
CURTAIN: Towards a Foundation Model for Neutrino Telescopes

Severin Magel
Harvard University
Technical University of Munich
severin.magel@tum.de

Felix J. Yu
Harvard University
felixyu@g.harvard.edu

Carlos A. Argüelles
Harvard University
carguelles@fas.harvard.edu

Abstract

Neutrino telescopes record more than a trillion photon detections per year, yet there does not exist a self-supervised learning (SSL) task which fully satisfies all three conditions: improving as the pretraining set grows, transferring to a different detector, and outperforming the best model trained without pretraining at all. We introduce CURTAIN, a temporal masking scheme and the first SSL task for neutrino telescopes that achieves all of these goals. Downstream performance monotonically improves with the number of pretraining events, as predicted by scaling laws. Our encoder builds on DeepIce, currently the leading model on the NuBench benchmarks direction reconstruction task. Pretraining it with CURTAIN reduces its median angular error by 45 % and its median energy error by 31 % compared to training from scratch. It also transfers to a detector geometry never seen in pretraining, reducing the median angular error by 36 % compared to training from scratch. At matched capacity and pretraining data, CURTAIN-pretrained encoders beat those pretrained by existing SSL tasks on both tasks and at every pretraining set size.

1 Introduction

Neutrinos interact with matter so rarely that they reach us in a straight line from their sources, carrying a direct view of the most violent processes in the universe [1]. That same rarity makes them hard to detect. A neutrino telescope instruments a km³-scale volume of glacial ice or deep water with vertical lines of optical modules (OMs) (Figure 1(a)). A neutrino interacting in the medium creates charged particles, which emit Cherenkov light as they travel. The OMs detect that light, registering the interaction as a variable-length spatio-temporal point cloud called an event. Each of its points, a pulse, records one or more detected photons and carries a position, an arrival time, and the amount of light detected, which commonly is referred to as charge.

Every physics analysis must estimate the neutrino’s direction and energy from this light pattern. Real data carry no labels, so estimators rely on high-fidelity Monte Carlo (MC) simulation: machine learning (ML) models are trained on pattern–target pairs from this forward model and applied to real detected patterns [2]. This simulation-based inference setup carries two costs. Every downstream task requires large simulation datasets, and generating them is expensive: the Cherenkov light of every event must be propagated photon by photon through hundreds of meters of ice or water. Any mismatch between the simulated and the real data distribution also biases the estimator in ways

