nature geoscience

ARTICLES

https://doi.org/10.1038/s41561-018-0272-8

Similarity of fast and slow earthquakes illuminated by machine learning

Claudia Hulbert[1]*, Bertrand Rouet-Leduc[1], Paul A. Johnson[1], Christopher X. Ren[1], Jacques Rivière[2], David C. Bolton[3] and Chris Marone[3]

Tectonic faults fail in a spectrum of modes, ranging from earthquakes to slow slip events. The physics of fast earthquakes are well described by stick–slip friction and elastodynamic rupture; however, slow earthquakes are poorly understood. Key questions remain about how ruptures propagate quasi-dynamically, whether they obey different scaling laws from ordinary earthquakes and whether a single fault can host multiple slip modes. We report on laboratory earthquakes and show that both slow and fast slip modes are preceded by a cascade of micro-failure events that radiate elastic energy in a manner that foretells catastrophic failure. Using machine learning, we find that acoustic emissions generated during shear of quartz fault gouge under normal stress of 1–10 MPa predict the timing and duration of laboratory earthquakes. Laboratory slow earthquakes reach peak slip velocities of the order of 1×10^{-4} m s^{-1} and do not radiate high-frequency elastic energy, consistent with tectonic slow slip. Acoustic signals generated in the early stages of impending fast laboratory earthquakes are systematically larger than those for slow slip events. Here, we show that a broad range of stick–slip and creep–slip modes of failure can be predicted and share common mechanisms, which suggests that catastrophic earthquake failure may be preceded by an organized, potentially forecastable, set of processes.

Tectonic faults slip in a variety of modes that range from earthquakes with strong shaking to transient slow slip and aseismic creep[1–7]. The mechanics of earthquakes are well described by frictional instability and fracture propagation with attendant elastic radiation[8,9]. Aseismic fault creep is also reasonably well understood as stable frictional shear driven by stress relaxation. However, seismic tremor and other modes of slow slip that include transient acceleration and self-driven propagation, known collectively as slow earthquakes, are not well understood, despite the large and rapidly growing number of observations that now extend to nearly every major tectonic fault system on Earth[6,10–12]. Central questions about the mechanics of slow earthquakes include (1) the mechanism(s) that limit fault slip speed and rupture propagation velocities during quasi-dynamic, self-propagating rupture[13], (2) the physics that allow earthquakes and slow slip events on the same fault segment[10,11], (3) how slow slips can precede and possibly trigger dynamic rupture[10,11,14] and (4) whether there exists a continuum of slip modes or whether rupture velocities are quantized[1,2,15].

Here, we address the question of whether slow and fast slip modes share similar mechanisms by studying acoustic signals emanating from laboratory faults. Our recent work shows that machine learning approaches can predict the timing of stick–slip failure[16] and the stress state in sheared layers of glass beads[17]. Here, we apply machine learning to investigate realistic fault zone material and a range of slip modes from fast to slow, and attempt to infer whether slow and fast earthquakes share similar mechanisms throughout the laboratory seismic cycle of loading and failure. We investigate whether seismic waves emanating from laboratory fault zones contain information about the duration and magnitude of an upcoming failure event and whether machine learning can also estimate the fault displacement history. Our experiments use fault gouge composed of quartz powder, which is similar in composition and particle size to natural, granular fault gouge. We find that laboratory slow earthquakes, which do not radiate high-frequency energy during slip, are preceded by a cascade of micro-failure events that radiate elastic energy throughout the seismic cycle and foretell catastrophic failure. For both slow and fast earthquakes, we find a mapping between fault strength and statistical attributes of the elastic radiation emitted throughout the seismic cycle. These data provide a method for reading the internal state of a fault zone, which can be used to predict earthquake-like failure for a spectrum of slip modes.

We analyse data from experiments conducted in the double-direct shear (DDS) geometry using a biaxial testing apparatus[18–20]. Two layers of simulated fault gouge are sheared simultaneously at constant normal load (in the range 1–10 MPa) and prescribed shear velocity (Fig. 1d). Details of the apparatus and testing procedures are provided in the Supplementary Information. The laboratory faults fail in repetitive cycles of stick and slip that mimic the seismic cycle of loading and failure on tectonic faults (Fig. 1). We vary the fault normal stress and loading stiffness, following recent works[21–23], to achieve a range of failure event rates corresponding to the spectrum of tectonic fault slip rates. Failure event durations and peak fault slip velocities range from 0.1 s and 0.1 m s^{-1}, for fast events, to >1 s and 1×10^{-5} m s^{-1}, for slow events, consistent with observations for tectonic faulting.

Our experiments include continuous records of fault zone elastic radiation from piezocrystals embedded in the forcing blocks of the DDS assembly[20,24]. The piezosensors are capable of resolving acoustic signals in the frequency band 0.02–2 MHz. We record acoustic emission signals throughout the seismic cycle for both fast and slow slip events. Faster events are associated with impulsive acoustic wave energy, whereas slow failure events are preceded by low-amplitude tremor. In both cases, the acoustic energy appears to be essentially featureless during the early stages of the laboratory seismic cycle (Fig. 1c).

We sheared layers of quartz powder at constant fault normal stress and prescribed slip velocities from 1 μm s^{-1} to 30 μm s^{-1}. Our

[1]Geophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA. [2]Department of Engineering Science and Mechanics, Pennsylvania State University, University Park, PA, USA. [3]Department of Geosciences, Pennsylvania State University, University Park, PA, USA. *e-mail: chulbert@lanl.gov

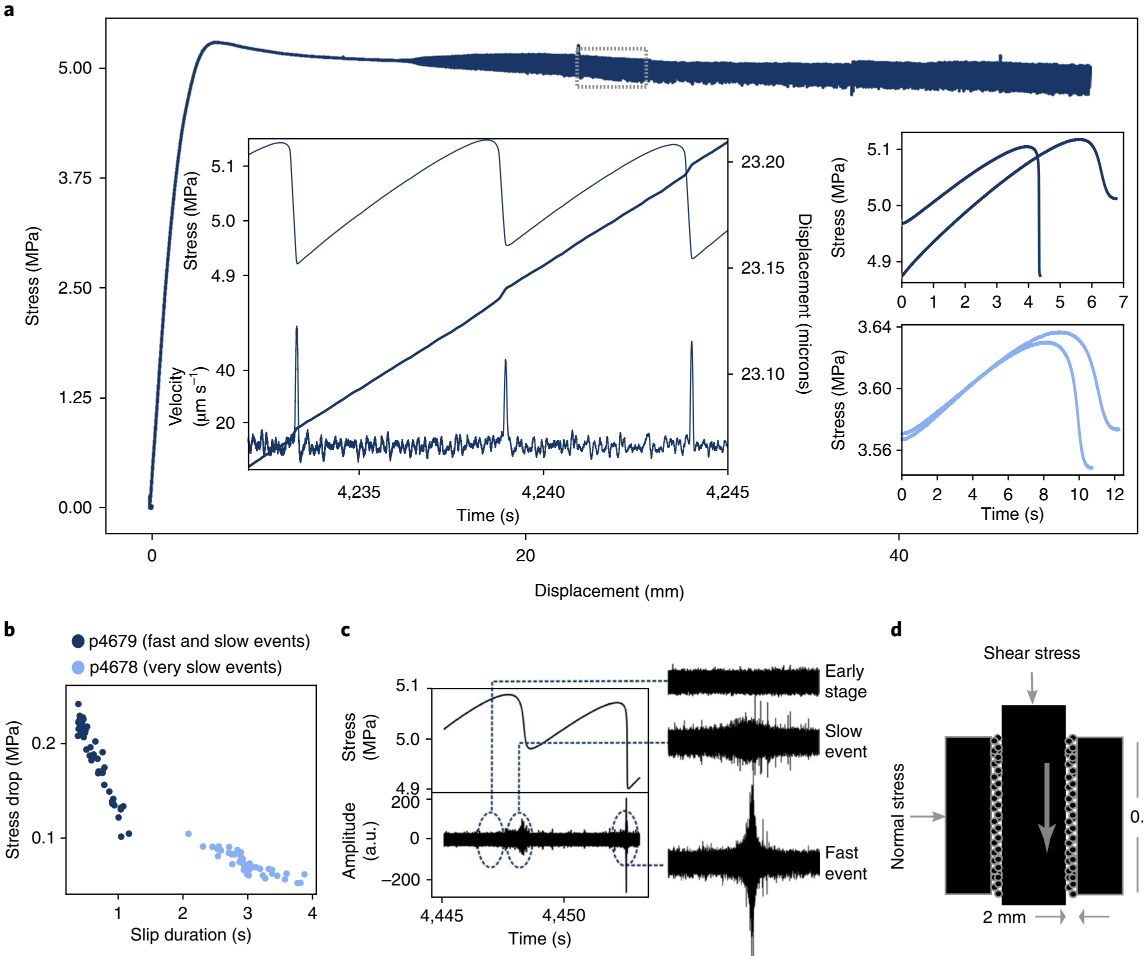

Fig. 1 | Laboratory experiments. a, Shear stress as a function of shear displacement for a complete experiment (p4679). Left inset, detail of shear stress, shear displacement and fault slip velocity (see the grey box in the main plot). Right insets, example failure events for this experiment (upper plot) and another experiment (p4678) with slower slip events (lower plot). **b**, Stress drop versus duration for all events. Stress drop decreases systematically with slip duration. **c**, Acoustic amplitude (in arbitrary units (a.u.)) and shear stress for two events, with zooms during load-up and for fast and slow failure. Elastic waves exhibit both tremor-like and impulsive signals for different slip modes. **d**, DDS configuration sheared at constant velocity $\dot{u}$, gouge layers and piezoceramic sensors (PZTs) used to record elastic waves from acoustic emissions.

experiments followed the procedures developed for a study of the spectrum of fault slip modes[21]. The ratio of loading stiffness to the critical frictional weakening rate (K/K_c) is close to 1, which produces a complex range of stick–slip and creep–slip failure modes[21,22].

We use machine learning techniques to analyse the acoustic emission records, following the approach of Rouet-Leduc and colleagues[16,17]. This approach uses a gradient boosted trees[25] algorithm, based on decision trees. We find that the timing, duration and magnitude of slow and fast laboratory earthquakes can be predicted with high fidelity.

