UNSUPERVISED LEARNING-BASED ANALYSIS OF HYDRAULIC FRACTURING-INDUCED SEISMICITY

A Dissertation

by

ADITYA CHAKRAVARTY

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Siddharth Misra
Committee Members,	Kan Wu
	Benchun Duan
	A. Daniel Hill
Head of Department,	Akhil Datta-Gupta

May 2023

Major Subject: Petroleum Engineering

ABSTRACT

Passive seismicity is crucial for optimizing hydraulic fracture treatments and reducing the risk of hydraulic fracturing-induced seismicity. Unsupervised machine learning is suitable for understanding fracturing-induced seismicity because it can process large amounts of data without prior knowledge or labeling and identify patterns and relationships in the data that isn't apparent to the human eye. Clustering and dimensionality reduction can be used to group similar earthquakes and reveal patterns in the distribution and evolution of seismicity over time. This can provide valuable insights into the behavior of the fractures and help identify the conditions that may lead to induced seismicity.

In case of meso-scale (~ 10 m) microseismicity, the analysis is based on Experiment 1 of the EGS Collab Project. A non-linear dimension reduction is applied on the high dimensional features extracted from seismic signals generating embeddings in low dimensions. The different groups obtained by clustering the low dimension embeddings reveal that different clusters are related to distinct fracture networks. A separate study on the same experiment focused on the simultaneous, wide-band hydrophone signals. Low frequency signals (2-80 Hz) were found to be strong related to the fluid injection. Workflows were developed to reliably locate the sources of the low frequency signals and their spatiotemporal distribution was correlated with the distribution of natural fractures. Additionally, workflows based on semi-supervised, graph-based label propagation are developed to reliably extend fracture labels to noisy and unlabeled microseismic data. Lastly, a field (~km) scale microseismic dataset containing high quality moment tensor information was used to explore relationship between the three-dimensional motion recorded at geophones and the deformation occurring at hydraulic fracture planes.

DEDICATION

This dissertation is dedicated to my family, friends and my partner, Japneet.

I am because you are.

ACKNOWLEDGEMENTS

I express my heartfelt gratitude to my PhD supervisor and mentor, Dr. Siddharth Misra. His work ethic and dynamism has been a constant source of inspiration for me. He showed incredible patience with me and was always there to guide and support me through difficult times. Forever indebted and incredibly fortunate to work under his supervision.

Equally fortunate I am to have crossed paths with the remarkable folks at Lawrence Berkeley National Laboratory. In no order of importance: Timothy Kneafsey, Patrick Dobson, Martin Schoenball, Chet Hopp, Veronica Rodriguez. All of them tremendous scientists who were extremely supportive, cheerful, and taught me a great deal about 'big science'.

I want to thank my PhD committee members: Drs. Kan Wu, Benchun Duan and Daniel Hill. They have been extremely supportive throughout my journey and given valuable inputs to my scientific work, in addition to being excellent teachers who taught seminal topics.

I am forever grateful to my colleague Rui Liu, without whose help I literally would not have passed the petroleum engineering academic courses, as well as the PhD qualifying examination.

My scientific training in the field of my PhD began at the University of Oklahoma, at the renowned Integrated Core Characterization Center (IC3) lab. My mentors Chandra Rai and Carl Sondergeld were the first to teach me the art of scientific research.

Last but not the least, I am grateful to the faculty and staff of Petroleum Engineering department and other offices at Texas A&M University who run this immense organization, Texas A&M University, and make it a life changing experience for students like me.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Siddharth Misra (chair, Petroleum Engineering), Dr. Kan Wu (Petroleum Engineering), Dr Benchun Duan (Petroleum Engineering), and Dr Daniel Hill (Petroleum Engineering).

All work conducted for the dissertation was completed by the student independently.

Funding sources

The work in this dissertation was funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences Geosciences, and Biosciences Division, under Award Number DE-SC0020675.

NOMENCLATURE

AE	Acoustic emission
DC	Double Couple
ISO	Isotropic
CLVD	Compensated Linear Vector Dipole
FFT	Fast Fourier Transform
GAI	Geomechanical Alteration Index
ML	Machine Learning
OT/OB-xx	Monitoring wells' names in the EGS Collab project
PCA	Principal Component Analysis
SNR	Signal-to-noise ratio
SRV	Stimulated Reservoir Volume
STFT	Short-time Fourier Transform
UMAP	Uniform Manifold Approximation Projection

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES Page

Page

Page

Page

LIST OF TABLES Page

1. INTRODUCTION

Fractures are ubiquitous and are either naturally or artificially generated. In petroleum and geothermal engineering, hydraulic fracturing is widely used to generate artificial fractures to enhance rock permeability. Hydraulic fracture characterization is of great significance because it is a crucial step before making forecasts and taking development measures. This is a huge challenge at present because of the large heterogeneities and uncertainties involved, especially after hydraulic fracturing and underground mining. In recent years progress has been made in monitoring technologies and assisted fracture interpretation and modeling approaches, such as seismic and microseismic monitoring, but many knowledge gaps persist. The response of mechanical and electromagnetic waves within geological media is complicated due to the inherent complexity of rocks - heterogeneous composition, presence of material discontinuities (fractures, voids) and variable void size and shape distribution.

Common to all these methods is the fact that they generate a large data volume. Moreover, human factors such as cognitive restrictions in vision, fatigue, subjectivity, and bias can introduce uncertainty in the outcomes.

Machine Learning (ML) methods are well suited for signal processing associated with seismic based fracture characterization. ML methods can be broadly divided into two groups: supervised (in which the target data label for set of measurements is known) and unsupervised (in which no a-priori information about the target label is known). Unsupervised ML methods like clustering can be applied to datasets to reveal physics previously unaccounted for by analytical models.

The problem statements driving the research contained in this thesis are outlined as follows:

1. Only the first phase arrivals are used for characterizing the stimulated reservoir volume (SRV) from microseismic data.
2. Except for specialized research, the three-component information is not utilized for microseismic-based hydraulic stimulation monitoring.
3. Poor understanding of spatial distribution of moment tensor types over the stimulated reservoir volume, and the predictive potential of seismic signal for moment tensor type.
4. Low frequency seismic signals are not utilized for characterization of SRV.
5. Difficulty in assigning fracture labels to event hypocenters for SRV interpretation.

The first chapter sets the precedent for the research directions in the dissertation. It gives a high-level overview of the challenges associated with passive seismic measurements and the opportunities unsupervised machine learning presents for tackling those challenges. The second chapter presents select geophysical theories of seismic wave propagation in fractured media, focal mechanisms associated with fracturing related rock deformation and concept of moment tensors. Third chapter gives a literature review of the research done till now that is related to the objectives of research objectives undertaken herein. The first part outlines the state of the art for seismic based hydraulic fracture characterization. The second part outlines a survey of unsupervised machine learning applied to seismological data analysis. Fourth chapter gives a detailed description of the experiments analyzed in laboratory scale, intermediate scale, and field scale hydraulic fracturing experiments. Fifth chapter discusses details about the data processing associated with experiments involved in analysis. The sixth chapter presents the results and conclusions of the analysis of the multi scale fracturing based on unsupervised learning.

2. RELEVANT THEORY FOR WAVE PROPAGATION IN FRACTURED MEDIA, FOCAL MECHANISMS AND MOMENT TENSORS

2.1 Wave transmission through fractured rock

To understand how fractures affect transmission of seismic waves, researchers can use a mathematical approach that treats the fracture as a boundary where there is a discontinuity in the displacement of particles. This boundary condition is characterized by a fracture stiffness, and it leads to a discontinuity in displacement across the boundary that is proportional to the ratio of the seismic stress to the specific stiffness. In addition to causing this discontinuity in displacement, the boundary behaves like a low-pass filter, meaning that it allows low-frequency waves to pass through while blocking high-frequency waves (Pyrak-Nolte et al., 2001).

The characteristics of a fracture that are important for understanding its effects on wave propagation include its length, the extent to which its faces are in contact, and the material filling the fracture. Fractures in rock can affect the behavior of waves passing through them, including the time it takes for a wave to pass through the fracture and changes to the wave's amplitude. These effects depend on the fracture parameters mentioned above, and on the angle at which the wave is incident on the fracture and the frequency of the wave. The way that the fracture influences the wave can be described using reflection and transmission coefficients, as shown in Figure 1, which are also dependent on these factors. As the medium through which the wave is traveling (the rock surrounding the fracture) has properties that vary with frequency and angle of incidence, it is possible to study these properties using seismic experiments and infer information about the fracture parameters by measuring the velocity and attenuation of the wave at different frequencies and angles of incidence (Boadu, 2007).

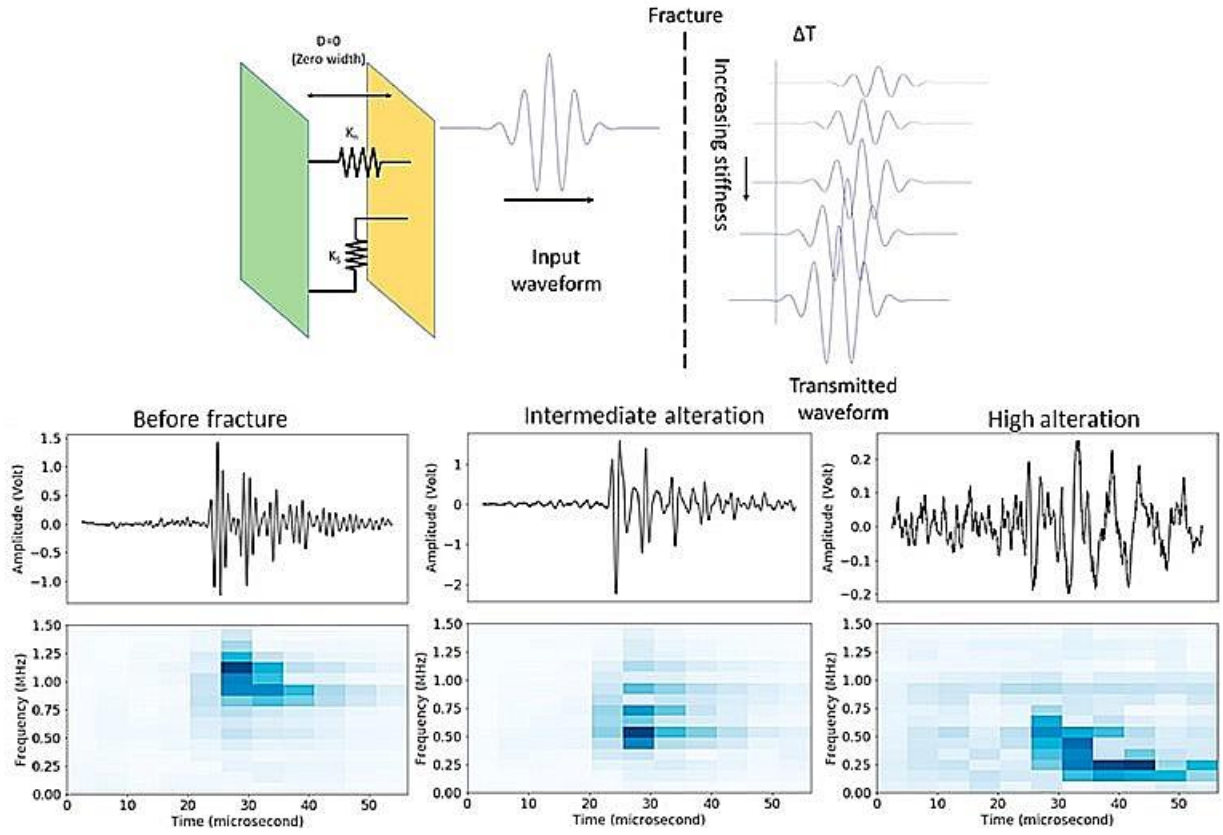

Figure 1: (Top) Schematic diagram of the displacement discontinuity model and the effect of fracture on the transmitted waveforms. (Bottom) Waveforms and corresponding short-time Fourier Transform spectrograms for transmitted waves after interaction with varying levels of fracturing damage encountered along the travel path.

2.2 Seismic energy radiation from crack displacement

When a force couple is applied to a small crack, the resulting displacement patterns are similar for both force couples. Therefore, it is difficult to distinguish between the two possible solutions when using observed waves to model the source, and the solution typically becomes a double-couple. The images below (Figure 2) show displacement results from a simple model of a small horizontal and vertical crack, with the arrows indicating the direction of the displacement.

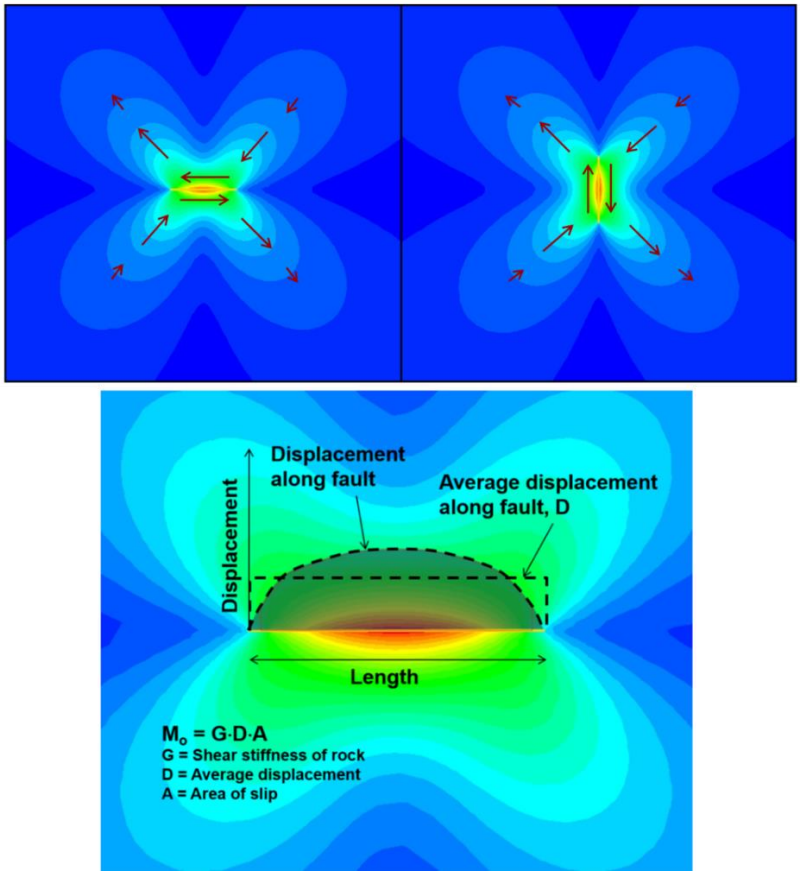

Figure 2: Crack model showing identical displacement pattern surrounding a vertical and horizontal crack (upper panel). (Lower panel) seismic moment calculation parameters highlighted. Hotter colors represent greater displacements (adapted from Beghini and Bertini, 1990).

The displacement field caused by a dislocation on a plane is like that produced by a double-couple. In a homogeneous and isotropic medium, the moment of a seismic event caused by a shear fracture on a plane can be represented by:

$M = G \times D \times A$ ..... (1)

Where M = Seismic moment, G = Shear stiffness, A = area of slip and D = displacement

The energy source parameter represents the energy that is radiated away from the source, rather than the total work done during the event. It is important to note that the elastic energy radiated

by a seismic event is just a fraction of the total work done by the source. A moment tensor is a way of representing the source of a seismic event. It is like the stress tensor, which describes the state of stress at a specific point. A moment tensor, on the other hand, describes the deformation at the source location that produces seismic waves (Figure 3).

The figure below (Figure 3) illustrates the similarity between stress and moment tensors. The moment tensor describes the deformation at the source based on generalized force couples, arranged in a 3x3 matrix. The matrix is symmetric, so there are only six independent elements (e.g., M12 = M21). The diagonal elements (e.g., M11) are called linear vector dipoles, which are equivalent to the normal stresses in a stress tensor. The off-diagonal elements are moments defined by force couples. To generate a moment tensor for a seismic event, it is necessary to use the Green's function. This function calculates the ground displacement recorded by the seismic sensor based on a known moment tensor (the "forward" problem). A moment tensor inversion is the process of using the inverse Green's function to determine the source moment tensor based on sensor data.

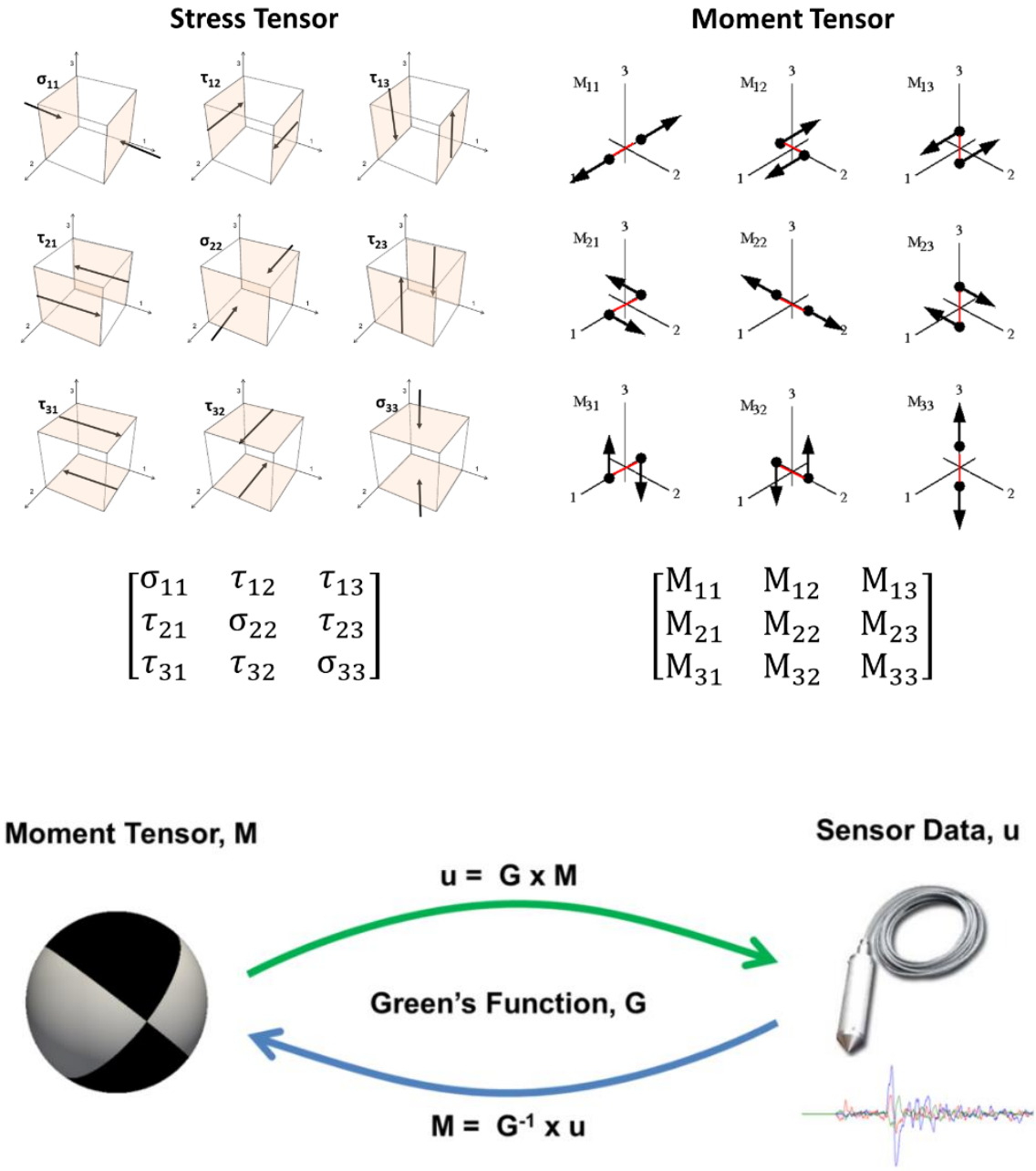

Figure 3: (above panel) Graphical representation of stress tensor and moment tensor and their respective matrix representations. (Bottom panel) Relationship between moment tensor, Greens function and the measured signal.

Focal mechanisms, which represent the type and orientation of faulting that occurred during an earthquake, can be calculated using a technique that tries to match the direction of P-wave arrivals recorded at each seismograph station. For a double-couple source mechanism (a type of

faulting that involves only shear motion on the fault plane), the first P-waves indicating compression should be in the quadrant containing the tension axis, while those indicating dilatation should be in the quadrant containing the pressure axis.

2.3 Moment tensor representation with beachball diagram

To construct a beach ball diagram, the moment tensor is used to determine the magnitude and direction of the first motion at each point on the surface of a sphere (Vavrychuk, 2011). Points where the motion is inward towards the source are colored white (red arrows), while points where the motion is outward away from the source are colored black (blue arrows). The border between white and black on the beach ball represents points where the motion is tangential (purple arrows), with the direction of motion across the border being white to black.

The figure following (Figure 4) illustrates the first ground motion on the beach ball surface, separated into radial and tangential components. The lengths of the radial and tangential arrows indicate the relative strength of the P and S waves, respectively. P-waves tend to be strongest in the middle of the white and black regions, while S-waves are strongest at the border between white and black.

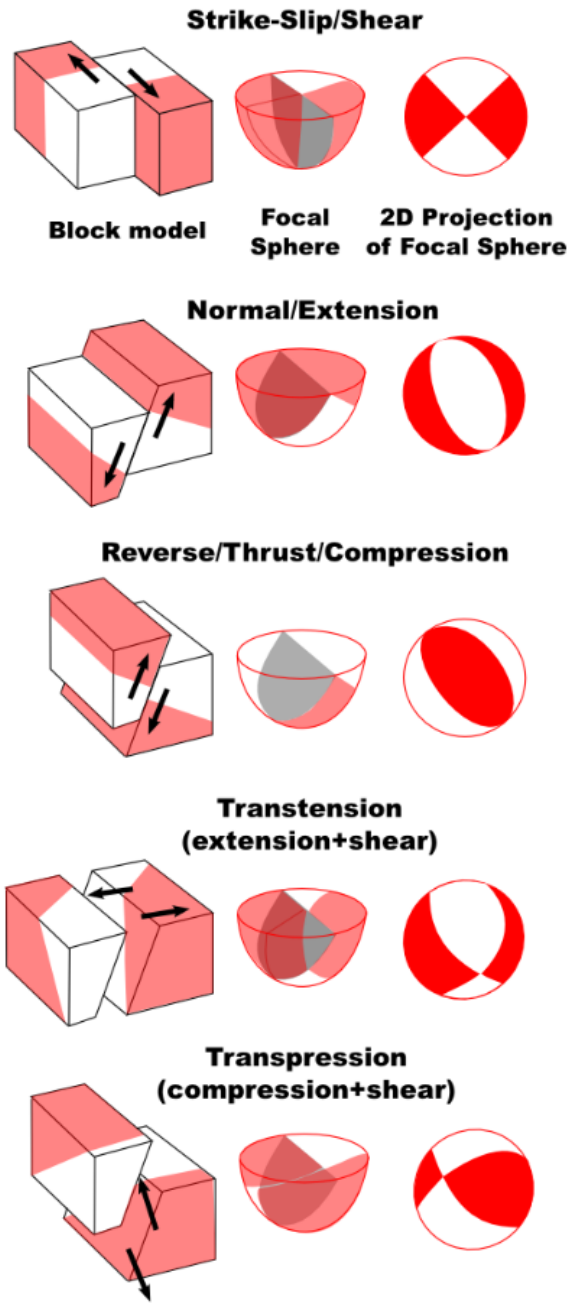

Figure 4: **Illustration of relationship between fault orientation and beachball diagram. In beach ball diagrams "red" holes radiates objects outward from the center of ball, while "white" pulls inward, towards the center.**

There is often ambiguity in determining the fault plane on which an earthquake occurred when calculating focal mechanisms using first-motion directions or certain waveform modeling

techniques. This is because there are two mathematically equivalent, orthogonal auxiliary planes that could represent the fault plane. Four examples are provided to demonstrate this ambiguity, with block diagrams showing the two possible types of fault motion that the focal mechanism could represent. The view angle for each diagram is 30 degrees to the left and above. The ambiguity can sometimes be resolved by comparing the two potential fault plane orientations to the alignment of small earthquakes and aftershocks. The first three examples involve purely horizontal (strike-slip) or vertical (normal or reverse) fault motion, while the fourth example involves fault motion with both horizontal and vertical components (oblique-reverse).

2.4 Caveats associated with Beachball representations of moment tensors

First P-waves may be recorded in the wrong quadrant. This can happen for several reasons: the algorithm may have misidentified the direction of the P-wave due to a lack of impulsive signal, the earthquake location and velocity model used to calculate the first P-wave arrivals may be incorrect, or the seismograph may be improperly wired such that up is recorded as down (which is rare). When calculating focal mechanisms using only first P-wave arrival directions, these incorrect observations can significantly affect the results. In some cases, multiple focal mechanism solutions may fit the data equally well depending on the quality and distribution of the first P-wave data.

2.5 Physical interpretation moment tensors and moment tensor decomposition

The moment tensor is a mathematical representation of the forces that caused an event, such as an earthquake. It can be difficult to interpret the geological or physical mechanism of the event based on the moment tensor alone. Therefore, the moment tensor is often decomposed into its

constituent elementary mechanisms by rotating the matrix to zero out the off-diagonal elements. This is like finding the principal axes of a stress tensor, which involves zeroing out the shear elements and leaving the normal stresses. As a result, every moment tensor can be expressed as three linear vector dipoles (orthogonal) that are rotated to a specific orientation. These dipoles are known as the P (pressure), B (neutral or null), and T (tension) principal axes.

2.5.1 Isotropic source

An isotropic source is a type of seismic source that produces waves that are equally strong in all directions. This means that the orientation of the P, B, and T axes has no meaning for an isotropic source, as the wave strength does not vary based on the orientation of the axes. An isotropic source only produces P-waves, which are pressure waves that travel through the Earth's interior. An isotropic source can be either an expansion or a contraction of the source volume. If the source is an expansion, such as an explosive event, the isotropic component is positive. This could be due to a confined blast or rock bulking. If the source is a contraction, such as an implosive event, the isotropic component is negative. This could be due to a pillar burst, buckling, or rock ejecting into a void. In the case of an implosive event, the first motion recorded by the waves will be towards the source, as they are traveling around a void.

2.5.2 Deviatoric Source

The deviatoric component of a moment tensor represents the part of the seismic source that causes displacement without changing the overall volume of the source. This means that there is an equal amount of movement in and out of the source. The deviatoric component is obtained by removing the isotropic component from the moment tensor. The deviatoric component is usually

caused by a general dislocation of a fault, which can involve both shear and normal displacement (although there is still no net volume change). To better understand the relative proportions of shear and normal displacement, the deviatoric component can be decomposed into two elemental sources: the DC (double couple) source and the CLVD (compound linear and volumetric dilation) source. The DC source represents a pure shear displacement, while the CLVD source represents a combination of shear and volumetric displacement.

2.5.3 Double Couple source

A double couple source refers to a type of source that is characterized by two pairs of forces that are equal in magnitude but opposite in direction and act along two orthogonal planes. The moment tensor of a double couple source has four independent parameters, which describe the magnitudes and orientations of the two pairs of forces. A double couple source generates seismic waves that are strongly directional, with maximum amplitude in two opposite directions. In contrast, an isotropic source generates seismic waves that are symmetric in all directions. The moment tensor of a double couple source mechanism is fully described by four independent elements of moment tensor matrix, that represent the magnitudes and orientations of the forces.

2.5.4 Compensated Linear Vector Dipole Source

A Compensated Linear Vector Dipole (CLVD) source is a type of deviatoric source that represents a normal displacement on a plane. The normal displacement from one linear vector dipole is compensated (hence the name) by opposing displacement from the other two linear vector dipoles, so there is no net volume change. An isotropic source is a seismic source that radiates seismic energy equally in all directions, and its moment tensor has a single nonzero

element, which represents the isotropic component of the seismic moment. In other words, an isotropic source is a spherically symmetric source that does not have any preferred direction of motion. On the other hand, a compensated linear vector dipole (CLVD) source is a seismic source that has a preferred direction of motion and exhibits a double-couple plus a compensated linear vector dipole pattern. The CLVD pattern describes the deformation of the seismic source that is related to the vertical stretching or squeezing of the earth's crust. The moment tensor of a CLVD source has four nonzero elements, which represent the double-couple component and the CLVD component of the seismic moment. The main difference between an isotropic source and a CLVD source in moment tensors is that the former represents a spherically symmetric source that does not have any preferred direction of motion, while the latter represents a source that has a preferred direction of motion and exhibits a deformation pattern related to the vertical stretching or squeezing of the earth's crust.

For a positive CLVD source, a single tensile dipole (stretching or pulling force) is compensated by two compressive dipoles (squeezing or pressing forces). A pure CLVD source would imply a Poisson's ratio of 0.5, a property shown by materials like chewing gum or toothpaste. There is no geological example of a pure CLVD source, but it can make sense as a mixed source event that includes both isotropic and CLVD components. This type of event mechanism may be dominant for confined pillar crushing events.

2.6 Hudson plot for representing moment tensor

Double-couple components of moment tensors can be represented using "beach balls," which show the orientation of the fault and the slip vector indicating the shear motion along the fault. Non-double-couple components of moment tensors are displayed in source-type plots. Moment

tensors occupy a "source-type space," which is a wedge in 3-dimensional space. The magnitude of the vector in this space is the scalar moment, and its direction indicates the type of source. To visualize the type of source, it is useful to plot the unit vectors of the source-type space in a 2-dimensional figure using certain projections. A source with pure or dominant shear faulting is located near the origin of coordinates on a source-type plot. An explosion or implosion source is located at the top or bottom vertex of the plot, respectively. Motion on a pure tensile or compressive crack is plotted at the margin of the plot. Points along the CLVD axis correspond to faulting on non-planar faults, and points in the first and third quadrants of the plot correspond to shear-tensile sources.

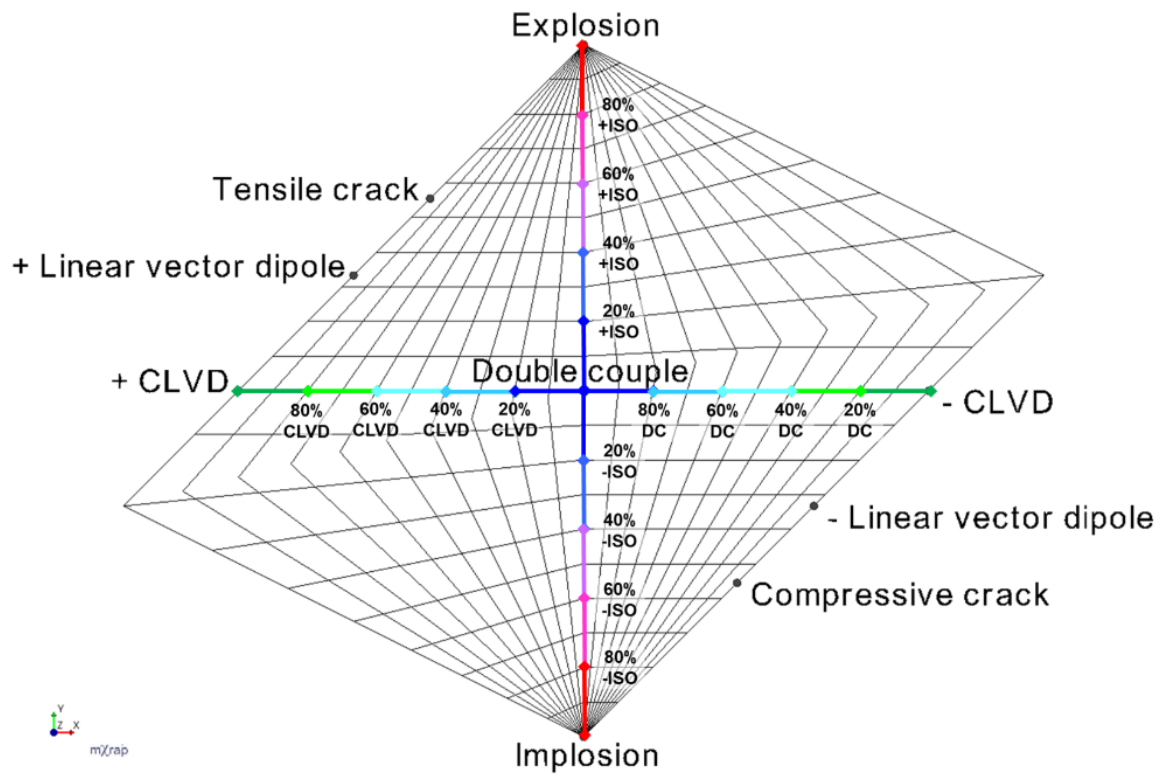

Figure 5: Hudson plot for representing moment tensors. See above text for explanation.

The Hudson plot can be used to visualize the decomposition of the moment tensor and see the relative proportions of the isotropic, DC, and CLVD elemental sources. The vertical axis of the chart represents the isotropic component, from -100% (implosion) to 100% (explosion). The horizontal axis represents the deviatoric decomposition, from +100% to -100% CLVD, with 100% DC in the center (0% isotropic, 0% CLVD). The outer border of the chart is the 0% DC line. The chart (Figure 5) can be used to understand the relative proportions of the different elemental sources in a seismic event.

3. LITERATURE REVIEW

3.1 Table 1: Summary of literature review of select key studies on the characterization of hydraulic fracture-induced seismicity at various length scales.

Author	**Method (scale)**	**Description**
Pyrak-Nolte et al., 1990	Seismic transmission (lab)	Seminal experiments on seismic transmission in dry and saturated fractured rock samples and influence on transmission coefficient and frequency spectra
Pater et al., 2001	Seismic transmission (lab)	Seismic transmission of lab scale sandstone sample post-fracturing to infer fracture extent by analyzing transmitted signal
Figuiredo et al., 2013	Seismic transmission (lab)	Effect of aligned fractures on the frequency of transmitted waves in synthetic samples
Damani et al., 2012	Scanning electron microscopy (lab)	SEM analysis of stimulated reservoir volume of Tennessee sandstone in micro and nano meter scale post fracturing
Bhoumick et al., 2017	Acoustic emission (lab)	Acoustic emission and shear wave transmission analysis of 6-inch samples both pre and post-fracturing
Leggett et al., 2022	Distributed acoustic sensing (lab)	Low-frequency DAS responses to propagating hydraulic fracture in lab scale plexiglass samples

Leggett et al., 2022	Distributed acoustic sensing (lab)	Thermal effects DAS responses to propagating hydraulic fracture in lab scale plexiglass samples
Kwatiek et al., 2018	Multi-modal sensing including active and passive seismic (meso)	Meso-scale (~10m) experimentation that allows dense instrumentation including active and passive seismic and rock core analysis in mine at depth of ~ 120 m
Ammann et al., 2018	Multi-modal sensing including active and passive seismic (meso)	Meso-scale (~10m) experimentation that allows dense instrumentation including active and passive seismic and rock core analysis at mine at depth of ~ 500 m
Schoenball et al., 2019	Multi-modal sensing including active and passive seismic (meso)	Meso-scale (~10m) experimentation that allows dense instrumentation including active and passive seismic and rock core analysis at mine at depth of ~ 1500 m
Neimz et al., 2021	Multi-modal sensing including active and passive seismic (meso)	Low frequency seismic signals associated with hydraulic fracturing
Boese et al., 2022	Multi-modal sensing including active and passive seismic (meso)	Low frequency seismic signals associated with hydraulic fracturing

Liu et al., 2020	Distributed acoustic sensing (field)	Stress and strain rate responses recorded by low frequency DAS for fracture propagation and fracture hit detection
Liu et al, 2021	Distributed acoustic sensing (field)	Algorithm and sensitivity analysis for hydraulic fracture width inversion from Low frequency DAS signals
Majer et al., 1997	Cross-well seismic survey (field)	Transmitter / receiver located in one or more of the wells to the seismically probe the region between two sensors.
Warpinski et al., 2001	Microseismicity (field)	Geometric analysis of hydraulic fracturing induced microseismicity
MacBeth, 2002	Seismic reflection (field)	Multi-component analysis for fracture induced seismic anisotropy
Burns et al., 2007	Seismic wave scattering (field)	Seismic wave scattering for estimating fracture geometry in field scale
Martinez-Garzon et al., 2017	Microseismic (field)	Analysis of moment tensors associated with fracturing-induced earthquakes in igneous and metamorphic rocks
He and Duan, 2019	Microseismicity (field)	Analysis of microseismic point clouds associated with fracture induced seismicity
Szafranski and Duan, 2022	Microseismicity (field)	Integrating machine learning and numerical simulation for analysis of microseismicity

Zhang et al., 2019	Microseismicity (field)	Moment tensors associated with field scale hydraulic fracturing in shale rocks
Chakravarty and Misra, 2021	Acoustic emission and seismic transmission (lab)	Wavelet fusion-based data fusion of active and passive seismic measurements in lab scale fractured rock sample.
Chakravarty et al., 2022	Multi-modal sensing including active and passive seismic (meso)	Low-frequency (2-80 Hz) seismic signal source location and correlation with microseismic and discrete fracture network.
Liu et al., 2021	Software simulation (milimeter scale)	Supervised classification of crack location and type using wavelet transform based features
Liu et al., 2022	Software simulation (milimeter scale)	Causal inference-based analysis of crack type and location using simulations
Liu et al., 2022	Software simulation (milimeter scale)	Supervised learning based on simulated signal analysis of multiple sensors for crack characterization
Jin and Misra, 2022	Software simulation (milimeter scale)	Modelling fatigue crack growth using reinforcement learning methods.