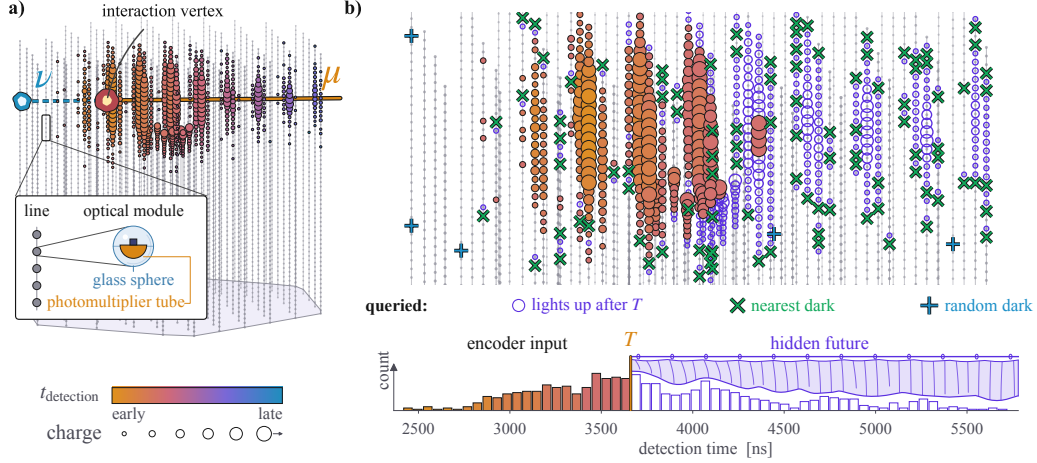


Figure 1: **(a)** A neutrino telescope, here on IceCube’s array geometry: a neutrino (ν) interacts and the outgoing muon (μ) lights OMs along its path. Marker color is arrival time, area is charge. **(b)** CURTAIN’s masking scheme on the same simulated event, cut at the threshold T . Filled circles are the encoder’s input. The model is queried at three kinds of positions: open circles are OMs that light up only after T (positives); green crosses are the dark OMs nearest to each positive (hard negatives); blue plus signs are dark OMs drawn at random (easy negatives). Below, the hit-time distribution, curtained off after T .

that are difficult to quantify [3]. Such mismatch is hard to avoid, because the detecting medium is a natural one that was never characterized in a laboratory. The optical properties of the glacial ice must be calibrated in situ and vary strongly with depth [4], while water detectors sit in a time-varying optical background of bioluminescence and radioactive decay [5].

Self-supervised learning (SSL) offers a way around both costs. Foundation models pretrained on large unlabeled datasets have transformed many fields, among them language, vision, and robotics [6–8]. Across the operating detectors, neutrino telescopes record more than a trillion pulses per year. This is a very conservative lower bound (Appendix A), and it already exceeds the 300 billion tokens GPT-3 was trained on [6]. A shared, information-rich representation of these patterns would improve every downstream analysis at once, and cut the number of simulated particle interactions each new analysis has to generate. Learning representations of events from real data can also improve the estimator’s generalization from MC to data.

SSL has already been applied to neutrino telescopes. PolarBERT [9] improves direction estimation from pretraining, reducing the reliance on labeled simulation, and Neptune [10] uses pretraining to mitigate MC mismodeling. We argue that neither is yet a foundation model for neutrino telescopes, which we take to require three properties: it improves as the pretraining set grows, it generalizes beyond a single detector, and it outperforms the best model trained without pretraining. PolarBERT’s input vocabulary enumerates the sensors of one detector, and neither method can pretrain the strongest supervised architecture. In our matched-capacity comparison, Neptune’s downstream performance does not improve with more pretraining data, and PolarBERT stays far behind a model trained from scratch on direction.

We introduce CURTAIN, an SSL pretraining method for neutrino telescope data that masks events in time rather than pulse by pulse. Its encoder–decoder architecture does not require passing masked pulses through the encoder, allowing the use of time- and position-dependent attention mechanisms. It is the first SSL task to satisfy all three properties.

2 CURTAIN

Like PolarBERT and Neptune, CURTAIN follows the masked modeling paradigm [7, 11]: hide part of an event and ask the model to recover it from the rest. We split each event at a randomly sampled threshold time T (Figure 1(b)): pulses before T are the visible context, and every OM lit only after

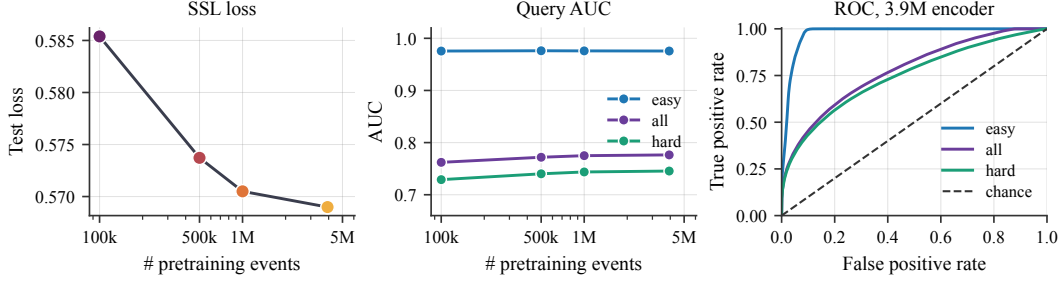


Figure 2: CURTAIN pretraining performance. **Left:** held-out SSL loss against the number of pre-training events. **Middle:** query ROC-AUC for separating the positives from the easy negatives, from the hard negatives, and from all negatives together. **Right:** ROC curves of the encoder pretrained on 3.9 million events.

