Generating parton-level events from CMS reconstructed events with Conditional Normalizing Flows

Antonio Petre on behalf of the CMS collaboration

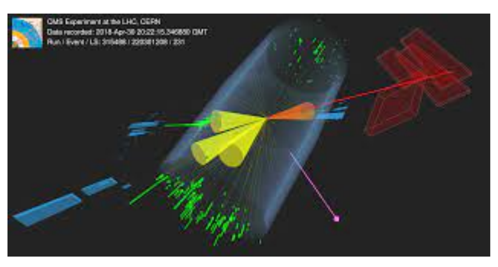
CMS

ETH zürich

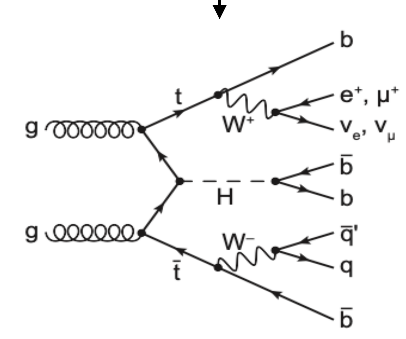
Matrix Element Method (MEM)

Matrix element method estimates the probability of a single reconstructed event $\vec{Y}$ to be generated by a physical process defined by θ parameters:

$$\mathcal{P}(\vec{X}_{reco}|\theta) \propto \int_{\phi} d\vec{X}_{hard}\, |M(\vec{X}_{hard}|\theta)|^2 \cdot Pdf \cdot W(\vec{X}_{reco}|\vec{X}_{hard})$$



$= \vec{X}_{reco}$ $\quad\quad \vec{X}_{hard} =$



Pros & Cons:

- ✓ It can be used for hypothesis testing or parameter estimation
- ✓ Maximizes the amount of theoretical information for the discriminator
- ✓ It is not bound to a specific process
- x Integral computation is very CPU demanding due to jet-parton matching (combinatorial problem)
- x Many approximations used to speedup the computation e.g. jet-parton alignment

Previous machine learning (ML) method for solving the problem:

- • Basic neural network architecture with 4-momenta of the reco-objects as inputs
- ✓ Fast evaluation and inference
- x Needs pre-computed MEM values to train the model (time expensive)
- x Embeds the approximations from the conventional MEM computation

Dataset and Run II Performance

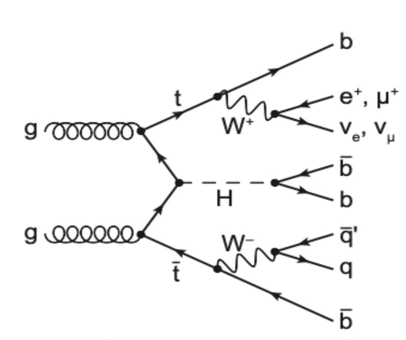
Process used: **single-lepton channel in ttH(→bb) with an additional radiation**

Events details:

- Pileup profile of LHC Run II (~30-50 simultaneous pp collisions)
- Full CMS detector simulation, including standard RUN II reconstruction

Data selection:

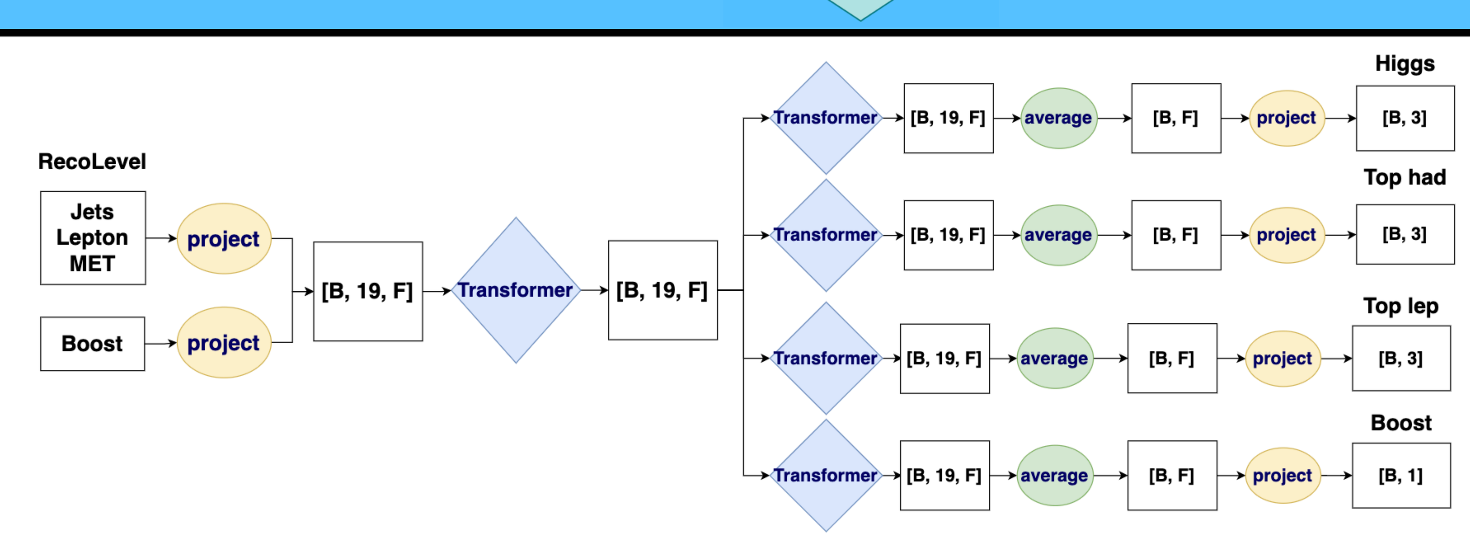
- At least 4 jets with $p_T > 30$ GeV and $|\eta| < 2.4$
- At least 3 jets identified as originating from a b-quark
- One prompt reconstructed lepton with $p_T > 30$ GeV
- MET > 20 GeV

