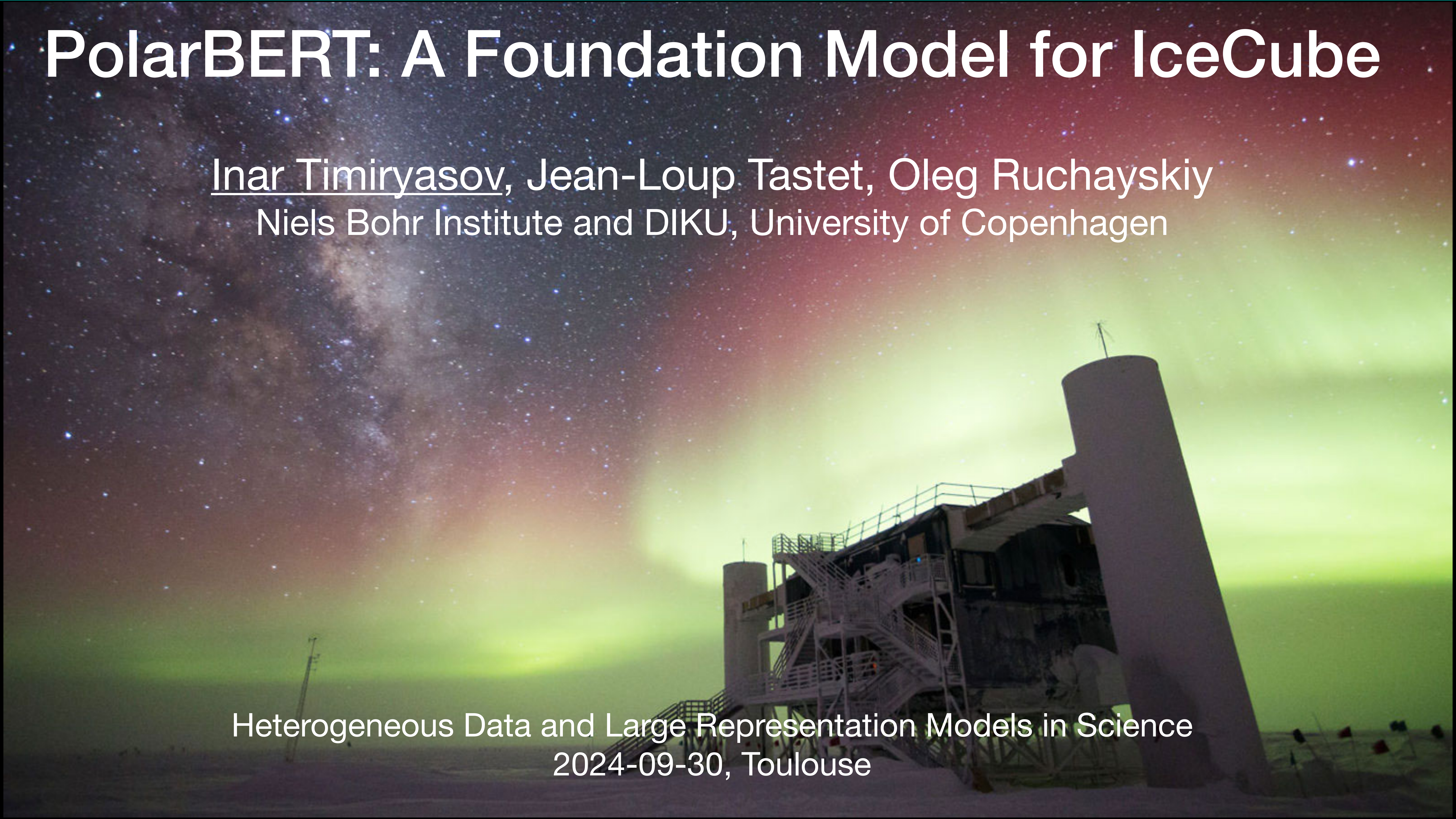
Co-financed by the Connecting Europe
Facility of the European Union

INTERACTIONS

UNIVERSITY OF
COPENHAGEN

PolarBERT: A Foundation Model for IceCube

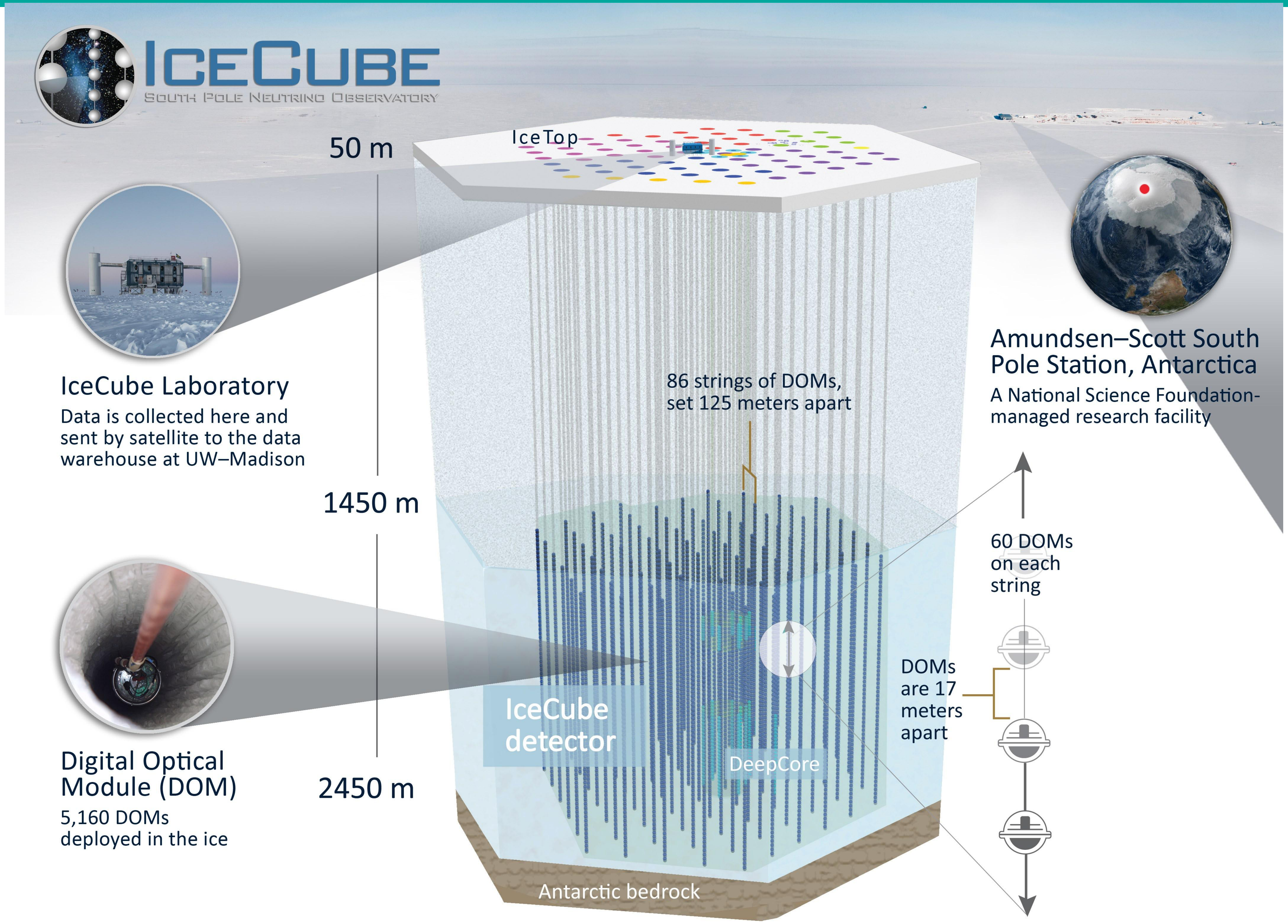
Inar Timiryasov, Jean-Loup Tastet, Oleg Ruchayskiy
Niels Bohr Institute and DIKU, University of Copenhagen



Heterogeneous Data and Large Representation Models in Science
2024-09-30, Toulouse

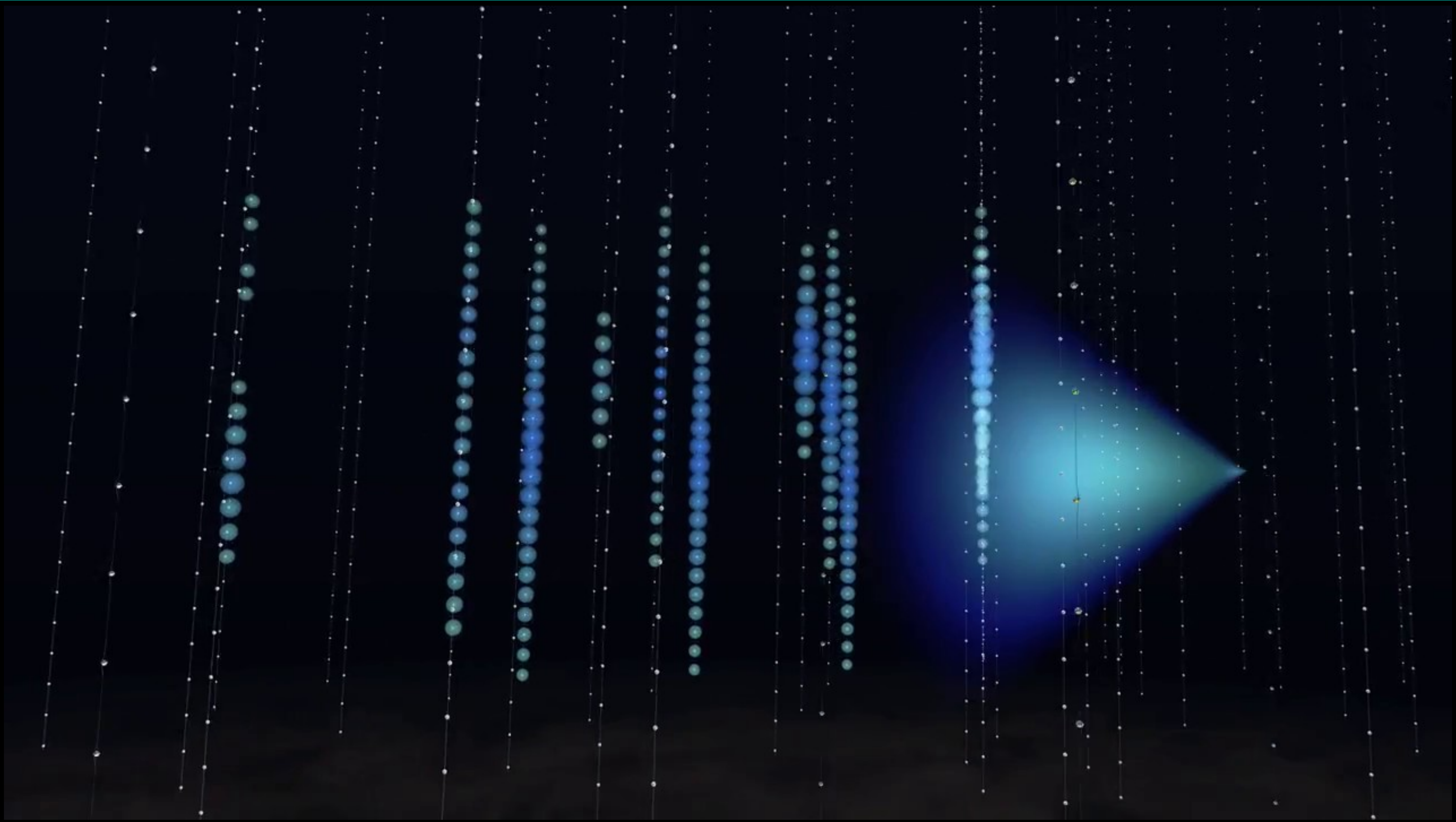
IceCube

- Neutrino telescope
- Located at the South Pole
- Detector volume: 1 cubic kilometer
- Oftentimes observes through Earth
- Public dataset from Kaggle Competition 130 million events



