Challenges and Progress towards Applying AI/ML methods on Experimental Data

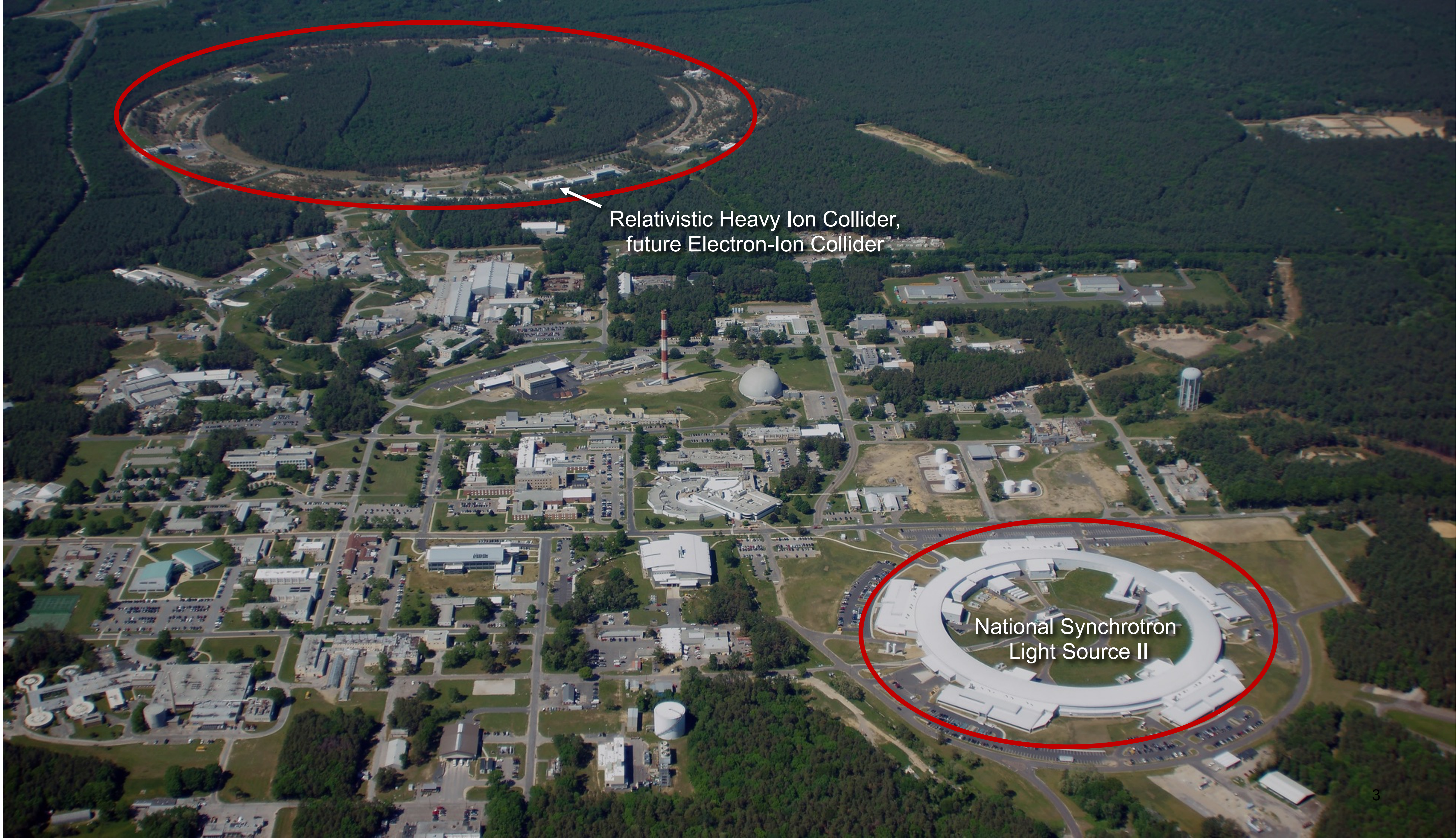
Yihui "Ray" Ren (yren@bnl.gov)
Computational Science Initiative (CSI), BNL

2023 RHIC/AGS Annual Users' Meeting, Aug. 1st – 4th, 2023

Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

@BrookhavenLab


Relativistic Heavy Ion Collider,
future Electron-Ion Collider
National Synchrotron
Light Source II


Computational
Science Initiative

Collaborators

- Timothy Rinn, Yeonju Go, Jin Huang, David Morrison
- Dmitrii Torbunov, Haiwang Yu, Brett Viren, Chao Zhang, Xin Qian
- Piotr Maj, Dominik Gorni, Soumyajit Mandal, Prashansa Mukim, Grzegorz Deptuch, Gabriella Carini
- Elizabeth Brost, Haider Abidi , Viviana Cavaliere, Michael Begel
- Yi Huang, Shubha Khrael, Meifeng Lin, Shinjae Yoo

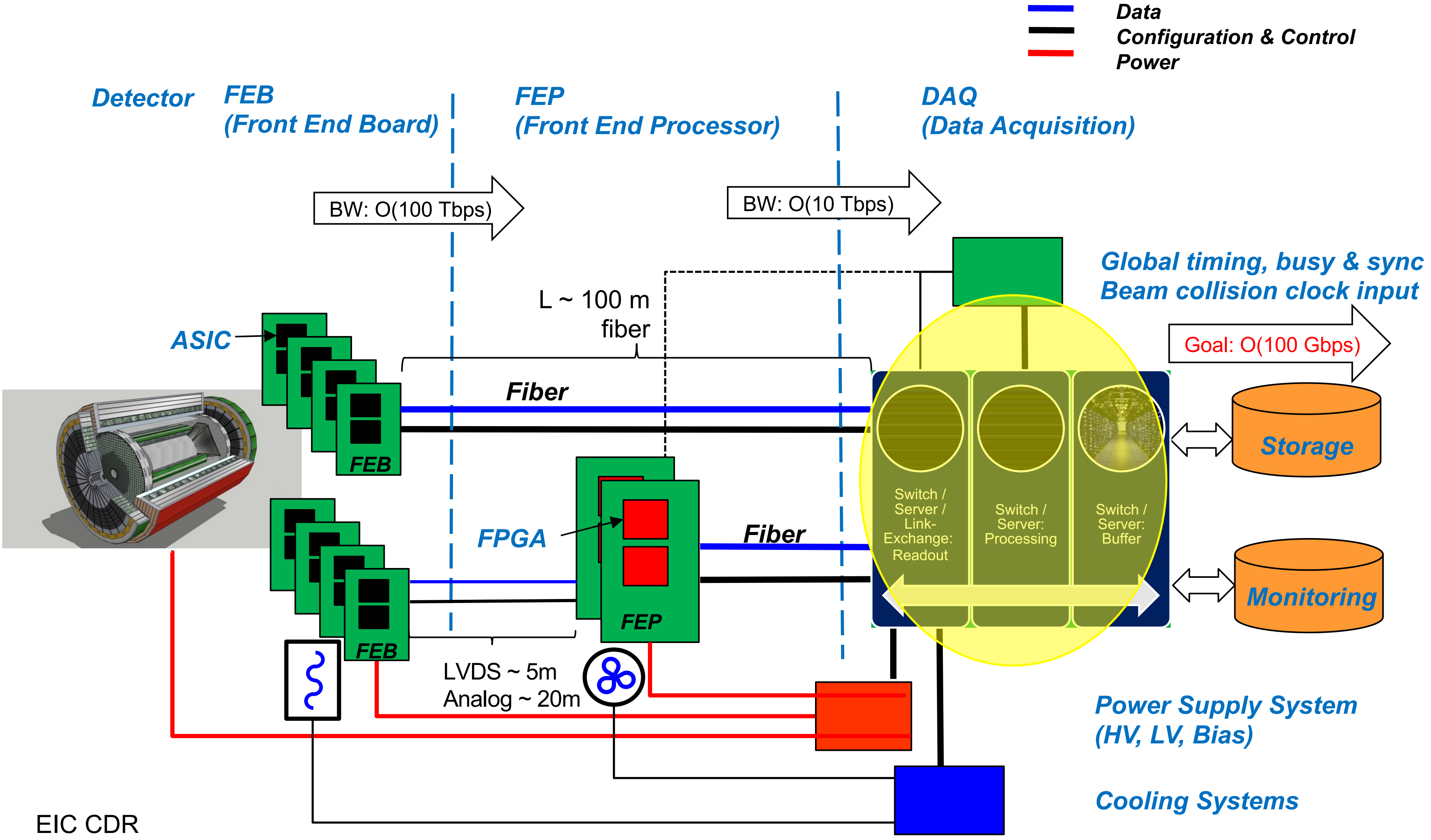
If there are any errors in the slides. I’m responsible.

Common Challenges

Why applying AI/ML to real experimental data so difficult?

- Real data come in with large volumes and fast.
- AI/ML is a data-driven method, real data do not have “ground truth” to train on.

Data Pipeline Diagram [EIC pre-CDR as example]



EIC CDR

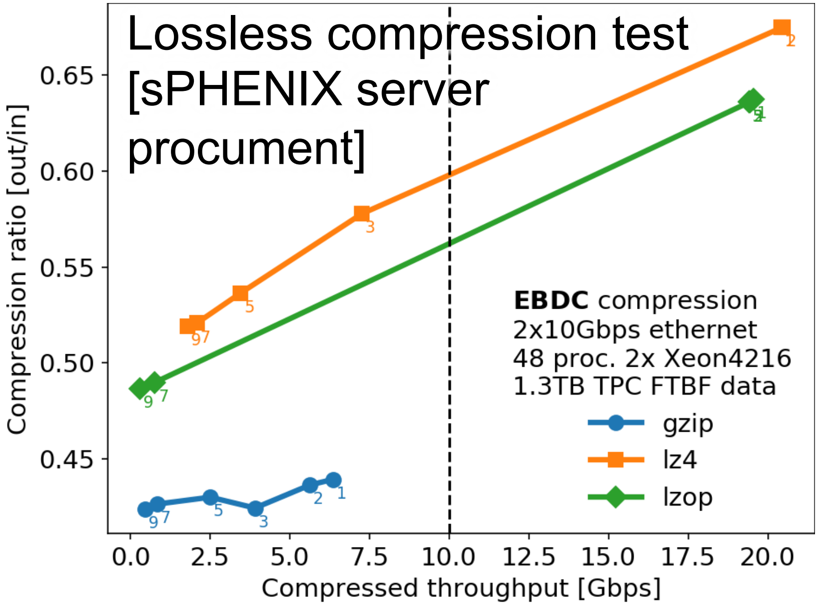
Online streaming data - compression

Lossless compression

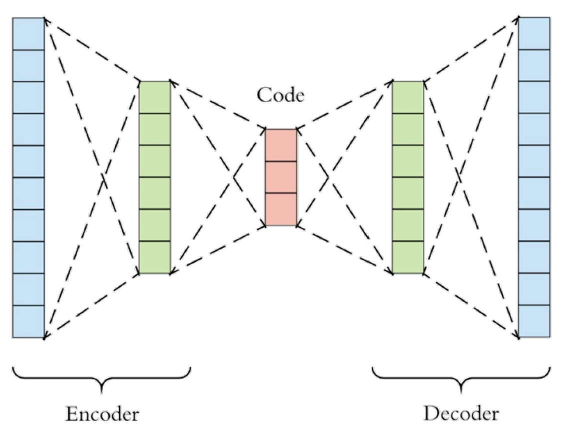
- Compress by ~1/2
- Well established fast compression algorithm

Lossy compression

- Opportunity for unsupervised machine learning based on data, e.g.
- Auto-encoder on ASIC for HGCal @ CMS [link]
- Bicephalous Convolutional Neural Encoder for zero-suppressed data (next)