The summary from Table 1 indicates that considerable research has been done in the field of fracture characterization at length scales ranging from laboratory (millimeter) to field (kilometer). Majority of research is based on signals recorded on geophones. Recent advances in

the field of fiber optics and its commercialization have led to increased adoption of fiber optics for both research and commercial purposes. With fiber optics (DAS) the key quantity of interest is the strain field around the wellbore (at which the fiber optic cable is installed). Most of the conventional seismology related to seismic based fracture characterization is on reflection seismology. Relatively less focus is on transmission seismology, largely due to the increased complexity and increased cost of data acquisition.

3.2 Table 2: Summary of literature review of application of unsupervised machine learning for seismological data analysis

Author	Dataset	Description
Holtzmann et al., 2017	Microseismic (field scale)	Applied non-negative matrix factorization to extract features from a set of earthquakes recorded at a geothermal field and determined clusters which corresponded to distinct periods and rates of fluid injection
Mousavi et al., 2019	Regional-scale earthquakes	Used deep learning features based on earthquake spectrograms to distinguish between local and tele-seismic signals.
Bolton et al., 2019	Acoustic emissions (lab scale)	Applied clustering on statistical features of continuous acoustic emissions recorded during a laboratory-scale friction stick-slip experiment

Ross et al., 2020	Regional-scale earthquakes	Showed an unsupervised method of estimating directivity large populations earthquakes, using frequency spectra. Directivity is focusing of wave energy along a discontinuity in the direction of rupture
Watson et al., 2020	Volcano-induced seismicity	Volcanic earthquake signals have been studied using machine learning to interpret the signals associated with different stages of eruptive cycles. Time-domain and statistical features were reduced in dimension using principal component analysis and k-means clustering applied to assign labels.
Johnson et al., 2020	Regional-scale seismicity	Used the spectral characteristics of continuous geophone signal combined with k-Means clustering to determine five types of signals recorded over the San Jacinto fault. They concluded that the non-tectonic signals primarily consist of distinct type of noise. The area under study was isolated and thus recoded minimal anthropogenic signals.
Chakravarty et al., 2021	Lab scale shear wave transmission data	Showed that machine learning methods can improve fracture imaging from measurements in laboratory-scale

		experiments using acoustic emissions and ultrasonic transmission.
Shi et al., 2021	Regional earthquakes	Used array-signal processing features to obtain the covariance matrix-based features like entropy, coherency, and variance to determine clusters in the principal component space and showed that the clusters were well correlated to the temporal evolution of the events.
Chakravarty et al., 2022	Meso-scale (~ 10 m) fracturing-induced seismicity data	Applied polarization analysis to three component accelerometer data to delineate different hydraulic fracture planes in a microseismic point cloud.

The summary from Table 2 indicates that very limited research exists in the domain of unsupervised machine learning in seismology. Most of the work done in the domain of unsupervised learning in seismology is associated with phenomena occurring in regional or field (kilometer) scale where the researchers have benefit of well characterized field information and high signal to noise ratio. Specifically, no reference exists in the field of unsupervised machine learning pertaining to signals from hydraulic fracturing. Chakravarty et al., 2021 was in the authors opinion, the pioneering study in the field of unsupervised machine learning on seismic signals associated with hydraulic fracturing.

4. DESCRIPTION OF EXPERIMENTS AND DATA[*]

Key points in this chapter:

1. Three sets of experiments are considered in this dissertation, each of a characteristic length scale. The first is a laboratory scale hydraulic fracturing setup which measured acoustic emissions and wave transmission.

2. The next experiment is a meso-scale (~ 10 m) hydraulic fracturing experiment conducted at depth of 1.5 kilometer. Accelerometer and hydrophone data are analyzed.

3. The last set is a field (kilometer) scale hydraulic fracturing in Duvernay Shale formation.

This research work analyses of data acquired from three separate experiments – all of which involve hydraulic fracturing. The length scale of the experiments spans six order of magnitude – ranging from millimeter-scale laboratory experiments to field scale hydrofracturing experiments. This section will cover the details of experimental setup and data acquisition.

4.1 Laboratory scale uniaxial hydraulic fracturing setup

This analysis is based on the measurements first reported by Bhoumick et al. (2017). Experiments were performed on two cylindrical Tennessee sandstone core blocks of dimensions: length 154 mm and diameter 152 mm. The schematic of the setup is shown in Figure 6. The plane containing the circular face of core block is termed the axial plane. The wellbore is designed perpendicular to axial plane. The two perpendicular planes containing the wellbore are

[*] Reprinted with permission from "Visualization of hydraulic fracture using physics-informed clustering to process ultrasonic shear waves" by Chakravarty, A., Misra, S., and Rai, C. S., *International Journal of Rock Mechanics and Mining Sciences, 137*, 104568. Copyright 2021 by Elsevier.

termed as the frontal planes. The experimental parameters and sample details are summarized in Table 3.

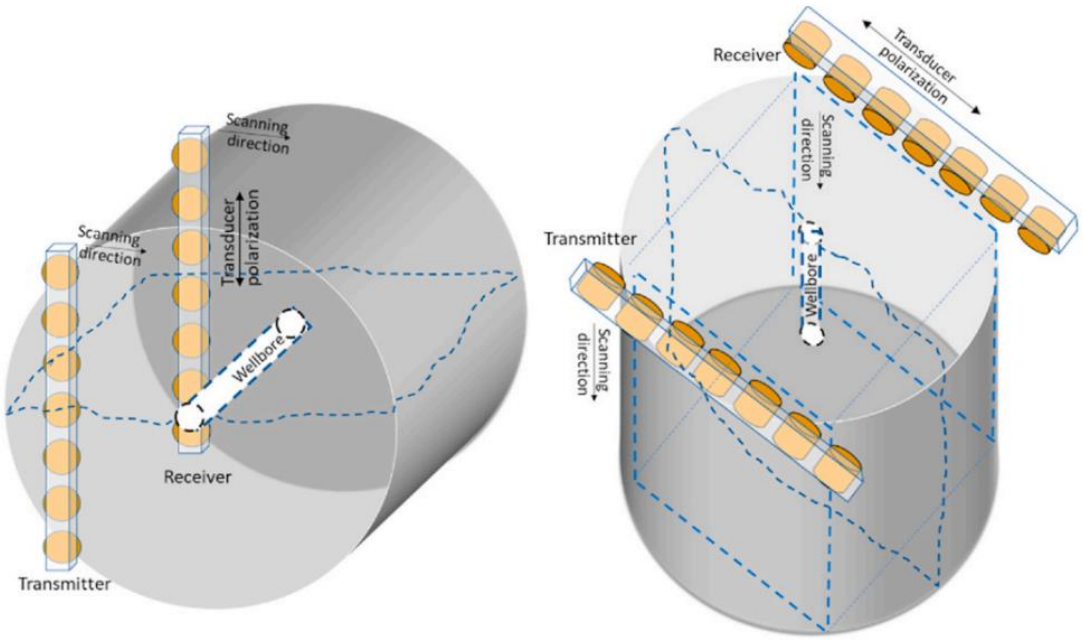

Figure 6: Schematic of transmitted shear waveform measurements in axial (left) and frontal (right) directions. Dotted lines in the right-side figure mark the portion of sample clipped to enable the scans. The irregular dotted contour in the figures is the expected outline of the primary hydraulic fracture.

The first step is circumferential velocity analysis to measure the P-wave velocity that corresponds to the direction of velocity anisotropy. In the next step, shear transmission waveforms were measured along axial and frontal planes using seven source/receiver pairs, as

Table 3: Experimental parameters associated with the laboratory scale setup

Sample Name	TSU6	TSU1
Data Available	SW transmission	SW transmission

		Acoustic emission	X-ray CT
Sample properties			
Length (mm)		154	NA
Diameter (mm)		152	NA
Porosity (%)		9.7	NA
Permeability (mD)		13	NA
Composition (wt %)		Quartz/Clay is 9:1	Quartz/Clay is 9:1
Experimental parameters			NA
Stress (psi)		870	NA
Injection rate (cc/min)		15	NA
Fracturing fluid		Water	Water
Breakdown pressure (psi)		2764	NA
Injection depth (mm)		80	NA
Borehole Depth (mm)		83	NA

shown in Figure 6. To ensure consistent sample-sensor contact between, the assembly is pressed on to the sample using air-driven actuators and honey is used as coupling medium. Before hydraulic stimulation, the transducers scan along the axial orientation of the sample (Fig. 6, left). The same scan is redone after fracturing. To facilitate shear wave transmission in the frontal orientation, flat surfaces are generated on the sample by clipping two portions of the sample of

length 0.5 inches from each side (Fig. 6, right).

All the seven transducer pairs touch the sample when in axial orientation, while only five sensor receiver pairs touch the frontal plane. The scan is performed at 1mm intervals that results in 133 measurement points along each of the two planes. Before fracturing, data is collected only in the axial orientation, and after fracturing data is collected for both axial and frontal directions. Two identical Tennessee sandstone samples, (TSU6 and TSU1) were analyzed in our study (Table 3). TSU6 has all the analyses described above but no X-ray tomography, and TSU1 only has axial transmission and acoustic emission and X-ray tomography.

4.2 Meso-scale hydraulic fracturing setup

The test site is the Sanford Underground Research Facility (SURF) located at Leads, South Dakota (Figure 7) located at a depth of 1500 meters. The data used in this analysis is from EGS Collab project (Schoenball et al., 2020). The setup comprises of six monitoring boreholes, one production and one injection wellbore. The monitoring boreholes contain active seismic sources, three-component accelerometers, hydrophones (pressure transducers), electrical resistivity probes, fiber optics - distributed temperature sensing (DTS), distributed strain sensing (DSS) and distributed acoustic sensing (DAS) and all instrumentation are cemented in place. On the injection well, notches were etched at target depths to encourage the fracture initiation from a well-defined location. Present analysis is on the data measured during the stimulation of a notch at depth of 50 m (from the wellhead) in the well E1-1 between 22 May and 24 May 2018. The experimental details are summarized in Table 4.

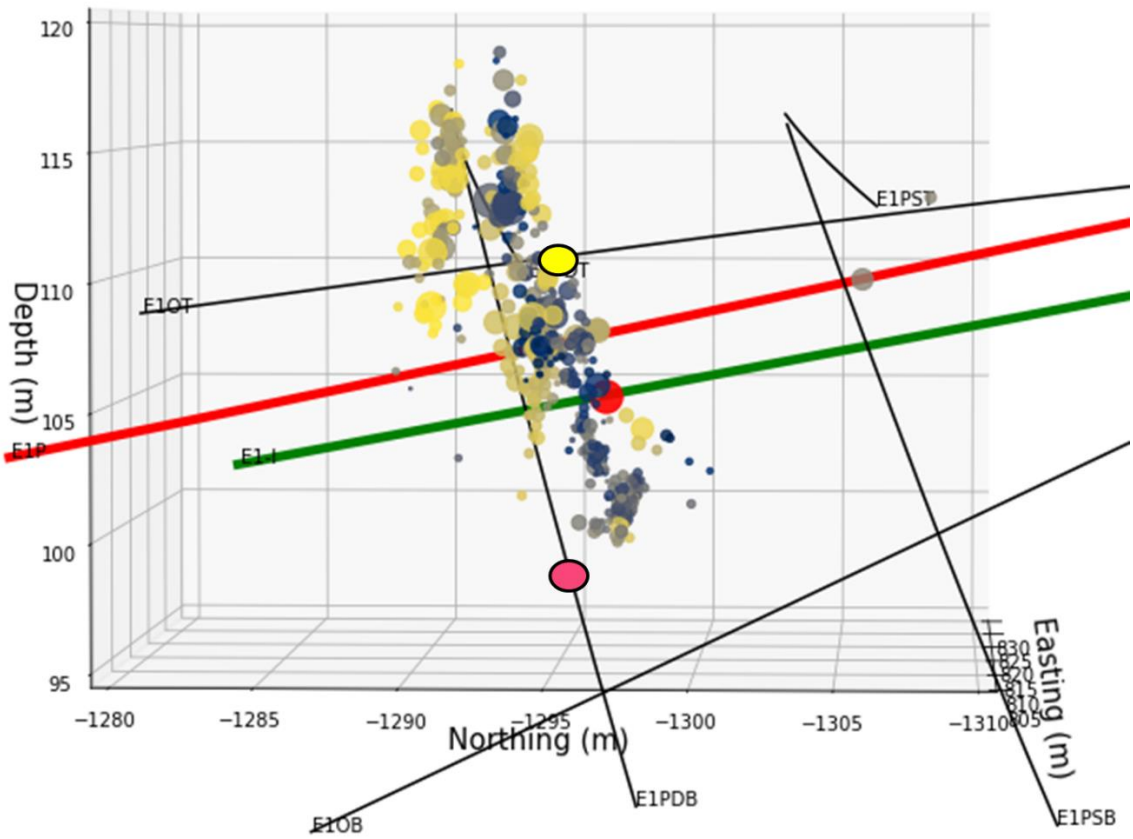

Figure 7: Location of the EGS Collab experiment 1 testbed at the Sanford Underground Research Facility in Leads, South Dakota. Facility structure and the layout of the injection well (green), production well (red) and monitoring boreholes. Passive seismicity measured during stimulation on the notch 50 m depth.

The continuous passive seismic signal is recorded by a set of accelerometers and pressure transducers installed in six monitoring boreholes. The raw data recorded by the arrays consists of 32-second-long continuous signals with a frequency of 100 kHz. There is a gap of 1.5 seconds between two sets of measurements. After removing artefacts associated with active seismic the signal is filtered between 3 kHz to 10 kHz to isolate the range of interest.

Table 4: Hydraulic protocol associated with the EGS Collab experiment 22 May to 24 May

Date	Injection interval	Located events	Maximum injection rate	Maximum injection pressure	Cumulative injected volume	Description
22-May-2018	E1-I, 50 m	37	0.2 L/min	26.0 MPa	2.2 L	Create nominally 1.5 m fracture
23-May-2018	E1-I, 50 m	129	0.4 L/min	26.8 MPa	23.6 L	Propagate fracture to nominally 5 m
24-May-2018	E1-I, 50 m	296	5 L/min	27.3 MPa	80.8 L	Propagate fracture to intersect E1-P

4.3 Field scale hydraulic stimulation setup

The data analyzed in this study originates from Tony Creek dual Microseismic Experiment (ToC2Me) which is a research-oriented field test conducted by the University of Calgary between October and December 2016 (Zhang et al., 2019). The data analyzed comes from three component geophones (Trillium Compact) buried at depths of 27 meter. The broadband sensors have a flat frequency response between 20 second and 100 Hz and the sampling frequency is 500 Hz. The data was collected from sensors and formatted into 60 second intervals in SEG2 file format. A sledgehammer source was used to calibrate the orientation of the three component geophones Individual events were detected using a cross correlation-based method (Eaton et al., 2017). The raw data used in this study are the filtered microseismic events which have a high signal to noise ratio and an associated moment tensor. The schematic of the layout is shown in Figure 8.

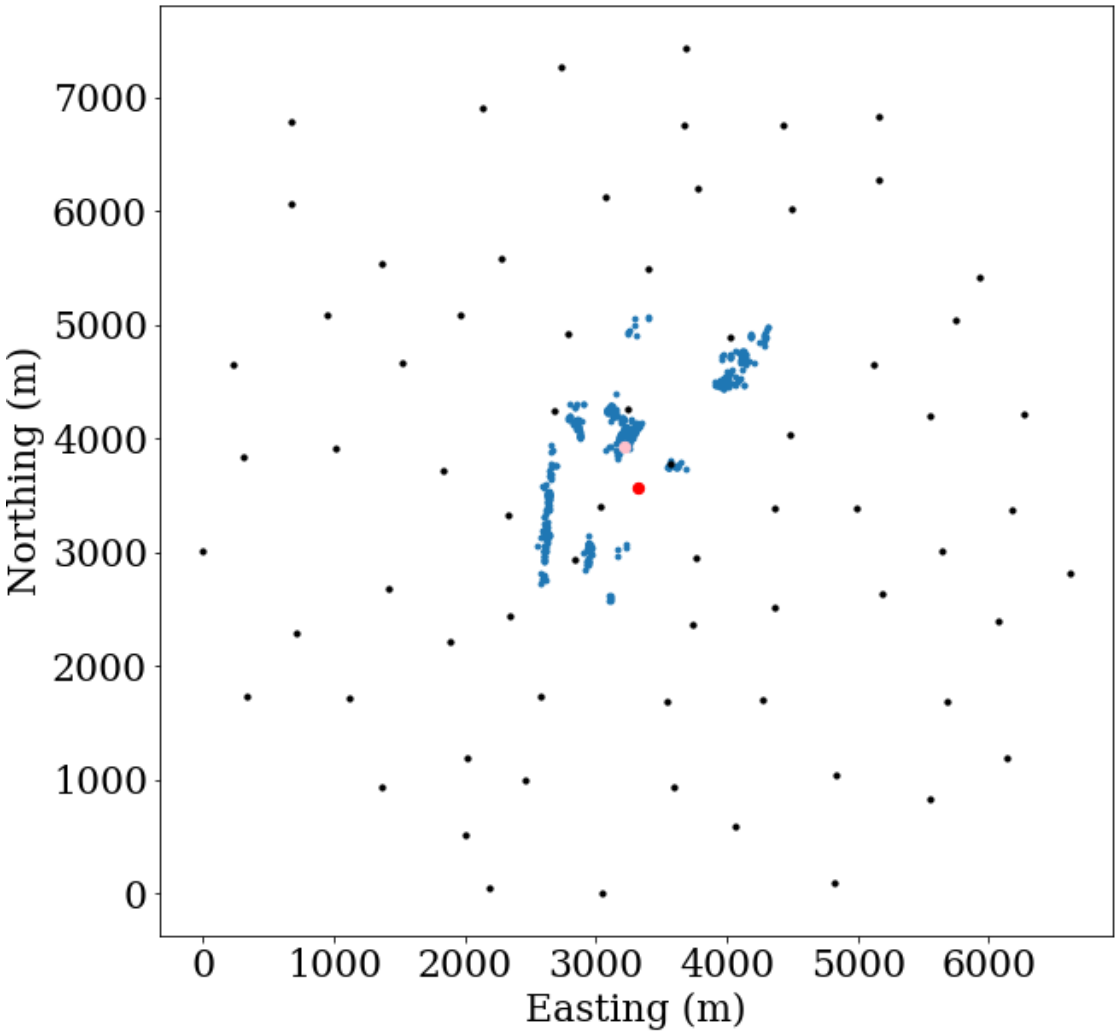

Figure 8: Schematic of the field scale hydraulic stimulation setup in Duvernay shale, Canada. The exact location is proprietary and undisclosed. Black dots represent geophones buried at 27-meter depth and blue dots represent the passive seismicity.

5. DATA PROCESSING METHODS AND UNSUPERVISED MACHINE LEARNING WORKFLOWS[*]

Key points in this chapter:

1. Lab scale: feature extraction from time frequency (short time Fourier transform) of signals. Workflow for obtaining labels that indicate fracture damage.

2. Meso-scale: Polarization feature extraction, clustering methods, clustering metrics, dimensionality reduction using UMAP, validation of UMAP representations using Wasserstein distances and Tau statistic. Low frequency signals source location. Filters for refining low frequency seismic source locations.

5.1 Lab-Scale Experimental Data Processing

In this study, the spectral energy density of the transmitted waveform was calculated using the coefficients of the short-time Fourier transform (STFT) spectrogram (examples of which are shown in Figure 9). To account for inconsistencies in identifying the first arrival of the stress wave that may be caused by scattering and reflection, the authors used the time of arrival of the first peak of the spectral energy as a surrogate. They also introduced a parameter called "J" to combine the effects of arrival time and transmission coefficient based on displacement-discontinuity theory. The J parameter is calculated as the ratio of the transmission coefficient to the arrival time of the first peak of spectral energy. The J parameter is found to decrease as the

[*] Reprinted with permission from "Visualization of hydraulic fracture using physics-informed clustering to process ultrasonic shear waves" by Chakravarty, A., Misra, S., and Rai, C. S., *International Journal of Rock Mechanics and Mining Sciences, 137*, 104568. Copyright 2021 by Elsevier.

specific stiffness of the fractures decreases, which is a direct indicator of the geomechanical alteration induced by hydraulic fracturing of the transmission zone.

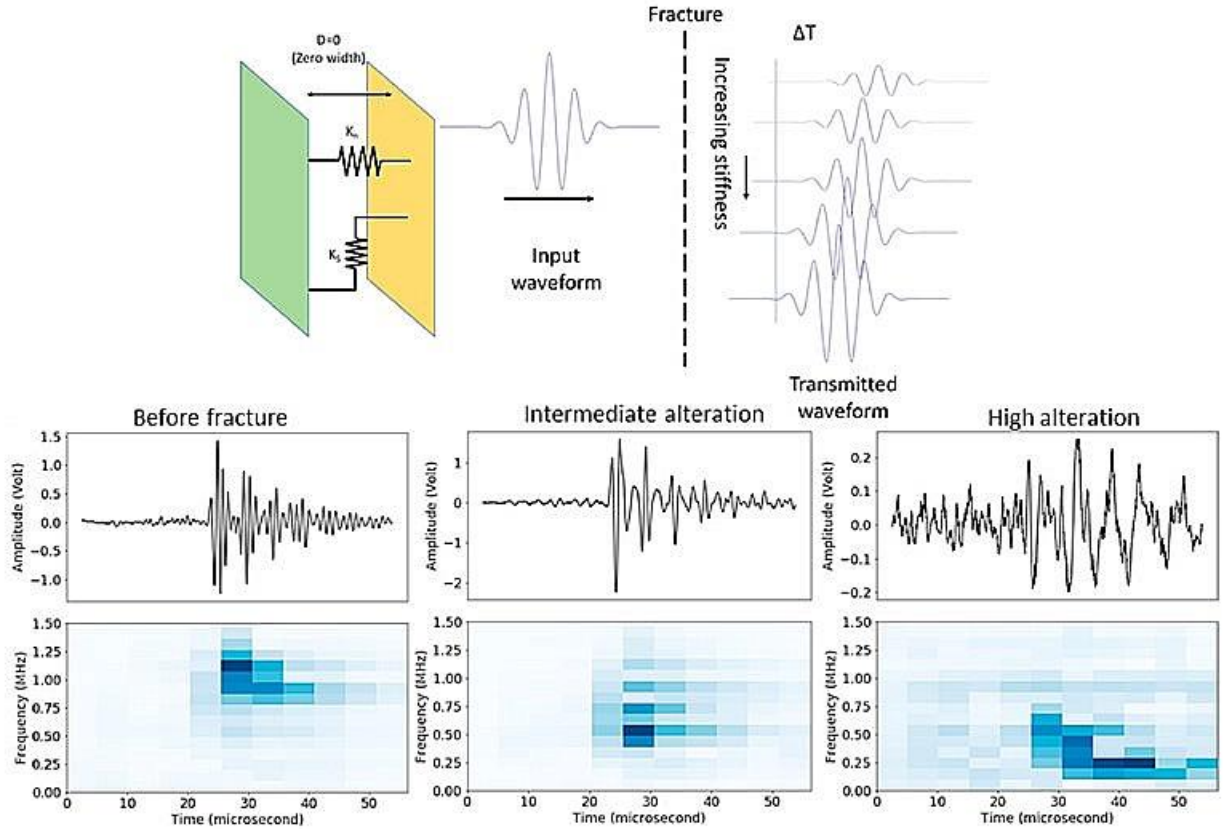

Figure 9: (Top) Schematic diagram of the displacement discontinuity model and the effect of fracture on the transmitted waveforms. (Bottom) Waveforms and corresponding short-time Fourier Transform spectrograms for transmitted waves after interaction with varying levels of fracturing damage encountered along the travel path.

To interpret the results of the clustering analysis, the authors assigned a geomechanical alteration index (GAI) to each cluster based on the median and range of J parameter values in the cluster. A GAI of 1 indicates the least alteration, which corresponds to high values of the J parameter, while progressively higher GAIs indicate higher levels of alteration represented by lower values of the J parameter. However, the authors caution that the J parameter should not be used on its

own to assess damage, because it does not consider the entire waveform, while the clustering analysis considers the entire waveform when grouping the data from different locations.

The process for obtaining physically consistent geomechanical alteration indices is described in Figure 10 as follows:

1. Transmitted shear wave signals are collected from samples before and after hydraulic fracturing.
2. The short-time Fourier transform (STFT) of the transmitted shear waveform features is extracted and subject to scaling and dimensionality reduction.
3. The processed data is input into a clustering method to identify clusters.
4. The identified clusters are made statistically consistent by using cohesion, separation, and silhouette scores to determine the optimal number of clusters.
5. The optimal clusters are assigned a physical meaning based on the J parameter, which combines the effects of arrival time and transmission coefficient based on displacement-discontinuity theory.
6. Geomechanical alteration indices (GAIs) are assigned to the clusters based on the median and range of J parameter values in the cluster, with a GAI of 1 indicating the least alteration and progressively higher GAIs indicating higher levels of alteration.

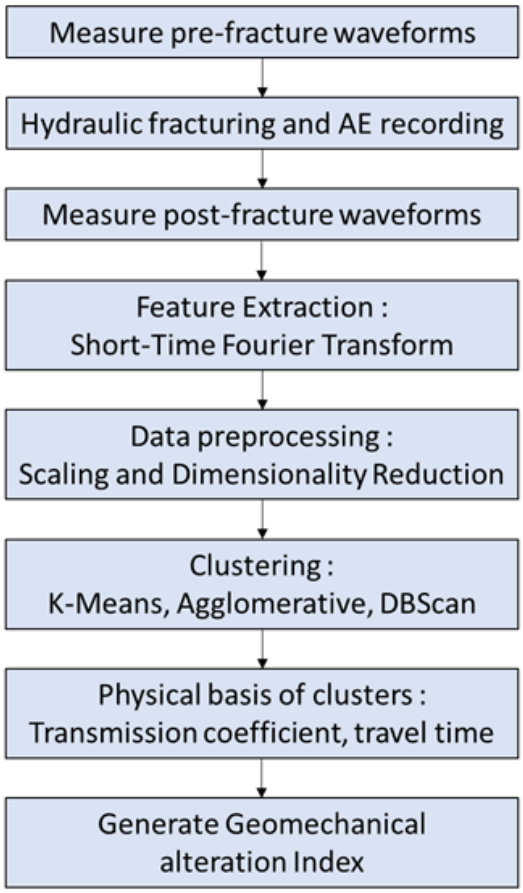

Figure 10: Workflow for quantifying the spatial distribution of geomechanical alteration due to hydraulic fracturing.

5.2 Meso-Scale Experiment Data Processing

The Obspy package is a widely used open-source software library for processing and analyzing seismic data in Python. It provides a range of tools and functions for handling and manipulating seismic data, including tools for filtering, detection, and visualization, and has been used for seismic data processing in this section.

Table 5: Hydraulic stimulation protocol under study for infrasound, stimulation carried out at the notch at 50-meter depth on the injection well E1-I.

Day (2018)	Injected volume	Description
24 May	75 L	Hydraulic fracturing
25 May – p1	77 L	Flow through fracture
25 May – p2	121 L	Flow through fracture

The short-time-average (STA) is a measure of the average amplitude of a seismic signal over a short time period, typically a few tens of milliseconds. The long-term-average (LTA) is a measure of the average amplitude of a seismic signal over a long time period, typically a few seconds or longer. The STA/LTA filter compares the STA to the LTA and looks for instances where the STA exceeds the LTA by a certain threshold. This indicates the presence of an impulsive signal or "trigger." The length of the triggers, in this case 8 ms, is an important parameter that determines the sensitivity of the STA/LTA filter. A shorter trigger length will result in a higher sensitivity but may also result in more false positives. The margins at the start and end of the triggers are adjusted to maintain a uniform size for each trigger, which can help with the consistency and accuracy of the analysis.

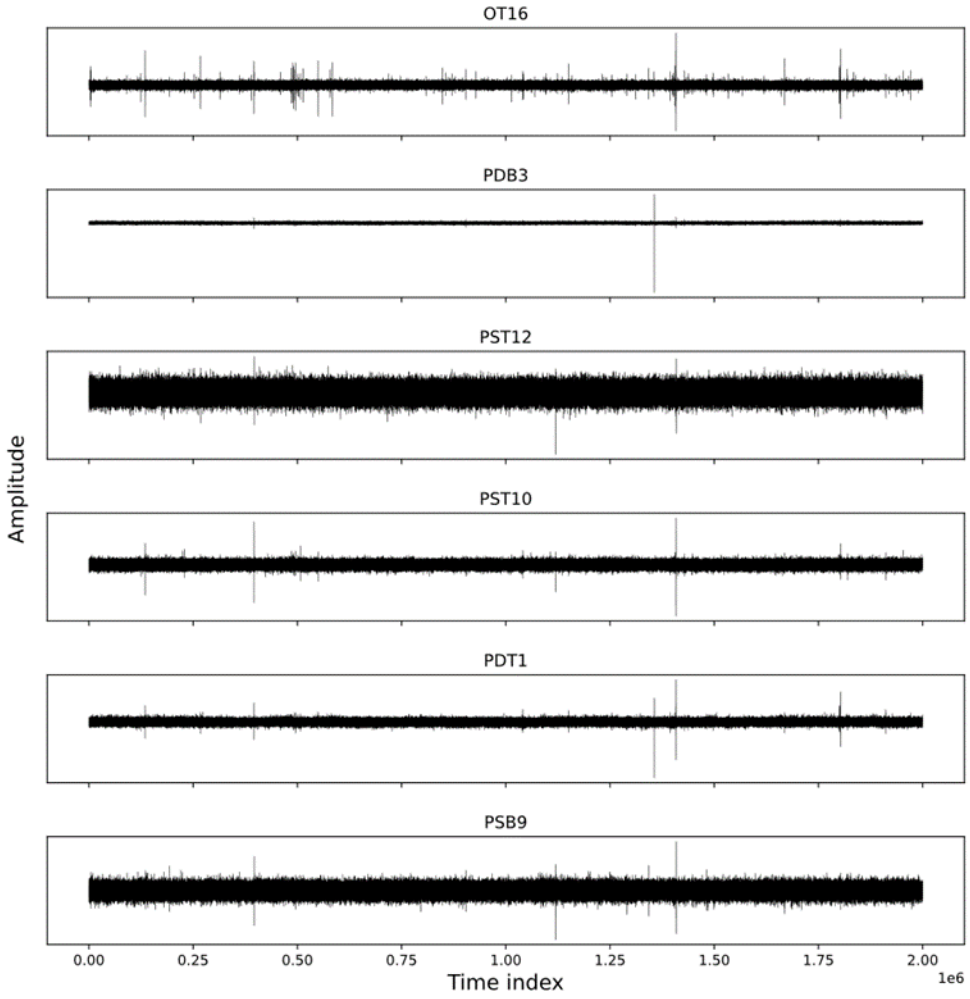

Figure 11: X-component signal of different three-component accelerometers showing different sensitivity of each accelerometer. OT16 has the highest signal to noise ratio.

The signal-to-noise ratio (SNR) is a measure of the relative strength of a signal compared to the noise present in the data. It is defined as the ratio of the energy of the signal to the energy of the noise. A higher SNR indicates a stronger signal relative to the noise, which can make it easier to detect and analyze the signal. In this case, the SNR is being calculated for triggers identified using the STA/LTA filter. The SNR is calculated as the ratio of the energy of each trigger to the energy of a noise sample of the same length. The SNR values for the different accelerometers show that one of the accelerometers (OT16) has a significantly higher SNR compared to the

others (Figure 11 and 12). This may be due to a higher sensitivity to ground motion or better coupling with the ground, which can result in a stronger signal. As a result, the analysis is focused on the OT16 sensor.

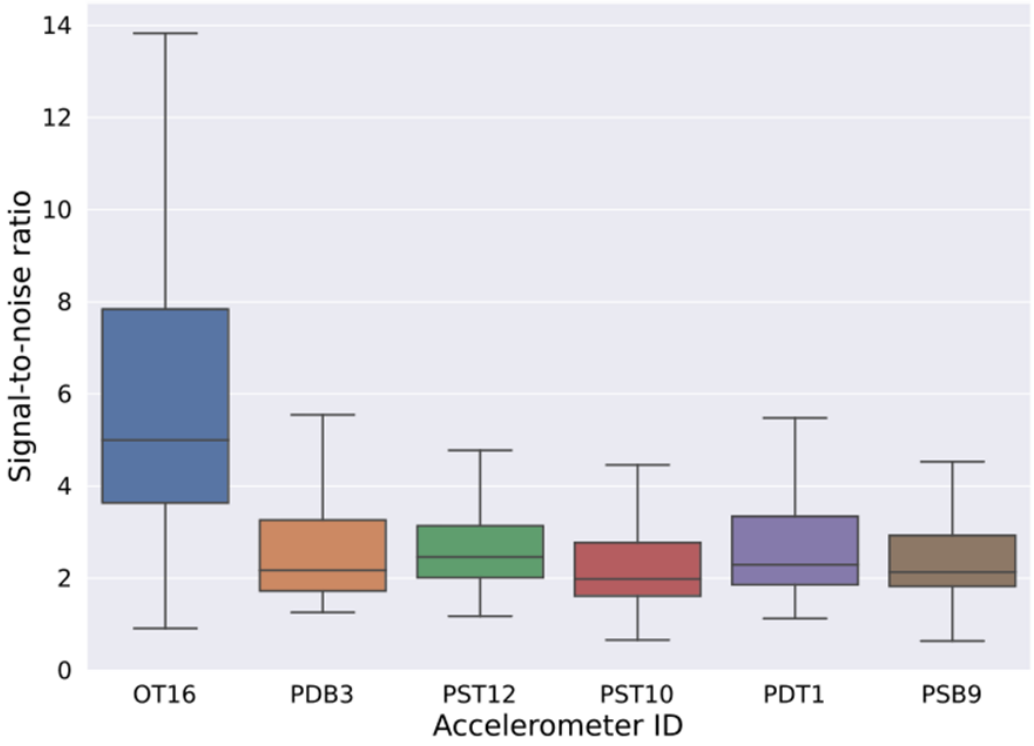

Figure 12: Difference in signal to noise ratio (SNR) of various accelerometers under consideration. The SNR of accelerometer OT16 is significantly higher than the other accelerometers.

The hodogram is a scatter plot that shows the relationships between the three components of a seismic signal, typically the acceleration in the x, y, and z directions (Figure 13). It can be used to visualize the polarization of the signal and understand the source mechanism of the seismic event. To extract features from the hodograms, the first step is to compute the covariance matrix of the three-component trigger signal. The covariance matrix is a 3x3 matrix that describes the correlations between the different components of the signal. It is useful in this case because

scattering distortions and seismic noise are usually uncorrelated among the three components, which can help with the analysis of noisy signals.

