AI4ePIC, Data Reduction, Faster Simulation and Data Enhancements

Yihui "Ray" Ren (yren@bnl.gov)

AI-CoDesign Group Leader, AI Department, Brookhaven National Laboratory

@BrookhavenLab

ePIC Collaboration Meeting, Villa Mondragone, Monte Porzio Catone (RM), Italy, Jan. 20-24, 2025

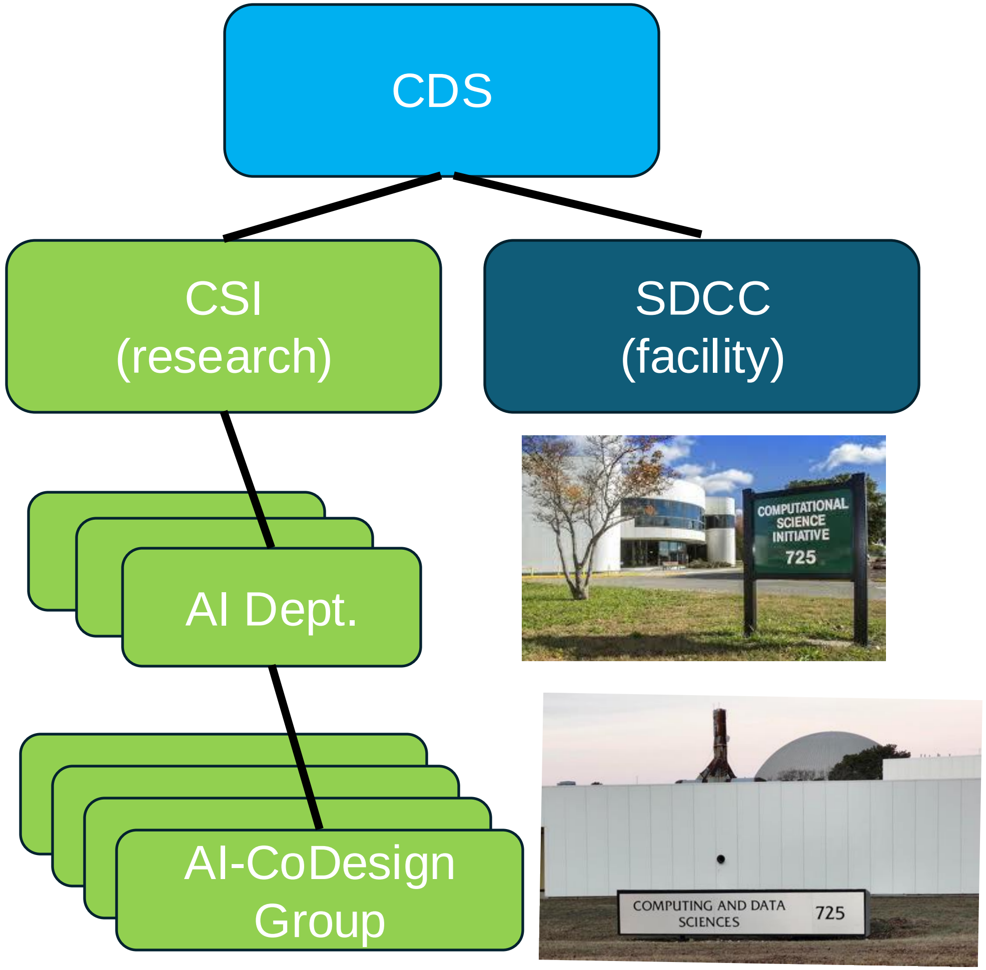

CT
NYC
Long Island
NJ
Atlantic Ocean
Relativistic Heavy Ion Collider,
future Electron-Ion Collider
(2.4 miles in circumference)
Computing and Data
Sciences (CDS)
National Synchrotron
Light Source II

Computing and Data Sciences (CDS)

AI Department is dedicated to AI research and its applications in scientific domains.

Has 27 researchers working on 49 projects collaboratively with domain scientists (as of 01/17/2025).

AI CoDesign group focuses on hardware-algorithm co-design and solving challenges in applying AI to real experiments.



Collaborators

___ highlights ePIC collaboration members

Fortunately working with:

(NP) Timothy Rinn, Yeonju Go, Evgeny Shulga, Joe Osborn, Jin Huang

(DUNE) Haiwang Yu, Brett Viren, Chao Zhang, Xin Qian

(ASIC) Soumyajit Mandal, Prashansa Mukim, Piotr Maj, Grzegorz Deptuch

(ATLAS) Elizabeth Brost, Haider Abidi, Viviana Cavaliere, Michael Begel

(CSI) Dmitrii Torbunov, Yi Huang, Shubha Khrael, Meifeng Lin, Shinjae Yoo

ePIC-related talks:

- SRO XII, 2024, "Neural Compression for sPHENIX Sparse TPC data", Link
- AI4EIC workshop, 2023, "Fast 2D BCAE for Compressing 3D TPC data", by Yi Huang. Link
- RHIC User meeting, 2023, "ML Technique Overview in Nuclear or High Energy Physics", Link
- AI4EIC workshop, 2022, "Tutorial on Graph Neural Networks" Link
- SRO IX, 2021, "Real-time machine learning at BNL CSI", Link

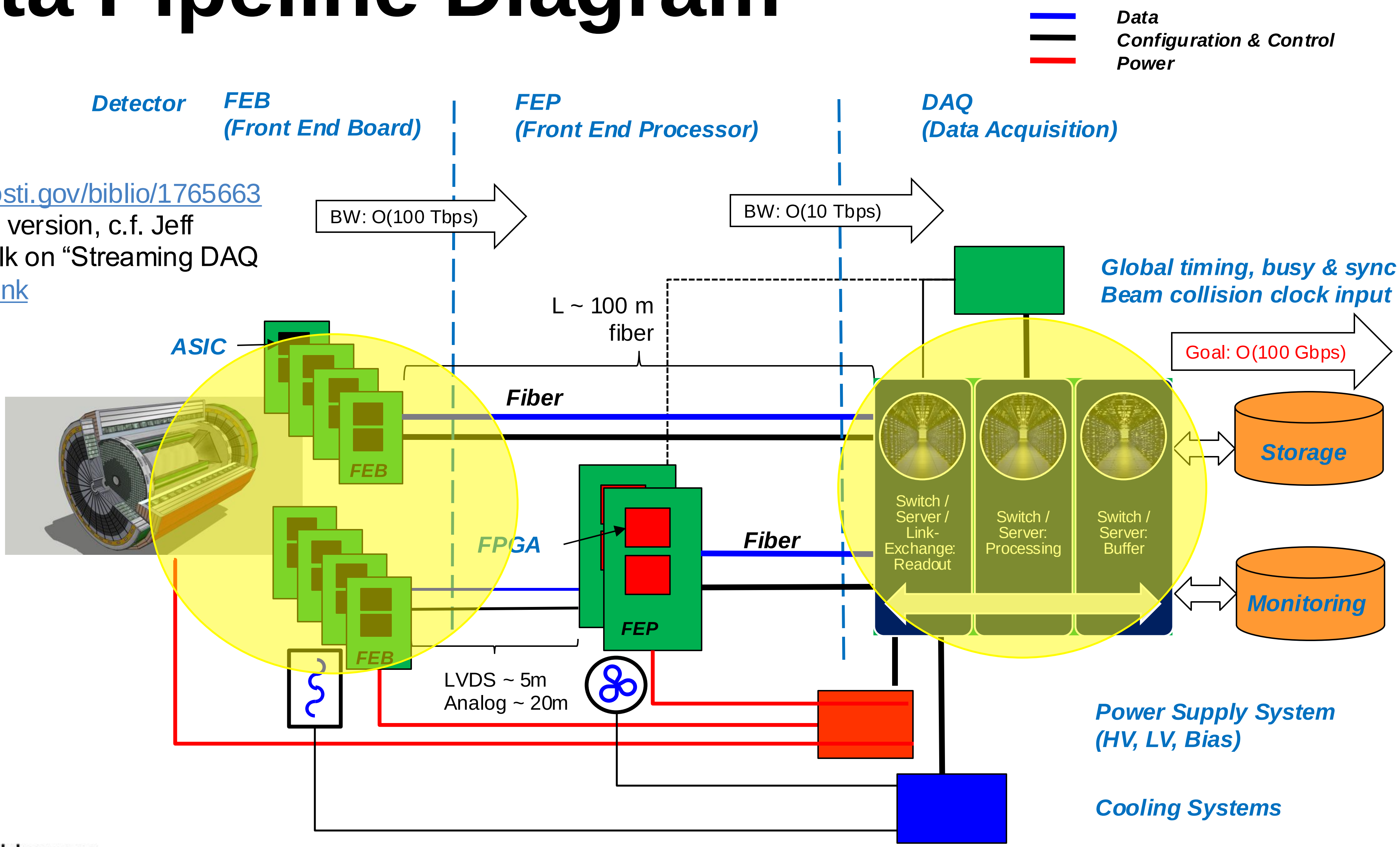
Common Challenges

- High data rate.
- Slow simulation.
- Simulation looks different from experiments. (domain shifting)

Can AI help?

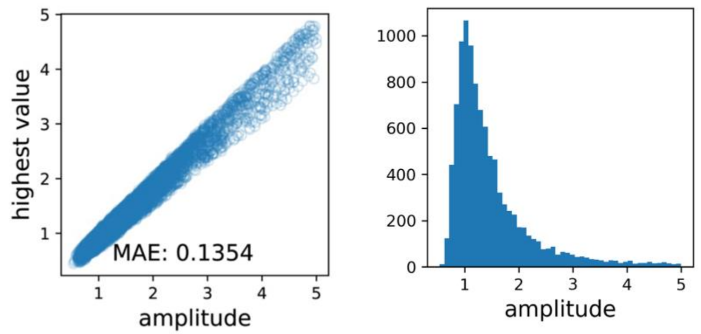
Data Pipeline Diagram

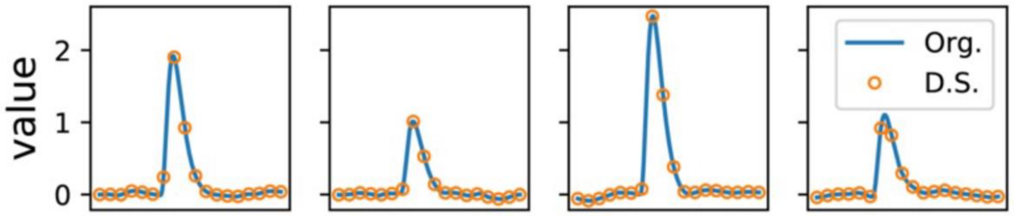
EIC CDR
https://www.osti.gov/biblio/1765663
For the latest version, c.f. Jeff Landgraf's talk on "Streaming DAQ Overview", Link

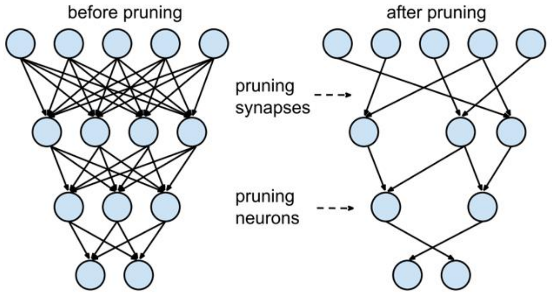


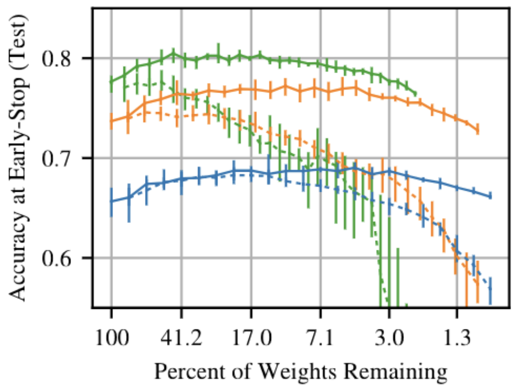
Finding Waveform Amplitude

- Simulated LGAD waveforms.
- Extremely low sample rate:
  - ~3 samples per peak.
- Goal: make network as small as possible.
- Lottery Ticket Hypothesis (pruning).
- Quantization-aware Training.
- MLP vs CNN.









