Warsaw University of Technology

ALICE

Machine learning approach to QA in TPC detector in ALICE experiment

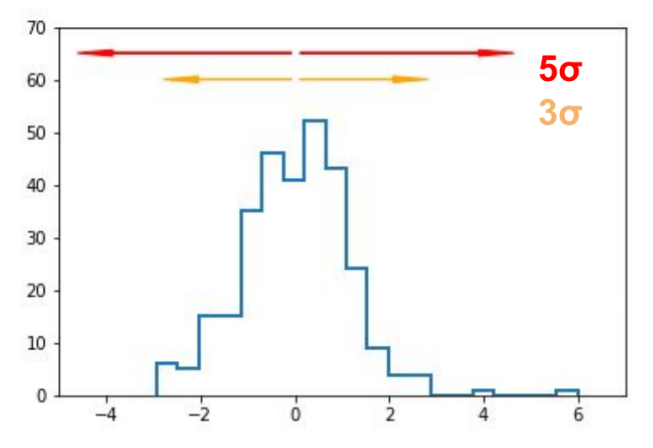
Piotr W. Nowak

in collaboration with:
Sebastian Bysiak (IFJ PAN)
Kamil Deja (WUT)
Jacek Otwinowski (IFJ PAN)
Marian Ivanow (GSI)

Current approach

For each run we compute values of monitored parameters, related to vertex position, multiplicity, MIP, DCA, etc.

For given period compute “robust” μ and σ of the obtained parameters distributions



For each parameter assign flags to each run

Combine low-level flags into high level and mark as good/bad

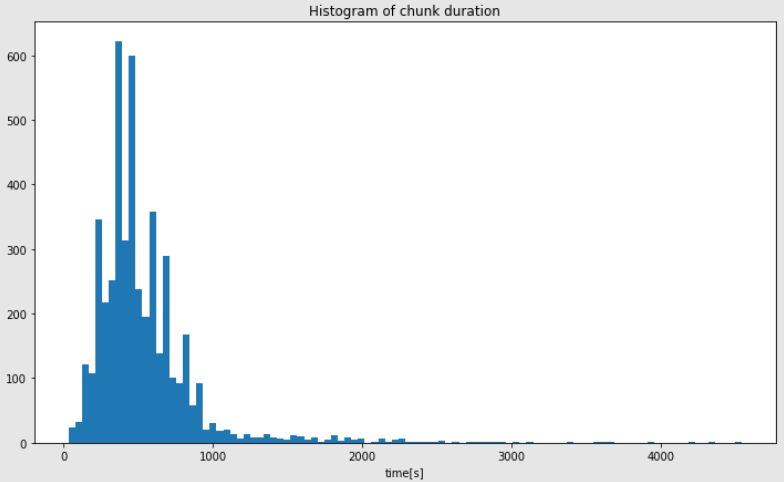
Overview

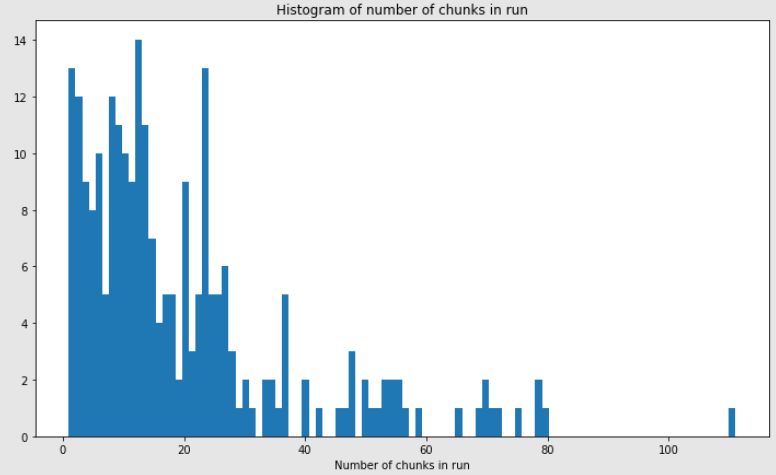
Now we compute this values for **shorter intervals of time**.

We call them **chunks**. They have different lengths (mostly around 8 min.)

And runs have different amount of chunks (mostly less than 30).

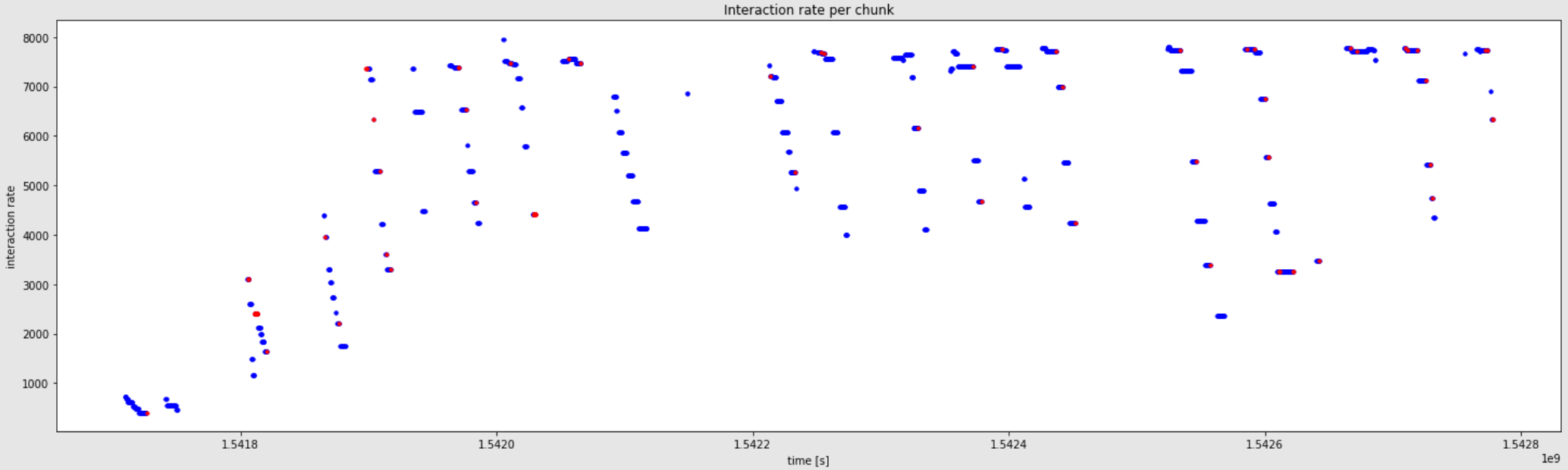
We compute data from 5 periods,
2 Pb-Pb colissions (LHC18q,LHC18r) and
3 p-p colissions (LHC18f,LHC18o,LHC18p)





Overview

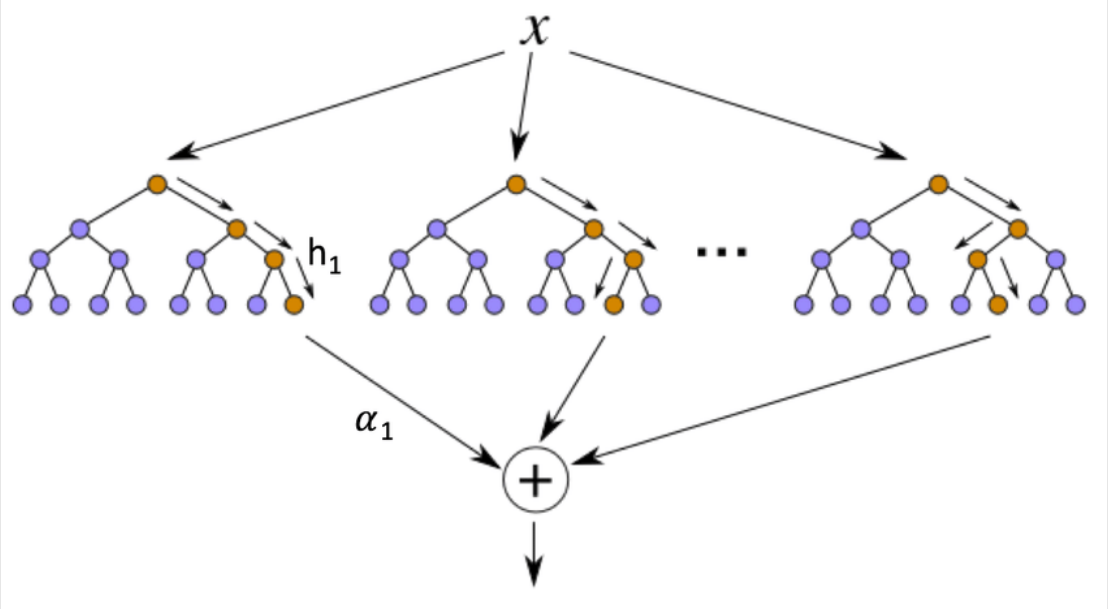
For every chunk calculate statistical values of monitored parameters and get warning flags.



Machine Learning techniques

Supervised learning – Boosted Decisions Trees

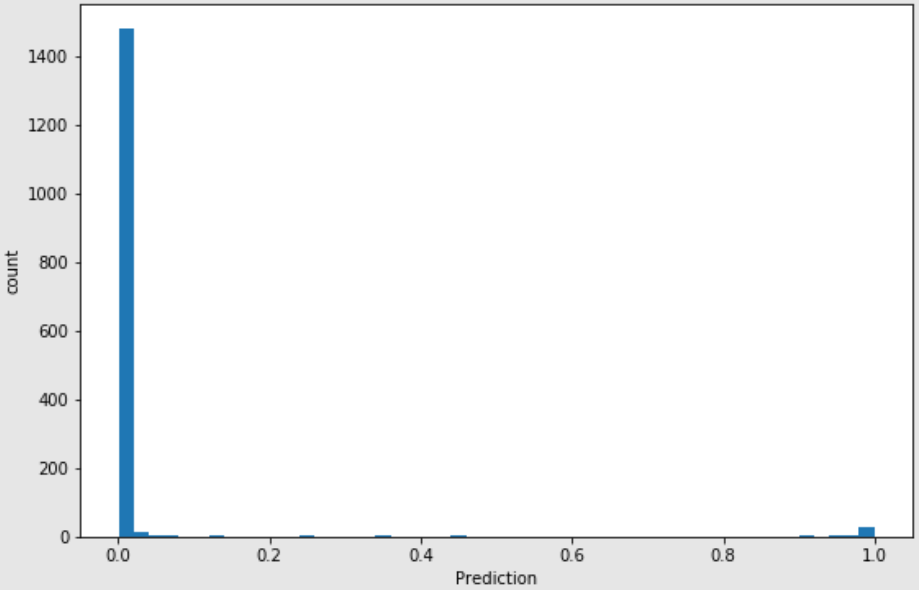
Supervised learning means that we have information about output.
One of the most popular supervised algorithms are decisions trees.

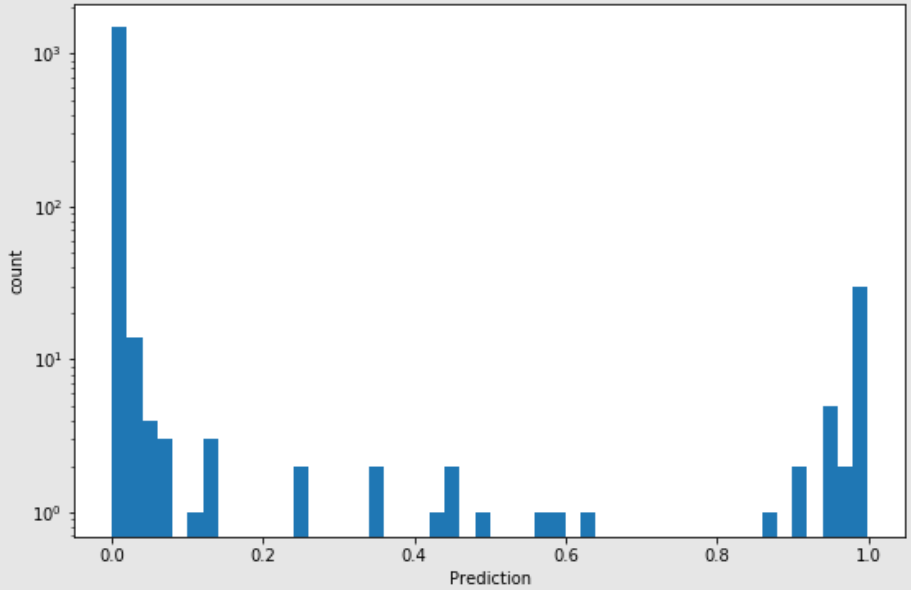


Machine Learning techniques

Supervised learning – Boosted Decisions Trees

Results (LHC18r,LHC18q)

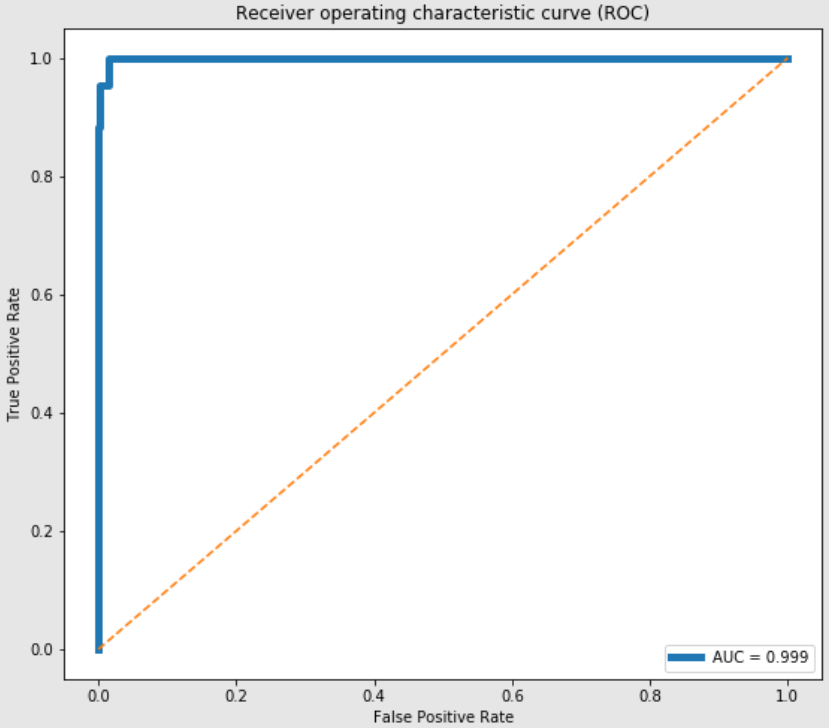


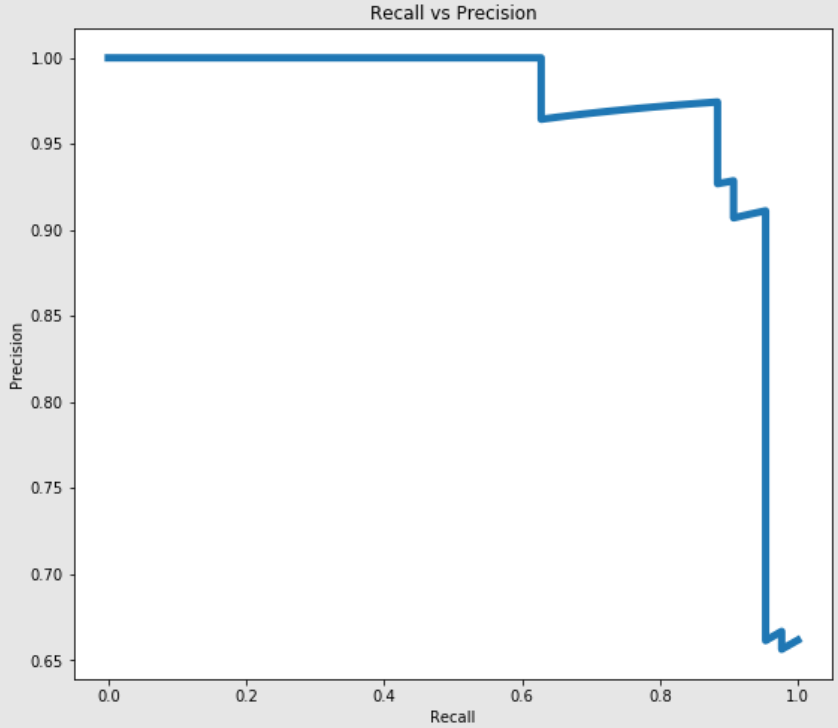


Machine Learning techniques

Supervised learning – Boosted Decisions Trees

Results (LHC18r,LHC18q)



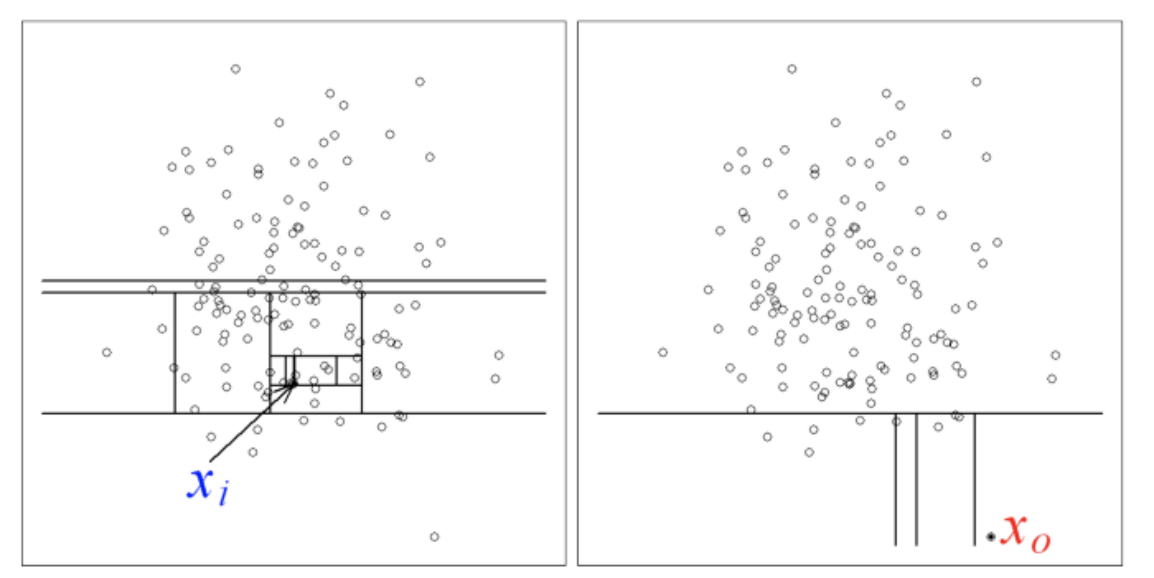


Machine Learning techniques

Unsupervised learning – Isolation forest

Unsupervised learning means that we don’t have information about output. Our method tries to find some relations, correlation or grouping data.

