Ideas for ML based unfolding

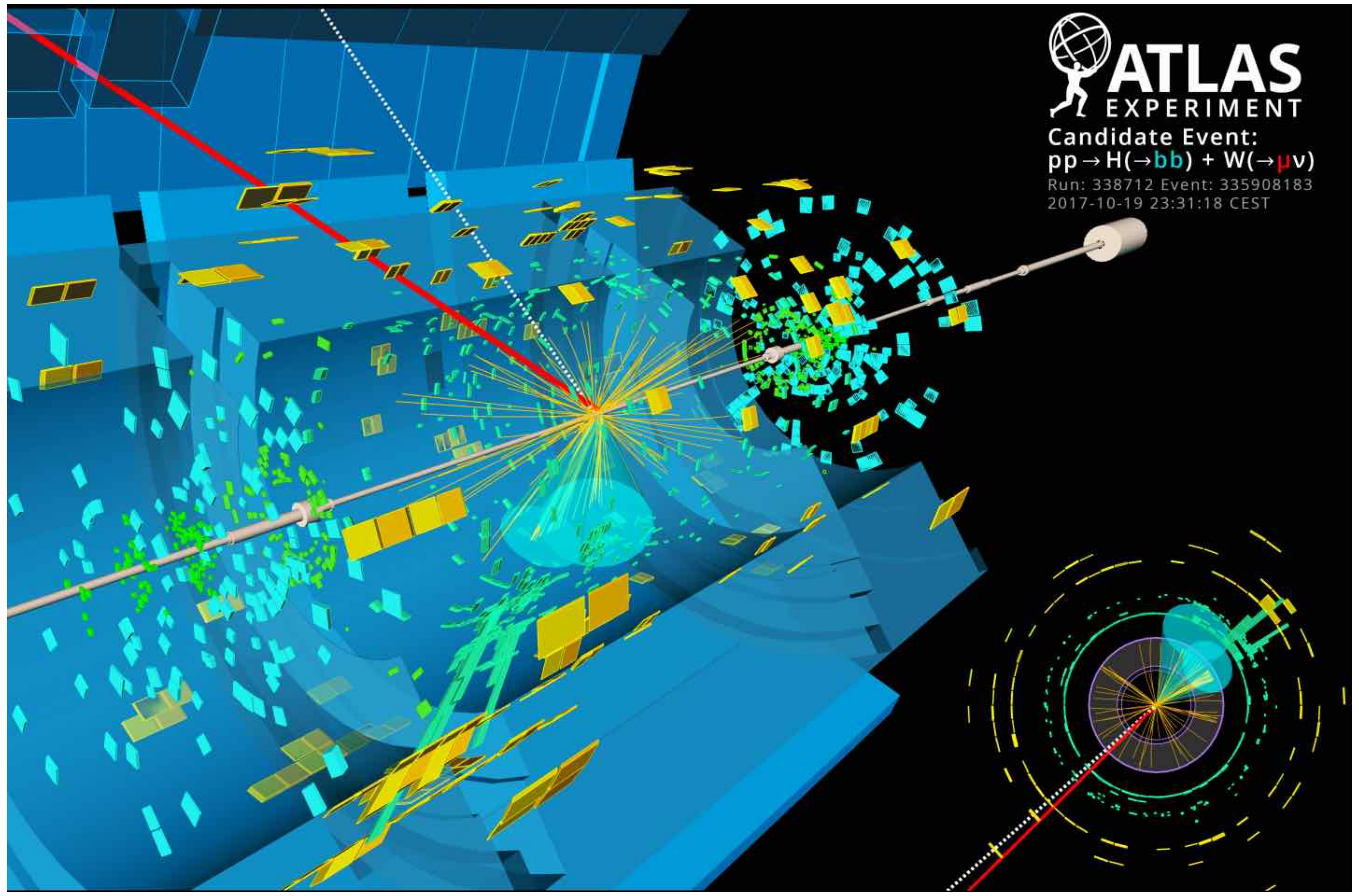
AI4EIC

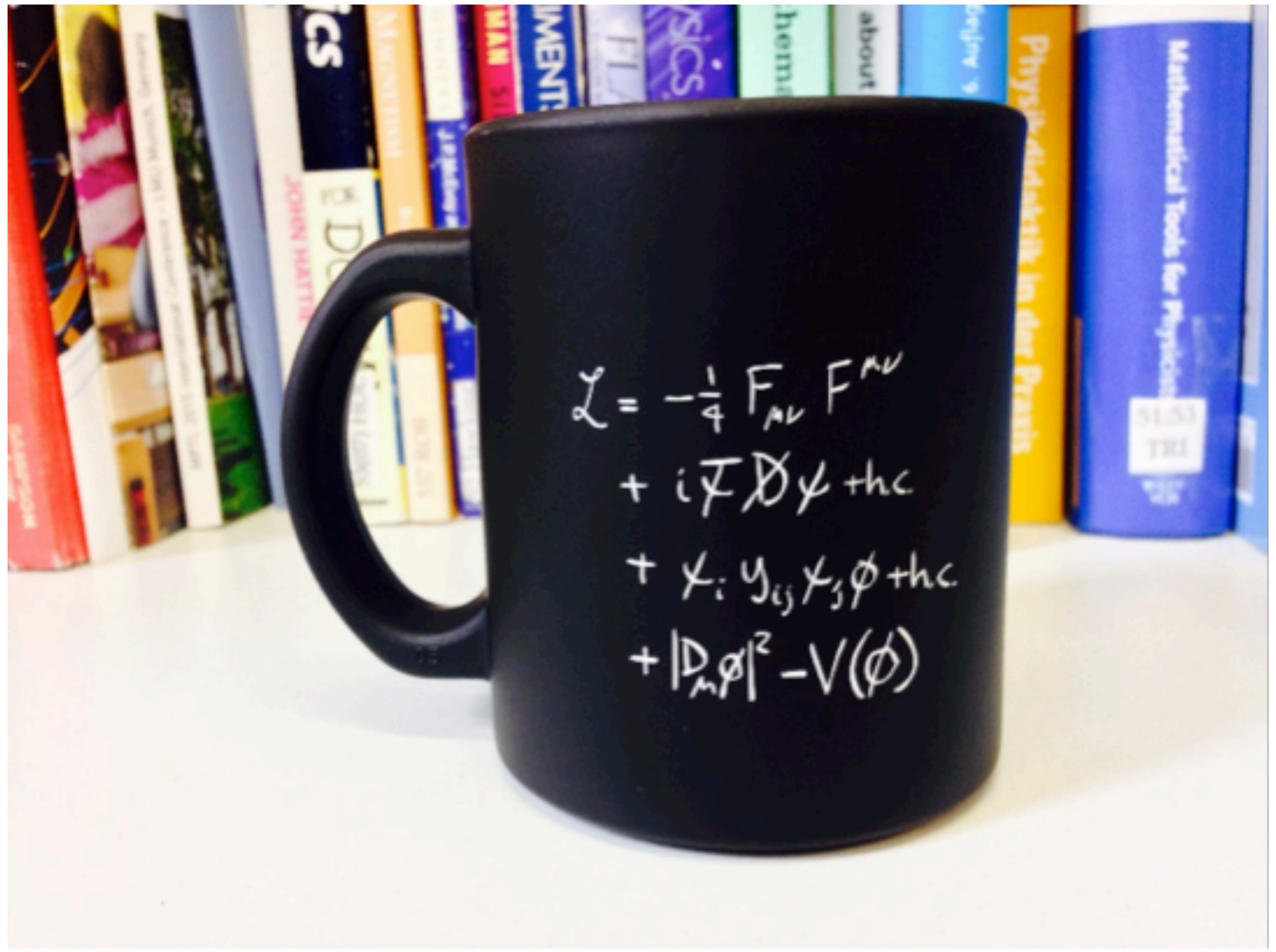
Anja Butter, LPNHE CNRS

LPNHE PARIS

Based on 2210.00019, 2105.09923, 2109.13243, 2006.06685, 1911.09107

A biased view on LHC physics





Setting
- Large Hadron Collider at CERN
- Proton collisions at 13 TeV
- **Huge** dataset ~1Pb/s before trigger selection

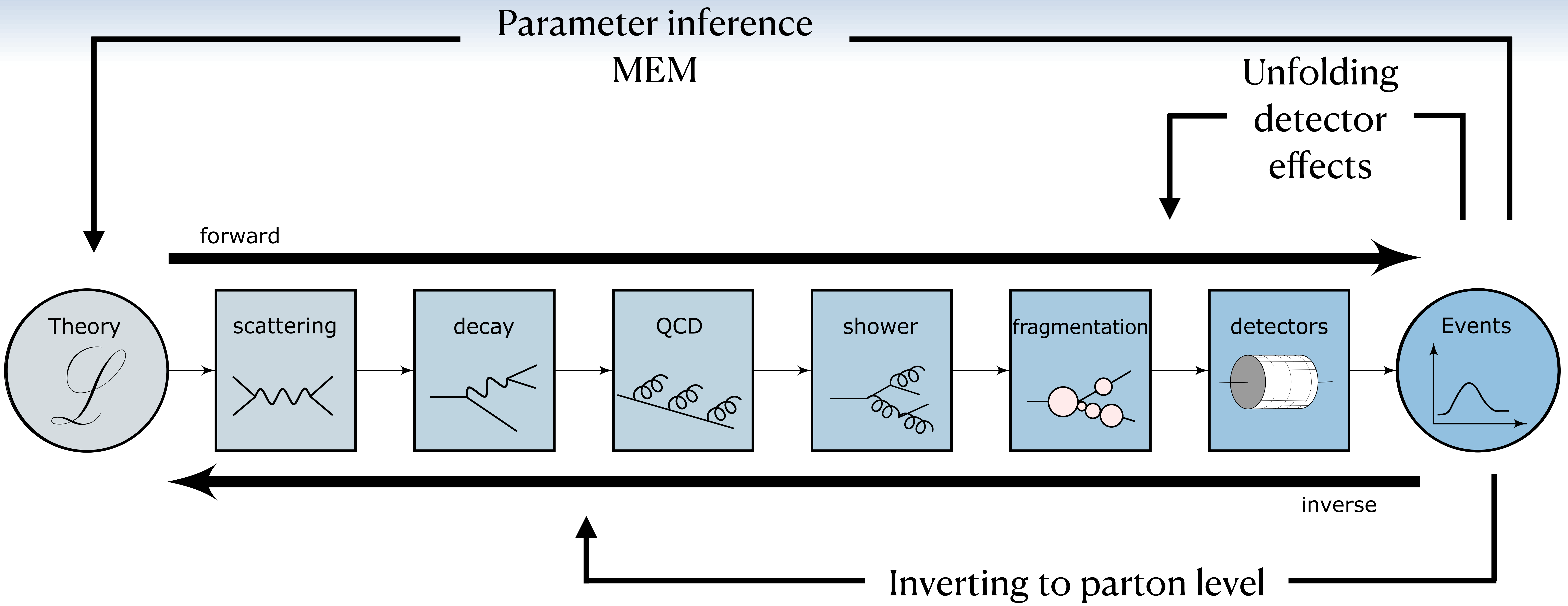
Goal
- Understand full dataset from **1st principles**
