Quantitative matching of forensic evidence fragments using fracture surface topography and statistical learning
Reviewer #1 (Remarks to the Author):

A typed PDF document was uploaded, along with a PDF version of the manuscript using Adobe “comments” to ask in-line text questions. They are copy/pasted here for convenience:

The following uses Nature's Criteria for Publication in order to evaluate the submission. The criteria are bulleted, followed by the reviewer's response:

Criteria for publication

- The quality of the data — whether they are technically sound, obtained with appropriate techniques, analysed and interpreted carefully, and presented in sufficient detail.

Response: Manner of data collection is well-described in the supplemental material, but the authors are encouraged to reference this in the actual manuscript since without this information, a reader will not realize that the supplement can be consulted for more details.

Fig. 3 is heavily utilized to explain the analytical procedure. However, increased clarity is requested in a few places. First, there is no discussion of “alignment” in image pairs, but there is reference to the use of FT information for image registration (yet it is not clear if the authors are using information from the power spectrum to achieve any form of registration).

A possible major detractor is the use of training versus testing data. After reading it is unclear if the image pairs used for model training are the same pairs used for object classification. If the sets are the same, then the reported classification accuracy may be inflated and independent testing/training sets are needed. If they are not the same, the reader would benefit from a very clear description of this difference.

- The level of support for the conclusions — whether sufficiently strong evidence is provided for the authors' claims and all appropriate controls have been included.

Response: Support of the utility of the method is largely a function of the testing/training issue formerly discussed. If dependent, then the results are possibly inflated and a second independent testing set should be added.

- The potential significance of the results — whether these results will be important to the field and advance understanding in a way that will move the field forward. (Note that posting of preprints and/or conference proceedings does not compromise novelty.)

Response: To the reviewer's knowledge, the exact approach used in this study has not been applied to forensic toolmark analysis in the past. Thus, the results are interesting and important. However, the use of objective methods to support toolmark associations is not a novel concept, and several other methods exist in the primary literature. In terms of a comparison of the proposed method versus those that already exist, since none have been compared using the exact same datasets, doing so would naturally be the next step in order to determine if the proposed work surpasses the performance of other objective metrics.

Elements of a reviewer report

- Key results-- Your overview of the key messages of the study, in your own words, highlighting what you find significant or notable. Usually, this can be summarized in a short paragraph.

Response: By collecting non-contact 3D images of fracture surfaced, followed by computation of the correlation coefficient in the Fourier space (power spectrum of FFT), selected frequency bandpasses can be used to describe the similarity of mated and non-mated pairs. The distribution (density) of these similarities can be modeled using a matrix-variate t distribution, which can be used to generate two conditional probabilities describing the likelihood of correlations being sampled from the mated and non-mated pairs. The LR (log LR) information can be transformed into a binary classification system, and false positive and negative rates generated in order to
evaluate the merits of the proposed approach.

• Validity --Your evaluation of the validity and robustness of the data interpretation and conclusions. If you feel there are flaws that prohibit the manuscript's publication, please describe them in detail.

Response: The reviewer was unable to resolve the use of the data sets. The abstract indicates fifty, and the text indicates two sets of 9 and two sets of 10 (18+20). The training illustrations/figures report at least one set of knives being used, 81 known matches (9 sets x 9 images). If the authors could clearly delineate testing/training/dependence/independence (which they do for a hold-one-out and a repeated imaging scenario) the reader could better evaluate the validity of the conclusions.

• Significance -- Your view on the potential significance of the conclusions for the field and related fields. If you think that other findings in the published literature compromise the manuscript's significance, please provide relevant references.

Response: The authors report an objective and numerical method for fracture matching that has not (to the reviewer's knowledge) been presented in the field of forensic to date. This is significant and informative, and would benefit from a future study illustrating how amenable it is to solving these types of problems when presented with additional datasets and material types.

• Data and methodology-- Your assessment of the validity of the approach, the quality of the data, and the quality of presentation. We ask reviewers to assess all data, including those provided as supplementary information. If any aspect of the data is outside the scope of your expertise, please note this in your report or in the comments to the editor. We may, on a case-by-case basis, ask reviewers to check code provided by the authors (see this Nature editorial for more information).

Response: The datasets are relatively small (at most 50 pairs). Clarity is needed to better understand training/testing question previously discussed. Some aspects of the methodology are unclear (alignment, etc.), and have been annotated using PDF Adobe comments in the actual manuscript. This reviewer was not familiar with density estimation using the prescribed matrix of correlation scores, but found the author’s description useful. The authors spent considerable time discussing tuning for degrees of freedom, but this detail is definitely outside the scope of my expertise and little comment (constructive or critical was offered in this section of the manuscript).

• Analytical approach -- Your assessment of the strength of the analytical approach, including the validity and comprehensiveness of any statistical tests. If any aspect of the analytical approach is outside the scope of your expertise, please note this in your report or in the comments to the editor.

Response: Limited statistical testing was performed (i.e., the modified KS test seemed valid as used). Instead, the focus was on ratios of conditional probabilities, and if binarized into a classification threshold, false positive and negative rates. These seemed fairly standard and appropriate.

• Suggested improvements -- Your suggestions for additional experiments or data that could help strengthen the work and make it suitable for publication in the journal. Suggestions should be limited to the present scope of the manuscript; that is, they should only include what can be reasonably addressed in a revision and exclude what would significantly change the scope of the work. The editor will assess all the suggestions received and provide additional guidance to the authors.

Response: Critical flaws were not identified, however, various clarifications are requested in order to improve the reader’s ability to understand nuances of the paper. These have been itemized above, and/or as actual comments in the manuscript --- and represent copyedit changes rather than substantial methodology corrections.

• Clarity and context -- Your view on the clarity and accessibility of the text, and whether the
results have been provided with sufficient context and consideration of previous work. Note that we are not asking for you to comment on language issues such as spelling or grammatical mistakes.

Response: Some clarifications are requested, but all are believed to be easily be addressed/corrected.

- Your expertise --Please indicate any particular part of the manuscript, data or analyses that you feel is outside the scope of your expertise, or that you were unable to assess fully.

Response: I was able to provide constructive feedback as a “typical forensic” reader. Although I am versed in image analysis, microscopy, error rates, Fourier transforms, extraction of information from the power spectral density, the use of correlation coefficients, and density estimation, I am not a statistician, and aspects of the density estimation modelling (selection of DF) are outside of my area of expertise.

Reviewer #2 (Remarks to the Author):

In "Quantitative Matching of Forensic Evidence Fragments using Fracture Mechanics and Statistical Learning," Thompson and co-workers assess whether two fracture surfaces can be recognized as paired using comparative microscopy and physical pattern analysis. Their contribution to this field is to identify an optimal range of search region sizes. The work is important and well executed, and the reviewer is enthusiastic.

The reviewer recommends addressing the following suggestions for the final version of this manuscript.

(1) The fracture process zone
The linkage to fracture mechanics process zone in the article is limited to the assertion that "It is established in fracture mechanics that the fracture process zone ahead of the crack tip typically extends to 2-3 times the grain size." This is attributed to Anderson’s book, but the reviewer had trouble finding this statement in the text, and suspects that it might be a rule of thumb. This is certainly not the case for highly ductile metals or single crystal turbine blades. Would it make sense to speak in terms of grain size in the paper, and then perhaps link to the fracture mechanics justification in the discussion? Otherwise, the assertion that this is based upon fundamental fracture mechanics probably requires a notch fracture test and careful strain analysis of the specific metals used to quantify the process zone size.

(2) Language in the paper
The language is loose in places, and the reviewer suggests that it be tightened. The whole paper could use trimming and rewording. The following statements need to be fixed:
- "The individual characteristics of the fracture process leave surface marks on both surfaces that can be identified in order to match fragments to each other reliably." (This is an imprecise statement)
- "This is the scale wherein the local stresses ahead of the crack tip reach a critical level, sufficient to overcome the intrinsic resistance of the material to fracture" (The authors, who include experts in fracture, will recognize this statement as being incorrect)

(3) The optimization of scanning area was not clear. Given that this is a key contribution of the paper, would it make sense to add a figure showing this explicitly?

(4) The metrics were for comparison were not clear. The reviewer was unfamiliar with the idea of "Log Odds" in Figure 6. What do they mean, and why not used percentage likelihood of a false positive match? Could the latter be added?

(5) Figure 9, and the false negative rate: Why are false negatives more likely with increasing
numbers of images?

Congratulations on this important and interesting work!

Reviewer #3 (Remarks to the Author):

Reviewer: Steve Lund, Statistical Engineering Division, NIST

Summary:

This paper’s primary contribution is an algorithm for evaluating whether two fragments originated from a common source before a fracturing event. Additionally, the manuscript proposes models to represent the distribution of matrices of correlations between images evaluated at different regions and length scales of a given fracture. The article also examines the influence of various modeling choices and algorithm settings on discrimination performance and is shown to perform well across a range of configurations. I believe this work helps address concerns raised by the NAS report “Strengthening Forensic Science in the United States: A Path Forward” and will be of interest to the forensics community. I recommend that this paper be published following minor revision.

Comments:
The writing in the abstract came across as less polished than the rest of the paper.

