Machine learning in astrophysics

Fall into ML
02.11.22

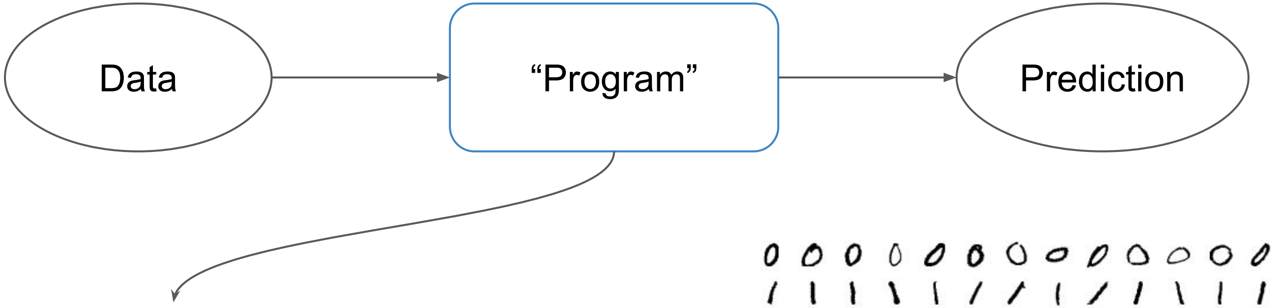
HSE UNIVERSITY

Ivan Kharuk, INR
ivan.kharuk@phystech.edu

Plan

- Why use ML?
- Telescope Array
  - Ways of thinking about data
  - Adjusting precision and recall
  - Validation of NN reliability
  - Making use of statistics
- Baikal-GVD
  - Choosing best data representation

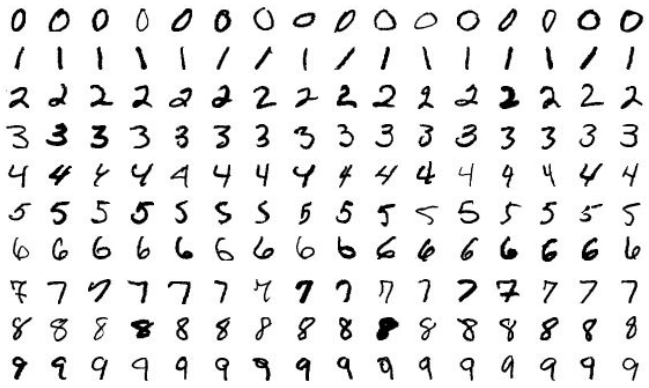
Why use ML?



Standard algorithms:

"Program" is a fixed algorithm, developed by a human.

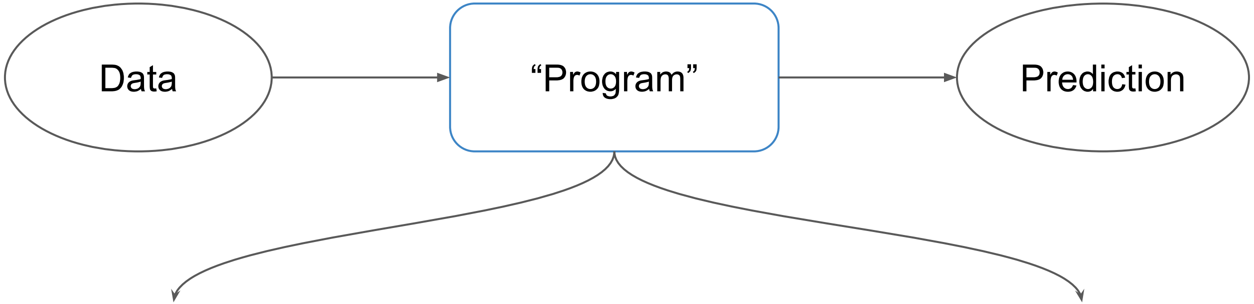
Best approach if the problem is exactly solvable



How to determine optimal features?

Why use ML?

Interpretable algorithms vs Powerful black box



Standard algorithms:

"Program" is a fixed algorithm, developed by a human.

Best approach if the problem is exactly solvable.

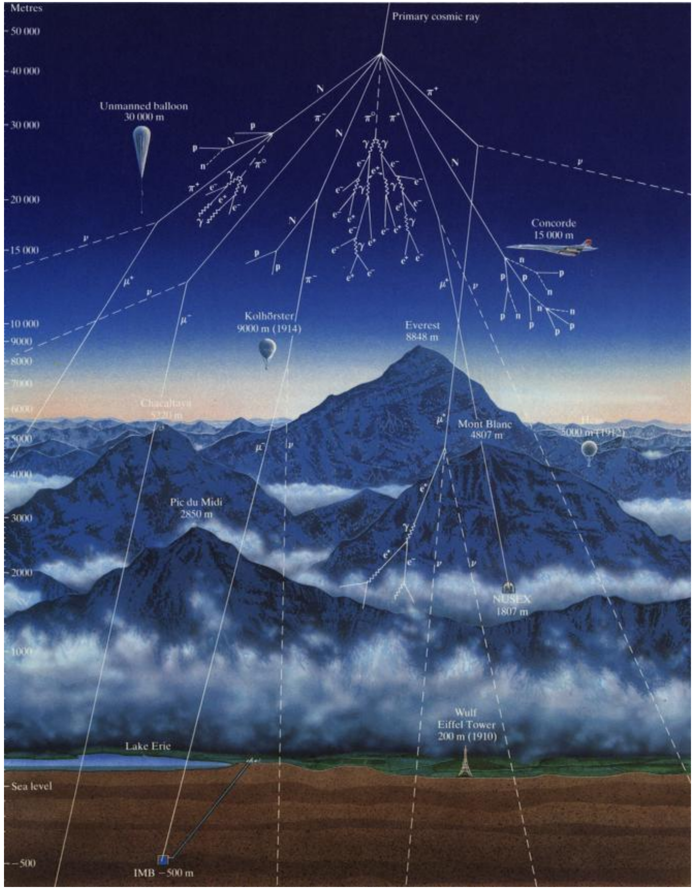
Machine learning:

"Program" is an algorithm, learning on examples to extract optimal features.

Programs that create optimal algorithms.

