FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics

Presenters:

Shuhang Li, Physics Department, Columbia University (sli7@bnl.gov)

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

Presented at *New York Scientific Data Summit (NYSDS)* 2025

Experimental Nuclear and Particle Physics

- Explores the fundamental building blocks of matter and the forces governing their interactions.

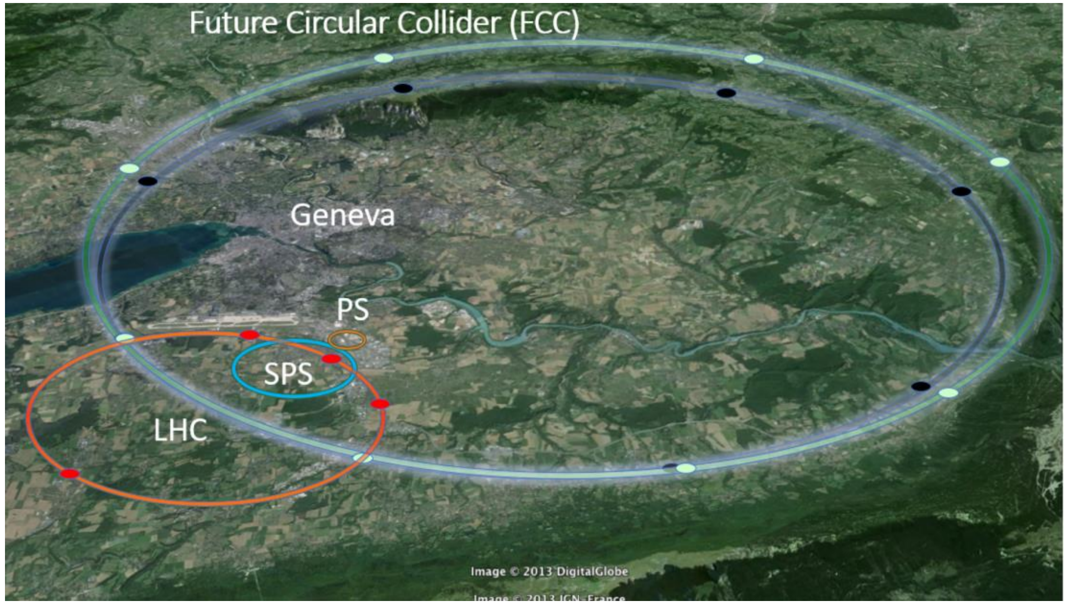
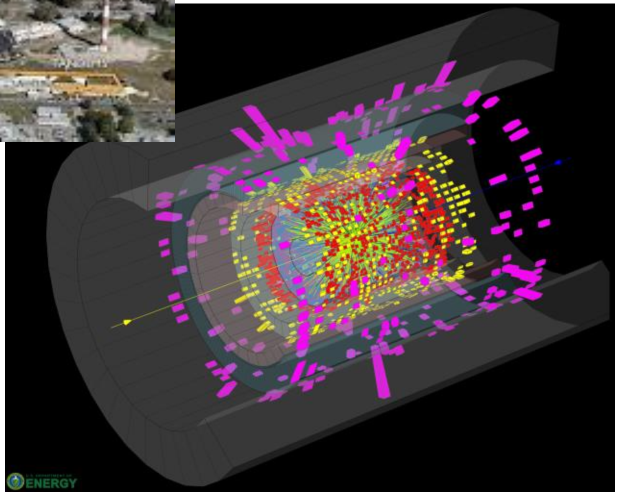
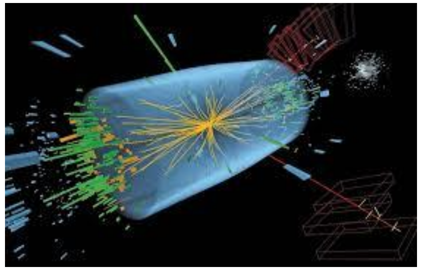


Image Credit: [ICMAB]

RHIC/EIC



LHC

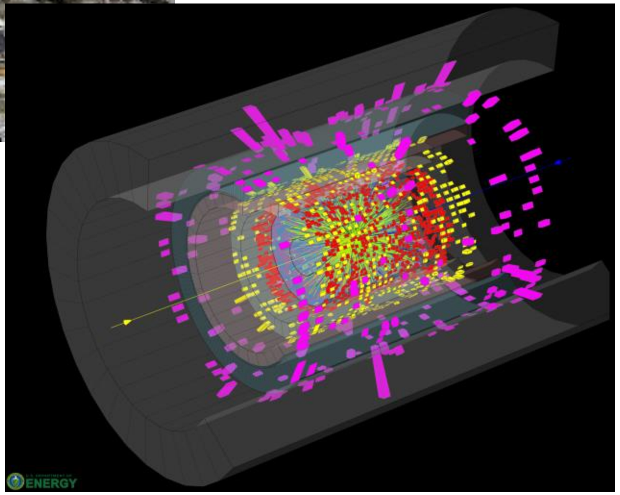


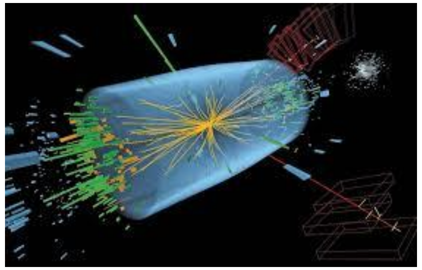
Brookhaven National Laboratory

Experimental High Energy Physics

- Explores the fundamental building blocks of matter and the forces governing their interactions.
- Large dataset
- Complicated analysis to reach physics result; Diverse tasks
- Various application with task specific ML models
- **Can foundation models provide a unified framework to accelerate discovery across diverse tasks?**

RHIC/EIC

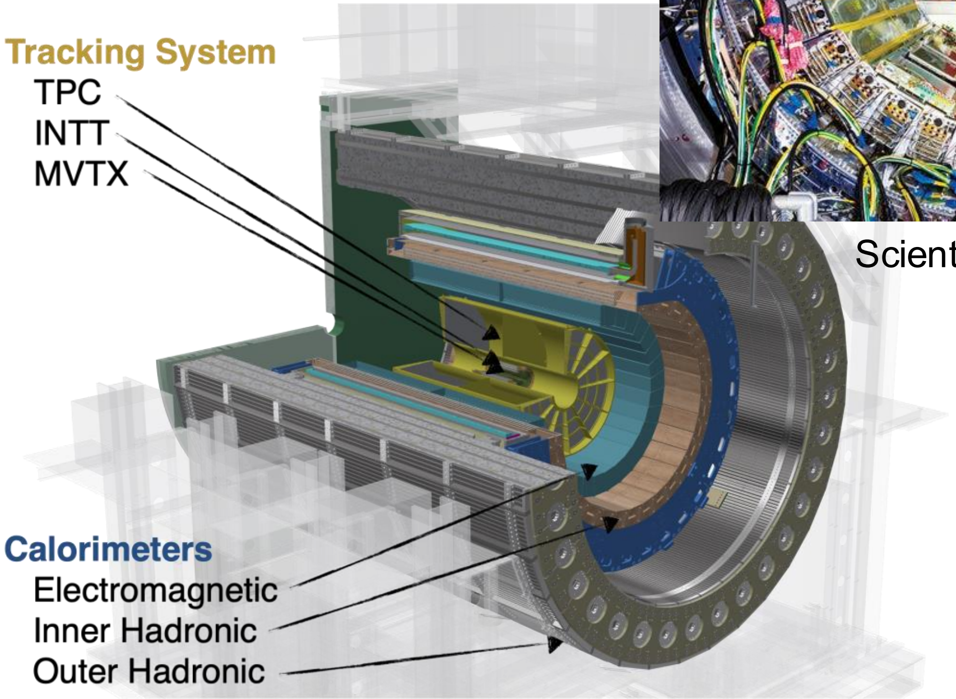


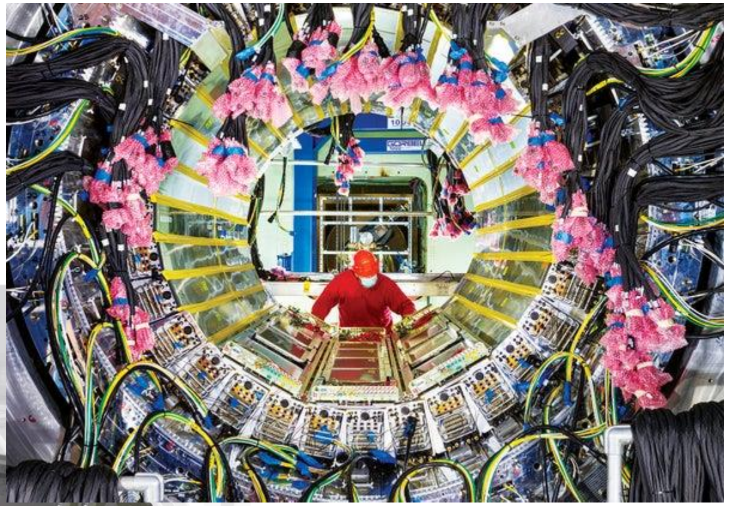


LHC

Brookhaven National Laboratory

sPHENIX at BNL





Scientific American, 03/01/2023

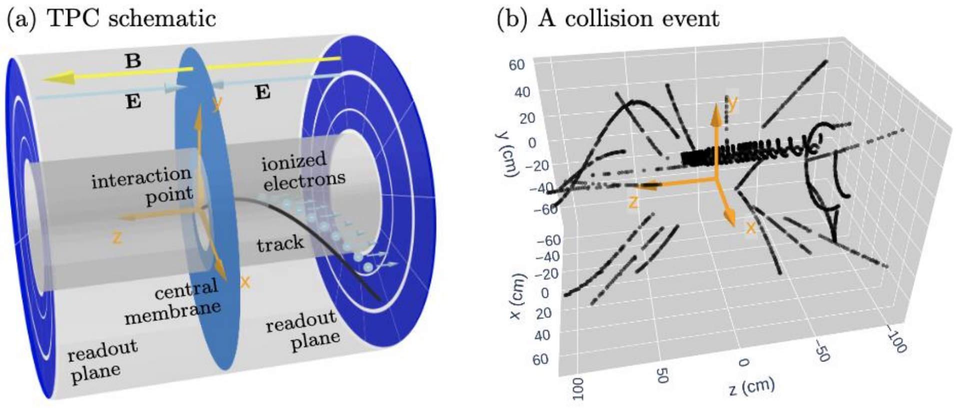
The largest particle collider in U.S.
Data taking began in 2023!
High-precision tracking system + Hermetic Electromagnetic & Hadronic calorimeters

