



AI/ML Workshop Summary

RHIC
RELATIVISTIC HEAVY ION COLLIDER
COLLISION EVENT
Au + Au
 $\sqrt{s_{NN}} = 200 \text{ GeV}$

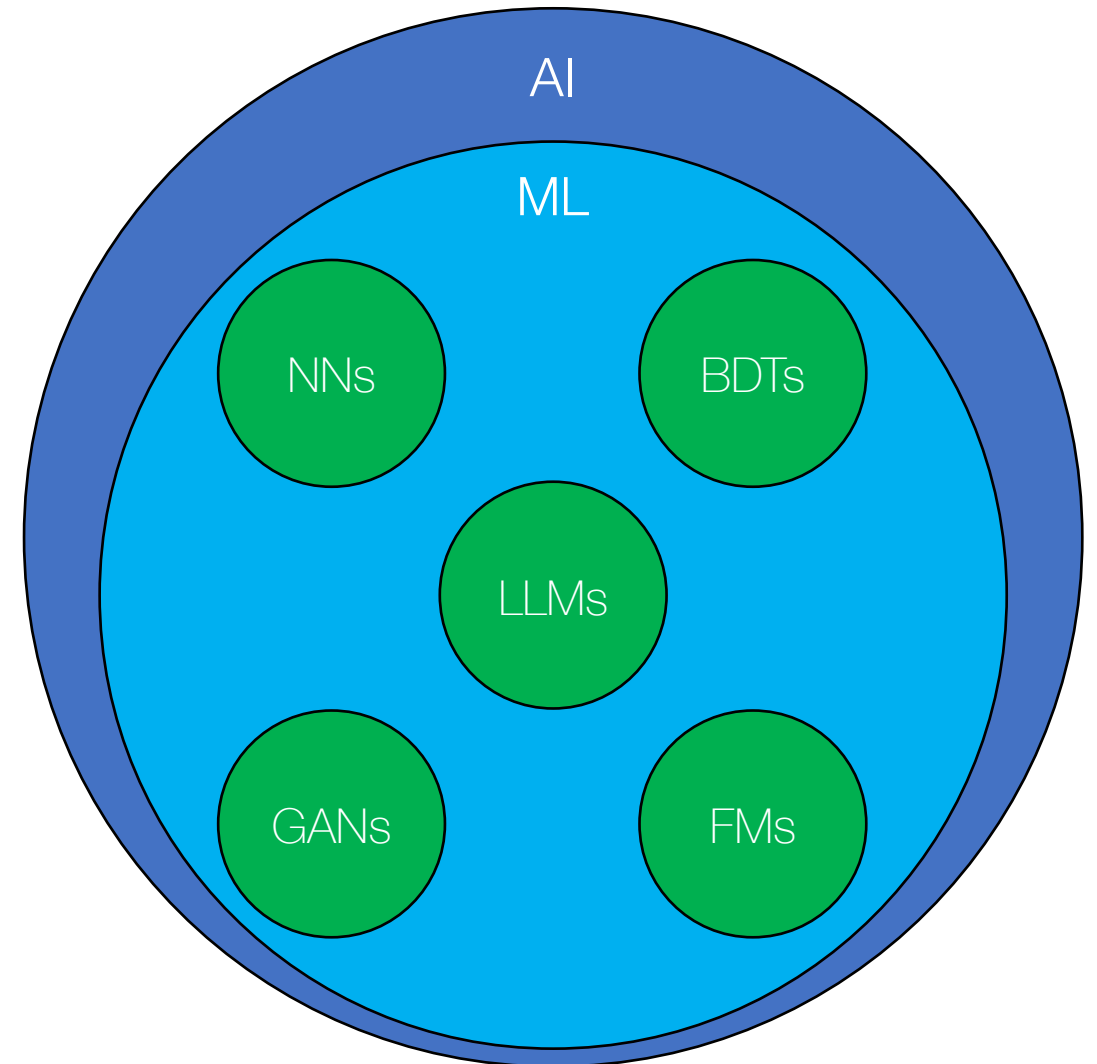
SPECTATOR VIEW
TIME: $3.2 \times 10^{-23} \text{ s}$
ENERGY DEPOSITION

Cameron Dean
Massachusetts Institute of Technology
RHIC/AGS Annual Users Meeting
May 13th 2026



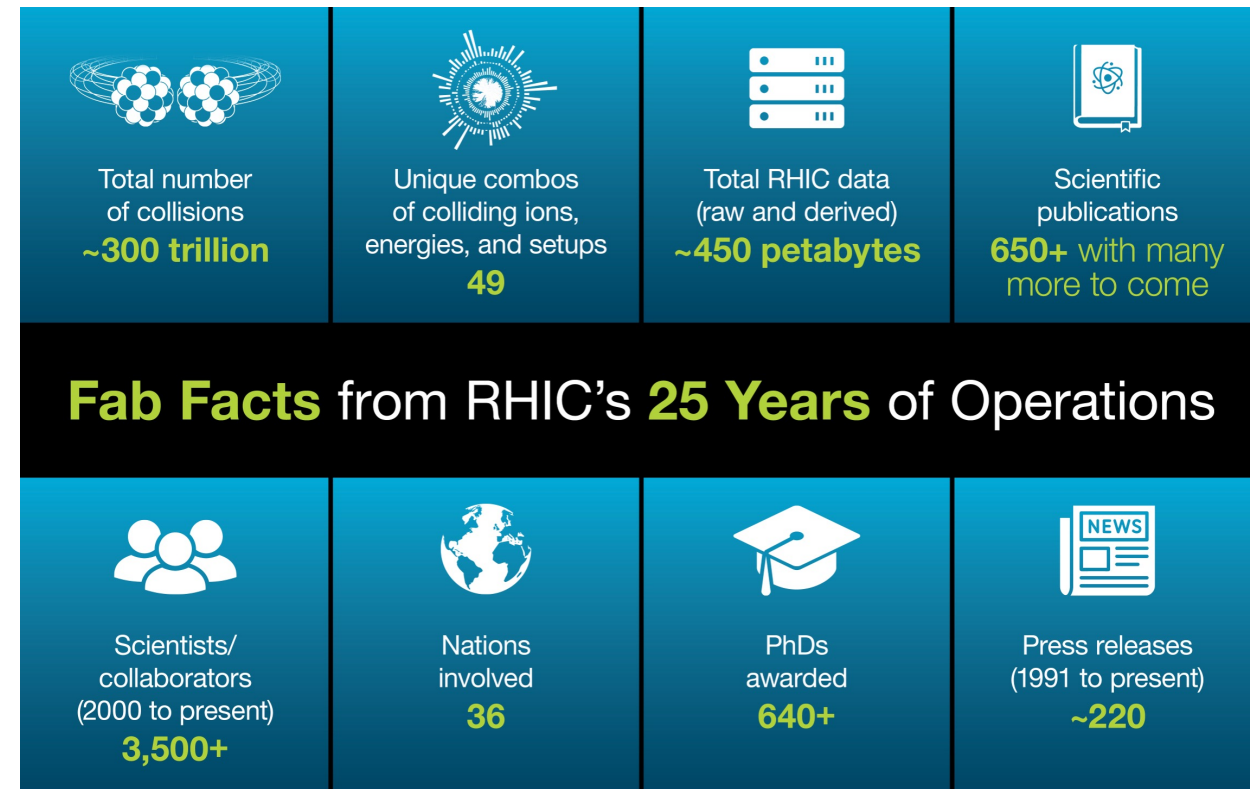
What is AI/ML?

- Artificial Intelligence is the top-level concept enabling decision making
- Machine Learning is a subset of AI concerned with training and applications
- The public often confuses AI with “Artificial General Intelligence”
 - This is when machines exhibit behavior we would describe as indistinguishable from humans
- Most ML algorithms use “supervised learning” (we pass the data and answer)
- “Unsupervised learning” is becoming more apparent (the input data is unlabeled)



Why do we want it for RHIC/EIC science?

- Our data samples are huge!
 - Costly and time consuming to process
 - Can this be optimized?
- RHIC is 26 years old
 - Lots of knowledge accumulated over the years
 - Original analysts/experts may no longer be in the field
 - Researchers can spend lots of time searching documents for relevant information
- RHIC science has not ended (there is the RHIC Analysis Group at BNL)
 - We need to preserve knowledge to complete the RHIC science mission
- The EIC is coming
 - How do we best optimize design/operation/analysis? This is the time to implement



<https://www.bnl.gov/newsroom/news.php?a=122794>

Workshop overview

- Held yesterday from 9am to 3pm
- We had 7 contributions
 - Unfortunately, an 8th contributor, Alexandr Prozorov, had to withdraw last week due to visa requirements
- Wide variety of topics, from EIC accelerator optimization to chat bots
- Good representation from labs and universities

Variable Rate Neural Compression for TPC Data

RHIC annual user meeting workshop
05-11-2026

UVCGAN-S: Unsupervised Background Subtraction of Calorimeter Jets in Heavy-Ion Collisions

Shuhang Li

Y. Go, D. Torbunov, Y. Huang, S. Li, T. Rinn, H. Yu, B. Viren, M. Lin, Y. Ren, D. Persepolis, J. Huang
arXiv:2510.23717 (2025)

COLUMBIA UNIVERSITY Brookhaven

Towards Foundation Models for Next-Generation Experiments at the EIC: A Perspective



Brookhaven National Laboratory
U.S. DEPARTMENT OF ENERGY

AI Surrogate Modeling for Fast Simulation of the Hydrodynamic Evolution of the Quark-Gluon Plasma

Computing and Data Sciences, BNL
Seungjun Lee

In collaboration with Y. Go, B. Schenke, C. Shen, Y. Ren, J. Huang, M. Chamizo-Llata

This work is funded under BNL-LDRD-26-46

RHIC/AGS Annual User's Meeting / AI/ML Workshop
05/11/2026

Brookhaven National Laboratory
U.S. DEPARTMENT OF ENERGY

SciBot: Your "AI Librarian" for Navigating RHIC's Fragmented Institutional Knowledge

2026 RHIC/AGS ANNUAL USERS' MEETING

Ankush Kanungo
for the BNL/DAP Team (M. Alf, Z. Dong, V. Garonne, J. Lauret, E. Lancon, A. Prozorov)

COLUMBIA UNIVERSITY Brookhaven

The EIC-BeamAI Collaboration: Preparing Machine Learning for the Electron-Ion Collider

