Brookhaven National Laboratory

U.S. DEPARTMENT of ENERGY

FM4NPP: Toward AI Foundation Models for Nuclear and Particle Physics

Yihui “Ray” Ren (yren@bnl.gov)
AI-CoDesign Group Leader, AI Department, Brookhaven National Laboratory

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

Annual RHIC & AGS Users' Meeting, BNL, NY, May 20-23, 2025

@BrookhavenLab

(Updated on June 1st, also presented at SCSP AI+Expo.)

An Extremely Brief History of NLP: From Handcrafted Rules to LLM

- 1980s: Symbolic methods and rule-based parsing.

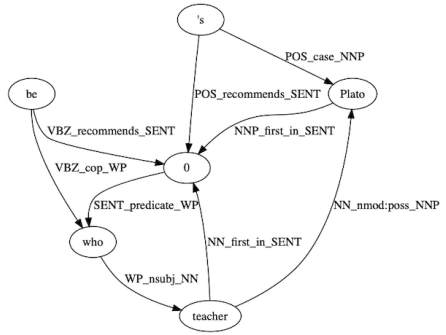


Image credit: Benny's blogpost Link

https://en.wikipedia.org/wiki/Natural_language_processing

An Extremely Brief History of NLP: From Handcrafted Rules to LLM

- 1980s: Symbolic methods and rule-based parsing.
- 2000s: Data-driven statistical methods. Growing data but weak models.

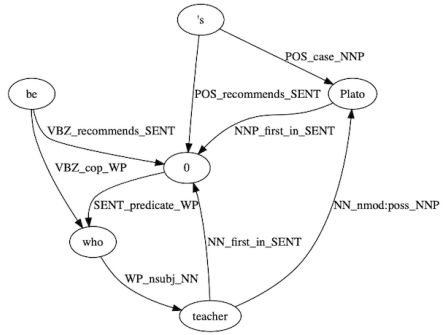


Image credit: Benny's blogpost Link

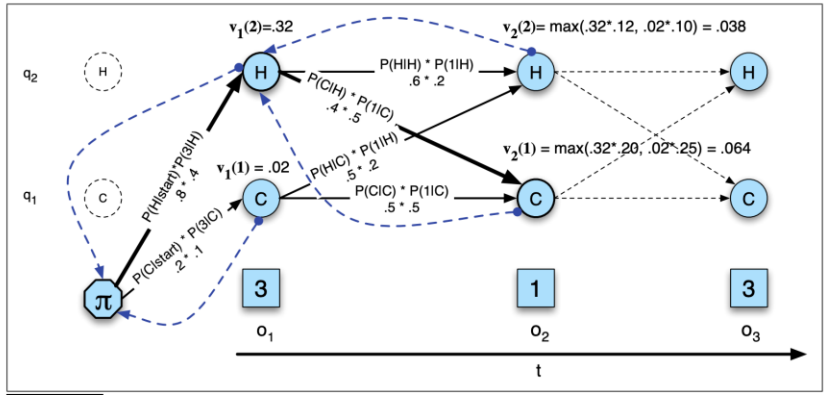


Image credit: Daniel Jurafsky & James H. Martin Link

https://en.wikipedia.org/wiki/Natural_language_processing

An Extremely Brief History of NLP: From Handcrafted Rules to LLM

- 1980s: Symbolic methods and rule-based parsing.
- 2000s: Data-driven statistical methods. Growing data but weak models.
- 2020s: LLM. Model capacity outpaces data for the first time.

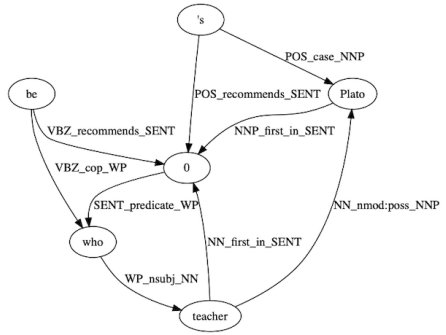


Image credit: Benny's blogpost Link

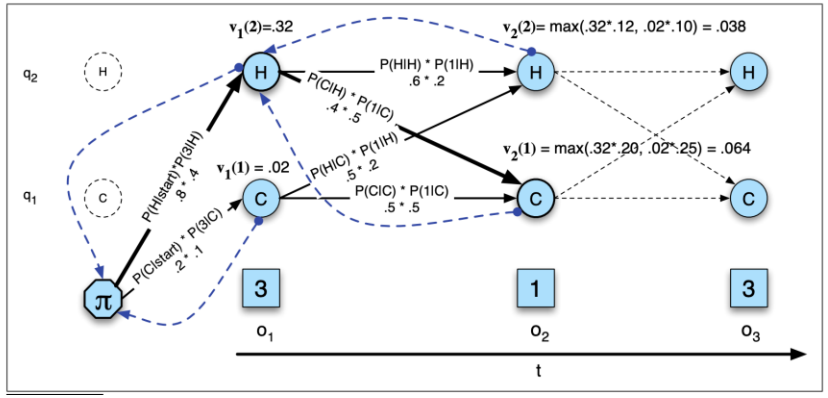


Image credit: Daniel Jurafsky & James H. Martin Link

https://en.wikipedia.org/wiki/Natural_language_processing

An Extremely Brief History of NLP: From Handcrafted Rules to LLM

- 1980s: Symbolic methods and rule-based parsing.
- 2000s: Data-driven statistical methods. Growing data but weak models.
- 2020s: LLM. Model capacity outpaces data for the first time.

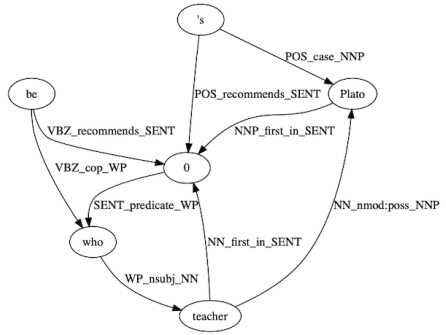


Image credit: Benny's blogpost Link

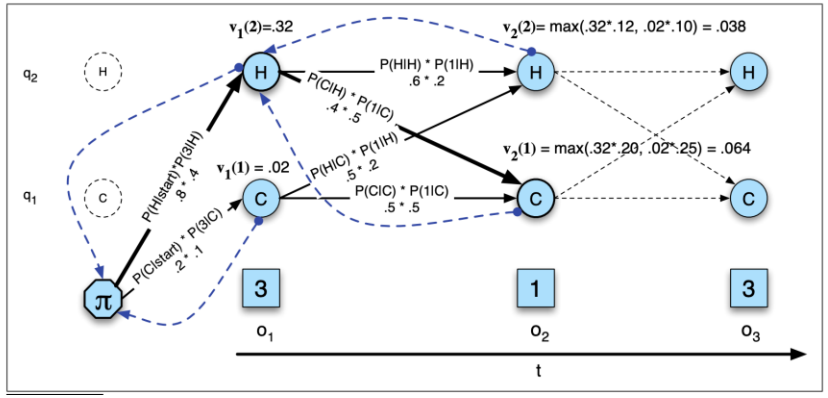


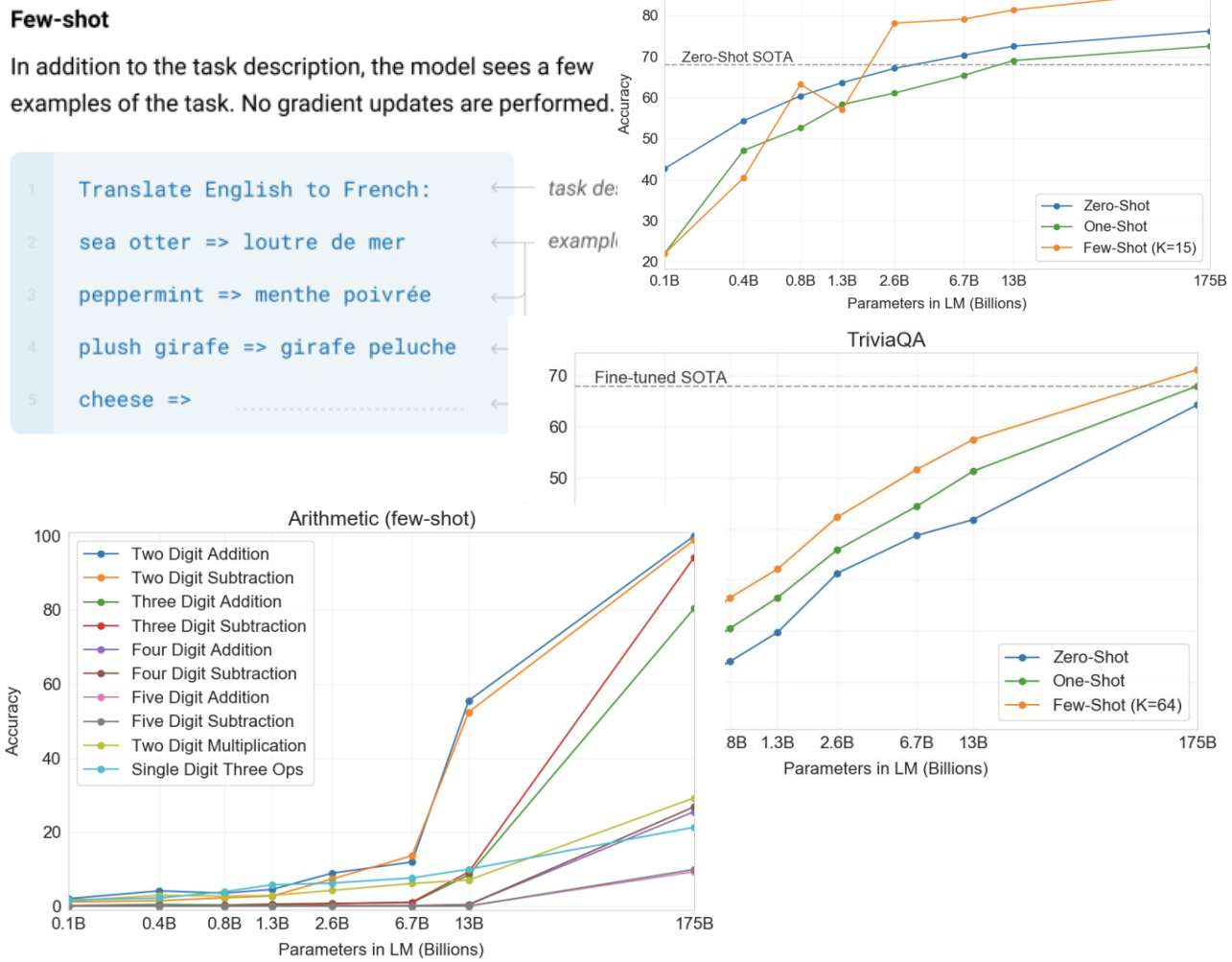
Image credit: Daniel Jurafsky & James H. Martin Link

Today, many science domains use data-driven AI methods.

