PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

ML-aided induced seismicity processing and interpretation for Geothermal Field Monitoring

Nori Nakata[12], Zhengfa Bi[1], Hongrui Qiu[3], Cheng-Nan Liu[4], Rie Nakata[1]

[1]Lawrence Berkeley National Laboratory, [2]Massachusetts Institute of Technology, [3]China University of Geosciences, [4]University of Utah

nnakata@lbl.gov

Keywords: Machine learning, induced seismicity, monitoring

ABSTRACT

The increasing demand for sustainable energy has intensified interest in geothermal energy, with enhanced geothermal systems (EGS) emerging as a key solution. Hydraulic fracturing, a critical technique in EGS, often induces microseismic events that provide valuable insights into subsurface fracture mechanisms and reservoir properties. This study examines the promising and evolving role of machine learning (ML) in analyzing induced seismicity in geothermal fields, focusing on applications including data acquisition, processing, and interpretation. Specifically, we present four key ML applications: wavefield inpainting for reconstructing missing seismic data, generative artificial intelligence (Gen-AI) for wavefield simulation, transformer-based phase picking for detecting small seismic events, and clustering methods to analyze the spatio-temporal evolution of seismicity. These approaches enhance the efficiency and accuracy of seismic monitoring, offering innovative tools for optimizing geothermal energy extraction. Additionally, the study highlights the importance of pre-processing and data augmentation in addressing challenges such as data scarcity and noise in seismic datasets. Techniques like spectral analysis, attenuation correction, noise injection, and the use of multiple time windows are key to obtaining reasonable results from ML models with limited amounts of data. This research provides insights into ML-aided induced seismicity processing and interpretation for seismic monitoring and reservoir management in geothermal fields, contributing to more efficient and sustainable energy operations while mitigating large induced seismic events.

1. INTRODUCTION

Geothermal field development modifies subsurface stress and strain conditions. This modification becomes even more severe with the increasing development of enhanced geothermal systems (EGS), where we apply hydraulic fracturing. The hydraulic fracturing can enhance permeability by altering subsurface stress conditions and induce seismicity (Majer et al., 2007; Zhang et al., 2023). Analyzing these seismic events provides valuable insights into fracture mechanisms and reservoir properties, which are essential for optimizing energy extraction and mitigating environmental risks (Zoback & Kohli, 2019).

Microseismic monitoring has become an essential component of geothermal operations. The spatial distribution of microseismic events during stimulation and production infers key properties such as future production potential, permeability, and allows to analyze the operation efficiency (Maxwell, 2014; Niemz et al., 2024). The waveforms of microseismicity contain additional features, such as focal mechanisms, which enable the differentiation between tensile and shear fractures. They also offer insights into the characteristics of fracture networks, such as their orientation and connectivity, which are critical for understanding reservoir behavior (Eisner et al., 2010; Warpinski et al., 2012).

Recent advancements in machine learning (ML) have been transforming the field of seismology, enabling the extraction of previously inaccessible information from vast datasets (Bergen et al., 2019; Kong et al., 2019; Xiong et al., 2021). In this study, we explore four innovative ML applications for analyzing induced seismicity in geothermal fields, covering key components of the entire process of data analysis, including recording, processing, and interpretation.

The first application is wavefield inpainting, which reconstructs missing seismic records during power outages, physical damage, or other disruptions. We can fill data gaps by learning patterns from earthquake records to create continuous seismic monitoring data. This approach is particularly beneficial for applications that rely on a well-sampled wavefield, such as microearthquake hypocenter relocation and full-waveform inversion, where spatially continuous data improve accuracy. The second application leverages generative artificial intelligence (Gen-AI) for wavefield simulation. Unlike traditional physics-based simulations, Gen-AI learns physical principles from data to create synthetic wavefields, which can then be used for tasks such as filling missing data, imaging seismic structures, and analyzing ground motion. The third application focuses on ML-based phase picking for detecting smaller seismic events, an essential task in geothermal fields. By developing a transformer-based network tailored to smaller-scale events, this method enhances the accuracy of phase detection. Finally, we delve into the interpretation of seismicity using clustering techniques. These methods provide insights into the spatio-temporal evolution of seismic events, enabling a more nuanced understanding of seismic behavior and its underlying mechanisms. Collectively, these advancements showcase the potential of ML to revolutionize the monitoring and interpretation of seismicity, offering more effective tools for managing and optimizing geothermal energy extraction.

2. FILLING DATA GAPS DUE TO MISSING STATIONS

2.1 Background and Motivations

Despite significant efforts to maintain seismometer networks, temporal gaps in recordings at certain stations are inevitable, and accessibility constraints often lead to spatial gaps in otherwise well-distributed arrays, such as using nodes (e.g., Li et al., 2018). To address this, we apply machine learning to reconstruct missing seismic data by leveraging algorithms similar to image inpainting, interpolation, or upscaling (Pathak et al., 2016). This approach enables us to fill data gaps within the array through interpolation or extend information beyond the array's perimeter via extrapolation, ultimately providing a more continuous and complete representation of seismic wavefields. We apply ML to fill the gaps based on existing records using algorithms similar to image inpainting/interpolation or upscaling (Pathak et al., 2016). This task involves both interpolation (filling in gaps within the data) and extrapolation (extending the data beyond measured points) when the missing station is at the edge of the seismic network to reconstruct a continuous and high-resolution representation of seismic wavefields.

Here, we propose an ML approach utilizing the Swin Transformer (Liang et al., 2021), a deep learning model architecture designed for hierarchical feature extraction and efficient long-range dependency modeling. The Swin Transformer employs a shifted window self-attention mechanism, which partitions the input into non-overlapping windows while allowing for inter-window interactions at each layer. This design enables the model to capture local and global waveform characteristics, making it well-suited for processing seismic wavefields with complex frequency content.

Unlike conventional convolutional neural networks (CNNs), which often rely on fixed receptive fields and struggle to model long-range dependencies efficiently, the Swin Transformer dynamically aggregates multi-scale information across different spatial and temporal resolutions. This hierarchical representation allows it to learn fine-scale waveform details while preserving broader structural patterns in the data. Given its success in image restoration, denoising, and super-resolution tasks, its application to time-series upscaling shows promise as a procedure for reconstructing high-resolution wavefields from sparse or incomplete data.

Additionally, although physics-based numerical methods rely on manual tuning of velocity models and adjustments to account for geological complexity, the Swin Transformer can learn waveform patterns directly from data. Furthermore, ML has demonstrated potential in handling large datasets efficiently, reducing the computational overhead. This ability enables the model to capture hidden features and complexities, offering a more flexible and adaptive solution as well as nearly real-time applications.

2.2 Data, Methods, and Results

The proposed methodology focuses on training the Swin Transformer for seismic waveform interpolation and extrapolation using data collected from a dense seismic array at the Cape Modern geothermal site in Milford, Utah (Nakata et al., 2024a, 2025). The array comprises 16 stations arranged in a 4-by-4 grid within a 60 × 60 m² area (inset in Figure 1). The model is trained solely on local earthquake data with magnitudes ranging from −1 to 1 and waveforms from 15 stations within the array to ensure robust validation. At the same time, it is tested on an excluded station and on events that were not part of the training set. The training data comprise waveforms from local seismic events captured by the dense array and filtered to a frequency range of 1–20 Hz. This frequency range is particularly suited for focusing on smaller-magnitude local earthquakes, where capturing fine waveform details is essential. Each waveform is divided into 4-second windows to standardize the input length for the model. Our algorithm does not normalize the wavefield amplitudes; hence, the transformer can learn both relative changes in waveform features and unique spatial site responses, thereby enhancing the model's ability to generalize across different seismic environments.

