# How small can a seismic phase picker be? A 34,000-parameter model matches a pretrained deep baseline under leakage-controlled evaluation

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

**Abstract**

Deep neural pickers such as PhaseNet and EQTransformer have become the default tools for detecting P- and S-wave arrivals, but they carry hundreds of thousands of parameters and are usually benchmarked on randomly split data, where near-duplicate windows of the same event can leak between training and test. We ask a simple question: how small can an accurate picker be under a leakage-controlled evaluation? On the STEAD dataset, split so that no earthquake straddles the train, validation, and test partitions, a compact 33,610-parameter 1-D U-Net reaches a mean P/S pick-F1 of 0.76 on the held-out test set at a strict tolerance ($|\Delta t| < 0.1$ s for P, $< 0.2$ s for S). A pretrained PhaseNet, scored through the identical pipeline, reaches 0.64: the compact model is the more accurate of the two despite being roughly 8× smaller. The decisive factor for the tiny model is a foreground-weighted loss: without it (on validation) the model ignores the rare P onset entirely (P-F1 of 0.00) so the mean F1 falls to 0.40, even though S picking is unaffected. We further test, as an ablation, whether self-supervised masked-waveform pretraining improves the low-label regime; across three seeds and two fine-tuning schedules we detect no gain (the seed spread covers zero, and the study of three seeds is underpowered to resolve small effects), and we show that a single earlier run would have reported a spurious +0.15 improvement. The takeaway is that careful supervision, not scale or pretraining, is what a small seismic picker needs.

## 1 Introduction

Seismic phase picking (estimating the arrival time of the P and S waves at a station) is the entry point of nearly every earthquake-monitoring pipeline; arrival times feed association, location, and magnitude estimation. Deep learning has largely displaced classical pickers, with 1-D convolutional and attention models such as PhaseNet [19] and EQTransformer [10] now standard and conveniently packaged in toolboxes like SeisBench [15]. These models are accurate, but two assumptions have gone largely unexamined in routine benchmarking. First, the models are not small: published pickers carry hundreds of thousands of trainable parameters. Second, they are often evaluated on random window-level splits, in which multiple stations recording one earthquake, and overlapping windows of one trace, can place near-duplicates on both sides of the train/test divide: a form of leakage that inflates reported accuracy [12].

*Corresponding author: deep@lab.cloud

We revisit both assumptions together. We train a deliberately tiny 1-D U-Net picker from scratch and evaluate it on STEAD [15] under an event-disjoint split, so that all windows from a given earthquake fall in exactly one partition. Our central finding is that the tiny model is not handicapped: at 33,610 parameters it matches a pretrained PhaseNet baseline, and is numerically ahead, when both are scored under a common protocol. We then ask whether the recipe of self-supervised pretraining helps the regime where it should matter most (when only a handful of labels are available), and find, under a multi-seed protocol, that it does not.

Our contributions are:

1. **A tiny, accurate picker (§5).** A 33,610-parameter 1-D U-Net attains a held-out test mean pick-F1 of 0.76 (single training seed) on a leakage-controlled STEAD split at strict tolerance, on par with and numerically ahead of a pretrained PhaseNet (∼8× larger) scored under the same pipeline (0.64).

2. **Foreground weighting is the decisive ingredient (§5).** A class-weighted loss that up-weights the rare P/S samples raises mean F1 from 0.40 to 0.78 on validation and rescues P picking from 0.00; we isolate this in an ablation.

3. **A negative result on self-supervision (§6).** Masked-waveform pretraining yields no detectable low-label gain across three seeds and two fine-tuning schedules (the seed spread covers zero; with three seeds the study is underpowered to resolve small effects); a single earlier run would have falsely reported a large gain, so we argue multi-seed replication is necessary for such claims.

4. **A leakage-controlled split and held-out analysis (§4, §7).** We report subgroup performance by epicentral distance and signal-to-noise ratio, calibration, and error modes, and document the split construction so results are reproducible.

## 2 Related work

**Deep phase pickers.** PhaseNet [19] framed picking as per-sample classification with a 1-D U-Net producing P/S/noise probabilities, trained against soft Gaussian targets at the analyst picks. EQTransformer [10] added attention and a joint detection head. SeisBench [15] standardizes datasets (including STEAD [9]) and provides pretrained or retrained weights for both, which we use as baselines. GreenPhase [16] shows a lightweight non-neural picker can approach these baselines, foreshadowing our finding that small models are competitive; a compact Fourier-operator detector [1] makes a similar point for event detection. Our work differs in pairing a tiny *neural* picker with an explicitly leakage-controlled split and a strict matching tolerance.