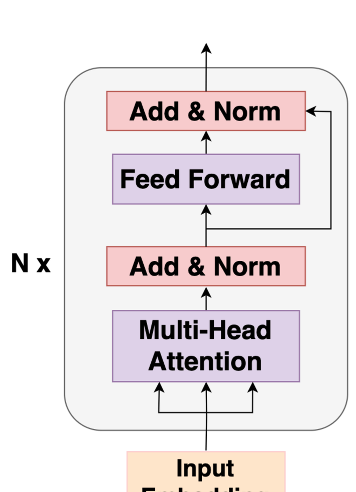


Run II ttH analysis: **MEM used for discriminating between signal and background**

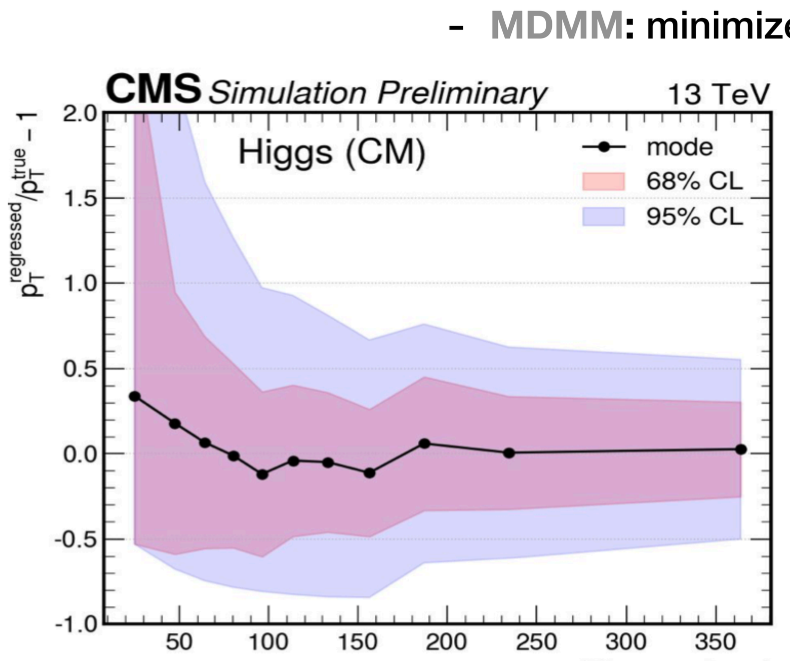
Computation performance: **~ 1 min/event → speedup needed**
normalizing flow is a good candidate

