

Event-Centric Earthquake Monitoring with a Multi-Station Transformer

Haoyu Wang^{*1,2} and Weiqiang Zhu^{1,2}

¹Department of Earth & Planetary Science, University of California, Berkeley

²Berkeley Seismology Laboratory, University of California, Berkeley

Abstract

Earthquake catalogs underpin studies such as fault structure, earthquake processes and seismic hazard, and their value depends on being both complete and reliable. Machine learning has improved individual tasks of earthquake monitoring, but processing stations independently discards the consistency of arrivals across the network, and running tasks in sequence fragments the information describing each earthquake. We present EQNet2, a geometry-aware multi-station Transformer that jointly predicts phase arrivals, first-motion polarities, source location and magnitude, and the association of arrivals into events. On a Northern California test dataset and on continuous waveforms from Ridgecrest and the San Francisco Bay Area, EQNet2 recovers more weak arrivals, produces fewer false detections, and associates more picks into located events. Its learned attention links information across stations along time–distance patterns consistent with seismic moveout. As seismic networks grow denser, exploiting multi-station coherence and event-centric prediction offers a path to more complete and reliable catalogs.

1 Introduction

Earthquake monitoring converts continuous waveforms recorded by a seismic network into catalogs that report the timing, location, magnitude, and source mechanism of individual earthquakes. These catalogs support research across seismology, from earthquake source processes and fault structure to seismic hazard. For example, dense and precisely located seismicity images fault geometry at depth (Poggiali, Chiaraluce, Ross, Zhu, & Marone, 2025; Ross, Idini, et al., 2019; Y. J. Tan et al., 2021). The distribution and migration of seismicity trace fluid movement and aseismic slip (Park, Mousavi, Zhu, Ellsworth, & Beroza, 2020; Ross & Cochran, 2021; Ross, Cochran, Trugman, & Smith, 2020; Suzuki et al., 2025; Wang, Wang, Peng, & Lei, 2026). Seismicity beneath volcanoes delineates the structures activated during unrest and the pathways of magma transport (Lomax, Anagnostou, Karakostas, Hicks, & Papadimitriou, 2025; X. Tan et al., 2025; Wilding, Zhu, Ross, & Jackson, 2023). Catalogs with a lower magnitude of completeness also resolve the statistics of earthquake sequences such as early-aftershock migration, precursory seismicity, and b-value variations (Herrmann, Piegari, & Marzocchi, 2022; Kwiatak et al., 2023; Peng & Zhao, 2009; Trugman & Ross, 2019). These results depended on both the quantity and the quality of the detected earthquakes. Missed detections raise the magnitude of completeness (Schorlemmer & Woessner, 2008; Woessner & Wiemer, 2005), false detections inflate event rates and add spatial scatter, and location errors blur fault structure and may introduce location artifacts (Hauksson & Shearer, 2005; Hauksson, Yang, & Shearer, 2012; Waldhauser & Ellsworth, 2000). Producing high-resolution catalogs that are complete, reliable, and precise therefore depends on continued advances in automated monitoring workflows.

Automated monitoring workflows convert continuous waveforms into event-level information through a sequence of tasks: picking P- and S-phase arrival times at each station, determining first-motion polarities, associating picks that originate from a common source, locating hypocenters, and estimating magnitudes and focal mechanisms. Deep learning has substantially improved each of these tasks (Mousavi & Beroza,

^{*}haoyu.wang@berkeley.edu

2022). Deep-learning phase pickers trained on millions of analyst-labeled seismograms now pick arrivals with near-analyst accuracy and detect weaker phases than traditional algorithms (Y. Chen, Savvaidis, Siervo, Huang, & Saad, 2024; Mousavi, Ellsworth, Zhu, Chuang, & Beroza, 2020; Ross, Meier, Hauksson, & Heaton, 2018; Saad et al., 2023; Soto & Schurr, 2021; Zhou, Yue, Kong, & Zhou, 2019; Zhu & Beroza, 2019). Convolutional classifiers determine first-motion polarities that previously required manual inspection, enabling focal mechanisms for abundant small events (Chakraborty et al., 2022; Cheng, Hauksson, & Ben-Zion, 2023; Hara, Fukahata, & Iio, 2019; Mousavi, Zhu, Ellsworth, & Beroza, 2019; Ross, Meier, & Hauksson, 2018). Machine-learning associators efficiently group large pick sets produced by deep-learning pickers into candidate events (McBrearty & Beroza, 2023; McBrearty, Delorey, & Johnson, 2019; Ross, Yue, Meier, Hauksson, & Heaton, 2019; Zhu, McBrearty, Mousavi, Ellsworth, & Beroza, 2022). Location models provide fast preliminary hypocenters (McBrearty & Beroza, 2022; Mousavi & Beroza, 2020; Zhu, Rong, Jie, & Wei, 2025), and transformer models determine focal mechanisms from more complete seismic information (Song, Meier, Ellsworth, & Beroza, 2026). Chaining these tools into pipelines has produced a new generation of high-resolution catalogs from dense and long-duration seismic recordings (Feng, Wang, Fang, Wu, & Yang, 2024; Kwiatak et al., 2023; X. Tan et al., 2025; Y. J. Tan et al., 2021; Wilding et al., 2023; M. Zhang, Liu, Feng, Wang, & Zhu, 2022; Z. Zhang et al., 2024; Zhu et al., 2023).

Despite these advances, two limitations of current automated monitoring workflows remain. The first is that most phase pickers still process each station independently (G. Chen & Li, 2022; Sun, Ross, Zhu, & Azizzadenesheli, 2023), whereas an earthquake is recorded by the network as a whole. Weak arrivals that are ambiguous on a single noisy seismogram become identifiable once their timing aligns with the moveout of arrivals at neighboring stations, and local noise that resembles a phase arrival on one trace has no coherent arrivals elsewhere in the network. Single-station pickers cannot use this consistency across the network, so they miss weak arrivals and pass isolated false picks to association. The second limitation is that the workflow is a unidirectional sequence of independently trained modules, so the workflow as a whole is not optimized for catalog quality (Si, Wu, Li, Wang, & Zhu, 2024; Zhu, Tai, Mousavi, Bailis, & Beroza, 2022). Each stage passes only discrete predictions to the next, without the waveforms or the uncertainties behind them, so errors made early are difficult to correct later and accumulate downstream.

Recent work has approached these two limitations. Multi-station pickers such as CubeNet (G. Chen & Li, 2022), EdgePhase (Feng, Mohanna, & Meng, 2022), and PhaseNO (Sun et al., 2023) take waveforms from many stations as joint input, so that the pick at one station is informed by the waveforms at the others. Multi-task models predict several attributes of the same event from a shared representation. PhaseNet+ (Zhu, Song, Wang, & Münchmeyer, 2025) adds first-motion polarity and origin-time estimation to phase picking, and an attention-based model picks arrivals and determines polarity simultaneously (J. Zhang, Li, & Zhang, 2023). The two directions have begun to converge. EQNet (Zhu, Tai, et al., 2022) performs picking and association jointly from multi-station input, and PLAN (Si et al., 2024) adds location. These studies show that automated monitoring improves both when information is shared across stations and when related tasks are predicted jointly. Extending this combination to the entire workflow could overcome both limitations at once, recovering weak arrivals and rejecting isolated noise through network coherence while keeping every measurement tied to the same earthquake.

In this study, we combine network-wide input with event-centric prediction in a single model, EQNet2, a multi-station Transformer that learns features from multiple stations and organizes its predictions around individual earthquakes (Figure 1). EQNet2 jointly predicts phase arrivals, first-motion polarities, source location and magnitude, and the association of arrivals into events from a shared representation of multi-station waveforms and relative station geometry. A learnable spatiotemporal relative position bias incorporates station geometry into the attention mechanism, allowing the model to accommodate varying station configurations. The model is trained on event-organized waveforms from the California Earthquake Event Dataset (Zhu, Wang, et al., 2025) and evaluated on a Northern California test dataset and on continuous waveforms from the 2019 Ridgecrest sequence and the 2025 Berkeley earthquake. Across these tests, EQNet2 recovers more weak arrivals, produces fewer false detections, and retains a larger fraction of its picks in located events.

2 Methods

2.1 Event-centric monitoring across the station network

EQNet2 performs the interconnected tasks of earthquake monitoring in one model that processes the whole network at once. It takes multi-station waveforms and station coordinates as input and jointly produces phase arrivals, first-motion polarities, station-wise event-arrival proxies, source-related estimates (travel time, S-P differential time, epicentral distance, backazimuth, depth, and magnitude), and network-wide association probabilities (Figure 1a). Rather than generating these outputs in separately optimized modules, EQNet2 decodes them all from one representation shared across stations and tasks, so each output refers to the same candidate events. Following PhaseNet+ (Zhu, Song, et al., 2025), the event-arrival proxy is defined as the temporal midpoint between the P- and S-phase arrivals. The proxy gives each trace a reference time at which the source-related estimates are extracted. Each predicted proxy further serves as the prompt for association, linking the arrivals at individual stations to candidate events.

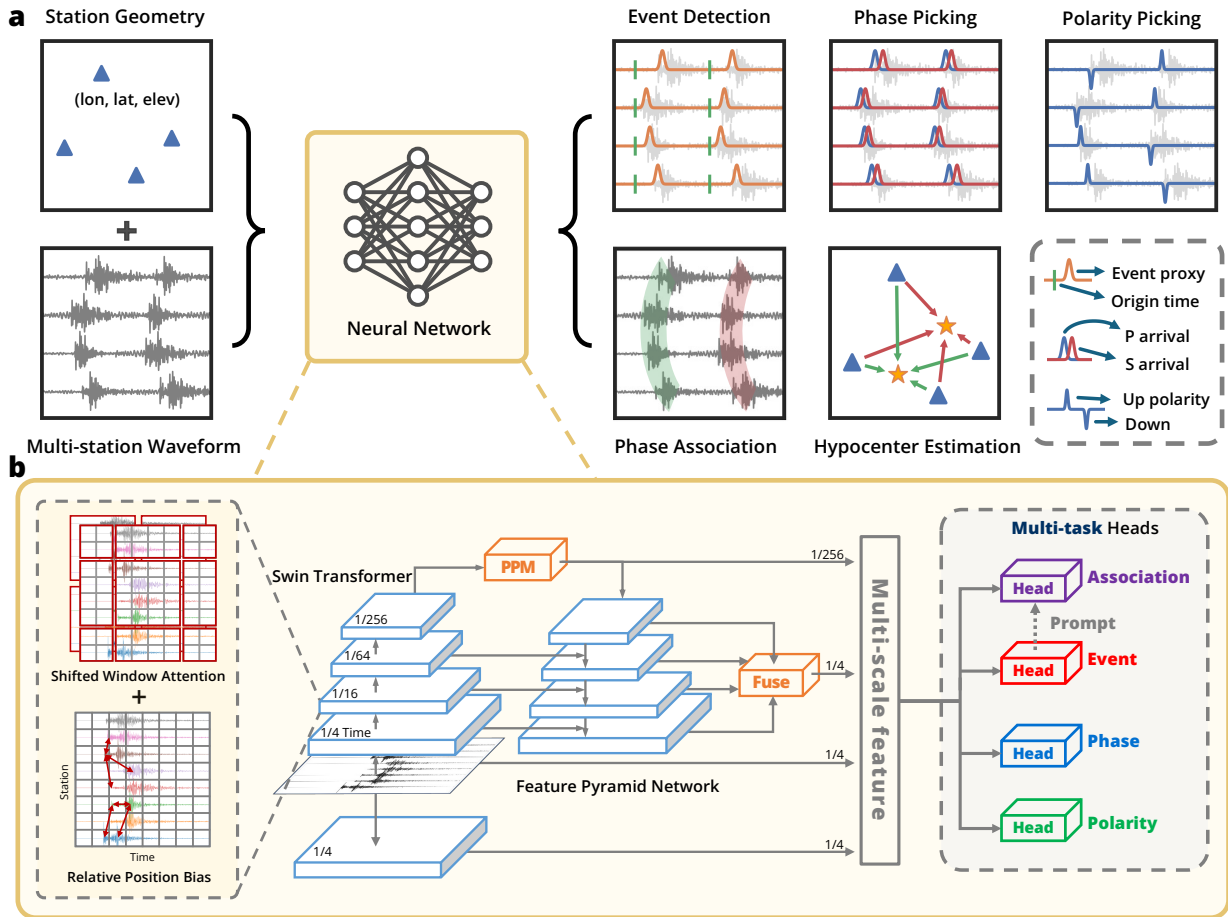


Figure 1: **Event-centric formulation and geometry-aware architecture of EQNet2.** **a**, Multi-station waveforms and relative station geometry are encoded into a shared representation from which phase arrivals, first-motion polarities, event-arrival proxies, source-related estimates, and network-wide association probability masks are jointly predicted for the same candidate events. **b**, Detailed structure of EQNet2. Multi-station waveforms are processed by a hierarchical Swin Transformer with spatiotemporal relative position bias and fused through a feature pyramid network before task-specific decoding. Predicted event-arrival proxies serve as prompts for event-specific association. PPM, pyramid pooling module (Zhao et al., 2017). Detailed architectures of the prediction heads are shown in Supplementary Fig. 1.

Figure 2 illustrates the joint outputs for event nc73866925 from the NCEDC 2023 test dataset. At each

station, EQNet2 predicts P- and S-phase arrivals, a first-motion polarity, and source-related estimates. The arrivals follow a coherent moveout across the network, and the station-wise origin-time estimates cluster around a common value. A single event-arrival prompt query groups these station-level predictions into one candidate event through the association probability mask. The analyses in Section 3 further quantify the accuracy of each prediction task and its effect on catalog construction.

2.2 Geometry-aware multi-station representation

EQNet2 pairs a shared backbone, which extracts features from all stations jointly, with parallel task-specific heads that decode these features into predictions (Figure 1b). The backbone is based on the shifted-window (Swin) Transformer (Z. Liu et al., 2022, 2021). It processes multi-station waveform tokens (short waveform segments) with windowed self-attention to balance modeling capacity and computational cost, where each attention window spans a range of time steps and a subset of stations (Supplementary Note 1). The multiscale features from the Swin backbone are fused by a feature pyramid network (FPN) module (Lin, Dollár, et al., 2017; Xiao, Liu, Zhou, Jiang, & Sun, 2018), producing representations suitable for tasks that require different temporal resolutions.

A key component of the backbone is the geometry-aware relative position bias (RPB) in each attention module. The relative arrival times of seismic phases at different stations depend on the geometry of the network relative to the source, whereas a standard attention mechanism (Vaswani et al., 2017) carries no geometric information. EQNet2 therefore adds a learnable bias term to the attention scores so that the exchange of information between waveform tokens can depend on their relative timing and station geometry. We incorporate the RPB term B into the attention computation:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(\cos(Q, K)/\tau + B) V, \quad (1)$$

where Q , K , and V are the query, key, and value feature matrices in the attention module, respectively, and τ is a learnable temperature parameter that scales the cosine similarity between query and key features. The bias term B is generated from the relative spatiotemporal coordinates between each pair of waveform tokens:

$$B = \mathcal{G}(\Delta t, R, \theta, \Delta z), \quad (2)$$

where Δt is the time difference between two waveform tokens; R is the great-circle distance between the corresponding stations; θ is the inter-station azimuth, represented by its sine and cosine; and Δz is the relative elevation. The function $\mathcal{G}$ is a two-layer multilayer perceptron (MLP). Each Transformer block has an independent $\mathcal{G}$, allowing different layers to learn distinct spatiotemporal interaction preferences. This design lets EQNet2 relate waveforms across stations under irregular network geometry, without a fixed station layout or ordering.

2.3 Event-centric prediction heads

Task-specific heads decode the shared multi-station representation into station-level and event-level outputs (Supplementary Fig. 1). The Phase Head outputs per-sample probabilities for noise, P-phase arrival, and S-phase arrival classes, trained as a segmentation task with Gaussian targets centered at manual picks (Zhu & Beroza, 2019). The Polarity Head outputs per-sample probabilities for upward, unknown, and downward first-motion categories, following PhaseNet+ (Zhu, Song, et al., 2025). Because the three polarity classes are ordered (Down, Unknown, Up), we formulate polarity prediction as an ordinal regression task (McCullagh, 1980).

The Event Head jointly performs event detection and source-related regression tasks. Its detection branch outputs a probability sequence for the event-arrival proxy, which serves as a reference time anchor for extracting phase arrivals and source-related estimates at each station. Separate regression branches predict travel time, S–P differential time, epicentral distance, backazimuth, depth, and magnitude. The regression branches additionally receive raw amplitude information because absolute and relative amplitudes constrain source location and magnitude.