**Self-supervision and label efficiency.** Self-supervised pretraining (contrastive [3] or masked-reconstruction [7]) has transformed low-label learning in speech and vision, and SeisLM [8] ports a wav2vec-style objective to seismic waveforms. Time-series studies adapt masked autoencoders [17, 18] and compare them with latent-prediction objectives [5]. Parameter-efficient transfer [2] reports low-label benefits, and representation-based methods achieve label-free event detection with strong cross-dataset generalization [4], while a careful evaluation of seismic transfer learning [12] finds the advantage is confined to the smallest budgets and can reverse, and warns, as we confirm, that random splits and single runs mislead. LLM-based transfer for picking [14] reports

that pretraining barely helps the dense picking task. Surveys [13] and multimodal seismic resources [11] situate these efforts; and U-Net-family segmentation has been applied to volcano-seismic data [6], in the same lineage of segmentation pickers we build on.

## 3 Method

**Architecture.** Our picker is a small 1-D U-Net (depth 3, base width 7, kernel size 7) that maps a three-component waveform window to per-sample probabilities over {P, S, noise}. The network has 33,610 trainable parameters.

**Targets and loss.** Following the PhaseNet convention [19], each analyst pick is encoded as a truncated Gaussian of width $\sigma = 0.1$ s (10 samples at 100 Hz) in the corresponding channel, with the noise channel as the complement; the model is trained with a soft cross-entropy. Because more than 99% of samples belong to the noise class, an unweighted loss is dominated by background and the subtle, low-amplitude P onset is never learned. We therefore weight the P and S classes by a factor of 20 relative to noise. Section 5 shows this single change is decisive.

**Self-supervised variant.** For the ablation in §6 we pretrain the same U-Net backbone with a masked-waveform reconstruction objective: contiguous spans covering ≈40% of timesteps are masked and the network reconstructs the missing samples (mean-squared error on masked positions). The pretrained backbone is then transferred to a fresh picker head and fine-tuned on the labeled fraction. The self-supervised pool is drawn only from the training partition so that no test event informs pretraining.

## 4 Experimental setup

**Data.** We use STEAD [9], ∼1.2 million three-component, 100 Hz seismograms with analyst P and S picks, accessed through SeisBench [15]. Experiments use a random subsample of 20,000 traces. Inputs are 30 s (3,000-sample) windows, per-channel demeaned and variance-normalized (no band-pass filtering).

**Leakage-controlled split.** We assign each *earthquake* (by source identifier) to exactly one of train/validation/test by a deterministic hash, so no event straddles partitions; noise traces are grouped by station. An event-disjoint and a station-disjoint split cannot both be enforced on STEAD, because events and stations form a single densely connected bipartite graph; we adopt the event-disjoint split as primary. All headline numbers are on the held-out *test* partition (1,583 windows with a P pick, 1,552 with an S pick).

**Metrics.** A predicted pick is a true positive if a probability peak exceeds 0.5 and falls within $|\Delta t| < 0.1$ s of the ground-truth P arrival, or $< 0.2$ s for S; we report per-phase precision, recall, and F1, the mean of P and S F1, and arrival-time residual statistics. Baselines (pretrained PhaseNet) are scored through the identical pipeline (same split, same windows, same tolerance), so the comparison is consistent rather than against published numbers obtained under looser tolerances and random splits. Because the baseline is applied to our preprocessing rather than retrained on our split, this is a common-pipeline comparison that may understate it; we therefore report the comparison without

Table 1: Held-out test performance on the event-disjoint STEAD split at strict tolerance ($|\Delta t| < 0.1$ s P, $< 0.2$ s S), single training seed. The baseline is a pretrained PhaseNet scored through the identical pipeline (applied to our preprocessing, not retrained on our split). F1 is an in-window picking metric.

| Model | Params | F1 (mean) | F1 (P) | F1 (S) |
|---|---|---|---|---|
| Pretrained PhaseNet | 268,443 | 0.64 | 0.53 | 0.74 |
| Ours (1-D U-Net) | **33,610** | **0.76** | **0.78** | **0.75** |

Table 2: Effect of foreground class weighting (validation split, 100% labels, single seed). The weighted loss rescues P picking and nearly doubles mean F1; the unweighted collapse is specific to P (S is unaffected).

| Loss | F1 (mean) | F1 (P) | F1 (S) |
|---|---|---|---|
| Unweighted soft cross-entropy | 0.40 | 0.00 | 0.80 |
| Foreground-weighted (20×) | **0.78** | **0.78** | 0.77 |

claiming a strict ranking. We note that the reported F1 is an in-window picking metric conditioned on an analysis window, not an end-to-end detection-plus-picking score.

