Data driven background estimation in HEP using Generative Adversarial Networks

Moriond EW 2023 - Young Scientist Forum - March 23rd, 2023

Victor Lohezic (victor.lohezic@cern.ch)
Fabrice Couderc, Julie Malclès, Özgür Şahin
IRFU - CEA Saclay

https://arxiv.org/abs/2212.03763

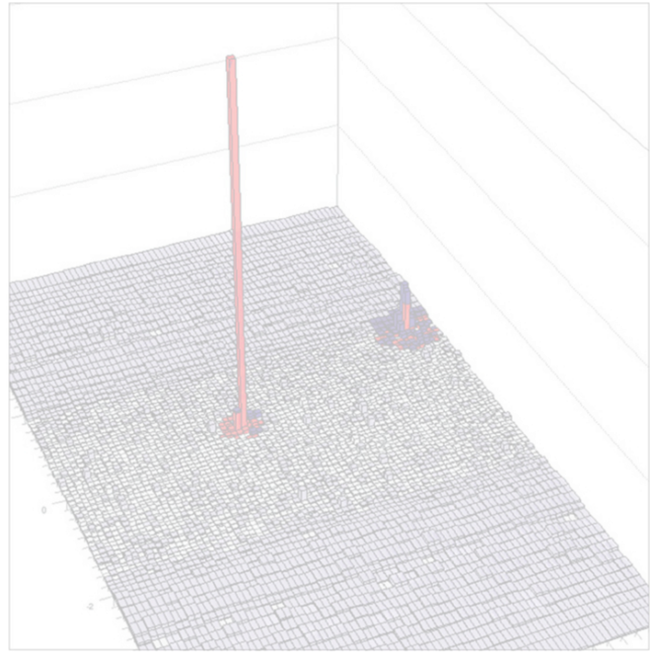
Publication accepted by EPJC

A data-driven estimation of background

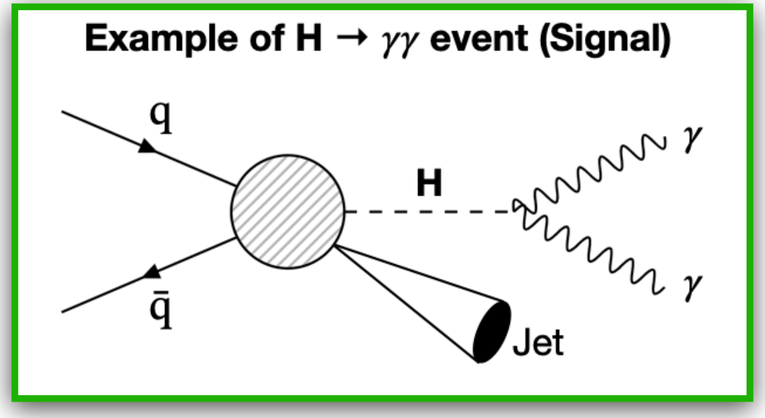
GAN based data-driven technique to estimate background processes with a misidentified object in collider events.

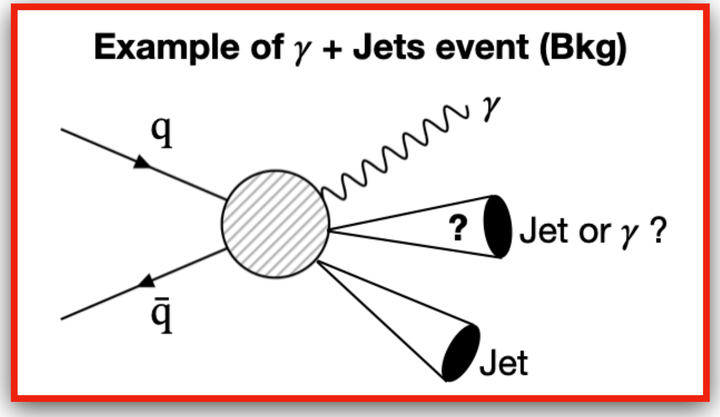
Showcase on the background of the H→γγ analysis thanks to a CMS Open Simulation

