# Learning to bin: differentiable approach for multi-dimensional discriminants in high-energy physics

Learning to bin (Eur. Phys. J. C 86, 1014 (2026))

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

ML4Jets, Vienna, 2026
14 September, 2026

Research Training Group Physics of the Heaviest Particles at the LHC

RWTH AACHEN UNIVERSITY

Gefördert durch:

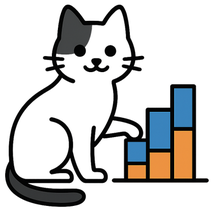
Bundesministerium für Forschung, Technologie und Raumfahrt

Physics Institute III A

RWTH AACHEN UNIVERSITY

# Overview

- Typical HEP analysis workflow deals with many collision events
  → Only a small subset contributes significantly to the measurement sensitivity
- We categorize events to optimize the sensitivity to a signal
  → Popular choice is to assign each event to the class with the largest multi-class score (using '*argmax*') and bin equidistantly in 1D
- Inference-aware methods such as INFERNO, Wunsch et al., and neos optimize learned summary statistics for the final inference task
- We present a method to directly optimize signal significance in multi-dimensional discriminants using **differentiable approach**
  → Learning to bin (Eur. Phys. J. C 86, 1014 (2026))
- Algorithms shared as Python package, lightweight plugin for analyses: gato-hep (**G**radient-based c**AT**egorization **O**ptimizer) available on PyPI, GitHub
  → Built on TensorFlow's automatic differentiation



gato-hep

TensorFlow

# Method: one-dimensional case

- Straightforward to replace non-differentiable cuts with sigmoids
- Multiple sigmoid functions can be combined to calculate weight for each bin

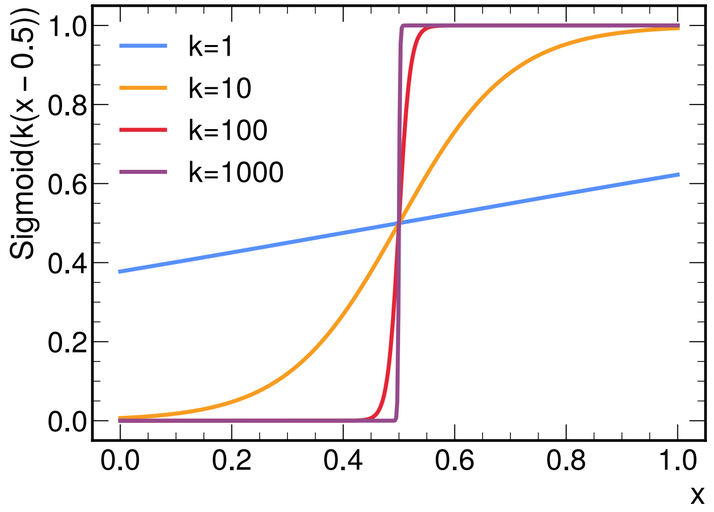
  → Event with observable $x$ and N bins (boundaries $e_{1,\ldots,N-1}$), the category membership weights are defined as:

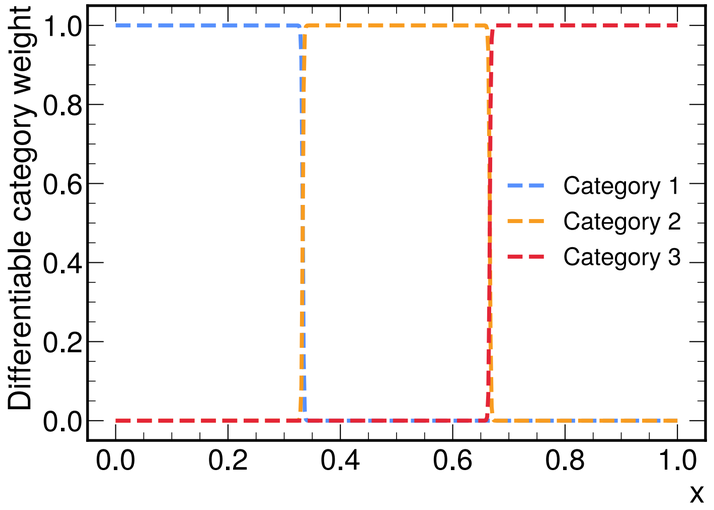
$$w_i(x) = \sigma\big(k(x - e_{i-1})\big) - \sigma\big(k(x - e_i)\big)$$

- The signal significance can be calculated as a function of the boundaries

  → Using asymptotic significance:

$$Z_A = \sqrt{2\left[(S+B)\ln\left(1+\frac{S}{B}\right) - S\right]}$$ (G. Cowan et al, 1007.1727)



# Method: multi-dimensional binning

- In multi-dimensional categorization problems (e.g. multi-classifier output or multiple 1D observables), more flexible solutions can be obtained using a **Gaussian Mixture Model (GMM)**
  → Cuts are not restricted to be rectangular

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

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

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

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

  → Learnable mean: $\mu_k$
  → Learnable covariance: $\Sigma_k$
  → Learnable mixture weight: $\pi_k$

