Machine learning at Baikal-GVD

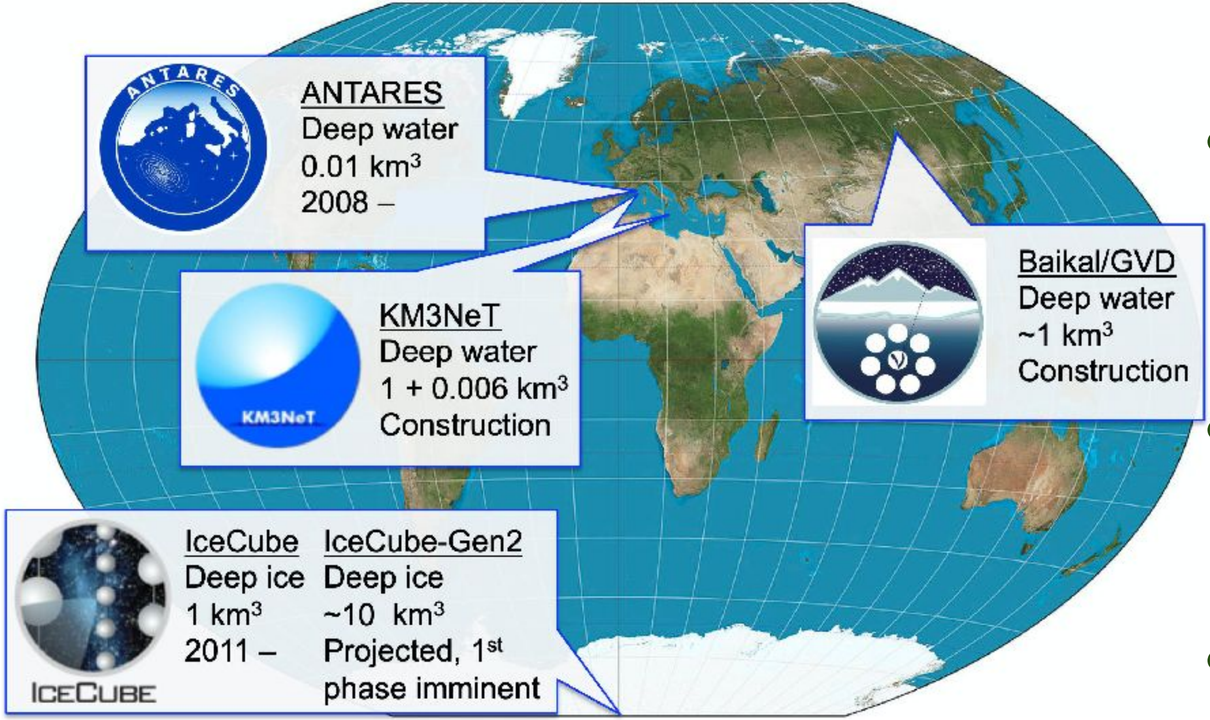
I. Kharuk (INR & MIPT), ivan.kharuk@phystech.edu

Outline

- Baikal-GVD experiment
- Machine-Learning applications
- Machine-Learning challenges