On page 15, the first paragraph in the section “Classification of a new object” and the corresponding equation suggest that the algorithm will produce a likelihood ratio. While the authors mention that these prior odds can be combined with any prior value p, they quickly jump to plugging in a prior probability of 0.5 and discussing the results in terms of posteriors and classification decisions for the rest of the paper. I think this presentation style risks causing readers to overlook the importance and challenge of picking a prior, and may leave readers committing the common error of transposing conditionals in which LRs are misunderstood as posterior odds. This concern is accentuated by the second sentence in the conclusion section (pages 23-24) "Our novel approach combines fracture mechanics with statistics and machine learning to quantify the probability that two candidate specimens are a match.” The provided algorithm does not quantify the probability that two candidate specimens are a match. This would require authoritative prior odds. Rather, the provided algorithm assigns a probability of the evidence under the proposition that the fragments came from the same source and a probability of the evidence under the proposition that the fragments came from different sources.

The last paragraph on page 15 begins "In the absence of prior information of the probability of a match, we are using an equal prior (p=0.5).” I would suggest rephrasing to something like "For the purpose of illustration, we chose an equal prior (p=0.5). In an actual criminal or civil case, choosing a prior match probability would require carefully considering any other evidence or relevant information previously presented, but such considerations are beyond the scope of this paper.”

When evaluating how many degrees of freedom to use in section 3, the authors seem to focus on discrimination without even mentioning calibration of the outputs. Skipping calibration would be fine if the authors described the outputs from model fitting as scores, but some discussion of calibration should be provided when suggesting the outputs have some intrinsic probabilistic meaning. For instance, Figure 6 shows model outputs for the same K-1-1 data as varying by 5+ orders of magnitude when using 3 degrees of freedom versus 20 degrees of freedom. Calibration would consider the question “which of these values more accurately describes how many times more often the underlying matrix of correlations occurs among truly matching fragments than among truly non-matching fragments?”. 

On page 10: “Each wavelength on the fracture surface has a population, on the frequency domain H(f), which is acquired using Fast Fourier Transform (FFT) operator.” It’s not clear what the
“population” is.

On page 16: "... and these LR results can be incorporated into a framework for evaluating the strength of evidence under different sets of assumptions." This is an awkward phrase, since "strength of evidence" is generally used to refer to LR.

On page 19: "... there is not perfect separation between all image pairs for the matches and non-matches. This can be noticed on Figure 7 where some image pairs have a correlation coefficient of less than 0.50 for the two bands of frequency analysis." According to the first sentence in the section "Reproducibility of results," Figure 7 only includes results from true matches, so this figure does not seem to provide an example of imperfect separation.

On page 21: "Guided by the results of Figure 10, it is apparent that we need at least 5 to 6 images for adequate discrimination." What is the criteria for discrimination to be adequate? Is the intention to suggest that if I have 9 images available, I only really need to use 5 or 6 of them? It seems more straightforward and appropriate to state that discrimination performance appears to improve with additional images and suggest that method performance is characterized for a given case using the available validation testing performed with the same number of images as are available in the given case.

On page 24: "Near perfect discrimination was achieved..." The sample size (i.e., number of comparisons for each of the populations being separated) should be summarized again here.
Authors’ Responses to Review Comments
Nature Communications manuscript NCOMMS-21-41063

We are thankful for the thorough and objective reviews we received on the previous version of our manuscript. The reviewers identified several ambiguities on how we expressed our framework for forensic analysis of fractured fragments utilizing basis of fracture mechanics and statistical analysis. We believe that we have addressed all the raised questions within the paper, and provided point-by-point explanation for the changes, as elaborated below. We believe that we have ended up with a substantially improved version of our manuscript for which we are very grateful to the reviewers. Furthermore, we also provide a document that tracks the changes that we made to the manuscript.

Reviewer #1 (Remarks to the Author):

• The quality of the data — “Manner of data collection is well-described in the supplemental material, but the authors are encouraged to reference this in the actual manuscript since without this information, a reader will not realize that the supplement can be consulted for more details. Fig. 3 is heavily utilized to explain the analytical procedure. However, increased clarity is requested in a few places. First, there is no discussion of “alignment” in image pairs, but there is reference to the use of FFT information for image registration (yet it is not clear if the authors are using information from the power spectrum to achieve any form of registration).”

  (a) Indeed, we noticed that we were not referencing the details in the supplementary materials, as we should have. We have corrected this deficiency. Specifically, we have revised both the supplement (dividing into sections that we now reference in the main paper regarding sample generation, imaging, image processing, alignments and tabulated results.

  (b) Fig 3.: We have revised Fig.3 and explained each module in details in the paper. We have also added another figure from the appendix and brought it into the main paper to explain the required K-images per fracture pair and the N-fractures pair that we analyzed.

  (c) Image alignments: We have added a full section in the supplementary material regarding image alignments. Also we added short details within the manuscript.

  Image alignment is a critical issue for the generalizability of the technique when there is a lack of clear reference, i.e. in the case of circular object like a bullet casing. For the cases studied in this work, we started with rectangular cross-sections, which naturally provide a reference coordinate. We utilized the sample edge as a reference and aligned the edges with the imaging field of view of the microscope. Then we moved a predetermined distance in the planner x- and y-coordinate to start the imaging sequence. Now there are three types of misregistration that can greatly affect the correlation estimations between a pair of images. Some of the issues encountered during the utilization of virtual comparison microscopy are:

    (i) Planner misregistration in the x- and y-coordinates. This is a very critical issue when comparing the images in their spatial (real) space, wherein an auto correlation must be utilized to adjust the
planar miss-shift. However, the procedure implemented in this work to utilize the spectral (frequency) space is very tolerant to planar miss-registration of up to 20% of the field of view. Actually, we have examined different planar-shift up to 400% of the FOV, and we found that the correlation would degrade by about 10% for up to 50% shift for the FOV.

(ii) Angular misregistration within the plane of observation (fracture plane). This is more critical when considering a set of overlapping k-images, wherein some image pairs may have reduced overlap. For our work, this was not an issue as we were studying fragments with good references for alignments. Beyond the presented work in this manuscript, we are implementing an alignment technique using the Fourier spectra on angular arcs within the two comparison bands to explore angular misregistration that would maximize the correlation coefficients for a pair of images. Then we run the same process for all k-image pairs on the pair of fragments to arrive at a global misregistration angle for the two surfaces. This process is conducted for pairs of fragments, regardless of their classification as true match or true non-match.

(iii) Fracture plane tilt: This is the case with a tortious fracture surface, encountered in semi-brittle or ductile material classes. In such case of tortuous path, the topological details of the surface will have high aspect ratio that might eclipse/ shadow other features. Furthermore, if large plasticity exists on the surface and the surface topology changes within several hundred microns with non-planner mean, acquiring 3D topological images would be much harder and additional mathematical treatment would be required, similar to comparison of cylindrical surface (e.g. cartridge cases).

“A possible major detractor is the use of training versus testing data. After reading it is unclear if the image pairs used for model training are the same pairs used for object classification. If the sets are the same, then the reported classification accuracy may be inflated and independent testing/training sets are needed. If they are not the same, the reader would benefit from a very clear description of this difference.”

We have clarified this issue and now exclude the training data when testing and are explicit about this in our description of the testing methodology. It should be also noted that, as we have further tested different classes of materials including ceramic articles, we were able to achieve the same level of discrimination while training on metallic articles with the same average grain size as the ceramic articles.

• The level of support for the conclusions — “Support of the utility of the method is largely a function of the testing/training issue formerly discussed. If dependent, then the results are possibly inflated and a second independent testing set should be added.”

We have addressed this in our response to the point above, wherein we used training sets on metallic rods to discriminate knife fragments and vis-versa. Furthermore, as noted above we can train on metallic samples and predict the ceramic fragment behavior as long as we have the same range of grain sizes.

• The potential significance of the results — “Response: To the reviewer’s knowledge, the exact approach used in this study has not been applied to forensic toolmark analysis in the past. Thus, the results are interesting and important. However, the use of objective methods to support toolmark associations is not a novel concept, and several other methods exist in the primary literature. In terms of a comparison of the proposed method versus those that already exist, since none have been compared using the exact
same datasets, doing so would naturally be the next step in order to determine if the proposed work surpasses the performance of other objective metrics.”

We totally agree with the reviewer that there are currently many objective methods that are being implemented, especially with 3D topological microscopy are becoming mainstream. Many of these objective methods rely on calculating a correlation score between two images or image subsets, and assign a score for the probability of the match. There are several unique attributes of the proposed objective method described in this paper. First, it is invoking the foundation of fracture mechanics to identify the proper length scales for proper objective comparison. These fracture-derived length scales define both the proper field of view for comparison and the required resolutions to identify the unique features on the surface. These details are typically handled on a trial and error bases. Second, building the decision process on a set of independent measurements on the surface and thereby enhancing the odds of the process identification and render the process more tolerant to measurements artifact. Third, using a set of k-images on the two surfaces to be more tolerant to topological inconsistencies due to missing grains and features on a pair of comparative surfaces and coupled with a statistical framework for decision-making.

At this stage, the paper focuses on the fundamental of the process and presenting its potential. A comparative study of other methods will be a great and worthy task for future work.

