Analysis of the TAIGA-HiSCORE Data using the Latent Space of Autoencoders

Yulia Dubenskaya[1], Stanislav Polyakov[4], Alexander Kryukov[1], Andrey Demichev[1], Pavel Volchugov[1], Elizaveta Gres[2], Dmitry Zhurov[3], Evgeny Postnikov[1], Alexander Razumov[1]

[1] *Skobeltsyn Institute of Nuclear Physics, Moscow State University*
[2] *Research Institute of Applied Physics, Irkutsk State University*
[3] *Irkutsk State University*
[4] *Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia*

The work was supported by RSF, grant no.24-11-00136.

07.2025

About TAIGA-HiSCORE

TAIGA-HiSCORE is a large-scale (10 km²) array of 1000 wide-angle non-imaging Cherenkov detectors, spaced ~100 m apart, designed to detect and study extensive air showers (EAS) from primary cosmic rays and gamma rays

TAIGA-HiSCORE enables the reconstruction of key EAS parameters: core position, arrival direction, and energy of the primary particle

TAIGA-HiSCORE data processing pipeline

Shower
↓
Cherenkov light
↓
Station signals
↓
Reconstruction pipeline
↓
Particle parameters

Each TAIGA-HiSCORE detector records:

- Light amplitudes (Cherenkov photon density)
- Arrival time differences (nanosecond precision across the array)

These parameters form the raw input for shower reconstruction

The fundamental challenge