We train the model for over 10,000 iterations using a loss function designed to penalize discrepancies in both amplitude and phase between the predicted and actual waveforms. This dual focus ensures that the upscaled waveform maintains fidelity not only in terms of signal strength but also in the timing of seismic arrivals, which is a key requirement for subsequent analysis. We use both qualitative and quantitative measures to evaluate the model's performance. Qualitatively, upscaled waveforms are visually compared with the ground-truth seismograms for each test event. Quantitatively, the mean squared error (MSE) is calculated to assess the fidelity of the reconstructed waveforms, capturing differences in both amplitude and phase. The testing phase uses data from earthquakes that were not part of the training set, allowing us to assess the model's ability to upscale waveforms without prior knowledge of site-specific geology.

The results of our study demonstrate that the Swin Transformer is not only an efficient and adaptive solution for filling data gaps but also excels at reconstructing high-resolution waveforms from sparse data (Figure 2). The model shows a strong correlation between the reconstructed wavefield and ground-truth seismograms, achieving high accuracy. While linear interpolation is one approach for filling data gaps, the Swin Transformer yields a substantially lower MSE. Moreover, our findings suggest that the Swin Transformer can serve as a cost-effective alternative to deploying high-density seismic arrays, as it accurately predicts wavefields within seconds. This efficiency could reduce reliance on expensive and logistically challenging seismic deployments. For areas with limited access, such as rugged mountainous regions or remote sites, this approach significantly enhances the practicality of seismic monitoring and exploration efforts.

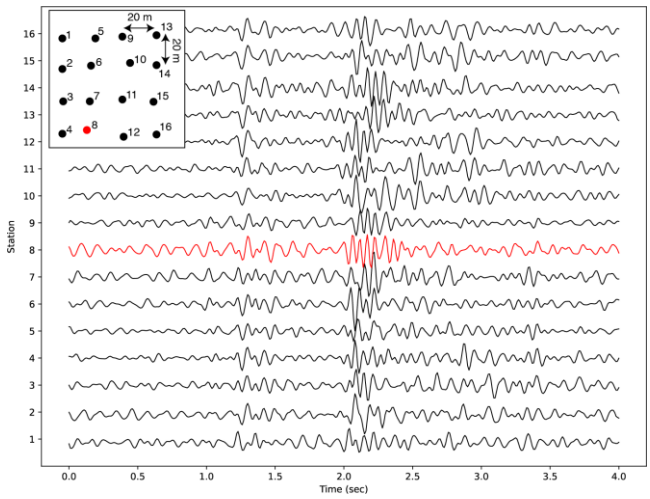

Figure 1. The processed seismograms from the array are displayed for the event of M1.06 that occurred on February 20, 2024. The waveform is bandpassed between 1-20 Hz. Each line represents the seismic waveforms recorded at each station, and the red line shows the test data (i.e., ground truth data). The inset shows the geometry of the receivers.

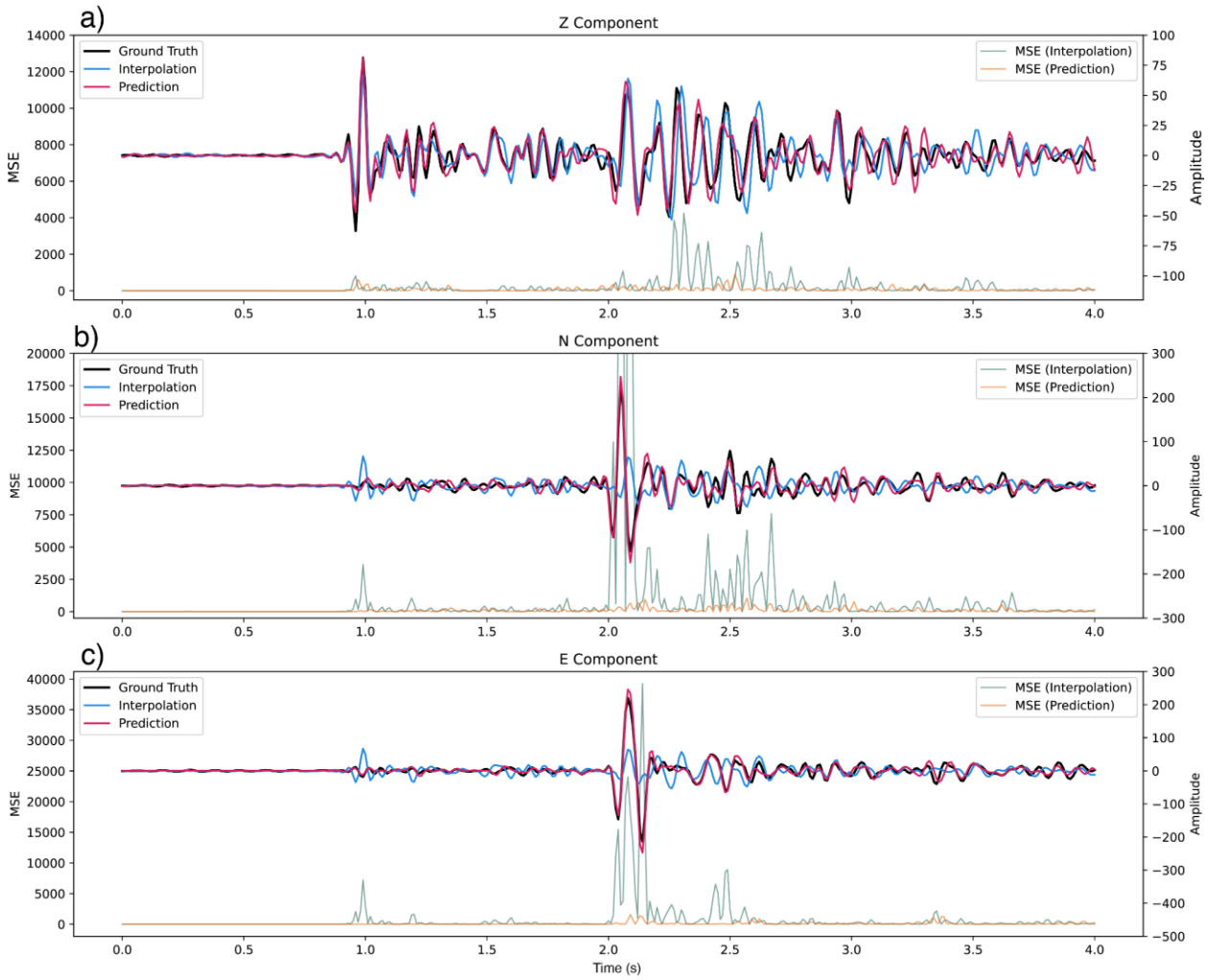

Figure 2. Generated waveforms using the Swin Transformer and linear interpolation compared to the ground truth data. Each panel illustrates the predictions of our model for each component: (a) vertical, (b) North-to-South, and (c) West-to-East. The ground truth data is represented in black, the Swin Transformer predictions in red, and the linear interpolation

predictions in blue. The MSE losses between the Swin Transformer and the ground truth are shown in orange, while the MSE losses for linear interpolation are represented in green.

3. WAVEFIELD SIMULATION USING GEN-AI

3.1 Conditional Generative Wavefield Modeling

The Swin Transformer–based algorithm presented above is currently used to fill data gaps using nearby records. In this section, we discuss a similar but more robust network to synthesize seismic wavefields using parameters such as source and receiver locations and source characteristics, which are analogous to those required for physics-based simulations. Instead of inputting velocity models, we adopt a data-driven approach, while allowing more flexible source and receiver geometries than in the example above to fill data gaps.

