Upcoming

- https://indico.desy.de/indico/event/25341/overview

Work-log

- https://trello.com/b/1hw2qil0/autoencoders-for-atlas
- https://docs.google.com/presentation/d/12_yRCl63H1VElDejEUteBqwXalWs0b9x3YcoKGFkkZA/edit?usp=sharing
- Presentation: https://docs.google.com/presentation/d/12_yRCl63H1VElDejEUteBqwXalWs0b9x3YcoKGFkkZA/edit?usp=sharing

Imp paths

- /eos/atlas/user/d/doglioni/Autoencoders_for_compression
- /afs/cern.ch/work/h/hgupta
- /afs/cern.ch/user/h/hgupta

Links

- https://mattermost.web.cern.ch/it-dep/channels/town-square
- https://cern.service-now.com/service-portal/
- https://hal.inria.fr/hal-02396279/document
- https://arxiv.org/abs/1707.08966
- https://arxiv.org/pdf/2006.04780.pdf

Ideas

- Check numerical deep compression works for loss and normalization
- Loss
    - L2
    - Relative error as loss
        - Might have to avoid divergence due to division
    - L1 + MSE

TODO

- Formulate custom norm parameters for PhenoML data
- Discuss how to add certain categories of particles to the test data

Wrapping up

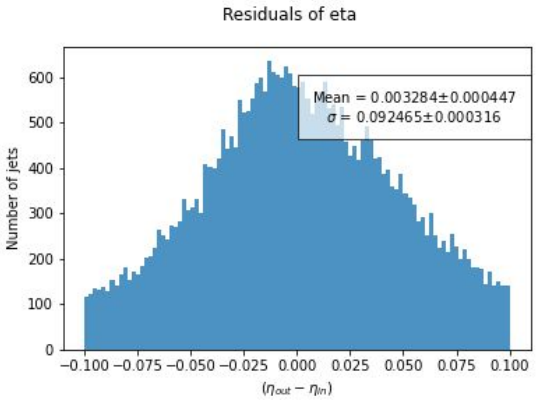
- Added models, training logs and testing codes to the github repo
- Merged with the master branch

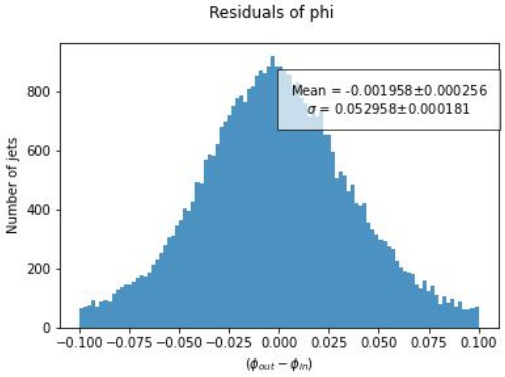
- Updated documentation for the repo
    - Merged again with the master

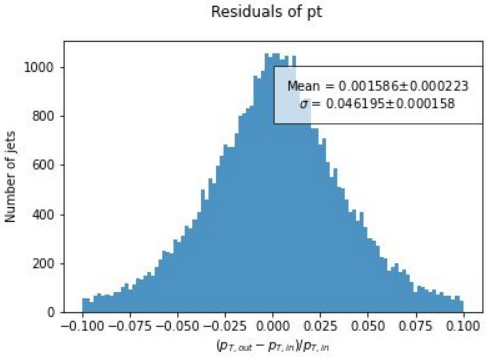
- Added rest of the plots to the drive
- Adding processed datasets from my laptop used for testing - phenoML processes csv files for different particles and created darkmachines chan3 csv file
    - Link: https://drive.google.com/drive/folders/12nX9uXFGyqJvH0eLRigIPUMGtM0l_6rv?usp=sharing
- List of commits:
    - https://docs.google.com/spreadsheets/d/1DFU6QOT6MXZKnYkXndzQlzxPZSPtLf-GB4Qk9BgQuT0/edit?usp=sharing
- Added the remaining plots in the slides
- TODO:
    - Report writing

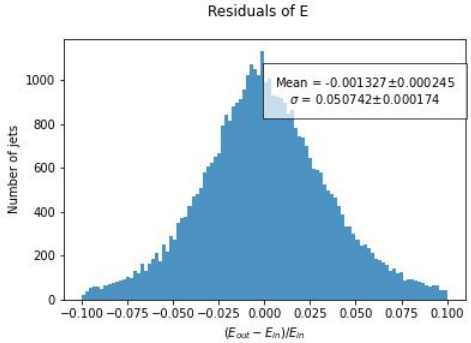
Aug 19

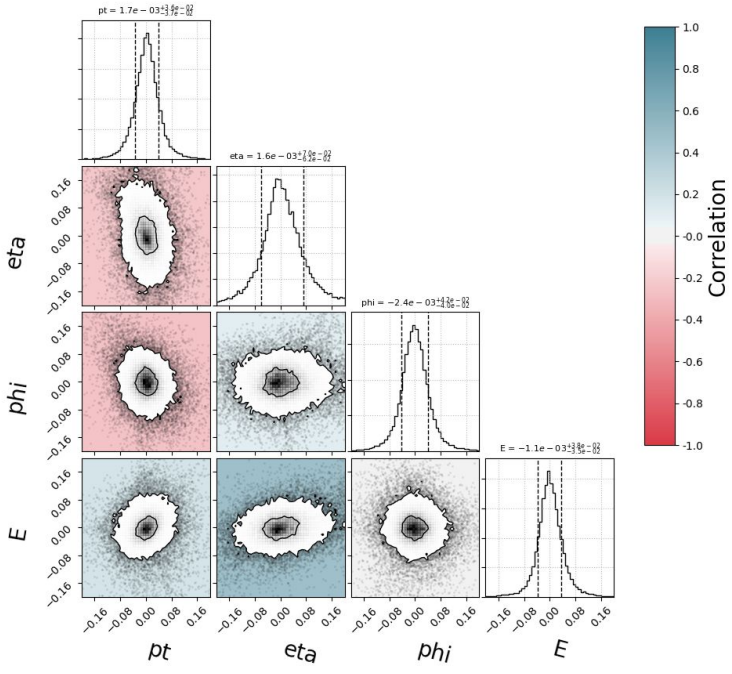
- Model trained on chan2a data and tested on chan3 data
    - MSE = [0.00039981445]


Residuals of eta
Mean = 0.003284±0.000447
σ = 0.092465±0.000316
Number of jets
$(\eta_{out} - \eta_{in})$


Residuals of phi
Mean = -0.001958±0.000256
σ = 0.052958±0.000181
Number of jets
$(\phi_{out} - \phi_{in})$

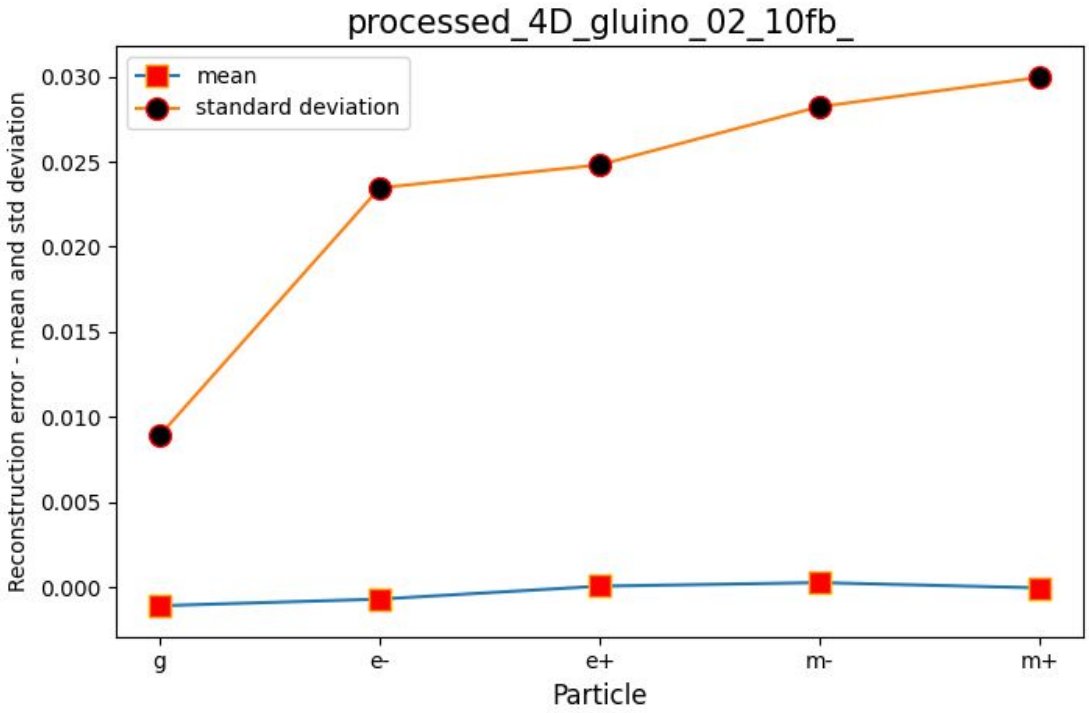

Residuals of pt
Mean = 0.001586±0.000223
σ = 0.046195±0.000158
Number of jets
$(p_{T,out} - p_{T,in})/p_{T,in}$

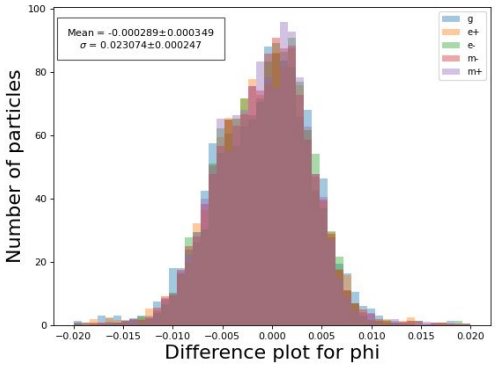

Residuals of E
Mean = -0.001327±0.000245
σ = 0.050742±0.000174
Number of jets
$(E_{out} - E_{in})/E_{in}$

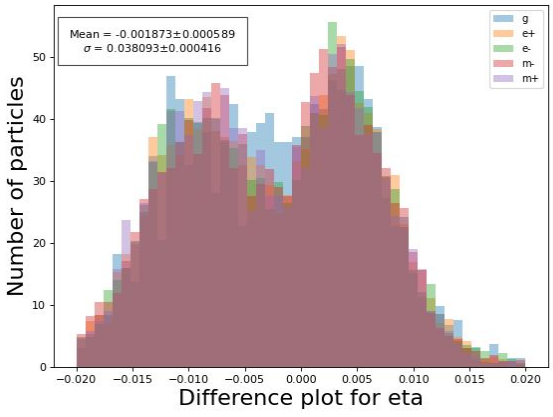

pt
eta
phi
E
Correlation

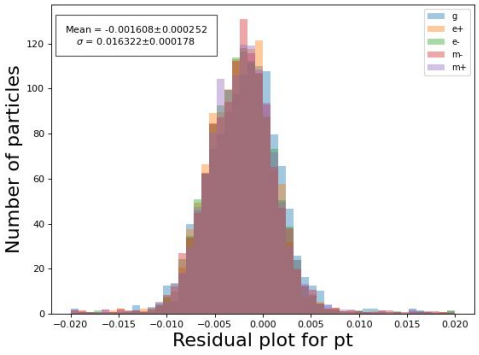
Aug 17

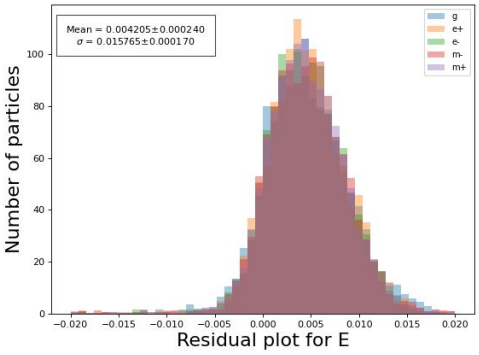
- Created plots for data distribution and particle distribution for darkmachines dataset
- Training a model with articles from chan2a and will test on chan3
- Changes in previous plots - **all plots uploaded on drive**


processed_4D_gluino_02_10fb_
mean
standard deviation
Reconstruction error - mean and std deviation
0.030
0.025
0.020
0.015
0.010
0.005
0.000
g
e-
e+
m-
m+
Particle


Mean = -0.000289±0.000349
σ = 0.023074±0.000247
g
e+
e-
m-
m+
Number of particles
Difference plot for phi

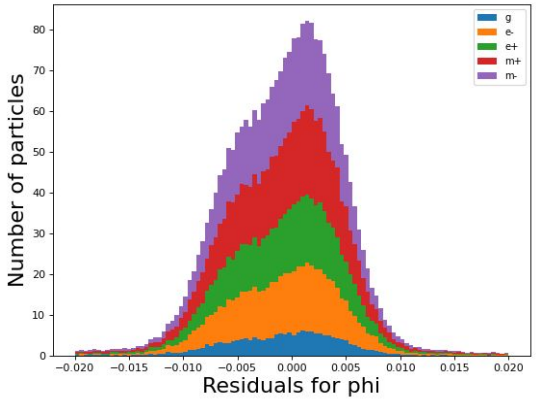

Mean = -0.001873±0.000589
σ = 0.038093±0.000416
g
e+
e-
m-
m+
Number of particles
Difference plot for eta