Reconstruct the primary particle's properties (energy, direction, etc.) from the spatial and temporal patterns of detector signals

The conventional approach: Why empirical parameters may not tell the whole story

Beyond amplitudes and timing, the current shower reconstruction method relies on auxiliary physical parameters:

- Pulse shape analysis: Rise time and width of the arrival time distribution
- Arrival time front curvature: Geometry of shower development
- Lateral distribution function: signal amplitude vs. core distance
- Cherenkov light density: Flux at 200m from shower core (R200)

While effective, these reconstruction parameters remain simulation-dependent, having been empirically optimized and validated through Monte Carlo studies

But is this the optimal way to extract particle properties?

How do we know if all the important information contained in the experimental data is preserved?

A machine learning approach to energy reconstruction

Traditional methods rely on empirical physical parameters (core position, amplitudes, etc.) to reconstruct particle energy

We propose replacing these physical parameters with parameters automatically learned from the data

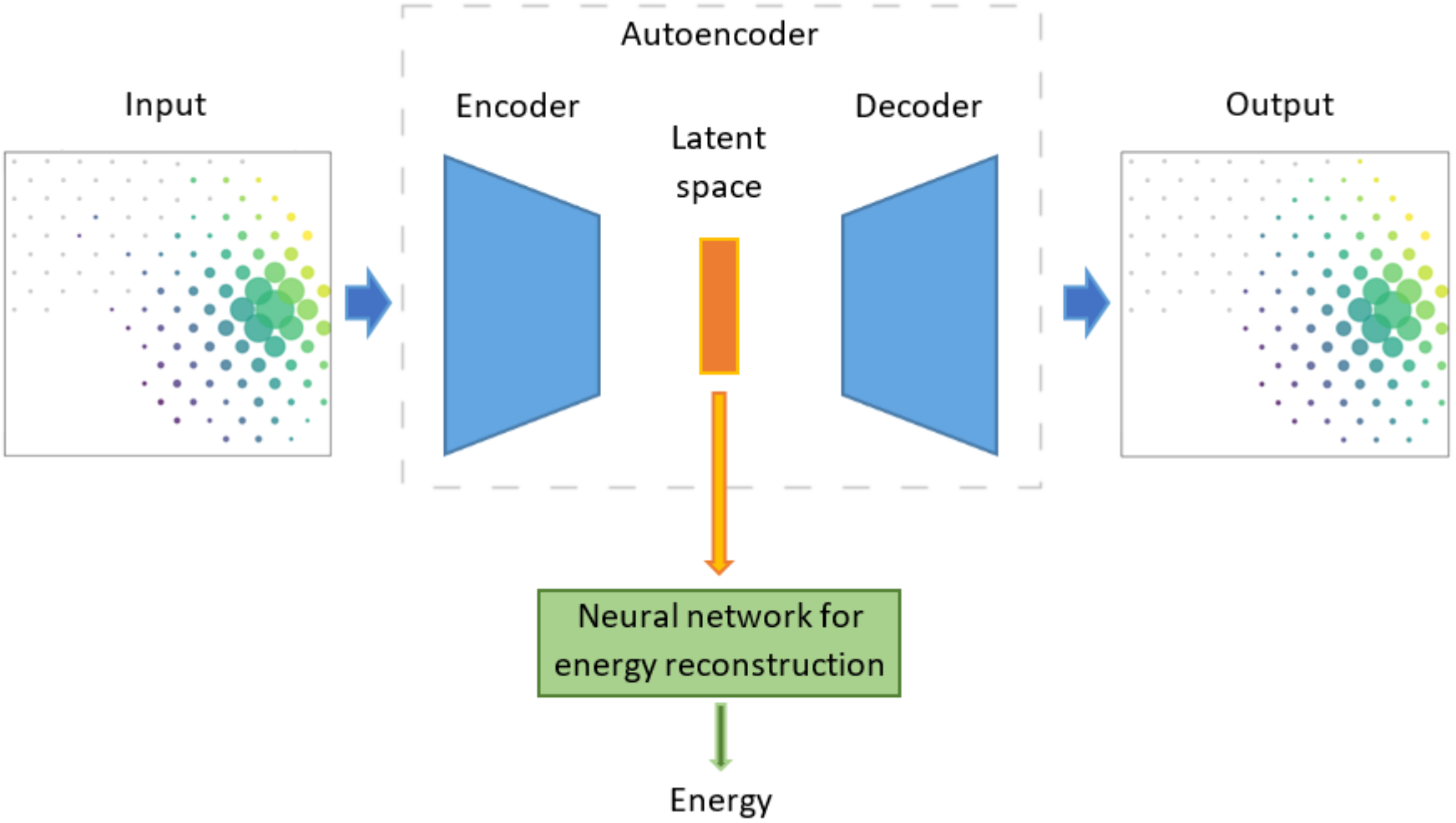
Thus, we train an autoencoder (AE) to extract latent features from TAIGA-HiSCORE data, then predict energy directly from this latent space

