Institute for Theoretical Particle Physics and Cosmology
RWTH AACHEN UNIVERSITY

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

# Kitchen Sink Anomaly Detection

CRC Young Scientist Meeting 2026

arXiv:2604.20965

**Lukas Lang**

23 Sep 2026

In collaboration with R. Das, M. Hein, G. Kasieczka, M. Krämer, R. Mastandrea, L. Moureaux, A. Mück, D. Shih

# Anomaly Detection

- The Standard Model of Particle Physics (SM) is the best theory we have so far
- Open questions going beyond the Standard Model (BSM)
  - Dark matter
- Studying BSM is one of the missions of the Large Hadron Collider (LHC)

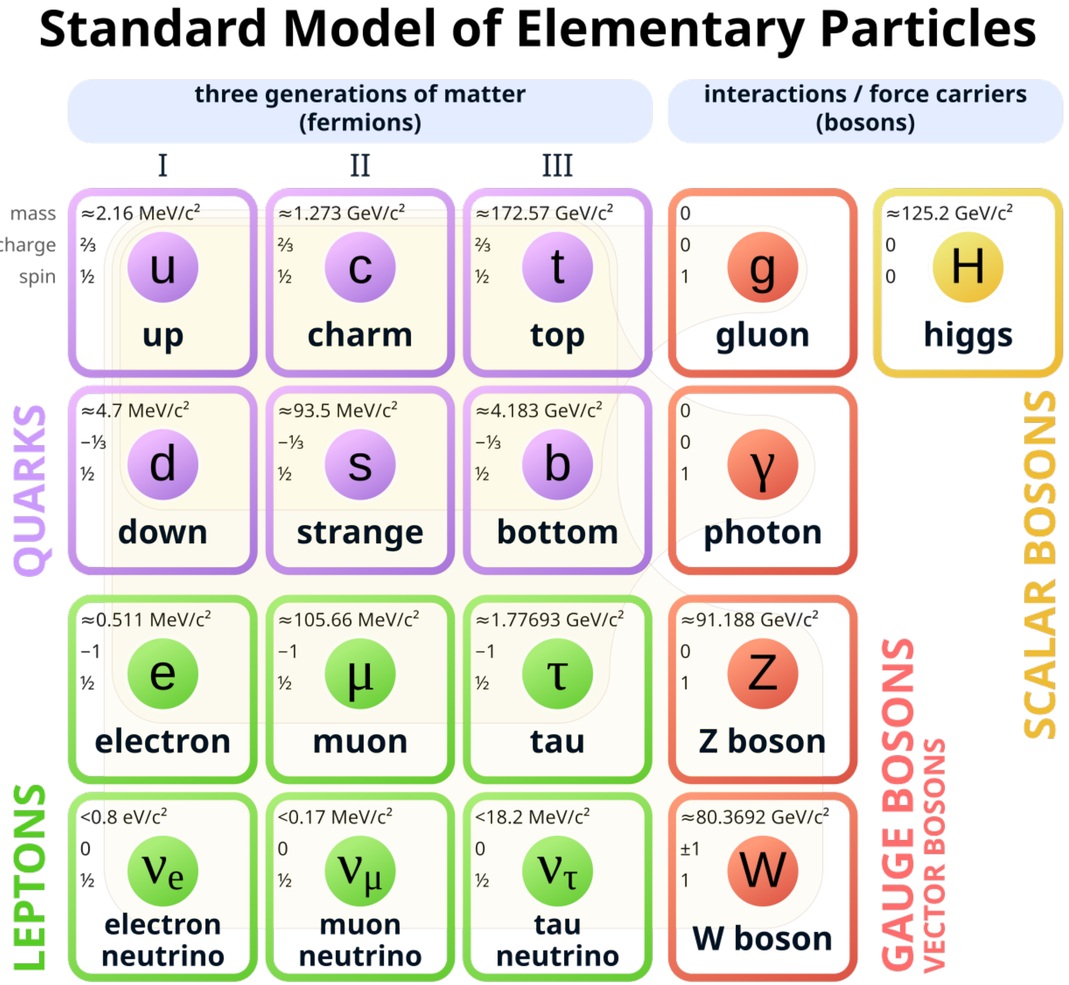
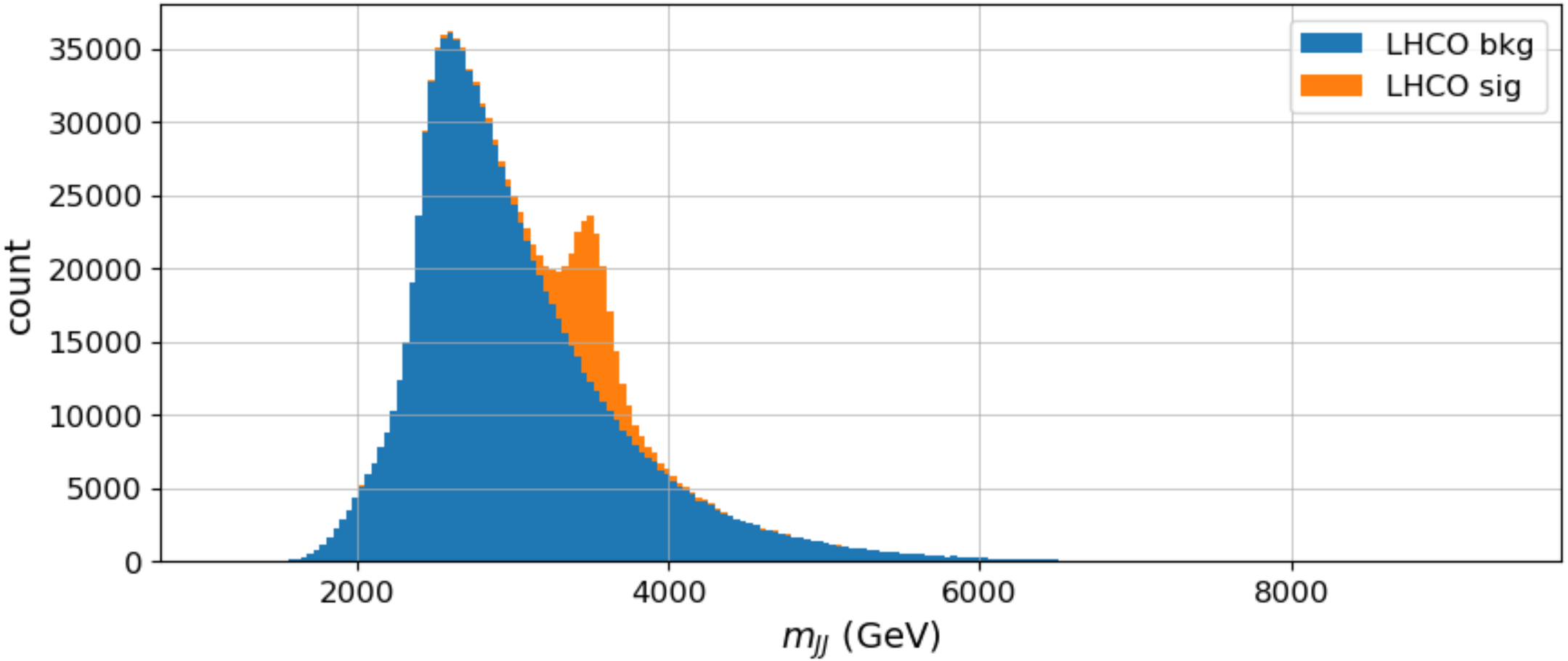


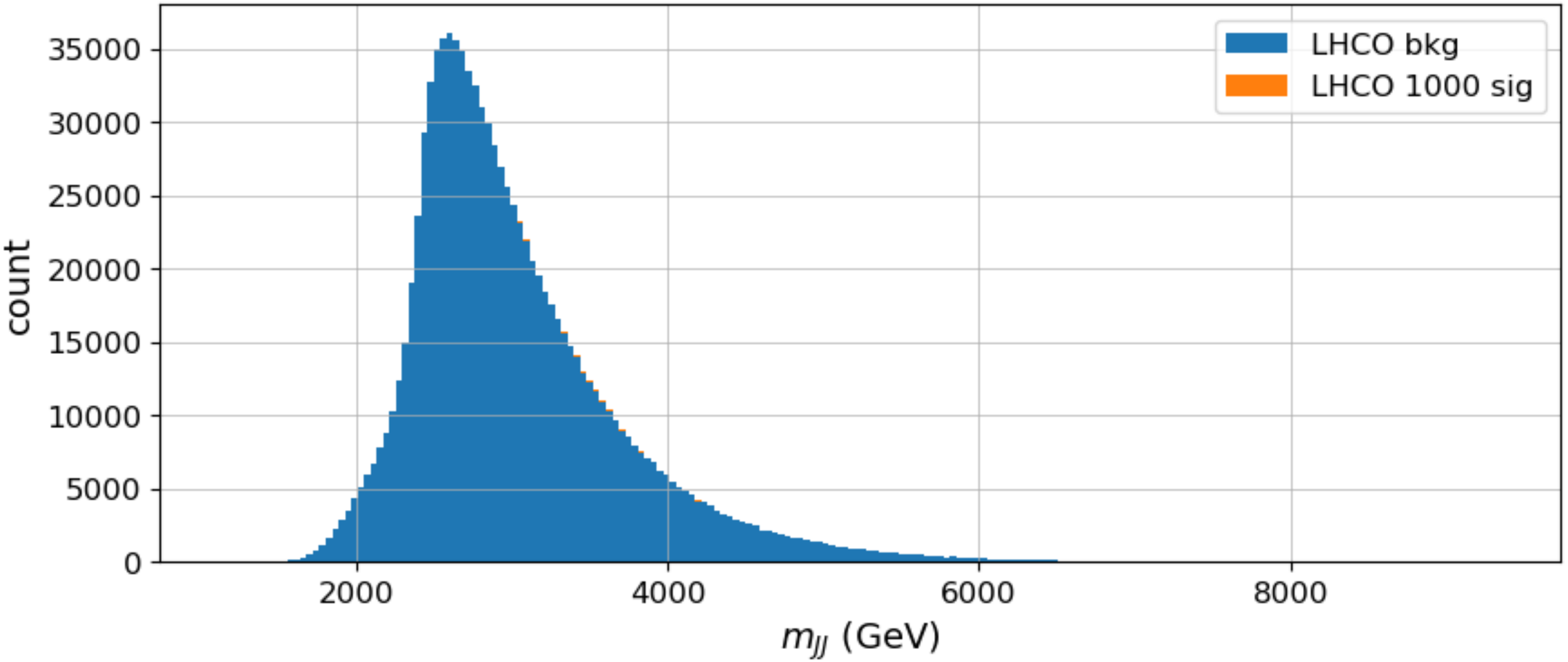

Image Wikipedia

# Searching for Resonances




[LHC Olympics 2020, arXiv:2109.00546]

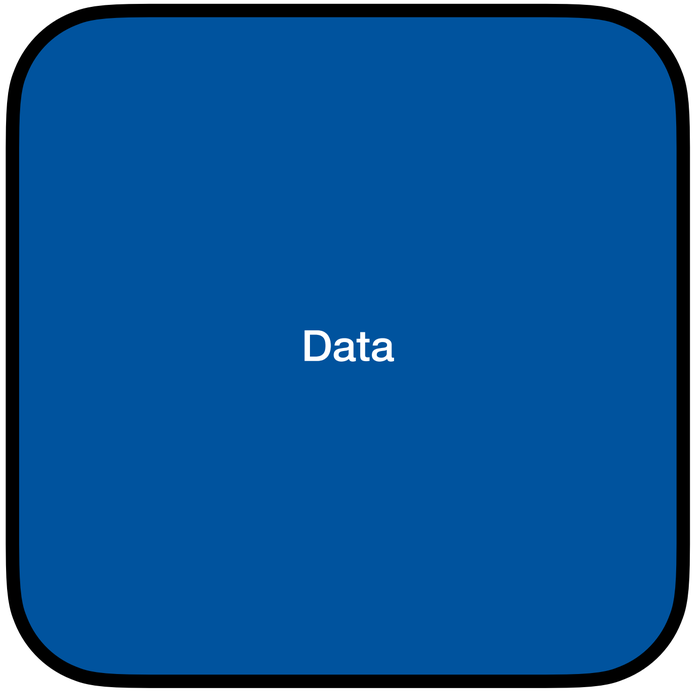
# Searching for Resonances



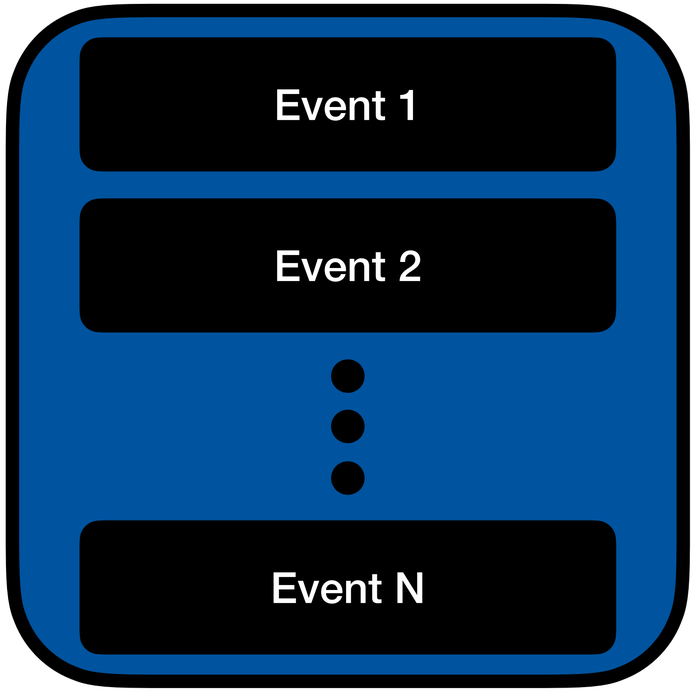

[LHC Olympics 2020, arXiv:2109.00546]