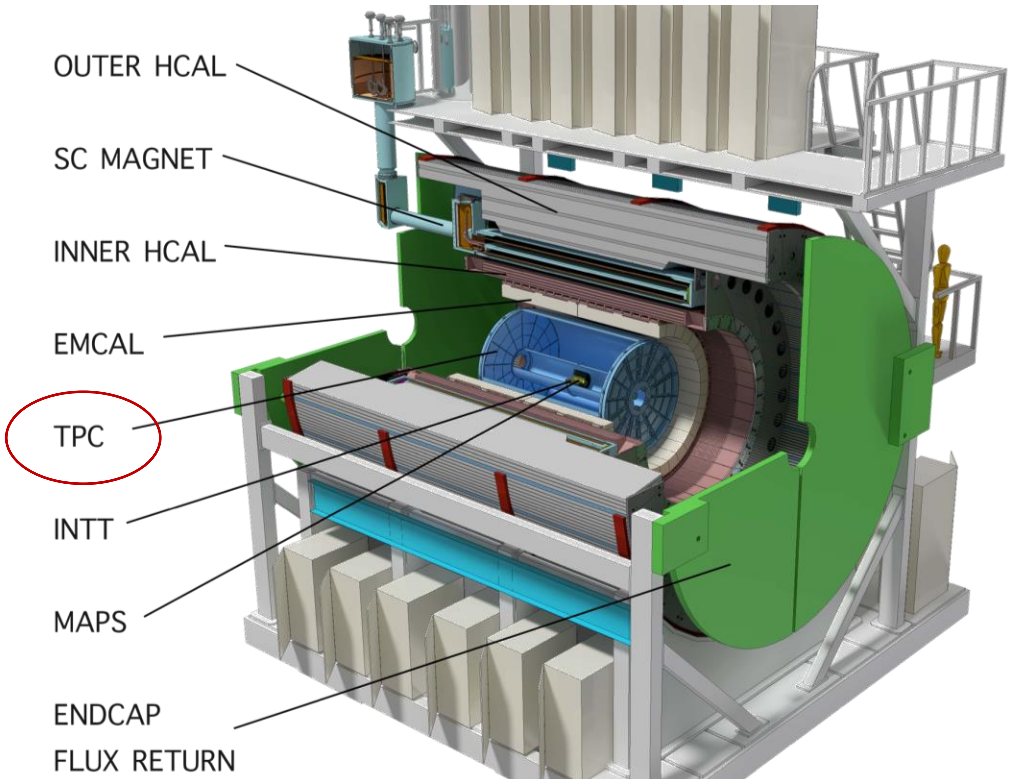
Simulated Dataset

sPHENIX Time Projection Chamber (TPC) spacepoints

Data:

- 10M events for self-supervised pre-training
- 75k events for downstream tasks supervised training.
- p+p collisions at $\sqrt{s}$ = 200 GeV
- Simulated using sPHENIX software tool chains.
- Publicly available: [TPCpp-10M]





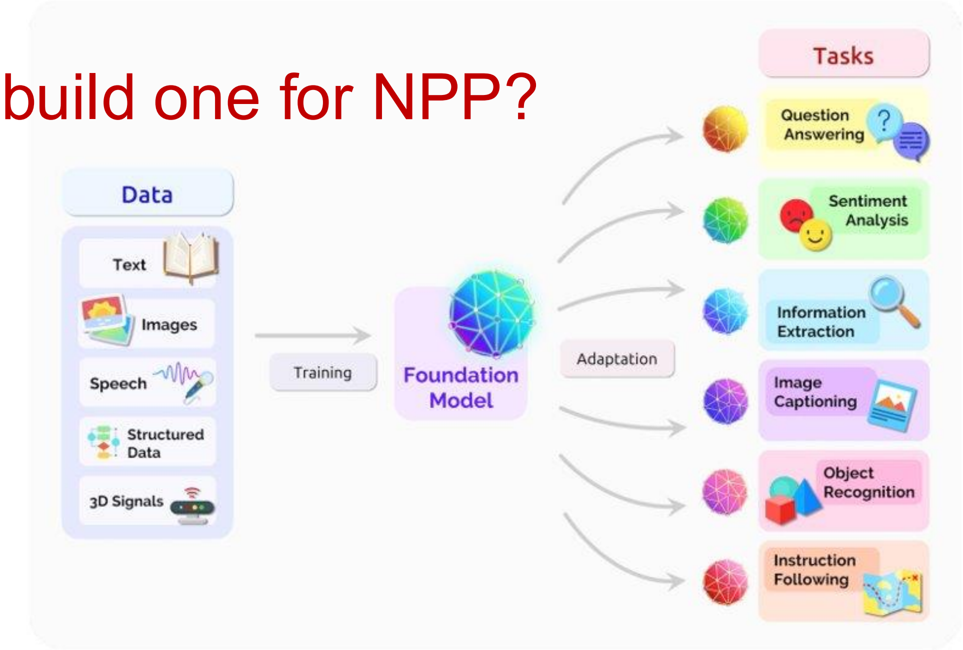
Brookhaven National Laboratory

Foundation Model

Can we build one for NPP?

Foundation Models (FMs) are envisioned as a counterpart to text-based LLM, but they can handle multiple types of data.

- Built on large-scale, primarily unlabeled data
- Capable of handling multiple modalities
- Trained via self-supervised learning on surrogate tasks
- Pre-trained and adaptable to diverse downstream applications
- Achieve state-of-the-art performance across application tasks
- Exhibit strong neural scaling behavior



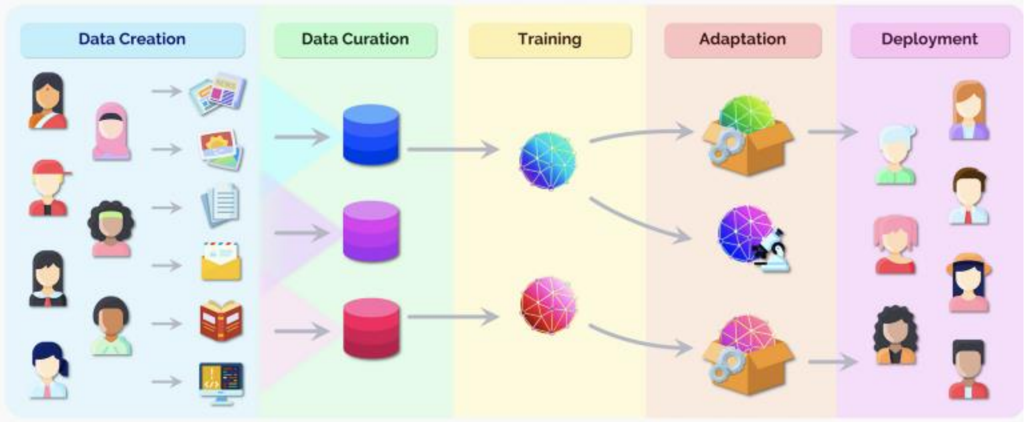


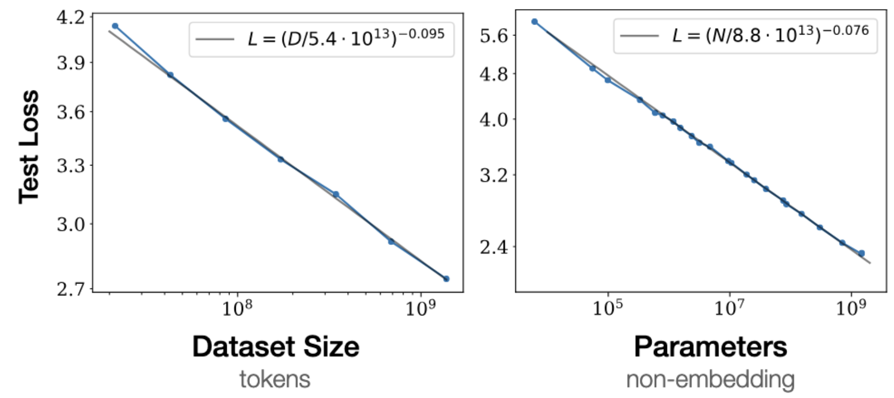
Image credit [1]

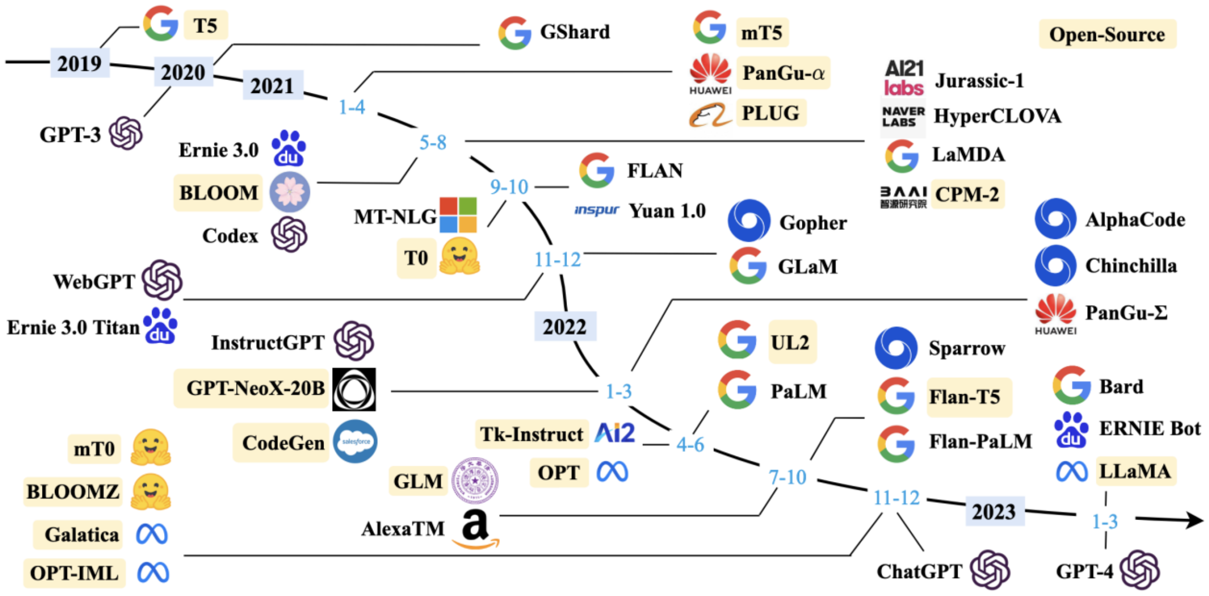
[1] Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).

