Exploring machine learning methods for unbinned data analysis of Drell-Yan phase space at CMS

BSc. David Gutiérrez Menéndez Dr. Fernando Guzmán Martínez

November 1, 2023

Higher Institute of Tecnology an Applied Sciences, Havana University

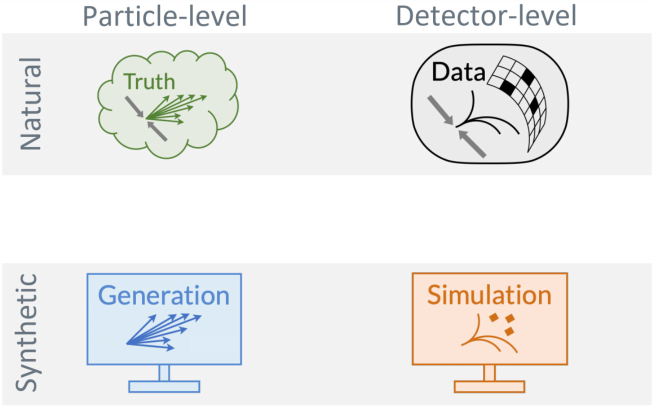
Introduction

Introduction

Particle detectors introduce smearing in measurements:

- Finite detector resolution
- Acceptance cuts (geometrical and physical)
- Binning effect in distributions

Goal: Use data (at detector level) and infer true desitributions (at particle level).

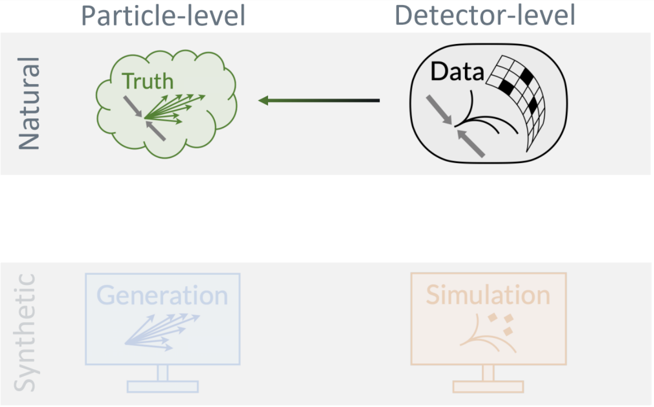


Introduction

Only then we can reliably compare the measurements against other sources:

- Semi-Inclusive theoretical predictions
- MC generator parameter tuning
- Other experiments

Unfolding (deconvolution).



Current Unfolding approach

Current Unfolding approach

Detector response can be modeled as a linear application.

$$x_{detector} = R * x_{truth} + x_{backg} \quad (1)$$

Where R represents the detector **response** as $R_{ij} = P(truthi|measurej)$.

Current Unfolding approach

Various approaches exist to unfolding:

- Naive R inversion

$$x_{truth} = R^{-1} * (x_{detector} - x_{backg}) \tag{2}$$

- Iterative Bayesian Unfolding (IBU): Given a **response** matrix and a prior distribution.

$$x_j^{(n)} = \sum_i P_{n-1}(truthi|measurej) * P(measurej) \tag{3}$$

Current Unfolding approach

Unfolding drawbacks:

- Works for binned data
- 1-dimensional by desing (very difficult to extend to higher dimensions)
- Does not account for relations between observables

Machine Learning approach

Distribution reweight.

- Using a binary classifier to distinguish between two distributions

- Likelihood ratio can be aproximated and used as weight function

$$w_{A\to B} = \frac{p_B(x)}{p_A x} \approx \frac{f(x)}{1 - f(x)} \tag{4}$$

- Distributions are binned for representation but weights are estimated per event