Simple auto-encode neural network

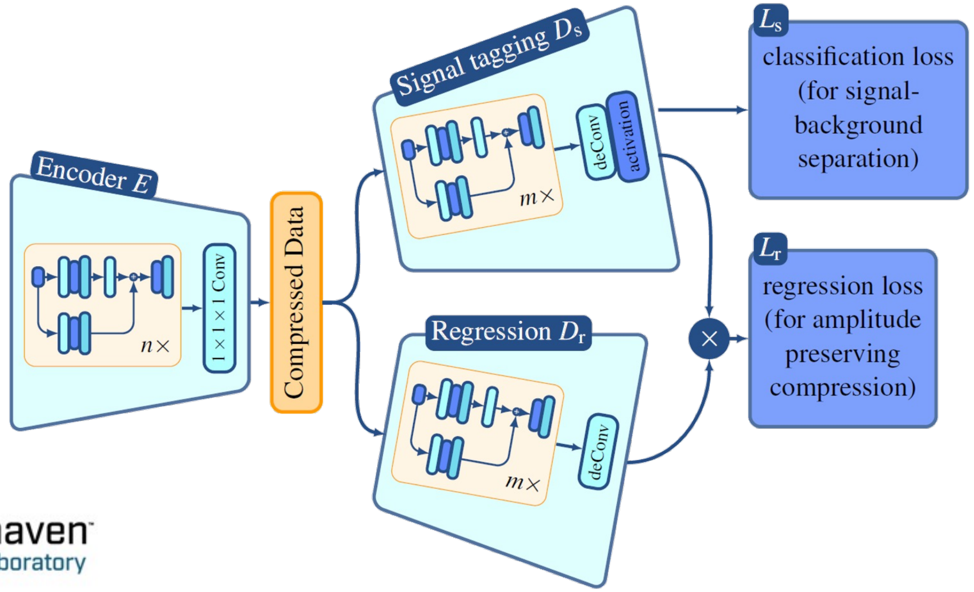


Bicephalous Convolutional Auto-Encoder for zero-suppressed data

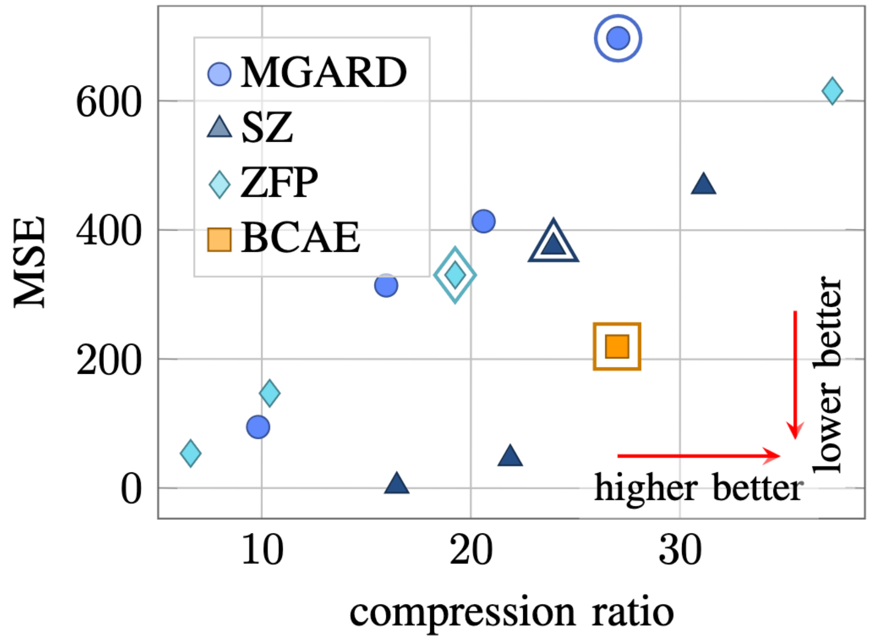
Some detector ADC data is challenging for Auto-Encoder, e.g. features such as zero-suppression cut off

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

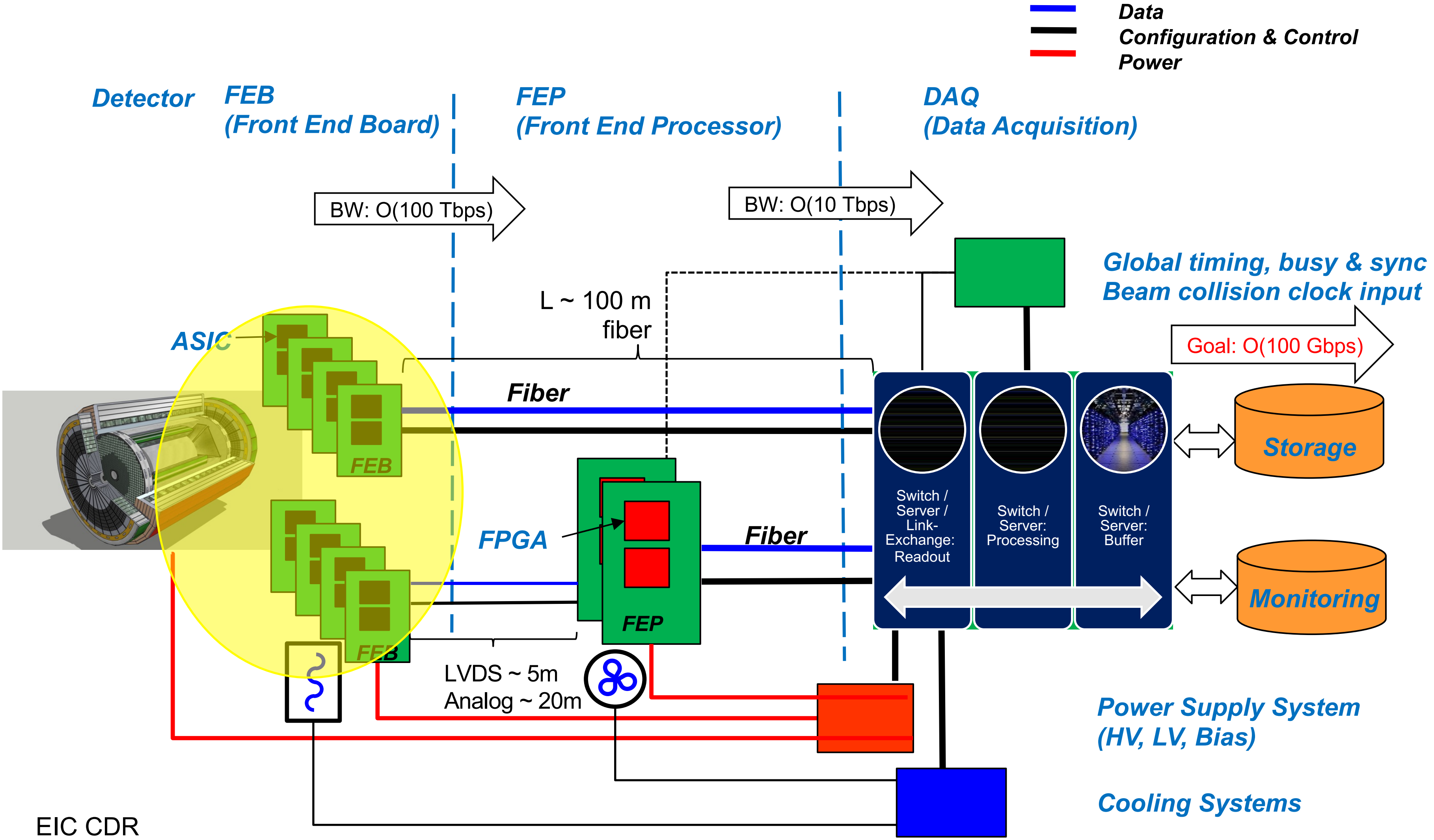
Ref: Y. Huang @ AI4EIC workshop [link], Paper [arxiv:2111.05423]



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



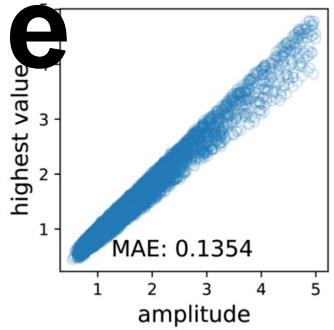
Data Pipeline Diagram

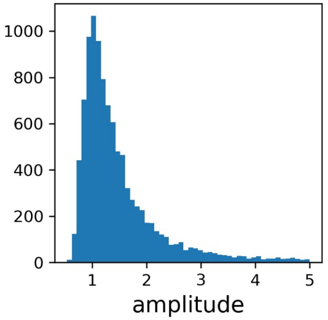


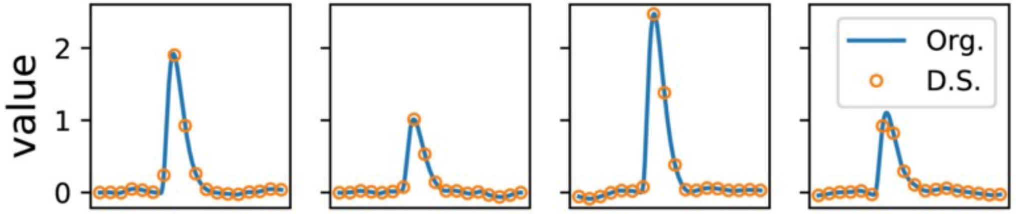
EIC CDR

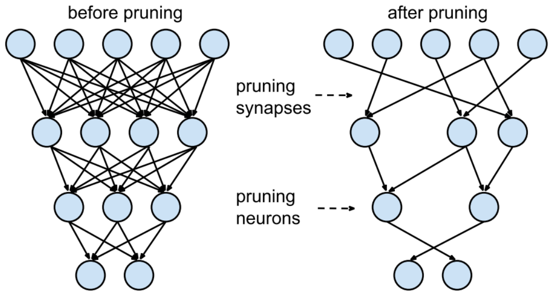
Finding Waveform Amplitude

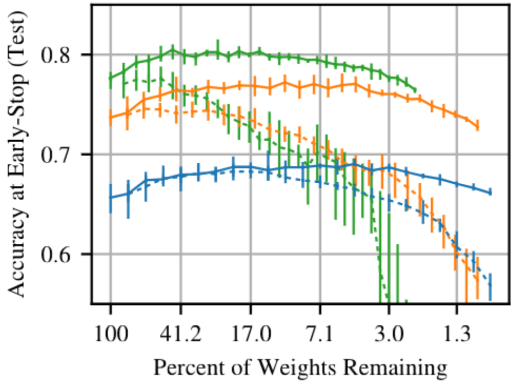
- Simulated LGAD waveforms.
- Goal: make network as small as possible.
- Lottery Ticket Hypothesis (pruning).
- Quantization-aware Training.
- MLP vs CNN.











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

Brookhaven National Laboratory

Results

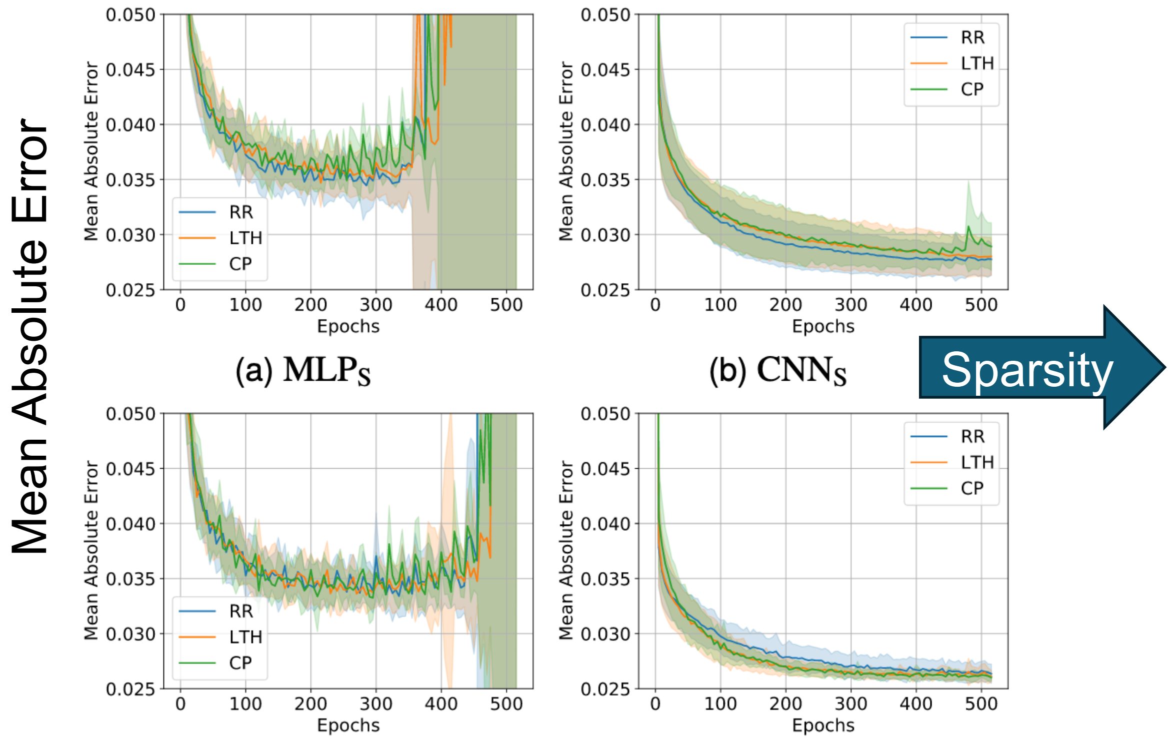
Not much difference between three reset methods. (RR, LTH, CP)

MLP can be pruned up to a point.

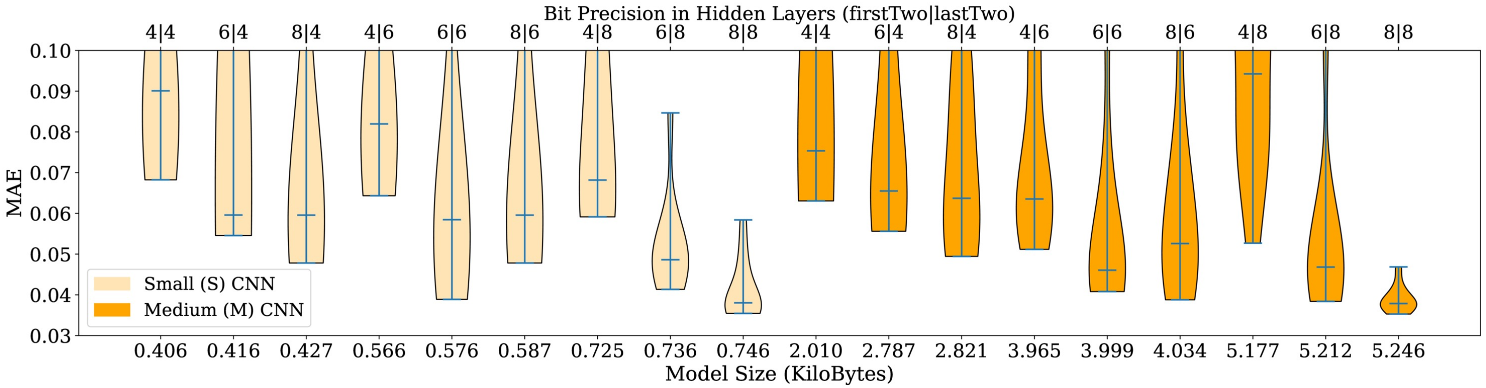
Larger MLP can be pruned further.

