J. Mech. Phys. Solids 174 (2023) 105249

Contents lists available at ScienceDirect

Journal of the Mechanics and Physics of Solids

journal homepage: www.elsevier.com/locate/jmps

Detection and reduction of systematic bias in high-throughput rupture experiments

Hou Wu [a,1], Xuhui Zhang [c,1], Yifan Zhou [a], Jose Blanchet [c,*], Zhigang Suo [b,*], Tongqing Lu [a,*]

[a] State Key Lab for Strength and Vibration of Mechanical Structures, International Center for Applied Mechanics, Department of Engineering Mechanics, Xi'an Jiaotong University, Xi'an 710049, P.R. China
[b] School of Engineering and Applied Sciences, Kavli Institute for Bionano Science and Technology, Harvard University, MA 02138
[c] Department of Management Science and Engineering, Stanford University, 475 Via Ortega, Stanford, CA 94305, United States

ARTICLE INFO

Keywords:
High-quality data
High-throughput experiment
rare-event rupture
Anderson-Darling test

ABSTRACT

Some high-throughput experiments aim to test many samples simultaneously under nominally the same conditions. However, whether a particular high-throughput experiment does so must be certified. We previously described a high-throughput experiment to study the statistics of rupture stretch of materials. In such an experiment, a large set of samples were tested simultaneously. We noticed a systematic bias that samples in different subsets give different statistical distributions of rupture stretch. Here we describe an approach to detect and reduce systematic bias in the high-throughput experiment. We divide the whole set of the data of the experiment into subsets, obtain the statistical distribution of rupture stretches for each subset, and compare the "closeness" of the distributions among the pairs of subsets by using the Anderson-Darling (A-D) test. We then try to reduce the systematic bias by improving the experimental design, such as increasing the rigidity of the frame and the firmness of the grips. With the improved experimental design, the newly obtained rupture data passes the A-D test. We use the new rupture data to fit the Weibull distribution. The improved high-throughput experiment greatly increases the accuracy of fitting and prediction of rare-event rupture stretch.

1. Introduction

One can roll a die one hundred times, or roll one hundred dice simultaneously. It is often tacitly assumed that the two experiments are equivalent. But this assumption is called into attention in our recent development of a high-throughput rupture experiment (Zhou et al., 2022). In such an experiment, a thousand samples are fabricated by 3D printing, and pulled simultaneously by a kinematic mechanism of one degree of freedom (Fig. 1a). We study the statistical distribution of rupture stretch of the one thousand samples. We notice a systematic bias that samples in different regions of the test structure give different statistical distributions of rupture stretch (Fig. 1b). This observation indicates that the samples are not tested under nominally the same conditions. As we will demonstrate, a high-throughput rupture experiment that does not test samples under nominally the same conditions may lead to a poor estimation of rupture stretch statistics.

* Corresponding authors.
E-mail addresses: jose.blanchet@stanford.edu (J. Blanchet), suo@seas.harvard.edu (Z. Suo), tongqinglu@mail.xjtu.edu.cn (T. Lu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jmps.2023.105249
Received 5 October 2022; Received in revised form 25 January 2023; Accepted 15 February 2023
Available online 26 February 2023

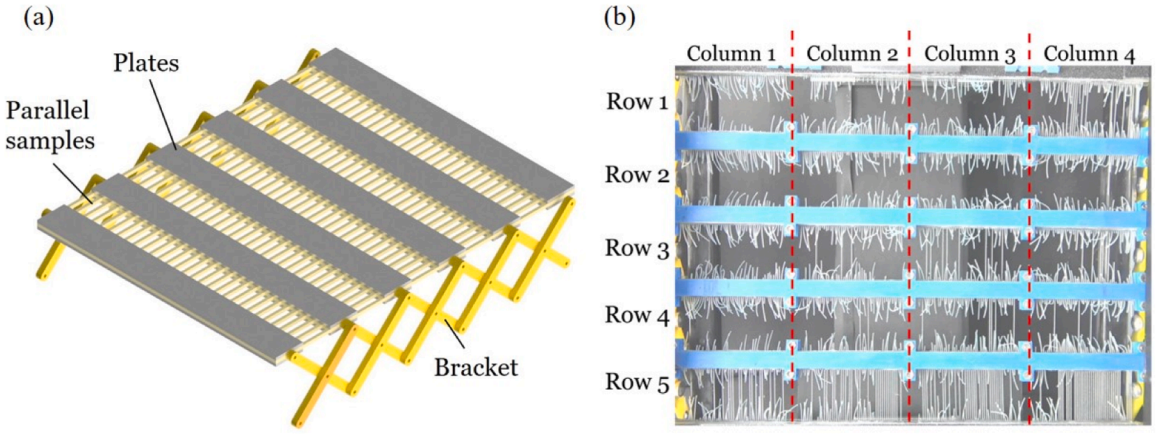


Fig. 1. Systematic bias in a high-throughput experiment. (a) Schematic of the high-throughput experiment. (b) A snapshot of the previous experiment (Zhou et al., 2022) shows that the fractions of ruptured samples have a systematic bias from row to row and from column to column.

Here we develop an approach to improve the high-throughput rupture experiment. We divide the whole set of the data of the previous experiment into subsets, and obtain the statistical distribution of rupture stretch for each subset. We compare the "closeness" of the distributions among the pairs of subsets by using the Anderson-Darling test (A-D test). Not surprisingly, the tests confirm substantial differences among the pairs of subsets. We then try to reduce the systematic bias by improving the experimental design, such as increasing the rigidity of the frame and the firmness of the grips. With the improved experimental design, the newly obtained rupture data passes the A-D test. We use the new rupture data to fit the Weibull distribution. The improved high-throughput experiment greatly increases the accuracy of fitting and prediction of rare-event rupture stretch.

Various high-throughput experiments have been developed in biology (Soon et al., 2013), pharmacy (Bajorath, 2002; Mennen et al., 2019), chemistry (Greenaway et al., 2018; Shevlin, 2017), and materials science (de Pablo et al., 2019; Hill et al., 2016; Ren et al., 2018; Sun et al., 2019). The high-throughput experiments to measure mechanical properties are very few. For example, high-throughput experiments are used to measure the elastic modulus of printed materials (Tweedie et al., 2005) or living cells (Darling and Di Carlo, 2015). Synthesizing and testing samples under nominally the same conditions have been challenging for almost all high-throughput experiments. For example, it is reported that a large amount of published data from the 1000 Genomes Project have a significant batch effect: subsets of the measurements have qualitatively different behavior. The effect can be caused by laboratory conditions, reagent lots and personnel differences that are unrelated to scientific variables (Leek et al., 2010). It is also reported that some small variations caused by protocol adaptations or errors in laboratorial routine analysis can lead to statistically different results in evaluating biofilms (Jorge et al., 2015). To reduce batch effects, one needs to improve both experiments and statistical analyses.

2. Statistical methods to test the systematic bias in a high-throughput experiment

Imagine that a high-throughput experiment has been run infinitely many times. Each run of the experiment generates a set of data of a large number of samples. Of these runs, one run is randomly selected, called a generic run. The data of the generic run can be divided into several subsets. Let $\{\lambda_1^1, \cdots, \lambda_{n_1}^1\}, \cdots, \{\lambda_1^k, \cdots, \lambda_{n_k}^k\}$ be the rupture stretches of k subsets, where n_i is the number of samples in the i-th subset. Denote cumulative distribution functions (cdfs) of the k subsets of samples by

$$F_{n_i}^i(\lambda) = \frac{1}{n_i}\sum_{j=1}^{n_i} 1_{\lambda_j^i \le \lambda}, i = 1, \cdots, k \tag{1}$$

The null hypothesis H_0 is that samples in the k subsets are independent and identically distributed (iid).