# Optimization and inference

- **Optimization**: the key is differentiable approximation
  - Obtained by replacing hard assignment 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:
    $T \to 0$ gives hard assignment
- **Inference**:
  - Event is assigned to the component with the largest score $s_k(x)$

# Toy dataset

- Toy events are sampled from multivariate normal distribution

$$\mathcal{N}(\mu_i, \Sigma)$$

→ Each process has specific means $\mu_i$ and a common covariance matrix

- One-dimensional case: consider **one signal + combined background**

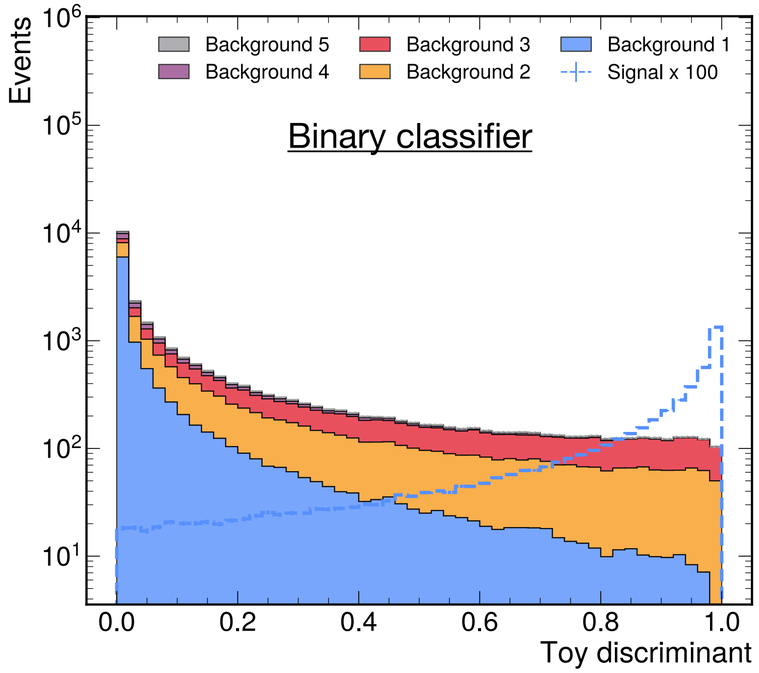
→ Build signal-to-background likelihood ratio and normalize to $[0,1]$

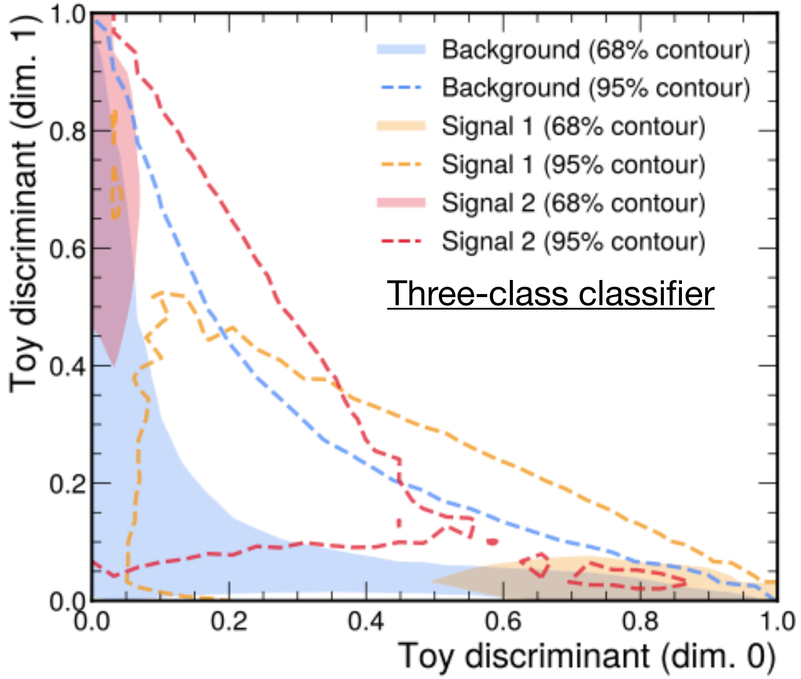
- Three-dimensional case: consider **two signal + combined background**

