Generative Models and Calorimeter Fast Simulation for the LHCb

Fedor Ratnikov for the team
HSF Simulation Meeting
Mar. 6, 2019

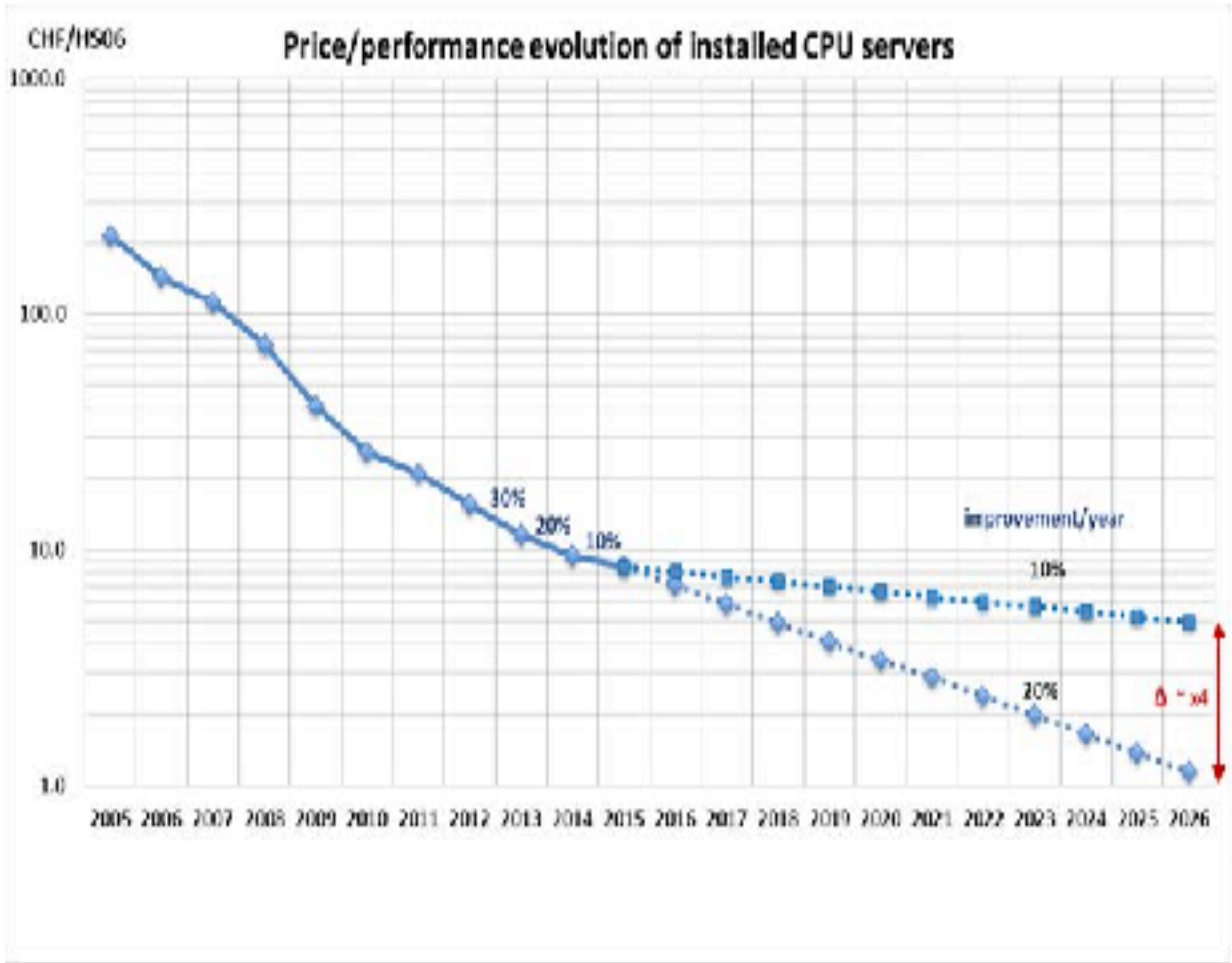
NRU Higher School of Economics,
Yandex School of Data Analysis

Library Approach

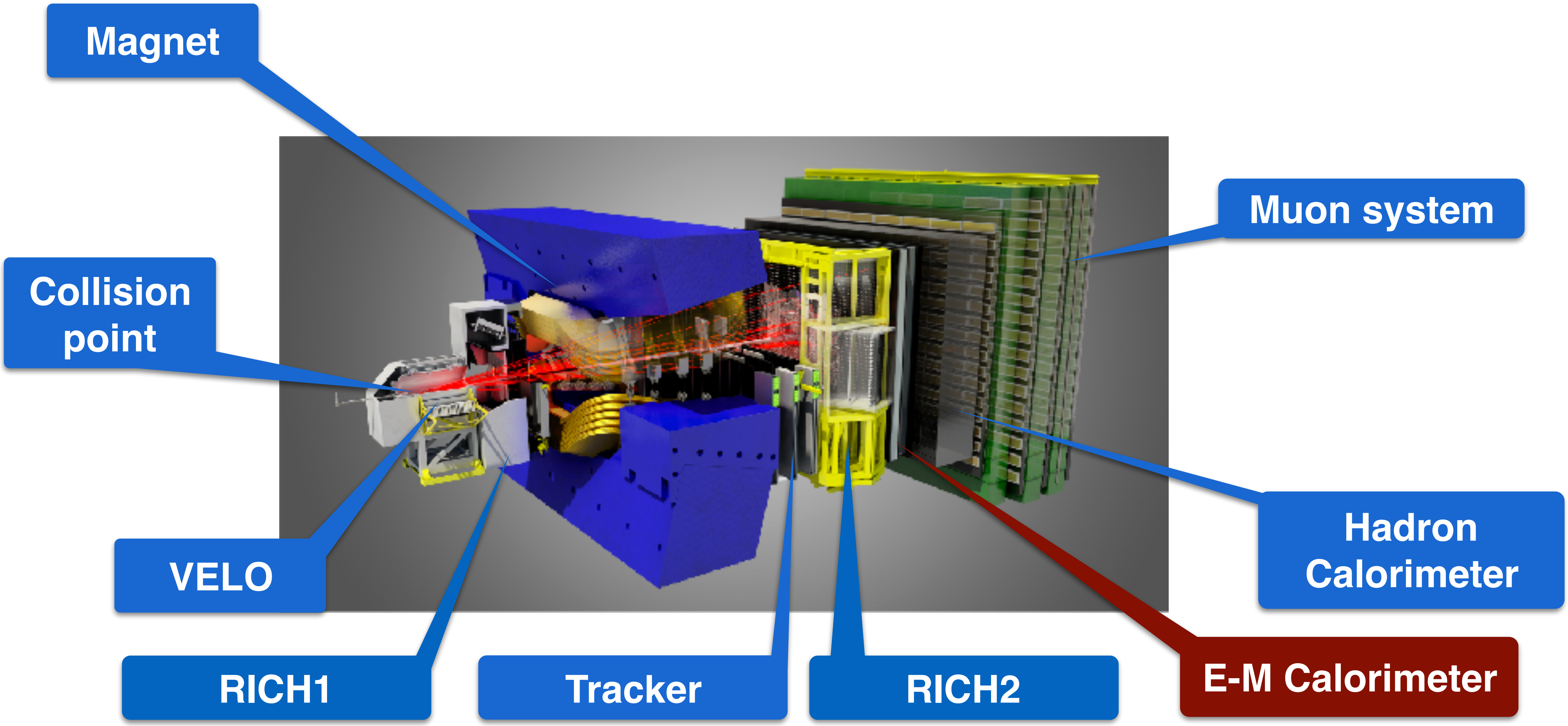
- We have train sample for the generative model
  - consistency with this train sample is a figure of merit for the generative model
- Objects of the train sample may be used for generation directly
  - remember KNN classification algorithm
  - k=1 - straightforward
    - the only drawback - search for the object with appropriate conditions in the (presumably huge) data library
  - k>1 - problem to interpolate between objects
    - short distance objects interpolation, more robust than global generation
- NB: this approach by construction uses full information which is contained in the training sample