# Machine Learning

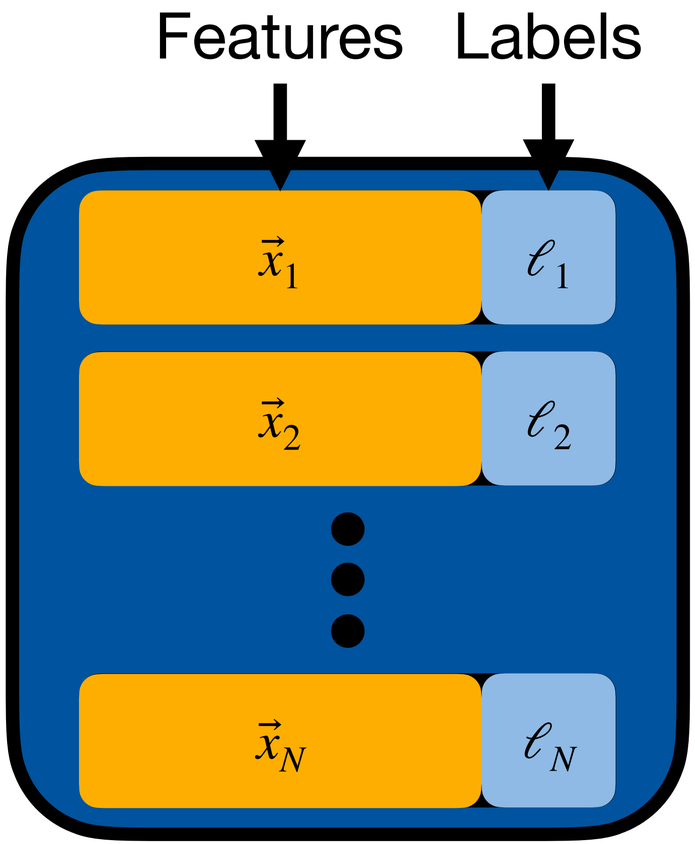
# Supervised Classification



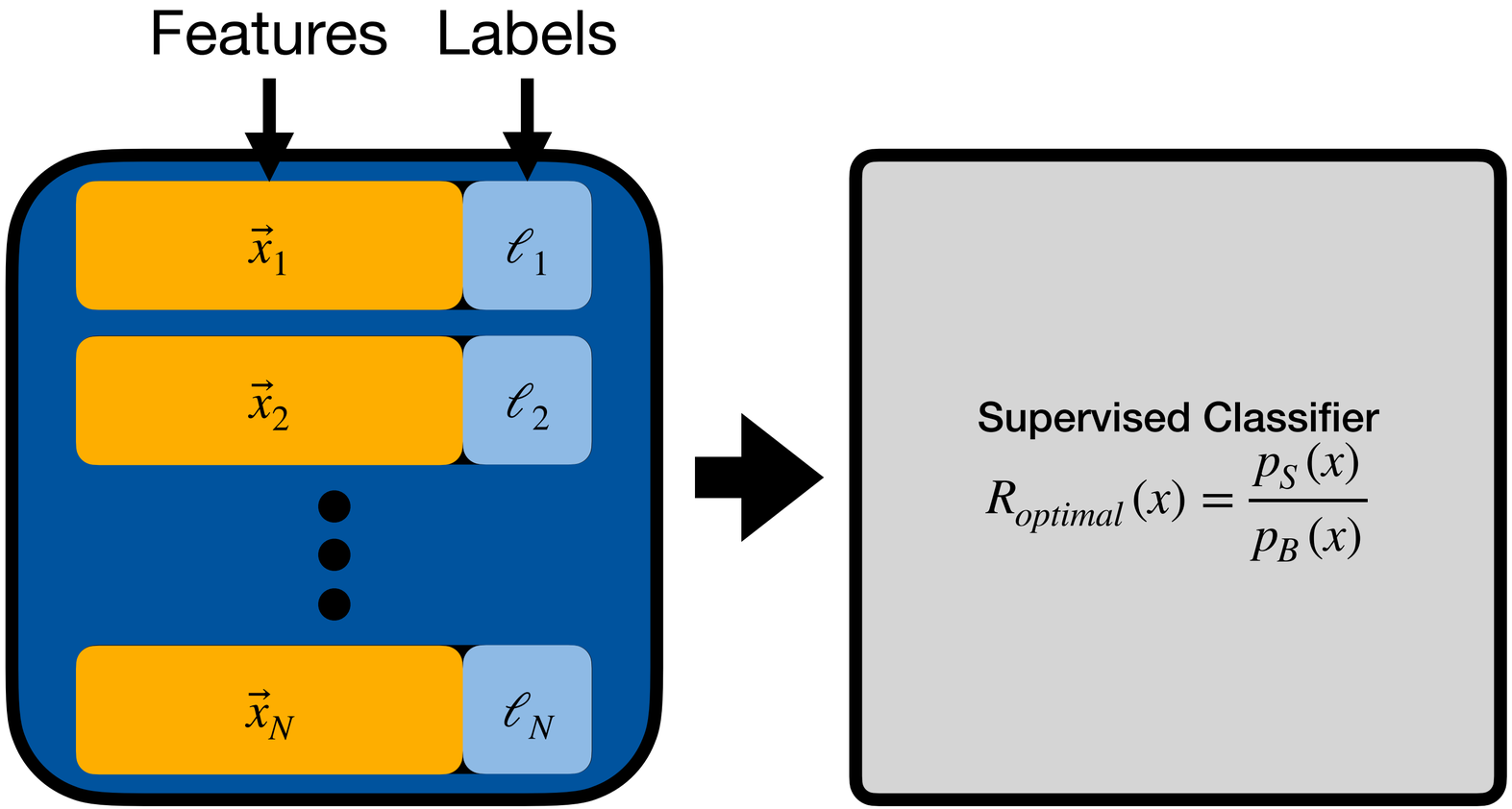
# Supervised Classification



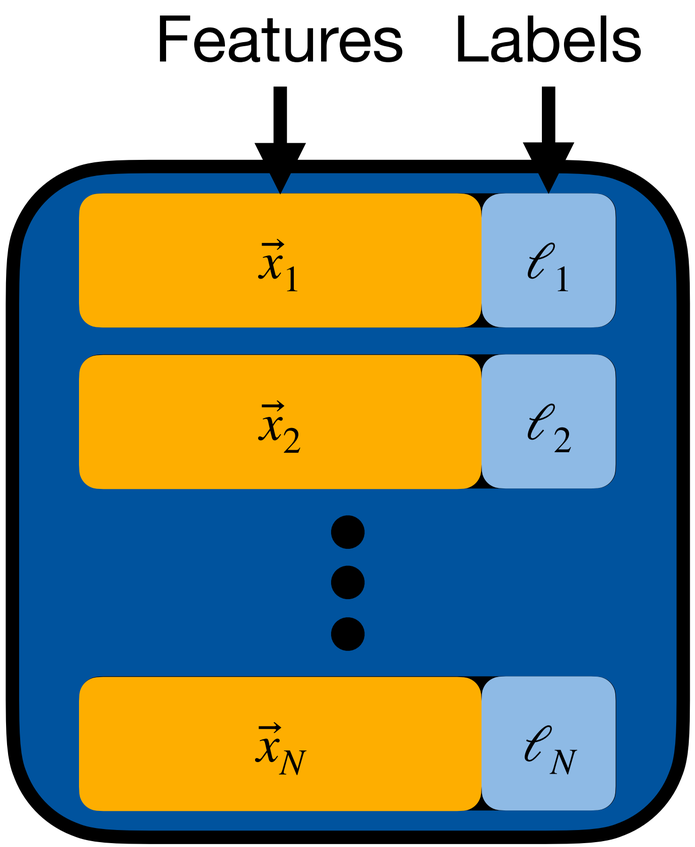
# Supervised Classification



# Supervised Classification

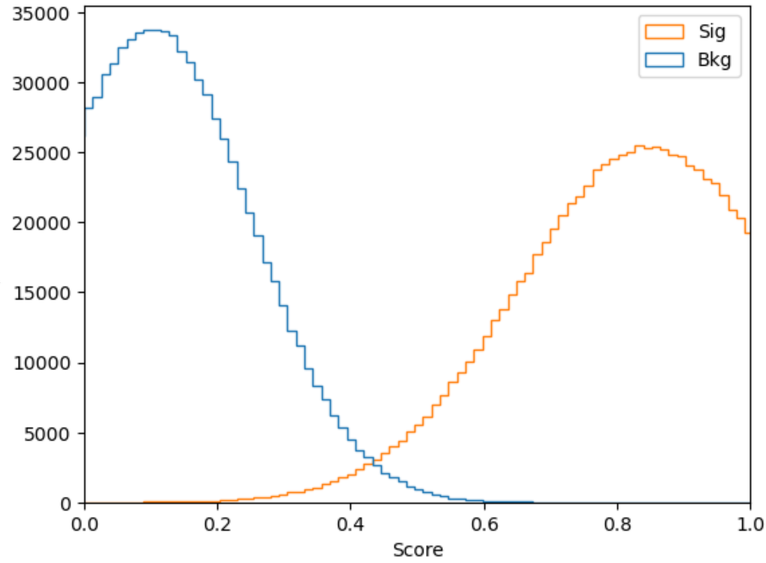


# Supervised Classification



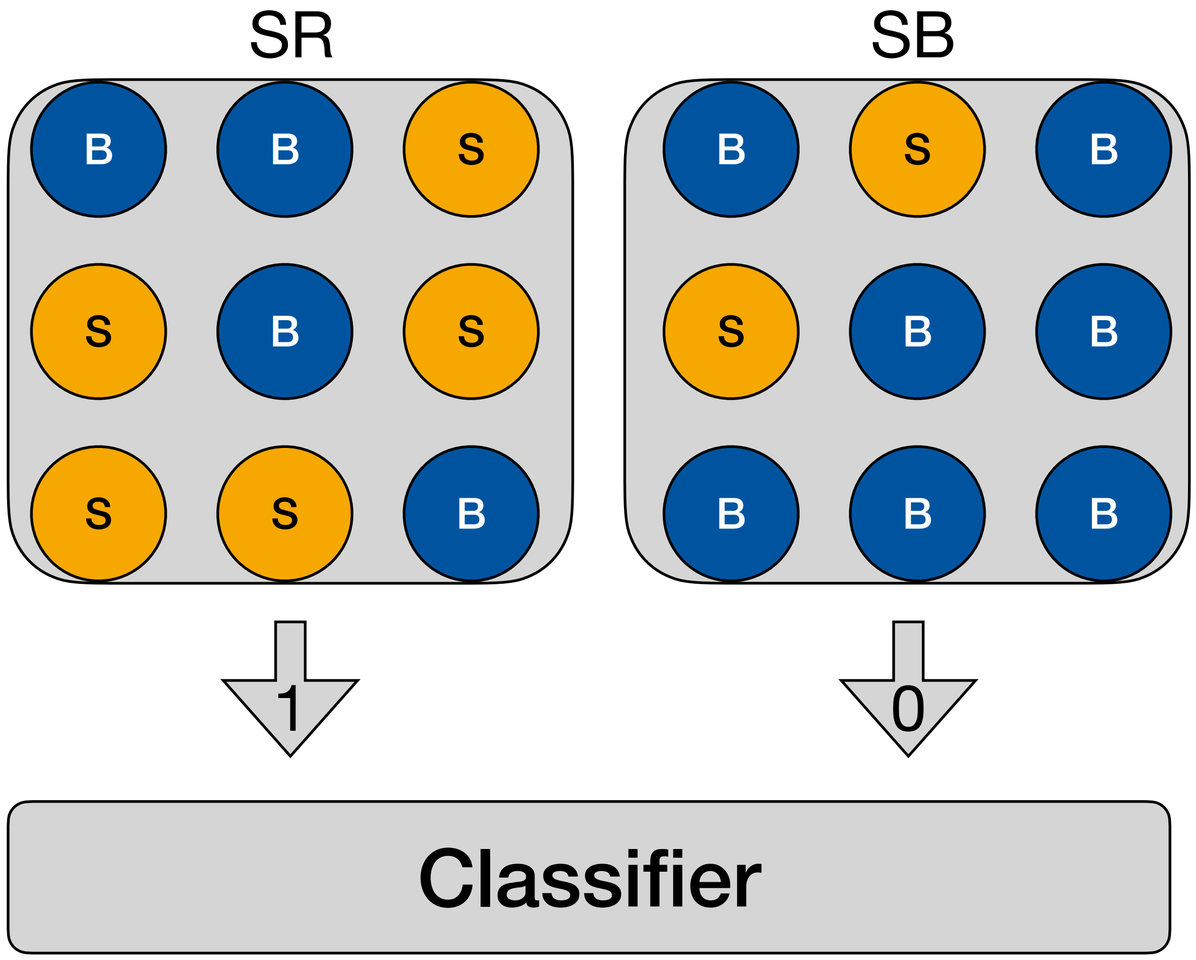

**Supervised Classifier**

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



# Nature does not Label Data

# Method: Clarification Without Labels (CWoLa)




Here ensemble of 50 Boosted Decision Trees

- Any monotonic function of the optimal classifier has the same decision boundaries as the optimal classifier

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

- Split into signal-enriched and background-enriched data set

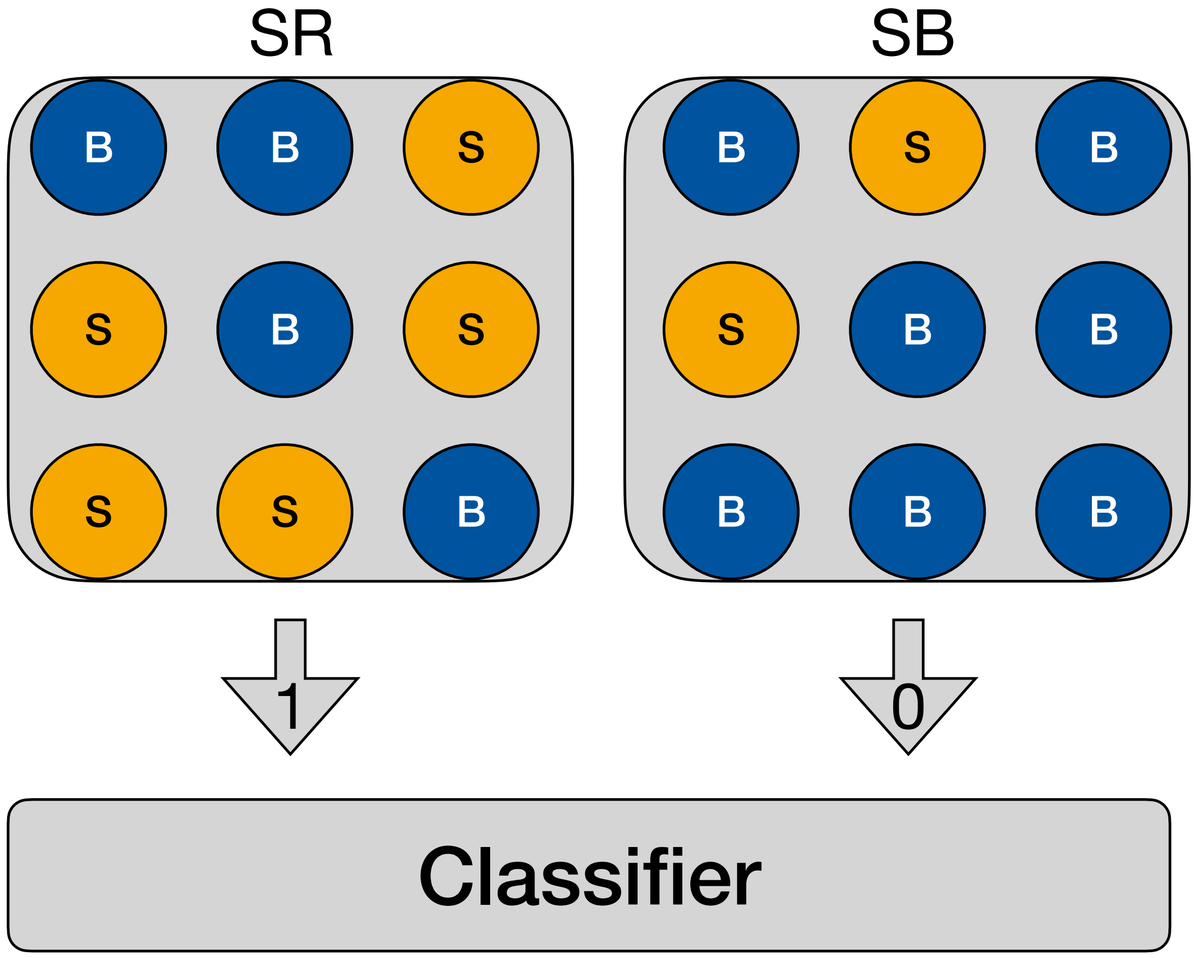
$$p_i(x) = f_i\, p_S(x) + (1 - f_i)\, p_B(x)$$

$$R_{CWoLa}(x) = \frac{p_1(x)}{p_2(x)} = \frac{f_1\, R_{optimal}(x) + (1 - f_1)}{f_2\, R_{optimal}(x) + (1 - f_2)}$$

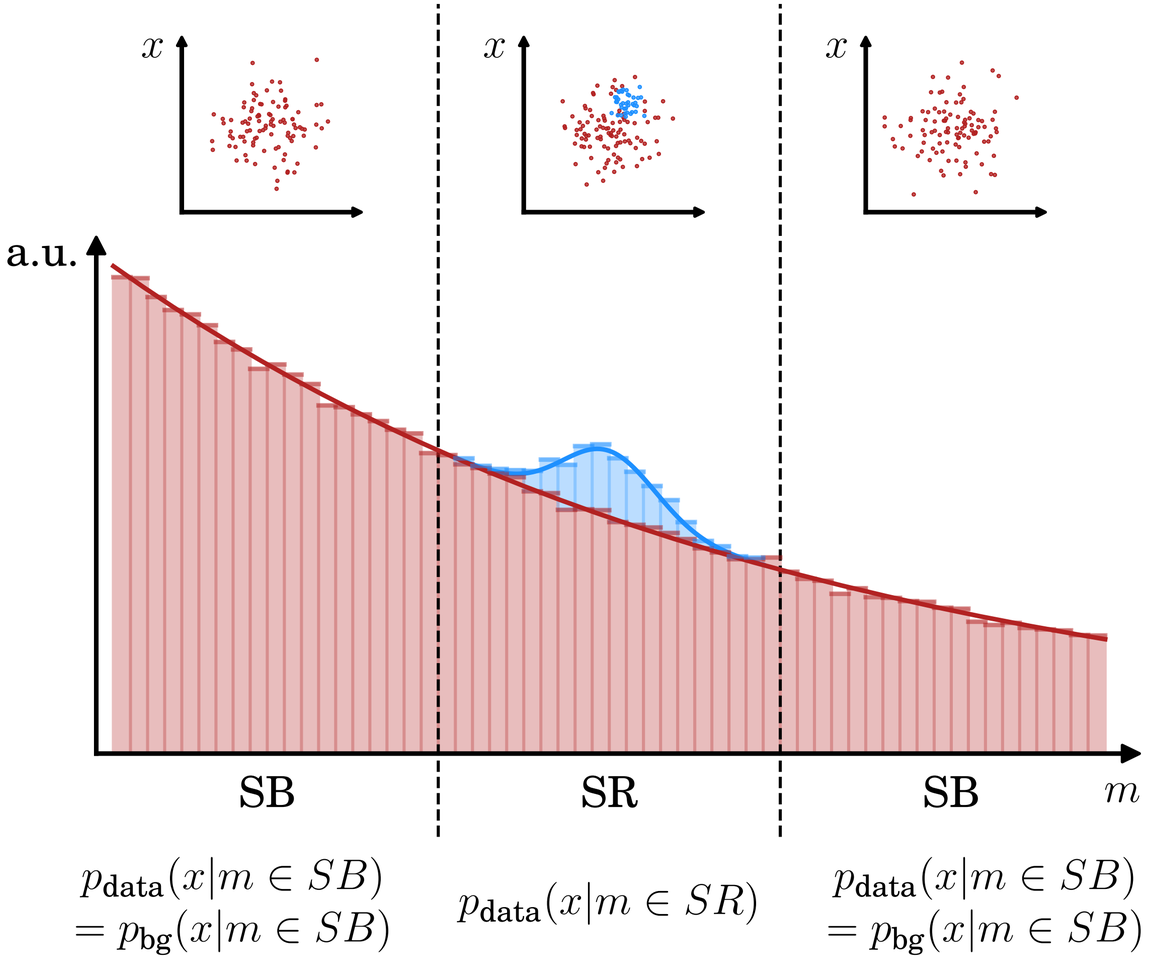
- Is monotonic for $f_1 > f_2$

[arXiv:1708.02949, arXiv:2109.00546, arXiv:1805.02664]