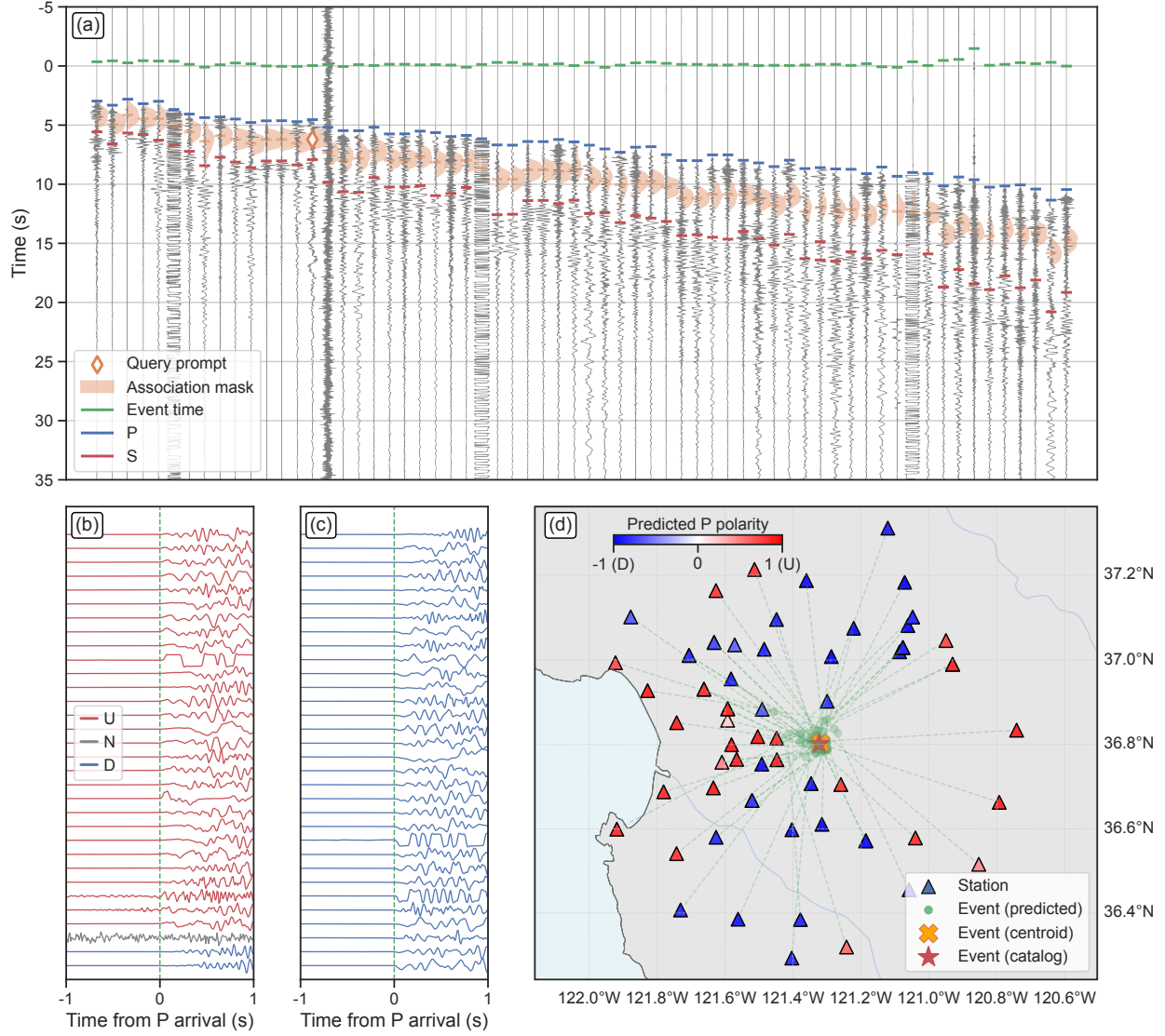


Figure 2: **Example of event-level outputs from EQNet2.** **a**, Vertical-component waveforms sorted by source-station distance. Blue and red ticks mark predicted P- and S-phase arrivals, respectively; green ticks mark station-wise origin-time estimates obtained from the predicted event-arrival proxies and travel times. The orange diamond denotes the query event-arrival proxy, and orange shading shows the corresponding association probabilities. **b,c**, Vertical-component waveforms aligned on the predicted P arrivals and colored by predicted first-motion polarities. Polarities are categorized as Up (U), Unknown (N), or Down (D) using a score threshold of 0.3. **d**, Seismic stations, station-wise epicentral estimates, their centroid, and the reference epicenter. Dashed lines indicate predicted backazimuths, and station colors indicate predicted first-motion polarities.

The Association Head groups arrivals into events directly in the same feature space used for detection and picking. Its design is inspired by prompt-based segmentation in the Segment Anything Model (SAM) (Kirillov et al., 2023; Ravi et al., 2024). Unlike the user-provided spatial prompts in SAM, the EQNet2 prompt is a model-predicted event-arrival proxy. Each candidate proxy is encoded as a query token that attends to backbone features across all stations (Supplementary Fig. 1), so the head has access to both waveform features and relative station geometry. Its station positional encoding follows the same relative-geometry design as the backbone RPB, maintaining geometric consistency between the two modules. The output is a network-wide association mask that gives the probability that each detected arrival belongs to the same event as the query (Figure 2). During inference, association masks are converted into candidate events by the graph procedure described in Supplementary Note 1.

2.4 Training data and optimization

Training a multi-station, multi-task model requires datasets with comprehensive annotations for all relevant tasks linked at the event level. We use the California Earthquake Event Dataset (CEED) (Zhu, Wang, et al., 2025), which organizes its records by event, each containing the associated station waveforms, phase picks, polarities, and hypocenter parameters. The Northern California Earthquake Data Center (NCEDC) subset of CEED through 2022 is used for training, and the NCEDC data from 2023 are held out for evaluation. The training subset contains approximately 325,000 events and 1.1M three-component waveform records. Each waveform record spans 120 s at 100 Hz. During training, we randomly select 16 waveform traces per event and crop each trace to a 40.96-s window with a random time shift. We also retain waveforms without manual labels when at least one trace from the same event contains labels. Such traces occur because not every station recording associated with an event has complete manual annotations, especially for weak phases or low-quality components. Retaining them allows the model to learn from the available waveform context while the supervised losses are applied only to annotated targets.

The model is trained with a composite multi-task loss that combines phase picking, event-arrival-proxy detection, polarity prediction, association, and source-related regression objectives; full loss definitions and optimization settings are provided in Supplementary Note 3. To improve generalization and robustness to complex seismic signals, we employ two data augmentation strategies during training: temporal event stacking and spatial event concatenation. Temporal event stacking superimposes waveform records from different events with randomized time offsets, simulating overlapping arrivals from closely spaced events in time. Spatial event concatenation combines stations from different events within the same training window, simulating cases in which different parts of the network record different nearby sources or unrelated events. These augmentations expose the model to temporally and spatially overlapping events, encouraging it to rely on network coherence when separating event-related signals from unrelated arrivals.

3 Results

3.1 Network information improves the picking of weak arrivals and the estimation of source parameters

We evaluated EQNet2 on fixed-length waveform windows from the NCEDC 2023 test dataset for phase picking, event detection, first-motion polarity, source-related estimation, and association. We compared EQNet2 with PhaseNet (Zhu & Beroza, 2019), PhaseNet+ (Zhu, Song, et al., 2025), and PhaseNO (Sun et al., 2023), as well as a single-station variant that retains the same task formulation but removes cross-station information. Because catalog annotations do not exhaustively label every arrival in each waveform window, a prediction outside the annotated event interval may correspond to an unlabeled neighboring event or an incomplete manual annotation rather than a true false positive. Phase-picking metrics are therefore evaluated only within an event-centered interval extending from 5 s before the catalog P arrival to three times the catalog S-P differential time after the P arrival.

For phase picking, EQNet2 achieves P-phase performance comparable to the leading baselines and the highest overall S-phase performance (Supplementary Fig. 2). EQNet2 also outperforms its single-station variant and produces the smallest standard deviations in the P- and S-phase picking residuals, while maintaining mean residuals comparable to those of the baselines (Supplementary Fig. 3; Supplementary Table 1). Together,

these results indicate that network information improves both recovery and timing precision. We stratified performance by SNR (Figure 3) and other parameters (Supplementary Note 4). EQNet2 achieves the highest recall for both phases in every SNR group, with the largest gains at low SNR. EQNet2 also achieves the highest F1 scores below 10 dB and comparable values in high SNR groups. Measured precision is lower for the multi-station models, partly because the selected evaluation time window can still include uncataloged events in dense scenarios (Supplementary Fig. 9).

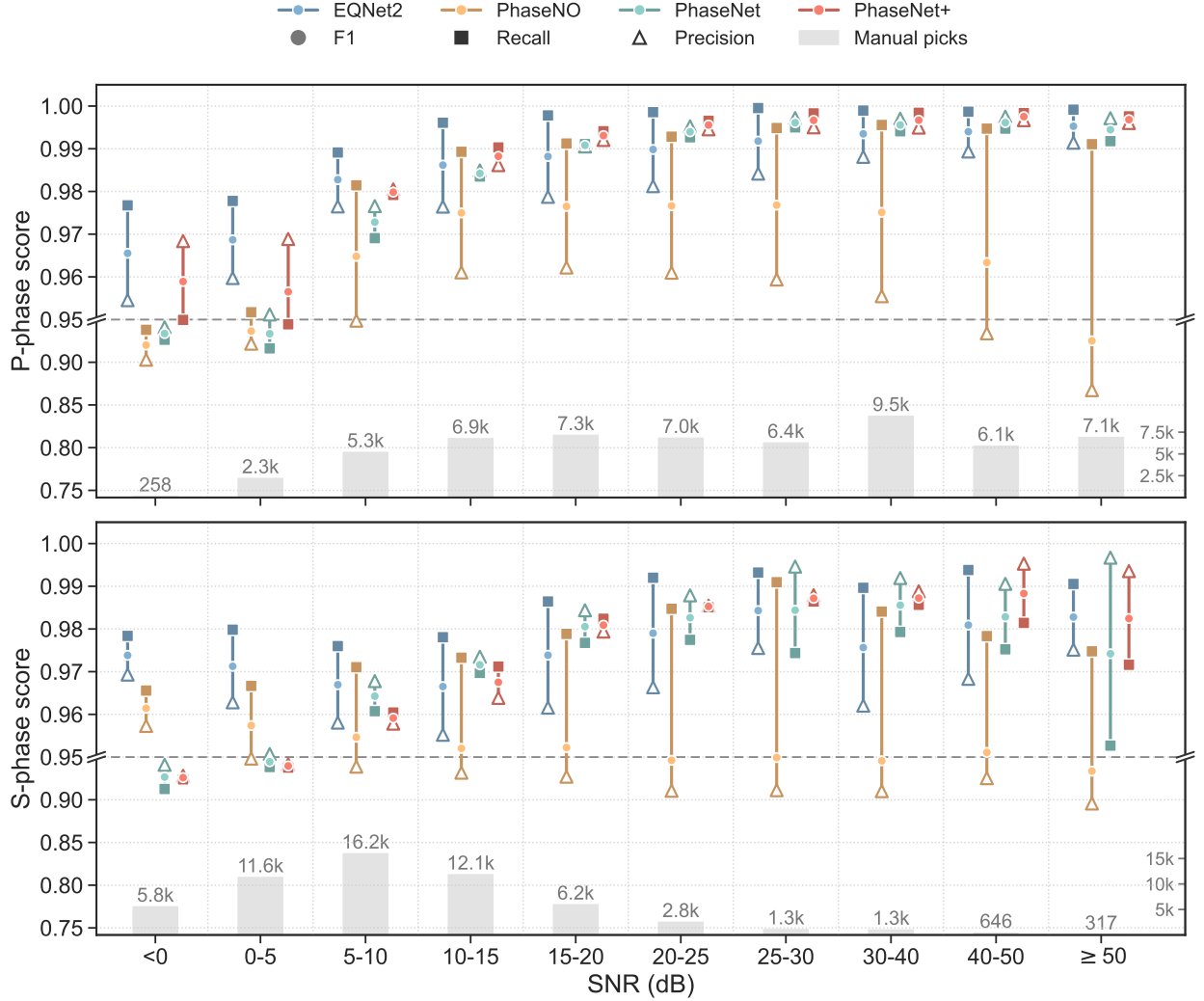


Figure 3: **Phase-picking performance across signal-to-noise-ratio groups.** Precision (triangles), recall (squares), and F1 score (circles) of P-phase (top) and S-phase (bottom) picking on the NCEDC 2023 dataset. The vertical axis is compressed below 0.95 (dashed line). Gray bars show the number of manual picks in each group (right axis).

Beyond phase arrivals, all compared models that include event detection achieve high detection rates, whereas the clearest differences appear in the source-related estimates (Supplementary Fig. 2; Supplementary Table 2). In particular, the standard deviation of the backazimuth residual decreases from 44.86° for the single-station variant to 7.43° for multi-station EQNet2 (Supplementary Fig. 3), consistent with backazimuth being constrained primarily by network geometry rather than any single waveform.

For first-motion polarity, we categorized the predictions of EQNet2 and PhaseNet+ using a threshold of 0.3 (Up > 0.3 , Down < -0.3 , Unknown otherwise). EQNet2 improves classification of the Up and Down classes relative to PhaseNet+ and the CNN model of Ross, Meier, and Hauksson (2018) (Supplementary

Fig. 4). Performance is lower for the “Unknown” class. Qualitative inspection suggests that this class is heterogeneous and includes many high-SNR waveforms with visually discernible polarities (Zhu, Song, et al., 2025). Both EQNet2 and PhaseNet+ tend to assign definitive labels (Up or Down) to a subset of these unknown samples, which lowers the Unknown-class accuracy without necessarily indicating an incorrect polarity estimate.

We also evaluated the prompt-based association module against spatiotemporal clustering of the predicted source-related estimates. For each test window, the association ratio measures the fraction of picked arrivals that are assigned to a predicted event group. When k of n picks in a window remain unassociated, the ratio equals $1 - k/n$, and the values therefore fall on discrete curves (Supplementary Fig. 5). The prompt-based module leaves picks unassociated in fewer windows, and its advantage grows for windows with many picks, where clustering more often fails to associate several arrivals. Predicted event-arrival proxies are thus effective prompts for gathering picks into events, leaving fewer picks unassociated.

To examine missed detections in dense-event scenarios, we inspected a randomly selected sample from the NCEDC 2023 test dataset containing two cataloged events (Supplementary Fig. 9). Among the four compared models, EQNet2 recovers the most complete set of P- and S-phase picks, including weak arrivals whose timing is consistent across stations. Its event-time predictions are also more complete and less scattered than those of PhaseNet+. EQNet2 nevertheless misses some event-arrival proxies and produces occasional isolated S-phase picks in this example. Overall, the example is consistent with the aggregate improvements in S-phase and low-SNR picking observed across the benchmark.

3.2 Continuous monitoring yields a more complete and spatially coherent Ridgecrest catalog

Fixed-window performance does not by itself establish catalog quality. We therefore test whether these improvements translate into end-to-end monitoring with continuous waveforms. The 2019 Ridgecrest earthquake sequence featured a M_w 7.1 mainshock on July 5, preceded by a M_w 6.4 foreshock on July 4, followed by a vigorous aftershock sequence that activated a complex fault structure at depth (Ross, Idini, et al., 2019; Shelly, 2020). The high event rate and structural complexity make it a challenging test for detection, association, and overall catalog quality. We processed continuous waveform data from July 4 to July 10, 2019, in a $1^\circ \times 1^\circ$ region centered at 117.504°W , 35.705°N . The data comprise 45 sensors from 16 CI stations, including HH, EH, HN, and BH channels. We compare the EQNet2 catalog with the Southern California Seismic Network (SCSN) catalog and with two catalogs built from the same data, one from PhaseNet+ picks associated by predicted origin time (Zhu, Song, et al., 2025) and one from PhaseNO picks associated with GaMMA (Zhu, McBrearty, et al., 2022). Associated picks were located using ADLoc (Zhu, Rong, et al., 2025) with the SCSN 1D velocity model (Hutton, Woessner, & Hauksson, 2010).

