Incorporating Physical Priors into Weakly-Supervised Anomaly Detection

Alkaid Cheng, Gup Singh, Ben Nachman

arXiv: 2405.08889 (PRL)

June 17, 2025

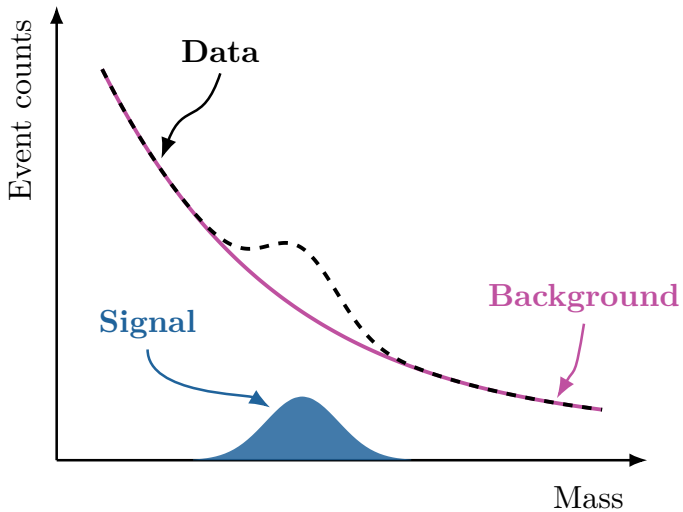
Two Complementary Approaches to Search for New Physics:

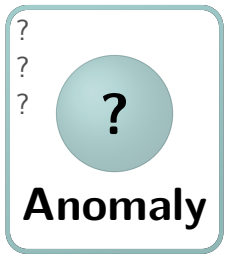
1. Dedicated Search (Majority of Analyses)

- Focus on **specific, well-motivated region of phase space**
- **More sensitive**, but **may not be looking at the right place**

2. Anomaly Detection

- Identify deviations **without strong model assumptions**
- **Less sensitive**, but **could search broadly across many possibilities**





Weakly Supervised Learning

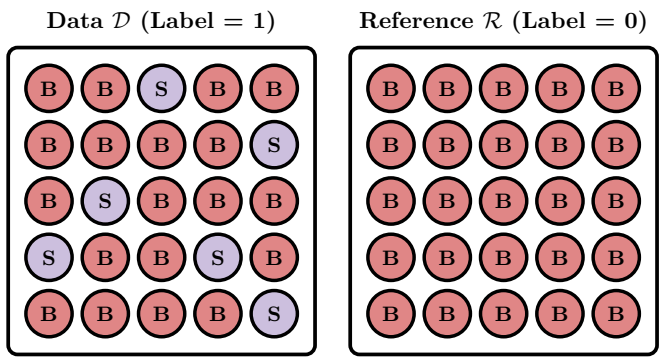
- No truth labels specified, but with some knowledge about their relative composition in the samples

CWoLA Method [1708.02949]

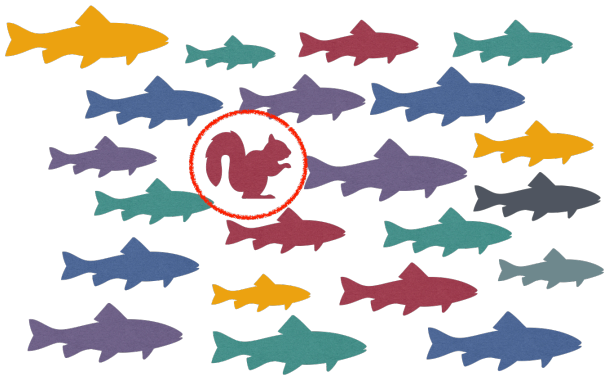
- The **optimal classifier between two mixed samples** (mixed signal and background) is also the **optimal classifier between signal (S) and background (B)**
- A convenient choice is:

  Data ($\mathcal{D}$): Observed data (may contain signal)

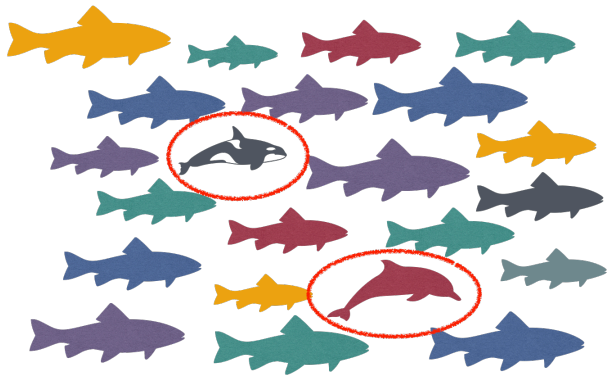
  Reference ($\mathcal{R}$): Pure background



Anomaly Detection - Weakly Supervised Approach



Unconstrained Anomaly



Fish-like Anomaly

Limitation of Traditional CWoLA Method

- Weakly classifier **does not know any physics about the data**
- **Model too flexible** to fit different forms of signal $\rightarrow$ **Poor sensitivity**
- Result is **difficult to interpret**: don't know what the anomalies correspond to
- Model easily **misguided by noisy/irrelevant features**

Prior Assisted Weak Supervision (PAWS) [2405.08889]

- Assume a **broad class of signals** that could describe the data
- Constrain the weakly supervised (WS) model to a **physically motivated manifold**
- **Balance** between performance and model independence

Core Idea of PAWS

- **Pre-train** a fully supervised classifier $f_{\text{FS}}(x)$ that serves as the **prior**
- **Parameterize the model by the physics parameters θ** of the signals so that it is sensitive to **generic signals from a broad parameter space**
- **Re-express** the weakly supervised model $f_{\text{WS}}(x)$ in terms of $f_{\text{FS}}(x, \theta)$, recognizing

$$P_D(x, \theta) = \mu P_S(x, \theta) + (1 - \mu) P_B(x)$$

$f_{\text{FS}}(\vec{x}, \theta)$ approximates the likelihood ratio:

$$f_{\text{FS}}(x, \theta) \approx \frac{P_S(x|\theta)}{P_S(x|\theta) + P_B(x)} \qquad \Lambda_{\text{FS}} \equiv \frac{P_S(x|\theta)}{P_B(x)} = \frac{f_{\text{FS}}(x, \theta)}{1 - f_{\text{FS}}(x, \theta)}$$

which gives

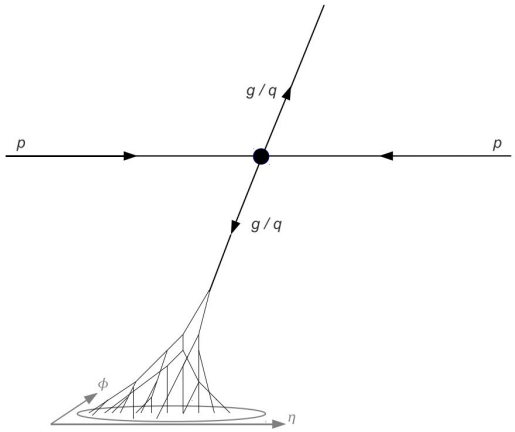
$$f_{\text{WS}}(x, \theta) \approx \frac{P_D(x|\theta)}{P_D(x|\theta) + P_R(x)} = \frac{\mu \Lambda_{\text{FS}}(x|\theta) + 1 - \mu}{\mu \Lambda_{\text{FS}}(x|\theta) + 2(1 - \mu)}$$