Georg Hoffstaetter de Torquat
For the EIC-BeamAI collaboration

Georg.Hoffstaetter@Cornell.edu
05/11/2026

Brookhaven National Laboratory
U.S. DEPARTMENT OF ENERGY

FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics

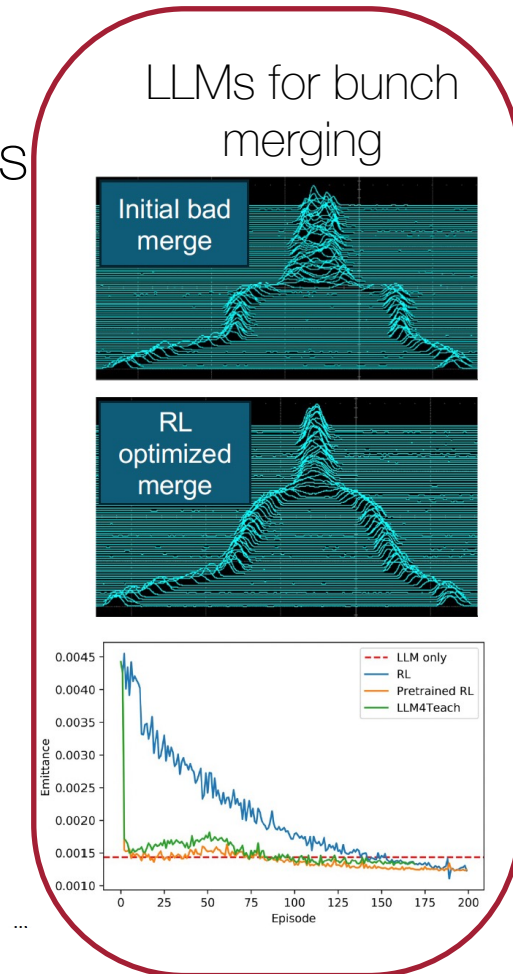
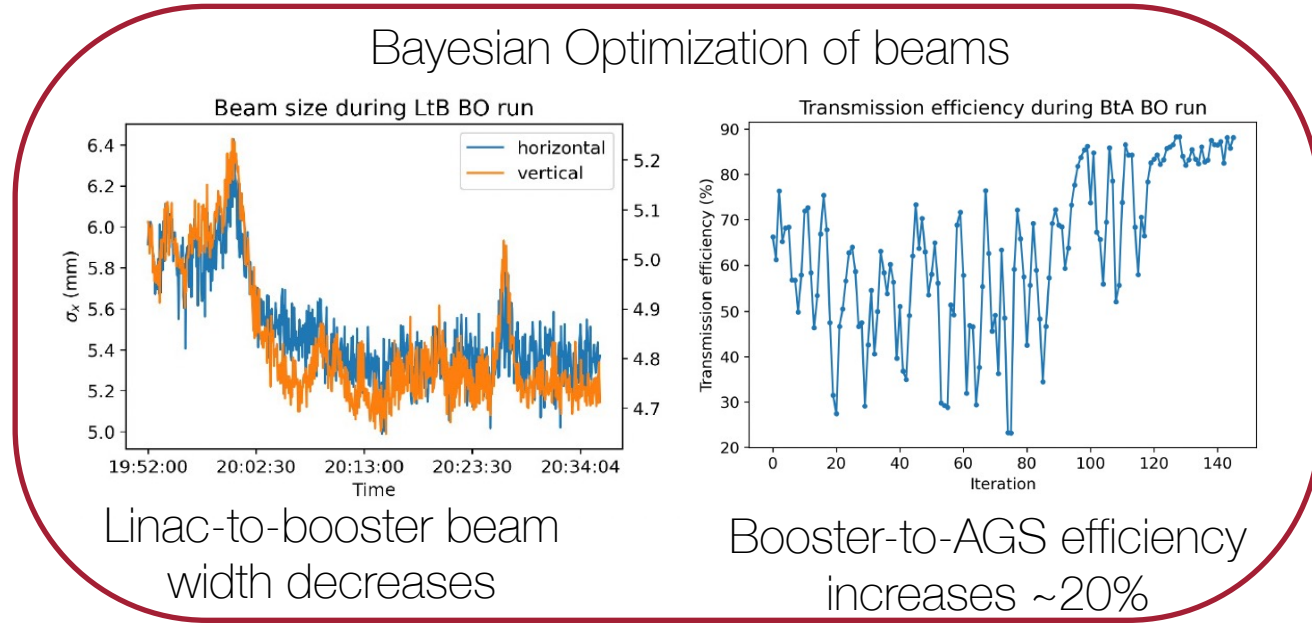
Published in ICLR 2026

Presenter:
David Park, AI Department, Brookhaven National Laboratory (dpark1@bnl.gov)

Presented at RHIC & AGS Annual Users Meeting 2026

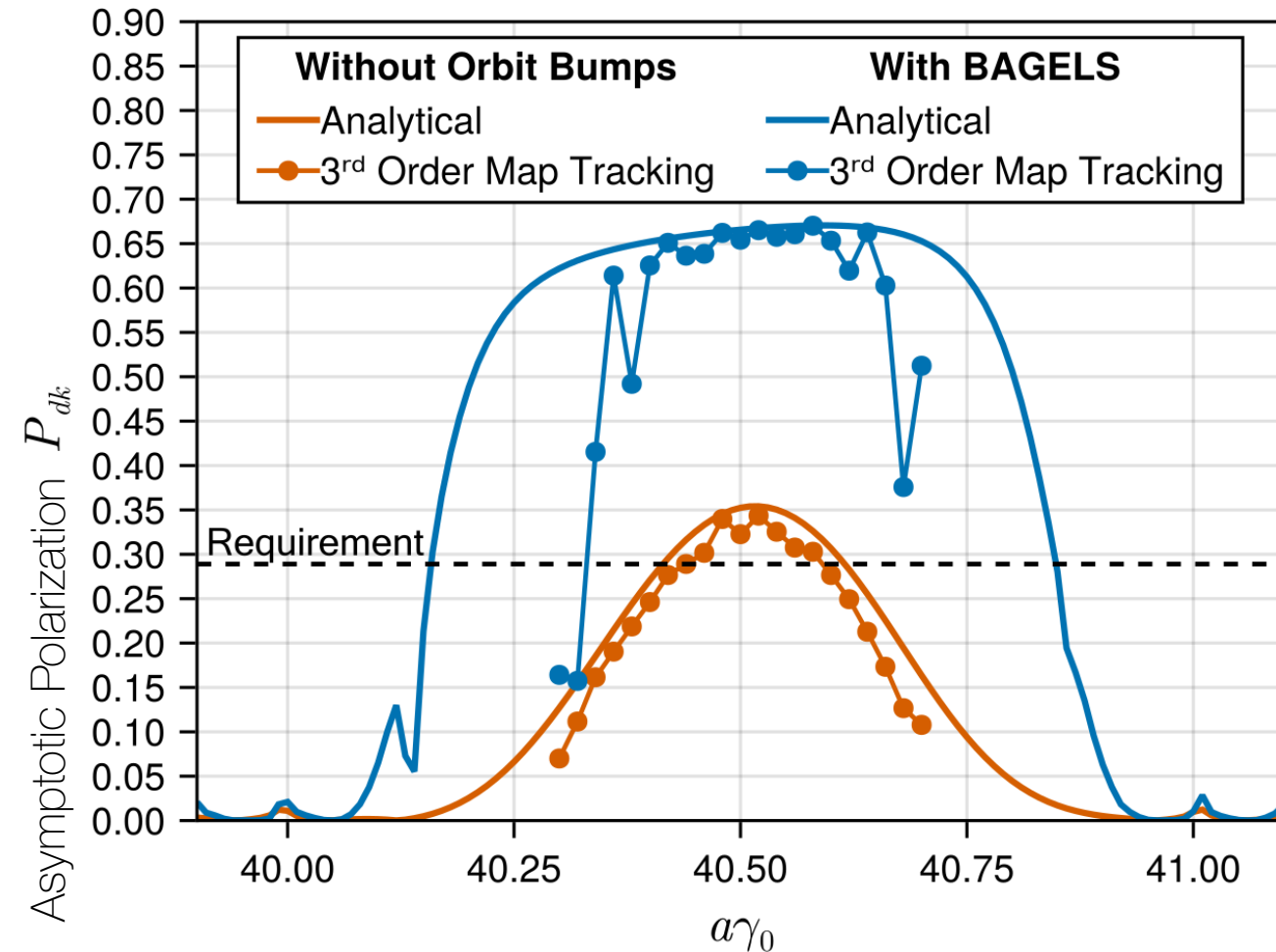
Accelerator optimization – G. Hoffstaetter

- Aim: improve brightness, polarization and efficiency
 - Uses BNL accelerator complex as testbench
- AI can be used to boost transfer efficiency
- Large Language Models can improve beam merging within a few iterations
- Digital Twins (simulated accelerators) can provide real-time feedback for EIC operations



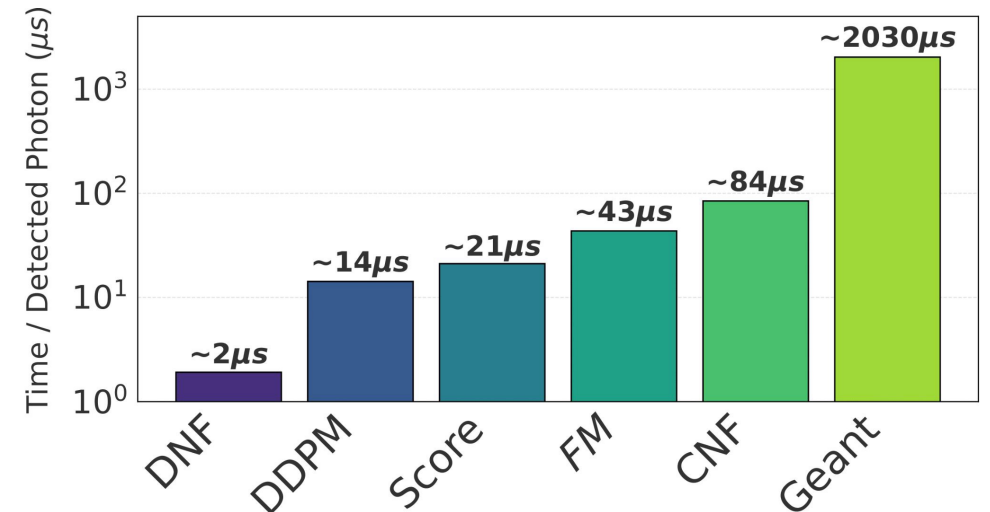
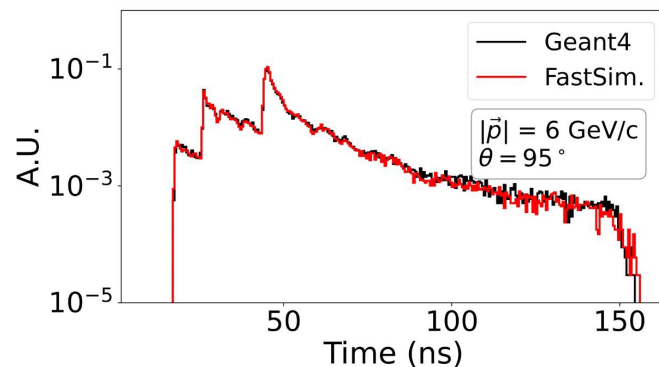
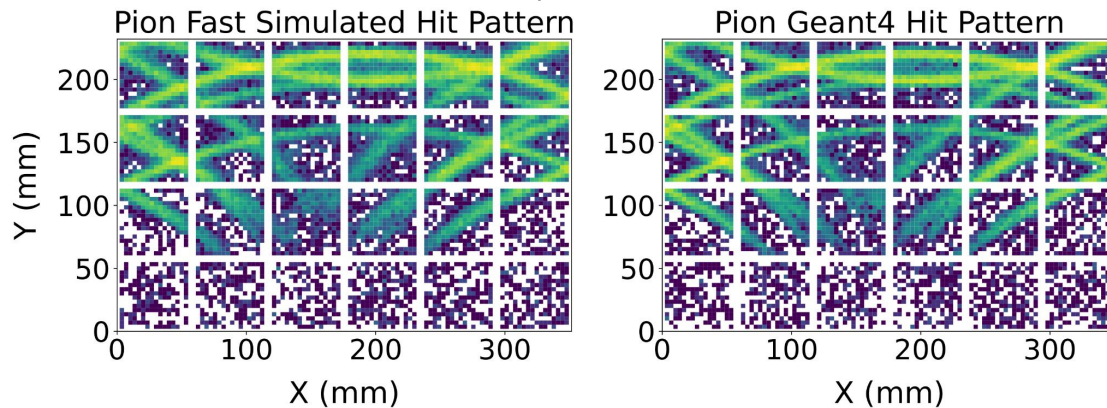
Beam optimization for the EIC – G. Hoffstaetter

- BAGELS = **B**est **A**adjustment **G**roups for **E**lectron **S**pin
- Used in the design of the EIC Electron Storage Ring
- For e^- polarization it “nearly double[s] the asymptotic polarization” for single IP config.
- Details how to introduce beam bumps to minimize radiative depolarization
- Can also be used to obtain bumps that alter other parameters but minimise polarization impact



