A Playbook for Scaling Scientific Foundation Models

Lav R. Varshney

Della Pietra Infinity Professor and Inaugural Director of AI Innovation Institute, **Stony Brook University**
(electrical and computer engineering, political science, and neurobiology)

Co-Founder and CEO, **Kocree, Inc.**

Chief AI Strategist and Computational Scientist, **Brookhaven National Laboratory**

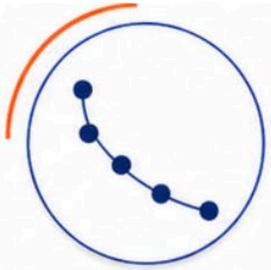
Joint work with Austin Ellis-Mohr (Illinois/Archean Sciences), Michael Nercessian (Berkely), Natalia Harguindeguy (Berkeley), Alexius Wadell (Michigan), Vignesh Sundaresha (Illinois), Yihui Ren (Brookhaven), and Adolfy Hoisie (Brookhaven)

Brookhaven National Laboratory

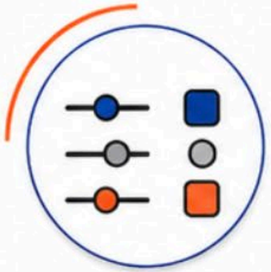
KOCREE

Stony Brook University

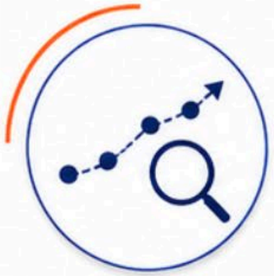
Outline



Scaling laws in biology, NLP, and science

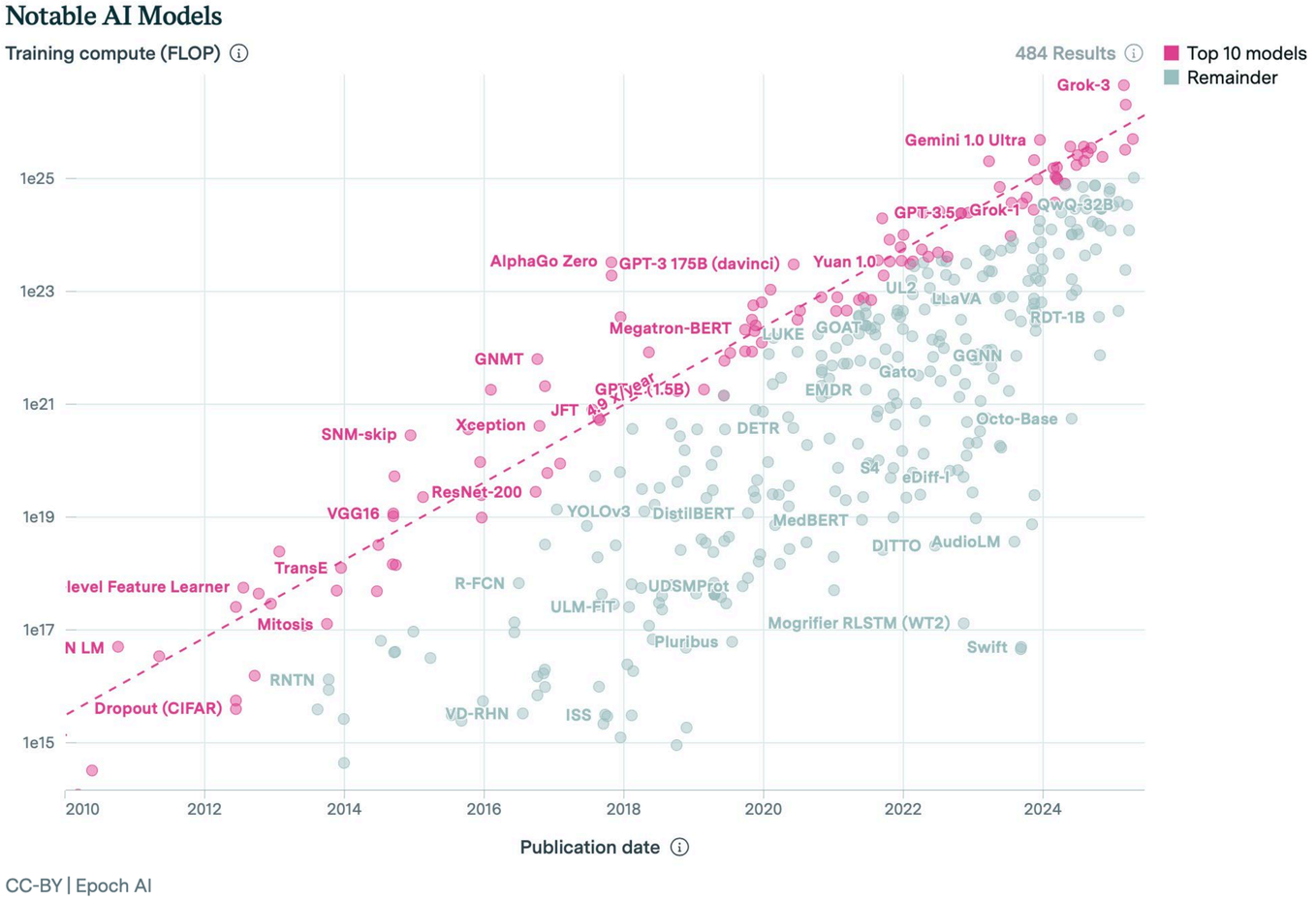


Domain specific design choices in SciFMs



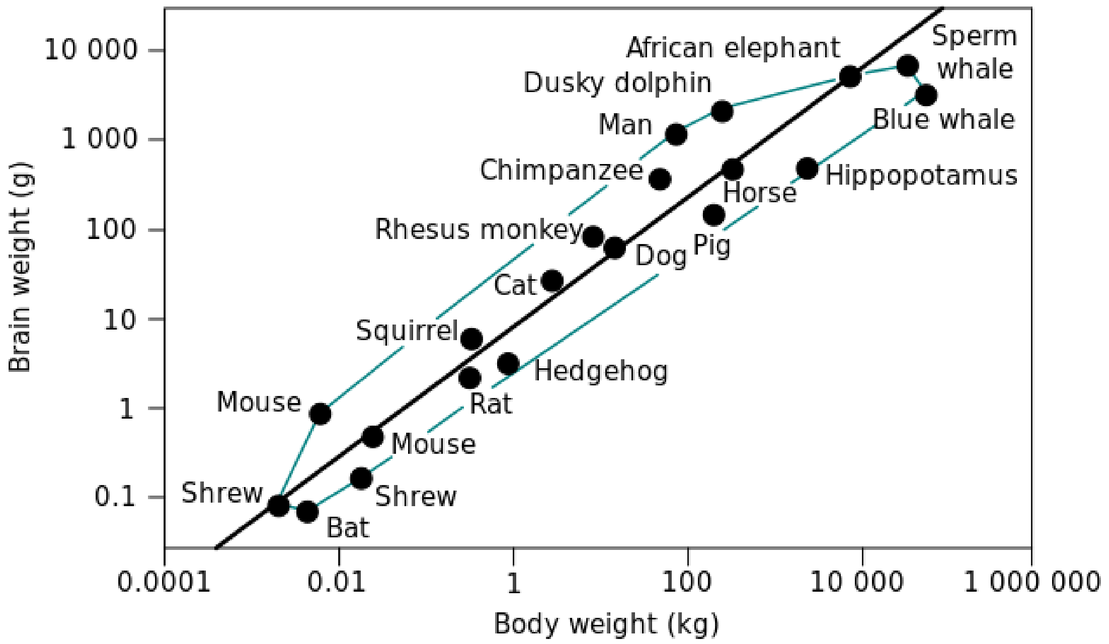
Small-run forecasting and theory-guided scaling

The hyperscaling paradigm



EpochAI, Accessed 2025.

