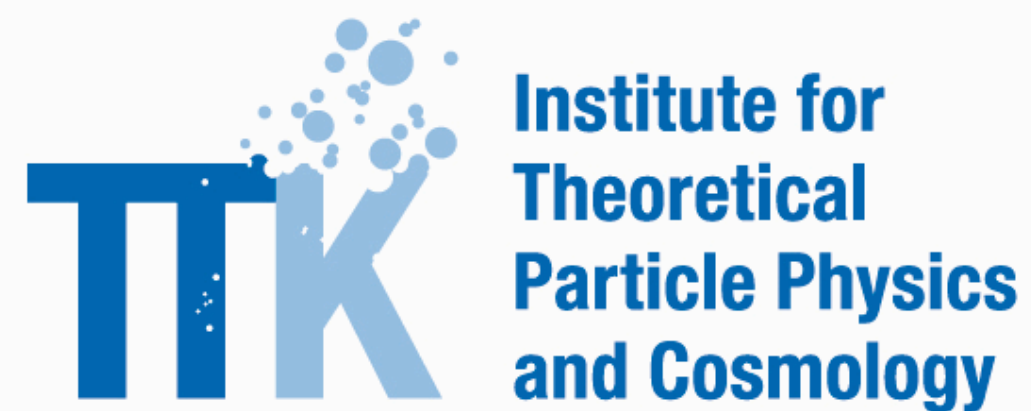


# Energy Flow Polynomials for More Model-Agnostic Anomaly Detection

**CRC Young Scientists Meeting 2025**



**Lukas Lang**

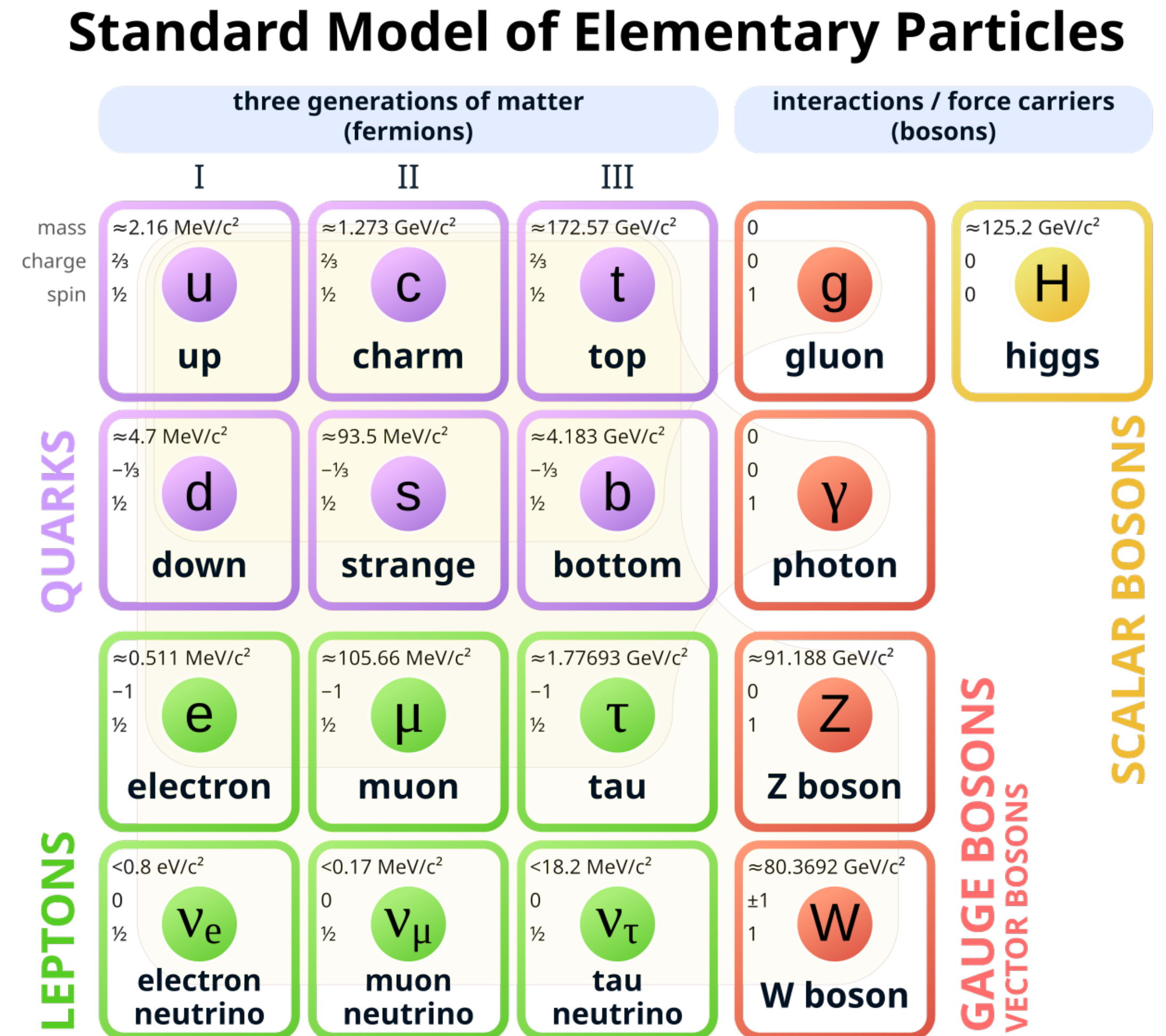
In collaboration with R. Das, M. Hein, G. Kasieczka, M. Krämer, R. Mastandrea, A. Mück, D. Shih  
July 23, 2025

# Energy Flow Polynomials for More Model-Agnostic **Anomaly Detection**

# Anomaly Detection

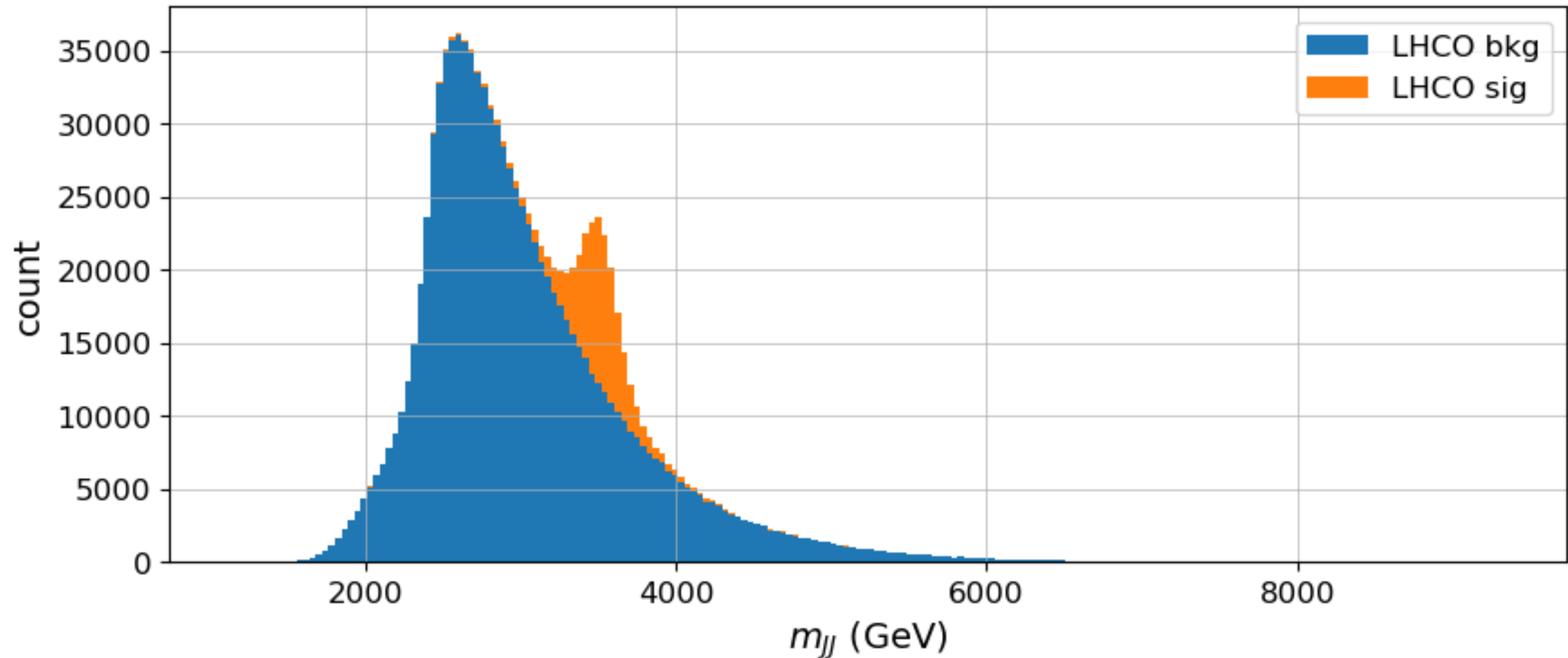
## The Search for New Physics

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



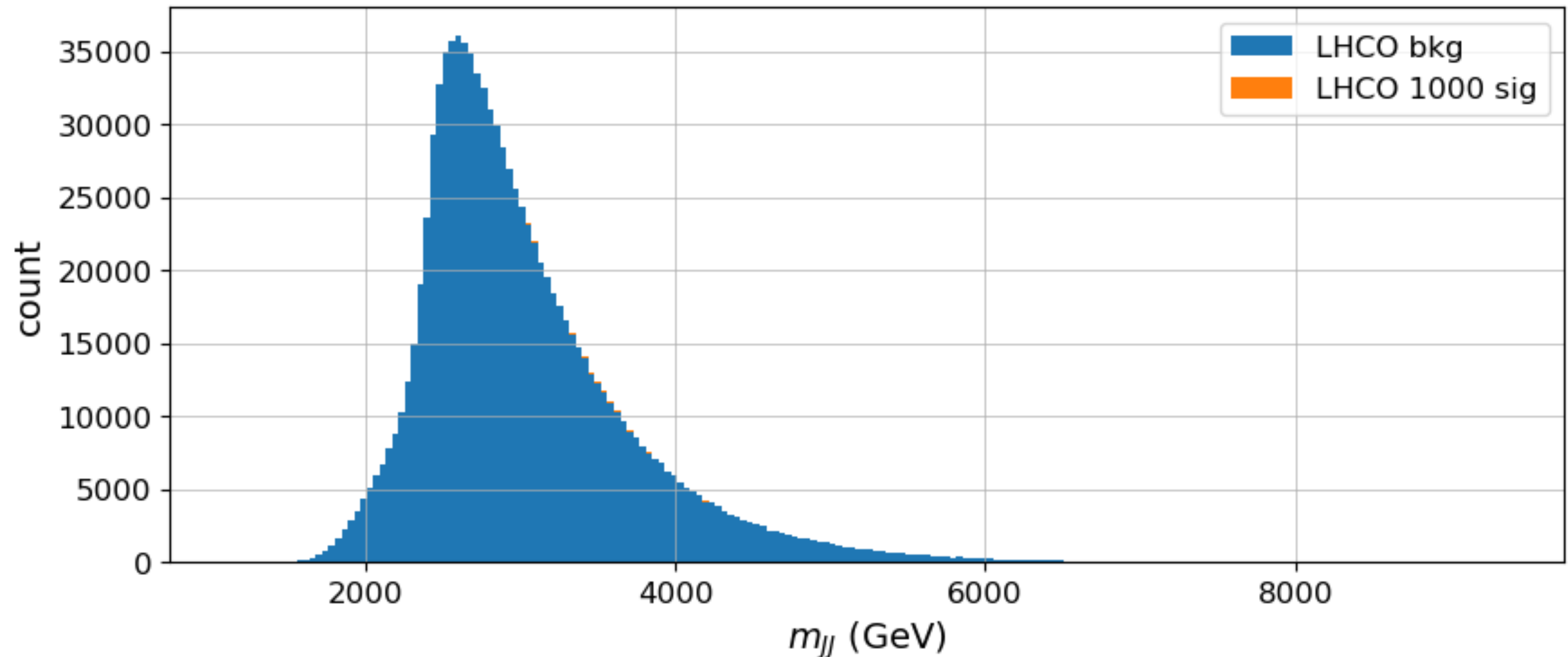
# Searching for Resonances

## LHC Olympics 2020 R&D Dataset



# Searching for Resonances

## LHC Olympics 2020 R&D Dataset



# Energy Flow Polynomials for More **Model-Agnostic Anomaly Detection**



# Classification Without Labels (CWoLa)

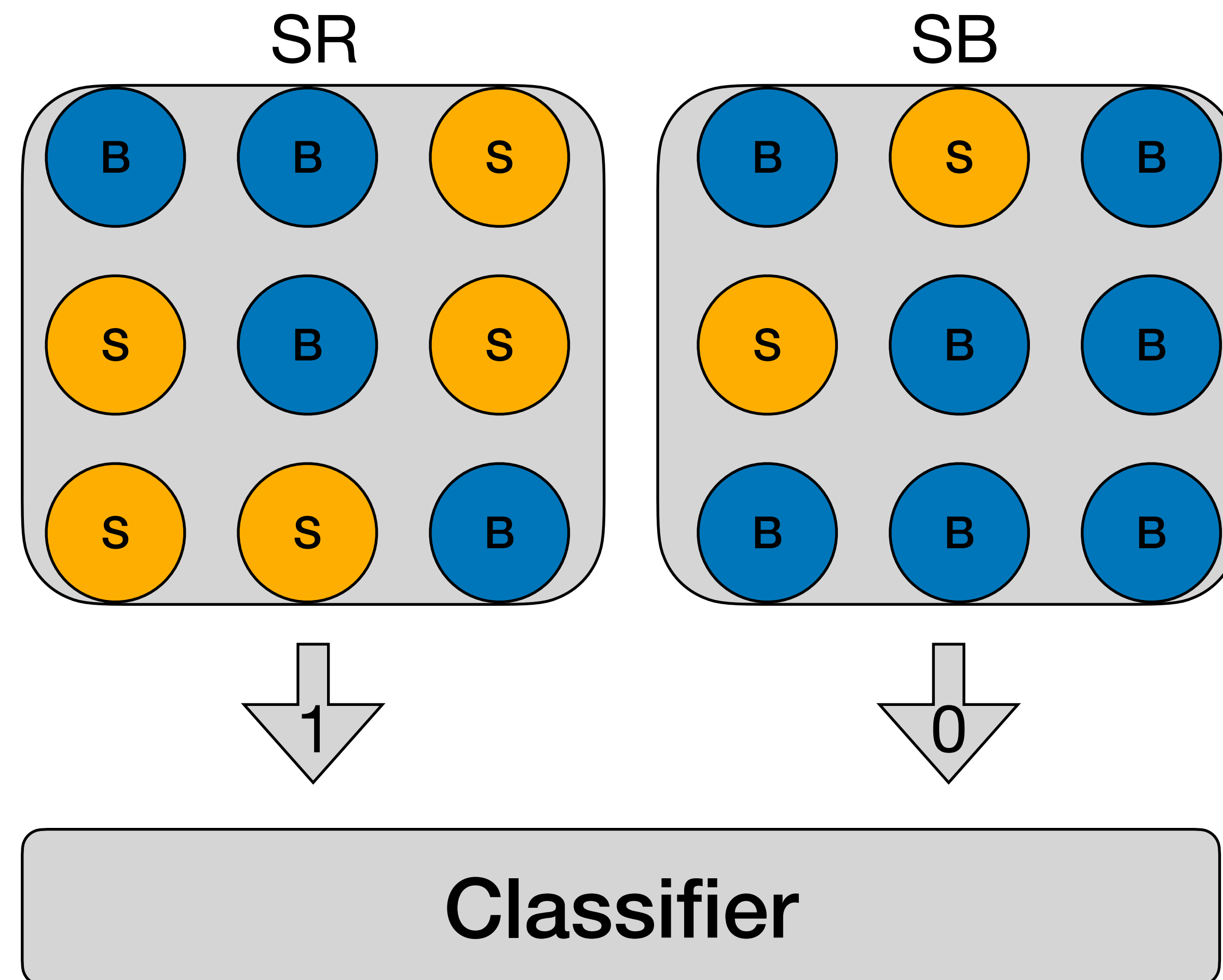
## Weakly Supervised Learning

- 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

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

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

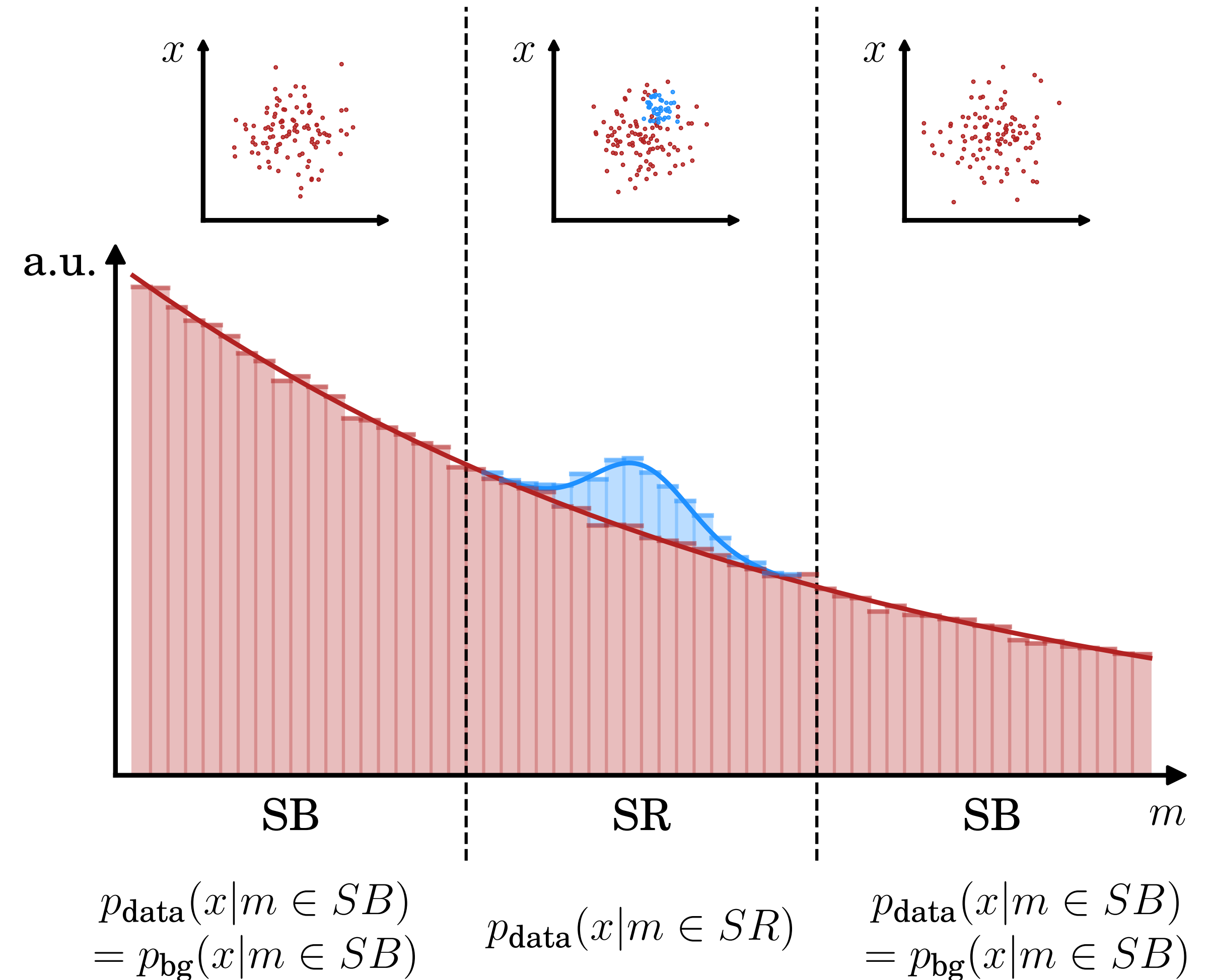
- Is monotonic for  $f_1 > f_2$



# CWoLa Hunting

## Weakly Supervised Learning

- Assuming a resonance in the invariant mass
- Defining a signal-enriched (SR) and background-enriched (SB) regions
- The Idealised Anomaly Detector (IAD)
  - SB = Pure background located in the SR
- The IAD is a more realistic case than full supervision