To model high-frequency waves efficiently without requiring detailed velocity models, we introduce Conditional Generative Modeling for wavefields (CGM-Wave; Bi et al., 2025), a cutting-edge generative modeling framework that builds upon our ground-motion modeling network, CGM-GM, which is based on variational autoencoders (VAE; Ren et al., 2024). Variational autoencoders have been widely applied in generative AI applications. CGM-Wave extends the CGM approach by integrating phase spectrum retrieval, enabling the model to capture both time-domain and frequency-domain characteristics of seismic events. Unlike its predecessor, CGM-GM, which directly compares waveforms, CGM-Wave employs a dual-domain reconstruction loss. In the time domain, the model compares generated and observed envelopes rather than raw waveforms, while in the frequency domain it compares amplitude spectra to match the spectral characteristics of real data. This strategy enables higher-resolution reconstruction of amplitude spectra, from which highly accurate phase information is subsequently retrieved using a convolutional neural network.

3.2 Wavefield Generation

We apply CGM-Wave to seismic data from The Geysers geothermal field. The Geysers geothermal field in northern California has been a major geothermal energy production site since the 1960s and is recognized as one of the world's largest and most productive geothermal energy resources. Decades of intensive exploitation have led to pronounced hydrothermal activity and a high rate of induced seismicity, with approximately 15,000 events recorded annually in the USGS catalog (Figure 3a). The occurrence of these events strongly correlates with steam production and fluid injection processes. While these processes drive energy production, they also pose challenges, including the potential for increased seismicity that could disrupt geothermal operations or affect nearby infrastructure. Addressing these risks requires advanced modeling techniques capable of capturing the complexity of the subsurface environment and the intricate interplay between natural and induced seismicity. In this context, generative modeling approaches are particularly promising, offering powerful tools to analyze and synthesize seismic data in challenging geothermal environments.

CGM-Wave is tested on a dataset of more than 30,000 seismic events recorded between 2020 and 2023 within The Geysers geothermal field, covering a 23 × 10 km area (Figure 3a). This dataset comprises nearly one million source–receiver pairs, providing high spatial coverage and data density, and thus represents an ideal test case for evaluating the robustness and applicability of the CGM framework. The model demonstrates strong performance in synthesizing seismic data, with generated waveforms closely matching real waveforms that were not used during training (Figure 3b). Although the overall waveform shapes are well reproduced, the amplitudes are slightly underestimated, likely because the network primarily focuses on phase reconstruction while preserving relative amplitude variations. Absolute amplitudes can be calibrated when a reference event is available to compute amplitude ratios between observed and generated waveforms. Time–frequency spectrograms further indicate that the generated wavefields exhibit energy distributions similar to those of the observed data.

These results highlight the effectiveness of CGM-Wave in handling complex subsurface heterogeneities and dense source–receiver geometries typical of geothermal regions. The ability to reliably synthesize seismic waveforms and spectral characteristics provides valuable tools for investigating induced seismicity and fracture mechanics within geothermal fields. By identifying subtle patterns in seismic data, CGM-Wave also shows promise for improving subsurface characterization in conjunction with conventional seismic imaging approaches.

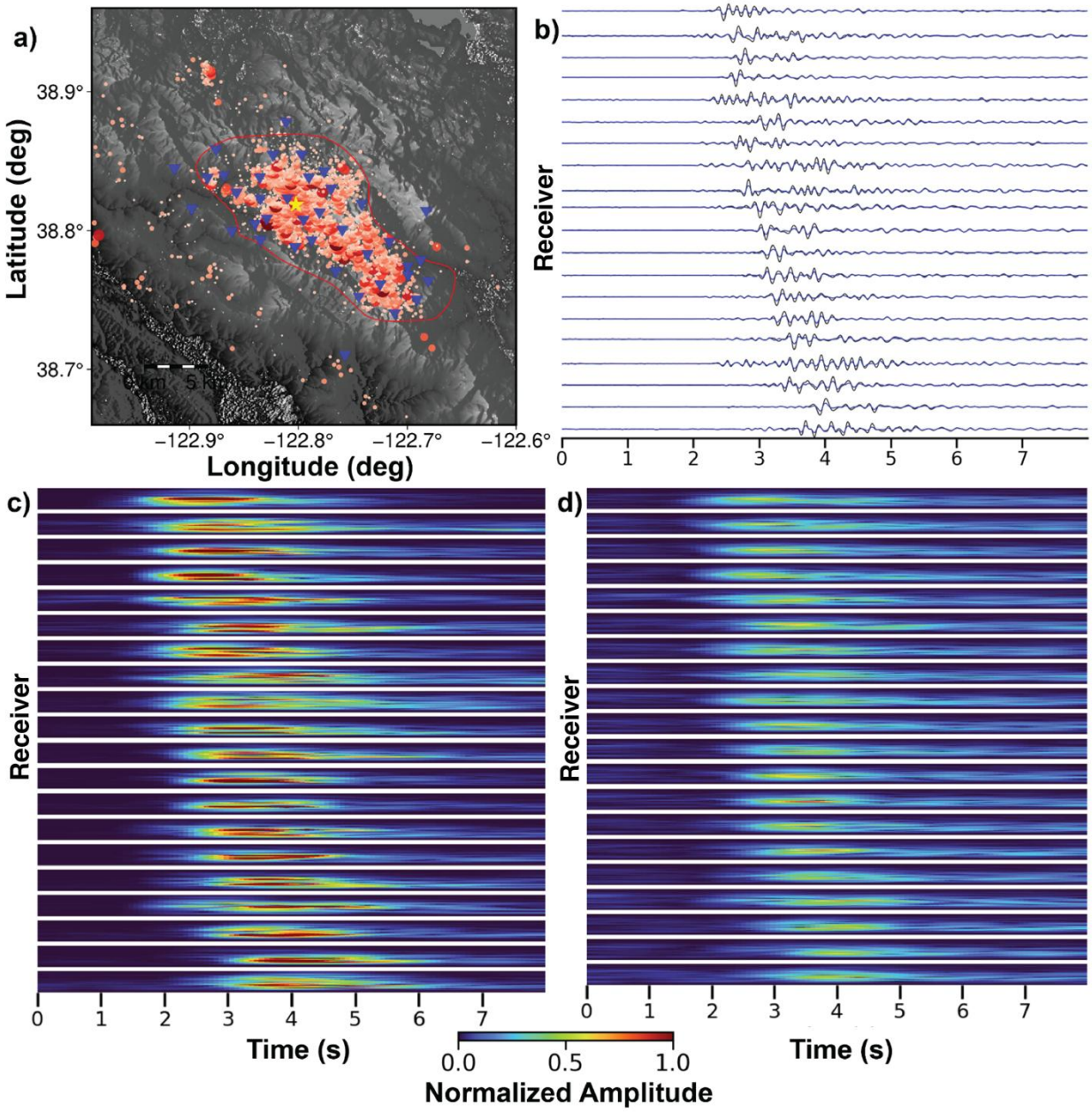

Figure 3. (a) Map of The Geysers geothermal field, with the red line outlining the reservoir. Blue triangles indicate seismic stations, while circles denote seismic events; circle color and size are proportional to event magnitude. A yellow star marks the seismic event used for validation. (b) Comparison of observed (black) and synthetic (blue) time-domain waveforms for a seismic event excluded from the training dataset, demonstrating model generalization. Each trace shows the recorded waveform at a station, aligned by distance from the earthquake. (c) Spectrograms of the observed waveforms for each trace, with frequency shown on the vertical axis in each subpanel. (d) Spectrograms of the synthetic waveforms generated by CGM-Wave.

4. PHASE PICKING FOR SMALL EARTHQUAKES

4.1 Motivation

Earthquake phase picking is a critical process in seismology, including geothermal studies, and serves as the foundation for accurate analysis of seismic events. Accordingly, many machine learning–based phase pickers have been developed and are now routinely used (e.g., Zhu et al., 2019; Mousavi et al., 2020; White et al., 2023). By identifying precise arrival times of P, S, and/or converted phases from seismic waveforms, phase picking enables the determination of key earthquake parameters such as event location, depth, and magnitude (e.g., Ng et al., 2024). These parameters are essential for understanding fault mechanics, characterizing subsurface structures, and monitoring seismic hazards. Because induced seismicity at geothermal fields typically involves small-magnitude events and network densities vary across sites, phase-picking tools must be broadly applicable or readily adaptable through transfer learning.