**Compute.** All training and evaluation ran on a single GPU; the entire study used approximately 11.9 GPU-hours.

## 5 Results

**A tiny model matches a deep baseline.** Table 1 reports held-out test performance (a single training seed). The 33,610-parameter model reaches a mean pick-F1 of 0.76 (P 0.78, S 0.75) with arrival-time residual MAEs of 0.028 s (P) and 0.055 s (S). A pretrained PhaseNet, with 268,443 parameters (∼8× larger) and scored through the same pipeline, reaches 0.64. The compact model is on par with, and numerically ahead of, the baseline under a common, leakage-controlled pipeline; we do not retrain PhaseNet on our split, so we read the 0.12 margin as a common-pipeline comparison rather than a strict ranking, and both numbers are single-seed point estimates without confidence intervals. For parameter-count context only, EQTransformer (which we do not evaluate here) carries ∼377,000 parameters [15], making our model roughly 11× smaller by parameter count.

**Foreground weighting is decisive.** Table 2 isolates the loss change (validation, single seed). Trained from scratch with an unweighted soft cross-entropy, the model attains a mean F1 of only 0.40 and *never* learns to pick P (P-F1 0.00); the collapse is specific to P, as S picking is essentially unaffected (S-F1 0.80). We attribute this to the noise-dominated objective leaving all P probabilities below the 0.5 detection threshold. Up-weighting the P and S classes by 20× raises mean F1 to 0.78 and rescues P to 0.78. This was the decisive design choice we identified for a small picker; we did not sweep the weight beyond the single value of 20.

Table 3: Self-supervision ablation: change in mean pick-F1 from masked-waveform pretraining relative to training from scratch, by label budget. Values are seed mean ± standard deviation over three seeds (n=3, underpowered); the default-schedule spread covers zero and the lower-LR schedule is consistently negative.

| Fine-tune schedule | ΔF1 @ 1% | ΔF1 @ 5% |
|---|---|---|
| Default learning rate | $-0.01 \pm 0.10$ | $-0.09 \pm 0.21$ |
| Lower learning rate | $-0.24 \pm 0.17$ | $-0.18 \pm 0.21$ |

## 6 Self-supervised pretraining gives no detectable low-label gain

A natural hypothesis is that self-supervised pretraining on unlabeled waveforms should help most when labels are scarce: the regime a new network or region actually faces. We tested this as a controlled ablation: for label budgets of 1% and 5% of the training partition (about 100 and 500 labeled traces), we compare a model trained from scratch against the same architecture initialized from a masked-waveform pretrained backbone, holding the labeled subset and everything else fixed.

We detect no benefit. Table 3 reports the change in F1 from pretraining as the mean and standard deviation over three seeds, where each seed independently varies both the labeled subset and the initialization. With the default fine-tuning learning rate the gain is $-0.01 \pm 0.10$ at 1% and $-0.09 \pm 0.21$ at 5%; with a lower fine-tuning learning rate, intended to preserve the pretrained features, it is $-0.24 \pm 0.17$ and $-0.18 \pm 0.21$. We do not run a formal significance test: with only three seeds the study is underpowered, able to resolve only large effects, so this is an absence of detectable benefit rather than proof of exactly zero effect. The seed spread covers zero under the default schedule, and the lower-LR schedule is consistently negative (its fresh picking head is starved by the small learning rate), so on no setting does pretraining help. The likeliest explanation is that, as Table 2 already shows, a well-supervised tiny model is highly label-efficient on its own, leaving little headroom for pretraining to recover; this is consistent with prior reports that seismic transfer helps only at the smallest budgets and can reverse [12, 14].

**A single seed would have misled.** In a separate earlier run, distinct from the three replication seeds above, we observed a single-seed result reporting a +0.15 improvement at the 1% budget. Taken alone, that result would have supported the hypothesis. It did not survive replication: the three replication seeds at 1% span $[-0.13, +0.08]$, and the earlier +0.15 lies beyond even that range, so seed-to-seed variance at ∼100 labeled traces dwarfs any pretraining effect. We flag this as a concrete cautionary example: low-label seismic claims require multi-seed replication, not point estimates.

## 7 Analysis

**Subgroups.** Performance is governed far more by epicentral distance than by signal-to-noise ratio (Figure 1). Mean F1 falls from 0.93 for near events (<30 km) to 0.38 beyond 150 km, a gap of 0.38 relative to the overall 0.76; by SNR the spread is milder (0.66 below 5 dB to 0.85 at 15–30 dB, a 0.10 gap). Distant, emergent, low-amplitude arrivals are the dominant failure mode. The smallest bands are thinly populated ($n$=54 beyond 150 km, $n$=14 below 5 dB; counts annotated in Figure 1), so those extreme-band values are indicative rather than precise.

