The Virtual Knapper:
A Realistic Software Program Approach to Virtually Recreate Early Stone Tool Forms

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Jordy Didier Orellana Figueroa

aus San Salvador, El Salvador

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	12.02.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	PD Dr. Claudio Tennie
2. Berichterstatter/-in:	Prof. Dr. Nicholas Conard
3. Berichterstatter/-in:	Prof. Dr. Andreas Maier

Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement n°714658; (STONECULT project).

The journey here has been a long one, but it all would not have been possible without the support, collaboration, encouragement, feedback, love, and friendship of so many people. I feel an immense privilege to have been able to work in this field, in academic research, to have met so many great people, so many great teachers and scientists; to have been able to work alongside them, to have been able to learn from them so much. When I think from the perspective of my younger self, it seems more like a dream than reality; never could I have imagined where I would be today. However, I also feel an immense privilege to have been able to bring this work to its completion in the first place; not always a given for many people. No amount of words would be enough to express my gratitude to each and every person that was a part of this journey with me, though I will endeavour to do so regardless.

Firstly, I give my deepest gratitude especially to my parents, who have always loved me, helped me, supported me, and encouraged me from so far away, and always unconditionally. Without a doubt I could have never been able to be where I am without them. I am both lucky and grateful that despite the many hardships they have endured, they are still present and in good health, and that I can share with them this accomplishment for which they have been pivotal. I am thankful also to the rest of my family, who were always by my side supporting me despite the large physical distance between us.

I wish to express my thanks to my supervisor, Dr Claudio Tennie, without whose teaching, encouragement, support, funding, and supervision, as well as the substantial research work carried out by Claudio to establish the STONECULT project, none of this would have been possible. I am thankful for the many pieces of advice so generously given on everything from simple but important presentation skills to advice on how to plan, structure, and perform rigorous scientific projects. Thank you for providing me the opportunity to work on this project, and for all the help you gave me along the way so I could succeed.

I would like to also thank my other supervisor, Prof. Nicholas J. Conard, who made me feel welcome and encouraged me to become immersed more deeply in archaeology, as well as arranging for my participation in an excavation in the Swabian Jura despite being a complete outsider to the field before starting my doctoral work. I extend my gratitude also to the other members of my doctoral examination committee, as well as Dr Shannon P. McPherron, who always supported and assisted me considerably during my doctoral work as if formally named as my supervisor, and who was extremely vital to the successful development of this project. I am also grateful to Dr Jonathan S. Reeves, whose guidance was made available whenever needed, and whose encouragement was never absent.

I am greatly thankful to the remaining members of my EVEREST Thesis Advisory Committee. To Prof. Isabel Valera, whose guidance and advice especially in scientific matters were crucial to the successful undertaking of my doctoral project. To Dr Korinna Allhoff, whose encouragement and support were also equally instrumental in the successful completion of this work.

To the late Prof. Harold L. Dibble, who was one of the key members of the STONECULT project, but regrettably passed away suddenly before we had a chance to meet in person. It is my hope that Prof. Dibble would have been pleased with the outcome of my research, and with the work that my colleagues in the STONECULT project—as well as his former students—went on to perform.

I am also very thankful to many other collaborators and colleagues, who in one way or another contributed to the sum total of this work. Special thanks to Mehdi S. M. Sajjadi and the rest of Prof. Valera's team for their useful advice on machine learning. To Dr Will Archer for the many discussions we had on flake prediction. To Dr Mark W. Moore, whose research was of key importance to this doctoral work. To Prof. Carel van Schaik, whose advice on navigating a scientific career was significantly influential for my own.

I am also so deeply grateful to all my dear friends and colleagues who made bad days less so, and good days even better; with whom I spent such a wonderful time, with whom I made such lovely memories, and without whom this all could not have been possible.

I am especially thankful to Dr Elisa Bandini, who was so incredibly helpful, patient, forgiving, caring, and supportive throughout the entire time we have known each other; who was always a great colleague and a great friend. I could not have made it through

without you, and I will be forever indebted to you, and by extension to your wonderful family too.

To Dr William D. Snyder, my dearest friend, who was always there whenever I needed help, and who made our time in the office all the more cheery, fun, and pleasant. There is a multitude of things I could say about how much you have helped me throughout the years in so many ways impossible to list, or how important you have been in bringing this doctoral dissertation to fruition, but I can summarise it all by saying you have always been an amazing friend. Thank you, from the bottom of my heart.

To Dr Patrick Cuthbertson, one of the first people I ever met when I arrived in Tübingen, and one who became a good friend despite the short time there was to share in person. To Kim-Louise Krettek, a kind friend who, despite the relatively short time we have known each other, has always supported me whenever I needed it. To my partner and best friend, Anuschka Marie Weiß, who has in innumerable ways helped me along this journey to its completion. To Dr Omar Rafael Regalado Fernández, a good friend who always brightens my mood, and whose friendship would send palaeontology-loving child me over the moon.

To Dr Tobias Massonne, Dr Pascal Abel, Robin Edds, Dr Julian Gieseke, Ruth Rey, Dr Chris Baumann, Ella Reiter, David Boysen, Tina Petersen, Alba Motes-Rodrigo, and Li Li. To my dearest friend Jake Scott, and to my many other friends that are scattered across the world. Thank you all so very much.

I fy nghartref mabwysiadol oddi cartref dw i caru mor annwyl, ac a losgaf i efo hiraeth iddo. I'w fryniau bytholwyrdd hardd a swynodd fi, ei glawogydd meddal a gysurodd fi, ei hanes chwerwfelys a hudodd fi, ei hen iaith a ysbrydolodd fi. I'r holl bobl hyfryd a gyfarfyddais i â nhw yn fy amser yno, hebddyn nhw byddwn i ddim hyd yn oed yma. Dw i'n diolch i chi i gyd o ddyfnderau fy nghalon. I'r holl atgofion llywy a greais i yno bydda i yn eu coleddu am byth. Efallai cariad di-alw, gan y wlad, gan y bobl; dw i'n gobeithio o leiaf maen nhw'n maddau fi'r gwallau a wnes i yma yn fy awydd i ddefnyddio ei hiaith. Gobeithio fydd fy llygaid yn syllu arnat ti eto yn fuan.

Եվ այն բազմաթիվ մարդկանց ընկերներին, որոնց ես դեռ ունեմ պատմությամբ լի ու երաժշտությամբ լի հողերից այդ հողերից, որոնք ես այնքան թանկով եմ փայփայում:

და ბევრ კეთილ მეგობარს, რომელიც შევიძინე კავკასიის მიდამოებში, რომლებმაც გამითბეს გული და ეს მოგზაურობა უფრო მხიარული გახადეს.

Ipal ne taketzalis ka teuk niwelituk nimumachtia, wan ipal ne culturah wan ne tukniwan ka teuk niwelituk nikishmati, ka kinekiket techichtekit muchi wan kinekiket kiwilwewat muchi iwan sujsul ihiyatuk. Ma uni nikpatakan se tunal. Ipal ne nunoywan, wan ne intejteku wan ne ijinan, wan ne intejteku wan ne ijinan. Ma ne tukniwan te kinhelkawakan ne inejnelwaw. Wan ma kinshinechchiwakan nuejerror tik intaketzalis. Ika ne mimet sujsul ejelkawatuk itan ne ital ne Kuskatan ka kinchiwaket te sujsul tekenha ken ne itemet itan ne ital ne Ultupai sujsul mujmumachtiatuk.

Table of Contents

Summary

Prehistoric stone tools are one of the most important types of evidence for the study of hominin evolution, dating back at least 2.6 million years. These tools were most commonly made through the (repeated) fracture of a stone to detach from it a flake, creating—for example—useful cutting edges on the flake surface. This process of flake removal is known as knapping or lithic reduction. One method used to study lithic reduction is its experimental replication by modern humans, making inferences from the process on the various behavioural, cognitive, or even social learning requirements needed to make these artefacts. Historically, many researchers have suggested that the earliest stone tools require—and thus are evidence for—copying of know-how information, which is intrinsic to modern human culture, but which is also not present in any extant non-human ape. On the other hand, recent research has advanced an alternative hypothesis: suggesting many of the earliest stone tools did not require know-how copying, but that they were manufactured through a simple set of rules and with a base set of skills closer to those of non-human apes than modern humans (the 'Zone of Latent Solutions' hypothesis). Nevertheless, in order to robustly test how changing certain variables (e.g. maximising for flake length or area) can affect the products of knapping, many replicable large-scale lithic replication experiments are necessary. These experiments generally require considerable amounts of time, material, and funds to undertake, from the lithic reduction itself, to the cataloguing, storage, measuring, and analysis of the products at each step of the reduction sequence, and are inevitably subject to knapper-derived biases (e.g. experience, motivation, and fatigue during knapping). With replication experiments that can last months, a fully computer-based analogue to rules-based lithic reduction could prove a powerful tool for the study of human behavioural, cognitive, and cultural evolution. Such software could be effective for testing various hypotheses on the factors that affect stone tool manufacture, such as the necessity of know-how copying, helping advance evolutionary science more broadly or as a tool for teaching and outreach, as well as serving as a base for the development of additional tools. This dissertation presents a proof of concept for a computer program based on a machine learning framework able to perform lithic reduction at a fraction of the time and cost, and without knapper-derived biases: a Virtual Knapper. In addition, this work also discusses the results of a proof of concept for another application based on the same machine learning framework: a Virtual Refitter.

Zusammenfassung

Prähistorische Steinwerkzeuge sind eines der wichtigsten Beweismittel für die Erforschung der Evolution der Homininen und reichen mindestens 2,6 Millionen Jahre zurück. Diese Werkzeuge wurden in der Regel durch (wiederholtes) Brechen eines Steins hergestellt, um einen Abschlag abzutrennen und so beispielsweise nützliche Schneidkanten auf der Oberfläche des Abschlags zu erzeugen. Dieser Prozess der Abtrennung von einem Abschlag wird als Knapping oder lithische Reduktion bezeichnet. Eine Methode zur Untersuchung der Steinbearbeitung ist die experimentelle Nachahmung durch moderne Menschen, um aus dem Prozess Rückschlüsse auf die verschiedenen verhaltensmäßigen, kognitiven oder sogar sozialen Lernanforderungen zu ziehen, die für die Herstellung dieser Artefakte erforderlich sind. In der Vergangenheit haben viele Forscher die Ansicht vertreten, dass die frühesten Steinwerkzeuge das Kopieren von Know-How-Information erfordern – und somit ein Beweis dafür sind – was der modernen menschlichen Kultur eigen ist und auch bei keinem lebenden nicht-menschlichen Menschenaffen vorkommt. Andererseits haben neuere Forschungen eine alternative Hypothese aufgestellt, die besagt, dass viele der frühesten Steinwerkzeuge nicht das Kopieren von Know-How erforderten, sondern dass sie nach einfachen Regeln und mit einem Grundstock an Fähigkeiten hergestellt wurden, die denen der nicht-menschlichen Menschenaffen ähnlicher sind als denen des modernen Menschen (die “Zone of Latent Solutions” Hypothese). Um jedoch zuverlässig zu testen, wie sich die Änderung bestimmter Variablen (z. B. die Maximierung der Abschlaglänge oder -fläche) auf die Abschlagsprodukte auswirkt, sind viele wiederholbare, groß angelegte lithische Replikationsexperimente erforderlich. Diese Experimente erfordern in der Regel einen beträchtlichen Zeit-, Material- und Finanzaufwand, angefangen bei der eigentlichen Steinbearbeitung bis hin zur Katalogisierung, Lagerung, Messung und Analyse der Produkte in jedem Schritt der Bearbeitungsfolge, und unterliegen zwangsläufig den durch die Steinschläger verursachten Verzerrungen (z. B. Erfahrung, Motivation und Ermüdung während der Bearbeitung). Mit Replikationsexperimenten, die sich über Monate hinziehen können, könnte sich ein vollständig computergestütztes Analogon zur regelbasierten Steinbearbeitung als leistungsfähiges Instrument für die Untersuchung des menschlichen Verhaltens, der kognitiven und kulturellen Evolution erweisen. Mit einer solchen Software könnten verschiedene Hypothesen über die Faktoren, die die Herstellung von Steinwerkzeugen beeinflussen, getestet werden, wie

z. B. die Notwendigkeit des Kopierens von Know-How, und sie könnte dazu beitragen, die Evolutionswissenschaft im weiteren Sinne voranzubringen. Des Weiteren könnte sie als Werkzeug für die Lehre und die Öffentlichkeitsarbeit sowie als Grundlage für die Entwicklung weiterer Werkzeuge dienen. In dieser Dissertation wird ein Machbarkeitsnachweis für ein auf einem maschinellen Lernverfahren basierte Computerprogramm vorgestellt, das in der Lage ist, lithische Abschläge zu einem Bruchteil der Zeit und Kosten und ohne von Steinschlägern verursachte Verzerrungen durchzuführen: ein virtueller Steinschläger. Darüber hinaus werden in dieser Arbeit auch die Ergebnisse eines Machbarkeitsnachweis für eine andere Anwendung auf der Grundlage desselben maschinellen Lernsystems vorgestellt: ein virtueller Refitter.

List of Publications

Accepted publications included as part of this dissertation

Appendix 1: **Orellana Figueroa JD**, Reeves JS, McPherron SP, Tennie C (2021) A proof of concept for machine learning-based virtual knapping using neural networks. Scientific Reports 11:19966. https://doi.org/10.1038/s41598-021-98755-6.

Appendix 2: **Orellana Figueroa JD**, Reeves JS, McPherron SP, Tennie C (*in press*) Virtual Knapping (and Refitting) with Neural Networks: Proofs of Concept. In: Kyriakidis P, Agapiou A, Leventis G (eds) CAA2021 Digital Crossroads. Proceedings of the 48th Conference on Computer Applications and Quantitative Methods in Archaeology. Tübingen University Press.

Publications from the doctoral work not included as part of this dissertation

Bellat M[*], **Orellana Figueroa JD**[*], Taghizadeh Mehrjadi R, Reeves JS, Scholten T[**], Tennie C[**] (*in prep.*) Machine Learning Applications in Archaeological Practices: A Review.

Snyder WD[*], Boysen D[*], **Orellana Figueroa JD**, et al (*forthcoming*) An overview of standardizable raw materials for controlled knapping experiments. Advances in Archaeological Practice.

Orellana Figueroa JD, Snyder WD (*in prep.*) A Blender-based protocol for rapid 2D image acquisition from 3D artifact data.

[*]: Shared first authorship

[**]: Shared last authorship

Declaration of Personal Contribution

I declare that this dissertation is my own work and has been composed by myself except where explicitly stated otherwise through reference or acknowledgement, and except where the collaborative publications specified above have been included as part of this work. My contribution and that of other collaborators of this work have been explicitly specified below, and in the form 'Declaration according to § 5 Abs. 2 No. 8 of the PhD regulations of the Faculty of Science: Collaborative Publications', which is attached further below.

Introduction. Figures 1 and 2 are used with permission from the specified sources. Figures 3, 4 and 5 were prepared by myself for a collaborative publication included in this work (Appendix 1), and are used with permission.

The work presented in Appendix 1 was previously published in *Scientific Reports* with the title 'A proof of concept for machine learning-based virtual knapping using neural networks' and authors **J.D. Orellana Figueroa**, J.S. Reeves, S.P. McPherron, and C. Tennie. All authors contributed to the conception of this work. S.P.M. created the software for the generation of the input data. I created all other software used in this work, processed and analysed the data, prepared all the figures, and composed the text of this work. J.S.R. performed an independent verification of the results of the machine learning model. All authors contributed with the editing and revisions of the text of this work.

The work presented in Appendix 2 is accepted for publication in *CAA2021 Digital Crossroads. Proceedings of the 48th Conference on Computer Applications and Quantitative Methods in Archaeology* with the title 'Virtual Knapping (and Refitting) with Neural Networks: Proofs of Concept' and authors **J.D. Orellana Figueroa**, J.S. Reeves, S.P. McPherron, and C. Tennie. All authors contributed to the conception of this work. J.S.R. originally conceived the idea of the virtual refitting machine learning model, which was developed further by all authors. S.P.M. created the software for the generation of the input data used for both the knapping and refitting machine learning models described. I created all other software used in this work, including that of the data generation for the toy model described. I processed and analysed the data, prepared all the figures, and composed the text of this work. All authors contributed with the editing and revisions of the text of this work.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Erklärung nach § 5 Abs. 2 Nr. 8 der Promotionsordnung der Math.-Nat. Fakultät
-Anteil an gemeinschaftlichen Veröffentlichungen-
Nur bei kumulativer Dissertation erforderlich!

Declaration according to § 5 Abs. 2 No. 8 of the PhD regulations of the Faculty of Science
-Collaborative Publications-
For Cumulative Theses Only!

Last Name, First Name: Orellana Figueroa, Jordy Didier

List of Publications

1. A proof of concept for machine learning-based virtual knapping using neural networks. Scientific Reports 11:19966. Orellana Figueroa JD, Reeves JS, McPherron SP, Tennie C (2021)
2. Virtual Knapping (and Refitting) with Neural Networks: Proofs of Concept. In: Kyriakidis P, Agapiou A, Leventis G (eds) CAA2021 Digital Crossroads. Orellana Figueroa JD, Reeves JS, McPherron SP, Tennie C (in press)
3.

Nr.	Accepted publication yes/no	List of authors	Position of candidate in list of authors	Scientific ideas by the candidate (%)	Data generation by the candidate (%)	Analysis and Interpretation by the candidate (%)	Paper writing done by the candidate (%)
1	Yes	See above	1.	65	60	90	80
2	Yes	See above	1.	65	75	90	85
3							

I confirm that the above-stated is correct.

11.09.2024

Date, Signature of the candidate

I/We certify that the above-stated is correct.

11 9 2024

Date, Signature of the doctoral committee or at least of one of the supervisors

Objectives and Expected Outcome of the Doctoral Research

The main objective for this doctoral research was to explore the viability of the development of a computer program analogue to lithic replication experiments, able to perform lithic reduction—as well as measure, analyse, and digitally store the resulting products—in a manner that is easily-shareable, reproducible, and faster than the real-life alternative, as well as validated against real-life lithics: the Virtual Knapper. If the development of such a program were to indeed prove viable, this doctoral project also sought to create a working proof of concept for the aforementioned Virtual Knapper program.

As there were likely several possible methods for the development of the Virtual Knapper, and as the viability of developing such a program with the requirements outlined above would have depended on the approach used, the various available methods needed to be carefully evaluated, and a suitable candidate selected for further development (and testing) of a proof of concept.

If the Virtual Knapper as envisioned proved viable, an additional objective was to explore alternative applications of the selected model for lithic research, including the creation of additional proofs of concept for the possible applications.

In summary, at the end of this doctoral project I expected to present a publicly available proof of concept for an accurate and valid virtual knapper framework, one that would be capable of serving as a valid analogue for real-life lithic replication experiments. In addition, if any additional application of the underlying framework or model for the Virtual Knapper was explored, a proof of concept for this application, along with the results obtained from it, would be presented as well.

1. Introduction

1.1. The Earliest Stone Tools.

Prehistoric stone tools (or *lithics*, or *lithic technology*) are the earliest form of hominin technology—and material culture—ever found in the archaeological record (Toth 1985, p. 101; Schick and Toth 1994, p. 78; Schick and Toth 2006, p. 1; de la Torre 2011, p. 1028).

Evidence of stone tools extends at least 2.6 million years in the past in East Africa (Semaw et al. 1997; Semaw et al. 2003; Braun et al. 2019; Plummer et al. 2023). In addition, there is some additional contested (Domínguez-Rodrigo and Alcalá 2016; Archer et al. 2020) evidence from the same region (the sites of Dikika and Lomekwi) dated to around 3.3 million years (Dikika: McPherron et al. 2010; Lomekwi: Harmand et al. 2015). However, there is general agreement that stone toolmaking was already part of the behavioural repertoire of some hominin populations by at least 2.6 million years (de la Torre 2011, p. 1033). For this reason, and also for the limited number of tools available for analysis, I will not include the evidence from Lomekwi and Dikika when speaking of the *earliest stone tools*.

Lithic technology, for its antiquity, its tendency to be preserved in the ground for millennia (being made mostly of inorganic materials), its commonality across so many hominin populations and its pervasiveness throughout most of hominin prehistory, history[1], and even down to the present day (Stout 2002; Sillitoe and Hardy 2003; Arthur 2010; Hayden 2015), make it one of the most important sources of evidence for studying hominin evolution.

The most common manufacture process of most of the earliest stone tools is called ‘flaking’, or ‘flake removal’, or also ‘knapping’ and ‘flintknapping’. I will henceforth use these terms interchangeably. The raw materials used in the process were volcanic rocks, such as flint, obsidian, chert, basalt, or quartzite (see e.g. Plummer 2004, pp. 120–121; Mussi et al. 2023). These rocks fracture in a special manner when

[1] Notably likely mentioned in the oldest complete work of historical writing that survives to this day: *’The Aethiopians … carried bows made of palm frond stems, … and short reed arrows, but instead of being of iron,* they are tipped with a sharp stone for carving seals*; …’* (Herodotus 1920, bk. VII, chap. 69, sec. 1; translation and emphasis mine).

dexterously struck with another hard object (for the earliest stone tools, normally a hammerstone; see de la Torre and Mora 2014, pp. 783–784), allowing for sharp pieces to detach from them, called 'flakes'. The toolmaker (or 'knapper') takes a stone made from one of these materials (the 'core'[2]), finds in it an appropriate surface near a sharp edge, which produces the necessary angles for the fracture to develop properly (the 'platform'), and proceeds to strike it to obtain the desired flake removal (Debénath and Dibble 1994, p. 10; Whittaker 1994, p. 14).

The earliest stone tools were primarily manufactured by repeatedly knapping a core (or flake), this process being known as 'core reduction' or 'lithic reduction' (Debénath and Dibble 1994, p. 10). The details of this process can vary across the large temporal range of stone toolmaking across hominin evolution; from different materials for the hammer (stones, antlers, wood), to vastly different reduction sequences. Thus, archaeologists broadly categorise lithics depending on the specifics of their manufacture. 'Oldowan' stone tools refer to earliest attested (and non-contested) lithic technology, with finds ranging between around 2.6 to 1.2 million years (Semaw et al. 2003; Semaw et al. 2020; Plummer et al. 2023). These Oldowan tools are mostly characterised by the use of water-worn pebbles, cobbles, or lumps of stone as cores, and simple(r) core reduction sequences, sometimes referred to as a flake-and-core technology (Braun and Hovers 2009, pp. 1–14; de la Torre 2011, pp. 1029–1030; cf. Kuman 2014, p. 5561). Following the Oldowan and first appearing at around 1.7 million years, are 'Acheulean' tools, which are characterised by what are known as 'Large Cutting Tools' or LCTs: forms such as hand axes (bifaces), cleavers, and picks (McNabb et al. 2004, p. 653). Nonetheless, Oldowan-like tools appear contemporaneously with Acheulean tools, even in the same contexts, and remained under production as the main type of tools in many parts of the world long after Acheulean tools became common in Africa (de la Torre et al. 2003; Moore and Braun 2009; Moyano et al. 2011; Lee 2013).

Because of this large temporal range of certain stone tool forms, some archaeologists prefer to use a different system of classification to speak of stone tools, which uses numbered 'Modes' rather than named *industries* (e.g. 'Oldowan'), with Mode 1 equivalent with Oldowan, and Mode 2 with Acheulean (Clark 1969, pp. 29–

[2] Sometimes these are also called 'blanks', though the latter generally refers to a shaped or detached piece of stone suited to be reduced further (Debénath and Dibble 1994, p. 10; Whittaker 1994, p. 20).

31). This allows for stone tools to be more easily grouped across time and space according to their manufacture techniques, rather than geography or dating. For this work, I shall focus on the earliest stone tools in the archaeological record, thus my focus will lie with Oldowan tools prior to the transition to Acheulean in Africa (see above), allowing me to explore several questions regarding the earliest stages of hominin evolution for which we have material culture.

1.1.1. Questions in Oldowan Research

There are several unanswered questions about various aspects of the Oldowan. I will discuss them briefly in this section, delving deeper into those relevant to the present work.

There are continuous debates on whether Oldowan hominins were opportunistic scavengers, active predators, or anything in between (see Schick and Toth 2006, p. 21; Kuman 2014, pp. 5564–5565), whether the Oldowan was static (e.g. De Oliveira et al. 2019, p. 2) or variable, and whether there might be a divide at around 2 Ma between an earlier Oldowan (i.e. *pre-Oldowan*) and a younger or *developed* Oldowan (de la Torre et al. 2003; Schick and Toth 2006, p. 17; Kuman 2014, pp. 5567–5568).