4.2 Architecture and Results

We develop a hierarchical encoder–decoder architecture designed to estimate probability distributions of P- and S-wave arrivals from three-channel seismic waveform data (Figure 4). The architecture integrates a feature-rich encoder with a streamlined decoder to ensure accurate detection of seismic arrivals while maintaining computational efficiency. The encoder employs a hierarchical structure comprising four sequential blocks (E1–E4), each operating at progressively coarser spatial scales corresponding to downsampling rates of 2, 4, 8, and 16. To preserve feature expressiveness during downsampling, the hidden dimensionality is expanded in the E1 and E3 blocks. Each block integrates ConvNeXt modules with inverted bottleneck structures to enhance feature extraction. The ConvNeXt block consists of a 1×1 input convolutional layer, a 3×3 depthwise convolutional layer, and two linear projection layers with a channel expansion ratio of four. This configuration separates spatial and channel-wise information processing, reducing computational overhead while retaining feature selectivity. Intermediate layers use Gaussian Error Linear Units (GELU) for nonlinearity and Global Response Normalization (GRN) to mitigate feature collapse through global aggregation, normalization, and calibration. These design choices ensure robust and scalable feature embedding.

Encoded features are passed through a simplified decoder comprising two convolutional blocks (D1 and D2) with 3×3 convolutional layers and GELU activations. To refine the spatial and temporal resolution of the output, a residual refinement block (R1) is incorporated, utilizing a 1×1 convolutional layer to project the final output. The decoder output represents the probabilities of P- and S-wave arrivals at each time step. These probabilities are post-processed to determine arrival times: the time step corresponding to the maximum P-wave probability identifies the P-wave arrival, while the time step with the peak S-wave probability determines the S-wave arrival. This approach enables robust arrival-time estimation even under noisy predictions.

To improve the robustness of phase detection for small-magnitude earthquakes, we introduce a data augmentation strategy. This augmentation is particularly important for geothermal fields, where induced seismicity often exhibits low signal-to-noise ratios (SNR) and complex wave propagation effects. Given the variability in seismic waveforms arising from differences in source characteristics, propagation paths, and site conditions, augmentation increases training diversity and mitigates overfitting. Our approach introduces controlled variations in both time and frequency domains to replicate realistic perturbations in seismic data acquisition.

In the time domain, random time shifts are applied to simulate arrival-time uncertainties caused by minor velocity variations or phase misalignment, preventing the model from overfitting to rigid arrival assumptions. Amplitude scaling is introduced to account for variations in source size, radiation patterns, and attenuation effects, forcing the network to focus on waveform shape rather than absolute amplitude and improving robustness across magnitudes. Additionally, time warping introduces nonlinear temporal distortions to mimic velocity heterogeneities and scattering effects observed in geothermal and structurally complex regions. Together, these transformations enhance resilience to misalignment and propagation-related distortions.

In the frequency domain, spectral perturbations are applied to simulate site-specific attenuation, frequency-dependent absorption, and instrumental response variations. Minor frequency shifts and selective filtering ensure that the network does not rely excessively on narrowband frequency content. Furthermore, noise injection emulates real-world seismic recording conditions by adding Gaussian noise or field-recorded background noise, improving detection performance in low-SNR environments where small-magnitude events may be obscured.

Rather than applying static augmentations, these variations are dynamically generated during training, ensuring that each epoch exposes the network to a diverse set of conditions. This continuous transformation prevents memorization of specific noise patterns or waveform characteristics, enhancing adaptability to unseen seismic data. Owing to this data augmentation strategy, only seven earthquakes are required to train the network while achieving reliable performance in event location and polarity estimation.

We apply and evaluate the trained network for phase picking using seismic wavefields recorded by 258 nodal sensors at the Cape Geothermal Field, Utah (Nakata et al., 2025; Figure 5). In this example, we focus exclusively on P-wave picking because of the complexity of S-wave arrivals, potentially caused by structural anisotropy. Among the 258 traces, some stations are located in close proximity (similar to the 4-by-4 array discussed in the previous section), resulting in wavefields that arrive nearly simultaneously (e.g., Traces 1–16 in Figure 5). The strong agreement between predicted and observed arrivals highlights the robustness and reliability of the model. The network effectively identifies P-wave arrivals under varying noise conditions and waveform complexities, making it well suited for seismic monitoring and hazard assessment. For traces with very low SNR (e.g., Traces 98–130), precise P-wave picking is more challenging; however, such errors can be readily suppressed by excluding picks associated with unrealistic apparent velocities. In practice, incorporating an explicit denoising step—such as combining multiple frequency bands to isolate signal signatures—could further enhance picking quality.

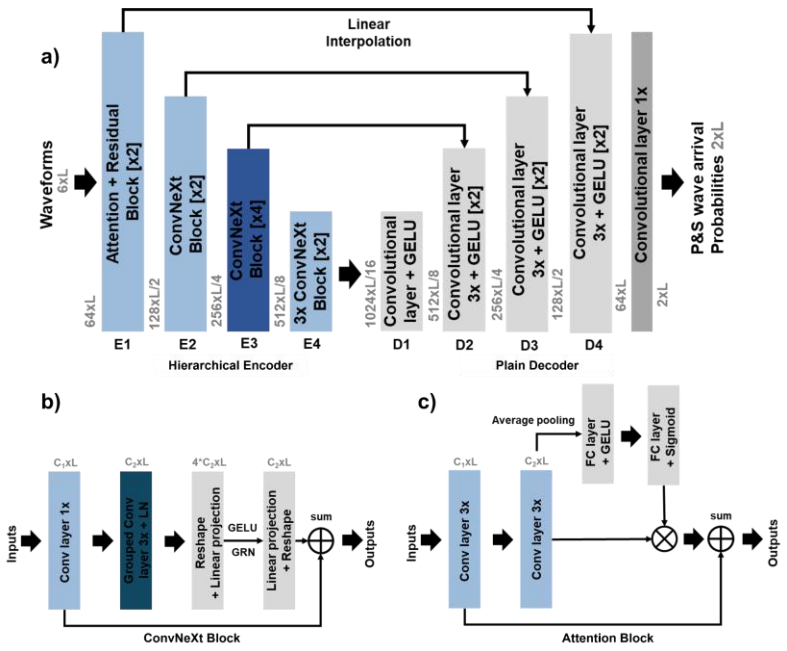

Figure 4. (a) Architecture of the proposed neural network for seismic waveform analysis. The hierarchical encoder comprises (b) ConvNeXt-based blocks and (c) Attention blocks with progressively increasing receptive fields and channel dimensions to extract multi-scale features. The decoder refines the encoded features through convolutional layers and GELU activation, producing the output probability distributions for P- and S-wave arrivals.

