



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

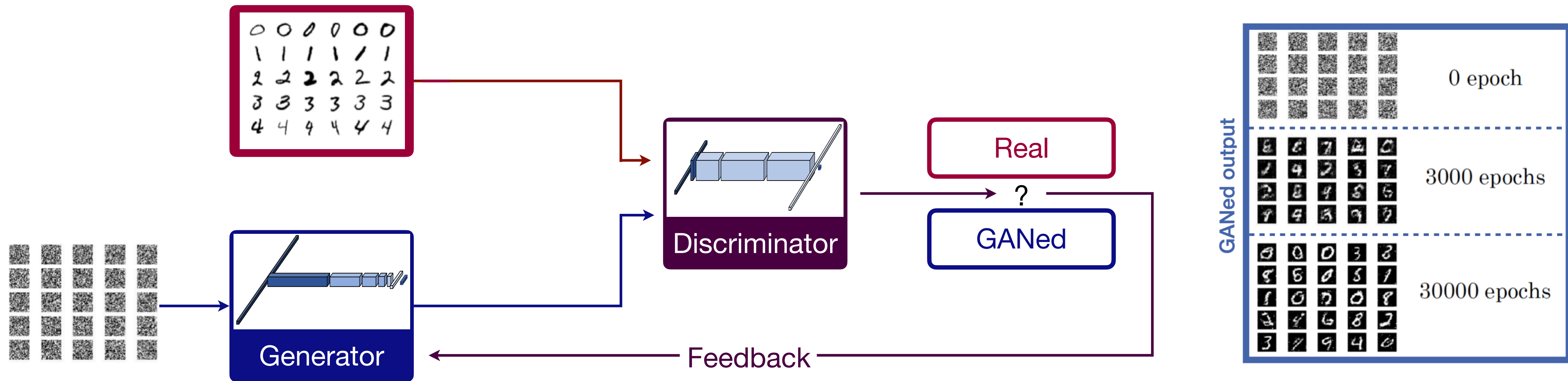
Monte Carlo simulations at the LHC

- 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

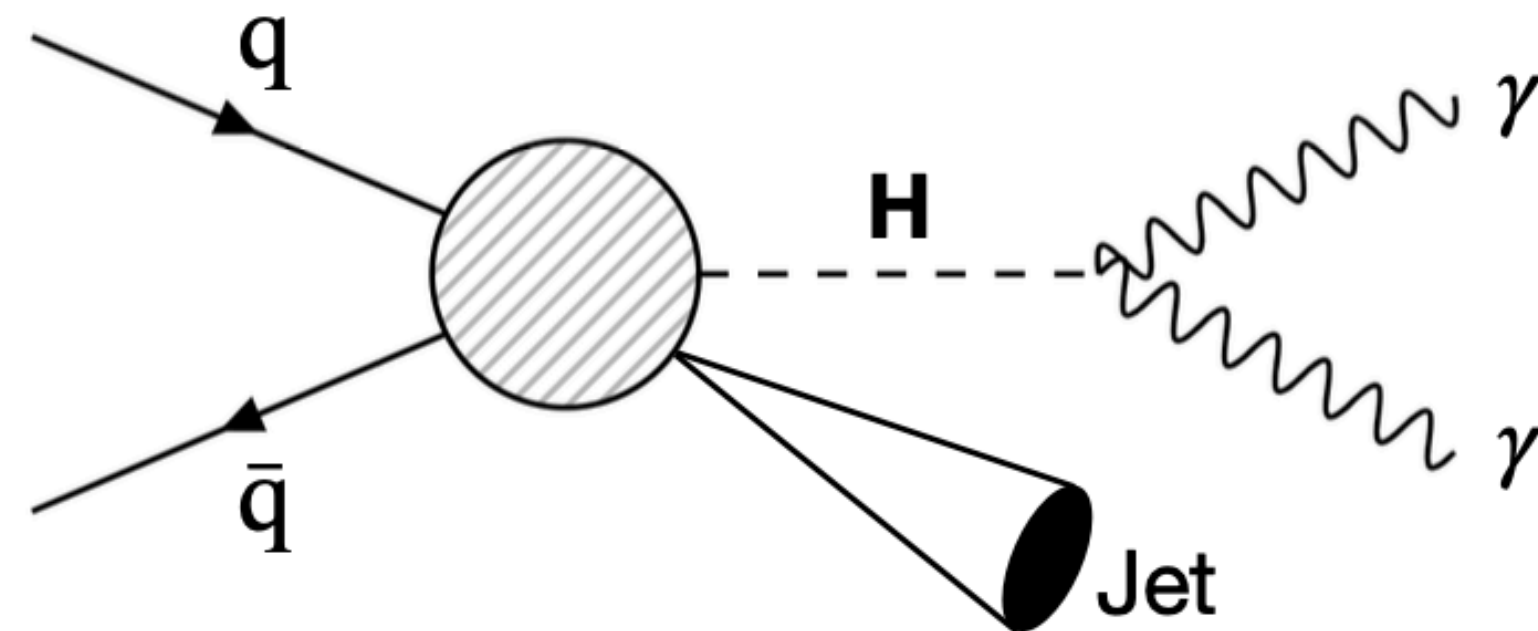
- Analysis specific datasets can be generate using generative ML algorithms such as GANs, AE, normalizing flows
 - Once trained, they are very efficient to evaluate.
 - However, as most of them trained on MC samples, they inherent some of the concerns from the last slides.
- What if we use these ML techniques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

GAN

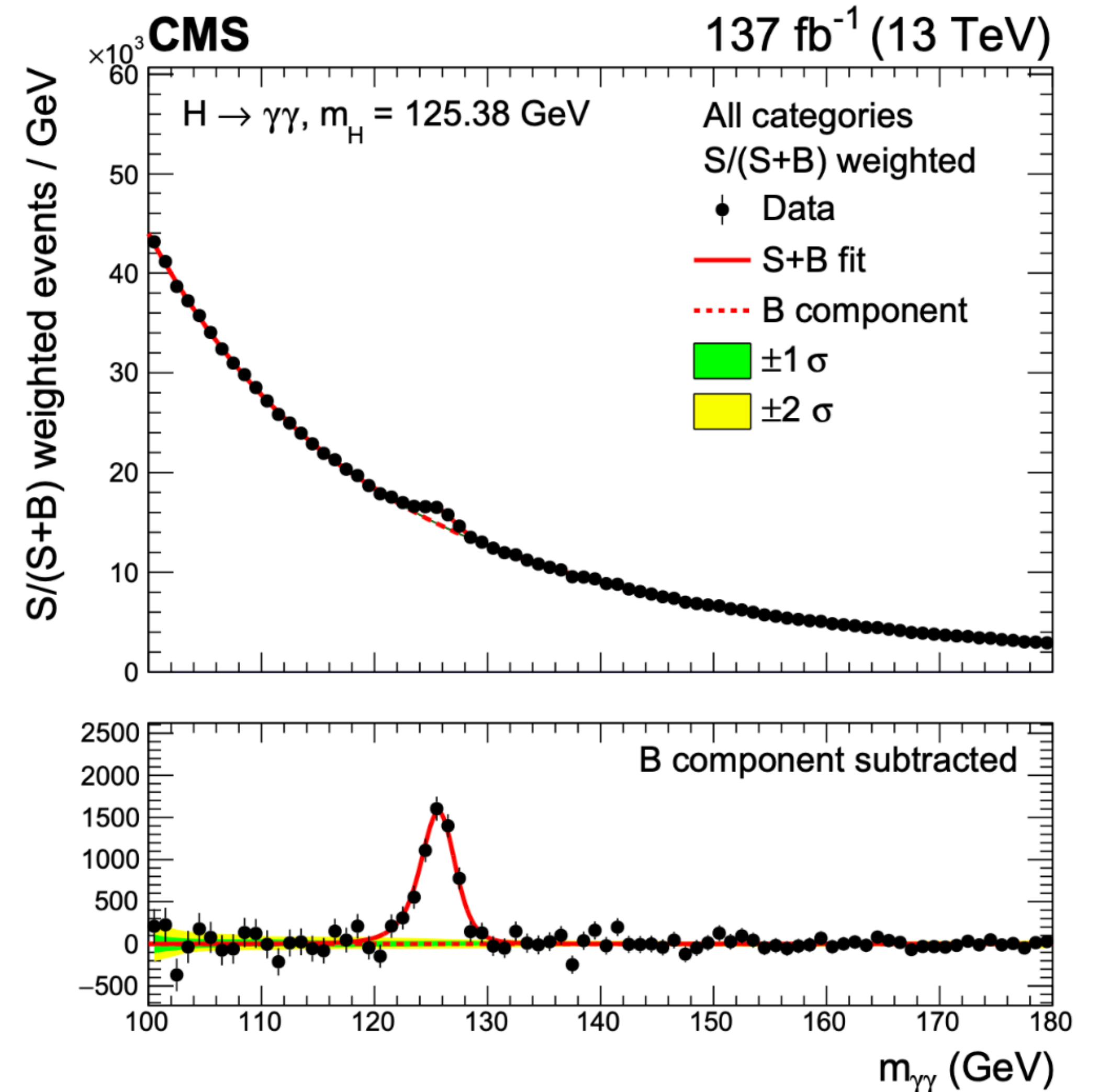


- Zero-sum game between generator and discriminator networks
- GANs are very difficult to train
 - but other generation techniques are no walk in the park either