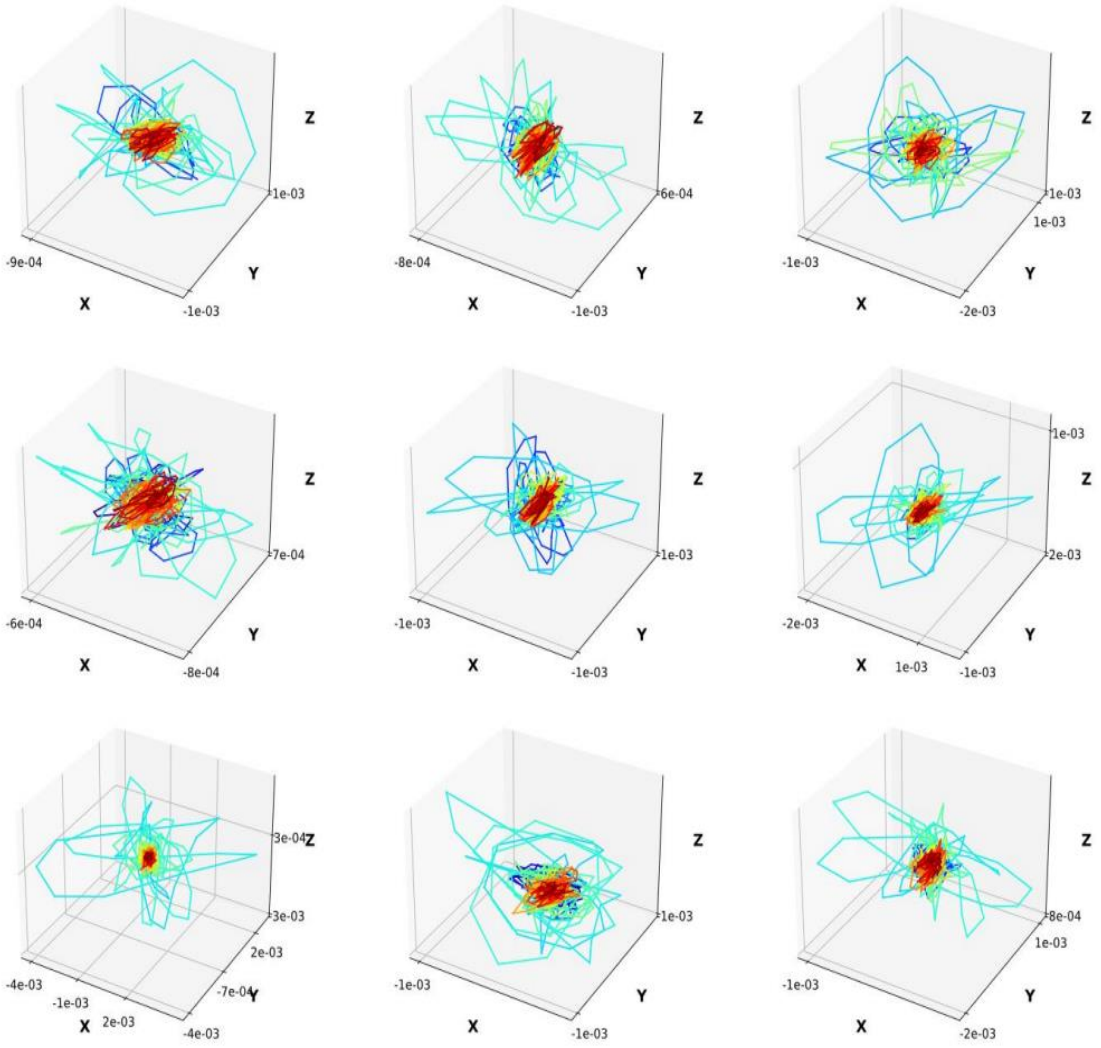

Figure 13: **Hodograms of triggers from accelerometer OT-16. Cooler colors indicate early time and hotter colors indicate later times.**

The covariance matrix can be factorized to yield the eigenvalues and eigenvectors, which can be used to define various polarization parameters. The eigenvalues are typically represented as λ1, λ2, and λ3, and the eigenvectors are represented as u1, u2, and u3. These parameters can be used

to describe the orientation, shape, and size of the hodogram and provide insight into the source mechanism of the seismic event.

Using the eigen values and vectors, polarization features can be defined as follows:

Azimuth $= arctan\left(\frac{u_{21}}{u_{11}}\right)$ ..... (1)

Incidence = $arccos\left(\sqrt{u_{11}^2 + u_{21}^2}/u_{31}\right)$ ..... (2)

Rectilinearity = $1 - \sqrt{\frac{\lambda_1}{\lambda_2}}$ ..... (3)

Planarity = $1 - \frac{2\lambda_3}{\lambda_1+\lambda_2}$ ..... (4)

In summary, the process described involves using the STA/LTA filter to detect impulsive signals or "triggers" in a continuous stream of seismic data. Each trigger has a duration of 8 ms and consists of 800 timesteps of data. Four features are extracted from each trigger using the hodogram and the covariance matrix of the three-component acceleration data. These four features are then used as inputs for unsupervised learning. In this case, the data from May 22 was used, resulting in a feature matrix with 815 rows (one for each trigger) and four columns (one for each feature). The feature matrix is then scaled using MinMax scaling to optimize the performance of unsupervised methods. MinMax scaling is well suited for this type of data because the features are either bound between 0 and 1 or between -180 and +180. MinMax scaling scales the data to a fixed range, typically between 0 and 1, which can help with the consistency and accuracy of the analysis.

5.3 Dimensionality Reduction: Uniform Manifold Approximation Projection

Manifold learning is a type of machine learning that involves projecting high-dimensional data onto a lower-dimensional manifold or surface. It is often used when there are non-linear relationships within the data and the goal is to represent the distribution of the data in a lower-dimensional space. The UMAP (Uniform Manifold Approximation and Projection) algorithm is a popular manifold learning method that can be used to project high-dimensional data onto a lower-dimensional manifold (McInnes et al., 2018). The main steps in the UMAP algorithm involve generating a graph of the data and finding a low-dimensional representation that optimizes an objective function. The main inputs for the algorithm include the target data for dimension reduction, the number of neighboring samples for localized approximation, the dimension of the reduced space, a layout control parameter (min-dist), and the number of optimization steps for the graph layout. The UMAP algorithm is based on a set of equations that govern the construction of the graph and the optimization of the objective function. These equations involve concepts such as the weights of the edges in the graph, the distances between data points, and the embedding of the data in the lower-dimensional space. Detailed descriptions of these equations and the mathematics behind the UMAP algorithm can be found in the original UMAP paper (McInnes et al., 2018). Key equations governing UMAP are as follows:

$$p_{i|j} = e^{-\left(\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right)} \quad \text{..... (5)}$$

$$k = 2^{(\sum_i p_{ij})} \quad \text{..... (6)}$$

$$q_{ij} = (1 + a(y_i - y_j)^{2b})^{-1} \quad \text{..... (7)}$$

$$CE(X,Y) = \sum_i \sum_j \left[p_{ij}(X) \log\left(\frac{p_{ij}(X)}{q_{ij}(X)}\right) + \left(1 - p_{ij}(X)\right) \log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right] \quad \text{..... (8)}$$

Eq (5) describes an exponential probability distribution that is used to measure the distances between pairs of points in the high-dimensional feature space (X). The distance between the points (Y) is a function of the distance of the i-th data point from its nearest neighbor (ρ).
Eq (6) defines the number of nearest neighbors (k) used in the UMAP algorithm. This value is used to construct a weighted k-neighbors graph of the data, which serves as the basis for the low-dimensional representation.

Eq (7) defines a family of curves that is used to model the distance probability in the low-dimensional space. These curves are based on the distances between points in the high-dimensional space and the number of nearest neighbors (k). Eq (8) defines the binary cross entropy (CE) loss function, which is used to project the data from the high-dimensional space (X) onto the lower-dimensional manifold. The loss function helps to optimize the low-dimensional representation of the data by minimizing the distance between the data points and their corresponding points on the manifold.
It's important to note that the number of nearest neighbors and the minimum distance metric used in the UMAP algorithm are not based on the physical distances between the data points in three dimensions, but rather on the distances on the nearest-neighbor graph derived from the data using the UMAP algorithm. This allows the UMAP algorithm to capture the underlying structure and relationships within the data, even when the data is non-linear or has complex relationships.

5.4 Clustering Methods

5.4.1 K-Means clustering

K-Means clustering is a widely used unsupervised machine learning algorithm for dividing a dataset into a predefined number of clusters (Everitt et al., 2011). It works by iteratively assigning each data point to the closest cluster centroid and then updating the cluster centroids based on the mean of the points in the cluster. The main steps in the K-Means algorithm are:

1. Initialize K cluster centroids randomly.
2. Assign each data point to the closest cluster centroid.
3. Update the cluster centroids to the mean of the points in the cluster.
4. Repeat steps 2 and 3 until convergence is reached (i.e., the cluster centroids do not change significantly).

The quality of the clustering solution is determined by the sum of the distances between the data points and their corresponding cluster centroids. The goal of the K-Means algorithm is to minimize this sum and obtain clusters that have minimum variance around the centroids.
K-Means is a fast and efficient algorithm for clustering large datasets, but it can be sensitive to the initial choice of cluster centroids and may not always produce the optimal solution. It is also limited to identifying clusters that are spherical in shape and equally sized.

5.4.2 Agglomerative clustering

Agglomerative clustering is a type of hierarchical clustering algorithm that works by iteratively merging the closest pairs of clusters until all the data points are in a single cluster (Everitt et al., 2011). It is a bottom-up approach, meaning that it starts by considering each data point as a

separate cluster and then progressively combines them into larger clusters. The main steps in agglomerative clustering are:

1. Initialize each data point as a separate cluster.
2. Calculate the distance between all pairs of clusters using a distance metric.
3. Merge the two closest clusters.
4. Update the distance matrix.
5. Repeat steps 2-4 until all the data points are in a single cluster.

The distance metric and linkage criteria used in agglomerative clustering determine how the clusters are merged and grouped. The most common distance metrics are Euclidean distance, Manhattan distance, and Cosine similarity, and the most common linkage criteria are single-linkage, complete-linkage, and average-linkage. Agglomerative clustering is a flexible method that can identify clusters of different shapes and sizes, but it can be computationally expensive for large datasets.

5.4.3 DBSCAN

DBSCAN (Density-based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that is used to identify clusters in a dataset (Everitt et al., 2011). It works by identifying dense regions in the feature space and marking the points within these regions as belonging to the same cluster. Points that lie in low-density regions are marked as noise. The steps in DBSCAN are:

1. Choose a point at random from the dataset and determine its density.

2. If the point is in a dense region, it is marked as a core point. All other points in the same region are also marked as core points.

3. All points that are reachable from the core points (i.e., within a certain distance, known as the Eps value) are added to the same cluster.

4. Repeat steps 1-3 until all points have been processed.

DBSCAN has two main parameters: Eps and MinPts. Eps is the maximum distance between two points to be considered in the same cluster, and MinPts is the minimum number of points required to form a cluster. DBSCAN is well suited for identifying clusters of different shapes and sizes, and it can handle datasets with noise and outliers. It does not require the user to specify the number of clusters in advance, but it can be sensitive to the choice of Eps and MinPts values.

5.4.4 Mean Shift clustering

Mean shift clustering is a non-parametric, unsupervised machine learning algorithm that is used to identify clusters in a dataset (Everitt et al., 2011). It works by shifting the data points towards the mean of the points in their local neighborhood until convergence is reached. The points that end up at the same location after the mean shift process become part of the same cluster. The mean shift algorithm has several parameters that can be adjusted to influence the clustering process. These include the kernel function (which determines the shape of the local neighborhood around each data point), the bandwidth (which controls the size of the neighborhood), and the convergence threshold (which determines when the mean shift process is complete). Mean shift clustering is well suited for data that is non-linearly distributed and has multiple modes (or clusters) in the feature space.

5.5 Clustering Metrics

5.5.1 Silhouette score

The silhouette score is a measure of the quality of a clustering solution (Rousseeuw, 1987). It is based on the concept of cohesion, which refers to the similarity of the data points within a cluster, and separation, which refers to the distance between different clusters. The silhouette score for a data point (s of datapoint i) is defined as the difference between the mean distance to the other points in the same cluster (a(i)) and the mean distance to the points in the nearest cluster (b(i)), divided by the maximum of these two values. Silhouette score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\left(b(i),\, a(i)\right)} \quad \text{..... (8)}$$

Cohesion is defined as the mean dissimilarity of the data point *i* with all other points in the same cluster. Low cohesion correlates with low values of intra-cluster distance *a*. Separation is the lowest mean dissimilarity of a data point *i* to other points in the clusters to which the datapoint *i* does not belong. High separation corresponds to high values of mean nearest cluster distance *b*. For every data point *i*, the intra-cluster distance is denoted as *a* and the mean nearest cluster distance is denoted as *b*. An effective cluster has high separation and low cohesion for each sample and consequently a high silhouette score. To ensure the statistical consistency of the clusters, it is important to determine the optimal number of clusters using a clustering method. The silhouette score can be used to evaluate the quality of a clustering solution and identify the optimal number of clusters. An effective clustering should have low cohesion and high separation, which will result in a high silhouette score. The silhouette score can be used to compare different clustering solutions and select the one with the highest score, which will generally correspond to the optimal number

of clusters. Figure 14 below outlines the workflow for part 1 of Meso-scale seismicity analysis that aims to delineate the distinct fracture planes starting from the unlabeled microseismic point cloud.

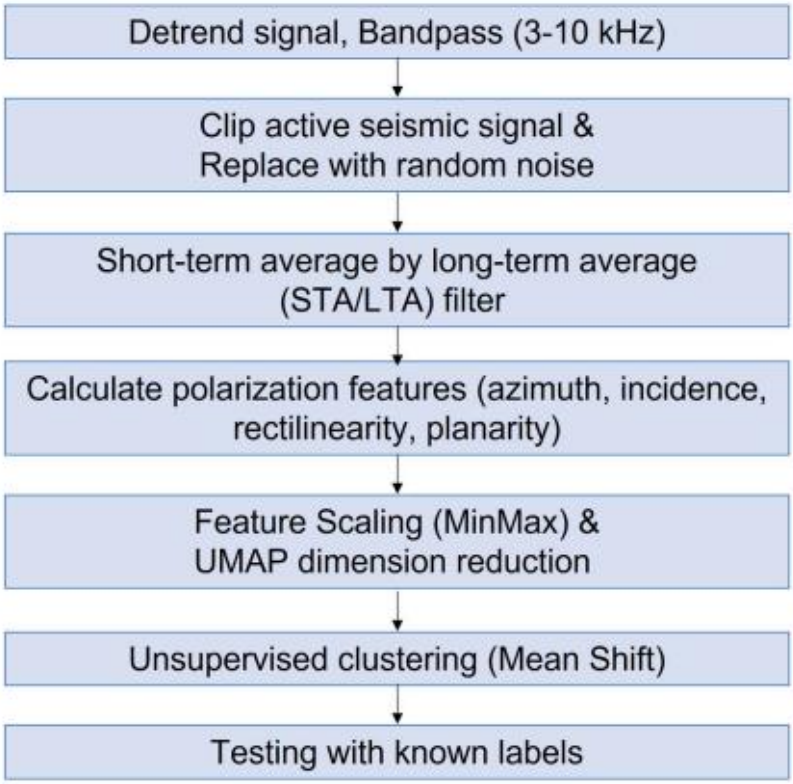

Figure 14: Unsupervised learning workflow from continuous passive seismic data

5.5.2 Calinski-Harabasz Index

The Calinski-Harabasz index, also known as the variance ratio criterion, is a measure of the quality of a clustering solution (Everitt et al., 2011). It is based on the ratio of the sum of between-cluster dispersion to the sum of inter-cluster dispersion for all clusters. Between-cluster dispersion refers to the variance within each cluster, while inter-cluster dispersion refers to the variance between the clusters. The Calinski-Harabasz (CH) index is calculated as:

Calinski-Harabasz index = $\frac{B(k-1)}{W(n-k)}$ ..... (9)

where B is the sum of between-cluster dispersion, k is the number of clusters, W is the sum of inter-cluster dispersion, and n is the total number of data points. The higher the Calinski-Harabasz index, the better the performance of the clustering solution. This measure can be used to compare different clustering solutions and identify the one with the highest index, which will generally correspond to the optimal number of clusters.

5.6 Calculating Distances Between Clusters of Different Sizes Using Wasserstein Distance

The Wasserstein distance, also known as the earth mover's distance, is a measure of the distance between two probability distributions (Ni et al., 2009). It is based on the idea of moving mass from one distribution to another, with the cost of moving each unit of mass equal to the distance it needs to be moved. The Wasserstein distance is defined as the minimum cost of moving mass from one distribution to the other. The Wasserstein distance is a popular measure in machine learning and data analysis because it can capture the underlying structure of the distributions, even when the distributions are very different. It is often used in image processing and natural language processing applications to measure the similarity between two distributions, such as the distributions of pixel intensities in an image or the distribution of word frequencies in a document. The Wasserstein distance can be expressed as follows:

Let u_1 and u_2 be probability measures and their cumulative distribution functions be $F_1(x)$ and $F_2(x)$. Optimal transport preserves the order of probability mass elements, so the mass at quantile q of u_1 moves to quantile q of u_2. The p-Wasserstein distance between u_1 and u_2 is expressed as

$$W_p(u_1, u_2) = (\int_0^1 | F_1^{-1}(q) - F_2^{-1}(q)|^p)^{1/p} \quad \text{..... (10)}$$

where F_1^{-1} and F_2^{-1} are the quantile functions (inverse cumulative distribution function). When p= 1, the formula becomes

$$W_p(u_1, u_2) = \int |F_1(x) - F_2(x)|dx \quad \text{..... (11)}$$

The Wasserstein distance (W_p) has several desirable properties, such as being able to handle discontinuities and outliers, and being able to capture subtle differences between distributions. It is also a metric, meaning that it satisfies the triangle inequality, which makes it well suited for use in clustering and classification tasks.

5.7 Quantifying Order of Distances Between Cluster Pairs Using Kendall-Tau Statistic

Concordant/discordant pairs can be defined as following: for two sets X and Y, if $X_i > X_j$ given $Y_i > Y_j$ for every i and j, then pairs are said to be concordant, else discordant.

The Kendall-Tau statistic is a measure of the ordinal association between two variables (Kendall, 1970). It is based on the concept of concordance, which refers to the degree to which the variables are ranked in the same order. The Kendall-Tau statistic is defined as the number of pairs of observations (x, y) for which x is ranked higher than y in one variable and y is ranked higher than x in the other variable, divided by the total number of pairs of observations.

The Kendall-Tau statistic is a non-parametric measure that is often used to evaluate the strength of an ordinal relationship between two variables (Kendall, 1970). It is particularly useful when the data is not normally distributed or when the relationship between the variables is not linear.

The Kendall-Tau statistic can take on values between -1 and 1, with values close to 1 indicating a

strong ordinal relationship between the variables, and values close to -1 indicating a strong inverse ordinal relationship. A value of 0 indicates no ordinal relationship between the variables.

The Kendall-Tau statistic is widely used in statistics and data analysis to evaluate the strength of an ordinal relationship between two variables. It is a robust measure that is not sensitive to the distribution of the data and can handle missing values. The tau statistic, τ [Kendall, 1970] is defined as the difference of concordant and discordant pairs to the total number of pairs.

$$\tau = \frac{Number\ of\ concordant\ pairs}{Number\ of\ discordant\ pairs} \quad \text{..... (12)}$$

All pairs perfectly concordant, tau statistic = 1; All pair perfectly discordant, tau statistic = -1

Figure z below summarizes the part 2 workflow based on UMAP that attempts to refine the representation learning by optimizing model hyperparameters and correspondence with input data.

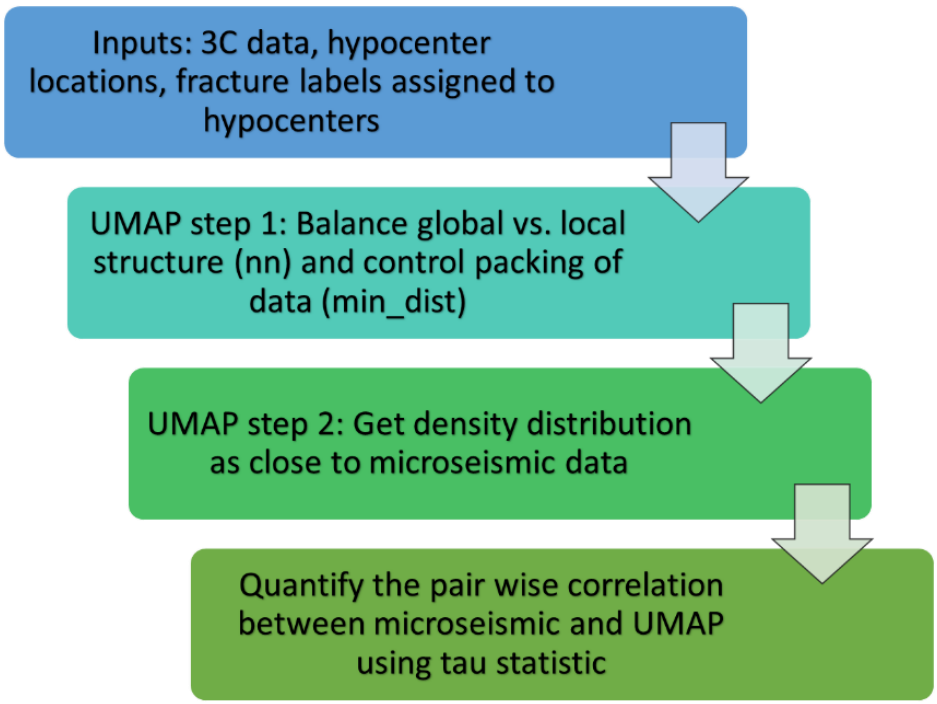

Figure 15: **Workflow for UMAP-based representation learning from microseismic data.**

5.8 Mesoscale Infrasound Processing and Source Location

To locate the source of emergent signals, such as tremor, a cross-correlation based grid search approach is used (Wech and Creager, 2008). This involves comparing the observed travel time lag, calculated from the cross-correlation of signals from different stations, with the calculated theoretical time lag between the stations using an input velocity model. The workflow is outlined in Figure 15. The source location is determined as the grid node with the minimum misfit between the observed and calculated time lags. The input for the algorithm includes hydrophone signals, a grid (as presented in Figure 16), and a velocity model. The grid is defined based on the extent of the hydrophone network, with an extension of 30% in both directions. The velocity model used is an isotropic velocity of 5.5 km/second for the compressional wave. The window length and overlap are important parameters in this process, and a window length of 1 second

with a 0.5 second overlap is used. The duration of the emergent signals is difficult to determine accurately due to uncertainty in detecting the first arrivals. To estimate the pulse duration, the STA-LTA filter is applied to a sample of hydrophone data, and an average value of one second is obtained as the pulse duration of the infrasound signals. This information is used to set the window length and overlap for the analysis.

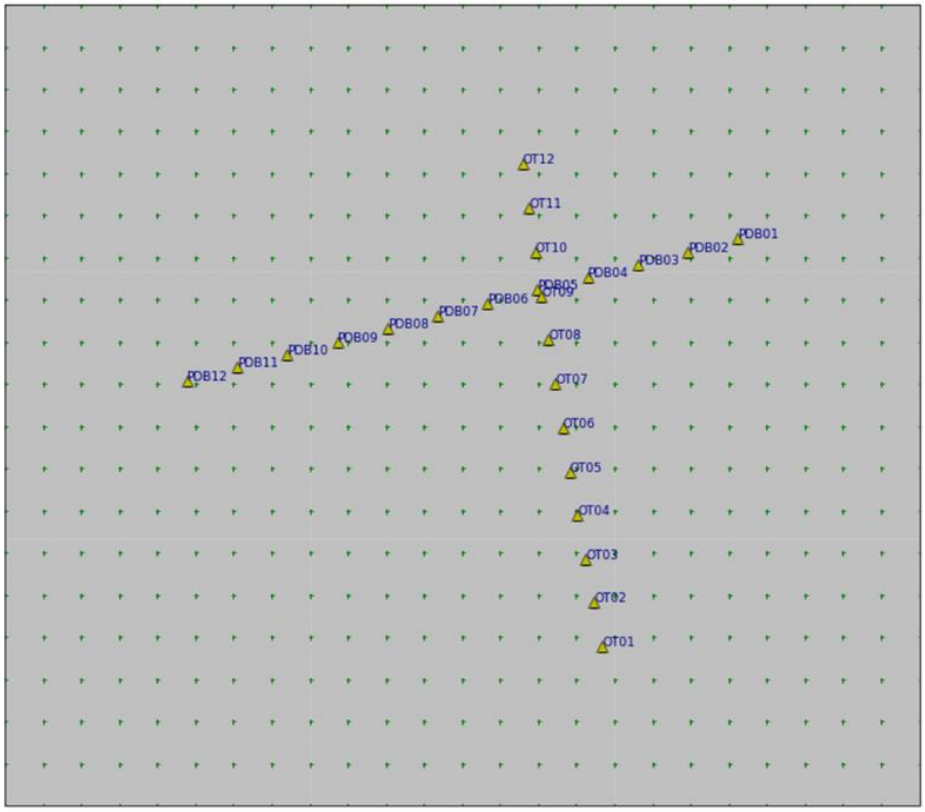

Figure 16: **Grid for the cross correlation-based infrasound location workflow**.

5.8.1 Postprocessing Steps on the Grid Search Output

To improve the accuracy of the cross correlation-based grid search location technique described in the paper, which is data-driven, filters are applied to remove false positives from the results. The filtering process includes the following steps:

1. To remove false positives from the results of the location technique, two filters are applied based on the normalized cross-correlation (CC) coefficient. The first filter removes signals with extremely high CC coefficients, which are likely to be correlated noise. The upper bound for this filter is set at 0.95. The second filter removes signals with very low CC coefficients, which are likely to be uncorrelated noise. The lower bound for this filter is set at 0.6. Windows with CC coefficients outside of these bounds are discarded to improve the accuracy of the location technique. This helps to eliminate false positives that may be caused by correlated noise or uncorrelated noise in the data.

2. The second filter applied to remove false positives from the location technique is based on the array beam power (Kvaerna and Doornbos, 1985). The relative power of the hydrophone array is calculated using the same window lengths and overlaps as used in the location algorithm. The resulting timestamps are then compared to the timestamps produced by the grid search-based location. Locations that have a normalized beam power lower than the noise floor of the beamforming output are discarded. A threshold value of 0.03 is used to effectively differentiate between located and non-located timestamps. This helps to remove locations that may have a relative power lower than the noise floor, which may be false positives. The differences between the beam power of the retained and discarded timestamps are shown in Figure 19.

3. The third filter applied to remove false positives from the location technique is based on bootstrapping. For each timestep, 20 iterations are performed using the cross-correlation-based locations. In each iteration, 5% of the cross correlograms are randomly removed and the resulting scatter is used as a measure of location uncertainty. The data points with the highest 10% of scatter values are discarded (Figure 20). These points represent locations showing the most scatter in the determined locations and are likely to be false positives.

4. The final filter applied to remove false positives from the location technique is based on the misfits obtained in the grid search algorithm. The misfit is defined as the difference between the maximum normalized cross-correlation function and the cross-correlation function corresponding to the located grid node. A large misfit indicates weak support from the modeled time lag (derived from the cross correlation) with the observed time lag. 50% of the data showing the highest misfit values is discarded. Despite losing half the data, the spatial coverage of the source locations shows little change, demonstrating the effectiveness of the misfit filter (Figure 21 and 22). This filter helps to improve the accuracy of the location technique by removing locations with high misfits, which may be false positives.

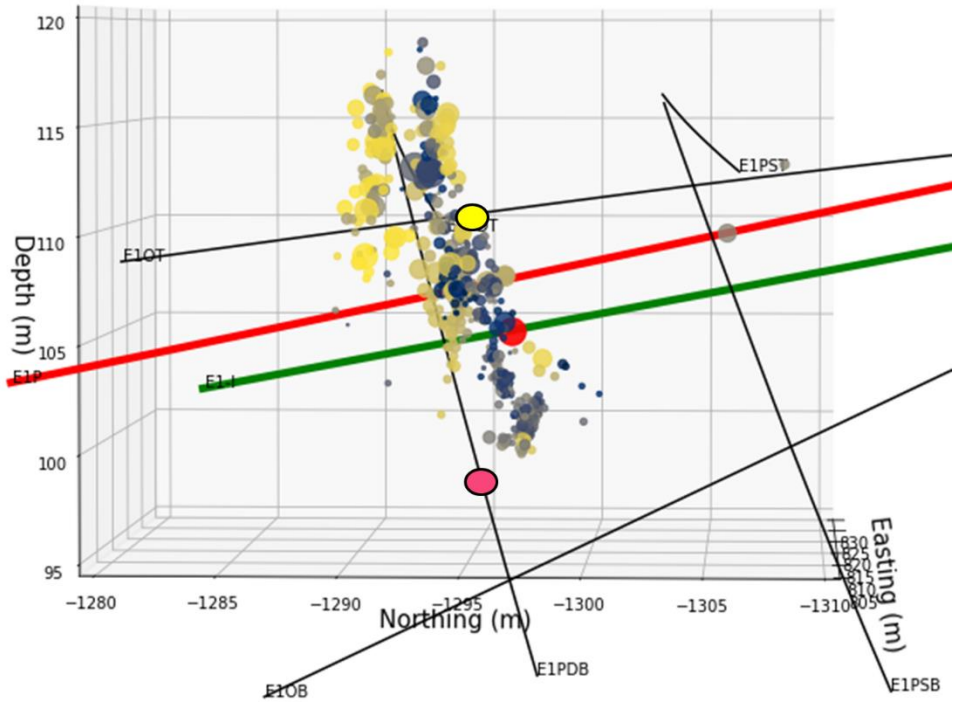

Figure 17: Relative orientation of wells E1-OT (yellow circle) and E1-PDB (pink circle) microseismic cloud with the injection and production wells in green and red respectively.

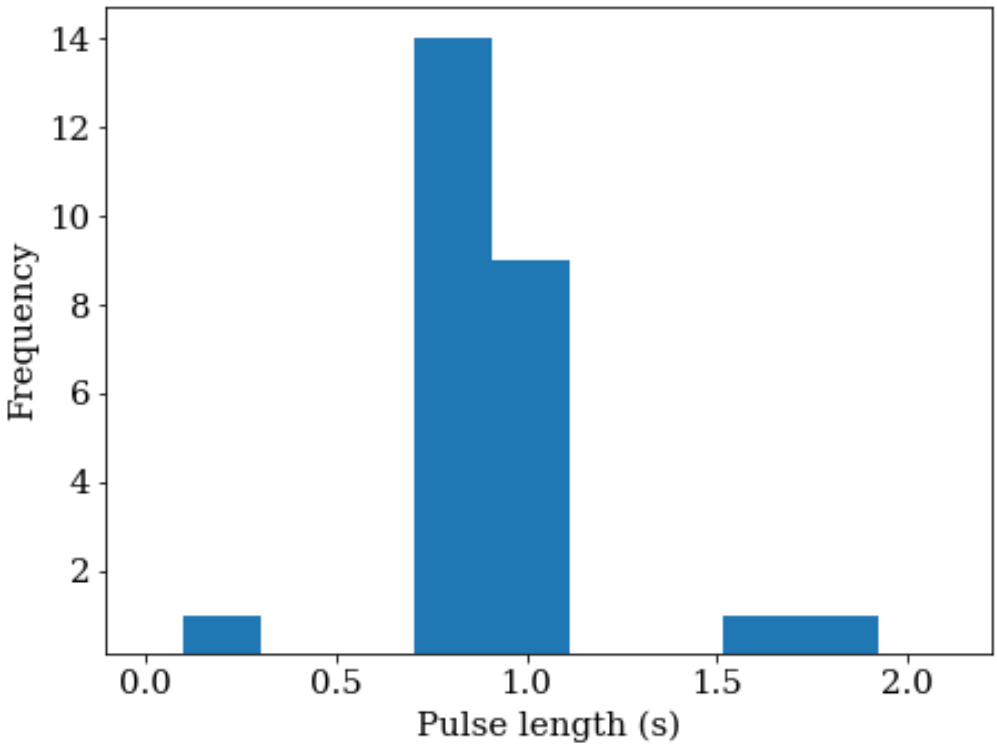

Figure 18: Histogram of infrasound pulse durations for 24 May hydrophone OT-03 obtained by applying STA LTA filter.

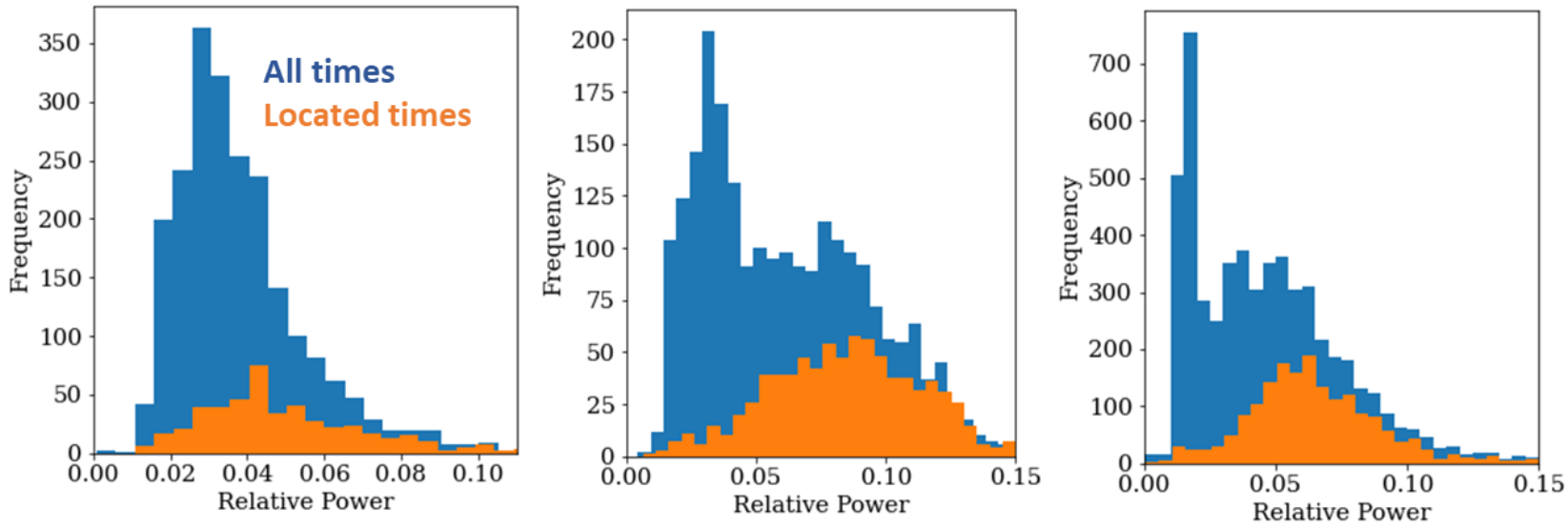

Figure 19: Histograms of beam power values showing variation of power values between located times (orange) and all data (blue) for 24 May (left), 25 May p1 (center) and 25 May p2 (right). A threshold of 0.03 is determined as the noise floor from beam relative power.

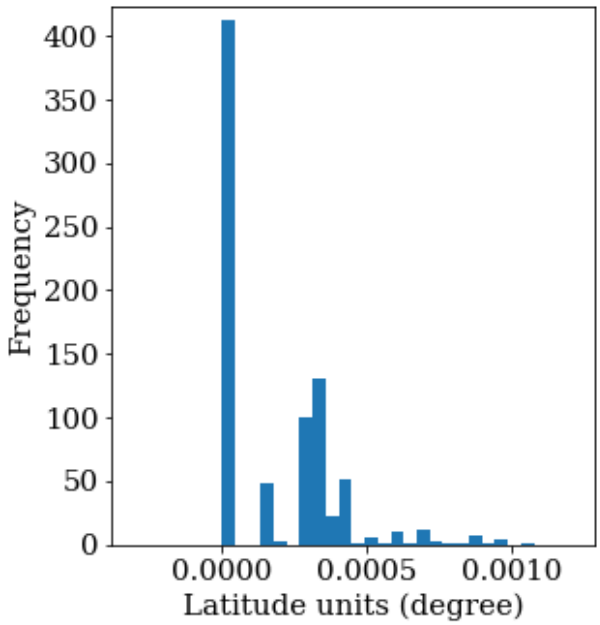

Figure 20: Horizontal scattering obtained from station bootstrapping, shown here for 24 May. The highest 10% of scattered values are discarded.