Figure 4 compares the temporal, magnitude, depth, and spatial distributions across four catalogs. The EQNet2 catalog contains 29,058 events, slightly more than the PhaseNO+GaMMA catalog (27,532), approximately 20% more than the PhaseNet+ catalog (24,318), and more than twice as many as the SCSN catalog (11,499). Relative to the SCSN catalog, EQNet2 recovers 89.3% of the events, compared with 87.1% for PhaseNO+GaMMA and 83.4% for PhaseNet+. At the pick level, 73.9% (1.44 M / 1.95 M) of the raw EQNet2 picks are associated into events and used for location, compared with 70.9% (1.14 M / 1.61 M) for PhaseNO+GaMMA and 59.2% (0.90 M / 1.52 M) for PhaseNet+. Thus, the larger EQNet2 catalog is supported by both higher recovery of SCSN events and a larger fraction of raw picks contributing to located events. The additional EQNet2 events concentrate at depths of 2–6 km and align with the orthogonal subfault structures identified by Ross, Idini, et al. (2019). The EQNet2 catalog delineates fault structures more clearly with less spatial scatter, particularly toward the southeastern end of the fault system.

Before absolute location, hypocenter estimates obtained directly from EQNet2 already capture the general fault geometry (Supplementary Fig. 10). Increasing station-window size from the 16 traces used during training to all 45 traces reduces spatial scatter and yields more concentrated event locations. The accompanying reduction in the number of associated events is consistent with fewer split events when more stations constrain each candidate event.

Using the EQNet2 first-motion polarities, we computed focal mechanisms with SKHASH (Skoumal, Hardebeck, & Shearer, 2024). The resulting compressional-axis (P-axis) azimuths are consistent with the dominant spatial patterns in the SCSN catalog (Yang, Hauksson, & Shearer, 2012), the catalog of Cheng et

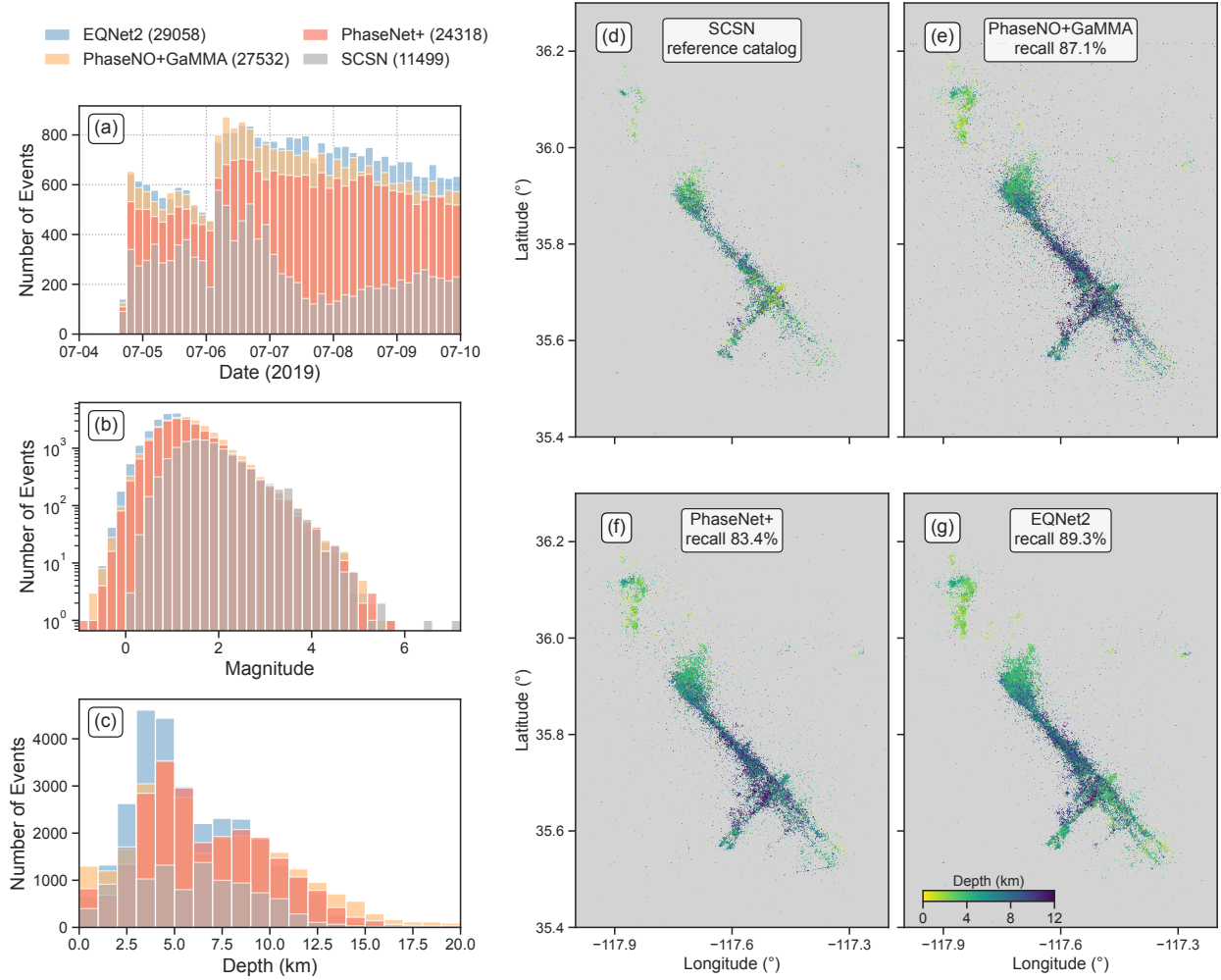


Figure 4: **Comparison of earthquake catalogs for the 2019 Ridgecrest sequence.** **a**, Temporal event frequency. **b**, Magnitude distribution. **c**, Depth distribution. **d–g**, Spatial distributions for SCSN, PhaseNO+GaMMA, PhaseNet+, and EQNet2. Event counts are shown in the legend, and labels in panels e–g report event recovery relative to the SCSN catalog. Colors in panels d–g indicate event depth.

al. (2023), and the PhaseNet+ catalog (Supplementary Fig. 11). EQNet2 produces more focal-mechanism solutions in the northern and central parts of the sequence, mainly within the interior of the station network where azimuthal coverage is more favorable.

3.3 Network coherence recovers weak arrivals and suppresses false detections

The Ridgecrest sequence primarily tests event recovery under dense seismicity. We next examine one day of continuous data around the 2025 Berkeley M 4.3 earthquake, where strong cultural (anthropogenic) noise provides a complementary test of whether network information can simultaneously recover weak event-related arrivals and reject non-earthquake signals. We processed waveforms from 69 stations within a $0.6^\circ \times 0.6^\circ$ region centered on the epicenter.

EQNet2 produces fewer raw P- and S-phase picks than PhaseNet+ (15,871 versus 30,659), but retains a substantially larger fraction after association and absolute location: 3,501 picks for EQNet2 (22.1%) and 1,000 picks for PhaseNet+ (3.3%) (Figure 5e). In the 2-hour window around the mainshock, many PhaseNet+ event predictions remain unassociated and scatter across the network, whereas the EQNet2 predictions concentrate into aftershocks shared across stations. The enlarged panels show the contrast at the phase level, where EQNet2 recovers P- and S-phase arrivals at more stations for each associated event, giving the aftershocks more complete arrival patterns (Figure 5a–d). Both models recover similar numbers of high-SNR arrivals, whereas the additional located EQNet2 picks are predominantly low-SNR arrivals supported by coincident detections at neighboring stations (Supplementary Note 6). Most unassociated picks from both models are isolated detections with no coincident pick elsewhere in the network, but PhaseNet+ produces 2.8 times as many isolated detections as EQNet2. Over the full day, EQNet2 locates 76 events, compared with 27 for PhaseNet+ and 3 in the USGS catalog. The relative advantage of EQNet2 also persists in a test with more distant stations around the M_w 7.1 Ridgecrest mainshock (Supplementary Fig. 12). Although both models degrade with increasing source–station distance, EQNet2 retains more arrivals that align across the network.

3.4 Learned attention across stations follows seismic moveout

The preceding tests establish the empirical benefits of network information. We next examine the station-to-station interactions learned by EQNet2. The learned relative position biases (RPBs), which bias the attention between two waveform segments according to their separation in time and space, change systematically with model depth (Figure 6; Supplementary Figs. 14 and 15). In shallow stages, high biases remain concentrated near small inter-station distances, emphasizing local information from the same or adjacent stations. In deeper stages, the high-bias patterns extend toward more distant stations and become increasingly slanted as the time span visible to each token grows. The slopes of some RPB heads correspond to apparent velocities of approximately 6–8 km/s, indicating that the model preferentially connects waveform segments whose time offsets match plausible travel times between stations.

Attention maps provide complementary event-specific evidence (Figure 7; Supplementary Fig. 16). When the model queries a segment containing a P arrival, it assigns high attention to P arrivals at other stations, particularly at nearby stations or stations with similar waveforms. These patterns recur across query stations, indicating that the model links arrivals sharing a common moveout rather than treating each station independently.

4 Discussion and Conclusions

EQNet2 performs the tasks of earthquake monitoring in a single model that learns from multiple stations jointly and organizes its predictions around individual earthquakes. In a fixed-window benchmark on Northern California data excluded from training, EQNet2 improves the picking of low-SNR arrivals. On continuous waveforms from the 2019 Ridgecrest sequence and the 2025 Berkeley earthquake, these gains translate into more recovered events, fewer false detections, and a substantially larger fraction of picks retained after association and location. The learned relative position biases form time–distance patterns consistent with seismic moveout, and the attention maps connect the arrivals of the same event across stations. These results show that combining network information with event-centric prediction improves both the completeness and the reliability of the catalog.

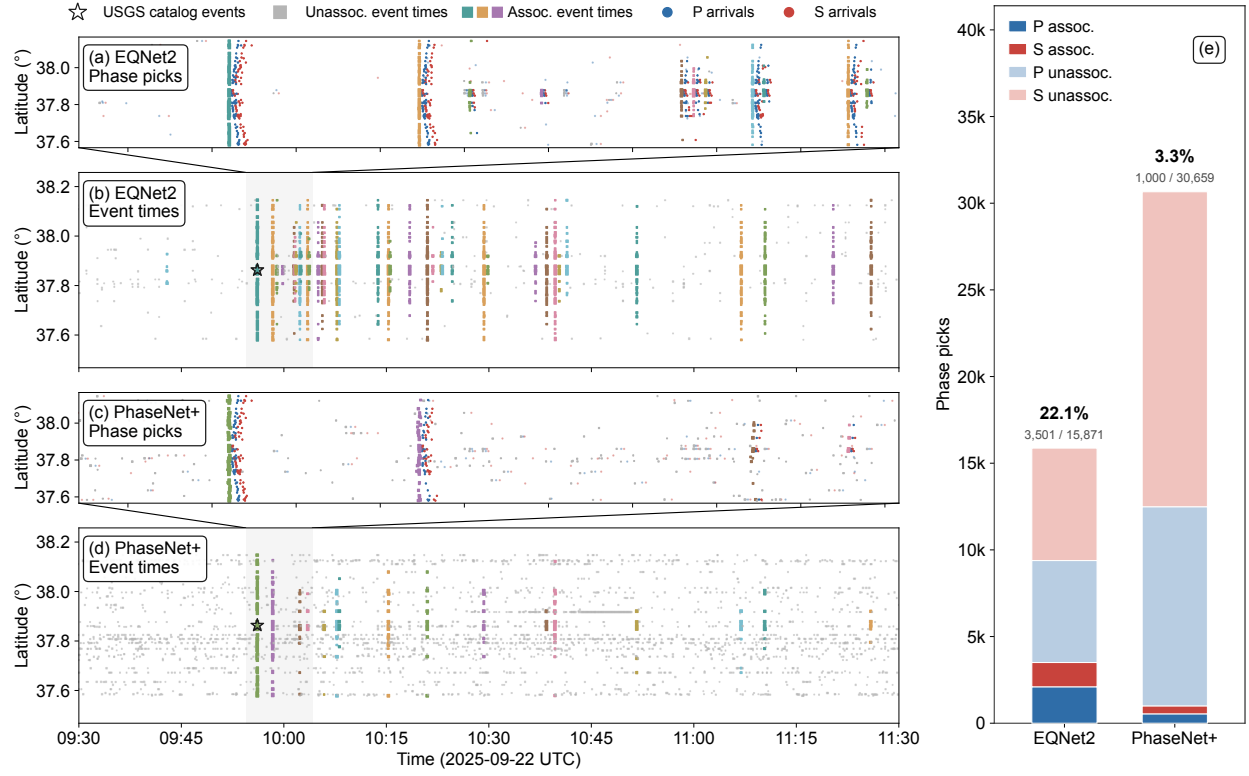


Figure 5: **Comparison of EQNet2 and PhaseNet+ for the 2025 Berkeley earthquake sequence.** **a–d**, Predictions in a 2-hour window around the 2025-09-22 Berkeley M 4.3 earthquake (event ID: nc75240492). Each square in panels b and d denotes a predicted event origin time at the corresponding station. Gray squares are unassociated predictions, colored squares indicate predictions associated with the same event, and the black star marks the USGS catalog event. Panels a and c enlarge the shaded interval and additionally show predicted P- and S-phase arrivals. **e**, Stacked counts of associated and unassociated P- and S-phase picks. Percentages indicate the fraction of raw picks retained after association and location.

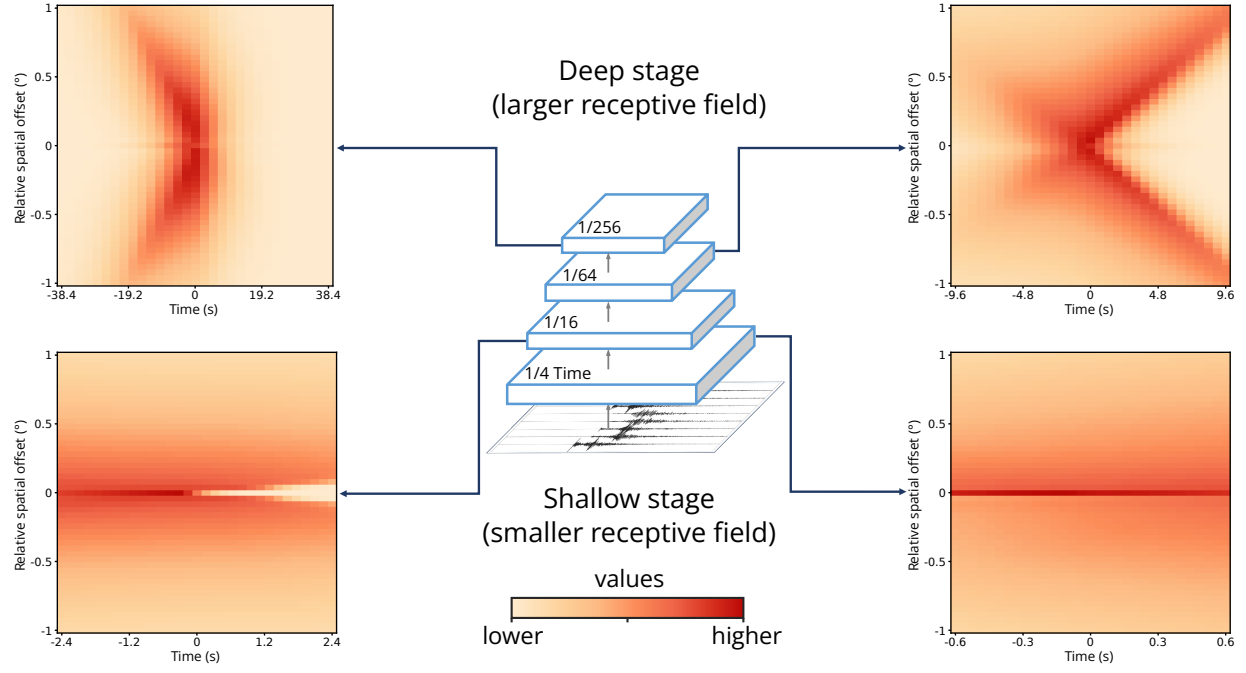


Figure 6: **Relative position biases across the EQNet2 backbone.** One attention head is shown for each of the four Swin Transformer stages. The horizontal axes denote relative time between waveform tokens, and the vertical axes denote the signed spatial offset of one station relative to the other, in degrees. Because the learned biases are nearly rotationally symmetric, the inter-station azimuth is fixed in the north–south direction and the relative elevation at zero. Darker colors indicate larger bias values. More visualizations are provided in Supplementary Figs. 14 and 15.