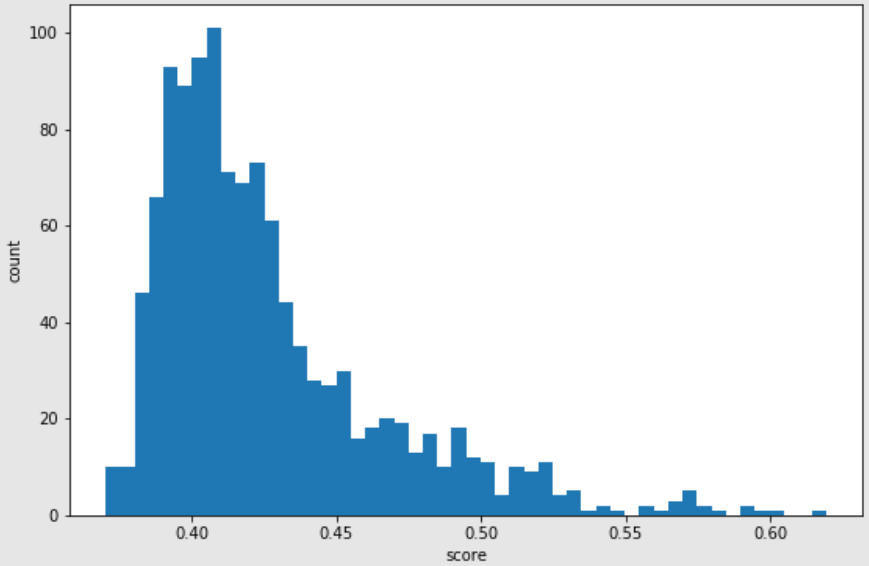
First unsupervised algorithm that we use was isolation forest.

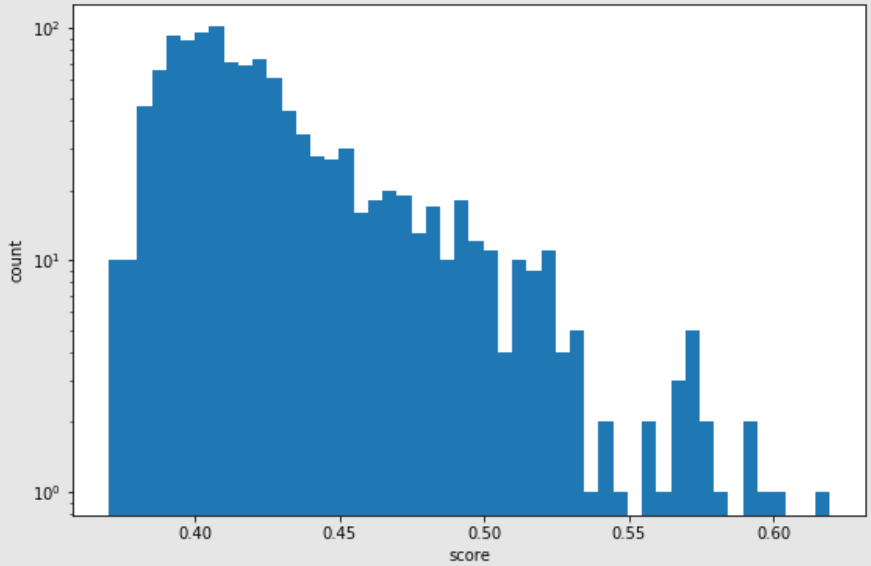


Machine Learning techniques

Unsupervised learning – Isolation forest

Results (LHC18r,LHC18q)

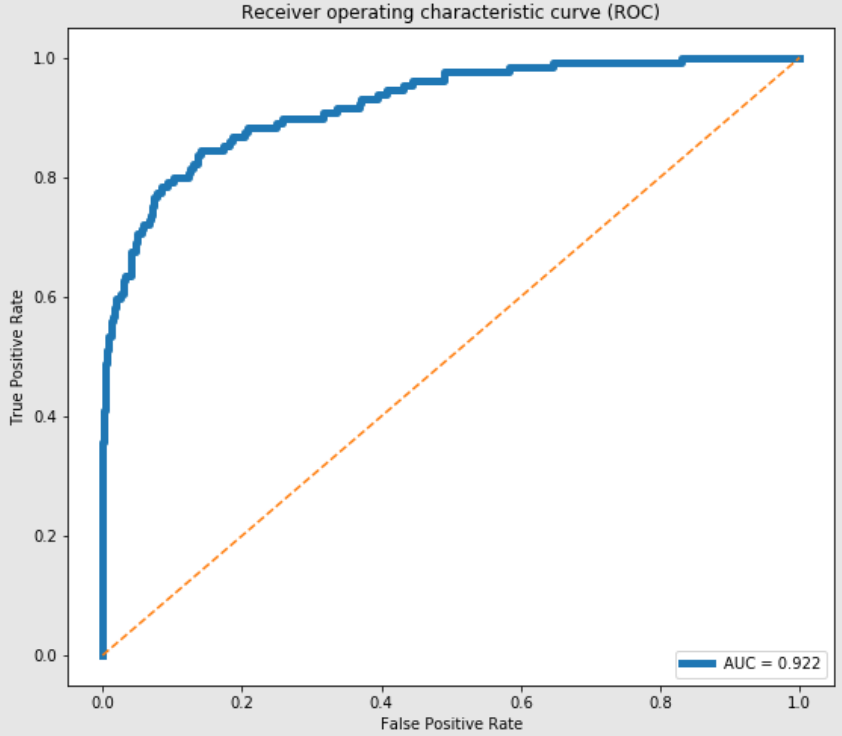


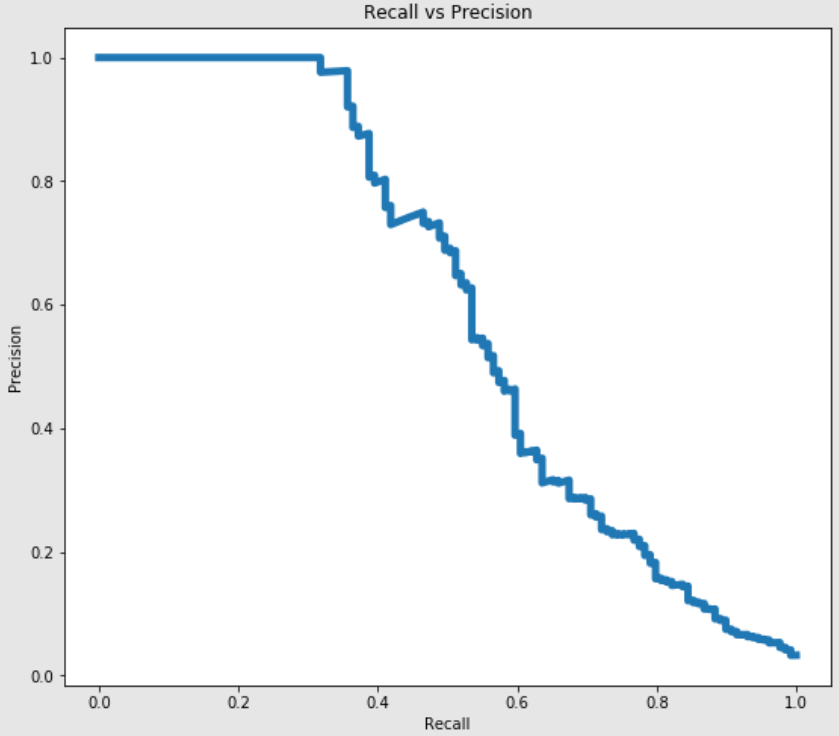


Machine Learning techniques

Unsupervised learning – Isolation forest

Results (LHC18r,LHC18q)

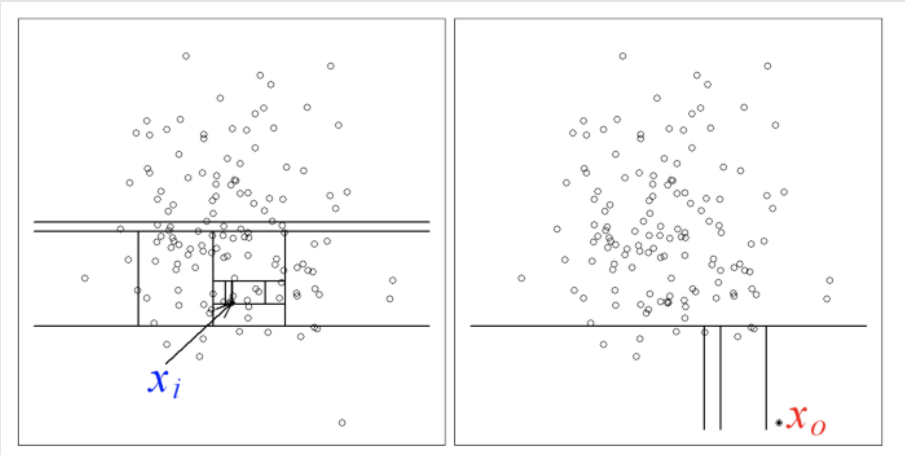




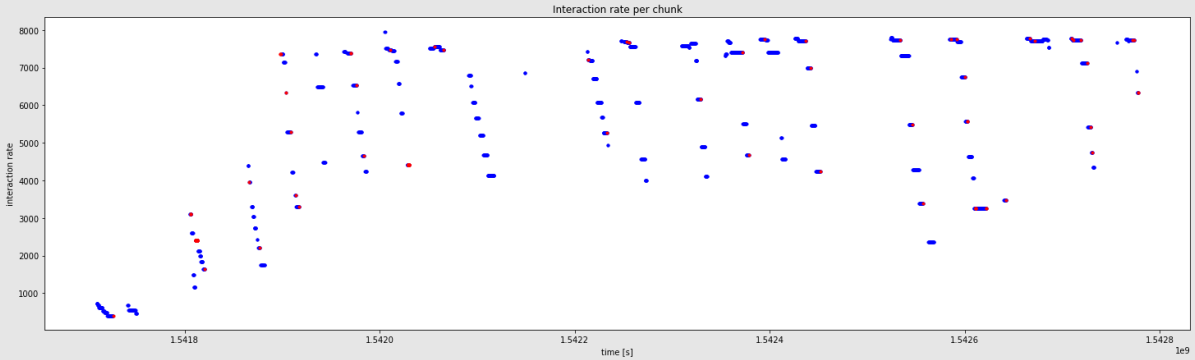
Machine Learning techniques

Unsupervised learning – Isolation forest

Problem is that isolation forest only score data by single parameter. It doesn't find corellations between parameters.

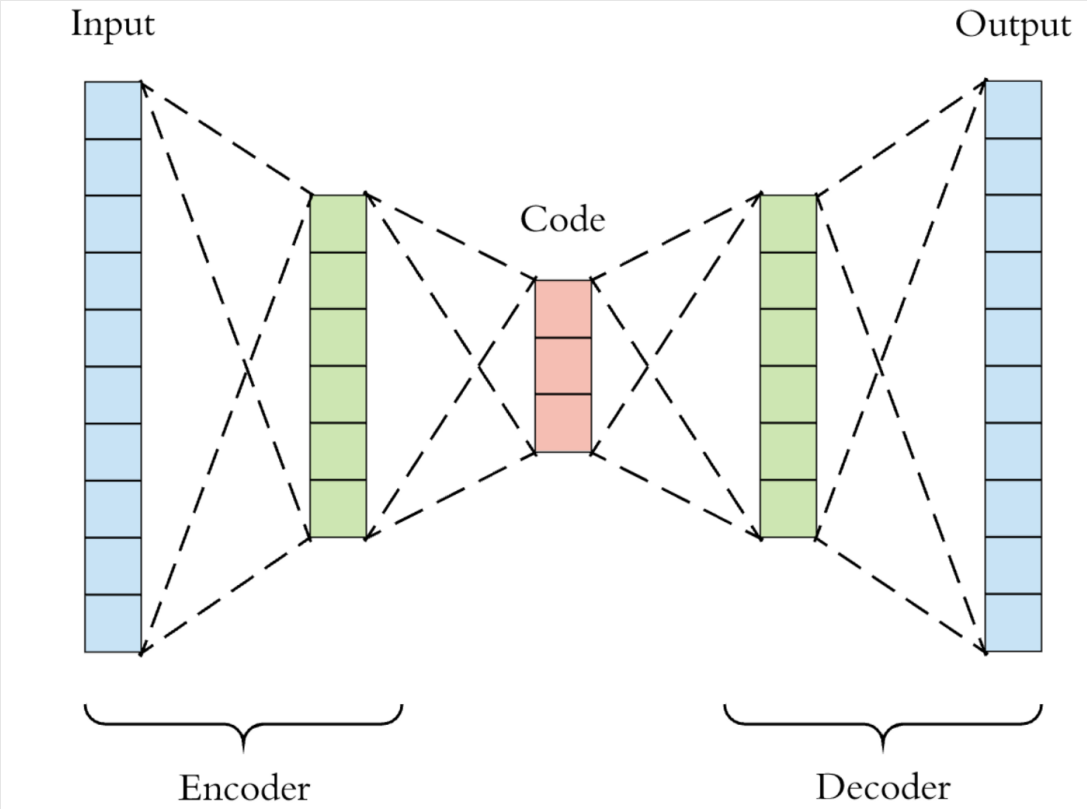


Lot of our parameters are corellated. For example with small luminosity lot of parameters change their values.

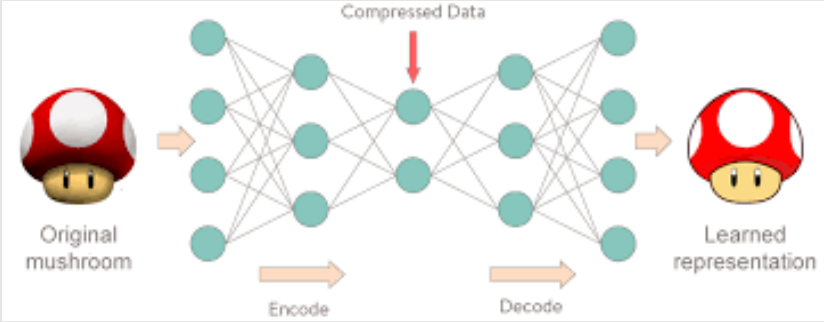


Machine Learning techniques

Unsupervised learning – Autoencoder

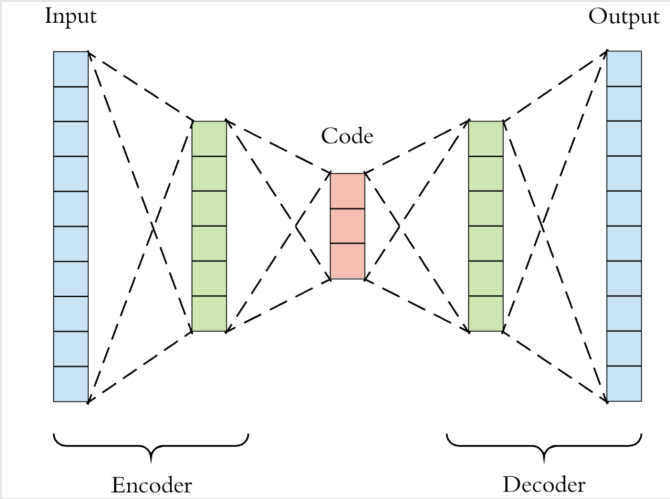


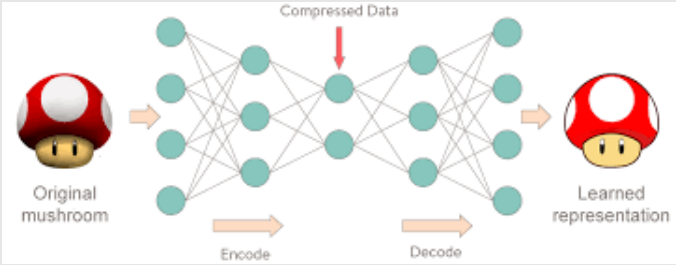
Autoencoder is technique that tries to represent (encode) input into smaller vector and then decoded it similar output.



Machine Learning techniques

Unsupervised learning – Autoencoder





Autoencoder can be used to reduce dimensionality of data or to anomaly detection.

Anomaly can be found by computing MSE of output and input. Anomaly should have trouble to decode unexpected values and that is represented by bigger MSE.