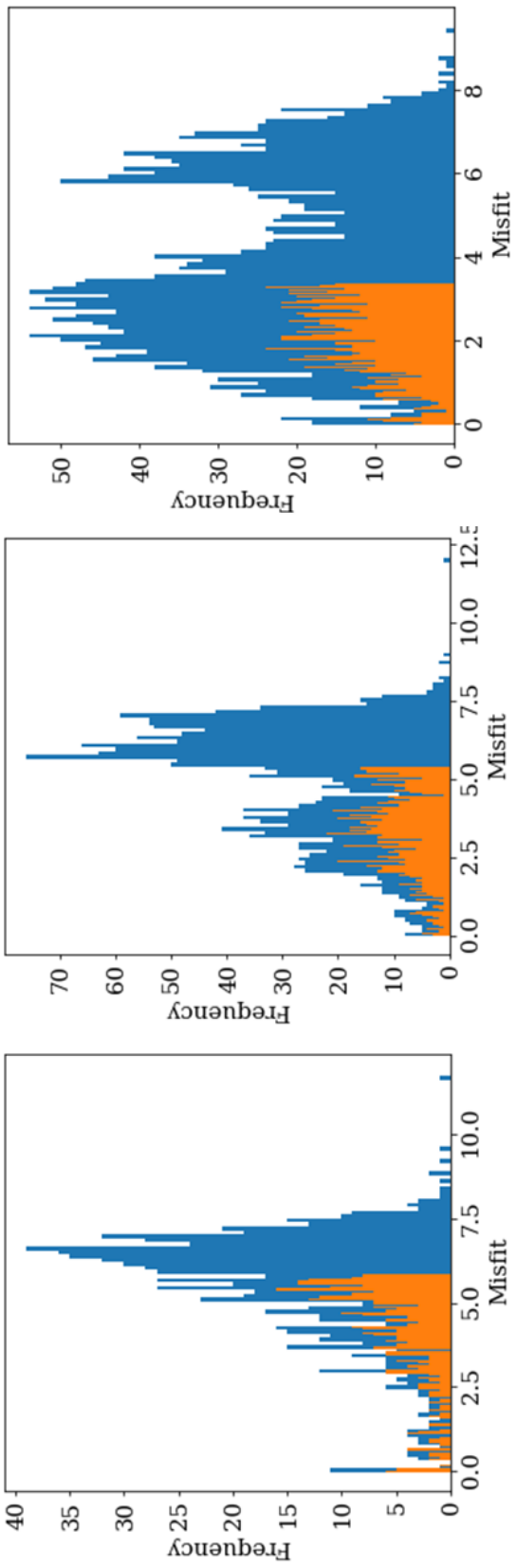

Figure 21: Histograms of misfit values obtained for 24, 25p1 and 25p2 experiments (left, center, and right respectively). Blue bars represent all misfit values and orange bars represent data with 50% of highest misfit values removed.

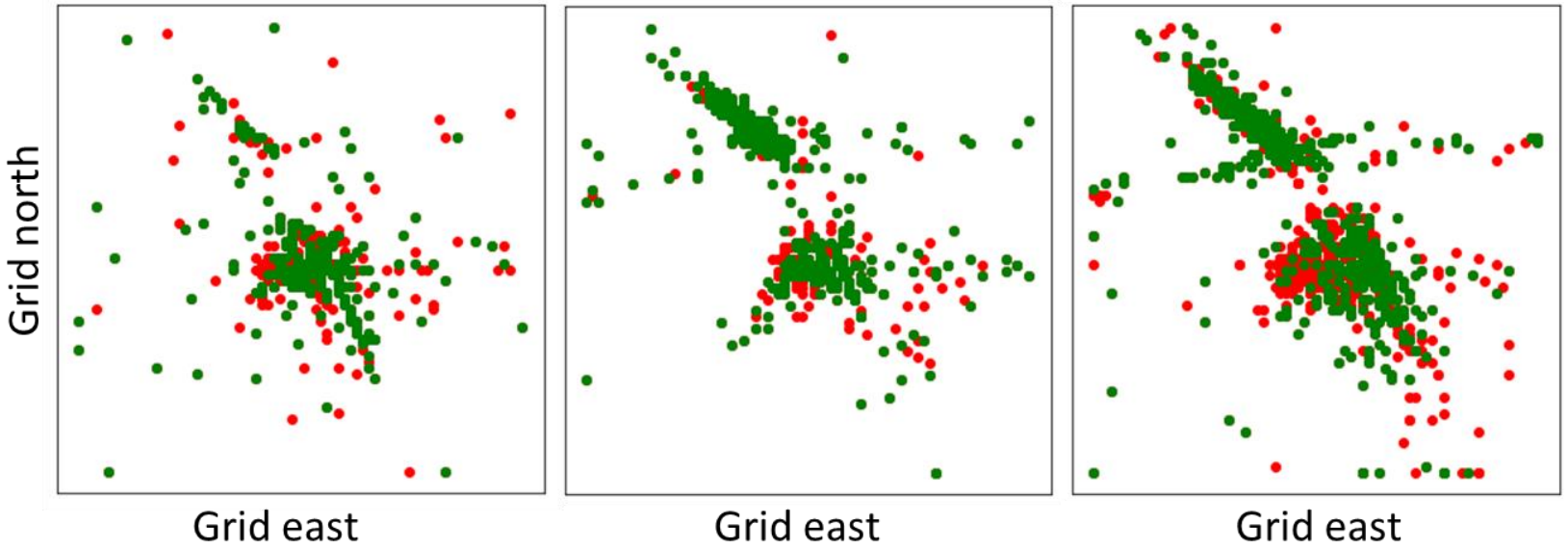

Figure 22: Effect of applying the misfit filter for infrasound locations. Pre and post filtered data shown in red and green respectively. X-axis denotes easting and Y-axis northing directions in the grid.

5.9 Semi Supervised Learning: Label Propagation

In semi-supervised learning, a machine learning algorithm is trained on a dataset that is a mixture of labeled and unlabeled data. The algorithm can use both types of data to make predictions. This can be useful when it is expensive or time-consuming to label a large dataset, as is often the case in natural language processing and computer vision tasks.

Semi-supervised learning is a type of supervised learning because the algorithm is still being trained to make predictions based on input data. However, it is "semi" supervised because it does not require as much labeled data as traditional supervised learning algorithms. This can make it a useful approach when labeled data is scarce. There are several techniques that can be used for semi-supervised learning, including self-training, co-training, and multi-view learning. (Dunham et al., 2020)

Label propagation on graphs is a technique for semi-supervised learning that can be used when the data points can be represented as a graph, with the edges between the points representing relationships between the points. The technique involves starting with a small number of labeled data points and propagating the labels to the unlabeled points by "smoothing" the labels across the edges of the graph. The process of label propagation involves iteratively updating the labels of the unlabeled points based on the labels of their neighbors. The labels of the unlabeled points are updated in such a way as to minimize the overall error in the labels of the entire graph. This process is repeated until the labels of the unlabeled points converge or until a maximum number of iterations is reached (Dunham et al., 2020). The specifics are outlined as follows:

1. It formulates the dataset X as a fully connected graph of n = l + u nodes (l = labelled, u = non-labelled) where each node corresponds to a data point in X. Y is the vector representing data

labels.

2. The edges connecting each pair of nodes have weights $w_{i,j}$, where $w_{i,j} = f(x_i, x_j)$ where f is the kernel function that gauges the proximity between a pair of data points.

3. If defines the extent to which two data points are like each other. This reflects the method's central idea that points closer to each other should have similar labels.

4. In label propagation, the probabilities Y are updated using a transition matrix with the rule $T(t+1) = TY(t)$ at the t^{th} iteration.

5. Every transition matrix element $T_{ij} \propto w_{ij}$ indicates the probability that the node j will be assigned the value Y_i of the node i.

Label propagation is a simple and effective technique for semi-supervised learning, and it has been successfully applied to a variety of tasks, including image classification and text classification. Label propagation is a probabilistic method and the uncertainty in the final labels obtained for a node can be quantified in terms of probability score.

The core information provided from microseismicity are the hypocenter locations. The locations the primary input for interpretation of conductive pathways for material transport in subsurface – for fluid flow between reservoir and production well in case of oil/gas or permeable pathways between injection and production well in case of geothermal resources. There is high uncertainty in the hypocenter locations due to the highly anisotropic nature of the subsurface, and the limited signal quality of signal. Additionally, the assignment of reliable fracture labels is possible only for a small percentage of events because of only a few events have sufficiently high signal to noise, and/or are favorably located for unambiguous assignment. For a large proportion of events, however, the assignment is ambiguous, and prone to variability due to interpreter bias.

Semi-supervised learning is the branch of machine learning focused on learning from small number of labelled data and massive number of unlabeled data – and therefore a strong fit for solving the problem. In the proposed approach two streams of readily available input data are considered: rough approximations of locations and the three-component geophone signal, as outlined in Figure 23. The data-driven workflow minimizes the human-induced bias related with associating fracture labels to microseismicity. By utilizing a rare dataset which has high number of events that are reliably associated with fracture planes, the data-driven method will be refined. As a result, the final product will be a generalizable, scalable workflow which can be applied to different length scales ranging from decameter scale experiments to km-scale field scale operations. Such a workflow can be useful in drilling optimization, well completions, or any other operations which requires good knowledge about the hydraulic fracture extent.

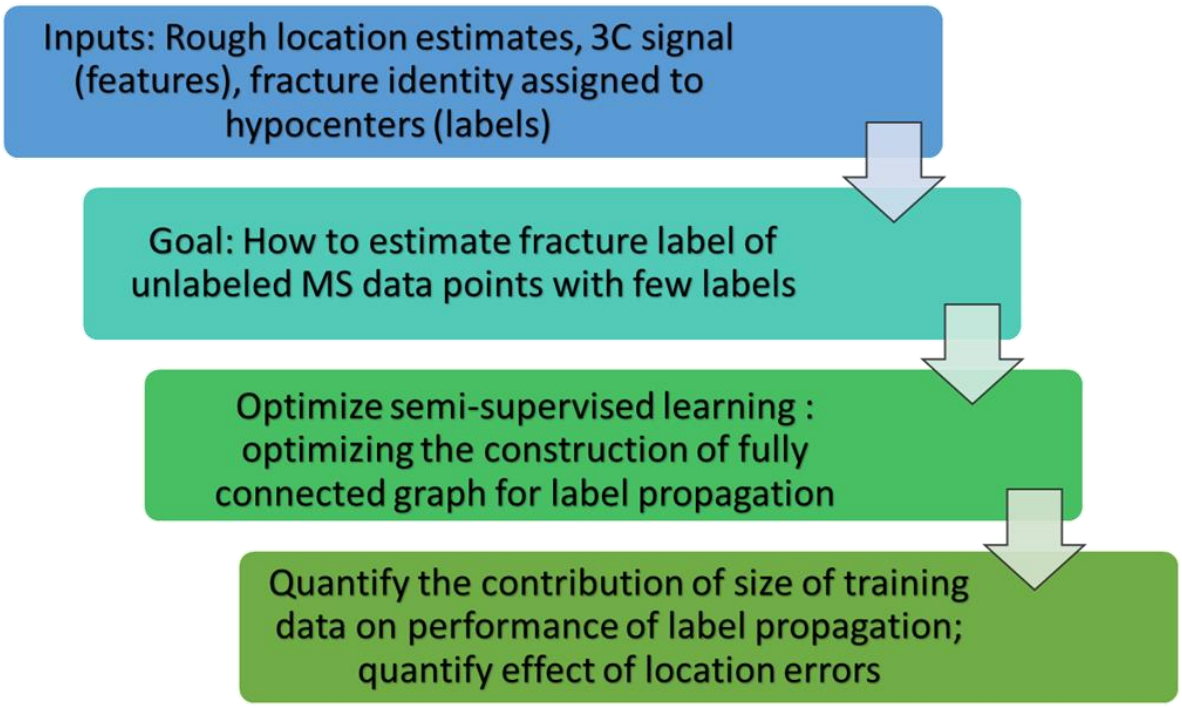

Figure 23: Workflow for semi supervised label propagation of meso scale Collab microseismic dataset.

5.9.1 Data description

Input features: three component Fourier transform spectra of microearthquake signals recorded at sensor OT-16. Target label: Fracture labels assigned to microseismic point cloud.

5.9.2 Workflow 1: Hyperparameter optimization for label propagation algorithm

The dataset is divided into training and testing sets with training set ratio of 0.3 and testing set size ratio of 0.7 and the splitting was stratified based on target labels. The key hyperparameter for label propagation algorithm are the kernel types and their respective sizes. The two kernel types are 1. K-nearest neighbor (‘KNN) and kernel size corresponding to the number of neighbors and 2. Radial basis function (‘RBF) and the corresponding kernel size being ‘gamma’. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. In each run either the RBF or the KNN kernel was chosen. The number of neighbors was varied as 2, 4, 7, 10, 15 and 20. 20 iterations were performed for each hyperparameter value, and the performance of label propagation algorithm was determined based on the average recall score of all fracture labels in the testing set. The performance of the label propagation algorithm was quantified on the recall score of the testing set. For KNN kernel, the highest recall score was achieved with number of neighbors =15 and for RBF kernel the highest recall was achieved with gamma =10. Overall, KNN kernel (with nn = 15) performed marginally better than RBF kernel (with gamma = 10) and was therefore chosen for subsequent analysis.

5.9.3 Workflow 2: Determining training data size on performance of label propagation.

This workflow is designed to answer the following questions: what is the smallest amount of training data which can yield good performance for a fixed about of testing data using label

propagation for the given microseismic dataset.

To this end, the testing data size was fixed at 0.6 and the training data size was varied between 0.05 to 0.4. the splitting was stratified based on the target variable. The label propagation was performed using KNN kernel with number of neighbors equal to 15. The performance was quantified based on the precision and recall values obtained for the testing set for each fracture label.

5.9.4 Workflow 3: Effect of locational error on performance of label propagation

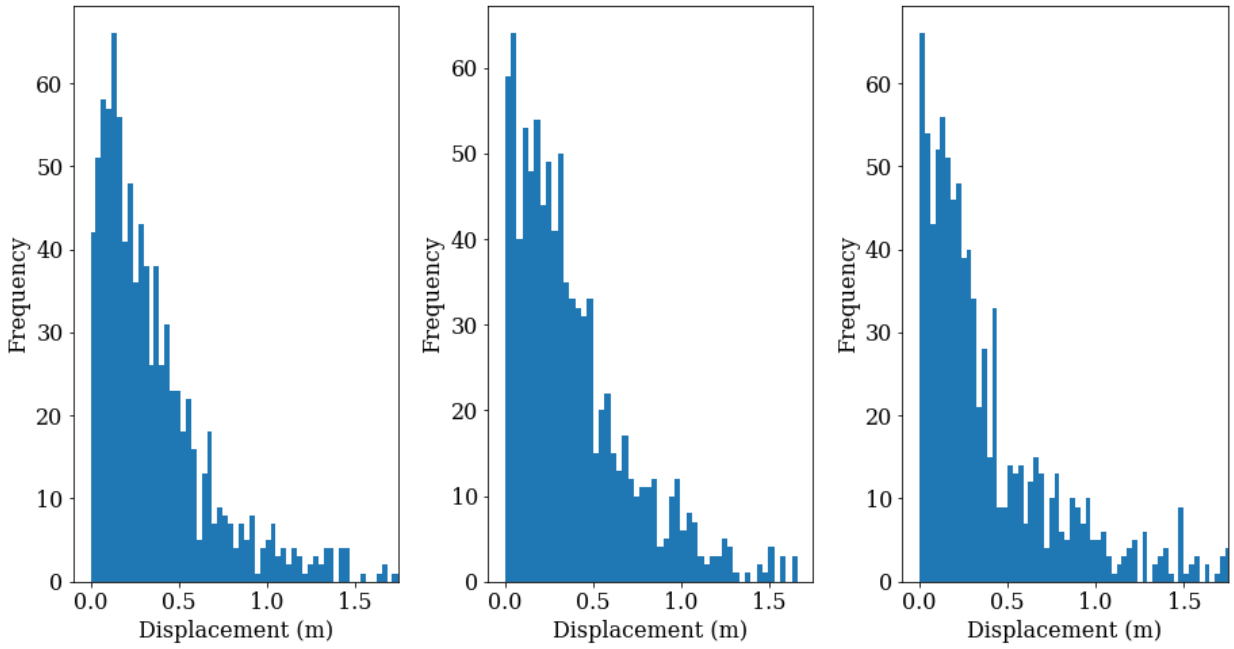

Figure 24: Absolute value of Displacement between final and rough estimate of microseismic locations in northing (left), easting (center) and depth (right) directions for EGS collab experiment 1 dataset.

The displacement between the rough estimate and the final position of each point in the microseismic cloud is used to calculate the mean and standard deviation in each direction. The histograms of the displacements in northing, easting and depth are shown in Figure 24. the standard deviation is calculated in each direction from this distribution. To introduce noise in the

locations, the standard deviation is multiplied by a factor, termed here as the 'Sigma multiplier factor' and added to the point. The sigma multiplier is varied from 0 to 3. Multiplier factor of 1 implies that one standard deviation error is added to the original location in each direction. A visualization for the locational error is shown in Figure 25. These locations, along with the three component signal spectra as used as input features for the UMAP-based dimension reduction. The reduced dimension embeddings obtained from UMAP are the dataset on which label propagation is performed. For a given sigma multiplication factor, the dataset is split into 0.1 training and 0.9 ratio and stratified based on the target variable. The label propagation algorithm is applied using knn kernel type and number of neighbors equal to 15. The performance for individual fracture plane (target label) is quantified based on the precision and recall obtained on the testing set.

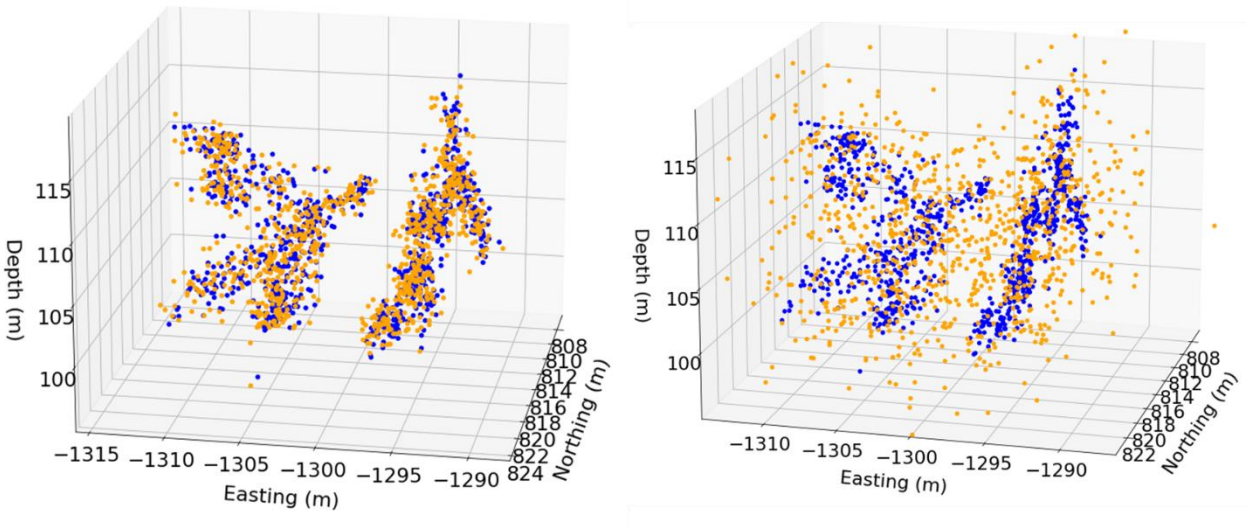

Figure 25: Microseismic locations, blue points indicate actual locations and orange points indicate points with added locational error based on standard deviation calculated between final locations and initial location estimates. (Left) case when multiplier factor =0.5 meaning orange points have half standard deviation error and (right) case when multiplier = 2 (twice standard deviation).

5.10 Relationship Between Geomechanical Deformation and Recorded Seismic Motion in Field Scale Hydraulic Fracturing Induced Seismicity

Conventional approaches of microseismicity interpretation are usually limited to spatial extend of the hypocenter locations. The effective drainage volume in case of hydrocarbon reservoirs or the effective heat exchange volume in case of geothermal reservoirs- is estimated on grounds of binning or shrink wrapping (Chakravarty & Misra et al., 2022) of the microseismic cloud. In this process however, no differentiation is made between different types of microseismic events. It is well understood that the mode of rock deformation reflects differences in the moment tensor of the seismicity (Martínez-Garzón et al., 2017). The specific mode of rock deformation during hydraulic fracturing manifests as differences in the radiation pattern of the corresponding seismicity. The signatures of the radiation pattern in turn manifest as differences in the moment tensor of the seismic event. It is established that certain classes of moment tensors like ISO and CLVD are more strongly correlated with the permeability enhancement compared than other classes (like DC) (Martínez-Garzón et al., 2017).

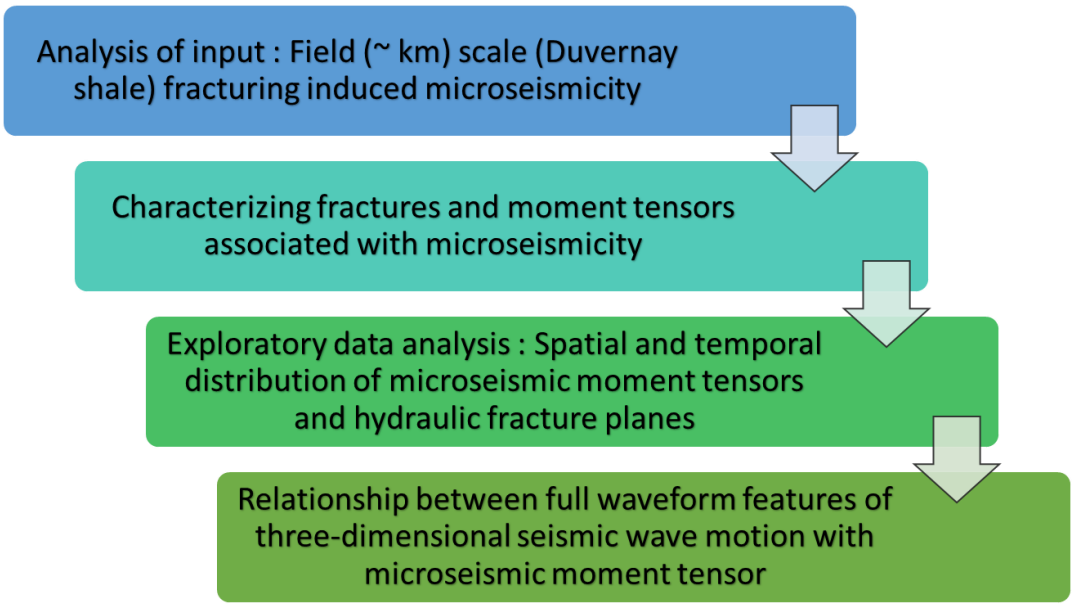

Figure 26: Workflow for establishing data-driven link between full waveform signal, moment tensor and fracture planes in field scale Duvernay shale microseismic dataset.

The joint analysis of moment tensors, spatial distribution of microseismicity and corresponding three component signal properties during hydraulic fracturing is therefore an important yet unexplored avenue of research. The absence of a well-defined analytical method for such analysis, and the presence of massive amounts of structured data makes this problem a strong candidate for applying machine learning methods. The outcomes of the proposed workflow (Figure 26) will deepen our understanding of the relationship between microseismicity and classes of subsurface rock deformation during hydraulic fracturing operations. While earlier it was limited to the spatial density of microseismic locations. It will reveal the contribution of different event classes to permeability enhancement, thereby enabling the usage of microseismicity to obtain a clearer picture of the subsurface conductive pathways.

6. RESULTS AND CONCLUSIONS[*]

6.1 Lab Scale Analysis

The first part of results addresses the following question:

"How to discover generalizable and physically relevant signatures of fracturing by analyzing the wave-transmission and microseismic-emission measurements using clustering methods?"

The key takeaways from this part of results are as follows:

1. The workflow preempts the need for picking arrival time of seismic waves; thereby, reducing the uncertainty associated with sonic/wave data analysis.

2. The energy and time-frequency contents of the waveforms are used to cluster the multipoint ultrasonic measurements. Following that, the tenets of the experimentally proven displacement discontinuity theory are applied ascribe physical meaning to the clusters by converting them into a geomechanical alteration index, which proves to be a robust measure of the hydraulic fracturing induced geomechanical alteration.

3. The results of the proposed workflow agree favorably with independent measurements from acoustic emission and X-ray computed tomography. The STFT of a waveform can accurately capture changes in the frequency, duration, and energy of a wave as it passes through a discontinuity. This is demonstrated by the difference in STFT-derived features of waveforms traveling through intact and fractured materials, as shown in a principal component space plot

[*] Reprinted with permission from "Visualization of hydraulic fracture using physics-informed clustering to process ultrasonic shear waves" by Chakravarty, A., Misra, S., and Rai, C. S., *International Journal of Rock Mechanics and Mining Sciences, 137*, 104568. Copyright 2021 by Elsevier. Reprinted with permission from "Unsupervised learning from three-component accelerometer data to monitor the spatiotemporal evolution of meso-scale hydraulic fractures" by Chakravarty, A. and Misra, S., *International Journal of Rock Mechanics and Mining Sciences, 151,* 105046. Copyright 2022 by Elsevier.

(Figure 27). The proposed method of feature extraction can effectively distinguish between waveforms that have passed through intact and fractured materials.

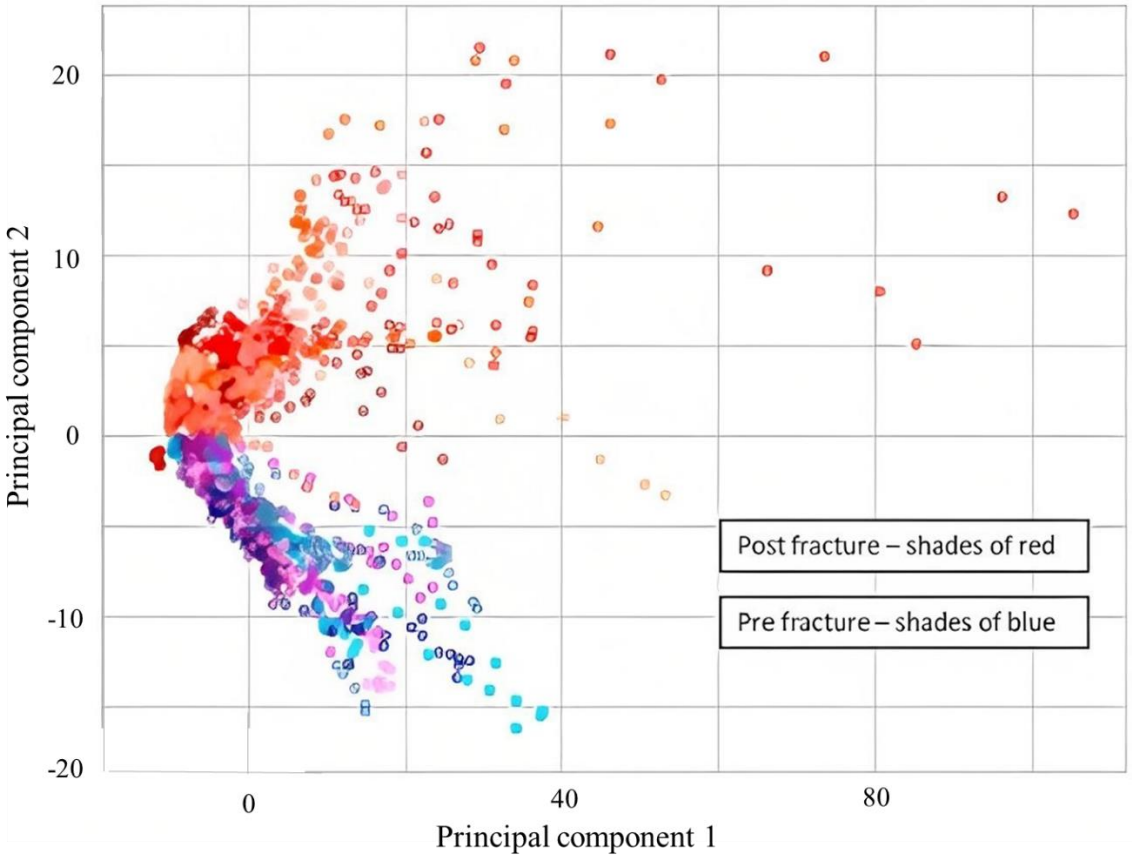

Figure 27: Projection of STFT-derived features from pre-fracture (blue shades) and post-fracture (red shades) in principal component space, projected in first (PC1) and second (PC2) principal components. The different shades of a color correspond to different transducers. This is evidence that STFT features in the principal component space can distinguish between shear transmission signals traveling through fractured material versus those through intact material.

Table 6 compares the mean silhouette scores of different clustering methods for different numbers of clusters. The table helps to identify the optimal clustering method and the number of clusters that will produce the most reliable results. In the current dataset, K-Means clustering has the highest silhouette scores, while Agglomerative clustering has slightly lower scores and tends

to be more computationally intensive. DBSCAN performs poorly when there are more than 2 clusters because the density of the data is concentrated in two main areas, beyond which the algorithm's effectiveness drops significantly (Table 6).

Table 6: Silhouette scores of various clustering methods for different cluster numbers obtained by processing the shear-waveform measurements after physically relevant feature extraction. The light grey filled boxes indicate the extremely poor clustering results, while the dark grey filled boxes indicate decent clustering performance.

Orientation	Number of Clusters	DBSCAN	K-Means	Agglomerative
Axial Plane	2	0.49	0.55	0.51
	3	-0.15	0.56	0.47
	4	-0.13	0.56	0.52
	5	-0.15	0.29	0.34
	6	-0.17	0.28	0.16
First Frontal Plane (wave transmission perpendicular to fracture)	2	-0.25	0.43	0.41
	3	-0.27	0.26	0.24
	4	-0.29	0.25	0.25
	5	-0.29	0.25	0.26
	6	-0.29	0.13	0.11
Second Frontal Plane (wave transmission parallel to fracture)	2	-0.23	0.75	0.75
	3	-0.30	0.71	0.73
	4	-0.33	0.24	0.22
	5	-0.34	0.26	0.23
	6	-0.34	0.21	0.22

Figure 28 uses violin plots to show the reliability and uncertainty of the clustering results as measured by the J parameter. These plots are used to determine the optimal method and number

of clusters, as well as to understand the physical meaning of each cluster. The plots visualize the distribution of each cluster as a function of the J parameter, with the middle horizontal line indicating the median value and the horizontal lines at the ends representing the 95th percentile. A robust clustering should have a clear separation between the 95th percentile values of different clusters, with minimal overlap between their spread in terms of the J parameter.

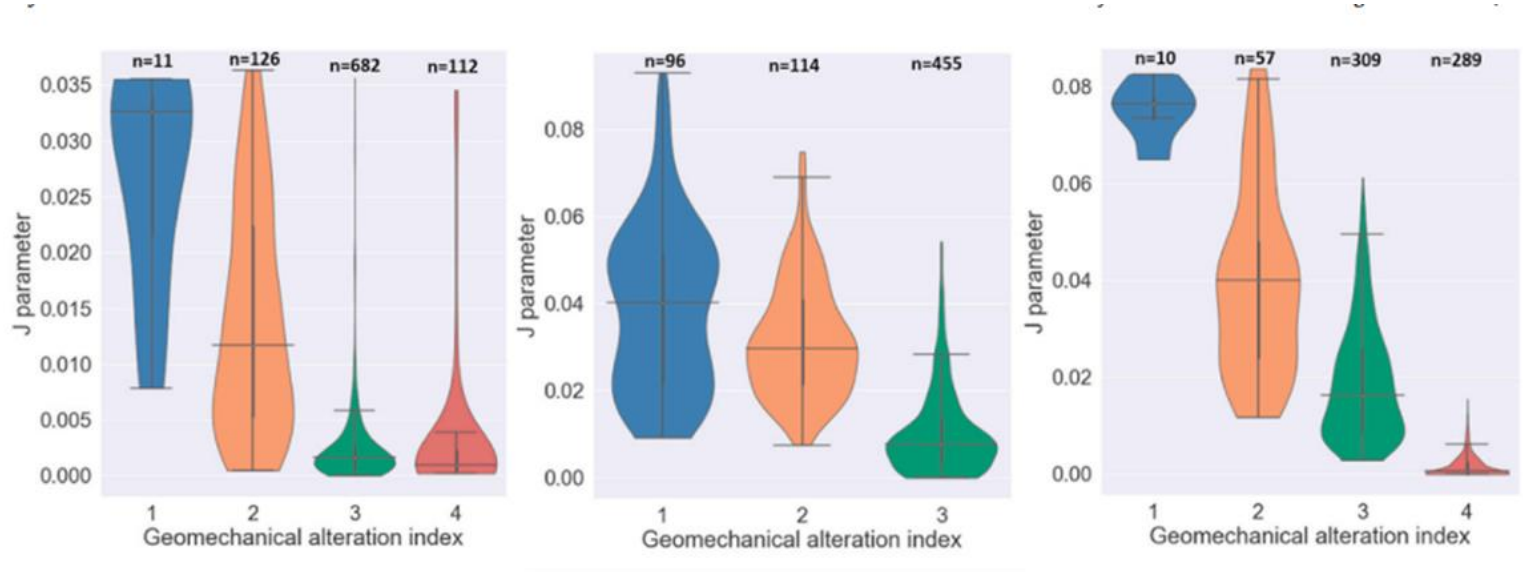

Figure 28: Violin plots of distributions of the newly developed *J* parameter for the axial (left), frontal (transmission is perpendicular to primary fracture, middle) and frontal plane (transmission is parallel to primary fracture, right) for sample TSU6. '*n*' is the number of data points (samples) that were assigned a specific cluster

By comparing the silhouette scores, median values, and 95th percentile values of the J parameter for various numbers of clusters and clustering methods (i.e., by considering both Figure 28 and Table 6), the optimal number of clusters is determined to be 4, 3, and 4 for the axial, frontal (perpendicular), and frontal (parallel) orientations, respectively. Among these clusters, Cluster 1 is the least robust in the axial and second frontal planes because it contains a small number of samples, while Clusters 3 and 4 in the axial plane are very similar but distinct from the other clusters. Cluster 4 in the second frontal plane is the most robust, followed by Cluster 3 in the first frontal plane. In general, the proposed method of feature extraction and physics-informed unsupervised approach is more effective at detecting areas with higher levels of geomechanical alteration. Figure 29 illustrates the output of the workflow for different orientations of sample TSU6, with hotter colors representing higher levels of geomechanical alteration. In the axial orientation, the damage is concentrated in the center plane of the sample, with the elongated region of red color representing the fracture length and thickness. The area with the highest GAI (geomechanical alteration index) coincides with the region of highest acoustic emission density. In the frontal plane with transmission parallel to the fracture, most of the alteration is found at the center of the sample, with the damage extending towards the lower right where there is a fracture outcrop on the surface. There is a high overlap between areas of high alteration and high acoustic emission density. In the frontal plane with transmission perpendicular to the fracture, the maximum alteration is found in the upper half of the sample, corresponding to the fracture width. There is also significant alteration in the lower right region, but this is not due to fracturing but rather to improper coupling between the sample and the transducers.

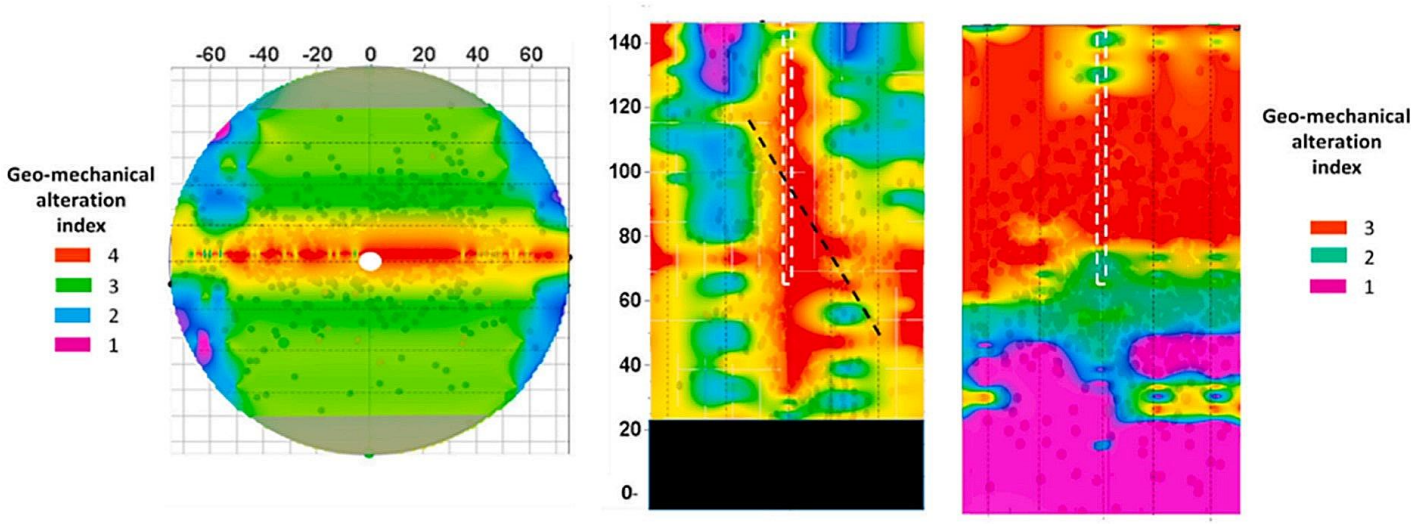

Figure 29: Maps of geomechanical alteration index (GAI) for the axial plane (left), first frontal plane (wave transmission perpendicular to fracture; middle) and second frontal plane (wave transmission parallel to fracture; right) orientations of sample TSU6. Grey regions show the zones not scanned by transducers.