• Key results—“Response: By collecting non-contact 3D images of fracture surfaces, followed by computation of the correlation coefficient in the Fourier space (power spectrum of FFT), selected frequency bandpasses can be used to describe the similarity of mated and non-mated pairs. The distribution (density) of these similarities can be modeled using a matrix-variate t distribution, which can be used to generate two conditional probabilities describing the likelihood of correlations being sampled from the mated and non-mated pairs. The LR (log LR) information can be transformed into a binary classification system, and false positive and negative rates generated in order to evaluate the merits of the proposed approach.”

We would like to thank the reviewer for grasping the main essence of the proposed framework and providing an excellent summary of the key results.

• Validity—“Response: The reviewer was unable to resolve the use of the data sets. The abstract indicates fifty, and the text indicates two sets of 9 and two sets of 10 (18+20). The training illustrations/figures report at least one set of knives being used, 81 known matches (9 sets x 9 images). If the authors could clearly delineate testing/training/dependence/independence (which they do for a hold-one-out and a repeated imaging scenario) the reader could better evaluate the validity of the conclusions.”

(a) The text has been amended to say, “Four different sets of samples were established with nine specimens in the two sets of knives and ten specimens in the two sets of steel rods for a total of 38 specimens.” The abstract has been amended to say thirty-eight instead of fifty.

(b) We have changed the method and updated the results in the paper so that we no longer include the training sets in the tests and amended the following text to clarify in Section 3: “Figure 6 shows the classifications obtained by training on each of the four datasets, represented by one of the color boxes, with all 9 images per sample and classifying on all the other sets of surfaces using the matrix-variate t distribution and a common degrees of freedom parameter, \( \nu = 3, 5, 10, 15, 20, \)
and 30, and prior probability of being a match of 0.5 (for example, training on the first set and testing on sets 2, 3, and 4, and continuing the same process with the other sets as the training set).”

• **Significance** — “Response: The authors report an objective and numerical method for fracture matching that has not (to the reviewer’s knowledge) been presented in the field of forensic to date. This is significant and informative, and would benefit from a future study illustrating how amenable it is to solving these types of problems when presented with additional datasets and material types.”

We appreciate the reviewer’s kind remarks and comments.

• **Data and methodology**— Your assessment of the validity of the approach, the quality of the data, and the quality of presentation. We ask reviewers to assess all data, including those provided as supplementary information. If any aspect of the data is outside the scope of your expertise, please note this in your report or in the comments to the editor. We may, on a case-by-case basis, ask reviewers to check code provided by the authors (see this Nature editorial for more information).

Response: The datasets are relatively small (at most 50 pairs). Clarity is needed to better understand training/testing question previously discussed. Some aspects of the methodology are unclear (alignment, etc.), and have been annotated using PDF Adobe comments in the actual manuscript. This reviewer was not familiar with density estimation using the prescribed matrix of correlation scores, but found the author’s description useful. The authors spent considerable time discussing tuning for degrees of freedom, but this detail is definitely outside the scope of my expertise and little comment (constructive or critical was offered in this section of the manuscript).

Following the previous comments, we have (a) clarified the sample size was 38 pairs, (b) added full section about the alignment in the supplementary materials, (c) addressed all the comments from the PDF adobe as noted below. Selecting the DF’s was one of the main issues we investigated to be sure of the model outcome.

• **Analytical approach** — Response: Limited statistical testing was performed (i.e., the modified KS test seemed valid as used). Instead, the focus was on ratios of conditional probabilities, and if binarized into a classification threshold, false positive and negative rates. These seemed fairly standard and appropriate.

We appreciate the reviewer’s comments.

• **Suggested improvements** — “Critical flaws were not identified, however, various clarifications are requested in order to improve the reader’s ability to understand nuances of the paper. These have been itemized above, and/or as actual comments in the manuscript --- and represent copyedit changes rather than substantial methodology corrections.”

We have made numerous change to the paper to respond to the reviewer’s comments. The details are given below in our point-by-point response to the comments that were given in the manuscript pdf file.
We have made numerous changes to the paper to respond to the reviewer’s comments. The details are given below in our point-by-point response to the comments that were given in the manuscript pdf file.

• Your expertise — “Response: I was able to provide constructive feedback as a “typical forensic” reader. Although I am versed in image analysis, microscopy, error rates, Fourier transforms, extraction of information from the power spectral density, the use of correlation coefficients, and density estimation, I am not a statistician, and aspects of the density estimation modelling (selection of DF) are outside of my area of expertise.”

Thanks you for your comments and suggestions. We have tried to improve the clarity of our presentation in places where you asked questions.

Responses to Reviewer 1’s comments in the manuscript pdf

We have corrected the sample set to be composed of four sets of 38 broken pairs in total.

We have changed to “microscopical”
"In the revision we have rewritten this paragraph so that the argument is not one-sided. "However, it should be noted that a considerable amount of prior work has been done to provide a quantitative and scientific basis for firearm and toolmark identification, for example, with the consecutive matching striae (CMS) method."

Figure 2(a) shows the topology. Figure 2(b) shows the height-height correlation which can identify this scale. We have clarified the scale and correlate it with the grain size in the paper.

In the revision we now say “The creation of this model can also be used to estimate probabilities of misclassification and compare to the empirically observed rates of misclassification. Conceptually, this is similar to forensic matching..."
Forensic fragments that might require surface comparison for identification encompasses a wide range of articles and material classes, including metallic, plastics and ceramics and glasses. Here we focused on metallic fragments and employed a common stainless steel rods, which cover a wide class of tool steels and pry bars, as well as knives. For developing the technique, we used these two classes of materials to show its validity and potential. At a later step of the work, a typical forensic fragment including metallic, plastics and ceramics could be examined and evaluated. However, at this stage we showed the potential of the work based on the studied articles. We have added the following statement at the beginning of the paragraph “To mimic forensic articles found in a crime sense and might undergo a comparative analysis, we consider…….”

with nine specimens in the two sets of knives and ten specimens in the two sets of steel rods. Each knife specimen was fractured at random, in a manner similar to Figure 1(a).

We now recognize the confusion. We have amended the text to read; “To mimic forensic articles found in a crime sense and might undergo a comparative analysis, we consider two main material classes: sets of rectangular rods of a common tool steel material (SS-440C) fractured under control tension and bending configurations (Figure S.2(b, d)), and sets of knives from the same manufacturer (Figure S.1(b), fractured at random employing the fixture shown in Figure S.1(a). Figure 1(a) shows a typical pair of fragments, generated for this study. The average grain size for both groups was approximately $d_g =25–35 \, \mu m$. Four different sets of samples were established with nine specimens in the two sets of knives and ten specimens in the two sets of steel rods. Further details about sample preparations are given in Section”
We see how these statements confused the reviewer. FFT is utilized to identify the proper bands for comparison as well as its tolerance for alignment within the plane of the image. We have added a full section about alignment in the supplementary materials. Also, we revised the entire section to remove the ambiguity “A series of k-overlapping surface height 3D topographic maps were acquired from the pairs of fracture surfaces (Figure 3; k = 9), and quantized using Fourier transform based spectral analysis as summarized in Figure 4 in (a) the image analysis step. The choice of overlap means there are three full independent sequential images on a surface. Multiple overlapping images were needed to overcome problems arising from missing grains between pairs of the fracture surface and/or the special circumstances of complex tortuous path of fracture. The effect of the number of images and overlapping ratio will be further discussed in Section 3. Additionally, having a super-image of stitched FOVs results in misregistration at the overlapping boundary of the stitched images, leading to an additional interfering frequency within the band of comparison”

Figure 3: Flow chart showing the image analysis steps and model fitting/calibration step, followed

Registration is achieved by picking a reference point relative to the sample edges. We use the sample edges for alignment with the microscope field of view. A detailed section is added in supplementary materials to cover these issues.
We have corrected these in the revision.

For a pair of fractured surfaces, the population of these features contains relevant information about the physical fracture processes present at each length scale (e.g. dimples and voids at the sub-microscale, and river marks at scales of tens of microns).

2. Compute the correlations of the frequency bands for the sets of images for an matching and non-matching surface pairs.

3. Use the Fisher’s z transformation on the correlation data to stabilize variance.

4. Fit models to describe the distribution of true matches and true non-matches, which account for the imperfect matching of features.

To respond to the model comment, the text has been amended to clarify that the model will be described in the next few paragraphs: “Fit models using a matrix-variate distribution as detailed later in this section to describe the distribution...”
The text has been amended to say: “Figure 5. The proposed method's discrimination ability can be judged from the clear separation of matching and non-matching surfaces within two separate clusters. The data in this illustration were derived from $N = 9$ base-tip pairs from fractured knives. A series of $k = 9$ overlapping images were taken from each base and tip fracture surface, resulting in $N \times 2k = 162$ total images (81 from the tips and 81 from the bases). Additional details are given in the S.1 for different data sets. In this example, image pairs for when the tip and base surfaces were from the same knife are true matches ($N \times k = 81$ matched-pairs), while those pairs for which the tip and base surfaces were from different knives are true non-matches ($N (N - 1) \times k = 648$ unmatched-pairs).”
non-matches, taking an ensemble of images from the surface helps perfectly discriminate the two classes.

What we meant by “ensemble of images” is that we have a series of images from the surface and we use a model which combines information from the multiple images. We have reworded this part of the paper to be clearer and made clarifications in other parts to highlight that this is how they are being used. 