- Precision measurements of the SM
- Find signs of new physics (eg dark matter)

How to get from data to theory with differential cross section measurements?

Inverting the simulation chain

Unfolding

Inverting the simulation chain



Why?

- ✦ Compare unfolded data to theory
- ✦ Compare experiments with each other
- ✦ Enable use by theorists → global analyses

Requirements

- ☐ High dimensional
- ☐ Bin independent
- ☐ Statistically well defined

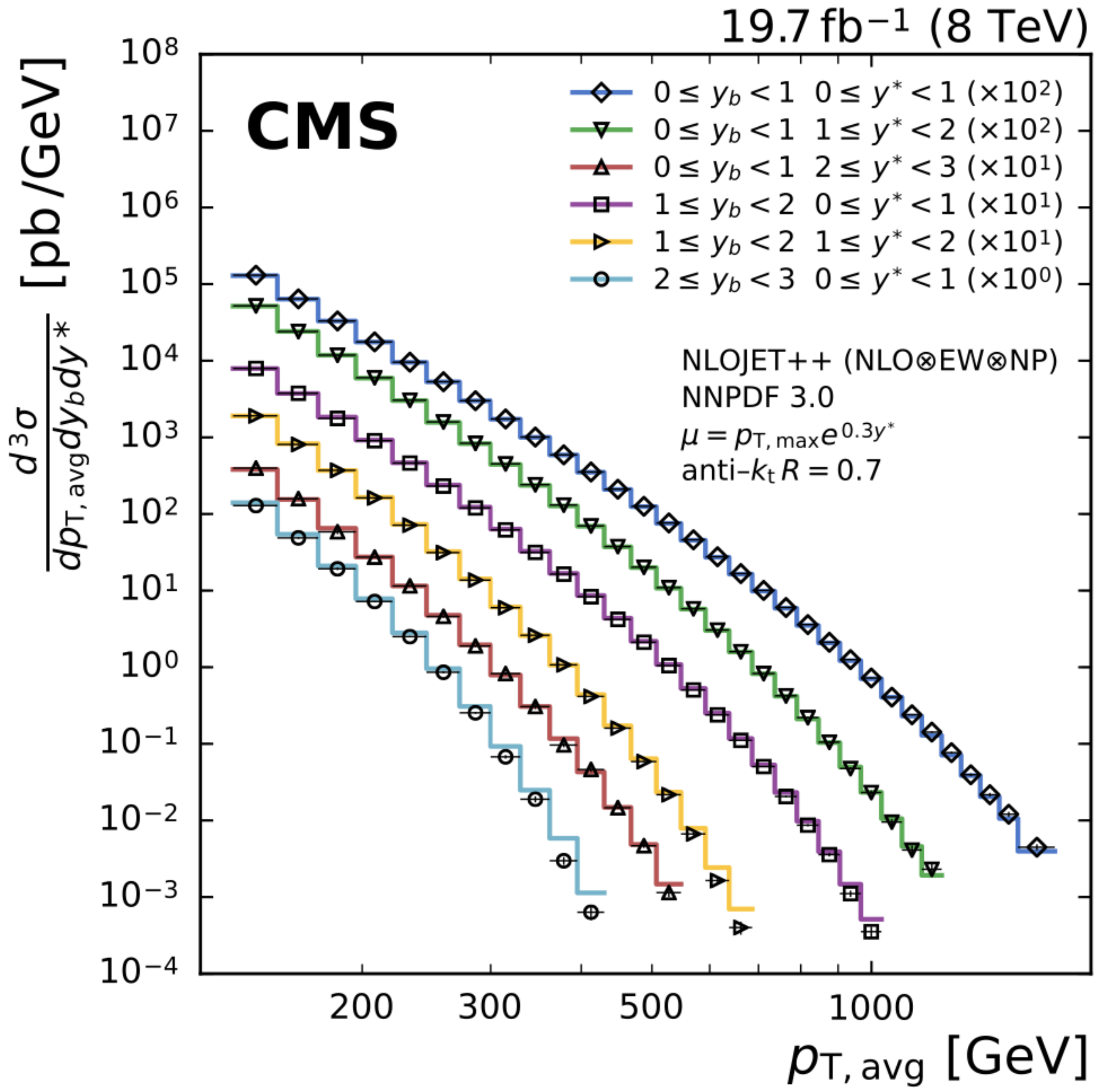
Why unbinned?

