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

INTERACTIONS

UNIVERSITY OF COPENHAGEN

# Foundation models for IceCube

Inar Timiryasov
Niels Bohr Institute, University of Copenhagen

Work in progress in collaboration with Oleg Ruchayskiy

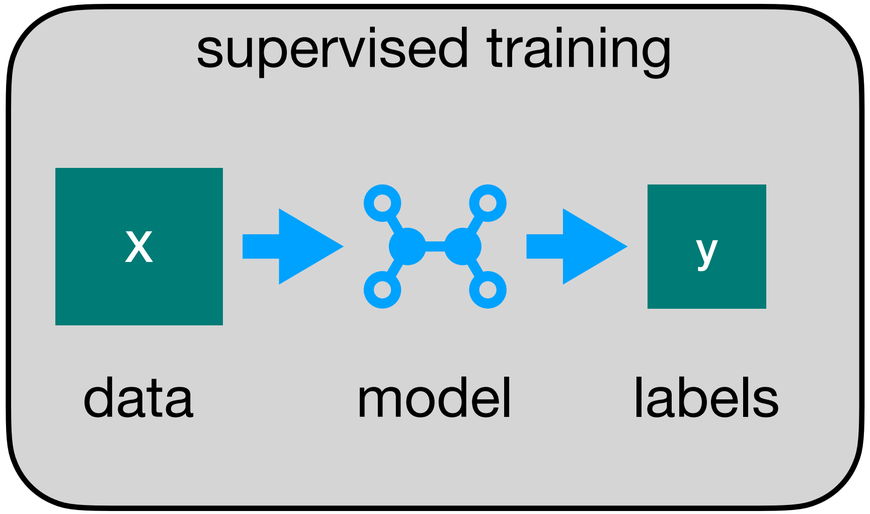
4th GraphNeT Workshop: Graph Neural Networks and Beyond

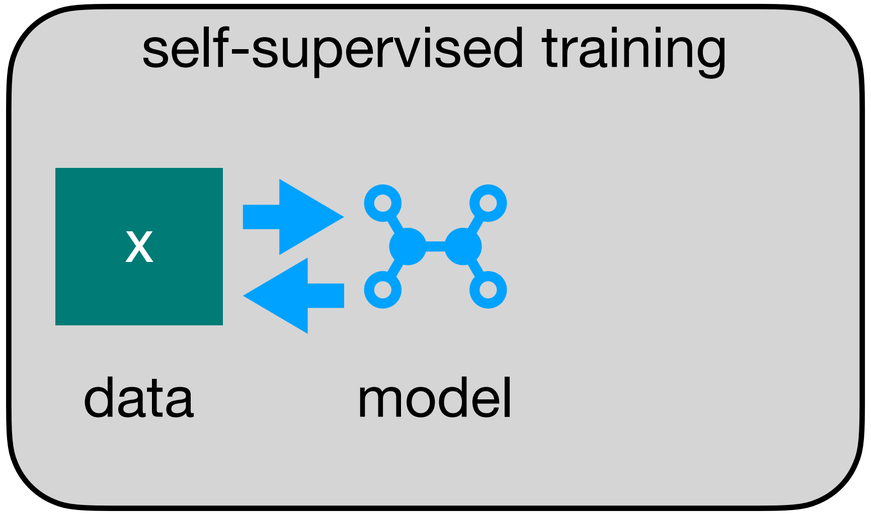
May 8th, 2024, TUM Institute for Advanced Study

# What do we mean by “foundation models”?

- Initially, the term has been coined for models like BERT and GPT-3
  2108.07258 “On the Opportunities and Risks of Foundation Models”

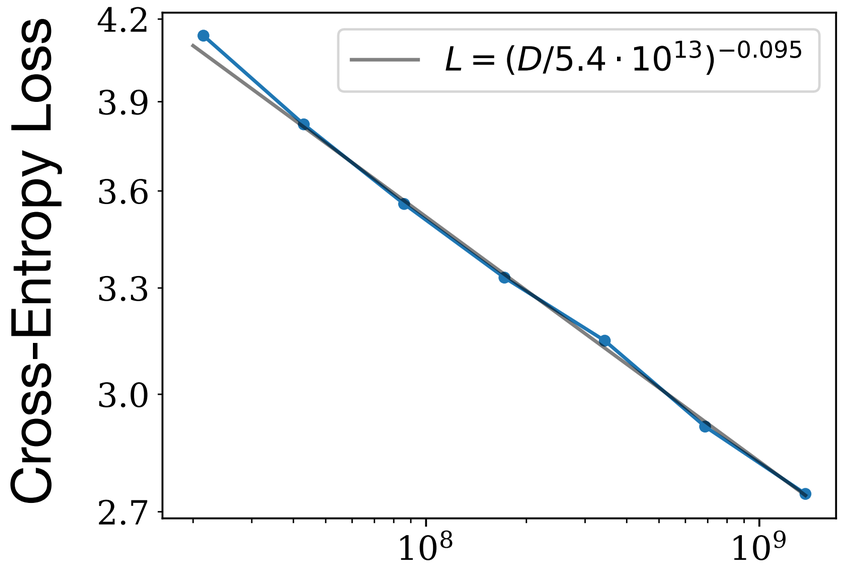
- Here, by foundational models we mean the models that are pretrained in a self-supervised way and can be fine-tuned for downstream tasks.



