Machine learning based simulation in reduced dimensions

Chitrakshee Yede

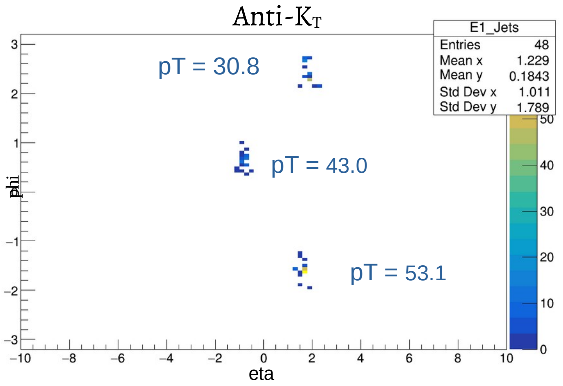
Mentors: Sourabh Dube, Arnab Laha (IISER, Pune)

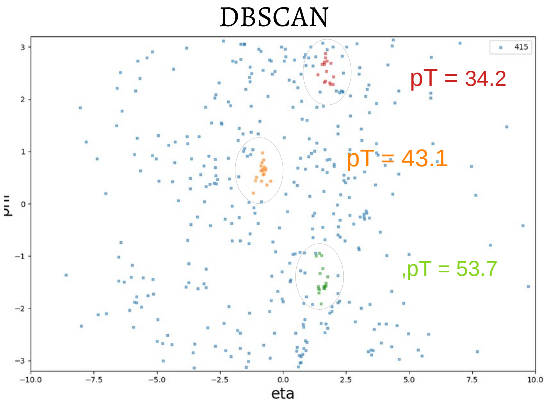
HSF- India trainee program

15th September, 2023

About me…

Bachelor of Science , Physics, 2021
Fergusson College (Autonomous),
Savitribai Phule Pune University, India
Project: The study of pileup and jet clustering algorithms at the LHC.

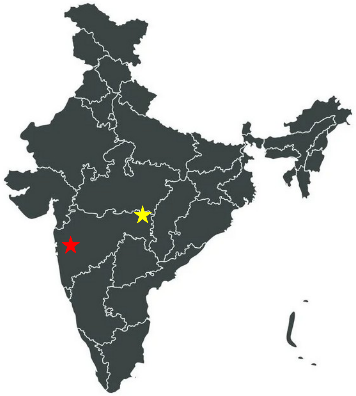


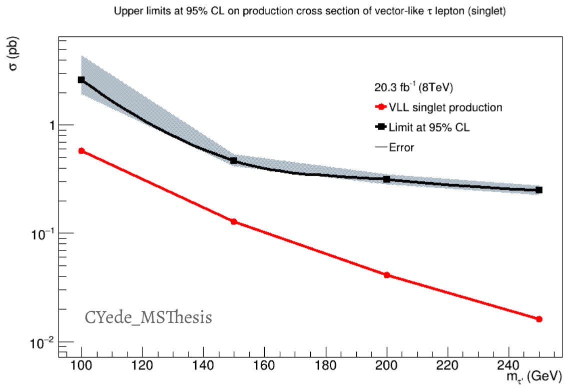


Jet Clustering Algorithms

Summer'21 Project: Examining 2D clustering of jets.

Master of Science , Physics, on-going
Savitribai Phule Pune University, India
Project: Constraining the vector-like tau model at $\sqrt{s}$ = 8 TeV.
(arXiv:1411.2921)





Simulating events for ML

Simulation is crucial in HEP!

- Particle Interaction Modeling
- Detector performance and design
- Event Generation
- Background Estimation
- Signal and background discrimination

PHYSICAL REVIEW D **105,** 112007 (2022)

Inclusive nonresonant multilepton probes of new phenomena at $\sqrt{s} = 13$ TeV

A. Tumasyan *et al.**
(CMS Collaboration)

Used BDTs to discriminate between the signal and the background

BDTs are trained using large simulation samples

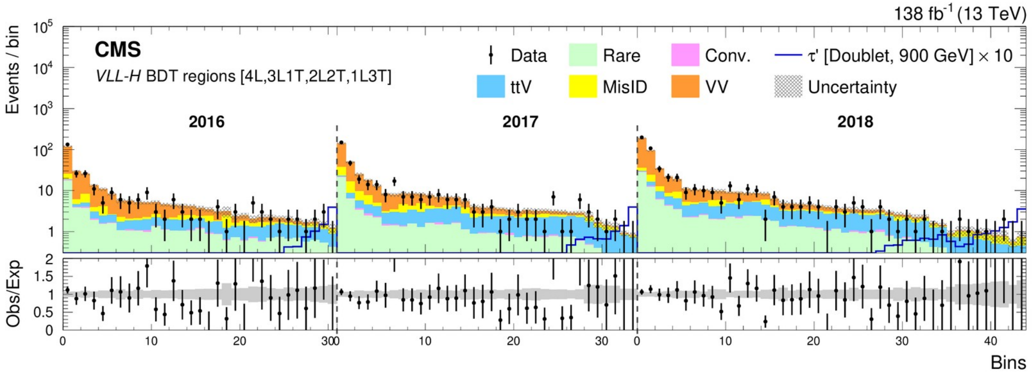
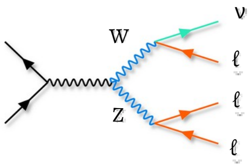


Fig.19, 10.1103/PhysRevD.105.112007

The Problem



Lets look at some numbers !!