Next step: specialized AI→ FM

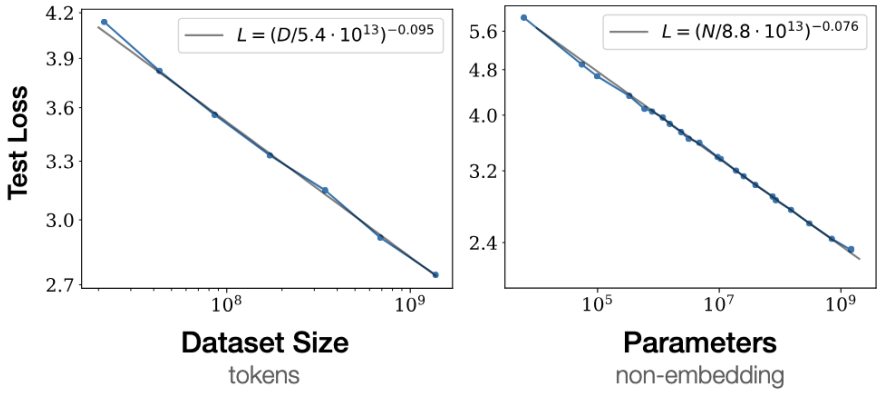
Main Features of LLM

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



Neural Scaling Behavior

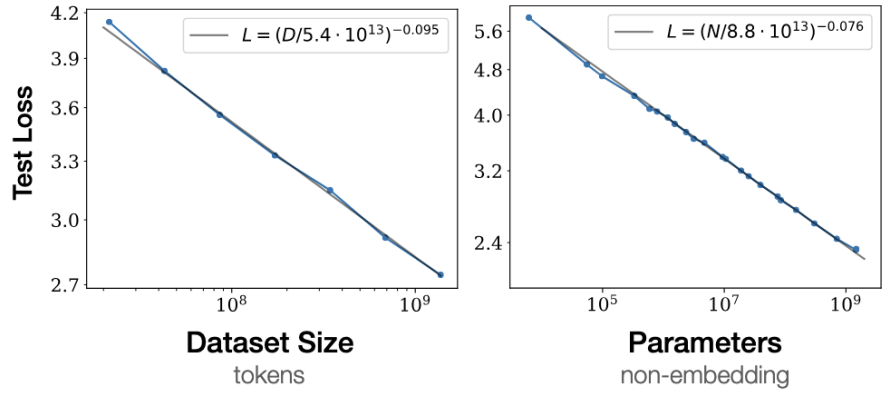
(2020) Neural Scaling Laws [1]



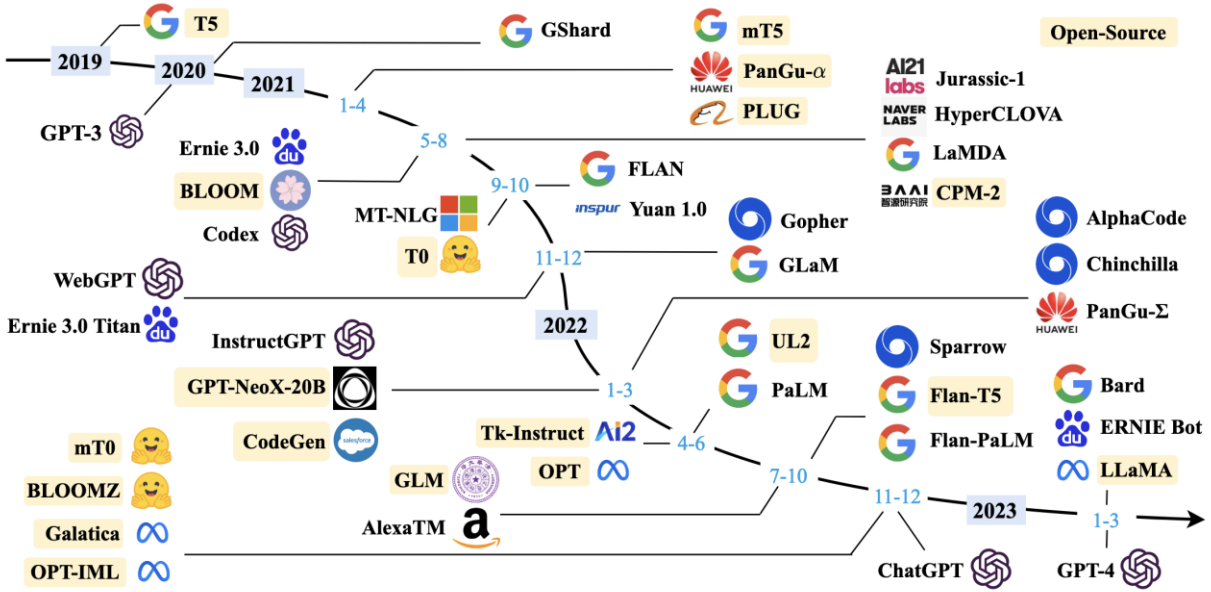
[1] Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv:2001.08361* (2020).

Neural Scaling Behavior

(2020) Neural Scaling Laws [1]



(20-23) LLM “Arms Race”



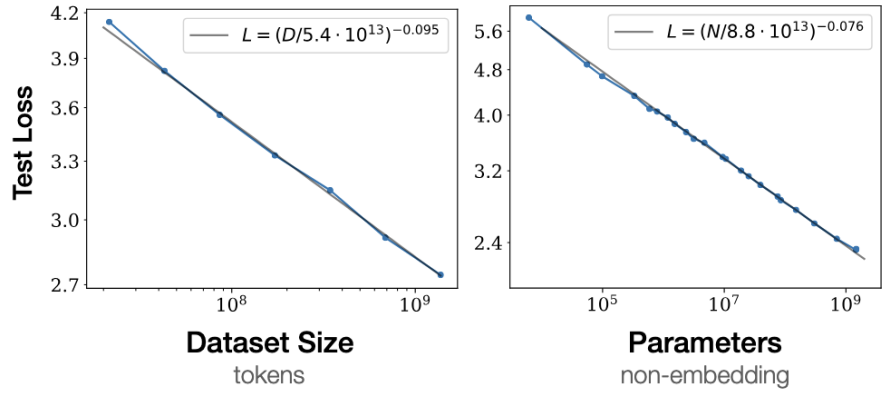
[1] Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv:2001.08361* (2020).

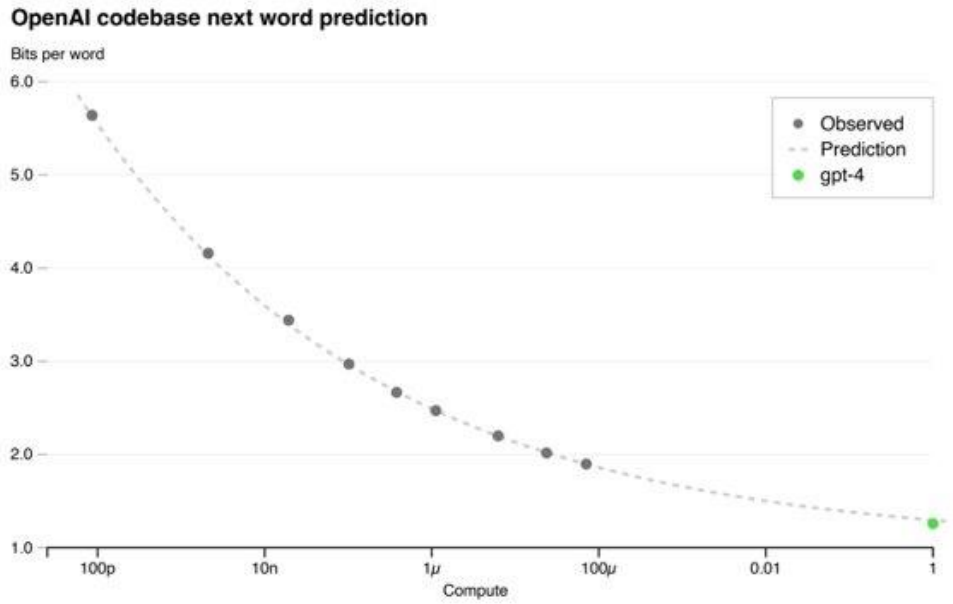
Neural Scaling Behavior

(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).

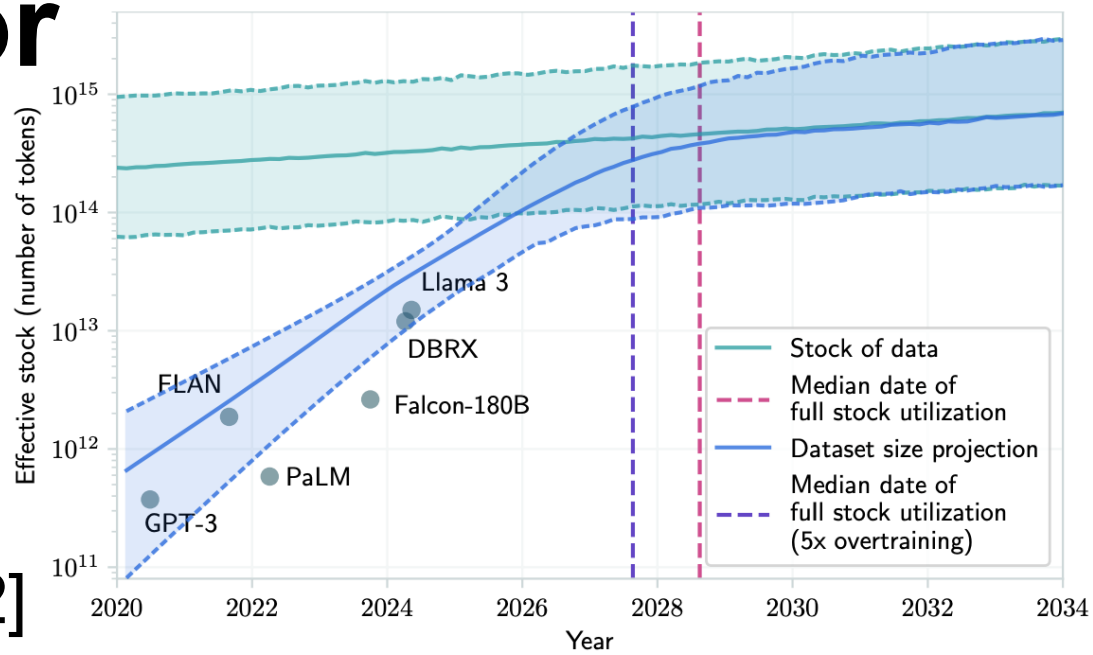
Neural Scaling Behavior

(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

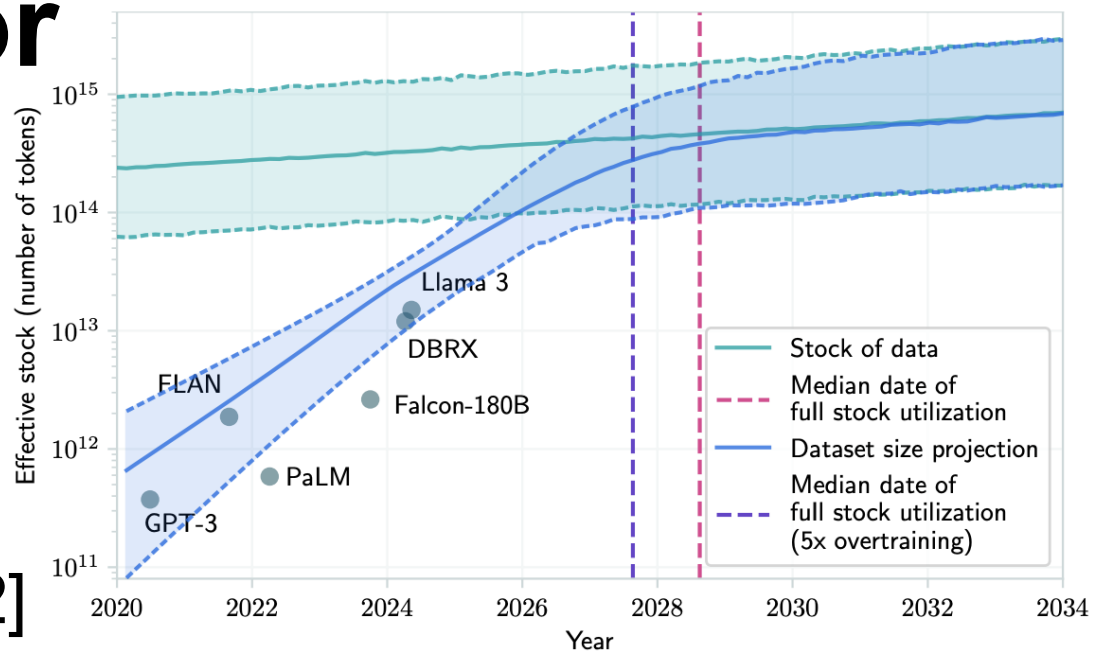
(2020) Neural Scaling Laws [1]

(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

Scientific Foundation Models

- Large amount of data (unlabeled).
- Self-supervised learning on surrogate task.
- Adaptable to multiple downstream tasks.
- Scaling behavior, bigger model, and more data lead to better performance.

Can we build one for NPP?