Conditional Transformer

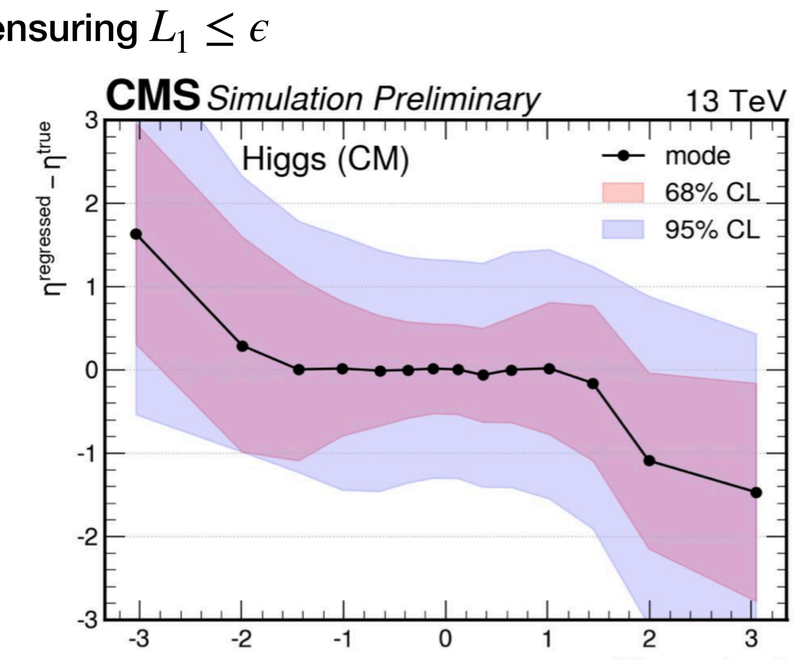




- The main building block is the **Transformer Encoder**
- The training data → partons or reco-objects in the laboratory frame
- The model was pretrained using modified differential multiplier method (MDMM):
  - Main loss L_0 is the Huber loss for partons and boost pz
    - Combines mean squared error for small errors and mean absolute error for large errors (less sensitive to outliers)
  - Second loss L_1 is the maximum mean discrepancy (MMD) loss to keep distributions coherent
    - Measures the distances between two probability distributions by comparing the kernel-based representation of their features
  - MDMM: minimize L_0 ensuring $L_1 \leq \epsilon$



Bias mode of the regression for Higgs p_T



Bias mode of the regression for Higgs η

References

New Method & Normalizing Flows

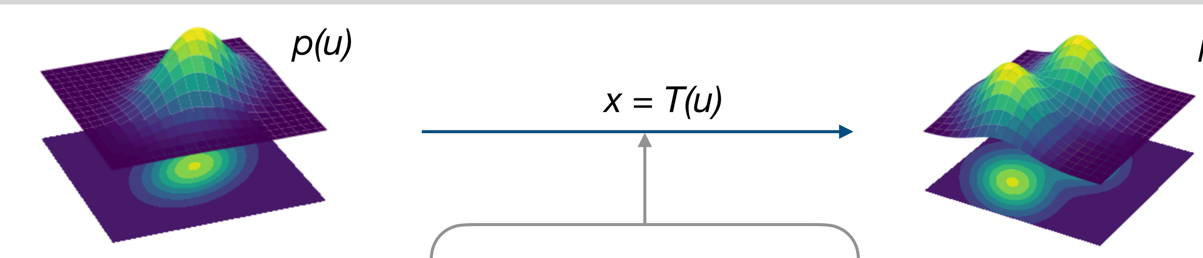
Our goal is to model the conditional probability of parton-level events given a reconstructed event using generative machine learning architectures, more specifically **normalizing flows:**

$$\int_{\phi} d\vec{X}_{hard}\; |M(\vec{X}_{hard}|\theta)|^2 \cdot Pdf \cdot W(\vec{X}_{reco}|\vec{X}_{hard})$$

Use importance sampling: $\vec{X}_{hard} \sim \mathcal{P}(\vec{X}_{hard}|\vec{X}_{reco},\theta)$

$\mathcal{P}(\vec{X}_{hard}|\vec{X}_{reco},\theta)$ found using **normalizing flows**

Flow models: Machine-learned maps (transformations) between probability distributions



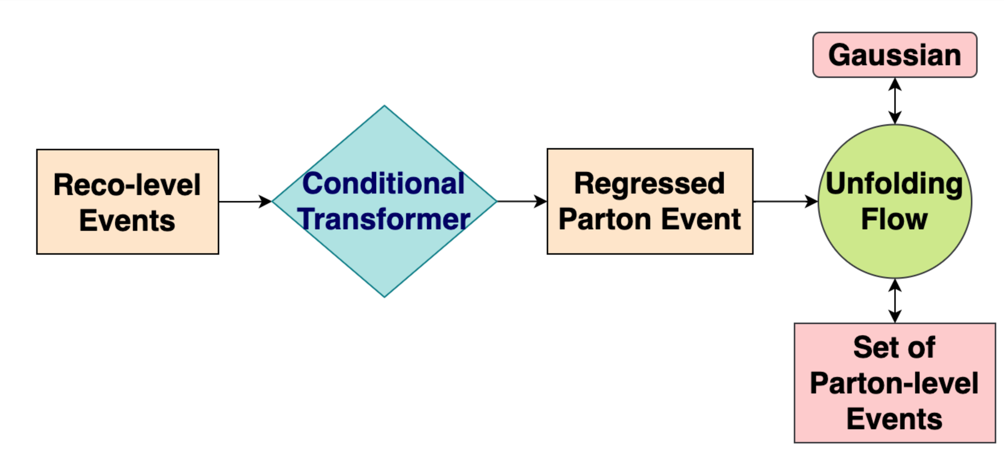
p(u) $\quad x = T(u) \quad$ p(x)