CNN can be sparsified greatly without loosing accuracy.

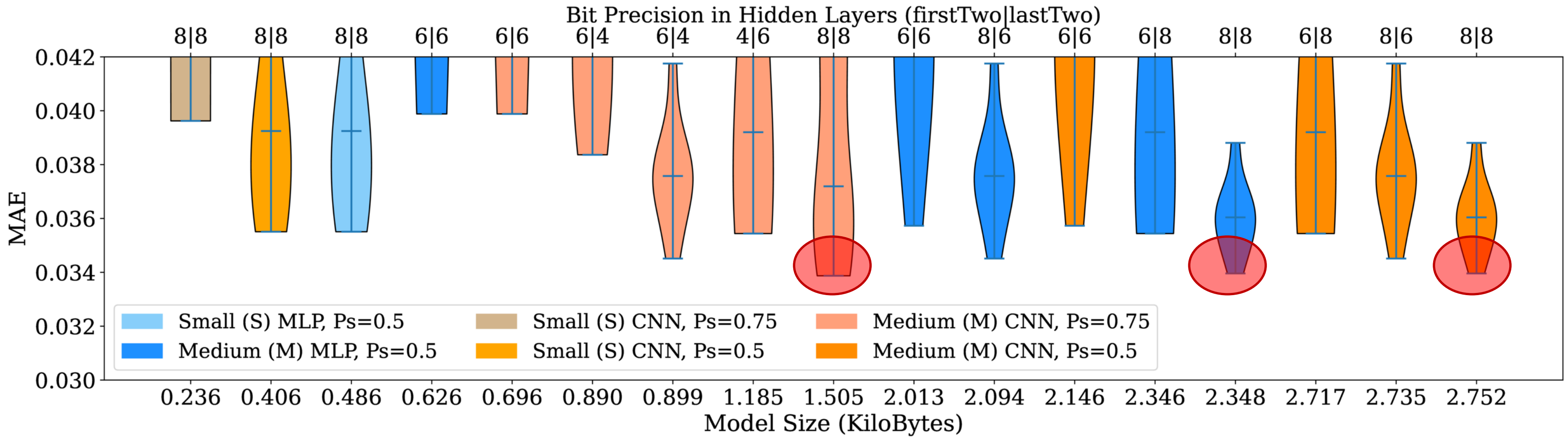
Pruning & Quantization



(a) $\mathrm{MLP_S}$ (b) $\mathrm{CNN_S}$ (d) $\mathrm{MLP_M}$ (e) $\mathrm{CNN_M}$

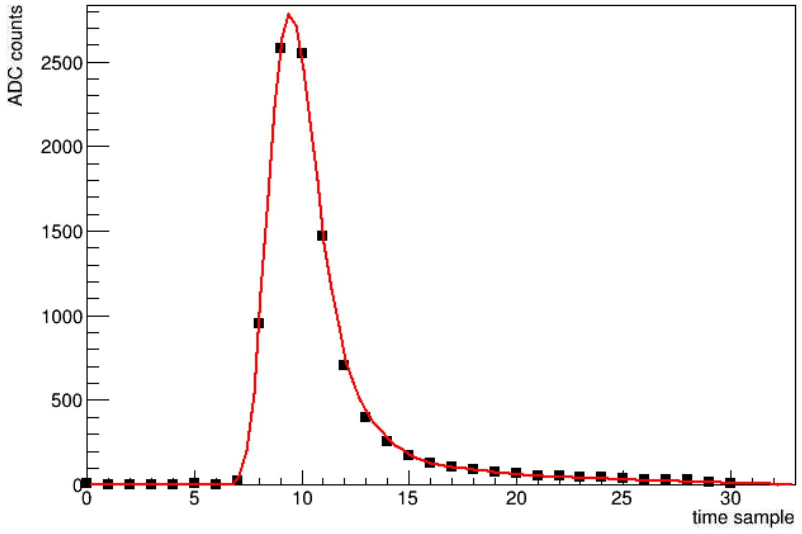


QAT+Pruning



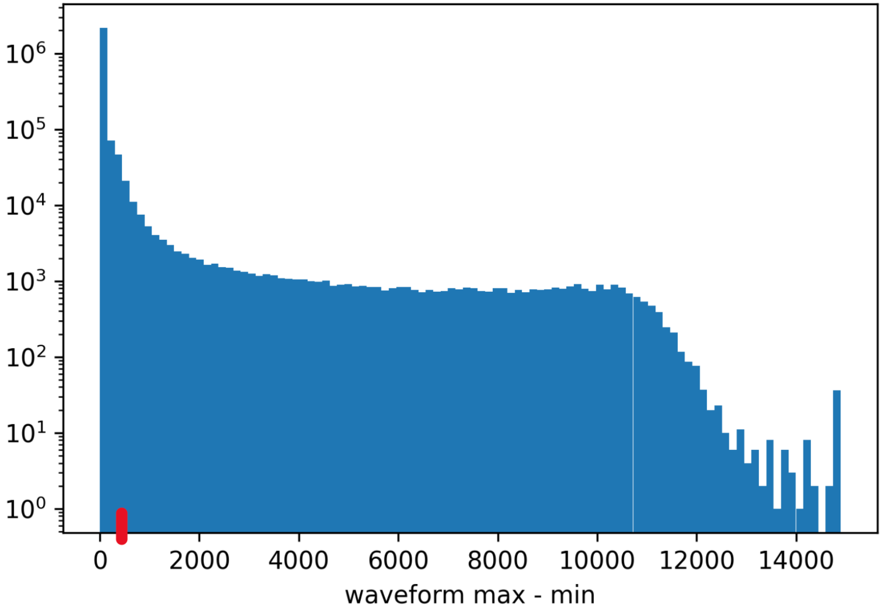
- Ref:
  - Y. Ren @Workshop IX on Streaming Readout [link]
  - Miryala, S., Mittal, S., Ren, Y., Carini, G., Deptuch, G., Fried, J., ... & Zohar, S. (2022). Waveform processing using neural network algorithms on the front-end electronics. Journal of Instrumentation, 17(01), C01039. [link]
  - Miryala, S., Zaman, M. A., Mittal, S., Ren, Y., Deptuch, G., Carini, G., ... & Katkoori, S. (2022, April). Peak prediction using multi layer perceptron (mlp) for edge computing asics targeting scientific applications. In 2022 23rd International Symposium on Quality Electronic Design (ISQED) (pp. 1-6). IEEE. [link]

sPHENIX Test-beam data



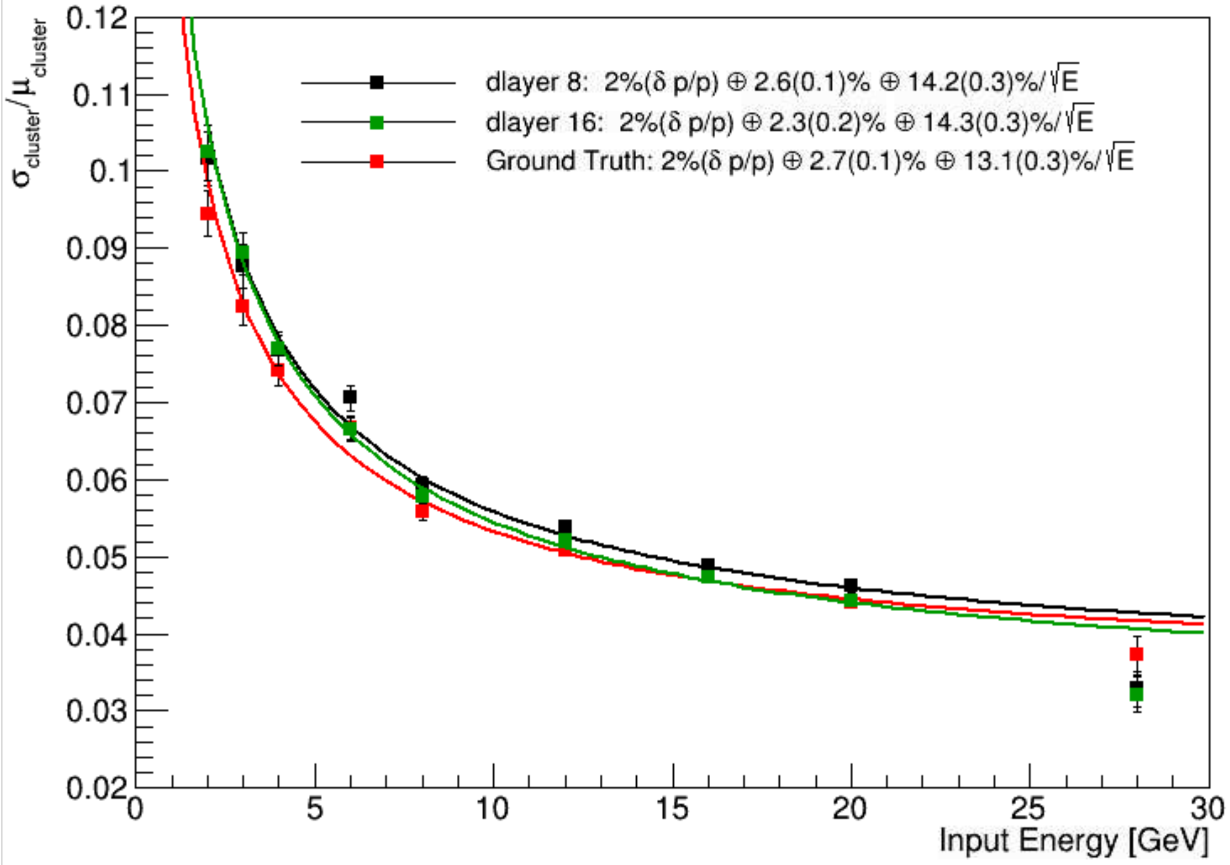
- sPHENIX EMCal 2018 test-beam data

  doi.org/10.1109/TNS.2020.3034643
- Trained on waveforms from 20 GeV incident electrons
- Ground truth (peak value) is provided by validated Template Fitting method.
- 3-layer CNN-1D models.



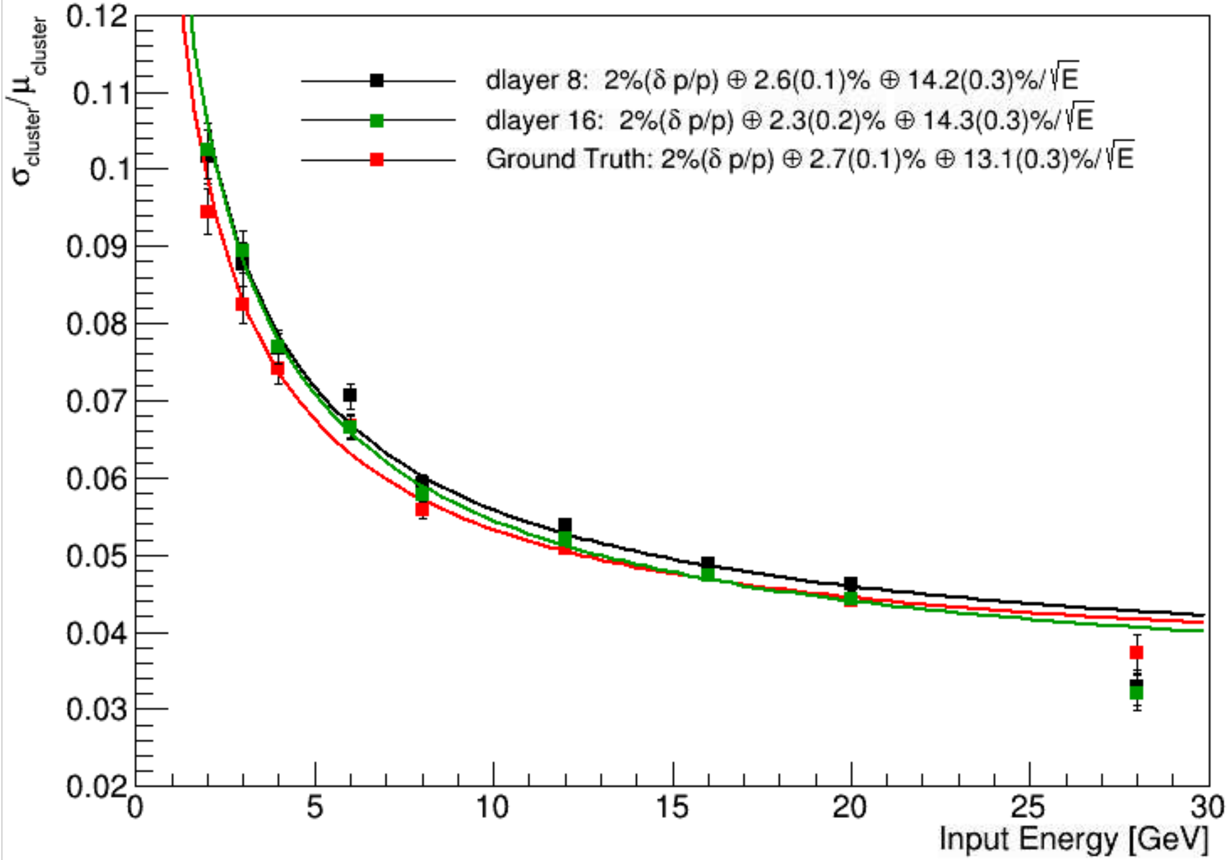
sPHENIX Test-beam data

- "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



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



The supervised AI/ML model’s accuracy is capped by the quality of data labels. How to go beyond?