This approach opens a path to a fully data-driven reconstruction pipeline, with future extensions to particle type and direction

Additionally, these latent features can be used in multimodal analysis

General architecture of the system

The energy of the primary particle is determined using latent features of the AE



Autoencoder as a data compressor

To make an AE extract the most relevant information from its input data, we train it as an independent neural network to reconstruct the energy with minimal error, using the latent features as a bottleneck

We train our AE on augmented Monte Carlo simulation data for gamma quanta with 58,600 events

After training, to reconstruct the energy we only need the encoder that compresses the original data and forms a latent space

The dimensionality of the AE latent space (N) is a parameter to be optimized

Input data and loss function for the AE

The positions of 121 HiSCORE stations are approximated by a 17x12 rectangular grid. The input data for the AE have 4 channels:

- amplitude A (number of photoelectrons)
- average time t of photoelectrons detected by the station
- standard deviation of photoelectron detection times
- trigger indicator

The AE reconstructs all these values using the loss function:

$$L = C_t L_t + C_A L_A + C_{\text{s.d.t.}} L_{\text{s.d.t.}} + C_{\text{trigger}} L_{\text{trigger}}$$

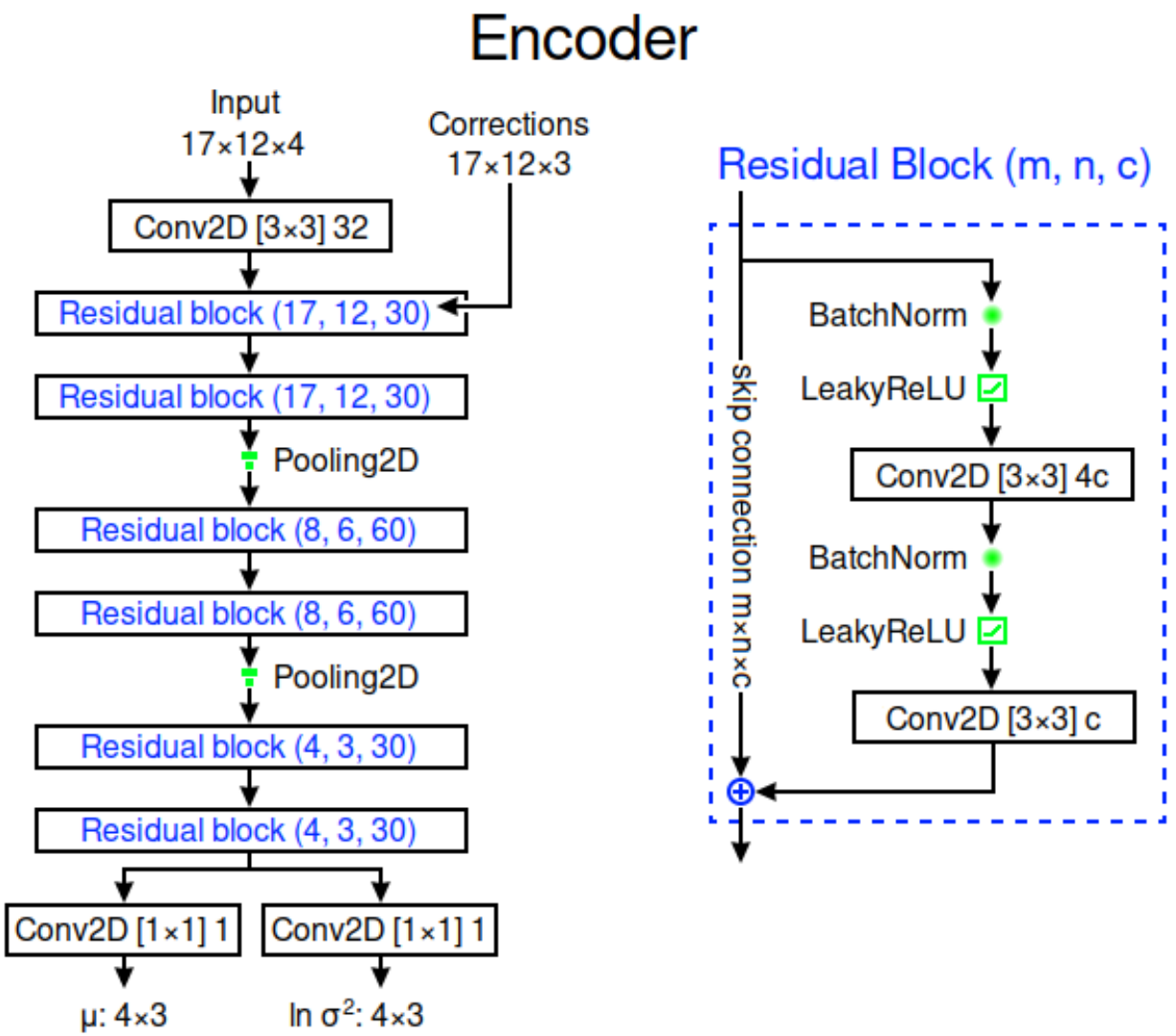
where L_{trigger} is binary cross-entropy, L_A and $L_{\text{s.d.t}}$ are masked MSE, and L_t is masked MSE multiplied by lg A

The coefficients are: $C_t = 10$, $C_A = 1$, $C_{\text{s.d.t.}} = 0.5$, $C_{\text{trigger}} = 0.0005$

Masked loss function components let AE keep nonzero values for the stations with the detected signal below the threshold (100 ph.e.) or missing stations, theoretically allowing AE to make physically plausible estimates for would-be detection times

Additionally, 17x12x3 station coordinate corrections are given as auxiliary inputs to both encoder and decoder

Architecture of the AE (technically, VAE)



