Accepted in EPSL May 2020

Predicting the proximity to macroscopic failure using local strain populations from dynamic in situ X-ray tomography triaxial compression experiments on rocks

J. McBeck[1], J. M. Aiken[2,3], Y. Ben-Zion[4], F. Renard[1,5]

[1] Physics of Geological Processes, The Njord Centre, Department of Geosciences, University of Oslo, Oslo, Norway.
[2] Center for Computing in Science Education, Department of Physics, University of Oslo, Oslo, Norway.
[3] Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan, USA.
[4] Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA.
[5] University Grenoble Alpes, University Savoie Mont Blanc, CNRS, IRD, IFSTTAR, ISTerre, 38000 Grenoble, France.

Highlights

- Machine learning successfully predicts the proximity to failure from strain fields.
- Dilatational strain predicts the proximity to failure better than shear strain.
- The 25^{th}-75^{th} percentile values of the strain population best predict failure proximity.

Abstract

Predicting the proximity of large-scale dynamic failure is a critical concern in the engineering and geophysical sciences. Here we use evolving contractive, dilatational, and shear strain deformation preceding failure in dynamic X-ray tomography experiments to examine which strain components best predict the proximity to failure. We develop machine learning models to predict the proximity to failure using time series of three-dimensional local incremental strain tensor fields acquired in rock deformation experiments under stress conditions of the upper crust. Three-dimensional scans acquired *in situ* throughout triaxial compression experiments provide a distribution of density contrasts from which we estimate the three-dimensional incremental strain that accumulates between each scan acquisition. Training machine learning models on multiple experiments of six rock types provides suites of feature importance that indicate the predictive power of each feature. Comparing the average importance of groups of features that include information about each strain component quantifies the ability of the contractive, dilatational and shear strain to predict the proximity of macroscopic failure. A total of 24 models of four machine learning algorithms with six rock types indicate that 1) the dilatational strain provides the best predictive power of the strain components, and 2) the intermediate values (25^{th}-75^{th} percentile) of the strain population provide the best predictive power of the statistics of the strain populations. In addition, the success of the predictions of models trained on one rock type and tested on other rock types quantifies the similarities and differences of the precursory strain accumulation process in the six rock types. These similarities suggest the potential existence of a unified theory of brittle rock deformation for a range of rock types.

Keywords

strain localization; tomography; machine learning; triaxial compression

1. Introduction

Predicting the timing of dynamic catastrophic failures is a fundamental problem for society and earthquake science. This question could be more straightforward to address if large faults were planar and surrounded by a purely elastic and homogeneous solid where strain energy only accumulates in the time between the episodic failures of the large faults. However, seismic and geodetic observations document that crustal volumes experience ongoing seismic and aseismic failure on a heterogeneous population of faults with various sizes, geometries and failure mechanisms (e.g., *Scholz*, 1990; *Ben-Zion*, 2008; *Hauksson et al*., 2012).

Under triaxial compression conditions, permanent deformation preceding macroscopic failure manifests by the decreasing slope of the axial strain-stress curve following a nominally linear phase, along with a corresponding reduction of the effective moduli toward macroscopic failure, and increasing transverse or radial deformation (e.g., *Brace et al.,* 1966; *Scholz et al.*, 1973; *Lyakhovsky et al*., 1997; *Hamiel et al*., 2004). Distributed dilation within the rock volume preceding dynamic rupture helps explain the emergence of these and other precursory signals.

The dilatancy-diffusion model posits that as fractures propagate and open in the presence of fluids in the crust, the increased pore space allows fluid to diffuse, producing larger ratios of fluid to solid volume (e.g., *Nur*, 1974). The increased ratio of fluid (air or liquid) to solid may reduce the P-wave velocity and decrease the effective elastic modulus, and dilatancy-enhanced fluid flow may change local hydrologic levels. Some field observations are consistent with these expectations (e.g., *Frank*, 1965; *Whitcomb et al*., 1973; *Aggarwal et al*., 1973; *Roeloffs*, 1988), but the general validity of these observations remain controversial (e.g., *McEvilly & Johnson*, 1974; *Bolt*, 1977; *Haase et al*., 1995). Observations of macroscopic volumetric strain during triaxial compression experiments indicate that rocks

tend to dilate before macroscopic failure under confining stresses representative of the upper crust (e.g., *Brace et al.*, 1966). Quantifying the behavior of rocks during the approach to macroscopic failure is a critical step toward constraining the precursory signals that reveal the proximity of the next large earthquake.

A key assumption of the earthquake preparation process is that both the dilatational and shear deformation evolve toward failure, and that both modes of deformation contain information critical for predicting the timing of rupture. For example, the second invariant of the strain rate tensor, which includes the volumetric and deviatoric strain tensor components, is often used as a proxy for long term seismic hazard (e.g., *Kreemer et al*., 2003). Earthquakes can rupture in a combination of modes, and so the reported accelerated seismic release rates before mainshock events (e.g., *Mogi*, 1969; *Papazachos*, 1973; *Sykes & Jaumé*, 1990) may involve different modes of deformation. However, increased rates of foreshocks surrounding normal faults, relative to thrust and strike-slip faults (*Abercrombie & Mori*, 1996), suggests that extensional deformation may provide less ambiguous signals than shear deformation before rupture. Similarly, the observed acceleration of the total fracture volume in X-ray tomography experiments suggests a process consistent with a critical phase transition (*Renard et al*., 2018; *Kandula et al*., 2019), but this tracking of the fracture volume does not provide information about the evolution of shear deformation.

Time series of 3D incremental strain component fields throughout X-ray tomography experiments (e.g., **Figure** 1) suggest that both the dilatational and shear strain components accelerate toward failure in some experiments of some rock types (e.g., *Renard et al*., 2019). The dilatational strain, as measured with the mean of the local incremental strain population, tends to accelerate in more experiments than the shear strain (*Renard et al*., 2019; *McBeck et al*., 2018, 2019, 2020). The consistency of the dilatational strain evolution suggests that shear strain may not provide information crucial for predicting the timing of macroscopic failure. In

this case, effective predictive analyses could concentrate on detecting the precursory signals indicative of dilatancy.

To determine which strain components are most likely to improve the ability of predicting the proximity to system-size failure, we analyze the time series of 3D local incremental strain fields captured in X-ray tomography triaxial compression experiments (e.g., **Figure** 1). Here, we characterize the proximity to failure by the ratio between the differential stress at a given time, and the differential stress at the macroscopic failure of the sample. The incremental strain fields are calculated throughout twelve *in situ* dynamic X-ray tomography triaxial compression deformation experiments on six rock types (**Figure** 2, **Table** S1). Machine learning models then use the statistics of the local contraction, dilation and shear strain populations to predict the proximity to failure (**Table** S2). We examine the feature importance and Shapley Additive Explanation (SHAP) values to determine which strain components and statistics have the best predictive power of the proximity to macroscopic failure. To assess differences in the predictive power of these strain components and statistics among different rock types, we train six unique models on datasets from two experiments on the same rock type (**Table** S1). To quantify the similarity of the precursory strain accumulation processes in these rock types, we test the success of the models with data from the same rock type used to train the model, and with data from different rock types. High degrees of success between models trained and tested on different rock types may reflect the similarity of the strain accumulation process in these rock types. To increase the robustness of the results, we repeat the analysis using four machine learning approaches, including random forest and gradient boosting (XGBoost) methods. This suite of analyses indicates that for the six analyzed rock types, the intermediate values of the dilatational strain populations provide the best predictive power of the proximity to macroscopic failure.

2. Methods

2.1. Experimental design and data

We deform rocks in an X-ray transparent deformation apparatus installed on beamline ID19 at the European Synchrotron and Radiation Facility. In each experiment, we increase the axial stress in steps of 0.5-5 MPa under constant confining stress between 5-35 MPa until the sample fails macroscopically (**Table** S3). At each stress step, we acquire a 3D X-ray tomogram at 6.5 μm per voxel-side resolution while the sample is under constant stress conditions inside the apparatus. Each scan requires about 2 minutes, and the final scan immediately precedes macroscopic failure of the rock. *McBeck et al*. (2020) describe in detail the experimental conditions and method of calculating the strain tensor fields used in the current analysis, so we only briefly describe these topics here. We calculate the differential stress, $\sigma_D = \sigma_1 - P_c$, where σ_1 is the axial stress and P_c is the confining pressure. The differential stress at failure is noted σ_F.