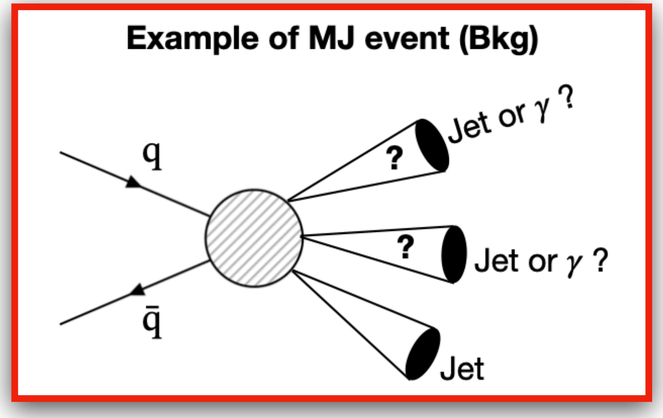
- Dominant backgrounds : γγ+Jets, γ+Jets, MultiJets (MJ)
- To distinguish between photons from primary interaction (prompt γ) and photons from jets hadronisation (non-prompt γ), reconstructed photons are given a score : the photonID (computed from shower shape and isolation variables)
- Strategy of the analysis is to train discriminants to separate background from signal and photonID is one of the key variables. However MC/Data agreement for γ+Jets and MJ samples is not satisfying and statistics is low…



Event display from CMS illustrating differences in shower shape between prompt (left) and non-prompt (right) photons





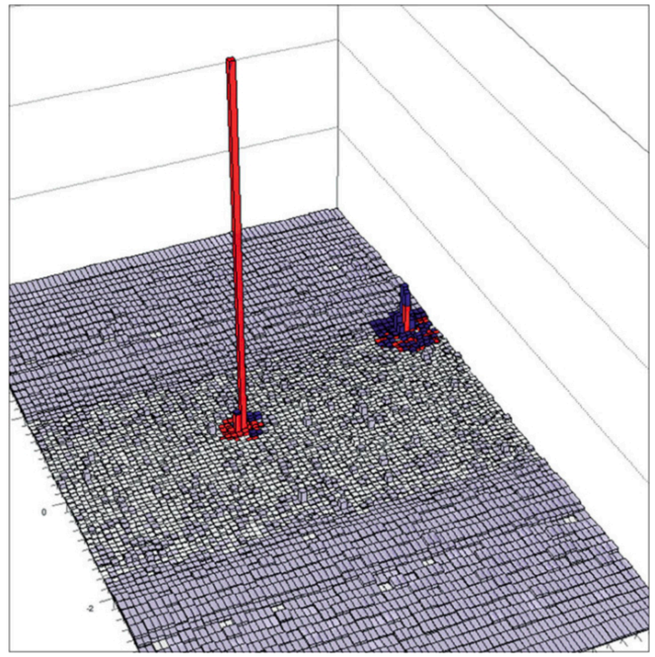


A data-driven estimation of background

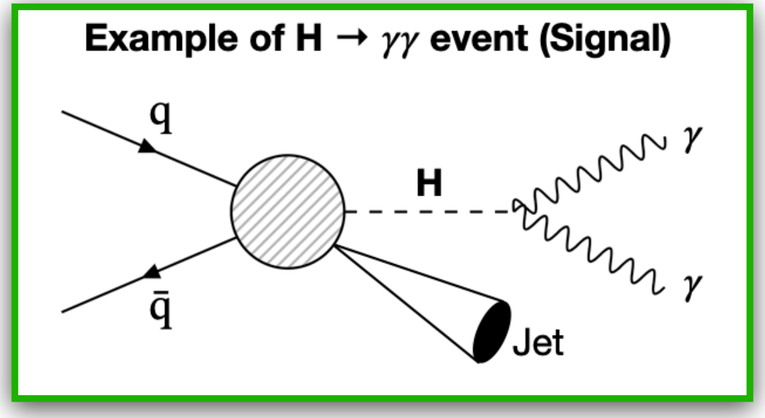
GAN based data-driven technique to estimate background processes with a misidentified object in collider events.

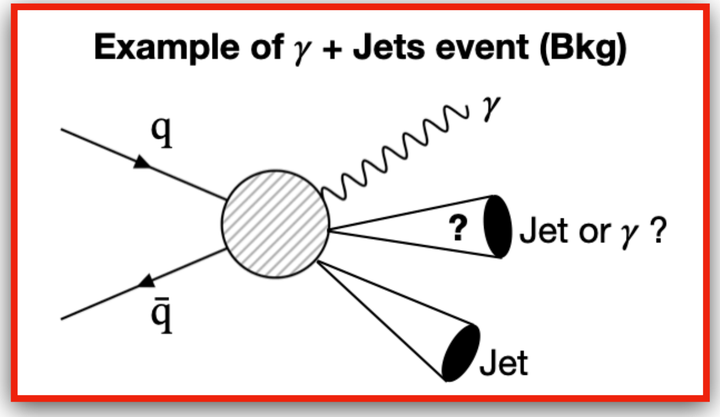
Showcase on the background of the H→γγ analysis thanks to a CMS Open Simulation