Core Idea of PAWS

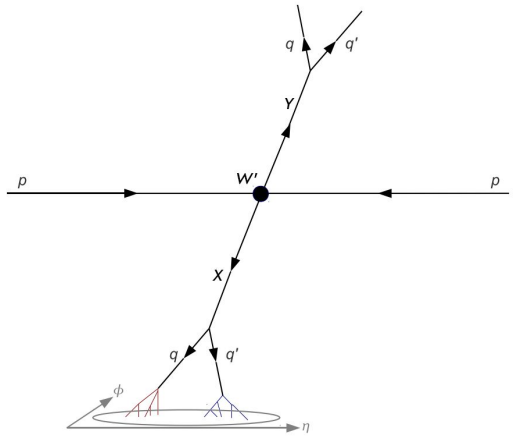
- **Pre-train** a fully supervised classifier $f_{\text{FS}}(x)$ that serves as the **prior**
- **Parameterize the model by the physics parameters θ** of the signals so that it is sensitive to **generic signals from a broad parameter space**
- **Re-express** the weakly supervised model $f_{\text{WS}}(x)$ in terms of $f_{\text{FS}}(x, \theta)$
- Freeze weights and biases of $f_{\text{FS}}(x, \theta)$ and let $f_{\text{WS}}(x, \theta)$ **learn the values of θ and μ**
- When θ matches the correct value corresponding to the signal, **physics knowledge from FS drives the performance of WS** $\rightarrow$ found the anomalous signal!

Physics Scenario: Heavy Resonance $W' \to XY$ in Dijet Final State

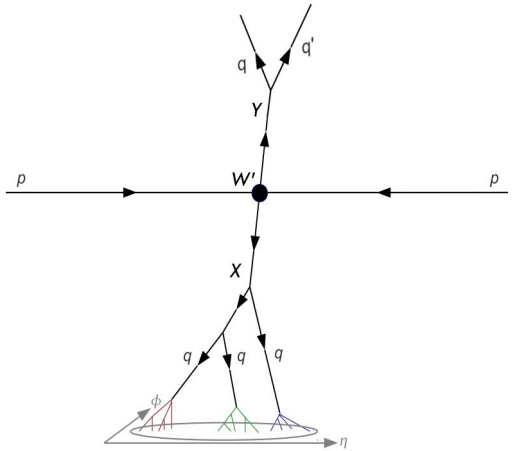
- Extension of LHC Olympics dataset with $W' \to X(qq)Y(qq)$ (2-prong) and $W' \to X(qqq)Y(qq)$ (3-prong) signals and QCD(qq) backgrounds



QCD (qq)
Background



$W' \to X(qq)Y(qq)$
2-prong Signal

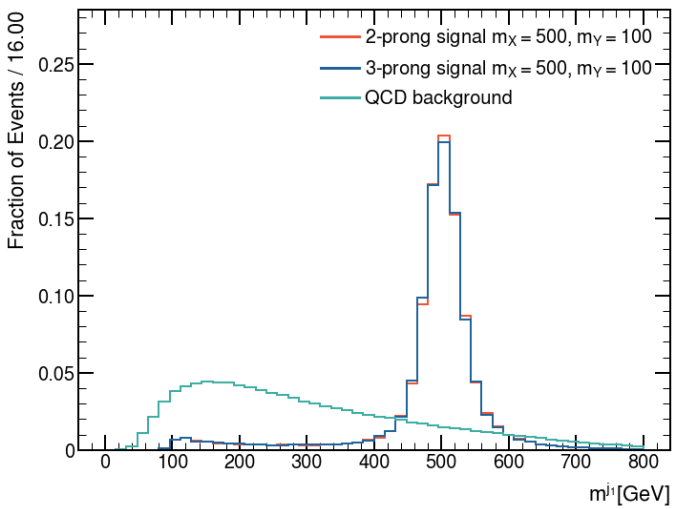


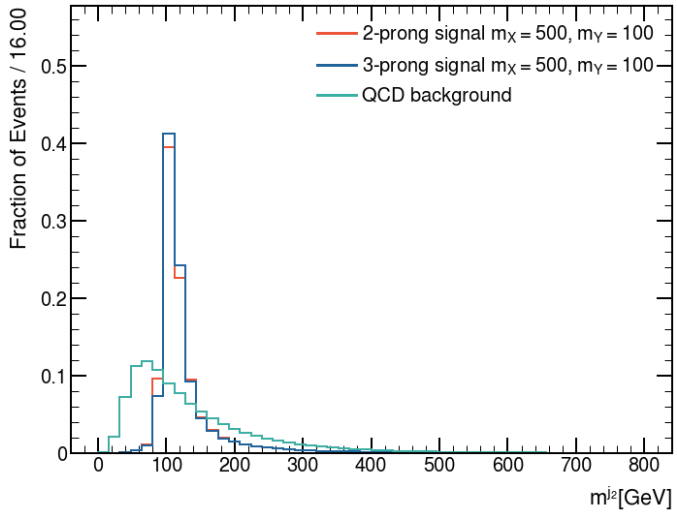
$W' \to X(qqq)Y(qq)$
3-prong Signal

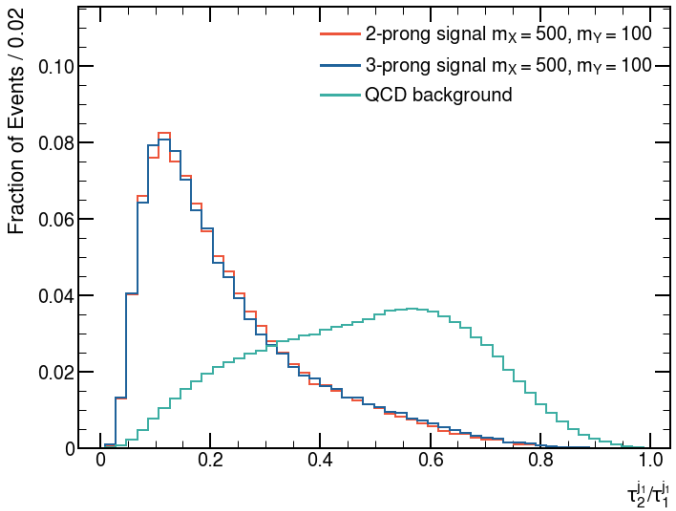
- Signals parameterised by $\theta = (m_X, m_Y)$ in the mass range of (50, 600) GeV with 50 GeV intervals
- Features $\vec{x} = \{m_{J1}, m_{J2}, \tau_{21}^{J1}, \tau_{21}^{J2}, \tau_{32}^{J1}, \tau_{32}^{J2}\}$