# Why foundation models?

- **Scale is important**
- Labeled data is expensive to produce
- Labeled data is usually MC
- In particles physics we are in the **exabyte era**




https://arxiv.org/abs/2001.08361

**Scaling Laws for Neural Language Models**


**Jared Kaplan** *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

**Sam McCandlish***
OpenAI
sam@openai.com

**Tom Henighan**
OpenAI
henighan@openai.com

**Tom B. Brown**
OpenAI
tom@openai.com

**Benjamin Chess**
OpenAI
bchess@openai.com

**Rewon Child**
OpenAI
rewon@openai.com

**Scott Gray**
OpenAI
scott@openai.com

**Alec Radford**
OpenAI
alec@openai.com

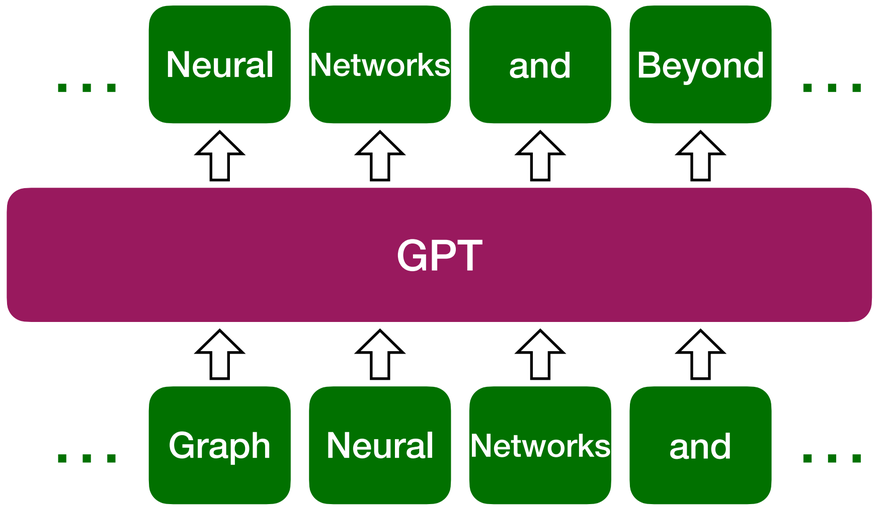
**Jeffrey Wu**
OpenAI
jeffwu@openai.com

**Dario Amodei**
OpenAI
damodei@openai.com

# Types of foundation models

**GPT**
(Generative pre-trained transformer)

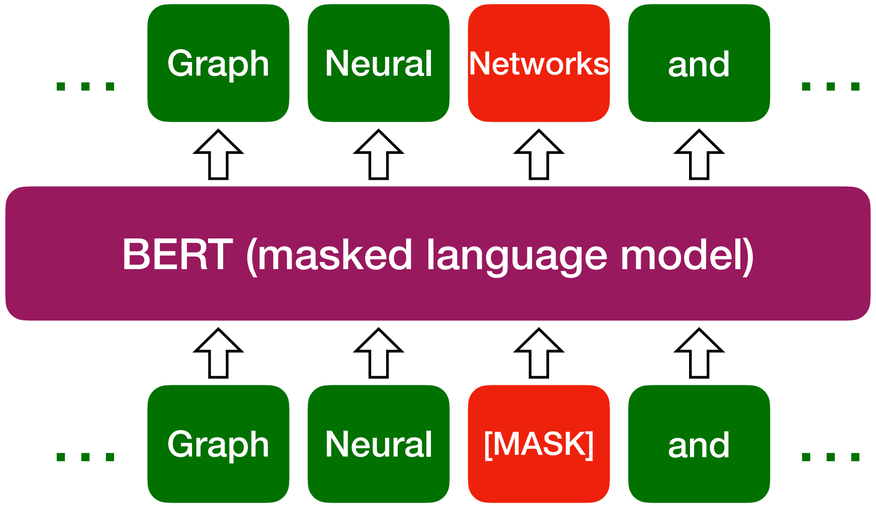
predict the next token




can be used to generate data

**BERT**
(Bidirectional Encoder
Representations from Transformers)

predict the masked tokens



# 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

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

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

- *.....*

there are also diffusion models for event generation

# Foundation models for IceCube

- Pulses sorted in time (like many kaggle solutions)
- Predict the masked ~~token~~ DOM

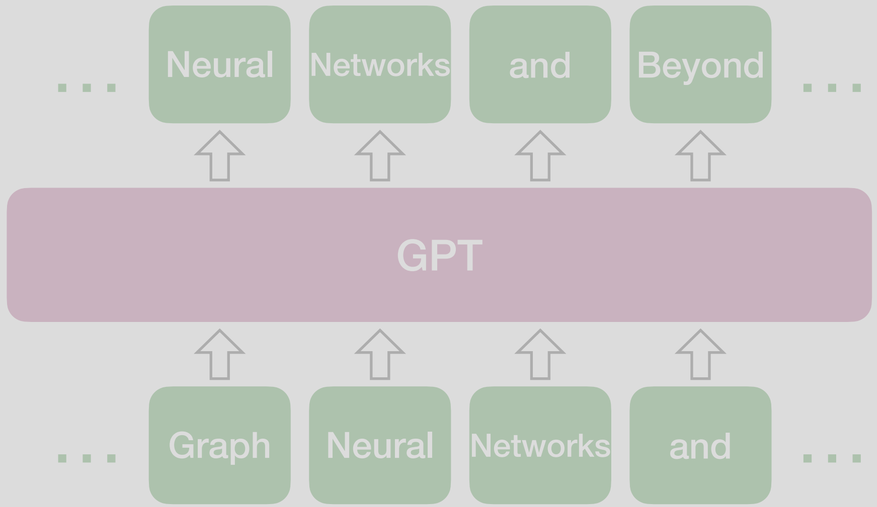
Why?

