Super(vised) CWoLa

Sam Klein, **Stephen Mulligan**, Tobias Golling
ML4Jets

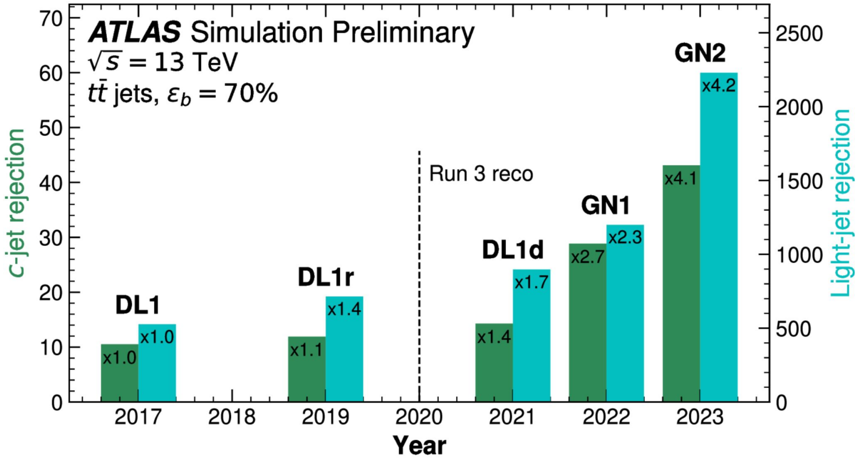
UNIVERSITÉ DE GENÈVE

Supervised ML trends

- Massive improvements in supervised learning
    - Architectural improvements
    - Lower level features

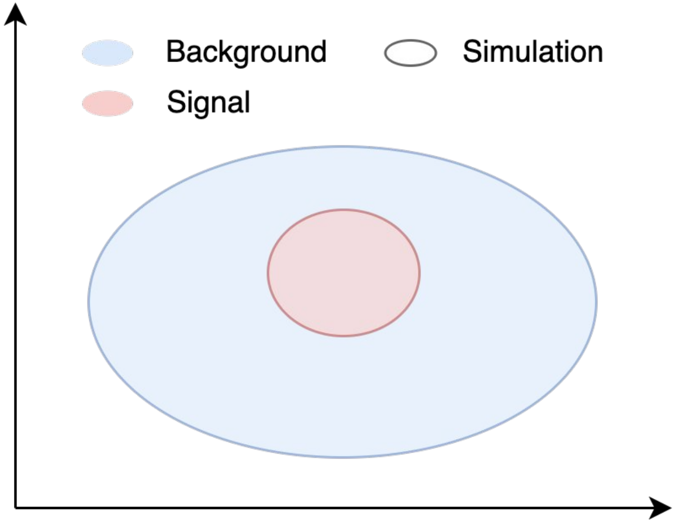
- Models **trained on simulation** but **applied to data**

- **Domain shift!**



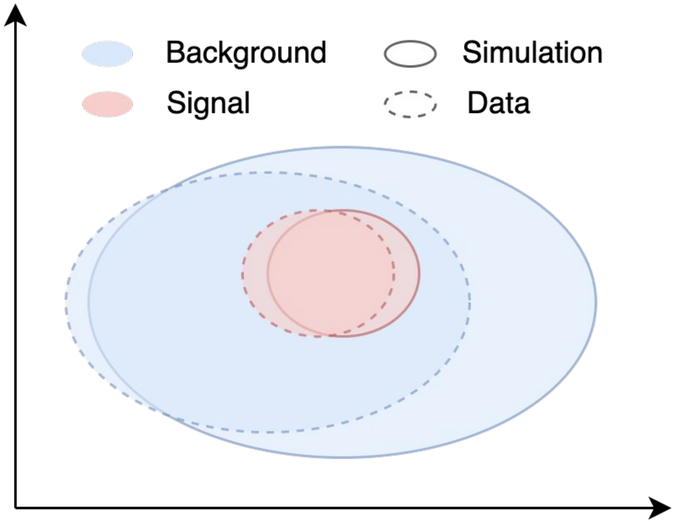
Supervised ML

- Focus on binary classification
- Signal vs background
- Want to enhance signal and reduce background
- Common task in ML 4 HEP