# Method: CWoLa Hunting

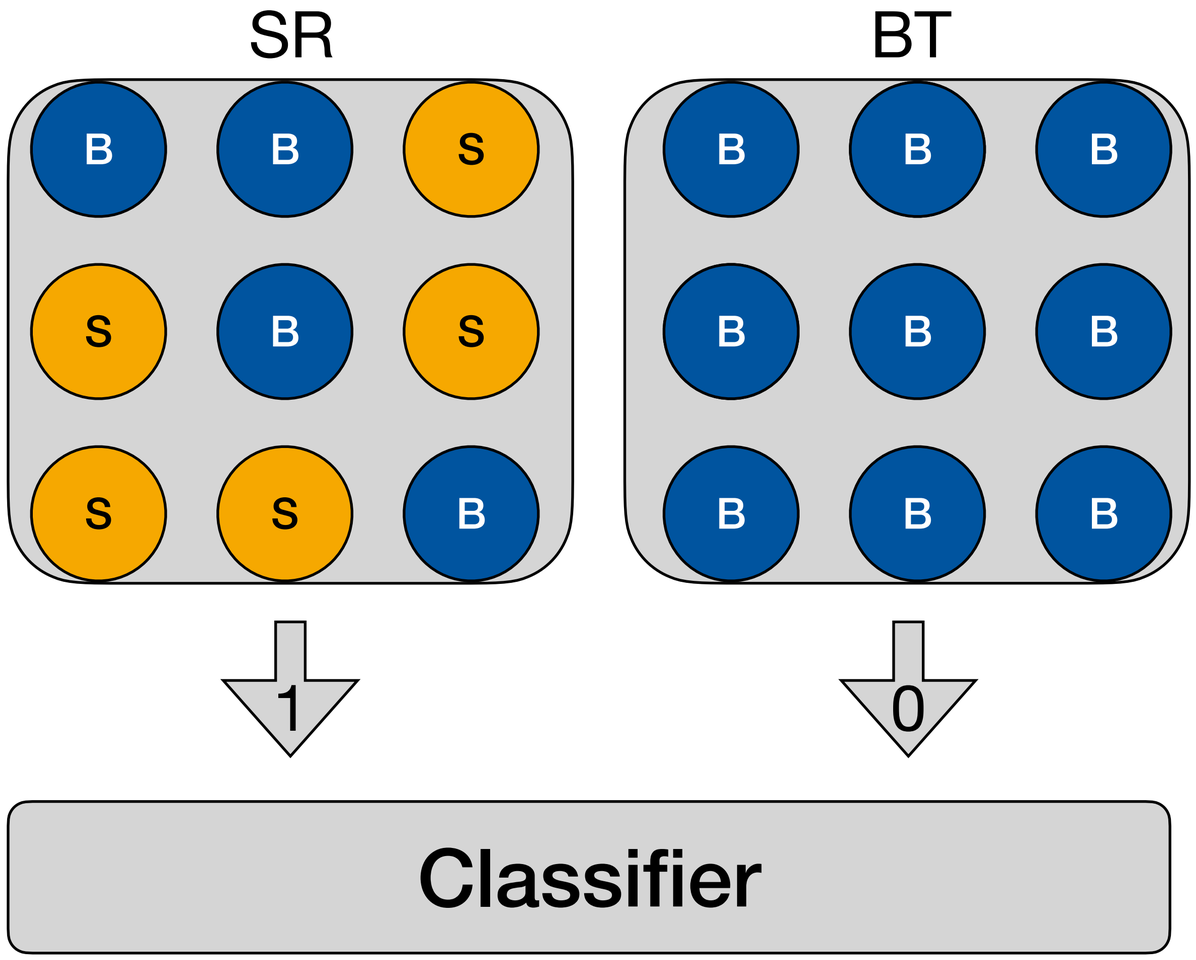



Here ensemble of 50 Boosted Decision Trees

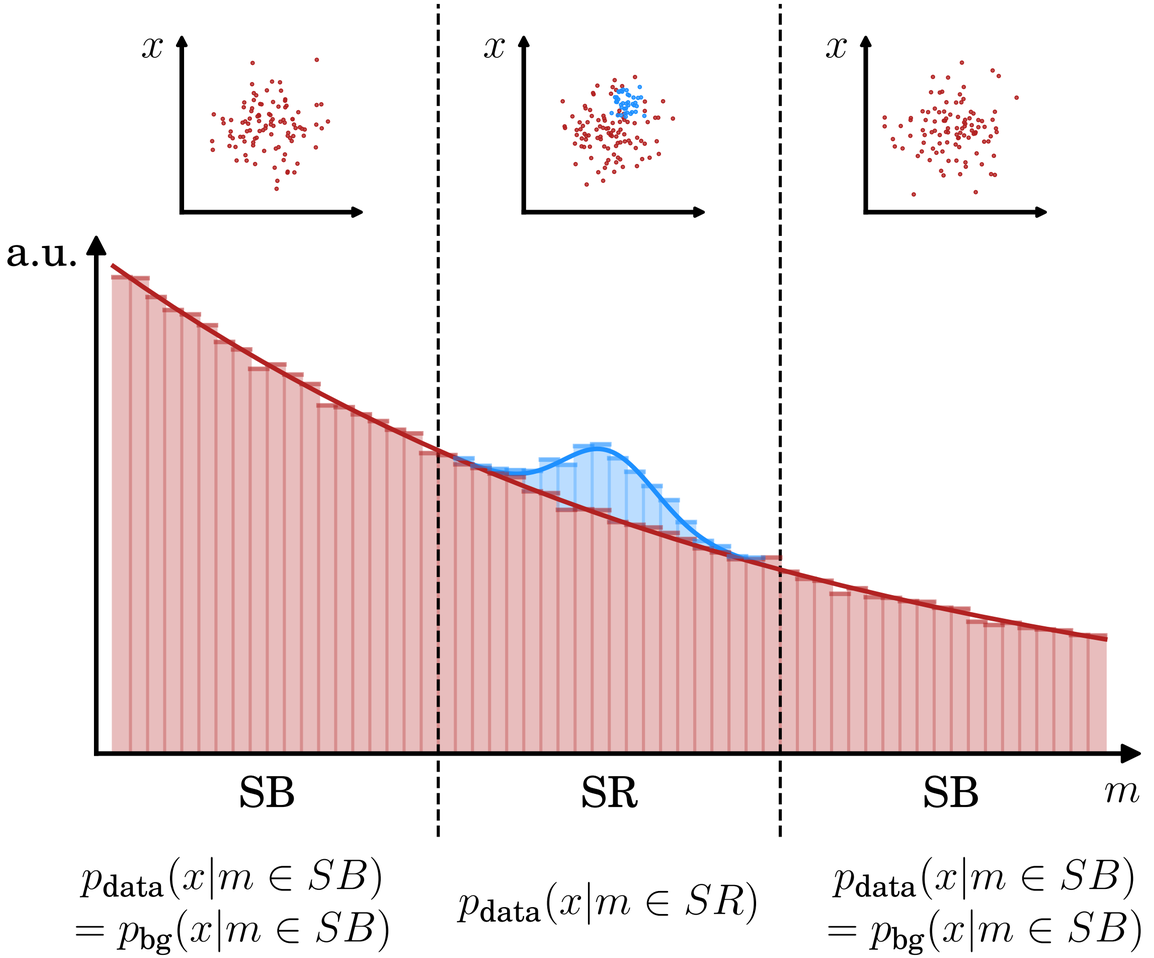



[arXiv:1708.02949, arXiv:2109.00546, arXiv:1805.02664]

# Method: Idealized Anomaly Detector (IAD)




Here ensemble of 50 Boosted Decision Trees

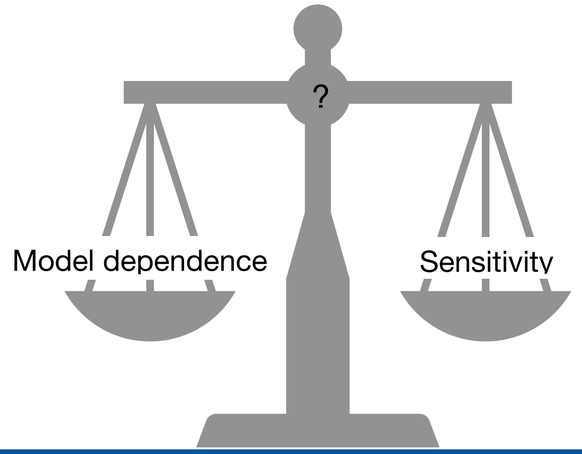



[arXiv:1708.02949, arXiv:2109.00546, arXiv:1805.02664]

# Model Agnostic Anomaly Detection

**Two Goals:**

- Model Agnostic:
  As little assumption on the signal as possible
- High Sensitivity:
  Find signals with low cross-section



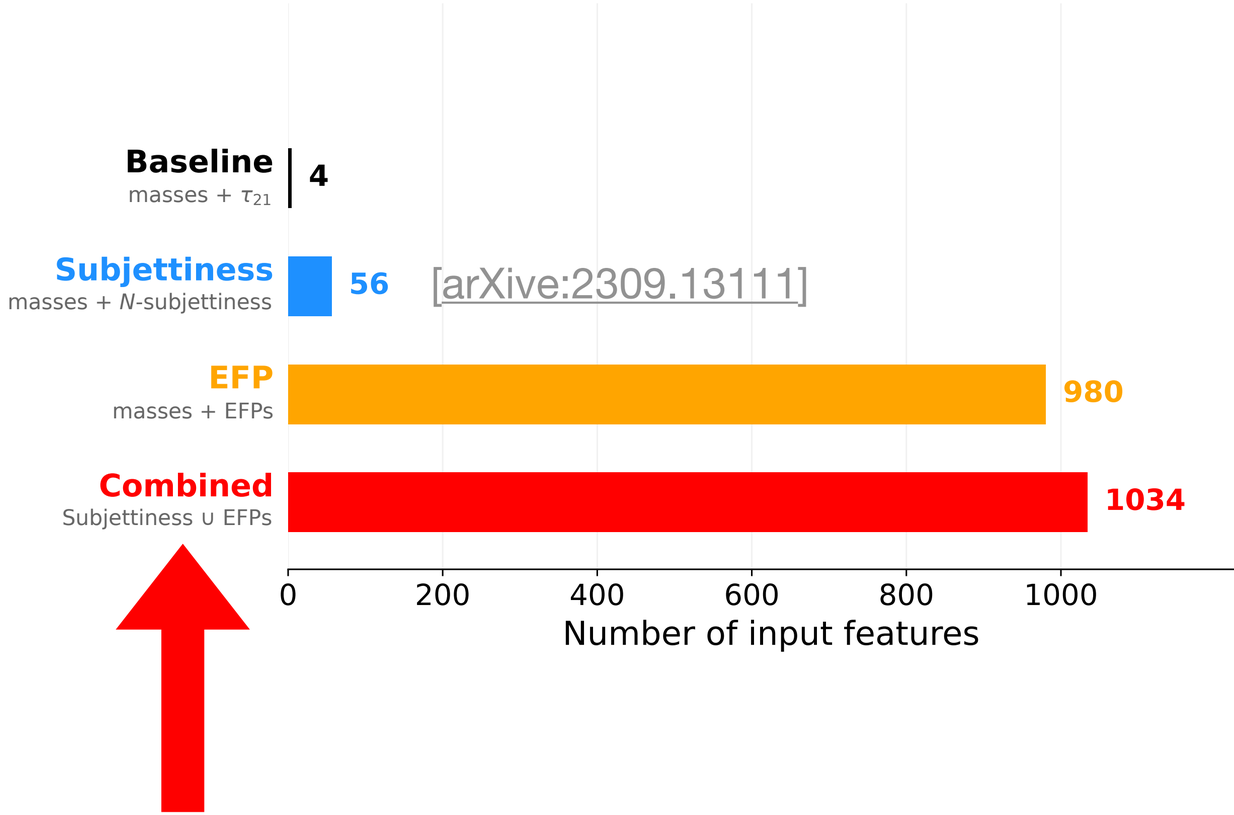

**Where do we make assumptions?**

- Method
  - Weakly supervised anomaly detection → contracting a background template (BT)
- Features
  - Choosing specific observables

# More Features to be More Model Agnostic

**Different families of features**

- Jet masses
- Substructure observables
  - N-Subjettiness
  - Energy Flow Polynomials (EFPs)




**The "kitchen sink" approach:**

➡ Use all the features you have

# Data: Six Signals, One Background

**Background:**

- Taken from the LHC Olympics R&D Dataset [arXiv:2101.08320]

  ➡ $10^6$ QCD dijet events + $\sim 6 \cdot 10^5$ QCD dijet events in SR

**Signals:**

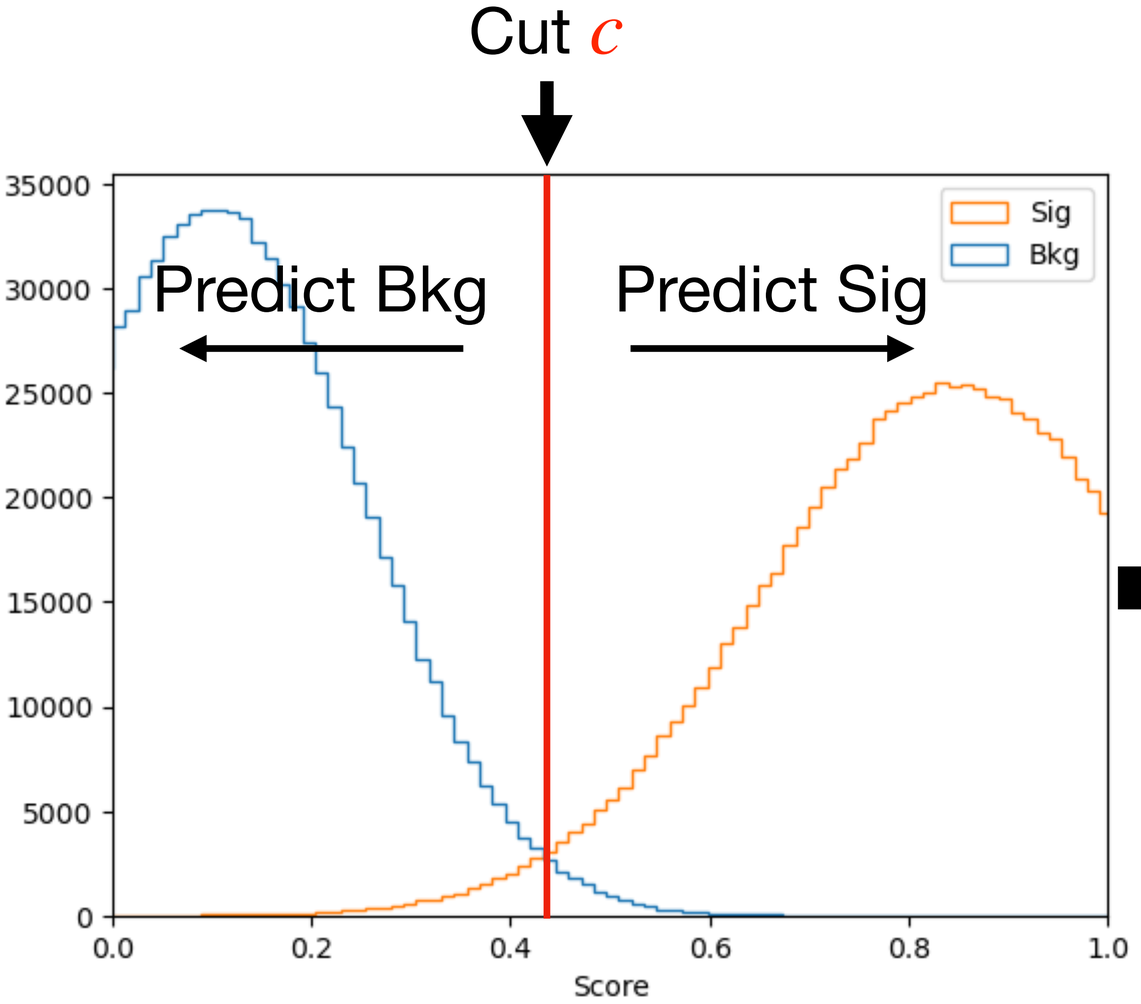
- Six different resonance signals inspired by [arXiv:2412.03747]

  ➡ Hadronic $A \to BC$ decays with $m_A = 3.5$ TeV

| Name | Prongs | Masses |
|---|---|---|
| LHCO 2-prong | $2+2$ | $m_Y = 100\,\text{GeV}$, $m_X = 500\,\text{GeV}$ |
| LHCO 3-prong | $3+3$ | $m_Y = 100\,\text{GeV}$, $m_X = 500\,\text{GeV}$ |
| $X \to YY' \to 4q$ | $2+2$ | $m_Y = 100\,\text{GeV}$, $m_{Y'} = 500\,\text{GeV}$ |
| $W_{KK} \to WR \to 3W$ | $2+4$ | $m_R = 500\,\text{GeV}$ |
| $Z' \to T'T' \to tZtZ$ | $5+5$ | $m_{T'} = 400\,\text{GeV}$ |
| $G_{KK} \to HH \to 4t$ | $6+6$ | $m_H = 400\,\text{GeV}$ |

