Foundation Models for Nuclear and Particle Physics (FM4NPP)

Joe Osborn

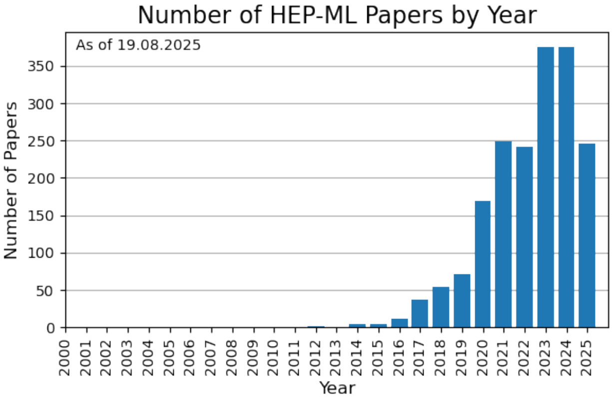
Brookhaven National Laboratory

Based on arXiv:2508.14087

September 16, 2025

ML/AI in HENP

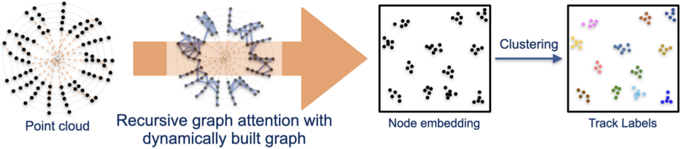
- ML/AI in HENP has seen rapid growth in the last 10 years
- Driven by a number of factors:
    - Available compute resources
    - Available dataset size
    - Developments in industry
    - Continued collaboration between HEP/NP/CS/Data-science



HepML-LivingReview

AI Models in HEP Tracking

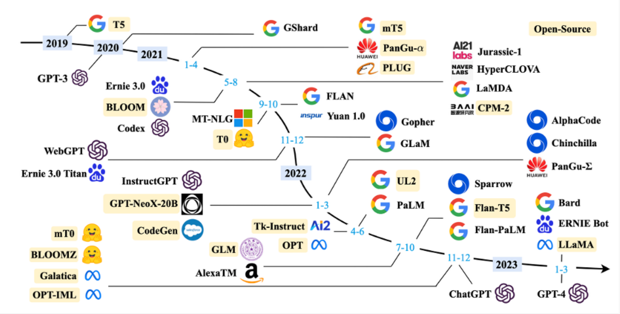
- HL-LHC and high rate nuclear physics experiments such as ALICE, sPHENIX, and ePIC have motivated R&D into new track reconstruction algorithms
- Classical algorithms, such as the Kalman Filter, are computationally expensive and difficult to parallelize
- GNNs are well suited for sparse data but face scalability difficulties



2407.13925

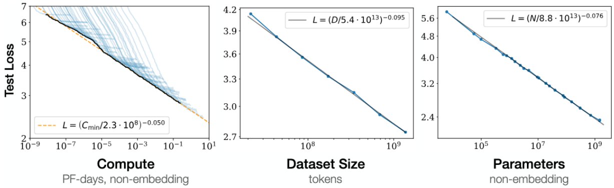
Scaling Large Language Models

- LLMs have received a lot of attention in the last decade
- Self-supervised auto-regressive pre-training (not reliant on labeled data)
- Pre-trained model can be extended for multiple downstream tasks
- (2020) Neural Scaling behavior demonstrated (2001.08361)
- (2020-2023) LLM "arms race"
- (2023) Scaling behavior holds for GPT-4 (2303.08774)



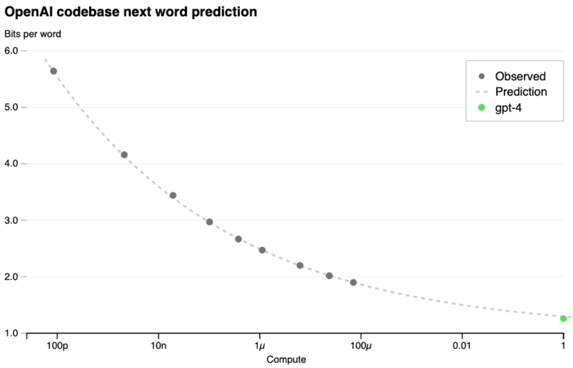
Scaling Large Language Models

- LLMs have received a lot of attention in the last decade
- Self-supervised auto-regressive pre-training (not reliant on labeled data)
- Pre-trained model can be extended for multiple downstream tasks
- (2020) Neural Scaling behavior demonstrated (2001.08361)
- (2020-2023) LLM "arms race"
- (2023) Scaling behavior holds for GPT-4 (2303.08774)