Neural Scaling Behavior

(2020) Neural Scaling Laws [1]

(20-23) LLM "Arms Race"





[1] Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv:2001.08361* (2020).

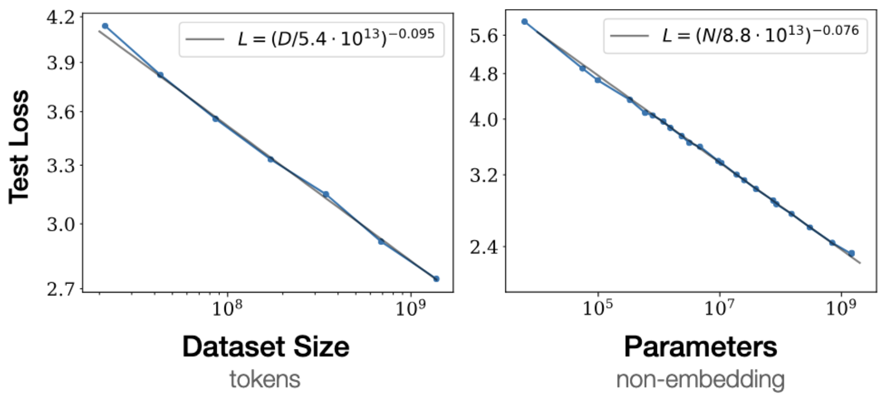
Brookhaven National Laboratory

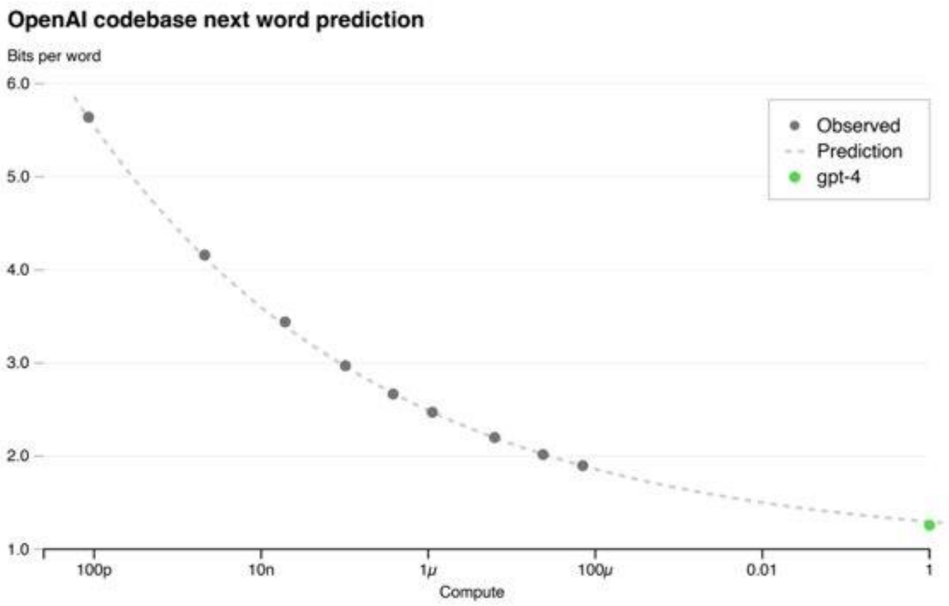
Neural Scaling Behavior

(2020) Neural Scaling Laws [1]

(20-23) LLM “Arms Race”

(2023) Scaling behavior holds for GPT-4 [2]





[1] Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv:2001.08361* (2020).
[2] Achiam, Josh, et al. "Gpt-4 technical report." *arXiv:2303.08774* (2023).

Brookhaven National Laboratory

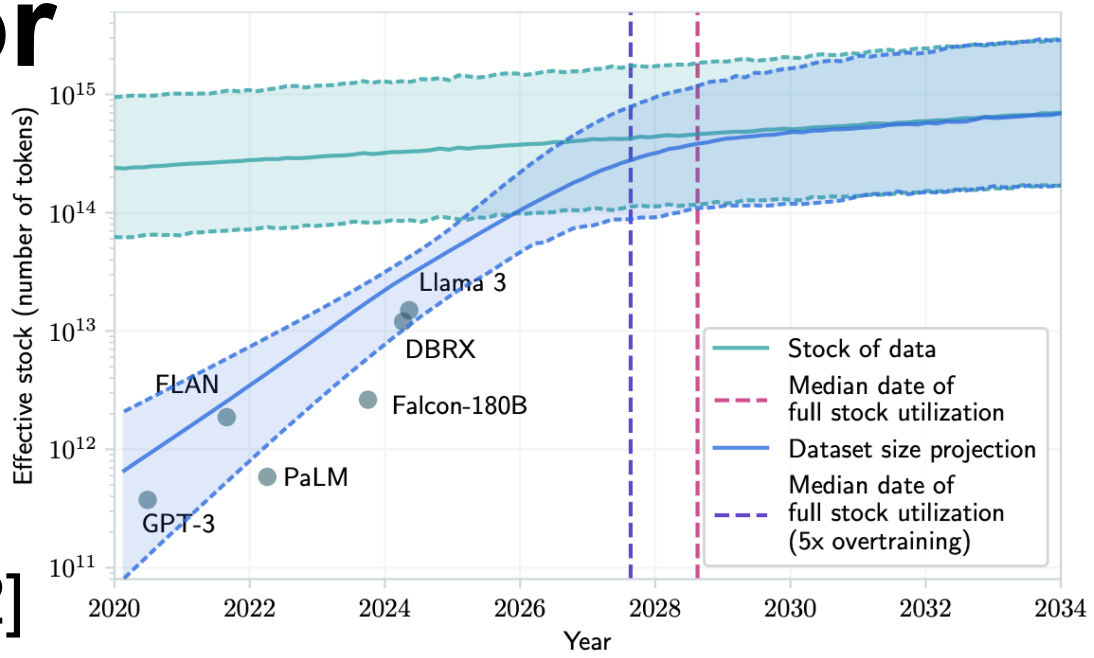
Neural Scaling Behavior

(2020) Neural Scaling Laws [1]

(20-23) LLM “Arms Race”

(2023) Scaling behavior holds for GPT-4 [2]

(2024) End of the scaling? [3,4]



Article | Open access | Published: 24 July 2024

AI models collapse when trained on recursively generated data

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal

Nature **631**, 755–759 (2024) | Cite this article

[1] Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv:2001.08361* (2020).
[2] Achiam, Josh, et al. "Gpt-4 technical report." *arXiv:2303.08774* (2023).
[3] Villalobos, Pablo, et al. "Will we run out of data? an analysis of the limits of scaling datasets in machine learning." arXiv preprint arXiv:2211.04325 1 (2022).
[4] Shumailov, Ilia, et al. "AI models collapse when trained on recursively generated data." Nature 631.8022 (2024): 755-759.

Neural Scaling Behavior

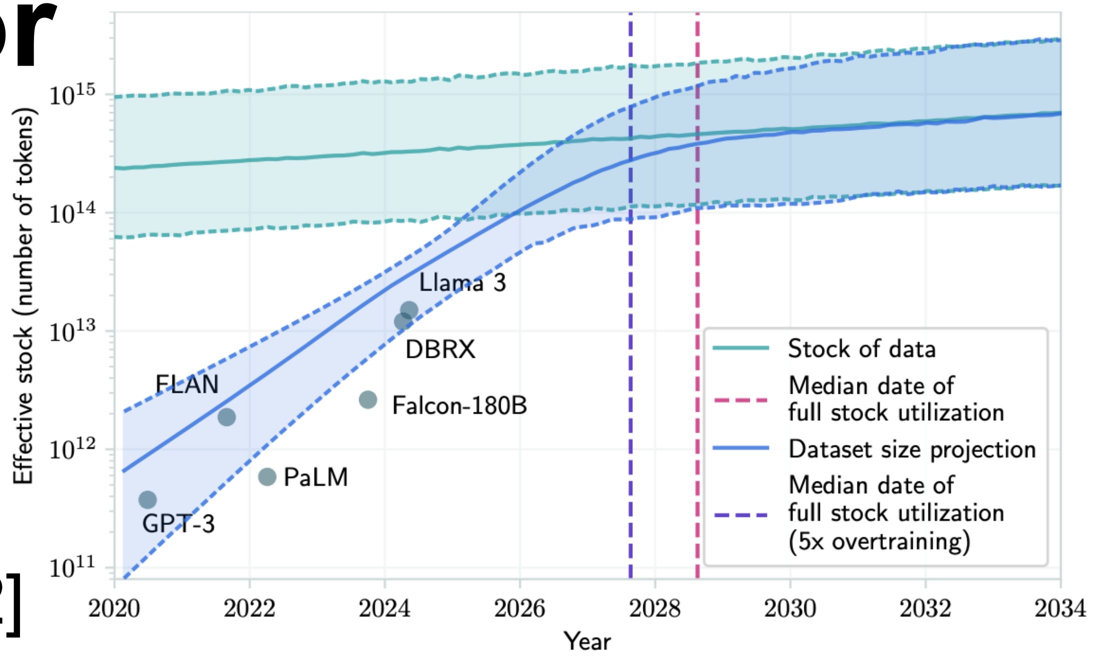
(2020) Neural Scaling Laws [1]

(20-23) LLM “Arms Race”

(2023) Scaling behavior holds for GPT-4 [2]

(2024) End of the scaling? [3,4]

Scientific data are “uncharted terrain”
Can we repeat the success of LLMs?