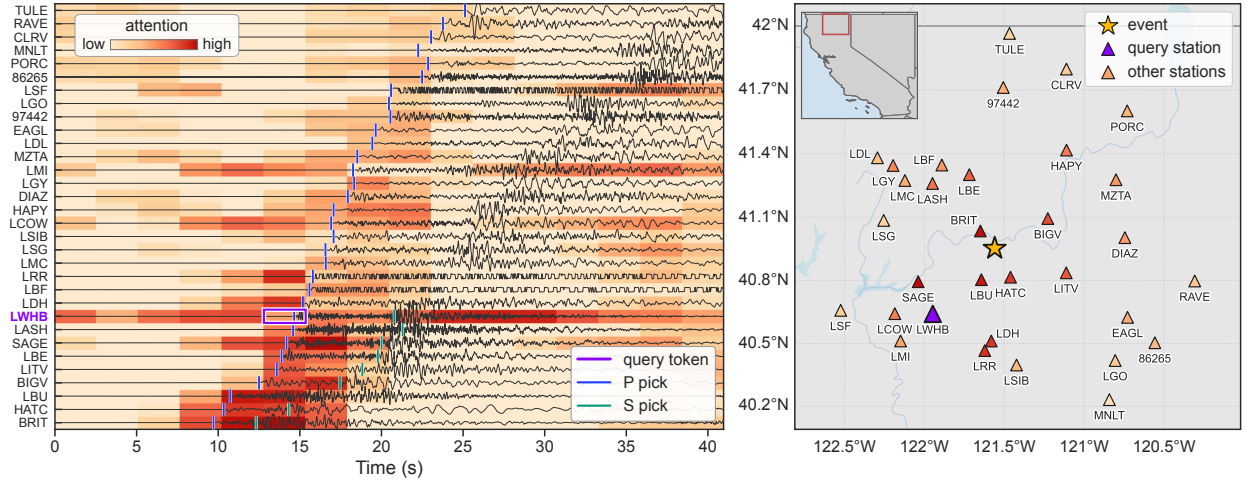


Figure 7: **Cross-station attention for a representative event.** The attention map for event nc73934641 from the last Swin Transformer block is shown together with the corresponding station distribution. The query token containing a P-phase arrival is enclosed by a purple rectangle, the query station is marked by a purple triangle, and darker colors indicate higher attention weights.

Multi-station modeling helps mainly by distinguishing weak arrivals that are coherent across the network from noise that is confined to a single station. A low-SNR arrival may be ambiguous at a single station but gains support from its consistent timing across the network. Conversely, cultural or instrumental noise can mimic a phase at one station but shows no consistent moveout across the network. The time–distance patterns in the learned relative position biases and the attention maps that connect the arrivals of the same event across stations indicate that the model has learned to exploit this coherence. Multi-station models have shown that network information can improve phase picking (G. Chen & Li, 2022; Feng et al., 2022; Sun et al., 2023). EQNet2 extends this by conditioning the attention between stations on their relative geometry throughout the shared backbone, so that the geometric relations between stations inform the picking, source, and association predictions alike.

Event-centric prediction contributes largely by optimizing the tasks jointly rather than in sequence, on a shared representation of each earthquake. Phase arrivals, first-motion polarities, event-arrival proxies, source-related estimates, and association masks are different measurements of the same earthquake. Predicting them from a shared representation lets the evidence for one measurement support the other measurements and encourages the outputs to be mutually compatible. In a sequential workflow, by contrast, independently optimized modules communicate primarily through discrete intermediate predictions, which discard the waveform evidence and its uncertainty and let early errors propagate. For example, the prompt-based association module takes a candidate event-arrival proxy as the prompt and predicts a network-wide association probability mask rather than treating association as a pairwise matching task (McBrearty et al., 2019; Ross, Yue, et al., 2019). Multi-task models have shown that related outputs can benefit from joint learning (Si et al., 2024; J. Zhang et al., 2023; Zhu, Song, et al., 2025; Zhu, Tai, et al., 2022). EQNet2 extends this by organizing all of its outputs around candidate events, so that the station-level measurements and the network-wide association and prediction refer to the same earthquake.

The computational efficiency of EQNet2 balances multi-station modeling, which increases the cost of each prediction, against event-centric prediction, which reduces the complexity of the whole monitoring workflow. Exchanging information across stations adds computation compared with processing each station independently, while the resulting features are shared across tasks and lead to more reliable detections. Event-centric prediction in turn reduces the complexity of the workflow by replacing several separate models and intermediate stages with one shared backbone and parallel task heads. EQNet2 is thus more expensive than lightweight single-station models but faster and more memory-efficient than other multi-station models such as PhaseNO under the same benchmark setting (Supplementary Table 3). In the Ridgecrest test, processing one day of continuous waveforms from 45 sensors takes about 10 minutes on a single NVIDIA A100 GPU, and post-processing all six days of output takes a few additional minutes. Model capacity and station-window size provide practical controls on the performance–efficiency trade-off. Smaller EQNet2 variants reduce runtime with a limited decrease in phase-picking performance on the NCEDC test dataset (Supplementary Table 3), whereas larger variants produce fewer raw picks and retain a larger fraction of them after association and location in continuous data (Supplementary Table 4). Enlarging the station window includes more stations in the attention for each event, which reduces split events and raises the fraction of picks retained (Supplementary Fig. 10; Supplementary Table 4), at the cost of higher memory use.

Several limitations of the present model point to future development. First, event-centric training depends on datasets in which the phase, polarity, source, and association labels are consistent for each event, and on augmentation that preserves this consistency; larger and more diverse datasets of this kind would improve generalization. Second, multi-task optimization remains challenging because picking, polarity, source-related estimation, and association losses can have different scales and convergence rates (S. Liu, Johns, & Davison, 2019). Third, applying the model to larger networks will require sparser station selection or hierarchical station windows to keep the computational cost manageable. In addition, some RPB heads are asymmetric in time (Figure 6), which we cannot yet explain physically. Understanding these learned patterns may help assess the reliability and generalization of the model.

This study shows that combining network coherence with event-centric prediction can recover low-SNR arrivals and suppress false detections, yielding more complete and more reliable catalogs. The attention learned across the network follows the physics of seismic wave propagation, suggesting that the model improves its predictions through a physically interpretable representation of the travel times between stations. The gains in completeness and reliability support moving from station-by-station and step-by-step workflows toward models that combine geometry-aware attention across stations with event-centric prediction. Future

work will focus on generalization across regions and network geometries and on scaling to larger and denser seismic networks, where the coherence across closely spaced stations could yield catalogs that resolve fault structure and earthquake processes in finer detail.

Data Availability

The CEED dataset used for training is available at Hugging Face (Zhu, Wang, et al., 2025). The continuous seismic data for the 2019 Ridgecrest earthquake sequence, the waveform-relocated SCSN catalog, and the refined focal mechanism solutions by Cheng et al. (2023) were downloaded from SCEDC (2013). The continuous seismic data for the 2025 Berkeley earthquake were downloaded from NCEDC (2014).

Code Availability

Our source code and model are publicly available at <https://github.com/AI4EPS/EQNet>.

Acknowledgements

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award DE-SC0026120. This research used the Lawrence Livermore computational cluster provided by the IT Division at Lawrence Berkeley National Laboratory (supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy User Facility (project m4781-2024). We thank Bo Rong, Xuancheng Huang, Junhao Song, and Savvas Marcou for their valuable suggestions on paper writing and figure style.

References

- Chakraborty, M., Cartaya, C. Q., Li, W., Faber, J., Rumpker, G., Stoecker, H., & Srivastava, N. (2022). PolarCAP – a deep learning approach for first motion polarity classification of earthquake waveforms. *Artificial Intelligence in Geosciences*, 3, 46–52.
- Chen, G., & Li, J. (2022). CubeNet: Array-based seismic phase picking with deep learning. *Seismological Research Letters*, 93(5), 2554–2569. doi: 10.1785/0220220147
- Chen, Y., Savvaidis, A., Siervo, D., Huang, D., & Saad, O. M. (2024). Near real-time earthquake monitoring in Texas using the highly precise deep learning phase picker. *Earth and Space Science*, 11(10), e2024EA003890. doi: 10.1029/2024EA003890
- Cheng, Y., Hauksson, E., & Ben-Zion, Y. (2023). Refined earthquake focal mechanism catalog for Southern California derived with deep learning algorithms. *Journal of Geophysical Research: Solid Earth*, 128(2), e2022JB025975.
- Feng, T., Mohanna, S., & Meng, L. (2022). EdgePhase: A deep learning model for multi-station seismic phase picking. *Geochemistry, Geophysics, Geosystems*, 23(11), e2022GC010453.
- Feng, T., Wang, R., Fang, L., Wu, J., & Yang, T. (2024). Microseismic constraints on the mechanical state of the Anninghe-Daliangshan fault zone, southeastern Tibetan Plateau. *Earth and Planetary Science Letters*, 639, 118765. doi: 10.1016/j.epsl.2024.118765
- Hara, S., Fukahata, Y., & Iio, Y. (2019). P-wave first-motion polarity determination of waveform data in western Japan using deep learning. *Earth, Planets and Space*, 71(1), 127.
- Hauksson, E., & Shearer, P. (2005). Southern California hypocenter relocation with waveform cross-correlation, part 1: Results using the double-difference method. *Bulletin of the Seismological Society of America*, 95(3), 896–903.
- Hauksson, E., Yang, W., & Shearer, P. M. (2012). Waveform relocated earthquake catalog for southern California (1981 to June 2011). *Bulletin of the Seismological Society of America*, 102(5), 2239–2244.

- Herrmann, M., Piegari, E., & Marzocchi, W. (2022). Revealing the spatiotemporal complexity of the magnitude distribution and b-value during an earthquake sequence. *Nature Communications*, 13(1), 5087.
- Hutton, K., Woessner, J., & Hauksson, E. (2010). Earthquake monitoring in southern California for seventy-seven years (1932–2008). *Bulletin of the Seismological Society of America*, 100(2), 423–446.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 4015–4026).
- Kwiatek, G., Martínez-Garzón, P., Becker, D., Dresen, G., Cotton, F., Beroza, G. C., ... Bohnhoff, M. (2023). Months-long seismicity transients preceding the 2023 M_W 7.8 Kahramanmaraş earthquake, Türkiye. *Nature Communications*, 14(1), 7534.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 936–944).
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 2980–2988).
- Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 1871–1880).
- Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., ... Guo, B. (2022). Swin transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 12009–12019).
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 10012–10022).
- Lomax, A., Anagnostou, V., Karakostas, V., Hicks, S. P., & Papadimitriou, E. (2025). The 2025 Santorini unrest unveiled: Rebounding magmatic dike intrusion with triggered seismicity. *Science*, 390(6775), eadz8538.
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations (ICLR)*.
- McBrearty, I. W., & Beroza, G. C. (2022). Earthquake location and magnitude estimation with graph neural networks. In *2022 IEEE international conference on image processing (ICIP)* (pp. 3858–3862).
- McBrearty, I. W., & Beroza, G. C. (2023). Earthquake phase association with graph neural networks. *Bulletin of the Seismological Society of America*, 113(2), 524–547.
- McBrearty, I. W., Delorey, A. A., & Johnson, P. A. (2019). Pairwise association of seismic arrivals with convolutional neural networks. *Seismological Research Letters*, 90(2A), 503–509.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2), 109–127.
- Mousavi, S. M., & Beroza, G. C. (2020). Bayesian-deep-learning estimation of earthquake location from single-station observations. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11), 8211–8224. doi: 10.1109/TGRS.2020.2988770
- Mousavi, S. M., & Beroza, G. C. (2022). Deep-learning seismology. *Science*, 377(6607), eabm4470. doi: 10.1126/science.abm4470
- 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. doi: 10.1038/s41467-020-17591-w
- Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019). Unsupervised clustering of seismic signals using deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, 16(11), 1693–1697.
- NCEDC. (2014). *Northern California Earthquake Data Center*. Dataset. UC Berkeley Seismological Laboratory. doi: 10.7932/NCEDC
- Park, Y., Mousavi, S. M., Zhu, W., Ellsworth, W. L., & Beroza, G. C. (2020). Machine-learning-based analysis of the Guy-Greenbrier, Arkansas earthquakes: A tale of two sequences. *Geophysical Research Letters*, 47(6), e2020GL087032.
- Peng, Z., & Zhao, P. (2009). Migration of early aftershocks following the 2004 Parkfield earthquake. *Nature Geoscience*, 2(12), 877–881.

- Poggiali, G., Chiaraluce, L., Ross, Z. E., Zhu, W., & Marone, C. (2025). Fault geometry and source mechanics of the Altotiberina fault system from a high-resolution machine-learning earthquake catalog. *Bulletin of the Seismological Society of America*, 115(5), 2181–2201.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., ... Feichtenhofer, C. (2024). SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Ross, Z. E., & Cochran, E. S. (2021). Evidence for latent crustal fluid injection transients in Southern California from long-duration earthquake swarms. *Geophysical Research Letters*, 48(12), e2021GL092465.
- Ross, Z. E., Cochran, E. S., Trugman, D. T., & Smith, J. D. (2020). 3D fault architecture controls the dynamism of earthquake swarms. *Science*, 368(6497), 1357–1361.
- Ross, Z. E., Idini, B., Jia, Z., Stephenson, O. L., Zhong, M., Wang, X., ... Jung, J. (2019). Hierarchical interlocked orthogonal faulting in the 2019 Ridgecrest earthquake sequence. *Science*, 366(6463), 346–351.
- Ross, Z. E., Meier, M.-A., & Hauksson, E. (2018). P wave arrival picking and first-motion polarity determination with deep learning. *Journal of Geophysical Research: Solid Earth*, 123(6), 5120–5129.
- 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.
- Ross, Z. E., Yue, Y., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2019). PhaseLink: A deep learning approach to seismic phase association. *Journal of Geophysical Research: Solid Earth*, 124(1), 856–869.
- Saad, O. M., Chen, Y., Siervo, D., Zhang, F., Savvaadis, A., Huang, G.-c. D., ... Chen, Y. (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.
- SCEDC. (2013). *Southern California Earthquake Center*. Dataset. California Institute of Technology. doi: 10.7909/C3WD3xH1
- Schorlemmer, D., & Woessner, J. (2008). Probability of detecting an earthquake. *Bulletin of the Seismological Society of America*, 98(5), 2103–2117.
- Shelly, D. R. (2020). A high-resolution seismic catalog for the initial 2019 Ridgecrest earthquake sequence: Foreshocks, aftershocks, and faulting complexity. *Seismological Research Letters*, 91(4), 1971–1978.
- Si, X., Wu, X., Li, Z., Wang, S., & Zhu, J. (2024). An all-in-one seismic phase picking, location, and association network for multi-task multi-station earthquake monitoring. *Communications Earth & Environment*, 5(1), 22. doi: 10.1038/s43247-023-01188-4
- Skoumal, R. J., Hardebeck, J. L., & Shearer, P. M. (2024). SKHASH: A Python package for computing earthquake focal mechanisms. *Seismological Research Letters*, 95(4), 2519–2526.
- Song, X., Meier, M.-A., Ellsworth, W. L., & Beroza, G. C. (2026). FocoNet: Transformer-based focal-mechanism determination. *Journal of Geophysical Research: Machine Learning and Computation*, 3(2), e2025JH000879. doi: 10.1029/2025JH000879
- Soto, H., & Schurr, B. (2021). DeepPhasePick: a method for detecting and picking seismic phases from local earthquakes based on highly optimized convolutional and recurrent deep neural networks. *Geophysical Journal International*, 227(2), 1268–1294. doi: 10.1093/gji/ggab266
- Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (Vol. 10553, pp. 240–248). Springer. doi: 10.1007/978-3-319-67558-9_28
- Sun, H., Ross, Z. E., Zhu, W., & Azizzadenesheli, K. (2023). Phase Neural Operator for multi-station picking of seismic arrivals. *Geophysical Research Letters*, 50(24), e2023GL106434. doi: 10.1029/2023GL106434
- Suzuki, R., Uchida, N., Zhu, W., Beroza, G. C., Nakayama, T., Yoshida, K., ... Hasegawa, A. (2025). The forearc seismic belt: A fluid pathway constraining down-dip megathrust earthquake rupture. *Science*, 389(6756), 190–194.
- Tan, X., Tramelli, A., Gammaldi, S., Beroza, G. C., Ellsworth, W. L., & Marzocchi, W. (2025). A clearer view of the current phase of unrest at Campi Flegrei caldera. *Science*, 390(6768), 70–75.
- Tan, Y. J., Waldhauser, F., Ellsworth, W. L., Zhang, M., Zhu, W., Michele, M., ... Segou, M. (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.
- Trugman, D. T., & Ross, Z. E. (2019). Pervasive foreshock activity across southern California. *Geophysical Research Letters*, 46(15), 8772–8781.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Waldhauser, F., & Ellsworth, W. L. (2000). A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California. *Bulletin of the Seismological Society of America*, 90(6), 1353–1368.
- Wang, P., Wang, B., Peng, Z., & Lei, X. (2026). Reactivation of basement faults by deep fluids during the 2017 Changdao earthquake swarm, Eastern China. *Communications Earth & Environment*, 7(1), 207.
- Wilding, J. D., Zhu, W., Ross, Z. E., & Jackson, J. M. (2023). The magmatic web beneath Hawai‘i. *Science*, 379(6631), 462–468.
- Woessner, J., & Wiemer, S. (2005). Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, 95(2), 684–698.
- Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 418–434).
- Yang, W., Hauksson, E., & Shearer, P. M. (2012). Computing a large refined catalog of focal mechanisms for southern California (1981–2010): Temporal stability of the style of faulting. *Bulletin of the Seismological Society of America*, 102(3), 1179–1194.
- Zhang, J., Li, Z., & Zhang, J. (2023). Simultaneous seismic phase picking and polarity determination with an attention-based neural network. *Seismological Research Letters*, 94(2A), 813–828. doi: 10.1785/0220220247
- Zhang, M., Liu, M., Feng, T., Wang, R., & Zhu, W. (2022). LOC-FLOW: An end-to-end machine learning-based high-precision earthquake location workflow. *Seismological Research Letters*, 93(5), 2426–2438.
- Zhang, Z., Liu, M., Tan, Y. J., Walter, F., He, S., Chmiel, M., & Su, J. (2024). Landslide hazard cascades can trigger earthquakes. *Nature Communications*, 15(1), 2878.
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2881–2890).
- 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.
- 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., 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.
- Zhu, W., McBrearty, I. W., Mousavi, S. M., Ellsworth, W. L., & Beroza, G. C. (2022). Earthquake phase association using a Bayesian Gaussian mixture model. *Journal of Geophysical Research: Solid Earth*, 127(5), e2021JB023249. doi: 10.1029/2021JB023249
- Zhu, W., Mousavi, S. M., & Beroza, G. C. (2020). Seismic signal augmentation to improve generalization of deep neural networks. In B. Moseley & L. Krischer (Eds.), *Machine learning in geosciences* (Vol. 61, pp. 151–177). Elsevier. doi: 10.1016/bs.agph.2020.07.003
- Zhu, W., Rong, B., Jie, Y., & Wei, S. S. (2025). Robust earthquake location using random sample consensus (RANSAC). *arXiv preprint arXiv:2502.10933*.
- Zhu, W., Song, J., Wang, H., & Münchmeyer, J. (2025). Towards end-to-end earthquake monitoring using a multitask deep learning model. *arXiv preprint arXiv:2506.06939*.
- Zhu, W., Tai, K. S., Mousavi, S. M., Bailis, P., & Beroza, G. C. (2022). An end-to-end earthquake detection method for joint phase picking and association using deep learning. *Journal of Geophysical Research: Solid Earth*, 127(3), e2021JB023283. doi: 10.1029/2021JB023283
- Zhu, W., Wang, H., Rong, B., Yu, E., Zuzlewski, S., Tepp, G., ... Allen, R. M. (2025). California earthquake dataset for machine learning and cloud computing. *arXiv preprint arXiv:2502.11500*.