Frankle, Jonathan, and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." *arXiv preprint arXiv:1803.03635* (2018).

Network Pruning

Not much difference between three reset methods. (RR, LTH, CP)

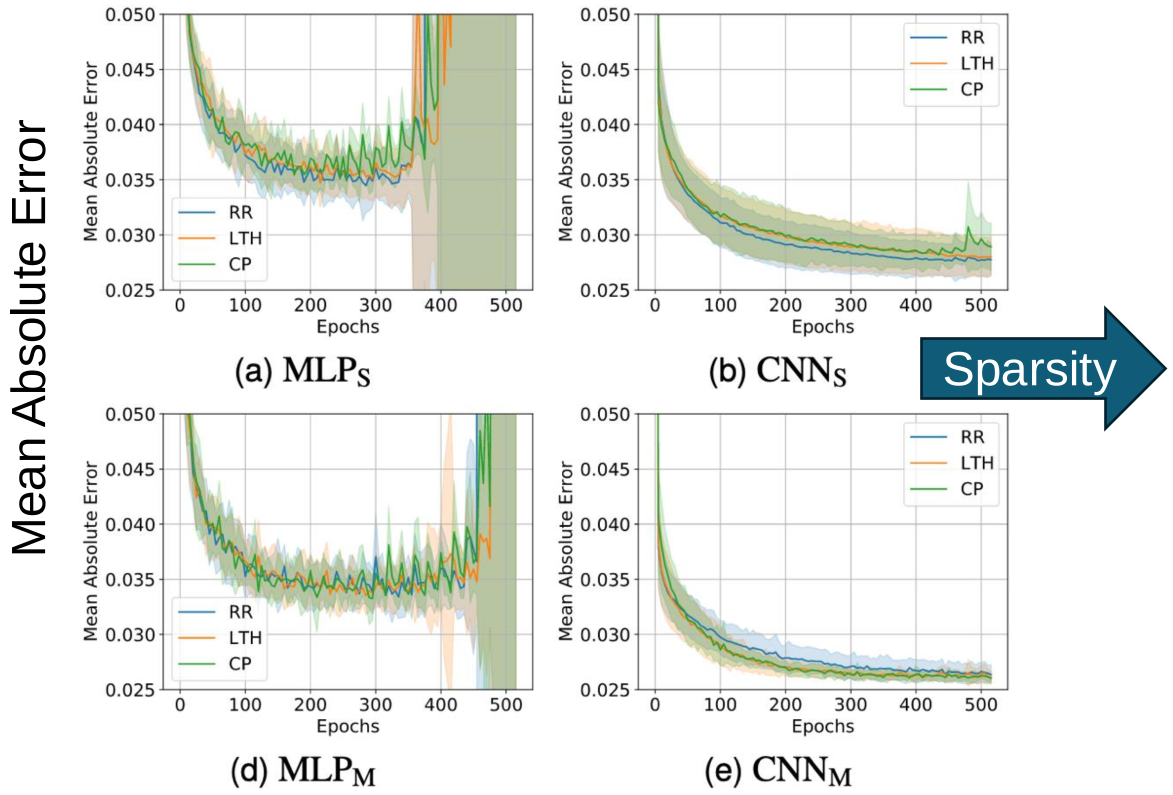
MLP can be pruned up to a point.

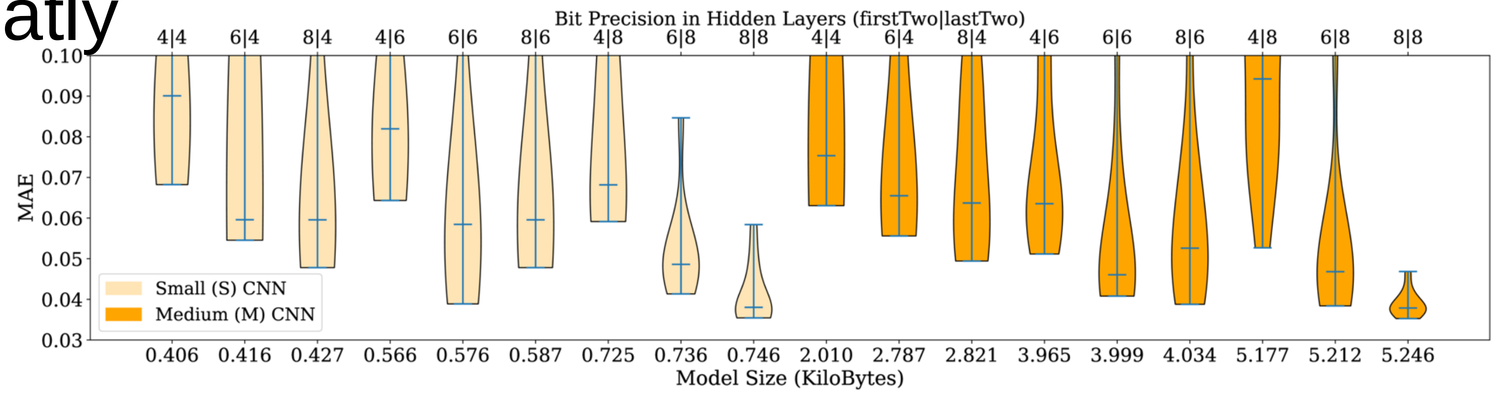
Larger MLP can be pruned further.

CNN can be sparsified greatly without loosing accuracy.

Pruning & Quantization

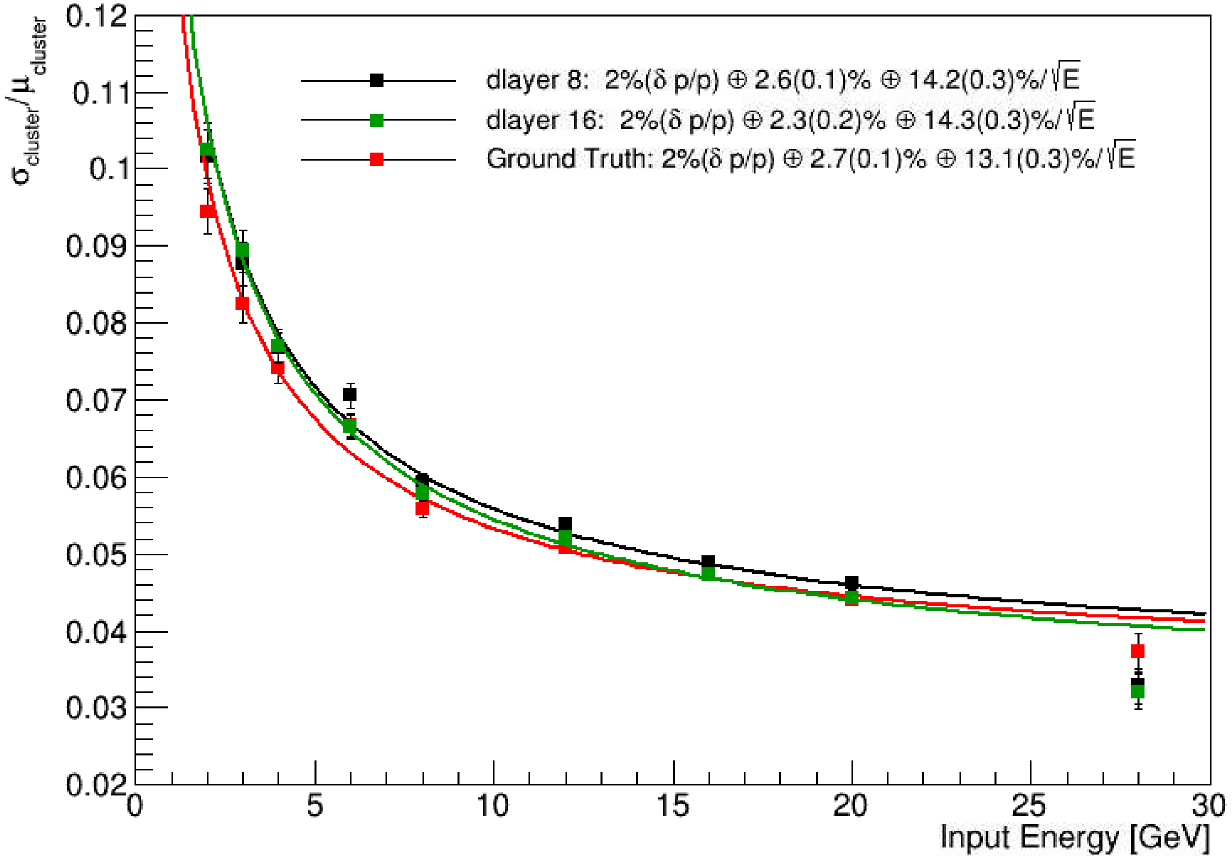
Y. Ren et al. (2022). Waveform processing using neural network algorithms on the front-end electronics. JINST, 17(01), C01039. [link]



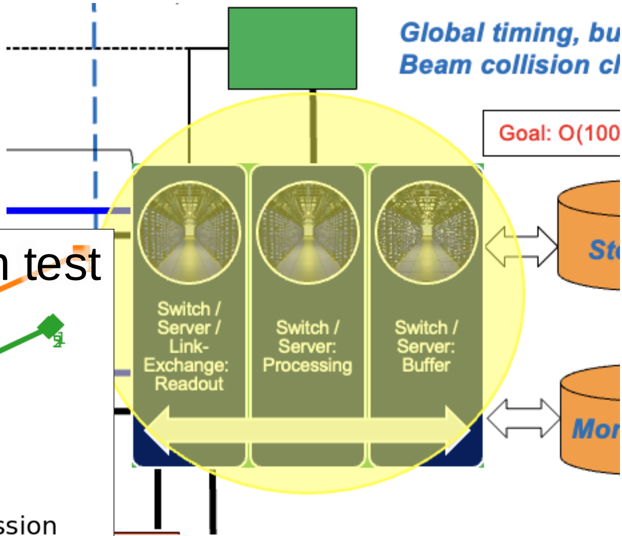


sPHENIX Test-beam data

- “dlayer 8/16”: channel size.
- y-axis is the fractional resolution (0.1 = a 10% sigma). The smaller the better.
- The CNN implementation has a larger resolution at low beam energies than more traditional approaches.
- Very similar performance observed in the region of 16-28 GeV
- It would be applicable for FPGA-based data reduction in the ePIC calorimeter too.