# LHC Olympics 2020 R&D Dataset

## LHC Olympics 2020 R&D datasets

2-prong signal

$$Z' \rightarrow XY$$

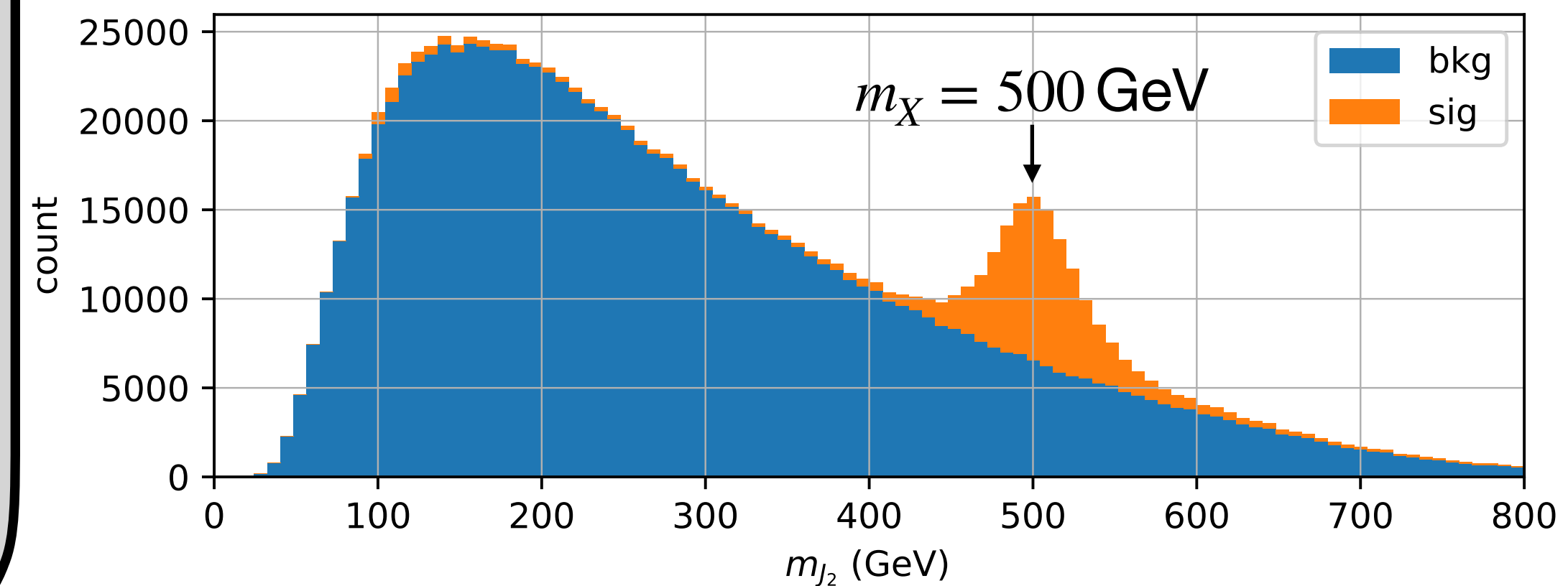
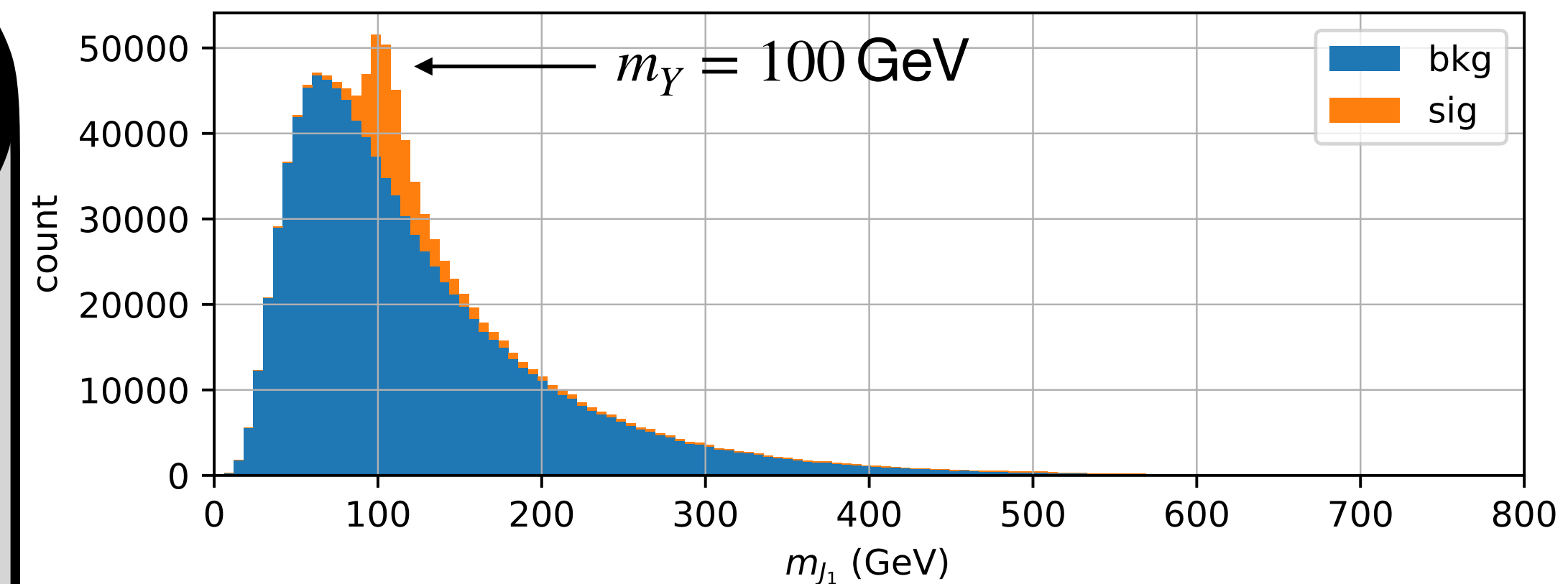
$$X \rightarrow qq \text{ \& } Y \rightarrow qq$$

3-prong signal

$$Z' \rightarrow XY$$

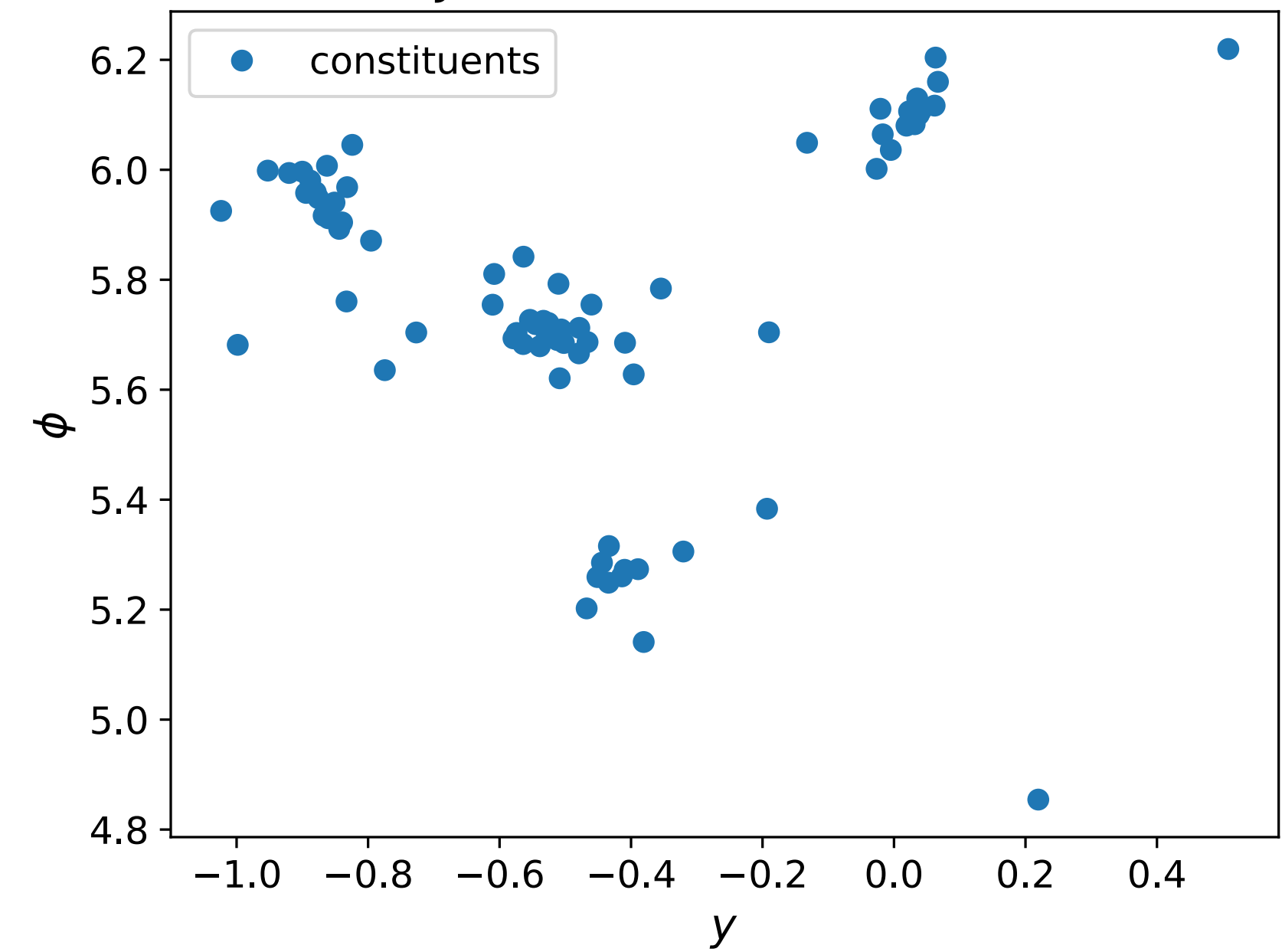
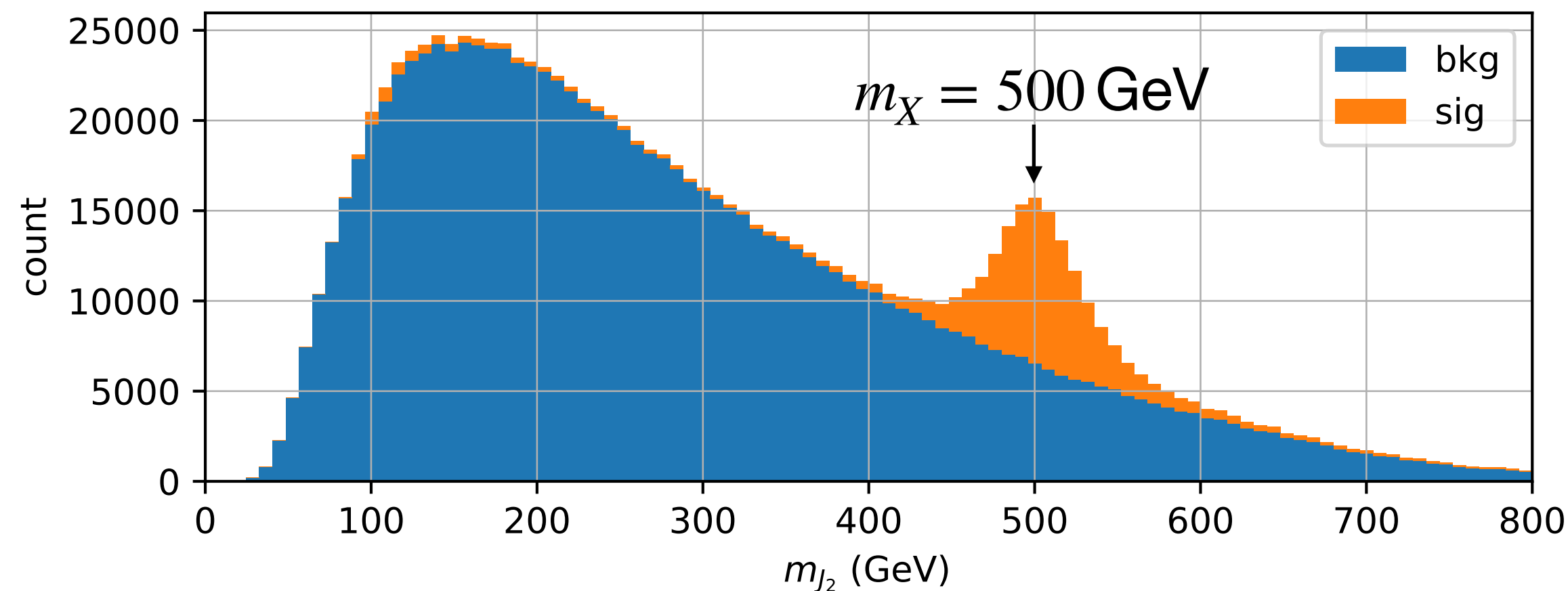
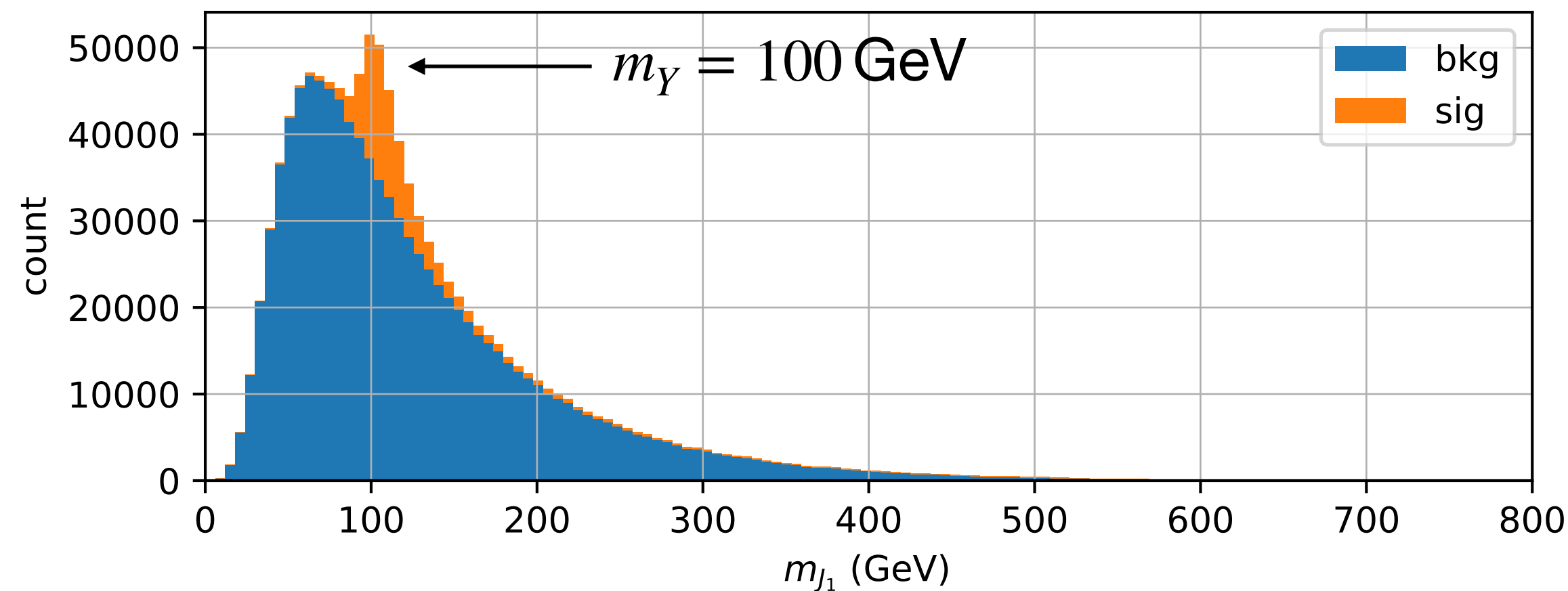
$$X \rightarrow qqq \text{ \& } Y \rightarrow qqq$$

Background QCD dijets



# Jet Observables

## Jet Mass



- Rapidity

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

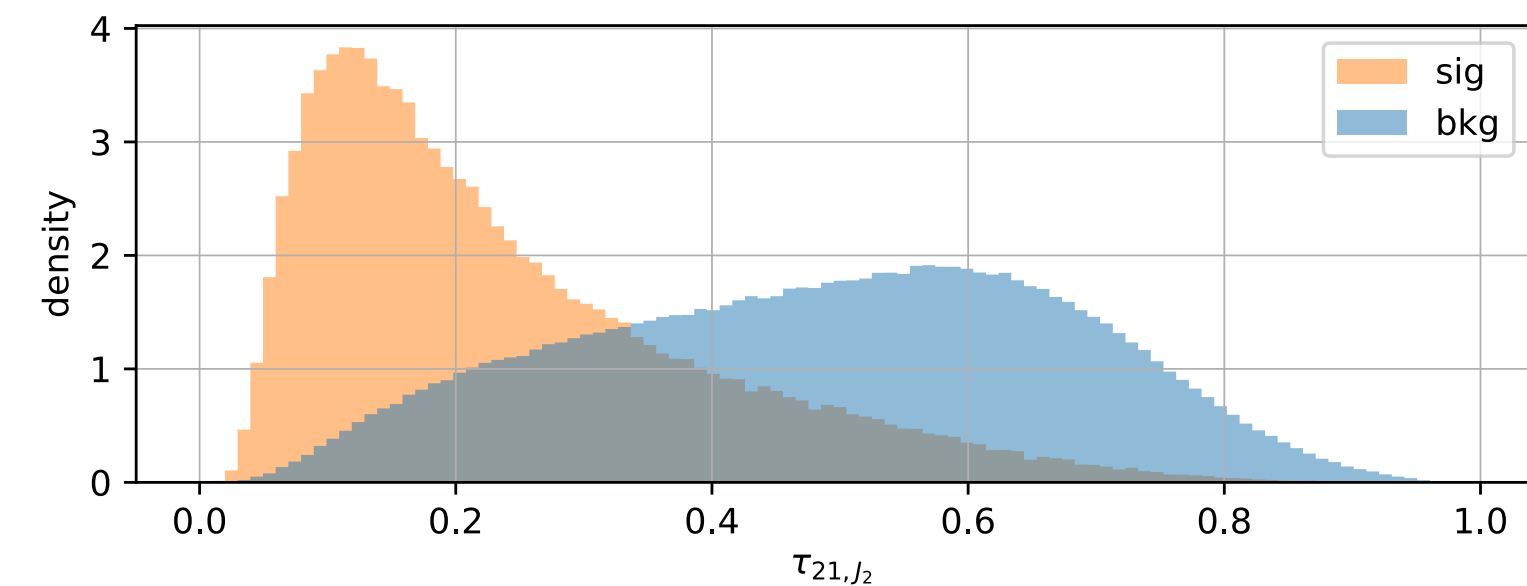
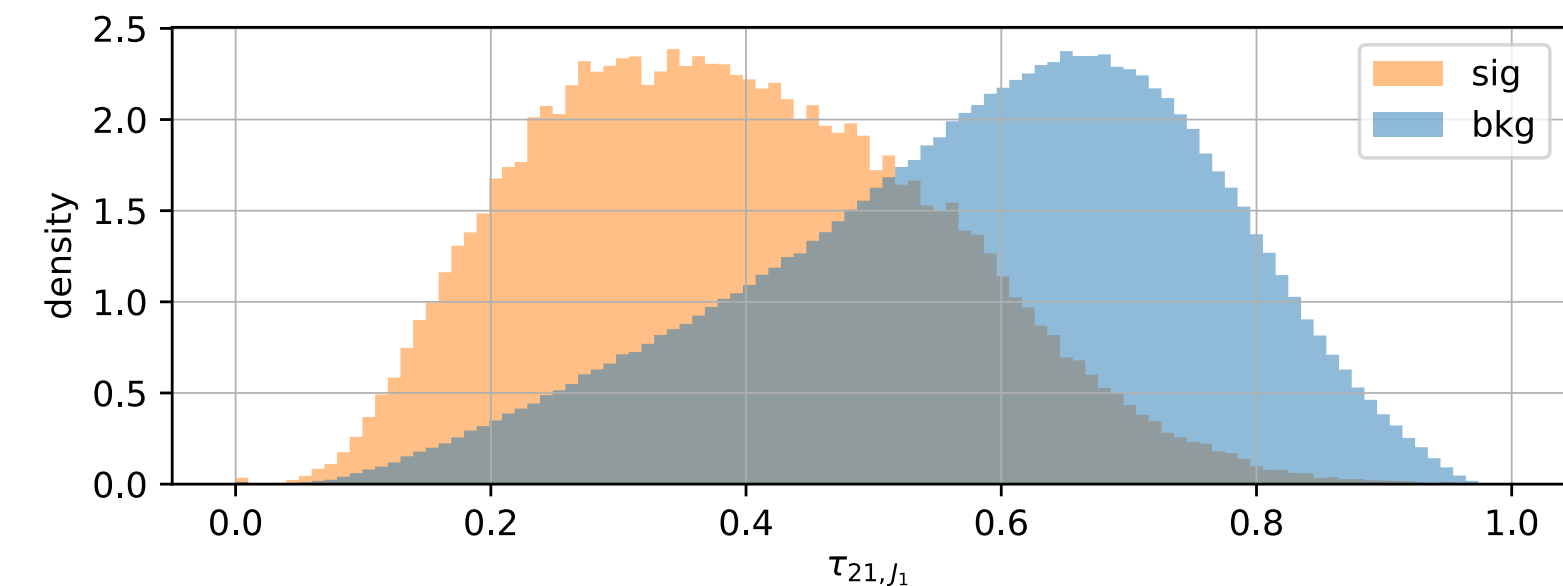
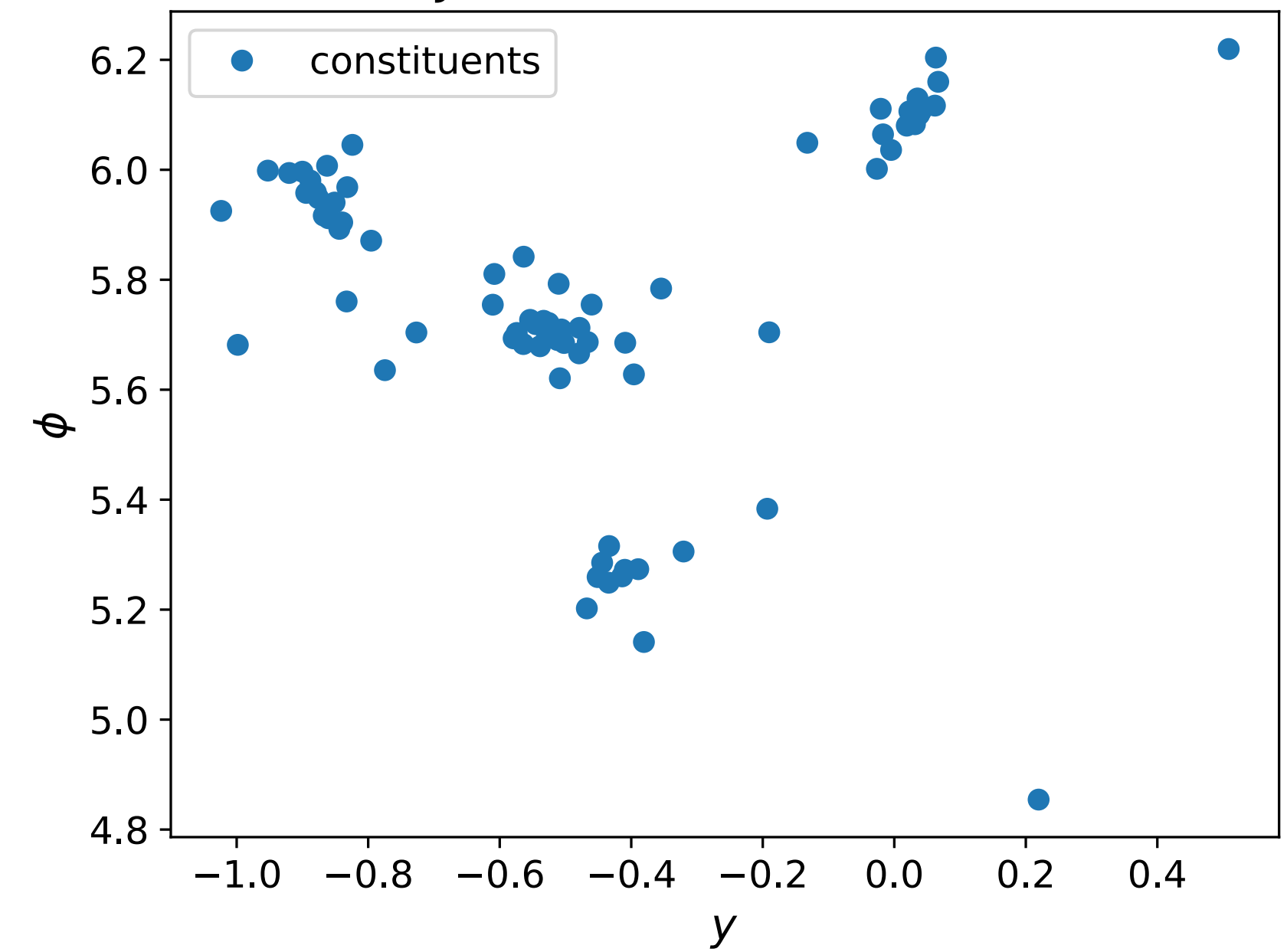
- Azimuthal angle

