9TH International Conference on New Frontiers in Physics 2020

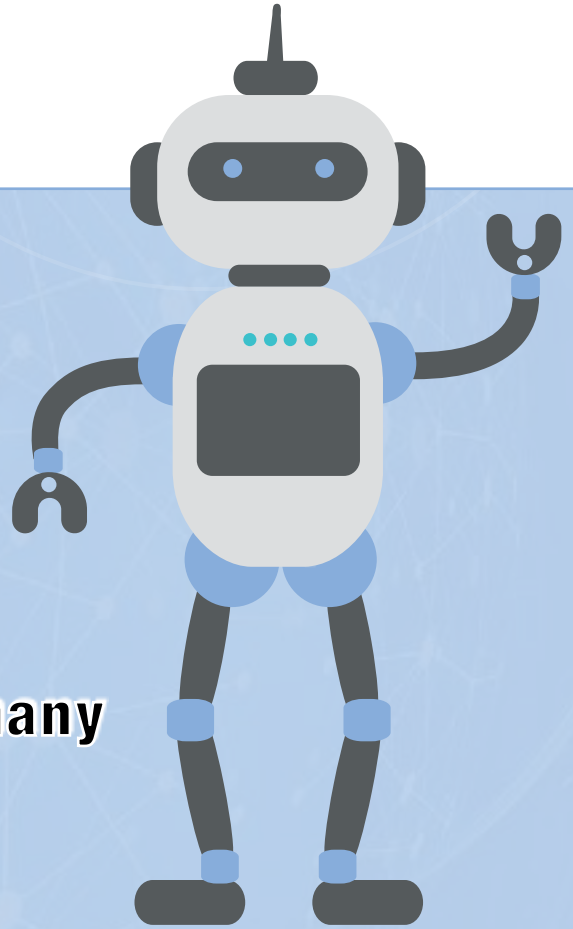
FIAS Frankfurt Institute for Advanced Studies

OUTLIER DETECTION IN NUCLEAR COLLISION USING UNSUPERVISED (DEEP-) LEARNING

Thaprasop P., et al. (2020). arXiv:2007.15830

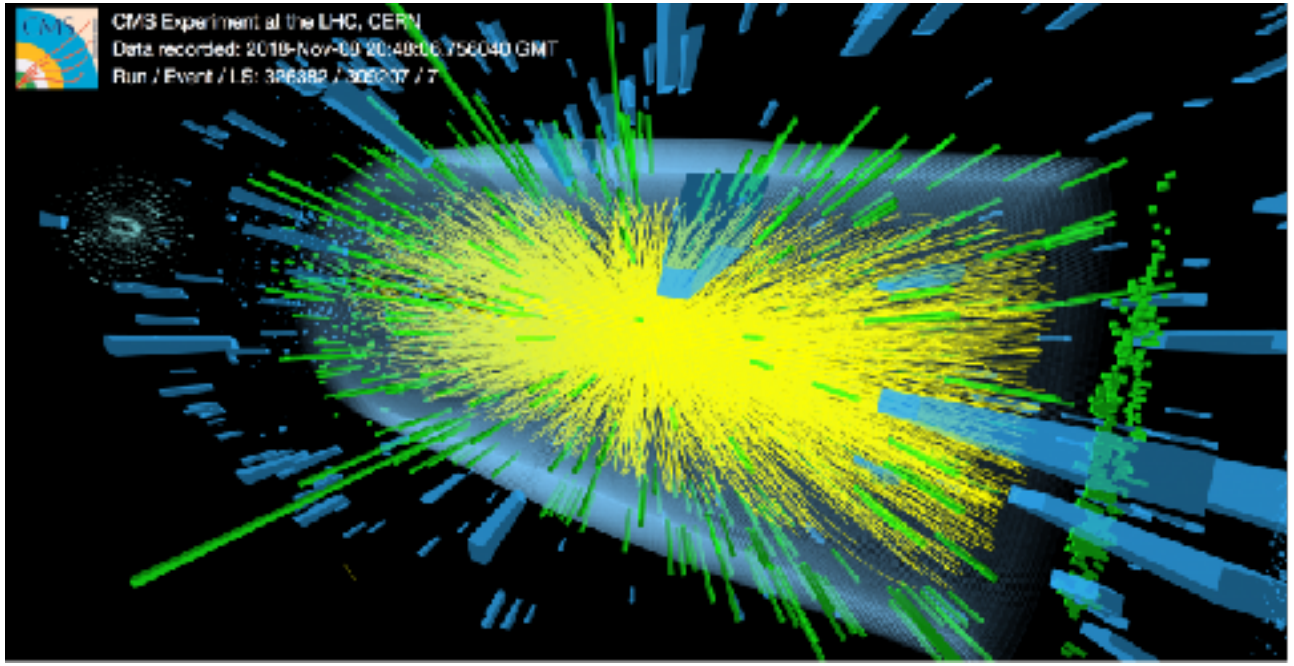
P. Thaprasop[1], K. Zhou[2], J. Steinheimer[2], and C. Herold[1]

1. School of Physics, Suranaree University of Technology, Thailand.
2. Frankfurt Institution for Advanced Studies (FIAS), Frankfurt, Germany



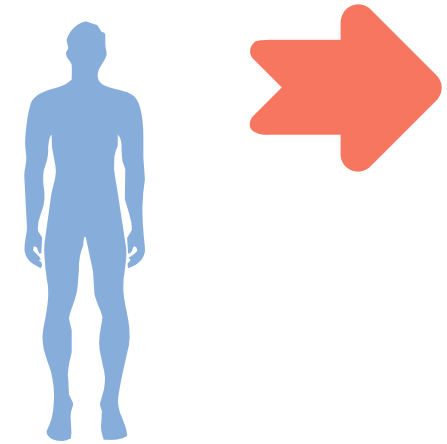
Presenter: Punnatat Thaprasop

05/09/20



Ref: cms.cern

Data



Human

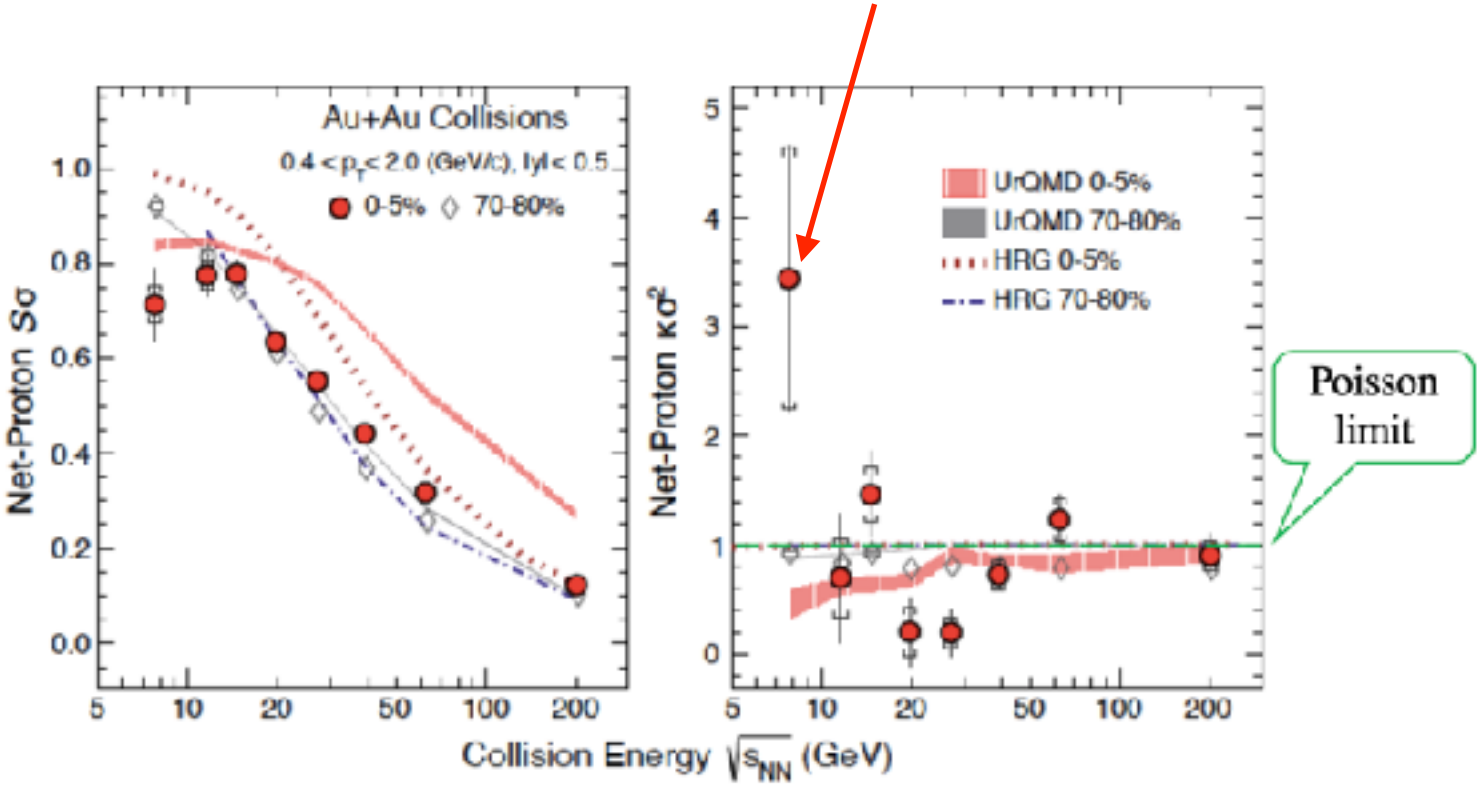
Machine Learning (ML)

Human's Results

ML's Results

Bias!

$\sqrt{S_{NN}} = 7.7$ **GeV. Signal for phase boundary, Imperfect centrality determination, Detector malfunction,…?**

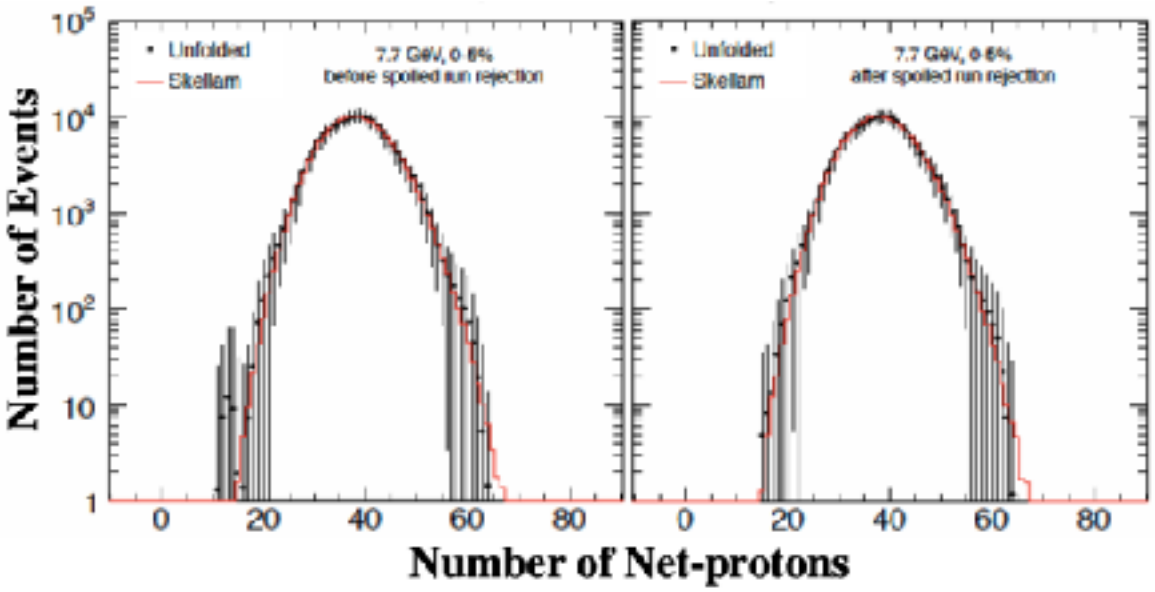


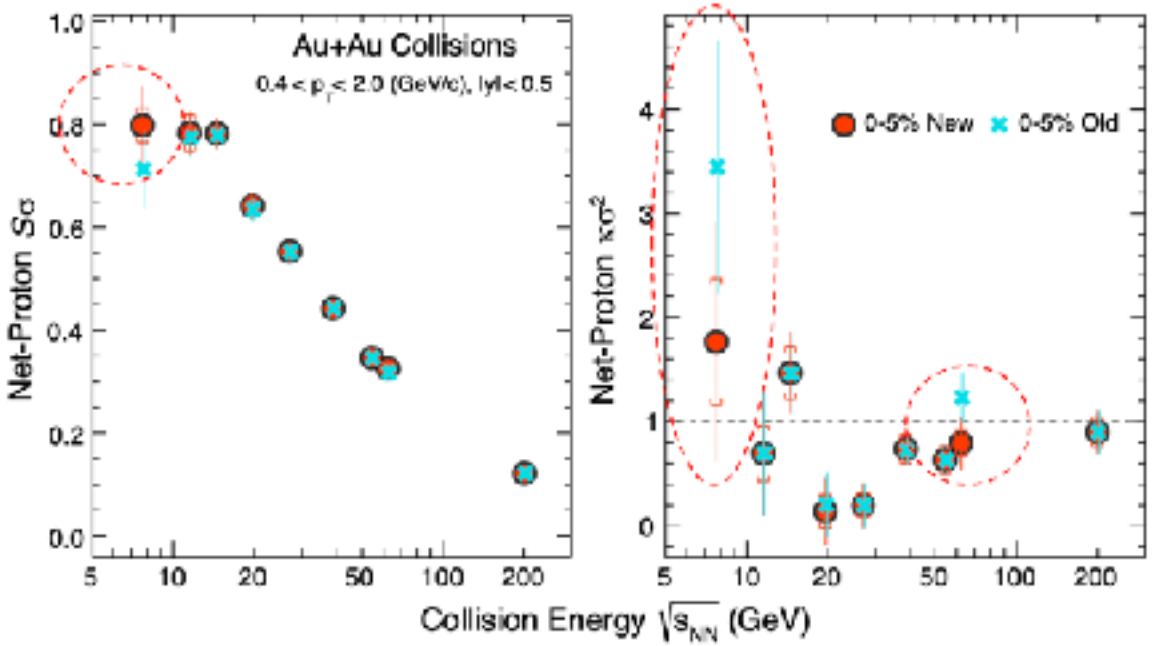
Efficiency Corrected Cumulant Ratios

$$\kappa\sigma^2 = \frac{C_4}{C_2} \qquad \kappa\sigma^2 = \frac{C_4}{C_2}$$

Ref: STAR: arXiv: 2001.02852

Ref: STAR: arXiv: 2001.02852





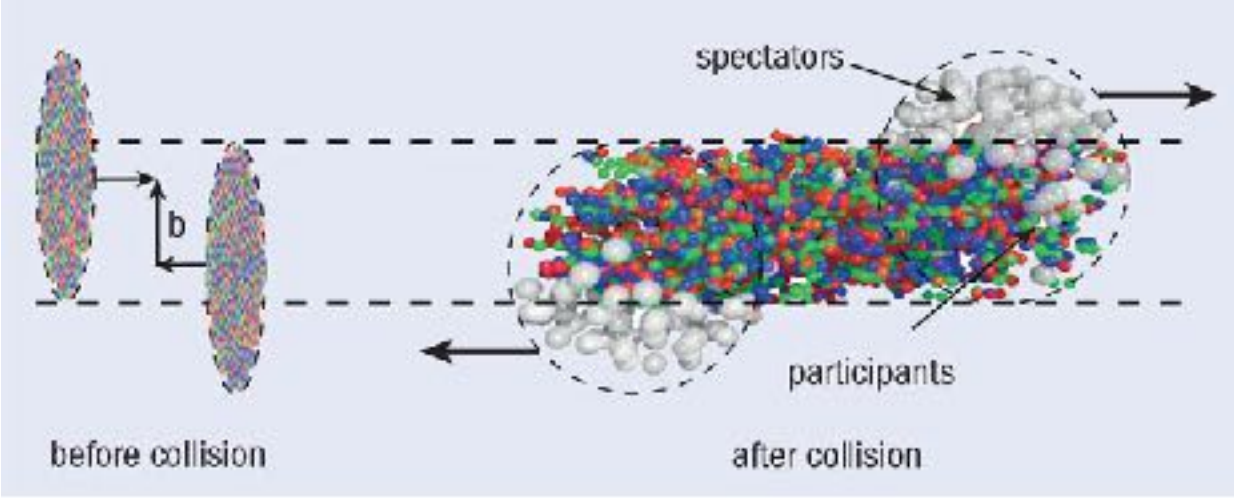
"Spoiled events" / "Outlier"

