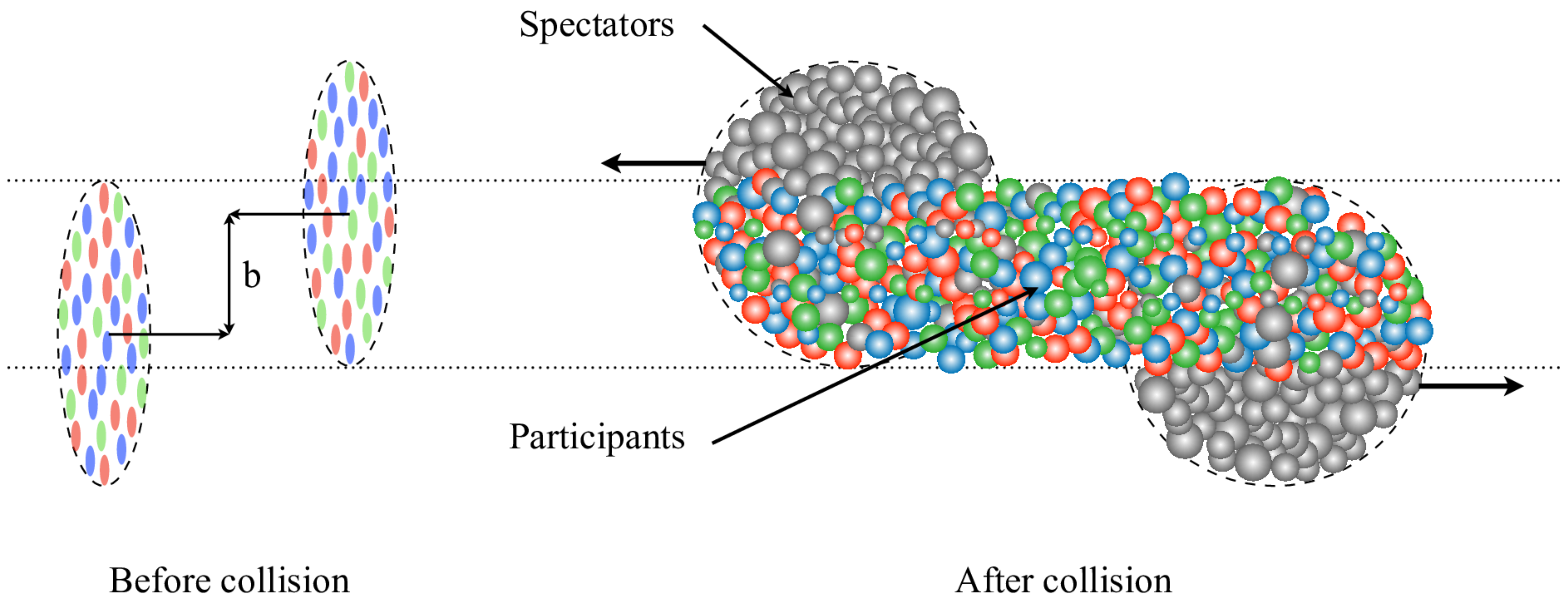
The use of new methods for processing data of a physical experiment. Application of machine learning methods on the NICA complex.

Centrality estimation in nucleus-nucleus collisions by machine learning algorithms

EVGENY ANDRONOV, VLADIMIR KOVALENKO, ANDREY SERYAKOV 28/08/2023 ST. PETERSBURG

Motivation

- In experiment one cannot strictly control initial conditions of a collision
- This leads to an inevitable ‘trivial’ contribution to all fluctuation measures which are of high importance in relativistic nuclear physics
- Typically, one groups events in the so-called centrality classes based on a model approximation of experimental data using one or another observable
- «Ideal» geometrical estimator of centrality is an impact parameter b:



from LHCb coll., *JINST* 17 (2022) P05009

$$c_b = \frac{1}{\sigma_{inel}} \int_0^b \frac{d\sigma}{db'} db'$$

centrality percentile

Goals of the study

- Centrality in real life - one has to map distribution of the centrality estimator to «ideal» measure (e.g. impact parameter)
- Main objective - improve the «quality» of centrality selection (by applying ML) in comparison to the standard methods

- Principal ingredients of the study - energy deposition in the modules of the Projectile Spectator Detector of the NA61/SHINE experiment
- Investigation of energy deposition «geometry» should provide additional information in comparison to the total energy deposited in the calorimeter
- Based on simulations of Li+Be, Ar+Sc and Pb+Pb collisions

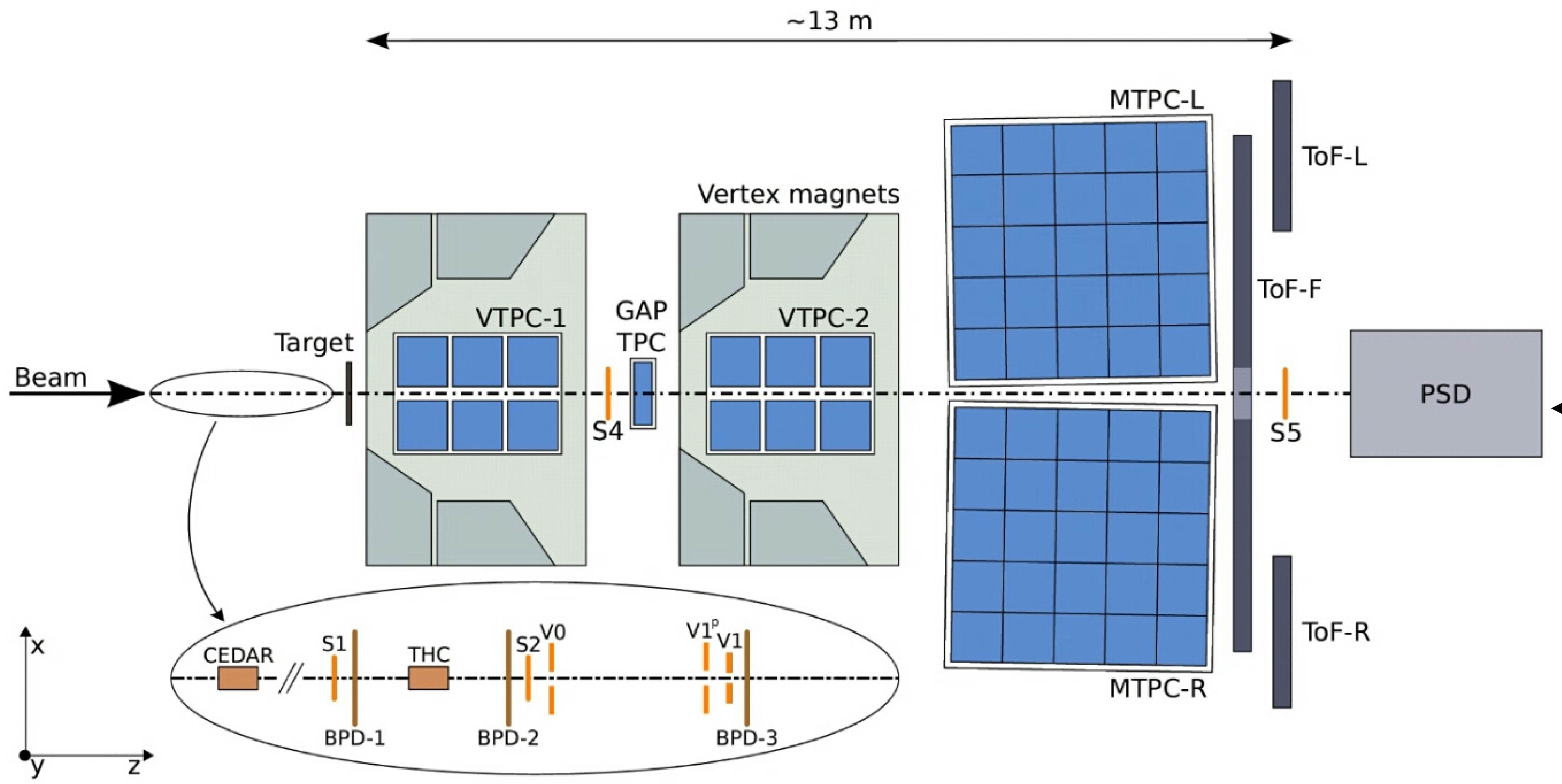
- with participation of former SPbU students: D. Uzhva, A. Zharov