→ Build softmax-like scores with three outputs summing to 1

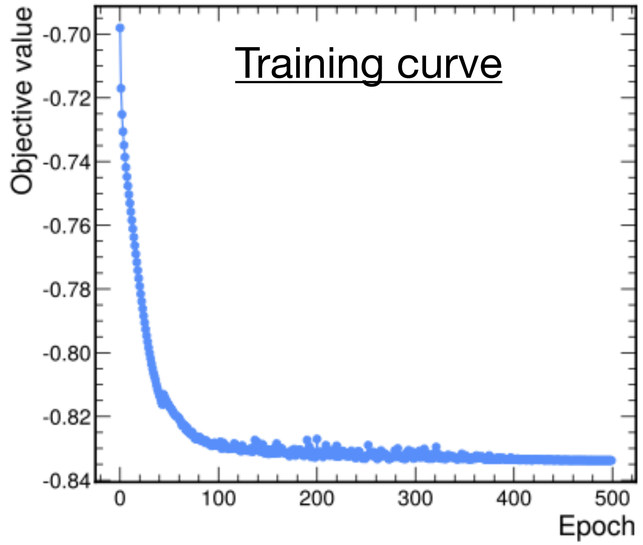
| Process | Scenario 1 $\mu_x^{(1)}$ | $\mu_y^{(1)}$ | $\mu_z^{(1)}$ | Scenario 2 $\mu_x^{(2)}$ | $\mu_y^{(2)}$ | $\mu_z^{(2)}$ | $\sigma$ [pb] |
|---|---|---|---|---|---|---|---|
| Signal 1 | 0.4 | −0.4 | 1.0 | 0.9 | −0.9 | 1.0 | 0.5 |
| Signal 2 | −0.4 | 0.4 | 1.0 | −0.9 | 0.9 | 1.0 | 0.1 |
| Background 1 | 0.0 | 0.0 | −0.5 | 0.2 | 0.2 | 0.5 | 100.0 |
| Background 2 | −0.2 | 0.0 | 0.0 | – | – | – | 80.0 |
| Background 3 | 0.1 | 0.4 | −0.3 | – | – | – | 50.0 |
| Background 4 | −0.1 | 0.6 | −0.2 | – | – | – | 20.0 |
| Background 5 | −0.2 | 0.1 | −0.1 | – | – | – | 10.0 |

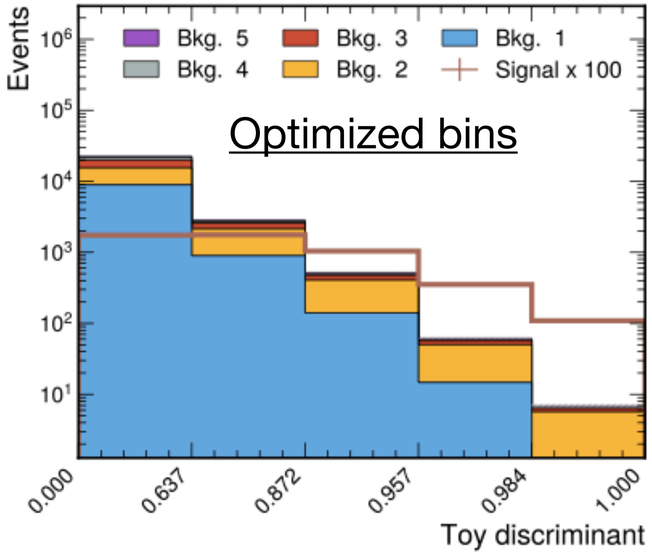
$$\Sigma = \mathrm{diag}(1,1,1) + 0.2\,(\mathbf{1} - \mathrm{diag}(1,1,1))$$