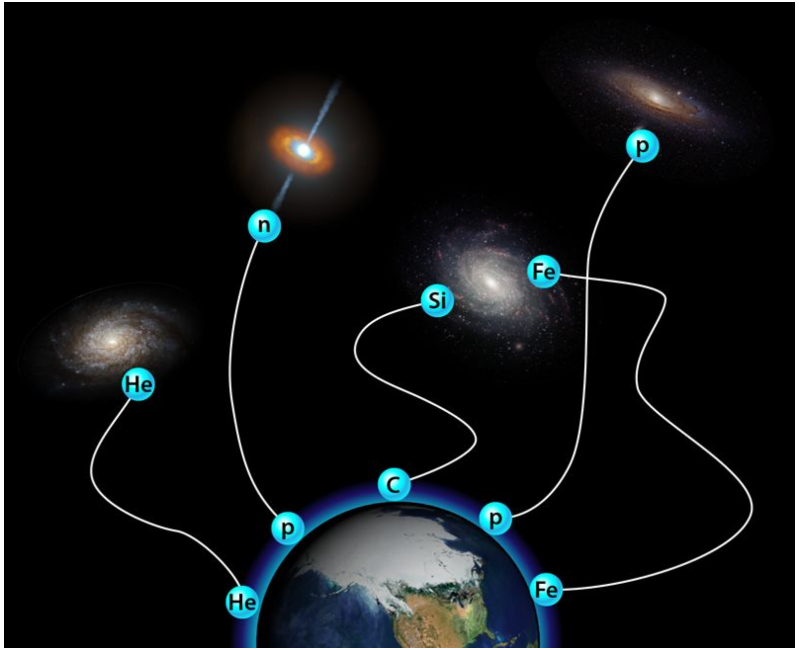
Telescope Array

What are cosmic rays



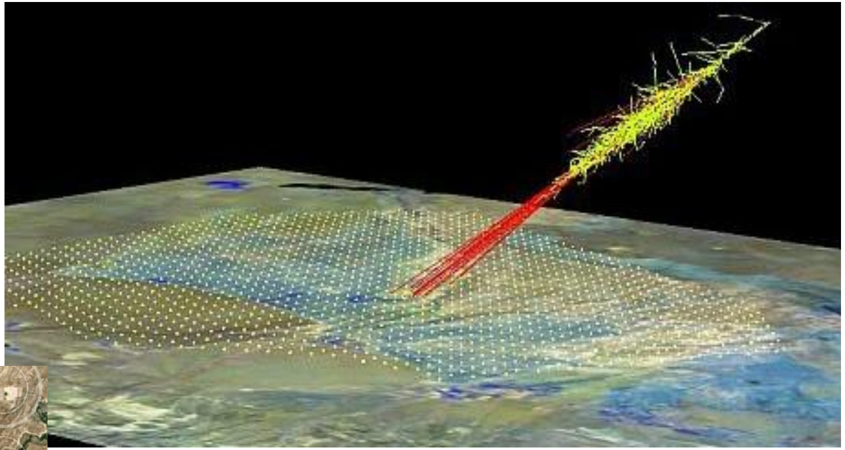
Indirect studies of cosmic objects:

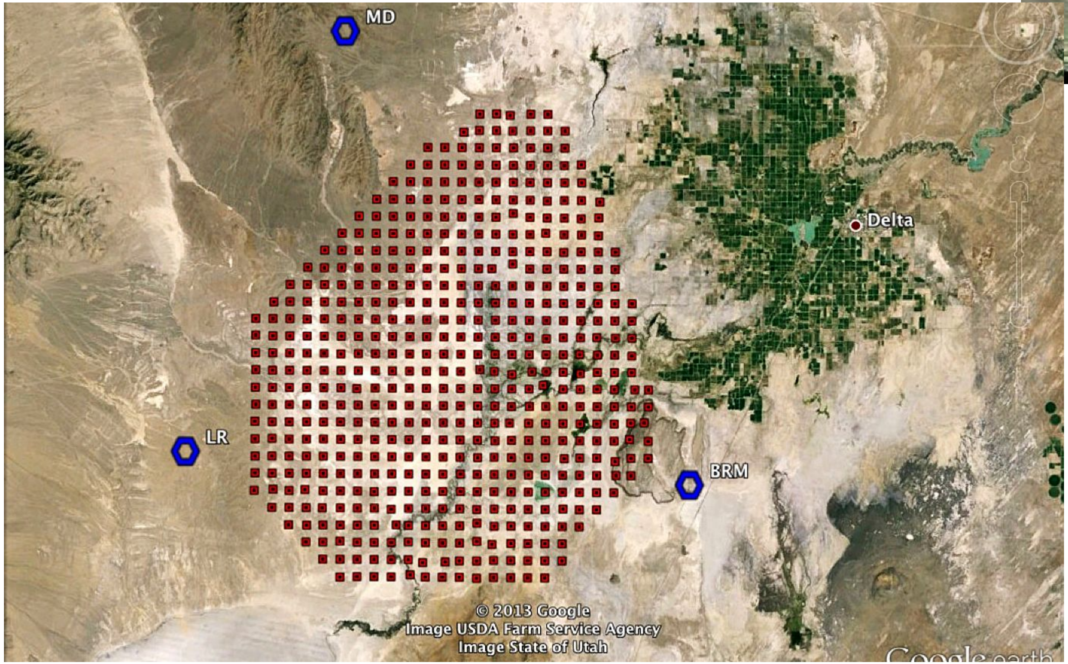
- Models of galaxies evolution
- Extremely-high energy physics
- Search for interesting objects

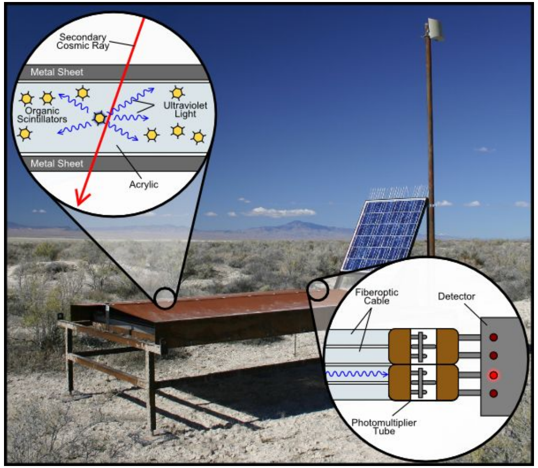


Telescope Array

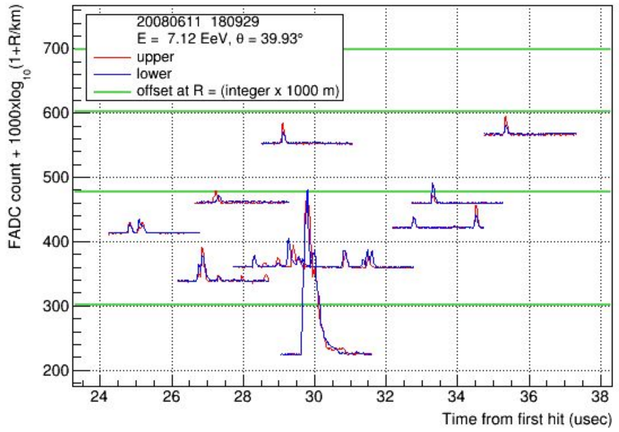
Largest cosmic ray observatory in Northern hemisphere
($700\ km^2$, 507 surface + 3 fluorescent detectors)