From the 3D tomograms of the 12 experiments, we perform digital volume correlation (DVC) in ten intervals of each experiment (e.g., **Figure** 2), following the approach of *McBeck et al*. (2020) with the code TomoWarp2 (*Tudisco et al*., 2017). Each interval is separated by approximately equal increments of cumulative macroscopic axial strain. The DVC calculations provide the 3D displacement fields between each scan acquisition, and thus the incremental strain tensor components. The positive and negative divergence represents the dilatational and contractive volumetric strains, respectively, and the magnitude of the angular velocity of the curl represents the shear strain.

2.2. Feature extraction

We train the models using data from all the experiments with the same rock type, producing six unique models of each rock type (**Table** S1). To extract information about the 3D incremental strain field, we subdivide each strain field into a grid of subvolumes that are

fixed in space from which we report statistics of the strain population. The 27 features of the data are a combination of one of the three strain components (dilation, contraction, shear strain), and a measurement of the population of these strain components within the subvolume (**Table** S2). Each subvolume in the cubic grid has a side length of 0.5 mm, about four times the DVC spatial resolution (0.13 mm), and is fixed in space. Therefore, each subvolume includes about 4^3 measurements of each of the three strain components. The strain components include contractive, dilatational and shear strains, as measured with the negative divergence, positive divergence, and magnitude of curl of the displacement field. The statistics include the 90^{th}, 75^{th}, 50^{th}, 25^{th}, and 10^{th} percentile, mean, standard deviation and sum of the strain population, and the number of measurements within a subvolume. For example, one feature is the 90^{th} percentile of the contraction values within a subvolume of the strain field. The number of measurements within a subvolume is the total number of strain values within the subvolume, which tends to increase for dilatational strains and decrease for contractive strains with loading. This number will be the same for the shear strain values throughout loading, and equal to about 4^3. We extract the number of measurements within a subvolume because previous work shows that the volume of rock that the dilatational and contractive strains occupy can accelerate toward failure (e.g., *Renard et al*., 2019), and so may signal the approach of macroscopic failure. Thus one subvolume produces one sample for the machine learning models, which contains 27 features. Testing the influence of using a smaller sub-volume size (0.2 mm) revealed only insignificant differences in the results.

2.3. Machine learning methods

We employ two machine learning methods: random forest classification (e.g., *Breiman*, 2001) and XGBoost (i.e., extreme gradient boosting) classification (e.g., *Chen & Guestrin*, 2016). We designed this analysis as a classification problem similar to time-dependent seismic hazard models that predict the rate of event occurrence (e.g., *Tiampo & Shcherbakov*,

151 2012). We synthesize results from the random forest and XGBoost methods to increase the
152 robustness of the conclusions. Both methods utilize an ensemble of decision trees as the base
153 estimator. **Text** S1 describes differences in the random forest and XGBoost methods. For
154 each machine learning method, we perform a grid search over the hyperparameter space to
155 find the appropriate set of hyperparameters for each model (*Géron*, 2017).
156 The models predict the proximity to system size failure using the macroscopic differential
157 stress applied to the rock at the time that the strain field was calculated, and the differential
158 stress at failure, σ_F. We use the differential stress an indicator of the distance until failure
159 because previous experimental studies have demonstrated that fractures develop in rocks
160 preceding macroscopic compressive failure as a critical phase transition relative to the
161 macroscopic σ_F (e.g., *Renard et al*., 2018; *Kandula et al*., 2019).
162 We use two different classification schemes for each of the two machine learning
163 methods, producing four machine learning approaches. We predict either two classes or four
164 classes of the proximity to failure. For the four prediction classes, we predict whether the rock
165 is experiencing <25%, 25-50%, 50-75% or >75% of σ_F. For the two prediction classes, we
166 predict whether the rock is experiencing <50%, or >50% of σ_F. For example, if the rock is
167 under 10 MPa differential stress, σ_D, and fails at σ_F=100 MPa, then $\sigma_D/\sigma_F = 0.10$ or 10%, so
168 this distance to failure is in the first prediction class in both the four- (<25%) and two-class
169 (<50%) prediction models. This classification approach enables comparing the model
170 accuracy to the accuracy expected from the random probability of selecting the correct
171 answer, providing a systematic means of quantifying the prediction success. The accuracy is
172 the number of correct answers out of the total number of correct and incorrect answers (e.g.,
173 *Müller & Guido*, 2016).
174 The time series of incremental strain fields are calculated with approximately equal
175 increments of macroscopic axial strain, so predicting the proximity to failure using the

176 macroscopic differential stress may result in different numbers of outcomes (stress classes)
177 depending on the non-linearity of the stress-strain relationship. Using more than four classes
178 results in several experiments having a different number of outcomes (predicted stress
179 classes). Such differences in the expected random probability between experiments reduces
180 the likelihood of achieving high transfer accuracies, when models are trained on data from
181 one rock type and tested with data from another rock type. For this reason, we do not use
182 more than four prediction classes.

183 With the 50% partitioning of the experiments into two stress classes, all of the
184 experiments have two classes of predictions. Consequently, the random probability of
185 selecting the correct answer is 1/2. With the 25% partitioning of the experiments into four
186 stress classes, all but two of the experiments have four stress classes. The shale experiments
187 include longer yielding phases preceding macroscopic failure than the other rock types, so
188 they only have three stress classes. Consequently, whereas the shale experiments produce a
189 random probability of 1/3, all the other experiments have four prediction classes with a
190 random probability of 1/4. We counted the number of outcomes in each class in the testing
191 datasets to ensure that the expected probabilities are the same as those in the testing datasets.
192 After counting the number of outcomes in each class, we found that the mean accuracies
193 across all the classes expected from a random guess in the testing datasets were the same as
194 those in the training datasets.

195 To determine which strain components and statistics signal the approach to macroscopic
196 failure, we examine the suites of feature importance of the trained models. The feature
197 importance quantifies the average information gain produced by encountering a feature. The
198 magnitude of improvement in the model prediction associated with a node and its
199 corresponding information is defined as the node impurity. The decrease in node impurity
200 weighted by the probability of reaching the node produces the feature importance (e.g.,

Müller & Guido, 2016). Higher feature importance indicates that the feature has more influence in fitting the data, so we interpret the importance as indicating the relative predictive power of features.

In addition to the feature importance, we use the Shapley Additive Explanation (SHAP) values to examine the impact of individual features on the feature space (*Lundberg & Lee*, 2017). SHAP values indicate the contribution of each feature on the prediction of the average (i.e., a completely featureless naïve model). They may explain single predictions for single measurements, and also for the collection of predictions in a data set (i.e., stress classes). SHAP values for a specific feature *S* are calculated by comparing results of models that do and do not include feature *S*. The approximate effect on all predictions may then be estimated by finding the mean SHAP value over the samples in the training data.

3. Results

3.1. Transfer and non-transfer success

First, we examine the success of the models when trained and tested on the same rock type, and with distinct datasets. In particular, we trained six unique models using 80% of the data from experiments on the same rock type: sandstone, basalt, monzonite, granite, shale, and limestone (**Figure** 3). We then tested the accuracy of the models with the remaining 20% of the data from each group of experiments (i.e., non-transfer accuracy), and with data from the other groups of experiments (i.e., transfer accuracy) (**Figures** 4, S2, S3, S4). We divided the data into training and testing datasets by random selection. Due to potential auto-correlation between the data in the training and testing datasets, the non-transfer accuracies may be higher than the transfer accuracies.

The four machine learning approaches include the two schemes of prediction with two and four classes of prediction, and two algorithms of the random forest and XGBoost

implementations. The four machine learning approaches and six rock types produce 24 models. The 12 models that predict two classes have a higher range of accuracy (0.76-0.89) than the 12 models that predict four classes (0.58-0.83), as expected from the random probability of selecting the correct answers of 0.50 and 0.25 for the two and four class prediction schemes, respectively (**Figure** 3). The XGBoost models tend to perform better, in terms of accuracy, than the random forest models. To determine which part of the strain accumulation process these models most successfully describe, we extract the non-transfer precision and recall of each prediction class for the 24 rock type models (**Figure** S2). In general, the precision scores indicate that the model predictions are similarly successful across all the prediction classes. The recall scores indicate that the models are more correct near the onset of loading.