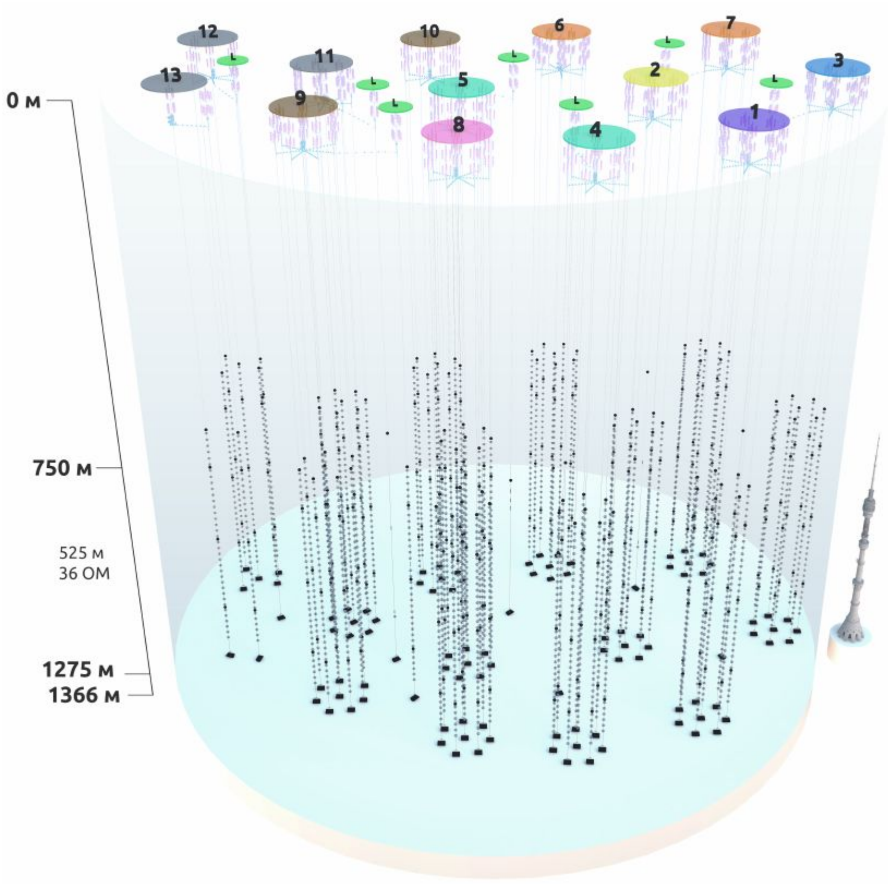
Baikal-GVD experiment

Baikal-GVD



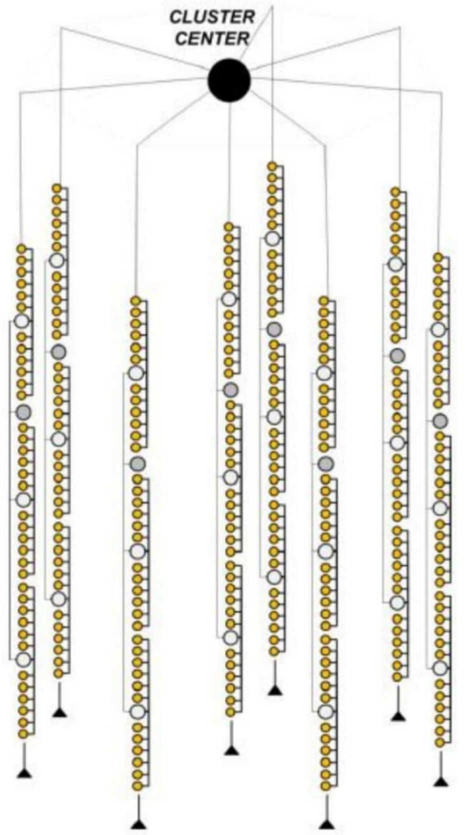
- Located in Lake Baikal, Russia
- Largest neutrino telescope in Northern hemisphere
- Small light scattering in water
- Effective exploitation:
  - New modules and maintenance at winter

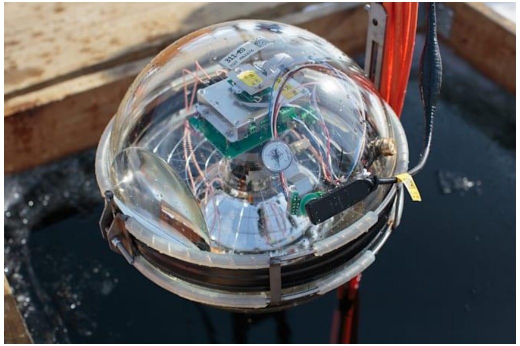
Baikal-GVD



- Instrumented volume: 0.7 km^3
- Modular structure:
  cluster diameter: 120 m
  inter-cluster distance: 300 m
- Target neutrino energies:
  TeV-PeV scale
- Register Cherenkov light from secondary particles

Baikal-GVD





8 strings per cluster:

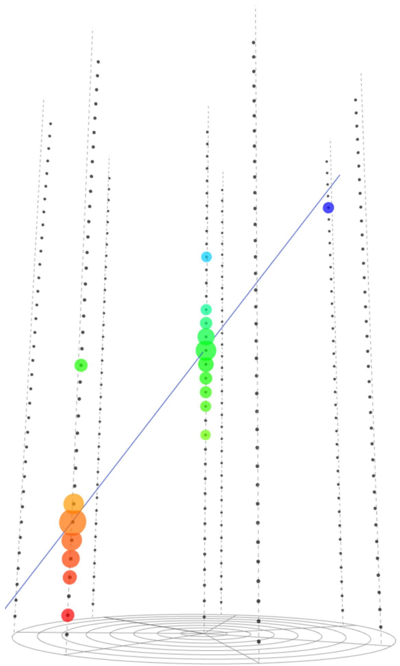
- ❏ Depth from 750 to 1275 m

36 optical modules per string:

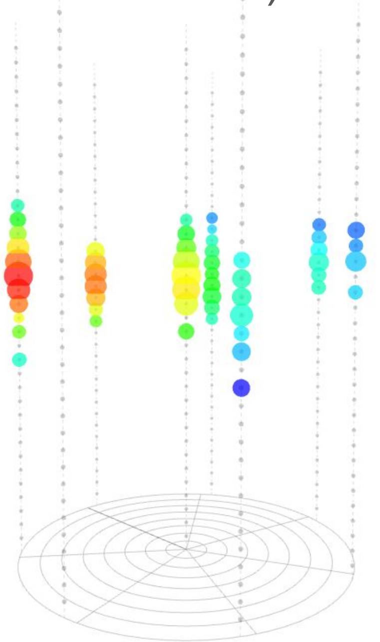
- ❏ 1 PMT looking downwards;
- ❏ Vertical spacing: 15 m
- ❏ Positioning accuracy: 20 cm
- ❏ Time synchronization : 2 ns.

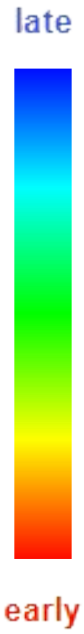
Baikal-GVD

Track-like (μ)



Cascade-like (EM or hadronic showers)





- Track-like Events
  - 200-300% energy resolution
  - 1° angular resolution
- Cascade-like Events:
  - ~20% energy resolution
  - 4° angular resolution

Machine Learning applications

Machine learning applications

(1)

Raw data → Identification of neutrino-induced events ⇢ Algorithmic reconstruction

Problem: 1 neutrino per 10^6-10^7 air showers

Suppress background by factor of 10 while keeping 98% neutrinos.