Anderson-Darling test (A-D test) is a non-parametric method to quantify the "closeness" between empirical cdfs of multiple subsets of samples (Scholz and Stephens, 1987). Define a test statistic A_{kN}^2 as

$$A_{kN}^2 = \sum_{i=1}^{k} n_i \int_{-\infty}^{\infty} \frac{\left(F_{n_i}^i(x) - H_N(x)\right)^2}{H_N(x)(1 - H_N(x))} dH_N(x) \tag{2}$$

where $N = n_i + \cdots + n_k$ is the combined sample size, and H_N is the cdf obtained by pooling the samples together. The test statistics A_{kN}^2 is a random quantity since $F_{n_i}^i, i = 1, \cdots, k$ are cdfs of k subsets of data of a randomly picked run. In reality, the experiment has only been run a small number of times. The cdf of an actual run of experiment is used to compute an actual value of the statistic according to Eq. (2), denoted by $\widehat{A}_{kN}^2$. Define the "P-value" as the probability that the A_{kN}^2 of the generic run exceeds the $\widehat{A}_{kN}^2$ of the actual run,

(a)

P-value calculated by A-D test between different rows			
(R1, R2)	0.0000	(R2, R4)	0.0000
(R1, R3)	0.0000	(R2, R5)	0.0000
(R1, R4)	0.0000	(R3, R4)	0.0871
(R1, R5)	0.0000	(R3, R5)	0.0037
(R2, R3)	0.0000	(R4, R5)	0.0668

(b)

P-value calculated by A-D test between different columns			
(C1, C2)	0.0084	(C2, C3)	0.0009
(C1, C3)	0.0000	(C2, C4)	0.0000
(C1, C4)	0.0000	(C3, C4)	0.0000

Fig. 2. Evaluate the systematic bias of the previous data. The *P*-value calculated by A-D test between (a) every two rows and (b) every two columns.

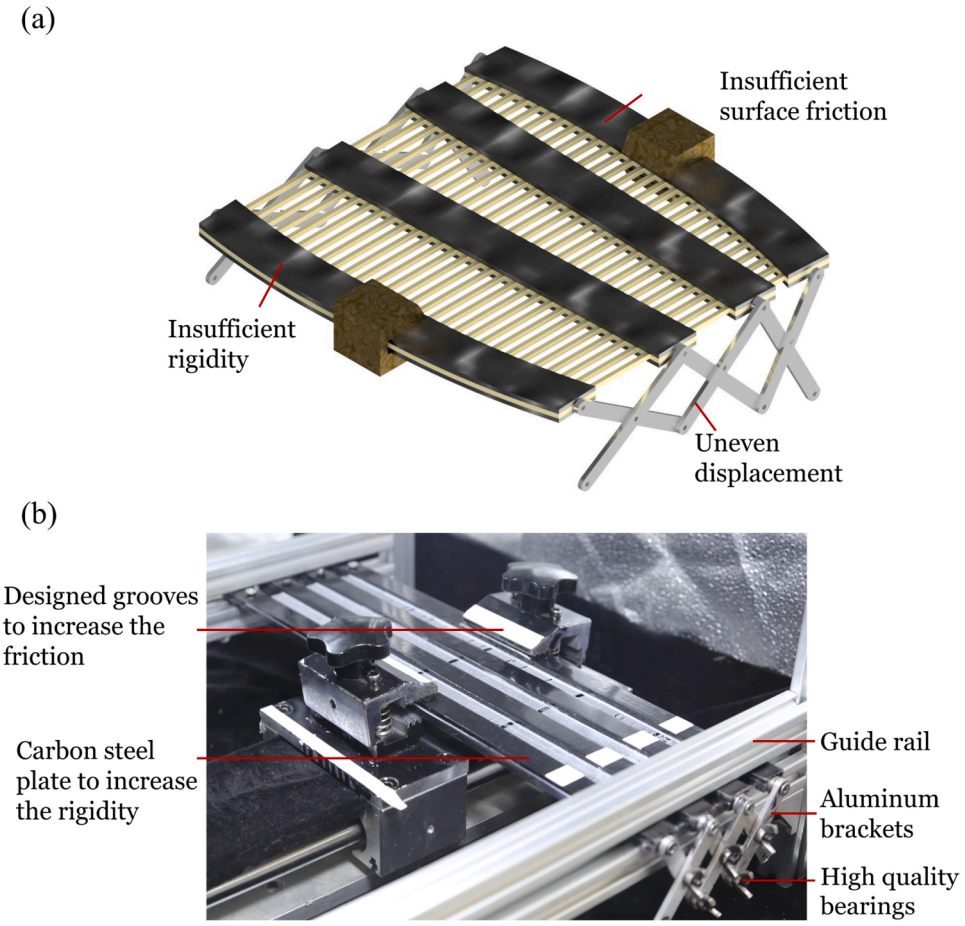


Fig. 3. Causes of systematic bias and methods to improve the experimental design. (a) In a high-throughput experiment, systematic bias is caused by insufficient rigidity of the plates, insufficient frictional force between the grips and the plates, and uneven displacement between different rows. (b) A photo of the improved experimental design.

$P(A^2_{kN} \geq \widehat{A}^2_{kN})$. Once the cdfs of the k subsets of samples are obtained, the P-value can be calculated using a function module "AnDarksamtest" available in MATLAB.

We select the significance level $\alpha = 0.1$ in this paper following the commonly used value in literature. If $P < \alpha$, we reject the null hypothesis H_0, and the samples in the k subsets are deemed not iid. If $P > \alpha$, we cannot reject the null hypothesis H_0, and the evidence is insufficient to deem that samples in the k subsets are not iid. The selection of significance level α is based on experience. Experimenters commonly specify the significance level, which is typically 0.01, 0.05, or 0.1. For example, Ho used the significance level 0.05 (Ho et al., 2019), and Petrone used the significance level 0.1 (Petrone et al., 2016). While we pick the significance level 0.1, we do not only

report the decision of acceptance and rejection but also report P-values of the tests, which is continuous. Generally two types of errors in a hypothesis test exist. Type I error is that the null hypothesis is true, but we reject it. Type II error is that the null hypothesis is false, but we accept it. For a finite sample size, it is usually impossible to make probabilities of both types of error small (Casella and Berger, 2021). The following trend often holds. The smaller the significance level, the lower the probability of making Type I error, and the higher the probability of making Type II error. One should choose the significance level according to which type of error causes more serious consequences in one's context. For our problem, our null hypothesis is that the samples of different subsets are iid. Making Type II error–namely, the data are not iid, but we falsely accept them as iid–is more serious for this problem. Once we accept the data are iid, we will proceed to use the predicted stretch for rare event rupture, for example, in design of structure. The failure of the material is of great practical consequence. Thus, we would rather err on being conservative, and choose a relatively high significance level of 0.1. We can let $k = 2$ and perform the pairwise two-subset test to find which subset has a clear bias from the others. We can also let k equal to the total number of subsets and perform a multiple-subset test to judge if there is a significant systematic bias in the whole data set.

There are multiple ways to test the null hypothesis. We have chosen the A-D test because the corresponding statistic given by Eq. (2) magnifies deviations in the tails more significantly than in the middle (Stephens, 1974). This feature is important in evaluating rare-event rupture stretch.