6.1.1 Assumptions and Limitations of presented workflow

There are a few limitations to using transmission measurements, such as:

1. Higher cost of deployment at the field scale because it is more difficult to place receivers to collect transmission data.
2. Reflection data is better for obtaining a 3D description of embedded structures, while transmission data is more suited for 2D characterization along the travel path.
3. Geological materials often have a high attenuation coefficient, which limits the use of higher frequencies and therefore the resolution of transmission measurements.

The current study also has some limitations:

1. The fracturing-induced alteration can only be visualized in 2D for the axial and two frontal planes.
2. The effects of the borehole on ultrasonic wave transmission have not been fully considered.
3. The effects of sample boundaries on wave transmission have not been fully incorporated into the proposed workflow.

When analyzing transmission data, any area of improper contact between the sample and transducer will show a high geomechanical alteration index.

6.1.2 Conclusions

1. This study employed a method that utilizes physically-relevant feature extraction and physics-based unsupervised learning to non-invasively show the changes in geological material caused by hydraulic fracturing. This approach eliminates the need for determining the arrival

time of stress waves, thus decreasing uncertainty in sonic/wave data analysis. The ultrasonic wavelength is on the same scale as mechanical discontinuities, so the proposed method does not rely on the effective medium theory. Instead, the energy and time-frequency aspects of the waveforms are utilized to group the ultrasonic shear-wave transmission measurements (Chakravarty et al., 2021)

2. This study applies the principles of the experimentally-verified displacement discontinuity theory to give physical significance to the clusters by converting them into a geomechanical alteration index, which serves as a reliable measure of changes caused by hydraulic fracturing. The non-invasive visualizations are consistent with measurements made using acoustic emission and X-ray computed tomography. Techniques such as short-time Fourier transform spectrogram, wave-transmission coefficient, silhouette score based on separation and cohesion, and dimensionality reduction were used to achieve the desired non-invasive visualization of the geomechanical alteration caused by hydraulic fracturing.

3. Propagation through fractures causes two notable changes: a decrease in transmission coefficient and an increase in the phase arrival time of the wave (slowness increases). The study uses whole signal waveform features as inputs for clustering and obtains clusters that do not have physical significance. These clusters are given physical meaning by defining a parameter (J) that accounts for the two physical changes. This parameter serves as a proxy for the mechanical alteration caused by cracks in the material. To spatially map the geomechanical alteration index, the clusters identified using the clustering method are first made statistically consistent using cohesion, separation, and silhouette score to determine the optimal number of clusters; then, the optimal clusters are assigned physical meaning based on the newly developed J parameter. Overall, the proposed method of physically relevant feature extraction and physics-informed

unsupervised learning is more effective in detecting regions that have undergone greater geomechanical alteration (Chakravarty et al., 2021).

6.2 Representation Learning Using UMAP Part 1: Native UMAP Implementation for Polarization Features From OT-16 Sensor

The second part of results addresses the following question:

"What is the best strategy for characterizing fracturing induced microseismic locations by analyzing the borehole-based 3-component accelerometer data using manifold learning methods?"

The key takeaways from this analysis are as follows:

1. Our study shows that the density-based clusters in the projected 3D space correspond to distinct types of hydraulically induced fracture zones in the reservoir volumes around the injection points.

2. The temporal evolution of these clusters is used to track the intensity and duration of reservoir stimulation (fracture creation and propagation) for the various types of fracture zones.

3. Considering the data from EGS Collab experiment 1, we showed that well-defined seismic polarization features from the microearthquake signal at a single station contain within the signatures of the fracture planes on which they lie. Hence, micro-seismic point cloud interpretation can be aided by the results of this workflow.

Using the identical STA/LTA (short-term average/long-term average) thresholds, 815 triggers were detected in 10 minutes of continuous recording on May 22 (Day 1), 4800 triggers were detected in 90 minutes of continuous recording on May 23 (Day 2), and 14000 triggers were

detected in 25 minutes of continuous recording on May 24 (Day 3). Figure 30 compares the injection rate with the trigger rate for the three days. The hypocenters (locations of the earthquake focus or origin) were determined by inverting the arrival times of high signal-to-noise ratio events detected simultaneously on multiple accelerometers and hydrophones.

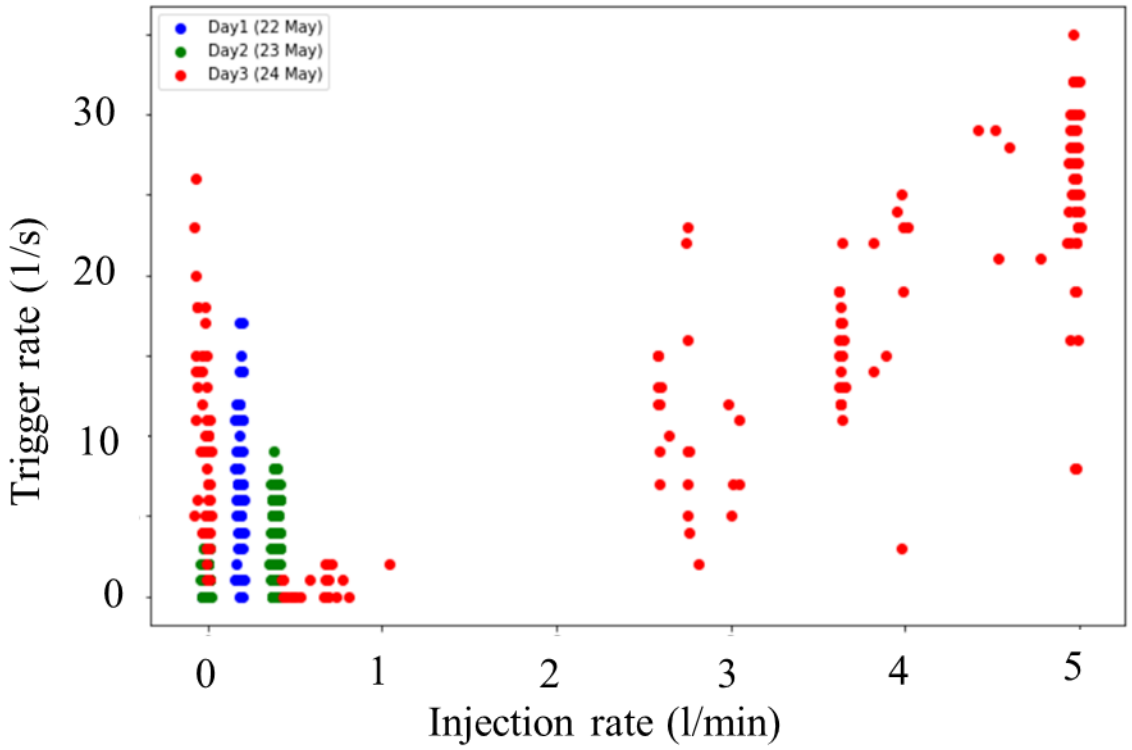

Figure 30: Different regimes of microseismicity recorded over 3 days. The x-axis represents injection rate and y-axis represents the number of triggers detected per second.

Day 1: Figure 31 and Figure 32 shows the distribution of 815 triggers in the 3D UMAP (uniform manifold approximation and projection) space for the May 22, 2018 (Day 1) stimulation. Four dominant clusters are observed in the May 22 dataset, with a minor fifth cluster near the embedding axis 3 corresponding to electronic noise signals. Due to the high planarity and rectilinearity of electronic noise, this cluster is relatively distant from the other clusters in the UMAP space.

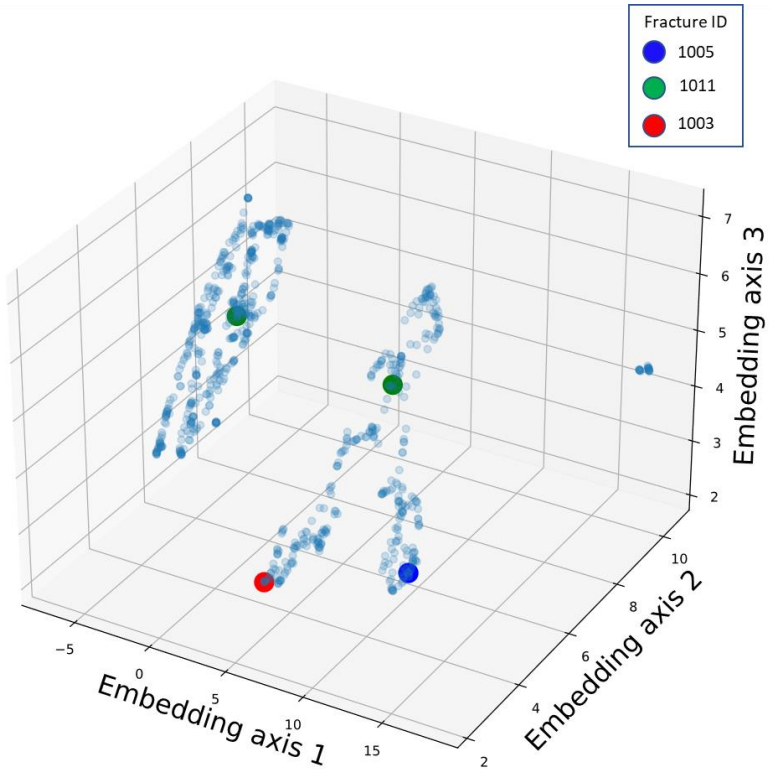

Figure 31: Locations of fracture-laden hypocenter triggers in the UMAP space for May 22. The minor fifth cluster near embedding axis 3 corresponds to the set of electronic noise signals. Due to the high planarity and rectilinearity of electronic noise, the corresponding cluster lies relatively distant from the other clusters. The four events lie on distinct clusters in the UMAP space with no two different fractures sharing the same clusters.

The projection of polarization features in the UMAP space is an effective unsupervised method for filtering high-amplitude electronic noise from the continuous record. Out of the 815 triggers detected on OT16 using STA/LTA, only 37 of them correspond to located events. Figure 31 shows 4 of the 37 events that have been assigned a fracture plane based on the cumulative information of the event hypocenters at the end of the stimulation in December 2018. These four events are located on distinct clusters in the UMAP space, with no two different fractures sharing the same cluster. This is likely because events located in distinct fracture planes tend to have distinct polarization features due to the surrounding media and their location in space and time.

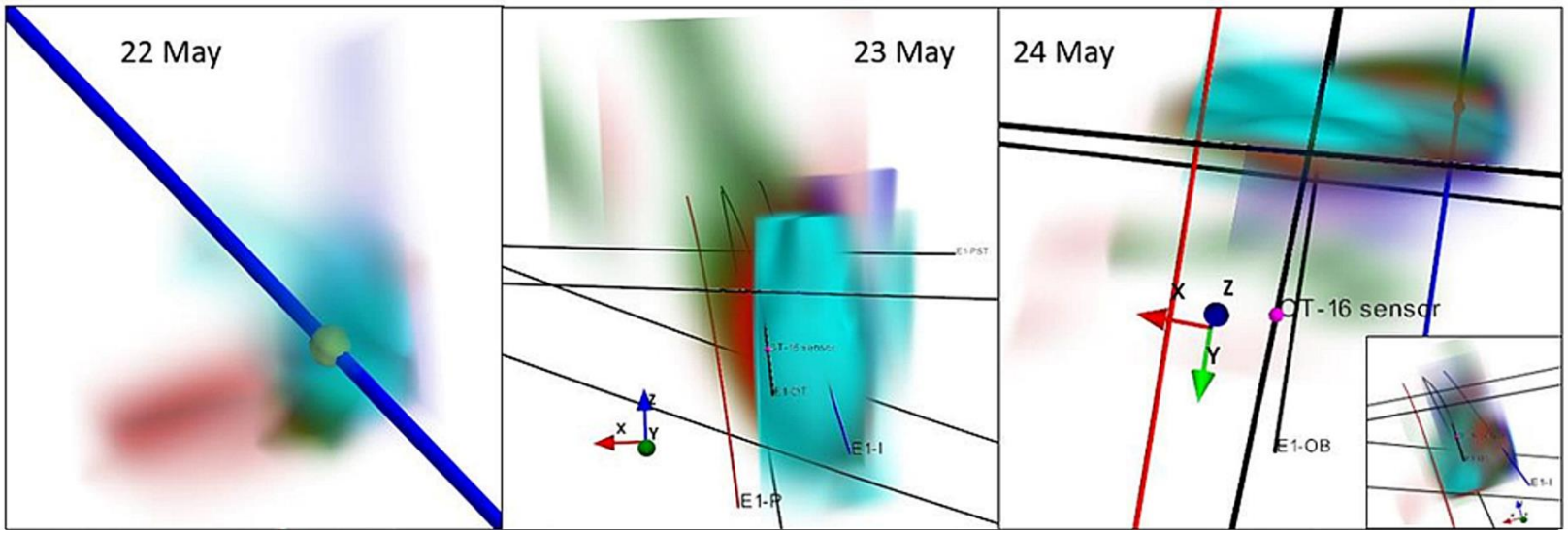

Figure 32: Fractured zones colored by cluster ID inferred from clustering of polarization features in three dimensional UMAP space on 22 May, 23 May, and 24 May 2018 (left, center and right, respectively). All the monitoring boreholes are shown by black lines. The inset on 24 May panel shows the fracture zones viewed subparallel to E1I and E1P.

The temporal variations in the clusters obtained are shown in Figure 33. Located microearthquakes, which are a subset of the triggers, are also assigned to clusters (second panel from bottom) based on their distance from the injection point in the notch at 164 feet in well E1-I. Cluster 4 is dominant at the early time, while cluster 3 is dominant at the late time. Clusters 1 and 2 are predominant at the middle time of the injection. Cluster 1 (blue) has the largest dimensions among the four fracture sets. Although cluster 2 (green) has relatively high activity throughout the injection, its extent is small and limited to the area around the injection point. The propagation of such fracture strands is influenced by factors such as the near-wellbore stresses and the virgin rock's geomechanical properties, including elastic anisotropy and pre-existing planes of weakness. Additionally, a stress gradient was created by the temperature difference between the mine shaft and the rock volume. The complex interactions between these factors result in the highly heterogenous propagation of fracture strands in space and time, causing different fracture branches to have disparate trajectories emanating from the injection point. This may explain why cluster 2 is limited to a small volume near the injection.

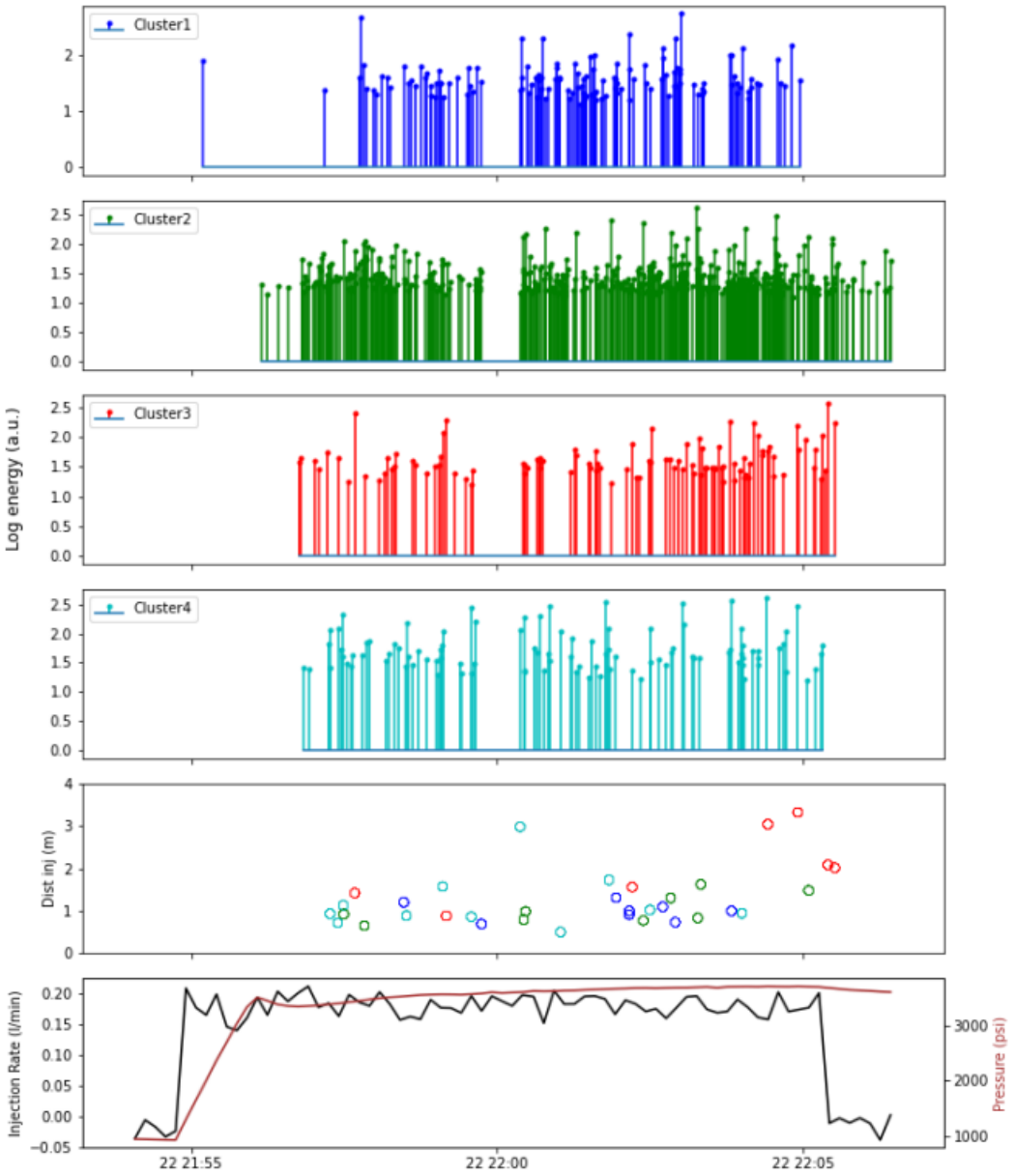

Figure 33: Temporal distribution of the fracture clusters created in the May 22 (Day 1) injection cycle (top four panels). Second last panel shows the time distribution of the 37 event hypocenters that were located out of 815 triggers. Bottom panel shows the hydraulic stimulation parameters, namely injection rate and pressure.

The spatial distribution of the 37 microearthquakes in the stimulated volume around the injection point that occurred on May 22 is shown in Figure 34. Each event is assigned a cluster label and color based on the proposed unsupervised learning workflow. There are 4 clusters in the figure.

Events with the same cluster label are in similar positions in space. Clusters 1 and 3 (blue and red, respectively) represent the two main branches of the hydraulic fracture, while Cluster 2 (green) is limited to the area beneath the injection points. Cluster 4 (cyan) corresponds to a fracture that has grown along the wellbore.

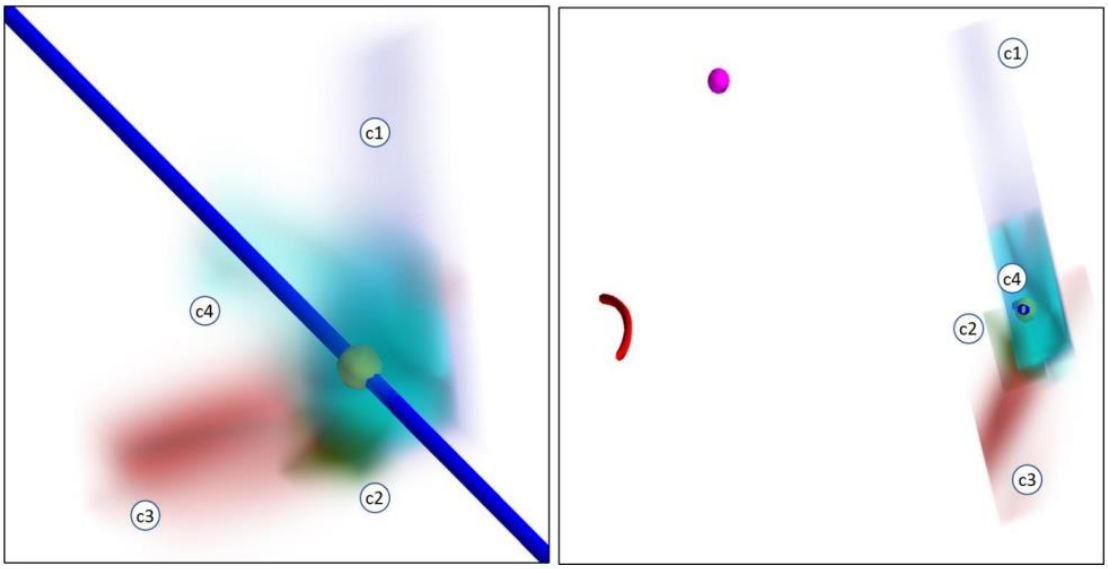

Figure 34: Close-up (left) and gun-barrel (right) view of the microseismic point density colored by cluster labels. The cluster labels are marked over their corresponding fracture branches. Red and blue indicate production and injection well, respectively. Pink sphere is OT 16 sensor. Green patch over the well represents the injection interval. The average distance between the injection and production well is 10 meters.

The recorded signal from a single sensor can be thought of as the result of combining a source function and a medium transfer function. The particle motion (and therefore the polarization features) represents the seismic wavefield at discrete times. Points with the same cluster label create similar seismic wavefields on the accelerometer. This suggests that microseismicity from different strands of fractures represents statistically different seismic wavefields. The examples from Days 1, 2, and 3 show that different strands of hydraulic fractures have different polarization signatures. The different clusters in the data represent different fracture branches

around the injection point. The temporal changes in the energy of triggers from different clusters can be used to track the growth of these fracture branches over time.

Day 2: On the second day of the study, 129 microearthquakes were recorded (Figure 32 and 35). The fluid injection lasted for about 60 minutes, during which the injection rate and pressure were gradually increased to create a fracture with a nominal radius of 5 meters. Initially, the seismicity was observed near the locations of the previous day, but after about 12 minutes, the earthquakes began to migrate downward and towards the injection well. After about 30 minutes, the seismicity shifted and moved closer to a monitoring borehole. This shift was also reflected in the temperature measurements taken by the distributed temperature sensing equipment. The measurements taken by the accelerometer in the borehole became extremely noisy after this point, possibly due to the vibrations caused by water jetting. The stimulated volume reached close to the production well, and different clusters of microearthquakes were observed in different areas, with many small secondary fractures being created rather than a single large one.

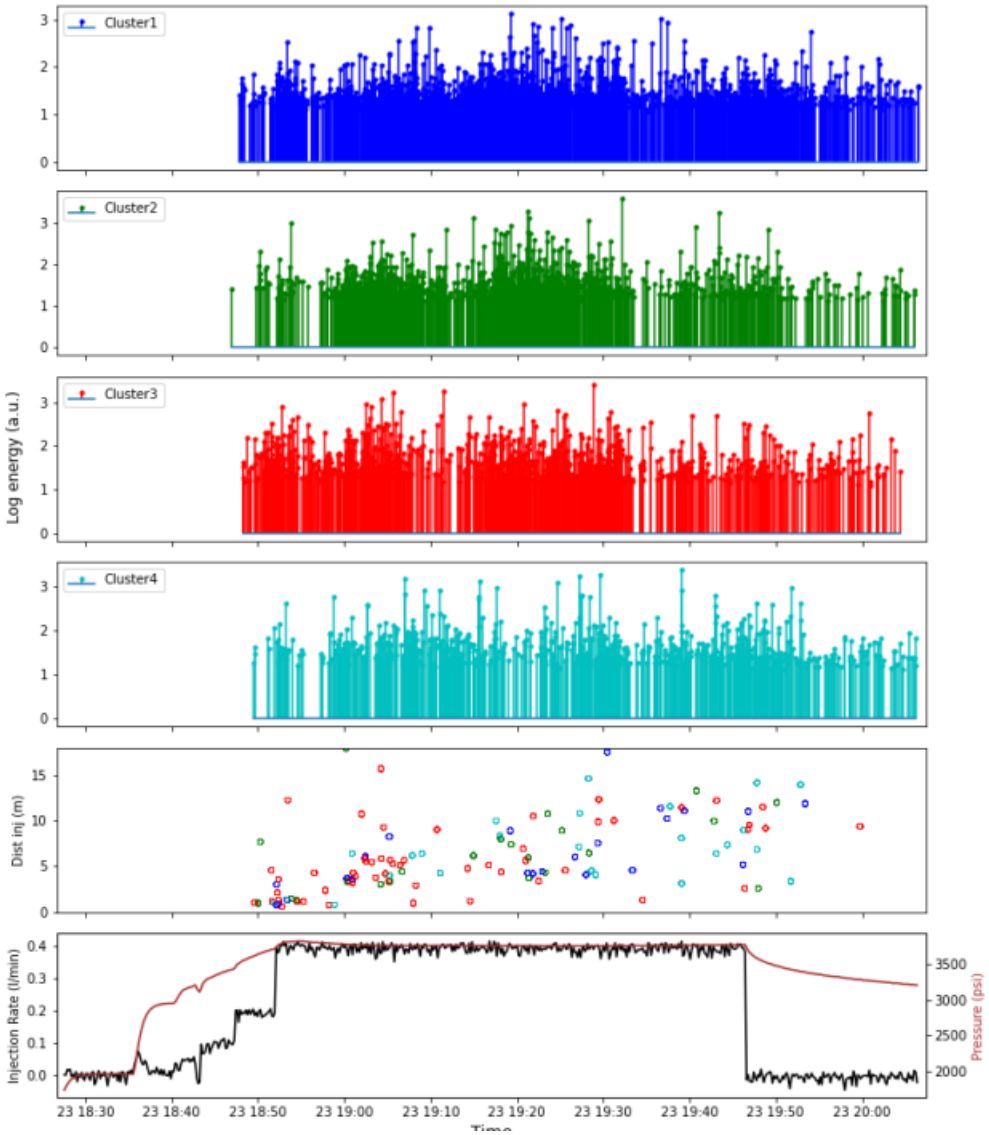

Figure 35: Temporal distribution of the fracture clusters created in the May 23 (Day 2) injection cycle (top four panels). Second last panel shows the time distribution of the 129 microearthquakes and their distance from the injection point (notch at 164 feet, E1-1). Bottom panel shows the hydraulic stimulation parameters (injection rate and pressure).

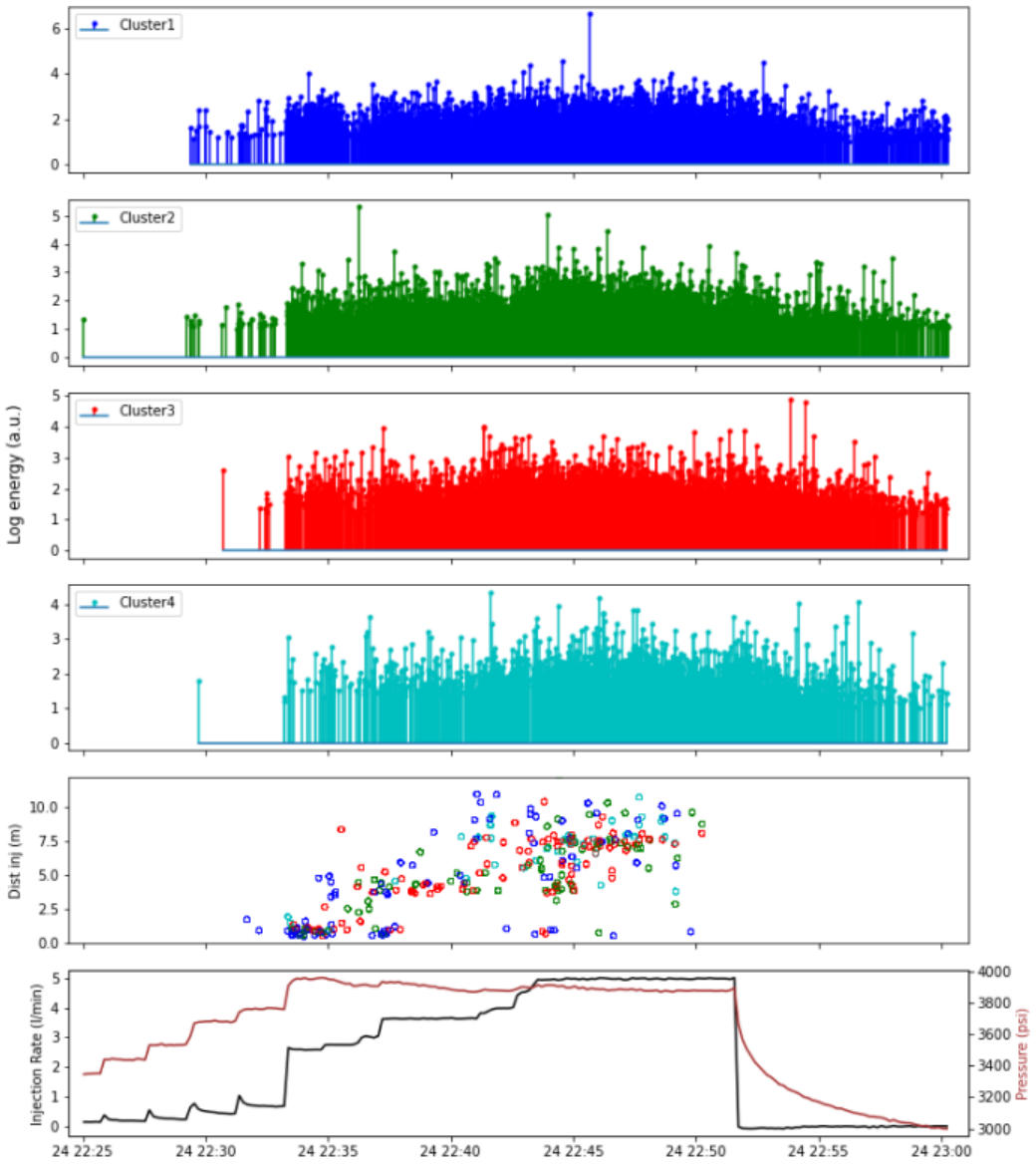

Figure 36: Temporal distribution of the fracture clusters created in the May 24 (Day 3) injection cycle (top four panels). Second last panel shows the time distribution of the 129 microearthquakes and their distance from the injection point (notch at 164 feet, E1-1). Bottom panel shows the hydraulic stimulation parameters (injection rate and pressure).

Day 3: On the third day of the study, a fluid injection was performed at a higher rate of 5 liters per minute until fracture breakthrough was achieved at the production borehole. The microearthquakes observed during this injection had different polarizations, which were related

to their spatial distribution. The higher flow rate of the injection on this day resulted in a much higher number of microearthquakes being recorded, with 296 events located and around 14,000 triggers recorded at the accelerometer in a span of 25 minutes. The microearthquakes (Figure 32 and 36) were divided into different clusters based on the timing of their occurrence, with the blue cluster being dominant at the beginning, the green and cyan clusters being dominant in the middle, and the red cluster being dominant at the end.

6.2.1 Conclusions

1. This study examines the use of an unsupervised manifold-learning method on multi-component accelerometer measurements obtained from a small-scale field experiment, which was designed to study the spatial and temporal changes in fractures caused by hydraulic fracturing in an enhanced geothermal system (Chakravarty and Misra 2021).
2. Considering signal-to-noise ratio, the study used three-dimensional particle motion data measured by a single accelerometer installed on the monitoring borehole surrounding the stimulated volume. The continuous data stream was divided into individual triggers to identify the signals associated with the hydraulically induced microseismic events.
3. However, a limitation of this approach is that all triggers are treated equally in the unsupervised learning process, which may include non-hydraulic fracturing seismic signals in the trigger dataset. These non-hydraulic signals can come from various sources such as personnel movement in the mining area, drilling machinery, local and regional seismicity, and sensor interference (Chakravarty and Misra 2021).
4. The study derived four polarization features (azimuth, incidence, rectilinearity, and planarity) from the identified signals. These features were then processed using uniform

manifold approximation methods to project the raw signals of the microseismic events onto a 3D space. The study found that the density-based clusters in the projected 3D space correspond to different types of fractures caused by hydraulic fracturing in the reservoir volumes around the injection points.

5. The projection of polarization features in the UMAP space is an efficient unsupervised method to filter out high-amplitude electronic noise from the continuous record. The study also showed that well-defined seismic polarization features from the microearthquake signal at a single station contain information about the fracture planes on which they lie, which can aid in the interpretation of microseismic point clouds.

6.3 Representation Learning Using UMAP Part 2: Refined UMAP Implementation for Microseismic Interpretation

The third part of results addresses the following question:

"Do the signals corresponding to the microseismicity induced by the fluid injection carry distinct signature of its underlying fracture plane?"

The key takeaways from this study are the following:

1. Our workflow involved feature extraction from a single sensor, followed by dimension reduction using the optimized UMAP algorithm (a novel development in this analysis), to generate embeddings that carried reliable signatures of the fracture planes.

2. We validated the strong correspondence between UMAP embeddings and microseismic coordinates. The first step of the validation computing calculating distances between pairs of differently sized clusters with Wasserstein distance.

3. This study showed that pressure signals from hydrophones also held diagnostic signatures, marking the first use of pressure transducer data in a non-marine setting.

The selection of appropriate sensor data features is crucial for successful representation learning. In this study, two sets of features were considered: (1) three-component Fourier spectra, which are derived from the Fast Fourier Transform of 3 ms-long trigger signals, and (2) short-time Fourier transform (STFT) of a single component of the trigger. While the three-component FFT can capture more detailed information about wave propagation, it is time-invariant, meaning it does not capture the time-varying nature of the signal. Additionally, the high frequency of the signal leads to low signal-to-noise ratios in the Fourier transform.

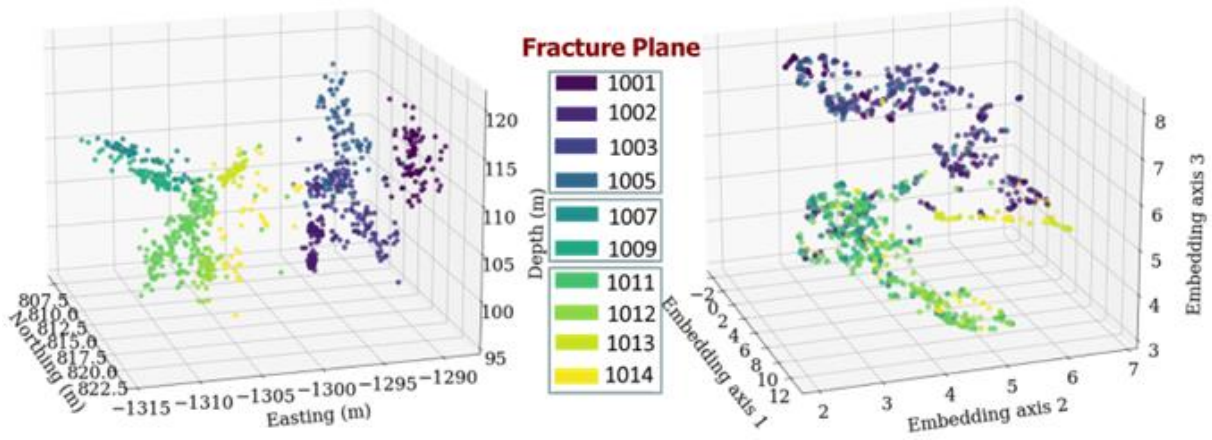

Figure 37: Microseismicity locations (left) and UMAP embeddings (right) derived from OT-16 accelerometer signals. Colors correspond to distinct fracture planes.