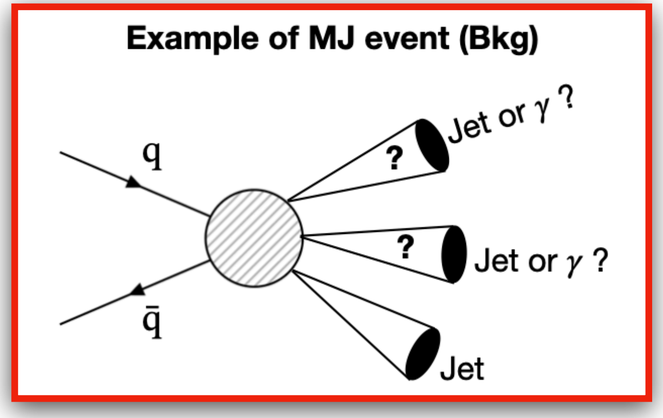
- Dominant backgrounds : γγ+Jets, γ+Jets, MultiJets (MJ)
- To distinguish between photons from primary interaction (prompt γ) and photons from jets hadronisation (non-prompt ~~γ~~), reconstructed photons are given a score : the photonID (computed from shower shape and isolation variables)
- Strategy of the analysis is to train discriminants to separate background from signal and photonID is one of the key variables. However MC/Data agreement for γ+Jets and MJ samples is not satisfying and statistics is low…

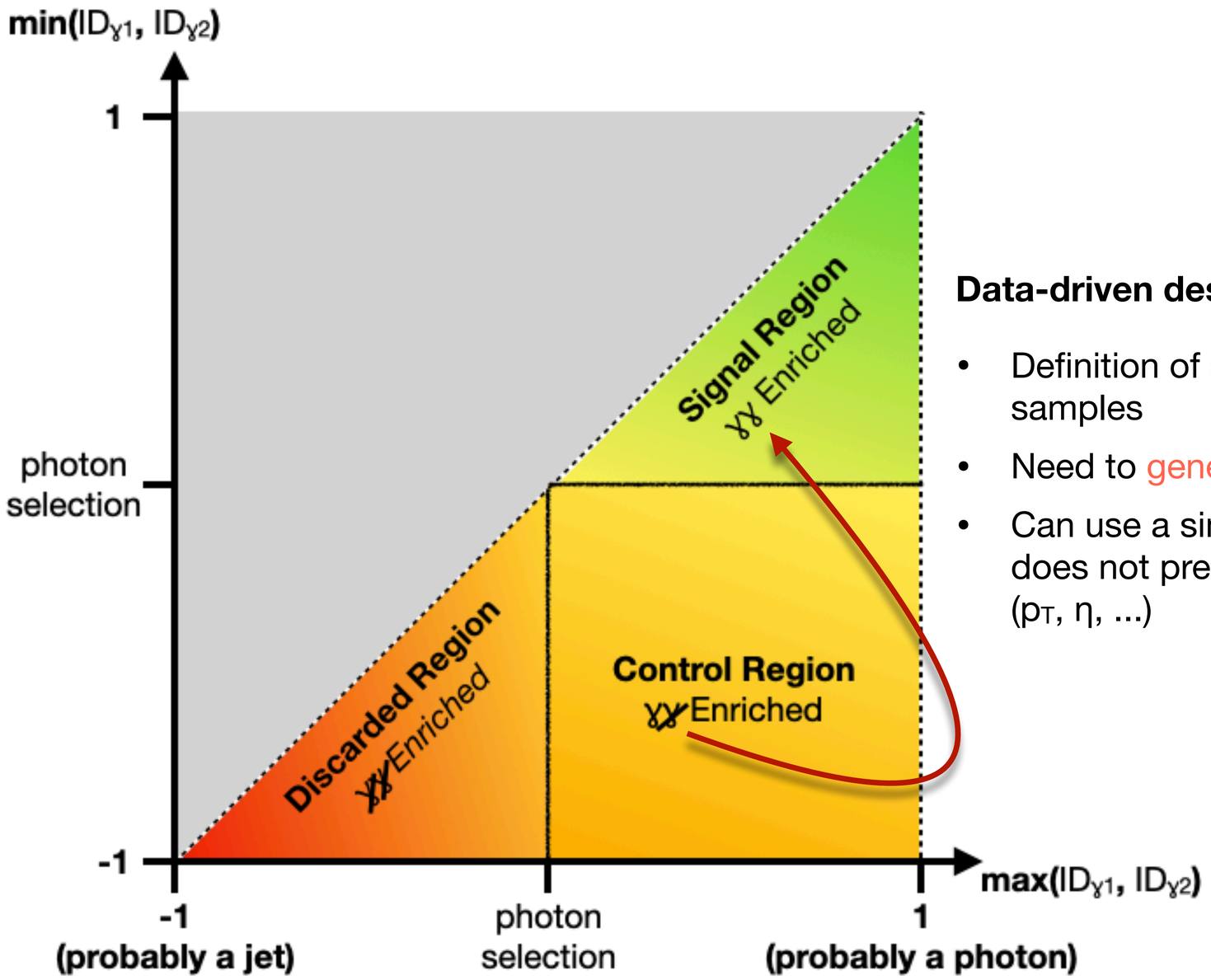


Event display from CMS illustrating differences in shower shape between prompt (left) and non-prompt (right) photons