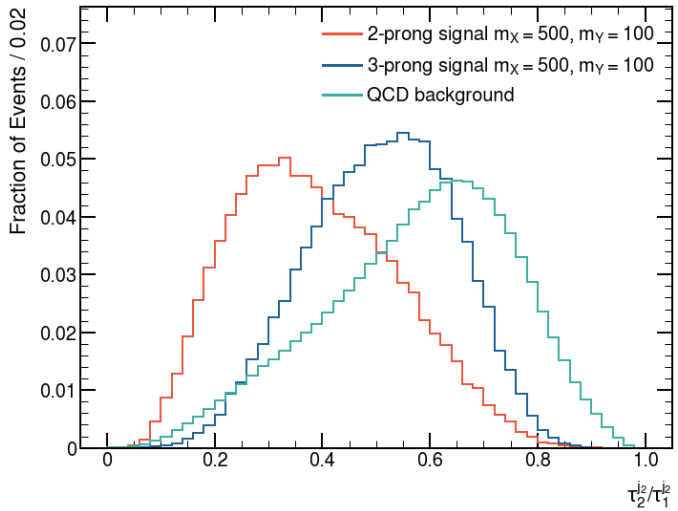
Feature Distributions

- Fully supervised model $f_{\text{FS}}(x, \theta)$ learns to separate signals from background and the parametrization of signals in terms of θ

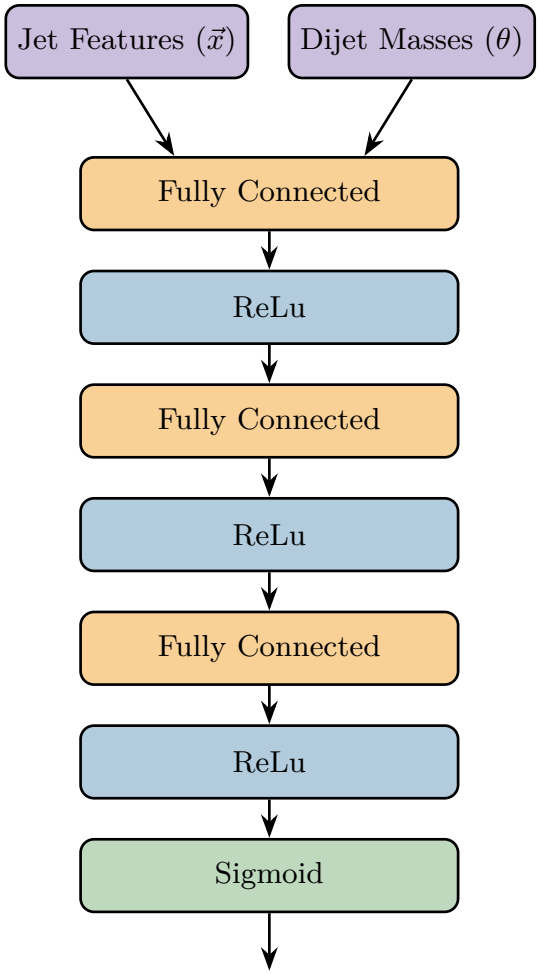








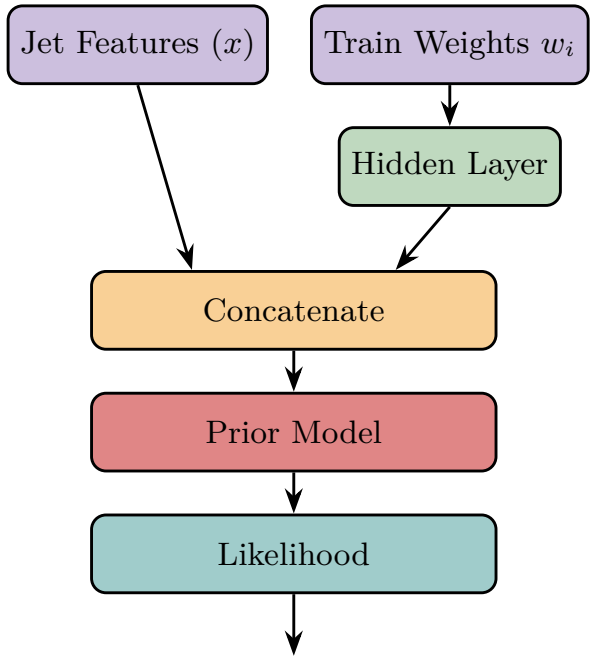
Model Architectures



Fully Supervised Classifier

(Prior Model)

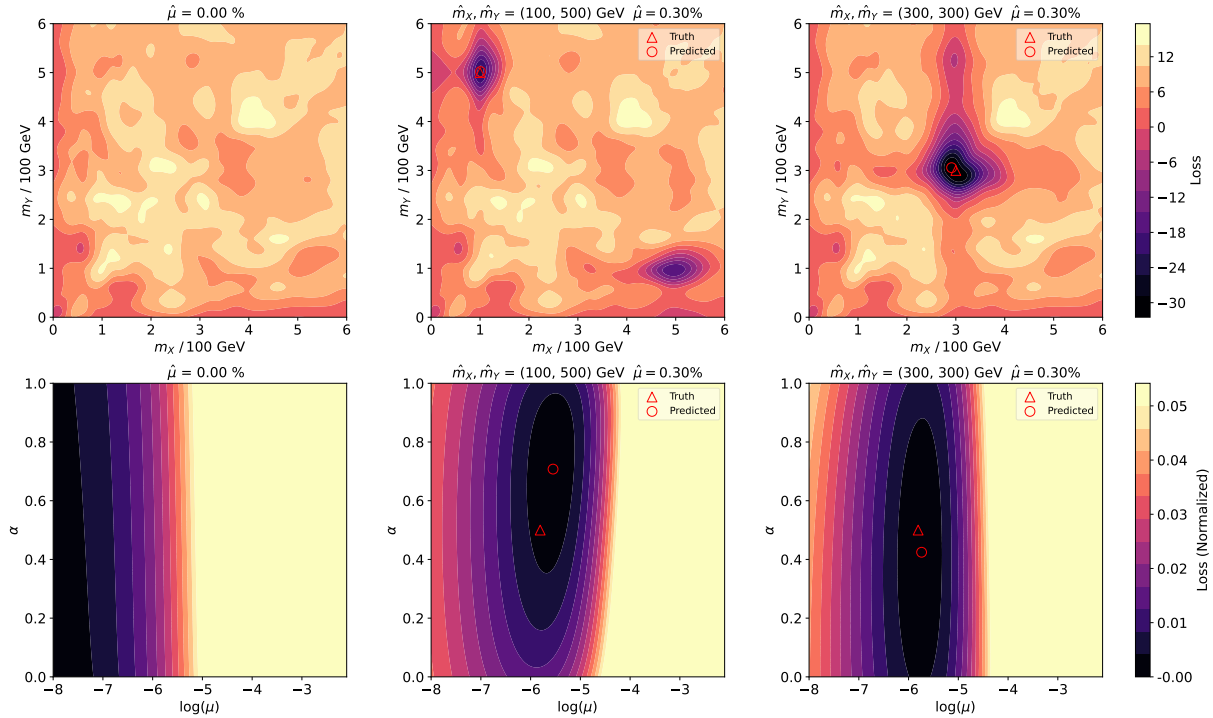
- Turn feature input θ into **learnable neural network weights** w_i
- Freeze supervised (prior) model



