FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics

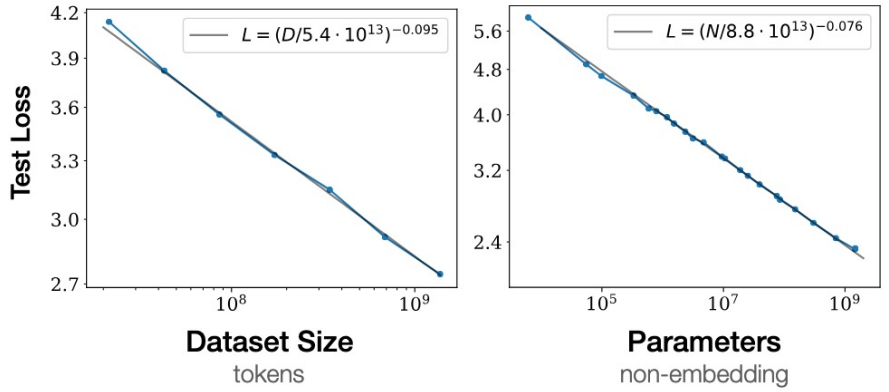
10/28/2025

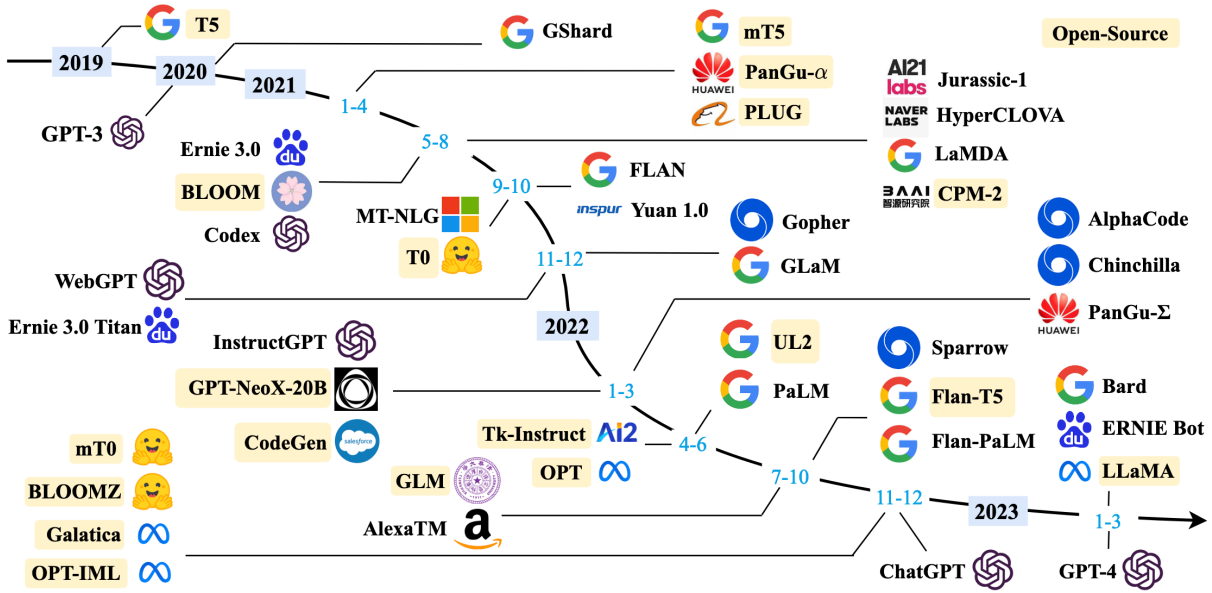
Shuhang Li

Team: Go, Yeonju; Huang, Jin; Huang, Yi; Li, Shuhang; Lin, Yuewei; Luo, Xihaier; Osborn, Joseph; Park, David; Ren, Yihui (Ray); Yoo, Shinjae; Yu, Haiwang

Scaling Large Language Models (LLM)

- Attention mechanism (2017), the simple building block (transformer model).
- Self-supervised auto-regressive pre-training (not reliant on labeled data).
- Pre-trained model can be used for multiple downstream tasks.
- (2020) Neural Scaling behavior [1].
- (2020-23) 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).

