Data driven background estimation in HEP using generative adversarial networks

arXiv:2212.03763

Artificial Intelligence and the Uncertainty challenge in Fundamental Physics 2023

Özgür Sahin
With Fabrice Couderc, **Victor Lohezic**, Julie Maclés

CEA Paris-Saclay

30 November 2023

Özgür Sahin

CEA Paris-Saclay / Irfu

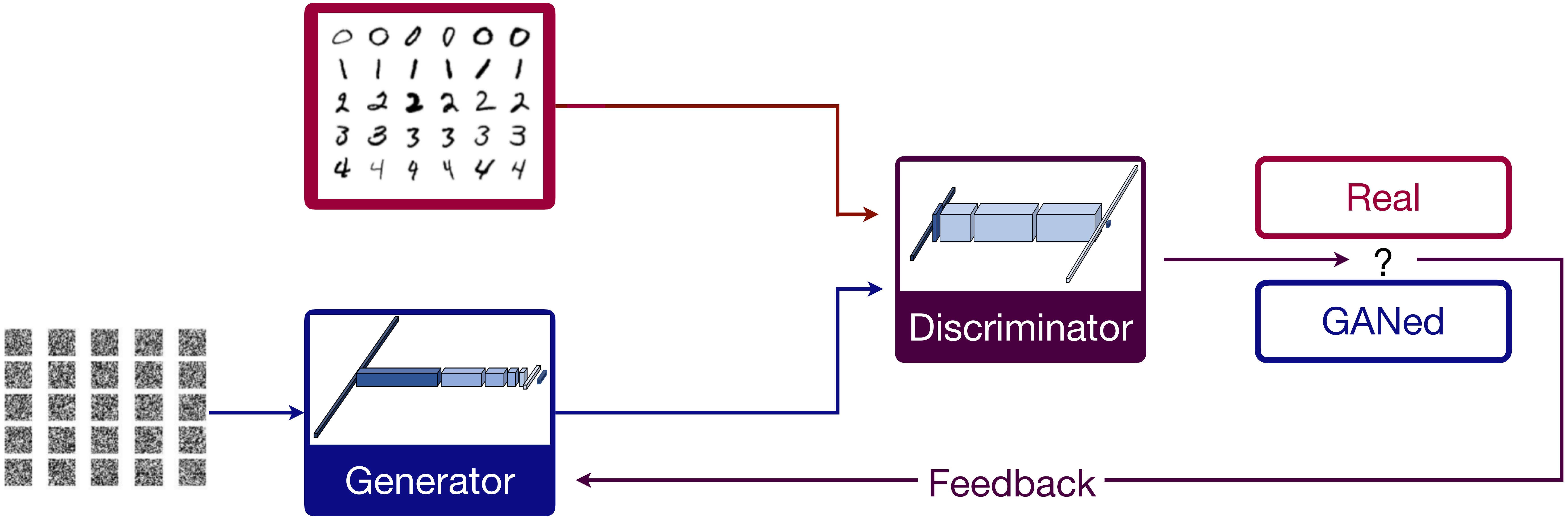
Monte Carlo simulations at the LHC

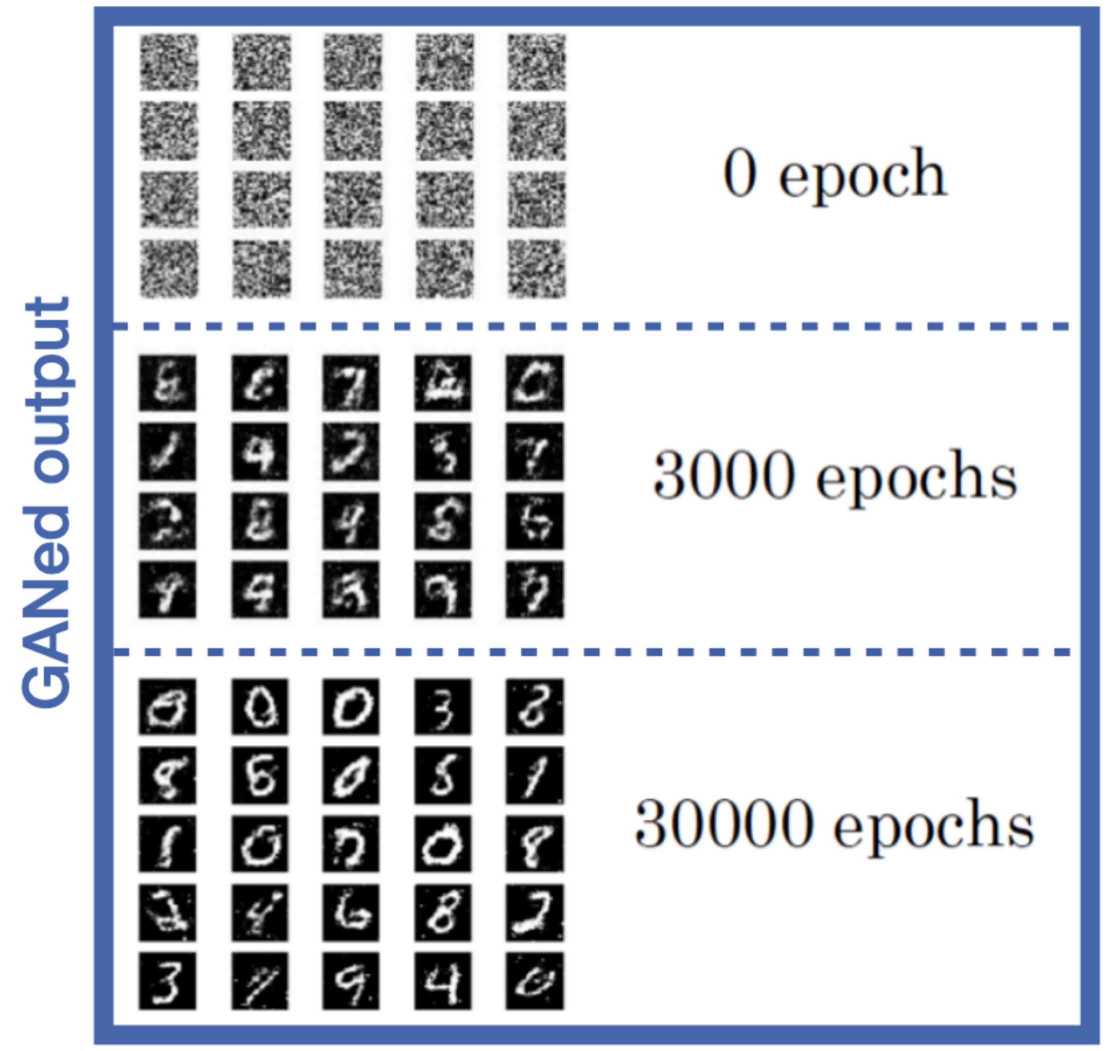
- Generating Monte Carlo (MC) simulations that accurately represent physics processes at the LHC experiments is challenging.
  - Specific scenarios, such as those involving misidentified objects, pose even greater challenges in terms of simulation accuracy.
  - The extensive use of ML algorithms necessitates a large number of training samples, which can significantly increase computational demands.
    - (News from Higgs 2023) Recently an ATLAS analysis demonstrated a notable enhancement due to improved simulation performance and statistics.

ML to rescue

- Analysis specific datasets can be generate using generative ML algorithms such as GANs, AE, normalizing flows
  - Once trained, they are very efficient to evaluate.
  - However, as most of them trained on MC samples, they inherent some of the concerns from the last slides.
- What if we use these ML techniques to obtain simulations (particularly for the background processes) in a data driven way.