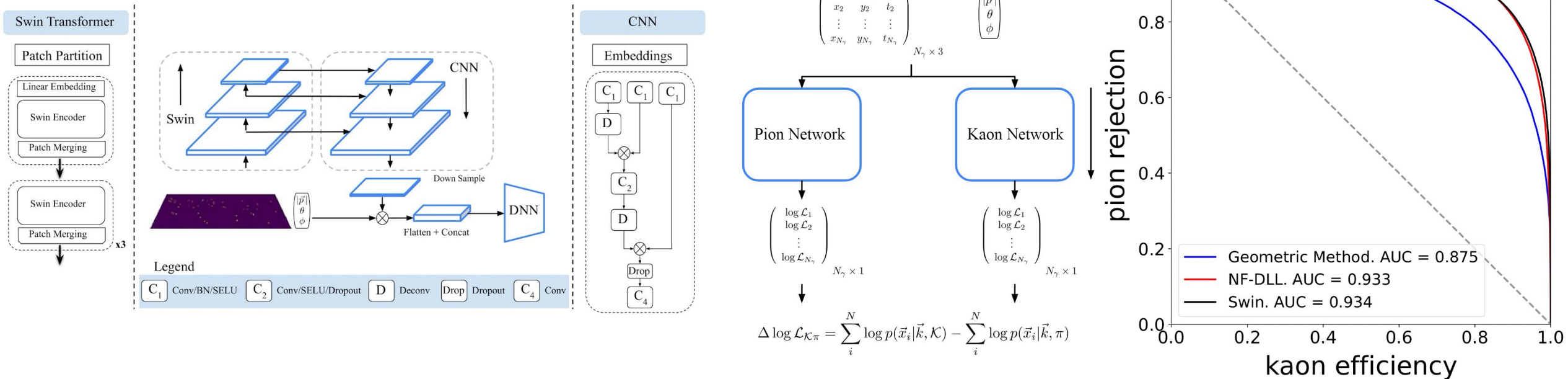
Fast-sim for the PID detectors - C. Fanelli

- Several architectures developed and tested for fast simulation of photons:
 - Normalizing Flows, Continuous Normalizing Flows, Conditional Flow Matching, Denoising Diffusion Probabilistic Models, Score Based Generative Models
- Demonstrate similar hit patterns but much faster generations



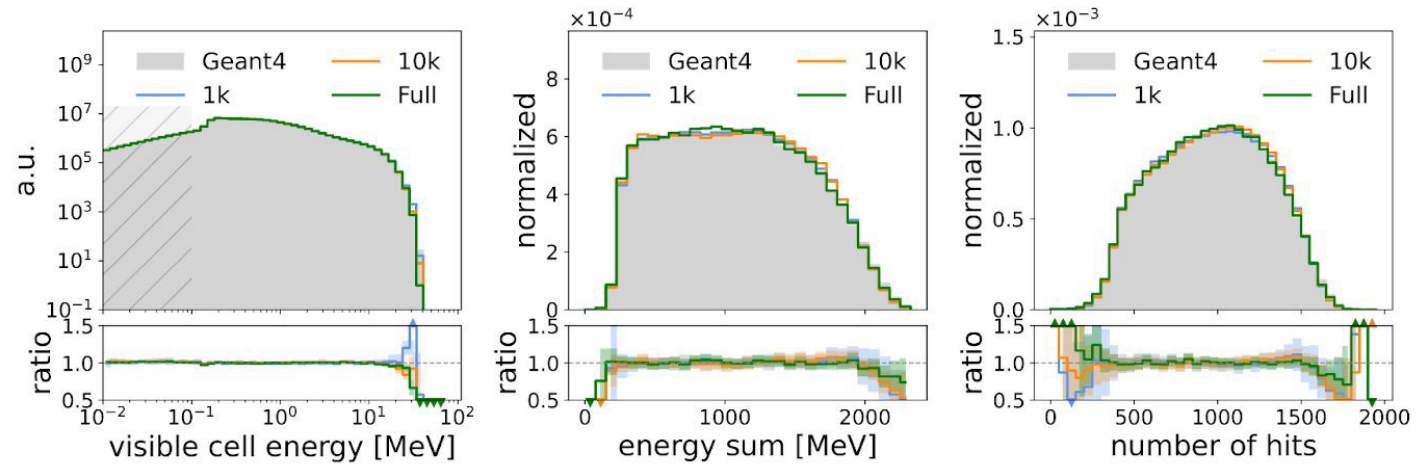
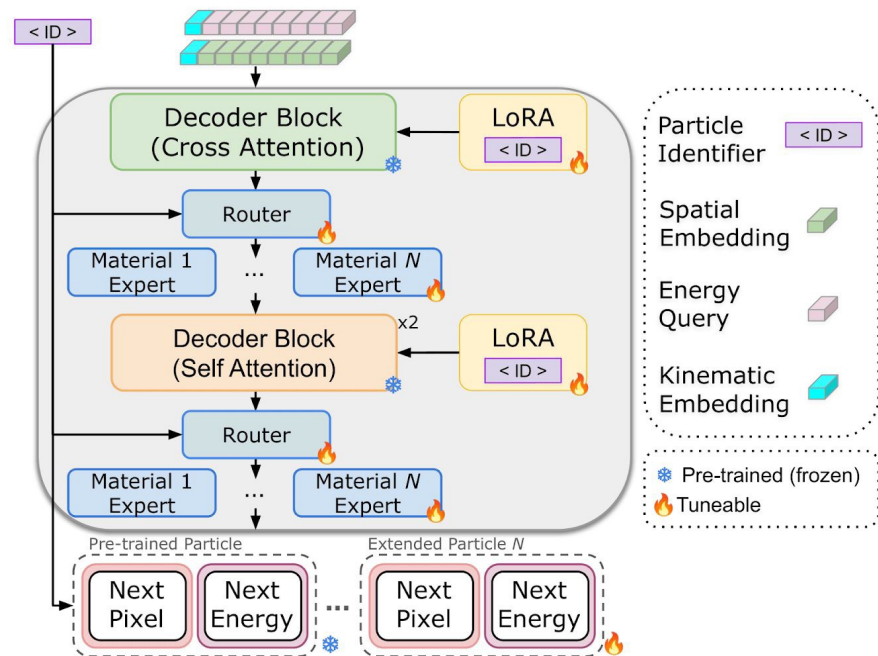
PID from fast sims- C. Fanelli

- Fast simulations can be used to train for PID by measuring log-likelihood for differed hypothesis



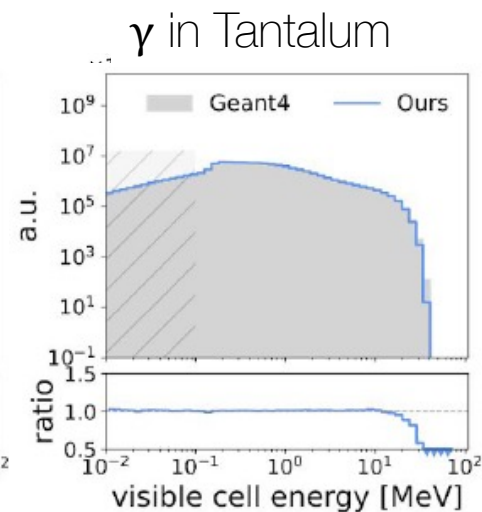
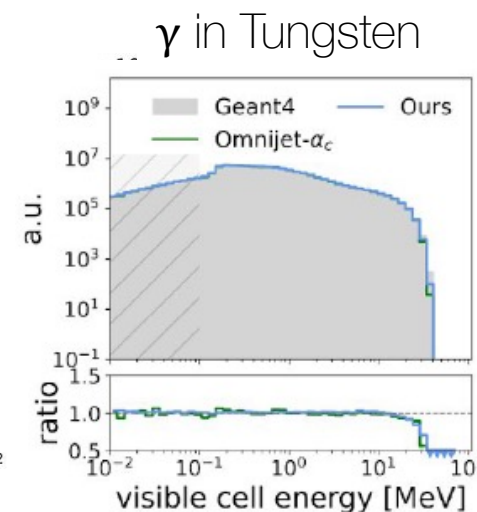
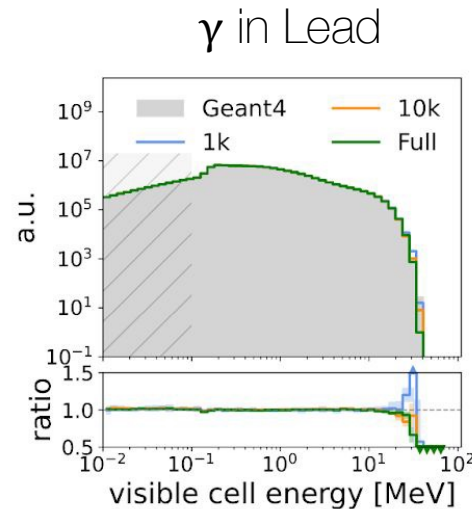
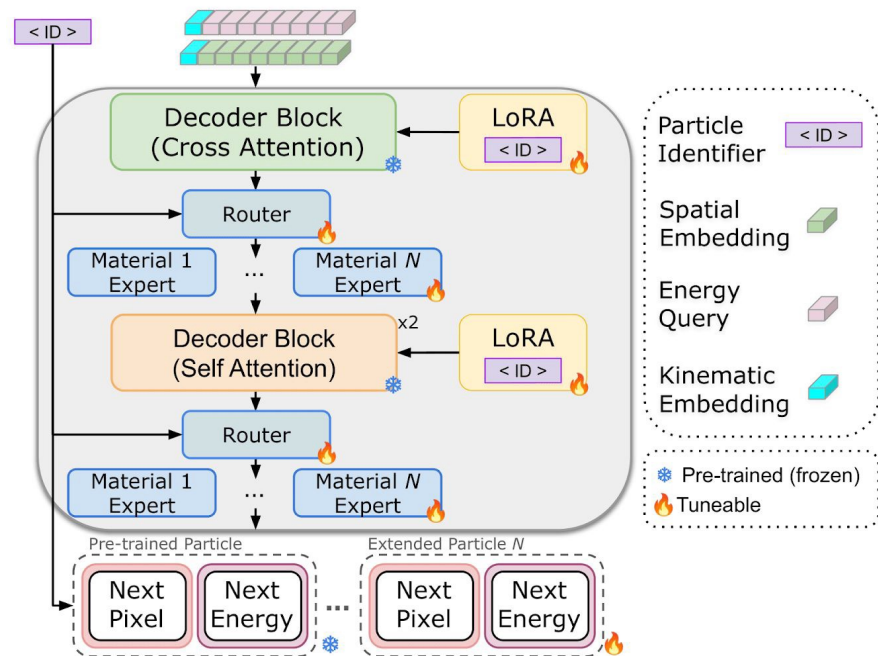
Foundation models - C. Fanelli

- Foundation Models are “self-supervised” models trained on large datasets
 - They can work on multiple tasks (instead of single tasks such as background rejection)
 - Aim is to apply them to different input data sources (different detectors, accelerators etc)
- Altering simulated designs often requires millions of simulated events to study response
- Generalizable FM for calorimeter simulations can converge in $\sim 10k$ generations



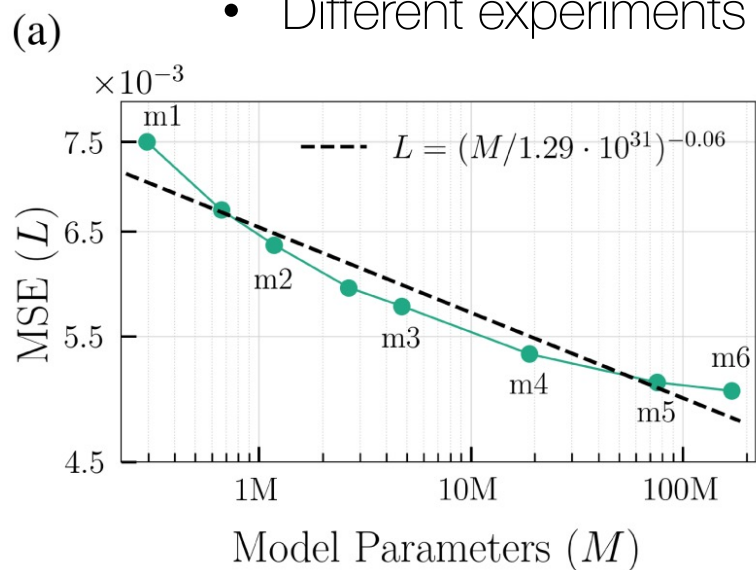
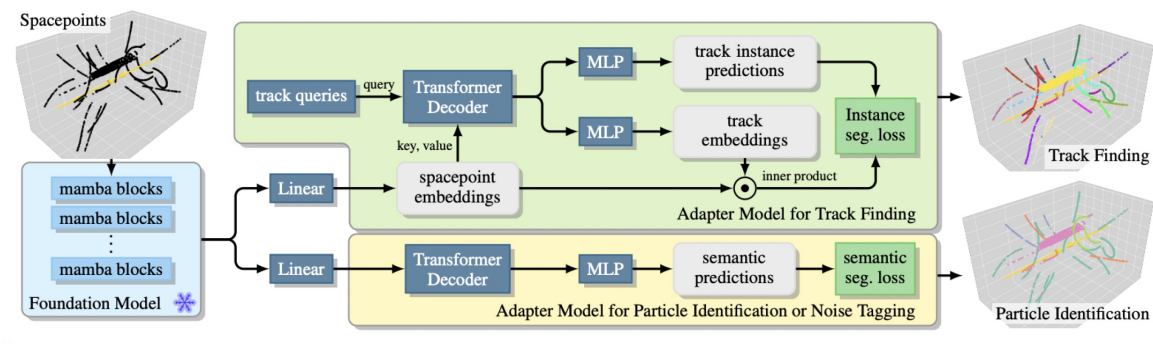
Foundation models - C. Fanelli