3. Evaluate the systematic bias of the previous data

In our previous high-throughput experiment, 1000 samples were tested simultaneously (Zhou et al., 2022). The samples are placed in five rows, with each row having 200 samples (Fig 1b). The five rows naturally divide the 1000 samples into five subsets. We calculate the cdf of each row and then use the A-D test to calculate the pairwise P-values $P_{R_{ij}}$ between the i-th row and the j-th row (Fig. 2a). However, all the P-values calculated using the data of these five rows are smaller than significance level, $\alpha = 0.1$,which indicates the distributions of the five rows are significantly different. The 1000 samples can also be divided into four subsets by columns. We calculate the cdf of each column and use the A-D test to calculate the pairwise $P_{C_{ij}}$ between the i-th column and the j-th column (Fig. 2b). All the $P_{C_{ij}}$ is smaller than α, which indicates the distributions of the four columns are also significantly different. With the A-D test we can further calculate the P-value of multiple subsets from the total sample set. The multiple subsets can be five rows ($k = 5$) or four columns ($k = 4$). In either way of dividing the subset, the calculated P_C, P_R are all close to zero, which indicates the distributions of these rows and columns are significantly different. These calculations conclude that the data in our previous high-throughput experiment does not pass the A-D test, providing ample evidence to reject the null hypothesis. We have also used the Kolmogorov-Smirnov test, and found a similar conclusion.

4. Improved high-throughput experiment

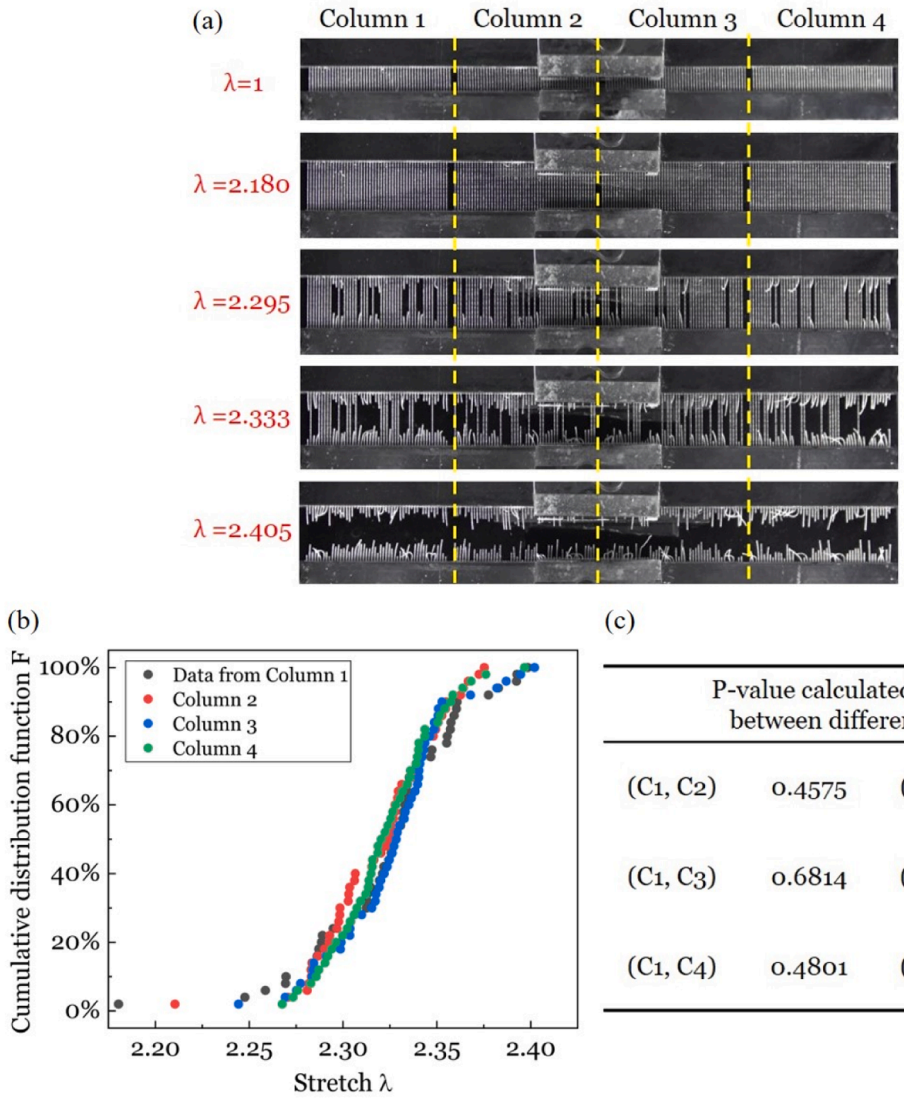
4.1. Experimental setup

The above statistical analysis has shown that the data obtained in the previous high-throughput experiment are not iid. We examine the previous experimental setup and find three main problems that cause the data inconsistency (Fig. 3a). First, the frictional force between the grips and the plates is insufficient. In the stretching process, if the samples on the left of the grips rupture more than those on the right, the plates slip and become tilted. Subsequently, the stretch applied to the left region is larger than that applied to the right region. The slippage in different rows can also be different. Second, the rigidity of the plates is insufficient. For samples in the top row and bottom row, in the stretching process, both sides of the plate may bend towards the samples so that the stretch applied to the samples at the middle part of the plate is larger than that applied to the samples on the sides of the plate. But for the rows in the middle, the bending effect is insignificant. Third, the brackets, consisting of bars and bearings, cannot transform displacement to different rows of samples equally due to the insufficient rigidity of the bars and the low transmission efficiency of the bearings. The three problems together lead to the systematic bias of the previous high-throughput experiment.

To resolve the three problems, we improve the experimental setup as follows. To increase the rigidity of the plates, we replace the material of the plates from aluminum to carbon steel, and increase the thickness and width of the plates from 5×5mm to 10×8mm. To avoid slip, we fabricate grooves in the middle of the plates to interlock with the grooves of the grippers. To ensure the kinematic mechanism to apply equal displacement to samples in different rows, we replace the plastic bars with aluminum bars, and apply lubricating oil to the bearings (Fig. 3b).

In stretching the multiple samples, we videotape the experiment, and identify the rupture of individual samples by image processing. The previous experiment was conducted in the ambient light, which changed with time and was disturbed by people nearby. The uncontrolled ambient light complicated the image processing. We improve the experiment as follows. Because the as-prepared samples are white, to increase contrast, we use a black and light-absorbing flannel as the background, and paint the kinematic mechanism black. We videotape the experiment in a studio, in which all background lights are blocked. To minimize shadowing, we place four light bulbs on the four corners of the studio.

We have also noticed another batch effect. In one run of the experiment, samples fabricated by the 3D printer have similar properties, such as rupture stretches. After the 3D printer has prepared samples for several runs of the experiment, the samples of different runs have different properties. The difference between samples in different runs is minimized once we clean the 3D printer after printing samples for each run of the experiment.