Generative Models at LHC

- About 80% of computing resources are used for MC simulation in HEP experiments
  - Calorimeter simulation is one of bottlenecks
  - RICH is the next in the row for LHCb detector
    - > 85% of simulation is taken by these
- Can not expect exponential rise of CPU performance
- Need work around for Run3 and HL-LHC
- Generative models trained on the detailed GEANT simulation may be a solution



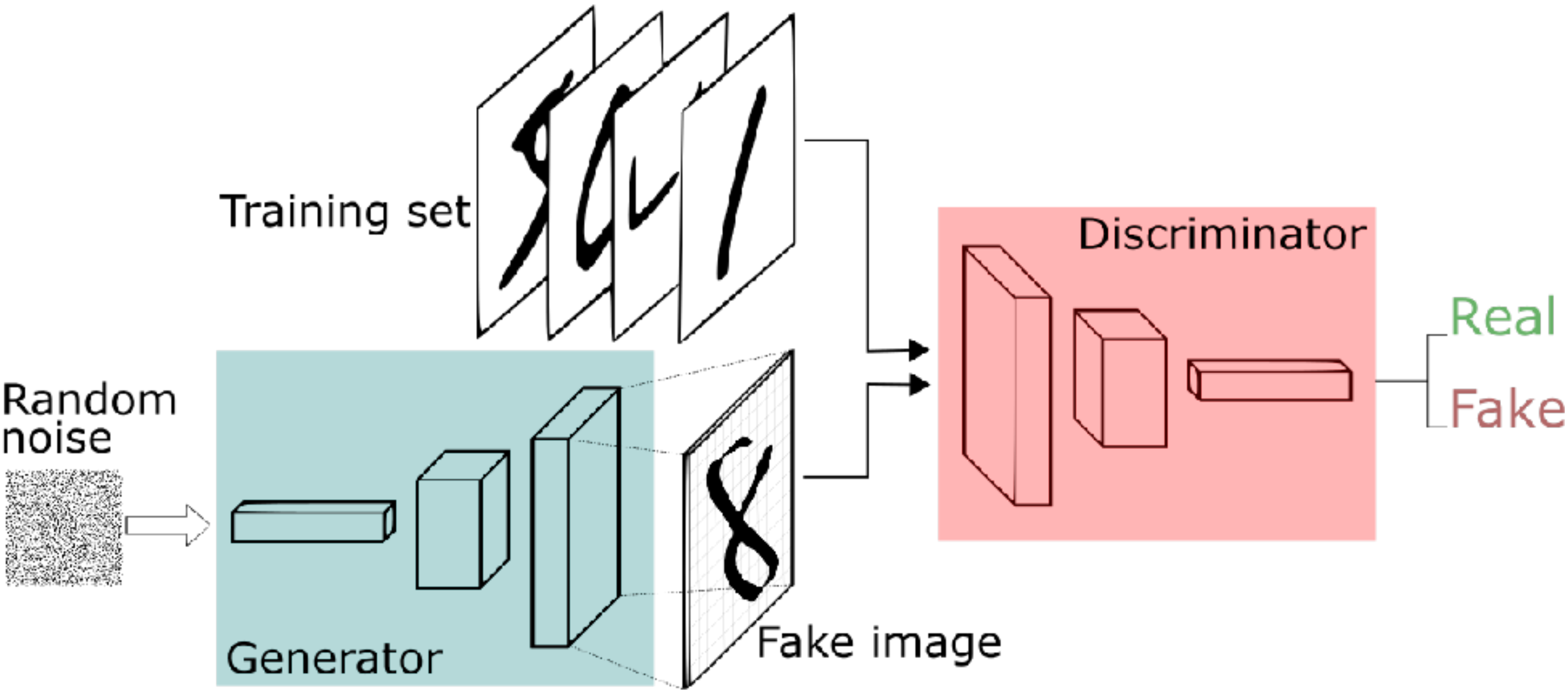
Example: Fast Simulation of the ECAL Response