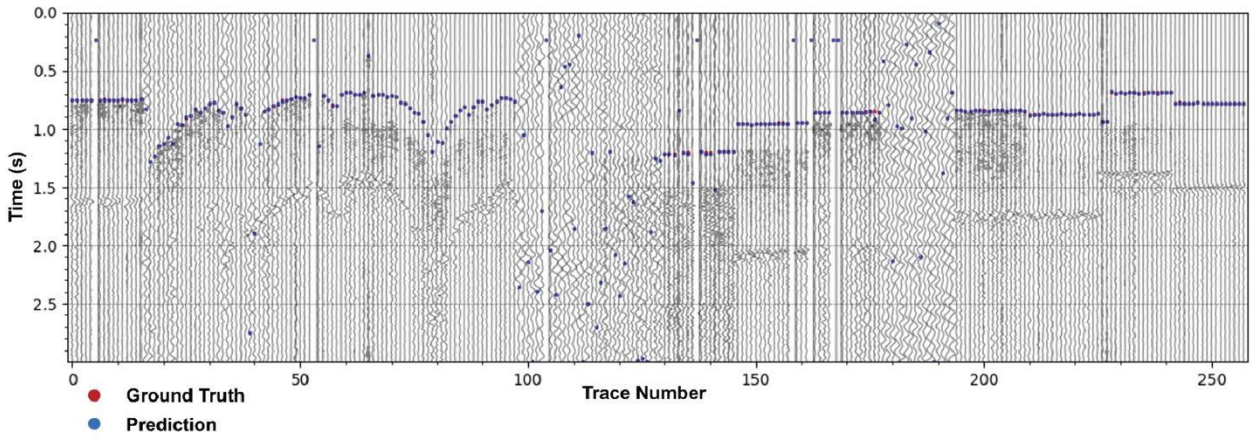

Figure 5. Performance of the trained network for phase picking on the nodal seismic records at the Cape Geothermal Field, Utah. The waveforms are aligned by the receiver numbers for deployment (Nakata et al., 2024b). The blue and red dots represent the predicted and manually picked (i.e., ground true) arrival times, respectively.

5. EARTHQUAKE CLUSTERING

5.1 Background and Dataset

The spectral properties of microseismic recordings provide valuable insights into fracture processes, as variations in failure and friction mechanisms influence the frequency content of seismic signals. Previous studies (e.g., Ohnaka & Mogi, 1981; Chitrala et al., 2013) have shown that events involving shear or complex mechanisms occupy similar frequency ranges, whereas compressive events tend to exhibit higher frequencies and tensile events generate the lowest average frequencies. Extensive research has employed moment tensor analysis to characterize shear and tensile events, particularly at The Geysers geothermal field and in other regions (e.g., Boyd et al., 2018; Martínez-Garzón et al al., 2014, 2017). However, applying moment tensor analysis to microseismic events induced by hydraulic fracturing presents significant challenges due to the typically low signal-to-noise ratio (SNR) and limited azimuthal coverage of observation networks, which hinder reliable moment tensor inversion. In such cases, identifying systematic and subtle differences in the temporal–spectral characteristics of microseismic signals (e.g., Holtzman et al., 2018; Ida et al., 2022) provides an alternative means of inferring fracture mechanisms and offers deeper insight into the dynamics of hydraulic fracturing.

To advance this capability, we present a systematic workflow for extracting reliable spectral characteristics from low-SNR microseismic recordings. This approach involves correcting power spectral densities (PSDs) to minimize variations caused by ray paths and propagation distances. We then apply K-means clustering, a widely used unsupervised learning algorithm, to identify distinct patterns in characteristic PSDs that reflect the influence of fracture mechanisms and medium properties. By grouping microseismic events based on spectral similarity, this method provides a robust framework for categorizing seismicity and linking spectral features to underlying fracture processes.

We apply the clustering workflow to microseismicity at the Utah FORGE site near Milford, Utah, a DOE research facility dedicated to advancing enhanced geothermal systems (EGS) technology by utilizing a high-temperature granitic reservoir. Despite challenges associated with high temperatures and in situ reservoir conditions, a major hydraulic stimulation was conducted in April 2022 in well 16A-32 to enhance reservoir permeability by creating fractures through high-pressure fluid injection. Seismic monitoring using a network of downhole geophones and surface stations recorded more than 10,000 microseismic events during the three-stage stimulation, of which over 2,500 events were reliably located to form a comprehensive seismic catalog. Here, we exclusively use data recorded by downhole geophones (Figure 6a–b), as small-magnitude events are often undetectable by surface stations.

5.2 Pre-Processing and K-Mean Clustering

As a preprocessing step, we calculate the characteristic power spectral density (PSD) for each microseismic event. Microseismic waveforms are first segmented using the reference catalog, as illustrated for a representative event in Figure 6. Both P and S waves are clearly observed across the downhole geophone arrays in wells 56, 58, and 78B. The downhole arrays in wells 58 and 78B each consist of eight sensors deployed at depths of approximately 2 km, with a station spacing of ~30 m. In contrast, only two sensors with poor data quality were operational in well 56 during the stimulation. Consequently, we focus our analysis on events from the third stimulation stage in well 16A-32, using data from wells 78B and 58, as the sensors in well 78B became operational only after the second stimulation stage.

Assuming that P and S waves propagate along straight ray paths between the source and receivers, we estimate P- and S-wave velocities (Vp and Vs) using beamforming analysis to optimize the alignment of recorded P- and S-wave arrivals across the downhole arrays (Figure 6f–h). The estimated Vp and Vs values yield a narrow Vp/Vs ratio distribution centered at ~1.7, consistent with the local velocity model (Zhang & Pankow, 2021).

To compute the PSDs of the P- and S-waveforms at each station, we first calculate the PSD of the truncated waveform for each component using multitaper spectral analysis (Prieto et al., 2009) and then sum the PSDs over all three components. The resulting single-station PSDs exhibit complex spectral patterns (Figure 6h), likely reflecting in situ effects such as ambient noise, sensor coupling and performance, and scattered waves generated by heterogeneous structures along the ray path (Oren & Nowack, 2017; Qin et al., 2023). To suppress these station-specific effects, we average the single-station PSDs on a logarithmic scale over all geophones within the same array (black curve in Figure 6h). We then average the array-mean PSDs from the two geophone arrays as a first-order correction for attenuation effects associated with differences in source–receiver distances.

We apply the K-means clustering algorithm to group one-dimensional spectral curves into distinct clusters, each represented by a centroid that characterizes the average spectral profile of events within that cluster. The objective of clustering is to identify common spectral patterns among microseismic events and to distinguish between different fracture mechanisms. The clustering procedure begins by randomly selecting K initial centroids from the dataset of N spectral curves. These centroids, denoted as $\{C_1(f), C_2(f), \ldots, C\_K(f)\}$, define the initial cluster profiles. For each spectral curve, the Euclidean (L2) distance to each centroid is computed, and the curve is assigned to the cluster with the smallest distance. The centroids are then updated as the mean spectral profile of all curves assigned to each cluster. This iterative process continues until the centroids converge, ensuring that events within each cluster share similar spectral characteristics.

In our analysis, each K-cluster represents a distinct group of microseismic events with similar spectral signatures, likely associated with different fracture processes. The choice of K strongly influences the clustering outcome: too few clusters may oversimplify the classification, whereas too many may introduce unnecessary subdivisions. We determine the optimal number of clusters using the elbow method, identifying the point beyond which additional clusters produce only marginal reductions in within-cluster variance. The choice of distance metric also affects the results. Clustering based on raw attenuation-corrected spectral profiles emphasizes differences in spectral peak amplitudes and frequencies, whereas clustering on logarithmically transformed profiles (in dB) highlights relative amplitude ratios. In this study, we cluster raw attenuation-corrected spectral profiles to extract distinct spectral features that provide insight into fracture mechanisms and reservoir properties.

5.3 Results and Interpretation

Figures 7a–d highlight the spectral characteristics and spatial distributions of microseismic event clusters identified through K-means clustering of the geometric-mean PSDs of P waves. Four clusters (C1–C4) are determined using the elbow method, with their spectral profiles shown in the top panels. Clustering is performed using PSDs on a linear scale, which provides clearer separation than clustering on a logarithmic scale. The centroids of the four clusters, which peak near 400 Hz, differ primarily in the location and amplitude of their dominant spectral peaks. From C1 to C3, the spectral peaks shift progressively toward higher frequencies, whereas C4 exhibits a comparatively flattened peak. The shaded regions indicating within-cluster variability are smaller than the differences between cluster centroids, confirming the robustness of the clustering results.