In addition, researchers remain unsure of which species of hominin manufactured the Oldowan stone tools. When Louis Leakey discovered the skull of a *Paranthropus/Australopithecus boisei* in close association with Oldowan tools at FLK *Zinj* in Oldupai, Leakey assigned it as the maker of the Oldowan tools there, and this was in turn used as evidence of their ancestry to modern humans. Yet a later find of a specimen that was more morphologically similar to modern humans than *P. boisei* in that same layer (as well as other sites) was classified by Leakey and colleagues as *Homo habilis* (‘skilful man’), and suddenly the latter was made the toolmaker, and *P. boisei* was made *H. habilis*’ victim (Leakey et al. 1964, p. 9; de la Torre 2011, pp. 1030–1031). However, with the large temporal range of Oldowan tools—more than 1 million years in East Africa (see above)—the number of hominin species present in the continent in that same timeframe (and therefore, of possible toolmaking species) is larger than the two mentioned above (*P. boisei* and *H. habilis*), and spans three currently recognised genera: *Homo*, *Australopithecus*, and *Paranthropus* (cf. Schick and Toth 2006, p. 18). Moreover, recently found Paranthropus fossils in close association with plant and animal processing, along with Oldowan assemblages, all

tentatively dated to around 2.9 million years (though with an ample margin of error), suggest that stone toolmaking behaviour was not limited to a single genus (Plummer et al. 2023). Evidence (in the form of a mandible) for a *Homo sp.* dated to around 2.8 Ma (i.e. pre-earliest Oldowan, Spoor et al. 2007), make the question of Oldowan toolmakers all the more complicated.

One considerable question facing the study of Oldowan lithics (let alone that of later periods) is that we do not know what variability in the morphology of the artefacts found is borne from stochastic or ecological variables such as the size of initial cores used, the availability of raw materials (and also their reduction intensity), functional requirements, or even knapper skill (but see Proffitt et al. 2022), nor how strong this influence was in shaping the artefacts we observe in the archaeological record (Toth 1985; Schick and Toth 2006, pp. 27–28, 30–31; Braun et al. 2008, 2009). If we cannot discern whether a different lithic reduction technique (or perhaps attempts at controlling variables such as angle of blow: Li et al. 2023b) was applied because of the necessity for more constrained raw material availability or because of a knapper's skill, determining the truth between claims of a static and variable/evolving Oldowan is difficult (Gallotti 2018, pp. 26–27). Further complications relate to the archaeological assemblages themselves: landscape usage (e.g. Oldowan *home bases* and lithic transport), site formation processes (e.g. sedimentation rates), and time and space averaging—such as a single layer spanning (and compressing) several thousand years hiding the months or years between the manufacture of two artefacts by multiple individuals with varying skill levels and reusing the same raw material—all contribute to the difficulty we have in uncovering all the relevant behaviours represented by lithic artefacts (Schick and Toth 2006, pp. 27–28, 33–34; Perreault 2019).

Different combination of parameters could result in indistinguishable lithic assemblages or stone tool forms, i.e. they suffer from equifinality (see Hiscock 2004). For instance, Toth (1985) suggested that certain Oldowan core forms *'can often grade into one another during reduction'* (Toth 1985, p. 107), whilst Schick and Toth (2006) suggested that *'the entire range of Oldowan forms can be produced by hard-hammer percussion, flaking against a stationary anvil, bipolar technique (placing a core on an hard object serving as an anvil and striking it with a hammer), and occasionally, throwing one rock against another'* (Schick and Toth 2006, p. 4).

Moreover, although the famous typology by Mary Leakey divides Oldowan tools into types, and some of these types have names that imply a function (e.g. 'choppers',

‘scrapers’, ‘awls’: Leakey 1971, pp. 4–8), a full list of uses for these tools cannot be made, even as cut and percussion marks in animal remains are often discussed in relation to hominin carnivory (including scavenging), and there is some additional indication of meat and plant processing activity from microwear analysis, the later perhaps for the manufacture of organic tools (Schick and Toth 2006, pp. 18–19; cf. Plummer et al. 2023). When it comes to tools made of organic materials, it is important to note that the evidence we can uncover through archaeological means is heavily biased towards non-organic and non-perishable components, such as stones and mineralised organic remains (e.g. fossilised bones). This renders extremely unlikely, due to the decay of organic material, the uncovering of much—if any—evidence of early prehistoric non-lithic technology.

Looking to other species of extant primates as a comparison point, we observe that many of them use stone tools in the wild, e.g. when digging tubers out of the ground, or cracking open nuts and shellfish, and that stone tool use is a permanent part of the behavioural repertoire of several primate populations (Malaivijitnond et al. 2007; McGrew 2010; Proffitt et al. 2016). Moreover, after human training, one Orangutan (*Pongo sp.*) named ‘Abang’ (Wright 2009) and at least four captive Bonobos (*Pan paniscus*)—‘Kanzi’, her half-sister ‘Panbanisha’, and the latter’s two sons, ‘Nyota’ and ‘Nathan’—were able to knap stone and obtain sharp flakes (Toth et al. 1993; Schick et al. 1999; Schick et al. 2006), sometimes using them as cutting tools (Schick et al. 1999; Schick et al. 2006) . Other species have also been observed—both in the wild and in captivity—detaching flakes from stone, and some of them using those flakes for cutting (in the wild: Proffitt et al. 2016; in captivity: Westergaard and Suomi 1995; see also Mercader et al. 2002; but see Schick and Toth 2006, pp. 24–25), with one study on a wild population of Macaques (*Macaca fascicularis*) exhibiting flaking behaviour during food processing, and producing an assemblage that Proffitt et al. (2023) suggest *‘would likely be identified as anthropogenic in origin’* (Proffitt et al. 2023).

However, archaeological evidence shows that Oldowan knappers produced and discarded an astounding number of flakes, cores, and hammers, that they were technically capable to exploit raw materials to controllably produce conchoidally fractured flakes with sharp edges, and intensively reduce cores; behaviours that are not reflected in any other extant primates (Schick and Toth 2006, pp. 24–26). The source of this difference between the toolmaking behaviours of early hominins and

those of extant primates is, on the other hand, a source of contention. Even the question of whether the Oldowan is indeed *special* (i.e. whether this difference represents a leap in the evolutionary history of the hominin clade or only a small step from what is currently known about our extant phylogenetic relatives) is not yet settled (e.g. Toth and Schick 2009; Wynn et al. 2011; Snyder and Tennie 2022).

As discussed above, environmental variables, such as raw material availability, core reduction intensity, tool functional requirements, landscape use, and time averaging in the archaeological record are important parameters that affect the assemblages that researchers discover during excavation, as well as the Oldowan tool forms themselves. Therefore, explanations grounded in environmental, biological, cognitive, and socio-cultural factors can all be provided as answers for the questions that lithic technology poses to the study of human evolution, both in and of themselves or as combinations of factors.

Biological causes have to consider the still unanswered (at least conclusively; see above) question of which species were responsible for Oldowan (or possibly even earlier) toolmaking, since many of the candidate species had different morphologies, and this might affect their technical ability to produce tools (e.g. arguments on lack of bipedalism, no precision grasping capability, smaller brain size, or differences in brain structures; cf. de la Torre 2011, pp. 1030–1032), if not their cognitive abilities to do so, which could allow some candidate species to be excised from that very list (cf. de la Torre 2011, pp. 1030–1032). Thus, researchers can explore the biological and morphological traits of different hominin fossils in relation to their capacity for stone toolmaking (cf. Susman 1988; de la Torre and Mora 2014; Kivell 2015; Kunze et al. 2022). However, many of these biological factors are interlinked with others such as cognitive and socio-cultural ones (Schick and Toth 2006, pp. 21–23), and as is the case with gene-culture coevolutionary theory (Feldman and Laland 1996), socio-cultural factors such as an increase in group size or a higher reliance in stone tools can, in turn, affect the biological evolution of a species (Laland 2008).

More recently, some archaeologists have proposed hypotheses that highlight the importance of certain putative cognitive developments in early hominins, such as those based on functional brain lateralisation (Stout et al. 2008), working memory (Wynn and Coolidge 2007; Haidle 2010; Wynn and Coolidge 2011; Stout et al. 2015), or goal abstraction and mental templates (Stout 2011).

Socio-cultural factors, on the other hand, have been associated with Oldowan toolmaking from its inception as a culture-historical grouping (Leakey 1971; de la Torre 2011; cf. Gallotti 2018, p. 13). After the advent of processualist archaeology, such culture-historical frameworks were slowly eschewed, and a more functional and technological framework followed (de la Torre 2011, p. 1032). Cultural and social factors, still nevertheless appear in different theories of hominin evolution and toolmaking as explanatory variables, such as the Social Brain Hypothesis (Gamble et al. 2011), as well as hypotheses advocating an early development, exaptation, or co-evolution of toolmaking and modern human-like culture (Toth and Schick 2009; Stout 2011), socially-transmitted artefact forms (Shipton and Nielsen 2015; Tennie et al. 2017, p. 663), teaching (Morgan et al. 2015), and language (Stout et al. 2008; Morgan et al. 2015).

Some researchers divide culture into several types, some of which seem to not be present in extant non-human apes, and consider Oldowan technology to not be radically more complex from the technological and cognitive capabilities of modern apes, or at least, not as radically complex as that of modern humans (Wynn and McGrew 1989; Wynn et al. 2011; Tennie et al. 2017, and partially the comment by de la Torre therein; Snyder and Tennie 2022; Tennie 2023; Paige and Perreault 2024). We must examine these putative socio-cultural drivers more carefully if any inferences of past hominin society, culture, and cognition from the study of lithic technology are to be properly considered.

1.1.2. Culture in the Oldowan

For cognitive and socio-cultural hypotheses, archaeologists often perform experiments that involve the emulation and replication of archaeological artefacts by modern knappers (cf. Eren et al. 2016), sometimes with the addition of neuroscientific analyses—such as functional imaging (e.g. PET, fMRI, fNRIS)—for the study of cognitive requirements for toolmaking (Stout and Chaminade 2007; Stout et al. 2008; Stout et al. 2011; Stout and Khreisheh 2015; Putt et al. 2017; Putt et al. 2019).

A different approach that researchers can use is the *chaîne opératoire* (French for *chain of operations*) or *reduction sequence* methodology. This technological approach involves, in the case of lithic production, a division of the process involved in toolmaking into a sequence of steps—a *chaîne opératoire*—often implying mental operations as

well as physical ones. Depending on the focus of the *chaîne opératoire*, the process of production can be understood to encompass everything from the acquisition of raw material to the discard of used flakes, to the technical aspects of the flake removal sequence of only a single core (Sellet 1993; Bar-Yosef and Van Peer 2009, pp. 104–105; see also Haidle 2010).

This methodology is a top-down down approach, as it presupposes that the manufacturers of these stone tools intentionally sought to arrive to the form of a specific end-product or deliberately followed the specific process as reconstructed. Although perhaps evident, it must be highlighted that *chaîne opératoire* analyses are by necessity dependent on the perception and conjecture regarding the goals of prehistoric knappers as envisioned by each individual researcher; thus, the methodology is highly subjective (Ruck et al. 2020; cf. studies on observer error in lithics by Pargeter et al. 2023; but see Odell and Odell-Vereecken 1980; Timbrell et al. 2022). It also presupposes that researchers are able to reconstruct the goals and operations of early hominins during stone tool manufacture, and if so, that the reconstructed goals, choices, and operations match those of prehistoric toolmakers (Proffitt and De La Torre 2014; cf. to some extent Eren and Meltzer 2024). The first supposition is *a priori* valid, whilst the latter is most likely epistemologically invalid (Bar-Yosef and Van Peer 2009, p. 105; Tostevin 2011, pp. 355–357).

The assumption of intentionality in stone tool forms—not exclusive to *chaîne opératoire* approaches—has been subject to scrutiny (Hiscock 2004; in an Acheulean context: Corbey et al. 2016). The previously mentioned evidence on the influence of ecological factors on final stone tool form and lithic assemblages, not to mention the problem of equifinality, throw an additional wrench into the problems with ‘desired end products’. Nevertheless, *chaîne opératoire* analyses have been used in the past to argue, based on the complexity of the sequence of steps envisioned for Oldowan and later lithic production, for the presence of modern human-like culture (i.e. *cumulative* culture; see below) in the hominin lineage from at least the Oldowan onwards (Stout 2011).

Cumulative culture is a form of culture where innovations can be maintained and accumulated across generations through social (i.e. non-genetic) transmission. This accumulation of socially transmitted innovations can lead to behaviours that are so complex they cannot be re-invented by one individual during their lifetime (i.e. ‘culture-dependent traits’: Reindl et al. 2017); what Tennie et al. (2009) termed the ‘ratchet

effect'. Due to the social nature of the transmission of information, cumulative culture also normally leads to rates of change that can vastly outpace those of genetic evolution, in the scale of decades or even years, rather than (only) generations (Perreault 2012; Tennie et al. 2018).

Modern human culture is cumulative, and one example that illustrates this is that of a common office chair. A swivelling, height-adjustable, wheeled, reclining office chair is a complex union of various parts and materials—which requires the acquisition, processing, and manufacturing of the materials themselves, and the tools and equipment necessary to obtain these materials, and the acquisition, processing, and manufacturing for those tools and equipment, *ad nauseam*—that no individual could re-invent it without culturally acquiring the required knowledge to do so. Arguably, it might not be possible for any modern human to acquire all that knowledge in one lifetime. Without this cultural transmission, the best an individual might be able to do is grab a comfortable enough wooden log and sit on it. The ratchet would have, in effect, slipped back to baseline, though it could be re-engaged again if innovations could once again be discovered, transmitted, and maintained (Tennie et al. 2009, p. 2405). Thus, finding an archaeological artefact that is too complex to be individually reinvented could be evidence that cumulative culture already existed when the artefact was made.

Nevertheless, there is an additional requirement beyond complexity alone, and that is that the behaviour (e.g. toolmaking) must be socially transmitted. One individual inventing a highly complex greeting consisting of several dozens of steps is not cumulative culture until that handshake is socially transmitted to other individuals, even though it would be highly complex and virtually impossible to re-innovate by another individual. In modern human cumulative culture, the main mechanism classes of social transmission that make cumulation possible are thought to be process copying, teaching, and prosociality (Tennie et al. 2009, 2018, 2020; Andersson and Tennie 2023). In the case of process copying, as well as specific types (i.e. copying variants; Tennie et al. 2020) of teaching, Andersson and Tennie (2023) have further specified that the ability to copy supraindividual *know-how* information across domains is the root of modern human cumulative culture, though presumably the need for special types of prosociality remains (Andersson and Tennie 2023).

As the name implies, *know-how* copying refers to the social acquisition of information on the *how* of a process (including technical processes) or behaviour, as opposed to e.g. the *what*, the *where*, and *why* (Tennie et al. 2017, 2020; Andersson

and Tennie 2023). Without the ability to copy *know-how* (which could happen via teaching of *know-how*), or with only a rudimentary ability to do so—perhaps limited to specific 'domains of variation' (Andersson and Tennie 2023, p. 4)—*know-how* information must be re-invented each time the process is undertaken, in a similar vein as to how one would have to assemble a chair lacking the instruction manual. If a species' ability to copy *know-how* was indeed limited, *know-how* information could not be accumulated culturally, and its complexity could not surpass the level of what individuals of that species could re-invent by themselves within their lifetime, what is known as the 'Zone of Latent Solutions' (cf. Tennie et al. 2009, 2020; Andersson and Tennie 2023). On the other hand, were the species' copying ability high enough, and the necessary scaffolding and motivation present (Andersson and Tennie 2023), the *know-how* information available in the behaviours of members of the species could increase culturally over time (i.e. it would be *cumulative*) to the point where this then culturally evolved *know-how* could no longer be innovated by an individual alone, leading to the type of culture visible in modern humans (Tennie and Call 2023; i.e. it would be *supraindividual*; no longer within the species' Zone of Latent solutions: Andersson and Tennie 2023).

Although humans are inarguably prolific copiers (e.g. Lyons et al. 2007; Clay et al. 2018), some researchers would assert that other great apes, such as chimpanzees, also can copy *know-how* in limited circumstances (Whiten et al. 2009). However, most empirical studies find that apes do not copy *know-how* (Tennie et al. 2006, 2012; Neadle et al. 2017; Motes-Rodrigo et al. 2022; Tennie and Call 2023; Andersson and Tennie 2023; Bandini and Tennie 2023; see also evidence for brain connectivity changes through human training of apes to make them copy: Pope et al. 2018). This is not to say that the Zone of Latent Solutions' account for non-human ape culture predicts an absence of all types of social learning; on the contrary, non-human apes have many social learning skills, and indeed it is these skills that are responsible for their cultures (Acerbi et al. 2022). While no study has yet conclusively confirmed the ability of non-human great apes to copy *know-how*, they do have the ability to *re-innovate* it (Bandini et al. 2021; Motes-Rodrigo et al. 2022; 2023), though if and only if it resides within their Zone of Latent Solutions (Tennie et al. 2009, 2017; Andersson and Tennie 2023). In this way, apes can be *triggered* (Acerbi et al. 2022) to show (re-innovated) *know-how* inside their Zone of Latent Solutions, as well as copy other types of information, such as *know-what* and *know-where* (Tennie and Call 2023; Andersson

and Tennie 2023), leaving thus open the question of if non-human apes cannot copy *know-how*, then *when* in our evolutionary history did the ability to copy *know-how*—and to do so well enough to accumulate cultural *know-how*—first appear?

Arguments on the *when* culture begins to accumulate range from around the time of the Oldowan and earlier (e.g. Stout 2011; Morgan et al. 2015), to much after (Snyder et al. 2022; Andersson and Tennie 2023; Paige and Perreault 2024). Although Oldowan tools are not completely identical across sites nor across time, the variability discernible from their morphology and manufacture pales in comparison to even the variability visible in modern human culture (Paige and Perreault 2024). This much is admitted by studies such as Stout (2011), which consider the lack of variability in the Oldowan to be the tail end of an already present and soon-to-expand hominin cumulative culture (along with the cognitive requirements needed for cumulative culture in the first place).

An alternative hypothesis seeks to provide a baseline with which to test any claims of any putative cultural (and related cognitive) signals by testing whether a specific behaviour or artefact could not be independently re-innovated by an individual (of the species we think exhibited this behaviour); thus, whether know-how copying would be required to manufacture the artefact in question (Tennie et al. 2017). If the manufacture of a lithic artefact were to be replicated with only a simple set of rules (e.g. Moore and Perston 2016) or by modern humans entirely naive to the task required (e.g. Snyder et al. 2022), rather than a technique that would necessitate *know-how* copying, it would be a strong signal that social transmission is not required (though it could have happened regardless). As it happens, recent experiments have provided important insights into the possibility that certain stone tool *types* could be independently re-innovated by prehistoric hominins.

1.1.2.1. A Simple Set of Rules

As stated above, researchers have already remarked on how the range of *forms* of Oldowan lithics can be arrived at from other forms and simple techniques (Toth 1985; Schick and Toth 2006, p. 4). Even the more complex reduction techniques in de la Torre et al. (2003) amount to the following of a simple reduction strategy of a few steps. In an attempt to explore whether a simple set of rules could produce stone tool *forms* (such as handaxes and cleavers) that are commonly ascribed to great cognitive or cultural leaps, Moore and Perston (2016) performed an experiment where a core

would be reduced by repeatedly knapping a random platform while aiming to maximise the amount of mass removed (through the knapper's own skill in determining the desired point of percussion), finding that stochastically-produced proto-bifaces appeared (though rarely) at certain stages of the reduction sequence.

Related experiments, examining the use of language, as well as brain activity during knapping of Oldowan and Acheulean tools (Putt et al. 2014; Putt et al. 2017; Putt et al. 2019) suggest that Oldowan toolmaking does not require more human-like cognitive abilities (i.e. those present in modern humans but not in other apes) nor language. A subsequent experiment by Snyder et al. (2022) of completely naive modern humans with no instructions also found that cultural transmission of *know-how* was not necessary for flake removal techniques common in the Oldowan, nor for bifacial reduction (Snyder et al. 2022).

These experiments are important as a *null model* of lithic production (cf. paradigmatic thinking in Eren and Meltzer 2024) from which we can examine the relative complexity of the production requirements in terms of cognition, social learning, or practice required to create existing stone tool forms.

However, such large-scale lithic replication experiments require considerable amounts of time to undertake, from planning (including the ethics approval process for human participants), raw material procurement, knapping, measurement, and analysis of the lithic products, as well as labelling and storage. The cost of the raw materials, as well as that of the participant rewards (if used) and any other member of the experiment (e.g. an experienced knapper) is also not insignificant. Furthermore, the requirements for transporting and storing the raw materials (both before and after knapping) can also limit the feasibility of undertaking a lithic replication experiment. For example, the study by Moore and Perston (2016) took several months to complete (Moore 2021), as the total number of flakes and cores (inspected at each stage of reduction for each core) analysed was over a thousand pairs; thus, more than a thousand stages of reduction.

The significant requirements for performing large-scale replication experiments make studies such as Moore and Perston (2016) and Snyder et al. (2022) difficult to undertake, let alone replicate. I thus sought to find an alternative that could be an accurate and valid analogue for real-life knapping at a portion of the time and cost; a piece of software capable of virtually removing flakes from a core, embedded in a framework that would allow for repeated flake removals based on specified parameters

in a fraction of the time and cost of real-life large-scale lithic experimentation: a *Virtual Knapper*.

1.2. The Case for the Virtual Knapper

A battery of similar tests to those of Moore and Perston (2016) could not only verify the results obtained, but also examine the range of conditions (e.g. degree of core reduction, average size of flake removals, reduction method) in which forms such as bifaces are most likely to stochastically appear. Moreover, a Virtual Knapper could, through a large number of virtual lithic experiments, examine the relative abundance of a specific lithic *form* in an assemblage, and to which other forms they are most closely *related* (*see* Toth 1985). In addition, every computerised experiment would lead to the creation of entire artificial assemblages of stone tools; assemblages that could, in turn, be examined at every stage of reduction, even back to the initial removals, rather than only in the final stages as is the case by necessity for archaeological assemblages. The virtual assemblages could be used—or even tailor-made with a specific experimental setup—to compare with archaeological data, with possible applications including the validation of different core reduction intensity and cortical ratio calculation methods.

Real-life experimental repeatability can only be achieved by using standardised blank shapes, which could be considered not entirely ecologically valid depending on their manufacture, since hominins did not have access to perfectly standardised forms (but see the discussion on river-worn cobbles in Snyder et al. *forthcoming*); a limitation that a computer program would not face, as researchers could digitally duplicate any and all unique cores of an entire dataset. A knapper also invariably learns as they knap more (cf. Pargeter et al. 2019), possibly making the earlier removals not representative of the knapper's skill at the end of the experiment. The above advantages would lead to a much faster turnaround of lithic experiments, and with the variability of a human knapper removed, a Virtual Knapper would also be far more consistent, leading to higher internal validity in experiments; with additional randomisation of knapping parameters to simulate a lack of skill (cf. Pargeter et al. 2020) also a possible addition.

However, for the Virtual Knapper to successfully replicate the process of flake removal, it would need to process 3D objects, a much more complex task than working

in two dimensions only, as it would need to predict a volume, rather than an area (or an outline). A program that worked *only* in two-dimensional space would be severely constrained, as the program would only be able to provide a flake outline (or an outline of a flake scar) on the core, rather than any information on the volumetric shape of the removal. In addition, the Virtual Knapper would need to be able to knap based on the 3D objects of any core, and use these 3D objects as inputs, regardless of whether these objects were 3D scans of real-life cores or made entirely within a computer environment using 3D modelling tools. The program would also need to include the information of the parameters of the flake removal, such as point of percussion, platform depth, platform surface interior angle, angle of blow, and hammer hardness (Li et al. 2023a). The Virtual Knapper could only have as inputs the 3D object and the flaking parameters and use these only to generate its predictions of the 3D objects of both the flakes removed and the modified core surfaces (which could be calculated by simply subtracting the predicted flake from the modified core).

In addition to flake prediction, a Virtual Knapper should also ideally automatically perform measurements and analyses of the lithics produced, with flake length, width, volume, thickness, area, and modified core volume as the simplest features to measure and calculate for the initial stages of the development of the program. In addition, knowing the raw material used for the virtual knapping—more specifically, its mean density—it is possible to calculate the mass of the lithic products by using the volume. With these few simple measurements, researchers would already have enough to perform some lithic analyses; e.g. percentage of core mass removed per flake removal, mass of flakes at each stage of reduction, or flake length compared to flake width. Other measurements and parameters, such as the measurement of cortex, or the number of flake scars, could be added in later stages of development.

The last important requirement for the Virtual Knapper is that the program would need to virtually knap a core, predict the resulting flake and modified core, measure and analyse the products, and it would need to perform this faster than the real-life equivalent of this process, as well as at high levels of validity and reliability. A useful target would be to say that the computer would need to perform the entire process of knapping, measuring, and analysing a single flake removal in a manner of seconds or minutes. Moreover, the duration of the process needs to be assessed for a computer with mid-range commercial specifications, rather than a high-performance workstation, or a distributed computing centre, since the goal is to make lithic research easy *and*

accessible to all researchers, not only those in possession of systems with high computing power, regardless of ever-increasing technological development. The diminished time and cost would thus allow for the exploration of broader questions of stone tool manufacture that would otherwise need to be examined with severely costly large-scale lithic replication experiments. Not to mention, of course, the possibility of using the Virtual Knapper as an outreach and teaching tool for prehistoric tool-making knowledge.

However, when developing the basic Virtual Knapper, the initial focus should be on reliably simulating hard-hammer percussion knapping first, since the technique is comparatively simple (i.e. only a single point of contact, rather than two points with an anvil), and is also characteristic of Oldowan tools (Schick and Toth 2006, p. 26).