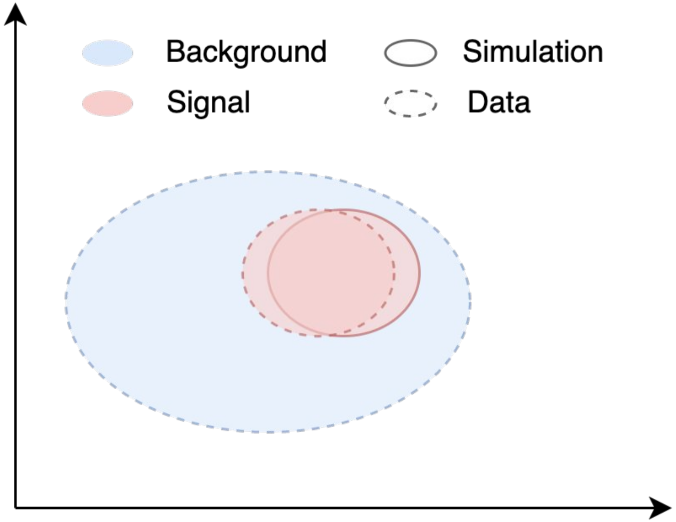
Domain shift

- Do models transfer?
    - Require calibrating
    - Lose efficiency
- Worse in high dimensions
    - Bigger shift for low level features
    - ML trend to 'go lower'…
- How to **reduce** the **impact** of shift?
    - Reduce sensitivity to mismodelling?
        - Adversarial attacks - Franck Rothen, 11:50 am
    - Reduce our **dependence on simulation?**



Super(vised) CWoLa

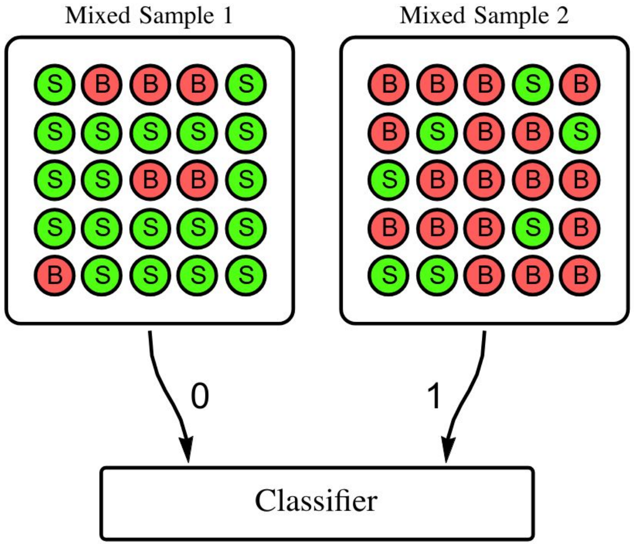
- **Can we drop background simulation?**
  - Will always be slightly mismodelled
  - If we don't need it, **drop it**



