### For a better Understanding of the Structure

Response by the authors: "- text "

Segmentation to make Reviews and Responses more readable: "————————-"

Due to many factors, the following articles are only added in the replies to the reviews: Kot, Tyszkiewicz, Leloch, et al. (2025); Linsel et al. (2025)

## Reviews

### Review by anonymous reviewer 1, 29 Oct 2024 12:21

The method presented here is of great value for the automatization of lithic scar pattern analysis. However, the aim of the study is not clear. Why is automatization necessary? What are the benefits? What the drawbacks? Are there other comparable methods? If so, what aspects of the new approach are better? I am in favor of publishing the method, however the manuscript needs major rewriting to incorporate the points/ questions raised above.

- The main goal of this article wasn't to solve an archaeological but rather a methodological problem. Hence, it has been presented and structured to introduce solely the method. The article was adjusted to incorporate more detail on relevant archaeological studies. However, how applicable it is for future research, needs to be tested on a broader scale.

- We added a section "related work" to give an overview of research conducted on lithic artifacts according to their operational sequences as well as state of the art analysis of 3D applications.

- Why the use and reuse of 3D models is important, see our newly accepted article (Linsel et al., 2025).

————————-

Furthermore, the manuscript needs restructuring: The introduction should be re-written. It contains basically only methods. I suggest rewriting the introduction with a discussion of the state of the art, as well as a focus on why an automated scar recognition is useful.

- The main theme of this article is not the automatic scar recognition method (segmentation) but rather a method to test the performance of a graph based approximation of the scar relations. However, this was partly due to the imprecision of phrasing, which was adjusted accordingly.

————————-

Part of the Materials is also Methods (and should be moved there), like Preprocessing and the Manual Segmentation sections.

- We moved these subsections to the "Method" section.

————————-

Detailed comments:

Lines 3-4: citation(s) is (are) required.

- Phrase was deleted.

Lines 9-16: I suggest to insert a figure to illustrate the concept of flaking and to label the individual features (ridges, scars, dorsal/ ventral etc.).

- Figure 1 was exchanged and now show also an annotated artifact with annotations of a scar and a ridge. But this article introduces a method and is hence not an introduction in lithic artifacts.

Lines 19-36: This (as well as the paragraph before) is methods. I suggest rewriting the introduction with the focus of why an automated scar recognition is useful.

- As mentioned before, it is not a paper about scar recognition but it is adjusted to why "scar relation predictions" are useful.

Table 1: Please indicate in the caption what the numbers referring to the publications mean.

- Rephrased and referenced (L51 - L53)

Lines 69-72: „Intriguing for this ...“ This sentence is very long and hard to understand. Please re-write.

- The sentence was removed because it didn't contribute to a better understanding.

Line 87: please use the plural of the French term chaîne opératoire.

- We decided to replace the term "chaîne opératoire" with "operational sequence" to make it easier to read.

Figure 1: Please add a legend for the color coding of the scar and ridge labels.

- It is uncommon to add legend to reflect ids (label).

Lines 107-108: „In recent years, the segmentation of artifact features on 3D models such as cutting edges has been based on ridge rather than scar segmentation (Pop, 2019 ; Schunk et al., 2023 ).“ would be rather part of the new introduction when reviewing the former approaches.

- Due to the restructuring and hence the minimal impact of these references, they are removed.

Line 188: not „edge retouching“ - rather: „edge retouch“ or „retouching the edge“

- Line was removed.

Line 189 (see also line 208): „cutting edge“ is a very specific functional edge, but tools are used for various functions. I suggest to call it „working edge“ or „functional edge“.

- „cutting edge“ → ”functional edge”

Line 190: Why denticulate? When talking about edge retouch (especially a continuous one) the most appropriate tool is a scraper.

- Line was removed.

Line 213: in which way is the simplified graph simplified?

- Already discussed in the previous subsection (”Graph Simplification”).

---

## Review by anonymous reviewer 2, 10 Nov 2024 11:37

The paper is very nicely written, with clear details of the technical aspects. I am excited to use the method proposed by the authors to annotate 3D flake models, as I think it could be immensely helpful for pre-processing data for use in machine learning.