	$N_{generated}$	$N_{training}$
2016	11.9M	124k
2017	10.8M	132k
2018	22M	233k

$N_{generated}$ - total number of WZ events generated, Madgraph + Pythia + Geant4 (CMSSW)
$N_{training}$ – number of WZ events used to train BDTs

Numbers from analysis described in paper, 10.1103/PhysRevD.105.112007

~ 1% of the generated events used for training

Event selection optimizes signal-to-background ratio, excluding a significant portion of background events, particularly when training a signal vs. background classifier.

Higher statistics leads to better training performance!

How can we achieve that?

Can generate more events! (Time-Consuming) ☹

or

Can generate partial events (only the required variables) !? ☺

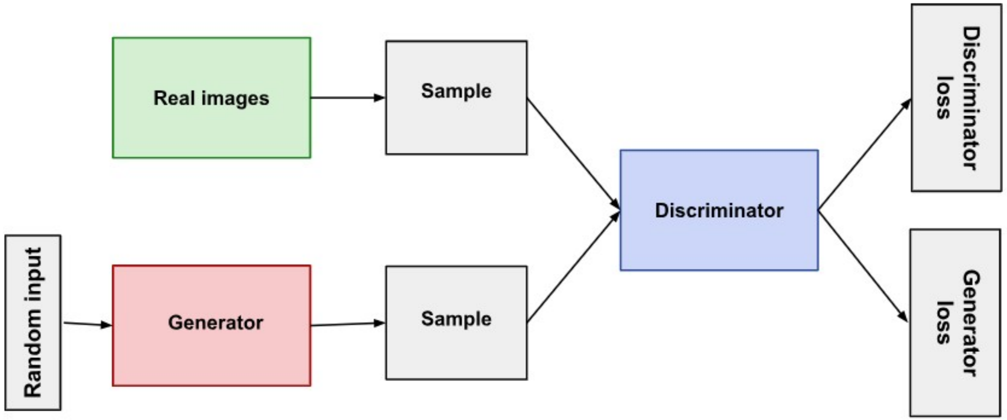
Generative Model

Generative Adversarial Networks (GANs)

- Two sub-models – Generator and Discriminator that competes with each other!
- Generator : Generates pseudo data
- Discriminator : Distinguishes between real and pseudo data
- This iteration continues until generator succeeds to fool the discriminator

Generates a distribution by sampling from latent space and learning the correlation of features space

LHC analysis-specific datasets with Generative Adversarial Networks (arXiv:1901.05282)
Generative models for fast simulation (J. Phys.: Conf. Ser. 1085 022005)
Particle Generative Adversarial Networks for full-event simulation at the LHC and their application to pileup description (arXiv:1912.02748)

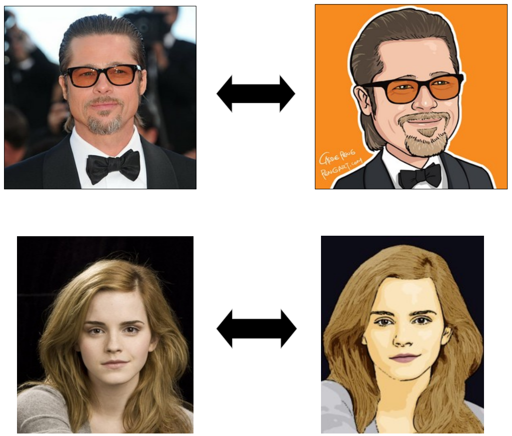


Complication: Slower and limited performance as number of features increases

Dimensionality Reduction + GANs

Suppose we obtain a lower dimension representation of the data

Generating at lower dimension is easier!



Goals:

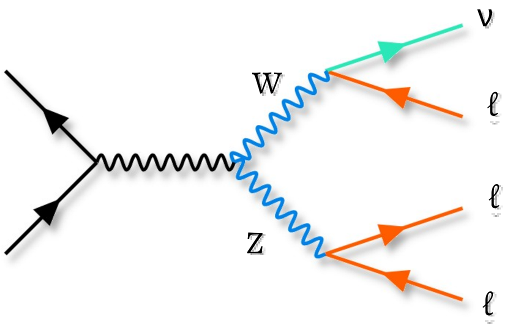
- Should be faster
- To design a user-friendly pipeline
- Should be operable on personal workstations

Specific event properties (True properties) → Low-dim representation → Use GAN to generate low-dim representation → Invert to get true properties (reconstructed properties)

Step 1 (Specific event properties → Low-dim representation)

Step 2 (Use GAN to generate low-dim representation)

