Bachelor Defence

Fréchet Distance Evaluation of Generative Models for Calorimeter Shower Simulations

7.10.2022 // Nana Marie Werther

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Motivation

- Limitations of the Standard Model
    - Gravitational force not described
    - Hierarchy problem
    - 95% unexplained Dark Matter
    - Neutrinos not massless

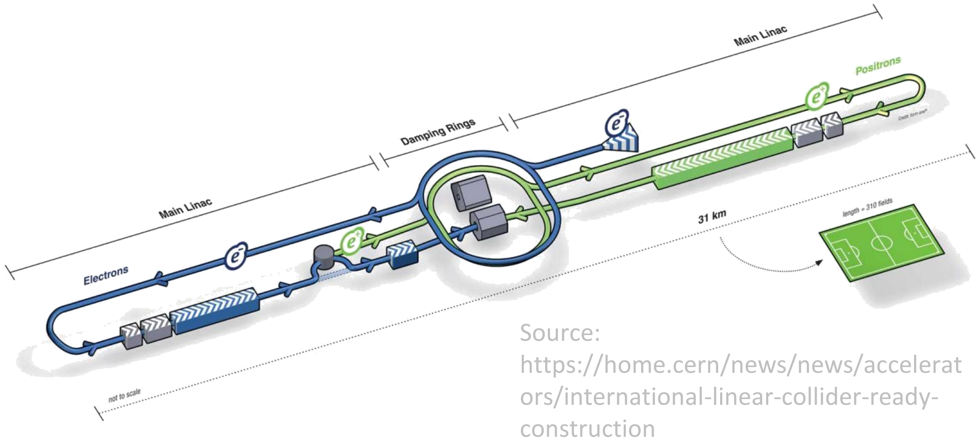
- Linear accelerator experiments for precision measurements (e.g. planned highly granular International Large Collider)

- Simulations for comparing theories with experimental data

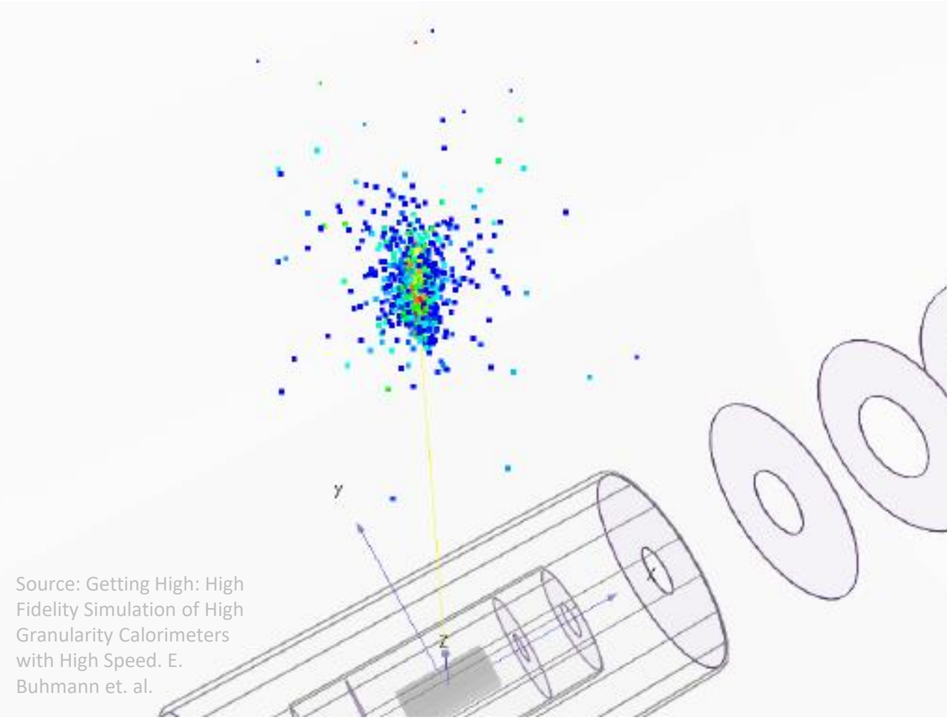
- Production of simulations increasingly costly
    - Higher luminosity
    - Larger amounts of pile-up

- Simulations amplified by Generative Adversarial Networks (GANs), currently evaluated by qualitative methods

→ Casestudy: Quantitative evaluation using the Fréchet Regression Distance (FRD)



Source: https://home.cern/news/news/accelerators/international-linear-collider-ready-construction



Source: Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed. E. Buhmann et. al.

Simulations for Future Collider Experiments

- Particle interaction processes simulated based on theory
- Some processes not well defined

→ statements via probabilities

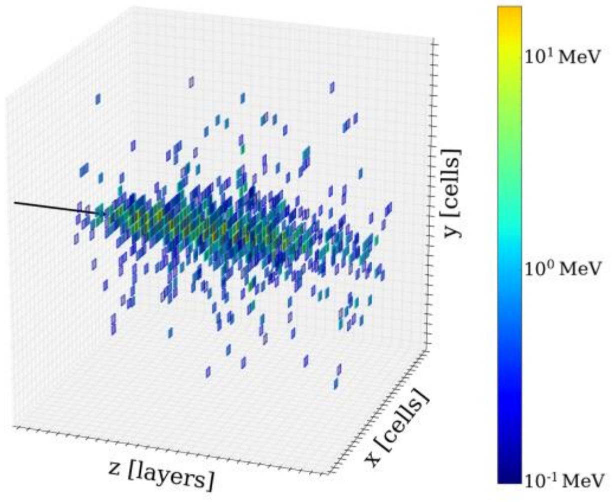
→ many simulated experiments necessary e.g. Monte Carlo Simulations

- Data sets are generated from simulated experiments
- Evaluation in unison for irregularities

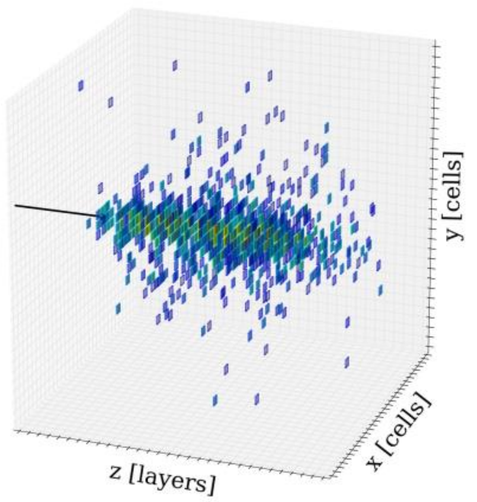
Training data set (Geant 4):

- SiW Ecal
- 30 active silicon layers in tungsten absorber stack (30x30x30 pixels)
- silicon sensor cells of 5x5 mm²

50 GeV Photon Shower Geant4



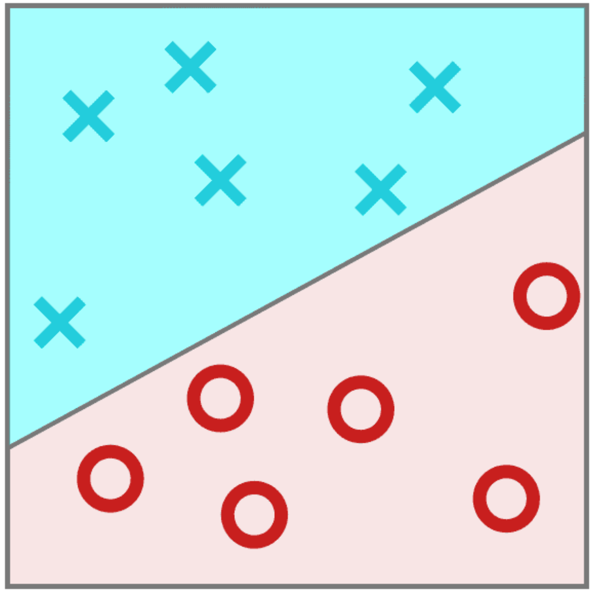
50 GeV Photon Shower BIB-AE



Source: Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed. E. Buhmann et. al.

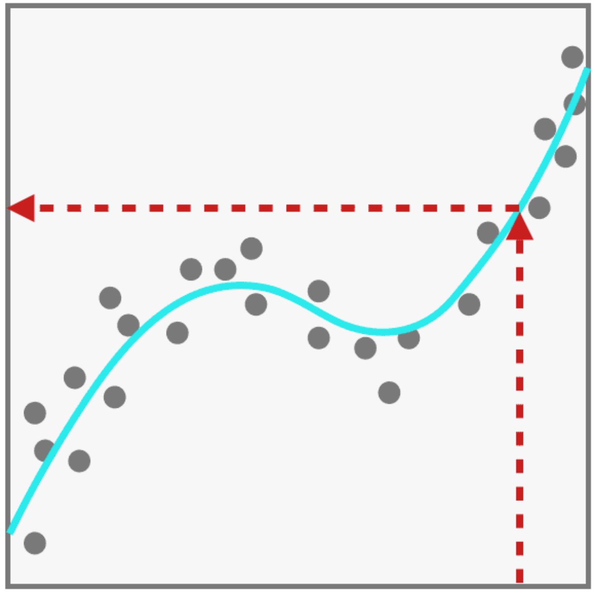
Machine Learning

Classification



Classifies input values into set classes. E.g. Inception V3

Regression



Regression networks predict an output (numeric value) based on given input.