Article | Open access | Published: 24 July 2024

AI models collapse when trained on recursively generated data

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal

Nature **631**, 755–759 (2024) | Cite this article

Brookhaven National Laboratory

Scientific Motivation

Proof of concept for an FM4NPP:

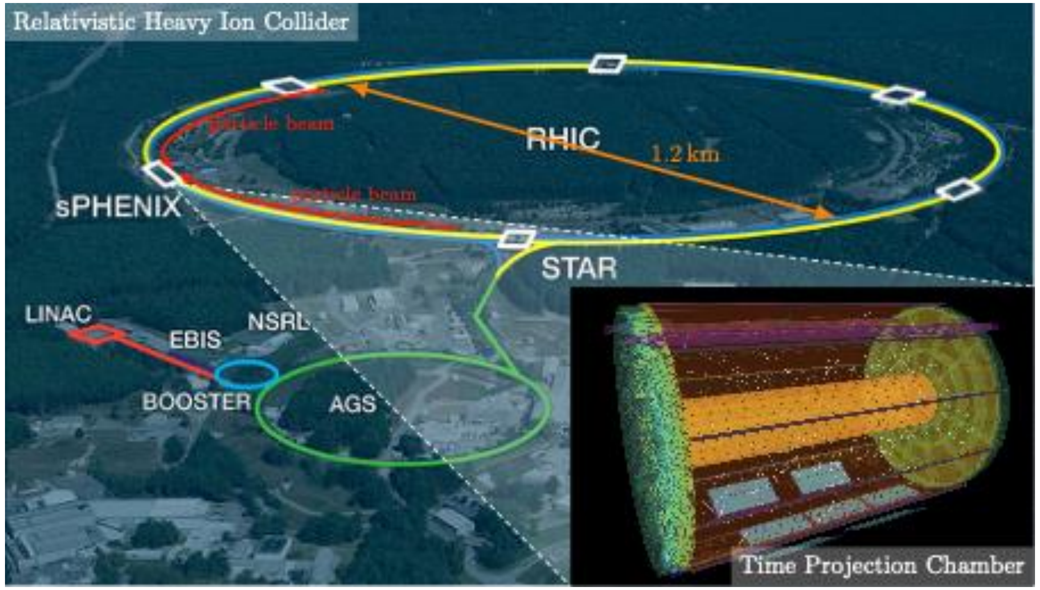
1. Neural scaling behavior
2. Generalizable to downstream tasks

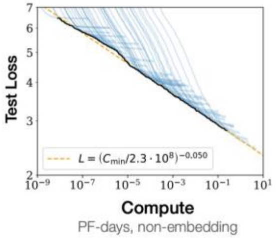
Two-stage approach:

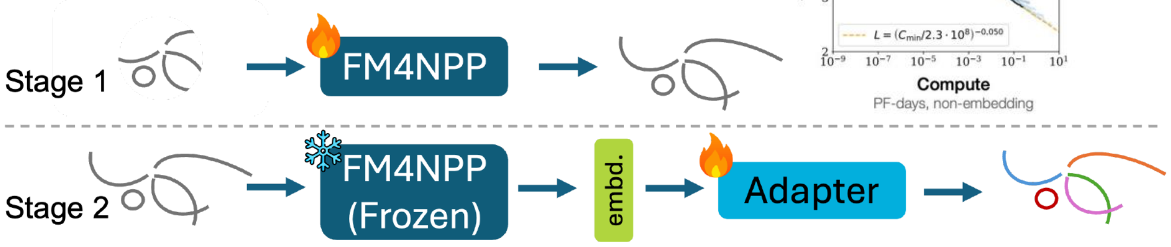
1. Large-scale FM pre-training on unlabeled data
2. Adapt the frozen FM for various tasks

Key Questions:

1. What is the right self-supervised learning task?
2. Will the FM pre-training scale?
3. Will the representation learned from FM useful?
4. Will larger FM also lead to better downstream tasks?
5. Will adapting from FM be more data efficient?

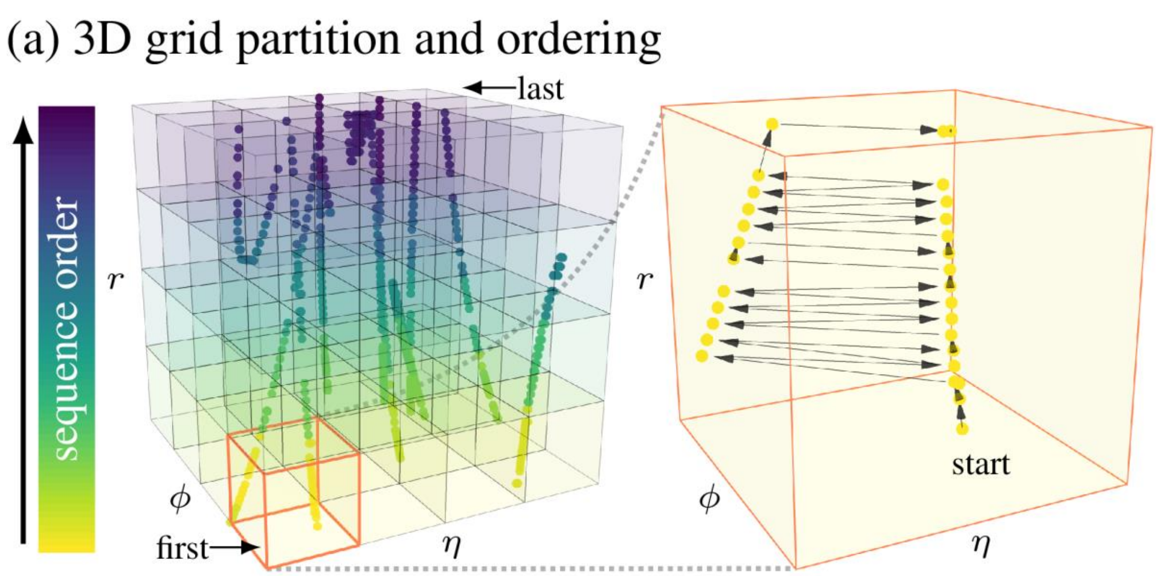


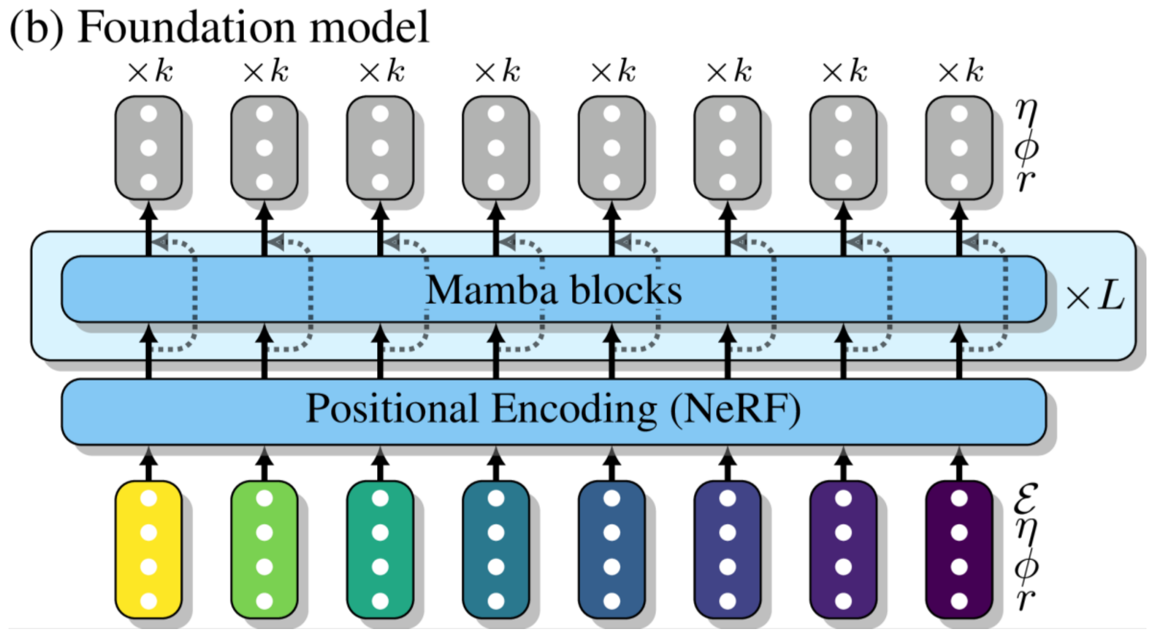




Key Innovations

- Serialization: **Hierarchical Raster Scan**
- Self-supervised learning: **Next k-nearest-neighbor Prediction**
- Adaptation to Mamba Model and large-scale training (e.g., μTransfer)

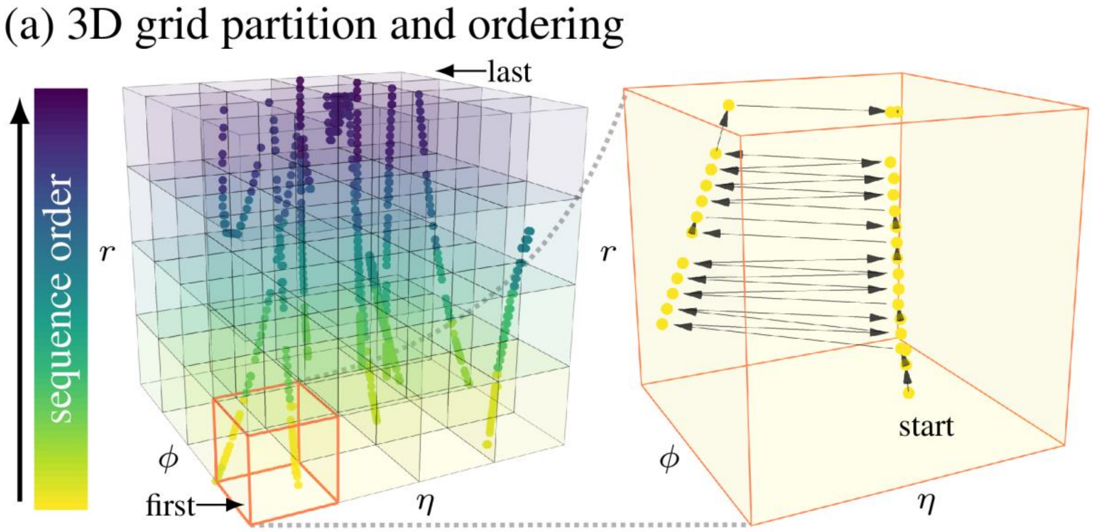


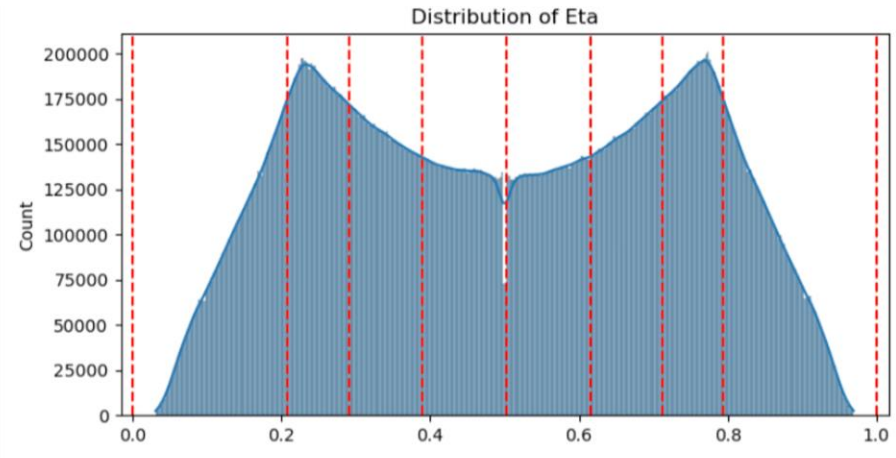


Hierarchical Raster Scan

HRS serializes spacepoints into a 1D sequence.