Allometric scaling: scale is often power-law



https://en.wikipedia.org/wiki/Brain-to-body_mass_ratio

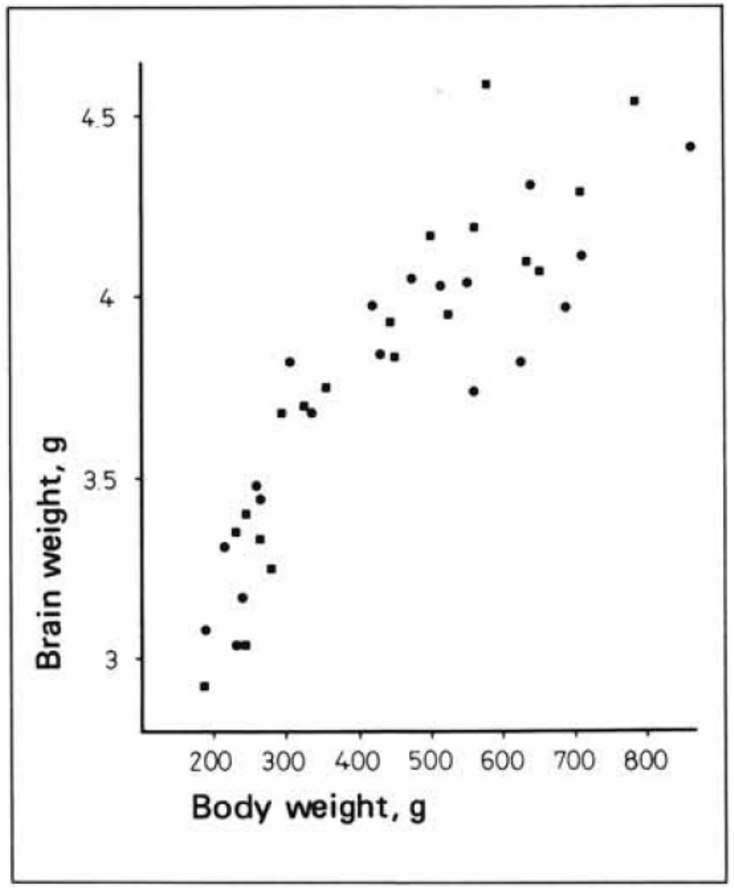
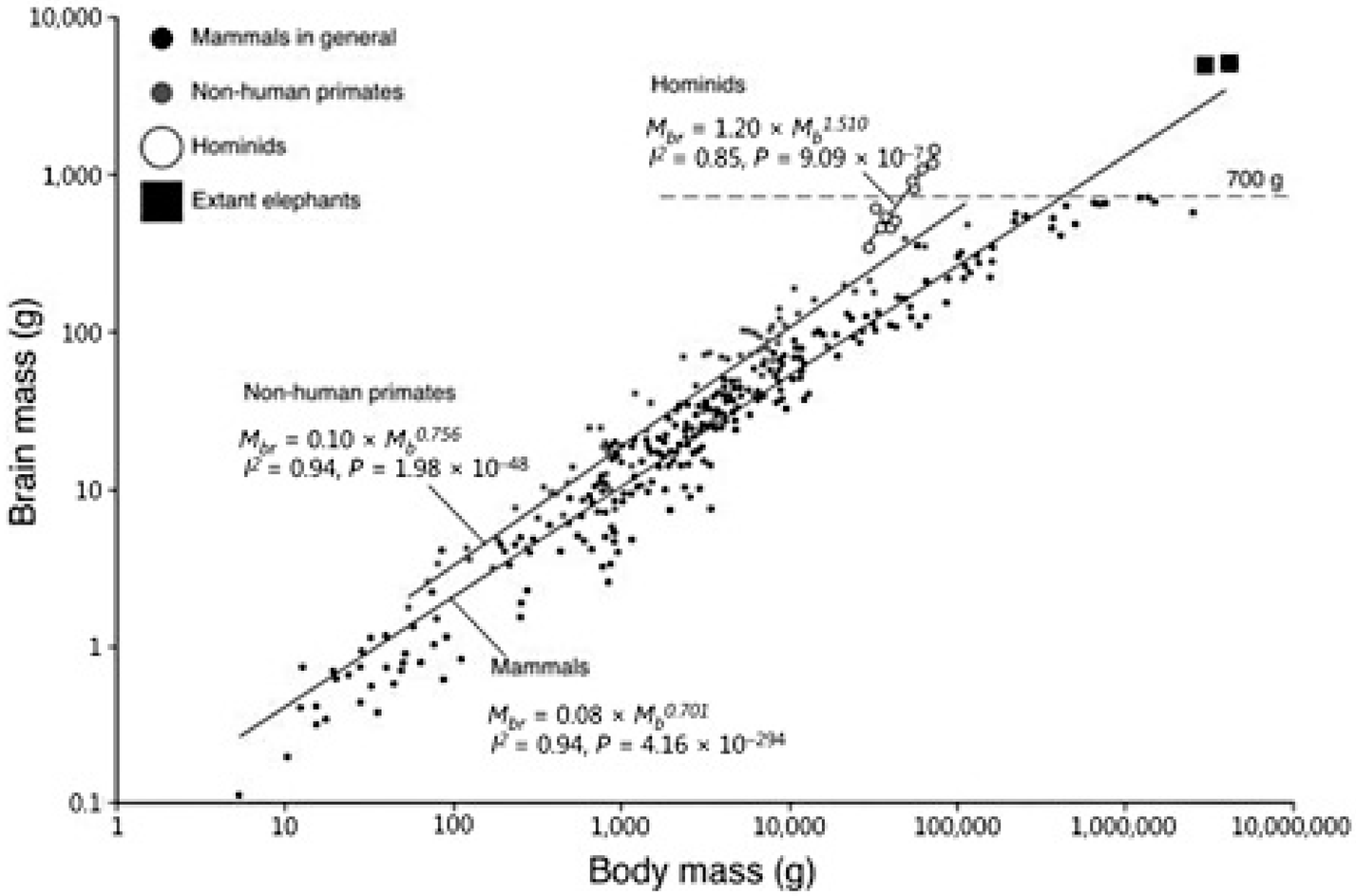


Fig. 1. Brain weights of guinea pigs *(Cavia cobaya)*

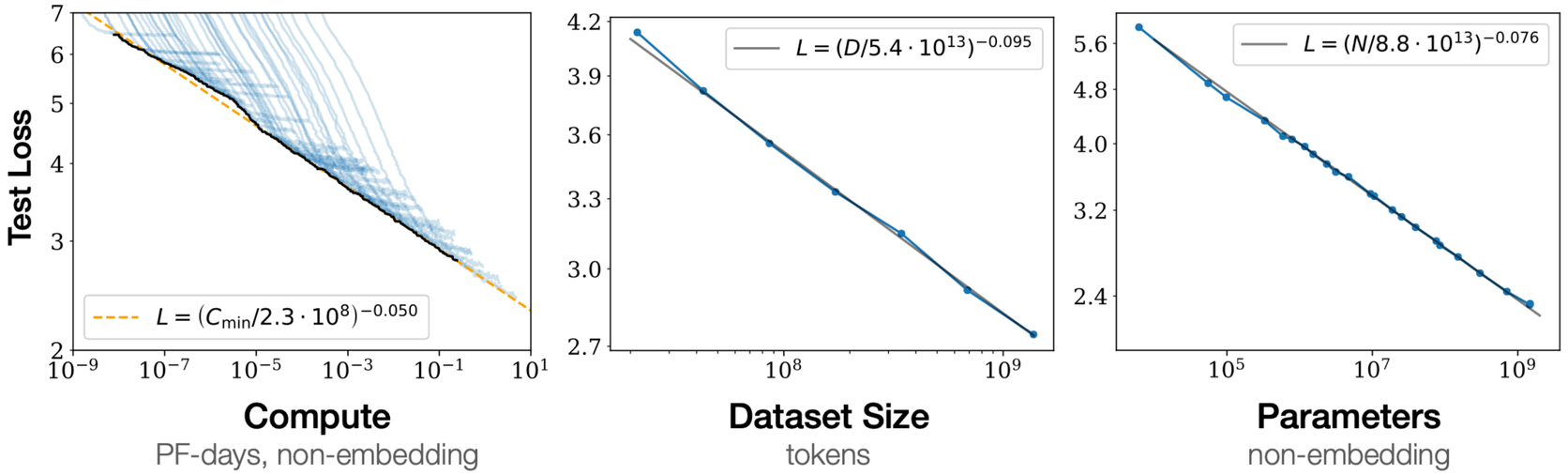
Stephan et al., **Folia Primatol.**, 1981.

Allometric scaling: exponents change by system



Manger et al., **Brain Behavior Evol**, 2013.

Characterized scaling laws in NLP



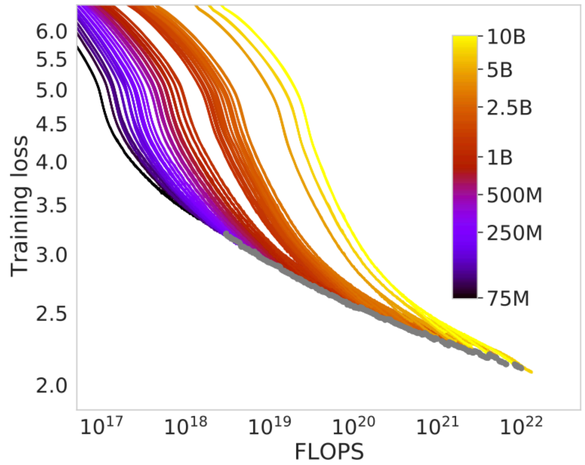
Kaplan et al., **arXiv**, 2020.