- Real data could be used for pretraining — labels are not needed
- Much larger models could be trained since the data is not a bottleneck
- Fine-tuning requires less data and hopefully doesn’t introduce strong bias towards MC

# Foundation models for IceCube

**GPT**
(Generative pre-trained transformer)

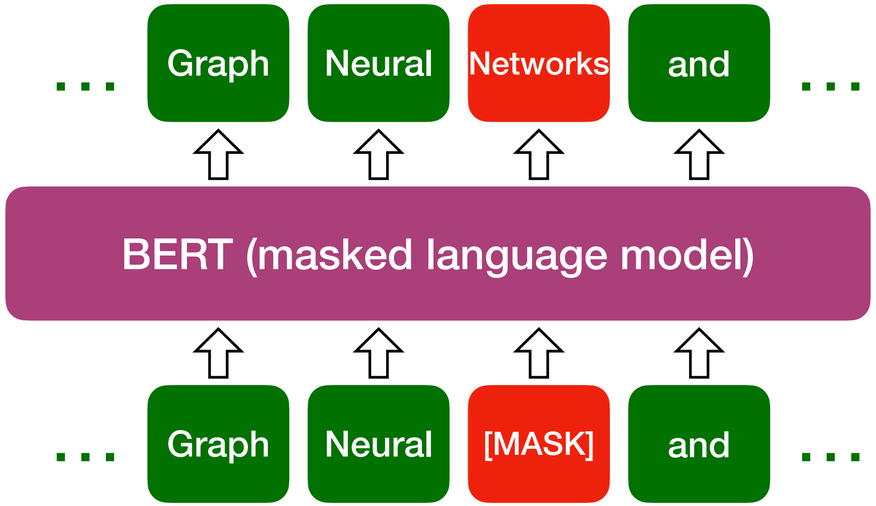
predict the next token




can be used to generate data

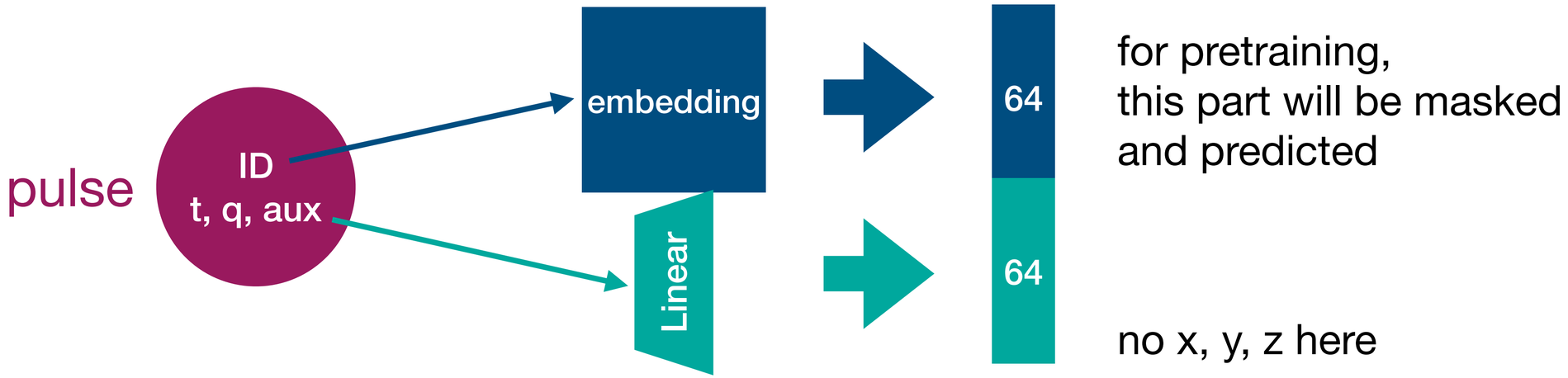
**BERT**
(Bidirectional Encoder
Representations from Transformers)