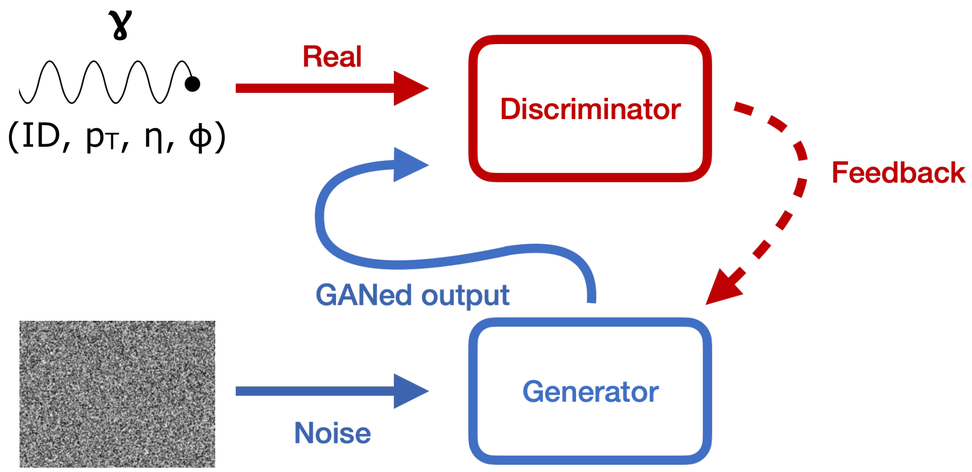




Data-driven description of background

- Definition of a control region in data to replace MC γ+Jets and MJ samples
- Need to generate a new min. photonID
- Can use a simple generation of min. photonID from a PDF but this does not preserve correlation with other observables in the event (p_T, η, ...)

GAN and optimisation procedure



Generative Adversarial Networks (GAN)

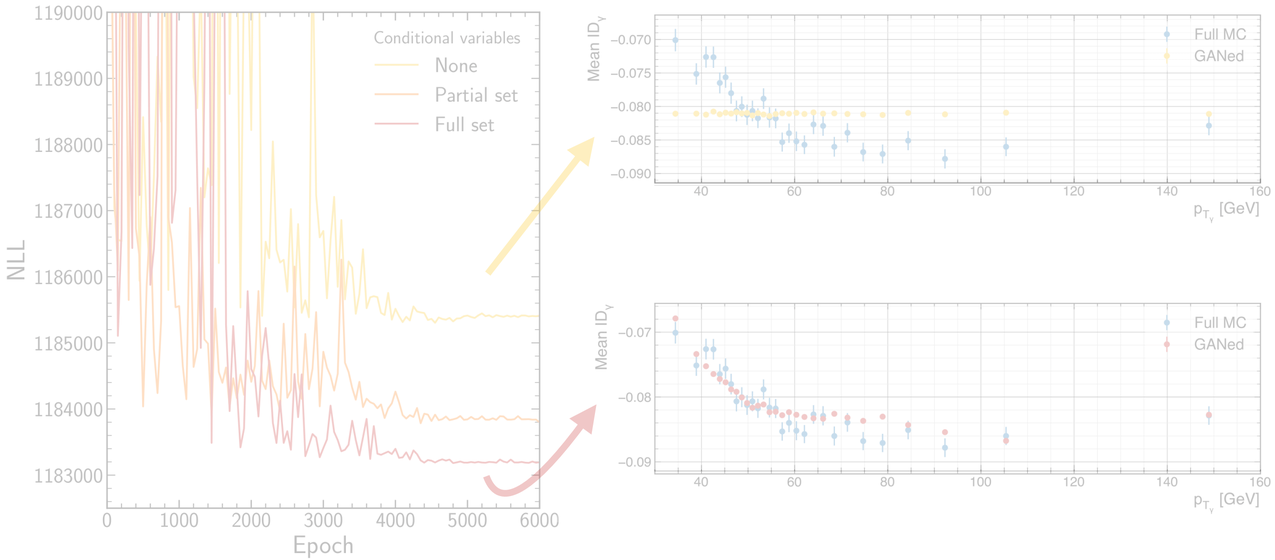
Goodfellow et al. suggested a model consisting of two neural networks competing against each other :