T is held out with all of its information hidden. The model is then queried at sensor positions outside the context, each query carrying nothing but a position, and must predict whether that sensor lights up after T . The queried positions are of two kinds: the held-out OMs, which do light up after T (positives), and OMs that stay dark for the whole event (negatives), taken either nearest to a positive (hard) or at random (easy). An additional auxiliary head regresses the arrival time. Appendix B gives the architecture and training details.

This masking scheme is in contrast to PolarBERT, which hides a random subset of pulses [9], and Neptune, which in the variant we compare against hides the arrival time of every pulse [10]; Appendix D compares the three schemes. We conjecture that masking temporally is what makes a representation transfer. A randomly masked pulse can be recovered by interpolating between its neighbors in space and time; masking everything after T removes that shortcut and forces the model to learn how light propagates through the detector.

This design is also why CURTAIN, unlike Neptune and PolarBERT, can pretrain the strongest supervised architecture. Both baselines are encoder-only in pretraining, so masked pulses must pass through the encoder’s attention blocks. DeepIce (second-place solution in Bukhari et al. [12]), the leading model on the direction reconstruction task of the NuBench benchmark [13], uses SpaceTime attention: a bias derived from the space–time distance between two pulses, undefined when a position or time is masked. CURTAIN’s encoder sees only visible pulses and its decoder is queried at the held-out positions, so any encoder whose attention bias depends on position or time can be pretrained this way, DeepIce included.

3 Experimental setup

All data used in this work comes from the NuBench datasets [13]. We use only the hexagon and flower_1 datasets, which resemble the IceCube [4] and KM3NeT-ARCA [5] geometries with simplified detector electronics, optical modules and optical medium, and restrict ourselves to track events, in which the incident neutrino creates a muon that travels through the detector.

We pretrain a DeepIce encoder using the CURTAIN SSL task. The encoder is downscaled to 2.2 million parameters to keep the computational cost manageable. We pretrain on 100 thousand, 500 thousand, 1 million, and 3.9 million events of the hexagon dataset using early stopping. We use a binary cross-entropy (BCE) loss on whether each queried sensor lights up after T , and a Huber loss on the first-hit time of the sensors that do. As baselines we pretrain Neptune and PolarBERT on the same data, at the same depth, model dimension and number of heads, so all schemes are compared at matched capacity.

We finetune the pretrained encoders on direction and energy reconstruction, on a training set of 100 thousand hexagon events. This split keeps only events with at least eight pulses, shares no event with any pretraining set, and has 200 thousand validation and 200 thousand test events. We use the von Mises–Fisher (vMF) negative log likelihood for direction and the mean squared error (MSE) on $\log_{10} E$ for energy. Both heads are parameterized differently from Ørsøe et al. [13], which we found to train more stably (Appendix C).

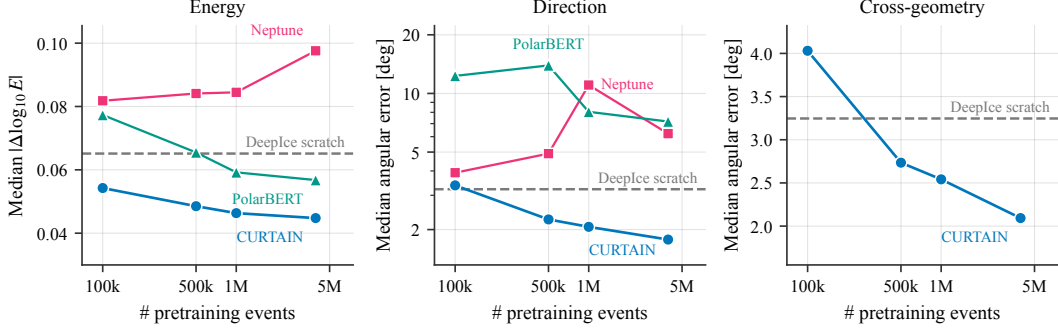


Figure 3: Downstream performance against the number of pretraining events, after finetuning on 100 thousand labeled events. **Left** and **middle**: median energy and angular error on hexagon, the geometry also used for pretraining, for CURTAIN and the Neptune and PolarBERT baselines; the angular-error axis is logarithmic. **Right**: median angular error on flower_1, a geometry never seen during pretraining, for CURTAIN only. Dashed lines: DeepIce trained from scratch on the same labeled events.

