VIRTUAL KNAPPING (AND REFITTING) WITH NEURAL NETWORKS: PROOFS OF CONCEPT

J.D. ORELLANA FIGUEROA[1], J.S. REEVES[1,2], S.P. MCPHERRON[3] AND C. TENNIE[1]

[1] Department of Early Prehistory and Quaternary Ecology, University of Tübingen
ext-contact@jorellanaf.com
[2] Technological Primates Research Group, Max Planck Institute for Evolutionary Anthropology
[3] Department of Human Evolution, Max Planck Institute of Evolutionary Anthropology

Abstract:

Recent advances in neural networks have brought about new opportunities for their application in archaeological research. Stone tools, due to their longevity and prevalence across most of prehistory, are a valuable source of evidence for archaeologists. For most of prehistory, stone tools were made by striking a core – often with a stone hammer – to produce flakes with sharp cutting edges. Modern experimental replication of such stone tools, as well as the refitting of prehistoric lithic material are both important methods for understanding in greater detail how prehistoric stone tools were manufactured, and by extension, what insights they can bring to our knowledge of human evolution. However, replication experiments can require considerable time and raw materials. Lithic experiments themselves are also difficult to control and replicate; e.g. as it is difficult to control many knapping variables. Refitting can be an even more time-consuming task, as archaeologists must find two matching pieces of stone amongst an entire assemblage of them. Here we discuss the development of a toy model and a recently published proof of concept for a virtual knapping framework capable of accurately predicting the shape of computer-generated flake removals from the surface information of the intact core. In addition, we present an early prototype for a virtual refitter as an extension of our virtual knapping framework. Both models, after additional development and validation, could become important tools for lithic experimentation and analysis, and provide more robust results with which to understand prehistoric stone tool production, and thus, human evolution.

Keywords: Archaeology, Computer Science, Cultural Evolution, Machine Learning, Neural Networks.

1. Introduction

Stone tools have been manufactured for least 2.6 million years (Braun et al. 2019; Semaw et al. 1997; Semaw et al. 2003), and their antiquity, the likelihood of their preservation for millennia, as well as their commonality across so much of human prehistory make lithic technology one of the main sources of evidence available for the study of human evolution.

These stone tools are in many cases marked by one attribute: sharp edges. Stone tool manufacture creates cutting edges on the stones that could then be used for tasks such as carcass and plant accessing and processing (Schick and Toth 2006a: 18–19). Making stone tools could be done through many different means, but for the earliest tools (i.e. *Oldowan* tools), it was generally accomplished by striking a stone (the *core*) with another stone (the *hammer*), a process also known as knapping; done for the earliest (i.e. *Oldowan*) stone tools mainly with the core and hammer held one in each hand (*freehand percussion*) (Schick and Toth 2006b: 4).

The act of repeatedly knapping a core is called 'core reduction' or 'lithic reduction', and it is in this way that stone tool forms were brought about, and assemblages of lithic products and by-products created.

One of the primary methods archaeologists use for understanding the history of human evolution is the experimental replication of prehistoric tool shapes by modern knappers (Eren et al. 2016). Researchers use the results of stone tool replication experiments, as well as the specifics of their experimental set-up to draw inferences about the various factors that influenced the shape of the lithic record. Some factors under analysis include past human culture and behaviour (Putt, Woods and Franciscus 2014; Snyder, Reeves and Tennie 2021; Tennie et al. 2017), cognition (Putt et al. 2017; Putt,

Wijeakumar and Spencer 2019), biology (Kivell 2015; Susman 1988), and landscape use (Braun et al. 2008).

However, replication experiments can require considerable time and raw materials, and are susceptible to human biases stemming from the difference across (e.g. skill) and within (e.g. fatigue, motivation) knappers. Thus, lithic experiments are also difficult to reproduce completely, as it is difficult to *control* many of the variables that affect the results of knapping. Some researchers have tried to address these issues by using standardized core shapes or by using a computer-controlled machine to knap (Dibble and Režek 2009; Magnani et al. 2014; Režek et al. 2011), exploring the effects of individual variables during flaking, but these methods then also require even more resources; i.e it is costly to exert experimental control.

In addition, many unknowns in our understanding of lithic production become obstacles for studying human evolution through the lens of lithic replication, and stone tools in general. For example, we do not know what the possible range of variability of even the earliest stone tools was (Braun et al. 2008, 2009; Braun and Hovers 2009), which makes it difficult to determine how strong the influence of ecological and stochastic variables (e.g. raw material availability, reduction intensity, cobble size) was on the lithics we observe in the archaeological record (Braun et al. 2008, 2009; Schick and Toth 2006a: 27–28, 30–31; Toth 1985).
Furthermore, various site formation processes (e.g. different sedimentation rates) as well as time and space averaging (e.g. the mixing of material days apart in a layer spanning 10 ka) affect every archaeological site, and they can obscure the actual processes and hominin behaviours that led to a certain lithic assemblage (Dibble et al. 2017; Perreault 2019; Schick and Toth 2006a: 27–28, 33–34).

Yet some of these unknowns can still be addressed, if perhaps only to examine how difficult it would be to untangle the individual effects of the many factors that shaped the lithic record in general (e.g. Braun et al. 2008; Moore and Perston 2016).

In addition, although the preservation of lithic material is indispensable for the study of human evolution, any possible early Palaeolithic technologies made out of organic material, like human (Allington-Jones 2015; Thieme 1997; Warren 1922) and non-human primate wood tools (McGrew 2010; Musgrave and Sanz 2018; Pruetz and Bertolani 2007) are likely lost, as are any additional insights that could have been gained from their study.
Even with the easier preservation of bone, the earliest definitive evidence of bone tools does not reach nearly as far back as stone tools do (Backwell and d'Errico 2001; d'Errico and Backwell 2003; Stammers, Caruana and Herries 2018), rendering early stone tools all the more important for our understanding of human evolution, and the limits of what they can tell us about hominin prehistory even more impactful.