Acoustic emissions predict fault friction and slip

Our analysis starts by using machine learning to estimate the frictional strength and slip behaviour of aperiodic stick–slip modes (Fig. 2). We rely exclusively on the continuous seismic signals recorded during shear, with the goal of inferring fault shear stress, shear displacement and gouge thickness. Model inputs consist of ~100 statistics of the seismic data calculated over a small, moving time window. Outputs are fault properties (stress, displacement, gouge thickness) over the same time window (see Supplementary Information for details). Model estimates are therefore instantaneous, as they do not make use of past or future signal history. One time window corresponds to ~5% of the average duration of one seismic cycle. This window is subdivided into two non-overlapping windows, and statistics of the seismic data are computed over both subwindows. We analyse two full experiments, with hundreds of slip events, and draw from several closely related experiments. One experiment contains both fast and slow events (p4679) and one contains only slow events (p4678). We build the machine learning model using the first half of each experiment (training set), and evaluate it on the remaining data (testing set), using the coefficient of determination R^2 as an evaluation metric. We build different models for each of the two experiments and each of the labels (stress, displacement and fault zone thickness). Hyperparameters are set using Bayesian optimization, by fivefold cross-validation. Details regarding statistical features of the acoustic data, model construction and model specifications are in Methods and the Supplementary Information.

We started using the full suite of statistical features for the machine learning model and then simplified our approach after the key features were identified. Figure 2a shows machine learning estimates of stress during a series of laboratory seismic cycles that include slow and fast events during aperiodic stick–slip (the most aperiodic suite of events we could find that includes slow slips: inter-event times vary from 3.1 s to 7 s), using all the statistical features. The top curve is the shear stress data, and the

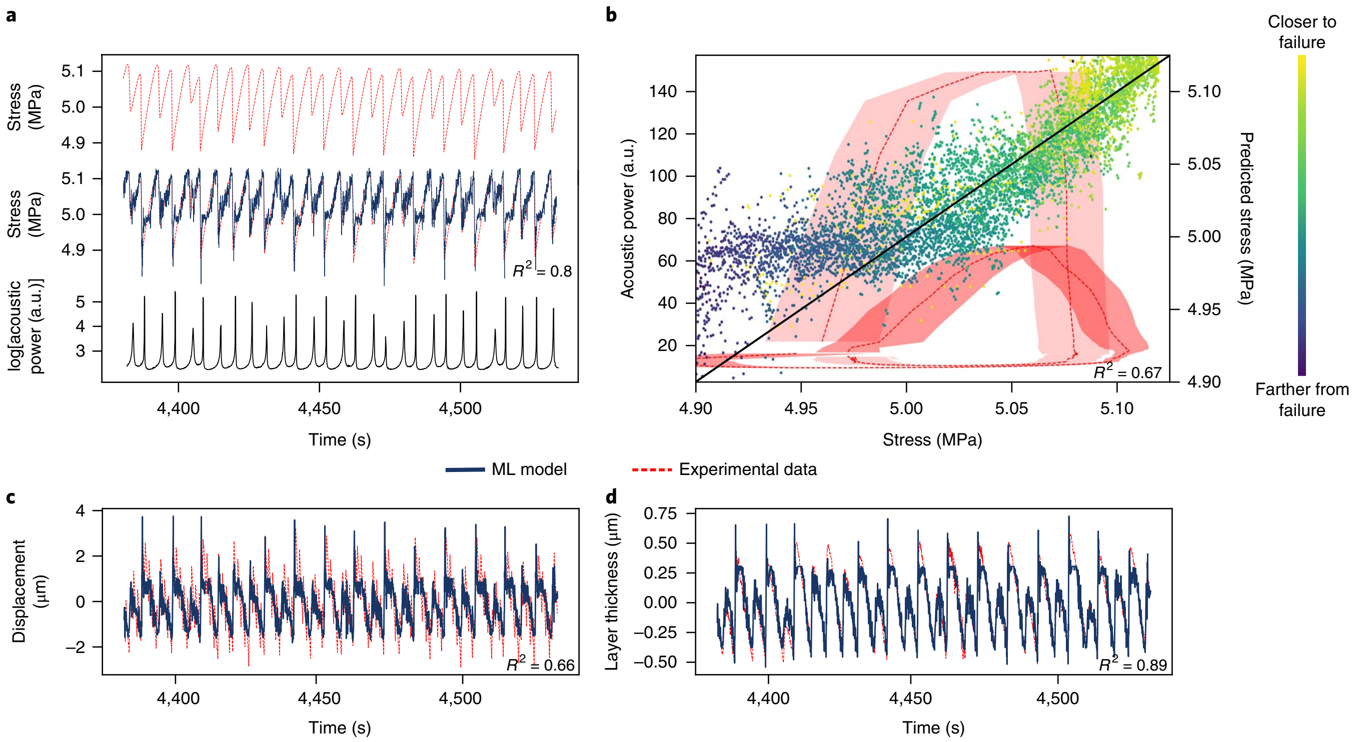

Fig. 2 | Detail of aperiodic stick–slip events showing alternating fast and slow slip. a, Shear stress versus time (top), measured shear stress and machine learning (ML) estimates versus time (middle) and acoustic power versus time (bottom). **b**, Scatter plot showing machine learning estimations of shear stress (right-hand *y* axis) using only acoustic power, as a function of measured shear stress. A perfect model would follow the black line. Red dashed lines show the relation of the acoustic power (left-hand *y* axis) to the average shear stress; the shaded envelopes represent 0.5 s.d. The acoustic power is systematically lower for smaller, slower failure events (darker red) than for larger stress failures, explaining the origin of the two curves. **c**,**d**, Machine learning estimations of fault displacement (**c**) and layer thickness (**d**), using all the features.

machine learning estimations (blue line) are shown at the same scale directly below, plotted over the shear stress data (red dashed line) for easy comparison. The machine learning estimations of stress derived from the seismic signal are generally noisier than the stress data themselves and overestimate stress in the early stages of the seismic cycle, immediately after failure, but match the stress well during both small and large events (Fig. 2a). Note that the machine learning model improves as the time to failure approaches.

Because we rely on an explicit machine learning algorithm (the algorithm makes explicit decisions on the basis of the values of the features), we can probe models to identify the most important features (see Methods). The full machine learning model estimates are more accurate, but in an effort to illuminate the underlying mechanics without the complexity of algorithmic details, we also discuss machine learning results obtained using only these best features. For estimates of shear stress (Fig. 2a) and fault zone strain (Fig. 2c,d), the best feature by far corresponds to the variance of the acoustic signal within a time window—that is, the acoustic power, consistent with our previous analyses of stick–slip in glass bead layers[16,17]. It is straightforward to rebuild a machine learning model from this single feature.

Figure 2b shows the relation between acoustic power and shear stress. The red dashed line is the mean shear stress over a cycle for the series of events shown in Fig. 2a, with the shaded region corresponding to 0.5 s.d. in stress. The alternating sets of slow and fast laboratory earthquakes define distinct loop patterns in this acoustic power versus stress space. The peaks in acoustic power occur roughly half-way through the failure events and correspond to the maximum fault slip velocity, followed by deceleration and the end of a laboratory earthquake. Note that the acoustic power is systematically lower for slower events, as indicated by the smaller loop. Faster events correspond to the larger loops.

Continuous seismic waves can also be used to determine fault slip and volumetric strain during the laboratory seismic cycle (Fig. 2c,d). Here, we show machine learning estimates (using all the features) of the detrended fault displacement history and layer thickness as a function of time for the failure events of Fig. 2a. In each case, we show relative changes in position, with positive values indicating slip in the shear direction (Fig. 2c) and layer compaction (Fig. 2d), respectively. The fault zone dilates during the loading portion of the stick–slip cycle and compacts during failure (Fig. 2d). For both measurements, the machine learning model matches the data well, although the displacements are small.

These results indicate that the continuous measurement of elastic energy emanating from the fault zone is imprinted with precise information regarding the current state of the fault. At any time during the laboratory seismic cycle, the acoustic signal can be used to estimate the stress state (Fig. 2a,b) and the fault displacement history (Fig. 2c,d).

Early signals contain a signature of the impending slip mode

Machine learning identifies specific patterns in the time series of acoustic power that enable precise estimates of the fault zone stress state and strain history during cycles of loading and failure. We further probe this connection by looking at the correlation between acoustic power early in the seismic cycle and the character of the impending failure event. Plotting the acoustic power during the first 10% of the seismic cycle versus the duration of the impending event shows a robust correlation (Fig. 3b). The acoustic signal generated

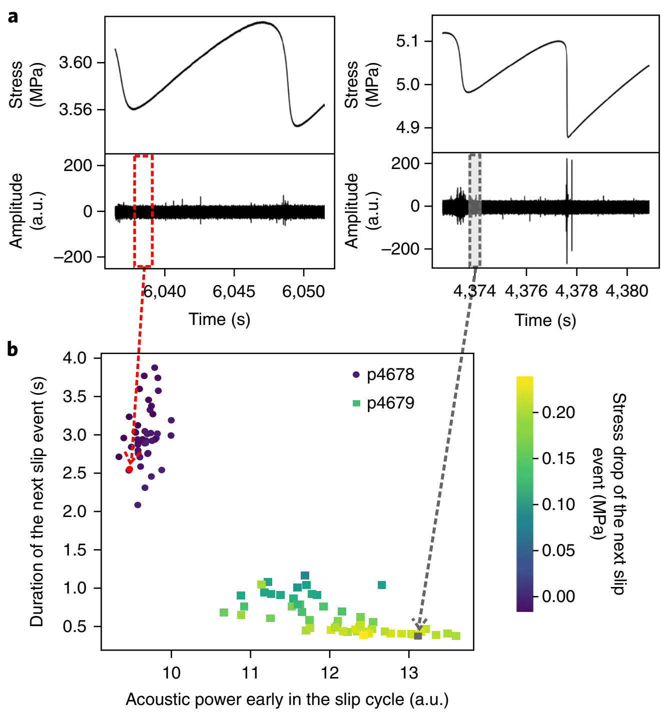

Fig. 3 | Acoustic signature foretells failure mode for laboratory events. **a**, Seismic cycles of slow (left) and fast (right) slip events. Top, shear stress. Bottom, elastic wave energy. Boxed regions show the period used for the measurement of acoustic power. **b**, Relation between acoustic power early in the seismic cycle versus slip duration and stress drop. Note that the fault zone emits more acoustic energy before fast events with larger stress drop compared with slow events with smaller stress drop. Even early in the seismic cycle, the acoustic signal contains a predictive signature of the impending failure magnitude.

in the early stages of what will be larger, faster laboratory earthquakes is systematically larger than that for slow slip events with smaller stress drop (Fig. 3b). If the stress cycle starts at a high (low) acoustic power, the cycle is likely to lead to a fast (slow) earthquake. The correlation holds for a broad spectrum of intermediate slip durations, providing supporting evidence for a continuum between fast and slow earthquakes that is associated with similar underlying physical processes. These initial values seem to be connected to the stress drop and energy release of the previous slip event (see Supplementary Information). The evolution of acoustic power over the seismic cycle exhibits a memory effect, such that fast events (with large stress drop and energy release) are followed by cycles that start with low acoustic power, whereas cycles following a slow event start at higher acoustic power (Fig. 2a). We hypothesize that high seismic energy early in the seismic cycle reflects a more locked and potentially unstable fault zone structure, with stronger asperity junction contacts and a granular texture, that is more likely to fail abruptly than other configurations.