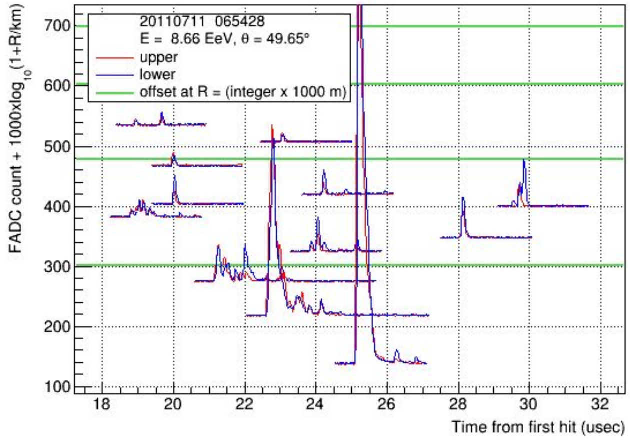






Stage 1

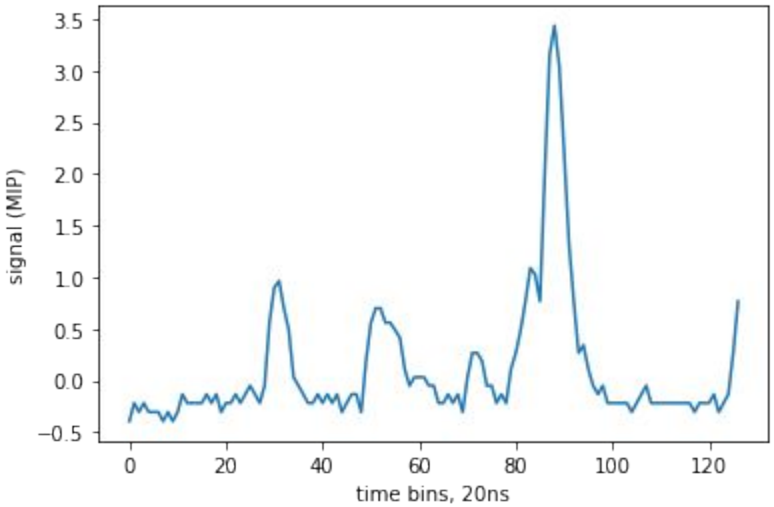


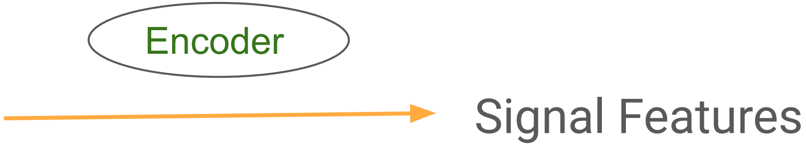


What is the best way to represent the data?

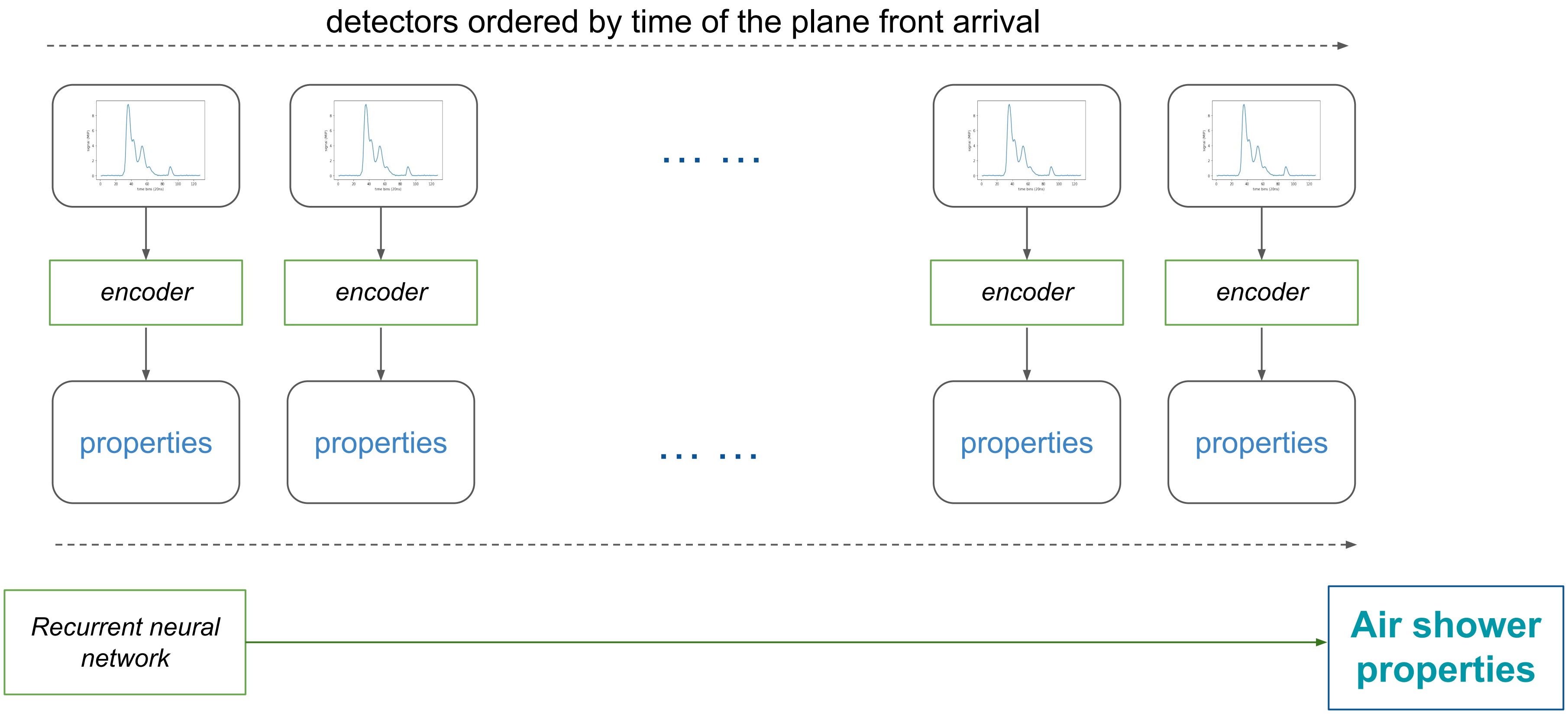
Waveforms: image or sequence?