Machine Learning approach

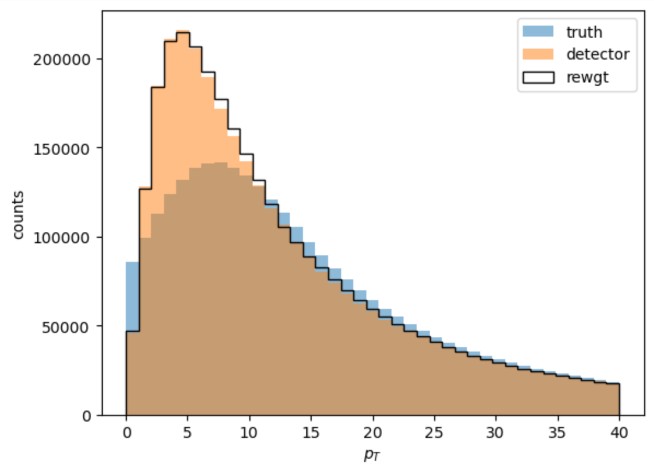


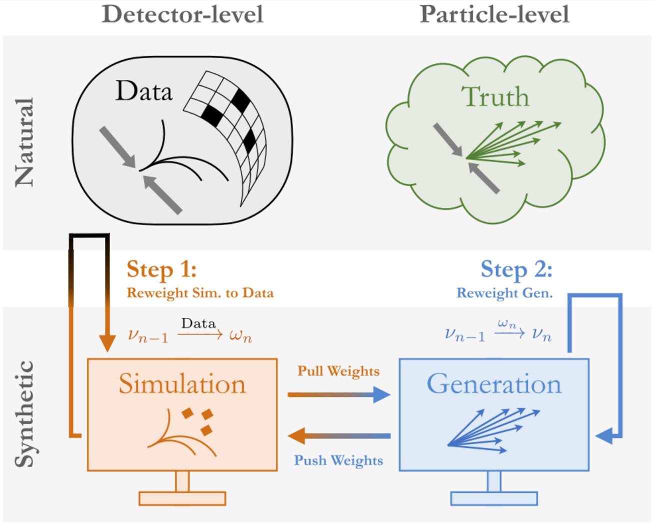
Figure 1: Transverse momentum distribution. $pp \to DY7TeV$.

Machine Learning approach

Advantages of ML reweight.

- Used for non-trivial distributions
- Easily extended to multiple dimensions
- Learns complex relations between observables

Iterative Neural Network Reweight: OmniFold.



Machine Learning approach

Iterative Neural Network Reweight: OmniFold.

- Continuous generalization of IBU
- Inherits all the advantages from ML reweight
- Unfolds experimental data using multiple dimensions (possibly full phase space)

Work in progess

Work in progess

Drell-Yan phenomenology analysis with CMS data.

- Explore HEP analysis viavility with **Python**
- Unlock ML ecosystem and methods
- Validation of NanoAODRun1 format for CMS data
- Improve accuaracy of infered distributions

Work in progess

Drell-Yan phenomenology analysis with CMS data.

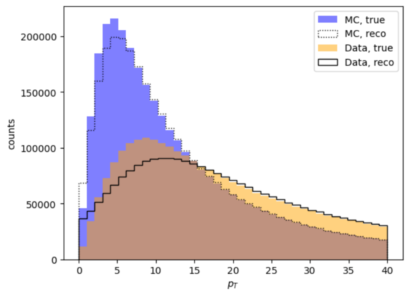


Figure 2: Syntetic and Natural distributions.

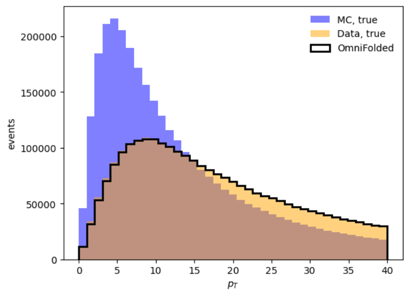


Figure 3: Result of reweighted distribution.

Work in progess

Future work.

- Full phase space unfolding
- Unbinned and data driven analysis
- Evaluate statistical and systematical uncertanties
- Search for more **robust** observables where current theoretical predictions struggle

Exploring machine learning methods for unbinned data analysis of Drell-Yan phase space at CMS

BSc. David Gutiérrez Menéndez Dr. Fernando Guzmán Martínez

November 1, 2023

Higher Institute of Tecnology an Applied Sciences, Havana University