In short, the fault emits characteristic seismic signals that tell us early on whether the system is heading toward a slow slip event or a fast earthquake. This signature exists for a broad range of slip behaviours, supporting the existence of a continuum between slip modes from fast failure to slow slip events.

Predicting failure time and magnitude

Previous work shows that the time to failure can be estimated from the continuous acoustic emission generated in a sheared layer of glass beads[16]. Here, we show that failure timing as well as the duration and magnitude of the complete spectrum of laboratory earthquakes, from slow to fast, can be predicted from machine learning using the elastic waves generated within the fault zone. Figure 4a shows data for a series of fast and slow laboratory earthquakes with bimodal stress drops. Also plotted are predictions of the time remaining before the next failure for this set of events, which include the full range of aperiodic slow slip events. The machine learning predictions of laboratory earthquake failure times are highly accurate, with $R^2 = 0.88$. Note that the model correctly predicts the time remaining before failure for both small and large stress drop events, and predicts well the two small outliers (at approximately 4,408 s and 4,415 s). Here again, each point of the prediction curve is derived from only a single small window of continuous seismic data, using the machine learning algorithm. The machine learning algorithm identifies acoustic power as the most important feature for predicting laboratory earthquakes (Fig. 4) and for estimating the fault stress, shear displacement and layer thickness (Fig. 2).

The machine learning predictions of failure event durations distinguish long- from short-duration events and thus slow versus fast laboratory earthquakes. Because the energy of the seismic signals contains quantitative information regarding the frictional stress at all times, the machine learning model is able to determine the fault's timing in the earthquake cycle. This provides the means to infer failure times and event durations long before the slip occurs (Figs. 3b and 4b).

We quantify these relations by plotting the measured versus predicted start of failure and end of failure for our entire suite of laboratory earthquakes (Fig. 4c). In addition, the measured versus predicted laboratory earthquake durations are shown in Fig. 4c, which demonstrates that the machine learning approach can predict the timing and duration of a broad range of laboratory earthquakes.

The accuracy of the failure time predictions also enables us to estimate future laboratory earthquake 'magnitudes' (Fig. 4d). We use the predicted inter-event times, the predicted durations of slip events, the acoustic power and the magnitudes of the preceding slip as features for a further machine learning analysis. Event amplitude prediction is more difficult than the failure time prediction, as the associated database is much smaller than the database built from scanning the continuous seismic signal (a few dozen versus several thousand data points when scanning the seismic signal—see Supplementary Information for more details).

We constructed 50 different machine learning models for both experiments; the average R^2 for the first experiment was 0.73 (for about 50 data points), and for the second experiment 0.42 (for about 30 data points). The predicted slip inter-event time and the predicted slip duration (as we may expect) seem to be the most important variables to make these predictions. For slow, long duration events, the energy release is lower, consistent with lower stress drop (Fig. 1b). As applied in nature, the predictions of acoustic energy release during laboratory earthquakes would correspond to predictions of maximum ground velocity associated with the passage of seismic events. Our data are consistent with observations showing that more highly stressed faults release greater seismic energy early in the seismic cycle, and it is interesting to speculate whether a similar correlation exists for tectonic faults.

Discussion and application to tectonic faulting

We study elastic radiation from laboratory fault zones and show that stick–slip failure can be predicted for a broad range of slip modes corresponding to the spectrum of tectonic faulting. For the laboratory seismic cycle, both slow and fast earthquakes are preceded by a cascade of micro-failure events that radiate elastic energy with a signature that foretells catastrophic failure. We show that acoustic emission signals generated from quartz fault gouge can be used to predict the timing, duration and magnitude of laboratory

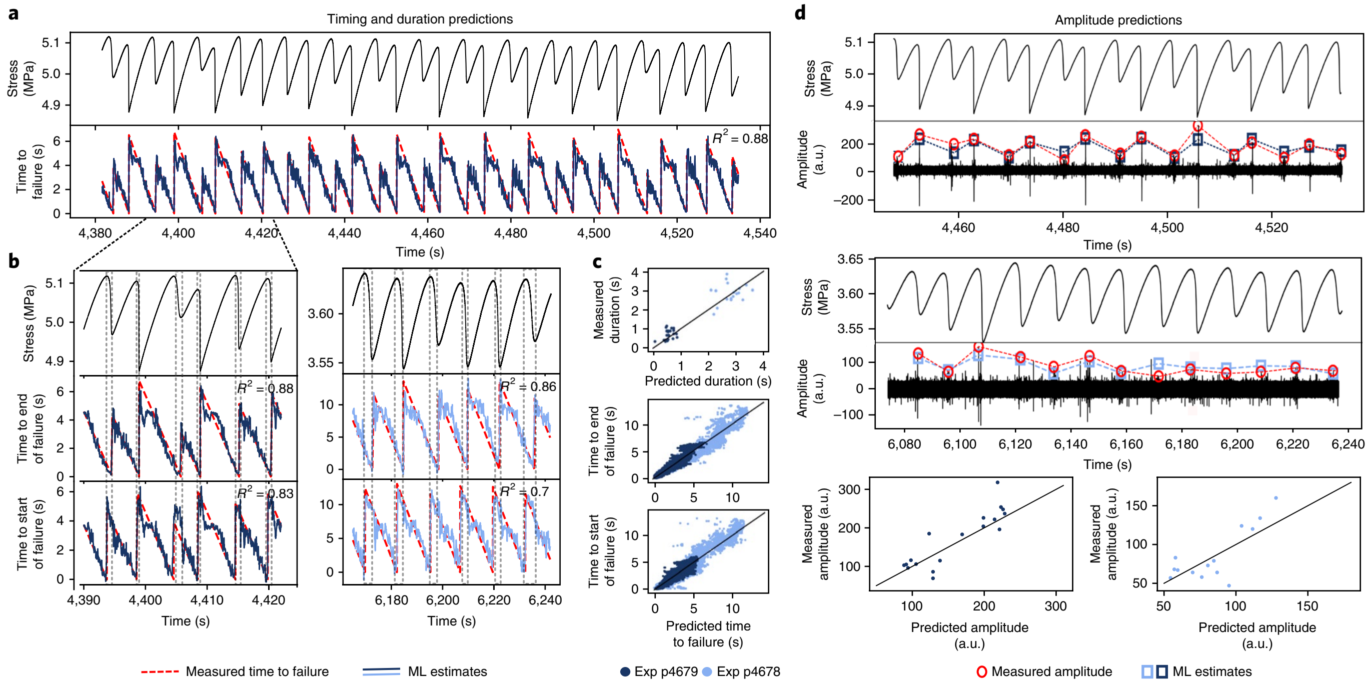

Fig. 4 | Laboratory earthquake prediction on testing set. a, A sequence of slip cycles (top) and the time to failure (bottom) for the events (red line) and associated predictions (blue line). **b**, Experiments with faster (left) and slower (right) laboratory earthquakes. In each set, the upper plot is shear stress, and the plots below show measured time to failure (red), with machine learning predictions (blue) of the end time (middle plot) and start time (lower plot) of the failure events. Grey bars delimit the predicted duration of each event. **c**, From top to bottom: predicted slip duration versus measured values, predicted times remaining before the end of the next slow slip and predicted times remaining before the beginning of the next slow slip. **d**, Amplitude predictions for p4679 (top) and p4678 (middle). Bottom plots show measured versus predicted amplitudes. Perfect predictions in **c** and **d** would follow the black diagonal line.

earthquakes. The laboratory slow earthquakes reach peak slip velocities of 10–100 μm s^{-1}, consistent with tectonic slow slip.

Acoustic signals generated before failure, early in the laboratory seismic cycle, are systematically different for larger, faster laboratory earthquakes than for slow slip events with smaller stress drop. Our work shows that a spectrum of frictional failure modes share common mechanisms and can be predicted from elastic energy emanating from the fault zone before failure.

Our recent work on episodic slow slip and tremor in Cascadia suggests that similar signals may also occur in Earth[26]. This analysis shows that the surface displacement of the Cascadia subduction zone can be estimated from continuous seismic data, in a way very similar to the displacement analysis presented here. Challenges for the application of our methodology to real data include dealing with much noisier data, and signals potentially coming from multiple faults. Future work will test whether catastrophic earthquake failure in Earth is also preceded by similar signals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41561-018-0272-8.