We next assess the ability of the models to predict the proximity to failure with data from experiments on rock types different than those used to train the models (i.e., transfer accuracy). The non-transfer accuracy reflects the success of the model in predicting the outcome with data from the same rock type. The transfer accuracy reflects the success of the model in predicting the outcome with data of different rock types. High transfer accuracies may indicate that the rock type used in the testing dataset fails with similar a strain accumulation process as the rock type used in the training data. Note that higher non-transfer accuracies may arise if samples used in the training dataset are moderately-highly correlated to samples used in the testing dataset. We show below that this caveat has minimal impact on the model success.

To compare the similarity of the strain accumulation processes in these rock types, we synthesize the accuracies of the models trained and tested with data from the 36 rock type pairs in each of the four machine learning approaches. First, we find the accuracy of the 36 rock type pairs for the four machine learning approaches (**Figures** S3, S4). Then we calculate

the difference between the accuracy and the random probability of selecting the correct answer, which is 0.50 and 0.25 for the two and four prediction classes, respectively. Then we calculate the mean of this score from the four approaches for the 36 combinations of training and testing rock types (**Figure** 4).

We expect that certain pairs of rock types will deform with similar strain accumulation processes. For example, we expect that sandstone and basalt, monzonite and granite, and shale and limestone will host similar strain accumulation processes because previous work has observed similar deformation mechanisms operating in these rocks (e.g., *Reches & Lockner*, 1994; *Menéndez et al*., 1996; *Lockner*, 1998; *Katz & Reches*, 2004; *Zhu et al*., 2010). The results are consistent with these expectations for the sandstone and basalt experiments, and shale and limestone experiments, but not the monzonite and granite experiments (**Figure** 4). In particular, the 1) sandstone and basalt, and 2) shale and limestone datasets have high transfer accuracy, as measured with the mean accuracy above the random probability, but the 3) monzonite and granite transfer accuracies are relatively lower. Whereas the difference in the non-transfer and transfer accuracies is 0.10-0.16 for the sandstone and basalt, and <0.15 for the limestone and shale, this difference is 0.15-0.20 for the monzonite and granite. Monzonite and granite are low porosity crystalline rocks comprised of quartz and feldspar, so we would expect them to fail in similar strain accumulation processes. In contrast to expectations, the transfer accuracy is >0.15 lower than the non-transfer accuracy for these rocks. The general similarity between the range of non-transfer and transfer accuracies suggest that the influence of auto-correlation on the non-transfer model performance was minimal.

3.2. Predictive power of strain components

The importance of the features decays from the most to the least important in all 24 models of the six rock types and four machine learning approaches (**Figure** S5). The feature

importance values are measured relative to other features in the model (e.g., *Müller & Guido*, 2016). However, the similar magnitudes of the feature importance across the models indicate that we may define thresholds in order to identify the highly important features of each model to then compare to other models. Consequently, we examine the average importance of features that have >25%, >50% and >75% importance of the maximum feature importance (**Figure** 5). To ensure that the magnitudes of the suites of feature importance of different models do not influence our results, we also examine the feature importance normalized by the maximum feature importance of each model. This normalization yields the same conclusions (described below) as analyzing the importance without normalization.

To assess which strain components (contraction, dilation, and shear strain) provide the best predictive power of the proximity to macroscopic failure, we examine the importance of features, grouped by strain component, that are identified as highly important. With the least restrictive inclusion of features (>25% of the maximum importance) (**Figure** 5), features that include information about all of the strain components are identified as important for each rock type. With the intermediate inclusion of features (>50%), a lower number of rock type models depend on the contraction and shear strain than with the >25% threshold. With the most restrictive inclusion of features (>75%), features that include information about the dilation are considered important for five of six of the rock types. Only the limestone models rely on information about the contraction, dilation and shear strain to predict failure with this 75% thresholding.

To assess which statistics of the strain populations provide the best predictive power about the proximity to macroscopic failure, we now examine the average importance of features that use these statistics. For each threshold of importance (25-75%), features using the sum, mean, and 25^{th}-75^{th} percentile of the strain population within each sampling subvolume have the highest importance (**Figure** 6, **Figure** S6). Features that use the extreme values of the strain

301 population, including the highest (90^{th} percentile) and lowest (10^{th} percentile) extremes, do
302 not rank as high in importance as the intermediate values. Previous machine learning analyses
303 have used the variance, or standard deviation, of acoustic emissions to detect approaching
304 failure in friction experiments (e.g., *Rouet-Leduc et al*., 2017; *Hulbert et al*., 2019). In our
305 analysis, the standard deviation of the strain components in a subvolume does not provide
306 significant predictive power of the timing of macroscopic failure (**Figure** 6, **Figure** S6).
307 To test the robustness of the results gleaned from the feature importance, we assess the
308 predictive power of the features using their Shapley Additive Explanation (SHAP) values
309 (*Lundberg & Lee*, 2017). Consistent with the analysis of the suites of feature importance, this
310 analysis suggests that features that include information about the dilation, and intermediate
311 and global values of the strain population provide the best predictive power (**Figure** S7).
312 To further explore the robustness of the results derived from the feature importance and
313 SHAP values, we train and test new models that only use features that include information
314 about the dilation, rather than all three strain components (**Figure** S8). The accuracies of the
315 models decrease by 0.01-0.10 when they only rely on the dilation. Both the feature
316 importance and SHAP values indicate that the statistics that provide the highest predictive
317 power tend to measure the intermediate values of the dilation population, including the mean,
318 median, sum, and number of values within a subvolume.
319 To help understand why the models select these sets of features as important, we examine
320 the evolution of the magnitude of the strain components throughout three characteristic
321 experiments (**Figure** 7). With the highest threshold of importance (>75%), these experiments
322 depend on features with information about the dilation (sandstone and monzonite), and
323 contraction (limestone), respectively (**Figure** 7). These strain components tend to have higher
324 mean values of the total population at a stress step than the strain components that are not
325 identified as important. In the sandstone experiment (**Figure** 7a), the mean of the population

of dilation values in each DVC calculation exceeds the mean of the shear strain and contraction after 0.6 normalized macroscopic axial strain. This trend holds for the dilation and contraction in the monzonite and limestone experiments, respectively, for at least 90% of the macroscopic axial strain. Interestingly, the trend of the dilation relative to axial strain shows a clear acceleration toward failure in the sandstone experiment (**Figure** 7a). However, the mean values of the strain components identified as important in the other experiments do not increase toward failure (**Figure** 7b-c).

We would expect that the strain components that help predict the proximity to macroscopic failure evolve over increasing differential stress (and axial strain) such that the strain population at low differential stress would be different from the strain population at higher differential stress. However, this expected trend is only evident in the sandstone experiment, at least by tracking the mean of the strain population throughout the rock core. Although the mean values of the dilation and contraction do not systematically increase in the monzonite and limestone experiments, the machine learning analyses reveal that these strain components can help predict the proximity to macroscopic failure. Thus, the machine learning analyses provide insights into the predictive power of these strain components that are not evident from this visual inspection of the evolution of the strain components.

4. Discussion

4.1. Quantifying the similarity of the precursory strain process in rocks

Previous studies characterized the micromechanics of deformation preceding macroscopic failure for different rock types (e.g., *Reches & Lockner*, 1994; *Menéndez et al*., 1996; *Lockner*, 1998; *Katz & Reches*, 2004; *Zhu et al*., 2010). The deformation mechanisms developed for different rock types, such as pore collapse and Hertzian fracturing in porous sedimentary rocks (e.g., *Menéndez et al*., 1996; *Zhu et al*., 2010), have not been applied to

other (e.g., low porosity) rock because the mechanisms generally depend on the microstructure of the rock. Machine learning analyses such as those employed here can clarify which rock types, from sandstone to granite, have similar strain accumulation processes preceding macroscopic failure. The similarity of processes may highlight rock types that share a dominant deformation mechanism preceding dynamic rupture.