Objective: To develop unsupervised learning model that are able to detect spoiled/outlier events

UrQMD

Au + Au

$\sqrt{S_{NN}} = 7.7$ **GeV**



Spoilers / Outliers

600 Peripheral (P) events

Impact parameter (b) = 7 fm

Backgrounds

184000 Central (C) events

Impact parameter (b) = 3 fm

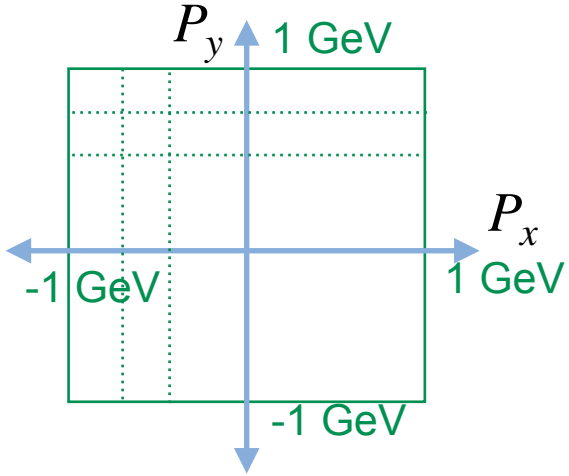
184600 events In total

Charge particle

Mid rapidity $|y| < 0.5$

Total = 184600 events

Histogram of charge particles in Transverse momentum

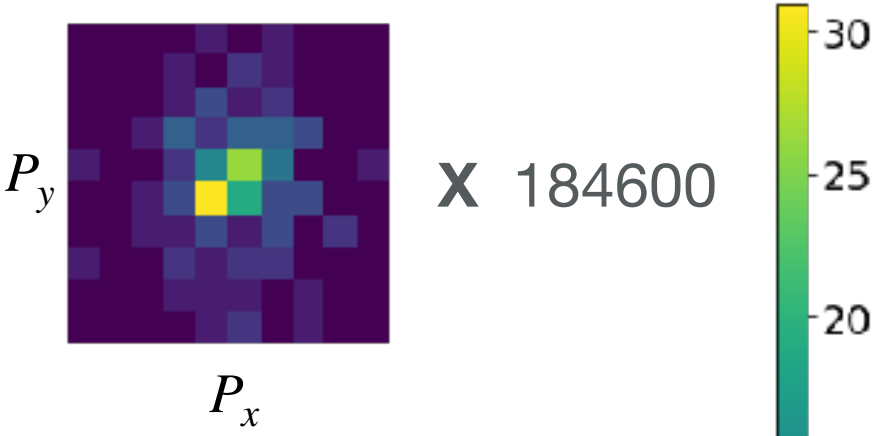


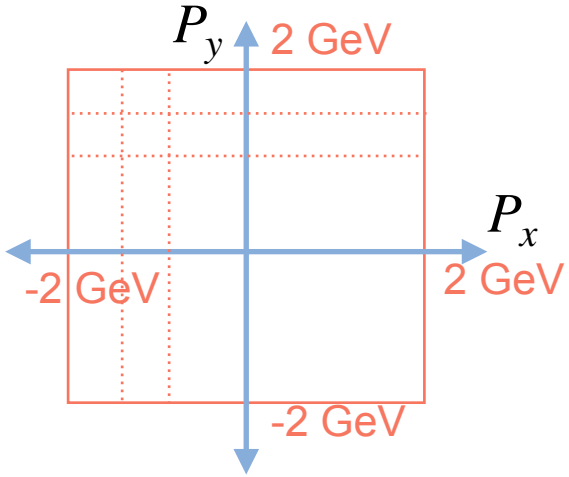
$$-1 \text{ GeV} \leq P_x \leq 1 \text{ GeV}$$

$$-1 \text{ GeV} \leq P_y \leq 1 \text{ GeV}$$

10 x 10 channels

$$\text{Bin width} = \frac{2}{10} = 0.2 \text{ GeV}$$



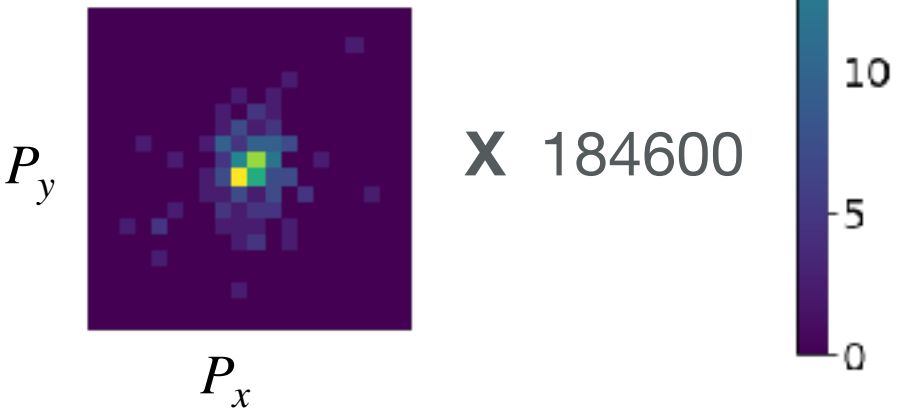


$$-2 \text{ GeV} \leq P_x \leq 2 \text{ GeV}$$

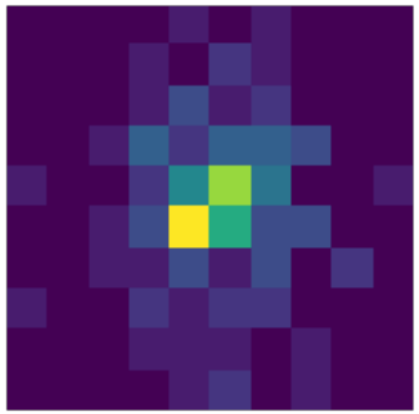
$$-2 \text{ GeV} \leq P_y \leq 2 \text{ GeV}$$

20 x 20 channels

$$\text{Bin width} = \frac{4}{20} = 0.2 \text{ GeV}$$



Momentum Feature

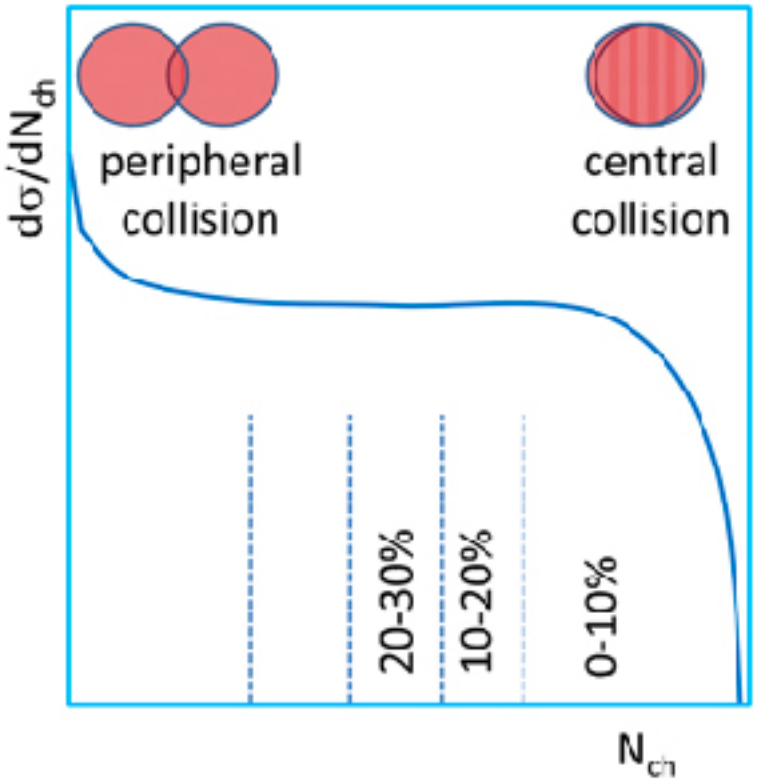


88

203

77

Ref: Chaudhuri A. K. (2014), A short course on relativistic heavy ion collisions.



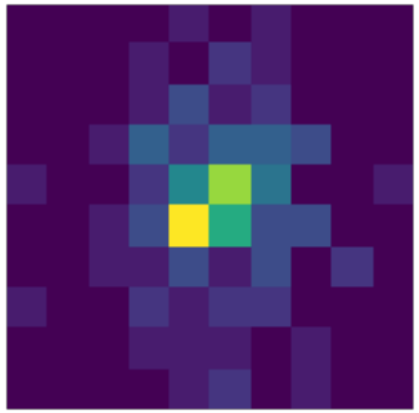
Peripheral

Central

Peripheral

Normalized

Momentum Feature

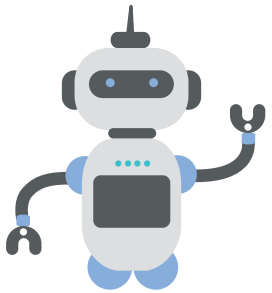


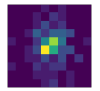
1

1

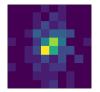
1

Machine Learning

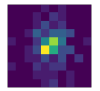




Peripheral

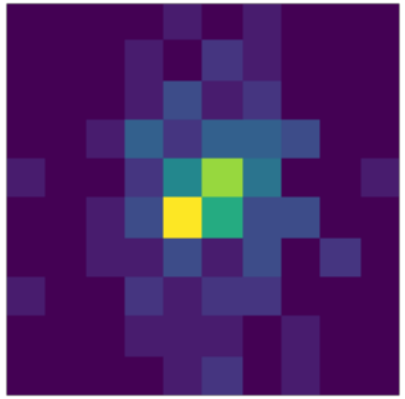


Central



Peripheral

Implement PCA



10x10

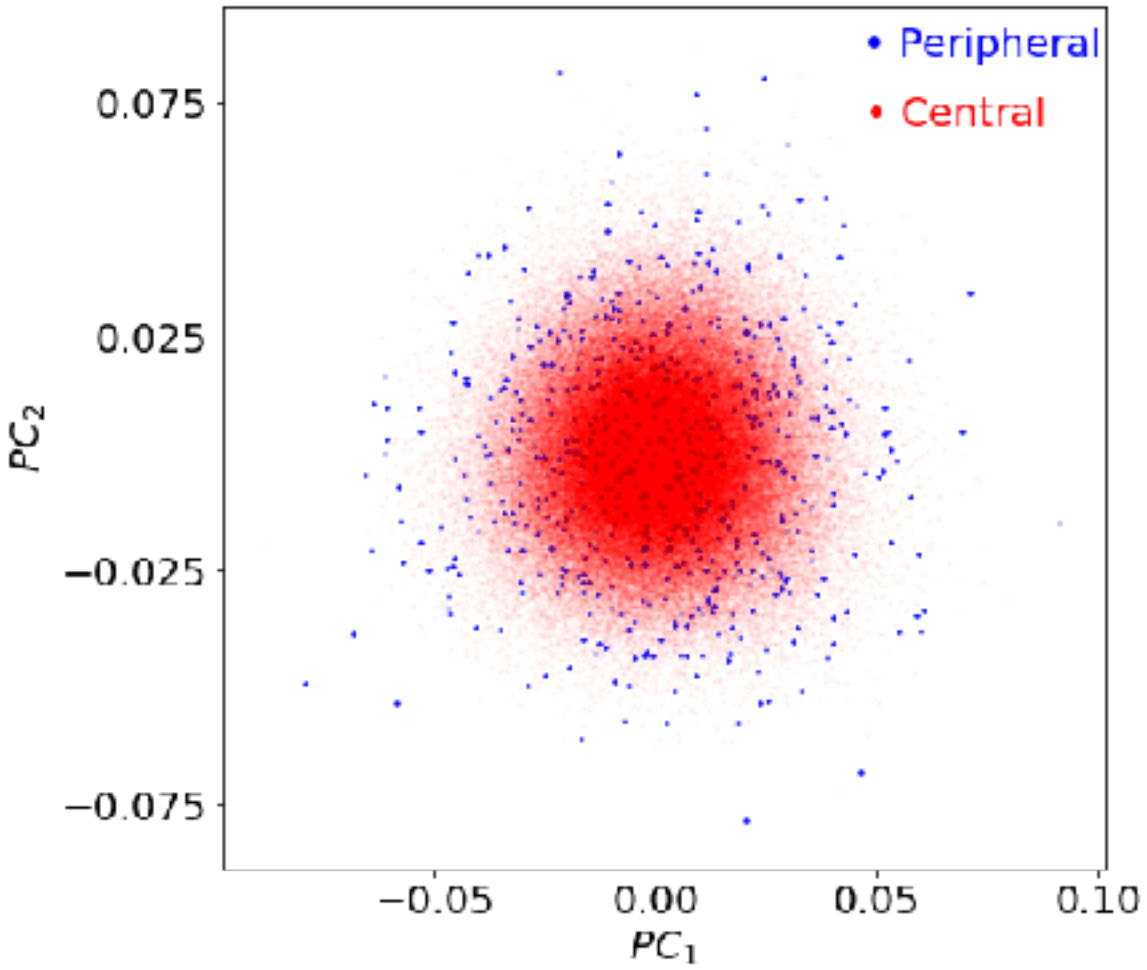
100 dimensional vector

$$X(1) = \begin{pmatrix} X_1(1) \\ X_2(1) \\ \vdots \\ X_{100}(1) \end{pmatrix}$$

PCA

$$X'(1) = \begin{pmatrix} PC_1(1) \\ PC_2(1) \\ \vdots \\ PC_{100}(1) \end{pmatrix}$$

10x10	20x20
184600 events	*184600 events*
Flatten	Flatten
100D vectors	400D vectors
PCA	PCA