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

# 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 ratio  $\tau_{NM} = \frac{\tau_N}{\tau_M}$  for  $N > M$



# **Energy Flow Polynomials** for More Model-Agnostic Anomaly Detection

# What are Energy Flow Polynomials?

## 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](mailto:pkomiske@mit.edu), [metodiev@mit.edu](mailto:metodiev@mit.edu), [jthaler@mit.edu](mailto: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:1712.07124v2 [hep-ph] 3 Apr 2018

# Mathematical representation

## Energy Flow Polynomials (EFPs)

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

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

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

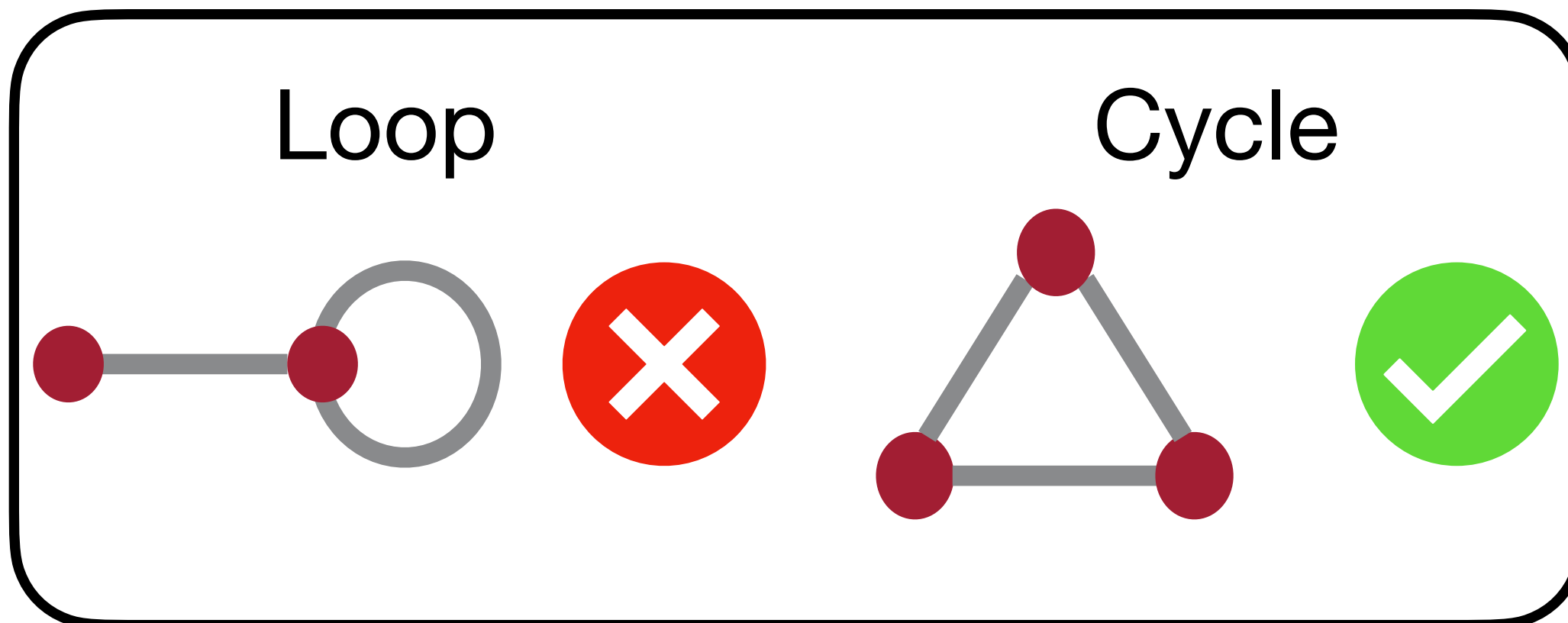
$$\Delta \phi_{ij} \equiv \phi_i - \phi_j$$



# Multigraph correspondence

## Energy Flow Polynomials (EFPS)

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



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

Vertex correspondence

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

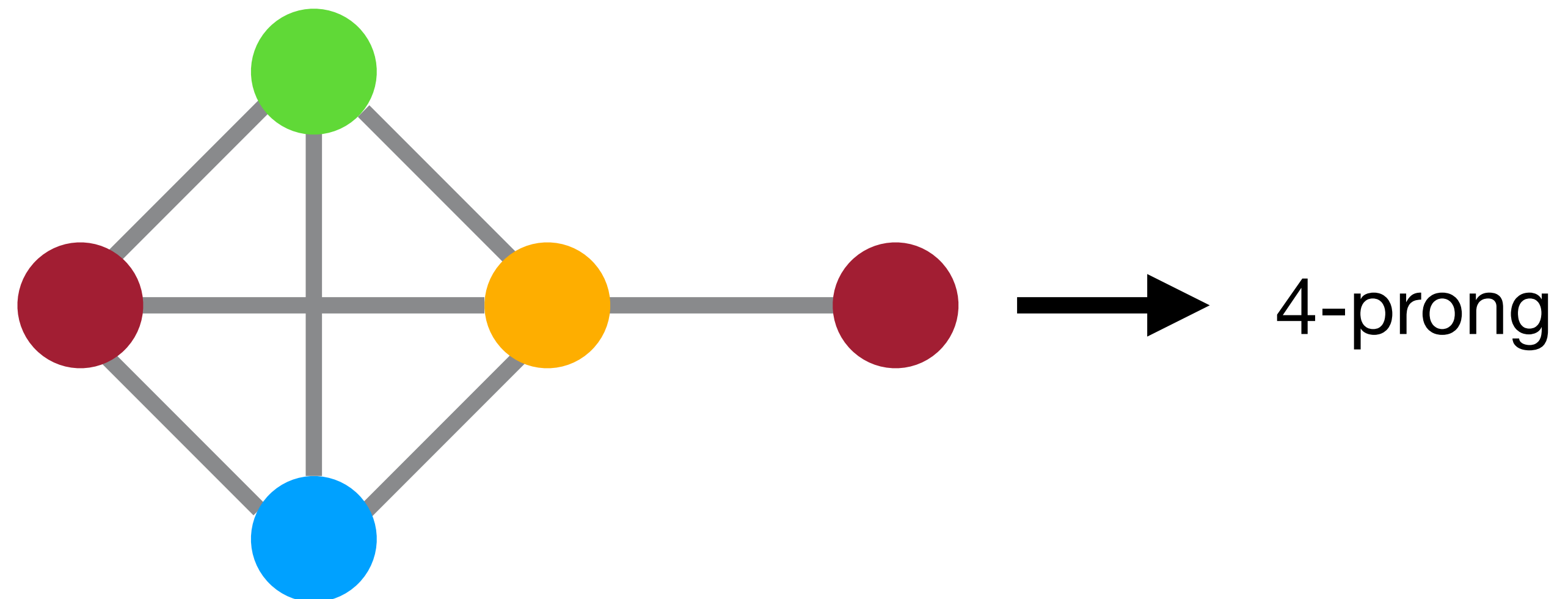
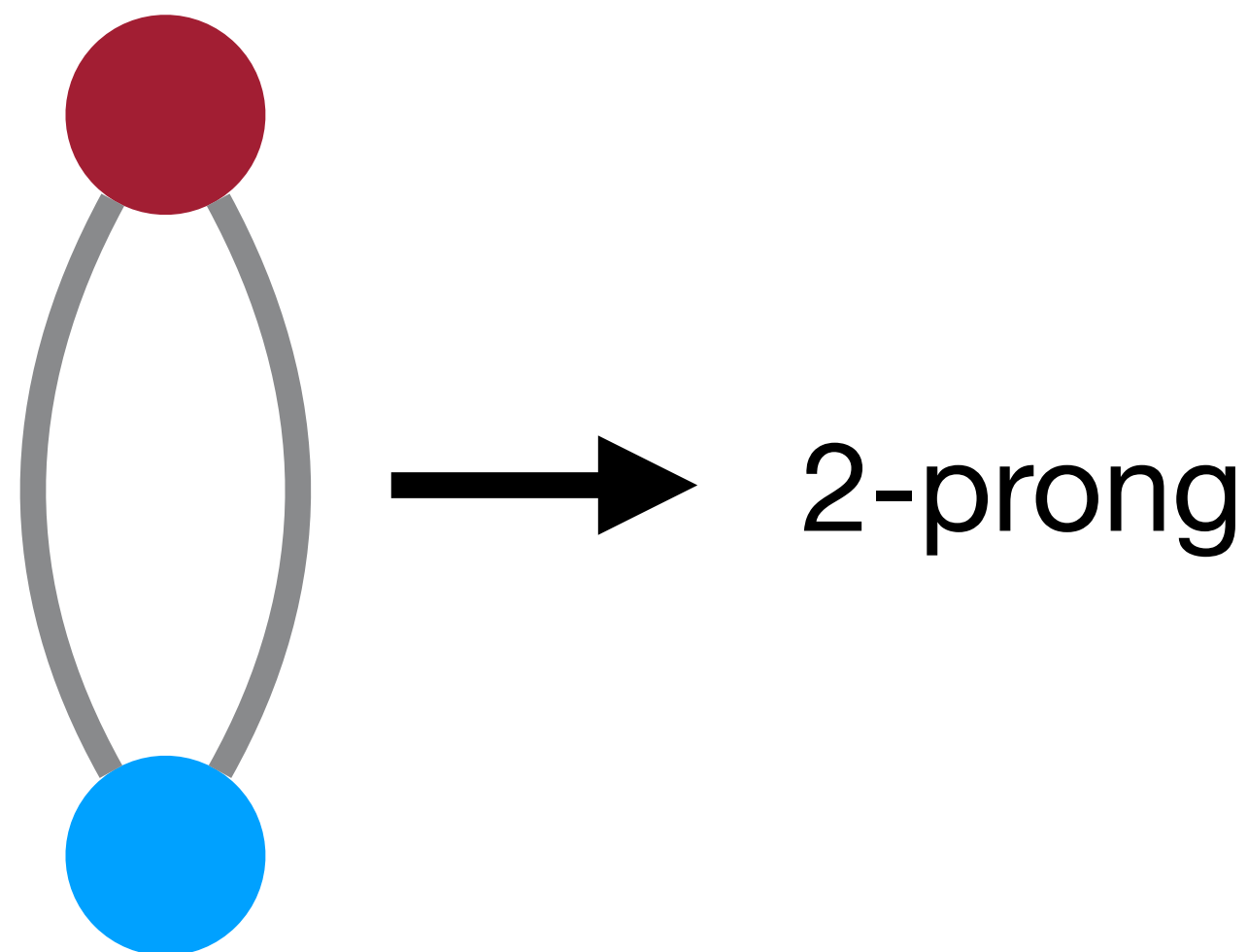
Edge correspondence

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

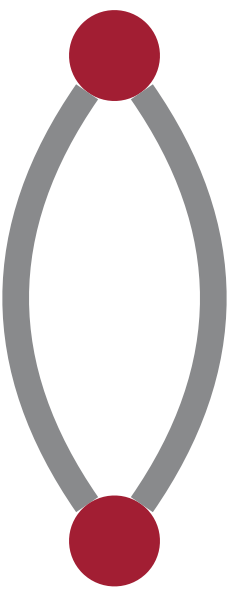
# Chromatic number

## Energy Flow Polynomials (EFPs)

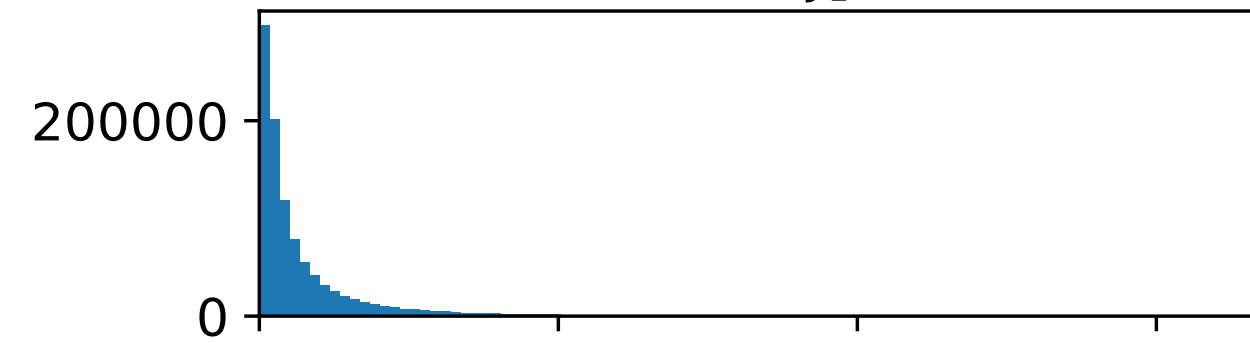
- The smallest number of colors needed to color vertices so that connected vertices do not have the same color
- The chromatic number corresponds to the number of separated prongs for which an EFP is first non-vanishing



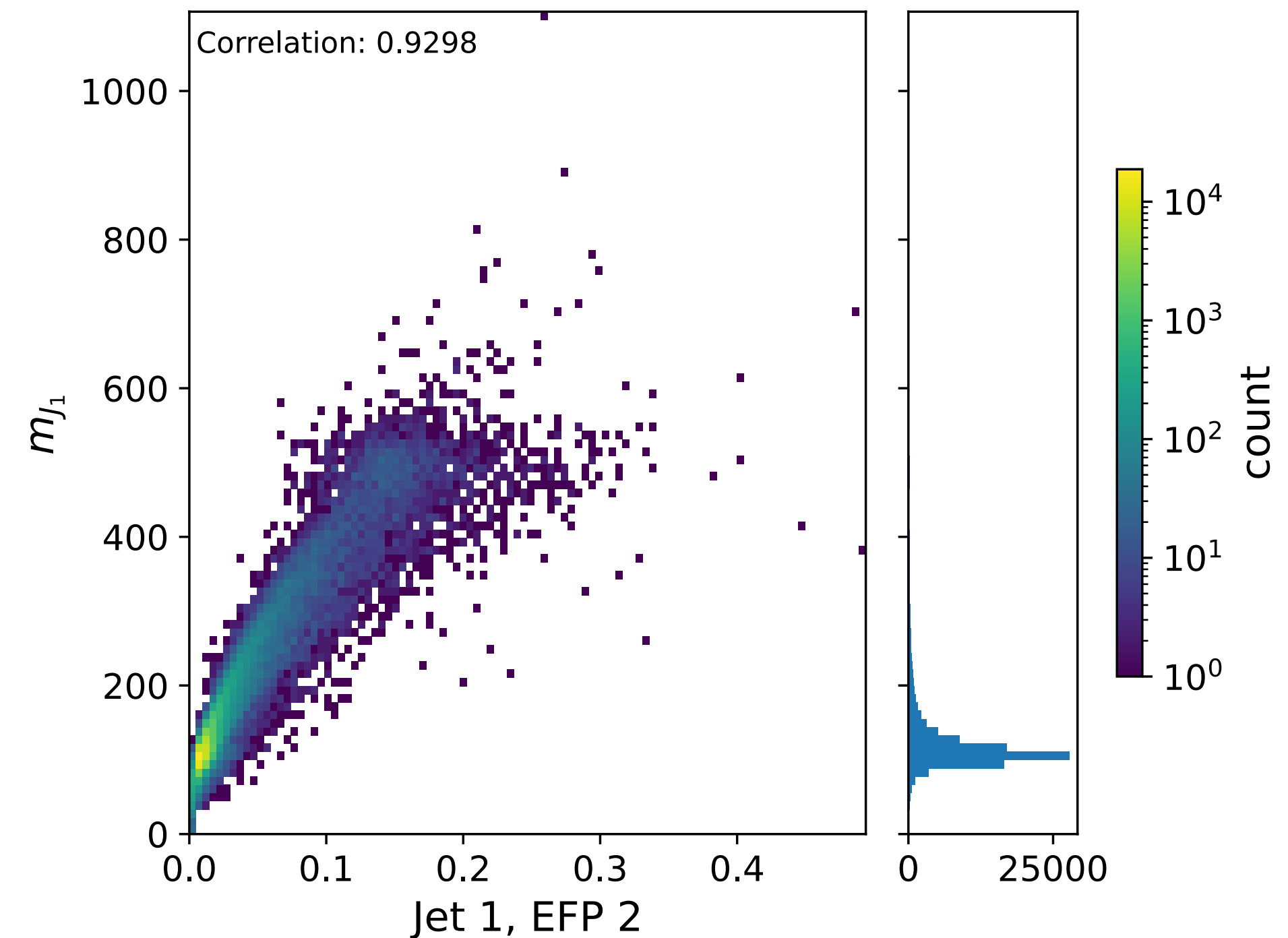
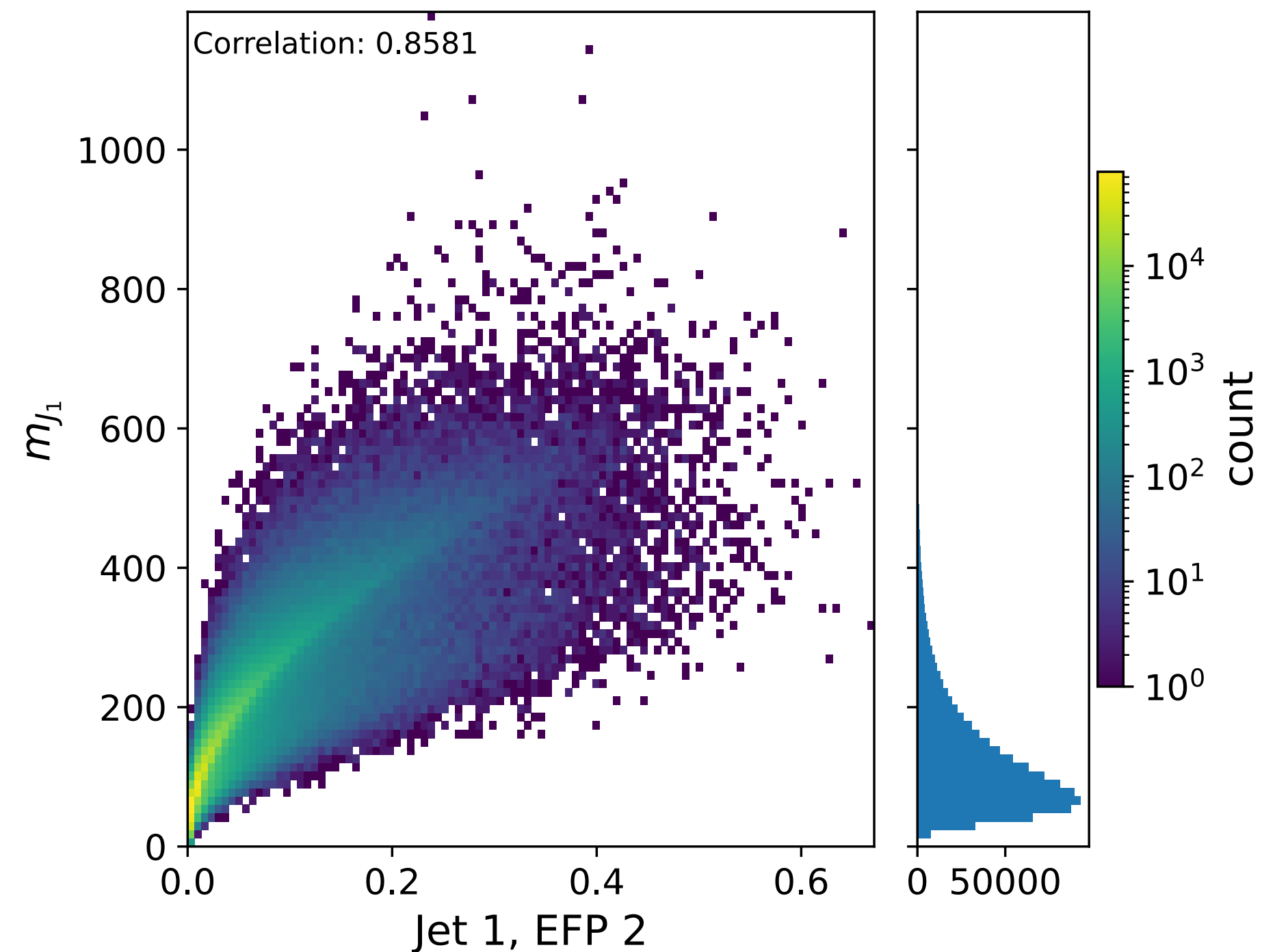
# Example: Jet Mass Energy Flow Polynomials (EFPs)