Higgs to diphoton

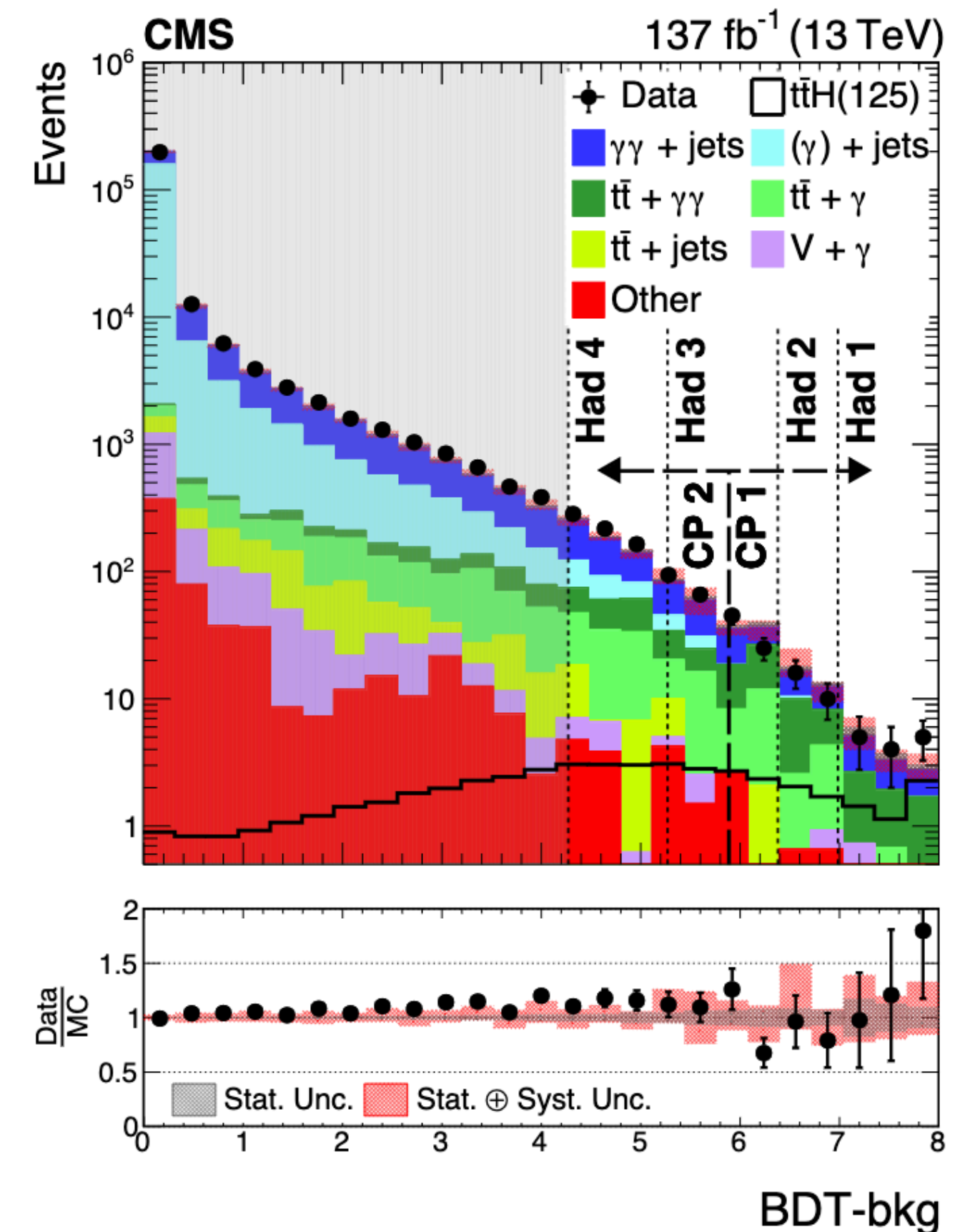


- Two isolated photons are selected using a dedicated photon ID.
- In CMS we use a fit on data side bands to estimate the background occupancy.
 - DNNs are trained on simulation samples.



Higgs to diphoton backgrounds

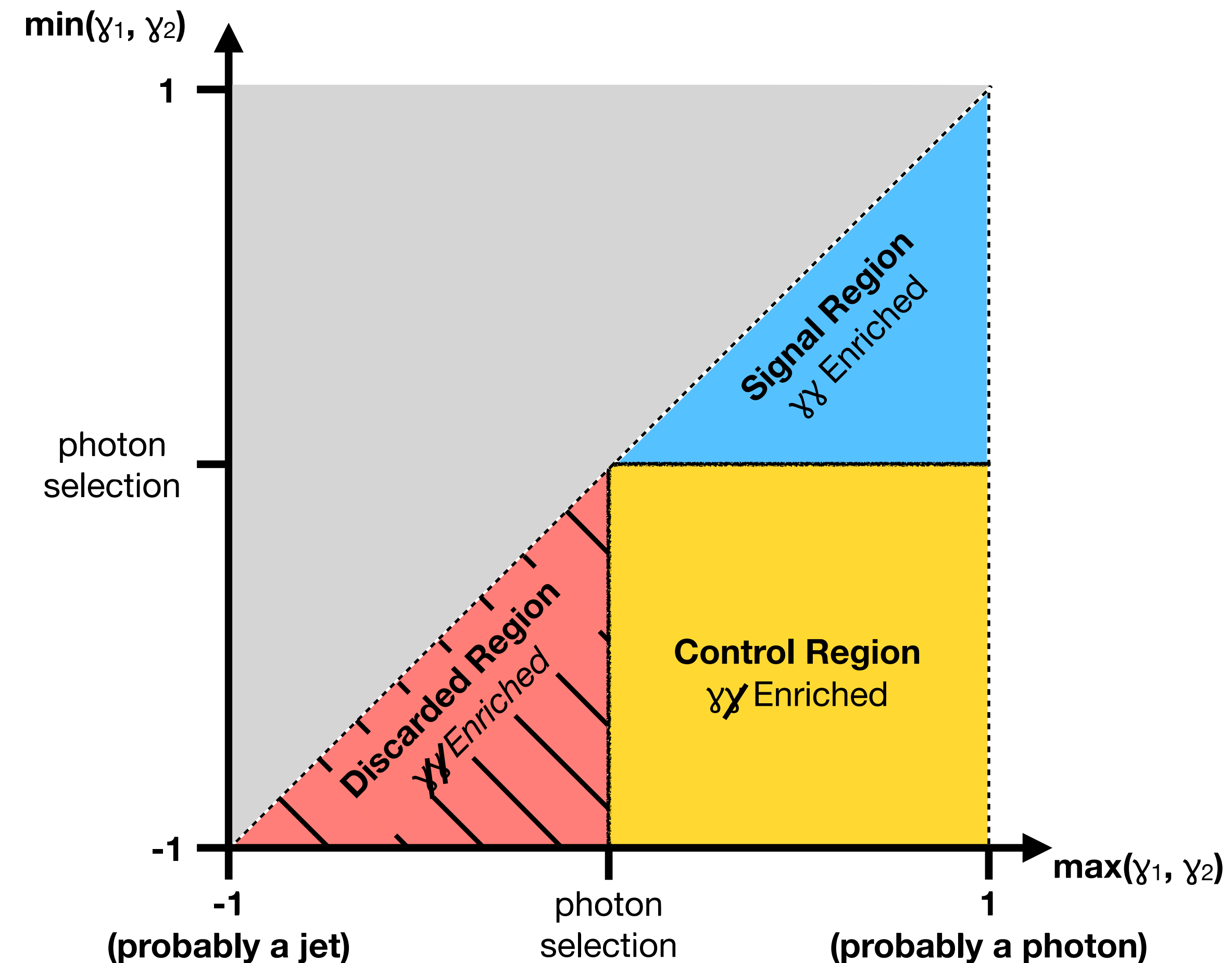
- In the $H \rightarrow \gamma\gamma$ analysis, dominant backgrounds are :
 $\gamma\gamma + \text{Jets}$, $\gamma + \text{Jets}$, Multi Jets (MJ)
- The agreement between Data and Monte Carlo (MC) simulated samples for $\gamma + \text{Jets}$ and MJ is not satisfying and the statistics is too low for the training of subsequent discriminants.
- Therefore, CMS chooses to have a data-driven technique to simulate $\gamma + \text{Jets}$ background process.