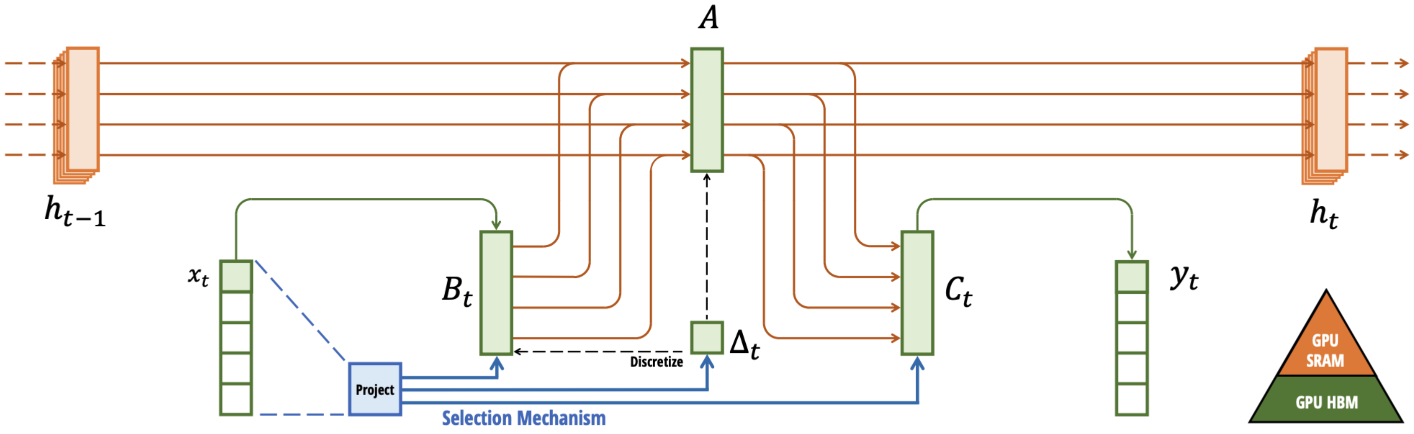
- First, divide the space into boxes, 6x8x8 in (r, η, ϕ).
- Within each box, order the points based on radius.
- Boxes are ordered from inner most box to the outermost box.

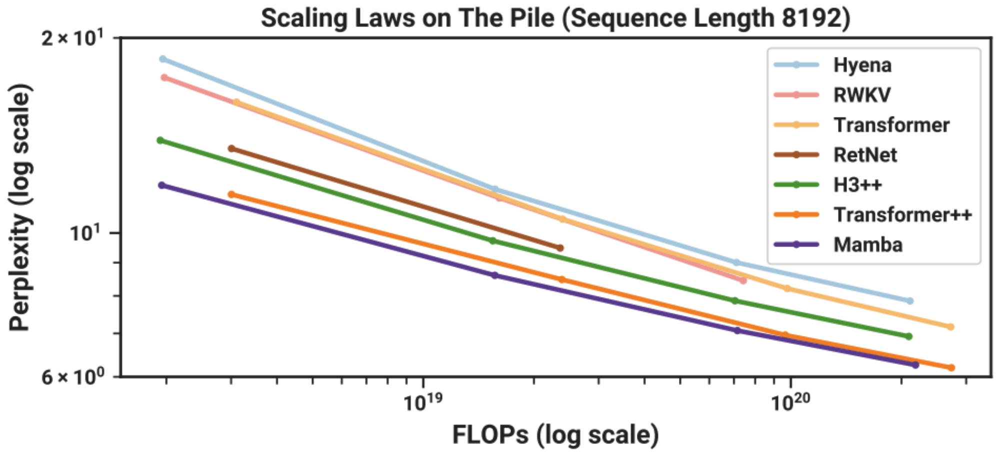




MAMBA: State Space Model (SSM)

- **Structured State Space Models (SSMs)** improve computational efficiency while maintaining long-range sequence modeling capabilities.
- **Continuous-time modeling:** Some variants of Mamba build on continuous-time formulations (in contrast to discrete-time models like RNNs).
- **Efficient implementation:** Mamba achieves linear time and memory complexity – something Transformers cannot do.
- We adapted the **Mamba** model [1].

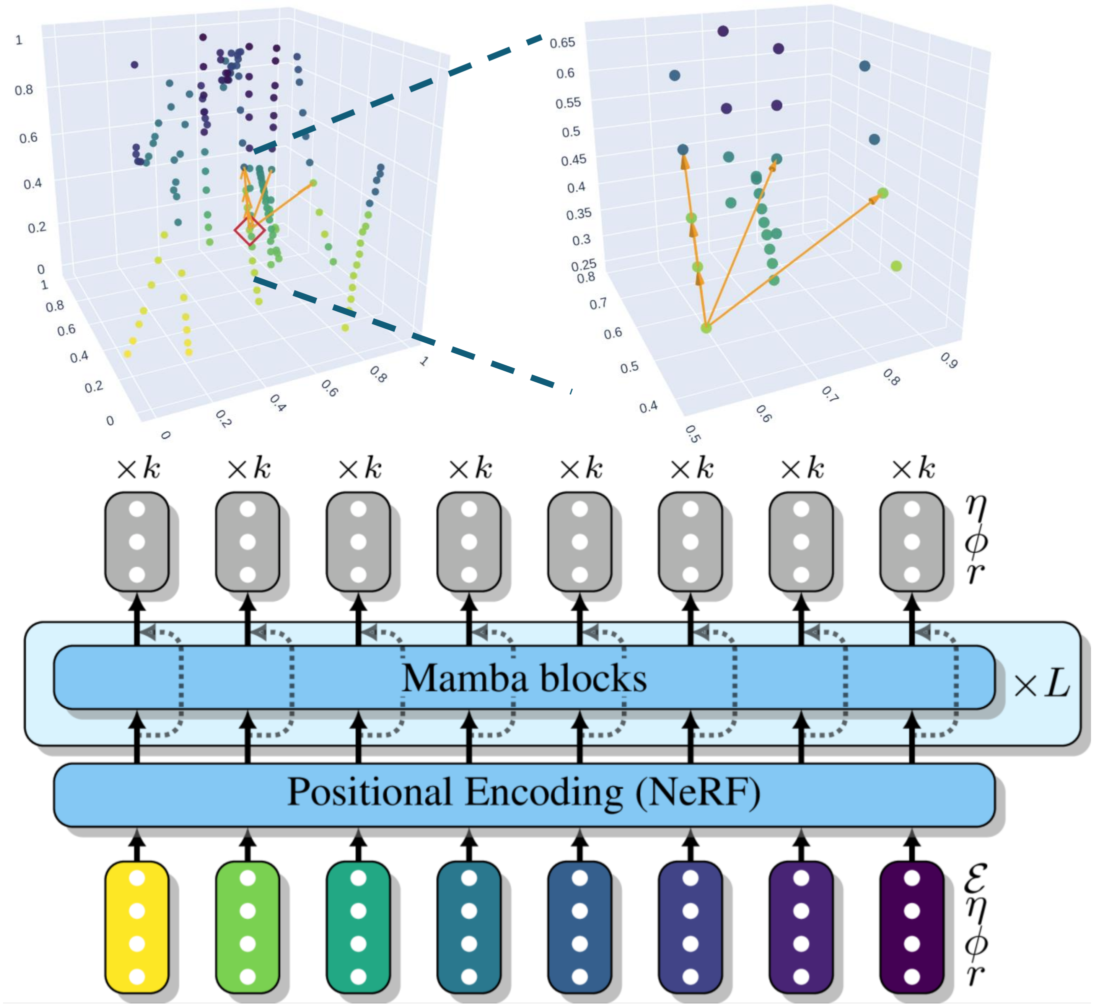




[1] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*.

Next k-nearest-neighbor Prediction

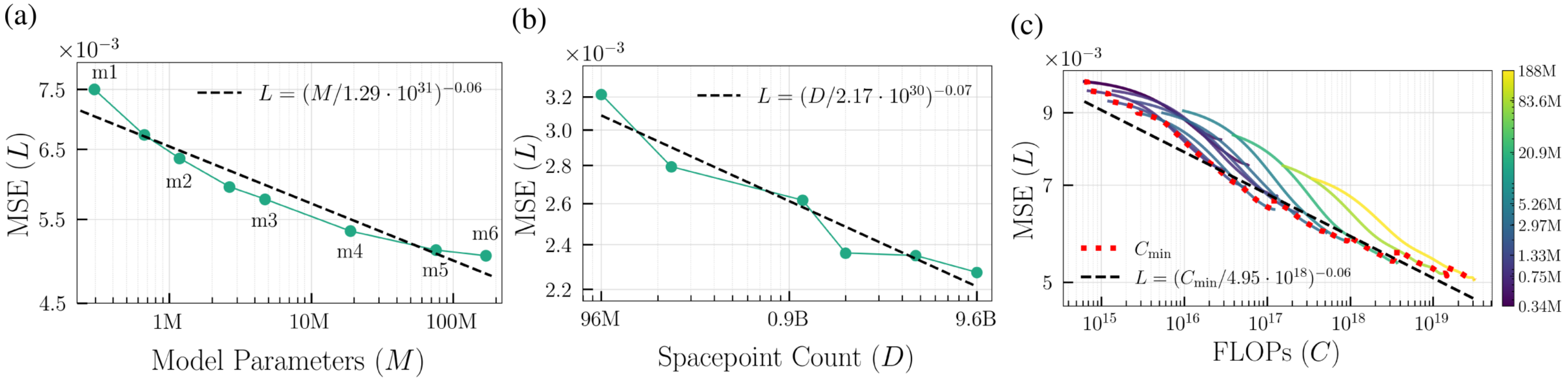
- HRS Serialization is not following a track. This means ordinary autoregression or “next token” prediction may be too random.
- Predicting next k-nearest neighbor maybe better as more points may fall into the same track.