Machine learning applications

Raw data

(1)

Identification of neutrino-induced events

Algorithmic reconstruction

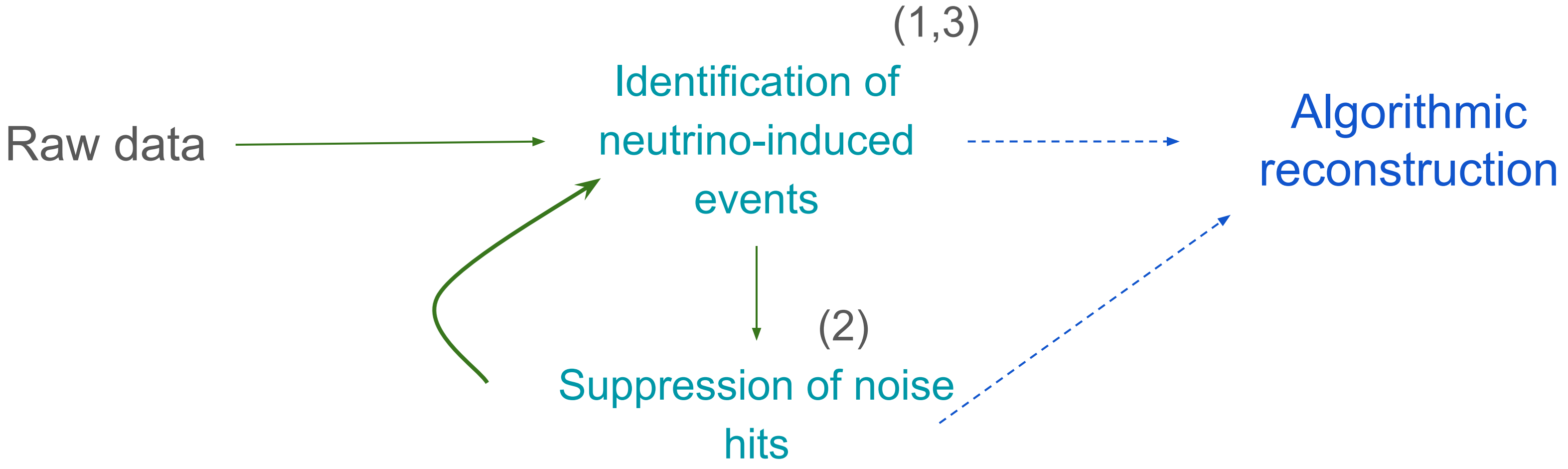
(2)

Suppression of noise hits

Problem: 85%-90% of collected hits are due to water luminescence

Reject OM activations due to noise

Machine learning applications



Problem: still may air showers

Identify neutrino candidates

Machine learning applications

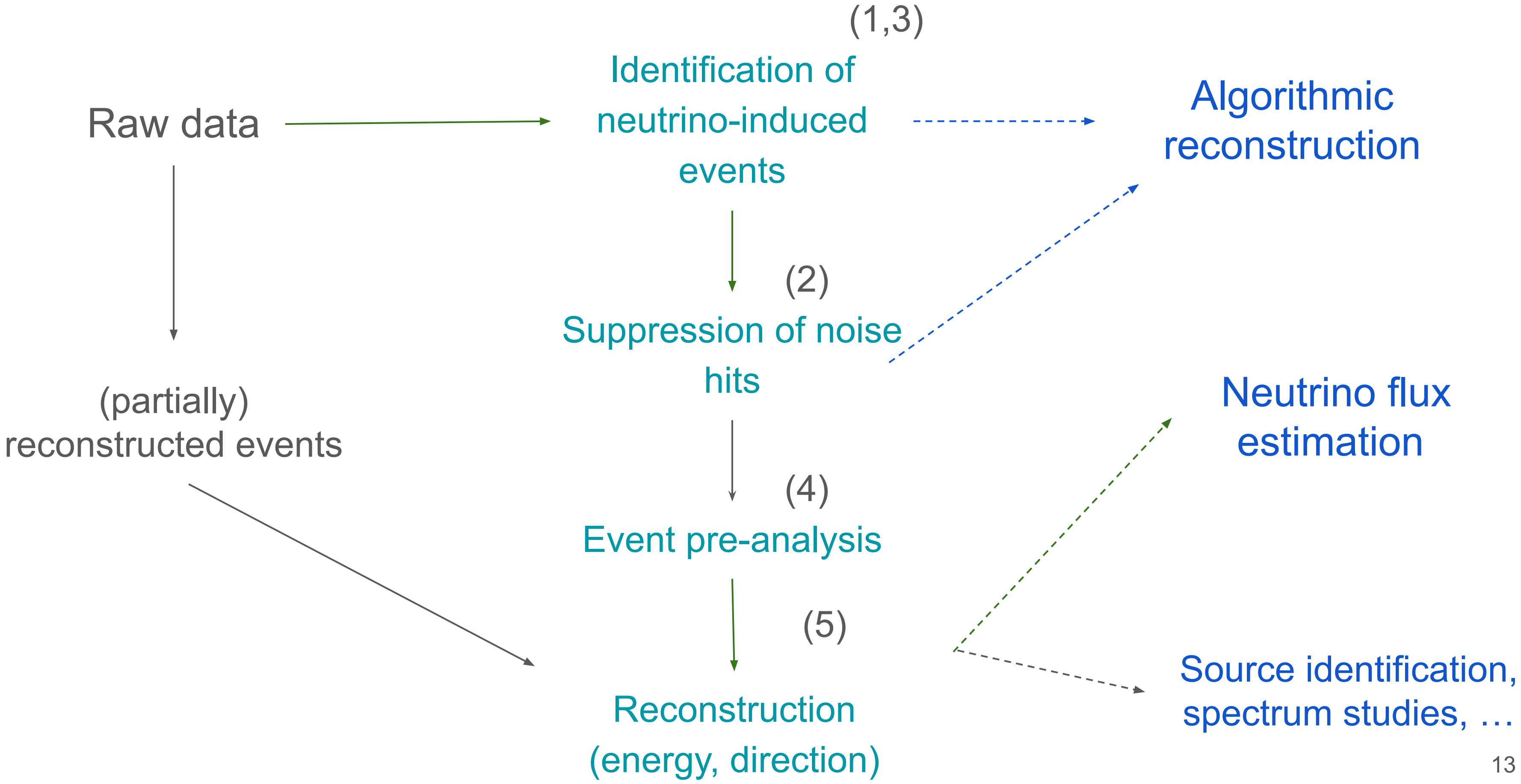
Raw data → (1,3) Identification of neutrino-induced events → Algorithmic reconstruction