$$\text{EFP}_2 = \text{Diagram} \approx \frac{2m_J^2}{p_{T,J}^2}$$


Jet 1, EFP 2 vs  $m_{J_1}$  Background



Jet 1, EFP 2 vs  $m_{J_1}$  Signal

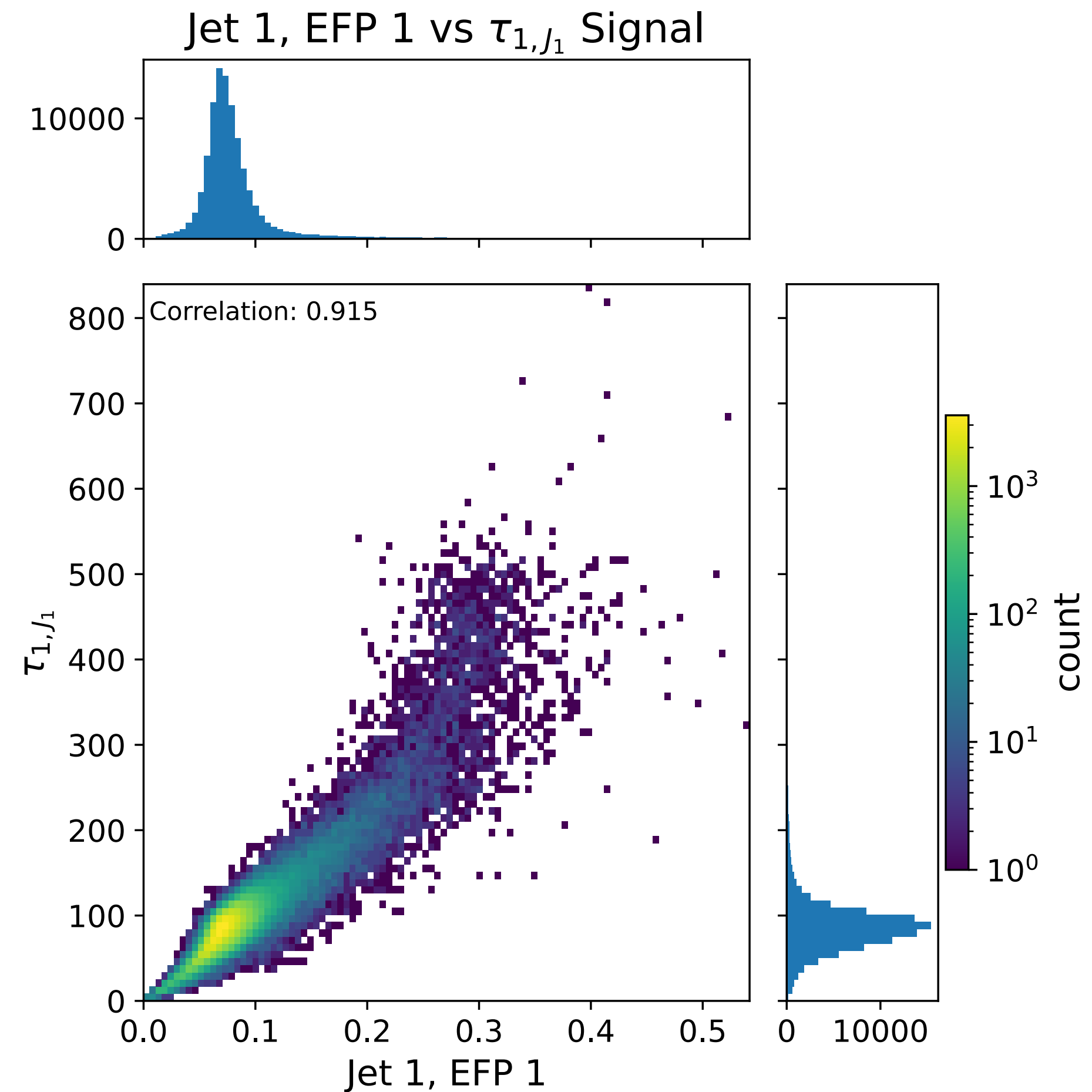
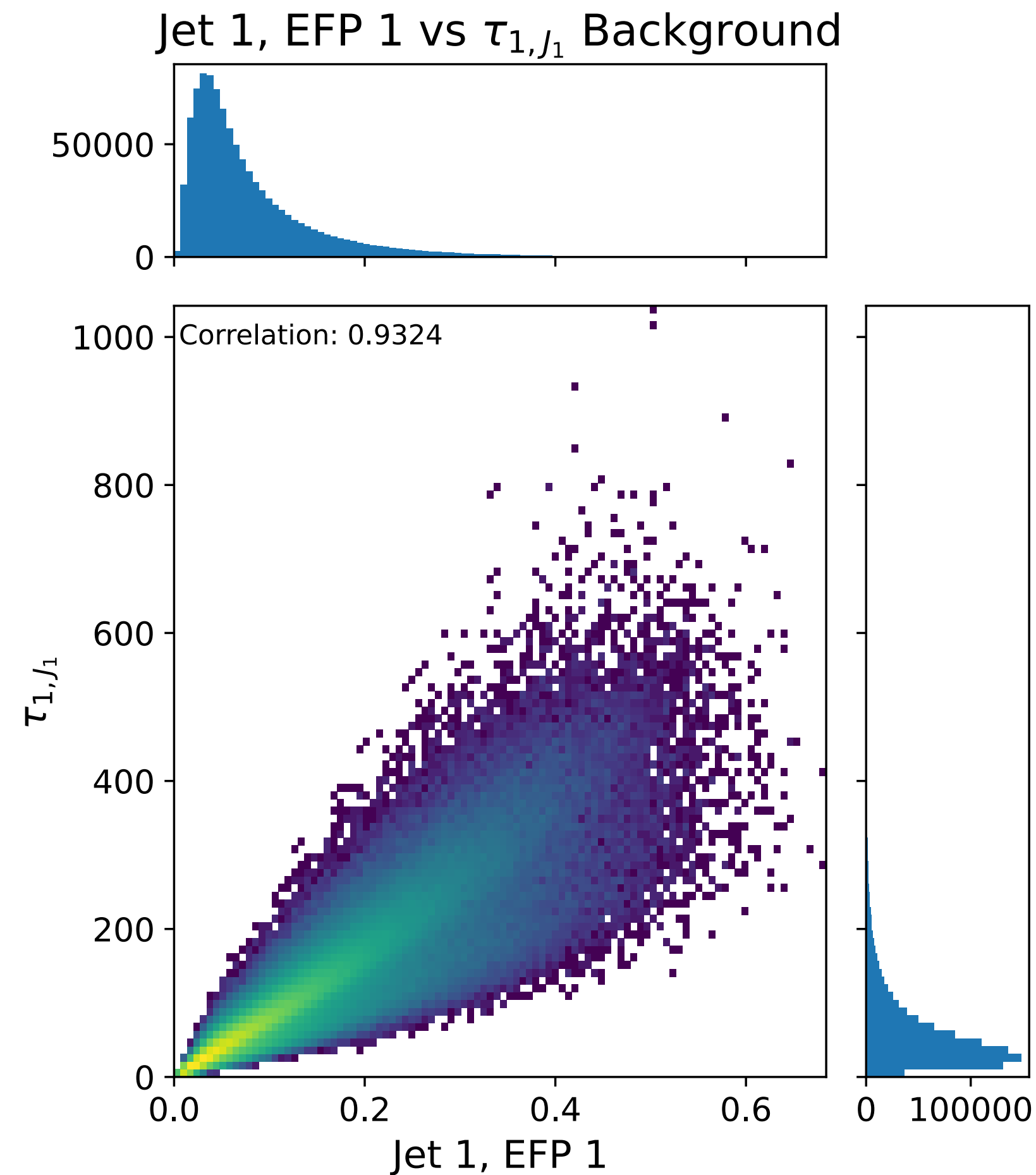


# Example: 1-Subjettiness

## Energy Flow Polynomials (EFPs)

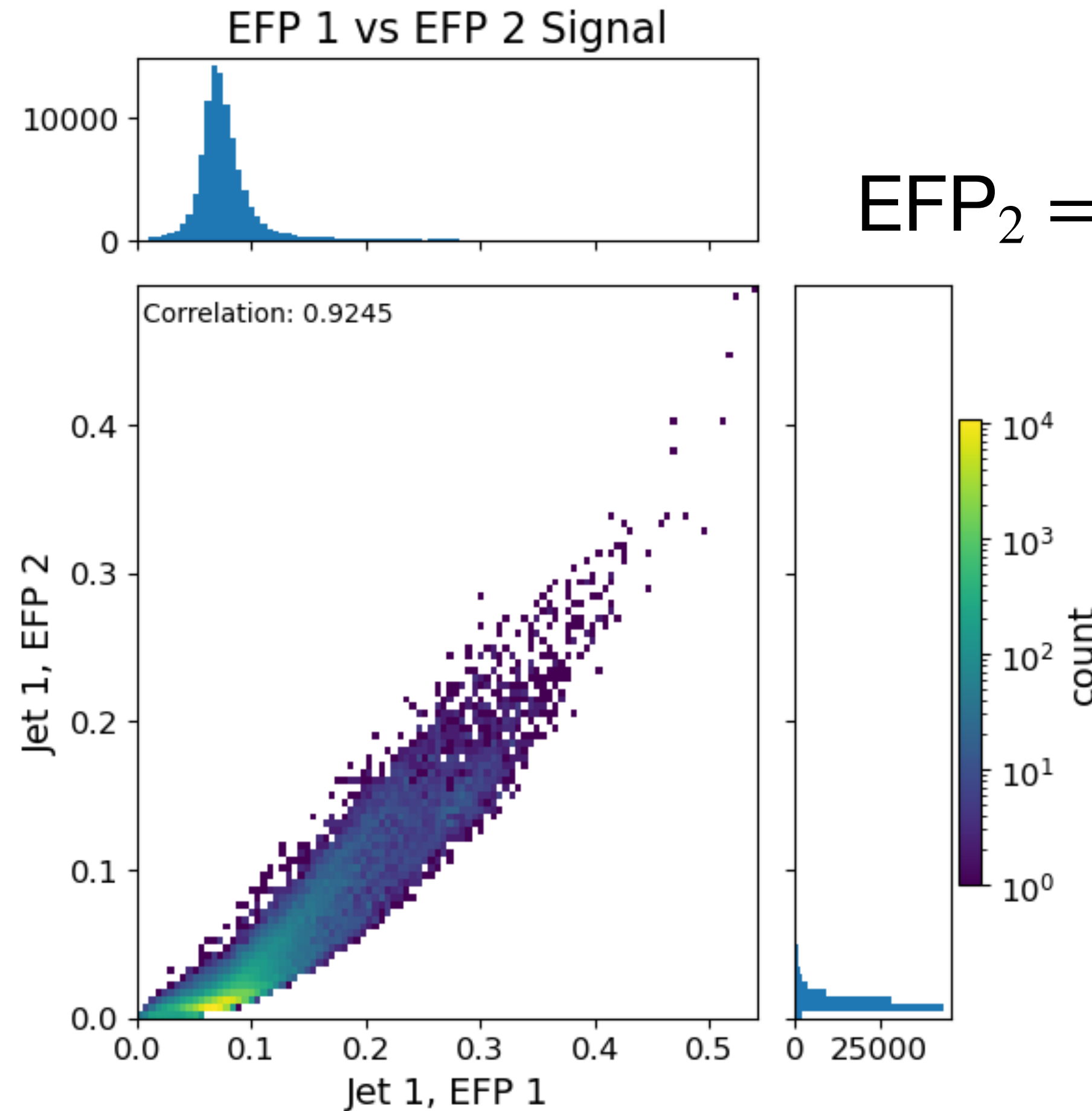
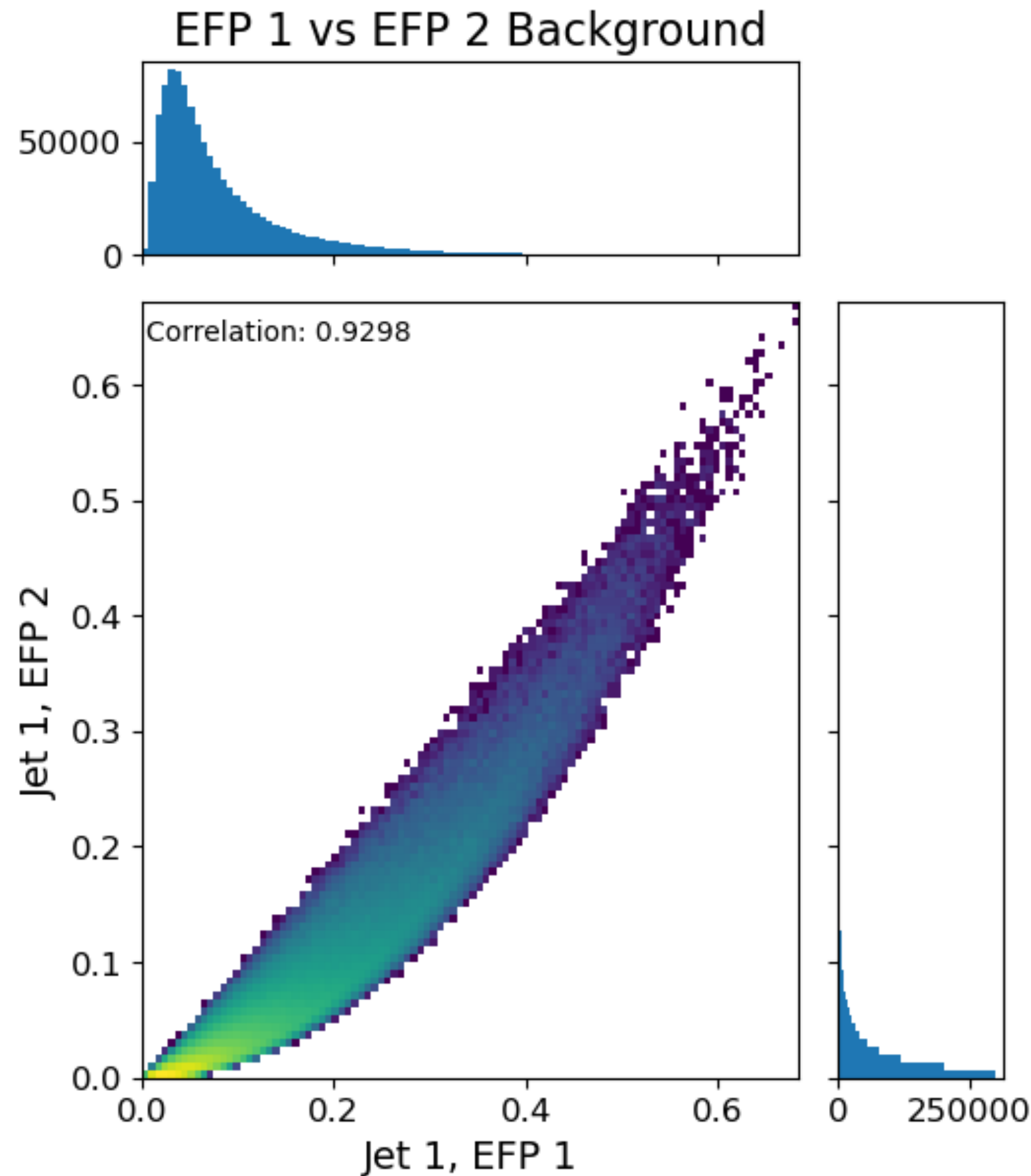
$$\text{EFP}_1 = \text{Diagram} = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}$$

The diagram represents a vertical line segment with two red circular nodes at its ends, connected by a thick grey line.