The STFT captures the time-varying nature of the signal, but it may not capture information from motion in the other two dimensions due to the large size of the feature vector. Despite this, the STFT was found to produce better UMAP embeddings, possibly because it captures the time-

varying nature of the signal that is absent in the time-invariant Fourier spectra. As a result, the STFT was chosen as the signal feature extraction method for the remainder of the study.

Another important reason for choosing STFT as the signal feature extraction method is that it is more generalizable to situations where sensors are only capturing a single channel of motion or a single quantity. In these cases, it is not possible to obtain three-dimensional features, so selecting single channel features allows the workflow to be applied more broadly. This makes the STFT method more suitable for a wide range of applications.

The previous analysis was conducted using a sensor (OT-16) located within the region of injection-induced microseismicity (Figure 37). However, in industrial applications of passive seismicity, sensors are often located at a considerable distance from the zone of seismicity. For example, in commercial oil and gas operations, sensors are typically located on the surface of the earth. To test the workflow in a meso-scale hydraulic fracturing context, the analysis was repeated using a sensor (OB-15) that was placed at a considerable distance from the stimulated reservoir volume, on a monitoring well parallel to the well containing the OT-16 sensor. The OB-15 sensor is in the same line of sight as OT-16 (as shown in Figure 38), and the azimuthal variations in the signal properties due to wave propagation in an anisotropic medium are expected to be minimal between the two sensors. This allows for a more realistic simulation of the conditions that are commonly encountered in industrial applications.

Since the OT-16 and OB-15 sensors are collinear, it is assumed that any variations in the time-frequency characteristics of signals measured at the two sensors are due solely to the distance from the source. This configuration allows for the assessment of the effect of distance on the UMAP embeddings derived from the signal features. The results shown in Figure 39 indicate

that the STFT features combined with UMAP dimension reduction produce robust and widely applicable embeddings, even when the sensors are located at a considerable distance from the source. The separation of different fracture planes in the UMAP space is as good as those obtained from the OT-16 sensor, which is in the middle of the microseismic cloud. This suggests that the STFT and UMAP method is suitable for a wide range of source-sensor relative distances.

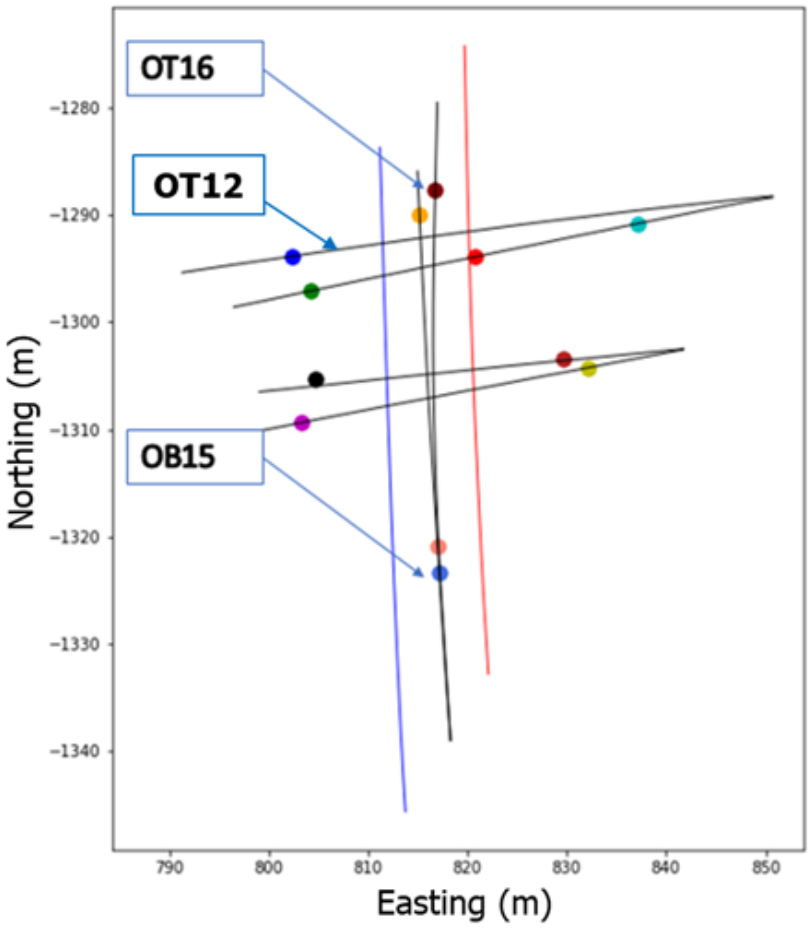

Figure 38: Location of OT-16 and OB-15 accelerometers and OT12 hydrophone in the EGS Collab experiment 1 testbed. OT-16 and OT-12 are located close to stage 1 microseismicity (around northing -1280 m) whereas accelerometer OB-15 is relatively distant from stage 1 microseismicity. The sensors are cemented in place inside the monitoring wells. Black lines represent monitoring wells. Blue and red lines indicate injection and production wells respectively.

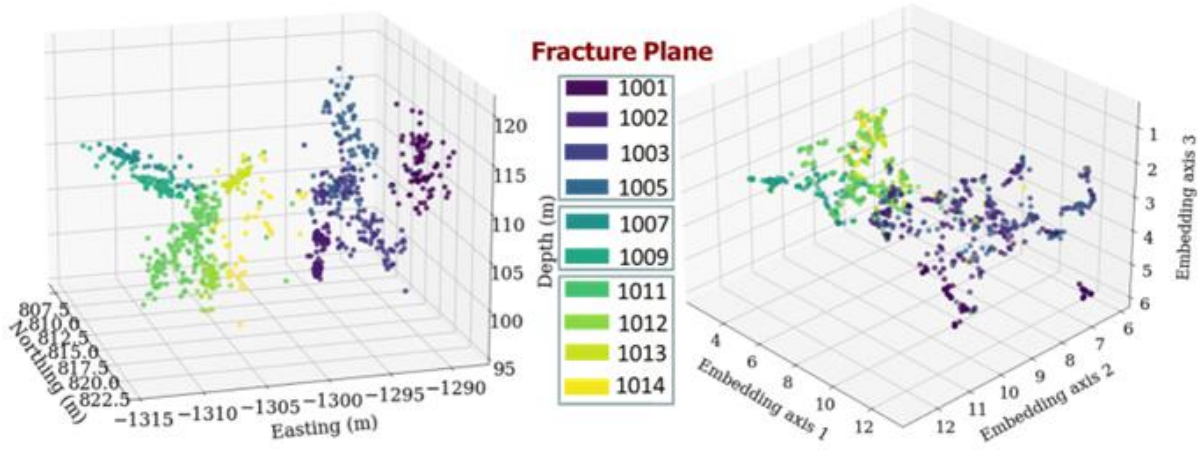

Figure 39: Microseismicity locations (left) and UMAP embeddings (right) derived from OB-15 accelerometer signals. Colors correspond to distinct fracture planes.

So far, we have demonstrated that the proposed workflow is effective for analyzing accelerometer signals that measure particle motion. In the following section, we extend the workflow to the analysis of pressure signals. Pressure waves are of scientific and commercial interest and are commonly used in underwater and marine environments. The use of pressure transducers in the EGS Collab experiment was an attempt to determine the suitability of hydrophones for measuring hydraulic fracturing-induced seismicity signals and related mechanical deformation signatures. It has been shown that pressure waves generated from subsurface fracturing phenomena can provide valuable information about fracture monitoring and characterization.

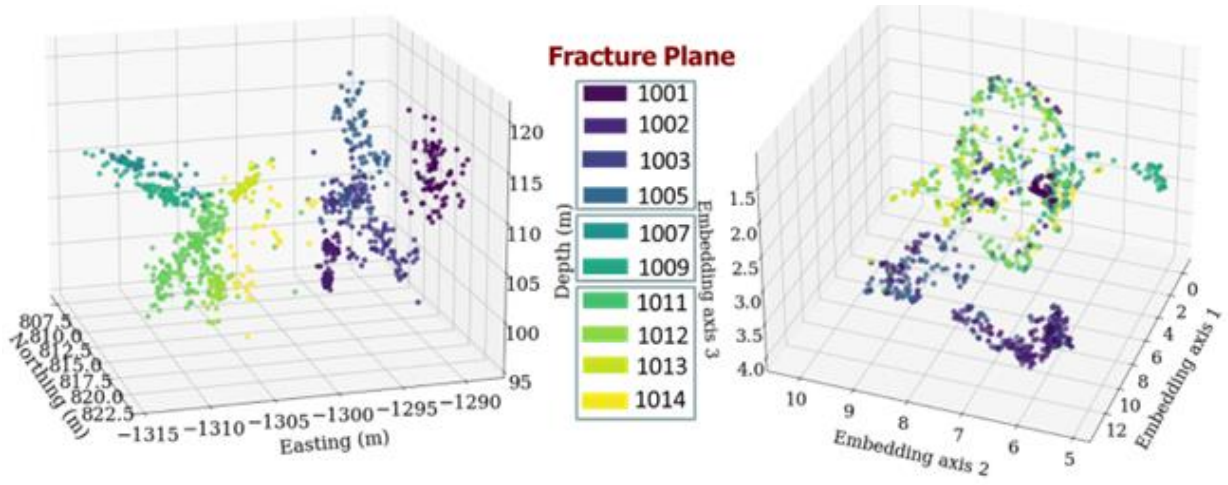

Figure 40: Microseismicity locations (left) and UMAP embeddings (right) derived from OT-12 hydrophone (pressure transducer) signals. Colors correspond to distinct fracture planes.

The propagation of mechanical waves is accompanied by the propagation of pressure waves that travel at the speed of P-waves in solid media. This pressure is like hydrostatic pressure due to wave propagation in fluids and can be measured using hydrophones or pressure transducers. In this study, a hydrophone (OT-12) was located on the same monitoring well as the accelerometer OT-16. Like accelerometers, the sensitivity of hydrophones is dependent on the coupling between the sensor and the medium (cement). The signal-to-noise ratio of hydrophones is generally lower than that of accelerometers, but hydrophones have a broader sensitivity response (from 2 Hz to 20 kHz). Figure 6 shows the corresponding UMAP embeddings derived from the signal features measured using the hydrophone OT-12. The embeddings (as shown in Figure 40) show excellent separation in terms of the underlying fracture planes, indicating that pressure waves also carry diagnostic signatures related to their location within a fractured network.

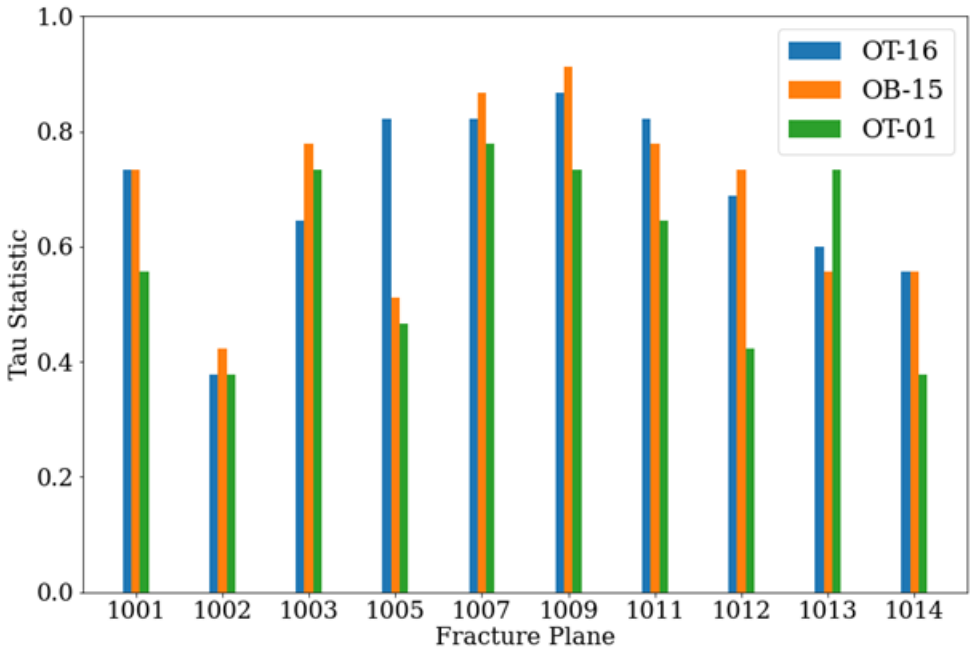

Figure 41: Kendall Tau statistic for three sensors corresponding to all the fracture planes considered in the study. OT-16 (highest SNR) shows best overall performance.

It is important to confirm the correspondence between pairwise distances in physical and UMAP space. In other words, for a given fracture plane, the order of distances to other fracture planes should be the same (or as close as possible) to the order of distance of corresponding data points in UMAP space. The "distance" in this context refers to the Wasserstein distance, rather than the Euclidean distance. Figure 41 shows the Tau statistic for the three sensors considered in this study. In general, the accelerometers have a better Tau statistic compared to the co-located hydrophone. Overall, the Tau statistic is above 0.5, indicating a strong correspondence between the pairwise fracture cluster "Wasserstein" distances in physical (cartesian) and UMAP space. Fracture plane 1002 shows the consistently lowest tau scores due to its geometric indistinguishability from planes 1001, 1003 and 1005. Further explanation is provided in the appendix.

6.3.1 Conclusions

1. Fractures are important for subsurface mass and energy transport and are crucial for safe and efficient energy operations in the hydrocarbon and geothermal industry. The networks of fractures created by hydraulic fracturing have complex geometries, and it is essential to confidently identify different branches of the fracture network for effective and consistent characterization of hydraulic fractures.

The current study aims to answer the question of whether passive seismic signals contain diagnostic signatures of their locations within the fracture network, and how much information about the underlying fracture planes can be extracted from the waveforms through representation learning. Another key contribution of the analysis is the development of a modified UMAP algorithm optimized for application in microseismic datasets with available fracture label information.

2. We used data from a unique hydraulic fracturing experiment that had detailed measurements to identify the location of fractures. We developed a process that involves analyzing data from a single sensor and reducing the amount of information using a specific algorithm called UMAP.

We also made changes to the UMAP algorithm to make it more effective for analyzing data from hydraulic fracturing operations. These changes involve choosing the most suitable settings for the algorithm, which can impact the results obtained from the sensor data.

3. We found that data from accelerometers at various distances from the seismic activity have accurate information about the location of fractures. Additionally, we discovered that pressure data from hydrophones also have useful information about the location of the microseismic event. This is the first time that pressure data from a non-marine setting has been used in this way. Overall, we showed that using an algorithm like UMAP and specifically

designed features can effectively uncover the structure of the signals in an unsupervised learning approach.

6.4 Low Frequency Seismic in EGS Collab Experiment 1

The fourth part of this thesis addresses the following question:

"Can near-infrasound and infrasound signals be used to better characterize the mechanical deformation processes associated with hydraulic fracturing?"

The key takeaways from this study are as follows:

1. The joint analysis of infrasound and microseismic encapsulates frequencies on the observable bounds of acquisition instrumentation (2 Hz to 15000 Hz). As a result, both high and low frequency fracturing phenomena driven by fluid injection are captured.

2. The joint data (microseismic and infrasound) reflects fluid injection-induced subsurface deformation that lies on a continuum - with one end representing of high frequency, small-scale shear slippage on fractures and the other end representing low frequency, large-scale void volume dilation or contraction.

3. It is hence concluded that microseismicity and infrasound signals contain complementary information about rock deformation due to fluid injection, and their joint analysis renders a more complete picture of the stimulated fractures in subsurface.

Two sets of hydrophones, each consisting of 12 hydrophones, were used to record infrasound and infrasound emissions during the injection of fluids. One set of hydrophones, E1-OT, was positioned perpendicular to the point cloud of microseismic activity and intersected it, while the other set, E1-PDB, was positioned subparallel to the cloud but did not intersect it (Figure 42).

The hydrophones in the E1-OT set were closer to the fluid-induced deformation and therefore recorded a stronger infrasound signal intensity than the E1-PDB hydrophones, which recorded a signal energy that was about five orders of magnitude weaker. On May 24, the fluid injection caused hydraulic fracture propagation until it intersected the production well, leading to a decrease in microseismicity. Later experiments mostly involved fluid flow through a fractured volume with a lower rate of microseismicity.

The change from fracture propagation to fluid flow through a fracture is reflected in the nature of the cumulative signal energy, with impulsive energy release (indicative of stick slip fracture propagation) dominant on May 24 and long-period infrasound tremors (indicative of long duration energy release) more common during fluid flow through fractures. A strong relationship was consistently observed between the cumulative injected volume and cumulative signal energy (Figure 43), suggesting that the infrasound signals are generated by fluid-driven processes. Importantly, both sets of hydrophones, regardless of their distance from the microseismic cloud, showed similar behavior in terms of the nature of the recorded energy, even though the E1-OT hydrophones recorded a much higher energy overall. This suggests that the energy recorded at different locations may have the same characteristics but differ in scale.

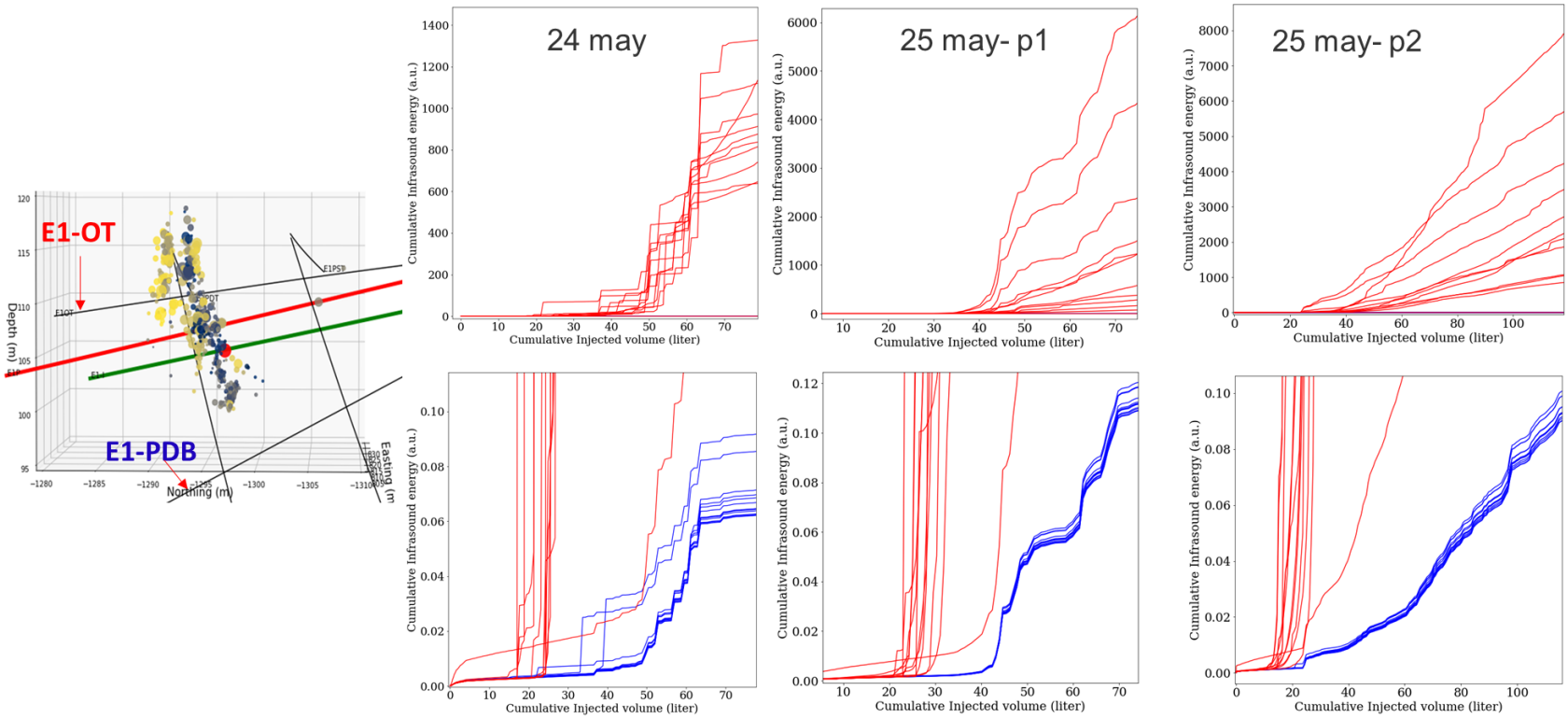

Figure 42: Dependence of fluid injection rate on infrasound energy release. E1-OT is situated perpendicular to the fractured zone and is intersected by it whereas E1-PDB is lies subparallel to the fracture and further away than E1-OT. With time, the gradient of cumulative infrasound energy for the hydrophone strings becomes progressively smoother.

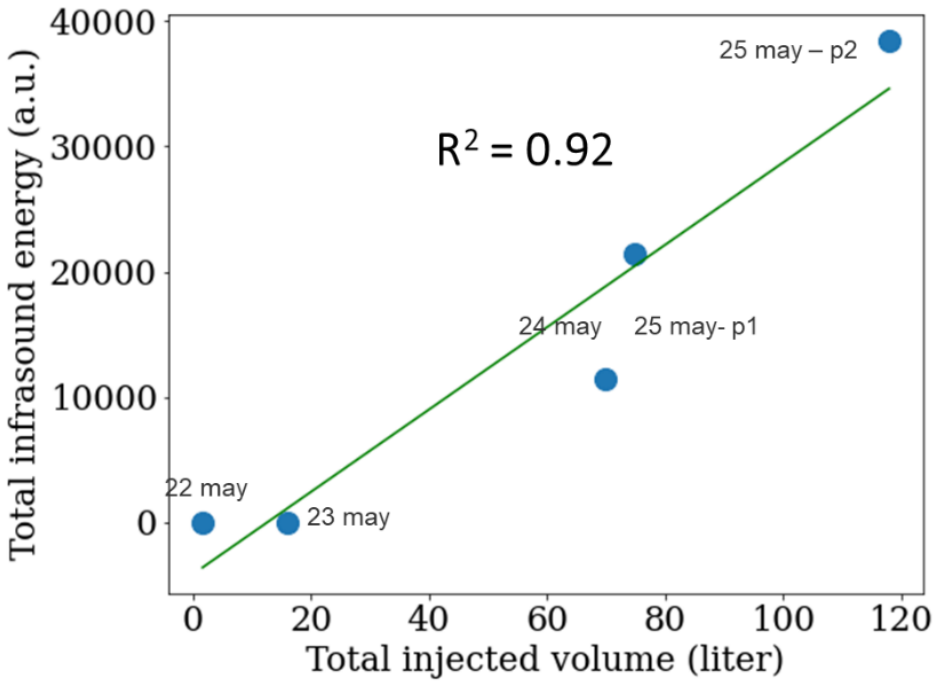

Figure 43: Dependence of cumulative infrasound (2-80 Hz) measured by combined hydrophone arrays (located on the monitoring wells E1-OT and E1-PDB).

On May 22nd and 23rd, the maximum rate at which the substance was injected was 200 mL/min and 400 mL/min, respectively, but only very weak infrasound signals were detected. On May 24th and the two parts of May 25th, when the maximum injection rate was 4.5 L/min, the strongest infrasound signals were detected. Figure 44 shows where the infrasound sources were located. After using filters to analyze the initial data from the grid search using cross correlation, a total of 322, 818, and 1117 infrasound sources were identified for the three stimulations.

6.4.1 Spatiotemporal variation of locations

The infrasound sources in Figure 44 show a temporal evolution with respect to the injection well. On the first day, the sources spread out perpendicular to the injection well, but around a certain time they become concentrated along a line sub parallel to the injection well. As the events

continue, the sources migrate northward and align in the same direction. The microseismicity shown in Figure 44 corresponds with the infrasound sources, with a point cloud situated at a specific distance from the injection point and overlapping with the infrasound sources north of the injection point. On the second day, the infrasound sources have a less diffuse distribution and are primarily located along an east-west trend south of the injection point, with a sparse group of sources on the north. At the start of injection on the second day, the infrasound sources fall on two sub parallel lineaments on either side of the injection point, being sub perpendicular to the injection well. Later events on this day align sub parallel to the injection well.

6.4.2 Joint analysis of Discrete Fracture Network and Infrasound Source Locations

Fusing complementary imaging techniques, such as active and passive seismic, can improve the imaging of fractures. Combining the analysis of high and low frequency components of deformation can also provide more information about fractures than can be obtained from individual methods. In the case of fluid injection, which can cause cracks to open in fractured rock, the injection of pressurized fluid can cause the crack to expand or contract like a diaphragm, producing mechanical waves. This process can generate both high frequency shear motion (microseismicity) and low frequency P-waves. This conceptual model of fluid-driven infrasound generation is illustrated in Figure 45.

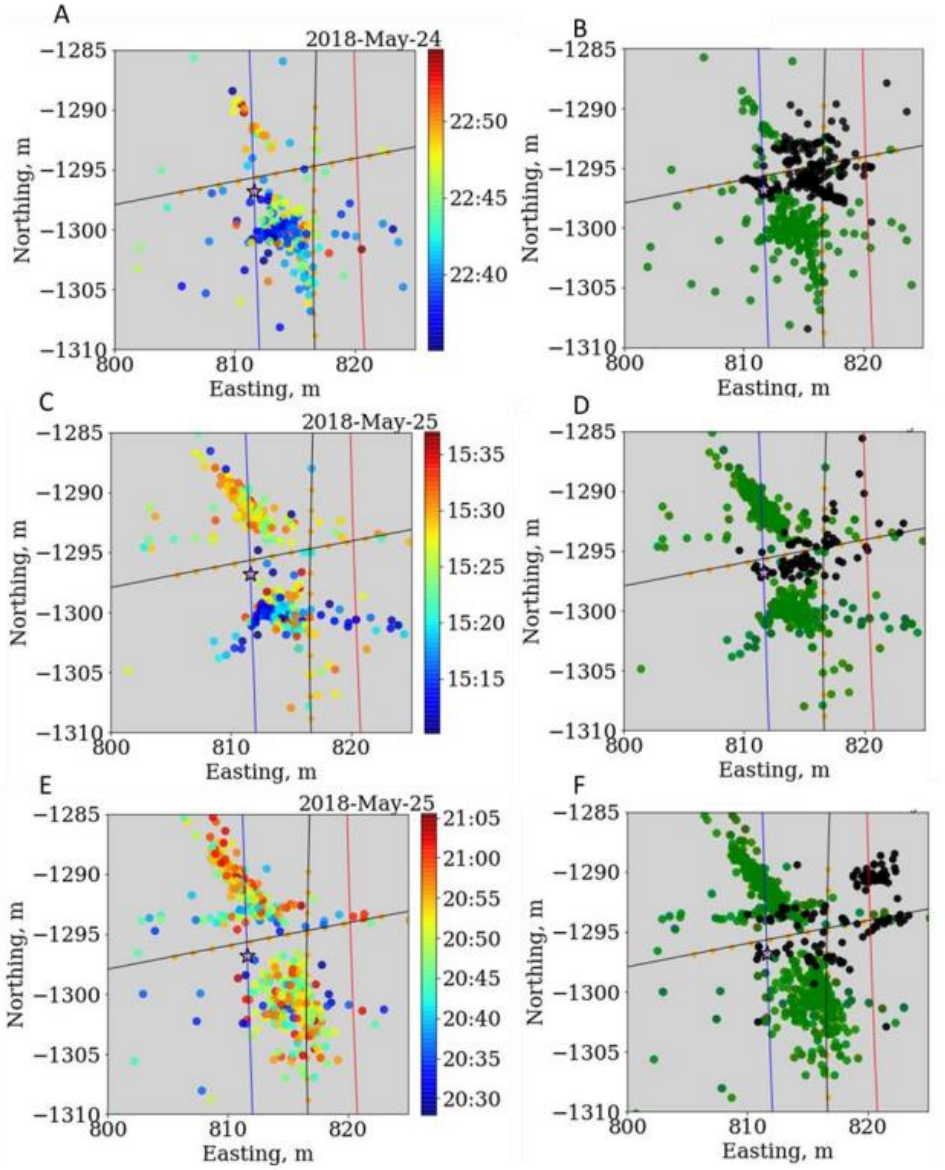

Figure 44: Infrasound source locations on a, b) May 24; c, d) May 25 part 1; and e, f) May 25 part 2. Colored points show infrasound while black points show the simultaneously recorded microseismicity. Blue and red lines indicate injection and production wells respectively. Pink star on the injection well E1-I marks the injection point. Black line subparallel to injection and monitoring wells is hydrophone string E1-PDB, and sub horizontal line is string E1-OT. Orange squares overlain on lines mark the hydrophone sensors emplaced in the monitoring wells.

Figure 45 shows the orientation and distribution of natural fractures in a discrete fracture network. These fractures, which are oriented roughly in the direction of the least horizontal stress and the injection well, are the most likely to be pressurized with fluid, as shown in Figure 4C.

The infrasound source has two main directions of activity, with the dominant direction being 140° counterclockwise from east and the minor direction being east-west. The east-west trending fractures were created through hydraulic fracturing. It is also noteworthy that a significant portion of the infrasound activity is not located near the production well, which suggests that there are different fluid pathways present. The area with microseismicity is where fluid interactions mobilize critically stressed cracks, causing shear motion, while the area with infrasound activity is likely to be pressurized natural fractures that generate low-frequency compressional motion. This complexity in the stimulated rock volume, as opposed to the idealized penny-shaped fracture, is supported by the high amount of fluid leak off observed in the fractured formation, as indicated by the large difference between injected and produced water volumes. The final conceptual model consists of the microseismic and low frequency seismic generation in highly fractured rock and intact rock.

In case of highly fractured rock, we find strong, significant occurrence of low frequency seismic, followed by generation of microseismicity. Moreover, since the media is highly fractured, the high frequency signals of microseismic are attenuated a much greater rate than they would be in an intact rack. So even if microseismic signals are generated, their measurement by sensors is diminished due to increased attenuation. On the other hand, the low frequency signals are attenuated far less, due to their lower frequency. This dual effect of fractures leads to significant infrasound measurement and relatively low microseismic measurement by distant sensors. In case of intact rock, there is negligible low frequency seismic generation and relatively greater occurrence of high frequency microseismic. Moreover, the attenuation of high frequency signals is far less in a intact rock compared to a fractured rock. So whatever microseismic signals are generated, a relatively large portion of them is measured by sensors located at a distance. Hence

in an intact environment, not only more high frequency signals are generated, but also their acquisition is far more compared to a fractured environment.

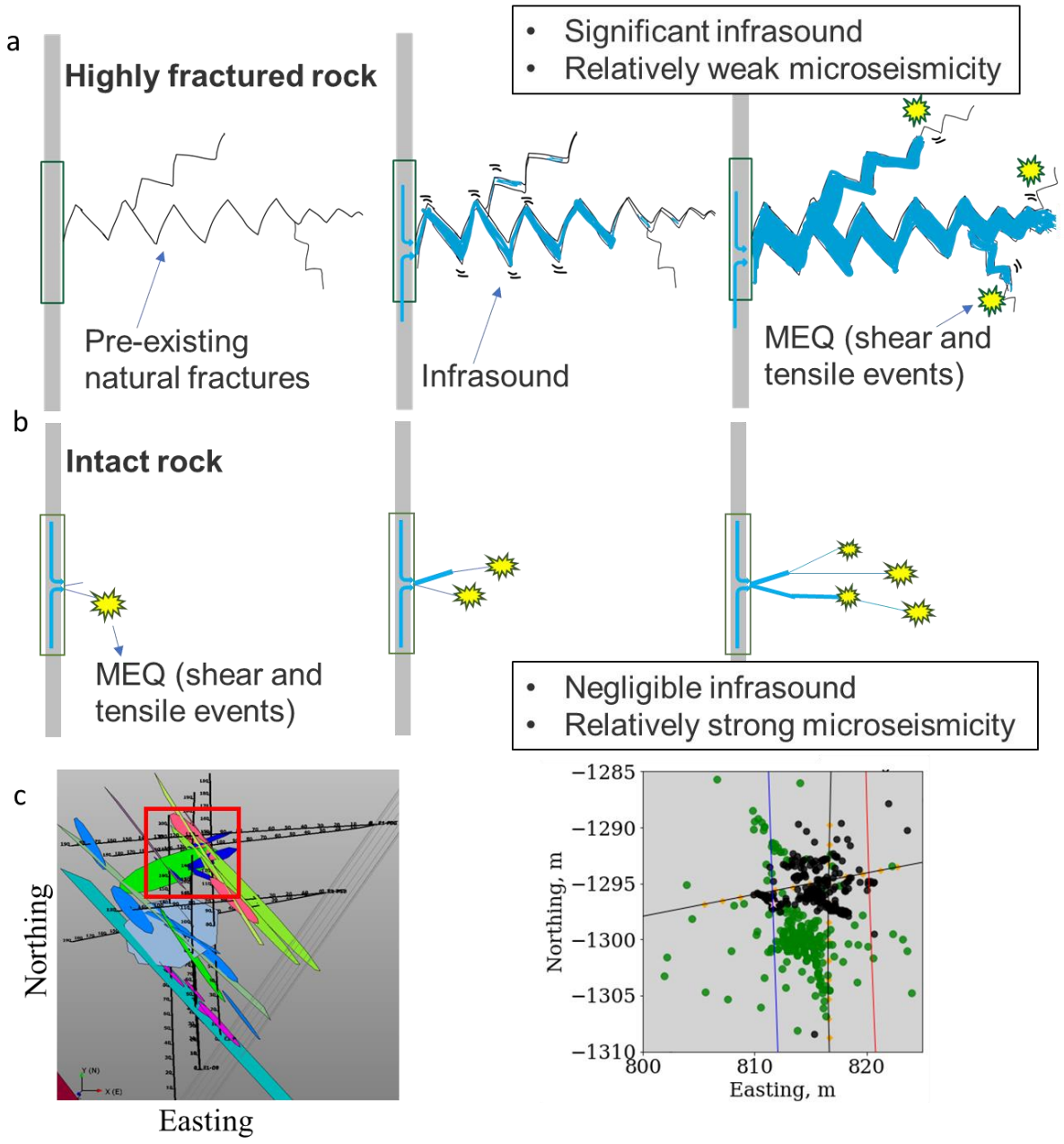

Figure 45: a) Schematic representation of the fluid-injection driven infrasound and microseismic energy release in a naturally fractured rock volume. MEQ's are microseismic events. b) Schematic representation of the fluid-injection driven infrasound and microseismic energy release in an intact rock volume. c) Comparison with microseismic and discrete fracture network (DFN). The interpreted network shows the orientation of

pre-existing natural fractures in the testbed, with large majority of the features inclined at 140° counterclockwise from east. Red box highlights the area of located infrasound activity. c) Their combined location cloud shows strong agreement with overall orientation inferred from the DFN.

6.4.3 Conclusions

1. The study analyzed low-frequency hydrophone signals captured during a hydraulic fracturing experiment at a depth of 1.5 km and a meso-scale of around 10 m. The hydrophone array was used to detect infrasound sources that had not been previously identified. The infrasound was found to be caused by the injection of fluid. Three different stimulations were conducted, resulting in a total of 322, 818, and 1117 infrasound source locations being identified.
2. The energy release at the beginning of the stimulation was found to be associated with the propagation of fractures, while later stages of the stimulation resulted in a smoother release of energy, which was linked to fluid flow in conduits causing tremor-like motions.
3. The study found that infrasound signals with a usable signal to noise ratio were only produced at high fluid injection rates. Since the infrasound is an emergent signal, traditional threshold-based methods for detecting the first arrival were found to be unreliable. To locate the infrasound source locations, the study used a data-driven cross-correlation-based grid search method. Four filtering steps were applied to improve the accuracy of the source location algorithm. These filters included: thresholds based on the array power, thresholds based on the misfit in the cross-correlation based grid searching, scatter in locations obtained from station bootstrapping, and upper and lower bounds on the normalized cross correlation coefficient.
4. After obtaining the final locations of infrasound sources, the study analyzed how these source locations evolved over the course of three episodes of fluid injection. It was observed that the infrasound hotspots shifted around the fluid injection point during the fracturing operations.

Some locations were found to produce exclusively one type of signal, while others produced both infrasound and microseismicity, indicating that these locations were experiencing both high and low frequency deformation from fluid injection. By comparing the spatiotemporal evolution of the infrasound sources to the microseismic sources and the discrete fracture network model, the study suggested that the pressurized fluid was causing the volume of the fractures to expand or contract depending on whether fluid was being injected or drained, and that this change in volume was generating compressional waves.