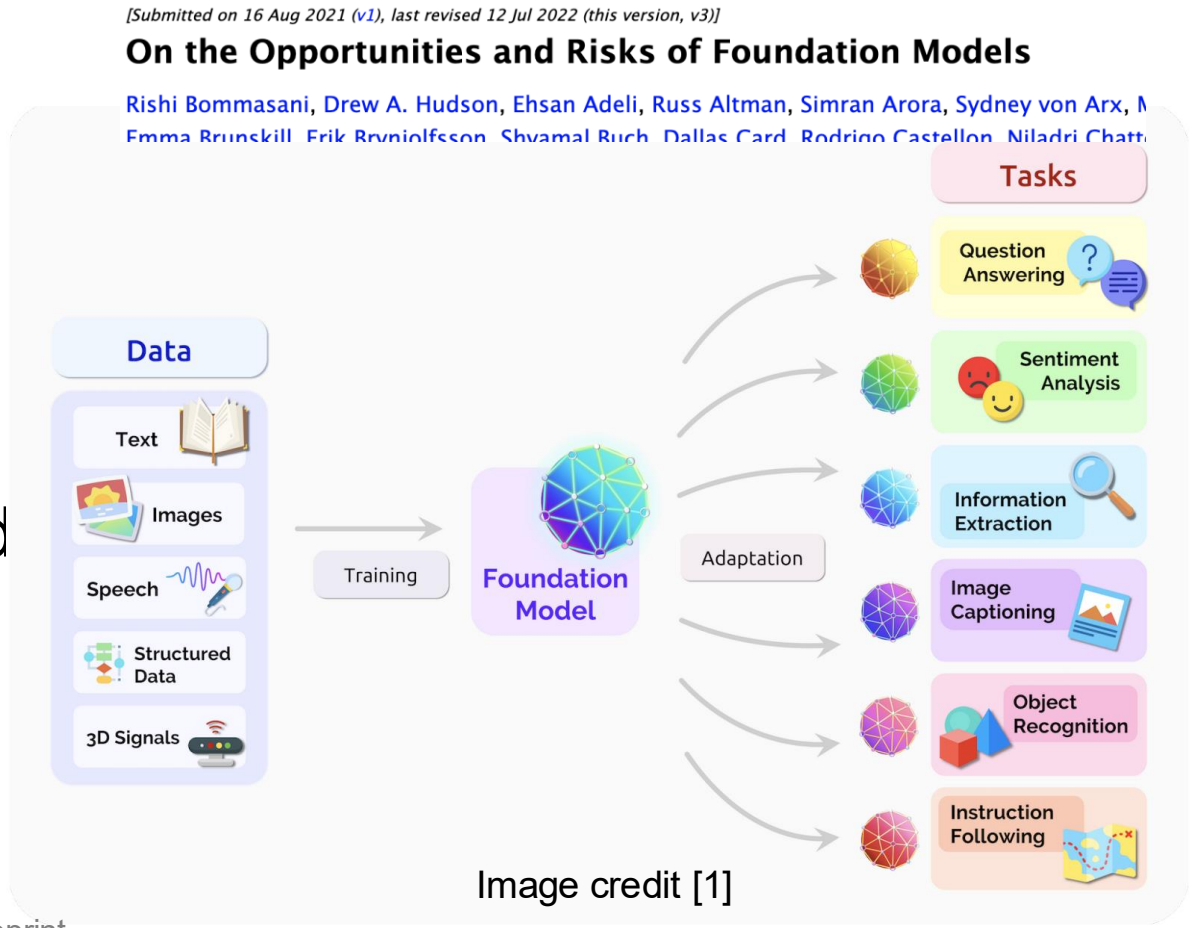
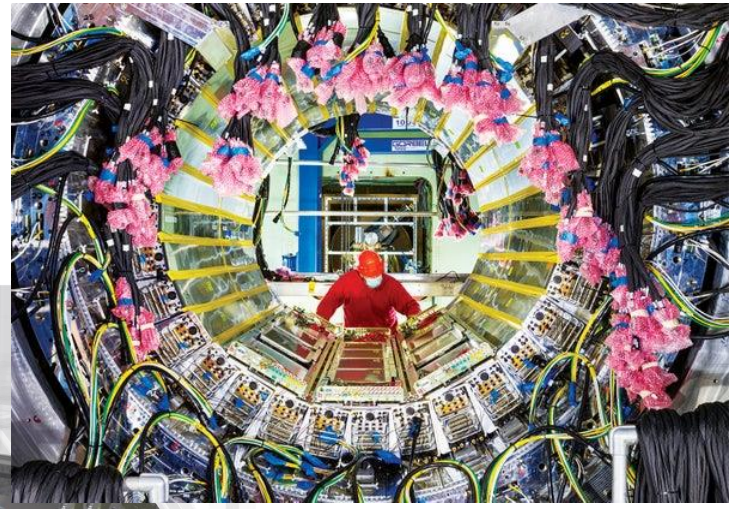


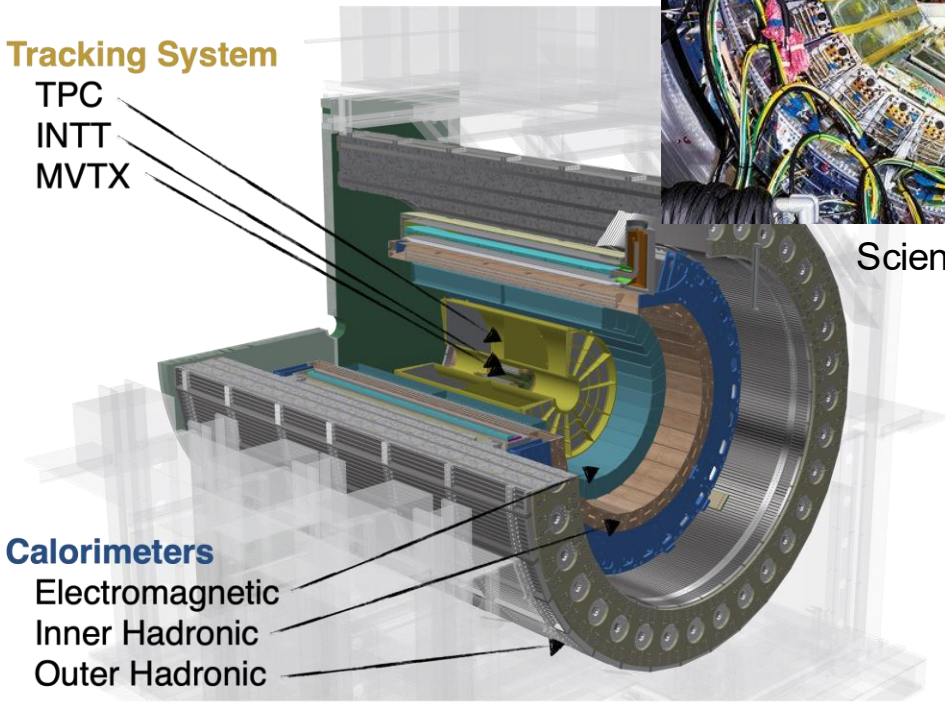
Image credit [1]

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

sPHENIX at BNL



Scientific American, 03/01/2023

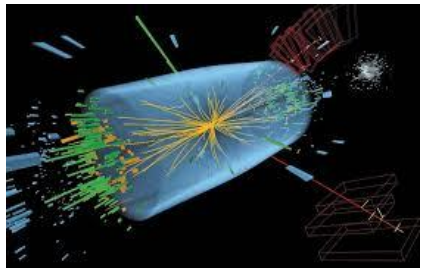


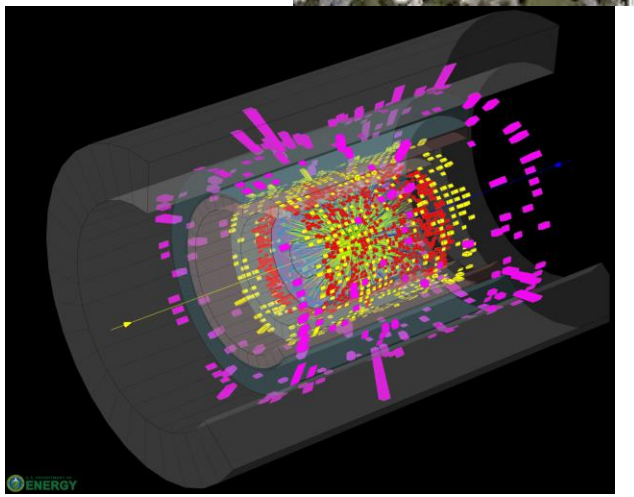
The largest particle collider in U.S.
Data taking began in 2023!
High-precision tracking system + Hermetic Electromagnetic & Hadronic calorimeters

A Perfect Fit!

- Broad impact: RHIC/EIC, LHC, future circular collider (FCC), etc.

LHC





RHIC/EIC

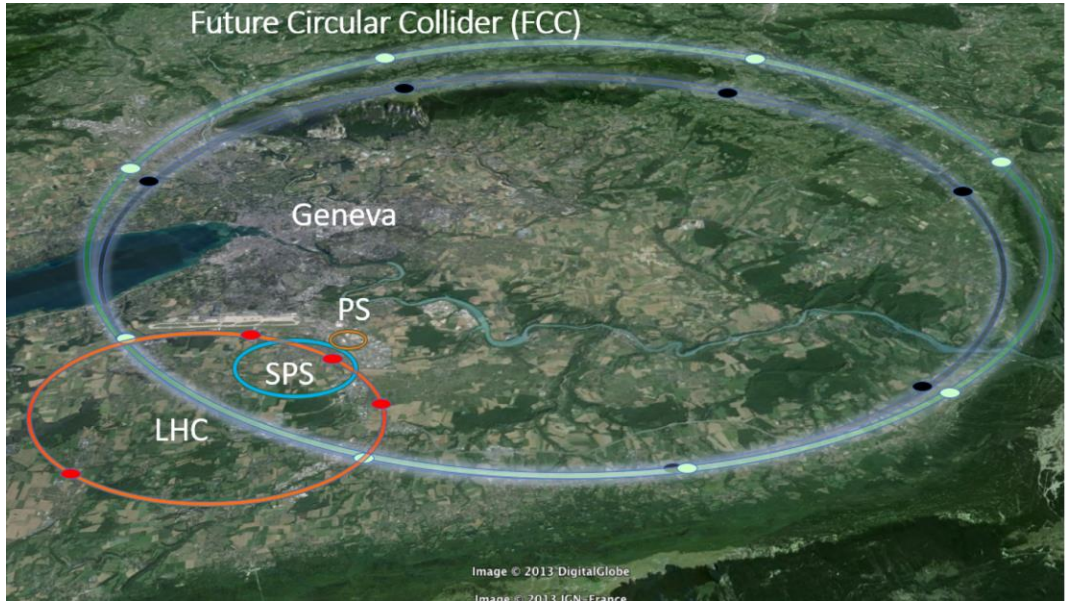
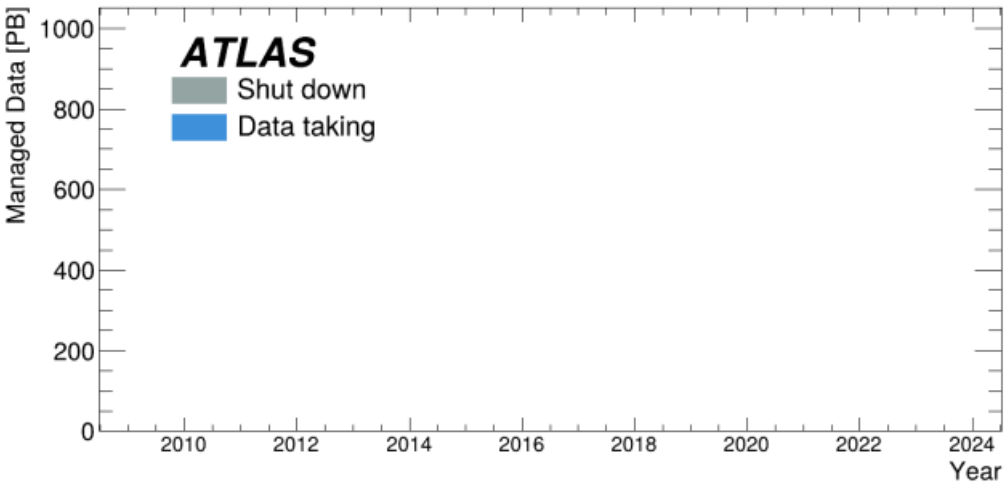


Image Credit: ICMAB

A Perfect Fit!

- Broad impact: RHIC/EIC, LHC, future circular collider (FCC), etc.
- A large amount of data, both simulations and experiments (e.g.,10M pp collision and 1M AuAu collision ~1 TB).
- Well-defined downstream tasks and well-established ways to measure the performance (e.g., particle tracking).
- A passionate team.