IceCube event

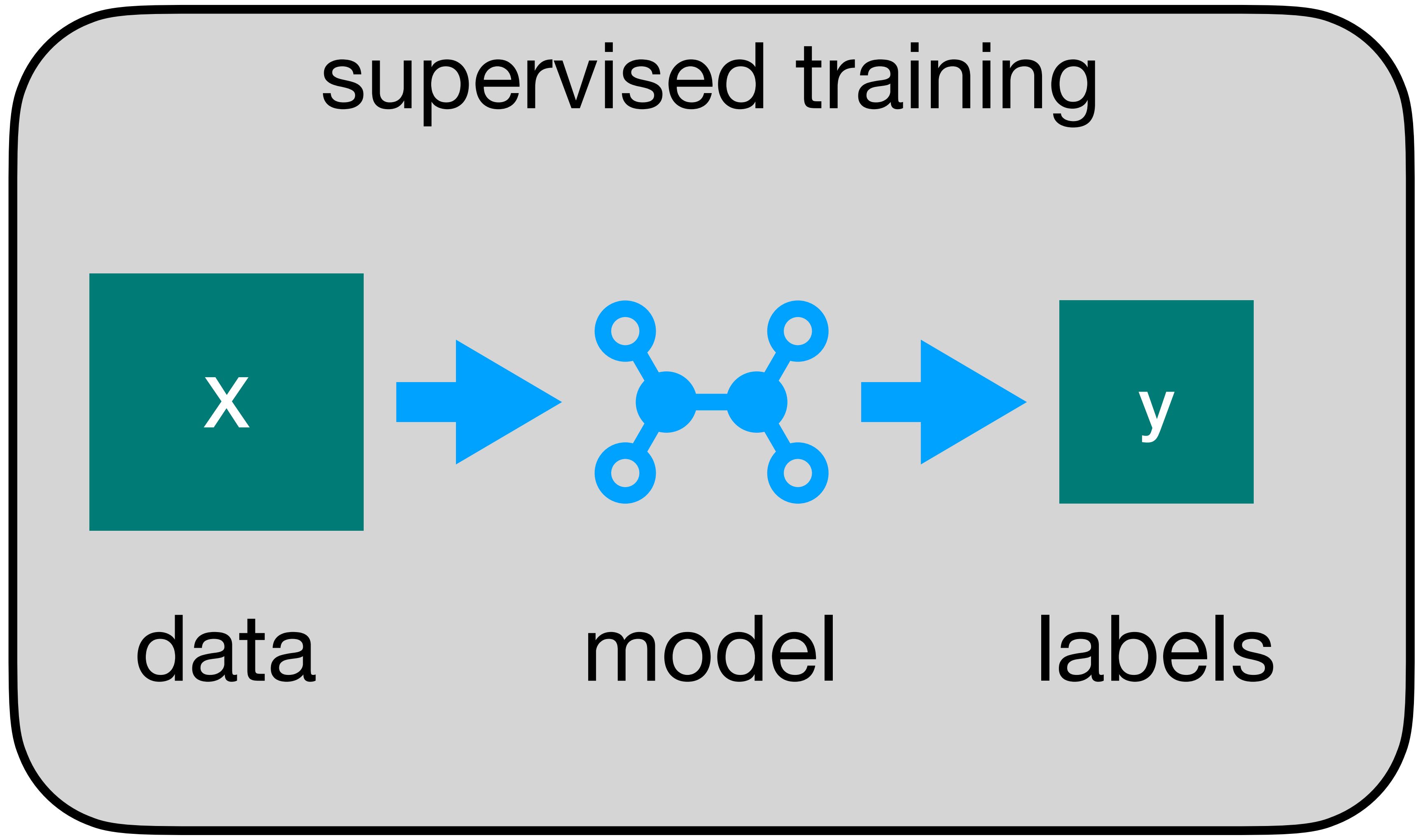
https://youtu.be/OXSqiPLn9CM?si=nnvKH0WpJgEWRn56

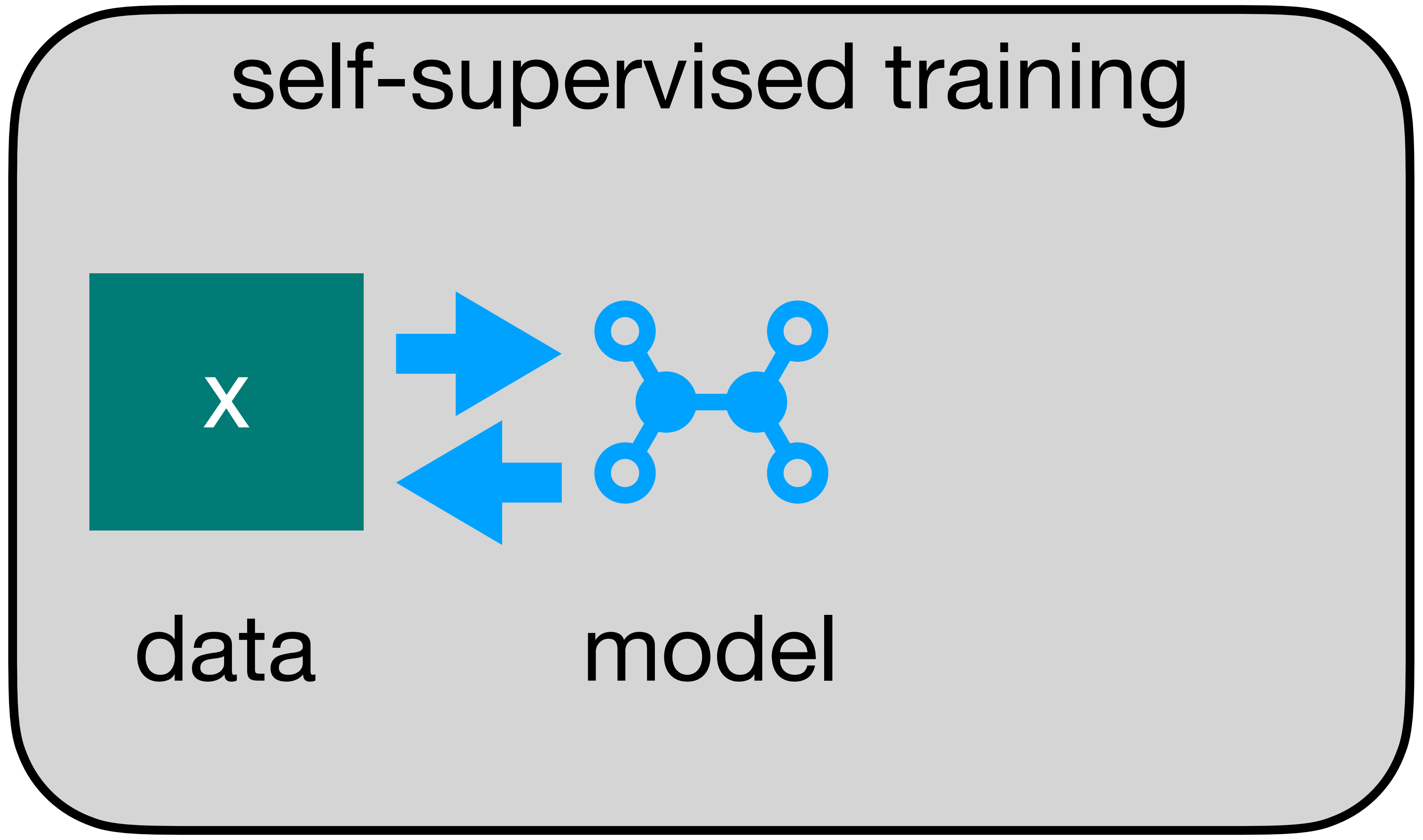


What do we mean by “foundation models”?

- Initially, the term has been coined for models like BERT and GPT-3
  [2108.07258](https://arxiv.org/abs/2108.07258) “On the Opportunities and Risks of Foundation Models”

- Here, by foundational models we mean the models that are pretrained in a self-supervised way and can be fine-tuned for downstream tasks.

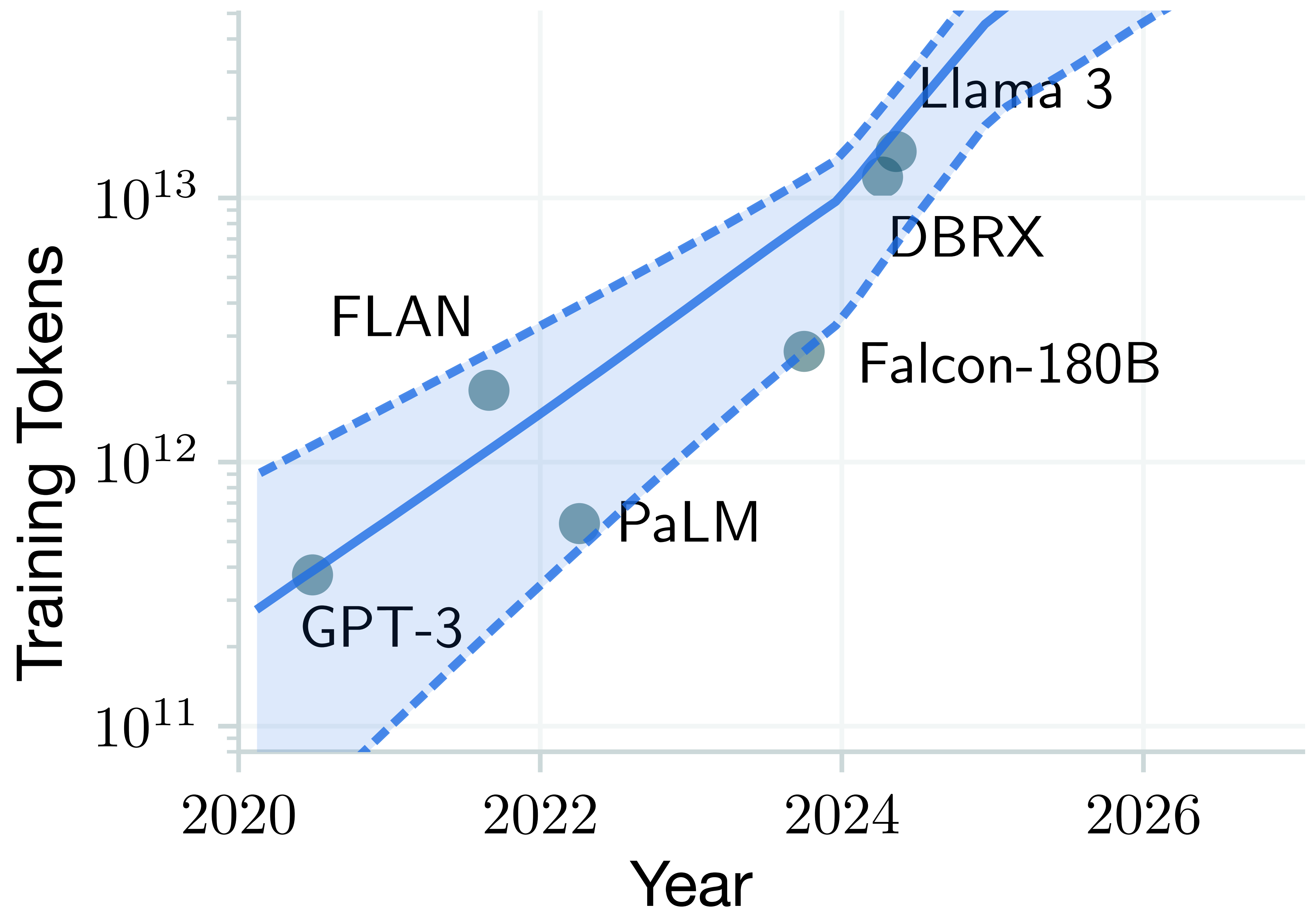




Success of self-supervise training

Outside physics:

- Labeled data is limited
- Unlabeled data is abundant (text, image, video)
- Led to genAI revolution



source:
2211.04325 "Will we run out of data? Limits of LLM scaling based on human-generated data"

● BERT - 3.3B tokens

1810.04805 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

Success of self-supervise training

- No signs of stopping!

Power

Probably the single biggest constraint on the supply-side will be power. Already, at nearer-term scales (1GW/2026 and especially 10GW/2028), power has become the binding constraint: there simply isn't much spare capacity, and power contracts are usually long-term locked-in. And building, say, a new gigawatt-class nuclear power plant takes a decade. (I'll wonder when we'll start seeing things like tech companies buying aluminum smelting companies for their gigawatt-class power contracts.[57])

https://situational-awareness.ai/
Leopold Aschenbrenner, June 2024

MICROSOFT / TECH / SCIENCE

Microsoft wants Three Mile Island to fuel its AI power needs

Photo by Andrew Caballero-Reynolds / AFP via Getty Images

/ Microsoft has signed a 20-year deal to exclusively access 835 megawatts of energy from a nuclear plant.

By Tom Warren, a senior editor and author of *Notepad*, who has been covering all things Microsoft, PC, and tech for over 20 years.

Sep 20, 2024 at 2:23 PM GMT+2

69 Comments (69 New)

https://www.theverge.com/2024/9/20/24249770/

Self-supervise training: Scaling Laws

Performance predictably improves with scale

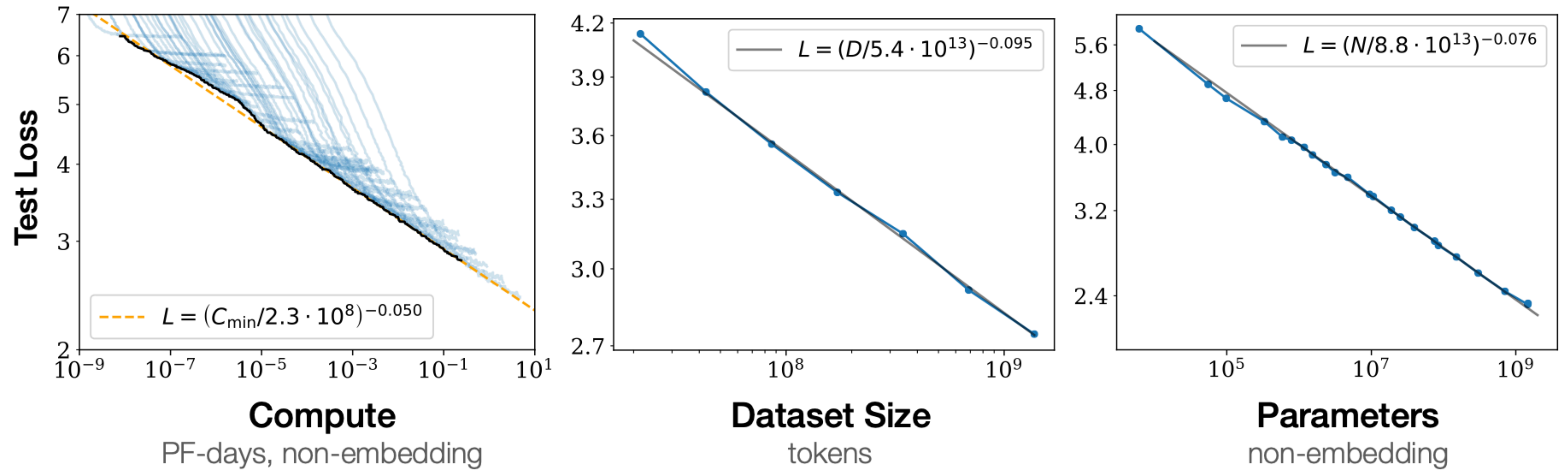


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

https://arxiv.org/pdf/2001.08361
Scaling Laws for Neural Language Models

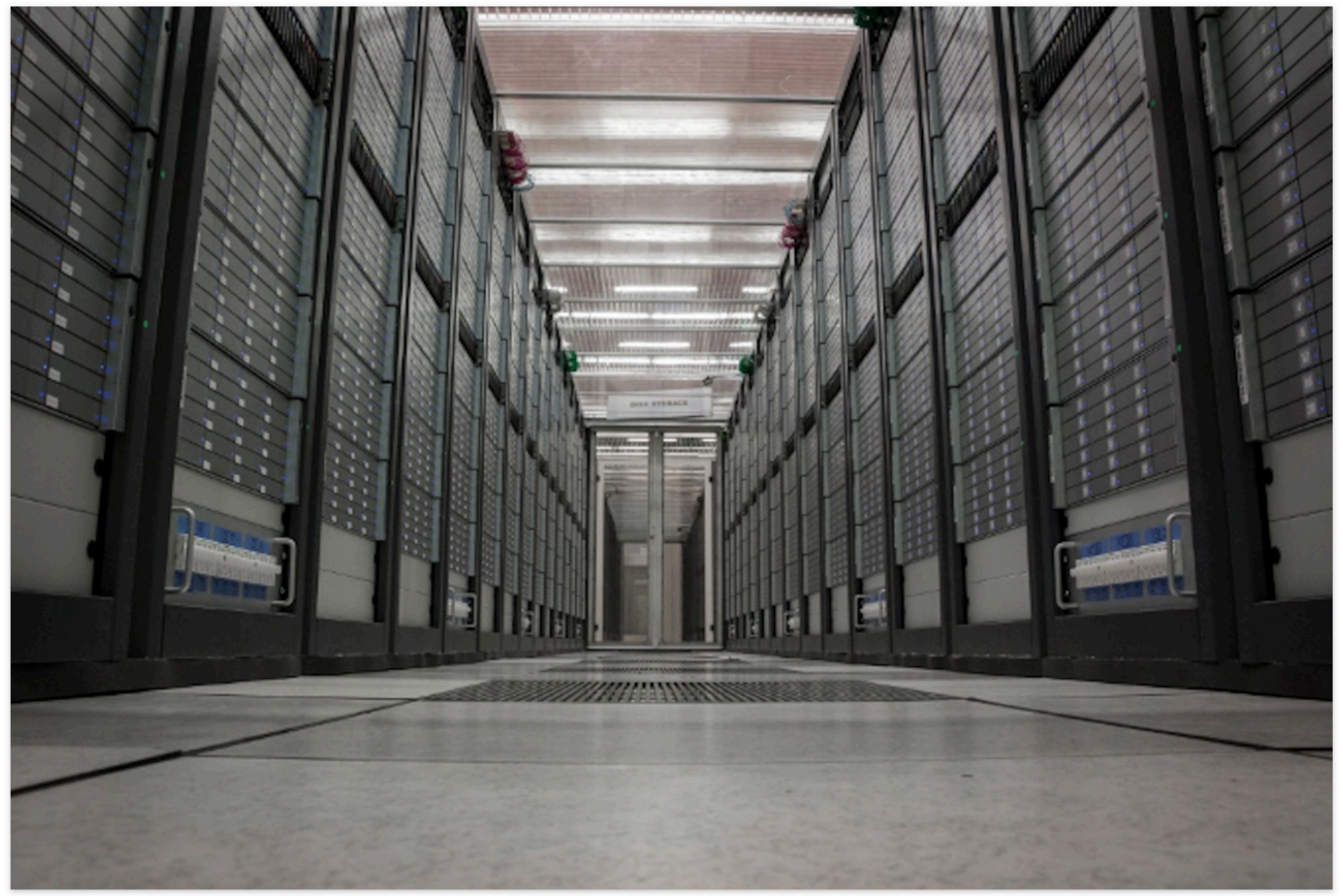
Self-supervised learning in physics

- High quality synthetic data
- Real data is extremely abundant
- Scaling has not been tested yet

An exabyte of disk storage at CERN

CERN disk storage capacity passes the threshold of one million terabytes of disk space

29 SEPTEMBER, 2023 | By Tim Smith



A fraction of the 111 000 devices that form CERN's data storage capacity. (Image: CERN)

source: https://home.cern/news/news/computing/exabyte-disk-storage-cern

Foundation models in particle physics

(a very incomplete list)

- **Pre-training strategy using real particle collision data for event classification in collider physics**
  https://arxiv.org/abs/2312.06909
  Tomoe Kishimoto, Masahiro Morinaga, Masahiko Saito, Junichi Tanaka

- **Finetuning Foundation Models for Joint Analysis Optimization**
  https://arxiv.org/abs/2401.13536
  Matthias Vigl, Nicole Hartman, Lukas Heinrich

- **Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models**
  https://arxiv.org/abs/2401.13537
  Lukas Heinrich, Tobias Golling, Michael Kagan, Samuel Klein, Matthew Leigh, Margarita Osadchy, John Andrew Raine

- **A Language Model for Particle Tracking**
  https://arxiv.org/abs/2402.10239
  Andris Huang, Yash Melkani, Paolo Calafiura, Alina Lazar, Daniel Thomas Murnane, Minh-Tuan Pham, Xiangyang Ju

- **OmniJet-α: The first cross-task foundation model for particle physics**
  https://arxiv.org/abs/2403.05618
  Joschka Birk, Anna Hallin, Gregor Kasieczka