- the "discriminator" sorts samples between real and generated ones - i.e. discriminates fakes
- the "generator" tries to produce samples which will fool the discriminator

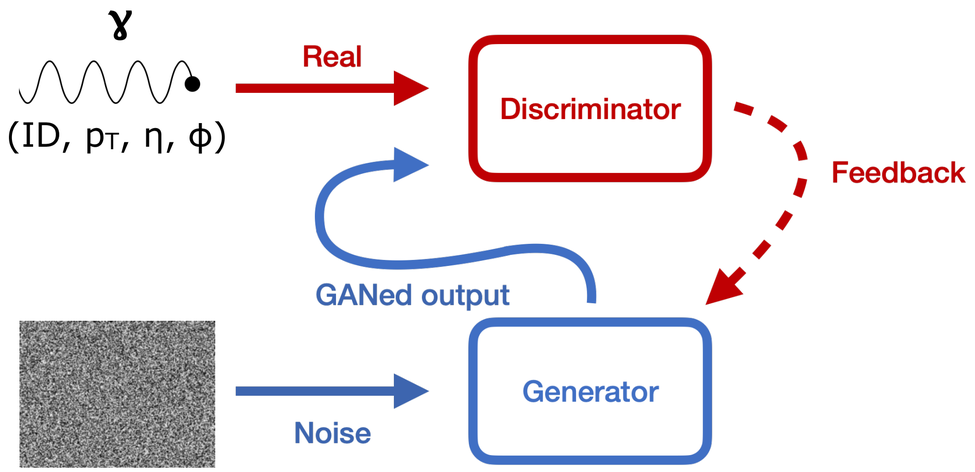
In our case we use a conditional version of a GAN and we train on the full non-prompt photon (ID, pT, η, ϕ)

➡ Due to nature of GAN, we need a metric in addition to the training loss. We define a Log-Likelihood metric which takes into account the correlations to monitor the performance of the GAN.

➡ Scan of many parameters of the training to find the optimal set



GAN and optimisation procedure



Generative Adversarial Networks (GAN)

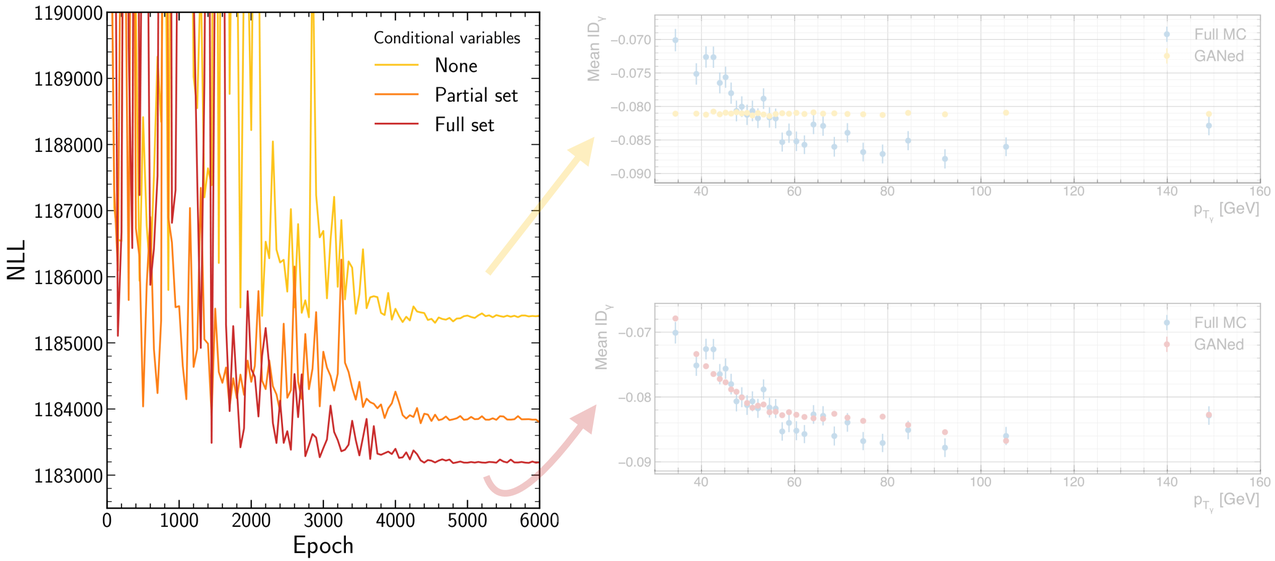
Goodfellow et al. suggested a model consisting of two neural networks competing against each other :

