Incorporating Physical Priors into Weakly Supervised Anomaly Detection

Chi Lung Cheng (University of Wisconsin-Madison)
Gup Singh (University of California Berkeley), Ben Nachman (Stanford)

Introdution

- **Anomaly Detection (AD)** identifies deviations in data without strong model assumptions.
- One paradigm of AD is **weakly supervised (WS) learning**, where no truth labels are specified, but with some knowledge about their composition.
  - One of the most widely used WS strategy is the **CWoLA method** which trains a classifier between a data sample and a reference sample (purely background). This strategy results in the optimal classifier between signal and background.
  - **Limitation of CWoLA**
    - Weakly classifier **does not know any physics about the data**
    - **Poor sensitivity** due to overly flexible model
    - Anomolies are **not interpretable**
    - Model easily **misguided by noisy/irrelevant features**

Prior Assisted Weak Supervision (PAWS)

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

Methodology

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

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

$f_{\text{FS}}(\vec{x},\theta)$ 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

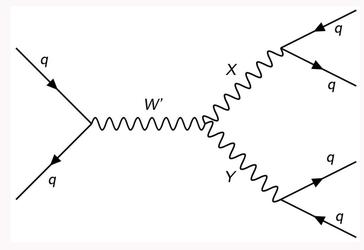
$$f_{\text{WS}}(x,\theta) \approx \frac{P_D(x|\theta)}{P_D(x|\theta)+P_R(x)} = \frac{\mu\Lambda_{\text{FS}}(x|\theta)+1-\mu}{\mu\Lambda_{\text{FS}}(x|\theta)+2(1-\mu)}$$

- Freeze weights and biases of $f_{\text{FS}}(x,\theta)$ and let $f_{\text{WS}}(x,\theta)$ **learn 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!

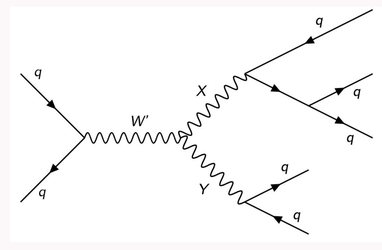
Datasets

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 backgrounds



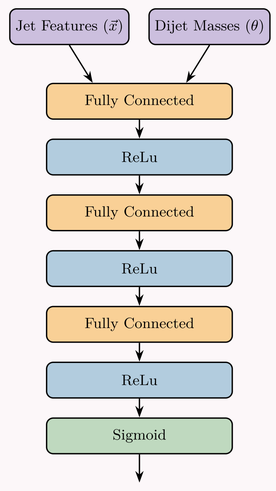
2-prong Signal



3-prong Signal

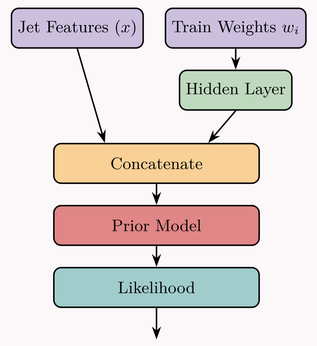
- Signals Parameters: $\theta = (m_X, m_Y)$ in the mass range of (50, 600) GeV with 50 GeV intervals (144 signals in total)
- Train Features: $x = \{m_{J_1}, m_{J_2}, \tau_{21}^{J_1}, \tau_{21}^{J_2}, \tau_{32}^{J_1}, \tau_{32}^{J_2}\}$
  (mass and n-subjetiness ratios of each jet)

Model Architectures



Supervised Classifier (Prior)