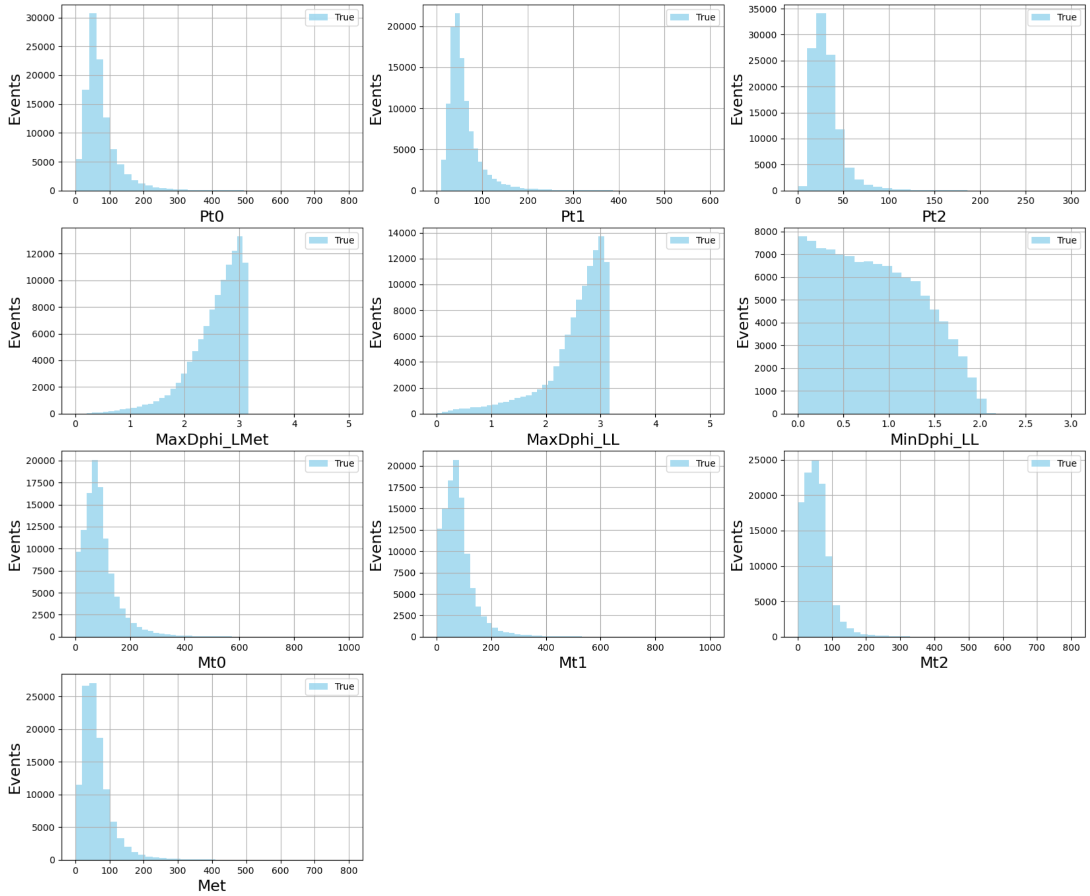
WZ Sample



WZ process is irreducible background for multilepton searches

Particle momenta, Missing pT, Transverse Mass, Angular information

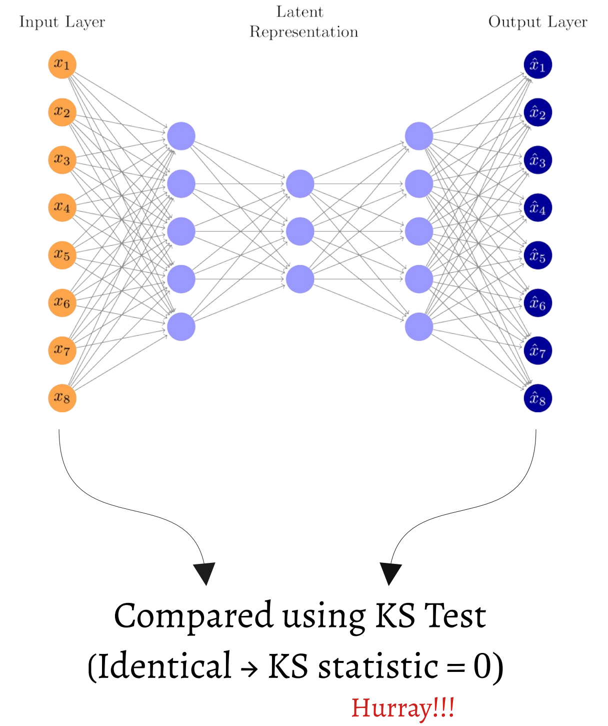
True variables



Autoencoders

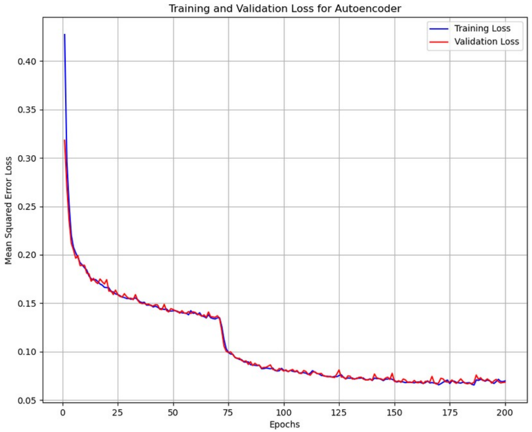
Step 1 : Implement a method that is reversible to obtain a lower dimensional representation of the true information.

- We are exploring autoencoders. Its a neural network architectures that aim to encode and then reconstruct input.
- WZ Sample – Training = 450k events
  Testing = 110k events
- We explored various neural network architectures and hyperparameter configurations.
- We utilize Kolmogorov-Smirnov (KS) test to assess the match between true and reconstructed data distributions. We report the mean KS statistic for the variables considered as a metric.