Standard CWoLa

- Classifier trained on two mixed samples M1 and M2

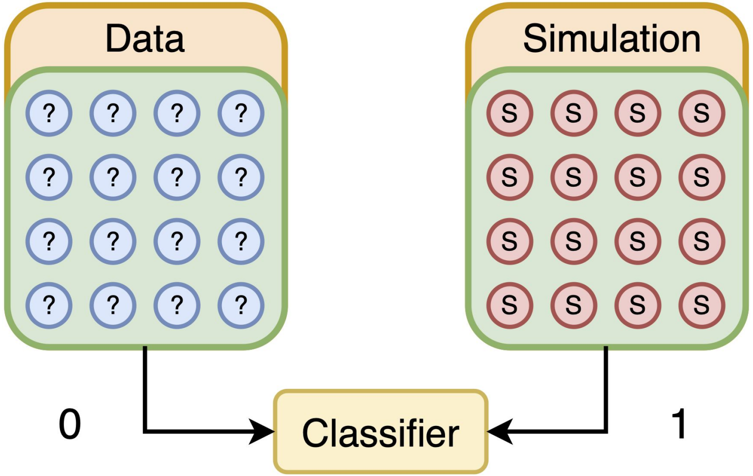
- **CWoLa Theorem**
  - The optimal classifier trained to distinguish M1 and M2 is also optimal for distinguishing S and B



1708.02949

Super(vised) CWoLa

- Data is an unknown mixture of signal and background
  - Could be pure background
  - Never background free
- Sample of simulated signal is **pure**
  - 100% signal samples
- CWoLa paradigm [1708.02949]
  - Label simulation one
  - Label data zero
  - Train **optimal classifier**

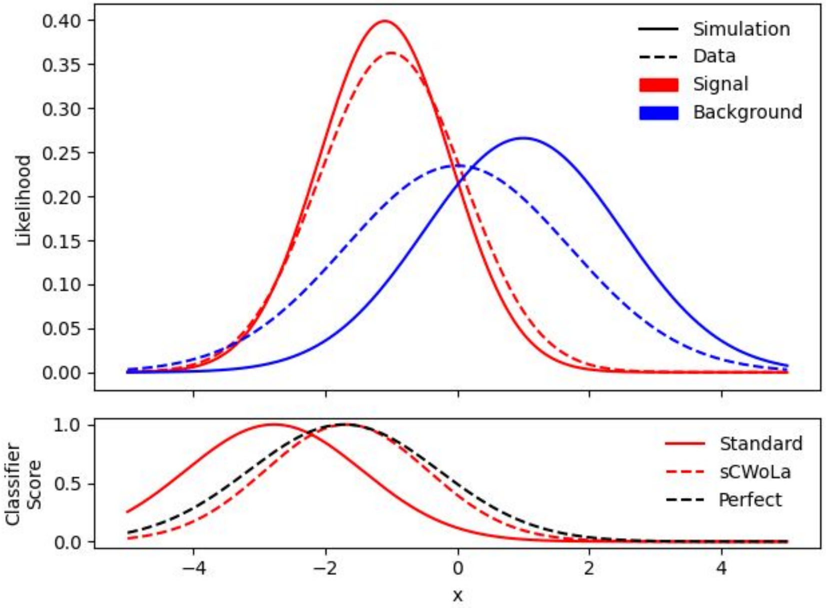


Super(vised) CWoLa

- Why does this work?
  - Classifier learns the likelihood ratio
  - Will learn the wrong likelihood ratio on simulated background!

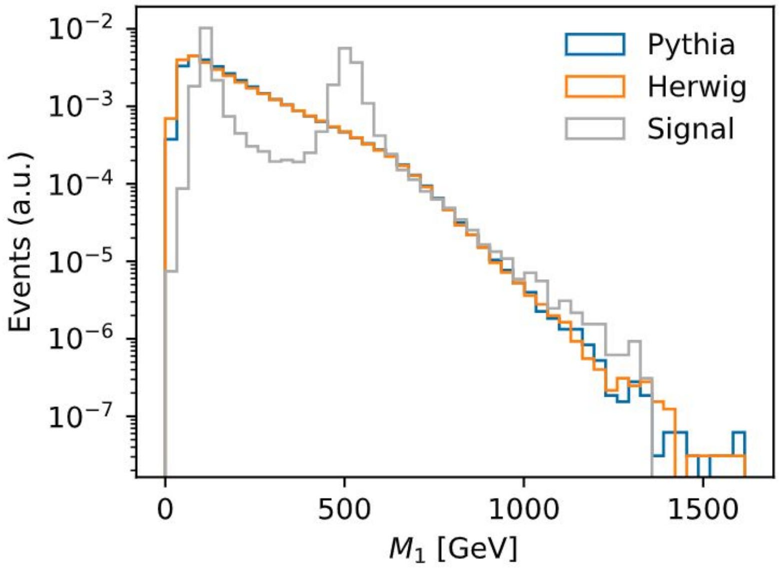
- **Enhance signal in data not simulation!**