Foundation Model

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

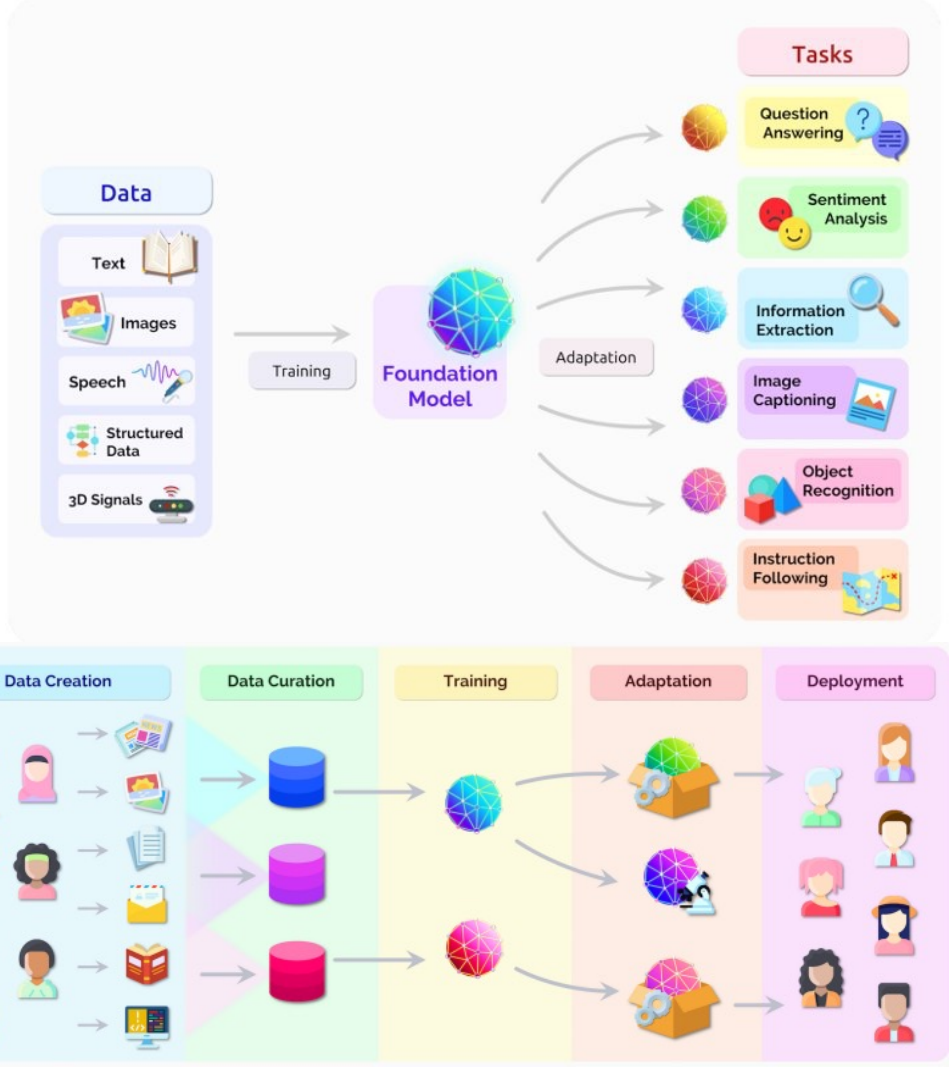


Image credit [1]

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

Scientific Motivation

Proof of concept for an FM4NPP:

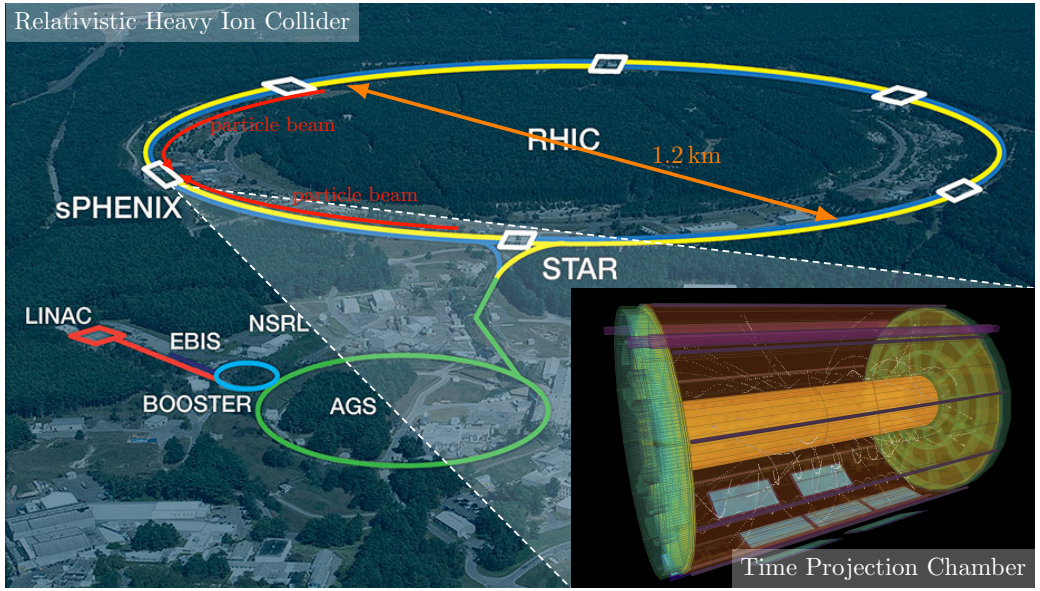
1. Neural scaling behavior
2. Generalizable to downstream tasks

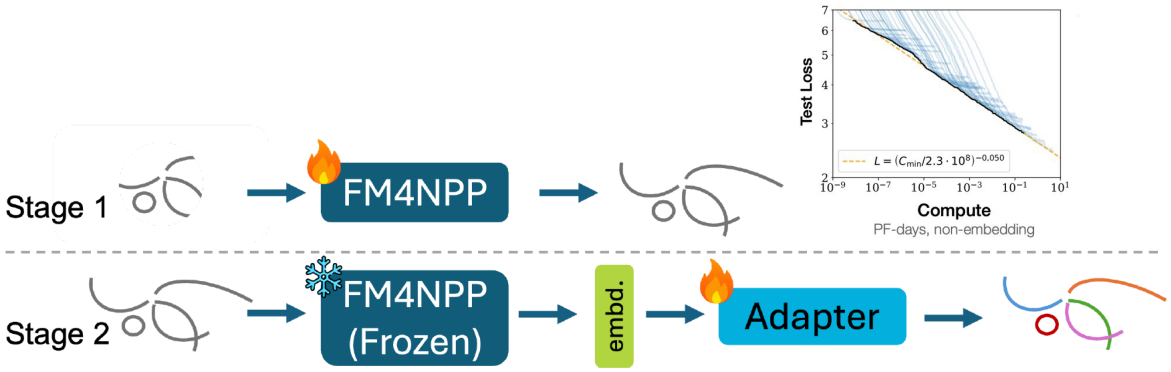
Two-stage approach:

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

Key Questions:

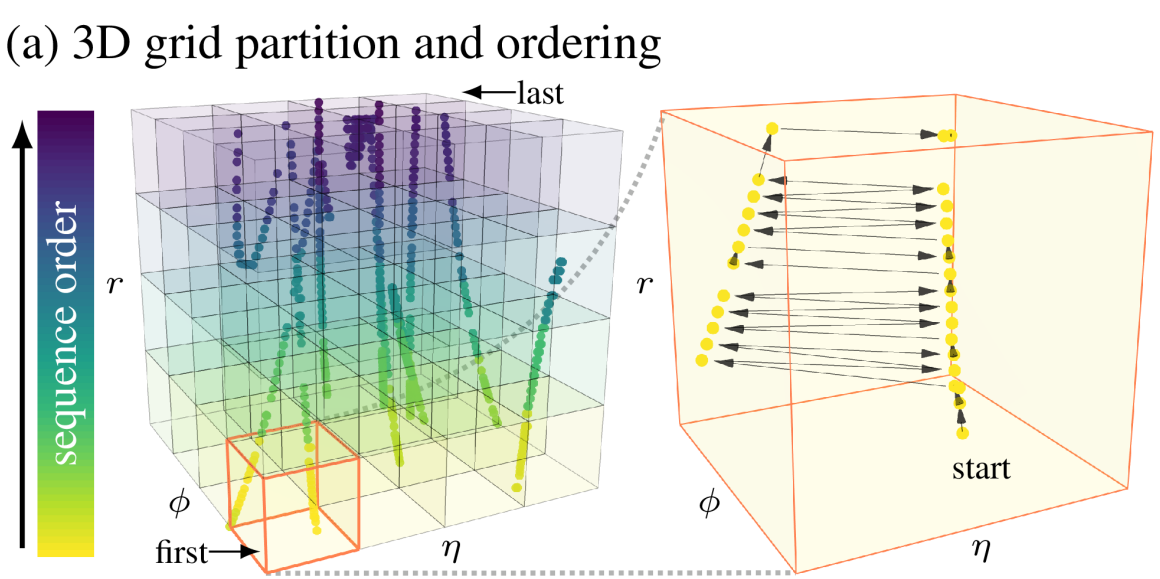
1. What is the right self-supervised learning task?
2. Will the FM pre-training scale?
3. Will the representation learned from FM useful?
4. Will larger FM also lead to better downstream tasks?
5. Will adapting from FM be more data efficient?