- **Inference-aware binning**
  - Currently: A priori bin choice
    → not optimal for post hoc analyses

- **Derivative measurements**
  - Given measurement of $x_0, \ldots, x_n$
    → enable post-hoc measurement of $f(x_0, \ldots, x_n)$

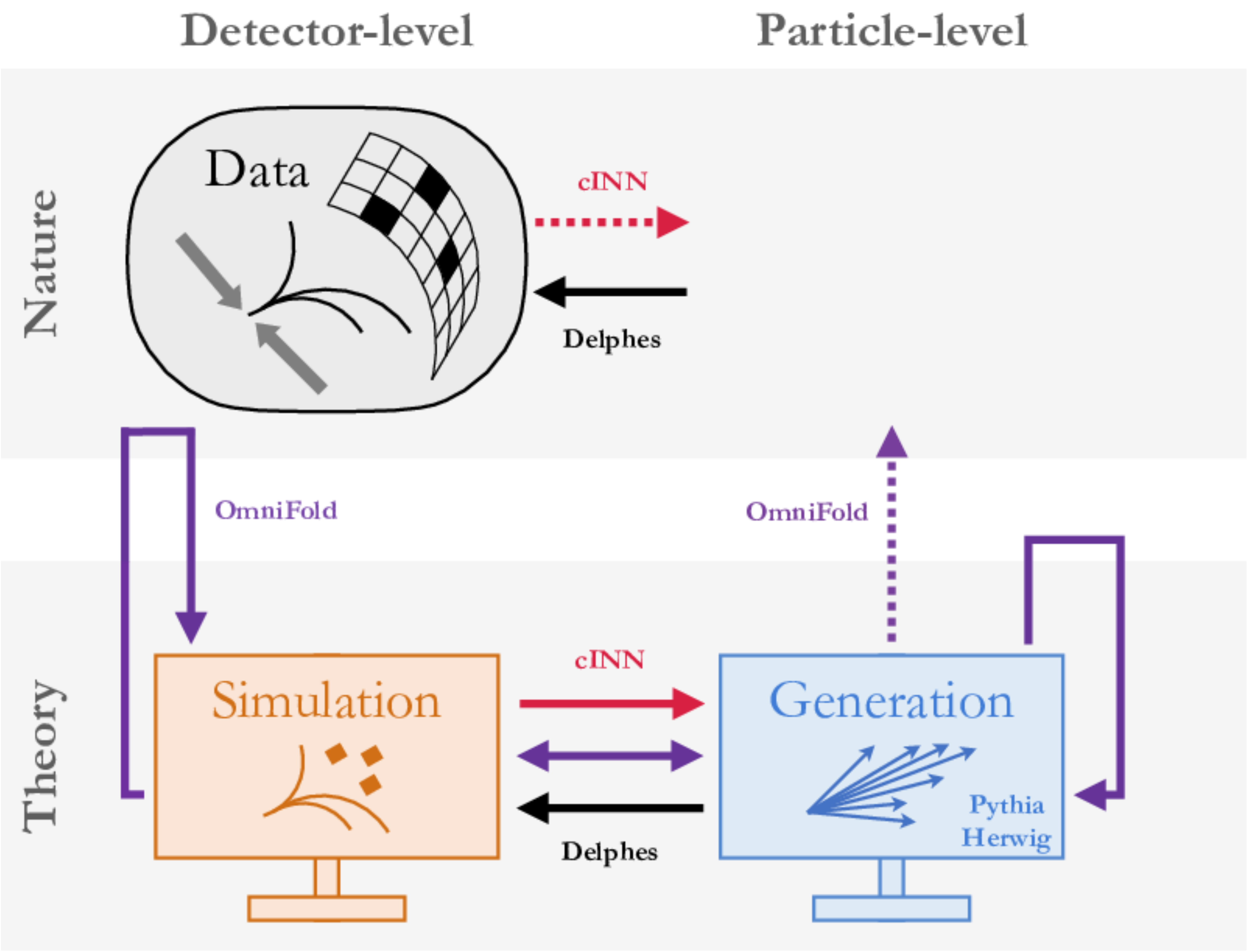
- Extension to **higher dimensions**
  - Understand correlations [jet p_T vs η]
  - ML-based unfolding extendable to many dimensions

- Example: Jet measurements
  - Currently up to 3D measurements
    → CMS dijet: $p_T^{\text{average}}, y*, y_b$
  - Challenging to go higher