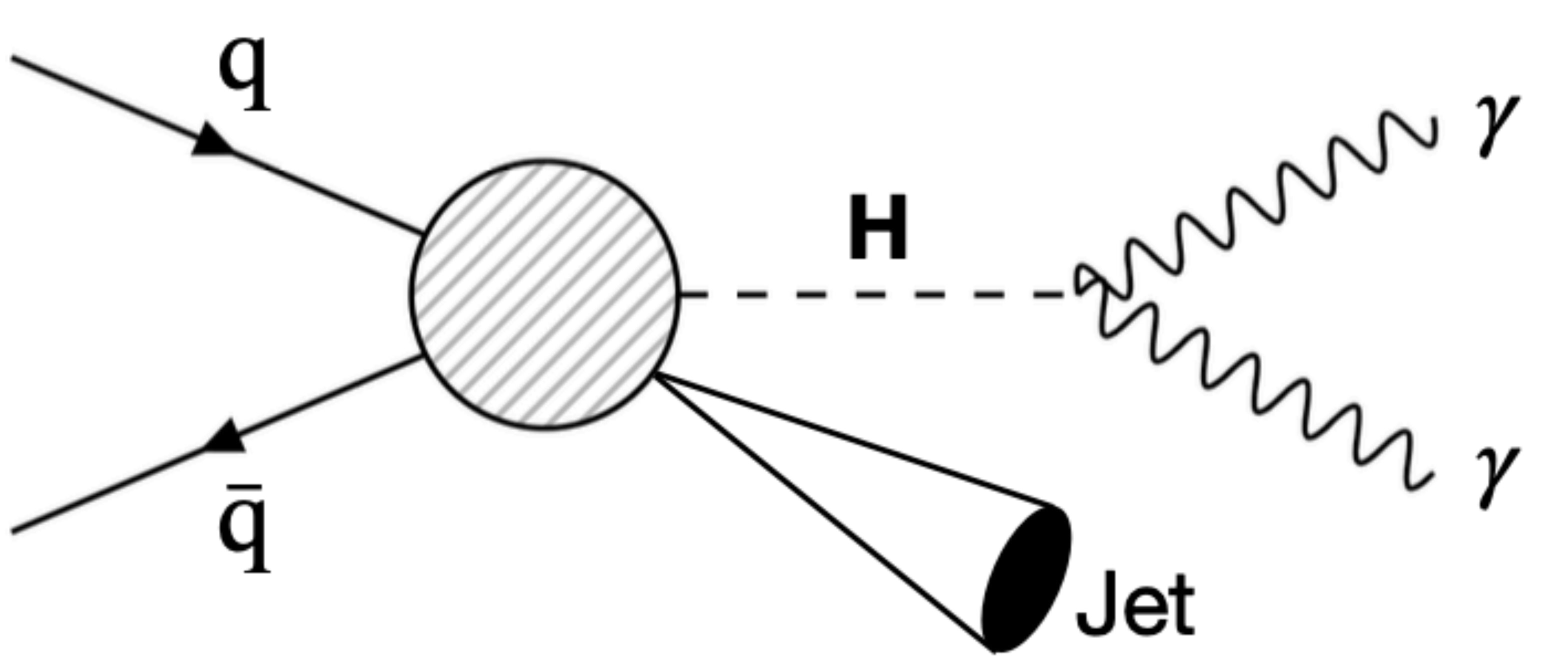
GAN

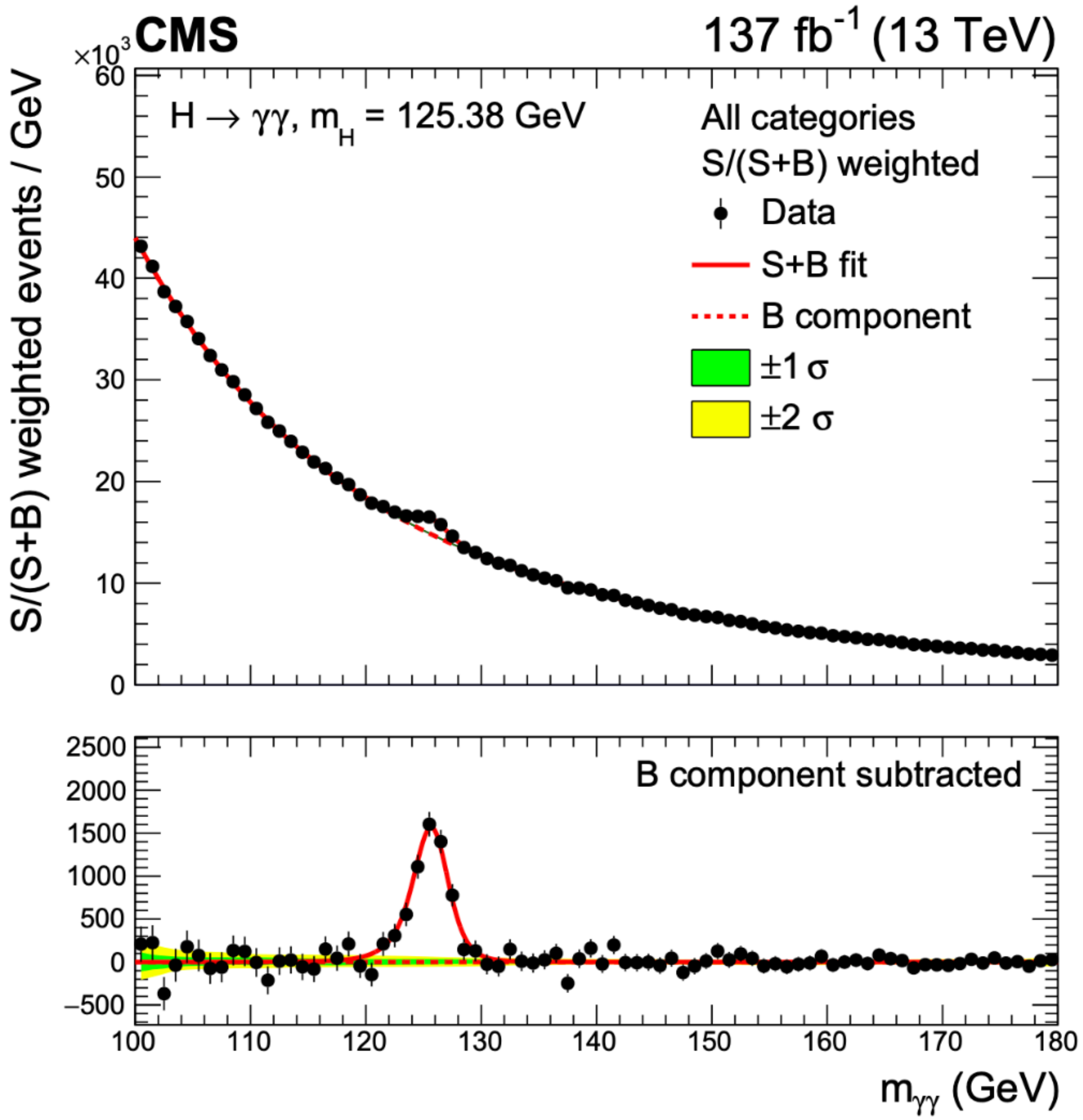




- Zero-sum game between generator and discriminator networks
- GANs are very difficult to train
  - but other generation techniques are no walk in the park either

Higgs to diphoton





- Two isolated photons are selected using a dedicated photon ID.
- In CMS we use a fit on data side bands to estimate the background occupancy.
  - DNNs are trained on simulation samples.

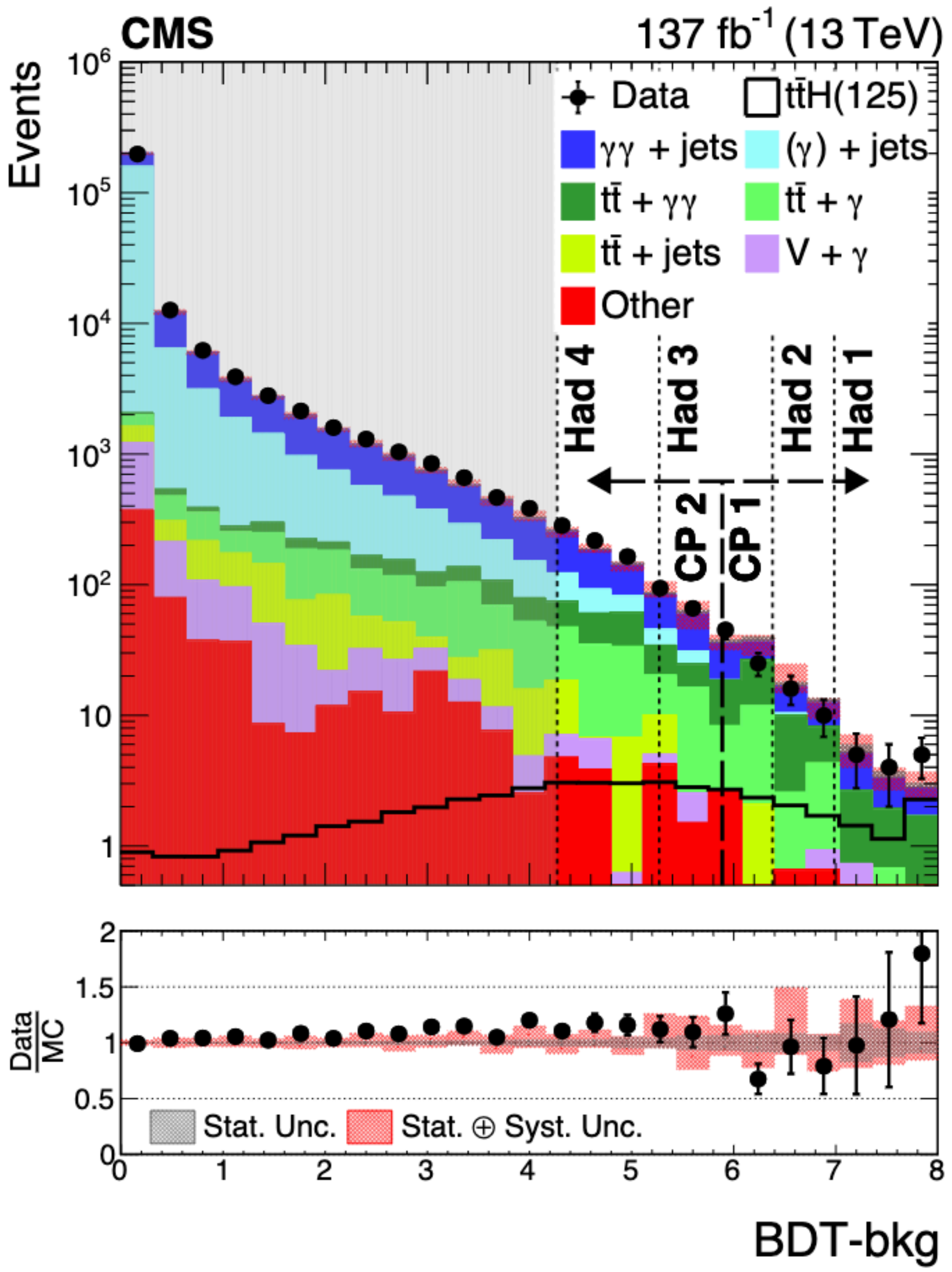
Higgs to diphoton backgrounds

- In the $H \rightarrow \gamma\gamma$ analysis, dominant backgrounds are :

  $\gamma\gamma$ + Jets, γ + Jets, Multi Jets (MJ)

  - The agreement between Data and Monte Carlo (MC) simulated samples for γ + Jets and MJ is not satisfying and the statistics is too low for the training of subsequent discriminants.

  - Therefore, CMS chooses to have a data-driven technique to simulate γ + Jets background process.