The encoder and decoder each have ~750k trainable parameters

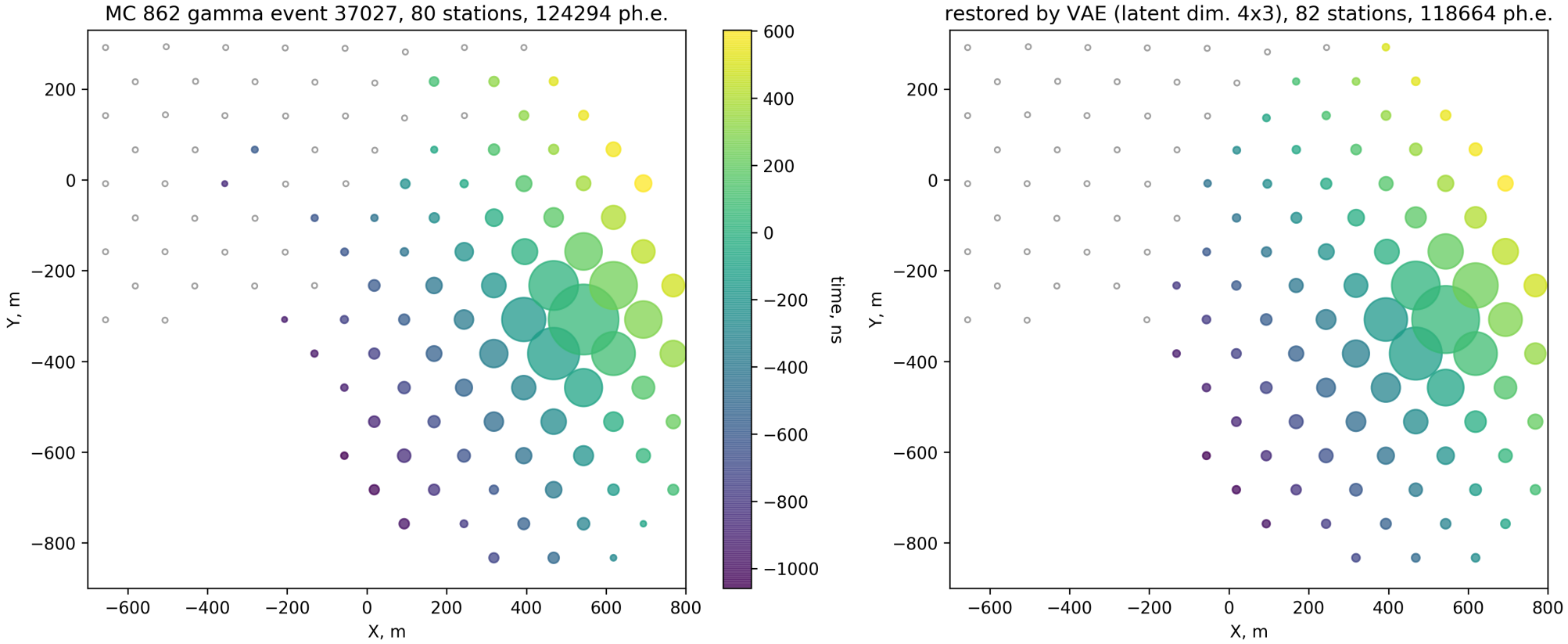
The decoder is mostly symmetrical with the encoder, with Conv2D layers replaced by Conv2DTranspose

All deconvolutional layers use 3x3 filters. The inputs are 4x3 latent features and the auxuliary data on coordinate corrections

The corrections are concatenated with the feature map at the beginning of the penultimate residual block

Results for the AE: Image reconstruction

A Monte-Carlo event restored by the AE with 12 latent parameters. Time is denoted by circle color and amplitude is denoted by size. Grey circles denote untriggered stations



Results for the AE: R^2 coefficients and errors

Coefficients of determination (R^2) for restored quasi-images depending on the latent space dimension N of the AE (N = 4, 6, 8, 12, 16, 20)

	4	6	8	12	16	20
time	0.99922	0.99973	0.99973	0.99975	0.99976	0.99976
amplitude	0.83241	0.91355	0.96229	0.97174	0.97374	0.9735
s.d. of times	0.514	0.53314	0.54794	0.58936	0.60801	0.62623
trigger	0.71928	0.744	0.74836	0.75272	0.76456	0.78352

Mean absolute errors for the AE with 12 latent features are:

amplitude: 97.1 photoelectrons
time: 5.87 nanoseconds
s.d. of time: 3.91 nanoseconds
trigger: 10.1 stations per event

Neural network for energy reconstruction

The neural network for energy reconstruction is a multi-layer perceptron

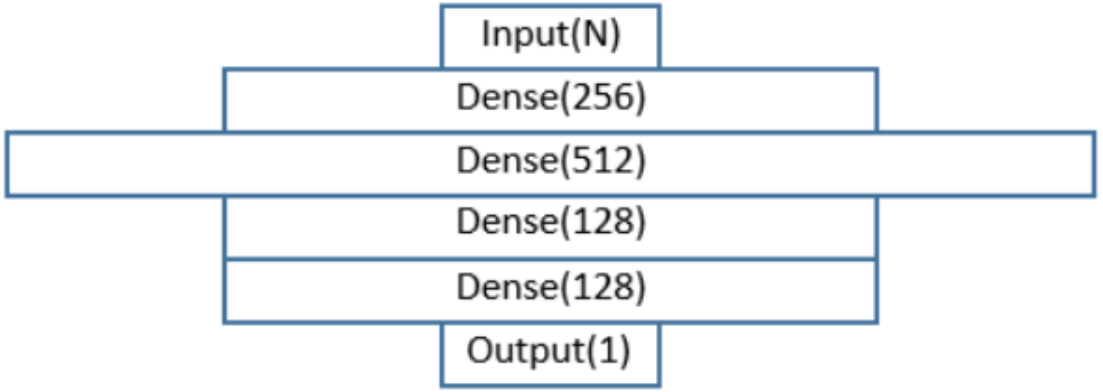
Input: N=4, 6, 8, 12, or 16 values according to the dimensionality of the latent space of the AE