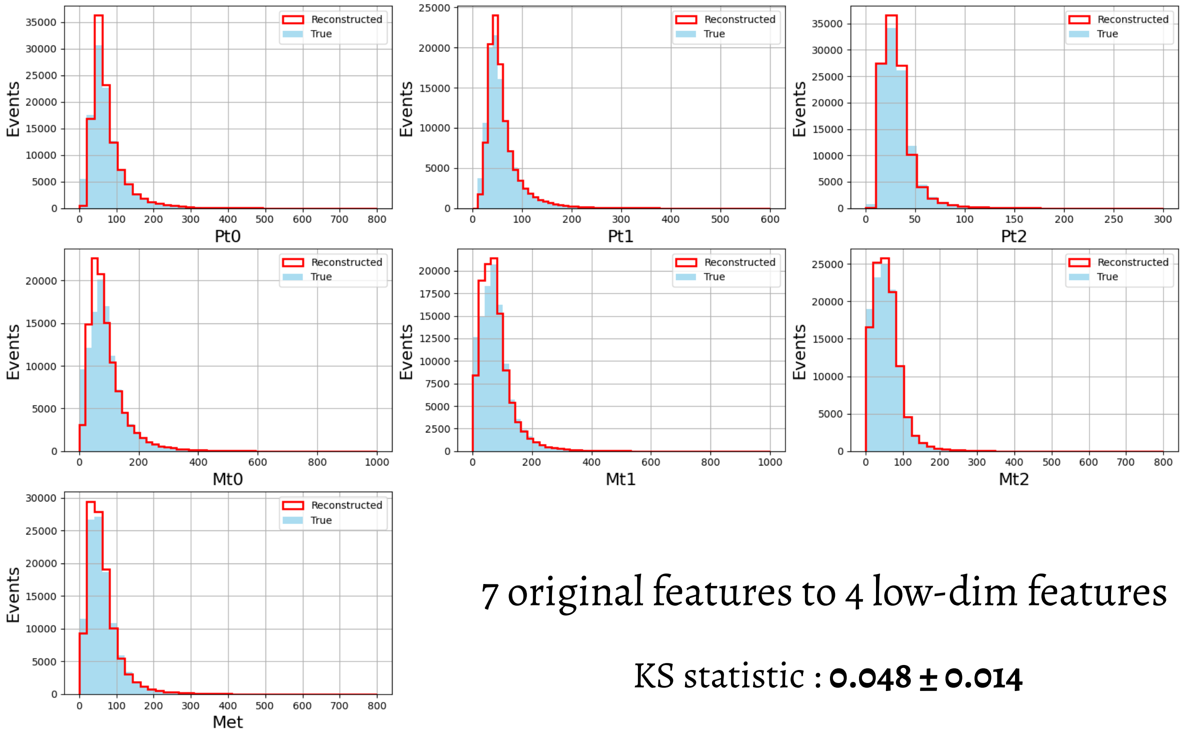


Results(1)

- Testing is done for 110k events
- NN Architecture :
  256/64/8→ 4→ 8/64/256
- Trainable parameters: 38,107
- Loss function : MSE
- Epochs = 200



True vs. Reconstructed



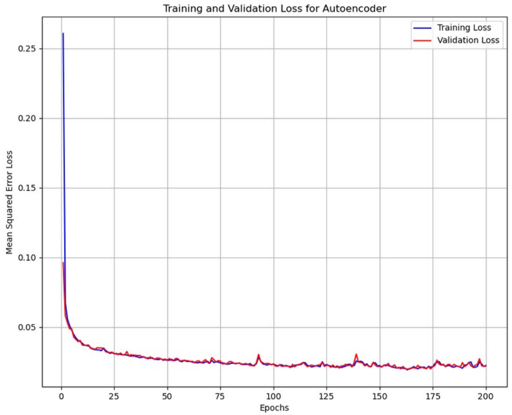
7 original features to 4 low-dim features

KS statistic : **0.048 ± 0.014**

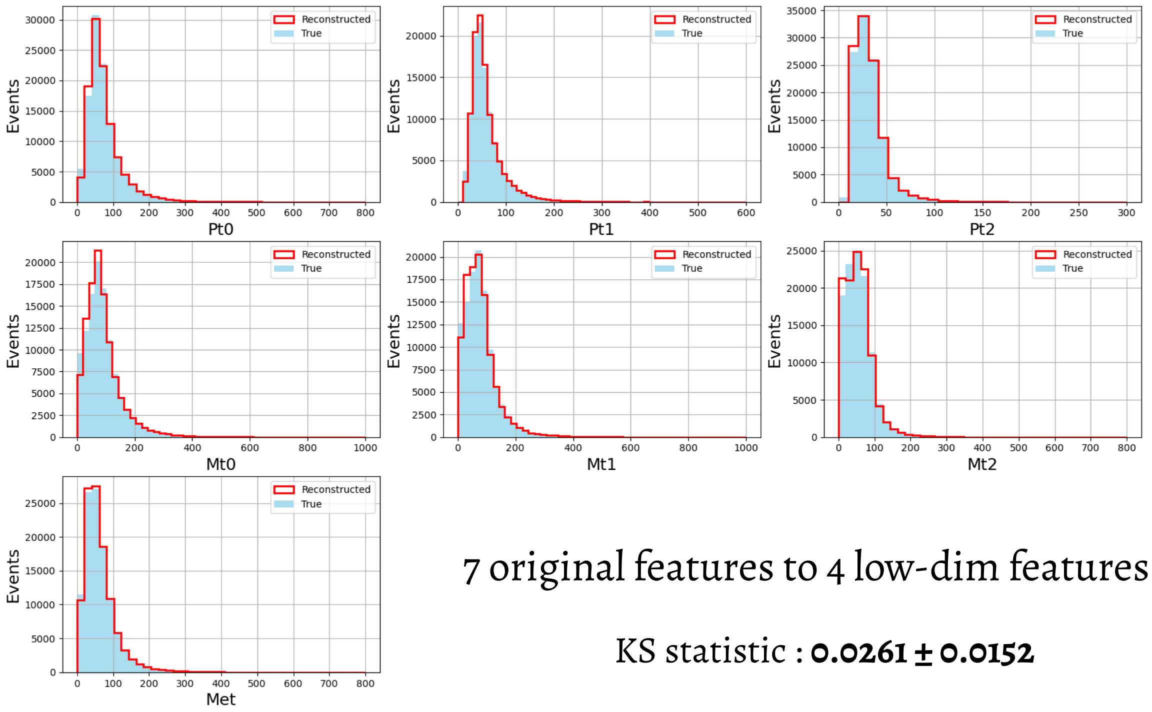
Results(2)