We additionally finetune the direction task on 100 thousand flower_1 events, with 20 thousand validation and 50 thousand test events, to test how well the pretrained representations generalize to an unseen geometry. All performance is evaluated on held-out sets drawn from the same distribution as the corresponding training set and disjoint from the pretraining events, reporting the median angular error, the angle between the reconstructed and the true direction, and the median $|\Delta \log_{10} E|$, evaluated in log space given the large dynamic range of neutrino events [1]. Each reported value comes from a single training run, so we give no estimate of run-to-run variation.

4 Results

4.1 Pretraining performance

Figure 2 shows the performance of CURTAIN in pretraining. The held-out SSL loss falls monotonically with pretraining set size, from 0.585 at 100 thousand events to 0.569 at 3.9 million (left panel in Figure 2), consistent with data scaling at fixed model size [14]. The easy negatives are already almost perfectly separated from the positives by the encoder pretrained on the smallest set (AUC = 0.976), and this separation does not improve with pretraining set size. The hard-negative AUC instead rises from 0.729 to 0.745, carrying the overall query AUC from 0.762 to 0.776.

4.2 Finetuning tasks

Direction reconstruction improves monotonically with pretraining set size (Figure 3, middle), reaching 1.78° against 3.22° from scratch, a 45 % reduction. Only the encoder pretrained on the smallest set, 100 thousand events, is slightly worse than training from scratch (3.37°). Energy also improves monotonically with pretraining set size (Figure 3, left), and unlike direction it beats training from scratch already at 100 thousand events (0.0543). It reaches a median $|\Delta \log_{10} E|$ of 0.0448 against 0.0651, a 31 % reduction and the smaller gain of the two. At matched capacity, pretraining data and compute budget, CURTAIN beats both baselines on both tasks at every pretraining set size.

This comparison is between complete methods: each baseline brings its own encoder and input embedding along with its pretraining objective, so it does not isolate the contribution of temporal masking. The comparison against the from-scratch DeepIce is the one that holds the architecture fixed. Neptune’s downstream performance does not improve with the pretraining set size. PolarBERT stays far behind the from-scratch DeepIce on direction at every size we tried (7.17° at best), and on energy only draws level with it at 500 thousand events and overtakes it from 1 million. Its weak direction performance is plausibly because PolarBERT’s OM-token embedding carries no geometric prior, so it needs large amounts of data to become competitive.

4.3 Cross-detector generalization

An encoder pretrained on hexagon also transfers to `flower_1`, a geometry it never saw during pretraining (Figure 3, right). With 100 thousand labeled events it improves on training from scratch, from 3.25° to 2.09° (36 %), using the encoder pretrained on 3.9 million events. A CURTAIN query carries only a sensor position, so nothing the encoder learns is tied to a sensor numbering, which makes transfer to another layout possible; whether temporal masking is what makes the transferred representation useful, as conjectured in Section 2, would require further ablation studies. Transfer also improves monotonically with pretraining set size, from 4.03° at 100 thousand events, still worse than training from scratch, to 2.09° at 3.9 million; pretraining overtakes training from scratch from 500 thousand events (2.73°).