I have a few comments, which are as follows:

In the Materials and Methods section, some aspects could benefit from further elaboration. I assume the paper's target audience includes not only experts in computer science but also archaeologists who may wish to apply this method in their own research. Due to my limited knowledge of computer vision, I found it challenging to fully follow the steps and, more importantly, the underlying logic behind each dataset preparation step.

- - The target audience of this article is computational archaeologists and computer scientists with expertise in 3D modelling, as this is a contribution to the CAA conference. Our new article (Linsel et al., 2025) was written with those with an archaeological background in mind.

---

For example, the preprocessing using GigaMesh could be clarified with more details on how the meshes were oriented and an overview of the protocols followed according to the GMOCF routine. Additional clarification on this routine itself would be helpful—specifically, whether it is intended to distinguish two connected surfaces using a ridge made by connected vertices.

- If the reader is not familiar with why this is important, we referenced the original article.

---

The Manual Segmentation section could also benefit from more detail. It was mentioned that both MeshLab and Blender were used, but it is unclear how the models were annotated in Blender. Were vertices on ridges marked manually to separate the scars, or did the authors follow some semi-automated processes? While Fig. 1 shows the segmentation steps in Blender, it is not clear how segmentation was actually conducted in the software.

- This is also not part of this article because while it is necessary, a 5000 word long article cannot close this gap.

- The annotation workflow is available on our platform (workflow).

---

It was also noted that flake scars were labeled, but the specific approach was not clear. For instance, were individual vertices and faces within each scar assigned a specific color to distinguish scars in later steps, or were scars labeled as shown in Fig. 6? If so, it may help to mention this earlier in the text. Providing additional details on these processes could benefit those interested in conducting similar 3D data processing.

- We simplified the terminology and only use scar labels in the new version of this article.

---

I apologize in advance for any misunderstanding of the graph, and if I may have missed relevant points in the main text. In the Parameter-based Directions section, the authors created a graph G with nodes S and R. Is this intended as a bipartite graph, with scars and ridges as nodes possessing distinct properties? While Fig. 8 illustrates the graph, the caption could benefit from additional clarification to explain the meaning of the node colors, the rationale behind node numbering, and which nodes represent S and which represent R.

- According to the graph notation, a graph can be defined by GRAPH = (NODES,LINKS/EDGES). Applying that logic, S (scars) are the nodes and R (ridges(connections)) are the edges connecting them.

---

I enjoy the study's succinct writing and clear organization, and I am eager to apply this method myself in the near future. I believe that by providing additional details for non-expert readers like myself, the authors could greatly enhance the paper's readability and encourage broader application of this method.

---

## 3rd Review

The work reported in this paper is very exciting, and I am eager to see it come to fruition. The results are intriguing, the approach to characterizing scars and their relationships is useful and novel, and I think the outlined future directions are generally sound. Below I provide an overview of the authors' approach as I understand it and discuss some remarkable aspects of their results; I then identify some issues with the manuscript that I think should be addressed.

Overview:

Building on their previous work, the authors use a graph approach to represent the scar arrangement visible on the surface of knapped stone artifacts. While this is not new, the authors propose and evaluate a method for directing such graphs (i.e., determining chronological relationships between scars) that is based on objective and quantifiable scar, ridge, and network (graph) properties. To this end they analyze 60 manually segmented 3D models of exper-

imental and archaeological artifacts and create directed graphs based on their computed properties; they then evaluate the accuracy of the resulting graphs against graphs produced by human analysts. The results of this evaluation are remarkable for several reasons. First, the accuracy of the directed graph models is surprisingly high across the board (Table 3; cf. L245) given a) the way it is computed (see Major issues #2 below) and b) that using a multivariate approach to determine edge direction likelihood will presumably yield an even better correspondence.

- The amount of observations (edges) and the number of parameter per node (of around 69) are way to high for a reliable analysis graph wise as f.i. a PCA needs to be performed with normal distributed, correlated variables with at least 5 observations per variable (more or less than 345 entries). Further, the heavy reduction of data, only using mean values of all vertex based approaches prevents a reasonable use of a multivariate analysis.