predict the masked tokens

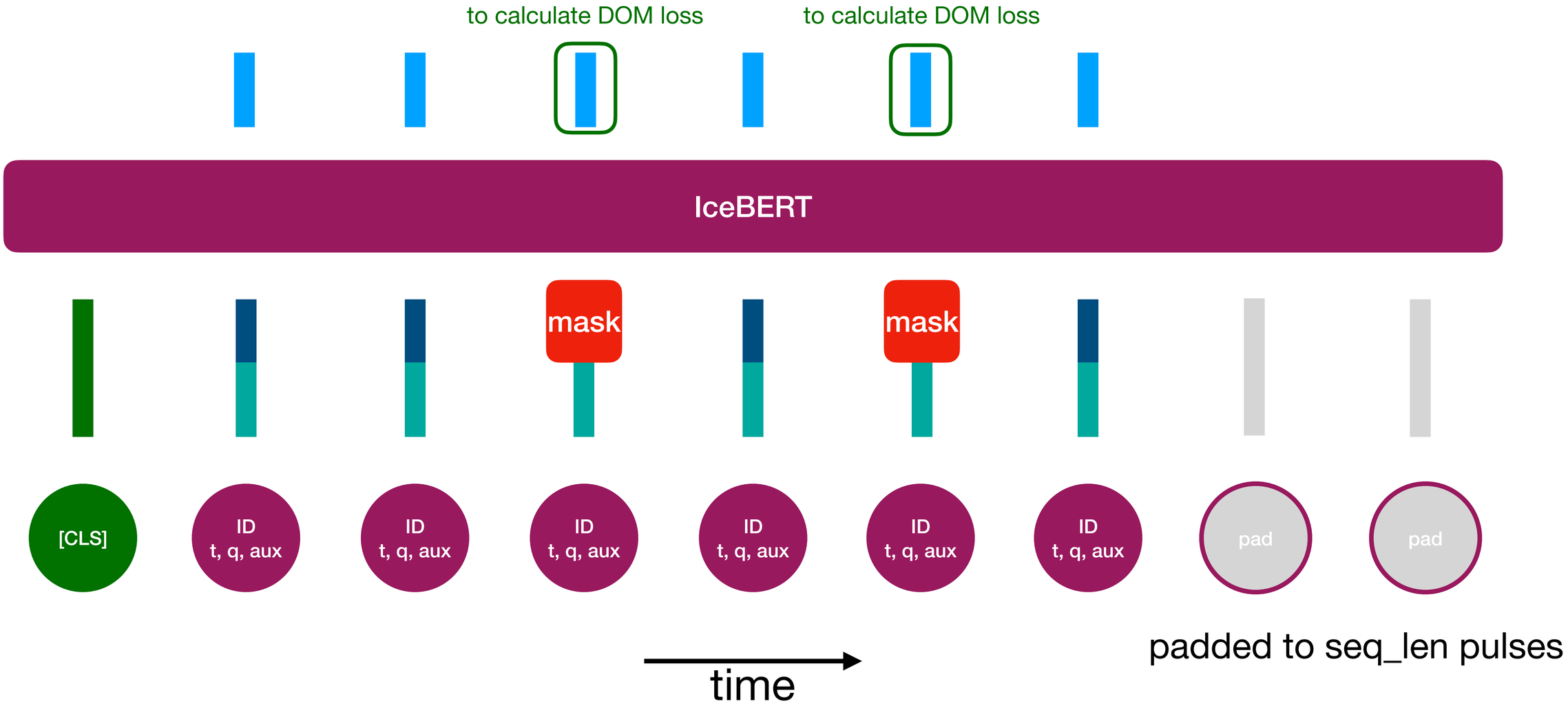


# Tokenization of pulses

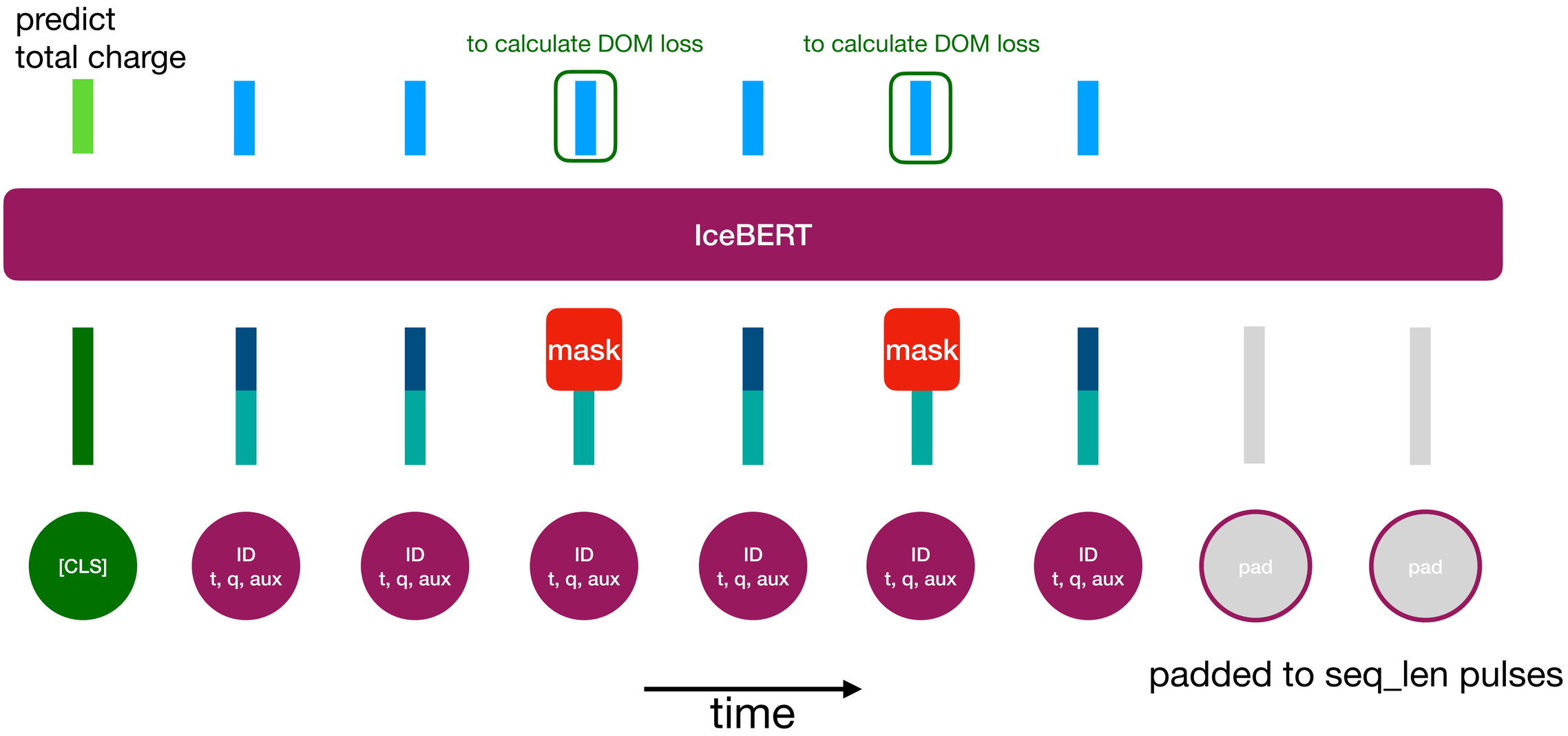
- A natural choice of 5160 “tokens” — DOM ids
- DOM ids are embedded into a higher dimensional space (64+)
- Time, charge, and other features are linearly transformed and concatenated to the DOM embeddings
- Explicit position is not used