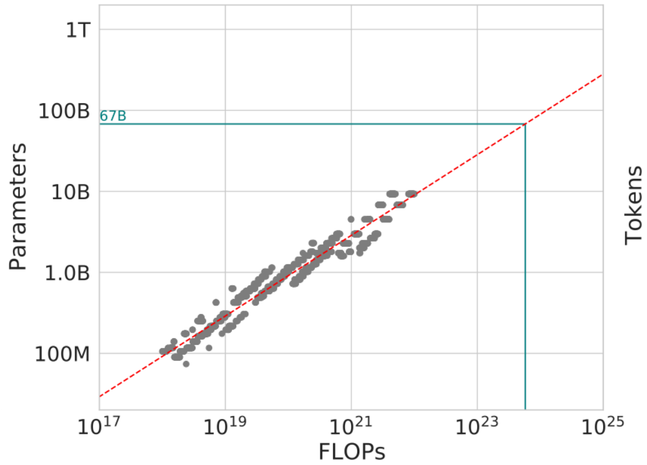
$$L \propto AN^{-\alpha} + BD^{-\beta}$$

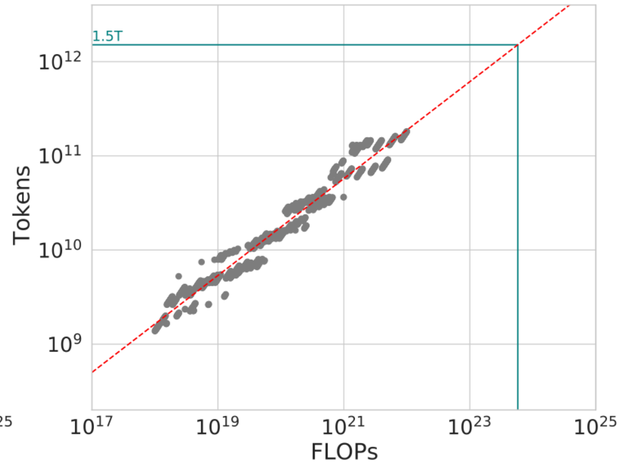
$$D_{\text{opt}} \propto (N_{\text{opt}})^{\frac{\alpha}{\beta}}$$

Hoffmann et al., **arXiv**, 2022.

- Empirically fit NLP scales with a power law

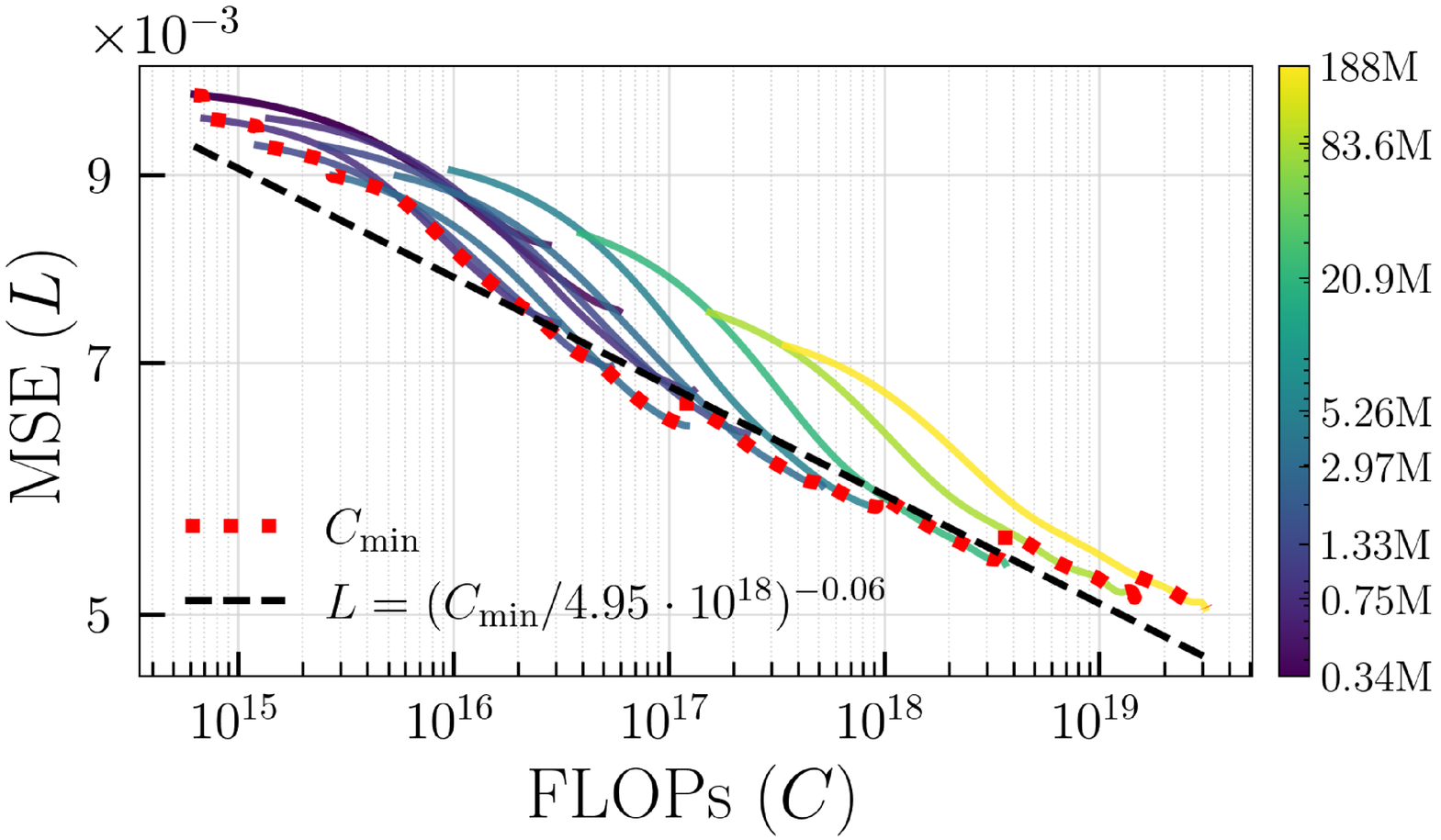






Physics SciFM scales with a power law

FM4NPP: A Scaling Foundation Model for Nuclear and Particle Physics



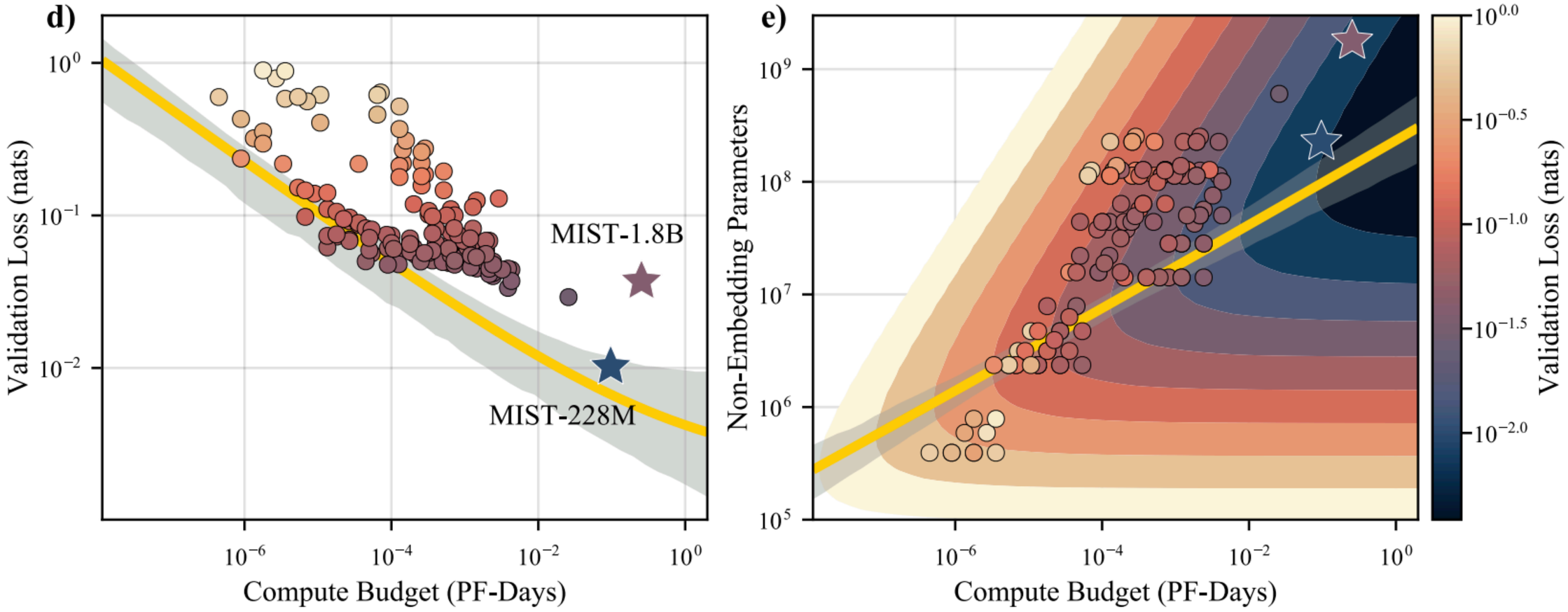
model	#trnbl para.	Particle Identification acc.↑	recall↑	pre.↑
SAGEConv	0.91M	0.7262	0.4563	0.6502
OneFormer3D	44.95M	0.7701	0.4897	0.5767
AdapterOnly	0.74M	0.6631	0.3387	0.6111
FM4NPP (m6)	0.74M	**0.9039**	**0.7652**	**0.8782**

Figure 2. SOTA performance on downstream tasks

Park et al., **arXiv**, 2025.