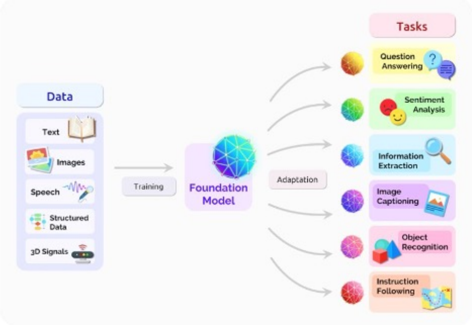
Scaling Large Language Models

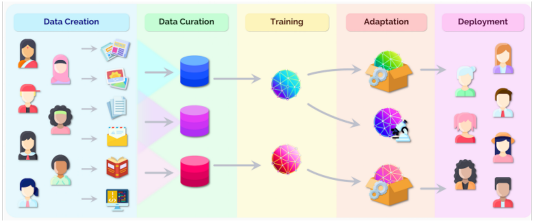
- LLMs have received a lot of attention in the last decade
- Self-supervised auto-regressive pre-training (not reliant on labeled data)
- Pre-trained model can be extended for multiple downstream tasks
- (2020) Neural Scaling behavior demonstrated (2001.08361)
- (2020-2023) LLM "arms race"
- (2023) Scaling behavior holds for GPT-4 (2303.08774)



Foundation Models

- Foundation models (FMs) are envisioned as a counterpart to text-based LLMs, but can handle multiple types of data
- Large scale, primarily unlabeled data
- Handle multiple modalities
- Trained via self-supervised learning
- Adaptable to diverse downstream applications
- Neural scaling behavior





Scientific FMs

- Many scientific domains are starting to explore application of FMs for their field
  - e.g. materials science, protein folding, bioinformatics. . .
- Perfect opportunity for high energy nuclear and particle physics
  1. Large amount of unlabeled data
  2. Many possible downstream reconstruction and analysis related tasks
  3. Opportunity for self-supervised learning
- Can we build a FM for NPP?

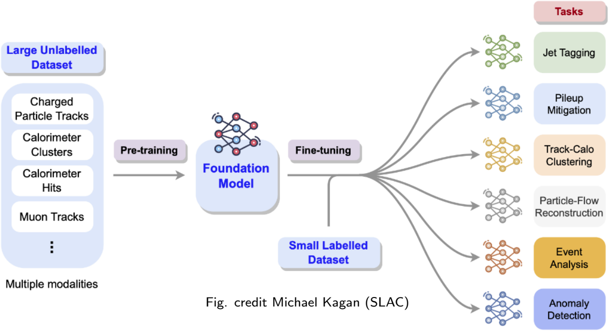
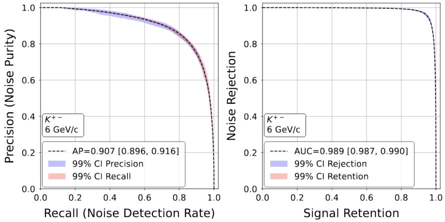


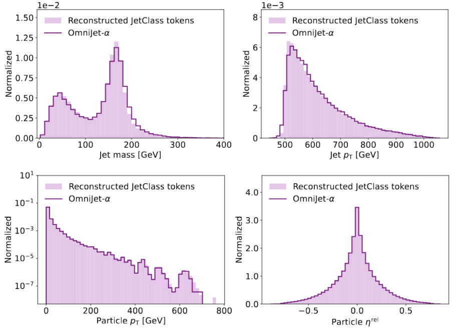
Fig. credit Michael Kagan (SLAC)

Foundation Model Development in HENP

- There has been a flurry of work in the last year studying (mostly) higher level objects
  - Examples : implications of FMs for physics (2501.05382) , jets (2412.10504, 2404.16091, 2403.05618), DIRC (2505.08736), and more
  - What about lower level reconstruction, which faces similar challenges between HEP and NP?
  - Can we demonstrate that a FM behaves as we expect it to?