Streaming data compression

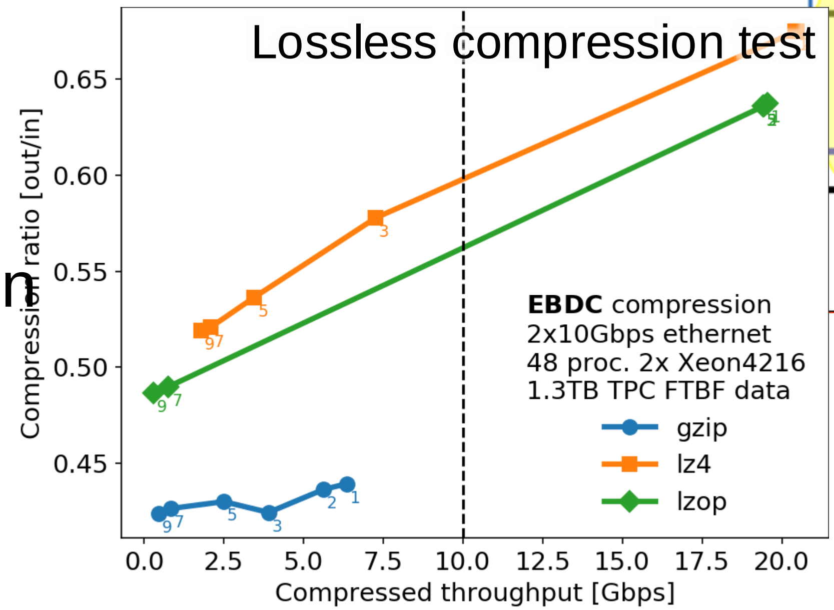


Lossless compression

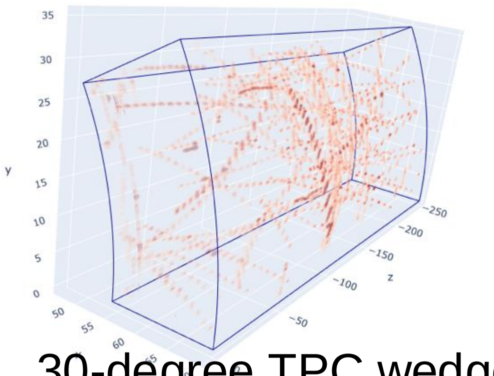
- Compress by ~1/2
- Well established fast compression algorithm

Lossy compression

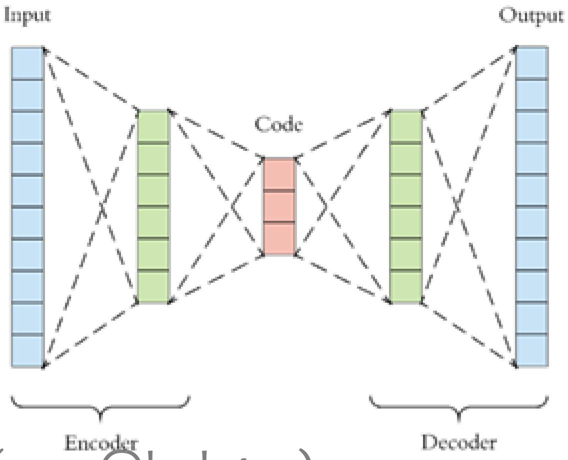
- Opportunity for unsupervised machine learning based on data.
- Bicephalous Convolutional Neural Encoder for zero-suppressed data (next)



Simple auto-encode neural network



30-degree TPC wedge

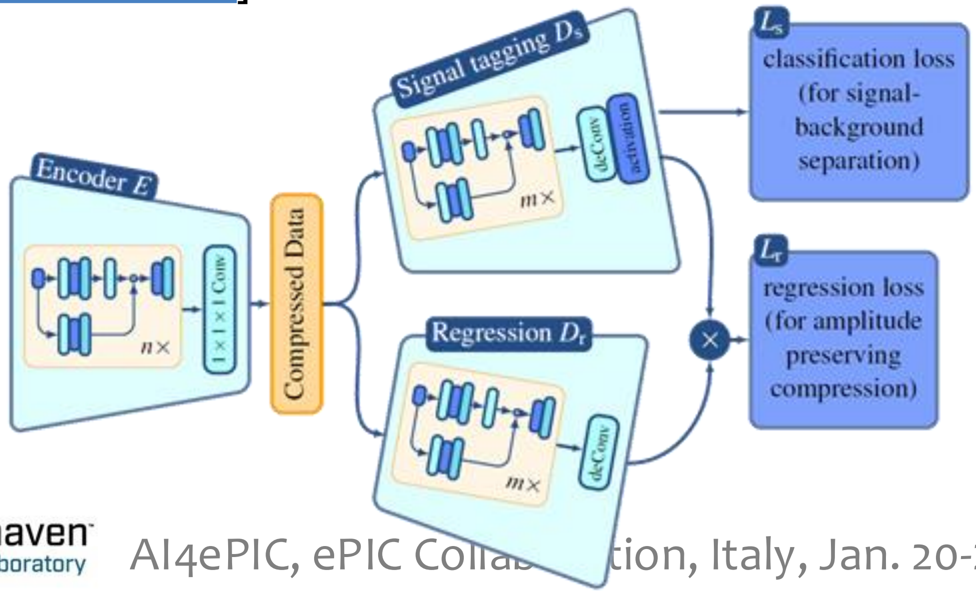


Bicephalous Convolutional Auto-Encoder for zero-suppressed data

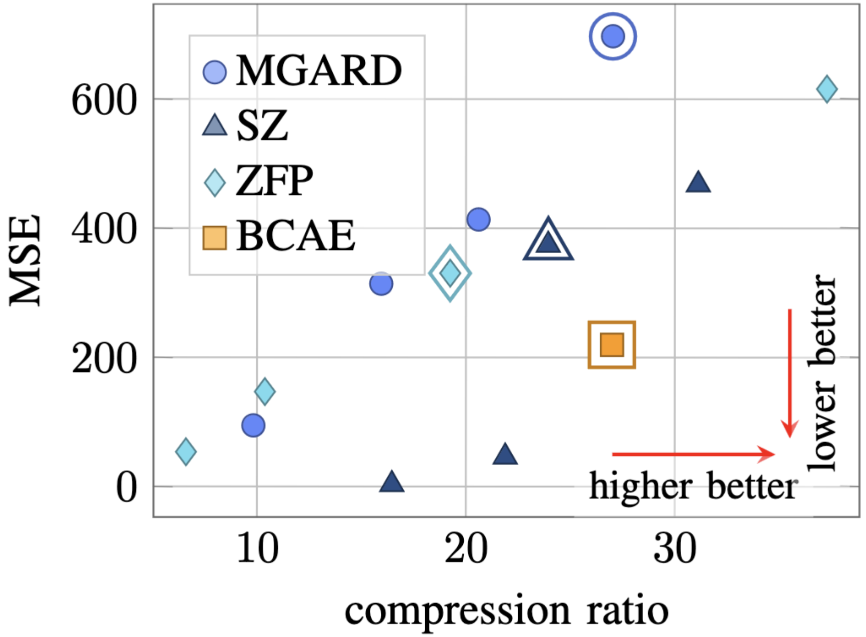
Some detector ADC data is challenging for Auto-Encoder, e.g. features such as zero-suppression cut off

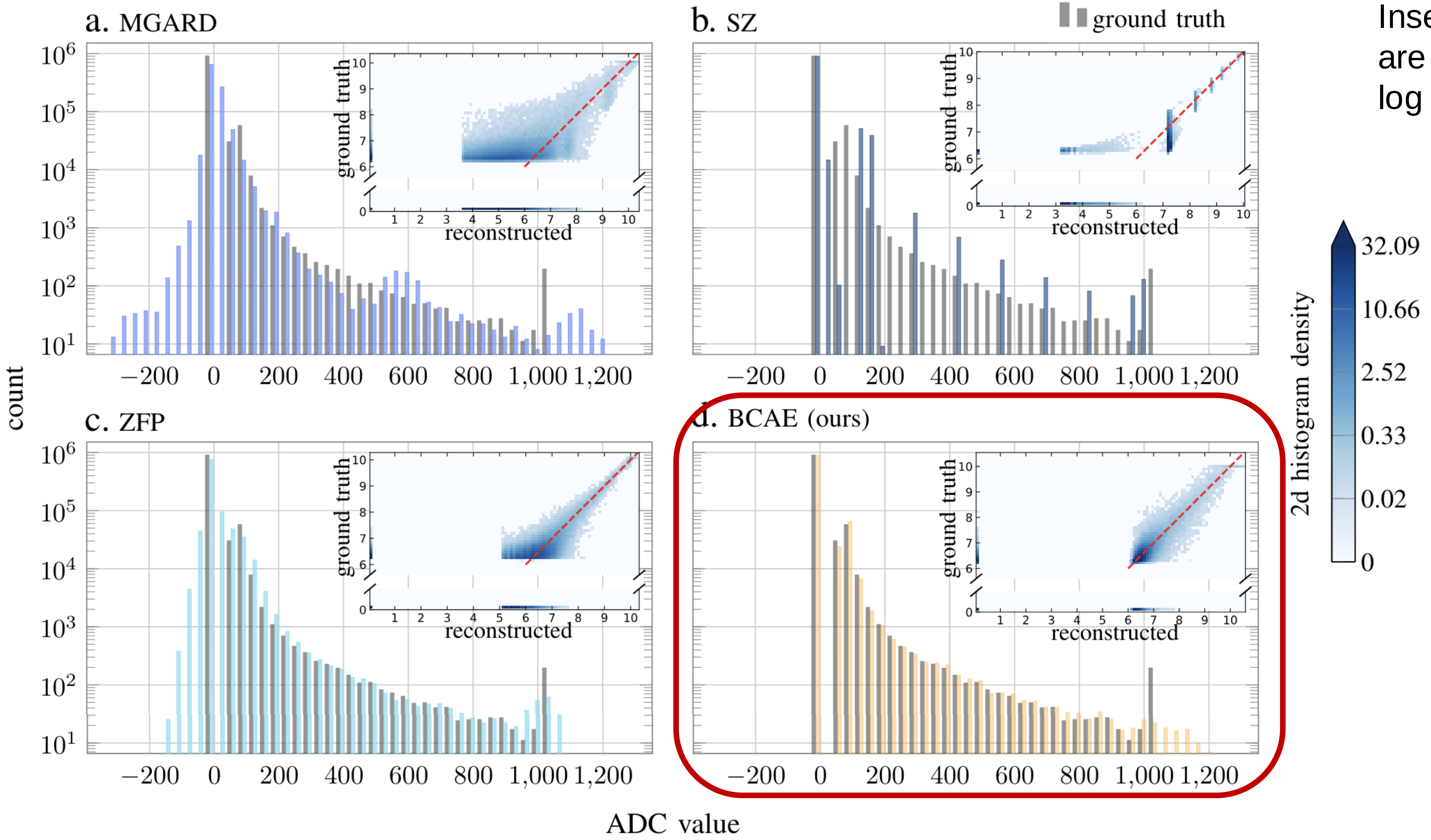
A dual-output auto encoder is designed to output both a region of interest and decompressed ADC. Possibility for further noise filtering

Ref: Y. Huang @ AI4EIC workshop [link], Paper [arxiv:2111.05423]



Compression comparison with published compressor tested on busiest sPHENIX TPC timeframes.
About 3000~4000 frames per second on A6000 GPU.