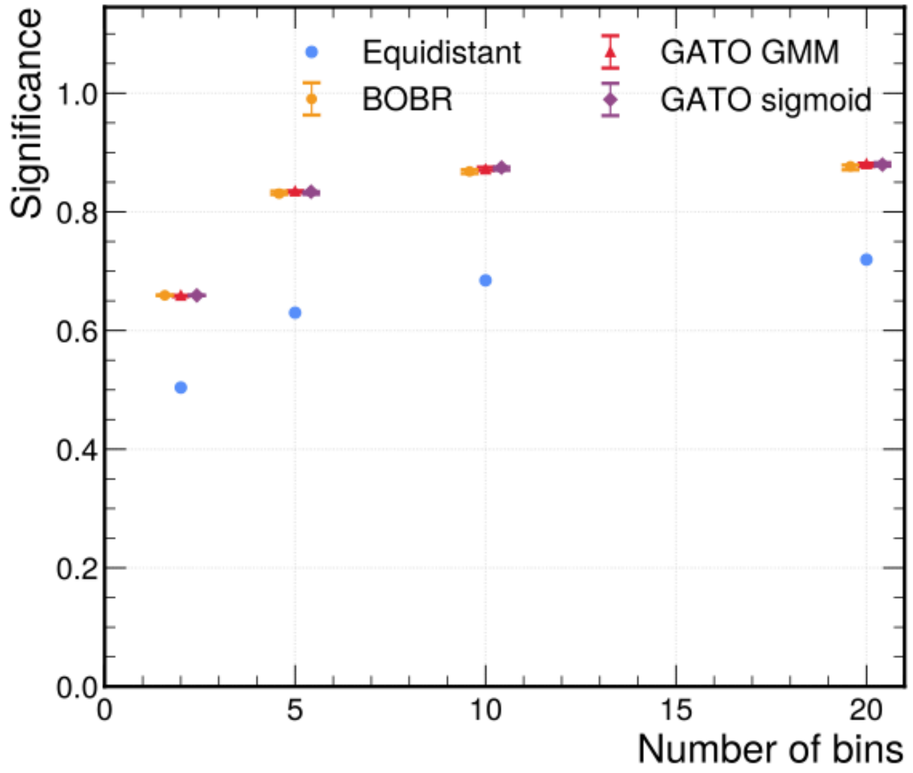


# Results: binary classifier



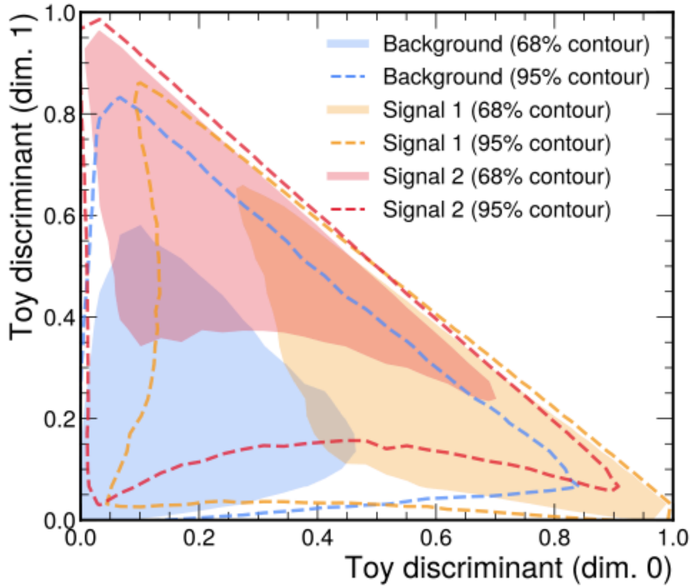



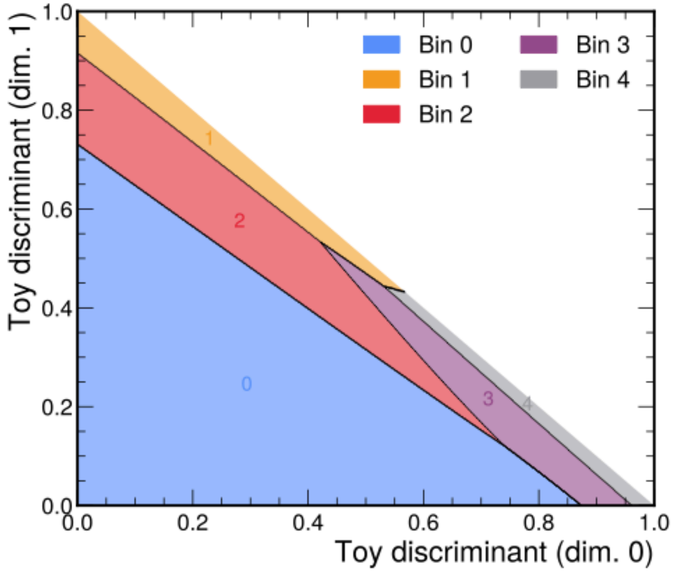



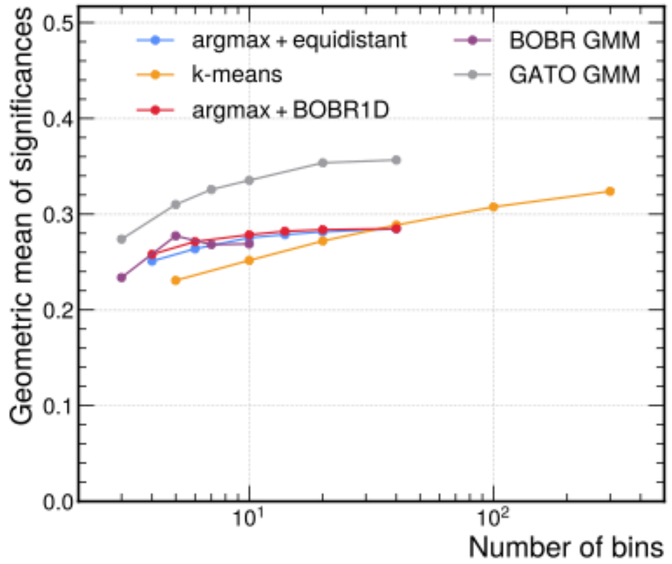

Gradient based method performs significantly better compared to equidistant binning

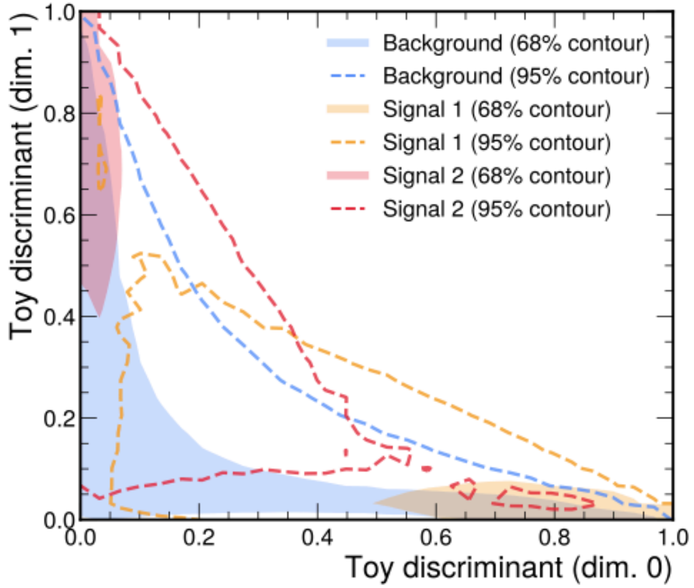
***B**ayesian **O**ptimization of **B**in bounda**R**ies (**BOBR**) also available as a package: [bobr-hep]*