- Foundation Models are “self-supervised” models trained on large datasets
 - They can work on multiple tasks (instead of single tasks such as background rejection)
 - Aim is to apply them to different input data sources (different detectors, accelerators etc)
- Altering simulated designs often requires millions of simulated events to study response
- Generalizable FM for calorimeter simulations can converge in $\sim 10k$ generations



Foundation models for TPC data – D. Park

- Tasks
 - Track finding
 - Particle Identification
 - Noise Finding
- Extensions
 - Additional subsystems
 - Different experiments

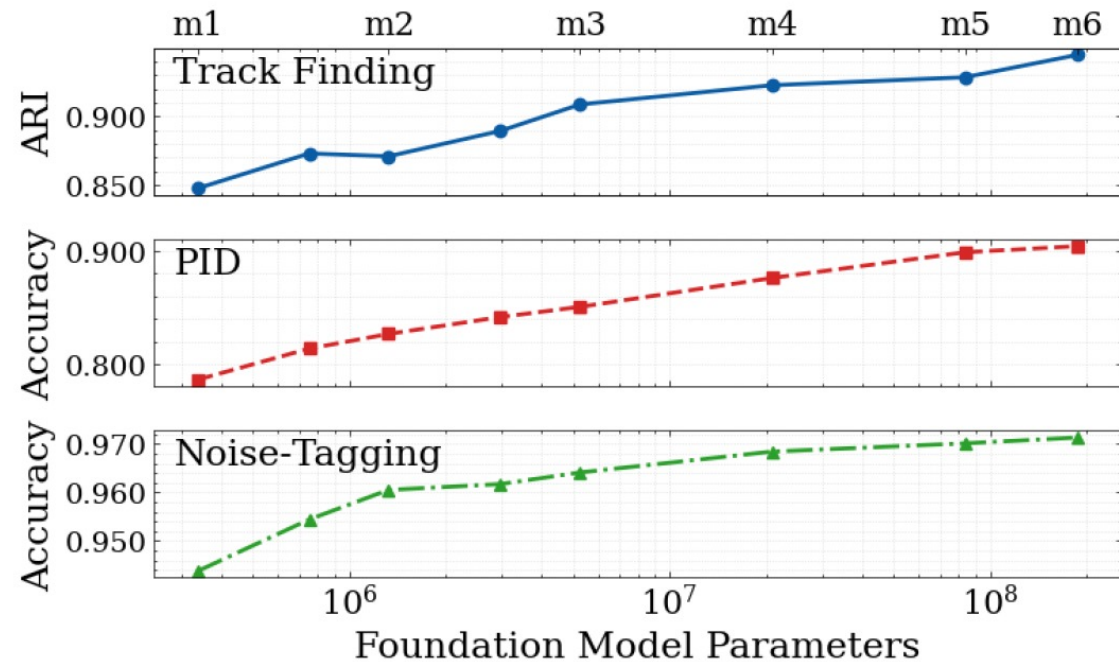
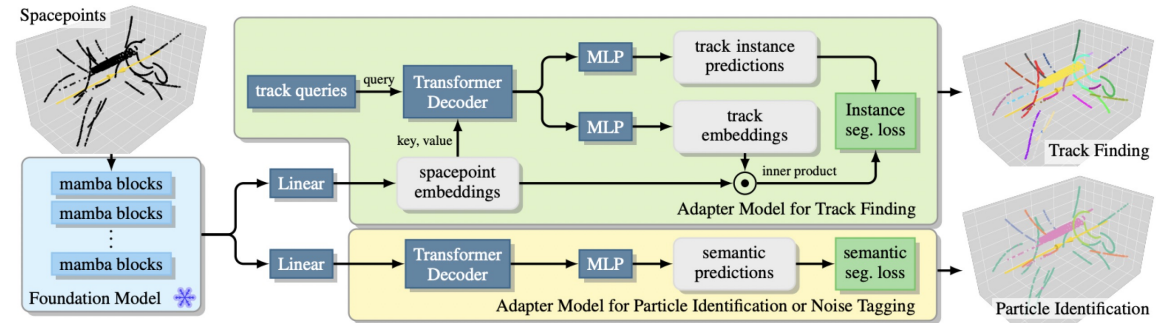
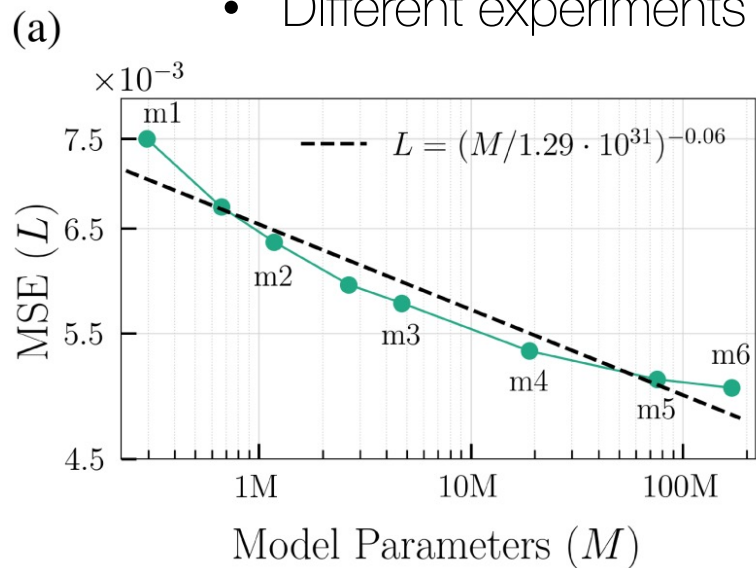


model	#trnbl para.	Track Finding		
		ARI↑	efficiency↑	purity↑
EggNet	0.16M	0.7256	74.19%	75.14%
Exa.TrkX	3.86M	0.8765	91.79%	66.42%
AdapterOnly	2.39M	0.7243	78.01%	64.54%
FM4NPP (m6)	2.39M	0.9448	96.08%	93.08%

model	#trnbl para.	Particle Identification			Noise Tagging		
		acc.↑	recall↑	pre.↑	acc.↑	recall↑	pre.↑
SAGEConv	0.91M	0.7262	0.4563	0.6502	0.9174	0.7227	0.8165
OneFormer3D	44.95M	0.7701	0.4897	0.5767	0.9646	0.9404	0.8948
AdapterOnly	0.74M	0.6631	0.3387	0.6111	0.9111	0.6215	0.8359
FM4NPP (m6)	0.74M	0.9039	0.7652	0.8782	0.9713	0.9367	0.9190

Foundation models for TPC data – D. Park

- Tasks
 - Track finding
 - Particle Identification
 - Noise Finding
- Extensions
 - Additional subsystems
 - Different experiments

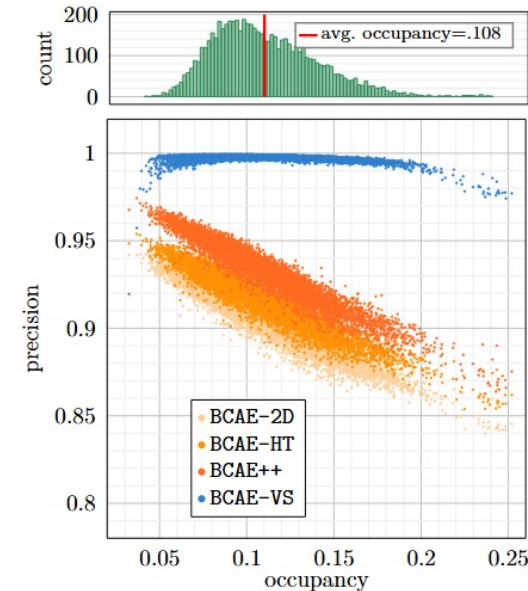
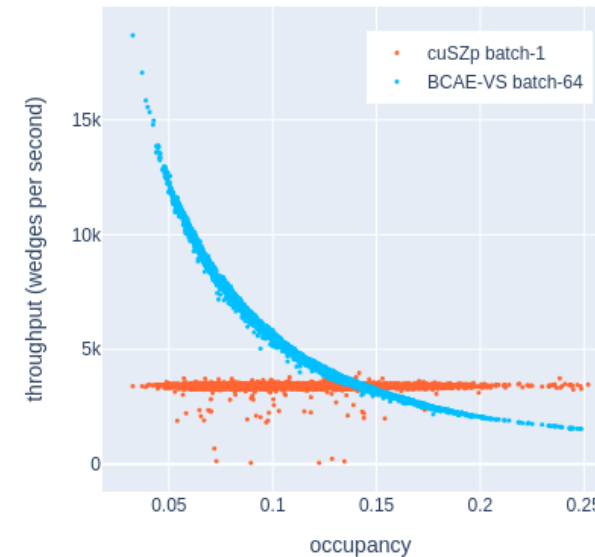


Compression for TPC data – Y. Huang

- TPC data contributed the most to the sPHENIX raw data (~90%)
 - Efficient compression has direct impact on storage, buffer requirements and more
- Clusters can be saved based on weight (importance)
 - Weight based on Mean Squared Error and nearest-neighbors
 - Cut on threshold weight
- Work performed on pp and AuAu data (plots from AuAu)
- BCAE-VS is an autoencoder with variable ratio compression for sparse input

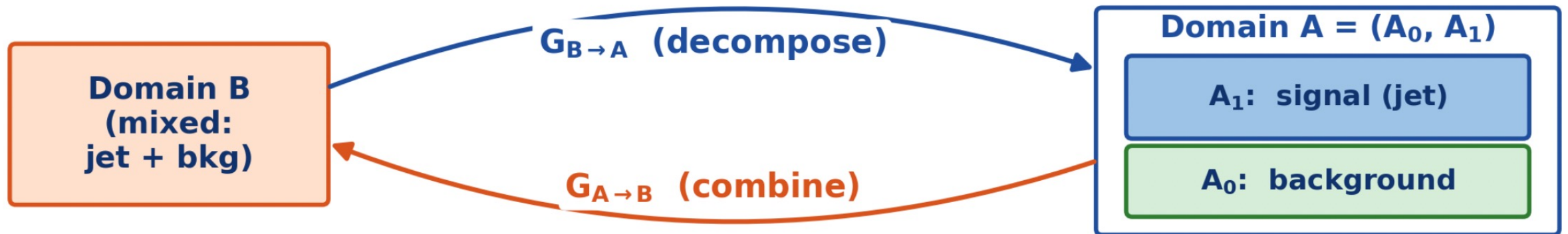
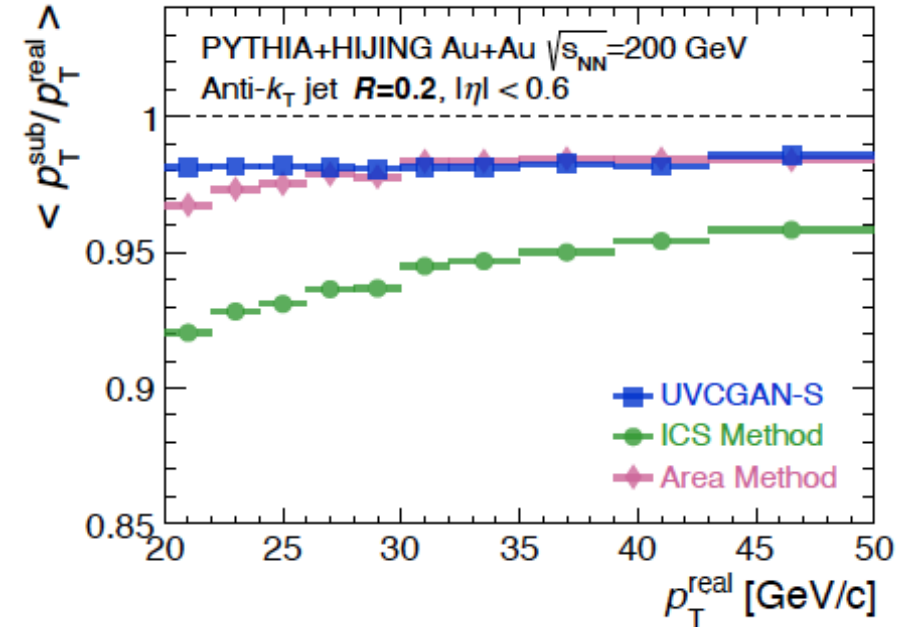
Table 1: Comparing compression algorithm with dense and sparse convolutions.