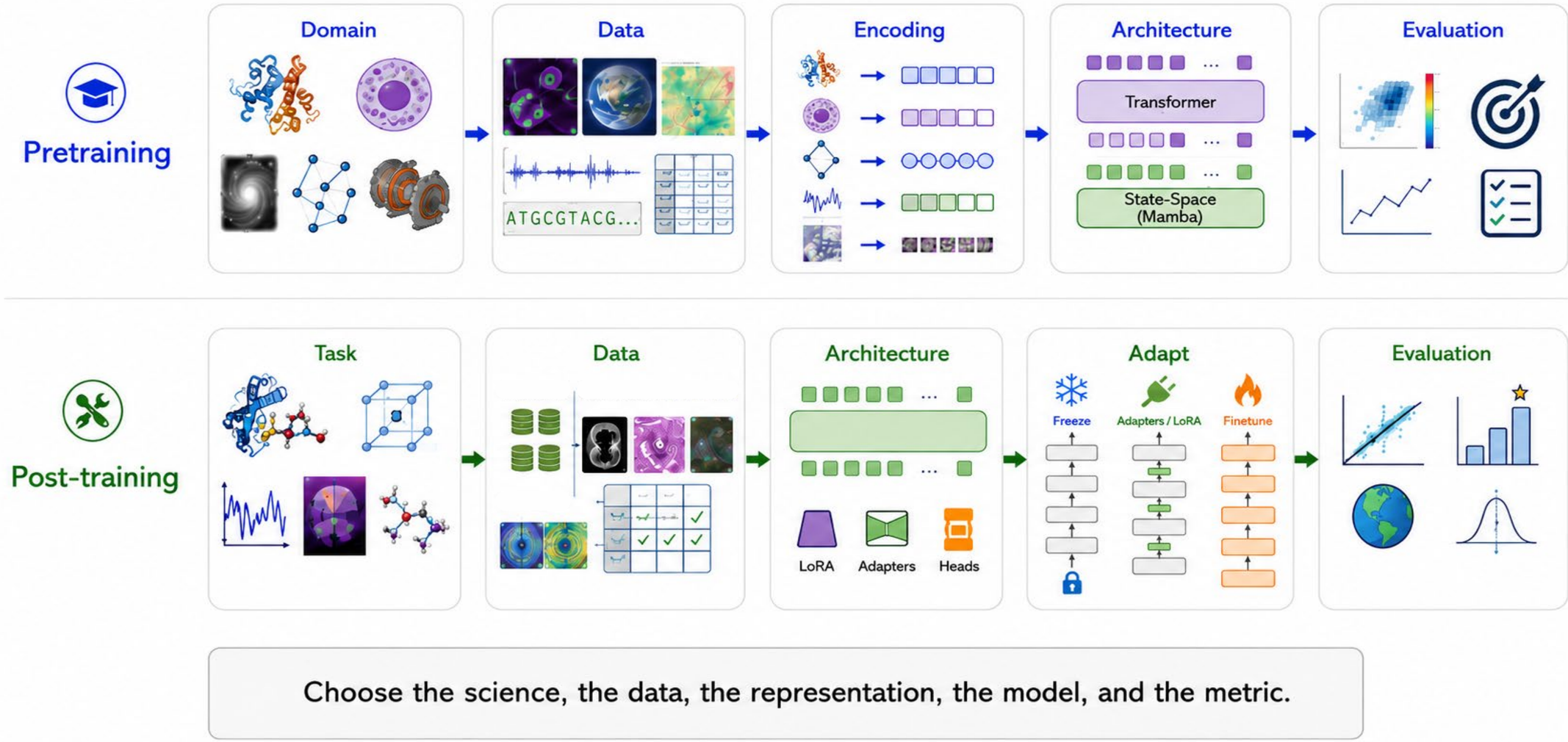
Chemistry SciFM scales with a power law

Foundation Models for Discovery and Exploration in Chemical Space

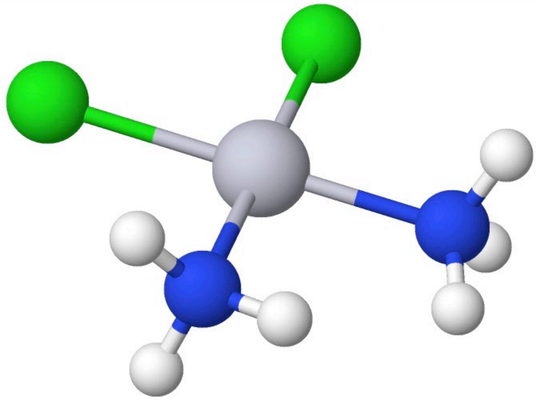


Wadell et al., **arXiv**, 2025.

A SciFM training workflow



Selecting representation: the “language” of chemistry



Wadell, Bhutani, Viswanathan, **arXiv**, 2025.

Cl[Pt@SP1](Cl)([NH3])[NH3]

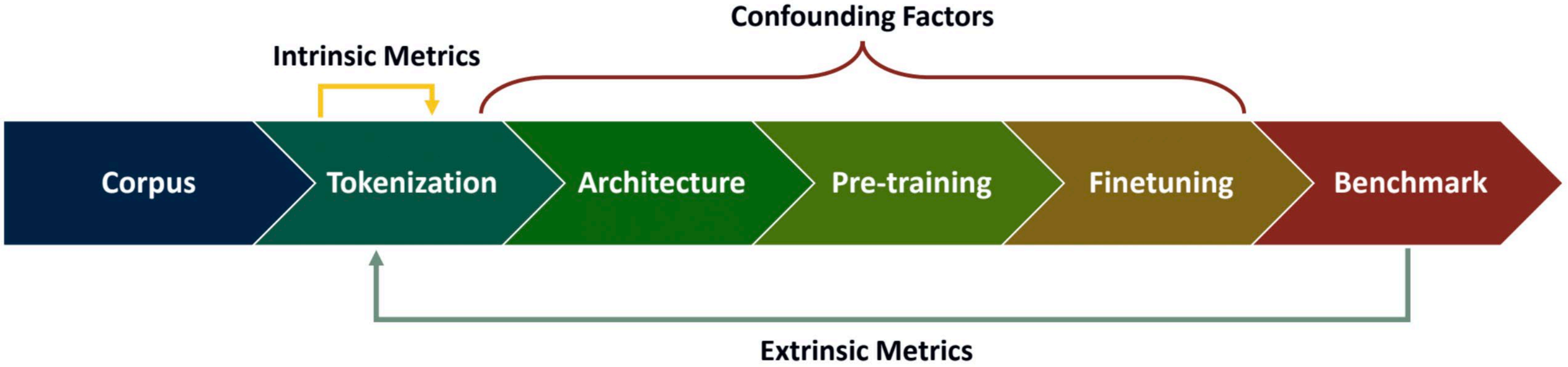
SMILES: Cl[Pt@SP1](Cl)([NH3])[NH3]

Atom: Cl [Pt@SP1] (Cl) ([NH3]) [NH3]

Smirk: Cl [Pt @SP 1] (Cl) ([N H 3]) [N H 3]

GPT-4o: Cl [Pt @ SP 1] (Cl) ([NH 3]) [NH 3]

Tokenization affects performance through the pipeline



N-Grams: $$P_n(x_i|x_{i-n+1},\ldots,x_{i-1}) = \frac{C(x_{i-n+1},\ldots,x_i)+1}{C(x_{i-n+1},\ldots,x_{i-1})+|V|}$$

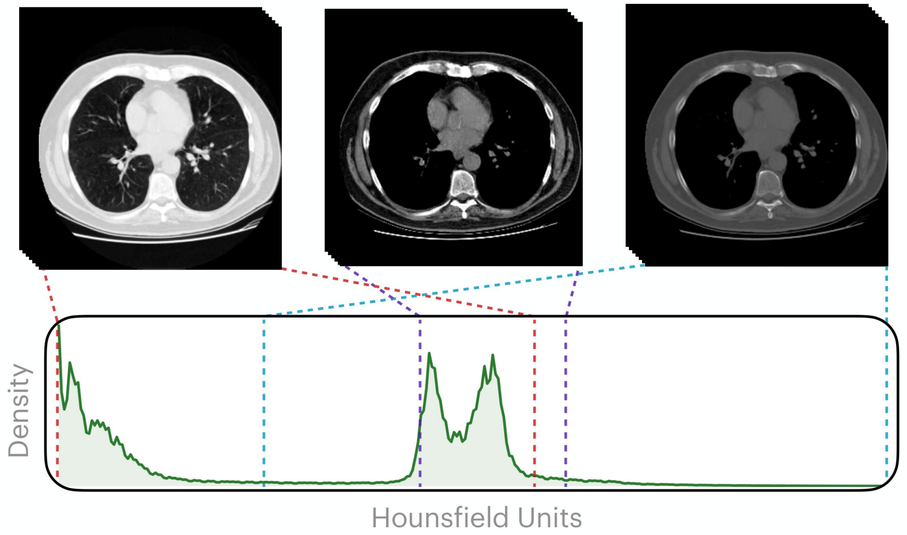
Wadell, Bhutani, Viswanathan, **arXiv**, 2025.