Even if the virtual knapping were only *as fast* as the real-life equivalent (which usually takes weeks if not months), it would still be considerable advantage, as (as mentioned previously) computers can be made to run an experiment continuously without interruption, and with multiple computers (or even multiple instances of the program running), there would still be a considerable improvement in speed for lithic experiments. However, even an arbitrary example goal of less than one minute for the program to complete both the knapping and the analysis of a single flake removal (see above) would allow for large-scale lithic experimentation to be orders of magnitude faster than real-life; from months for a single experiment (Moore 2021), to less than a week (10,080 minutes in a week). With a faster system that only needed fractions of a second to complete the process, several dozen large-scale replication experiments could be completed in a single eight-hour workday, using only one machine and only one instance of the program: more than 28,800 flake removals.

Although the concept of a high-performance virtual knapping program is highly promising, and the possibilities allowed are ample, the development of such a piece of software had many important challenges that needed to be overcome first. The most challenging obstacle is that 3D object data is more difficult to use, process, and modify, than 2D data. Furthermore, in most cases, digital 3D objects (including those digitised by a 3D scanner) consist of and describe only the outer surface of the object in question: the hull. The internal structure is not considered, as it is not necessary when displaying the object using computer graphics, the main purpose of most digital 3D objects. The inclusion of points describing the internal volume of an object makes displaying the object more resource-intensive (if only due to the larger amount of data

to process) with little-to-no benefit, as only the outer surface will be visible (displaying transparent and translucent objects does not necessarily require internal structure either). However, 3D objects with internal structure (volume meshes) do exist and are used for physics simulation such as Finite Element Analysis (see below), but the most common 3D scanning equipment (excluding CT scanners) are *surface scanners*, which do not capture the internal structure of objects, so the addition of internal volumetric data would be an additional task to accomplish. Without an internal structure for the core objects, however, the drawing of a flake's (or flake scar's) boundary in 3D space would require the *creation and positioning* of the points that would describe the flake volume, rather than a simple selection of the points that would form the flake scar.

There are numerous possible approaches that alleviate the issue of dealing with 3D volumes. An approach based on 3D geometric morphometrics (*3DGM*, an analysis of 3D object morphology using standardised landmarks, e.g. Shott and Trail 2010; Herzlinger et al. 2017) could be effective. One previous study has shown that 3DGM can predict the distal shape of a flake highly accurately using only the information of the platform shape (Archer et al. 2017). However, since the Virtual Knapper would not know *a priori* the shape of the platform of the detached flake (since this is part of what it must predict), there would be a need to develop a larger pipeline that worked in two stages: flake platform shape prediction, then overall flake shape prediction. The alternative would be to simply develop a new method to predict flake shape using the core shape, rather than follow the geometric morphometric avenue.

The use of 2D images (one or multiple of them) to describe the 3D shape of the lithic material through either the use of *voxels* (dividing the object into small cubes, in the same way that images are divided into small squares or *pixels*), a tomographic approach (slices across the volume), or a projection approach (e.g. images across all six orthographic projections), were another possibility explored (and the last eventually selected) for this work (see "Appendix 1": Orellana Figueroa et al. 2021; and "Appendix 2": Orellana Figueroa et al. *in press*).

Regardless of how the 3D data was processed, however, I first needed to define and describe the underlying technology for the Virtual Knapper framework, as this would constrain any other methodology used for processing the cores and knapping variables, as well as predict the resulting flakes. I discuss the three main avenues considered for this work in the next section.

1.3. A Predictive Model

1.3.1. Physics-Based Model

Conchoidal fracture is a type of fracture that can occur when brittle materials—such as those used for stone flaking (e.g. glass, obsidian, and flint)—break. In order to predict the resulting shape and volume of the flake(s) and core(s) of a flake removal in a three-dimensional volume of stone, the Virtual Knapper would (under this physics-based approach) require the use of methods such as finite element or boundary element analyses (henceforth *FEAs* and *BEAs*, respectively).

These types of analyses are, in most cases, very computationally costly. FEAs, for example, would require the construction of *meshes* by way of dividing the desired object into a set of discrete points, both on the surface and the insides of the volume of stone to simulate knapping on. In essence, this would be equivalent to re-creating a large piece of stone with building blocks and having every individual corner of each brick acting as one of the points in this building *mesh* of blocks (when multiple corners touch together, this would be only a single point). The model would then need to solve partial differential equations for every individual point in the volume. This is one of the main factors that make FEAs highly resource intensive, as even a simple mesh with 32 points per spatial dimension (32 × 32 × 32) would have a total of 32,768 points that would need to be calculated. By comparison, the simplified core meshes used in the proof of concept ("Appendix 1": Orellana Figueroa et al. 2021) have an estimated average of around ten thousand points for the surface data alone (i.e. without any points describing its internal structure). BEAs are, in many cases, more computationally efficient than FEAs, as they only require calculating over the *surface* (i.e. the boundary), rather than the entire *volume* (Hahn and Wojtan 2015, p. 151:3). Nevertheless, the computation of the points in either method must be repeated at each time step (e.g. every simulated microsecond) to be able to calculate the spread of forces and the resulting fracture (as this is how these methods operate), which adds another factor that must be considered when attempting to run a simulation.

In addition, fracture analyses with FEAs often require 'notches, pre-cracks or kerfs' when simulating (cf. Bilgen et al. 2018; e.g. Ni et al. 2018; and Hahn and Wojtan 2016), but this is not often possible to use when one wishes to simulate a flake removal, since one would need to know *a priori* of any flaws in the material for conchoidal fracture to

initiate (Cotterell and Kamminga 1987, p. 658). The use of a pre-crack in the simulation could also affect its accuracy, as the parameters of the initial fracture drive its own propagation through the material, being itself a new possible source of error.

The most relevant results—i.e. physics simulations of conchoidal fracture without pre-cracks—published at time of writing were those by Bilgen et al. (2018) and Kopaničáková and Krause (2020), but even these were run a cluster of 24 computer nodes, each with two CPUs with ten cores each (total of 480), as well as 64 GB of memory per node. Even in the performance metrics from Kopaničáková and Krause (2020), run on a single node with 20 cores and 64 GB of memory, the machine could only provide results after approximately more than three minutes[3]. Nevertheless, since one of the main goals for obtaining a computer model that simulates knapping was to be able to use it to undertake experiments similar to Moore and Perston (2016), this would require the use of at least one thousand separate flaking simulations (a total of slightly more than two days non-stop computation with the same setup as Kopaničáková and Krause 2020). Simply replicating the Moore and Perston (2016) experimental setup would not be enough, however, as the Virtual Knapper should also include the ability to modify the parameters—or rules—of core reduction used, and correctly predict what their effects on the final result would be; i.e. the *full* experiment would consist of multiple iterations of Moore and Perston (2016), each with a different set of rules applied during knapping. For this *full* experiment, one would then require the simulation of several sets of at least one thousand knapping events, each of which would take several days to complete, unless obtaining access to a computing cluster.

Although even the aforementioned calculated time is a large advantage over real-life experiments, there was an additional issue. The test performed by both Bilgen et al. (2018) and Kopaničáková and Krause (2020) was only based on an outwards displacement of a surface being exerted on the simulated object and in this way causing the fracture, rather than a collision (between the hammer and core). It is unclear whether the methods cited above could also model a collision scenario more pertinent to knapping, or if they can do so without large changes to the code that underlies the physics simulation.

[3] Assuming the 7.9M degrees of freedom for the conchoidal test are equivalent to the performance chart in Fig. 17a of (Kopaničáková and Krause 2020).

Thus I needed to consider alternative approaches to physics simulations—hoping however that improvements in the near future would make these a pathway for the study of lithic reduction—and move on to the second possible path to a predictive 3D computer model of flaking: a model built on empirical data.

1.3.2. Empirically Grounded Model

Building an empirically grounded (i.e. based on data obtained from real-life experiments) model of stone flaking requires understanding how each of the different variables that play a role in every conceivable flake removal effect the parameters of the resulting flakes (e.g. how does angle of blow affect flake volume, mass, length, etc.). These variables are numerous (e.g. platform depth, angle of blow, hammer hardness, platform shape), and the resulting parameters even more so (e.g. flake length, flake mass, flake area, flake shape, bulb volume), but in order to build an empirical model, one would have to be able to understand how all of these interact with one another to construct the Virtual Knapper.

Moreover, since modern knapping replication experiments normally involve a human actor, precisely calculating and controlling for the different variables is difficult in a living model. An experienced knapper may be able to strike the core at a specific point with a high degree of precision (Pargeter et al. 2020), but even so, measuring variables such as the angle of blow (the angle between the applied force and the surface of the platform), the applied force, or the hammer velocity, would require additional steps in data acquisition and measurements (e.g. the use of high-quality video footage, which requires appropriate set-up and large computing storage capacity for the resulting hours of video). In addition, maintaining variables constant whilst examining the effects of a single independent experimental variable is impossible even for the most expert of knappers, and human error in a strike cannot be corrected. If one wished to examine the effect of only angle of blow in flake formation, for instance, the knapper would have to strike the surface of the stone at the precise location, with the same velocity, and applying the same amount of force *every single time*. Any slight deviation would introduce error in the end-result, and the effect of the single variable to measure (angle of blow) would not be as easily ascertained due to the possible interacting effects of the other variables that the knapper failed to (or rather *could not*) control.

Some knapping experiments have been able to avoid some of these difficulties by using programmable knapping machines that can perform repeated and consistent flake removals based on programmable parameters (reviewed in Li et al. 2023a). Machine-based controlled flaking experiments allow researchers to examine and measure the effects of the studied variables, as well as better control each individual parameter with more precision than could be possible with any human knapper (e.g. Magnani et al. 2014). Machine knapping experiments began in the 1980s with the plate glass (i.e. a thick sheet of glass) flaking experiment by Dibble and Whittaker (1981).

Before delving into detail on these machine knapping experiments, however, I must first mention one additional knapping variable that has thus far gone unnoticed, but one which is crucial to control as it could strongly influence the results of any flake removal: the morphology of the knapped core. Brittle stones commonly used as knapping blanks are never identical to one another in nature, and only with modern tools is shaping them into standardised forms at all possible (Snyder et al. *forthcoming*). As such, experimenters wishing to control for core shape are faced with a trade-off between what is archaeologically and naturally (i.e. externally) valid, and a more rigorous experimental setup (i.e. internally valid; cf. Eren et al. 2016).

For the first machine flaking experiment by Dibble and Whittaker (1981), controlling the core shape was accomplished by using plate glass, with its thin edges serving as the platform for knapping. This, however, severely constrained the platform width to the width of the plate glass itself (i.e. a few millimetres), a setting which is not applicable to most archaeological finds (Dibble 1997, p. 151), and especially not to Oldowan stone tools. Later on, however, Dibble and Režek (2009) moved from plate glass to casted glass core shapes that more closely resemble archaeological material, and from which different standardised core shapes can be made (Režek et al. 2011; Lin et al. 2013; Magnani et al. 2014; Leader et al. 2017; Lin et al. 2018, 2022; Dogandžić et al. 2020; McPherron et al. 2020; Li et al. 2023a; see also Fig. 1). The results of a recent study comparing the results of the machine knapping experimental setup on different materials suggest that glass itself is a valid material to evaluate knapping performance, showing that only the force needed to detach a flake is different between materials (Dogandžić et al. 2020).

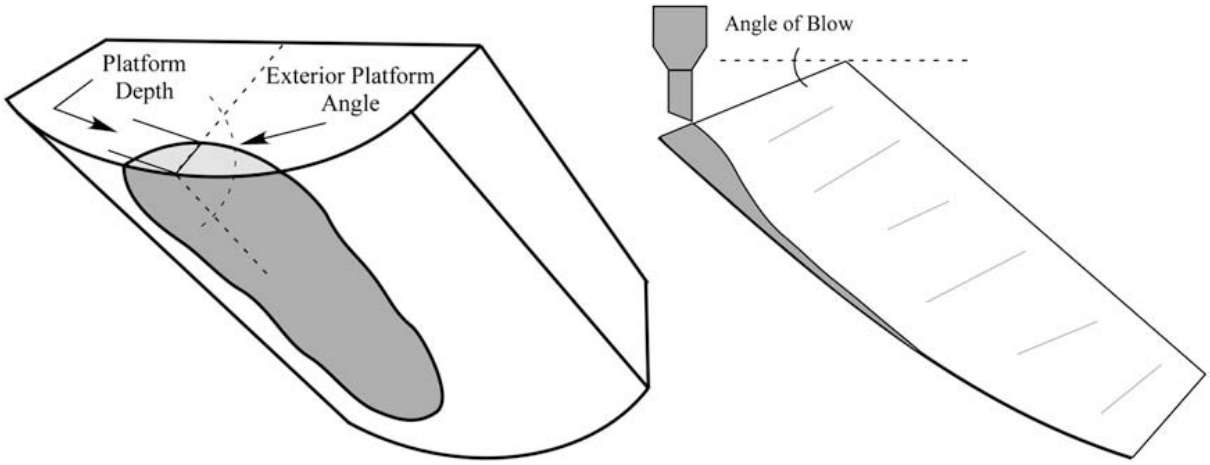

Figure 1: Standardised casted glass core from Dibble and Režek (2009). Used with permission.

With the ability to control for all these parameters, experimenters now had the ability to examine the effects of individual variables—as well as interactions between one and more variables—with respect to flake shape, all whilst controlling for most others. The two main groups of interactions that researchers explored in machine-assisted flaking were related to the influence of the platform surface, as well as the influence of the parameters of the hammer strike.

The variables used to examine the influence of the platform surface included the platform depth (PD), which the distance from the edge of the platform to the point of percussion (POP; i.e. the point of impact of the hammer), as well as the exterior platform angle (EPA), which is the angle between the surface of the platform and the surface from which the flake detaches, and most recently, the platform surface interior angle (PSIA), the angle that forms between the point of percussion and the edges of the platform (Dibble and Režek 2009; McPherron et al. 2020; Li et al. 2023a). The variables used to examine the influence of the percussive blow or hammer strike include the applied force from the hammer, the angle of blow (AOB), the hardness of the hammer (e.g. soft wood vs hard steel) and its velocity (Li et al. 2023a). These variables are perhaps many of the most relevant and most easily controlled by a knapper when striking a core.

Through controlled experiments such as those performed by Dibble and Režek (2009), researchers were able to uncover the effects the control variables have in flake formation by measuring the parameters of the resulting flake they were interested in (e.g. width, length, mass). Performing additional experiments with additional variables

such as hammer hardness, the results from Magnani et al. (2014) showed how different variables could have equivalent effects on the same parameters in the resulting flakes. For instance, a steeper angle of blow, smaller platform depth, and a softer hammer all increased the value of the width divided by platform depth (Magnani et al. 2014). In a word, many of these variables are equifinal (see above). Moreover, in order to capture the sum of the interactions between variables, a large number of parameters for the resulting flakes had to be measured (e.g. the width by thickness, the length by thickness, the area of the flake by thickness), as there are highly complex interactions during knapping, and any one controlled variable could affect many of these parameters in turn (see Fig. 2).

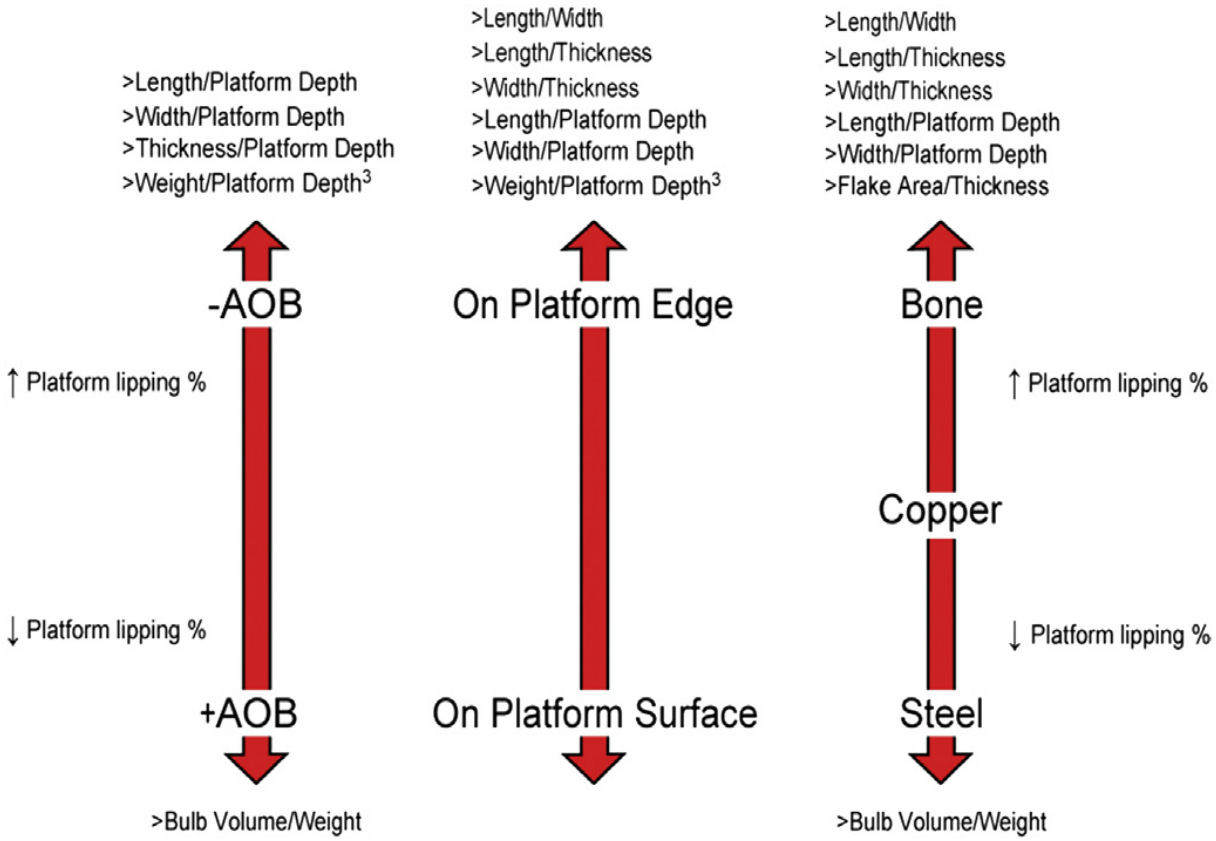

Figure 2: A diagram of the different interactions of flaking variables and parameters in the resulting flakes, from Magnani et al. (2014). Used with permission.

Although great leaps in understanding have been accomplished through machine-based controlled flaking experiments, not every variable and not every interaction has been measured yet. For instance, as stated earlier, with the ability to control the morphology of the cores used for the knapping experiments, researchers have

examined the effects of different extreme core surfaces (Režek et al. 2011). However, even though the results obtained by Režek et al. (2011) have shown the various effects core surface morphology has in the resulting flake shape, the wide range of possible variability in core shapes in the archaeological record (as well as the natural world) remains untested (i.e. a lack of external validity, especially when it comes to all possible core morphologies; cf. Eren et al. 2016). Moreover, considerable time is required to perform these machine flaking experiments on standardised core surfaces alone, suffering from the same time, cost, and storage constraints as any other lithic replication experiment.

When it comes to creating a Virtual Knapper model based on current empirical machine-flaking experimental data, the limited amount of information on the effects of core morphology rendered available the use of only to the core shapes that have been used in past experiments. If one wished to use other core morphologies, one would be forced to assume that these do not affect resulting flake shape differently than what has been observed in previous experiments (e.g. Režek et al. 2011), and that it would be possible to extrapolate from the experimental cores the effects that other—less regular—core surfaces have.

In summary, if one wished to build a predictive 3D computer model (i.e. a *Virtual Knapper*) for stone flaking based on empirical findings from controlled experiments, there existed already an advantage in that researchers have uncovered much data on the effects of different knapping variables on flake shape, but for some variables (e.g. influence of different core surfaces) there is no data yet. Nevertheless, there still could have been several knapping variables overlooked, or ones whose existence was unknown, or while known, not known to affect flake formation (e.g. the recent introduction of platform surface interior angle as an important variable in flake formation: McPherron et al. 2020).

Due to the small amount of available data, which restricted the knowledge of interacting parameters only to a small set of very specific core shapes, it would have to be a formidable challenge to proceed with an empirically based predictive 3D computer model of stone flaking, as the success of such an approach to other core shapes was uncertain. Nonetheless, I must also admit that a different view is valid as well. Many researchers working with machine-based knapping experiments have already provided data on the variables that drive flake formation, as well as their effects on the measured properties of resulting flakes. I could, therefore, have attempted to

explore the possible results I could obtain with the available data in terms of building a predictive model of stone flaking, which could perhaps also have paved the way for future research in the process (i.e. when even more data is available).

Nevertheless, there was another, somewhat related approach that I decided to explore first, one that could build a model able to look at all the necessary variables at once, identify their interactions—as well as the strength of these interactions—and take all of these into account to accurately simulate a flake removal: machine learning.

1.3.3. Machine Learning Models

Machine learning is a term that spans a diverse family of algorithms used to formulate mathematical models of a set of data to be studied (Hastie et al. 2009, xi-xii; Fernandes de Mello and Antonelli Ponti 2018, 1–5). Machine learning is a branch of artificial intelligence and computer science, but its methods have found success in many fields where variables are numerous and interactions are complex, such as facial recognition (e.g. Lawrence et al. 1997), image classification (e.g. Ciresan et al. 2011), object (e.g. Barik and Mondal 2010) and species identification (e.g. Kumar et al. 2012; Mäder et al. 2021; Carvalho et al. 2022), and many more (e.g. Hastie et al. 2009, pp. 1–6).

These examples in which machine learning has been successfully applied parallel the problem of multitudinous factors with uncertain effects present in developing the Virtual Knapper. For instance, one could be able to avoid the issue of not considering all the variables one needed to consider when building the base of the Virtual Knapper, as many machine learning algorithms are formulated and have the specific goal of predicting outcomes from variable interactions unknown to modellers, e.g. in recognising the variables that best predict certain specific medical conditions even whilst these are unclear or unknown to humans (Wang et al. 2014; e.g. van Ginneken et al. 2015). Moreover, since the (immediate) goal—the creation of a computer model of knapping—was prediction, and less so the understanding of the variables and interactions involved in knapping (this latter is *a priori* necessary for the empirical approach), machine learning was likely an ideal avenue for accelerating results in the prediction aspect without requiring a complete (or even partial) understanding. This is because machine learning algorithms are, by definition, able to learn how to model a set of data through *the study of the data itself* (also known as *training*), rather than

through the manual (i.e. human) programming of the parameters and interactions of the desired model, as is the case with physics simulations and empirically grounded models.

For this reason, I explored the application machine learning methods to the problem of obtaining predictive 3D computer model for stone flaking, as these approaches have proven successful in the past for similar problems.

However, machine learning methods have a perhaps idiosyncratic data requirement, not common to many other statistical methods. Since machine learning models learn to model data from the data itself, a large input (training) dataset must be available for the model to be able to generalise; i.e. for its predictions to be accurate for new data, rather than only data that is similar to the input data. I discuss the data requirements of the machine learning-based Virtual Knapping framework below.

1.3.3.1. Data Requirements

As a baseline, a machine learning Virtual Knapper would require a dataset of matching cores and flakes with which the model should be trained. This implies, more specifically, a dataset composed of many sets made up of one flake, and the core from which it was detached, and from which no newer flakes have been knapped (i.e. its surface was not subsequently modified). As a computer model would require digital data, these cores and flakes would need to be digitised. The simplest procedure for achieving this would be to use a 3D scanner to obtain digital meshes (3D models) of the real-world cores and flakes. The goal of the Virtual Knapper would be to predict the flake that would be removed given a certain core and a point of percussion (if not more variables). Therefore, the correct procedure to obtain these data would be to scan the core first, then knap it *only once*, and scan the resulting flake. However, although not necessarily required, having an additional scan for the modified (post-flaking) core could also prove helpful to more easily and more accurately locate the point of percussion on the unmodified core.

However, I must clarify that the cores do not necessarily need to be fully cortical, and the flakes do not necessarily need to be the initial removal. A single core's entire reduction sequence could be a valid—and indeed, desirable—input, if the matching core and flake models for each stage of the reduction were available. Even if only the core and flake scans from the initial and final removal of a long reduction sequence were available (an extreme example), they could be used as inputs if these stages

have available their matching unmodified core and removed flake models. In summary, *any* pair of unmodified core and subsequent removed flake is a valid input, regardless of the stage of core reduction.

One additional requirement, however, was that the models of the cores and flakes should be aligned so that they are oriented and positioned in the same way (see "Appendix 2": Orellana Figueroa et al. *in press*). This is to say that—ideally—the flake model should be refitted into the modified core model (assuming here it is available), and these two together, when superimposed, would have their surfaces line up exactly to form the intact core (see Fig. 3). Moreover, it could also be useful to align the entire set of core and flake models with each other as well, since there exists in the point of percussion one of the few morphological landmarks that remains constant between all cores (cf. Archer et al. 2017) and could therefore be useful for the model's training as well as posterior analyses. This not only meant that the cores would need to be located in 3D space in a way that makes the point of percussion be at the same coordinates in 3D space for all cores, but also that the platform surface should follow a standard alignment across all the meshes (see Fig. 4).