---

Second, the difference between the experimental and archaeological (GdF) datasets in terms of the effects of graph simplification is striking and raises several questions about the datasets. For instance, are the graphs of comparable complexity? Are the experimental pieces retouched? What, exactly, is being removed by the process of simplification on the experimental artifact graphs, and how much 'simpler' is the result? Third, the performance of the 'Surface Area' property, which consistently produces graphs within 4% of the best and is easy to compute, makes me wonder if it is worth considering any of the other attributes when employing univariate approaches. Along these lines, I find the relative uniformity in the performance of the various attributes in the archaeological (GdF) dataset notable, and I don't understand how the statement on L257-258 (or L272-274) is warranted except in the case of the simplified self-created dataset. It would have been interesting to see what accuracy values can be expected if manual graphs created by multiple analysts were to be compared.

To me the results presented here don't warrant much excitement about the graph simplification approach that is proposed, but I do agree with the authors that other forms of graph simplification should be explored in the future (e.g., L288-290).

---

**Major issues:**

Several limitations of the study are discussed by the authors to varying degrees, including the less-than ideal datasets that are used. Leaving these aside, my main criticisms pertain to how the work is presented at the conceptual level and how the resulting graphs are evaluated for accuracy:

1. Explanations and justifications: Technically, the procedure for creating undirected graphs and then directing them makes sense. The computation of various attributes makes sense as well, as does the graph simplification. At a conceptual level, however, I somewhat struggled to understand what is going on. A few examples: a. Why is the graph directed using individual attributes instead of a weighed combination? This should be justified, not least because

it may have implications for the interpretation of the results (e.g., accuracy estimates could be too conservative).

- This study reflects the state of the art at the point of the deadline and hence the research is in an ongoing development. That been said, of course a weighted combination could potentially improve the results and could have further implications. However, many of these parameters are newly implemented and hence an in details investigation of all parameters individually seemed to be necessary.

---

b1. What is the purpose of simplifying the graph by removing retouch scars?

- As you correctly mentioned later: manual annotations vary between researchers and hence their of detail. This leads to an over- and undersegmentation of scars. We are currently developing an automatic labelling of monitored cores, which enables us to create very precise approximations of the real ground truth data. To prevent over-promising, like shown in Grosman (2016) (L 81 - L 86), we decided to hint on it (L 350 - 351) but not share this part of our research.

---

b1.1: Why bother (manually) segmenting retouch scars in the first place?

- Because otherwise the selected scope of detail restricts future research, f.i. enriching these scars with more detailed information like whether it is actual retouch or post-depositional edge damage, so these can be detected automatically. Therefore, not segmenting retouch scars would hinder the very research needed to solve the problem defined in b3.

---

b2. The possibility of automatically identifying retouched edges is exciting, and I can envision several applications down the road, but I'm not sure I see the point when the segmentation still relies on manual input (I don't think this approach would work well with the kinds of automated methods currently available).

- Without proper annotation data, we are not able to distinguish between actual retouch and post-depositional edge.

---

b3. The procedure for identifying retouch also seems unable to distinguish between actual retouch and post-depositional edge damage.

- Compare b1.

---

c. Table 2 lists nine variables, but only five of these are linked to properties that the reader has come across by that point in the manuscript (e.g., in Table 1). What is the theoretical basis for the inclusion of the other four? They should be explained in terms of how they relate to knapping behaviours and mechanics. For instance, what are the archaeological interpretations of the network properties? Even for variables that are linked to archaeologically determined

properties (e.g., 'Curvature along Polylines' – 'RRP-1') a discussion seems warranted. For instance, what is the envisioned interpretation of the second IIoP (L143-145) in terms of the concavity of the scar (definition of RRP-1 on Table 1)? What about the sampling of surface attributes near scar borders?

- Without controlled experiments, which produce reliable data, with clear relation of scars and knapping step, the theoretical implications are way too vast to be explored in this article. We have only 60 datasets, with highly varying results.

- The complexity of working with 3 different data sources (mesh, polyline, Graph) with 9 parameters, many applied on different scale, result in huge amount of data, which need to be handles by experts in computer science. Theoretical implications, however, especially if no data is available, should be provided by experts in archaeology.