Weakly Supervised Classifier

Loss Landscapes

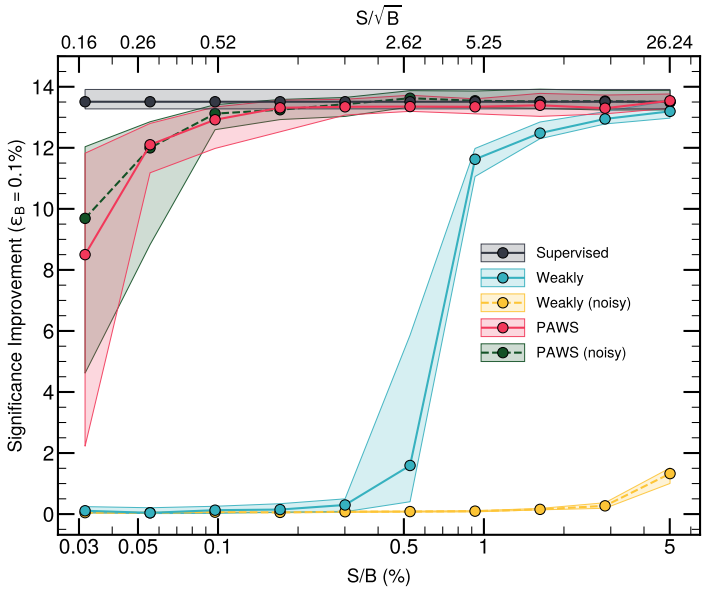
- Scanning **average loss values on test samples in a 2D parameter space**, i.e. the physics parameters (m_X, m_Y).



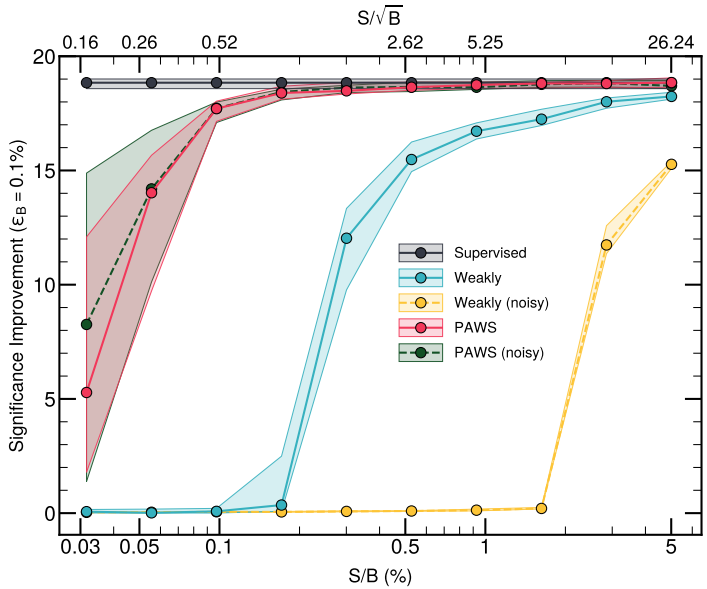
- With **no signal injected**, the learned μ **is consistent with 0**
- With small injected signal of 0.3% at $(m_X, m_Y) = (300, 300)$ GeV and $(100, 500)$ GeV, **minimum loss appears at the correct mass and** μ as guided by the prior model

Significance Improvement Characteristic (SIC)

- **SIC =** $\varepsilon_S/\sqrt{\varepsilon_B}$ = Factor by which the **signal significance increases for a cut on the classifier output** at a given ε_B