- ECAL takes the most time in the LHCb event simulation

GAN

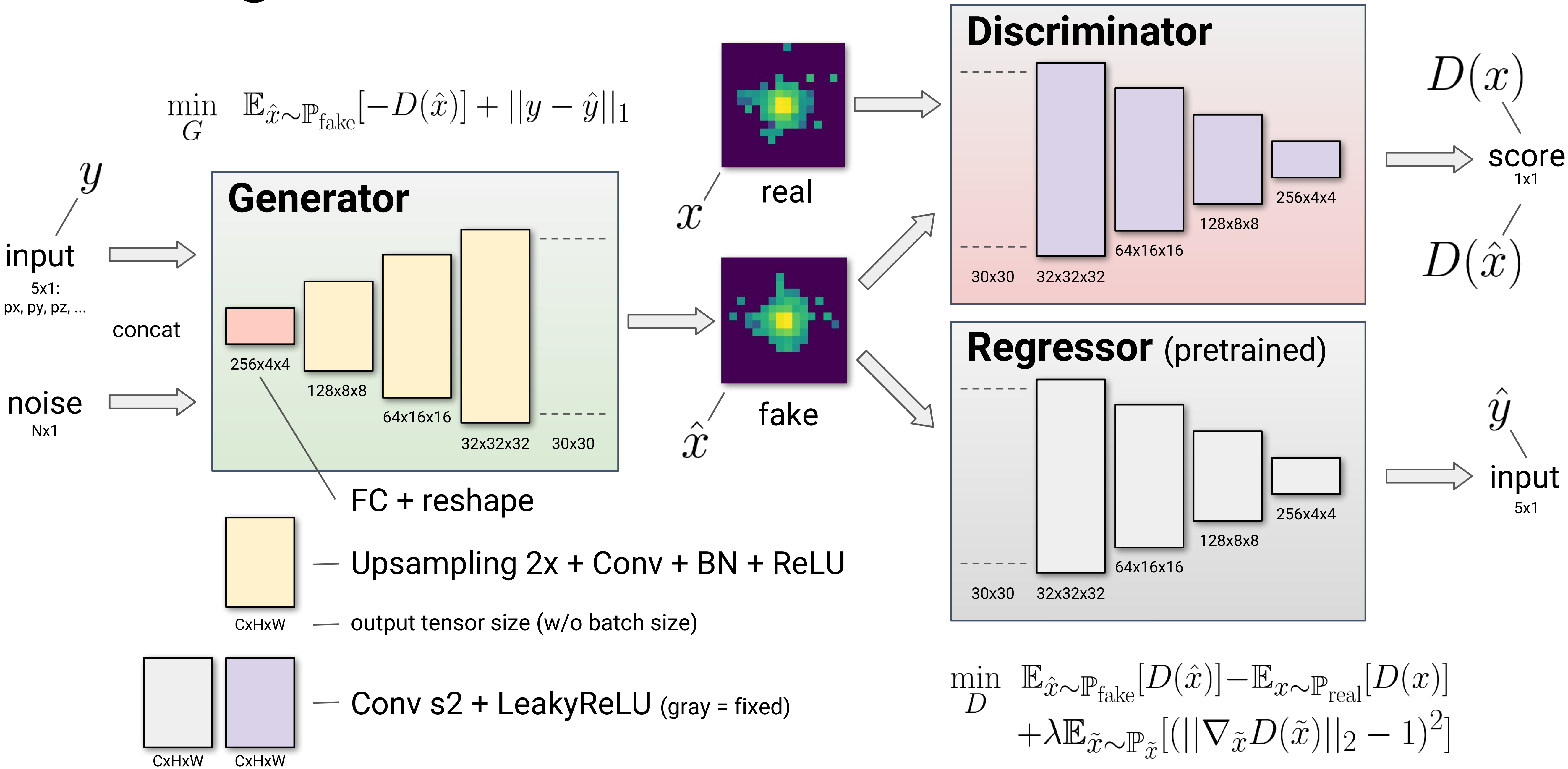
https://medium.freecodecamp.org/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394



- Implicit p(xly), sampling only

LHCb ECAL Fast Simulation: GAN

Training scheme



$$\min_{G} \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\text{fake}}}[-D(\hat{x})] + ||y - \hat{y}||_1$$

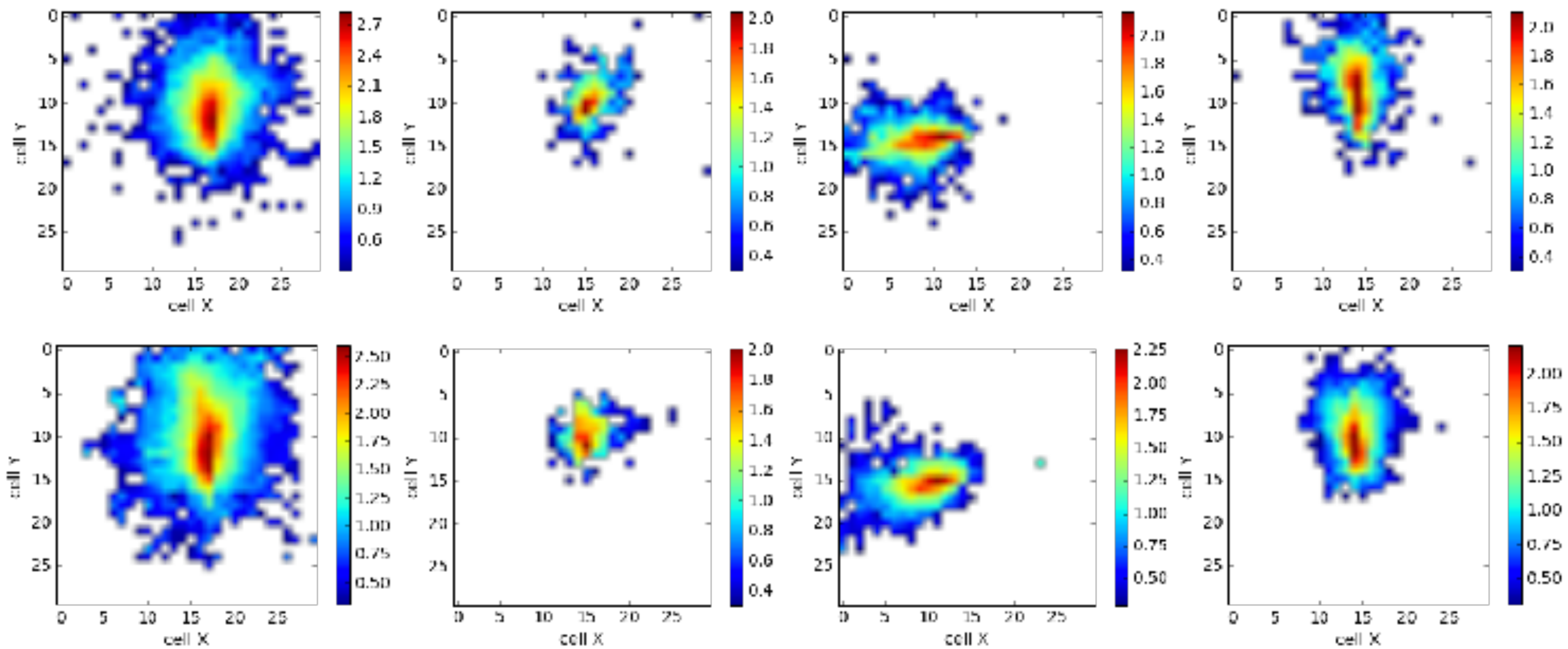
$$\min_{D} \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\text{fake}}}[D(\hat{x})] - \mathbb{E}_{x\sim\mathbb{P}_{\text{real}}}[D(x)] + \lambda\mathbb{E}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}[(||\nabla_{\tilde{x}} D(\tilde{x})||_2 - 1)^2]$$

LHCb ECAL Simulation

GEANT Simulated

log10(cell energy)