Inset figures are in log-log scale.

Fast BCAE-2D

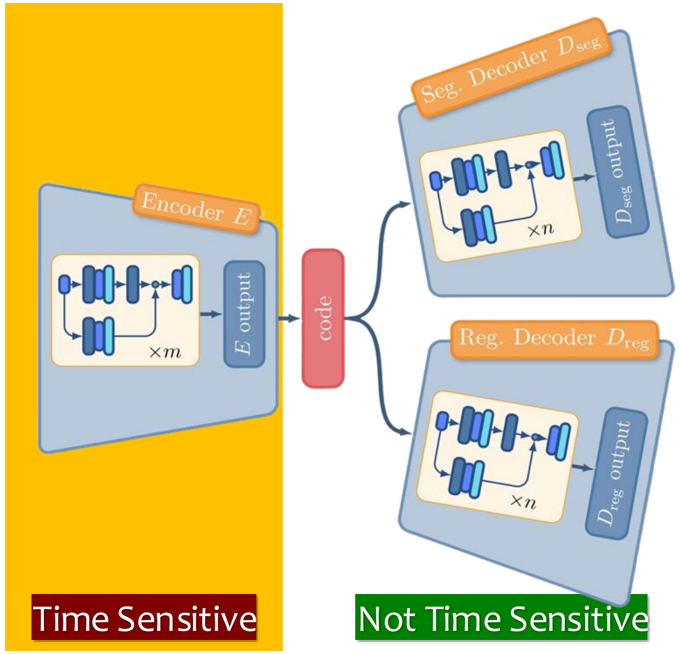
For real-time compression, only the speed of the Encoder matters.

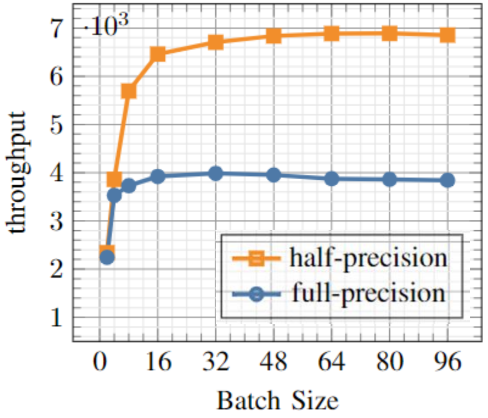
Is it possible to trade off the size of Encoder with the size of Decoder? YES!

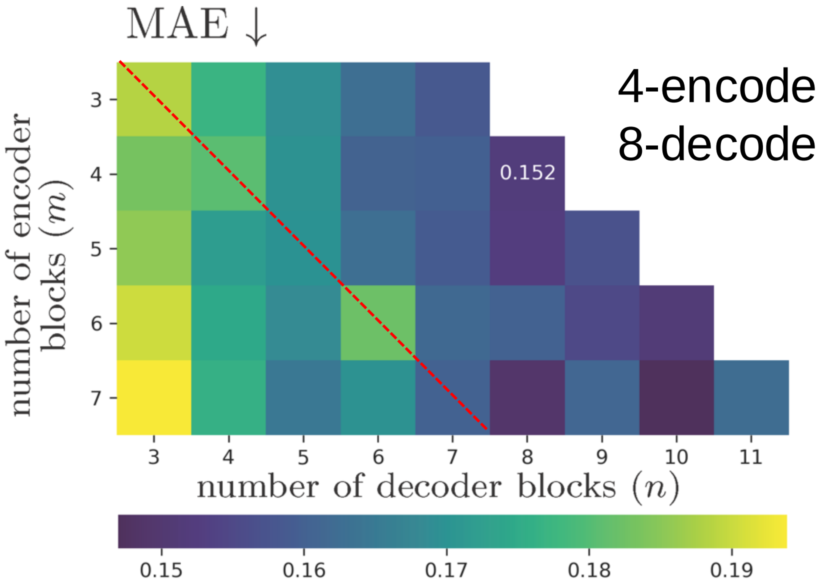
7k TPC wedges /s.

Best Paper Runner-up

Huang, Yi, Yihui Ren, Shinjae Yoo, and Jin Huang. "Fast 2D Bicephalous Convolutional Autoencoder for Compressing 3D Time Projection Chamber Data." In *Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, pp. 298-305. 2023.LINK

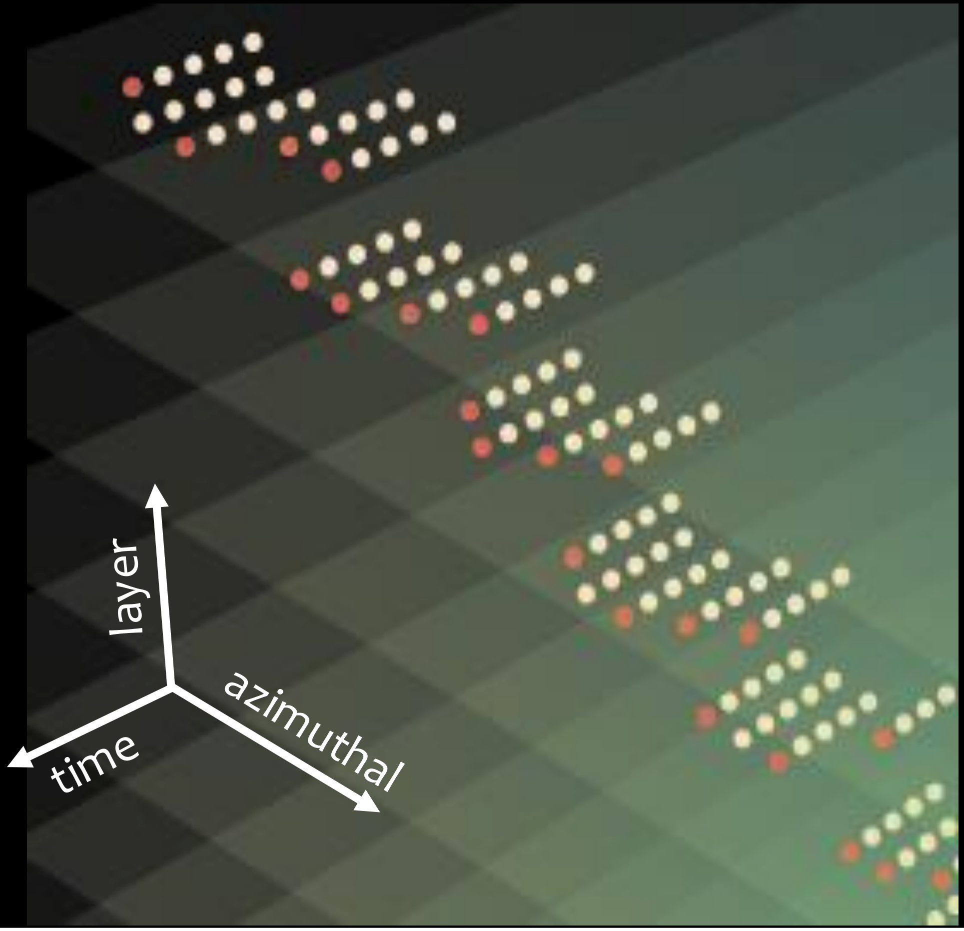






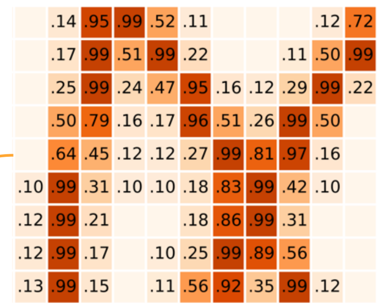
4-encoder layers
8-decoder layers

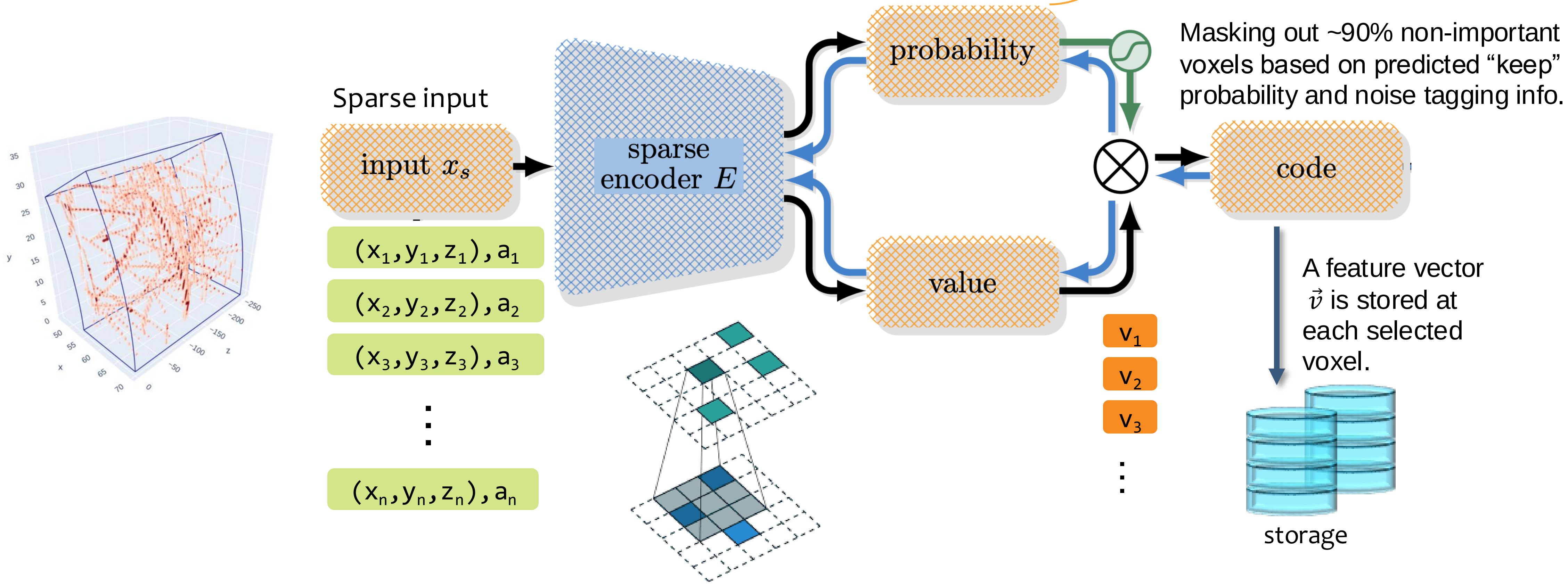
BCAE-VS: Bicephalous Convolutional Autoencoder with Variable ratio Compression for Sparse input