Supplementary Information

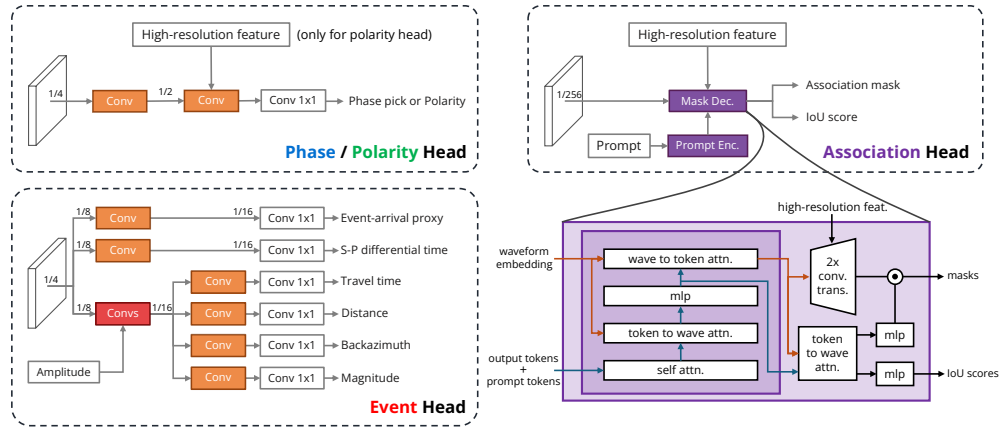
Supplementary Note 1: Details of model structure

Input waveforms are normalized and divided into short time segments (tokens). These tokens are then processed by a series of Swin Transformer blocks with progressive temporal downsampling, producing a hierarchy of multi-scale features. The key mechanism within these blocks is shifted window attention. Window attention is computed within local windows spanning both time and stations to reduce cost relative to global attention, and shifting the windows between successive blocks lets information flow across window boundaries. This design suits large seismic networks, where attention over all stations would be costly and the relevant information is expected to come mainly from nearby stations with similar waveforms.

The multi-scale backbone features are fused by an FPN module (Lin, Dollár, et al., 2017; Xiao et al., 2018). Since the backbone already models cross-station interactions through attention, the FPN and most task-specific heads primarily operate along the time dimension using one-dimensional convolutional modules.

The task heads are shown in Supplementary Fig. 1. The Phase, Polarity, and Event Heads consist mainly of 1D convolutional layers. To improve temporal precision, the Polarity Head additionally uses shallow high-resolution features from the backbone. The source-related estimates, such as travel time and epicentral distance, are read from the regression output sequences at the predicted event-arrival proxy time. Station-wise origin times are then obtained by subtracting the predicted travel time from the proxy time, and station-wise epicenters by projecting the predicted epicentral distance along the predicted backazimuth from each station.

The Association Head uses a prompt-based attention mechanism to predict association masks from backbone features. During inference, the Association Head is applied to all detected arrivals. To convert these masks into candidate events, we construct a graph whose nodes are detected event-arrival proxies, and edges represent association probabilities. Each connected component of the graph is treated as one candidate event, which merges queries belonging to the same earthquake.



Supplementary Fig. 1: Structure of task heads. The Association Head is modified from the Segment Anything Model (Kirillov et al., 2023; Ravi et al., 2024); the other heads mainly consist of 1D convolution modules.

Supplementary Note 2: Ordinal regression for polarity picking

The first-motion polarity can be classified into three classes: Down, Unknown, and Up, which have an inherent order. Because the three classes are ordered, we formulate polarity picking as ordinal regression and convert the model output into class likelihoods (McCullagh, 1980).

For a binary classification problem, the likelihoods of classes 0 and 1 can be $p_0 = 1 - \sigma(x)$ and $p_1 = \sigma(x)$, where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function and x is the predicted logit. For a general k -class classification, the problem can be solved by extending the sigmoid function to the softmax function and the scalar logit to a k -dimensional logit. However, for ordinal problems, we can interpret the binary case

as $p_0 = p(x \leq 0) = \sigma(0 - x)$ and $p_1 = p(x > 0) = 1 - p(x \leq 0)$. For a more general case, with cut points $\{c_0 = -\infty, c_1, c_2 = \infty\}$, binary classification can be described as

$$\begin{aligned} p_0 &= p(c_0 \leq x \leq c_1) = \sigma(c_1 - x) - \sigma(c_0 - x) = \sigma(c_1 - x) \\ p_1 &= p(c_1 < x \leq c_2) = \sigma(c_2 - x) - \sigma(c_1 - x) = 1 - \sigma(c_1 - x) \end{aligned} \quad (1)$$

Extending to the multinomial case, the likelihood of ordinal class k can be defined as

$$p_k = p(c_{k-1} < x \leq c_k) = \sigma(c_k - x) - \sigma(c_{k-1} - x) \quad (2)$$

The model is optimized by minimizing the negative log-likelihood of these class probabilities. For polarity picking, x is the output logit, the classes are ordered as Down, Unknown, and Up, and the cut points are $c_1 = -\theta$ and $c_2 = \theta$, with $c_0 = -\infty$ and $c_3 = \infty$, where θ is a hyperparameter.

Supplementary Note 3: Loss functions and training details

The total loss is a weighted sum of task losses:

$$\mathcal{L} = \lambda_{\text{phase}} \mathcal{L}_{\text{phase}} + \lambda_{\text{event}} \mathcal{L}_{\text{event}} + \lambda_{\text{pol}} \mathcal{L}_{\text{pol}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{asso}} \mathcal{L}_{\text{asso}} \quad (3)$$

It includes cross-entropy for phase picking and event-arrival-proxy detection, a combination of focal (Lin, Goyal, Girshick, He, & Dollár, 2017) and dice loss (Sudre, Li, Vercauteren, Ourselin, & Cardoso, 2017) for association, negative log-likelihood for polarity determination, and smooth L1 loss for regression tasks. To handle data imbalance, additional weighting is applied to regression targets with uneven distributions, such as depth and magnitude. Rarer target values, including larger magnitudes, are assigned higher weights in the corresponding loss terms. We also apply masks to polarity and regression losses so that only relevant event-related time windows contribute to optimization. Because earthquake catalogs are incomplete and some true arrivals may be unlabeled, we avoid treating all unlabeled samples as negative examples. For phase picking and event-arrival detection, masked loss regions are used to reduce the penalty on potentially unlabeled arrivals. We also randomly sample points from noise windows to balance the ratio of positive and negative samples during training.

We employ the AdamW optimizer (Loshchilov & Hutter, 2019) with a base learning rate of 0.001, a weight decay of 0.05, and a batch size of 40, with a cosine decay learning rate scheduler and a one-epoch warm-up. Training the model for 33 epochs takes approximately 24 hours using 4 NVIDIA A100 GPUs, depending on data-loading efficiency.

We use two data augmentation strategies. The first is temporal event stacking, which is widely used in single-station model training (Zhu, Mousavi, & Beroza, 2020). We adapt the network-level stacking strategy used in PhaseNO (Sun et al., 2023). For a given primary event, we randomly sample a secondary event sharing at least three stations, scale its waveforms by a random factor to simulate magnitude differences, adjust the magnitude label accordingly, apply a random time shift, and stack the scaled waveforms onto the primary event at shared stations. This simulates temporally dense event distributions common in earthquake sequences. The second strategy is spatial event concatenation, designed to simulate scenarios where signals are recorded by a subset of the network. For a given primary event, we randomly select a secondary event of similar magnitude and sample a random subset of stations from both. The combined data are concatenated along the station dimension, creating inputs that encourage the Association Head to distinguish nearby but distinct events.

Supplementary Note 4: More metrics on the NCEDC 2023 dataset

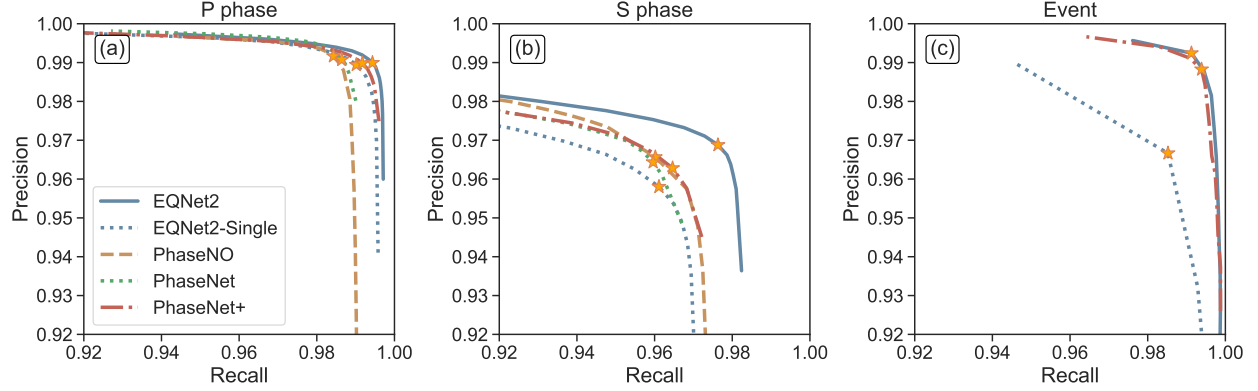
We evaluate picking and detection performance using precision, recall, and F1 score. Precision measures the fraction of predicted positives that are correct, whereas recall measures the fraction of reference positives that are recovered:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (4)$$

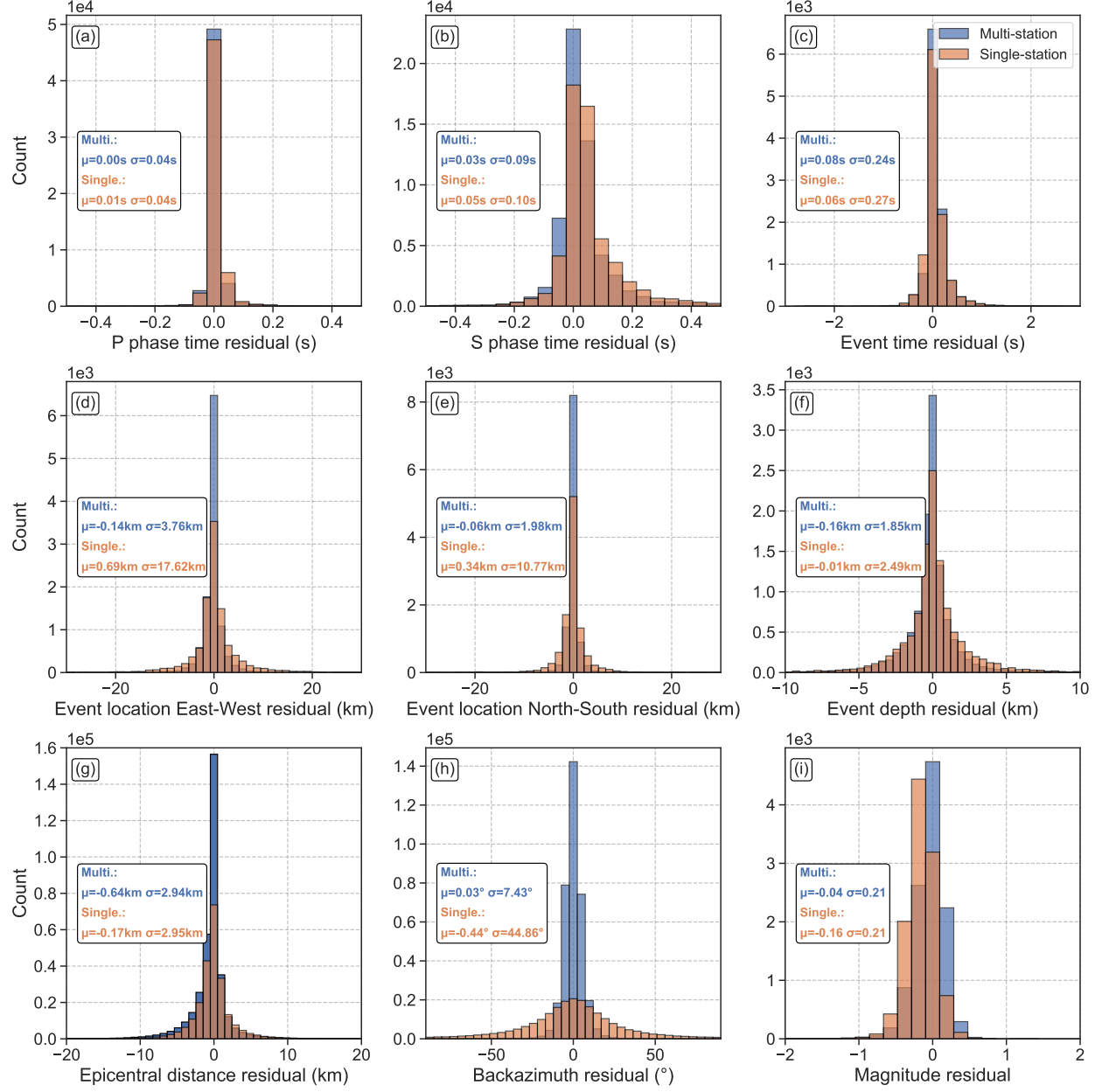
The F1 score is the harmonic mean of precision and recall:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5)$$

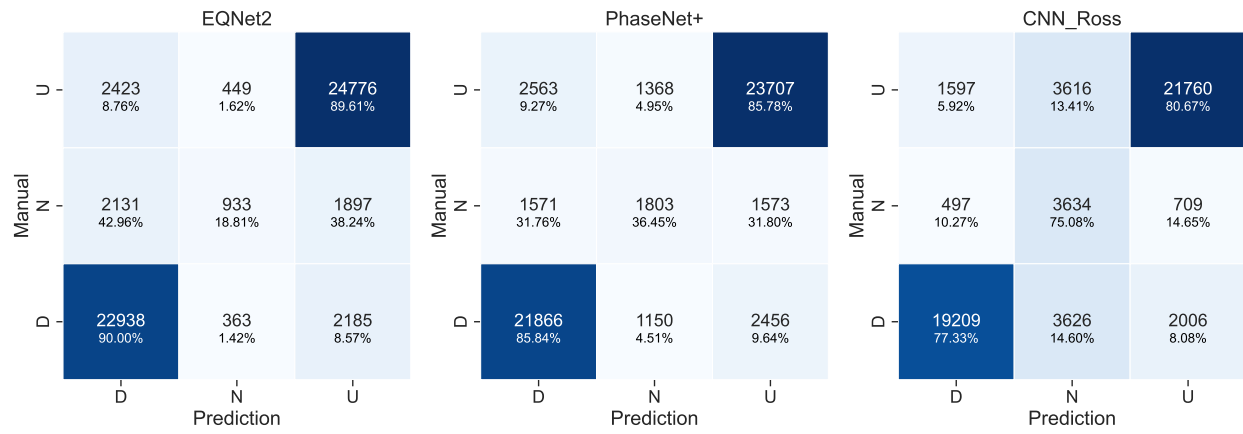
For phase picking, a predicted phase arrival is counted as a true positive (TP) if it is matched to a reference arrival of the same phase type with an absolute time residual smaller than 0.5 s. For event detection, a predicted event is counted as a TP if the absolute origin-time residual is smaller than 3 s. Unmatched predictions are counted as false positives (FP), and unmatched reference arrivals or events are counted as false negatives (FN).



