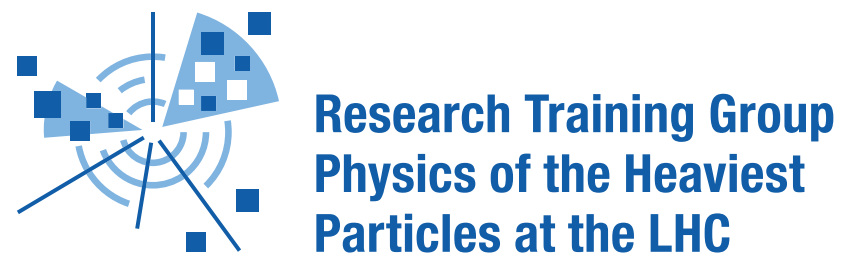


Analysis optimisation

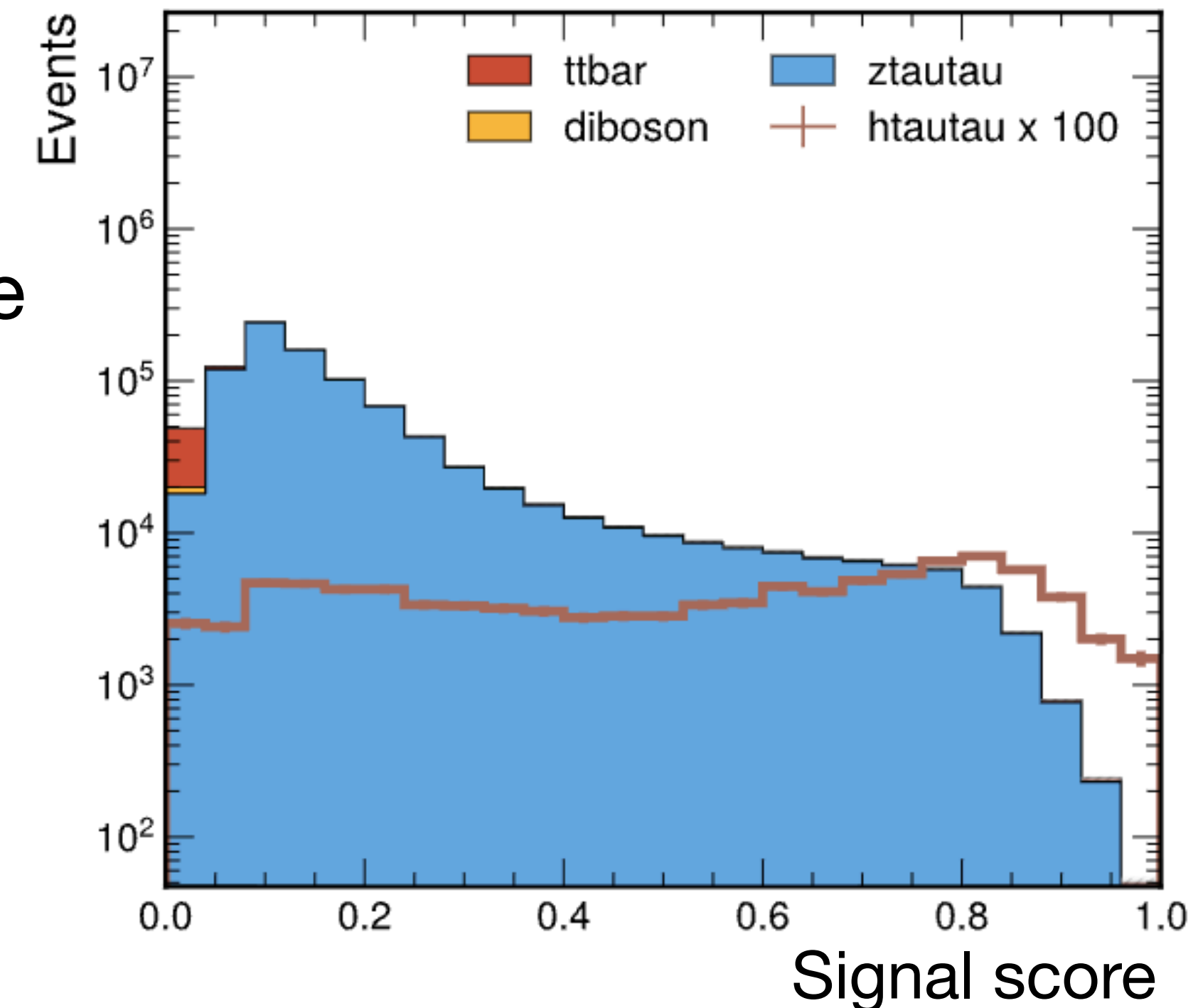
Johannes Erdmann, Nitish Kumar Kasaraguppe
III. Physikalisches Institut A, RWTH Aachen University