- the "discriminator" sorts samples between real and generated ones - i.e. discriminates fakes
- the "generator" tries to produce samples which will fool the discriminator

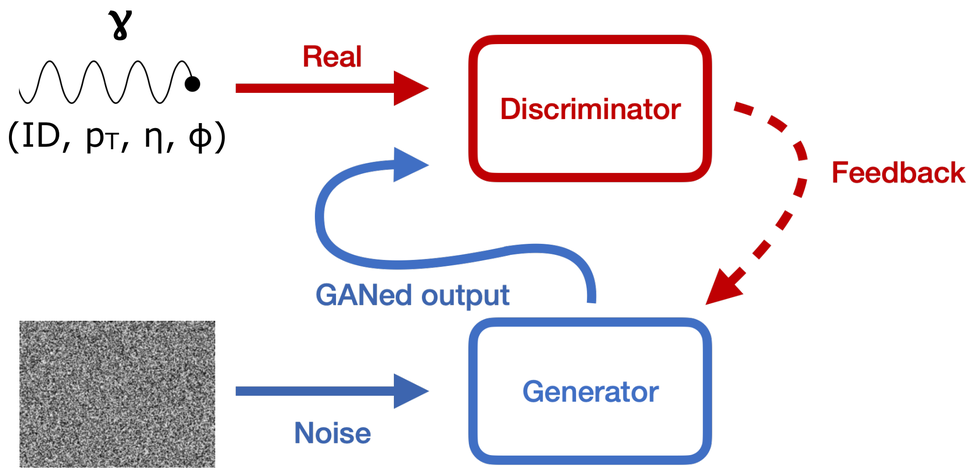
In our case we use a conditional version of a GAN and we train on the full non-prompt photon (ID, pT, η, ϕ)

➡ Due to nature of GAN, we need a metric in addition to the training loss. We define a Log-Likelihood metric which takes into account the correlations to monitor the performance of the GAN.

➡ Scan of many parameters of the training to find the optimal set



GAN and optimisation procedure



Generative Adversarial Networks (GAN)

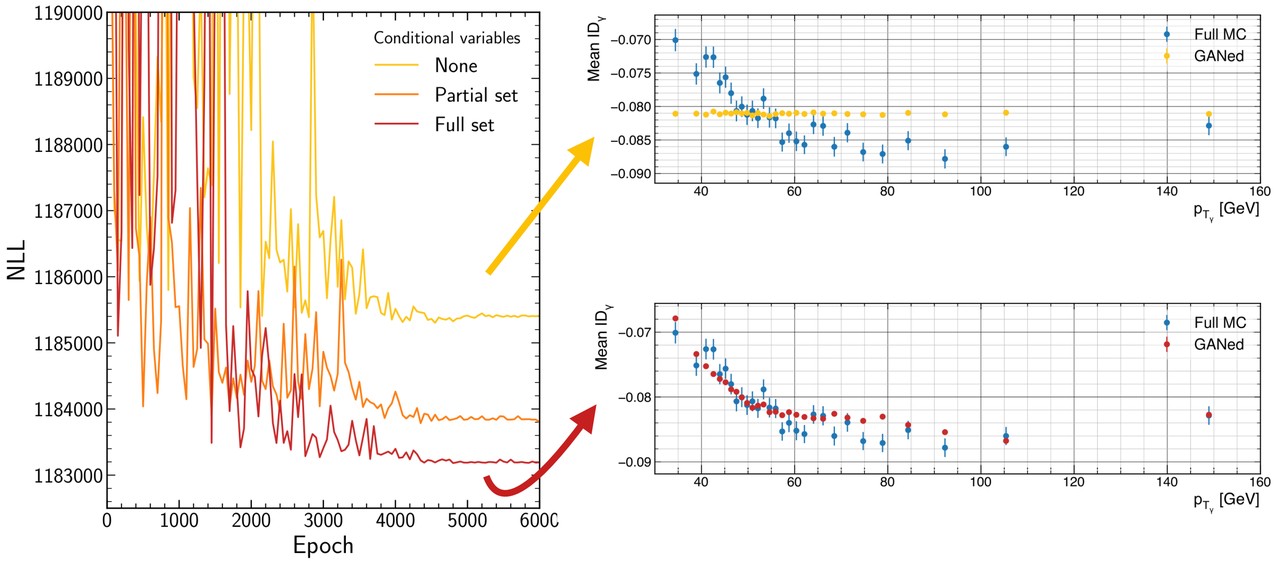
Goodfellow et al. suggested a model consisting of two neural networks competing against each other :

