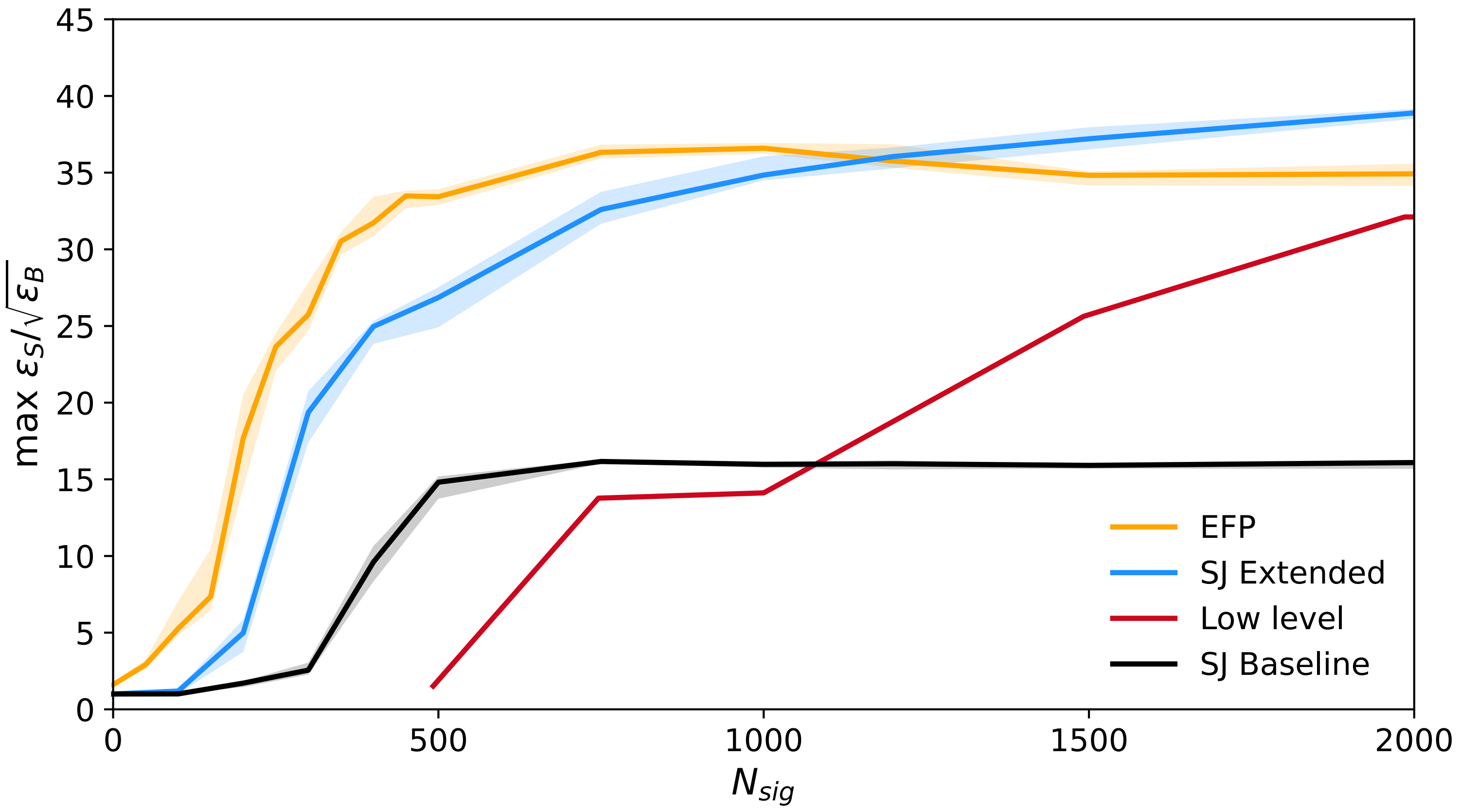
Comparison of Feature Sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.
"Identifying Anomalous Events Using Low-Level LHC Data", Master Thesis of Joep Geuskens (2023)
Master Thesis of Lukas Lang (2024)



Comparison of Feature Sets

"Tree-Based Algorithms for Weakly Supervised Anomaly Detection" [2309.13111], T. Finke, **MH** et. al.
"Identifying Anomalous Events Using Low-Level LHC Data", Master Thesis of Joep Geuskens (2023)
Master Thesis of Lukas Lang (2024)



Conclusion

Summary

- EFPs are useful for anomaly detection
- By choosing the right observables, we can…

  →be more model agnostic

  →be sensitive to lower signal cross sections

Outlook

- Understand why EFPs work so well

  →Currently using interpretable ML methods
- Test EFPs for other signal types

  →Currently working on semi-visible jet
- Test EFPs in more realistic setup