Locate the most valuable signals, and compress by down-selecting the signals

BCAE-VS Architecture

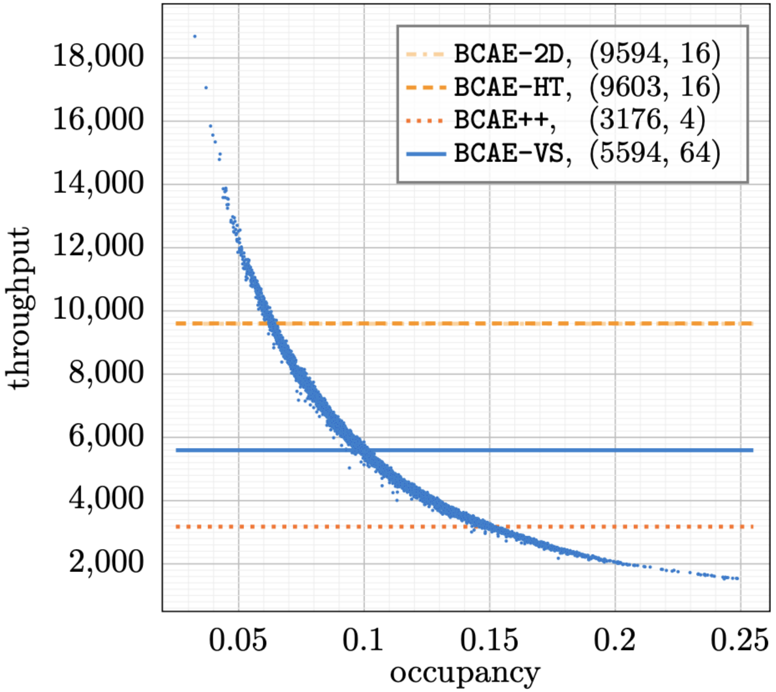


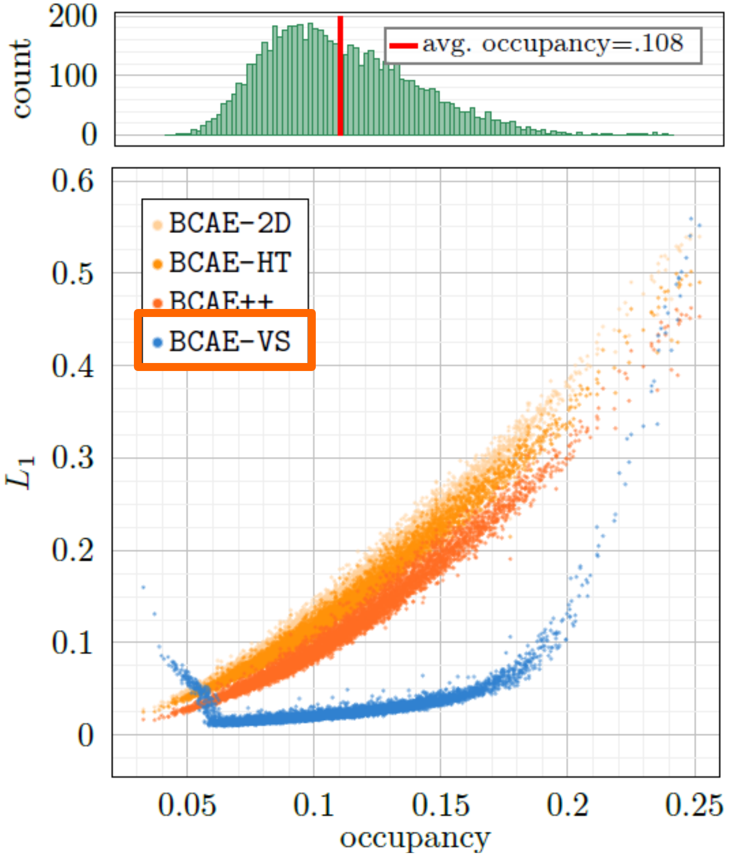


BCAE-VS Performance

Preliminary Results

- Higher compression ratio.
- Smaller model (382 parameters)
- Variable compression rate and speed depending on sparsity.
- This could be applicable to ePIC data reduction such as the dRICH detector, C.f. Alessandro Lonardo (INFN)'s talk (Link), and far backward trackers.





model	comp. ratio ↑	L_1 ↓	L_2 ↓	PSNR ↑	recall ↑	precision ↑	encoder size	throughput ↑
		reconstruction performance					efficiency	
BCAE-2D	31	.152	.862	20.6	.907	.906	169k	**9.6k**
BCAE-HT (3D)	31	.138	.781	20.8	.916	.915	9.8k	**9.6k**
BCAE++ (3D)	31	.112	.617	21.4	.936	.934	226k	3.2k
BCAE-VS	**34**	**.028**	**.089**	**26.0**	**.988**	**.996**	**382**	5.6k

Huang, Yi, et al. "Variable Rate Neural Compression for Sparse Detector Data." *arXiv preprint arXiv:2411.11942* (2024). LINK

Bridging Sim and Real: Unpaired AI-driven Data Translation

No Label constraint: AI/ML is a data-driven method. Often trained on simulation data with ground truth. But real data do not have "ground truth" to train on.

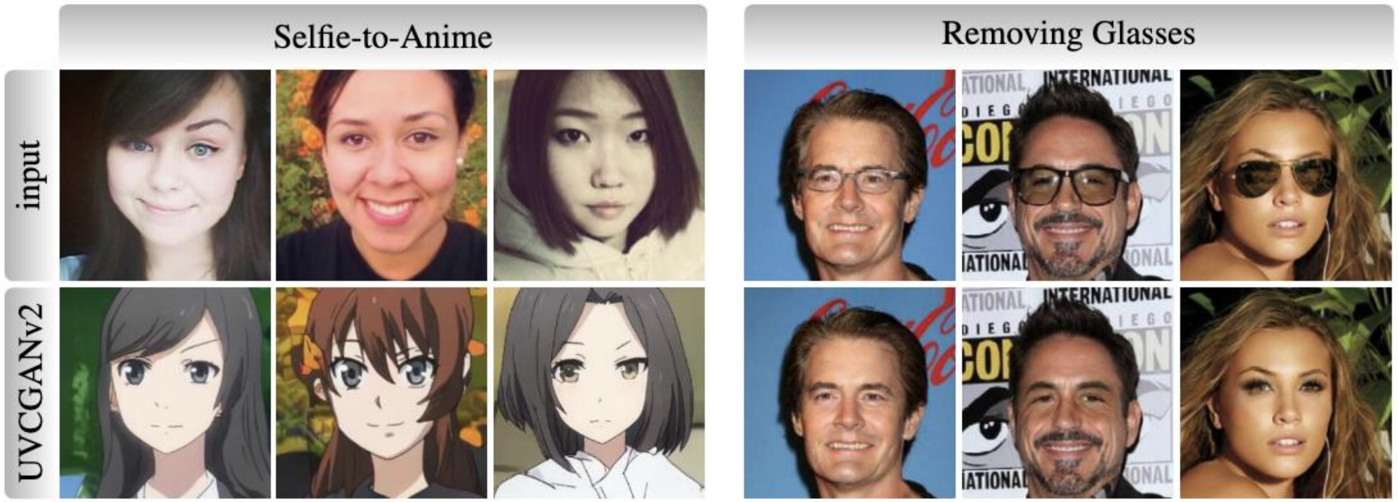
Unpaired constraint: since the ground truth of the experimental data is unknown, it's impossible to generate matched simulation images.

How to use simulation data with ground truth and unlabeled experimental data

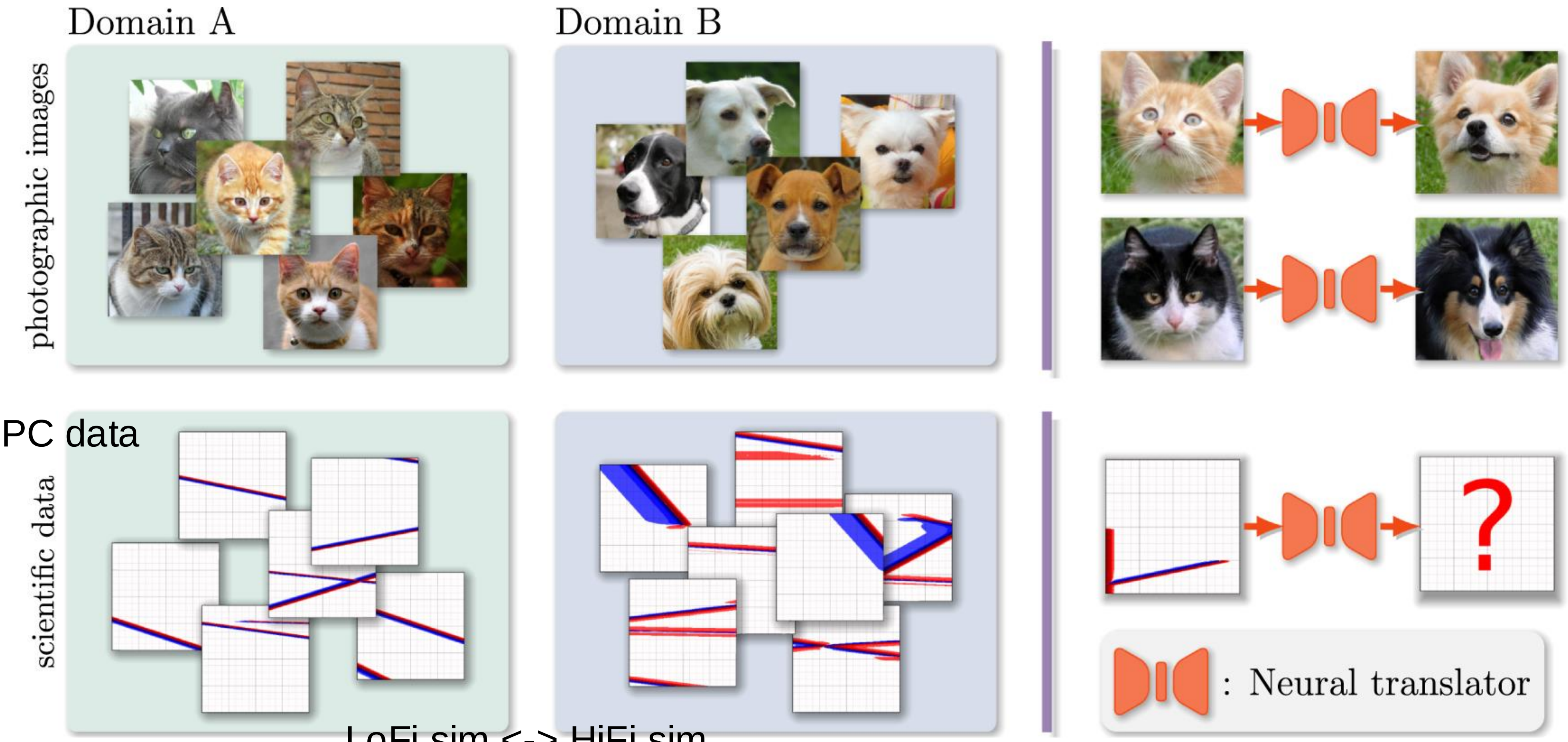
Our UVCGAN results on open dataset

Note the “feature consistency” (hair color, head orientation, unchanged background, etc.)