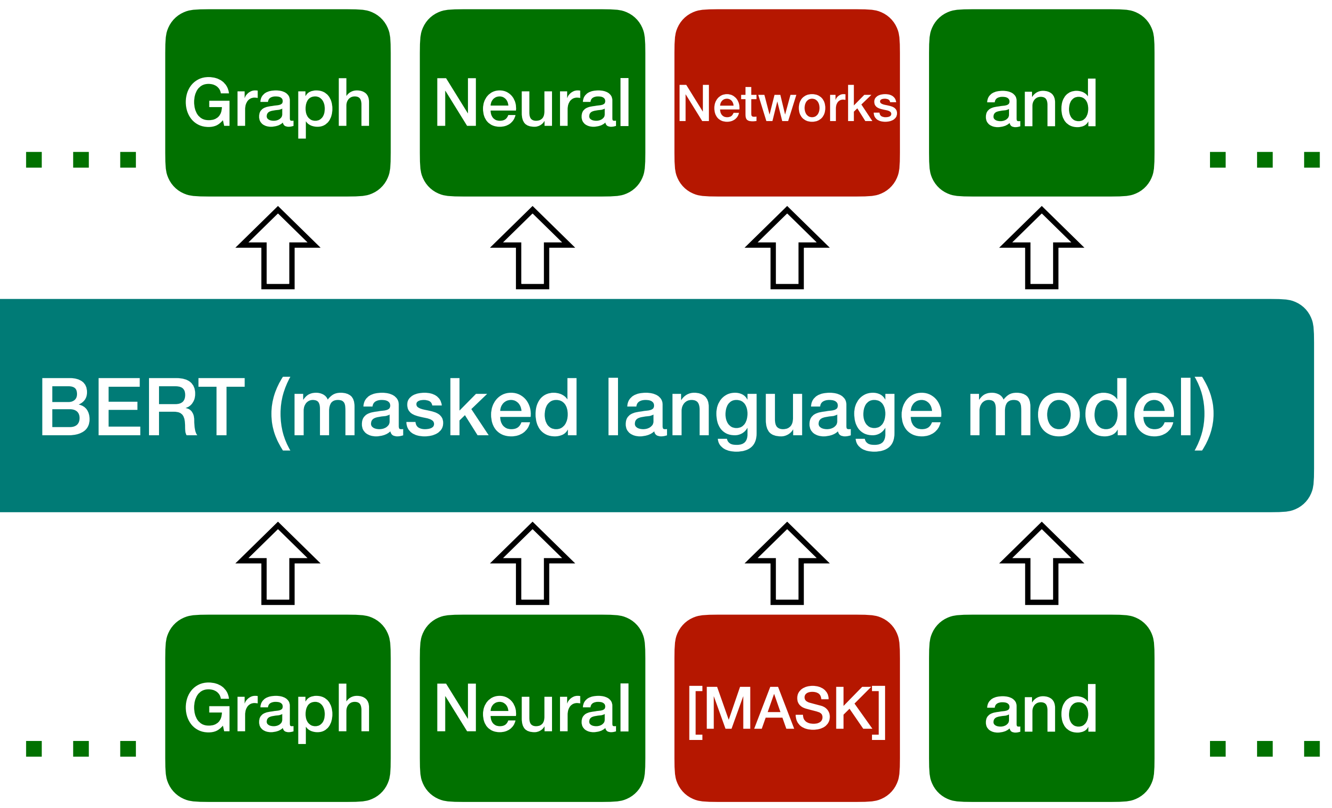
- **Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models**
  https://arxiv.org/abs/2403.07066
  Philip Harris, Michael Kagan, Jeffrey Krupa, Benedikt Maier, Nathaniel Woodward

- **OmniLearn: A Method to Simultaneously Facilitate All Jet Physics Tasks**
  https://arxiv.org/abs/2404.16091
  Vinicius Mikuni, Benjamin Nachman

Challenges of self-supervise learning in particle physics

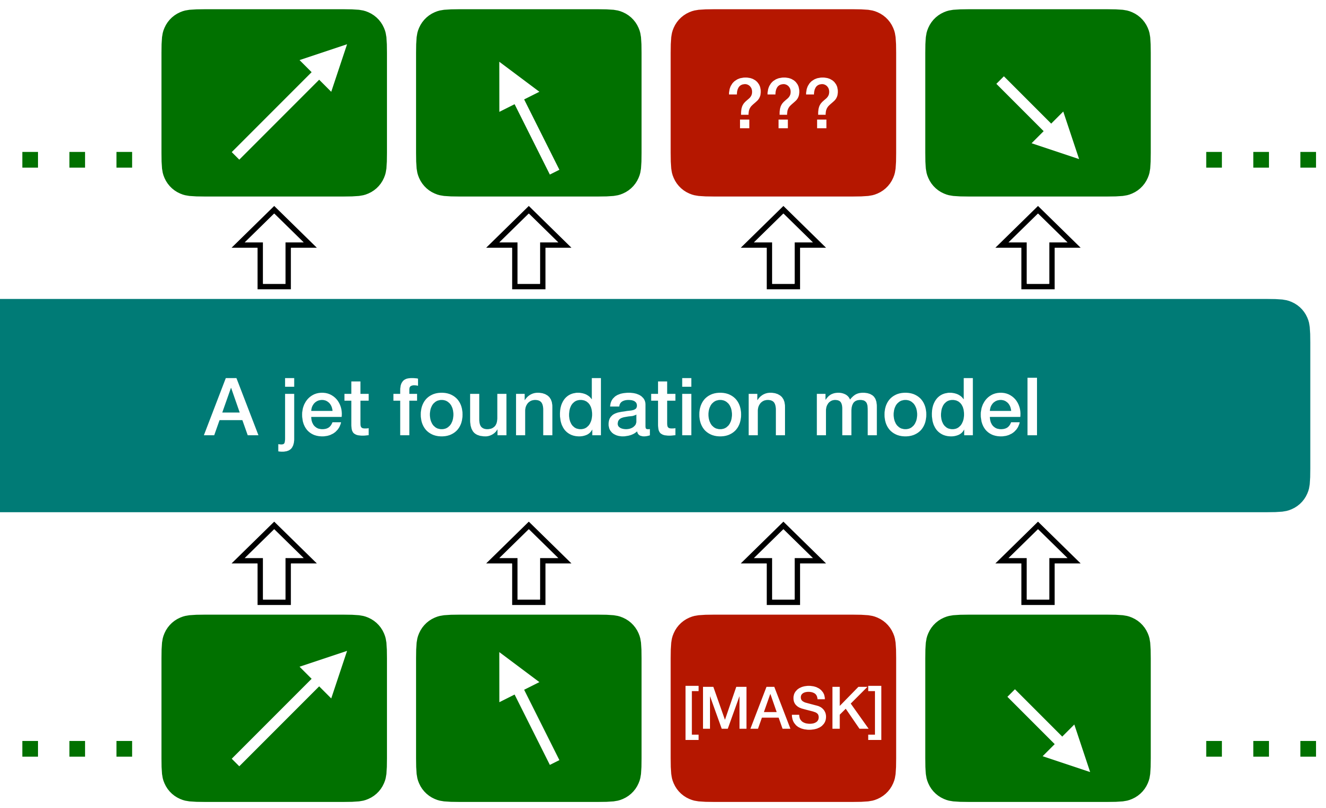
BERT
(Bidirectional Encoder Representations from Transformers)

predict the distribution of a token from a discrete set



A jet foundation model

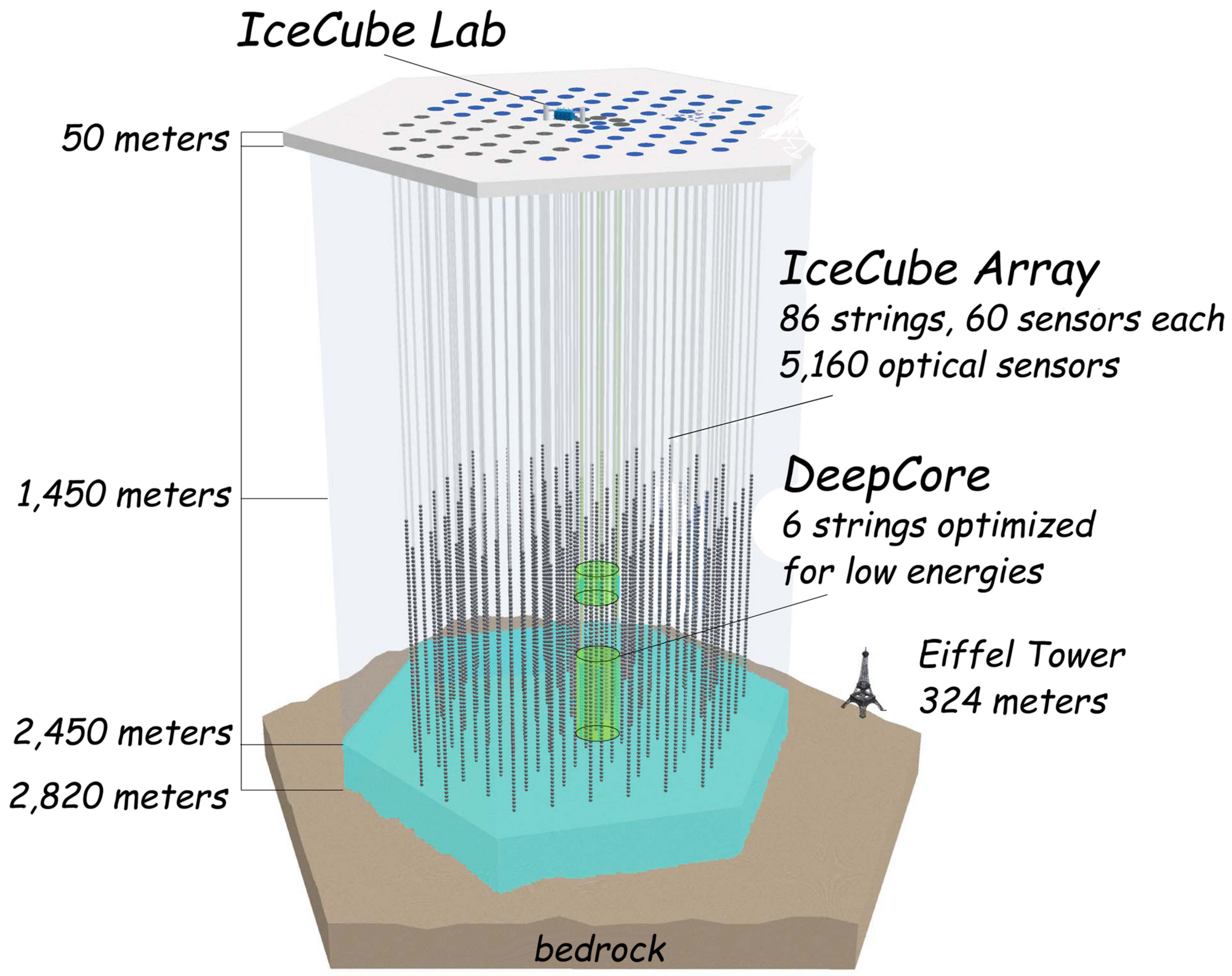
How to predict a continuous 4-vector?