Modality-native tokenization in radiology

Pillar-0: A new frontier for radiology foundation models

- Radiology volumes occupy a wide intensity range, and certain structures are optimally viewed in subspaces of this range.
- **Pillar-0** emulates radiology workflows by projecting volumes into multiple *windows*, with each window occupying a different slice of the intensity range.

3D CT volume



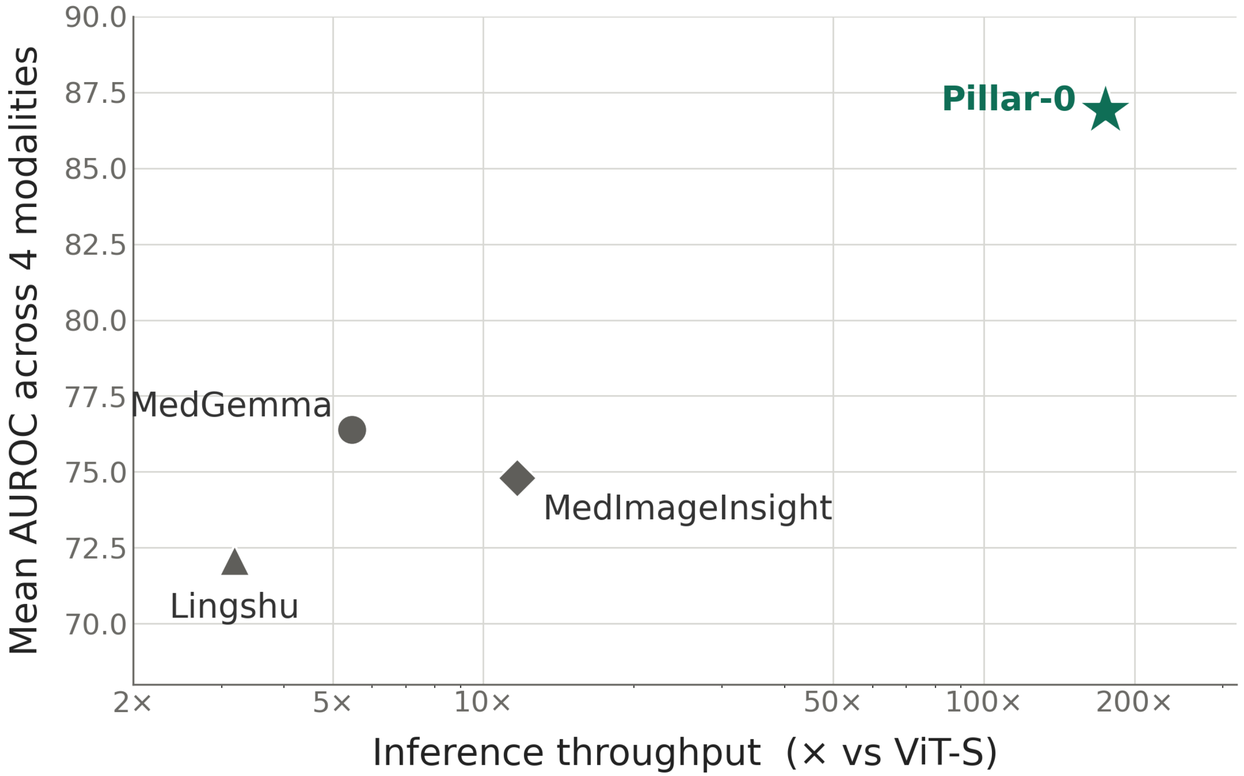
+4.6 AUROC compared to existing standard (single-channel) on abdomen-pelvis CT downstream task (clinical findings classification; 82.2 vs 77.6).

Agrawal, Liu, Lian, Nercessian, Harguindeguy, et al., **arXiv**, 2025.

3D radiology needs volume-aware architecture

Pillar-0: A new frontier for radiology foundation models

- Radiology volumes (512x512x768) can be thousands of times larger than a 224x224 ImageNet image
- Full 3D resolution with standard architectures is intractable leading to other radiology FMs processing them slice-by-slice
- **Pillar-0** adopts a multi-scale attention and progressive downsampling architecture (Atlas) to make full-resolution 3D modeling tenable and performant

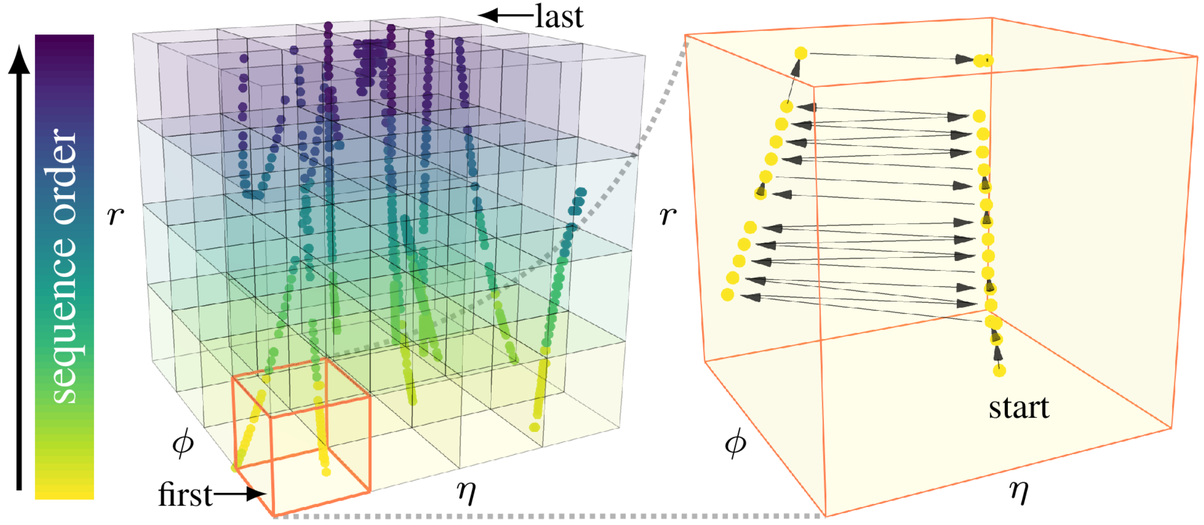


Agrawal, Liu, Lian, Nercessian, Harguindeguy, et al., **arXiv**, 2025.