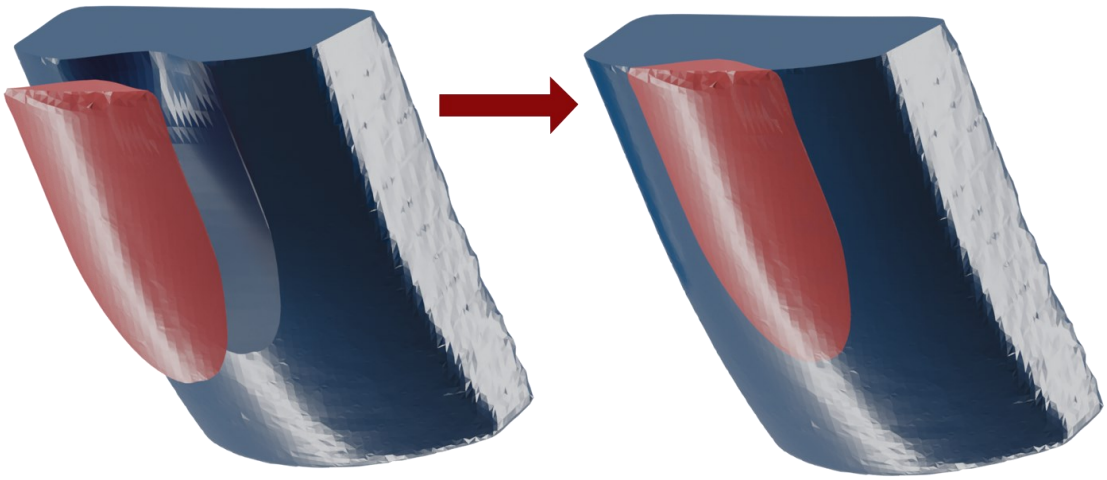

Figure 3: Example of an unmodified core model (blue) and a flake model (red) refitted and aligned, matching the exact shape of the unmodified core from which the flake was knapped. Figure 1 (cropped) from Orellana Figueroa et al. (2021) licensed under CC BY 4.0.

Figure 4: Example of cores post-flake removal, aligned with the point of percussion (white dot) in the same physical location along all cores. Since these core shapes have their top platform surface perfectly flat, one can easily align all cores so that the platform surface is perfectly horizontal. Figure 5 from Orellana Figueroa et al. (2021) licensed under CC BY 4.0.

Having addressed the *what* of the data needed for building a Virtual Knapper able to digitally simulate knapping, I should focus on *how much*. An important consideration to consider is that cores shapes can have a great degree of variation; virtually infinite, so long as the right platform angle exists. Capturing this variability would necessarily require a large amount of data, otherwise it is quite possible that the machine learning model would not generalise; i.e. that it would only be able to knap the core shapes it had trained for, but not different ones.

For machine learning models to *generalise* (i.e. to make predictions outside the range of data already seen during training) the training data must contain a large proportion of the variability of the larger set of data that one wishes to predict. These can range from hundreds to millions of training inputs (e.g. Ridnik et al. 2021), depending on the specific architecture of the machine learning model. It is likely that some several hundred unique core morphologies could be the minimum for the amount of data required in order to capture enough variability for the model to predict resulting flakes for the widest possible range of core shapes. In addition, the presence of flakes scars in the core surface would also alter the core surface morphology, thus affecting

prediction. One would then need to add a requirement for the input data to contain several hundred core and flake pairs in different stages of a reduction sequence. The precise amount would be difficult to calculate, owing to not knowing how well the machine learning model could generalise from the given inputs *a priori*. It is unclear whether the expedient use of complete reduction sequences of only a few hundred cores would capture enough real-world variability with which to train the Virtual Knapper model, or whether it would be necessary to expand the dataset even further for the model to simulate knapping with even more core shapes.

Regarding raw material considerations for the input dataset, as I discussed above, recent research has suggested that only the force necessary to detach a flake is significantly different across materials (Dogandžić et al. 2020). This implies that the type of raw material is not important in terms of training data necessary, and that the model did not need to be limited to a single type of raw material if it were able to account for force applied (or assume that all knapping strikes are successful removals), nor that the already high amount of cores and flakes described above needed to be multiplied for each different raw material if the Virtual Knapper were to also be able to correctly predict across raw materials.

It must be remarked, however, that these theoretical hundreds of cores and flakes discussed here would need to be real-life cores and flakes, and these would need to have been digitised through a 3D scanner. If these data were not already available, then obtaining them would take a considerable amount of time. Even with an optimistic estimate of an hour per hundred core and flake scans using the latest available protocols (not available at the time this work was undertaken; Falcucci 2022; Göldner et al. 2022), it would still require hundreds of hours at least to complete, as each item would need to be aligned at the point of percussion (see Fig. 3), but also scans would need to be taken for each stage of core reduction, a process that is not as easily parallelisable unless hundreds of cores are knapped in parallel, and each only knapped once. With the protocols available at the time this work was undertaking, however, the scanning and processing of lithics would take considerably longer, perhaps dozens if not hundreds of times more (assuming only a single artefact could be scanned at a time as opposed to the hundred or so in Göldner et al. 2022).

Regrettably, though some cores and flake scans were ready for use, such comprehensive set of 3D scanned core and flake data was not yet readily available, and creating such a large dataset would be worthy of its own long-term project, as well

as outside the scope of this dissertation. Since no such dataset existed, a *full* Virtual Knapper model could not be created, but designing a proof of concept for the possible data processing pipeline or framework for its development by using the limited data available would still be a crucial step in the creation of the program. Building and evaluating the framework would provide information on the feasibility of developing the full model and signal the need for the creation of a dataset of digital cores and flakes.

The proof of concept for Virtual Knapper framework would then need to use a much smaller set of data, with the alternative being computer-generated data made to resemble the real inputs as closely as possible. Limiting the input data to only a handful of cores but still including entire reduction sequences would allow testing—on this more limited dataset—of the machine learning methods envisioned for the full dataset, as well verify the possible potential of the chosen approach.

However, machine learning methods designed to predict one 3D shape based on another available during the development of the proof of concept (e.g. Dai et al. 2017; Najibi et al. 2020) were not suitable for the problem of virtual knapping. In order to use machine learning as a foundation of the Virtual Knapper framework, I sought to encode the 3D information of the core surface into 2D images, for which algorithms to predict one image based on another already existed (Isola et al. 2017; Nguyen et al. 2019). The input data was based therefore on an abstraction of the problem of predicting a removed flake from an intact core by transferring the problem from 3D space into two dimensions.

To encode the 3D surface of the core into a 2D image, I used depth maps (also known as *z-buffers*), which are conceptually similar to monochrome digital elevation map rasters, where the areas of higher elevation are encoded as being lighter, whilst the areas of lower elevation are encoded as being darker (see Fig. 5). I would capture the depth maps of the core surface as well as the dorsal surface of the flake itself. Superimposing the depth map of the dorsal surface of the flake on the depth map of the core surface, I could then obtain a depth map representing the core surface prior to that specific removal. With the depth maps of the intact and the modified core surfaces, I calculated a *difference map* showing the volume of mass removed with the flake removal by subtracting the modified core depth map from the intact core depth map. With the difference map and the depth map of the intact core surface, I could calculate the 3D shape of the flake, as well as the depth map of the modified core (see “Appendix 2”: Orellana Figueroa et al. *in press*). Therefore, if the machine learning

model were able to predict an accurate difference map from the core surface depth map alone, I could calculate a predicted 3D flake shape and predicted modified core; thus removing the 3D-to-2D level of abstraction from the predicted result.

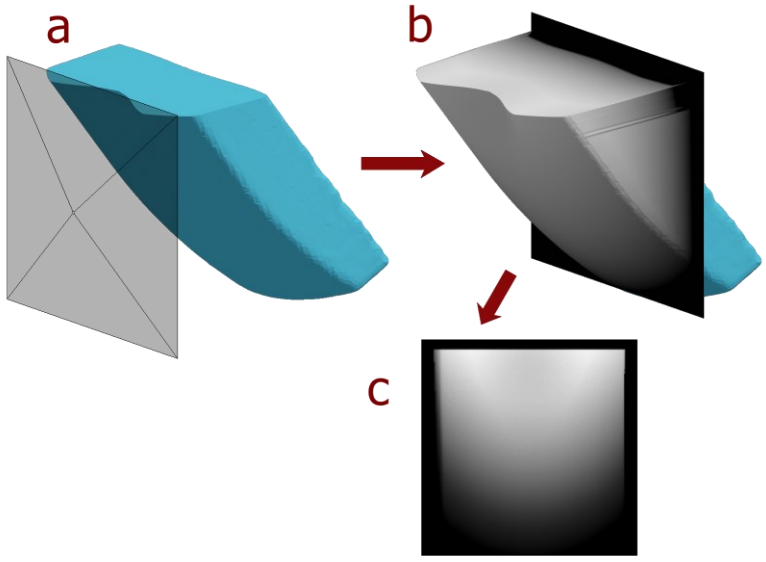

Figure 5: Diagram showing simplified depth map capture procedure and result. Figure 2 from Orellana Figueroa et al. (2021) licensed under CC BY 4.0.

To test if neural network and depth map approach could be successful, I first tested the developed framework with a toy model using very simple computer-generated data made to look how the depth maps of real cores and flakes could look like ("Appendix 2": Orellana Figueroa et al. *in press*). The success of the toy model would allow me to test the framework with more realistic 3D data independently generated by Shannon P. McPherron and subsequently processed by myself, using also a more advanced neural network architecture to serve as a proof of concept for a Virtual Knapper program ("Appendix 1": Orellana Figueroa et al. 2021). The successful development of the Virtual Knapper proof of concept would also lead to an extension of the software to predict not the products of knapping, but to predict the core from which a flake was removed: a proof of concept for a *Virtual Refitter*.

A discussion of the results of the toy model and Virtual Refitter proof of concept described in "Appendix 2" (Orellana Figueroa et al. *in press*), and the Virtual Knapper proof of concept described in "Appendix 1" (Orellana Figueroa et al. 2021), follows in the following chapter.

2. Results and Discussion

In this doctoral dissertation, I explored the various available methods for the creation of a computer-based framework that serves as fast, reliable, reproducible, and valid analogue for lithic replication experiments, dubbed the Virtual Knapper. I discussed the three potential avenues for the development of the Virtual Knapper program: physics-based computer simulations, predictive models based on empirical data, and machine learning-based methods. Although by far the most accurate method available for simulating knapping, physics simulations had the limitation that the fastest methods published at the time this work was undertaken dealt with fracture based on surface deformation, rather than collision (as is the case for knapping). In addition, basing a Virtual Knapper model of the empirical results of controlled experiments (e.g. Dibble and Režek 2009; Režek et al. 2011; Magnani et al. 2014; McPherron et al. 2020; Lin et al. 2022) is difficult, since despite the considerable efforts by many researchers, the effects of various variables and interactions during knapping have yet to be exhaustively explored, especially as all machine flaking experiments (e.g. Dibble and Režek 2009; Režek et al. 2011; Magnani et al. 2014; Dogandžić et al. 2020; McPherron et al. 2020; Lin et al. 2022) have thus far focused on the single initial removal only, and only from very simplified core surfaces.

The third option available for the development of the Virtual Knapper framework, and the one that I chose for this dissertation, was to apply machine learning methods to the problem of virtual knapping simulation. This method front-loaded most of the processing time needed to the training of the model, rather than during use, and since a machine learning-based program can be shared with its model already trained, the process of virtual knapping would require little time and processing power. In addition, one family of machine learning algorithms—neural networks—have found success when applied to various problems where the variables involved are numerous, and their interactions complex, a situation comparable to the problem of simulating knapping virtually.

The initial proof of concept for a possible machine learning-based virtual knapping framework was a simplified toy model that operated on a simplified analogy for flake removals (“Appendix 2”: Orellana Figueroa et al. *in press*). As there was no appropriate set of 3D-scanned lithics data, and creating a dataset of scanned lithics would be outside the scope of this doctoral work, I had to first test any possible Virtual Knapper

framework candidate using a more limited dataset. Since flake removal is essentially the removal of a specific volume (the flake) from another volume (the core), and a Virtual Knapper would predict the removed flake from a given core, I thought to raise an analogy in terms of an explosive volcanic eruption, where the toy model would predict the volume removed (analogous to the flake) from a volcano (analogous to the core) given a specific eruption (analogous to the strike of a hard hammer). The toy model was thus named *Krakatau* ("Appendix 2": Orellana Figueroa et al. *in press*).

The analogy was inspired through the exploration of the use of depth maps to encode the 3D data of lithics into 2D images that could be more easily input to a neural network (see "Introduction"). As discussed in the introduction to this dissertation, depth maps are a 3D graphics analogue of height maps, the latter being used in geographic information systems to encode the height information of a surface; essentially encoding the 3D information of a surface into two dimensions. I generated data based on random sampling of two statistical distributions, using a chi-squared distribution for the vertical axis of the image, and a normal distribution for the horizontal axis (see especially Fig. 2 in "Appendix 2": Orellana Figueroa et al. *in press*). These generated data were made to be similar to what an ideal depth map of the glass cores from Dibble and Režek (2009), which I was to be likely using as an input dataset in a more developed model. The neural network was trained to predict an image of the resulting *removed volcano volume* (analogous to a flake) from an image of the *pre-eruption volcano* (analogous to the unknapped core). The toy model proved successful at accurately predicting the resulting flake analogues from the core surface analogues alone, providing an encouraging signal that further development with the current methodology could provide favourable results.

Since both the neural network architecture used and the dataset generated were both quite simple, an additional, more elaborate model and dataset needed to be tested to more confidently suggest that the proposed Virtual Knapper framework could form the basis (along with a comprehensive dataset of 3D cores and flakes) of a *valid* Virtual Knapper program; one able to perform flake removals and obtain results accurate to real-life knapping.

In "Appendix 1" (Orellana Figueroa et al. 2021), I described the proof of concept for the Virtual Knapper framework with the use of 3D data based on real-life cores independently generated through a Python script made by Shannon P. McPherron, my processing and conversion of the data to 2D images able to capture the 3D surface

information (depth maps), and the training and testing of a Conditional Generative Adversarial Neural Network (CGAN) architecture. The machine learning-based model was successful in predicting the shape of flake removals based only on the information of the 3D surface of the core captured using the depth maps, which also encoded within (by virtue of the way they were obtained; see “Appendix 1”: Orellana Figueroa et al. 2021) the platform depth and the location of the point of percussion. The model was trained with 70% (*n* = 1407) of the core and flake dataset (*n* = 2010), with the remaining 30% (*n* = 603) of the intact core depth maps (i.e. before flake removal) being used to evaluate the model’s ability to predict the shape of the corresponding flake removal. The results showed a high prediction accuracy of the resulting flake length, cube root of flake volume, and overall flake shape, with slightly less accuracy for its prediction of flake width. Moreover, the program was able to build 3D models of the predicted flakes (as well as the modified cores) automatically, thus allowing the products of the Virtual Knapper model to serve as inputs once again (albeit only manually so far), and also allowing for core reduction to theoretically take place almost entirely within the Virtual Knapper framework.

Additional training runs with lower training-testing ratios (i.e. 50%–50%, 30%–70%, and 10%–90%, respectively) provided comparable, albeit slightly less accurate results, with the 10%–90% split predictably obtaining the lowest prediction accuracy for all metrics, with flake width prediction failing completely, likely due to the larger noise in the predicted flakes being mistaken for actual data by the simple flake area detection algorithm used for the metrical analysis. The flake width prediction accuracy of the next smallest split (30% training, 70% testing), although low, remained acceptable for a model trained with such a small dataset, and was a better indication of the performance of the model with few (but not unreasonably few) training data (*n* = 603). Another co-author of the publication (Jonathan S. Reeves) also carried out an independent replication of the data processing, and the training and testing of the model using a different workstation, obtaining an equivalent outcome. This suggested that the model was robust, and that the accurate results obtained with the main dataset split were representative of the accuracy of the model at flake prediction. The neural network model, being a program capable of removing a flake from a core, is the fundamental core of the Virtual Knapper framework, and is also the most complex development task to accomplish for its creation.

Moreover, although the data was computer-generated, and only a single primary core shape (i.e. the morphology of the core did not change except in the exterior platform angle) was used, the exterior platform angle of the knapped core platform varied from 51° to 90° in steps of 1°, for a total of 40 distinct (albeit, very similar) initial cores available for reduction. Moreover, whilst the cores used were based on those used in Dibble and Režek (2009), the data generation process involved some initial removals not horizontally centred on the core (i.e. struck slightly to one side, rather than in the middle of the core). This is a departure from the machine knapping experiments from which the design of these cores originates (Dibble and Režek 2009; Režek et al. 2011; Magnani et al. 2014; Leader et al. 2017; McPherron et al. 2020; Lin et al. 2022; Li et al. 2023a), as these experiments only struck cores at the horizontal centre.

In addition, the input dataset also contained sets of cores and flakes that had been repeatedly *knapped* during data generation (i.e. a flake was removed from an already modified core), thus also providing the Virtual Knapper model with cores at various stages of the reduction sequence, rather than only initial removals, another important departure from the machine knapping experiments, where only one flake was removed per core (Li et al. 2023a; but see Leader et al. 2017, which used a bevel or concavity to emulate a previous flake removal near the platform surface). The inclusion of these heavily reduced cores provides the model with (additional) much needed variability in core morphology that is also present in real knapping. However, I must remark that, during data generation, the virtual cores were not rotated before flake removal, the point of percussion was simply shifted in position horizontally, the same procedure used for the off-centre initial flake removals. The data generation process for flake removal was thus quite basic and resembled the unifacial simple exploitation strategy from de la Torre et al. (2003).

Despite the limitations in the dataset, the success of the CGAN model for flake shape prediction suggests that the method is suitable for the development of a Virtual Knapper program trained (with a dataset of 3D-scanned real-life knapped cores and flakes) to knap a core to obtain results accurate to real-life knapping; i.e. a *valid* Virtual Knapper. The proof of concept has shown the possible minimum training dataset size required to obtain accurate results; however, the findings only apply to a single initial core morphology with some variations from multiple and off-centre removals.

The necessary addition of different core shapes would certainly increase the dataset size requirements, though it is possible that the increase of the latter is not linearly correlated with the increase of the former; i.e. that adding 100 times more core shapes will not require a hundred-fold increase in dataset size, but perhaps only fifty-fold, or even ten-fold, with further increase providing only diminishing returns. However, since the neural network model makes a predicted flake, rather than a predicted modified core (this is done indirectly, see “Appendix 1”: Orellana Figueroa et al. 2021), it might be possible that the variability of the core surface further away from the flake area would make less of an impact on flake shape; thus reducing the amount of core shapes the program needs to learn from. This, of course, is only speculation, as one cannot know *a priori* how much variability in core shape needs to be included during training for the model to be able to generalise its predictions across *all* core shapes with acceptable levels of accuracy. It is possible for the model to learn to model how core shape and platform variables affect the resulting flake *generally* by looking at only a small set of data, but the reverse (needing considerably more data than shown here) is also true, and the place between those two extremes where the truth lies is yet to be discovered.

Furthermore, other knapping variables that were not considered in the proof of concept (such as the angle of blow and hammer hardness) must also be included for the development of a more *valid* Virtual Knapper. Nevertheless, ensuring that predictions can be accurate for different angles of blow will require additional data from flake removals where the angle of blow is known. The inclusion of failed removals and battering (discussed briefly in “Appendix 1”: Orellana Figueroa et al. 2021), as well as the inclusion of pressure flaking in the model would also require additional data. Other special cases of knapping, such as core shattering and anvil technique would require additional research to find out a possible implementation option. In addition, one particular special case, namely imperfections in the core mass, such as faults, cracks (also from failed removals), or material intrusions, that is likely impossible to add to the model from 3D scanned real-life data alone, unless the internal structure of a real knapped core were manually modelled, and even then, the validity of the handmade 3D model would be an important confound when measuring prediction accuracy, requiring e.g. intra- and inter-rater reliability analyses. Most 3D scanners and digital 3D objects are based only surface data, and the internal structural properties of objects are not commonly captured or modelled by 3D scanners; one advantage of physics

simulations over the chosen method. In the current depth map-based methodology, internal structural imperfections could also not be captured through depth maps either, as many of these imperfections, by definition, are present not only on the surface of the rock (what depth maps do capture), but also inside its volume (what depth maps do not capture). Thus, additional protocols for depth map capture would need to be developed to be able to encode imperfection in the core mass (e.g. by using colour data for such faults in the material).

It might also be necessary that different models be trained specifically for each different hammer type, which would imply multiplying the dataset size requirements by the number of different hammer types the Virtual Knapper should account for. In addition, different raw material could also require a similar process, increasing the data requirements even further. However, one recent study has suggested that different raw materials (when machine-knapped) do not result in significantly different flake shapes, so long as enough force is applied (Dogandžić et al. 2020); implying that perhaps that there is no need for the training of unique models for every unique raw material type.

An alternative, complimentary approach would be to perform data generation in the same way as was done for the proof of concept, but with a large range of core shapes for the initial training of the model (rather than only using similar cores as was the case here), whilst also ensuring the generated flakes are as similar as possible to real-world flakes. Once trained with the virtual data, the model could use an additional (but much smaller) dataset obtained from real-life data and train the model once more with the real-life data. This two-step approach of re-training a previously trained model is known as transfer learning (Aggarwal 2018, pp. 43–44), and allows it to take the advantage granted by training with a large dataset, and apply it to a similar, more relevant dataset, to obtain better results with fewer data compared to if it only used the second dataset. Transfer learning might lower the requirements for the size of the real-world core and flake dataset considerably and speed up the development of a *valid* Virtual Knapper program. However all the previous scenarios were optimistic scenarios; the pessimistic scenario would be that it would indeed require thousands of real cores and flakes for the training dataset.

Of course, to verify the accuracy of the Virtual Knapper with real-world data, there would still be a need to build a testing dataset of real cores and flakes to test the program's predictions against. In fact, one could already test the Virtual Knapper model as it currently stands (i.e. trained with computer-generated data, no angle of blow,

singular raw material, etc.) once a real-life testing dataset (i.e. scanned cores before and after flake removal) were available to verify the model's predictions. Moreover, once testing was complete, the testing dataset could then be added to the training dataset, if necessary, with additional and newer testing data taking its place for further validation.

If the model failed when tested with real-life data, one would then continue the development process step-by-step, adding more data for training (firstly, computer-generated, then real), and testing at each step in development to confirm improvement. However, once the model was able to accurately predict the knapping of a wide variety of real cores, there would be an even clearer signal for the success of the Virtual Knapper framework, and thus, of the completion of the initial stage of the development of a virtual alternative for real-world lithic experiments. At this stage, researchers would be able to use the Virtual Knapper and be guaranteed accurate virtual knapping and would also have the possibility of reducing cores virtually; a *valid* Virtual Knapper. The Virtual Knapper could also serve as a pedagogic tool for students or people interested in archaeology who wish to learn about knapping, core reduction, and lithic analysis, especially where skilled knappers or raw materials are scarce. The implementation of specific additional parameters (e.g. angle of blow, failure to knap) could be required to determine this initial Virtual Knapper as complete, whilst other ones may only be optional and left for future development.

A *valid* Virtual Knapper (i.e. tested on real-life data and with highly comparable results to real-world knapping) would be the base for a more comprehensive future virtual lithic experimentation program; i.e. a *full* Virtual Knapper. The *full* Virtual Knapper would, using the framework of the *valid* Virtual Knapper as a core, ideally provide researchers with the ability to perform large-scale replication experiments entirely within a computer environment.

However, the step from a *valid* Virtual Knapper (itself the next step from the current proof of concept) to a virtual experimentation program will not be small, even when the core functionality—and the most complex problem—of virtual flake removal was successfully implemented. The program would require many additional features along with (ideally) a graphical user interface to set-up large-scale or batch replication experiments, as well as provide summaries of their results. The overarching goal of this dissertation was to begin development for a tool capable of facilitating experimental lithic research. Thus, for the *full* Virtual Knapper to fulfil its purpose, it

must serve as a strictly (virtual) analogue to real-world lithic experimentation; which means the *full* Virtual Knapper must be able to replicate real-life stone tool experiments as realistically as possible for researchers to be able to run replication experiments of as many forms and methodologies as they could in real-life (e.g. reduce cores based on the various methods in de la Torre et al. 2003 to compare with archaeological assemblages).

Nevertheless, additional features would be required for the automation of experiments, such as automatic platform detection (thus, also automated platform depth, platform surface interior angle, and exterior platform angle measurements), automatic core rotation (so the core is aligned as it should be for depth map capture based on the currently devised protocol), and rule-based platform selection (e.g. allowing the experimenter to select the range of platform depths, exterior platform angles, platform surface interior angles, and angles of blow the program can choose from to knap). These are not simple features to develop, and could be considered entire projects on their own, but as the code and data for the Virtual Knapper would be free and open-source, other researchers interested and able to contribute would be free to do so, thus leveraging the power of the wider research community for the development of a powerful tool for lithic experimentation. The capabilities of the envisioned *full* Virtual Knapper cannot be overstated. The ability to not only perform several thousand flake removals in a matter of days or even hours (see “Introduction”), but to also have the products measured, analysed, and stored digitally as 3D objects in that same amount of time would constitute an improvement of orders of magnitude when compared with traditional lithic experimentation.

However, even without these automation features, knapping experiments could be manually undertaken through a graphical user interface. A window displaying a core, where the user can click on the surface to mark the point of percussion (and also input the angle of blow) and have the resulting flake and modified core predicted for the next removal in the reduction would be sufficient for fast lithic experimentation, considering that the products would be measured, analysed, and stored automatically. Moreover, including an aspect of randomness during point of percussion selection would be a comparatively simple task to implement. A feature such as this would allow users to obtain a set of flakes and cores when struck at a slightly different random position, providing a window into the range of variability of products during core reduction, and could also help simulate novice knapper skills (or lack thereof, cf. Pargeter et al. 2020)

during core reduction. The user interface would also serve as a tool for teaching the importance of various knapping variables for core reduction, such as exterior platform angle, platform surface interior angle (see for all these variables Li et al. 2023a) to novice knappers.