'Base' distribution
Easy-to-evaluate
Easy-to-sample

Flow transformation:
- Invertible
- Tractable Jacobian
- Many params. for optimization

'Target' distribution
Multi-modal

Our Strategy



Base distribution of the flow

Model $\mathcal{P}(\vec{X}|\vec{Y},\theta)$

Parton-level points used for MEM integral
(Target distribution)

Reco-level Events → Jets + Lepton + MET : p_T, η, ϕ, b-tag score, SPANET output
SPANET: ML architecture which predicts jet-parton assignment

Regressed Parton Event → Higgs + two tops + additional radiation : p_T, η, ϕ

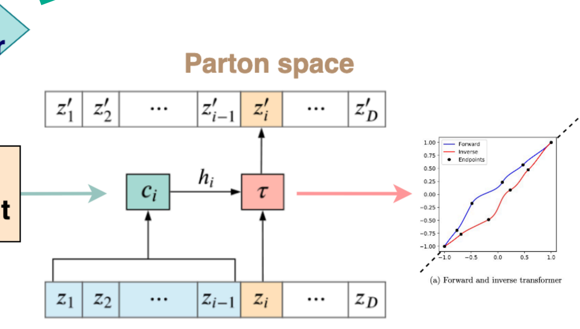
Conditional Transformer → Regress the **parton-level event** for a given **reco-level event**
Extracts a latent information vector which conditions the **Unfolding Flow**

Unfolding Flow →
- ✓ Generates plausible phase-space points compatible with reco-objects
- ✓ Reduces assumptions on partons' directions
- ✓ Handles events with out-of-acceptance final state objects and multiple jet multiplicities

Unfolding Flow

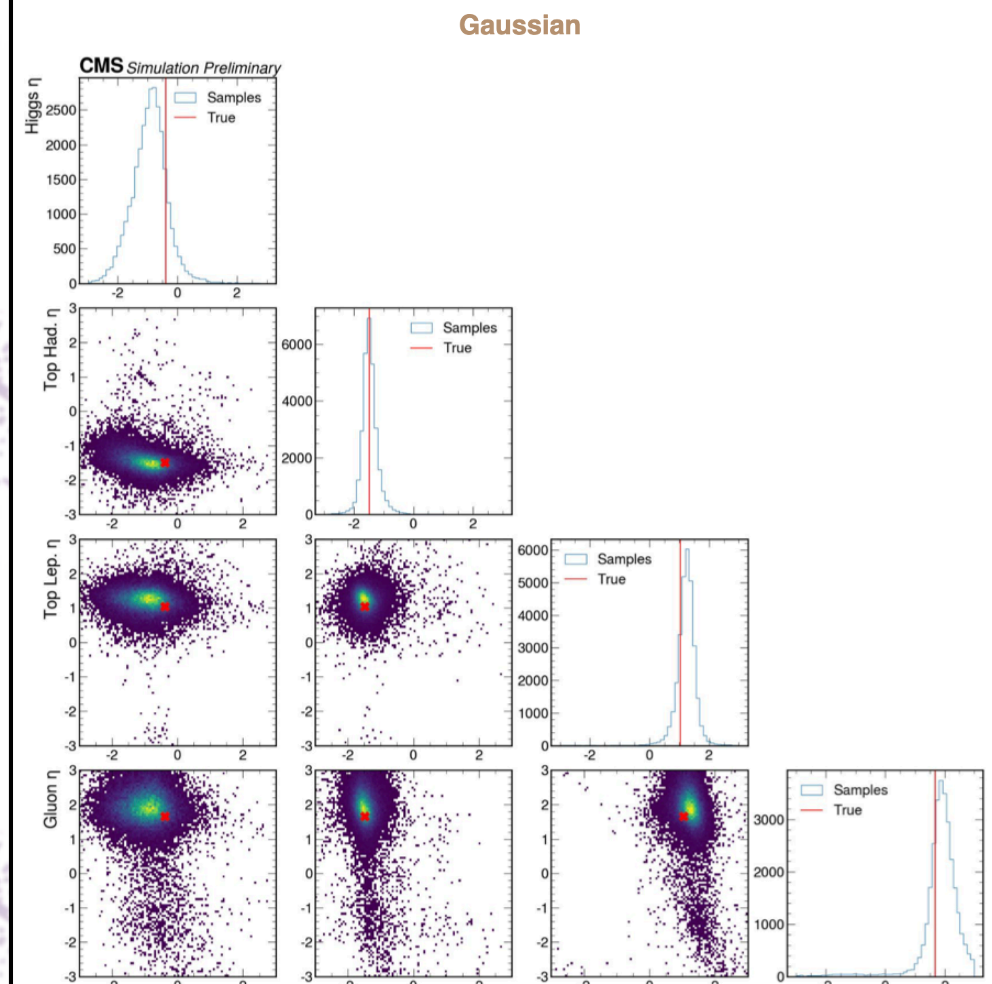
Implemented using **rational quadratic splines (RQS)** with **autoregressive blocks**