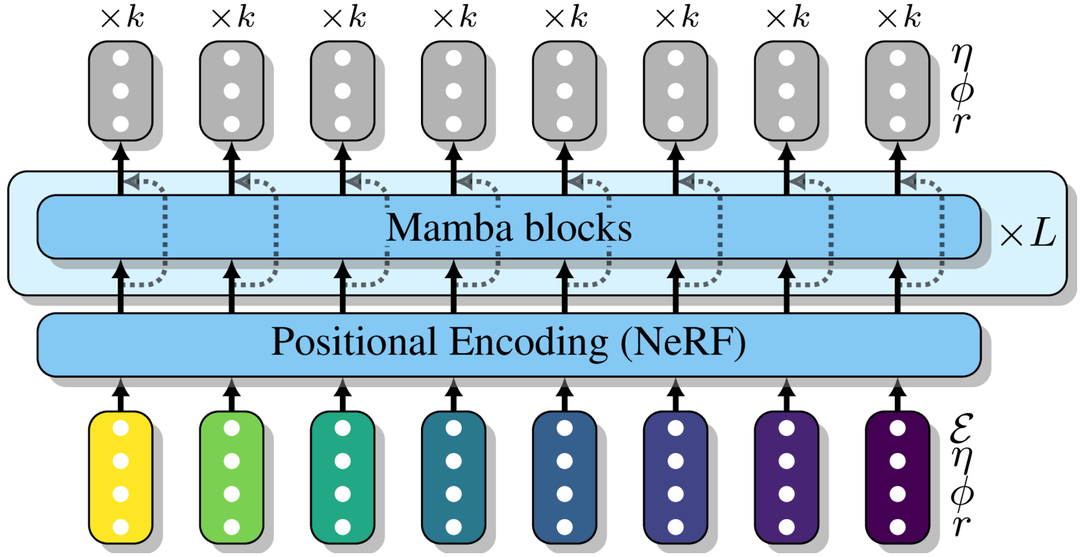
Architecture choice for FM4NPP

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

(a) 3D grid partition and ordering



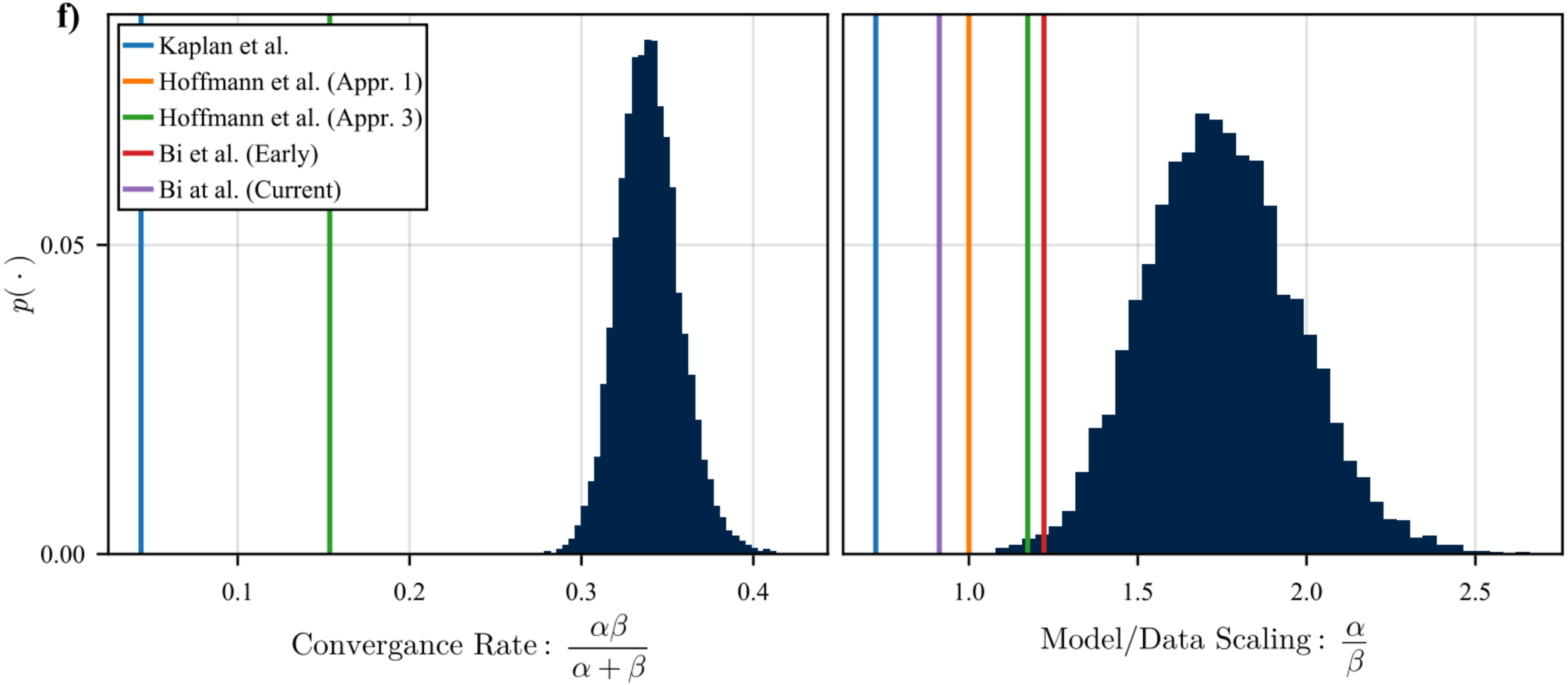
(b) Foundation model



Park et al., **arXiv**, 2025.

SciFM scaling exponents can differ from NLP

Foundation Models for Discovery and Exploration in Chemical Space



Wadell et al., **arXiv**, 2025.

SciFM scaling exponents can differ from NLP

nature communications

Article

https://doi.org/10.1038/s41467-023-42992-y

Exploiting redundancy in large materials datasets for efficient machine learning with less data

Received: 30 April 2023

Accepted: 26 October 2023

Published online: 10 November 2023

Check for updates

Kangming Li[1], Daniel Persaud[1], Kamal Choudhary[2], Brian DeCost[2], Michael Greenwood[3] & Jason Hattrick-Simpers[1,4,5,6]

Extensive efforts to gather materials data have largely overlooked potential data redundancy. In this study, we present evidence of a significant degree of redundancy across multiple large datasets for various material properties, by revealing that up to 95% of data can be safely removed from machine learning training with little impact on in-distribution prediction performance. The redundant data is related to over-represented material types and does not mitigate the severe performance degradation on out-of-distribution samples. In addition, we show that uncertainty-based active learning algorithms can construct much smaller but equally informative datasets. We discuss the effectiveness of informative data in improving prediction performance and robustness and provide insights into efficient data acquisition and machine learning training. This work challenges the "bigger is better" mentality and calls for attention to the information richness of materials data rather than a narrow emphasis on data volume.

UMA: A Family of Universal Models for Atoms

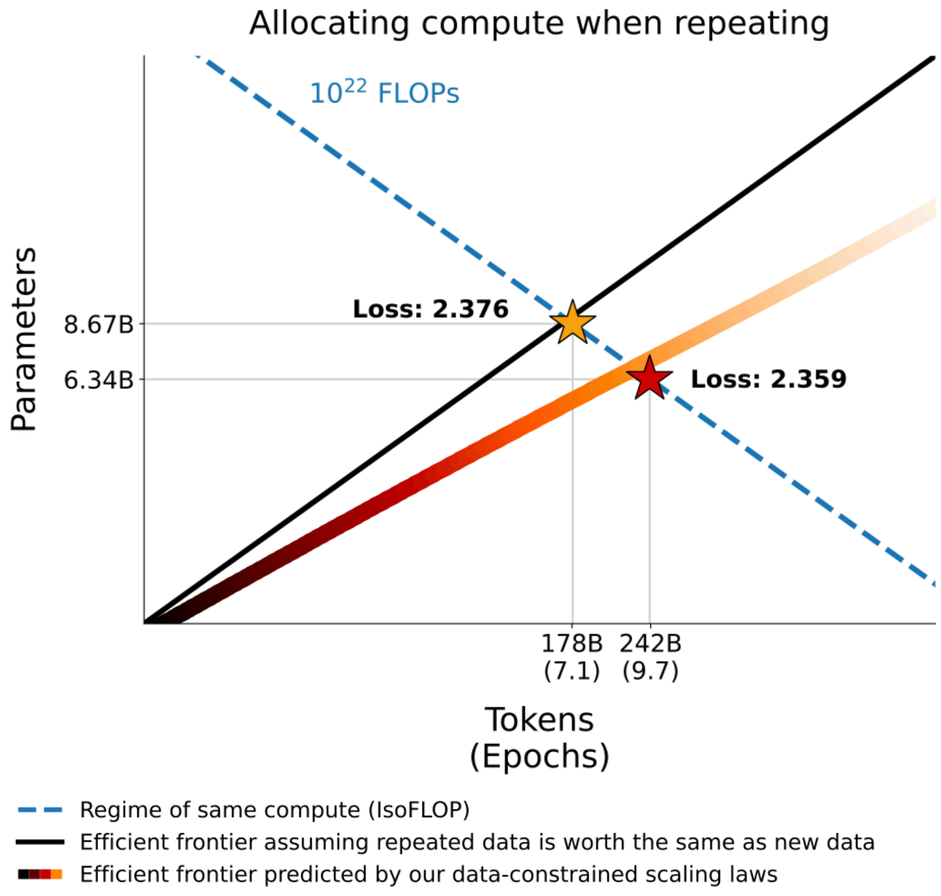
Brandon M. Wood[1,†*], Misko Dzamba[1,*], Xiang Fu[1,*], Meng Gao[1,*], Muhammed Shuaibi[1,*], Luis Barroso-Luque[1], Kareem Abdelmaqsoud[2], Vahe Gharakhanyan[1], John R. Kitchin[2], Daniel S. Levine[1], Kyle Michel[1], Anuroop Sriram[1], Taco Cohen[1], Abhishek Das[1], Ammar Rizvi[1], Sushree Jagriti Sahoo[1], Zachary W. Ulissi[1], C. Lawrence Zitnick[1,†]

[1]FAIR at Meta, [2]Department of Chemical Engineering, Carnegie Mellon University
*Co-first Author, †Co-corresponding Author

Table 8 Power Law Coefficients determined from fitting Equations 4 and 6. Error bounds are determined by bootstrap sampling 1000 times and taking the 10th and 90th percentile values, quoted in brackets.