- **Assumption - Signal vs data likelihood ratio will be closer**



Experiments

- Use LHCO R&D dataset
  - Take Pythia as a proxy for data
  - Take Herwig as a proxy for simulation
  - Use Pythia signal
    - Not considering signal mismodelling
- Use high and low level features
  - High level - jet mass, subjettiness
  - Low level - p_T, $\Delta\eta$, $\Delta\phi$
- Consider different amounts of signal contamination



Experiments

- Use in the context of a resonant new physics search
  - Strong performance is not the only requirement, often have auxiliary requirements

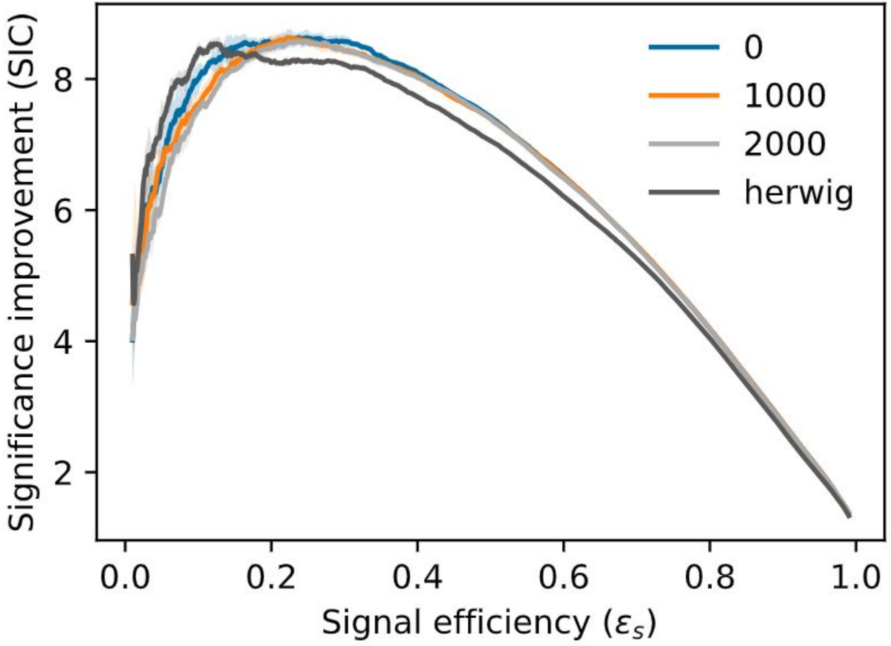
- E.g can we also decorrelate classifier from M_{JJ}?
  - Necessary for background estimate
  - Or just use data directly…
  - Using histograms

High level features

- Herwig
  - Standard approach - trained on simulated background vs signal, evaluated on data