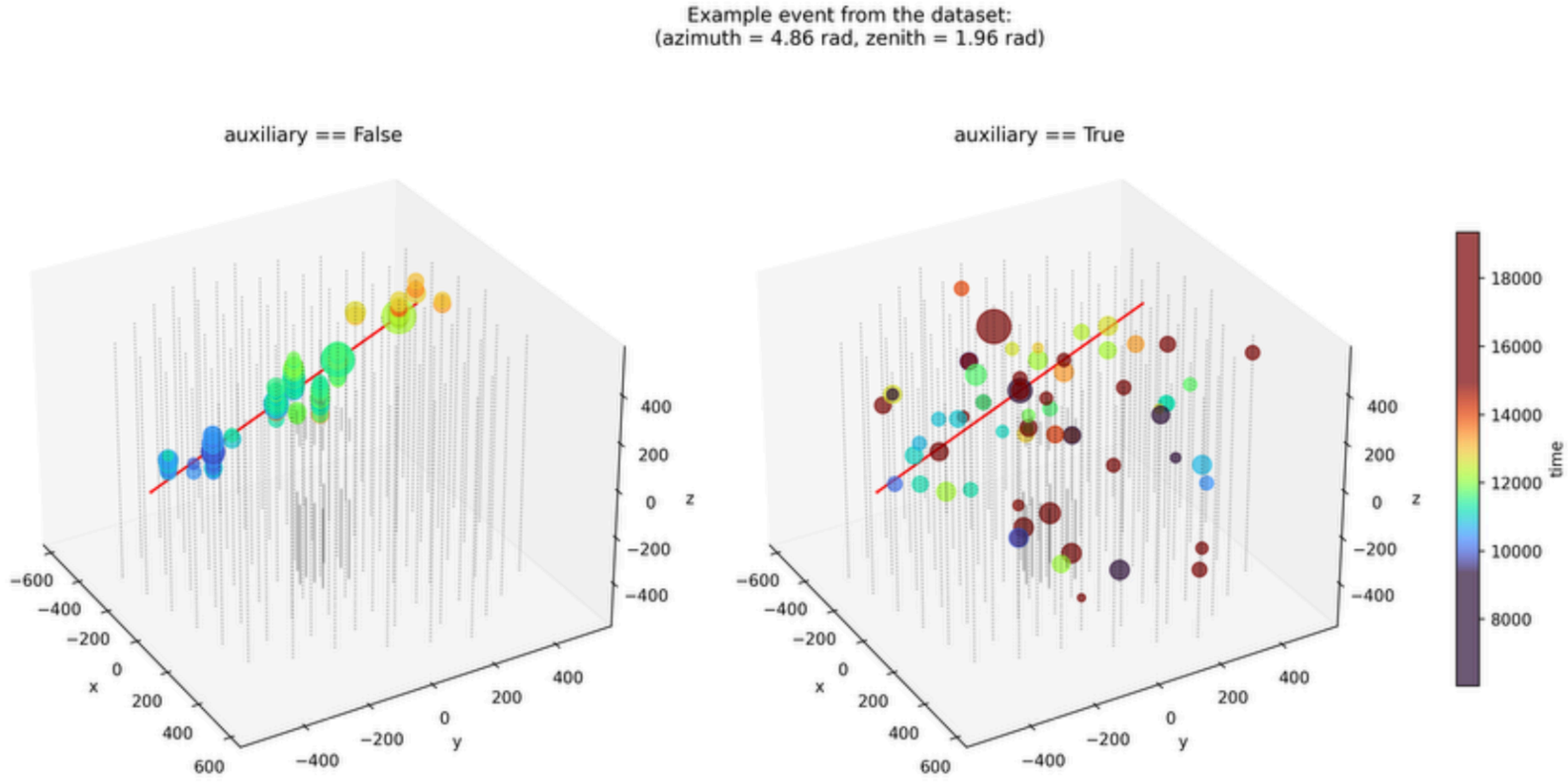
Usually lossy discretization:
- VQ-VAE (2401.13537, 2403.05618)
- pixelization (2402.10239)

Challenges of self-supervise learning in particle physics

- How to predict a continuous 4-vector?
- Usually lossy discretization:
  - VQ-VAE (2401.13537, 2403.05618)
  - pixelization (2402.10239)
- How to sort 4-vectors?
- IceCube
  - 5160 DOMs — natural "tokenization"
  - Pulses have timestamps



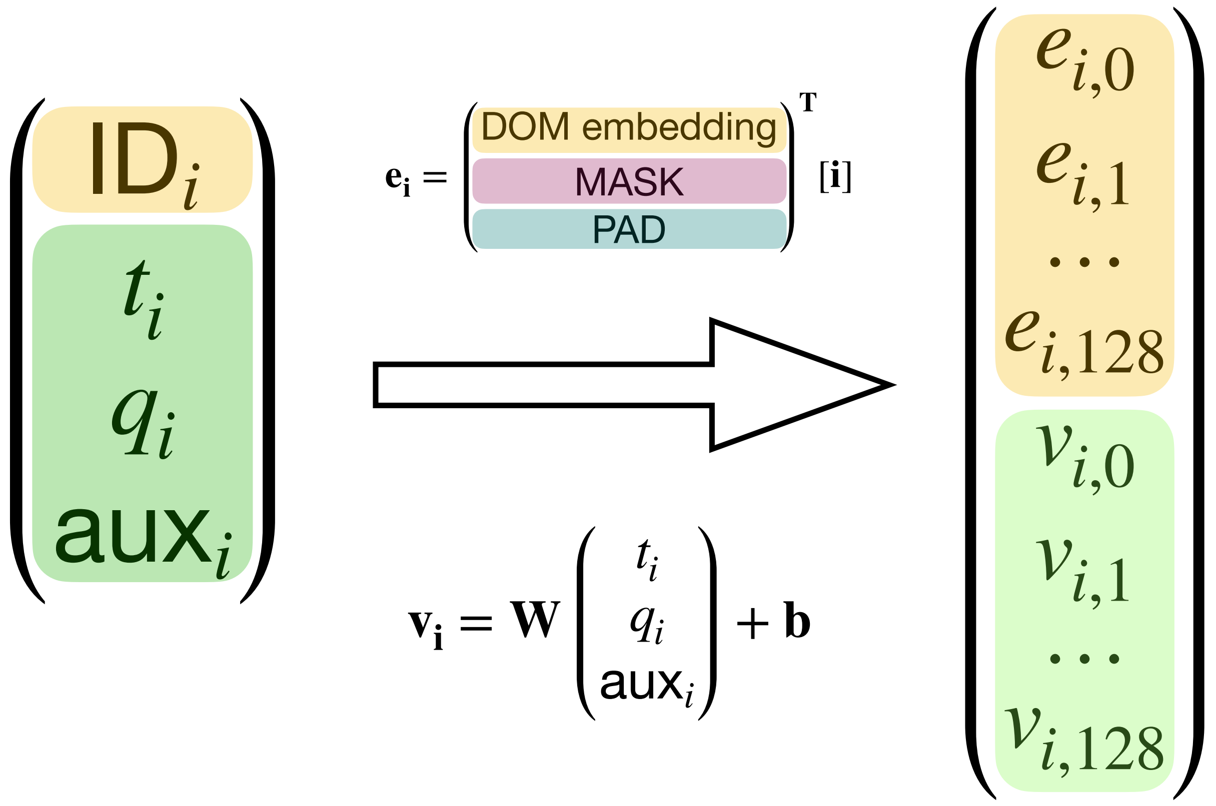
IceCube event



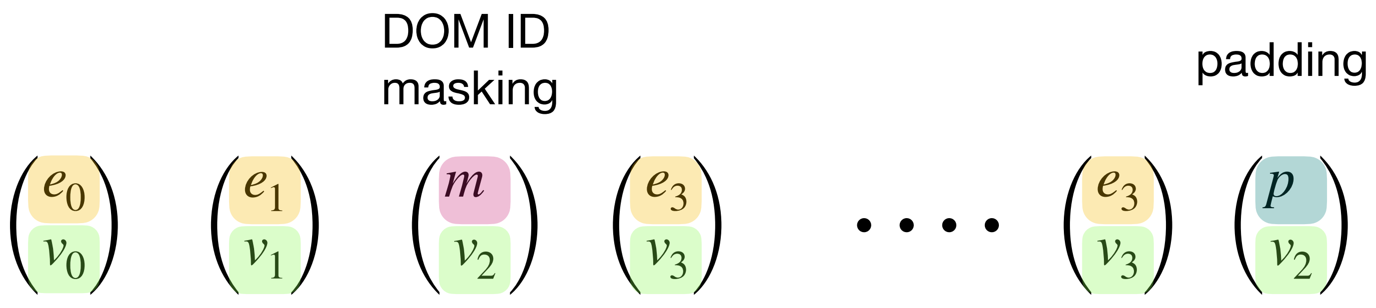
IceCube Embedding

pulses (arranged by time)

pulse embedding

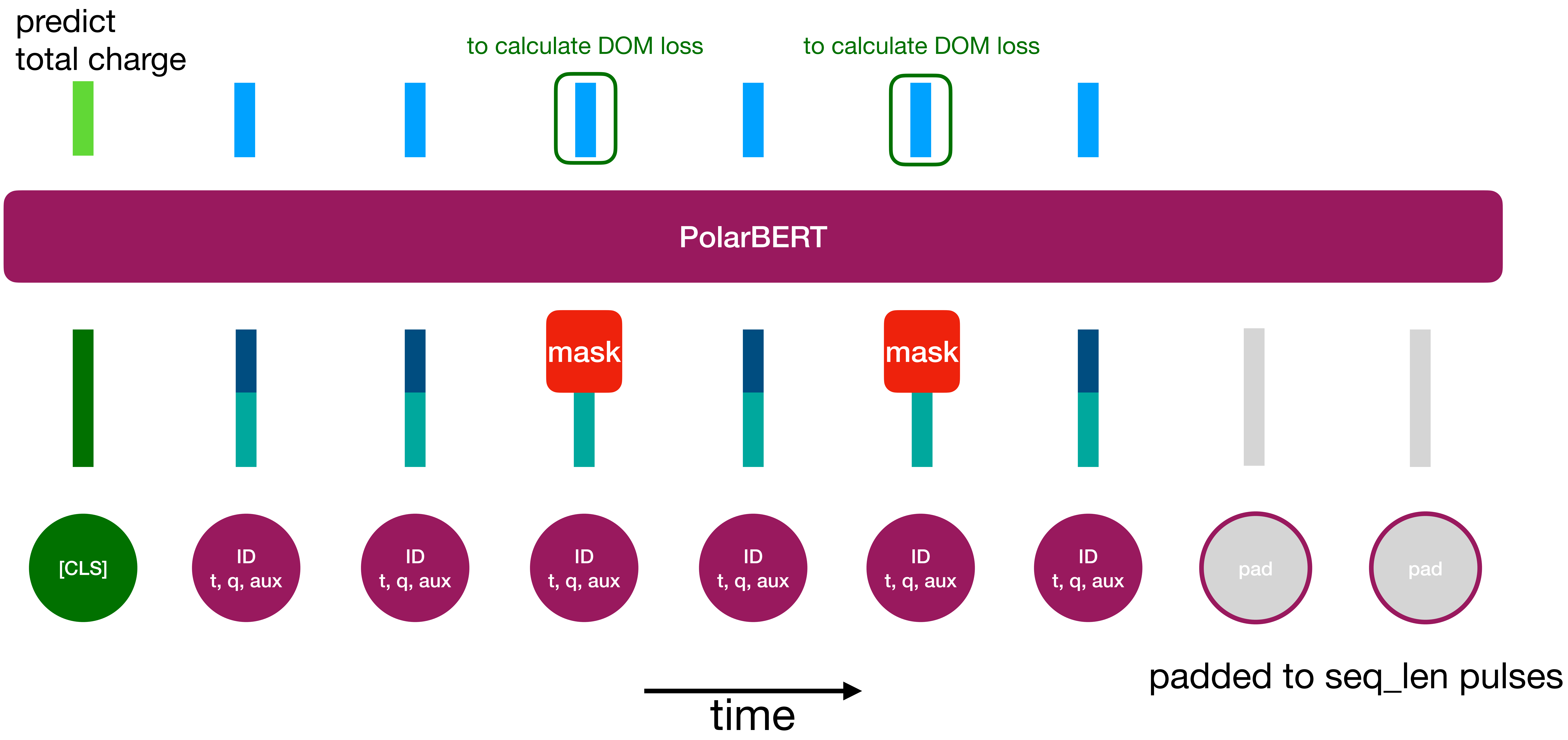


No position data!