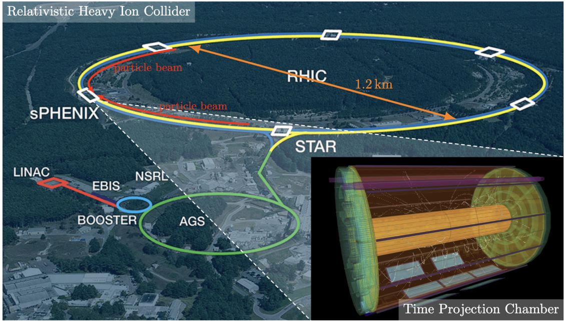
2505.08736



2403.05618

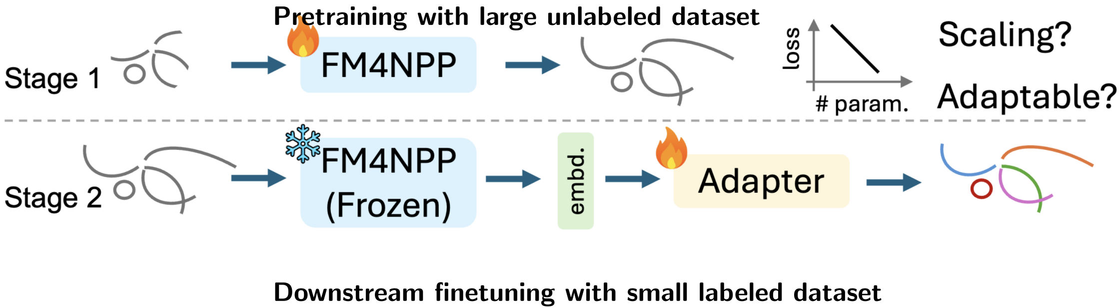
Physics Motivation

- Can we apply a FM to high energy nuclear/particle physics data?
  - Will the FM pre-training exhibit scaling behavior?
  - Can the FM learn additional downstream physics related tasks? What are the right tasks?
  - Does a larger model lead to improved physics performance?
  - ...
- Initial proof of concept for FM4NPP:
  1. Neural scaling behavior
  2. Generalizable to downstream tasks



FM4NPP Goal

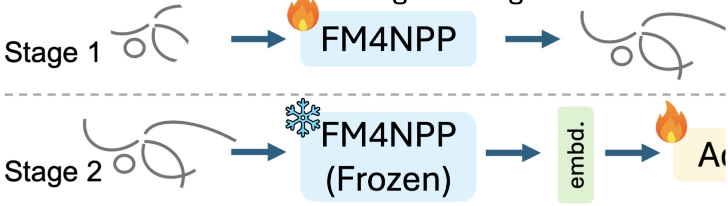
1. Neural scaling behavior (characteristic of all FMs)
2. Generalizable to downstream tasks

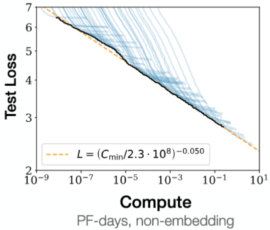


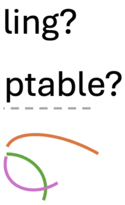
FM4NPP Goal

1. Neural scaling behavior (characteristic of all FMs)
2. Generalizable to downstream tasks

Pretraining with large unlabeled data

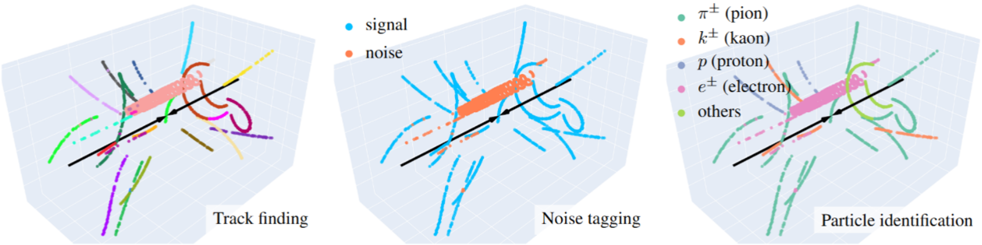






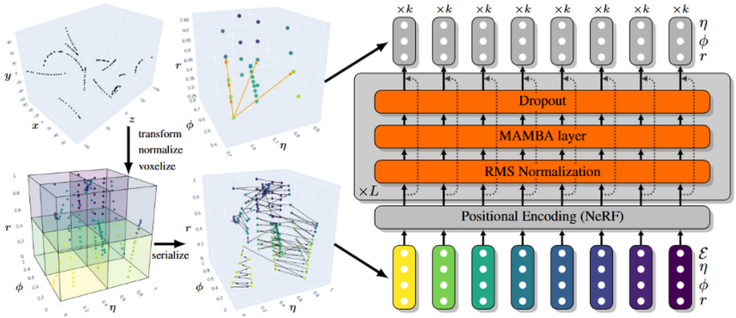
Downstream finetuning with small labeled dataset