We may expect that 1) sandstone and basalt, 2) monzonite and granite, and 3) shale and limestone have similar dominant strain accumulation processes. Fontainebleau sandstone and Mt. Etna basalt are both porous rocks associated with failure mechanisms that include stress concentrations at the edges of pores and/or grains that promote fracture nucleation, and the potential of fracture termination at the edges of pores and/or grains (e.g., *Stanchits et al*., 2006; *Zhu et al*., 2010). Monzonite and granite are both low porosity crystalline rocks comprised of similar minerals with lower degrees of local mechanical heterogeneity (i.e., compressive strength and stiffness) than sedimentary rocks with higher porosity. Shale and limestone are associated with compactive failure mechanisms, such as pore-collapse and compaction bands, and accommodating larger magnitudes of inelastic deformation prior to macroscopic failure (e.g., *Railsback*, 1993; *Baud et al*., 2000; *McBeck et al*., 2018; *Huang et al*., 2019). However, both shale and limestone can exhibit macroscopic dilation under triaxial compression deformation (e.g., *Baud et al*., 2000).

The high transfer accuracies of the developed machine learning models suggest fundamental similarities of the strain accumulation process in the examined rock types (**Figure** 4). Some trends in the transfer accuracies are expected: the 1) sandstone and basalt, and 2) shale and limestone models and datasets produce high transfer accuracies (**Figure** 4). Some trends in the transfer accuracies are unexpected. For example, the granite and monzonite models and datasets do not have high transfer accuracy. One explanation of this difference could be the varying ranges of confining stresses applied to the granite (5-10 MPa)

and monzonite (25-35 MPa) experiments. Another difference between Westerly granite and monzonite is their mean grain sizes. The Westerly granite used here has a grain size in the range 100-200 μm, whereas monzonite has a larger mean grain size of 450 μm (*Aben et al.,* 2016). This difference in grain sizes may cause fractures to be impeded at grain boundaries in the granite at higher rates than in the monzonite. This higher rate of impedance could result in more distributed fracture networks in the granite than the monzonite, which may produce differing degrees of strain localization preceding dynamic failure in the rocks. These varying degrees of strain localization, in turn, may produce the lower transfer accuracies of models trained and tested with the local strain data of the monzonite and granite experiments.

The high transfer accuracies, and related similarity of the suites of feature importance, for the 1) sandstone and basalt, 2) shale and limestone, and 3) monzonite, shale and limestone datasets suggest that a unified general theory may be able to describe the strain accumulation process in these groups of rock types.

4.2. Predicting the proximity to failure with the intermediate dilatational strain

The ranking of the predictive importance among the strain components and statistics is similar for the four machine learning approaches and six rock types (**Figures** 6, S7, S8). This consistency provides confidence in the overall conclusion that the intermediate values of the dilatational strain provide the best predictive power of the strain components and statistics. We now tentatively extend the observations made from the experimental data to implications for detecting the earthquake preparation process. Although earthquakes have been idealized as dominantly frictional phenomena, interseismic strengthening of fault zones and surrounding crustal volume may increase both the cohesive and frictional strength (e.g., *Qiu et al*., 2020). This strengthening may then cause earthquakes to involve the fracture of partially healed fault zone material and the surrounding crust, and then subsequent relocalization followed by frictional sliding. Consequently, macroscopic failure under triaxial compression performed in

these experiments may represent the relocalization deformation stage in a partially healed cohesive fault zone prior to an earthquake, before frictional slip along a well-defined principal slip zone begins.

The conclusion that the intermediate values of the local strain have greater predictive power than the extreme values of the local strain suggests that in order to predict seismic hazard, it is important to examine data recorded in the volume of crust adjacent to the main faults that localize the highest magnitudes of strain. Such analyses should consider the seismicity and deformation in crustal volumes around major faults to estimate the strain value within the 25^{th}-75^{th} percentile of the population, rather than the extreme values that eventually localize along the principal slip zones. The conclusion that the dilatational strain provides the highest predictive power, relative to the contraction and shear strain, for a diverse range of rock types and four machine learning approaches, suggests that previous methods that track the influence of dilatancy (e.g., *Whitcomb et al*., 1973) may be able to constrain the timing of earthquake rupture (e.g., **Figure** 8).

5. Conclusions

Proposed theories of the micromechanical deformation of rock leading to macroscopic failure tend to depend on the examined rock type (e.g., *Katz & Reches*, 2004). Developing a general method that predicts the timing of dynamic rupture for a wide variety of rock types, from porous sandstone to lower porosity granite, would represent a fundamental advance in rock physics. Moreover, an improved ability to forecast system-size ruptures would have significant societal benefits. Our machine learning analyses suggest that groups of rock types have similar strain accumulation processes preceding macroscopic failure. The similarity of these processes may highlight a dominant mechanism of deformation preceding system-size rupture. The models suggest that the precursory strain accumulation process is similar in 1)

sandstone and basalt, 2) shale and limestone, and 3) monzonite, shale and limestone (**Figures** 4, S4). These results indicate that a wide range of rock types accumulate precursory strain with similar expressions before macroscopic failure.

Our suite of machine learning analyses suggests that the intermediate values (25^{th}-75^{th} percentile) of the local dilatational strain provide the highest predictive power of the proximity to macroscopic failure for a diverse group of rock types including basalt, granite and shale under the stress conditions of the upper crust (e.g., **Figure** 6). This result suggests that tracking deformational processes that produce dilatancy (i.e., opening fractures), and the corresponding geophysical signatures such as variations in seismic velocities and density, can provide accurate estimates of the onset of macroscopic failure without relying on explicit information about the shear deformation. The results of this work provide incentives to enhance monitoring efforts focused on effects of dilatancy in crustal volumes around large faults.

Acknowledgements

We thank Elodie Boller, Paul Tafforeau, and Alexander Rack for providing advice on the design of the tomography setup, Benoit Cordonnier for experimental expertise, and Sanchez Technology for building the deformation apparatus. This study was funded by the Norwegian Research Council (grant 272217) and the U.S. Department of Energy (award DE-SC0016520). The experimental data (incremental strain fields) are available on UNINET Sigma2 (the National Infrastructure for High Performance Computing and Data Storage in Norway) (*McBeck*, 2020). Editor Rebecca Bendick and two anonymous reviewers helped improve this manuscript.

References

Aben, F. M., Doan, M.-L., Mitchell, T. M., Toussaint, R., Reuschlé, T., Fondriest, M., Gratier, J.-P., and Renard, F. (2016). Dynamic fracturing by successive coseismic loadings leads to pulverization in active fault zones, *Journal of Geophysical Research: Solid Earth*, 121, 2338– 2360, doi:10.1002/2015JB012542.

Abercrombie, R. E., & Mori, J. (1996). Occurrence patterns of foreshocks to large earthquakes in the western United States. *Nature*, *381*(6580), 303-307.

Aggarwal, Y. P., Sykes, L. R., Armbruster, J., & Sbar, M. L. (1973). Premonitory changes in seismic velocities and prediction of earthquakes. *Nature*, *241*(5385), 101-104.

Baud, P., Schubnel, A., & Wong, T. F. (2000). Dilatancy, compaction, and failure mode in Solnhofen limestone. *Journal of Geophysical Research: Solid Earth*, *105*(B8), 19289-19303.

Ben-Zion, Y. (2008). Collective Behavior of Earthquakes and Faults: Continuum-Discrete Transitions, Evolutionary Changes and Corresponding Dynamic Regimes, *Rev. Geophysics*, 46, RG4006, doi:10.1029/2008RG000260.

Brace, W. F., Paulding, B. & Scholz, C. (1966), Dilatancy in the fracture of crystalline rocks, *Journal of Geophysical Research*, 71, 3939-3953, doi:10.1029/JZ071i016p03939

Breiman, L. (2011). Random forests. *Machine learning* 45.1 5-32.

Bolt, B. (1977). Constancy of P travel times from Nevada explosions to Oroville Dam station 1970-1976, *Bull. Seism. Soc. Am*. 67, 27-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (785-794).

Frank, F.C., 1965. On dilatancy in relation to seismic sources. *Reviews of Geophysics*, *3*(4), 485-503.

Haase, J.S., Shearer, P.M. & Aster, R.C. (1995). Constraints on temporal variations in velocity near Anza, California, from analysis of similar event pairs, *Bull. Seismol. Soc. Am*., 85, 194-206.