Torbunov, D., Huang, ... & Ren, Y. (2023). UVCGAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 702-712). [LINK]
Torbunov, D., Huang, Y., Tseng, H. H., Yu, H., UVCGAN v2: An Improved Cycle-Consistent GAN for Unpaired Image-to-Image Translation *arXiv:2303.16280*. [LINK]



Will Unpaired Image Translation preserve physics quantity?



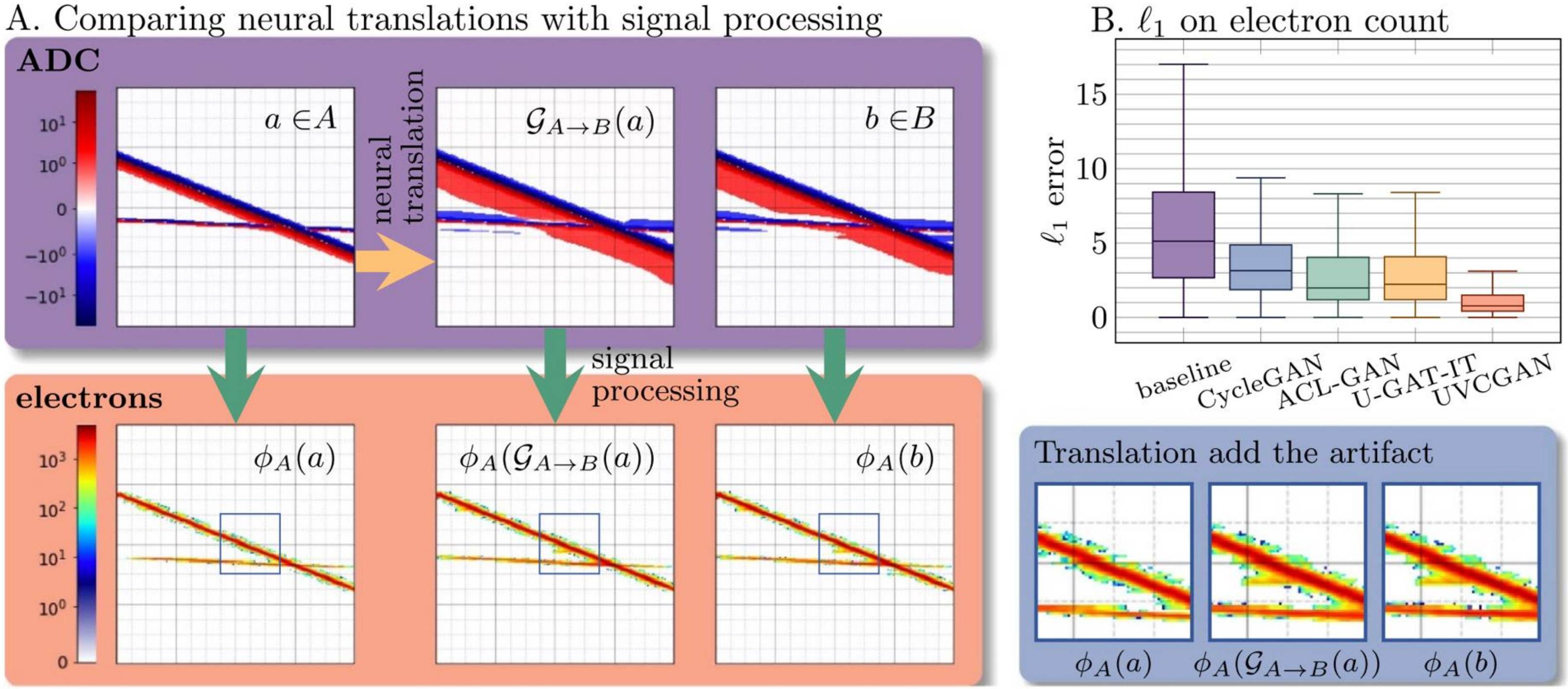
LArTPC data

LoFi sim <-> HiFi sim

Results

Because the translation is unsupervised and unpaired between data domains; it can be adapted to any downstream tasks.

This is ready for use in the ePIC simulation augmentation, happy to collaborate.



Domain shift mitigation for a supervised learning algorithm

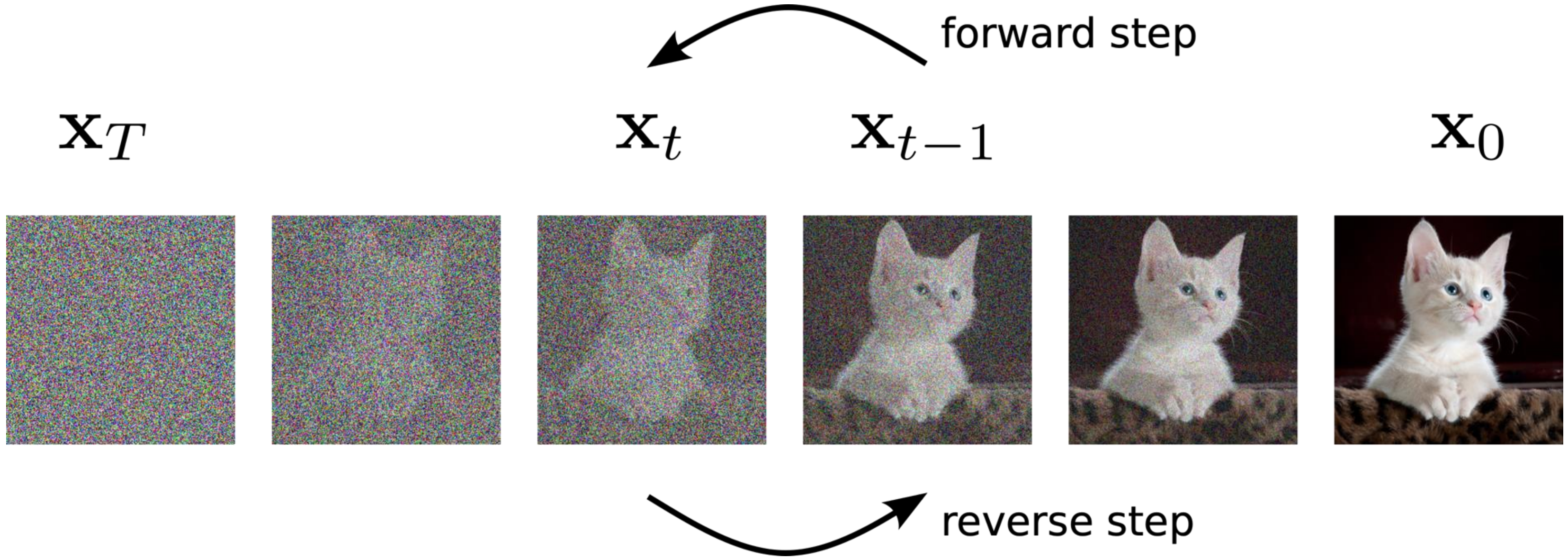
	E_A	E_B	E_{CycleGAN}	$E_{\text{ACL-GAN}}$	$E_{\text{U-GAT-IT}}$	E_{UVCGAN}
B	0.390	0.211	0.222	0.223	0.257	0.216
	Trained on A tested on B	Trained on B tested on B				Trained on B' tested on B

Huang, Y., Torbunov, D., Viren, B., Yu, H., Huang, J., Lin, M., & Ren, Y. (2024). Unpaired image translation to mitigate domain shift in liquid argon time projection chamber detector responses. *Machine Learning: Science and Technology*, *5*(4), 045021. LINK

Denoising Diffusion Probabilistic Model (DDPM)

c.f. Yeonju Go's ACAT24 talk (LINK)

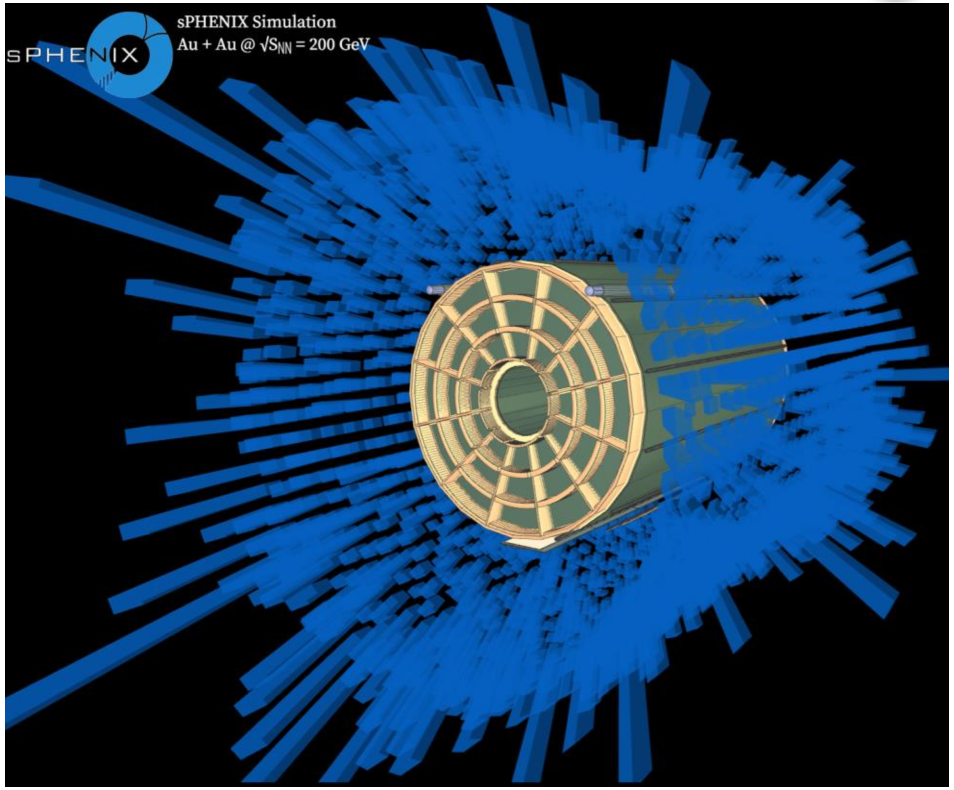
- DDPM provides ***high quality data from random noise***
- **Forward** process: add random gaussian noise
- **Reverse** process: use neural network and generate data
- In real application, O(1,000) steps are used