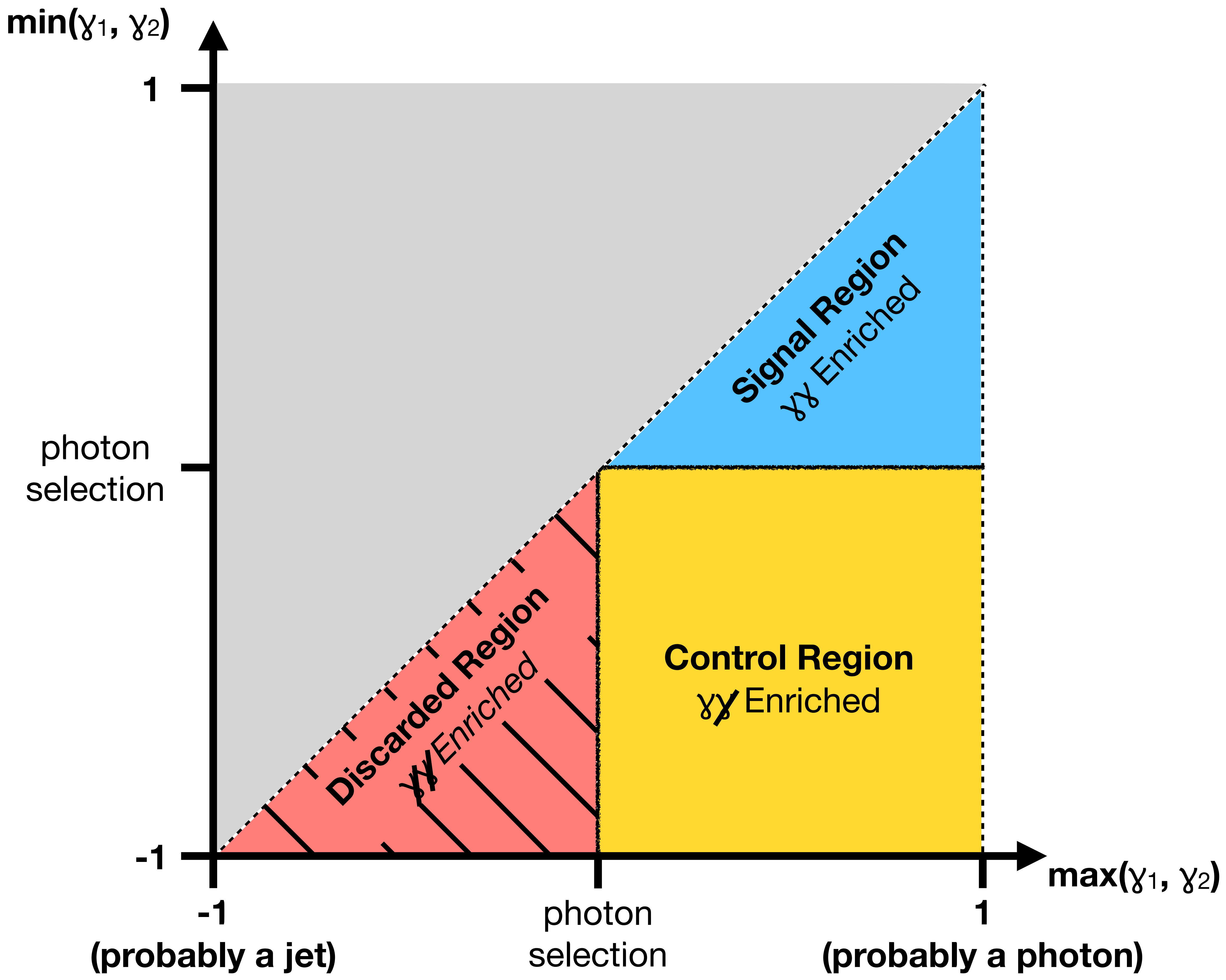


CMS collaboration
arXiv:2003.10866

Özgür Sahin
CEA Paris-Saclay / Irfu

Simulation of γ + Jets

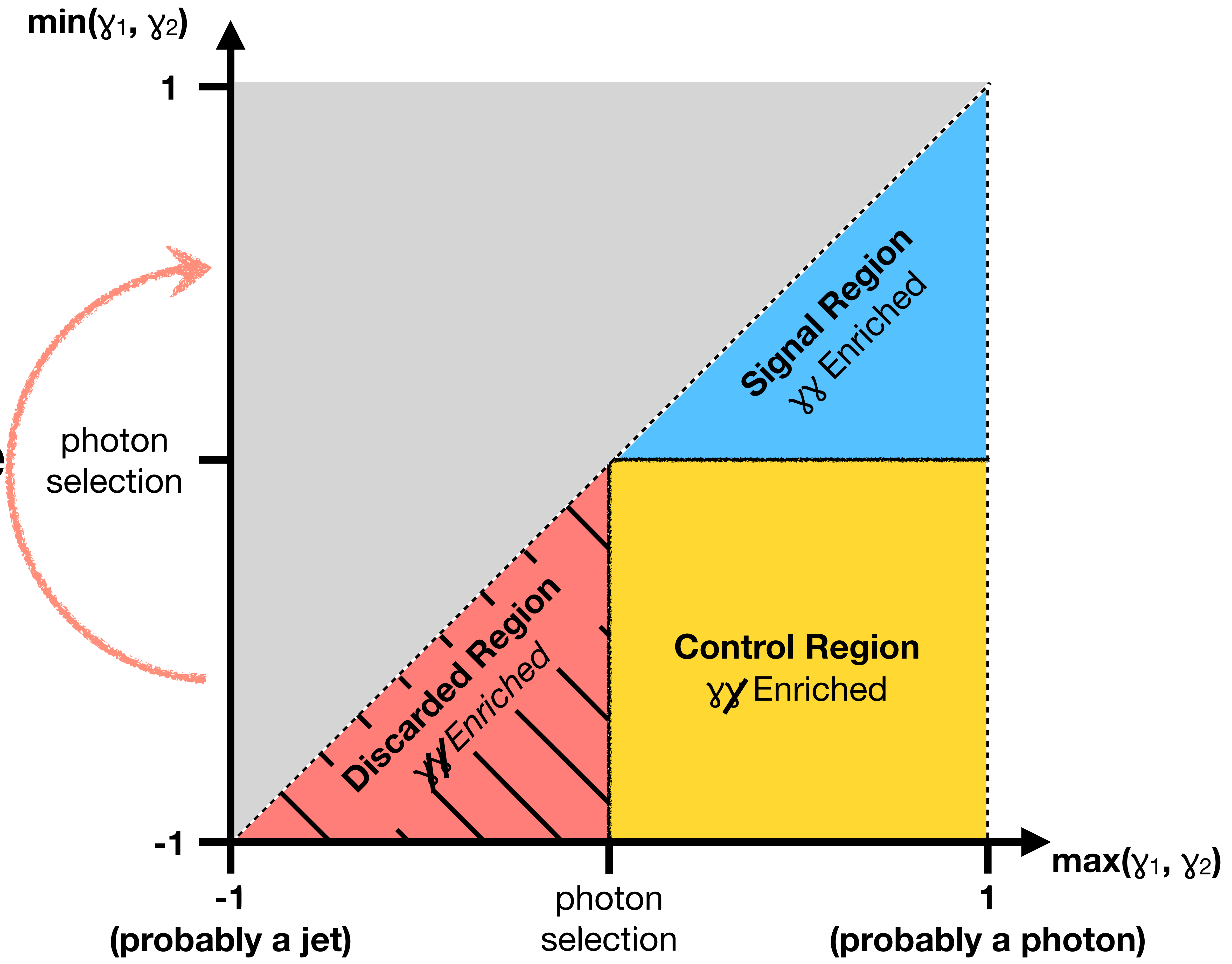
- The idea is simple:
  - We can simulate γ + Jets by recycling the data events that failed the Photon ID requirements.



Özgür Sahin
CEA Paris-Saclay / Irfu

Simulation of γ + Jets

- The idea is simple:
  - We can simulate γ + Jets by recycling the data events that failed the Photon ID requirements.
  - We remove the photon ID failing the selection and replacing by a distribution following the distribution of MC γ + Jets passing these selection criteria.