5 Conclusion

We introduced CURTAIN, an SSL pretraining task for neutrino telescopes that temporally splits an event at a randomly sampled threshold time and requires the model to predict which sensors light up after that time. We asked three things of a foundation model, and CURTAIN delivers all three at the scale we tested. Its performance improves monotonically with the size of the pretraining set, in the pretext task and in both downstream tasks. The representations generalize beyond a single detector, reaching a median angular error of 2.09° against 3.25° from scratch on a geometry never seen during pretraining, given 100 thousand labeled events. And it outperforms the best model trained without pretraining, cutting the median angular error by 45 % and the median $|\Delta \log_{10} E|$ by 31 %. At matched capacity and pretraining budget, CURTAIN also beats both Neptune and PolarBERT on both downstream tasks by a large margin. All of this was shown in simulation, with simplified detector electronics, optical modules and optical medium, from single training runs, and after full finetuning on 100 thousand labeled events; real data, label efficiency, and an ablation that separates temporal masking from the encoder remain open. CURTAIN was pretrained on 3.9 million simulated events, less than 10^{-4} of what IceCube records in a year. Pretraining at that scale, and on real rather than simulated data, is the natural next step. In the future we plan on jointly training on different detectors and exploring model-size and compute-constrained scaling laws of the SSL and finetuning tasks.

Acknowledgments and Disclosure of Funding

SM is supported by a PROMOS scholarship from the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD), awarded by the Technical University of Munich. FJY is supported by the NSF Graduate Research Fellowship under Grant No. 2140743. CAA are supported by the Faculty of Arts and Sciences of Harvard University, the National Science Foundation (NSF, CAREER Grant No. 2239795), the John Templeton Foundation (Grant No. 63651), the Research Corporation for Science Advancement, and the David & Lucile Packard Foundation.

References

- [1] Maurizio Spurio. *Probes of Multimessenger Astrophysics: Charged Cosmic Rays, Neutrinos, γ -Rays and Gravitational Waves*. Astronomy and Astrophysics Library. Springer International Publishing, Cham, 2018. ISBN 978-3-319-96853-7 978-3-319-96854-4. doi: 10.1007/978-3-319-96854-4.
- [2] R. Abbasi, M. Ackermann, J. Adams, et al. A convolutional neural network based cascade reconstruction for the IceCube Neutrino Observatory. *Journal of Instrumentation*, 16(07): P07041, July 2021. ISSN 1748-0221. doi: 10.1088/1748-0221/16/07/P07041.
- [3] M.G. Aartsen, M. Ackermann, J. Adams, et al. Efficient propagation of systematic uncertainties from calibration to analysis with the SnowStorm method in IceCube. *Journal of Cosmology and Astroparticle Physics*, 2019(10):048, October 2019. ISSN 1475-7516. doi: 10.1088/1475-7516/2019/10/048.

- [4] M.G. Aartsen, M. Ackermann, J. Adams, et al. The IceCube Neutrino Observatory: Instrumentation and online systems. *Journal of Instrumentation*, 12(03):P03012, March 2017. ISSN 1748-0221. doi: 10.1088/1748-0221/12/03/P03012.
- [5] S Adrián-Martínez, M Ageron, F Aharonian, et al. Letter of intent for KM3NeT 2.0. *Journal of Physics G: Nuclear and Particle Physics*, 43(8):084001, August 2016. ISSN 0954-3899, 1361-6471. doi: 10.1088/0954-3899/43/8/084001.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [7] Kaiming He, Xinlei Chen, Saining Xie, et al. Masked Autoencoders Are Scalable Vision Learners . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01553. URL <https://doi.ieeeecomputersociety.org/10.1109/CVPR52688.2022.01553>.
- [8] NVIDIA, Johan Bjorck, Fernando Castañeda, et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots, March 2025.
- [9] Inar Timiryasov, Jean-Loup Tastet, and Oleg Ruchayskiy. PolarBERT: A Foundation Model for IceCube. In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2024. URL https://ml4physicalsciences.github.io/2024/files/NeurIPS_ML4PS_2024_259.pdf.
- [10] Felix J. Yu, Nicholas Kamp, and Carlos A. Argüelles. Reducing Simulation Dependence in Neutrino Telescopes with Masked Point Transformers, January 2026.
- [11] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, et al. Masked Autoencoders for Point Cloud Self-supervised Learning, March 2022.
- [12] Habib Bukhari, Dipam Chakraborty, Philipp Eller, et al. IceCube – Neutrinos in Deep Ice The Top 3 Solutions from the Public Kaggle Competition, October 2023.
- [13] Rasmus Ørsøe, Stephan Meighen-Berger, Jeffrey Lazar, et al. NuBench: An open benchmark for deep learning-based event reconstruction in neutrino telescopes. *Journal of Instrumentation*, 21(05):T05001, May 2026. ISSN 1748-0221. doi: 10.1088/1748-0221/21/05/T05001.
- [14] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling Laws for Neural Language Models. <https://arxiv.org/abs/2001.08361v1>, January 2020.
- [15] Yury Malyshev. Baikal-GVD neutrino telescope: Design reference 2022. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1050:168117, May 2023. ISSN 0168-9002. doi: 10.1016/j.nima.2023.168117.
- [16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, et al. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL <https://doi.org/10.1145/3503250>.
- [17] Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019.
- [18] Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=2M9CUYnBA>.