# Example: Different EFPs

## Energy Flow Polynomials (EFPs)



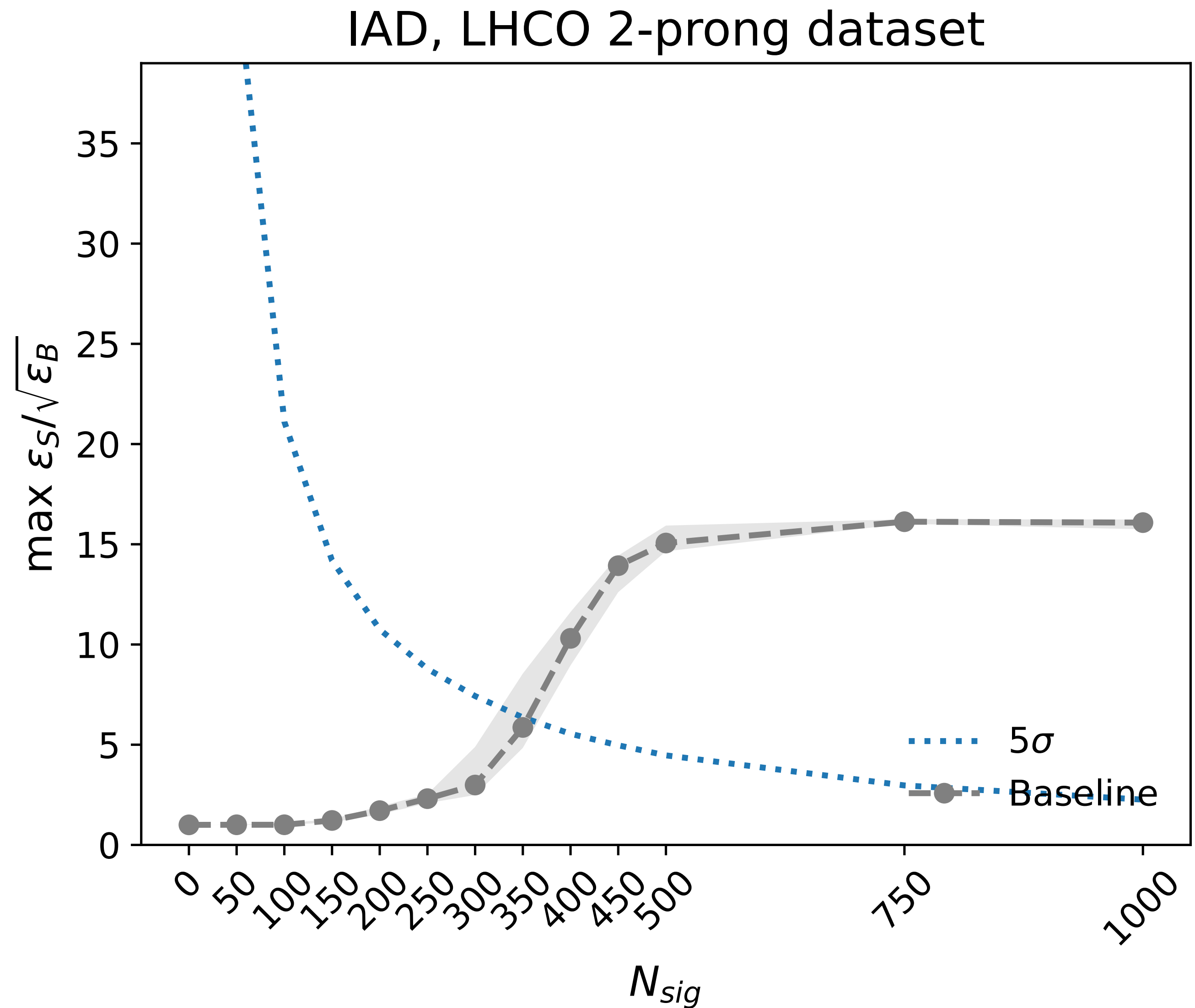
$$\text{EFP}_1 = \text{Diagram 1} = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}$$

$$\text{EFP}_2 = \text{Diagram 2} = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}^2$$

# Idealised Anomaly Detector (IAD)



# Evaluating Classifiers



# Evaluating Classifiers

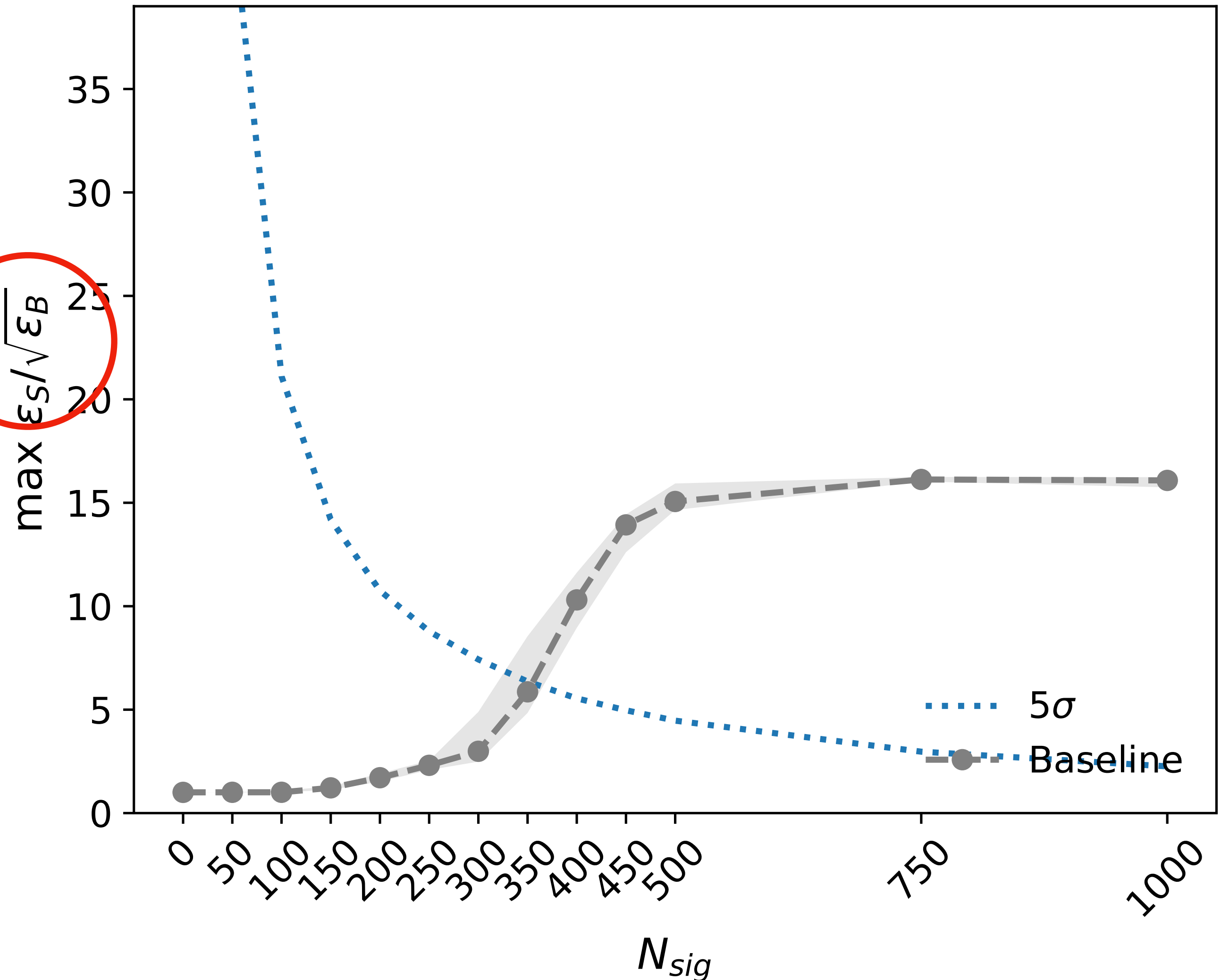
Poisson Significance Improvement Characteristic (SIC)

$$\text{SIC} \cdot \sigma_{\text{poisson}} = \sigma_{\text{cut}}$$



$\max \varepsilon_s / \sqrt{\varepsilon_B}$

IAD, LHCO 2-prong dataset



# Evaluating Classifiers

Poisson Significance  
Improvement Characteristic  
(SIC)

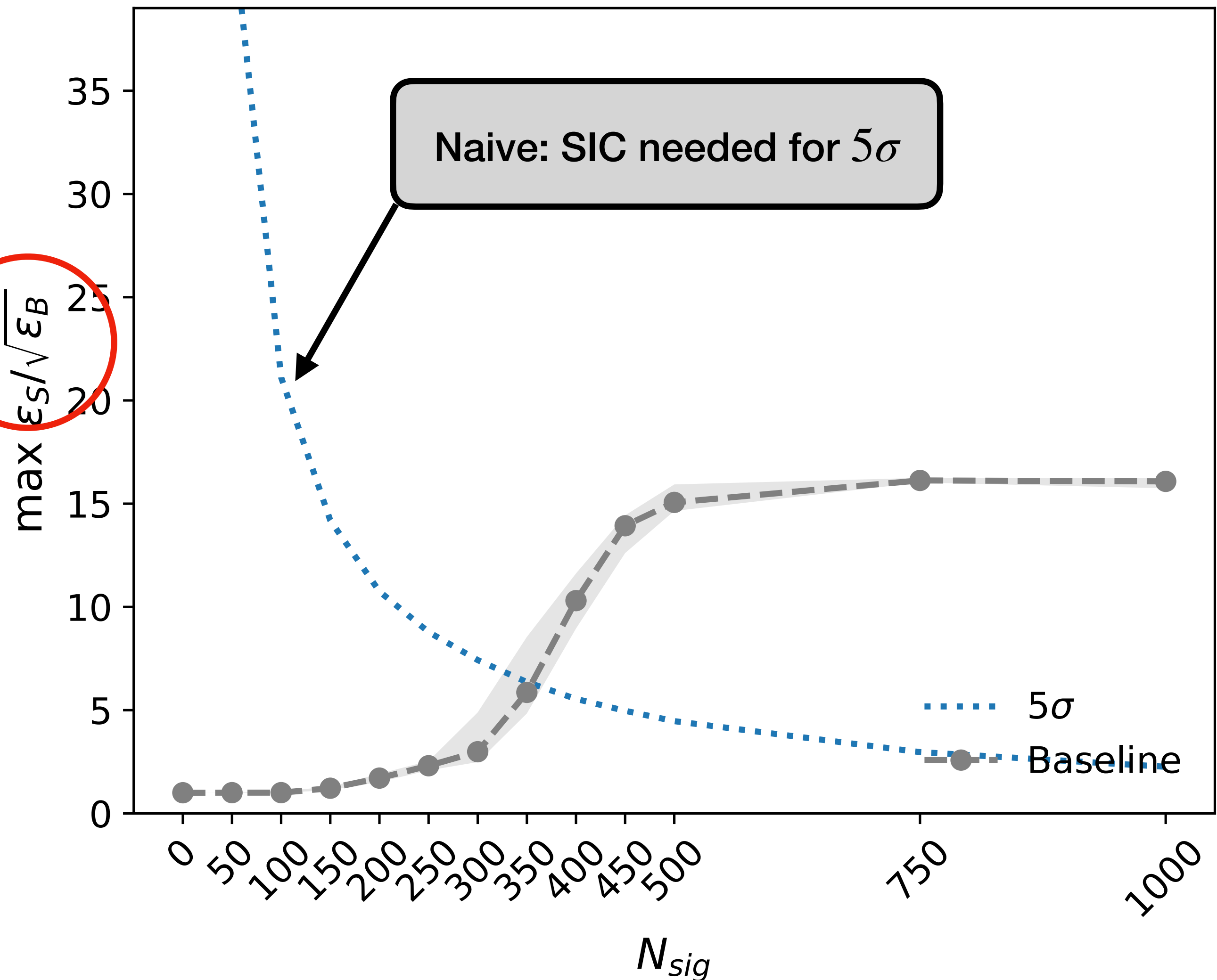
$$\text{SIC} \cdot \sigma_{\text{poisson}} = \sigma_{\text{cut}}$$

Baseline:

$$\{m_{J_1}, \Delta m_J, \tau_{21,J_1}, \tau_{21,J_2}\}$$

$$\max \varepsilon_s / \sqrt{\varepsilon_B}$$

IAD, LHCO 2-prong dataset



# Signal Injections

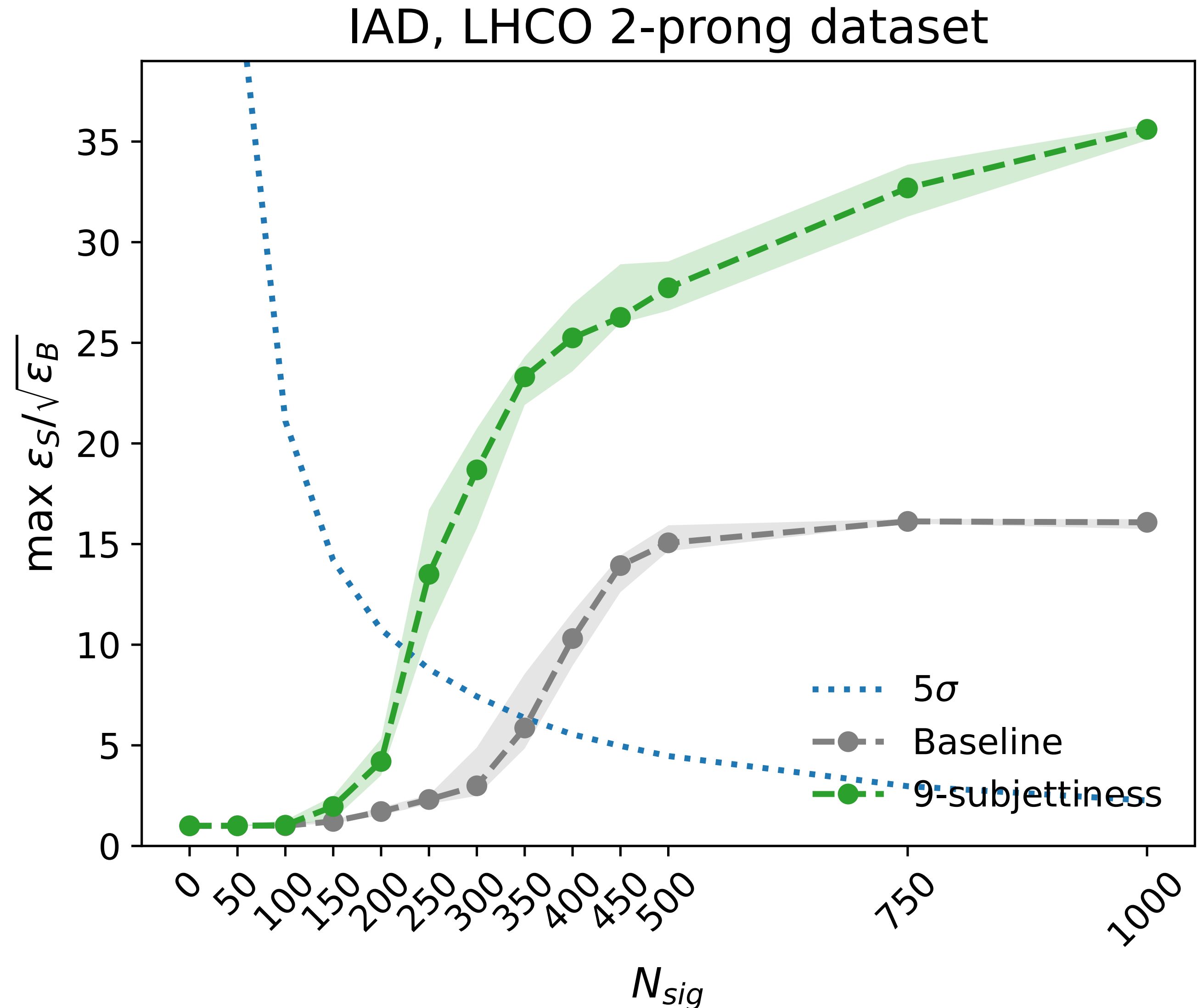
Baseline:

$$\{m_{J_1}, \Delta m_J, \tau_{21,J_1}, \tau_{21,J_2}\}$$

9-Subjettiness:

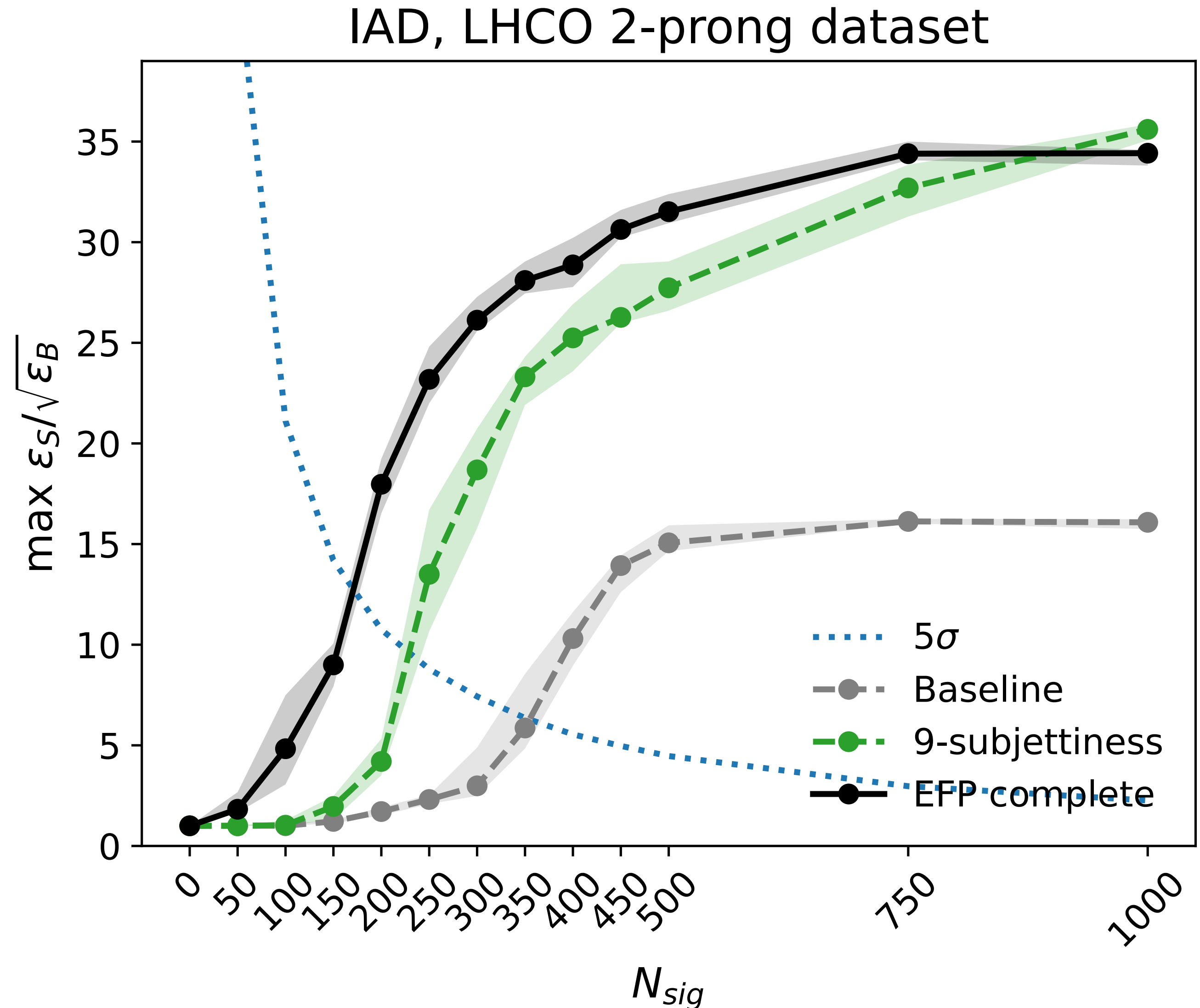
$$\{m_{J_1}, \Delta m_J, \tau_{N,J_1}, \tau_{N,J_2}\}$$

for  $N \leq 9$



# Signal Injections

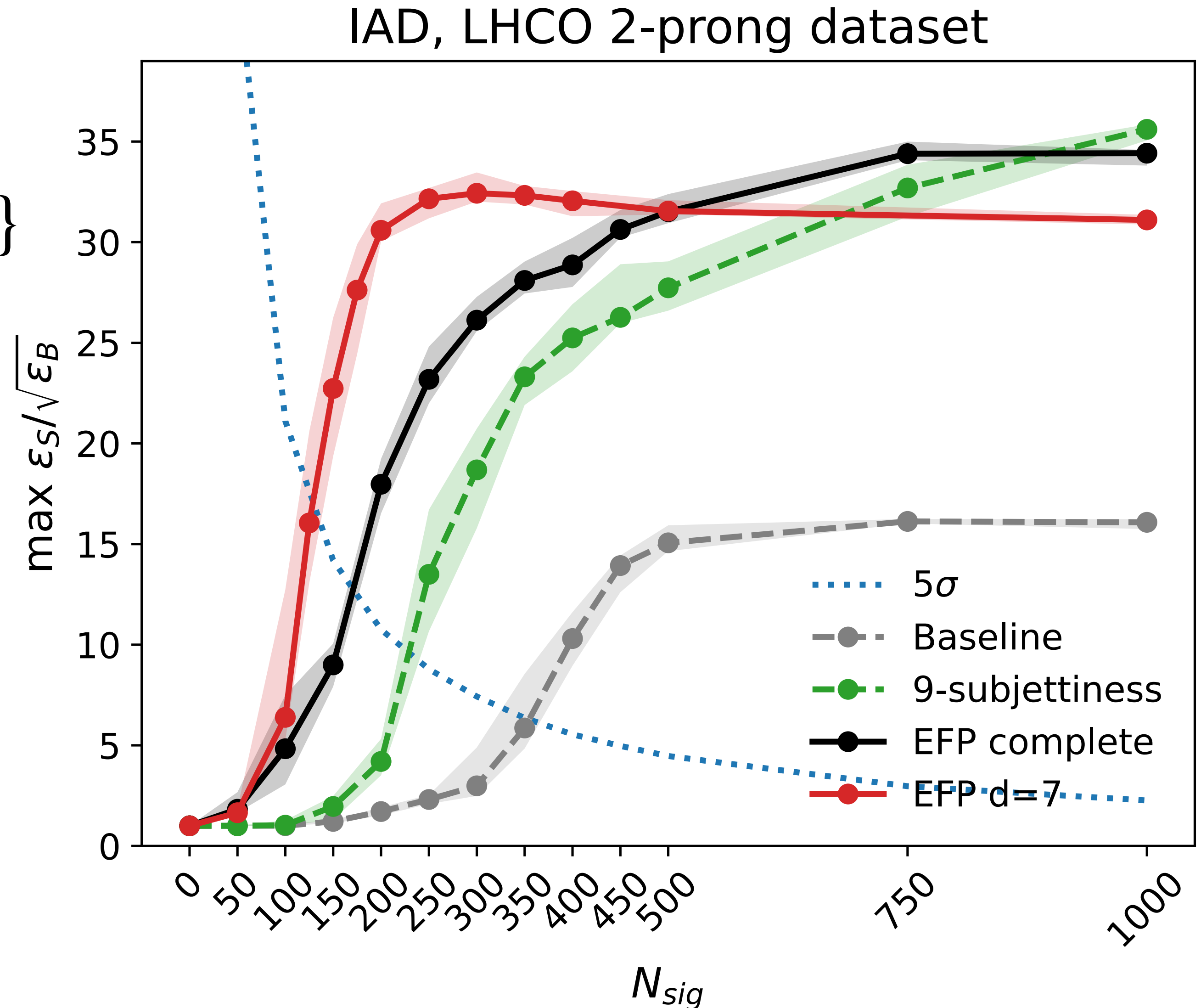
EFP Complete:  
 $\{m_{J_1}, \Delta m_J, \text{EFP}_{i,J_1}, \text{EFP}_{i,J_2}\}$   
 for  $i \in [1, 489]$