Received: 17 January 2018; Accepted: 12 November 2018;
Published online: 17 December 2018

References

1. Ide, S., Beroza, G. C., Shelly, D. R. & Uchide, T. A scaling law for slow earthquakes. *Nature* **447**, 73–76 (2007).
2. Peng, Z. & Gomberg, J. An integrated perspective of the continuum between earthquakes and slow-slip phenomena. *Nat. Geosci.* **3**, 599–607 (2010).
3. Rowe, C. D. & Griffith, W. A. Do faults preserve a record of seismic slip: a second opinion. *J. Struct. Geol.* **78**, 1–26 (2015).
4. Shelly, D. R. Complexity of the deep San Andreas fault zone defined by cascading tremor. *Nat. Geosci.* **8**, 145–151 (2015).
5. Ide, S. Characteristics of slow earthquakes in the very low frequency band: application to the Cascadia subduction zone. *J. Geophys. Res. Solid Earth* **121**, 5942–5952 (2016).
6. Wallace, L. M. et al. Large-scale dynamic triggering of shallow slow slip enhanced by overlying sedimentary wedge. *Nat. Geosci.* **10**, 765–770 (2017).
7. Frank, W. B., Rousset, B., Lasserre, C. & Campillo, M. Revealing the cluster of slow transients behind a large slow slip event. *Sci. Adv.* **4** (2018).
8. Brace, W. F. & Byerlee, J. D. Stick–slip as a mechanism for earthquakes. *Science* **153**, 990–992 (1966).
9. Scholz, C. H. *The Mechanics of Earthquakes and Faulting* (Cambridge Univ. Press, Cambridge, 2002).
10. Veedu, D. M. & Barbot, S. The Parkfield tremors reveal slow and fast ruptures on the same asperity. *Nature* **532**, 361–365 (2016).
11. Obara, K. & Kato, A. Connecting slow earthquakes to huge earthquakes. *Science* **353**, 253–257 (2016).
12. Radiguet, M. et al. Triggering of the 2014 M_w7.3 Papanoa earthquake by a slow slip event in Guerrero, Mexico. *Nat. Geosci.* **9**, 829–833 (2016).
13. Svetlizky, I., Bayart, E., Cohen, G. & Fineberg, J. Frictional resistance within the wake of frictional rupture fronts. *Phys. Rev. Lett.* **118**, 234301 (2017).
14. Kato, A. et al. Propagation of slow slip leading up to the 2011 M_w9.0 Tohoku-Oki earthquake. *Science* **335**, 705–708 (2012).
15. Gomberg, J., Wech, A., Creager, K., Obara, K. & Agnew, D. Reconsidering earthquake scaling. *Geophys. Res. Lett.* **43**, 6243–6251 (2016).
16. Rouet-Leduc, B. et al. Machine learning predicts laboratory earthquakes. *Geophys. Res. Lett.* **44**, 9276–9282 (2017).
17. Rouet-Leduc, B. et al. Estimating fault friction from seismic signals in the laboratory. *Geophys. Res. Lett.* **45**, 1321–1329 (2018).
18. Marone, C. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.* **26**, 643–696 (1998).
19. Johnson, P. A. et al. Acoustic emission and microslip precursors to stick–slip failure in sheared granular material. *Geophys. Res. Lett.* **40**, 5627–5631 (2013).
20. Kaproth, B. M. & Marone, C. Slow earthquakes, preseismic velocity changes, and the origin of slow frictional stick–slip. *Science* **341**, 1229–1232 (2013).

21. Leeman, J., Saffer, D., Scuderi, M. & Marone, C. Laboratory observations of slow earthquakes and the spectrum of tectonic fault slip modes. *Nat. Commun.* **7**, 11104 (2016).
22. Scuderi, M., Marone, C., Tinti, E., Di Stefano, G. & Collettini, C. Precursory changes in seismic velocity for the spectrum of earthquake failure modes. *Nat. Geosci.* **9**, 695–700 (2016).
23. Leeman, J. R., Marone, C. & Saffer, D. M. Frictional mechanics of slow earthquakes. *J. Geophys. Res. Solid Earth* **123**, 7931–7949 (2018).
24. Rivière, J., Lv, Z., Johnson, P. & Marone, C. Evolution of *b*-value during the seismic cycle: insights from laboratory experiments on simulated faults. *Earth Planet. Sci. Lett.* **482**, 407–413 (2018).
25. Friedman, J. et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **28**, 337–407 (2000).
26. Rouet-Leduc, B., Hulbert, C. & Johnson, P. A. Constant chatter of the Cascadia megathrust revealed by machine learning. *Nat. Geosci.* https://doi.org/10.1038/s41561-018-0274-6 (2018).

Acknowledgements

We thank Institutional Support (LDRD) and DOE Fossil Energy for funding the work at Los Alamos, and the National Science Foundation and the LANL-CSES program for funding the work at Penn State. We thank J. Gomberg, A. Delorey, I. McBrearty, R. Guyer, C. Lee and J. Leeman for discussions and comments.

Author contributions

C.H., B.R.-L. and C.X.R. conducted the machine learning analysis. J.R., D.C.B., P.A.J. and C.M. conducted the experiments. P.A.J. and C.M. supervised the project. C.H., C.M. and P.A.J. wrote the manuscript along with all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41561-018-0272-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Experimental apparatus. We use a biaxial deformation machine to shear fault zones at constant normal stress[27]. The DDS configuration consists of two granular layers (each layer is constructed to be precisely 3 mm thick, with nominal contact area of 10 cm × 10 cm) sandwiched in a three-block assembly. We use quartz powder (MIN-U-SIL-40, US Silica) to simulate granular fault gouge and rough, steel forcing blocks. A block of poly(methylmethacrylate) is placed in series with the central block of the DDS configurations to match the fault system stiffness with the critical frictional weakening rate, so that $K/K_c \approx 1$. The central block is 15 cm high so that the frictional area remains constant throughout shear. The steel forcing blocks of the DDS assembly have 0.8-mm-deep and 1-mm-wide teeth perpendicular to the shear direction, to prevent sliding along the bounding surface. Cellophane tape is used to contain the layer during sample construction. A thin rubber sleeve is placed at the base of the DDS assembly to contain the layer and minimize loss of material during geometric spreading as an experiment proceeds[28]. Normal stress is applied to the nominal frictional area and maintained by servohydraulic control, while the central forcing block is driven at a constant shear velocity. Horizontal and vertical forces are measured using strain gauge load cells with resolution of ±10 N, while the corresponding displacements are measured with direct-current displacement transducers with ±0.1 μm precision. Elastic waves are recorded using broadband piezoelectric sensors (~0.02–2 MHz frequency bandwidth) embedded in each side forcing block of the DDS configuration. The piezoceramic sensors (PZTs) are close to the gouge layers (~2 mm). Forces and displacements are continuously recorded at 1 kHz, while seismic activity is recorded at 4 MHz. Important data sets for our use here include the continuous seismic data, the shear stress, the normal stress and the shear and normal components of the layer strain. The bulk friction is obtained from shear and normal stress. The shear stress signal is used in the training procedure of the seismic signal to mark where a slip event occurs. During the machine learning testing procedure, the algorithm sees only acoustical data that it has not seen before.

We focus on two experiments, p4678 and p4679, and a series of supporting runs that are closely related to an extensive set of tests reported earlier[21]. We study a broad range of conditions using many stick–slip events.

Conditions for p4678 were: normal stress: 5 MPa; average shear stress: 3.6 MPa; average inter-event time: 12.1 s.

Conditions for p4679 were: normal stress: 7 MPa; average shear stress: 5 MPa; average inter-event time: 5.3 s. This experiment exhibits a range of stick–slip behaviours including dual behaviour, with alternating slower (smaller stress magnitude of the stress drop) and faster (larger) events. Figure 1 shows steady-state frictional shear transitions to stick–slip sliding at shear displacement of 13.8 mm (shear strain of 5.7). Spikes in stress at 23 mm and 46 mm are transient strengthening caused by brief pauses in shear loading. The shear loading rate is 10 μm s^{-1} until 37 mm and then drops to 5 μm s^{-1} and 3 μm s^{-1}; note the corresponding increase in stress drop magnitude for lower velocity. In general, higher normal stress and lower driving velocity produces failure events with larger stress drop and shorter duration (Fig. 1b). For conditions near the stability boundary defined by $K/K_c = 1$, we observe aperiodic failure and dual behaviour with alternating smaller and larger events (Fig. 2). Fault slip velocity during failure scales with stress drop, and thus the dual behaviour: sequences represent alternating faster and slower events. We chose to focus on the acoustic emission data from these events because they correspond to the most complex conditions, with aperiodic stick–slip cycles and a range of stress drop magnitudes and slip modes.

Step-by-step description of data analysis and machine learning methods.

(1) Preparing the data for machine learning analysis. This first step consists of transforming the data into a format easy to compute features on. The raw experimental data are in a binary format, with acoustic data and mechanical data (stress, layer thickness, displacement) separated. Therefore, we begin by creating files of acoustic data and mechanical data with a common time base. As the acoustic and mechanical data have different sampling rates, we create files of 1.33 s duration with both acoustic and mechanical data.

(2) Defining the features of the acoustic data. This step is critical, as it extracts the characteristics (features) of the acoustic data that the machine learning model will use. Here, each file of acoustic data is scanned through moving time windows. The length of one time window corresponds to ~5% of the laboratory seismic cycle (0.3 s for p4679, 0.5 s for p4678). These time windows are further subdivided into two non-overlapping contiguous subwindows (each subwindow therefore represents ~2.5% of the average slip cycle). We chose to use two subwindows because many features are symmetric in the slip cycle. With the information from two subwindows, the algorithm is able to differentiate between the loading and the slipping part of the cycle. We compute machine learning features for each of these subwindows. These time windows thus become a list of statistical features that describe the data in a condensed manner (see Supplementary Information for a detailed list of features).

(3) Preparing the labels. The machine learning model will be tasked to find a mapping between characteristics of the acoustic data and a bulk physical property of the fault (stress, displacement, time to failure, which we refer to as labels). In this step, we prepare the time series of these bulk physical properties such that each time window of acoustic data can be labelled with the corresponding bulk physical properties. Each time window of acoustic data is thus labelled with the average shear stress (or layer thickness, displacement) during the same time window as the file of acoustic data. Alternatively, the acoustic files can be labelled with the time remaining before the next slip event: we use the PeakUtils Python library on the shear stress data to pick the failure times, and we label each time window of acoustic data with the time remaining before the next pick.