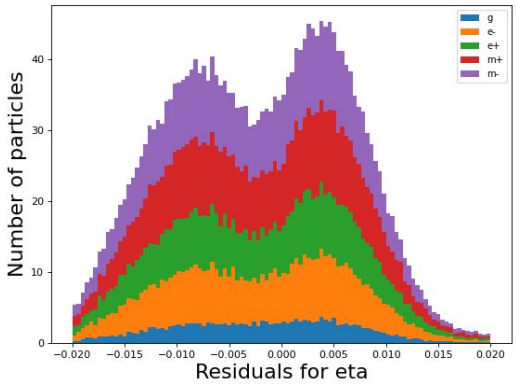

Mean = -0.001608±0.000252
σ = 0.016322±0.000178
g
e+
e-
m-
m+
Number of particles
Residual plot for pt

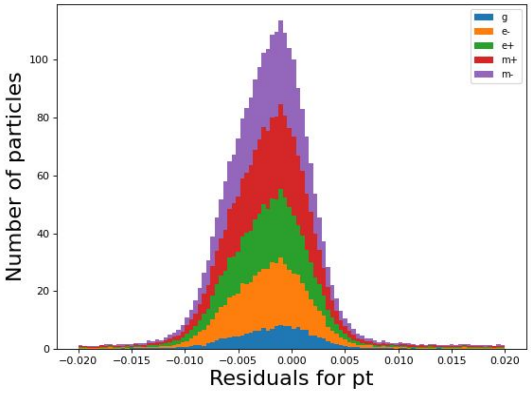

Mean = 0.004205±0.000240
σ = 0.015765±0.000170
g
e+
e-
m-
m+
Number of particles
Residual plot for E

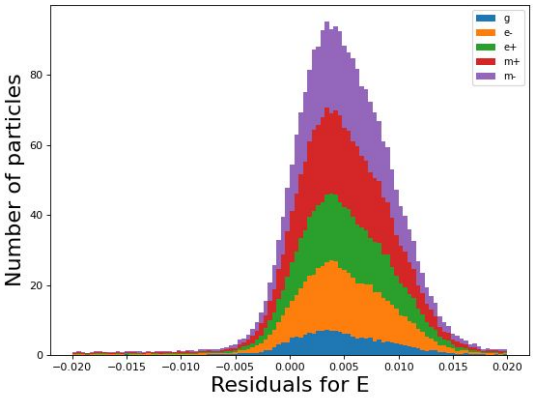
Aug 13

- Stop_02
- https://drive.google.com/drive/folders/1JjEZEBYCSaQcGrIuKhNetv5o0d0no9SN?usp=sharing

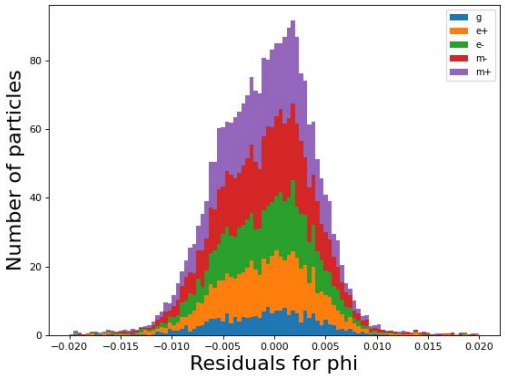


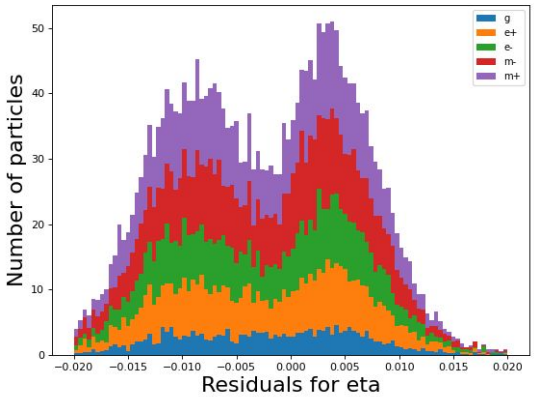


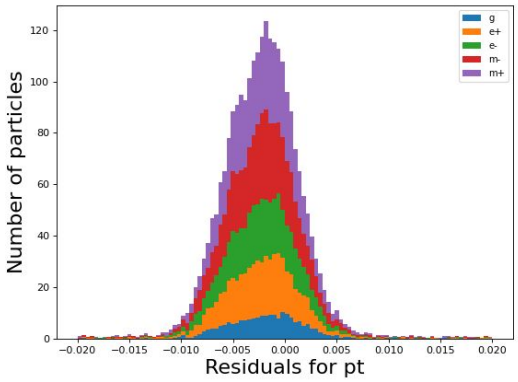


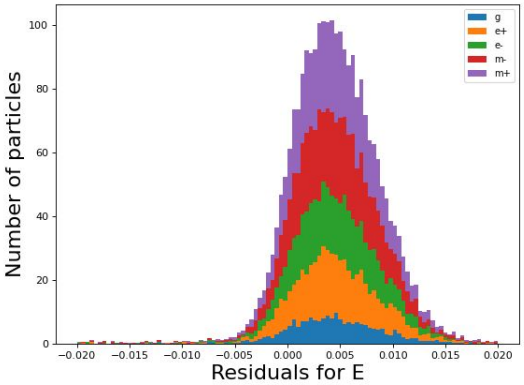


- gluino_02
- https://drive.google.com/drive/folders/1FQxUQbWehHbrTzhEJ3Hb5Vo-JJuau80U?usp=sharing









Aug 11

- stop_02_p_p_to_t1_t1~0_5.69774999996_48.csv
  - e+
    - MSE: [8.7379005e-05]
  - e-
    - [8.628796e-05]
  - m+
    - [9.5003474e-05]
  - m-
    - [0.000113431066]
  - g
    - [1.1838656e-05]
- gluino_02
  - e+
    - [6.45536e-05]
  - e-
    - [5.0873867e-05]
  - m+

        - [9.127547e-05]
    - m-
        - [7.6756885e-05]
    - g
        - [1.2573055e-05]

Aug 5

- Data distribution for different process in phenoML data can be found here:
    - https://drive.google.com/drive/folders/1geNKLyFq2vIyomvFkBGpFWUzLYwncYcz?usp=sharing

Aug 4

- Added the plots in slides
    - SM processes: Slide 55 - 57
    - BSM processes: Slide 58 - 61

Aug 3

- Plots for particle distribution for different sm and bsm data files in the PhenoML dataset

July 28

- List of files in the dataset:
- SM
    - 2gam_10fb.csv
    - 4top_10fb.csv
    - atop_10fb.csv
    - atopbar_10fb.csv
    - gam_jets_10fb.csv
    - njets_10fb.csv
    - single_higgs_10fb.csv
    - single_top_10fb.csv
    - single_topbar_10fb.csv
    - ttbar_10fb.csv
    - ttbarGam_10fb.csv
    - ttbarHiggs_10fb.csv
    - ttbarW_10fb.csv
    - ttbarWW_10fb.csv
    - ttbarZ_10fb.csv
    - Wgam_10fb.csv
    - w_jets_10fb.csv
    - wtop_10fb.csv
    - wtopbar_10fb.csv
    - ww_10fb.csv

    - Zgam_10fb.csv
    - z_jets_10fb.csv
    - ztop_10fb.csv
    - ztopbar_10fb.csv
    - zw_10fb.csv
    - zz_10fb.csv
- BSM
    - gluino_01_p_p_to_go_go_0_0.2013275_21.csv
    - gluino_02_p_p_to_go_go_0_0.0508105_30.csv
    - gluino_03_p_p_to_go_go_0_0.0144098_39.csv
    - gluino_04_p_p_to_go_go_0_0.00442036_48.csv
    - gluino_05_p_p_to_go_go_0_0.00143275_84.csv
    - gluino_06_p_p_to_go_go_0_0.0004843405_66.csv
    - gluino_07_p_p_to_go_go_0_0.000168185_75.csv
    - stop_01_p_p_to_t1_t1~0_26.7494500003_39.csv
    - stop_02_p_p_to_t1_t1~0_5.69774999996_48.csv
    - stop_03_p_p_to_t1_t1~0_1.2483025_75.csv
    - stop_04_p_p_to_t1_t1~0_0.0200922000001_84.csv
    - Zp_technicol_01_0.3865.csv
    - Zp_technicol_02_0.12206.csv
    - Zp_technicol_03_0.044272.csv
    - Zp_technicol_04_0.017957.csv
    - Zp_technicol_05_0.00807869999999.csv

July 27

- Tested the pretrained model on the three datasets created from ttbar
    - 1. Remove events where there is at least one lepton (Slides 48 -49)
    - 2. All events but only jet data (Slides 50-51)
    - 3. All data (Slides 52-54)

July 26

- Disk space issue, removed some old dataset
- Creating datasets from ttbar - taking only first 100k events

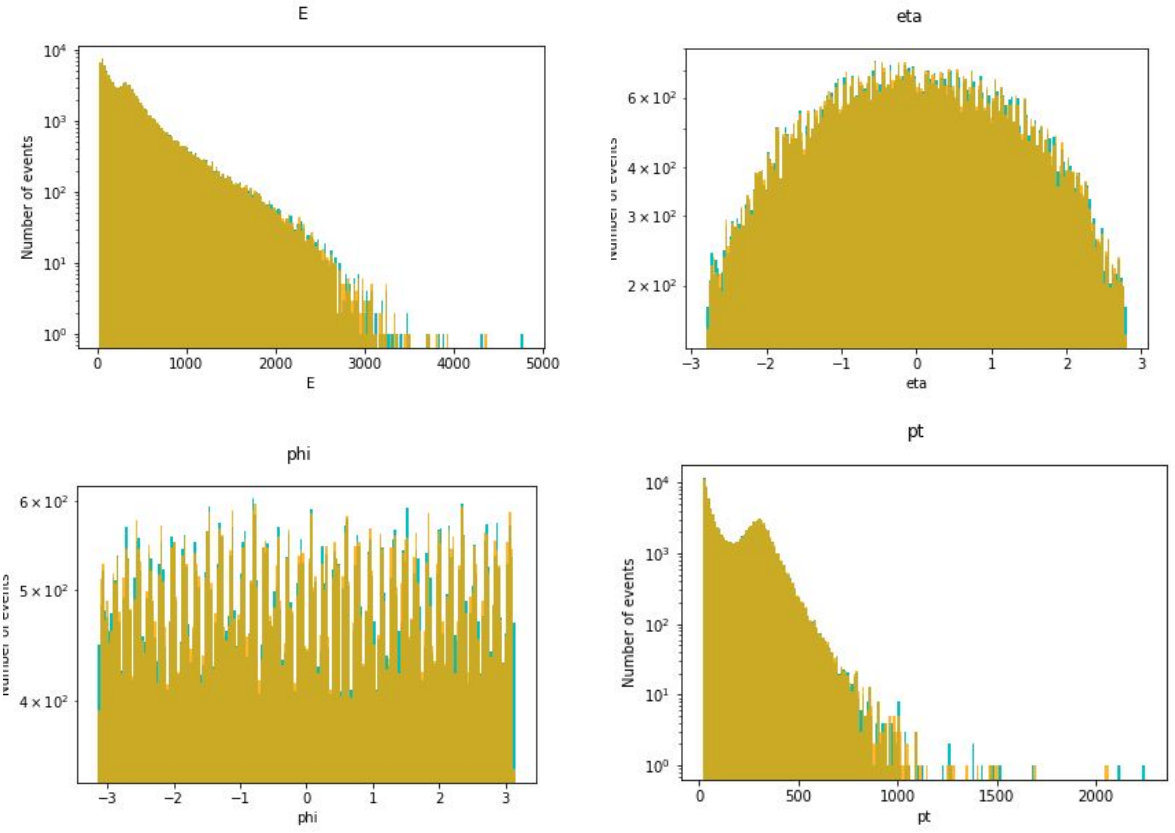
July 25

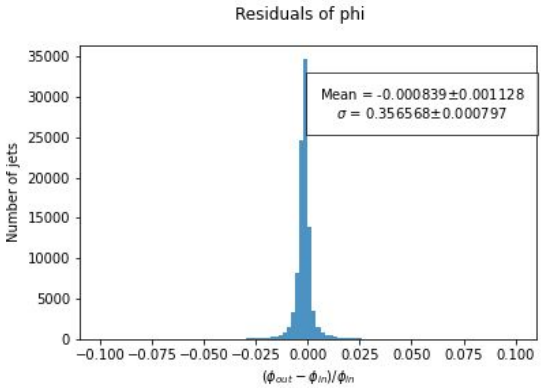
- Extracted all files from sm.tgz
- Creating the 3 types of dataset for ttbar
- Getting “disk quota exceeded” error