time-series (padded to fixed length)

Pretraining

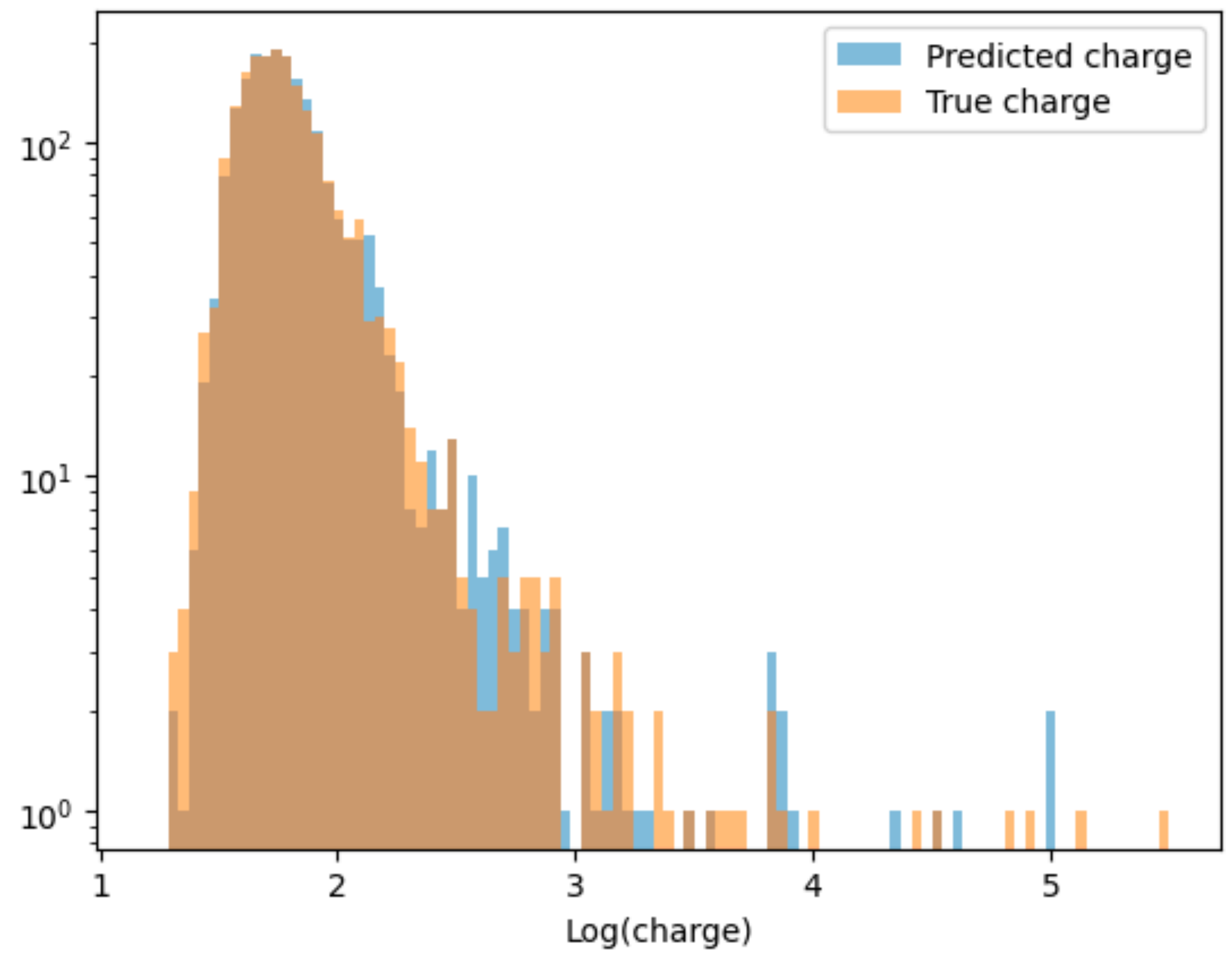


Pretraining: DOM loss

- The detection process is inherently stochastic
- We cannot predict the next DOM with certainty
- Similarly to LLMs, we use cross-entropy
  (but other option are possible: Earth Mover's Distance, Chamfer distance)
- DOM-loss: $L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_i)$, the sum over N masked doms
- Use only aux=false (HLC) pulses! aux=true pulses are impossible to predict.

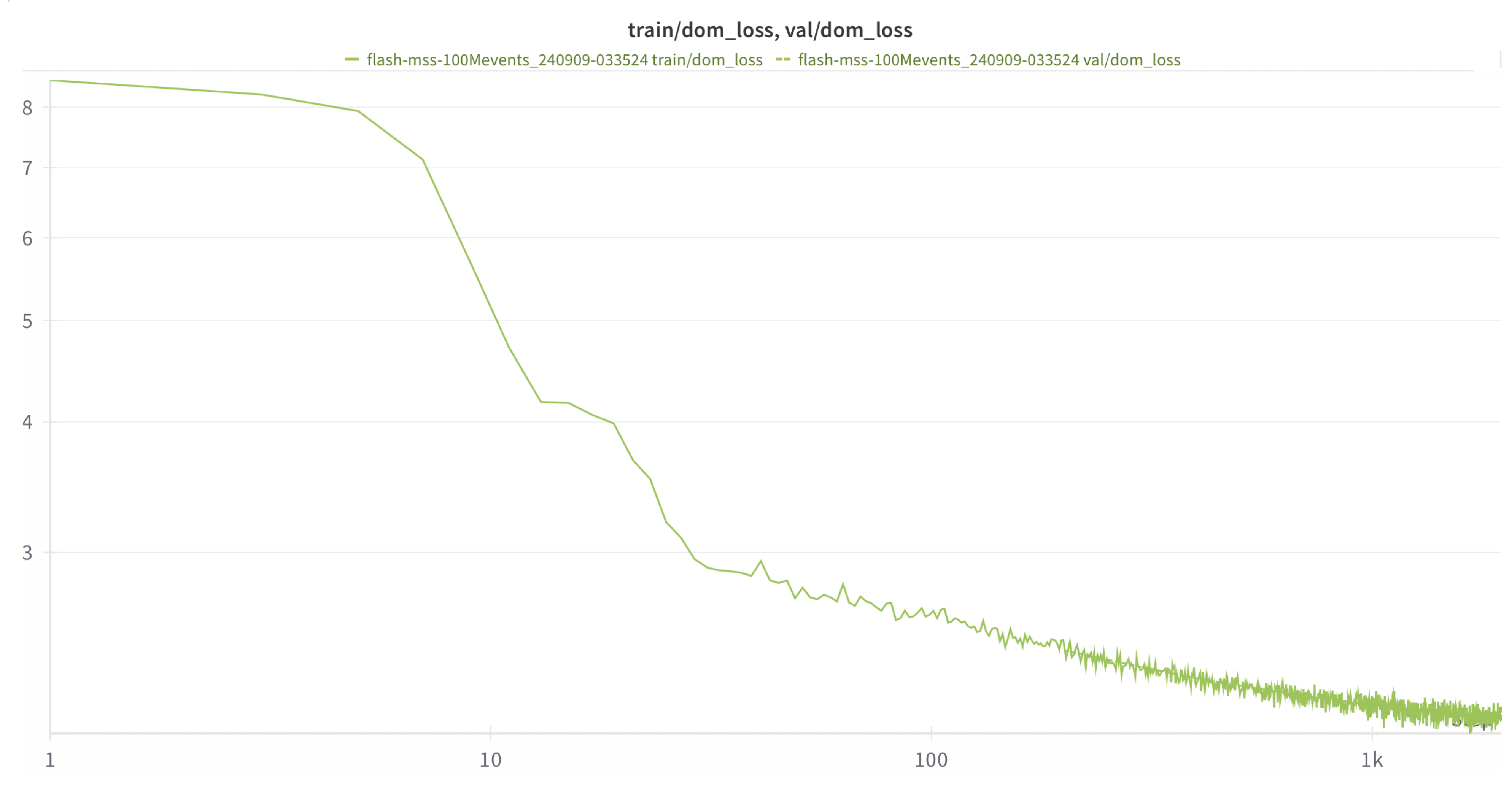
Pretraining: regression loss

- The model has to learn how to collect useful information in [CLS] embedding for the future use on downstream tasks.
- We need some feature that is not directly accessible to the model, but can be obtained from the data (no labels)
- Candidates: the total charge of the event, center of charge
- We subsample the events, and the charge is provided as a log
- Charge prediction loss: MSE($\log$(total charge))



PolarBERT: Foundation Model For IceCube

- Backbone: transformer (could be GRU, Mamba)
- Pretraining:
    - Subsample events to seq_len (currently 128)
    - input: (DOM embedding) $\bigoplus$ (projection of features)
    - loss function = DOM-loss + $\lambda \times$ charge-prediction-loss
- Fine-tuning for downstream tasks
- IceCube kaggle MC data for both pretraining and finetuning (studies using real data can be only published by the collaboration)