(c)

P-value calculated by A-D test between different columns			
(C1, C2)	0.4575	(C2, C3)	0.4515
(C1, C3)	0.6814	(C2, C4)	0.7401
(C1, C4)	0.4801	(C3, C4)	0.5726

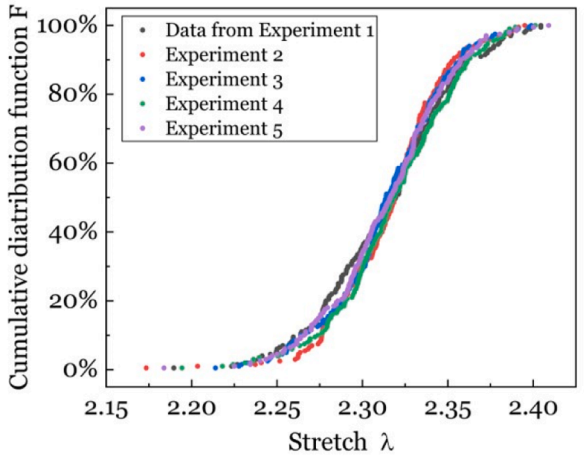
Fig. 4. A high-throughput rupture experiment of one row of 200 samples. (a) Snapshots of the 200 samples at several applied stretches. The 200 samples are divided into four columns (C1, C2, C3, C4). (b) The fractions of ruptured samples in each column are used to calculate an empirical cumulative distribution function of rupture stretches. (c) The *P*-value calculated by A-D test between every two columns.

4.2. Experimental procedure

The 600 samples are evenly distributed in three rows (Fig. 3b). Each sample is in the shape of a dog bone, with the size of the central part being $12 \times 0.5 \times 0.5$ *mm*. The distance between each sample is 1 *mm*. The three rows of samples are connected with rectangular bars of dimension $310 \times 15 \times 0.5$ *mm*. A 3D geometric description of the 600 samples and four rectangular bars is created in a CAD package, which is turned into .stl files for the Object 350 printer (Connex3, Stratasys). The printer operates in the digital printing mode, with a resolution of height 30μm. The 600 samples are printed with an elastomer (Agilus30Clear). The four rectangular bars are printed with a thermoplastic (VeroCyan).

Before the experiment starts, we place six limit blocks with a length of 12mm on both sides of the three-row samples to ensure that the initial stretch of each sample is 1.0. We stretch the 600 samples monotonically with a stretch rate of 0.1/min until all the samples rupture. The experimental process is videotaped and the rupture stretch of each sample is identified by image processing as follows. We use commercial software Potplayer to capture the video frame by frame. The time interval of two frames is 1/24s. The applied stretch is taken to be the current distance between the grippers divided by the initial distance between the grippers. Each image is then processed using MATLAB into a 1920×1080 matrix. Each element in the matrix represents the grayscale of a pixel. The elements of the matrix are then binarized: 255 for a pixel of a grayscale nearly white, and 0 for a pixel of a grayscale nearly black. Each sample in the image has its own position information, represented by a sub-matrix. During stretching, the upper and lower boundaries of each sample

(a)



(b)

P-value calculated by A-D test between different experiments			
(E1, E2)	0.1367	(E2, E4)	0.2649
(E1, E3)	0.3558	(E2, E5)	0.3926
(E1, E4)	0.3193	(E3, E4)	0.3753
(E1, E5)	0.6057	(E3, E5)	0.7628
(E2, E3)	0.5300	(E4, E5)	0.4020

Fig. 5. A high-throughput rupture experiment of one row of 200 samples is run five times. (a) The rupture stretches of samples in each run of experiment are used to calculate an empirical cumulative distribution function of rupture stretches. (b) The *P*-value calculated by A-D test between every two runs.

change, while the left and right boundaries are fixed. We select the material particles along the vertical line in the middle. The sub-matrix reduces to a vector. For each sample, we sum the elements of the vector and get one number. When a sample is broken, the sum is close to zero. When a sample is unbroken, the sum is a large number. By observing the change of the sums, we can obtain the position of each individual ruptured sample and the time when rupture of the sample occurs. The code for image recognition is in the supporting information.

5. Experimental results

5.1. A-D test on the samples of different columns in one row

We first test 200 samples in one single row in one run of experiment. We divide the 200 samples into 4 columns. In the experiment, the 200 samples are pulled from the undeformed state $\lambda = 1$ until all samples rupture. The rupture stretch of each sample is identified by image processing. Five snapshots of the experiment are shown in Fig. 4(a). At $\lambda = 2.180$, none of the 200 samples rupture. At $\lambda = 2.295$, 12, 11, 8, and 9 samples in the four columns rupture. At $\lambda = 2.333$, 31, 33, 29, and 32 samples in the four columns rupture. At $\lambda = 2.405$, all samples rupture. At a given stretch, the cumulative distribution function $F(\lambda)$ is calculated as the number of ruptured samples divided by the total number of samples (50). Fig. 4(b) plots the cdfs of the rupture stretch of the four columns.

We perform the A-D test using the four cdfs of the four columns. The calculated *P*-value between every two columns $P_{C_{ij}}$ are all larger than the given significance level, $\alpha = 0.1$(Fig. 4c), which indicates the distributions of the four columns are close. We calculate the *P*-value of multiple subsets ($k = 4$)from the total sample set. The calculated $P_C = 0.7460$ is much larger than the significance level $\alpha = 0.1$. Therefore, we cannot reject the null hypothesis H_0, and the samples in these four columns are not inconsistent with the iid assumption.

5.2. A-D test on the samples of one row in different runs of experiment

The experiment of 200 samples in a single row is run five times. Fig. 5(a) plots the cdfs of the rupture stretch of 200 samples of one row in five runs. We use the A-D test to calculate the pairwise $P_{E_{ij}}$ between every two runs (Fig. 5b). All the *P*-values are larger than the given significance level, $\alpha = 0.1$. We can also calculate the *P*-value of multiple subsets ($k = 5$) from the total sample set. The calculated $P_E = 0.5058$ is much larger than the significance level, $\alpha = 0.1$. Therefore, we cannot reject the null hypothesis H_0, and the samples in these five runs of experiment are not inconsistent with the iid assumption.

5.3. A-D test on the samples in different rows

We test 600 samples distributed in three rows in one run of experiment. The 600 samples are pulled from $\lambda = 1$ until all samples rupture. Four snapshots of the experiment are shown in Fig. 6(a). At $\lambda = 2.283$, 25, 27, and 29 of the three rows of samples rupture. At $\lambda = 2.331$, 106, 105, and 140 samples in the three rows rupture. It can be seen that at this stretch the samples in the third row rupture more than those in the other two rows. At $\lambda = 2.413$, all samples rupture. Fig. 6(b) plots the cdfs of the rupture stretch of the three rows.