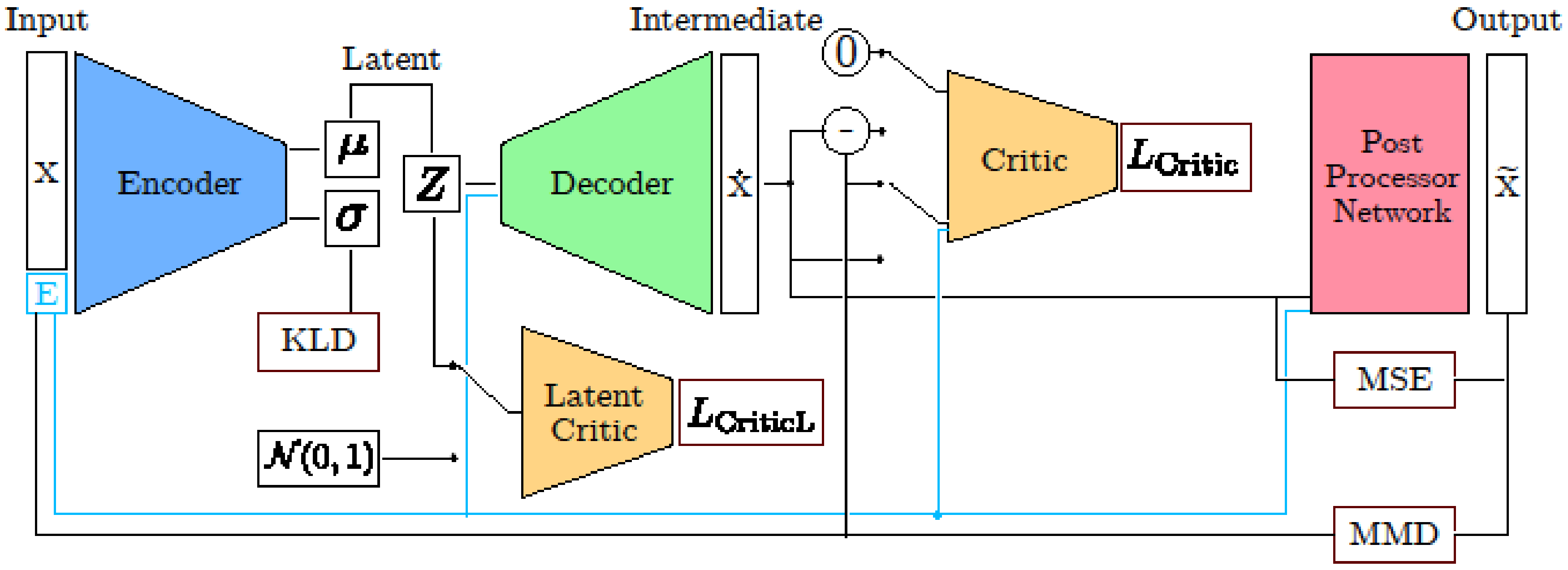
Source: https://www.sharpsightlabs.com/blog/regression-vs-classification/

Generation

Generation networks generate similar but not identical based on training data. E.g. VAE, DCGAN

Bounded-Information-Bottleneck Autoencoder (BIB-AE)

- Generative model to **accurately model of calorimeter simulations**
- Combination of GAN and Variational Autoencoder (VAE) and a few additional concepts
- Encodes input photon showers into latent space, newly generated showers are sampled from latent space
- Post Processor network for fine-tuning hit energies
- No explicit loss function as with the VAE

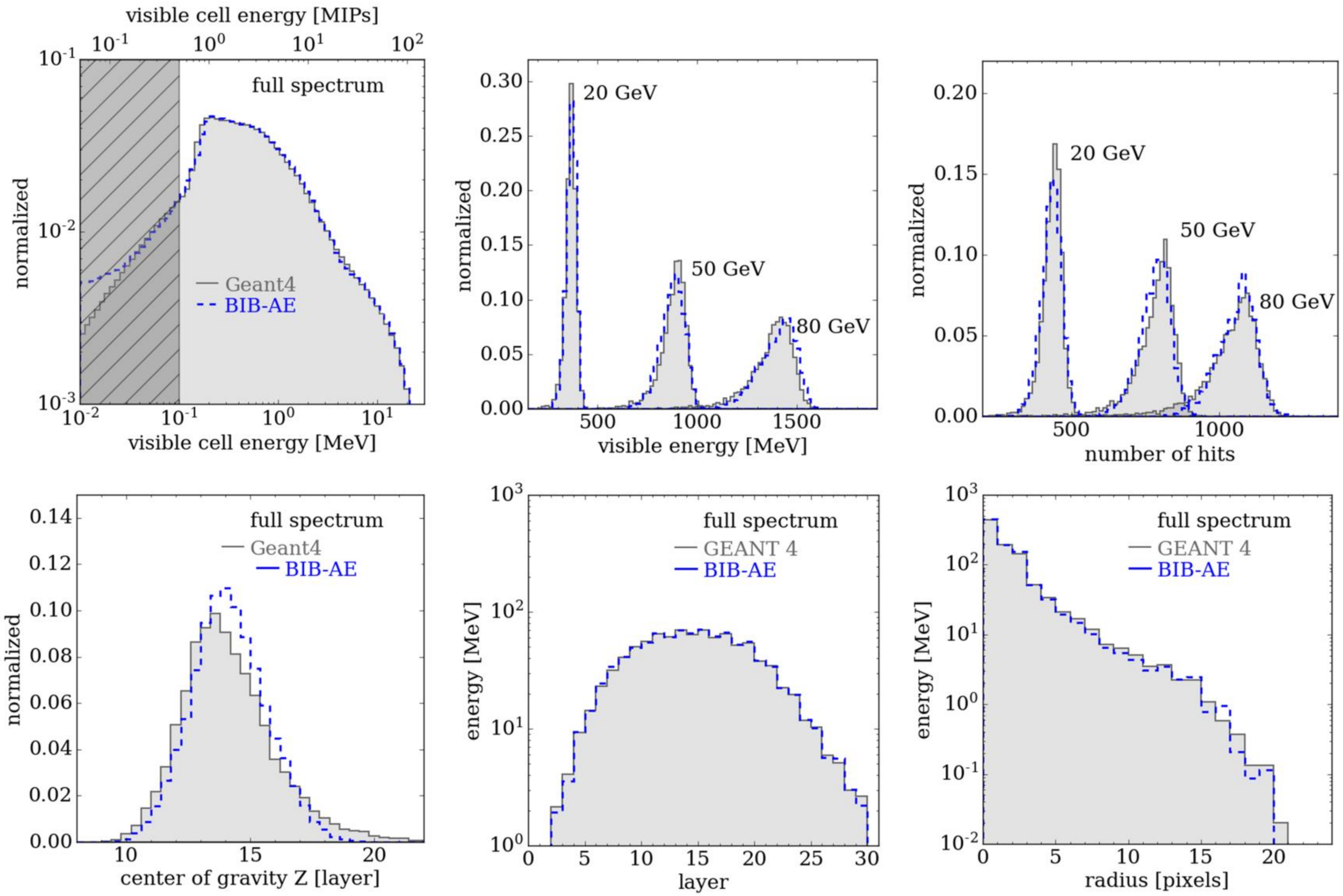


Source: Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed. E. Buhmann et. al.

Validation of Generative Networks

Previously
Generative Networks validated by (visual) qualitative evaluation

Process
- Generate large amount of showers after each epoch
- Visual (biased) evaluation of e.g. histogram distributions
- Automated qualitative evaluation with Fidelity Score (FS)



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Validation of GANs: Fidelity Score

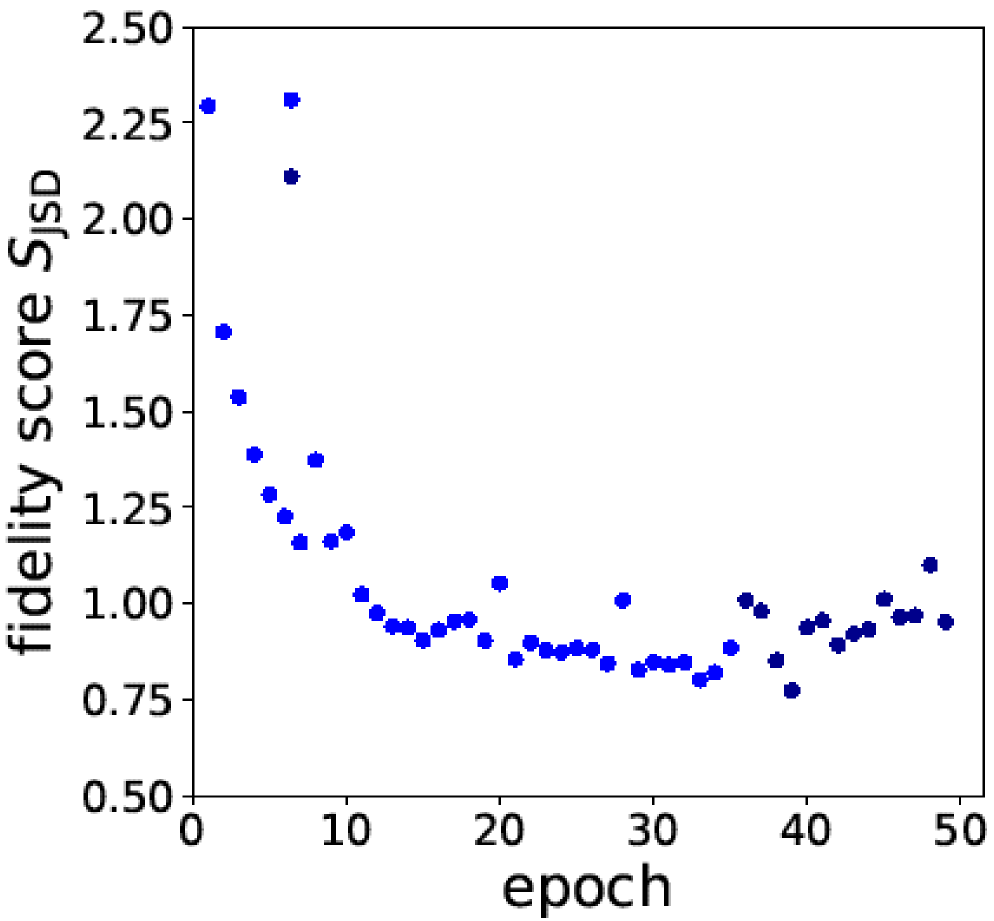
- Fidelity Score summarizes the models performance across several relevant observables
- Weighted sum of Jensen-Shannon divergance (JSD) between Geant4 truth and generations results