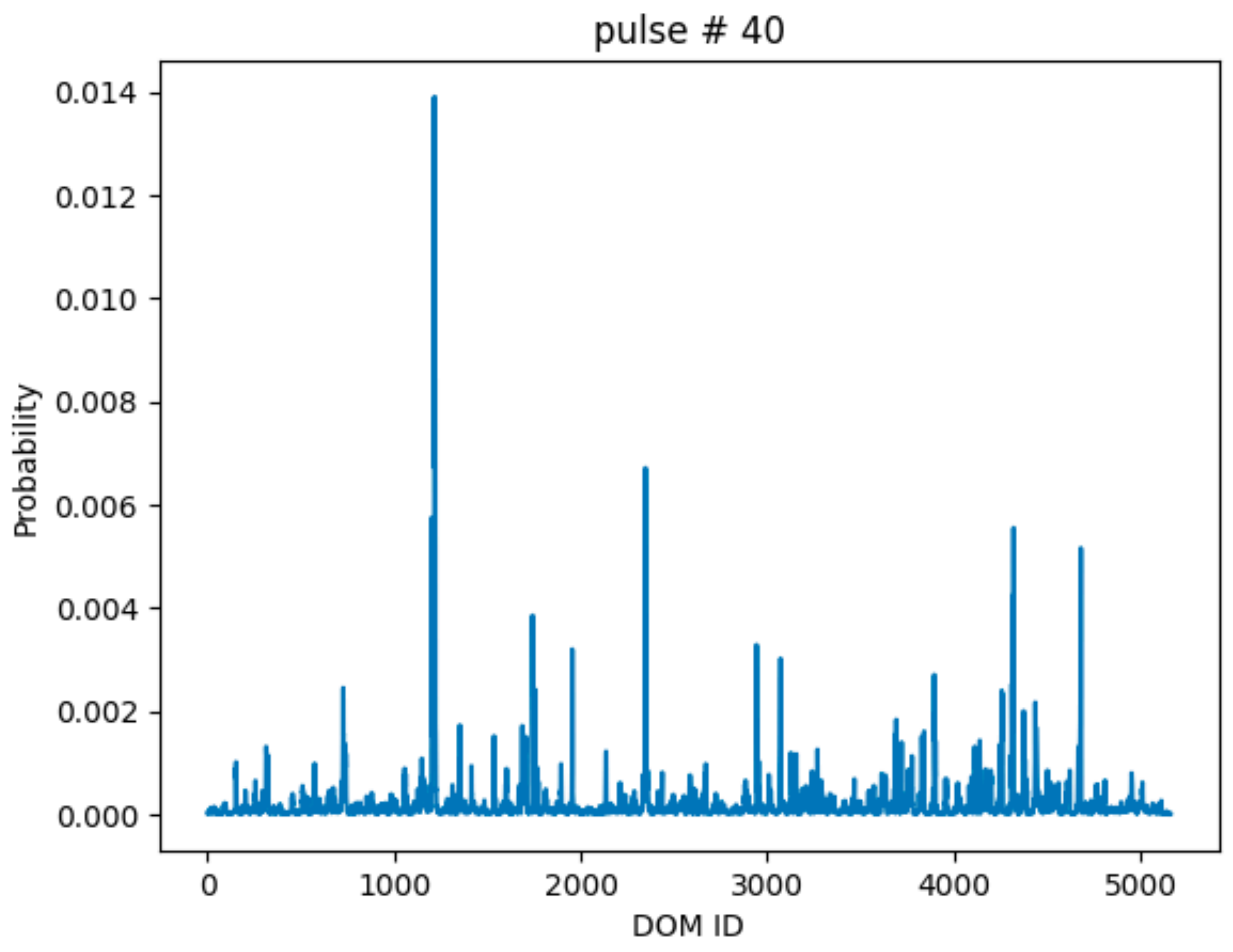


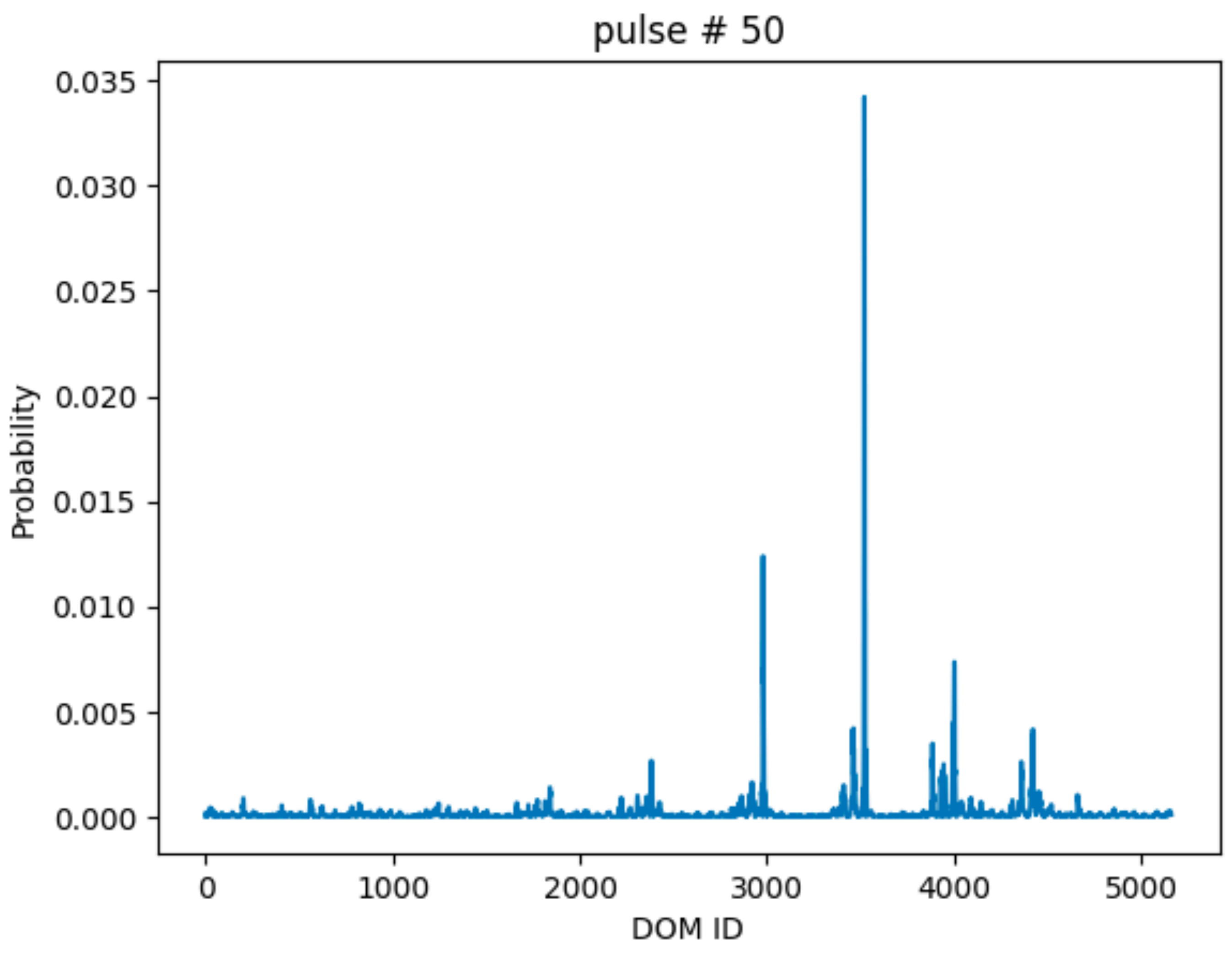
100 million events

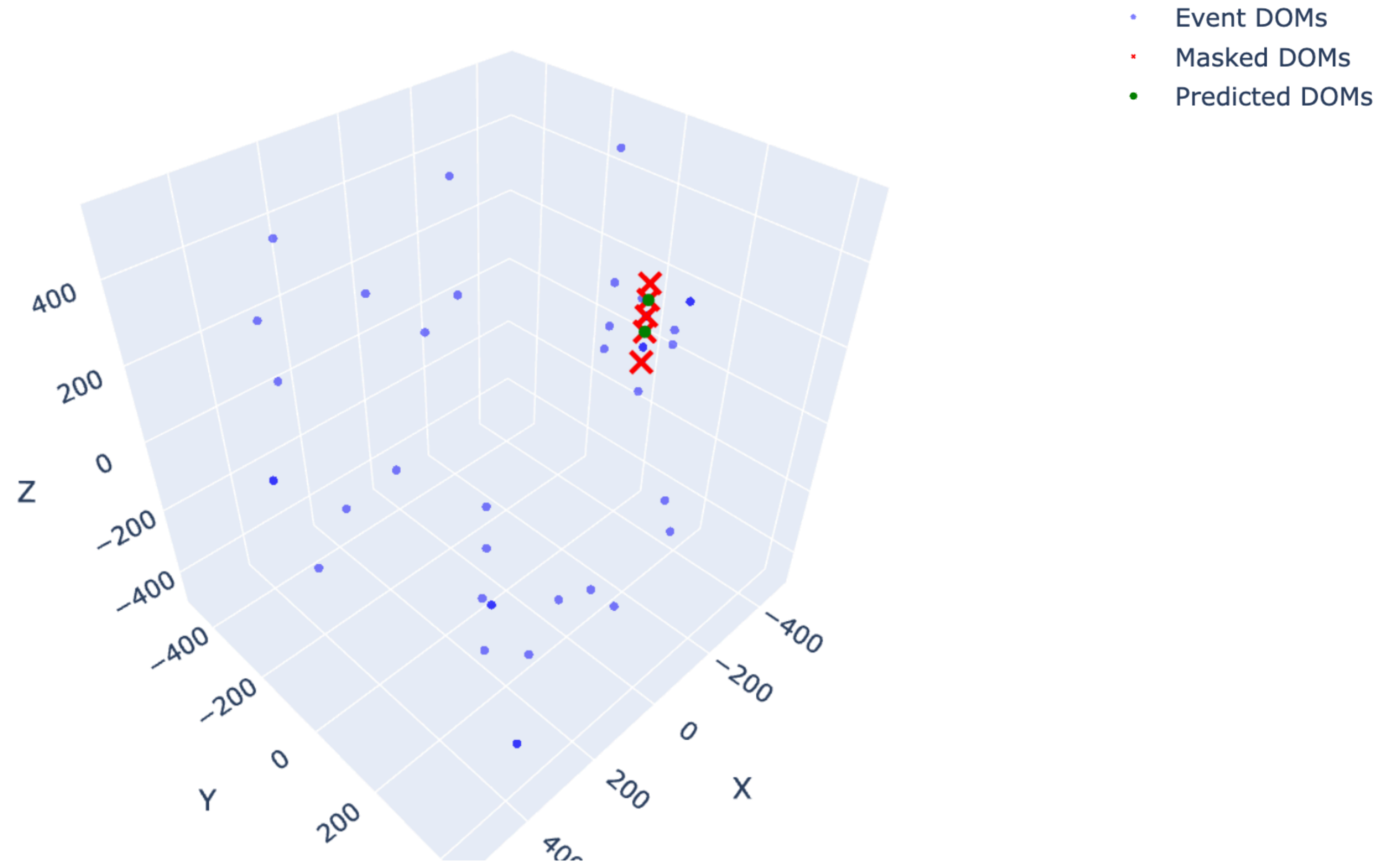
BERT: 3,300M tokens
PolarBERT: 127,000M “tokens”
(100M events x 127 pulses)