BND Graduate School
11 September, 2026



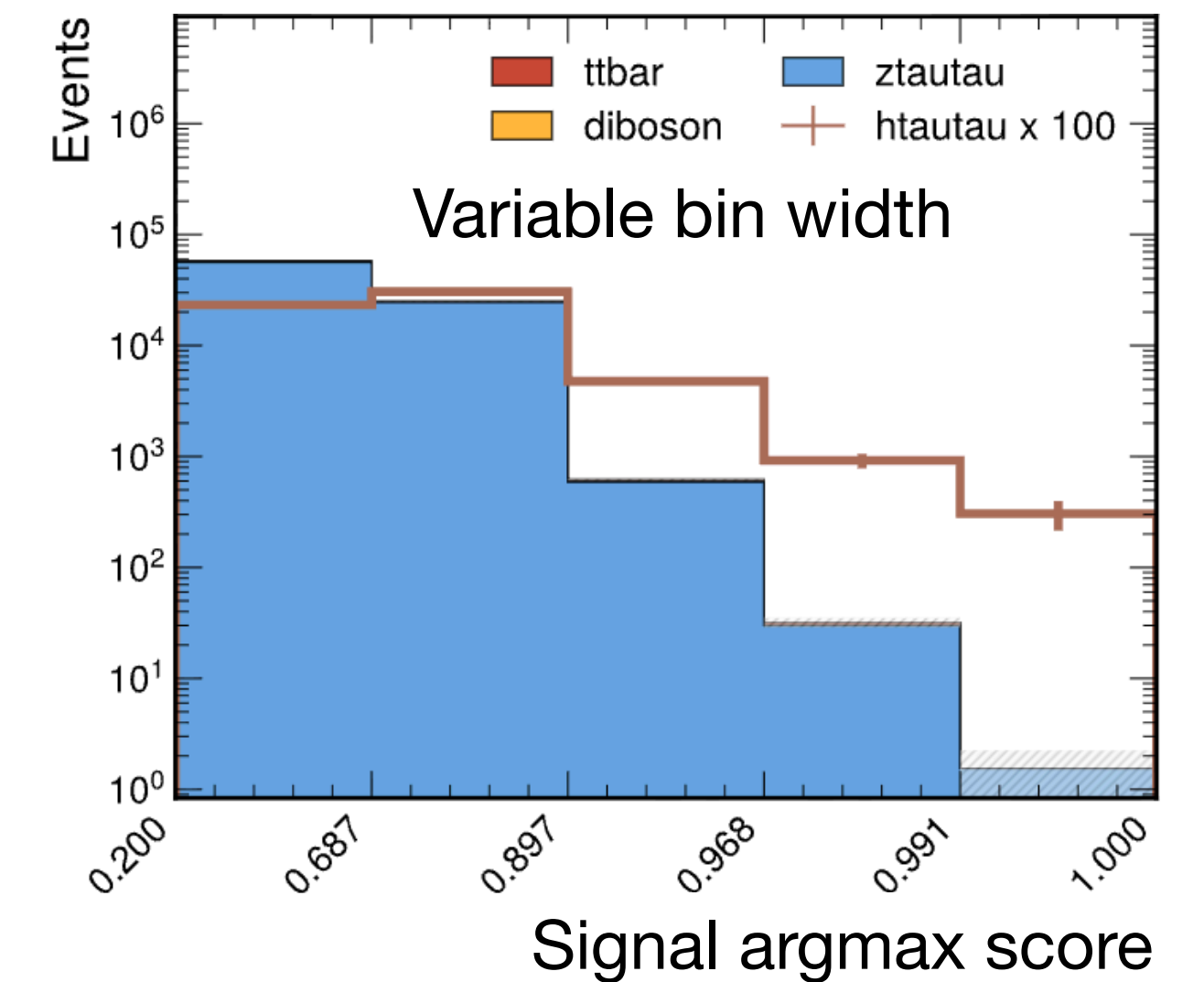