5. The study also found a strong agreement between the fracture orientations and infrasound source locations based on the discrete fracture network model of the testbed before fracturing. The study suggests that the pressurization of natural fractures is the most likely mechanism for generating infrasound. The observation that infrasound corresponds to fluid flow also indicates that a significant portion of the injected fluid is diverted away from the intended location, such as the production well. However, it is important to note that the location method used in the study only outputs the location of infrasound sources in two dimensions, which is a limitation of the study.

6. It is widely understood that microseismicity only captures a small fraction of the input hydraulic energy and only partially images the fracture network. By combining the analysis of infrasound and microseismic data, the study can capture both high and low frequency fracturing phenomena driven by fluid injection. The joint data reflects a continuum of fluid injection-induced subsurface deformation, with one end representing high frequency, small-scale shear slippage on fractures, and the other end representing low frequency, large-scale void volume dilation or contraction. Thus, it is concluded that microseismicity and infrasound signals contain

complementary information about rock deformation due to fluid injection, and their joint analysis provides a more complete picture of the stimulated fractures in subsurface.

6.5 Semi-Supervised Label Propagation (EGS Collab)

The fifth part of this thesis addresses the following question:

"How to apply semi-supervised learning methods on microseismic locations with limited labelled data and massive unlabeled data to improve the passive seismicity-based monitoring of the subsurface during hydraulic fracturing?"

The key takeaways from this study are follows:

1. Graph-based label propagation is applied to the reduced features derived from the microseismic signals to extend identities of fracture planes to unlabelled hypocenter locations.

2. The performance of label propagation weakens at the smallest sizes of training data i.e., 5 % and stabilizes near the 20 % in terms of both precision and recall.

3. Precision and recall values stay somewhat constant until one standard deviation after which both the quantities show a considerable downward shift. It is thus concluded that the algorithm is well tolerant to locational errors of up to one standard deviation in each direction.

The first part of result refers to determining the smallest training data size that yields optimal performance in terms of precision. Tabular plots are chosen to represent results as for every run (i.e., train/test data size), 20 iterations are performed, each iteration considering a shuffled and stratified sample. The model used KNN kernel with number of neighbors hyperparameter set at 15. See methods section for workflow describing in detail the selection procedure of kernel type and kernel size. Since the objective is to determine size with highest median value and low

variability, we consider the ratio of median and variance (MVR). The test fraction showing highest ratio of median, and variance is deemed the optimal test fraction for the label propagation. The performance of label propagation seems to deteriorate as the at the smallest sizes of training data i.e., 5 % and stabilizes between 15 and 25 % (Figure 46 and 47) in terms of both precision and recall.

	Test_fraction	Mean	Median	Variance	MVR
0	0.050000	0.954949	0.955996	0.000881	1084.619125
1	0.100000	0.955122	0.954196	0.000520	1836.037875
2	0.150000	0.953167	0.953846	0.000158	6018.707341
3	0.200000	0.955425	0.967998	0.000362	2675.195408
4	0.250000	0.954360	0.953125	0.000419	2274.935189
5	0.300000	0.954669	0.953846	0.000185	5145.437227
6	0.350000	0.954112	0.953125	0.000215	4433.087784
7	0.400000	0.957496	0.954885	0.000248	3845.337189

Figure 46: Tabular results showing the variation in precision of label propagation algorithm with fixed testing data size of 0.6 for fracture plane 1002. The MVR (median to variance ratio) is considered to choose the optimal test fraction, here seen as 15%.

	Test_fraction	Mean	Median	Variance	MVR
0	0.050000	0.638324	0.636318	0.004978	127.822514
1	0.100000	0.667671	0.657277	0.003712	177.075099
2	0.150000	0.651310	0.647313	0.002808	230.549930
3	0.200000	0.656791	0.651389	0.003713	175.422241
4	0.250000	0.657492	0.662236	0.003491	189.693385
5	0.300000	0.670017	0.666667	0.003147	211.867720
6	0.350000	0.670983	0.651899	0.003006	216.842924
7	0.400000	0.675097	0.666667	0.002903	229.608251

Figure 47: Tabular plots showing the variation in precision of label propagation algorithm with fixed testing data size of 0.6 for fracture plane 1005. The optimal test fraction is deemed as 15% that corresponds to the highest median to variance ratio (MVR).

The second part of analysis addresses the effect of locational uncertainty on the performance of the label propagation algorithm. Error is introduced on the microseismic locations which is then used as an input feature along with three component signal spectra for UMAP based dimension reduction. These locations, along with the three component signal spectra as used as input features for the UMAP-based dimension reduction. The reduced dimension embeddings obtained from UMAP are the dataset on which label propagation is performed. For a given sigma multiplication factor, the dataset is split into 0.1 training and 0.9 ratio and stratified based on the target variable. The label propagation algorithm is applied using knn kernel type and number of neighbors equal to 15. The performance for individual fracture plane (target label) is quantified based on the precision and recall obtained on the testing set.

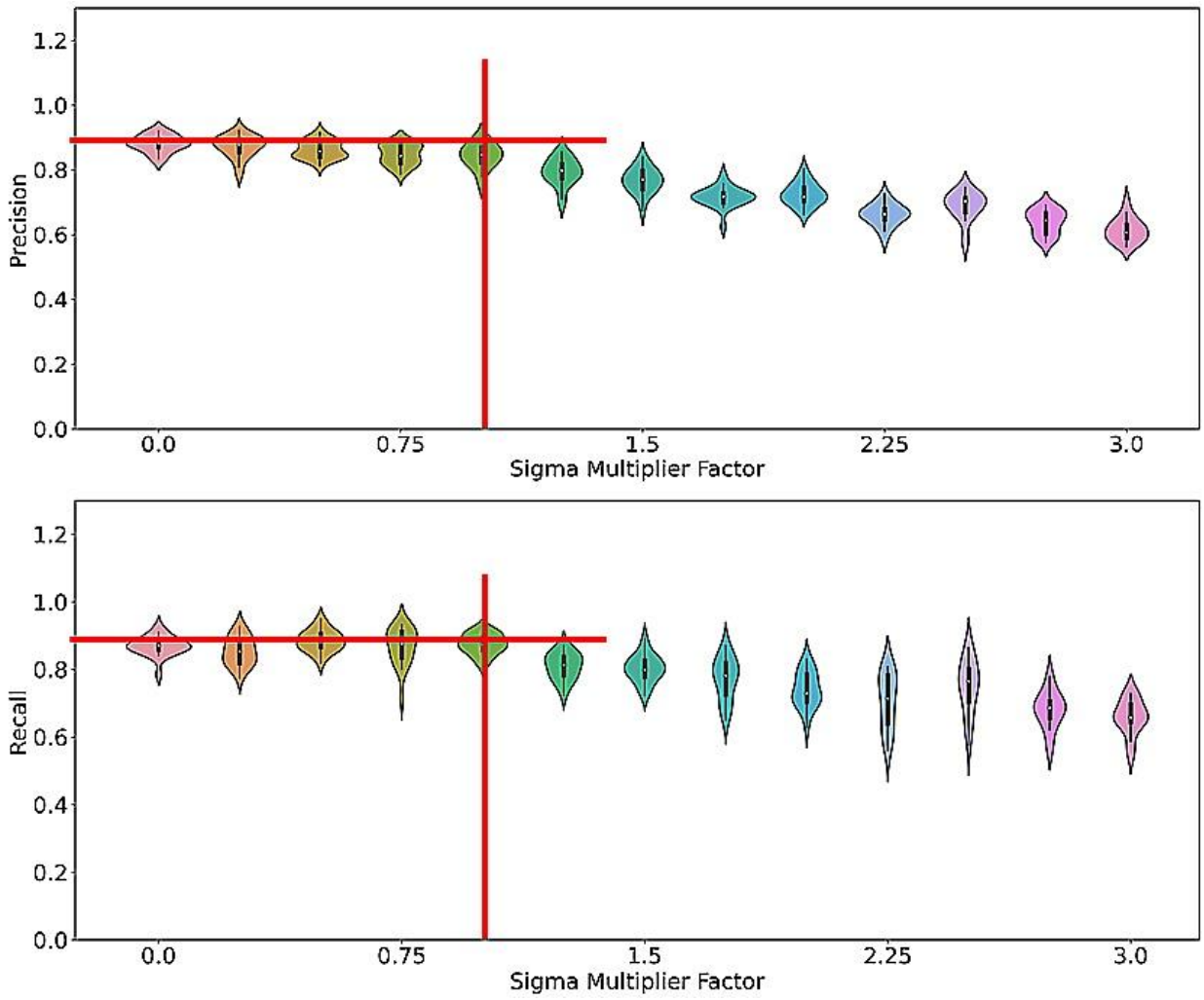

Figure 48: Violin plots showing the effect of locational error on precision and recall of label propagation. Results are shown for fracture plane 1011 using 10 % training data.

Figure 48 shows the effect of sigma multiplier factor on the precision and recall of the label propagation algorithm. A multiplier factor of 1 implies one standard deviation error being added to the original location in each direction. The precision and recall for a given fracture plane reduces as the multiplier factor is increased from 0 to 3. It is observed that the precision and recall values stay somewhat constant until one standard deviation after which both the quantities show a considerable downward shift. It is thus concluded that the algorithm is well tolerant to locational errors of up to one standard deviation in each direction.

6.5.1 Explaining differences in the performance of label propagation on fracture clusters

The differences in the precision and recall of label propagation algorithm for different fracture planes depends on the geometrical distribution of the UMAP embedding of input data. The input data comprises of signal features and the locations. Therefore, the physical location of microseismic point clouds has an important contribution to the overall performance of label propagation. The likelihood of correct label propagation depends on the geometry (shape) of the point clouds and its position and orientation relative to the other fracture planes.

For every cluster of points corresponding to different fracture planes, the following properties are calculated:

1. Density: calculated as the number of points divided by the sum of pair wise Euclidean distances for every point in the cluster
2. Minimum and sum Wasserstein distances: See calculation of Wasserstein distance in the methods. Minimum distance is the Wasserstein distance to the closest cluster and sum is defined as the summation of pair wise Wasserstein distances to all the other clusters. Low minimum distance greater chances of interference between nearby cluster whereas higher minimum distance implies lesser interference and hence greater confidence in correct label assignment by the label propagation algorithm. Greater sum of Wasserstein distance also implies greater confidence in correct label assignment by the label propagation algorithm and vice versa.
3. The degree of planarity is quantified the residual of the least squares plane fit to the cluster of points. Higher residual implies lesser planarity and vice versa.

Table 7: Quantifying the density, inter-cluster distances and geometry (planarity) of different hydraulic fracture clusters comprised of microseismic point clouds

Fault ID	Density	Min WD	Sum WD	Residual
1001	0.294298	7.265011	142.484174	17.152570
1002	0.398813	5.949746	96.774637	9.101500
1003	0.229644	5.442346	97.247523	15.697979
1005	0.212581	5.442346	106.999150	21.658784
1007	0.316281	2.417573	111.814211	5.054897
1009	0.266666	2.417573	104.348146	4.566540
1011	0.210551	4.059514	92.804626	32.695144
1012	0.486272	4.059514	102.735965	10.494594
1013	0.779211	7.340233	86.257510	3.376093
1014	0.179906	7.268234	86.625853	18.152496

Table 7 summarizes the results of geometric analysis. Fault ID's 1005, 1007 and 1009 show the poorest performance in the label propagation algorithm and show least values for minimum Wasserstein distances. Low minimum Wasserstein distances is a strong indicator that neighboring fracture planes are overlapping or intersecting with these fracture planes. As a result, there is relatively higher probability of incorrect label assignment by algorithm that considers the location of points, along with other quantities.

Similarly, the measured signal properties of closely laying, or overlapping and intersection fractures planes are also bound to have greater similarity compared to signals from two points that lie further apart. Both the features (location and signal properties) are bound to highly similar for fracture planes which show low minimum Wasserstein distances. The minimum Wasserstein distance, therefore, is a reliable metric to explain the output of difference microseismic fracture planes.

6.5.2 Conclusions

1. Smallest size of training data for a fixed test data size: to quantify the least training data size, the testing data size was fixed at 0.6 and the training data size was varied between 0.05 to 0.4. the splitting was stratified based on the target variable. The label propagation was performed using KNN kernel with number of neighbors equal to 15. The performance was quantified based on the precision and recall values obtained for the testing set for each fracture label. The performance of label propagation weakens at the smallest sizes of training data i.e., 5 % and stabilizes near the 20 % in terms of both precision and recall.

2. Both graph-based label propagation kernel types ‘RBF’ and ‘KNN’ yield similar results for semi supervised label propagation task i.e., estimating fracture plane from the microseismic features.

3. Effect of locational uncertainty on the performance of semi-supervised algorithm performance: Error is introduced on the microseismic locations which is then used as an input feature along with three component signal spectra for UMAP based dimension reduction. For a given sigma multiplication factor, the dataset is split into 0.1 training and 0.9 ratio and stratified based on the target variable. The label propagation algorithm is applied using knn kernel type and number of neighbors equal to 15. It is observed that the precision and recall values stay somewhat constant until one standard deviation after which both the quantities show a considerable downward shift. It is thus concluded that the algorithm is well tolerant to locational errors of up to one standard deviation in each direction.

4. Explaining differences in the performance of label propagation on different fractures: Fault ID’s 1005, 1007 and 1009 show the poorest performance in the label propagation algorithm and show least values for minimum Wasserstein distances. Low minimum Wasserstein distances is a strong indicator that neighboring fracture planes are overlapping or intersecting with these

fracture planes. As a result, there is relatively higher probability of incorrect label assignment by algorithm that considers the location of points, along with other quantities. Similarly, the measured signal properties of closely laying, or overlapping and intersection fractures planes are also bound to have greater similarity compared to signals from two points that lie further apart. Both the features (location and signal properties) are bound to highly similar for fracture planes which show low minimum Wasserstein distances.

6.6 Uncovering Relationship Between Geomechanical Deformation and Wave Motion Through Clustering

The sixth part of this thesis addresses the following question:

"How to characterize regions with similar and dissimilar geomechanical deformation by analyzing the microseismic data captured by geophone array? Do dissimilar geomechanical alterations exist near each other within SRV? Do the different classes of microseismic events in a fracture plane exhibit relatively similar 3D motion?"

The key takeaways from this study are as follows:

1. Moment tensor properties are they information correspond to micro seismically-derived geomechanical deformation in a hydraulic stimulation.

2. Clustering followed by analysis of descriptive statistics of the moment tensors shows differences in the faulting style operating at different parts of the stimulative reservoir volume.

3. One dominant class of moment tensor is present in each strand of hydraulic fracture network.

4. Strong statistical correlation is seen between the three-dimensional wave motion and class of moment tensors. The correlation is quantified using Calinski-Harabasz index. Additionally, it is observed that three component wave motion is a better diagnostic of moment tensor class compared to polarization features of signal.

Input data: 530 hypo central locations and their signals recorded on 69 three component geophones located at depth of ~ 30 m and spread evenly above the SRV. The average depth of events is ~ 4 km. The 530 locations are showing little temporal grouping (Figure 49) and are clustered based on their density distribution to form four groups that represent spatially distinct fracture sets as shown in Figure 50. Zhang et al. 2021, have determined the moment tensors of the 530 locations to a high confidence and the detailed of their workflow can be found in (Zhang et al., 2019). The first part of the analysis is to answer the following research question: 1.) How to find hydraulically fractured regions that are geomechanically similar and those that are geomechanically dissimilar by analyzing the microseismic data captured by geophone array? and 2) Do dissimilar geomechanical alterations and damages exist near each other within SRV?

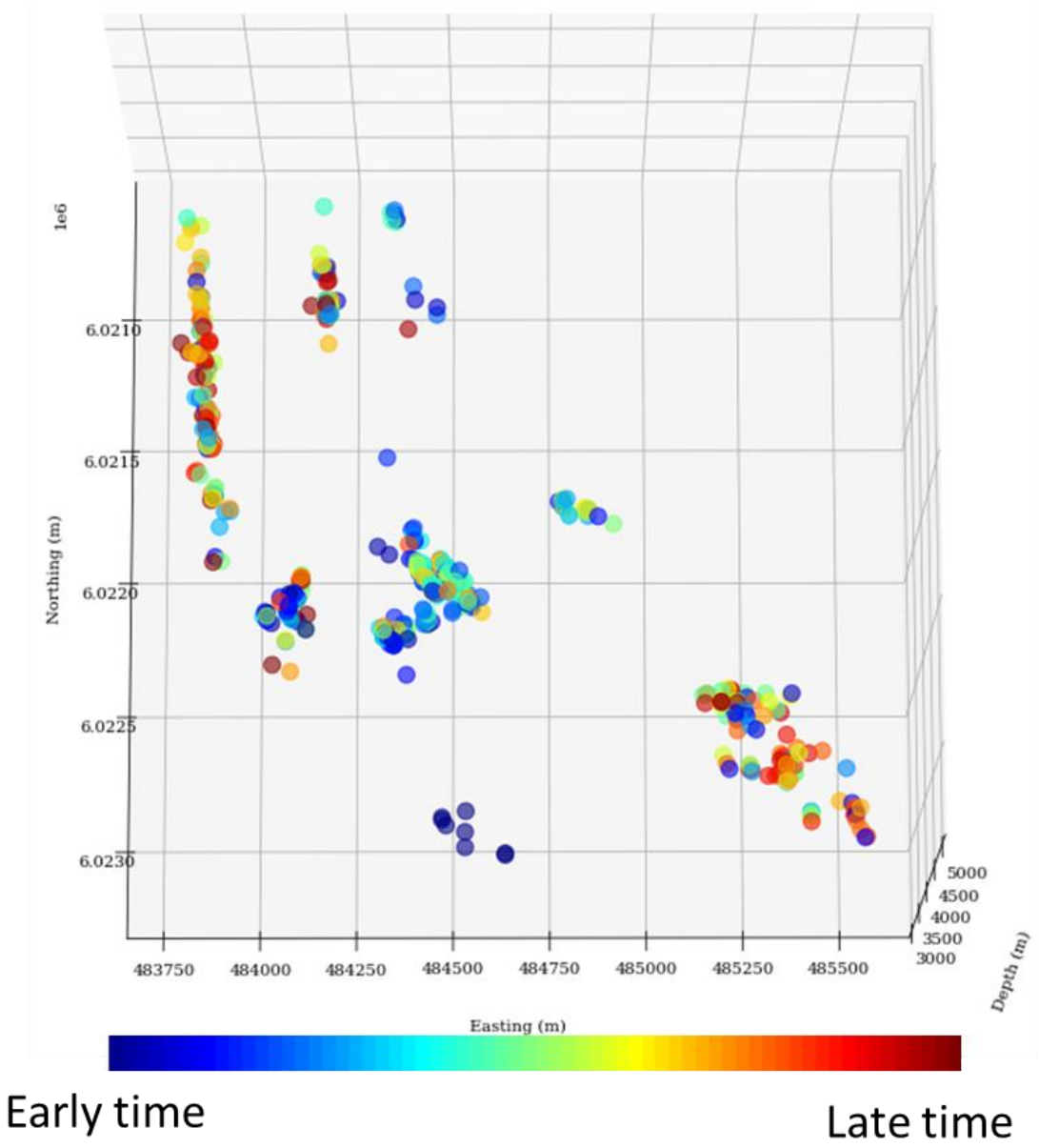

Figure 49: Plan view of the microseismicity recorded in the Toc2Me experiment in the Duvernay shale formation in 2018. Cooler colors show early event and hotter colors show later events. No clear spatiotemporal trend is observed in the microseismicity.

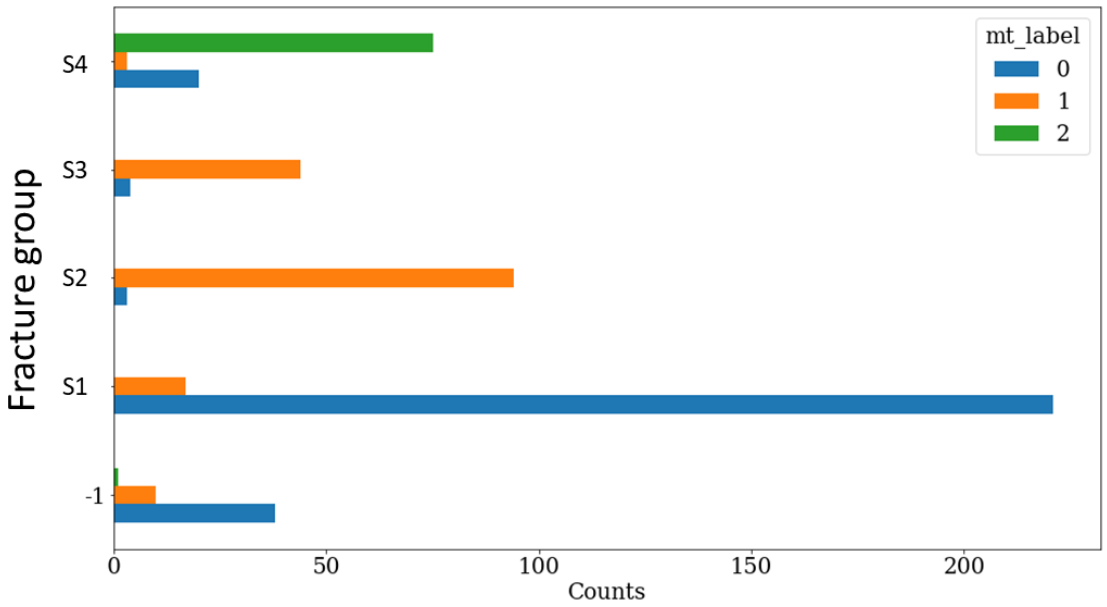

Figure 50: Fracture clusters in the Duvernay shale field scale microseismicity. Four spatially distinct clusters are present, represented by different colors. Grey points represent non-clustered events.

The moment tensor associated with an earthquake is represented by a symmetric 3 X 3 matrix. This implies that are the six distinct elements that uniquely define a matrix. The six elements are taken as input features for clustering moment tensor data. The scaling of inputs is done using RobustScaler functionality in Scikit-learn. Scaled data is embedded in three dimensions using UMAP (see previous methods sections for details on UMAP) and clustered using DBSCAN to obtain three clusters. The geophysically descriptive statistics of three clusters now follow.

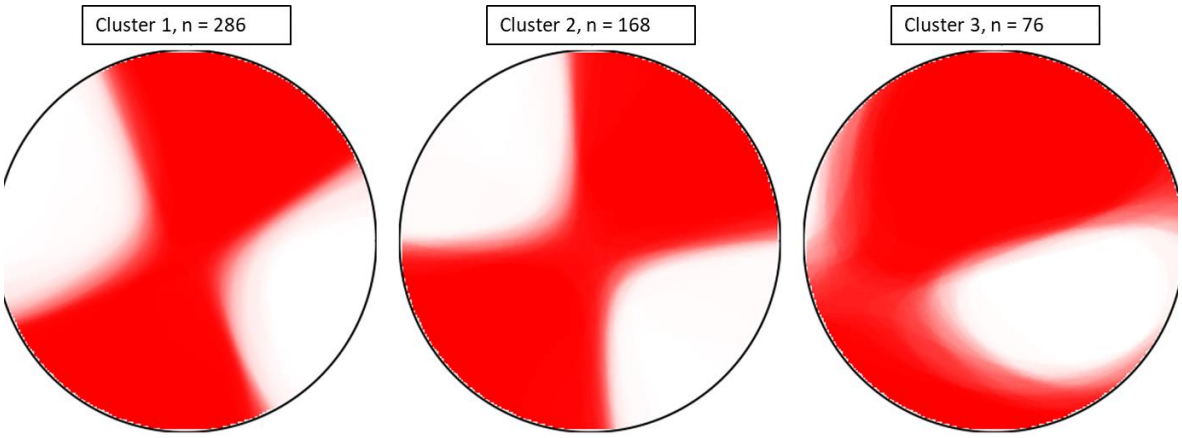

Figure 51: Fuzzy beachball representation of three clusters of moment tensors of the Toc2Me dataset. Within a single beachball, the fuzzy appearance is due to overlaying of several beachballs. Note the very small amount of intra cluster variability. The orientation of red and white axes on the beachball show relative position of faulting axes in the corresponding earthquake.

First part of descriptive statistics for the moment tensors are the corresponding beachball diagrams as shown in Figure 51. The moment tensor can be visualized using a beach ball diagram, which consists of a circle divided into quadrants, with each quadrant representing one of the three principal axes of the earth (X, Y, and Z). The orientation of the fault planes is depicted by the placement of the colors within the quadrants. For example, if the red quadrant is on the bottom and the white quadrant is on the right, this indicates that the earthquake was caused by a thrust fault that occurred along the X axis and a strike-slip fault that occurred along the Y axis. In the second step, the moment tensor is decomposed into isotropic and deviatoric components and deviatoric is further divided into DC and CLVD components using the methodology described in (ref) and implemented for the current dataset using the Pyrocko library in Python.

Label 2 shows much higher values of isotropic component compared to Labels 0 and 1, as shown in Figure 52. Isotropic components are associated with volumetric expansion which are

obviously tied to the dilation caused from fluid injection in subsurface voids. On the other hand, Labels 0 and 1 show much higher proportion of double couple component. Double couple motion is associated with more typical shear mechanisms that occur in tectonic earthquakes.

The CLVD moment tensor is used to describe faulting that is not purely isotropic (symmetrical) or pure shear (asymmetrical). Instead, it represents a combination of these two types of faulting, with the amount of each type being proportional to the magnitude of the three independent terms. Label 2 also shows the highest CLVD ratio and Label 1 shows the lowest CLVD ratio.

The deviatoric component of the moment tensor is the part of the moment tensor that represents the deviation of the faulting from a purely isotropic (symmetrical) state. It is calculated by subtracting the hydrostatic (isotropic) component of the moment tensor from the full moment tensor.

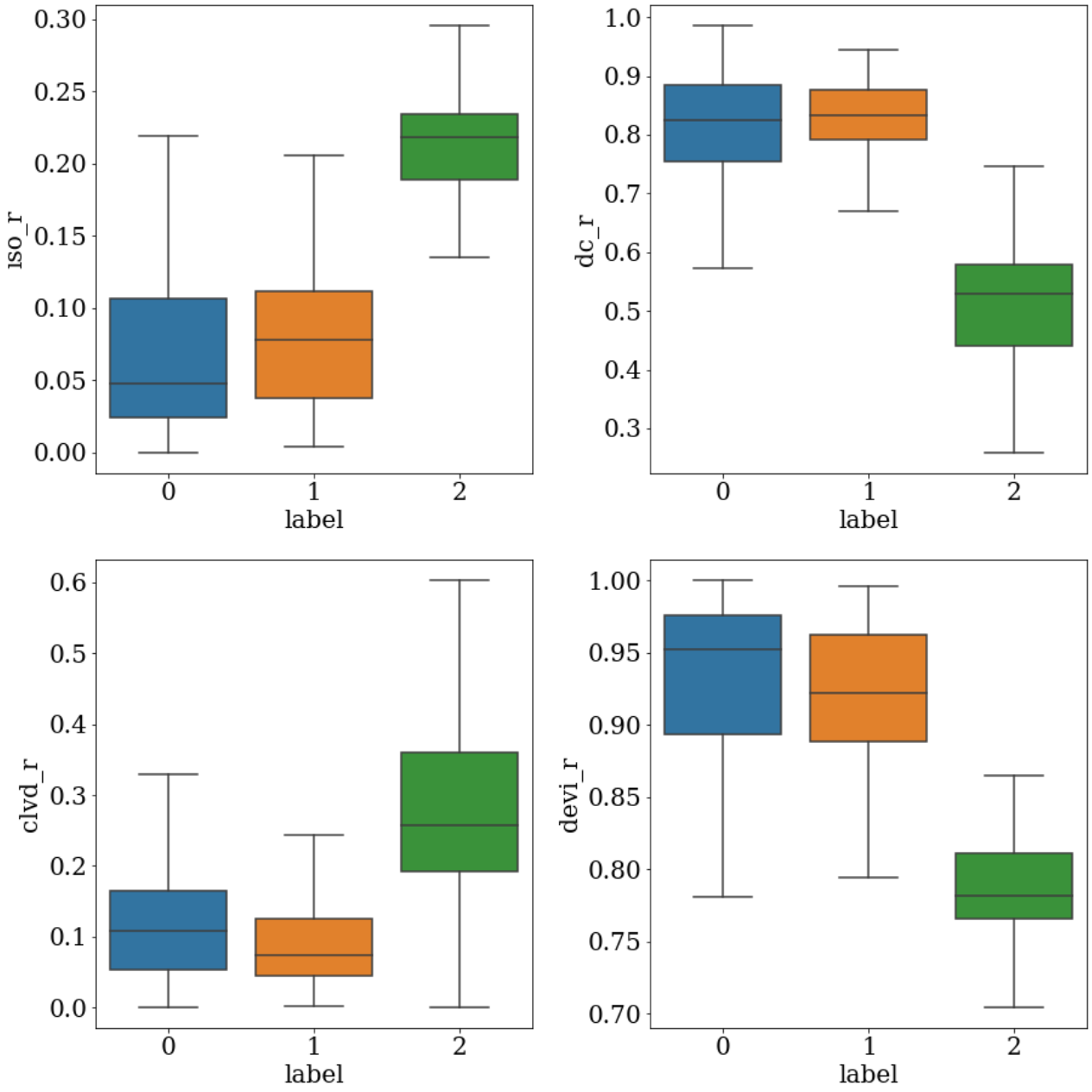

Figure 52: Box plots showing cluster-wise variation in the isotropic and deviatoric components of the moment tensor. Deviatoric components are further divided into double couple (DC) and compensated linear vector dipole (CLVD) components.

The hydrostatic component of the moment tensor represents the part of the faulting that is caused by uniform, isotropic stress, such as that caused by the weight of the overlying rocks. The deviatoric component, on the other hand, represents the part of the faulting that is caused by non-

uniform, anisotropic stress, such as that caused by tectonic forces.

The deviatoric component of the moment tensor is often used to study the anisotropic (asymmetrical) nature of earthquakes and the tectonic forces that caused them.

As label 2 shows the highest isotropic ratio, in turn it also shows the lowest deviatoric composition among the clusters. Both Labels 0 and 1, representing a largely double couple mechanism, show high ratios of deviatoric component in the moment tensors.

Overall, cluster analysis of moment tensor information establishes that label 2 has the highest expression of hydraulic stimulation induced volumetric expansion. Labels 0 and 1 are representative of more typical, tectonic earthquakes- events which are greater in proportion compared to label 0 but not directly associated with the volumetric expansion. Label 0 events are the type which are responsible for fracture creation whereas labels 1 and 2 are representative of seismicity which is more associated with fault reactivation. The population analysis also makes it clear that events associated with fault reactivation are greater in number than actual fracture-generating events in a field scale hydraulic fracture experiment. This observation can be explained as the experiment occurs in a shale hosted sedimentary formation. Being shaley in nature, there is an extensive fracture network that is critically stressed and amenable to activation with little to no hydraulic stimulation.

After the geomechanical characterization of the moment tensors, we now consider the spatial distribution of the different moment tensor types. Figure 53 shows the fracture-wise distribution of moment tensor types. Every fracture label has one clear majority of moment tensor type. Hence, it is clear from figure that at least in the present experiment, there is one dominant moment tensor type in each fracture cluster. -1 indicates the non-clustered event locations.

Fracture label 3 shows the most diverse population amongst the lot- it shows the presence of all three moment tensor types where all the other fracture clusters show at most two cluster types.

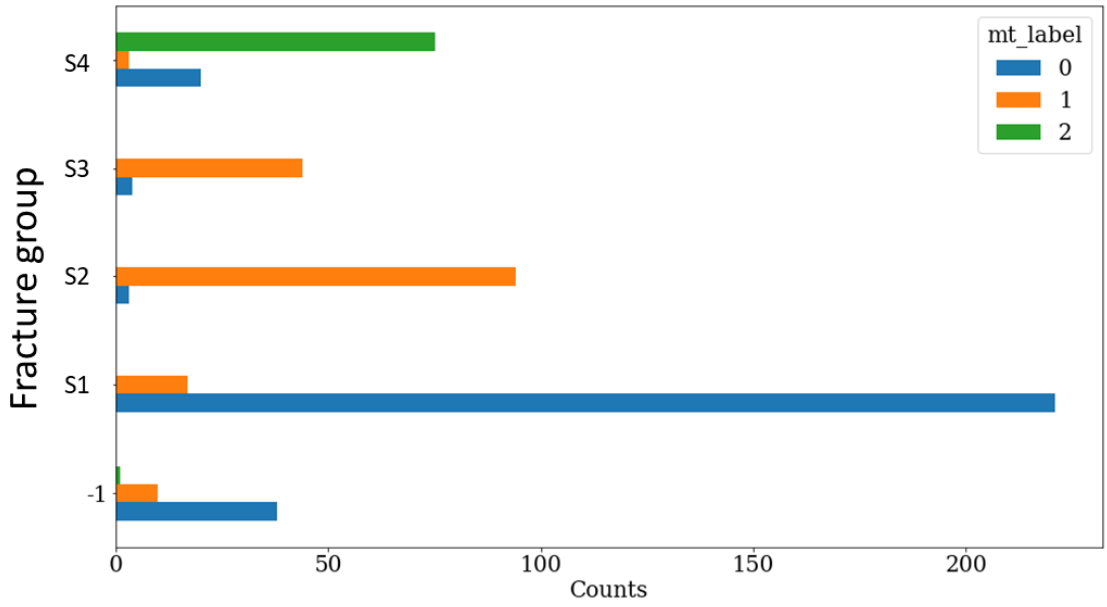

Figure 53: Distribution of moment tensor clusters in different fracture zones in the Toc2Me experiment. MT type 1 is the most widely spread whereas MT type 2 is the least widely spread moment tensor cluster. Overall, one fracture has one dominant moment tensor type implying that one fracture plane usually has a dominant deformation style that generates seismicity during a hydraulic fracturing operation.

Moment tensor type 0 – which is the most volumetric type in the population shows dominant presence in fracture set S1 (blue colored points in Figure z). Moment tensor type 1 is the geographically most widely spread cluster, with its presence in all the locational clusters and dominant presence in 2 out of 4 locational clusters. Note that moment tensor type 1 also has the largest population in the dataset. Moment tensor type 2 is geographically the least widely spread cluster in the dataset with it being present in one locational cluster (Fracture set) `– in which it is also the dominant population.

Figure 54 is the corollary of Figure 53, in that it shows the fracture set composition of moment tensor types i.e., which fracture planes carry any given moment tensor. Fracture set S4 (frac_label 3) is the most homogenous fracture plane in that it contains almost only moment tensor type 2 events - the events showing maximum volumetric component. Moment tensor type 0 and 1, that are more representative of double couple type events (more typical of tectonic type) events have a more heterogeneous distribution.

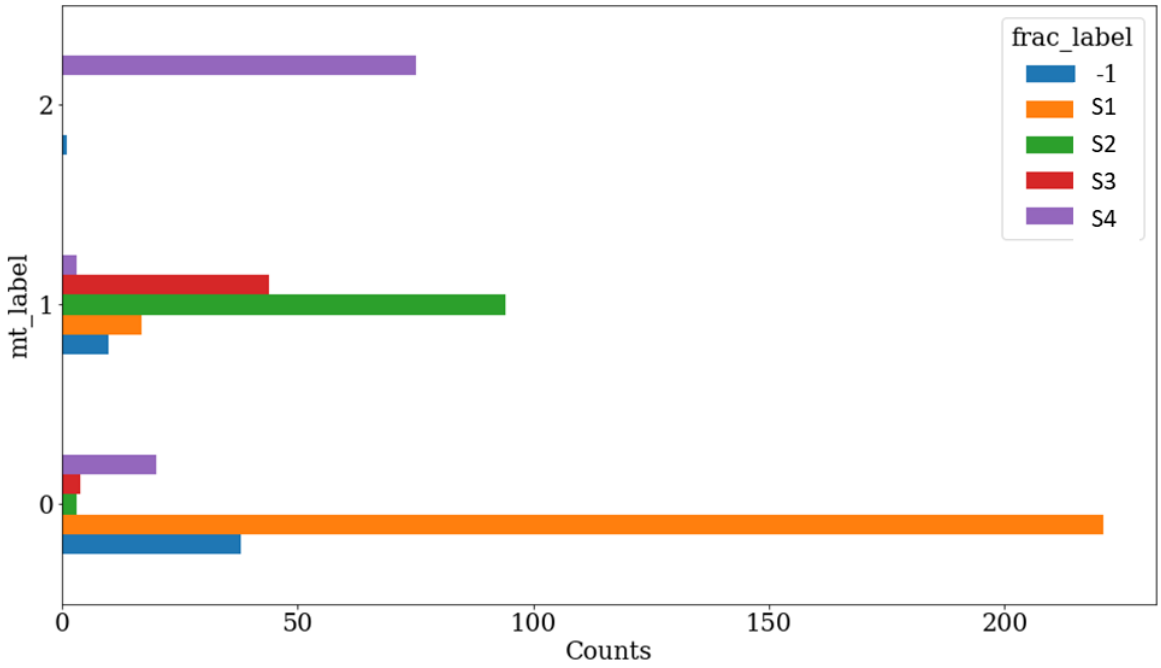

Figure 54: Distribution of fracture clusters for different moment tensor classes. Colors represent different fracture clusters.