With a highly automated system, reproducibility would be uncomplicated, since an entire experiment could be easily repeated by copying some configuration settings file as stored by the software, as well as the dataset of 3D objects used as cores. Although reproducibility might appear more difficult with the more *manual* experiment methodology described above, it would, on the contrary, be trivial for a program to store the locations of the selected points of percussion (simple *x*, *y*, *z* coordinates), along with the range of values for knapping parameters that were selected during knapping (simple numerical values), and the 3D objects of the cores knapped at every step (any desired file type available for storing 3D objects). Virtual lithic experiments could, therefore, be made to be highly reproducible.

Thanks to the ability for computers to perfectly copy digital data, standardising cores across experiments would not require artificial ideal cores (although this would still be possible), since an entire input dataset of core shapes could be copied, shared, and used across different independent experiments. Moreover, the creation of any desired or ideal core shapes would be possible using one of the many 3D graphics software available, thus avoiding having to grind, cast, or shape raw material to ideal shapes (Snyder et al. *forthcoming*). Thus, the Virtual Knapper would also be ideal for simple and widespread data-sharing, as all the cores used for virtual experimentation would be stored digitally, as would the resulting products (flakes and modified cores) obtained from the Virtual Knapper. With the digitally stored data, large repositories of virtually knapped cores and flakes could be quickly accumulated. The program could also store the flakes and cores for each reduction stage, rather than only the core at the end of the reduction, providing researchers with high-resolution data that would require a large amount of time to collect with real-life experiments (i.e. 3D scanning every flake removal and the modified core at every stage of core reduction). Previous studies have demonstrated the importance of analysing *intermediate* stages of reduction for lithic research (Toth 1985; Moore and Perston 2016), so this advantage of virtual knapping should not be overlooked.

The lithics obtained through virtual core reduction could also be helpful for lithicists, especially when, with widespread data-sharing, a large and varied set

assemblages could be available for study. These compiled assemblages would be more exhaustive than any one single lithic experiment or an (usually incomplete) archaeological assemblage and could be used to study the possible range of variability that exists with one or more specific reduction techniques (see “Introduction”). These assemblages could also be used to compare and analyse with archaeological material, and if no available shared virtual assemblage was suitable for a specific analysis, performing a set of virtual core reduction more closely aligned with the necessary requirements for the desired analyses could be completed in a week or less (see “Introduction”). The resulting assemblage could then be shared to an open repository for other researchers to explore further.

In addition, even now only with the proof of concept, the measurement of the products of virtual knapping is fully automated for several variables, as are the simple metrical analyses presented in “Appendix 1” (Orellana Figueroa et al. 2021). Although it is likely that more development would be channelled to the measuring and analysis capabilities of the *full* Virtual Knapper than has been for the current proof of concept, I have shown that the automation of these processes is possible, and is considerably faster than their real-life alternatives, needing less than a minute to measure 603 flakes and cores (“Appendix 1”: Orellana Figueroa et al. 2021). The short time taken for measuring and analysing lithics would be an advantage when working with datasets that include all steps of all core reductions, as it would be possible to analyse the entire sequence in little time, as well as provide the data for desired *intermediate* steps as quickly as for the final assemblage. A *full* Virtual Knapper, although still requiring considerable development, would nevertheless be a powerful tool at the disposal of stone tool researchers, and would help examine questions of lithic production that only large-scale lithic replication experiments could address.

Hearkening back to the current progress with the Virtual Knapper proof of concept (with computer-generated data), it is possible to state that it can already become a useful tool for lithic studies. Since the framework only requires for its inputs an oriented 3D mesh (with its position in 3D space based on the desired point of percussion), the point of percussion, and the measured exterior platform angle to be able to predict simple flake removals, even a basic *manual* (i.e. not automated) lithic replication experiment could already be undertaken with the developed software, though with the large caveat that its prediction accuracy has not yet been validated with real-world knapped cores and flakes.

The implementation of an uncomplicated user interface that allows experimenters to manually rotate cores in 3D space and select the desired points of percussion for the program to knap could be added to the current program, though this would be necessarily more constrained than ideal for the user, since the alignment of the core for depth-map generation is highly specific. With the user interface, the cores used would need to be *manually* rotated to specifically follow the same pattern as the cores the model was trained with (see “Appendix 1”: Orellana Figueroa et al. 2021; and “Appendix 2”: Orellana Figueroa et al. *in press*); that is to say, with the platform surface perfectly perpendicular, and the core positioned to capture what is to become the dorsal surface of the removed flake. In a *full* Virtual Knapper, the rotation for proper alignment would be done automatically by the program, with the user only needing to select the point of percussion. Nevertheless, even with the manual rotation requirements in place, the program could already serve as a prototype for the more complete Virtual Knapper.

If rotation were not required (e.g. using only the unifacial simple partial exploitation of de la Torre et al. 2003), then the model would not require any functionality for core rotation, and e.g. simple unifacial reduction experiments could be run repeatedly with the use of stochasticity for the selection of the core and the location of the point of percussion instead, in the same way as the flake removals were carried out for the data generation for the proof of concept; not unlike how a more feature-rich *full* Virtual Knapper would perform automated experiments. A user interface to select the point of percussion (that does not allow core rotation) could be added, and the paradigm described above of allowing for some randomness to influence the actual location of the point of percussion based on the user input and a desired range of deviation, in a way similar to a novice’s lack of precise knapping control (Pargeter et al. 2020), could also be implemented as a prototype for the more complex Virtual Knapper. However, there are important steps for the development of the program that cannot be skipped, since although a user interface for visualising the 3D objects would increase usability, it is important to first clearly verify that the framework is a valid analogue to real-world knapping to prevent fruitless investment into developing easy-of-use features for a program that could ultimately not be valid, even if the current results do not suggest this could be the case (if not now, then with further training).

The preliminary success of the proof of concept, however, provided clear encouragement to investigate alternative applications for the machine learning

architecture. One such application was to use the model, but to invert the input data and desired output, using the modified core as an input to predict the likely refitting flake. In “Appendix 2” (Orellana Figueroa et al. *in press*) I describe a very early proof of concept for a possible future Virtual Refitter program.

The CGAN model, in this instance, was trained not to predict the likely flake removed based on a depth map of an intact core, but rather the likely flake that would refit a specific flake scar in a core. As discussed in “Appendix 2” (Orellana Figueroa et al. *in press*), the model used the same CGAN neural network, the same core and flake depth maps dataset, and the same training parameters (70% of the data (*n* = 1401) for training and the remaining (*n* = 603) for testing), but the input data was the depth maps of the knapped core, and the output was the predicted dorsal surface of the flake. For testing, once again the model only received the modified core depth maps (i.e. without the matching flake depth maps). Once all the predictions were complete (*n* = 603), the model’s predicted depth maps were each compared (through a basic prediction error algorithm) to the flake dorsal surface depth maps of the testing dataset (*n* = 603), making a list of which flakes in the dataset most resembled the predicted flake, and sorting it by ranking the predicted and actual depth maps’ similarity from highest to lowest. The better the model was able to predict the flake that matched the flake scar in the modified core, the more similar that predicted depth map would be to the depth map of the matching flake, and the higher in the ranking it would be located. Ideally, the model’s prediction would match most closely with the actual refitting flake, and the latter would be placed at the top of the similarity ranking, whilst if the prediction was not very accurate, the actual refitting flake would be further down in the ranking.

The results obtained showed a high prediction accuracy for finding the matching flakes in the top 10 positions of the ranking (73.43%), though the accuracy decreased down to 56.11% to find the matching flake in the top three positions, and a one in three chance of placing the matching flake in the first position in the ranking (34.16%). Thus it is safe to conclude that the model is able to provide a shortlist of possible matching refits from a large dataset (in this case, *n* = 603) with enough accuracy that the refit (if it indeed was present in the assemblage; see below) would be found in most cases.

However, as mentioned in “Appendix 2” (Orellana Figueroa et al. *in press*), the proof of concept for the Virtual Refitter has many limitations. Most notably perhaps is the very rigid requirement of matching rotation, position, and scale across the dorsal flake surface depth maps, requiring considerable data preparation for its use, rendering

it difficult to use in an actual fieldwork scenario. Fragmentary or broken flake removals are also impossible to account for, as the model assumes a single removal per aligned flake scar used as input. The Virtual Knapper concept would lead to orders of magnitude increases in speed for lithic experimentation, but the prototype described here would likely require more time and effort to digitise the data to simply use the model than it would to manual core and flake matching, or even full refitting. Moreover, in a fieldwork scenario, the prototype would only be able to predict the flake removals of intact flake scars, and on archaeological cores, these would mean the final flake removals. Although it would be possible to repeatedly predict possible refits, attach them to the core, scan it, and perform another prediction on the previously occluded flake scars; if a core's reduction sequence was incomplete, then the program would only be able to assist refitting up to the first missing flake removal. It may also perhaps not be possible to use the program if the sequence had too many missing elements (not uncommon in archaeological assemblages, e.g. Fig. 5.1 in Hovers 2012), or even if the *débitage* removed during knapping would render otherwise intact flake scars too incomplete (e.g. Fig. 2 in Proffitt and De La Torre 2014).

For a Virtual Refitter to be practical to use, it should not require precise 3D scanning and alignment of every single piece of material and should ideally be able to make use of the surface colour and texture data of the flake or core to find matching refits, an important aspect of manual refitting, as it provides useful information when finding the corresponding flake removal (López-Ortega et al. 2020). This would require a new training paradigm, perhaps involving randomly (and programmatically) rotating and moving the cores and flakes before capturing the depth maps to have the model learn to make predictions and match refits even when the entire dataset is not all perfectly aligned. Machine learning-based software is already available in mobile phone applications for other domains (e.g. Pangti et al. 2021; Mäder et al. 2021; Carvalho et al. 2022), and a putative mobile-based Virtual Refitter could use the phone's camera to acquire both the desired flake scar to find a refit for as well as the entire assemblage of flakes to search in, providing its refit predictions on the fly. However, the current prototype is a far cry from this ultimate goal, and much more work will be required to develop a program with all these capabilities, but the results provide a sign that a similar approach could be an important part of a possible future Virtual Refitter.

The outcome of this research, preliminary as it is, nevertheless constitutes additional progress to the use and development of machine learning applications in archaeological research, an area of research that—especially in recent years—has slowly begun to reap the benefits of these new methodologies (Bellat et al. *in prep*.[4]; Fiorucci et al. 2020; Bickler 2021; Calder et al. 2022; Argyrou and Agapiou 2022). These technologies could help address pressing questions or solve important problems in the field of archaeology could lead to important breakthroughs in the future. In fact, this research was not the first to apply machine learning or neural networks to the study of prehistoric stone tools, even if it was the first for the problem of virtual knapping (and refitting). Grove and Blinkhorn (2020) used an ensemble of neural networks trained to classify Middle and Later Stone Age (MSA and LSA) lithics to explore what typological indicators are the strongest makers for MSA and LSA lithics, finding that although there is some continuity between the two, there remains clear differentiating markers. However, even more research of applications in the field of lithic analysis has been published since (Emmitt et al. 2022; Bustos-Pérez and Ollé 2024). Machine learning methods have also seen applications in study of bone surface modifications (Byeon et al. 2019), site prospection and identification (Verschoof-van der Vaart and Lambers 2019; Caspari and Crespo 2019; Orengo and Garcia-Molsosa 2019; Orengo et al. 2020; Guyot et al. 2021), automated painted rock art identification and extraction (Jalandoni et al. 2022), and many more (Bellat et al. *in prep*.; Fiorucci et al. 2020; Bickler 2021; Calder et al. 2022; Argyrou and Agapiou 2022). A very recent—and perhaps the most dramatic example—is the virtual *unrolling* and decipherment through various cross-disciplinary methods (of which machine learning forms a strong part of) of extremely fragile carbonised papyri from Herculaneum, buried by the 79 CE eruption of Vesuvius (Parsons 2023; Vesuvius Challenge 2024). This is without mentioning the successful implementation of machine learning in many other fields for solving a large variety of problems (see “Introduction”).

As any other (especially newer) technology, however, it is important that researchers are careful with the methodologies they use in their applications, and that they be careful in not exaggerate the capabilities, nor underrate the caveats and limitations that come with applying machine learning methods to archaeological

[4] a publication from the present doctoral work and sharing first authorship, though not included as part of this dissertation.

questions (Bellat et al. *in prep*.; Calder et al. 2022). Furthermore, although the use of machine learning in this doctoral work does not enter into direct ethical concerns, such as with applications in autonomous driving (Geisslinger et al. 2021; Bachute and Subhedar 2021), criminal sentencing (Ávila et al. 2020; Ryberg 2024; Kawamleh 2024), or policing (Babuta and Oswald 2021), it is nonetheless important to note that even though machine learning might not be any more ethically questionable than any other statistical method in concept, actual implementations in the real world can negatively (and justifiably) sway the perception of such methods in other fields. Being well-conscious of the reasons outlined above, I have sought to discuss in as much clarity and detail as possible the limitations of both proofs of concept and the selected methodology in general, along with the ample amount of work remaining for the development of a *full* Virtual Knapper and Refitter.

I must, despite these limitations, especially highlight the potential utility of the *full* Virtual Knapper for the field of lithics research and even our understanding of our evolution. Stone tool replication experiments remain one of the main methods used by lithic archaeologists to study the history of hominin evolution, but the amount of time and labour required to perform large scale and replicable lithic experiments—and the independent replication of any such experiments—makes difficult the scientific examination of the questions of what stone tools can tell us about human evolution. However, it also hinders the amount, accessibility, and reproducibility of the lithic replication experiments that can be undertaken, as well as the strength of the conclusions drawn from them (see Eren and Meltzer 2024). With the Virtual Knapper, researchers would have the ability to study wider questions of stone tool production, such as those addressed by experiments like that of Moore and Perston (2016). It could also serve as a valuable tool for education and outreach (cf. Sánchez-Martínez et al. 2024), providing an interactive (computer-based) environment to discover knapping and prehistoric stone tools, as well as the effect of different variables in flake removal, much more accessible than—though not in any way a replacement for—real-life knapping.

However, whilst a single set of results alone could disprove any high-level hypothesis, performing additional experiments replicating, validating, and expanding on existing findings are one of the key pillars of the scientific method. For instance, undertaking a replication of the experiment by Moore and Perston (2016) would likely require again several months for every single trial (Moore 2021), as well as labelling,

transporting, sorting, and analysing over one thousand lithic artefacts (Moore and Perston 2016). The viability of undertaking a single experiment alone might be acceptable, but performing a battery of similar experiments is likely infeasible for most scholars, thus limiting the scope of the questions that can be viably explored with experimental stone tool replication.

In addition, the conclusions that researchers can draw from studies that cannot be replicated for one or more reasons (e.g. blank shape is completely unknown) are inherently weaker than if their findings could be corroborated independently (Eren and Meltzer 2024). There exist highly controlled—and highly reproducible—experiments such as the machine glass flaking experiments (Li et al. 2023a), where almost all knapping variables are controlled, down to the exact core shape, the knapping force applied, and the angle of the blow, replacing an inexact human knapper with a machine. However, examining the influence on lithic manufacture of every single knapping variable (including the surface morphology of an already-reduced core), and every combination of every variable and reduction sequence, would likely take entire generations of researchers to accomplish. In addition, the replicability of lithic experiments is predicated on reporting the exact shapes and sizes of the original cores used, as well as following a specific and precise core reduction strategy (e.g. clearly defining where, how, and how many flakes to remove), as well as perhaps additional methodological constraints, which increase the investment required for a single experiment.

Moreover, the study of stone tools often focuses on defined tool forms or reduction sequences, or a typology of lithics based on their shapes, and from this, making inferences on the behavioural, cognitive, anatomical, or socio-cultural evolution in hominins (Toth 1985; Gowlett 2009; e.g. de la Torre and Mora 2014; Shipton and White 2020). However, the true range of variability in flake shape that can come about from altering one or two knapping parameters is still not known, even less when we wish to know their effect during an entire core reduction sequence. Assigning a behavioural or cultural process as the cause of the lithic variability in the archaeological record is fraught with caveats, since reducing a core based on a simple set of rules can result in similar lithic products as those found archaeologically, and without resorting to higher-level explanations (Moore and Perston 2016; Tennie et al. 2017).

If researchers were able to translate their lithic experiments to a virtual environment, where these experiments could be undertaken at a fraction of the real-

world cost, completed in less time, and the results were comparable to the results real-life knapping, the possibility of undertaking a large number of highly reproducible lithic replication experiments would be available for many archaeologists. The proof of concept for such a program as described in this dissertation, the Virtual Knapper, is not only a successful demonstration of the possible capabilities (admittedly with considerable work pending) of a more developed machine learning-based virtual knapping framework, having shown results with high accuracy and robustness, but it could also become the first step in a possible revolution in the study of early stone tools and their relation to hominin cognitive, behavioural, and cultural evolution.

References

Acerbi A, Snyder WD, Tennie C (2022) The method of exclusion (still) cannot identify specific mechanisms of cultural inheritance. Sci Rep 12:21680. https://doi.org/10.1038/s41598-022-25646-9

Aggarwal CC (2018) Neural networks and deep learning: A textbook. Springer, Cham

Andersson C, Tennie C (2023) Zooming out the microscope on cumulative cultural evolution: 'Trajectory B' from animal to human culture. Humanit Soc Sci Commun 10:402. https://doi.org/10.1057/s41599-023-01878-6

Archer W, Aldeias V, McPherron SP (2020) What is 'in situ'? A reply to Harmand et al. (2015). J Hum Evol 142:102740. https://doi.org/10.1016/j.jhevol.2020.102740

Archer W, Pop CM, Režek Z, et al (2017) A geometric morphometric relationship predicts stone flake shape and size variability. Archaeol Anthropol Sci 10:1991–2003. https://doi.org/10.1007/s12520-017-0517-2

Argyrou A, Agapiou A (2022) A Review of Artificial Intelligence and Remote Sensing for Archaeological Research. Remote Sensing 14:6000. https://doi.org/10.3390/rs14236000

Arthur KW (2010) Feminine Knowledge and Skill Reconsidered: Women and Flaked Stone Tools. Am Anthropol 112:228–243. https://doi.org/10.1111/j.1548-1433.2010.01222.x

Ávila F, Hannah-Moffat K, Maurutto P (2020) The seductiveness of fairness: Is machine learning the answer? – Algorithmic fairness in criminal justice systems. In: The Algorithmic Society. Routledge

Babuta A, Oswald M (2021) Machine learning predictive algorithms and the policing of future crimes: Governance and oversight. In: Predictive Policing and Artificial Intelligence. Routledge

Bachute MR, Subhedar JM (2021) Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms. Machine Learning with Applications 6:100164. https://doi.org/10.1016/j.mlwa.2021.100164

Bandini E, Motes-Rodrigo A, Archer W, et al (2021) Naïve, unenculturated chimpanzees fail to make and use flaked stone tools. Open Res Europe 1:20. https://doi.org/10.12688/openreseurope.13186.2

Bandini E, Tennie C (2023) Naïve, adult, captive chimpanzees do not socially learn how to make and use sharp stone tools. Sci Rep 13:22733. https://doi.org/10.1038/s41598-023-49780-0

Barik D, Mondal M (2010) Object identification for computer vision using image segmentation. In: 2010 2nd Int. Conf. Educ. Technol. Comput. Shanghai, pp V2-170-V2-172

Bar-Yosef O, Van Peer P (2009) The *Chaîne Opératoire* Approach in Middle Paleolithic Archaeology. Curr Anthropol 50:103–131. https://doi.org/10.1086/592234

Bellat M, Orellana Figueroa JD, Taghizadeh Mehrjadi R, et al (*in prep*.) Machine Learning Applications in Archaeological Practices: A Review

Bickler SH (2021) Machine Learning Arrives in Archaeology. Adv archaeol pract 9:186–191. https://doi.org/10.1017/aap.2021.6

Bilgen C, Kopaničáková A, Krause R, Weinberg K (2018) A phase-field approach to conchoidal fracture. Meccanica 53:1203–1219. https://doi.org/10.1007/s11012-017-0740-z

Braun DR, Aldeias V, Archer W, et al (2019) Earliest known Oldowan artifacts at >2.58 Ma from Ledi-Geraru, Ethiopia, highlight early technological diversity. Proc Natl Acad Sci USA 116:11712–11717. https://doi.org/10.1073/pnas.1820177116

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

Braun DR, Plummer TW, Ditchfield PD, et al (2009) Oldowan Technology and Raw Material Variability at Kanjera South. In: Hovers E, Braun DR (eds) Interdisciplinary approaches to the Oldowan. Springer, Dordrecht, Netherlands, pp 99–110

Braun DR, Tactikos JC, Ferraro JV, et al (2008) Oldowan reduction sequences: Methodological considerations. Journal of Archaeological Science 35:2153–2163. https://doi.org/10.1016/j.jas.2008.01.015

Bustos-Pérez G, Ollé A (2024) The quantification of surface abrasion on flint stone tools. Archaeometry 66:247–265. https://doi.org/10.1111/arcm.12913

Byeon W, Domínguez-Rodrigo M, Arampatzis G, et al (2019) Automated identification and deep classification of cut marks on bones and its paleoanthropological implications. Journal of Computational Science 32:36–43. https://doi.org/10.1016/j.jocs.2019.02.005

Calder J, Coil R, Melton JA, et al (2022) Use and Misuse of Machine Learning in Anthropology. IEEE BITS Inform Theory Mag 1–13. https://doi.org/10.1109/MBITS.2022.3205143

Carvalho M de A, Marcato J, Martins JAC, et al (2022) A deep learning-based mobile application for tree species mapping in RGB images. International Journal of Applied Earth Observation and Geoinformation 114:103045. https://doi.org/10.1016/j.jag.2022.103045

Caspari G, Crespo P (2019) Convolutional neural networks for archaeological site detection – Finding 'princely' tombs. J Archaeol Sci 110:104998. https://doi.org/10.1016/j.jas.2019.104998

Ciresan DC, Meier U, Masci J, et al (2011) Flexible, High Performance Convolutional Neural Networks for Image Classification. In: Twenty-Second Int. Jt. Conf. Artif. Intell.