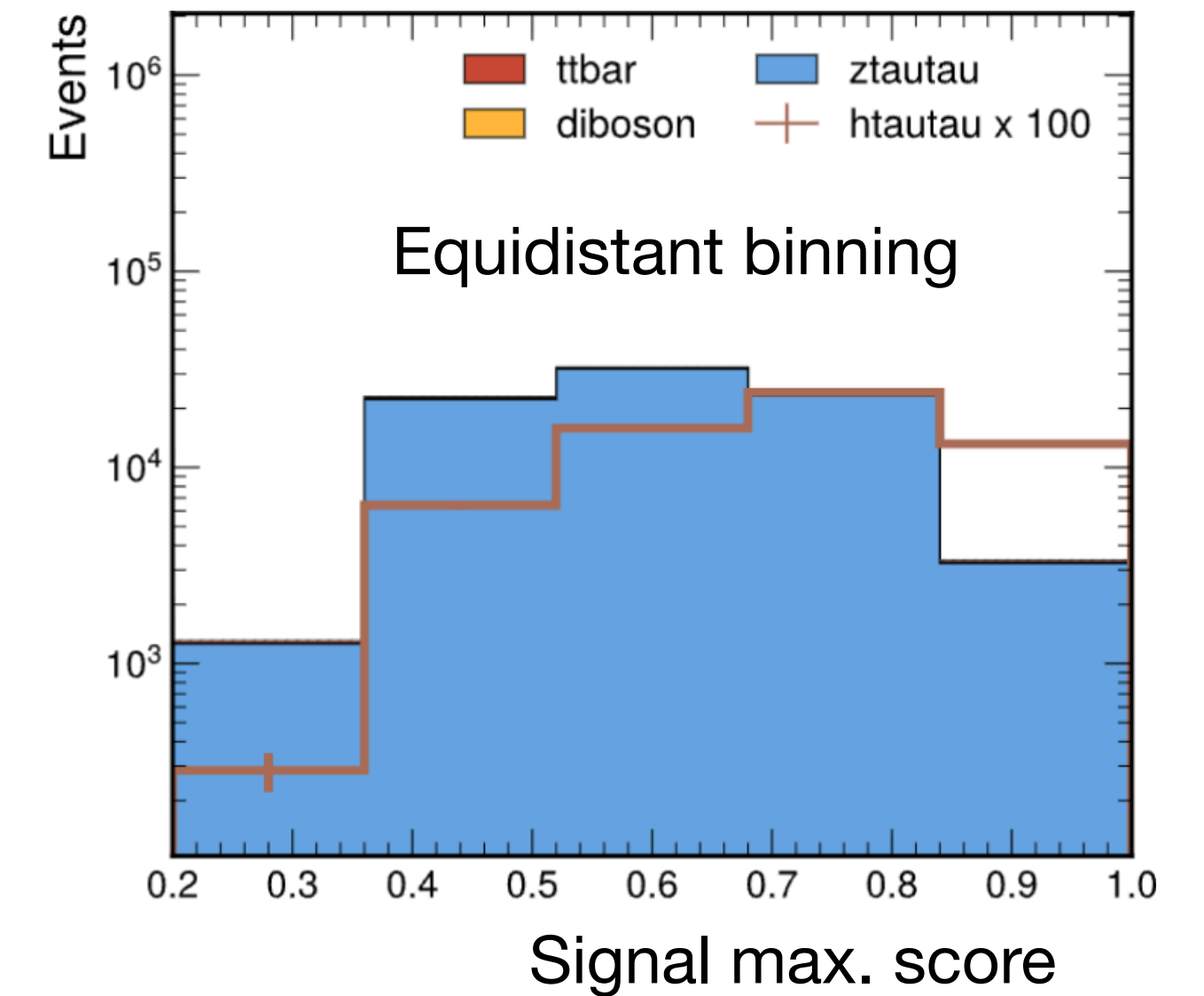
Overview