$(m_X, m_Y) = (100, 500)$ GeV

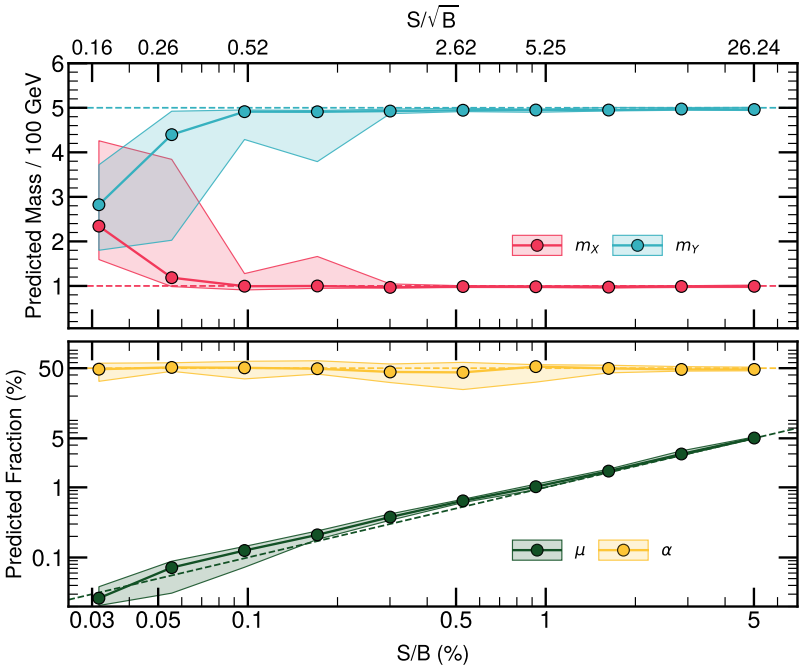


$(m_X, m_Y) = (300, 300)$ GeV

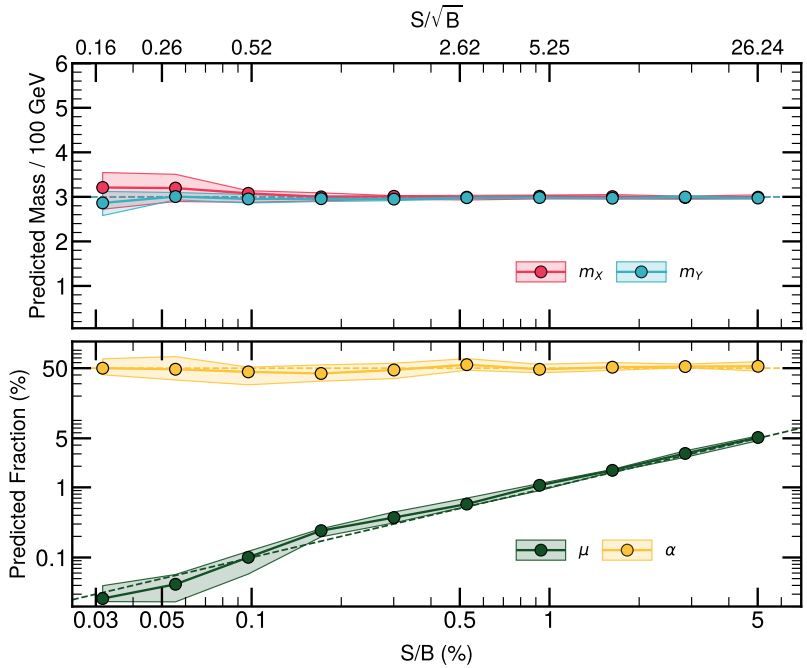
- When correct θ are chosen, **PAWS matches performance of FS model**
- PAWS achieves **sensitivity to signals 10 times weaker (0.03%)** compared to classical weakly-supervised approach
- Performance **unaffected by noisy features**

Parameter Prediction

- PAWS also **gives the values of the physics parameters** (m_X, m_Y, signal fraction μ and branching ratio α)
- The **learned parameters are physical and thus directly interpretable** when the anomaly is in the pre-trained model class



$(m_X, m_Y) = (100, 500)$ GeV



$(m_X, m_Y) = (300, 300)$ GeV

Conclusion and Outlook

Conclusion

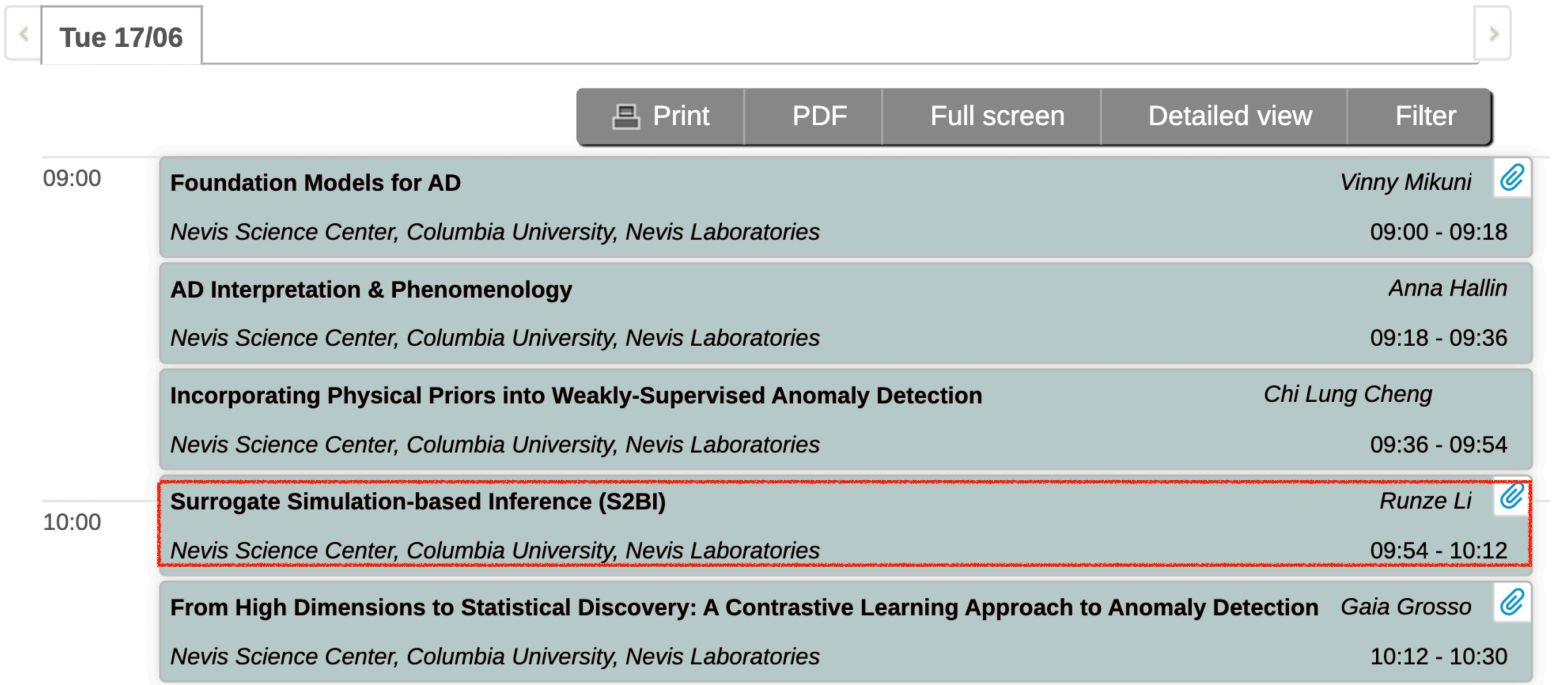
- PAWS introduces a way to **inject physical priors into weakly supervised training**, pushing **state-of-the-art sensitivity** for anomaly detection
- **Interpretable physical parameters** and **robustness against noisy features**

Outlook

- **Unbinned, High-Dimensional Statistical Inference**: Turn WS model directly into a likelihood estimator
- **Generative Background**: Replace simulated background with data-driven background
- **Scaling Up:** Higher dimensional parameter spaces (θ) to cover more complex theories.
- **Beyond the Priors:** Study the sensitivity to anomalies that are *similar* but not exactly in the pre-trained class.
- **Beyond Physics:** This method of "prior-assisted" learning is general. It could be applied to anomaly detection in other scientific fields like astrophysics or materials science.