---

I think an in-depth discussion of the meaning of the resulting directed graphs (e.g., Fig. 9) in terms of reduction behaviours and chronology (e.g., start and end points) would have been very helpful. The challenges of inferring these from scar arrangements (e.g., Kot, Tyszkiewicz, and Gryczewska (2024)– cited by the authors) deserve recognition (e.g., multiple scars may result from a single hit, two scars may be adjacent and their order may be known yet they may be separated by several steps in the reduction sequence, and two non-adjacent scars may have been removed one after the other), and here they seem to be largely glossed over.

- We addressed now included more details on the graph and that these graphs are partly ordered (e.g. L290 - 300), which is also the result of Kot, Tyszkiewicz, and Gryczewska (2024) and Kot, Tyszkiewicz, Leloch, et al. (2025). However, this work is not an archaeological investigation of the meaning of graph modelling but rather an article introducing a new state of an ongoing research project with a methodological goal.

- Many points, listed like "multiple scars may result from a single hit" is based on direct observation during a knapping process like Kot, Tyszkiewicz, and Gryczewska (2024) or on additional interpretations. The latter, which applies to the datasets of this article, results in additional bias. These contractions, if the correct properties for determining the scar relation has been found, should not result in differences in the interpretation of scar relations. Hence, this contraction is relevant for converting a scar graph model more to a strike based model, it is not relevant for the scar graph models.

---

This is unfortunate, since a more in-depth consideration may have resulted in other and perhaps ***more meaningful approaches*** to graph simplification being considered.

- As you are later notice: archaeological data in the present form of drawings and operational sequences, are highly subjective. This study had the goal to explore the possibility of a simple graph simplifications. More complex simplifi-

cations will be explored in future studies but you need to keep in mind, for this study, we work with huge datasets leading to graph models with multiple GB in size. Handling these datasets is in itself complex. ***More meaningful approaches***, like contraction nodes in sequences, would need exemplary datasets consisting of closely monitored annotated 3D models with OS data and without any human interpretation. They would result in more complex analysis and question like: How do you handle parameters based on mesh or polyline data? Do you analysis the all ridge polylines, segment them in those with direct contact with other? Questions not answerable in the first of a kind study.

---

2. Evaluation: I am not convinced by the evaluation procedure used by the authors. First, it is important to point out that for none of the artifacts is the full, true reduction sequence known. Simply put, there is no 'ground truth dataset' here (cf. line 226); this is acknowledged to some extent on p. 14 (L278-282), but almost as an afterthought.

- We adjusted all references to the dataset and added multiple lines to indicate that these a preliminary results (e.g. L23-27; L345-346).

2b. Second, no theoretical justification is provided for the evaluation function presented in Eq. 9 (L226-228). Consider the following temporal sequences of events denoted by letters: 1) A→B→C→D→E→F, 2) A→D→B→C→E→F. Assuming (1) is the true sequence, the accuracy of the second sequence according to the proposed formula would be 40% (i.e., 2/5); however, one could also look at the second sequence as being 80% correct (i.e.,4/5), as the only false pairwise sequence is D→B (A→D is true, since D did happen after A). I also wonder if the % accuracy as calculated here is correlated with the number of connections within a given graph (probably not a desirable outcome).

- It seems like we "over-promised" by referencing operational sequences, while conducting research on the "temporal relation of scars". Hence, we adjusted the terminology in the article so we are now only speaking about the "temporal relation of scars".

- We completely agree, that the evaluation is in need of improvement, but that is an inherent problem of the method. Similar evaluation methods are commonly used, most recently reversely applied as error rate in Kot, Tyszkiewicz, Leloch, et al. (2025) to validate new data against the data presented in Kot, Tyszkiewicz, and Gryczewska (2024). But one aspect need to be highlighted, Kot, Tyszkiewicz, Leloch, et al. (2025)(published after the first round of reviews) not reference the attributes or properties used for the manual direction prediction, making our approach even more transparent than similarly used evaluation methods in archaeology.

- And as addressed (L290 - L300), if the ideal parameter or combination is found, it results in a partly ordered graph, what applies also to is the case for most refittings compare Kot, Tyszkiewicz, Leloch, et al. (2025).