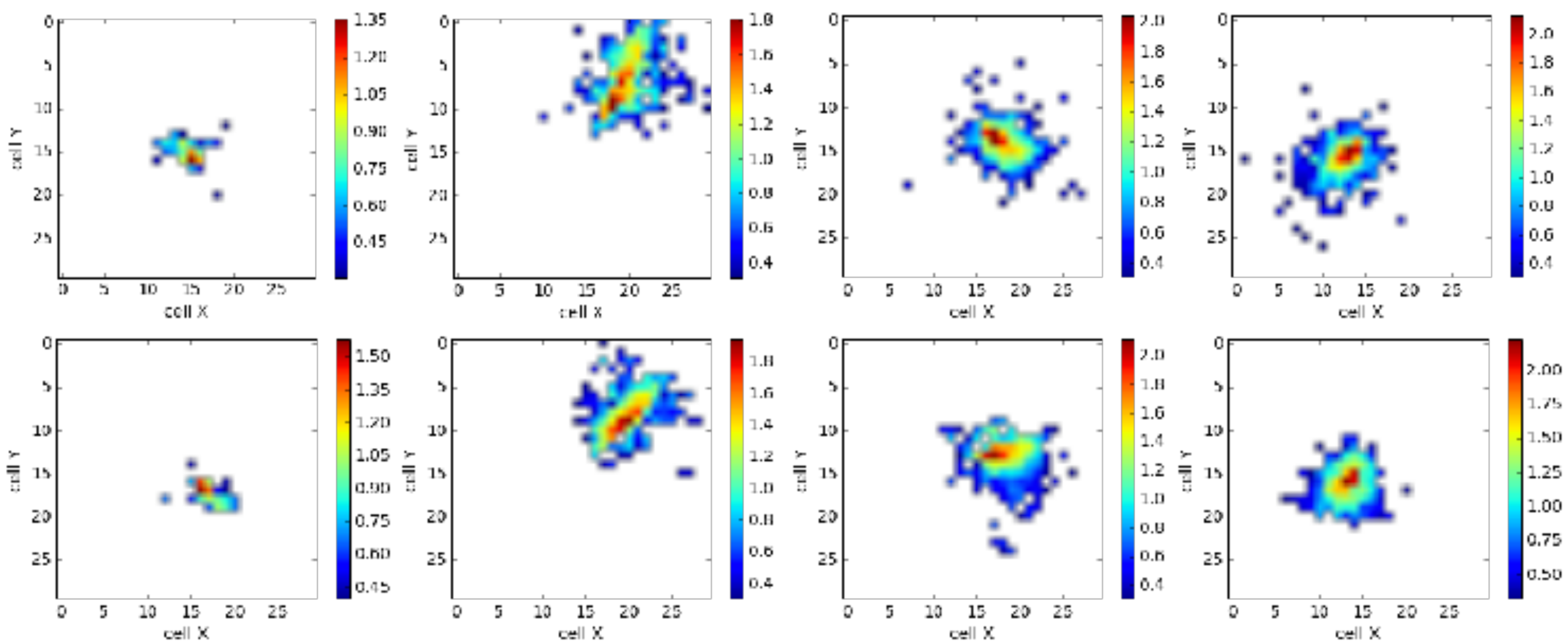
GAN Generated



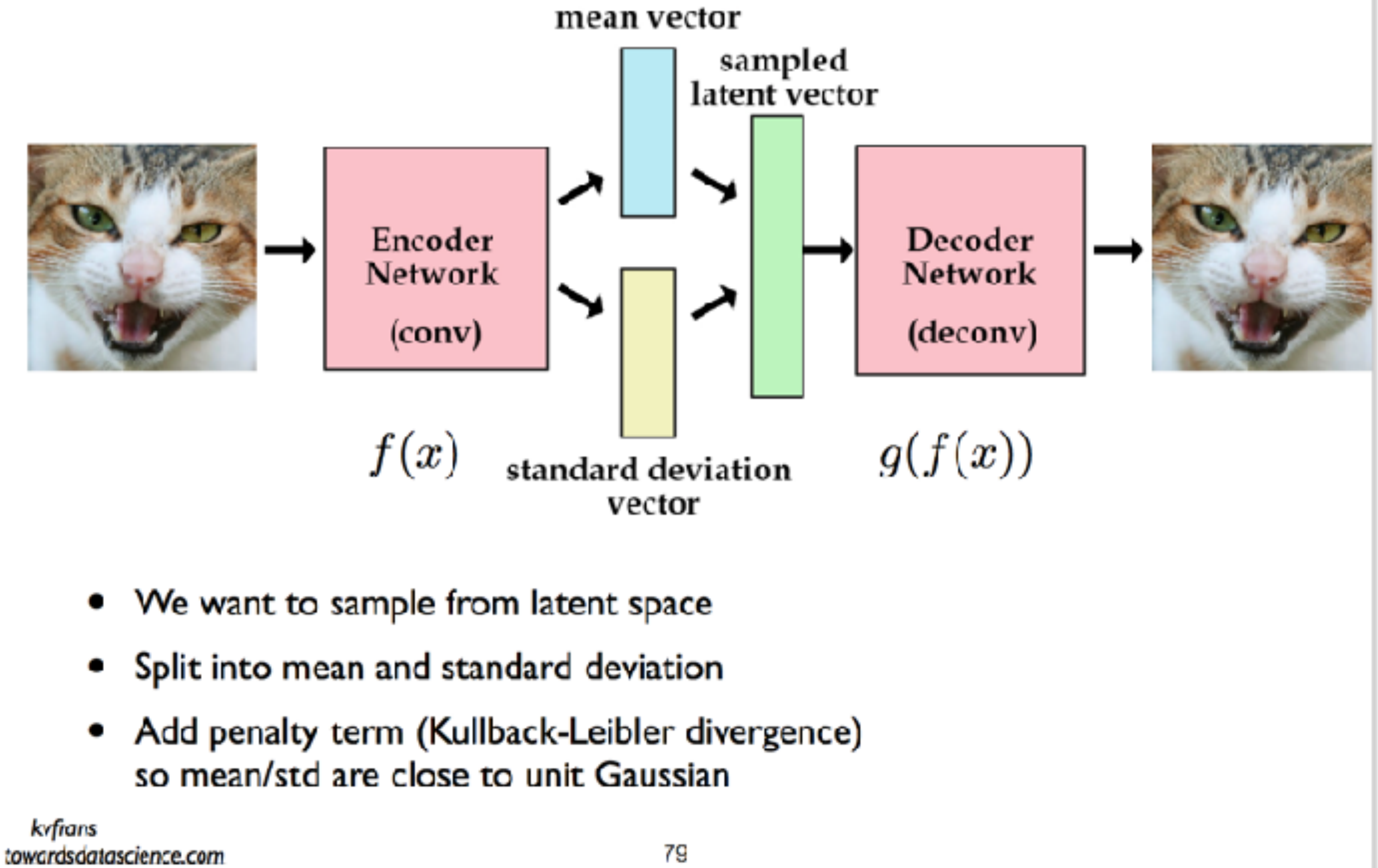
GEANT Simulated

log10(cell energy)