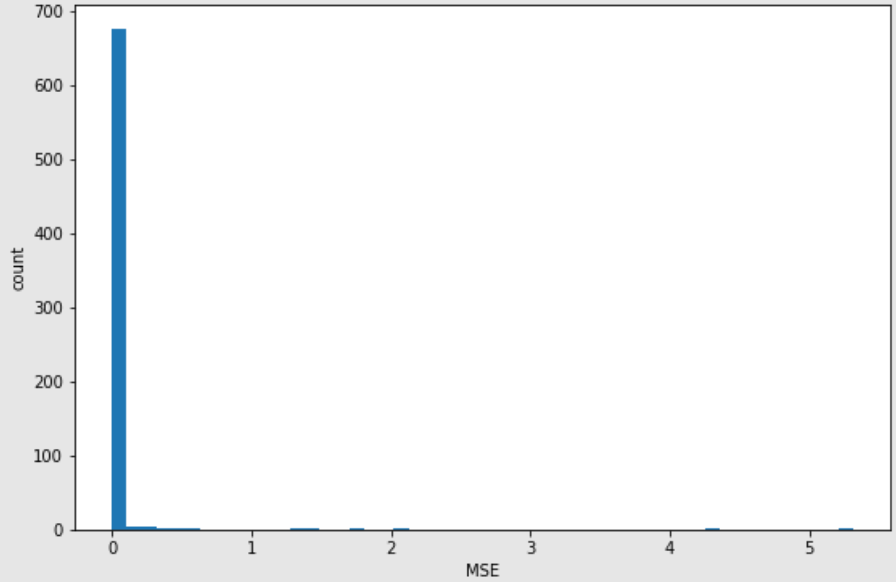
Our model have architecture

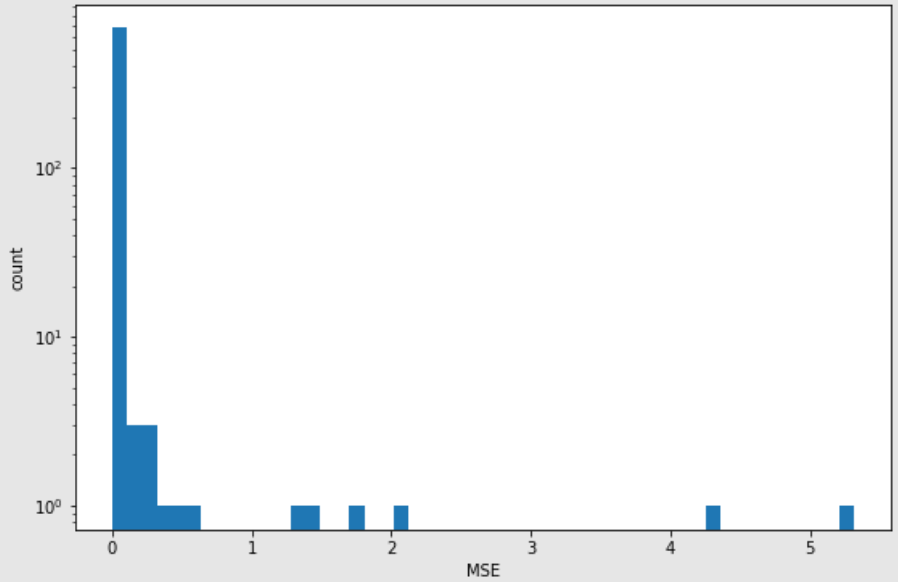
97 → 512 → 128 → 64 → 128 → 512 → 97

Machine Learning techniques

Unsupervised learning – Autoencoder

Results (LHC18r,LHC18q)

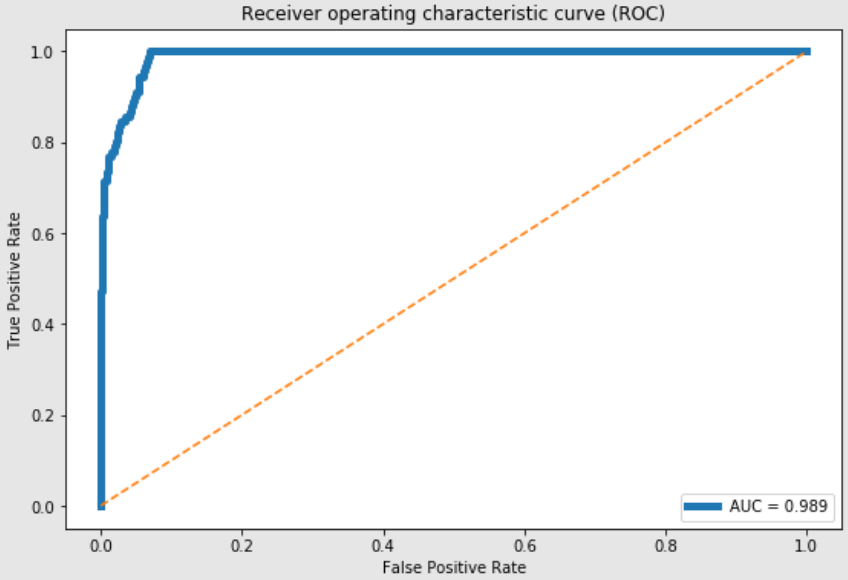


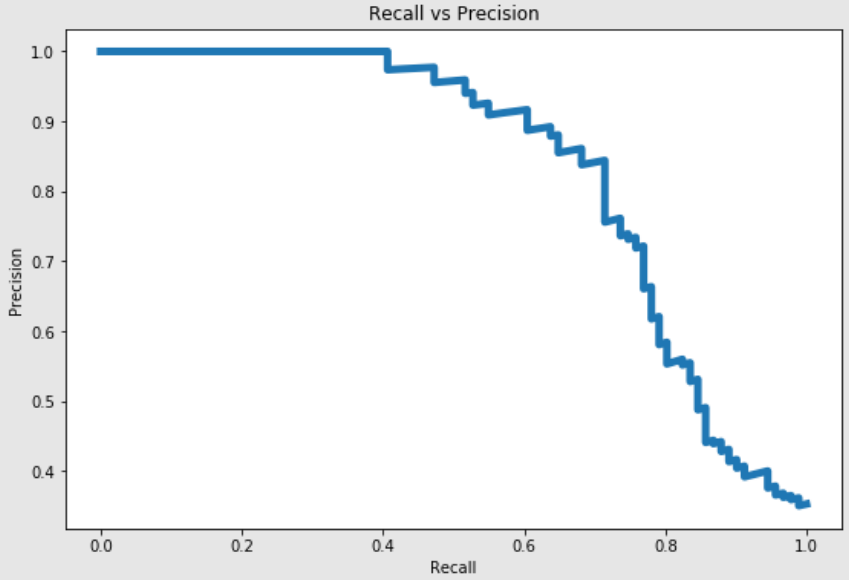


Machine Learning techniques

Unsupervised learning – Autoencoder

Results (LHC18r,LHC18q)

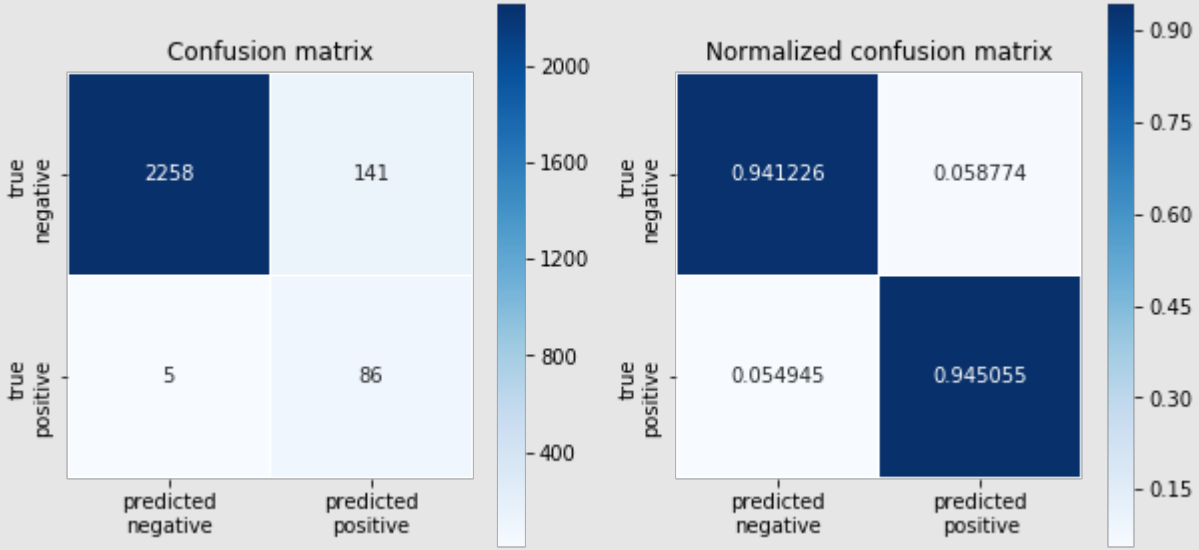




Machine Learning techniques

Unsupervised learning – Autoencoder

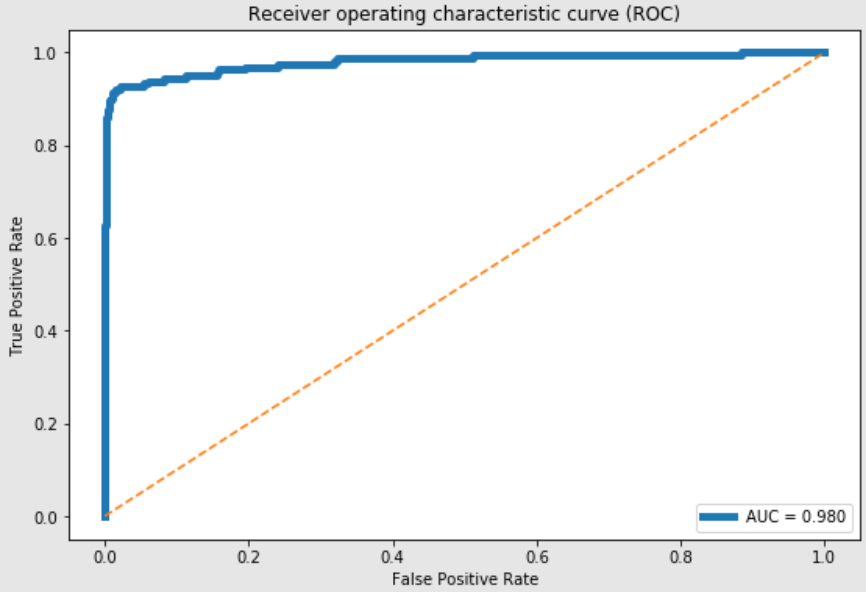
Results (LHC18r,LHC18q)

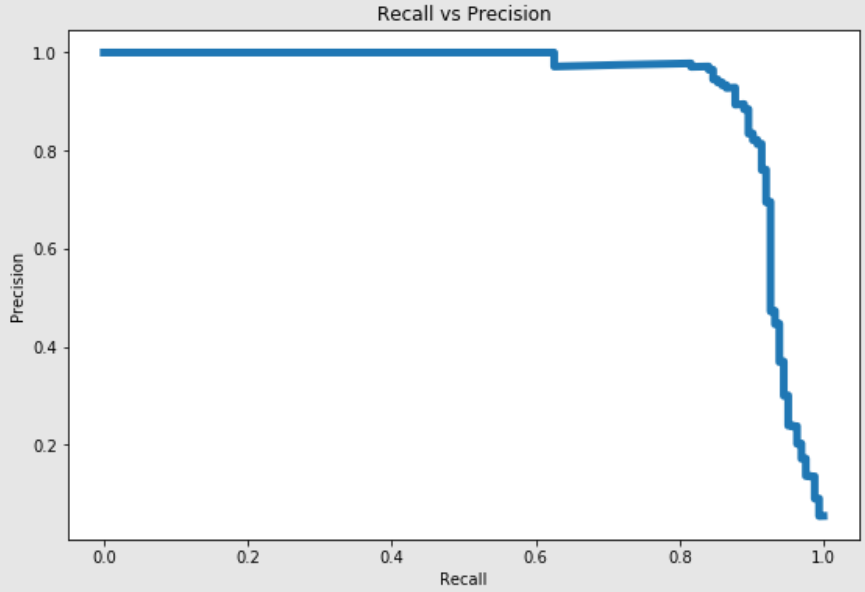


Machine Learning techniques

Unsupervised learning – Autoencoder

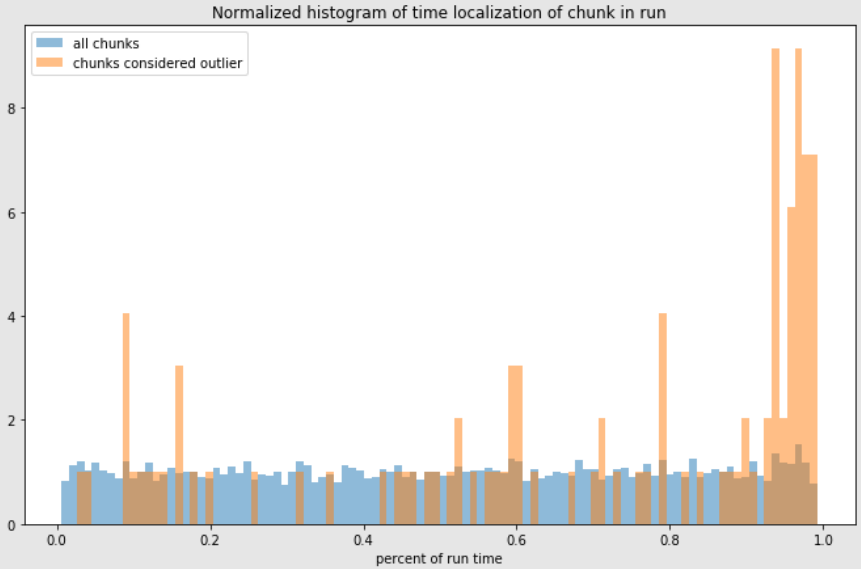
Results (LHC18f,LHC18o,LHC18p)

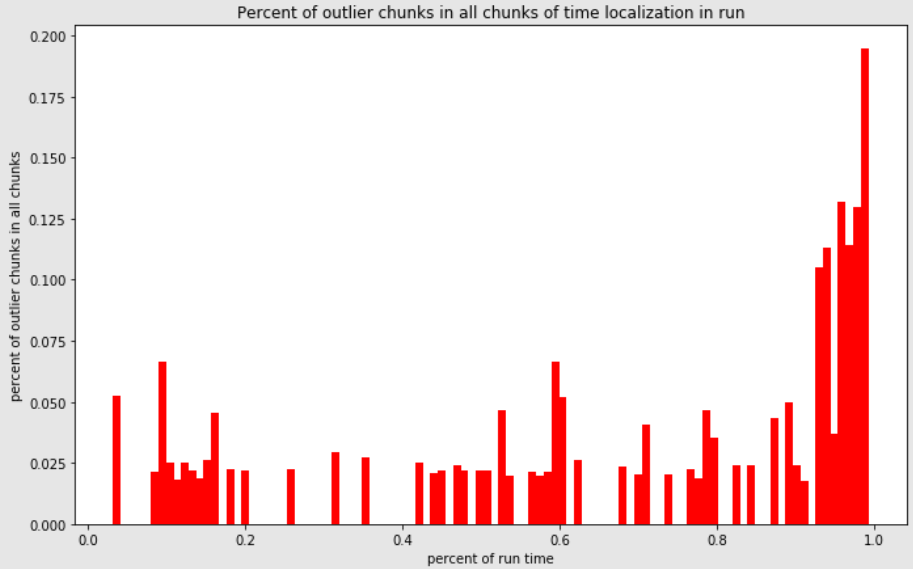




Validation

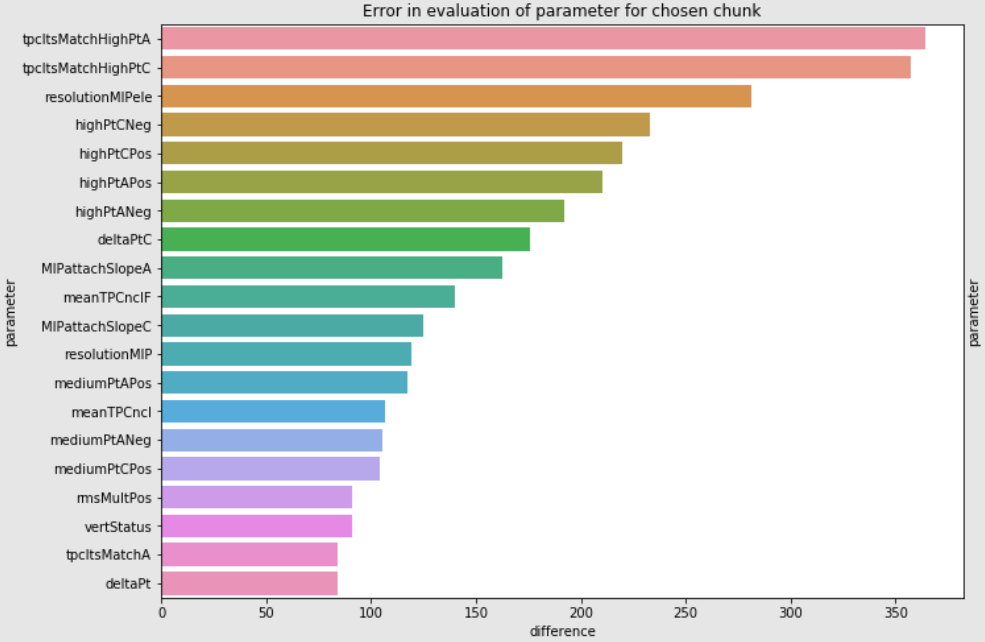
To validate our results we can look at time profile of our data. (LHC18q,LHC18r)

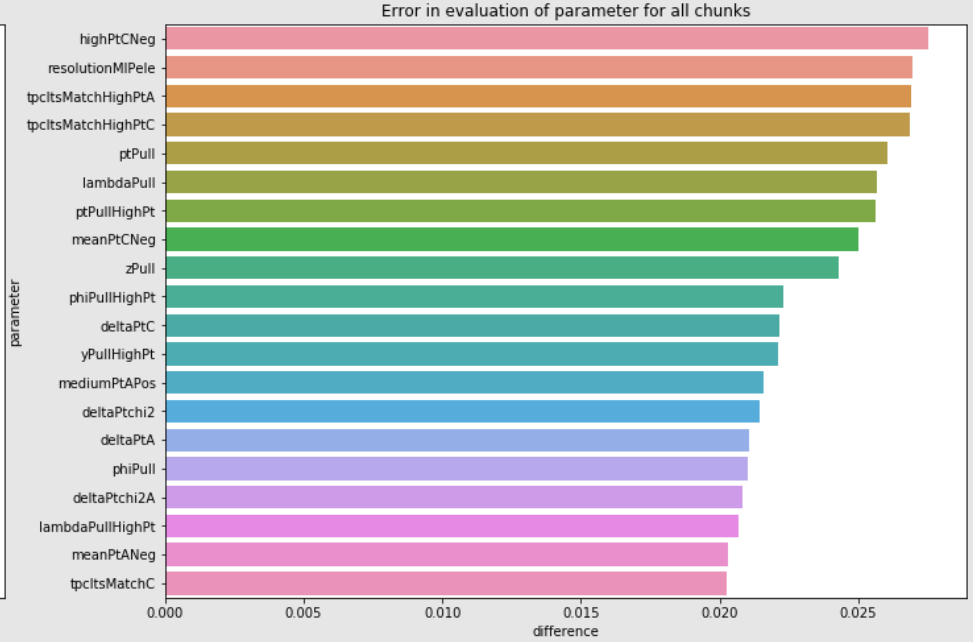




Validation

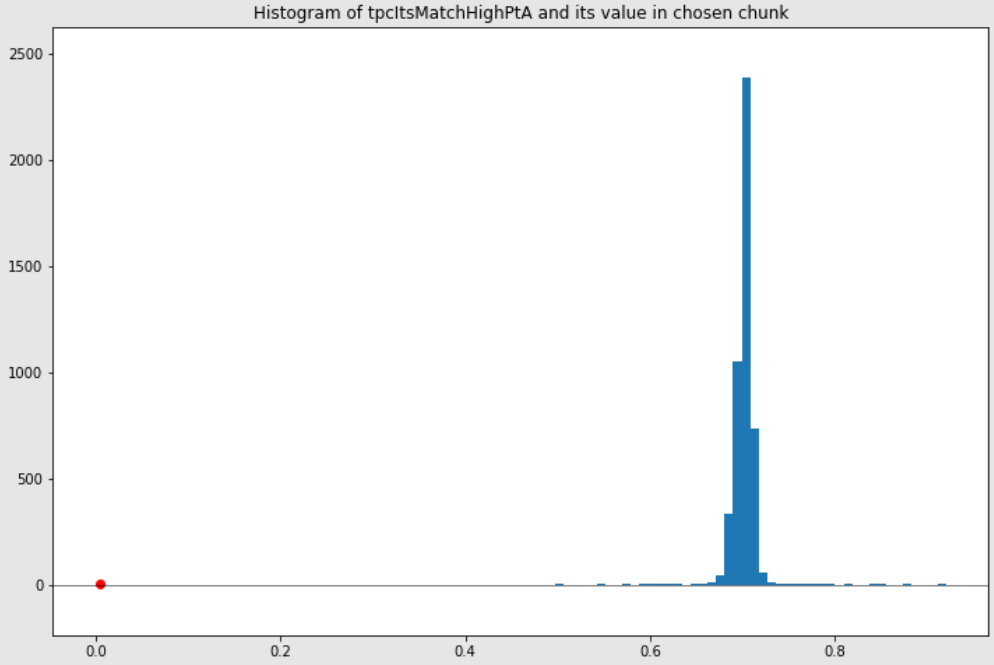
Our model predict outlier based on MSE of reconstructed parameters. (LHC18q,LHC18r)