Identification of neutrino-induced events → (2) Suppression of noise hits → Algorithmic reconstruction

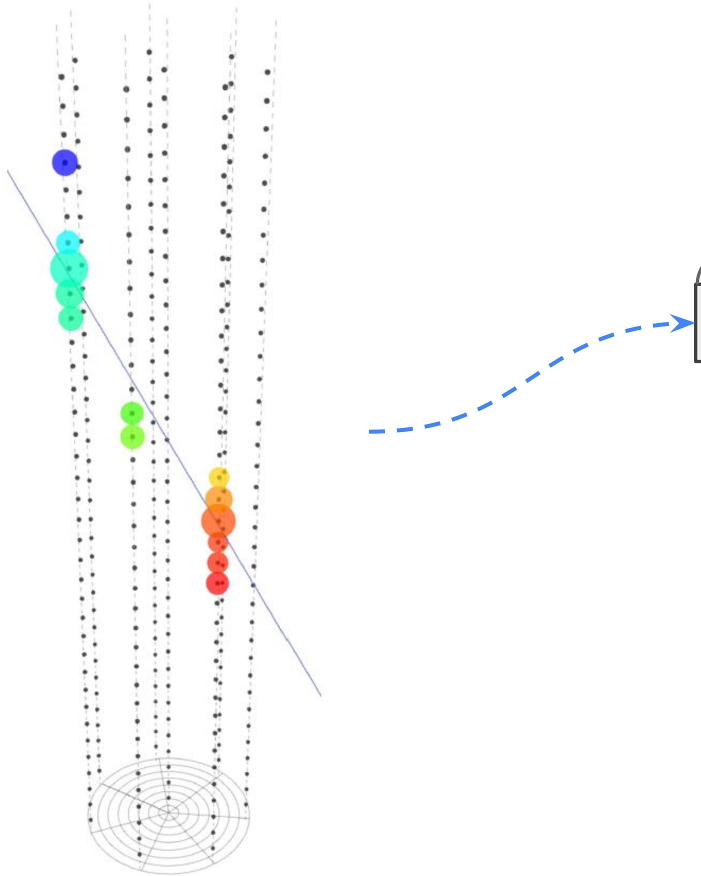
Suppression of noise hits → (4) Event pre-analysis

Understand event structure:
segment hits, estimate track length

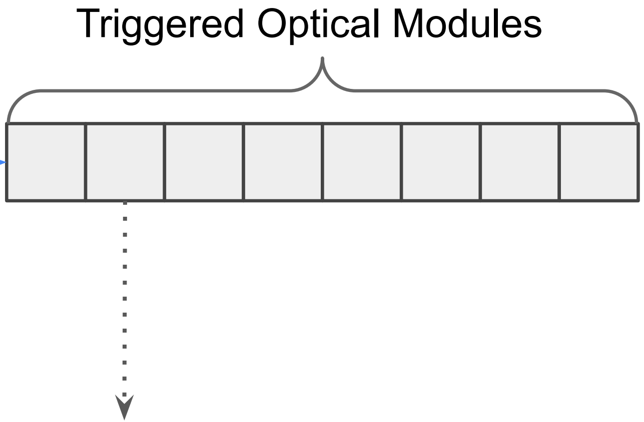
Machine learning applications



Data representation



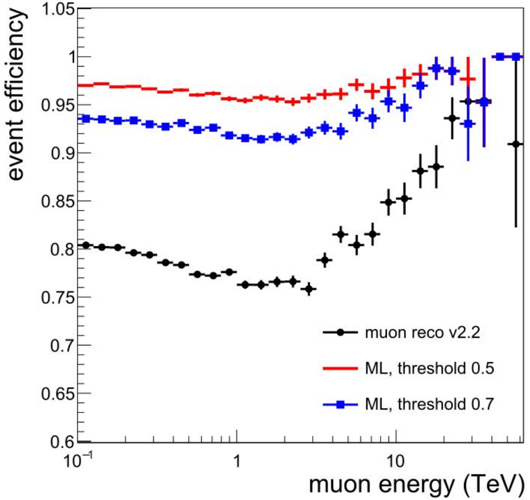
After experiments, decided to use Transformers