Both high costs and large time investments make large-scale lithic experimentation difficult to undertake, let alone reproduce, such as with studies similar to that of Moore and Perston (2016) – one example of a larger-scale experiment which showed how more complex patterns of core reduction could appear stochastically. A lack of reproducibility, as in other fields, and the many requirements for carrying out replication experiments are an important limitation to stone tool research, and indirectly, the study of human evolution. Without the possibility of feasibly addressing broader questions, such as the stochastic *appearance* of complex knapping sequences, or the equifinality of different tool forms, our insights will necessarily be sparse, and will be based for some time on only a handful of experiments. To tackle the constraints of lithic experimentation, it

would therefore be beneficial to find an alternative that could be considerably faster, but still be a suitable – i.e. valid – substitute for real-life knapping.

One possible solution that we ourselves pursued was to simulate the process of knapping in a computer environment, where the process of raw material collection and storage, and the measurement and analysis of lithics, could all be accomplished virtually – and thus, cheaply – in a matter of minutes.
Moreover, since the data used in the program would be already digital, making copies and sharing entire datasets, even one containing tens of thousands of lithics, would be comparatively effortless, and would allow researchers to feasibly reproduce lithic experiments with little cost involved, as researchers could digitally make a perfect copy of any and all unique cores used in an experiment. In addition, knapping software would not suffer from fatigue, lack of motivation, or require rest (or even sleep), unlike a real-life knapper, rendering it capable of continuously knapping for hours and days, if necessary. Depending on underlying programming, the program could also remain at a constant *skill level*, which is not the case in human knappers, who will likely learn over time. Furthermore, whilst different human knappers will inevitably all have varying levels of skill as well, a computer program could be perfectly identical to its copies instead.

In summary, a computer-based model for fast and accurate virtual knapping simulation (externally validated against an archaeological or experimental dataset) could allow lithic experiments at a fraction of the time and resources, and also eschewing the issue of various real-life knapper biases. In addition, a virtual knapping program would permit experiments to be more easily reproduced, allow more effective data sharing, and provide the possibility of generating large virtual lithic assemblages that could be studied and compared with additional archaeological and experimental data.

Recent advances in machine learning – and especially artificial neural networks – have allowed for researchers to explore a wide range of applications across numerous fields of science and technology (Kumar et al. 2012; e.g. Schwarting, Alonso-Mora and Rus 2018; van Ginneken et al. 2015). In the last few years, machine learning methods have also been applied to archaeological research (Grove and Blinkhorn 2020; Lambers, Verschoof-van der Vaart and Bourgeois 2019; Orengo et al. 2020).
Neural networks are useful for problems where the data is highly dimensional, where there are a large number of variables, and where these variables have complex interactions that render modelling the data using more traditional methods difficult. The primary goal of a virtual knapping program (i.e. the prediction of the shape of a flake from that of an intact core) is one such problem.
We sought to explore the capabilities of neural networks to serve as the basis for a proof of concept for a *virtual knapper* program.

2. A Computer-Based Alternative

2.a Proposed framework

Machine learning models to predict one 3D shape from another 3D shape are still limited in scope, as common applications of machine learning using 3D data include object recognition, segmentation (Ahmed et al. 2019), human pose estimation (e.g. Marin-Jimenez et al. 2018), shape reconstruction (e.g. Soltani et al. 2017), and inpainting (Wang et al. 2017).
The lack of established methods for predicting a 3D object from another remains a limitation in how straightforward a virtual knapper framework could be, as it would require a workaround that allowed both

machine learning and 3D data to work together. In this case, the workaround was to first consider the problem in the realm of predicting one 2D image from another.

The first candidate for 2D image prediction was an *encoder-decoder network*, also known as an *autoencoder* (Nguyen et al. 2019).
In order to be able to use this architecture specifically for simulating knapping, however, we needed to encode the 3D surface of the core into a 2D image. Images which perform this function are already common in the field of GIS, wherein 2D rasters can encode terrain elevation information, which can then be re-projected into three dimensions, and serve as the basis for digital elevation models, sometimes known as *heightmaps*.
The surface morphology of our 3D cores and flakes could therefore theoretically be mapped to 2D images in a manner similar to heightmaps, with what are known in the field of computer graphics as *depth maps*, as they encode the depth of the object's surface in three-dimensional space.

We would align the core so the point of percussion would be in the same location for every core: at the centre of the image, at the exact same height, and at the exact same depth for every core. Depth maps could be clipped, or be set-up to have a maximum depth, beyond which any object or part of any object would not be visible in the depth map image, and we would use the platform depth to define the maximum depth value for each core. We also envisioned that the depth map would be captured with the platform surface perpendicular to the image, and as horizontal as possible.

In order to test the feasibility of this framework, we developed a simplified toy model. The goal of the model was to generate input data that would be comparable to the ideal processed input data of a virtual knapper program; i.e. the depth maps of cores in a standard orientation. With these data, it would train and evaluate a simple autoencoder machine learning architecture to predict the resulting flake shape from the input core depth map alone.

2.b Initial Toy Model (*Krakatau Deepfake*)

Our toy model was conceptualized as an explosive volcanic eruption model in order to describe its functionality in less technical terms. The model was thus named *Krakatau*, in reference to the explosive 1883 eruption of the volcano of the same name, which radically altered the topography of the area.

Following the depth map–heightmap analogy, we could then imagine a heightmap for the landscape of a volcano, and the goal would be to predict how the volcanic eruption would affect the landscape; i.e. to predict the heightmap of the post-eruption landscape. This would give us the information of the volume and distribution of the lost material of volcano, as the difference between the pre- and post-eruption landscape (i.e. the amount of volume of material that was lost), so that when superimposing the lost material on top of the post-eruption volcano, we would obtain once more the shape of the pre-eruption volcano. Therefore, any one data point could be re-created with the remaining two (see Fig. 1).