LHC

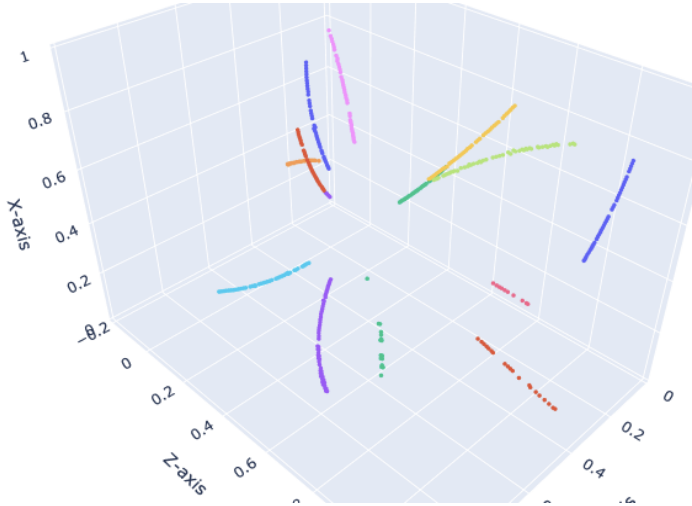


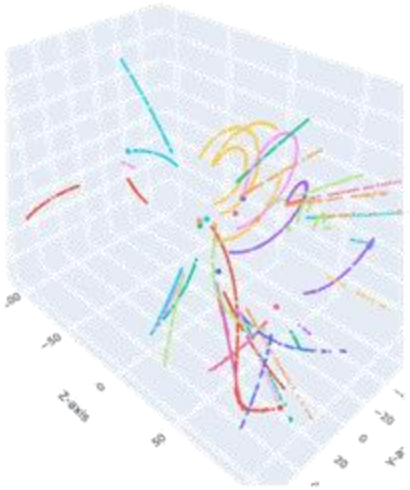
2024/07/10: ATLAS data-taking milestone: 1EB of data

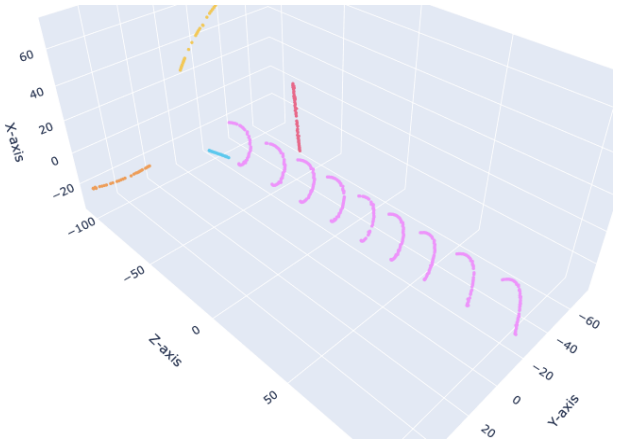
…Yet Quite Challenging

- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
- What neural network architecture to use?
- Will performance scale with the size of the network and data?
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…

Some samples of proton-proton collision







…Yet Quite Challenging

- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
  - Autoregressive, just like GPT
  - Mask reconstruction
  - Artificial denoising
  - Zigzag puzzles
- What neural network architecture to use?
- Will performance scale with the size of the network and data?
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…

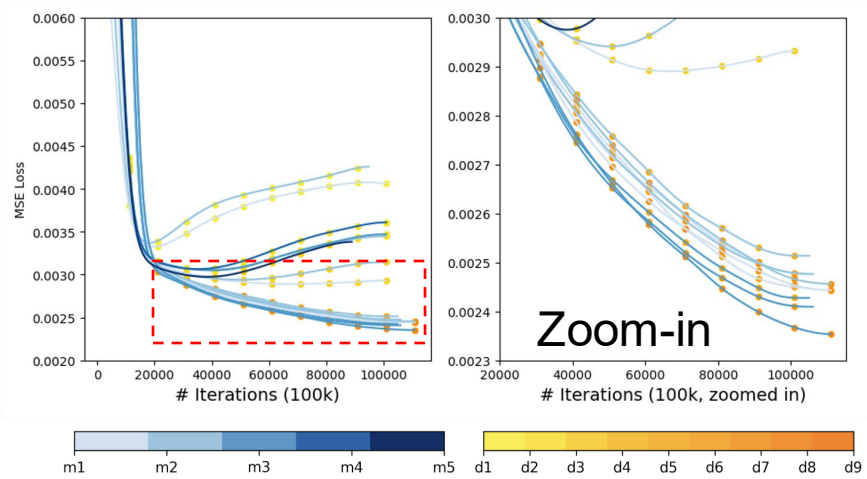
…Yet Quite Challenging

- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
- What neural network architecture to use?
- Will performance scale with the size of the network and data?
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…

- PointGPT-type
- Implicit Neural Network
- State Space Model.
- etc.

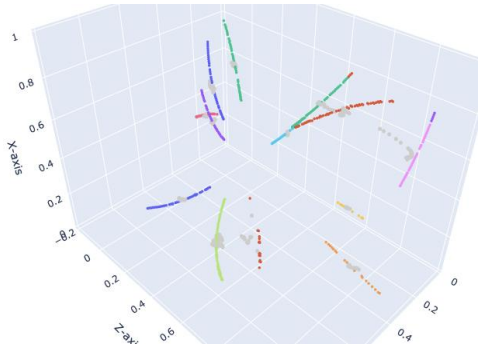
…Yet Quite Challenging

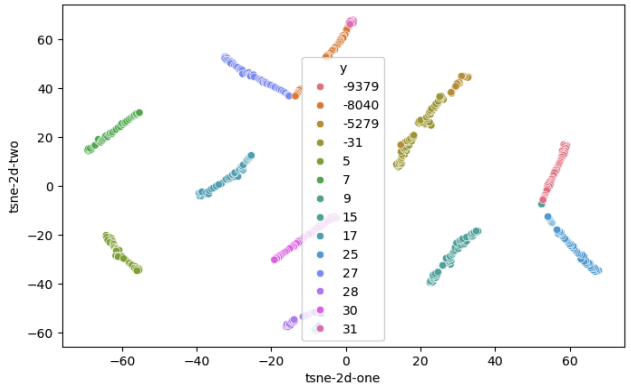
- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
- What neural network architecture to use?
- Will performance scale with the size of the network and data? ⟵ Hopefully
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…



…Yet Quite Challenging

- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
- What neural network architecture to use?
- Will performance scale with the size of the network and data?
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…

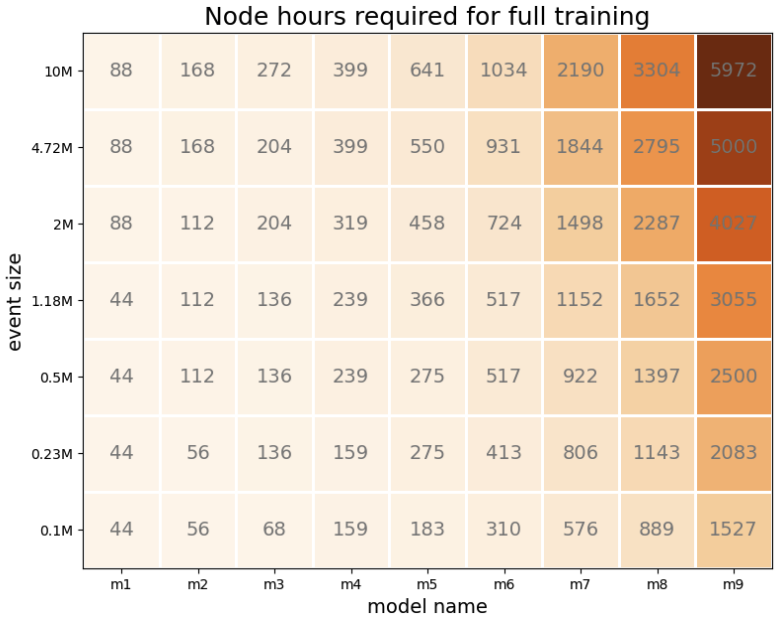




⟵ Very Likely

…Yet Quite Challenging

- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
- What neural network architecture to use?
- Will performance scale with the size of the network and data?
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…

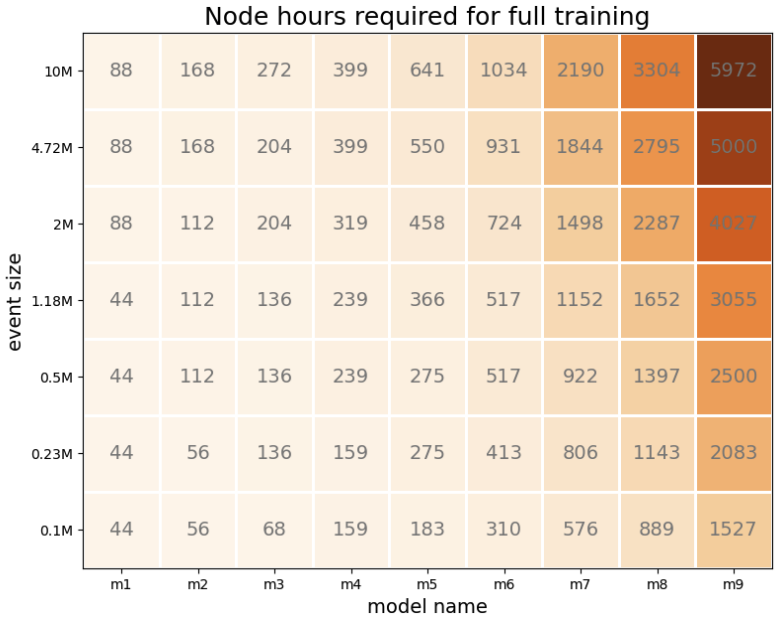


Estimated total node hours: 58,134

DOE HPCs:
Perlmutter (LBNL)
Frontier (ORNL)
SciServer (BNL)

…Yet Quite Challenging

- Data are 3D points, not language or image. How to formulate a self-supervised learning task?
- What neural network architecture to use?
- Will performance scale with the size of the network and data?
- Will the learned representations be useful for downstream tasks?
- Last but not the least, how much computation does it cost? c.f. GPT-3 costs ~$4.6 million.
- The list goes on and on…



Estimated total node hours: 58,134

DOE HPCs:
Perlmutter (LBNL)
Frontier (ORNL)
SciServer (BNL)

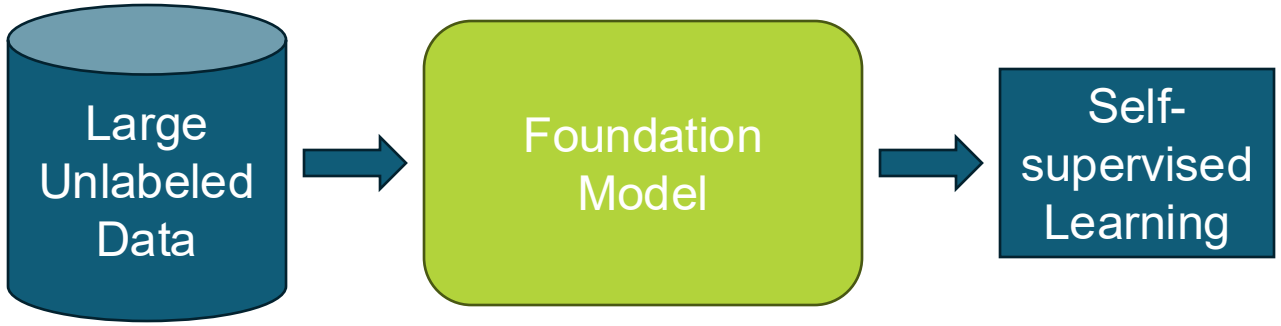
Goal

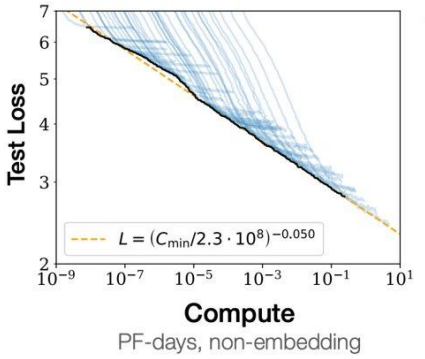
Proof of concept for an FM:

A. Neural scaling behavior

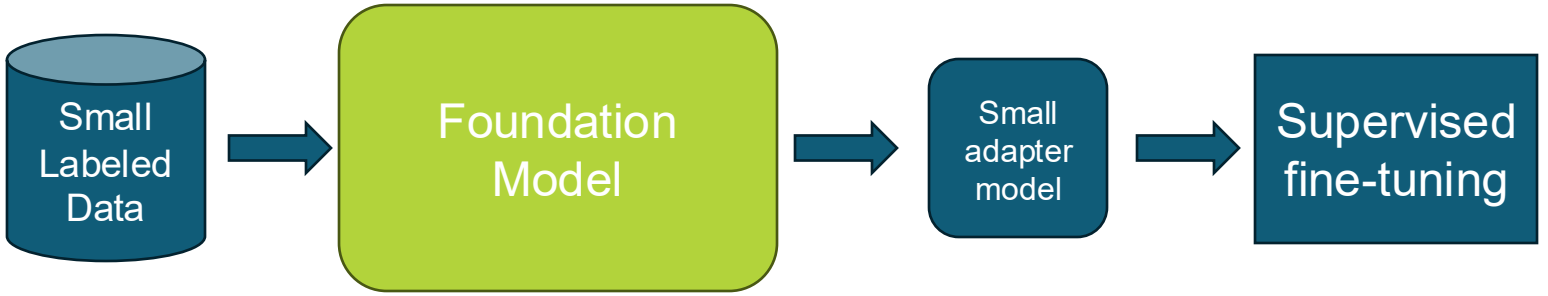
B. Generalizable to downstream tasks → e.g., Particle tracking application

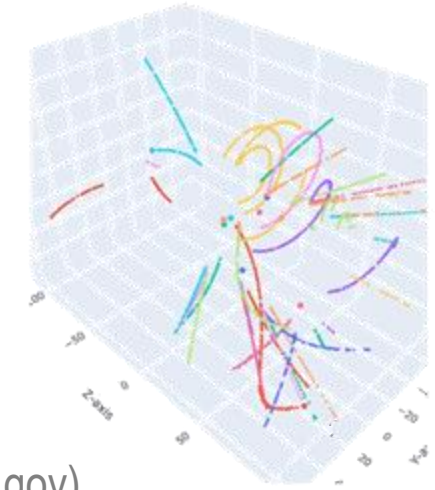
A.





B.

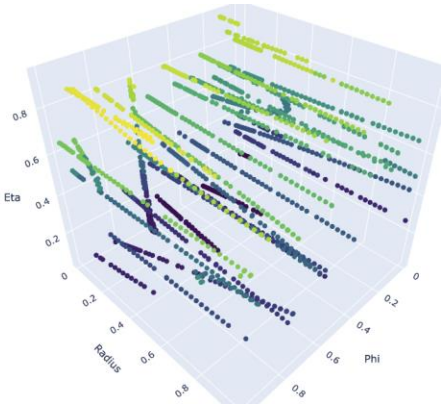


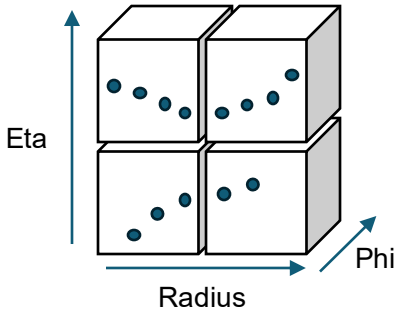


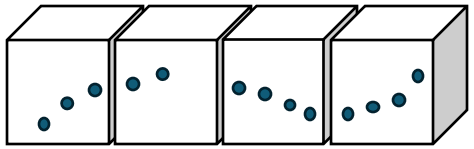
Self-supervised Task (WIP)

- Serialization: From 3D points to a sequence of points.
  - Divide the space into 3D cubes.

Radius -> Eta -> Phi (REP)

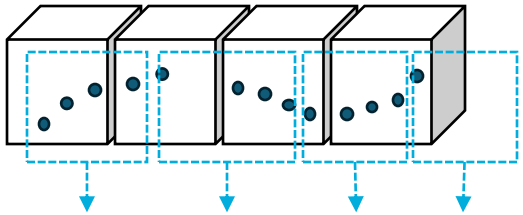






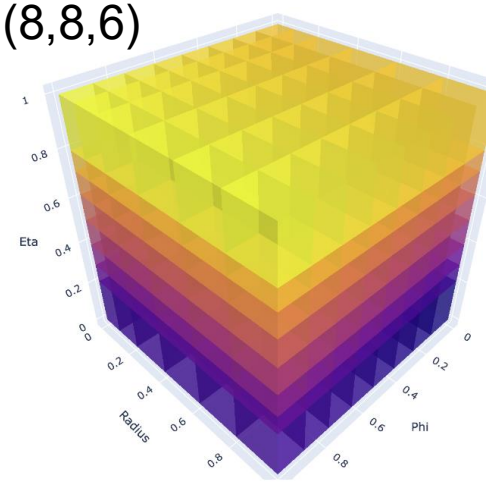
Within a cell, ordered by Radius

Grouping points sequentially (e.g., group_size = 4)



masking

3D cells of (8,8,6)



Neural Scaling (WIP)

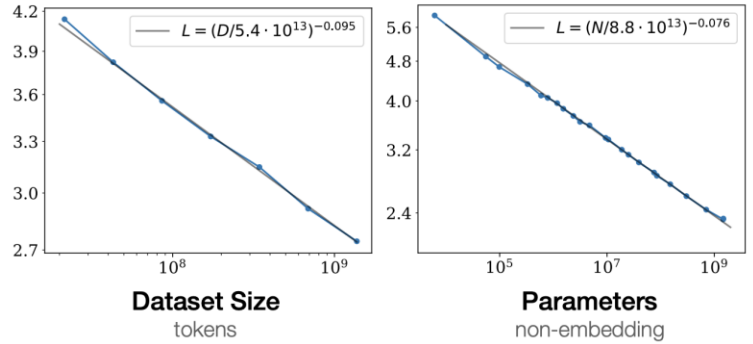
Model sizes

name	dim	layers (T [B/E/D])	heads (B/E/D))	Parameters
m1	128	12 [6 / 3 / 3]	4 / 2 / 2	**2.38 M**
m2	196	12 [6 / 3 / 3]	4 / 2 / 2	**5.56 M**
m3	256	12 [6 / 3 / 3]	8 / 4 / 4	**9.48 M**
m4	384	12 [6 / 3 / 3]	8 / 4 / 4	**21.3 M**
m5	512	14 [8 / 3 / 3]	8 / 4 / 4	**44.14 M**
m6	768	14 [8 / 3 / 3]	8 / 4 / 4	**99.24 M**
m7	1024	18 [10 / 4 / 4]	16 / 4 / 4	**226.74 M**
m8	1536	20 [12 / 4 / 4]	16 / 4 / 4	**509.98 M**
m9	2048	20 [12 / 4 / 4]	16 / 4 / 4	**1.01 B**

Dataset sizes

name	# events
d1	0.1M
d2	0.17M
d3	0.32M
d4	0.56M
d5	1.0M
d6	1.78M
d7	3.16M
d8	5.62M
d9	10M

Scaling for LMM



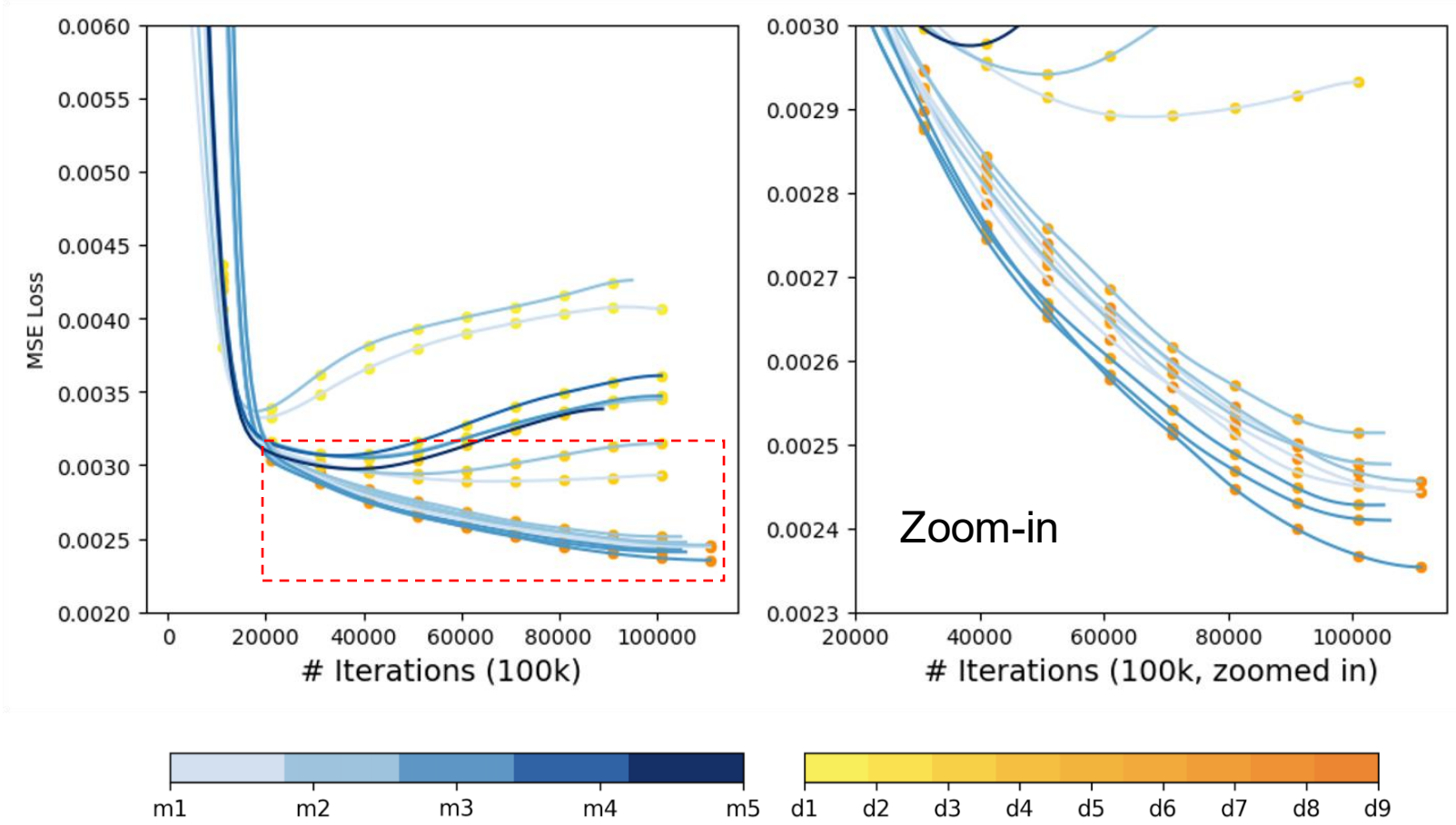
Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.

PointGPT-type Model Scaling (WIP)

Blue curves represent model size, m1 to m5 (2.38M to 44.14M parameters).

Yellow marks represent data size, d1 to d9 (10k events to 10M events).

Previous observation that the dataset is scaling remains valid, but the model is saturated quickly.