Hamiel, Y., Liu, Y., Lyakhovsky, V., Y. Ben-Zion, Y. and Lockner, D. (2004). A Visco-Elastic Damage Model with Applications to Stable and Unstable fracturing, *Geophys. J. Int.*, 159, 1155-1165.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hauksson, E., Yang, W., & Shearer, P. M. (2012). Waveform relocated earthquake catalog for southern California (1981 to June 2011). *Bulletin of the Seismological Society of America*, 102(5), 2239-2244.

Huang, L., Baud, P., Cordonnier, B., Renard, F., Liu, L., & Wong, T.-F. (2019). Synchrotron X-ray imaging in 4D: Multiscale failure and compaction localization in triaxially compressed porous limestone, *Earth and Planetary Science Letters*, *528*, https://doi.org/10.1016/j.epsl.2019.115831.

Hulbert, C., Rouet-Leduc, B., Johnson, P. A., Ren, C. X., Rivière, J., Bolton, D. C., & Marone, C. (2019). Similarity of fast and slow earthquakes illuminated by machine learning. *Nature Geoscience*, *12*(1), 69.

Géron, A., (2017). Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, Inc., Boston, MA.

Kandula, N., Cordonnier, B., Weiss, J., Dysthe, D. K., & Renard, F. (2019). Dynamics of microscale precursors during brittle compressive failure in Carrara marble, *Journal of Geophysical Research, 124*, 6121-6139, doi: 10.1029/2019JB017381.

Katz, O., & Reches, Z. E. (2004). Microfracturing, damage, and failure of brittle granites. *Journal of Geophysical Research: Solid Earth*, *109*(B1), B01206, doi:10.1029/2002JB001961.

Kreemer, C., Holt, W. E., & Haines, A. J. (2003). An integrated global model of present-day plate motions and plate boundary deformation. *Geophysical Journal International*, *154*(1), 8-34.

Lockner, D. A. (1998). A generalized law for brittle deformation of Westerly granite. *Journal of Geophysical Research: Solid Earth*, *103*(B3), 5107-5123.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).

Lyakhovsky, V., Y. Ben-Zion and A. Agnon (1997). Distributed Damage, Faulting, and Friction, *J. Geophys. Res*., 102, 27635-27649.

Menéndez, B., Zhu, W., & Wong, T. F. (1996). Micromechanics of brittle faulting and cataclastic flow in Berea sandstone. *Journal of Structural Geology*, *18*(1), 1-16.

McBeck, J. (2020). The mixology of precursory strain partitioning approaching brittle failure in rocks [Data set]. *Norstore*. https://doi.org/10.11582/2020.00002.

McBeck, J., Ben-Zion, Y., & Renard, F. (2020). The mixology of precursory strain partitioning approaching brittle failure in rocks. *Geophysical Journal International*, https://doi.org/10.1093/gji/ggaa121.

McBeck, J. A., Cordonnier, B., Vinciguerra, S., & Renard, F. (2019). Volumetric and shear strain localization in Mt. Etna basalt. *Geophysical Research Letters*, *46*(5), 2425-2433.

McBeck, J., Kobchenko, M., Hall, S. A., Tudisco, E., Cordonnier, B., Meakin, P., & Renard, F. (2018). Investigating the onset of strain localization within anisotropic shale using digital volume correlation of time-resolved X-ray microtomography images. *Journal of Geophysical Research: Solid Earth*, *123*(9), 7509-7528.

McEvilly, T. V. and L. R. Johnson (1974). Stability of P and S velocities from central California quarry blasts, *Bull. Seism. Soc. Am*. 64, 343-353.

Mogi, K. (1969). Some feature of recent seismic activity in and near Japan (2), activity before and after great earthquakes. *Bull. Earthq. Res. Inst., Univ. Tokyo*, *47*, 395-417.

Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc., Sebastopol, CA.

Nur, A. (1974). Matsushiro, Japan, earthquake swarm: Confirmation of the dilatancy-fluid diffusion model. *Geology*, *2*(5), 217-221.

Papazachos, B. C. (1973). The time distribution of the reservoir-associated foreshocks and its importance to the prediction of the principal shock. *Bulletin of the Seismological Society of America*, *63*(6-1), 1973-1978.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Mathieu Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,

534 Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011). Scikit-learn: Machine
535 learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

536 Qiu, H., Hillers, G. and Ben-Zion, Y. (2020). Temporal changes of seismic velocities in the
537 San Jacinto Fault zone associated with the 2016 Mw5.2 Borrego Springs earthquake,
538 *Geophys.*, *J. Int.*, 220, 1536–1554, doi: 10.1093/gji/ggz538.

539 Railsback, L. B. (1993). Contrasting styles of chemical compaction in the Upper
540 Pennsylvanian Dennis Limestone in the Midcontinent region, USA. *Journal of Sedimentary*
541 *Research*, *63*(1), 61-72.

542 Reches, Z. E., & Lockner, D. A. (1994). Nucleation and growth of faults in brittle
543 rocks. *Journal of Geophysical Research: Solid Earth*, *99*(B9), 18159-18173.

544 Renard, F., McBeck, J., Cordonnier, B., Zheng, X., Kandula, N., Sanchez, J. R., Kobchenko,
545 M., Noiriel, C., Zhu, W., Meakin, P., & Fusseis, F. (2019). Dynamic In Situ Three-
546 Dimensional Imaging And Digital Volume Correlation Analysis to quantify strain localization
547 and fracture coalescence in sandstone. *Pure and Applied Geophysics*, *176*(3), 1083-1115.

548 Renard, F., Weiss, J., Mathiesen, J., Ben-Zion, Y., Kandula, N., & Cordonnier, B. (2018).
549 Critical evolution of damage toward system-size failure in crystalline rock. *Journal of*
550 *Geophysical Research: Solid Earth*, *123*(2), 1969-1986.

551 Roeloffs, E. A. (1988). Hydrologic precursors to earthquakes: A review. *Pure and Applied*
552 *Geophysics*, *126*(2-4), 177-209.

553 Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A.
554 (2017). Machine learning predicts laboratory earthquakes. *Geophysical Research*
555 *Letters*, 44(18), 9276-9282.

556 Scholz, C. H. (1990). The mechanics of earthquakes and faulting, *Cambridge*.

557 Scholz, C. H., Sykes, L. R., & Aggarwal, Y. P. (1973). Earthquake prediction: a physical
558 basis. *Science*, *181*(4102), 803-810.

559 Stanchits, S., Vinciguerra, S., & Dresen, G. (2006). Ultrasonic velocities, acoustic emission
560 characteristics and crack damage of basalt and granite. *Pure and Applied Geophysics*, *163*(5-
561 6), 975-994

562 Sykes, L. R., & Jaumé, S. C. (1990). Seismic activity on neighbouring faults as a long-term
563 precursor to large earthquakes in the San Francisco Bay area. *Nature*, *348*(6302), 595-599.

564 Tiampo, K. F., & Shcherbakov, R. (2012). Seismicity-based earthquake forecasting
565 techniques: Ten years of progress. *Tectonophysics*, *522*, 89-121.