- Typical HEP analysis workflow deals with many collision events
→ Only a small subset contributes significantly to the measurement sensitivity
- Events are divided into analysis categories or bins to improve the sensitivity
- Modern analysis rely on classifiers to distinguish signal from background
→ Categories or bins could be based on the classifier scores
→ The choice of categories can significantly impact the expected sensitivity
- We study these effects using FAIR Universe $H \rightarrow \tau\tau$ dataset:
[Zenodo](#), [Paper](#)



Binning optimisation

- Categories or bins is often chosen manually
→ Popular choice is to apply *argmax* projections of multi-class scores and bin equidistantly in 1D
- You can further optimise using variable bin width on *argmax* projections
- Common metric used for binning optimization is asymptotic significance: $Z_A = \sqrt{2 \left[(S + B) \ln \left(1 + \frac{S}{B} \right) - S \right]}$ (G. Cowan et al, [1007.1727](#))



Differentiable binning optimisation

- In multi-dimensional binning problems, more flexible solutions can be obtained using **Gaussian Mixture Model (GMM)**

- Define N **Gaussians** in d -dimensional space, one per bin k :

$$\ell_k(x) = \mathcal{N}(x \mid \boldsymbol{\mu}_k, \Sigma_k).$$

- The binning is performed such that an event is assigned to the component with the largest score $s_k(x)$ (π_k is the mixture weight)

$$s_k(x) = \log \ell_k(x) + \log \pi_k.$$

- **Optimisation:** the key is differentiable approximation

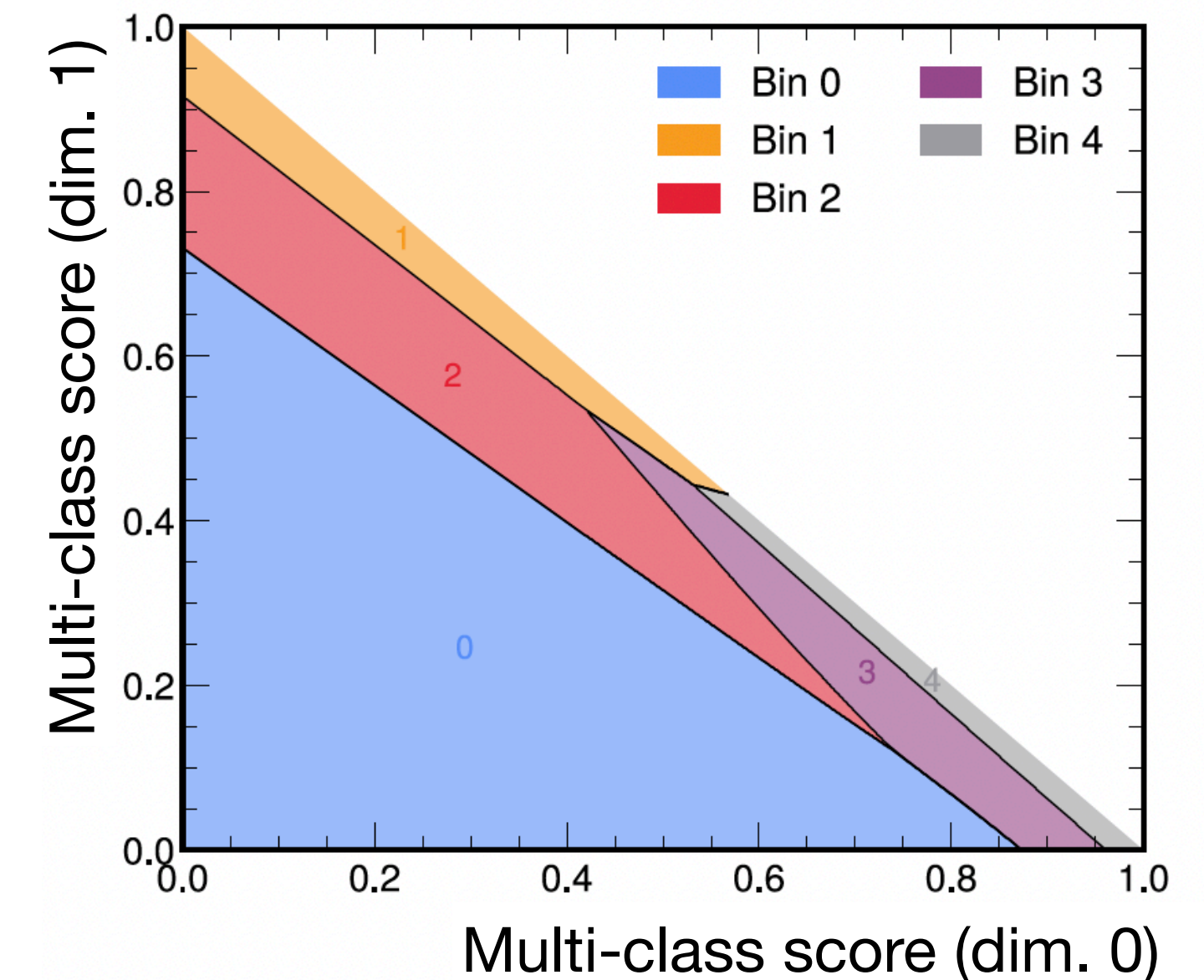
- Obtained by replacing discrete assignment of event with temperature scaled softmax of $s_k(x)$

$$\gamma_k(x) = \text{softmax}_k \left(\frac{s_k(x)}{T} \right)$$

→ Each bin carries a fraction of each event

- Temperature T controls level of approximation: $T \rightarrow 0$ gives hard assignment