# Metric

# Who doe we Know if a Classifier is Good?



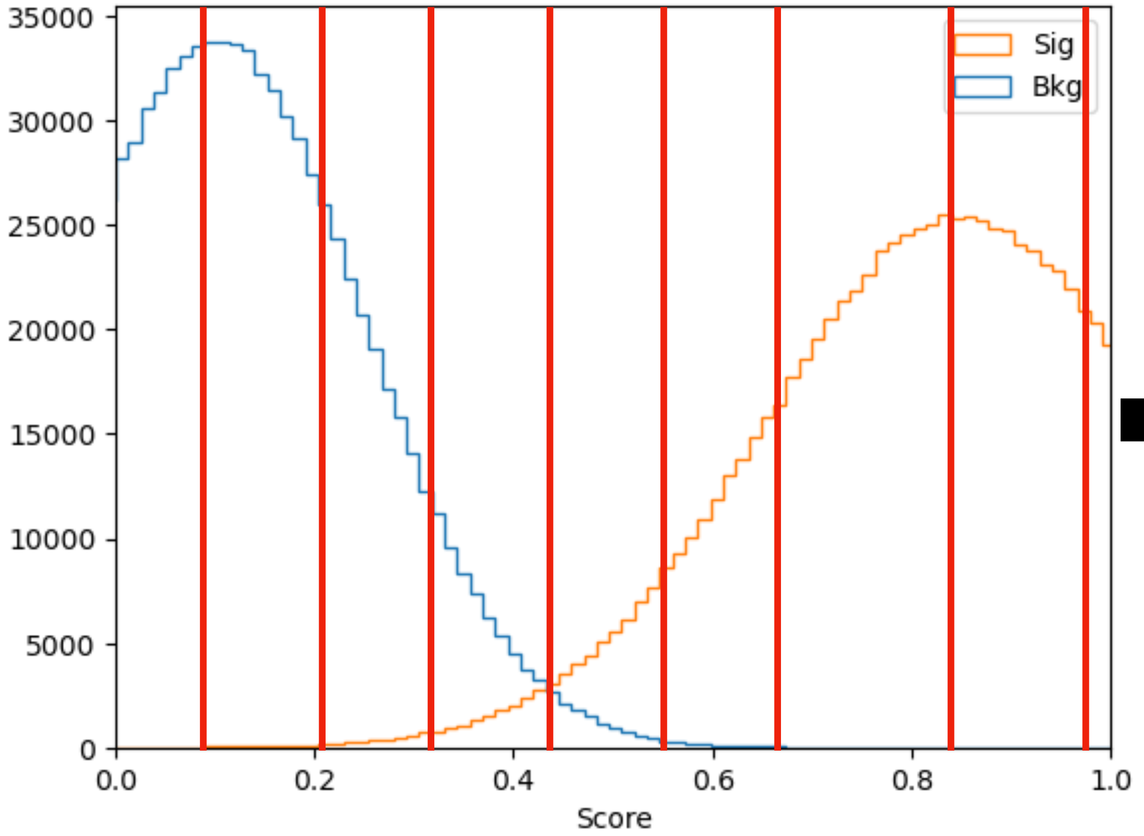

## Signal efficiency

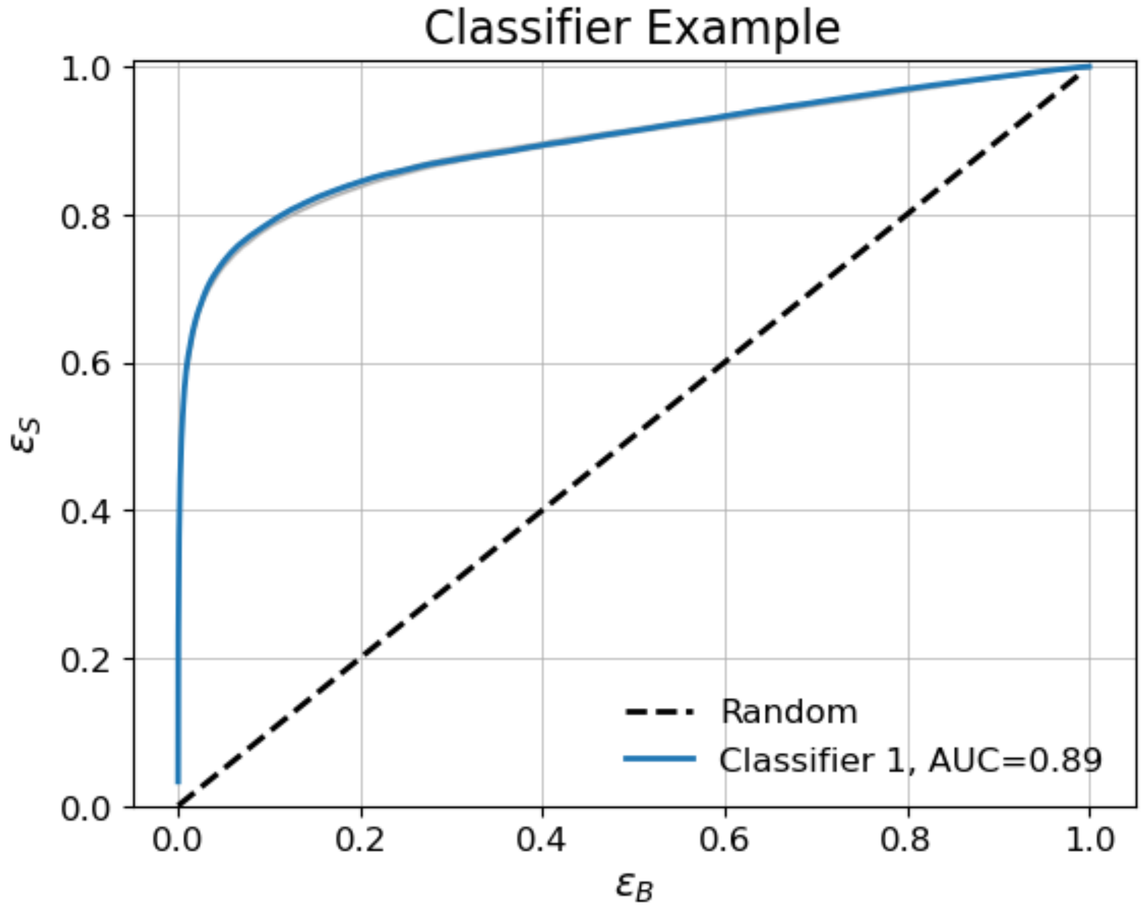
$$\epsilon_S = \int p_S(\vec{x})\Theta(h(\vec{x})-c)d\vec{x} = \frac{TP}{TP+FN}$$

## Background efficiency

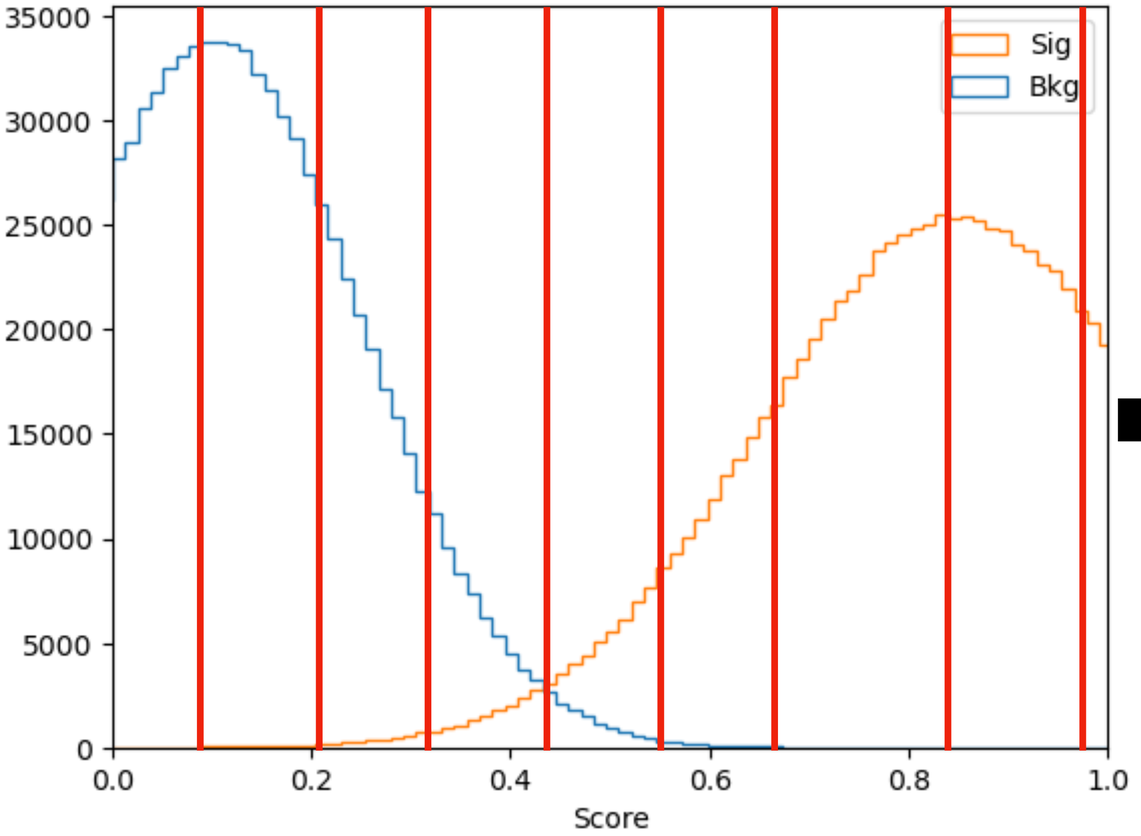
$$\epsilon_B = \int p_B(\vec{x})\Theta(h(\vec{x})-c)d\vec{x} = \frac{FP}{TN+FP}$$

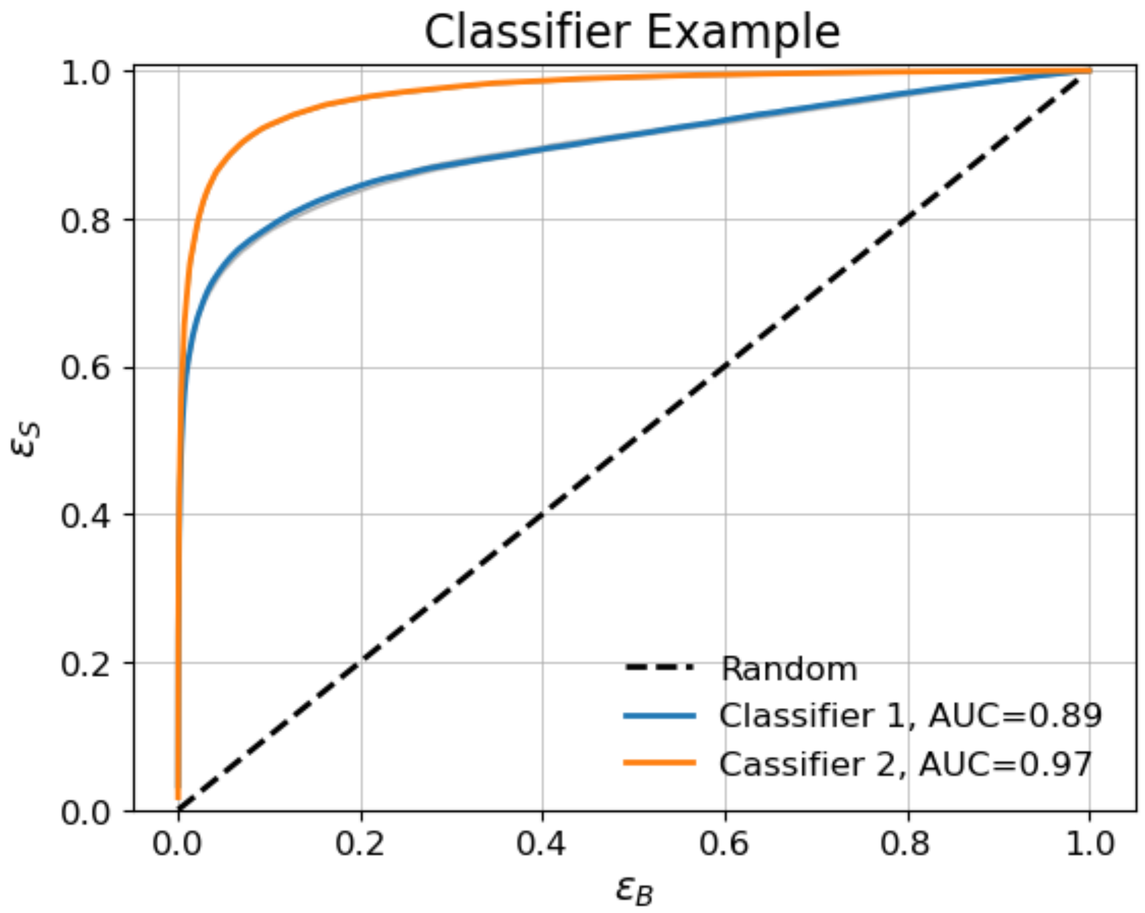
# Receiver operating characteristic (ROC) curve



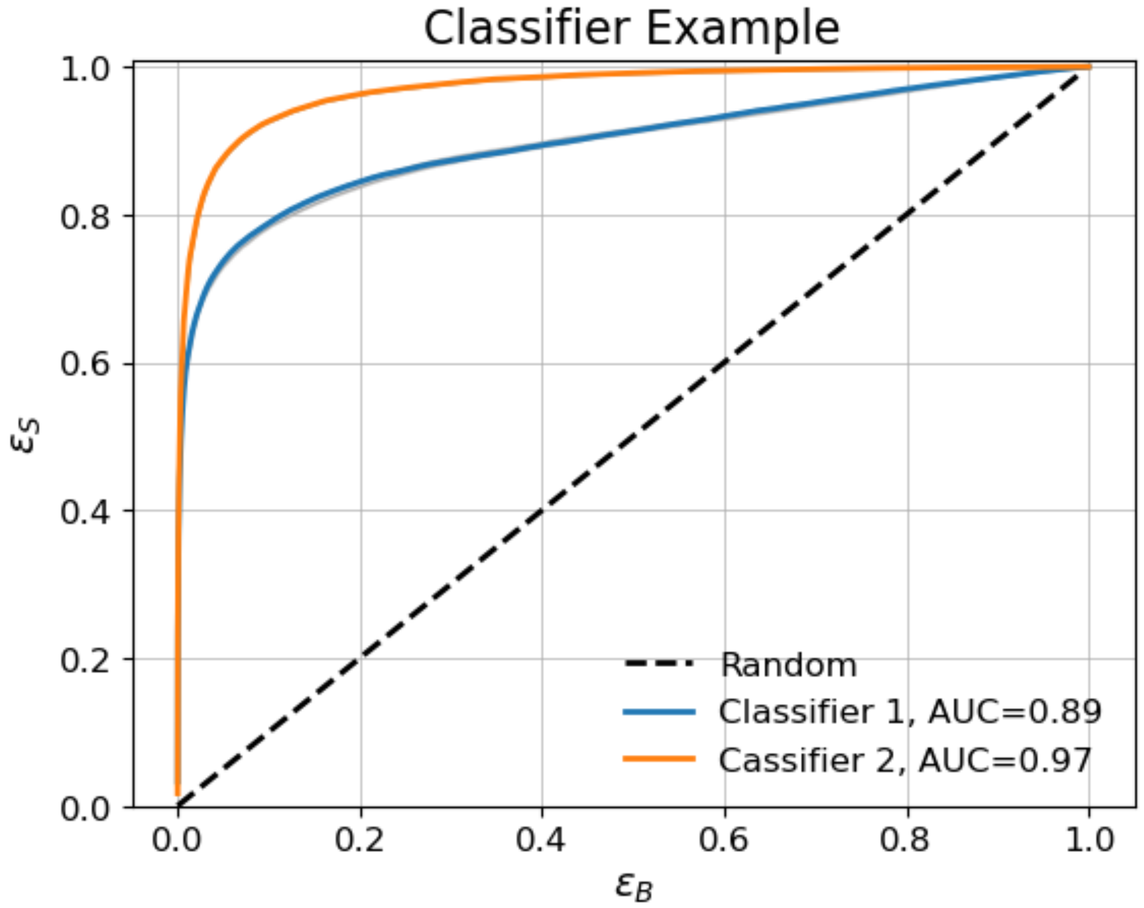
# Receiver operating characteristic (ROC) curve

