















## RESEARCH ARTICLE

10.1029/2024JH000220

# From Labquakes to Megathrusts: Scaling Deep Learning Based Pickers Over 15 Orders of Magnitude

**Key Points:**

- We propose a workflow to enhance the performance of pre-trained seismic picking models on out-of-distribution data sets without retraining
- Data rescaling and prediction ensembling can strongly augment pre-trained seismic phase-picking models
- Rescaling makes seismic phase picking models trained on local seismicity directly applicable to quakes spanning over 15 orders of magnitude

Peidong Shi<sup>1</sup> , Men-Andrin Meier<sup>1</sup> , Linus Villiger<sup>1</sup> , Katinka Tuinstra<sup>1</sup> , Paul Antony Selvadurai<sup>1</sup> , Federica Lanza<sup>1</sup> , Sanyi Yuan<sup>2</sup> , Anne Obermann<sup>1</sup> , Maria Mesimeri<sup>1,3</sup> , Jannes Münchmeyer<sup>4</sup> , Patrick Bianchi<sup>1</sup> , and Stefan Wiemer<sup>1</sup> 

<sup>1</sup>Swiss Seismological Service, ETH Zürich, Zürich, Switzerland, <sup>2</sup>College of Geophysics, China University of Petroleum Beijing, Beijing, China, <sup>3</sup>Bedretto Underground Laboratory for Geosciences and Geoenergies, ETH Zürich, Zürich, Switzerland, <sup>4</sup>University of Grenoble Alpes, University of Savoie Mont Blanc, CNRS, IRD, University of Gustave Eiffel, ISTerre, Grenoble, France

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

P. Shi,  
peidong.shi@sed.ethz.ch

**Citation:**

Shi, P., Meier, M.-A., Villiger, L., Tuinstra, K., Selvadurai, P. A., Lanza, F., et al. (2024). From labquakes to megathrusts: Scaling deep learning based pickers over 15 orders of magnitude. *Journal of Geophysical Research: Machine Learning and Computation*, 1, e2024JH000220. <https://doi.org/10.1029/2024JH000220>

Received 2 APR 2024

Accepted 3 SEP 2024

**Author Contributions:**

**Conceptualization:** Peidong Shi, Men-Andrin Meier, Linus Villiger

**Data curation:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan, Anne Obermann, Maria Mesimeri, Patrick Bianchi

**Formal analysis:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan,

**Abstract** The application of machine learning techniques in seismology has greatly advanced seismological analysis, especially for earthquake detection and seismic phase picking. However, machine learning approaches still face challenges in generalizing to data sets that differ from their original training setting. Previous studies focused on retraining or transfer-learning models for these scenarios, but require high-quality labeled data sets. This paper demonstrates a new approach for augmenting already trained models without the need for additional training data. We propose four strategies—rescaling, model aggregation, shifting, and filtering—to enhance the performance of pre-trained models on out-of-distribution data sets. We further devise various methodologies to ensemble the individual predictions from these strategies to obtain a final unified prediction result featuring prediction robustness and detection sensitivity. We develop an open-source Python module **quakephase** that implements these methods and can flexibly process input continuous seismic data of any sampling rate. With **quakephase** and pre-trained ML models from SeisBench, we perform systematic benchmark tests on data recorded by different types of instruments, ranging from acoustic emission sensors to distributed acoustic sensing, and collected at different scales, spanning from laboratory acoustic emission events to major tectonic earthquakes. Our tests highlight that rescaling is essential for dealing with small-magnitude seismic events recorded at high sampling rates as well as larger magnitude events having long coda and remote events with long wave trains. Our results demonstrate that the proposed methods are effective in augmenting pre-trained models for out-of-distribution data sets, especially in scenarios with limited labeled data for transfer learning.

**Plain Language Summary** Machine learning has revolutionized earthquake detection and arrival time picking, relying on vast amounts of accurately labeled data for model development and training. However, when faced with new, unique data sets, the lack of labeled information poses a significant challenge. In this study, we introduce a method to enhance the performance of pre-trained machine learning models on such exotic data sets, even in the absence of labeled data. Our approach does not involve creating new models; instead, it focuses on enhancing and aggregating existing pre-trained models to tackle the quandary of missing labeled data. Our comprehensive benchmark tests underline that machine learning models, initially trained for tectonic earthquakes, can be effectively repurposed to analyze events from labquakes and tiny induced earthquakes to megathrusts recorded by various instruments and at various sampling frequencies.

## 1. Introduction

In recent years, we have witnessed a remarkable surge in the application of machine learning (ML) techniques in earthquake monitoring and characterization, especially in the fields of earthquake detection and phase picking (Mousavi & Beroza, 2023; Münchmeyer et al., 2022; Ross et al., 2018; Saad et al., 2020; Shi et al., 2021; Wang et al., 2019; S. Yuan et al., 2018; Zhu, Hou, et al., 2023). ML has established itself as a standard methodology for analyzing seismic data and constructing enhanced high-resolution earthquake catalogs, thereby significantly advancing seismological studies and fostering new insights into earthquake dynamics and hazard assessment (Münchmeyer, 2024; Saad et al., 2021; Spallarossa et al., 2021; Tan et al., 2021).

Despite these advancements, challenges persist in the application of ML techniques within the field of seismology. A notable challenge arises from the inherent limitations of ML models to the statistical distribution of their training data sets. While ML models exhibit excellent performance on data sets with statistical features akin

© 2024 The Author(s). Journal of Geophysical Research: Machine Learning and Computation published by Wiley Periodicals LLC on behalf of American Geophysical Union.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Anne Obermann, Jannes Münchmeyer, Stefan Wiemer  
**Funding acquisition:** Stefan Wiemer  
**Investigation:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan, Anne Obermann, Maria Mesimeri, Jannes Münchmeyer, Patrick Bianchi, Stefan Wiemer  
**Methodology:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Jannes Münchmeyer  
**Project administration:** Stefan Wiemer  
**Resources:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan, Anne Obermann, Maria Mesimeri, Patrick Bianchi  
**Software:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Jannes Münchmeyer  
**Supervision:** Stefan Wiemer  
**Validation:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan, Anne Obermann, Jannes Münchmeyer  
**Visualization:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan, Maria Mesimeri  
**Writing – original draft:** Peidong Shi  
**Writing – review & editing:** Peidong Shi, Men-Andrin Meier, Linus Villiger, Katinka Tuinstra, Paul Antony Selvadurai, Federica Lanza, Sanyi Yuan, Anne Obermann, Maria Mesimeri, Jannes Münchmeyer, Patrick Bianchi, Stefan Wiemer

to the training data, their efficacy may significantly degrade when confronted with data sets possessing distinct properties—a phenomenon commonly referred to as out-of-distribution (OOD) generalization in the ML community (Arjovsky, 2021; Hendrycks & Gimpel, 2016).

In the domain of seismology, the majority of available open-source labeled data sets are predominantly focused on tectonic earthquakes with magnitudes ranging from 0 to 5, while small-scale events (e.g., labquakes) and megathrust events of different frequency ranges, epicentral distances, and noise conditions are missing or under-represented in most data sets (Chen et al., 2024; Mousavi et al., 2019; Ni et al., 2023; Woollam et al., 2022). Consequently, prevalent ML models designed for earthquake detection and phase picking are trained exclusively on these tectonic earthquakes (Mousavi et al., 2020; Ross et al., 2018; Wang et al., 2019; Woollam et al., 2022; Zhu & Beroza, 2019). This specialization renders abundant small earthquakes and rare major earthquakes outside the norm, thereby constituting the de facto OOD range for these pre-trained models. Given the potential devastation associated with major tectonic earthquakes, their characterization demands a high level of confidence. Simultaneously, understanding the rupture behaviors of small-scale rock failure events, including labquakes from rock-physics and underground laboratories and injection-induced microseismic events, though non-hazardous, remains pivotal for comprehending rupture dynamics in controlled environments (Selvadurai, 2019; Villiger et al., 2020). Consequently, there is a compelling need for robust ML approaches capable of reliably characterizing both ends of this seismic spectrum.

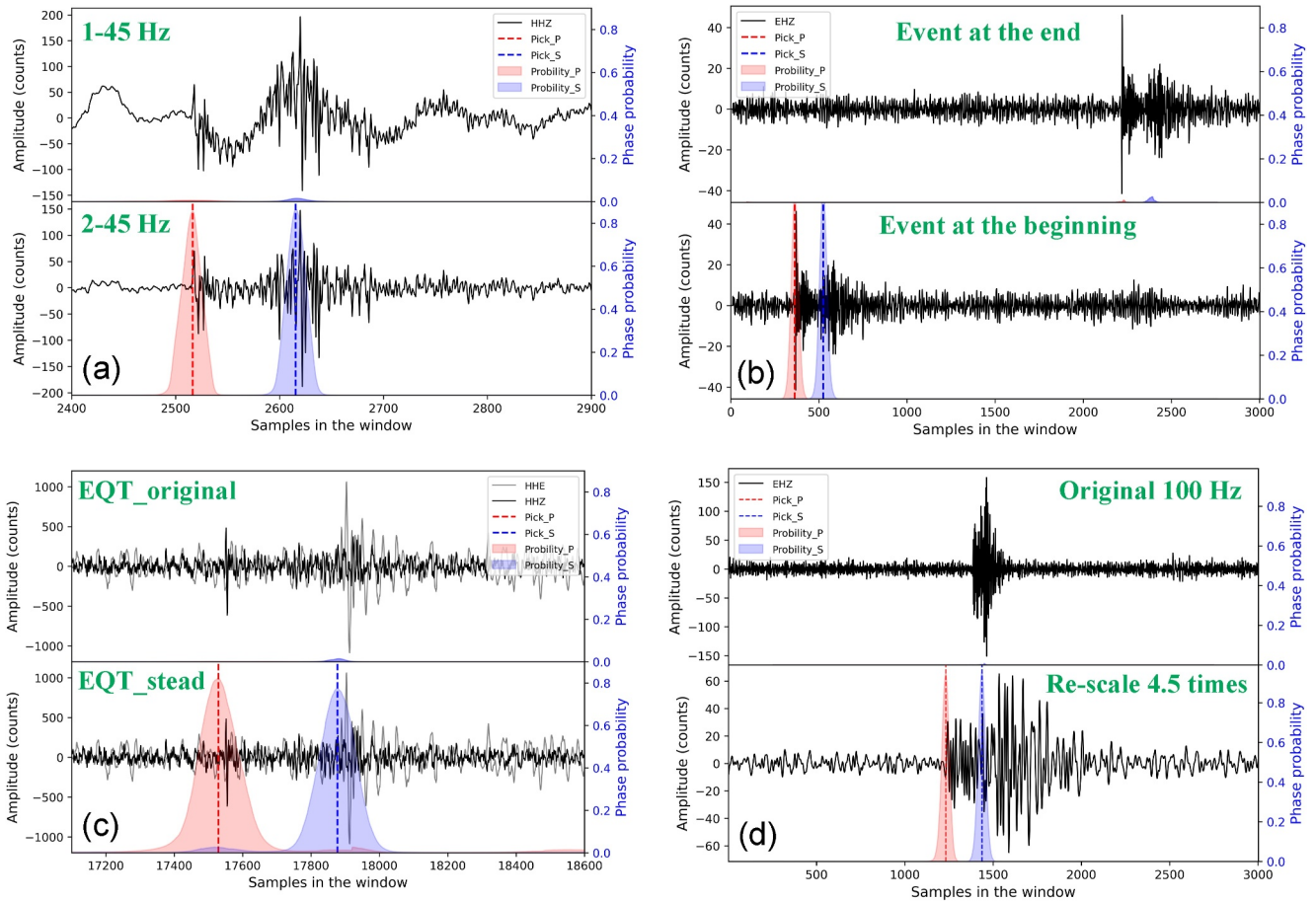
To deal with the generalization issue of ML on OOD data sets, previous studies have focused on employing transfer-learning and re-training techniques to enhance the generalization capability of pre-trained models (Chai et al., 2020; Lapins et al., 2021). However, these approaches require high-quality labeled data for model fine-tuning, a resource that is often scarce and demands extensive manual compilation efforts. This limits the applicability and performance of ML models. Recently, some studies have employed short sliding window steps to finely scan seismic data and combined results from different models to increase the model performance without additional training data (Park et al., 2023; C. Yuan et al., 2023). Nevertheless, these attempts are still limited by the capabilities of pre-trained models and are not universally applicable to data sets of varying scales or distinct waveform and epicentral distances, such as those encountered when studying, for example, labquakes from rock-mechanics and underground laboratories or microseismic events recorded in geothermal projects.

