PolarBERT: A Foundation Model for IceCube

Inar Timiryasov
Niels Bohr Institute, University of Copenhagen, Denmark
inar.timiryasov@nbi.ku.dk
Jean-Loup Tastet
Department of Computer Science, University of Copenhagen, Denmark
jeta@di.ku.dk
Oleg Ruchayskiy
Niels Bohr Institute, University of Copenhagen, Denmark
oleg.ruchayskiy@nbi.ku.dk

Abstract

The IceCube Neutrino Observatory instruments a cubic kilometer of Antarctic ice with optical sensors to detect the light emitted by neutrino interactions. Its data are used to reconstruct the direction, energy, and type of neutrinos for particle physics and astrophysics research. Although deep learning has been successfully applied to these reconstruction tasks, existing methods are typically supervised and require extensive labeled Monte Carlo simulations. In this work, we develop PolarBERT, a foundation model for IceCube, pre-trained without labels in a self-supervised manner. We show that it can be fine-tuned for neutrino directional reconstruction in a sample-efficient way, and that performance improves with a larger pre-training dataset.

Keywords: Foundation Models, Self-Supervised Learning, Neutrino Physics, IceCube

1 Introduction

The IceCube Neutrino Observatory is a cubic-kilometer detector at the South Pole, composed of 5,160 digital optical modules (DOM) embedded in Antarctic ice Abbasi et al. [2009]. These DOMs detect Cherenkov light from charged particles produced by neutrino interactions, with the primary goal of identifying cosmic ray sources and studying neutrino properties. Traditionally, IceCube data analysis has relied on computationally expensive maximum likelihood methods Aartsen et al. [2014].

Recently, deep learning methods have shown promise for IceCube data analysis Micallef [2021], Abbasi et al. [2022], Søgaard et al. [2023]. However, these supervised models depend on vast amounts of synthetic Monte Carlo data, which are resource intensive to generate. IceCube collects approximately 70 billion events annually Tilav et al. [2020], and although most are atmospheric muons, they share fundamental physics with the signal events. This abundance of real, unlabeled data motivates a self-supervised learning approach. As a first step

AI in Science
Conference 2025
EU2025DK

towards this vision, we propose PolarBERT, a foundation model for IceCube trained on simulated data via a "masked DOM prediction" task, analogous to BERT's masked token prediction Devlin et al. [2018].

The concept of foundation models is gaining traction in particle physics, particularly for jet reconstruction at the Large Hadron Collider Finke et al. [2023], Vigl et al. [2024], Birk et al. [2024], see also Barman et al. [2025]. The main challenge is data representation. Approaches often involve converting raw data into discrete tokens using techniques such as VQ-VAE Heinrich et al. [2024], Birk et al. [2024]. For IceCube, the 5,160 DOMs provide a natural tokenization scheme. Our model treats DOMs as tokens for the purpose of the pulse modeling task used for pretraining, but supports two input schemes: i) embedding DOMs as tokens or ii) linearly embedding their (x, y, z) positions, with the latter providing slightly better performance. Meanwhile, continuous features, such as pulse time and charge, are linearly embedded after a transformation/normalization step. This hybrid approach allows us to pre-train a powerful model that can learn detector characteristics directly from data and then be fine-tuned for various downstream tasks like neutrino direction reconstruction.

2 Data and Model

IceCube Data. We use a public Monte Carlo dataset released by the IceCube Collaboration for a Kaggle competition Eller [2023], which contains 131 million simulated neutrino events. Each event consists of a series of pulses, each characterized by its time, charge, originating DOM ID, and a quality flag (*auxiliary*). As event sizes vary significantly (from 2 to 100,000 pulses), we down-sample or pad them to a fixed sequence length of 128. We prioritize high-quality pulses (*auxiliary* = `False`) during this process, a strategy that has proven effective Bukhari et al. [2023].

Model Architecture. PolarBERT is an encoder-only Transformer model inspired by BERT Devlin et al. [2018]. As shown in Figure 1, we process the pulse data using a hybrid embedding strategy. Each of the 5,160 DOM IDs is mapped to a trainable embedding vector or an affine transformation of its coordinates. Continuous features (time, charge, auxiliary flag) are rescaled and linearly projected into a separate vector. These two vectors are concatenated to form the final input representation for the Transformer. We do not use explicit positional encodings, as the pulse timestamps already provide temporal ordering.

Pre-training and Fine-tuning. We use a pre-training objective similar to masked language modeling. We mask a fixed fraction of the primary pulses in each sequence and train the model to predict the original DOM IDs. To help the model learn a useful representation for downstream tasks, we add an auxiliary regression task: predicting the logarithm of the total charge of all pulses in the event. The total loss is the sum of the cross-entropy loss for DOM prediction and the MSE loss for the total charge prediction.