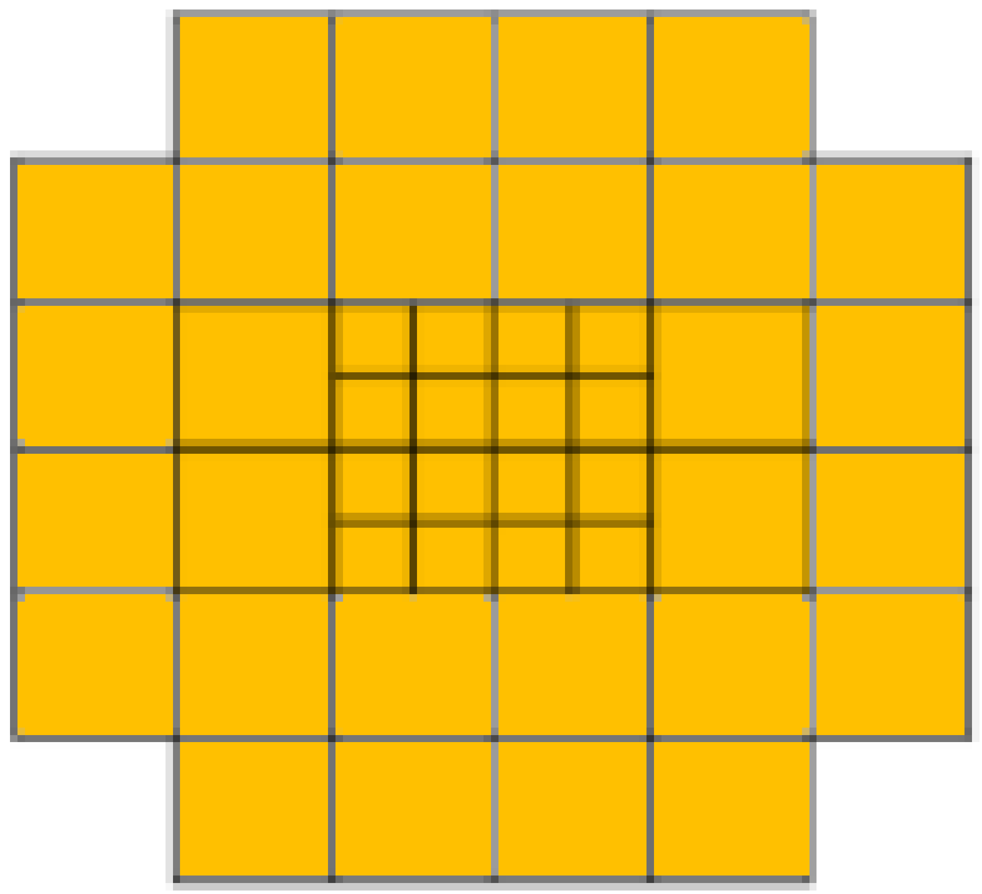
- Some additional details can be found in
  - A. Seryakov, D. Uzhva, Phys.Part.Nucl. 51 (2020) 3, 331-336
  - E. Andronov, report at «Nucleus-2021» (20-25 Sept 2021)

Review

- Rather hot topic: useful for NA61, MPD, BM@N and higher coll. energy experiments 🔥 🔥 🔥
- Broad discussion in the literature (follow «A living review of machine learning for particle physics»):
  - N. Mallick et al., *Phys.Rev.D* 103 (2021) 9, 094031 - dN/deta and <pT> as features for BDT to extract b
  - P. Xiang et al., *Chin.Phys.C* 46 (2022) 7, 074110 - pT of all particles in event for CNN and DNN to extract b
  - A. Saha et al., *Phys.Rev.C* 106 (2022) 1, 014901 - pT spectra for BDT, kNN to extract b and eccentricity
  - D. Basak, K. Dey, *Eur.Phys.J.A* 59 (2023) 7, 174 - pT-eta-phi spectra for CNN and DNN to extract Npart
  - N. Karpushkin et al., *Phys.Part.Nucl.* 53 (2022) 2, 524-530 - calorimeter response for NN and autoencoder to extract b
- Methods from the first 4 references are not easily applicable when you study fluctuations of some observable in your «central» detector (e.g. multiplicity fluctuations measured with TPC) as they introduce autocorrelation bias
- In this case it is preferable to work with forward detectors

Projectile spectator detector of NA61/SHINE





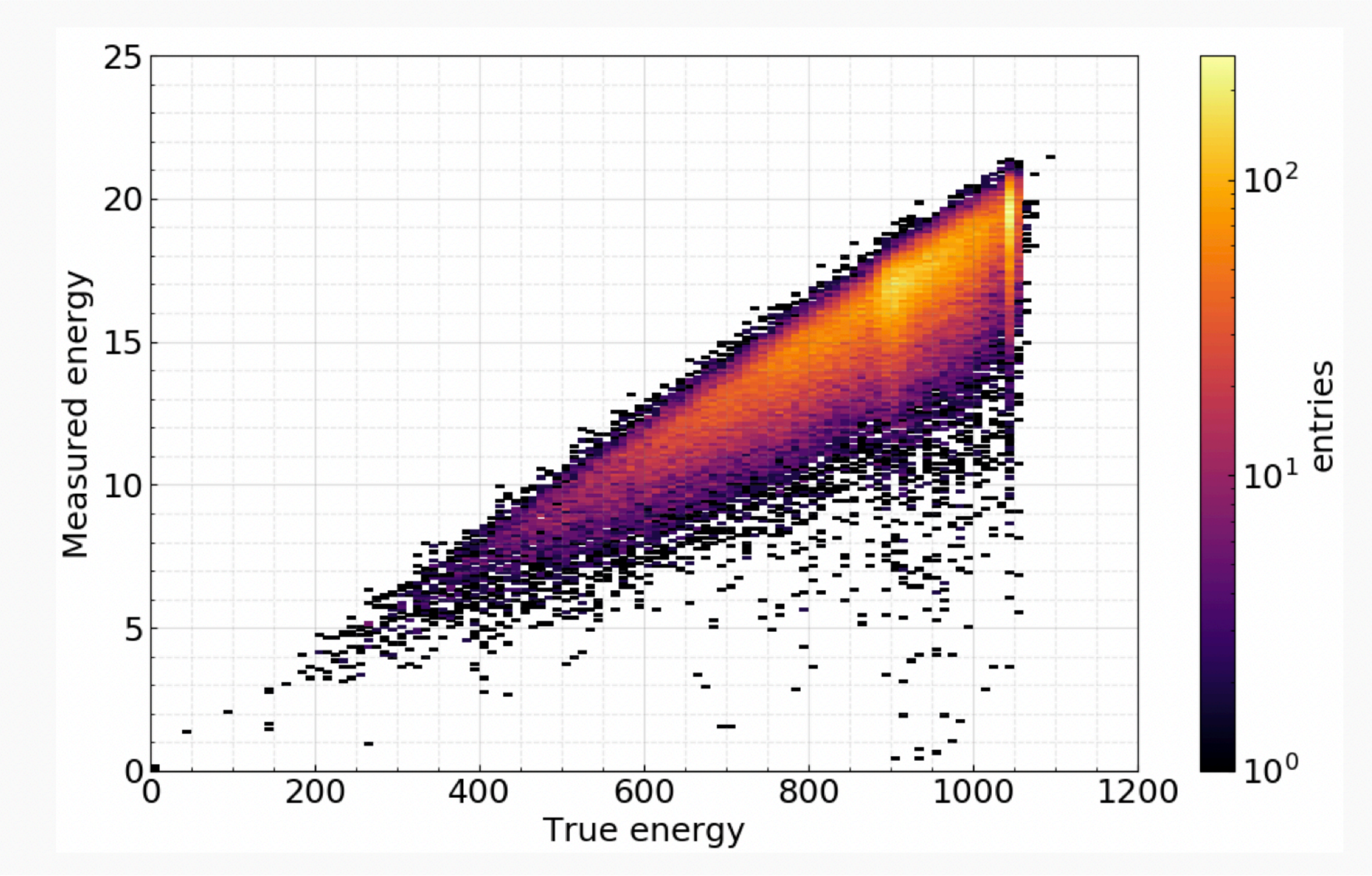
- 16 central 10*10 cm2; 28 peripheral 20*20 cm2
- 10 sections in deep
- potentially 440 features
- similar structure of calorimeters in BM@N, MPD