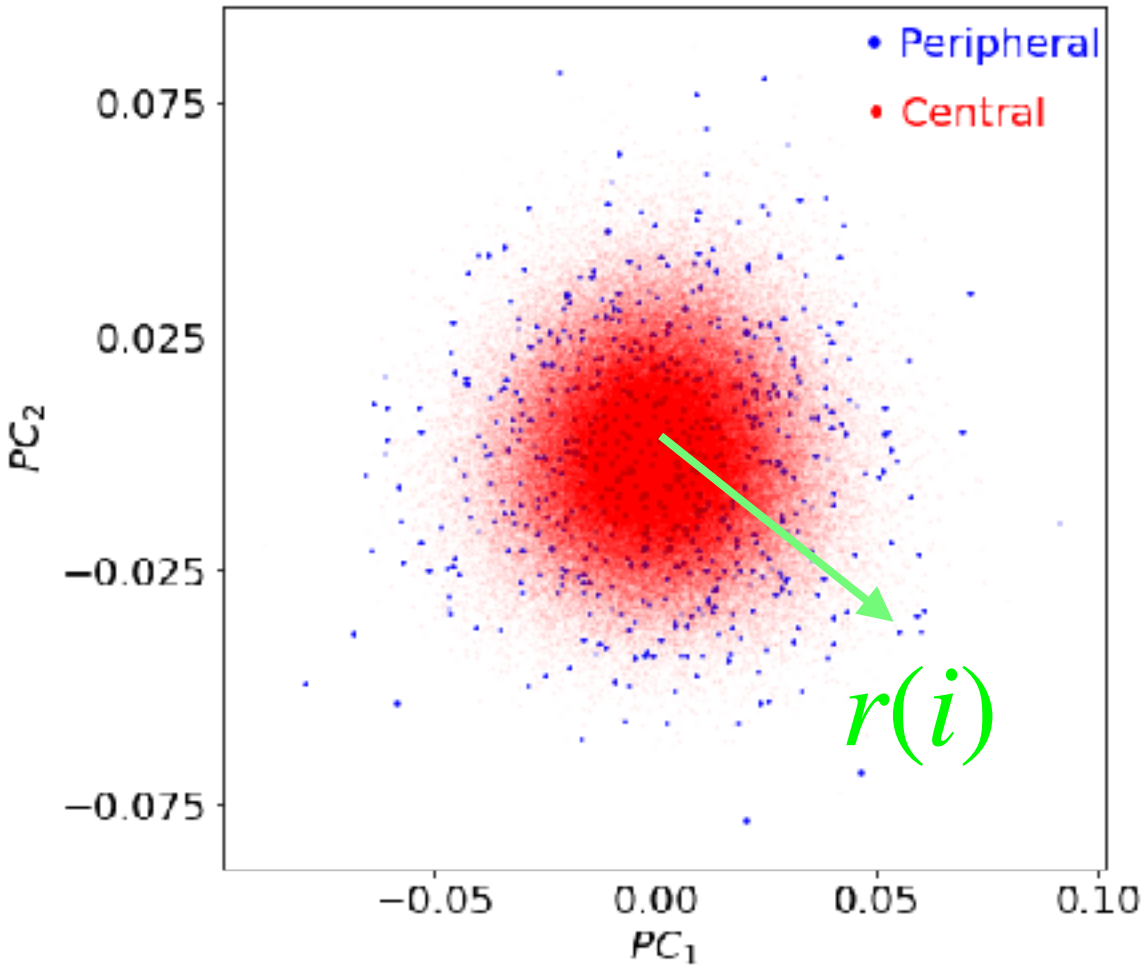


Peripheral events spread over the larger area

Input - 100D vectors

PCA transform

2D vectors



PCA radius comparison

$$r(i) = \sqrt{X^2_{PC_1}(i) + X^2_{PC_2}(i) + X^2_{PC_3}(i) + \ldots + X^2_{PC_m}(i)}$$

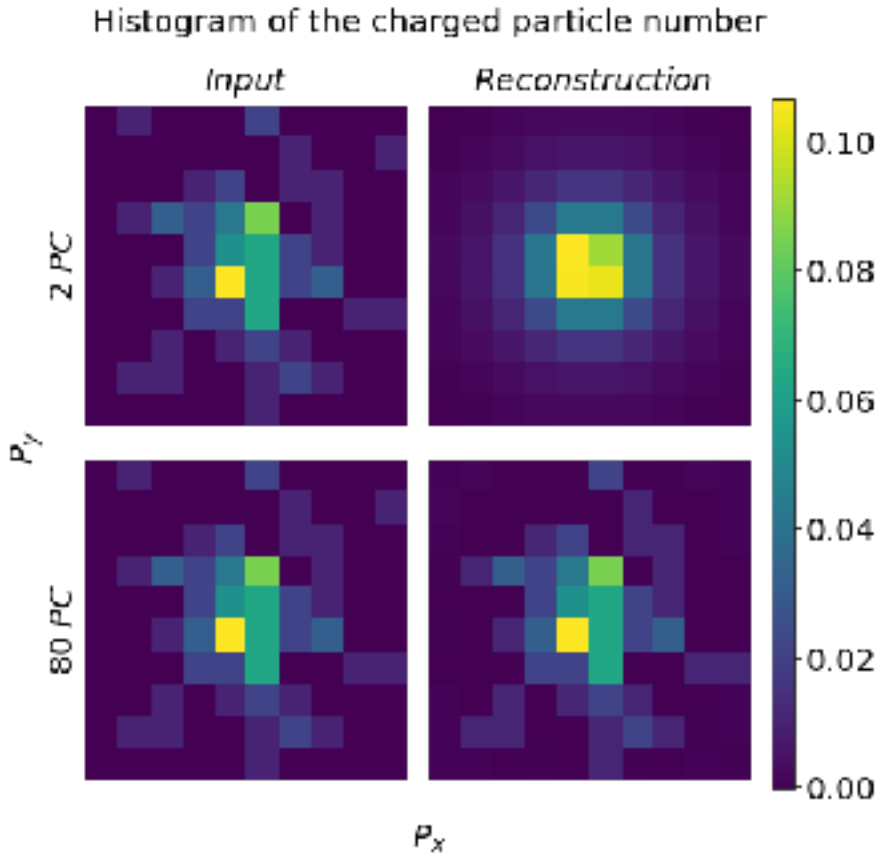
$$r(i) = \sqrt{X^2_{PC_1}(i) + X^2_{PC_2}(i)}$$

PCA reconstruction error

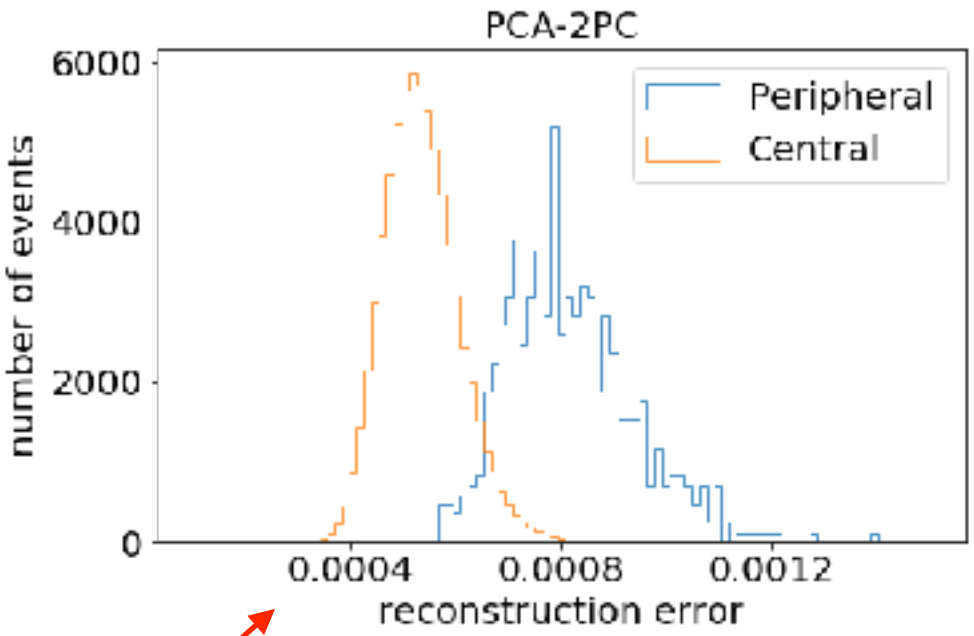
Input - 100D vectors

2/80 PC

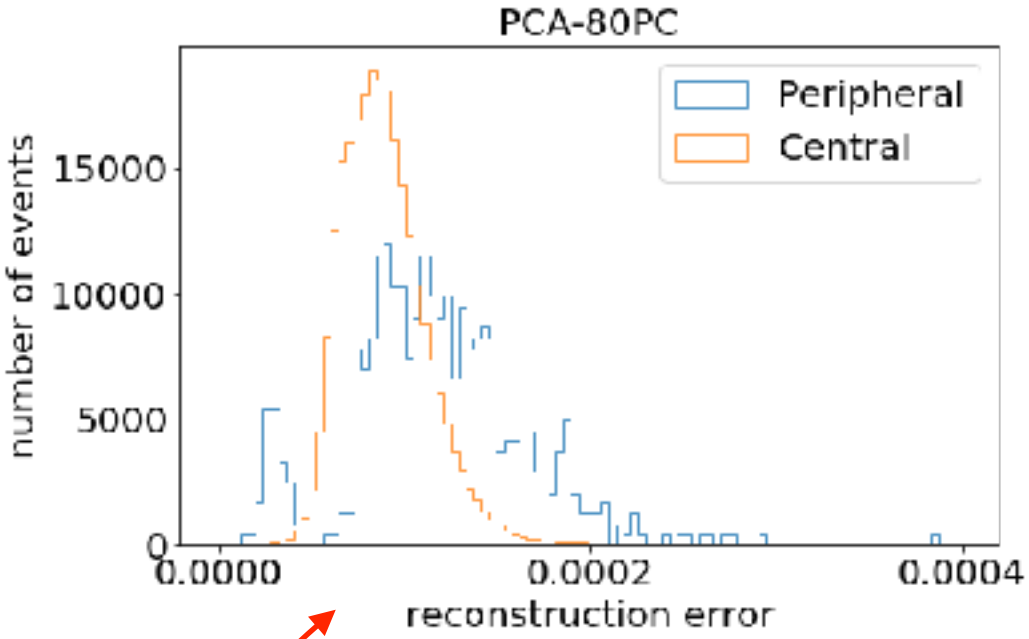
Output - 100D vectors



$$RE(i) = \frac{1}{N}\left(\sum_{j=1}^{N}[X_{rec_j}(i) - X_j(i)]^2\right)^{\frac{1}{2}}$$

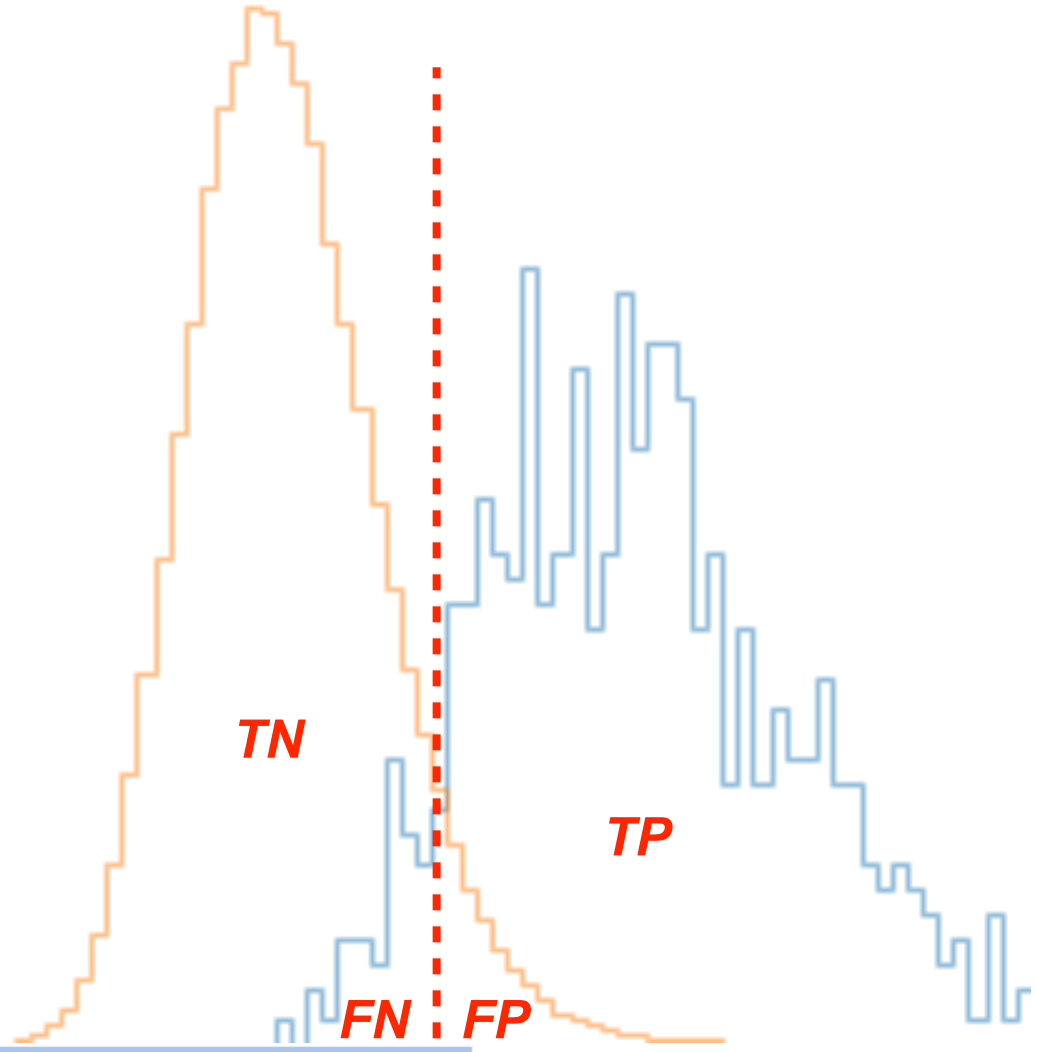


Larger value —> Better separation)
Small overlapped



Smaller value —> Worse separation
Large overlapped

How to quantify the performance of the model?



Receiver Operating Characteristics (ROC) curve

TP (True Positive): correctly classified signal

FP (False Positive): wrongly classified signal

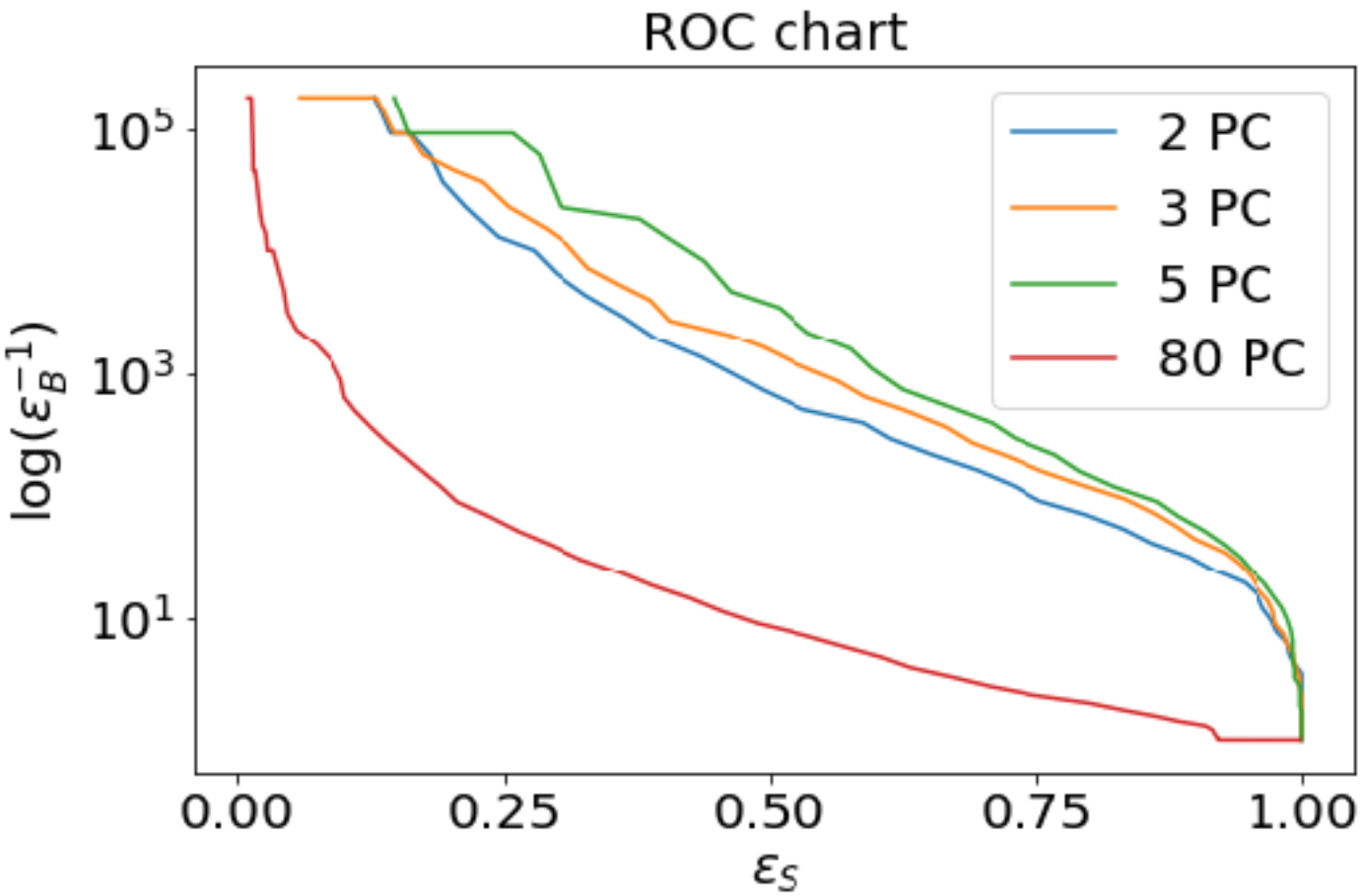
FN (False Negative): wrongly classified backgrounds