Activation functions: leakyReLU

Loss function: MSE

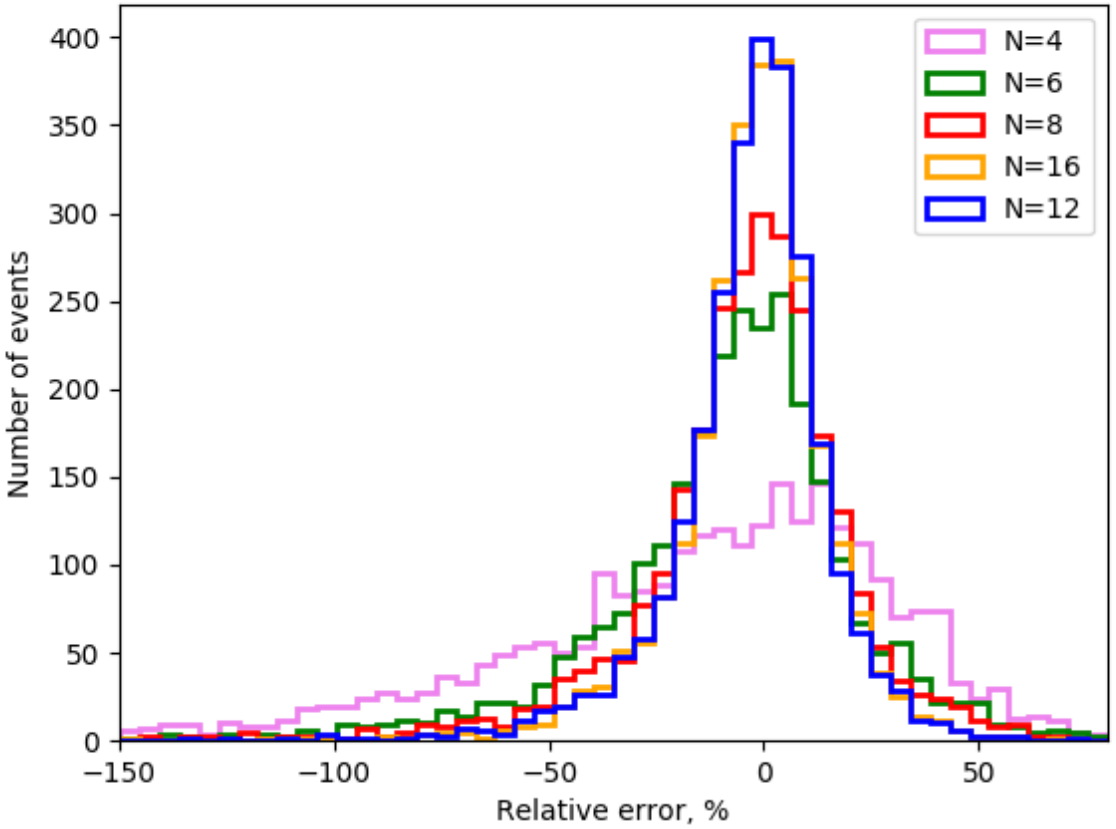
Output: the energy of the primary particle (1 value)

Network architecture



We train this network on 35,500 sets of latent features along with the energy values of the corresponding Monte Carlo gamma quanta events