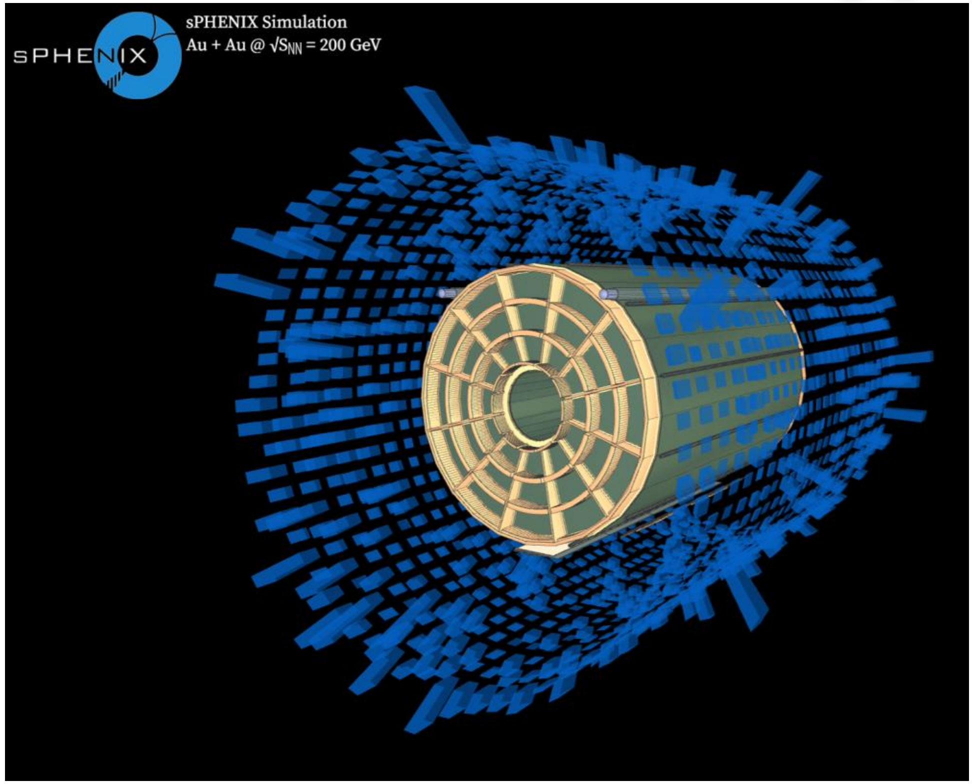
Heavy Ion Collision Event

- **HIJING** Monte Carlo event generator for Au+Au collisions at $\sqrt{s_{\mathrm{NN}}}$ 200=GeV
- **Geant4** full detector simulation with the sPHENIX geometry

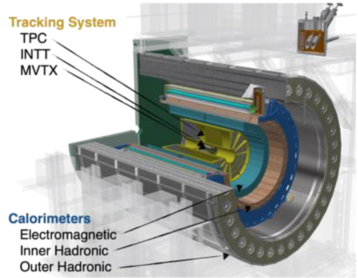
Head-on collision (0-10% Centrality)

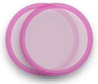


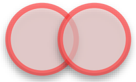
Side collision (40-50% Centrality)

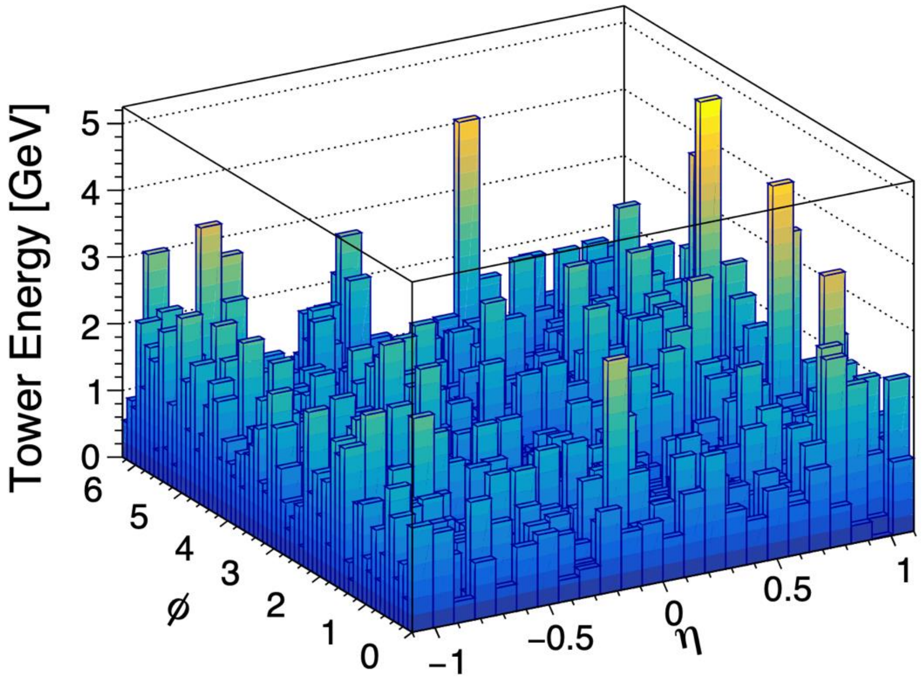


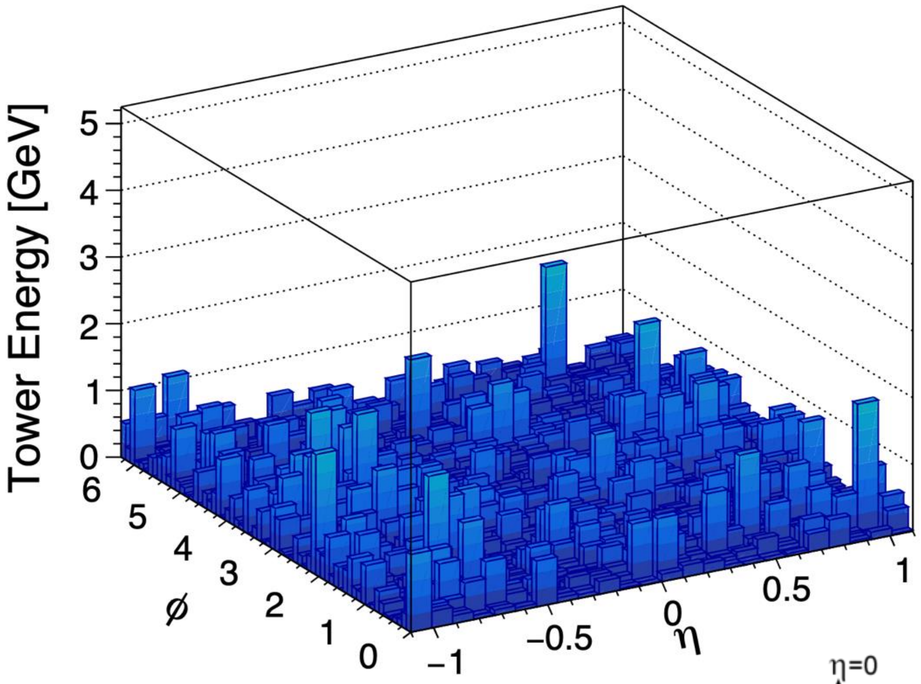
Tower Distributions



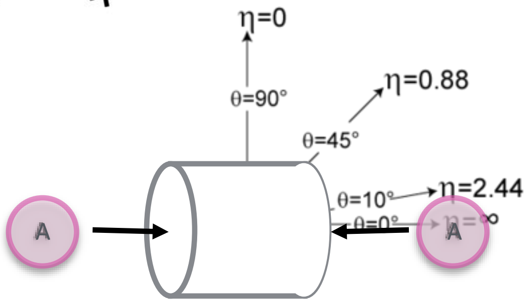
0-10% Centrality 

40-50% Centrality 



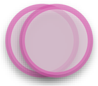


- Full calorimeter **towers** (Electromagnetic + Inner hadronic + Outer hadronic)
  - $-1.1 < \eta < 1.1, 0 < \phi < 2\pi$
  - (24 x 64) bins in (η, ϕ)



Display of Generated Events

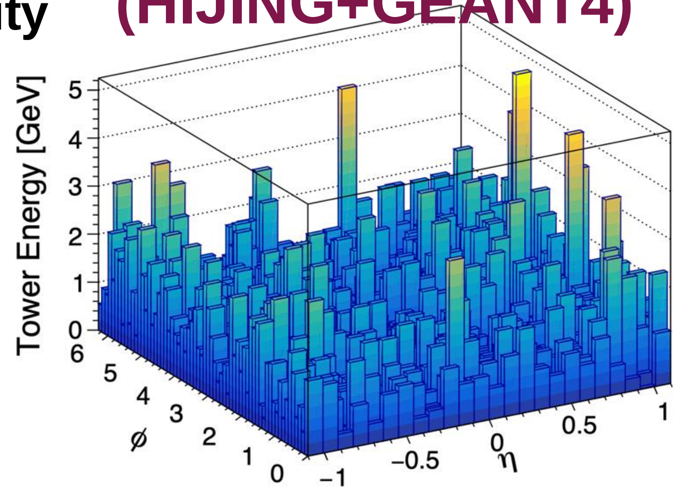
0-10% Centrality

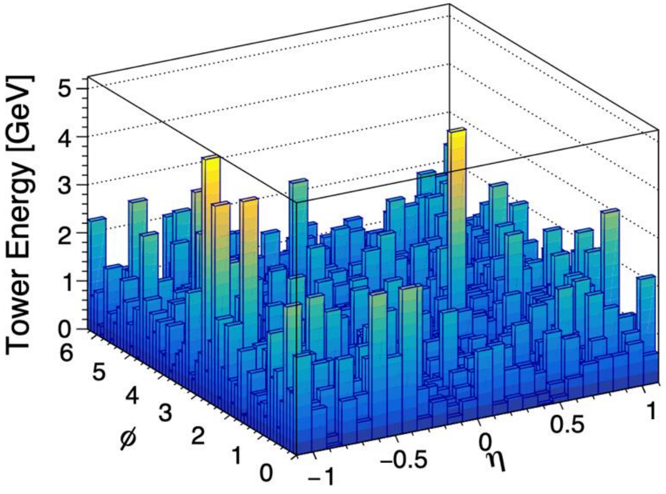


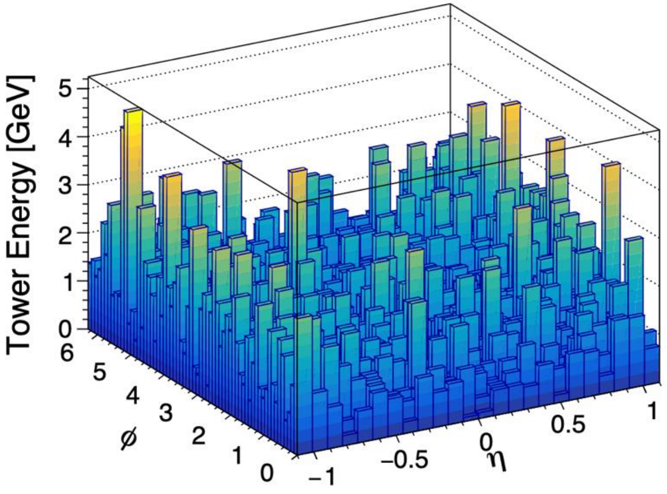
Training sample (HIJING+GEANT4)

