Foundation Models for Nuclear and Particle Physics (FM4NPP)

Joe Osborn

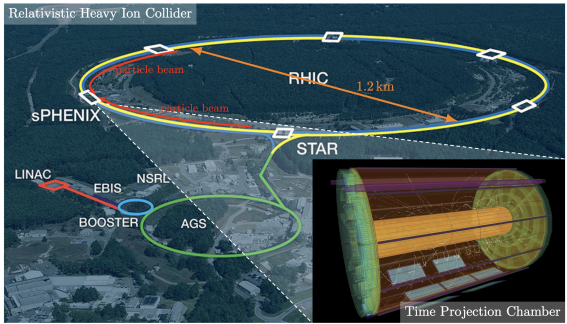
Brookhaven National Laboratory

September 16, 2025

Based on:
arXiv:2508.14087, accepted to ICLR
IJMPA, Vol. 41, 15 2650086

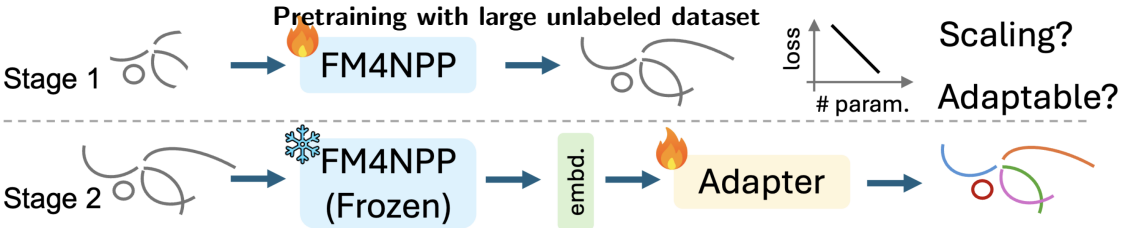
Physics Motivation

- 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

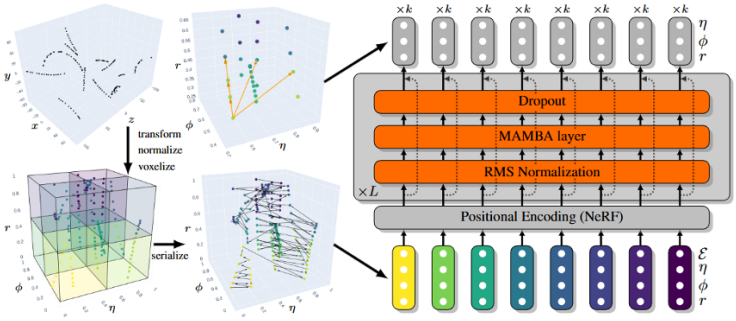


FM4NPP Goal

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

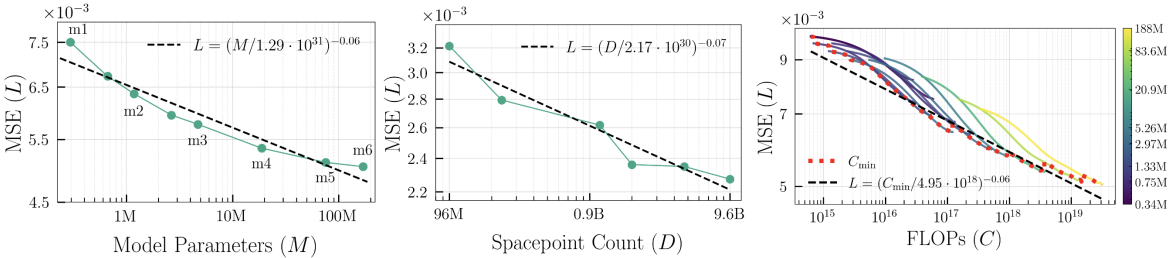


Data Pretraining



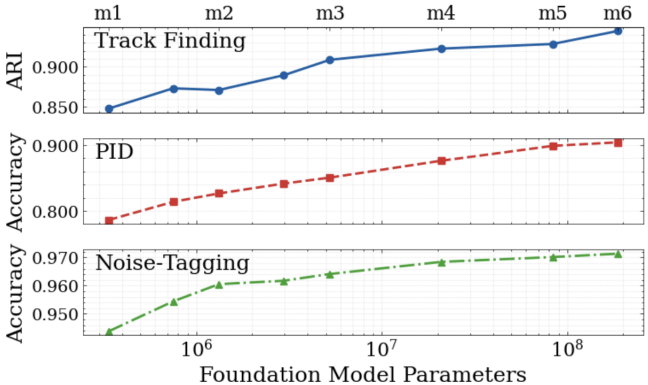
- Data: $\sqrt{s} = 200$ GeV p+p minimum bias PYTHIA events, simulated through sPHENIX Geant4 geometry
- Pretraining using unlabeled TPC spacepoints in a self supervised surrogate task
- Work in normalized (η, ϕ, r) space by voxelizing TPC
- Hierarchical serialization
- Auto-regressive style self-supervised task: predict position of k-nearest neighbor with larger r