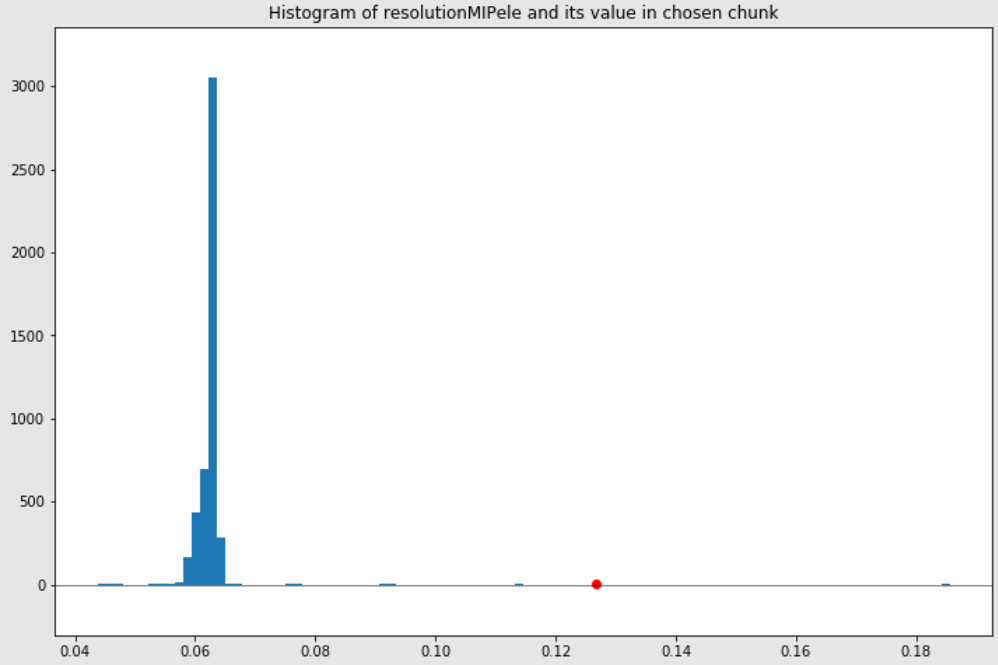




Validation

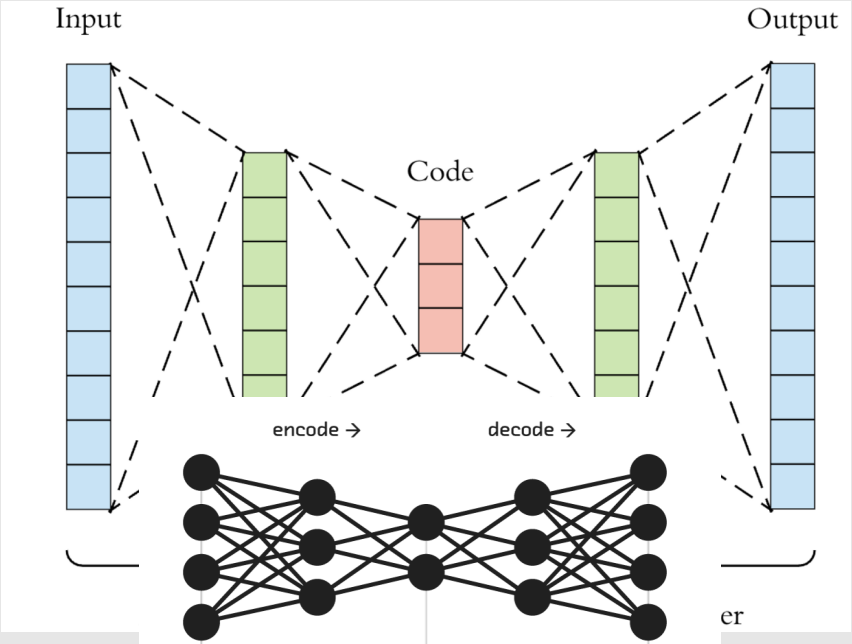
By looking into the parameters that our outlier example have biggest MSE of reconstruciton we can see that it's values are in the edge of distribution.





Analysis

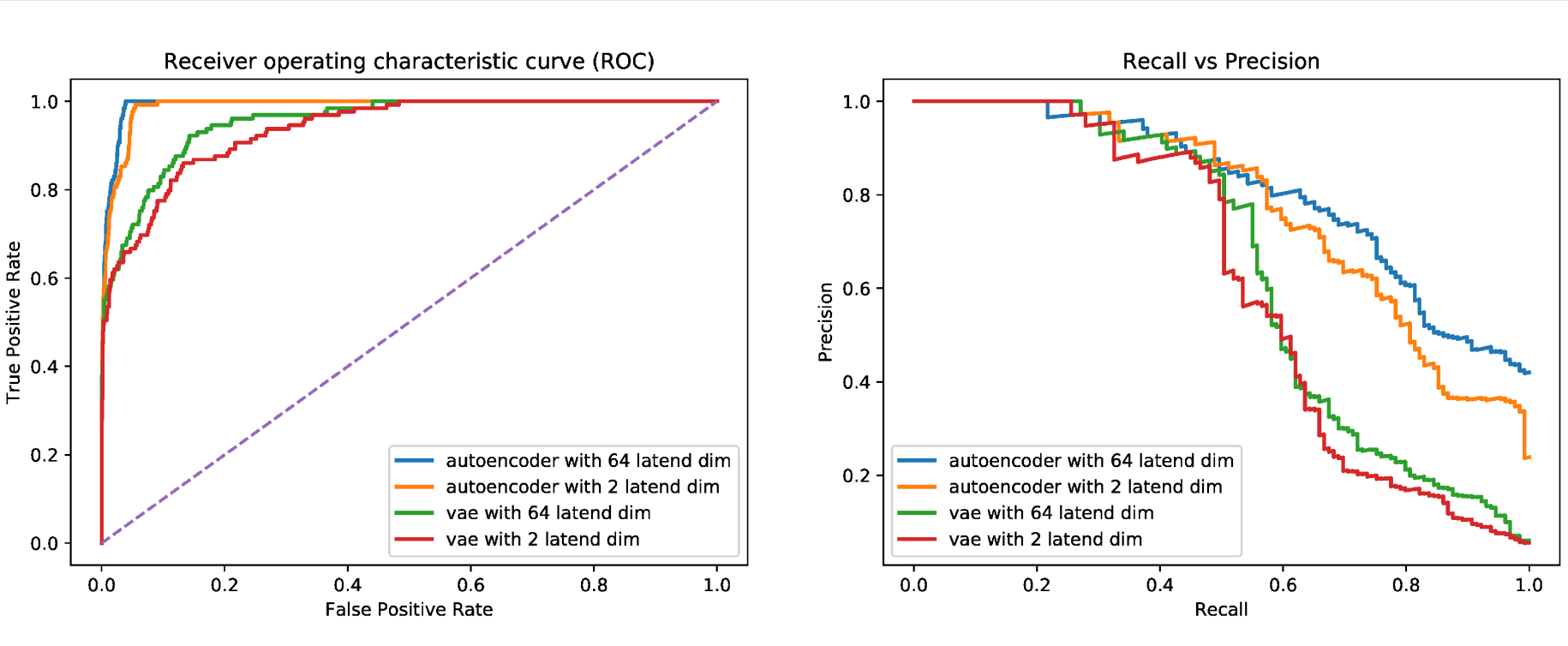
What if we use very small latent dimention?



Variational Autoencoder is generative model similar to autoencoders. The difference is that we try to represent our data with to vectors (mean and sigma) and then having gausian distribution described with this 2 parameter we sample (randomly select from distribution) value that is give to decoder.

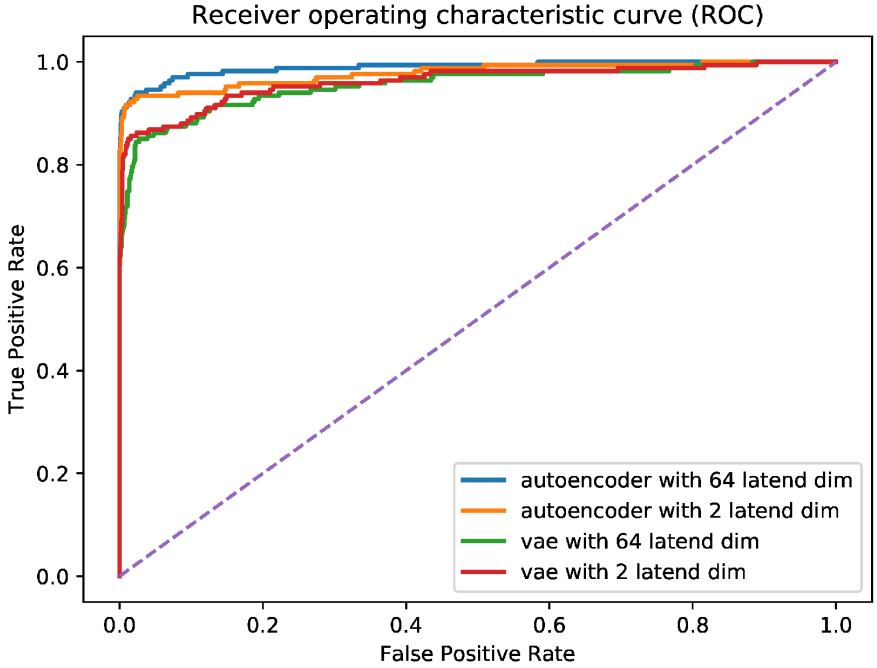
Analysis

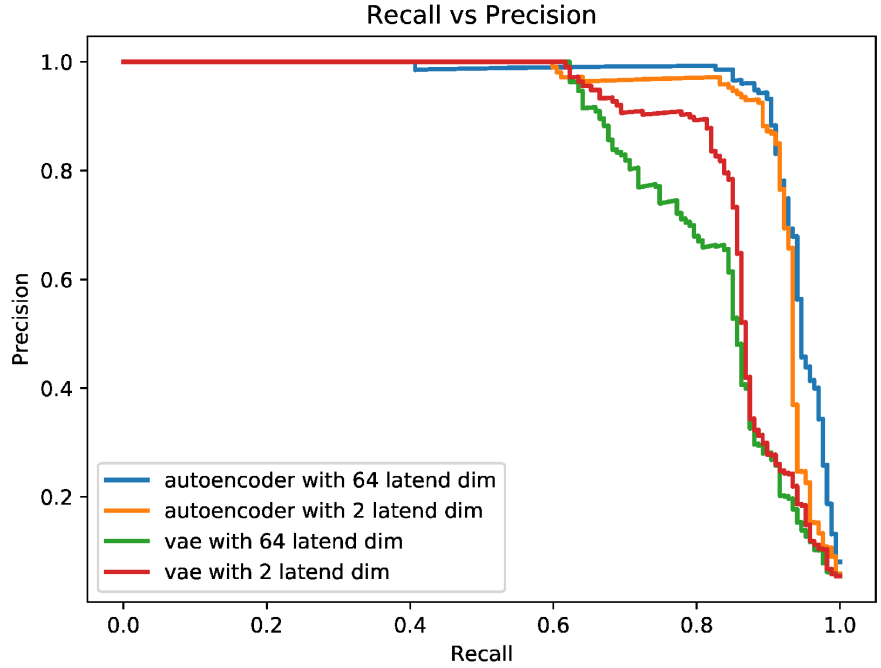
Autoencoder and variational autoencoder with different latent dimention size (LHC18q,LHC18r)



Analysis

Autoencoder and variational autoencoder with different latent dimention size (LHC18f,LHC18o,LHC18p)

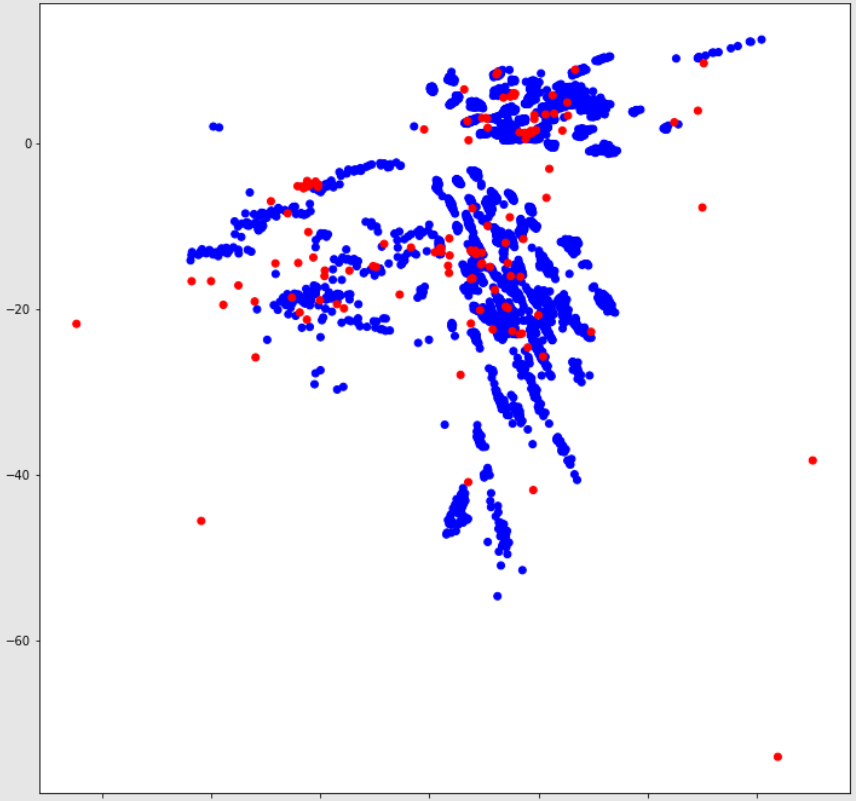




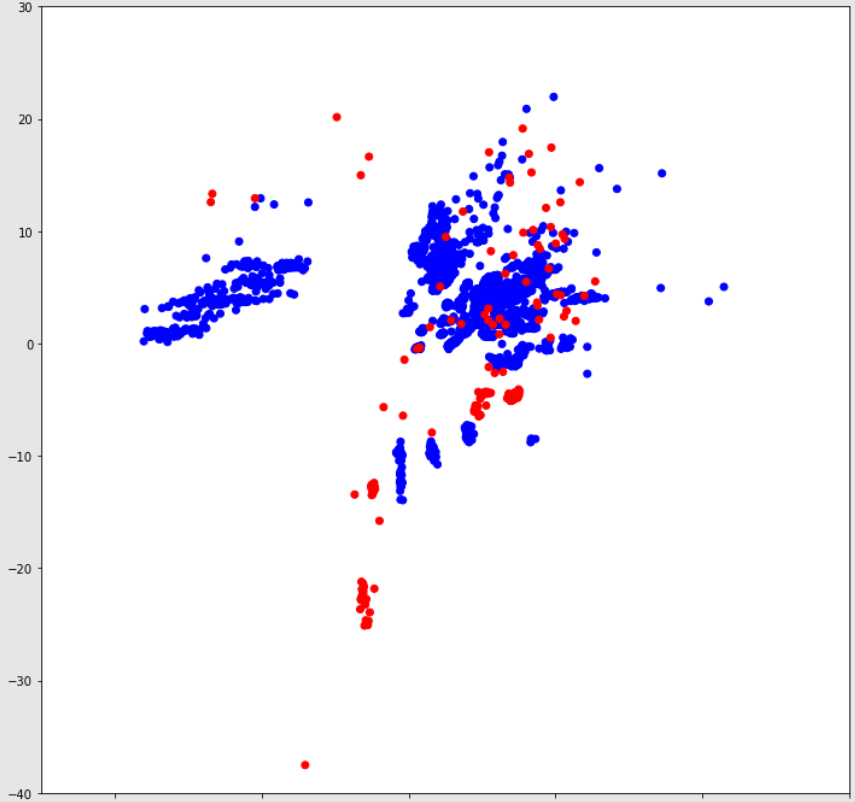
Analysis

Autoencoder

(LHC18q,LHC18r)



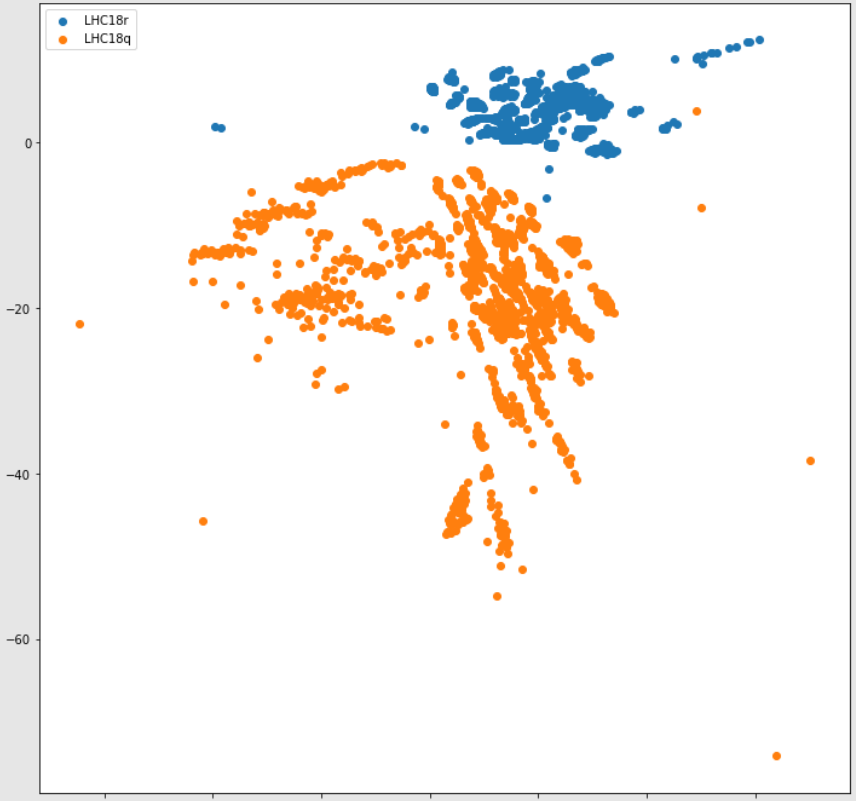
(LHC18f,LHC18o,LHC18p)