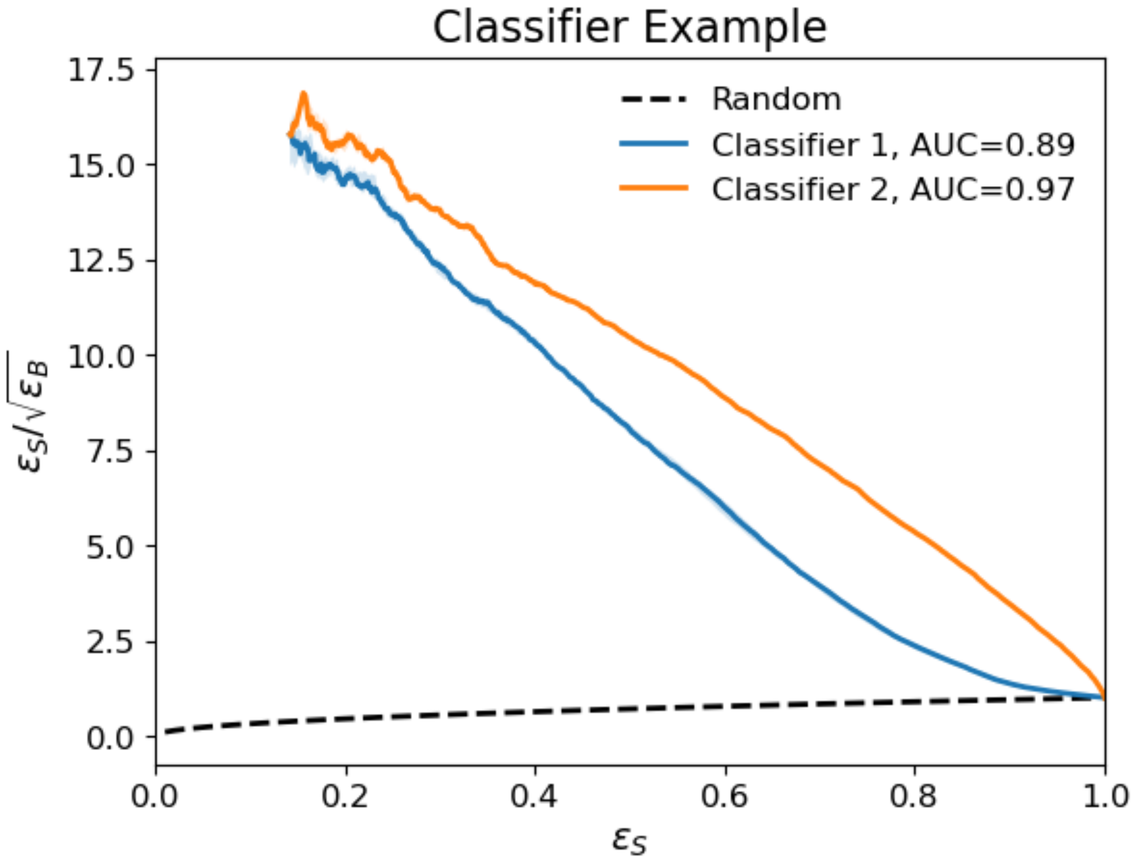


# Significance Improvement Characteristic (SIC) curve

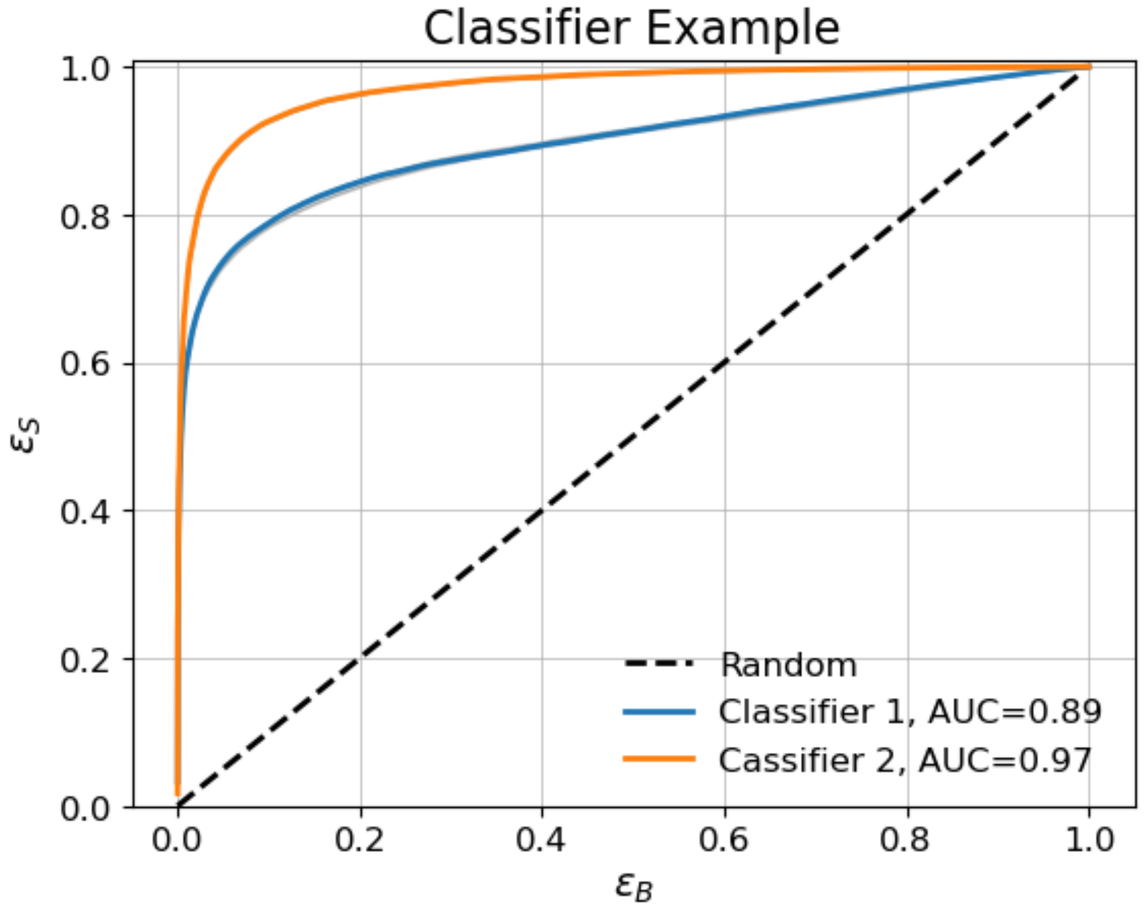
$$\sigma_{\text{Classifier}} \approx \frac{\epsilon_S}{\sqrt{\epsilon_B}} \frac{S}{\sqrt{B}} = \text{SIC} \cdot \sigma_{\text{initial}}$$

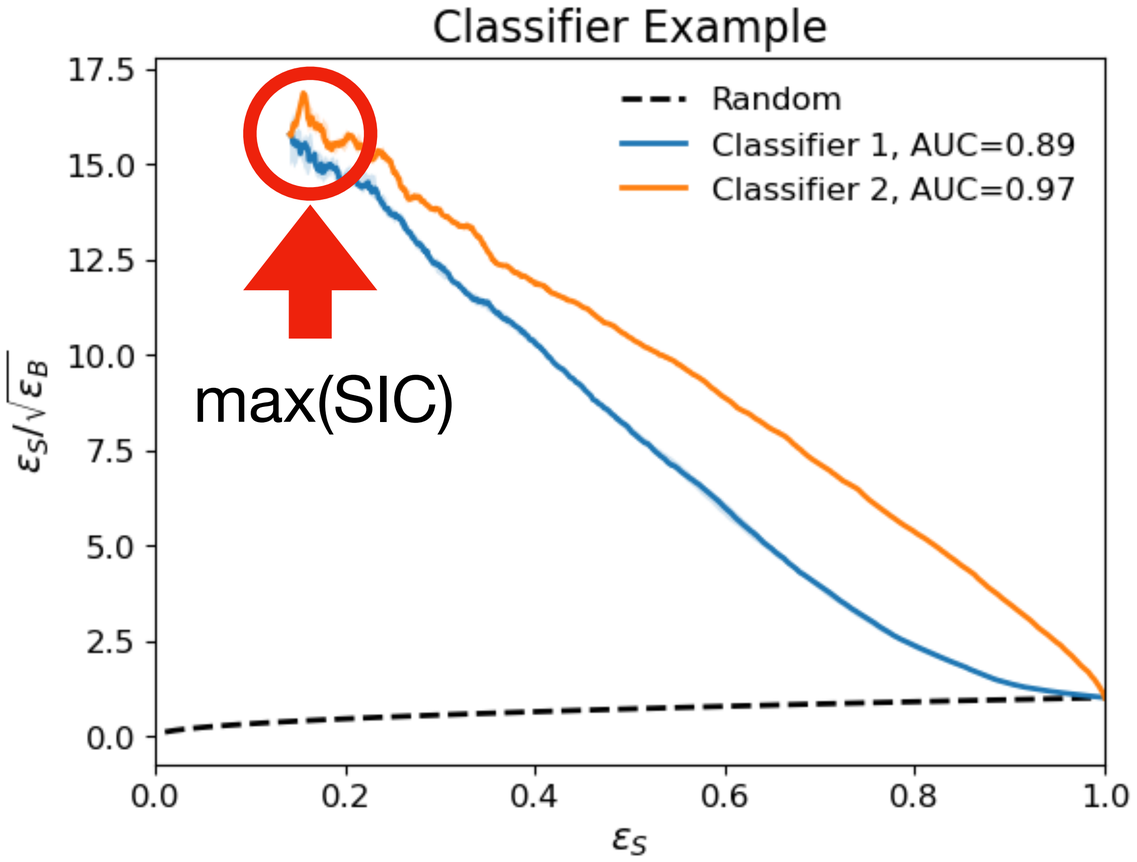


# Significance Improvement Characteristic (SIC) curve

$$\sigma_{\text{Classifier}} \approx \frac{\epsilon_S}{\sqrt{\epsilon_B}} \frac{S}{\sqrt{B}} = \text{SIC} \cdot \sigma_{\text{initial}}$$