Dataset



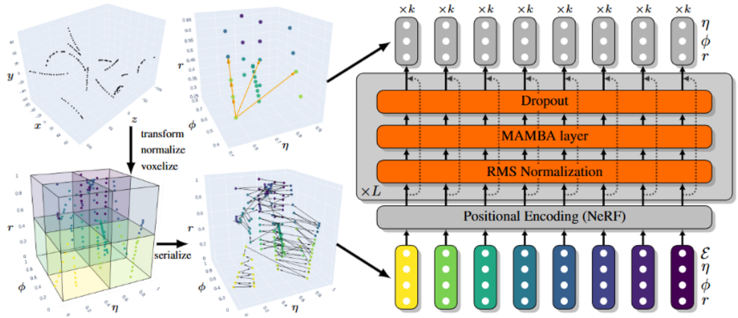
- Generated $\sqrt{s} = 200$ GeV $p+p$ minimum bias PYTHIA events, simulated through sPHENIX Geant4 geometry and reconstructed spacepoints in the sPHENIX TPC using singularity container
- Identify 3 downstream tasks: track finding, noise tagging, and particle identification

Data Pretraining



- Pretraining using unlabeled spacepoint data in a self supervised surrogate task
- Work in normalized (η, ϕ, r) space by voxelizing TPC
- Hierarchical serialization
- Auto-regressive style self-supervised task: predict position of k-nearest neighbor with larger r

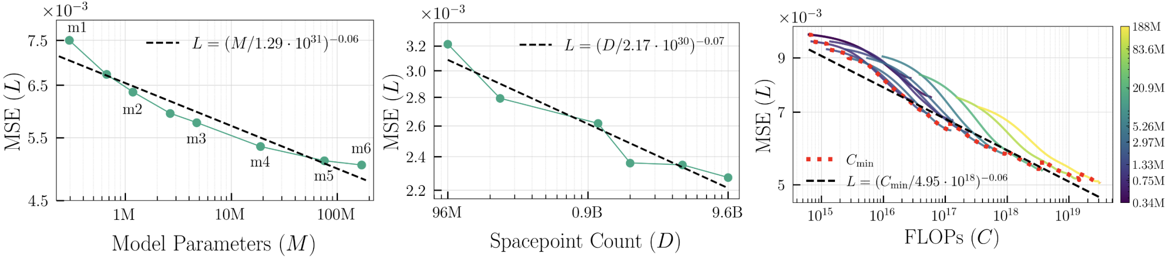
Scaling in Pretraining



- Using pretraining model with parameter size up to almost 200M with 12M minimum bias p+p events
- Tested multiple model and dataset sizes
- Trained on Perlmutter at NERSC for over 10k GPU hours

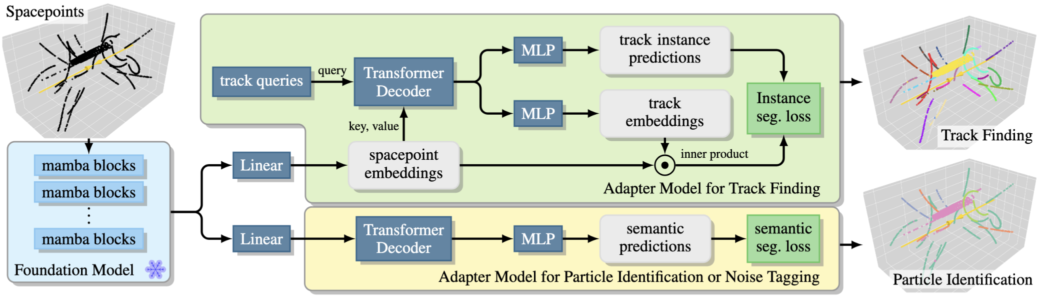
	m1	m2	m3	m4	m5	m6
Model Width	64	128	256	512	1024	1536
Model Params	0.34M	1.3M	5.3M	21M	84M	188M
NVIDIA GPU	H100 80GB		A100 80GB			
Num GPUs	1	1	4	8	24	64
Train Hrs	10	12	20	32	50	72