GAN Generated

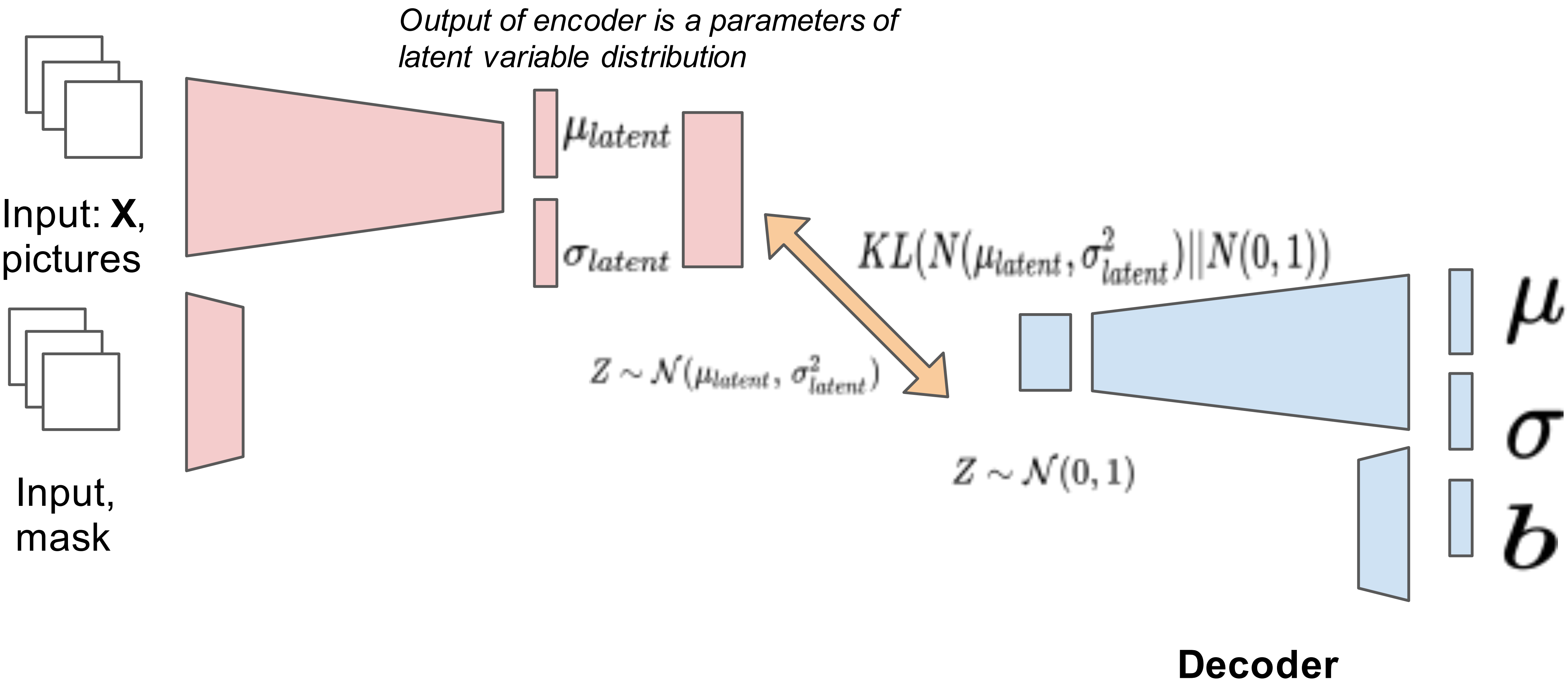


Variational Autoencoder



- VAE allows calculate p(x|y) explicitly
  - NB: GAN only allows sampling from p(x|y)
- ... but smaller size of latent dimensions
  - blurry objects

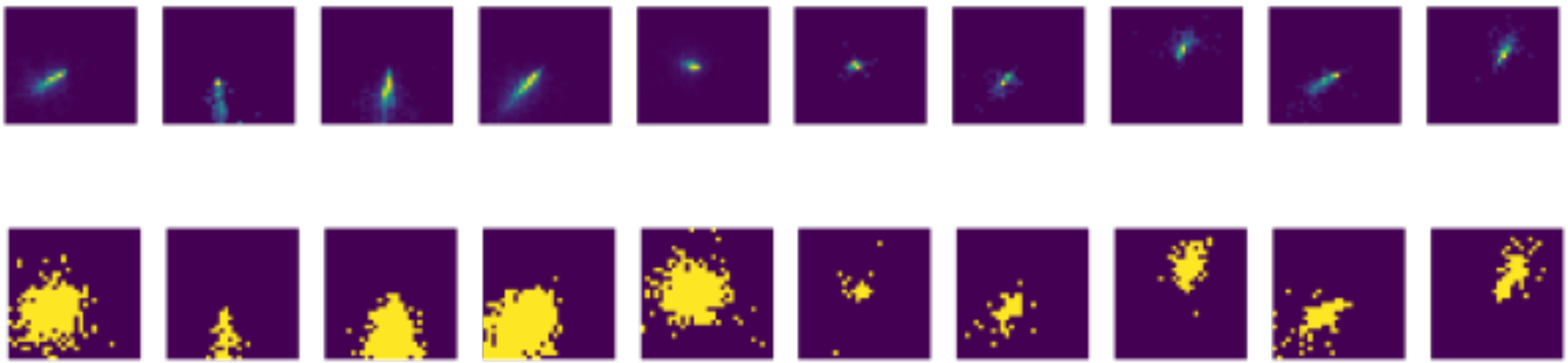
LHCb ECAL Fast Simulation: VAE



$$Loss = KL(N(\mu_{latent}, \sigma^2_{latent})||N(0,1)) + Logprob(X, (\mu, \sigma)) + Logprob(mask, b)$$