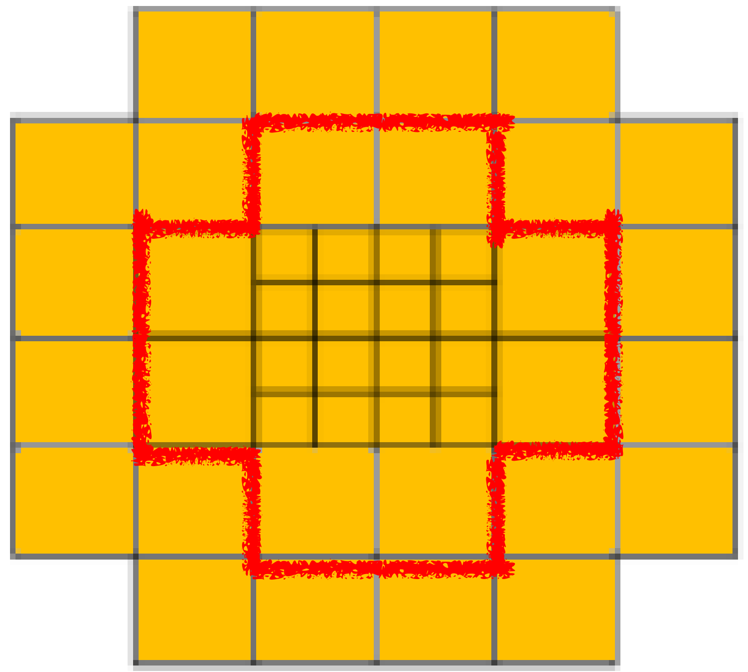
NA61/SHINE coll., *JINST* 9 (2014) P06005