We perform the A-D test using the three cdfs of the three rows (Fig. 6c). The $P_{R_{12}}$ is 0.4975, which is larger than the significance level

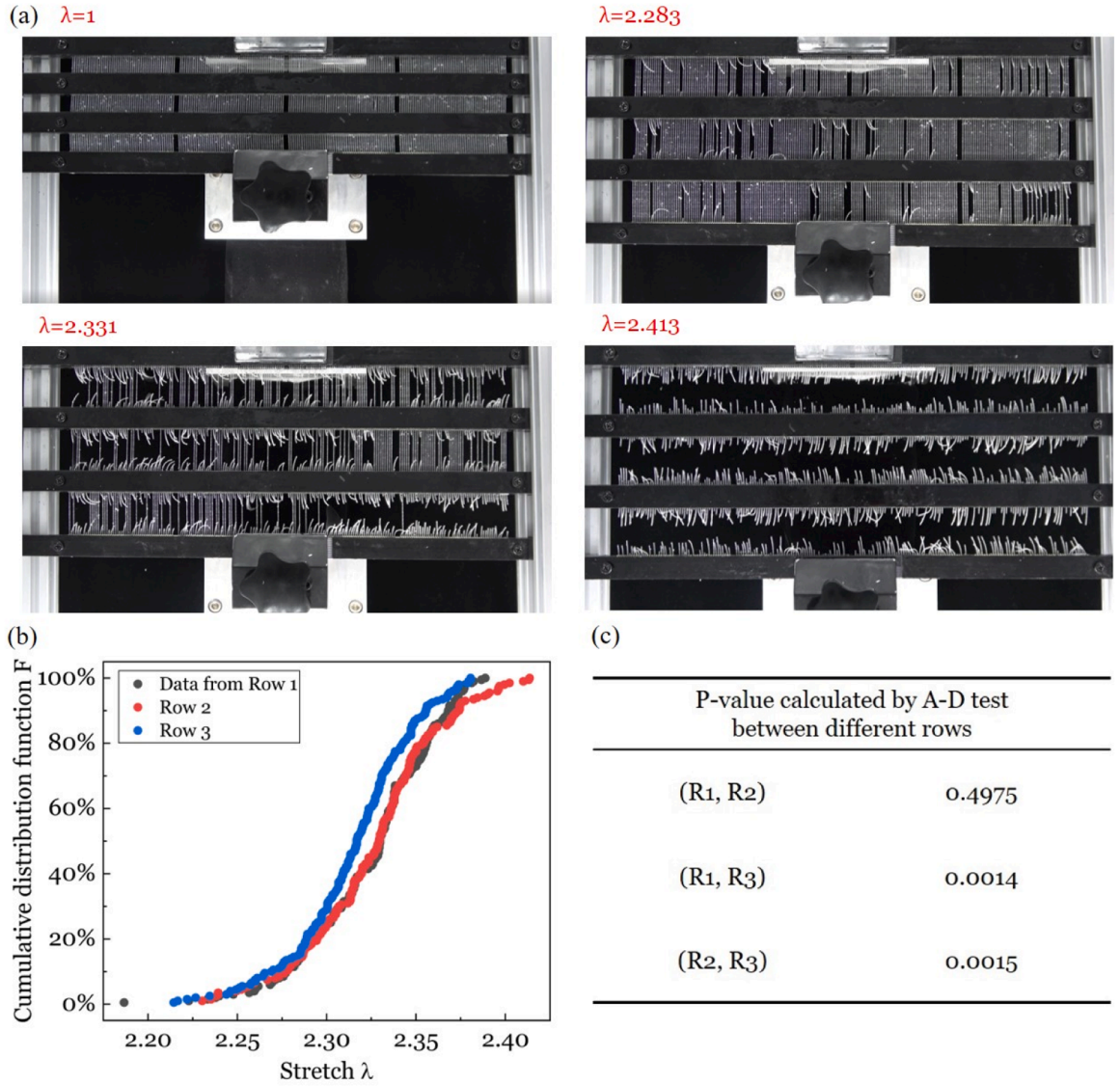


Fig. 6. A high-throughput rupture experiment for three rows of 600 samples. (a) Snapshots of the 600 samples at several applied stretches. The 600 samples are placed in three rows (R1, R2, R3). (b) The fractions of ruptured samples in each row are used to calculate an empirical cumulative distribution function of rupture stretches. (c) The *P*-value calculated by A-D test between every two rows.

$\alpha = 0.1$ and indicates the distributions of these two rows are similar. But the *P*-values calculated using the data of Row 3 are smaller than the significance level α, which indicates the distribution of this row is significantly different from the other two rows. We also calculate the *P*-value of multiple subsets $(k = 3)$ from the total sample set. The calculated $P_R = 0.0011$ is smaller than the significance level α. Therefore, we reject the null hypothesis H_0, and the samples in the different rows are inconsistent with the iid assumption. The data from the three rows do not pass the A-D test, which indicates the evenness of the displacement applied to different rows need to be further improved by experimental design.

6. Predict rare events with experimental data

We use the data obtained from the improved experimental setup to predict rare events of rupture. We aggregate the rupture data of 1000 samples from the five runs of the experiment on one single row in Fig. 5. We plot the cdf of the rupture stretch of the 1000 samples, each ruptured sample corresponding to a red dot in the $F - \lambda$ plane (Fig. 7a). According to a common practice in fracture mechanics (Doremus, 1983), we fit the measured cdf to the three-parameter Weibull distribution (Coles et al., 2001; Weibull, 1951).

$$F(\lambda) = 1 - \exp\left[-\left(\frac{\lambda - \alpha}{\beta}\right)^{\gamma}\right] \tag{3}$$

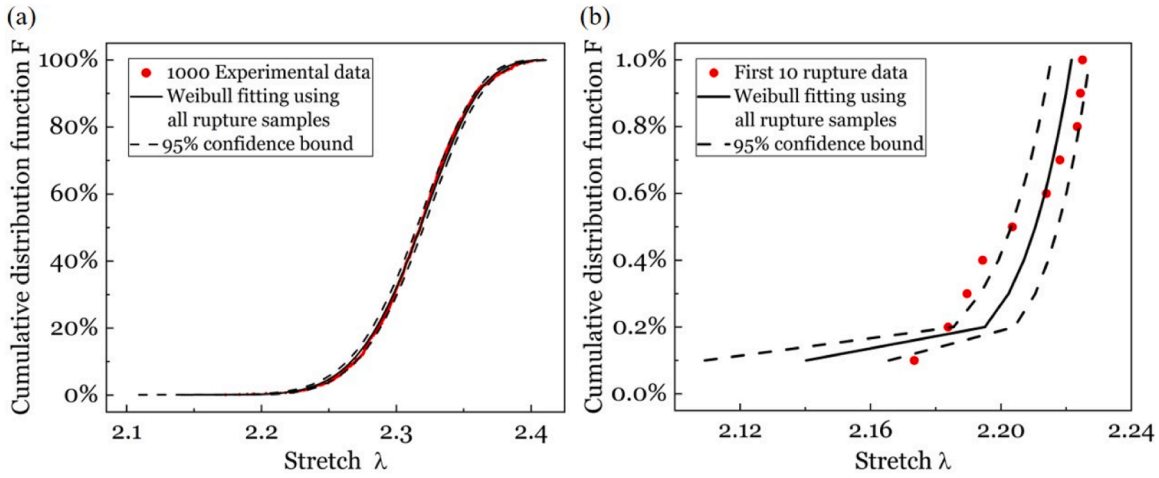


Fig. 7. Weibull fit with the rupture stretch of the 1000 samples. (a) The 1000 experimental data from the five runs of the experiment are used to calculate the cdf of rupture stretches (red dots). The cdf is fitted by the Weibull distribution (black solid curve). Also plotted is the 95% confidence interval (dashed black curves). (b) The zoom-in of the plot in the range $0 < F(\lambda) < 1\%$.