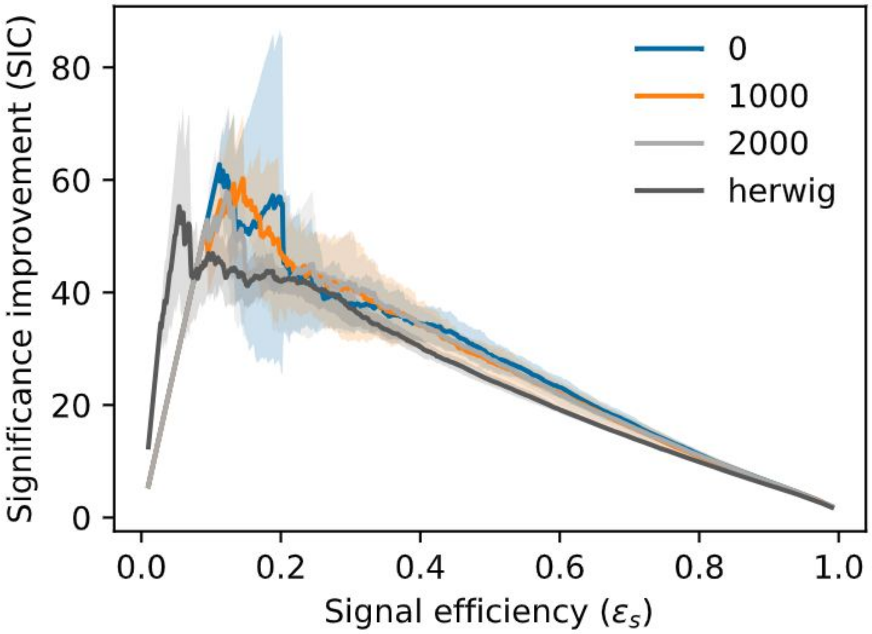
- sCWoLa slightly outperforms

- Mismodelling in high level features is small



Low level features

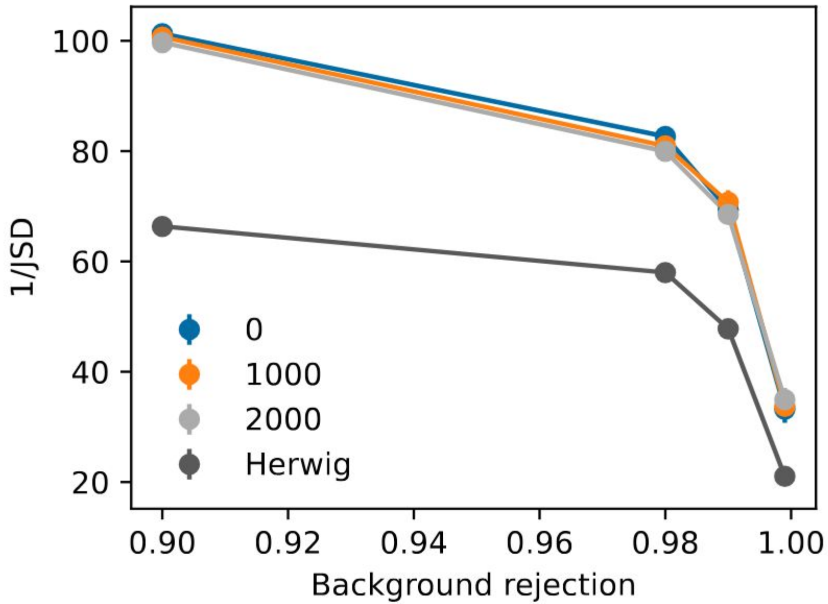
- Mismodelling increases
- sCWoLa outperforms