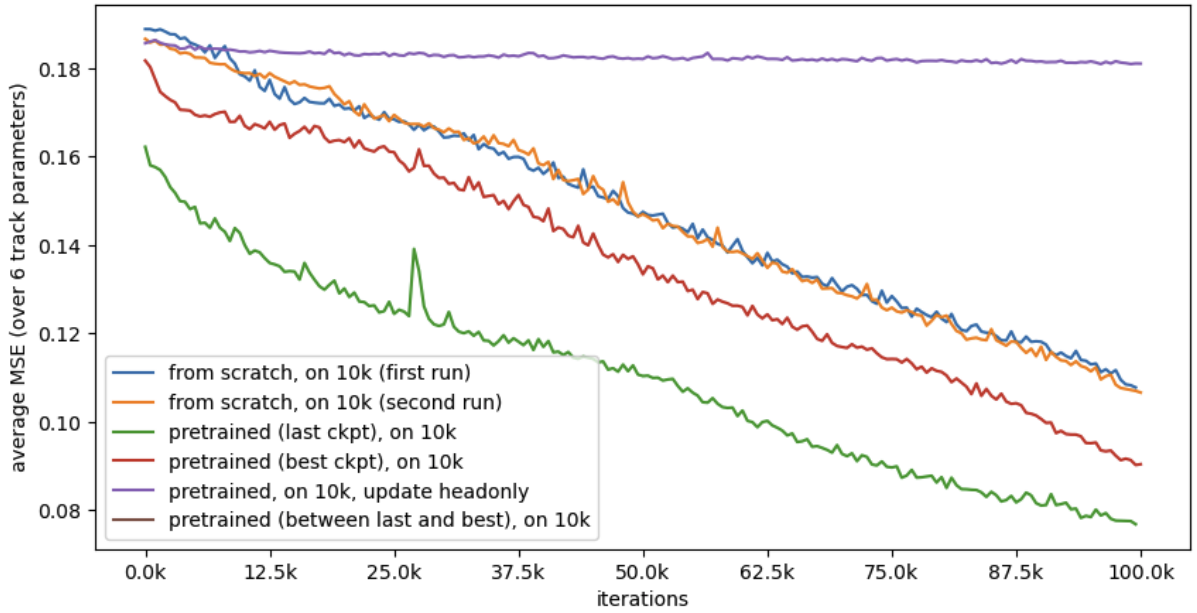


However…

Blue curves represent model size, m1 to m5 (2.38M to 44.14M parameters).

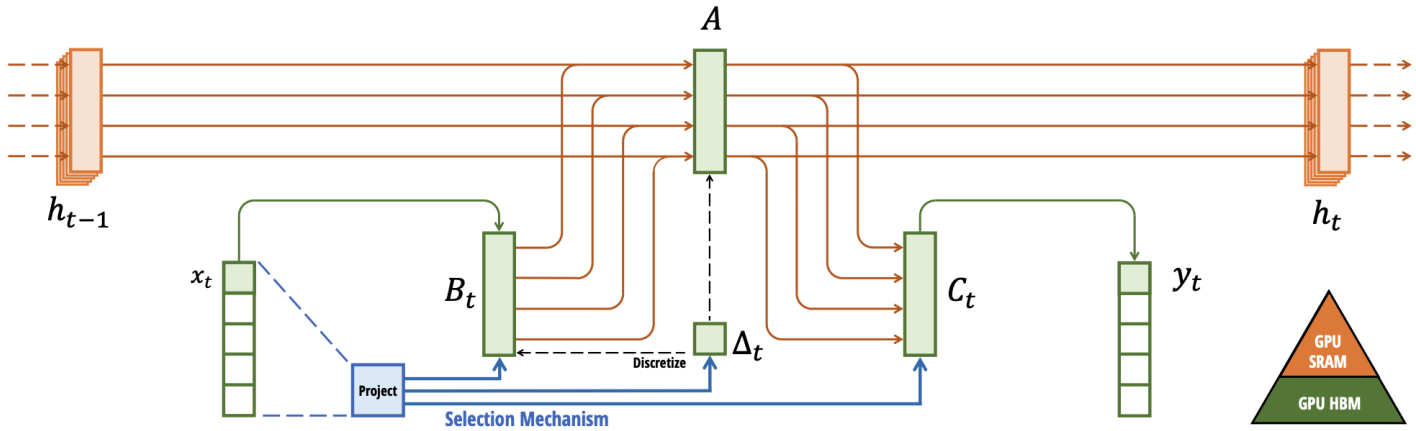
Yellow marks represent data size, d1 to d9 (10k events to 10M events).

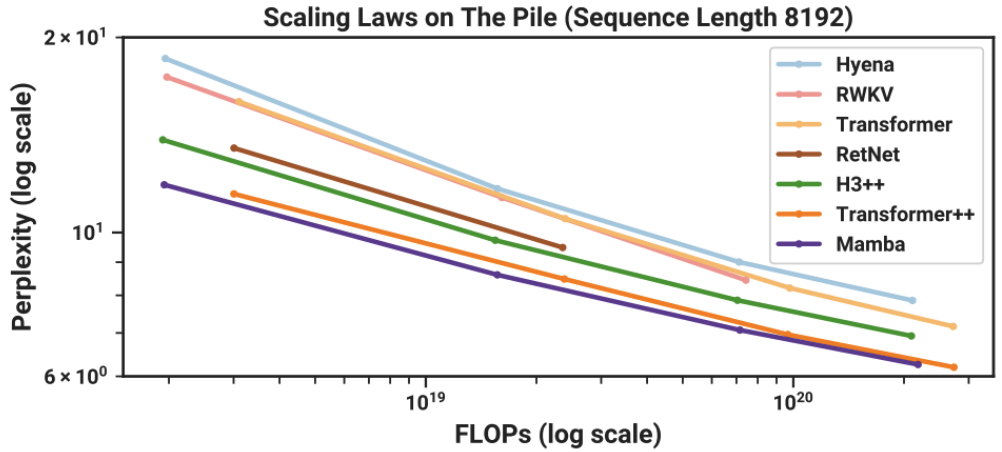
Previous observation that the dataset is scaling remains valid, but the model is saturated quickly.



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].





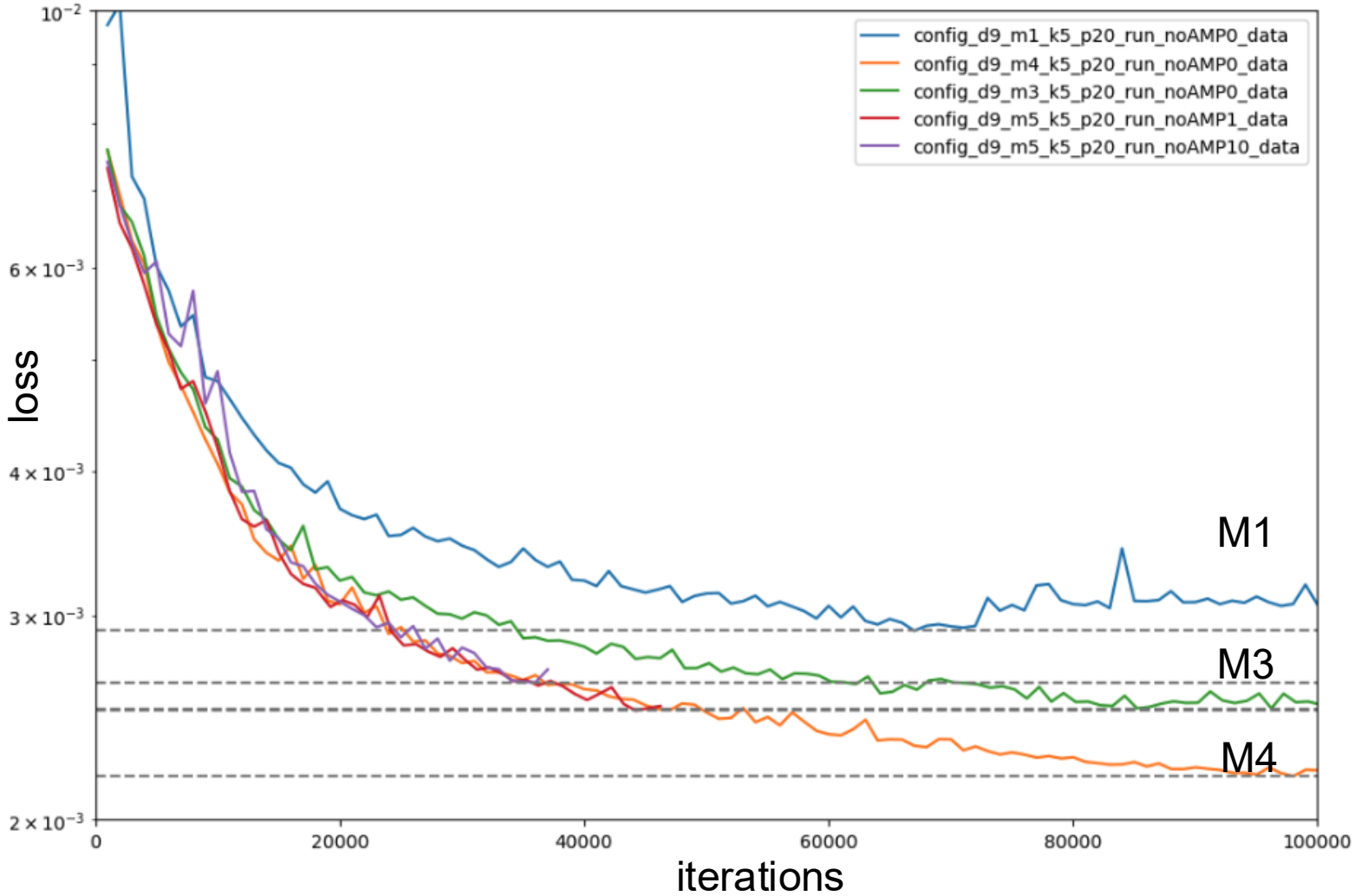
[1] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*.

Mamba Scaling (WIP)

MAMBA without AMP

- Takes about 48 hrs for 100,000 iterations.
- Clear scalability up to m4 then converged again at m5.

- M1: 5M
- M3: 20M
- M4: 80M
- M5: 320M: red – 18, purple – 12 layers

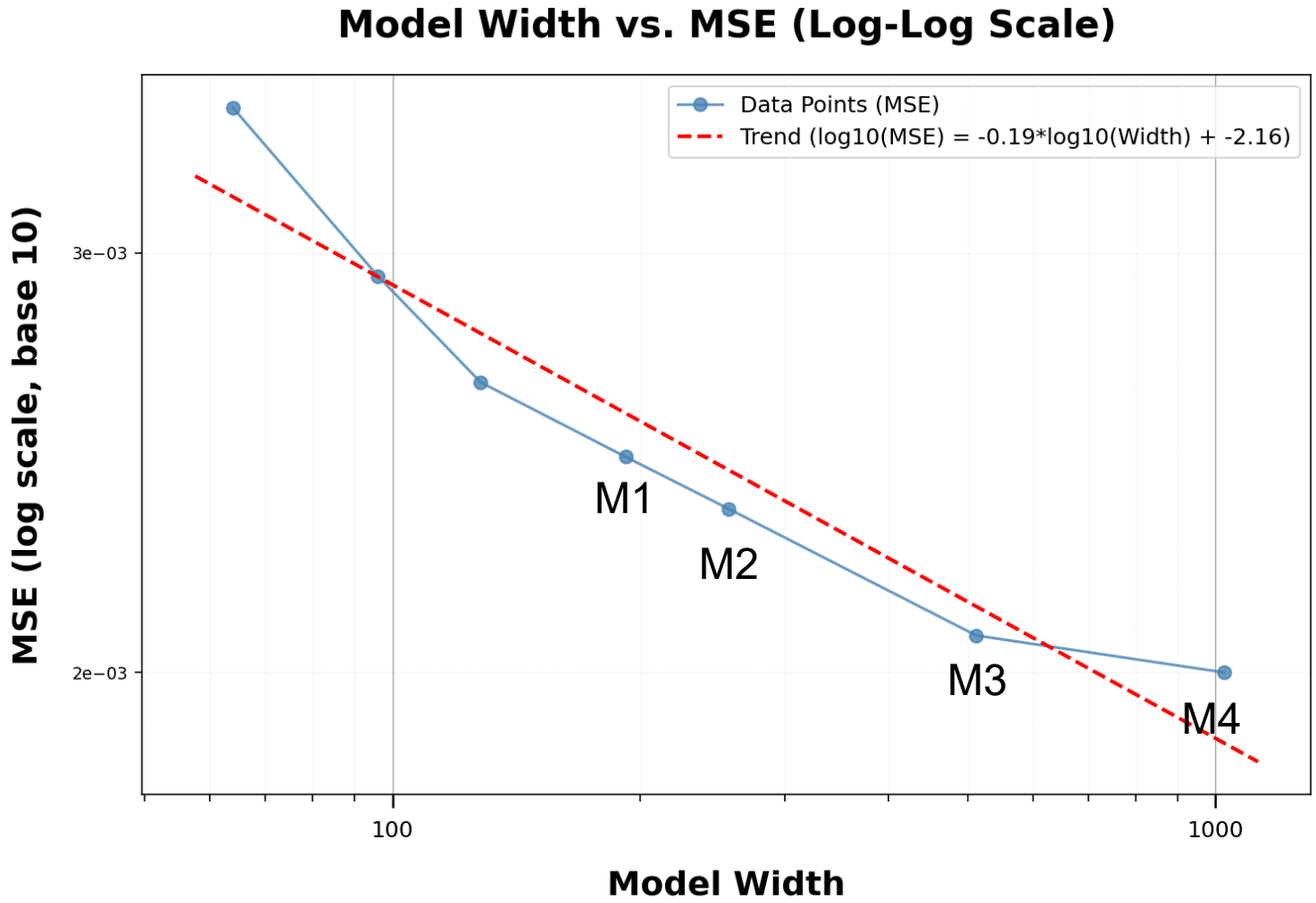


Mamba Scaling (WIP)

MAMBA without AMP

- Takes about 48 hrs for 100,000 iterations.
- Clear scalability up to m4 then converged again at m5.