model	comp. ratio \uparrow	reconstruction performance					efficiency	
		$L_1 \downarrow$	$L_2 \downarrow$	PSNR \uparrow	recall \uparrow	precision \uparrow	encoder size	throughput \uparrow
BCAE-2D	31.1	.152	.862	20.6	.907	.906	169k	9.6k
BCAE-HT (3D)	31.1	.138	.781	20.8	.916	.915	9.8k	9.6k
BCAE++ (3D)	31.1	.112	.617	21.4	.936	.934	226k	3.2k
BCAE-VS	34.3	.030	.101	25.6	.986	.997	382	5.6k



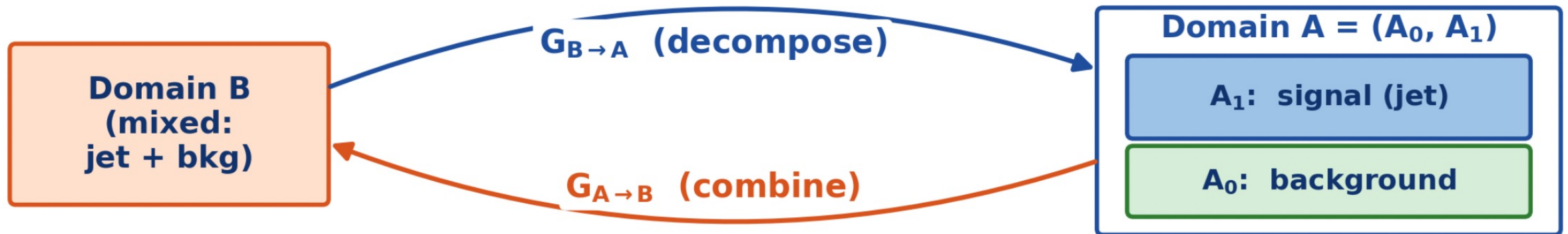
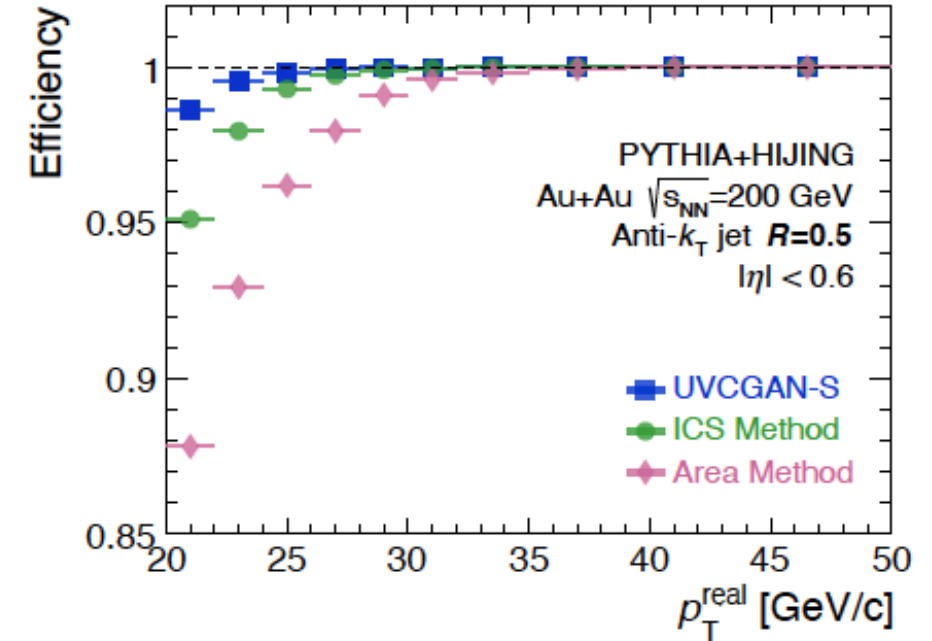
Background subtraction in jets – S. Li

- Generative Adversarial Networks consist of two competing models
 - Discriminator tries to separate classes (signal and background)
 - Generator, G , tries to trick the first model with generated samples
 - Discriminator gets better at spotting fakes
- ICS = Iterative Constituent Subtraction
- “sub” is background subtracted, “real” is truth



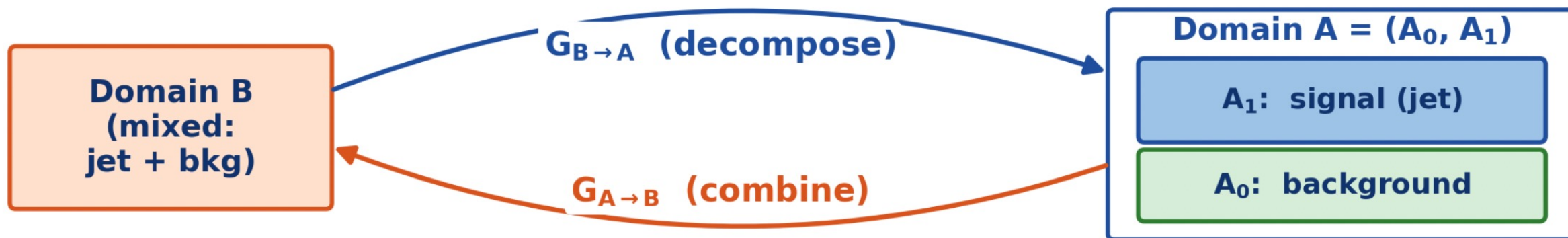
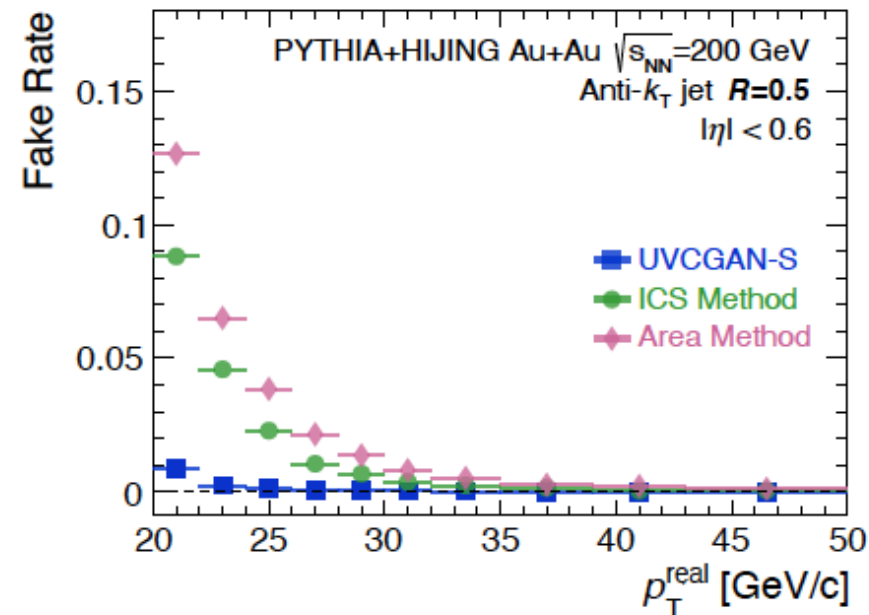
Background subtraction in jets – S. Li

- Generative Adversarial Networks consist of two competing models
 - Discriminator tries to separate classes (signal and background)
 - Generator, G , tries to trick the first model with generated samples
 - Discriminator gets better at spotting fakes
- ICS = Iterative Constituent Subtraction
- “sub” is background subtracted, “real” is truth



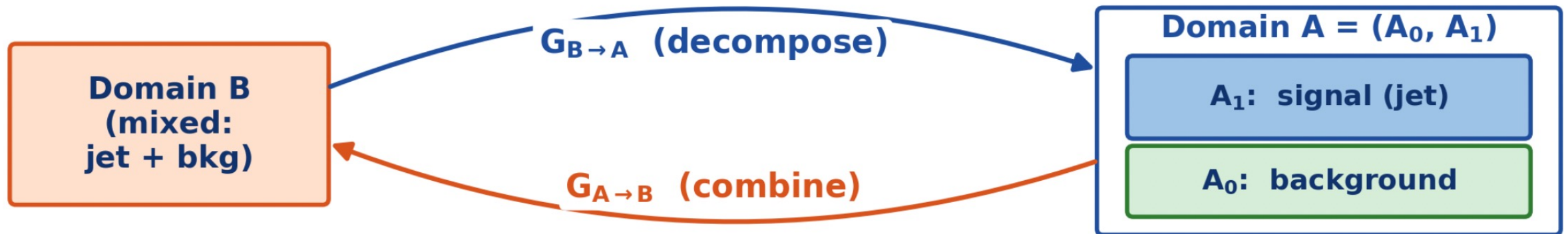
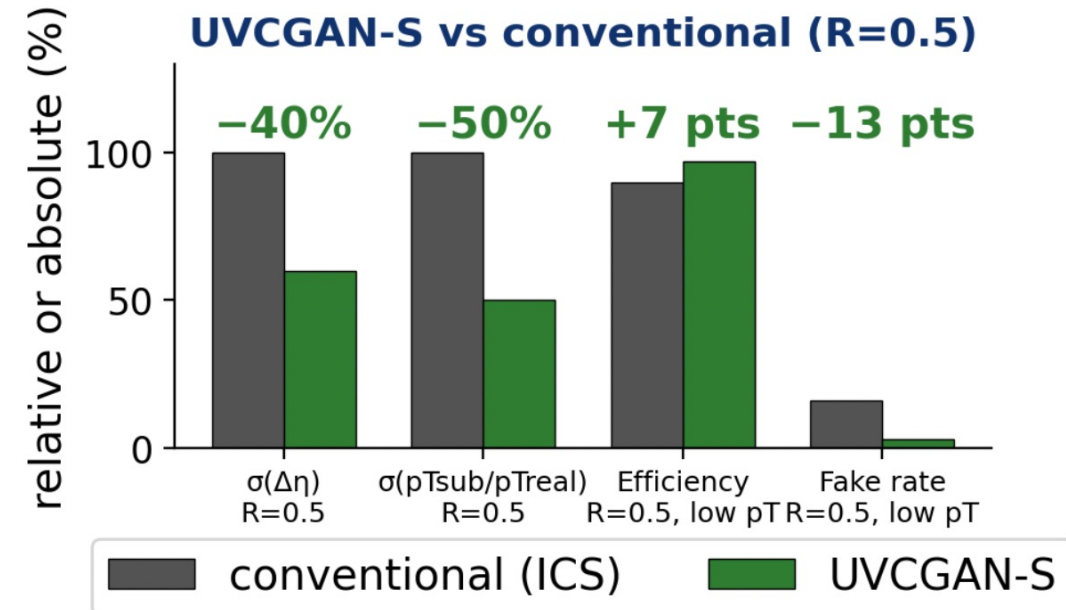
Background subtraction in jets – S. Li