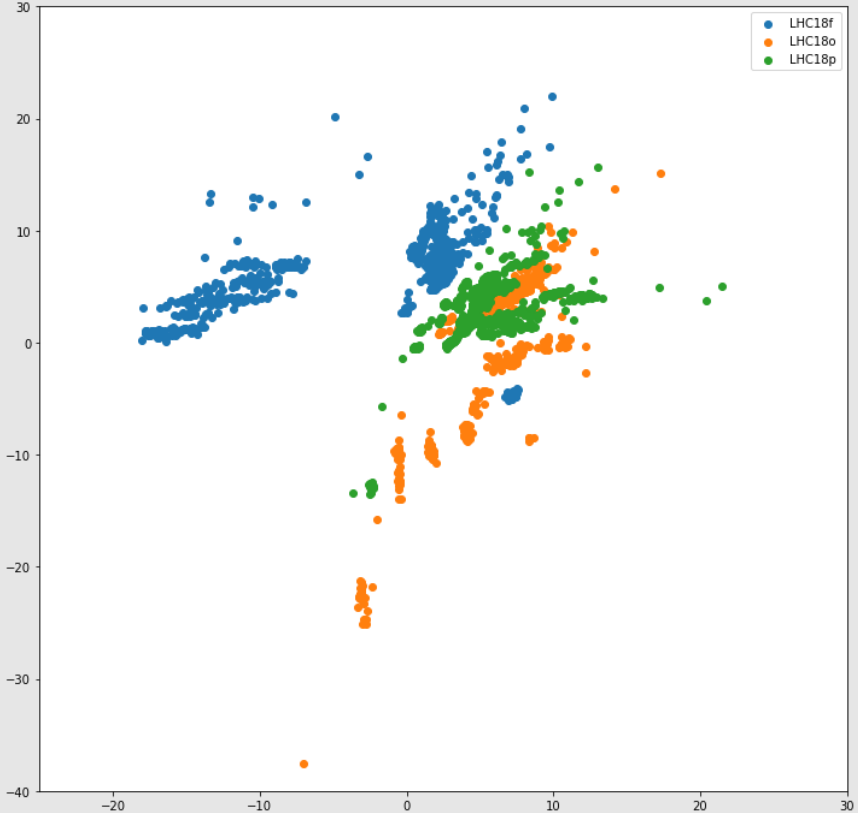
Analysis

Autoencoder

(LHC18q,LHC18r)



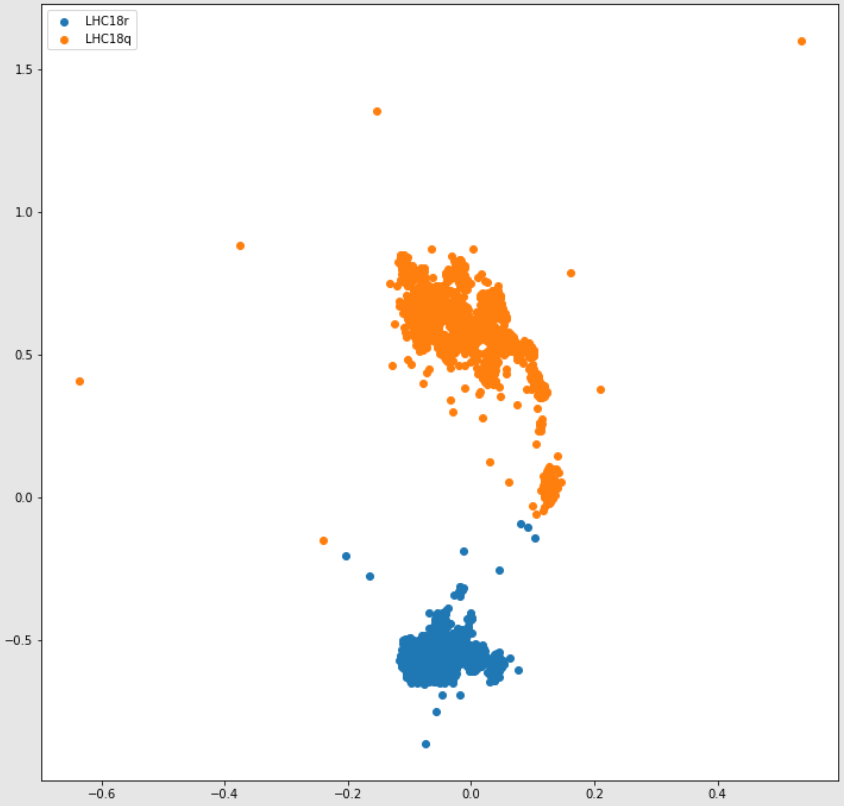
(LHC18f,LHC18o,LHC18p)



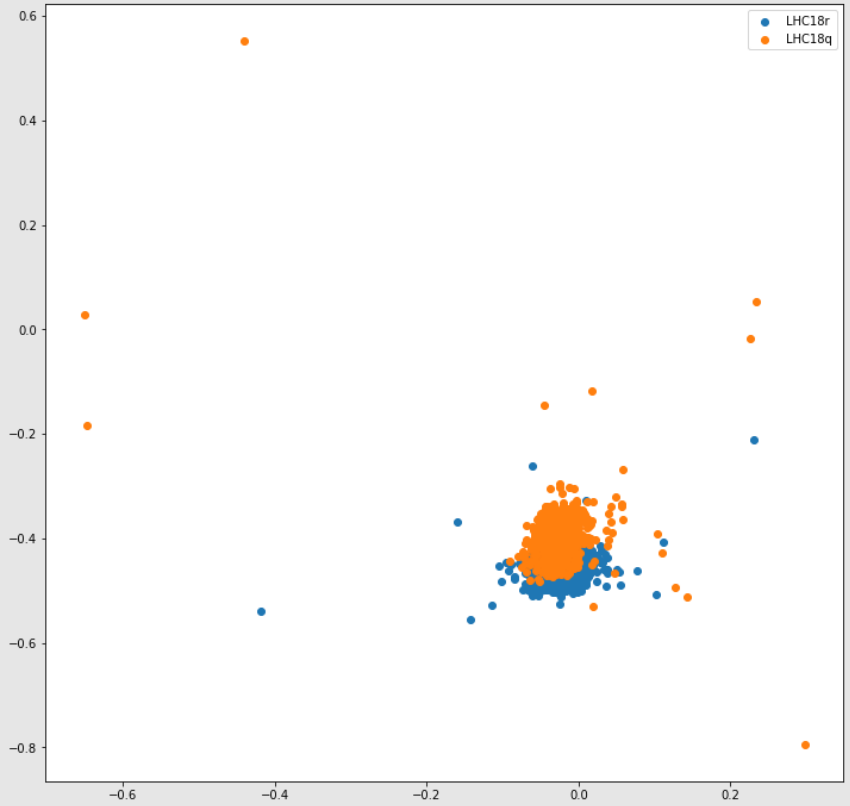
Analysis

Variational Autoencoder (LHC18q,LHC18r)

Mean

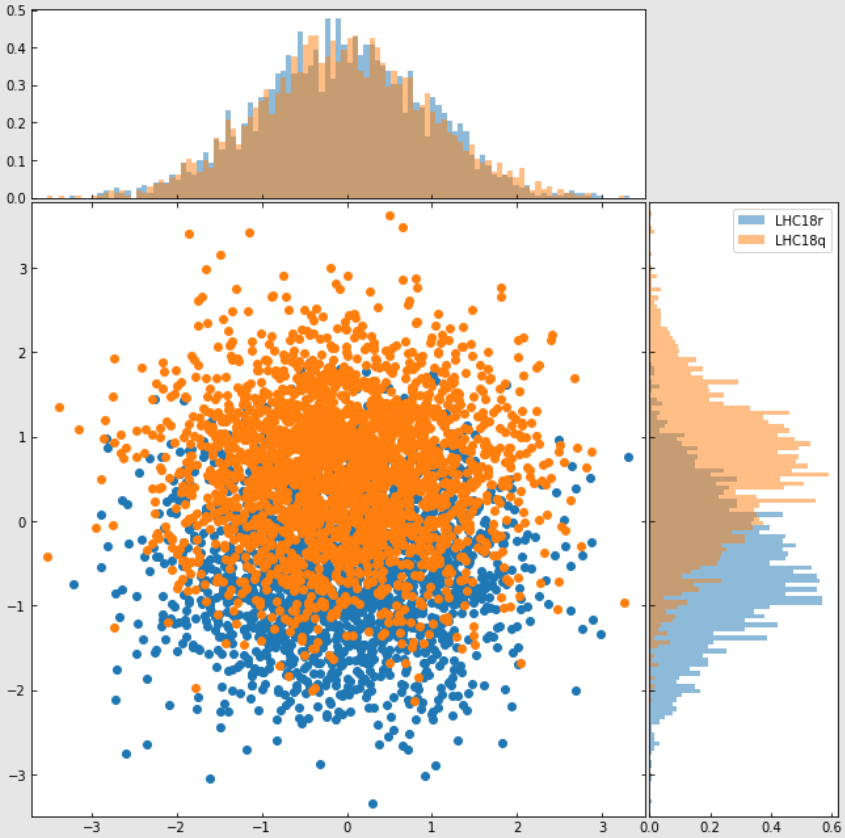


Sigma



Analysis

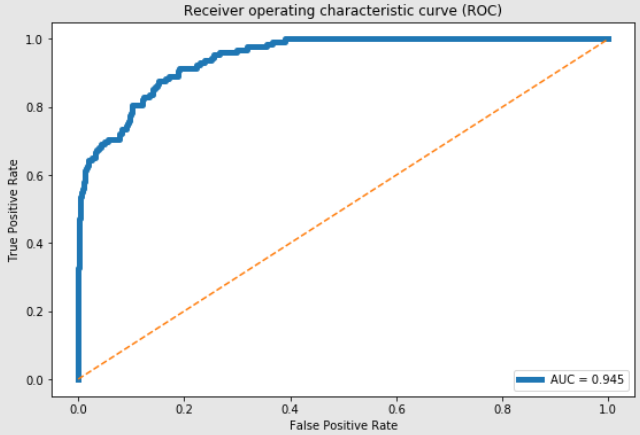
Variational Autoencoder (LHC18q,LHC18r)

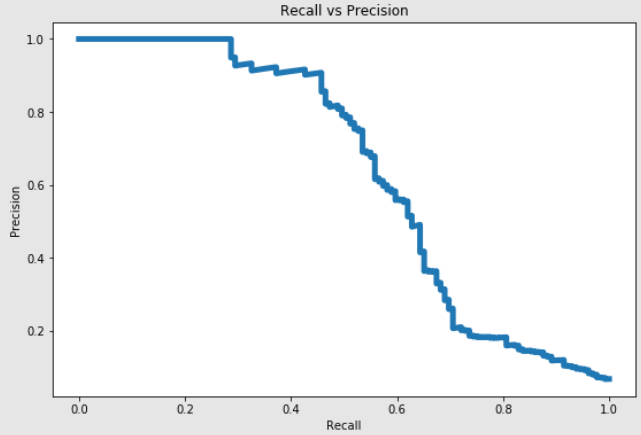


Analysis

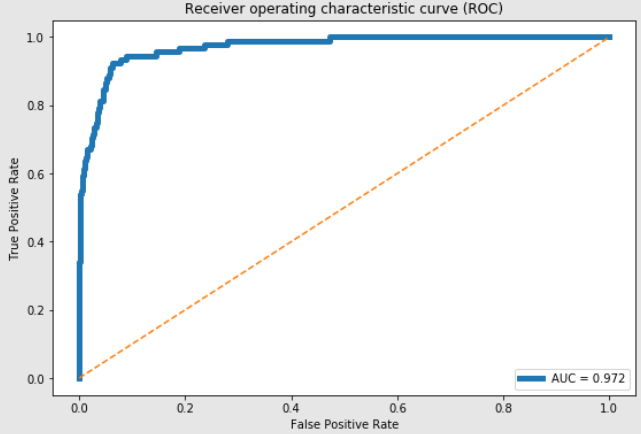
Variational autoencoder

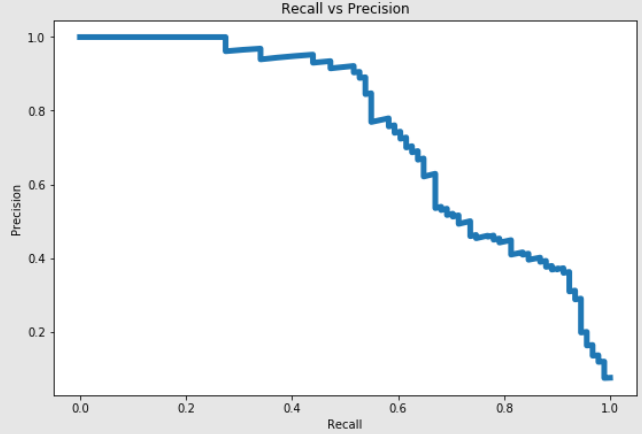
Latent dim = 2
(LHC18q,LHC18r)





Latent dim = 2
LHC18q

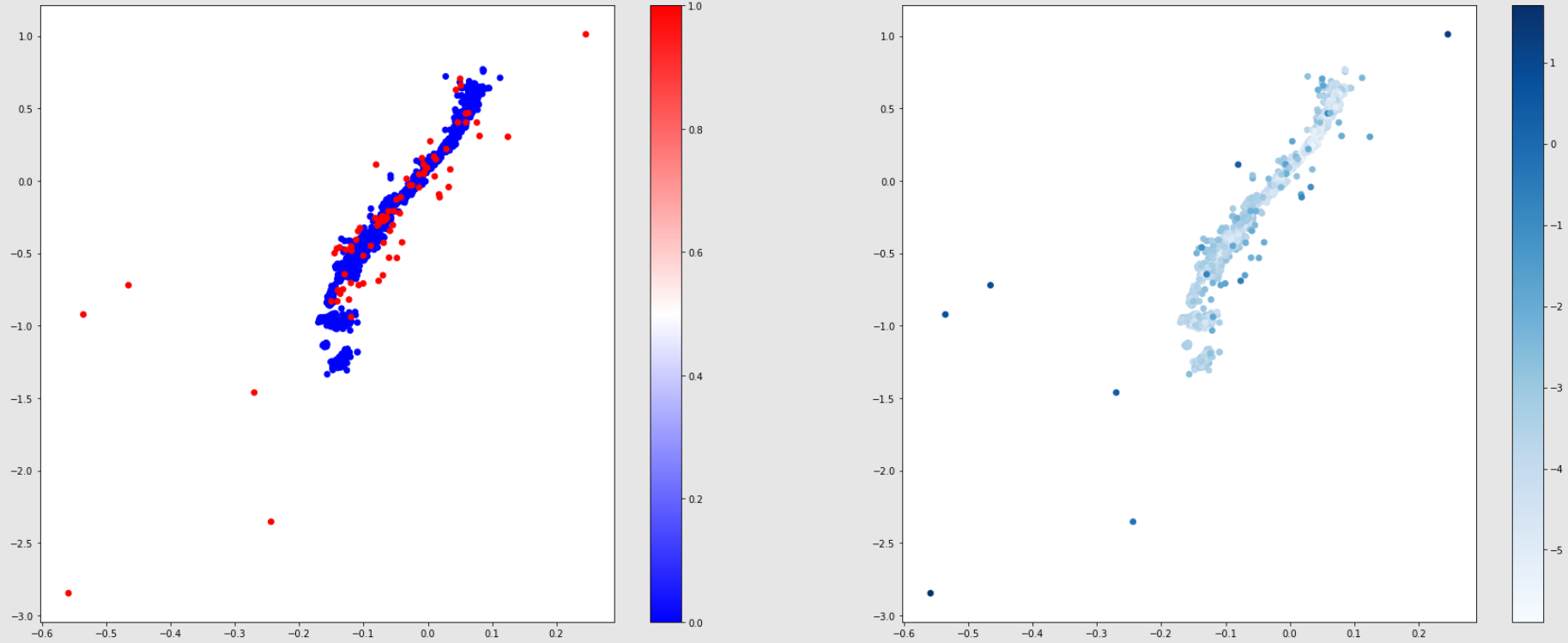




Analysis

Variational Autoencoder (LHC18q)

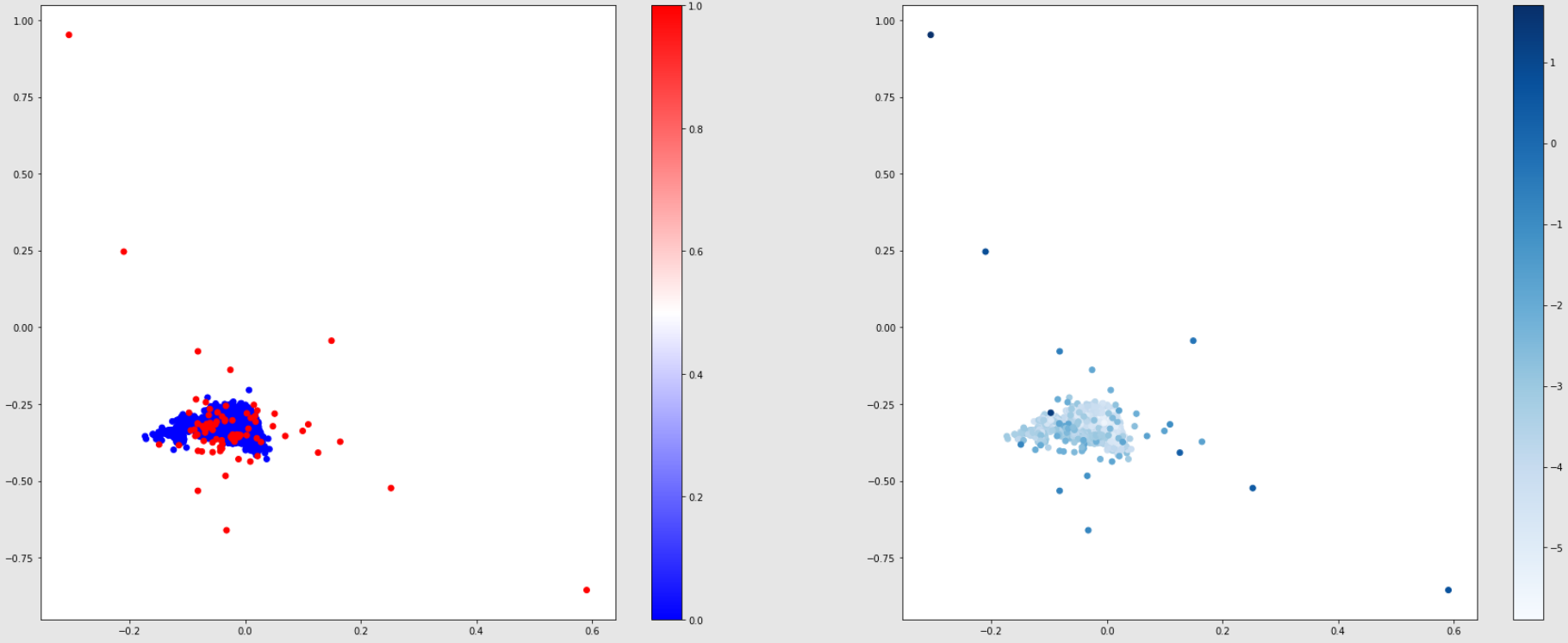
mean



Analysis

Variational Autoencoder (LHC18q)