In this paper, we develop new methodologies and tools to address the OOD generalization challenges encountered by pre-trained ML models, thus enabling their application across scales for which they were not originally trained. In subsequent sections, we present the techniques developed to enhance the generalization ability of ML models and outline methodologies for ensembling individual ML predictions to derive a unified result. We also introduce an open-source Python package, **quakephase**, which integrates these techniques and ensemble methods to streamline the processing of continuous data across different scales using various pre-trained ML models from SeisBench (Woollam et al., 2022). To demonstrate the efficacy of our proposed methods, we conduct systematic benchmark tests on seismic data spanning diverse scales—from labquakes with a magnitude as low as  $-8$  to major tectonic earthquakes with a magnitude of  $7$ —recorded by various instruments, including AE sensors, geophones, broadband seismometers, strong motion instruments, and distributed acoustic sensing (DAS). Our tests demonstrate the effectiveness of the proposed methodologies in significantly improving the performance of pre-trained models on OOD data sets, extending their applicability from the regional scales to the entire spectrum of earthquake magnitudes for which ML models are not originally trained and applicable.

## 2. Method

In this section, we present a comprehensive suite of approaches designed to enhance the performance and generalization ability of pre-trained ML models on OOD data sets. The proposed methodologies encompass (a) data rescaling, (b) model aggregation, (c) filtering, and (d) time window shifting (Figure 1).

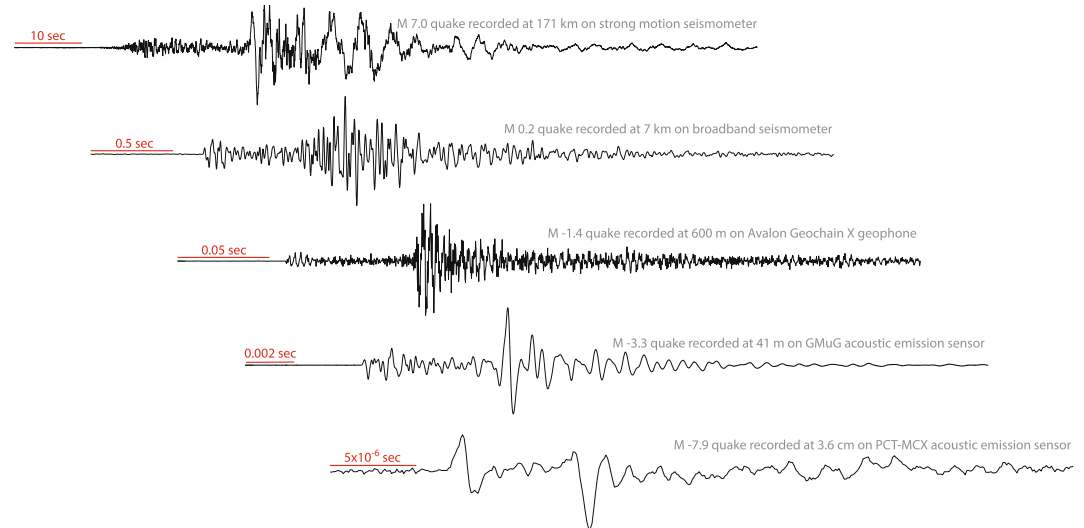
**Rescaling:** The inherent self-similarity observed in earthquake processes, where similar waveforms are emitted at different frequency ranges irrespective of source scales (Ide, 2019; Kwiatek et al., 2011; Manthei & Plenkers, 2018; Meier et al., 2017; Prieto et al., 2004; Selvadurai, 2019), forms the basis for our rescaling approach (Figure 2). While earthquake coda waves are highly dependent on the scattering properties of the crust, the direct P- and S-phases, which are frequently used for locating local or regional earthquakes, are primarily influenced by source processes (e.g., the rise time, source time function, and frequency content of the sources). These direct phases exhibit a lower degree of waveform complexity and variability compared to coda waves and usually



**Figure 1.** Strategies to enable pre-trained ML models transferable to OOD and exotic data sets. Each subfigure contrasts the ML predictions generated by the original model (EQTransformer or PhaseNet) using default parameters in the top panel, with the augmented prediction outcomes obtained by applying four proposed strategies: (a) filtering, (b) shifting, (c) model ensemble (S-phase arrivals are more clear on horizontal component waveform plots which are shown in gray color), and (d) rescaling, showcased in the bottom panel. Phase probabilities and picks generated by ML are plotted and overlaid on event waveforms, with P-phases highlighted in red and S-phases in blue.

demonstrate self-similarity across large and small earthquakes (Figure 2). A comparison of earthquake waveforms across the multiple scales evaluated in this study confirms the apparent self-similarity feature of the sources, despite the wide range of source magnitudes, sampling frequencies, and recording instruments used (Figure 2 and Table 2). These features are learned by ML models during training, provided that the training data contains sufficient variability. The current mainstream ML phase-picking models directly process earthquake waveform samples of fixed length for identifying and classifying the P- and S-phases without considering the frequency or source scale properties (Mousavi et al., 2020; Ross et al., 2018; Zhu & Beroza, 2019). We leveraged this approach as a scale-independence property using the apparent self-similarity property of earthquake processes. To this end, we develop a rescaling approach to enhance ML model performance. This involves first resampling the recorded seismic waveforms ( $s_o$ ) to a target sampling rate ( $s_i$ ), and subsequently feeding the resampled waveform into the used ML model with fixed-length input data samples (Table S1 in Supporting Information S1) for phase picking (Figure 1d).

Due to the fixed-length input data samples of ML models, when up-sampled (target sampling rate larger than original data sampling rate, i.e., upscaling), features of earthquake waveforms are magnified, providing the ML model with a zoom-in into waveform characteristics (Movie S1). On the other hand, when down-sampled (target sampling rate smaller than original data sampling rate, i.e., downscaling), features of earthquake waveforms are compressed, providing the ML model with zoomed-out wave trains (Movie S2). Lowpass filtering will be applied to avoid aliasing when applying down-sampling. This usually limits a lower down-sampling threshold we can reach, below which earthquake signals might be invisible. Finally, the rescaling rate is defined by  $R = \frac{t_m}{t_i} = \frac{s_i}{s_m}$ ,

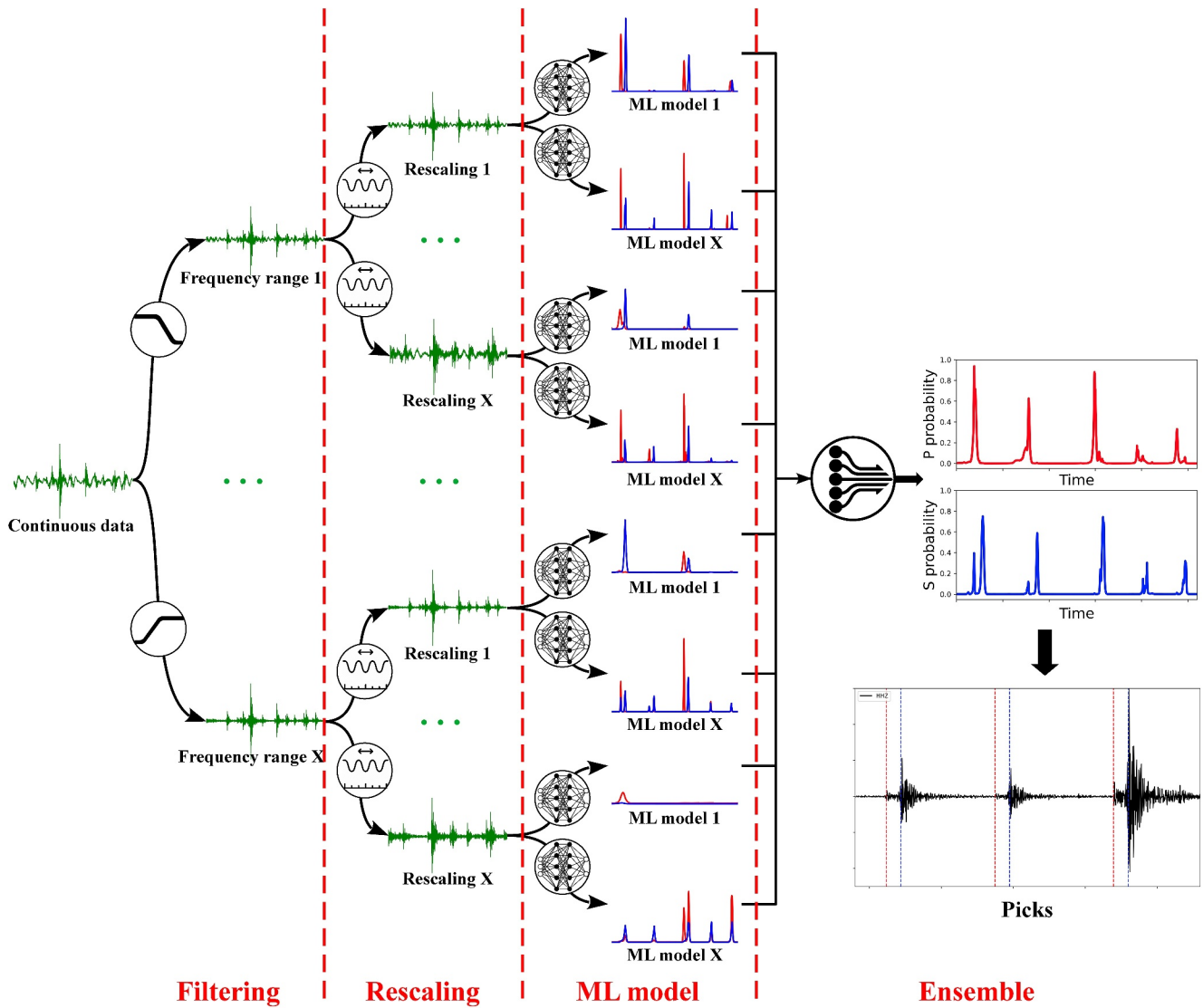


**Figure 2.** Raw seismograms of earthquakes across a wide range of magnitudes (from M 7.0 to M  $-7.9$ ) analyzed in this study (Table 2). The seismograms are displayed from top to bottom as follows: (1) M 7.0 Kumamoto earthquake recorded by a strong motion sensor, (2) M 0.2 induced earthquake in the Iceland Hengill geothermal field captured by a broadband seismometer, (3) M  $-1.4$  microseismic event at the Utah FORGE geothermal site recorded by a three-component downhole geophone, (4) M  $-3.3$  pico-earthquake induced by fluid injection at the Bederreto underground lab, detected by an acoustic emission sensor, and (5) M  $-7.9$  laboratory-generated acoustic emission event recorded by an acoustic emission sensor. The respective time scales, magnitudes, source-receiver distances, and sensor types are annotated alongside each seismogram.

where  $t_m$  and  $t_i$  are the time window length of the ML model designed during the model training process (e.g., 30 s for PhaseNet, Table S1 in Supporting Information S1) and the time window length of the actual input data respectively, and  $s_m$  and  $s_i$  are the data sampling rate used for training the ML model (100 Hz for PhaseNet, see Table S1 in Supporting Information S1) and the designated data sampling rate of the actual input data respectively. When rescaling is applied, it primarily impacts two components of the input data for ML models. The first and more apparent component is the P-to-S duration, represented by the number of samples between the P-arrival and the S-arrival. The second, often overlooked component, is the number of samples per phase cycle, which can be interpreted as frequency content as perceived by ML models. Together, these adjustments enable the successful application or improved performance of pre-trained models on OOD data sets, where these models previously failed. This approach allows flexible manipulation of earthquake waveforms, providing ML models with either enlarged or compressed representations of earthquake waveforms to simulate the diverse waveforms encountered during model training. From the perspective of the model, this rescaling means making the model sensitive to higher frequencies (upscaling) or lower frequencies (downscaling) than the original training data.

**Model aggregation:** ML-based seismic phase picking models have been built using different model architectures and trained on different data sets (Woollam et al., 2022). These models exhibit distinct performances when applied to a specific data set (Münchmeyer et al., 2022). Acknowledging the variability in performance among ML-based seismic phase picking models, we suggest to combine the prediction results from many different models. Combining prediction results from multiple models gives us more possibilities to make reliable predictions even when some models are not working properly due to generalization issues (Figure 1c, C. Yuan et al. (2023)).

**Filtering:** Many ML models are trained on seismic recordings filtered within specific frequency ranges (Table S1 in Supporting Information S1). Applying the same frequency range as the training data to field data sets may be reasonable; however, ambient noise conditions and varying signal frequencies, particularly for small events, necessitate careful consideration. Ambient noise conditions present significant varieties at different sites. In addition, for small events recorded with a high sampling rate (e.g., labquakes and microseismic events), the frequency range of recorded signals (kHz or MHz) can be much higher than that used during training. Filtering can significantly impact model performance, particularly in the presence of strong noises within certain frequency ranges (Figure 1a). We thus recommend choosing a suitable frequency range or combining results from different frequency ranges to optimize the model performance.