CMS collaboration
arXiv:2003.10866

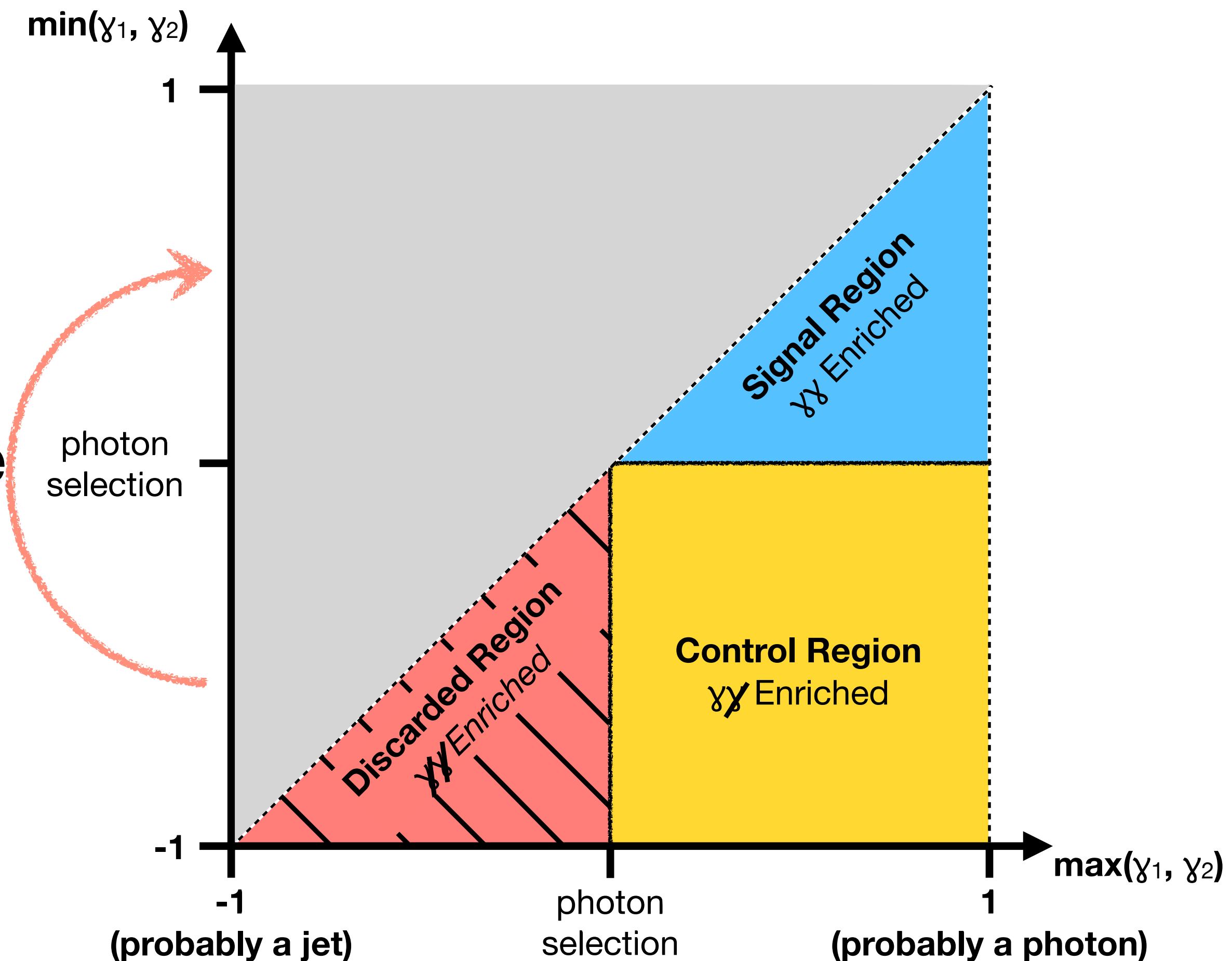
Simulation of γ + Jets

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



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.



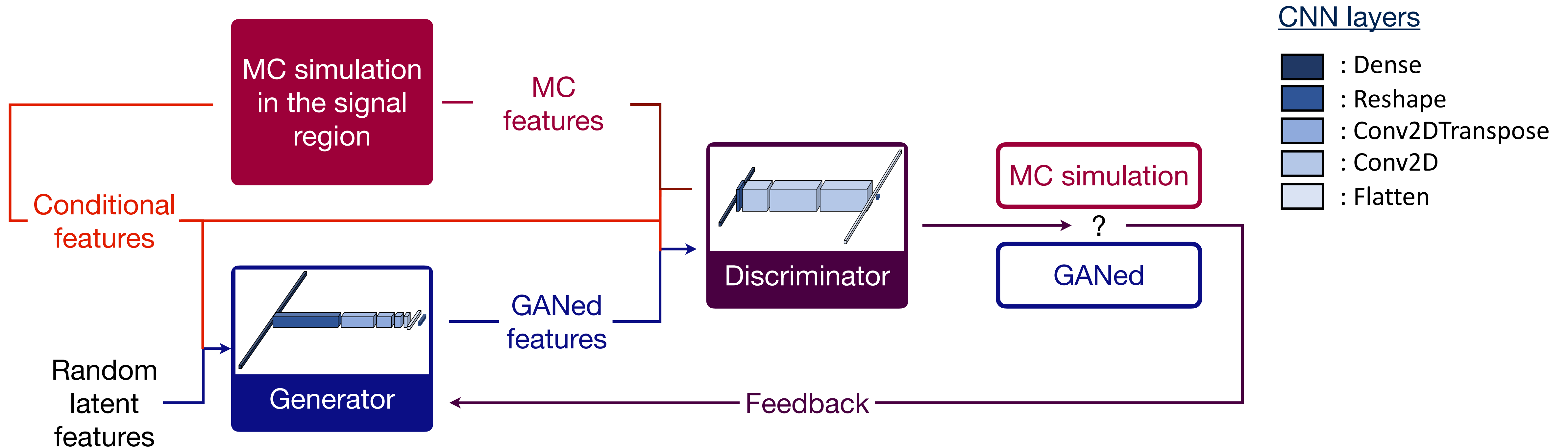
Simulation of γ + Jets

- 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.
 - This method completely ignores any correlation and kinematical differences.

Z axis is the distance correlation.
Gabor et al. [arXiv:0803.4101](https://arxiv.org/abs/0803.4101)

$\Delta\phi_{\gamma\gamma}$	-0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.07	0.01	0.08	0.09	0.01	0.01	1.00
$\Delta\eta_{\gamma\gamma}$	-0.01	0.26	0.29	0.01	0.01	0.24	0.51	0.01	0.01	0.01	0.09	0.06	0.13	1.00	0.01
$m_{\gamma\gamma}$	-0.02	0.34	0.02	0.01	0.02	0.34	0.03	0.01	0.09	0.01	0.08	0.15	1.00	0.13	0.01
$p_{T\gamma\gamma}$	-0.05	0.31	0.03	0.01	0.02	0.74	0.03	0.01	0.41	0.01	0.97	1.00	0.15	0.06	0.09
$\frac{p_{T\gamma\gamma}}{m_{\gamma\gamma}}$	-0.05	0.38	0.03	0.01	0.02	0.66	0.03	0.01	0.38	0.01	1.00	0.97	0.08	0.09	0.08
N_{Vtx}	-0.02	0.01	0.01	0.01	0.05	0.01	0.01	0.01	0.22	1.00	0.01	0.01	0.01	0.01	0.01
N_{Jets}	-0.01	0.07	0.01	0.01	0.02	0.25	0.01	0.01	1.00	0.22	0.38	0.41	0.09	0.01	0.07
ϕ_γ	-0.01	0.01	0.01	0.68	0.02	0.01	0.01	1.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
η_γ	-0.01	0.02	0.45	0.01	0.08	0.04	1.00	0.01	0.01	0.01	0.03	0.03	0.03	0.51	0.01
$p_{T\gamma}$	-0.04	0.14	0.03	0.01	0.03	1.00	0.04	0.01	0.25	0.01	0.66	0.74	0.34	0.24	0.02
ID_γ	-0.01	0.02	0.01	0.02	1.00	0.03	0.08	0.02	0.02	0.05	0.02	0.02	0.02	0.01	0.01
ϕ_γ	-0.01	0.01	0.01	1.00	0.02	0.01	0.01	0.68	0.01	0.01	0.01	0.01	0.01	0.01	0.01
η_γ	-0.03	0.09	1.00	0.01	0.01	0.03	0.45	0.01	0.01	0.01	0.03	0.03	0.02	0.29	0.01
$p_{T\gamma}$	-0.05	1.00	0.09	0.01	0.02	0.14	0.02	0.01	0.07	0.01	0.38	0.31	0.34	0.26	0.02
ID_γ	1.00	0.05	0.03	0.01	0.01	0.04	0.01	0.01	0.01	0.02	0.05	0.05	0.02	0.01	0.01
ID_γ															
$p_{T\gamma}$															
η_γ															
ϕ_γ															
N_{Jets}															
N_{Vtx}															
$\frac{p_{T\gamma\gamma}}{m_{\gamma\gamma}}$															
$p_{T\gamma\gamma}$															
$m_{\gamma\gamma}$															
$\Delta\eta_{\gamma\gamma}$															
$\Delta\phi_{\gamma\gamma}$															