Generator Based Inference (GBI) - PAWS

- Follow up work (2506.00119) to be presented by **Runze**

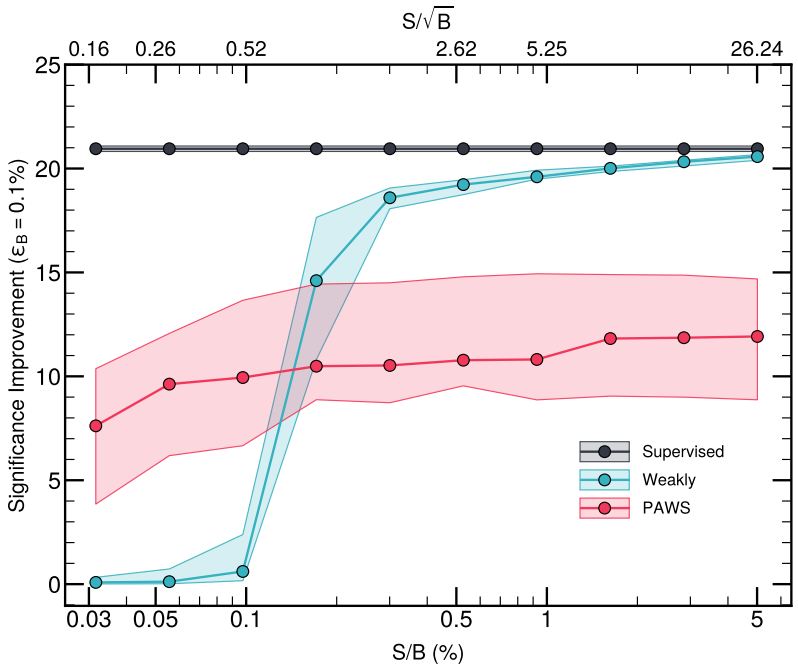


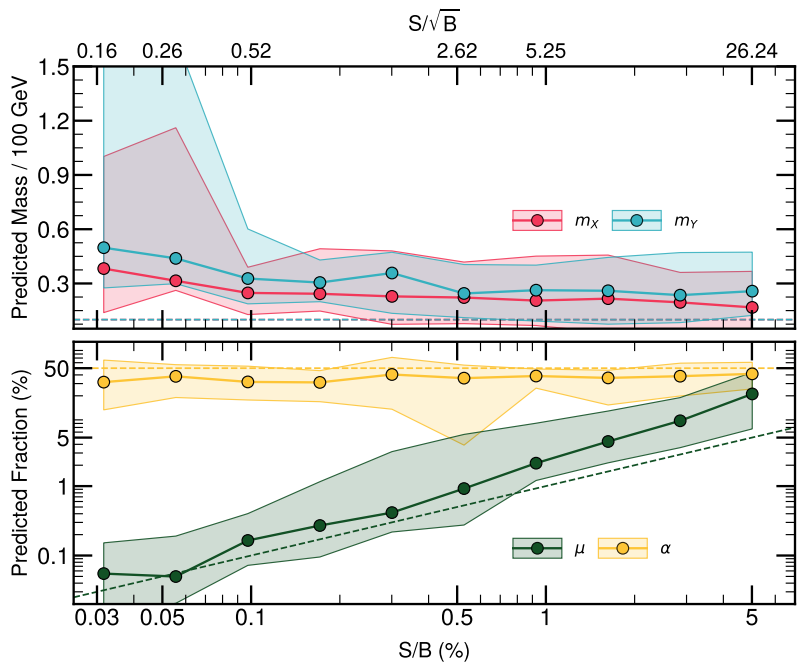
Acknowledgment

Thank You!

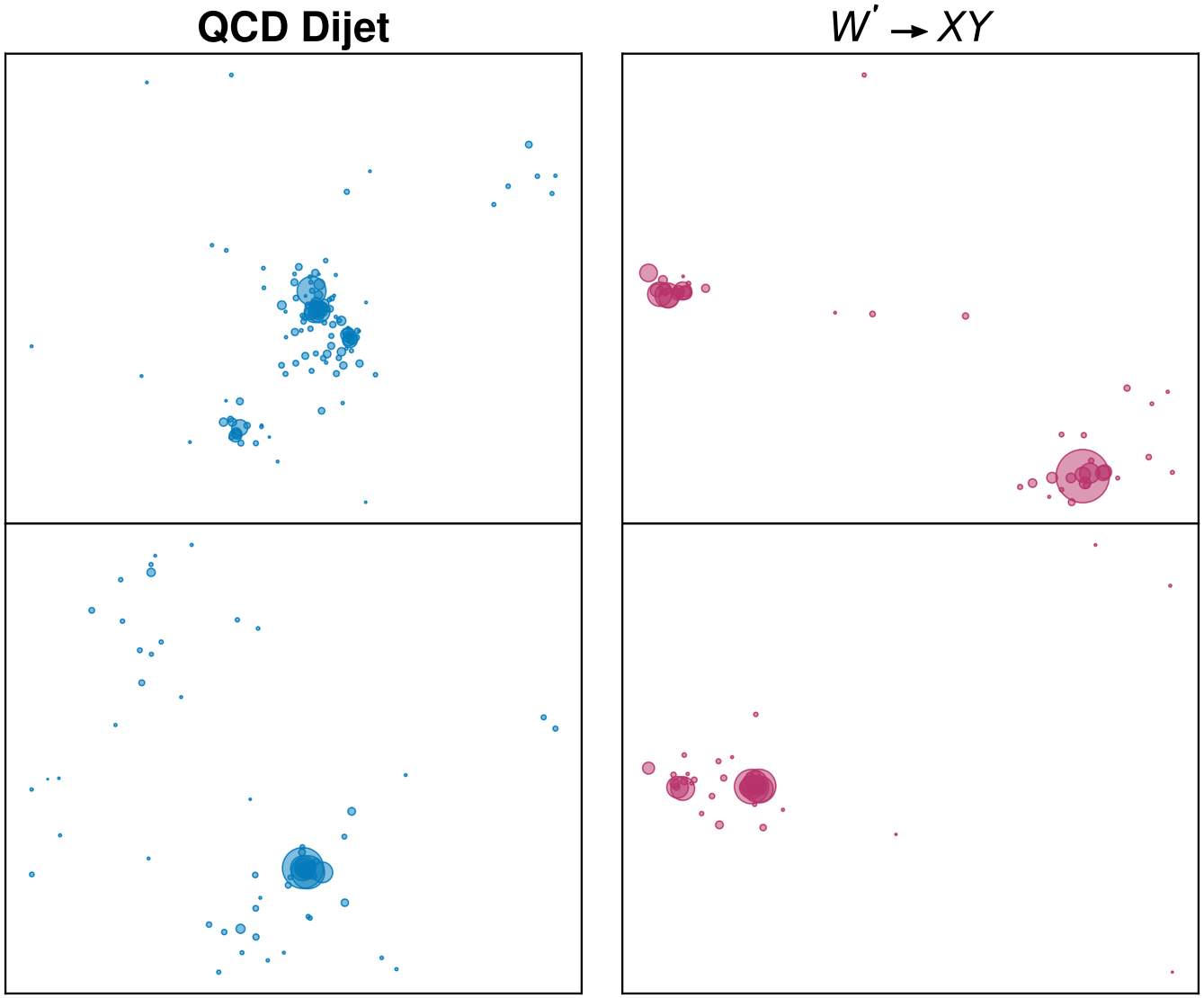
Out-of-Prior Case: (m_X, m_Y) = (10, 10) GeV

- Still able to offer some degree of improvement to sensitivity
- Unable to reach supervised performance (prior does not match data)