To further improve the discrimination, considering multiple k-observations from the same surface would distinguish it from the non-matches, since the other observations on that surface are well-separated from the non-matches. In the current framework, we take the information from every pair of images and collectively based on the model, a decision is driven accounting for the fact that the images are not independent, i.e. overlapping and coming from the same fracture surface. Role of imaging repetition or overlap, may improve the signal to noise ratio.

Accordingly, we propose using a matrix-variate distribution to model the densities of the matching and non-matching populations and specifically, a matrix-variate distribution (MVMD).

This has now been included in part 4 on page 12. “Fit models using a matrix-variate distribution as detailed later in this section…”

prior is used, this expression can also be converted to a likelihood ratio (LR), which is a common method in forensic applications, and these LR results can be incorporated into a framework for evaluating the strength of evidence under different sets of assumptions. Classification decisions can then be made under the rules of evidence relevant to the case. In this discussion, we will make classification decisions using a cutoff value of 0.5.

The text has been amended to say, “In this discussion, we will make classification decisions using a cutoff value for the probability of 0.5$ (or a likelihood ratio of 0).”
As mentioned above, we have changed the methodology to exclude the set used for training from the tests and amended the text to describe the new results: “Figure 6 shows the classifications obtained by training on each of the four datasets, represented by one of the color boxes, with all 9 images per sample and classifying on all the other sets of surfaces using the matrix-variate $t$ distribution and a common degrees of freedom parameter, $\nu = 3, 5, 10, 15, 20,$ and 30. The output given in terms of the log-odds of being a match – log-odds larger than zero ($p = 0.5$) indicate classification as a match. While initially there are no false positives or false

needed to optimize classification performance, we started by training models using all nine images on each training set as before. We again used the MxVT model with $\nu = 3, 5, 10, 15, 20,$ and 30, and then tested them on subsets of consecutive images of size $k$, for $k = 2, 3, \ldots, 9$ with the model reduced to considering only the selected images and the training set for each model excluded from testing.”
We have changed the method and amended the text clarify that we exclude the training set from consideration when testing: “We train classifiers on the same sets as before, except using 5 images with 50% overlap instead of 9 images with 75% overlap and then test the models by classifying pairs of surfaces using all possible subsets of those images on the surface of sizes 2, 3, 4, and 5. When restricted to the case of 50% overlap, there is only...
Among the broad range of training sample sets, this domain of distinctive individuality was found

We have removed the word “broad” and made other changes to the wording to address this comment “Among the range of training sample sets, this domain of distinctive individuality was found to be persistent and easily identified.”.

Reviewer #2 (Remarks to the Author):

In "Quantitative Matching of Forensic Evidence Fragments using Fracture Mechanics and Statistical Learning," Thompson and co-workers assess whether two fracture surfaces can be recognized as paired using comparative microscopy and physical pattern analysis. Their contribution to this field is to identify an optimal range of search region sizes. The work is important and well executed, and the reviewer is enthusiastic.

Thank you for your kind remarks on our paper.

The reviewer recommends addressing the following suggestions for the final version of this manuscript.

(1) The fracture process zone

The linkage to fracture mechanics process zone in the article is limited to the assertion that "It is established in fracture mechanics that the fracture process zone ahead of the crack tip typically extends to 2-3 times the grain size." This is attributed to Anderson's book, but the reviewer had trouble finding this statement in the text, and suspects that it might be a rule of thumb. This is certainly not the case for highly ductile metals or single crystal turbine blades. Would it make sense to speak in terms of grain size in the paper, and then perhaps link to the fracture mechanics justification in the discussion? Otherwise, the assertion that this is based upon fundamental fracture mechanics probably requires a notch fracture test and careful strain analysis of the specific metals used to quantify the process zone size.

The authors would like to thank the reviewer for her/his insight. We have omitted the reference to fracture process zone in the overview section and referenced everything in terms of grain sizes. Then we added a full section in the result and discussion "Imaging scale for comparison" where we discussed the link to fracture mechanics of brittle materials, "We found that this scale is about 2-3 times the average grain size of the material. Interestingly, this is the average fracture process zone in brittle and semi-brittle material wherein the crack tip process zone typically extends to 2-3 times the grain size, or around 50–75 μm for the tested material system. This is the scale wherein the intrinsic material resistance to fracture is reached. Accordingly, the surface characteristic becomes unique and non-self-affine at a larger scale, where it is eclipsed by the interference of the fracture process zone length scale."
(2) Language in the paper
- The language is loose in places, and the reviewer suggests that it be tightened. The whole paper could use trimming and rewording.

The paper has been revised and many places were rewarded as can be seen in the difference-file

- The following statements need to be fixed: -- "The individual characteristics of the fracture process leave surface marks on both surfaces that can be identified in order to match fragments to each other reliably." (This is an imprecise statement)

We have revised this statement to read: “We focus on fracture matching, the forensic discipline of determining whether two pieces came from the same fractured object. The fracture mechanisms leave surface marks on both surfaces that could be utilized for matching fragments.”

- “This is the scale wherein the local stresses ahead of the crack tip reach a critical level, sufficient to overcome the intrinsic resistance of the material to fracture” (The authors, who include experts in fracture, will recognize this statement as being incorrect)

The authors acknowledges the strict definition of the fracture process zone. The statement has been revised to read; “This is the scale wherein the intrinsic material resistance to fracture is reached”

(3) The optimization of scanning area was not clear. Given that this is a key contribution of the paper, would it make sense to add a figure showing this explicitly?

We have added three sections to address the area scanning and its implications:

(a) in Method Overview section “The height-height correlation function at this transition scale captures the uniqueness of the fracture surfaces, so we can use that function’s behavior in setting the observation scales (i.e., field of view, FOV, and imaging resolution) for comparing matching and non-matching surfaces to produce a statistical model of each topological class’s behavior for use in classification. This imaging scale should be greater than about 10-times the self-affine limit scale to avert signal aliasing. Further, we can further combine multiple observations at different length scales or topological frequencies (around the self-affine saturation scale and readily segmented in the frequency space) of a single surface into one model in order to improve the ability to discriminate between surfaces of the same class or materials and manufacturing processes (for instance, individualization of a pry tool from a similar batch of identical tools).”

(b) In “2. Materials and Method: Sample Generation and Imaging” Section: “Utilizing the results of the height-height correlations of Figure 2(b), the transition scale commences at around 50–70 μm to become non self-affine and saturate, rendering a required imaging FOV of about 500μm. For the examined material systems, this scale amounts to 2–3 times the grain size, and the FOV should cover 20–30 grain-diameters. Accordingly, an optical magnification of 20X is employed, providing a 550 μm FOV and 0.55 μm/pixel resolution (Figure 2(a)).

(c) In “ 3. Results and discussion: Imaging Scale for comparison” Section: “When comparing characteristic features on a fractured surface, identifying the proper magnification and FOV are critical. An optical image obtained by high magnification and a small field of view, will possess an indistinguishable characteristic. This is the range where surface roughness shows self-affine nature as noted on Figure 2(b). In this range, the material intrinsic local fracture mechanism show similar topological surface features over the fractured surface (e.g. local dimples and voids). On the contrary,
employing lower magnifications will result in a lower power of identifying the class-characteristics of the surface. However, we showed that the transition scale of the height-height correlation function captures the uniqueness of the fracture surfaces. We found that this scale is about 2-3 times the average grain size of the material.”

(4) The metrics were for comparison were not clear. The reviewer was unfamiliar with the idea of "Log Odds" in Figure 6. What do they mean, and why not used percentage likelihood of a false positive match? Could the latter be added?
In the revision we have explained the relationship between log-odds and added indicators of probability to the plots (Figures 7, 9, and 11).

(5) Figure 9, and the false negative rate: Why are false negatives more likely with increasing numbers of images?
At 7, 8, and 9 images, we only have 1 false negative in those cases (see table S.1). The rate appears to increase because the denominator is decreasing. There is still a question of why there is 1 rather than zero for these values. We have amended the text to give a little context here: “For values of k > 4, only 20 and 30 DF have false negatives (specifically, they each have one false negative result).”

Congratulations on this important and interesting work!
Thank you again for your comments.

Reviewer #3 (Remarks to the Author):

Summary:
This paper’s primary contribution is an algorithm for evaluating whether two fragments originated from a common source before a fracturing event. Additionally, the manuscript proposes models to represent the distribution of matrices of correlations between images evaluated at different regions and length scales of a given fracture. The article also examines the influence of various modeling choices and algorithm settings on discrimination performance and is shown to perform well across a range of configurations. I believe this work helps address concerns raised by the NAS report “Strengthening Forensic Science in the United States: A Path Forward” and will be of interest to the forensics community. I recommend that this paper be published following minor revision.

The authors would like to than the reviewer for the kind comments.

Comments:
(1) The writing in the abstract came across as less polished than the rest of the paper.

The abstract has been reworded.