A A conservative lower bound on the number of pulses recorded per year

No collaboration publishes a pulse count for its trigger stream, so we bound it from below using the trigger conditions themselves. Every trigger fixes a minimum number of sensors that must register light before an event is written out. Multiplying that minimum by the trigger rate and one year of livetime therefore gives a very conservative floor on the number of pulses recorded per year.

	Trigger minimum	Rate	Pulses/year
IceCube [4]	≥ 8 OMs (SMT8)	2.5 kHz	6×10^{11}
Baikal-GVD [15]	≥ 2 adjacent OMs	1.1 kHz	7×10^{10}
KM3NeT [5]	≥ 5 L1 hits	40 Hz	6×10^9
Total			$\geq 7 \times 10^{11}$

Some 89 % of the total is IceCube’s. The true figure is far larger, because the bound credits each event with only the minimum number of sensors its trigger demands, whereas a typical event lights many more, and with only a single pulse per sensor, whereas a lit sensor often records several. It also counts only the three detectors listed. The 7×10^{11} is therefore a very conservative floor, not an estimate.

B Architecture and training details

This appendix documents the exact configuration behind every CURTAIN result in this paper.

Pulse cap. Events with more than $N_{\max} = 768$ pulses are subsampled uniformly without replacement to $N_{\max}$; smaller events are used in full. This bounds the $\mathcal{O}(N^2)$ attention cost and is applied at every stage, with a fresh subsample per capped event, and to all three encoders, so every comparison is cap-matched. For CURTAIN the cap precedes the temporal split, so visible pulses and queried sensors come from the same set. It is well above the median event of 137 pulses and binds on 22 % of test events, the high-energy tail (maximum 13,344).

Temporal masking. The cutoff T is drawn so that a uniformly random fraction of the event’s pulses falls before it, re-drawn every epoch, so the model sees every masking ratio equally often. Hard negatives are the nearest dark sensor to each positive, with a smaller number drawn at random from the sensors that stay dark.

DeepIce encoder. Visible pulses (x, y, z, t, q) , with times relative to the charge-weighted mean time of the visible pulses, are embedded by a Fourier feature encoder and processed by a reduced DeepIce encoder: $N_{\text{spacetime}} = 2$ multi-head attention blocks whose attention logits carry an additive relative space–time interval bias (SpaceTime bias), followed by $N_{\text{vanilla}} = 3$ standard pre-norm Transformer blocks (multi-head attention, feed-forward, residual connections and normalization). We use $d_{\text{model}} = 128$ with 8 heads of width 16, for 2.2 M parameters. The backbone outputs one token per visible pulse. See Fig. 4 for a sketch of the architecture.