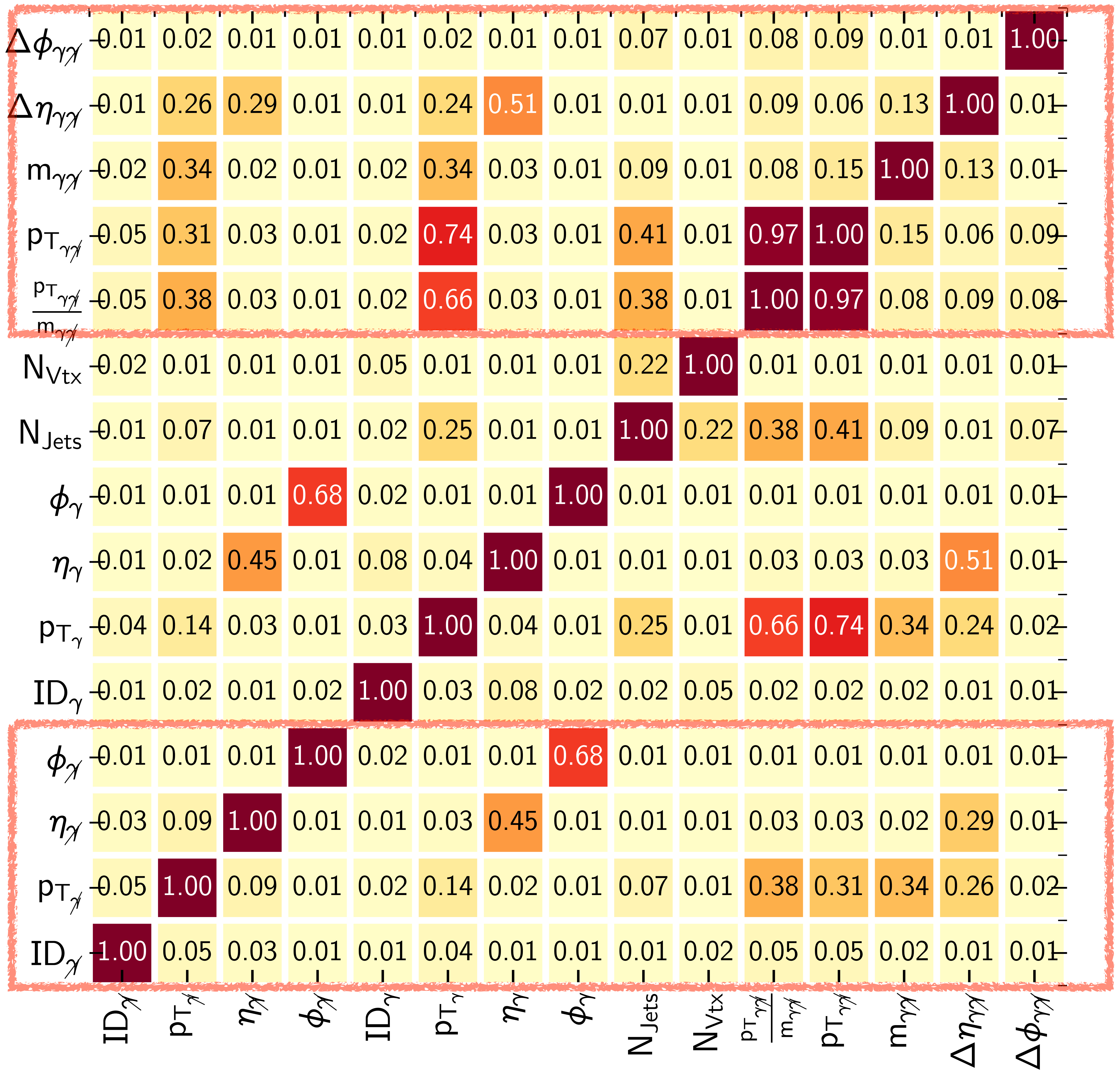


Özgür Sahin

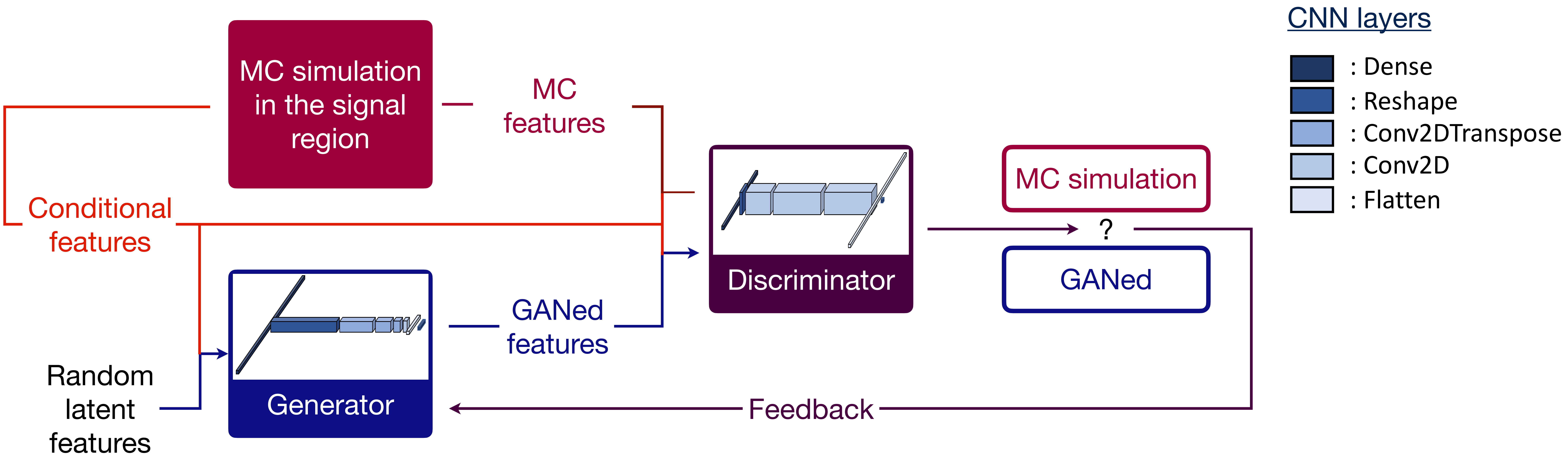
Simulation of γ + Jets

Z axis is the distance correlation.
Gabor et al. arXiv:0803.4101

- The idea is simple:
  - We can simulate γ + Jets by recycling the data events that failed the Photon ID requirements.
  - We remove the photon ID failing the selection and replacing by a distribution following the distribution of MC γ + Jets passing these selection criteria.
  - This method completely ignores any correlation and kinematical differences.



GAN to rescue



- Based on the DC-GAN architecture we employ a conditional GAN.
- Rather than generating the photon ID, we generate a new fake photon.

Özgür Sahin

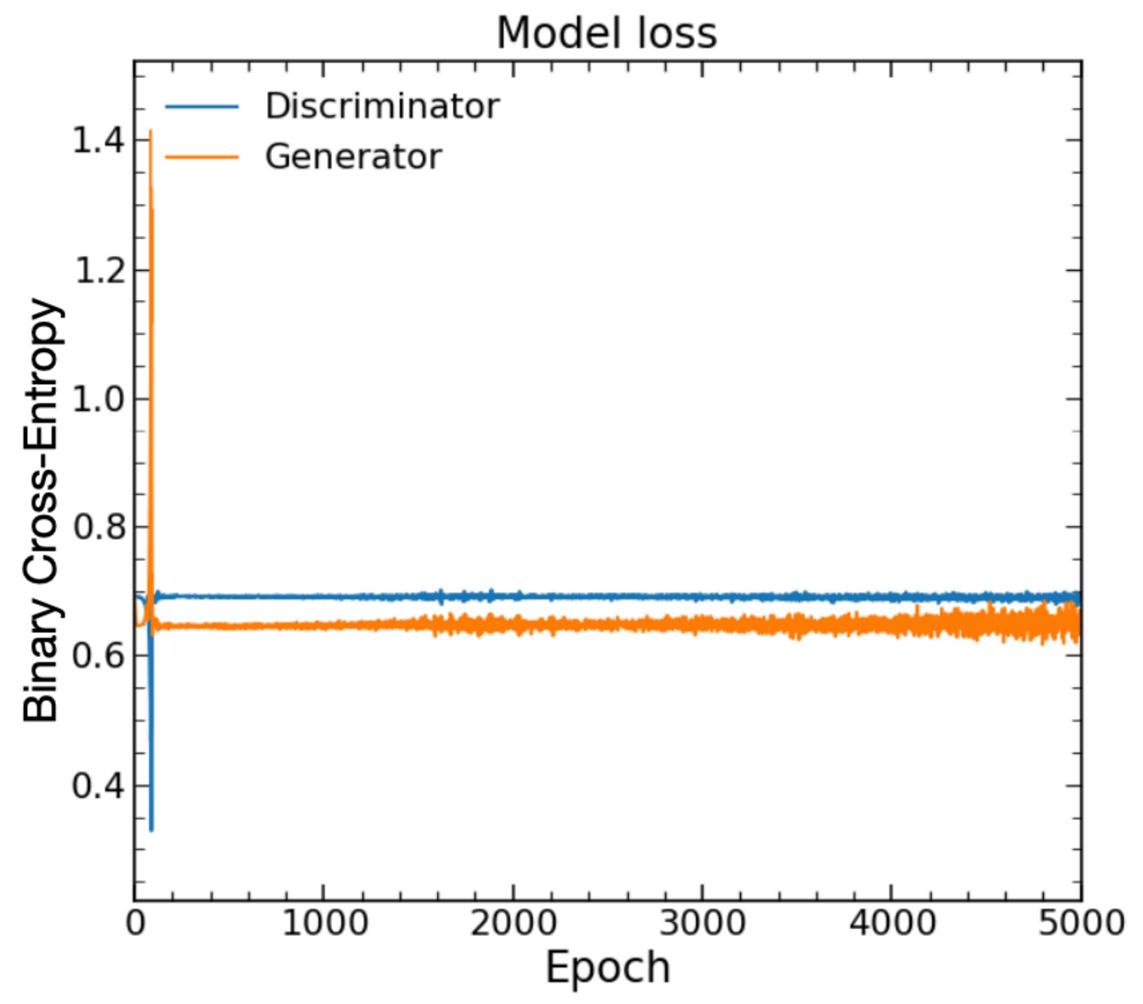
Training GAN

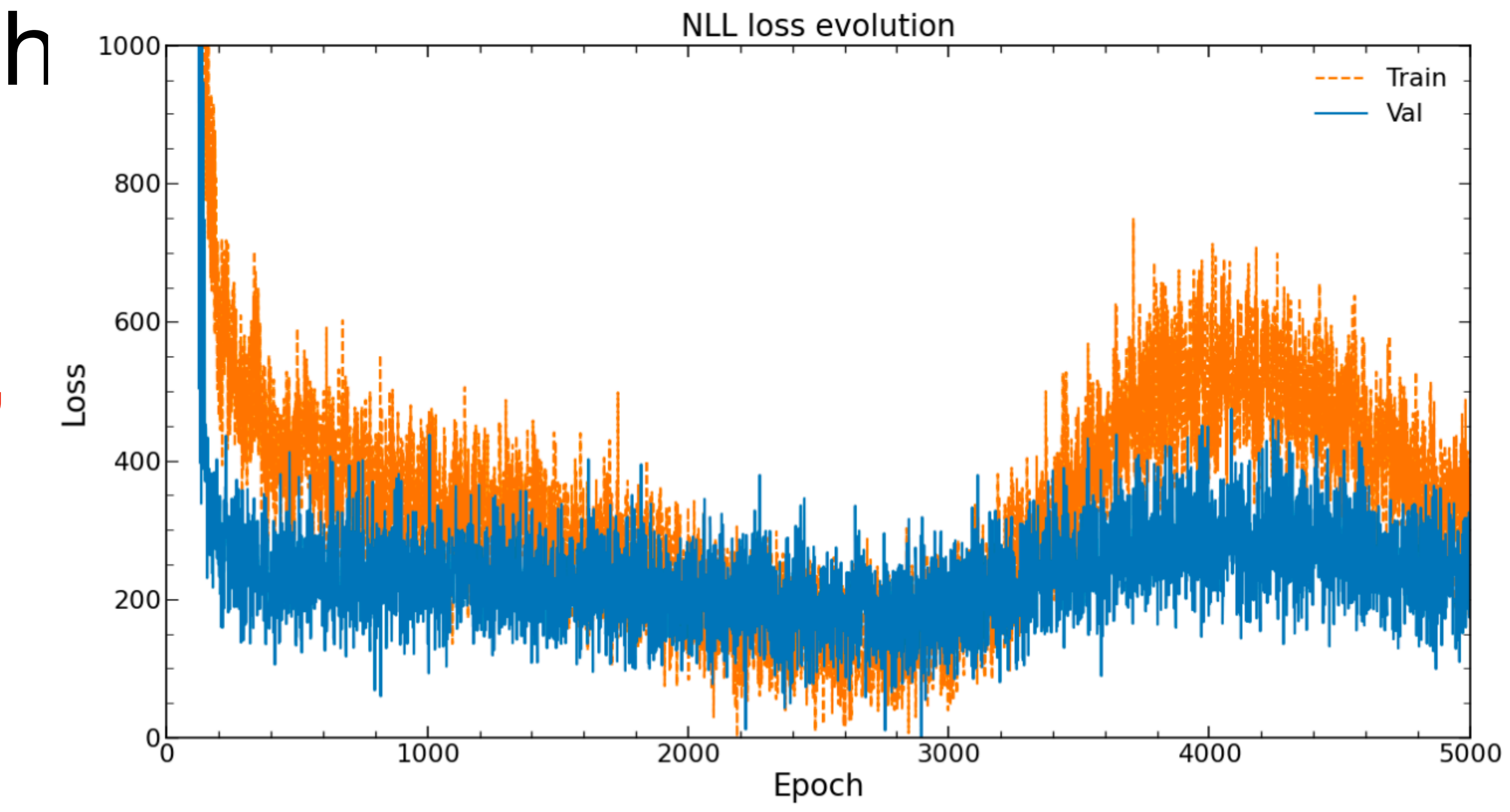
- However, the very idea of two adversary networks in GAN is that their loss are balanced.