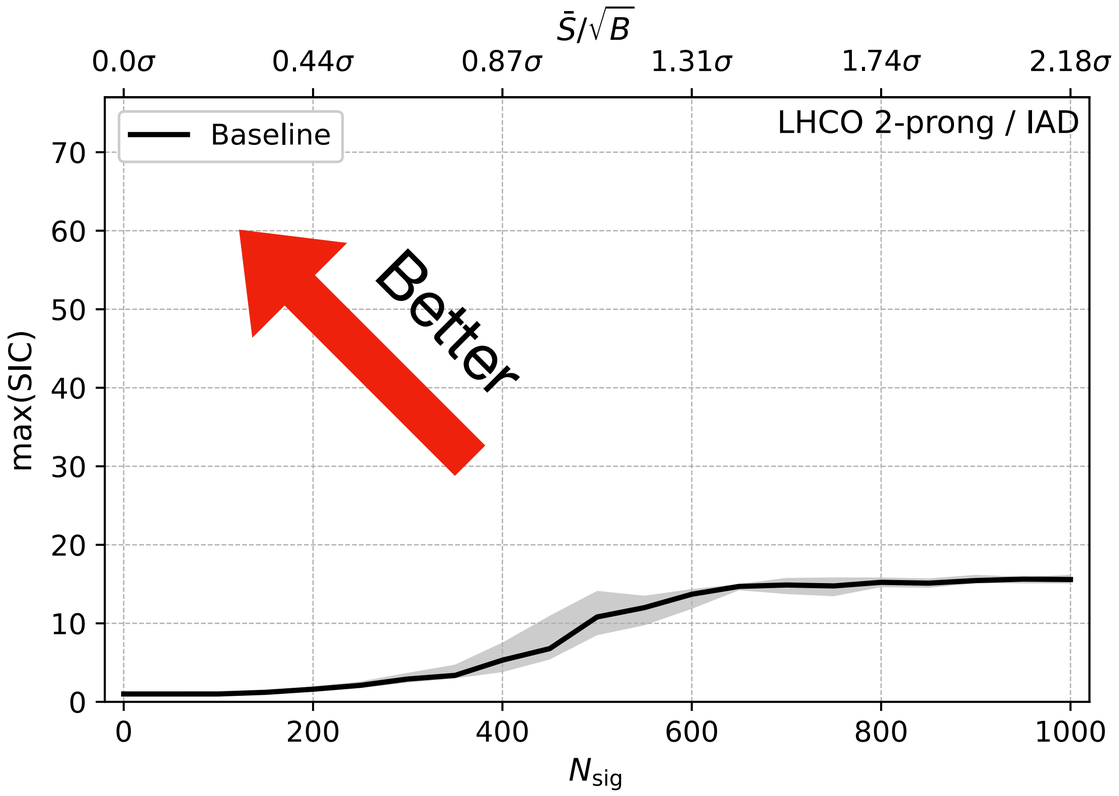


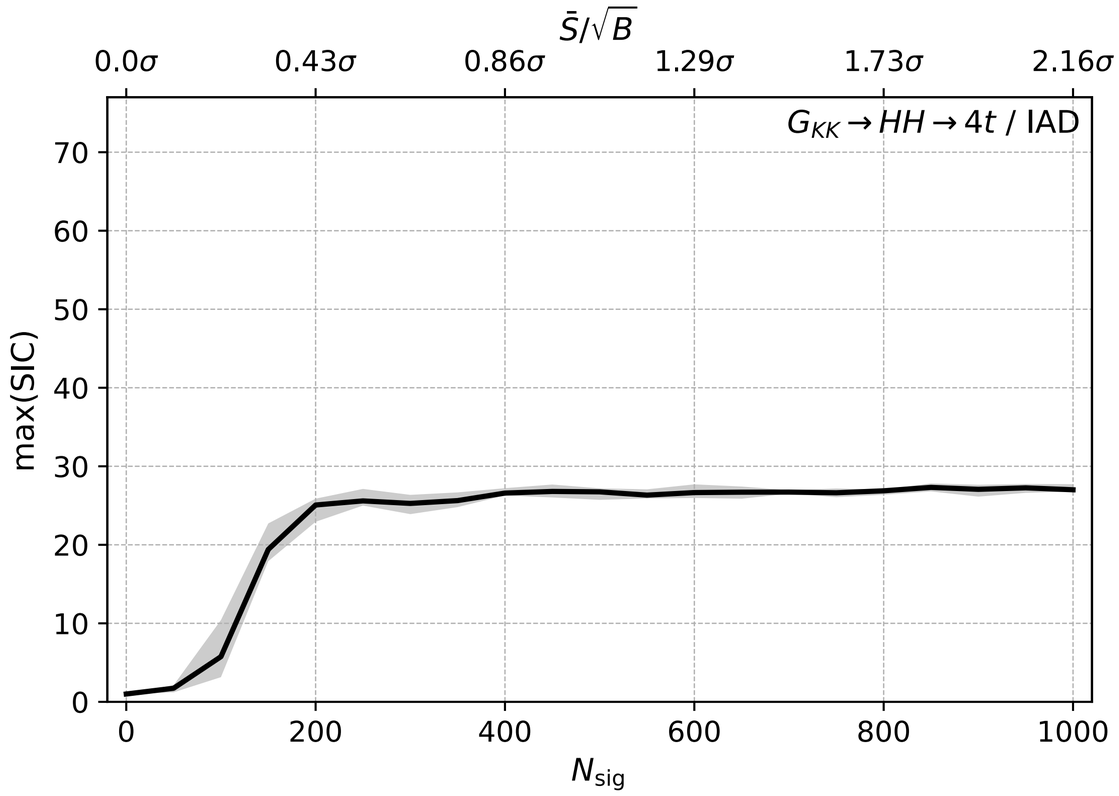
# Results

# Baseline: Some Signals are Easier

## LHCO 2-prong

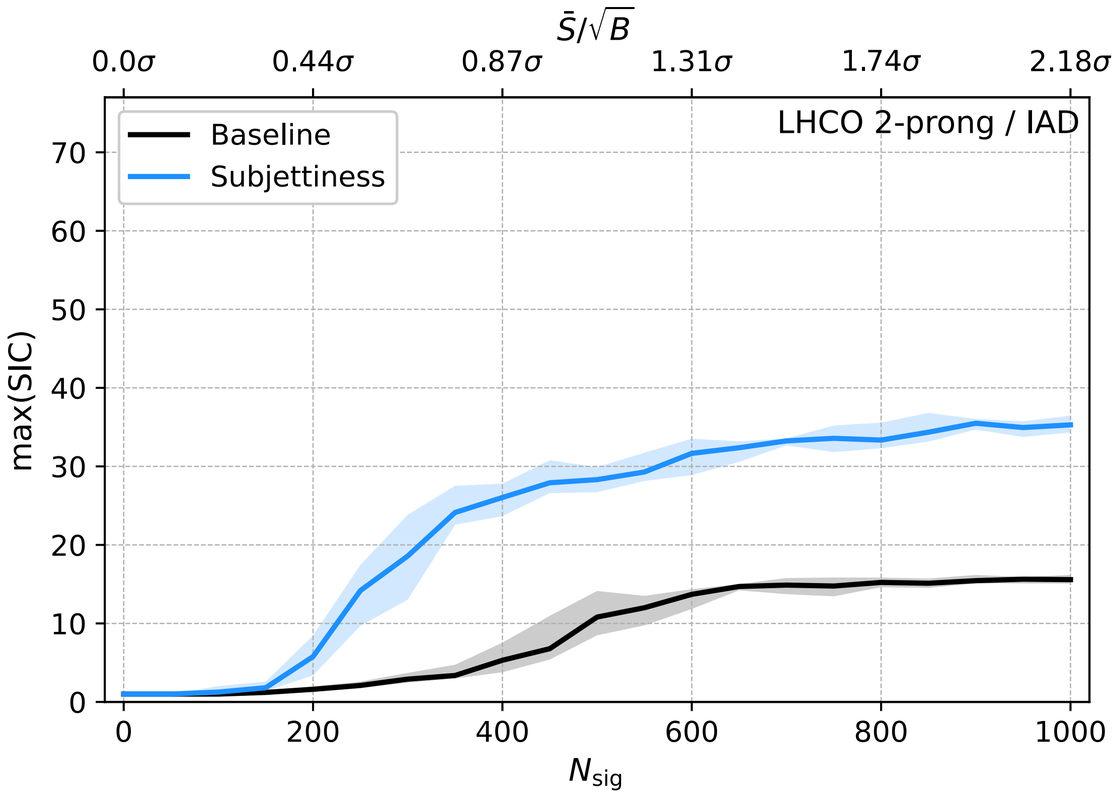



## $G_{KK} \to HH \to 4t$

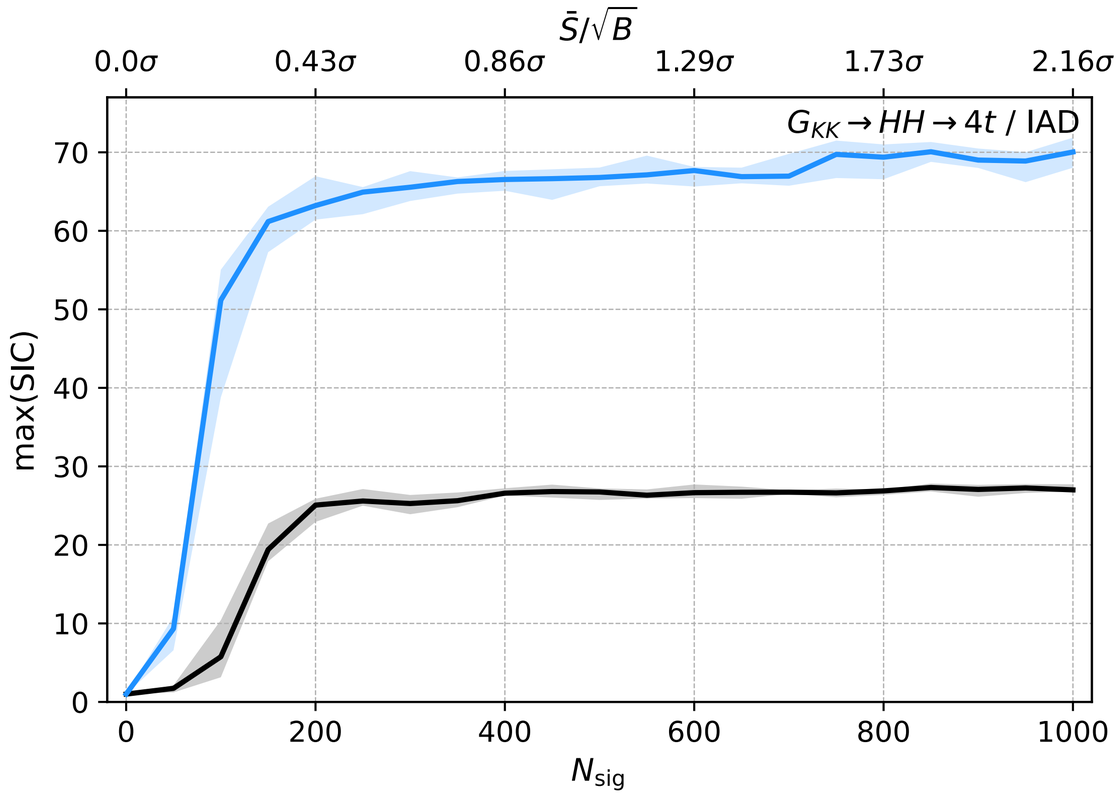


# Subjettiness: A Versatile Feature Set

## LHCO 2-prong

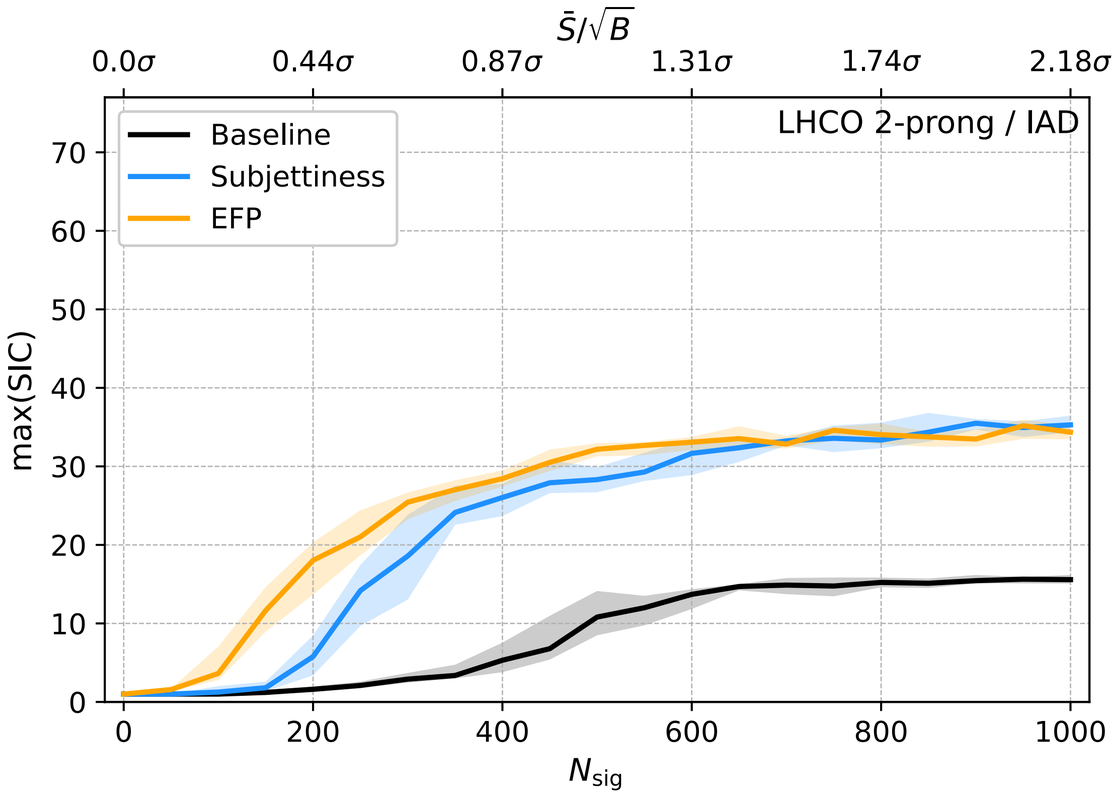



## $G_{KK} \to HH \to 4t$

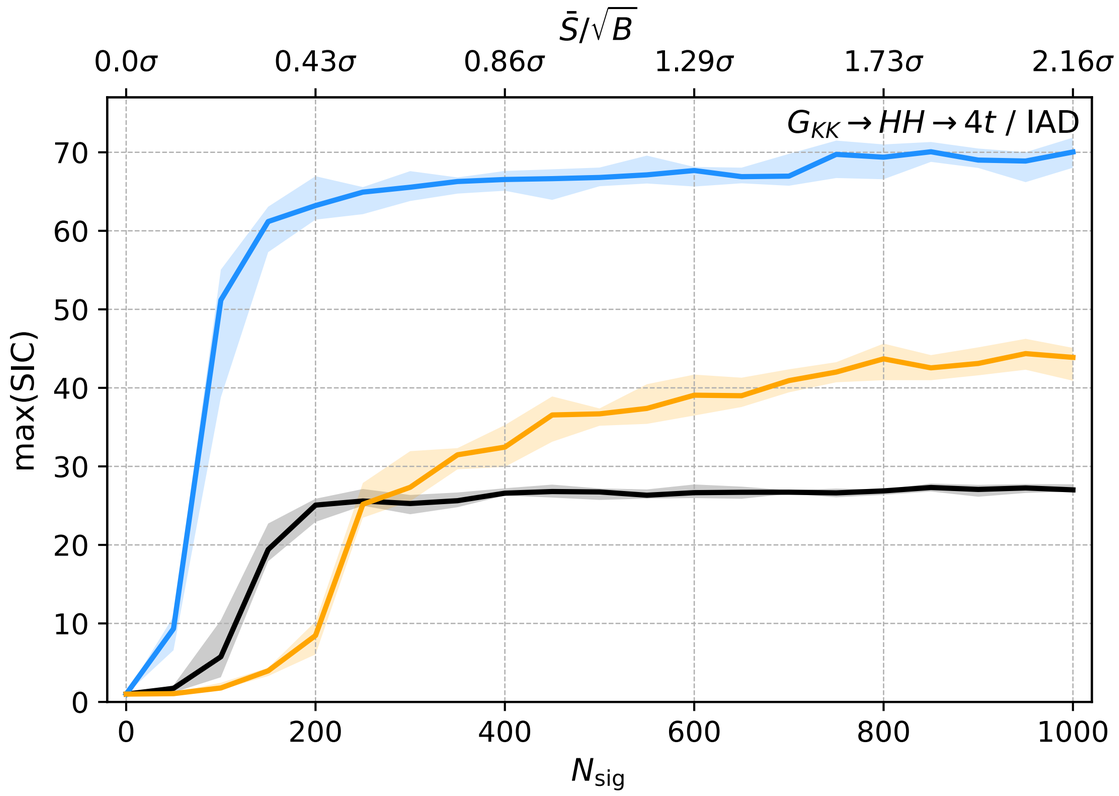


# EFP: Sometimes Great, Sometimes Not

## LHCO 2-prong

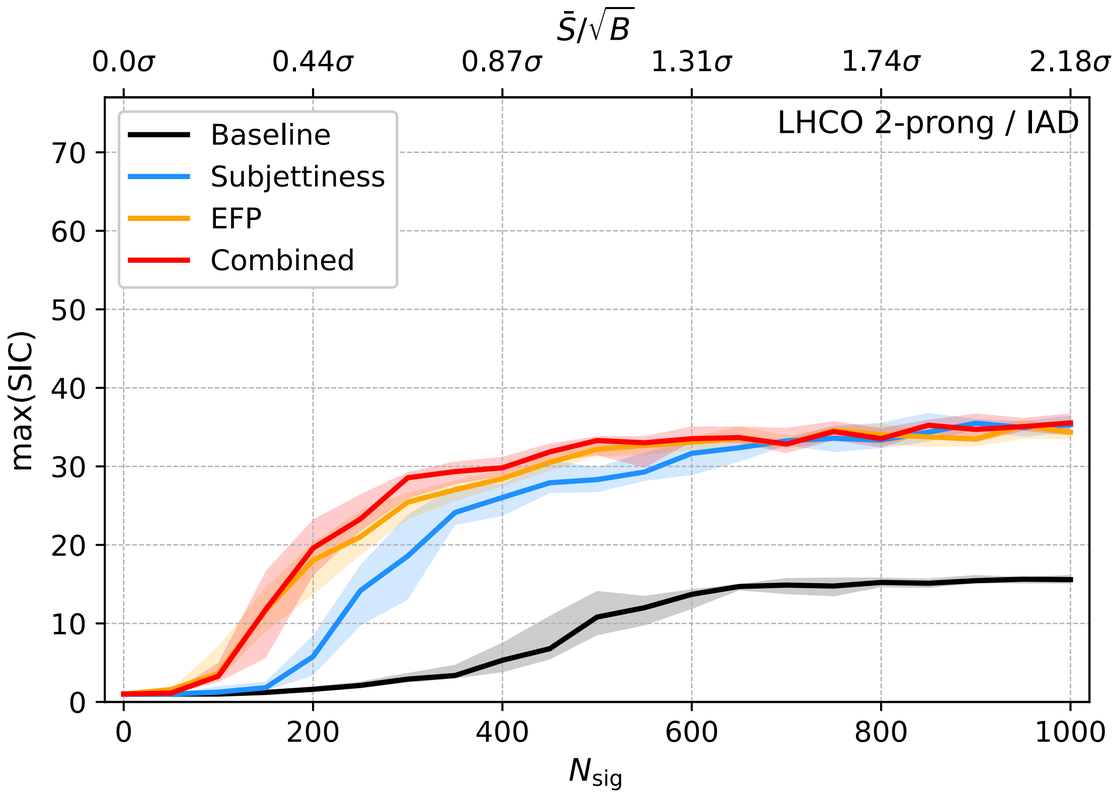



## $G_{KK} \to HH \to 4t$

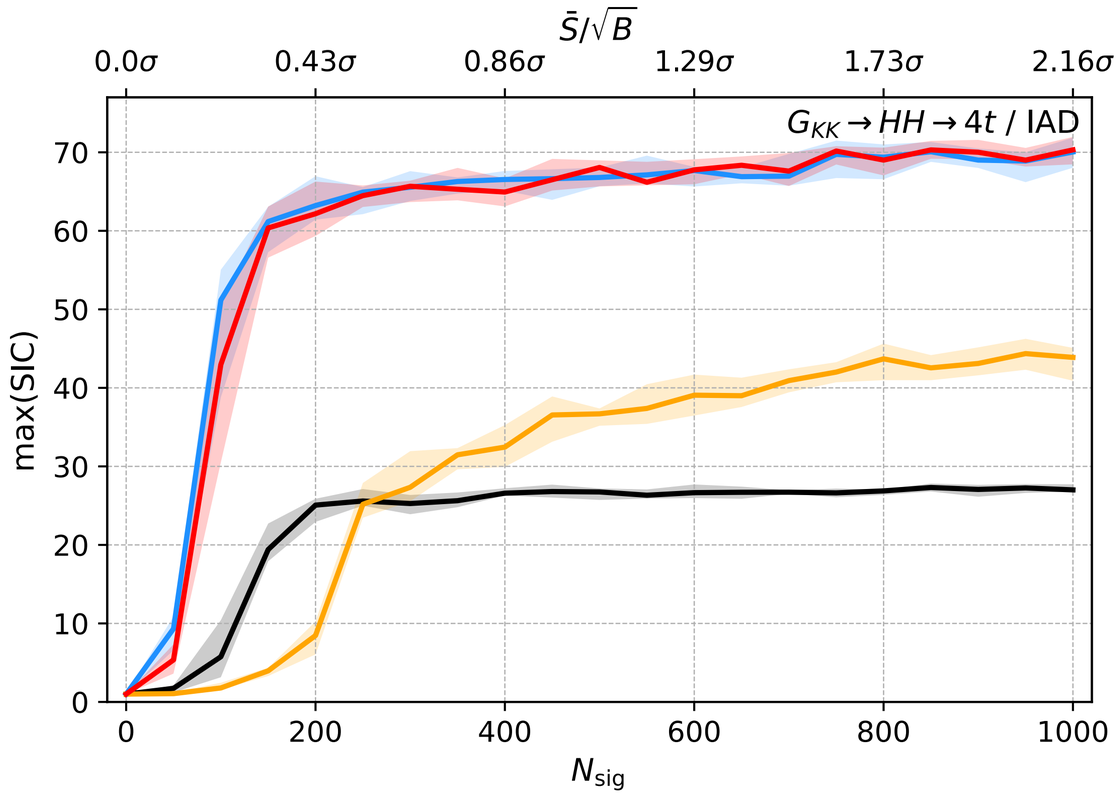


# Combined: Most Performant and Model Agnostic

## LHCO 2-prong

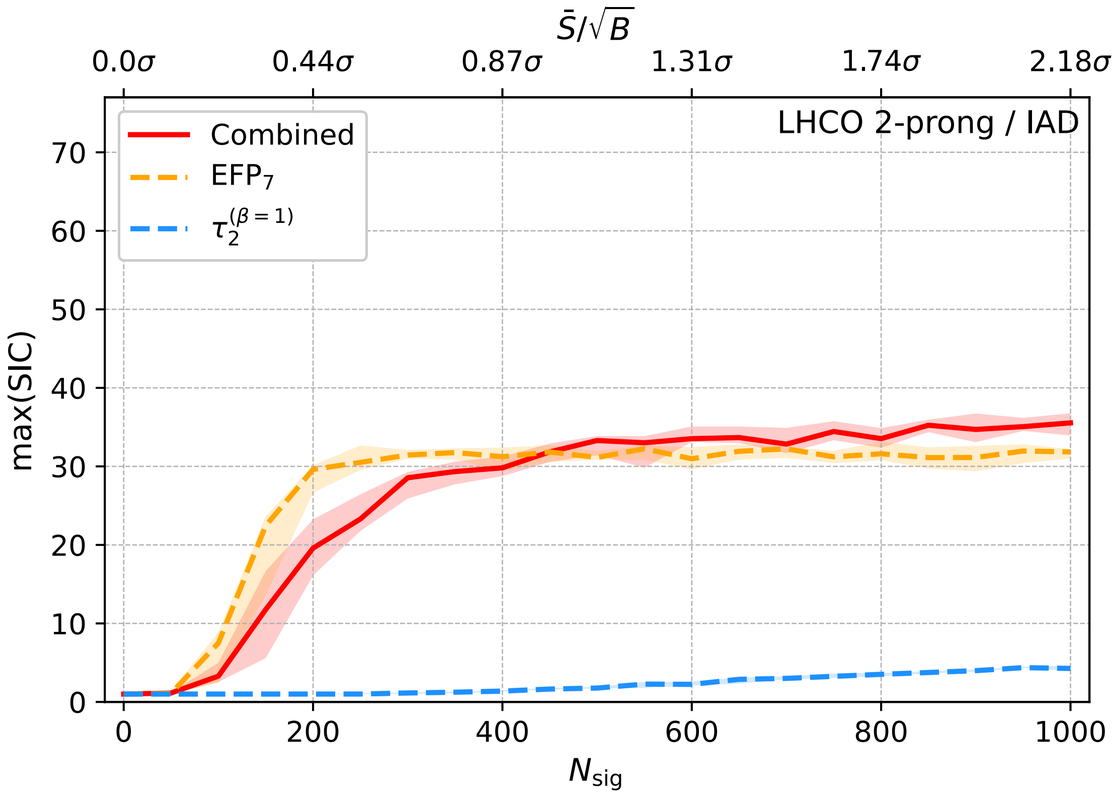



## $G_{KK} \to HH \to 4t$

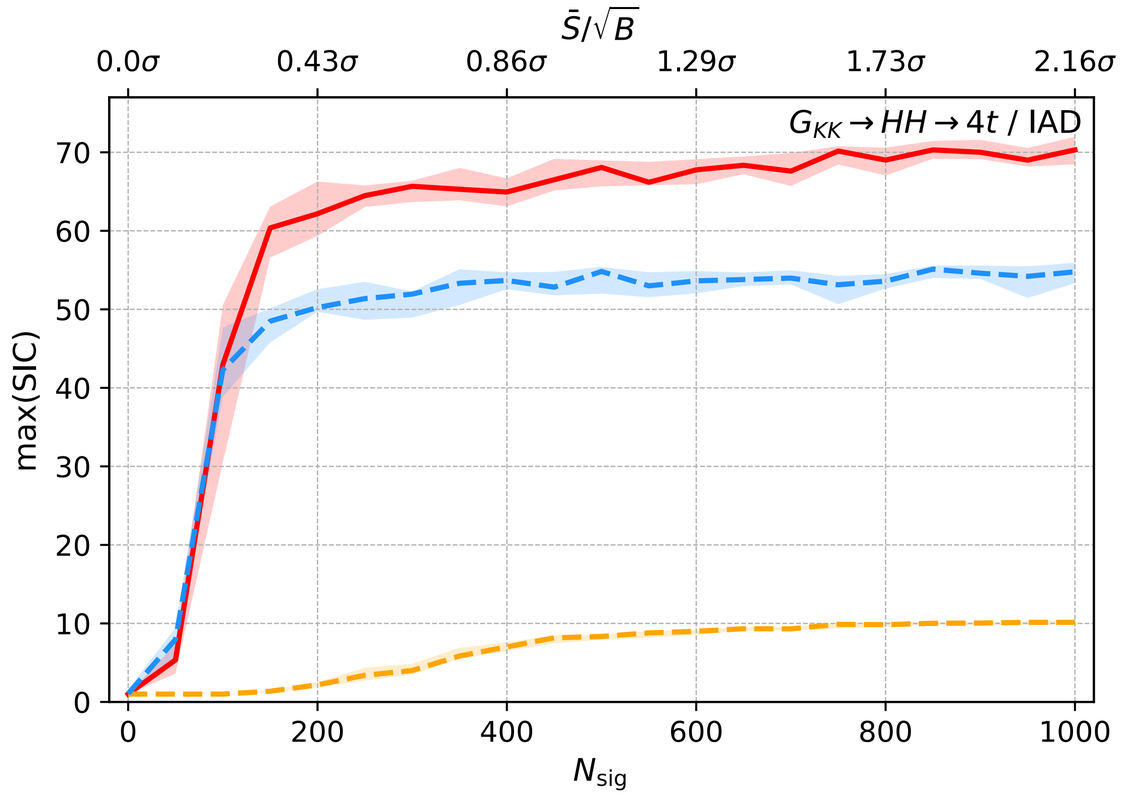


# Individual Substructure Feature Drives the Performance

## LHCO 2-prong

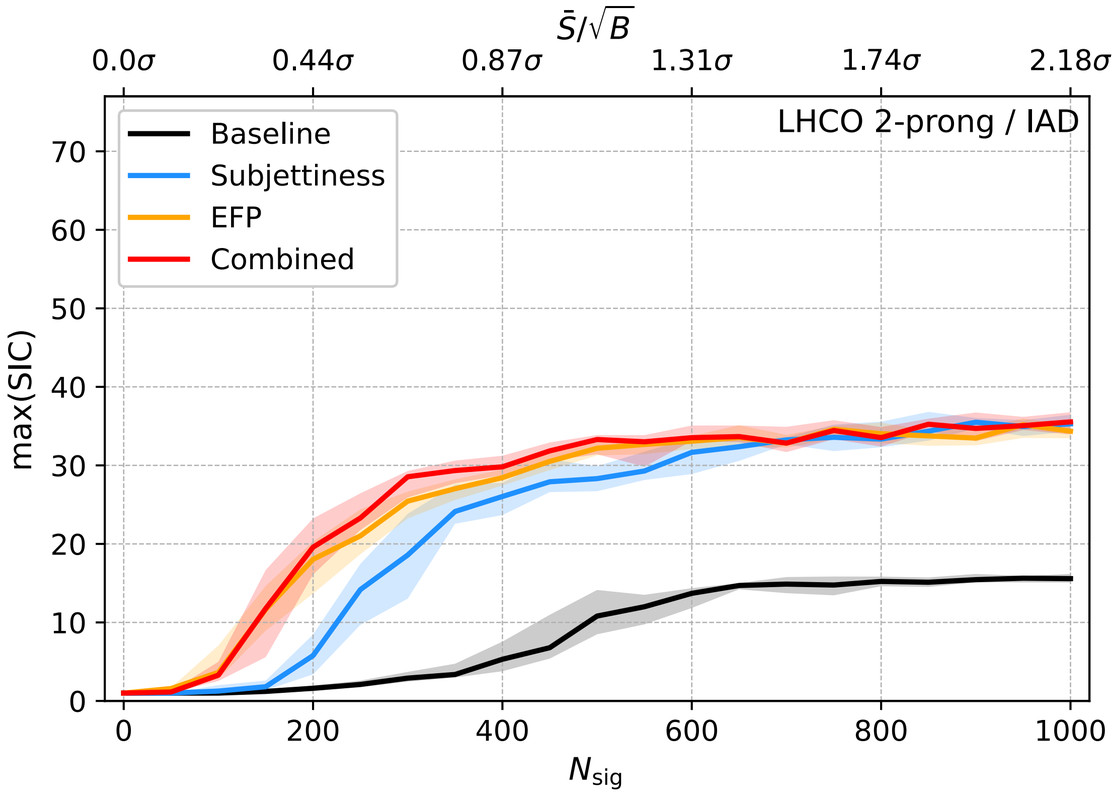



## $G_{KK} \to HH \to 4t$



# Discovery Potential: What do We Need for $5\sigma$ ?

## LHCO 2-prong



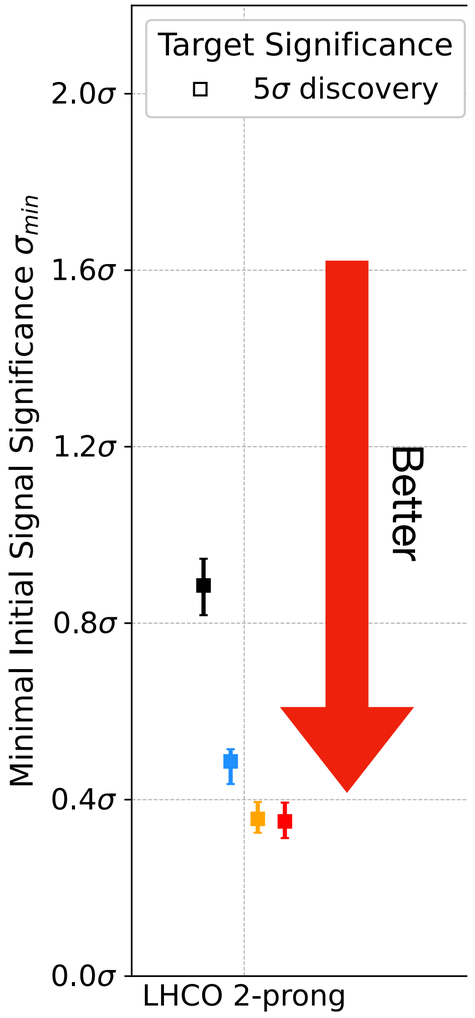

Initial Significance:

$$\sigma = S/\sqrt{B}$$

Significance with Classifier, Asimov estimate: $\mathcal{S}$