ML unfolding methods

High-dimensional & bin independent



M. Arratia et al. [2109.13243]

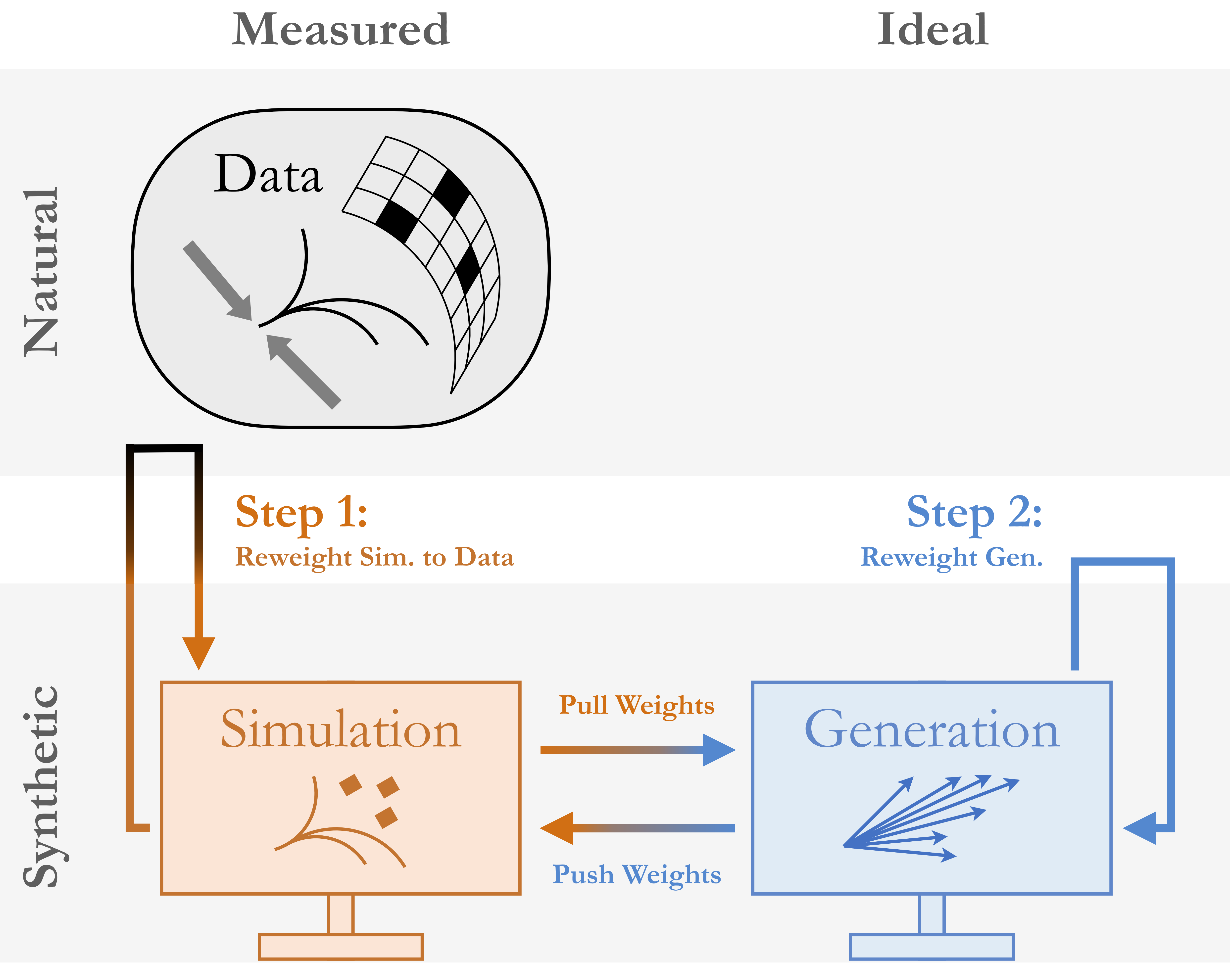
Classifier based aproach
Output: reweighted distribution of MC events

Density based approach
Output: probability density per unfolded event

VAE alternative:
OTUS by J. N. Howard et al.[2101.08944]

GAN+classifier:
MLEG by Y.Alanazi, et al. [2008.03151]

Classifier based approach



M. Arratia et al. [2109.13243]

Classifier based aproach
Output: reweighted distribution of MC events

Density based approach
Output: probability density per unfolded event

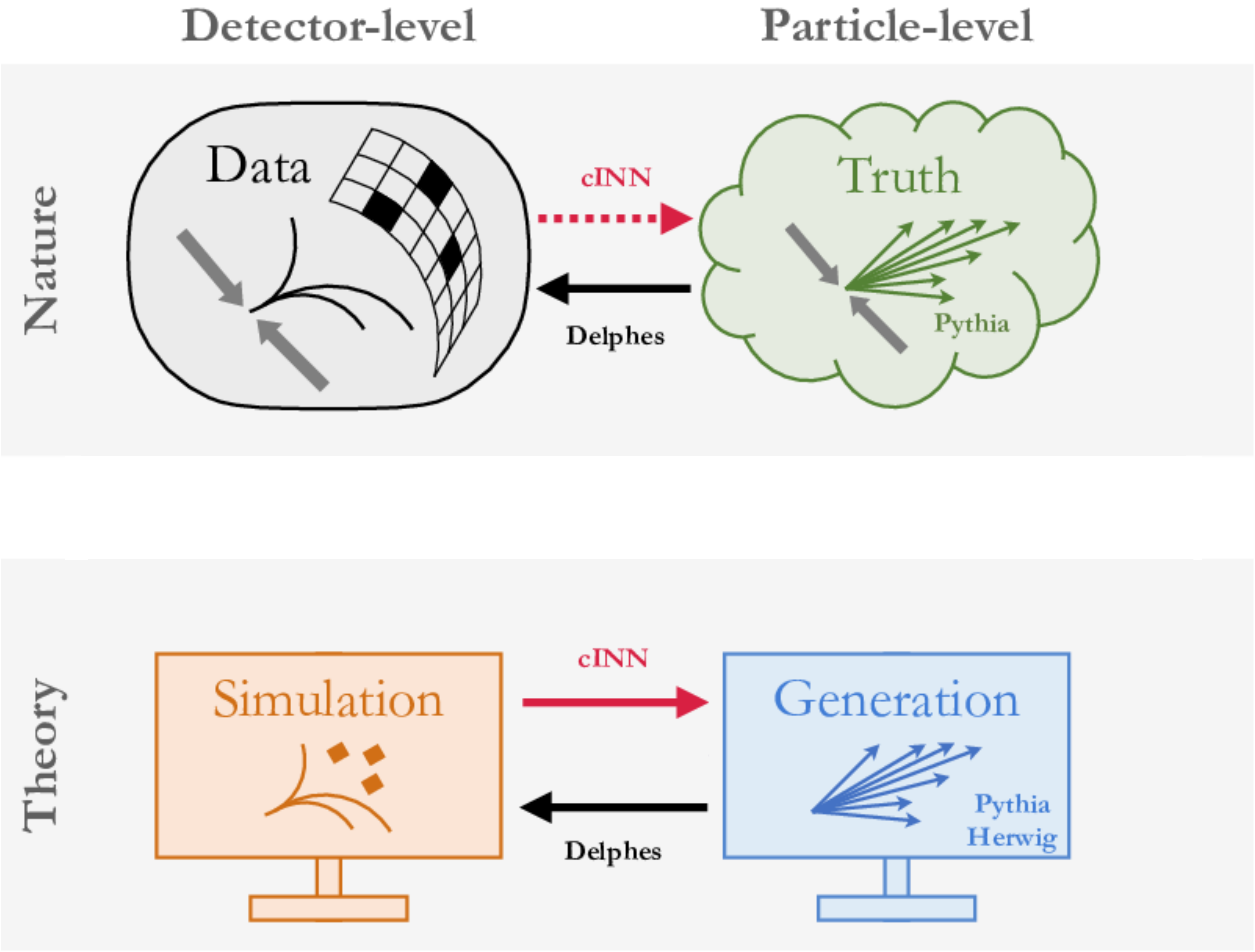
VAE alternative:
OTUS by J. N. Howard et al.[2101.08944]

GAN+classifier:
MLEG by Y.Alanazi, et al. [2008.03151]

For more details check out the tutorial!