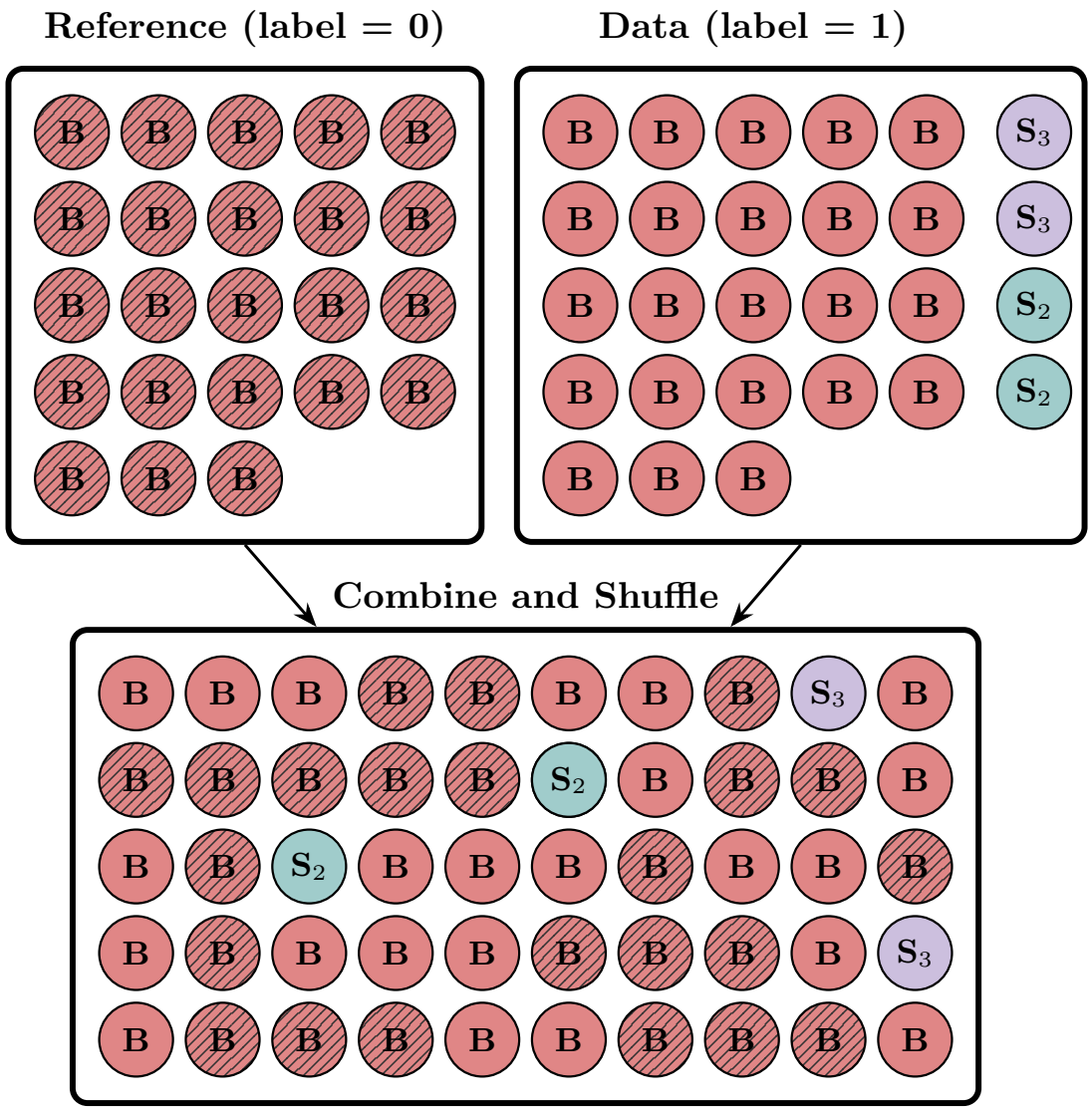




Jet Substructure of Dijet Signals and Background

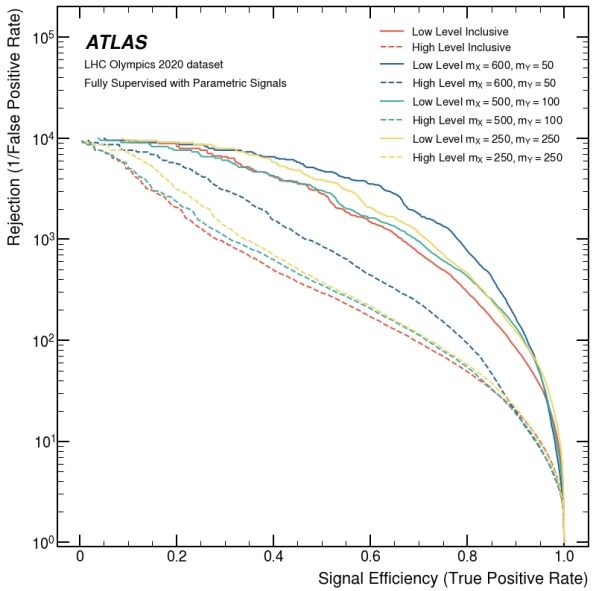


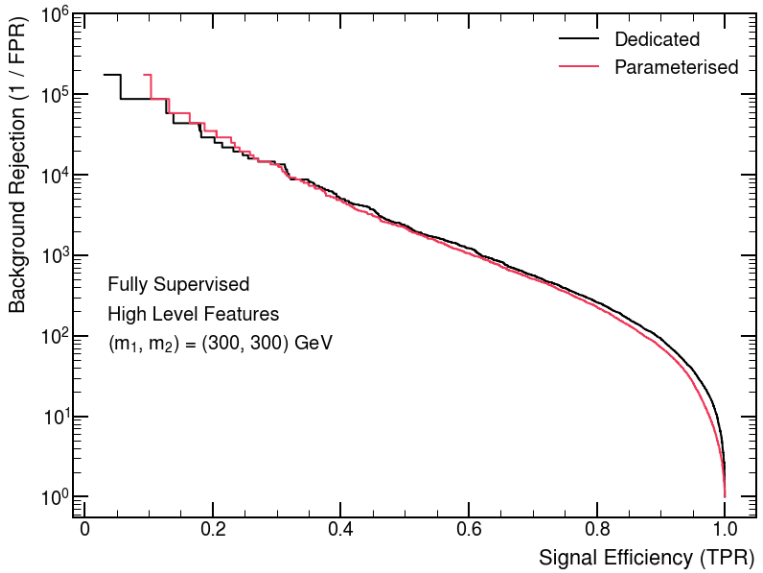
Appendix: Weakly Supervised Dataset



Appendix: Supervised Model Performance

- (Left Plot) High-dimensional data readily improves performance of pretrained model
- (Right Plot) Training a parameteric classifier (on θ) does not degrade discriminating power for individual θ