A set of twenty thousand heightmaps consisting of *volcanoes*, *post-eruption volcanoes*, and *material lost from eruption* were generated using Python 3 (Van Rossum and Drake 2009), as well as the NumPy (Oliphant 2006) and Matplotlib (Hunter 2007) libraries. The heightmaps of the *volcanoes* were generated to resemble idealized depth maps of standardized cores from Dibble and Režek (2009), and the process used to obtain the remaining two images of the set was to generate the heightmaps for the lost material, and subtract it from the volcano heightmaps,

obtaining the post-eruption volcano heightmaps.
The method used for generating the heightmaps was to plot from two probability distributions to create a 2D surface, with one distribution providing the shape of the x-axis, and the other, the shape of the y-axis. When combining two probabilities in two-dimensional space, a 2D probability distribution surface emerges, which could then be used as a heightmap (see Fig. 2).

The x-axis distribution shape was based on a normal distribution, and the y-axis shape on a non-central chi-squared distribution. The standard deviation of the former and the λ value of the latter were randomized for every heightmap. In addition, the maximum height of each volcano was also randomized, to simulate knapping different platform depths for each core.
The mean of the normal distribution was set to the horizontal centre of the image, whilst the y-axis distribution was shifted down a few pixels to leave a small gap at the top of the image.
In more technical terms, the data generation program took 2400000 random samples from each distribution and plotted them in a 2D histogram with 256x256 bins (see Fig. 3).

To build and train the neural network, we used the Tensorflow library (Abadi et al. 2016) with Python 3 (Van Rossum and Drake 2009), as well as the NumPy (Oliphant 2006) and Matplotlib (Hunter 2007) libraries. The neural network architecture used was a shallow autoencoder network.
The autoencoder was trained with 15000 heightmaps from our dataset (75%). The model was trained for a total of 150 epochs, and subsequently tested with the remaining 5000 *volcano* heightmaps, obtaining predictions of their respective *material lost* heightmaps. We used the predicted heightmaps to predict the *post-eruption volcano* heightmaps. Finally, the predicted *post-eruption volcano* heightmaps were compared to their matching *actual* heightmaps to measure the model's accuracy using the mean root-mean-square error (RMSE) across all predictions.

We obtained an RMSE of less than 0.1, which indicated a high accuracy of prediction of the *post-eruption* – as well as the *material removed* – heightmaps. As the range of the data was [0, 1], the RMSE was less than 10% of the range of the data. The error was considerably small, which was a promising sign that a similar framework could be applicable to the prediction of flakes using 3D data.

Nevertheless, this dataset lacked considerable amounts of variability, as all the cores and flakes had very similar shape, with the primary difference being how *stretched* this basic topography was. The use of a very simple RMSE loss function during training affected prediction results, as it led to a smoothing and averaging of the reconstructed images, rendering each prediction more of a slightly varying average of all *material lost* heightmaps, rather than individual predictions of each *eruption*.
It was clear that despite the promising results, a more robust machine learning algorithm would be necessary for more accurate results.

To overcome the limitations of the toy model, we proceeded to build a system that could more robustly test our framework by building a more complex model that would use 3D computer-generated cores and flakes, rather than the more abstract 2D data generated for Krakatau.

2.c A Proof of Concept with Computer-Generated 3D Cores and Flakes

For the virtual knapping proof of concept, we used a conditional generative adversarial (neural) network (CGAN) architecture for the machine learning model, and generated

a dataset of 3D cores and flakes (n = 2010) from 3D models of glass cores similar to those used in Dibble and Režek (2009), as described in our main publication (Orellana Figueroa et al. 2021). Note that all the details on the methods used and results obtained can be found in the main publication; below we will merely summarize the main aspects of the proof of concept.

With the generated 3D cores and flakes, we applied the depth map generation methodology we had conceptualized for our toy model's data generation, including a standardized location for the point of percussion, and making platform surface perpendicular to the image. Since, however, our 3D data only consisted of modified cores and their refitting flakes, we had to capture the depth maps of the dorsal flake surfaces and superimpose them on the depth maps of the modified cores to calculate those of the core surface prior to knapping.

Using the depth maps of the intact cores we trained our CGAN to predict the depth maps of the volume removed, which could together be used to calculate the modified core surface (i.e. the flake scar), as we did for the *Krakatau* model. With the intact and the predicted modified core surfaces, we could then create 3D models of the predicted flake removals, which we could visually compare to the original cores in our dataset.

More statistical analyses were also undertaken. We calculated the mean RMSE and mean Normalized RMSE (NRMSE) for the prediction of the shape of the flake removals, as well as the R^2 of the predicted vs. actual flake length, width, and the cube root of the flake volumes.

We trained our CGAN with 70% (n = 1801) of our total dataset for 150 epochs (with a total training time of approximately 150 minutes), reserving the remaining 30% (n = 603) for holdout testing. The prediction of the 603 flake removals, as well as the analyses, and the generation of the 3D models of the predicted flakes took less than 10 minutes in total; a clear signal of how fast and efficient a virtual knapping program could be for performing lithic replication experiments.

The trained model had a high prediction accuracy in flake length ($R^2 = 0.85$), volume ($R^2 = 0.77$), and was reasonably accurate when predicting flake width ($R^2 = 0.58$); with an R^2 value of 1.00 implying perfect prediction. For the prediction of overall flake shape, we obtained a mean RMSE of 0.028 and a mean NRMSE of 3.7%; with RMSE and NRMSE values of 0.00 implying perfect prediction (Orellana Figueroa et al. 2021).

In addition, the predicted 3D flakes were in many cases remarkably similar to the original flakes in the testing dataset (see Fig. 4), though the visual comparison should remain only a crude exercise for now, due to the manual resizing required to make the predicted flake match the scale of the original one (see Orellana Figueroa et al. 2021: 10).