(4) Creation of a machine-learning-friendly database. This step puts the data in a format that will actually be used by the models. Each line of the database contains the features describing each time window of acoustic data, and the corresponding average stress (or other label) during the same time. Each line i is therefore a list $\{x^i_{1,1}, x^i_{1,2}, \ldots, x^i_{1,D}, x^i_{2,1}, x^i_{2,2}, \ldots, x^i_{2,D}, y^i\}$, with $x^i_{1,j}$ the jth feature of the first half (first subwindow) of the ith time window of acoustic data, $x^i_{2,j}$ the jth feature of the second half of the ith time window of acoustic data, D the total number of features and y^i the average stress (or other label) during this time window.

(5) Overlapping the time windows. Each time window overlaps by 90% with the previous time window. Thus, we scan the data in increments corresponding to 10% of the window size, while building features and the corresponding labels, and then add the resulting list of labelled features to the database at each time increment.

(6) Train–test split. The database built as described above is a time series of features of the acoustic data and a corresponding average bulk physical property of the system. As such, the train–test split must be of two contiguous pieces, owing to the autocorrelation of the system: one line of our database (the features of a time window of acoustic data and corresponding label) is similar to the following window, especially considering that the windows overlap. Therefore, a random train–test split is not appropriate at all. For each of the two experiments, we use the first contiguous half of the data for training and the second contiguous half of the data for testing.

(7) Tuning the hyperparameters of the model. Before actually creating a model relating features of the acoustic data to the corresponding bulk property of the fault, we have to determine the space of functions that will be explored as possible models. This is done by tuning the hyperparameters of our model. These hyperparameters control how vast the explored function space will be during training, typically by setting how smooth the explored functions are. In the case of the gradient boosted trees models used here, the hyperparameters determine the power of expression of the trees that constitute the model (see 'Hyperparameter optimization' in the Supplementary Information for details). The hyperparameters of the model are determined to maximize the performance of trial models in cross-validation: a set of hyperparameters is used to model a subset of the training data and evaluated on its performance on the rest of the training data.

(8) Model training. With a training database at hand and the complexity of our model determined (the hyperparameters), we can train our final model. Model training consists of obtaining best fits of the training data, given the complexity of the model. The previous step of optimizing this complexity on subsets of the training data ensures that this best model will not overfit the data. Indeed, too complex a model (prone to overfitting) would do poorly in cross-validation. Our final model is an ensemble of decision trees, formed from a series of yes/no decisions based on a given list of features, to arrive at a modelled (predicted) label. We give more details on decision trees and our particular implementation, gradient boosted trees (XGBoost), in the Supplementary Information.

(9) Assessing our final model. The performance of the final model is assessed using the testing set, the second half of the data. The metric used here, R^2, compares the squared errors of the model with the squared errors of the null model, a model that always predicts the average value of the label. The same metric is used during the cross-validation step above, and provided the data are similar in training and testing (they follow the same distribution), the performance in cross-validation is a good proxy for the performance in testing. A different model is made for each bulk physical property (shear stress, time to failure and so on).

(10) Feature importance. Once we have an accurate final model, we can look for the best features identified by our model, to try to understand how the model reached its estimations.

Code availability. We are unable to make the computer code associated with this paper available at this point, but we aim to make it available in the future. Please contact C.H. for details.

Data availability

The data are available from the Penn State Rock Mechanics laboratory (www3.geosc.psu.edu/~cjm38/).

References

27. Karner, S. L. & Marone, C. The effect of shear load on frictional healing in simulated fault gouge. *Geophys. Res. Lett.* **25**, 4561–4564 (1998).

28. Scott, D. R., Marone, C. J. & Sammis, C. G. The apparent friction of granular fault gouge in sheared layers. *J. Geophys. Res. Solid Earth* **99**, 7231–7246 (1994).

nature geoscience

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41561-018-0272-8

In the format provided by the authors and unedited.

Similarity of fast and slow earthquakes illuminated by machine learning

Claudia Hulbert[1]*, Bertrand Rouet-Leduc[1], Paul A. Johnson[1], Christopher X. Ren[1], Jacques Rivière[2], David C. Bolton[3] and Chris Marone[3]

[1]Geophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA. [2]Department of Engineering Science and Mechanics, Pennsylvania State University, University Park, PA, USA. [3]Department of Geosciences, Pennsylvania State University, University Park, PA, USA. *e-mail: chulbert@lanl.gov

Supplementary Information for
Machine Learning Predictions Illuminate Similarity of Fast and Slow Laboratory Earthquakes

Claudia Hulbert[1], Bertrand Rouet-Leduc[1], Paul A. Johnson[1],
Christopher X. Ren[1], Jacques Rivière[2], David C. Bolton[3], Chris Marone[3]
[1]Los Alamos National Laboratory, Geophysics Group, Los Alamos, New Mexico, USA
[2]Department of Engineering Science and Mechanics, Pennsylvania State University,
University Park, PA 16802, USA
[3]Department of Geosciences, Pennsylvania State University, University Park, Pennsylvania, USA

This PDF file includes:

Supplementary Information

Figs. S1 to S5

Supplementary Information

Signal and noise in experimental apparatus

We studied the influence of noise as part of an effort to 1) verify that our signals were not contaminated by noise associated with the acoustic sensors or testing machine, and 2) to assess the extent to which the lab results might extend to tectonic faults with additional complexity and noise. We note that the lab data contain several of the same types of noise sources encountered in Earth. That is, electrical noise in the analog signals, noise in the analog to digital conversion, and noise associated with fluctuations in control conditions (stress, shear rate) and small AE events. In addition, we tested the hypothesis that our measurements are contaminated by noise in the acoustic sensors and/or noise that derives from the testing machine by comparing the signals for 1) active shearing under full servo control (our typical conditions during the experiment), 2) when the machine is running and normal stress is applied but shearing is stopped, and 3) when the machine is running but both loading rams are locked, so there is neither servo control on the normal stress nor motion of the ram that applies shearing. We found that machine-related noise, from cases 2 and 3 does not have predictive ability to estimate friction nor layer thickness, and thus we rejected this hypothesis.

Feature construction

Feature extraction follows the description given above and in previous work,[1] in which a moving time window is used to analyze the continuous seismic data. We rely on features with useful physical meaning to make our predictions. The features can be separated into three main categories:

- signal energy: we use several higher order moments of the acoustic data to capture the evolution of the signal's energy. Within each time window we compute the signal 1-4 moments, as well as the variance, skewness and kurtosis (centered moments).

- precursors: during slow slip and fast lab earthquakes, the system enters a critical state when close to failure, where fault zones emit relatively strong acoustic emissions, compared to earlier in the lab seismic cycle. We rely on thresholds to capture precursory activity before each laboratory event.

• Fourier analysis: we build several features based on Fourier analysis in order to analyze the frequency content of the signal (Figure S4). These features correspond to integrals of the power spectrum over different frequency bands.

Feature selection: once the features are constructed for ML, we perform feature selection in order to select the most promising variables. This is useful in particular when there are several acoustic sensors and therefore several channels of acoustic data to analyze in this case the total number of features can become very large. We use recursive feature elimination (RFE), based on another decision tree approach, to select a sub-set of features that we use as input to the ML analysis. RFE has been shown to be particularly robust in presence of correlated predictors.[2]

Hyperparameter optimization:

Hyperparameters determine the parameter space that will be explored as possible models, *i.e.* they control how vast the explored function space will be during training. For instance, a model with large tree depth, able to split the data into many partitions, will be more complex than a model with small tree depth. The choice of the complexity of the model is crucial for our final performance: if it is too complex, we may overfit the data and end up with a poor performance in testing; a model too simple may also perform poorly. However, we do not know beforehand which complexity is good for a given dataset. We need to evaluate this by building and comparing many different models during the training phase. Models are compared by cross-validation: they are iteratively built on subsets of the training data, and evaluated on the remaining training data. They are then compared, the best model is identified, and its hyperparameters are selected for the final model.

XGBoost is characterized by a large number of hyper-parameters to tune, and the performance of the models is heavily dependent on these hyperparameters. Therefore we rely on an optimization procedure to tune these hyper-parameters. More specifically, we rely on the *gp_minimize* function from textitskopt to do so, that implements the EGO Bayesian optimization procedure.[3]

The algorithm proceeds as following. The main idea is to approximate the function textitf that relates the value of the hyper-parameters to the performance of the model. This function is approximated through Gaussian processes (GP) - i.e. by choosing a Gaussian process prior. At first, several points are drawn randomly: hyper-parameters are drawn at random, the corresponding XGBoost regressions are built, and their associated performance is saved. Performance is measured by 5-fold cross-validation on the training set. Once these random points are built, they serve as initial database, and a first GP model is created that relates the hyper-parameters to the performance of the regression.

The choice of the following hyperparameters to try is guided by this GP model in an intelligent fashion. With x_t^+ the best point obtained so far (associated to the highest performance), the choice of the next candidate point x to try is given by maximizing the expected improvement in performance EI:

$$x_{t+1} = \mathrm{argmax_x EI(x)}$$

with EI the expected increase in performance:

$$EI(x) = E[f(x) - f(x_t^+)]$$

Once a new datapoint is tried in this way, it is added to the database and improves the approximation of the function textitf. The procedure continues until a maximum number of iterations is reached; the best point, and therefore the best hyper-parameters found, is returned. Then these hyperparameters can be used to build a final model on the training set, and evaluate it on the testing set. Note that any optimization procedure could be used to tune the hyper-parameters, but we chose the EGO method because it is efficient in high-dimensions, is not prone to getting stuck in local optima, and requires few iterations to converge. An alternative standard procedure is to to a grid-search, but this is prohibitive for a high number of hyper-parameters. The Bayesian optimization loop optimizes the following hyperparameters of the XGBoost models: max_depth (maximum depth of a tree), learning_rate (step size shrinkage), n_estimators (number of trees), gamma (minimum loss reduction to create new partition), min_child_weight (minimum leaf weight), subsample (ratio to bootstrap dataset), colsample_by_tree (ratio of features to consider), reg_alpha (weight of the l_1 penalization), reg_lambda (weight of the l_2 penalization). Following the hyperparameter optimization, the hyperparameters of each of the models presented in the main text are the following:

- exp. p4679, stress: max_depth:16, learning_rate:0.0610953, n_estimators:365, gamma:0, min_child_weight:2, subsample:0.4384717, colsample_by_tree:0.8308638, reg_alpha:0.1, reg_lambda:0.1.
- exp. p4679, displacement: max_depth:1, learning_rate:0.3814755, n_estimators:24, gamma:0, min_child_weight:3, subsample:0.2613578, colsample_by_tree:0.9119756, reg_alpha:0, reg_lambda:0.
- exp. p4679, layer thickness: max_depth:2, learning_rate:0.1084846, n_estimators:337, gamma:0, min_child_weight:2, subsample:0.57292082, colsample_by_tree:0.6349378, reg_alpha:0.1, reg_lambda:0.05
- exp. p4679, time remaining before the beginning of the next event: max_depth:20, learning_rate:0.0443410, n_estimators:113, gamma:0, min_child_weight:4, subsample:0.3354982, colsample_by_tree:0.6588662, reg_alpha:0.1, reg_lambda:0.1.
- exp. p4679, time remaining before the end of the next event: max_depth:8, learning_rate:0.0804079, n_estimators:268, gamma:0, min_child_weight:4, subsample:0.2265177, colsample_by_tree:0.8962034, reg_alpha:0.05, reg_lambda:0.
- exp. p4678, time remaining before the beginning of the next event: max_depth:14, learning_rate:0.0101759, n_estimators:489, gamma:0, min_child_weight:4, subsample:0.4561674, colsample_by_tree:0.8059474, reg_alpha:0, reg_lambda:0.1.
- exp. p4678, time remaining before the end of the next event: max_depth:15, learning_rate:0.04142359, n_estimators:254, gamma:0, min_child_weight:4, subsample:0.7138402, colsample_by_tree:0.76217941, reg_alpha:0, reg_lambda:0.
- exp. p4679, amplitude predictions: max_depth:13, learning_rate:0.6374797, n_estimators:57, gamma:0, min_child_weight:4, subsample:0.5498936, colsample_by_tree:0.2368331, reg_alpha:67, reg_lambda:62.
- exp. 4678, amplitude predictions: max_depth:14, learning_rate:0.3857817, n_estimators:240, gamma:0, min_child_weight:2, subsample:0.2569217, colsample_by_tree:0.5476946, reg_alpha:5, reg_lambda:3.

Training a model

We tried several machine learning algorithms on this problem. The results that we report here rely on gradient boosted trees (and in particular the XGBoost library[4]), because this approach led to the best performance.

During the training phase, the algorithm has access to both the statistical features derived from the seismic signal and the label (fault friction, displacement rate or time remaining before failure), and attempts to build a model relating the two. Features are exclusively constructed from a small moving window scanning the continuous seismic data.

Once the model is built, it is evaluated in the testing phase, over data that the algorithm has never seen (testing set). In this phase, the algorithm has access only to the statistical features derived from the seismic signal, and never sees the regression label also measured during the experiment (friction, displacement rate, or time remaining before failure). This label is only used to evaluate the performance of the model estimates.

The training phase corresponds to building the structure of our decision trees. The following describes how this structure is found, and how it can be used afterwards in the testing phase.

- Decision trees

Here we give a brief overview of regression trees,[5] as we will rely exclusively on regressions in what follows. A decision tree is built by sequentially, creating nodes that partition the data. To generate each node, the data available at this current node are split into two subsets corresponding to right/left branches. Choosing a branch direction corresponds to determining the feature X_m and the associated threshold c used to partition the data into these two subsets. This corresponds to selecting the split that partitions the data available at the current node j into two subsets that are maximally dissimilar to each other with respect to the label of the regression (here the label is either the time remaining before the next failure, or the magnitude of the next event).

In more simple terms, one decision tree is created on the training set as follows: at first, we consider the whole training dataset, and try to partition it in two by selecting a threshold over one of the features - one threshold corresponds to one decision in the tree. For instance: is the variance higher than 10? This creates a split, *i.e.* two branches in the tree. If the answer is no, go to the left branch of the tree; if yes, go to right branch. This corresponds to the first node of the tree. At this point, each of the two partitions are also divided in two, again by selecting a threshold over one of the features: *e.g.* is the kurtosis higher than 3? The procedure continues iteratively, until a criteria is reached (maximum depth of the tree, minimum number of samples within each of the final partitions, *etc.*). These final partitions are the leaves of the tree. From these, we can get an estimation of the label (friction, displacement, time remaining before failure). In the training set, we have access to the values of the label. A regression tree associates one label value to each of the leaves; for one leaf, this value corresponds to the average of the labels associated to all the datapoints that fall in that leaf. This is the end of the training phase for one tree. Now the structure of the tree is fixed (*i.e.* the final partitions, and their associated label value), and it can be evaluated on the testing set. For this purpose, the testing data is divided in the same partitions, following the same suite of yes/no branch decisions. In the testing data, we do not have the values of the label. But according to which leaves the datapoints fall on, they will be assigned the value associated to that given leaf at the end of the training phase; this corresponds to the model's estimates.

The construction of the tree structure boils down to how to create a split (a decision): at each iteration, which is the best feature to select, and what threshold value of that feature best partitions the data? The criteria used for this purpose is the maximum reduction in (empirical) variance between the data available at the current node, and the two subsets of data partitioned by the split. More specifically, with j the current node of the tree, S_j the labels of the subset of data available at the current node, N_j the number of data points in S_j, $N_{j,\mathrm{L}}$ and $N_{j,\mathrm{R}}$ the number of data points in the left and right subsets $S_{j,\mathrm{L}}$ and $S_{j,\mathrm{R}}$ generated by the split, the criterion for a possible split s is:

$$\Delta\mathrm{Var}(\mathrm{s},\mathrm{j}) = \mathrm{Var}(\mathrm{S_j}) - \frac{\mathrm{N_{j,L}}}{\mathrm{N_j}}\mathrm{Var}(\mathrm{S_{j,L}}) - \frac{\mathrm{N_{j,R}}}{\mathrm{N_j}}\mathrm{Var}(\mathrm{S_{j,R}}) \tag{1}$$

The split selected is the split that maximizes this variance reduction criterion. This criterion ensures that the data within each of the two subsets generated by the split are as homogeneous as possible, while these two subsets are as heterogeneous as possible one from another. Different thresholds are tried for each of the features, based on a histogram. The best feature and the best threshold value are used to create the node.

• Brief overview of gradient boosted trees

One single decision tree is an estimator with high variance - which means that two different trees built from the same data may end up with very different models and performances, and that a few new datapoints may have a large impact on one model. To alleviate this issue, *ensembles of decision trees* are often considered rather than single trees. In ensembles, many trees are built together, and their individual estimations are combined into one single model. By doing so, we obtain a more robust estimator with lower variance. Several ensemble tree methods can be used: random forests, extra trees, gradient boosted trees, *etc.*. For the results presented here, we use gradient boosted trees. Gradient boosted trees rely on many decision trees. The trees are built sequentially. We start by building one single tree as described above; the resulting model is probed, and a new tree is built to minimize the error of this first model. The procedure goes on, and each new tree is built to minimize the errors of all the preceding trees taken together. Once this procedure is finished, the model is a combination of all the estimates of the individual trees, with weights associated to each tree.

More specifically, each new tree is added to the ensemble such that the error is minimized:

$$(h_t, \alpha_t) = \mathrm{argmin}_{h,\alpha} \sum_{i=1}^{n} l\left(y_i, H_{t-1}(x_i) + \alpha h(x_i)\right) = \mathrm{argmin}_{h,\alpha} l(y, H_{t-1} + \alpha h)$$

with h_t the new tree at step t, α_t its coefficient in the ensemble, $H_{t-1} = \sum_{j=1}^{t-1} \alpha_j h_j$ the ensemble at step $t-1$, and l the loss function, the squared error in our case.

We rely on the XGBoost implementation of gradient boosted trees.[4] The performance of this algorithm is very sensitive to the choice of hyper-parameters. Therefore, instead of relying on a simple grid search to set the hyperparameters, we rely on a more sophisticated approach based on Bayesian optimization, described in the paragraph above. We use this method to optimize hyper-parameters based on 5-fold cross-validation.

Assessing the model's performance:

Once a model is built on the training set, and evaluated on the testing set, we need to assess how good these estimates are. At the end of the testing phase, we end up with a series of estimated labels (stress, displacement, time remaining before failure, *etc.*), associated to each of the datapoints that belong to the testing set. We need to compare these estimated labels $\hat{y}_i$ to the true labels y_i, $i = 1, ..., N_{test}$, with N_{test} the size of the testing dataset. For this purpose, we rely on the coefficient of determination, R^2:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where $\bar{y}$ denotes the average of y. A perfect model would have an R^2 of 1. A model that would always predict the average label would have an R^2 of 0. A model with a negative R^2 would perform worse than always predicting the average label. Therefore a model that has and R^2 score above 0 learned something about the data; the closer the score is to 1, the better.

Feature importance:

The use of an ensemble method based on decision trees allows us to report the most important features in our model and therefore enables us to gain physical understanding for this problem - which, besides their higher performance, was one of the primary reasons behind our choice of relying on these particular models. We use the gradient boosted trees F-score to measure the importance of each feature, which represents the number of times a given feature was selected at a tree node by the algorithm. When building a node in a tree, a random subset of features is given to the algorithm. Among this subset of features, the algorithm picks the best possible one to make a split (build a node), by looking at how well this given feature partitions the data. A good feature will result in a large reduction in empirical variance (as described above). Therefore, if the same feature is chosen many times, it means that it is crucial to build the final model. The F score counts how many times each feature was chosen to create a node in the model; this gives a ranking criterion for the importance of the features. The importance of the most predictive features identified by the algorithm to estimate failure times is shown in Figure S1. The evolution of these most important features in time is shown in Figure S2.

Important features:

The features identified by the algorithm as most important for generating these predictions are systematically the same. To show this, we ran 100 iterations of the algorithm on each experiment, selecting hyper-parameters via the EGO procedure using a random seed. This allows us to build 100 different models, each one generating highly accurate predictions. Figure S1 summarizes the 3 strongest features identified by the algorithm for each of these 100 models.