ML unfolding methods

High-dimensional & bin independent



M. Arratia et al. [2109.13243]

Classifier based aproach
Output: reweighted distribution of MC events

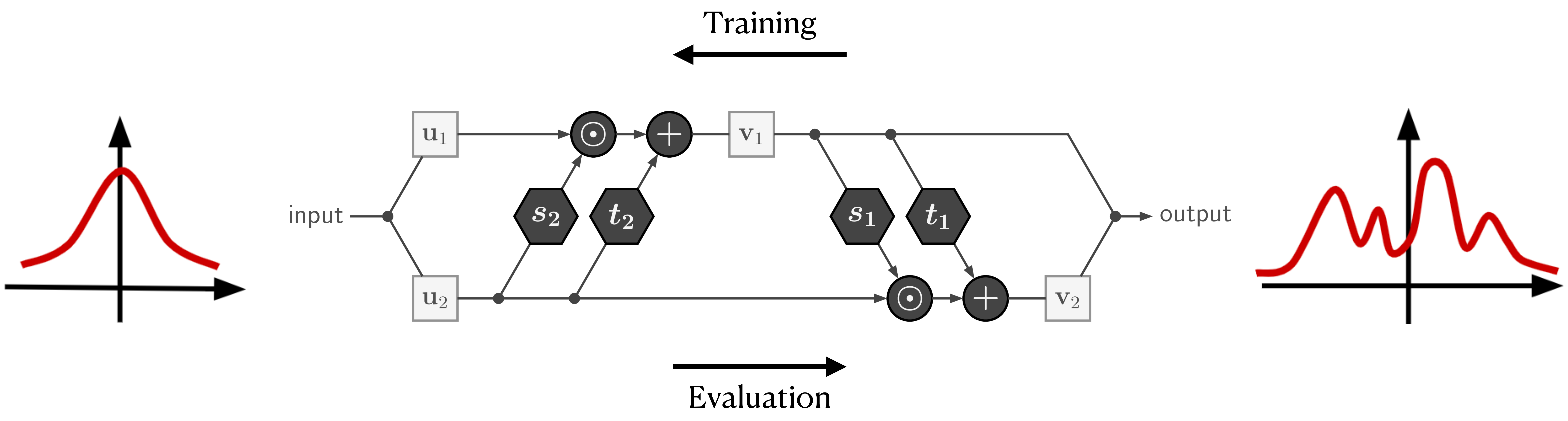
Density based approach
Output: probability density per unfolded event

VAE alternative:
OTUS by J. N. Howard et al.[2101.08944]

GAN+classifier:
MLEG by Y.Alanazi, et al. [2008.03151]

Normalizing flows

Invertible networks for complex transformations



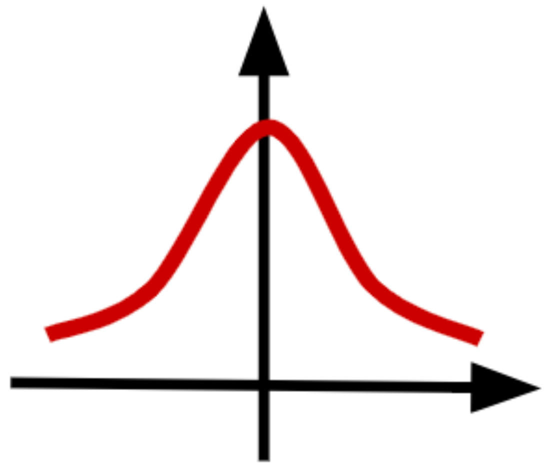
+ Bijective mapping
+ Tractable Jacobian → $p_x(x) = p_z(z) \cdot J_{NN}$
+ Fast evaluation in both direction

Training on samples x
→ Maximize the log-likelihood

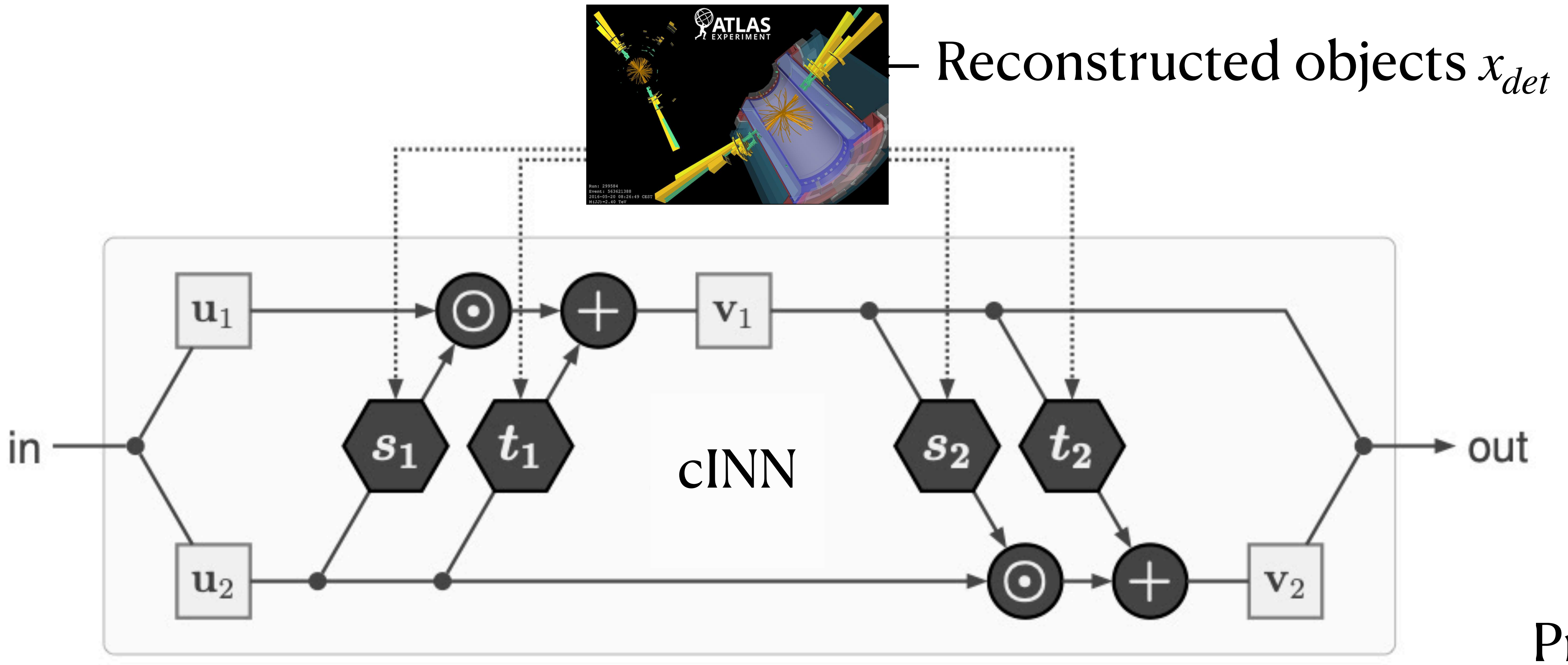
$$\begin{aligned}\mathscr{L} &= \log p(\theta\,|\,x) \\ &= \log p(z\,|\,\theta) + \log J_{NN} + p(\theta)\end{aligned}$$