Clark G (1969) World prehistory: A new outline, Second edition. Cambridge at the University Press, London

Clay Z, Over H, Tennie C (2018) What drives young children to over-imitate? Investigating the effects of age, context, action type, and transitivity. Journal of Experimental Child Psychology 166:520–534. https://doi.org/10.1016/j.jecp.2017.09.008

Corbey R, Jagich A, Vaesen K, Collard M (2016) The Acheulean Handaxe: More Like a Bird's Song Than a Beatles' Tune? Evol Anthropol Issues News Rev 25:6–19. https://doi.org/10.1002/evan.21467

Cotterell B, Kamminga J (1987) The Formation of Flakes. Am Antiq 52:675–708. https://doi.org/10.2307/281378

Dai A, Qi CR, NieBner M (2017) Shape Completion Using 3D-Encoder-Predictor CNNs and Shape Synthesis. In: 2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR. IEEE, Honolulu, HI, pp 6545–6554

de la Torre I (2011) The origins of stone tool technology in Africa: A historical perspective. Philos Trans R Soc B Biol Sci 366:1028–1037. https://doi.org/10.1098/rstb.2010.0350

de la Torre I, Mora R (2014) The Transition to the Acheulean in East Africa: An Assessment of Paradigms and Evidence from Olduvai Gorge (Tanzania). J Archaeol Method Theory 21:781–823. https://doi.org/10.1007/s10816-013-9176-5

de la Torre I, Mora R, Domínguez-Rodrigo M, et al (2003) The Oldowan industry of Peninj and its bearing on the reconstruction of the technological skills of Lower Pleistocene hominids. J Hum Evol 44:203–224. https://doi.org/10.1016/S0047-2484(02)00206-3

De Oliveira E, Reynaud E, Osiurak F (2019) Roles of Technical Reasoning, Theory of Mind, Creativity, and Fluid Cognition in Cumulative Technological Culture. Hum Nat. https://doi.org/10.1007/s12110-019-09349-1

Debénath A, Dibble HL (1994) Lower and middle paleolithic of Europe. University Museum, University of Pennsylvania, Philadelphia

Dibble HL (1997) Platform Variability and Flake Morphology: A Comparison of Experimental and Archaeological Data and Implications for Interpreting Prehistoric Lithic Technological Strategies. Lithic Technol 22:150–170. https://doi.org/10.1080/01977261.1997.11754540

Dibble HL, 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. J Archaeol Sci 36:1945–1954. https://doi.org/10.1016/j.jas.2009.05.004

Dibble HL, Whittaker JC (1981) New experimental evidence on the relation between percussion flaking and flake variation. J Archaeol Sci 8:283–296. https://doi.org/10.1016/0305-4403(81)90004-2

Dogandžić T, Abdolazadeh A, Leader G, et al (2020) The results of lithic experiments performed on glass cores are applicable to other raw materials. Archaeol Anthropol Sci 12:44. https://doi.org/10.1007/s12520-019-00963-9

Domínguez-Rodrigo M, Alcalá L (2016) 3.3-Million-Year-Old Stone Tools and Butchery Traces? More Evidence Needed. PaleoAnthropology 2016:46–53. https://doi.org/10.4207/PA.2016.ART99

Emmitt J, Masoud-Ansari S, Phillipps R, et al (2022) Machine learning for stone artifact identification: Distinguishing worked stone artifacts from natural clasts using deep neural networks. PLOS ONE 17:e0271582. https://doi.org/10.1371/journal.pone.0271582

Eren MI, Lycett SJ, Patten RJ, et al (2016) Test, Model, and Method Validation: The Role of Experimental Stone Artifact Replication in Hypothesis-driven Archaeology. Ethnoarchaeology 8:103–136. https://doi.org/10.1080/19442890.2016.1213972

Eren MI, Meltzer DJ (2024) Controls, conceits, and aiming for robust inferences in experimental archaeology. Journal of Archaeological Science: Reports 53:104411. https://doi.org/10.1016/j.jasrep.2024.104411

Falcucci A (2022) MicroStone: Exploring the capabilities of the Artec Micro in scanning stone tools

Feldman MW, Laland KN (1996) Gene-culture coevolutionary theory. Trends in Ecology & Evolution 11:453–457. https://doi.org/10.1016/0169-5347(96)10052-5

Fernandes de Mello R, Antonelli Ponti M (2018) Machine Learning: A Practical Approach on the Statistical Learning Theory. Springer Nature Springer, Cham

Fiorucci M, Khoroshiltseva M, Pontil M, et al (2020) Machine Learning for Cultural Heritage: A Survey. Pattern Recognition Letters 133:102–108. https://doi.org/10.1016/j.patrec.2020.02.017

Gallotti R (2018) Before the Acheulean in East Africa: An Overview of the Oldowan Lithic Assemblages. In: Gallotti R, Mussi M (eds) The Emergence of the Acheulean in East Africa and Beyond. Springer International Publishing, Cham, pp 13–32

Gamble C, Gowlett J, Dunbar R (2011) The Social Brain and the Shape of the Palaeolithic. CAJ 21:115–136. https://doi.org/10.1017/S0959774311000072

Geisslinger M, Poszler F, Betz J, et al (2021) Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk. Philos Technol 34:1033–1055. https://doi.org/10.1007/s13347-021-00449-4

Göldner D, Karakostis FA, Falcucci A (2022) Practical and technical aspects for the 3D scanning of lithic artefacts using micro-computed tomography techniques and laser light scanners for subsequent geometric morphometric analysis. Introducing the StyroStone protocol. PLOS ONE 17:e0267163. https://doi.org/10.1371/journal.pone.0267163

Gowlett JAJ (2009) Artefacts of apes, humans, and others: Towards comparative assessment and analysis. J Hum Evol 57:401–410. https://doi.org/10.1016/j.jhevol.2009.04.011

Grove M, Blinkhorn J (2020) Neural networks differentiate between Middle and Later Stone Age lithic assemblages in eastern Africa. PLOS ONE 15:e0237528. https://doi.org/10.1371/journal.pone.0237528

Guyot A, Lennon M, Lorho T, Hubert-Moy L (2021) Combined Detection and Segmentation of Archeological Structures from LiDAR Data Using a Deep Learning Approach. J Comput Appl Archaeol 4:1. https://doi.org/10.5334/jcaa.64

Hahn D, Wojtan C (2016) Fast approximations for boundary element based brittle fracture simulation. ACM Trans Graph 35:1–11. https://doi.org/10.1145/2897824.2925902

Hahn D, Wojtan C (2015) High-resolution brittle fracture simulation with boundary elements. ACM Trans Graph 34:151:1–151:12. https://doi.org/10.1145/2766896

Haidle MN (2010) Working-Memory Capacity and the Evolution of Modern Cognitive Potential: Implications from Animal and Early Human Tool Use. Curr Anthropol 51:S149–S166. https://doi.org/10.1086/650295

Harmand S, Lewis JE, Feibel CS, et al (2015) 3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya. Nature 521:310–315. https://doi.org/10.1038/nature14464

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: Data mining, inference, and prediction, 2nd ed. Springer, New York, NY

Hayden B (2015) Insights into early lithic technologies from ethnography. Philosophical Transactions of the Royal Society B: Biological Sciences 370:20140356. https://doi.org/10.1098/rstb.2014.0356

Herodotus (1920) Herodotus, with an English Translation, 1st edn. Harvard University Press, Cambridge, MA

Herzlinger G, Goren-Inbar N, Grosman L (2017) A new method for 3D geometric morphometric shape analysis: The case study of handaxe knapping skill. Journal of Archaeological Science: Reports 14:163–173. https://doi.org/10.1016/j.jasrep.2017.05.013

Hiscock P (2004) Slippery and Billy: Intention, Selection and Equifinality in Lithic Artefacts. Camb Archaeol J 14:71–77. https://doi.org/10.1017/S0959774304230050

Hovers E (2012) Invention, Reinvention and Innovation: The Makings of Oldowan Lithic Technology. In: Elias S (ed) Origins of Human Innovation and Creativity. Elsevier, pp 51–68

Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-Image Translation with Conditional Adversarial Networks. In: 2017 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR. pp 5967–5976

Jalandoni A, Zhang Y, Zaidi NA (2022) On the use of Machine Learning methods in rock art research with application to automatic painted rock art identification. Journal of Archaeological Science 144:105629. https://doi.org/10.1016/j.jas.2022.105629

Kawamleh S (2024) Algorithmic evidence in U.S criminal sentencing. AI Ethics. https://doi.org/10.1007/s43681-024-00473-y

Kivell TL (2015) Evidence in hand: Recent discoveries and the early evolution of human manual manipulation. Philos Trans R Soc B Biol Sci 370:20150105. https://doi.org/10.1098/rstb.2015.0105

Kopaničáková A, Krause R (2020) A recursive multilevel trust region method with application to fully monolithic phase-field models of brittle fracture. Computer Methods in Applied Mechanics and Engineering 360:112720. https://doi.org/10.1016/j.cma.2019.112720

Kuman K (2014) Oldowan Industrial Complex. In: Smith C (ed) Encyclopedia of Global Archaeology. Springer New York, New York, NY, pp 5560–5570

Kumar N, Belhumeur PN, Biswas A, et al (2012) Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In: Fitzgibbon A, Lazebnik S, Perona P, et al. (eds) Comput. Vis. – ECCV 2012. Springer, Berlin, Heidelberg, pp 502–516

Kunze J, Karakostis FA, Merker S, et al (2022) Entheseal Patterns Suggest Habitual Tool Use in Early Hominins. PaleoAnthropology Vol. 2022 No. 2 (2022): PaleoAnthropology. https://doi.org/10.48738/2022.ISS2.61

Laland KN (2008) Exploring gene–culture interactions: Insights from handedness, sexual selection and niche-construction case studies. Philosophical Transactions of the Royal Society B: Biological Sciences 363:3577–3589. https://doi.org/10.1098/rstb.2008.0132

Lawrence S, Giles CL, Tsoi AC, Back AD (1997) Face recognition: A convolutional neural-network approach. IEEE Trans Neural Netw 8:98–113. https://doi.org/10.1109/72.554195

Leader G, Abdolahzadeh A, Lin SC, Dibble HL (2017) The effects of platform beveling on flake variation. Journal of Archaeological Science: Reports 16:213–223. https://doi.org/10.1016/j.jasrep.2017.09.026

Leakey LSB, Tobias PV, Napier JR (1964) A New Species of The Genus Homo From Olduvai Gorge. Nature 202:7–9. https://doi.org/10.1038/202007a0

Leakey MD (1971) Excavations in Beds I and II: 1960-1963. Cambridge Univ. Press, Cambridge

Lee HW (2013) The Persistence of Mode 1 Technology in the Korean Late Paleolithic. PLoS One 8: https://doi.org/10.1371/journal.pone.0064999

Li L, Lin SC, McPherron SP, et al (2023a) A Synthesis of the Dibble et al. Controlled Experiments into the Mechanics of Lithic Production. J Archaeol Method Theory 30:1284–1325. https://doi.org/10.1007/s10816-022-09586-2

Li L, Reeves JS, Lin SC, et al (2023b) Did Early Pleistocene hominins control hammer strike angles when making stone tools? Journal of Human Evolution 183:103427. https://doi.org/10.1016/j.jhevol.2023.103427

Lin SC, Rezek Z, Abdolahzadeh A, et al (2022) The mediating effect of platform width on the size and shape of stone flakes. PLOS ONE 17:e0262920. https://doi.org/10.1371/journal.pone.0262920

Lin SC, Rezek Z, Dibble HL (2018) Experimental Design and Experimental Inference in Stone Artifact Archaeology. J Archaeol Method Theory 25:663–688. https://doi.org/10.1007/s10816-017-9351-1

Lin S, Režek Z, Braun D, Dibble H (2013) On the Utility and Economization of Unretouched Flakes: The Effects of Exterior Platform Angle and Platform Depth. Am Antiq 78:724–745. https://doi.org/10.7183/0002-7316.78.4.724

López-Ortega E, Morales JI, Ollé A, Rodríguez-Álvarez XP (2020) Avoiding the Blue and Black/White and Gold Argument: An Automated Colour Reference System Applied to Lithic Refit Processes. J Archaeol Method Theory 27:245–270. https://doi.org/10.1007/s10816-019-09426-w

Lyons DE, Young AG, Keil FC (2007) The hidden structure of overimitation. Proc Natl Acad Sci 104:19751–19756. https://doi.org/10.1073/pnas.0704452104

Mäder P, Boho D, Rzanny M, et al (2021) The Flora Incognita app – Interactive plant species identification. Methods Ecol Evol 12:1335–1342. https://doi.org/10.1111/2041-210X.13611

Magnani M, Režek Z, Lin SC, et al (2014) Flake variation in relation to the application of force. J Archaeol Sci 46:37–49. https://doi.org/10.1016/j.jas.2014.02.029

Malaivijitnond S, Lekprayoon C, Tandavanittj N, et al (2007) Stone-tool usage by Thai long-tailed macaques (Macaca fascicularis). Am J Primatol 69:227–233. https://doi.org/10.1002/ajp.20342

McGrew WC (2010) Chimpanzee Technology. Science 328:579–580. https://doi.org/10.1126/science.1187921

McNabb J, Binyon F, Hazelwood L (2004) The Large Cutting Tools from the South African Acheulean and the Question of Social Traditions. Current Anthropology 45:653–677. https://doi.org/10.1086/423973

McPherron SP, Abdolahzadeh A, Archer W, et al (2020) Introducing platform surface interior angle (PSIA) and its role in flake formation, size and shape. PLoS ONE 15:e0241714. https://doi.org/10.1371/journal.pone.0241714

McPherron SP, Alemseged Z, Marean CW, et al (2010) Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia. Nature 466:857–860. https://doi.org/10.1038/nature09248

Mercader J, Panger M, Boesch C (2002) Excavation of a Chimpanzee Stone Tool Site in the African Rainforest. Science 296:1452–1455. https://doi.org/10.1126/science.1070268

Moore MW (2021) Personal Communication

Moore MW, Braun DR (2009) Homo Floresiensis and The African Oldowan. In: Hovers E, Braun DR (eds) Interdisciplinary Approaches to the Oldowan. Springer Netherlands, Dordrecht, pp 25–37

Moore MW, Perston Y (2016) Experimental Insights into the Cognitive Significance of Early Stone Tools. PLOS ONE 11:e0158803. https://doi.org/10.1371/journal.pone.0158803

Morgan TJH, Uomini NT, Rendell LE, et al (2015) Experimental evidence for the co-evolution of hominin tool-making teaching and language. Nat Commun 6: https://doi.org/10.1038/ncomms7029

Motes-Rodrigo A, McPherron SP, Archer W, et al (2022) Experimental investigation of orangutans' lithic percussive and sharp stone tool behaviours. PLOS ONE 17:e0263343. https://doi.org/10.1371/journal.pone.0263343

Motes-Rodrigo A, Tennie C, Hernandez-Aguilar RA (2023) Bone-related behaviours of captive chimpanzees (Pan troglodytes) during two excavating experiments. Primates 64:35–46. https://doi.org/10.1007/s10329-022-01033-w

Moyano IT, Barsky D, Cauche D, et al (2011) The archaic stone tool industry from Barranco León and Fuente Nueva 3, (Orce, Spain): Evidence of the earliest hominin presence in southern Europe. Quaternary International 243:80–91. https://doi.org/10.1016/j.quaint.2010.12.011

Mussi M, Mendez-Quintas E, Barboni D, et al (2023) A surge in obsidian exploitation more than 1.2 million years ago at Simbiro III (Melka Kunture, Upper Awash, Ethiopia). Nat Ecol Evol 7:337–346. https://doi.org/10.1038/s41559-022-01970-1

Najibi M, Lai G, Kundu A, et al (2020) DOPS: Learning to Detect 3D Objects and Predict Their 3D Shapes. In: 2020 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR. IEEE, Seattle, WA, USA, pp 11910–11919

Neadle D, Allritz M, Tennie C (2017) Food cleaning in gorillas: Social learning is a possibility but not a necessity. PLOS ONE 12:e0188866. https://doi.org/10.1371/journal.pone.0188866

Nguyen TT, Nguyen CM, Nguyen DT, et al (2019) Deep Learning for Deepfakes Creation and Detection. ArXiv190911573 Cs Eess

Ni T, Zhu Q, Zhao L-Y, Li P-F (2018) Peridynamic simulation of fracture in quasi brittle solids using irregular finite element mesh. Eng Fract Mech 188:320–343. https://doi.org/10.1016/j.engfracmech.2017.08.028

Odell GH, Odell-Vereecken F (1980) Verifying the Reliability of Lithic Use-Wear Assessments by 'Blind Tests': The Low-Power Approach. J Field Archaeol 7:87–120. https://doi.org/10.1179/009346980791505545

Orellana Figueroa JD, Reeves JS, McPherron SP, Tennie C (2021) A proof of concept for machine learning-based virtual knapping using neural networks. Sci Rep 11:19966. https://doi.org/10.1038/s41598-021-98755-6

Orellana Figueroa JD, Reeves JS, McPherron SP, Tennie C (*in press*) Virtual Knapping (and Refitting) with Neural Networks: Proofs of Concept. In: Kyriakidis P, Agapiou A, Leventis G (eds) CAA2021 Digit. Crossroads Proc. 48th Conf. Comput. Appl. Quant. Methods Archaeol. Tübingen University Press

Orengo HA, Conesa FC, Garcia-Molsosa A, et al (2020) Automated detection of archaeological mounds using machine-learning classification of multisensor and multitemporal satellite data. PNAS 117:18240–18250. https://doi.org/10.1073/pnas.2005583117

Orengo HA, Garcia-Molsosa A (2019) A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery. J Archaeol Sci 112:105013. https://doi.org/10.1016/j.jas.2019.105013

Paige J, Perreault C (2024) 3.3 million years of stone tool complexity suggests that cumulative culture began during the Middle Pleistocene. Proc Natl Acad Sci 121:e2319175121. https://doi.org/10.1073/pnas.2319175121

Pangti R, Mathur J, Chouhan V, et al (2021) A machine learning-based, decision support, mobile phone application for diagnosis of common dermatological diseases. J Eur Acad Dermatol Venereol 35:536–545. https://doi.org/10.1111/jdv.16967

Pargeter J, Brooks A, Douze K, et al (2023) Replicability in Lithic Analysis. Am Antiq 88:163–186. https://doi.org/10.1017/aaq.2023.4

Pargeter J, Khreisheh N, Shea JJ, Stout D (2020) Knowledge vs. Know-how? Dissecting the foundations of stone knapping skill. Journal of Human Evolution 145:102807. https://doi.org/10.1016/j.jhevol.2020.102807

Pargeter J, Khreisheh N, Stout D (2019) Understanding stone tool-making skill acquisition: Experimental methods and evolutionary implications. J Hum Evol 133:146–166. https://doi.org/10.1016/j.jhevol.2019.05.010

Parsons S (2023) Hard-Hearted Scrolls: A Noninvasive Method for Reading the Herculaneum Papyri. PhD thesis, University of Kentucky

Perreault C (2012) The Pace of Cultural Evolution. PLoS ONE 7:e45150. https://doi.org/10.1371/journal.pone.0045150

Perreault C (2019) The quality of the archaeological record. The University of Chicago Press, Chicago

Plummer T (2004) Flaked stones and old bones: Biological and cultural evolution at the dawn of technology. Am J Phys Anthropol 125:118–164. https://doi.org/10.1002/ajpa.20157

Plummer TW, Oliver JS, Finestone EM, et al (2023) Expanded geographic distribution and dietary strategies of the earliest Oldowan hominins and Paranthropus. Science 379:561–566. https://doi.org/10.1126/science.abo7452

Pope SM, Taglialatela JP, Skiba SA, Hopkins WD (2018) Changes in Frontoparietotemporal Connectivity following Do-As-I-Do Imitation Training in Chimpanzees (*Pan Troglodytes*). J Cogn Neurosci 30:421–431. https://doi.org/10.1162/jocn_a_01217

Proffitt T, Bargalló A, de la Torre I (2022) The Effect of Raw Material on the Identification of Knapping Skill: A Case Study from Olduvai Gorge, Tanzania. J Archaeol Method Theory 29:50–82. https://doi.org/10.1007/s10816-021-09511-z

Proffitt T, De La Torre I (2014) The effect of raw material on inter-analyst variation and analyst accuracy for lithic analysis: A case study from Olduvai Gorge. Journal of Archaeological Science 45:270–283. https://doi.org/10.1016/j.jas.2014.02.028

Proffitt T, Luncz LV, Falótico T, et al (2016) Wild monkeys flake stone tools. Nature 539:85–88. https://doi.org/10.1038/nature20112

Proffitt T, Reeves JS, Braun DR, et al (2023) Wild macaques challenge the origin of intentional tool production. Sci Adv 9:eade8159. https://doi.org/10.1126/sciadv.ade8159

Putt SSJ, Wijeakumar S, Spencer JP (2019) Prefrontal cortex activation supports the emergence of early stone age toolmaking skill. NeuroImage 199:57–69. https://doi.org/10.1016/j.neuroimage.2019.05.056

Putt SS, Wijeakumar S, Franciscus RG, Spencer JP (2017) The functional brain networks that underlie Early Stone Age tool manufacture. Nat Hum Behav 1:0102. https://doi.org/10.1038/s41562-017-0102

Putt SS, Woods AD, Franciscus RG (2014) The Role of Verbal Interaction During Experimental Bifacial Stone Tool Manufacture. Lithic Technol 39:96–112. https://doi.org/10.1179/0197726114Z.00000000036

Reindl E, Apperly IA, Beck SR, Tennie C (2017) Young children copy cumulative technological design in the absence of action information. Sci Rep 7: https://doi.org/10.1038/s41598-017-01715-2

Režek Z, Lin S, Iovita R, Dibble HL (2011) The relative effects of core surface morphology on flake shape and other attributes. J Archaeol Sci 38:1346–1359. https://doi.org/10.1016/j.jas.2011.01.014

Ridnik T, Ben-Baruch E, Noy A, Zelnik-Manor L (2021) ImageNet-21K Pretraining for the Masses. https://doi.org/10.48550/ARXIV.2104.10972

Ruck L, Holden C, Putt SSJ, et al (2020) Inter- and Intra-rater Reliability in Lithic Analysis: A Case Study in Handedness Determination Methodologies. J Archaeol Method Theory 27:220–244. https://doi.org/10.1007/s10816-019-09424-y

Ryberg J (2024) Criminal Justice and Artificial Intelligence: How Should we Assess the Performance of Sentencing Algorithms? Philos Technol 37:9. https://doi.org/10.1007/s13347-024-00694-3

Sánchez-Martínez J, Calmet K, Moreno JM, Gilabert XR (2024) Virtual reconstruction of stone tool refittings by using 3D modelling and the Blender Engine: The application of the ‘ReViBE’ protocol to the archaeological record. PLOS ONE 19:e0309611. https://doi.org/10.1371/journal.pone.0309611

Schick KD, Toth N (1994) Making silent stones speak: Human evolution and the dawn of technology. Simon & Schuster, New York, NY

Schick KD, Toth N (2006) An Overview of the Oldowan Industrial Complex: The sites and the nature of their evidence. In: Schick KD, Toth NP (eds) The Oldowan: Case studies into the earliest Stone Age. Stone Age Institute, Gosport, IN, pp 3–42

Schick KD, Toth N, Garufi G, et al (1999) Continuing Investigations into the Stone Tool-making and Tool-using Capabilities of a Bonobo (Pan paniscus). Journal of Archaeological Science 26:821–832. https://doi.org/10.1006/jasc.1998.0350

Schick KD, Toth N, Semaw S (2006) A Comparative Study of the Stone Tool-Making Skills of *Pan*, *Australopithecus*, and *Homo Sapiens*. In: Schick KD, Toth NP (eds) The Oldowan: Case studies into the earliest Stone Age. Stone Age Institute, Gosport, IN, pp 155–222

Sellet F (1993) Chaine Operatoire; The Concept and Its Applications. Lithic Technology 18:106–112. https://doi.org/10.1080/01977261.1993.11720900

Semaw S, Renne P, Harris JWK, et al (1997) 2.5-million-year-old stone tools from Gona, Ethiopia. Nature 385:333–336. https://doi.org/10.1038/385333a0

Semaw S, Rogers MJ, Quade J, et al (2003) 2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia. J Hum Evol 45:169–177. https://doi.org/10.1016/S0047-2484(03)00093-9

Semaw S, Rogers MJ, Simpson SW, et al (2020) Co-occurrence of Acheulian and Oldowan artifacts with Homo erectus cranial fossils from Gona, Afar, Ethiopia. Sci Adv 6:eaaw4694. https://doi.org/10.1126/sciadv.aaw4694

Shipton C, Nielsen M (2015) Before Cumulative Culture: The Evolutionary Origins of Overimitation and Shared Intentionality. Hum Nat 26:331–345. https://doi.org/10.1007/s12110-015-9233-8

Shipton C, White M (2020) Handaxe types, colonization waves, and social norms in the British Acheulean. J Archaeol Sci Rep 31:102352. https://doi.org/10.1016/j.jasrep.2020.102352

Shott MJ, Trail BW (2010) Exploring New Approaches to Lithic Analysis: Laser Scanning and Geometric Morphometrics. Lithic Technol

Sillitoe P, Hardy K (2003) Living Lithics: Ethnoarchaeology in Highland Papua New Guinea. Antiquity 77:555–566. https://doi.org/10.1017/S0003598X00092619

Snyder WD, Boysen D, Orellana Figueroa JD, et al (*forthcoming*) An overview of standardizable raw materials for controlled knapping experiments. Adv Archaeol Pract

Snyder WD, Reeves JS, Tennie C (2022) Early knapping techniques do not necessitate cultural transmission. Sci Adv 8:eabo2894. https://doi.org/10.1126/sciadv.abo2894

Snyder WD, Tennie C (2022) What kind of culture did early hominin toolmakers have?