- Wasserstein or KL based losses are expected overcome this:

  - We could not achieve the precision on both distributions and correlations

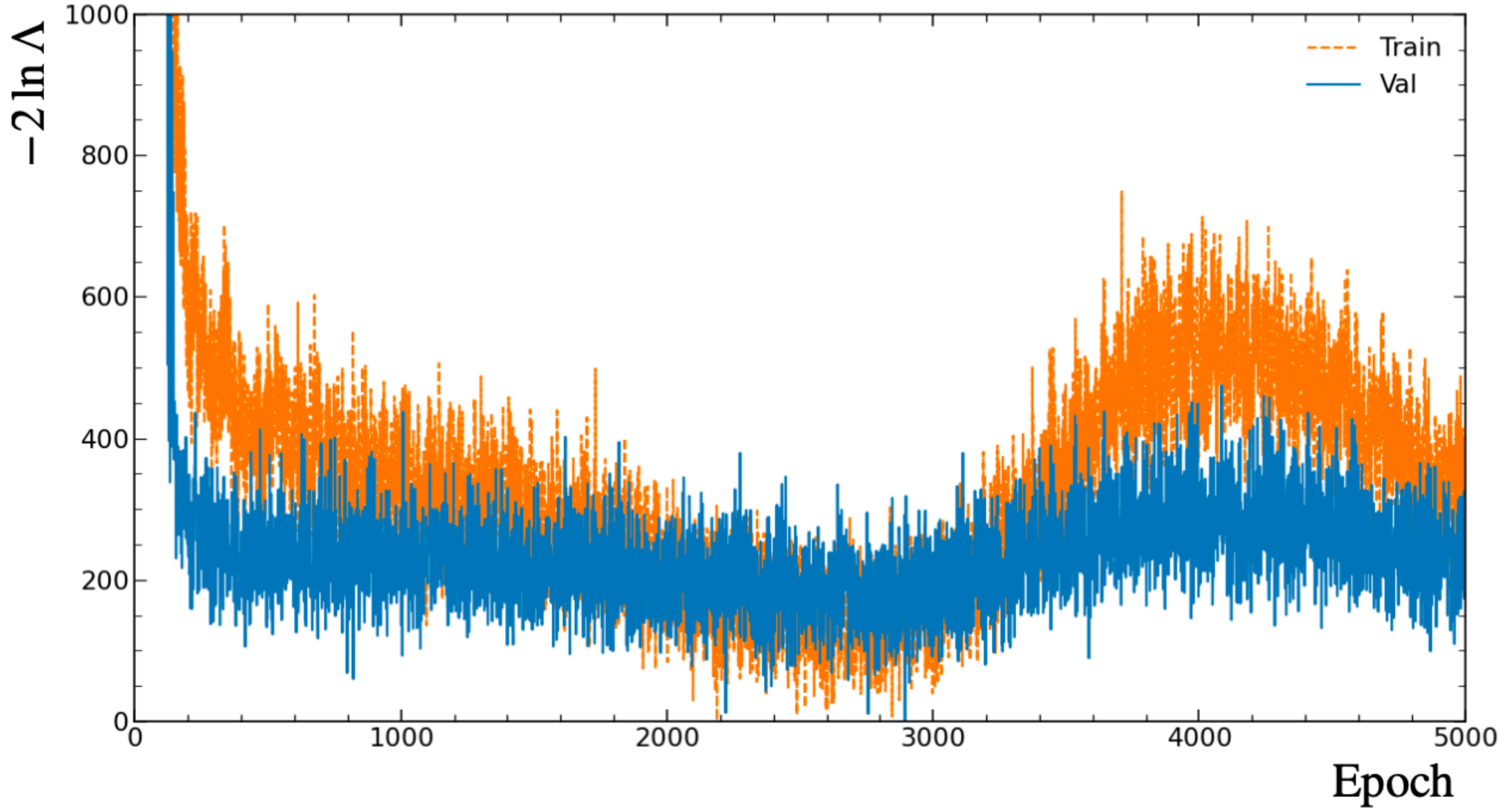
  - Rather than implementing a modified loss, use a sample-wise log likelihood metric.