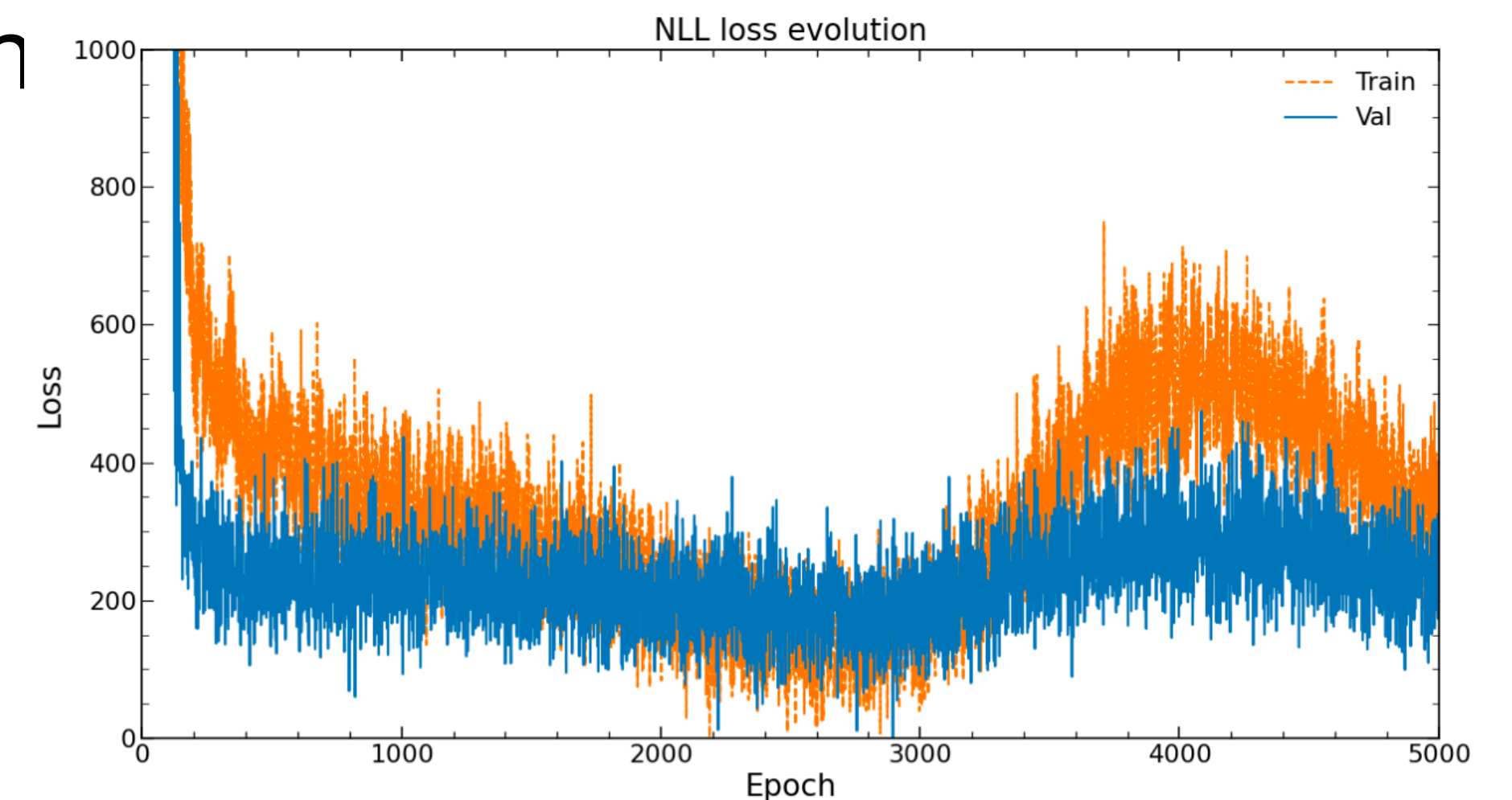
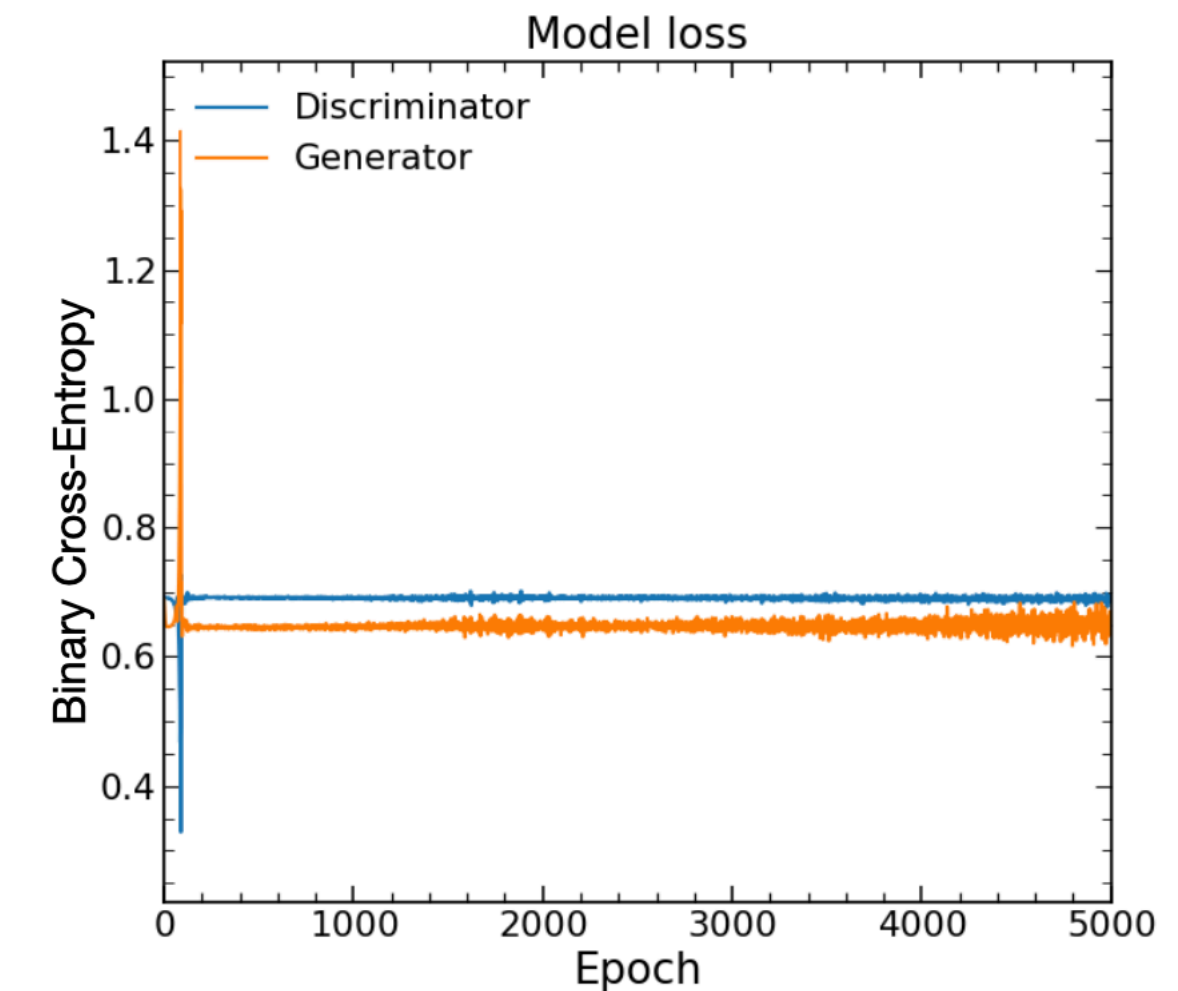
GAN to rescue