- M1: 5M
- M3: 20M
- M4: 80M
- M5: 320M: red – 18, purple – 12 layers

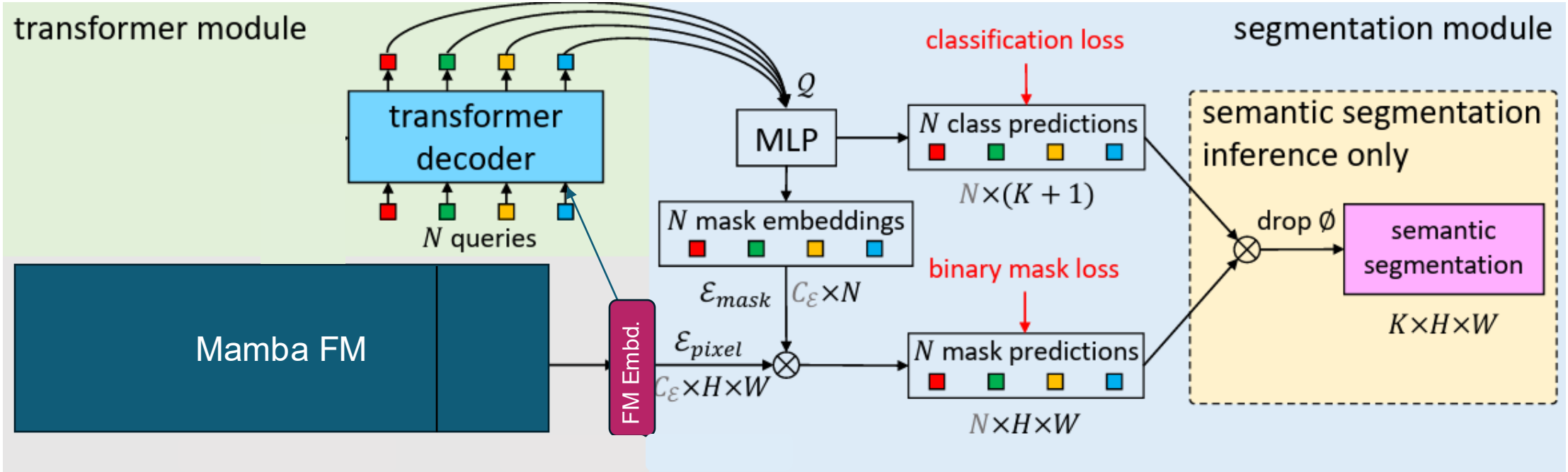


What Can We Do with Learned Representation? (WIP)

Adapting an instance segmentation "head":

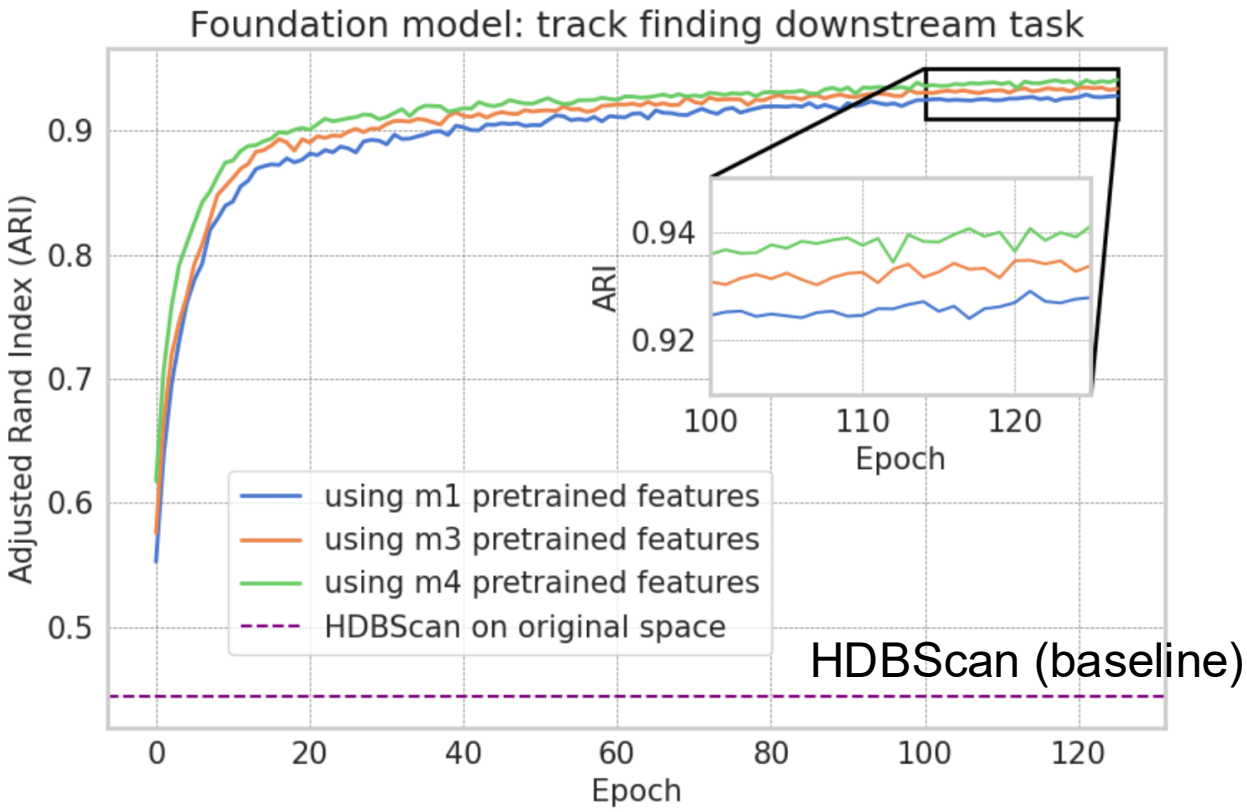
- Freeze the weights in FM.
- Adapt a MaskFormer [1] and train on FM embeddings.

[1] Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, *34*, 17864-17875.



Preliminary Results on Tracking (WIP)

- ✅ Larger models lead to better performance.
- The Adjusted Rand Index (**ARI**) is a widely used metric for evaluating the similarity between two clustering assignments
  - **+1**: Perfect agreement
  - **0**: Random labeling
  - **< 0**: Worse than random

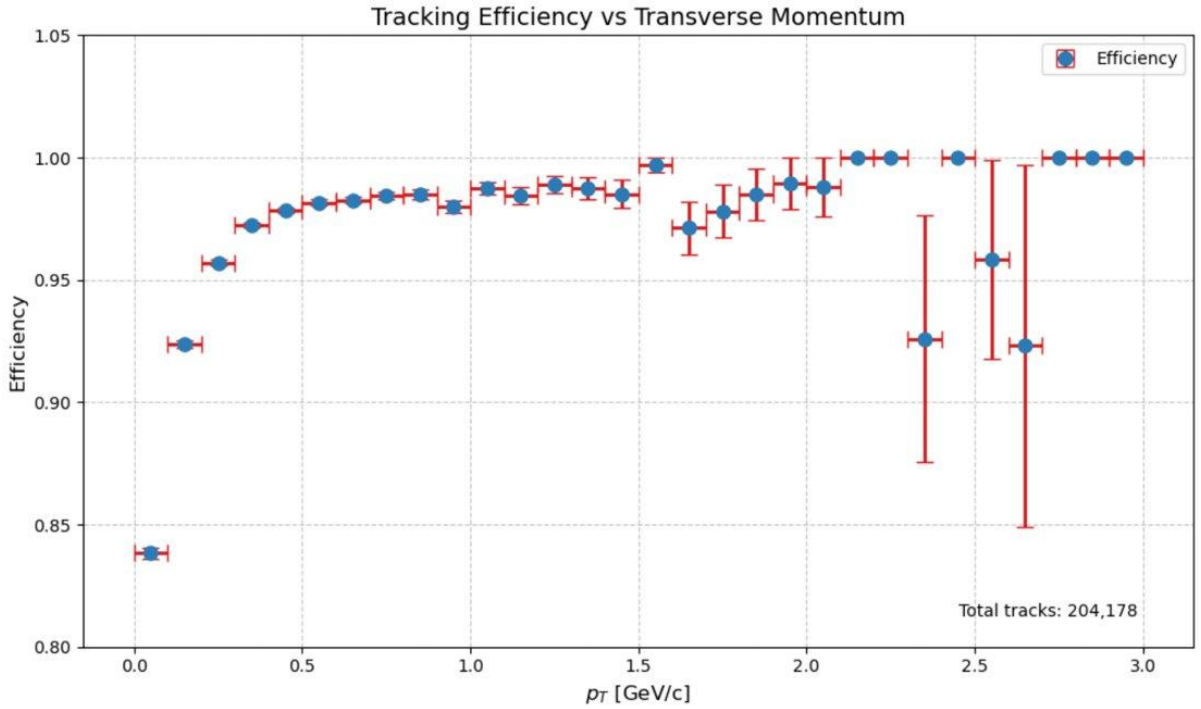


Preliminary Results on Tracking (WIP)

✅ 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.

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

Preliminary Results on Tracking (WIP)

Comparison with state-of-the-art methods (EggNet (GNN)[1], TrkX (GNN)[2])

(we will expand this table to include more models.)

	ARI	overall hit efficiency	overall hit purity	tracking efficiency		tracking purity
				overall	(> 1 GeV)	
Foundation Model	0.9424	97.52%	98.33%	95.98%	99.26%	93.02%
FM Adapter Head	0.8723	94.02%	95.13%	89.65%	95.67%	82.48%
EggNet	0.7256	93.01%	92.34%	74.19%	79.23%	75.14%
ExaTrkx	0.8052	96.49%	96.35%	87.76%	93.07%	86.08%

[1] Calafiura, Paolo, et al. "EggNet: An Evolving Graph-based Graph Attention Network for Particle Track Reconstruction." *arXiv:2407.13925* (2024).
[2] Ju, Xiangyang, et al. "Performance of a geometric deep learning pipeline for HL-LHC particle tracking." The European Physical Journal C 81 (2021): 1-14.

A Few Takeaways

Foundation Model: Pre-trained on a significant amount of data then can be used in many downstream tasks.

What FM needs: A large amount of data, well-defined downstream tasks, performance metrics, and passionate researchers. :D

A Few Takeaways

Foundation Model: Pre-trained on a significant amount of data then can be used in many downstream tasks.

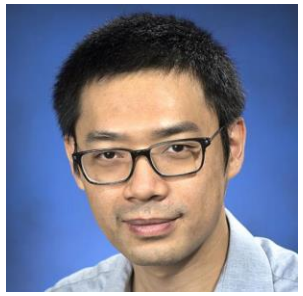
What FM needs: A large amount of data, well-defined downstream tasks, performance metrics, and passionate researchers. :D

Is this the Model-T moment?
Can we measure it?
If so, what can we do about it?

Image Credit: ChatGPT

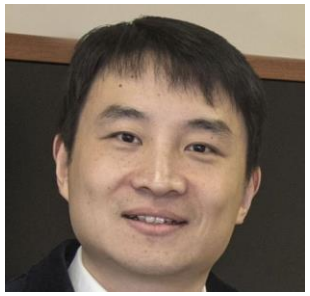
Thank you!

Yihui “Ray” Ren (yren@bnl.gov)



The Passionate Team.