- Generative Adversarial Networks consist of two competing models
 - Discriminator tries to separate classes (signal and background)
 - Generator, G , tries to trick the first model with generated samples
 - Discriminator gets better at spotting fakes
- ICS = Iterative Constituent Subtraction
- “sub” is background subtracted, “real” is truth

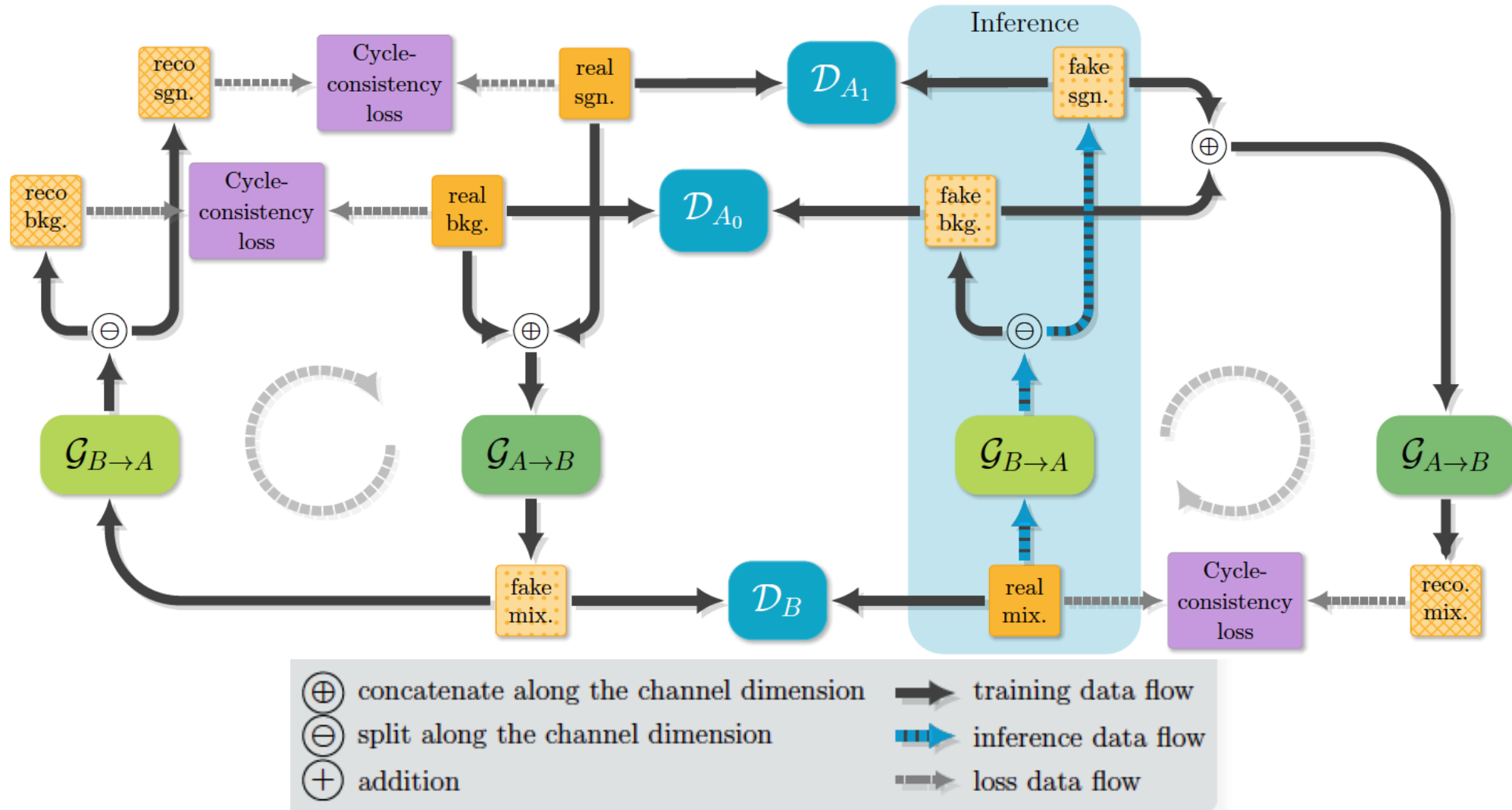


Background subtraction in jets – S. Li

- Generative Adversarial Networks consist of two competing models
 - Discriminator tries to separate classes (signal and background)
 - Generator, G , tries to trick the first model with generated samples
 - Discriminator gets better at spotting fakes
- ICS = Iterative Constituent Subtraction
- “sub” is background subtracted, “real” is truth

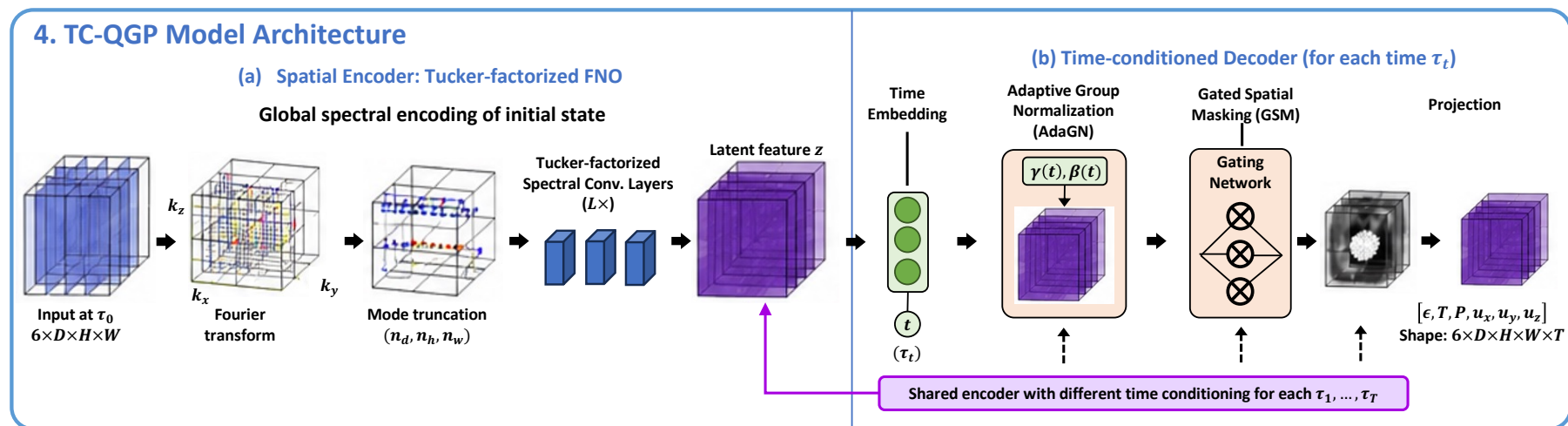


Background subtraction in jets – S. Li



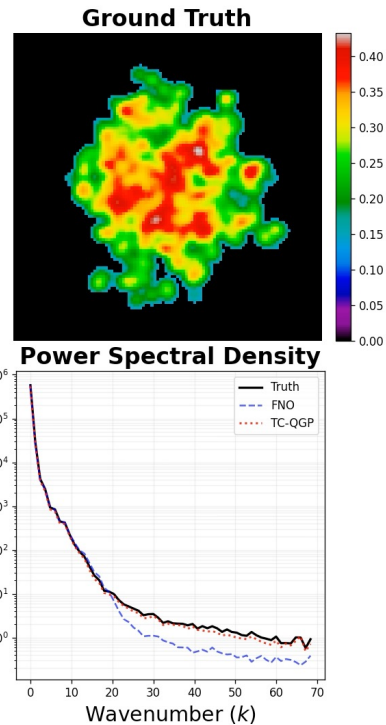
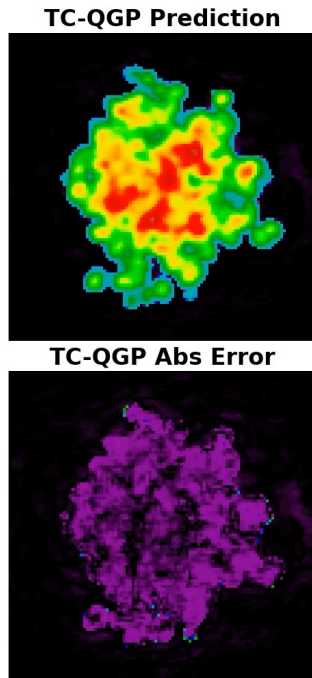
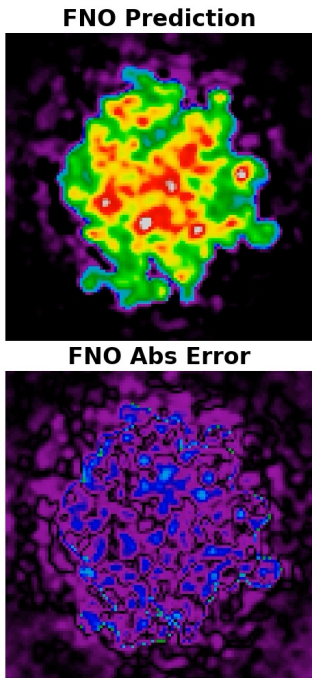
Fast simulations of QGP – S. Lee

- Major bottleneck for heavy ion simulations is QGP hydrodynamic evolution
 - Time and memory intensive
- Typically performed using Fourier Neural Operators
 - Uses Fast Fourier Transform to solve Partial Differential Equations (think time-domain to frequency domain)
- Loss function defined as: $\mathcal{L} = \frac{1}{T} \sum_{t=1}^T [\mathcal{L}_{L^2}(\hat{f}_t, f_t) + \lambda \mathcal{L}_{L^2}(\nabla \hat{f}_t, \nabla f_t)]$
 - First term penalizes pointwise relations and the second penalizes gradient differences



Fast simulations of QGP – S. Lee

$\tau = 0.60 \text{ fm}/c$



Temperature

- Accuracy – Mean Relative Error (%)

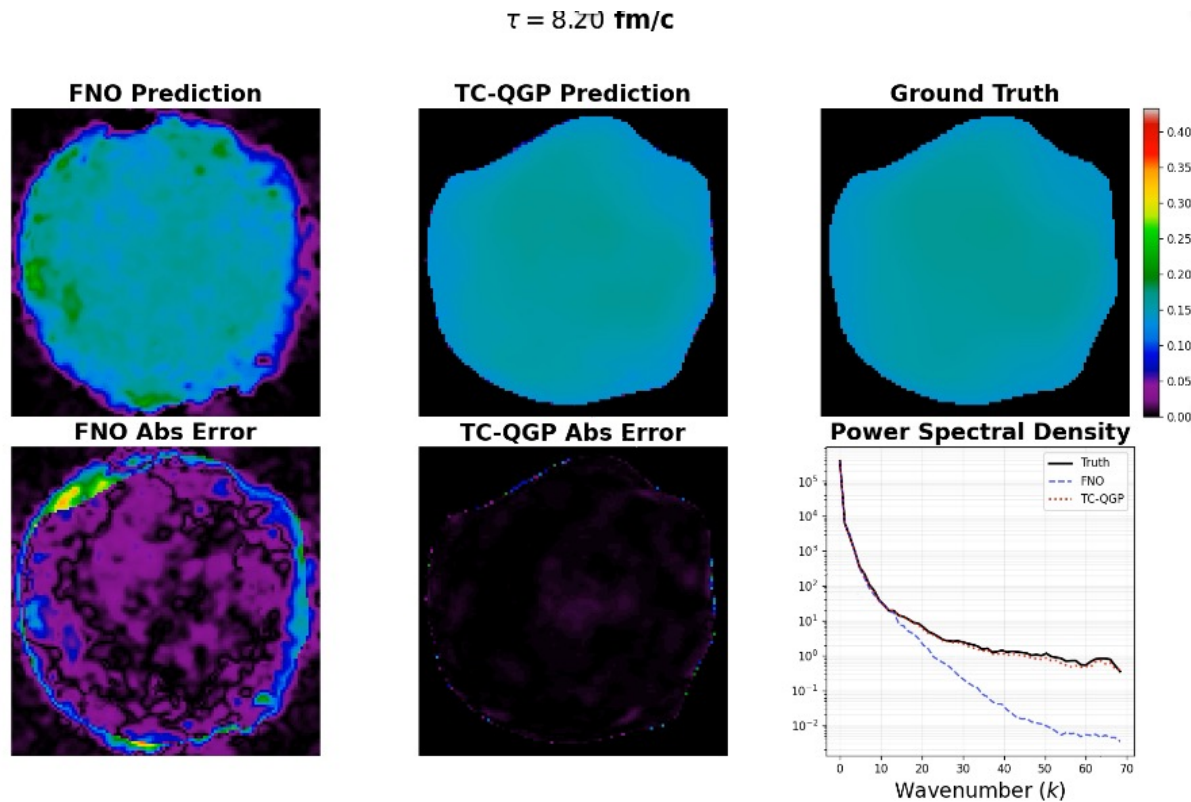
	FNO	TC-QGP	↓
ϵ	17.44	4.08	4.3×
T	13.89	1.66	8.4×
P	20.32	4.54	4.5×
u_x	28.10	8.52	3.3×
u_y	27.73	7.80	3.6×
u_z	31.44	5.44	5.8×
All	23.15	5.34	4.3×