Overall, the results suggest our virtual knapping framework was successful in accurately predicting the shape of flake removals by observing only the depth maps encoding the information of the intact core surface.

3. Future Applications: A Virtual Refitter?

Refitting flakes to their matching core scars is yet another important tool for lithic studies, as it allows archaeologists to reconstruct the reduction sequence of the lithic material in an archaeological site, allowing them to make inferences regarding the methods of flake reduction used, the transport of material, and site formation processes of archaeological sites (Schick and Toth 2006a: 30–31). Refitting, however, is a time-consuming task, as archaeologists must sift through possibly large amounts of lithic material in an attempt to find two matching pieces. Furthermore, a flake may refit to a core, but only once the

intermediate refits have been found; however, it is possible that those intermediate refits are not present in the site, and that cortical pieces, which are very useful guides for refitting, may also not present, making the refitting process far more difficult and time-consuming (Schick, Toth and Semaw 2006: 212).

The success of our proof of concept could thus open still more possibilities. The concept behind the virtual knapping framework could be applied in a different manner, attempting to predict the possible *matching* flake from the flake scar of a modified core. We developed the idea into another proof of concept (see below); a very early and still very limited prototype for a program capable of pairing matching flakes and cores to aid in the refitting process, which could ultimately pave the way for the development of a *virtual refitting* program.

With the depth maps of the dorsal surfaces of the flakes from the computer-generated dataset – already captured for the virtual knapping framework – and using the same modified core depth maps captured earlier, we trained another CGAN to predict the flake dorsal surface. After training the model with 70% (n = 2020) of the modified core and dorsal flake depth map dataset, we used the remaining 30% (n = 606) of cores as the testing dataset. Importantly, our testing input included only the depth maps of the modified cores, and none of the flakes; whilst the model provided the predicted depth map of the dorsal flake surface as output. However, the dataset used still contained some core and flake pairs that were later removed for the reported runs of the virtual knapping proof of concept, explaining the minor discrepancy in dataset size between the two results.

After the predicted flake depth maps were obtained, we compared them with every depth map in our testing dataset, and *ranked* the latter by how similar they were to the predicted depth map.

The model was able to nevertheless provide accurate results, predicting the matching flake well enough that for 34.16% of the testing dataset (207/606) the most closely-matching flake was the actual refitting flake (see Fig. 5 for a visualization).
The model was able to put the matching refit within the predicted top three most likely refit for 56.11% of the testing dataset (340/606). For the top five, this increased to 64.19% of the dataset (389/606). For the top 10 results, it was further increased to 73.43% (445/606).
For comparison, the probability of randomly placing the matching refit in the top 10 is only 1.65%, whilst placing it as the matching refit (i.e. most likely match) is ten times lower.

The results suggest that our prototype virtual refitter is able to *narrow* the list of possible flakes that would fit the flake scar on a modified core from an entire assemblage down to a short list of ten with a high degree of accuracy; suggesting also that the broader problem of automated refitting could be solvable using machine learning methods. A full virtual refitter, if similarly able to narrow down the search space of possible flakes, could become a very useful tool to reduce the time and difficulty of real-life refitting.

However, we must emphasize that our virtual refitter is currently still extremely limited in both scope and functionality. Firstly, the program does not address the issue of fragmentary flakes, nor does it make use of colour and texture information in the lithics, an important aspect for refitting. The flakes and cores must also follow an extremely strict alignment paradigm, rendering the program as it stands right now impossible to use without 3D scanning and aligning every piece perfectly, a difficult and time-consuming task when building a training dataset for the virtual knapping model, but prohibitive for the single purpose of matching a core and a flake in an archaeological assemblage.

In addition, attempts at finding a matching flake would likely be hindered by the fact that the intermediate stages of the flake may no longer exist. Any flake scar that was partially or fully occluded by a subsequent flake removal (and its flake scar) could not be used with our current prototype, as the partiality of the scar would cause the model to predict a partial flake to match with, which would not be accurate. In essence, as of this stage, this prototype can only work with intact flake scars, which will necessarily only come from the last flakes removed from the cores, and partial flake scars could only be used if the core is refit, and if the refit is able to restore the full flake scar. If even one flake of the core reduction sequence were not present in the assemblage, as is common with archaeological assemblages, the system could likely become impossible to use.

Moreover, there are use cases where predicting a core from a flake is much more desirable than the reverse, thus likely requiring many more data for training than the simpler prediction of the dorsal surface of a flake from a modified core we used here. Trying to arrive at a true virtual refitter from this prototype requires important additions and improvements so as to have a more practical process that is able to fulfil the same all-important goal as our virtual knapping program; namely, a faster and easier process than its real-life equivalent.

4. Discussion

We have shown that machine learning, and more specifically, neural networks, have the potential to become important building blocks in new tools for archaeological research. We conceived a machine learning-based framework for virtual knapping and tested its suitability for future exploration – initially – with a toy model using computer-generated 2D data as idealized depth maps of 3D cores and flakes, and after obtaining promising results, subsequently with a more developed proof of concept using computer-generated 3D cores and flakes and their respective depth maps.

The results from the virtual knapping proof of concept allowed us to consider additional applications for the framework. This led to the conceptualization for the possible development of an automated virtual refitting tool capable of finding – from a flake scar on a core – the most likely matches for a refitting flake from an entire dataset of them (a virtual refitter). Although still rather basic in methodology, we nevertheless obtained results that showed that the model was quite accurate at finding the matching refit within the 10 most likely flakes it suggested.

However, there are important limitations for all the proofs of concept presented here. Firstly, none of our approaches assumed differences in raw materials; in fact, neither the data generation nor the neural networks took raw material into account at all, though it could be theoretically possible to encode raw material information into our depth maps (e.g. by using false-colour, rather than monochrome images).