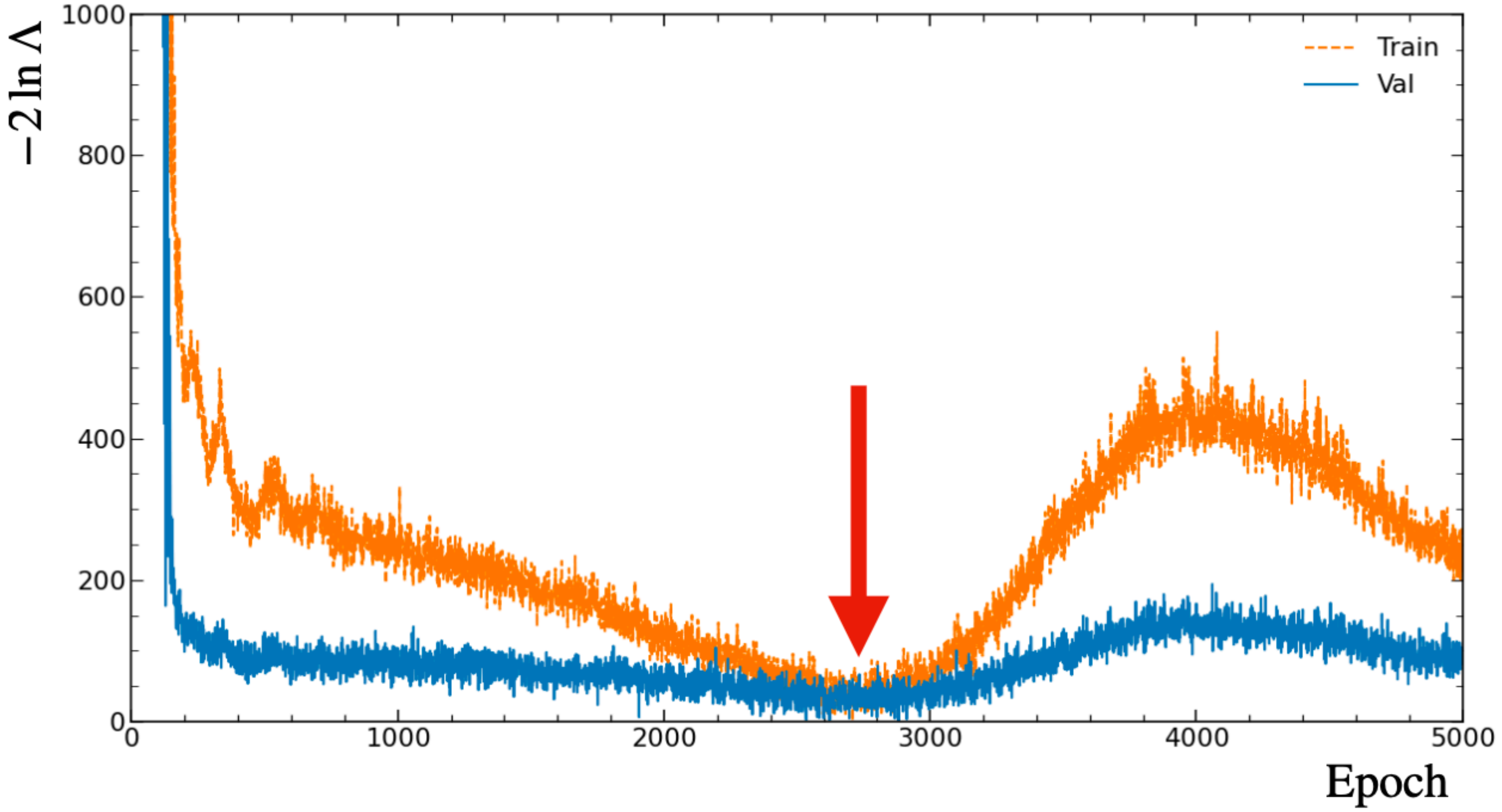


Picking the best model

1 generation per event

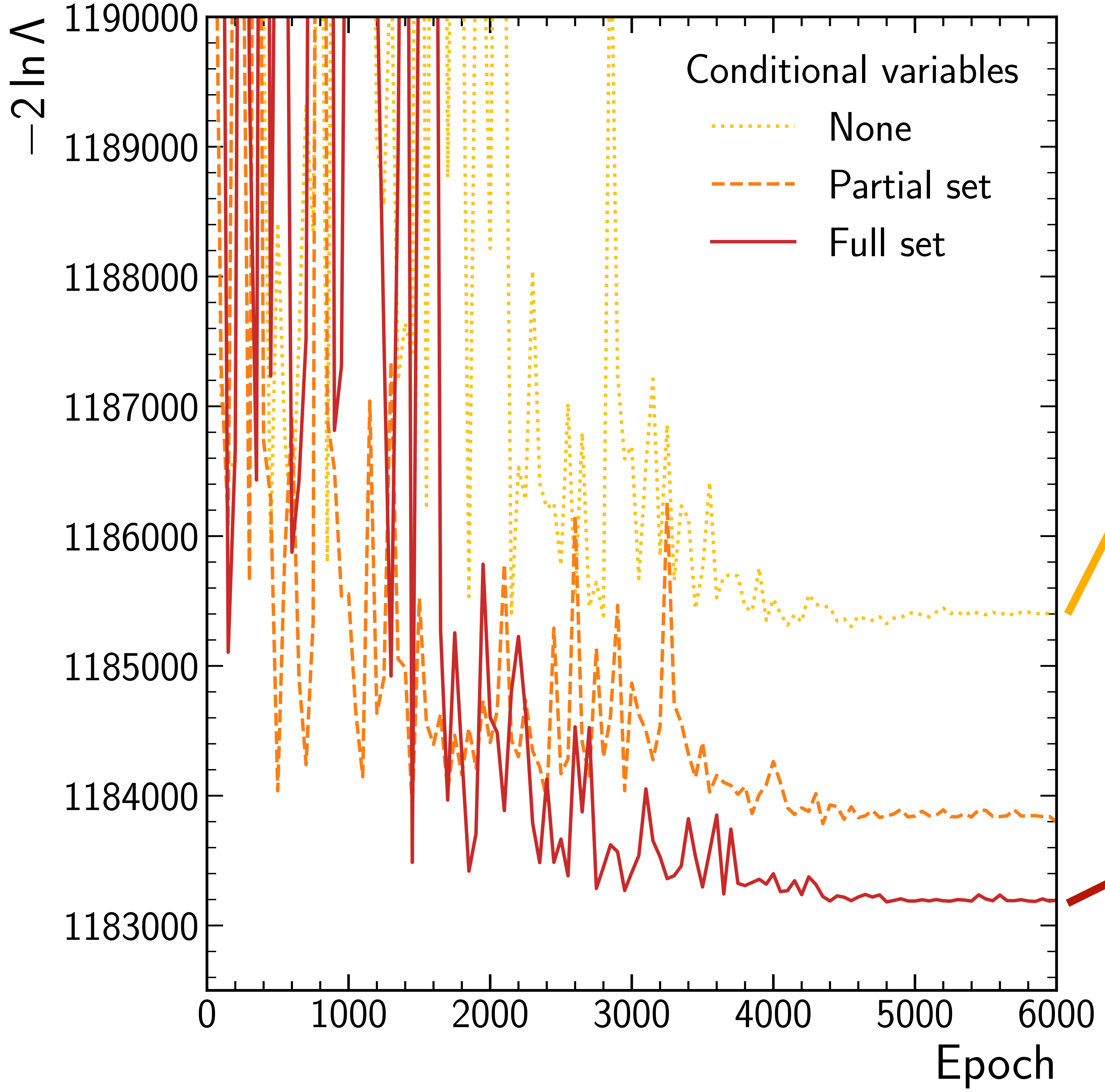


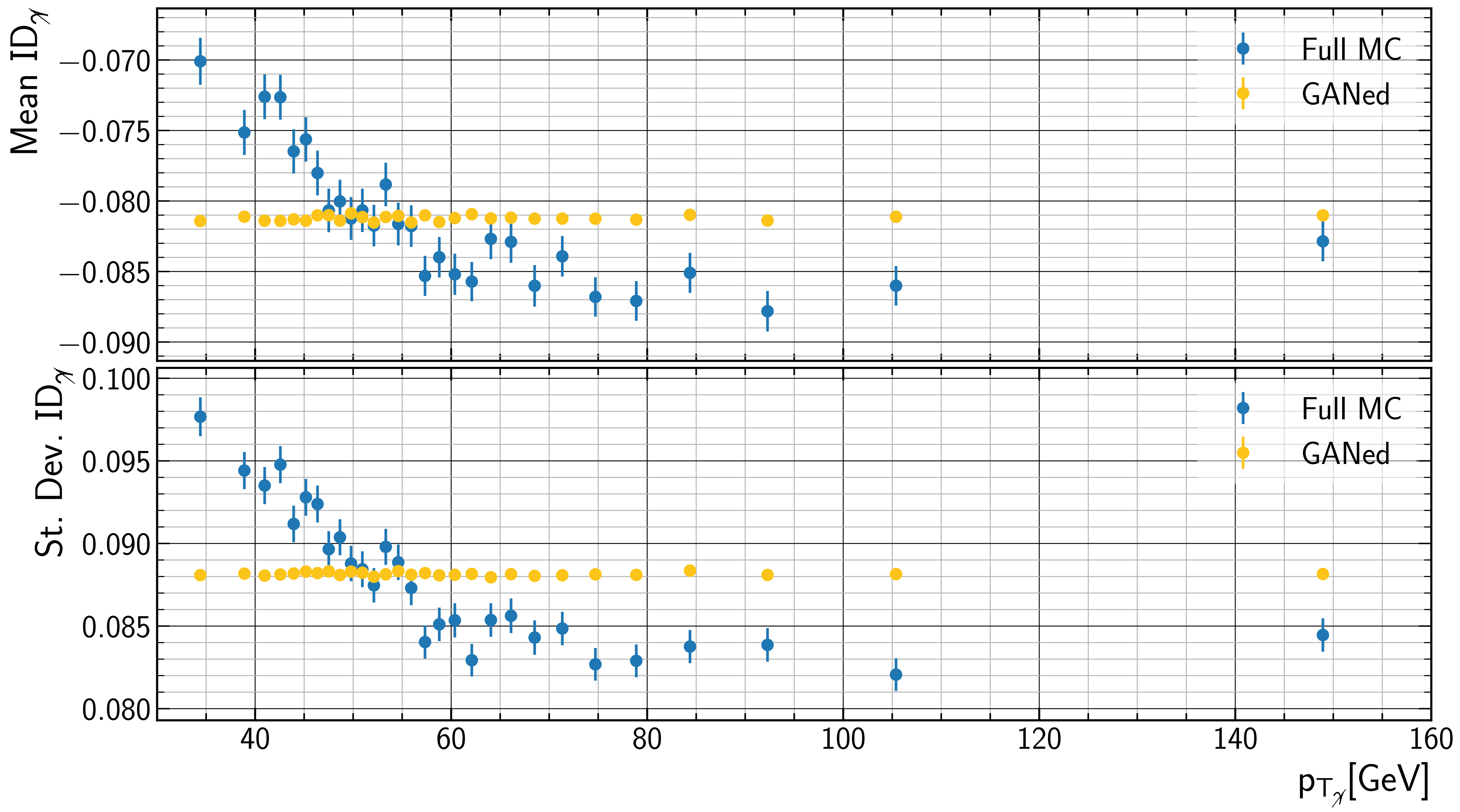
10 generations per event

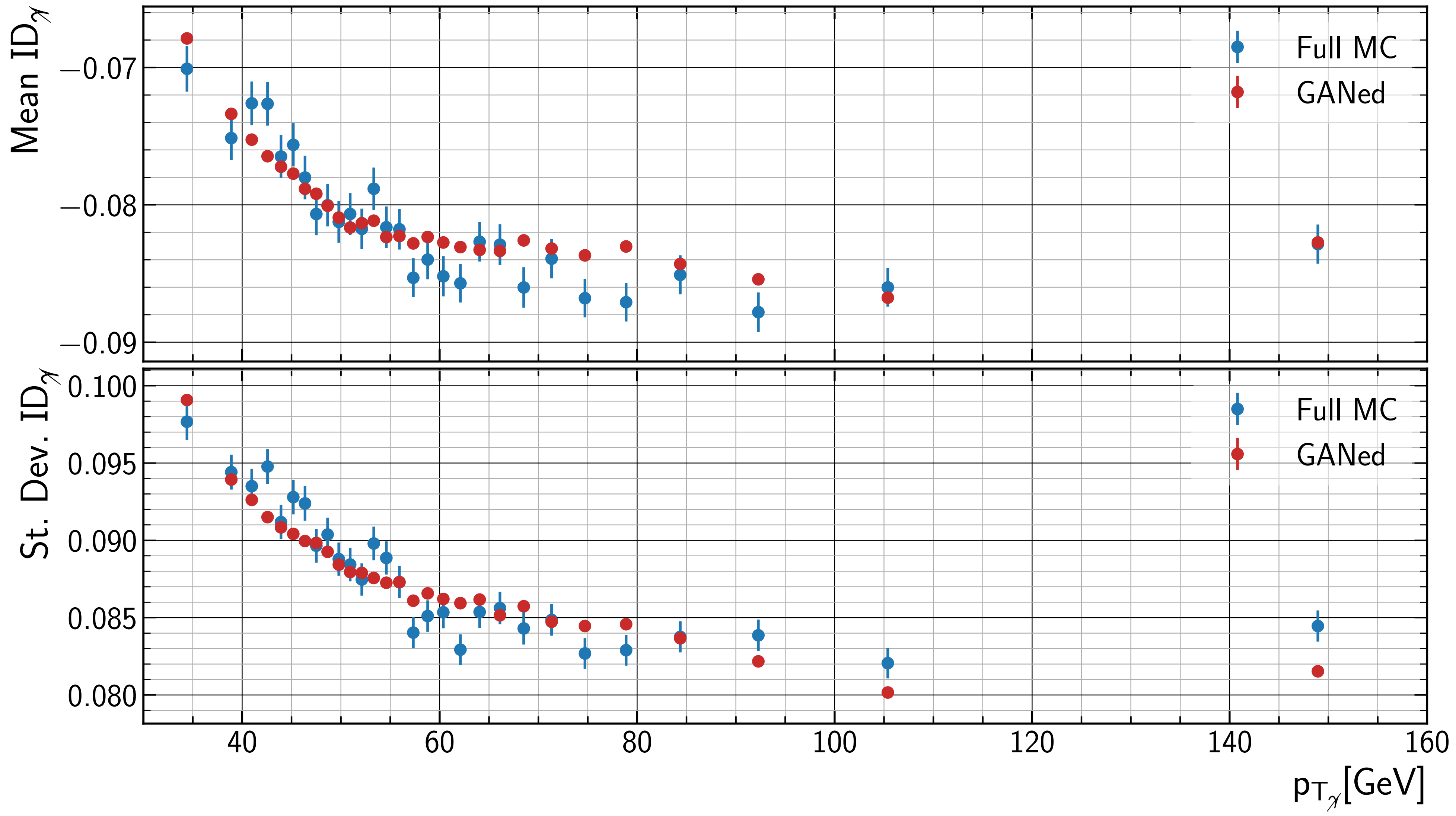


- The performance metric is heavily fluctuating:
  - Significant uncertainty in the selection of the best model.
  - Repeated generation of the events help us reduce these fluctuations.