**Figure 3.** The quakephase workflow delineates the data processing flow from continuous seismic data to the final ensembled phase probabilities and phase picks. The workflow encompasses four primary processing modules: filtering, rescaling, application of the ML models, and ensembling. The final picks are extracted from the unified phase probabilities ensembled from the prediction results of different strategies, utilizing a designated ensemble method such as PCA, max, etc.

*Shifting:* Most ML models make predictions based on input windows with a fixed number of samples (prediction window, Table S1 in Supporting Information S1). Sensitivity to the positioning of event waveforms within fixed-length input samples has been observed in many ML models (Figure 1b, Park et al. (2023)). In practice, employing a sliding-window approach with appropriate overlap ratios can enhance model performance by exposing seismic events in multiple prediction windows at different positions. A high overlap ratio can ensure that a seismic event appears in many different sections of the prediction window thus increasing model performance, but at the expense of higher computational cost (Figure 1b). Overlapping ratios can significantly affect the model performance, particularly for models lacking generalization ability on new data sets (Park et al., 2023).

These approaches can be effectively ensembled to produce more robust or sensitive predictions. We devise and introduce various ensembling methods, including principal component analysis (PCA, Abdi and Williams (2010)), semblance (Staněk et al., 2015; C. Yuan et al., 2023), and ensemble maximum (Woollam et al., 2022; C. Yuan et al., 2023) to amalgamate the predicted phase probabilities from the various strategies and parameters into a unified prediction (Figure 3 and Figure S1 in Supporting Information S1; Table 1). The final unified phase probabilities can be employed to extract and associate phase picks (Münchmeyer, 2024) or to locate

**Table 1**  
*Ensemble Methods Utilized in **Quakephase** for Integrating Phase Probabilities From Different Predictions Into a Unified Final Prediction*

Ensemble	Equation	Applicable scenarios
PCA	$Prob(t) = PCA_{n=1}(Prob_i(t))$	Balance between sensitivity and reliability
max	$Prob(t) = \max_i(Prob_i(t))$	Maximize sensitivity
semblance	$Prob(t) = \frac{(\sum_i Prob_i(t))^2}{N \sum_i Prob_i^2(t)}$	Balance between sensitivity and reliability
median	$Prob(t) = \text{median}_i(Prob_i(t))$	Balance between sensitivity and reliability
mean	$Prob(t) = \text{mean}_i(Prob_i(t))$	Balance between sensitivity and reliability
prod	$Prob(t) = \prod_i Prob_i(t)$	Maximize reliability
min	$Prob(t) = \min_i(Prob_i(t))$	Maximize reliability

*Note.*  $Prob_i(t)$  represents the  $i$ -th ML probabilities curve from a specific combination of model and parameters (filtering and rescaling).  $N$  is the total number of ensemble members (the combination of ML models and strategy parameters).  $Prob(t)$  is the final ensembled probabilities at each time sample

seismic events via back-projection (Shi, Angus, et al., 2019; Shi, Nowacki, et al., 2019; Shi et al., 2022). Of these ensemble methods, the maximum ensemble takes the maximum probability among the predictions from different model and parameter combinations at each time sample as the final ensemble output (Table 1), thus focusing on improving model sensitivity and event detection rates. In comparison, the PCA ensemble first performs principal component analysis across all ensemble members (ML predictions from all the combinations of different models and parameters) and adopts the first principal component which maximizes the data variance at each sample as the final ensemble output (Table 1). In this way, the models producing more consistent prediction results among each other will be automatically given more weights than the models generating outlier-like predictions during ensembling. This achieves a balance between sensitivity and prediction robustness because significant probabilities in the final outputs will require large enough probabilities from most of the ensemble members (Figure 3 and Figure S1 in Supporting Information S1). Other implemented ensemble methods and the corresponding calculation methodologies and applicable scenarios are listed in Table 1. In practical applications, we recommend selecting an ensemble method based on specific monitoring requirements, such as enhancing detection sensitivity or improving picking robustness, and performing a benchmark test on a subset of the applied data set where ensemble performance and strategy parameters can be cross-checked. The model augmentation strategies together with ensemble methods can better characterize seismic phases, especially for tiny earthquakes with low signal-to-noise ratios (SNR).

We develop a Python-based module, **quakephase** that implements the aforementioned strategies and ensemble methods to promote model performance (Figure 3 and Figure S1 in Supporting Information S1). The input seismic data will be initially filtered into various frequency ranges. Subsequently, the filtered data will undergo resampling/rescaling according to the preset rescaling rates. Rescaled data will be segmented based on the ML prediction window size and overlapping ratio before being input into different ML models. The final phase probabilities will be determined by integrating the ML predictions from the various models using a predefined ensemble method (Figure 3). **Quakephase** can process continuous seismic data of any sampling rate and generate seismic phase picks and/or unified continuous phase probabilities according to the user-defined combinations of the various strategies and a chosen ensemble method (Figure 3 and Figure S1 in Supporting Information S1). **Quakephase** integrates the numerous pre-trained ML models available from SeisBench (Woollam et al., 2022), providing flexibility in aggregating a multitude of popular original and re-trained models in the seismology community (e.g., PhaseNet (Zhu & Beroza, 2019), EQTransformer (Mousavi et al., 2020), GPD (Ross et al., 2018)). In the following sections, we assess the performance of **quakephase** and benchmark the proposed strategies using diverse OOD or challenging data sets that have not been used in model training (Table 2 and Figure 2). These exotic data sets cover a wide magnitude range of seismic events (from  $M_w$  -8 to 7) and comprise data recorded using different sampling rates (from 100 Hz to 10 MHz) and instruments (e.g., geophones, seismometers, acoustic emission sensors, and distributed acoustic sensing).

**Table 2**  
*Comprehensive Overview of the OOD or Challenging Data Sets Evaluated in This Study*

Data set	Magnitude	Sampling rate	Instrument	Mode	Samples
Kumamoto	7.0	100 Hz	Accelerometer	Segments	407
COSEISMIQ	-0.6–3.8	100 Hz	Broadband seismometer	Segments	38,205
VI-EDA	-1.4–0.8	100 Hz	Broadband seismometer	Continuous	–
DAS	0.0	2,000 Hz	Downhole DAS	Segments	965
FORGE	-2.2–0.6	4,000 Hz	Downhole geophone	Continuous	–
Reflection survey	-2.1–-1.5	1,000 Hz	Surface geophone	Segments	1,747
BULGG	-4.6–-2.1	200 KHz	AE sensor	Segments	10,212
Labquake	-7.8–-7.0	10 MHz	AE sensor	Segments	16,113

*Note.* Some representative event waveforms from the various data sets are presented in Figure 2.

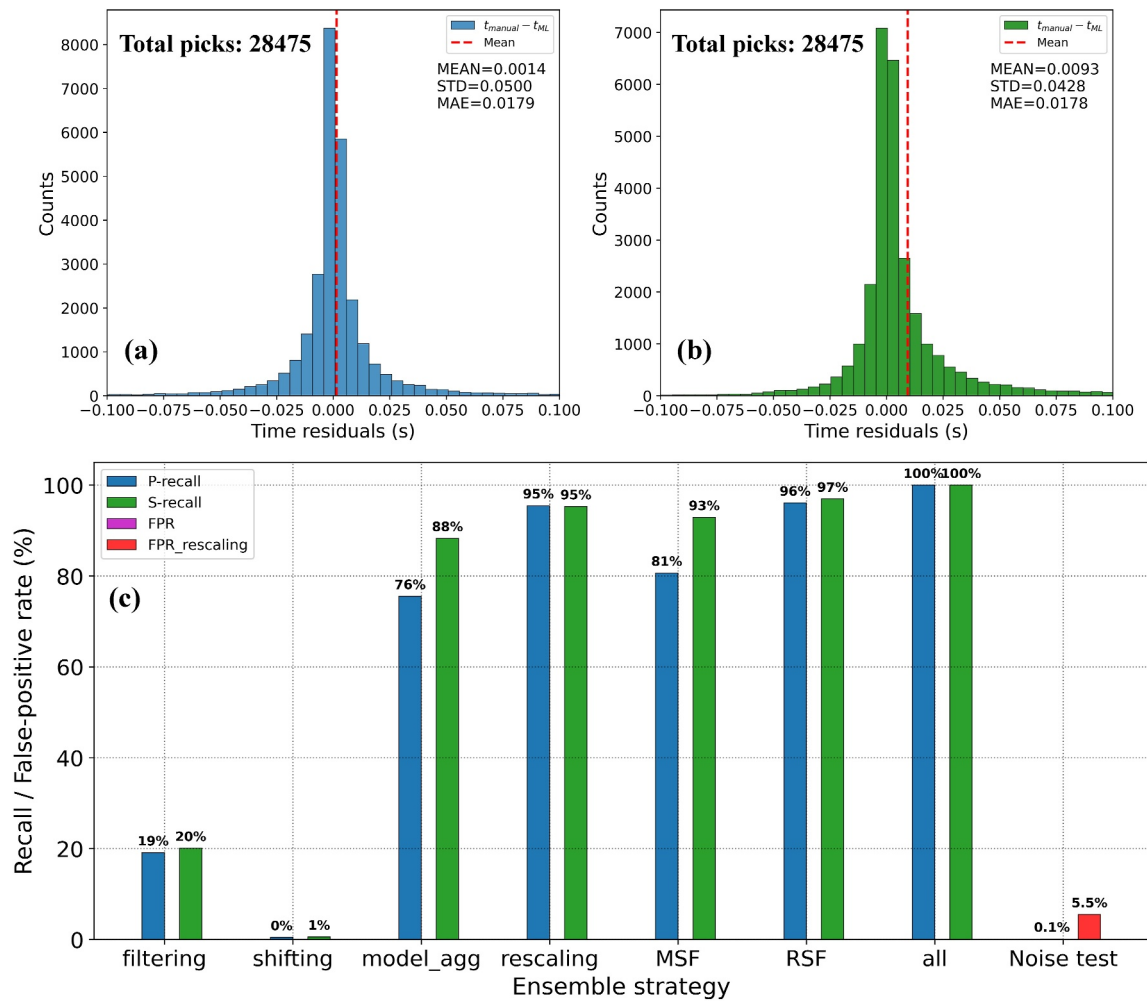
### 3. Applications and Benchmarks on Out-Of-Distribution Data Sets

#### 3.1. Natural and Induced Earthquakes in the Hengill Geothermal Field

To monitor induced earthquakes associated with geothermal production, a dense seismic network (COSEISMIQ) consisting of approximately 50 broadband and short-period seismometers was deployed in the seismically active Hengill geothermal field in the Reykjanes Peninsula of Iceland (Grigoli et al., 2022; Obermann, Wu, et al., 2022). During the full operation of the COSEISMIQ network (from 2018-12-01 to 2021-01-31), over 10,000 seismic events including both natural and induced earthquakes have been captured with a sampling rate of 100 Hz (Grigoli et al., 2022; Nooshiri et al., 2022; Swiss Seismological Service (SED) at ETH Zurich, 2018). We analyze the detected natural and induced earthquakes using a popular pre-trained ML model EQTransformer (Mousavi et al., 2020) and observe that EQTransformer is unable to pick a total of 28,475 event waveforms with the default parameter settings (picking threshold 0.1, refer to Table S1, Figure S2 in Supporting Information S1), which accounts for approximately 15% of all event waveforms (totaling around 200,000). Several factors may lead to the failure application of pre-trained models on this data set. Firstly, seismic events of smaller magnitudes possess distinct spectrum properties and waveform characteristics and usually exhibit a lower SNR, making it challenging for the EQTransformer to generalize effectively to this event type. Moreover, some induced microseismic events occurring in the shallow crust are located close to monitoring stations, resulting in short P-to-S times and wave trains. Their waveform features and the shorter inter-phase times fall outside the distribution of model training sets, which leads to the failure of pre-trained models.

We conduct benchmark tests to assess the efficacy of our proposed strategies in enhancing model prediction performance against the original EQTransformer model without any augmentation strategies on this data set (Figure 4). Our tests reveal that shifting and filtering marginally improve model performance by 20% and 1% enhancement, respectively (Figure 4). In contrast, model aggregation and rescaling exhibit substantial improvements on the same data set, improving recall rate (event detection rate) from 0 to approximately 80% and 95%, respectively (Figure 4). By combining all four strategies (ensemble nine rescaling rates, five models, four frequency bands, and higher overlapping ratio, Table S2 in Supporting Information S1), we achieved a 100% recall rate with high picking accuracy, defined as the majority of picks falling within five samples (of original data sampling rate, hereafter the same) relative to manual picks, using a pre-trained ML model that previously failed (Figure 4, Figure S2 in Supporting Information S1). Notably, employing rescaling alone yields a high recall rate and picking accuracy (over 95% recall rate), underscoring its effectiveness as the most crucial factor in augmenting ML model performance.