JSD(P || Q) = 1/2 * KL(P || M) + 1/2 * KL(Q || M)

M = 1/2 * (P + Q)

KL: Kullback-Leibler Divergence

"||" operator indicates "*divergence*"



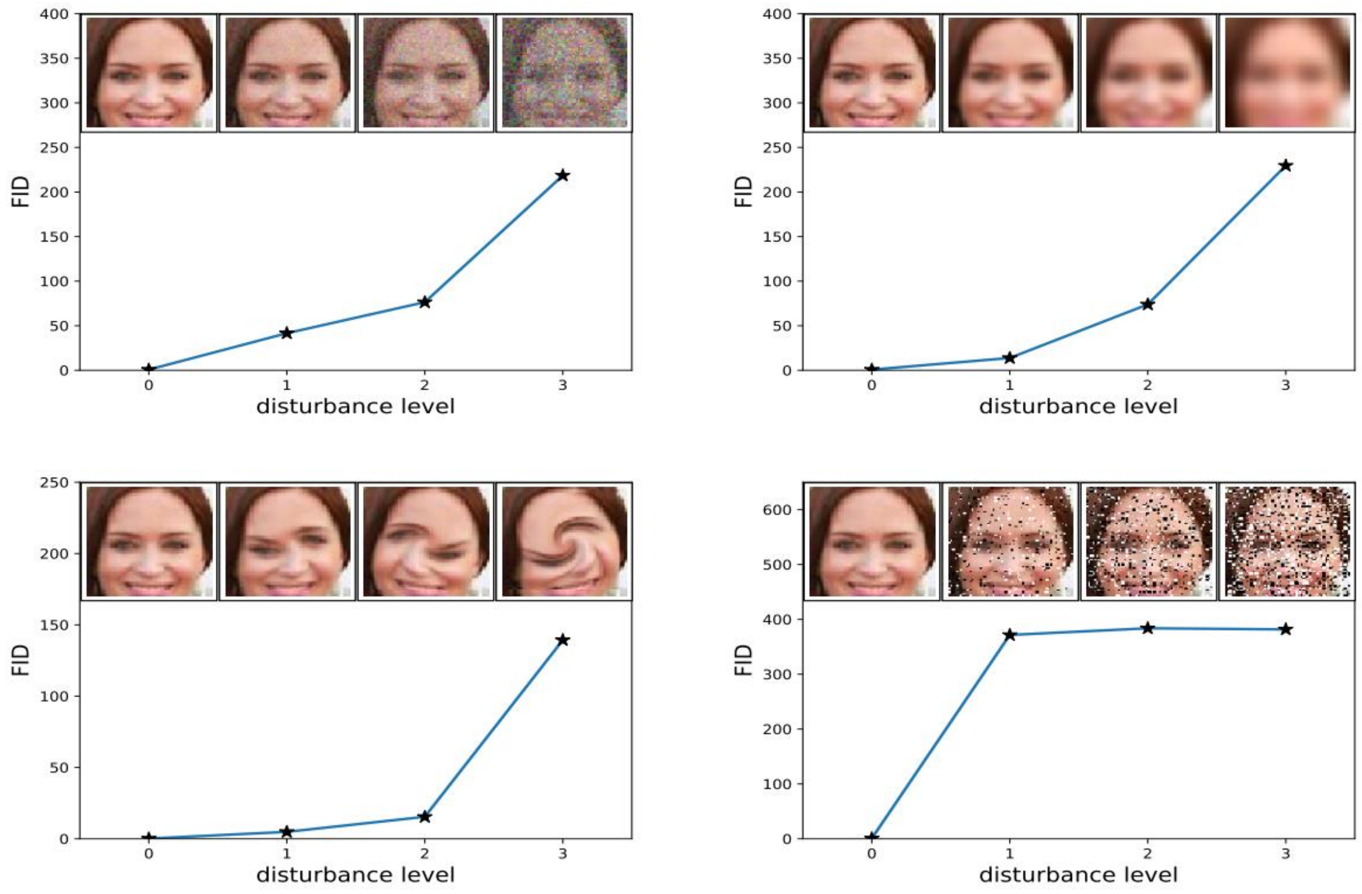
Quantitative Validation of GANs: Fréchet Inception Distance (FID)

Fréchet Inception Distance

- Utilizes Inception V3 network to calculate the FID score using computer vision specific features (activations) from second to last layer (global spatial pooling layer)
- Activations summarized into multi-variant Gaussians by calculating means µ and covariance ∑
- Difference between distributions calculated using Wasserstein-2 distance (Fréchet Distance)

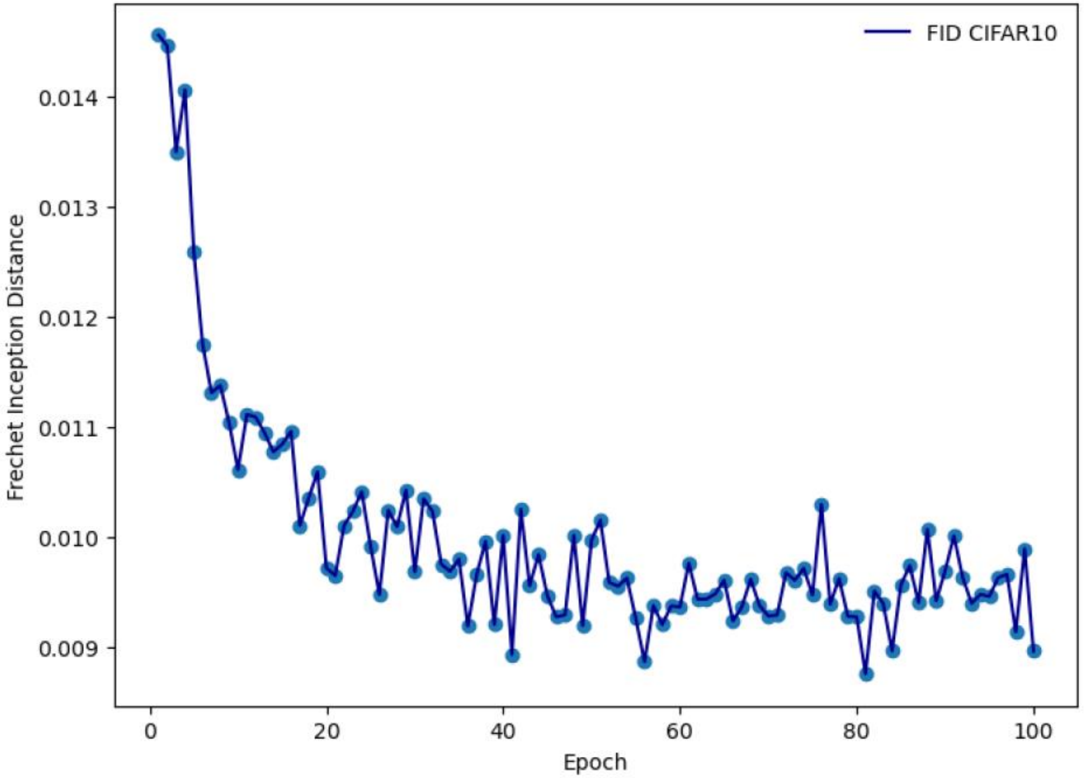
→ Score = 0.0 means images are identical

$$\mathrm{FID} = ||\mu_{\mathrm{real}} - \mu_{\mathrm{generated}}||_2^2 + \mathrm{tr}\left(\Sigma_{\mathrm{real}} + \Sigma_{\mathrm{generated}} - 2\left(\Sigma_{\mathrm{real}}\Sigma_{\mathrm{generated}}\right)^{\frac{1}{2}}\right)$$



FID for the CIFAR10 data set

- GAN trained with CIFAR10 data set
  (60.000 colored images, 10 classes, 32x32 pixels)
- Dimensionality reduction through trained Inception V3 network (2.048 activations)
- FID calculated with GAN generated images