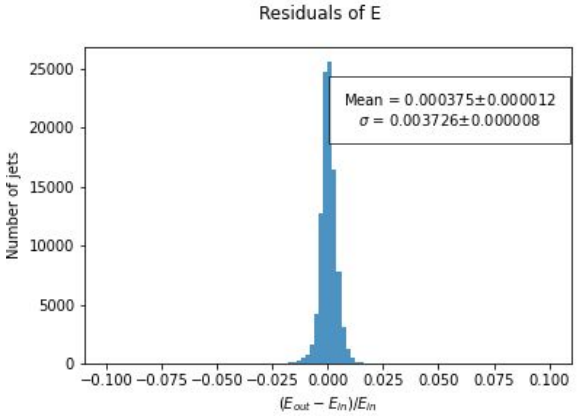
July 24

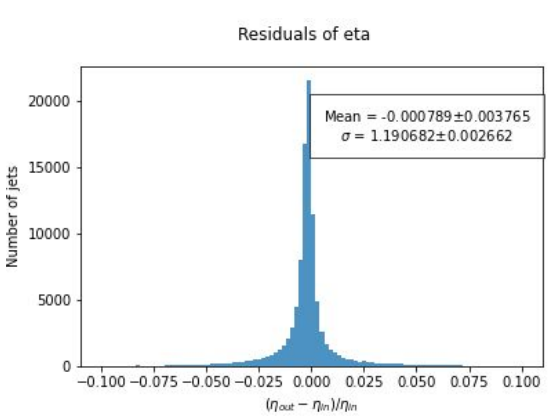
- Models take around 3 days to train for the 500MB dataset

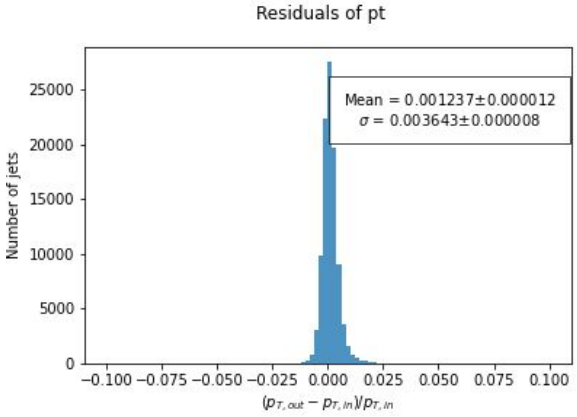
    - Maybe use only half of the training data?
    - Changed figures and numbers for full dataset in slides - (Slides 36-37)
    - Not much difference between MSE on test-set
- Tested the model in atop
    - Created 3 versions of the data
        - 1. Remove events where there is at least one lepton (Slides 38-39)
        - 2. All events but only jet data (Slides 40-41)
        - 3. All data (Slides 42-44)
- Tested the model in atopbar - all events and all particles
    - (Slides 45-47)
- Plots in slides and drive
- Figured out the issue with missing ttbar file
    - Untar takes a long time and login expires :/
    - Submitted a job on HTCondor to decompress the tar file
- Custom normalisation
    - MSE on test-set: [2.4856113e-06]
    - Plots

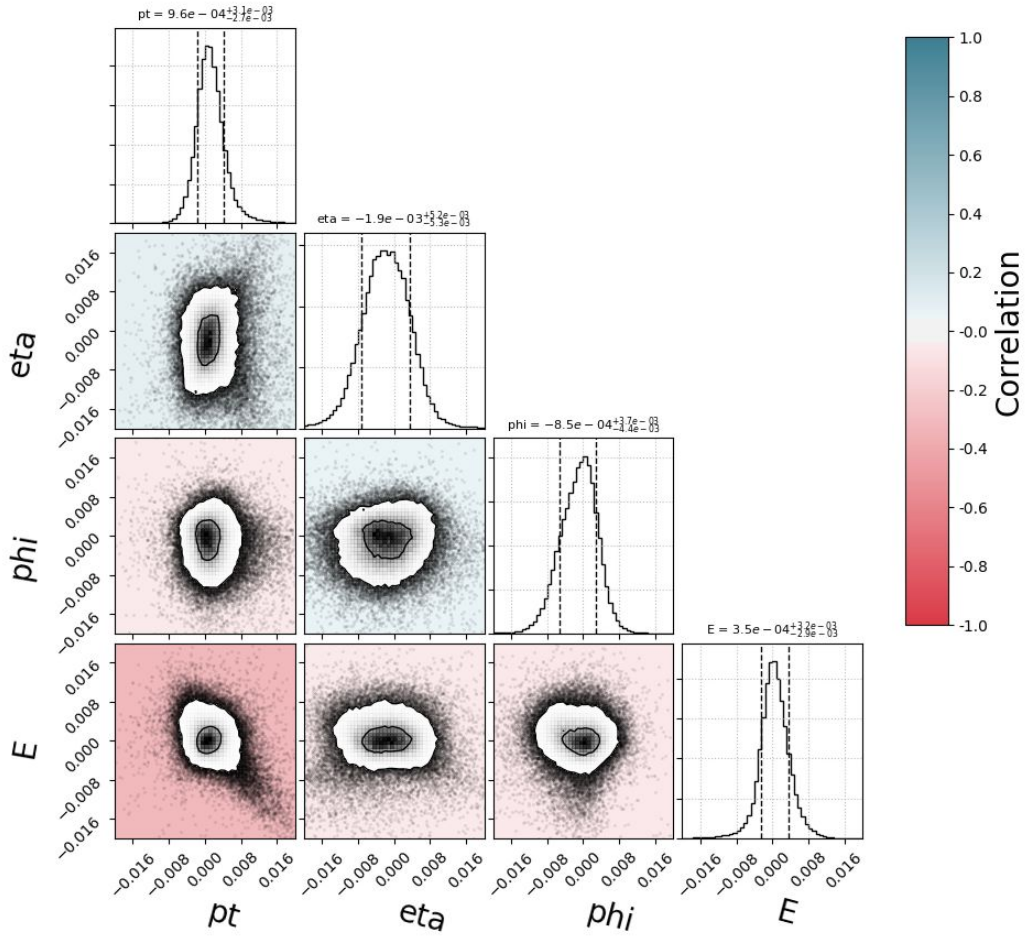



Residuals of phi
Mean = -0.000839±0.001128
σ = 0.356568±0.000797
Number of jets
$(\phi_{out} - \phi_{in})/\phi_{in}$

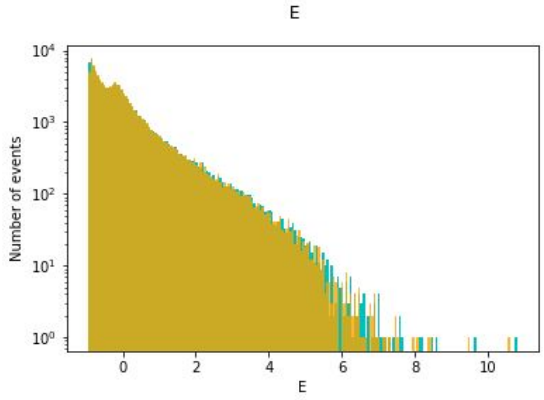

Residuals of E
Mean = 0.000375±0.000012
σ = 0.003726±0.000008
Number of jets
$(E_{out} - E_{in})/E_{in}$

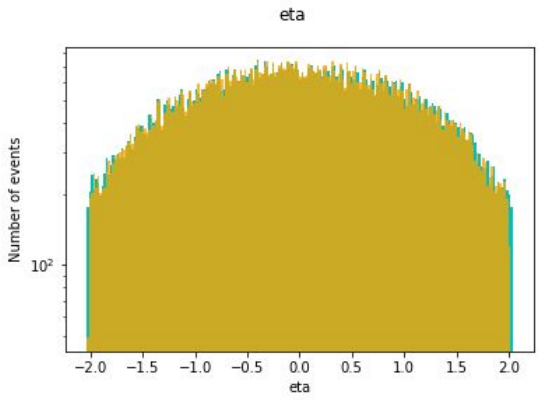


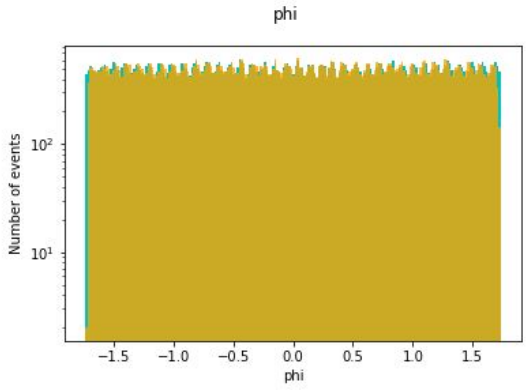


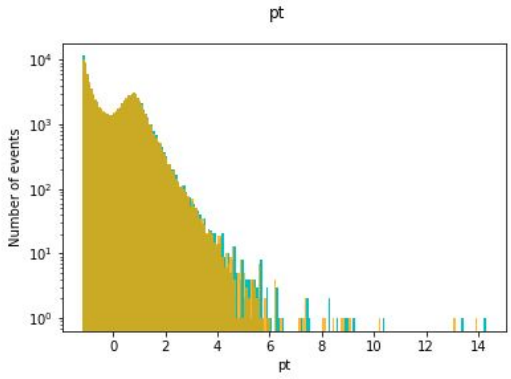


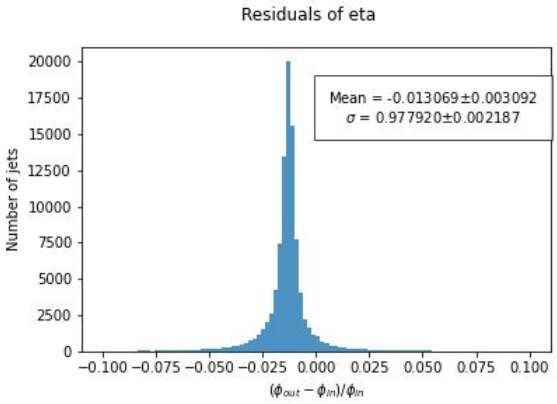
- Standard normalisation
  - MSE on test-set: [0.00019922169]
  - Plots

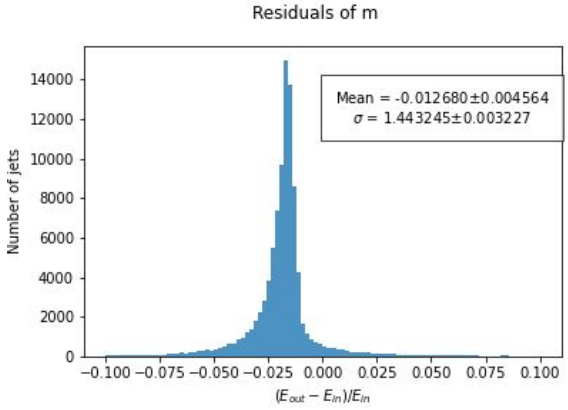

E
Number of events
E

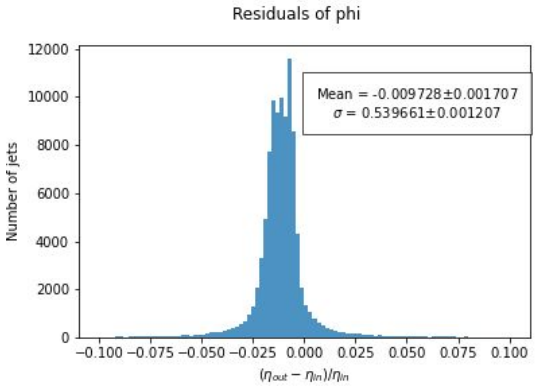

eta
Number of events
eta

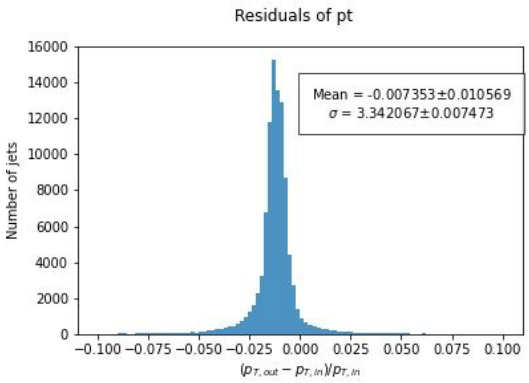

phi
Number of events
phi

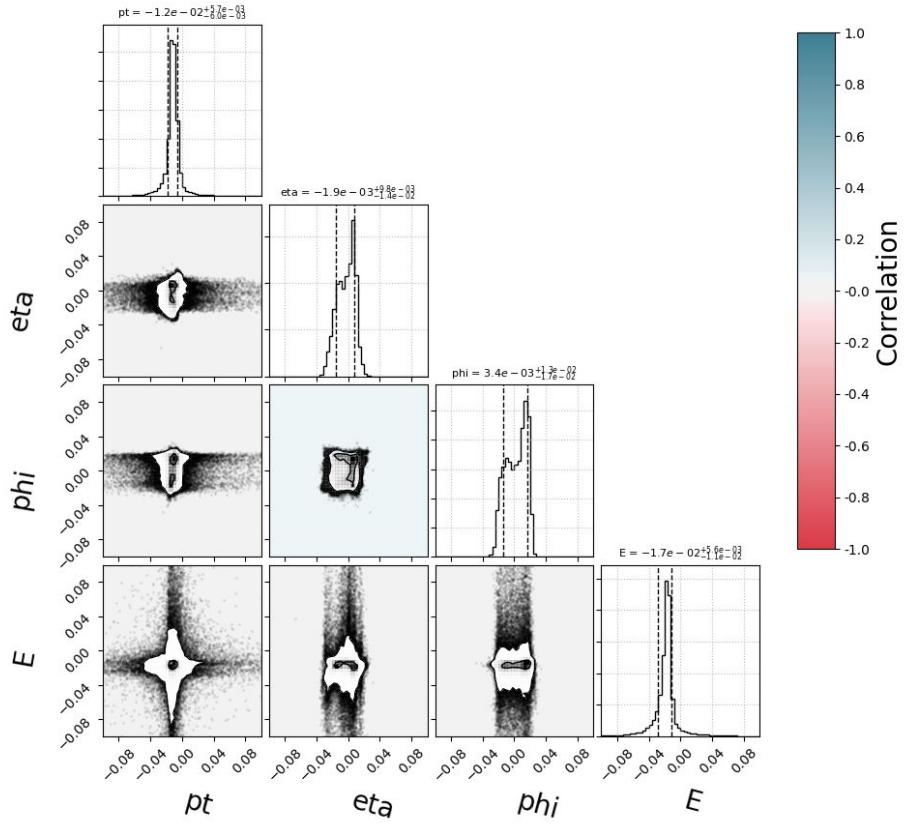

pt
Number of events
pt


Residuals of eta
Mean = -0.013069±0.003092
σ = 0.977920±0.002187
Number of jets
$(\phi_{out} - \phi_{in})/\phi_{in}$