Neural Scaling Behavior



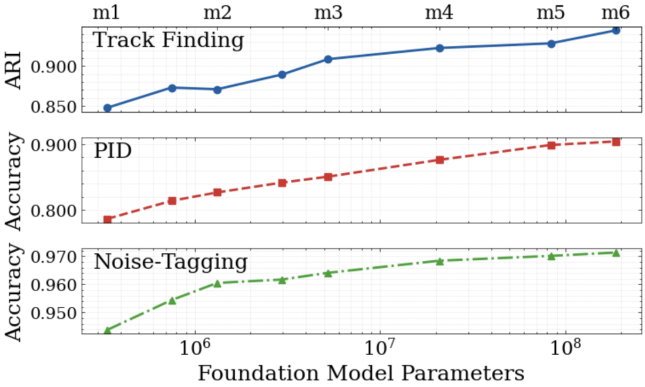
- Log-log scale of MSE loss vs model size shows clear scaling behavior
- Consistent behavior with scaling observed in LLMs
- Model m6 begins to saturate (due to lack of training data?)

Adapter Architecture



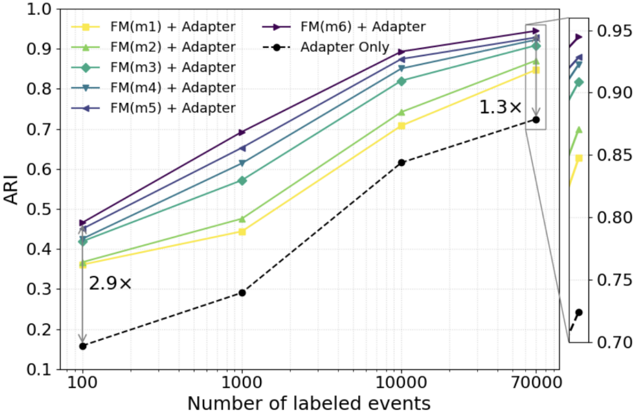
- FM weights from pretraining are frozen
- Lightweight downstream adapter models
  - Track-finding - transformer decoder inspired by instance segmentation model like Maskformer
  - PID/noise tagging - single attention layer + MLP head for per point prediction

Downstream Task Performance



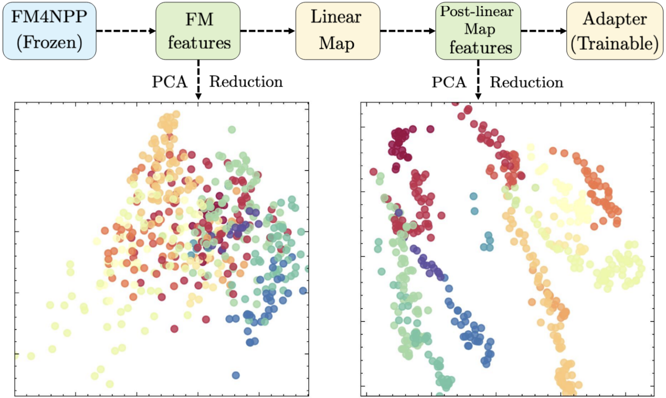
- Each downstream task improves in overall performance with model size
  - → Larger pretrained FMs produce better downstream task performance
- The largest model has an accuracy or ARI of 90% or larger

Track Finding Performance



- Larger pretrained FM outperform smaller ones
  - → Larger FMs contain richer information and can be generalized easier
- The FM pretraining improves the adapter only performance by ∼30%

FM Visualization



- Raw FM embeddings exhibit no clear separation among particle tracks
  - → Representations are task agnostic
- After applying a linear projection, well separated clusters (corresponding to different tracks) emerge
- The FM encodes general purpose representations