Generated images from Epoch 5

Generated images from Epoch 74

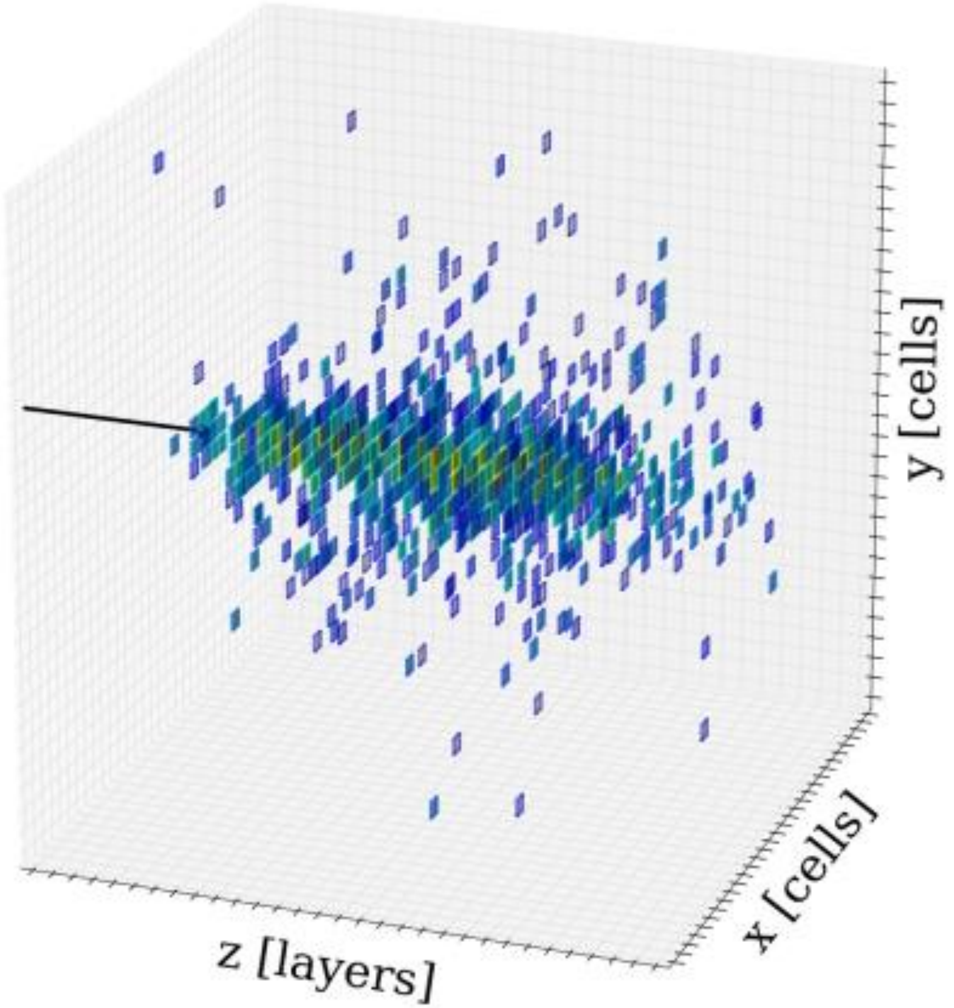
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fréchet Inception Distance for the BIB-AE data set

- Inception V3 expects 2D images with 3 color channels
- The 49.800 images of the test set generated by BIB-AE are 3D

→ Dimensionality reduction by summing over the z-axis
→ Color channel added

- Applying Inception V3 leads to 2.048 activation features to calculate FID score

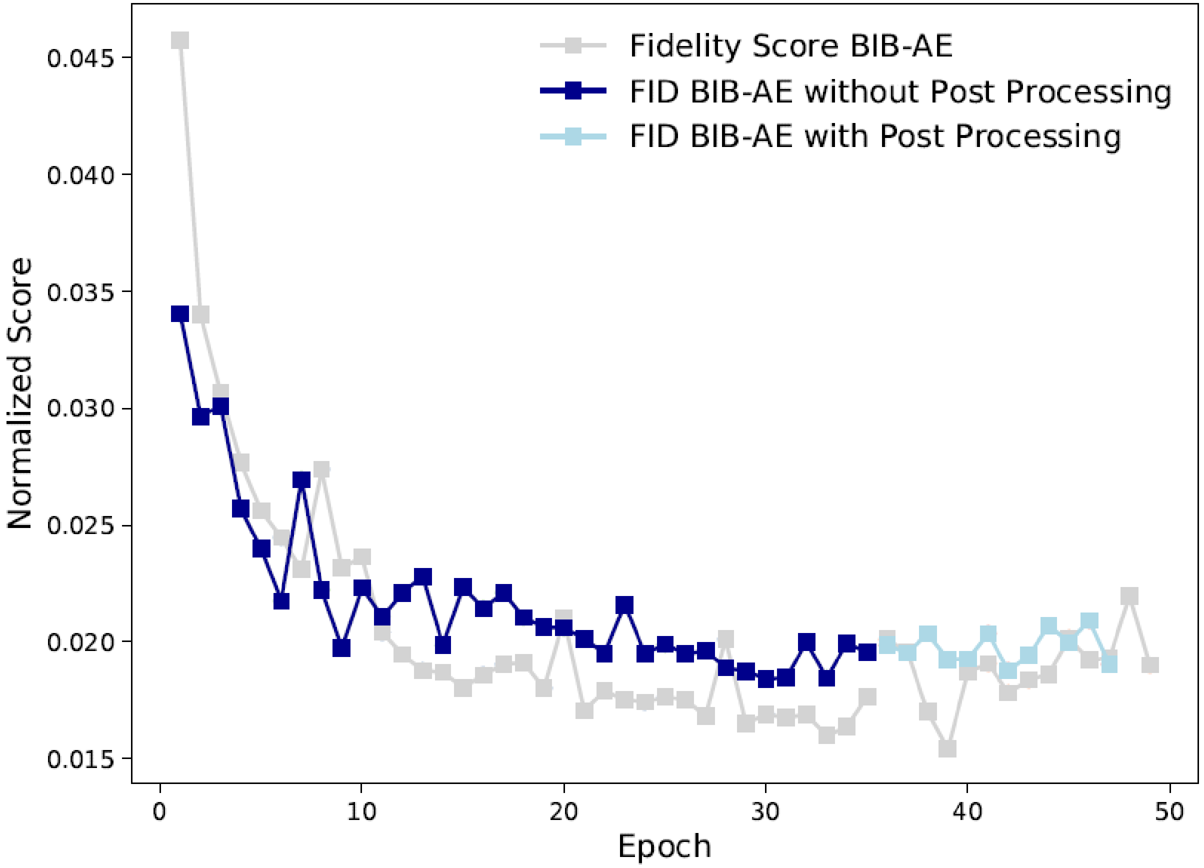


Fréchet Inception Distance for the BIB-AE data set

- Results generally correlate with Fidelity Score
- Downward progression can be observed but previosly chosen best Epoch 39 can not be identified by local minimum

→ Summing over z-axis leads to loss of information

FID Score for BIB-AE data set **summed over z-axis**
[49800, 3, 30, 30]



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

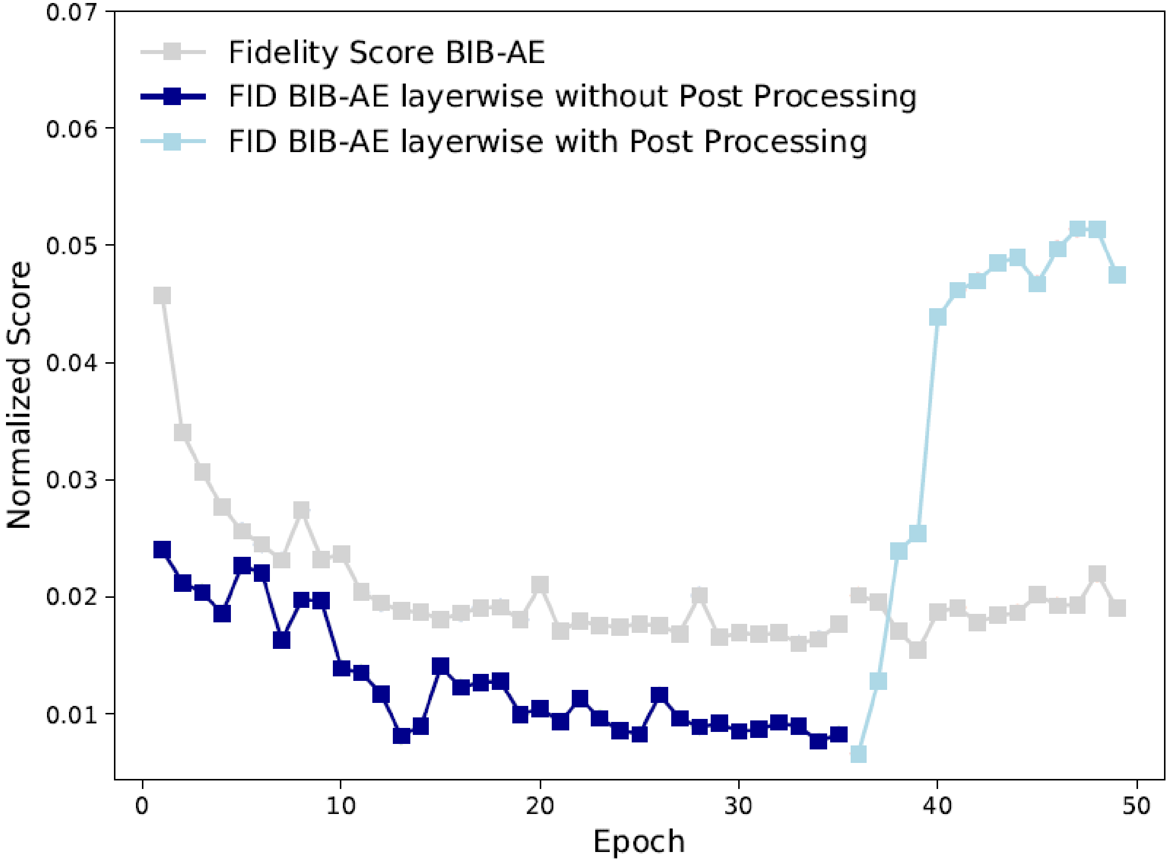
Fréchet Inception Distance for the BIB-AE data set

FID score now calculated layerwise:

- Iterating over z-axis, calculating score for each layer
- Mean FID score calculated from all layers

→ Unexpected FID score for later Epochs
→ FID drastically increases as soon as Post Processing is applied
→ Investigation of a physically better motivated network

Mean FID Score for BIB-AE data set **iterating over z-axis** [49800, 3, 30, 30, 30]



Fréchet Regression Distance (FRD)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FRD for the BIB-AE data set using the Regressor

Layer Type	Size	Parameters	Comment
Convolutional layer	16x(3,3,3)	432	no bias
Layer Normalization	(14,14,14)	2,744	
Leaky ReLU			$\alpha = 0.02$
Convolutional layer	32x(3,3,3)	13,824	no bias
Layer Normalization	(6,6,6)	216	
Leaky ReLU			$\alpha = 0.02$
Convolutional layer	16x(2,2,2)	4,096	no bias
Fully connected layer	100	200,100	
Leaky ReLU			$\alpha = 0.02$
Fully connected layer	1	101	
ReLU			
Total parameters:		224,473	