Case №1 - Li7+Be9@150A GeV/c

SHIELD MC model with GEANT4 simulations of PSD response

A. Dementyev and N. Sobolevsky, http:// www.inr.troitsk.ru/shield/introd-eng.html

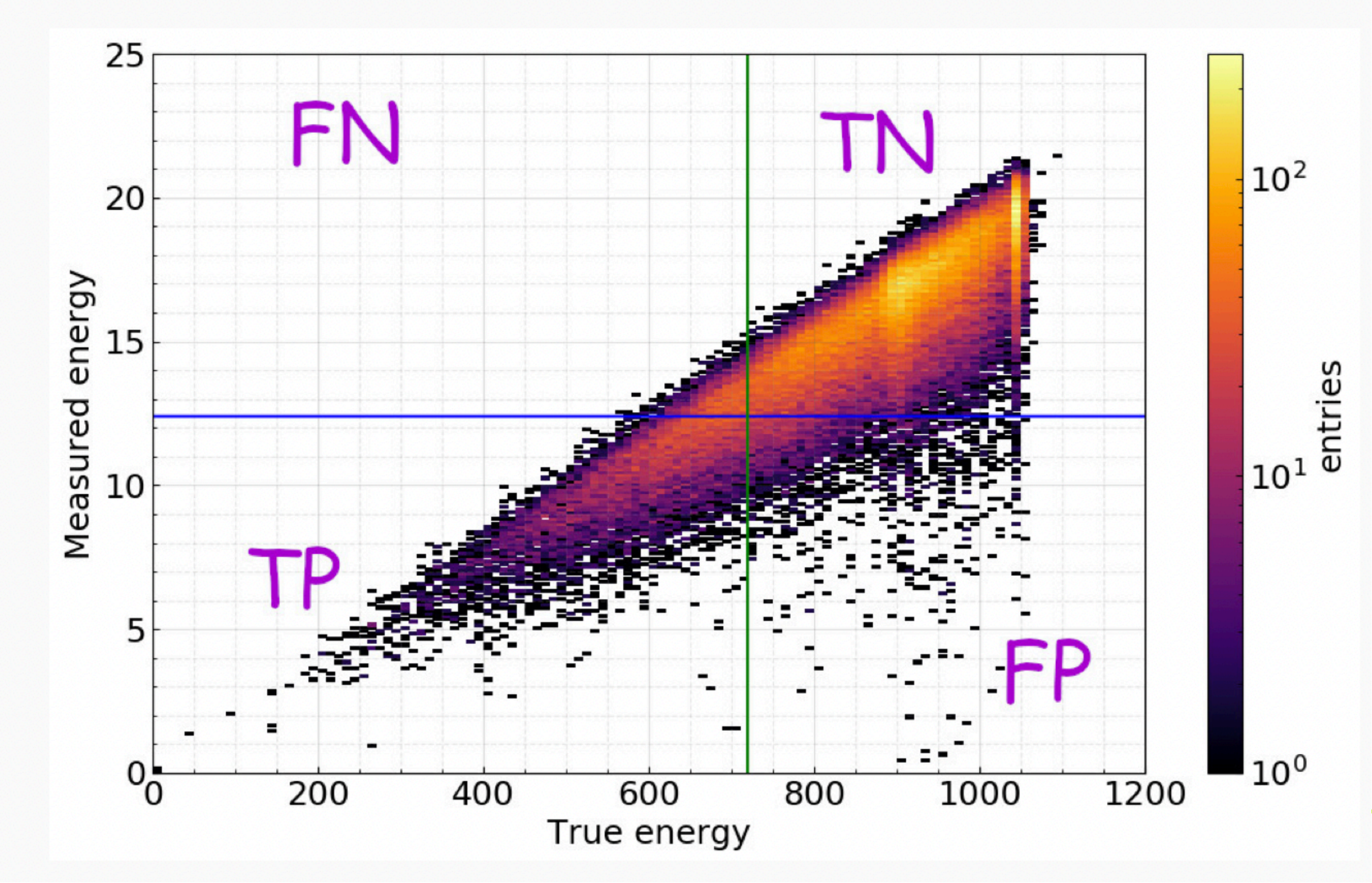


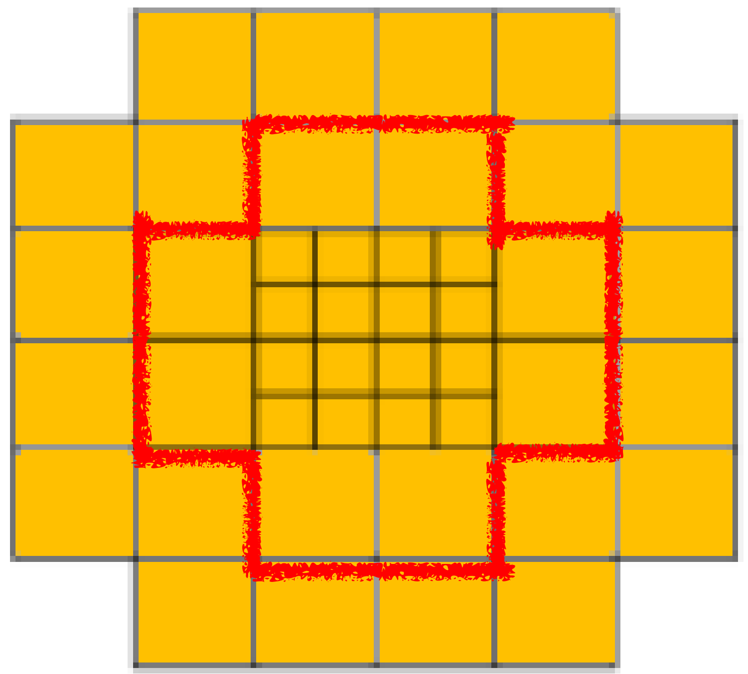


Case №1 - Li7+Be9@150A GeV/c

Standard method to select 15% of most central events would lead to accuracy

$$\varepsilon = \frac{TP + TN}{TP + TN + FP + FN} \approx 0.93$$

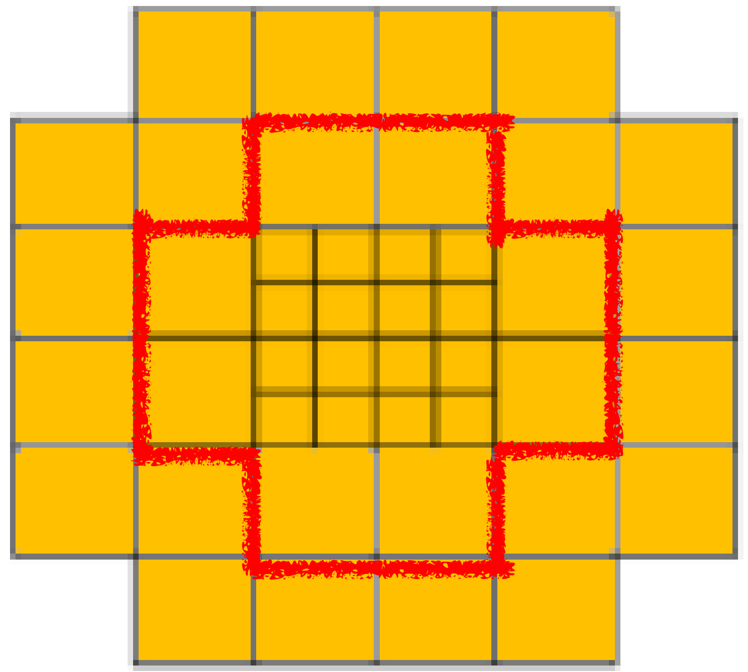


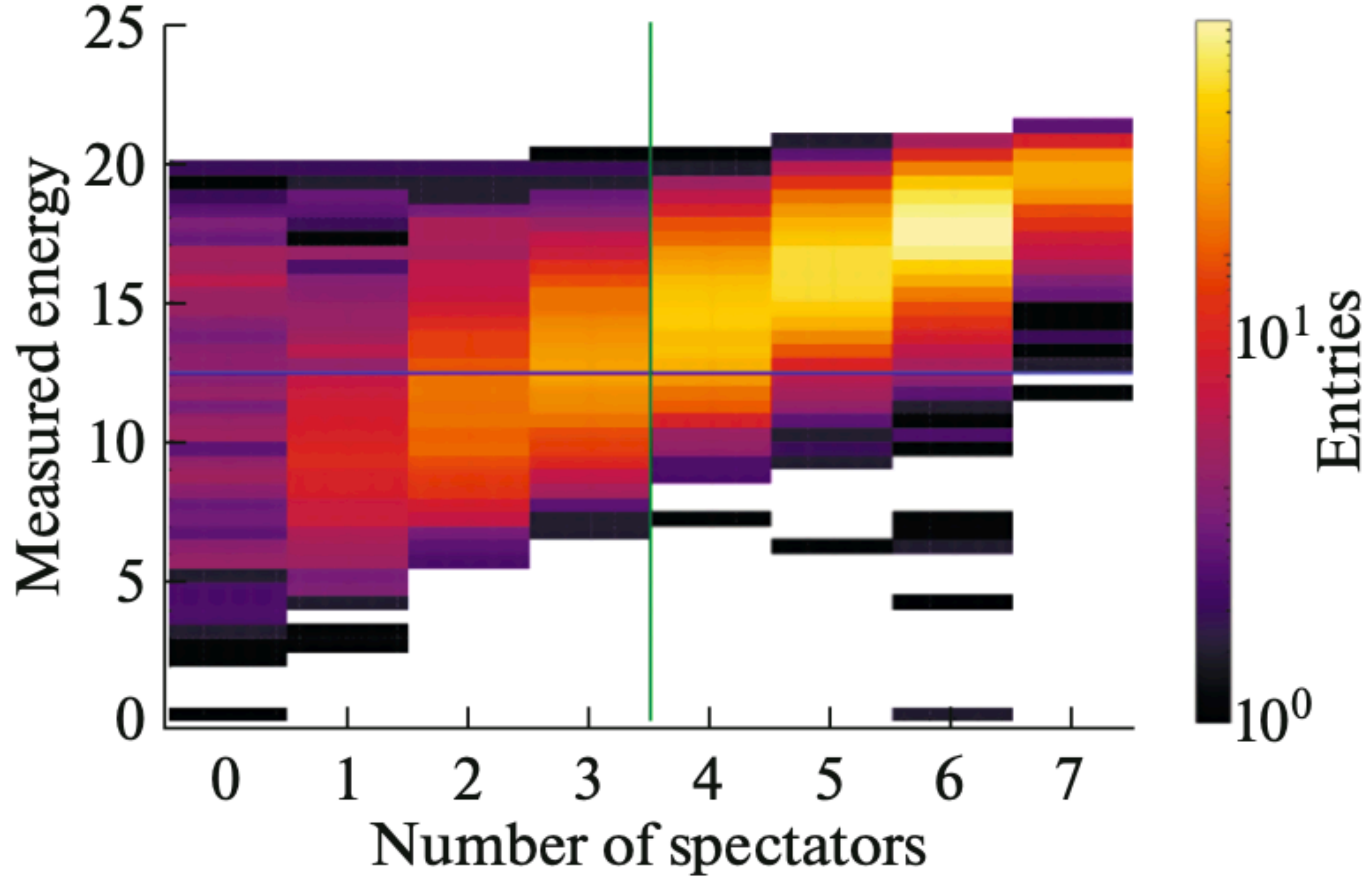


Case №1 - Li7+Be9@150A GeV/c

The same cut-based procedure applied to the number of nucleon spectators (instead of the forward energy)

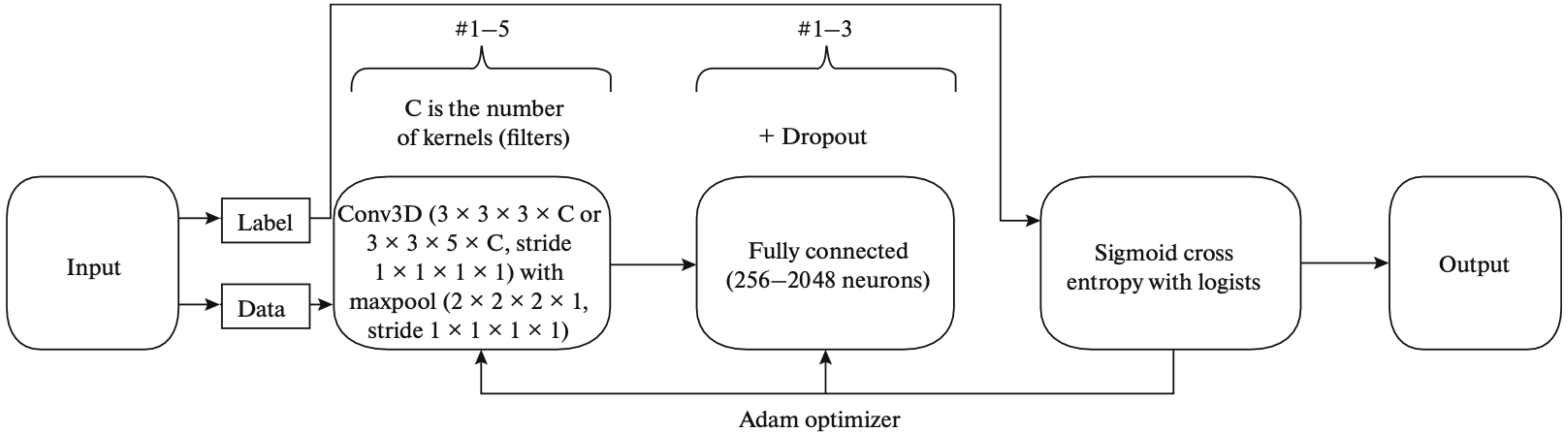
leads to accuracy $\varepsilon = \dfrac{TP + TN}{TP + TN + FP + FN} \approx 0.867$





Case №1 - Li7+Be9@150A GeV/c

- Two CNNs were trained on this dataset: one was trained on the true forward energy, another - on the number of nucleon-spectators



- CNN application slightly improved the accuracy of centrality class selection:

	E_{true}	N_{spec}
Cut-based	93.0%	86.7%
CNN	93.7%	92.8%

Case №1 - Li7+Be9@150A GeV/c

- How can event selection influence the «final» results:

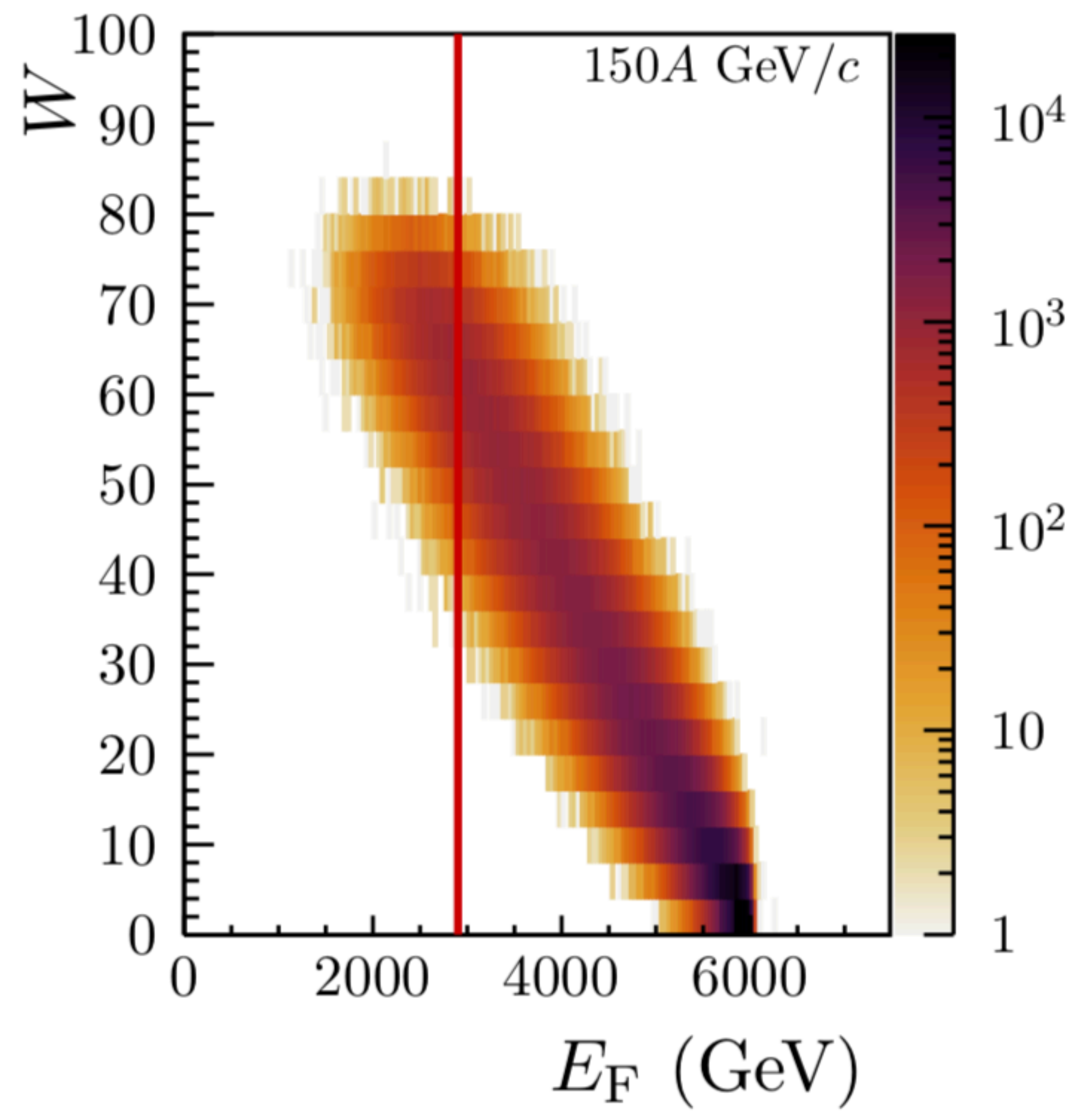
Centrality based on forward energy		
	$\langle N \rangle$	$\omega\,[N]$
Ideal case	19.59	1.88
Cut based method	18.56	2.65
CNN	18.69	2.49
Centrality based on number of forward spectators		
	$\langle N \rangle$	$\omega\,[N]$
Ideal case	15.69	3.67
Cut based method	18.56	2.65
CNN	16.36	3.30