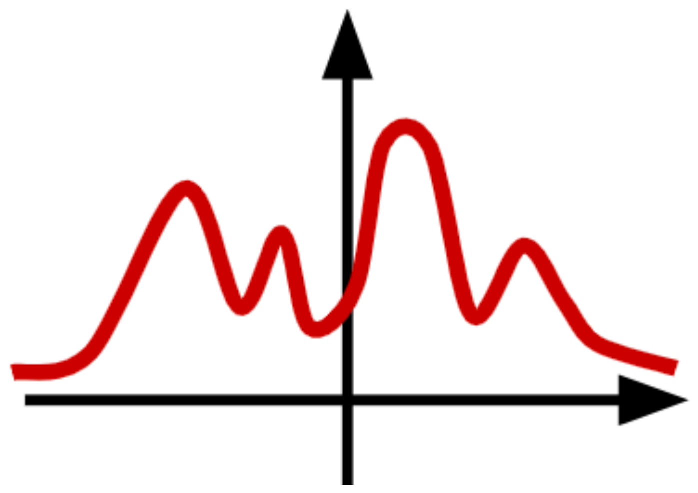
cINN unfolding

Given a reconstructed event:
What is the probability distribution at particle level?



Normal distribution
$z \sim \mathcal{N}$





Probability distribution
$x \sim P_{\text{parton}}$

Training

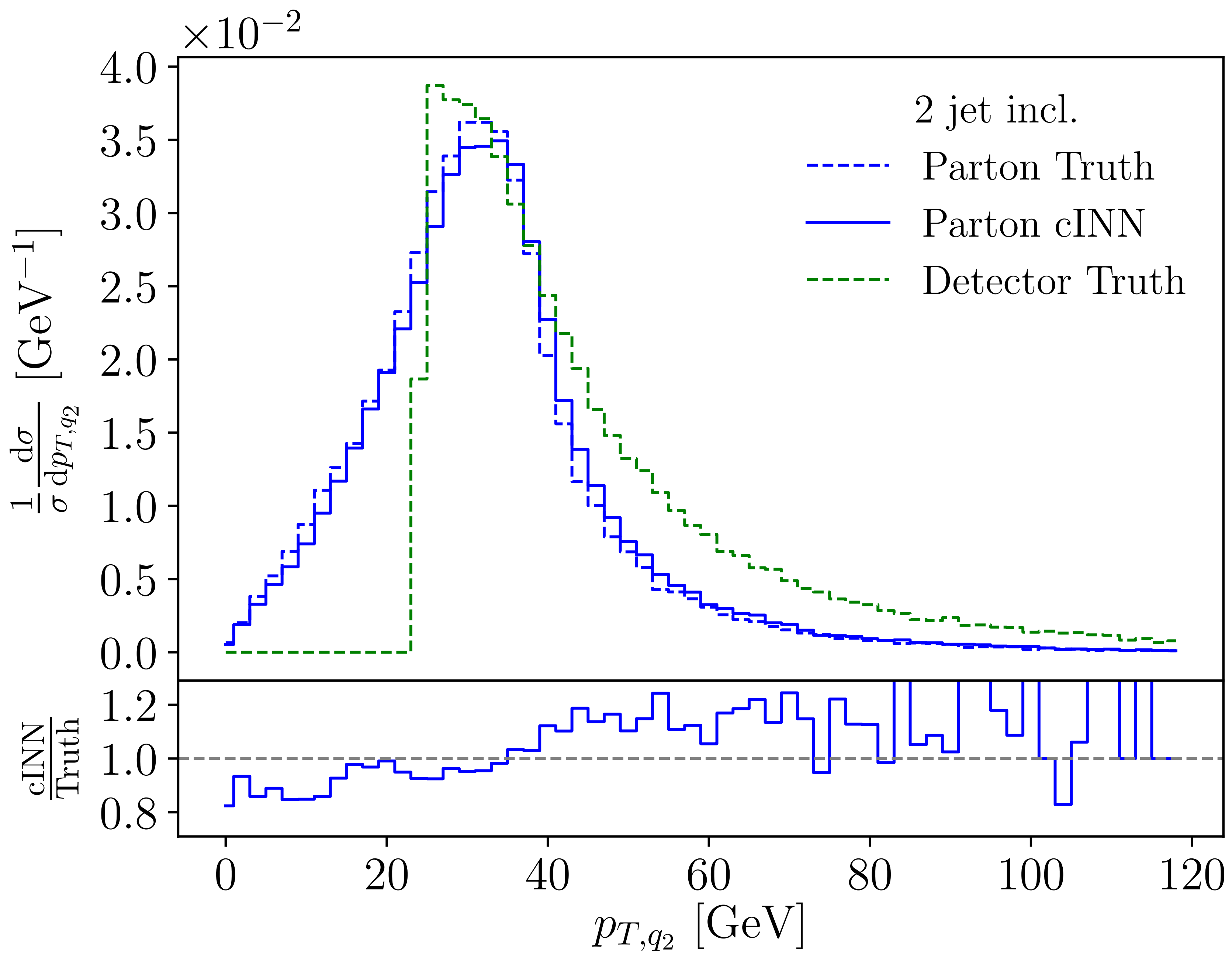
Unfolding

$$\begin{aligned} \mathscr{L} &= \log p(\theta \,|\, x_{part}, x_{det}) \\ &= \log p(z(x) \,|\, \theta) + \log J_{NN} + p(\theta) \end{aligned}$$