# Pretraining



# Pretraining

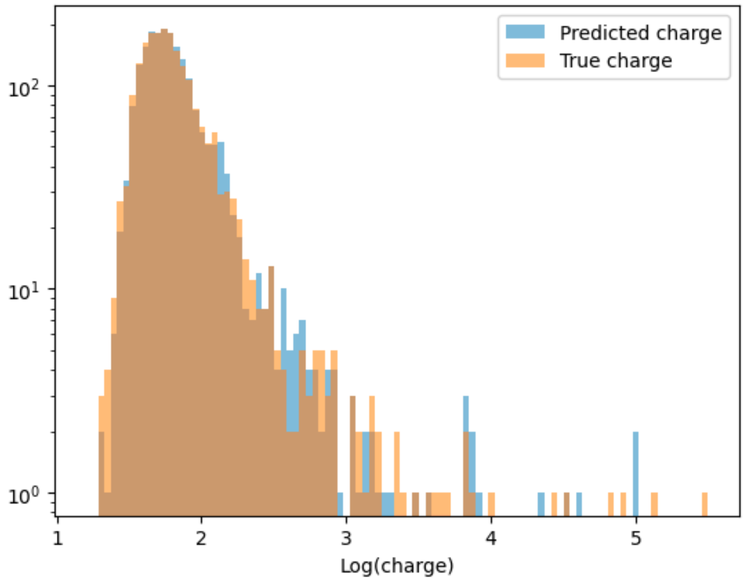


# Pretraining: DOM loss

- 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

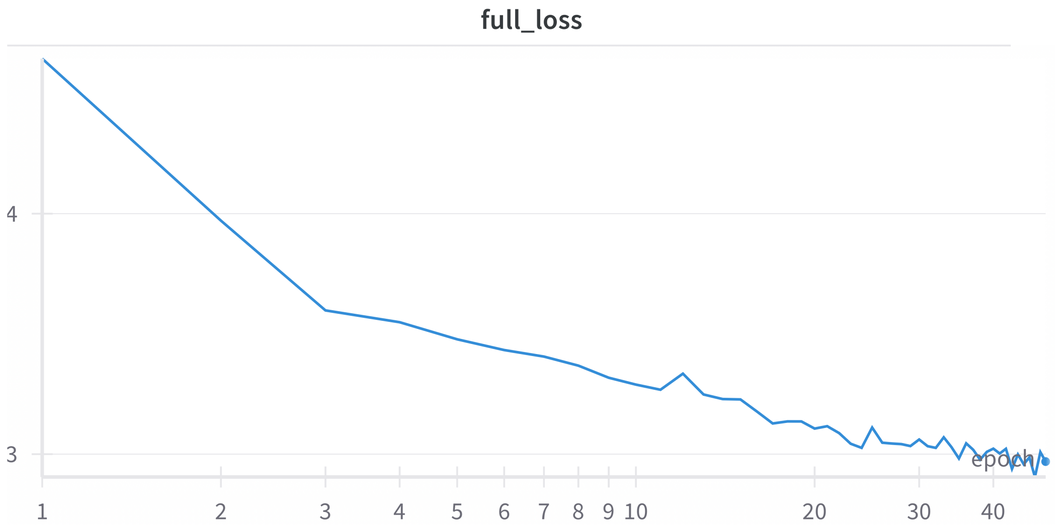
- The model has to learn how to collect useful information in [CLS] embedding for the future use on downstream tasks.
- We need some feature that is not directly accessible to the model, but can be obtained from the data (no labels)
- Candidates: the total charge of the event, center of charge
- We subsample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