Query head (decoder). Held-out sensor positions are lifted by a NeRF-style Fourier position embedding [16] and attend to the encoded event through a single cross-attention block: queries come from the embedded positions, keys and values from the per-pulse tokens, followed by a feed-forward layer and a linear projection to the per-query outputs. The head adds 0.23 M parameters, for 2.5 M during pretraining, and is discarded afterwards, so only the 2.2 M encoder transfers downstream.

Pretraining objectives. Each query is scored by two objectives,

$$\mathcal{L} = \mathcal{L}_{\text{occ}} + \lambda_t \mathcal{L}_{\Delta t}, \quad (1)$$

where $\mathcal{L}_{\text{occ}}$ is a BCE on whether the queried sensor lights up after T , and $\mathcal{L}_{\Delta t}$ is a Huber loss on the first-hit time of hit sensors. The time term applies to hit queries only and λ_t is chosen as 1.

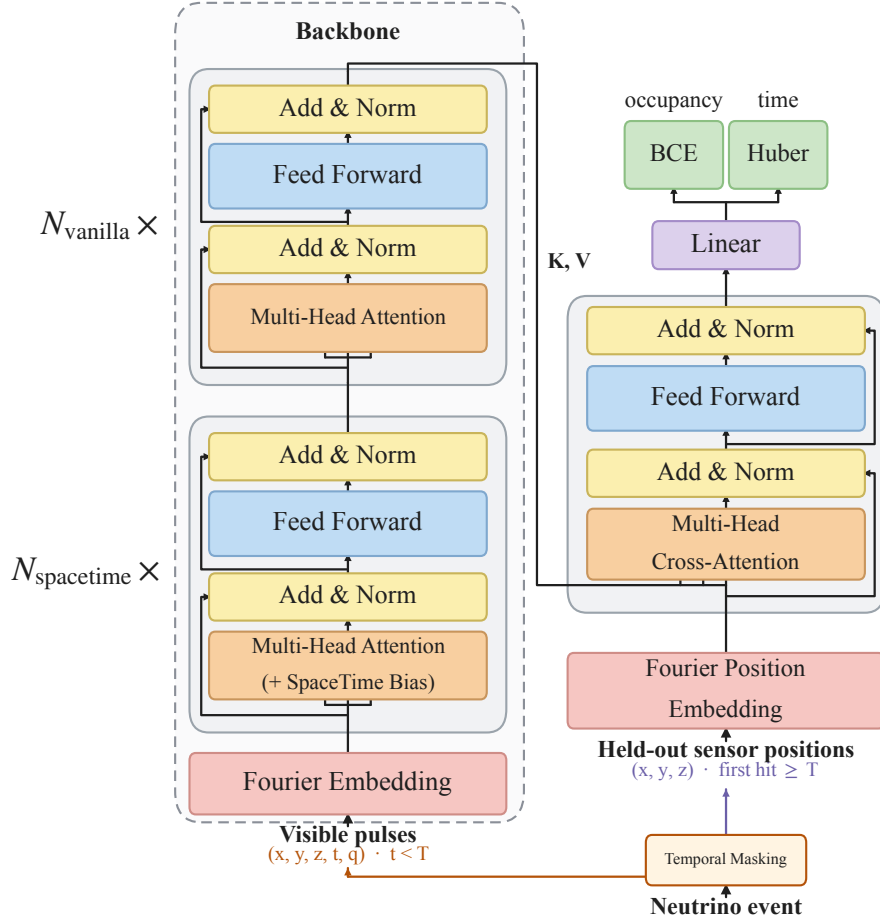


Figure 4: The CURTAIN pretraining architecture. Visible pulses enter the encoder backbone on the left; held-out sensor positions are embedded separately on the right and cross-attend to the encoded event, so the encoder never sees a mask token and the SpaceTime bias stays well defined.

Optimization. We train with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, no weight decay) at a learning rate of 1.4×10^{-3} for a batch size of 256 events, with no warmup; the learning rate is halved whenever the held-out SSL loss has not improved for 10 epochs. Training runs for at most 200 epochs, with early stopping on the same loss at patience 15. Gradients are clipped at a global norm of 1.0 and we train in full float32 on one to four NVIDIA H200s, taking 330 GPU hours for the largest pretraining set (3.9 million events).