High Level Decorrelation

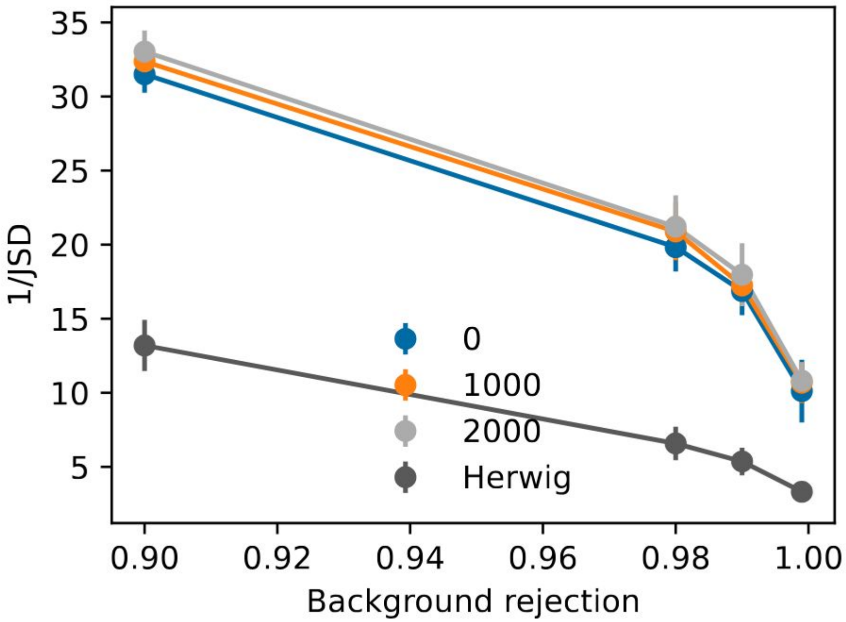
- Herwig - consistently lower 1/JSD

- Over-reliance on mass to achieve comparable performance



Low Level Decorrelation

- Similar trend at low level



Summary

- Simple method to **train classifiers without simulated background**

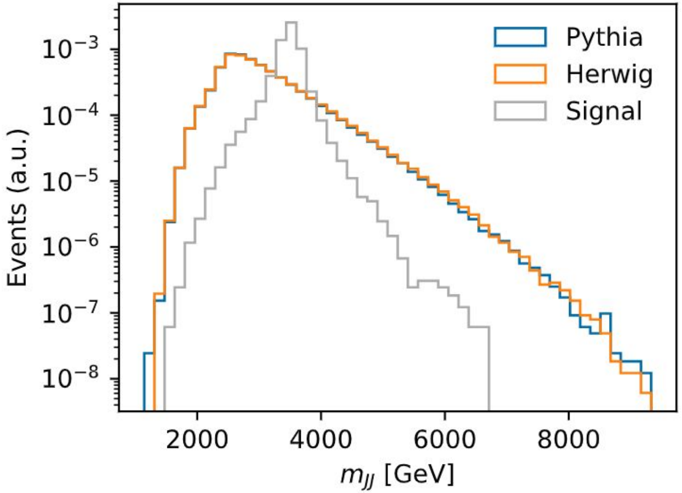
- With a data driven background estimate for M_{JJ} could allow for dedicated searches with no background simulation

- Simple idea which needs to be fully explored in real settings!

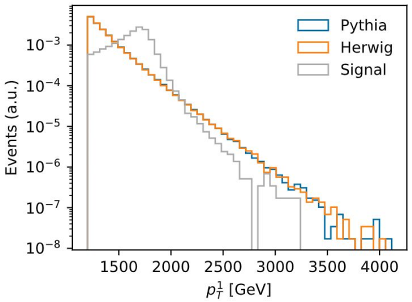
Backup

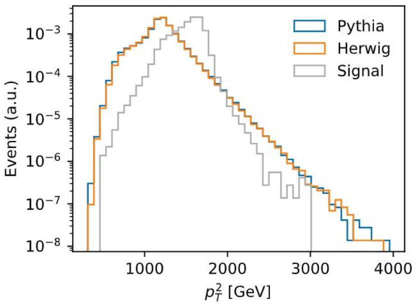
Dataset

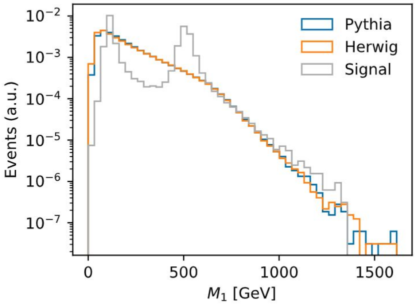
- Background - QCD Dijets with p_T = 1.3 TeV
- Signal - $W' \rightarrow XY$ with $m_{W'}$ = 3.5 TeV , m_X = 500 GeV , m_Y = 100 GeV
- Pythia 8 and Herwig++
- Delphes 3.4.1 with standard CMS detector card
- Fastjet with anti-kt, jet radius of 1