To assess the impact of rescaling and ensemble schemes on picking precision and false positive rate, we extract 9,730 noise samples extracted from the same COSEISMIQ data set and apply **quakephase** to these noise samples both with and without rescaling (Figure 4c). Without rescaling, **quakephase** identified 5 P-picks and 7 S-picks using a 0.1 picking threshold, resulting in a false positive rate of 0.1%. Upon applying rescaling, **quakephase** detects 331 P-picks and 208 S-picks with the same picking threshold, yielding a false positive rate of 5.5%. Importantly, the implementation of rescaling does not significantly increase the false positive rate, even when employing a notably low picking threshold of 0.1. Concurrently, it markedly enhances the picking sensitivity



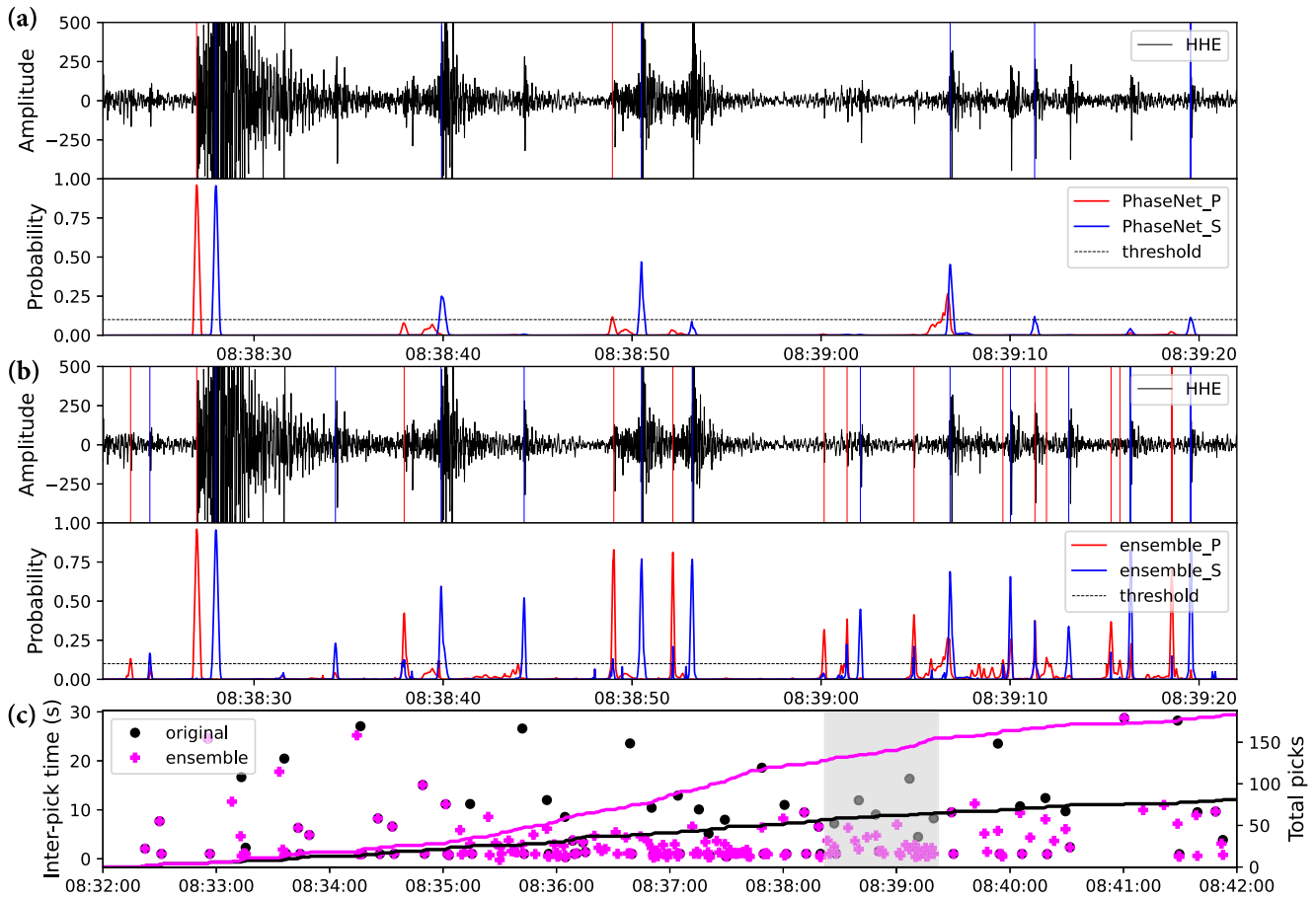
**Figure 4.** Performance evaluation on the COSEISMIQ data set involving 28,475 event waveforms, with both P and S manual picks serving as ground-truth labels for benchmarking (Figure S2 in Supporting Information S1). (a) Distribution of P-phase ML picking residuals compared to manual picks. (b) Distribution of S-phase ML picking residuals relative to manual picks. (c) Recall rate, defined as the proportion of successful picks within a 0.5-s difference relative to manual picks, assessed across various strategies or combinations thereof. **model\_agg** denotes the model aggregation strategy, **MSF** indicates the combination of model aggregation, shifting, and filtering strategies, while **RSF** represents the combination of rescaling, shifting, and filtering strategies. **all** signifies the integration of all four strategies, resulting in 100% recall rate for all picks, whose picking accuracy is depicted in (a) and (b). The last histogram in (c) shows the false positive rate (FPR), defined as the proportions of the false P- and S-picks relative to all tested noise samples, for noise samples. During benchmarking of individual strategies or combinations thereof, other parameters remain consistent with the default settings (refer to Table S2 in Supporting Information S1).

(recall rate) from 0 to well over 95% (Figure 4c). Furthermore, we notice that the majority of false positive picks exhibit relatively low picking probabilities, typically slightly above 0.1. These false picks can be easily eliminated by either increasing the picking threshold (employing a threshold of 0.3 would effectively eliminate most false positives, resulting in a false positive rate of less than 1%) or through subsequent event location processes, such as phase association (Münchmeyer, 2024). These findings demonstrate that rescaling is a robust and effective method to increase picking sensitivity and enhance model performance while not compromising the picking precision.

### 3.2. Event Burst of Short Inter-Event Times

During seismic monitoring of intense earthquake sequences, seismic events characterized by overlapping phases or short inter-event times pose a considerable challenge for successful detection and accurate arrival time picking. To evaluate the efficacy of our proposed strategies and the **quakephase** toolkit in handling such scenarios, we apply them to a 10-min continuous data set recorded by the aforementioned COSEISMIQ network (Obermann, Sánchez-Pastor, et al., 2022). This data set features an event burst comprising tens of events occurring within a few minutes, presenting abundant overlapping phase arrivals (Figure 5 and Figure S3 in Supporting Information S1).





**Figure 5.** Quakephase prediction results obtained from 10 min of continuous data (08:32:00 to 08:42:00 on 2021-07-30) recorded by station VI.EDA within the COSEISMIQ network's extended operation time. (a) Waveform plots and original ML prediction outcomes using PhaseNet with default parameters. (b) Waveform plots and **quakephase** prediction outcomes with ensemble strategies. The upper panel showcases Z-component waveforms and ML picks, while the bottom panel exhibits corresponding ML phase probabilities alongside the picking threshold within a 1-min timeframe (08:38:22 to 08:39:22; a comprehensive plot of the entire time range can be found in Figure S3 in Supporting Information S1), highlighted in gray in subplot (c). (c) The inter-pick time (the time difference between the current pick and the previous pick) and the accumulated number of picks for both the original ML prediction and the ML prediction with rescaling in **quakephase**. The original model identifies 81 phase arrivals, while the model with rescaling detects 183 phases (refer to Figure S3 in Supporting Information S1).

In practical applications, the computational overhead associated with ensembling prediction results from all different strategies and parameter space can be considerable, potentially exceeding the computational expense of a single run of the original model by several folds, depending on the number of models employed. This compromises the computational efficiency of ML models, especially for real-time seismic monitoring. In contexts where timely processing is imperative, we suggest applying benchmark tests on a small subset of data. This will allow us to identify optimal parameter ranges and well-performed ML models beforehand, thereby streamlining the exploration of parameter, strategy, and model combinations during subsequent analyses on larger data sets or continuous streams of data. Building upon the benchmark results outlined in the preceding section which shares the same region and monitoring array, in this section, we opt to ensemble results from the rescaling range spanning from 1 to 20 (specifically: 1, 2, 3, 4, 5, 8, 10, 15, 20) and over two frequency bands (no filtering and 1–50 Hz; for detailed parameters, refer to Table S3 in Supporting Information S1). These parameter ranges have been empirically demonstrated to produce effective outcomes within the tested magnitude scale and region. We choose the original PhaseNet model (Zhu & Beroza, 2019) as the base model for phase picking, benchmarking its results with and without the rescaling approach which is demonstrated to be the most effective strategy.

Employing rescaling and a max ensemble approach, the phase probabilities generated by **quakephase** surpass those of the original model, leading to more reliable phase picks and successful detection and classification of many overlapping phases (Figures 5a and 5b). Consequently, the total number of phase picks after using rescaling

has doubled (183 picks vs. 81 picks), and the average inter-phase time decreases from approximately 7 to 3 s compared to the original model (Figure 5c). We conduct manual verification of all **quakephase** picks, confirming 58 accurate P-picks and 85 accurate S-picks displaying clear phase arrivals. The remaining 40 phase picks (some as shown in Figure 5b) have lower SNR and may occasionally appear to be false positives without clear phase arrivals. However, due to low SNR and interference from overlapping arrivals, it is difficult to verify whether they are actual false positives or real phase arrivals from small events. In addition, the energy from overlapping phase arrivals or coda waves makes it extremely challenging to detect and classify the phase arrivals from the weaker event even for manual picking (Figure 5b). Since we adopt a max ensemble approach to prioritize the detection rate, increasing the picking threshold can be used to claim more robust picks.

### 3.3. Microseismic Events Induced by Hydraulic Stimulation

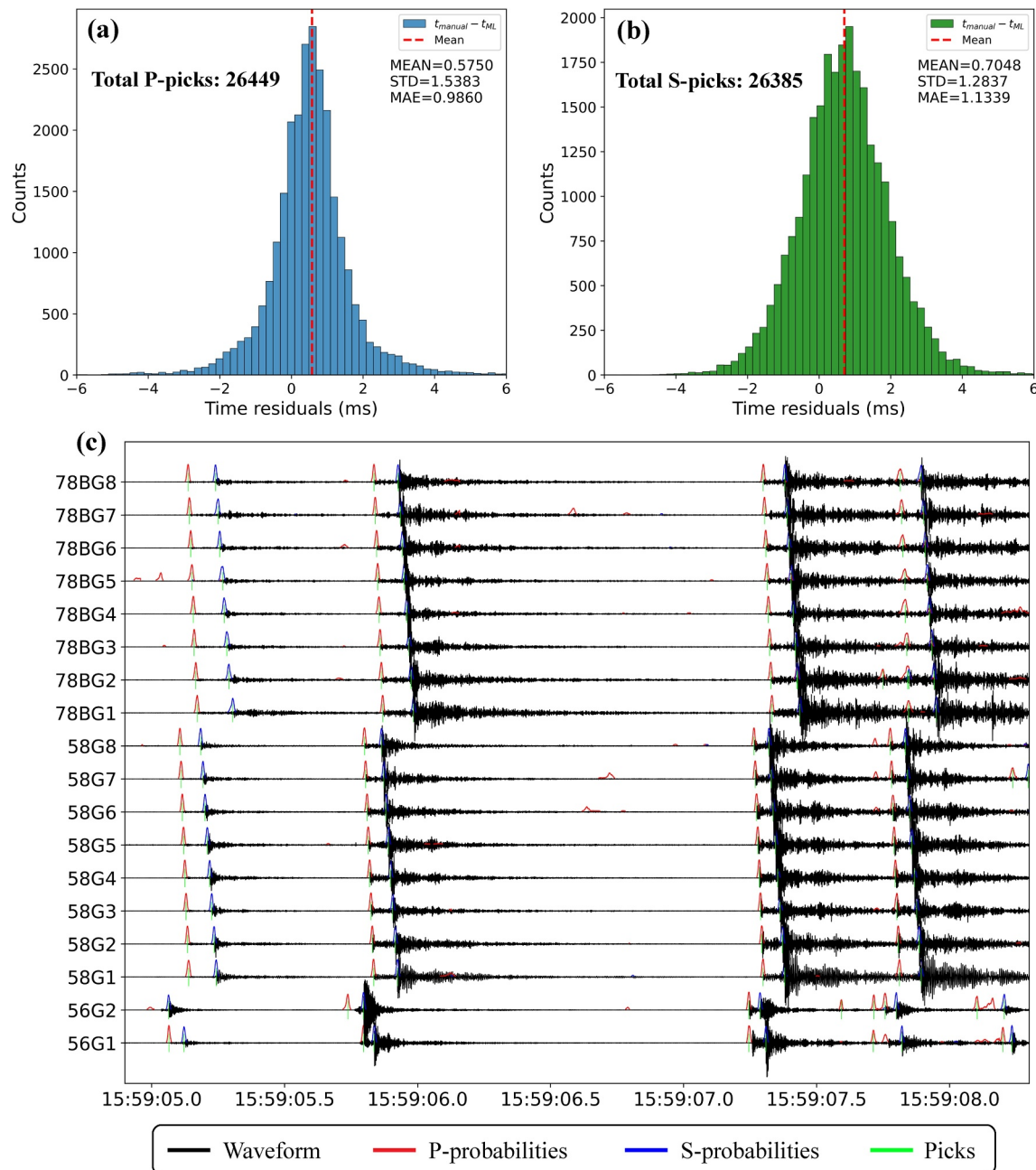
To assess the capacity of **quakephase** in enhancing model generalization on data acquired across different instruments and sampling rates, we apply it to process continuous microseismic recordings collected by three-component downhole geophones with a sampling rate of 4,000 Hz. A total of 18 geophones were deployed in three deep monitoring wells to capture microseismic events induced by hydraulic stimulation at the Utah FORGE geothermal site (Figure 6; Moore et al. (2019); Pankow et al. (2020); Rutledge et al. (2022); Niemz et al. (2024)). During 4 days of monitoring, approximately 300 GB of microseismic recordings were collected by the downhole geophones. During hydraulic stimulation, numerous microseismic events are induced with an event rate climaxing to 200 per min. To effectively handle the short inter-event intervals and overlapping arrivals of these microseismic events, we adopt a high overlapping ratio of 0.98. To enable the rapid processing of continuous microseismic recordings during monitoring, we only use EQT-stead (Woollam et al., 2022) as the base ML model and utilize a fixed rescaling rate of 60, which are tested and validated on perforation shot recordings conducted prior to stimulation. In this situation, an ensemble approach is not required during the processing, which eases the computation demand. The dominant frequency range of the recorded microseismic events is around 100–1,000 Hz (Figure S4 in Supporting Information S1). To mitigate anthropogenic noise from fluid injection, we apply bandpass filtering in the range of 100–1,800 Hz.