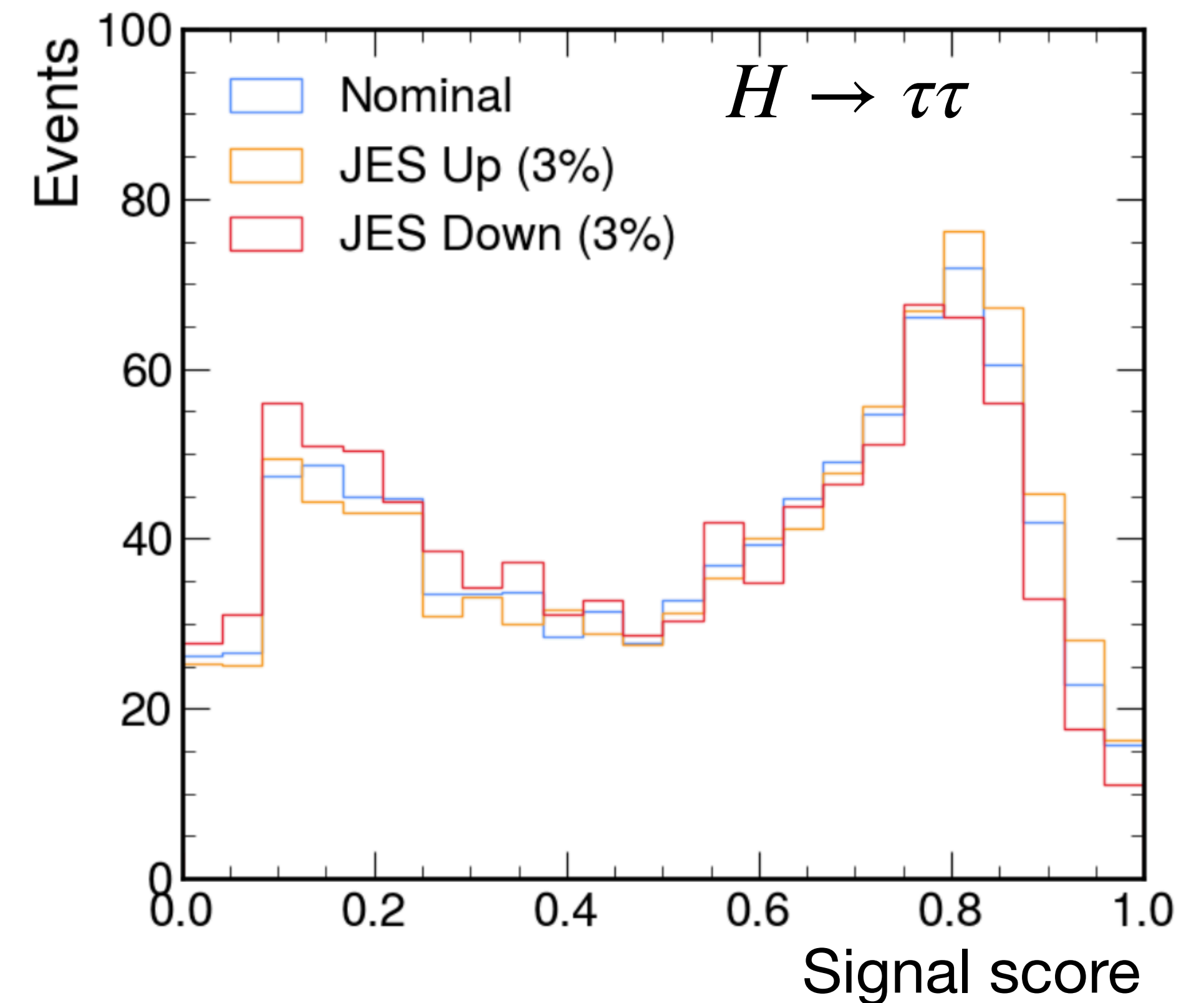
- **Inference:** Event is assigned to the component with the largest score $s_k(x)$



Gradient-based c**A**tegorization **O**ptimizer (GATO)
[Eur. Phys. J. C 86, 1014 \(2026\)](#), [PyPI](#), [GitHub](#)