TN (True Negative): correctly classified backgrounds

$$\epsilon_s = \frac{TP}{TP + FN}$$

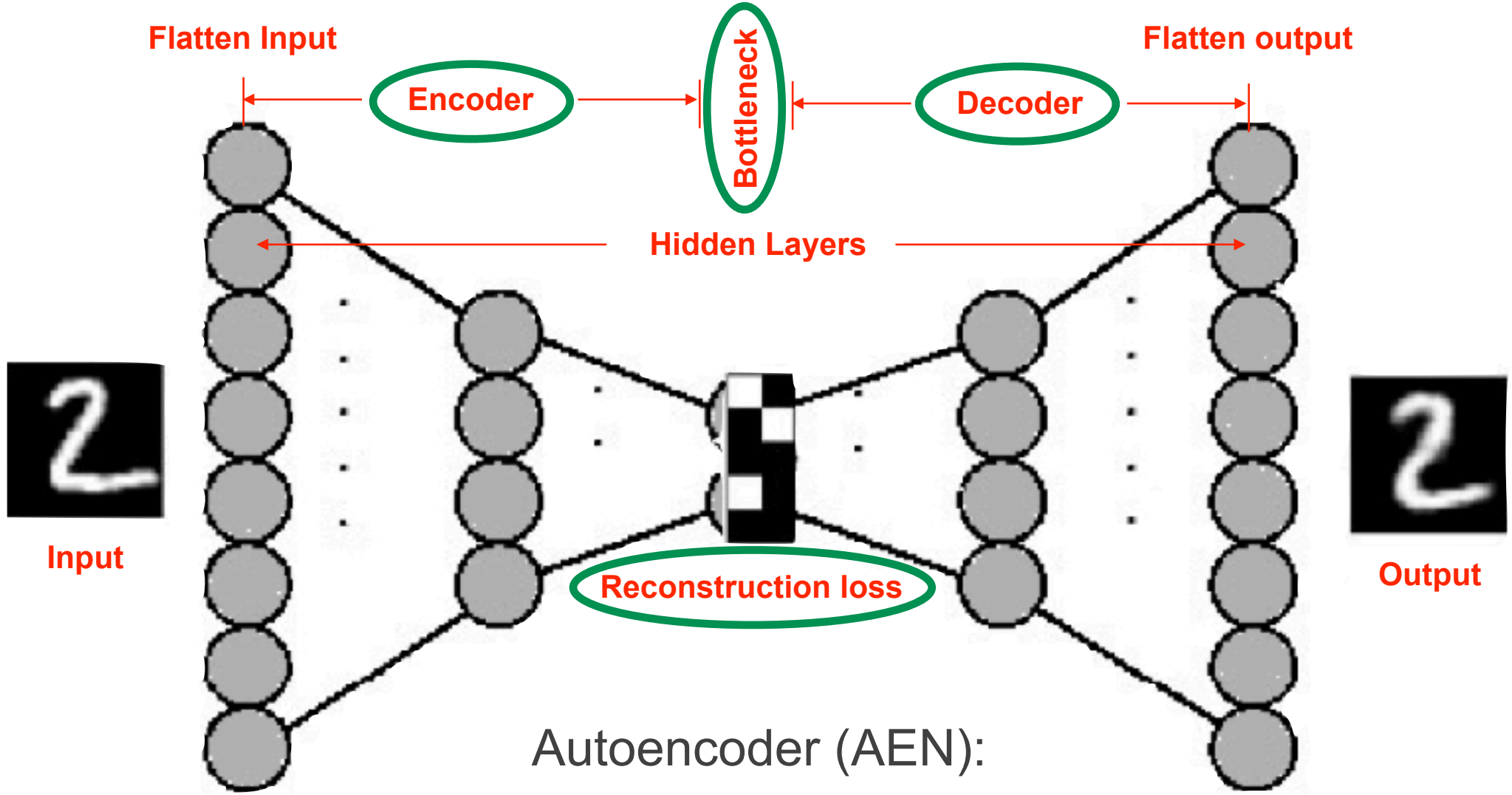
$$\epsilon_b = \frac{FP}{FP + TN}$$

Reconstruction Error (RE)



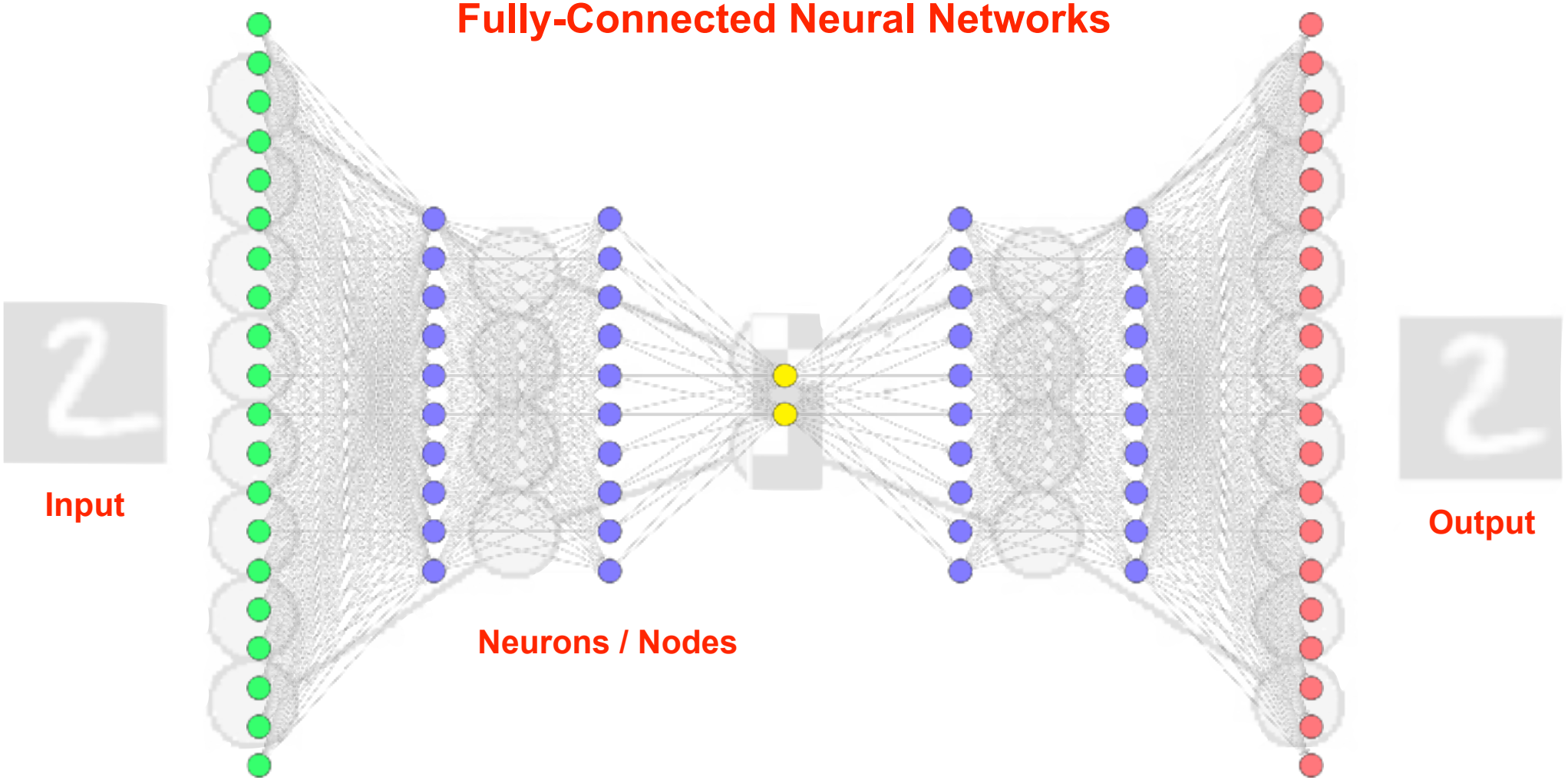
To identify good separation capabilities:

5 PC > 3 PC > 2 PC >> 80 PC

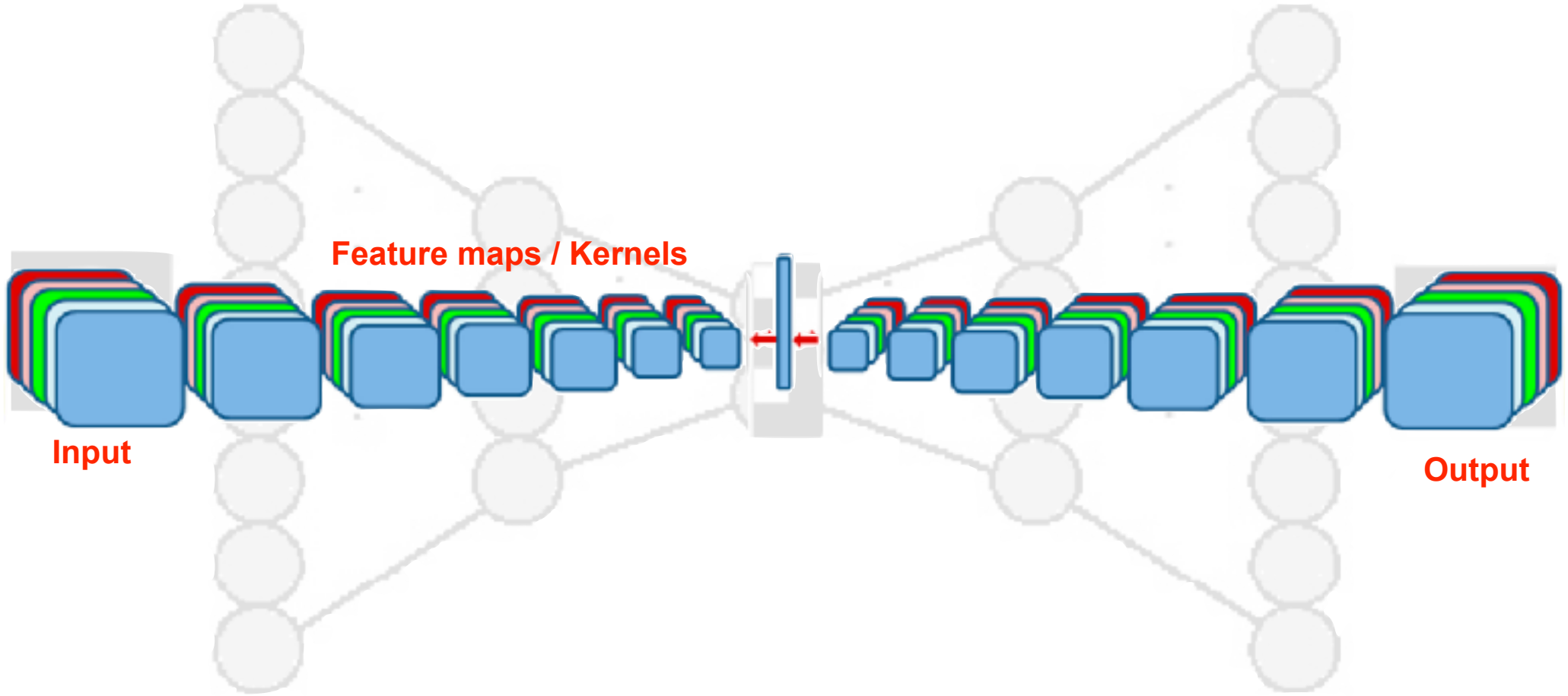


Autoencoder (AEN):
artificial neural network data coding algorithm

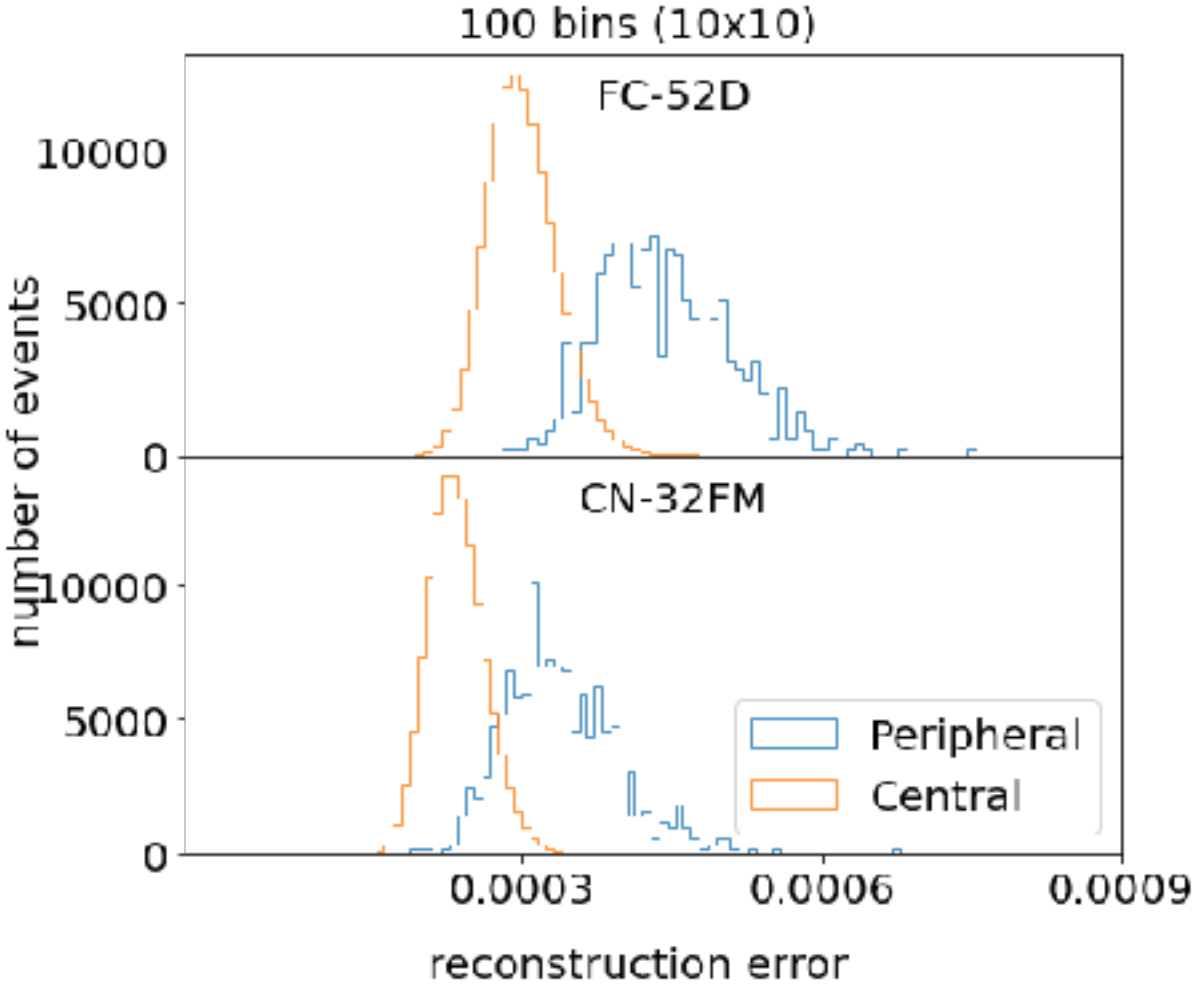
Fully-Connected Neural Networks

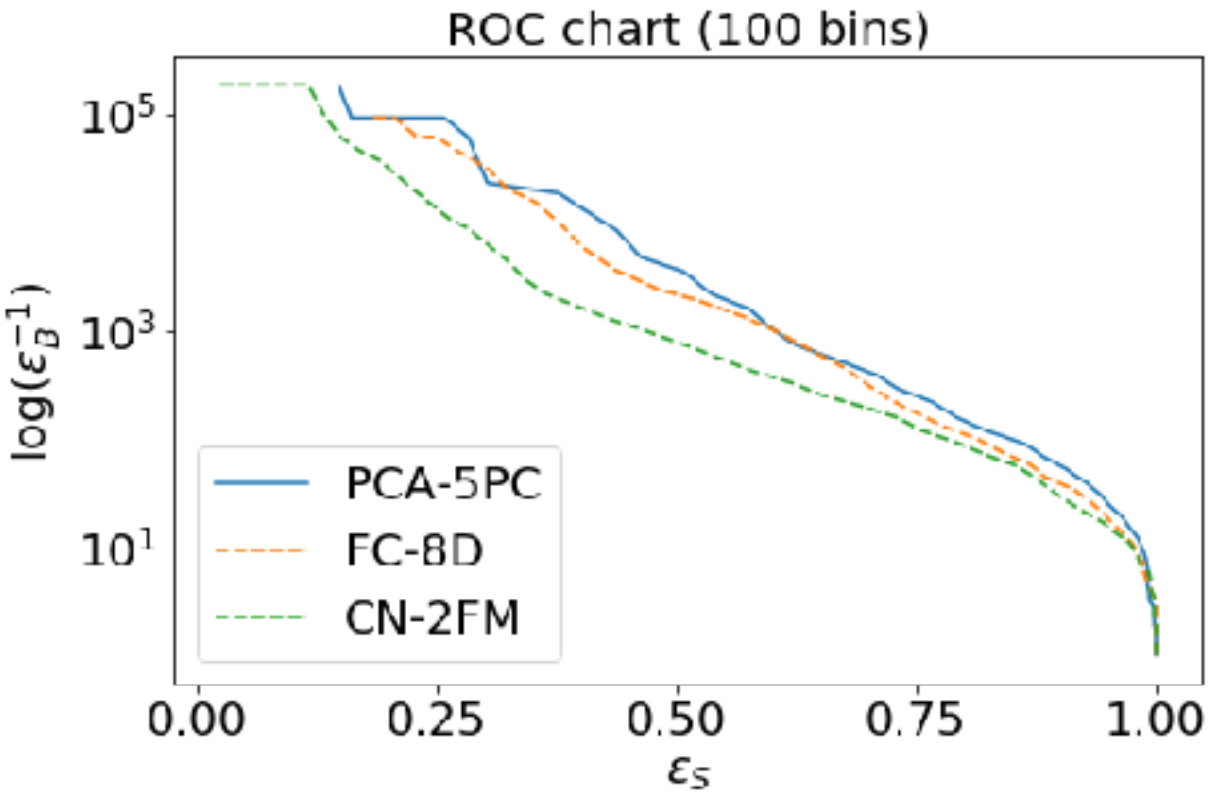


Convolutional Neural Networks

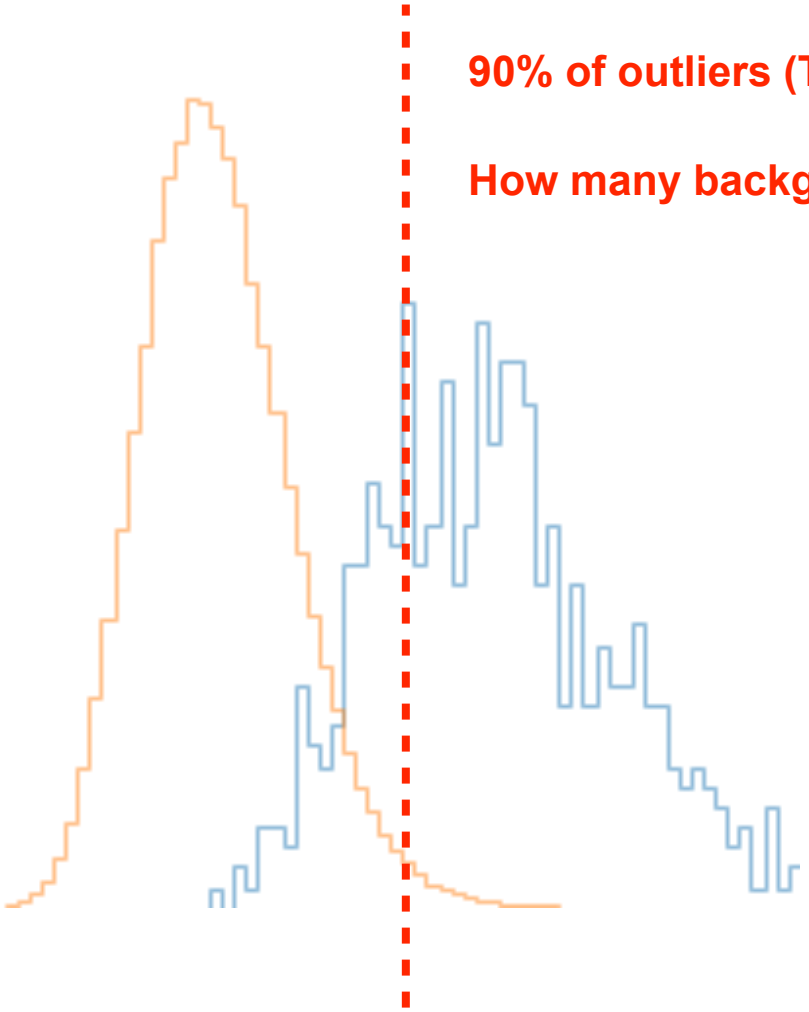


$$RE(i) = \frac{1}{N}\left(\sum_{j=1}^{N} [X_{rec_j}(i) - X_j(i)]^2\right)^{\frac{1}{2}}$$





PCA VS AEN	
PCA	**AEN**
linear approach	non-linear approach
low computing resources	low/high computing resources
less flexible	more flexible



90% of outliers (TP = 540)

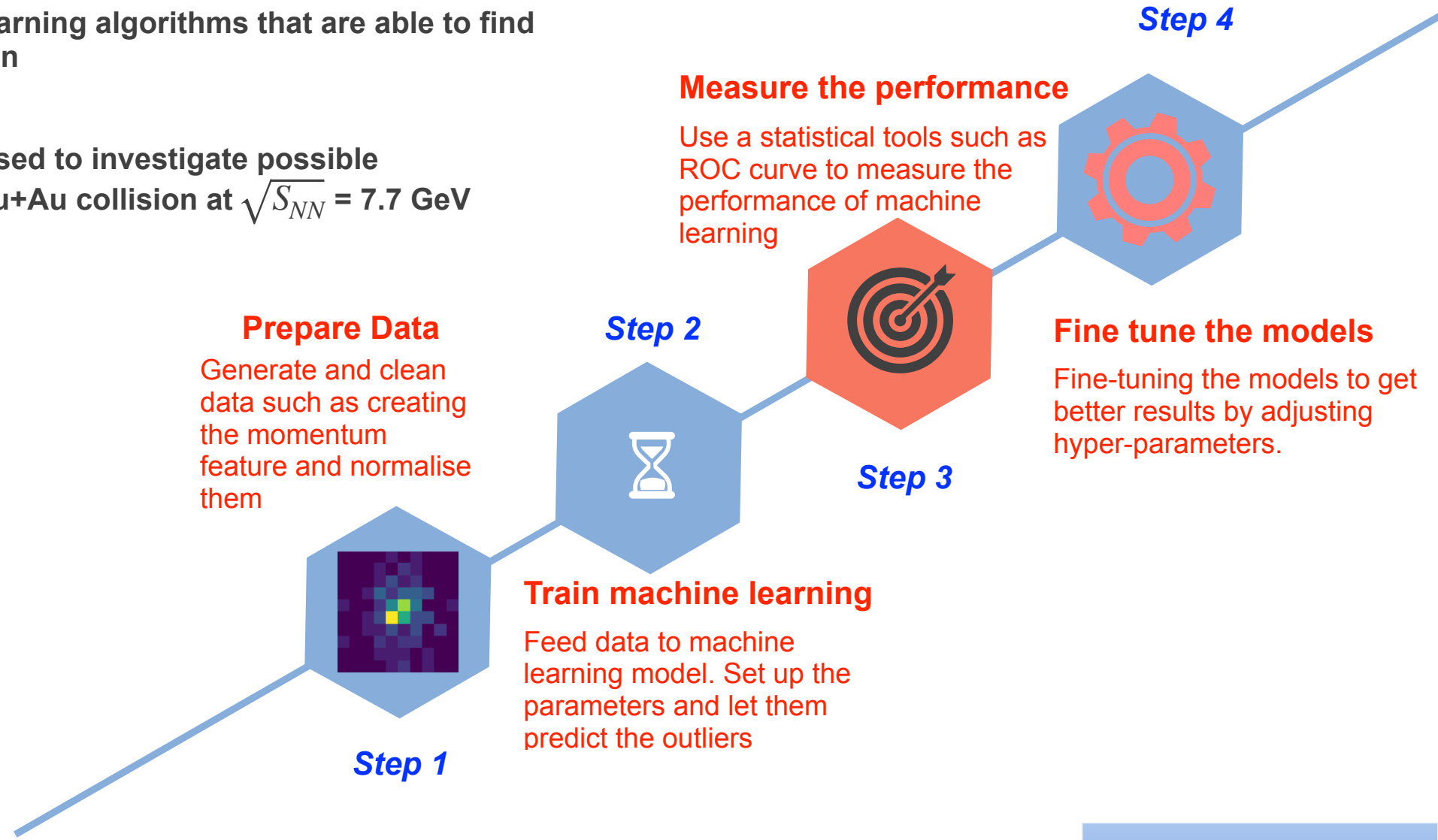
How many backgrounds mix in? (FP = ?%)

	100 bins (10x10)		400 bins (20x20)	
PCA	2 PC	3.5%	2 PC	1.9%
	3 PC	2.2%	3 PC	1.1%
	5 PC	1.6%	5 PC	0.8%
AEN	FC-2D	3.3%	FC-2D	1.8%
	FC-3D	2.5%	FC-3D	1.2%
	FC-8D	2.0%	FC-8D	0.9%
AEN	CN-1FM	4.7%	CN-1FM	3.4%
	CN-2FM	3.6%	CN-2FM	8.1%
	CN-4FM	4.2%	CN-4FM	1.4%

A lower % means better performance !!

Summary

- We developed machine learning algorithms that are able to find outliers in nuclear collision
- Such algorithms can be used to investigate possible interesting physics e.g. Au+Au collision at $\sqrt{S_{NN}}$ = 7.7 GeV



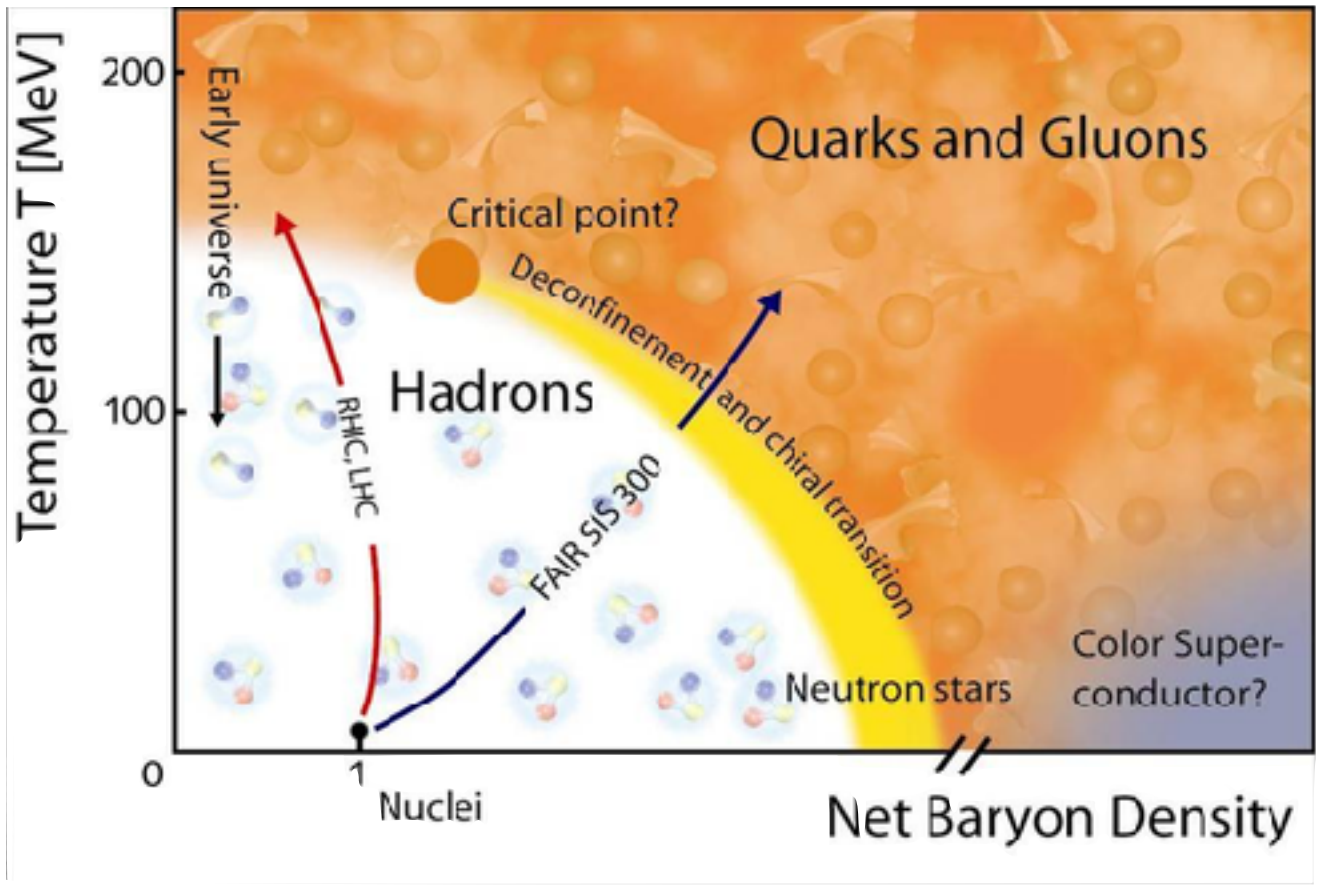
Conclusion

- **The principle components (or bottleneck component) keep both noise and significance, for both algorithms, we need to fine tune for a particular number of hyper-parameters to get the best results**

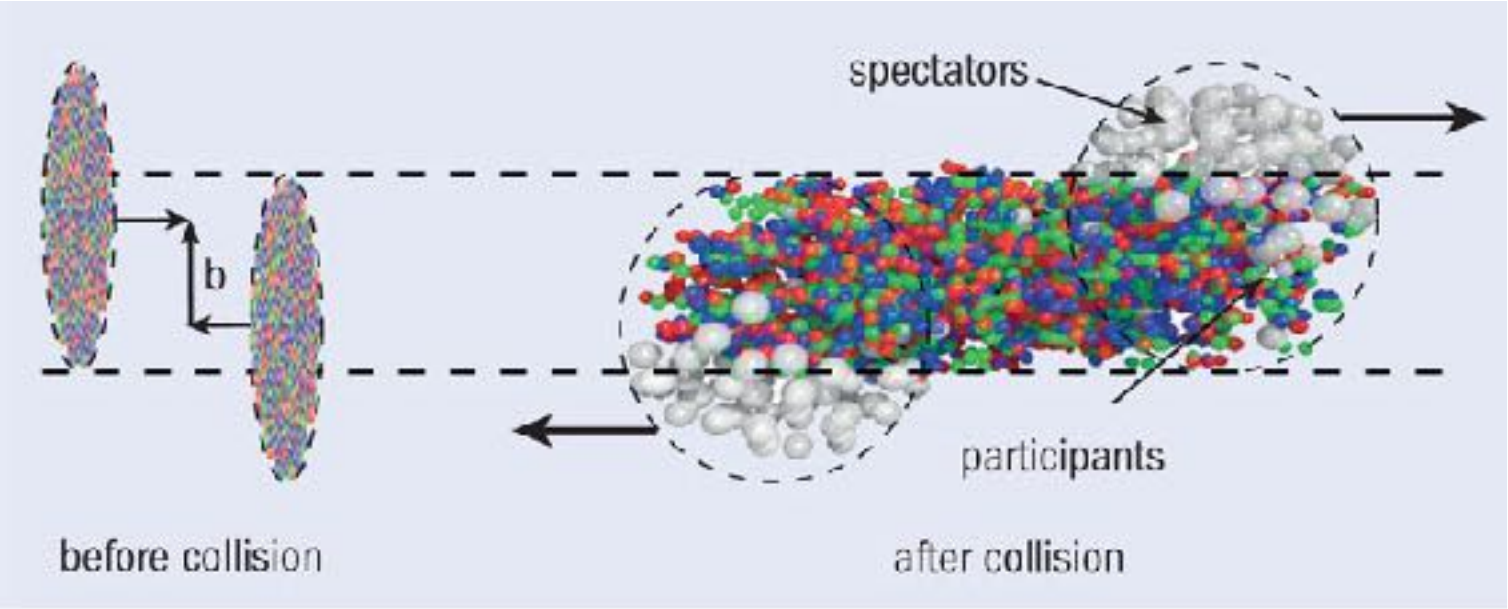
Acknowledgement

- Dr. Christoph Herold
- Dr. Jan Steinheimer
- Dr. Kai Zhou
- DPST scholarship
- FIAS