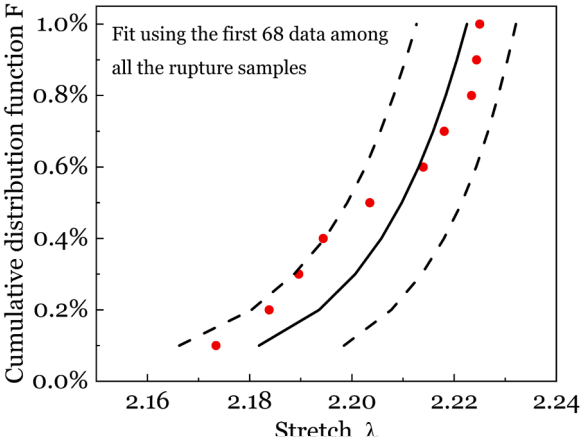


Fig. 8. The cdf of the 1000 experimental data up to $F = 1\%$ (10 red dots). The first 68 ruptured samples are selected using the peak-over-threshold method to analyze the rare-event rupture. The cdf of the 68 ruptured samples is fit to the Weibull distribution (solid black curve). Also plotted is the 95% confidence interval (dashed black curves).

where α, β, and γ represent the location, scale, and shape of the Weibull distribution, respectively. We use the maximum likelihood method to obtain the fitting parameters of the Weibull distribution as α=2.0974, β=0.2331, and γ=6.9036 (Millar, 2011). The fitting curve is marked with a black and solid curve (Fig. 7a). For a given value of the cdf, we further plot the 95% confidence interval marked with two black dashed curves (Fig. 7a). The 95% confidence interval is narrow in the entire range of rupture stretch and most of the experimental data lie within the 95% confidence interval, which indicates that the Weibull distribution fits well to the entire range of the experimental data. Fig. A1 shows the Weibull distribution does not fit well with the data reported in our previous paper (Zhou et al., 2022).

To discuss the rare-event rupture, we magnify the plot in Fig. 7(a) in the range $0 < F(\lambda) < 1\%$ to Fig. 7(b). Here we focus on the first 10 ruptured samples. In this range, some experimental data lies outside the 95% confidence interval. That is, the Weibull fit using the data of all 1000 samples is unable to predict the rare events with a high confidence even if we use the data of the improved experiment. To achieve a prediction for rare-event rupture with a higher accuracy and confidence, we should not use the entire range of data, but instead use the data close to the left tail. We do so by applying the peak-over threshold method (Coles et al., 2001). The method focuses on the statistics of the left tail by imposing a threshold stretch. We apply the method to our data, and the threshold stretch is chosen as $\lambda = 2.26$. Of the 1000 experimental results, 68 samples rupture when the stretch is less than the threshold stretch. We use these 68 data to fit the Weibull distribution, and also plot the fitting curve and the 95% confidence interval (Fig. 8). With this method, all the experimental data in the range $0 < F(\lambda) < 1\%$ fall within the 95% confidence interval. For example, we specify a rare event by the cumulative probability $F(\lambda) = 0.1\%$, corresponding to the first ruptured sample among the 1000 samples. For the rare event of "0.1%

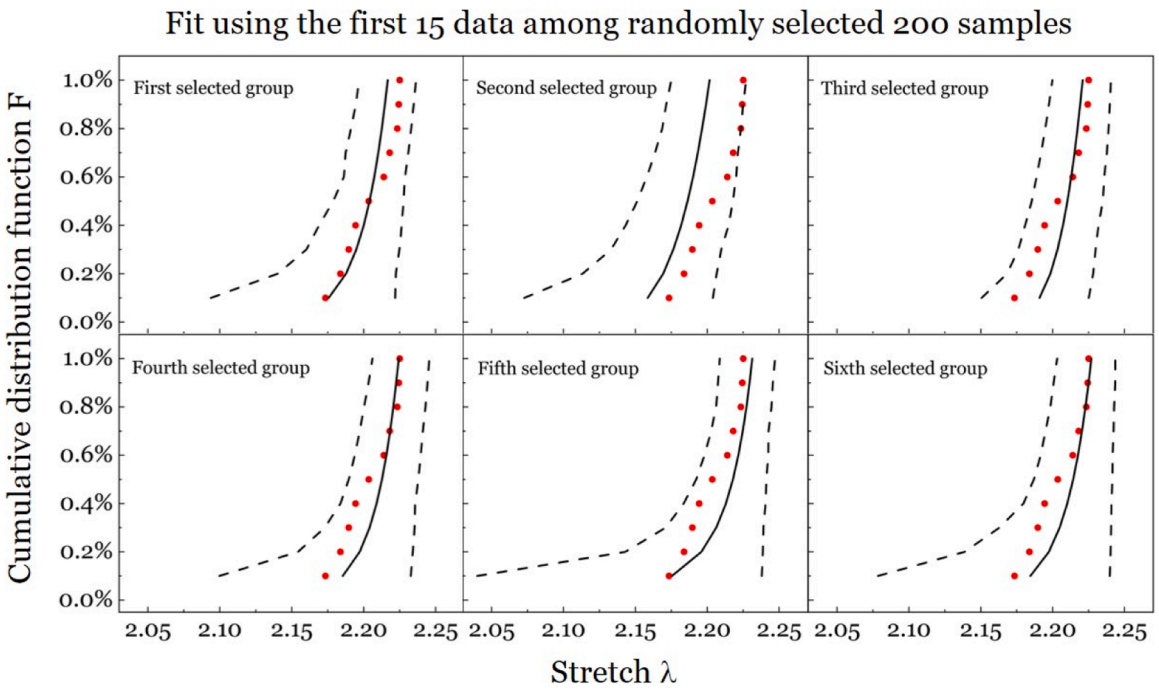


Fig. 9. Predict rare-event with a smaller data set. From the 1000 tested samples randomly select 200 samples. Of the 200 samples, the first 15 ruptured samples are chosen to fit the Weibull distribution, and are used to calculate the 95% confidence interval. Also included are the measured cdf of the first 10 ruptured samples among the 1000 samples. This procedure is repeated six times.

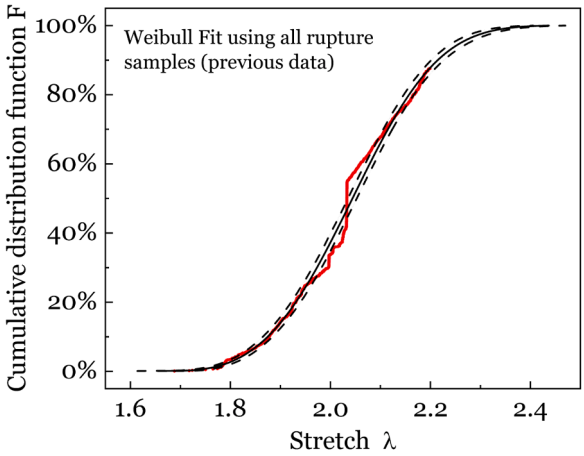


Fig. A1. The previous rupture stretches are used to calculate the cumulative distribution function (cdf), $F(\lambda)$. Each ruptured sample corresponds to a red dot in the $F-\lambda$ plane. The measured cdf is fit to the Weibull distribution (black solid line). Also plotted is the 95% confidence interval (two black dashed lines)

rupture", the measured rupture stretch is $\lambda = 2.173$, the Weibull fit is $\lambda = 2.182$ and the 95% confidence interval is $2.166 < \lambda < 2.198$. This high level of confidence, as well as the narrow range of stretch, is likely to satisfy most applications. Fig. A2 shows the prediction of rare events is not as good using the data obtained in our previous experiment (Zhou et al., 2022).