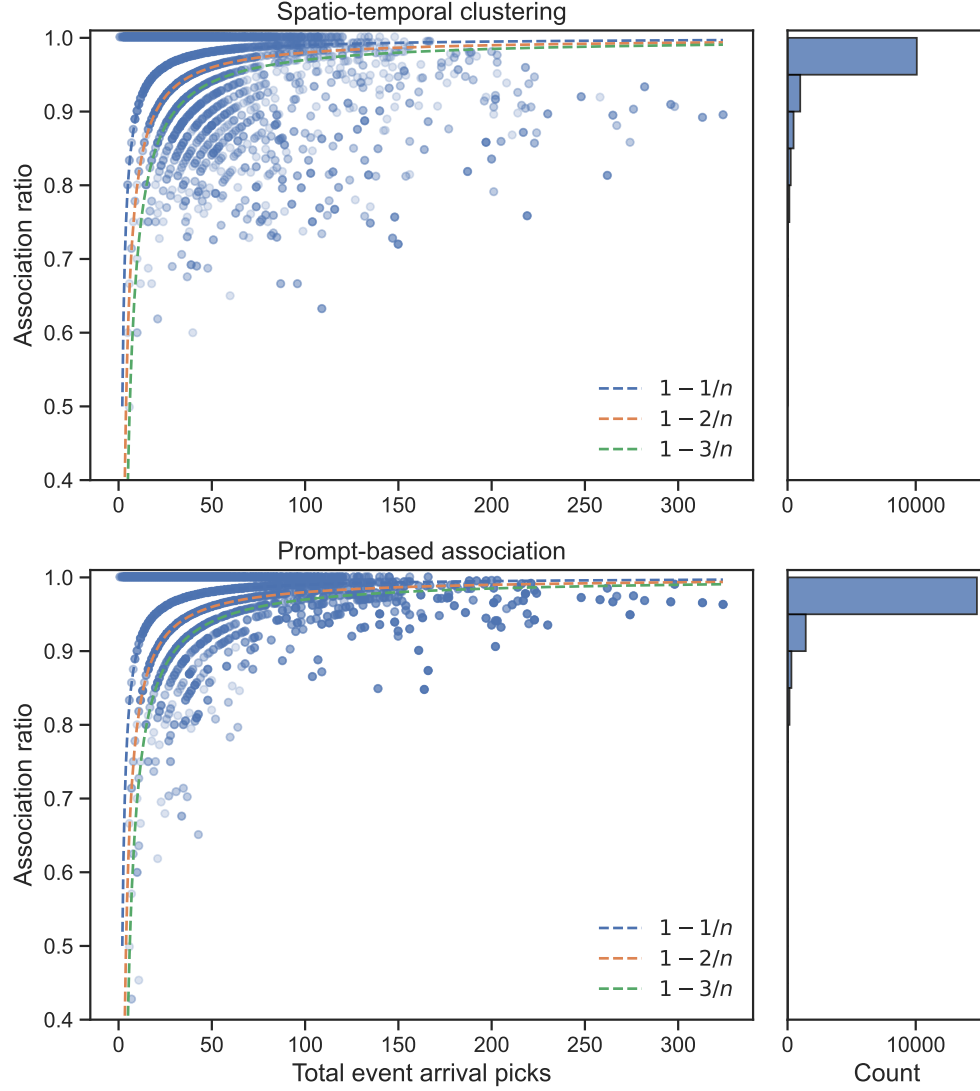
Supplementary Fig. 2: Precision–recall curves of phase picking and event detection on the NCEDC 2023 dataset. Yellow stars mark the highest F1 score for each model.



Supplementary Fig. 3: Phase picking and event detection residual distributions of EQNet2 and its single-station variant on the NCEDC 2023 dataset.



Supplementary Fig. 4: Confusion matrix for P-phase first-motion polarity determination on the NCEDC 2023 dataset. The three classes, D, N, and U, denote Down, Unknown, and Up polarities, respectively.



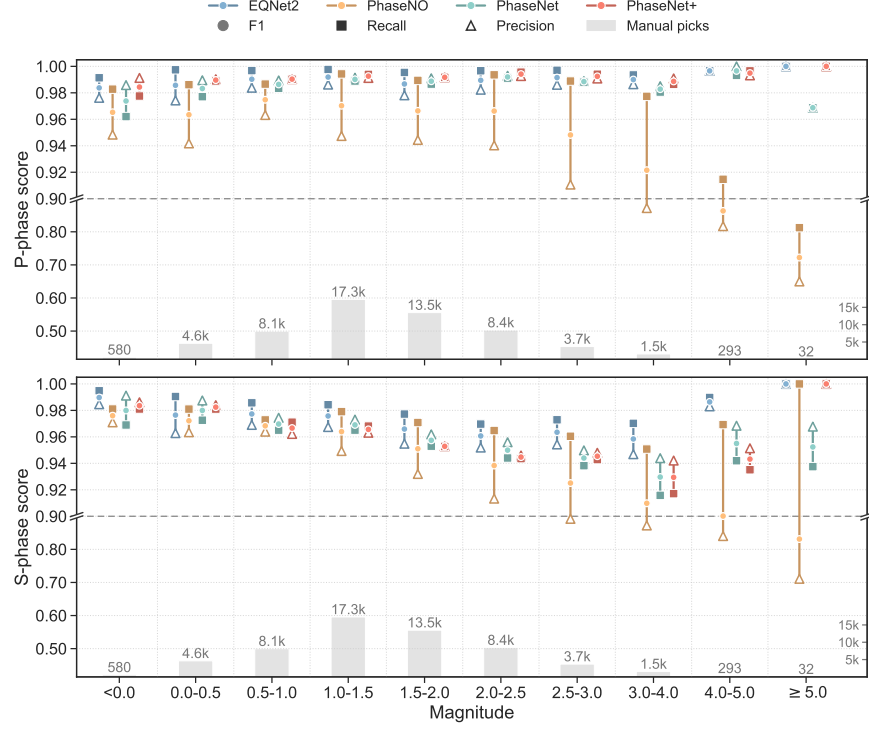
Supplementary Fig. 5: Window-wise association ratio on the NCEDC 2023 dataset by spatiotemporal clustering and prompt-based method. The association ratios are expected to concentrate on curves whose equations follow $1 - \frac{k}{n}$, where n is the total number of signals and k is the number of unassociated ones.

Supplementary Table 1: Phase-picking metrics and time-residual statistics at a score threshold of 0.3.

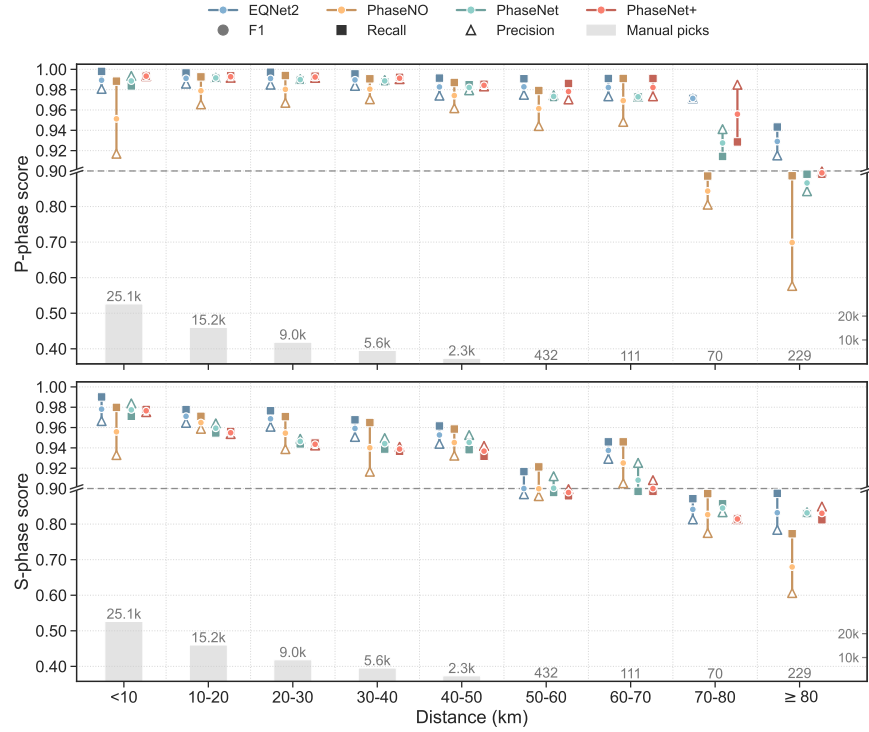
		Precision	Recall	F1	Mean (s)	Std (s)
EQNet2	P-phase	0.982	0.997	0.990	0.004	0.038
	S-phase	0.961	0.980	0.971	0.027	0.088
EQNet2-Single	P-phase	0.979	0.995	0.987	0.007	0.039
	S-phase	0.949	0.967	0.958	0.049	0.101
PhaseNO	P-phase	0.942	0.990	0.965	0.003	0.050
	S-phase	0.937	0.972	0.954	0.010	0.100
PhaseNet	P-phase	0.991	0.986	0.989	−0.001	0.039
	S-phase	0.967	0.957	0.962	0.016	0.103
PhaseNet+	P-phase	0.988	0.993	0.991	−0.001	0.038
	S-phase	0.967	0.959	0.963	0.011	0.100

Supplementary Table 2: Event detection residual statistics at threshold 0.3.

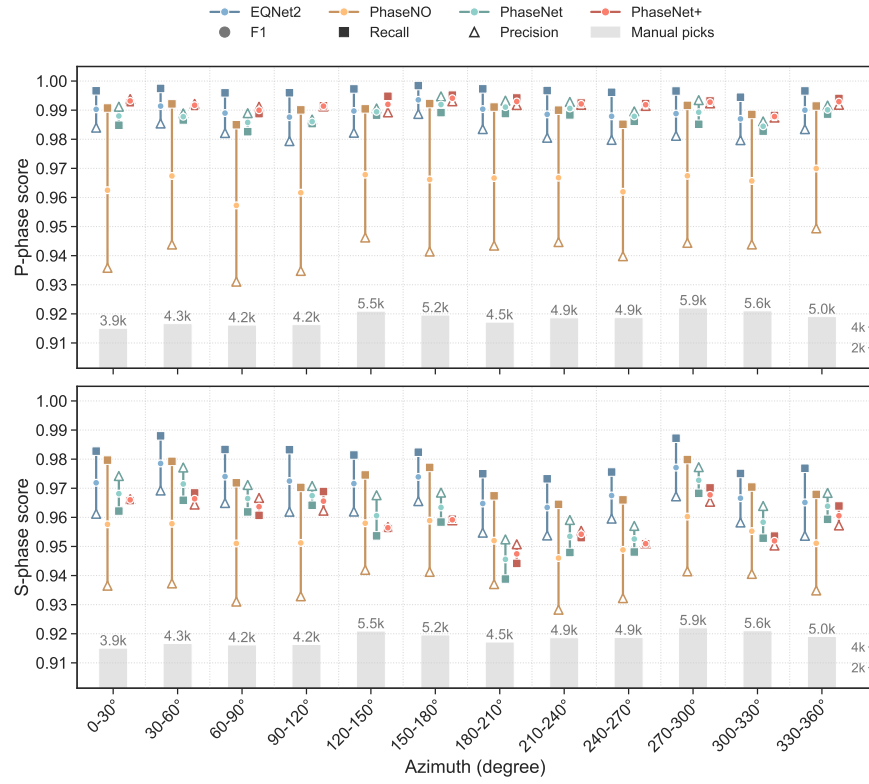
Model	Detection Metrics			Origin Time (s)		Distance (km)		Backazimuth (deg)	
	Precision	Recall	F1	Mean	Std	Mean	Std	Mean	Std
EQNet2	0.911	0.999	0.953	0.075	0.243	−0.638	2.941	0.027	7.428
EQNet2-Single	0.726	0.997	0.840	0.062	0.278	−0.174	2.954	−0.443	44.863
PhaseNet+	0.926	0.999	0.961	0.004	0.264	—	—	—	—



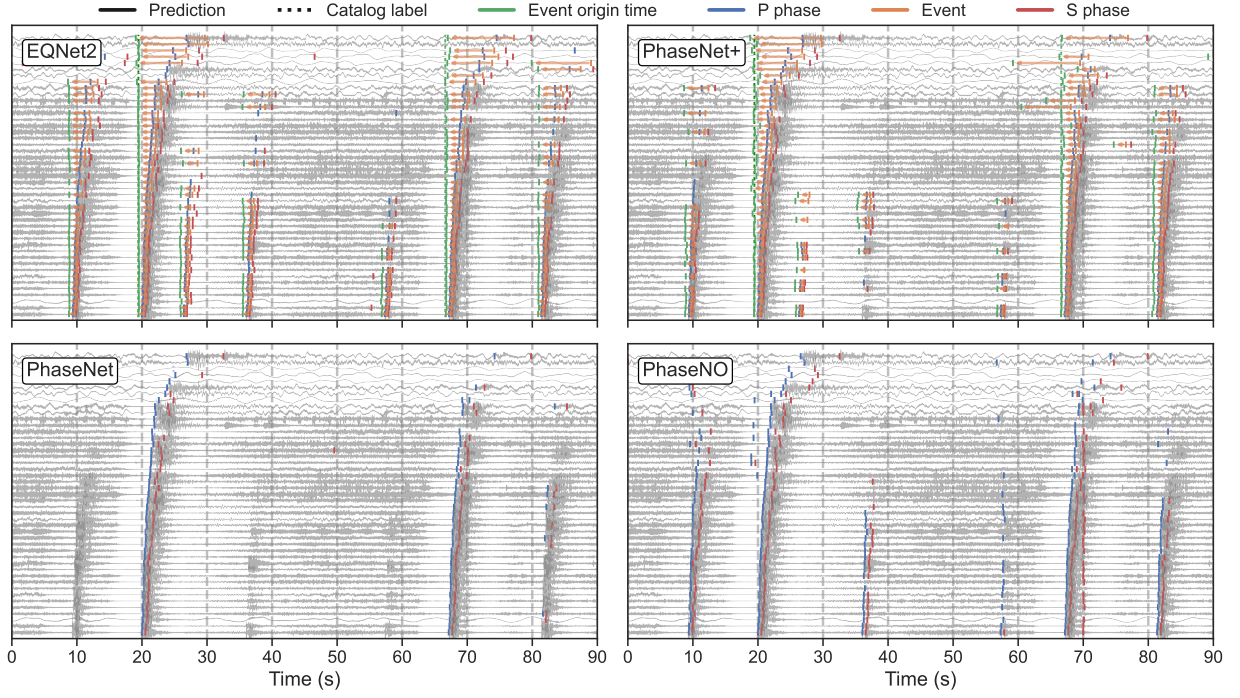
Supplementary Fig. 6: Phase picking F1, precision and recall scores for P- and S-phase arrival picking on the NCEDC 2023 dataset across magnitude groups. Symbols are the same as Figure 3 in the main text.