For both experiments, and whether we predict the beginning or the end of the slip event, the strongest features remain the same. In particular, the variance of the acoustic signal during the current window (window N) is by far the best feature identified. Other strong features include the variance in the previous non-overlapping window

(window N-10), and the kurtosis of the signal (windows N and N-10). In two models out of 400, the algorithm also relies on the frequency content of the signal to make its predictions. Because higher order moments are tightly linked to the energy of the signal, the fact that the algorithm mostly relies on the variance and kurtosis for its predictions shows that this energy follows a very precise pattern during the stress cycle. In particular, the variance increases progressively faster as failure approaches, and decreases once the end of the slip is near (c.f., Figures 3 and 4 of the main text). By using the variance in the current window, and in the previous non-overlapping window, the algorithm is able to distinguish between the loading and slipping parts of the cycle. The evolution in time of the strongest features is shown in Figure S2. This very specific pattern is what allows us to make accurate predictions for both the beginning and the end of the slip event.

Estimations of fault displacement and gouge thickness:

When a model of fault displacement and layer thickness is built from only acoustic power (in a way similar to Figure 2 for shear stress in the main text), similar loop patterns emerge (Figure S3). Here again these patterns are different for slow and fast earthquakes; the smaller loops correspond to slow slip events, while the larger ones correspond to fast events. Starting from the end of a large event, the red shaded regions trace out a counter clockwise pattern, with low acoustic power early in the seismic cycle and an increase when failure begins and stress begins to drop (Figure 2 and S3). The first peak in acoustic power begins at higher stress and is smaller in amplitude, corresponding to the smaller stress drop events. This is followed by another cycle, with increasing stress and low acoustic signal power until stress reaches a peak and the acoustic signal strength increases dramatically to define the larger cycle in Figure 2. The loop trajectories for stress, displacement, and layer thickness, are built as follows. First, we cut the time series of acoustic power, such that each piece corresponds to one slip cycle. We then resample these pieces such that they all have the same number of datapoints. It is then straightforward to compute the mean trajectory (and standard deviation) in time for the slip cycles that correspond to fast and slow events. Because the algorithm has access to two sub-windows of acoustic power to make one estimation, it is able to differentiate between the loading and slipping phases of the cycle. Moreover, the variance builds up at a lower pace for slow than for fast slip events, which allows it to distinguish between slow and fast slip cycles.

Magnitude predictions:

Once the time predictions are completed, we build a new database to predict the lab earthquake magnitude. This database only includes one line of data per slip event, and therefore is much smaller than the database used to predict slip timing and duration constructed by scanning the acoustic data (a few tens of data values vs tens of thousands) which makes the analysis harder. The database includes several simple statistics (mean, variance, min, max) of the predicted inter-event slip time, the predicted time duration, and the variance of the acoustic signal, calculated at around half of the stress cycle when the predictions start to become very accurate. It also includes the magnitude of the last slip event. Therefore our magnitude predictions for the next slip event rely on predicted times and acoustic data at around half of the current stress cycle, which allows us to predict magnitude a few seconds in advance (one cycle typically lasts from 3 to 14 seconds depending on the experiment). Specifically, we predict the maximum half-peak amplitude of the absolute value of the acoustic signal for the next event, $A = \max(\text{abs}(ac))$, which can be used in turn to calculate magnitude. These predictions are made roughly at the middle of the current slip cycle. We rely on another gradient boosted trees regression to make the predictions. Figure 6 shows predictions for two sets of slow slip events. The red circles correspond to the true experimental amplitudes and the blue squares show the predicted values. Event magnitude M can be obtained from $M = \log(A)$, with A the maximum half peak amplitude. Because we are measuring the signal adjacent to the fault zone we do not account for the distance from the source. We use the first 70% of data as training set, and the last 30% as testing set. We rely on a similar approach as for time predictions, using reinforcement learning to set the gradient boosted trees' hyperparameters. We do not perform feature selection in this case, as the number of features is already small.

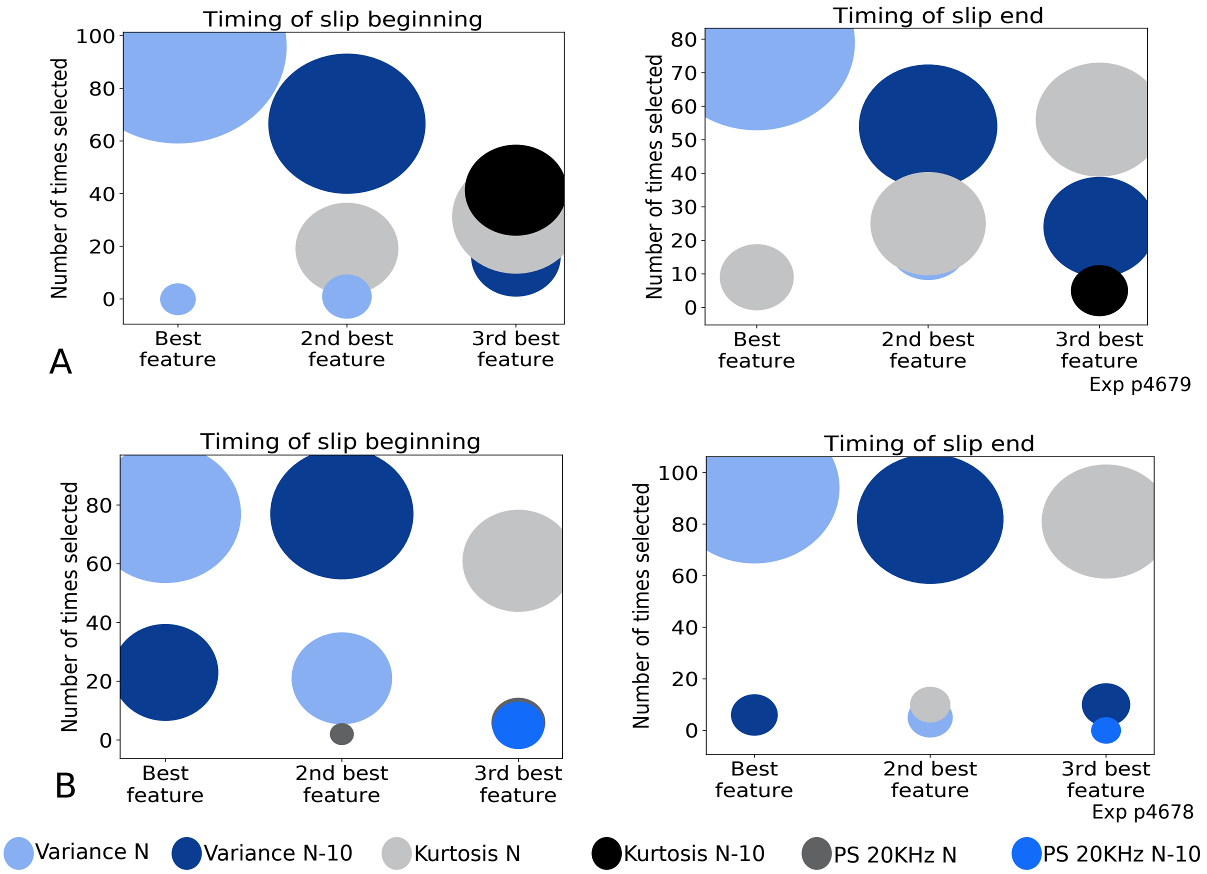

Figure S1. Strongest features for failure timing, identified by building 100 different models. For each plot, the y axis shows the number of models (among the 100) in which a particular feature was identified as the most predictive, 2nd most predictive, and 3rd most predictive. The area of each circle represents the feature importance, as measured by the gradient boosted trees' F-score. Plots on the left show the models built to predict the beginning of the slip event, and plots on the right those built to predict the end of the slip event. Panels (**A**) and (**B**) correspond to the two different experiments. The most important features identified in all of these models remain systematically the same.