Spoor F, Leakey MG, Gathogo PN, et al (2007) Implications of new early Homo fossils from Ileret, east of Lake Turkana, Kenya. Nature 448:688–691. https://doi.org/10.1038/nature05986

Stout D (2002) Skill and Cognition in Stone Tool Production: An Ethnographic Case Study from Irian Jaya. Current Anthropology 43:693–722. https://doi.org/10.1086/342638

Stout D (2011) Stone toolmaking and the evolution of human culture and cognition. Philos Trans R Soc B Biol Sci 366:1050–1059. https://doi.org/10.1098/rstb.2010.0369

Stout D, Chaminade T (2007) The evolutionary neuroscience of tool making. Neuropsychologia 45:1091–1100. https://doi.org/10.1016/j.neuropsychologia.2006.09.014

Stout D, Hecht E, Khreisheh N, et al (2015) Cognitive Demands of Lower Paleolithic Toolmaking. PLoS ONE 10:e0121804. https://doi.org/10.1371/journal.pone.0121804

Stout D, Khreisheh N (2015) Skill Learning and Human Brain Evolution: An Experimental Approach. Camb Archaeol J 25:867–875. https://doi.org/10.1017/S0959774315000359

Stout D, Passingham R, Frith C, et al (2011) Technology, expertise and social cognition in human evolution: Technology and cognition in human evolution. Eur J Neurosci 33:1328–1338. https://doi.org/10.1111/j.1460-9568.2011.07619.x

Stout D, Toth N, Schick K, Chaminade T (2008) Neural correlates of Early Stone Age toolmaking: Technology, language and cognition in human evolution. Philosophical Transactions of the Royal Society B: Biological Sciences 363:1939–1949. https://doi.org/10.1098/rstb.2008.0001

Susman RL (1988) Hand of Paranthropus robustus from Member 1, Swartkrans: Fossil evidence for tool behavior. Science 240:781–784. https://doi.org/10.1126/science.3129783

Tennie C (2023) The Earliest Tools and Cultures of Hominins. In: Tehrani JJ, Kendal J, Kendal R (eds) The Oxford Handbook of Cultural Evolution, 1st edn. Oxford University Press

Tennie C, Bandini E, van Schaik CP, Hopper LM (2020) The zone of latent solutions and its relevance to understanding ape cultures. Biol Philos 35:55. https://doi.org/10.1007/s10539-020-09769-9

Tennie C, Caldwell C, Dean LG (2018) Culture, Cumulative. In: Callan H (ed) The International Encyclopedia of Anthropology. John Wiley & Sons, Ltd, Oxford, UK, pp 1–7

Tennie C, Call J (2023) Unmotivated Subjects Cannot Provide Interpretable Data and Tasks with Sensitive Learning Periods Require Appropriately Aged Subjects: A Commentary on Koops et al. (2022) 'Field experiments find no evidence that chimpanzee nut cracking can be independently innovated'. AB&C 10:89–94. https://doi.org/10.26451/abc.10.01.05.2023

Tennie C, Call J, Tomasello M (2006) Push or Pull: Imitation vs. Emulation in Great Apes and Human Children. Ethology 112:1159–1169. https://doi.org/10.1111/j.1439-0310.2006.01269.x

Tennie C, Call J, Tomasello M (2009) Ratcheting up the ratchet: On the evolution of cumulative culture. Philos Trans R Soc B Biol Sci 364:2405–2415. https://doi.org/10.1098/rstb.2009.0052

Tennie C, Call J, Tomasello M (2012) Untrained Chimpanzees (Pan troglodytes schweinfurthii) Fail to Imitate Novel Actions. PLoS ONE 7:e41548. https://doi.org/10.1371/journal.pone.0041548

Tennie C, Premo LS, Braun DR, McPherron SP (2017) Early Stone Tools and Cultural Transmission: Resetting the Null Hypothesis. Curr Anthropol 58:652–672. https://doi.org/10.1086/693846

Timbrell L, Scott C, Habte B, et al (2022) Testing inter-observer error under a collaborative research framework for studying lithic shape variability. Archaeol Anthropol Sci 14:209. https://doi.org/10.1007/s12520-022-01676-2

Tostevin GB (2011) Reduction Sequence, Chaîne Opératoire, and Other Methods: The Epistemologies of Different Approaches to Lithic Analysis Levels of Theory and Social Practice in the Reduction Sequence and Chaîne Opératoire Methods of Lithic Analysis. PaleoAnthropology 351:375

Toth N (1985) The oldowan reassessed: A close look at early stone artifacts. Journal of Archaeological Science 12:101–120. https://doi.org/10.1016/0305-4403(85)90056-1

Toth N, Schick K (2009) The Oldowan: The Tool Making of Early Hominins and Chimpanzees Compared. Annu Rev Anthropol 38:289–305. https://doi.org/10.1146/annurev-anthro-091908-164521

Toth N, Schick KD, Savage-Rumbaugh ES, et al (1993) Pan the Tool-Maker: Investigations into the Stone Tool-Making and Tool-Using Capabilities of a Bonobo (Pan paniscus). Journal of Archaeological Science 20:81–91. https://doi.org/10.1006/jasc.1993.1006

van Ginneken B, Setio AAA, Jacobs C, Ciompi F (2015) Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: 2015 IEEE 12th Int. Symp. Biomed. Imaging ISBI. pp 286–289

Verschoof-van der Vaart WB, Lambers K (2019) Learning to Look at LiDAR: The Use of R-CNN in the Automated Detection of Archaeological Objects in LiDAR Data from the Netherlands. J Comput Appl Archaeol 2:31–40. https://doi.org/10.5334/jcaa.32

Vesuvius Challenge (2024) Vesuvius Challenge 2023 Grand Prize awarded: We can read the scrolls! https://web.archive.org/web/20240428101504/https://scrollprize.org/grandprize. Accessed 2 May 2024

Wang H, Roa AC, Basavanhally AN, et al (2014) Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. JMI 1:034003. https://doi.org/10.1117/1.JMI.1.3.034003

Westergaard GC, Suomi SJ (1995) The stone tools of capuchins (Cebus apella). Int J Primatol 16:1017–1024. https://doi.org/10.1007/BF02696114

Whiten A, McGuigan N, Marshall-Pescini S, Hopper LM (2009) Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. Philosophical Transactions of the Royal Society B: Biological Sciences 364:2417–2428. https://doi.org/10.1098/rstb.2009.0069

Whittaker JC (1994) Flintknapping: Making and Understanding Stone Tools, 1st ed. University of Texas Press, Austin

Wright RVS (2009) Imitative Learning of a Flaked Stone Technology-The Case of an Orangutan. Mankind 8:296–306. https://doi.org/10.1111/j.1835-9310.1972.tb00451.x

Wynn T, Coolidge FL (2007) Did a small but significant enhancement in working memory capacity power the evolution of modern thinking. In: Mellars P, Boyle K, Bar-Yosef O, Stringer C (eds) Rethinking the human revolution. McDonald Institute for Archaeological Research, pp 79–90

Wynn T, Coolidge FL (2011) The Implications of the Working Memory Model for the Evolution of Modern Cognition. Int J Evol Biol 2011:741357. https://doi.org/10.4061/2011/741357

Wynn T, Hernandez-Aguilar RA, Marchant LF, Mcgrew WC (2011) 'An ape's view of the Oldowan' revisited. Evol Anthropol Issues News Rev 20:181–197. https://doi.org/10.1002/evan.20323

Wynn T, McGrew WC (1989) An Ape's View of the Oldowan. Man 24:383–398. https://doi.org/10.2307/2802697

Appendix 1:
Orellana Figueroa, et al. (2021)

A Proof of Concept for Machine Learning-Based Virtual Knapping Using Neural Networks

Abstract

Prehistoric stone tools are an important source of evidence for the study of human behavioural and cognitive evolution. Archaeologists use insights from the experimental replication of lithics to understand phenomena such as the behaviours and cognitive capacities required to manufacture them. However, such experiments can require large amounts of time and raw materials, and achieving sufficient control of key variables can be difficult. A computer program able to accurately simulate stone tool production would make lithic experimentation faster, more accessible, reproducible, less biased, and may lead to reliable insights into the factors that structure the archaeological record. We present here a proof of concept for a machine learning-based virtual knapping framework capable of quickly and accurately predicting flake removals from 3D cores using a conditional adversarial neural network (CGAN). We programmatically generated a testing dataset of standardised 3D cores with flakes knapped from them. After training, the CGAN accurately predicted the length, volume, width, and shape of these flake removals using the intact core surface information alone. This demonstrates the feasibility of machine learning for investigating lithic production virtually. With a larger training sample and validation against archaeological data, virtual knapping could enable fast, cheap, and highly-reproducible virtual lithic experimentation.

Authors: Jordy Didier Orellana Figueroa[1*], Jonathan Scott Reeves[1,3], Shannon P. McPherron[2] & Claudio Tennie[1,2]

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

Introduction

Knapped stone tools provide an abundant and long-lasting record of past behaviours and cognition of prehistoric humans on an evolutionary time scale. As a result, the stone artefact record is one of main pillars upon which our understanding of human evolution—and the evolution of human behaviour and cognition—is built. This understanding comes from building inferential links between formal and technological variation observed in the archaeological record and the behavioural, cognitive, and evolutionary processes that lead to its formation[1–8]. However, these links are not always apparent from the stone tools themselves, even in the earliest lithic technologies[8–20], where the archaeological record is primarily comprised of simpler core and flake tools[10,21–24]. Therefore, archaeologists rely on experimental approaches to replicate stone artefacts under test conditions to determine whether factors such as function[25,26], raw material availability[27], skill[28], technique[29–31], cognition[28,32–36], or culture and social learning[33,34,36–38] played a role in the production (and subsequent discard) of knapped stone tools.

Replication experiments produce insights into the archaeological record but come with some limitations. For one, replication experiments are necessarily affected by the knapper's own conscious and unconscious biases, their knapping experience, their expertise in the manufacture of certain tool forms, and their range of knowledge of various knapping techniques[39]. In addition, replication experiments cannot be easily reproduced, as many variables cannot be controlled under traditional experimental setups with modern knappers, whilst using a different knapper could introduce an additional variable not under control. Some experimenters partly address these issues by standardising the blanks (i.e. cores or flakes). Standardising raw materials can be done by sawing blocks of material into particular shapes, casting standardised shapes in materials like ceramic or glass, and more recently, by 3D milling of materials into particular shapes. In addition, some experimenters have also begun using machine-controlled knapping, focusing on searching for first principles in knapping by isolating the effect of specific variables on flake production[40–43]. However, standardising blanks and building machine-controlled flaking apparatuses comes with a substantial increase in the amount of time and resources required to prepare, measure, and store the experimental equipment and materials. The need for time and resources is further

amplified in the first principles approach, as the number of different experiments needed to investigate the effect of multiple interacting variables is substantial.

One alternative that may circumvent the potentially vast resource and time limitations of traditional lithic experimentation entirely, or otherwise reduce costs, is to develop simulations of stone tool reduction in a digital environment. More specifically, a piece of software able to accurately virtually simulate flaking in three dimensions—comparably similar to actual knapping—would allow for fast, inexpensive, and replicable experiments. Doing so would provide a means to carry out stone tool production experiments in a controlled and reproducible environment for less time and money. The virtual knapping program would also be unaffected by any biases that individual human knappers may have in traditional experiments, and which are hard to control (given also that these biases may in some cases still be unknown).

If knapping could be done virtually, and it were—at least in some cases—a valid substitute for actual knapping, it would serve as a less resource-expensive and more feasible alternative for lithic experiments. Variation in flake shape arises out of a large constellation of parameters that are difficult to systematically test. Having a computer-based model where individual variables could be isolated and examined programmatically would not only increase the speed of what is currently a lengthy process, but could also help us further understand cause and effect relationships of different variables and the interactions between them.

In addition, there would be fewer material requirements, also in terms of long-term storage and transport, since cores could be shaped entirely within a computer, and infinitely duplicated and knapped (and re-knapped), allowing for increased dataset sizes and greater reproducibility. The software could be used to create virtual assemblages testable against actual lithic experiments, examining the influence of certain variables during lithic reduction, or more exhaustively uncovering the possible range of variability of specific reduction techniques. Moreover, the reproducibility and robusticity inherent within a well-made virtual knapping program could even counterbalance some of the error during simulation. A well-crafted virtual knapping program would also be free of human knapper biases entirely, allowing experiments undertaken with it to be more controlled, more reproducible, and perhaps more representative compared to traditional lithic experiments.

A single virtual simulation would ideally take considerably less time to reduce a set of cores than a human knapper would, and even many measurements on the resulting

lithics could be automated and performed at a fraction of the time within the software, given that the products would already be digitised. It would also be much more reproducible than current knapping experiments, especially as the (virtual) knapper's biases could be kept identical for all experiments. Currently, this is not possible to a similar degree due to factors such as differences across knappers (e.g. different skill levels, different modern traditions of knapping) and even within them (e.g. changing motivation, energy, concentration, learning during the experiment).

Here we provide an attempt for a proof of concept of a framework for a virtual knapper using a machine learning approach based on neural networks applied to programmatically created 3D inputs (cores and flakes). Our approach generated a predicted 3D flake and modified core as an output from an intact (i.e. unknapped) core. Our approach proved capable of reliably and validly predicting the length, width, volume, and overall shape of a flake removal from the surface of a core given the point of percussion. We therefore conclude that we successfully created a proof of concept—pathway—for a virtual knapper.

Predicted flakes from a more complete virtual knapper—e.g. using the approach outlined here—could form the basis for (virtual) lithic assemblages to compare with archaeological data, which could also allow archaeologists to examine how the different knapping variables affect the resulting assemblages, and to examine important inferences on the various biological, environmental, and sociocultural factors that could have played a role in the formation of the archaeological assemblages we find in the present; thus, also informing a large part of our understanding of human evolution.

Machine learning.

Arguably, the most intuitive approach for virtual knapping would be physics-based simulations of conchoidal fracture—a type of fracture underlying stone knapping—that would likely require the use of mathematical methods such as finite element analysis (*FEA*). Although the application of FEA for virtual knapping is an important avenue to explore, simulating conchoidal fractures is a resource intensive process, and even the most recent research uses high-performance cluster computers to run simulations[44,45], especially if we wished to simulate more realistic—hence complicated—knapping scenarios. Simulations wishing to examine the effects that different reduction

sequences have on the resulting assemblages, or whether and how some tool *forms* can come about through the reduction of other *forms*[24,29] require large amounts of flake removals and changing of knapping variables, making a FEA approach not entirely viable.

However, FEAs are only one of many approaches available to tackle the development of a virtual knapping program. To address all of the requirements we had set forth for a virtual knapper, we chose to base our method on neural networks. In a similar way as to how neural networks have allowed for drastically increasing the resolution of images in a fraction of the time it takes for computers to render them traditionally[46,47], we sought for our neural network framework to predict a flake removal virtually in a fraction of the time it takes for physics-based simulations.

The primary goal for the virtual knapping program was to be a tool that could reliably perform a virtual replication experiment in a very short time without requiring large amounts of computational resources. To this end, a virtual knapper program should also be able to run on an office computer system, not unlike common agent-based modelling software tools, but it should also accurately simulate real stone flaking—focusing, as a starting goal, on hard-hammer percussion knapping (i.e. flakes removed using a hand-held hammerstone to strike the core) of a single raw material type.

Machine learning is a technique that allows computers to build a model of a set of data automatically by analysing the data and *learning* from it, without requiring the user to manually set-up or adjust the model's parameters[48,49]. The advantage of machine learning-based modelling is that it allows for the bulk of the computational processing—i.e. the *training* of the machine learning model—to be completed prior to the model's practical use; normally requiring only a very small fraction of the computing time needed to train the model in the first place.

Machine learning is a broad field, and encompasses a wide range of methods and algorithms. One such family of algorithms are artificial neural networks, which are broadly based on a simplified model of inter-connected biological neurons[50,51]. Artificial neural networks learn iteratively by a process known as *training*: the network makes predictions from the input data, then evaluates the error in prediction with a mathematical function, and adjusts its neurons and the strength of their connections in order to improve future predictions[51].

Artificial neural networks have gained prominence in recent years, as they are advantageous for highly-dimensional data with large numbers of variables and complex interactions. This advantage is even more important for problems where these interactions are difficult to formulate with traditional statistical modelling, or even when we do not know which variables and interactions are important. For instance, human vision is very good at recognising objects, but programming—or mathematically describing—an algorithm to recognise objects in images would be extremely difficult when done traditionally, but can even surpass human performance in specific scenarios[51,52]. Applications of neural networks include autonomous driving[53], recommendation algorithms[54], and computer-aided medical diagnosis[55,56].

One disadvantage of machine learning, however, is that it often requires a large amount of training data. For our envisioned framework, we required 3D models of a large number of core and flake combinations (i.e. a flake and the core from which it was removed). Such a dataset is not (yet) publicly available, and we did not have the resources to create it ourselves. Moreover, for the initial evaluation of our approach, we sought to avoid adding unnecessary complexity by limiting the shape of the initial cores in our dataset, since—due to the *bias-variance trade-off*—additional variability in a dataset usually requires a larger dataset for the model not to *overfit* to the particular training dataset, performing poorly with new data[51]. In the meantime, we opted instead for programmatically-generated cores and flakes. These have the advantage of being quickly generated with a constrained amount of variability, and if a machine learning model can successfully predict the flakes from this data set, then predicting flakes from a larger more varied data set could likely only be a question of additional training data, as the cores and flakes we used here were based on empirical findings from previous machine-controlled knapping experiments[40]. Unlike previous machine-controlled knapping experiments, however, our flakes were not restricted to a single removal for each core, as we also removed flakes from already knapped cores during data generation (see Fig. 1).

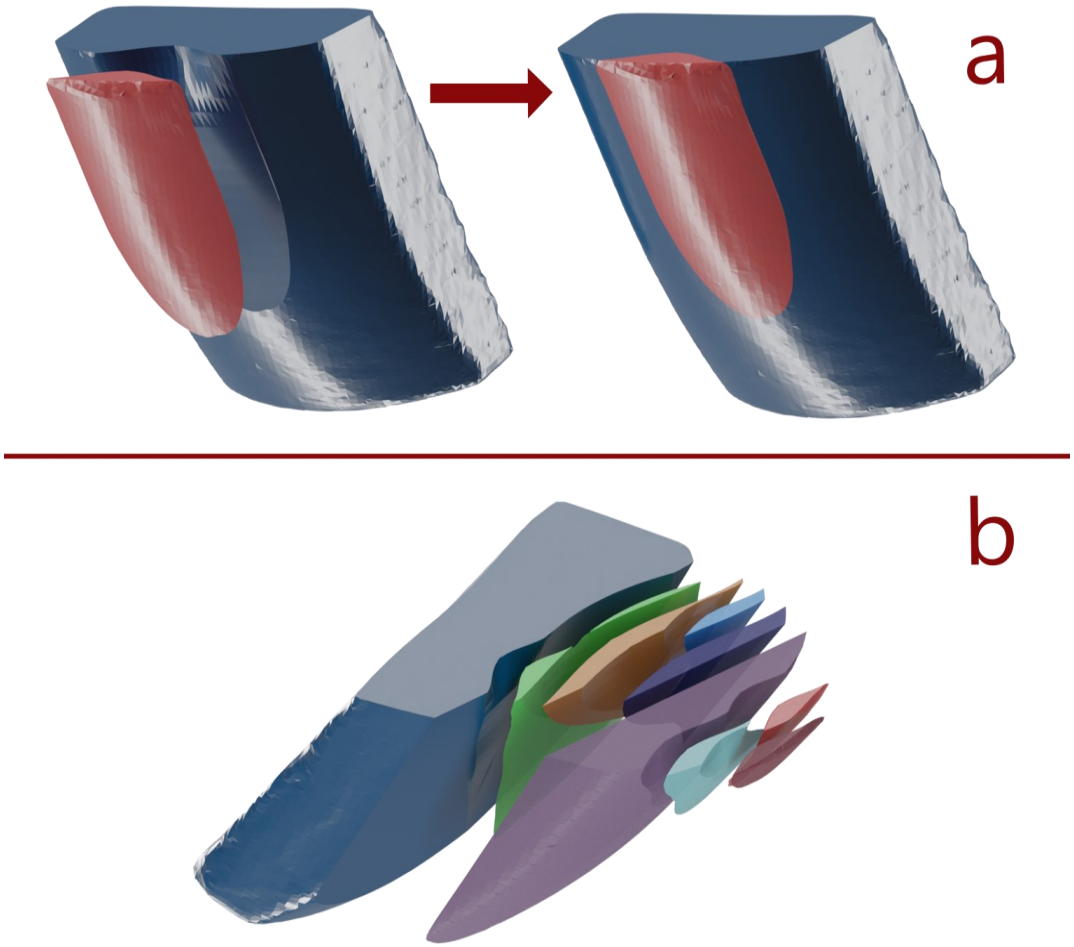

Figure 1. Example of a core and removed flakes from the input dataset. (**a**) The training dataset consists of pairs of flaked cores (blue) and their matching flakes removals (red), oriented such that together they represent the complete core prior to flaking, much like a refit. (**b**) Some of the flake and core pairs were generated in different stages of reduction (see “Methods”). This is an illustration of a generated reduction sequence. Note that, in the dataset, each flake has a matching modified core model as well.

Image-to-image translation.

Neural network algorithms that predict one 3D shape from another are rare or remain limited in their application[57,58]. However, predictions from 2D datasets are far more common. Here, we circumvent this problem by representing our 3D datasets as a two dimensional surfaces to apply image-to-image translation.

Image-to-image translation is a task in which a neural network model converts (or *translates*) one type of picture to another type altogether. Examples include converting

a picture of a landscape taken during the day into a picture of the same landscape at night, converting a line drawing into a photorealistic image, predicting the colourised version of a black and white image, or converting a diagram of a façade into a photorealistic image of a building.

However, since our input consisted of 3D objects, not (2D) images, we needed to encode the information of the relevant surfaces of the 3D cores and flakes into an image. In order to accomplish this task, we made use of depth maps on our 3D cores and flakes.

Depth maps.

Depth maps (or *z-buffers*) are images that encode the distance (or *depth*) between a view point in 3D space from where the depth map is captured, and the 3D surfaces visible from that same point (see Fig. 2). Depth maps are very similar in concept to digital elevation models, which capture the elevation of a portion of the Earth's surface (a 3D property), and encode it into a 2D image whose colours (or raster values) represent different elevations. Depth maps can be conceptualised as a less-restricted form of elevation maps, with the depth map's maximum allowed depth analogous to the lowest surface elevation of a digital elevation model, and the distance between the surface of the object and the view point as analogous to the elevation of the terrain's surface.

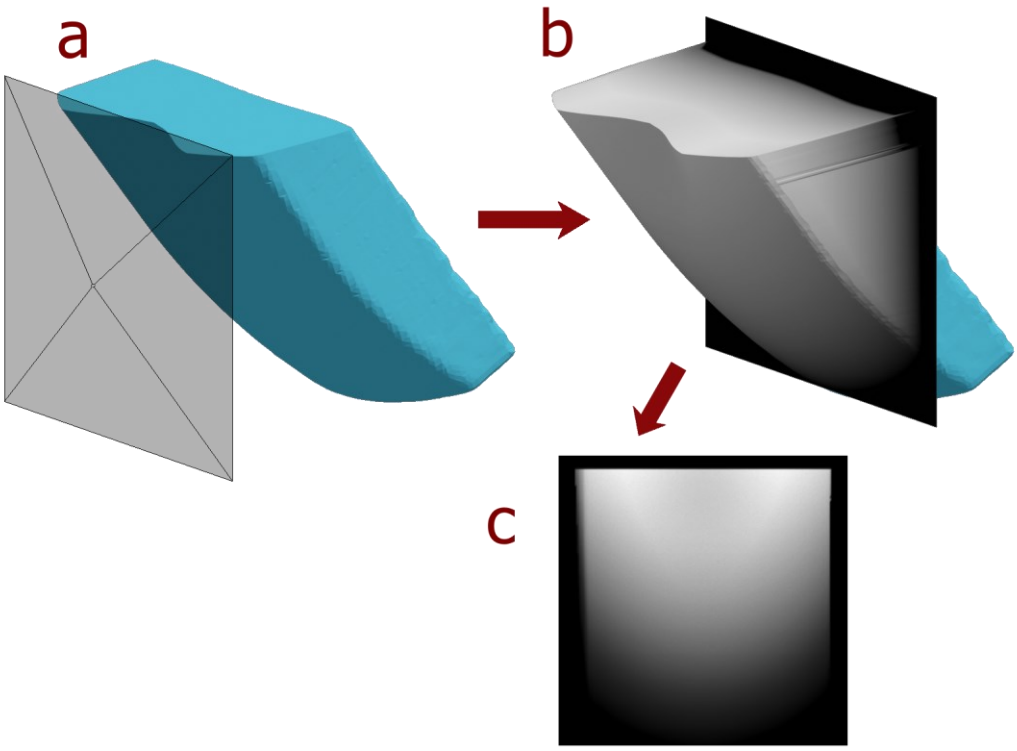

Figure 2. Example of the standard orientation for depth map capture as displayed using a 3-D model of a knapped core, which has a near-perfectly flat platform surface. Note that the platform surface is aligned horizontally with respect to the depth map. Note also that, though difficult to see, the point of percussion is aligned to be in the exact centre of the image. (**a**) 3-D mesh of the core with camera (left) to capture depth map image. (**b**) Depth map rendered into a 3-D surface superimposed over the original core mesh. The depth map's frame is located at the maximum depth we set when captured. Anything deeper than the maximum depth is rendered as pure black in the image. (**c**) Resulting depth map image.

Conditional generative adversarial network (CGAN).

The conditional generative adversarial network (CGAN) architecture consists of a *discriminator* model, which learns to distinguish between the real outputs of our dataset and fake outputs created by a *generator* model, the second part of the CGAN. The generator model learns to create outputs that are realistic enough to fool the discriminator into believing they are real based on the input images. The training process becomes an iterative adversarial contest in which, as the training progresses, the generator becomes better at fooling the discriminator, and the discriminator, in turn,

becomes better at detecting the generator's predicted output. The training ideally culminates in a generator model trained to create outputs that are as close to the real outputs as possible, and able to provide highly accurate predictions under non-training circumstances.

The CGAN performs image-to-image translation by mapping the unmodified core depth maps (input) to the resulting flake volume depth maps (output); what is, in essence, an abstraction of the task of predicting flakes from cores. The predicted flake depth maps obtained as outputs can be then used to obtain the modified core depth map, and with these, calculate the 3D flakes and modified cores using the 3D model of the unmodified core (which would be available in a standard use-case).

Results

The CGAN predicted the depth maps of the flake volumes removed (n = 603) in under 2 minutes, giving an average of less than 200 ms per individual flake prediction. The length, width, volume, and flake shape error calculation of all the predicted depth maps took less than 3 s, giving an average of less than 5 ms per individual prediction (see "Methods" for information on workstation specifications). The CGAN obtained a high degree of accuracy in all measured metrics.

An R^2 of 1.00 (higher is better) and root-mean squared-error (RMSE; see "Methods") of 0.00 (lower is better) would indicate perfect prediction accuracy. For its prediction of flake length, our model obtained an R^2 of 0.85, with an RMSE of 9.15 pixels (see Fig. 3a), but a lower R^2 of 0.58 for its prediction of flake width, with an RMSE of 8.50 pixels (see Fig. 3b). The prediction of the flake's cube root volume obtained an R^2 of 0.77 with an RMSE of 0.76 (see Fig. 3c; see "Methods" for the lack of unit of measurement), indicating a high prediction accuracy by the CGAN.