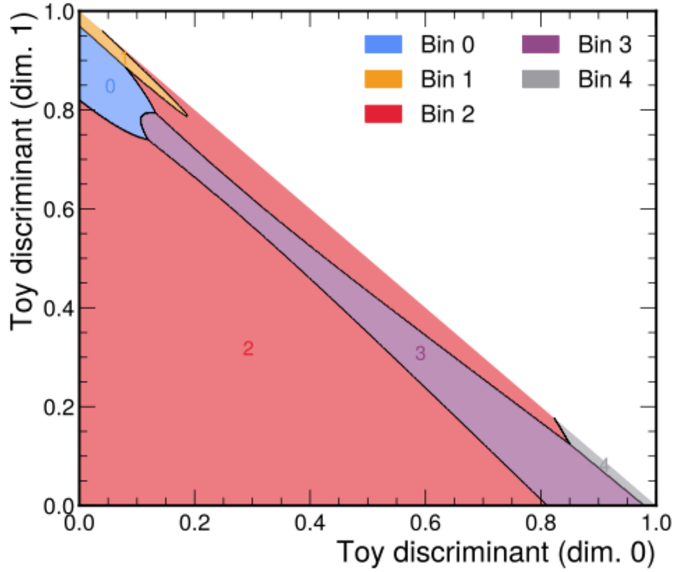
# Results: three dimensional case

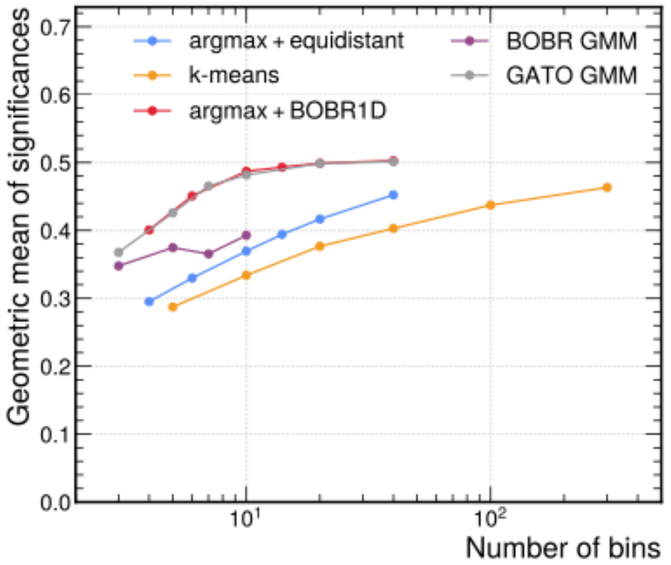



*K-means method: S. Diekmann et al.*

# Penalties during optimization

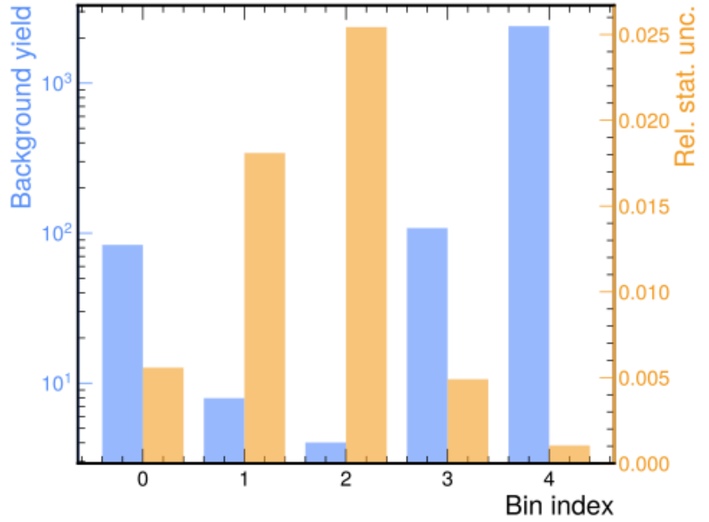
- We use penalties to discourage bins with zero expected background yield or large statistical uncertainty
  - Low background yield penalty:

$$P_{\text{low}}(B) = \sum_{k=1}^{K} \left[\max(0,\, B_{\text{min}} - B_k)\right]^2$$