Parameter	Dense	MoLE
α	0.61 (0.57,0.65)	0.56 (0.49,0.59)
β	0.39 (0.35, 0.43)	0.44 (0.39, 0.43)
α/β	1.56	1.27

Data quality changes optimal scaling



Muennighoff *et al.*, **NeurIPS**, 2023.

Approach	Coeff. a where $N_{\text{opt}}(M_{\text{opt}}) \propto C^a$	Coeff. b where $D_{\text{opt}} \propto C^b$	α/β
OpenAI (OpenWebText2)	0.73	0.27	-
Chinchilla (MassiveText)	0.49	0.51	1.22
Ours (Early Data)	0.450	0.550	0.91
Ours (Current Data)	0.524	0.476	0.73
Ours (OpenWebText2)	0.578	0.422	

Better data quality improves data scaling

Table 4 | Coefficients of model scaling and data scaling vary with training data distribution.

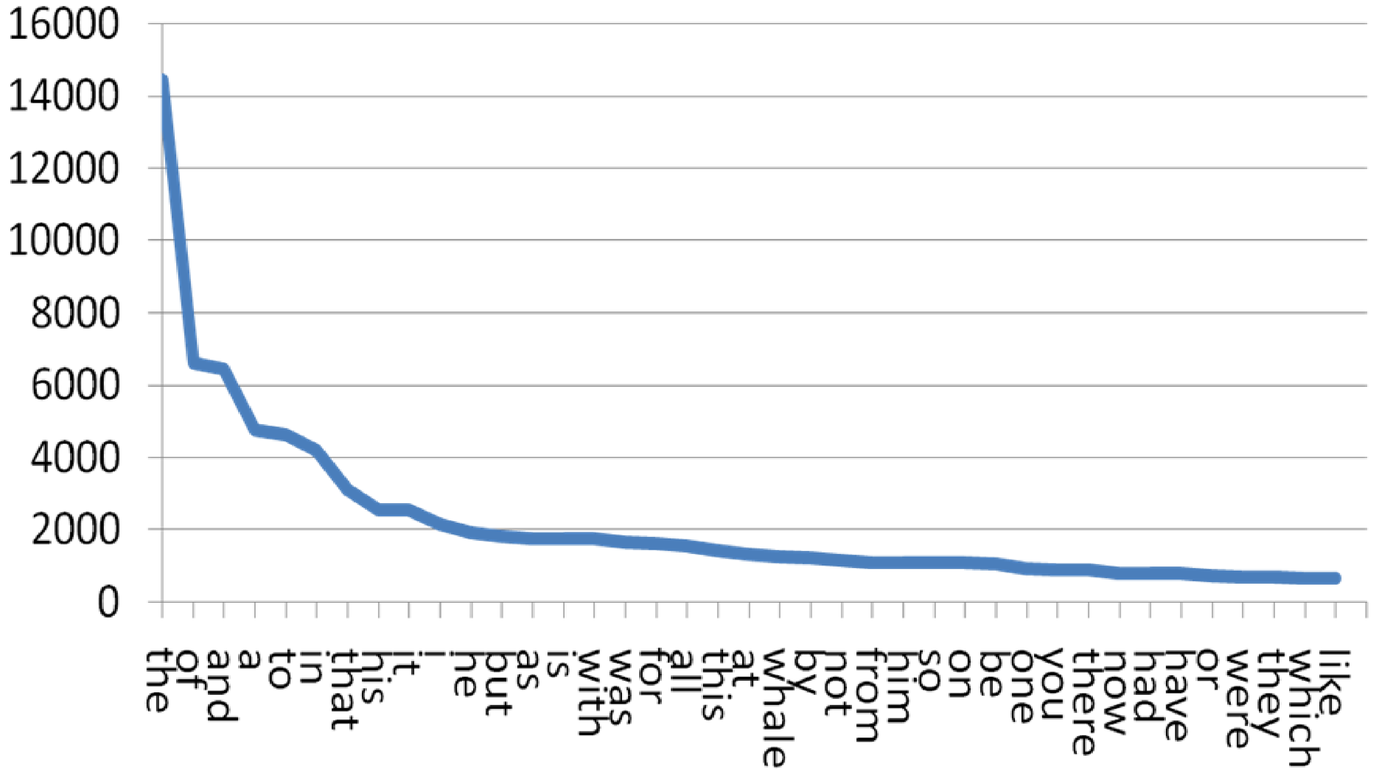
Bi *et al.*, **arXiv**, 2024.

$$L \propto AN^{-\alpha} + BD^{-\beta}$$

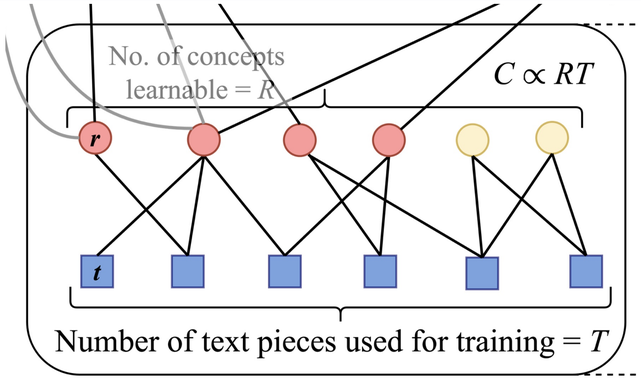
$$D_{\text{opt}} \propto (N_{\text{opt}})^{\frac{\alpha}{\beta}}$$

Concept distributions may shape scaling

- Why are model/data scaling exponents different?
- Data to concept distribution hypothesis
- Zipf-1 language law



Can theory help us generate prospective scaling laws?



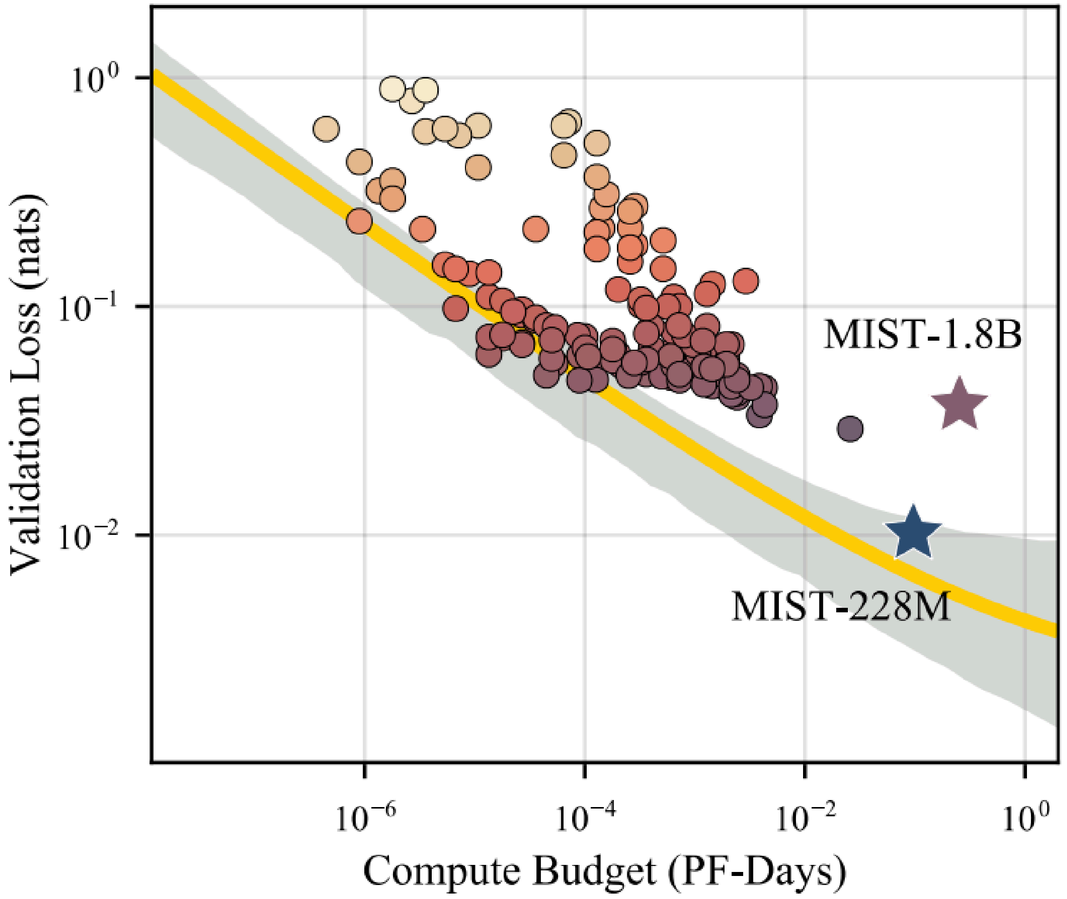
Theorem 1 (Super-linear Data Scaling under Concept Concentration)
Consider a bipartite graph between T text nodes and R concept nodes. Suppose the probability of a text node connecting to a concept $r \in \{1, \ldots, R\}$ follows a Zipf distribution:

$$p_r = \frac{r^{-\alpha/\beta}}{H_R^{\alpha/\beta}}, \quad \alpha/\beta \geq 0,$$