Supplementary Fig. 7: Phase picking F1, precision and recall scores for P- and S-phase arrival picking on the NCEDC 2023 dataset across epicentral-distance groups. Symbols are the same as Figure 3 in the main text.



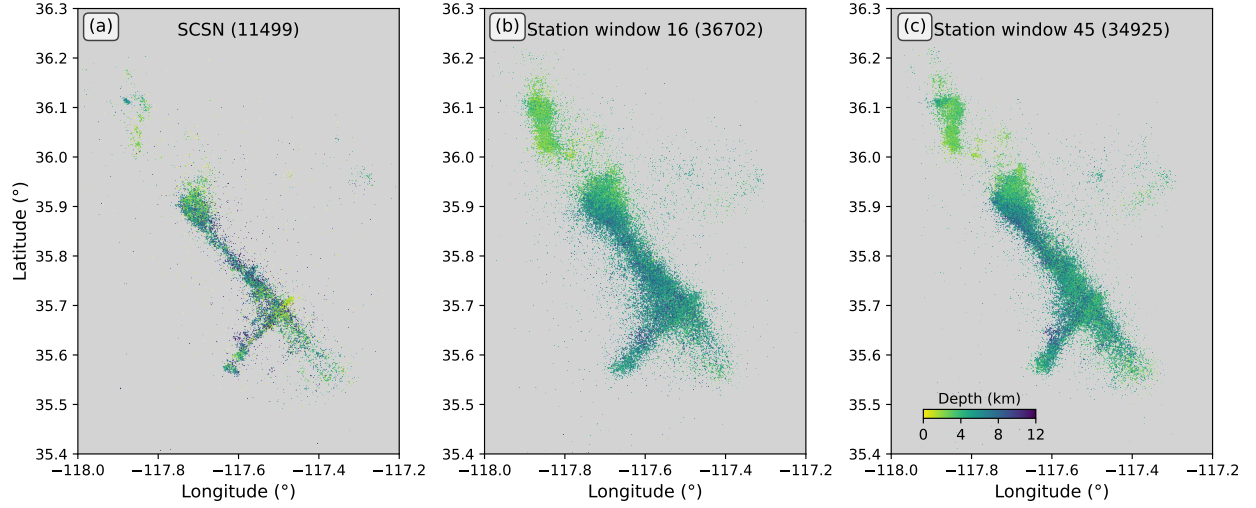
Supplementary Fig. 8: Phase picking F1, precision and recall scores for P- and S-phase arrival picking on the NCEDC 2023 dataset across azimuth groups. Symbols are the same as Figure 3 in the main text.



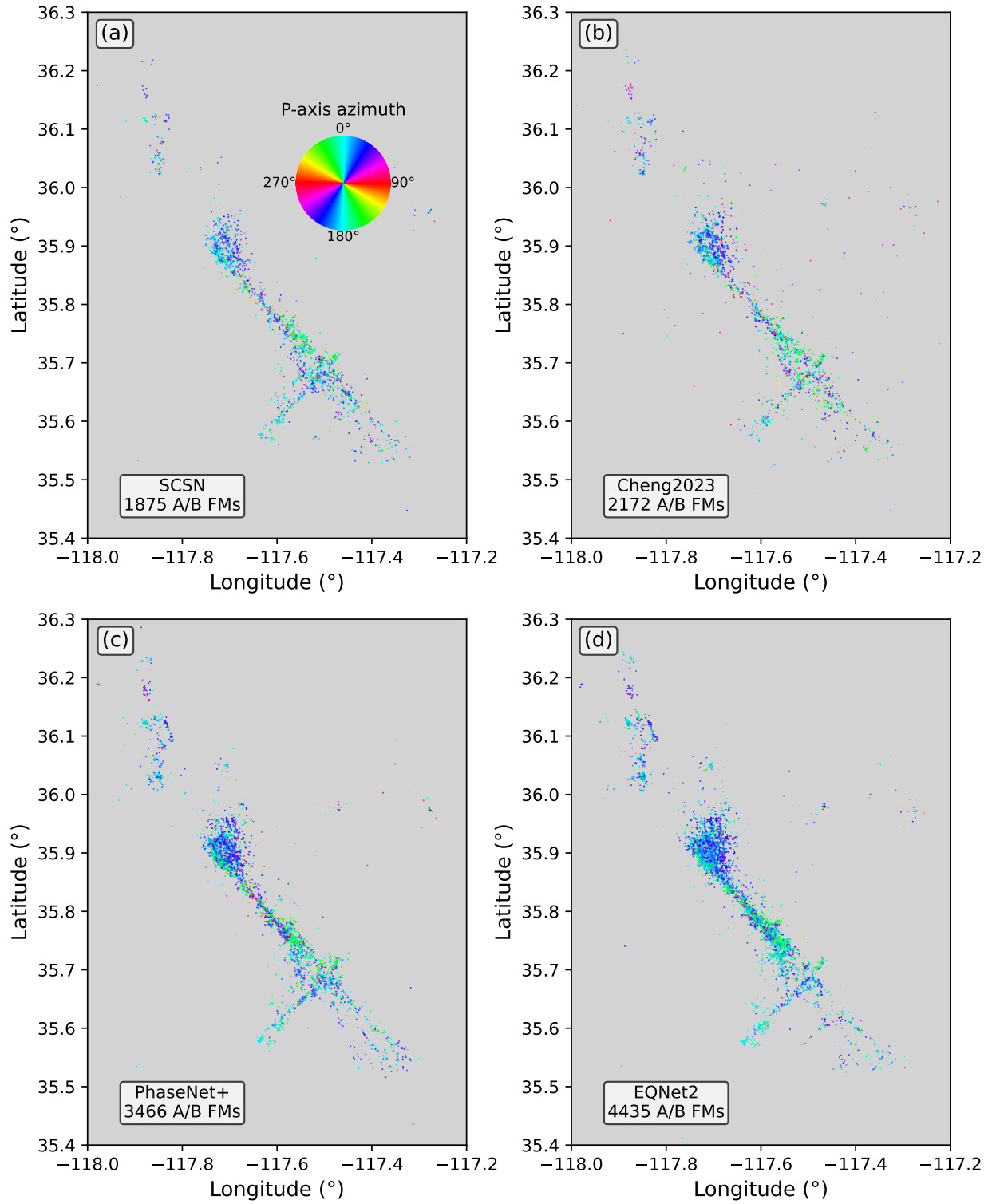
Supplementary Fig. 9: A dense-event window from the NCEDC 2023 test dataset. Predicted P- and S-phase arrivals from EQNet2, PhaseNet+, PhaseNet, and PhaseNO are shown for a 90-s window containing two cataloged events, nc73848531 and nc73848536. Solid lines denote predictions and dotted lines denote catalog annotations; blue and red indicate P- and S-phase arrivals, green indicates origin-time estimates, and orange indicates event-arrival proxies. Event-level predictions are available only for EQNet2 and PhaseNet+.

Supplementary Note 5: Additional results for the 2019 Ridgecrest earthquake sequence

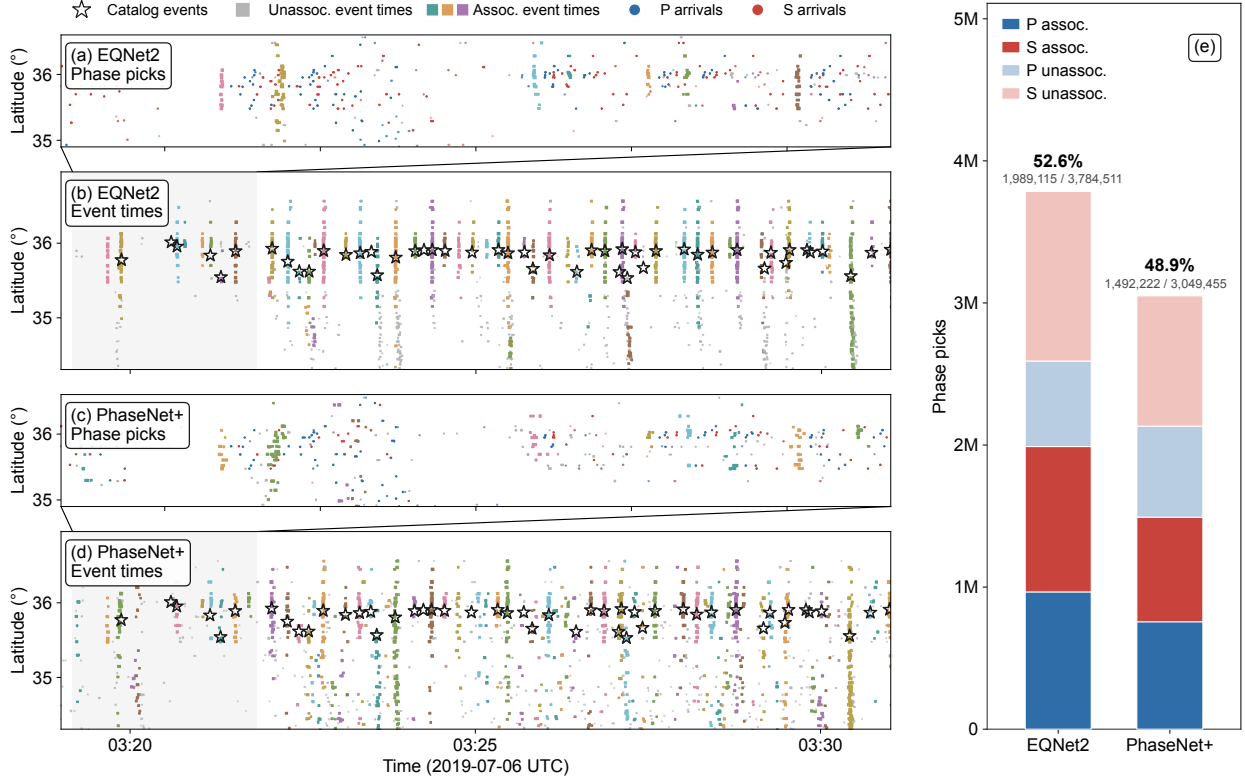
This note provides additional Ridgecrest results: preliminary locations from the model outputs before absolute location, focal-mechanism comparisons, and a pick-level comparison around the M_w 7.1 mainshock.



Supplementary Fig. 10: Epicenter locations predicted by EQNet2 with different station window sizes before the absolute location procedure: (a) SCSN (reference), (b) 16, (c) 45 (all sensors).



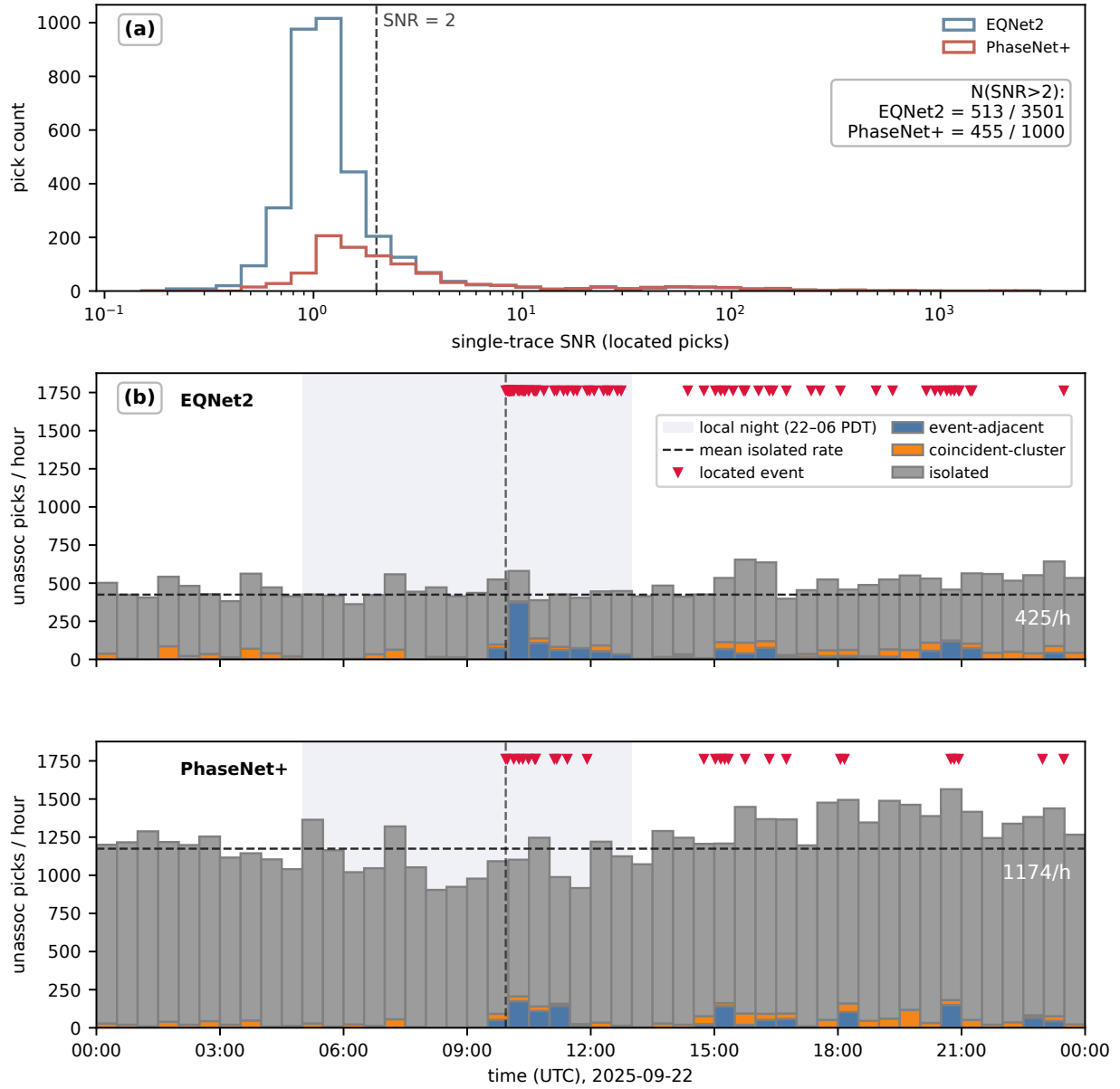
Supplementary Fig. 11: Spatial distribution of P-axis azimuths from focal mechanism solutions: (a) SCSN catalog, (b) catalog of Cheng et al. (2023), (c) PhaseNet+ catalog, and (d) EQNet2 catalog.



Supplementary Fig. 12: Comparison of EQNet2 and PhaseNet+ for the 2019 Ridgecrest earthquake sequence. (a–d), Predictions in a time window selected from the 2019 Ridgecrest sequence around the M_w 7.1 mainshock (event ID: ci38457511). Each square in panels b and d denotes a predicted event origin time at the corresponding station. Gray squares are unassociated predictions, colored squares indicate predictions associated with the same event, and the black star marks the SCSN catalog event. Panels a and c enlarge the shaded interval and additionally show predicted P- and S-phase arrivals. (e) Stacked counts of associated and unassociated P- and S-phase picks. Percentages indicate the fraction of raw picks retained after association and location.

Supplementary Note 6: Examination of Berkeley earthquake detections

Throughout, we refer to picks output by the models as *raw*, picks kept by the associator as *associated*, and picks used in the final location as *located*, and we verify the events each model locates. Among the final located picks, the two models pick a similar number of strong arrivals above an SNR of 2 (513 for EQNet2 and 455 for PhaseNet+), while EQNet2 locates roughly 2,500 more low-SNR arrivals; each coincides with a median of 40 same-model picks at nearby stations (± 2 s) and is thus supported by network coherence rather than single-trace amplitude (Supplementary Fig. 13a). Among the unassociated picks, both models have a similar, small number that are adjacent to a located event or coincident across stations, but the isolated remainder runs at 425 per hour for EQNet2 against 1,174 for PhaseNet+ and varies little between day and night (Supplementary Fig. 13b), marking it as a steady detector background unrelated to the sequence. Raising the pick-score threshold clears some of this background but barely changes the gap in association ratio. EQNet2’s located picks therefore rest on coherent waveform evidence, whereas most of PhaseNet+’s larger raw output consists of isolated detections that never form events.