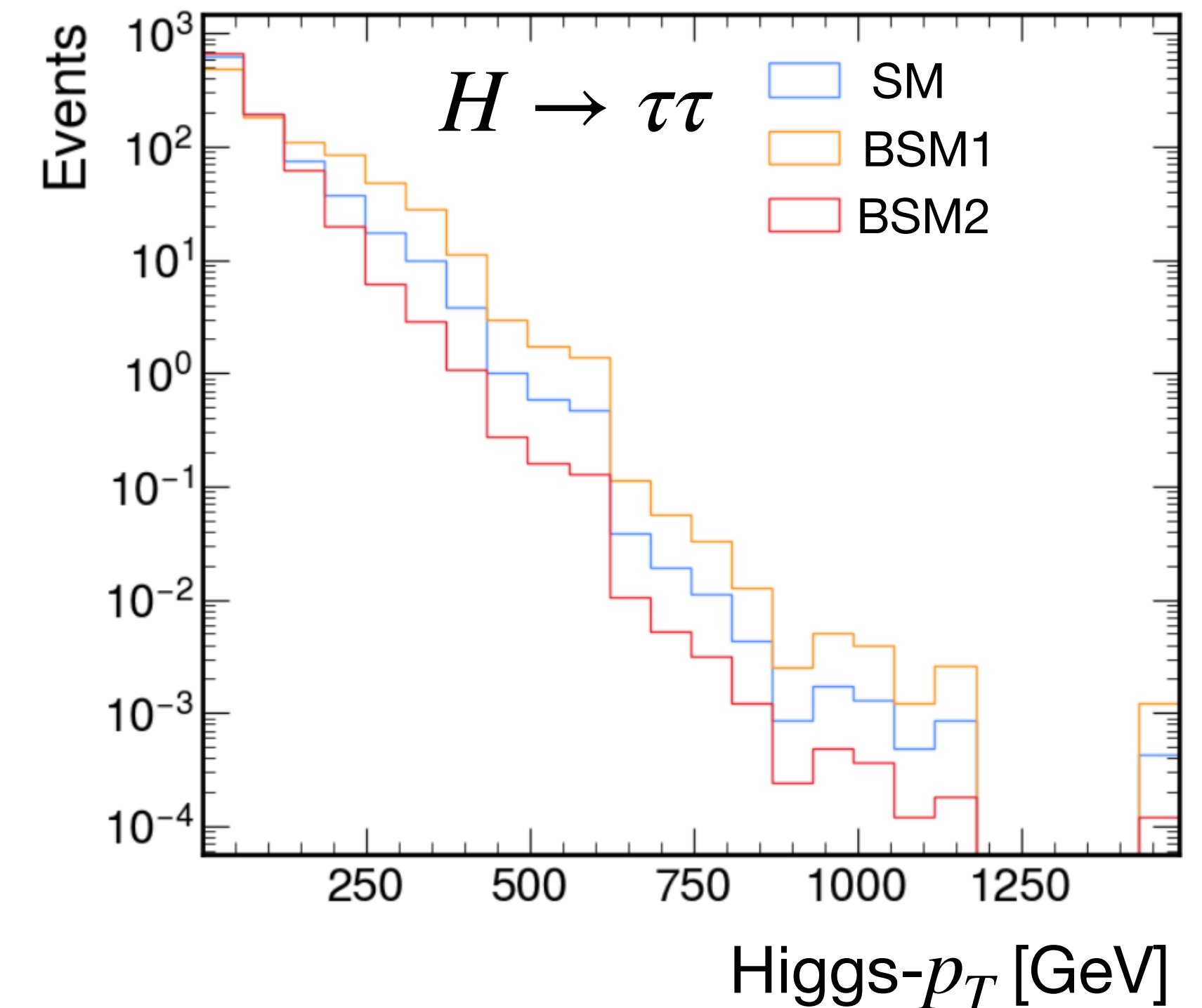
Project I

- Jet Energy Scale (JES) variations shift jet momenta and can propagate to variables used for training the classifier
→ This can change the classifier score and bin assignment
- Bins optimised only for nominal simulation may lose sensitivity under systematic variations
- **Objective**: Optimise a single bin configuration that maintains high signal significance under **nominal**, **JES-up**, **JES-down variations**



Project 2

- Different signal hypotheses can produce different Higgs- p_T spectrum
- Bins optimised for one signal hypotheses can be suboptimal for another
- **Objective**: Optimise a single bin configuration that has optimal significance **for all the signal hypotheses**



Project 3

- A useful analysis needs not only signal-enriched bins, but also regions that help constrain important backgrounds
- In $H \rightarrow \tau\tau$, $t\bar{t}$ is a relevant background and can populate a distinct region of multi-class score space
- **Objective**: Simultaneously optimise **signal regions** for $H \rightarrow \tau\tau$ and a **control region enriched in $t\bar{t}$**