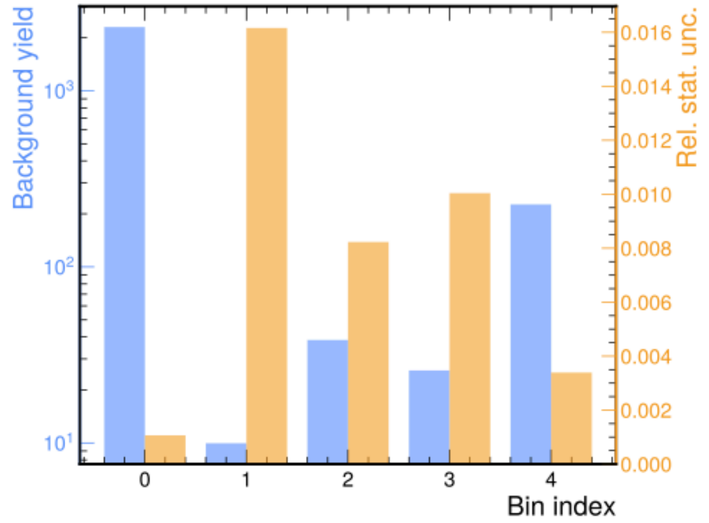
  - Large relative background statistical uncertainty penalty:

$$P_{\text{unc}}(B) = \sum_{k=1}^{K} \left[\max(0,\, \sigma_{\text{rel},k} - r)\right]^2$$

Low background yield penalty of one event

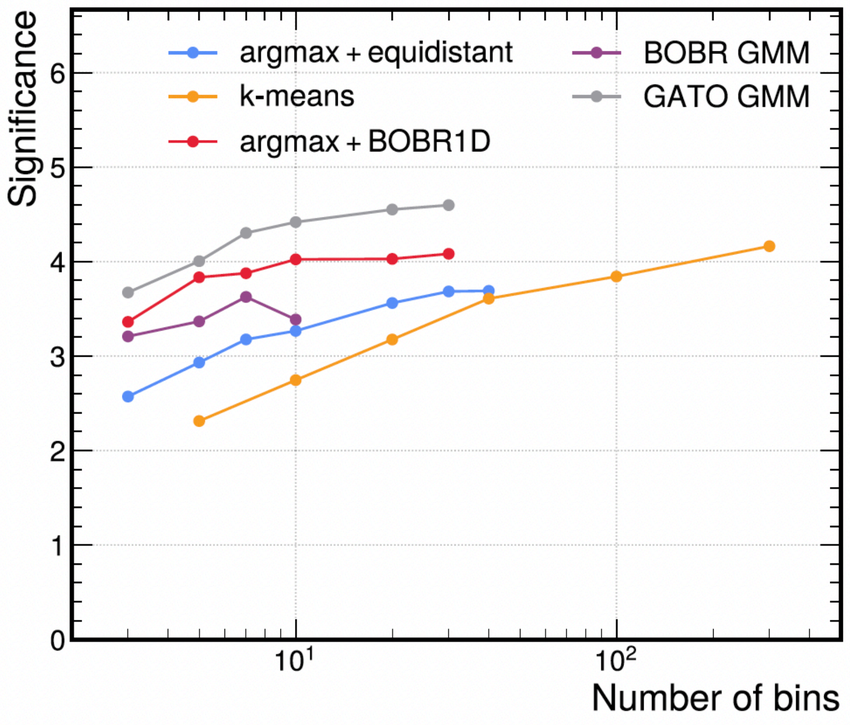



Low background yield penalty of ten events



# Performance on realistic dataset

- We also studied the performance on realistic FAIR universe $H \to \tau\tau$ public dataset
- A multi-classifier is trained to distinguish $H \to \tau\tau$ (signal) from background processes $Z \to \tau\tau$, $t\bar{t}$ and diboson ($VV$)
- Bins are optimized to improve the significance of $H \to \tau\tau$ signal
- GATO models with GMM parameterization achieve the highest significance for all the considered bin counts



# Summary

- Novel approach to differentiable multidimensional binning optimization based on Gaussian Mixture Models: Learning to bin (Eur. Phys. J. C 86, 1014 (2026))

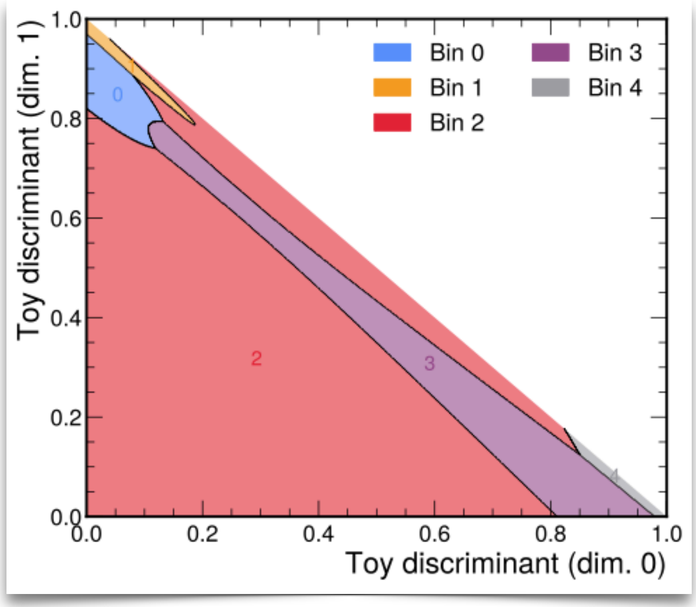
  → **Can also be used as a piece in fully differentiable analysis pipeline**