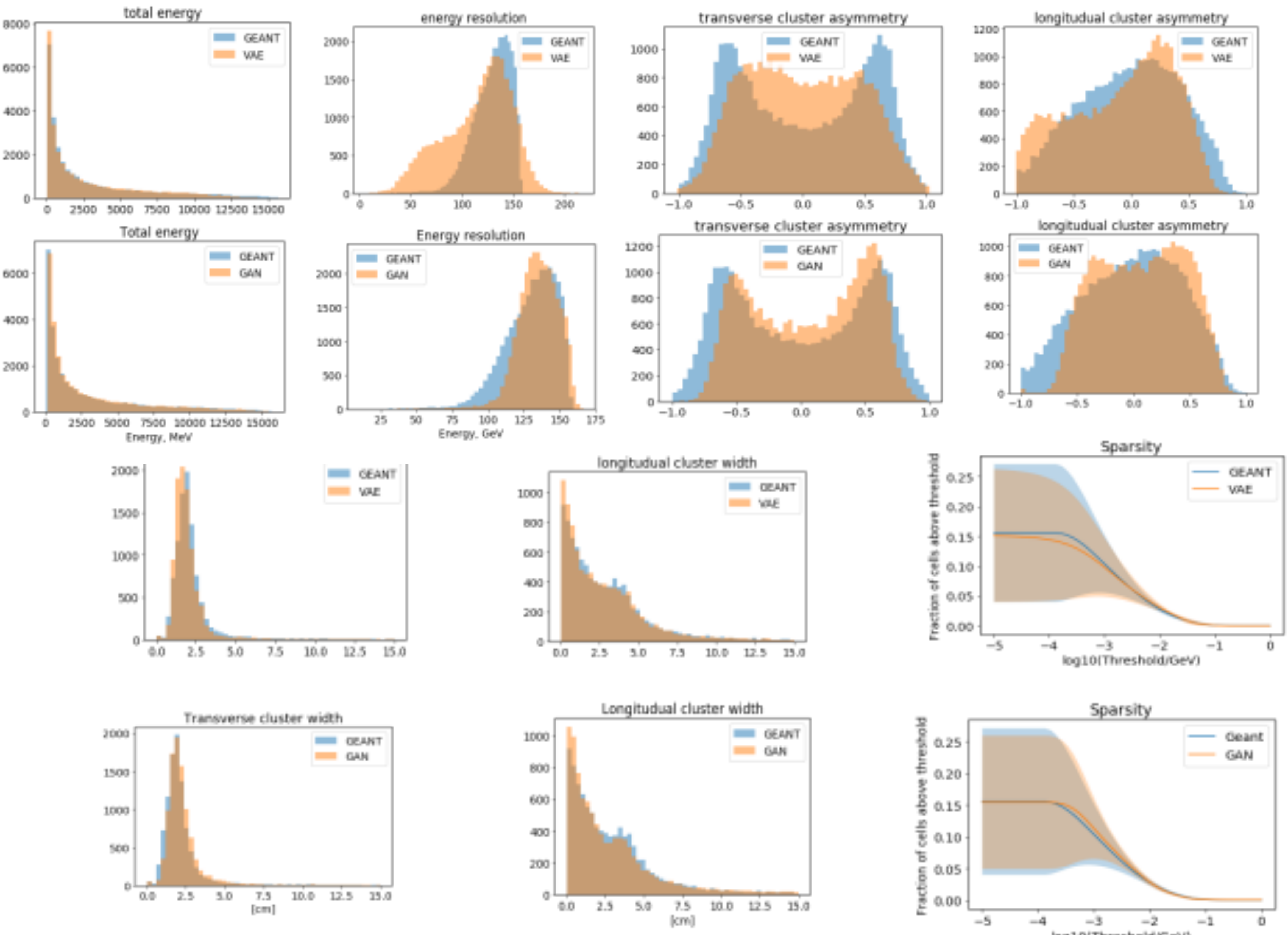
VAE in 5D

GEANT Simulated



VAE Simulated

ECAL Single Cluster Properties



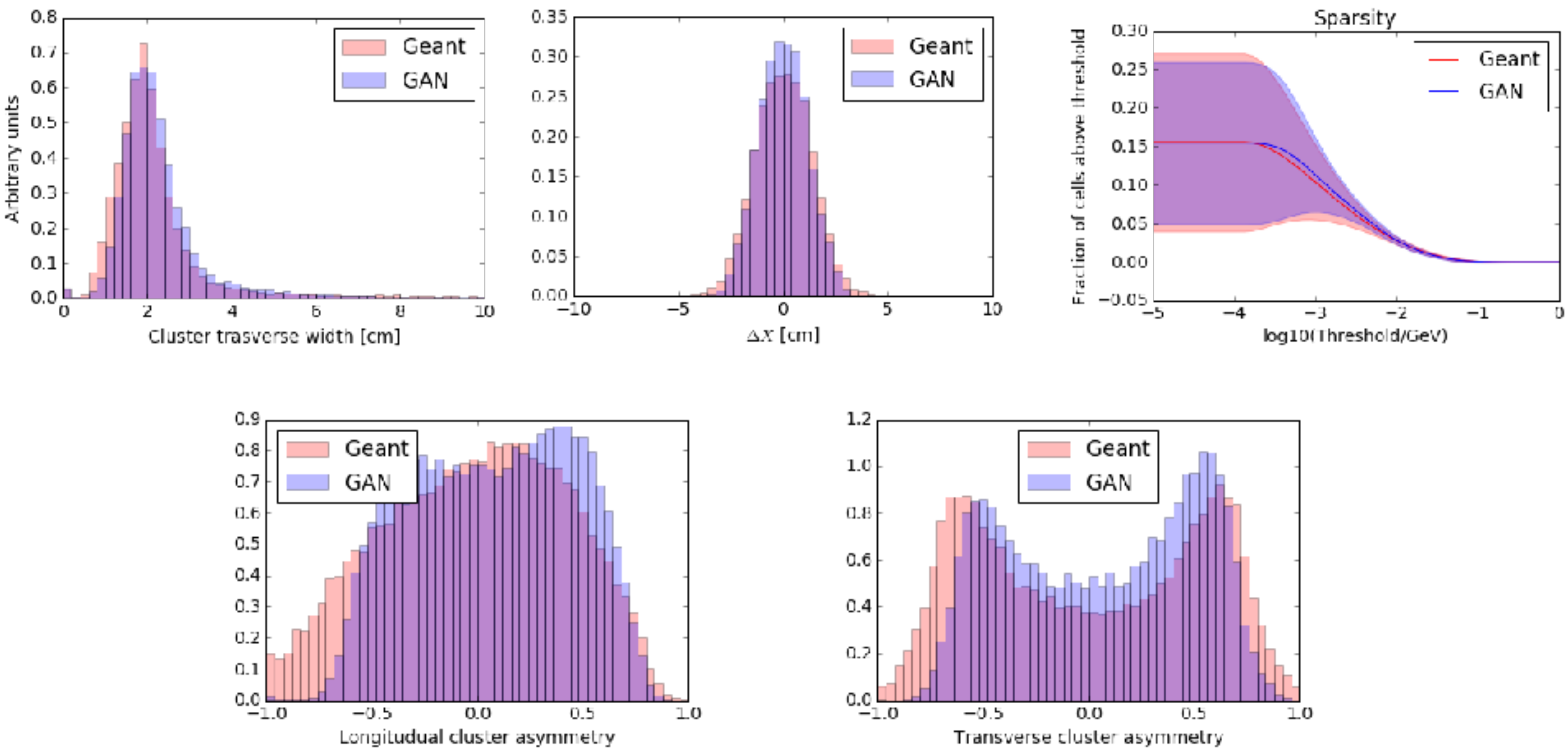
VAE

GAN

VAE

GAN

Primary and Marginal Distributions



- ◇ Is hard to fit marginal distributions
  - ◇ unless the model is aware that those are important for us

Natural Requirements

- For image generation we are usually happy if the result looks like it is desired
- In science we need the result to reasonably well match the given set of requirements. This target set is driven by scientific considerations to reach the ultimate scientific goal
  - e.g. we could want $E^2-p^2=m^2$ for generated particles
- Explicit control to satisfy requirements is preferable
  - e.g. exclude E from generated features, set it explicitly from generated p