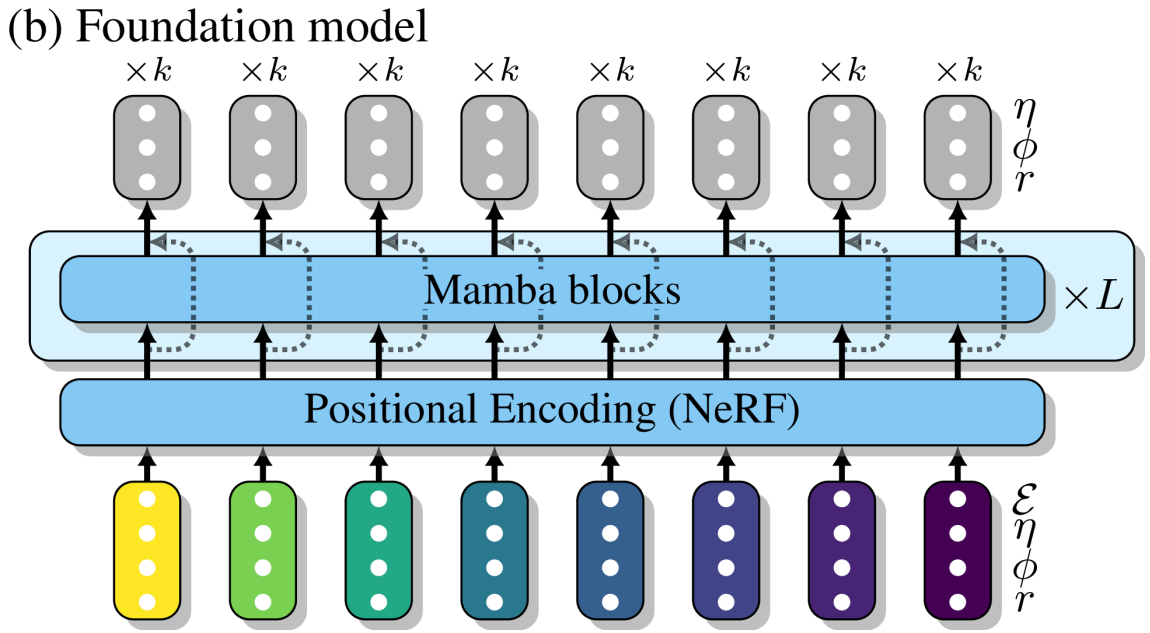




Pre-Training

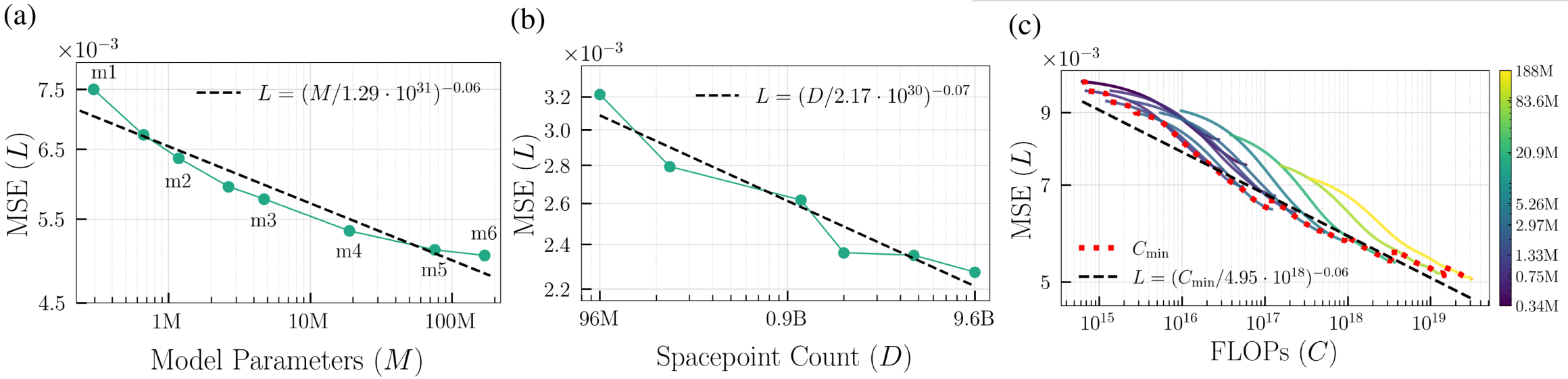
- Serialization: **Hierarchical Raster Scan**
- Self-supervised learning: **Next k-nearest-neighbor Prediction**
- Adaptation to Mamba Model and large-scale training (e.g., μTransfer)