weights still updated during **Unfolding Flow** training

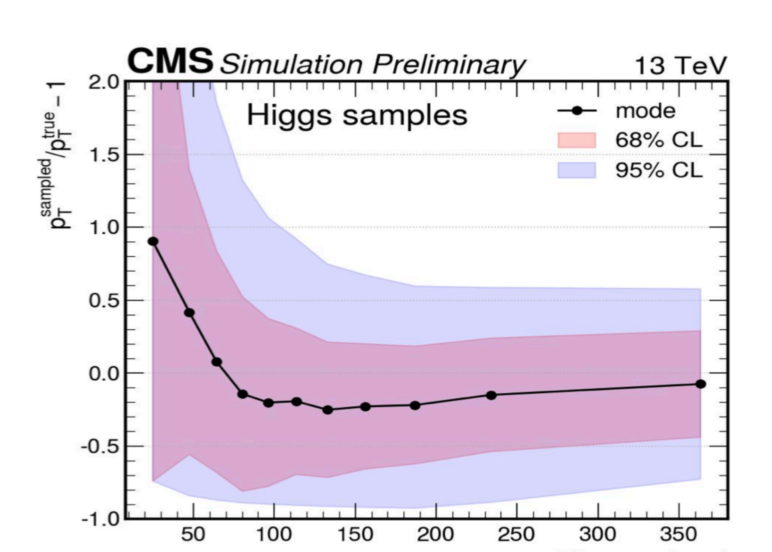


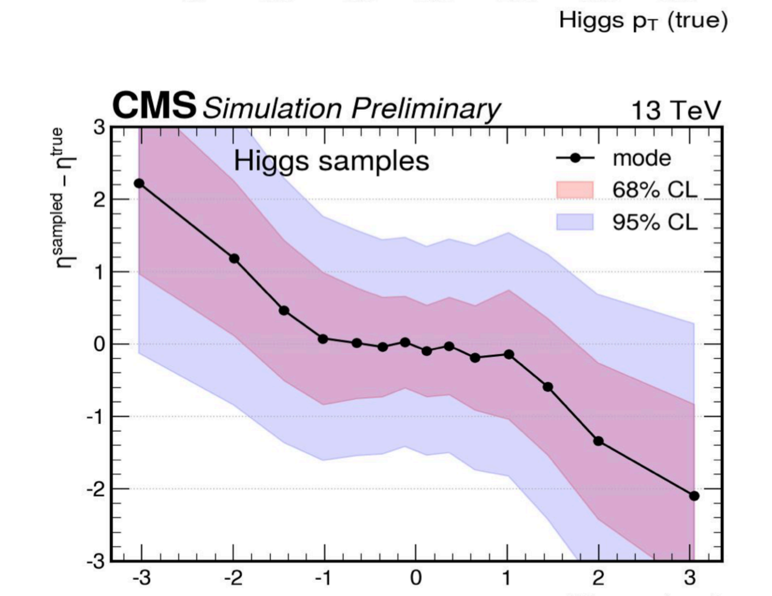
Training:

- **maximum likelihood**: evaluate the density of the true partons from the signal MC
- **sampling parton-level events** from the flow and comparing them with the target



Sample 20k parton-level events for one reco-event
Super fast (less than 1 second)





Sample 30 parton events for 1.5M reco-events
Check the quality of the sampled partons