Neural Scaling Behavior



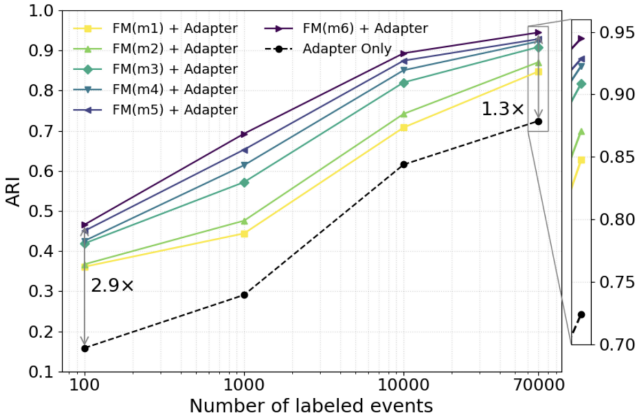
- Log-log scale of MSE loss vs model size shows clear scaling behavior
- Consistent behavior with scaling observed in LLMs
- Model m6 begins to saturate (due to lack of training data?)

Downstream Task Performance



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

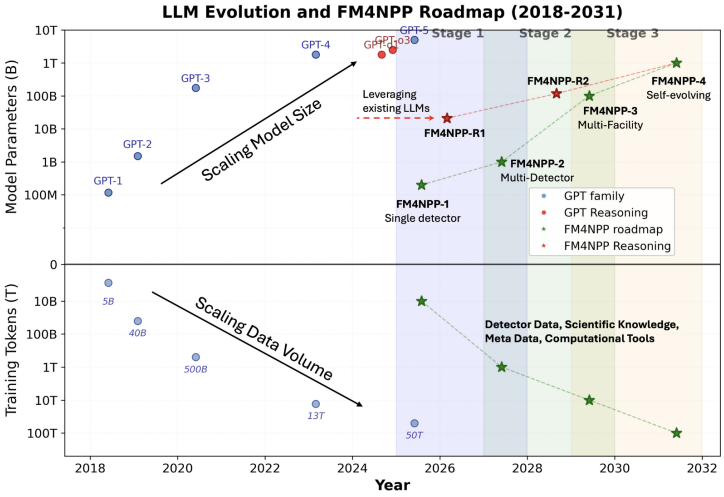
Track Finding Performance



- 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%

Future Work

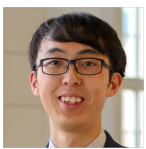
- FM4NPP evolution to follow similar model size scaling as (e.g.) ChatGPT
- LDRD for FY27/28 focused on multi-modal data with silicon and calorimeters
- Potential for future funding to explore domain shift studies with data from other facilities (ATLAS, DUNE, LEP/HERA, EIC)

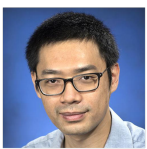


Y. Ren, JDO et al. IJMPA, Vol. 41, 15 2650086

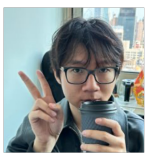
FM4NPP Team

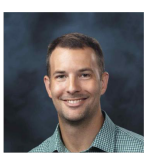


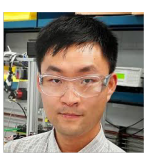


(AI Dept.). David Park Yi Huang Xihaier Luo Yuewei Lin Shinjae Yoo Yihui “Ray” Ren