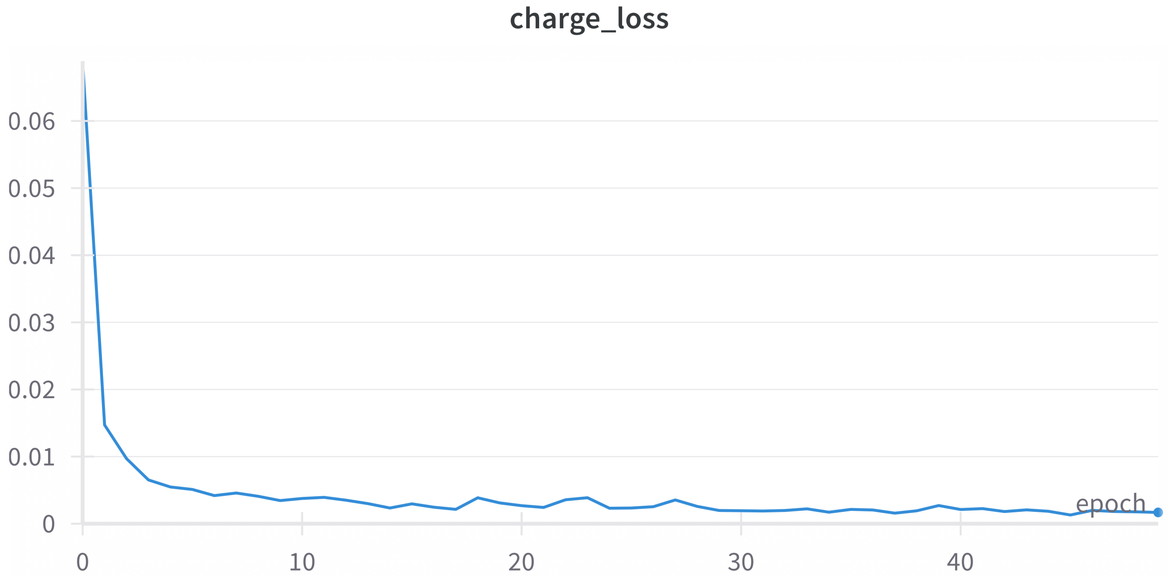


# Foundation model for IceCube

- Backbone: transformer, GRU, Mamba
- Pretraining:
  - Subsample events to seq_len (currently 128)
  - input: (DOM embedding) $\oplus$ (projection of features)
  - loss function = DOM-loss + $\lambda \times$ charge-prediction-loss
- Fine-tuning for downstream tasks

# Foundation model for IceCube



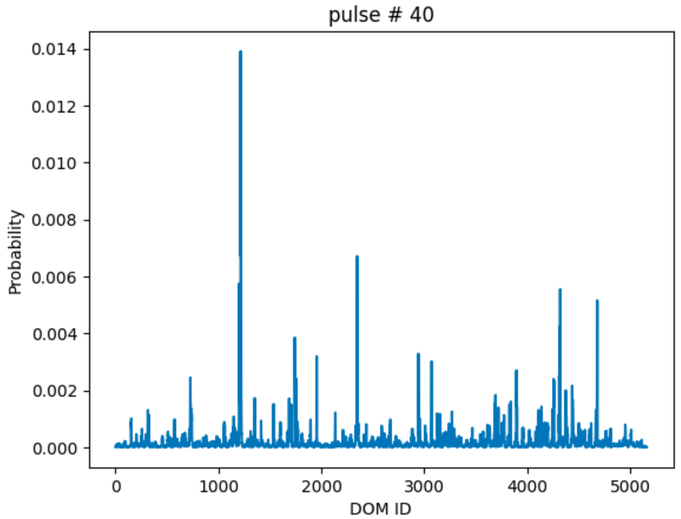



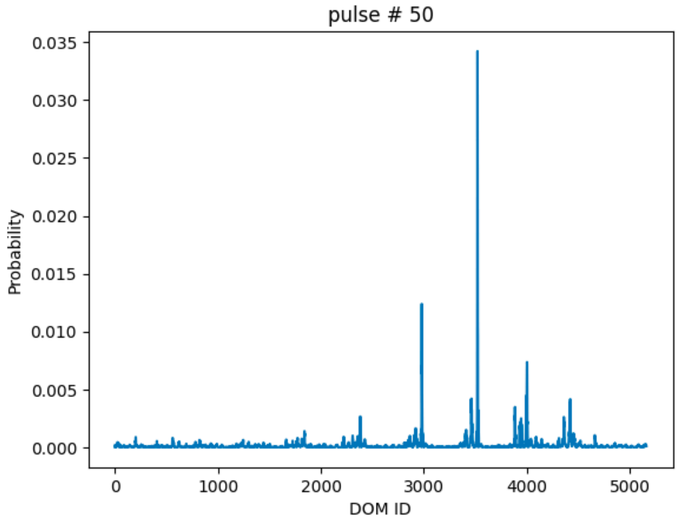

every epoch — 1 million events

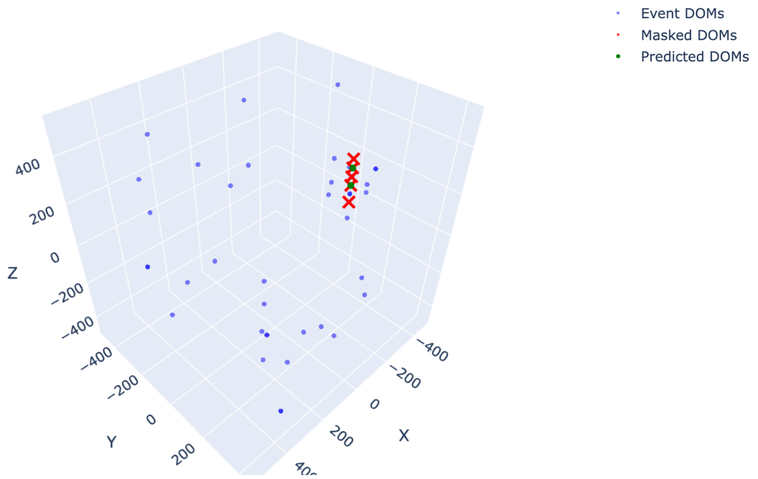
Self-supervised pretraining works!