- Cost (single A100, batch 1)

	FNO	TC-QGP
Parameters	382.4 M	356.6 M
GPU Memory	16,585 MB	4,459 MB
Inference Time	108.8 ms	230.4 ms

vs MUSIC: ~20 hours per event → TC-QGP: ~0.2 s per event

Fast simulations of QGP – S. Lee



Temperature

- Accuracy – Mean Relative Error (%)

	FNO	TC-QGP	↓
ε	17.44	4.08	4.3×
T	13.89	1.66	8.4×
P	20.32	4.54	4.5×
u_x	28.10	8.52	3.3×
u_y	27.73	7.80	3.6×
u_z	31.44	5.44	5.8×
All	23.15	5.34	4.3×

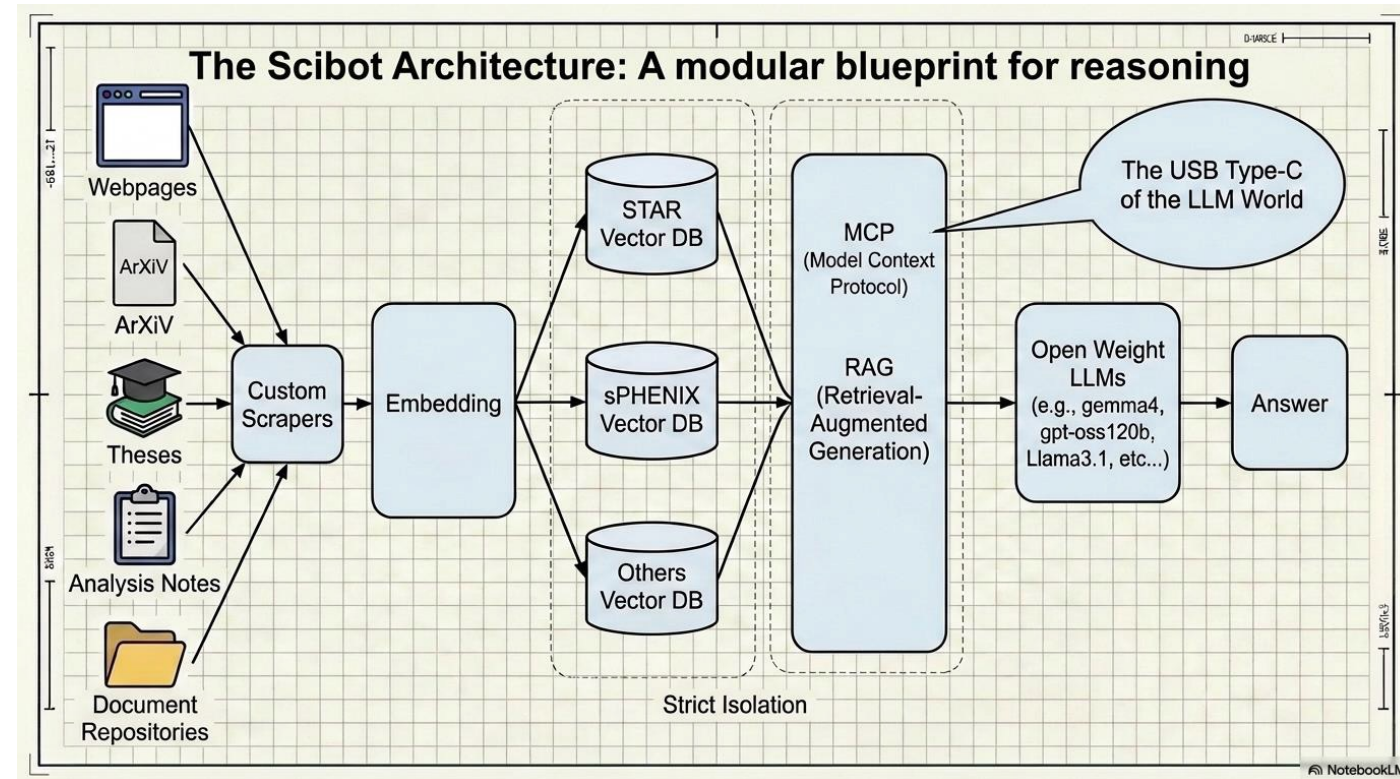
- Cost (single A100, batch 1)

	FNO	TC-QGP
Parameters	382.4 M	356.6 M
GPU Memory	16,585 MB	4,459 MB
Inference Time	108.8 ms	230.4 ms

vs MUSIC: ~20 hours per event → TC-QGP: ~0.2 s per event

Chat bots for BNL data – A. Kanuganti

- Chat bots are becoming increasingly popular (think ChatGPT)
- How can scientists take advantage?
 - Some data is restricted to collaborators
- Lots of research time can be spent in literature review
- What if chat bots can parse theses, analysis notes, presentations and news releases



Chat bots for BNL data – A. Kanuganti

ChatGPT >



How can I analyze jets in 200 GeV Au+Au collisions from 2014 in a RHIC STAR analysis. Explain in 250 words, cite each claim, restrict to RHIC domain only and keep the response grounded with zero fabrication.

A typical jet analysis in 200 GeV Au+Au collisions from the 2014 run at the [STAR Collaboration](#) follows a well-established workflow grounded in RHIC-specific reconstruction and heavy-ion techniques.

First, events are selected using minimum-bias or high-tower triggers, with vertex and centrality cuts derived from charged-particle multiplicity in the [Time Projection Chamber](#) and forward detectors (standard STAR event selection procedures). Charged tracks (TPC) and neutral energy ([Barrel Electromagnetic Calorimeter](#)) are combined to form particle candidates, ensuring full jet reconstruction within STAR acceptance ($|\eta| \leq 1$).

Jets are reconstructed using sequential recombination algorithms such as anti- k_r via [FastJet](#), typically with resolution parameter $R = 0.2$ – 0.4 . This choice balances capture of jet energy against large underlying-event background in Au+Au (standard in RHIC jet analyses). The dominant complication is the heavy-ion background: a large, fluctuating soft particle density. This is corrected using area-based subtraction ($p \times A$), where p is the median background density estimated event-by-event (as established in RHIC heavy-ion jet measurements).

Detector effects and residual background fluctuations are handled through embedding: simulated p+p jets are embedded into real Au+Au events and reconstructed, allowing response matrices to be built. These matrices are then unfolded (e.g., iterative Bayesian methods) to obtain corrected jet spectra (standard STAR unfolding workflow).

Chat bots for BNL data – A. Kanuganti

STAR Chatbot



Choose a model

gpt-oss-120b

Chat



How can I analyze jets in 200 GeV Au+Au collisions from 2014 in a RHIC STAR analysis



Steps to analyze 200 GeV Au+Au jets from the 2014 STAR run

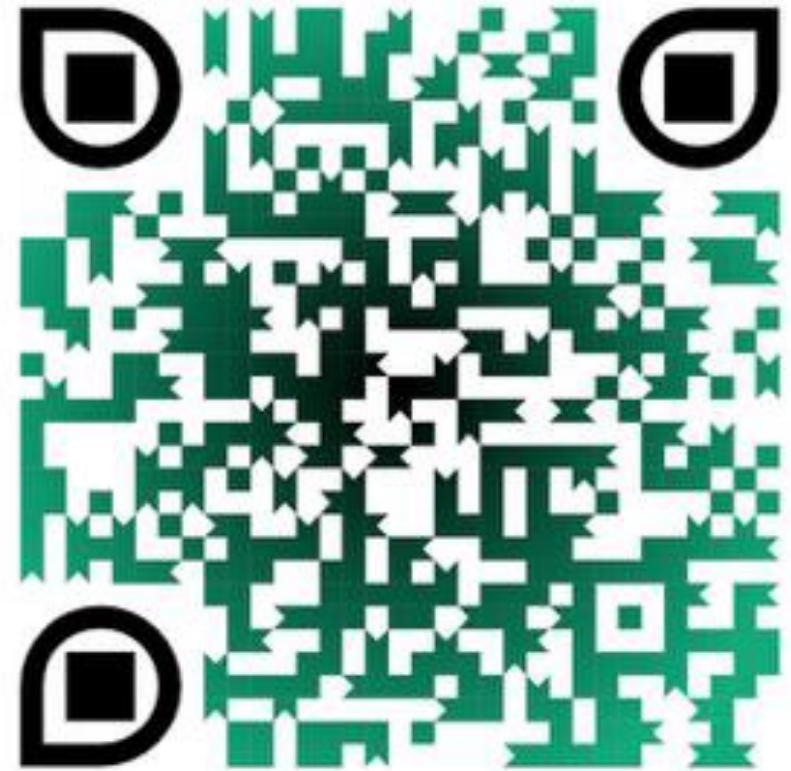
- **Data set & event selection**
 - Use the Au+Au $\sqrt{s_{nn}}=200$ GeV data recorded in 2014 by STAR [4].
 - Select events with a high-energy Barrel Electromagnetic Calorimeter (BEMC) tower ($E_T > 4.3$ GeV) to define the *signal* jet sample; employ minimum-bias events for background estimation [4].
 - Apply the standard STAR vertex cut $|vz| < 30$ cm for inclusive jet analyses ($|vz| < 70$ cm can be used for γ_{dir} + jet studies) [7].
- **Detector inputs**
 - Charged tracks from the Time Projection Chamber (TPC) (full azimuth, $|\eta| < 1$).
 - Neutral energy from BEMC towers, with double-counting corrections for electrons and hadronic deposits [6].
- **Jet reconstruction**
 - Run the anti- k_T algorithm (FastJet) with radius parameter $R=0.4$ (or $R=0.3$ for tighter jets) [4, 5].
 - Include both charged tracks ($p_T \geq 200$ MeV/c) and calibrated BEMC towers in the clustering [5].
- **Background subtraction**
 - Compute the event-wise median background density ρ and correct each jet's raw transverse momentum: $p_T^{corr} = p_T^{rec} - \rho \cdot A$ (A = jet area) [5].
 - Use a mixed-event technique to model the combinatorial jet background and subtract it statistically from the measured spectra [4].
- **Bias mitigation**
 - To obtain a trigger-bias-free jet sample, study recoil jets opposite a high- p_T hadron or photon trigger; the recoil jet population is largely unbiased [4].