Brookhaven National Laboratory

Motivation

What AI/ML can do, without labeled training data?

Can we leverage prior knowledge (i.e. simulations)?

How to tackle the gap or discrepancy between simulation and experiments?

“All models are wrong, but some are useful”. George E. P. Box

Simulations:

- Can get the fundamentals correct,
- Inexpensive to run,
- Freedom of choosing parameters.

Experiments:

- Evidence for scientific advancement,
- Very expensive to run,
- “Ground truth” unknown

Motivation

Domain A
(Simulation)

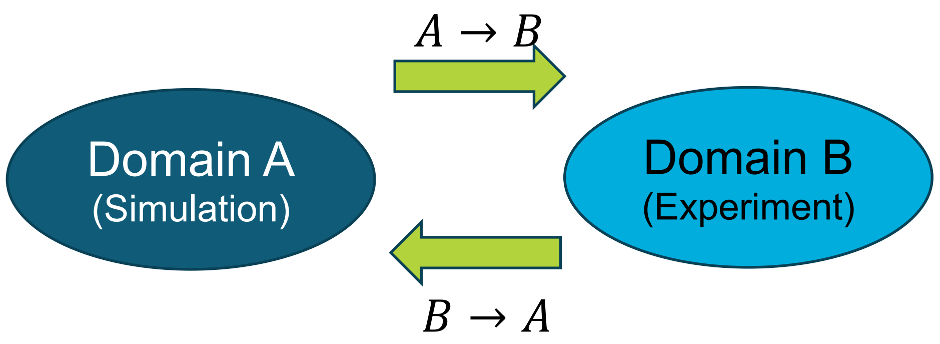
Domain B
(Experiment)

- Cause: difference between two data distributions (“domain shift”)
- Existing remedies:
  - Data Augmentation. (Heuristics, domain-agnostic, use case-dependent.)
  - Domain Adaptation. (Task-specific, require trained model & data annotation.)
  - Transfer Learning. (Require data annotation.)

Brookhaven National Laboratory

Task-agnostic Data Translation

Directly translate or enhance simulation data to make them more realistic.
Ideally, the ground truth is retained during the translation, and systematic difference is bridged.



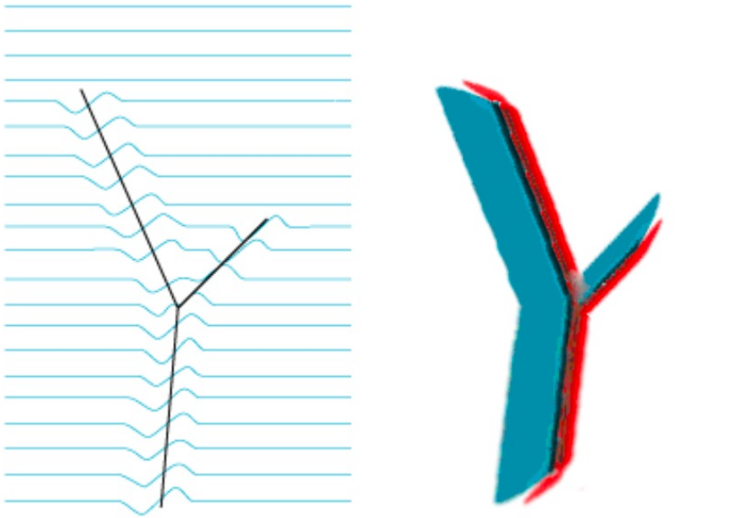
- $A \rightarrow B$: “Augmented High-Fidelity Simulation” that can produce “labeled” data.
- $B \rightarrow A$: “Data Cleaning” that can remove noise of experiment data.
- Analysis tools (w/ human-intelligence) can have better and more data to work with.
- ML models have labeled data to train and are easier to transfer to the real data.

DUNE and LArTPC

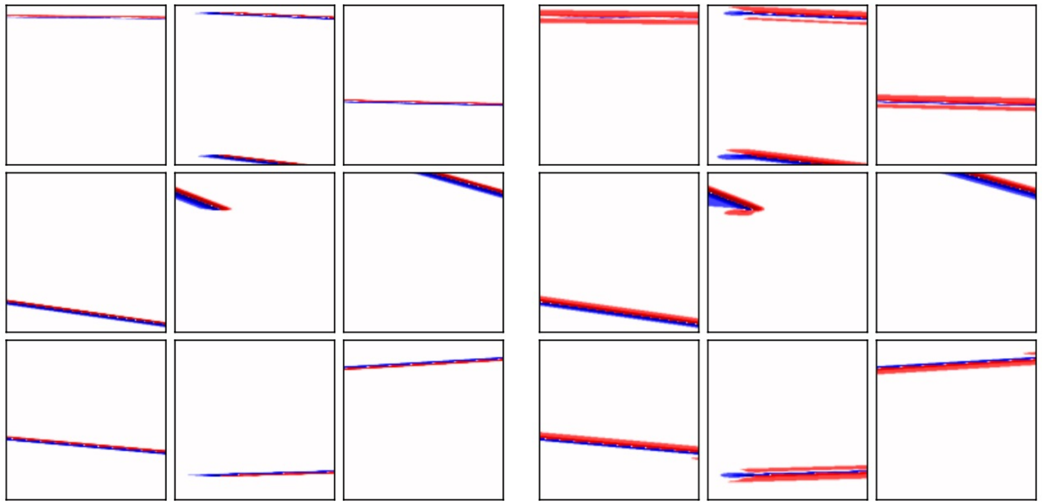
Before doing this on real data, we would like to study a task under well-understood settings:

1. **Domain A** – simplified detector response, where a cloud of electrons is read only by the nearest wire.
2. **Domain B** – realistic detector response, where a cloud of electrons can produce excitations in multiple wires.

Also applicable for gas-medium TPC in both SPHENIX and STAR.



(a) Schematic Waveforms (b) Actual Data Sample

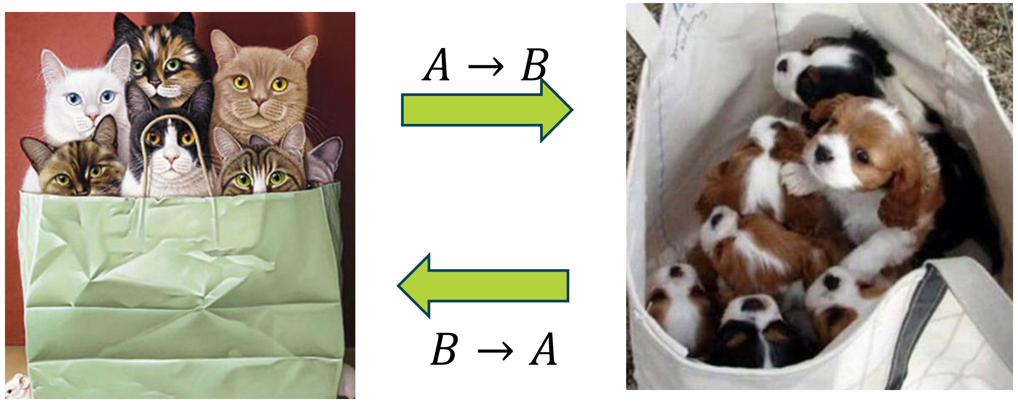


(a) **Domain A** (b) **Domain B**

Brookhaven National Laboratory

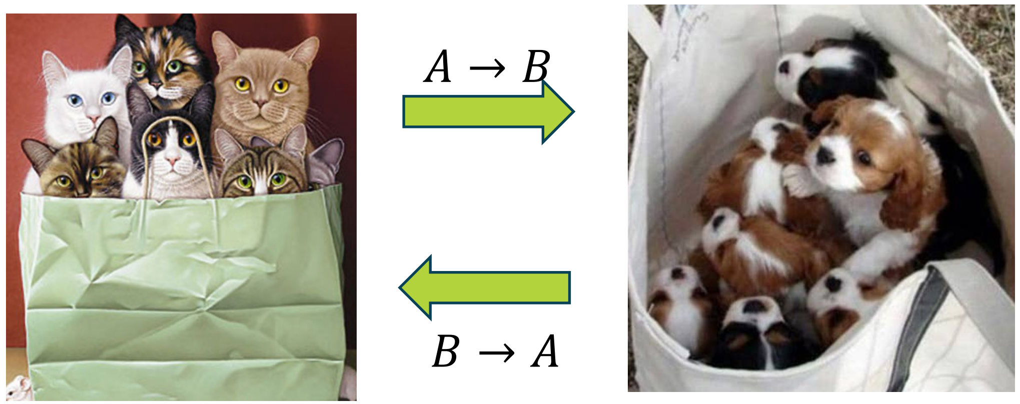
Unpaired Image-to-image translation

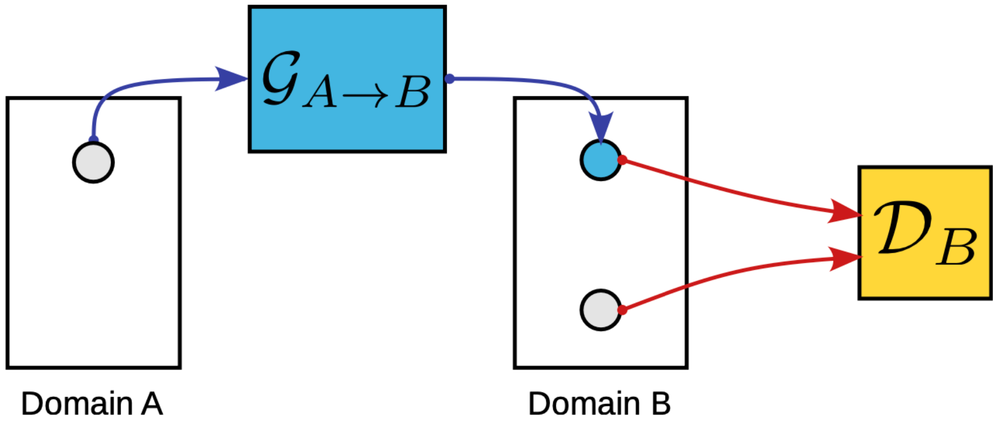
Unpaired constraint: since the ground truth of the experimental data is unknown, it’s impossible to generate matched simulation images.



<u>Unpaired</u> Image-to-image translation

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



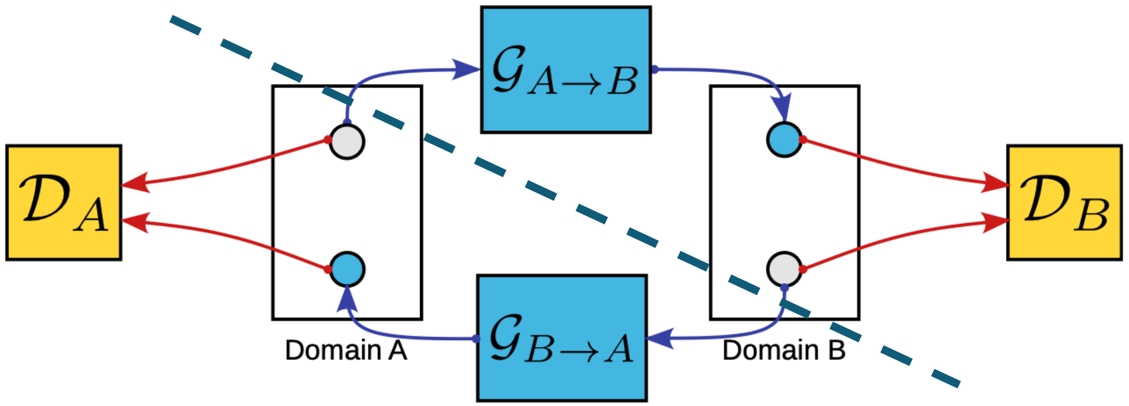


A popular way for generative tasks is GAN. However, GAN is prone to "mode collapse".

<u>Unpaired</u> Image-to-image translation

CycleGAN connects two sets of Generator and Discriminator.

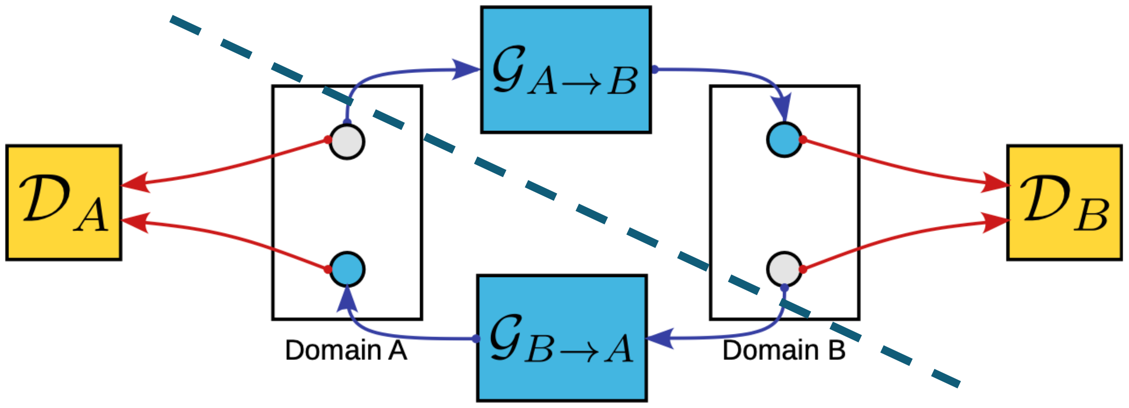
* *"Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks"* Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Proceedings of the ICCV 2017



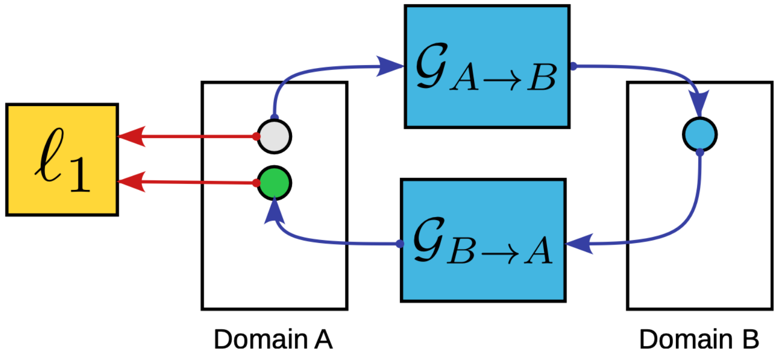
<u>Unpaired</u> Image-to-image translation

CycleGAN connects two sets of Generator and Discriminator.