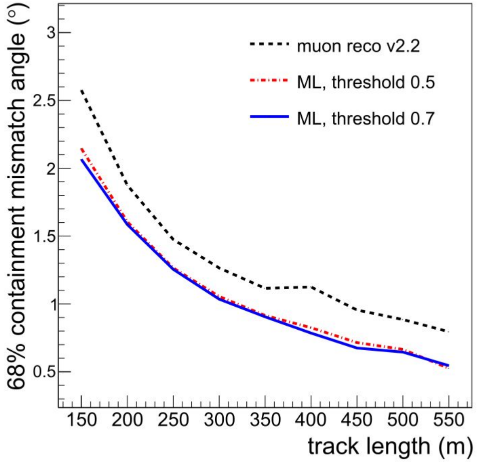


- Position
- Activation time
- Registered charge

Results

- Preserve 98% neutrinos while reducing background by factor of 10
- Improved OM noise suppression
  - Hence improved reconstruction metrics
- Improved angular and energy resolution with NNs
- Neutrino flux estimation:
  - Possible to identify 10±3 neutrino events on top of 10^6 EAS



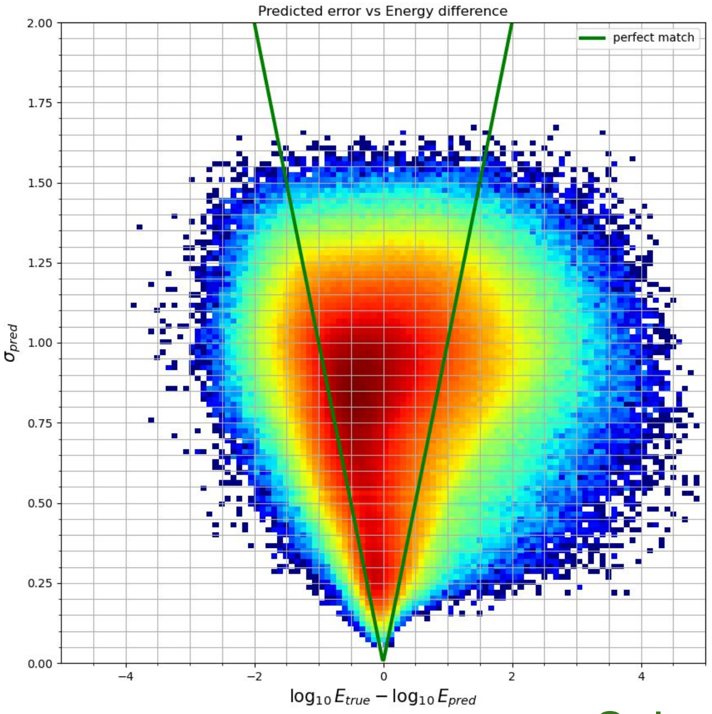


ML-based event quality cuts

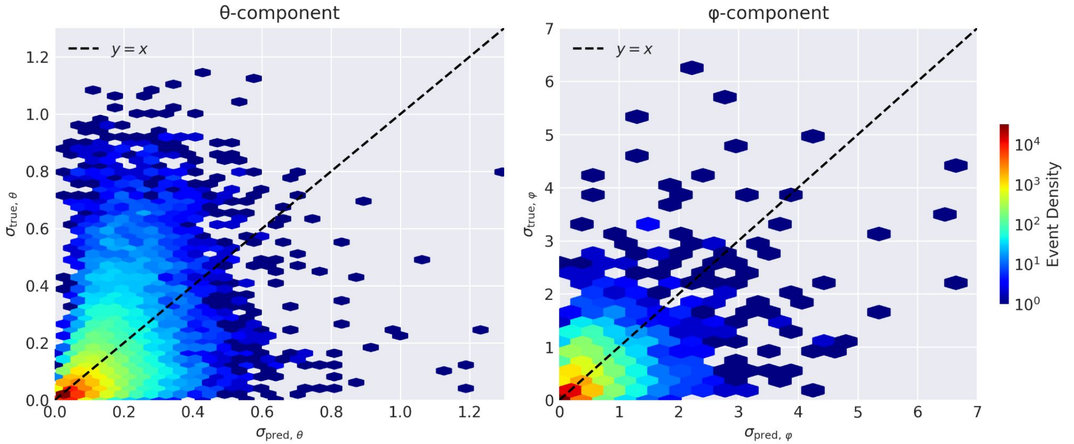
$$\text{Loss} = L_{reco} + \ln\sigma^2 + L_{reco} / \sigma^2$$

Train NN to estimate uncertainty!

Energy reconstruction uncertainty



Arrival direction reconstruction uncertainty



Select events by NN uncertainty in loosen quality cuts

Increased event statistics and reconstruction accuracy

Machine Learning challenges

The problem

Neural networks require data to train on. But what if such data is unavailable/biased?

Need to use simulations and rely on generalization / perfect MC.

Many sources of MC-data imperfections:

- Simplified MC simulations, geometry
- Detector noises and specifics
- Unknown systematics