---

**More minor things:**

I would encourage the authors to consider the following suggestions in possible revisions to this manuscript:

1. I appreciate the mathematical notation, but as I read the paper I often wondered if some of it may be unnecessary (e.g., L98, L128-129), particularly since at times it also seems misleading (e.g., L115 – there are no possible circumstances under which a scar can be Sl = M). Consider providing a brief plain English explanation for formulas such as Eq. 9.

- We adjusted many terms and have exchanged most of the abbreviations with written out versions.

- To simplify the specifics of our approach, we decided to use "scar label" as an umbrella term for all forms of labels, not distinguishing between cortex, scar etc. We exchanged all references to "labels". Under this premises " label" can indeed be Sl = M. But we are still using the framework provided as "relation between scars", because to introduce "the relation between surface features/labels" as new method would result, in our opinion, in an even more cryptic article.

---

2. Please clarify what is meant by "the mean value of all parameters. . . " on L214. A scar has a single value for surface area, for instance, so I don't understand what was averaged.

- That is completely correct and we adjusted it.

---

2b. In fact, I think that the entire paragraph should be clarified – how many unique graphs were created for each 3D model?

- This can be difficult to understand, but all parameter predictions only result in directed ridges, which could form individual graph models. Hence, only two graph models, the original and the simplified, exist.

---

3. Discuss the experimental dataset in more detail. This is needed because the presence or absence of retouch on the artifacts from this dataset may explain some of the differences in performance (Table 4).

- As mentioned, this paper represents only a state of the art, not the final results and this is not the goal of this article: We wanted to present a new method, which was only done by our working group and further archaeologically relevant interpretations are at the moment highly speculative. The amount of data is way too low to create a reliable hypothesis.

---

4. Ensure that all acronyms are explained on first use, and that all information is adequately contextualized. The following are some examples where this is an issue, but the list is not meant to be exhaustive:

- To simplify the reading of the article, we decided to write all acronyms out, which are not frequently used (GMOCF; CO; OS).
- The site acronyms (ROB; GdF) are now added to the main text.

---

a. L43: MSII – first used on this line and not defined (the abstract doesn't count).
- Adjusted

---

b. L43: What is RSP-1, and why is this property not approximated in this study? Note: RSP is defined on L32, but not this specific property.
- Argumentation is added.

---

c. L58: What does CO stand for?
- Deleted

---

d. L96: What does GMOCF stand for?
- Deleted

---

e. Figure 2 label: What does ROB stand for? Is this from the experimental collection?
- Added

---

f. L31: Why were these attributes separated into 10 properties?
- Explained here: L45-53

---

g. L32: What are the binary properties noted on line 32? Why can't they be derived directly from a segmented artifact? Why are they important?
- Explained her: L329-332.

---

h. L48: Why are these properties not yet included in the approach? I can guess, but it would be nice if the study was better contextualized.
- Because there exists no ground truth data to detect them. This is a good example, why we "bother segmenting retouch scars" and little details "in the first place" (Major issues: b1.1). Details matter.

---

5. Provide more (and more consistent) detail in the Figure and Table captions. For instance:

a. some figures depicting artifacts show their IDs (e.g., Figure 2), others don't (e.g., Figure 1 or Figure 3) – why?

- IDs added

---

b. Table 1 caption: What do the numbers mean? If they refer to suggested importance (e.g., 1-5, with 5 being least important), why are the same numbers listed for multiple properties in the same column?

- As referenced in L 56-57/ Tab 1:subcaptions, it references the list position, therefore the importance.

---

c. Table 3 caption: explain why some of the text appears in bold, even if it is relatively obvious.

- For better readability of highest performing parameters (common practise in ).

---

6. For greater clarity, consider listing the actual variables used in Tables 3, 4 and Figures 6, 7. For example, IIoP k is discussed in the text (L239) yet it is not listed in the tables (e.g., Table 3).

- IIoP's are added in the table.

---

7. Consider including additional information on Figure 10 (similar to what is shown in Figure 5c, but with edge directions indicated) to make comparisons with Figure 9 easier.