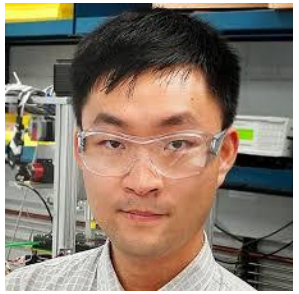



(AI Dept.) David Park, Yi Huang, Xihaier Luo, Yuewei Lin, Shinjae Yoo







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

Questions?

Yihui “Ray” Ren (yren@bnl.gov)

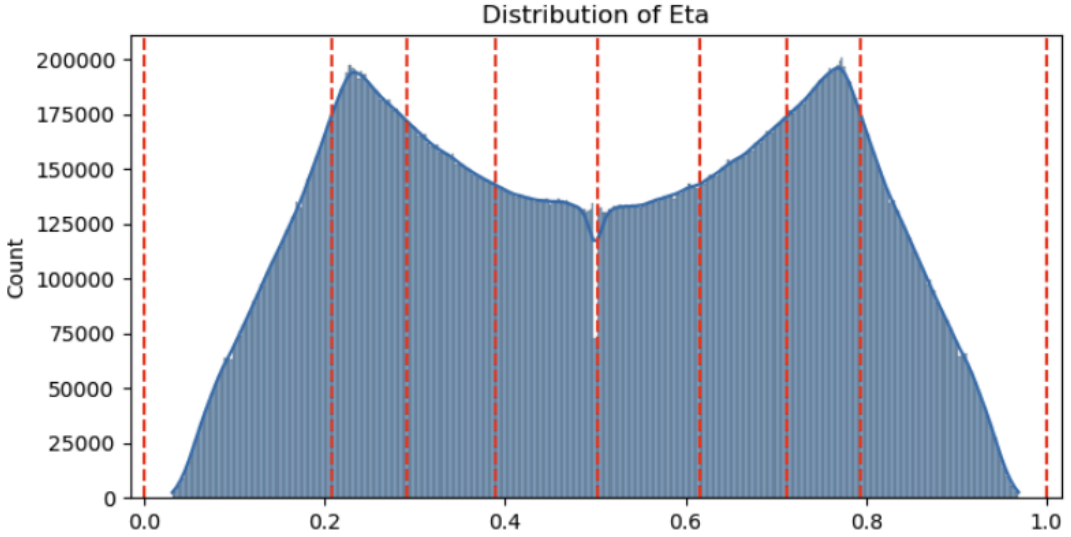
Acknowledgment

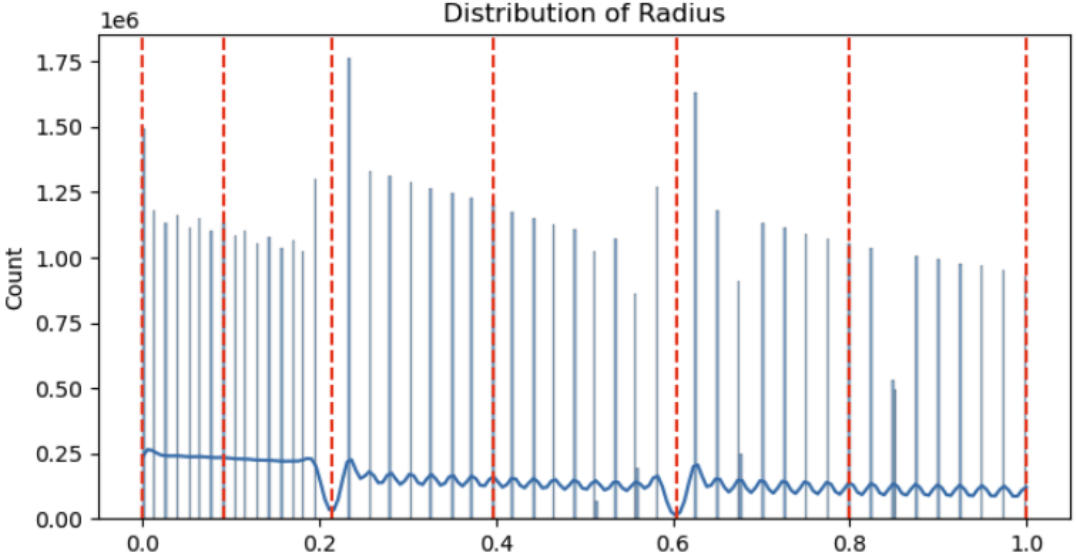
This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award DDR-ERCAP-m4722.

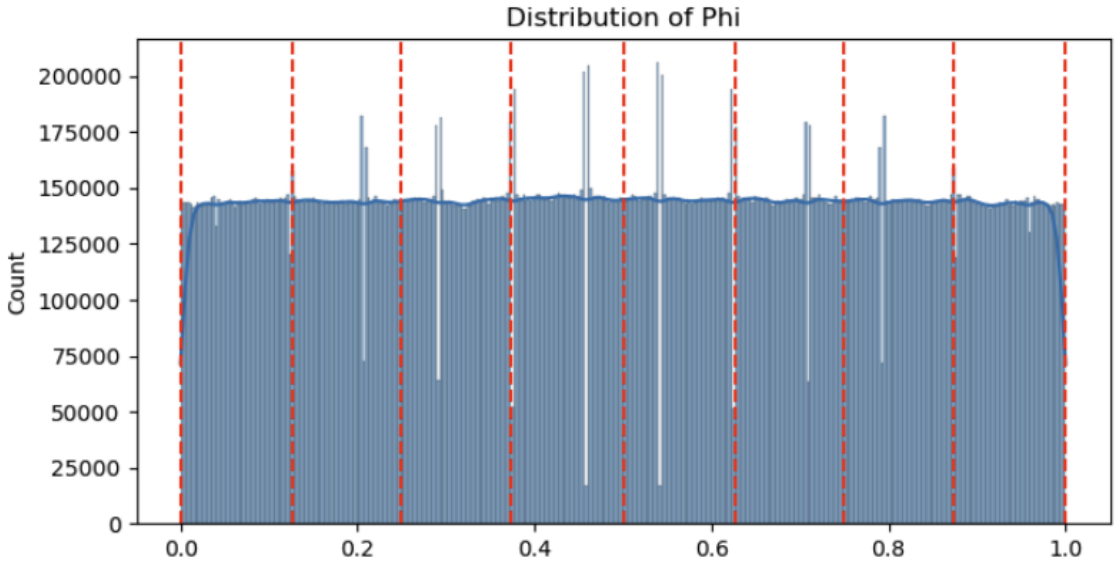
The LDRD Program at Brookhaven National Laboratory, sponsored by DOE’s Office of Science under Contract DE-SC0012704, supported this work.

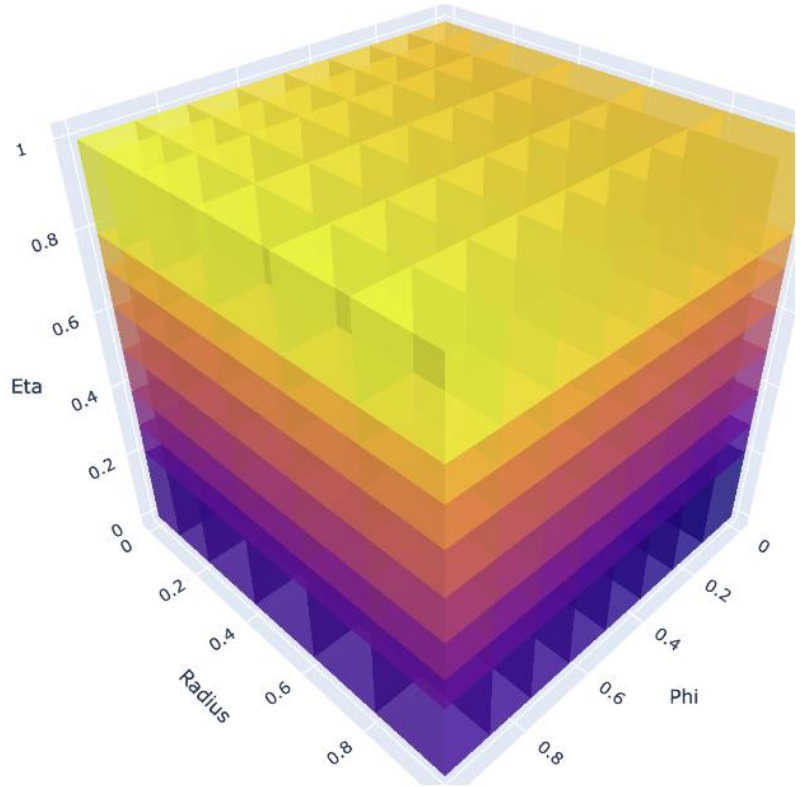
Voxelization parameters

Number of bins: (Eta, Phi, Radius) = (8, 8, 6)

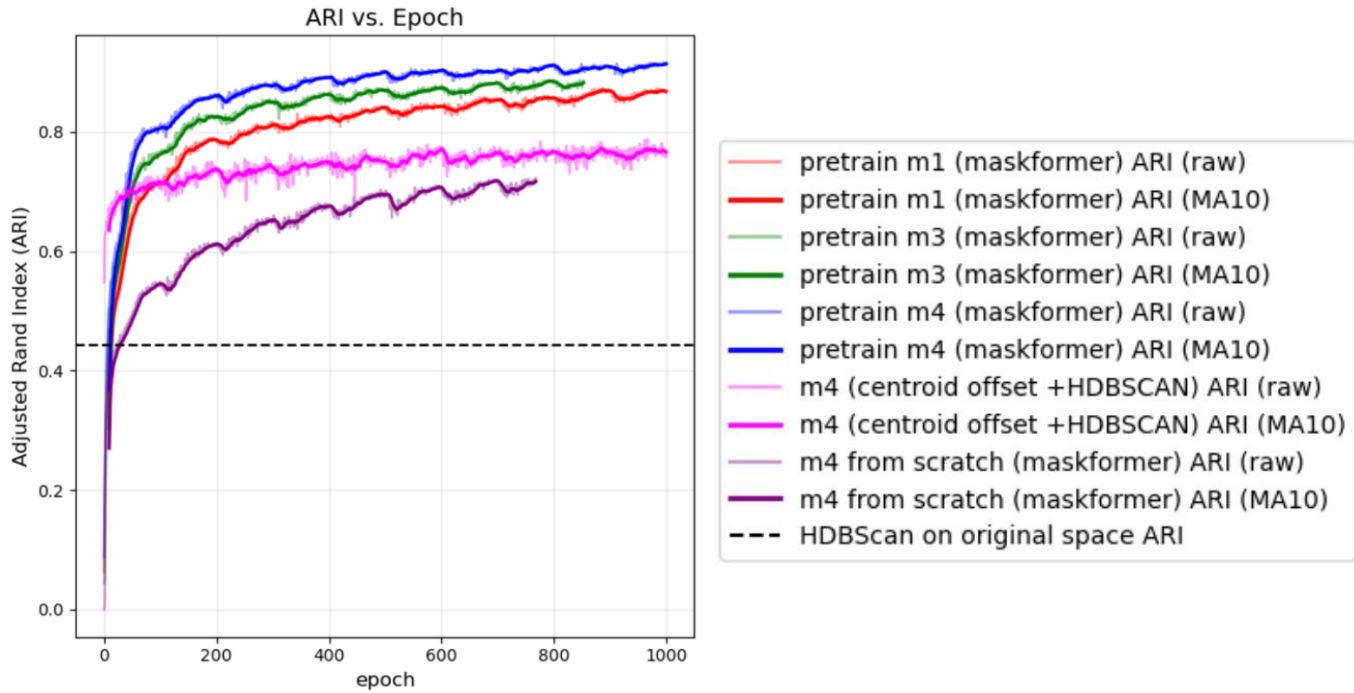








3D cells of (8,8,6)


ARI vs. Epoch
Adjusted Rand Index (ARI)
epoch
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
pretrain m1 (maskformer) ARI (raw)
pretrain m1 (maskformer) ARI (MA10)
pretrain m3 (maskformer) ARI (raw)
pretrain m3 (maskformer) ARI (MA10)
pretrain m4 (maskformer) ARI (raw)
pretrain m4 (maskformer) ARI (MA10)
m4 (centroid offset +HDBSCAN) ARI (raw)
m4 (centroid offset +HDBSCAN) ARI (MA10)
m4 from scratch (maskformer) ARI (raw)
m4 from scratch (maskformer) ARI (MA10)
HDBScan on original space ARI