Even if histograms closely match, NN can be over-specialized to MC data

NNs are bad at extrapolating

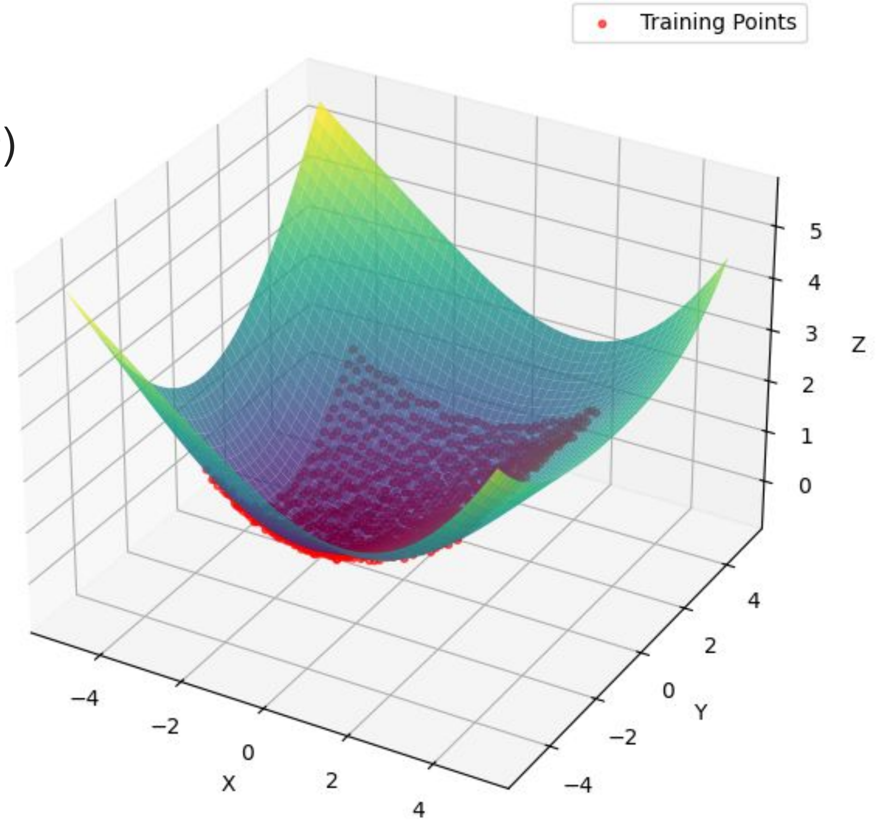
We want to fit the function:

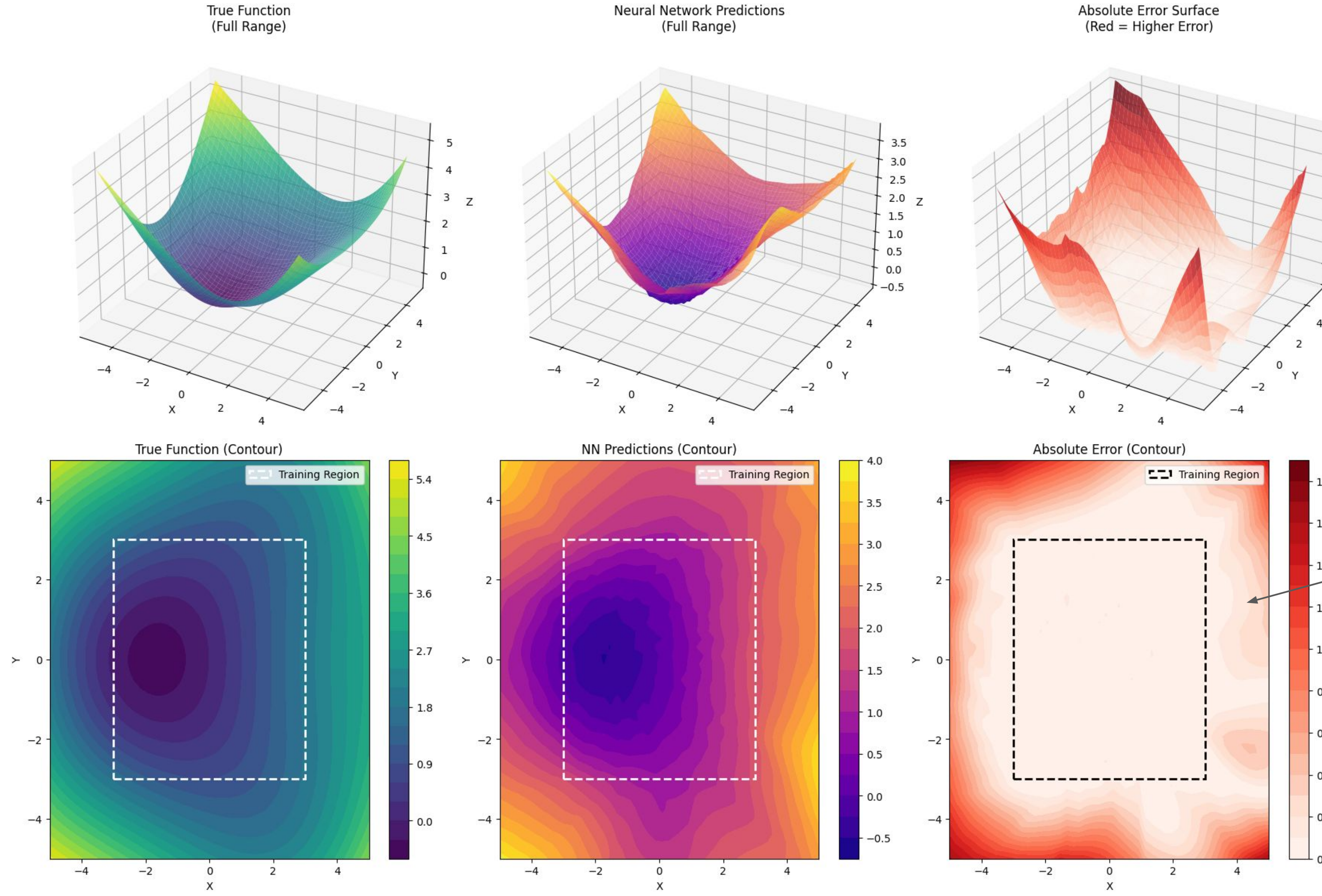
$f(x,y) = \sin(0.5 * x) * \cos(0.5 * y) + 0.1 * (x^2 + y^2)$

with MLP: 4 hidden layers,

128 neurons per layer.