One important limitation with the current approach for virtual knapping (both for the initial toy model and the more robust proof of concept) is the assumption made that all flake removals would be successful. Failed flake removals are common, especially with novices (Pargeter et al. 2020), and are thus important to include in a future virtual knapping framework.

One possible approach would be to use all-black depth maps to signal that no flake mass has been removed from the core. Implementing additional knapping variables could also be encoded into depth maps in a similar manner to the platform depth and exterior platform angle. Other knapping variables, such as hammer hardness, could also be integrated into a virtual knapper through the training of different models for hard and soft hammers.

These limitations, although important, could likely be solved with additional

development on the virtual knapping framework presented here, and should not wholly detract from the success obtained with its proof of concept.

We have shown that neural networks could be used to simulate flake removal, and we can begin evaluating our model's performance on actual core and flake pairs. The creation of such a dataset will require considerable effort, but will provide the model with additional validity (Lin, Rezek and Dibble 2018).
Furthermore, transfer learning could be applied to the currently trained model by allowing it to take advantage of the training already performed with the large generated training data, but also made more accurate to real-world data by training it once again on a smaller – more valid – dataset of actual cores and flakes, eschewing the need of creating very large training datasets of 3D-scanned lithics.
It must be stressed, however, that the approach taken for virtual knapping is not a replacement for other approaches, but rather has been – and must be – complemented by work from other groups, such as those working with machine knapping (Dibble and Režek 2009; Dogandžić et al. 2020; Magnani et al. 2014; Režek et al. 2011), and vice-versa.

Experimental stone tool replication remains an important part of lithic studies, as they allow modern archaeologists to study the influence of different variables during the knapping process, such as technique or raw materials.
However, the substantial raw material and time requirements, as well as the biases from across and within knappers, make traditional lithic experiments difficult to reproduce. A tool that could allow for fast, inexpensive, digital, and singularly-biased lithic reduction could not only provide more robust results, but also be able to generate large and easily-shareable virtual assemblages that could be used as a comparison with archaeological or experimental assemblages, as well as explore how differences across knappers affect the products of lithic reduction.
Our virtual knapping framework, trained on a larger – and more valid – dataset could serve as a very important tool for lithic studies, helping researchers better address questions of prehistoric stone tool production.

Furthermore, our idea for a virtual refitter, although more limited in the scope of application compared to a virtual knapper, would nevertheless also serve as an important tool for archaeologists, especially those that must analyse archaeological lithic assemblages.
Refitting can be highly time-consuming (more so than knapping), and could be well-served by an automated computerized tool to assist the work. A program capable of not only finding a flake matching a flake scar, but also to piece it back together in 3D, recreating also the original core digitally (and finding new refits in turn), could become indispensable for lithic analysis in the field or in the lab.
The initial virtual refitting model presented here, although still very simple and very limited, shows both the potential and challenges of applying computational models such as neural networks for archaeological research.

List of Figures

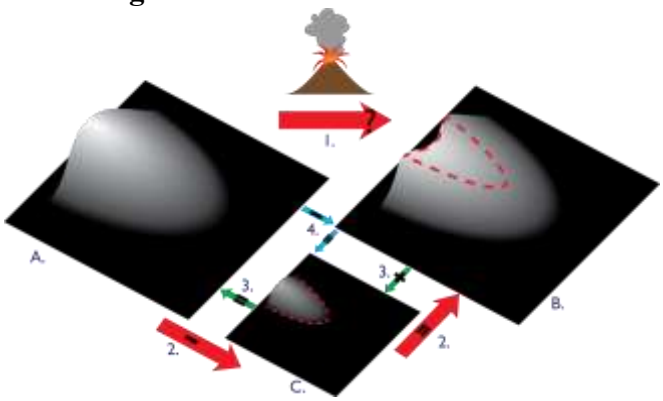

Figure 1. Diagram showing an overview of the Krakatoa model data concept. A: Original volcano heightmap. B: Post-eruption volcano heightmap. C: Heightmap of the volume of material removed from the original volcano during eruption. 1: Process of eruption turns *A* into *B*, the latter is what we wish to predict from the former. 2: When we subtract the lost material from the original volcano surface (essentially, what the eruption does), we obtain the modified volcano surface (red arrows). 3. When we perform the inverse operation, and add the lost material back on to the volcano post-eruption, we reconstruct the original surface of the volcano landscape (green arrows). 4: The heightmap of *C* can be obtained by subtracting *B* from *A* (blue arrows). Note that, in practice, all the heightmaps are images with the same dimensions.

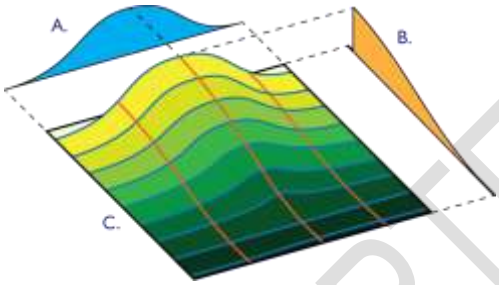

Figure 2. Diagram depicting the generation of the heightmaps through the use of two probability distributions. A: Normal distribution used for the shape of the x-axis. B: Non-central chi-squared distribution used for the shape of the y-axis. C: Combination of both distributions in two dimensions, generating a surface resembling an idealized depth map of a core.

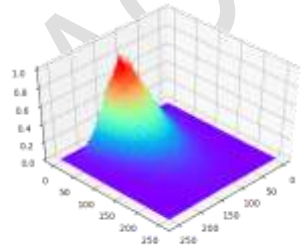

Figure 3. Example of a normalized *volcano* heightmap. Note that the maximum height of the *volcano* was set during generation to a random value – in this case, slightly above 0.9.