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

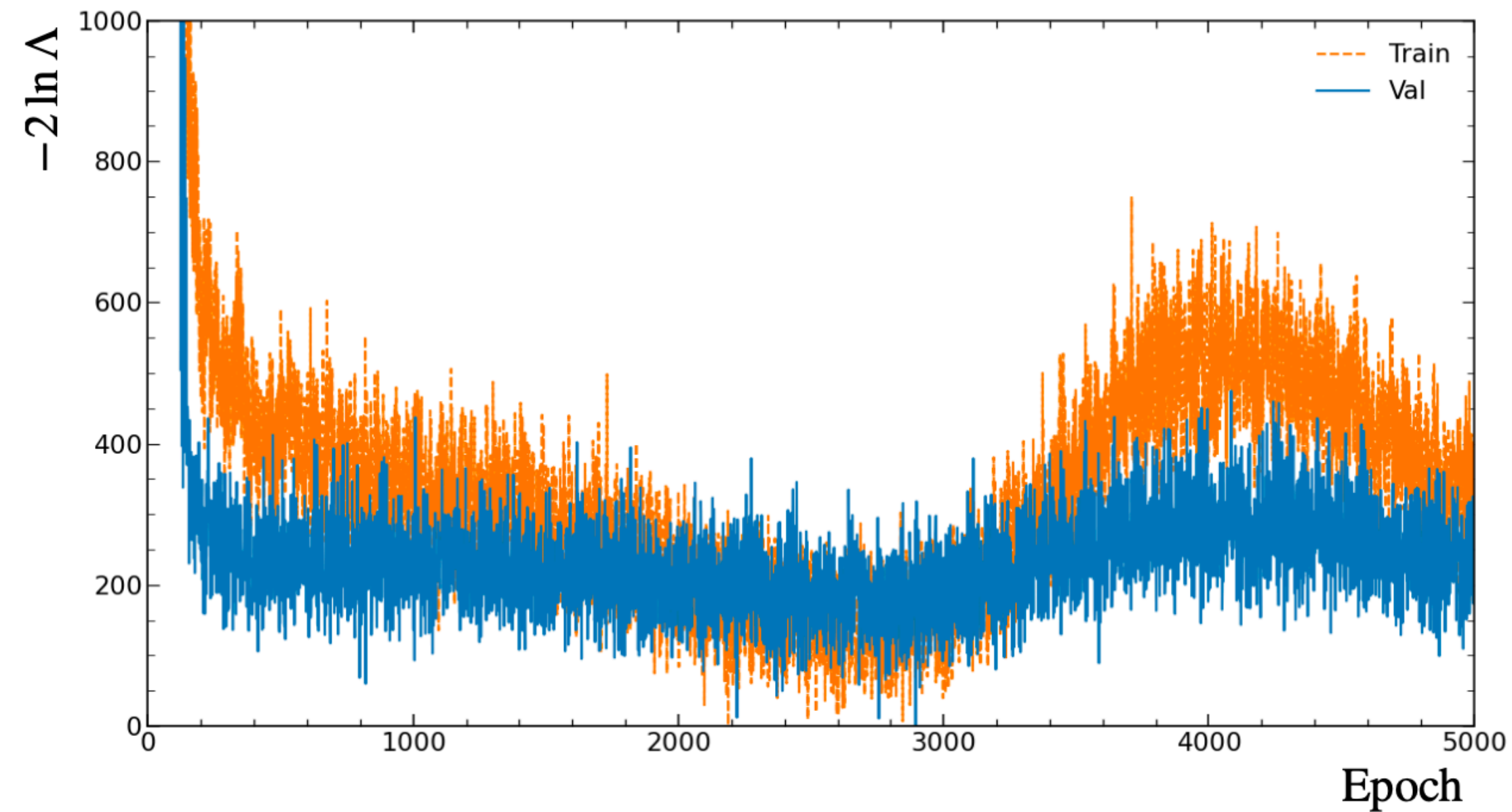
Training GAN

- 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.**

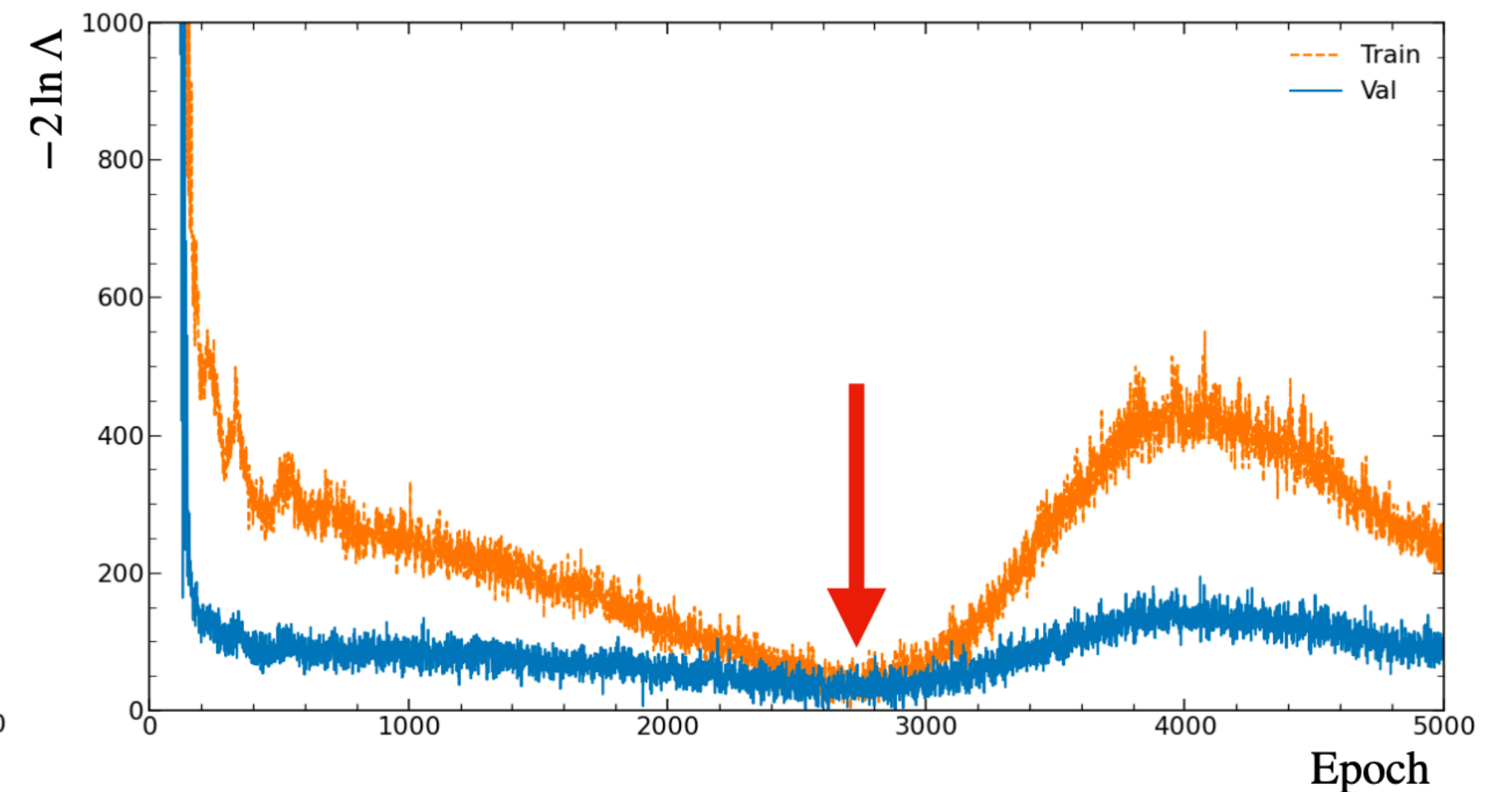