# Signal Injections

- $\{m_{J_1}, \Delta m, EFP_{7,J_1}, EFP_{7,J_2}\}$

$$EFP_7 = \begin{array}{c} \bullet \\ | \\ 7 \\ | \\ \bullet \end{array} = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}^7$$



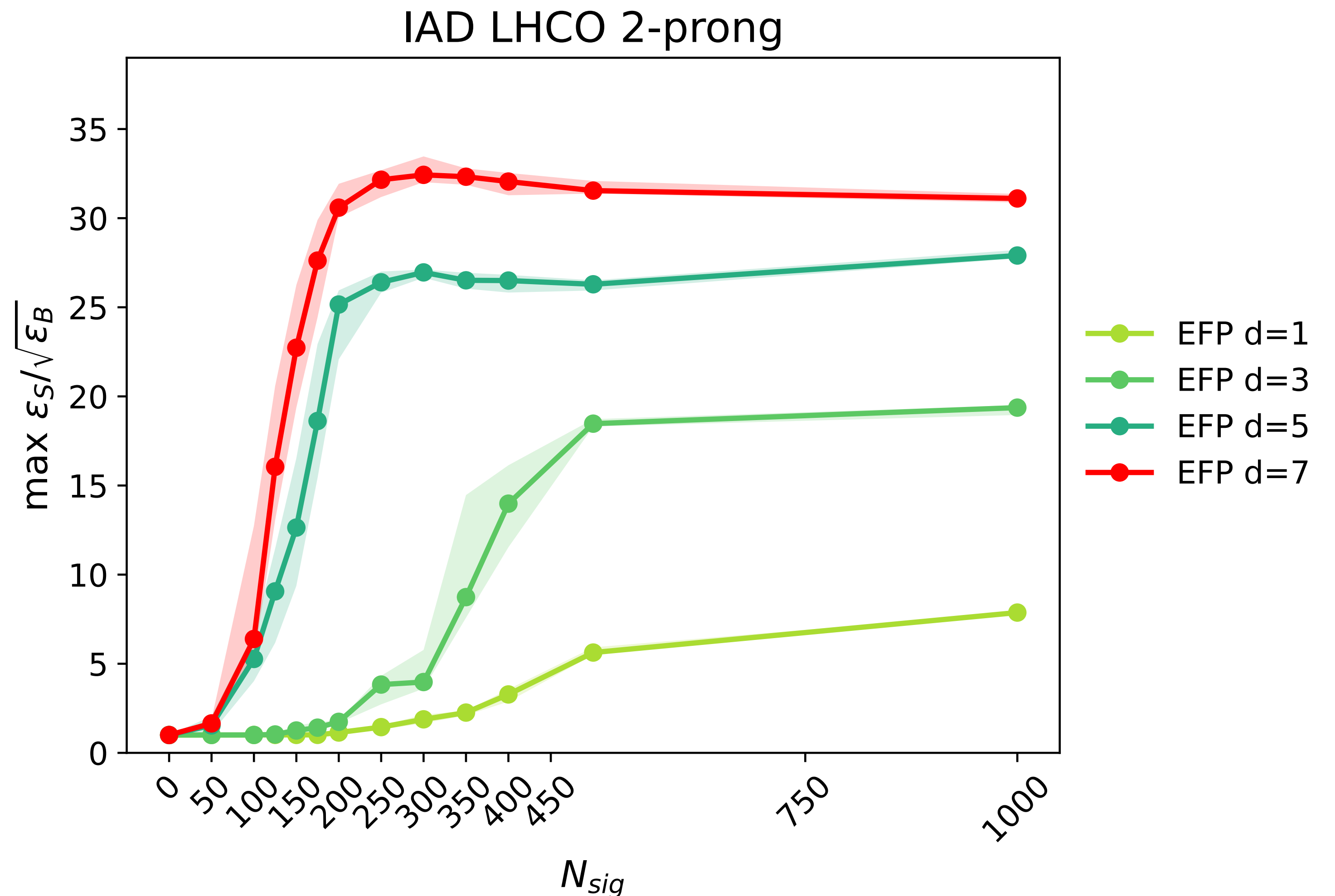


# Signal Injections

## EFP 7

$$d = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}^d$$

- In the EFP complete set we have EFPs with different structures
- The complexity is limited by computational power
- We are considering all EFPs with 7 edges or more ( $d \leq 7$ )

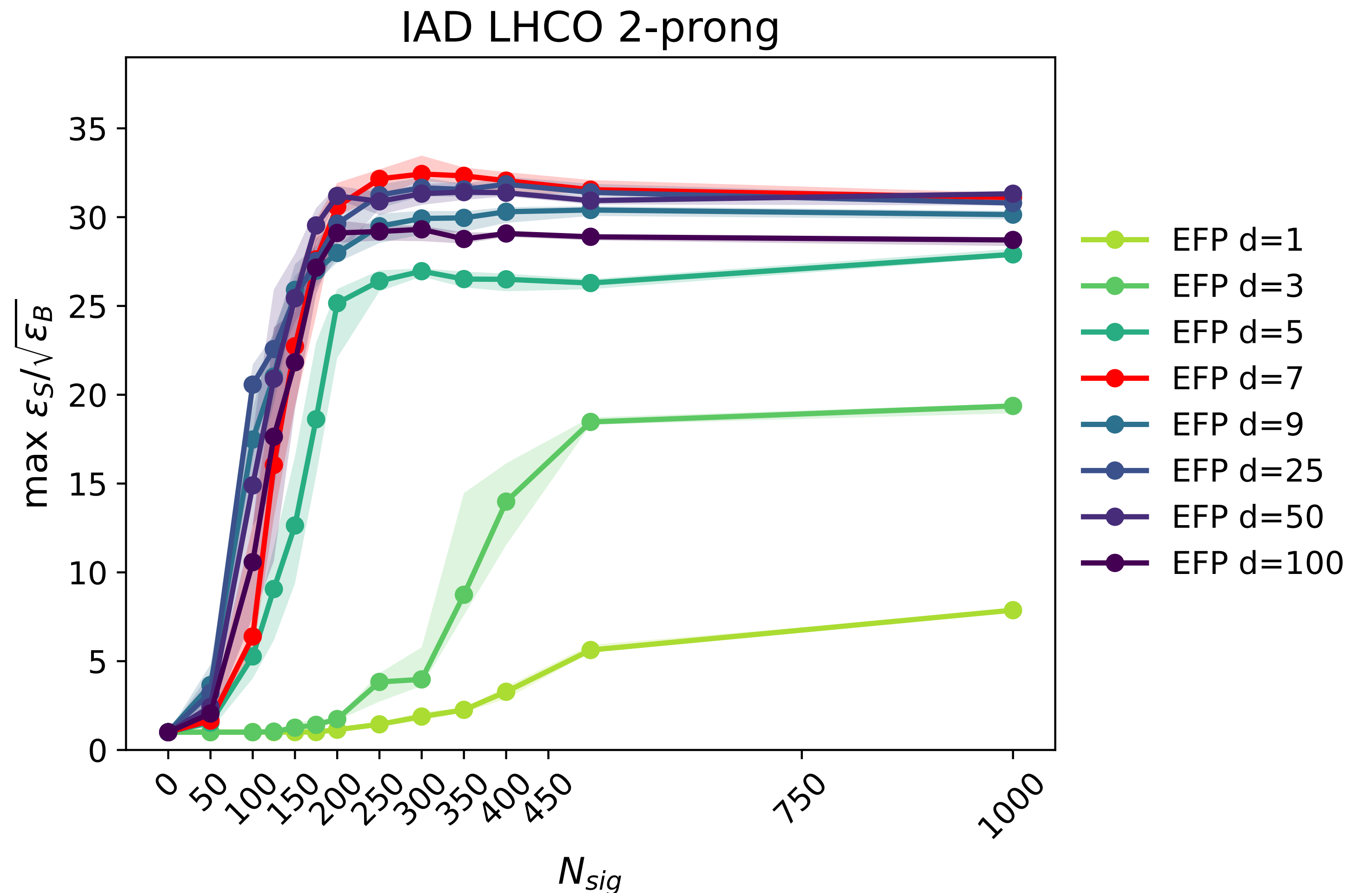


# Signal Injections

## EFP 7

$$d = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}^d$$

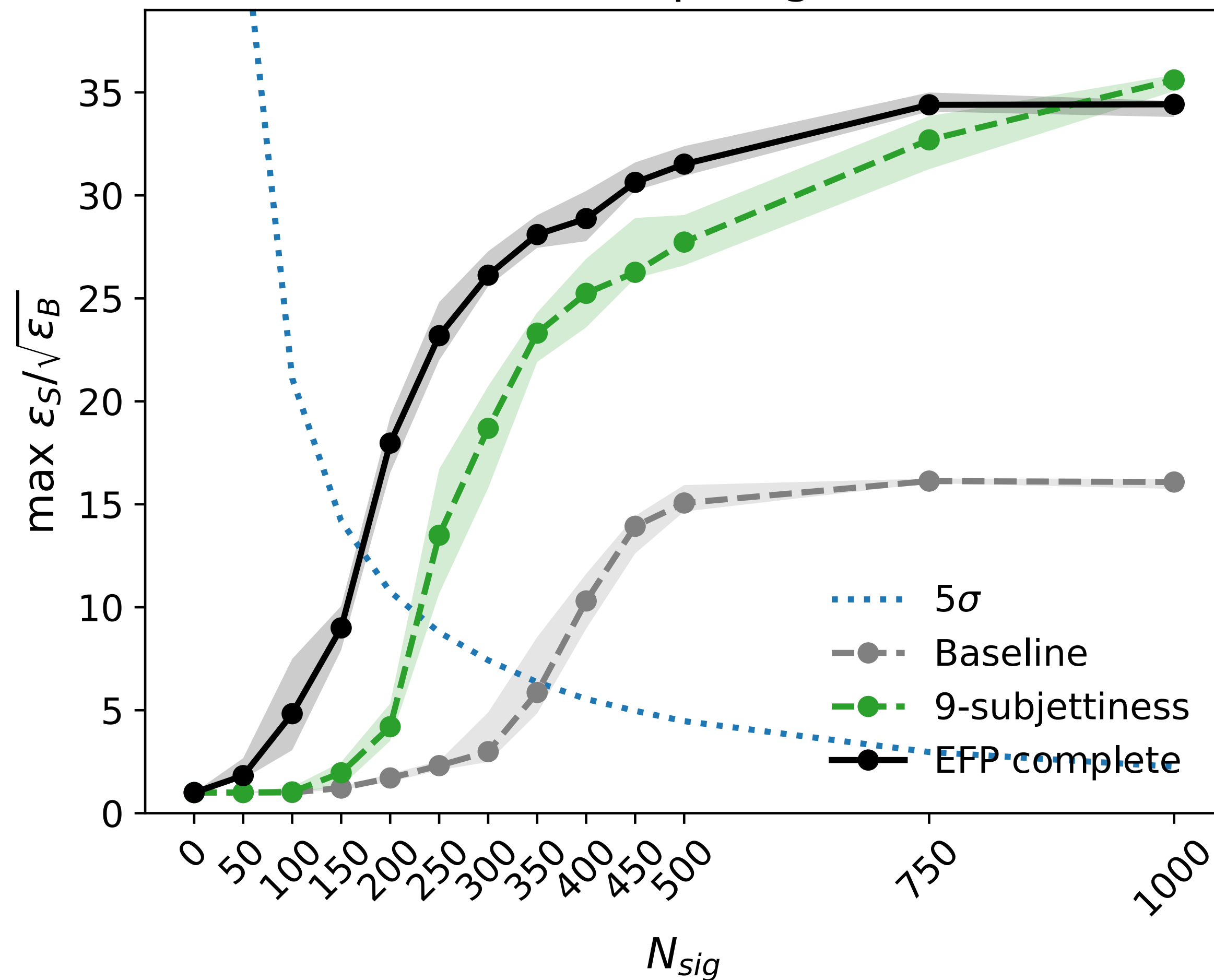
- EFP 7 is the most expressive in the EFP complete set for the LHCO 2-prong signal
- Higher number of edges do not improve the sensitivity significantly