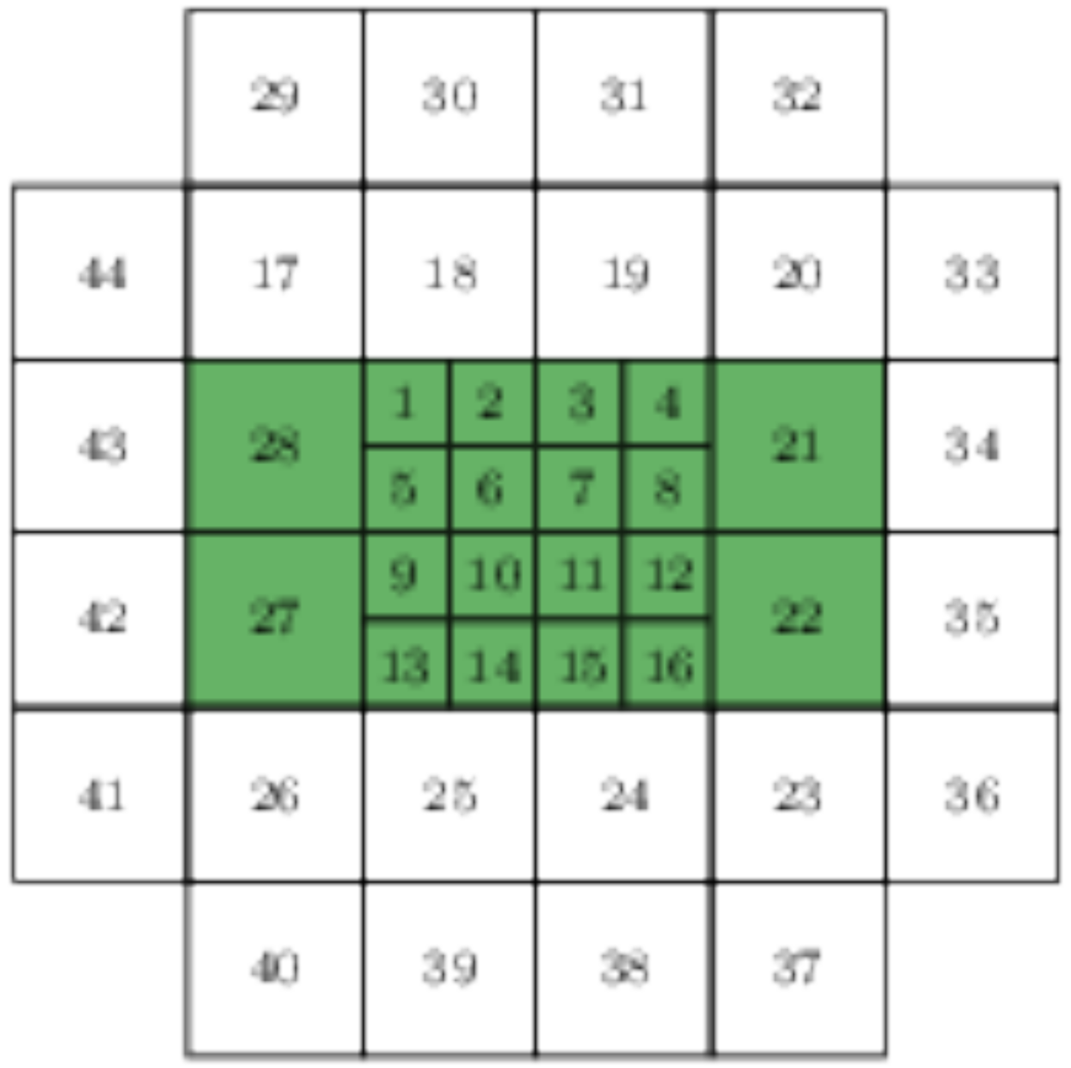
$\langle N \rangle$ average particle multiplicity

$$\omega[N] = \frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle}$$ scaled variance

Case №2 - Ar40+Sc45@150A GeV/c

EPOS1.99 MC model with GEANT4 simulations of PSD response





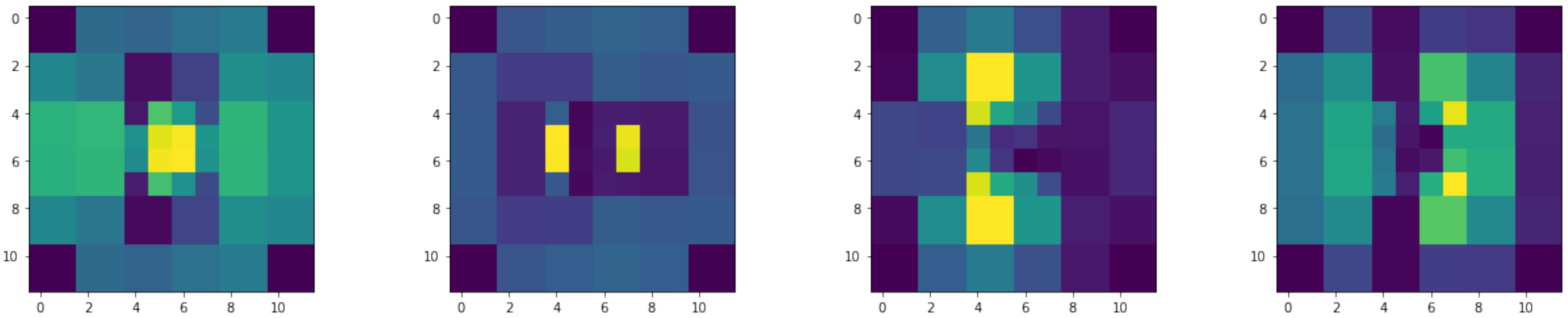
For 6% most central:

$$TPR = \frac{TP}{P} = 0.8918 \quad FPR = \frac{FP}{N} = 0.007$$

$$Prec = \frac{TP}{TP + FP} = 0.8921$$

Case №2 - Ar40+Sc45@150A GeV/c

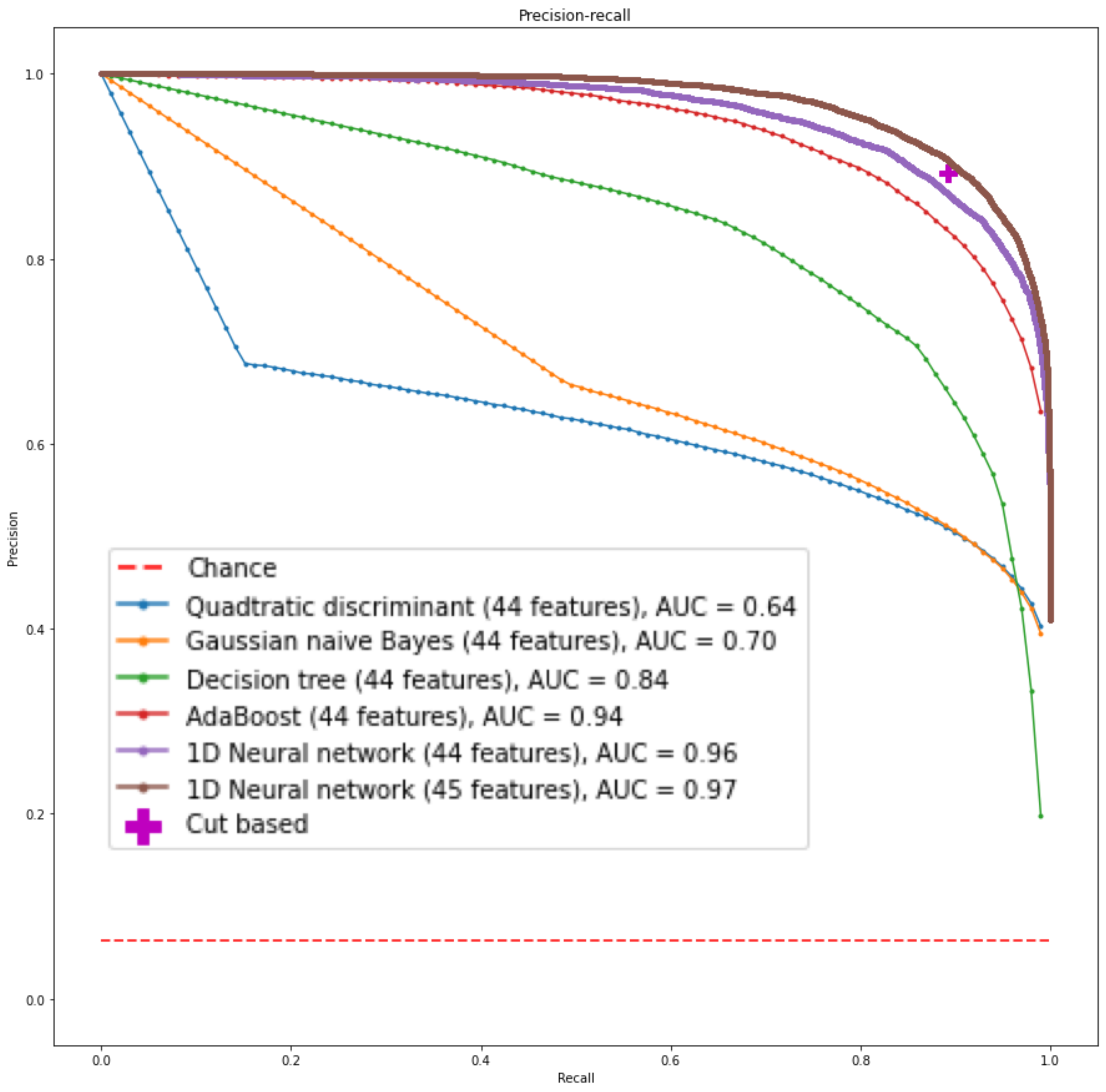
- To understand significance of different modules we performed principal component analysis for a set of all 44 modules:

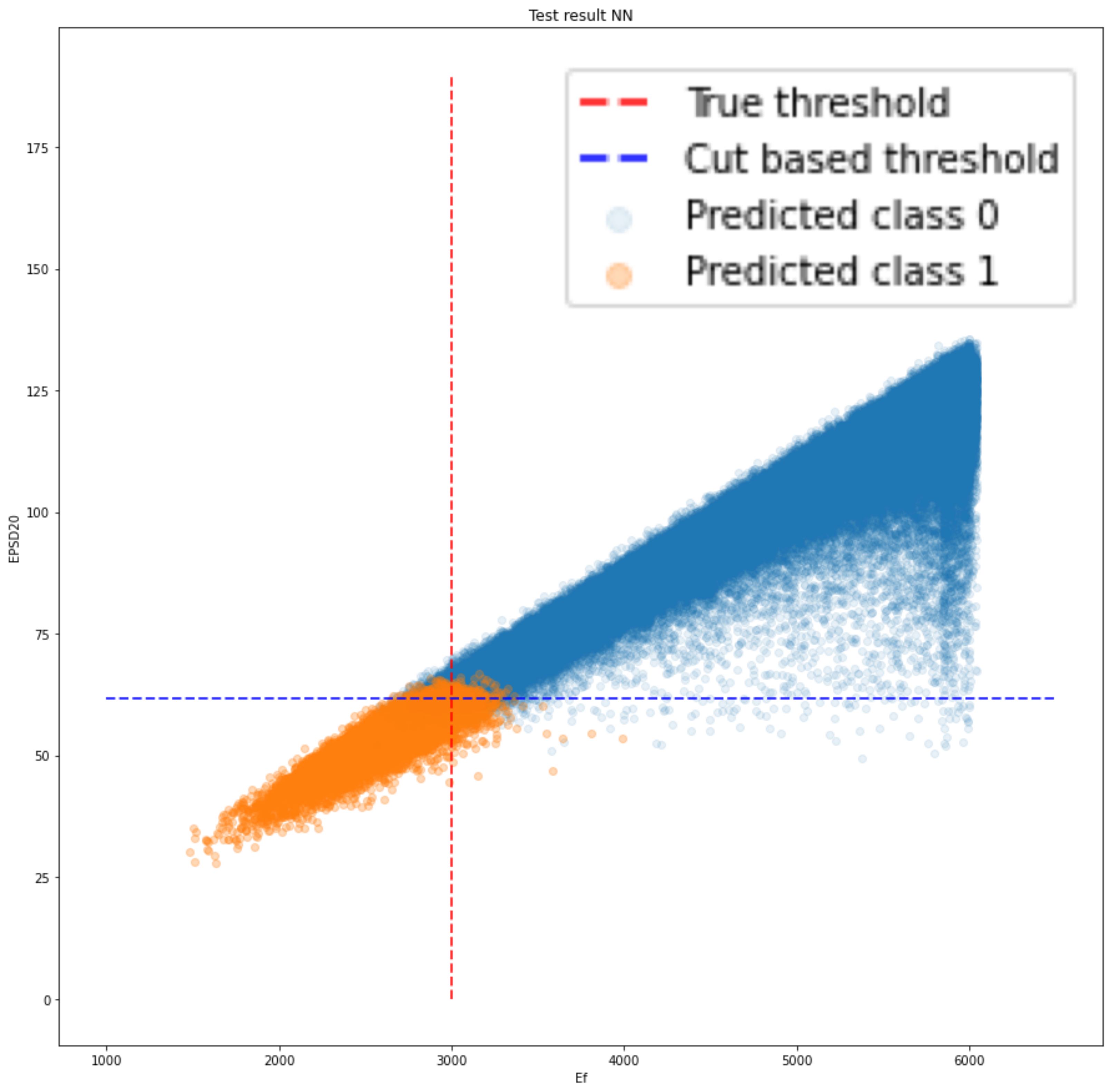


- Four most significant principal components illustrate the major role of small central modules with a ring of closest large modules.

Case №2 - Ar40+Sc45@150A GeV/c

- A number of classifying machine learning algorithms has been trained to distinguish between central and peripheral Ar+Sc collisions