Using **quakephase**, we automatically detect and pick over 30,000 microseismic events from continuous recordings. Random visual verification confirms that the vast majority of the detected events (exceeding 99%) are real microseismic events. We compared our detections with a manual catalog (Dyer et al., 2022) and matched all events and picks reported in the manual catalog (around 3,000 events and 26,000 picks). A systematic comparison between ML picks and manual picks reveals exceptionally high ML picking accuracy, with the majority of ML picks within eight samples of the manual results (Figure 6). It is noteworthy that despite the input data frequency range (100–1,800 Hz) being entirely distinct from the range used to train the ML model (1–50 Hz), the rescaling approach effectively stretches input event waveforms, aligns with the scale range of the training data set, and augment the pre-trained model to generalize effectively on this OOD data set.

### 3.4. Labquakes in Rock Physics Experiments and Underground Laboratories

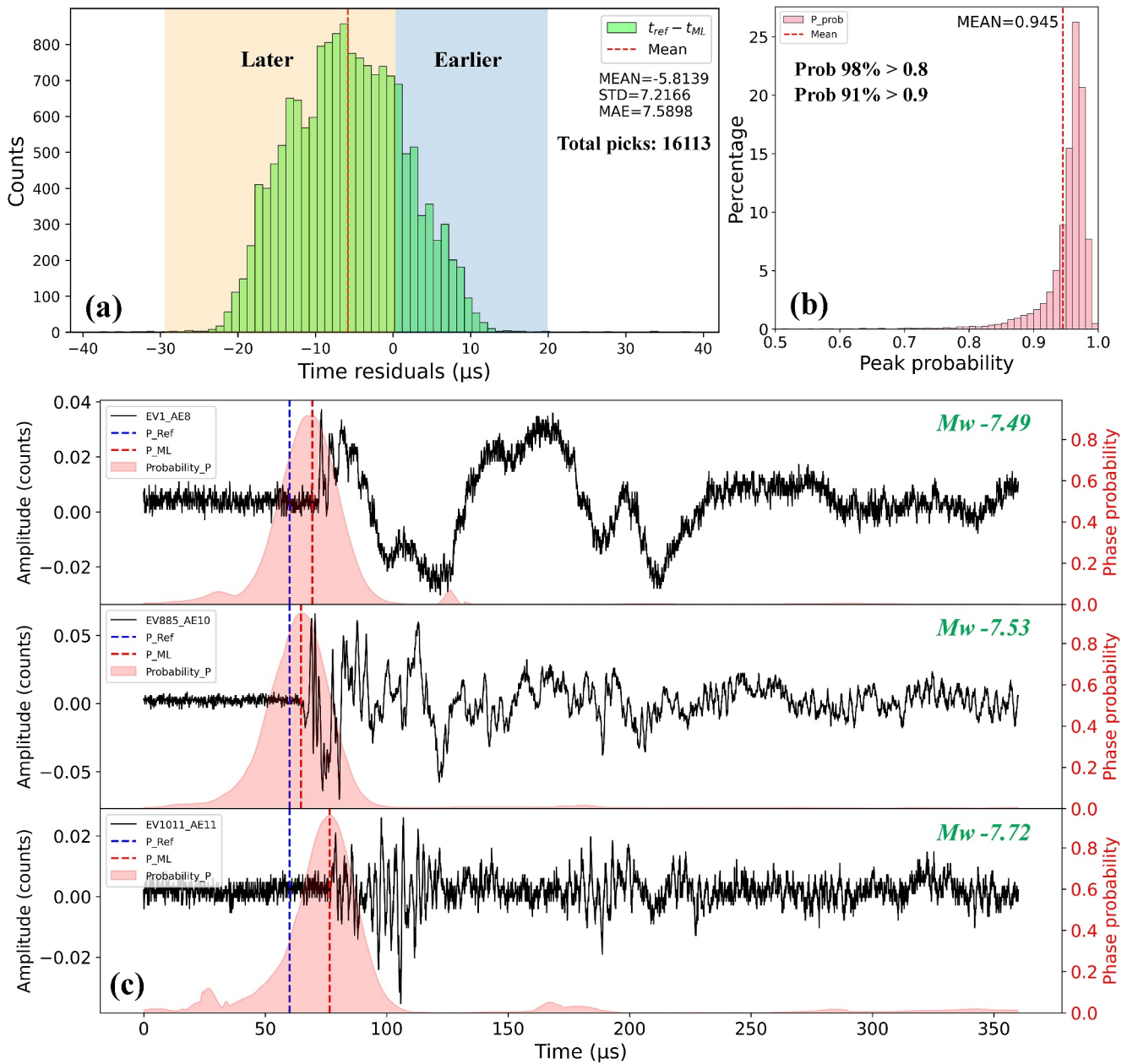
Small earthquakes ( $M < -2$ ), also occasionally referred to as acoustic emissions (AE), are generated during high-pressure laboratory rock physics experiments and hydraulic injection experiments in underground laboratories (Amann et al., 2018; Kwiatek et al., 2011; Selvadurai, 2019; Villiger et al., 2020). At present, there are not many dedicated ML models for characterizing these small-scale events or labeled ground-truth data sets to train such models compared to tectonic earthquakes (Trugman et al., 2020). Therefore, we explore whether **quakephase** with rescaling and ensemble methodologies can be directly applied for characterizing labquakes and pico-seismic events. We process waveform snippets from 1,016 labquake events collected by 16 single-component AE sensors operating under the piezoelectric principal. Each sensor has a sampling rate of 10 MHz and was acquired continuously during a triaxial rock deformation experiment performed on a sample of Berea sandstone. Magnitude ranges were determined using absolute calibration methods using a transfer plate (Selvadurai et al., 2022) and were found to range from  $-7.8$  to  $-7.0$ . Our labquake data set comprises a total of 16,113 reference P-phase picks automatically extracted using an AIC picker (Kurz et al., 2005).

To characterize labquakes utilizing pre-trained models originally trained on tectonic earthquakes, we test the ensemble parameters of **quakephase** (e.g., ML models, rescaling rates, and frequency ranges) on a few labquake samples. Finally, we employ four ML models, three rescaling rates, and four frequency ranges (Table S3 in



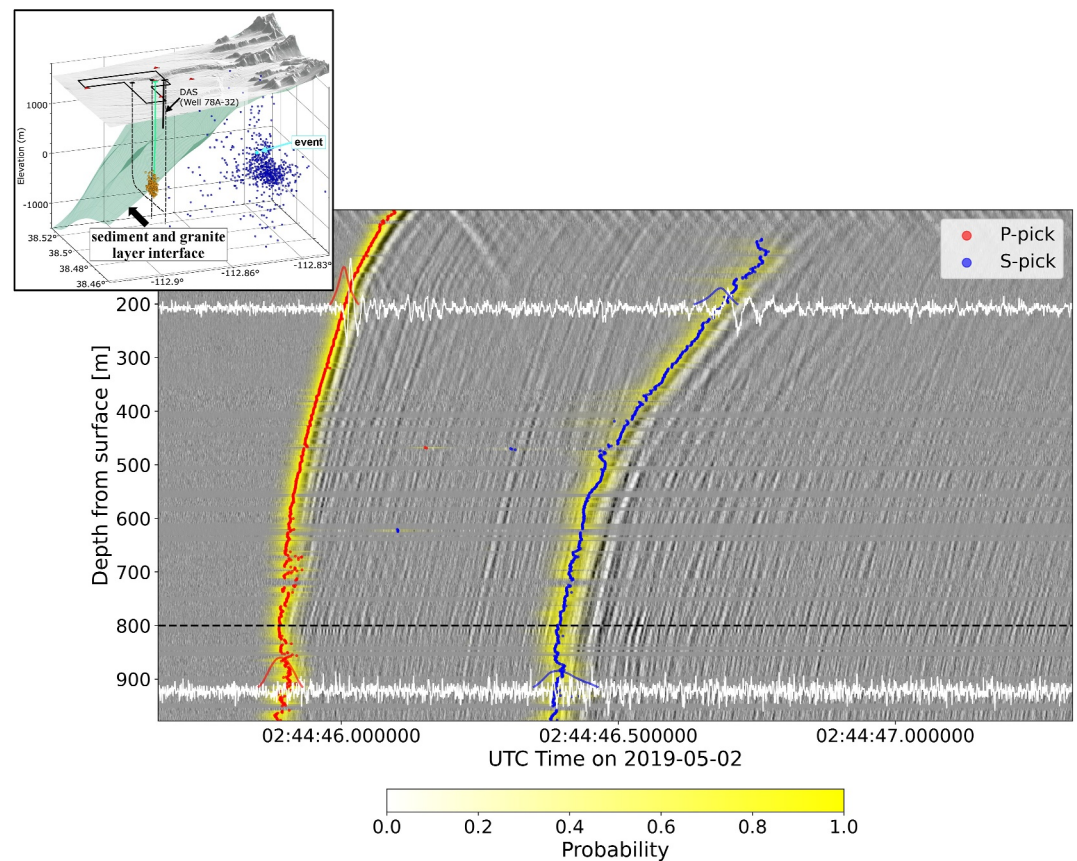
**Figure 6.** Performance evaluation of **Quakephase** on the Utah FORGE microseismic data set. (a) Histogram illustrating the distribution of P-phase picking residuals relative to manual picks, with a total of 26,449 matched P-phase picks between ML picks and manual picks. (b) Histogram depicting the distribution of S-phase picking residuals relative to manual picks, with a total of 26,385 matched S-phase picks identified between ML picks and manual picks. (c) An example record section spanning approximately 3.5 s, displaying horizontal component waveforms alongside predicted P- and S-phase probabilities from all 18 geophones.

Supporting Information S1) and ensemble their prediction results via **quakephase**. Comparative analysis against the reference picks, coupled with visual verification, underscores the superior accuracy of ML picks over reference picks (Figure 7). The majority of ML picks are within 20 microseconds relative to the automatic reference picks. Despite **quakephase** tends to pick slightly later than the reference picks, manual verification validates the superior accuracy and consistency of ML picks, aligning more closely with exact phase arrivals across most scenarios (Figure 7). Furthermore, more than 90% of ML phase picks exhibit probabilities exceeding 0.9, indicating the high performance and precision of the ML predictions (Figure 7b).