Selecting the dimension of the latent space



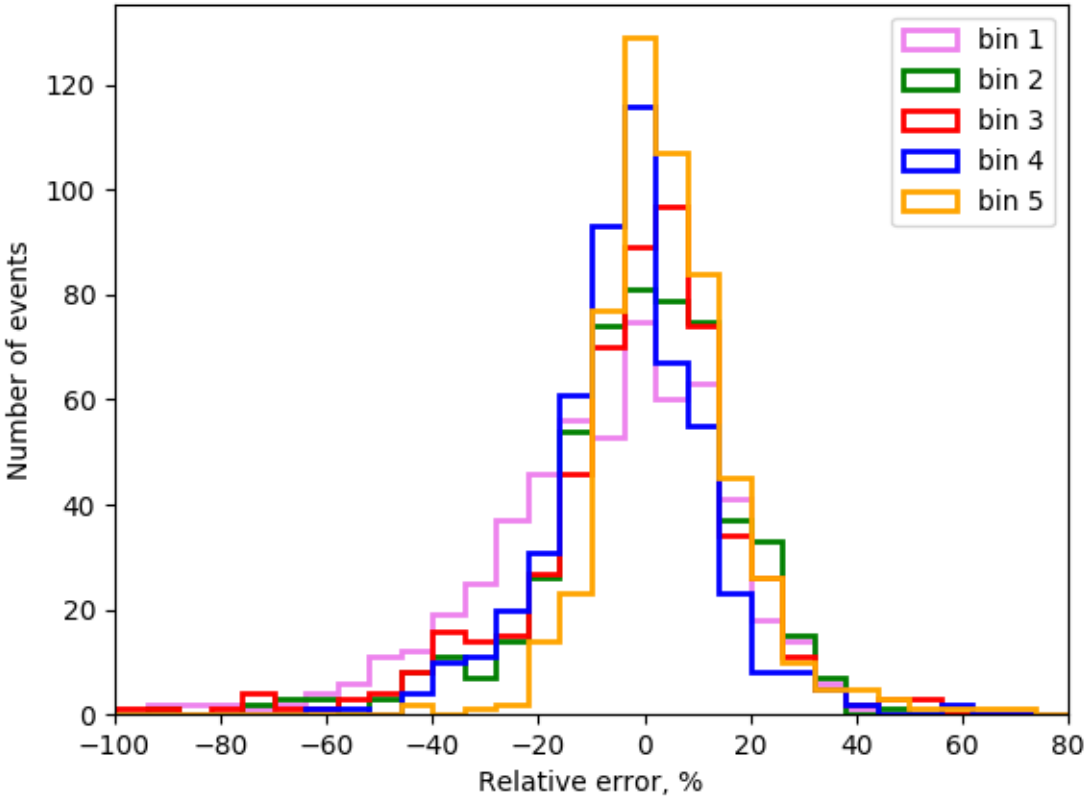
The distribution of the relative error of energy reconstruction for different values of N (dimension of the latent space):

$$Relative\ error = \frac{Energy_t - Energy_r}{Energy_t} * 100\%$$

$\mathrm{Energy_t}$ – real energy value
$\mathrm{Energy_r}$ – predicted (reconstructed) energy value

	N=4	N=6	N=8	N=12	N=16
mean(Relative Error), %	-20.4	-9.4	-4.7	-2.8	-2.4
std(Relative Error), %	52.5	34.4	25.1	18.9	19.1