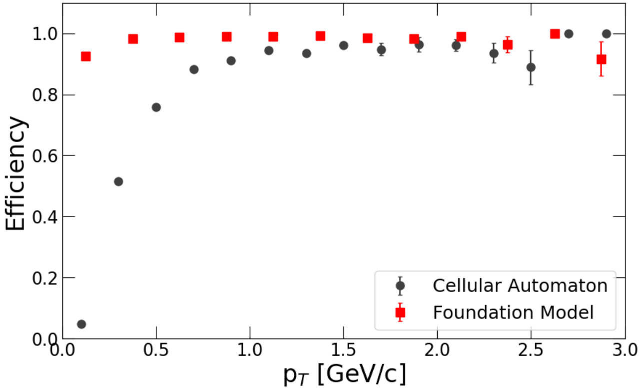
Comparisons to Other Models

- Adapted additional models in the literature to this data set
- We confirm the performance gain is from the FM pre-training by comparing to the "Adapter-only" case
- The FM outperforms all models we tested against on this dataset

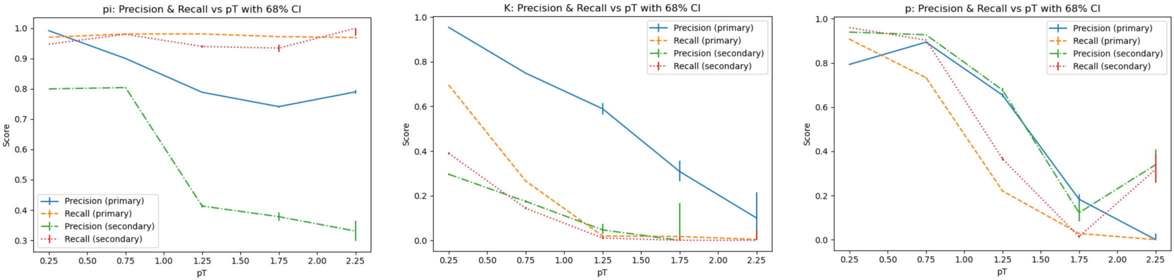
	Track finding				PID			Noise Tagging		
model	ARI↑	efficiency↑	purity↑	model	acc.↑	recall↑	precision↑	acc.↑	recall↑	precision↑
EggNet	0.7256	74.19%	75.14%	SAGEConv	0.7262	0.4563	0.6502	0.9174	0.7227	0.8165
Exa.TrkX	0.8765	91.79%	66.42%	OneFormer3D	0.7701	0.4897	0.5767	0.9646	**0.9404**	0.8948
Adapter Only	0.7243	78.01%	64.54%	Adapter Only	0.6631	0.3387	0.6111	0.9111	0.6215	0.8359
FM4NPP	**0.9395**	**95.85%**	**92.73%**	FM4NPP	**0.8993**	**0.7589**	**0.8689**	**0.9717**	0.9367	**0.9190**

Track Finding Performance

- Ongoing work to benchmark against "traditional" algorithms
- Comparison of track finding efficiency is significantly better than Cellular Automaton based seeding algorithm
- Note - FM is all truth tracks with at least 5 clusters, CASeeder is only primaries with at least 20 clusters in TPC acceptance (!)

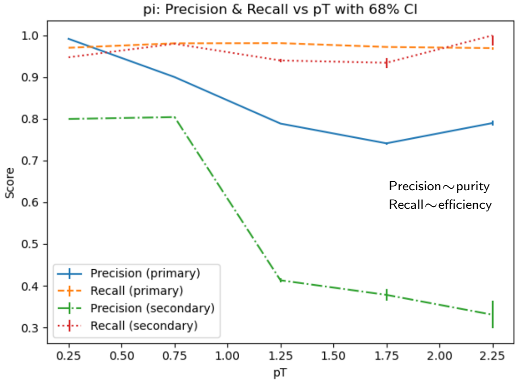


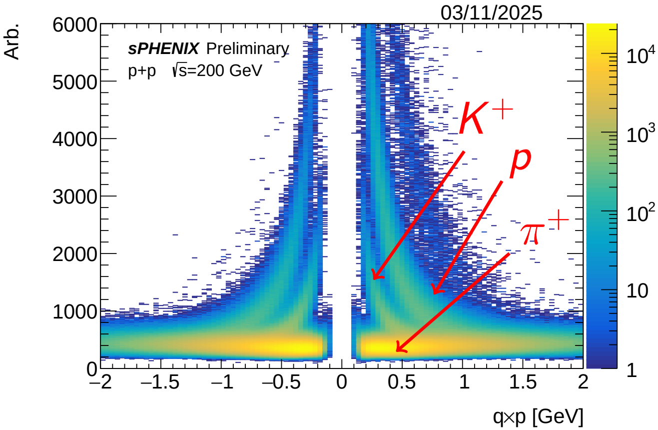
PID Performance



- The FM does a very good job at identifying pions and not misidentifying other particles as pions
- The FM has a strong p_T dependence in its ability to identify K/p (and not misidentify other particles as K/p)