* "*Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks*" Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Proceedings of the ICCV 2017



$$\mathcal{G}_{B\to A}(\mathcal{G}_{A\to B}(X_A)) == X_A$$

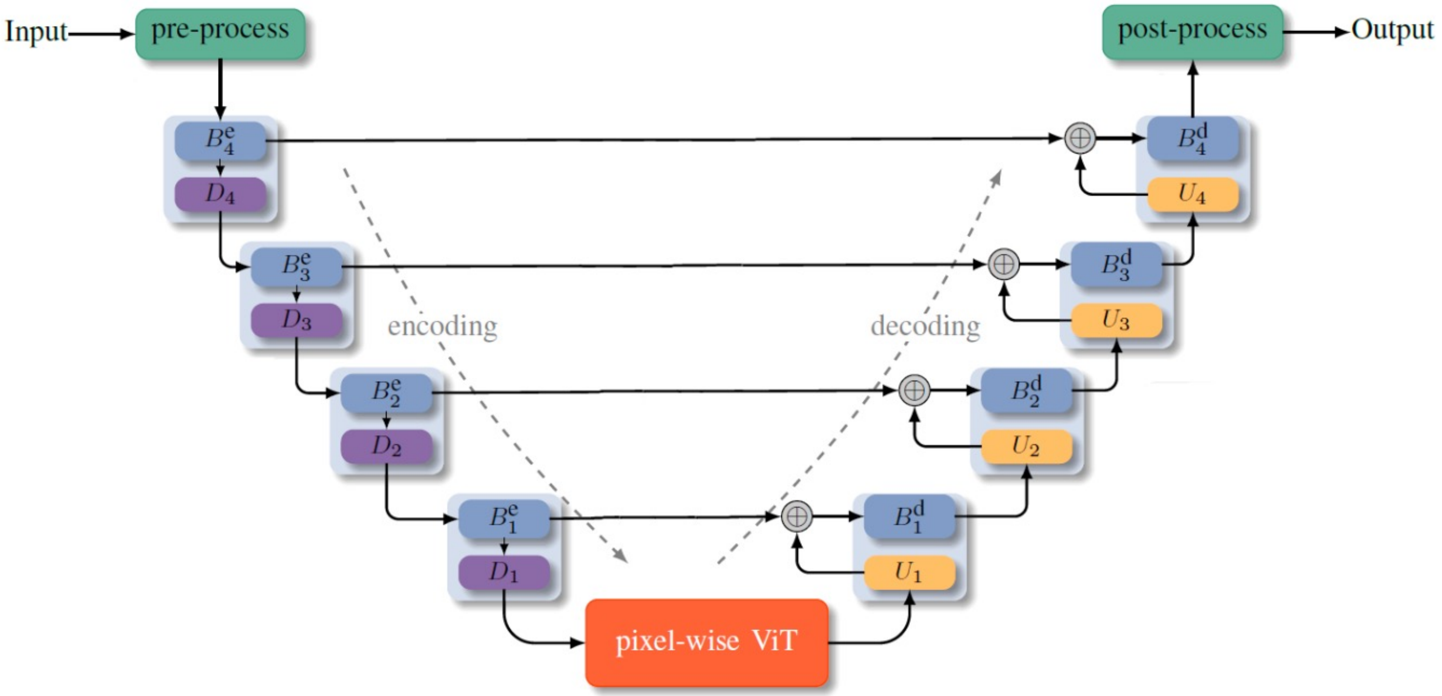


Unet-ViT-CycleGAN (UVCGAN)

Adding a ViT block at the bottleneck of the Unet improves long-range pattern learning.

Self-supervised pre-training.

And other tricks.



"UVCGAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation", [arxiv: 2203.02557]

Brookhaven National Laboratory

UVCGAN fixes rough edges

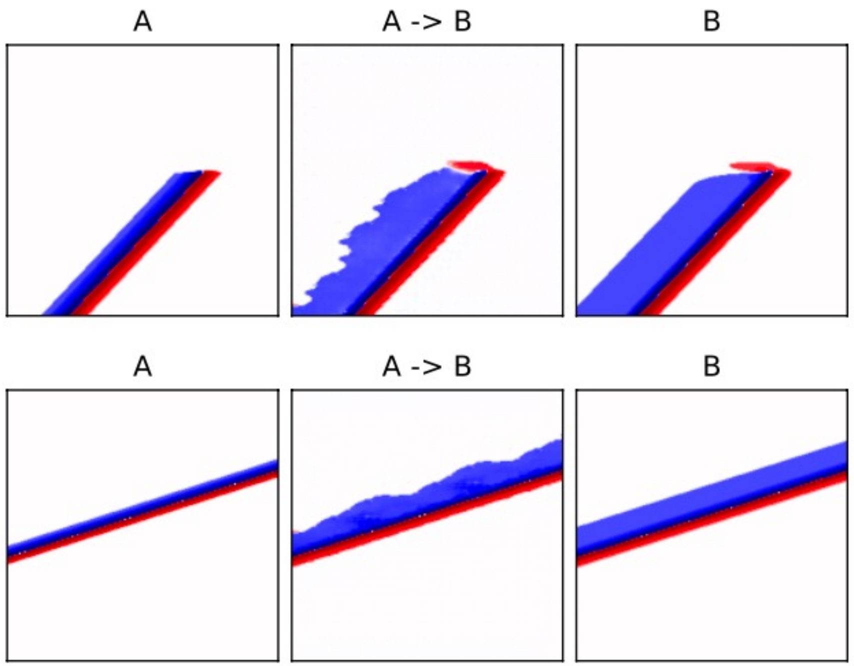


Figure: Default CycleGAN Generator

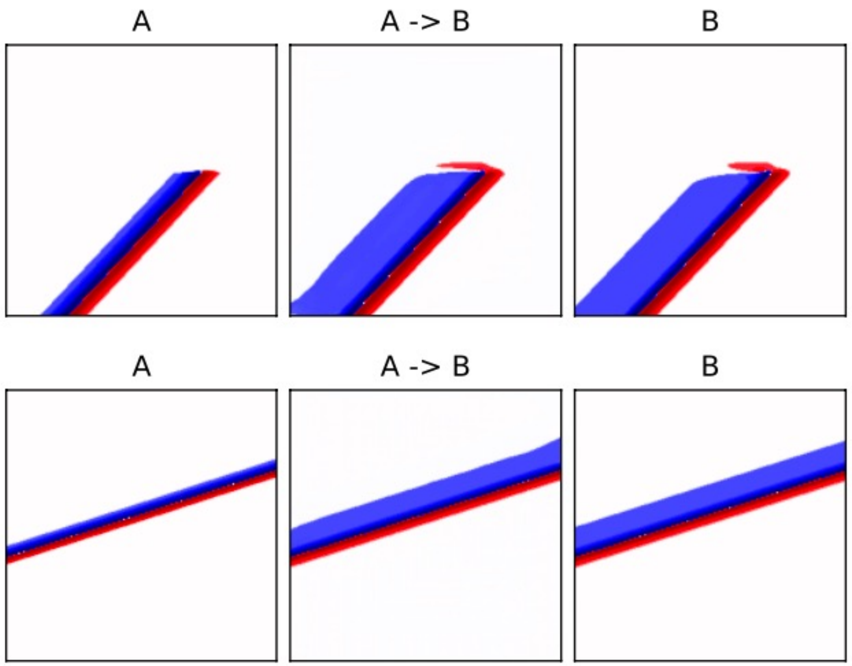


Figure: New UNet-ViT Generator

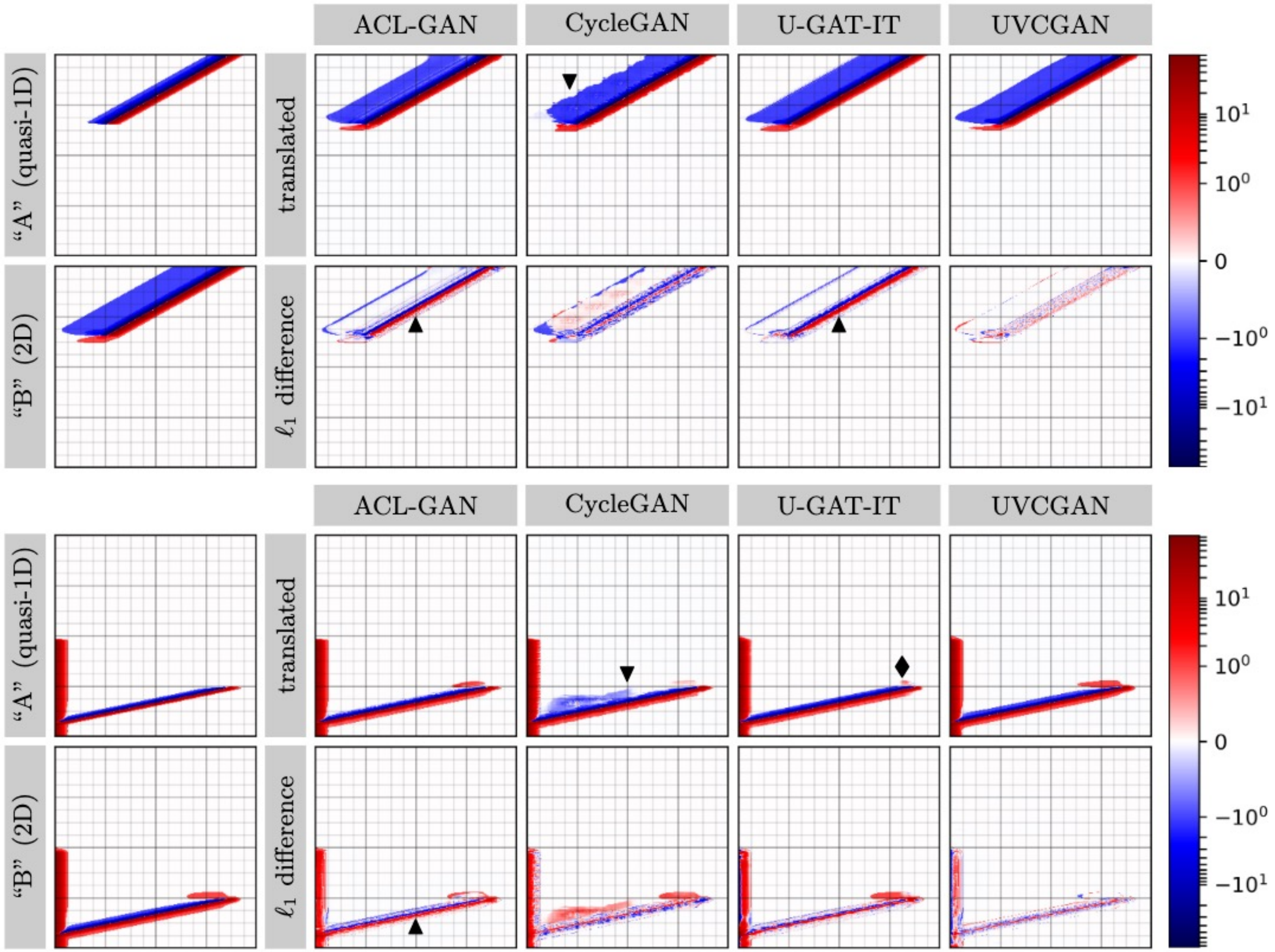
Results

We have compared our model (UVCGAN) vs advanced models:

1. ACL-GAN arXiv:2003.04858
2. CycleGAN arXiv:1703.10593
3. U-GAT-IT arXiv:1907.10830

algorithm	"A" to "B" ℓ_1	"A" to "B" ℓ_2	"B" to "A" ℓ_1	"B" to "A" ℓ_2
ACL-GAN	0.083	0.566	0.039	0.121
CycleGAN	0.074	0.180	0.061	0.159
U-GAT-IT	0.078	1.187	0.073	1.161
UVCGAN	**0.030**	**0.033**	**0.025**	**0.027**

A->B



arxiv: https://arxiv.org/abs/2304.12858 (under revision)
data released: https://zenodo.org/record/7809108#.ZDV0B-zMKvB