For the downstream task of direction reconstruction, we replace the predic-

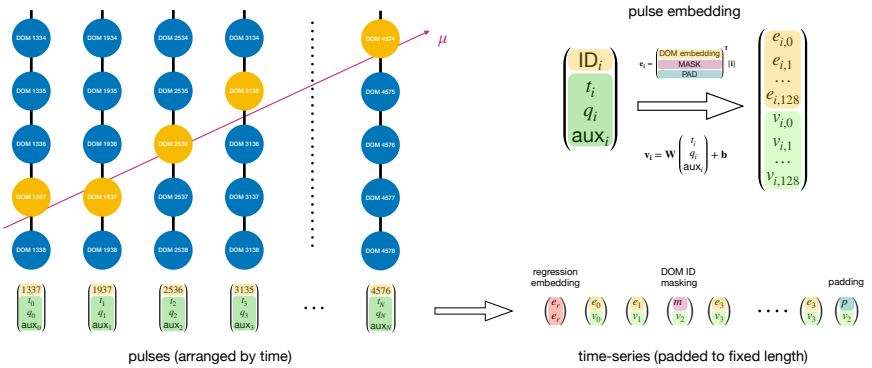

Figure 1: A sketch representing how pulses are embedded into the model space. Each DOM ID has a trainable embedding, alongside special embeddings for masking and padding. Continuous features like time and charge are linearly projected into the model space after being mapped to an appropriate range.

tion head with a small MLP that maps the output representation to a three-dimensional unit vector representing the neutrino direction. We then fine-tune the entire model using a mean angular error loss.

3 Model performance

PolarBERT is implemented in PyTorch with `FlashAttention` Dao et al. [2022]. We conducted a series of experiments to evaluate how its performance scales with the number of training steps and the sizes of the pre-training and fine-tuning datasets.

In order to consistently evaluate the pre-training performance at every epoch, we employed a trapezoidal learning rate schedule, saving checkpoints regularly throughout training, before annealing each of them with a cosine schedule. Figure 2 shows the cross-entropy validation loss as a function of training steps for an 8M-parameter model. The loss L exhibits a clear scaling-law (see Kaplan et al. [2020], Hoffmann et al. [2022]) with the number of pre-training events seen D, consistent with the fit $L \approx 1.904 + 21557 \cdot D^{-0.62}$.

To evaluate the impact of pre-training on downstream performance, we then fine-tuned models pre-trained for one epoch on datasets of varying sizes (5M, 10M, 20M, and 40M events) on the task of neutrino direction reconstruction. Figure 3 shows the test mean angular error as a function of the numbers of pre-training and fine-tuning events. We can observe two clear trends: i) for any given pre-training size, the error decreases as more labeled data is used for fine-tuning and, ii) models pre-trained on larger datasets perform better when fine-tuned on a fixed number of events. We observe, however, diminishing returns when scaling pre-training from 20M to 40M events, possibly due to the limited capacity of the model. These experiments demonstrate the effectiveness

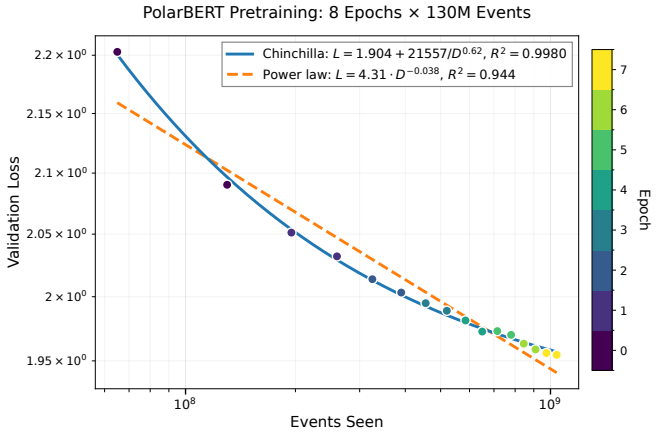

Figure 2: The validation loss as a function of the number of pretraining events follows a clear scaling trend.

of scaling up the unlabeled pre-training dataset for improving sample efficiency and downstream performance on scientific tasks.