Training data region (MC): x,y $\in$ [-3, 3]





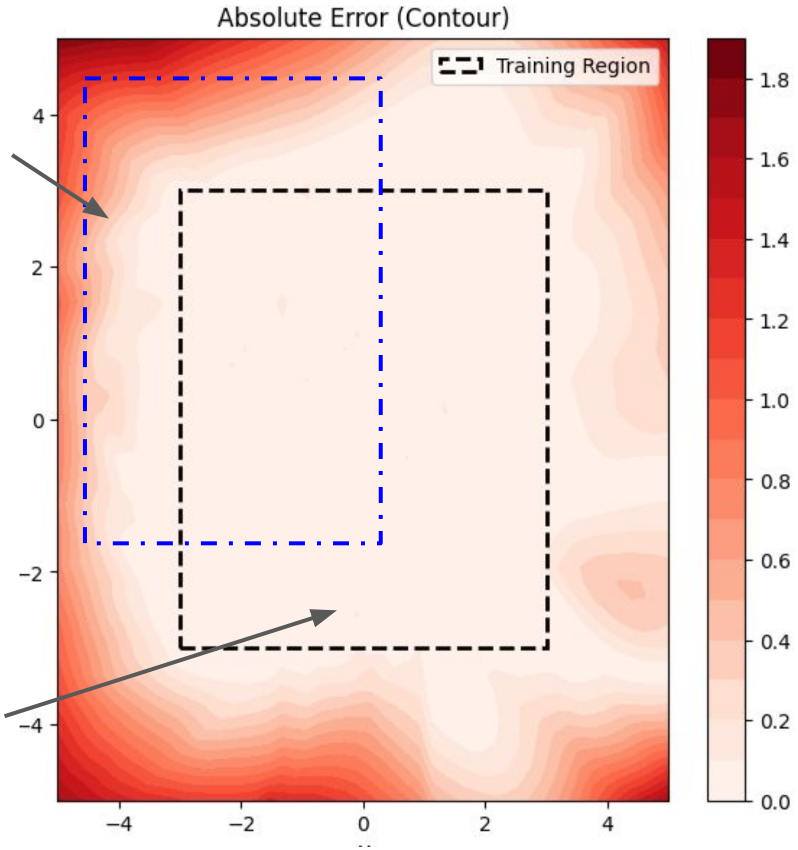
Unreliable area with increased error

Connection to cross-calibration

Actual data phase space, **P_{real}**

shift due to MC imperfections

MC phase space, **P_{MC}**



NN: $f : P_{MC} \rightarrow E_{expected}$

Calibration (bias correction):
$g(f): P_{real} \rightarrow E_{expected}$

But:
$\Delta f(P_{real})$ is highly nonlinear, hence calibration must be nonlinear as well
Also:
Increased error (need to be properly estimated)

Solution: Domain adaptation

arXiv:1409.7495

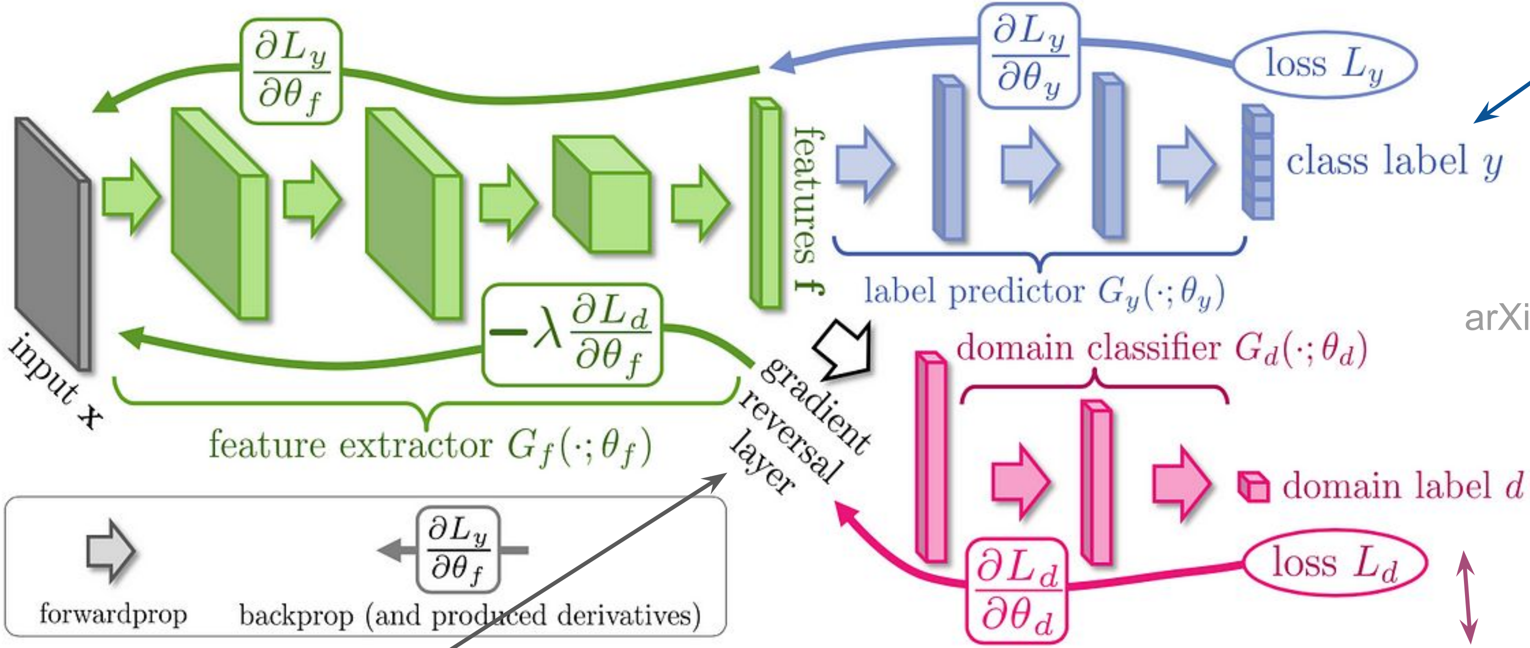
Identify and learn domain-invariant features