Generated (DDPM)

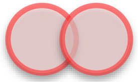
Generated (GAN)

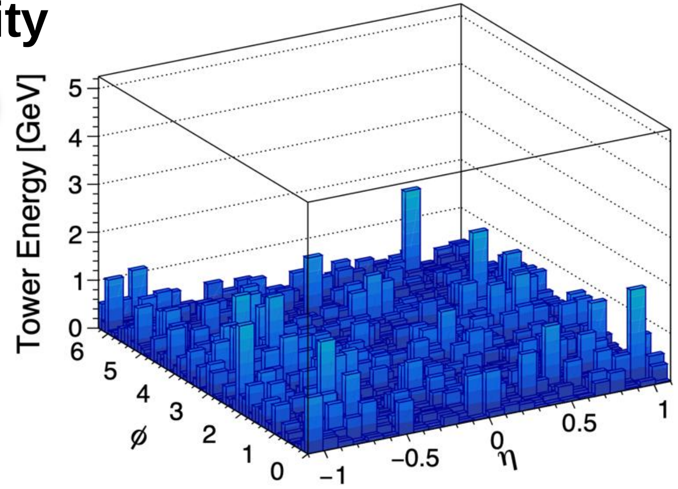


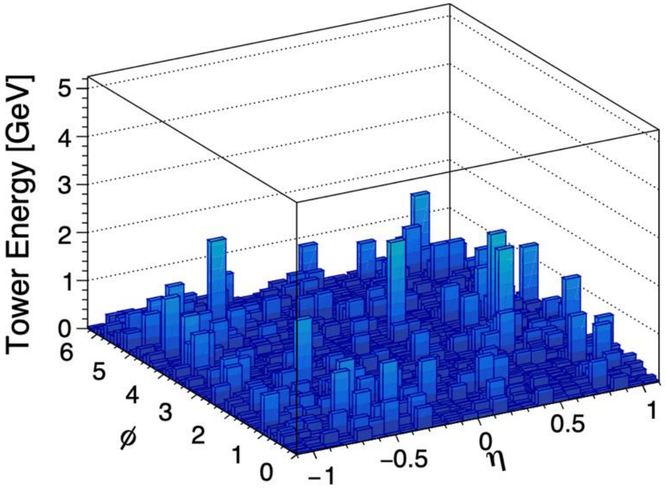


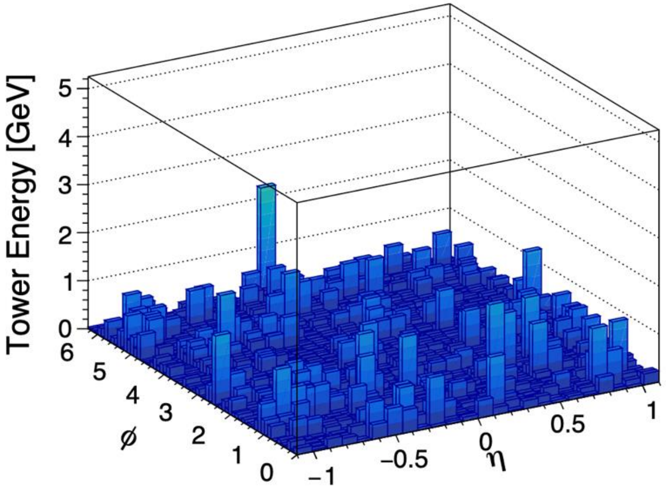


40-50% Centrality

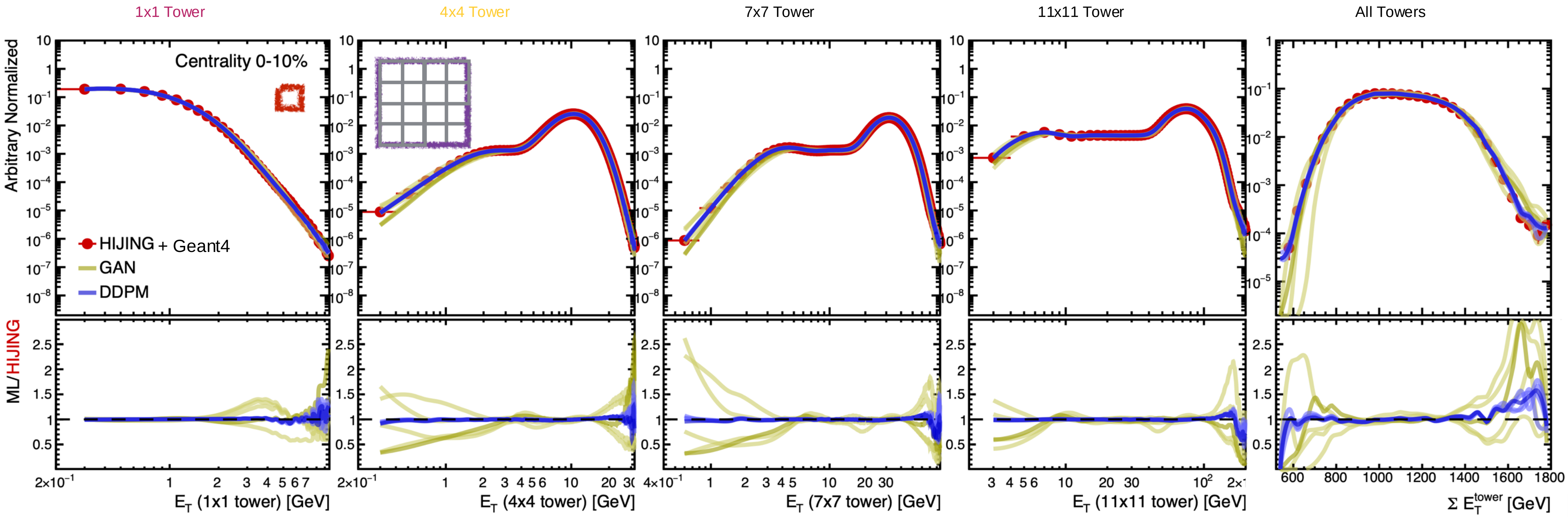


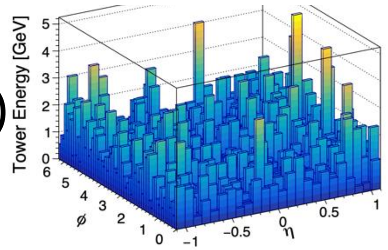






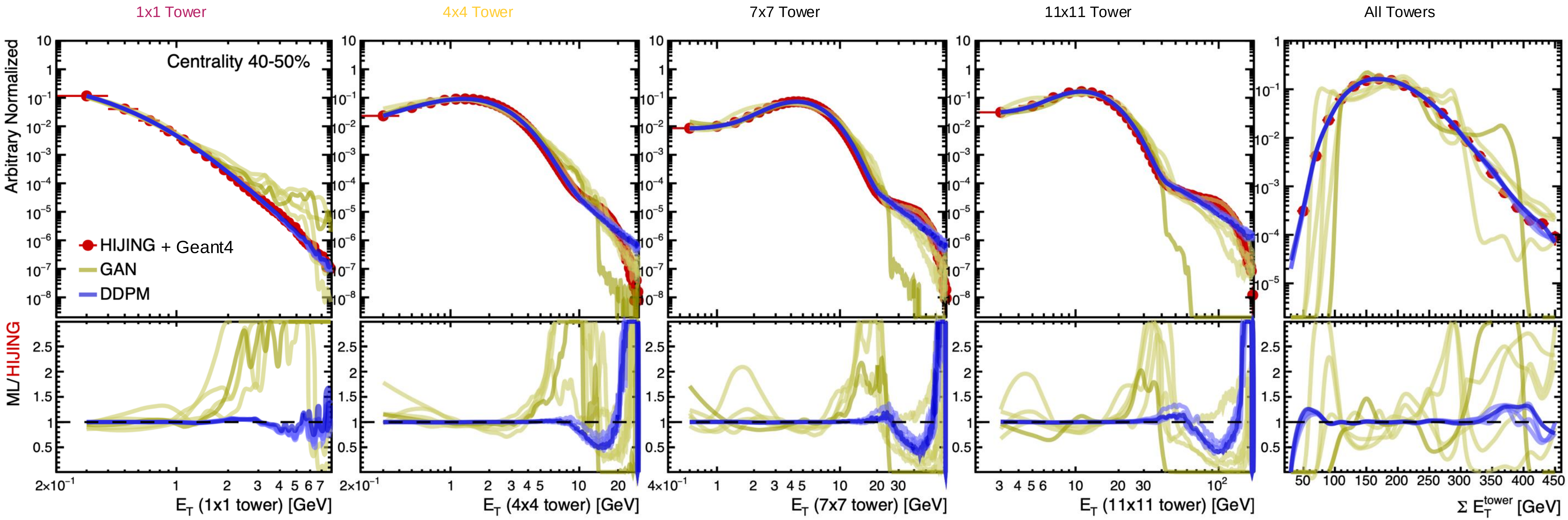
Performance: Transverse Energy (0-10%)





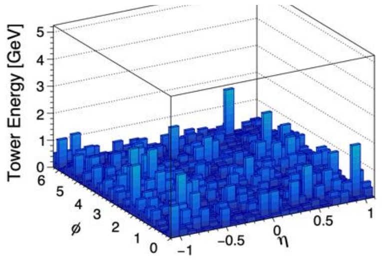
- Each model is retrained 5 times with different random seeds
- HIJING+Geant4 used as training data (600k events) and testing data (100k events)
- Both DDPM and GAN reproduce the data distribution where the data are abundant
- DDPM outperforms GAN in overall distribution w/ great stability and accuracy

Performance: Transverse Energy (40-50%)



- DDPM outperforms GAN
- 100x Speedup v.s. Geant4
- Potentially for ePIC, generate a large set of background events (e.g. synchrotron radiation)

Go, Yeonju, Torbunov, D, et al. "Effectiveness of denoising diffusion probabilistic models for fast and high-fidelity whole-event simulation in high-energy heavy-ion experiments." *Physical Review C*, *110*(3), 034912. LINK



Conclusion

- High data rate. (BCAEs neural compression models for sparse data, ePIC’s dRHIC and far backward detectors.)
- Slow simulation. (DDPM-based generative models for faster simulation.)
- Simulation $\leftrightarrow$ experiments. (unpaired translation UVCGAN)

Thank you!

(NP) Timothy Rinn, Yeonju Go, Evgeny Shulga, Joe Osborn, Jin Huang

(DUNE) Haiwang Yu, Brett Viren, Chao Zhang, Xin Qian

(ASIC) Soumyajit Mandal, Prashansa Mukim, Grzegorz Deptuch, Piotr Maj, Gabriella Carini

(ATLAS) Elizabeth Brost, Haider Abidi, Viviana Cavaliere, Michael Begel

(CSI) Dmitrii Torbunov, Yi Huang, Shubha Khrael, Meifeng Lin, Shinjae Yoo

Feel free to contact me:

Yihui "Ray" Ren (yren@bnl.gov)