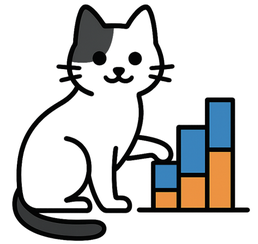
- We presented optimization for signal significance

  → Objective could also be any function of signal and background yields

  → Systematic effects could be included with a differentiable inference-aware objective that accounts for nuisance parameters

- gato-hep shared as Python package, lightweight plugin for analyses



gato-hep

# Backup Slides

# Toy dataset

- Each physics process $i$ (signals and backgrounds) is represented as three-dimensional feature vector
- Toy events are sampled from multivariate normal distribution

$$\mathcal{N}(\mu_i, \Sigma)$$

→ Each process has specific means $\mu_i$ and a common covariance matrix

- The total background likelihood for an events is:

$$p_{\text{bkg.}}(X) = \sum_{i \in \{\text{bkg. } 1,\ldots,\text{bkg. } 5\}} \frac{\sigma_i}{\sum_{j \in \{\text{bkg. } 1,\ldots,\text{bkg. } 5\}} \sigma_j} \mathcal{N}(X \mid \mu_i, \Sigma)$$

- For 1D:

$$L(X) = \frac{p_{\text{sig. } 1}(X)}{p_{\text{bkg.}}(X)}, \quad D(X) = \frac{L(X)}{1 + L(X)} \in [0,1]$$

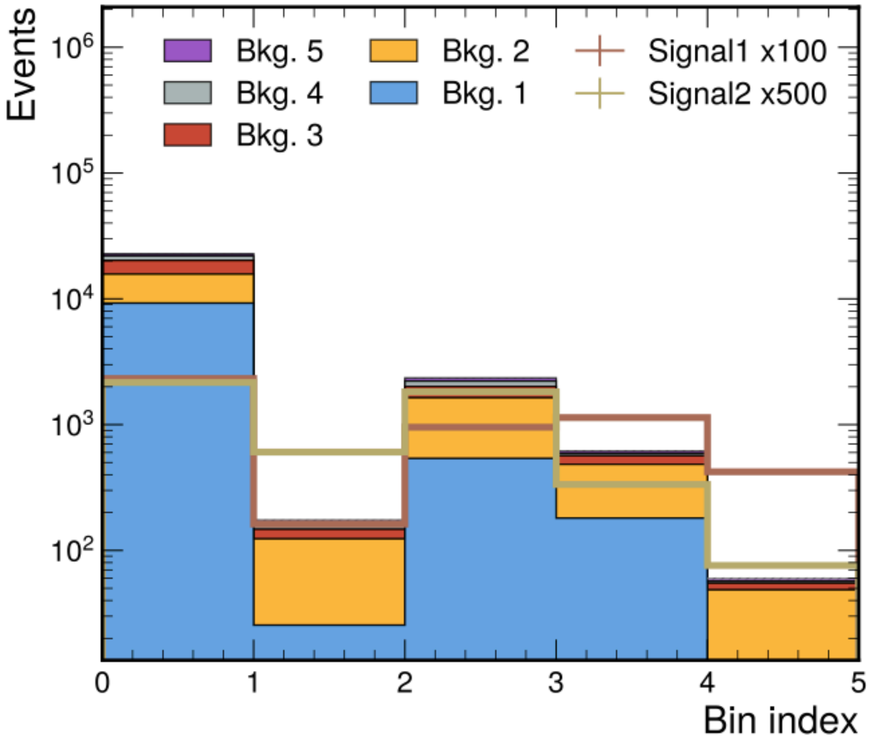
- For 3D:

$$S_k(X) = \frac{p_k(X)}{p_{\text{sig. } 1}(X) + p_{\text{sig. } 2}(X) + p_{\text{bkg.}}(X)},$$