Next in analysis it the temporal evolution of moment tensor classes. Figure 55 shows the time evolution of moment tensor types during the hydraulic fracture experiment that lasted for two months from early October 2016 to late November 2016. Moment tensor types 0 and 1 dominate the early part of the injection (early October to early November). Another important feature is

that there is hardly any overlap between the moment tensor classes. This implies that only one type of mechanical deformation is active at any given time during hydraulic fracturing.

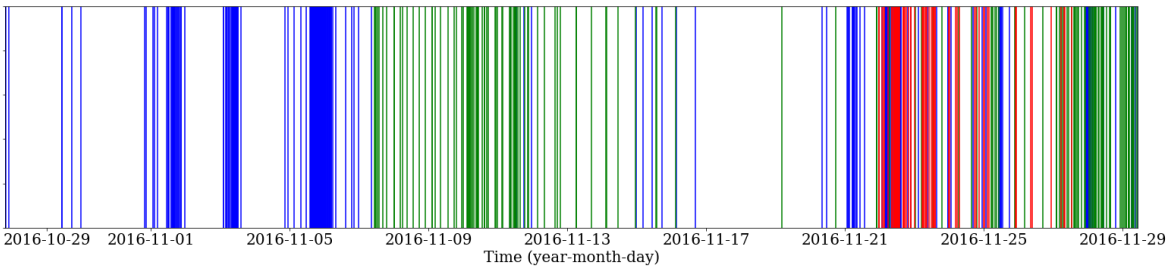

Figure 55: Temporal evolution of moment tensor types for the Toc2me experiment that lasted from early October 2016 to late November 2016. The classes are well separated with little to no overlap in the early part of the experiment. During the later stages (November) there is significantly more overlap between the moment tensor classes. Double couple type events are active at early times and isotropic type (volumetric expansion) type events are more active exclusively at the later stages of the experiment.

The start of the experiment is dominated by moment tensor type 0 that extends from early October to 6 November 30 (Figure 55). Over this interval the distribution is uneven, with the events clustered in in tight groups and not actively outside those clusters. The second regime of experiment begins around 6 November and extends till 15 November. Only the moment tensor type 1 is active during this time. We note that both the clusters are double couple type events that are associated with tectonic type earthquakes. The second half of the experiment is dominated by the moment tensor type 2 events which represent volumetric expansion of the subsurface voids. There is also significantly more overlap between the classes in the second half. The geophysical interpretation of this pattern is that fault reactivation is the dominant form of seismicity during the initial part of the stimulation. These events are indicative of reactivation of the extensive

fracture network in the subsurface whereas fracture dilation and fracture propagation are more active at later stages.

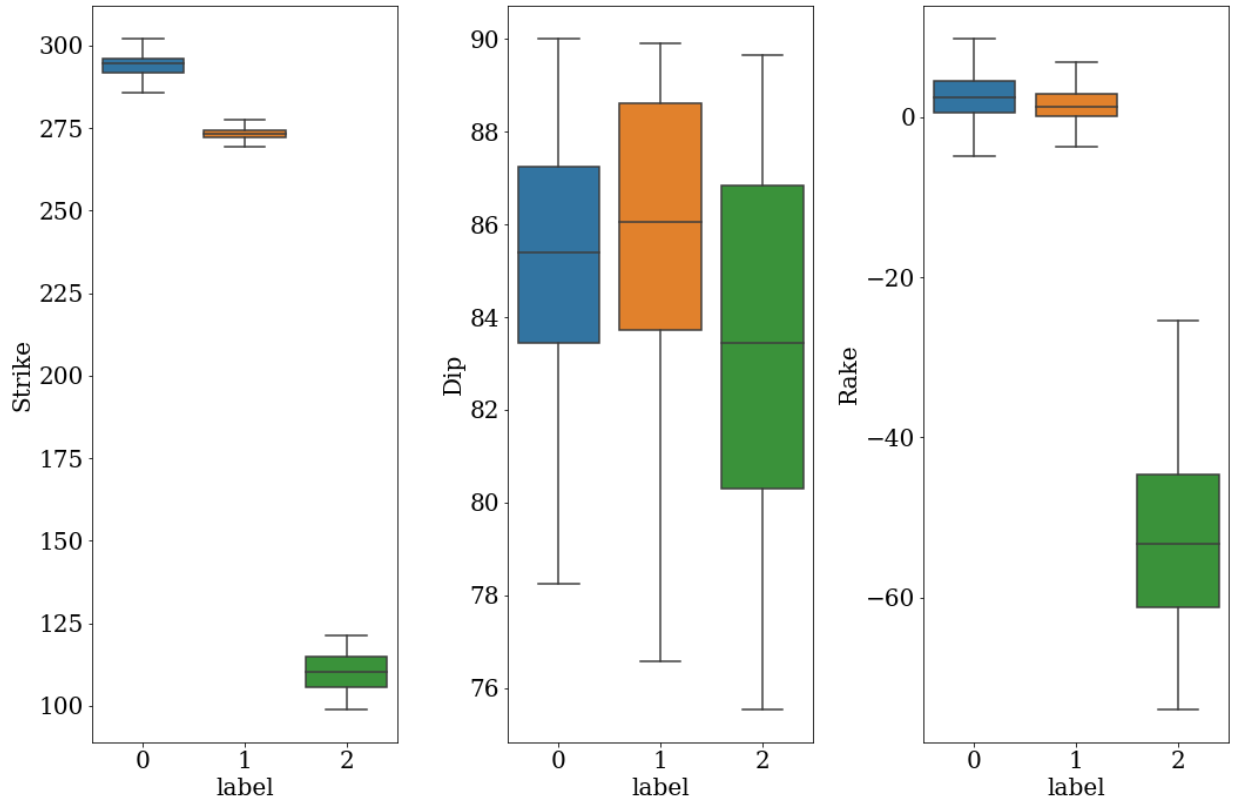

Figure 56: Fault dip, strike and rake determined for the different moment tensor classes.

The fault dip and strike refer to the orientation and inclination of a fault plane, while rake refers to the angle between the fault plane and the direction of motion on the fault (Figure 56). The fault dip is the angle at which the fault plane slopes downward from the horizontal, while the fault strike is the direction of the line formed by the intersection of the fault plane with the horizontal plane (Aki and Richards, 1980). The rake is the angle between the fault plane and the direction of motion on the fault, and it can be positive (meaning the fault motion is towards the dip direction) or negative (meaning the fault motion is away from the dip direction). Strike of fault plane is significantly different for three moment tensor classes. The dip angle for all three

moment tensor classes does not show as drastic differences as other angles. The rake angle of moment tensor cluster type 2 is significantly different from the remaining two classes.

The final part of analysis is for addressing the question: How to find hydraulically fractured regions that are geomechanically similar and those that are geomechanically dissimilar by analyzing the microseismic data captured by geophone array? Do the different microseismic events in a fracture plane exhibit relatively similar 3D motion?

To that end, different sets of signal features were used as inputs for generating embeddings in UMAP space, and then the labels derived from moment tensor classes were imposed on the embeddings. The match of the moment tensor classes with the intrinsic clustering of the embeddings in UMAP space was quantified using the Calinski-Harabasz Index (see chapter 3 for detailed theoretical background of Calinski-Harabasz Index). Five different sets of signal features were tested. The features which yield the highest CH index are deemed the most indicative of the differences observed in their corresponding moment tensors. For every one of 530 earthquakes, the measured signal from each one of 69 geophones the following features were considered:

1. Vertical component short-time Fourier transform (termed as STFT)
2. Three-component Fourier transform spectra (termed as 3C FFT)
3. P-phase polarization features (azimuth, incidence, rectilinearity and planarity)
4. S-phase polarization features (azimuth, incidence, rectilinearity and planarity)
5. P + S -phase polarization features (azimuth, incidence, rectilinearity and planarity)
6. Combined polarization and Fourier spectra

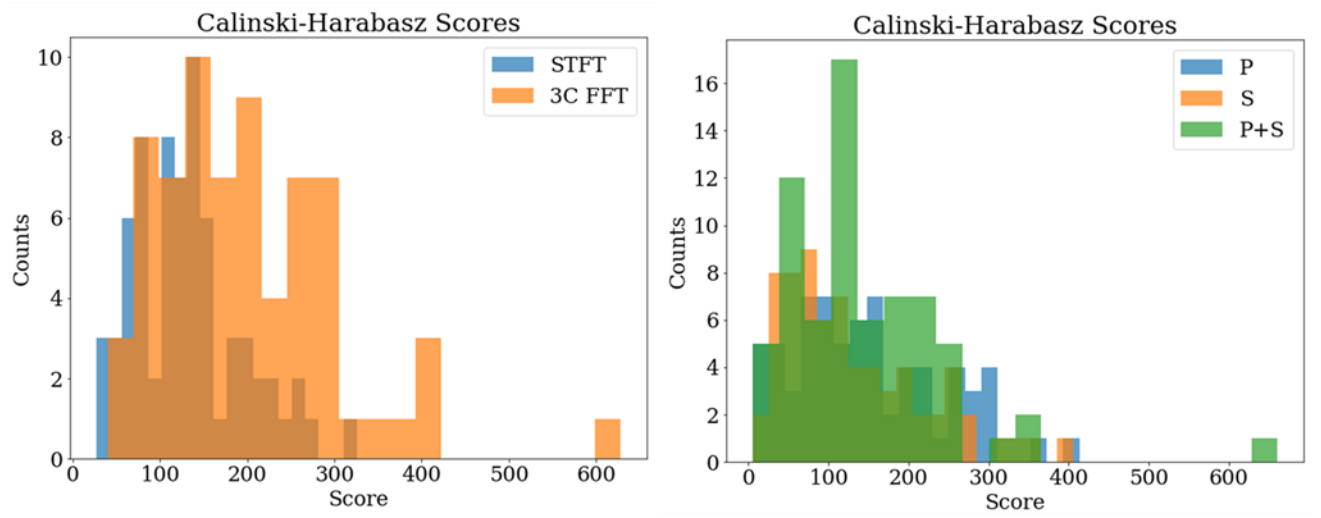

Figure 57: Calinski-Harabasz scores for time frequency features (left) and polarization features (right). In case of time frequency features, three component Fourier spectra features show higher CH scores and hence indicate better clustering match with moment tensor clusters, compared to vertical component STFT features. In case of polarization, combined P+ S polarization features showed higher statistically averaged scores over individual phase polarization features. Results comparing 3C Fourier spectra and combined polarization and 3C Fourier spectra are shown in the Appendix.

Overall, time-frequency features yielded higher scores compared to polarization features as shown in Figure 57. For time-frequency features, three-component Fourier spectra features show higher CH scores and hence indicate better clustering match with moment tensor clusters, compared to vertical component STFT features. In case of polarization, combined primary and shear phase polarization features showed higher statistically averaged scores over individual phase polarization features. Therefore, the variations in three-component signal features are the most indicative of the variation on the moment tensor types. Table 8 shows the summary of the Calinski-Harabasz scores obtained from the list of wave motion features from 70 sensors.

Table 8 Calinski-Harabasz scores obtained from wave motion features from 70 sensors

S. No.	Feature name	Median Calinski-Harabasz score
1.	Vertical STFT	126.02
2.	3C FFT	187.16
3.	P-phase polarization	133.69
4.	S-phase polarization	106.29
5.	P + S -phase polarization	128.39
6.	Combined polarization and 3CFFT	143.18

Based on the observations, we provide the answers to the question asked at the outset:

How to find hydraulically fractured regions that are geomechanically similar and those that are geomechanically dissimilar by analyzing the microseismic data captured by geophone array?

Answer: Geomechanical differences can be reliably quantified based on moment tensors acquired from high quality phase arrivals of microseismic signals acquired from geophone array.

Do the different microseismic events in a fracture plane exhibit relatively similar 3D motion?

Answer: Yes, different classes of microseismic events are shown to exhibit their signature wave motion. This notion is supported from observations from data of 70 geophones surrounding the stimulated reservoir volume, as quantified by high Calinski-Harabasz scores (median score=187) once the wave motion features are embedded in UMAP space and overlain by cluster labels. The final step of the analysis investigates the spatial distribution of the Calinski-Harabasz scores determined for different sensors.

This is done to check if there is any systematic variation in the azimuth or the distance and the goodness of fit between the moment tensor types and signal features. Figure z shows the spatial distribution of CH scores for all the geophones in the Toc2Me experiment. For single vertical component short time Fourier transform features, it appears that the scores tend to increase from the northeast to the southwest (Figure 58). The highest score (~320) is shown by the geophone located at the south-western extremity of the sensor distribution. For three component Fourier spectra, the trend is different and more subtle. The CH score tends to increase from south-west to the northeast. This orientation is different from that observed from single component features. The highest score is shown by sensor at the north-western extremity of the geophone distribution. For the polarization features, no clear directional pattern is observed. Overall, it is concluded that distance of the sensors from the region of seismicity does not have a direct impact on the control on the moment tensor types.

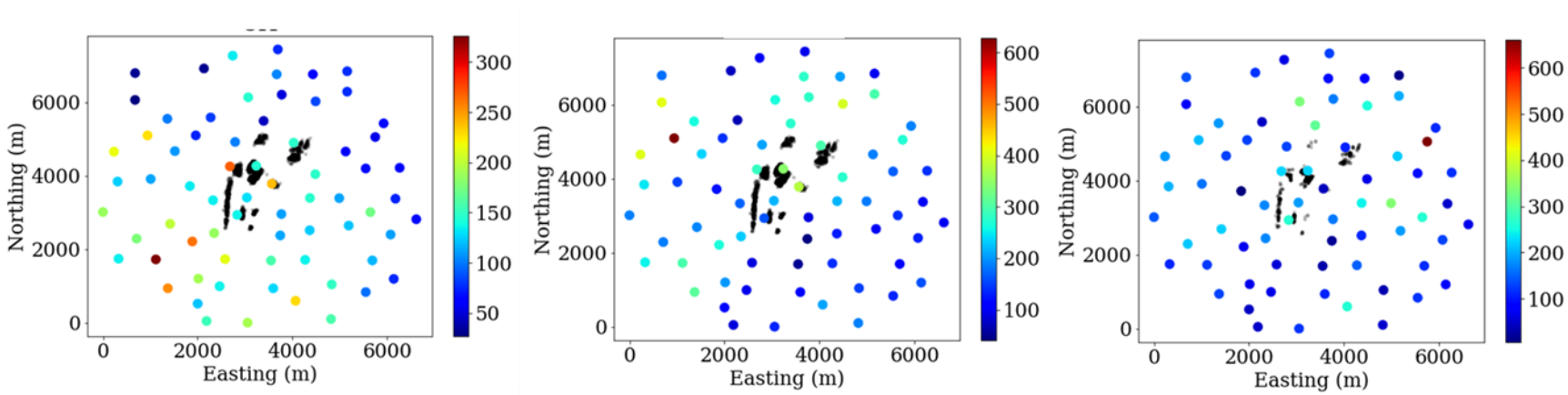

Figure 58: spatial distribution of clustering goodness scores (CH scores). For single vertical component short time Fourier transform features, it appears that the scores tend to increase from the northeast to the southwest. For three component Fourier spectra, the trend is different and more subtle. The CH score tends to increase from south-west to the northeast.

6.6.1 Conclusions

1. There is no spatiotemporal pattern observed in the 11-month long microseismicity generated in a field (km) scale hydraulic fracturing experiment. The microseismic events form distinct clusters over space. Using UMAP for dimension reduction, the moment tensors associated with the microseismic events show three clusters using the density-based clustering algorithm. The clusters obtained from combination of UMAP and density-based clustering show statistical differences in terms of the moment tensor decompositions described in terms of DC, CLVD and ISO, as well as the fracture orientations described in terms of dip, strike and rake of the faults described by the moment tensors.
2. Each cluster of hydraulic fractures is associated with a predominant moment tensor type. This is evident from the fact that each fracture label is linked to one specific majority moment tensor type. The label -1 in the results represents event locations that are not clustered. Among the clusters, label 3 stands out as having the most varied population, as it includes all three moment tensor types, while the other clusters exhibit a maximum of two types.
3. The moment tensor type 0, which is the most volumetric type in the population, is primarily found in the fracture set S1 (as indicated by the blue points in Figure z). The moment tensor type 1 is the most geographically widespread among the clusters, with its presence in all locational clusters and a dominant presence in 2 out of 4 locational clusters. Additionally, it has the largest population in the dataset. On the other hand, moment tensor type 2 is the least geographically widespread among the clusters, being present only in one locational cluster (Fracture set) and being the dominant population in that cluster.
4. Fracture set S4 (frac_label 3) is the most consistent in terms of its moment tensor type, with almost all its events being of moment tensor type 2, which is characterized by a maximum

volumetric component. On the other hand, moment tensor type 0 and 1, which are more closely associated with double couple type events (more typical of tectonic-related events) have a more varied distribution.

5. Based on the results overall, the time-frequency features performed better than the polarization features in terms of clustering match with the moment tensor clusters. Specifically, among the time-frequency features, the three-component Fourier spectra features yielded higher CH scores and thus demonstrated a better match with the moment tensor clusters, compared to the vertical component STFT features.

6. In the case of polarization, the combination of primary and shear phase polarization features demonstrated higher statistically averaged scores compared to the individual phase polarization features. This suggests that variations in the three-component signal features are the most indicative of variations in the moment tensor types.

REFERENCES

A. Damani, A. Sharma, C. H. Sondergeld., and C. S Rai, "Mapping of Hydraulic Fractures under Triaxial Stress Conditions in Laboratory Experiments using Acoustic Emissions", presented at the *SPE Annual Technical Conference and Exhibition*, San Antonio, TX, USA, October 8-10, 2012.

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

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (2010),

Boese, C. M., Kwiatek, G., Fischer, T., Plenkers, K., Starke, J., Blümle, F., Janssen, C., & Dresen, G. (2022). Seismic monitoring of the STIMTEC hydraulic stimulation experiment in anisotropic metamorphic gneiss. *Solid Earth, 13*(2), 323–346. https://doi.org/10.5194/SE-13-323-2022

Bolton, D. C., Marone, C., Shokouhi, P., Rivière, J., Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2019). Characterizing acoustic signals and searching for precursors during the laboratory seismic cycle using unsupervised machine learning. In *Seismological Research Letters* (Vol. 90, Issue 3, pp. 1088–1098). Seismological Society of America. https://doi.org/10.1785/0220180367

Boadu, F. K., & Long, L. T. (1996). Effects of fractures on seismic-wave velocity and attenuation. Geophysical Journal International, 127(1), 86–110. https://doi.org/10.1111/J.1365-246X.1996.TB01537

Bean, C., I. Lokmer, and G. O'Brien (2008), Influence of near-surface volcanic structure on long-period seismic signals and on moment tensor inversions: Simulated examples from Mount Etna, J. Geophys. Res., 113, B08308, doi:10.1029/2007JB005468.

Beghini, M., & Bertini, L. (1990). Fatigue crack propagation through residual stress fields with closure phenomena. Engineering Fracture Mechanics, 36(3), 379–387. https://doi.org/10.1016/0013-7944(90)90285-O

C. Macbeth, "Multicomponent VSP analysis for applied seismic anisotropy". Pergamon, 2002

C.J. de Pater, J. Groenenboom, D.B. van Dam, and R. Romijn, "Active seismic monitoring of hydraulic fractures in laboratory experiments", *International Journal of Rock Mechanics and Mining Sciences*, vol. 38, no. 6, pp 777-785, Aug. 2001.

Chakravarty, A., & Misra, S. (2022). Unsupervised learning from three-component accelerometer data to monitor the spatiotemporal evolution of meso-scale hydraulic fractures. *International Journal of Rock Mechanics and Mining Sciences*, *151*, 105046. https://doi.org/10.1016/J.IJRMMS.2022.105046

Chakravarty, A., Misra, S., & Rai, C. S. (2021). Visualization of hydraulic fracture using physics-informed clustering to process ultrasonic shear waves. *International Journal of Rock Mechanics and Mining Sciences*, *137*, 104568. https://doi.org/10.1016/j.ijrmms.2020.104568

Chakravarty, A., & Misra, S. (2021). Hydraulic fracture mapping using wavelet-based fusion of wave transmission and emission measurements. *Journal of Natural Gas Science and Engineering*, *96*, 104274. https://doi.org/10.1016/J.JNGSE.2021.104274

Chakravarty, A., Misra, S., Stenftenagel, J, Wu, K., Duan, B., (2022), Hydraulic Fracturing-driven Infrasound Signals – A New Class of Signal for Subsurface Engineering (in review).

D. Burns, M. Willis, M. Toksöz, and L. Vetri. "Fracture Properties from Seismic Scattering", *The Leading Edge*, vol. 26, no. 9, pp 1186-1196, Sep. 2007.

Dunham, M. W., Malcolm, A., & Kim Welford, J. (2020). Improved well-log classification using semisupervised label propagation and self-training, with comparisons to popular supervised algorithms. *Geophysics*, *85*(1), O1–O15.

E. Majer, J Peterson, T Daley, L Myer, J Queen, P D'Onfro, and W Rizer, "Fracture detection using crosswell and single well surveys". *Geophysics*, vol. 62 no. 2, pp 495–504, Feb. 1997.

Fred Kofi Boadu, Leland Timothy Long, Effects of fractures on seismic-wave velocity and attenuation, *Geophysical Journal International*, Volume 127, Issue 1, October 1996, Pages 86–110, https://doi.org/10.1111/j.1365-246X.1996.tb01537.x

He, Z., & Duan, B. (2018). Dynamic Study on the Fracture Interaction and the Predominant Frequency of the Induced Microseismic Signals During Hydraulic Fracturing: DOI: 10.14800/IOGR.428. *Improved Oil and Gas Recovery*, *2*. https://doi.org/10.14800/IOGR.428

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. *Science Advances*, *4*(5). https://doi.org/10.1126/sciadv.aao2929

J. J. S. de Figueiredo, J. Schleicher, R. R. Stewart, N. Dayur, B. Omoboya, R. Wiley, A. William, "Shear wave anisotropy from aligned inclusions: ultrasonic frequency dependence of velocity and attenuation", *Geophysical Journal International*, vol. 193, no. 1, pp 475–488, Feb. 2013.

Jia, S., Jia, S., Jia, S., Jia, S., Deng, X., Deng, X., Deng, X., Deng, X., Xu, M., Xu, M., Xu, M., Xu, M., Zhou, J., & Jia, X. (2020). Superpixel-level weighted label propagation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(7), 5077–5091. https://doi.org/10.1109/TGRS.2020.2972294

Jin, Y., & Misra, S. (2022). Controlling mixed-mode fatigue crack growth using deep reinforcement learning. Applied Soft Computing, 127, 109382. https://doi.org/10.1016/J.ASOC.2022.109382

Kendall, M.G., "Rank Correlation Methods" (4th Edition), Charles Griffin & Co., 1970

Kwiatek, G., Martínez-Garzón, P., Plenkers, K., Leonhardt, M., Zang, A., von Specht, S., Dresen, G., & Bohnhoff, M. (2018). Insights Into Complex Subdecimeter Fracturing Processes Occurring During a Water Injection Experiment at Depth in Äspö Hard Rock Laboratory, Sweden. *Journal of Geophysical Research: Solid Earth*, *123*(8), 6616–6635. https://doi.org/10.1029/2017JB014715

Kwiatek, G., Martínez-Garzón, P., Plenkers, K., Leonhardt, M., Zang, A., von Specht, S., et al. (2018). Insights into complex subdecimeter fracturing processes occurring during a water injection experiment at depth in Äspö hard rock laboratory, Sweden. Journal of Geophysical Research: Solid Earth, 123, 6616–6635. https://doi.org/10.1029/2017JB014715

Leggett, S. E., Zhu, D., & Hill, A. D. (2022). Thermal Effects on Far-Field Distributed Acoustic Strain-Rate Sensors. *SPE Journal*, *27*(02), 1036–1048. https://doi.org/10.2118/205178-PA

Leggett, S., Reid, T., Zhu, D., & Hill, A. D. (2022). Experimental Investigation of Low-Frequency Distributed Acoustic Strain-Rate Responses to Propagating Fractures. *Society of Petroleum Engineers - SPE Hydraulic Fracturing Technology Conference and Exhibition, HFTC 2022*. https://doi.org/10.2118/209135-MS

Liu, Y., Jin, G., Wu, K., & Moridis, G. (2021). Hydraulic-Fracture-Width Inversion Using Low-Frequency Distributed-Acoustic-Sensing Strain Data—Part I: Algorithm and Sensitivity Analysis. *SPE Journal*, *26*(01), 359–371. https://doi.org/10.2118/204225-PA

Liu, Y., Wu, K., Jin, G., Moridis, G., Kerr, E., Scofield, R., & Johnson, A. (2021). Fracture-Hit Detection Using LF-DAS Signals Measured during Multifracture Propagation in Unconventional Reservoirs. *SPE Reservoir Evaluation & Engineering*, *24*(03), 523–535. https://doi.org/10.2118/204457-PA

Liu, R., & Misra, S. (2022). A generalized machine learning workflow to visualize mechanical discontinuity. Journal of Petroleum Science and Engineering, 210, 109963. https://doi.org/10.1016/J.PETROL.2021.109963

Li, H., Misra, S., & Liu, R. (2021). Characterization of mechanical discontinuities based on data-driven classification of compressional-wave travel times. International Journal of Rock Mechanics and Mining Sciences, 143, 104793. https://doi.org/10.1016/J.IJRMMS.2021.104793

Liu, R., & Misra, S. (2022). Monitoring the propagation of mechanical discontinuity using data-driven causal discovery and supervised learning. Mechanical Systems and Signal Processing, 170, 108791. https://doi.org/10.1016/J.YMSSP.2021.108791

LJ Pyrak-Nolte, LR Myer, and NGW Cook, "Transmission of seismic waves across single natural fractures", *J. Geophys. Res*., vol 95, no. B6, pp 8617– 8638, Jun. 1990.

M. L. Jost, R. B. Herrmann; A Student's Guide to and Review of Moment Tensors. Seismological Research Letters 1989;; 60 (2): 37–57. doi: https://doi.org/10.1785/gssrl.60.2.37

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*(2), 800–809. https://doi.org/10.1002/2016GL071963

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/arxiv.1802.03426

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. https://doi.org/10.1109/LGRS.2019.2909218

Narayan, A., Berger, B., & Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology 2021 39:6*, *39*(6), 765–774. https://doi.org/10.1038/s41587-020-00801-7

Ni, K., Bresson, X., Chan, T., Esedoglu, S., Ni, K., Bresson, · X, Chan, · T, Bresson, X., Chan, T., & Esedoglu, S. (2009). Local Histogram Based Segmentation Using the Wasserstein Distance. *International Journal of Computer Vision 2009 84:1*, *84*(1), 97–111. https://doi.org/10.1007/S11263-009-0234-0

Niemz, P., Dahm, T., Milkereit, C., Cesca, S., Petersen, G., & Zang, A. (2021). Insights Into Hydraulic Fracture Growth Gained From a Joint Analysis of Seismometer-Derived Tilt Signals and Acoustic Emissions. *Journal of Geophysical Research: Solid Earth*, *126*(12), e2021JB023057. https://doi.org/10.1029/2021JB023057

NR. Warpinski, SL Wolhart, and CA Wright, "Analysis and Prediction of Microseismicity Induced by Hydraulic Fracturing", presented at the *SPE Annual Technical Conference and Exhibition*, New Orleans, LA, USA, September 30–October 3, 2001.

ObsPy: A Python Toolbox for Seismology, Seismological Research Letters, 81 (3), 530-533.

P. Bhoumick, C. Sondergeld, and C. S. Rai, "Mapping Hydraulic Fracture in Pyrophyllite Using Shear Wave", presented at the *U.S. Rock Mechanics/Geomechanics Symposium*, Seattle, WA, USA, June 17-20, 2018.

Ross, Z. E., Trugman, D. T., Azizzadenesheli, K., & Anandkumar, A. (2020). Directivity Modes of Earthquake Populations with Unsupervised Learning. *Journal of Geophysical Research: Solid Earth*, *125*(2), e2019JB018299. https://doi.org/10.1029/2019JB018299

Rousseeuw, P.J.. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", J. Comput. Appl. Math, vol. 20, pp 53–65, Nov. 1987.

Schoenball, M., Ajo-Franklin, J. B., Blankenship, D., Chai, C., Chakravarty, A., Dobson, P., Hopp, C., Kneafsey, T., Knox, H. A., Maceira, M., Robertson, M. C., Sprinkle, P., Strickland, C., Templeton, D., Schwering, P. C., Ulrich, C., Wood, T., Ajo-Franklin, J., Baumgartner, T., … Zoback, M. D. (2020). Creation of a Mixed-Mode Fracture Network at Mesoscale Through Hydraulic Fracturing and Shear Stimulation. *Journal of Geophysical Research: Solid Earth*, *125*(12), e2020JB019807. https://doi.org/10.1029/2020JB019807

Shi, P., Seydoux, L., & Poli, P. (2021). Unsupervised Learning of Seismic Wavefield Features: Clustering Continuous Array Seismic Data During the 2009 L'Aquila Earthquake. *Journal of Geophysical Research: Solid Earth*, *126*(1), e2020JB020506. https://doi.org/10.1029/2020JB020506

Szafranski, D., & Duan, B. (2022). A Workflow to Integrate Numerical Simulation, Machine Learning Regression and Bayesian Inversion for Induced Seismicity Study: Principles and a Case Study. *Pure and Applied Geophysics*, *179*(10), 3543–3568. https://doi.org/10.1007/S00024-022-03140-7/FIGURES/15

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/S10994-019-05855-6/FIGURES/5

Vavryčuk, V. (2011). Tensile earthquakes: Theory, modeling, and inversion. Journal of Geophysical Research: Solid Earth, 116(B12), 12320. https://doi.org/10.1029/2011JB008770

Watson, L. M. (2020). Using unsupervised machine learning to identify changes in eruptive behavior at Mount Etna, Italy. *Journal of Volcanology and Geothermal Research*, *405*, 107042. https://doi.org/10.1016/j.jvolgeores.2020.107042

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 20

Z. Li, Y. Kang, W. Lv, W. X. Zheng and X. -M. Wang, "Interpretable Semisupervised Classification Method Under Multiple Smoothness Assumptions With Application to Lithology Identification," in IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 3, pp. 386-390, March 2021, doi: 10.1109/LGRS.2020.2978053.

Zhang, H., Eaton, D. W., Rodriguez, G., & Jia, S. Q. (2019). Source-Mechanism Analysis and Stress Inversion for Hydraulic-Fracturing-Induced Event Sequences near Fox Creek, AlbertaSource-Mechanism Analysis and Stress Inversion. *Bulletin of the Seismological Society of America*, *109*(2), 636–651. https://doi.org/10.1785/0120180275

APPENDIX

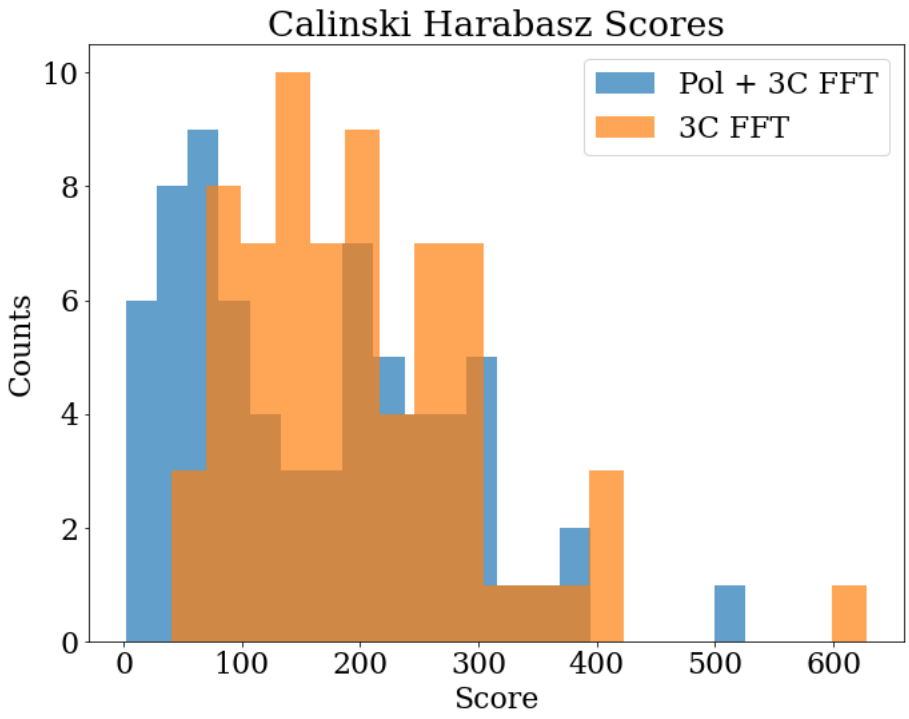

Figure A1: Calinski-Harabasz scores for the comparison of combined polarization and three component Fourier spectra against three component Fourier spectra for 70 sensors in the Toc2Me microseismic dataset. The median scores of different feature sets are summarized in Table 8.

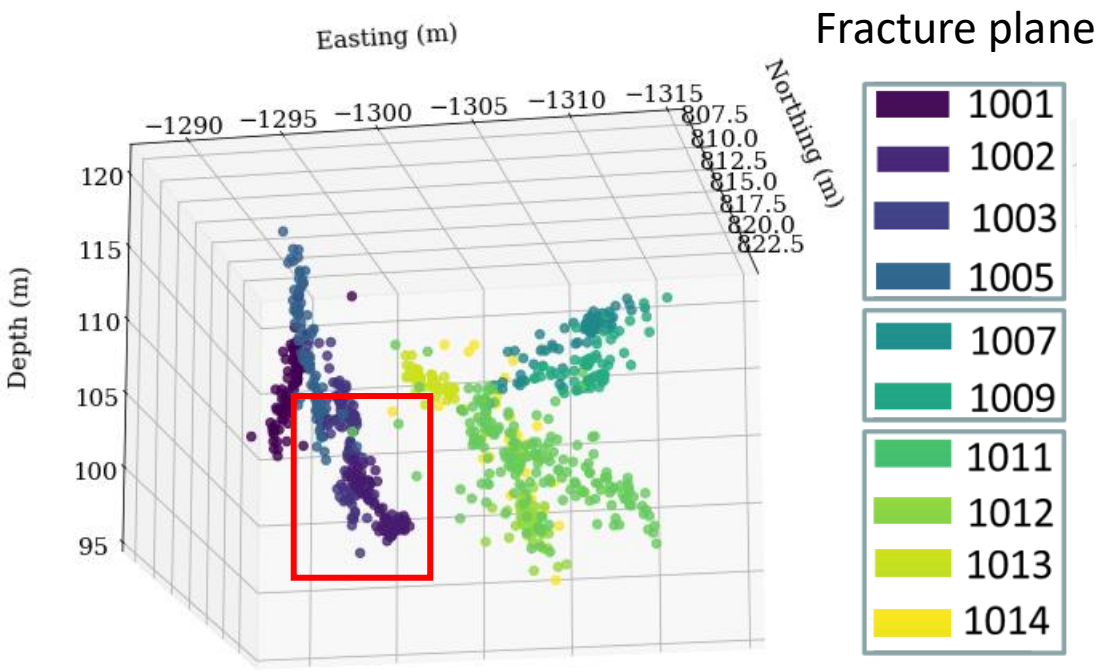

Figure A2: Highlighted location of fracture plane 1002 (red box). The most likely explanation for the low Tau statistic of fracture plane 1002 is that its location is tightly overlapping with 1003, 1001 and 1005. It is only through independent observations like core measurements that it is distinguished from the other three planes. Note that other three plans have distinct azimuths and dips which makes them easily distinguishable from geometric analysis alone, unlike fracture plane 1002.