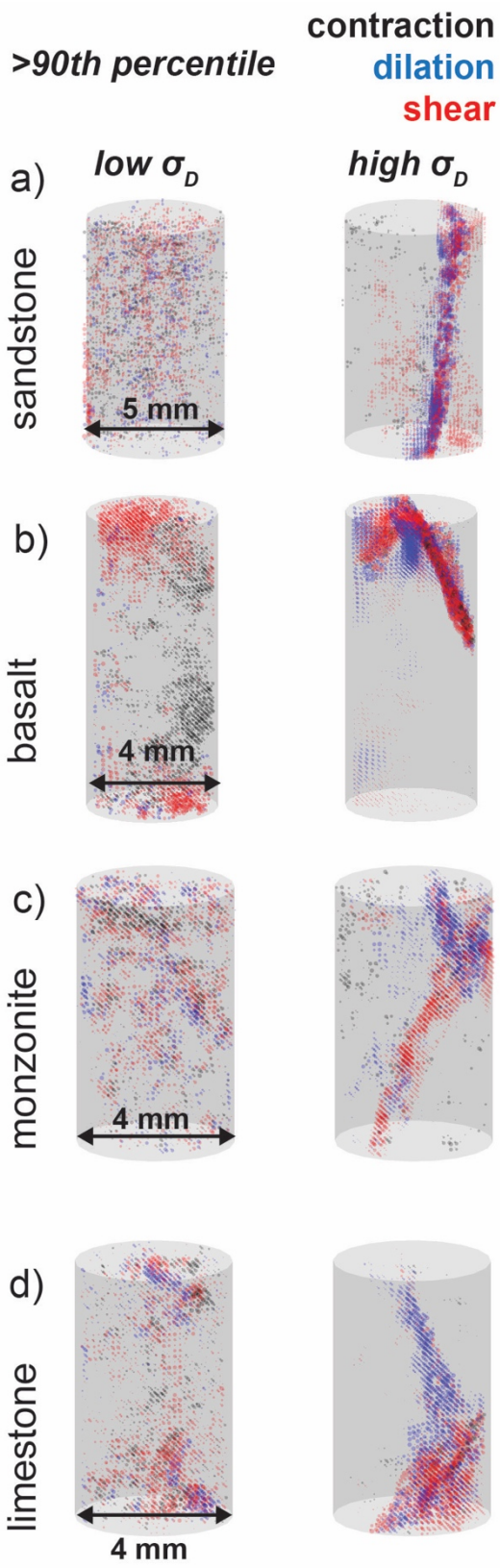
566 Tudisco, E., Andò, E., Cailletaud, R., & Hall, S.A. (2017). TomoWarp2: a local digital
567 volume correlation code. *SoftwareX*, 6, 267–270.

568 Whitcomb, J. H., Garmany, J. D., & Anderson, D. L. (1973). Earthquake prediction: Variation
569 of seismic velocities before the San Francisco earthquake. *Science*, *180*(4086), 632-635.

570 Zhu, W., Baud, P. & Wong, T.F. (2010). Micromechanics of cataclastic pore collapse in
571 limestone. *Journal of Geophysical Research: Solid Earth*, *115*(B4),
572 doi:10.1029/2009JB006610.

Figure 1

Characteristic snapshots of the strain fields in a) sandstone, b) basalt, c) monzonite and d) limestone under lower and higher differential stress, σ_D. The dots show where the value of the contraction (black), dilation (blue) and shear strain (red) is greater than the 90th percentile of the population of the given strain component. The localization of strain components evolves from lower to higher σ_D, suggesting that the strain fields may indicate the proximity to macroscopic failure.



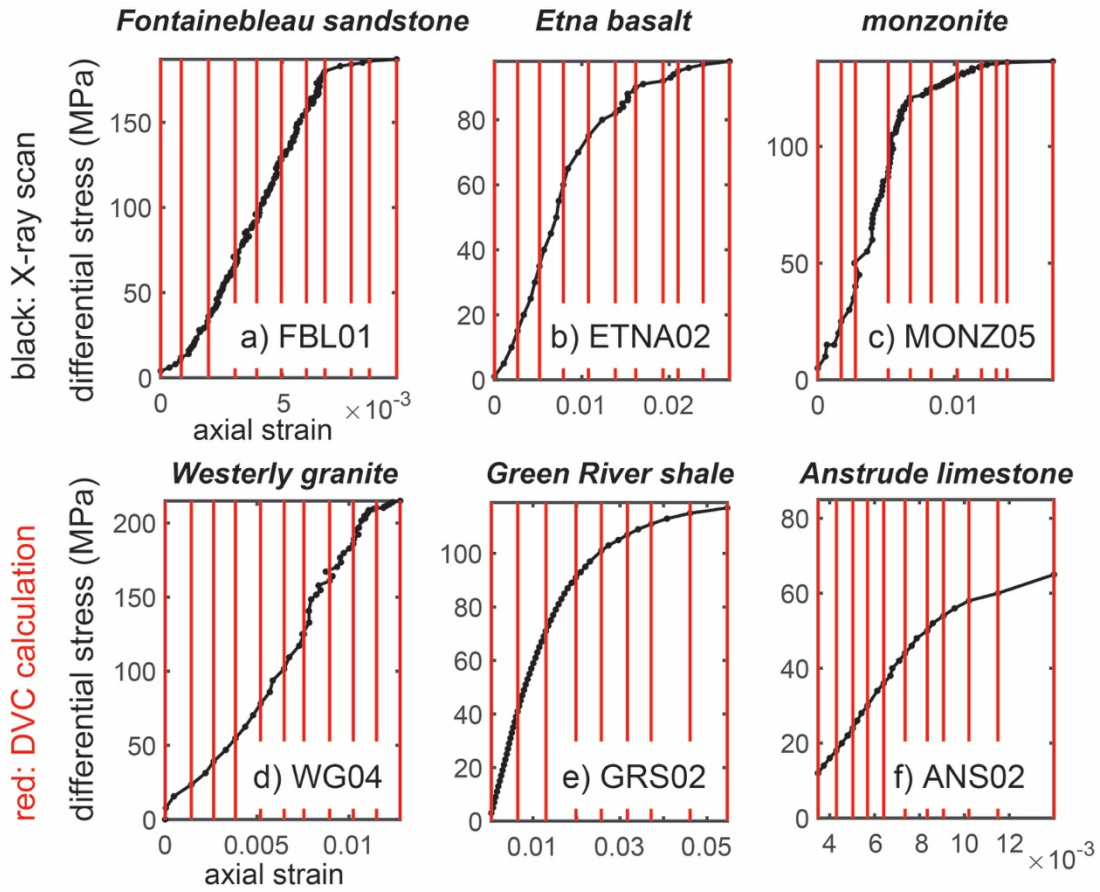


Figure 2

Loading history and conditions of digital volume correlation calculations for experiments on sandstone (a), basalt (b), monzonite (c), granite (d), shale (e), and limestone (f). Black dots show the differential stress and axial strain when each tomogram (i.e., 3D density field) was acquired. Red lines show the loading conditions of the tomograms used in each digital volume correlation calculation. In each experiment, macroscopic failure occurs <0.1 MPa after the maximum differential stress.

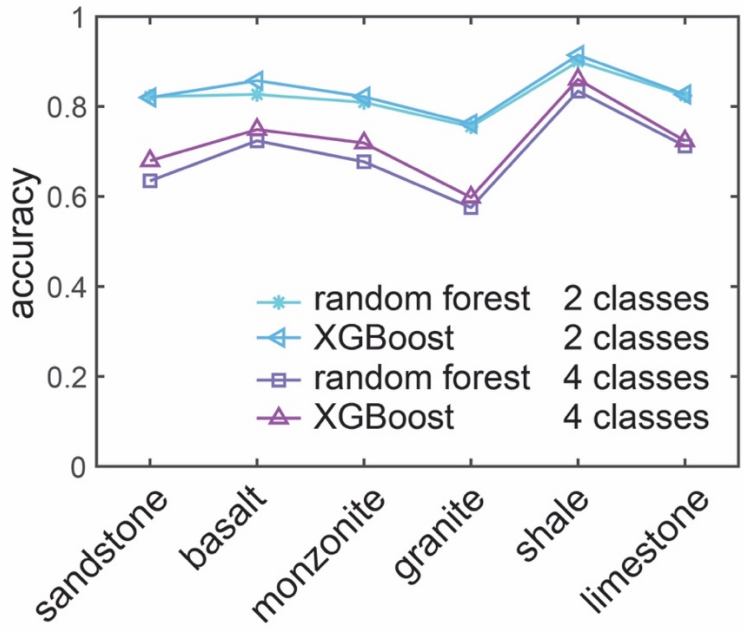


Figure 3

Accuracy of the six rock type models for the four machine learning approaches, including predicting two and four classes of the proximity to failure, with the random forest and XGBoost algorithms. The predictions of two classes have a higher range of accuracy (0.76-0.89) than the predictions of four classes (0.58-0.83). When compared to the random probability of selecting the correct answer, the predictions of four classes are more correct than the predictions of two classes. With two and four classes of predictions, the random probability is 1/2 and 1/4, respectively, so the accuracy above the random probability is higher for the four class predictions than the two class predictions. The XGBoost models tend to perform slightly better than the random forest models.