Results

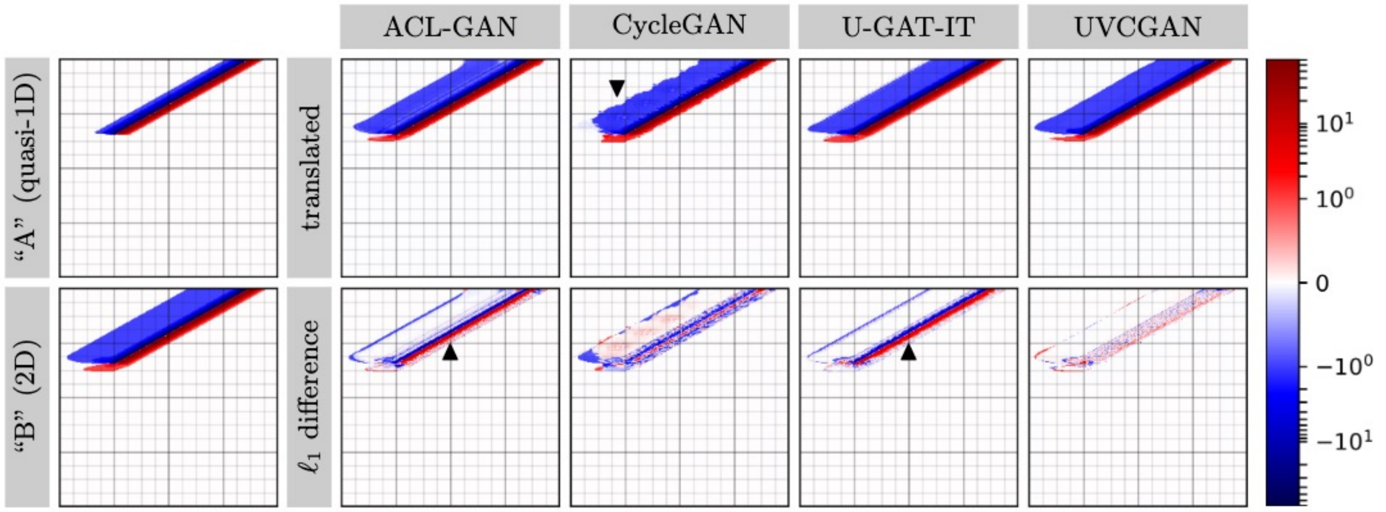
We have compared our model (UVCGAN) vs advanced models:

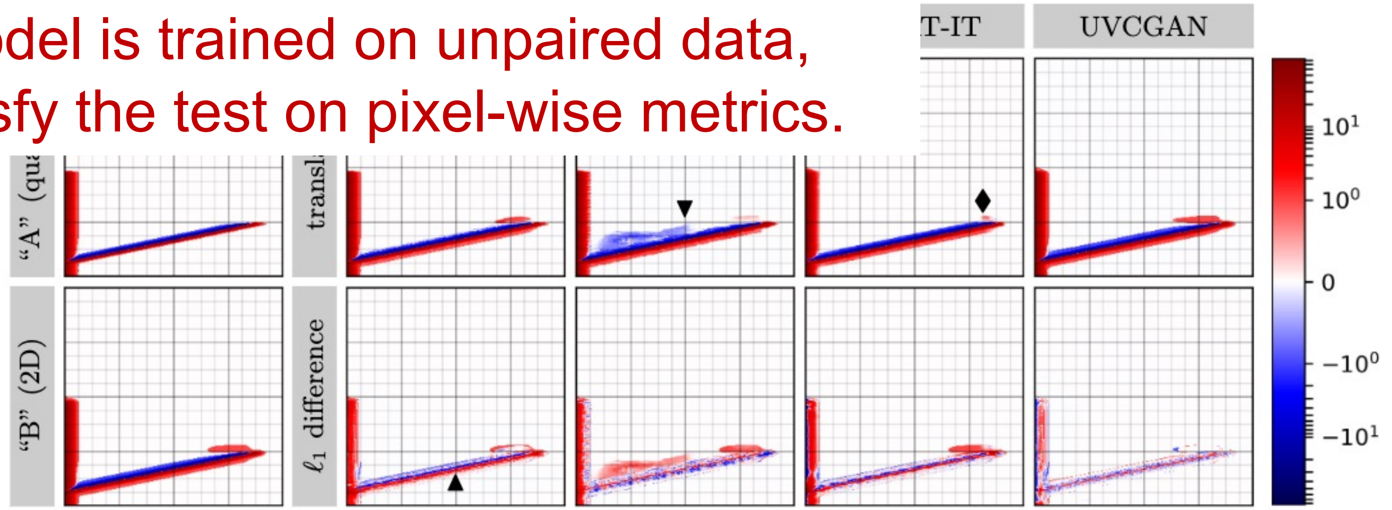
1. ACL-GAN arXiv:2003.04858
2. CycleGAN arXiv:1703.10593
3. U-GAT-I

Key Takeaway: The model is trained on unpaired data, but translation can satisfy the test on pixel-wise metrics.

algorithm	"A" to "B" ℓ_1	"A" to "B" ℓ_2	"B" to "A" ℓ_1	"B" to "A" ℓ_2
ACL-GAN	0.083	0.566	0.039	0.121
CycleGAN	0.074	0.180	0.061	0.159
U-GAT-IT	0.078	1.187	0.073	1.161
UVCGAN	**0.030**	**0.033**	**0.025**	**0.027**

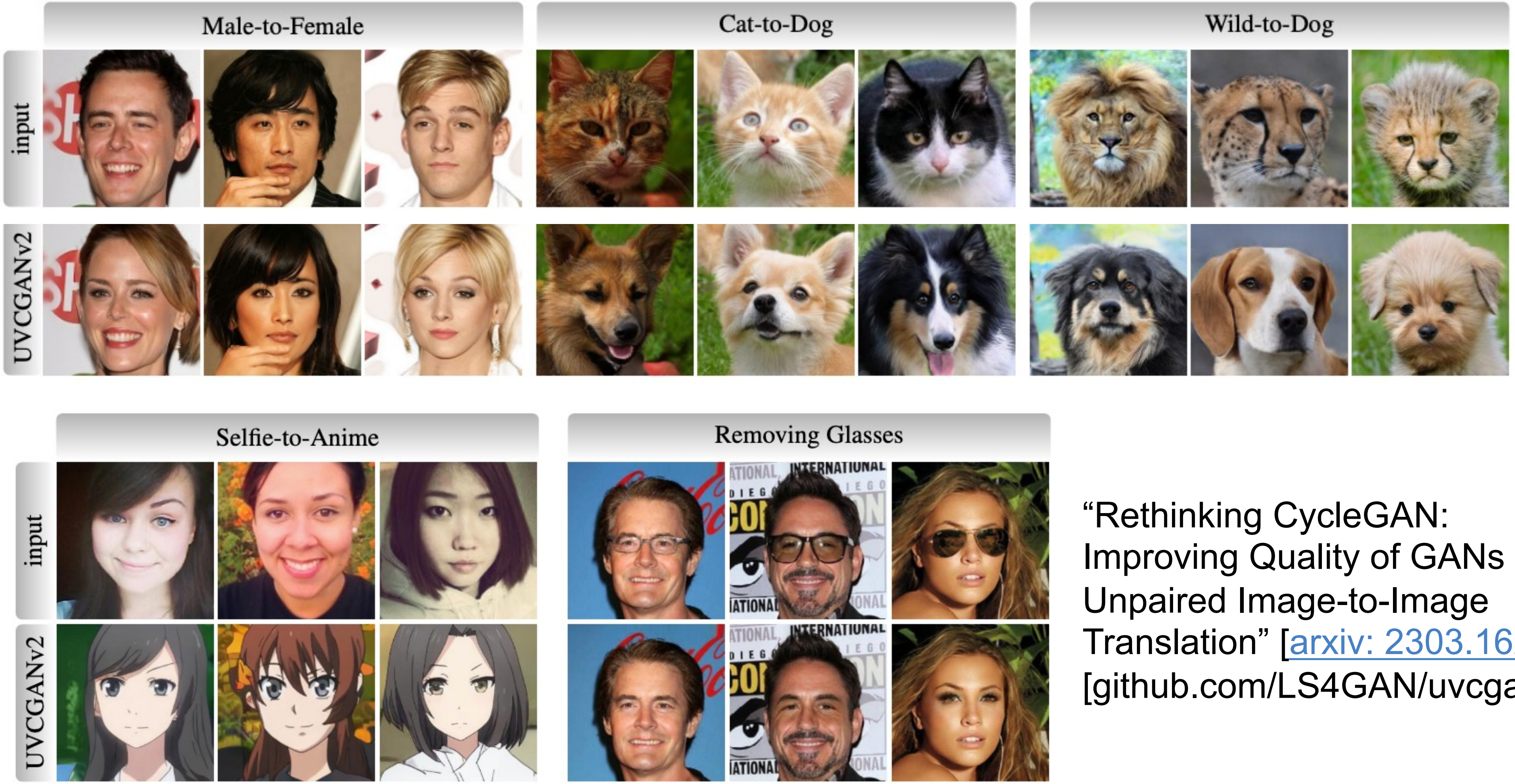
A->B





arxiv: https://arxiv.org/abs/2304.12858 (under revision)
data released: https://zenodo.org/record/7809108#.ZDV0B-zMKvB

UVCGAN-v2, on open-benchmark data

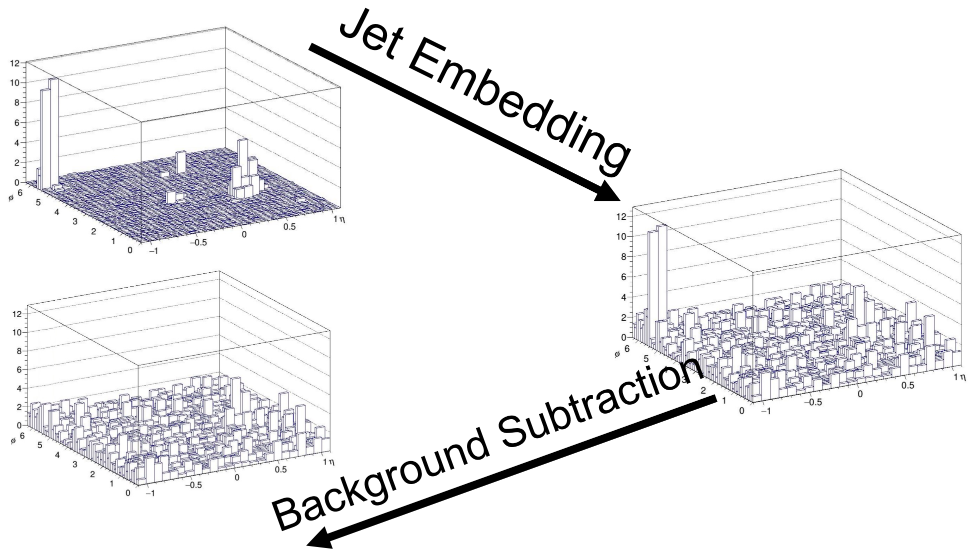


“Rethinking CycleGAN: Improving Quality of GANs for Unpaired Image-to-Image Translation” [arxiv: 2303.16280] [github.com/LS4GAN/uvcgan2]

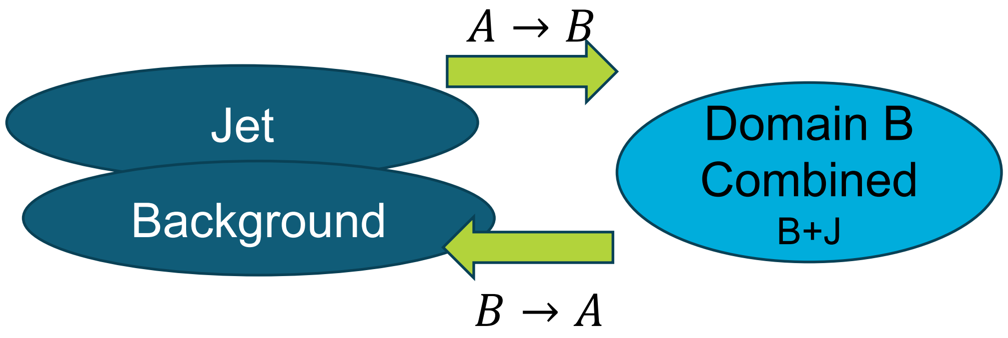
Jet Data Generation

Data Generation:

- Domain A background and jet samples were generated using Pythia and HIJING respectively.
- Generated events are then passed through a geant mock up of the sPHENIX calorimeter system to better reproduce real measurements.