The spatial distributions of the clusters, shown in map view (second row of Figures 7a–d), reveal distinct patterns. Events in clusters C3 and C4 are concentrated within a symmetric ~120 m–wide zone around the stimulation well, with C3 events clustering closer to the wellbore. In contrast, events in clusters C1 and C2 extend beyond this zone, with C2 showing a preference for regions with positive Y

values. These spatial differences suggest that clusters C1 and C2 are likely associated with stress redistribution over a broader area, whereas clusters C3 and C4 are more directly related to injected fluid and are confined to a smaller region. Cross-sectional views show no pronounced differences among the clusters, implying that fracture mechanisms are influenced anisotropically.

Figure 7e illustrates the spatiotemporal evolution of microseismic events for each cluster, while Figure 7f shows the seismicity rate as a function of time. During the stimulation period, C4 events dominate the seismicity, while each cluster exhibits a distinct temporal pattern. Clusters C3 and C4 show a sharp decline in seismicity rate after approximately four hours, consistent with our hypothesis that these events are directly associated with fluid injection (hereinafter referred to as "fluid-related clusters"). In contrast, the more sustained seismicity observed in clusters C1 and C2 suggests continued influence of stress redistribution over time (hereinafter referred to as "stress-related clusters").

Temporally, clusters C1 and C2 exhibit a secondary increase in activity approximately three hours after stimulation onset, consistent with mechanisms dominated by stress redistribution induced by fluid injection. Meanwhile, the steady decline in clusters C3 and C4 supports the interpretation that these events result from fracturing processes directly driven by injected fluids. Overall, the microseismicity initially appears to be driven by the creation of new fractures near the injection well due to elevated pore pressure, transitioning to fluid migration as the dominant mechanism after roughly three hours.

The combined spatial and temporal distributions of the clusters further enhance our understanding of the underlying fracture processes. In regions associated with fluid-related clusters (C3 and C4), indicators of high fluid flow are observed (Figures 7e–f), suggesting that optimal placement of a second deviated well should target these zones rather than the entire microseismic cloud. In addition, the asymmetric spatial patterns observed in stress-related clusters (C1 and C2) point to directional weaknesses or spatial variations in rock susceptibility to stress perturbations.

We emphasize that these interpretations are based on a preliminary spatiotemporal analysis of event clusters grouped by their array-mean PSDs. While the results provide valuable insight into fracture dynamics during hydraulic stimulation, further investigation incorporating additional observations and more detailed mechanistic analyses will be presented in a forthcoming paper dedicated to microseismic clustering.

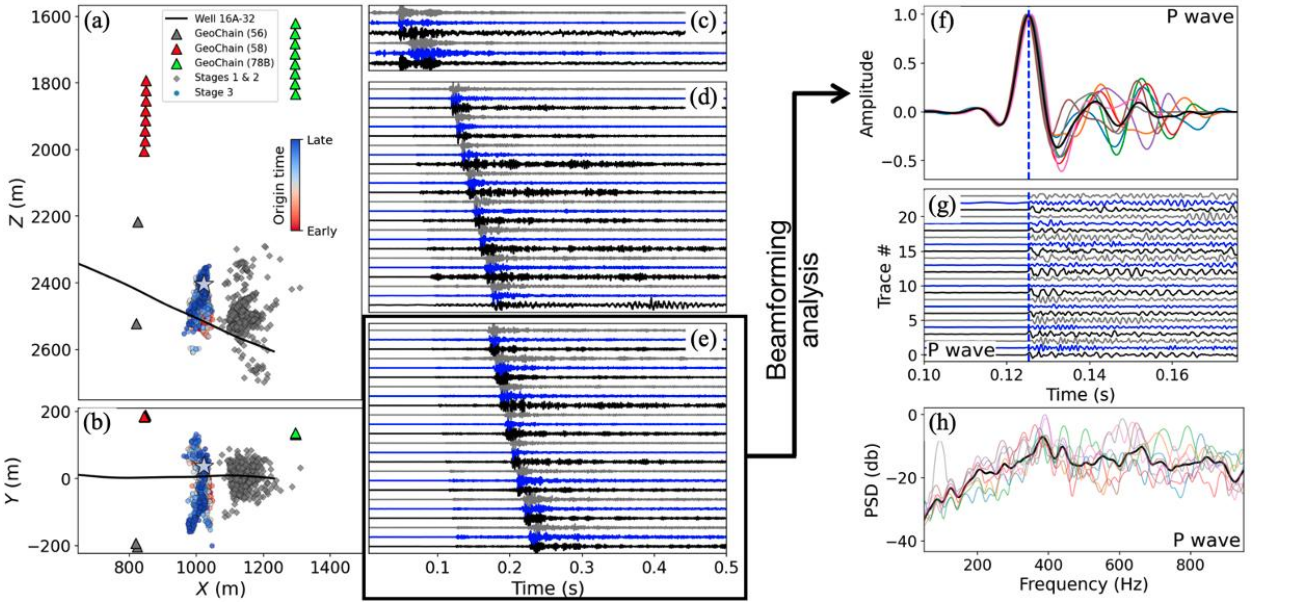

Figure 6. (a) Cross-section view and (b) map view of microseismic event distribution. Events in each stimulation stage are color-coded by origin time. The waveforms for a representative event (star) recorded by downhole geophone arrays in wells (c) 56, (d) 58, and (e) 78B are shown, truncated to a 0.5-second window from the event origin time. Different waveforms components are represented in distinct colors. Panels (f)-(g) illustrate the beamforming analysis for P waves recorded by well 78B, while panel (h) presents the corresponding single-station PSDs (thin colored curves) and the array-mean PSD (thick black curve). The beamforming analysis determines the optimal P-wave slowness by grid-searching to align the characteristic functions, defined as the time gradient of envelope functions summed over three components (panel f). The shallowest geophone serves as the reference for this analysis. Blue dashed lines in panels (f)-(g) indicate the resulting P picks, aligned according to the optimal P-wave velocities. The same approach can be applied to analyze S waves.