**Figure 7.** Analysis of Labquake results. (a) Histogram presenting the distribution of P-phase picking residuals relative to automatic reference picks, with a total of 16,113 picks matched and compared. The majority of ML picking times slightly lag behind the reference picks, but turn out to be more accurate than the automated reference picks. (b) Distribution of P-phase probabilities for the ML picks. (c) Waveforms and ML picking results for three example labquake events with magnitudes ranging from  $-7.72$  to  $-7.49$ . Phase probabilities are overlaid on the waveform plots in pink.

In addition, we extend our analysis to the AE data set recorded at 200 KHz sampling rate acquired from hectometer-scale fluid injection experiments conducted in the Bedretto Underground Laboratory for Geosciences and Geoenergies (BULGG) (Figures S5 and S6 in Supporting Information S1; Bröker et al., 2024; Obermann et al., 2024; Plenkers et al., 2023). ML picks are predominantly identified within 10 samples and 50 sample points relative to the manual P- and S-picks respectively, and exhibit an average phase probability of 0.9 and 0.6 for P- and S-picks (Figures S5 and S6 in Supporting Information S1). The results further affirm that ML models pre-trained on tectonic events are capable of characterizing these tiny events (magnitudes range from  $-4.6$  to  $-2.1$ ; SNR ranges from 0.09 to 7,344, SNR calculated by taking the maximum absolute amplitude ratio between the defined signal window and the noise window) with the aid of the proposed approaches especially rescaling



**Figure 8.** A seismic profile displaying DAS recordings within a vertical borehole, alongside corresponding ML prediction results at each trace. The seismic event captured has a magnitude of 0. Red dots denote P picks, while blue dots represent S picks. P- and S-phase probabilities are overlaid on the plots in yellow with the depth of the color representing the probability values. Additionally, two traces are highlighted along with their corresponding phase probabilities (white for waveforms, pink for P-phase probabilities, and light blue for S-phase probabilities). The dashed line denotes the interface separating the upper sediment layer and the underlying granite layer. Notably, a few picking outliers at depths around 470 and 620 m originate from large amplitude anomalies observed in certain traces. The inserted subfigure positioned at the top-left corner depicts the borehole DAS monitoring setup, along with the spatial distribution of the sediments and granite interface which leads to converted phases and lower SNRs in the DAS recordings.

(Figures S5 and S6 in Supporting Information S1). Notably, it is intriguing to observe that S-phases can be reliably identified and classified using solely one-component AE recordings despite the phase probabilities of S-picks being generally lower than the P-picks (Figures S5 and S6 in Supporting Information S1).

### 3.5. Data Set Recorded by Distributed Acoustic Sensing and Reflection Seismic Survey

In recent years, there has been a growing interest within the seismology community in leveraging DAS observations to study earthquake processes and Earth's interior structure (Lellouch et al., 2021; Li et al., 2023; Lindsey et al., 2019; Obermann, Sánchez-Pastor, et al., 2022). Examples of recent applications of ML on DAS data include signal recovery and data denoising (Chen, 2024; van den Ende et al., 2021), as well as an adaptation of PhaseNet for earthquake detection and picking (Zhu, Biondi, et al., 2023). We analyze a microseismic recording collected from DAS deployed in a vertical borehole near the hydraulic stimulation zone at the Utah FORGE geothermal site (Figure 8, Pankow (2022), processed according to the workflow of Tuinstra et al. (2024)). The DAS monitoring system has a gauge length of 10 m, a channel spacing of 1.02 m and samples the wavefield at 2,000 Hz. Using PhaseNet as the base model, we assemble its predictions from five rescaling rates and three frequency ranges (Table S3 in Supporting Information S1, Figure 8). The final ensemble results demonstrate the ability of **quakephase** and pre-trained models to reliably pick P and S arrivals for traces with relatively high SNR (Figure 8). Remarkably, S arrivals can be detected simply from the uni-axial DAS recordings, albeit with slightly lower phase

probabilities compared to P arrivals. A slight delay in P-picks is observable due to interference from converted P-to-S phases at the layer interface around the depth of 800 m (see the granite contact in the inset of Figure 8). Additionally, we perform a similar benchmark test employing **quakephase** to pick the first arrivals of a 3D reflection seismic survey conducted at the Utah FORGE site and obtain satisfactory results (Figure S7 in Supporting Information S1, Miller (2018)). The examples from DAS and 3D seismic reflection surveys illustrate the effectiveness of the proposed approaches in augmenting existing models for different types of sensors, such as single-component geophones and distributed acoustic sensors, across various sampling rates (1,000–4,000 Hz). However, for these densely spatially sampled recordings, models that effectively utilize spatial coherency would be more suitable and efficient when properly trained (S. Yuan et al., 2018; Zhu, Biondi, et al., 2023).

### 3.6. 2016 M7.0 Kumamoto Earthquake

Large tectonic earthquakes recorded by strong motion sensors and teleseismic events typically have long coda or longer P-to-S times and are thus difficult to fit into a single ML prediction window (usually 30–60 s long, Table S1 in Supporting Information S1). In addition, the scarcity of large earthquakes in the training data set, coupled with their distinct waveform characteristics, may make pre-trained ML models unsuitable for characterizing such events, especially for simultaneously capturing and identifying both P- and S-phases from the same event. Here, we test whether **quakephase** can be used to pick dominant phase arrivals of the 2016 M7.0 Kumamoto earthquake, utilizing data collected by the strong motion sensors of the K-NET and KiK-net networks in a triggering mode at 100 Hz (Figure S8 in Supporting Information S1, National Research Institute for Earth Science and Disaster Resilience (2019)). Our tests with a rescaling rate of 0.04–1 and ensembling demonstrate that the P and S arrivals of large tectonic earthquakes can be reliably identified and picked using the original PhaseNet model (Figure S8, Table S3 in Supporting Information S1). We further evaluate the picking performance of **quakephase** by comparing it with the PhaseNet-INSTANCE model, which is trained on a data set including events with larger epicentral distance (Woollam et al., 2022). Through rescaling, **quakephase** demonstrates superior robustness in picking both P- and S-waves, particularly in effectively capturing S-phases that are frequently overlooked by the PhaseNet-INSTANCE model (Figure S9 in Supporting Information S1). A similar test on the 2011 M9.1 Tohoku earthquake further confirms the effectiveness of rescaling for processing seismic data with long wave trains that exceed the prediction window length (Figure S10 in Supporting Information S1). However, caution is advised when analyzing the picking results due to the complex phase arrivals associated with large events at significant distances (Figure S10 in Supporting Information S1).

## 4. Discussion

Our comprehensive benchmark tests underscore that rescaling and model aggregation are pivotal elements in significantly enhancing ML model performance on OOD data sets. Model aggregation can compensate for the generalization deficiency of a specific model by aggregating results from diverse models. However, this approach implies that at least one model is applicable, and its applicability is confined to a limited scale range akin to the training data set. On the other hand, rescaling enables models to interpret data with characteristic frequency content outside their training range by implicitly using scale-invariant properties of seismic signals. This way, the rescaling approach empowers ML models pre-trained on tectonic earthquakes to adeptly characterize labquakes, induced microseismic events, and major tectonic earthquakes across a wide spectrum of source scales, sampling rates, and instruments. We posit that the variety in the training data set and the inherent self-similarity of earthquakes and earthquake records (Figure 2) are two fundamental aspects facilitating this adaptability.

Ensembling results from different filtering ranges will be necessary when recorded data contain natural or anthropogenic noise in frequency ranges that are not represented in the training data set. While our benchmark tests suggest that shifting alone does not enhance model performance (Figure 4), previous studies highlight the potentially significant impact of overlapping ratios (related to shifting) on model performance (Park et al., 2023). In practice, to achieve a balance between efficiency and performance, we recommend adopting an overlapping ratio larger than 0.5 to ensure each data sample undergoes processing by the model at least twice.

For seismic network operators, achieving robustness and minimizing false positive rates during earthquake monitoring are paramount to avoiding false alarms while vigilantly monitoring infrequent large events. Limited by the training data sets which are predominantly rich in small seismic events (magnitudes from 0 to 5) and lack enough representatives from large events, ML models sometimes present unstable performance in detecting and

picking large events. However, for operational seismic monitoring, these large but rare events are critical and should be detected with a high confidence level. The false positive rate of ML detection is intricately linked to the chosen threshold, and a too-low threshold may trigger numerous false detections, increasing the burden of manual inspection. In this situation, a higher picking threshold is favored to improve the detection precision. On the other hand, if we aim to obtain a high-resolution catalog containing abundant low-magnitude events, such as for imaging detailed fault structures, a lower threshold is preferred to increase the recall rate and include as many events as possible. In this context, false positive events can be eliminated during phase association or source back-projection location, where multiple coherent picks across the monitoring network are required (Münchmeyer, 2024; Shi et al., 2022). In practice, if benchmark data sets are available, an optimal threshold that balances precision and recall can be determined by systematically evaluating metrics such as the F1 score (Münchmeyer et al., 2022). Our proposed model augmentation approaches and ensemble methods offer flexible solutions for more reliable seismic event detection and phase arrival picking, particularly in OOD applications. For operational earthquake monitoring, where robustness and reliability are paramount, we suggest applying the rescaling approach to accommodate potential OOD events, aggregating several well-performing models, and assembling their predictions using the PCA ensemble method.

The final ensembled prediction results of **quakephase** depend on the selection of various hyper-parameters of the proposed approaches, including rescaling rate, aggregated ML model, filtering range, and overlapping ratio. Rescaling and model aggregation can significantly impact the detection rate (Figure 4). Ensembling more rescaling rates and models can increase the recall rate, but may also introduce additional false positives that need to be eliminated during subsequent processing. As with other ML model parameters, such as the picking threshold, optimal hyper-parameters are typically obtained through benchmark tests, for example, using F1 score, on a sub-dataset with exhaustive ground-truth labels without missing picks (Münchmeyer et al., 2022). However, prior knowledge can aid in parameter determination.

For rescaling, a baseline rescaling rate can be derived from the original data sampling rate and the sampling rate of the trained data set (a ratio between the two), with which untreated raw seismic recordings will be used as model input for prediction. Generally, rescaling rates spanning within one order of magnitude relative to the baseline rate can be explored to enhance model performance, assuming the original sampling rate is adequate and can capture dominant source information. For instance, data recorded at 8,000 Hz would have a baseline rescaling rate of 80 for ML models pre-trained on 100 Hz data sets. Consequently, rescaling rates between 8 and 800 can be examined for model augmentation, with a preference for within four times the baseline rate to avoid excessive computational load (i.e., rescaling rates from 20 to 320). For large or remote events whose wave train cannot fit into a single prediction window or contain significant low frequencies, exploring lower rescaling rates (down-scaling) is more appropriate. Additionally, for events oversampled in the time domain at excessively high sampling rates, such as the labquakes in Section 3.4, exploring lower rescaling rates can improve model prediction performance (Figure 7). For small events with wave trains that are relatively short compared to the ML prediction window, such as the small induced events in Sections 3.1 and 3.2, adopting higher rescaling rates than the baseline rate (up-scaling) is necessary to ensure better prediction performance. For large regional or remote events, rescaling rates can extend lower by two or three orders of magnitude, as shown in Section 3.6 Figure S8 in Supporting Information S1 (down to a rescaling rate of 0.04) and Movie S2 (down to a rescaling rate of 0.006).

For model aggregation, we recommend integrating a few high-performing models that will not incur excessive false positives. Pre-screening all available models on a few event segments can help identify the most reliable and effective models. Regarding filtering, it is favored to use frequency ranges with high SNR, especially when prior information is available to assess the noise and source spectra content. For overlapping ratio, an overlapping ratio of 0.5, which ensures each sample is processed by two prediction windows, is generally sufficient. However, in scenarios where phase arrivals overlap or exhibit short inter-phase times, increasing the overlapping ratio may enhance model prediction performance (Sections 3.2 and 3.3).