- Testing is done for 110k events
- NN Architecture :

  512/128/64/8→ 4→ 8/64/128/512
- Trainable parameters: 157,147
- Loss function : MSE
- Epochs = 200



True vs. Reconstructed



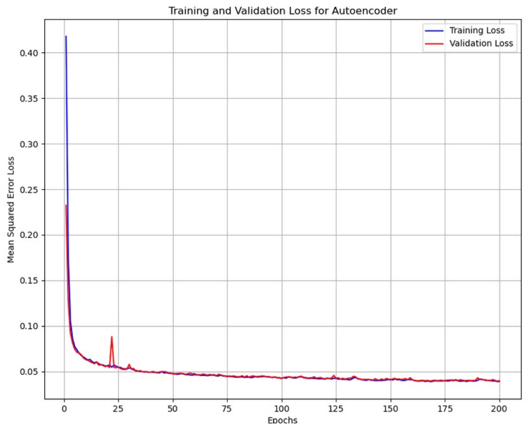
7 original features to 4 low-dim features

KS statistic : **0.0261 ± 0.0152**

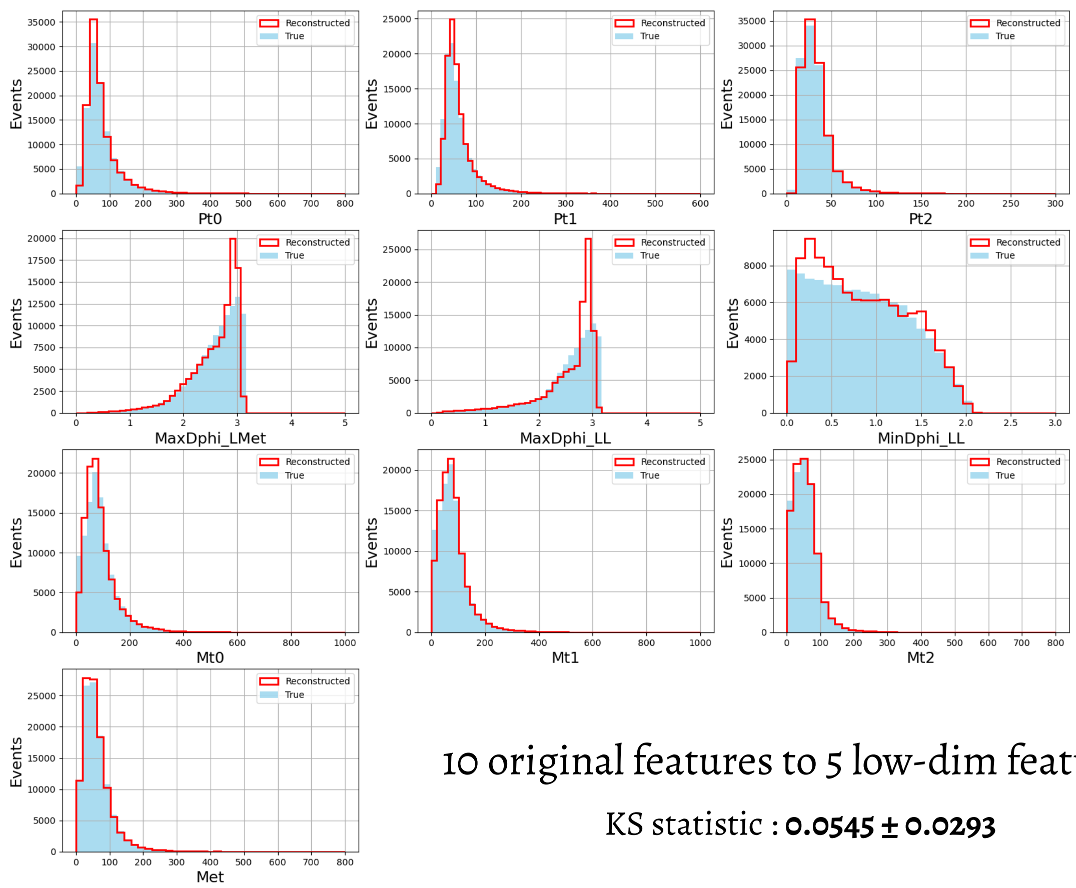
Results(3)

- Testing is done for 110k events
- NN Architecture :

  512/128/64/8→ 5→ 8/64/128/512
- Trainable parameters: 160,239
- Loss function : MSE
- Epochs = 200



True vs. Reconstructed



10 original features to 5 low-dim features

KS statistic : **0.0545 ± 0.0293**

Next Steps

- STEP 1 : Freeze process of making lower dimension representations.