Why conditional GAN?

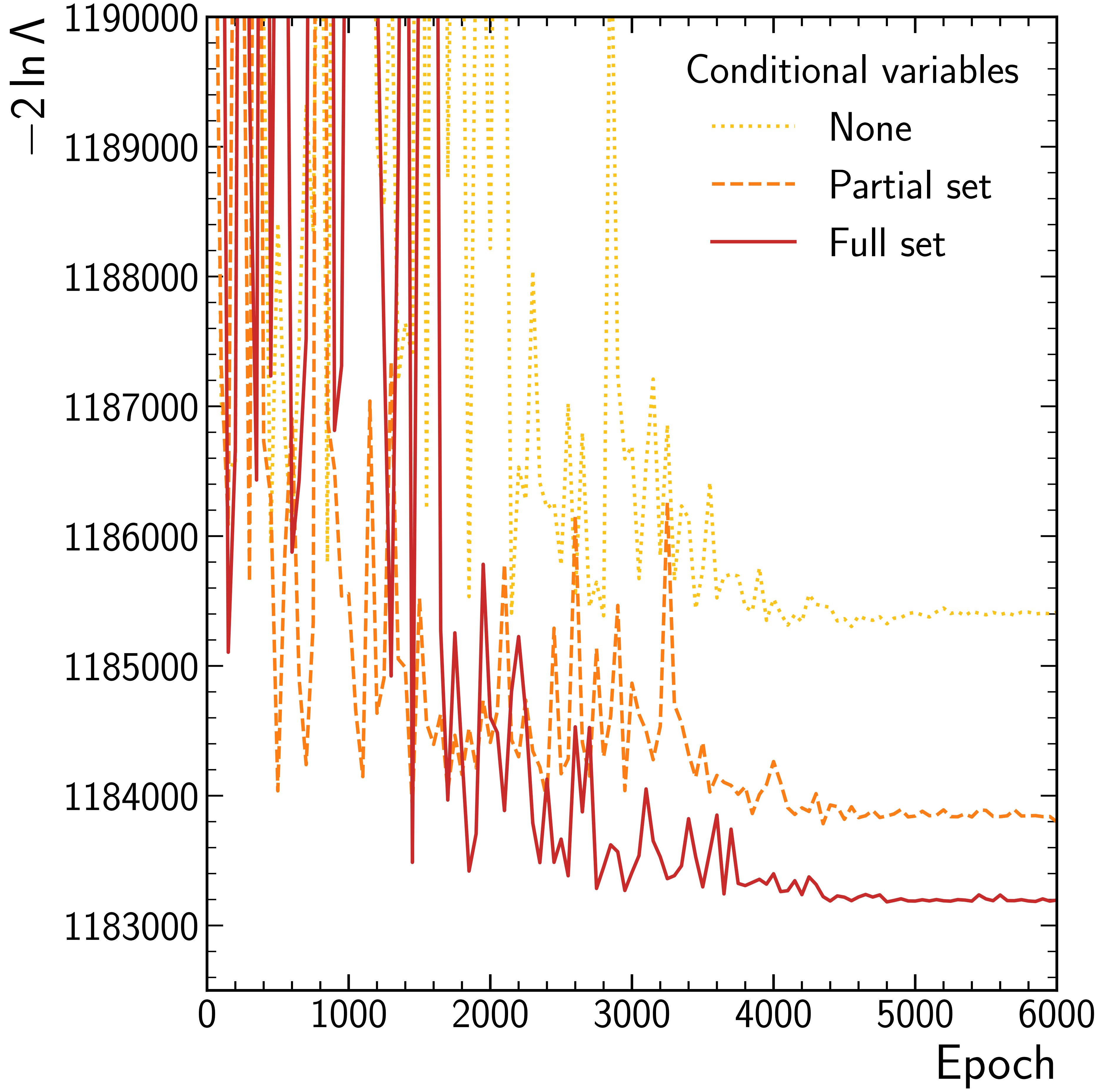




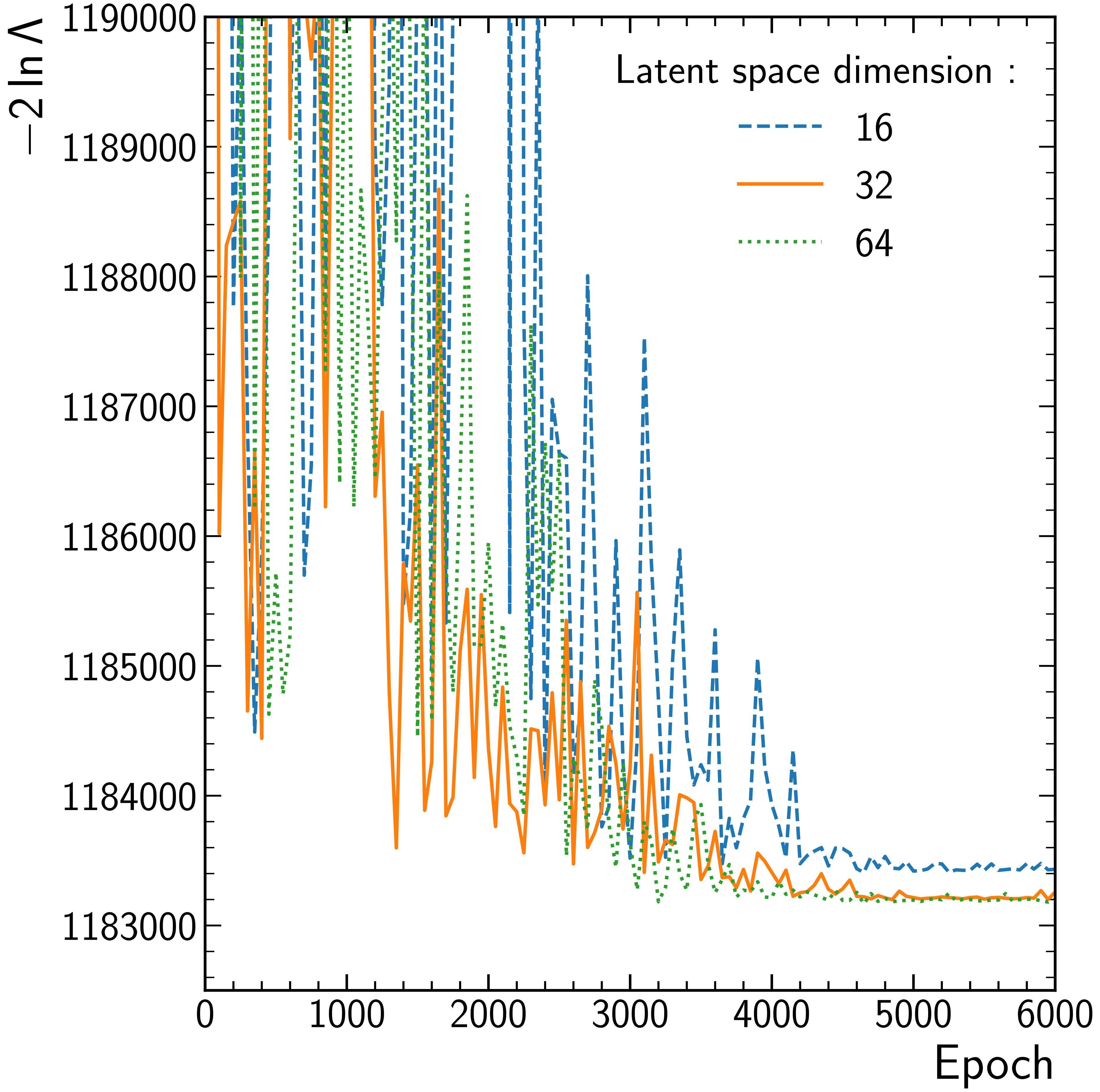


Why conditional GAN?