Interpreting the DOM Loss

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_i)$$



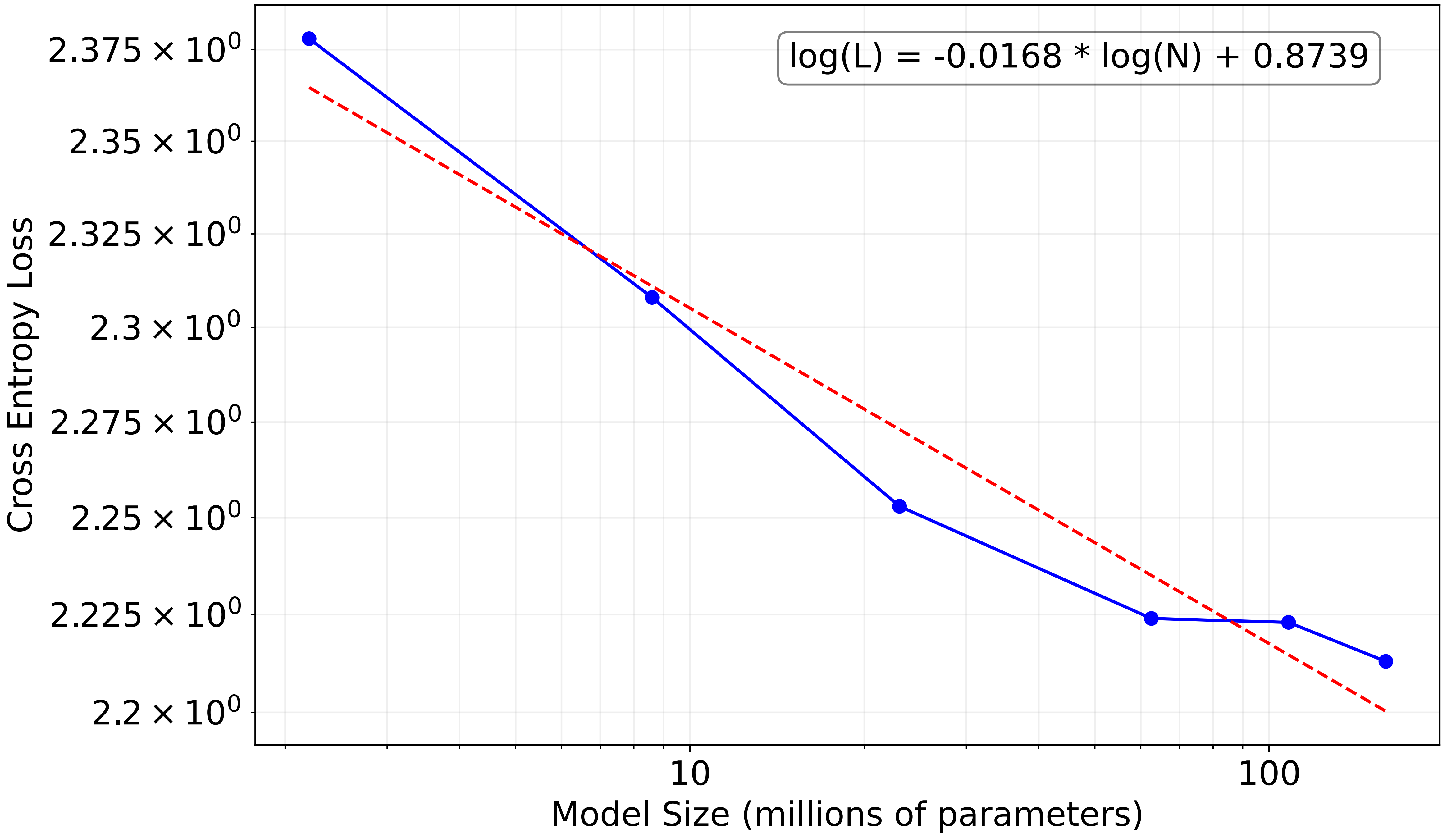




some uncertainty about the string and the DOM

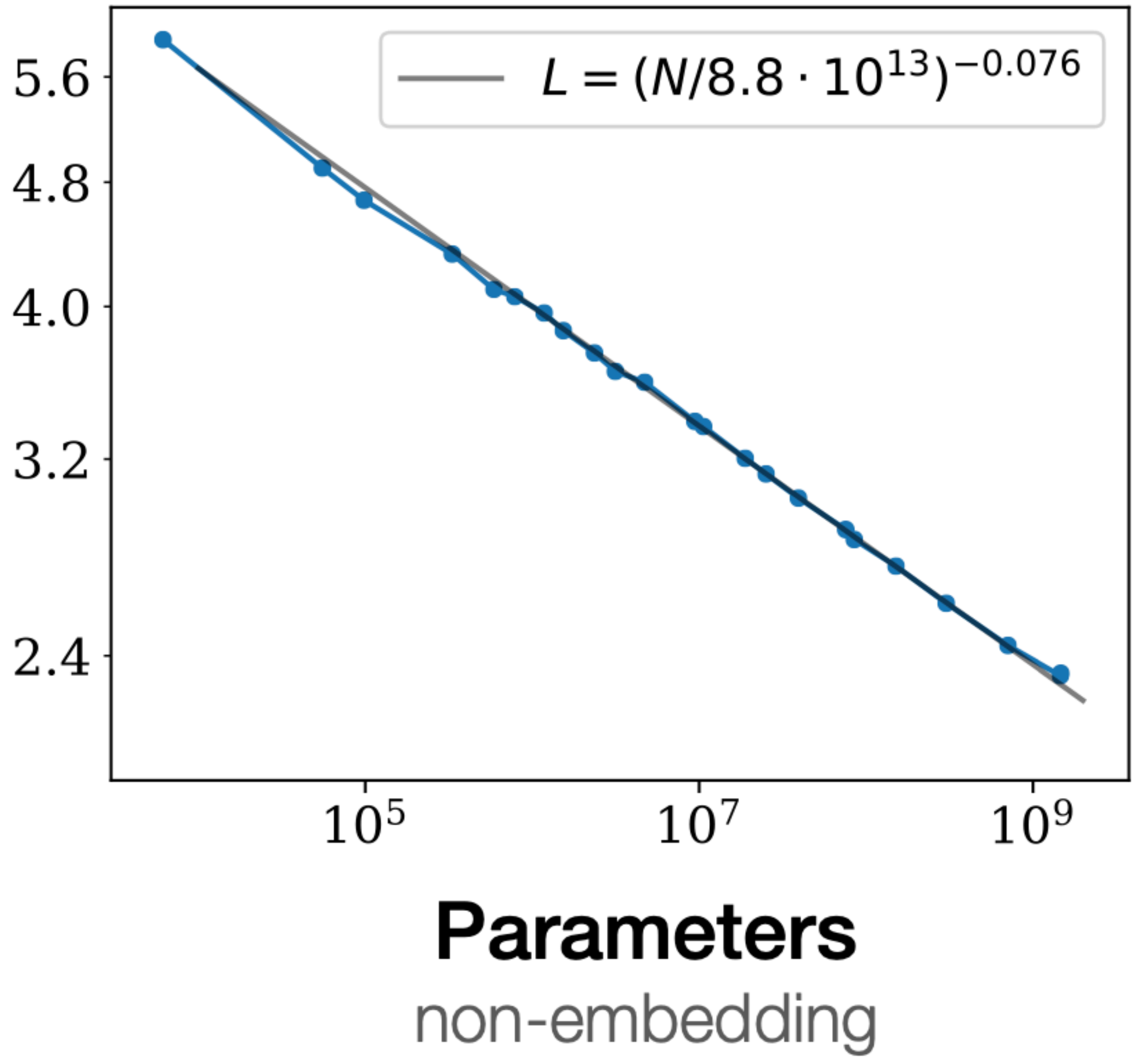
Model Size Scaling

PolarBERT



Models trained on 10M neutrino events

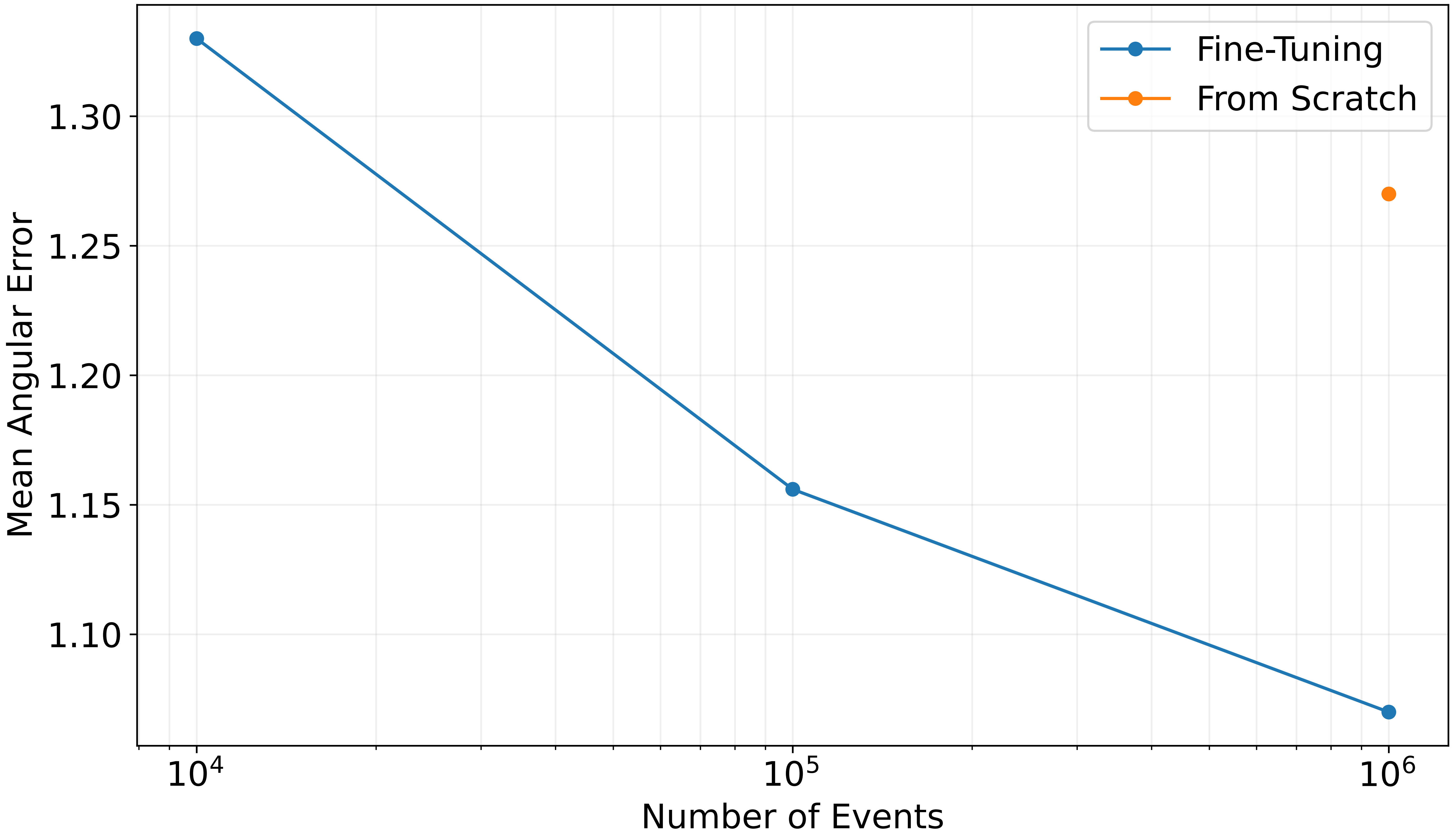
LLMs



Models trained to convergence

Kaplan et all, 2020

Finetuning (Directional Reconstruction)



- Pretrained model can be successfully fine-tuned on a downstream task
- We add a “prediction head”: an MLP to the [CLS] embedding output
- Train resulting model with direction labels
- Fine-tuning is sample efficient
- Allows to experiment with the architecture of the fine-tuned model

Future Steps

- Prometheus data for fine tuning (different labels)
  A few million events
- Dataset size scaling — what are the returns from scaling in particle physics?
- Pretraining for more than one epoch (cf 2305.16264 “Scaling Data-Constrained Language Models”)
- A more systematic study to address specific architecture choices

Conclusions

- The hybrid embedding approach and masking strategy are effective in capturing relevant information from unlabeled data.

- A clear scaling law in pre-training performance, similar to that seen in large language models.

- There are significant improvements in sample efficiency and performance when fine-tuning the pre-trained model compared to training from scratch.