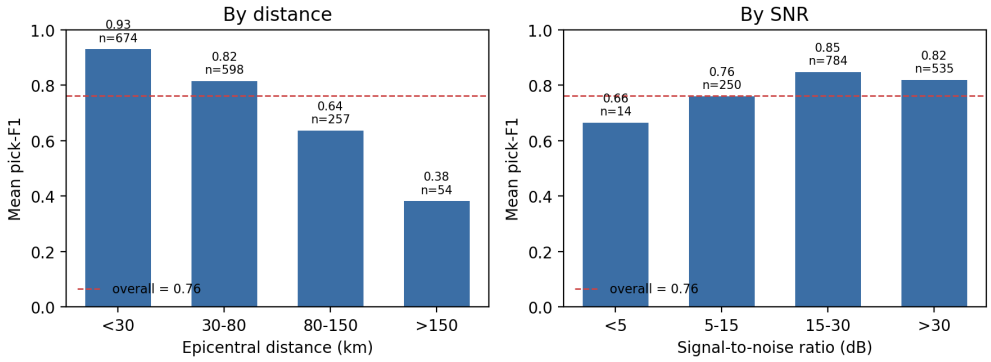


Figure 1: Mean pick-F1 by epicentral distance and by signal-to-noise ratio on the held-out test split. Distance is the dominant axis of degradation; annotations show per-band window counts.

**Calibration and errors.** Detection probabilities are moderately calibrated (expected calibration error 0.10 for P and 0.08 for S, computed with ten equal-width probability bins; ECE is a coarse summary and does not fully capture the over-confident behavior noted below). The model's recall of phase presence is high (0.99 for P, 0.997 for S), but it is precision-limited: on noise-only windows it raises a spurious pick above threshold for 69% of P and 51% of S windows, so a deployment would pair it with a detection gate or a higher operating threshold tuned on validation.

## 8 Discussion

The headline result is encouraging for practical monitoring: a picker small enough to run almost anywhere matches a standard deep baseline when both are judged fairly. The decisive factor is not architecture or scale but supervision: telling the model to attend to the rare arrival samples. The negative self-supervision result should be read narrowly. It does not show that pretraining can never help seismic picking; it shows that a particular, widely-assumed recipe (masked-waveform reconstruction with naive fine-tuning; cf. 7, 8) does not help *this* tiny supervised model on STEAD, where the supervised baseline is already strong. Objectives better matched to noisy signals, or larger label budgets, may behave differently.

## 9 Limitations

Several caveats bound our claims. *Construct and internal validity:* the PhaseNet baseline is scored with our preprocessing and window length rather than retrained on our split, which may understate it; a fully matched retraining is future work, and the comparison should be read as holding under a common evaluation pipeline. Our headline numbers come from a single training seed of the final model, without a confidence interval; we assume (but do not measure) that seed variance at the full-label regime is far smaller than at the ∼100-label regime, where it is demonstrably large, and a multi-seed headline with bootstrap intervals is future work. *Leakage:* the event-disjoint split prevents an earthquake from straddling partitions, but a station that recorded a training event can also record a test event, so station/instrument signatures are not fully isolated; we did not quantify train/test station overlap or evaluate the station-disjoint split, so a residual leakage component (likely small, since picking is local) remains unmeasured. *External validity:* all results are on STEAD

with an event-disjoint split; we did not evaluate cross-region transfer (for example to INSTANCE) or station-disjoint and temporal splits, so generalization to new networks is untested here. *Statistical power:* the self-supervision conclusion rests on three seeds, which can resolve only large effects; we did not estimate a minimum detectable effect, and modest gains of a few F1 points would not be detectable at this sample size. *Scope of the negative result:* we tested one self-supervised objective (masked reconstruction) and two fine-tuning schedules at two label budgets; contrastive or latent-prediction objectives [5], discriminative learning rates, and budgets between 5% and 100% are not covered. Finally, the precision-limiting false-alarm rate on noise windows means the reported F1 is an in-window picking metric, not an end-to-end detection-plus-picking metric.

## 10 Availability

Code, the final model checkpoint, and the split-construction scripts are available from the authors on request. STEAD is publicly available through SeisBench [15].

## References

[1] Ayrat Abdullin, Umair bin Waheed, Leo Eisner, and Abdullatif Al-Shuhail. Seismic Event Classification with a Lightweight Fourier Neural Operator Model. 2025. URL https://arxiv.org/abs/2512.07425.