Traditional Approach

- High dimensional data is hand-crafted into **low dimensional summaries** (e.g. histograms) with simple cut-based event selections

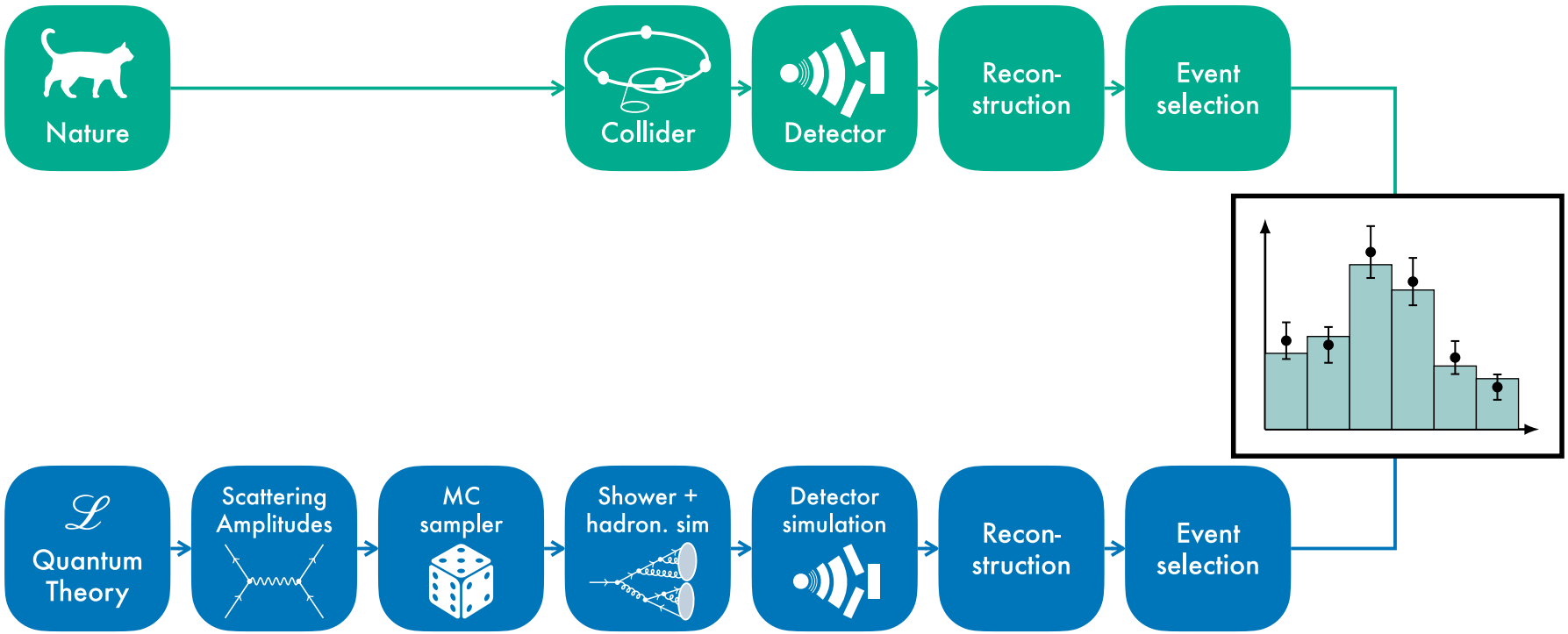
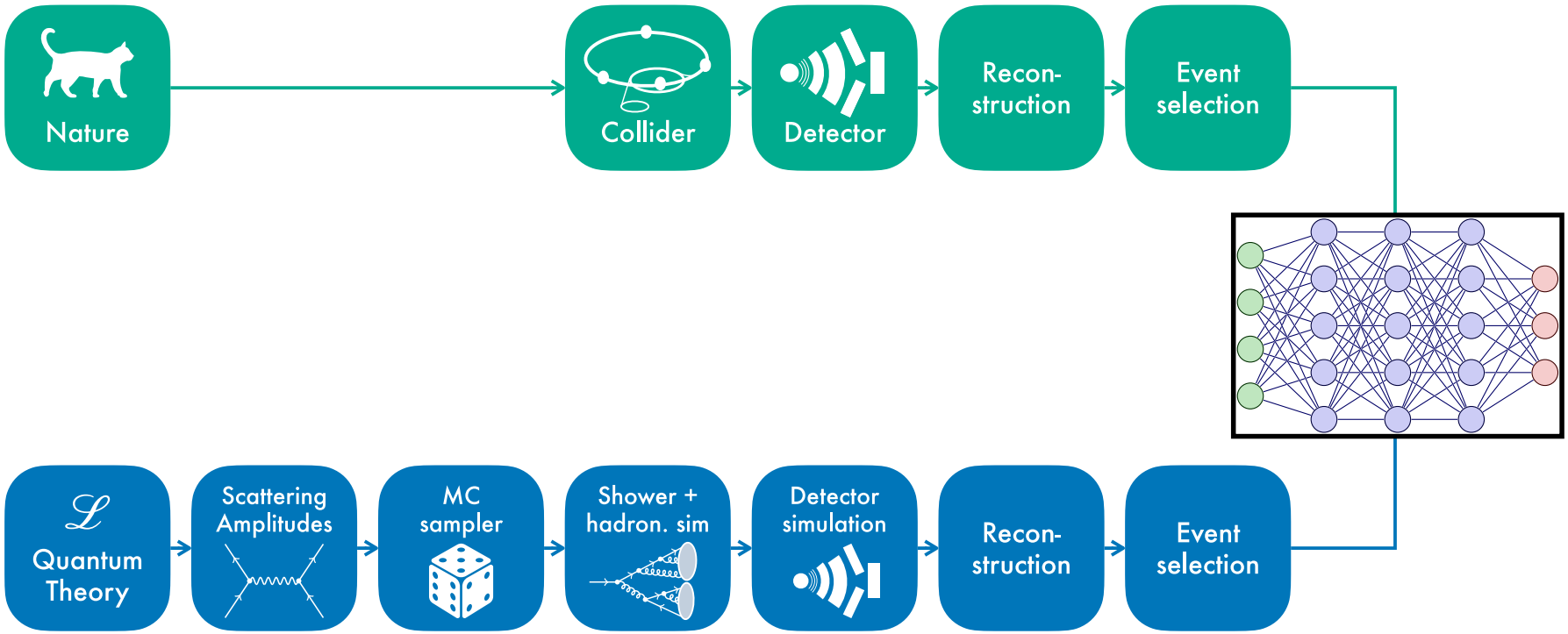


Figure Credit: Ramon

ML Approach

- Directly compare data and simulation in the natural **high-dimensional space**



Appendix: Resonant Search

- In reality, signal is often buried in a huge pile of background
- Any bump in the data would be washed away by statistical fluctuations and/or uncertainties
- Solution: **Train a classifier to distinguish signal and background** using the high-dimensional features