- 128 bins (20ns each) of the signal in upper and lower detectors





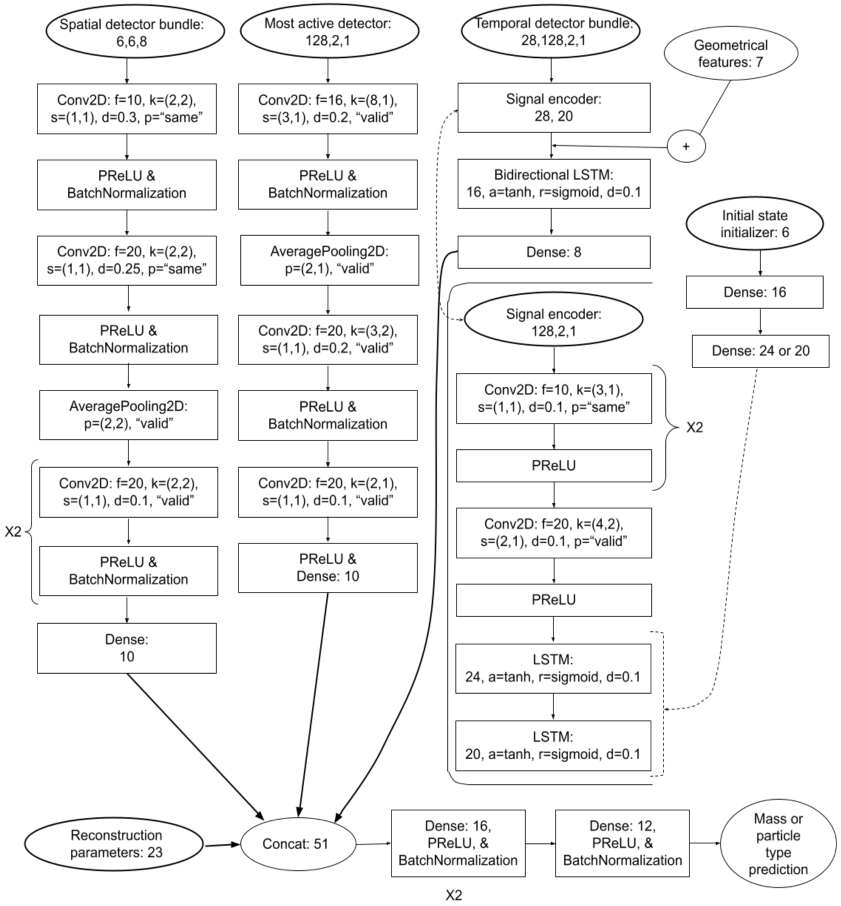
Event representation



Neural network

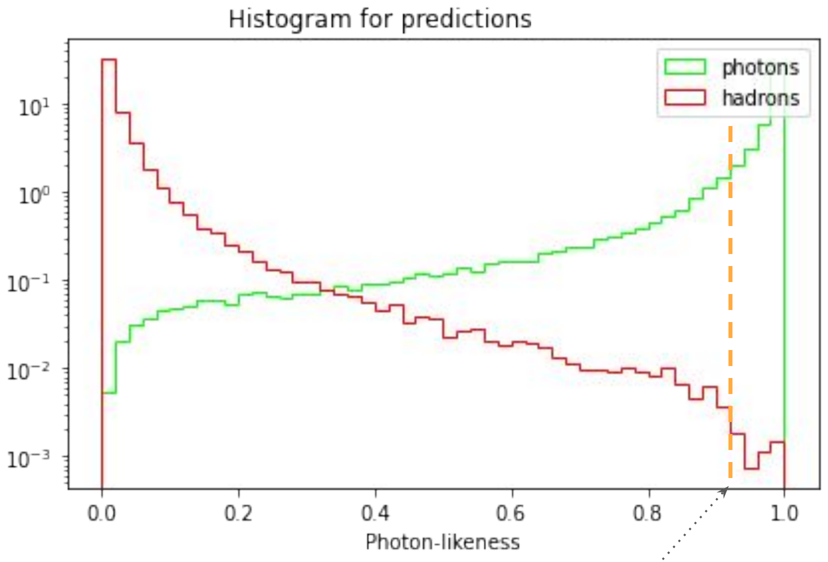
Neural network's blocks:

- Spatial detectors bundle (geometrical features)
- Strongest waveform (signal specifics)
- Temporal detector bundle (overall information)
- Reconstruction parameters (high-level information)



Stage 2-a

NN prediction ξ ∈[0;1]:
0 - hadron, 1 - photon

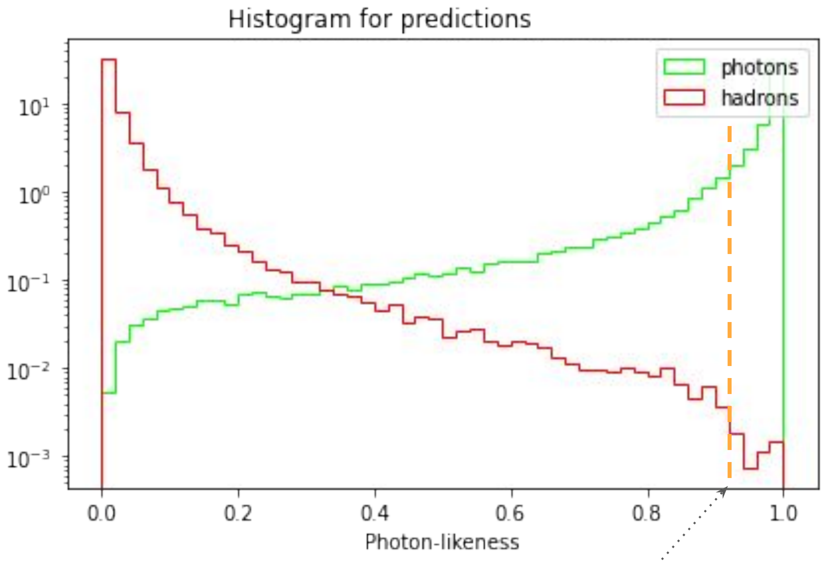


threshold

What is the right balance between true and false positives?

Optimizing predictions

NN prediction ξ ∈[0;1]:
0 - hadron, 1 - photon



threshold

Cut optimization: strongest sensitivity in absence of photons in data
(~ minimizing (false photons)/(true photons))

Requires special loss functions
(hand-made, *focal loss*)

Stage 3

NN must be insensitive to unphysical details.