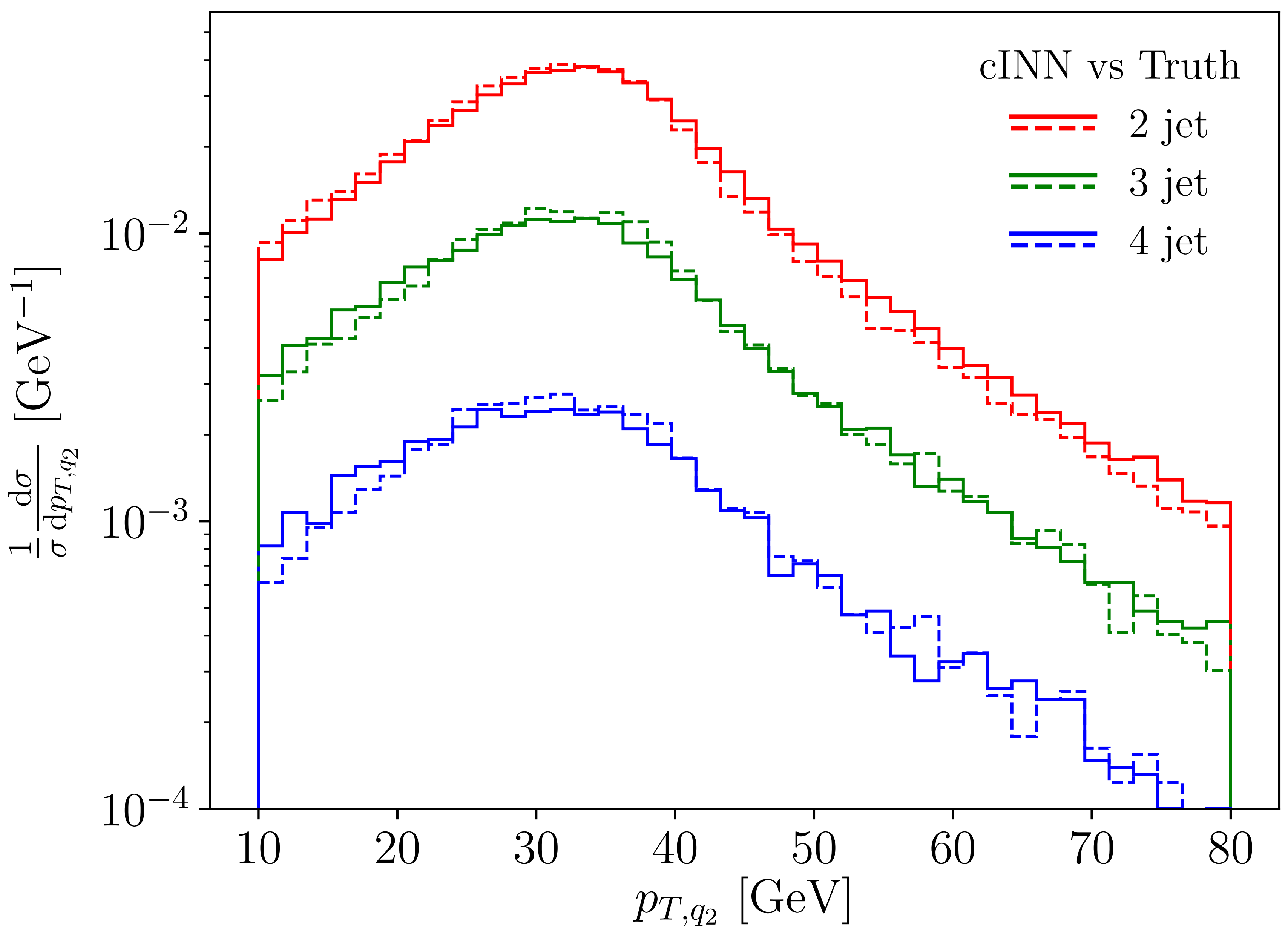
Inverting inclusive distributions

$pp > WZ > q\bar{q}l^+l^-$+ ISR → 2/3/4 jet events

Training on inclusive dataset



Evaluate exclusive 2/3/4 jet events

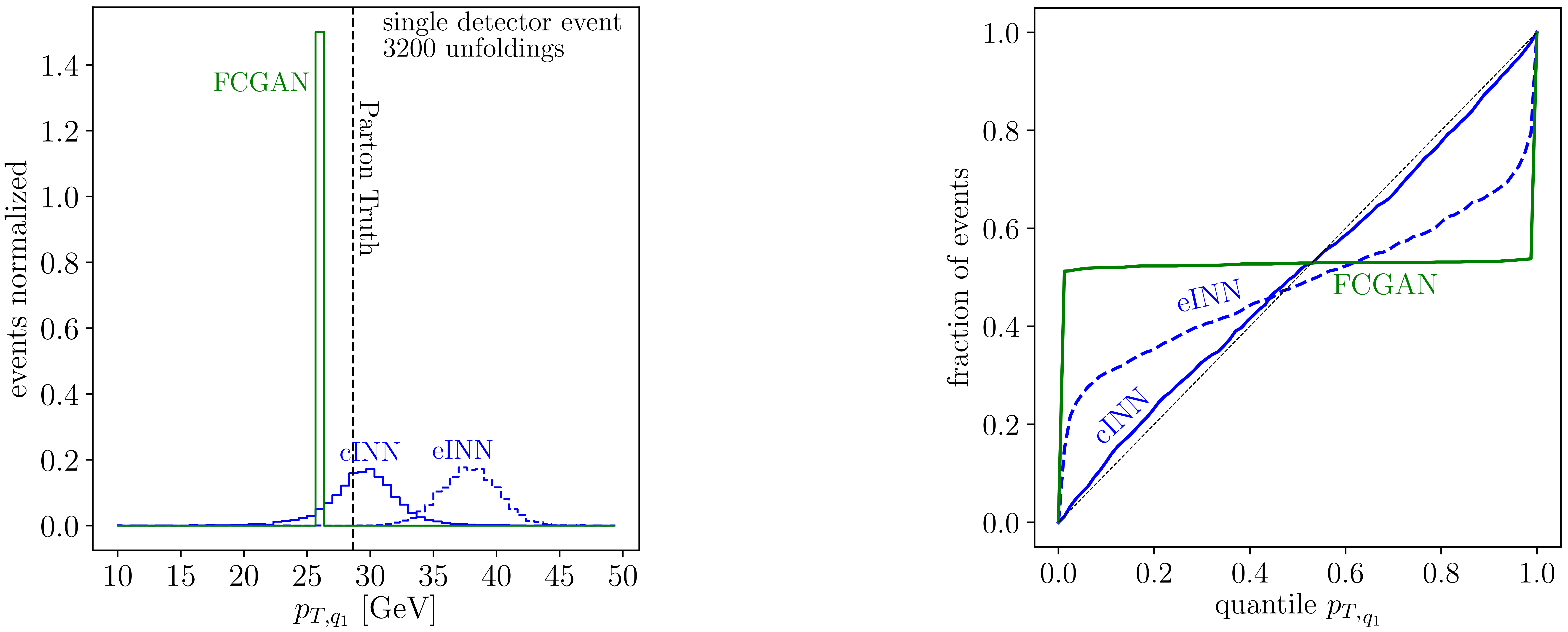


- ☑ High-dimensional
- ☑ Bin-independent
- ☐ Statistically well defined ?

M. Bellagente et al. [2006.06685]

Event-wise unfolding

No deterministic mapping!
Check calibration of probability density for individual event unfolding



- High-dimensional
- Bin-independent
- Statistically well defined

M. Bellagente et al. [2006.06685]

General challenges along the way

- Unbinned acceptance & background subtraction
  - Classifier based: reweighting with negative weights
  - Density based: Learn difference between full and background distribution

- Large local weight fluctuations
  - Assign constant weight within local phase space patch
  - Store local average weight squared $\tilde{w} = \langle w^2 | x \sim x_i \rangle$

- Uncertainties
  - Vary aspect of simulation and repeat unfolding procedure
  - Closure tests to estimate bias of pprocedure

    $\rightarrow$ Nothing conceptually new wrt. standard unfolding

    $\rightarrow$ Control challenging in high-dimensional approach

How to published unfolded data?