(2) On page 15, the first paragraph in the section “Classification of a new object” and the corresponding equation suggest that the algorithm will produce a likelihood ratio. While the authors mention that these prior odds can be combined with any prior value p, they quickly jump to plugging in a prior probability of 0.5 and discussing the results in terms of posteriors and classification decisions for the rest of the paper. I think this presentation style risks causing readers to overlook the importance and
challenge of picking a prior, and may leave readers committing the common error of transposing conditionals in which LRs are misunderstood as posterior odds. This concern is accentuated by the second sentence in the conclusion section (pages 23-24) “Our novel approach combines fracture mechanics with statistics and machine learning to quantify the probability that two candidate specimens are a match.” The provided algorithm does not quantify the probability that two candidate specimens are a match. This would require authoritative prior odds. Rather, the provided algorithm assigns a probability of the evidence under the proposition that the fragments came from the same source and a probability of the evidence under the proposition that the fragments came from different sources.

We have rewritten this section to emphasize that the use of 0.5 as a prior probability is illustrative, that a choice of prior will depend on many other factors, and that instead a likelihood ratio can be produced: “For the purposes of illustrating the method, we are using an equal prior probability of being a match or non-match (i.e., p=0.5 or a log-likelihood ratio of 0). In an actual criminal or civil case, choosing a different prior match probability would require carefully considering any other evidence or relevant information previously presented, but such considerations are beyond the scope of this paper. The sentence in question from the conclusion now also highlights the importance of the choice of prior: “Our novel approach combines fracture mechanics with statistics and machine learning to quantify, given a prior probability, the posterior probability that two candidate specimens are a match.”

(3) The last paragraph on page 15 begins “In the absence of prior information of the probability of a match, we are using an equal prior (p=0.5).” I would suggest rephrasing to something like “For the purpose of illustration, we chose an equal prior (p=0.5). In an actual criminal or civil case, choosing a prior match probability would require carefully considering any other evidence or relevant information previously presented, but such considerations are beyond the scope of this paper.”

We have rewritten the text to incorporate this.

(4) When evaluating how many degrees of freedom to use in section 3, the authors seem to focus on discrimination without even mentioning calibration of the outputs. Skipping calibration would be fine if the authors described the outputs from model fitting as scores, but some discussion of calibration should be provided when suggesting the outputs have some intrinsic probabilistic meaning. For instance, Figure 6 shows model outputs for the same K-1-1 data as varying by 5+ orders of magnitude when using 3 degrees of freedom versus 20 degrees of freedom. Calibration would consider the question “which of these values more accurately describes how many times more often the underlying matrix of correlations occurs among truly matching fragments than among truly nonmatching fragments?”.

We have added a new subsection to the Results section, “Calibration of output probabilities” to address this concern and evaluate the calibration of all the predictions made by all of the models. The vast majority of the model classifications are correct classifications with probabilities of being a match of either <0.001 for non-matches or >0.999 for matches. The relative lack of samples in the middle range makes it hard to judge the calibration. The lowest probability of a match among the true matches was 0.3709. Among the various models, the 99th percentile of the predictions for non-matches was, in the worst case, 0.1437. Only outliers overlapped in middle range. We note that our evaluation of the calibration is limited by the sample size in the experiment---with more samples and more observations with match probabilities between 0.1 and 0.9, a better evaluation of the calibration could be made.
(5) On page 10: “Each wavelength on the fracture surface has a population, on the frequency domain H(f), which is acquired using Fast Fourier Transform (FFT) operator.” It’s not clear what the “population” is.

We changed “population” to “distribution”: “H(f), which is acquired using a Fast Fourier Transform (FFT) operator. For example, grain size has a distribution of frequencies across the spectrum rather than one specific frequency. Similarly, other microscopic fracture features have a range of spectral distributions50, 51.”

(6) On page 16: “…and these LR results can be incorporated into a framework for evaluating the strength of evidence under different sets of assumptions.” This is an awkward phrase, since “strength of evidence” is generally used to refer to LRs.

The intention here is to note that we can consider this model and compare to other models – we make several parametric assumptions here, such as using an AR(1) model, using a t-distribution, whereas other choices would result in different likelihood ratios, though this is perhaps better expressed as a statement about uncertainty about the strength of the evidence. We have amended the text to say, “and these LR results can be incorporated into a framework for expressing the uncertainty about the strength of evidence under different sets of assumptions.”

(7) On page 19: “…there is not perfect separation between all image pairs for the matches and non-matches. This can be noticed on Figure 7 where some image pairs have a correlation coefficient of less than 0.50 for the two bands of frequency analysis.” According to the first sentence in the section “Reproducibility of results,” Figure 7 only includes results from true matches, so this figure does not seem to provide an example of imperfect separation.

This should refer to figure 4, not figure 7. This has been fixed in the text.

(8) On page 21: “Guided by the results of Figure 10, it is apparent that we need at least 5 to 6 images for adequate discrimination.” What is the criteria for discrimination to be adequate? Is the intention to suggest that if I have 9 images available, I only really need to use 5 or 6 of them? Its seems more straightforward and appropriate to state that discrimination performance appears to improve with additional images and suggest that method performance is characterized for a given case using the available validation testing performed with the same number of images as are available in the given case.

We have amended this statement to say, “Guided by the results of Figure 10, it is apparent that we need at least 5 to 6 images for error-free discrimination in the sample sets and performance improves with additional images.”

(9) On page 24: “Near perfect discrimination was achieved…” The sample size (i.e., number of comparisons for each of the populations being separated) should be summarized again here.

We have amended this to clarify, “Near-perfect discrimination was achieved in the four training sets totaling 38 samples to classify...”
REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

I uploaded the PDF with inserted comments (seemed most efficient manner to communicate since manuscript pages do not have line numbers). The authors provided a detailed response to original comments which has clarified several items. A few minor issues/copy edits were noted and are identified in the attached/uploaded PDF.

Reviewer #2 (Remarks to the Author):

The manuscript is much strengthened in the current revision.

It would still benefit, however, from reframing, rewriting, and additional explanation.

The core of the manuscript is use of "Fracture Mechanics and Statistical Learning," but the use of fracture mechanics is entirely limited to the following sentence:

"We found that this scale is about 2-3 times the average grain size of the material. Interestingly, this is the average fracture process zone in brittle and semi-brittle material wherein the crack tip process zone typically extends to 2-3 times the grain size, or around 50–75 µm for the tested material system."

I noted last time that this rule of thumb was incorrectly attributed to Anderson's text (ref 36). The authors have removed the attribution and listed the reference in the next sentence ("This is the scale wherein the intrinsic material resistance to fracture is reached.36"; the authors are aware that this is also an incorrect description of what a fracture process zone is, and it needs to be corrected.) The link to fracture mechanics does not appear strong enough for it to appear in the title or abstract as a foundation of the work. The presentation of this rule of thumb as a foundation for a criminal conviction needs to be reconsidered; this reviewer will push back against this idea without substantially stronger evidence from the authors. The authors have looked at one stainless steel and one other steel, and found that a couple of grains is a "big enough" search region for these two. However, any use of the authors' approach for a different metal would seem to require experimental recalibration; the state of the art in fracture mechanics does not, to this reviewer's knowledge, allow one to predict the uniqueness of a process zone in a metal without first performing an experiment.

Additional caveats would seem to be that the mode of loading (in terms of linear elastic fracture mechanics, the authors seem to have looked at a mode I loading only) and the presence of pre-existing flaws could fundamentally change the conclusions of the authors. This needs to be articulated clearly.

Finally, the statistics in the revised figures seem to indicate that there are ranges of applicability of the authors' approach. The empirical probabilities shown in Figure 13 that are greater than one and less than zero require explanation. More broadly, articulating cases such as these where the work of the authors breaks down to give unreliable predictions would seem to be an important addition.

Reviewer #3 (Remarks to the Author):

The authors have largely addressed the comments from my previous review. However, a couple comments still apply to the revised manuscript. I retain the recommendation that this paper be published following minor revision.

Please see the attached pdf for detailed comments.
The authors have largely addressed the comments from my previous review. However, a couple comments still apply to the revised manuscript.

(2) On page 15, the first paragraph in the section “Classification of a new object” and the corresponding equation suggest that the algorithm will produce a likelihood ratio. While the authors mention that these prior odds can be combined with any prior value \( p \), they quickly jump to plugging in a prior probability of 0.5 and discussing the results in terms of posteriors and classification decisions for the rest of the paper. I think this presentation style risks causing readers to overlook the importance and challenge of picking a prior, and may leave readers committing the common error of transposing conditionals in which LR\( s \) are misunderstood as posterior odds. This concern is accentuated by the second sentence in the conclusion section (pages 23-24) “Our novel approach combines fracture mechanics with statistics and machine learning to quantify the probability that two candidate specimens are a match.” The provided algorithm does not quantify the probability that two candidate specimens are a match. This would require authoritative prior odds. Rather, the provided algorithm assigns a probability of the evidence under the proposition that the fragments came from the same source and a probability of the evidence under the proposition that the fragments came from different sources.

We have rewritten this section to emphasize that the use of 0.5 as a prior probability is illustrative, that a choice of prior will depend on many other factors, and that instead a likelihood ratio can be produced: “For the purposes of illustrating the method, we are using an equal prior probability of being a match or non-match (i.e., \( p = 0.5 \) or a log-likelihood ratio of 0). In an actual criminal or civil case, choosing a different prior match probability would require carefully considering any other evidence or relevant information previously presented, but such considerations are beyond the scope of this paper. The sentence in question from the conclusion now also highlights the importance of the choice of prior: “Our novel approach combines fracture mechanics with statistics and machine learning to quantify, given a prior probability, the posterior probability that two candidate specimens are a match.”