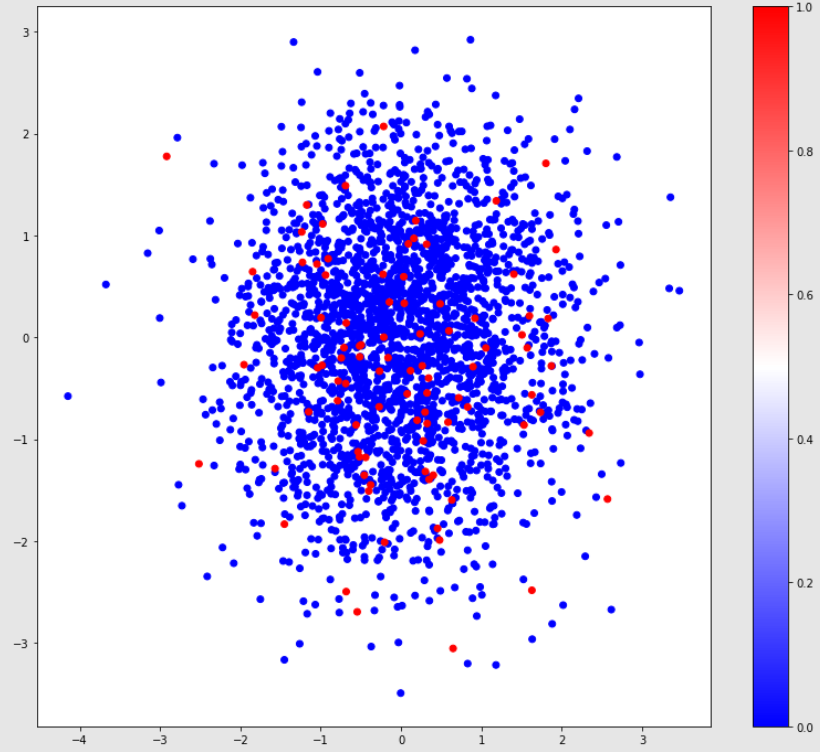
sigma

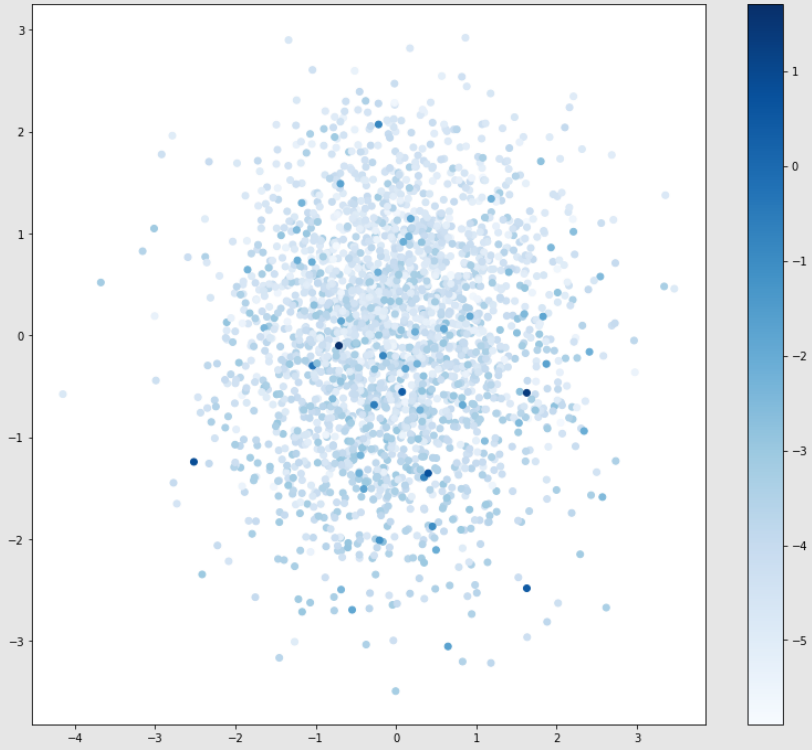


Analysis

Variational Autoencoder (LHC18q)

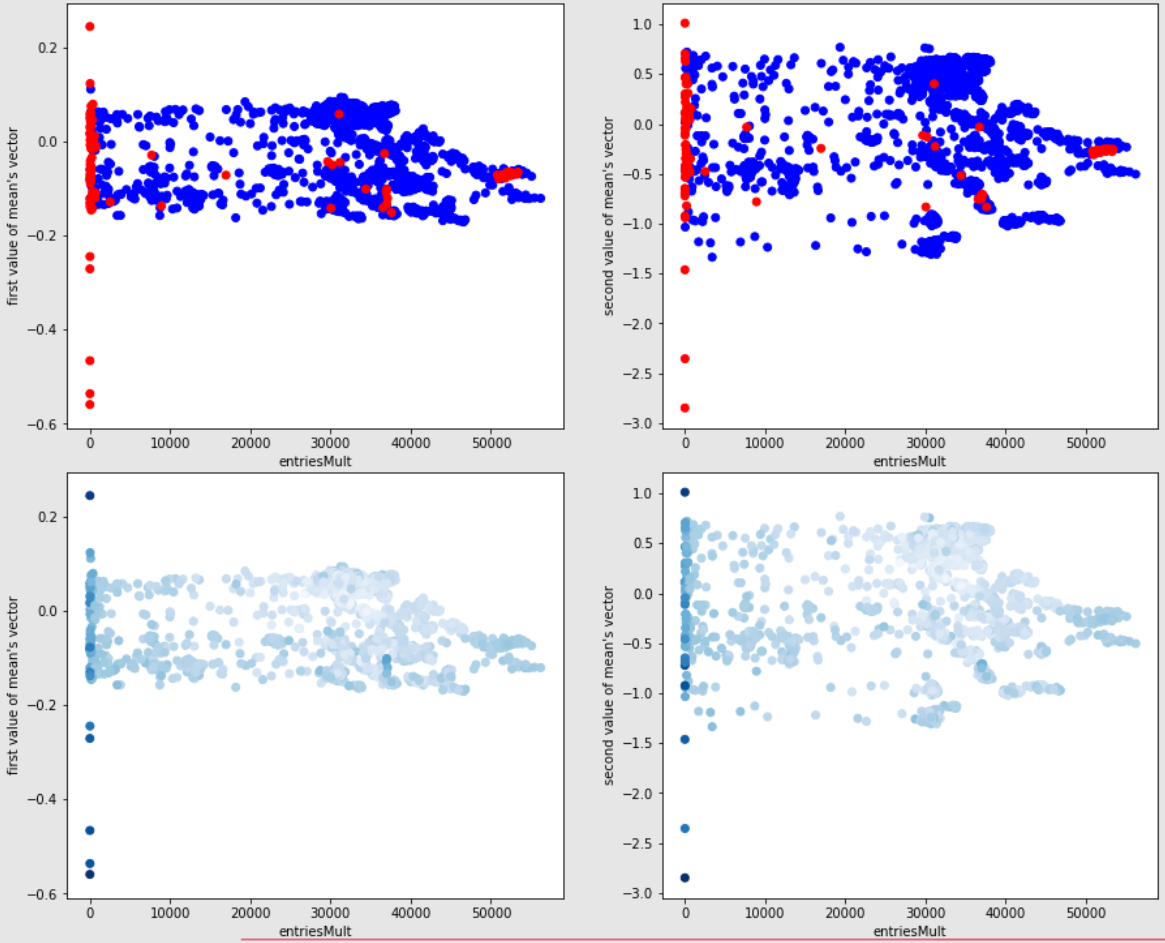
sampling





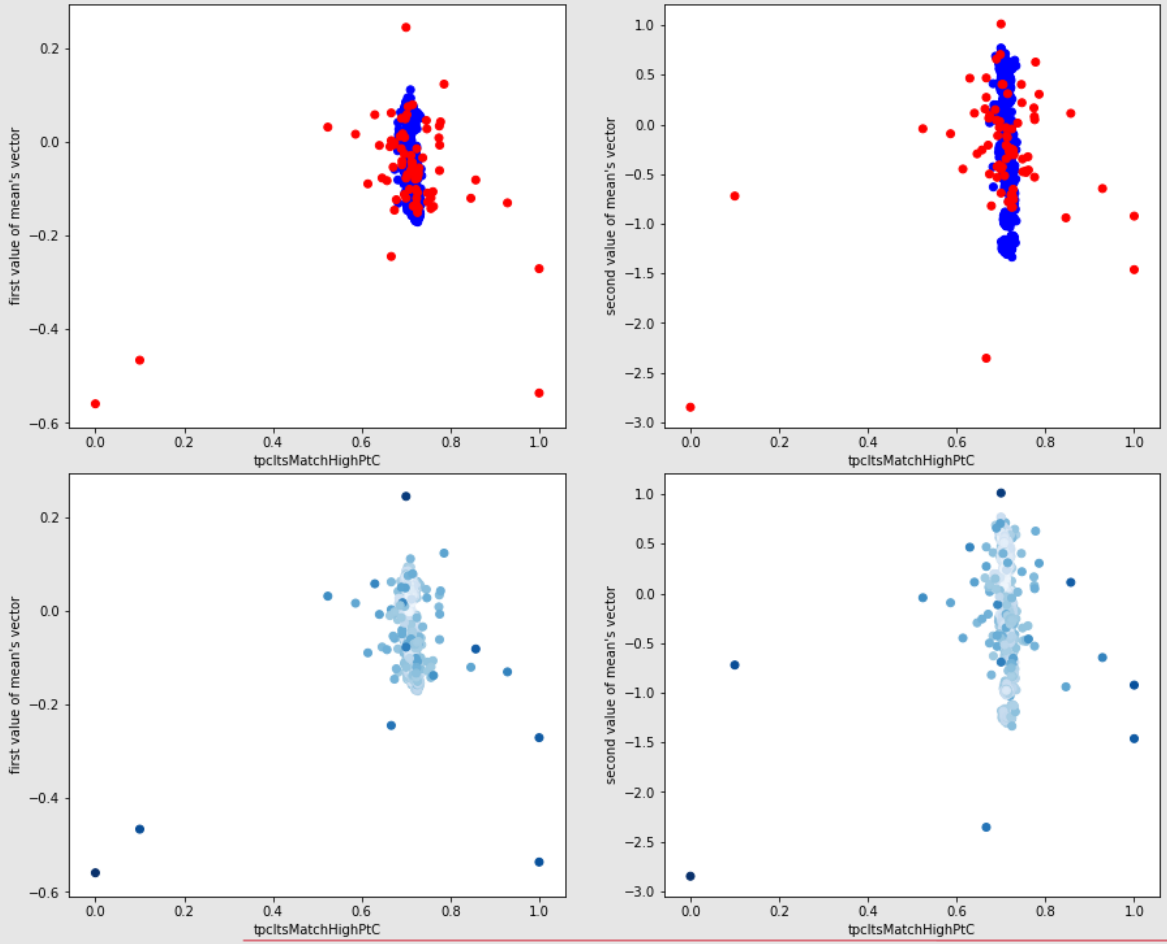
Analysis

Variational Autoencoder (LHC18q)



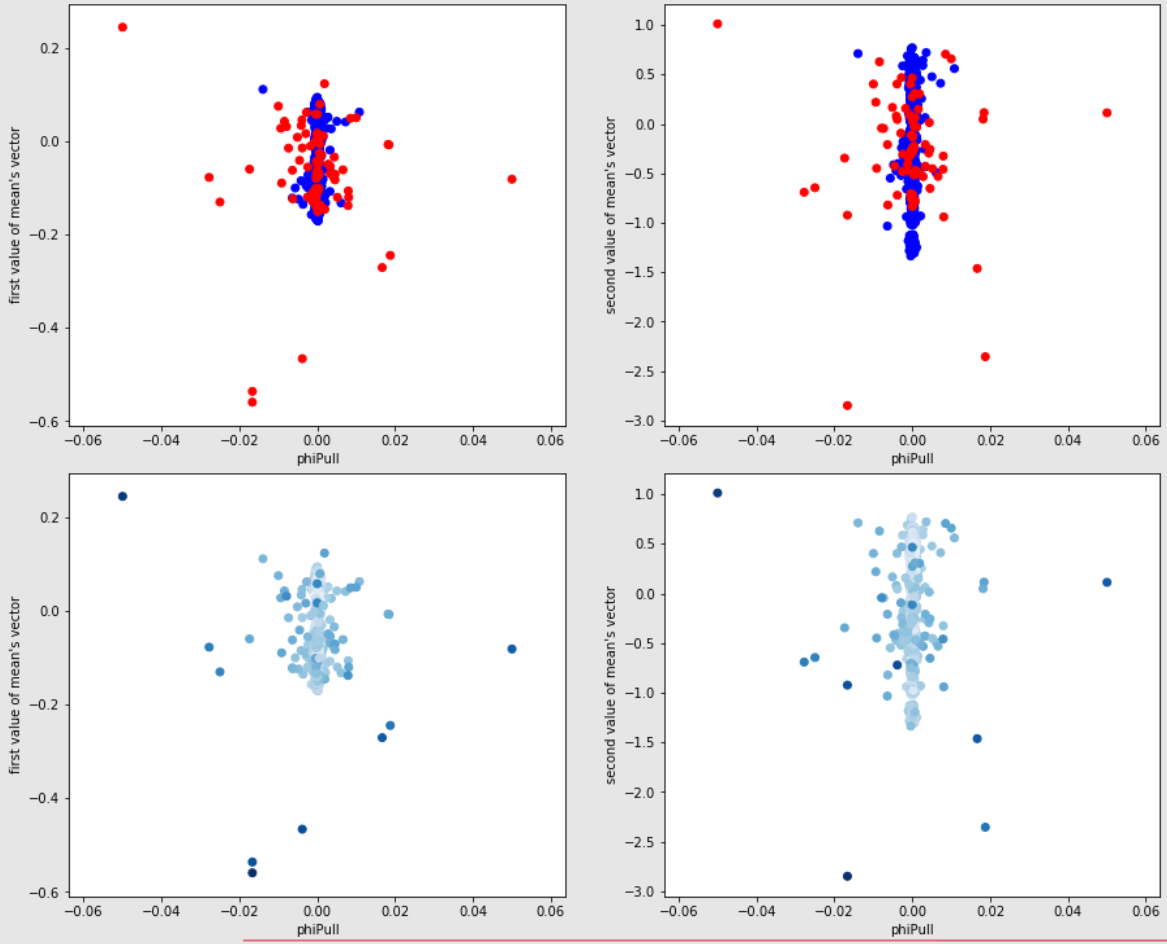
Analysis

Variational Autoencoder (LHC18q)



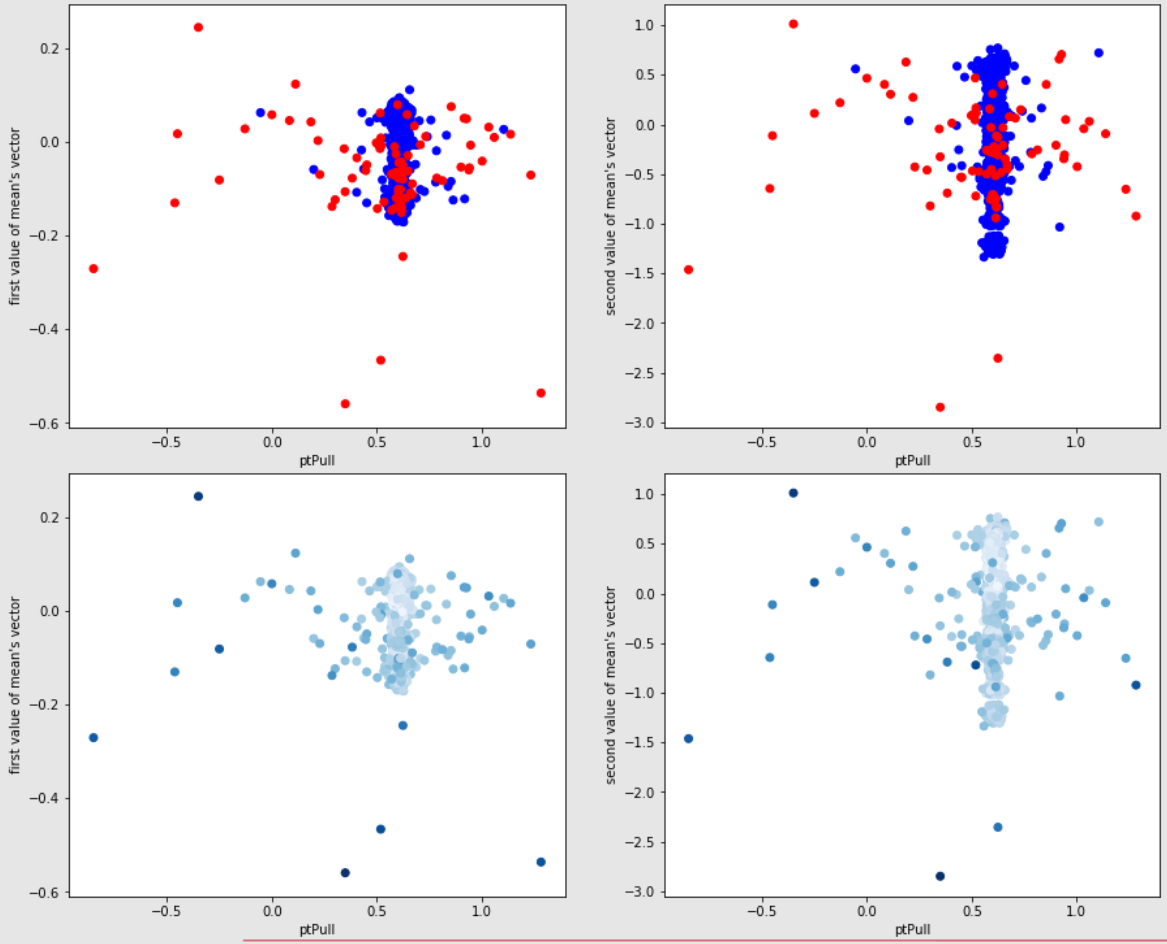
Analysis

Variational Autoencoder (LHC18q)