Architecture of the regression network “Regressor” used in Getting High Paper

Regression Network „Regressor“

- Implemented in PyTorch
- Batch size = 64
- Learning rate (Adam optimizer) = 0.001
- L1 Loss (Mean Absolute Error)

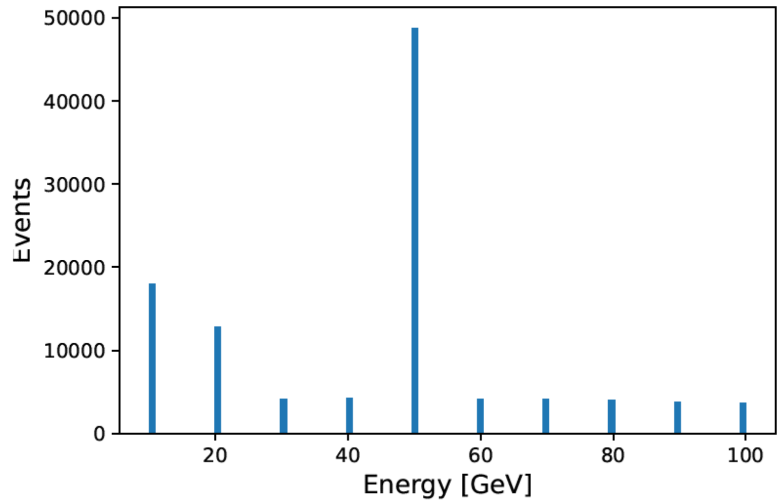
Training:

- Training set of 80.000 events
- Validation set of 20.000 events
- Energy range: 10 GeV to 100 GeV
- 200 Epochs
- Best Epoch selected by lowest validation loss

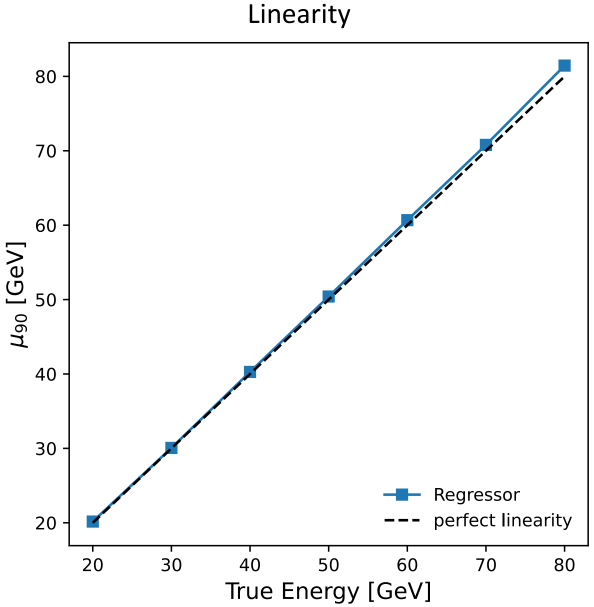
Validation of Regressor network

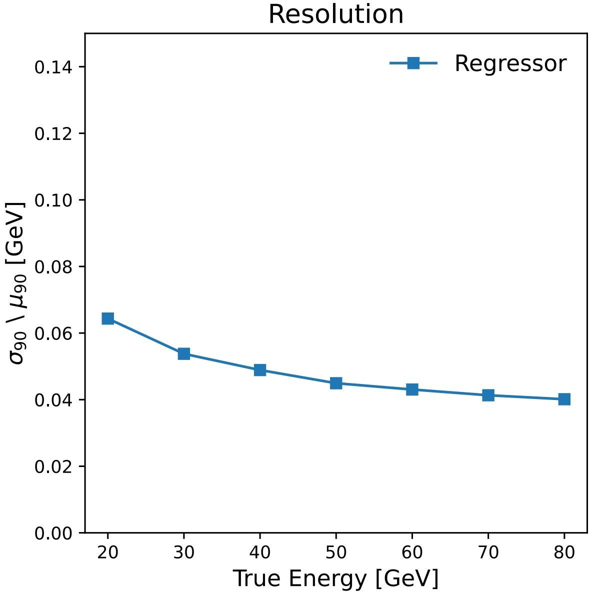
- Reconstructed Gaussian-like distributions centered around true values show ability to recognize true energies of test set
- Validate showers with discrete energies (20-90 GeV)
- Calculate mean (μ_{90}) and the root-mean-square (σ_{90}) of the 90 percent core of the distribution for all sets of showers

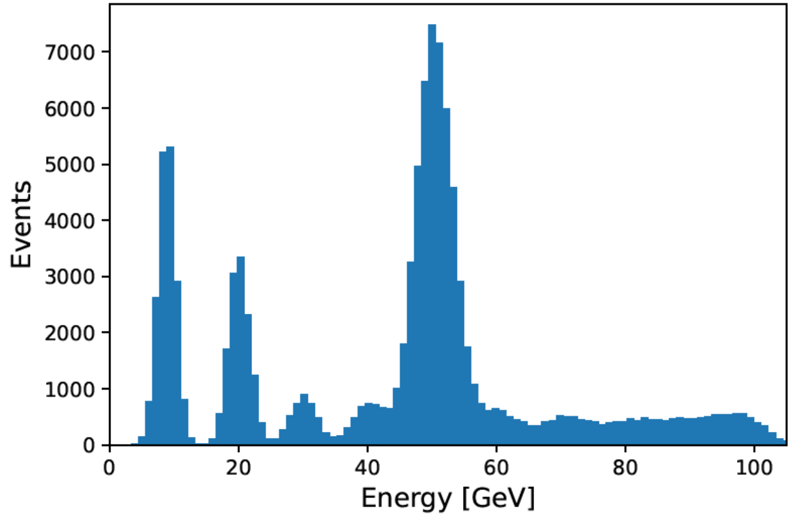
Linearity = mean plotted against true energies
Resolution = relative width σ_{90}/μ_{90} plotted against true energies



Incident photon energy in the test data set (Geant4)





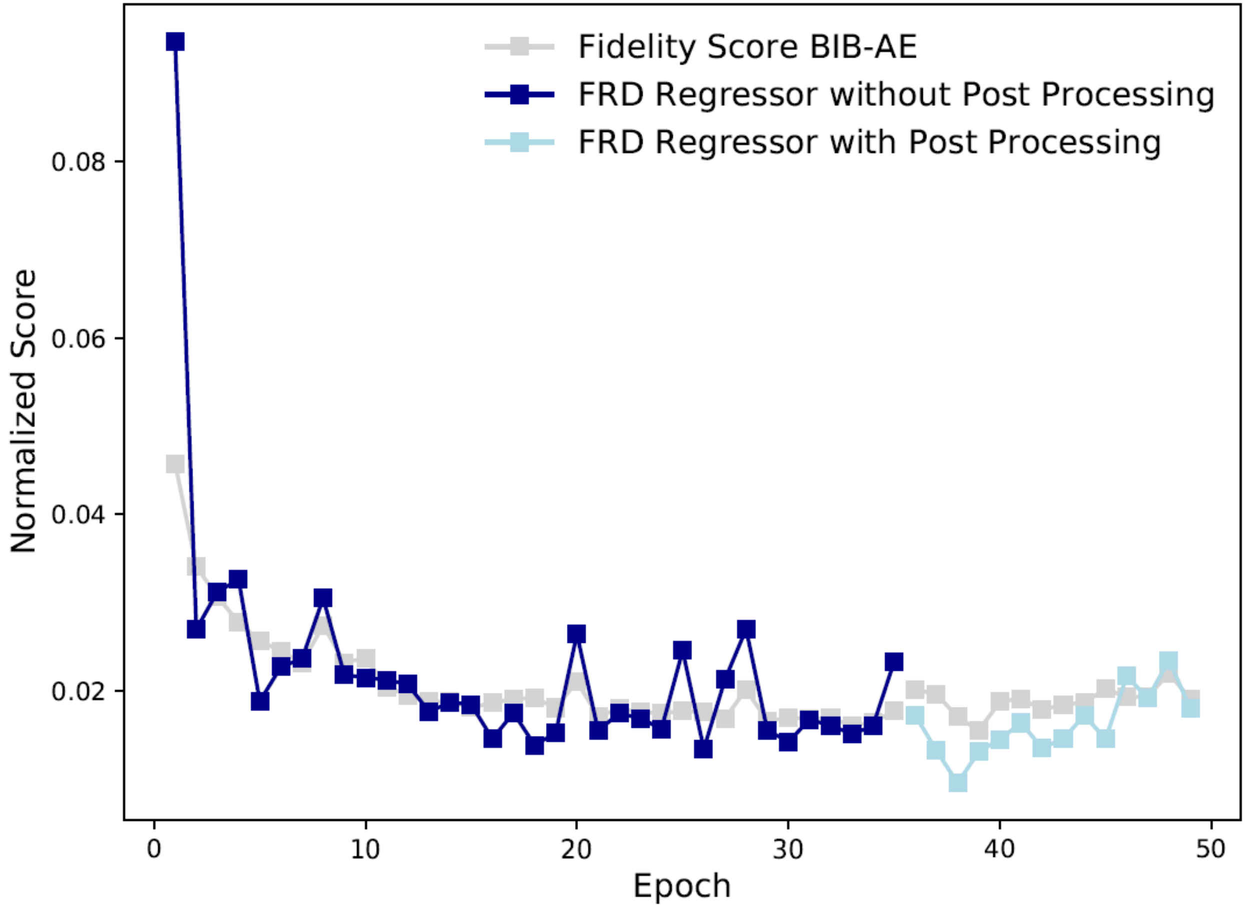


Single energy reconstruction of the test set with the Regressor

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FRD for the BIB-AE data set using the Regressor