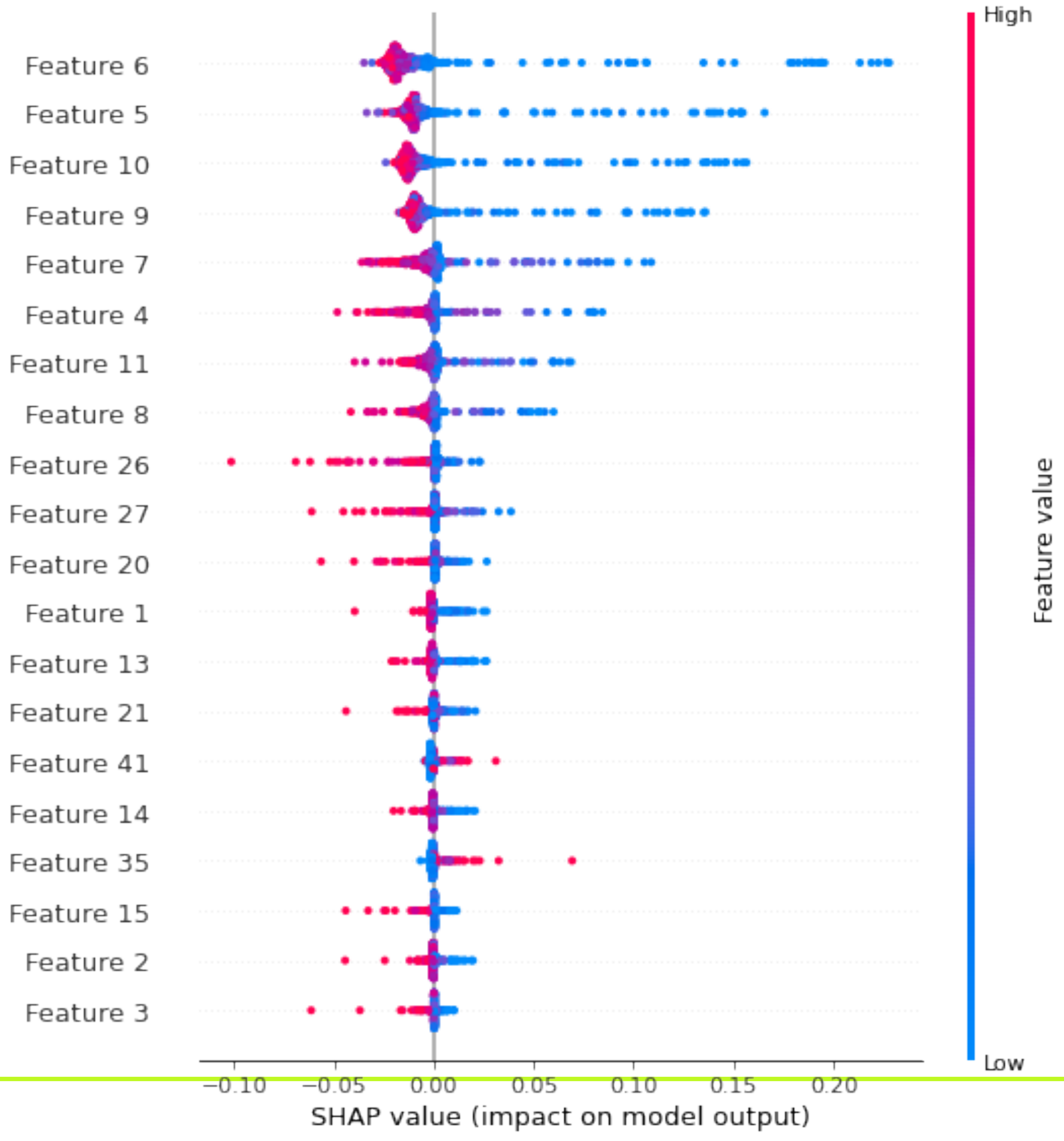




- Classifiers based on 44 features do not increase precision for a given recall (TPR) in comparison to cut based

Introduction of charged hadron multiplicity as the 45th feature allows to increase precision by ~1%.

Case №2 - Ar40+Sc45@150A GeV/c



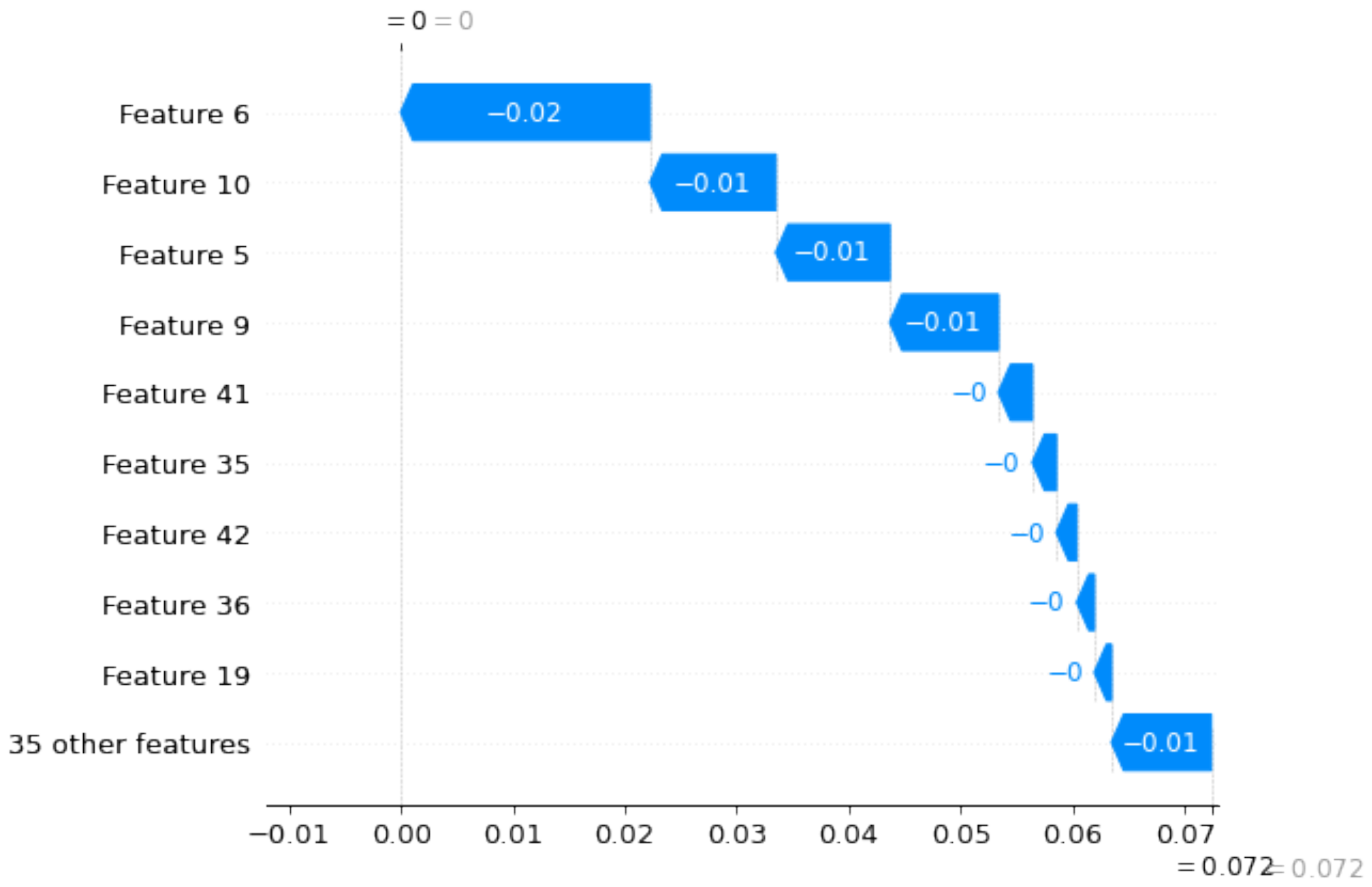
How to estimate feature importance for the neural net?

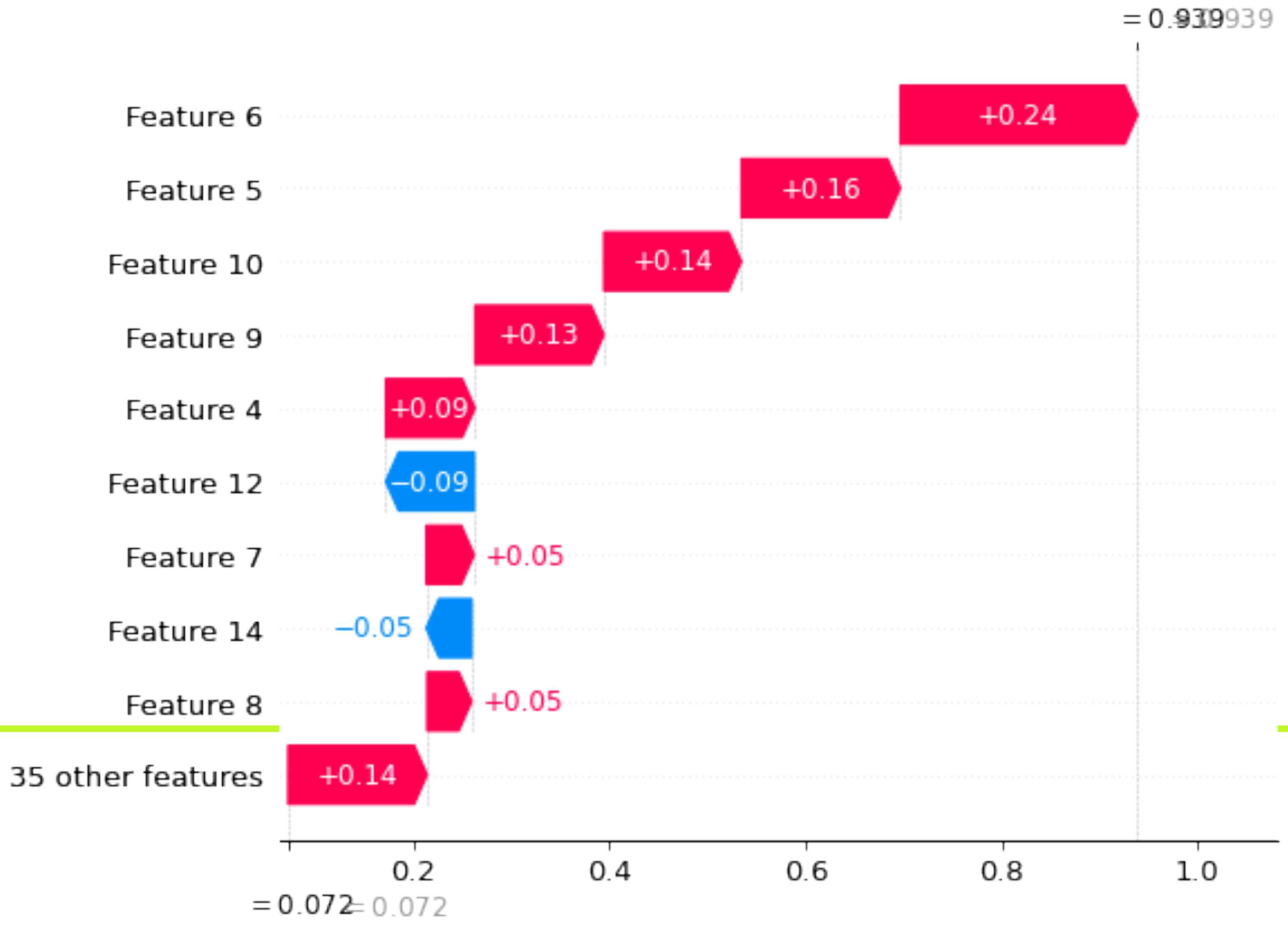
Using concept from the game theory - Shapley values