The ensemble of different augmentation approaches will increase the computational load compared to the direct application of pre-trained models. The implementation of various approaches (model aggregation, rescaling, and filtering) are independent of each other and can thus be embarrassingly parallelized. Currently, our focus is on developing and implementing these approaches to enhance the performance of pre-trained models. Parallelization is simply achieved over different model prediction windows (related to the overlapping ratio) and is implemented internally within SeisBench, while the combination of other strategies is executed sequentially. Consequently, for

the current version of **quakephase**, the running time will theoretically increase linearly with the ensemble parameters used (ML models, frequency ranges, and rescaling rates). We provide theoretical estimates of the run time:  $T = N_{ml} \times N_{freq} \times \sum_{i=1}^{N_{rs}} \left( \frac{S_m \times R_i}{S_d} \right) \times T_0$ , where  $T$  is the run time with ensemble strategies,  $T_0$  is the run time of directly applying the original model with the same overlap ratio,  $N_{ml}$  is the number of aggregated ML models,  $N_{freq}$  is the number of adopted frequency ranges,  $N_{rs}$  is the number of rescaling rates used,  $S_m$  is the frequency sampling rate of the pre-trained model (usually 100 Hz),  $R_i$  is the  $i$ -th rescaling rate, and  $S_d$  is the frequency sampling rate of the raw data. A benchmark test of run time was performed on the 10-min continuous data of the COSEISMIQ network, as presented in Section 3.2. A single run of the original PhaseNet model with a 0.99 overlapping ratio on the 10-min, 100 Hz three-component data takes 2.1 s. Running **quakephase**, which ensembles two models (PhaseNet-original and PhaseNet-stead), two rescaling rates (1 and 2), and two frequency ranges (raw data and 1–50 Hz), consumes 23.2 s. This is very close to the theoretical run times according to the previous equation (12 times the duration of a single run).

For model aggregation, we currently leverage the SeisBench framework, which offers convenient and standardized access to a range of open-source seismic picking models (Woollam et al., 2022). To this end, the current available models include PhaseNet (Zhu & Beroza, 2019), EQTransformer (Mousavi et al., 2020), and GPD (Ross et al., 2018), among others (Woollam et al., 2022). Additionally, these models have been retrained on different data sets—such as DiTing (Zhao et al., 2023), ETHZ (Swiss Seismological Service (SED) at ETH Zurich, 1983), INSTANCE (Michellini et al., 2021), Iquique (Woollam et al., 2019), LENDB (Magrini et al., 2020), NEIC (Yeck et al., 2021), OBS (Bornstein et al., 2024; Niksejel & Zhang, 2024), STEAD (Mousavi et al., 2019), and volpick (Zhong & Tan, 2024)—enabling the direct application of a diverse set of seismic picking models within the **quakephase** environment. We also notice several seismic picking models that are not yet accessible via SeisBench, including PickNet (Wang et al., 2019), PpkNet (Zhou et al., 2019), EQCCT (Saad et al., 2023), and PhaseNO (Sun et al., 2023). In the future, we plan to integrate more open-access models into **quakephase**/SeisBench, thereby enriching the diversity of models available for seismic phase picking.

## 5. Conclusions

ML-based methodologies for earthquake detection and phase picking are becoming popular, yet their application is limited by their generalization ability on OOD data sets lacking labeled data for re-training or transfer-learning. To address this issue, in this paper we introduced four strategies: rescaling, model aggregation, filtering, and shifting, aimed at augmenting the performance of pre-trained models on OOD data sets. The rescaling technique, a cornerstone of this approach, can substantially enhance model performance through the magnification and compression of the original waveforms, serving as a pivotal factor for extending model applicability to data recorded by diverse instruments and sampling rates. Applying these strategies alone or in combination will produce corresponding prediction results of the input data. To synthesize these predictions into a unified outcome, we further develop various approaches, each designed to feature either detection sensitivity or robustness. We develop a Python-based open-source package **quakephase** (<https://github.com/speedshi/quakephase>) that implements these strategies and ensemble methods. **Quakephase** provides users with flexibility in selecting the combination of model augmentation strategies and ensemble methods tailored to their specific requirements, whether maximizing detection rate or enhancing detection robustness. Benchmark tests on diverse OOD data sets, spanning different instruments (from DAS, AE sensors to broadband seismometers), sampling rates (ranging from 10 MHz to 100 Hz), and source scales (magnitude spanning from  $-8$  to  $7$ ), affirm the effectiveness of the proposed methodologies. The proposed methods together with the developed tool empower the comprehensive characterization of earthquakes across various scales, fostering the broader utilization of ML techniques in a wide range of earthquake monitoring contexts.

## Data Availability Statement

The seismic data of the COSEISMIQ network evaluated in Sections 3.1 and 3.2 are available from Swiss Seismological Service (SED) at ETH Zurich (2018) and can be downloaded through the European Integrated Data Archive (EIDA) Web Services (<https://www.orfeus-eu.org/data/eida>) with network code 2C, OR, and VI. A description of the COSEISMIQ data set can be found in Grigoli et al. (2022) and on the COSEISMIQ project website (<http://www.coseismlq.ethz.ch/en/dissemination/stations>). The Utah FORGE microseismic data



recorded by downhole geophones and DAS data, which are evaluated in Sections 3.3 and 3.5, are available at the Utah FORGE Data Distribution website (<https://constantine.seis.utah.edu/datasets.html>) and Pankow (2022). The 3D reflection seismic survey assessed in Section 3.5 can be accessed via Miller (2018). The 2016 M7.0 Kumamoto strong motion data used in Section 3.6 are available from National Research Institute for Earth Science and Disaster Resilience (2019) and the relevant website (<https://www.kyoshin.bosai.go.jp>) with network code K-NET and KiK-net.

**Acknowledgments**

The authors would like to thank the editors and the three anonymous reviewers for their constructive comments, which helped improve the quality of the paper. This work was supported by the De-Risking Enhanced Geothermal Energy project (Innovation for DEEPs <http://deepgeothermal.org>). DEEP is subsidized through the Cofund GEOTHERMICA (Project No. 200320-4001), which is supported by the European Union's HORIZON 2020 programme and various National Funding Agencies for research, technological development, and demonstration under Grant 731117. The contribution of Men-Andrin Meier was partially supported by Grant C22-10 (HighFEM) of the Swiss Data Science Center (SDSC), Ecole Polytechnique Fédérale de Lausanne and ETH Zürich. The contribution of Anne Obermann was partially supported by the Geothermica project Derisking exploration for geothermal plays in magmatic environments (DEEPEN; Grant 03EE4018).

**References**

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: Computational Statistics*, 2(4), 433–459. <https://doi.org/10.1002/wics.101>

Amann, F., Gischig, V., Evans, K., Doetsch, J., Jalali, R., Valley, B., et al. (2018). The SEISMO-hydronechanical behavior during deep geothermal reservoir stimulations: Open questions tackled in a decameter-scale in situ stimulation experiment. *Solid Earth*, 9(1), 115–137. <https://doi.org/10.5194/se-9-115-2018>

Arjovsky, M. (2021). Out of distribution generalization in machine learning. <https://doi.org/10.48550/arXiv.2103.02667>

Bornstein, T., Lange, D., Münchmeyer, J., Woollam, J., Rietbrock, A., Barcheck, G., et al. (2024). PickBlue: Seismic phase picking for ocean bottom seismometers with deep learning. *Earth and Space Science*, 11(1), e2023EA003332. <https://doi.org/10.1029/2023ea003332>

Bröker, K., Ma, X., Gholizadeh Doonechaly, N., Roskopf, M., Obermann, A., Rinaldi, A. P., et al. (2024). Hydromechanical characterization of a fractured crystalline rock volume during multi-stage hydraulic stimulations at the BedrettoLab. *Geothermics*, 124, 103126. <https://doi.org/10.1016/j.geothermics.2024.103126>

Chai, C., Maceira, M., Santos-Villalobos, H. J., Venkatakrishnan, S. V., Schoenball, M., Zhu, W., et al. (2020). Using a deep neural network and transfer learning to bridge scales for seismic phase picking. *Geophysical Research Letters*, 47(16), e2020GL088651. <https://doi.org/10.1029/2020gl088651>

Chen, Y. (2024). SigRecover: Recovering signal from noise in distributed acoustic sensing data processing. *Seismological Research Letters*, 95(3), 1976–1985. <https://doi.org/10.1785/0220230370>

Chen, Y., Savvaids, A., Saad, O. M., Dino Huang, G.-C., Siervo, D., O’Sullivan, V., et al. (2024). TXED: The Texas earthquake dataset for AI. *Seismological Research Letters*, 95(3), 2013–2022. <https://doi.org/10.1785/0220230327>

Dyer, B., Karvounis, D., & Bethmann, F. (2022). Utah FORGE: Updated seismic event catalogue from the April, 2022 stimulation of well 16A (78)-32 [Dataset]. <https://doi.org/10.15121/1908927>

Grigoli, F., Clinton, J. F., Diehl, T., Kaestli, P., Scarabello, L., Agustsdottir, T., et al. (2022). Monitoring microseismicity of the Hengill geothermal field in Iceland. *Scientific Data*, 9(1), 220. <https://doi.org/10.1038/s41597-022-01339-w>

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Ide, S. (2019). Frequent observations of identical onsets of large and small earthquakes. *Nature*, 573(7772), 112–116. <https://doi.org/10.1038/s41586-019-1508-5>

Kurz, J. H., Grosse, C. U., & Reinhardt, H.-W. (2005). Strategies for reliable automatic onset time picking of acoustic emissions and of ultrasound signals in concrete. *Ultrasonics*, 43(7), 538–546. <https://doi.org/10.1016/j.ultras.2004.12.005>

Kwiatk, G., Plenkers, K., Dresen, G., & Group, J. R. (2011). Source parameters of picoseismicity recorded at mponeng deep gold mine, South Africa: Implications for scaling relations. *Bulletin of the Seismological Society of America*, 101(6), 2592–2608. <https://doi.org/10.1785/0120110094>

Lapins, S., Goitom, B., Kendall, J.-M., Werner, M. J., Cashman, K. V., & Hammond, J. O. (2021). A little data goes a long way: Automating seismic phase arrival picking at Nabro volcano with transfer learning. *Journal of Geophysical Research: Solid Earth*, 126(7), e2021JB021910. <https://doi.org/10.1029/2021jb021910>

Lellouch, A., Schultz, R., Lindsey, N. J., Biondi, B., & Ellsworth, W. L. (2021). Low-magnitude seismicity with a downhole distributed acoustic sensing array—Examples from the FORGE geothermal experiment. *Journal of Geophysical Research: Solid Earth*, 126(1), e2020JB020462. <https://doi.org/10.1029/2020jb020462>

Li, J., Kim, T., Lapusta, N., Biondi, E., & Zhan, Z. (2023). The break of earthquake asperities imaged by distributed acoustic sensing. *Nature*, 620(7975), 800–806. <https://doi.org/10.1038/s41586-023-06227-w>

Lindsey, N. J., Dawe, T. C., & Ajo-Franklin, J. B. (2019). Illuminating seafloor faults and ocean dynamics with dark fiber distributed acoustic sensing. *Science*, 366(6469), 1103–1107. <https://doi.org/10.1126/science.aay5881>

Magrini, F., Jozinović, D., Cammarano, F., Michelini, A., & Boschi, L. (2020). Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale. *Artificial Intelligence in Geosciences*, 1, 1–10. <https://doi.org/10.1016/j.aig.2020.04.001>

Manthei, G., & Plenkers, K. (2018). Review on in situ acoustic emission monitoring in the context of structural health monitoring in mines. *Applied Sciences*, 8(9), 1595. <https://doi.org/10.3390/app8091595>

Meier, M.-A., Ampuero, J., & Heaton, T. H. (2017). The hidden simplicity of subduction megathrust earthquakes. *Science*, 357(6357), 1277–1281. <https://doi.org/10.1126/science.aan5643>

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., & Lauciani, V. (2021). INSTANCE—the Italian seismic dataset for machine learning. *Earth System Science Data*, 13(12), 5509–5544. <https://doi.org/10.5194/essd-13-5509-2021>

Miller, J. (2018). Utah FORGE: 2D and 3D seismic data [Dataset]. *Energy and Geoscience Institute at the University of Utah*. <https://doi.org/10.15121/1452746>

Moore, J., McLennan, J., Allis, R., Pankow, K., Simmons, S., Podgorney, R., et al. (2019). The Utah frontier observatory for research in geothermal energy (FORGE): An international laboratory for enhanced geothermal system technology development. In *44th Workshop on Geothermal Reservoir Engineering* (pp. 11–13).