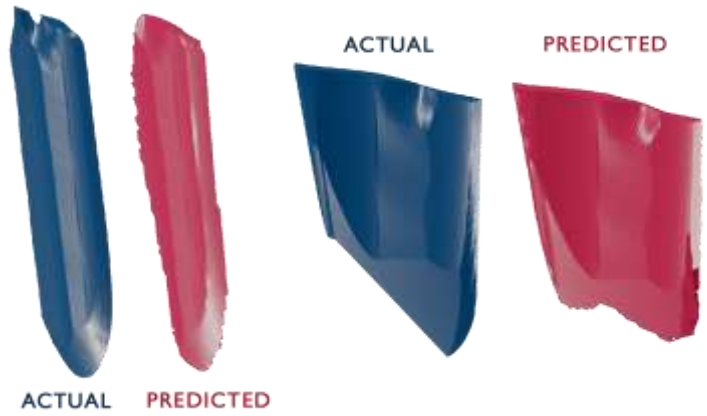

Figure 4. Side-by-side comparison of two predicted flakes with their counterparts in the testing dataset.

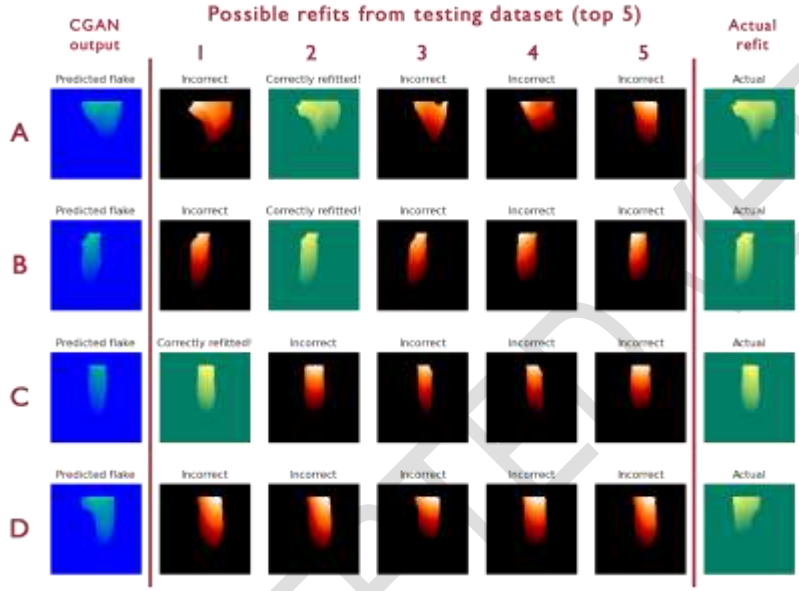

Figure 5. Visualization of the output of the virtual refitter prototype. On the left (*CGAN output*) is the predicted shape of the flake. In the middle are shown the top five possible refits selected by the program. On the right (*Actual refit*) is the actual refitting flake being sought. The refitter placed the correct refit as the second most likely match for A and B, and in the most likely match for C. The refitter could not place the actual refit in any of the top five positions for D.

References

Abadi, M, Agarwal, A, Barham, P, Brevdo, E, Chen, Z, Citro, C, Corrado, GS, Davis, A, Dean, J, Devin, M, Ghemawat, S, Goodfellow, I, Harp, A, Irving, G, Isard, M, Jia, Y, Jozefowicz, R, Kaiser, L, Kudlur, M, Levenberg, J, Mane, D, Monga, R, Moore, S, Murray, D, Olah, C, Schuster, M, Shlens, J, Steiner, B, Sutskever, I, Talwar, K, Tucker, P, Vanhoucke, V, Vasudevan, V, Viegas, F, Vinyals, O, Warden, P, Wattenberg, M, Wicke, M, Yu, Y and Zheng, X. 2016 TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467*.

Ahmed, E, Saint, A, Shabayek, AER, Cherenkova, K, Das, R, Gusev, G, Aouada, D and Ottersten, B. 2019 A survey on Deep Learning Advances on Different 3D Data Representations. *arXiv:1808.01462 [cs]*.

Allington-Jones, L. 2015 The Clacton Spear: The Last One Hundred Years. *Archaeological Journal* 172(2): 273–296. DOI: https://doi.org/10.1080/00665983.2015.1008839.

Backwell, LR and d'Errico, F. 2001 Evidence of termite foraging by Swartkrans early hominids. *Proceedings of the National Academy of Sciences* 98(4): 1358–1363. DOI: https://doi.org/10.1073/pnas.98.4.1358.

Braun, DR, Aldeias, V, Archer, W, Arrowsmith, JR, Baraki, N, Campisano, CJ, Deino, AL, DiMaggio, EN, Dupont-Nivet, G, Engda, B, Feary, DA, Garello, DI, Kerfelew, Z, McPherron, SP, Patterson, DB, Reeves, JS, Thompson, JC and Reed, KE. 2019 Earliest known Oldowan artifacts at >2.58 Ma from Ledi-Geraru, Ethiopia, highlight early technological diversity. *Proceedings of the National Academy of Sciences* 116(24): 11712–11717. DOI: https://doi.org/10.1073/pnas.1820177116.

Braun, DR and Hovers, E. 2009 Introduction: Current Issues in Oldowan Research. In: Hovers, E and Braun, DR (eds.). *Interdisciplinary approaches to the Oldowan*. Vertebrate paleobiology and paleoanthropology series. Dordrecht, Netherlands: Springer. pp. 1–14.

Braun, DR, Plummer, TW, Ditchfield, PD, Bishop, LC and Ferraro, JV. 2009 Oldowan Technology and Raw Material Variability at Kanjera South. In: Hovers, E and Braun, DR (eds.). *Interdisciplinary approaches to the Oldowan*. Vertebrate paleobiology and paleoanthropology series. Dordrecht, Netherlands: Springer. pp. 99–110.