Shapley values of features estimated for Keras sequential convoluted neural network again indicate significant role of central modules.

Case №2 - Ar40+Sc45@150A GeV/c

Shapley values for single events classified as peripheral (left) and central (right). For both events the dominating feature is module 6.





Case №2 - Ar40+Sc45@150A GeV/c

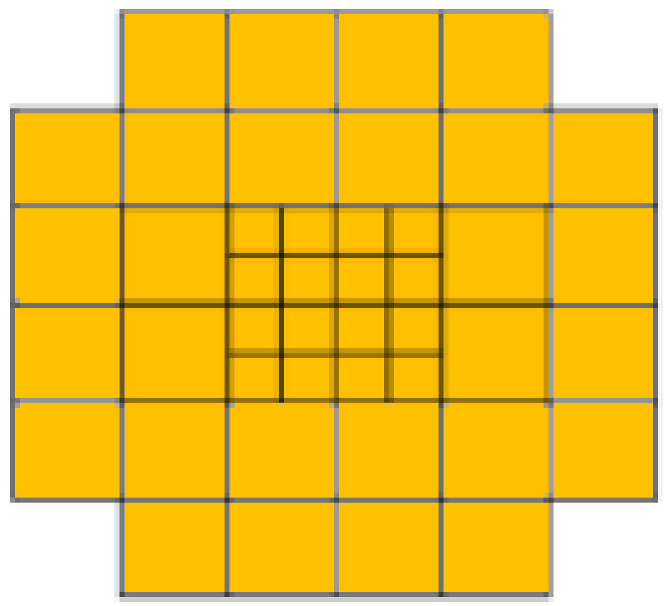
Gplearn package commonly used in genetic programming allows to perform symbolic classification. The idea of algorithm is fitting data with random functions of input features constructed using selected set of mathematical operators.

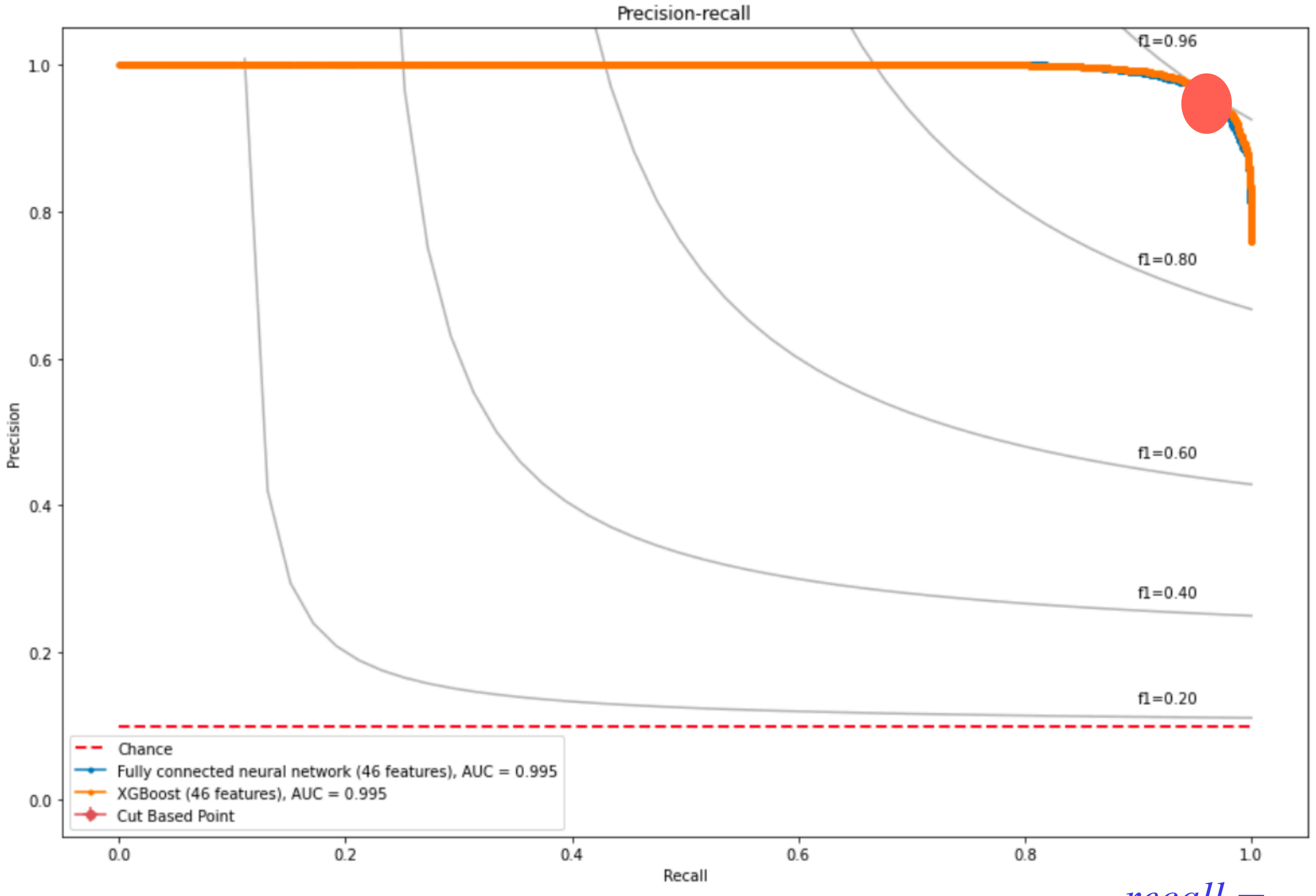
Symbolic classifier has been applied to this dataset with principal components as input features.

Optimal classifying function was found to be $13 \cdot PC_0 - PC_1 - 24.027$ with TPR=0.93, FPR=0.01, Prec=0.83

Case №3 - Pb+Pb@13A GeV/c

- DCM-QGSM-SMM model was used to generate MC data
- Energy depositions in all modules were selected as features
- Only XGBoost and fully-connected NN were tested in this analysis up-to-now
- No visible advantage in terms of precision-recall diagram in comparison to standard method can be seen.





$$precision = \frac{TP}{TP + FP}$$

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

Summary

- Multiple machine learning techniques have been applied to improve centrality selection in Li+Be, Ar+Sc and Pb+Pb collisions using energy deposition in the Projectile Spectator Detector.
- Best performance was obtained for light ion collisions while for Ar+Sc and Pb+Pb improvement in quality of selection was moderate.
- Additional analysis on feature importance in Ar+Sc collision indicated the major role of central modules in this procedure which limits its potential for application in collider experiments for heavy-ion collisions.

Thank you for your attention!

Supported by Saint Petersburg State University, project ID: 94031112.
We thank to the support and help from all the members of the CERN NA61/SHINE Collaboration

Back-up

Li+Be details of CNN

- Two classes: 0-3 and 4-7 spectators (15.8% centrality), 80000 events
- The best perfomance was obtained with the dropout rate parameter set as 0.1 (only 10% of FC neurons remain unzeroed)
- 1 conv layer with 128 features (3x3x5)
- 1 max pool (2x2)
- 1 FC layer with 1024 neurons
- Learning rate 5*1e-4
- Batch size 100