Frechet Regression Distance for BIB-AE data using the "Regressor" regression network



Calculated with second to last layer (global special pooling layer with 100 activations)

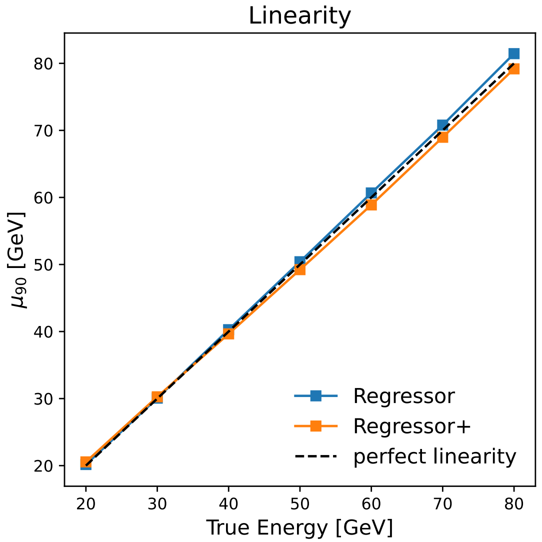
FRD Score:

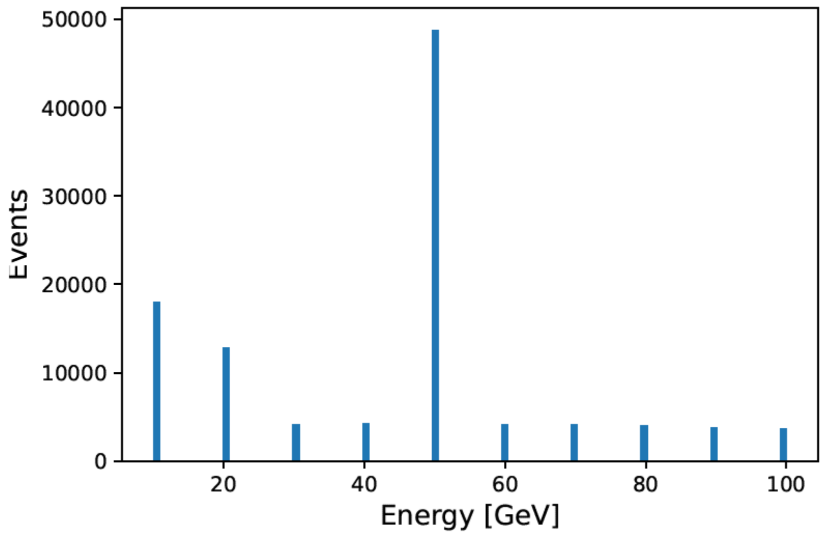
- Steep decrease during first 15 Epochs
- Slight oscillation from Epoch 15 to 35
- From Epoch 36 (Post Processing) further decrease and minimum at Epoch 38

→ Regressor picks up on best few Epoch (previous best Epoch 39) but does not reach quite the same results

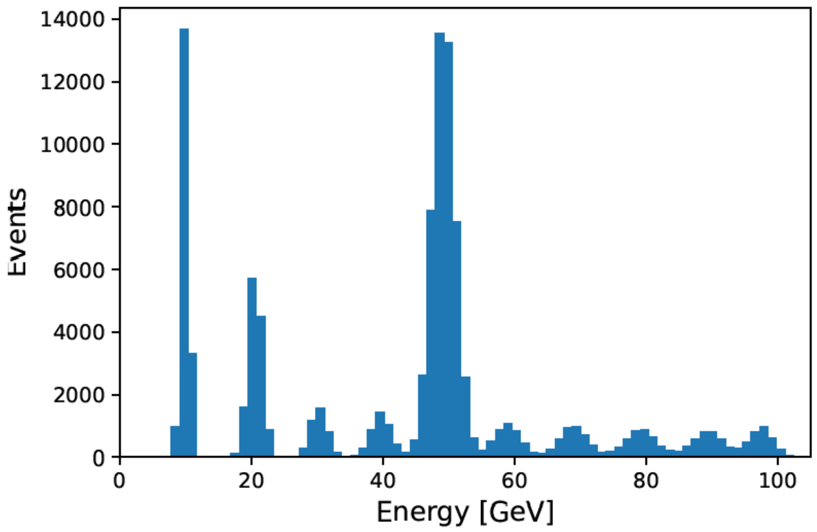
Validation of Regressor+ network

- Regressor+ with more Layers and Batch Normalization
- Reconstruction of the energy labels by the Regressor show the networks ability to recognize the true energies of the test set
- Linearity and Resolution better then Regressor

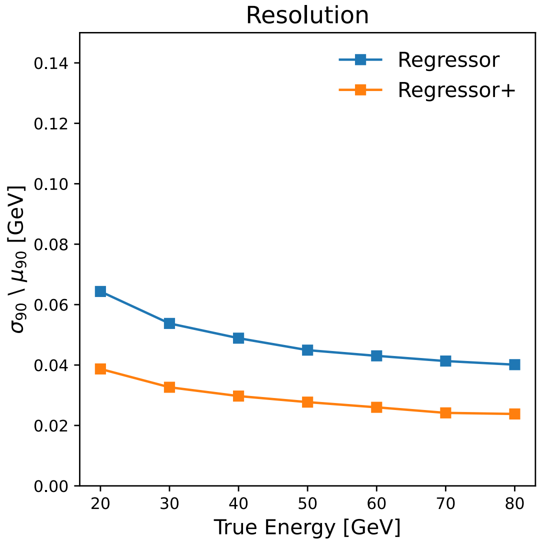




Incident photon energy in the test data set (Geant4)



Single energy reconstruction of the test set with the Regressor+



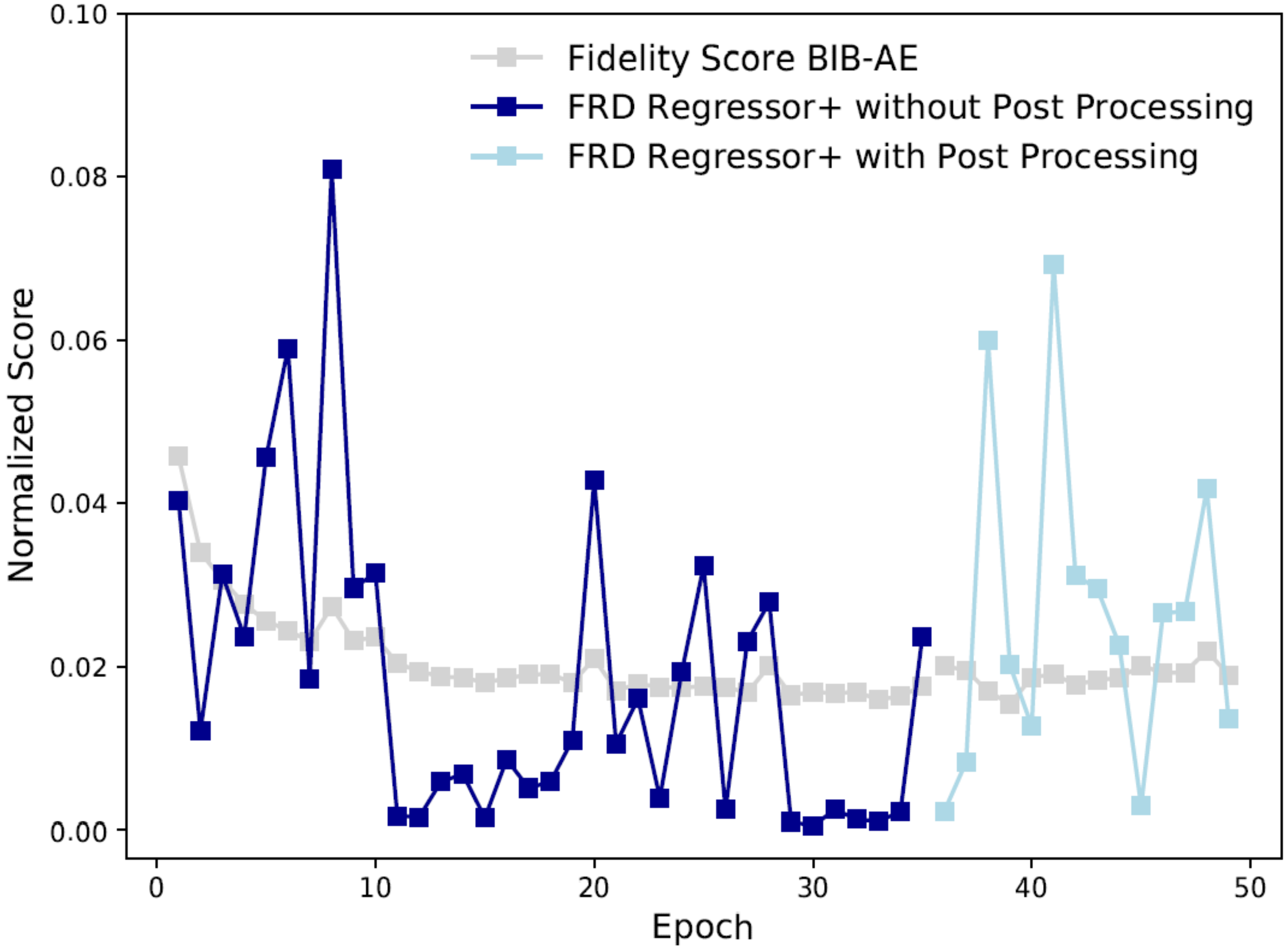
FRD for the BIB-AE data set using the Regressor+

FRD Score:

- Noisy until Epoch 10
- Then decreases while remaining noisy
- With Post Processing on average higher values with distinctly higher valued outliers