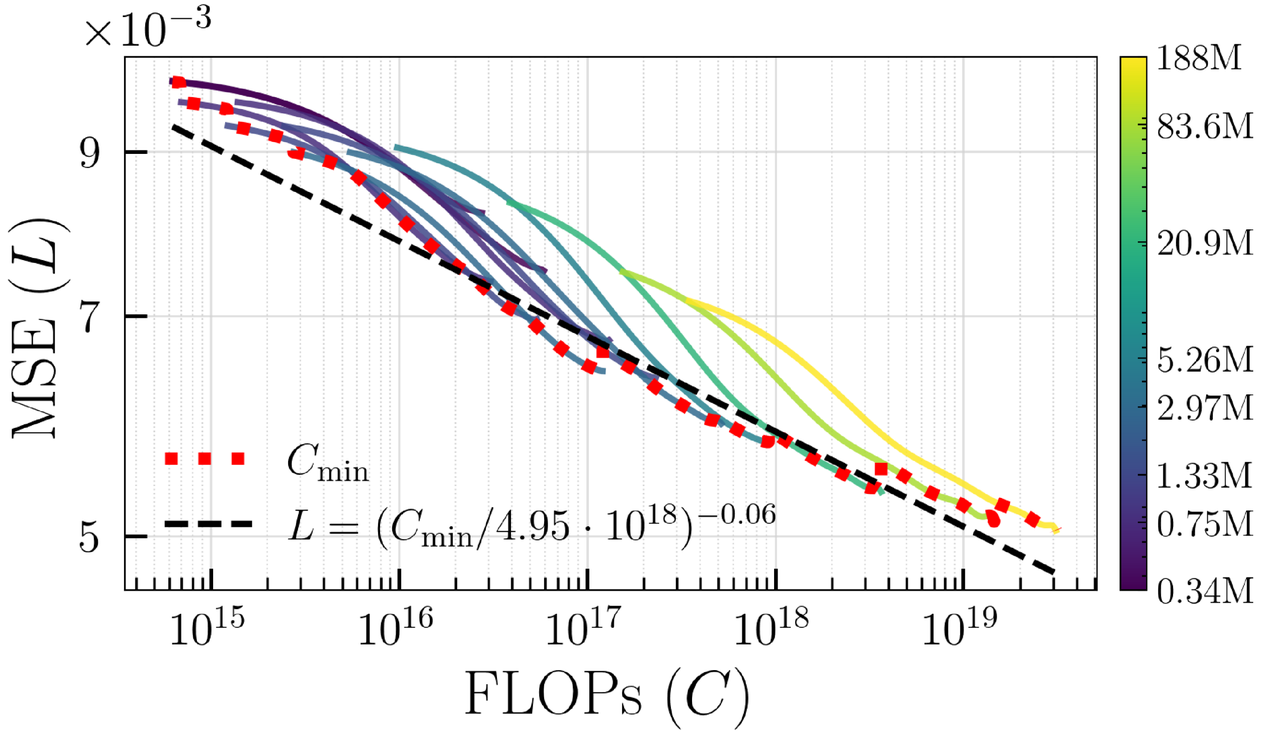
where $H_R^{(\alpha/\beta)} = \sum_{r=1}^{R} r^{-\alpha/\beta}$ normalizes the distribution. Then, to learn $R(1-\delta)$ concepts via iterative decoding, up to any $\delta \in (0,1)$ tolerance, the number of required text nodes T must grow faster than R, satisfying

$$T \propto \begin{cases} R, & 0 \leq \alpha/\beta < 1 \\ R^{\alpha/\beta}, & \alpha/\beta > 1 \end{cases} \implies D \propto \begin{cases} N, & 0 < \alpha/\beta < 1 \\ N^{\alpha/\beta}, & \alpha/\beta > 1 \end{cases}$$

Small runs can guide large runs



Wadell et al., **arXiv**, 2025.

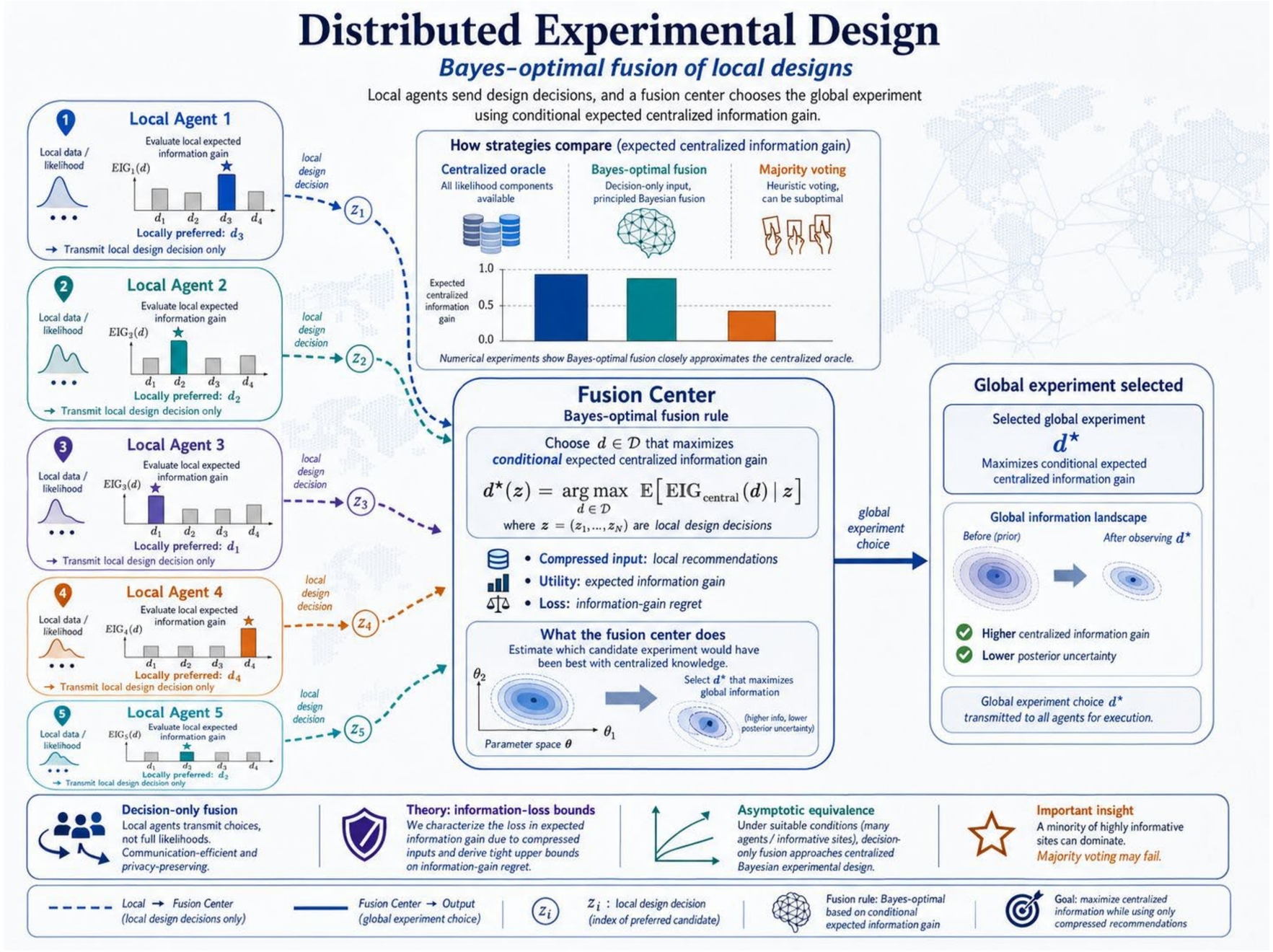


Park et al., **arXiv**, 2025.

Experimental design for training foundation models

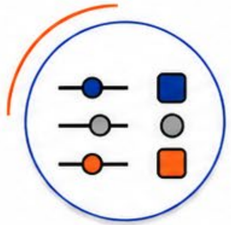
Generalization rather than information gain

[Nagananda, Varshney, and Varshney]



Summary

Scaling laws can inform SciFM design



Domains & datasets change scaling and constraints

Can small runs and theory inform scaling?

Downstream performance is the real story