- No

---

8. Introduction: a. Consider providing a short but explicit discussion of the advantages of working with 3D models. The use of 3D models should be justified.

- We added a section "Related Work" to contextualized this article with other 3D approaches.

- Due to the parallel schedule of this and a second article (Linsel et al., 2025), in which discuss reuse of data including 3D models and contextualized these in the context of manual techniques .

---

b. Provide more details on how this work fits within what is clearly a wider research agenda and how it builds on previous work.

- See 8a.

9. Abstract:

a. Lines 4-5: "These models, developed using [MSII] curvature" – where is this discussed in the text?

- Deleted

b. Line 8: I would suggest either qualifying this sentence or expanding on this idea of automation potential in the main text.
- Deleted

c. On the last line broad applicability is mentioned, but I don't think the statement is well supported by the results obtained here (consider, for example, the differences in the performance of the graph simplification procedure between the experimental and GdF artifacts).
- We meant that the approach is applicable in the sense of "you can compare scar properties for the first time" across different datasets. It was never intended to solve the complete problem, which will be a task for the next decade.

---

Other line items:
L97, L99: Explain these conceptually (i.e., what they are meant to accomplish, and why that is necessary).
- As in L97 originally stated, it ensures that the mesh is a differentiable manifold.
- L99: More details are added.

---

L112-113: Clarify why MeshLab was replaced with Blender.
- done: L148-151

---

L105-106: Some clarification may be warranted here. 'Adjacent' here (and based on the illustration on Figure 2) seems to imply that ridge vertices are not actually assigned to a scar surface. Is that correct? If so, how many vertices are excluded (i.e., how 'wide' is the ridge segment)?
- No, all vertices are included in the annotation. For a better illustration, the width was artificially widened.

---

An alternative representation is that of overlapping (i.e., same coordinates) mesh vertices, shared by two or more adjacent scars, which is what seems implied on L170-171.
- No, all data is explicit.

---

L149: A second approach to what, exactly? Also, consider replacing 'relies similar' with 'is similar'.
- Rephrased: L183-184

---

L179: Consider replacing "hence it's ridges" with "hence its edges".
- done

---

Table 2: There seems to be enough space to spell out what the different properties refer to (e.g., RSP-2), as done in the first column of Table 1. I think this would make the table easier for readers to digest.

- Done ()

---

L39: Presumably Linsel et al., (2024) refers to the 2023 publication listed in the references? Or does the entry on L344 need revising?

- Linsel et al., (2024) = L344

---

L120: "according to the . . . " should be replaced with "according to established . . . ".

- Rephrased: L134-L137

---

L146: "euclidean should be Euclidean"

- changed

---

L207: This is an incomplete sentence. Rephrase.

- Rephrased: L233-234

---

L233: "display" should read "displays"

- changed

---

L262: "then" should read "than"

- changed

---

L271: Rephrase – I find this sentence confusing. How does a concept get combined with data? - Figure 9, subplot (b): "Simplfied" should read "Simplified"

- changed

# References


Grosman L (2016). Reaching the Point of No Return: The Computational Revolution in Archaeology. *Annual Review of Anthropology* 45, 129–145. ISSN: 1545-4290. DOI: https://doi.org/10.1146/annurev-anthro-102215-095946.

Kot M, J Tyszkiewicz, and N Gryczewska (2024). Can we read stones? Quantifying the information loss in flintknapping. *Journal of Archaeological Science* 161, 105905. ISSN: 0305-4403. DOI: https://doi.org/10.1016/j.jas.2023.105905.

Kot M, J Tyszkiewicz, M Leloch, N Gryczewska, and S Miller (2025). Reliability and validity in determining the relative chronology between neighbouring scars on flint artefacts. *Journal of Archaeological Science* 175, 106156. ISSN: 0305-4403. DOI: https://doi.org/10.1016/j.jas.2025.106156.

Linsel F, JP Bullenkamp, and H Mara (2025). Reusing 3D Measurement Data of Lithic Artifacts to Develop Analytical Methods. In: *NWDVA 2024 Bochum: Digitale Archäologie*. Accepted 10-02-2025.