Autoencoder	Trainable parameter	KS_statistic
7 to 2	157,113	0.074 ± 0.021
7 to 4	38,107	0.048 ± 0.014
7 to 4	157,147	0.026 ± 0.015
10 to 5	6,191	0.075 ± 0.042
10 to 5	160,239	0.054 ± 0.029

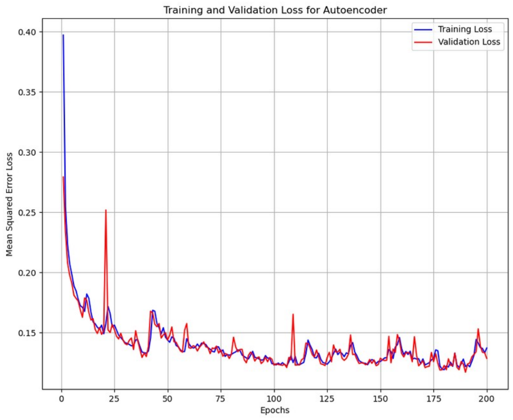
- STEP 2 : To construct a GAN to generate at lower dimension (we may explore variational autoencoders (VAEs)).
- STEP 3 : Establish viability of output through detailed comparison between true and generated data and correlation studies to check if generated data carries sufficient physics information

Backup

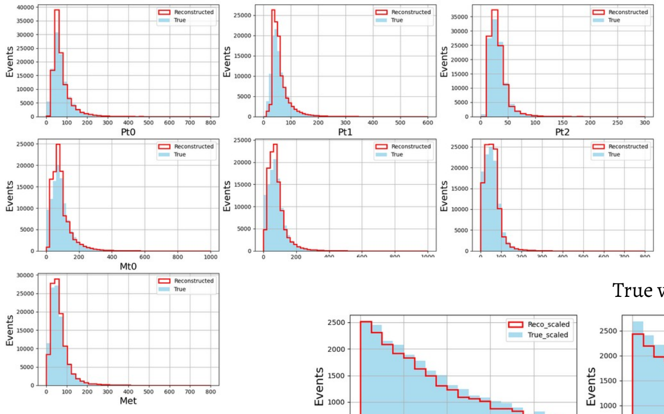
Results(4)

- Testing is done for 110k events
- NN Architecture :

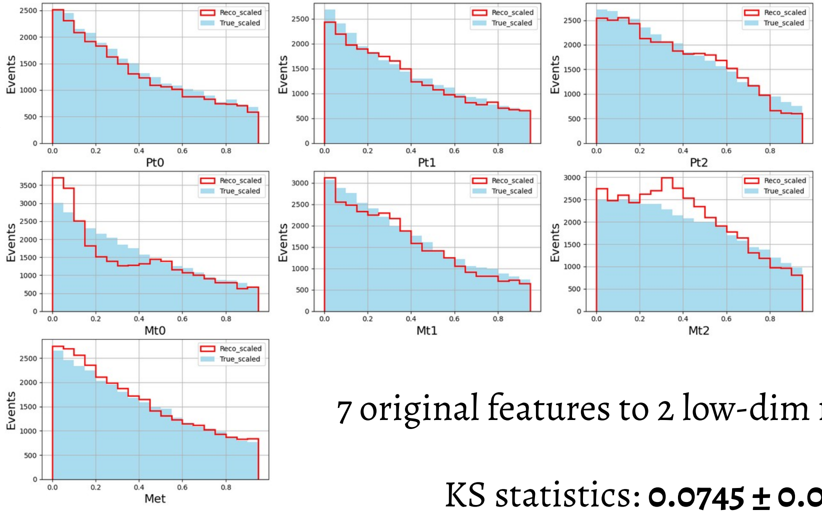
  512/128/64/8→ 2→ 8/64/128/512
- Trainable parameters: 157,113
- Loss function : MSE
- Epochs = 200



True vs. Reconstructed



True vs. Reconstructed (scaled)



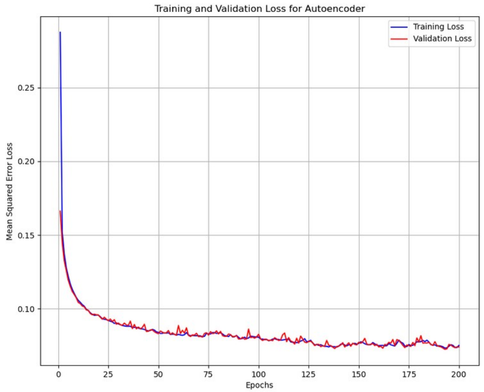
7 original features to 2 low-dim features