Simulations (MC) are subject to errors and imprecision:

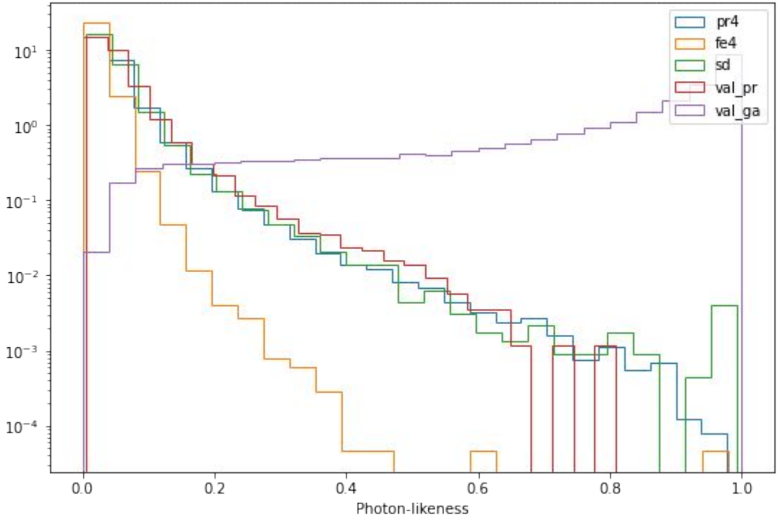
- Not (properly) working detectors
- Simulation errors and specifics
- Limitations of simulating the hardware response

How to make sure that NNs predictions are reliable?

Cross-checks

NN must be insensitive to unphysical details.

val_* - simulated with QGSJET II-03
*4 - simulated with QGSJET II-04
sd - surface detectors data



Make cross-checks:

- against standard algorithms
- between MC and real data

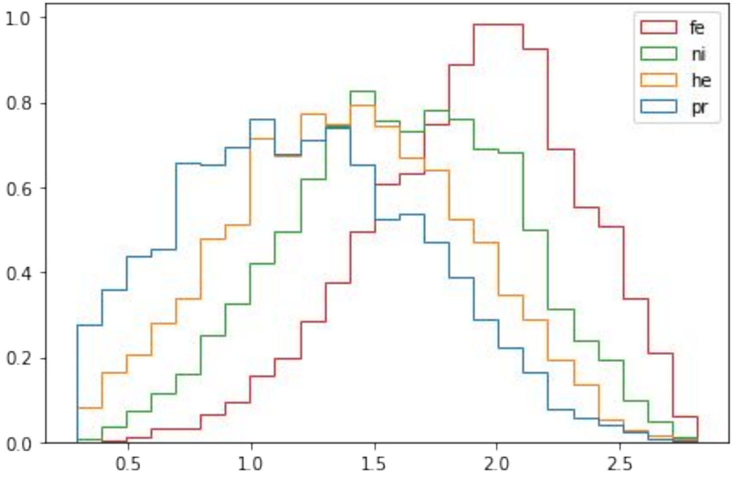
Discrepancies often can be resolved by:

- various dropouts
- masks
- noise sampling

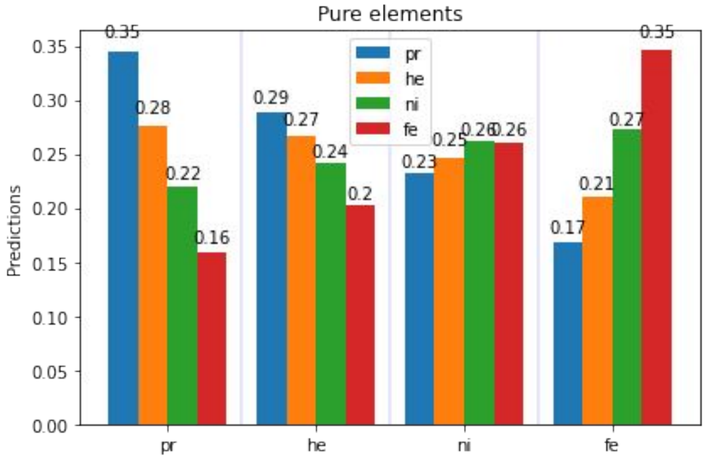
Stage 2-b

Evolution of air showers is stochastic.
Data may be similar for different primaries

Can we do better on ensembles of events?

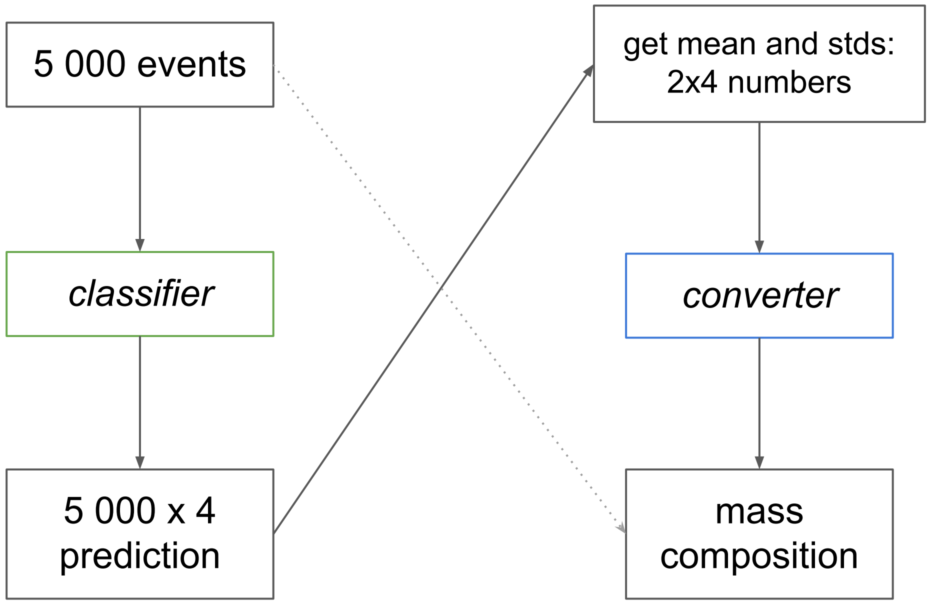


~39% success in 4-class model



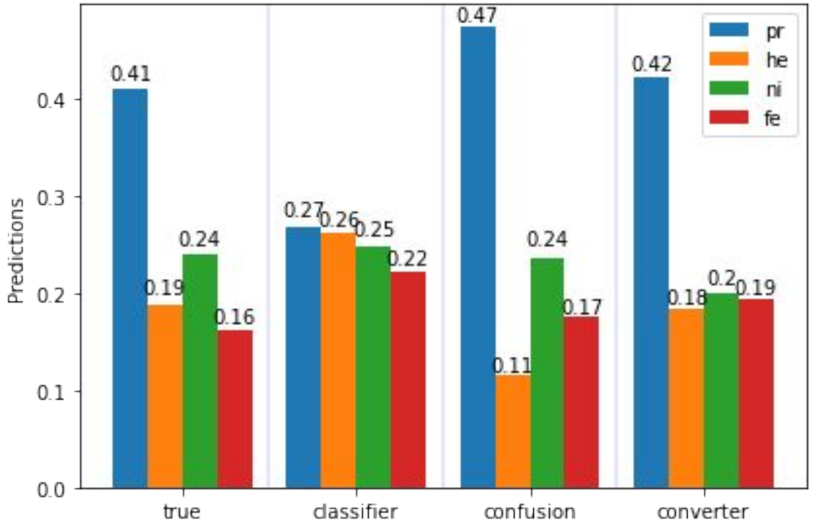
Making use of statistics

We are interested in obtaining mass composition of an ensemble of events!