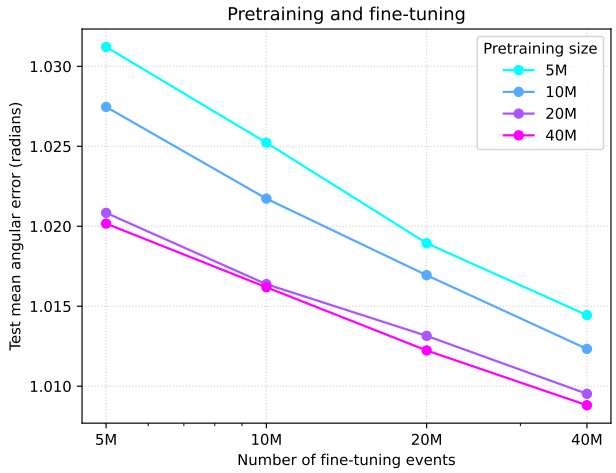

Figure 3: Downstream performance on neutrino direction reconstruction. The mean angular error (lower is better) decreases as the fine-tuning dataset size increases. Models pre-trained on larger datasets (20M and 40M) achieve lower error, demonstrating the benefit of large-scale pre-training. For comparison, the 2nd place model on Kaggle, trained for 4-5 epochs on the full 130M dataset, reached an angular loss of 0.97.

4 Conclusions

PolarBERT is a first step towards a foundation model for the IceCube Neutrino Observatory. It demonstrates the feasibility of self-supervised learning for neutrino event reconstruction. Our findings show that: i) a hybrid embedding and masked DOM prediction strategy is effective for learning from low-level detector data; ii) the pre-training performance follows a predictable scaling law; and iii) scaling up the unlabeled pre-training dataset significantly improves sample efficiency and downstream performance. This work suggests that pre-training on the vast amounts of real, unlabeled data available to IceCube could lead to powerful models. Future work will focus on scaling up the models and applying them to real experimental data. We also plan to study transfer to other simulated datasets Lazar et al. [2024].

Broader Impact Statement

This work demonstrates the potential of foundation models in experimental physics and may encourage similar research in other scientific domains. Our finding that performance scales with data and model size has important implications for the future of large-scale scientific data analysis.

Acknowledgements

We thank Erik Dam, Troels C. Petersen, and Raghavendra Selvan for valuable discussions, as well as the anonymous referees for their constructive feedback. This work was supported by a research grant (VIL57416) from VILLUM FONDEN. The work of IT was partially supported by the Carlsberg foundation and by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 847523 'INTERACTIONS'. Computational resources for this work were partially provided by the SCIENCE AI Centre of Copenhagen University.

References

M. G. Aartsen et al. Energy Reconstruction Methods in the IceCube Neutrino Telescope. *JINST*, 9:P03009, 2014. doi: 10.1088/1748-0221/9/03/P03009.

R. Abbasi et al. The IceCube Data Acquisition System: Signal Capture, Digitization, and Timestamping. *Nucl. Instrum. Meth. A*, 601:294–316, 2009. doi: 10.1016/j.nima.2009.01.001.

R. Abbasi et al. Graph Neural Networks for low-energy event classification & reconstruction in IceCube. *JINST*, 17(11):P11003, 2022. doi: 10.1088/1748-0221/17/11/P11003.

K. G. Barman et al. Large Physics Models: Towards a collaborative approach with Large Language Models and Foundation Models. 1 2025.

J. Birk, A. Hallin, and G. Kasieczka. OmniJet-α: The first cross-task foundation model for particle physics. 3 2024.

H. Bukhari, D. Chakraborty, P. Eller, T. Ito, M. V. Shugaev, and R. Ørsøe. IceCube – Neutrinos in Deep Ice The Top 3 Solutions from the Public Kaggle Competition. 10 2023.

T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

P. Eller. Public Kaggle Competition "IceCube – Neutrinos in Deep Ice". In *38th International Cosmic Ray Conference*, 7 2023.

T. Finke, M. Krämer, A. Mück, and J. Tönshoff. Learning the language of QCD jets with transformers. *JHEP*, 06:184, 2023. doi: 10.1007/JHEP06(2023)184.

L. Heinrich, T. Golling, M. Kagan, S. Klein, M. Leigh, M. Osadchy, and J. A. Raine. Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models. 1 2024.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

J. Lazar, S. Meighen-Berger, C. Haack, D. Kim, S. Giner, and C. A. Argüelles. Prometheus: An open-source neutrino telescope simulation. *Comput. Phys. Commun.*, 304:109298, 2024. doi: 10.1016/j.cpc.2024.109298.

J. Micallef. Using convolutional neural networks to reconstruct energy of GeV scale IceCube neutrinos. *JINST*, 16(09):C09019, 2021. doi: 10.1088/1748-0221/16/09/C09019.

A. Søgaard et al. GraphNeT: Graph neural networks for neutrino telescope event reconstruction. *J. Open Source Softw.*, 8(85):4971, 2023. doi: 10.21105/joss.04971.

S. Tilav, T. K. Gaisser, D. Soldin, and P. Desiati. Seasonal variation of atmospheric muons in IceCube. *PoS*, ICRC2019:894, 2020. doi: 10.22323/1.358.0894.

M. Vigl, N. Hartman, and L. Heinrich. Finetuning Foundation Models for Joint Analysis Optimization. 1 2024.