Brookhaven National Laboratory

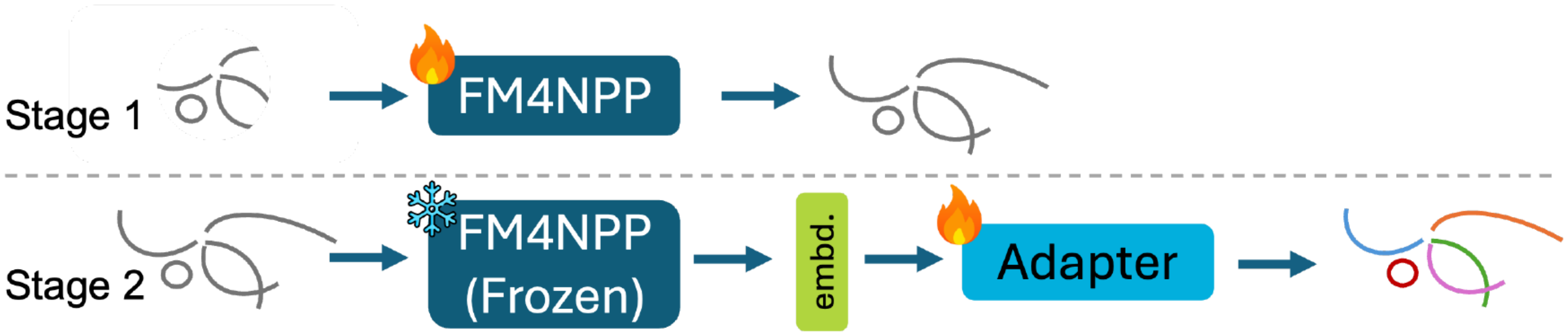
Neural Scaling Behavior

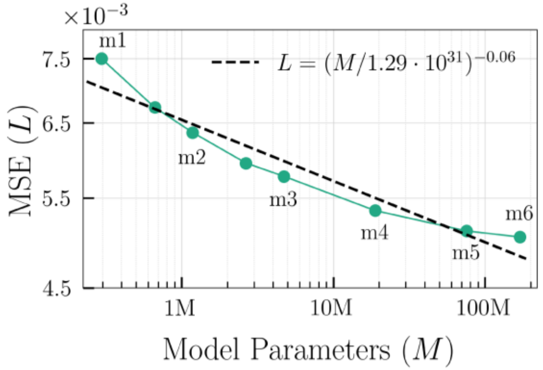
	Model Sizes					
	m1	m2	m3	m4	m5	m6
Model Width	64	128	256	512	1024	1536
Model Params	0.34M	1.3M	5.3M	21M	84M	188M



- Log-log scale of MSE loss versus # Model Parameters, # Spacepoints and Compute
- Model m6 begins to saturate (may be due to lack of training data).

Will the FM Features be Useful?



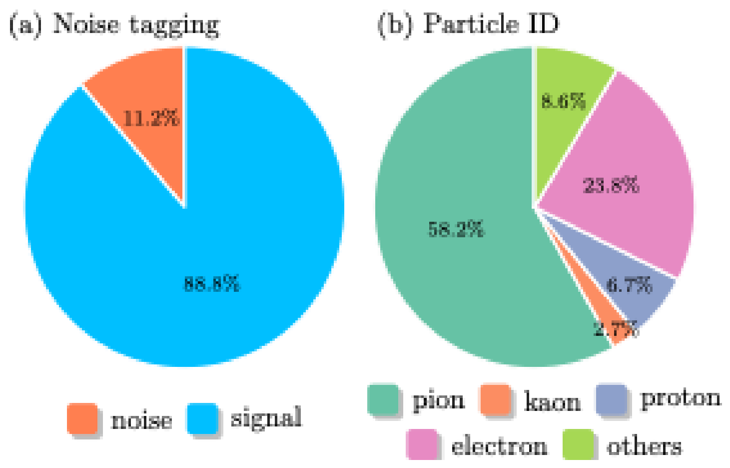


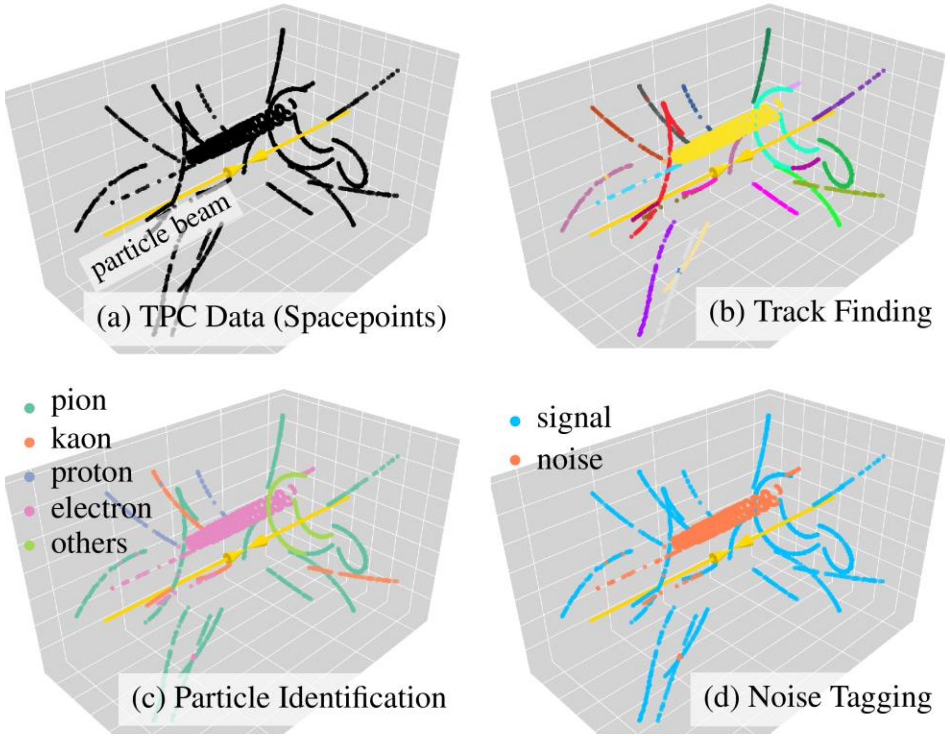
1. ✅ What is the right self-supervised learning task?
2. ✅ Will the FM pre-training scale?
3. Will the representation learned from FM be useful?
4. Will larger FM also lead to better downstream tasks?
5. Will adapting from FM be more data efficient?

Downstream Tasks

Three **downstream tasks**:

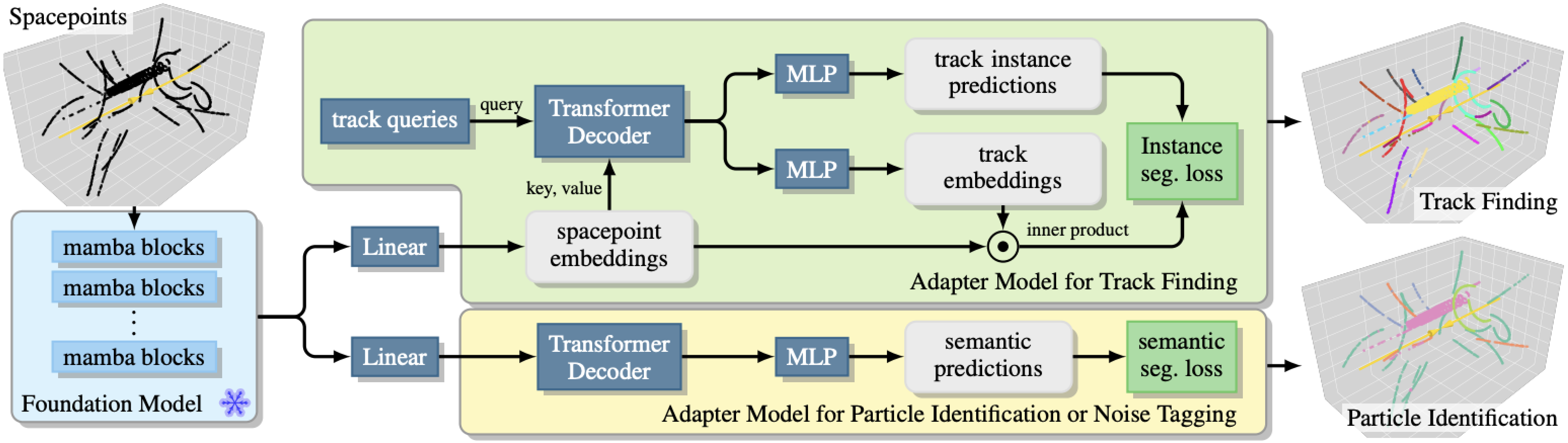
1. **Track Finding**: group spacepoints from the same particle together
2. **Noise Tagging**: classify spacepoints from low momentum secondary particles ("noise")
3. **Particle Identification**: classify spacepoints based on their particle species





Brookhaven National Laboratory

Adapting FM for Downstream Tasks



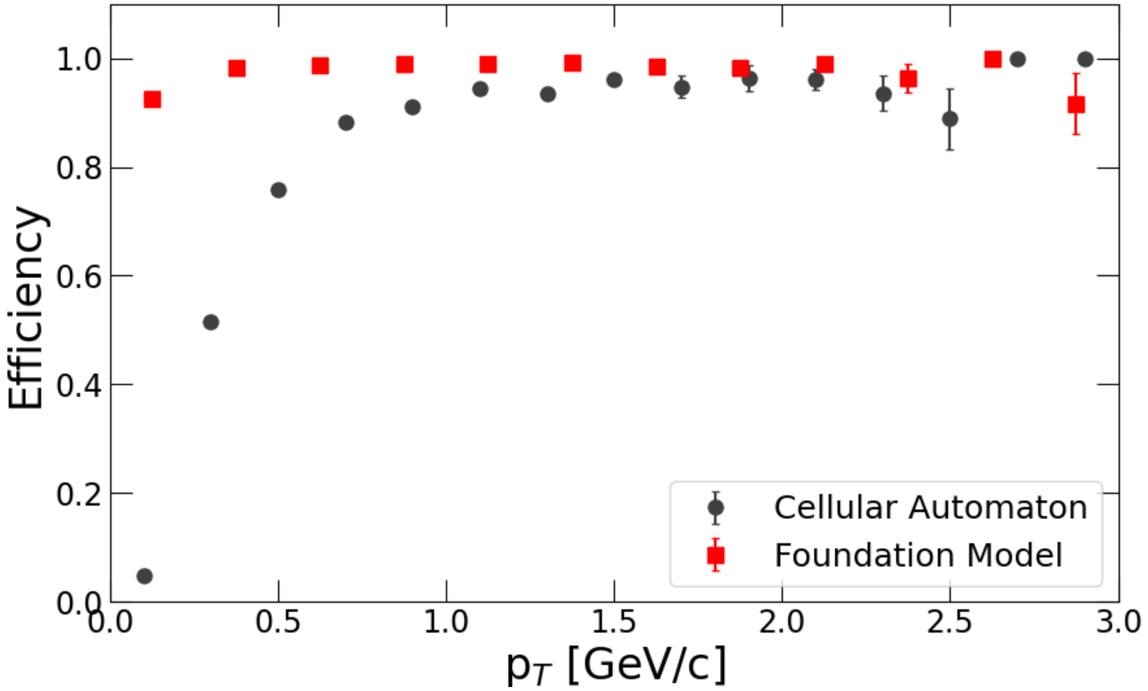
1. FM weights are frozen.
2. Lightweight Adaptor models are trainable on labeled data.
3. Evaluated on three downstream tasks: Track Finding, Particle Identification and Noise Tagging.