Braun, DR, Tactikos, JC, Ferraro, JV, Arnow, SL and Harris, JWK. 2008 Oldowan reduction sequences: Methodological considerations. *Journal of Archaeological Science* 35(8): 2153–2163. DOI: https://doi.org/10.1016/j.jas.2008.01.015.

d'Errico, F and Backwell, LR. 2003 Possible evidence of bone tool shaping by Swartkrans early hominids. *Journal of Archaeological Science* 30(12): 1559–1576. DOI: https://doi.org/10.1016/S0305-4403(03)00052-9.

Dibble, HL, Holdaway, SJ, Lin, SC, Braun, DR, Douglass, MJ, Iovita, R, McPherron, SP, Olszewski, DI and Sandgathe, D. 2017 Major Fallacies Surrounding Stone Artifacts and Assemblages. *Journal of Archaeological Method and Theory* 24(3): 813–851. DOI: https://doi.org/10.1007/s10816-016-9297-8.

Dibble, HL and Režek, Z. 2009 Introducing a new experimental design for controlled studies of flake formation: Results for exterior platform angle, platform depth, angle of blow, velocity, and force. *Journal of Archaeological Science* 36(9): 1945–1954. DOI: https://doi.org/10.1016/j.jas.2009.05.004.

Dogandžić, T, Abdolazadeh, A, Leader, G, Li, L, McPherron, SP, Tennie, C and Dibble, HL. 2020 The results of lithic experiments performed on glass cores are applicable to other raw materials. *Archaeol Anthropol Sci* 12(2): 44. DOI: https://doi.org/10.1007/s12520-019-00963-9.

Eren, MI, Lycett, SJ, Patten, RJ, Buchanan, B, Pargeter, J and O'Brien, MJ. 2016 Test, Model, and Method Validation: The Role of Experimental Stone Artifact Replication in

Hypothesis-driven Archaeology. *Ethnoarchaeology* 8(2): 103–136. DOI: https://doi.org/10.1080/19442890.2016.1213972.
Grove, M and Blinkhorn, J. 2020 Neural networks differentiate between Middle and Later Stone Age lithic assemblages in eastern Africa. *PLOS ONE* 15(8): e0237528. DOI: https://doi.org/10.1371/journal.pone.0237528.
Hunter, JD. 2007 Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9(3): 90–95. DOI: https://doi.org/10.1109/MCSE.2007.55.
Kivell, TL. 2015 Evidence in hand: Recent discoveries and the early evolution of human manual manipulation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370(1682): 20150105. DOI: https://doi.org/10.1098/rstb.2015.0105.
Kumar, N, Belhumeur, PN, Biswas, A, Jacobs, DW, Kress, WJ, Lopez, IC and Soares, JVB. 2012 Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In: Fitzgibbon, A, Lazebnik, S, Perona, P, Sato, Y, and Schmid, C (eds.). *Computer Vision ECCV 2012*. Lecture Notes in Computer Science. 2012. Berlin, Heidelberg: Springer. pp. 502–516. DOI: https://doi.org/10.1007/978-3-642-33709-3_36.
Lambers, K, Verschoof-van der Vaart, W and Bourgeois, Q. 2019 Integrating Remote Sensing, Machine Learning, and Citizen Science in Dutch Archaeological Prospection. *Remote Sensing* 11(7): 794. DOI: https://doi.org/10.3390/rs11070794.
Lin, SC, Rezek, Z and Dibble, HL. 2018 Experimental Design and Experimental Inference in Stone Artifact Archaeology. *J Archaeol Method Theory* 25(3): 663–688. DOI: https://doi.org/10.1007/s10816-017-9351-1.
Magnani, M, Režek, Z, Lin, SC, Chan, A and Dibble, HL. 2014 Flake variation in relation to the application of force. *Journal of Archaeological Science* 46: 37–49. DOI: https://doi.org/10.1016/j.jas.2014.02.029.
Marin-Jimenez, MJ, Romero-Ramirez, FJ, Muñoz-Salinas, R and Medina-Carnicer, R. 2018 3D human pose estimation from depth maps using a deep combination of poses. *arXiv:1807.05389 [cs]*.
McGrew, WC. 2010 Chimpanzee Technology. *Science* 328(5978): 579–580. DOI: https://doi.org/10.1126/science.1187921.
Moore, MW and Perston, Y. 2016 Experimental Insights into the Cognitive Significance of Early Stone Tools Petraglia, MD (ed.). *PLOS ONE* 11(7): e0158803. DOI: https://doi.org/10.1371/journal.pone.0158803.
Musgrave, S and Sanz, C. 2018 Tool Use in Nonhuman Primates. In: Callan, H (ed.). *The International Encyclopedia of Anthropology*. Oxford, UK: John Wiley & Sons, Ltd. pp. 1–7. DOI: https://doi.org/10.1002/9781118924396.wbiea2063.
Nguyen, TT, Nguyen, CM, Nguyen, DT, Nguyen, DT and Nahavandi, S. 2019 Deep Learning for Deepfakes Creation and Detection. *arXiv:1909.11573 [cs, eess]*.
Oliphant, TE. 2006. *A guide to NumPy*. Trelgol Publishing USA.
Orellana Figueroa, JD, Reeves, JS, McPherron, SP and Tennie, C. 2021 A Proof of Concept for Machine Learning-Based Virtual Knapping Using Neural Networks. *Open Science Framework Preprints* DOI: https://doi.org/10.31219/osf.io/9uybv.
Orengo, HA, Conesa, FC, Garcia-Molsosa, A, Lobo, A, Green, AS, Madella, M and Petrie, CA. 2020 Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. *Proceedings of the National Academy of Sciences* 117(31): 18240–18250. DOI: https://doi.org/10.1073/pnas.2005583117.
Pargeter, J, Khreisheh, N, Shea, JJ and Stout, D. 2020 Knowledge vs. Know-how? Dissecting the foundations of stone knapping skill. *Journal of Human Evolution* 145: 102807. DOI: https://doi.org/10.1016/j.jhevol.2020.102807.