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

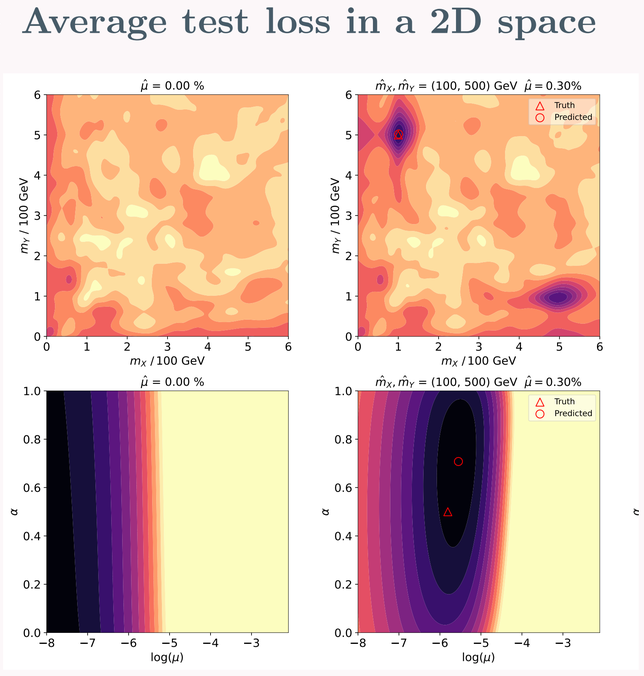


Weakly Classifier

Results

Loss Landscapes

- **Average test loss in a 2D space**

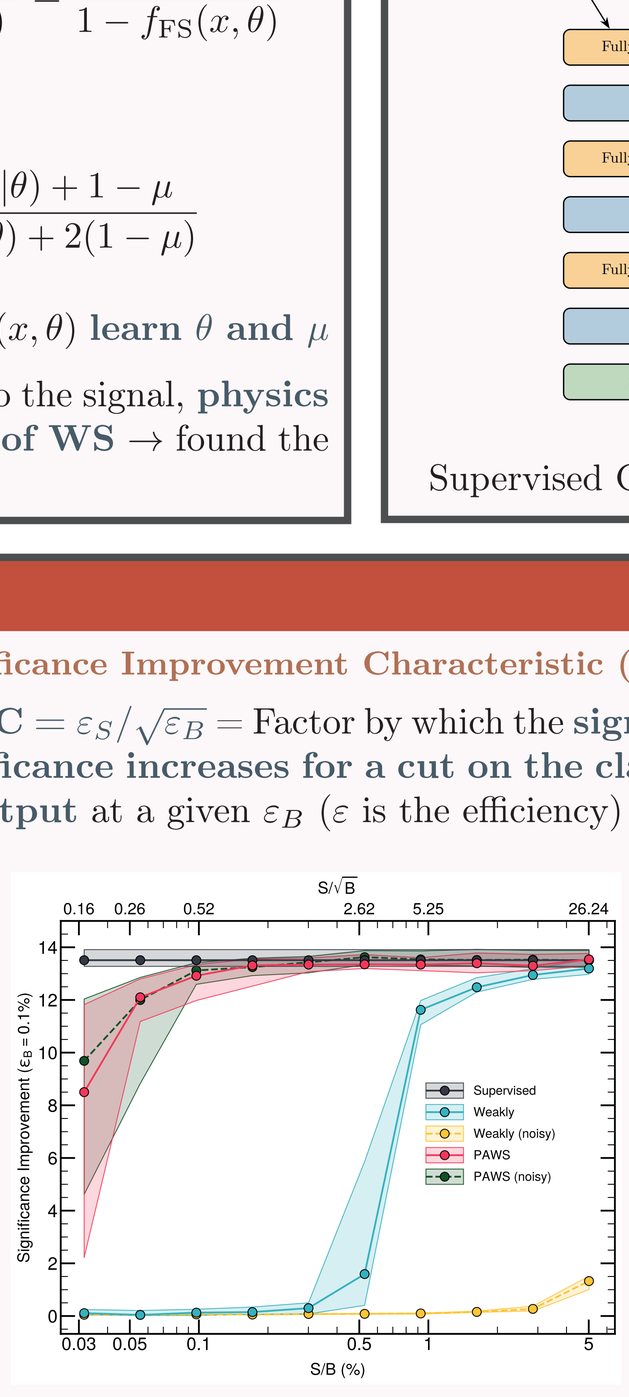


Top: (m_X, m_Y) Bottom: (μ, α)

- With **no signal injected**, the learned μ **is consistent with 0**
- With small injected signal of 0.3% at $(m_X, m_Y) = (100, 500)\,\text{GeV}$, **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 (ε is the efficiency)



$(m_X, m_Y) = (100, 500)\,\text{GeV}$

- 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)

$(m_X, m_Y) = (100, 500)\,\text{GeV}$

- The **learned parameters are physical and directly interpretable** when the anomaly is in the pre-trained model class

arXiv: 2405.08889