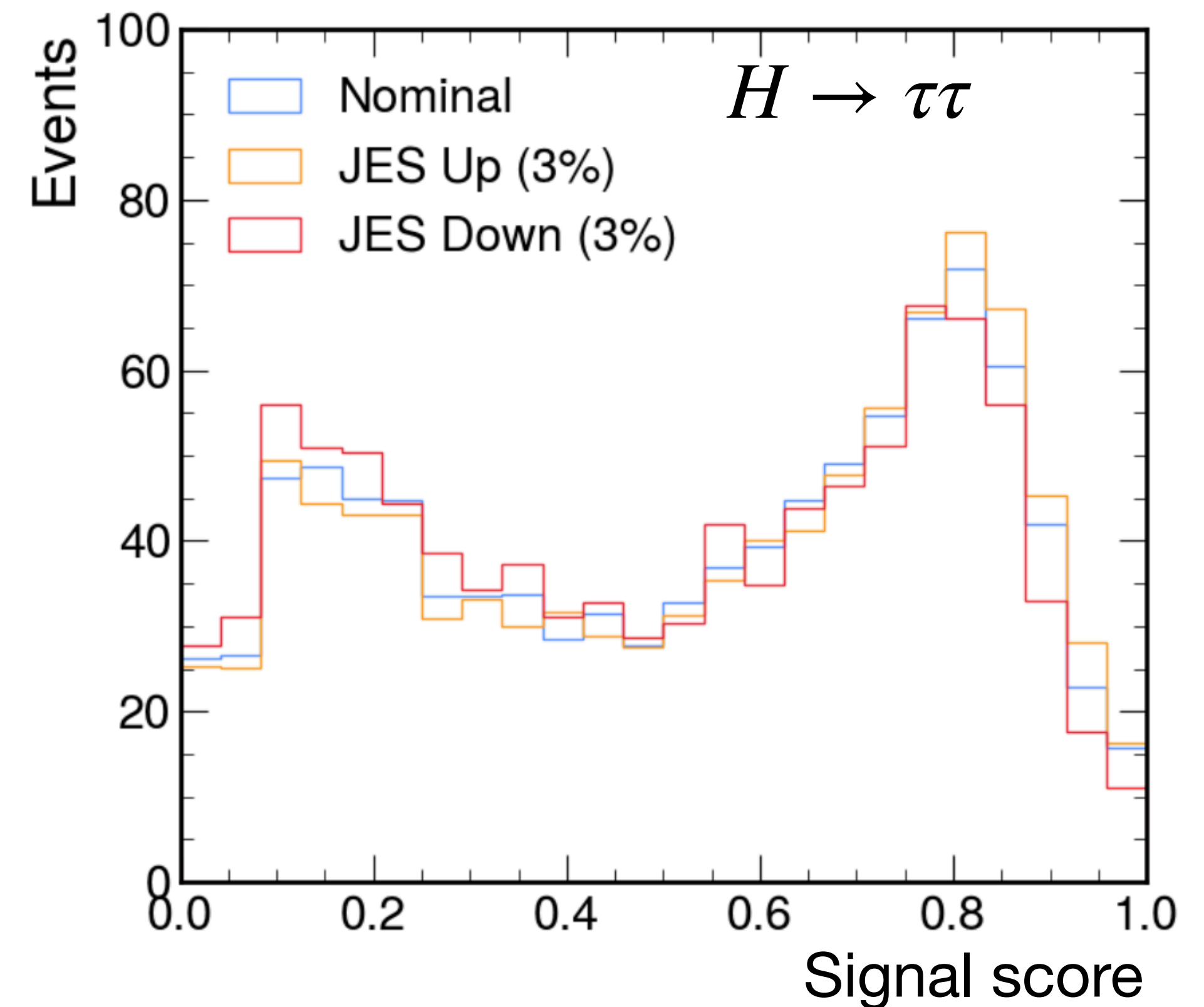
Additional details

- The dataset you get is only a part of the dataset of the full available from FAIR Universe $H \rightarrow \tau\tau$ dataset: [Zenodo](#), [Paper](#)
- A multi-classifier is already trained and the scores are also available along with the dataset
 - Details of training are available in Appendix B of [Eur. Phys. J. C 86, 1014 \(2026\)](#)
 - Details on data and example script is available at: https://github.com/nitishkumarkv/analysis_optimisation_BND_school
- For the projects, using gato-hep is not necessary; **you are free to use any method or approach you prefer**

Additional information

Project I

- Jet Energy Scale (JES) variations shift jet momenta and can propagate to variables used for training the classifier
→ Here the JES is varied up and down by 3%
- The nominal dataset is available as:
gato_like_scores_nominal.pkl
- The up and down dataset is available as:
gato_like_scores_jes_{up/down}.pkl
- A simple starting point for optimisation would be to optimise for geometric mean of significances for the nominal and up/down systematics