The sentence added at the very end of Section 2 on page 18 has been modified from what was suggested by changing the phrase “... choosing a prior match probability” to “... choosing a different prior match probability.” Inserting the word “different” substantially and adversely affects the meaning of the suggested sentence. The inserted sentence suggests one can assume equal prior probabilities without carefully considering other evidence or relevant information previously presented, which is equivalent to saying one can, by default, interpret likelihood ratios as posterior odds - the exact erroneous sentiment the suggested sentence was supposed to help prevent.

Additionally, the second to last sentence in section 2 on page 18 now reads:

“For the purposes of illustrating the method, we are using an equal prior probability of being a match or non-match (i.e., \( p = 0.5 \) or a log-likelihood ratio of 0).”

The phrase “log-likelihood ratio” should be replaced with “log prior odds.” The current phrasing commits the error of transposing the conditional (treating hypothesis odds as equivalent to a likelihood ratio).
(8) On page 21: “Guided by the results of Figure 10, it is apparent that we need at least 5 to 6 images for adequate discrimination.” What is the criteria for discrimination to be adequate? Is the intention to suggest that if I have 9 images available, I only really need to use 5 or 6 of them? It seems more straightforward and appropriate to state that discrimination performance appears to improve with additional images and suggest that method performance is characterized for a given case using the available validation testing performed with the same number of images as are available in the given case.

We have amended this statement to say, “Guided by the results of Figure 10, it is apparent that we need at least 5 to 6 images for error-free discrimination in the sample sets and performance improves with additional images.”

Page 25 contains the sentence: “All of this suggests that choosing a value near \( v = 10 \) and \( k \geq 5 \) images is sufficient for classification.” This sentence should also be modified accordingly.
Authors’ Responses to Review Comments
Nature Communications manuscript NCOMMS-21-41063A

We are thankful for the additional inputs the reviewers have provided. We have addressed all the raised questions within the paper, especially those of Reviewer 2, and provided point-by-point explanation for the changes, as elaborated below. Furthermore, we have also provided a file with tracked changes that we made to the manuscript.

Reviewer #1 (Remarks to the Author):

I uploaded the PDF with inserted comments (seemed most efficient manner to communicate since manuscript pages do not have line numbers). The authors provided a detailed response to original comments which has clarified several items. A few minor issues/copy edits were noted and are identified in the attached/uploaded PDF.

Thank you for your comments. All Suggested corrections have been addressed in the revised manuscript.

Reviewer #2 (Remarks to the Author):

The manuscript is much strengthened in the current revision.

It would still benefit, however, from reframing, rewriting, and additional explanation.

The core of the manuscript is use of "Fracture Mechanics and Statistical Learning," but the use of fracture mechanics is entirely limited to the following sentence:

"We found that this scale is about 2-3 times the average grain size of the material. Interestingly, this is the average fracture process zone in brittle and semi-brittle material wherein the crack tip process zone typically extends to 2-3 times the grain size, or around 50–75 \( \mu \text{m} \) for the tested material system."

I noted last time that this rule of thumb was incorrectly attributed to Anderson's text (ref 36). The authors have removed the attribution and listed the reference in the next sentence ("This is the scale wherein the intrinsic material resistance to fracture is reached.36"; the authors are aware that this is also an incorrect description of what a fracture process zone is, and it needs to be corrected.) The link to fracture mechanics does not appear strong enough for it to appear in the title or abstract as a foundation of the work. The presentation of this rule of thumb as a foundation for a criminal conviction needs to be reconsidered; this reviewer will push back against this idea without substantially stronger evidence from the authors. The authors have looked at one stainless steel and one other steel, and found that a couple of grains is a "big enough" search region for these two. However, any use of the authors' approach for a different metal would seem to require experimental recalibration; the state of the art in fracture mechanics does not, to this reviewer's knowledge, allow one to predict the uniqueness of a process zone in a metal without first performing an experiment.
The authors would like to thank the reviewer for his/her insight. We have revised the manuscript title and focused the manuscript on the observation scale for comparison which is derived from the transition of roughness to become non-self-affine. We showed that such scale has been linked to both the material fracture toughness and to the intrinsic characteristic length ahead of the crack tip for a wide range of materials.

“The fracture surface topography contains many unique features over a wide range of length scales, which generally provide substantial information on damage initiation and propagation. The material microstructure controls the micromechanisms of fracture and the microscopic crack growth path, while the loading direction sets the macroscopic crack trajectory. Mandelbrot et al. first showed the self-affine nature of fractured surfaces and relating its roughness via the exponent characterizing the scale invariant properties ‘fractal dimension’ to the material resistance to fracture. The self-affine nature of the fracture surface roughness has been experimentally verified for a wide range of materials (metals, ceramics and glasses) and static and dynamic loading conditions. A key finding is the variation of such surface descriptors when measured parallel to the crack front and along the direction of propagation. The cut-off length scale of the self-affine behavior was suggested as a unique length scale to characterize the microscale fracture process in ductile and brittle/semi-brittle materials.”

For the specific case of hardened metals that undergoes cleavage fracture, we noted that the transition length scale of 2-3 grain size derived from the non-self-affine length scale is similar to the characteristic distance proposed by Ritchie, Knott, and Rice (1973, RKR), over which the local stresses ahead of the crack tip exceed the fracture stress. Furthermore, we cited the work of Curry and Knott (1978) to show the stochastic nature of the cleavage process due to the randomness of the critical fracture-triggering site. Such randomness further helps ascertain the individuality of each fracture surface. Further, we limited our discussion to the class of hardened material (e.g. cutting tools and pry bars) which will exhibit cleavage fracture. In the current work we have examined two different sets of similar class of materials, that is commercially available knives and 440 stainless steel, which would undergo cleavage fracture. We have added Figure 3(a) which shows the topology of the fractured knife in bending. In the section of the method overview we added the following discussion after the presentation of the transition length scale of 2-3 grain sizes.

“In the field of Fracture mechanics, it is postulated that cleavage failure occurs when the local stress ahead of the crack tip exceeds the fracture strength of the material over a characteristic distance, equal to two grain diameter. As noted earlier, the self-affine cut-off scale is actually the process zone scale in cleavage fracture of materials. Furthermore, the cleavage fracture process zone is statistical in nature, as a finite volume of the material ahead of the crack tip should include a local defect to nucleate cleavage crack. This very critical argument to show the statistical basis for the individuality of the fracture surface. Two nominally identical articles from the same material lot might show very different local fracture surface topology and fracture failure strength because of the statistical randomness of the microscopic spatial location of the critical fracture-triggering particle. These unique microscopic feature signatures exist on the entire fracture surface as it is influenced by three primary factors; namely the material microstructure, the intrinsic material resistance to fracture, and the direction of the applied load. This work explores the existence of such a length scale and the
We have identified that the critical length scale is those couple of grains, as observed from the roughness transition and the similarity to RKR theory. However, we do NOT image the sample at such scale. We claim that this is the characteristic scale that is being imprinted on the surface. To capture such scale, we have to at least capture 10 periods of such scale within the imaging field of view. As such the FOV is set to 20 grain diameters. We have clearly articulated tis in both the abstract and mentioned it again several times within the text. Specifically, we have added:

“ABSTRACT: For the examined class of hardened alloys typical for cutlery and tool steel, we found that this scale is about two grain-diameter, which is similar to characteristic distance for cleavage fracture resistance of materials, rendering the required imaging scale to be about 20 grain diameter.”

“Section2: Materials and Methods; Sample Generation and Imaging: For the examined class of hardened alloys typical for cutlery and tool steel, we found that this scale is about two grain-diameter, which is similar to characteristic distance for cleavage fracture resistance of materials, rendering the required imaging scale to be about 20 grain diameter.

However, any use of the authors' approach for a different metal would seem to require experimental recalibration; the state of the art in fracture mechanics does not, to this reviewer's knowledge, allow one to predict the uniqueness of a process zone in a metal without first performing an experiment.

We think there is a slight misperception because of our initial wording of this section. We are not claiming any uniqueness of the process zone, which we agree with the reviewer needs to be calibrated for each class of material. What we are claiming is the existence of a characteristic scale, which is stochastic in nature. When one captures enough average of such scale (about 10 periods), then frequency content of such scale is unique for such fracture surface. For example we refer the reviewer to Fig.7, where we find clear separation between the matched and non-matched pairs of images for the frequency range of 5-10 and 10-20 mm⁻¹. At frequencies higher than this range, the material intrinsic local fracture mechanism show similar topographical surface features over the fractured surface (e.g. local cleavage river patterns and/or dimples and voids). Further, we highlight that this scale is similar to that proposed by RKR theory of 2-3 grain diameter. To this extent, the only calibration needed is the identification of the roughness transition length scale by height-height correlation in order to set the imaging scale.

Additional caveats would seem to be that the mode of loading (in terms of linear elastic fracture mechanics, the authors seem to have looked at a mode I loading only) and the presence of pre-existing flaws could fundamentally change the conclusions of the authors. This needs to be articulated clearly.