Residuals of m
Mean = -0.012680±0.004564
σ = 1.443245±0.003227
Number of jets
$(E_{out} - E_{in})/E_{in}$



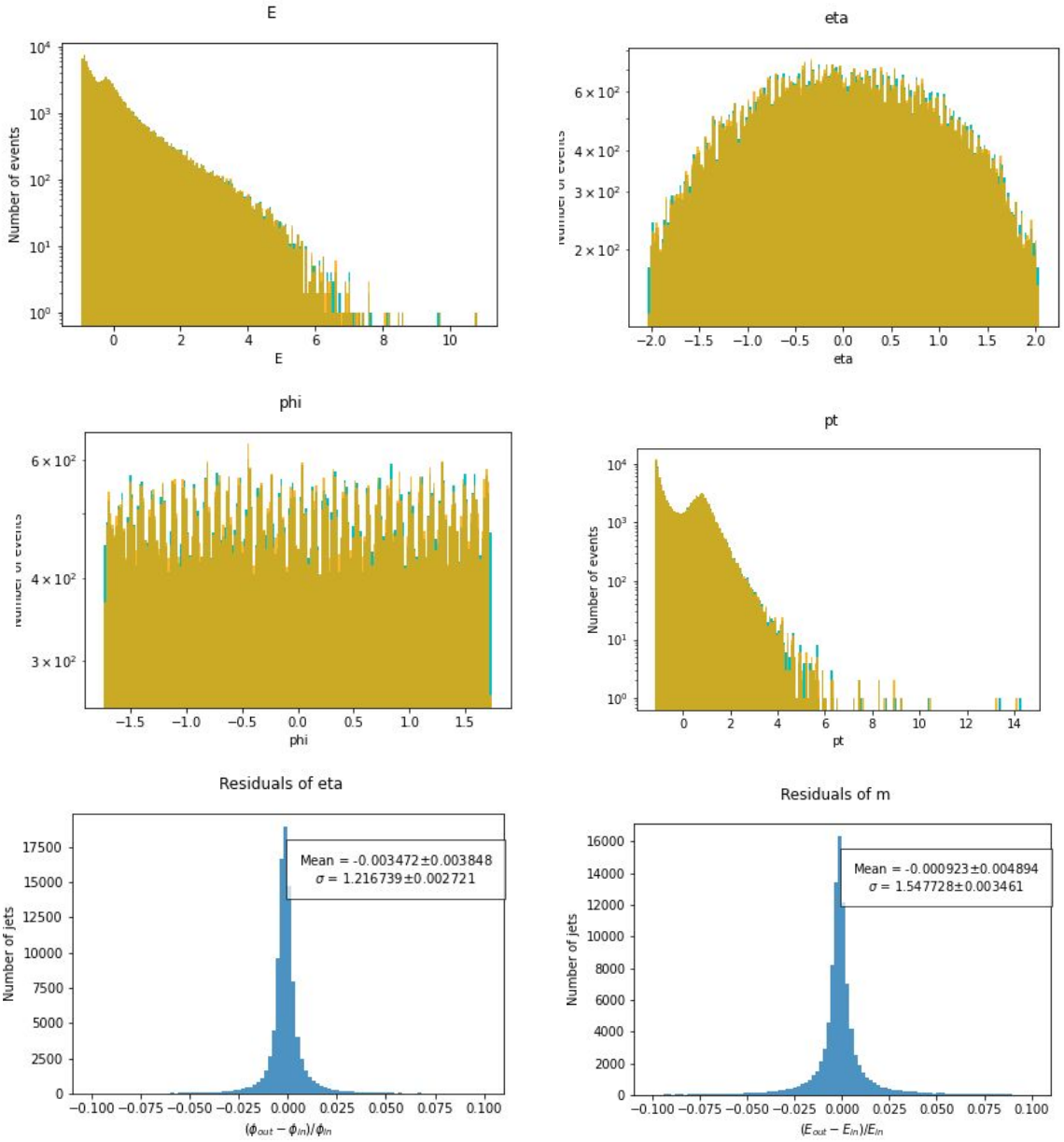


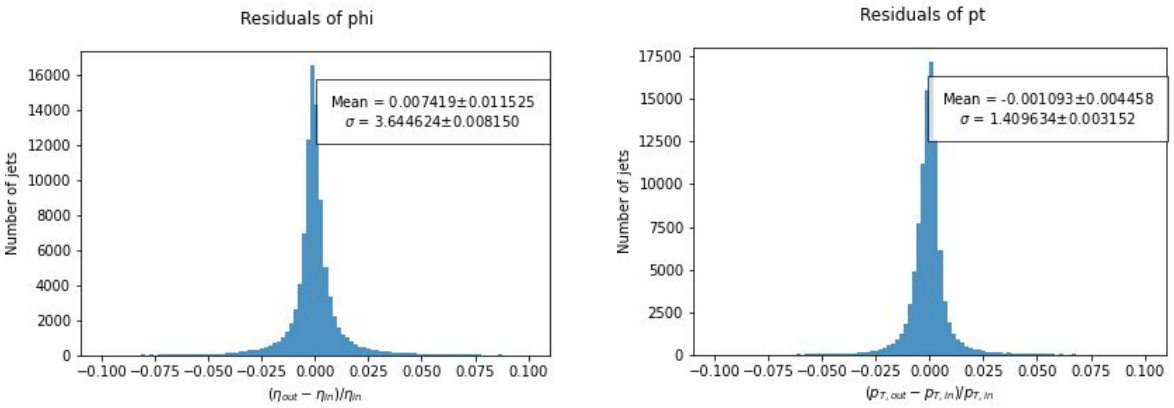


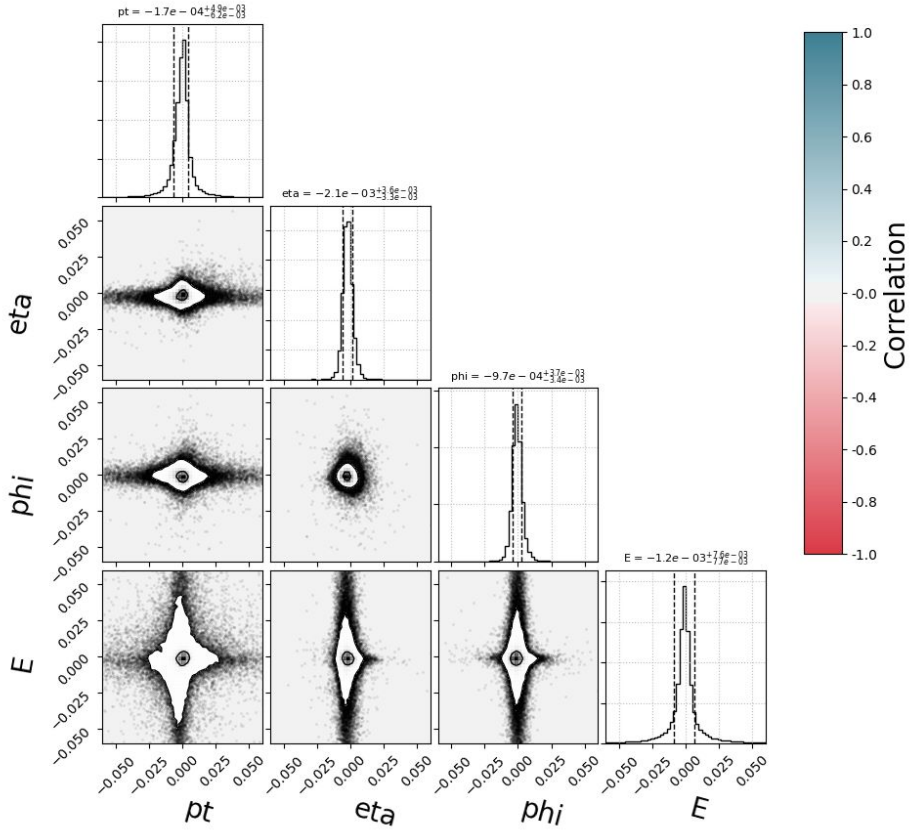
July 20

- Results show that model trained with half train-set is better than the full train set
    - Either the data is conflicting
    - Or the full train model did not train properly
        - The codes timed out after 1 day - maybe not enough epochs were done
        - No way to find out, as of now
- Resubmitted the jobs for training with full train-set

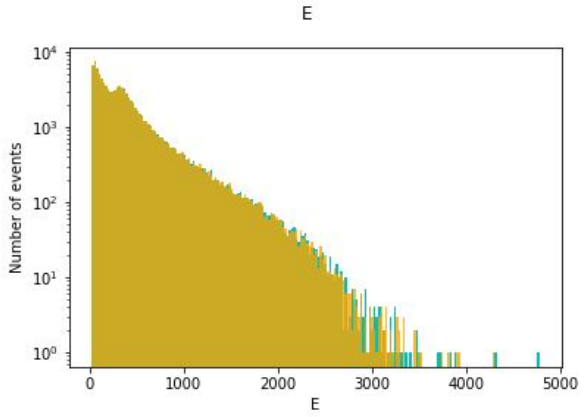
    - Extended the max time to 3 days
- Test results with half the training data
- **Std-norm**
    - Test-set MSE : [2.7026193e-05]
    - Plots:

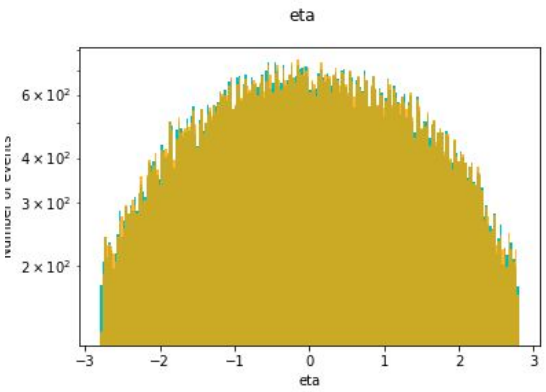


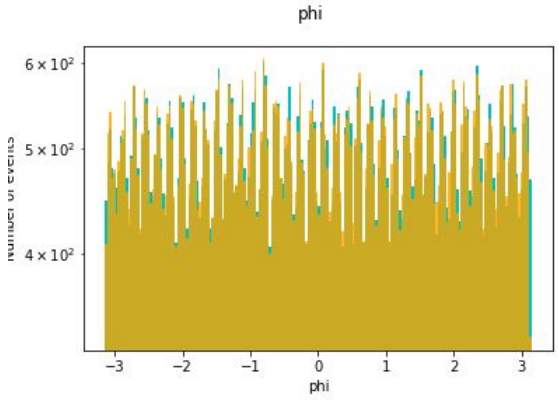


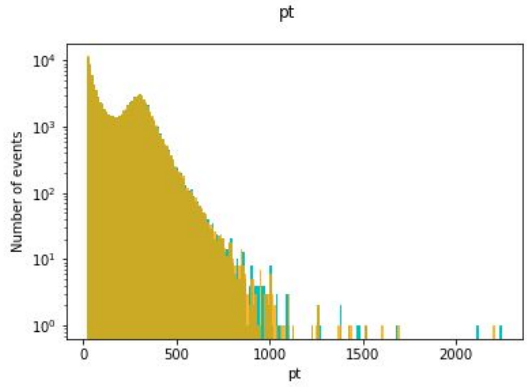


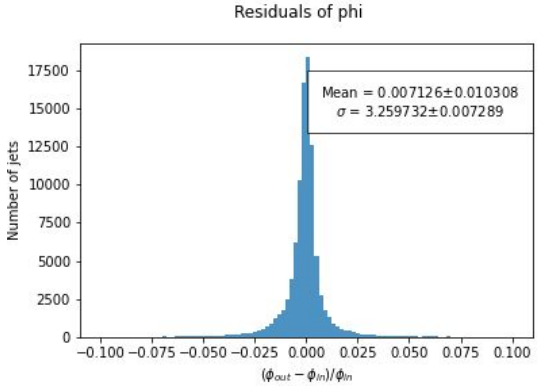
- **Custom-norm**
  - Test-set MSE : [8.7358785e-06]
  - Plots:

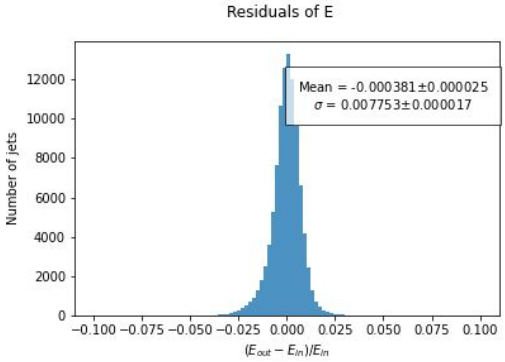

E
Number of events
E

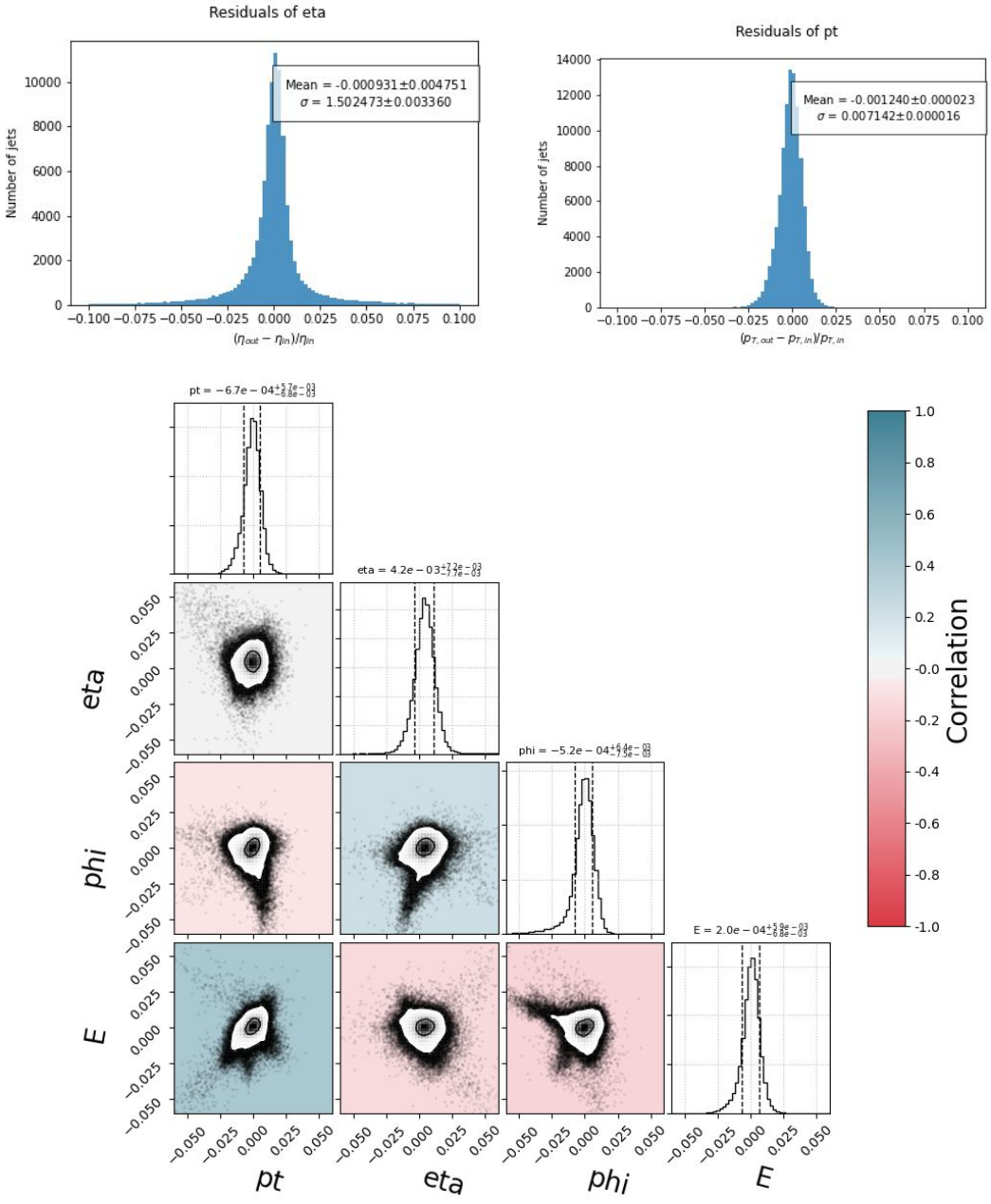

eta
eta


phi
phi


pt
Number of events
pt


Residuals of phi
Mean = 0.007126±0.010308
σ = 3.259732±0.007289
Number of jets
(φout − φin)/φin


Residuals of E
Mean = -0.000381±0.000025
σ = 0.007753±0.000017
Number of jets
(Eout − Ein)/Ein



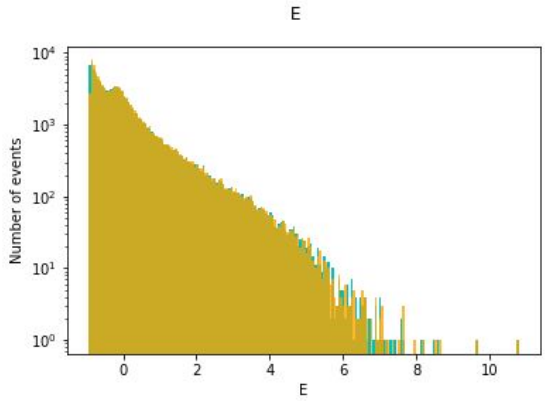
July 17

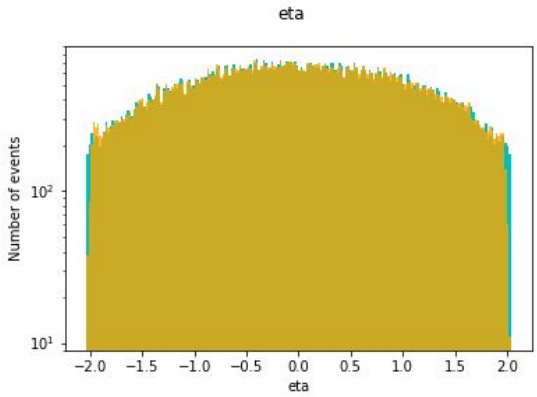
- Job exited - incomplete training
- Resubmitted

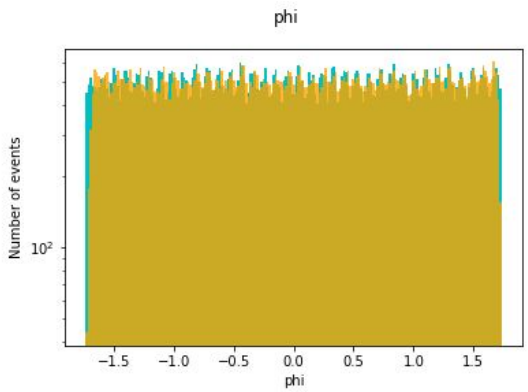
July 16

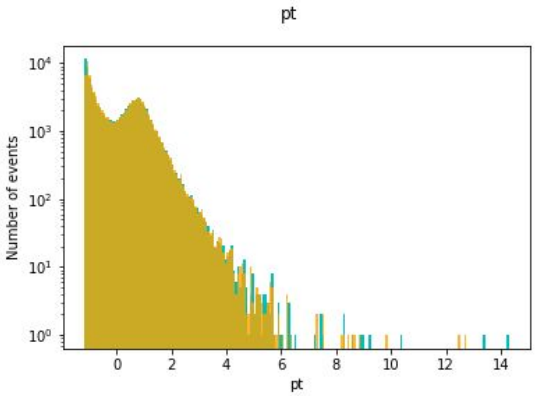
- Submitted jobs for training with half (of 500MB) training data

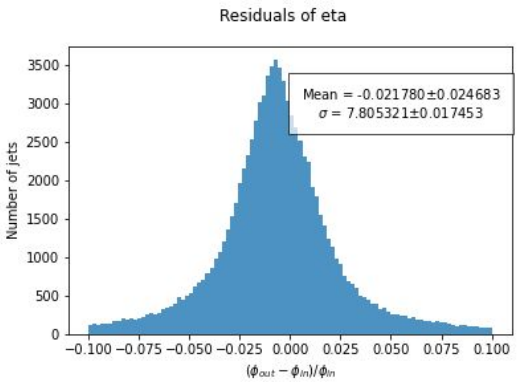
- Results with ~500MB of training data:
  - **Std-norm**
    - Test-set MSE :[0.0008520562]
    - Plots:

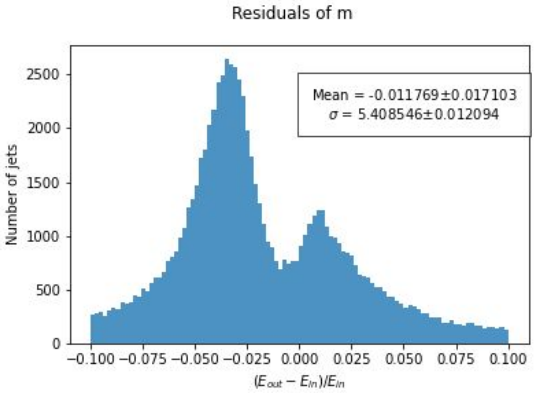


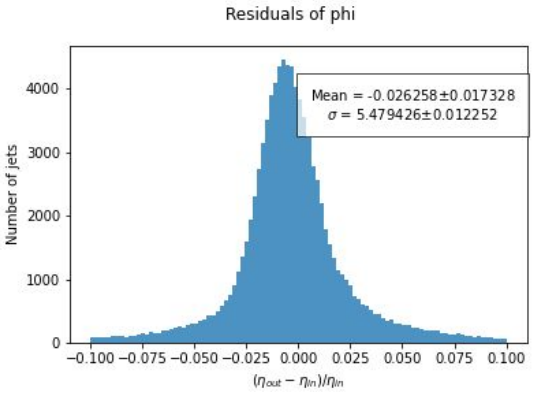


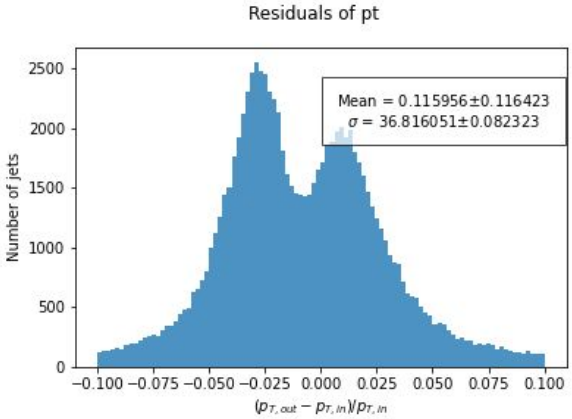


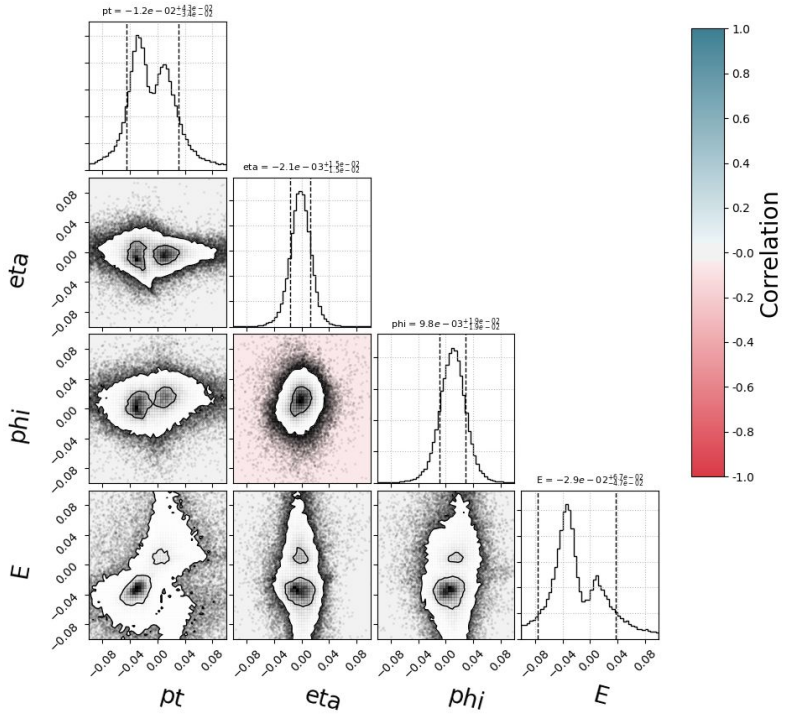




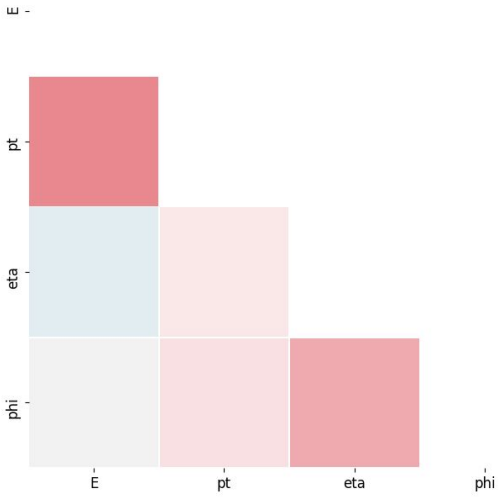


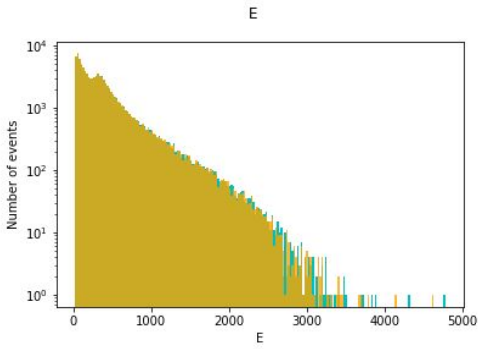


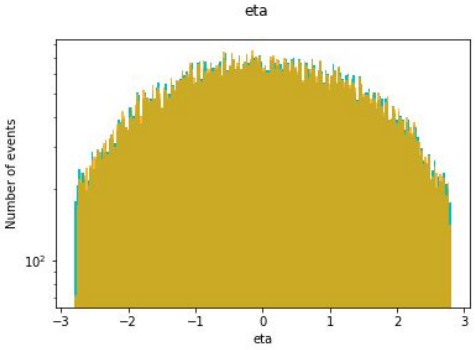


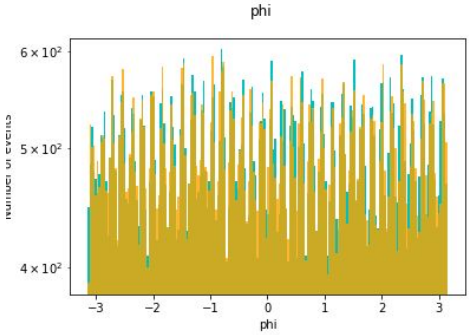


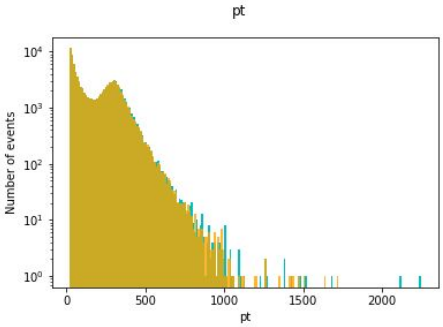
- **Custom-norm**
  - Test-set MSE : [3.0818668e-05]
  - Plots:

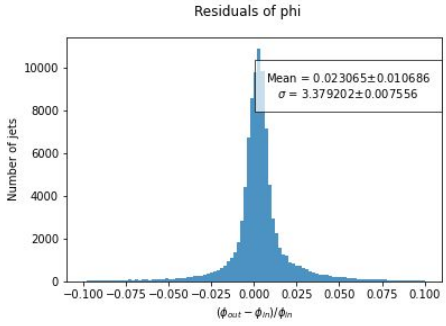

E
pt
eta
phi
E
pt
eta
phi
1.00
0.75
0.50
0.25
0.00

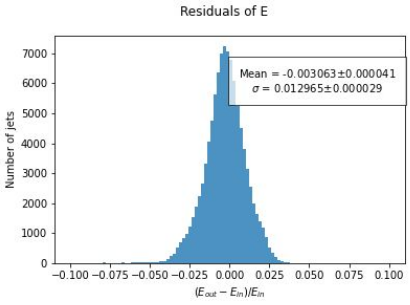

E
Number of events
E

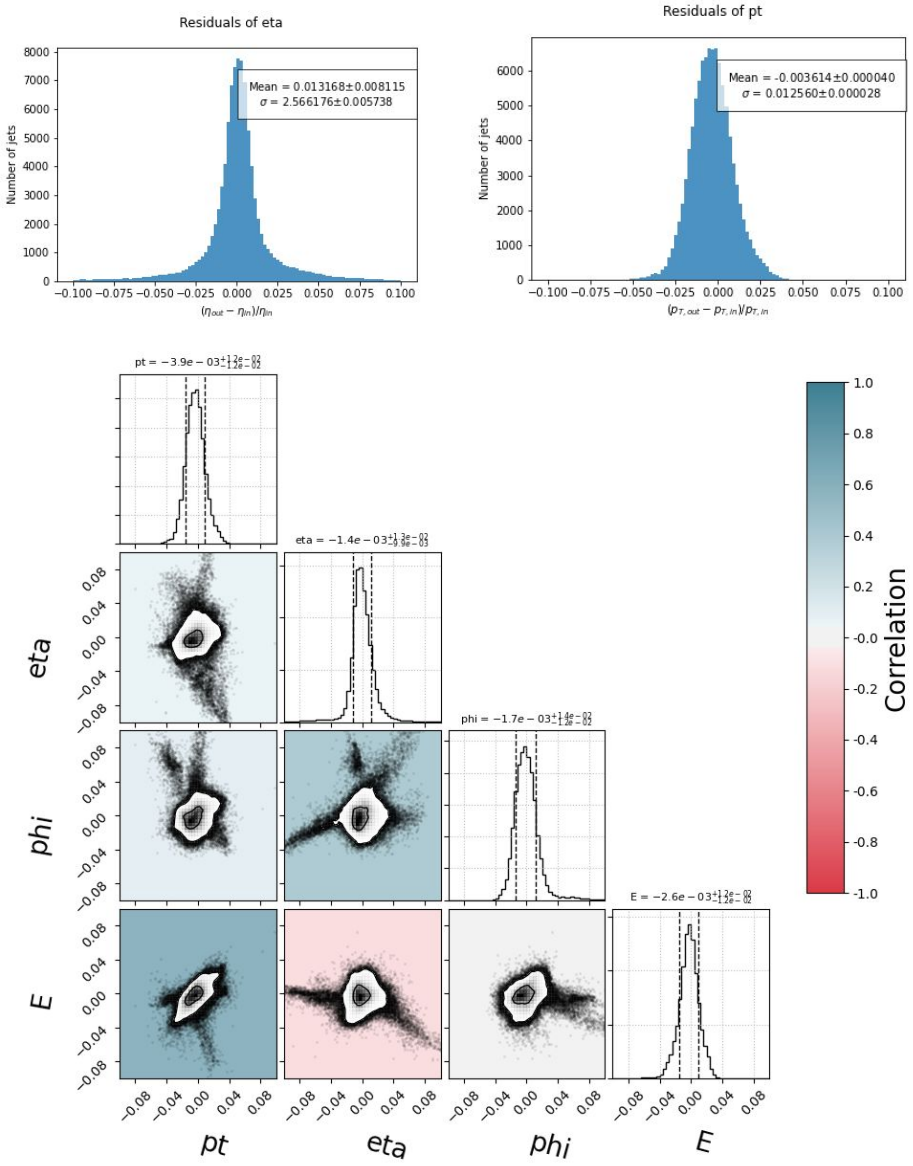

eta
Number of events
eta


phi
phi


pt
Number of events
pt


Residuals of phi
Mean = 0.023065±0.010686
σ = 3.379202±0.007556
Number of jets
$(\phi_{out} - \phi_{in})/\phi_{in}$


Residuals of E
Mean = -0.003063±0.000041
σ = 0.012965±0.000029
Number of jets
$(E_{out} - E_{in})/E_{in}$

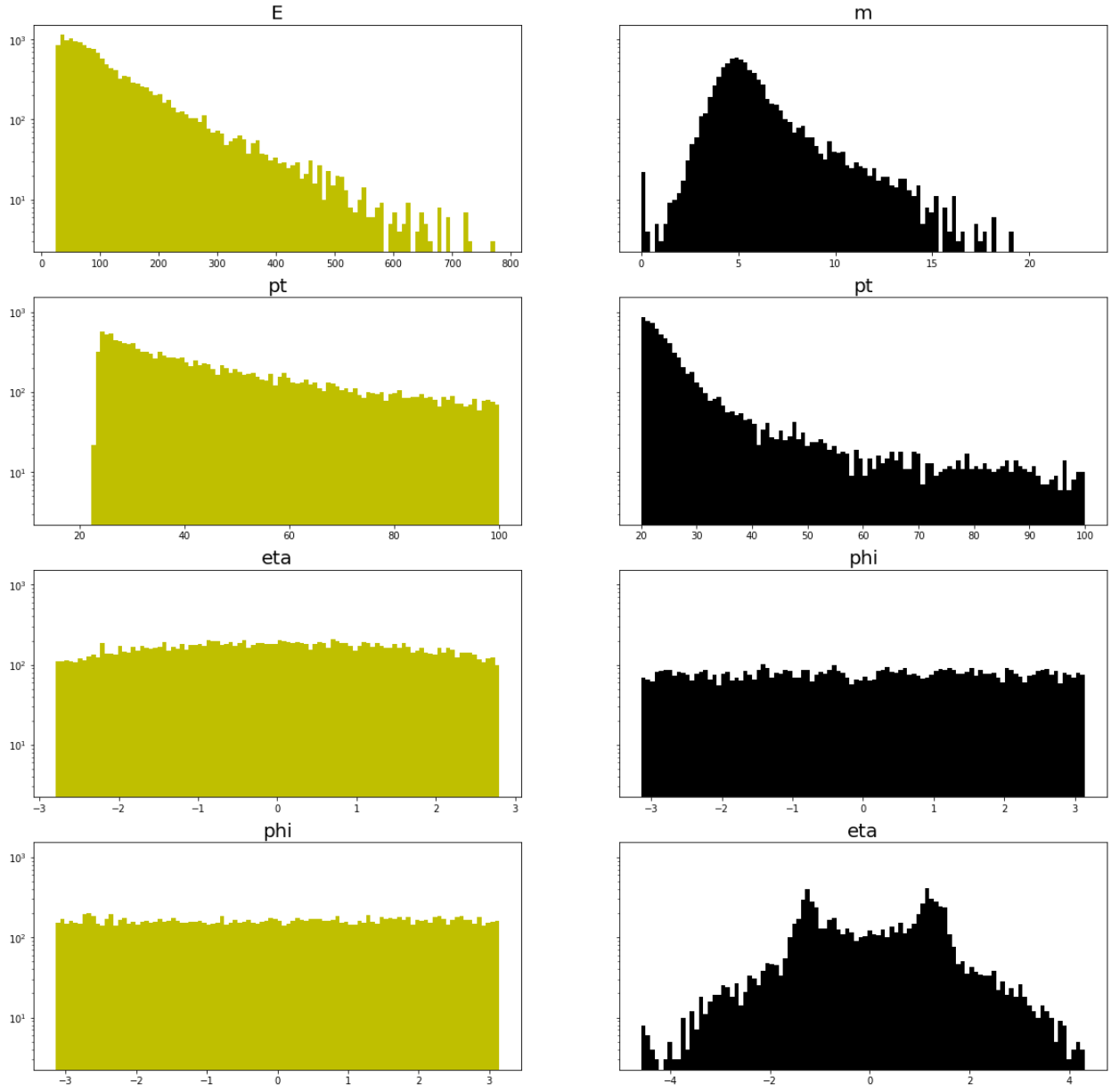


July 15

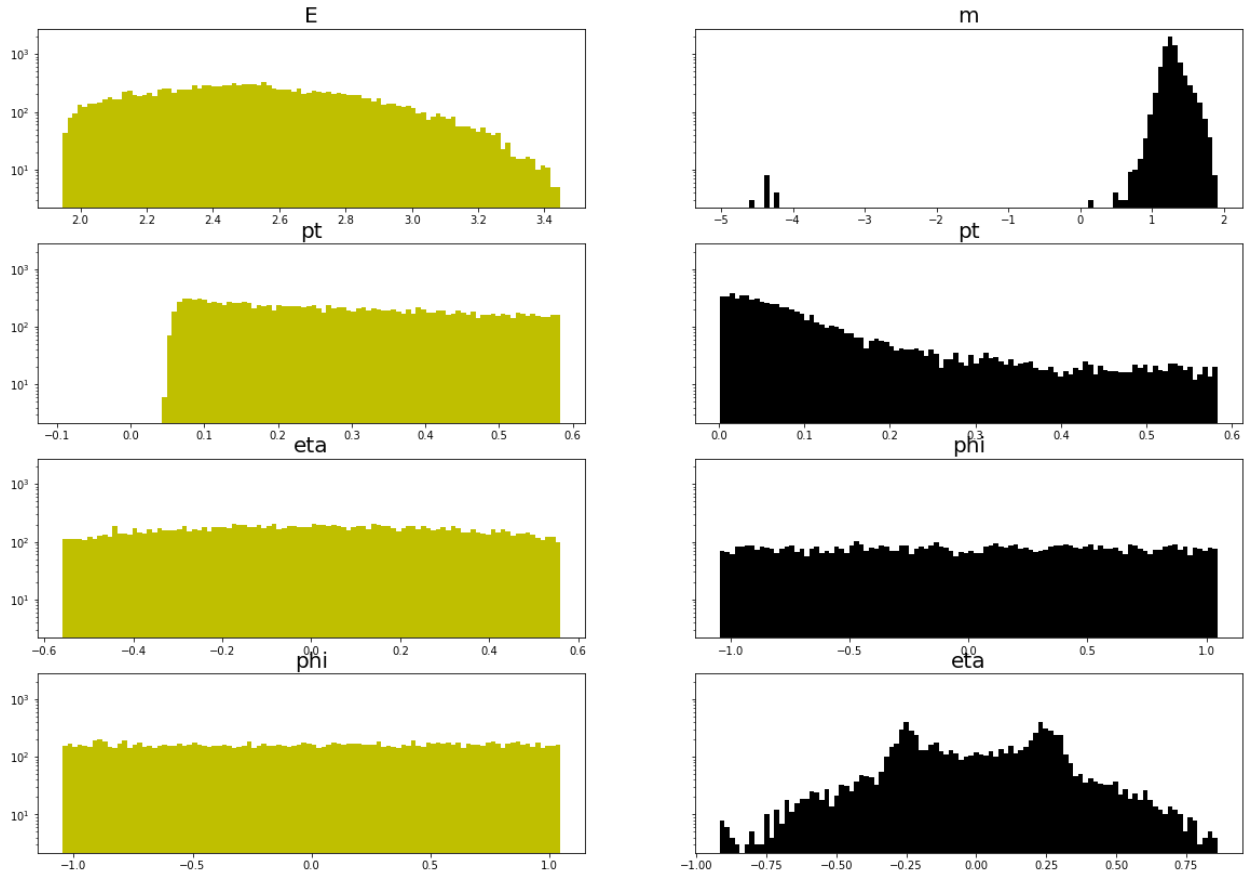
- Submitted a job to process 4D data
- Added the script to train 4D-3D model
- Submitted job to train models with std and custom norm
    - custom norm has only log - no sub or div