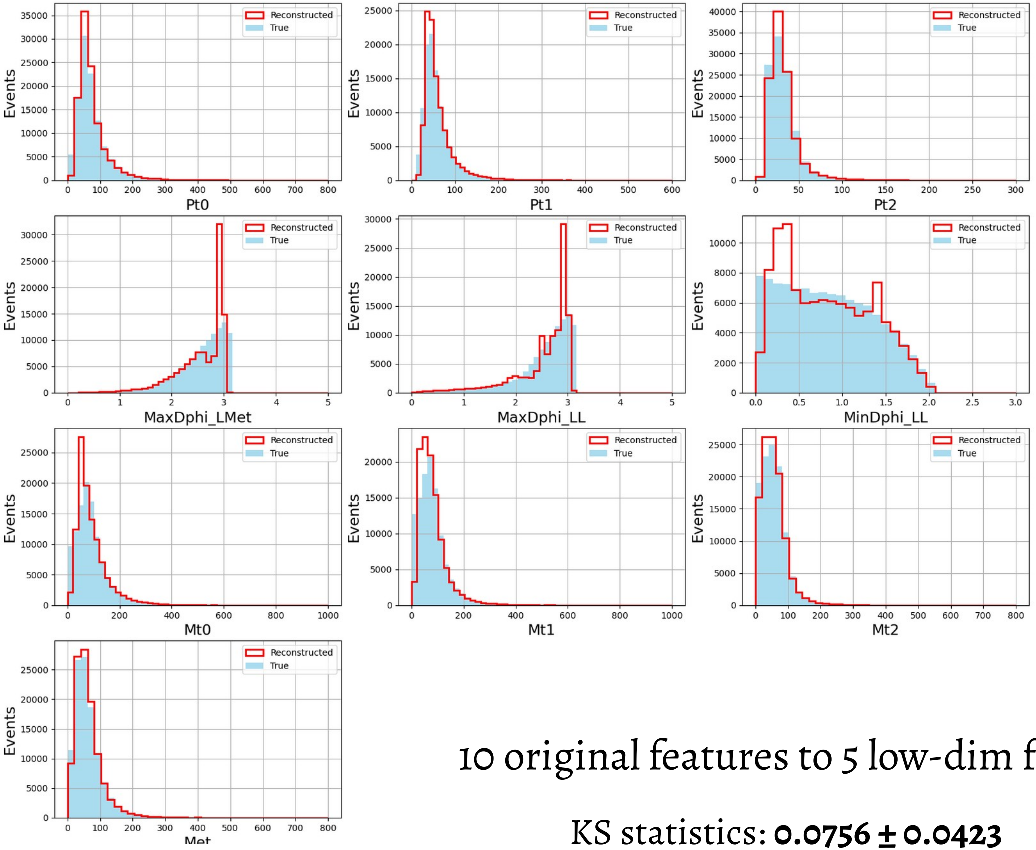
KS statistics: **0.0745 ± 0.0216**

Results(5)

- Testing is done for 110k events
- NN Architecture :

  64/32/8→ 5→ 8/32/64
- Trainable parameters: 6,191
- Loss function : MSE
- Epochs = 200
- Batch_size = 500