Perreault, C. 2019. *The quality of the archaeological record*. Chicago: The University of Chicago Press.
Pruetz, JD and Bertolani, P. 2007 Savanna Chimpanzees, Pan troglodytes verus, Hunt with Tools. *Current Biology* 17(5): 412–417. DOI: https://doi.org/10.1016/j.cub.2006.12.042.
Putt, SSJ, Wijeakumar, S and Spencer, JP. 2019 Prefrontal cortex activation supports the emergence of early stone age toolmaking skill. *NeuroImage* 199: 57–69. DOI: https://doi.org/10.1016/j.neuroimage.2019.05.056.
Putt, SS, Wijeakumar, S, Franciscus, RG and Spencer, JP. 2017 The functional brain networks that underlie Early Stone Age tool manufacture. *Nature Human Behaviour* 1(6): 0102. DOI: https://doi.org/10.1038/s41562-017-0102.
Putt, SS, Woods, AD and Franciscus, RG. 2014 The Role of Verbal Interaction During Experimental Bifacial Stone Tool Manufacture. *Lithic Technology* 39(2): 96–112. DOI: https://doi.org/10.1179/0197726114Z.00000000036.
Režek, Z, Lin, S, Iovita, R and Dibble, HL. 2011 The relative effects of core surface morphology on flake shape and other attributes. *Journal of Archaeological Science* 38(6): 1346–1359. DOI: https://doi.org/10.1016/j.jas.2011.01.014.
Schick, KD and Toth, N. 2006a An Overview of the Oldowan Industrial Complex: The sites and the nature of their evidence. In: Schick, KD and Toth, NP (eds.). *The Oldowan: Case studies into the earliest Stone Age*. Stone Age Institute publication series. Gosport, IN: Stone Age Institute. pp. 3–42.
Schick, KD and Toth, NP (eds.). 2006b. *The Oldowan: Case studies into the earliest Stone Age*. Stone Age Institute publication series no. 1. Gosport, IN: Stone Age Institute.
Schick, KD, Toth, N and Semaw, S. 2006 A Comparative Study of the Stone Tool-Making Skills of *Pan*, *Australopithecus*, and *Homo Sapiens*. In: Schick, KD and Toth, NP (eds.). *The Oldowan: Case studies into the earliest Stone Age*. Stone Age Institute publication series. Gosport, IN: Stone Age Institute. pp. 155–222.
Schwarting, W, Alonso-Mora, J and Rus, D. 2018 Planning and Decision-Making for Autonomous Vehicles. *Annual Review of Control, Robotics, and Autonomous Systems* 1(1): 187–210. DOI: https://doi.org/10.1146/annurev-control-060117-105157.
Semaw, S, Renne, P, Harris, JWK, Feibel, CS, Bernor, RL, Fesseha, N and Mowbray, K. 1997 2.5-million-year-old stone tools from Gona, Ethiopia. *Nature* 385(6614): 333–336. DOI: https://doi.org/10.1038/385333a0.
Semaw, S, Rogers, MJ, Quade, J, Renne, PR, Butler, RF, Domínguez-Rodrigo, M, Stout, D, Hart, WS, Pickering, T and Simpson, SW. 2003 2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia. *Journal of Human Evolution* 45(2): 169–177. DOI: https://doi.org/10.1016/S0047-2484(03)00093-9.
Snyder, WD, Reeves, JS and Tennie, C. 2021 Early knapping techniques do not necessitate cultural transmission. *Open Science Framework Preprints* DOI: https://doi.org/10.31219/osf.io/ph6gw.
Soltani, AA, Huang, H, Wu, J, Kulkarni, TD and Tenenbaum, JB. 2017 Synthesizing 3D Shapes via Modeling Multi-view Depth Maps and Silhouettes with Deep Generative Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017. Honolulu, HI: IEEE. pp. 2511–2519. DOI: https://doi.org/10.1109/CVPR.2017.269.
Stammers, RC, Caruana, MV and Herries, AIR. 2018 The first bone tools from Kromdraai and stone tools from Drimolen, and the place of bone tools in the South African Earlier Stone Age. *Quaternary International* 495: 87–101. DOI: https://doi.org/10.1016/j.quaint.2018.04.026.

Susman, RL. 1988 Hand of Paranthropus robustus from Member 1, Swartkrans: Fossil evidence for tool behavior. *Science* 240(4853): 781–784. DOI: https://doi.org/10.1126/science.3129783.
Tennie, C, Premo, LS, Braun, DR and McPherron, SP. 2017 Early Stone Tools and Cultural Transmission: Resetting the Null Hypothesis. *Current Anthropology* 58(5): 652–672. DOI: https://doi.org/10.1086/693846.
Thieme, H. 1997 Lower Palaeolithic hunting spears from Germany. *Nature* 385(6619): 807–810. DOI: https://doi.org/10.1038/385807a0.
Toth, N. 1985 The oldowan reassessed: A close look at early stone artifacts. *Journal of Archaeological Science* 12(2): 101–120. DOI: https://doi.org/10.1016/0305-4403(85)90056-1.
van Ginneken, B, Setio, AAA, Jacobs, C and Ciompi, F. 2015 Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. April 2015. pp. 286–289. DOI: https://doi.org/10.1109/ISBI.2015.7163869.
Van Rossum, G and Drake, FL. 2009. *Python 3 reference manual*. Scotts Valley, CA: CreateSpace.
Wang, W, Huang, Q, You, S, Yang, C and Neumann, U. 2017 Shape Inpainting Using 3D Generative Adversarial Network and Recurrent Convolutional Networks. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. October 2017. Venice: IEEE. pp. 2317–2325. DOI: https://doi.org/10.1109/ICCV.2017.252.
Warren, SH. 1922 The Mesvinian Industry of Clacton-on-Sea, Essex. *Proceedings of the Prehistoric Society of East Anglia* 3(4): 597–602. DOI: https://doi.org/10.1017/S0958841800024765.