We are thankful to the reviewer for raising this issue. We have performed additional set of twisting the knife to fracture to accommodate mode-I and mode-III loading. we performed the same imaging process and used the tensile and bending data set to train the model. Again we managed to discriminate the entire data set with very high probability. We have added Fig.3 to show the subtle differences in topology for bending and twist samples. We added a full section to examine the capabilities of the proposed framework, as follows:
“Examining the frame work capabilities on a twisted-fracture knife set All examined sets of fractured articles were tested in tension or bending. This is mode-I cleavage fracture where in the crack propagation direction is normal to the loading axis. The fracture surface showed topographical features normal to the fracture surface, similar to those shown in the SEM image of Figure 3(a). However for a general forensic article such as a knife or a pry tool, an edge could be broken due to bending and twisting of the article. This would impose a mixed mode of loading including mode-I opening and mode-III twisting of a crack. To understand the effect of external loading mode on the generality of the proposed analysis framework, as set nine knives from the same manufacturer as the previously used sets were fractured at random using the same fixture shown in Figure S1(b) and forming the twist test set shown in Figure S1(e). A typical twisted knife fracture topology is very different at both the macro and micro scales. At the macro scale, the crack trajectory is no longer planar with curvilinear or twisted trajectory as those shown in Figure S1(e). At the micro scale, the SEM image of Figure 3(b) shows twisted topology in the plane of the crack that is very different than those of mode-I loading. These unique texture would probably further enhance the individuality of the fracture surface. We will attempt to examine the validity of the analysis protocol on such general case of fractured articles. The twisted knife set was imaged using the same process as in Section 2 and the same magnification of 20X. However, due to the excessive tortuosity of the crack path, five images (k = 5) with 75% overlap between adjacent images were employed. Using the models previously trained in Section 2 on the four training sets loaded in tension or bending, and restricted to 5 images and setting the degrees of freedom v = 10. The results for this shown in Figure 15 are similar to those obtained in Figure 8 despite the use of a different external loading of mode-I tensile cleavage fracture. The true match cases where identified with probability exceeding 99.999% and the true non-match were identified with probability not exceeding 0.05% for all the different training sets. This suggests the scale of comparison, derived from the self-affine saturation scale of the surface topology and its correspondence to the critical distance for cleavage fracture for critical load to commence is more general and tied to the grain scale for hardened tool materials failing by cleavage fracture. This results is far more reaching. As long as the cleavage fracture is the dominant mode of failure, a single robust training data set under simplified loading conditions would be sufficient to help in discriminating; (i) articles that were exposed to complex external loading (i.e. mixed mode of fracture). (ii) articles from different classes of materials, but share the same grain size distributions, and (iii) articles with different grain sizes would only require changes of the FOV to include 20-grain diameter and comparison frequency bands corresponding to the 2–4 and 4–8 grain size ranges. It is conceivable to extend these results to glassy metals, polymers and ceramics, that undergoes cleavage and/or brittle or semi-brittle fracture. In such cases, the critical microstructure scale would be the characteristic fracture topology features such as river and herringbone patterns. Though, additional experimental verification are needed for these classes of non-crystalline materials.”
Figure 3: Detailed SEM analysis of the a typical fracture surface for (a) bent broken knife broken in bending, showing topological details normal to the imaging plane, and (b) twist broken knife in torsion, showing in-plane swirl textures.

Figure 15: A plot of the classification results for a generality verification data set of nine knives broken by twisting with five images per surface. The models were trained with on different training bending and tensile fracture sets using five images with the degrees of freedom setting $\nu = 10$. 
Finally, the statistics in the revised figures seem to indicate that there are ranges of applicability of the authors' approach. The empirical probabilities shown in Figure 13 that are greater than one and less than zero require explanation. More broadly, articulating cases such as these where the work of the authors breaks down to give unreliable predictions would seem to be an important addition.

We are very thankful to the reviewer for recognizing this gross error in plotting. The plot was generated by local regression smoother (LOESS) without constraints for each model. The previously shown probabilities greater than 1 was fitting of the model outcome which was clustering at the both ends of the near 1 for match and near 0 for non-match. In using LOESS we constrained the limits of the fit so as not to induce such fitting error in the revised figures.

Reviewer #3 (Remarks to the Author):

The authors have largely addressed the comments from my previous review. However, a couple comments still apply to the revised manuscript. I retain the recommendation that this paper be published following minor revision. Please see the attached pdf for detailed comments.

The authors would like to thank the reviewer for the kind comments. All suggested corrections have been addressed in the revised manuscript.
REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In the last revision, I largely suggested copy edits. The authors' response indicates that they made all suggested updates. However, this does not appear to be true in the track-changed document, nor the final submission (authors still write "crime sense" which I think should be "crime scene" and "microscopic inspection" which I think should be "microscopical inspection" -- the former means a very small inspection versus the latter which means inspection of small features).

Either way, electing not to make these editorial changes is under the purview of the authors, and does not negate publication, but I still assert the authors are in error and should reconsider why they were ignored.

Reviewer #2 (Remarks to the Author):

The authors continue to improve their interesting and important manuscript, and the reviewer is confident that it will reach publishable form soon.

The reviewer believes the stakes to be high for this manuscript as the article is targeted at a jury and misrepresentations of fracture mechanics papers from a half century ago could thus have a range of undesirable effects. The reviewer apologises for demanding high levels of precision.

The authors now cite Ritchie et al (1973) as a source for a fracture mechanics basis of the current paper's assertion that searching over a few grains is appropriate for surface matching. This is not appropriate because those authors note that this observation fails at room temperature and is applicable only to the onset of fracture and not the growth of a crack (see their discussion on page 404 of that paper). The authors misrepresent a cryogenic analysis of fracture initiation to be relevant to a room temperature analysis of fracture propagation.

The authors also misrepresent Curry and Knott as suggesting that the relationship between micro damage and crack initiation, as studied by Curry and Knott, has some bearing on the overall crack surface. Curry and Knott only study crack initiation in the paper cited.

The authors cite much other work, including their own work on non-crystalline polymers, to assert a basis in fracture mechanics for the 2-3 grain law. This is of course absurd. While some fracture surfaces such as fiberglass have features that can be matched together quite well, quantitative fracture surface matching has not been performed in fracture mechanics before; this is why the current article is worthy of publication in Nature Communications. However, the efforts to form a foundation from this literature for a "very critical argument to show the statistical basis for the individuality of the fracture surface" has not been successful.

As a side note, the authors probably meant to list four "microscopic feature signatures" instead of three-- imperfections to the microstructure should probably be listed.

Reviewer #3 (Remarks to the Author):

The authors have addressed my previous comments, and I have no further comments. I recommend that this paper be accepted for publication.
Authors’ Responses to Review Comments
Nature Communications manuscript NCOMMS-21-41063A

We are thankful for the additional inputs the reviewers have provided. We have addressed all the raised questions within the paper, especially those of Reviewer 2, and provided point-by-point explanation for the changes, as elaborated below. Furthermore, we have also provided a file with tracked changes that we made to the manuscript.

Reviewer #1 (Remarks to the Author):
There was a typo that was overlooked during editing; "crime sense" which I think should be "crime scene".
   - Correction has been implemented.

"microscopic inspection" which I think should be "microscopical inspection" -- the former means a very small inspection versus the latter which means inspection of small features).
   - Thank you for pointing the confusion about the use of the word; "microscopic inspection of the fracture surfaces by examiners can reliably validate matches". The corrected statement reads; "inspection via a microscope of the fracture surfaces by examiners can reliably validate matches."

Reviewer #2 (Remarks to the Author):
The authors continue to improve their interesting and important manuscript, and the reviewer is confident that it will reach publishable form soon.
   - The authors thanks the reviewer for the kind comments and will attempt to provide additional clarifications as warranted.

The reviewer believes the stakes to be high for this manuscript as the article is targeted at a jury and misrepresentations of fracture mechanics papers from a half century ago could thus have a range of undesirable effects. The reviewer apologizes for demanding high levels of precision.

   - The authors would like to remind the reviewer that this is not a fracture mechanics paper, where the authors attempts to provide universal relationships between the fractal dimension and the fracture toughness of crack-growth resistance. Such an attempt is clearly unlikely and completely beyond the scope of the paper.

   - The core scope of the paper is an attempt to examine the usefulness of the fractal concepts in quantitative fractography and the use of such scale to define the imaging range where two-pairs of fracture surfaces are coming from the same fracture event. In this work, we utilized the limit where the fractal dimension is constant to set the critical scale for imaging on the fracture surface. This limit corresponds to the pertinent metallurgical dimensions on the fracture surface. Aside from the RKR-model (JMPS, 1973) that the principal stress ahead of the crack exceeds a local fracture stress over a microstructurally significant distance (about two grain diameter for mild steel tested at extreme low temperature of -196°C), Dauskardt, Haubensak and Ritchie (Acta Met,
The authors' analysis of critical observation showed that fractal dimension is constant over a range of the order of twice to three times the grain size range for transgranular cleavage fracture (Fig. 9a), about twice the grain size range for intergranular fracture (Fig. 11(a)), and of the order of the grain size for the quasicleavage fracture (Figs. 12a, 13a). It is clear that for these different cleavage mechanisms, the extent of the fractal character of the fracture surface shows good correlation with microstructure and fractographic features, that is the grain size. However, for different classes of fracture mechanisms by microvoid coalescence, the extent of the fracture character extends to twice the microvoid spacing (Fig. 10a).