Mousavi, S. M., & Beroza, G. C. (2023). Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences*, 51(1), 105–129. <https://doi.org/10.1146/annurev-earth-071822-100323>

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1), 3952. <https://doi.org/10.1038/s41467-020-17591-w>

Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). STANford Earthquake Dataset (STEAD): A global data set of seismic signals for AI. *IEEE Access*, 7, 179464–179476. <https://doi.org/10.1109/access.2019.2947848>

- Münchmeyer, J. (2024). PyOcto: A high-throughput seismic phase associator. *Seismica*, 3(1). <https://doi.org/10.26443/seismica.v3i1.1130>
- Münchmeyer, J., Woollam, J., Rietbrock, A., Tilmann, F., Lange, D., Bornstein, T., et al. (2022). Which picker fits my data? A quantitative evaluation of deep learning based seismic pickers. *Journal of Geophysical Research: Solid Earth*, 127(1), e2021JB023499. <https://doi.org/10.1029/2021jb023499>
- National Research Institute for Earth Science and Disaster Resilience. (2019). NIED K-NET, KiK-net [Dataset]. *National Research Institute for Earth Science and Disaster Resilience*. <https://doi.org/10.17598/NIED.0004>
- Ni, Y., Hutko, A., Skene, F., Denolle, M., Malone, S., Bodin, P., et al. (2023). Curated Pacific Northwest AI-ready seismic dataset. *Seismica*, 2(1). <https://doi.org/10.26443/seismica.v2i1.368>
- Niemz, P., McLennan, J., Pankow, K. L., Rutledge, J., & England, K. (2024). Circulation experiments at Utah FORGE: Near-surface seismic monitoring reveals fracture growth after shut-in. *Geothermics*, 119, 102947. <https://doi.org/10.1016/j.geothermics.2024.102947>
- Niksejel, A., & Zhang, M. (2024). OBSTransformer: A deep-learning seismic phase picker for OBS data using automated labelling and transfer learning. *Geophysical Journal International*, 237(1), 485–505. <https://doi.org/10.1093/gji/ggae049>
- Nooshiri, N., Bean, C. J., Dahm, T., Grigoli, F., Kristjándóttir, S., Obermann, A., & Wiemer, S. (2022). A multibranch, multitarget neural network for rapid point-source inversion in a microseismic environment: Examples from the Hengill Geothermal Field, Iceland. *Geophysical Journal International*, 229(2), 999–1016. <https://doi.org/10.1093/gji/ggab511>
- Obermann, A., Rosskopf, M., Durand, V., Plenkers, K., Bröker, K., Gholizadeh Doonechaly, N., et al. (2024). Seismic response of hectometer-scale fracture systems to hydraulic stimulation in the Bedretto Underground laboratory, Switzerland. *Journal of Geophysical Research: Solid Earth*. <https://doi.org/10.1029/2024JB029836>
- Obermann, A., Sánchez-Pastor, P., Wu, S.-M., Wollin, C., Baird, A. F., Isken, M. P., et al. (2022). Combined large-N seismic arrays and DAS fiber optic cables across the Hengill geothermal field, Iceland. *Seismological Research Letters*, 93(5), 2498–2514. <https://doi.org/10.1785/0220220073>
- Obermann, A., Wu, S.-M., Ágústadóttir, T., Duran, A., Diehl, T., Sánchez-Pastor, P., et al. (2022). Seismicity and 3-D body-wave velocity models across the Hengill geothermal area, SW Iceland. *Frontiers in Earth Science*, 10, 969836. <https://doi.org/10.3389/feart.2022.969836>
- Pankow, K. (2022). Utah FORGE DAS seismic data 2022 [Dataset]. Retrieved from <https://gdr.openei.org/submissions/1393>
- Pankow, K., Mesimeri, M., McLennan, J., Wannamaker, P., & Moore, J. (2020). Seismic monitoring at the Utah Frontier observatory for research in geothermal energy. In *Proceedings of the 45th Workshop on Geothermal Reservoir Engineering, Stanford, CA, USA* (pp. 10–12).
- Park, Y., Beroza, G. C., & Ellsworth, W. L. (2023). A mitigation strategy for the prediction inconsistency of neural phase pickers. *Seismological Society of America*, 94(3), 1603–1612. <https://doi.org/10.1785/0220230003>
- Plenkers, K., Reinicke, A., Obermann, A., Gholizadeh Doonechaly, N., Krietsch, H., Fechner, T., et al. (2023). Multi-disciplinary monitoring networks for mesoscale underground experiments: Advances in the bedretto reservoir project. *Sensors*, 23(6), 3315. <https://doi.org/10.3390/s23063315>
- Prieto, G. A., Shearer, P. M., Vernon, F. L., & Kilb, D. (2004). Earthquake source scaling and self-similarity estimation from stacking P and S spectra. *Journal of Geophysical Research*, 109(B8). <https://doi.org/10.1029/2004jb003084>
- Ross, Z. E., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, 108(5A), 2894–2901. <https://doi.org/10.1785/0120180080>
- Rutledge, J., Dyer, B., Bethmann, F., Meier, P., Pankow, K., Wannamaker, P., & Moore, J. (2022). Downhole microseismic monitoring of injection stimulations at the Utah FORGE EGS Site. In *ARMA US Rock Mechanics/Geomechanics Symposium* (pp. ARMA–2022).
- Saad, O. M., Chen, Y., Siervo, D., Zhang, F., Savvaidis, A., Huang, G.-C. D., et al. (2023). EQCCT: A production-ready Earthquake detection and phase picking method using the compact convolutional transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–15. <https://doi.org/10.1109/tgrs.2023.3319440>
- Saad, O. M., Hafez, A. G., & Soliman, M. S. (2020). Deep learning approach for earthquake parameters classification in earthquake early warning system. *IEEE Geoscience and Remote Sensing Letters*, 18(7), 1293–1297. <https://doi.org/10.1109/lgrs.2020.2998580>
- Saad, O. M., Huang, G., Chen, Y., Savvaidis, A., Fomel, S., Pham, N., & Chen, Y. (2021). Scalodeep: A highly generalized deep learning framework for real-time earthquake detection. *Journal of Geophysical Research: Solid Earth*, 126(4), e2020JB021473. <https://doi.org/10.1029/2020jb021473>
- Selvadurai, P. A. (2019). Laboratory insight into seismic estimates of energy partitioning during dynamic rupture: An observable scaling breakdown. *Journal of Geophysical Research: Solid Earth*, 124(11), 11350–11379. <https://doi.org/10.1029/2018jb017194>
- Selvadurai, P. A., Wu, R., Bianchi, P., Niu, Z., Michail, S., Madonna, C., & Wiemer, S. (2022). A methodology for reconstructing source properties of a conical piezoelectric actuator using array-based methods. *Journal of Nondestructive Evaluation*, 41(1), 23. <https://doi.org/10.1007/s10921-022-00853-6>
- Shi, P., Angus, D., Rost, S., Nowacki, A., & Yuan, S. (2019). Automated seismic waveform location using multichannel coherency migration (MCM)—I: Theory. *Geophysical Journal International*, 216(3), 1842–1866. <https://doi.org/10.1093/gji/ggy132>
- Shi, P., Grigoli, F., Lanza, F., Beroza, G. C., Scarabello, L., & Wiemer, S. (2022). MALMI: An automated earthquake detection and location workflow based on machine learning and waveform migration. *Seismological Society of America*, 93(5), 2467–2483. <https://doi.org/10.1785/0220220071>
- Shi, P., Nowacki, A., Rost, S., & Angus, D. (2019). Automated seismic waveform location using multichannel coherency migration (MCM)—II: Application to induced and volcano-tectonic seismicity. *Geophysical Journal International*, 216(3), 1608–1632. <https://doi.org/10.1093/gji/ggy507>
- Shi, P., Seydoux, L., & Poli, P. (2021). Unsupervised learning of seismic wavefield features: Clustering continuous array seismic data during the 2009 L'Aquila earthquake. *Journal of Geophysical Research: Solid Earth*, 126(1), e2020JB020506. <https://doi.org/10.1029/2020jb020506>
- Spallarossa, D., Cattaneo, M., Scafidi, D., Michele, M., Chiaraluce, L., Segou, M., & Main, I. (2021). An automatically generated high-resolution earthquake catalogue for the 2016–2017 central Italy seismic sequence, including P and S phase arrival times. *Geophysical Journal International*, 225(1), 555–571. <https://doi.org/10.1093/gji/ggaa604>
- Staněk, F., Anikiev, D., Valenta, J., & Eisner, L. (2015). Semblance for microseismic event detection. *Geophysical Journal International*, 201(3), 1362–1369. <https://doi.org/10.1093/gji/ggv070>
- Sun, H., Ross, Z. E., Zhu, W., & Azzizadenesheli, K. (2023). Phase neural operator for multi-station picking of seismic arrivals. *Geophysical Research Letters*, 50(24), e2023GL106434. <https://doi.org/10.1029/2023gl106434>
- Swiss Seismological Service (SED) at ETH Zurich. (1983). National seismic networks of Switzerland [Dataset]. <https://doi.org/10.12686/SED/NETWORKS/CH>
- Swiss Seismological Service (SED) at ETH Zurich. (2018). COSEISMIQ—Control SEISmicity and manage induced earthQuakes [Dataset]. <https://doi.org/10.12686/sed/networks/2c>

- Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., et al. (2021). Machine-learning-based high-resolution earthquake catalog reveals how complex fault structures were activated during the 2016–2017 central Italy sequence. *The Seismic Record*, *1*(1), 11–19. <https://doi.org/10.1785/0320210001>
- Trugman, D. T., McBrearty, I. W., Bolton, D. C., Guyer, R. A., Marone, C., & Johnson, P. A. (2020). The spatiotemporal evolution of granular microslip precursors to laboratory earthquakes. *Geophysical Research Letters*, *47*(16), e2020GL0888404. <https://doi.org/10.1029/2020gl0888404>
- Tuinstra, K., Grigoli, F., Lanza, F., Rinaldi, A. P., Fichtner, A., & Wiemer, S. (2024). Locating clustered seismicity using distance geometry solvers: Applications for sparse and single-borehole das networks. *Geophysical Journal International*, *238*(2), 661–680. <https://doi.org/10.1093/gji/ggae168>
- van den Ende, M., Lior, I., Ampuero, J.-P., Sladen, A., Ferrari, A., & Richard, C. (2021). A self-supervised deep learning approach for blind denoising and waveform coherence enhancement in distributed acoustic sensing data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(7), 3371–3384. <https://doi.org/10.1109/tnnls.2021.3132832>
- Villiger, L., Gischig, V. S., Doetsch, J., Krietsch, H., Dutler, N. O., Jalali, M., et al. (2020). Influence of reservoir geology on seismic response during decameter-scale hydraulic stimulations in crystalline rock. *Solid Earth*, *11*(2), 627–655. <https://doi.org/10.5194/se-11-627-2020>
- Wang, J., Xiao, Z., Liu, C., Zhao, D., & Yao, Z. (2019). Deep learning for picking seismic arrival times. *Journal of Geophysical Research: Solid Earth*, *124*(7), 6612–6624. <https://doi.org/10.1029/2019jb017536>
- Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., et al. (2022). SeisBench—A toolbox for machine learning in seismology. *Seismological Research Letters*, *93*(3), 1695–1709. <https://doi.org/10.1785/0220210324>
- Woollam, J., Rietbrock, A., Bueno, A., & De Angelis, S. (2019). Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network. *Seismological Research Letters*, *90*(2A), 491–502. <https://doi.org/10.1785/0220180312>
- Yeck, W. L., Patton, J. M., Ross, Z. E., Hayes, G. P., Guy, M. R., Ambruz, N. B., et al. (2021). Leveraging deep learning in global 24/7 real-time earthquake monitoring at the National Earthquake Information Center. *Seismological Research Letters*, *92*(1), 469–480. <https://doi.org/10.1785/0220200178>
- Yuan, C., Ni, Y., Lin, Y., & Denolle, M. (2023). Better together: Ensemble learning for earthquake detection and phase picking. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–17. <https://doi.org/10.1109/tgrs.2023.3320148>
- Yuan, S., Liu, J., Wang, S., Wang, T., & Shi, P. (2018). Seismic waveform classification and first-break picking using convolution neural networks. *IEEE Geoscience and Remote Sensing Letters*, *15*(2), 272–276. <https://doi.org/10.1109/lgrs.2017.2785834>
- Zhao, M., Xiao, Z., Chen, S., & Fang, L. (2023). DiTing: A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology. *Earthquake Science*, *36*(2), 84–94. <https://doi.org/10.1016/j.eqs.2022.01.022>
- Zhong, Y., & Tan, Y. J. (2024). Deep-learning-based phase picking for volcano-tectonic and long-period earthquakes. *Geophysical Research Letters*, *51*(12), e2024GL108438. <https://doi.org/10.1029/2024gl108438>
- Zhou, Y., Yue, H., Kong, Q., & Zhou, S. (2019). Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks. *Seismological Research Letters*, *90*(3), 1079–1087. <https://doi.org/10.1785/0220180319>
- Zhu, W., & Beroza, G. C. (2019). PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273.
- Zhu, W., Biondi, E., Li, J., Yin, J., Ross, Z. E., & Zhan, Z. (2023). Seismic arrival-time picking on distributed acoustic sensing data using semi-supervised learning. *Nature Communications*, *14*(1), 8192. <https://doi.org/10.1038/s41467-023-43355-3>
- Zhu, W., Hou, A. B., Yang, R., Datta, A., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2023). QuakeFlow: A scalable machine-learning-based earthquake monitoring workflow with cloud computing. *Geophysical Journal International*, *232*(1), 684–693. <https://doi.org/10.1093/gji/ggac355>