Since the quality of data is improved in the present experiment, we test if we can predict the rare-event of rupture with a satisfactory accuracy using a smaller data set than the whole data set. To represent a smaller data set, from the 1000 tested samples we randomly select 200 samples six times. Each time we use the peak-over-threshold method to get a threshold stretch, below which 15 samples rupture out of the 200 samples. We use the first 15 rupture data to fit the Weibull distribution, and also calculate the 95% confidence interval (Fig. 9). Note that the six selections are random and, in each selection the 200 samples are different. We find that the fitting results of the six selections are consistent: all the experimental data fall within the 95% confidence interval. By comparison, the predictions of rare events using 200 samples in our previous experiment are not good (Fig. A3). This confirms that when the quality

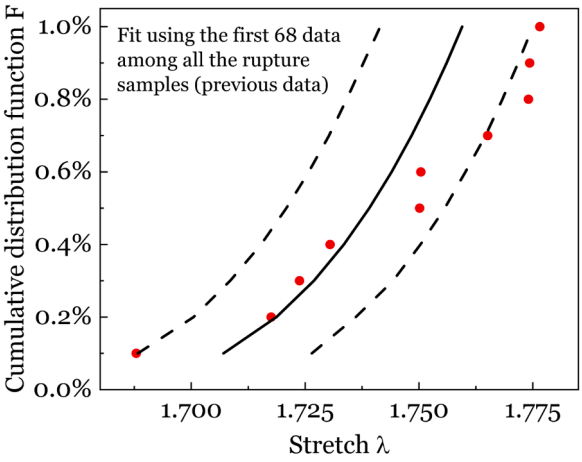


Fig. A2. The measured cdf up to the first 68 previous ruptured samples is fit to the Weibull distribution. Also plotted are the 95% confidence interval, as well as the measured cdf of the first 10 ruptured samples.

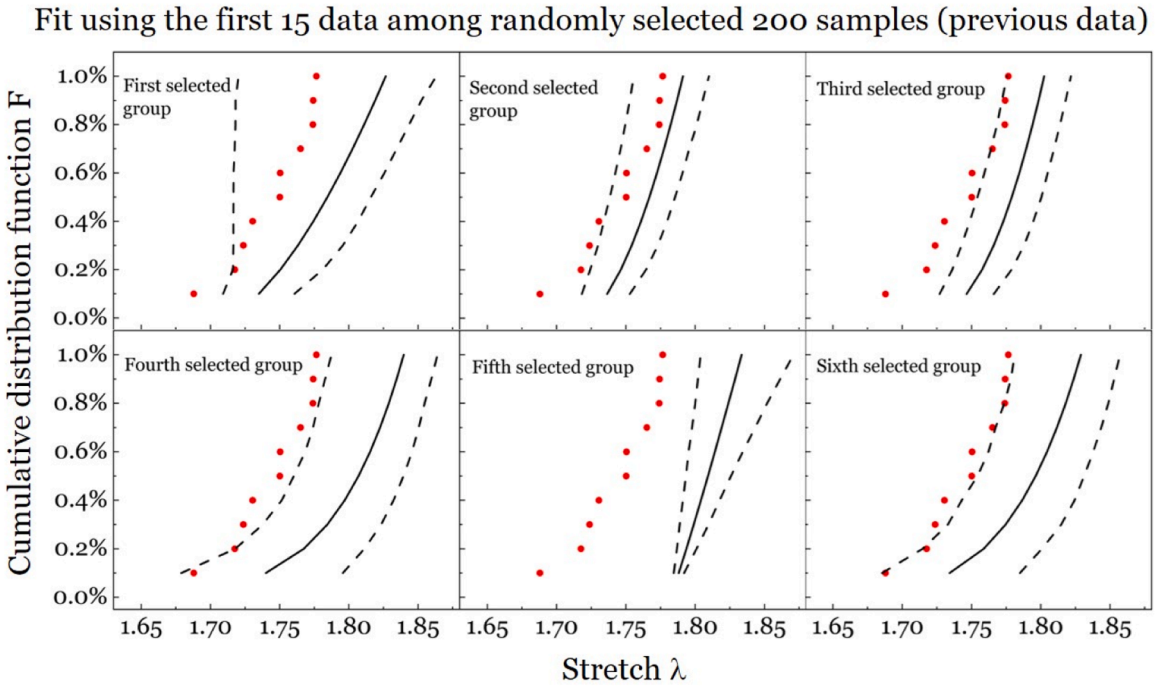


Fig. A3. Predict rare-event with a smaller data set. From the 1000 previous rupture samples randomly select 200 samples. Of the 200 samples, the first 15 ruptured samples are chosen to fit the Weibull distribution, and are used to calculate the 95% confidence interval. Also included are the measured cdf of the first 10 ruptured samples among the 1000 samples. This procedure is repeated six times.

of the data is improved, the number of data needed to predict rare-event rupture with a high accuracy and confidence can be reduced.

7. Conclusions

We use the statistical methods, the A-D test, to show that the data of rupture stretch in our previous high-throughput experiment are not iid. We improve the experimental setup such as increasing the rigidity of the frame and the firmness of the grips. With the improved experimental setup, we run the high-throughput rupture experiment again. When each run of the experiment only contains one row of samples, the data of rupture stretch in different columns of one row are not inconsistent with the iid assumption. Similarly, the data of different runs of the experiment are not inconsistent with the iid assumption. However, when each run of the experiment contains three rows of samples, the data in different rows are inconsistent with the iid assumption. We use the left tail of the data of 1000 samples

obtained by five runs of single-row experiment to fit the Weibull distribution, and predict the rare-event rupture at small stretches. The data from the improved experiment predict rare-event rupture stretch with a high accuracy and confidence. Furthermore, we show that the number of data needed to predict rare-event rupture can be reduced. This work provides a procedure to improve the quality of data in a high-throughput experiment: run the high-throughput experiment, use statistical methods to test if the data of the high-throughput experiment are consistent with the iid assumption, find out the sources of systematic bias, improve the experimental setup, and iterate.

CRediT authorship contribution statement

Hou Wu: Validation, Formal analysis, Writing – original draft. **Xuhui Zhang:** Formal analysis, Writing – original draft. **Yifan Zhou:** Validation, Writing – review & editing. **Jose Blanchet:** Writing – review & editing, Validation, Supervision. **Zhigang Suo:** Writing – review & editing, Validation, Supervision. **Tongqing Lu:** Writing – review & editing, Validation, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work at Xi'an Jiaotong University is supported by NSFC (No.11922210). The work at Stanford and Harvard is supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397.

Appendix A

Weibull distribution fitting

We use the software R-studio to fit the cdf of the measured rupture stretches to the Weibull distribution. In particular, we use the maximum-likelihood estimation method to determine the parameters of the Weibull distribution. The likelihood function is defined by the Weibull density:

$$L_n(\lambda_1, \cdots, \lambda_n; \theta) = \prod_{i=1}^{n} f(\lambda_i; \theta) \tag{4}$$

where L_n is the likelihood function based on n iid observations, λ_i is the rupture stretch for the i-th ruptured sample, $f(\lambda_i)$ is the probability density function of Weibull distribution evaluated at λ_i, θ is the underlying true parameters (α,β,γ). We seek for an estimator $\widehat{\theta}$ that maximizes the value of $L_n(\lambda_1, \cdots, \lambda_n; \theta)$. We carry out the maximum-likelihood fitting with the default numerical-optimization method using the function "optim" in R-studio. After the fitting procedure, we can obtain both the fitted parameters $\widehat{\theta}$ and the observed information matrix L, where L approximates the expected curvature of the log-likelihood surface.

The sampling distribution of the fitted parameters $\widehat{\theta}$ approximately obeys the multivariate normal distribution $N(\theta, L^{-1})$(Coles, 2001). And we construct the sampling distribution of the fitted stretches by the inverse function of Eq. (3). The confidence interval is obtained by truncating the sample distribution. For example, to plot the 95% confidence interval, the confidence limits are the lower and upper 2.5% quantiles of the sampling distribution.