July 14

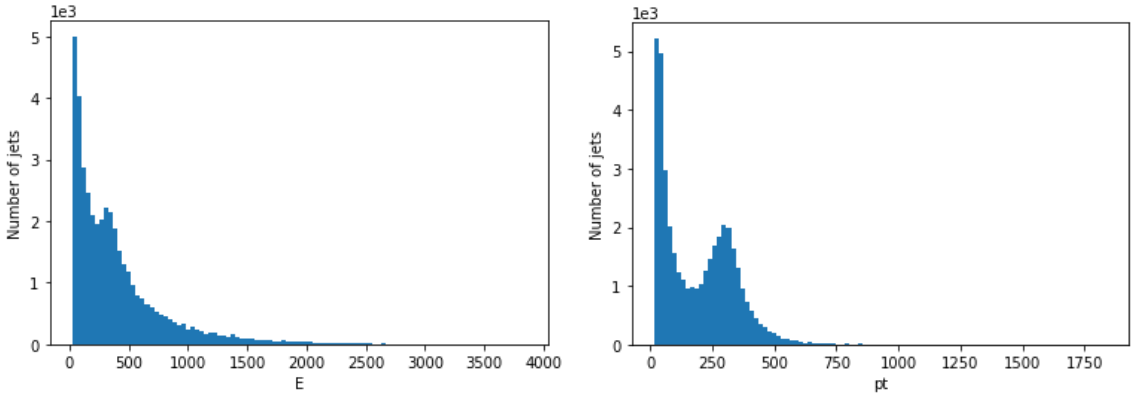
- Comparison of data distribution between phenoML and evaluation data
- w/o normalization

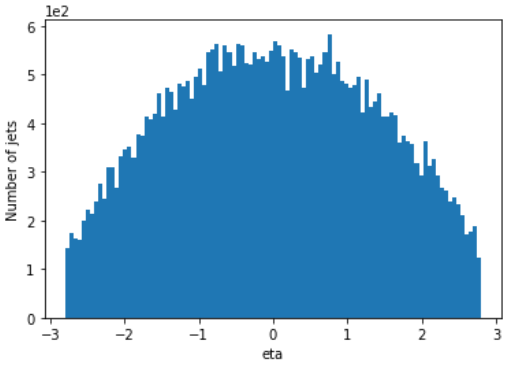


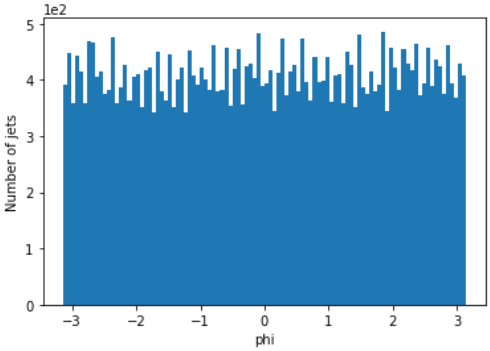
- After custom normalization



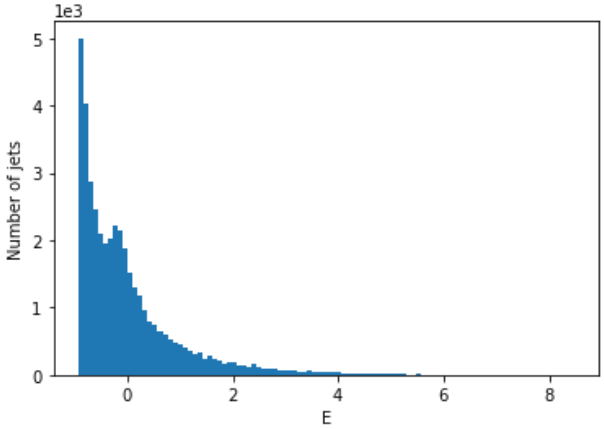
- GSoC data
  - 111778 - train
  -
- Created a smaller file for njets
  - Submitted job to extract as events - with zero padding
- Writing script to read as 4D
- Plots for 4D data distribution (10k samples)
  - E and pt % 1000.

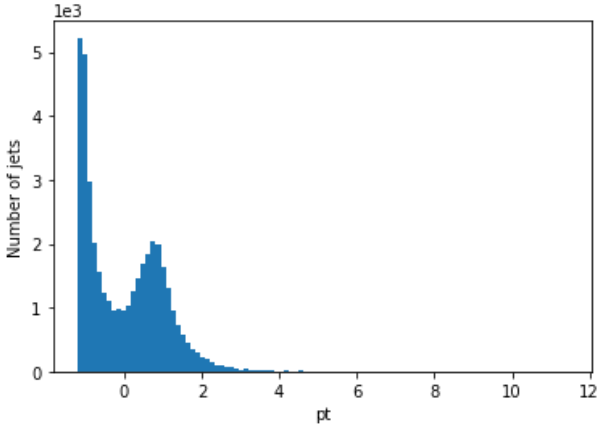


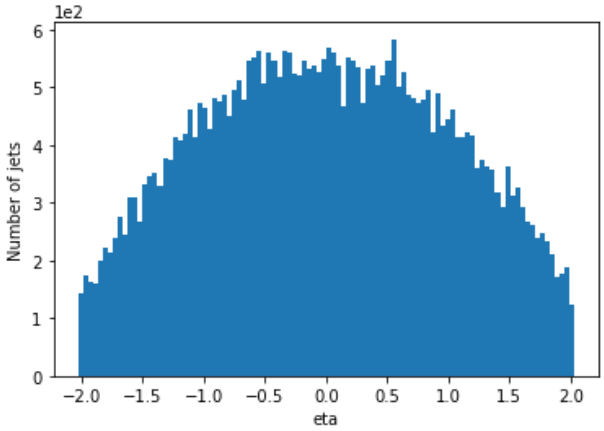


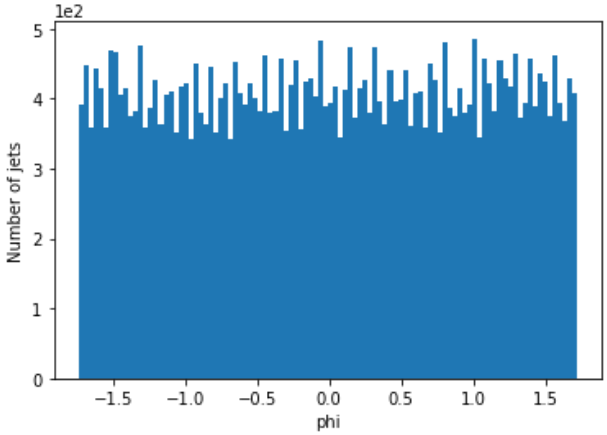


- Standard normalisation

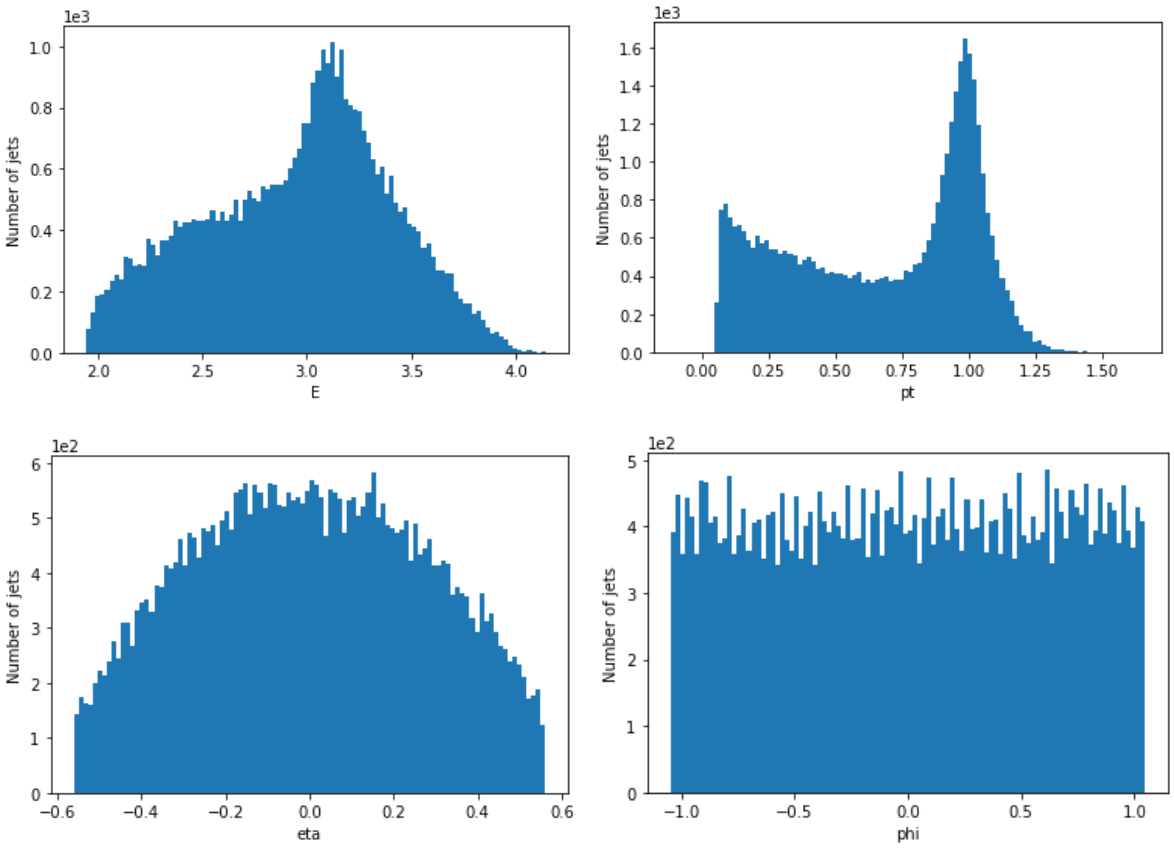








- Custom norm for AOD data -
  - data['eta'] = data['eta'] / 5
  - data['phi'] = data['phi'] / 3
  - data['m'] = np.log10(data['m'] + 1) / 1.8
  - data['pt'] = (np.log10(data['pt']) - 1.3) / 1.2



July 13

- Job with 6 CPUs terminated
    - Memory usage by the code - 17909 MBs
- Resubmitted
    - Process still getting killed -_-

July 12

- Code to process the data not getting queued
    - reduced the no. of CPUs and resubmitted

July 11

- Updated the README for autoencoder repo: merged instructions for install_libs file
    - Updated setup.py file with few package related details
        - https://github.com/Autoencoders-compression-anomaly/AE-Compression-pytorch/commit/e0df729225258f03d63d5f53148ef16f83f6b93f
- Added the code to read the data on GitHub

    - https://github.com/Autoencoders-compression-anomaly/collider-unsupervised-learning/tree/honey_dev/process_data/
- Wrote and committed the code for training a model on the phenoML data
    - https://github.com/Autoencoders-compression-anomaly/AE-Compression-pytorch/tree/honey_dev/examples/phenoML

July 10

- Reading data code
    - Check about jet and b-jet
    - Which approach to take the data - 40 vs 4 (mix events)?
        - Trying both for now
            - **4D - 3D for now**
    - **What is the 3,5 in reshape for?**
    - Normalization?
        - Taking std for now
            - **Formulate custom norm parameters**
- Submitted the code to process data on condor

July 9

- Dataset at https://zenodo.org/record/3685861/files/sm.tgz?download=1
    - Downloaded in AFS

Jul 2-3

- Going through the documentation of PhenoML dataset
    - https://arxiv.org/pdf/2002.12220.pdf
- Going through some physics-related literature provided by Rebeca