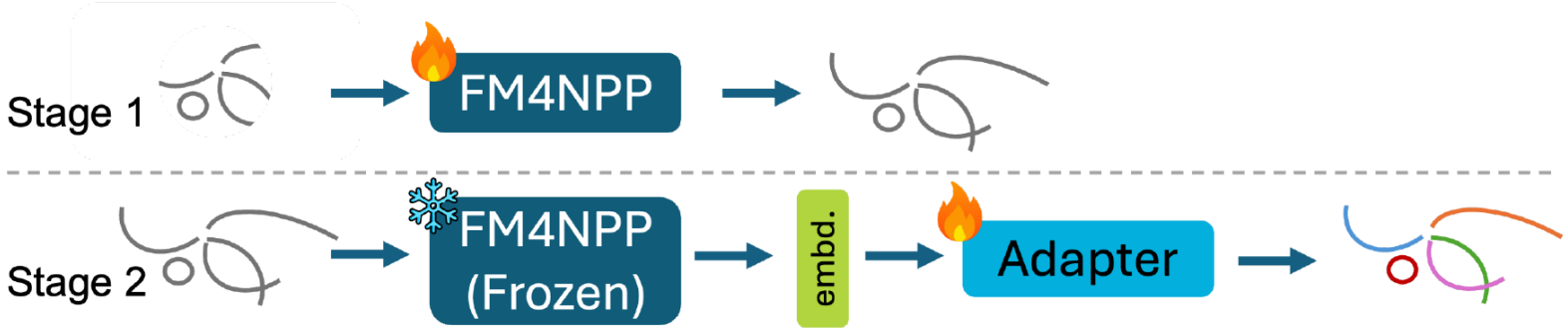
Neural Scaling Behavior

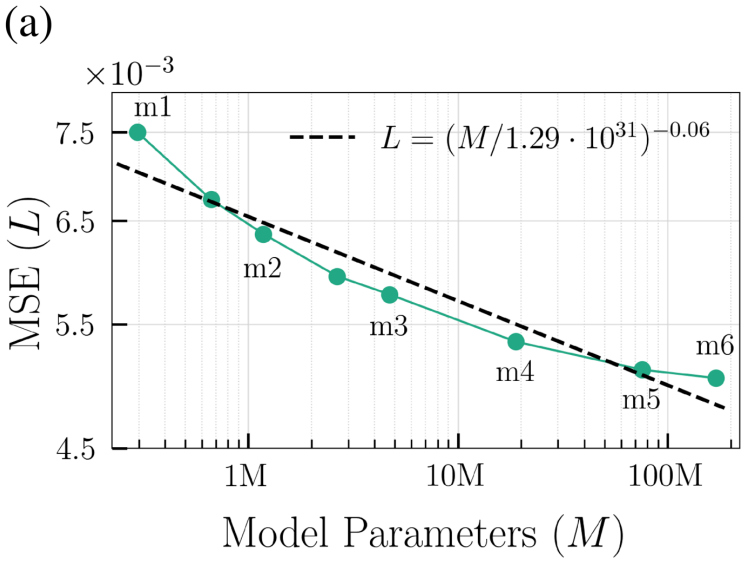
	Model Sizes					
	m1	m2	m3	m4	m5	m6
Model Width	64	128	256	512	1024	1536
Model Params	0.34M	1.3M	5.3M	21M	84M	188M



- 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).

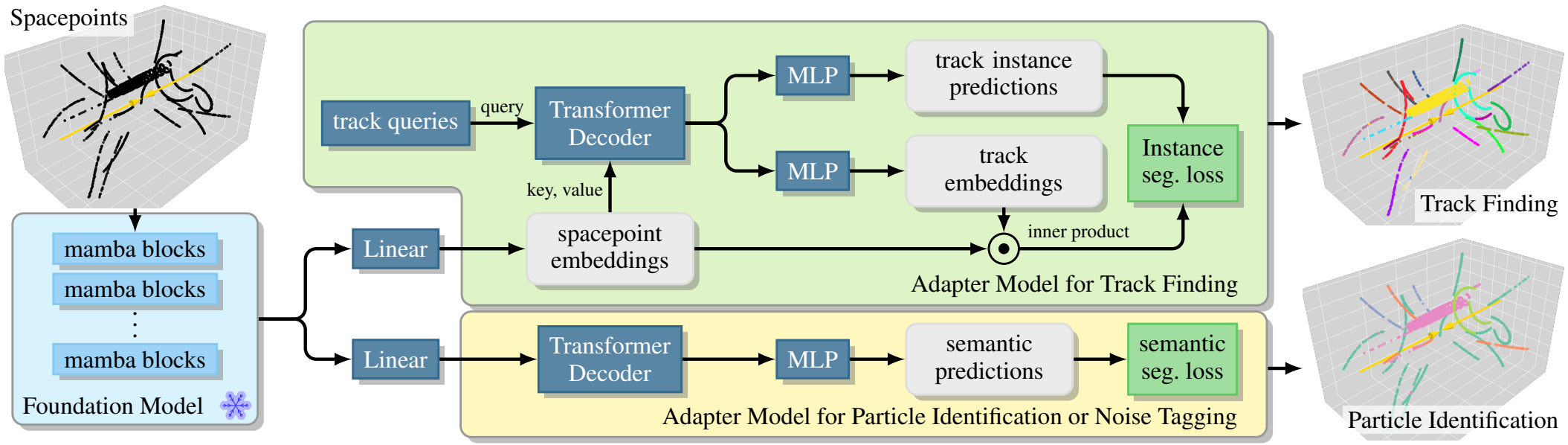
Will the FM Features be Useful?





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?

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.