→ Regressor+ did not give better estimate of generation fidelity

→ varying aspects of architecture tested



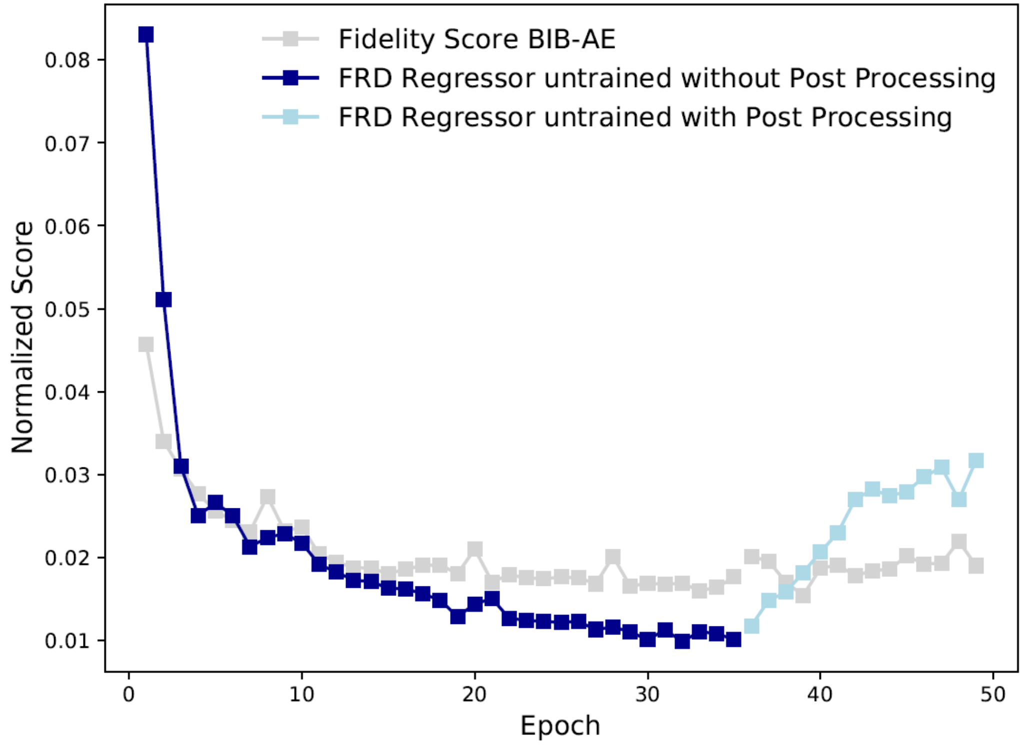
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Debugging of Regressor+

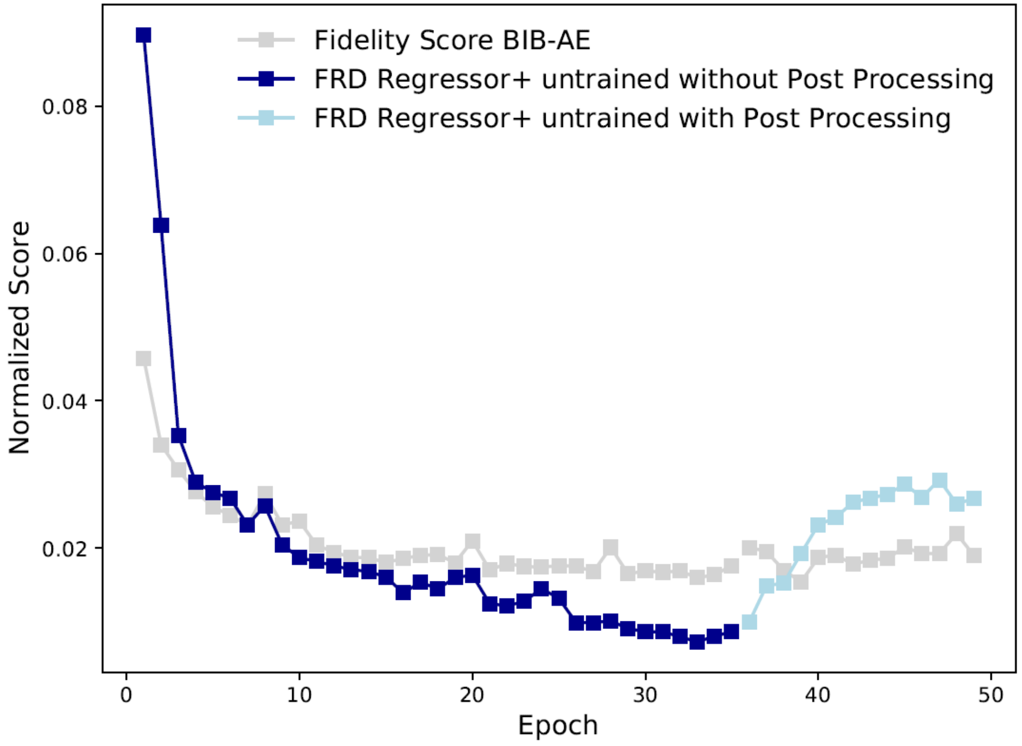
- FRD calculated for untrained Regressor and untrained Regressor+ to determine effect on outcome

→ Increase in FRD score after introduction of Post Processing otherwise seemingly good