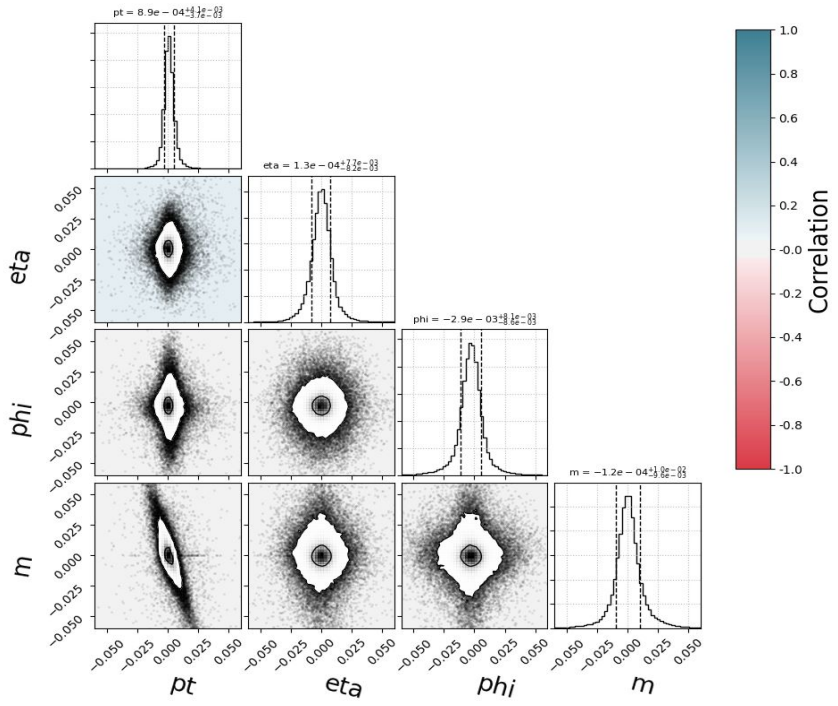
Jul 1

- Updated the presentation as per the suggestions
    - Included few additional slides from the evaluation just for completeness
- Wrote a doc for using Gpus on HTCondor: https://docs.google.com/document/d/1zCjPpN80zq57bktLlqsSv9H1tI2lR9_Hr3vNGWkEGgg/edit?usp=sharing

28 Jun

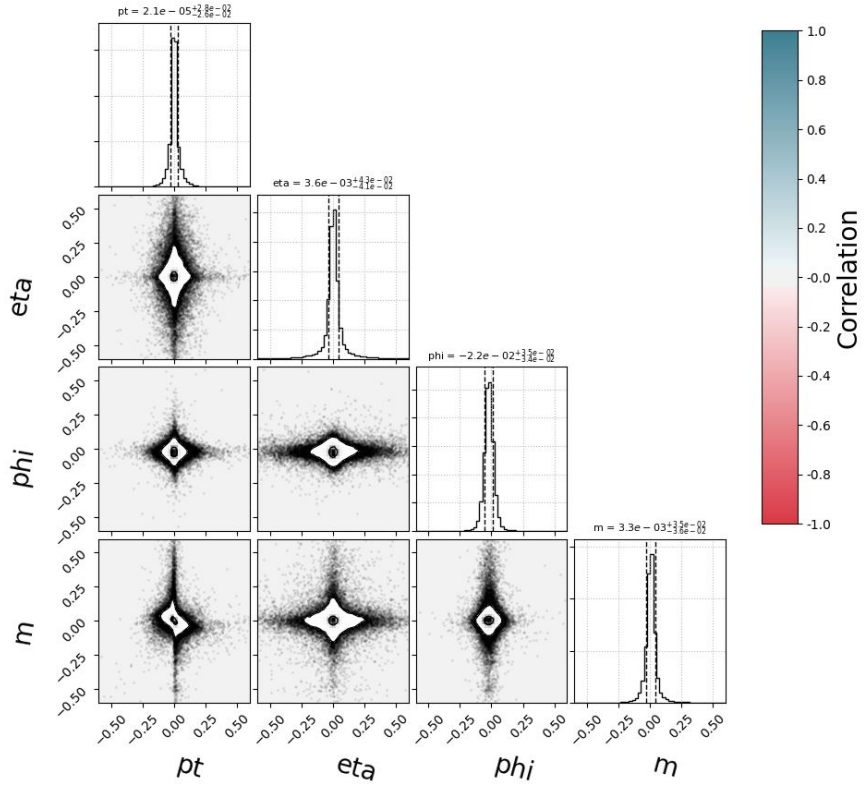
- Updated presentation can be found here: Worklog slides
- Added I/O and corr plots to trello

27 Jun

- Working on the presentation
- Created an Ipynb to test the model and create the plots
    - Fixed the scale for Corr plots
- Tested the custom_norm model
    - https://drive.google.com/drive/folders/1wZ7KwhTXHdX755ct2uVMGgdebO8M20tD?usp=sharing
    - 

- Tested the data for scaled-normalization

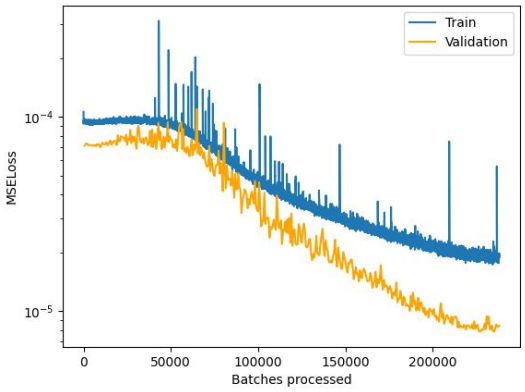
- https://drive.google.com/drive/folders/1m2PN_Ytt_97mAhClOljLELtQKPwhbF2Z?u



sp=sharing

25 Jun

- Custom norm for 100 (previously trained) + 400ep
  - Total time taken: 1275.949 minutes ~ 21:15 hours
  - Time taken per epoch: 3.1898 minutes
  - Validation MSEloss: 7.844627e-06, Training MSELoss: 1.814027e-05



  - **~500ep looks good to train the model**

        - Training MSE ~1e-05/1e-06 can be taken as the stopping point
- Pushed the recent HTCondor files to honey_dev branch

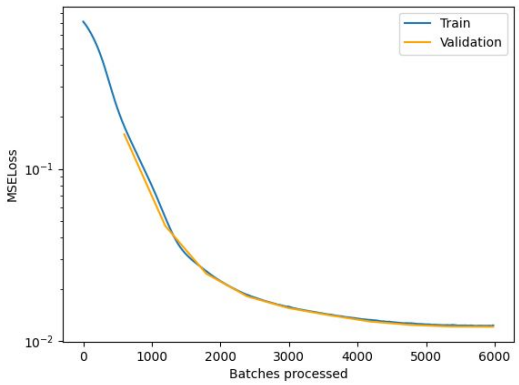
23 Jun

- Custom norm code for 500ep on HTCondor
    - previously trained for 100, retraining that model for 400ep
- TODO
    - Create plots for the trained models
    - Add images to Trello
    - Check webfest
    - Modify code to take LR as an argument and launch the job using multiple LRs
    - More complete presentation with work in this first month
        - Includes work that you've done to document the HTCondor/modifications to the code
    - Summary of results

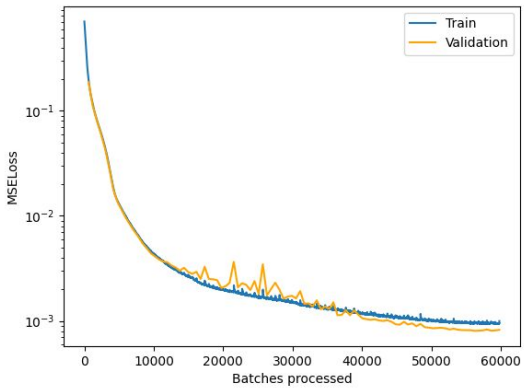
22 Jun

- Wrote scripts to process and scale the data in HTCondor
    - Some issue in copying the created files
    - Added an explicit transfer in the sub file
    - Pushed the codes on GitHub and merged prior codes with master
- Processed the root files
- Was getting TLE for GPU codes
    - +JobFlavour = "workday" → increase the time limit of the job fixed it
    - https://batchdocs.web.cern.ch/local/submit.html
    - condor_rm to remove jobs
- **Pretrained**
    - Eric's thesis has plots for 14D latent space
        - 27-200-200-200-14-200-200-200-27
    - No pre-trained model available for this configuration
    - Model mentioned in the GitHub notebook - 27-400-400-200-18-200-400-400-27.pth for 20D
    - Used pretrained - 200_custom_norm_over_night_all.pth model
        - Residual and corner plots in ppt [link to the presentation, trello]
    - lth_thesis_project/jet_by_jet_compression/aod_compression/aod_custom_normalization_and_test.ipynb
    - Input Data plots match the ones in the thesis
    - MSE = **8.3398e-05** on test-set

- [Scaled data] For 10eps on GPU@HTCondor - **23**mins,
    - Validation MSELoss: 1.210103e-02 Training MSELoss: 1.226083e-02

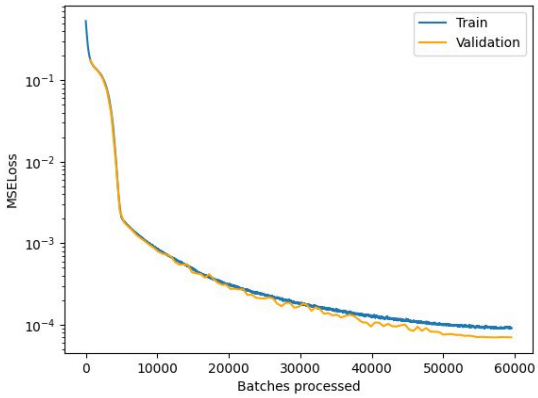


- Considering model **[27, 400, 400, 200, 20, 200, 400, 400, 27]**
- Need to do around **~580** epochs for 350k batches
- [Scaled data] Trained the code for 100ep, LR@1e-4 on HTCondor (everything from now on is GPU)
    - Training time: **4:20 hours**
    - Validation MSELoss: **8.236064e-04** Training MSELoss: 9.428300e-04



- [Custom-norm data] Trained the code for 100ep, LR@1e-4 on HTCondor
    - Training time: **4:42 hours**

- Validation MSELoss: **7.017471e-05** Training MSELoss: 9.206531e-05



19 Jun

- Started working on the AOD files
- Run - 364292
- The file mentioned in the notebooks - DAOD_TRIG6.16825104._000035.pool.root.1
  - 2260895
- The file in the folder - *_000079* and *_000263*
  - 079 - 6117135
  - Check if file matters
    - no

18 Jun

- Had a meeting with Lucas about the doubts regarding HTCondor
  - AFS is visible to the nodes, only the contents of the current folder changes
- Wrote a quick start guide to run a sample python script on HTCondor
  - Link: https://docs.google.com/document/d/1chrlFBSHY6bq46_U5uwGbzEposSyamMLqsX99605UUo/edit?usp=sharing

16 Jun

- Weekly meeting:
  - Got access to the datasets
  - Caterina confirmed from Erik Wallin that the 27d datasets are what Eric Wulff used.
  - Checked the timeline

12,15 Jun

- Worked on tutorials for HTCondor
  - https://batchdocs.web.cern.ch/tutorial/exercise10.html

- Numpy one gives as error while installation
- virtualenv and tf works well
- Docker - error: docker image tensorflow/tensorflow:latest-gpu not found
- Tried converting the 27D train script to run on HTCondor
    - fastai requires python 3
        - Script fix: python -m virtualenv -p python3 myvenv
        - Packages to install: fastai; (optional) pandas, corner

11 Jun

- Fixed missing file issue in Autoencoders-compression-anomaly /AE-Compression-pytorch
- Changed the README file to correct the repo link
- Reading up on Batch concepts
- Checked Eric's thesis for the number of epochs used for training.
  For 4D data: 10@10−7; 10@10−4 and 2000@10−6.
  For 27D. he performed a grid search. The plot in his thesis shows that around 350k batches were processed (for 18D latent variable), so this can be taken as a starting point.
- The best performing network from the grid search was trained using a learning rate of 10−2, wd= 0.01 and bs = 4096
- Modified and trained for 100ep using Erik's python script
- Created a new branch on github to record changes
- Modified 27D_train to load a model and retrain

10 Jun

- Training on 27D data for 350 more epochs
- Read on HTCondor
- Mailed Erik for clarification on data - waiting for a reply

9 Jun

- started on 27D data:
    - https://github.com/erwulff/lth_thesis_project/blob/master/jet_by_jet_compression/aod_compression/train_on_aod.ipynb
        - Abs rel error on old checkpoint = 0.0007828
        - MSE on old checkpoint= 8.3398e-05
        - Abs relative error on model trained with 10 epochs: 0.017212
        - Abs relative error on model trained with 150 epochs= 0.00700411

8 Jun

- Got the data
- Setting up lxplus - reading stuff
- Links

    - http://information-technology.web.cern.ch/services/lxplus-service
    - https://lxplusdoc.web.cern.ch/
    - Batch service - https://batchdocs.web.cern.ch/index.html

5 Jun

- Created the plots for L1 vs MSE

4 Jun

- Still training AE_3D_200_RELU_BN_L1_custom_norm, completed 4k epochs - **Done**

3 Jun

- Training with L1 on evaluation data - laptop, jupyter

2 Jun

- Requested for data files
    - https://github.com/Autoencoders-compression-anomaly/AE-Compression-pytorch/issues/2
- Issue raised to add some missing files
    - https://github.com/Autoencoders-compression-anomaly/AE-Compression-pytorch/issues/1
- Add L1 experiments somewhere