QUESTIONS

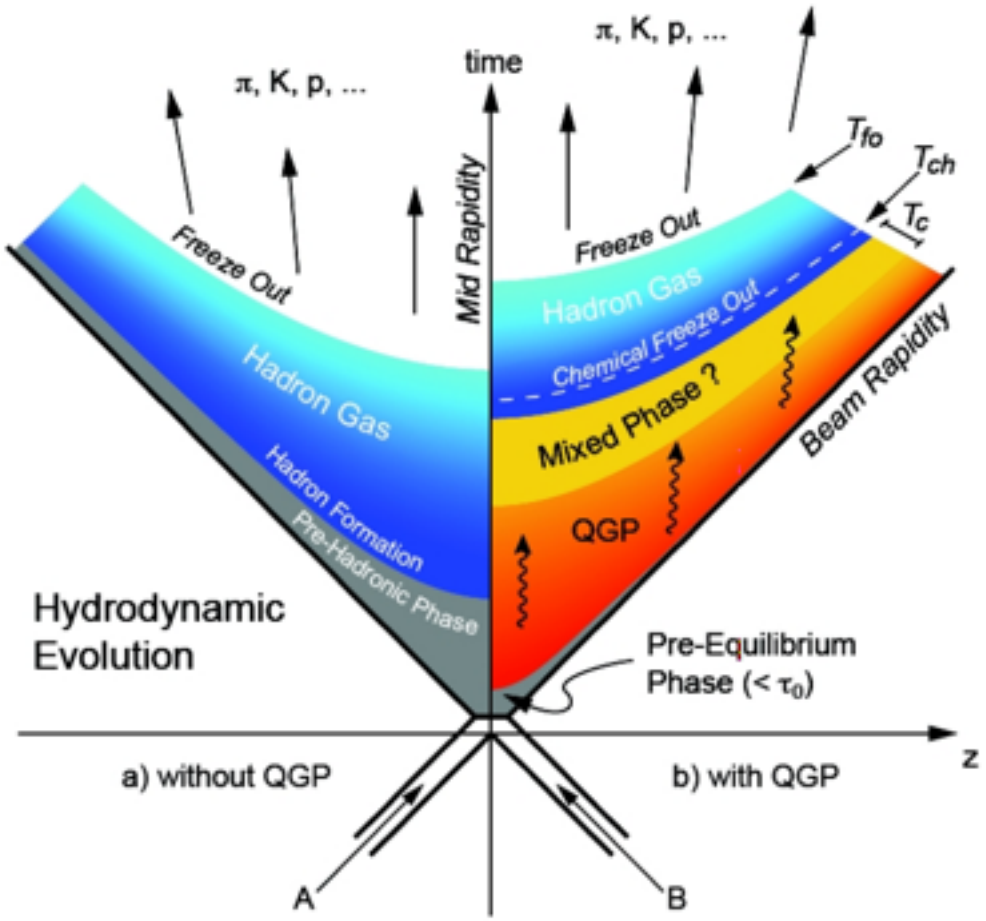


Ref: Bicudo, Pedro & Cardoso, M. & Cardoso, Nuno. (2011).
QCD confinement and chiral crossovers, two critical points.



Simple model of nuclei collision.
Two nuclei got the effect of Lorentz contraction along the beam axis

Ref: cerncourier



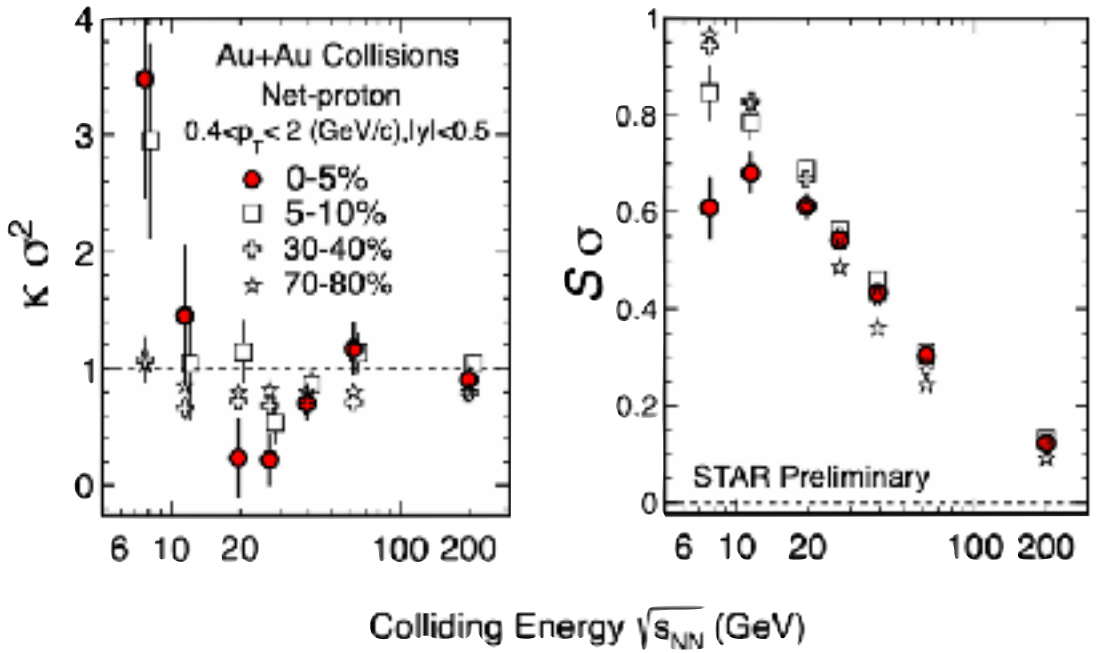
Ref: particlesandfriends

$t \lesssim 1$ fm/c: **Pre-equilibrium**

$t \sim 1/10$ fm/c: **Thermalization**

$t \sim 20$ fm/c: **Hadronization**

$t \gtrsim 20$ fm/c: **Thermal freeze-out**



Ref: Luo, X. (2015b). Energy dependence of moments of net-proton and net-charge multiplicity distribution at star

Efficiency Corrected Cumulant Ratios

$$\kappa\sigma^2 = \frac{C_4}{C_2} \qquad S\sigma = \frac{C_3}{C_2}$$

$$m = N_p - N_{\bar{p}}$$

$$C_1 = < m >$$

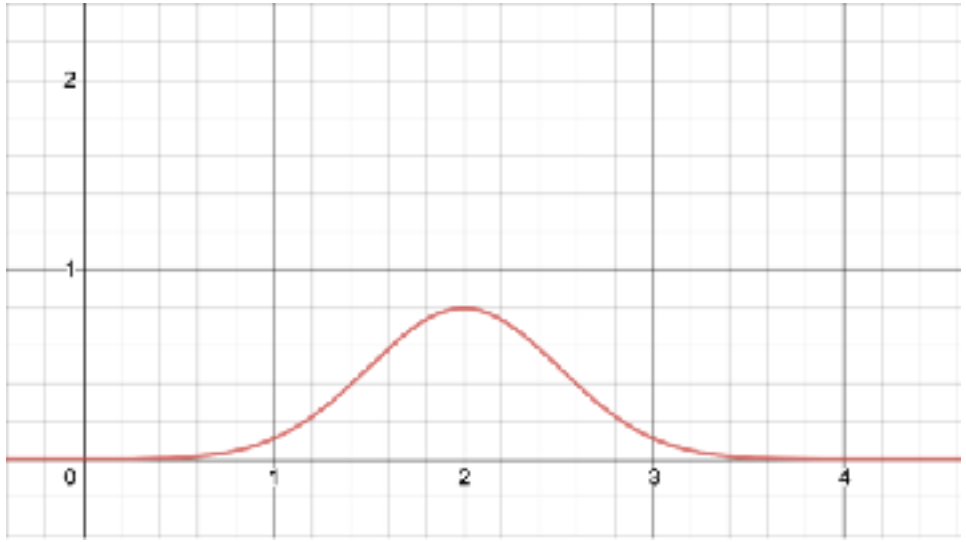
$$C_2 = < (m - < m >)^2 >$$

$$C_3 = < (m - < m >)^3 >$$

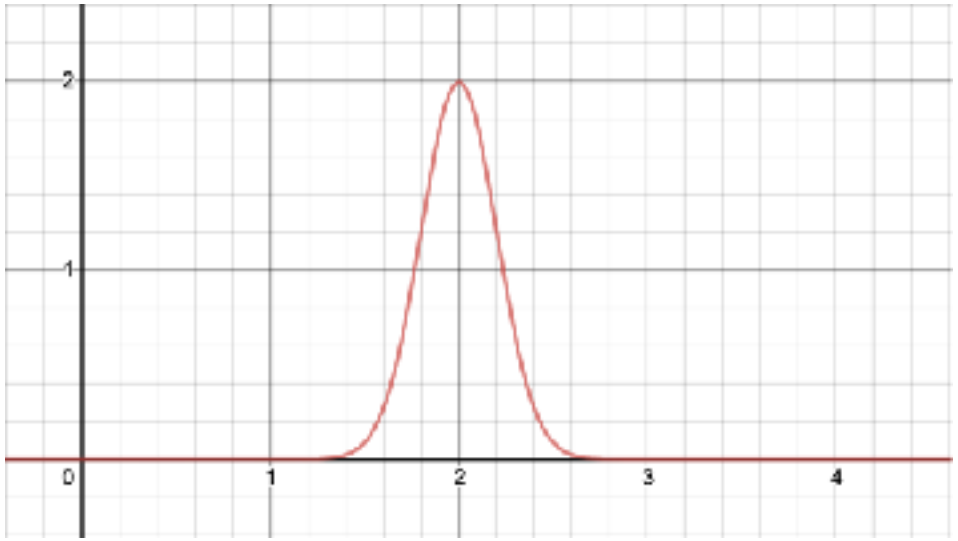
$$C_4 = < (m - < m >)^4 > - 3 < (m - < m >)^2 >^2$$

Gaussian function

$$f(x) = \frac{e^{-\frac{(x-\mu)}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