Enforcing Important Statistics

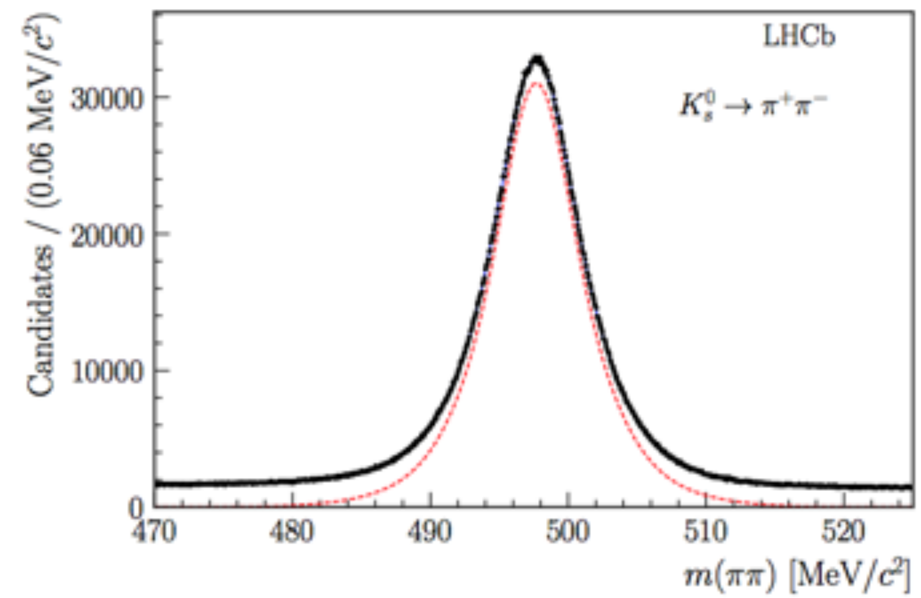
- No generative model is ideal
  - some deviations from the original distribution remain
- Model tends to learn primary statistics of generated objects
- In physics applications we mostly need our model to learn some particular statistics which may be marginal to the generated object
  - e.g. cluster shape fluctuations for fast calorimeter simulation
- Can enforce these statistics by explicit adding them to the los
  - can’t we?

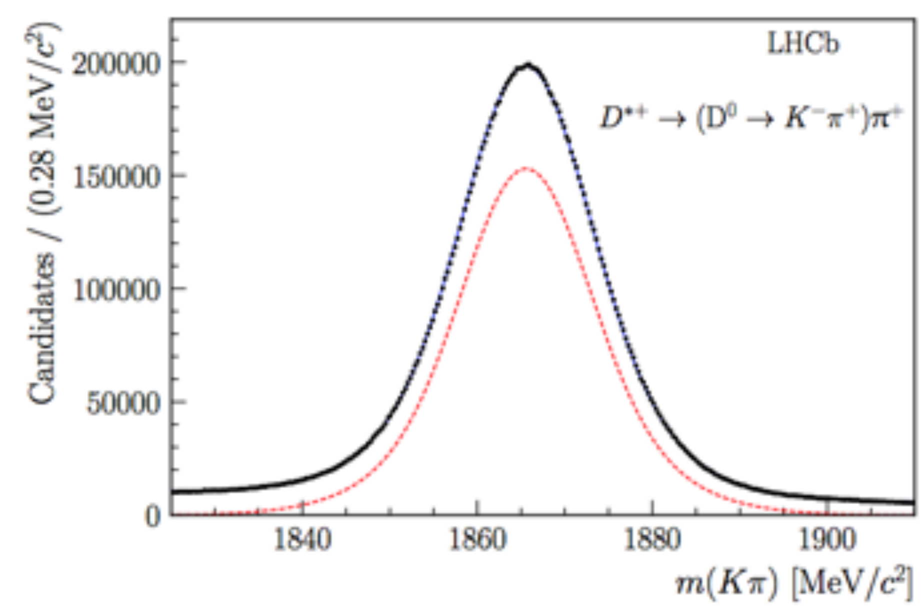
Enforcing Important statistics

- Can enforce statistics by explicit adding them to the los
  - can't we?
- By adding statistics into the loss we do enforce match for these statistics
  - most likely by the price of overtraining these particular statistics
    - … and we lose handle to validate quality of generator on this statistics
- Still can remove those statistics from loss, and see how far they would deviate
  - figure of merit for generating this statistics