Main Results

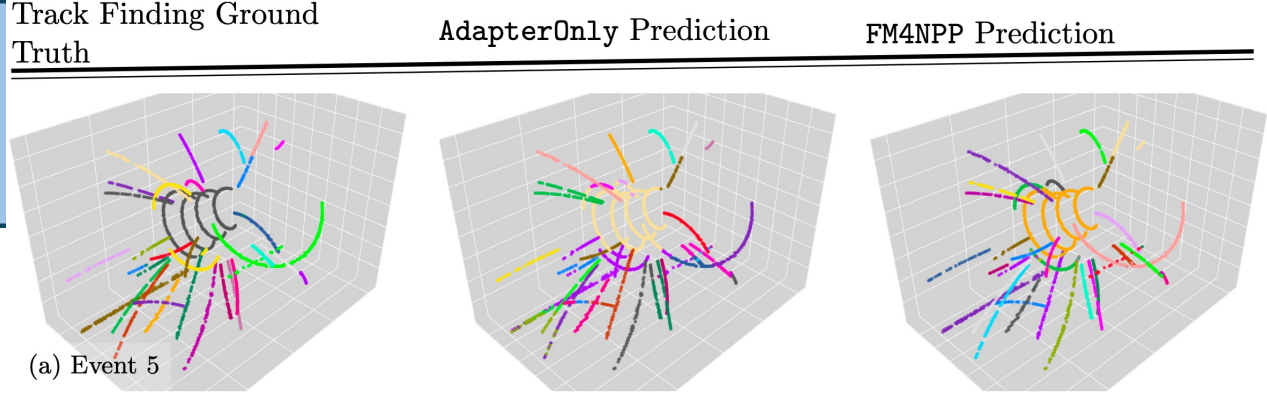


Figure 1. Example Output for Track Finding

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.

Will the representation learned from FM useful?

YES ✅

		Track Finding		
model	#trnbl para.	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%**

		Particle Identification			Noise Tagging		
model	#trnbl para.	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**