Preliminary Results on Tracking (WIP)

✅ High tracking efficiency across Pt.

Tracking efficiency (adapted from TrackML [1]). Tracks are uniquely matched to particles by the double majority rule:

- For a given track, the matching particle is the one where the absolute majority (strictly more than 50%) of the track points belong.
- Track should have the absolute majority of the points of the matching particle.



Typically, particle physicists focus on high-momentum tracks with filtering. Here, there is no filtering for the FM result, while CAseeder require primaries with 20 spacepoints that are in acceptance

[1] Calafiura, P. "TrackML: A High Energy Physics Particle Tracking Challenge, in the proceedings of the 14th International Conference on e-Science, Amsterdam, Netherlands."

Main Results

1. Our FM4NPP approach outperforms all comparative models on all three downstream tasks.
2. We confirm the performance gain is from FM pre-training by comparing with the “AdapterOnly” model.

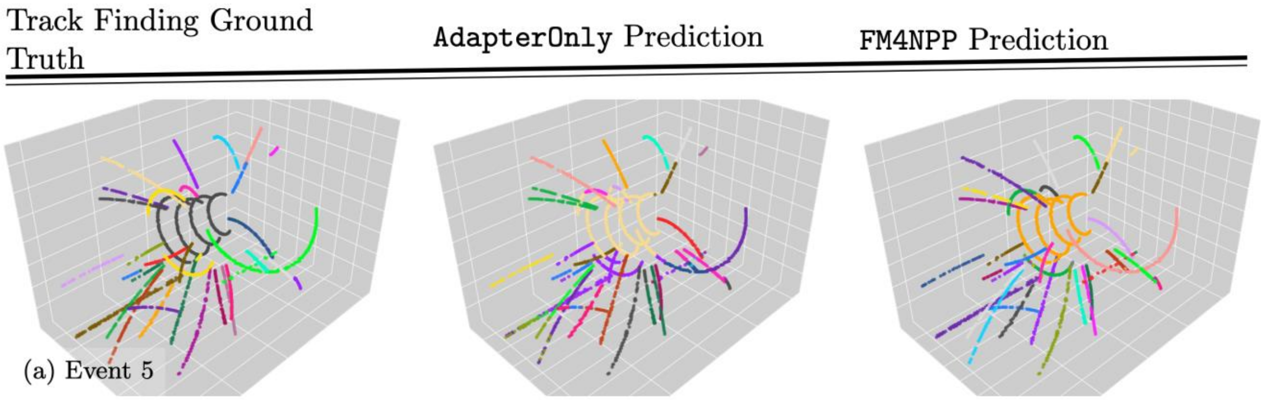


Figure 1. Example Output for Track Finding

model	#trnbl para.	Track Finding ARI↑	Track Finding efficiency↑	Track Finding 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%**

Will the representation learned from FM useful?
YES ✅

model	#trnbl para.	Particle Identification acc.↑	Particle Identification recall↑	Particle Identification pre.↑	Noise Tagging acc.↑	Noise Tagging recall↑	Noise Tagging 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**

Brookhaven National Laboratory

More Efficient and Better Models

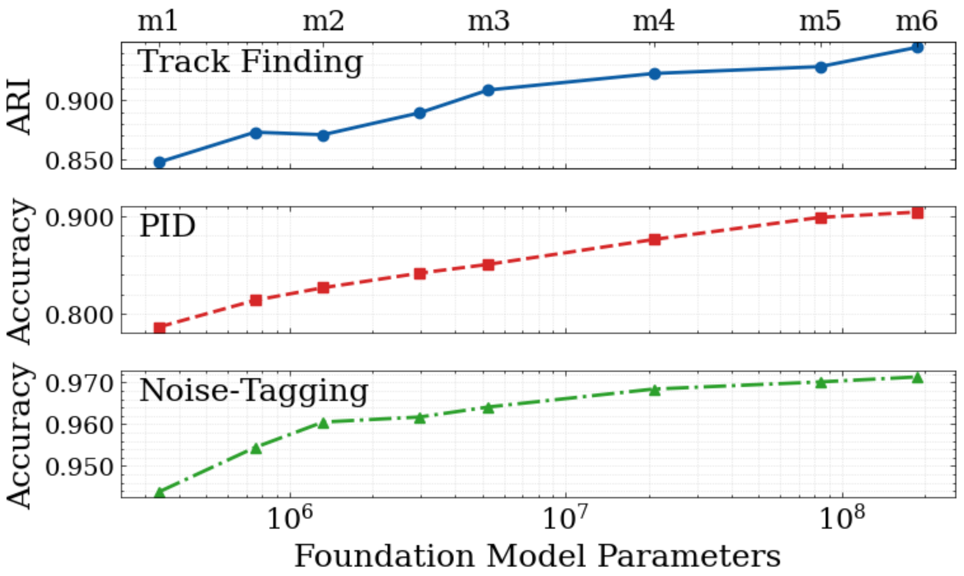


Figure 1. Larger Models lead to better Accuracy.

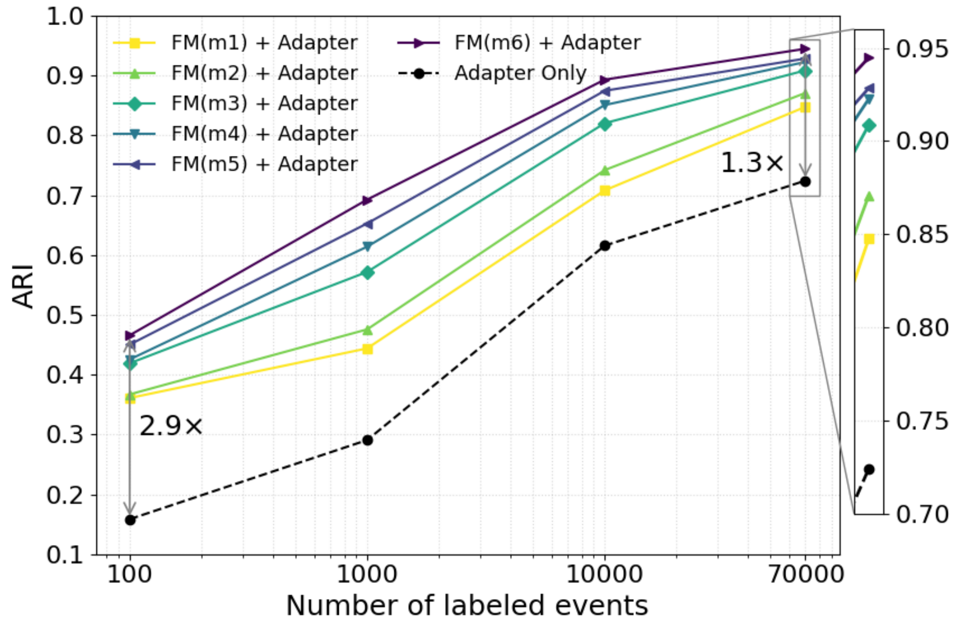
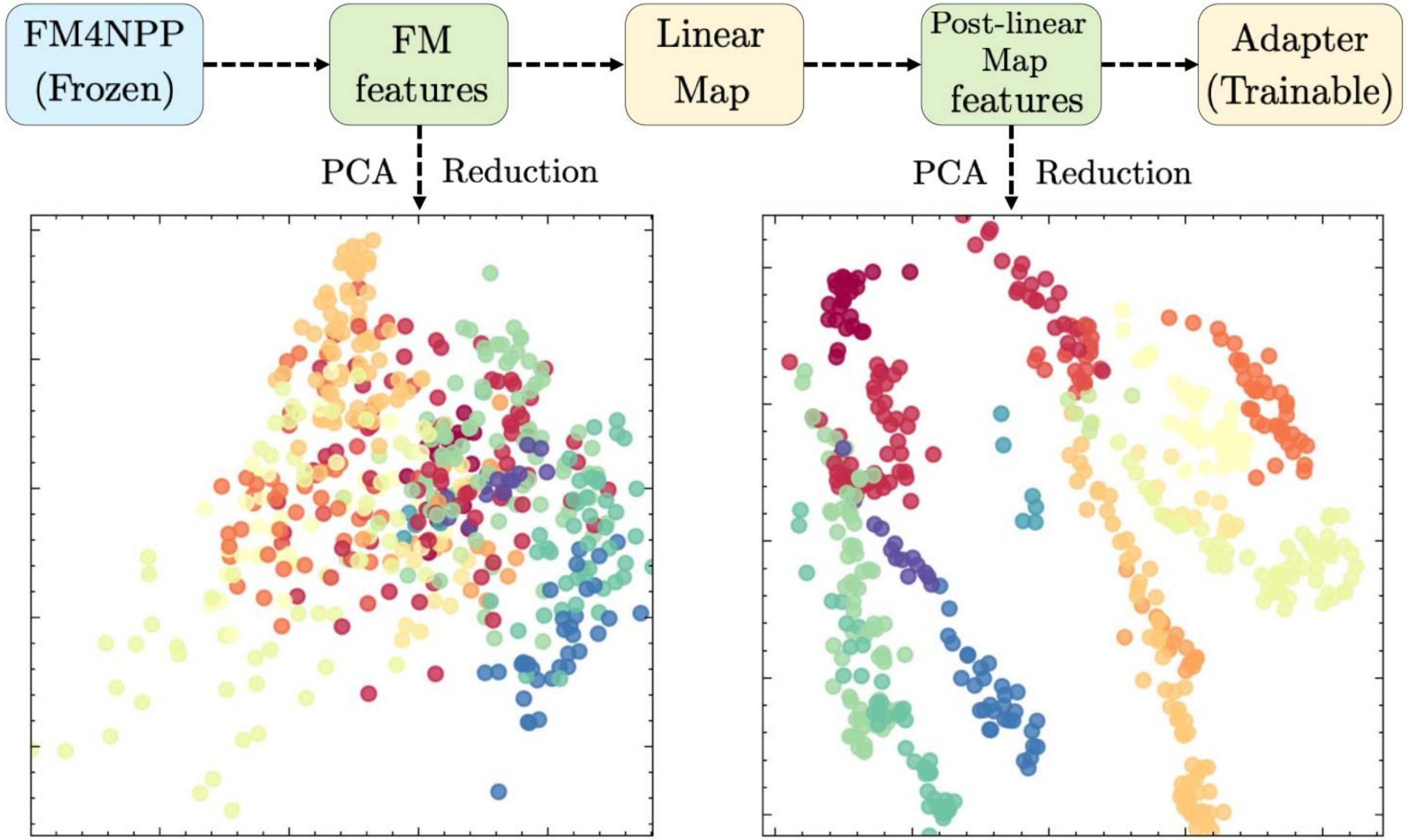


Figure 2. Larger Models offer better Data Efficiency.