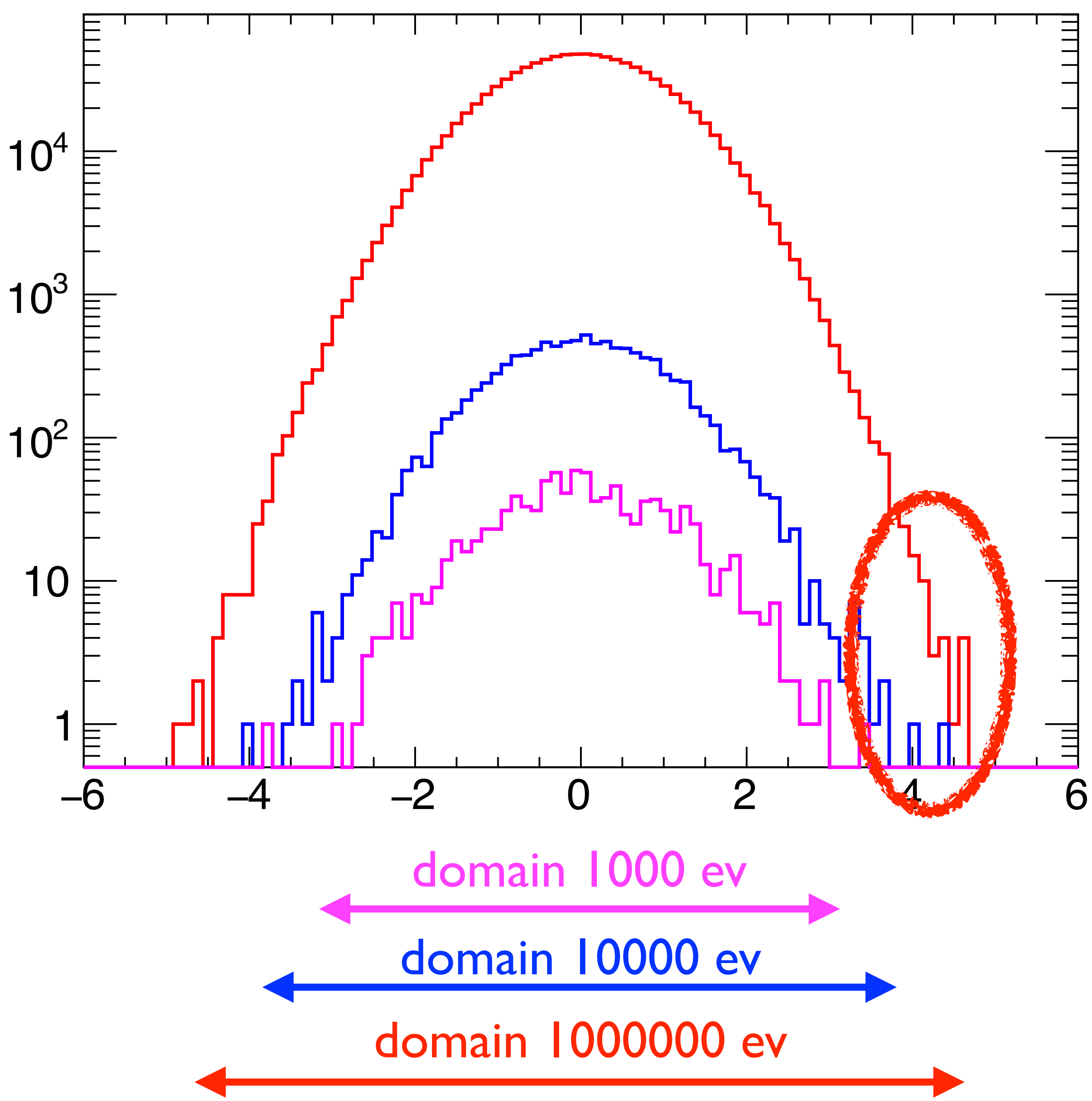
Generative Models Trained on Real Data

- Real data samples, even calibration, are never 100% clean
  - contamination from events with different labels/conditions
- Can not determine label of particular object uniquely
  - however can statistically determine fractions of different labels
- Can use weighted samples to train WGAN and CramerGAN





Completeness



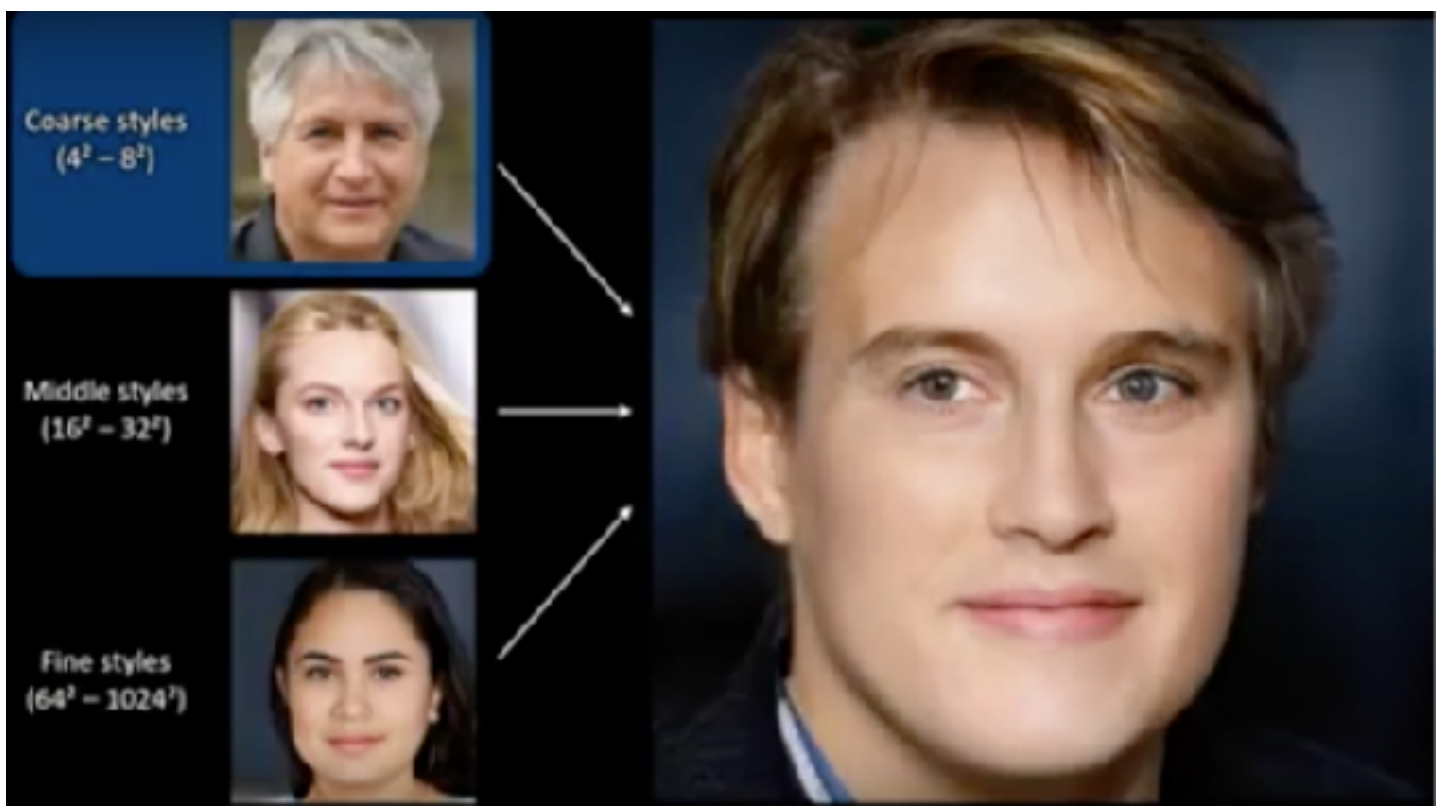
- Domain for the generative model is driven by the training sample
  - model can not extend beyond the train domain even if produces high statistics
    - until explicitly set to behave beyond train domain

Decomposition

- Quality of the generative models is limited by the size of the train data sample
  - generative models may not give profit for producing statistically correct big data sets
    - no information beyond the train sample is available

Decomposition

- Quality of the generative models is limited by the size of the train data sample
  - generative models may not give profit for producing statistically correct big data sets
    - no information beyond the train sample is available



- Not quite if we can decompose generative model into separate components
  - random combinations of different components may drastically increase variativity

Decomposition

- Quality of the generative models is limited by the size of the train data sample
  - generative models may not give profit for producing statistically correct big data sets
    - no information beyond the train sample is available
- Not quite if we can decompose generative model into separate components
  - random combinations of different components may drastically increase variativity
- E.g. fast simulation of the calorimeter response
  - generator is trained on 10^6 incident particles
  - $\sim$50 particles in the calorimeter per event
  - total variativity $\sim(10^6)^{50} = 10^{300}$!

Quality Metric

- No generative model is ideal
  - some deviations from the original distribution remain
- Minor deviations are not that important e.g. for image generation
- Minor deviations may be a big deal for physics generative models
  - e.g. we could want $E^2 - p^2 = m^2$ for generated particles to be precise
- Ultimate generative model quality metric is comparing final physics result obtained using generative model, and the one obtained using train data
  - accuracy is limited by the size of the train data

Conclusions

- Surrogate generative models demonstrate extraordinary progress in current years
- Fast simulation for LHC detectors in Run 3 is a natural target
  - fast simulation of calorimeters is a primary target
- Generative models need attention ensure scientifically solid results
  - completeness of generated sample
  - satisfying boundary conditions, control of scientifically important but marginal statistics
  - evaluating quality of the model, propagate model imperfections to systematic uncertainties of the final scientific result
- We developing different approaches for fast generation of calorimeters in LHCb
  - results look promising, but not production quality yet