Conclusion

- Variety of presentations
 - Demonstrates ML is being used in multiple aspects of nuclear physics; from experimental designs, data analysis and knowledge preservation
- A clear trend
 - Lab and university personnel are leading innovation and overcoming restrictions from commercial products
- A common point
 - Hardware availability can be a limiting factor
 - Deployment of operational algorithms at EIC will require dedicated computing power
 - Algorithm development is not limited by scientist ingenuity but by available compute

Thank you for your attention

Try the SciBot!!!
scibot.csi.bnl.gov

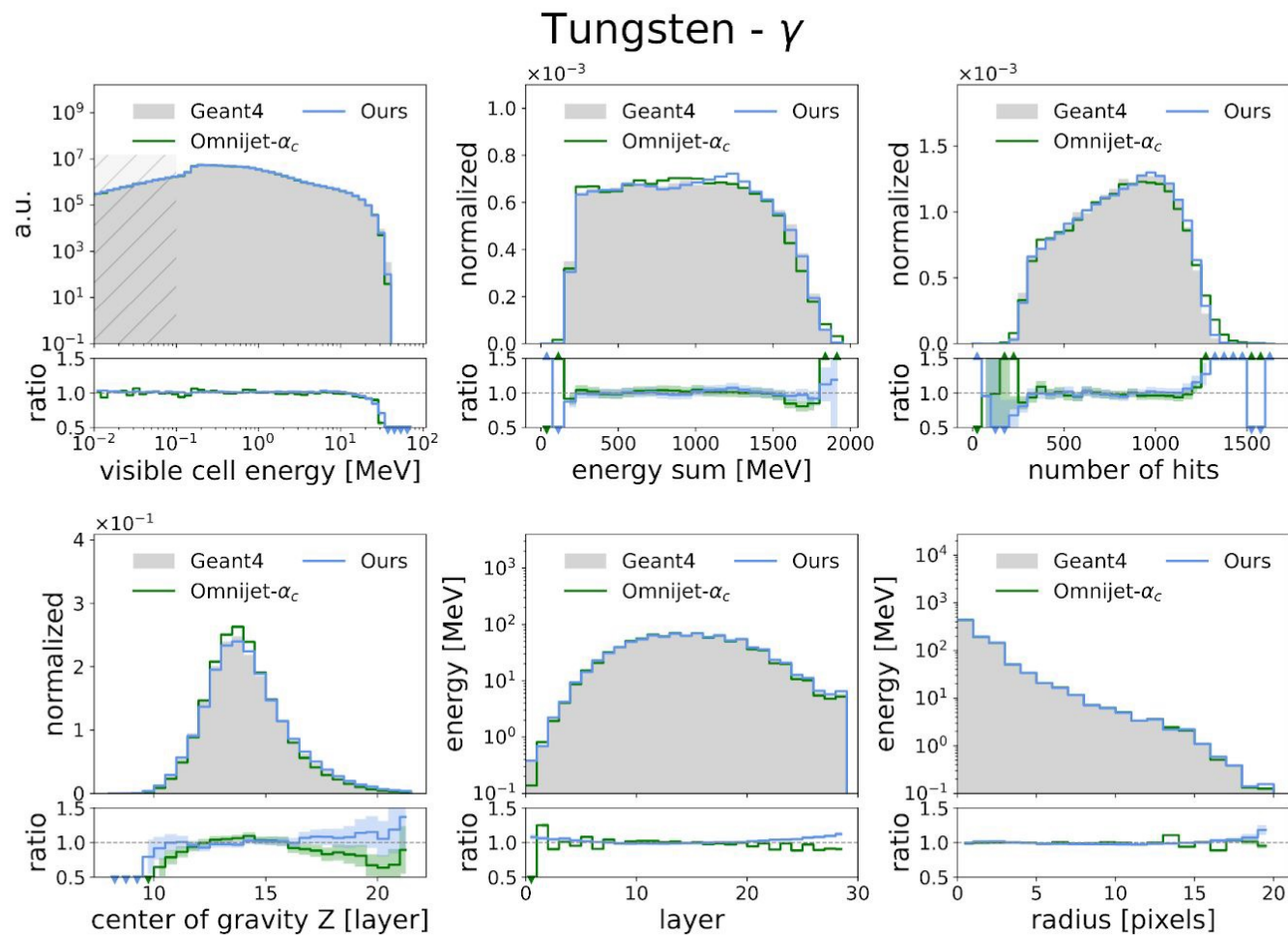


Backup

Foundation models for calorimetry

Baseline model, photons in tungsten

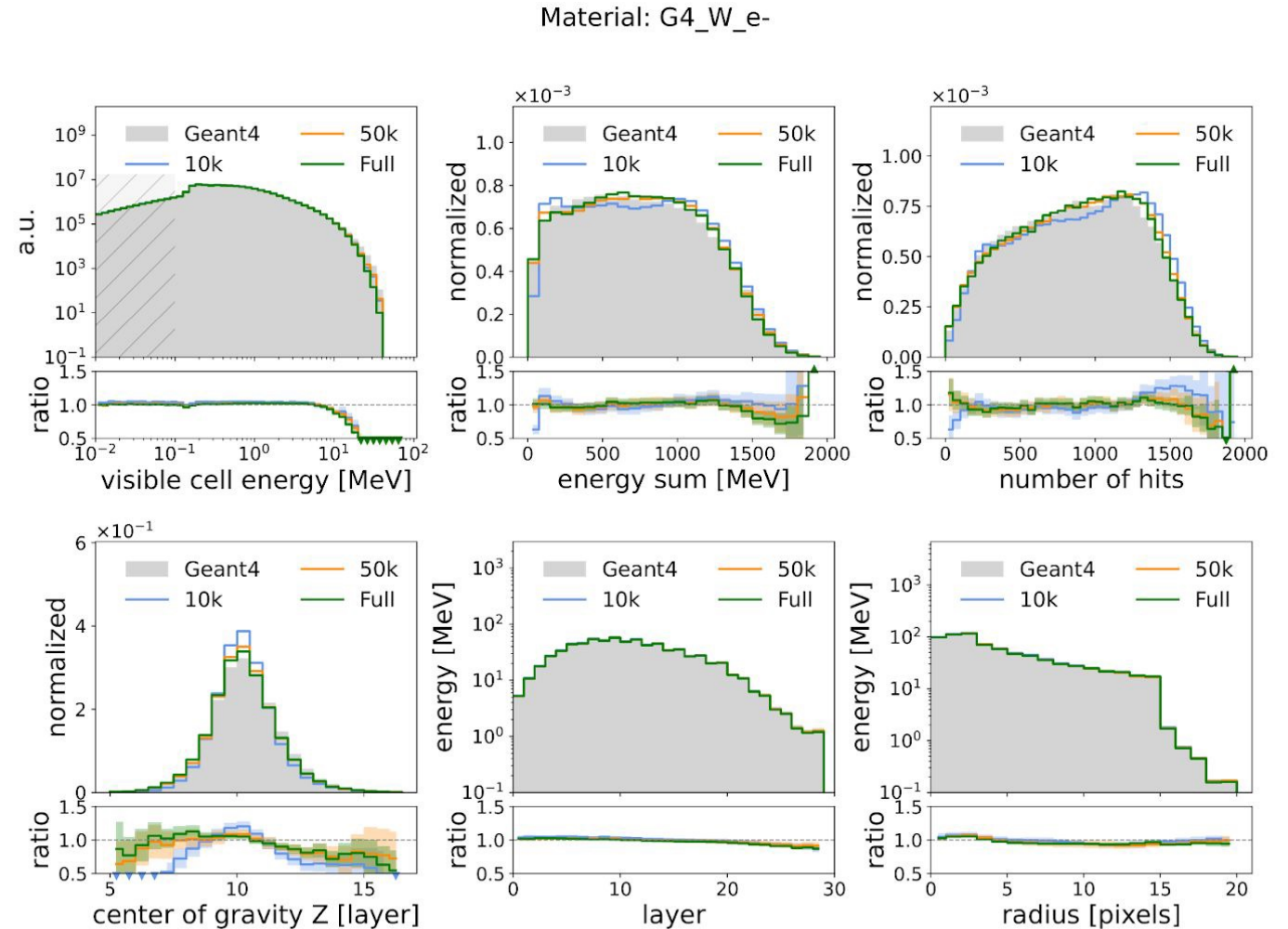
- Clockwise from top left
 - Visible cell energy
 - Total energy sum per shower
 - Hit multiplicity
 - Radial energy profile relative to the shower centroid
 - Average energy deposition per calorimeter layer
 - Longitudinal center of gravity



Foundation models for calorimetry

New model, electrons in tungsten

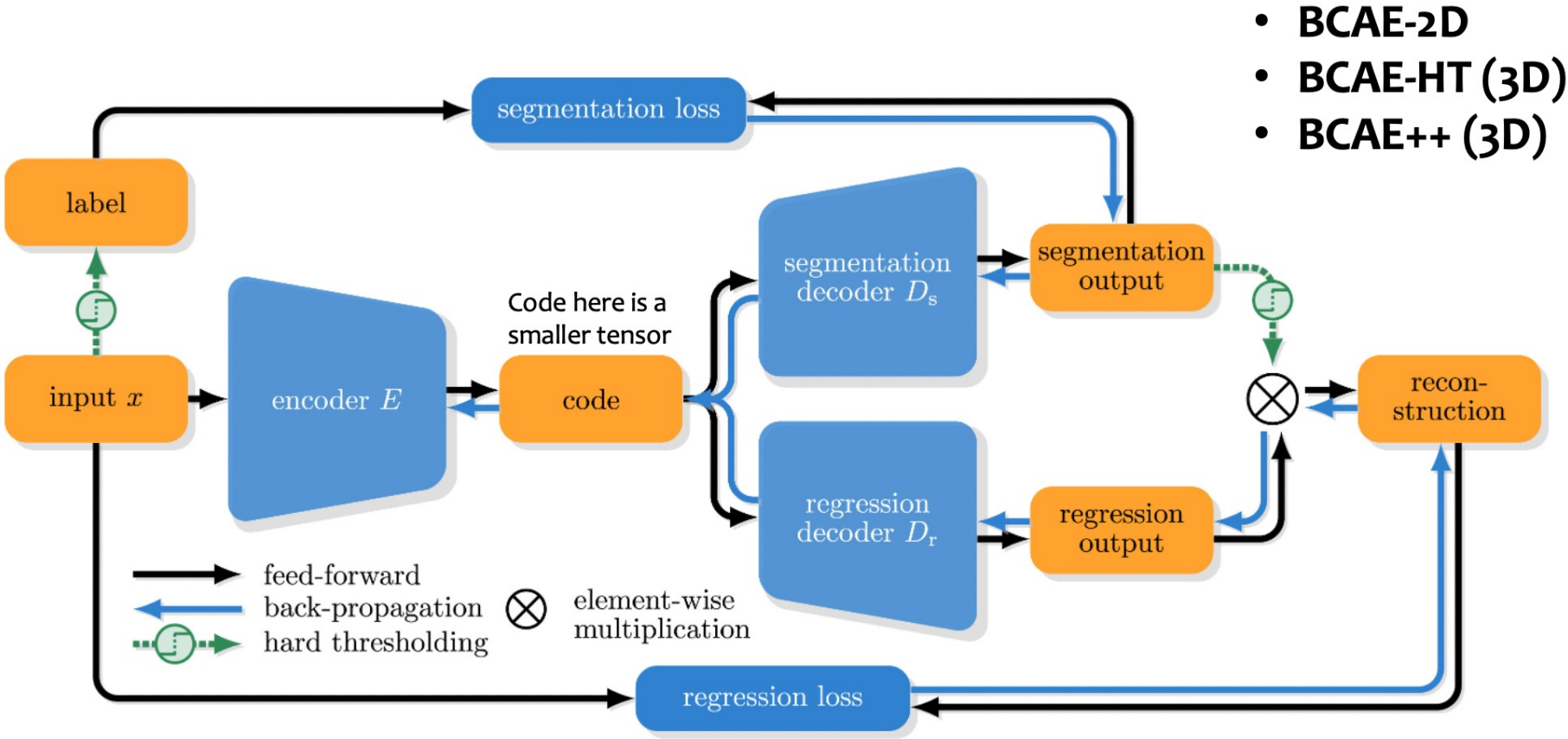
- Clockwise from top left
 - Visible cell energy
 - Total energy sum per shower
 - Hit multiplicity
 - Radial energy profile relative to the shower centroid
 - Average energy deposition per calorimeter layer
 - Longitudinal center of gravity



Compression - Autoencoders

The BCAE Family – Compression by Size

Bicephalous Convolution Autoencoder



Compression - Autoencoders

BCAE-VS – Compression by Importance

Variable ratio Compression for Sparse input