Solve for $\sigma_{min}$:

$$\max(\mathcal{S})(\sigma_{min}) = 5\sigma$$



# The Combined Set is the Best!

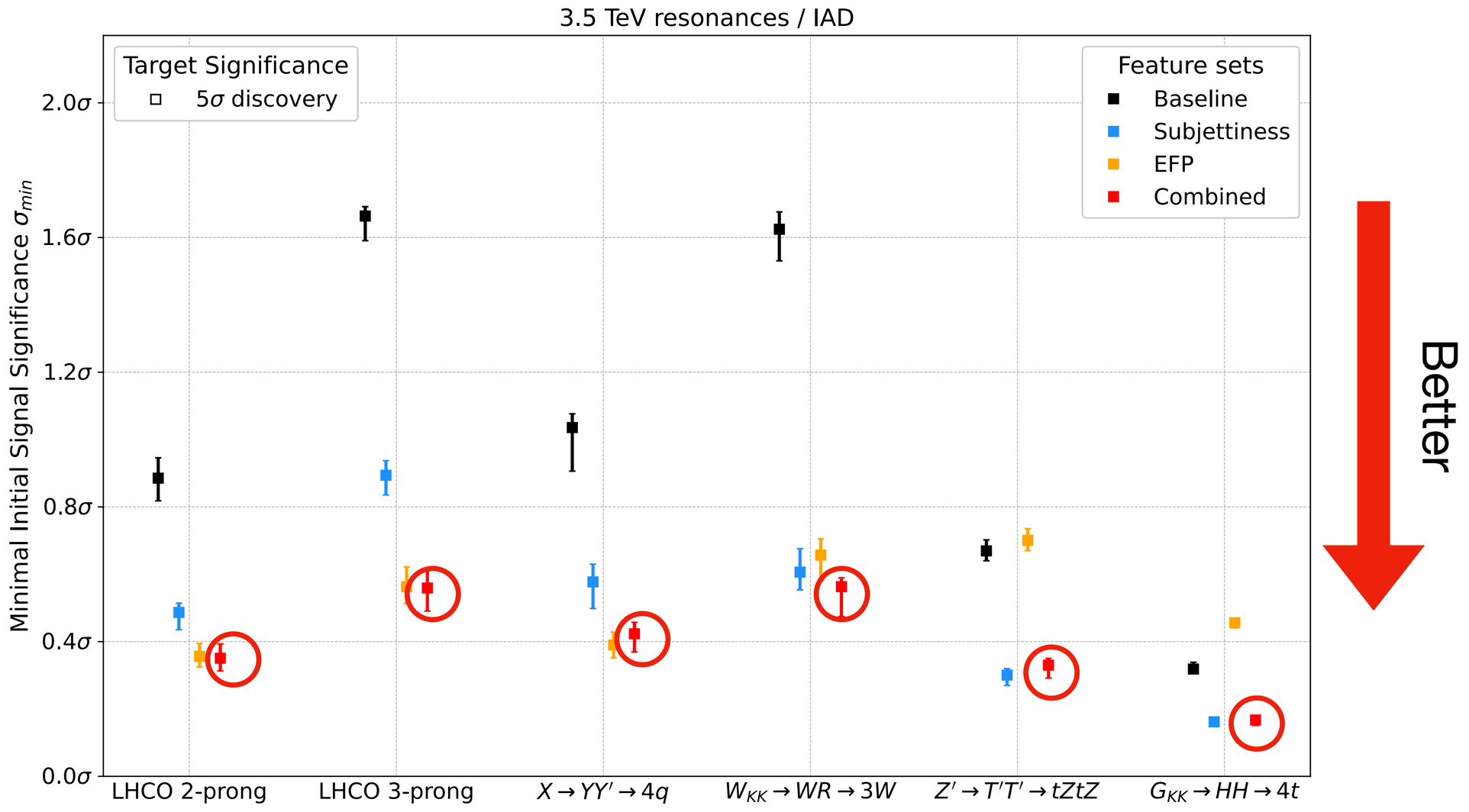
Initial Significance:

$$\sigma = S/\sqrt{B}$$

Significance with Classifier $\mathcal{S}$

Solve for $\sigma_{min}$:

$$\max(\mathcal{S})(\sigma_{min}) = \sigma_t$$

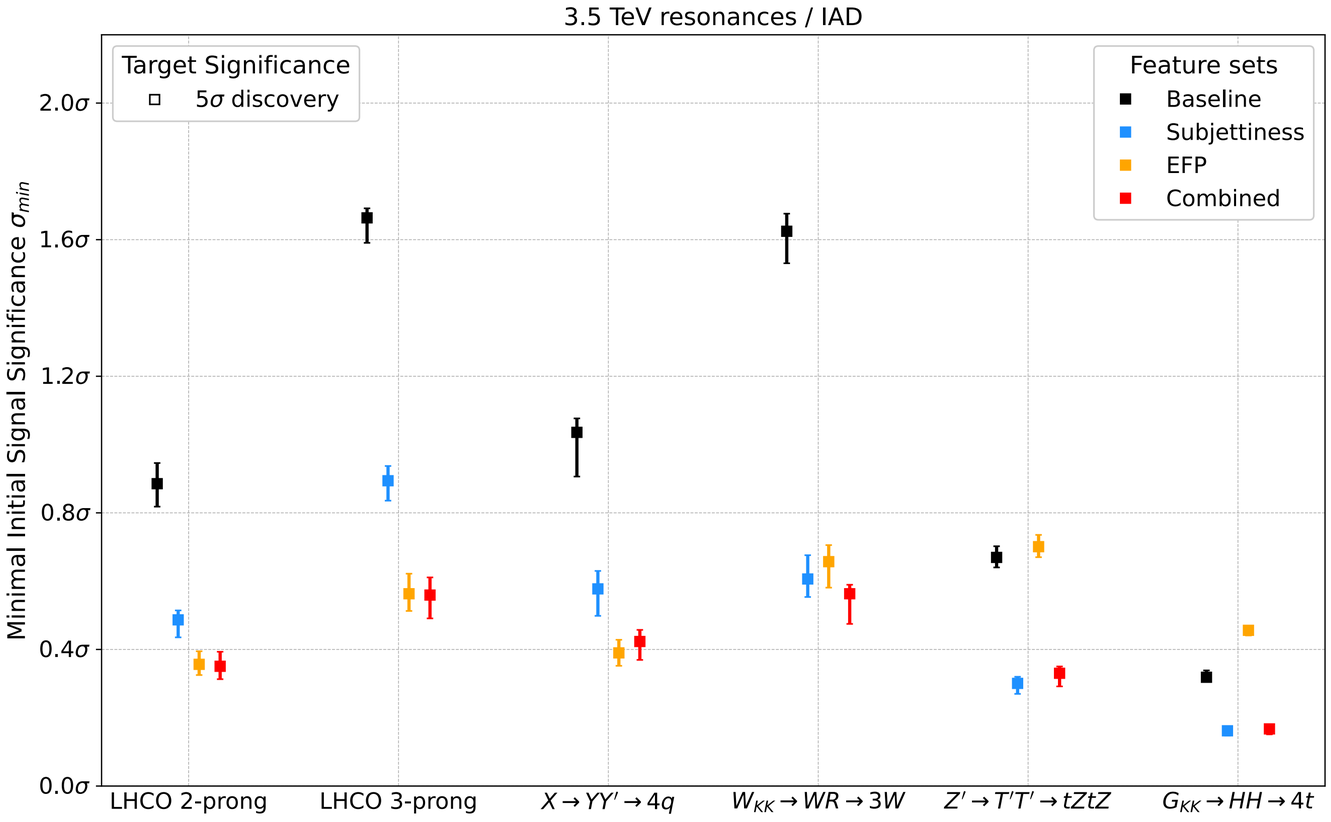


GET THE DATA

# Conclusion

READ THE PAPER

- New signal benchmarks compatible with the LHCO background
- The Combined set yielding on average a factor of $\sim 2.5$ better discovery potential than the Baseline
- CWoLa hunting produces similar results as the IAD case

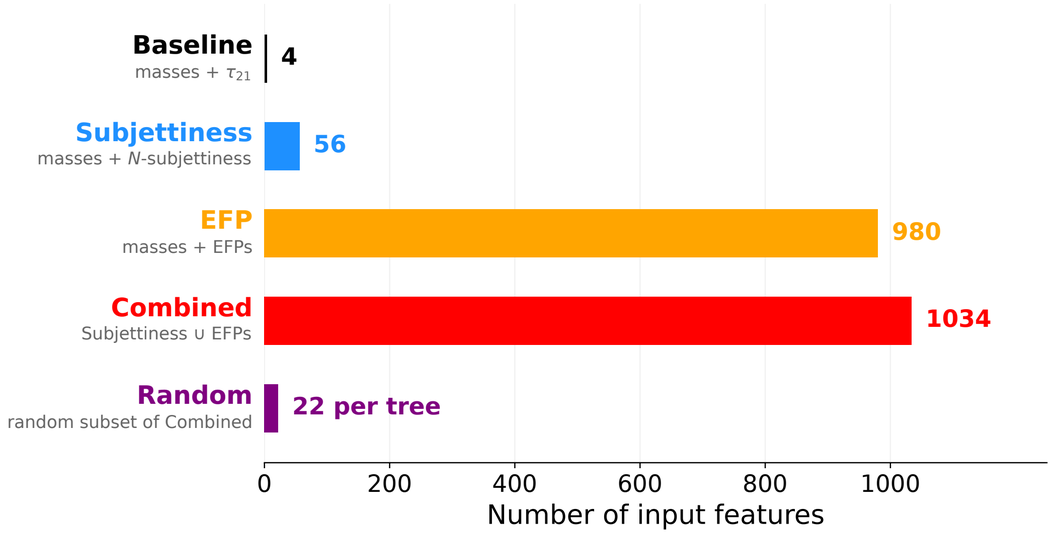


# Backup

# Reducing Training Time with the Random Set

## The Problem:

- Training time of a BDT scales linearly with the number of input features

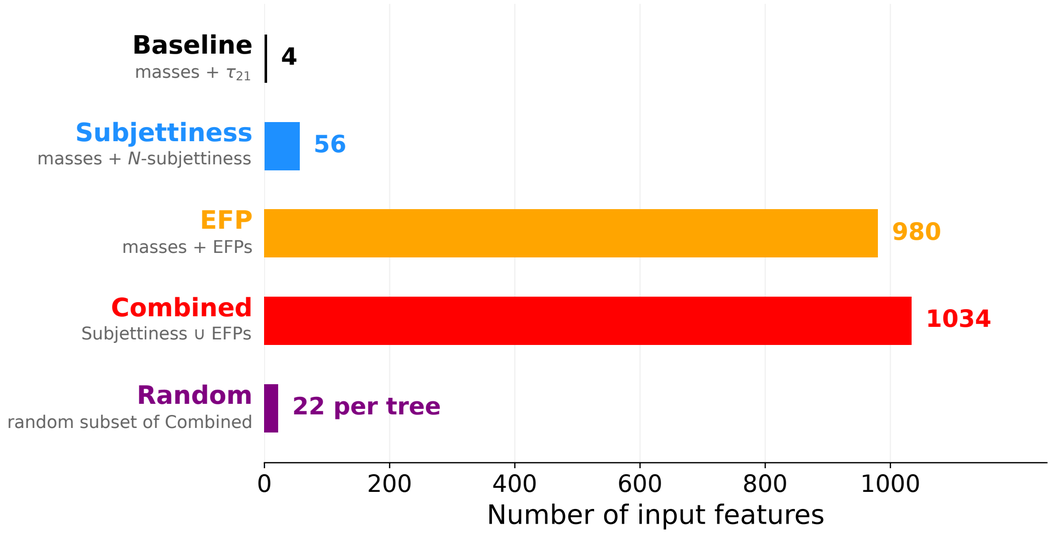


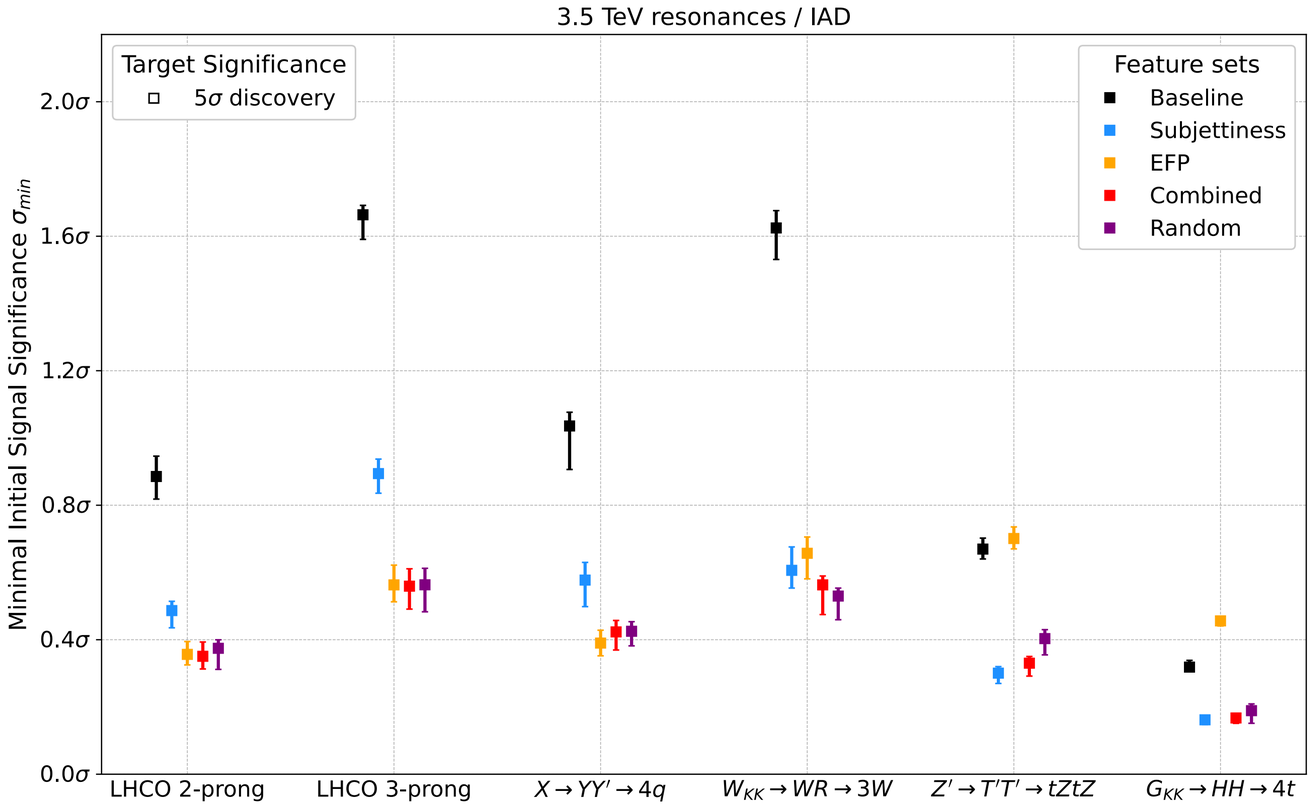

## The Idea:

- Leveraging the ensemble in our classifier architecture
- Training each ensemble member on a different random subset of features
- So 50 BDTs each trained on 22 randomly drawn features from the Combined set