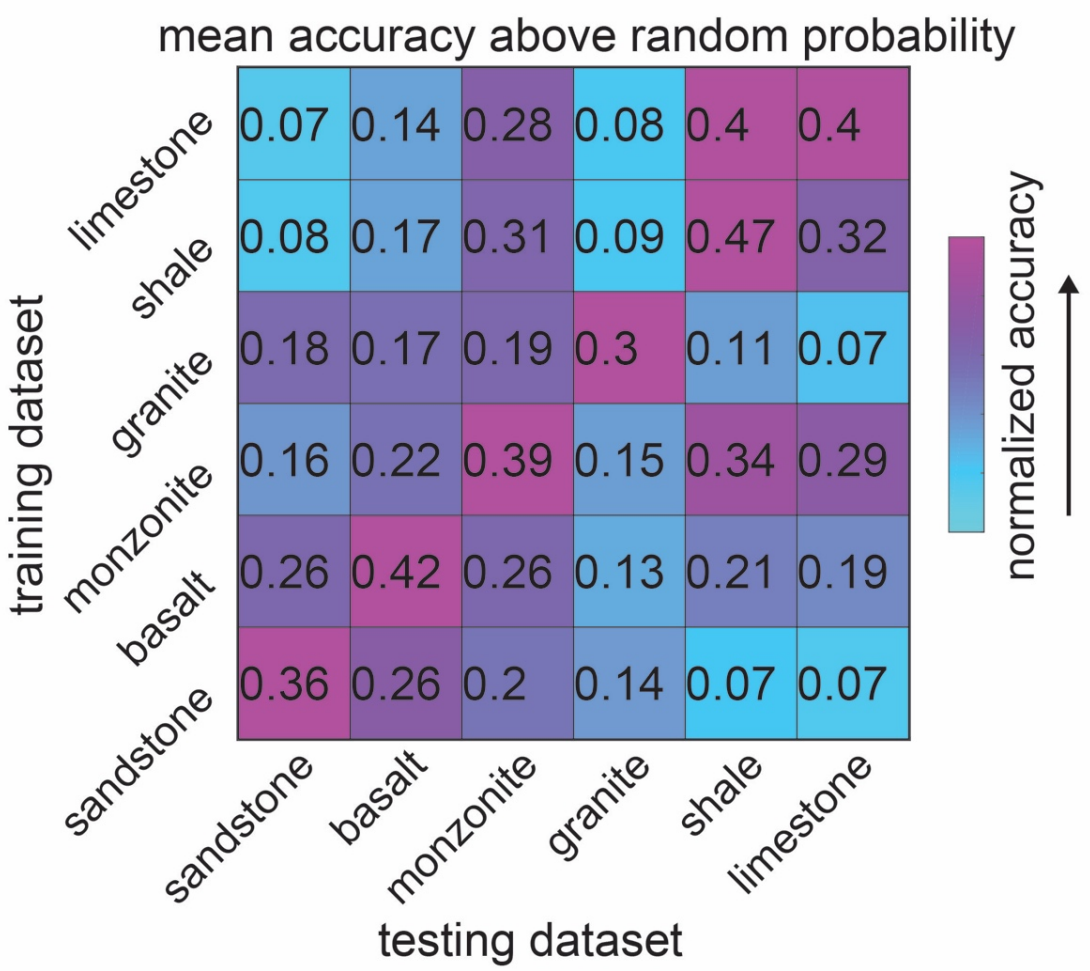


Figure 4

Cumulative accuracy of models when trained and tested on the same and different rock types. To estimate the mean accuracy above the random probability (shown here with the text in the boxes), we first calculate the accuracy above the random probability for the four machine learning approaches (e.g., **Figures** S3, S4), and then we calculate the mean of this score from the four approaches for the 36 combinations of training and testing rock types. The rows indicate the rock type used to train the model. The columns indicate the rock type used to test the model. The color is the mean accuracy above the random probability normalized by the maximum value in each row. We expect to find high transfer accuracies for the 1) sandstone and basalt, 2) monzonite and granite, 3) and limestone and shale. However, we only observe this trend for 1) and 3). The other unexpected trend is the high transfer accuracies for monzonite, shale and limestone.

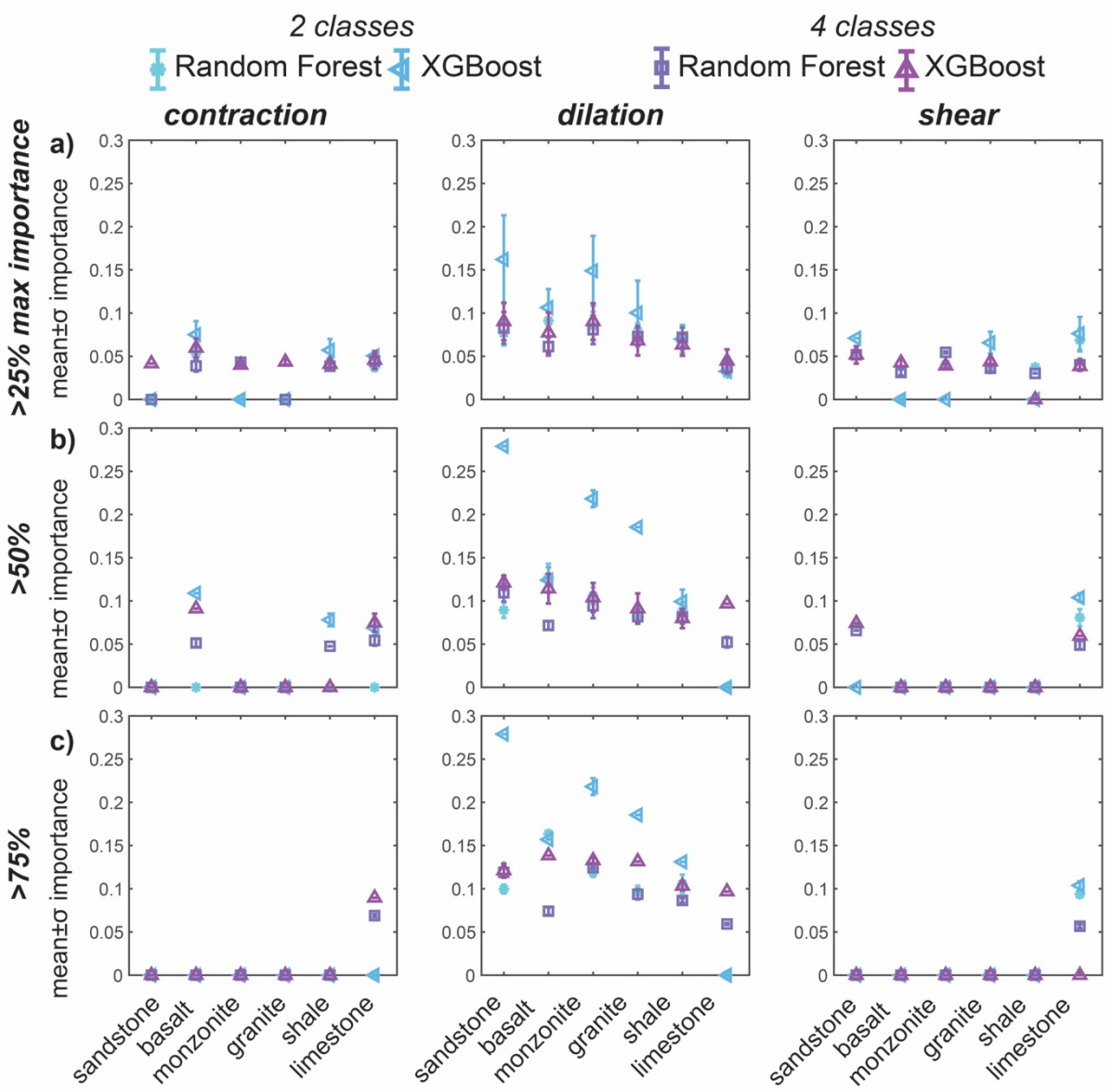
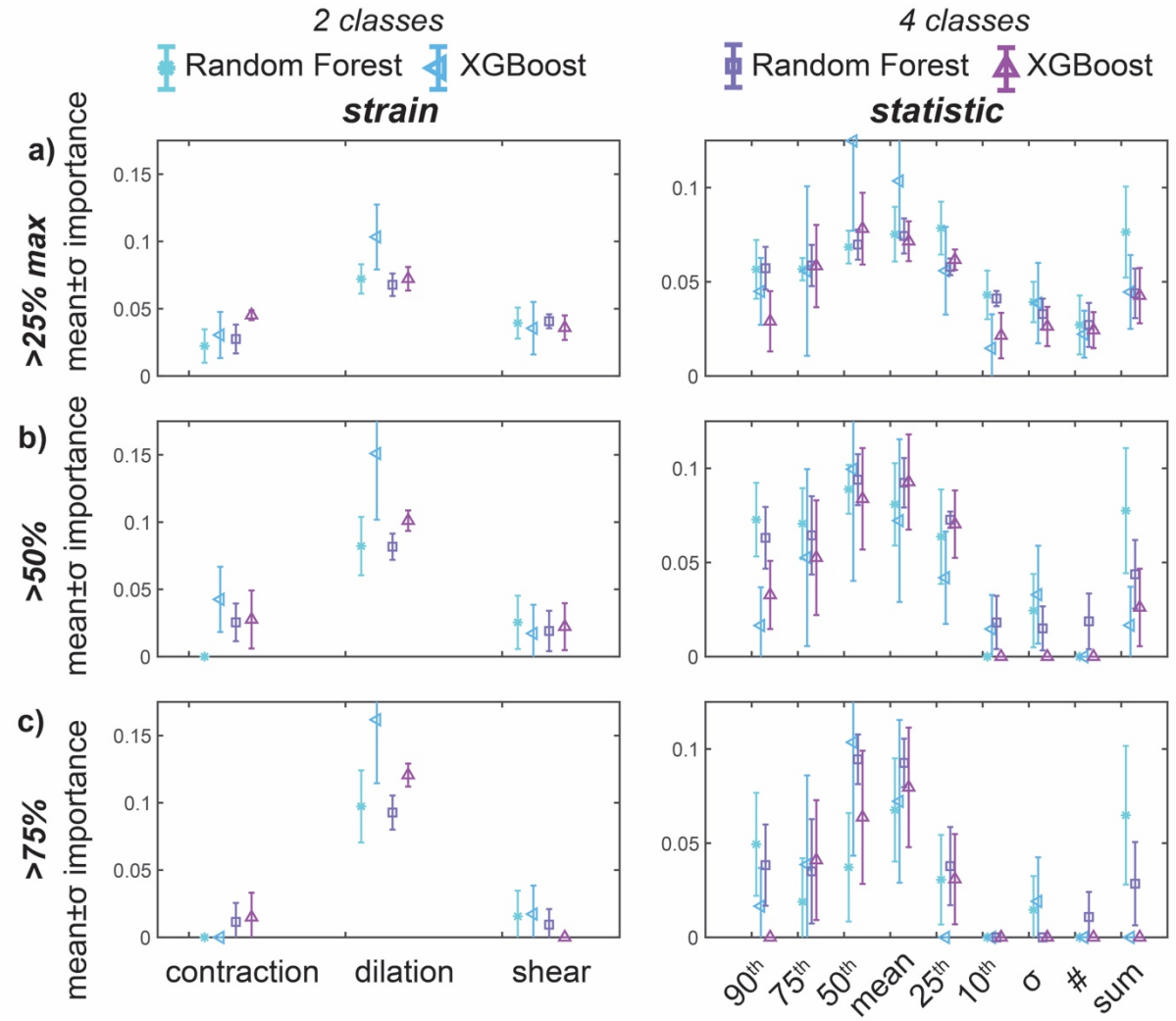