Picking the best model

1 generation per event

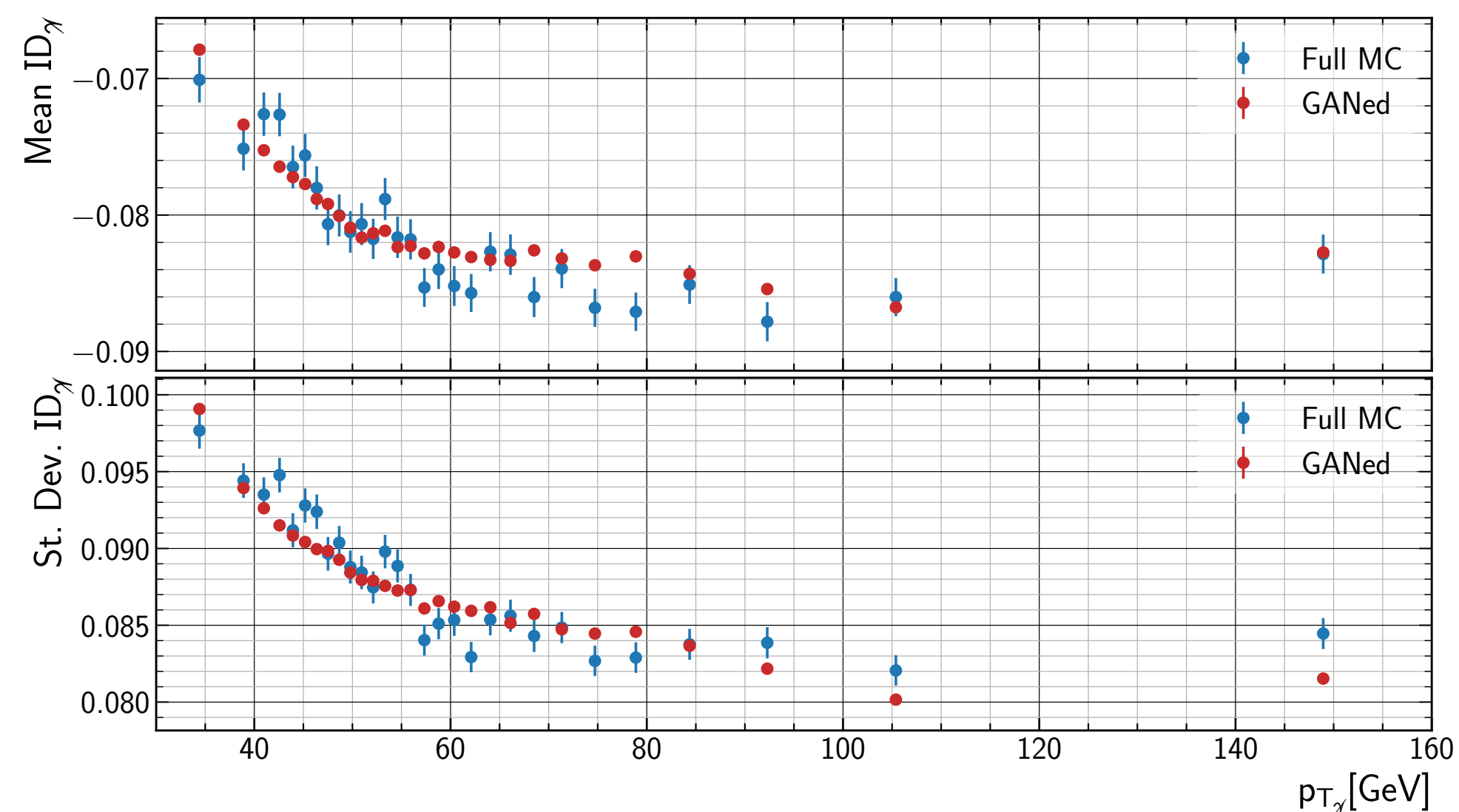
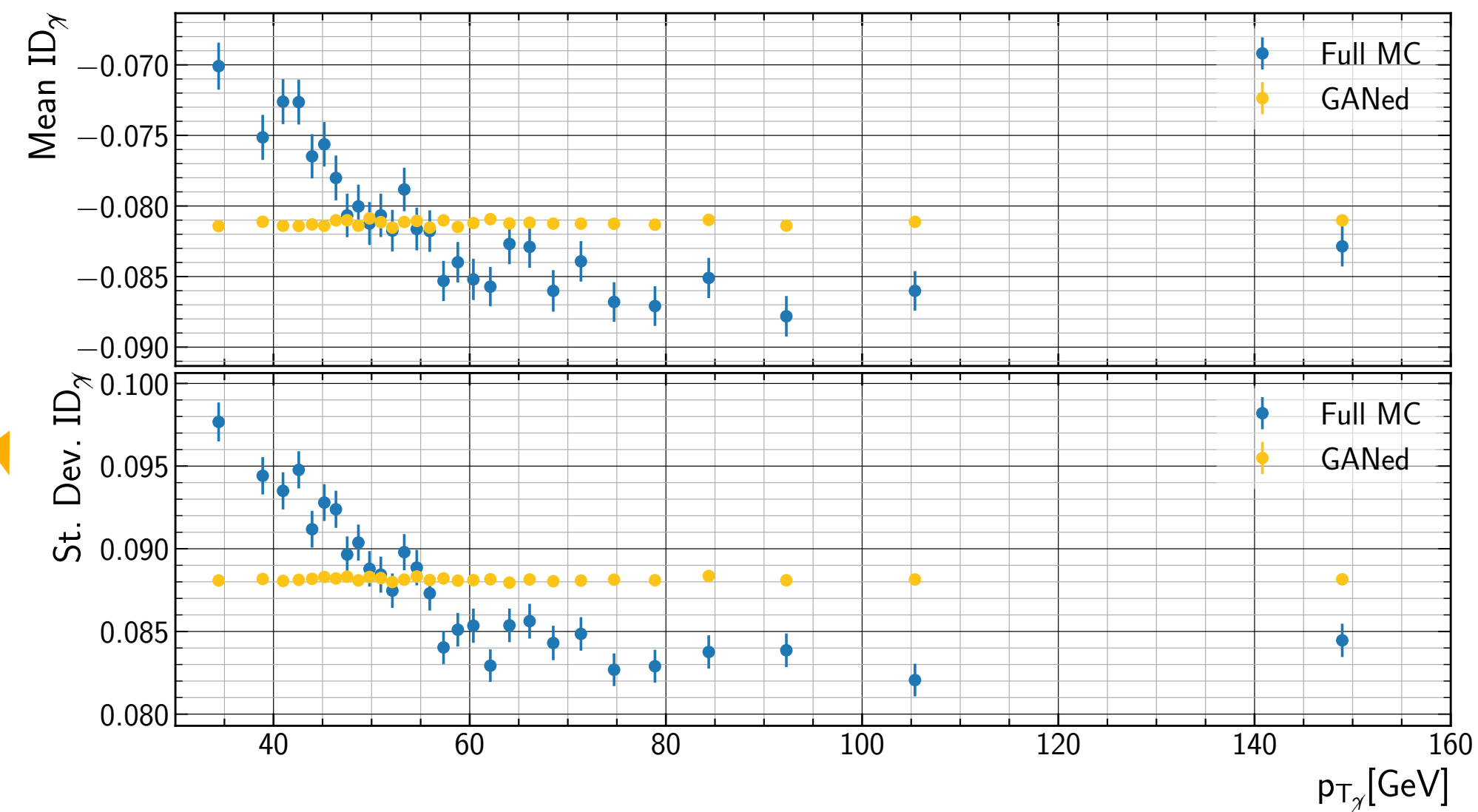
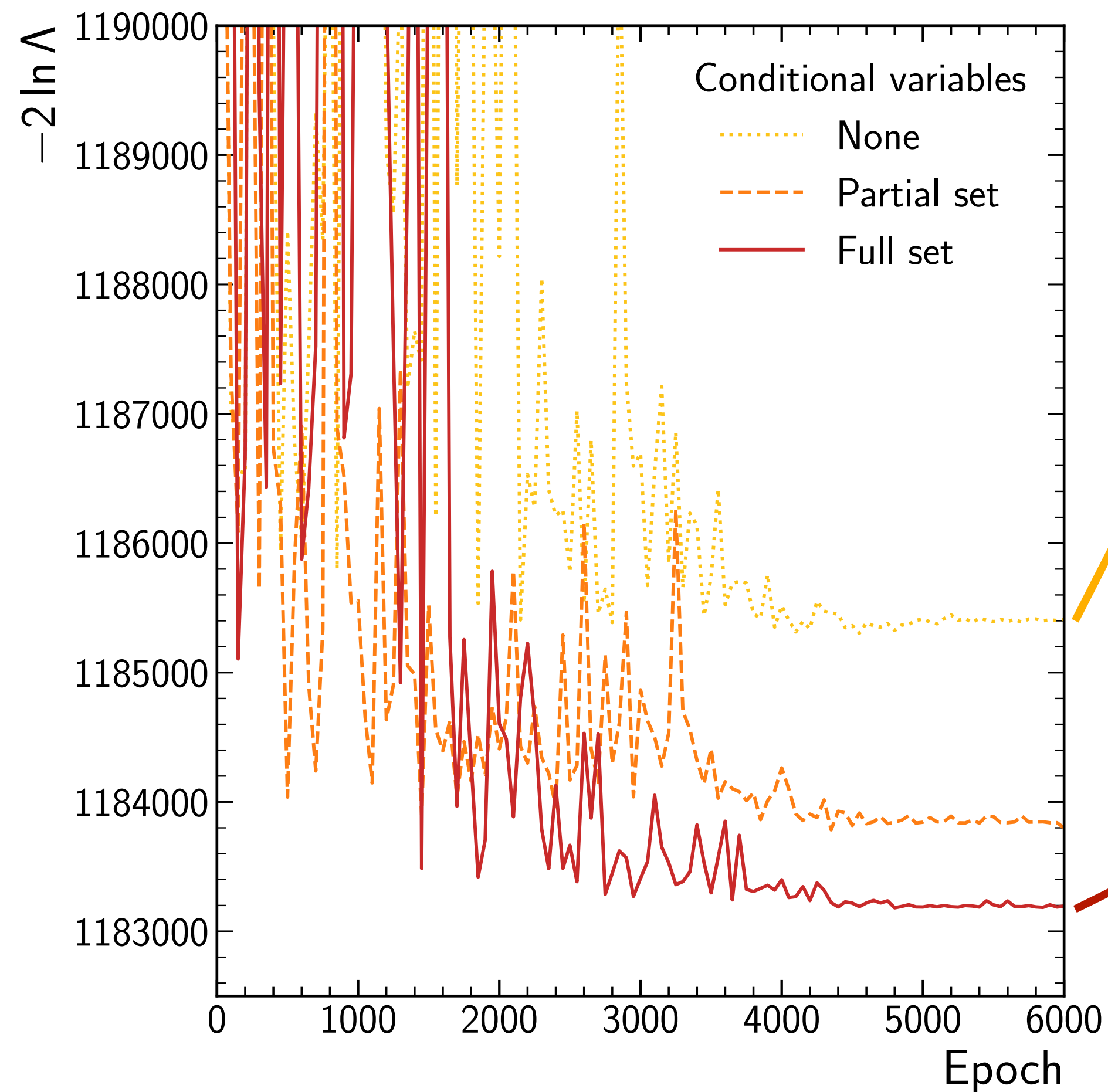