More Efficient and Better Models

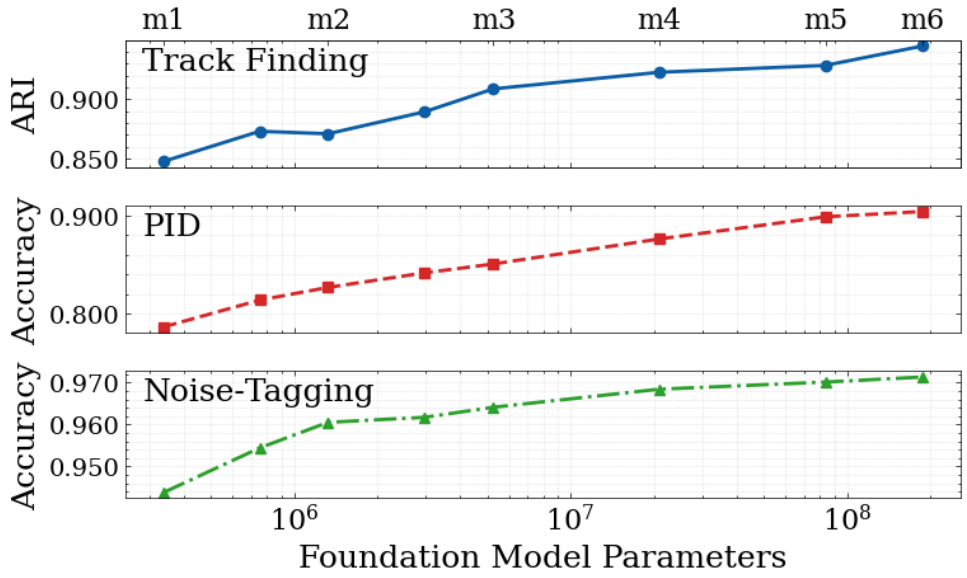


Figure 1. Larger Models lead to better performance

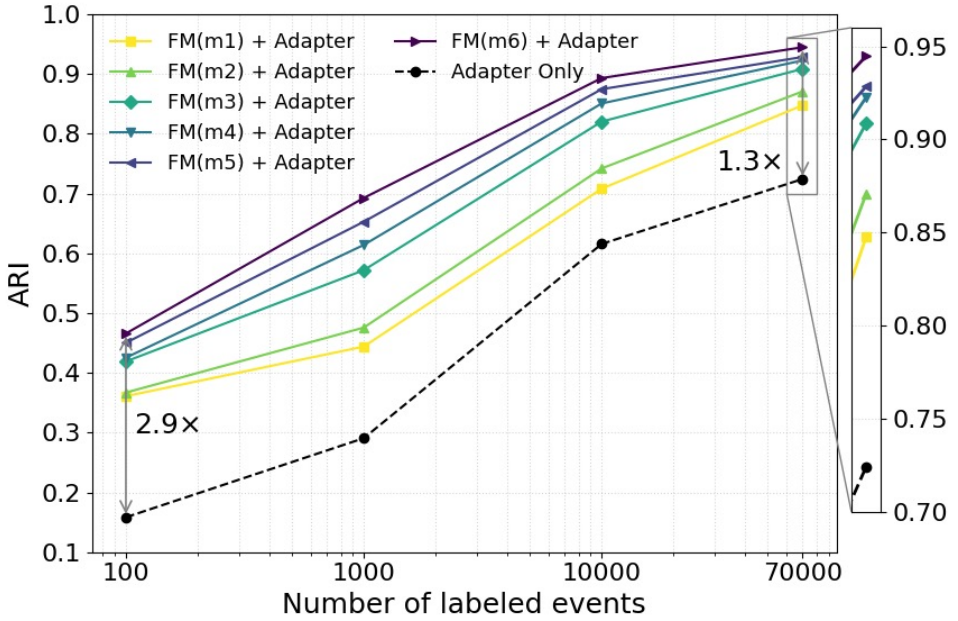


Figure 2. Larger Models offer better Data Efficiency.

Will larger FM also lead to better downstream tasks?

YES ✅

Will adapting from FM be more data efficient?

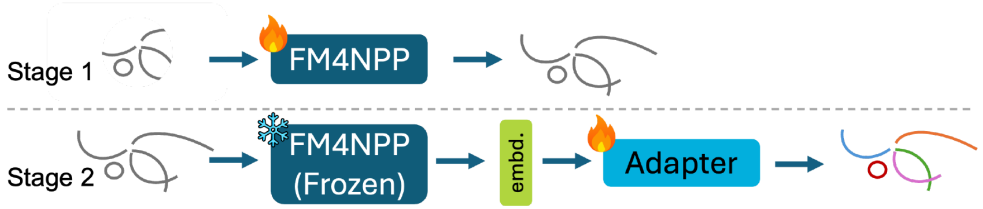
YES ✅

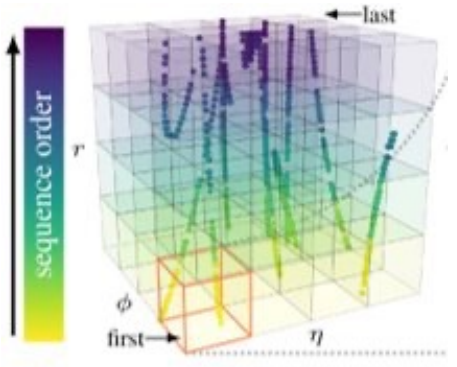
Conclusions and Future Work

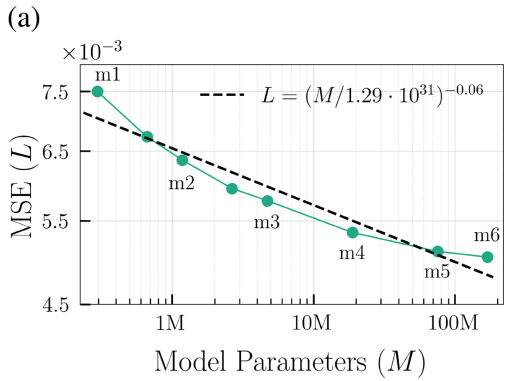
1. We have demonstrated a scalable and adaptable FM4NPP approach that can leverage big data and big computation.
2. The FM4NPP achieves a new state of the art on three downstream tasks.

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, etc.)
4. Incorporating other modalities: detector submodules, simulation, meta data, etc.







- ✅ **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?**

[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 SSRN: http://dx.doi.org/10.2139/ssrn.5389206

Backup slides

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