Converter is the second neural network, which improves *classifier* predictions for ensembles of events

Making use of statistics

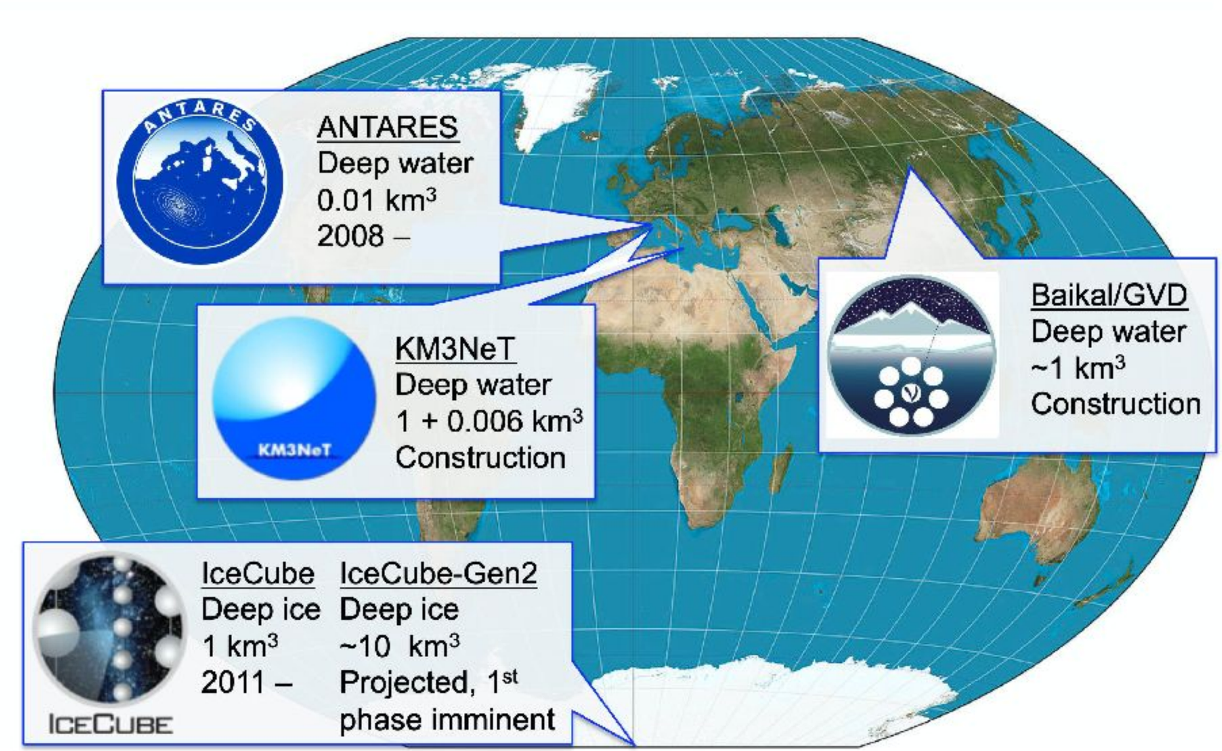


	proton	helium	nitrogen	iron
classifier	0.1	0.14	0.12	0.09
converter	0.03	0.07	0.06	0.02

Error: mean absolute error (averaging over events) on 2000 ensembles

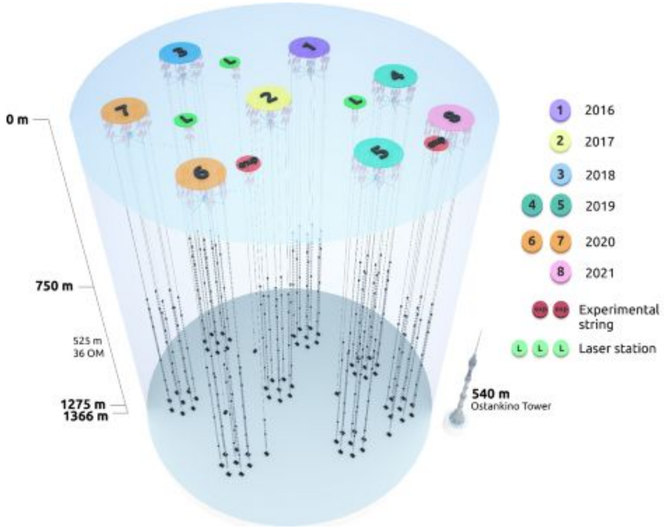
Baikal-GVD

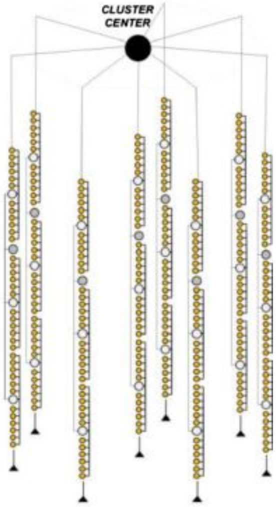
Baikal-GVD



ANTARES
Deep water
0.01 km³
2008 –
KM3NeT
Deep water
1 + 0.006 km³
Construction
Baikal/GVD
Deep water
~1 km³
Construction
IceCube
Deep ice
1 km³
2011 –
IceCube-Gen2
Deep ice
~10 km³
Projected, 1st phase imminent
ICECUBE
KM3NeT
ANTARES

Baikal-GVD

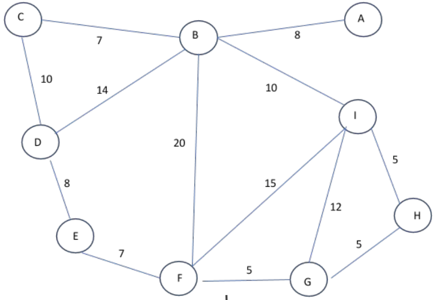

0 m
750 m
525 m
36 OM
1275 m
1366 m
540 m
Ostankino Tower
2016
2017
2018
2019
2020
2021
Experimental string
Laser station


CLUSTER CENTER

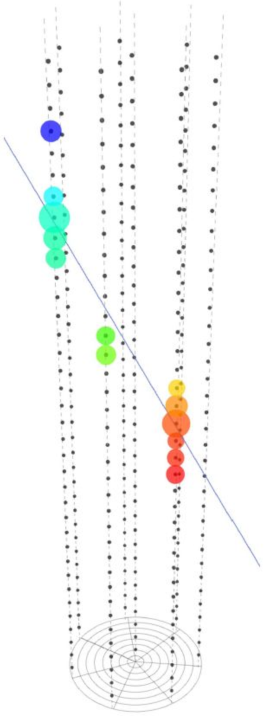
Baikal-GVD: data and its representation

Data representation should be **task-specific**.