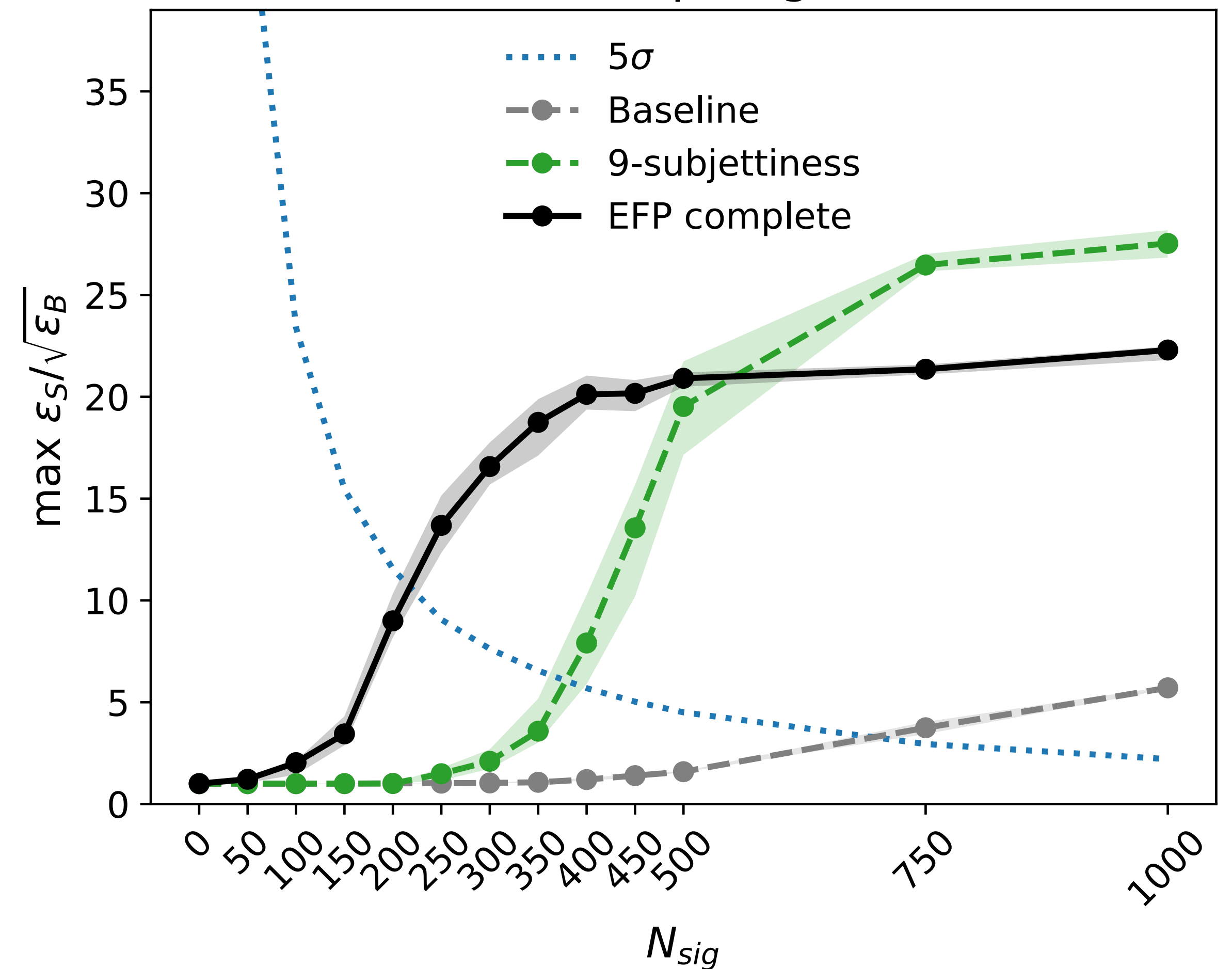
# Signal Injections

## Different Signals

IAD, LHC0 2-prong dataset



IAD, LHC0 3-prong dataset

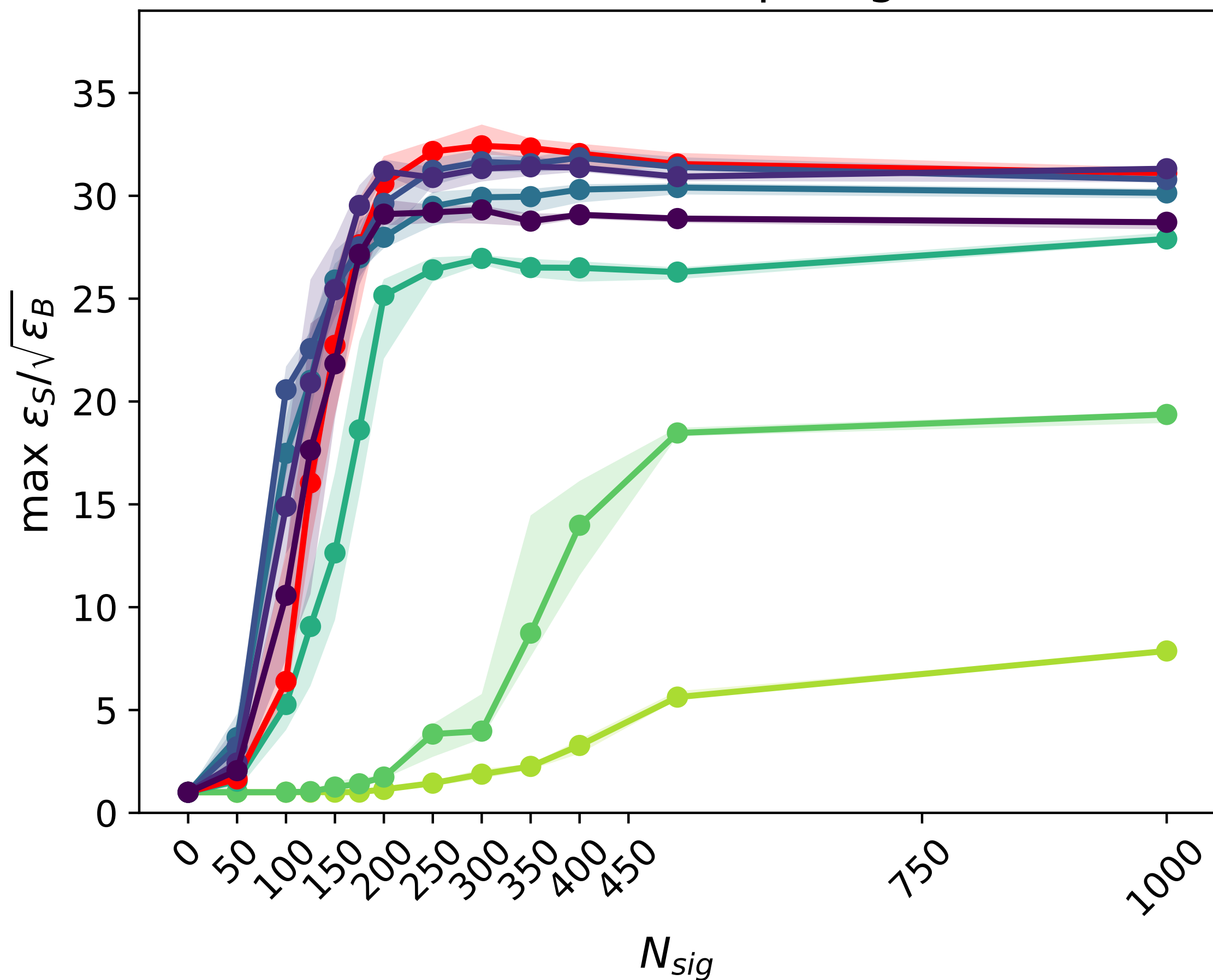


# Signal Injections

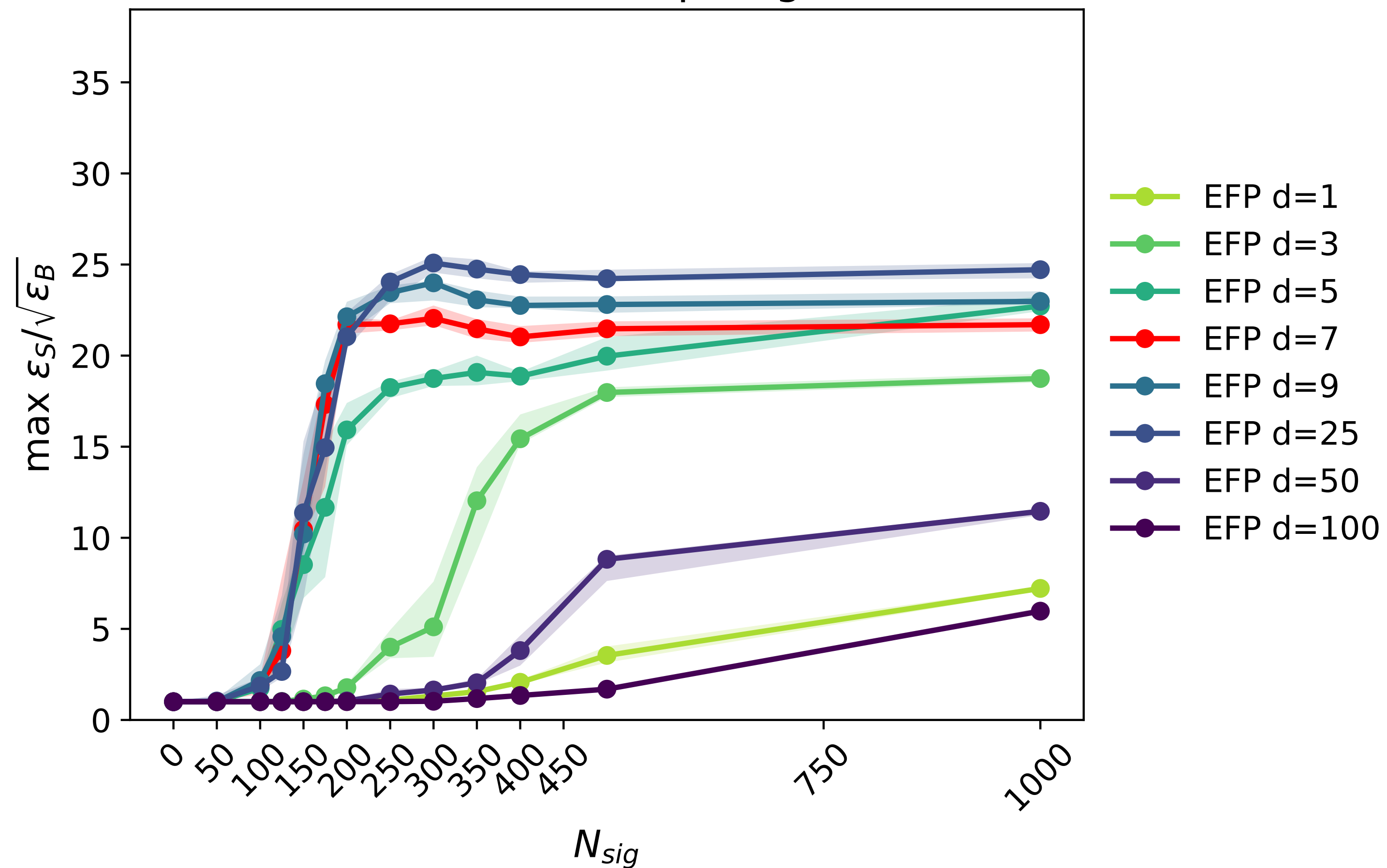
## EFP 7 for 3-prong LHCO Signal

$$d = \sum_{i_1=1}^M \sum_{i_2=1}^M z_{i_1} z_{i_2} \theta_{i_1 i_2}^d$$

IAD LHCO 2-prong

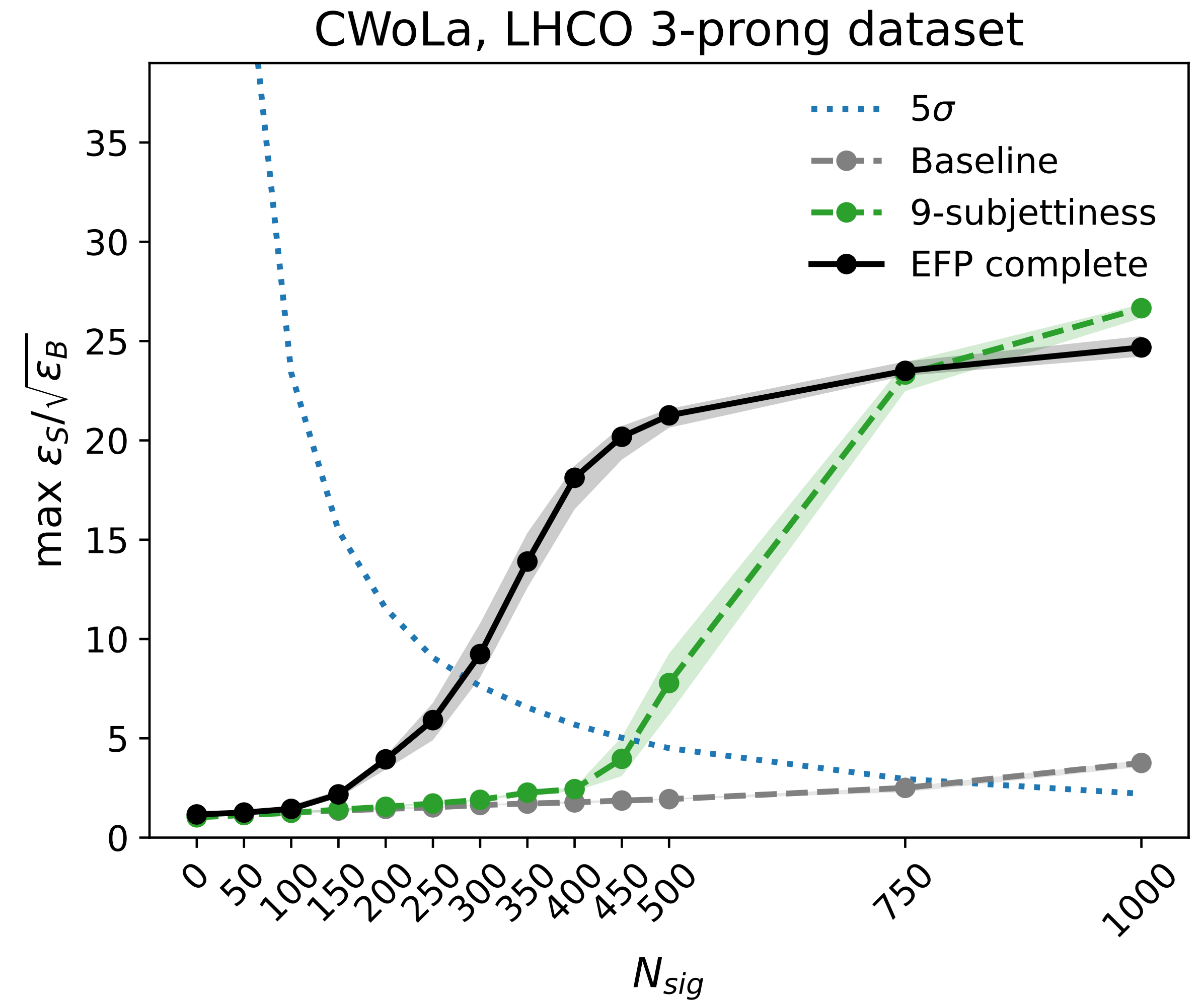
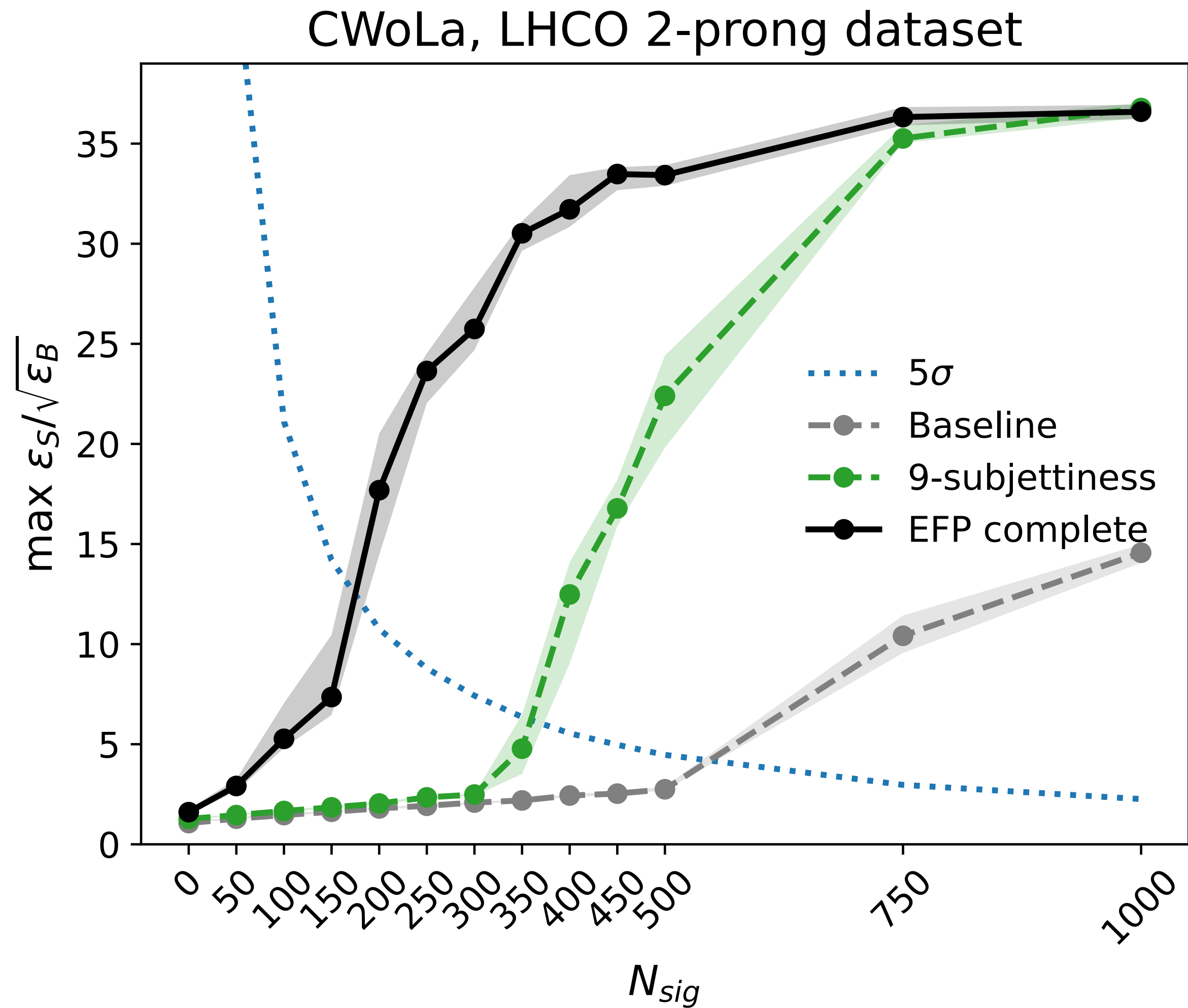


IAD LHCO 3-prong



# CWoLa Hunting

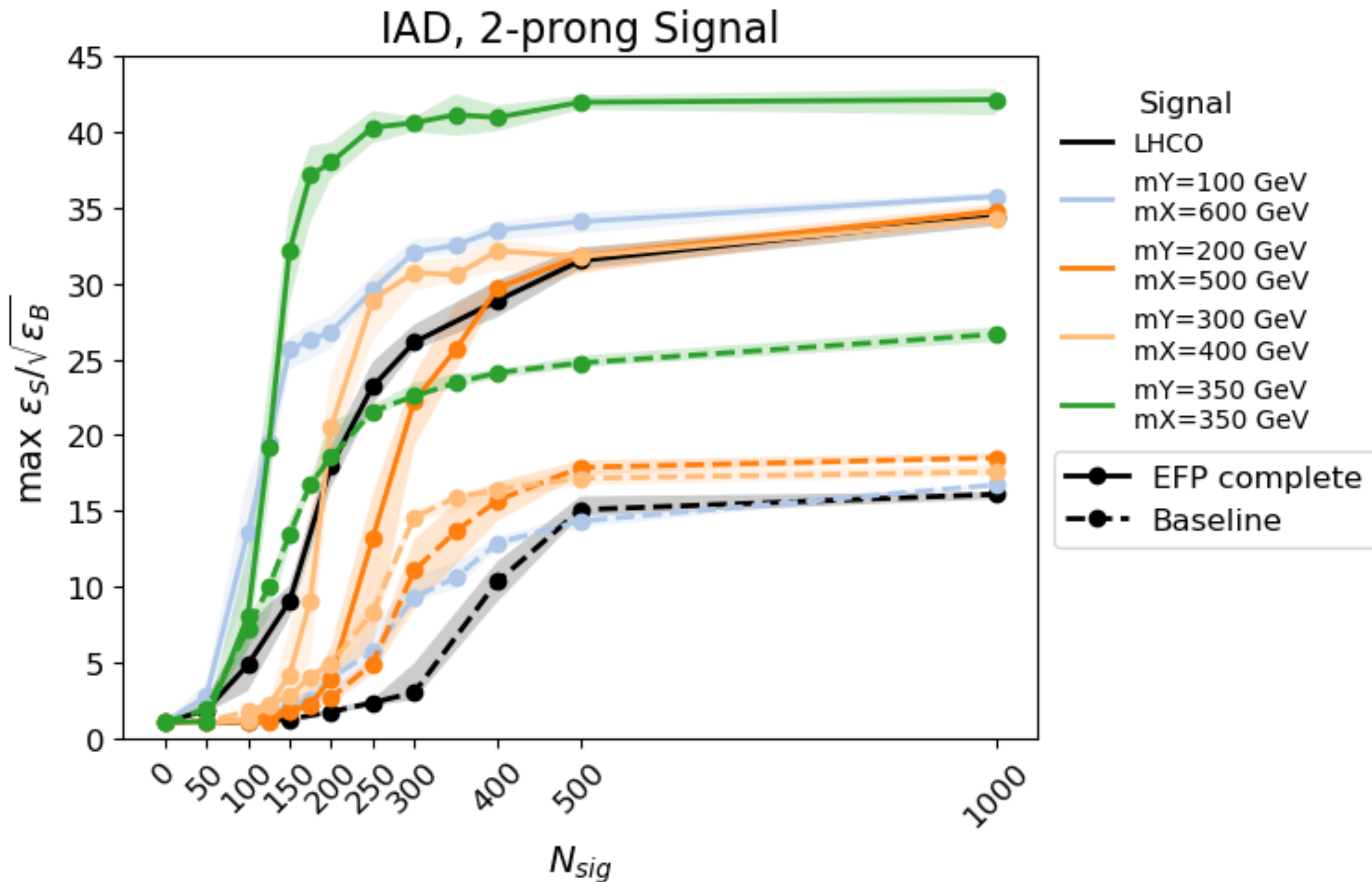
# Signal Injections





# Changing the Masses

## Performance



EFP Complete:

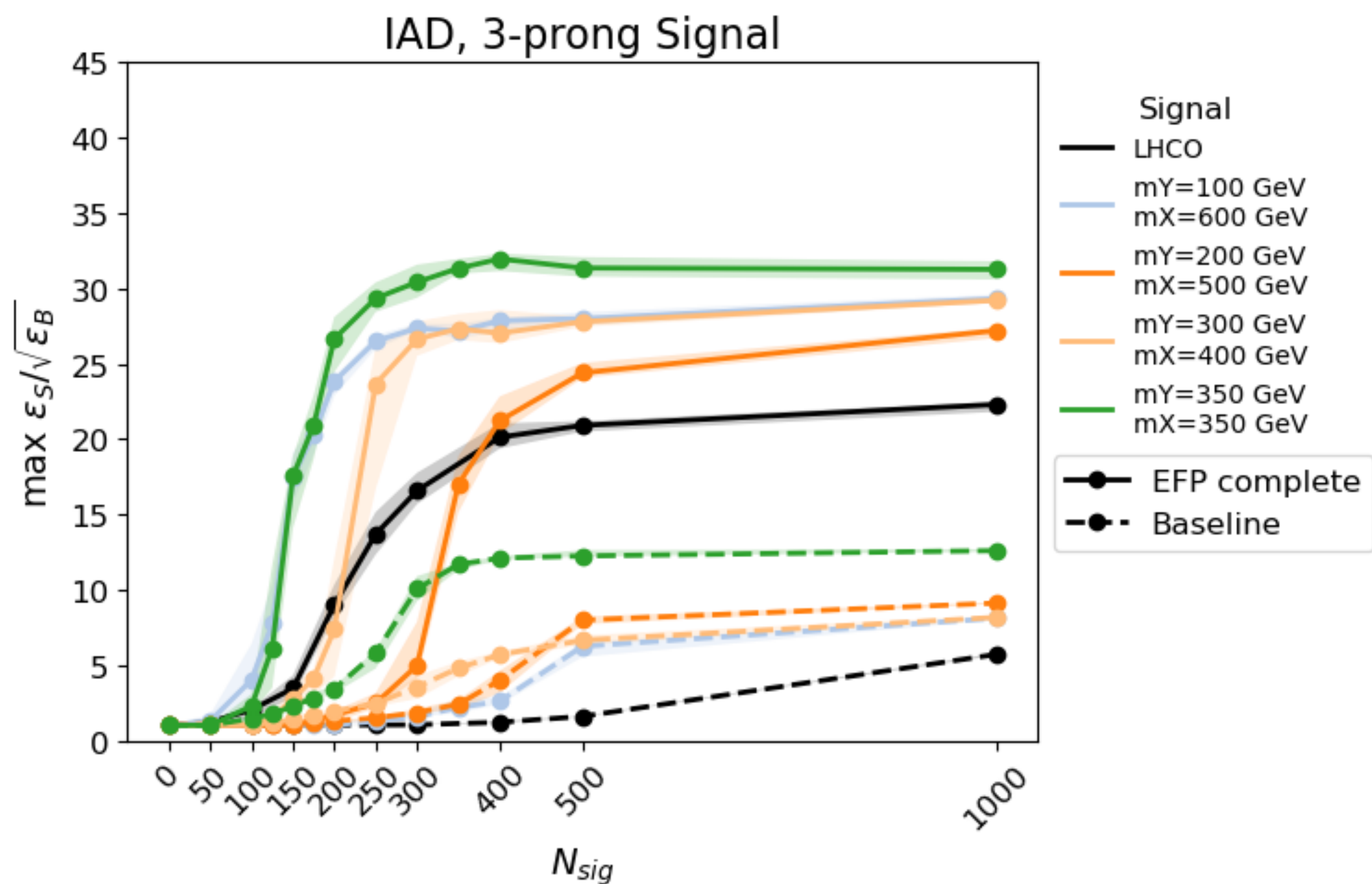
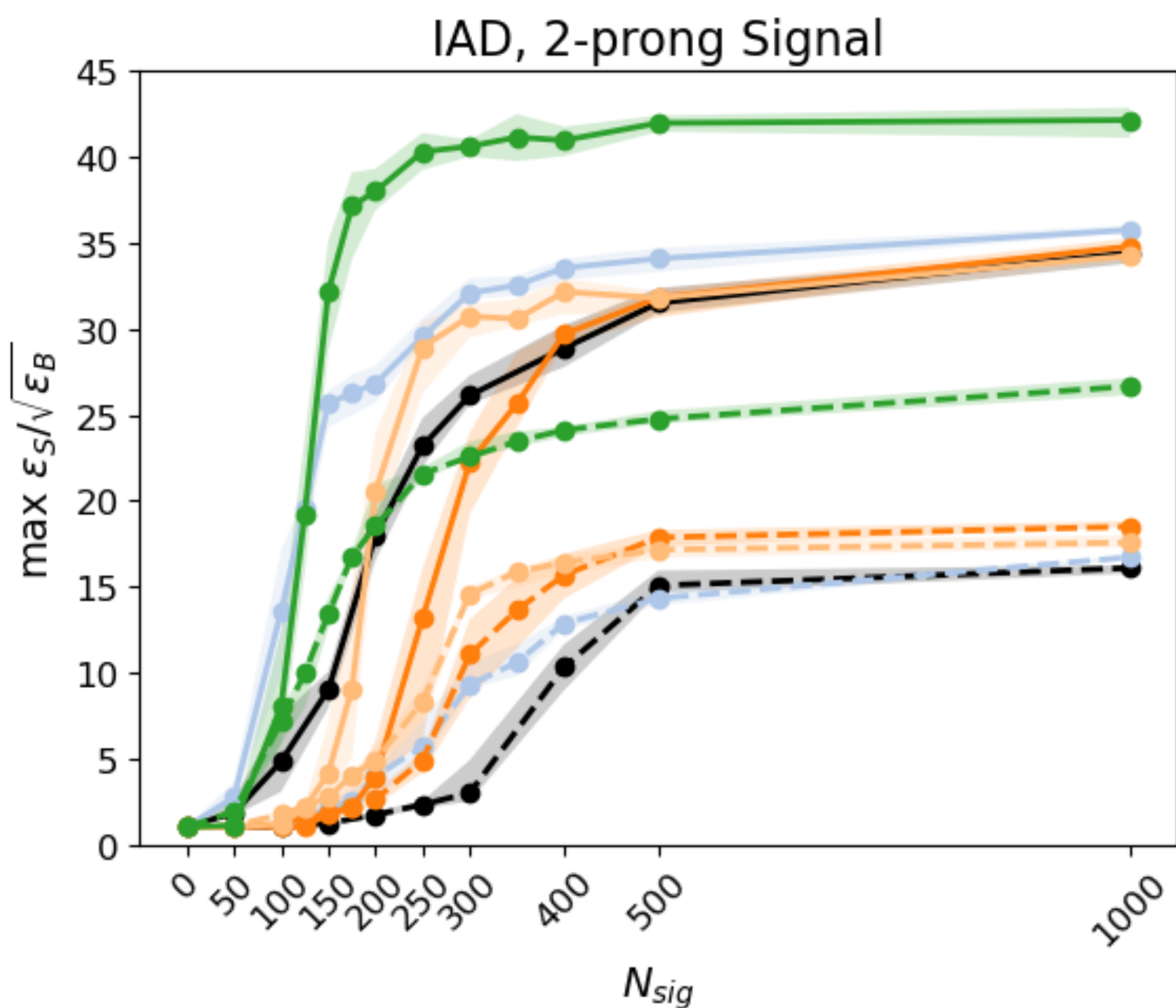
$\{m_{J_1}, \Delta m_J, \text{EFP}_{i,J_1}, \text{EFP}_{i,J_2}\}$  for  
 $i \in [1, 489]$