Baselines. The Neptune and PolarBERT baselines use the same optimizer, schedule and budget.

C Finetuning details

For downstream tasks the query head is removed and the backbone is finetuned end-to-end with a fresh task head, using the same input convention as pretraining. For the DeepIce encoders, pre-trained and from scratch alike, the event embedding is the mean of the encoder’s output pulse tokens; Neptune and PolarBERT keep the readout of their original implementations. The task head is a layer normalization followed by a single linear layer, computed in float32. The energy task head predicts the logarithmic neutrino energy. The direction task head predicts four values: the three Cartesian components of the neutrino direction, normalized to a unit vector, and $\log_{10}(\kappa - 1)$, where κ is the concentration parameter of the vMF distribution.

Table 1: SSL pretraining schemes for neutrino telescopes. Neptune’s slashed entries read as track direction / cascade classification; our baseline uses the track-direction variant.

	PolarBERT	Neptune	CURTAIN
Which pulses	Random 25 % of pulses	All / random 75 %	Every pulse after a threshold time T
Which information	The OM token only	Time / time and space	All pulse information
How reconstructed	Cross-entropy over the OM vocabulary, plus an auxiliary total-charge regression	Smooth-L1 regression of the masked coordinates	Binary discrimination of active OMs against sampled negatives, plus an auxiliary first-hit-time regression
Transformer type	Encoder only	Encoder only	Encoder–decoder
Cross-detector	New vocabulary required	Untested	Demonstrated

For the pretrained encoders, the backbone is frozen for the first two epochs so that only the new task head is trained; the from-scratch DeepIce is trained end-to-end from the start. All finetuning runs use a batch size of 256 and a learning rate of 5×10^{-4} without weight decay, reduced on plateau (ReduceLRonPlateau from PyTorch [17]), and an exponential moving average of the weights [18] with a decay rate of 0.999. We validate every 300 steps and stop once the validation loss has not improved for 8 validations, or after at most 100 thousand steps. The cross-geometry finetune of the encoder pretrained on 3.9 million events additionally uses a 2,000-step learning-rate warm-up.

D Pretraining schemes compared

All three schemes hide part of an event and train the model to recover what was hidden, but they differ along four axes: which pulses are hidden, which of their information is hidden, what the model must then predict, and which architectures the scheme admits. Table 1 summarizes them.

PolarBERT. A random 25 % of pulses are masked. Each pulse is represented by a concatenation of a token identifying the OM that recorded it and a multi-layer perceptron (MLP) on the pulse features. PolarBERT hides the OM identifying token. The model then recovers it by cross-entropy over the vocabulary of OMs, with an auxiliary regression of the total charge. Because the vocabulary enumerates the sensors of one detector, carrying the pretrained model to another geometry means learning a new vocabulary.

Neptune. What is masked depends on the downstream task. For track direction the arrival time of every pulse is hidden; for cascade classification a random 75 % of pulses have both time and position hidden. In either case the model regresses the masked coordinates with a smooth-L1 loss. Our baseline uses the track-direction variant, with the arrival time of every pulse hidden. The scheme has no per-detector vocabulary and so carries over to another geometry in principle, but cross-detector transfer has not been reported.

CURTAIN. Every pulse after a sampled threshold time is hidden and all of its information discarded. Rather than reconstruct the hidden pulses, the model is queried at sensor positions and must separate OMs that light up after the cut from OMs that stay dark for the whole event. A query carries only a position, so the scheme depends on no sensor numbering and transfers to a detector it was not pretrained on (Figure 3, right).

Architecture. PolarBERT and Neptune mask pulses in place and use encoder-only transformers, so the masked pulses pass through the encoder’s attention. CURTAIN’s encoder sees only the visible pulses and its decoder is queried separately, leaving the encoder free of mask tokens and free to use an attention bias defined on space–time distance.