Graph

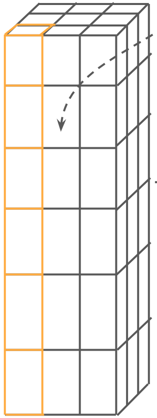


Graph updating protocols



Geometric

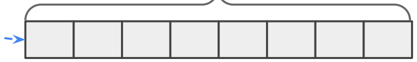
Detectors readings



3D convolutions

Causal

112 “optical modules”



Time ordering

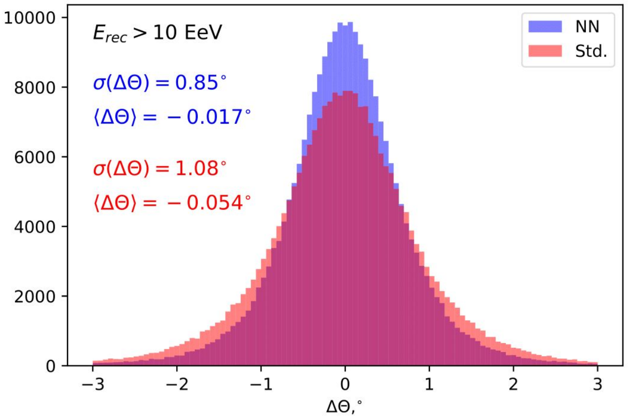
1D convolutions and RNNs

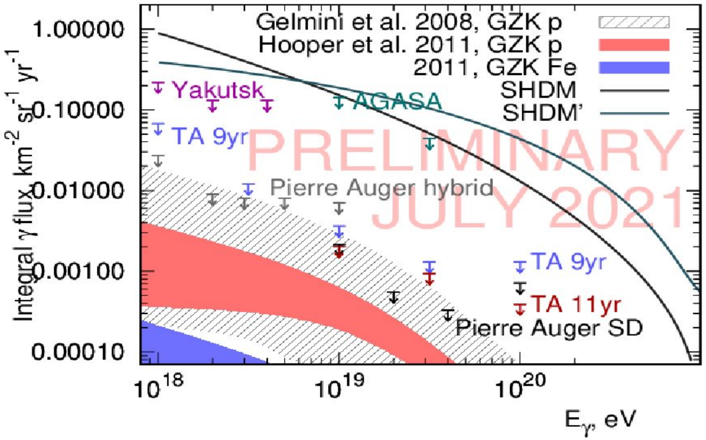
Thank you for attention!

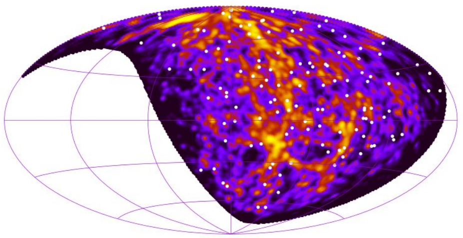
Appendix

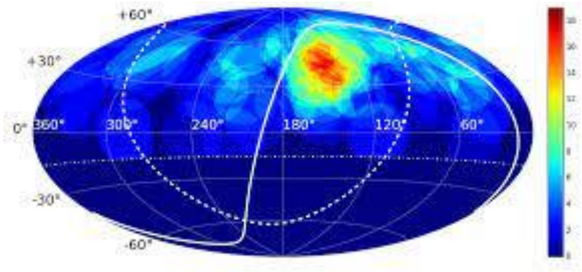
Other applications

One can estimate primary particle's:
mass, energy, and incoming direction



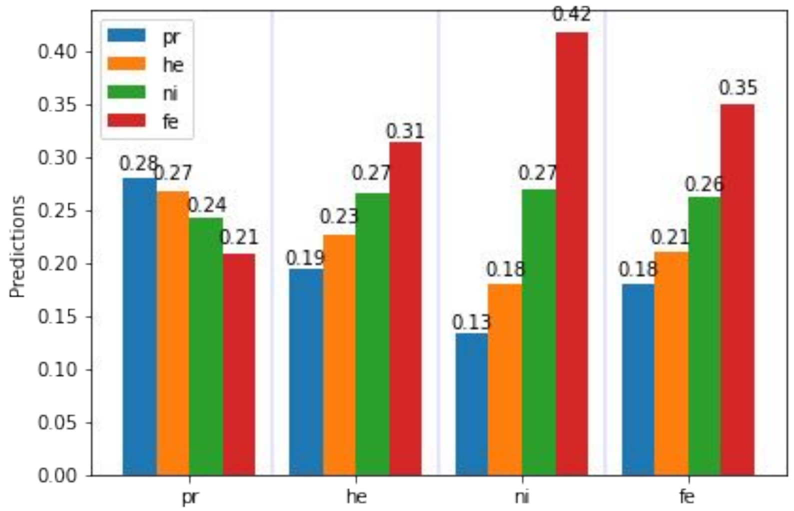




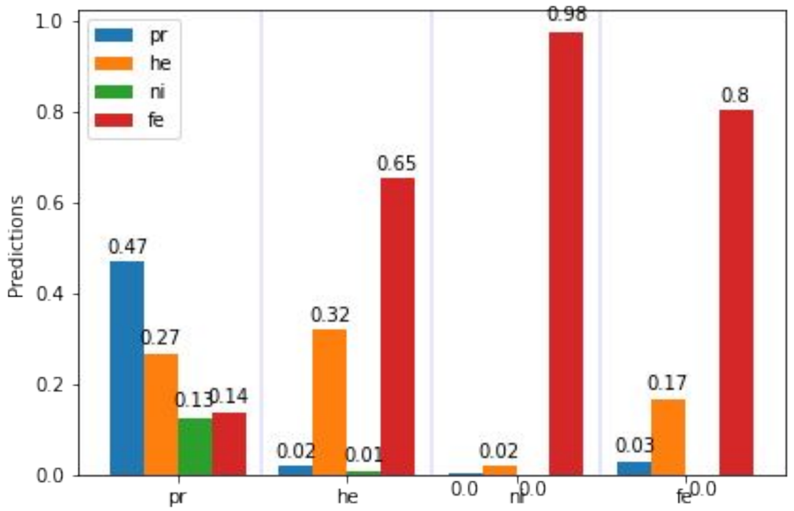


Model dependence

Neural network, trained on QGSJET II-03, observing events generated with QGSJET II-04:



Classifier predictions



Converter predictions

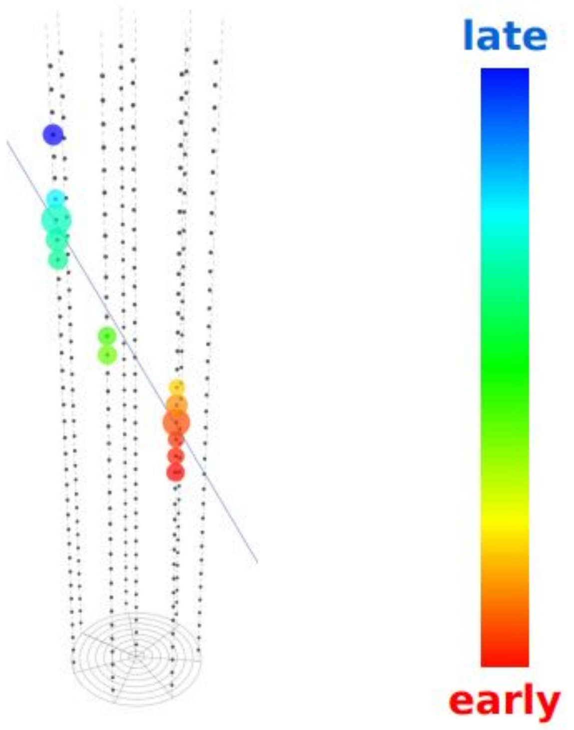
High systematic error: up to 100%

Baikal-GVD: tasks

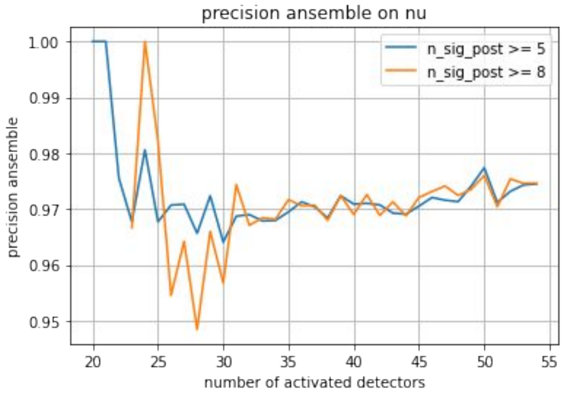
Task 1: Detectors are located underwater →
Signal-noise separation

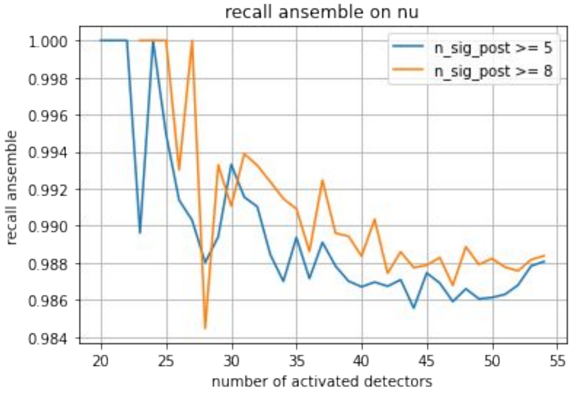
Task 2: Detector is sensitive to muons and neutrinos →
Identifying neutrino events

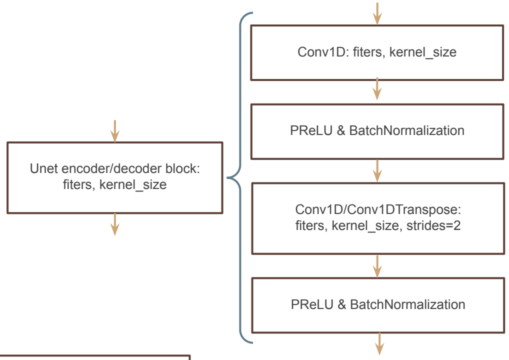
Taks 3-...: Given detectors' data,
Reconstruct the energy, arriving directions, etc.

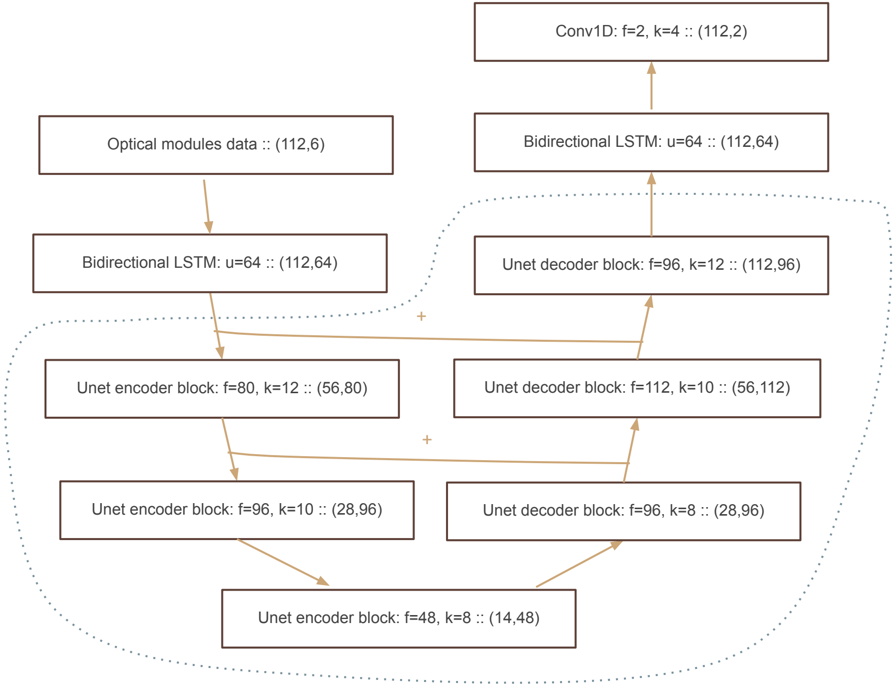


Signal-noise separation









Other applications:

- Obtaining posterior distribution of model parameters
  (intervertebral neural networks, arXiv:1808.04730, 2110.09493)

- Unsupervised clustering
  (deep adaptive image clustering, self-organizing maps)