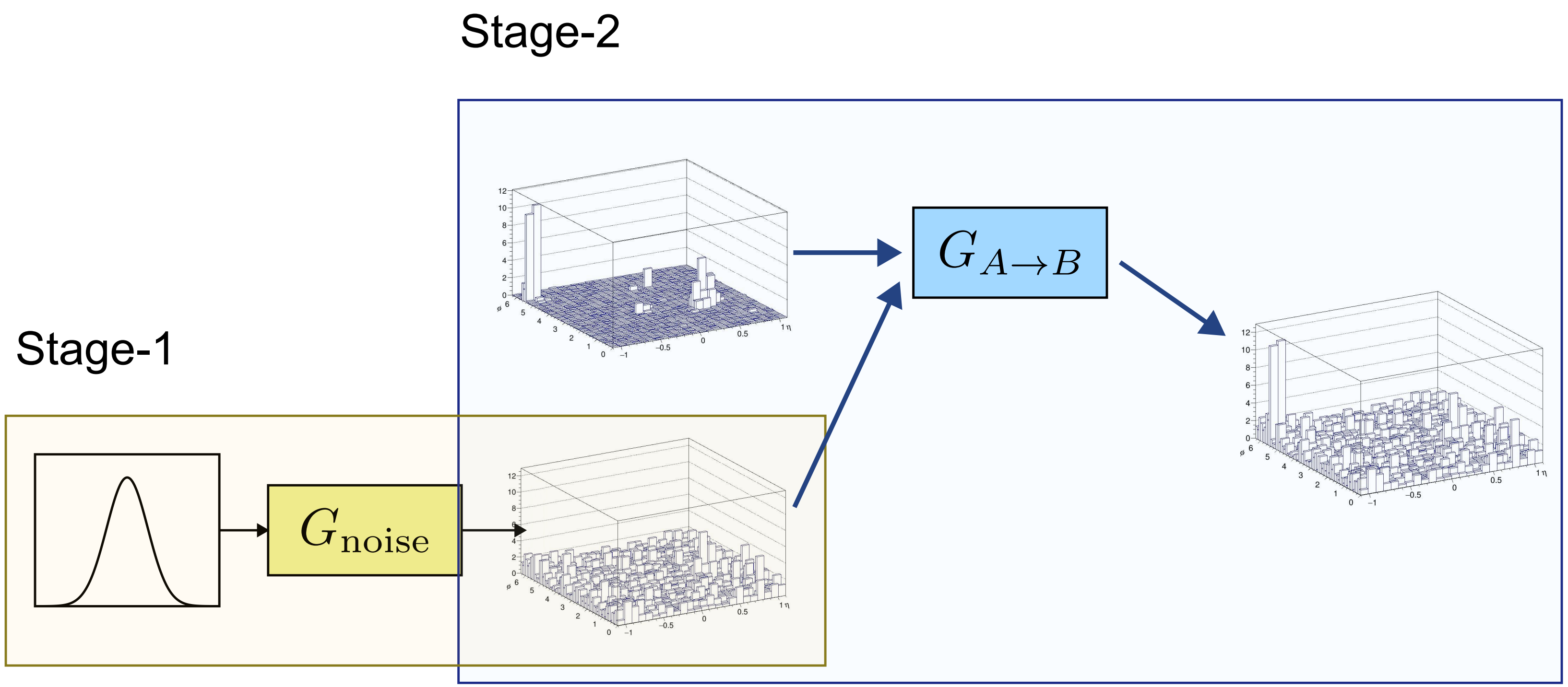


- Domain **A:**
  - Heavy Ion Background (HIJING, 0-10% centrality events)
  - Jets (Pythia, Flows Afterburner, etc.)
- Domain **B:**
  - Samples are combined with a straight addition of the energy depositions.
  - In future, +M, and real experimental data.
- Instances from A and B are Unpaired.



Another talk: “Interpretable Machine Learning Methods for to Jet Background Subtraction in Heavy Ion Collisions”
Speaker: Mr Tanner Mengel (University of Tennessee)

Two-Stage Approach

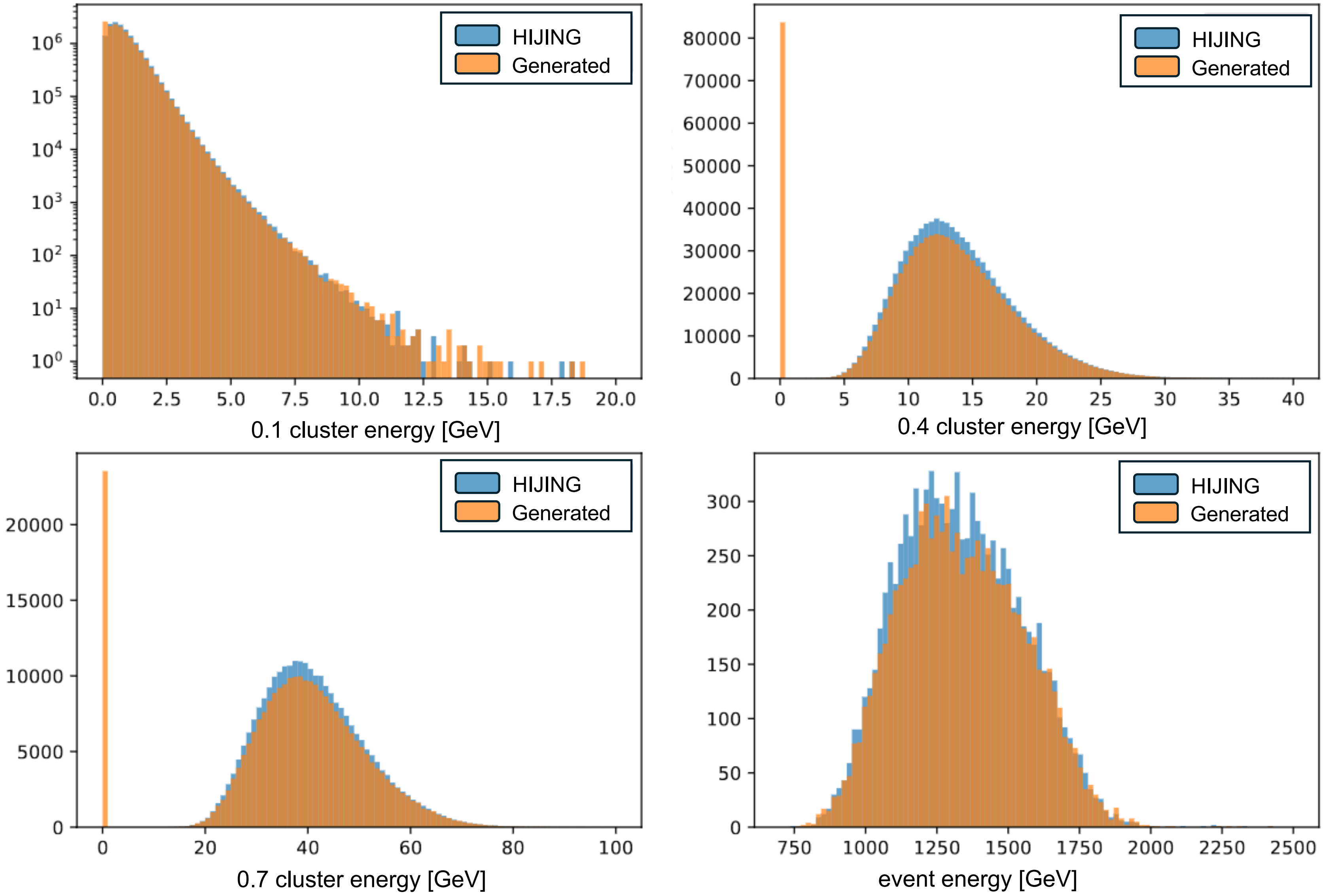


Results on Background Generation

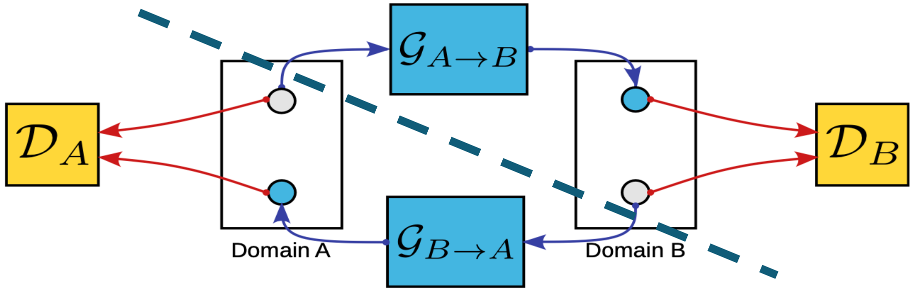
Preliminary

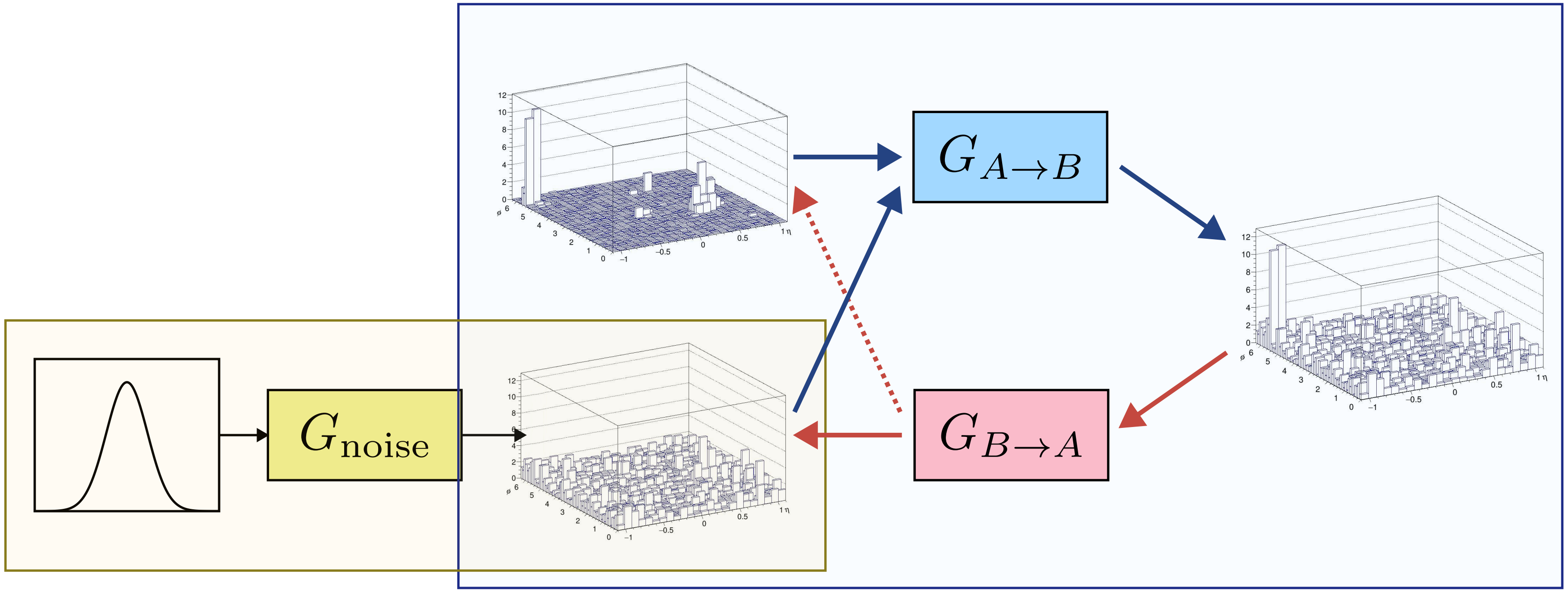
Comparing Distributions:

- 1x1
- 4x4
- 7x7
- event-level

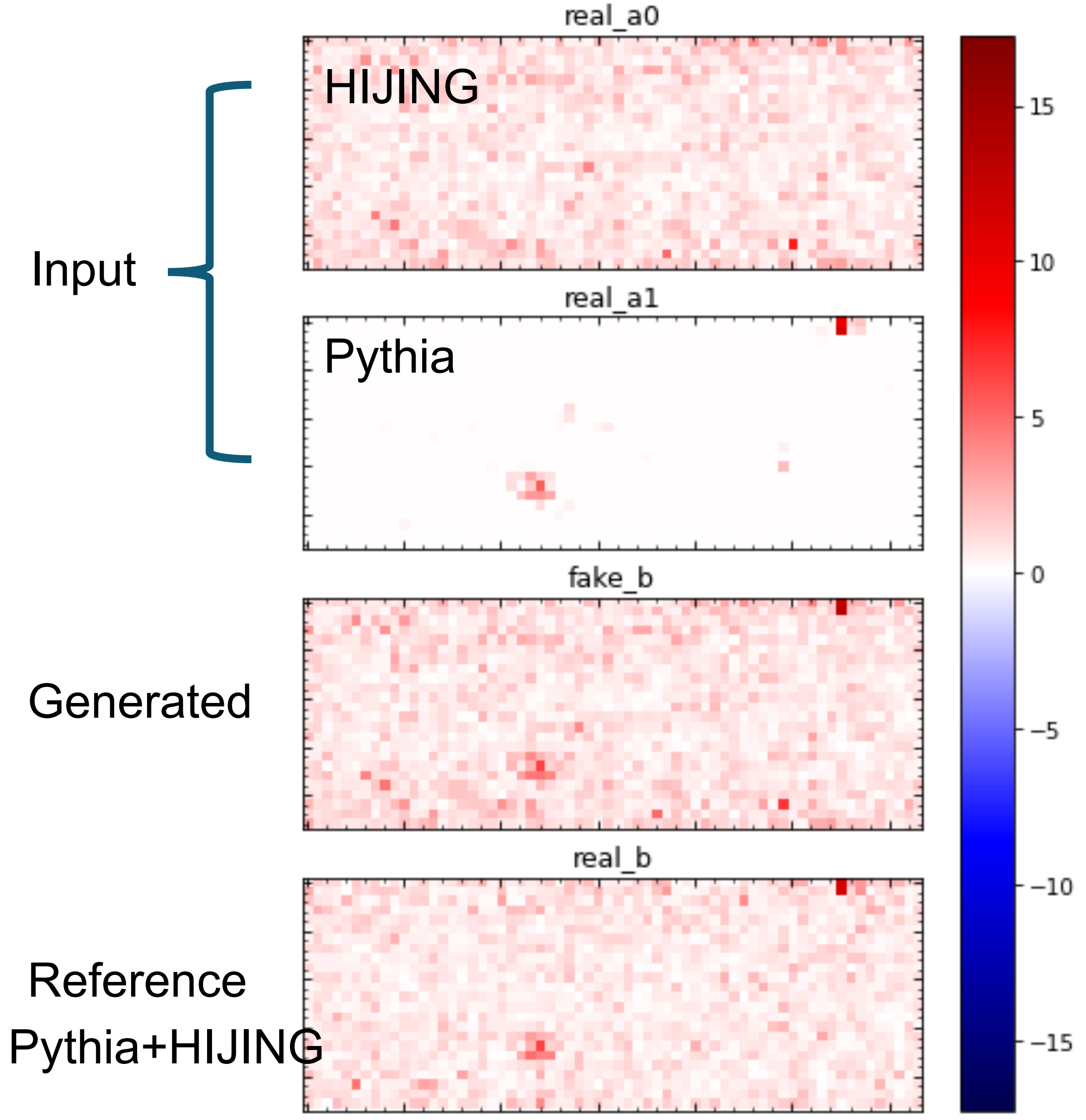


Asymmetric CycleGAN

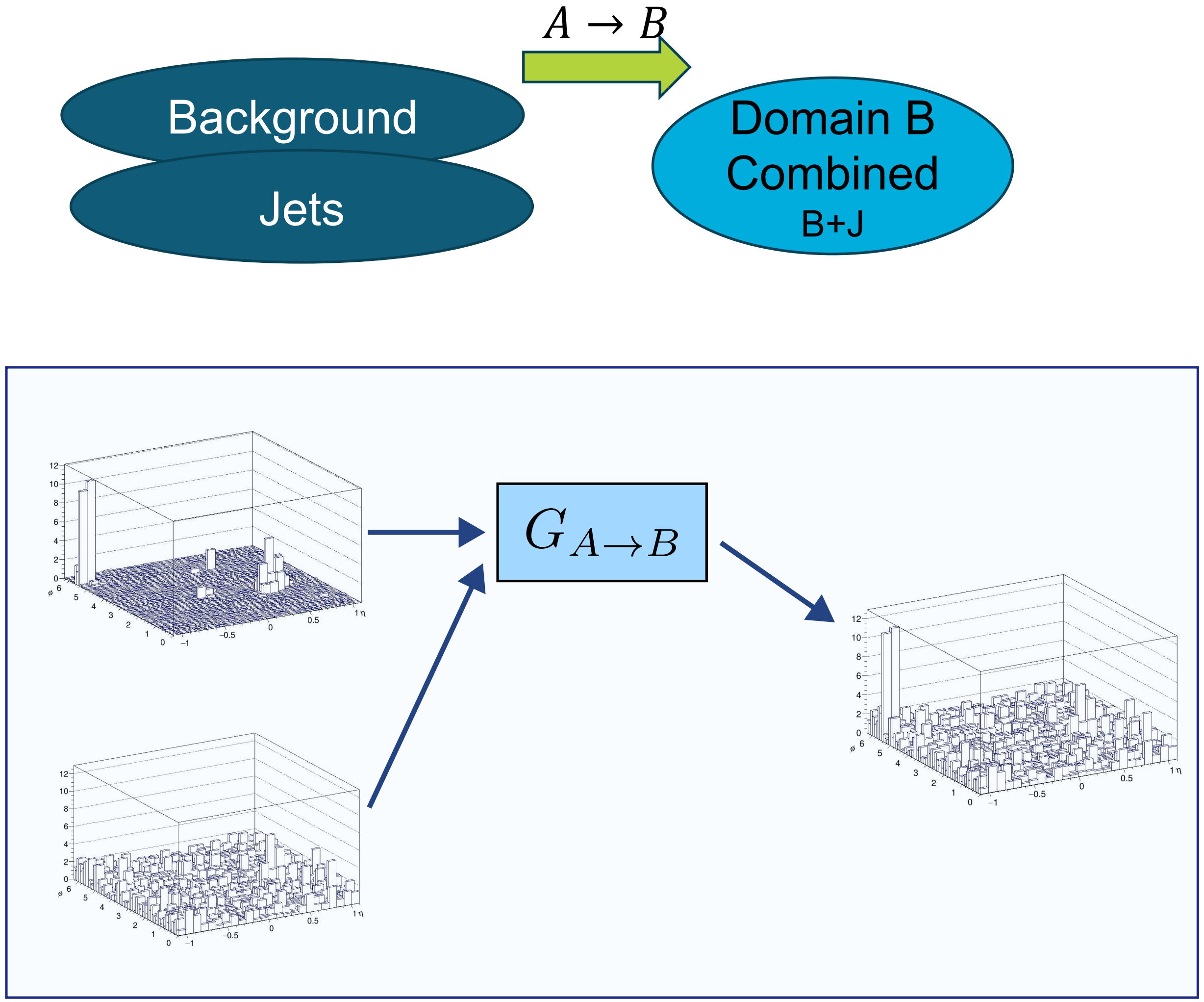




Preliminary Results on $A \leftrightarrow B$



(a) $A \to B$



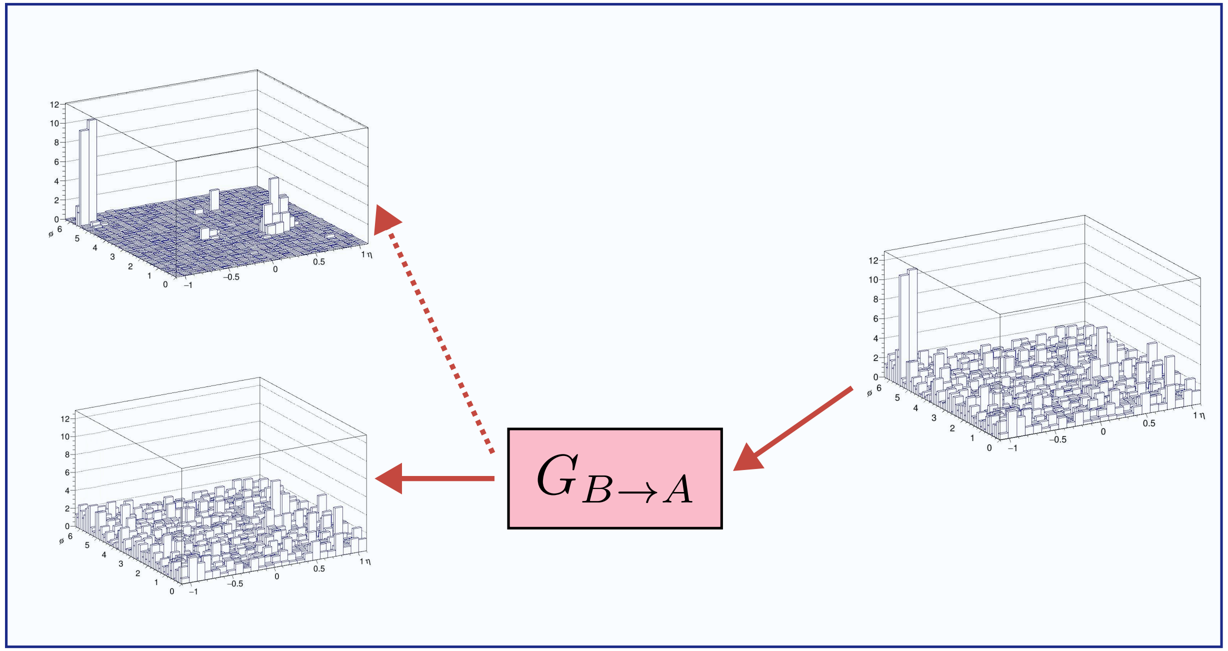
Preliminary Results on $A \leftrightarrow B$

Background

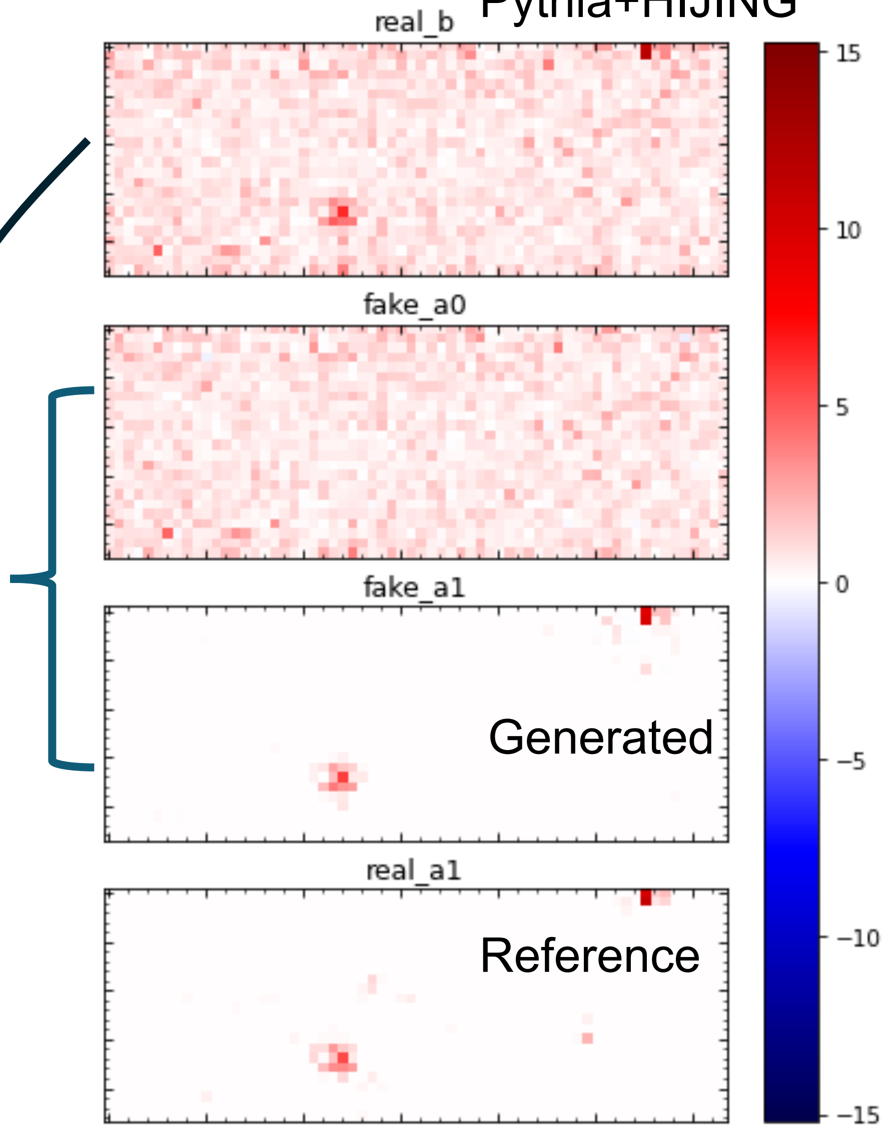
Jets

Domain B Combined B+J

$B \rightarrow A$



Pythia+HIJING



(b) $B \rightarrow A$

Future Study

- These are very early results that we are excited to share.
- In future, we will incorporate “media modification” (Jet+Backgrounds).
- Any suggestions and comments would be very helpful.
- How to validate when we apply this to real experimental data?
- Any other constraints we should consider?

Collaborators & Ack.

- Timothy Rinn, Yeonju Go, Jin Huang, David Morrison
- Dmitrii Torbunov, Haiwang Yu, Brett Viren, Chao Zhang, Xin Qian
- Piotr Maj, Soumyajit Mandal, Prashansa Mukim, Grzegorz Deptuch, Gabriella Carini
- Elizabeth Brost, Haider Abidi , Viviana Cavaliere, Michael Begel
- Yi Huang, Shubha Khrael, Meifeng Lin, Shinjae Yoo

- LDRD-19-028 High-Throughput Advanced Data Acquisition for eRHIC, Particle Physics and Cosmology Experiments
- LDRD-21-023 Towards Edge Computing: A Software and Hardware Co-Design Methodology for ASIC-based Scientific Neuromorphic Computing
- LDRD-21-029 Bridging the Gap between Scientific Simulations and Experiments
- with Cycle-Consistent Generative Models
- LDRD-22-018 Real-time Image Classification using Machine Learning
- LDRD-23-048 Real-time Information Distillation on Novel AI Hardware

References

1. Huang, Yi, et al. "Efficient data compression for 3d sparse tpc via bicephalous convolutional autoencoder." *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2021.
2. Miryala, Sandeep, et al. "Waveform processing using neural network algorithms on the front-end electronics." *Journal of Instrumentation* 17.01 (2022): C01039.
3. Miryala, Sandeep, et al. "Peak prediction using multi layer perceptron (MLP) for edge computing asics targeting scientific applications." *2022 23rd International Symposium on Quality Electronic Design (ISQED)*. IEEE, 2022.
4. Torbunov, Dmitrii, et al. "Uvcgan: Unet vision transformer cycle-consistent GAN for unpaired image-to-image translation." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.
5. Huang, Yi, et al. "Unsupervised Domain Transfer for Science: Exploring Deep Learning Methods for Translation between LArTPC Detector Simulations with Differing Response Models." *arXiv preprint arXiv:2304.12858* (2023).
6. Torbunov, Dmitrii, et al. "Rethinking CycleGAN: Improving Quality of GANs for Unpaired Image-to-Image Translation." *arXiv preprint arXiv:2303.16280* (2023).

Thank You!

Yihui “Ray” Ren
<yren@bnl.gov>

We are hiring!

Google BNL CSI Jobs