Figure 5

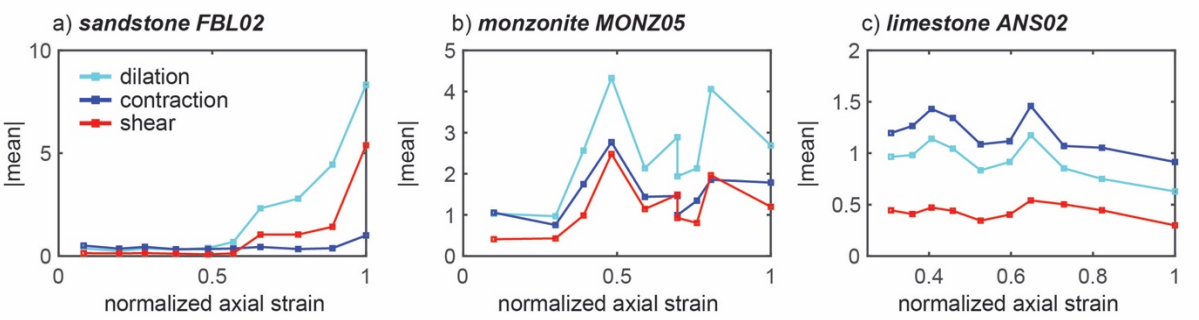
Importance of groups of features that include information about the contraction (left), dilation (center), and shear strain (right), for each rock type model and machine learning approach, for features with importance >25% (a), >50% (b), and >75% (c) of the maximum importance of each model. At the lowest threshold (a, 25%), most of the models depend on features with each of the strain components, but the features with dilation tend to have the highest importance. At the highest threshold (c, 75%), all of the models except the limestone models depend only on the dilatation strain. For all of the other rock types and the machine learning approaches, features with dilation have the highest predictive power of the strain components.



627

628 **Figure 6**

629 Importance of groups of features with each strain component (left) and statistic (right), averaged
630 over all the rock types for each machine learning approach, for features with importance >25%
631 (a), >50% (b), and >75% (c) of the maximum importance of each model. With increasingly
632 important groups of features, dilation increases in average importance, while contraction and
633 shear strain decrease in average importance. The intermediate values of the populations (25^{th}-
634 75^{th}) tend to have the highest importance of the statistics.



635

636 **Figure 7**

637 Evolution of the magnitude of incremental strain components throughout three experiments: a)
638 sandstone experiment FBL01, b) monzonite experiment MONZ05, c) limestone experiment
639 ANS02, shown as the mean of the incremental strain population divided by the incremental
640 macroscopic axial strain (vertical axis), relative to the normalized macroscopic axial strain
641 (horizontal axis). The strain components identified as highly important in each rock type model
642 (dilation for sandstone and monzonite, contraction for limestone) tend to have higher mean
643 values than the other strain components. Only the dilation in the sandstone experiment (a)
644 shows a systematic evolution with increasing axial strain, suggesting that this component best
645 predicts the proximity to failure, as confirmed by the machine learning methods. In contrast,
646 the dilation in the monzonite (b) and contraction in the limestone (c) experiments do not show
647 a systematic evolution with distance toward failure, although the machine learning methods
648 identify these strain components as the best predictors of the proximity to failure.

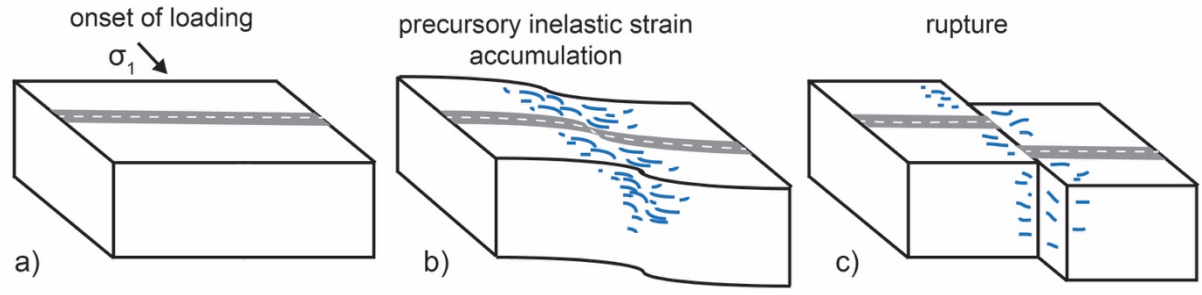


Figure 8

Development of precursory inelastic dilatational strain toward catastrophic rupture. The intermediate values of the dilatational strain population provide the best predictive power about the proximity to macroscopic failure in triaxial compression experiments at the conditions of the upper crust in 24 models with six rock types and four machine learning approaches.