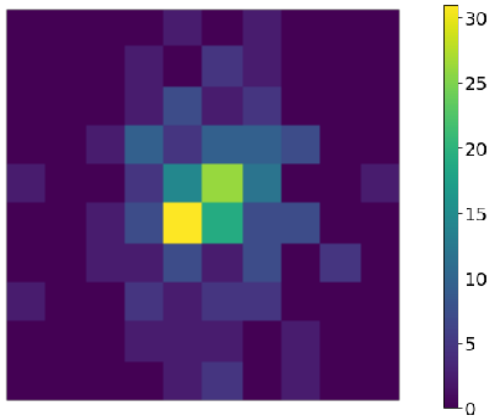


$\mu = 2$ $\sigma = 0.4$

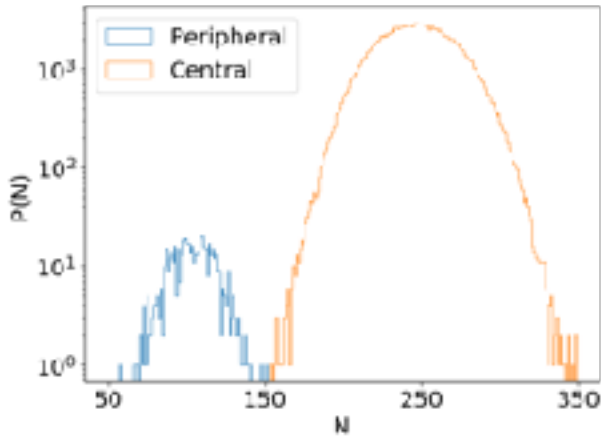


$\mu = 2$ $\sigma = 0.2$

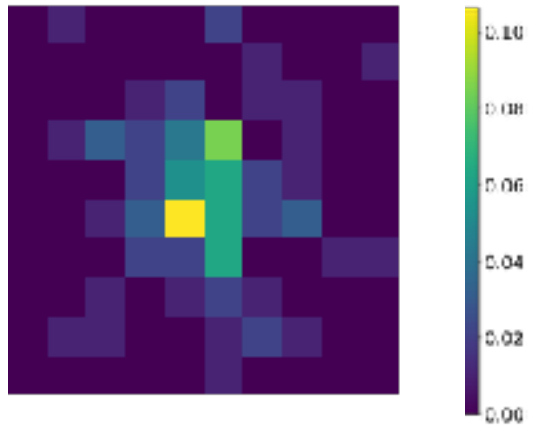
Momentum Feature



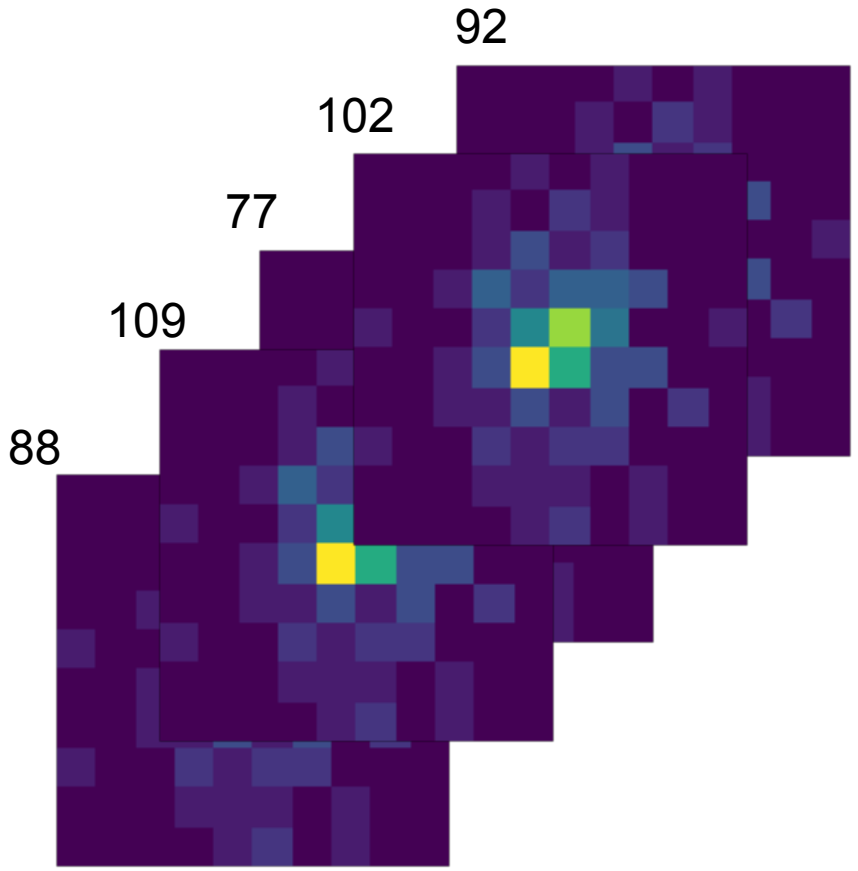
Histogram of Charged Particle



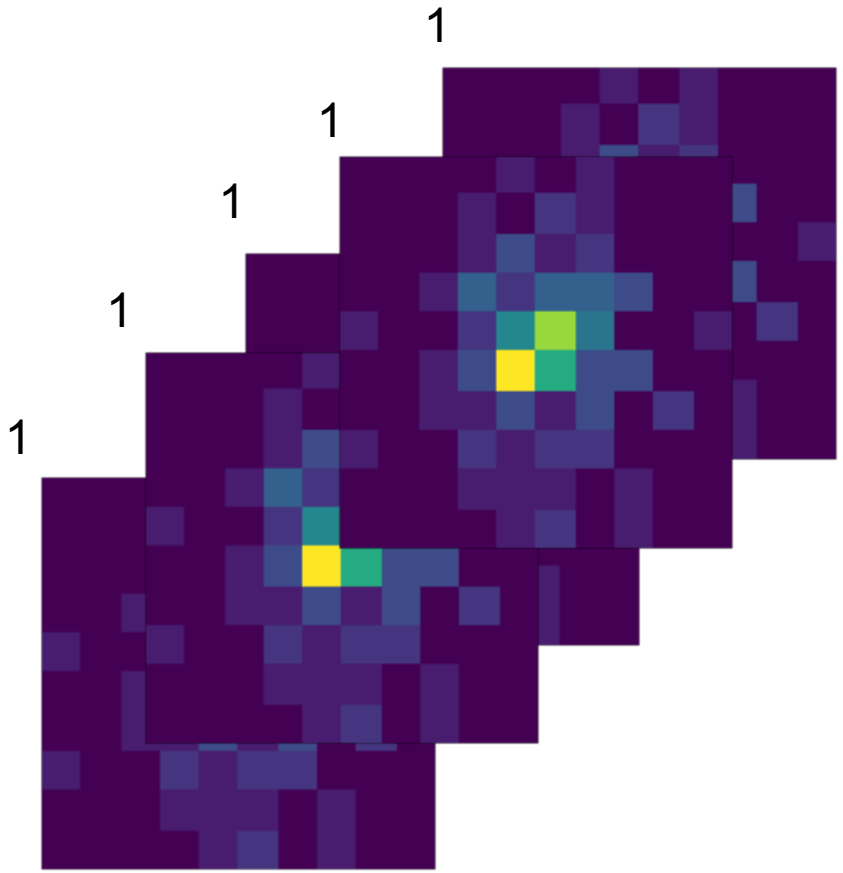
Normalized Momentum Feature

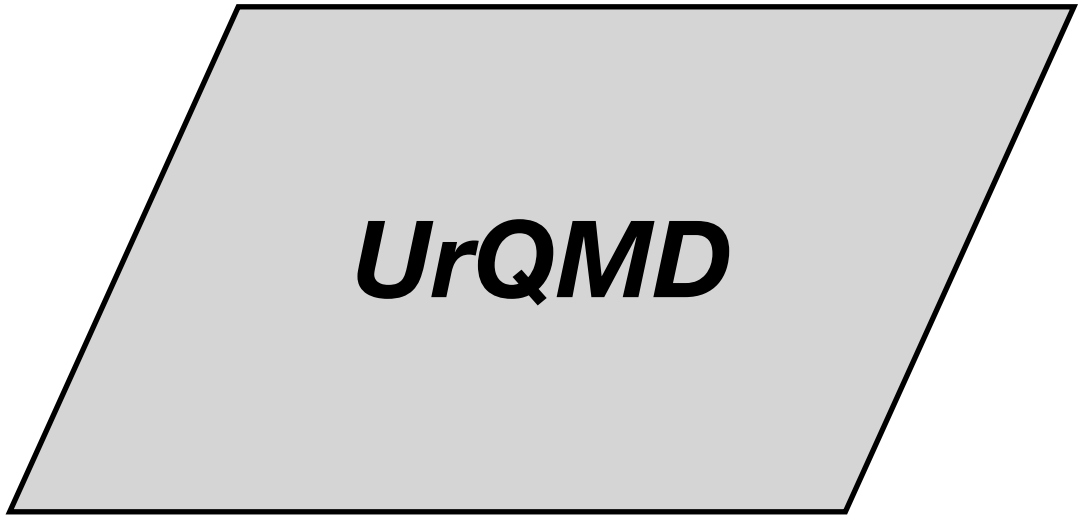


Non-normalized



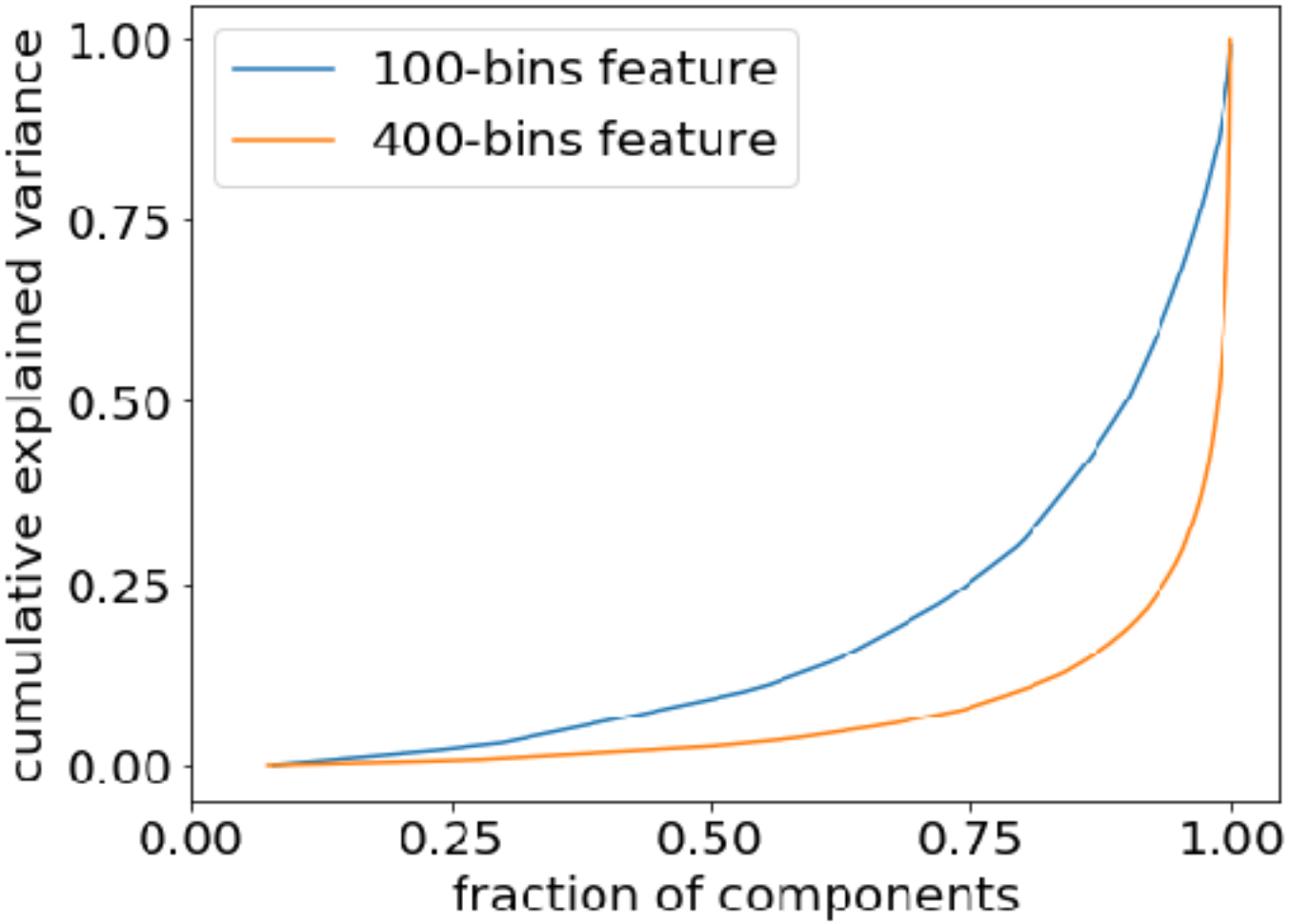
Normalized



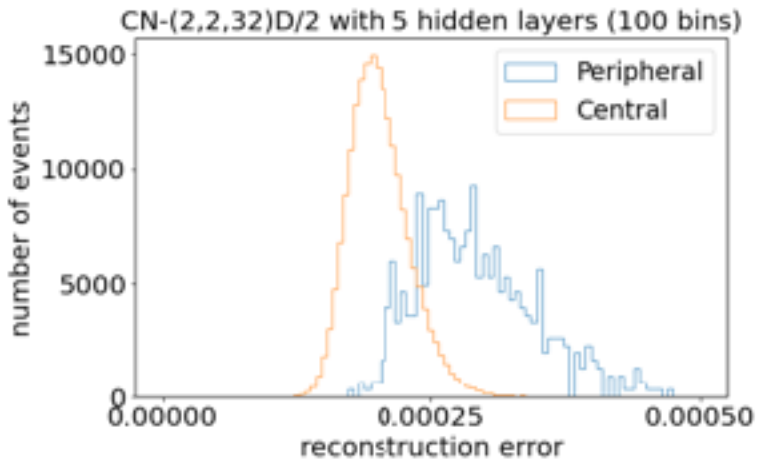


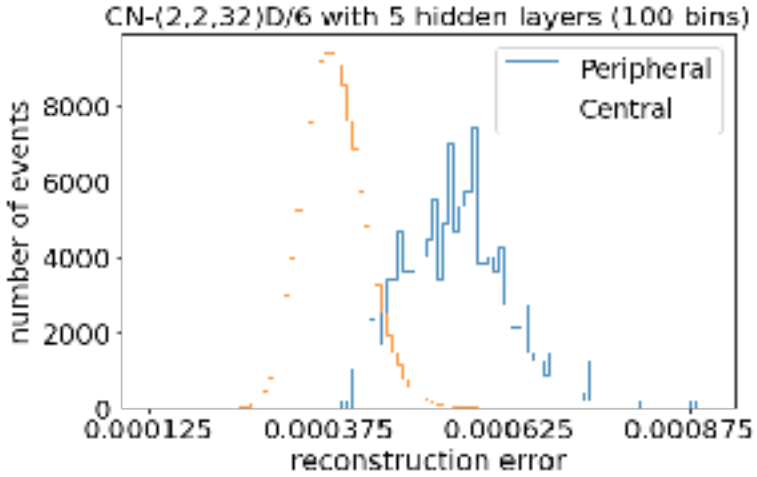
Ultra-relativistic Quantum Molecular Dynamics

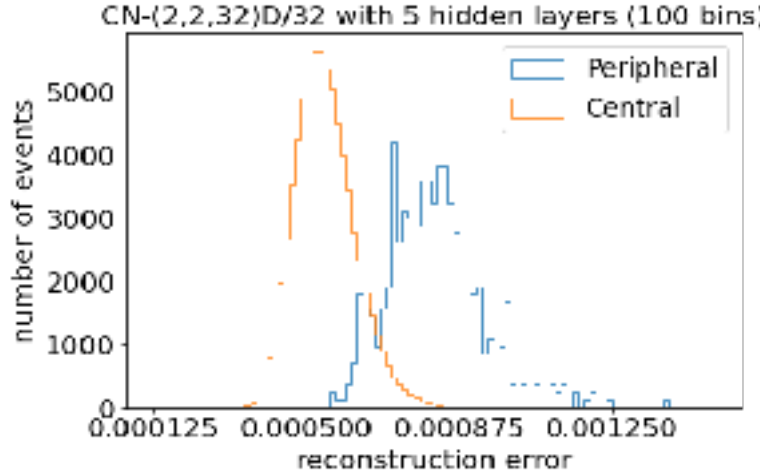
It is a simulation package based on
Monte Carlo solution of
Boltzmann transport equations.

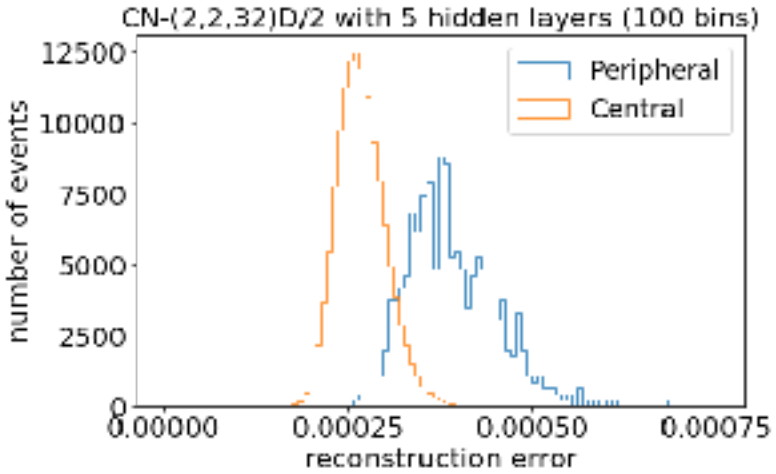


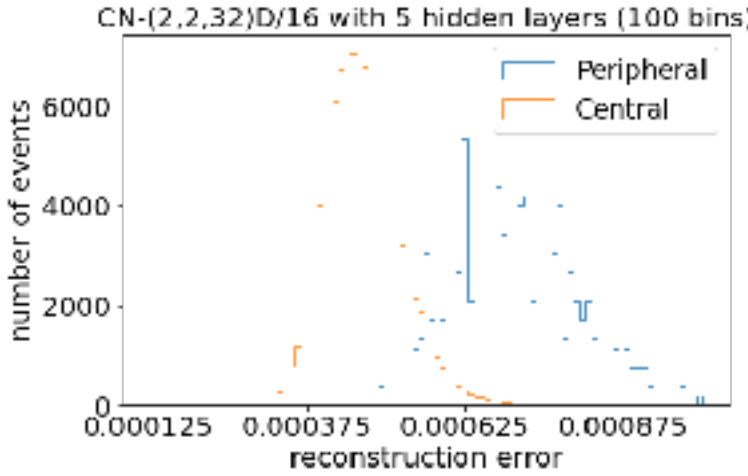
$$\sigma^2_{cev} = \frac{\sum_{i=1}^{l} \sigma^2_{PC_i}}{\sum_{i=1}^{l_{max}} \sigma^2_{PC_i}}$$

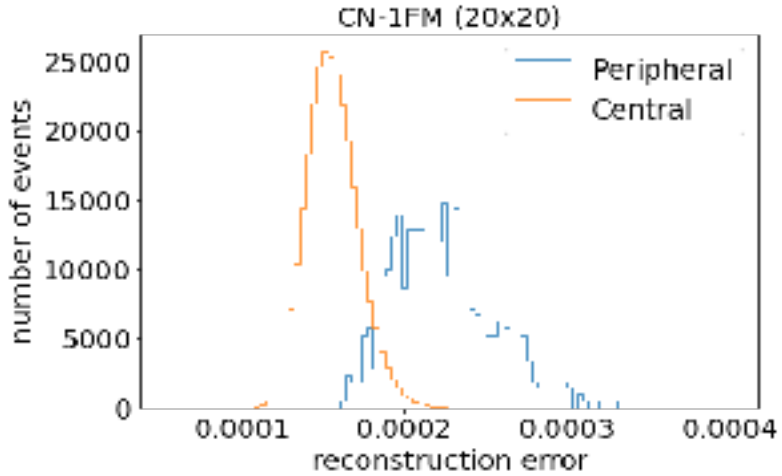

CN-(2,2,32)D/2 with 5 hidden layers (100 bins)
Peripheral
Central
number of events
reconstruction error
0.00000
0.00025
0.00050
0
5000
10000
15000