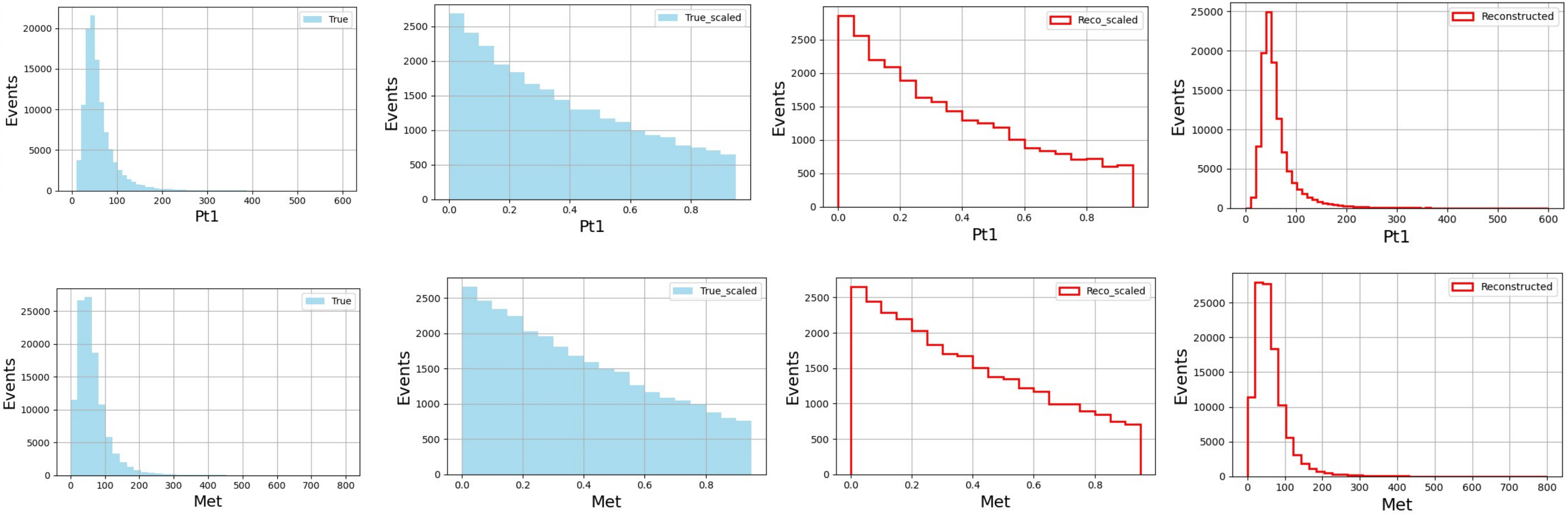


10 original features to 5 low-dim features

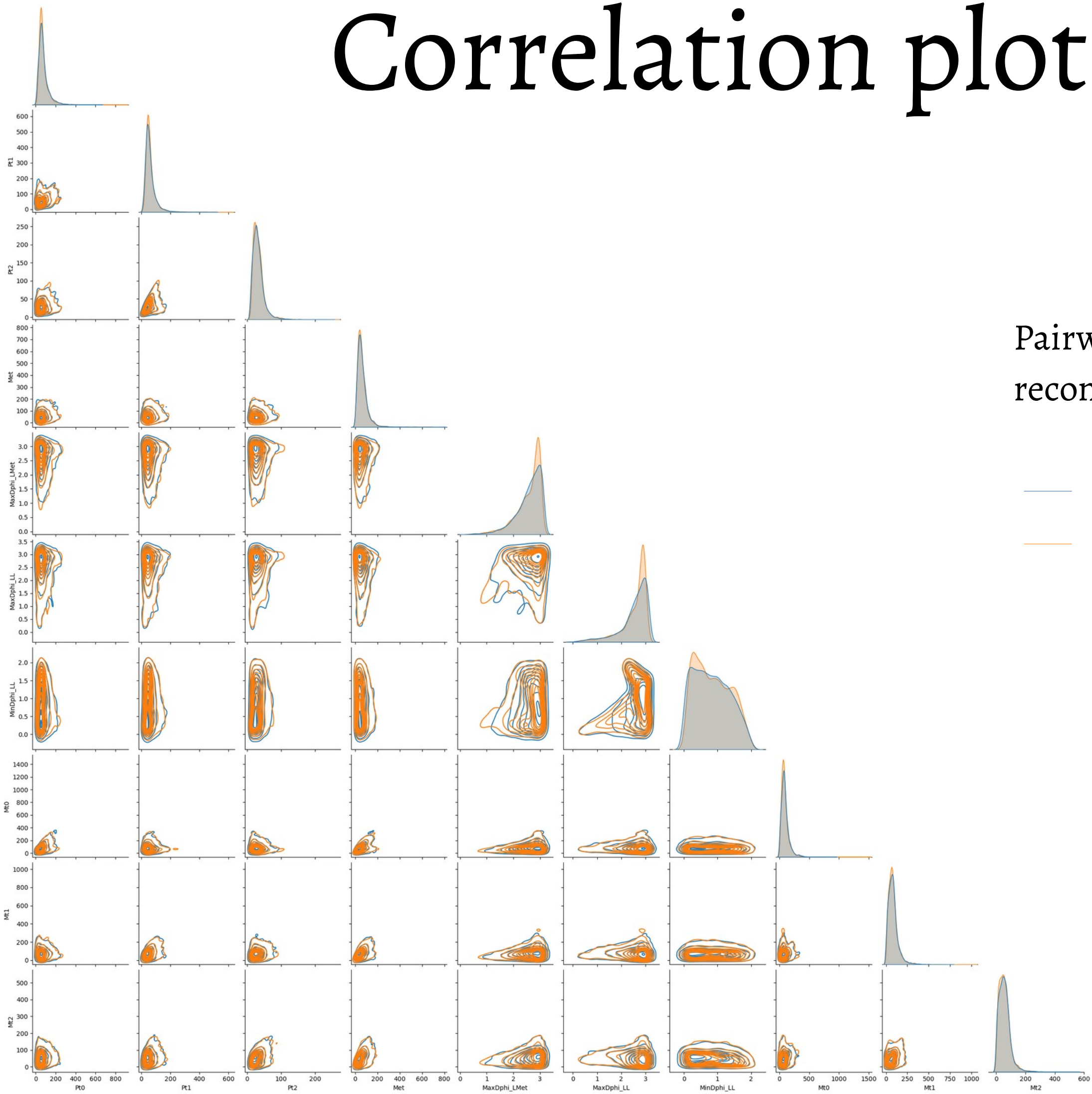
KS statistics: **0.0756 ± 0.0423**

Scaling

Scaler used → StandardScaler



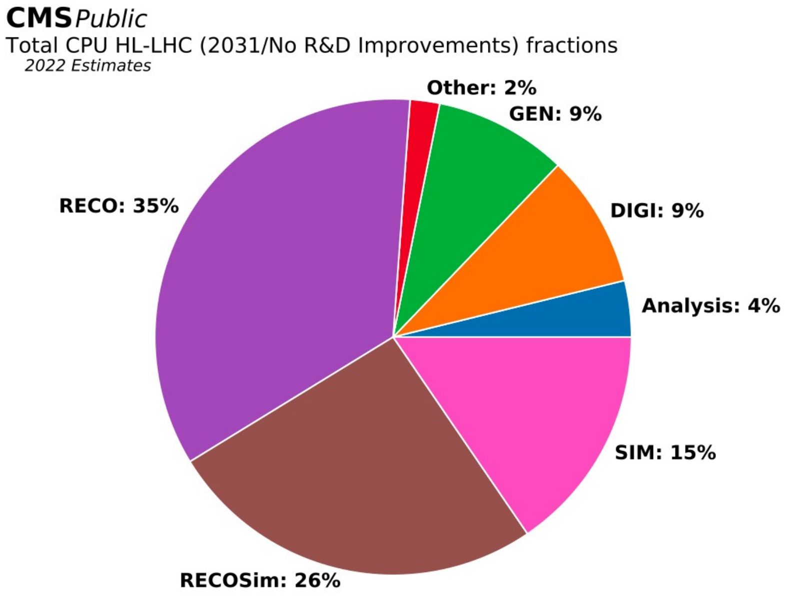
Correlation plot



Pairwise correlation plot of true and reconstructed variables

- Original
- Reconstructed

Usage of CPU



CMS generating more and more simulation is a harder strain on the resources.

Source : https://cds.cern.ch/record/2815292?ln=en