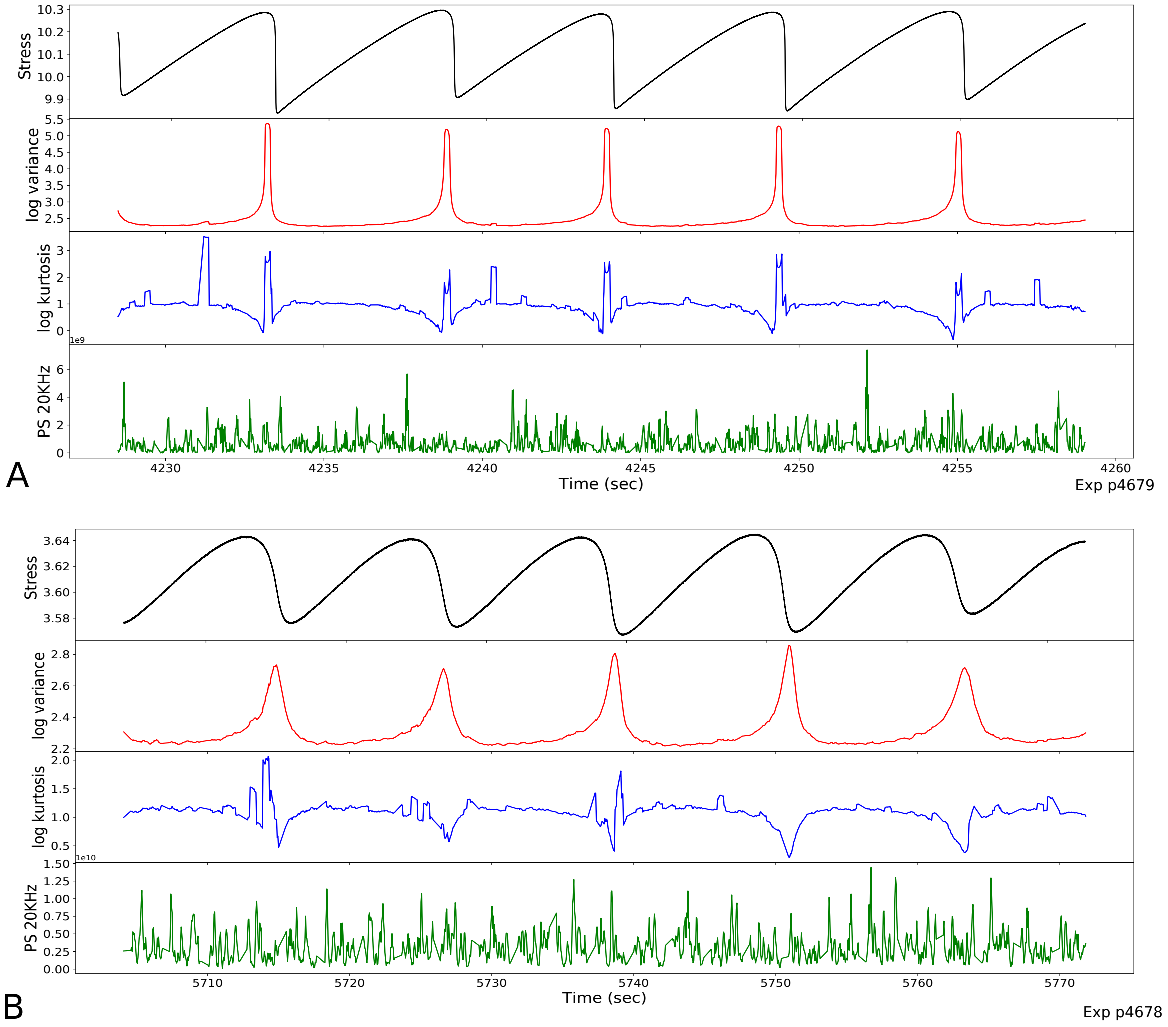

Figure S2. Evolution of the three most predictive features identified by the machine learning algorithm, over five stress cycles. The variance (power) of the acoustic signal increases slowly at the beginning of the stress cycle, takes off at the beginning of the slip event, then decreases precipitously near the end of the slip event. In contrast, the kurtosis increases near the end of the slip event, and decreases as a slip approaches. Peaks in the kurtosis often occur near failure, linked to increased precursory activity. This shows that the energy of the acoustic signal follows a very precise pattern during the stress cycle. Panels (**A**) and (**B**) correspond to the two different experiments.

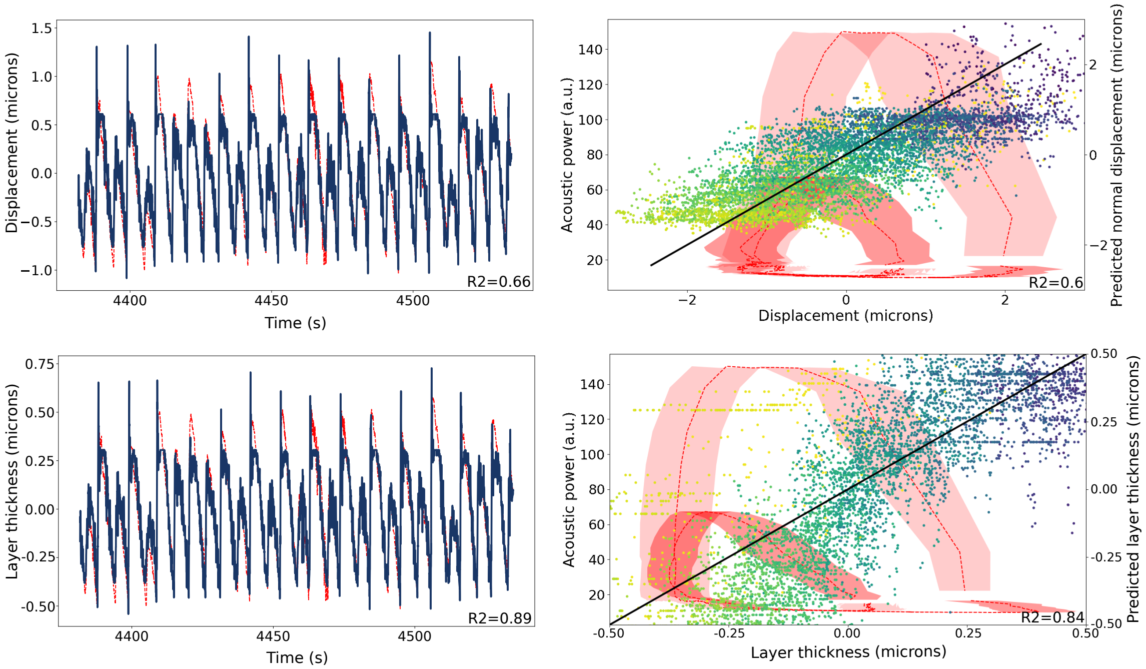

Figure S3. Illustration of the 'loop' patterns followed by the acoustic power when plotted against displacement and layer thickness (respectively A and B, right). Again, we see smaller loops associated to the slow events, and larger loops associated to the fast events. The left hand-side plots show the associated estimations of the models using all features (same as Figure 3 of the manuscript). The scatter plots on the right hand-side (right axis) show the estimations using only the acoustic power. Here too, these estimation become better and better as failure gets closer (scatter plot colors correspond to time remaining before failure: brighter means imminent failure).

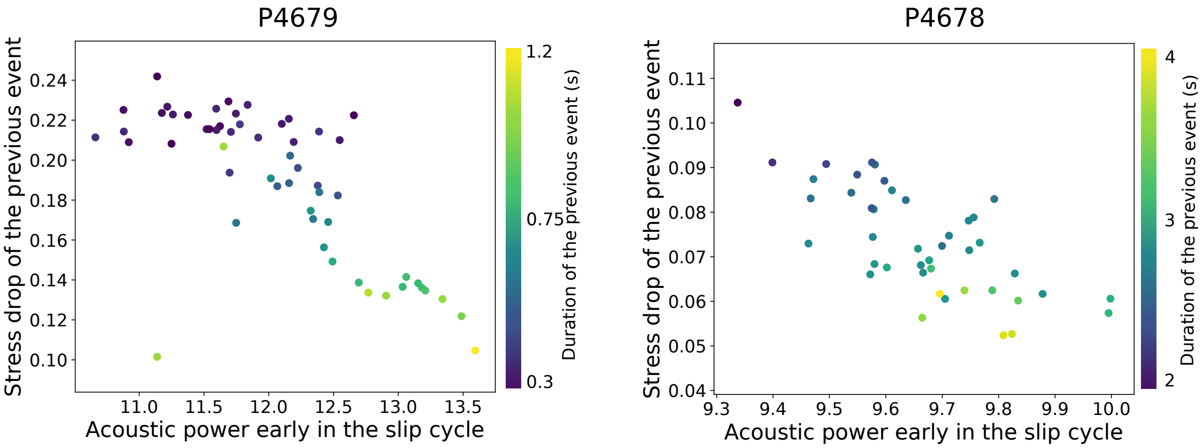

Figure S4. Connection between acoustic power early in the slip cycle, and the stress drop and duration of the previous event.

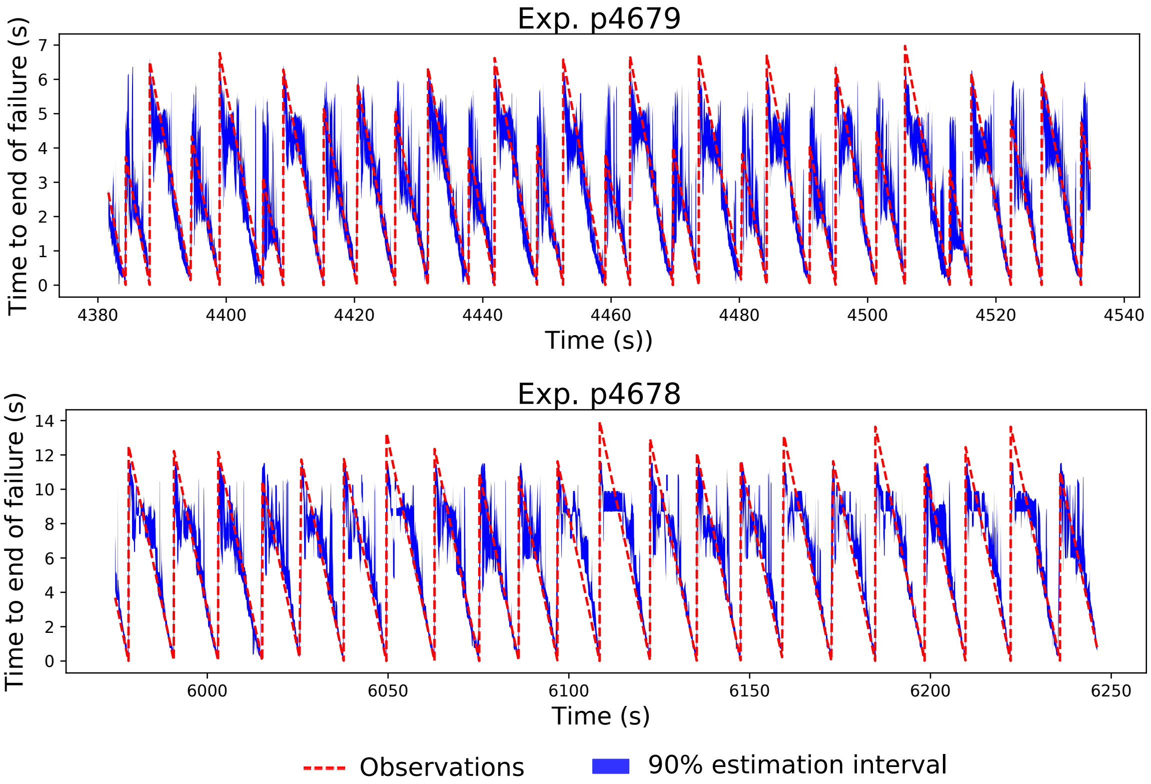

Figure S5. Model uncertainty reduces as the end of the current event approaches. The estimation intervals shown are for a random forest model, that allows us to easily compute such a measure. The estimation interval shown here means that 90% of the trees that compose the forest made an estimate within the blue shaded region. For both experiments, the models have greater confidence in their estimations of failure times when failure gets closer.

References

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

[2] B. Gregorutti, B. Michel, P. Saint-Pierre, Correlation and variable importance in random forests. *Statistics and Computing* **27**, 659–678 (2017).

[3] D. R. Jones, M. Schonlau, W. J. Welch, Efficient global optimization of expensive black-box functions. *Journal of Global optimization* **13**, 455–492 (1998).

[4] T. Chen, C. Guestrin, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (ACM, 2016), pp. 785–794.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees (1999).