- the "discriminator" sorts samples between real and generated ones - i.e. discriminates fakes
- the "generator" tries to produce samples which will fool the discriminator

In our case we use a conditional version of a GAN and we train on the full non-prompt photon (ID, pT, η, ϕ)

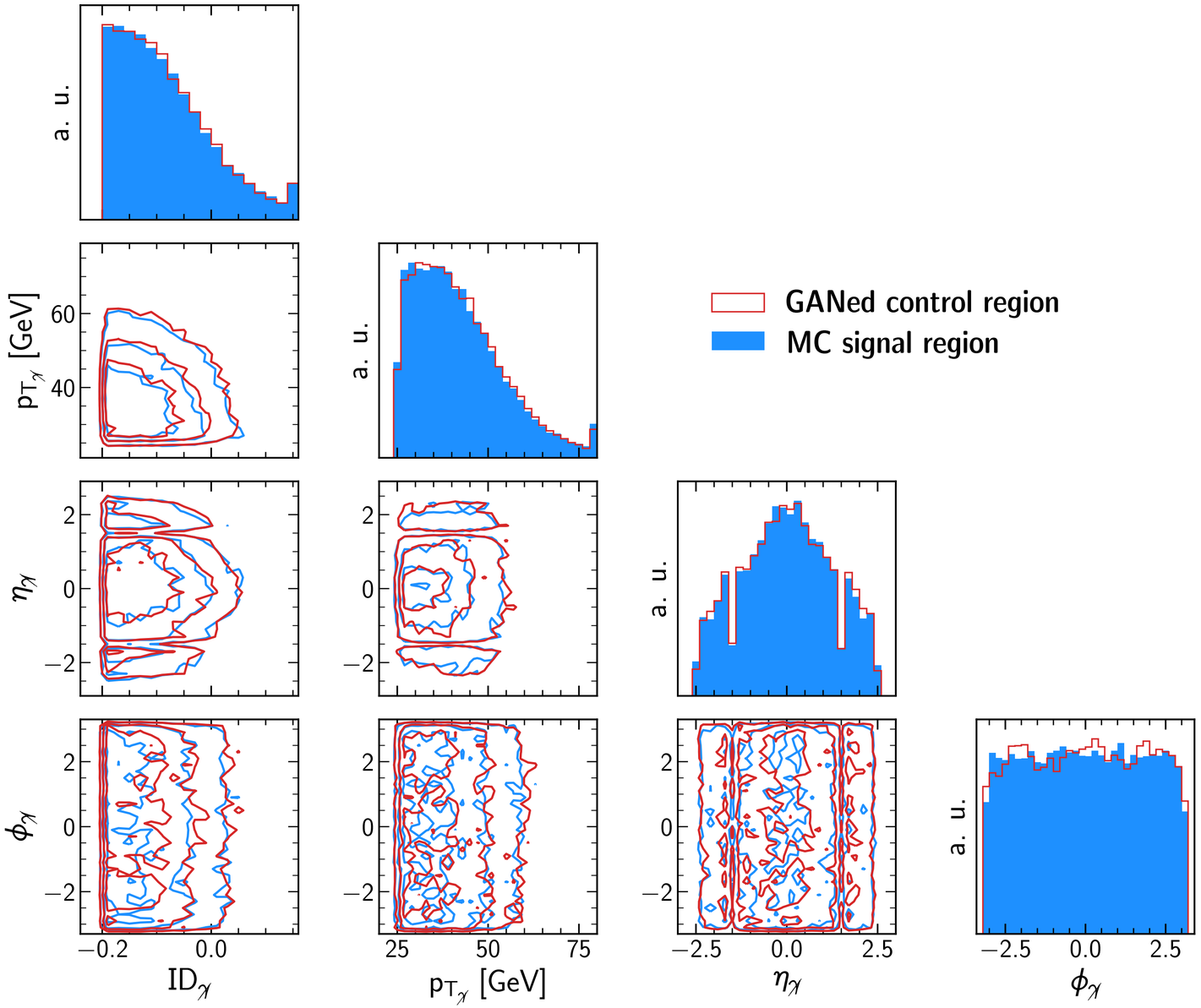
➡ Due to nature of GAN, we need a metric in addition to the training loss. We define a Log-Likelihood metric which takes into account the correlations to monitor the performance of the GAN.