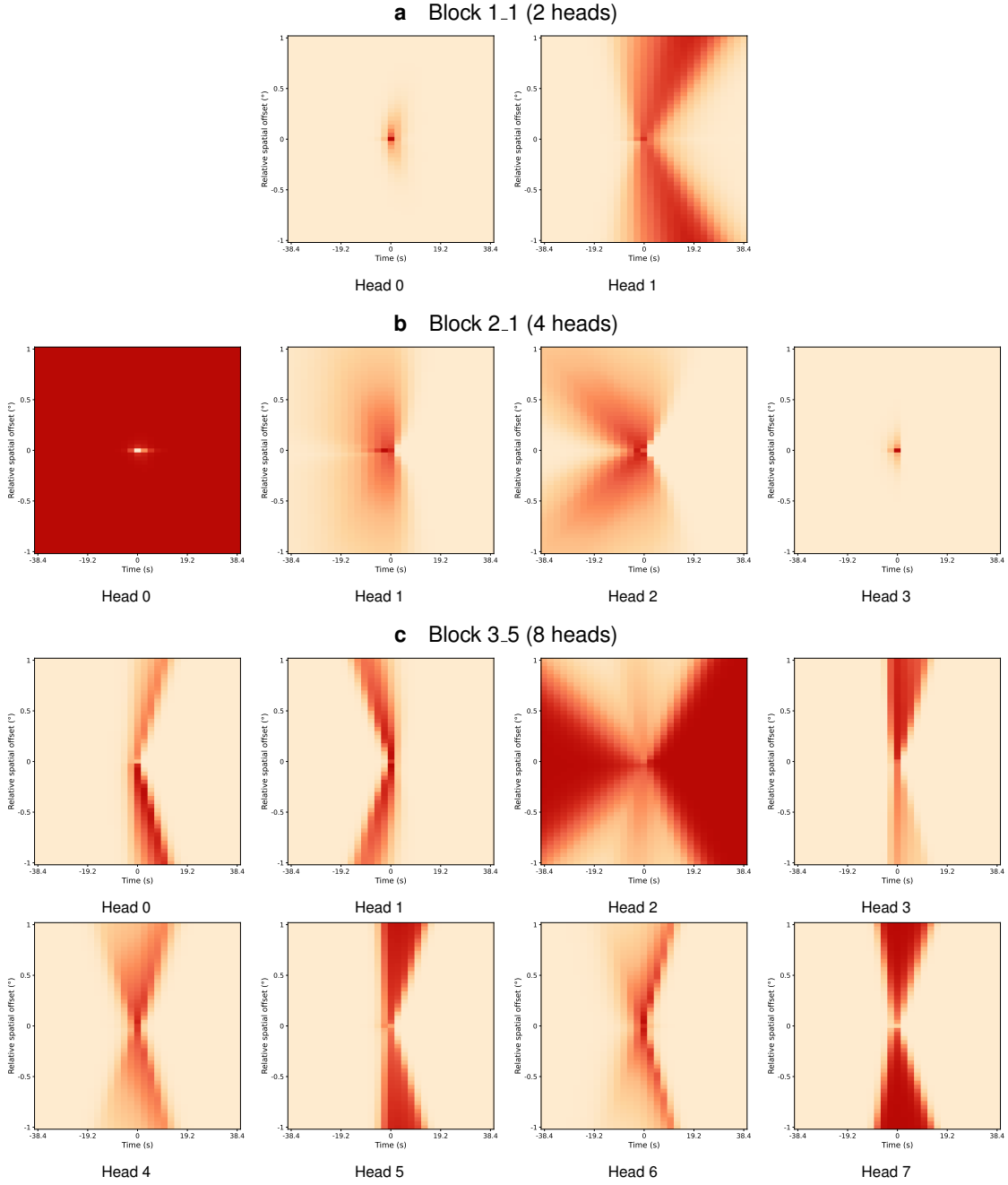


Supplementary Fig. 13: Pick-level evidence for the Berkeley one-day comparison. (a) Counts of located picks as a function of single-trace signal-to-noise ratio for EQNet2 and PhaseNet+. (b) Unassociated picks through the day, separated into picks adjacent to a located event, picks coincident across stations, and isolated picks.

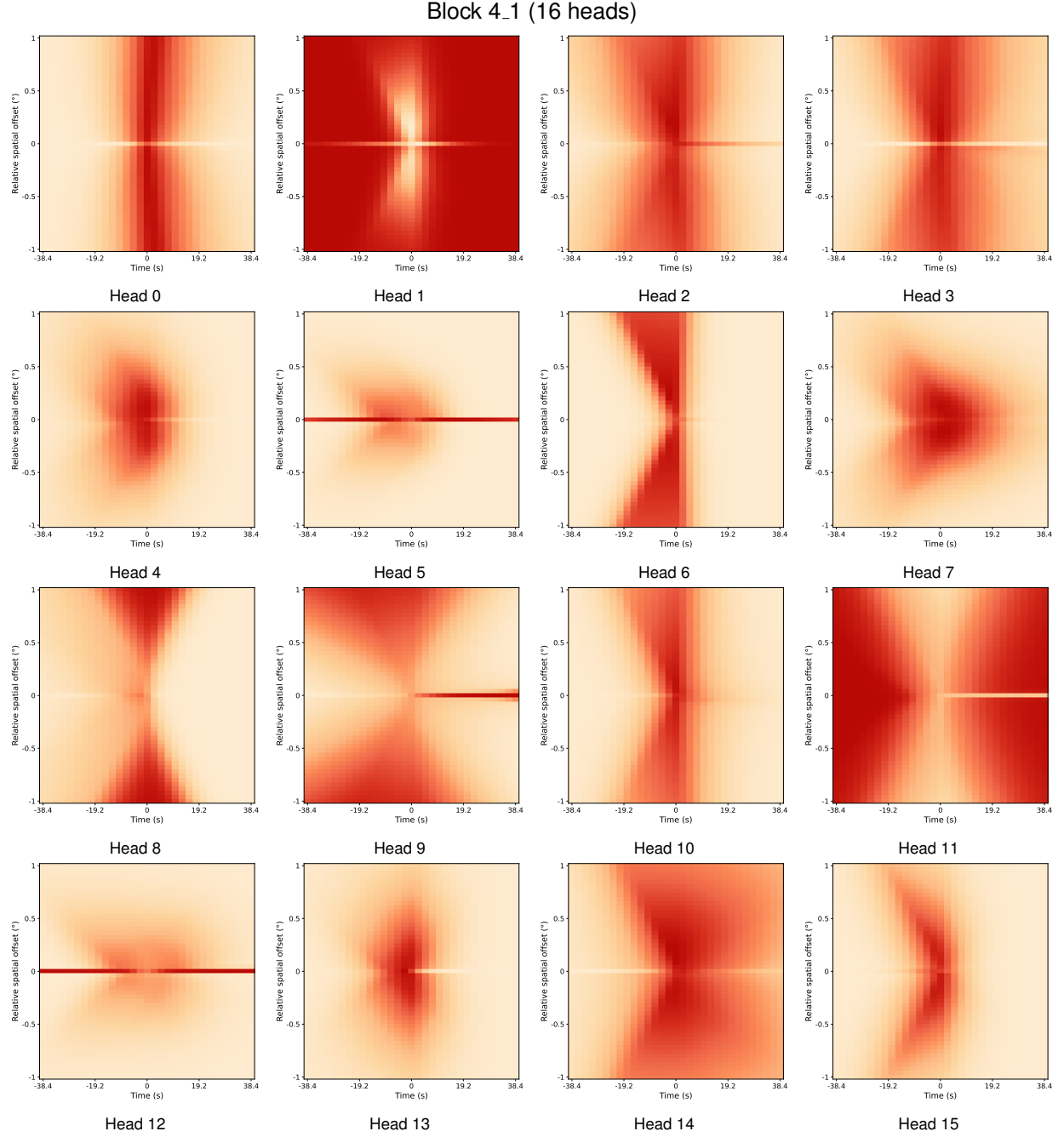
Supplementary Note 7: Additional visualizations of learned RPBs and attention maps

To examine how cross-station relationships are represented in the backbone, we visualize the learned relative position bias (RPB) for every attention head in the four backbone stages (Supplementary Figs. 14 and 15). The heads exhibit distinct patterns, including spatially localized, temporally broad, and slanted structures. Some of the slanted structures are consistent with moveout-like time–distance relationships, but the pattern is not shared by all heads. The bias ranges fall into two approximate groups, 0–8 and 0–16. Each head is therefore normalized independently for visualization.

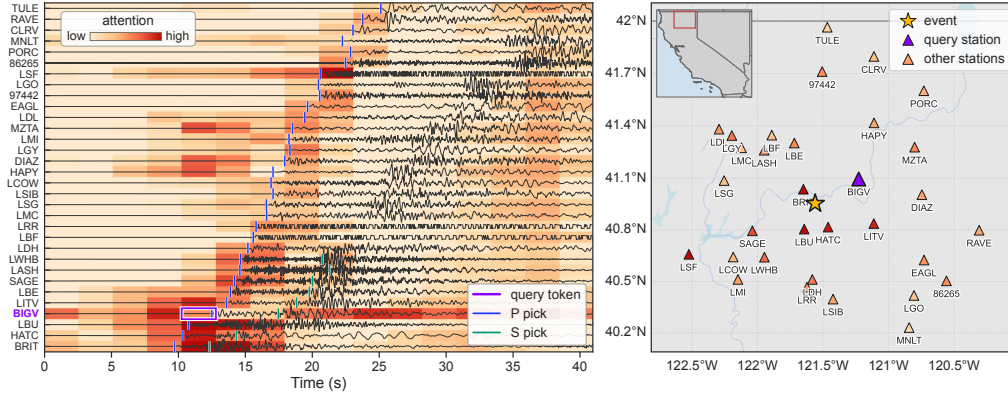
We additionally compare attention maps from the deepest backbone stage for three query stations located in different parts of the network (Supplementary Fig. 16). All three maps are from the same event and are intended to illustrate how attention varies with query-station geometry rather than to evaluate generalization across events.



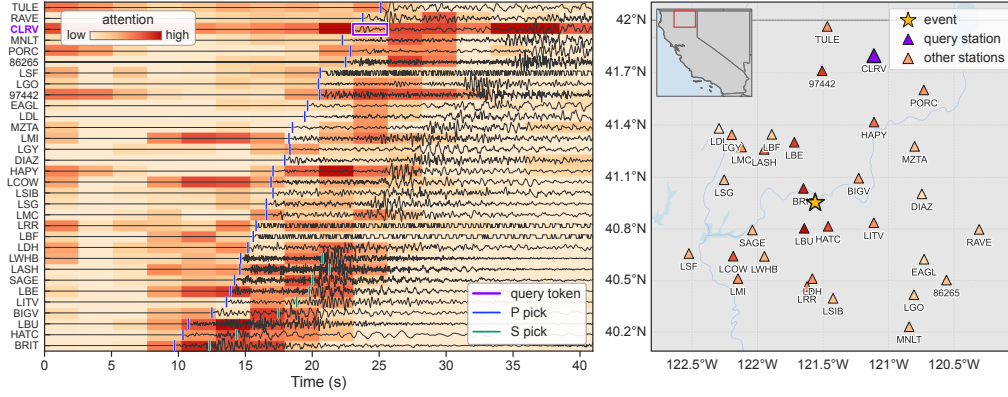
Supplementary Fig. 14: Relative position bias for all attention heads in the shallow and intermediate backbone stages. Panels **a–c** show blocks 1_1, 2_1, and 3.5, respectively. The horizontal axes denote relative time, the vertical axes denote relative spatial offset, and darker colors indicate larger bias within each head. Because the bias ranges differ among heads, with approximate ranges of 0–8 or 0–16, every panel is normalized independently.



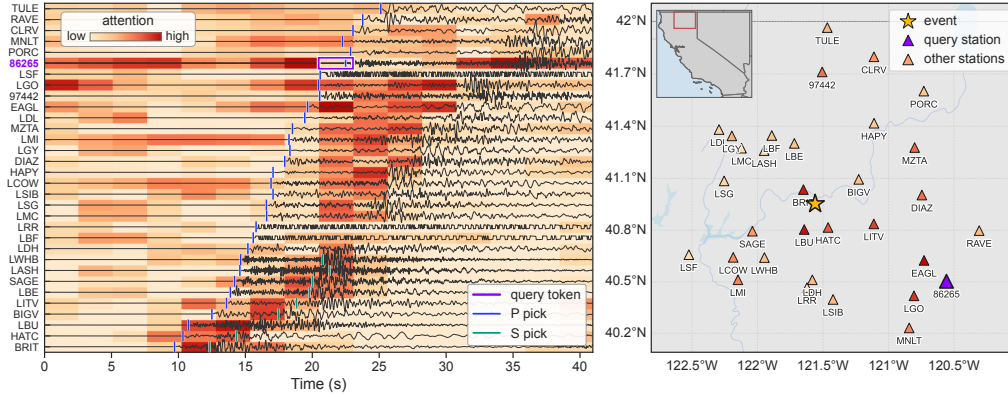
Supplementary Fig. 15: Relative position bias for all 16 attention heads in the deepest backbone stage (block 4.1). The horizontal axes denote relative time, the vertical axes denote relative spatial offset, and darker colors indicate larger bias within each head. The diversity of patterns suggests specialization among attention heads. Because the bias ranges differ among heads, with approximate ranges of 0–8 or 0–16, every panel is normalized independently.



(a) Query station BK.BIGV.00.HH.



(b) Query station BK.CLRV.00.HH.



(c) Query station CE.86265.10.HN.

Supplementary Fig. 16: Additional attention maps for event nc73934641 from the deepest Swin Transformer block. The three query stations were selected to represent different positions within and near the edge of the station network. The token containing the query P arrival is enclosed by a purple rectangle, the query station is indicated by a purple triangle, and darker colors indicate higher attention weights.

Supplementary Note 8: Computational cost evaluation

GFLOPs of models are calculated for inputs comprising 16 stations, three components, and 4096 time samples. The time and memory costs are tested using an NVIDIA RTX 6000 Ada GPU with data from the NCEDC 2023 dataset. The sublinear scaling of runtime with the number of traces for PhaseNet+ and EQNet2 may reflect code execution overhead. Association ratios are calculated for the Berkeley earthquake case with different EQNet2 model sizes and station-window sizes for attention. The association ratio is defined as the number of picks after association and location divided by the number of raw picks.

Supplementary Table 3: Computational cost and performance scaling statistics. EQNet2-Base corresponds to the default EQNet2 model used in the main-text experiments.

Category	Metric	PhaseNet+	EQNet2-Tiny	EQNet2-Small	EQNet2-Base	PhaseNO
Model size	GFLOPs	2.72	3.49	10.69	37.75	167.23
	Params (M)	1.55	1.02	4.02	15.97	2.37
Runtime	Time (ms), 16 sta.	3.7	24.4	24.9	31.9	167.2
	VRAM (GB), 16 sta.	0.14	0.25	0.27	0.42	6.88
	Time (ms), 32 sta.	5.1	26.9	29.3	54.9	685.3
	VRAM (GB), 32 sta.	0.27	0.49	0.53	0.76	27.25
P picking	Precision	0.988	0.981	0.982	0.982	0.942
	Recall	0.993	0.996	0.996	0.997	0.990
	F1	0.991	0.988	0.989	0.989	0.965
S picking	Precision	0.967	0.957	0.960	0.961	0.937
	Recall	0.959	0.976	0.979	0.981	0.972
	F1	0.963	0.966	0.970	0.971	0.954

Supplementary Table 4: Catalog-level association ratio (%) in the Berkeley earthquake case for different EQNet2 model sizes and station-window sizes for attention. Each cell reports the association ratio, followed by the number of located picks over the number of raw picks (located/raw). The 69-station configuration uses the full station set.

Station window size	EQNet2-Tiny	EQNet2-Small	EQNet2-Base
18	6.06%	7.68%	9.33%
	2,801 / 46,206	2,967 / 38,625	2,683 / 28,762
36	8.66%	11.24%	12.73%
	2,976 / 34,361	3,139 / 27,921	3,020 / 23,727
69	13.06%	17.60%	22.06%
	3,417 / 26,171	3,838 / 21,803	3,501 / 15,871