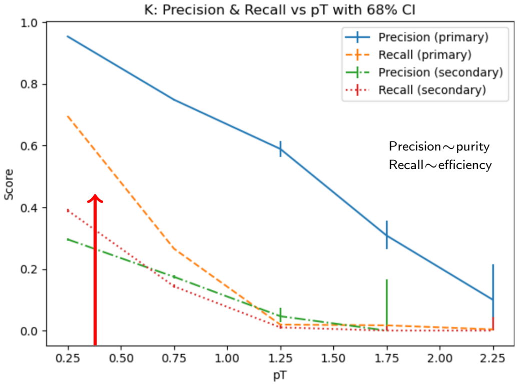
PID Performance

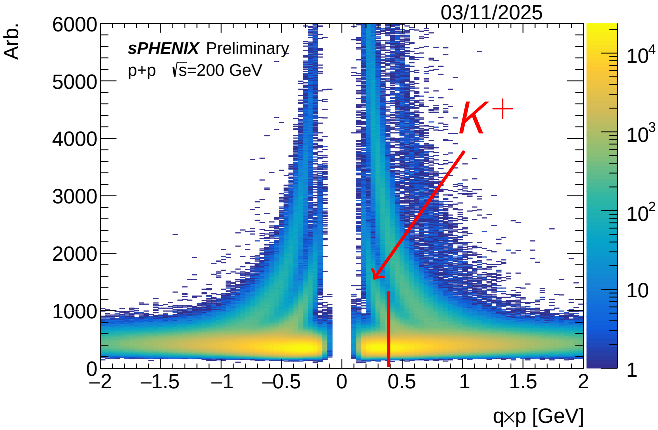




- Pion misidentification increases with p_T
- Kaon misidentification strongly degrading around $\sim$ 400 MeV
- Proton misidentification strongly degrading around $\sim$ 800 MeV

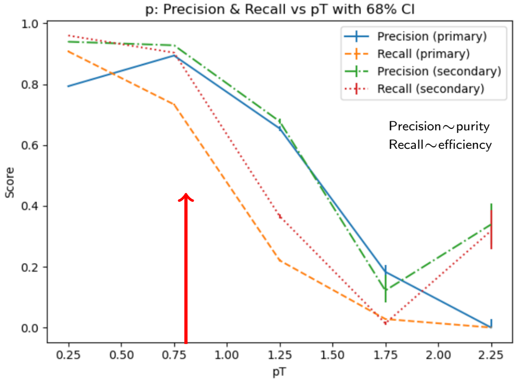
PID Performance

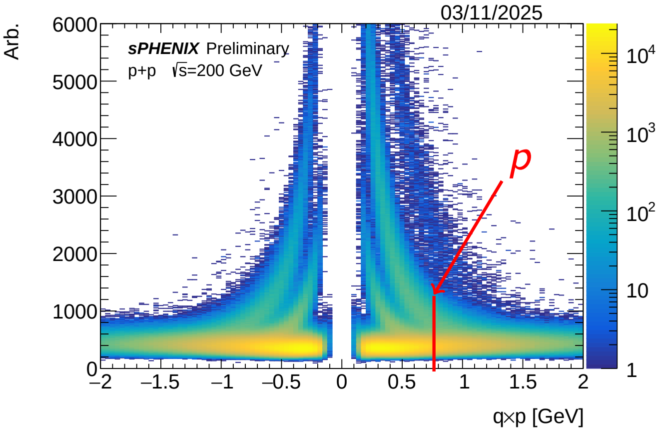




- Pion misidentification increases with p_T
- Kaon misidentification strongly degrading around $\sim$ 400 MeV
- Proton misidentification strongly degrading around $\sim$ 800 MeV