➡ Scan of many parameters of the training to find the optimal set



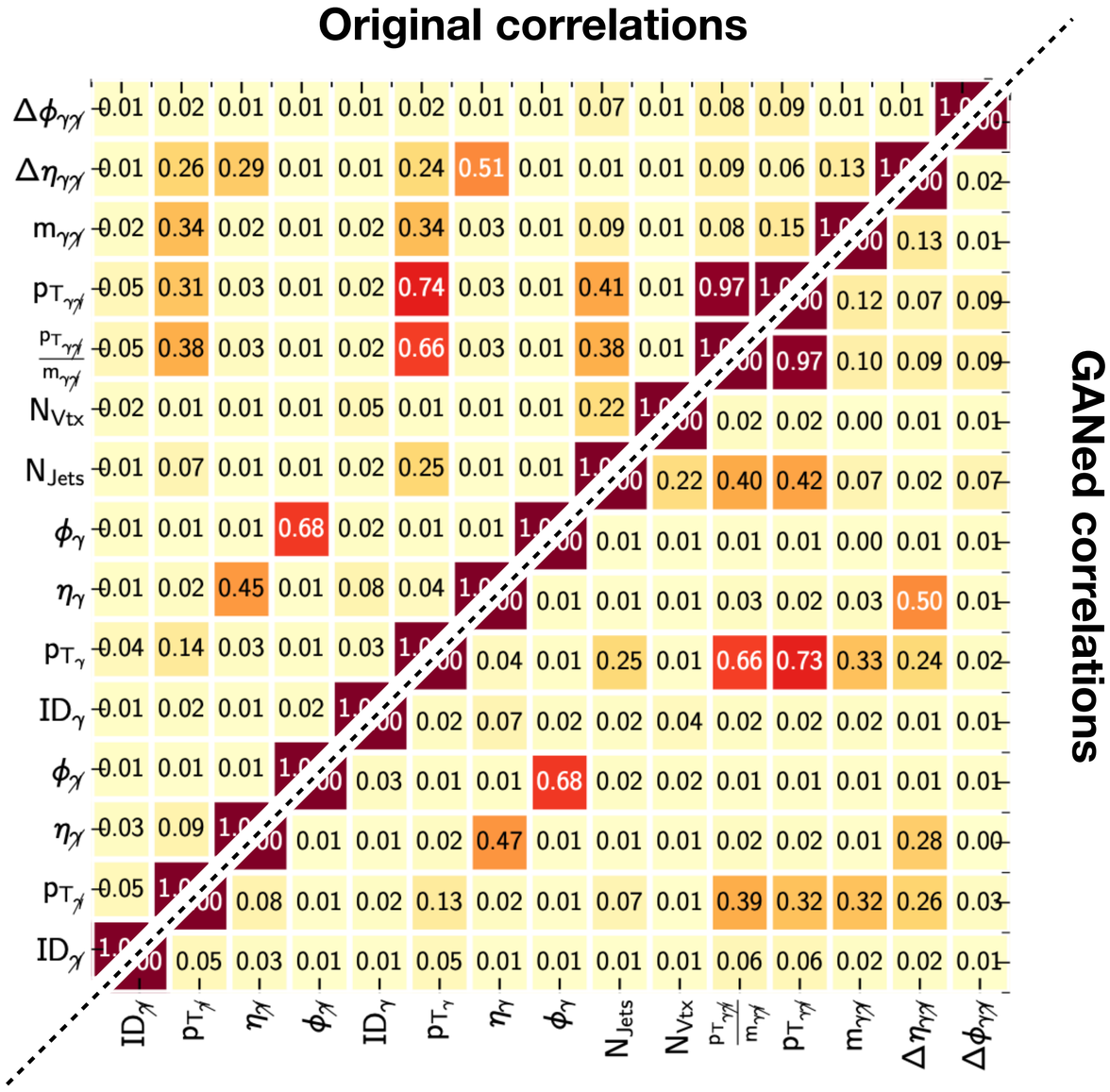
Results

- GAN is able to generate a full misidentified object that would pass the selection criteria *(see 1D distributions on diagonal)*
- GAN learns correlations between observables of the objects *(see contours on off-diagonal plots)* but also correlations with the rest of the event *(see distance correlation coefficients matrix)*
- This method could be used as a general tool to generate other objects for other use cases



Results

- GAN is able to generate a full misidentified object that would pass the selection criteria *(see 1D distributions on diagonal)*
- GAN learns correlations between observables of the objects *(see contours on off-diagonal plots)* but also correlations with the rest of the event *(see distance correlation coefficients matrix)*
- This method could be used as a general tool to generate other objects for other use cases



Data driven background estimation in HEP using Generative Adversarial Networks

Moriond EW 2023 - Young Scientist Forum - March 23rd, 2023

Victor Lohezic (victor.lohezic@cern.ch)
Fabrice Couderc, Julie Malclès, Özgür Şahin
IRFU - CEA Saclay

Thank you !

https://arxiv.org/abs/2212.03763

Publication accepted by EPJC