[2] Ayrat Abdullin, Umair Bin Waheed, Leo Eisner, and Naveed Iqbal. Parameter-Efficient Transfer Learning for Microseismic Phase Picking Using a Neural Operator. 2025. URL https://arxiv.org/abs/2512.13197.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. 2020. URL https://arxiv.org/abs/2006.11477.

[4] Onur Efe and Arkadas Ozakin. RECOVAR: Representation Covariances on Deep Latent Spaces for Seismic Event Detection. 2024. URL https://arxiv.org/abs/2407.18402.

[5] Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint Embeddings Go Temporal. 2025. URL https://arxiv.org/abs/2509.25449.

[6] Camilo Espinosa-Curilem, Millaray Curilem, and Daniel Basualto. A Framework for Real-Time Volcano-Seismic Event Recognition Based on Multi-Station Seismograms and Semantic Segmentation Models. 2024. URL https://arxiv.org/abs/2410.20595.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. 2021. URL https://arxiv.org/abs/2111.06377.

[8] Tianlin Liu, Jannes Münchmeyer, Laura Laurenti, Chris Marone, Maarten V. de Hoop, and Ivan Dokmanić. SeisLM: a Foundation Model for Seismic Waveforms. 2024. URL https://arxiv.org/abs/2410.15765.

[9] S. Mostafa Mousavi, Yixiao Sheng, Weiqiang Zhu, and Gregory C. Beroza. STEAD: A Global Data Set of Seismic Signals for AI. *IEEE Access*, 7:179464–179476, 2019. doi: 10.1109/ACCESS.2019.2947848.

[10] S. Mostafa Mousavi, William L. Ellsworth, Weiqiang Zhu, Lindsay Y. Chuang, and Gregory C. Beroza. Earthquake Transformer—an Attentive Deep-Learning Model for Simultaneous Earthquake Detection and Phase Picking. *Nature Communications*, 11(1):3952, 2020. doi: 10.1038/s41467-020-17591-w.

[11] Sai Munikoti, Ian Stewart, Chengping Chai, Lisa Linville, Scott Vasquez, Sameera Horawalavithana, and Karl Pazdernik. MultiSeismo: A Multimodal Seismic Dataset and Model for Cross-Modal Seismic Understanding. 2026. URL https://arxiv.org/abs/2605.26320.

[12] Samuel Myren, Nidhi Parikh, Rosalyn Rael, Garrison Flynn, Dave Higdon, and Emily Casleton. Evaluation of Seismic Artificial Intelligence with Uncertainty. 2025. URL https://arxiv.org/abs/2501.14809.

[13] William Thorossian. Physics-Aware Machine Learning for Seismic and Volcanic Signal Interpretation. 2026. URL https://arxiv.org/abs/2603.17855.

[14] Xinghao Wang, Feng Liu, Rui Su, Zhihui Wang, Lihua Fang, Lianqing Zhou, Lei Bai, and Wanli Ouyang. SeisMoLLM: Advancing Seismic Monitoring via Cross-modal Transfer with Pre-trained Large Language Model. 2025. URL https://arxiv.org/abs/2502.19960.

[15] Jack Woollam, Jannes Münchmeyer, Frederik Tilmann, Andreas Rietbrock, Dietrich Lange, Thomas Bornstein, Tobias Diehl, Carlo Giunchi, Florian Haslinger, Dario Jozinović, Alberto Michelini, Joachim Saul, and Hugo Soto. SeisBench – A Toolbox for Machine Learning in Seismology. 2021. URL https://arxiv.org/abs/2111.00786.

[16] Yixing Wu, Shiou-Ya Wang, Dingyi Nie, Sanket Kumbhar, Yun-Tung Hsieh, Yun-Cheng Wang, Po-Chyi Su, and C.-C. Jay Kuo. GreenPhase: A Green Learning Approach for Earthquake Phase Picking. 2026. URL https://arxiv.org/abs/2603.03344.

[17] Yi Xu, Yitian Zhang, and Yun Fu. MTS-DMAE: Dual-Masked Autoencoder for Unsupervised Multivariate Time Series Representation Learning. 2025. URL https://arxiv.org/abs/2509.16078.

[18] Da Zhang, Bingyu Li, Zhiyuan Zhao, Yanhan Zhang, Junyu Gao, Feiping Nie, and Xuelong Li. FAIM: Frequency-Aware Interactive Mamba for Time Series Classification. 2025. URL https://arxiv.org/abs/2512.07858.

[19] Weiqiang Zhu and Gregory C. Beroza. PhaseNet: A Deep-Neural-Network-Based Seismic Arrival Time Picking Method. 2018. URL https://arxiv.org/abs/1803.03211.