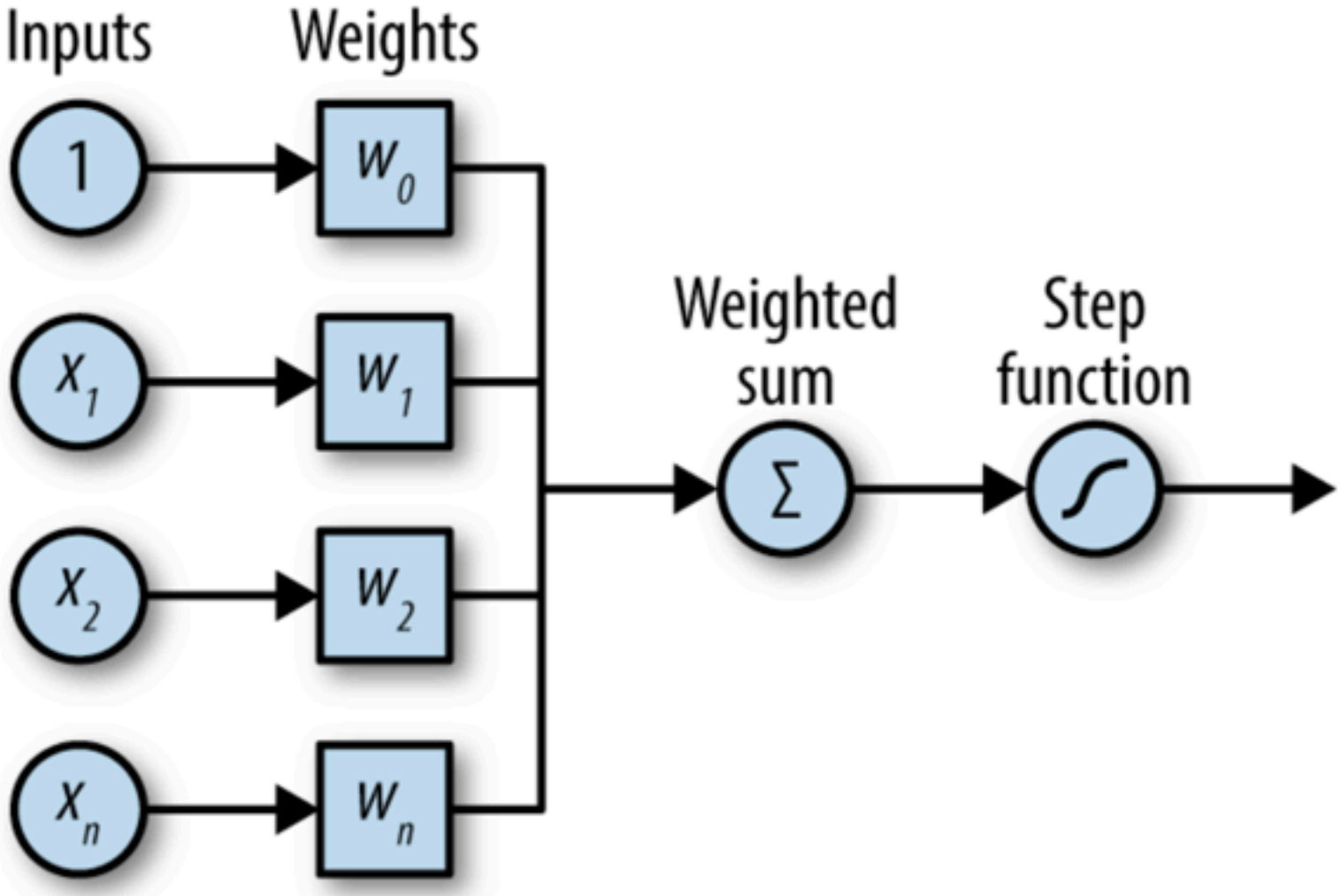

CN-(2,2,32)D/6 with 5 hidden layers (100 bins)
Peripheral
Central
number of events
reconstruction error
0.000125
0.000375
0.000625
0.000875
0
2000
4000
6000
8000

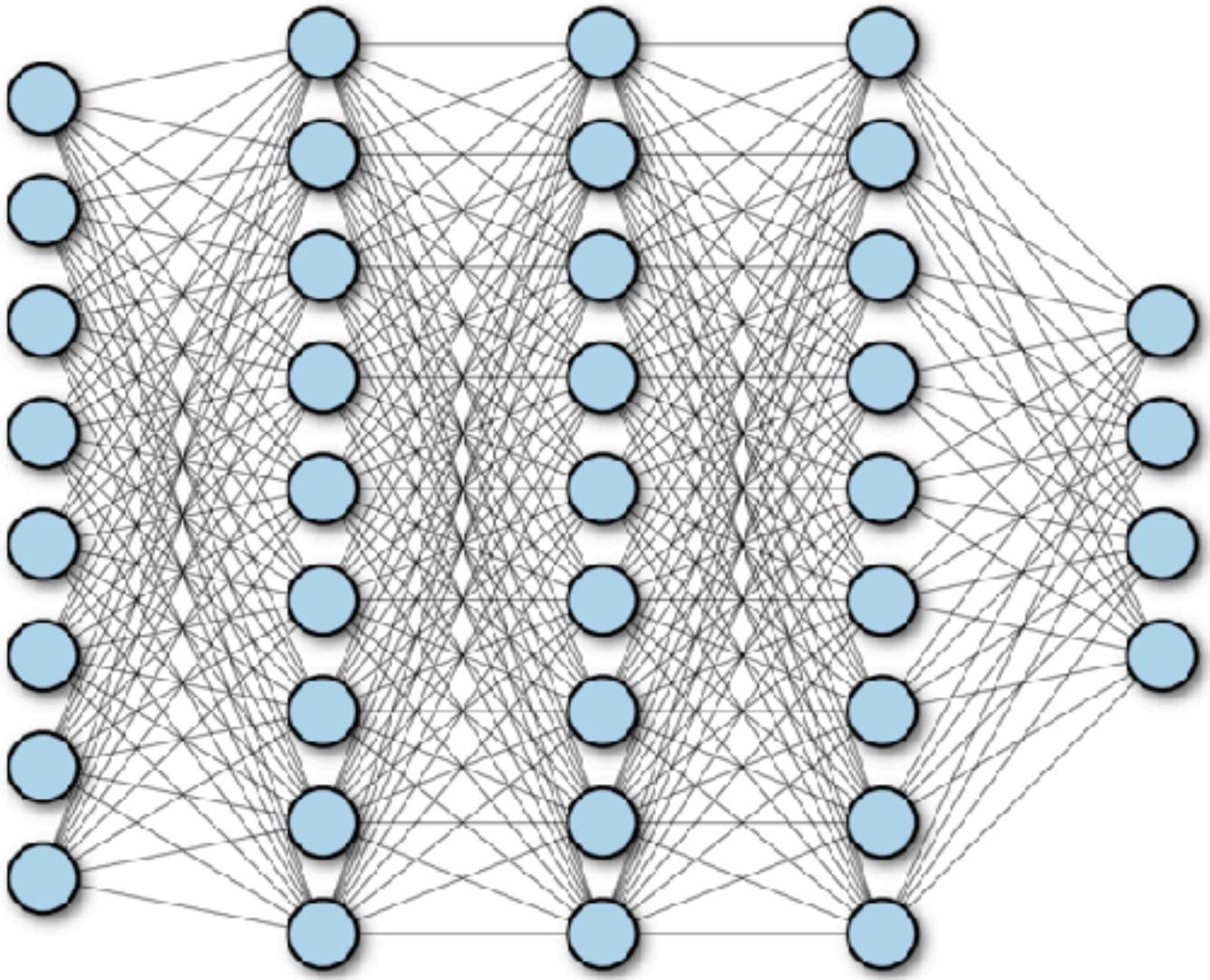

CN-(2,2,32)D/32 with 5 hidden layers (100 bins)
Peripheral
Central
number of events
reconstruction error
0.000125
0.000500
0.000875
0.001250
0
1000
2000
3000
4000
5000

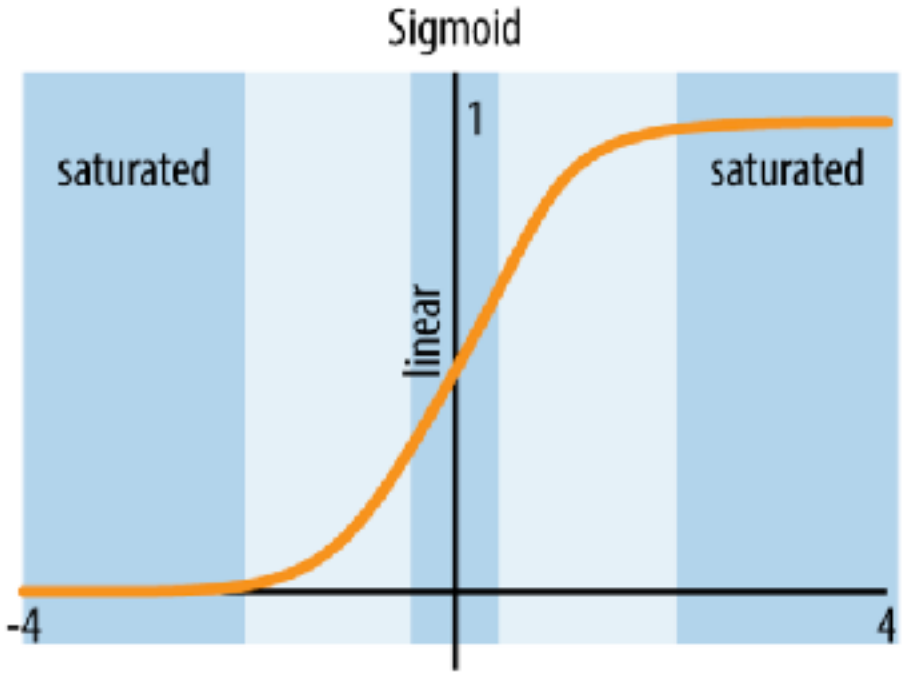

CN-(2,2,32)D/2 with 5 hidden layers (100 bins)
Peripheral
Central
number of events
reconstruction error
0.00000
0.00025
0.00050
0.00075
0
2500
5000
7500
10000
12500

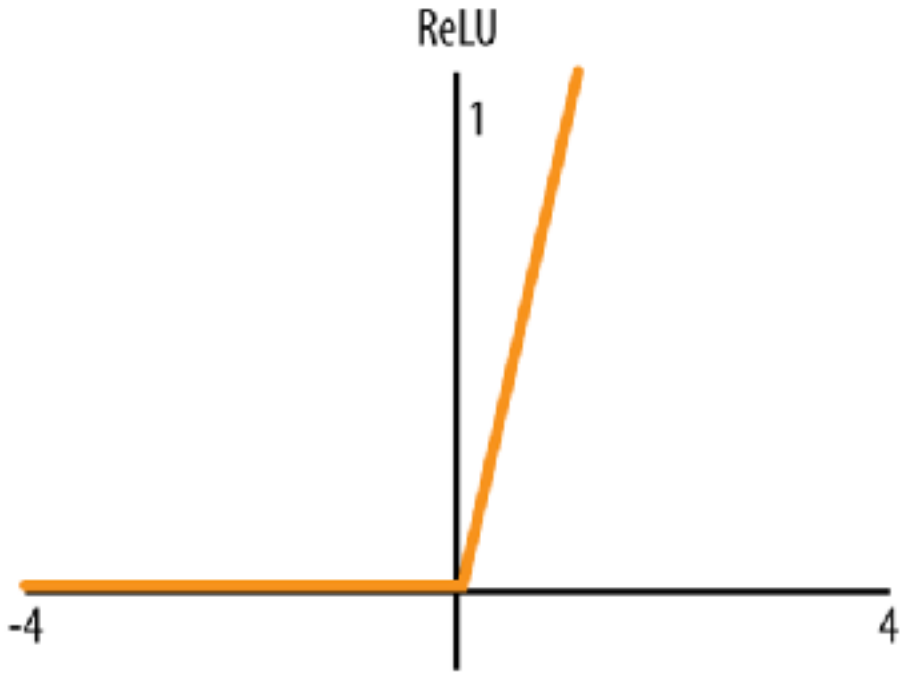

CN-(2,2,32)D/16 with 5 hidden layers (100 bins)
Peripheral
Central
number of events
reconstruction error
0.000125
0.000375
0.000625
0.000875
0
2000
4000
6000

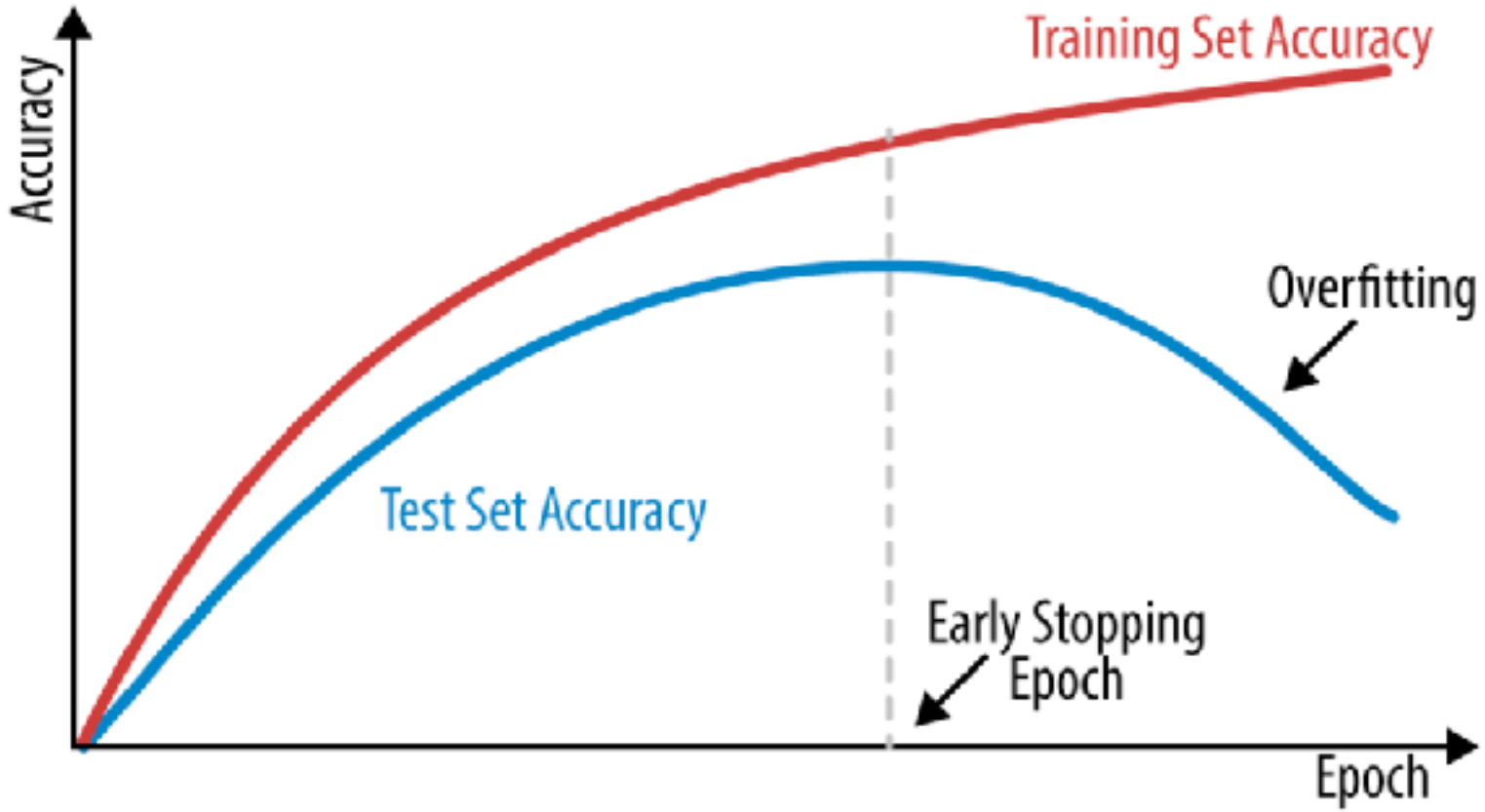

CN-1FM (20x20)
Peripheral
Central
number of events
reconstruction error
0.0001
0.0002
0.0003
0.0004
0
5000
10000
15000
20000
25000


Inputs
Weights
1
x_1
x_2
x_n
w_0
w_1
w_2
w_n
Weighted sum
Σ
Step function




Sigmoid
saturated
saturated
linear
1
-4
4


ReLU
1
-4
4


Training Set Accuracy
Accuracy
Overfitting
Test Set Accuracy
Early Stopping
Epoch
Epoch