1. Submission file
   - Yaml [HEPData]
2. Data file $(1 + n_{\text{uncert.}}) \times (n_{\text{dim}} + 2) \times n_{\text{samples}}$

Example: density-based approach (no event-weights)
$\vec{x}_1, \vec{x}_2, ..., \vec{x}_N$
$1, 1, ..., 1$
$1, 1, ..., 1$

Example: classifier-based approach
$\vec{x}_1, \vec{x}_2, ..., \vec{x}_N$
$w_1, w_2, \ldots, w_N$
$w_1^2, w_2^2, \ldots, w_N^2$

3. Optional (but recommended)
   - Network architecture (ONNX) + weights

Beyond unfolding: Enabling the MEM

2210.00019

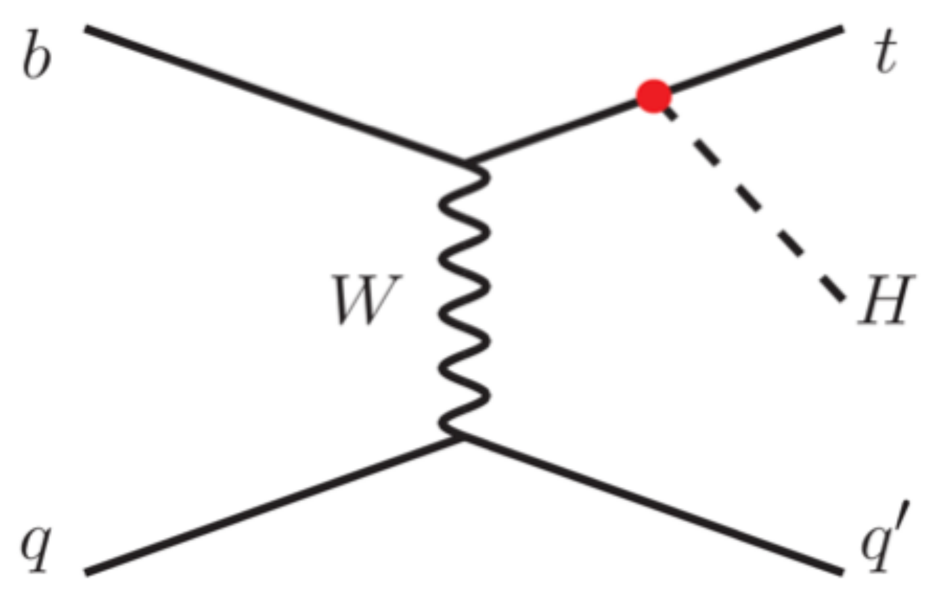
Matrix element method is based on untractable likelihood

$$p(x_{\text{reco}}|\alpha) = \int \mathrm{d}x_{\text{hard}}\; \underbrace{p(x_{\text{hard}}|\alpha)}_{\text{diff. CS}}\; \underbrace{p(x_{\text{reco}}|x_{\text{hard}},\alpha)}_{\text{estimate with network}}$$

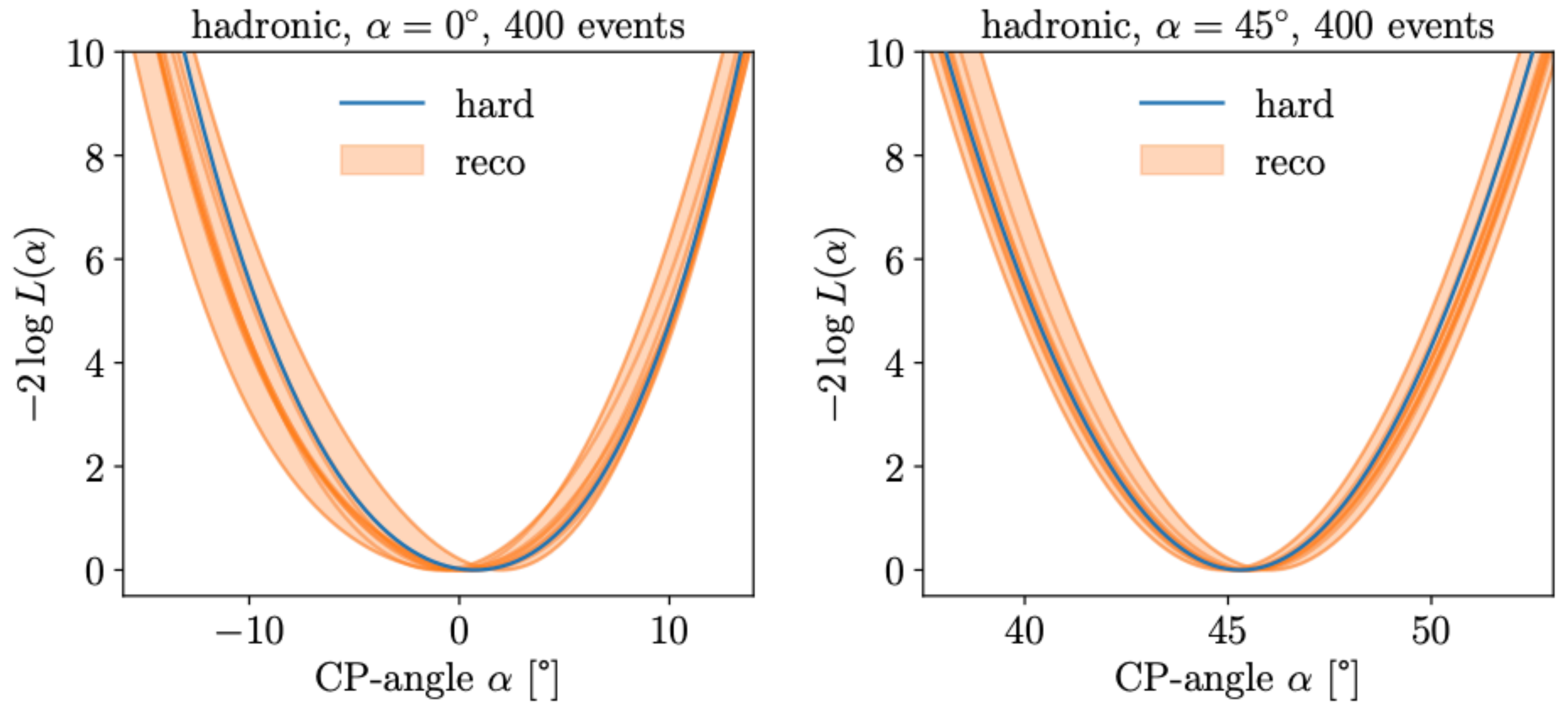
Problem: integration over full phase space of the hard scattering

Solution: Use unfolding cINN to sample x_{hard}

$$p(x_{\text{reco}}|\alpha) = \left\langle \frac{1}{q(x_{\text{hard}})}\, p(x_{\text{hard}}|\alpha)\, p(x_{\text{reco}}|x_{\text{hard}},\alpha) \right\rangle_{x_{\text{hard}}\sim q(x_{\text{hard}})}$$



Single Higgs production
with anomalous non-CP conserving Higgs coupling



Outlook

Unbinned measurements allow for higher **flexibility**

→ analysis dependent bin optimization
→ derivative measurements
→ high dimensional

Several possible solutions, eg. **classifier or density based**

First proposal how to publish unbinned measurements

Format: event + weights + local weight uncertainty

github: ramonpeter/UnbinnedMeasurements

Future: Better unfolding enables better analyses like **MEM**