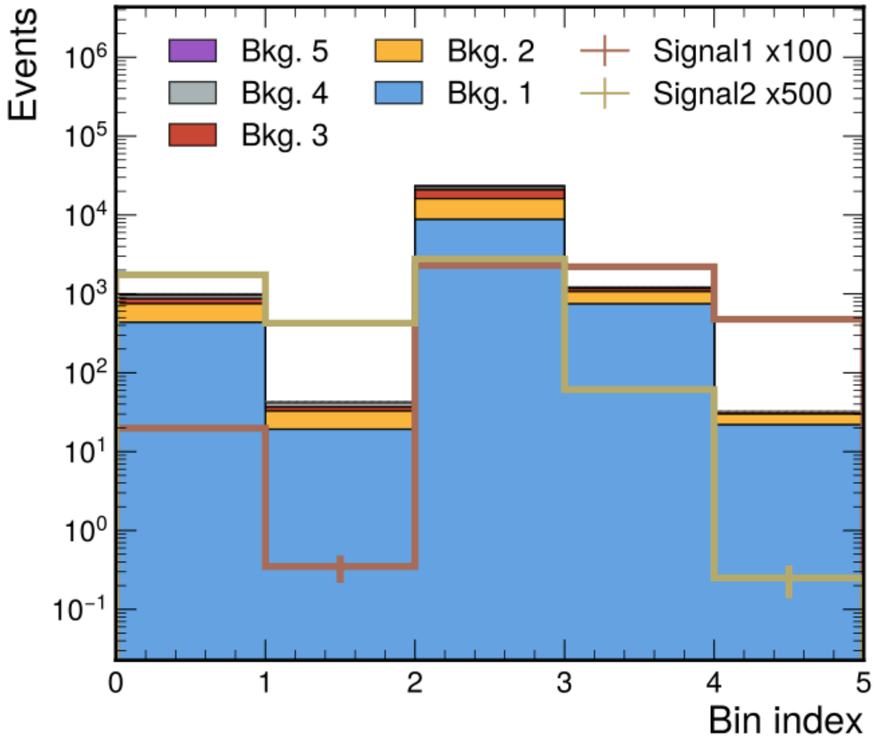
| | Scenario 1 | | | Scenario 2 | | | |
|---|---|---|---|---|---|---|---|
| Process | $\mu_x^{(1)}$ | $\mu_y^{(1)}$ | $\mu_z^{(1)}$ | $\mu_x^{(2)}$ | $\mu_y^{(2)}$ | $\mu_z^{(2)}$ | $\sigma$ [pb] |
| Signal 1 | 0.4 | −0.4 | 1.0 | 0.9 | −0.9 | 1.0 | 0.5 |
| Signal 2 | −0.4 | 0.4 | 1.0 | −0.9 | 0.9 | 1.0 | 0.1 |
| Background 1 | 0.0 | 0.0 | −0.5 | 0.2 | 0.2 | 0.5 | 100.0 |
| Background 2 | −0.2 | 0.0 | 0.0 | – | – | – | 80.0 |
| Background 3 | 0.1 | 0.4 | −0.3 | – | – | – | 50.0 |
| Background 4 | −0.1 | 0.6 | −0.2 | – | – | – | 20.0 |
| Background 5 | −0.2 | 0.1 | −0.1 | – | – | – | 10.0 |

$$\Sigma = \text{diag}(1,1,1) + 0.2\,(\mathbf{1} - \text{diag}(1,1,1))$$

# Optimized GMM bins for three dimensional case



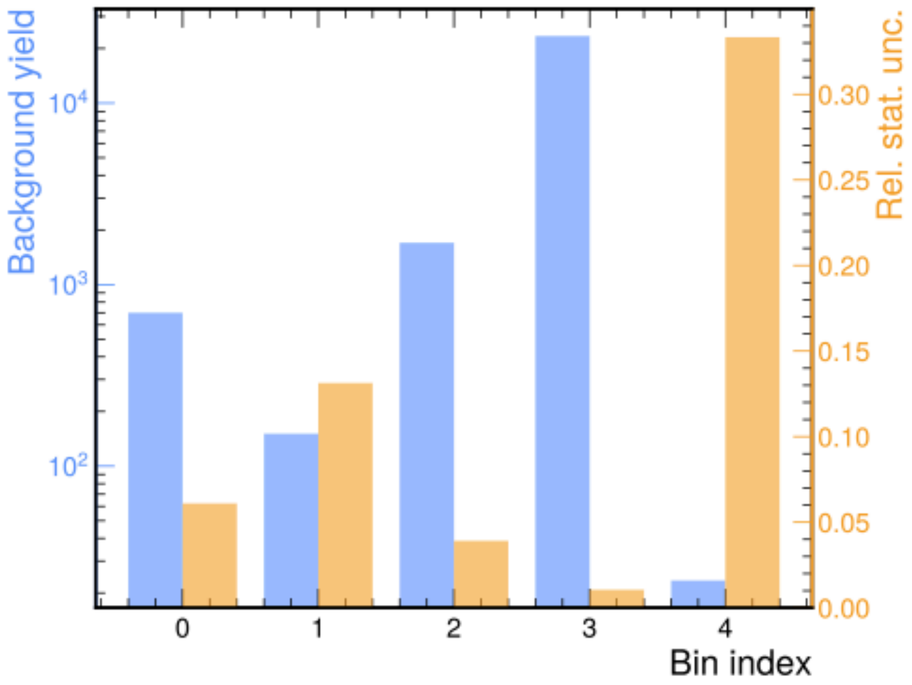

(e) First scenario




(f) Second scenario

# Penalties for large relative background uncertainty

Relative background uncertainty of 50%




Relative background uncertainty of 10%