In our work on AISI 440C tempered stainless steel, we found that this range is about two to three grain size as shown in Fig. 2 of the current proposal. This is the range where the surface roughness deviate from self-similarity, and the character of the features becomes unique for identifications. It should be noted also that Dauskardt, Haubensak and Ritchie, 1990 have shown two ranges of fractal dimensions for transgranular cleavage fracture, one in the range of 1-10 microns corresponding to the cleavage step. This range of cleavage step will be non-unique as it will be found in all surfaces of the same alloy that exhibit cleavage failure. The second range is in the order of 2-3 grain. This is the range we exploited here to explore the uniqueness of each fracture surface pairs, while embracing the stochastic nature of the grain size and the randomness of the critical fracture-triggering site.

As we identified the critical scale of about 2-3 grain size, then sampling theories dictate that the observation window should be about 10-times the critical wavelength to avoid signal aliasing. This limitation requires the imaging scale to be greater than 20-grain size. This is one of the most critical scales that should be considered to set the comparison scale either in a single imaging step or through montage at higher magnification to retain the resolution.

The Abstract is corrected to read; “In the case of the examined class of hardened alloys, which are common in cutlery and tool steel, the identified scale can be related to the microstructure size-scales and is found to be approximately two to three times the grain diameter. This scale closely relates to the characteristic distance necessary for the initiation of cleavage fractures in semi-brittle and hardened metallic alloys. Consequently, the imaging scale required is approximately 20 times the grain-diameter.”

Additional edits are marked on the difference file, on pages 9-11

The authors now cite Ritchie et al (1973) as a source for a fracture mechanics basis of the current paper’s assertion that searching over a few grains is appropriate for surface matching. This is not appropriate because those authors note that this observation fails at room temperature and is applicable only to the onset of fracture and not the growth of a crack (see their discussion on page 404 of that paper). The authors misrepresent a cryogenic analysis of fracture initiation to be relevant to a room temperature analysis of fracture propagation.

The author acknowledges the analysis of Ritchie et al 1973 of a ductile steel alloy that has ductile to brittle transition temperature of about DBTT=−95°C, where fracture occurs before general yielding due to slip-induced cleavage. For the current examined alloy of AISI 440C stainless steel, a common alloy for cutlery and knives, the alloy has up to 1.2% carbon to make it hard and remains sharp. Such carbon content also shift the DBTT to be above the room temperature (Armstrong
and Warner, 1956). Accordingly the current alloy behave in a very similar manner to examined alloy and testing conditions of RKR-theory.

- While RKR-critical microstructure scale is commonly referenced in literature without much rebuke, we have explained the similarity of the failure mechanisms here to avert any further confusion.

- The following text has been added in Page 10 of the diff file:

“It is important to note that the reported fractographic details are reported for mild steel, examined at extremely low temperature, below the ductile to brittle transition temperature (DTBTT) of (~95°C), where fracture occurs before general yielding due to slip-induced cleavage. For the current examined alloy of AISI 440C stainless steel, a common alloy for cutlery and knives, the alloy has up to 1.2% carbon content in order to make the alloy hard and remains sharp. Such carbon content also shifts the DBTT to be above the room temperature47. Accordingly, it is no surprise that the examined alloy in the form of rods of knives at room temperature show similar fractal character to mild steel alloys tested below their DTBTT34,45.

The authors also misrepresent Curry and Knott as suggesting that the relationship between micro damage and crack initiation, as studied by Curry and Knott, has some bearing on the overall crack surface. Curry and Knott only study crack initiation in the paper cited.

- The authors acknowledge the confusion of the author about the discussion of Curry and Knott (1978) to show the stochastic nature of the cleavage process initiation due to the randomness of the critical fracture-triggering site and how we reference it into propagating crack. Similarly, the reviewer questioned the relevance of RKR-theory for initiation and the analysis the authors carried out for a propagating crack. There are two aspects here that we would like to bring it to the attention of the reviewer.
1. The cleavage crack might nucleate but the specimen may not experience complete fracture. When a microcrack is initiated at hard particle and propagated within a grain ahead of the crack tip, it may be arrested if there is insufficient global driving forces to continue crack propagation. Accordingly, persistence of the requirement of reaching critical stress over a microstructure distance will be maintained for continued crack propagation until the macroscopic crack reach an unstable propagation domain.
2. It is common in the analysis of fractal dimension and fractal analysis of fracture surfaces (i.e. propagating crack) to correlate the extent of fractal behavior and microstructure and proctographic features. The link is typically carried out by invoking the RKR-of exceeding fracture stress over microstructure significant distance, and the statistical interpretation of Curry and Knott, 1978 and Lin, Evans and Ritchie, 1986, (wherein the this characteristic dimension represents the location from the crack tip of the most probable initial fracture event and reflects the mutual competition of high stresses close to the tip and an increasing probability of finding a "weak link" (e.g. large brittle particle) to initiate fracture away from the tip [23]; Duskardt et al, 1990).

- To clarify such confusion, the text has been edited to reflect these issues as noted below

“This critical scale is required for cleavage crack initiation. However, it is apparent that such critical scale is also embedded in the topography of the fracture surface. When a microcrack is initiated at a hard-
particle, it may be arrested if there is insufficient global driving forces to continue crack propagation 46. Accordingly, the requirement of reaching critical stress over a microstructure critical distance will be maintained for continued crack propagation until the macroscopic crack reach an unstable propagation domain, and thereby set-forth the critical fractal scale on the topology of the fracture surface.”

Correction of citation of Curry and Knott:

“Two nominally identical articles from the same material lot might show very different toughness (resistance to fracture) and failure strength values. By extension, we speculate also that such statistical differences will result in different local fracture surface topology because of the statistical randomness of the microscopic spatial location of the critical fracture-triggering particle.”

The authors cite much other work, including their own work on non-crystalline polymers, to assert a basis in fracture mechanics for the 2-3 grain law. This is of course absurd. While some fracture surfaces such as fiberglass have features that can be matched together quite well, quantitative fracture surface matching has not been performed in fracture mechanics before; this is why the current article is worthy of publication in Nature Communications. However, the efforts to form a foundation from this literature for a “very critical argument to show the statistical basis for the individuality of the fracture surface” has not been successful.

- Again, the authors notice how the reviewer got confused between fracture mechanics basis and fractal characteristics observed from quantitative fractography analysis of different fracture surfaces. The authors have rewarded this section and removed such ambiguity. The authors made sure that there is no claim of universal relationships between the fractal dimension and fracture toughness of crack-growth resistance and the fractal scale limits should be examined for each material class and then related to the microstructure scale for the topology of the fracture surface.

As a side note, the authors probably meant to list four "microscopic feature signatures" instead of three-

- imperfections to the microstructure should probably be listed.

- The text has been amended to reflect the general attributes of parameters that affect cleavage susceptibility and their extension to the topology of the fracture surface as relevant to uniqueness in the current work;

“Susceptibility of cleavage fracture is sensitive to microstructure (grain size and carbide population), yield strength, stress state (triaxiality), and environment (temperature and radiation). Similarly, the fracture topography will exhibit unique microscopic feature signatures that exist on the entire fracture surface. We extend these parameters to generally include the material microstructure, the intrinsic material resistance to fracture, the direction of the applied load, and the statistical distribution of imperfections within the microstructure.”

Reviewer #3 (Remarks to the Author):

The authors have addressed my previous comments, and I have no further comments. I recommend that this paper be accepted for publication.
The authors would like to thank the reviewer for the kind comments.
REVIEWERS' COMMENTS

Reviewer #4 (Remarks to the Author):

I reviewed the responses to reviewer comments and relevant associated parts of the manuscript and supporting materials. I should note that while I have a good understanding of fracture mechanics principles based on my expertise in rock fracture, I do not have direct experience with the fracture of metals. My comments should be interpreted accordingly.

In response to the note about the relevance of the work of Ritchie et al. (1973) and differences in temperature, the authors essentially note that it is not the absolute temperature that is relevant for purposes of comparison, but rather than temperature relative to the brittle-ductile transition. I find this to be a reasonable argument.

I was not able to personally get access to a copy of the paper by Curry and Knott. However, in reviewing the responses to the reviewer’s corresponding comment and the corresponding changes to the manuscript, it appears that the remaining citations to Curry and Knott are now more directly focused on the question of crack initiation. The extension to the fracture surface topology is reframed as speculation on the part of the authors, which appears consistent with the reviewer’s comment.

Regarding the second-to-last comment by Reviewer #2, the overall modifications seem appropriate. In particular, the changes made to some of the text that de-emphasize a strong fundamental link in the literature between fracture mechanics theory and the proposed fracture matching concepts appear to have addressed this comment, at least in part. The authors might consider changing "In the field of Fracture mechanics, it is postulated that..." to "Some fracture mechanics studies have demonstrated that..." to make the statement less sweeping in nature.

The response to the final comment by Reviewer #2 appears appropriate, as the corresponding text has been clarified.

During my review, I noted this sentence fragment on page 11: "This very critical argument to show the statistical basis for the individuality of the fracture surface". In view of some of the other changes to the text, the authors might consider removing this altogether (and adding some sort of transitionary phrase to the start of the following sentence).

Overall, it appears that the authors have made a good faith effort to address the comments from the reviewers. It is not obvious to me that further modifications are necessary, but if the previous reviewers believe there are remaining issues, my recommendation would be to give the authors an opportunity to address those.