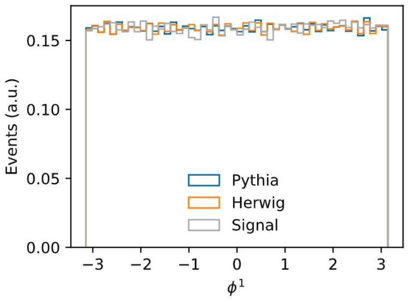


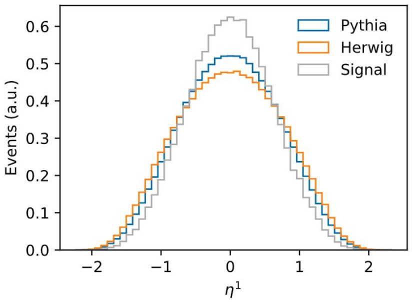
Dataset - High Level

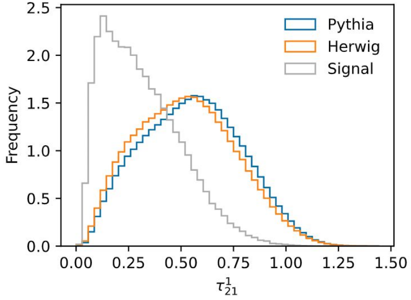


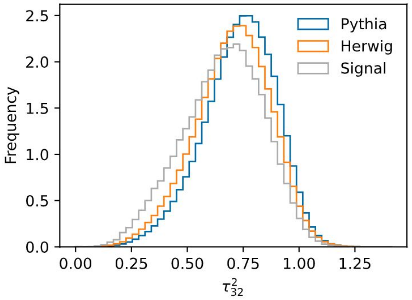




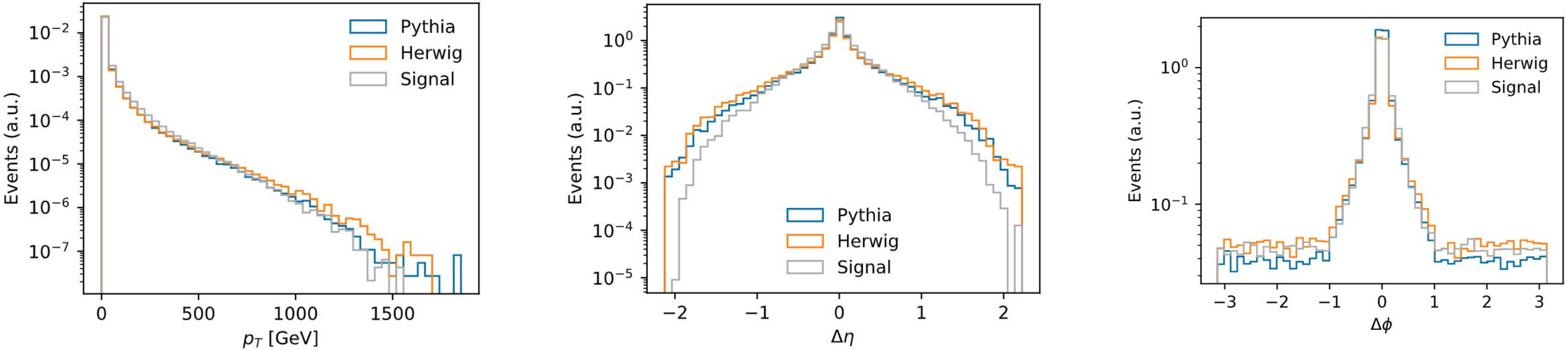








Dataset - Low Level



Models

- Low level - BDTs
- High Level - Transformers