10 generations per event



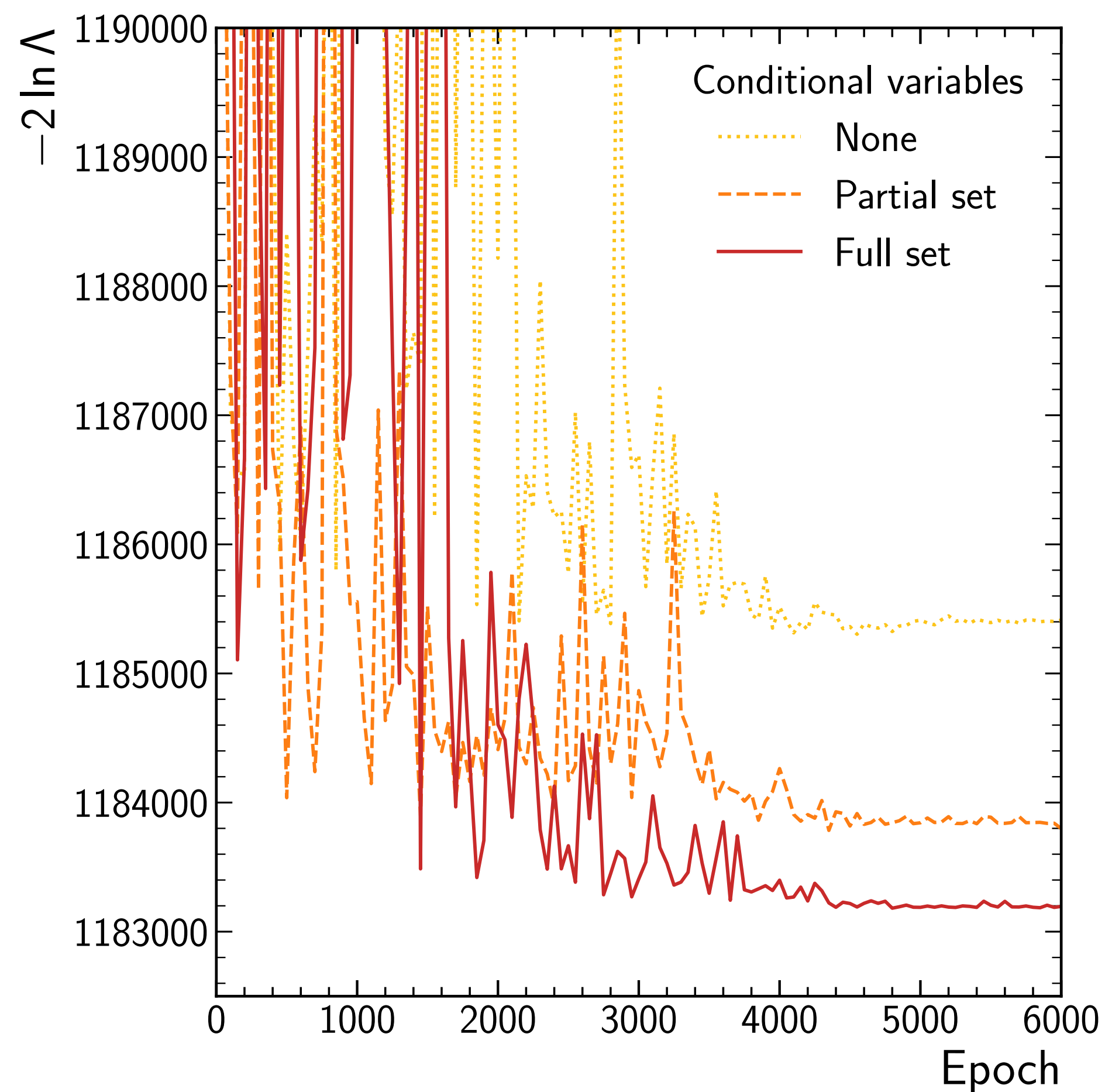
- The performance metric is heavily fluctuating:
 - Significant uncertainty in the selection of the best model.
 - Repeated generation of the events help us reduce these fluctuations.

Why conditional GAN?

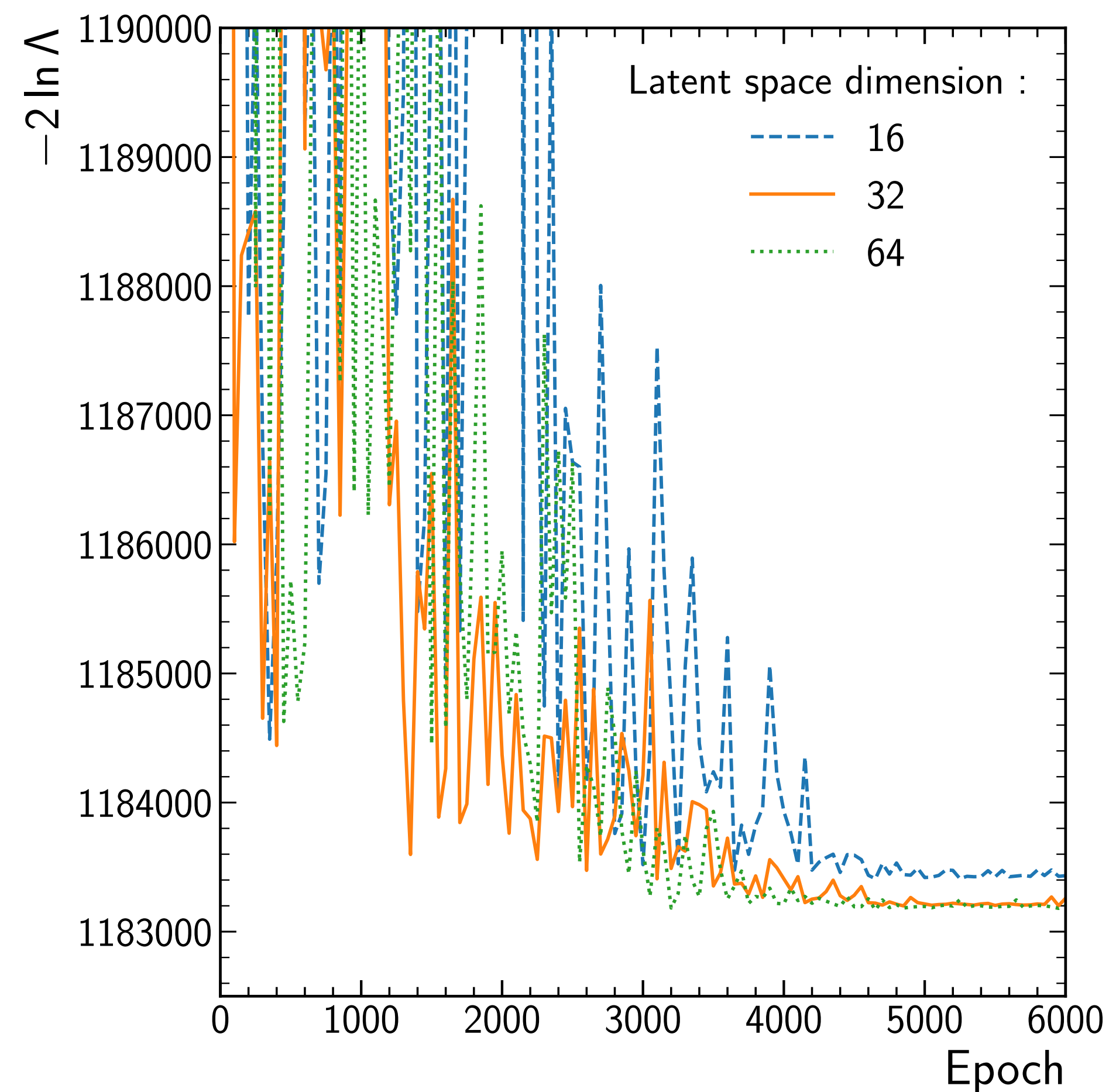


Why conditional GAN?