Peak-over-threshold method

To fit the Weibull distribution with the data near the rare event, we use the peak-over-threshold method (Coles, 2001). The Weibull distribution now corresponds to the distribution of minimum

$$M_n = \min_{1 \le i \le n} \lambda_i \tag{5}$$

Following (Coles, 2001), we consider the point process on R^2 plane

$$N_n = \left\{ \left(\frac{i}{n+1}, \lambda_i \right) : 1 \le i \le n \right\} \tag{6}$$

A point process on a set A is a stochastic rule for the occurrence and position of point events. The scaling in the first ordinate ensures that the time axis is always mapped to $(0,1)$; the second ordinate represents the rupture stretch. Consider a region of the form

$$A = [0,1] \times (-\infty, \lambda_{\max}) \tag{7}$$

for some rupture stretch that smaller than threshold stretch $\lambda_{\max}$. Here we introduce the "intensity measure $\Lambda(A)$" of the process, which gives the expected number points in set A. The point process N_n can be approximated by a Poisson point process with intensity measure on $A = [0,1] \times (-\infty, \lambda_{\max})$, which is given by

$$\Lambda(A) = (1-0)\left(\frac{\lambda_{\max} - \alpha}{\beta}\right)^{\gamma} \tag{8}$$

According to the likelihood of Poisson point process (Coles, 2001), we can obtain the likelihood function by using intensity measure $\Lambda(A)$

$$L_n\left(\lambda_{(1)}, \cdots, \lambda_{(n)}\right) = \exp\left[-\left(\frac{\lambda_{\max} - \alpha}{\beta}\right)^{\gamma}\right] \cdot \prod_{i=1}^{n_r}\left(\frac{\gamma}{\beta}\right) \cdot \left(\frac{\lambda_{(i)} - \alpha}{\beta}\right)^{\gamma-1} \tag{9}$$

where $\lambda_{(i)}$ is the rupture stretch ordered such that $\lambda_{(i)} \le \lambda_{(j)}$ if $i \le j$. Here $\lambda_{\max}$ is the threshold stretch and n_r is the total number of observations with stretches smaller than $\lambda_{\max}$. To select the threshold stretch $\lambda_{\max}$ used in the Poisson point process modeling (Coles, 2001), we use the empirical mean residual life plots, which is defined as:

$$E = \frac{1}{n_u}\left(\sum_{i=1}^{n_u}\left(u - \lambda_{(i)}\right) : u \le \lambda_{\max}\right) \tag{10}$$

where E is the empirical mean excess residual, u is any given threshold, n_u is the number of observations that is smaller than u. We choose $\lambda_{\max}$ that for $u \le \lambda_{\max}$ the pairs (u, E) are approximately linear in the empirical mean residual life plots. We carry out the maximum-likelihood fitting using the function "fitpp'' and the function "mrlplot'' from the package "POT'' in R-studio. Notice that the package "POT'' is intended for the extreme-maxima modeling, whereas our interest is the modeling of the extreme minima. We therefore modify the stretch values to their negative values before passing the data set into these functions, to transform our original problem into an equivalent extreme maxima modeling problem.

Appendix B

References

Bajorath, J., 2002. Integration of virtual and high-throughput screening. Nat. Rev. Drug Discov. 1, 882–894.

Casella, G., Berger, R.L., 2021. Statistical inference. Cengage Learn.

Coles, S., Bawa, J., Trenner, L., Dorazio, P., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer.

Darling, E.M., Di Carlo, D., 2015. High-throughput assessment of cellular mechanical properties. Annu. Rev. Biomed. Eng. 17, 35–62.

de Pablo, J.J., Jackson, N.E., Webb, M.A., Chen, L.-Q., Moore, J.E., Morgan, D., Jacobs, R., Pollock, T., Schlom, D.G., Toberer, E.S., 2019. New frontiers for the materials genome initiative. npj Comput. Mater. 5, 1–23.

Doremus, R.H., 1983. Fracture statistics - a comparison of the normal, Weibull, and Type-I extreme value distributions. J. Appl. Phys. 54, 193–198.

Greenaway, R., Santolini, V., Bennison, M., Alston, B., Pugh, C., Little, M., Miklitz, M., Eden-Rump, E., Clowes, R., Shakil, A., 2018. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. Nat. Commun. 9, 1–11.

Hill, J., Mulholland, G., Persson, K., Seshadri, R., Wolverton, C., Meredig, B., 2016. Materials science with large-scale data and informatics: unlocking new opportunities. Mrs Bull. 41, 399–409.

Ho, J.C., Michalak, A.M., Pahlevan, N., 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. Nature 574, 667–670.

Jorge, P., Lourenco, A., Pereira, M.O., 2015. Data quality in biofilm high-throughput routine analysis: intralaboratory protocol adaptation and experiment reproducibility. J. AOAC Int. 98, 1721–1727.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739.

Mennen, S.M., Alhambra, C., Allen, C.L., Barberis, M., Berritt, S., Brandt, T.A., Campbell, A.D., Castanon, J., Cherney, A.H., Christensen, M., Damon, D.B., de Diego, J. E., Garcia-Cerrada, S., Garcia-Losada, P., Haro, R., Janey, J., Leitch, D.C., Li, L., Liu, F.F., Lobben, P.C., MacMillan, D.W.C., Magano, J., McInturff, E., Monfette, S., Post, R.J., Schultz, D., Sitter, B.J., Stevens, J.M., Strambeanu, J.I., Twilton, J., Wang, K., Zajac, M.A., 2019. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. Org. Process. Res. Dev. 23, 1213–1242.

Millar, R.B., 2011. Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB. John Wiley & Sons.

Petrone, C., Magliulo, G., Manfredi, G., 2016. Mechanical properties of plasterboards: experimental tests and statistical analysis. J. Mater. Civil Eng. 28.

Ren, F., Ward, L., Williams, T., Laws, K.J., Wolverton, C., Hattrick-Simpers, J., Mehta, A., 2018. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. Sci. Adv. 4, eaaq1566.

Scholz, F.W., Stephens, M.A., 1987. K-sample Anderson-Darling tests. J. Am. Stat. Assoc. 82, 918–924.

Shevlin, M., 2017. Practical high-throughput experimentation for chemists. ACS Med. Chem. Lett. 8, 601–607.

Soon, W.W., Hariharan, M., Snyder, M.P., 2013. High-throughput sequencing for biology and medicine. Mol. Syst. Biol. 9, 640.

Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. 69, 730–737.

Sun, S., Hartono, N.T., Ren, Z.D., Oviedo, F., Buscemi, A.M., Layurova, M., Chen, D.X., Ogunfunmi, T., Thapa, J., Ramasamy, S., 2019. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. Joule 3, 1437–1451.

Tweedie, C.A., Anderson, D.G., Langer, R., Van Vliet, K.J., 2005. Combinatorial material mechanics: high-throughput polymer synthesis and nanomechanical screening. Adv. Mater. 17, 2599. -+.
Weibull, W., 1951. A statistical distribution function of wide applicability. J. Appl. Mech.-Trans. Asme 18, 293–297.
Zhou, Y.F., Zhang, X.H., Yang, M., Pan, Y.D., Du, Z.J., Blanchet, J., Suo, Z.G., Lu, T.Q., 2022. Article high-throughput experiments for rare-event rupture of materials. Matter 5, 654-+.