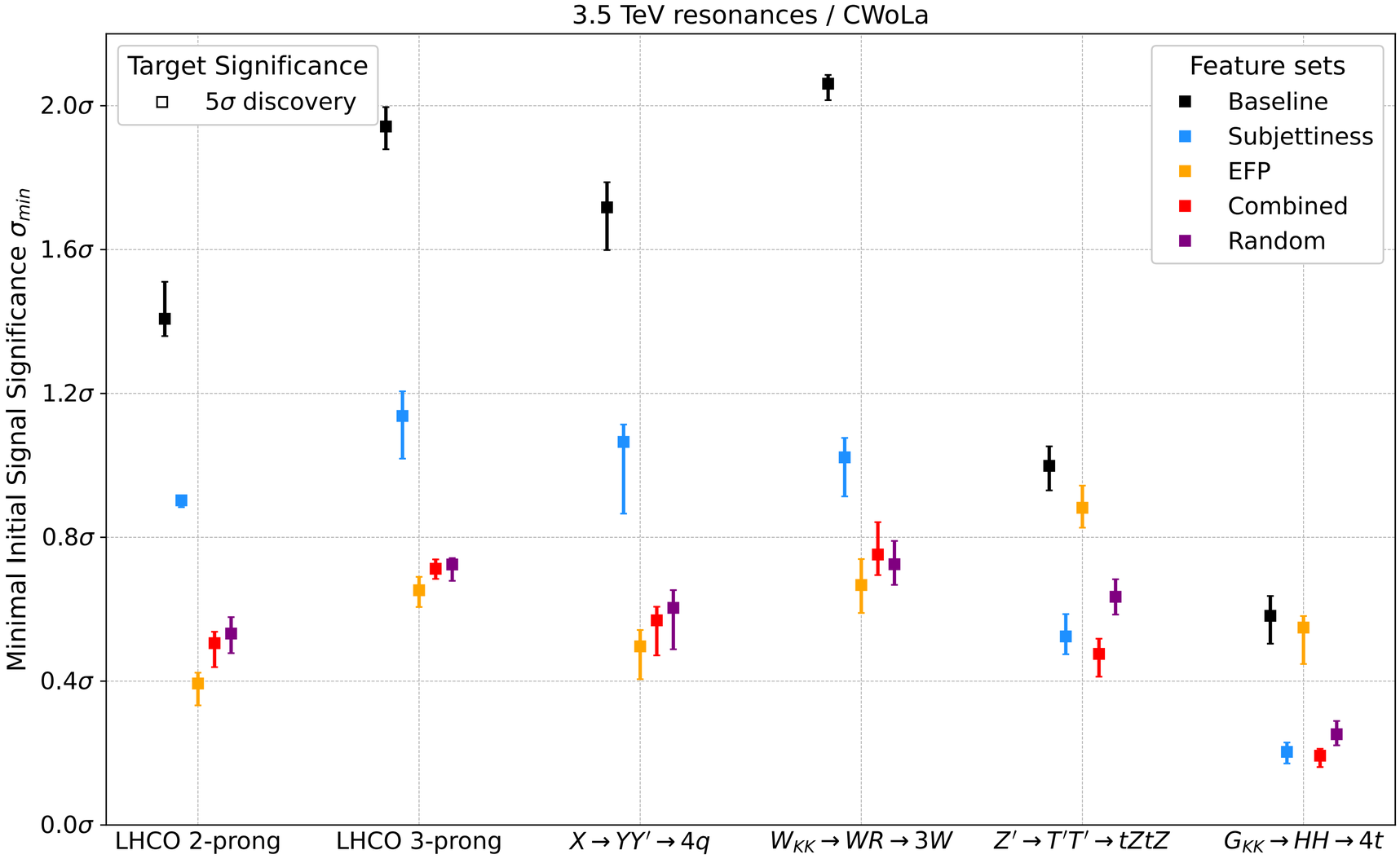
# Random: Same Performance But Faster

- Training a BDT on 22 features instead of 1034
  → $\sim 20$ times faster run time
- Performance similar to the Combine set

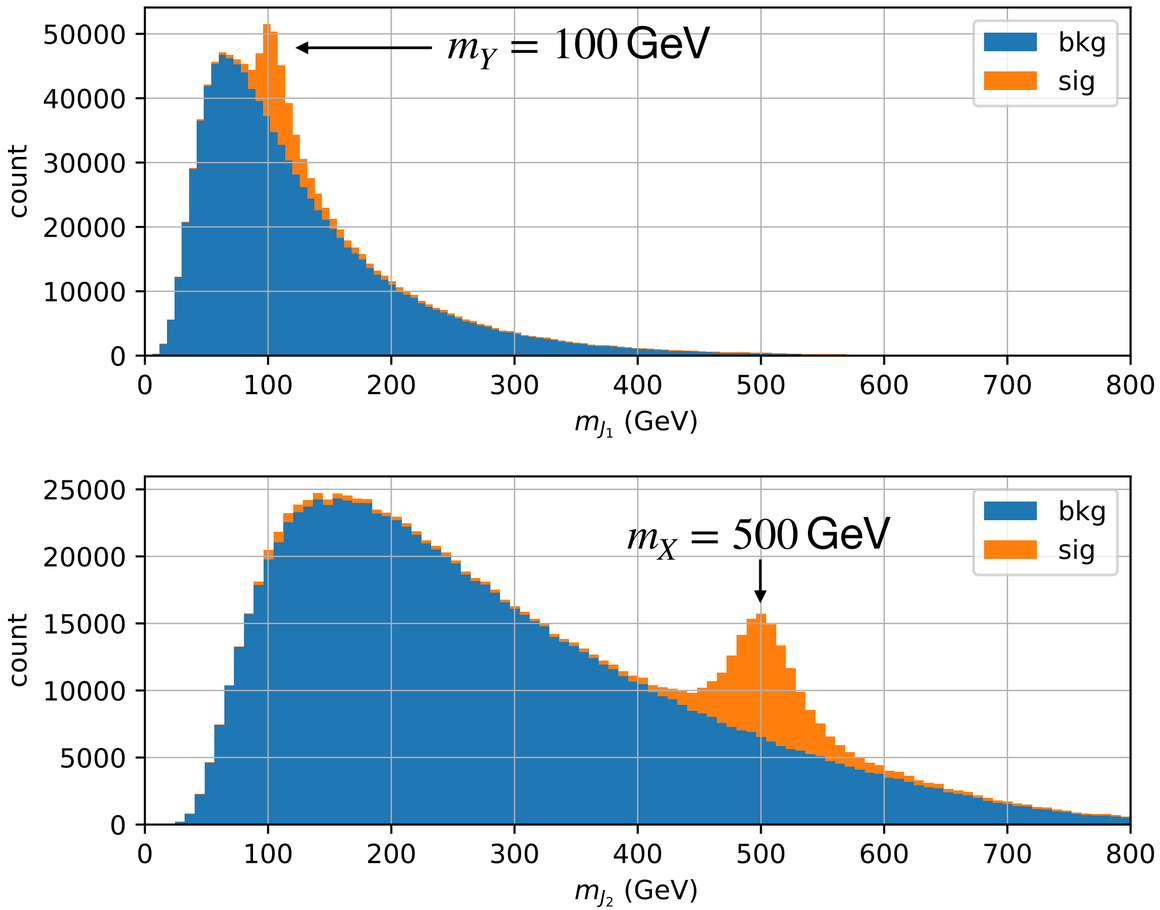


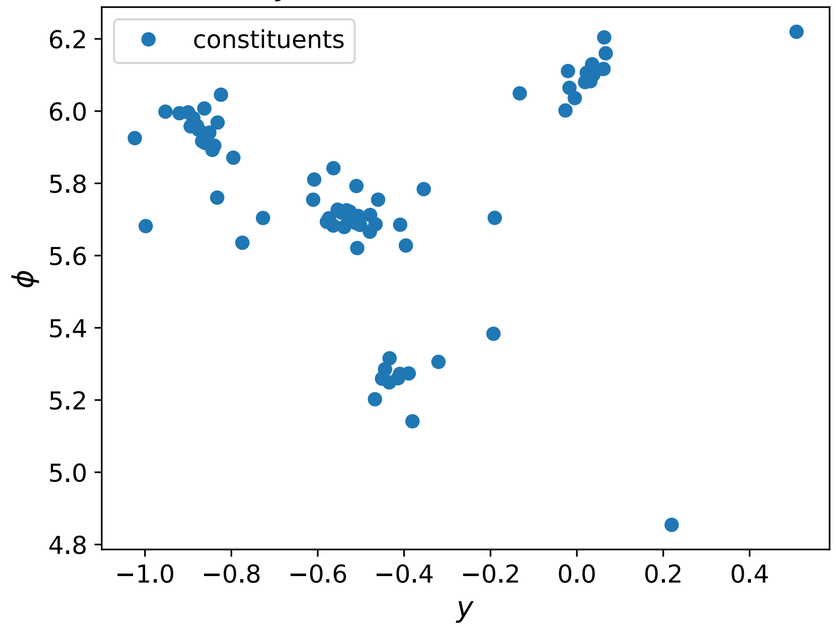



# CWoLa Results



# Jet Observables




- Rapidity

$$y = \frac{1}{2}\log\left(\frac{E + p_z}{E - p_z}\right)$$

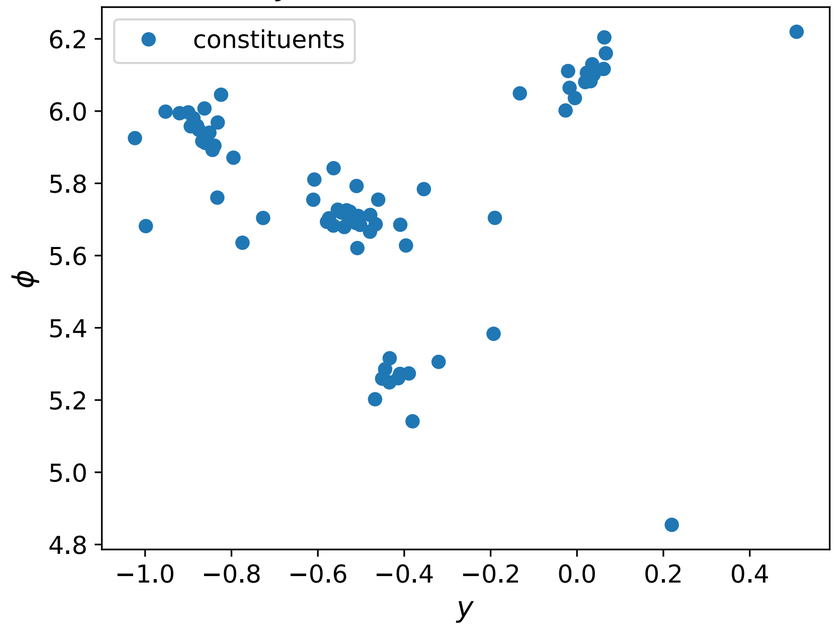
- Azimuthal angle

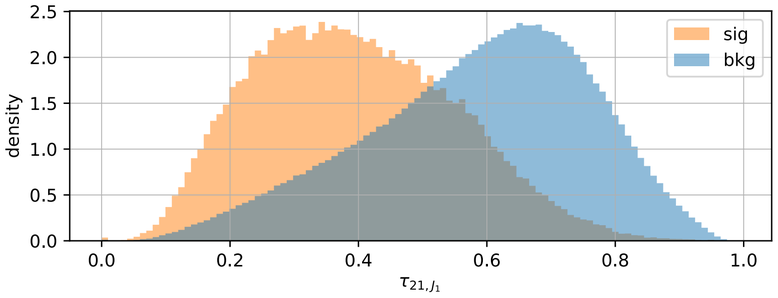
$$\phi = \arctan\left(\frac{p_y}{p_x}\right)$$

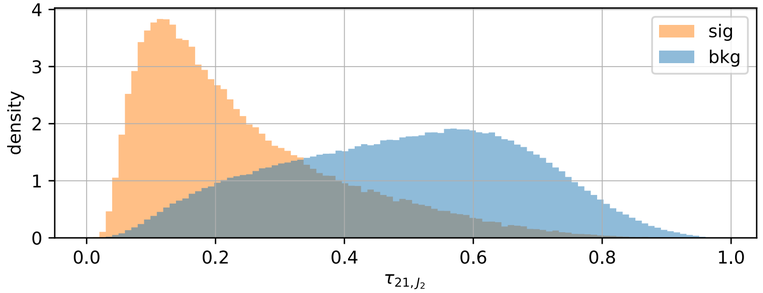
# Jet Observables

## N-Subjettiness $\tau_N$

- Probing jets for a specific number of subjets N (or less)
- Idea being to cluster the constituents of a jet around N jet candidate axis
- One also often uses the subjettiness ration $\tau_{NM} = \frac{\tau_N}{\tau_M}$ for $N > M$



# What are Energy Flow Polynomials?

arXiv:1712.07124v2 [hep-ph] 3 Apr 2018

## Energy flow polynomials:
## A complete linear basis for jet substructure

**Patrick T. Komiske, Eric M. Metodiev, Jesse Thaler**

*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

*E-mail:* pkomiske@mit.edu, metodiev@mit.edu, jthaler@mit.edu

ABSTRACT: We introduce the energy flow polynomials: a complete set of jet substructure observables which form a discrete linear basis for all infrared- and collinear-safe observables. Energy flow polynomials are multiparticle energy correlators with specific angular structures that are a direct consequence of infrared and collinear safety. We establish a powerful graph-theoretic representation of the energy flow polynomials which allows us to design efficient algorithms for their computation. Many common jet observables are exact linear combinations of energy flow polynomials, and we demonstrate the linear spanning nature of the energy flow basis by performing regression for several common jet observables. Using linear classification with energy flow polynomials, we achieve excellent performance on three representative jet tagging problems: quark/gluon discrimination, boosted $W$ tagging, and boosted top tagging. The energy flow basis provides a systematic framework for complete investigations of jet substructure using linear methods.

[arXiv:1712.07124]

# Mathematical representation

Energy Flow Polynomial for a graph G

$$EFP_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell}$$

Energy fraction

Angular distance between particle $i_k$ and $i_\ell$

## Hadronic colliders

$$z_i = \left(\frac{p_{T,i}}{p_{T,J}}\right)^{\kappa} \qquad \theta_{ij} = \left(\Delta y_{ij}^2 + \Delta\phi_{ij}^2\right)^{\frac{\beta}{2}}$$

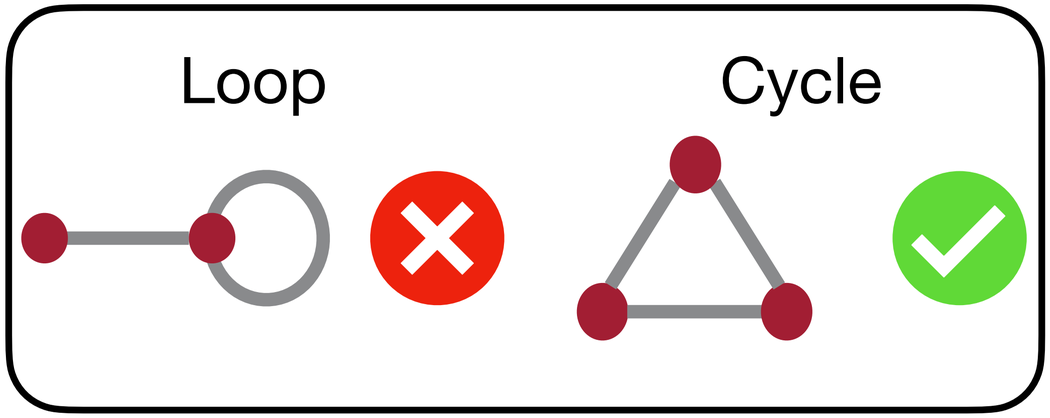
$$p_{T,J} \equiv \sum_{i=1}^{M} p_{T,i} \qquad \Delta y_{ij} \equiv y_i - y_j \qquad \Delta\phi_{ij} \equiv \phi_i - \phi_j$$

[arXiv:1712.07124]

# Multigraph correspondence

$$EFP_G = \sum_{i_1=1}^{M} \cdots \sum_{i_N=1}^{M} z_{i_1} \cdots z_{i_N} \prod_{(k,\ell)\in G} \theta_{i_k i_\ell}$$

- A multigraph is composed of vertices $(N)$ which are connected by multiple edges $(k, \ell)$
- Only loop-less multigraphs relate to EFPs




Vertex correspondence

$$j \iff \sum_{i_j=1}^{M} z_{i_j}$$

Edge correspondence

$$k \text{ —— } \ell \iff \theta_{i_k i_\ell}$$


[5] arXiv:1712.07124