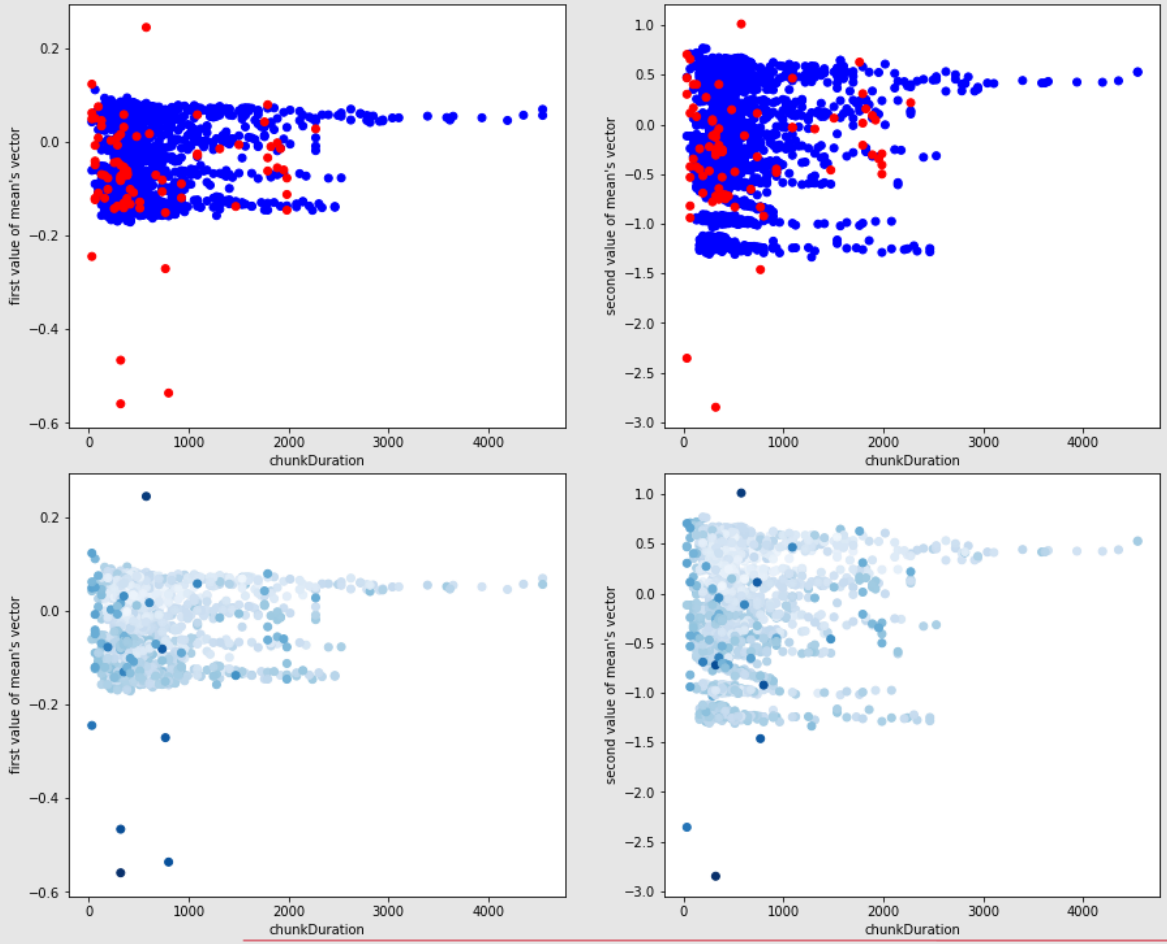
Analysis

Variational Autoencoder (LHC18q)



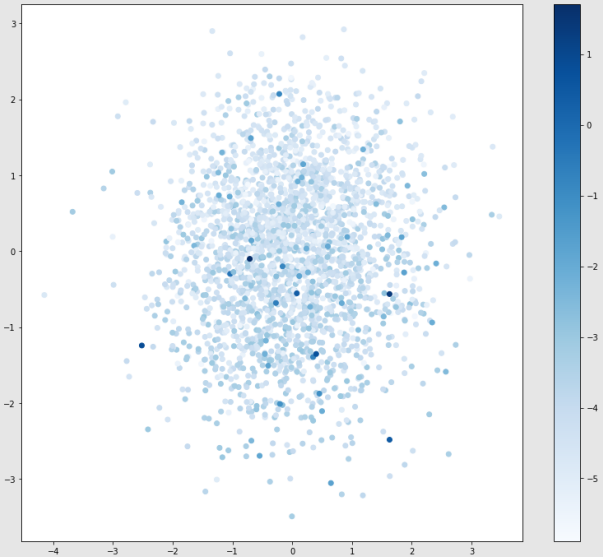
Analysis

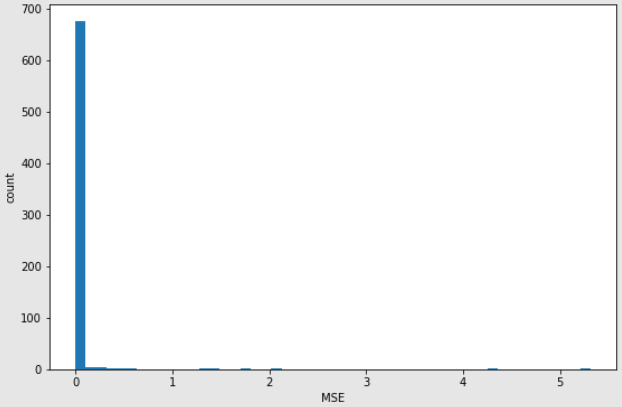
Variational Autoencoder (LHC18q)

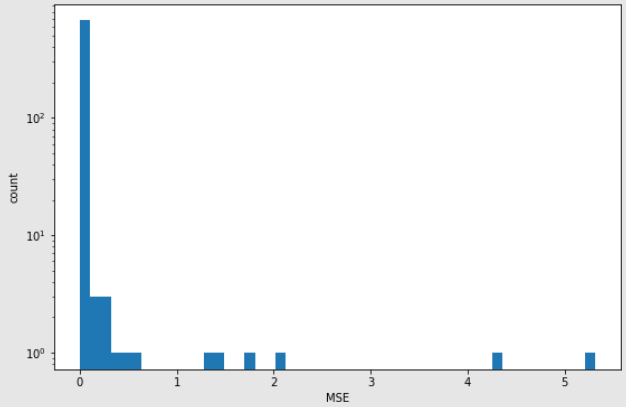


Analysis

Resemblance of autoencoder and variational autoencoder models



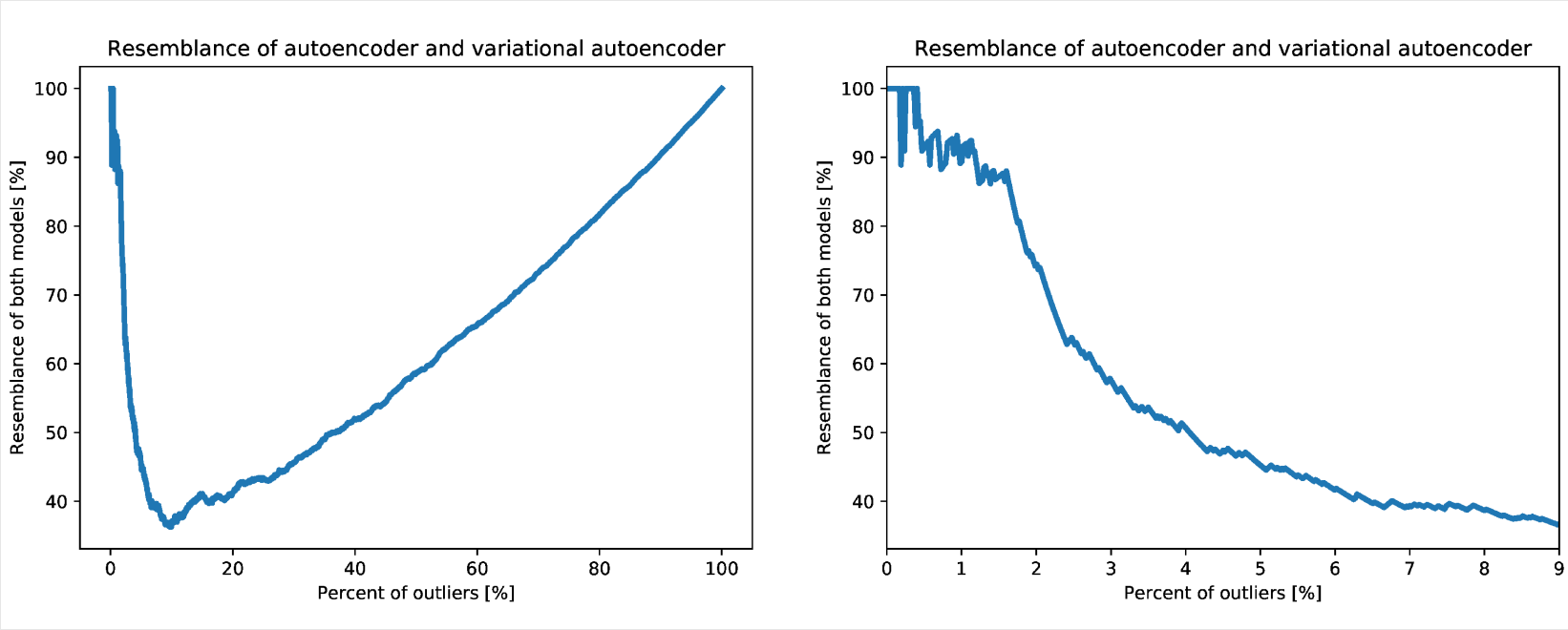




chunk id	autoencoder rank	vae rank
1565	1	1
1934	2	3
2489	3	2
1235	4	4
1359	5	5
6589	6	7
7846	7	6
4897	8	8

Analysis

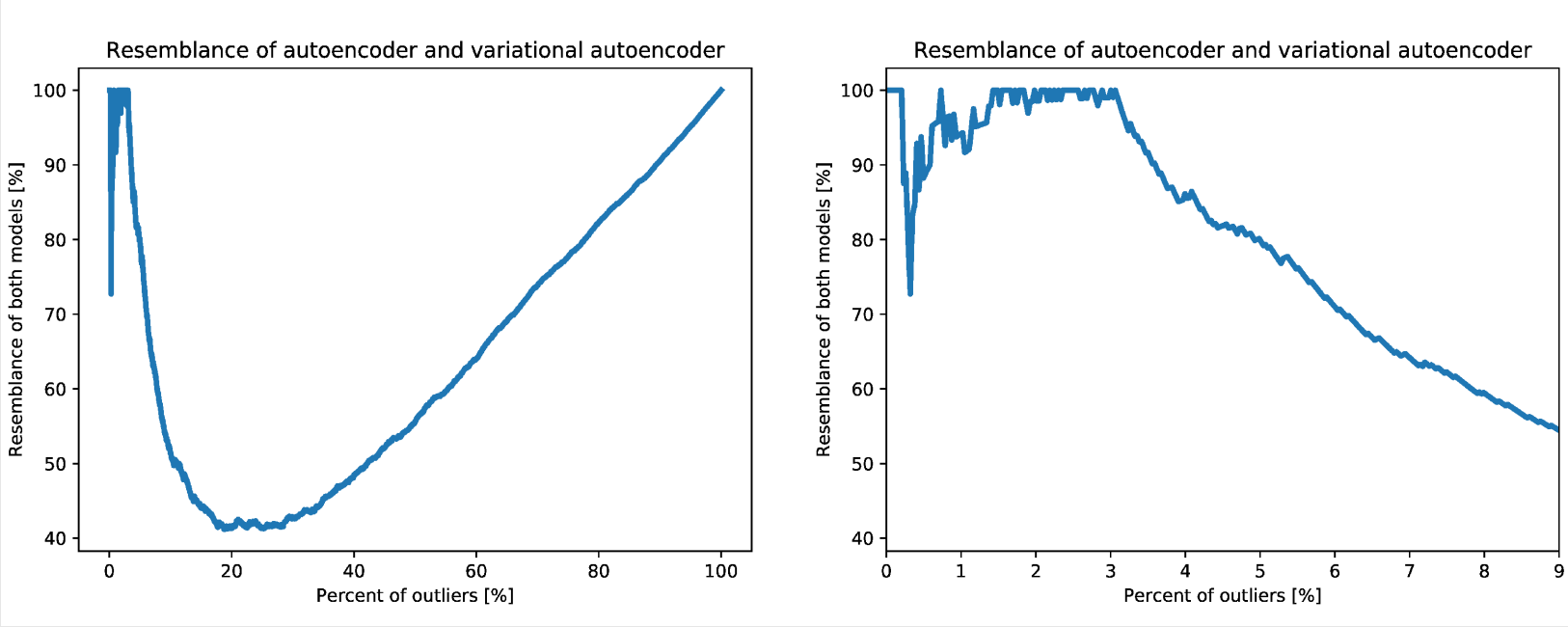
Resemblance of autoencoder and variational autoencoder models (LHC18q,LHC18r)



Current method “find” 2.75% outliers

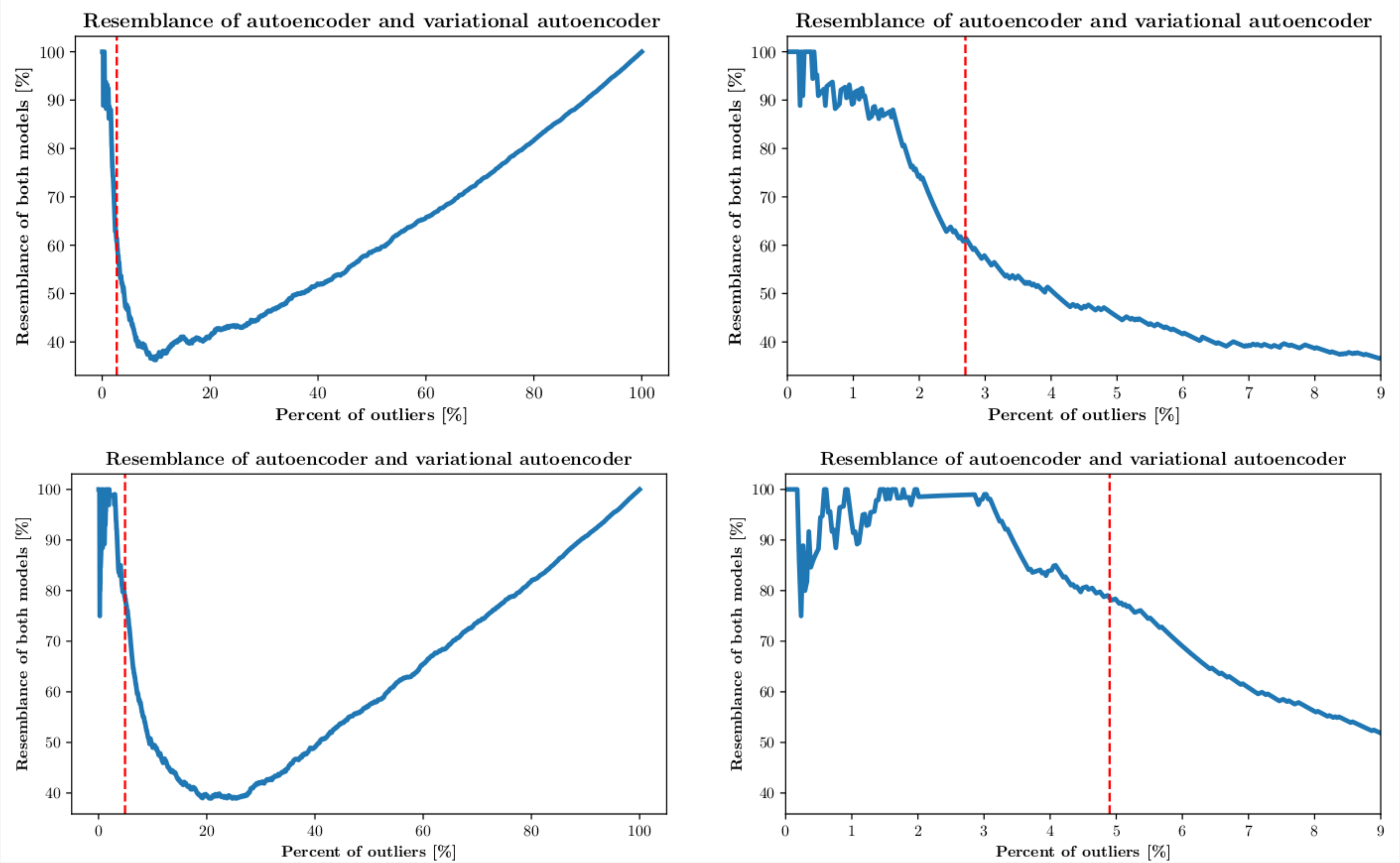
Analysis

Resemblance of autoencoder and variational autoencoder models (LHC18f,LHC18o,LHC18p)



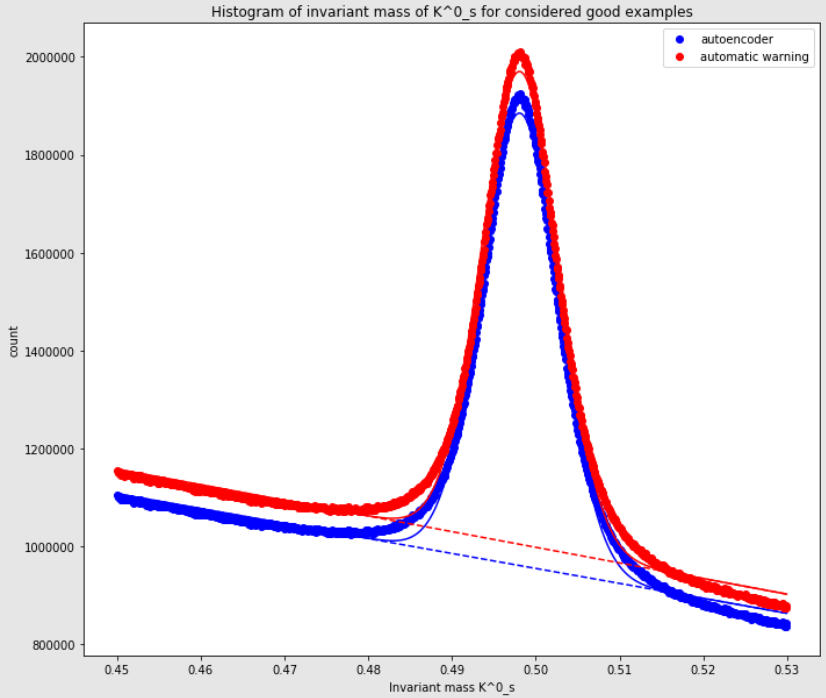
Current method “find” 4.87% outliers

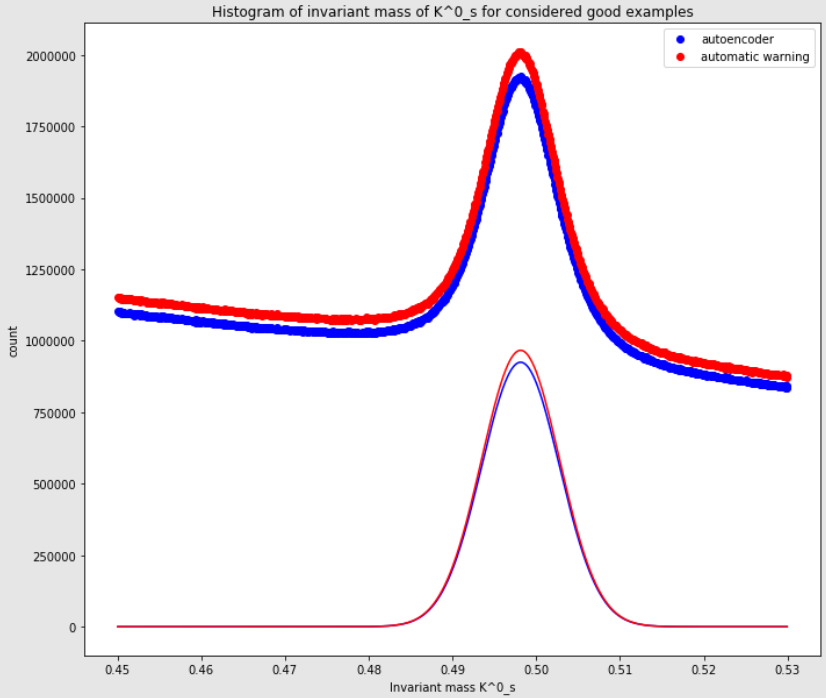
Analysis


Resemblance of autoencoder and variational autoencoder
Resemblance of both models [%]
Percent of outliers [%]
Resemblance of autoencoder and variational autoencoder
Resemblance of both models [%]
Percent of outliers [%]
Resemblance of autoencoder and variational autoencoder
Resemblance of both models [%]
Percent of outliers [%]
Resemblance of autoencoder and variational autoencoder
Resemblance of both models [%]
Percent of outliers [%]

