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 experimental 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. 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).
b. What is the purpose of simplifying the graph by removing retouch scars? Why bother (manually) segmenting retouch scars in the first place? 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). The procedure for identifying retouch also seems unable to distinguish between actual retouch and post-depositional edge damage.
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?

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

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. 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).

More minor things:

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

1. I appreciate the math 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 S_I = M). Consider providing a brief plain English explanation for formulas such as Eq. 9.
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. In fact, I think that entire paragraph should be clarified – how many unique graphs were created for each 3D model?
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).
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:
    a. L43: MSII – first used on this line and not defined (the abstract doesn't count).
    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.
    c. L58: What does CO stand for?
    d. L96: What does GMOCF stand for?
    e. Figure 2 label: What does ROB stand for? Is this from the experimental collection?
    f. L31: Why were these attributes separated into 10 properties?
    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?
    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.
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?
    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?
    c. Table 3 caption: explain why some of the text appears in bold, even if it is relatively obvious.

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).
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.
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.
    b. Provide more details on how this work fits within what is clearly a wider research agenda and how it builds on previous work.
9. Abstract:
    a. Lines 4-5: "These models, developed using [MSII] curvature" – where is this discussed in the text?
    b. Line 8: I would suggest either qualifying this sentence or expanding on this idea of automation potential in the main text.
    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).

<u>Other line items</u>:

- L97, L99: Explain these conceptually (i.e., what they are meant to accomplish, and why that is necessary).
- L112-113: Clarify why MeshLab was replaced with Blender.
- 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)? 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.
- L149: A second approach to what, exactly? Also, consider replacing 'relies similar' with 'is similar'.
- L179: Consider replacing "hence it's ridges" with "hence its edges".
- 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.
- L39: Presumably Linsel et al., (2024) refers to the 2023 publication listed in the references? Or does the entry on L344 need revising?
- L120: "according to the …" should be replaced with "according to established …".
- L146: "euclidean should be Euclidean"
- L207: This is an incomplete sentence. Rephrase.
- L233: "display" should read "displays"
- L262: "then" should read "than"

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