Energy reconstruction for different energy bins

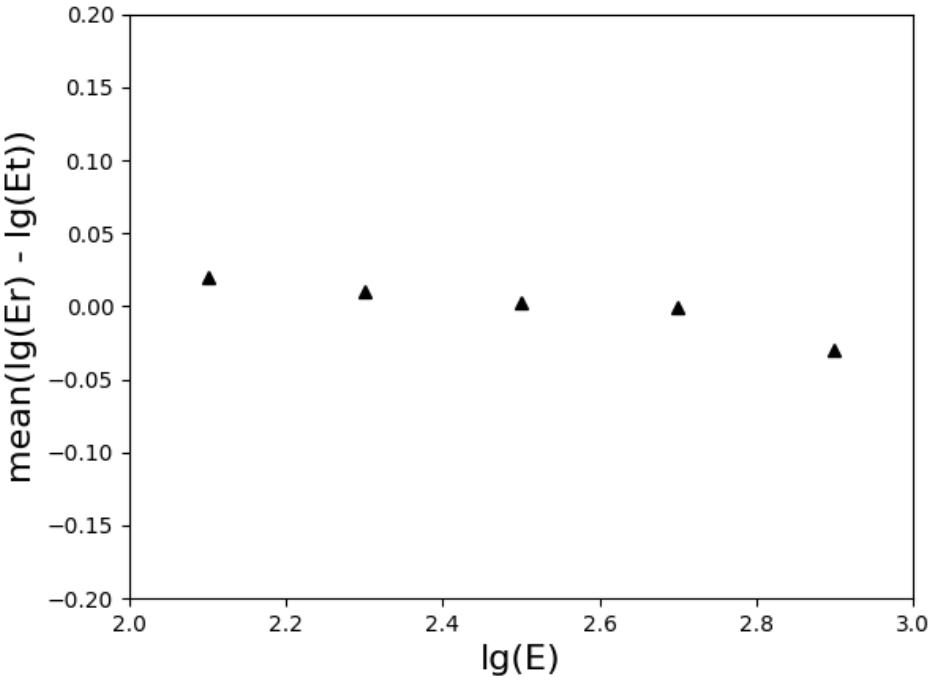


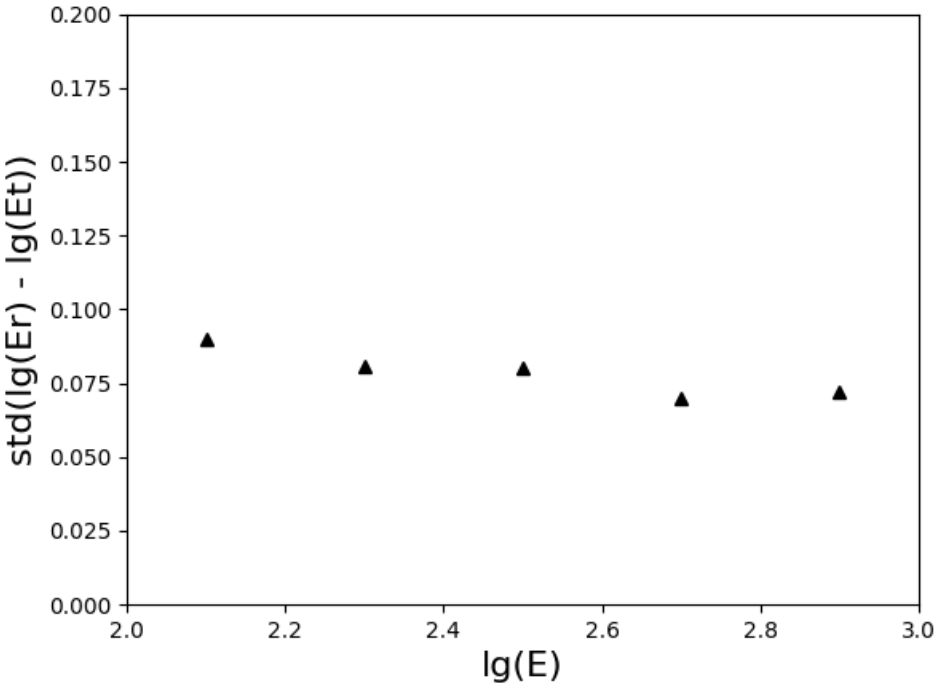
The distribution of the relative error of energy reconstruction for N=12 for different energy bins

The energy bins are intervals used to group energy values, the bin widths are chosen so that each bin contains approximately the same number of events

	bin 1 100-160 TeV	bin 2 160-260 TeV	bin 3 260-420 TeV	bin 4 420-650 TeV	bin 5 650-1000 TeV
mean(Relative error), %	-8.1	-1.2	-2.3	-2.7	4.7
std(Relative error), %	24.2	19.2	20.4	14.7	13.1

Energy reconstruction: Comparison with the results obtained using conventional method





Following standard methodology, the absolute error is calculated as:

$Error = \lg(E_r) - \lg(E_t)$, E_r – reconstructed energy value, E_t – real energy value

Our results for the mean absolute error and its standard deviation match the conventional method's values, while demonstrating even better performance at low energies. This is a promising result, however, the task of comparing the results requires further study

Conclusion

We demonstrate that the latent space of the autoencoder can be used to reconstruct the parameters of the primary particle

Using the example of determining the energy of the primary particle for TAIGA-HiSCORE data, it is shown that the latent space dimension of 12 features is a reasonable choice

The standard deviation of the relative energy reconstruction error is 20-25% for energies from 100 to 400 TeV and 13-15% for energies from 400 to 1000 TeV, which is comparable with the results obtained by the conventional methods

We expect that the developed methods will be useful in multimodal analysis, which will further improve the accuracy of energy reconstruction

Thank you for attention!