Available data for training (source domain)

Data of interest (target domain)

Domain adaptation technique

Train to recognize digits



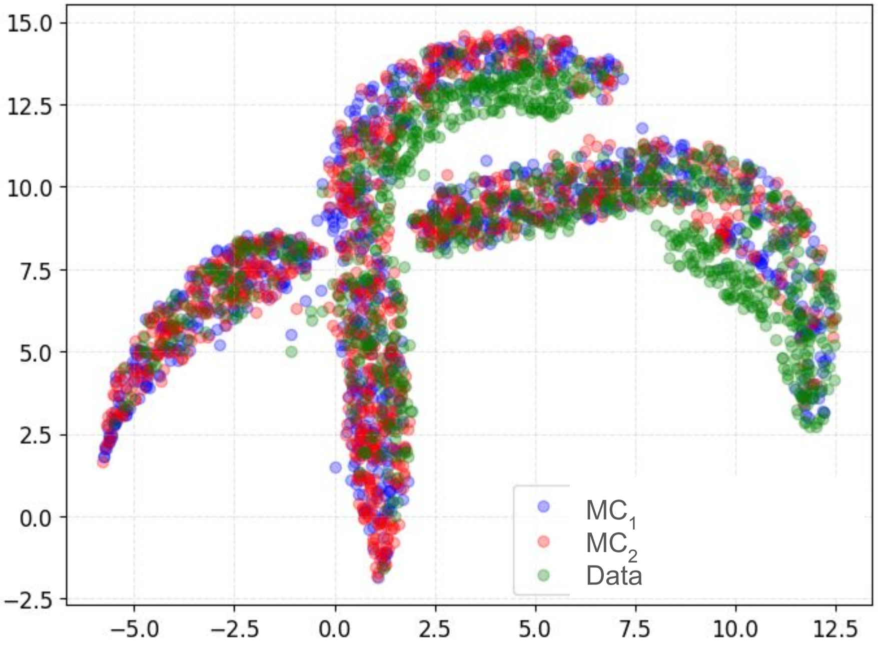
Source: arXiv:1409.7495

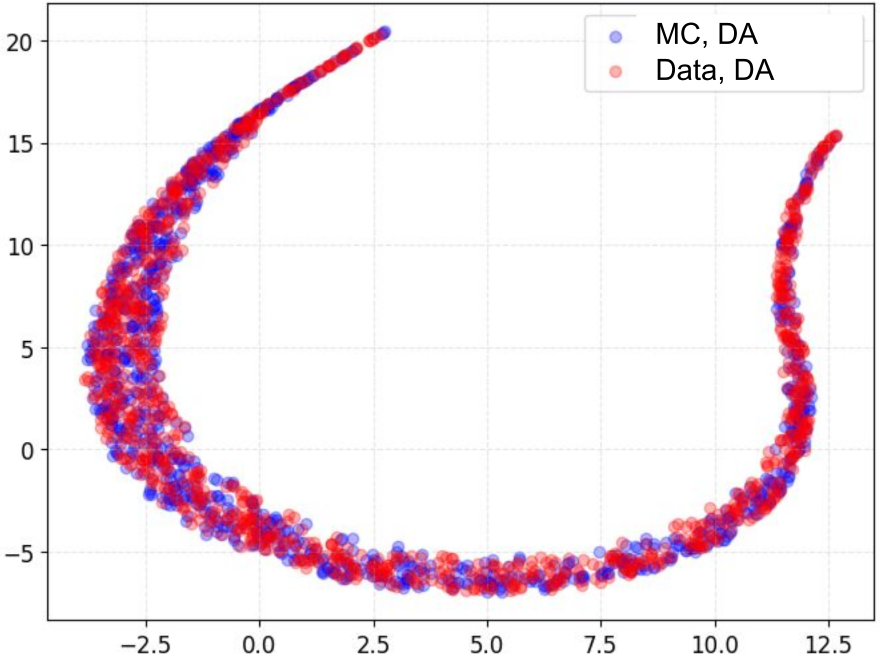
Exclude domain-specific features

vs

UMAP comparison

Project features extracted by NN to 2D using UMAP for visual comparison.





Applications in physics

- CMS@LHC (2405.13778)
- Cherenkov Telescope Array Large-Sized Telescope (2308.12732)
- Photometric Classification of Supernovae (1810.06441)
- Strong Gravitational Lens Analysis (2410.16347)

Conclusion

- Neural networks improve reconstruction accuracy
  - And are fast!
- Transformer-encoder is powerful architecture suitable for many experiments
- Neural network can become specialized/biased to MC
  - Domain adaptation as solution
  - But there are some pitfalls…