# Interpreting the loss

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

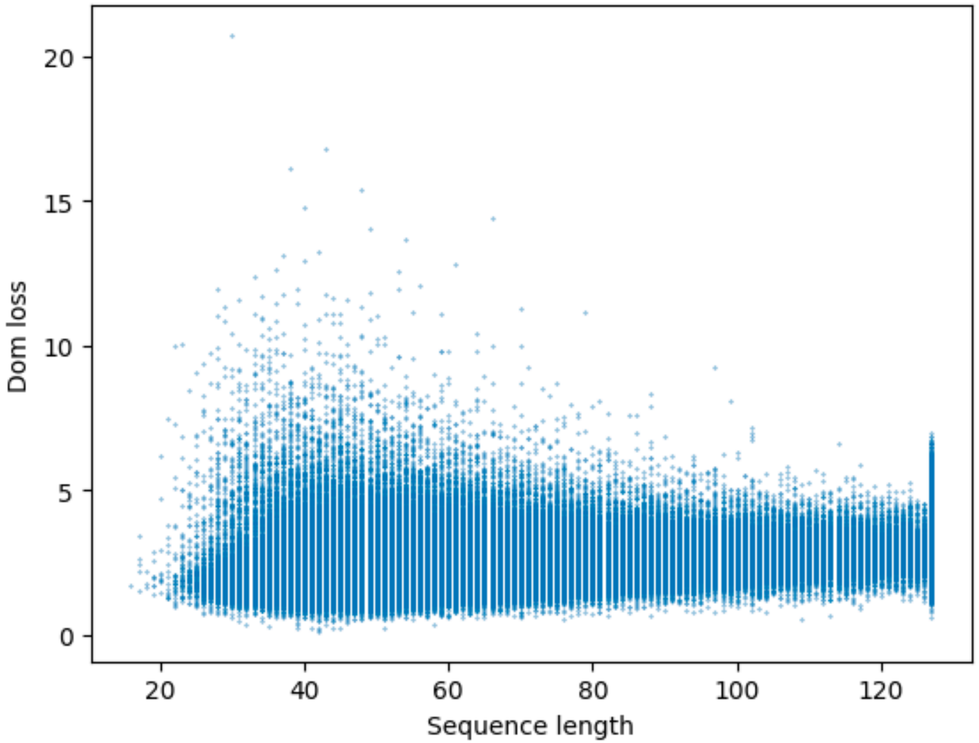


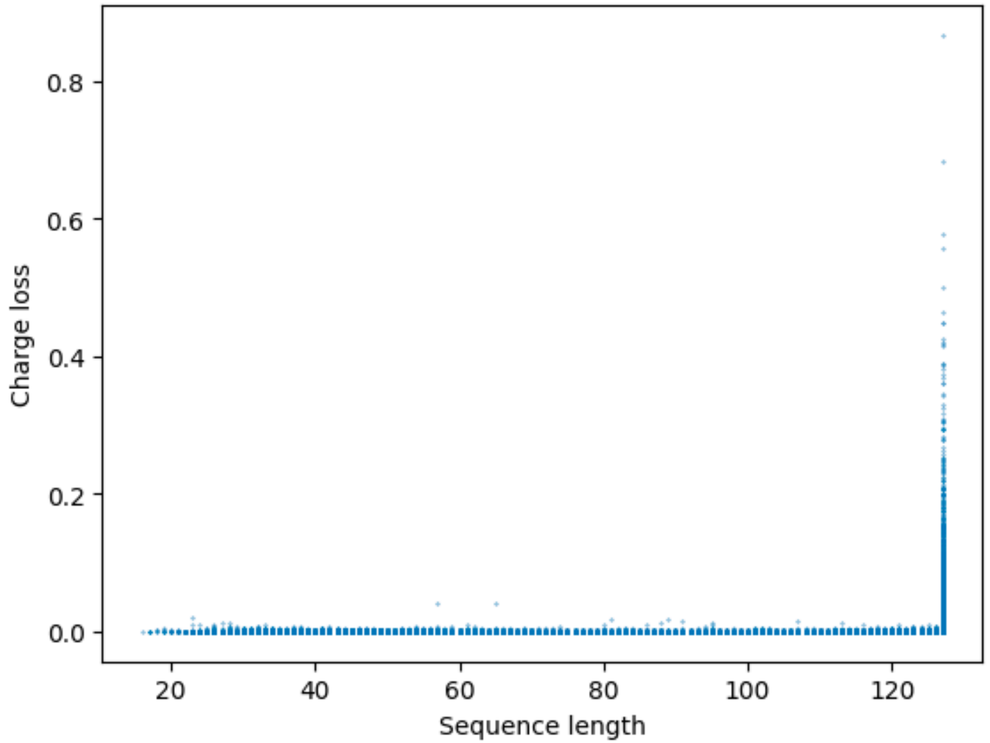




some uncertainty about the string and the DOM

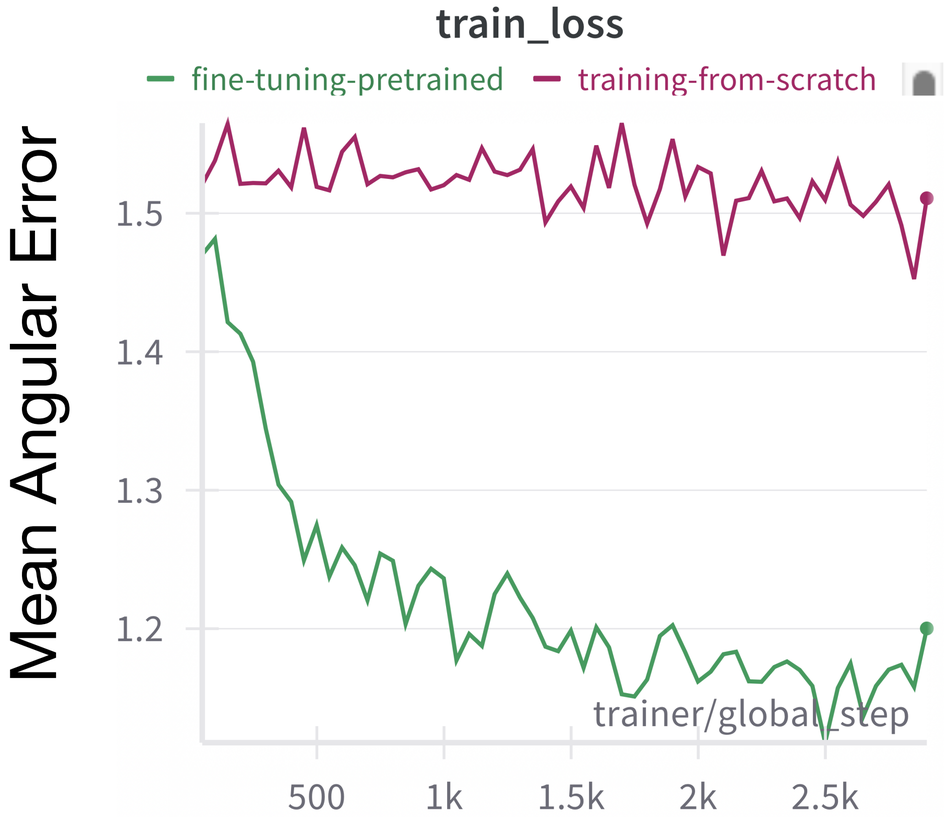
# Loss and Sequence Length

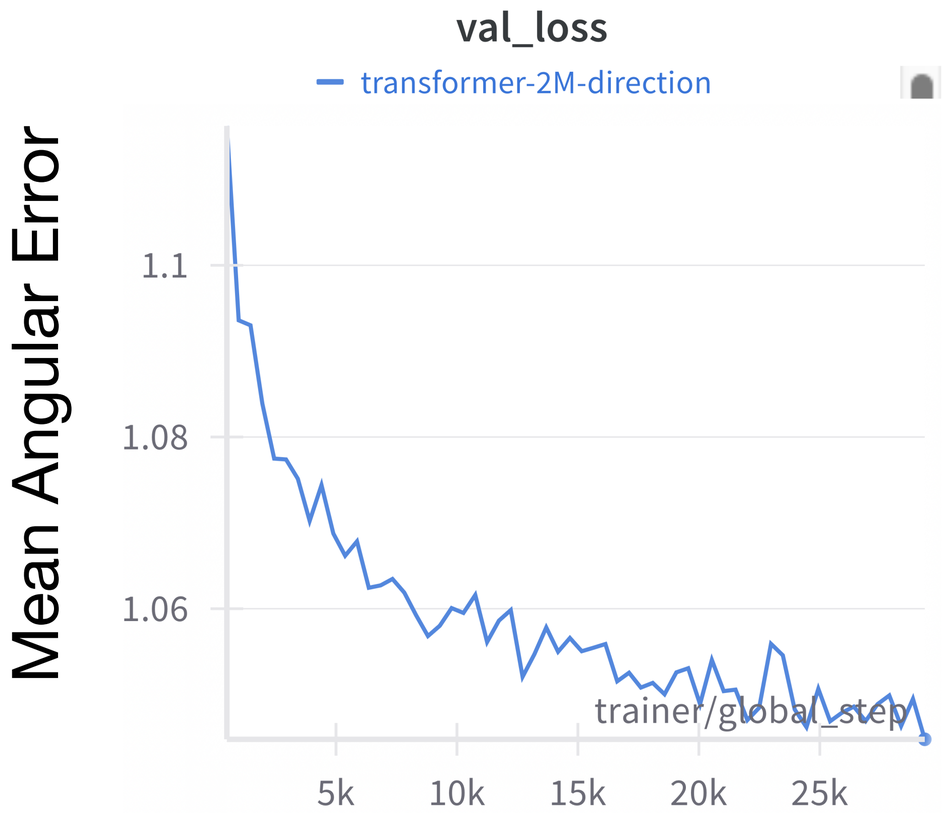



truncation is problematic

# Fine-tuning


train_loss
fine-tuning-pretrained
training-from-scratch
Mean Angular Error
1.5
1.4
1.3
1.2
trainer/global_step
500
1k
1.5k
2k
2.5k


val_loss
transformer-2M-direction
Mean Angular Error
1.1
1.08
1.06
trainer/global_step
5k
10k
15k
20k
25k

# Challenges

- Labeled data (I am not a member of IceCube)

  Looking forward for Prometheus data!

- Not specific to foundation models: proper ablations and architecture search.

  Difference between a transformer and a GRU becomes apparent after ~4 hours of training. Transformers and Mamba perform similarly.

- Subsampling is problematic.

# Outlook

- Self-supervised pretraining — masked DOM prediction — works extremely well across different backbones (transformer, RNN, Mamba) and sizes.
- Foundation models can be trained on the real data and fine-tuned for the downstream tasks.
- The nearest plan — using Prometheus data as “MC” and IceCube kaggle data as “real data”