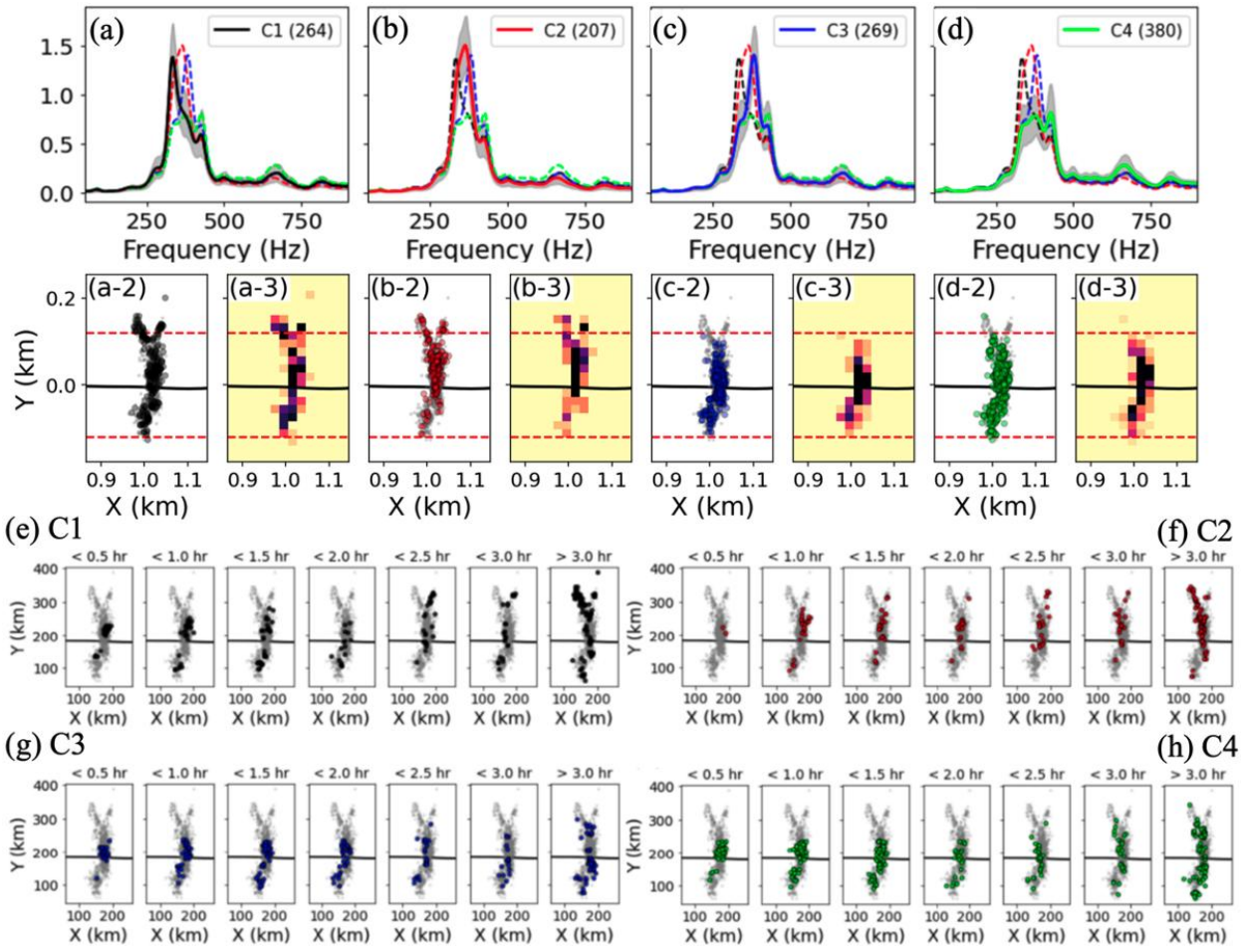

Figure 7. Clustering results of P-wave PSDs for the Utah FORGE dataset. (a)-(d): The top panels display the centroid spectral profiles (solid color curves), with the number of events indicated in the legend. The gray shaded area encompasses 80% of the spectral curves within the cluster, representing the inherent variability among events. To better illustrate differences between clusters, centroid PSDs from other clusters are shown as dashed curves in their corresponding colors. Second row panels (a-2 to d-2, a-3 to d-3): The spatial distributions of microseismic events for each cluster are presented, with (a-2 to d-2) showing event locations as scatter plots and (a-3 to d-3) displaying density plots. These visualizations highlight spatial clustering differences, with some clusters concentrated near the injection well and others more widely distributed. (e-1) to (e-4) illustrates the spatiotemporal distributions of events in each cluster. (f-1) and (f-2): Seismicity rate as a function of time, shown in absolute numbers and percentage, respectively.

6. DISCUSSION

6.1 Pre-Processing

Although machine learning models ideally learn features directly from raw data, limited data availability, noise, and the inherent variability of seismic records pose significant challenges. To address these issues, preprocessing steps are often required. In this study, we apply preprocessing methods such as spectral analysis, attenuation correction, and bandpass filtering to obtain reliable results across different applications. These procedures ensure that ML models receive clean, well-structured, and optimized inputs, thereby enhancing their ability to generalize across diverse seismic events and regions.

Preprocessing, however, also introduces challenges and limitations. It relies heavily on expert judgment to design and implement appropriate procedures, creating dependence on specialized knowledge. Moreover, preprocessing can inadvertently introduce data-selection biases based on expert preferences, potentially affecting both model performance and interpretability. The scarcity of high-quality annotated datasets remains a persistent challenge for supervised learning. Consequently, alternative strategies are needed to reduce reliance on extensive preprocessing. One such strategy is data augmentation, which has emerged as a powerful approach for mitigating the limitations associated with small or insufficient datasets.

6.2 Data Augmentation

Data augmentation has become a critical strategy for overcoming the shortage of high-quality seismic data. By synthetically expanding the size and diversity of datasets, data augmentation enables machine learning models to achieve improved performance. This process not

only compensates for data scarcity but also mitigates overfitting by introducing variability into the training data. Techniques such as waveform perturbation, which simulate natural variability through random time shifts, amplitude scaling, and noise injection, are particularly effective. These methods generate training samples that more closely reflect real-world conditions, thereby improving model robustness and adaptability. More advanced approaches, such as generative modeling exemplified by CGM-Wave, further extend these capabilities by synthesizing realistic seismic waveforms and time–frequency spectra. These synthetic data products are especially valuable for training ML models for tasks such as phase picking and seismic event clustering, providing richer datasets for these critical applications.

Despite its advantages, data augmentation alone cannot fully eliminate the need for preprocessing or manual intervention. Many augmentation techniques still rely on preprocessed data as their starting point, and fully automated workflows that bypass these steps remain challenging to implement. The development of end-to-end learning frameworks represents an important step toward addressing this limitation. Such frameworks aim to process raw seismic data directly, reducing dependence on manual preprocessing while streamlining the overall analysis workflow. This shift has the potential to improve both efficiency and reproducibility in seismic data analysis. In addition, integrating physics-informed machine learning models offers another promising avenue. By embedding domain-specific knowledge into the learning process, these models strike a balance between data-driven inference and physical consistency, thereby reducing reliance on preprocessing and enhancing model interpretability and reliability.

The combined use of data augmentation, end-to-end learning frameworks, and physics-informed approaches represents a transformative shift in machine learning–based seismic data analysis. Together, these advances can alleviate challenges related to data scarcity and manual preprocessing, enabling more accurate, efficient, and scalable workflows. However, continued research and development are required to fully realize their potential, particularly in creating systems that seamlessly integrate raw data, domain expertise, and advanced algorithms. As these approaches mature, they are likely to reshape how seismic data are prepared and analyzed, opening new opportunities for addressing complex geophysical problems.

7. CONCLUSIONS

We present the promising role of machine learning in each step of induced seismicity data analysis and interpretation for geothermal energy exploration and development. Specifically, we demonstrate the use of interpolation and generative AI to fill data gaps and synthesize new seismic wavefields. We then apply ML to improve the accuracy of automated phase picking and to significantly reduce the time required for manual phase picking. Finally, we propose an ML-aided interpretation framework based on earthquake clustering using spectral characteristics to reveal hidden features likely related to the physical mechanisms driving seismicity.

Preprocessing and data augmentation are crucial for preparing induced seismicity data for ML applications. Although ML models ideally operate directly on raw data, practical challenges such as data scarcity, noise, and waveform variability often prevent this. Consequently, the development of effective preprocessing schemes and data augmentation strategies is essential for extracting meaningful features, reducing data complexity, and optimizing inputs for successful ML applications in geothermal fields. This is particularly important during early project stages, when seismicity is sparse and conventional ground-motion models or reservoir imaging approaches are not yet feasible.

8. ACKNOWLEDGMENT

This work was supported by the U.S. Department of Energy, Hydrocarbons and Geothermal Energy Office, under Award Number DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory, as well as by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. We are grateful to Fervo Energy, Calpine, and Utah FORGE for sharing data and/or allowing us to deploy seismic sensors in their fields. C.-N. Liu was also supported by the National Science Foundation's internship program.

REFERENCES

Aki, K., & Richards, P. G. (2002). Quantitative seismology.

Bergen, K.J., Johnson, P.A., Hoop, M. V. De & Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science. doi:10.1126/science.aau0323

Bi, Z., Nakata, N., Nakata, R., Ren, P., Wu, X. & Mahoney, M. W., 2025. Advancing data-driven broadband seismic wavefield simulation with multi-conditional diffusion model, arXiv:2501.14348.