Impact of the conditional features

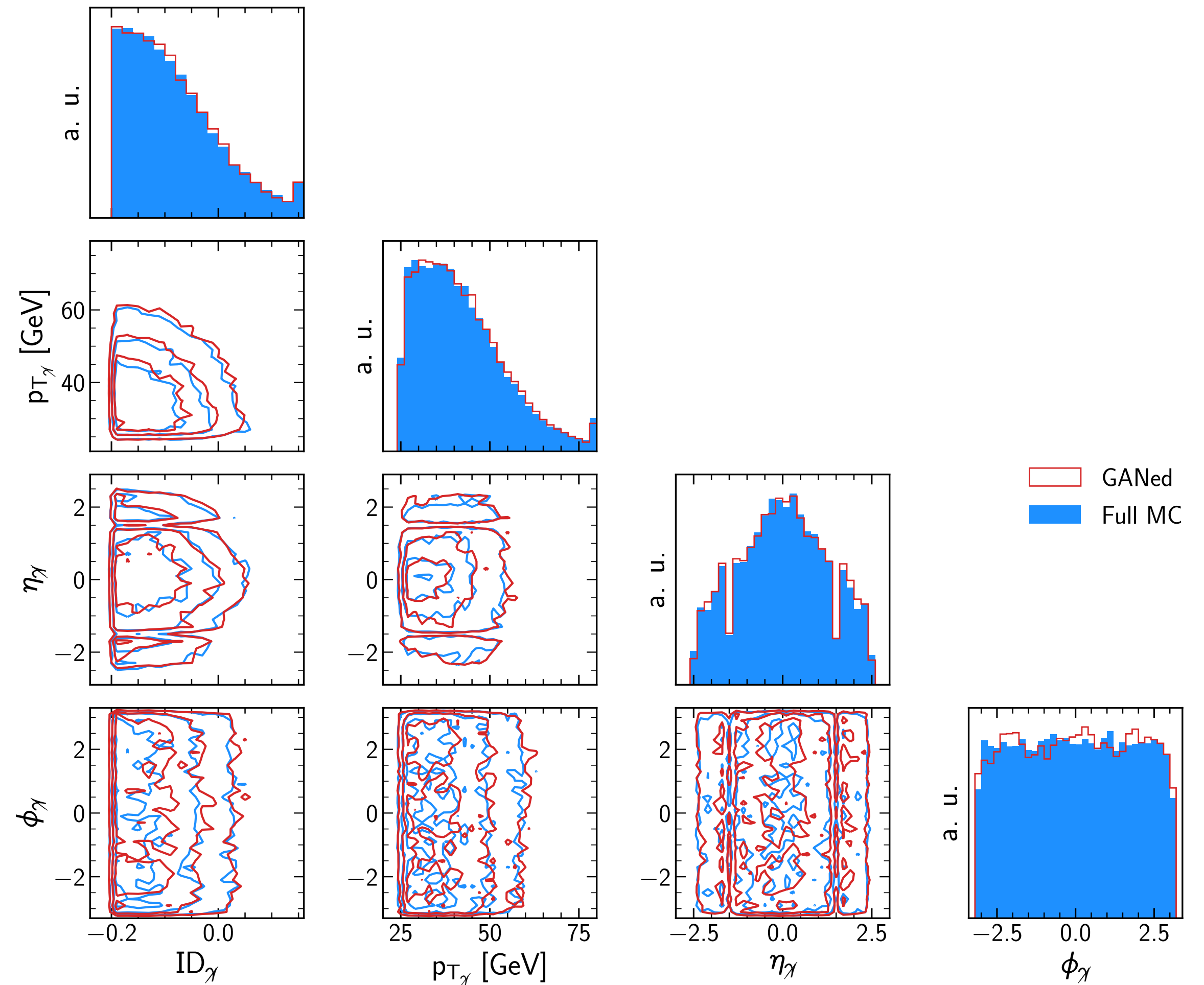


Impact of the random latent space



Putting all together

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



Correlations

$\Delta\phi_{\gamma\gamma}$	-0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.07	0.01	0.08	0.09	0.01	0.01	1.00
$\Delta\eta_{\gamma\gamma}$	-0.01	0.26	0.29	0.01	0.01	0.24	0.51	0.01	0.01	0.01	0.09	0.06	0.13	1.00	0.01
$m_{\gamma\gamma}$	-0.02	0.34	0.02	0.01	0.02	0.34	0.03	0.01	0.09	0.01	0.08	0.15	1.00	0.13	0.01
$p_{T,\gamma\gamma}$	-0.05	0.31	0.03	0.01	0.02	0.74	0.03	0.01	0.41	0.01	0.97	1.00	0.15	0.06	0.09
$\frac{p_{T,\gamma\gamma}}{m_{\gamma\gamma}}$	-0.05	0.38	0.03	0.01	0.02	0.66	0.03	0.01	0.38	0.01	1.00	0.97	0.08	0.09	0.08
N_{Vtx}	-0.02	0.01	0.01	0.01	0.05	0.01	0.01	0.01	0.22	1.00	0.01	0.01	0.01	0.01	0.01
N_{Jets}	-0.01	0.07	0.01	0.01	0.02	0.25	0.01	0.01	1.00	0.22	0.38	0.41	0.09	0.01	0.07
ϕ_γ	-0.01	0.01	0.01	0.68	0.02	0.01	0.01	1.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
η_γ	-0.01	0.02	0.45	0.01	0.08	0.04	1.00	0.01	0.01	0.01	0.03	0.03	0.03	0.51	0.01
$p_{T,\gamma}$	-0.04	0.14	0.03	0.01	0.03	1.00	0.04	0.01	0.25	0.01	0.66	0.74	0.34	0.24	0.02
ID_γ	-0.01	0.02	0.01	0.02	1.00	0.03	0.08	0.02	0.02	0.05	0.02	0.02	0.02	0.01	0.01
ϕ_χ	-0.01	0.01	0.01	1.00	0.02	0.01	0.01	0.68	0.01	0.01	0.01	0.01	0.01	0.01	0.01
η_χ	-0.03	0.09	1.00	0.01	0.01	0.03	0.45	0.01	0.01	0.01	0.03	0.03	0.02	0.29	0.01
$p_{T,\chi}$	-0.05	1.00	0.09	0.01	0.02	0.14	0.02	0.01	0.07	0.01	0.38	0.31	0.34	0.26	0.02
ID_χ	1.00	0.05	0.03	0.01	0.01	0.04	0.01	0.01	0.01	0.02	0.05	0.05	0.02	0.01	0.01
	ID_γ	$p_{T,\gamma}$	η_γ	ϕ_γ	ID_γ	$p_{T,\gamma}$	η_γ	ϕ_γ	N_{Jets}	N_{Vtx}	$\frac{p_{T,\gamma\gamma}}{m_{\gamma\gamma}}$	$\frac{p_{T,\gamma\gamma}}{m_{\gamma\gamma}}$	$p_{T,\gamma\gamma}$	$m_{\gamma\gamma}$	$\Delta\eta_{\gamma\gamma}$

$\Delta\phi_{\gamma\gamma}$	-0.01	0.03	0.00	0.01	0.01	0.02	0.01	0.01	0.07	0.01	0.09	0.09	0.01	0.02	1.00
$\Delta\eta_{\gamma\gamma}$	-0.02	0.26	0.28	0.01	0.01	0.24	0.50	0.01	0.02	0.01	0.09	0.07	0.13	1.00	0.02
$m_{\gamma\gamma}$	-0.02	0.32	0.01	0.01	0.02	0.33	0.03	0.00	0.07	0.00	0.10	0.12	1.00	0.13	0.01
$p_{T,\gamma\gamma}$	-0.06	0.32	0.02	0.01	0.02	0.73	0.02	0.01	0.42	0.02	0.97	1.00	0.12	0.07	0.09
$\frac{p_{T,\gamma\gamma}}{m_{\gamma\gamma}}$	-0.06	0.39	0.02	0.01	0.02	0.66	0.03	0.01	0.40	0.02	1.00	0.97	0.10	0.09	0.09
N_{Vtx}	-0.01	0.01	0.01	0.02	0.04	0.01	0.01	0.01	0.22	1.00	0.02	0.02	0.00	0.01	0.01
N_{Jets}	-0.01	0.07	0.01	0.02	0.02	0.25	0.01	0.01	1.00	0.22	0.40	0.42	0.07	0.02	0.07
ϕ_γ	-0.01	0.01	0.01	0.68	0.02	0.01	0.01	1.00	0.01	0.01	0.01	0.01	0.00	0.01	0.01
η_γ	-0.01	0.02	0.47	0.01	0.07	0.04	1.00	0.01	0.01	0.01	0.03	0.02	0.03	0.50	0.01
$p_{T,\gamma}$	-0.05	0.13	0.02	0.01	0.02	1.00	0.04	0.01	0.25	0.01	0.66	0.73	0.33	0.24	0.02
ID_γ	-0.01	0.02	0.01	0.03	1.00	0.02	0.07	0.02	0.02	0.04	0.02	0.02	0.02	0.01	0.01
ϕ_χ	-0.01	0.01	0.01	1.00	0.03	0.01	0.01	0.68	0.02	0.02	0.01	0.01	0.01	0.01	0.01
η_χ	-0.03	0.08	1.00	0.01	0.01	0.02	0.47	0.01	0.01	0.01	0.02	0.02	0.01	0.28	0.00
$p_{T,\chi}$	-0.05	1.00	0.08	0.01	0.02	0.13	0.02	0.01	0.07	0.01	0.39	0.32	0.32	0.26	0.03
ID_χ	1.00	0.05	0.03	0.01	0.01	0.05	0.01	0.01	0.01	0.01	0.06	0.06	0.02	0.02	0.01
	ID_γ	$p_{T,\gamma}$	η_γ	ϕ_γ	ID_γ	$p_{T,\gamma}$	η_γ	ϕ_γ	N_{Jets}	N_{Vtx}	$\frac{p_{T,\gamma\gamma}}{m_{\gamma\gamma}}$	$\frac{p_{T,\gamma\gamma}}{m_{\gamma\gamma}}$	$p_{T,\gamma\gamma}$	$m_{\gamma\gamma}$	$\Delta\eta_{\gamma\gamma}$

- 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.