Project 2

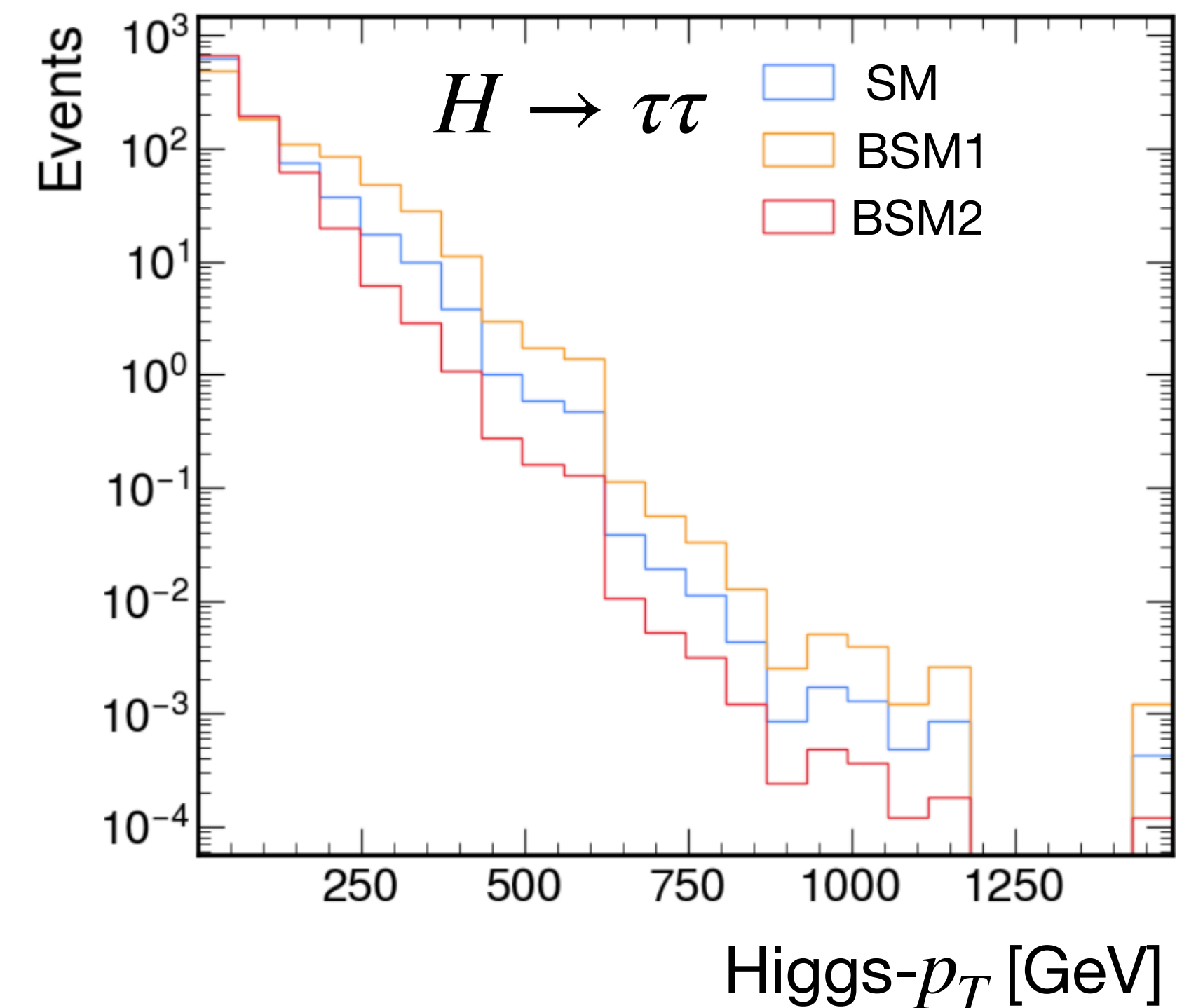
- Different signal hypotheses are produced by varying the weights of signal Higgs- p_T using a function

$$w_{hard}(p_T) = 1 + 0.6 \tanh\left(\frac{p_T - 180}{80}\right)$$

$$w_{soft}(p_T) = 1 - 0.6 \tanh\left(\frac{p_T - 180}{80}\right)$$

→ Each weights are normalized to have same sum of weights as nominal samples

- Use the dataset: *gato_like_scores_with_BSM_weights.pkl*
→ The data frame for $H \rightarrow \tau\tau$ already includes three weight columns: '*weight*', '*weight_hard*', '*weight_soft*'
- A good starting point is to optimise for geometric mean of significances for different signal hypotheses



Project 3

- A useful analysis needs not only signal-enriched bins, but also regions that help constrain important backgrounds
- **Objective**: Simultaneously optimise signal regions for $H \rightarrow \tau\tau$ and a control region enriched in $t\bar{t}$
- Use the nominal dataset: *gato_like_scores_nominal.pkl*
- A good starting point is to optimise for geometric mean of significances for $H \rightarrow \tau\tau$ and $t\bar{t}$ and assign $t\bar{t}$ enriched bin as the control region