Bouchon, M. (1981). A simple method to calculate Green's functions for elastic layered media. Bulletin of the Seismological Society of America, 71(4), 959-971.

Boyd, O. S., Dreger, D. S., Gritto, R., & Garcia, J., 2018. Analysis of seismic moment tensors and in situ stress during Enhanced Geothermal System development at The Geysers geothermal field, California. Geophysical Journal International, 215(2), 1483–1500. doi:10.1093/gji/ggy326.

Chitrala, Y., Moreno, C., Sondergeld, C. & Rai, C., 2013. An experimental investigation into hydraulic fracture propagation under different applied stresses in tight sands using acoustic emissions. J Pet Sci Eng, 108. doi:10.1016/j.petrol.2013.01.002

Dahlen, F., & Tromp, J. (1998). Theoretical global seismology. In Theoretical Global Seismology. Princeton university press.

Eisner, L., Williams-Stroud, S., Hill, A., Duncan, P. & Thornton, M., 2010. Beyond the dots in the box: Microseismicity-constrained fracture models for reservoir simulation. Leading Edge (Tulsa, OK), 29. doi:10.1190/1.3353730

Holtzman, B.K., Paté, A., Paisley, J., Waldhauser, F. & Repetto, D., 2018. Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field. Sci Adv, 4. doi:10.1126/sciadv.aao2929

Ida, Y., Fujita, E. & Hirose, T., 2022. Classification of volcano-seismic events using waveforms in the method of k-means clustering and dynamic time warping. Journal of Volcanology and Geothermal Research, 429. doi:10.1016/j.jvolgeores.2022.107616

Kong, Q., Trugman, D.T., Ross, Z.E., Bianco, M.J., Meade, B.J. & Gerstoft, P., 2019. Machine learning in seismology: Turning data into insights. Seismological Research Letters, 90. doi:10.1785/0220180259

Chenyu Li, Zefeng Li, Zhigang Peng, Chengyuan Zhang, Nori Nakata, Tim Sickbert; Long‐Period Long‐Duration Events Detected by the IRIS Community Wavefield Demonstration Experiment in Oklahoma: Tremor or Train Signals?. Seismological Research Letters 2018;; 89 (5): 1652–1659. doi: https://doi.org/10.1785/0220180081

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1833-1844).

Martínez-Garzón, P., Kwiatek, G., Sone, H., Bohnhoff, M., Dresen, G., & Hartline, C., 2014. Spatiotemporal changes, faulting regimes, and source parameters of induced seismicity: A case study from The Geysers geothermal field. Journal of Geophysical Research: Solid Earth, 119, 8378–8396. doi:10.1002/2014JB011385.

Martínez-Garzón, P., Kwiatek, G., Bohnhoff, M., & Dresen, G., 2017. Volumetric components in the earthquake source related to fluid injection and stress state. Geophysical Research Letters, 44, 800–809. doi:10.1002/2016GL071963.

Majer, Ernest L., Roy Baria, Mitch Stark, Stephen Oates, Julian Bommer, Bill Smith, and Hiroshi Asanuma. "Induced seismicity associated with enhanced geothermal systems." Geothermics 36, no. 3 (2007): 185-222.

Maxwell, S.C., 2014. Microseismic Imaging of Hydraulic Fracturing : Improved Engineering of Unconventional Shale Reservoirs. Society of Exploration Geophysicists.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y. & Beroza, G. C. Earthquake transformer-an attentive deep-learning model for simultaneous earthquake detection and phase picking. Nat. Commun. 11, 3952 (2020).

Nakata, N., H. Chang, S.-M. Wu, Z. Bi, L.-W. Chen, F. Soom, H. Gao, A. Titov, and S. Dadi (2025) Fracture characterization revealed by Microseismicity at Cape Modern Geothermal Field, Utah, 50th workshop on Geothermal Reservoir Engineering, SGP-TR-229.

Nakata, N., Wu, S.-M., Hopp, C., Jung, Y., Luo, L., Lisabeth, H., Sonnenthal, E., Smith, T., Robertson, M., Vasco, D. W., Dadi, S., Microseismicity observation and characterization at Cape Modern, Utah, Geothermal Rising Conference Transactions, 48, 2024b.

Nakata, N., S.-M. W., C. Hopp, M. Robertson, and S. Dadi (2024a) Microseismicity observation and characterization at Cape Modern and Utah FORGE, 49th workshop on Geothermal Reservoir Engineering, SGP-TR-227.

Raymond Ng, Xiaowei Chen, Nori Nakata, Jacob I Walter, Precise relative magnitude measurement improves fracture characterization during hydraulic fracturing, Geophysical Journal International, Volume 238, Issue 2, August 2024, Pages 1040–1052, https://doi.org/10.1093/gji/ggae204

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. doi:10.1016/j.geothermics.2024.102947

Ohnaka, M. & Mogi, K., 1981. Frequency dependence of acoustic emission activity in rocks under incremental, uniaxial compression. Bulletin of the Earthquake Research Institute, 56, 67–89.

P. Olasolo, M.C. Juárez, M.P. Morales, Sebastiano D´Amico, I.A. Liarte (2016) Enhanced geothermal systems (EGS): A review, Renewable and Sustainable Energy Reviews, 56, 133-144.

Oren, C. & Nowack, R.L., 2017. Seismic body-wave interferometry using noise autocorrelations for crustal structure. Geophysical Journal International Geophys. J. Int, 208, 321–332. doi:10.1093/gji/ggw394

Prieto, G.A., Parker, R.L. & Vernon, F.L., 2009. A Fortran 90 library for multitaper spectrum analysis. Comput Geosci. doi:10.1016/j.cageo.2008.06.007

Qin, L., Qiu, H., Nakata, N., Deng, S., Levander, A. & Ben-Zion, Y., 2023. Variable Daily Autocorrelation Functions of High-Frequency Seismic Data on Mars. Seismological Research Letters, 94. doi:10.1785/0220220196

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2536-2544.

Ren, P., Nakata, R., Lacour, M., Naiman, I., Nakata, N., Song, J., Bi, Z., Malik, O. A., Morozov, D., Azencot, O., Erichson, N. B., Mahoney, M. W., 2024, Learning Physics for Unveiling Hidden Earthquake Ground Motions via Conditional Generative Modeling, arXiv, 2407.15089.

Warpinski, N.R., Du, J. & Zimmer, U., 2012. Measurements of hydraulic-fracture-induced seismicity in gas shales. SPE Production and Operations, 27. doi:10.2118/151597-PA

White, M.C.A., Sharma, K., Li, A. et al. Classifying seismograms using the FastMap algorithm and support-vector machines. Commun Eng 2, 46 (2023). https://doi.org/10.1038/s44172-023-00099-8

Xiong, N., Qiu, H. & Niu, F., 2021. Data-Driven Velocity Model Evaluation Using K-Means Clustering. Geophys Res Lett, 48. doi:10.1029/2021GL096040

Zhang, H. & L. Pankow, K., 2021. High-resolution Bayesian spatial autocorrelation (SPAC) quasi-3-D Vs model of Utah FORGE site with a dense geophone array. Geophys J Int, 225. doi:10.1093/gji/ggab049

Zhang, Z., White, M.C.A., Bai, T., Qiu, H. & Nakata, N., 2023. Characterizing Microearthquakes Induced by Hydraulic Fracturing with Hybrid Borehole DAS and Three-Component Geophone Data. IEEE Transactions on Geoscience and Remote Sensing, 61, Institute of Electrical and Electronics Engineers Inc. doi:10.1109/TGRS.2023.3264931

Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic arrival- time picking method. Geophysical Journal International, 216(1), 261–273.

Zoback, M.D. & Kohli, A.H., 2019. Unconventional Reservoir Geomechanics. Unconventional Reservoir Geomechanics. doi:10.1017/9781316091869