Validation

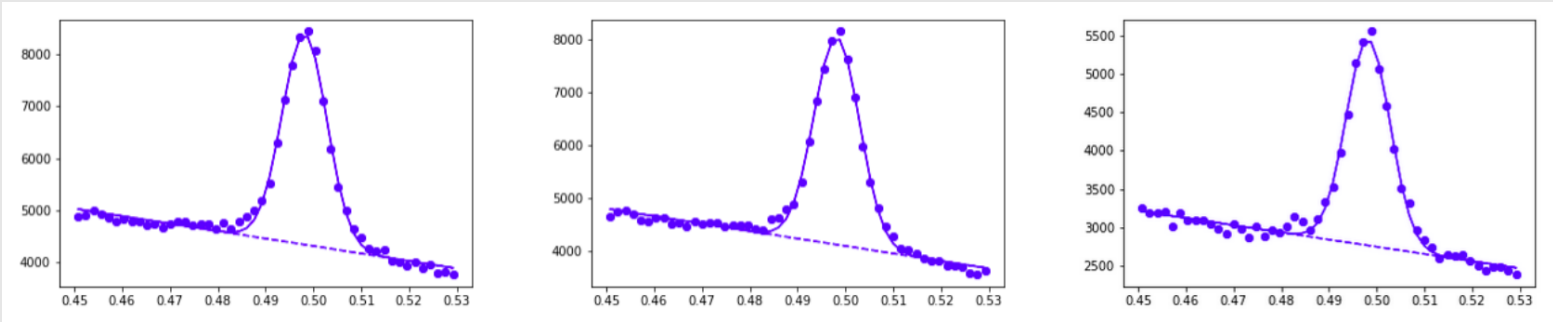
We also look at measured invariant mass K^0_s

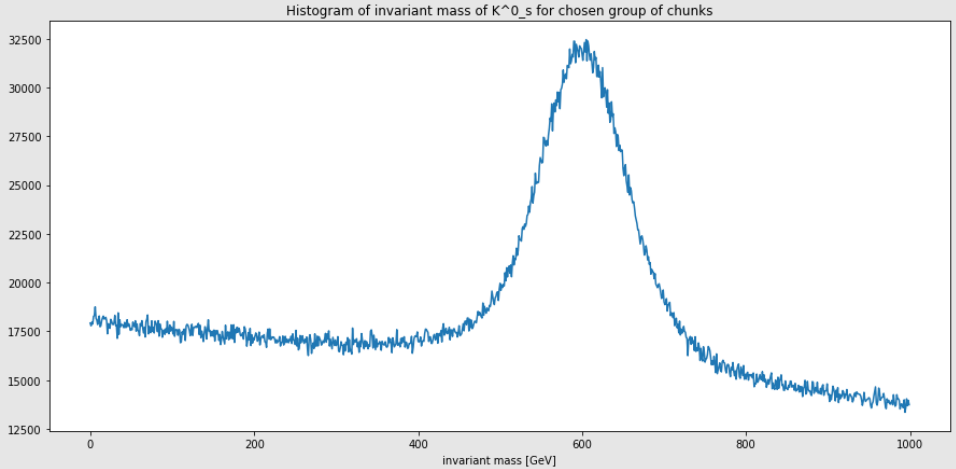




Validation

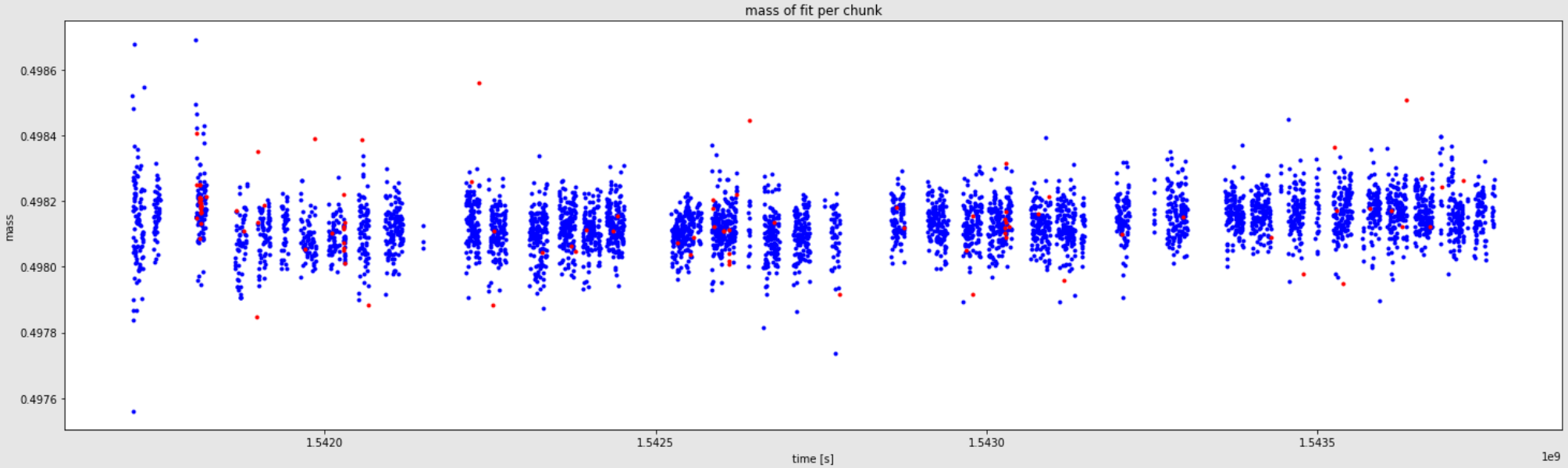
For every chunk we fit peak of invariant mass K^0_s





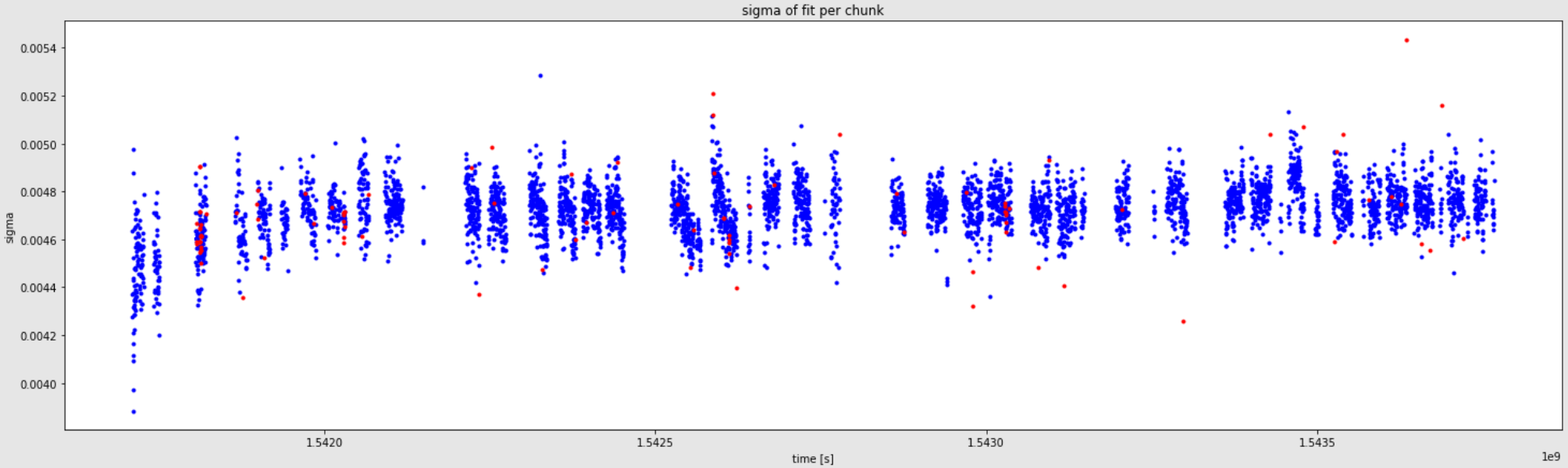
Validation

Fit results (LHC18q,LHC18r)



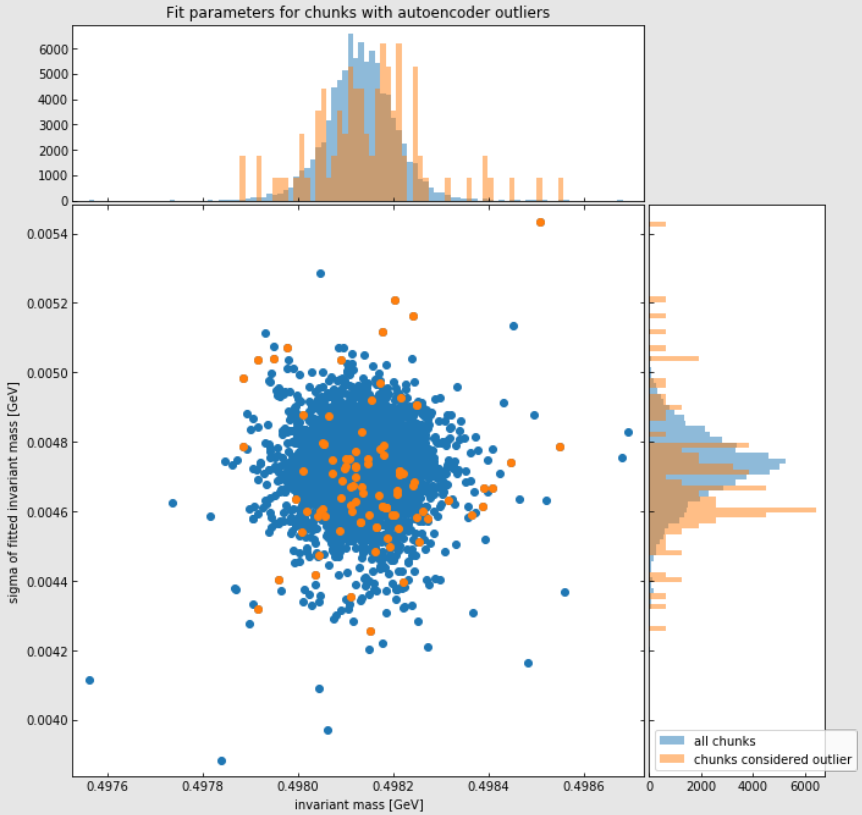
Validation

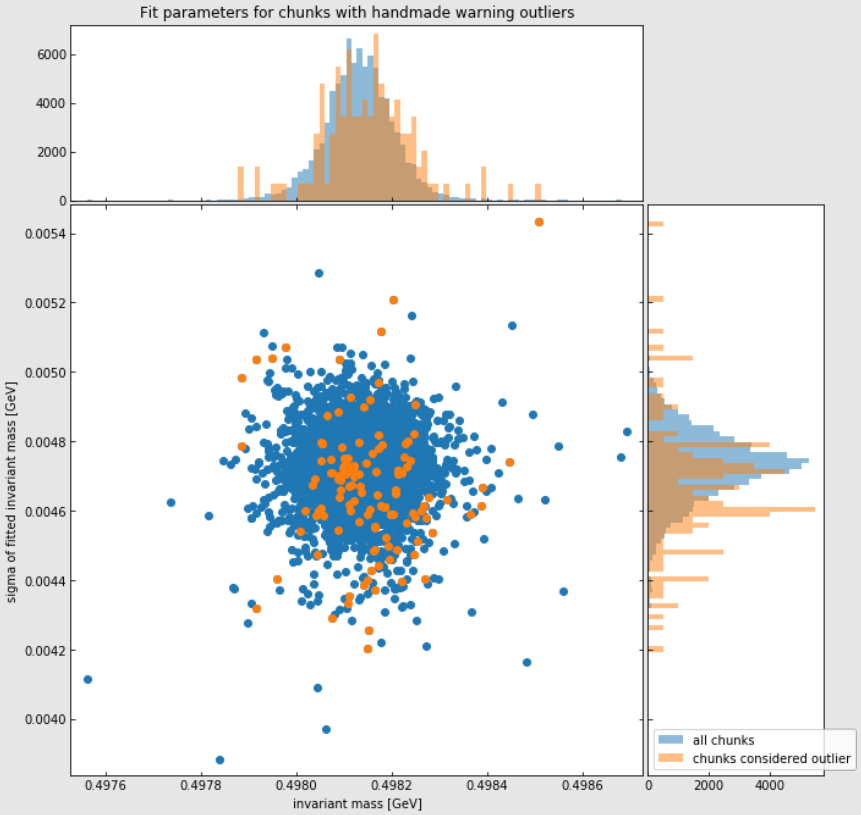
Fit results (LHC18q,LHC18r)



Validation

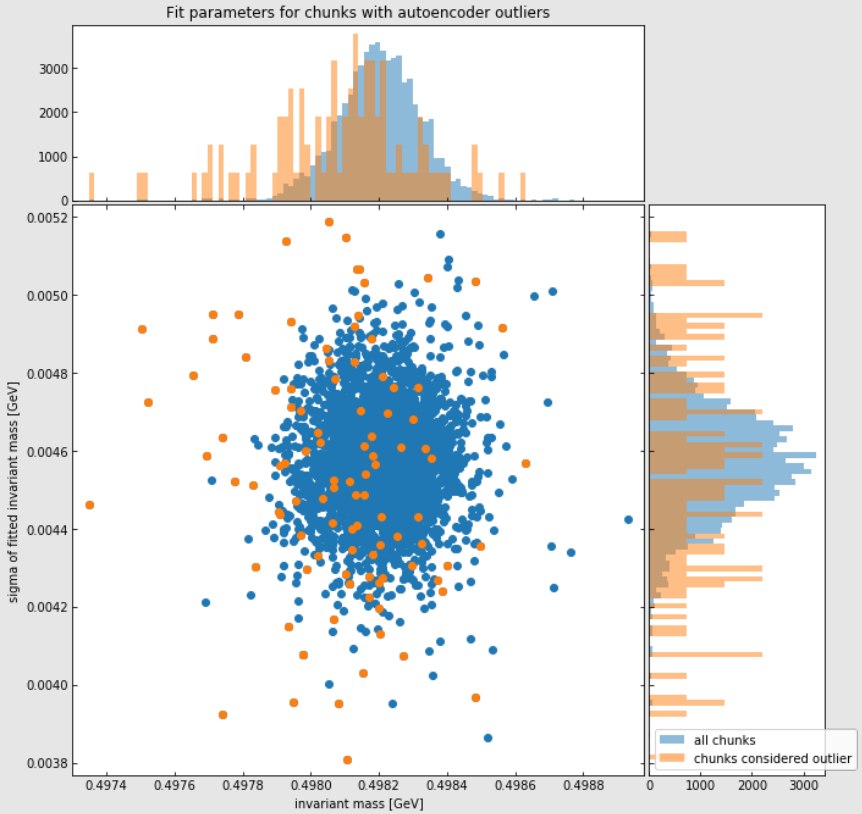
Fit results (LHC18q,LHC18r)

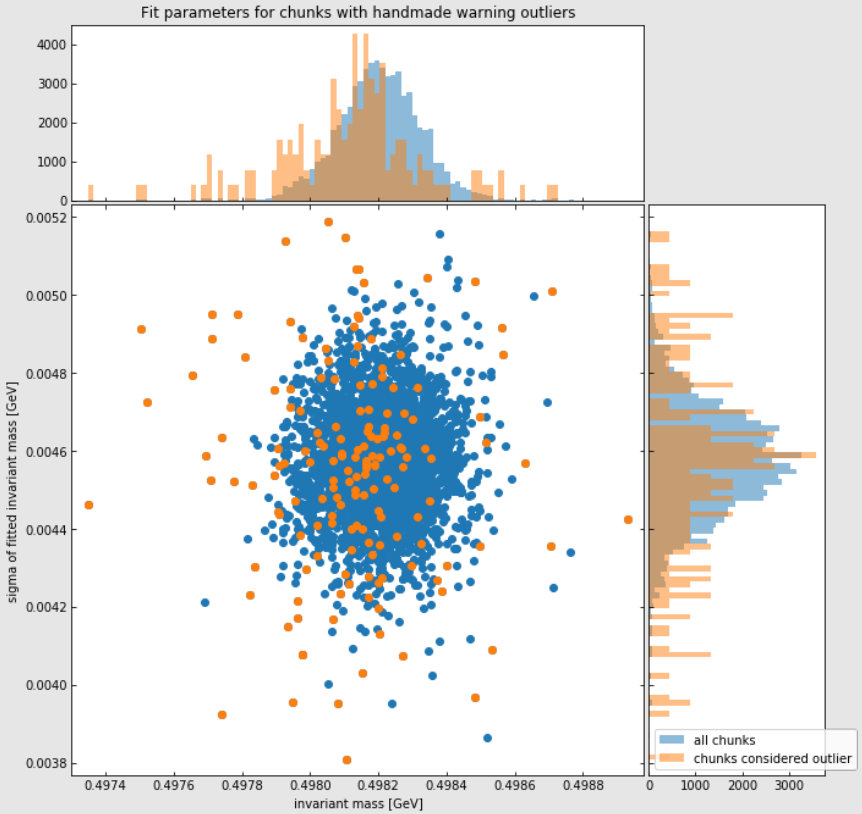




Validation

Fit results (LHC18f,LHC18o,LHC18p)





Summary

- Unsupervised Machine Learning algorithms can detect outliers
- In contrast to current methods instead of division outlier/inlier we have probability of being outlier
- Autoencoders and Variational Autoencoders give new ways to visualize data
- Validation using invariant mass fit