Impact of the conditional features

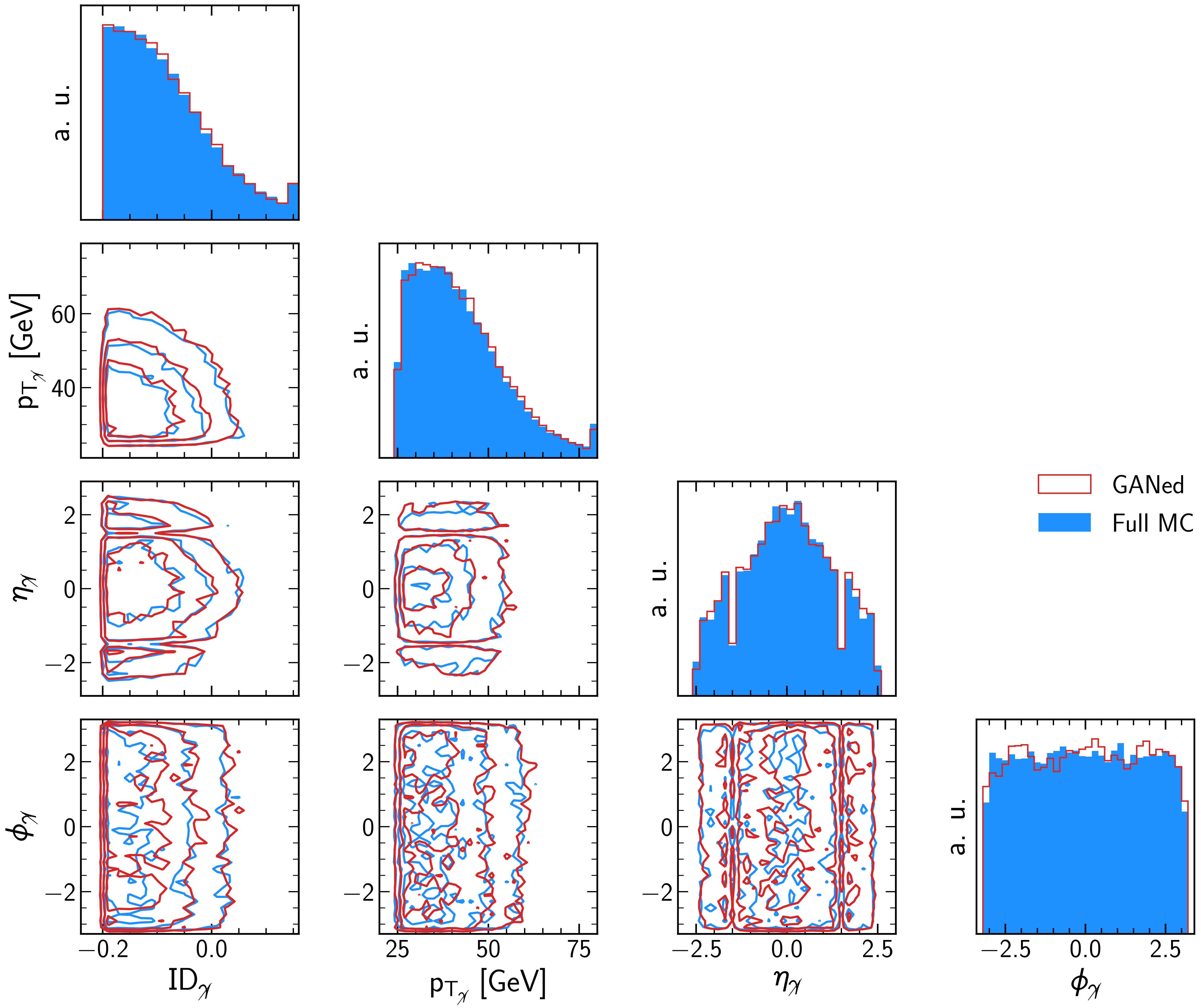


Impact of the random latent space

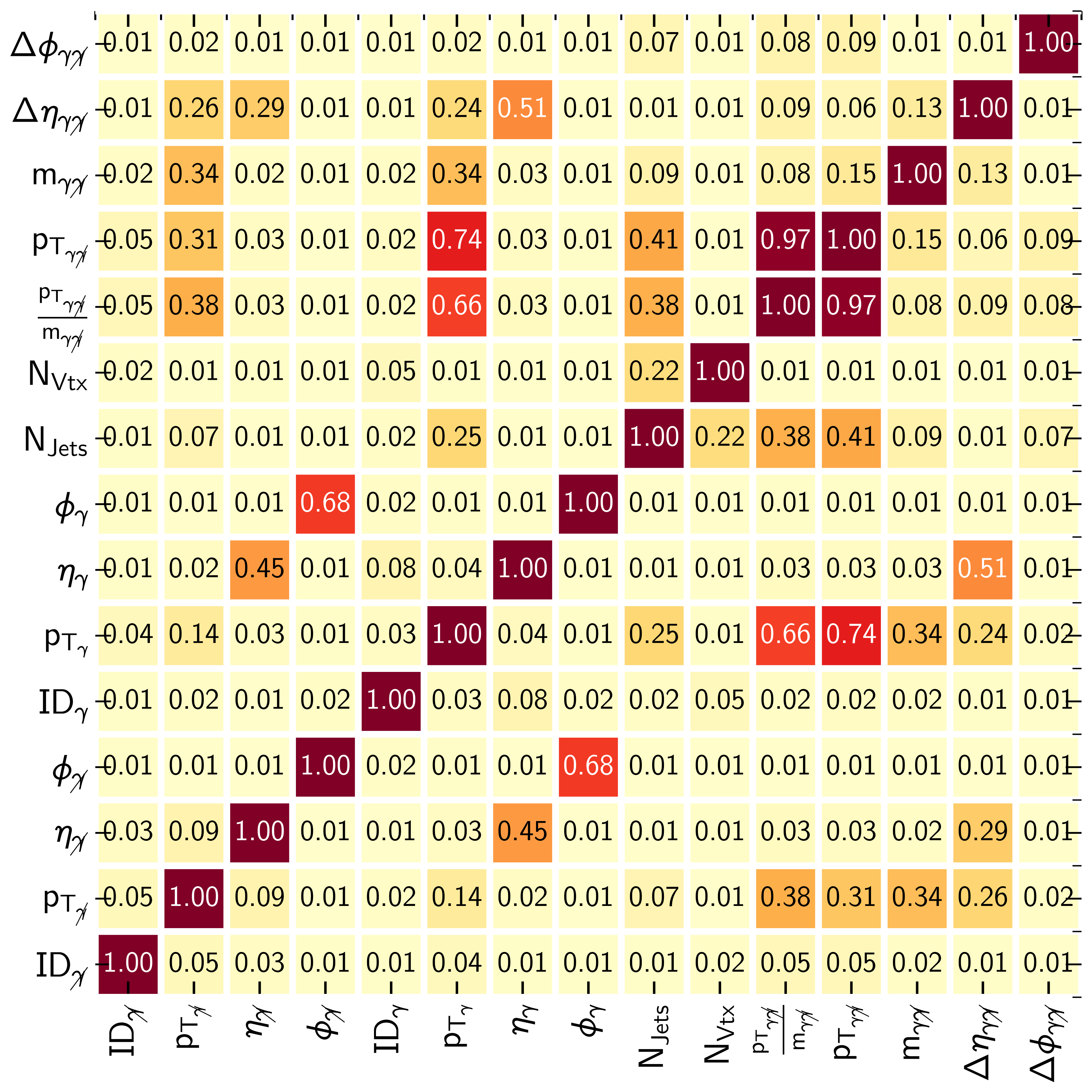


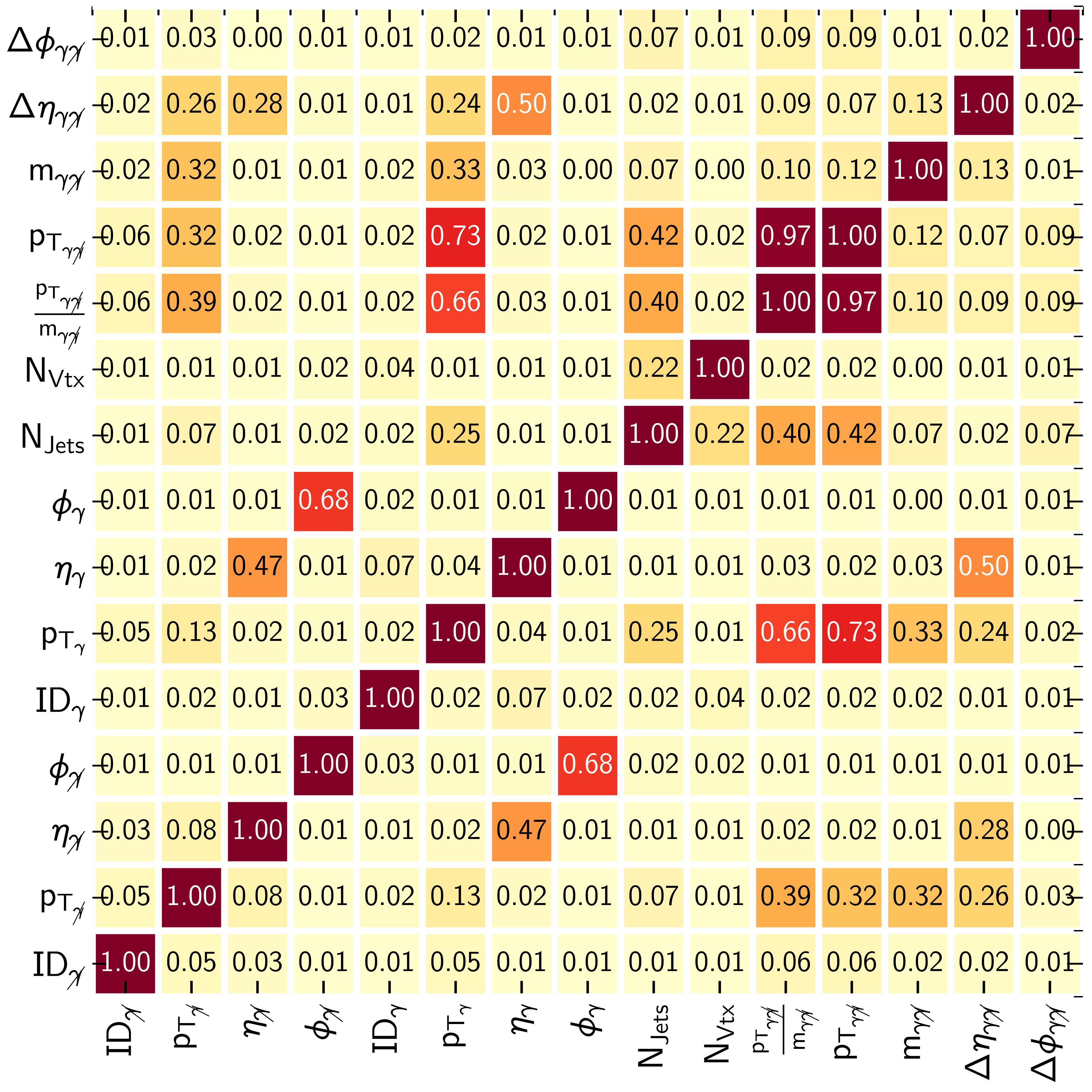
Putting all together

- With the conditional Generative AI we not only produce individual features, but also mimic the correlations nicely.



Correlations





- Z axis is the distance correlation.

Does it really work?

- Can the network generate from the 'control region' sample, a fake photon that behaves similarly to 'signal region' photon.