Untrained Regressor

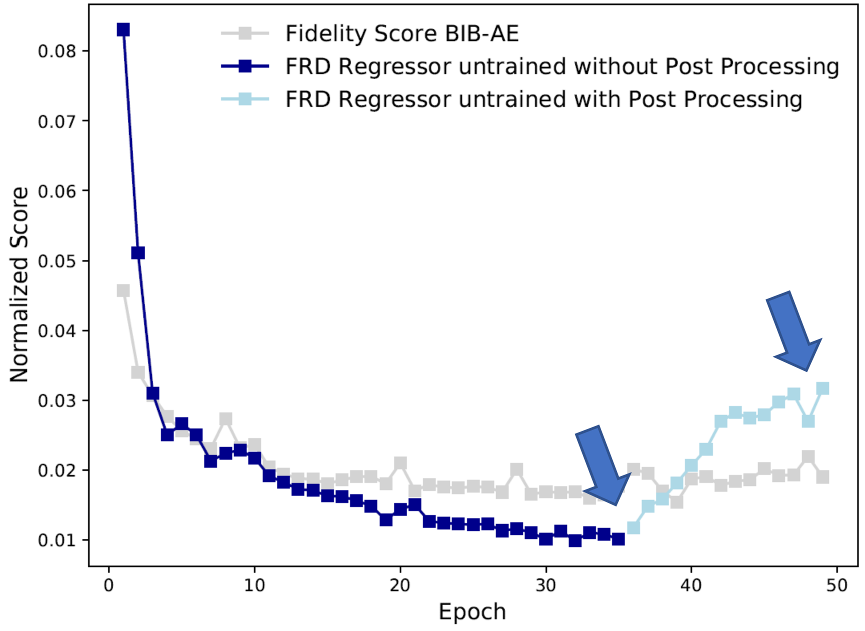


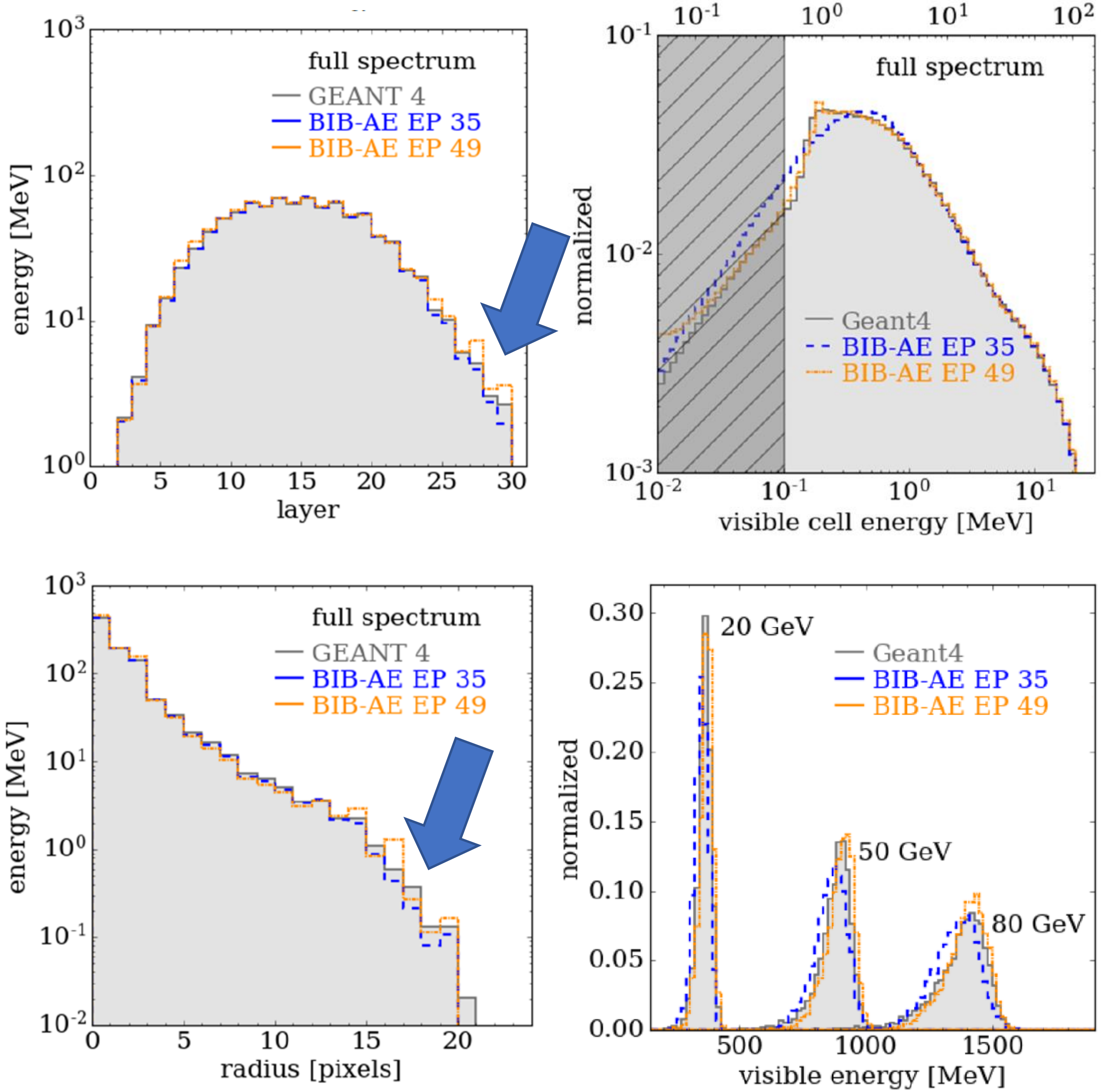
Untrained Regressor+



Debugging of Regressor+

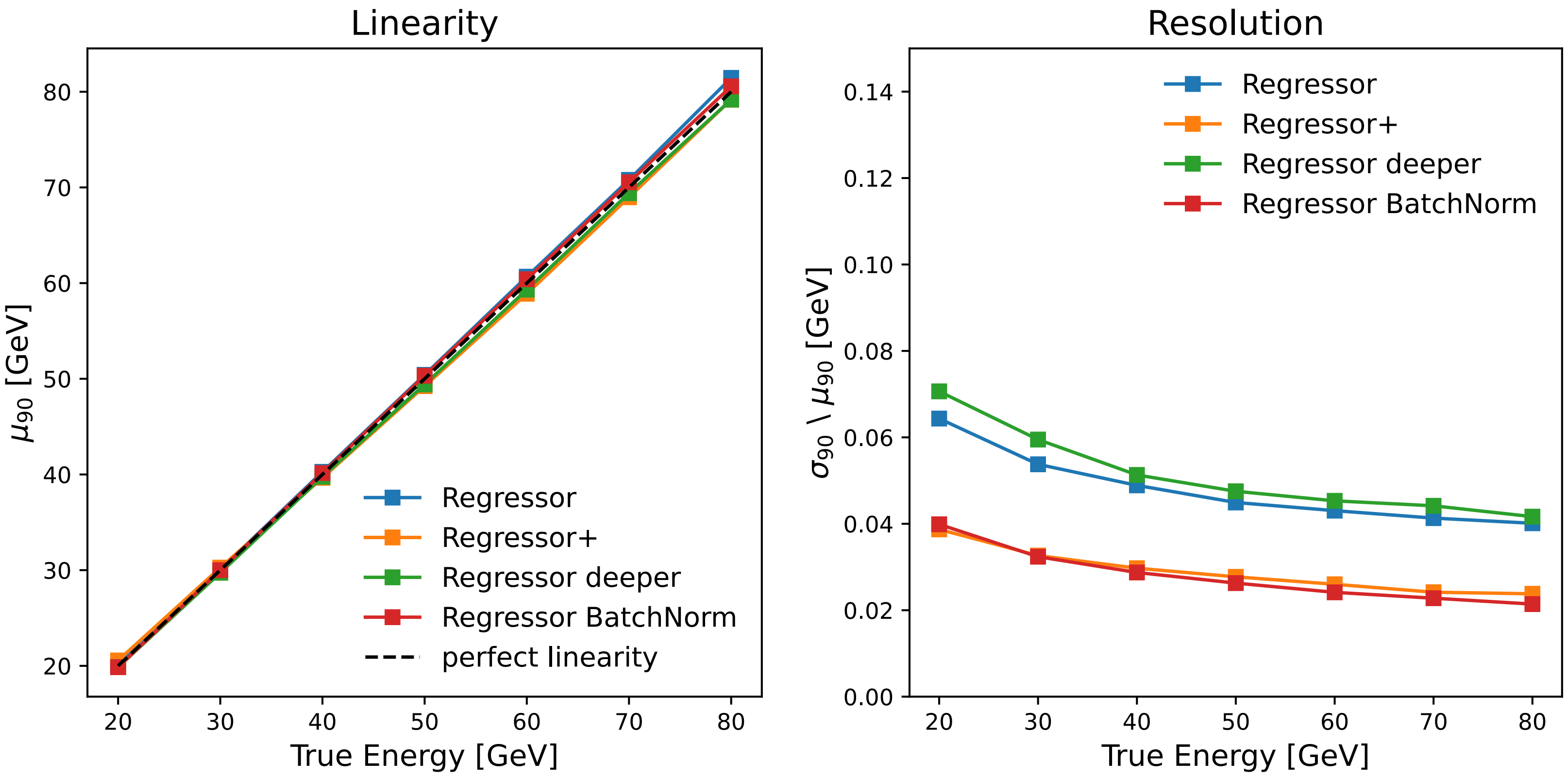
- Post Processing deposits more energy in previous underestimated rear layers and outer corner of calorimeter
- While feature like MIP peak modelled correctly, it has effect on attributes not relevant for manual observation, that Regression network takes into consideration





Debugging of Regressor+

Investigating further by adjusting aspects of the Regressor:



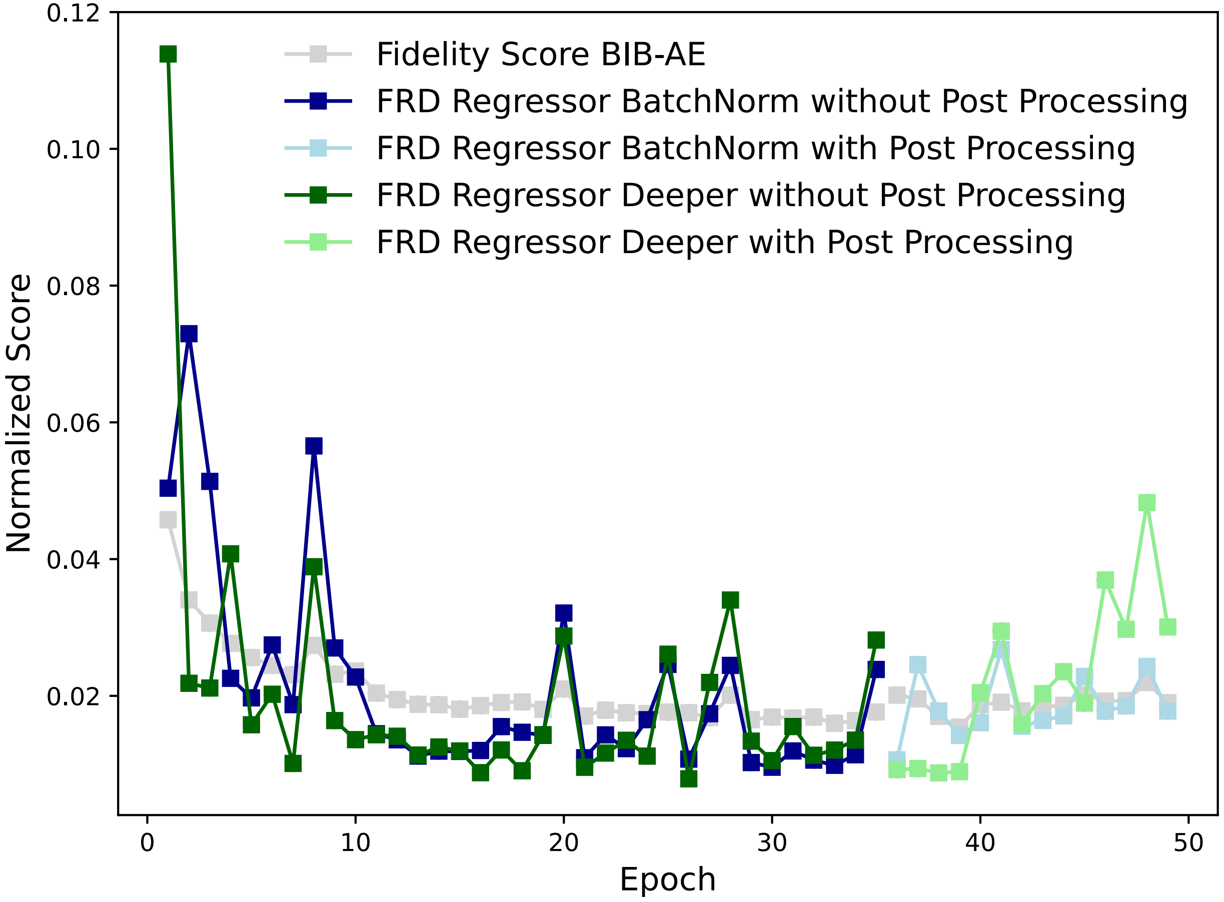
→**Regressor+** has higher resolution then Regressor

→**Regressor with Batch Normalization** has best resolution

→**Deeper network** has worst resolution

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Debugging of Regressor+



Regressor with Batch Normalization

- FRD score noisy, with particular low value for Epoch one
- Oscillates less as soon as the post processing is applied and almost aligns with Fidelity Score
- Not able to clearly identify the previously best epoch 39 with the lowest score

Deeper Regressor

- FRD score quite noisy
- Previously best epoch 39 characterized by very low FRD value, while not the global minimum
- With Post Processing volatile but steep increase like untrained Regressor and Regressor+

Results and Conclusion

- Regressor+, especially due to Batch Normalization, provides a **higher resolution**, but appears to lead to a **worse FRD score**.
- Network might place importance on different features
- Leading to better plots of relevant observables

Possible crosschecks and solutions:

- Train just shower core [15x15x15] cells

→ Other features might gain significance

- Layerwise relevance propagation, e.g. heatmapping
- Improve generative model to yield low FRD and Fidelity Score