Baseline:

$\{m_{J_1}, \Delta m_J, \tau_{21,J_1}^{\beta=1}, \tau_{21,J_2}^{\beta=1}\}$

# Changing the Masses

## Performance

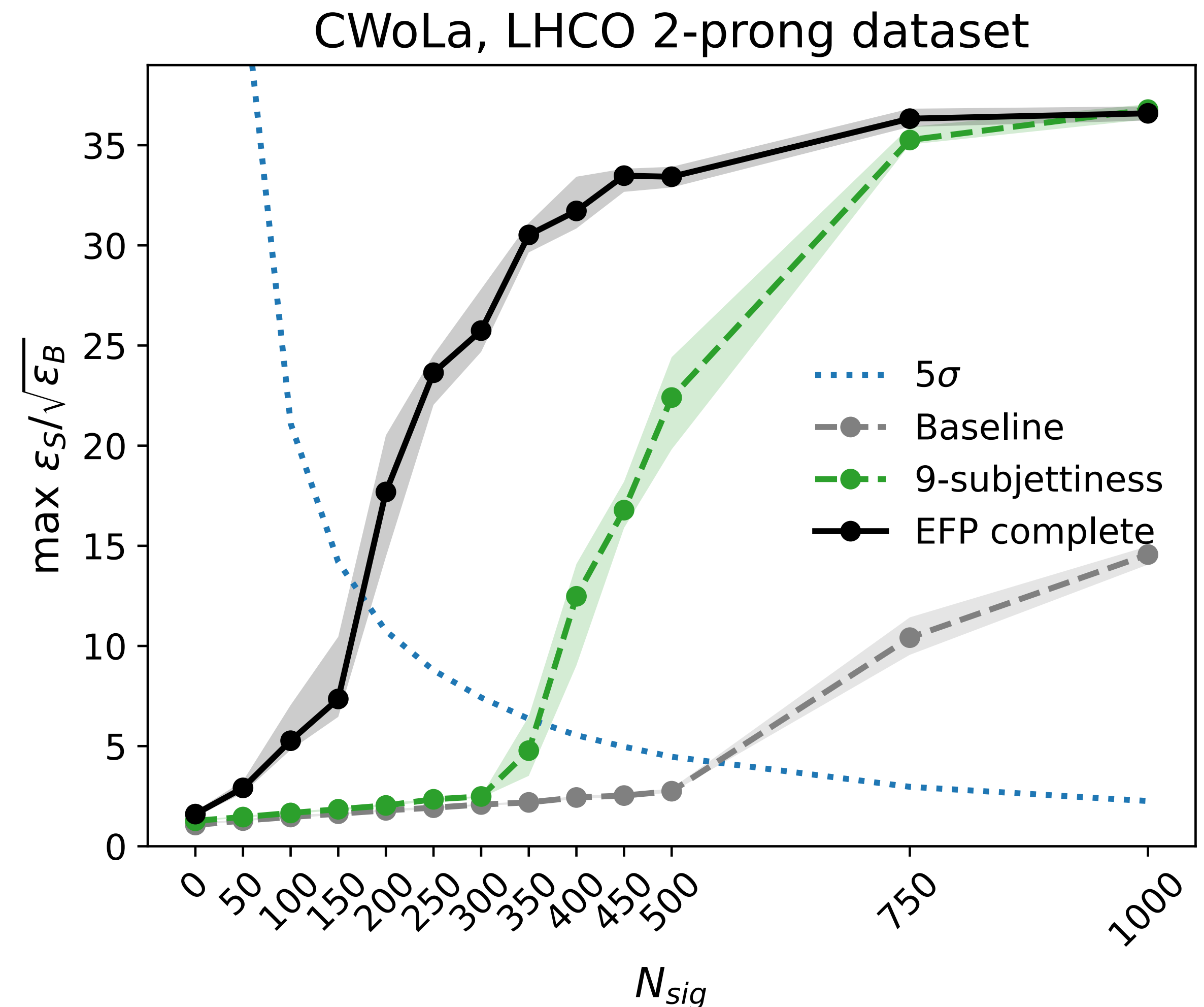


# Conclusion

- EFPs are a powerful tool for anomaly detection
- **EFPs outperform N-Subjettiness** feature sets for low signal injections
- **Improvement of sensitivity**
- EFPs perform well for **different masses**

## Outlook:

- Investigate **more complex signals** (e.g. 2+4-prong, 5+5-prong)

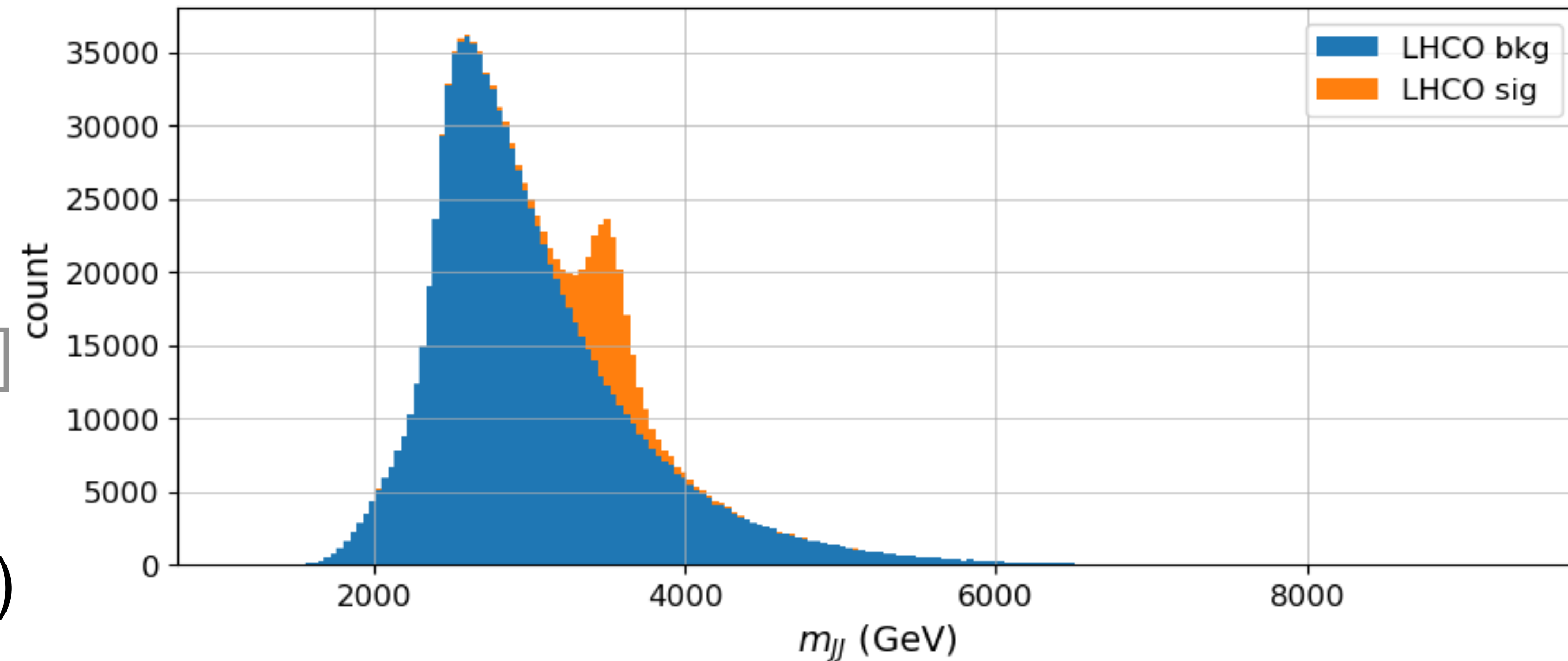


# Backup

# LHC Olympics 2020 R&D Dataset

## Dijets

- 1.1 million simulated dijet events [2]
  - 1 million QCD **background**
  - 100k **signal**
- 610k extra QCD **background** [3]
- The **signal** consists of resonant production of a new  $Z'$  (3.5 TeV)

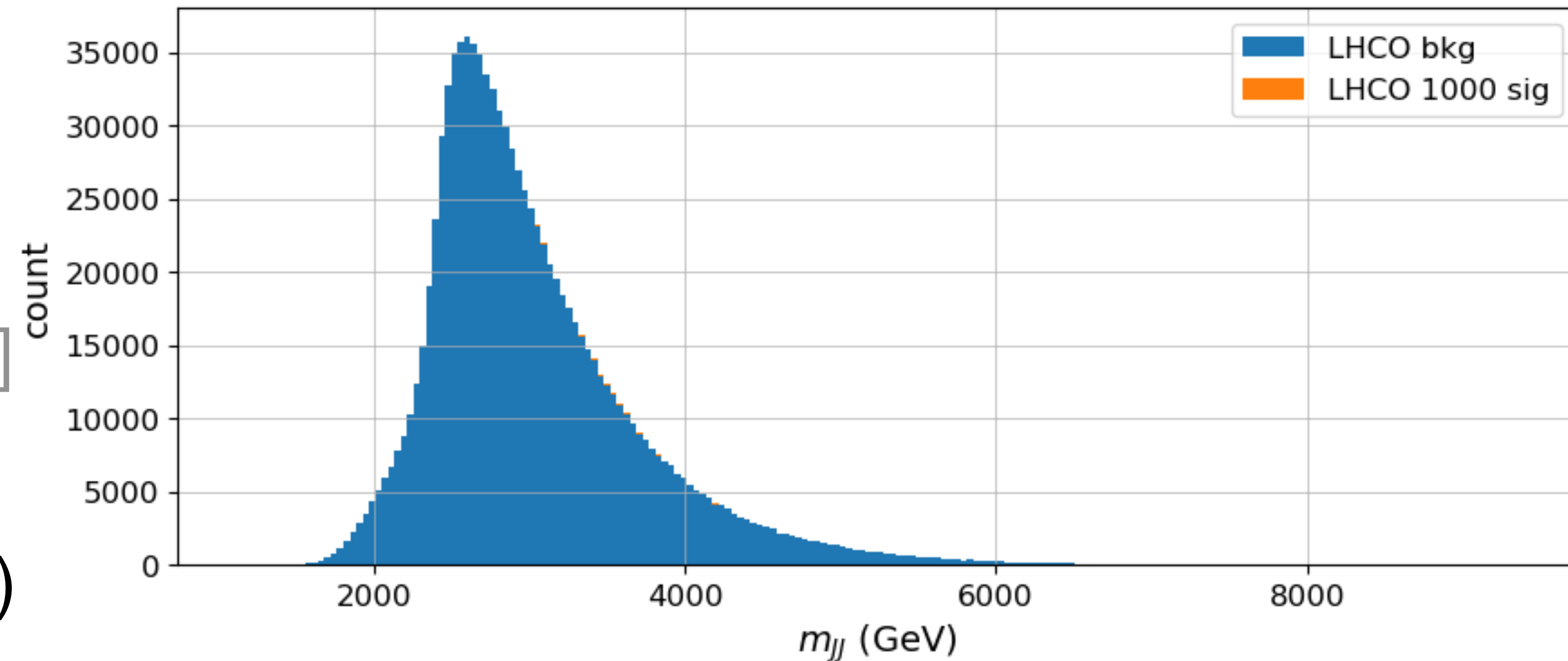




# LHC Olympics 2020 R&D Dataset

## Dijets

- 1.1 million simulated dijet events [2]
  - 1 million QCD background
  - 100k signal
- 610k extra QCD background [3]
- The signal consists of resonant production of a new  $Z'$  (3.5 TeV)



# LHC Olympics 2020 R&D Dataset

## Dijets

### LHC Olympics 2020 R&D datasets

2-prong signal  
100 000 events

$$Z' \rightarrow XY$$

$$X \rightarrow qq \text{ \& } Y \rightarrow qq$$

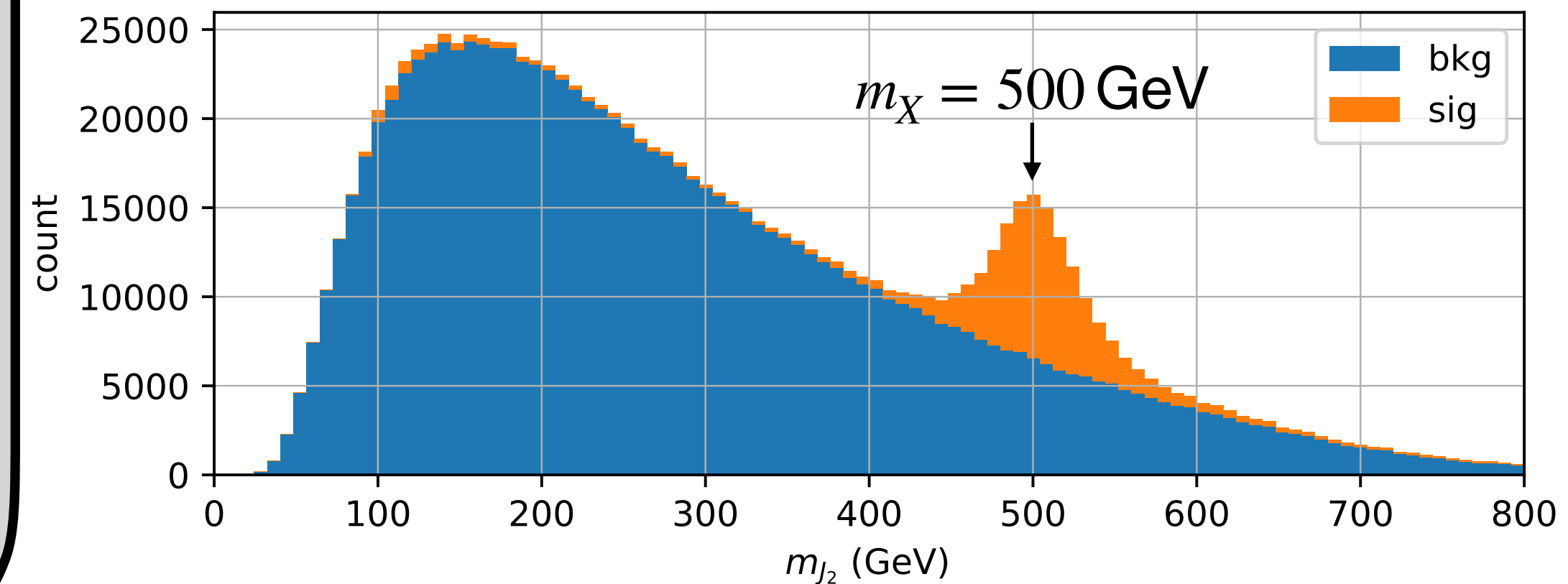
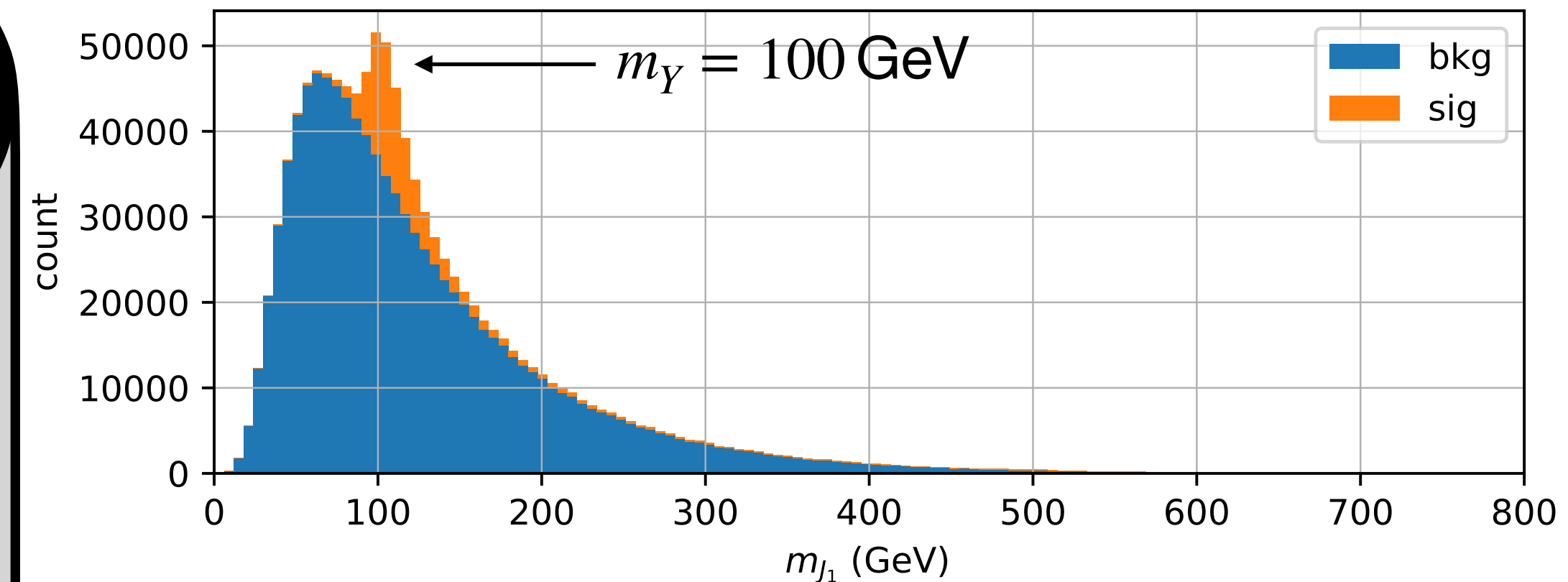
3-prong signal  
100 000 events

$$Z' \rightarrow XY$$

$$X \rightarrow qqq \text{ \& } Y \rightarrow qqq$$

Background events  
1 000 000 QCD dijets

+ 610 000 extra QCD dijet events with  $m_J \in \text{SR}$





# The CMS Signals

## Motivation

- We want to test if EFPs are a natural choice for weakly supervised anomaly detection
- Therefore, we are going beyond the LHCO R&D signals
- The CMS Signals are ([arXiv:2412.03747v1](https://arxiv.org/abs/2412.03747v1)):
  - $X \rightarrow YY' \rightarrow 4q$  (2+2-prong)
  - $W_{kk} \rightarrow WR \rightarrow 3W$  (2+4-prong)
  - $Z' \rightarrow T'T' \rightarrow tZtZ$  (5+5-prong)
  - $Y \rightarrow HH \rightarrow 4t$  (6+6-prong)

# Setup

- Similar to the setup of [7]
- GBDT => [HistGradientBoostingClassifier](#)
  - Limit the maximum number of leaves to 31
  - Set the maximum number of iterations to 200
  - Early stopping with a patience of 10
  - Validation fraction of 0.5
- Ensemble over N=50 individual GBDTs