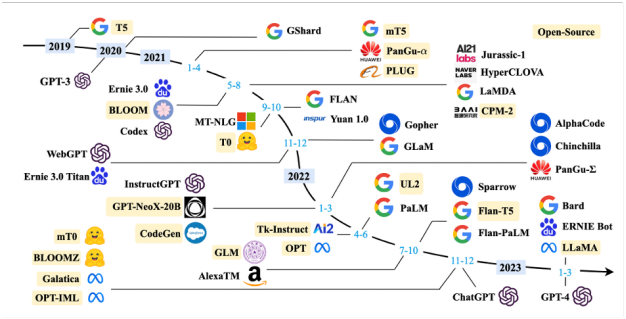


(Phys Dept.) Shuhang Li (Columbia) Haiwang Yu Joe Osborn Yeonju Go Jin Huang

Extras

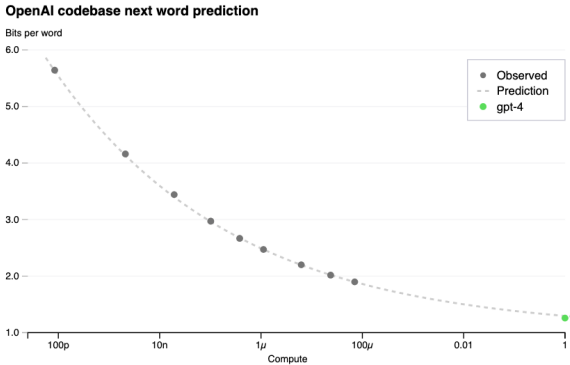
Scaling Large Language Models

- 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)



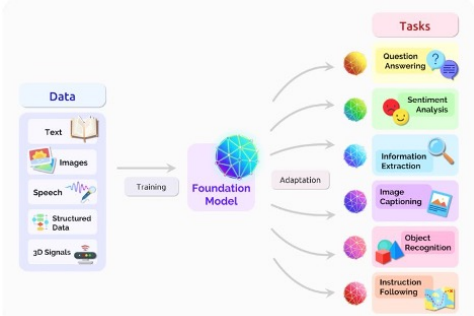
Scaling Large Language Models

- 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)

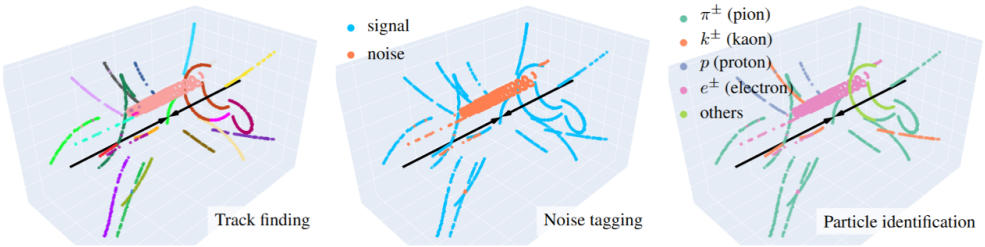


Foundation Models

- 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

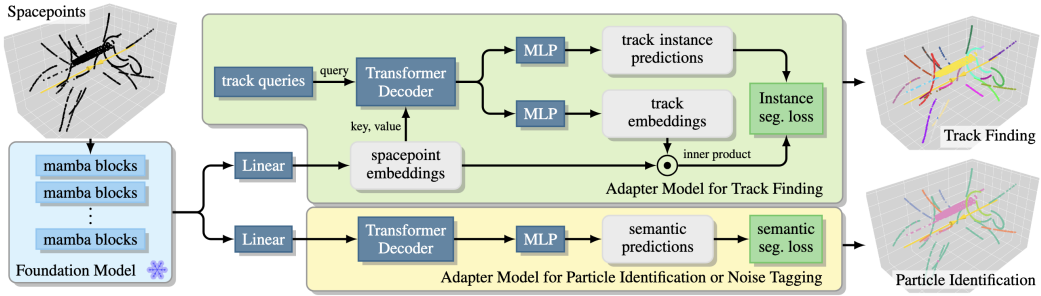


Dataset



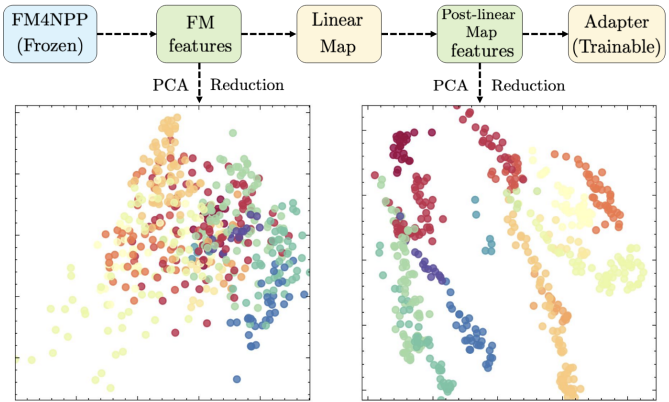
- Generated $\sqrt{s} = 200$ GeV p+p minimum bias PYTHIA events, simulated through sPHENIX Geant4 geometry and reconstructed spacepoints in the sPHENIX TPC using singularity container
- Identify 3 downstream tasks: track finding, noise tagging, and particle identification

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

FM Visualization



- Raw FM embeddings exhibit no clear separation among particle tracks
  - $\rightarrow$ 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**