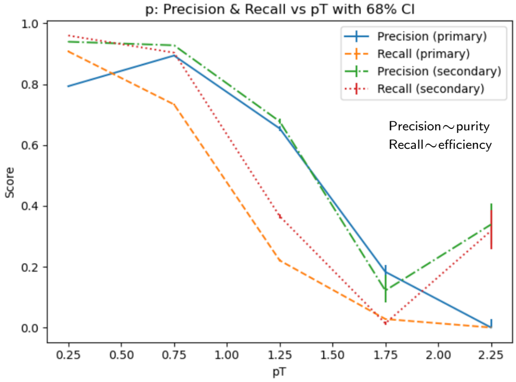
PID Performance

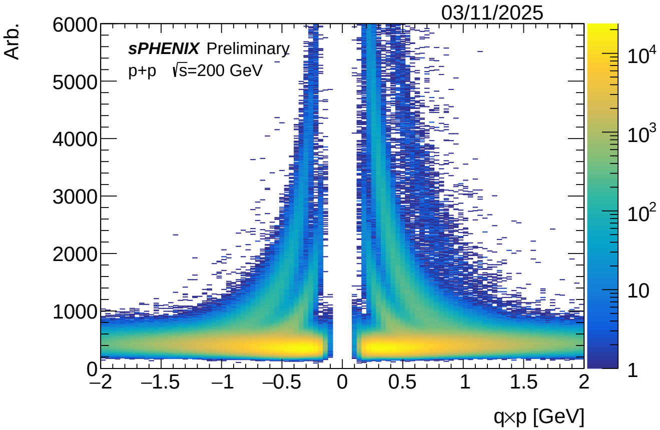




- Pion misidentification increases with p_T
- Kaon misidentification strongly degrading around $\sim$ 400 MeV
- Proton misidentification strongly degrading around $\sim$ 800 MeV

PID Performance

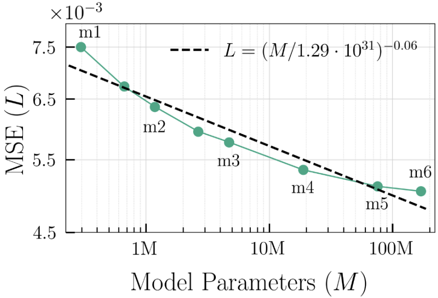


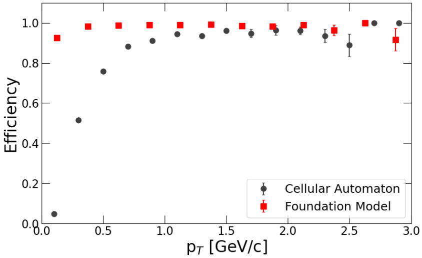


- Pion misidentification increases with p_T
- Kaon misidentification strongly degrading around $\sim$ 400 MeV
- Proton misidentification strongly degrading around $\sim$ 800 MeV

FM4NPP learns about the energy loss characteristics of the TPC and uses it for PID!

Conclusions



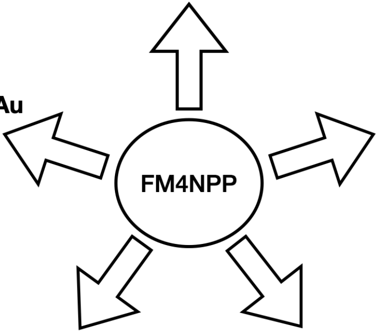


- We demonstrated that a FM trained on sPHENIX TPC spacepoints scales as expected based on other LLMs/FMs
- The pre-trained FM improves the performance of downstream track finding, PID, and noise tagging
- When the FM model size is larger, better downstream performance is achieved
- This model surpasses the performance of other GNN based models in the literature

Future Work

Further analysis downstream tasks (e.g. heavy flavor ID, B jet tagging…)

Realistic beam environments (Au+Au collisions, p+p with full pileup)

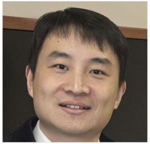
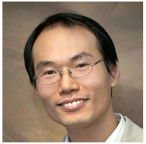
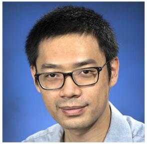


Different experiments and/or beam conditions (transfer learning)

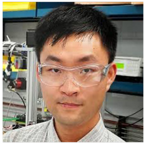
Holistic reconstruction (e.g. particle flow)

Additional sPHENIX subsystems (e.g. silicon, calos)

FM4NPP Team

(AI Dept.). David Park Yi Huang Xihaier Luo Yuewei Lin Shinjae Yoo Yihui “Ray” Ren

(Phys Dept.) Shuhang Li (Columbia) Haiwang Yu Joe Osborn Yeonju Go Jin Huang