Will larger FM also lead to better downstream tasks?
YES ✅

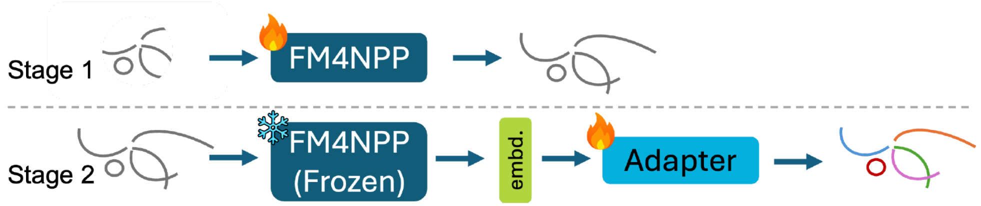
Will adapting from FM be more data efficient?
YES ✅

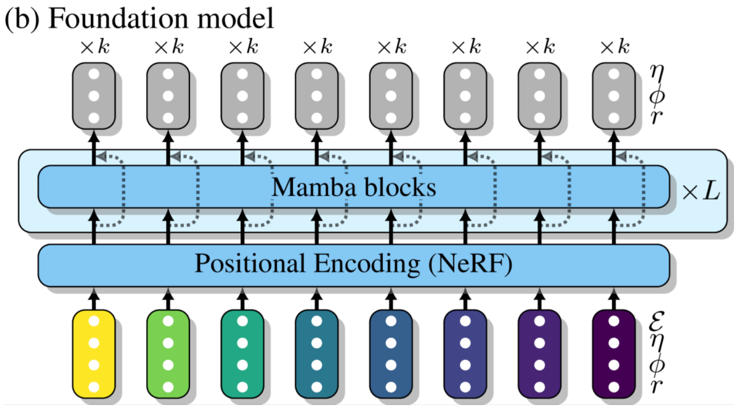
FM Features are Task-Agnostic But, Task-relevancy is one linear map away!



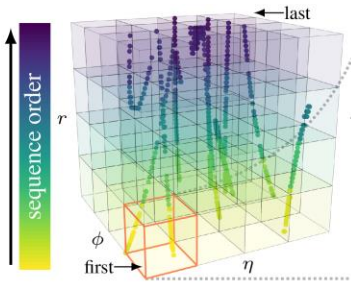
Conclusions

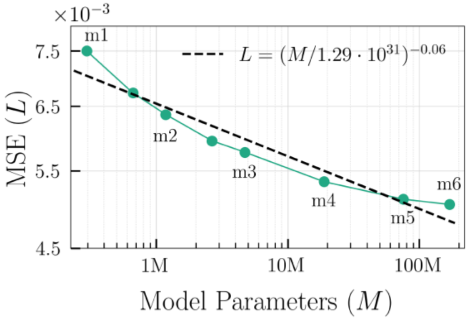
1. We have demonstrated a scalable and adaptable FM4NPP approach that can leverage big data and big computation.
2. The FM4NPP achieves a new state of the art on three downstream tasks.





- ✅ What is the right self-supervised learning task?
- ✅ Will the FM pre-training scale?
- ✅ Will the representation learned from FM useful?
- ✅ Will larger FM also lead to better downstream tasks?
- ✅ Will adapting from FM be more data efficient?



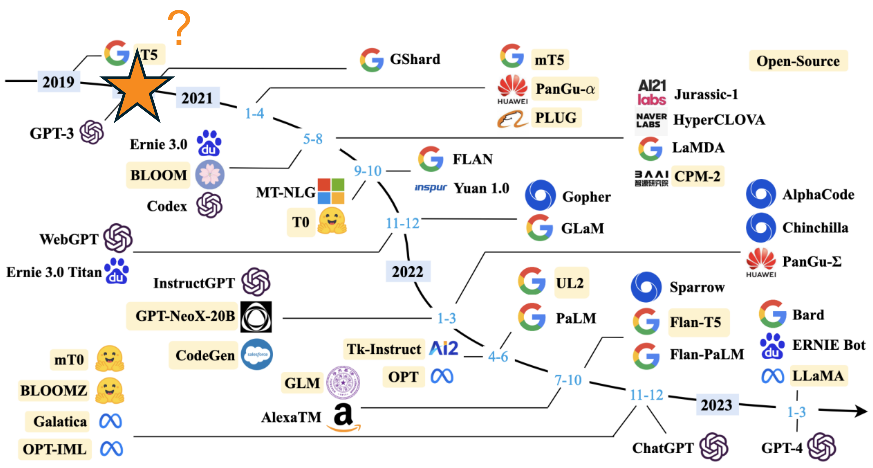


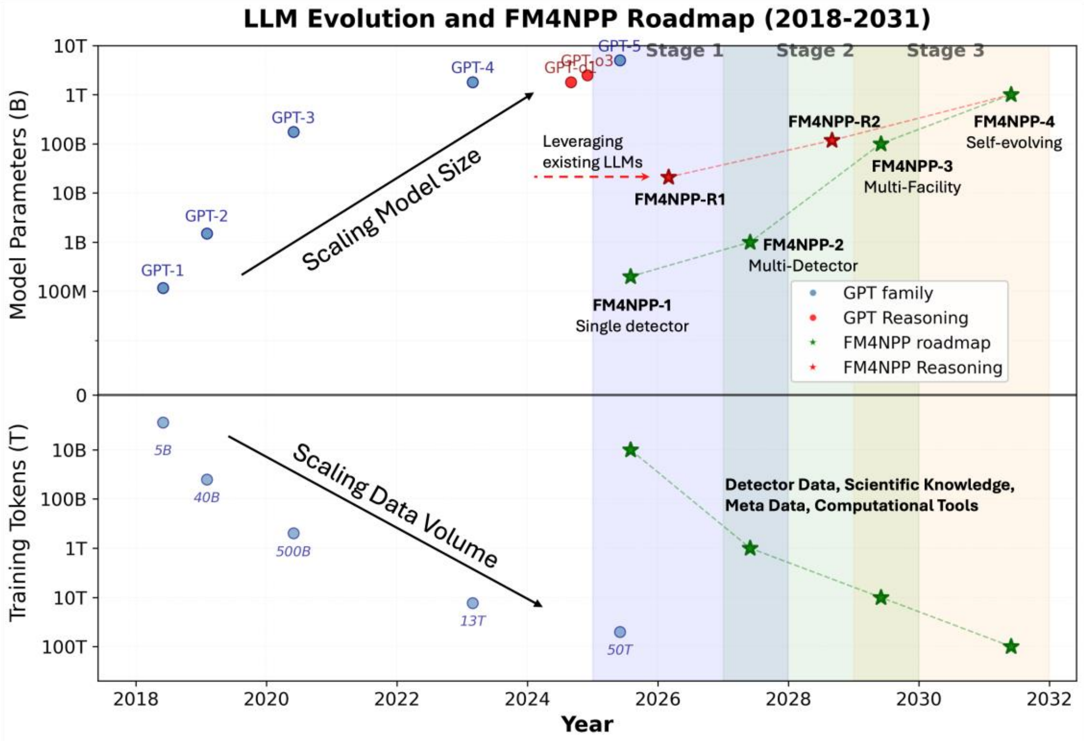
Brookhaven National Laboratory

[1] Park, David K. and Li, Shuhang and Huang, Yi and Luo, Xihaier and Yu, Haiwang and Go, Yeonju and Pinkenburg, Christopher and Lin, Yuewei and Yoo, Shinjae and Osborn, Joseph D. and Huang, Jin and Ren, Yihui, *FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics*. Available at arxiv: https://arxiv.org/abs/2508.14087

Future Work

1. Scaling to bigger models, incorporating more data, and validating on other tasks.
2. Exploring other architectures and self-supervised learning tasks.
3. Expanding to other experiments in NP and HEP (ATLAS, CMS, Bella II, etc.)
4. Incorporating other modalities: detector submodules, simulation, meta data, etc.





Brookhaven National Laboratory

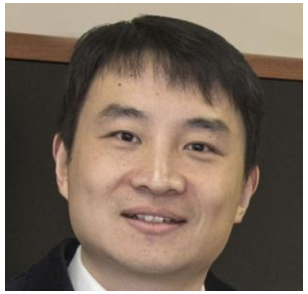
Acknowledgement

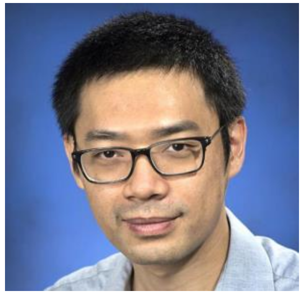
- This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Brookhaven National Laboratory, LDRD 25-045, which is operated and managed for the U.S. Department of Energy (DOE) Office of Science by Brookhaven Science Associates under contract No. DE-SC0012704.
- Shuhang Li was partially supported by the DOE Office of Science through the Office of Nuclear Physics under Award No.~DE-FG02-86ER40281.
- Yihui Ren, Xihaier Luo and Shinjae Yoo were partially supported by the DOE Office of Science through the Office of Advanced Scientific Computing Research and the Scientific Discovery through Advanced Computing (SciDAC) program.
- This research also utilized resources of the National Energy Research Scientific Computing Center (NERSC) under the ``GenAI@NERSC'' program. NERSC is a DOE Office of Science User Facility with Award No. DDR-ERCAP0034059. The authors are grateful to the NERSC staff for their support, particularly Shashank Subramanian and Wahid Bhimji.

Thank you!

The Passionate Team.



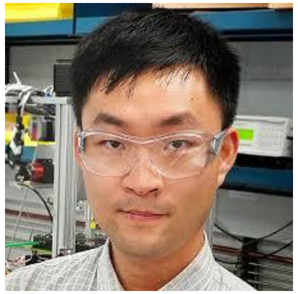





(AI Dept.) David Park, Yi Huang, Xihaier Luo, Yuewei Lin, Shinjae Yoo, Yihui “Ray” Ren







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