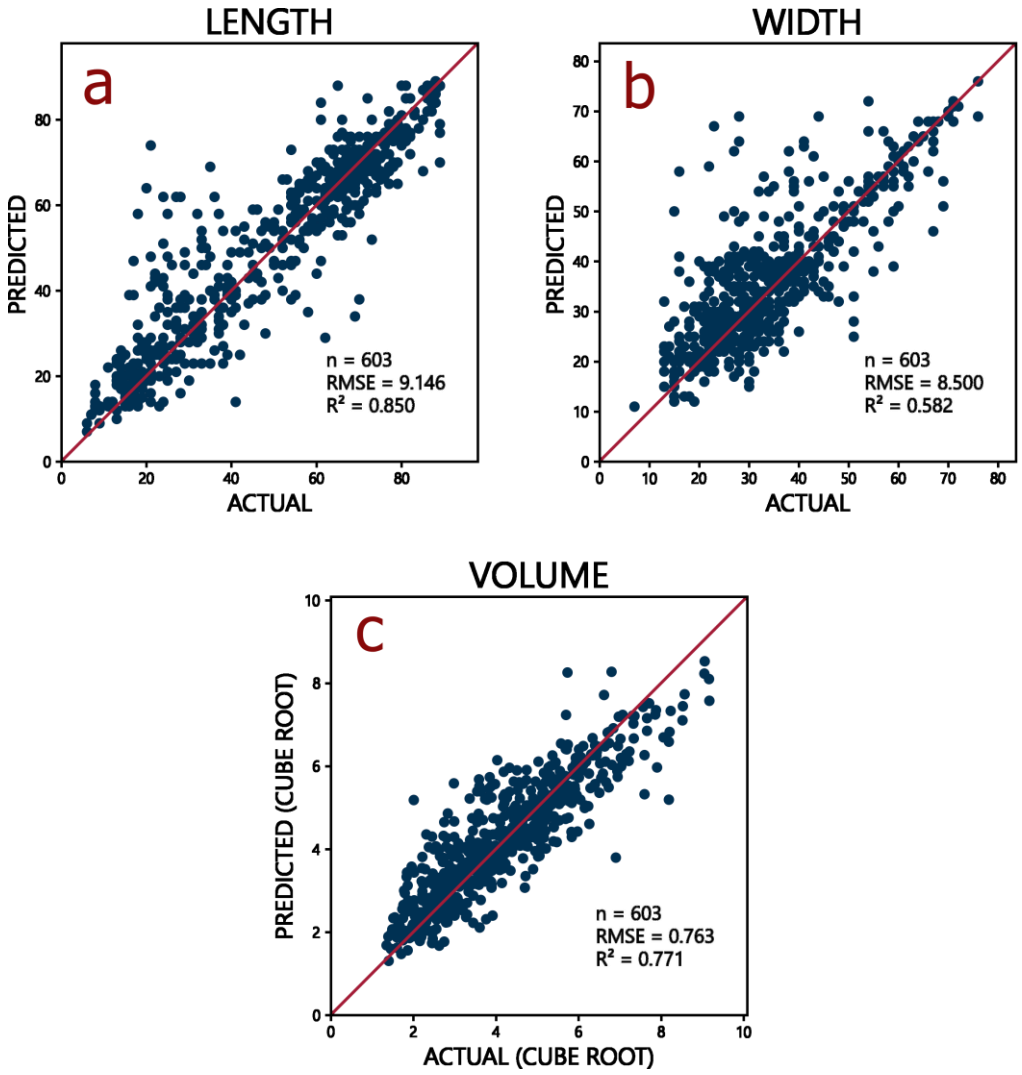

Figure 3. (**a**) Plot of predicted length vs actual length of testing dataset flakes. (**b**) Plot of predicted width vs actual width of testing dataset flakes. (**c**) Plot of predicted cube root of volume vs actual cube root of volume of testing dataset flakes.

In terms of flake shape prediction, we calculated an average mean absolute error (MAE; see "Methods") of 0.024 across all flake predictions. The interval for the data (the range of all *possible* values) was [0, 1], which suggests very low error across predictions. Even when considering the interval for the *actual*—rather than the *possible*—data values of our testing dataset ([0.00, 0.75]), or that of our prediction dataset (i.e. [0.00, 0.52]), the average error remained considerably low, at less than 5% of the interval.

We obtained a very low average RMSE of 0.028 across all flake predictions, but the average normalised root-mean squared-error (NRMSE; see "Methods") was higher, at 0.213, or 21.3%. The higher value of the NRMSE is expected due to the way

it was calculated, which would weigh errors in smaller flakes proportionally much higher than the same amount of error in more voluminous flakes. Our alternate NRMSE calculation ($NRMSE_2$), calculated across all flakes, rather than the average of individual NRMSEs (see “Methods”) had a much lower value of 0.037. Using visual inspection, we can state that the shape of the predicted flakes had a (qualitative) striking resemblance to their respective original input flakes (see Fig. 4). The generation of the 3D models of the predicted flakes from the depth maps took less than 2 minutes; less than 200 ms per individual predicted flake.

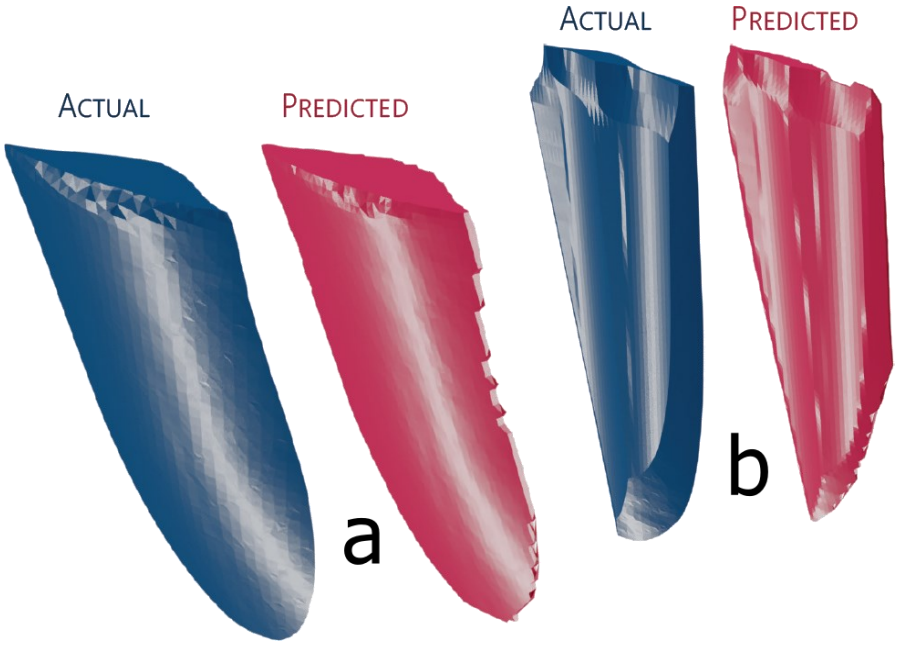

Figure 4. Comparison of two actual vs predicted flakes (**a**,**b**). Note that the size and depth of the predicted flake model was manually scaled to match the size of the actual flake model, though this does not alter the overall shape of the flake (see “Methods” section).

A second, independent, training run on the same workstation obtained very similar results (R^2 of length = 0.85, R^2 of volume = 0.74, R^2 of width = 0.55, average MAE = 0.024, RMSE = 0.028, NRMSE = 0.221, $NRMSE_2$ = 0.037).

The model remained reasonably accurate with different training dataset sizes, except in width prediction, where prediction accuracy went down significantly; though this seem to have been related to other issues (see “Discussion”). The lowest results were obtained with the training dataset size of 10% of the total dataset (training n = 201, testing n = 1809), with a flake length prediction R^2 of 0.66, and RMSE of 13.26; a

flake width prediction R^2 of 0.06, with RMSE of 12.98, and cube root of flake volume prediction R^2 of 0.20 and RMSE of 1.010. We also calculated an average MAE of 0.036, an average RMSE of 0.044, an average NRMSE of 0.314 (or 31.4%), and an $NRMSE_2$ of 0.056.

Discussion

Lithic replication experiments are an important component of human evolutionary research, but replication experiments require considerable material, storage, and time resources to be effective, and being subject to human biases and differences between and within knappers, these experiments become difficult—if not impossible—to reproduce. Even when knapping experiments are replicated, their validity may be affected by knapper's biases and differences. Here we have used machine learning and programmatically-generated core and flake inputs to produce a proof of concept for a virtual knapping program. Such a program would improve the reproducibility of experimental replication studies by being conducted in a digital environment. In addition, by removing a large portion of the biases (and differences between knappers) brought about by the use of human knappers for replication experiments, a virtual knapping framework could allow researchers to more easily examine the influence of different knapping variables and their interactions in shaping the archaeological lithic assemblages; experiments that would be much more prohibitive to undertake in a real-world environment, even with real-life machine-knapping experiments. Moreover, with a singularly-biased computer model, such experiments would be much more controlled and scientific, as the results would not be biased by human factors (e.g. the knapper's mood, stamina, motivation, or even different knappers), which could even allow researchers to examine the effect of knapper biases and differences between knappers on lithic reduction.

With the accurate results of our proof of concept framework, we can start evaluating the performance and efficacy of the approach on more complex datasets that better approximate the real world. However, while it is true that the core shapes used varied primarily in the exterior platform angle (the angle between the platform where the flake is struck and the core surface where the flake is removed), some flakes were taken from an initially smooth core surface and some flakes were taken from a core surface made irregular by the removal of previous flakes. Irregular core surfaces

are more like those found in the vast majority of actually knapped cores. The next step for the evaluation of the framework is to build a model based on actual core and flake pairs, which will first require a large investment in 3D scanning of material, but will add important variability, and in doing so, will increase the external validity of the model[59].

This new approach to virtual knapping could also take advantage of what is known as transfer learning, where a model, already trained with a large dataset, can be additionally trained with a similar, albeit more specific and smaller dataset without sacrificing accuracy in prediction. This type of training could be applied to our model, capturing the benefits of the large numbers of realistic data we generated, as well as requiring a lower dataset size for training with actual flakes and cores.

While it is possible that other variables not measured here, or used for the data generation, contribute to the shape of actual flakes, the framework could be extended to incorporate any number of significant new knapping variables either through the acquisition a broader dataset, or through additional neural network models. Striking a core in the same place with the same exterior platform angle but with a different hammer or angle of blow would produce different flakes. If the effects of these other variables were known, then the core and flake data generation program could be made to include them; otherwise, experimental data sets that include these variables would have to be knapped, scanned and included in the model. An additional solution could involve the training of a predictive model specifically for hard hammer percussion and a separate model specifically for soft hammer percussion. Simulation experiments could then be conducted by virtually knapping identical cores with the two separate models to compare their outcomes. Other variables, such as raw material properties, could be tackled in a similar fashion.

We emphasise that this machine learning approach does not intend to fully replace others; rather, it can work in conjunction with other approaches that seek to understand flake formation[40–43,60,61]. The more we can understand flake formation in general, the better we can build a machine learning model to simulate knapping, since we will know which types of variability are important to introduce and which type are not.

Currently, our proof of concept does not yet have the capacity to detect whether a strike would result in a successful flake removal or a failure to detach one. Our data generation assumed successful flaking in all cases; consequently, the model would be over-confident in removing flakes that in actuality would not be possible to remove, adding error during virtual lithic experiments. A simple solution, considering the

prediction of the neural network is based on a map of volume removed, is to build a dataset of knapping scenarios where no flake would be detached, and use a blank flake *volume removed* depth map to signal the failure to detach a flake. After training with a dataset that includes failed removals, the model could, theoretically, be able to also predict both failed and successful flake removals.

Based on our results, even with the limitations outlined above, we can conclude that a machine learning-based virtual knapper, using actual knapped 3D cores and flakes as input, is—in principle—a feasible approach to building a complete program for virtual lithic experimentation. This we have showed in our proof of principle study here. The main obstacle to a valid and reliable simulation currently lies in access to high quality core and flake 3D datasets of sufficient size. If a more complete virtual knapping were to prove successful at flake prediction once a sufficiently large and varied dataset of actual cores and flakes was available as input, we would have obtained a framework for widespread, fast, and cost-effective virtual lithic experimentation that could be independently verified as reliable and valid (as this proof of concept was) and become and efficient equivalent to actual knapping. Such a program could also serve as a teaching tool for novice knappers for learning how different knapping variables (e.g. platform depth) affect flake removals. A virtual knapper could be used to perform large-scale lithic experimentation virtually at a fraction of the time and cost, without knapper biases, and would be independently replicable.

Methods

Data generation.

Using Python 3[62] and the PyMesh library[63], we programmatically generated a core and flake dataset. As a starting point, we used a 3D scan of an actual glass core used in controlled machine-knapping experiments[40,42]. We then removed flakes from this core in a manner similar to these controlled experiments. These flakes are simplified versions of the actual flakes removed in[40], but they conform to the basic properties of flaking and flake morphology. For the initial 405 flake removals, we only knapped one flake from each core, varying platform depths and exterior platform angles. These two variables are known to play a large part in determining flake outcomes[40], and so by varying them systematically, we were able to produce a variety of flakes.

After the initial 405 flake removals, we also varied the horizontal location along the core edge where the flake was removed. This introduced some asymmetries into the core surface. After an additional 344 flake removals (totalling 749 with the previous 405), we also began removing flakes from already-flaked cores to introduce additional variability in the core surface morphology (see Fig. 1) for an additional 1506 data points.

After removing some cases with errors (e.g. missing surfaces, negative platform depth) through a visual inspection and by programming error checks in the depth map generation code (see Supplementary Data SI1), we ended with a total of 2010 sets of 3D models consisting of a modified (i.e. knapped) core and a flake—both positioned and oriented uniformly based on the point of percussion (see Fig. 5), and together forming the unmodified (i.e. un-knapped) core (see Supplementary Data SI5; Fig. 1). All 3D models were stored as .ply files, and the platform parameters for each flake removal were stored as a .csv file.

Figure 5. All core and flake models follow a standard orientation, in which their platform surfaces are aligned on the same (horizontal) plane, and all models are centred with the point of percussion (white) in the same location in 3D space. The point of percussion varied by changing the distance from the core edge (platform depth), as well as its horizontal position along the platform edge (off-centre). The differently coloured cores represent cores with different exterior platform angles.

Depth map generation.

Using Python 3[62] as well as the Open3D[64] and NumPy libraries[65], we captured depth maps of the topology of the 3D core surface, which we could input into a neural network trained for image-to-image translation (see Supplementary Data SI1)—with the assumption that our captured depth maps encoded enough information of the core surface morphology to allow for accurate predictions of resulting flakes by the model. The depth maps were captured with a dimension of 128×128 pixels. The depth maps captured the surface from which the flake was detached (i.e. the surface with the flake scar), aligning the platform surface of each core (and flake) perpendicularly to the view point, as well as aligning the point of percussion to be in the horizontal centre—and in the same vertical position—in every depth map. The 3D shapes were projected orthographically to the depth map to avoid angle foreshortening from a perspective projection, in case this was to be detrimental to the model's prediction accuracy.

In addition, the maximum depth was calculated based on the platform depth and exterior platform angle (all obtained thanks to knowing the location of the point of percussion) to also encode those variables into the depth map itself; the deeper the platform and the more acute the angle, the larger the maximum depth. The depth maps were normalised to an interval of [0, 1], with the maximum depth set to 0, and the point closest to the view point set to a value of 1.

Although the input data only contained already-knapped cores and the last flake removed, the two together were used to generate the depth map of the core prior to flake removal. Since both the flakes and cores were already aligned in 3D space, the core before flaking could be reconstructed.

With the initial core (unmodified) depth map obtained, we calculated a map of the difference between the modified (flaked) and the unmodified core surface, which shows the volume taken from the core by the knapping of the flake. In our model, we used the volume removed as the desired predicted output of our neural network, rather than a depth map of the flake's ventral or dorsal surface, since the dorsal flake surface is already encoded in the unmodified core depth map, and the ventral surface, in that of the modified core. Thus, we can obtain the shape of the flake removal by calculating the difference between the modified and unmodified core surface depth maps, and we can, in turn, calculate the modified core surface depth map by subtracting the volume removed from the unmodified core surface depth map. In a standard use case

scenario, we would only have the unmodified core surface depth map as an input to the neural network model, which would output a predicted volume removed depth map, with which we could obtain the modified (flaked) core surface and the flake removed.

Neural network training and testing.

With the depth maps of our generated cores and flakes, we built a conditional generative adversarial network (CGAN) for image-to-image translation[66] following the implementation in the TensorFlow documentation[67] using Python 3[62] and the TensorFlow 2 library (see Supplementary Data SI2)[68].

We shuffled the order of our depth map pairs and split our depth map dataset (n = 2010) into two smaller subsets: 70% for the training dataset (n = 1407), and 30% for the testing dataset (n = 603). The training data was shuffled once more when creating the Tensorflow Dataset object.

We trained the CGAN for 150 epochs (see “Supplementary Information S1” for code). Our input was the unmodified core depth maps of the training dataset, and we provided the CGAN with the volume removed depth map as the desired output to learn to predict. The training was done on an Asus Vivobook Pro 17 laptop (N705UD), with a 4-core 8-thread Intel Core i7-8550U CPU, 16 GB of DDR4 RAM, and a dedicated NVIDIA GeForce GTX 1070 GPU. The training process took approximately 2.5 hours using the NVIDIA GPU as a CUDA platform.

After training was completed, we moved to testing the trained model. We input only the unmodified core depth maps from our dataset into our CGAN to obtain a dataset of predicted flake volume depth maps. Prediction for all 603 depth maps took less than 2 minutes total.

Data analysis.

After converting the 3D models of the cores and flakes into 2D depth maps, splitting these into a training and testing dataset, as well as feeding the latter to our neural network to predict flake removals, we measured the predicted depth maps and compared them with the matching depth maps from our output testing dataset (see Supplementary Data SI3).

To calculate prediction accuracy, we compared the predicted flake volume depth maps with those of our testing dataset. Since our analyses were performed on the

depth maps, rather than the 3D objects, the prediction metrics had pixels for units, rather than metric units such as centimetres. We applied common basic quantitative lithic analyses to compare the predicted and testing dataset, and examine the prediction accuracy.

We compared the length, width, and cube root of volume of the flakes across datasets. In order to evaluate the accuracy in predicted flake shape, we calculated the average mean absolute error (MAE), average root-mean squared-error (RMSE), and normalised root-mean squared-error (NRMSE, normalised by the range of values for each testing depth map) between the predicted and actual flake depth map images.

To calculate our metrics, we first set a cut-off threshold to eliminate low-level noise in the predicted depth maps. We used different threshold values (0.1, 0.05, 0.01, and 0.005), but observed that the value of 0.01 provided the best results across all training runs, and was therefore the one used in the reporting of results. We first found all the pixels with values higher than our noise threshold for both testing and predicted flakes, and assigned this area of the image as the flake. For our linear measurements, we used the width and length of this area to calculate flake length and width for both predicted and actual flakes. Therefore, the RMSE for the prediction accuracy for these metrics have pixels as units. To calculate the volume, we summed the *elevation* values of each pixel in the image that was above the noise threshold. It is difficult to assign an actual unit to the depth data, as it is based on abstract and normalised 3D Cartesian distance units; therefore, we reported the RMSE for the volume—as well as the flake shape accuracy metrics—as unit-less.

To prevent artificially reducing the error by using image pixels that contained no data (thus increasing the total number of data points with low values, and reducing the mean error), we calculated the error only for the part of the image that contained either the predicted or actual flake. Areas of the depth map that only had noise or had a value of zero were not used for the calculation. We calculated the difference in each pixel between the predicted and actual depth maps, then calculated the MAE, RMSE, and NRMSE of each flake prediction, with each pixel representing one data point. Once we had obtained the MAEs, RMSEs, and NRMSEs of every individual flake prediction, we calculated the averages for each metric, which we report in our results. Finally, we also calculated a different average NRMSE ($NRMSE_2$) by taking the average RMSE previously calculated, and normalising it by dividing it by the range of testing data values ($y_{max} – y_{min}$), rather than normalising it per flake prediction.

We additionally calculated the RMSE of the prediction using our own code, as well as the coefficient of determination (R^2) between the CGAN's predictions and the testing data using the scikit-learn library's metrics.r2_score function[69].

On a reviewer's request, we performed the calculation of all previously described metrics separately for initial versus subsequent removals (i.e. the first removal from an intact core, and removals from non-intact cores). Since there was no *a priori* labelling of either initial or non-initial flake removals, JDOF visually inspected all cores and compiled a list of initial flake removals. Although great care was taken to include all initial flake removals—and only initial flake removals—there could have been some that were missed, but we considered our labelling was thorough enough that the results would remain valid.

According to the results from these separate analysis (see Supplementary Data SI5), the model had a higher prediction accuracy with initial removals when compared to non-initial removals (e.g. length prediction $R^2 = 0.925$ vs. 0.806), even as the initial flakes were less numerous (n = 243) than flakes from subsequent removals (n = 360). The higher accuracy with initial flakes was true for all metrics, save for width prediction, where the prediction for initial removals was considerably lower compared to that of subsequent removals, with an R^2 of 0.197 vs. 0.596. The pattern was constant for the models trained with different fractions of the data except for the model trained with 10% of the data, which was instead more accurate with non-initial removals (e.g. length prediction $R^2 = 0.785$ vs. 0.591). However, for the analysis of the initial flake removals, the width prediction R^2 was calculation as a negative value (the width prediction for the 10% run was quite low already), which is a possibility with the scikit-learn function used, and suggests that specific model was worse than a constant model.

With the addition of the processing time for the separate analyses, the time taken for the analysis of the 603 predictions was approximately doubled from the original 3 seconds (with the singular analysis) to approximately 6 seconds total (with both the singular and separate analyses).

Finally, using Python 3[62] and the Open3D[64] and NumPy[65] libraries (see Supplementary Data SI4), we transformed the predicted depth maps to predicted 3D models of flakes to perform an additional visual comparison between predicted and actual shape.

These analyses were performed in a custom-built desktop computer, with a 6-core 12-thread AMD Ryzen 5 3600 CPU, and 16 GB of DDR4 RAM.

Due to the current depth-mapping algorithm, in order to produce the visualisation in Fig. 4, we had to manually scale down (i.e. make the model smaller in all dimensions) and reduce the depth (make the model smaller in the z-dimension) of the predicted flakes to match the models of their respective actual flake through visual inspection. The resizing process does not affect flake shape, nor its width and length, and serves as a useful visualisation of the possible accuracy of our framework, even if it is not mathematically precise. Future iterations of the program could allow the resizing of the predicted flake 3D model automatically using the precise scale of the 3D model of the actual flake with some modification of the framework's code. Moreover, the depth map generation could be done using a perspective, instead of an orthographic projection, as we observed that reconstructing the 3D model was more difficult using our remeshing method.

Additional tests.

We trained and evaluated the CGAN using different training dataset sizes to examine the robusticity of our framework: 10% (n = 201), 30% (n = 603), and 50% (n = 1005); see SI5.

Data availability

The dataset generated and analysed during the current study, as well as the code used for the modelling and analysis are available in an Open Science Framework repository: https://doi.org/10.17605/OSF.IO/ANQZF.

Code availability

The code used for the processing and analysis of the generated dataset are available in an Open Science Framework repository: https://doi.org/10.17605/OSF.IO/ANQZF.

References

1. Foley, R. & Lahr, M. M. On stony ground: Lithic technology, human evolution, and the emergence of culture. Evol. Anthropol. Issues News Rev. 12, 109–122 (2003).

2. Wynn, T., Hernandez-Aguilar, R. A., Marchant, L. F. & Mcgrew, W. C. 'An ape's view of the Oldowan' revisited. Evol. Anthropol. Issues News Rev. 20, 181–197 (2011).

3. Stout, D. Stone toolmaking and the evolution of human culture and cognition. Philos. Trans. R. Soc. B Biol. Sci. 366, 1050–1059 (2011).

4. Bar-Yosef, O. & Van Peer, P. The Chaîne Opératoire approach in Middle Paleolithic Archaeology. Curr. Anthropol. 50, 103–131 (2009).

5. Gallotti, R. Before the Acheulean in East Africa: An overview of the Oldowan Lithic Assemblages. In The Emergence of the Acheulean in East Africa and Beyond (eds Gallotti, R. & Mussi, M.) 13–32 (Springer International Publishing, 2018). https://doi.org/10.1007/ 978-3-319-75985-2_2.

6. Muller, A., Clarkson, C. & Shipton, C. Measuring behavioural and cognitive complexity in lithic technology throughout human evolution. J. Anthropol. Archaeol. 48, 166–180 (2017).

7. Muller, A. & Clarkson, C. Identifying major transitions in the evolution of lithic cutting edge production rates. PLoS ONE 11, e0167244 (2016).

8. Dibble, H. L. et al. Major fallacies surrounding stone artifacts and assemblages. J. Archaeol. Method Theory 24, 813–851 (2017).

9. Tennie, C., Premo, L. S., Braun, D. R. & McPherron, S. P. Early stone tools and cultural transmission: Resetting the null hypothesis. Curr. Anthropol. 58, 652–672 (2017).

10. de la Torre, I. The origins of stone tool technology in Africa: A historical perspective. Philos. Trans. R. Soc. B Biol. Sci. 366, 1028–1037 (2011).

11. de Torre, I. & Mora, R. Remarks on the current theoretical and methodological approaches to the study of technological strategies of early humans in Eastern Africa. In Interdisciplinary Approaches to the Oldowan (eds Hovers, E. & Braun, D. R.) 25–37 (Springer Netherlands, 2009). https://doi.org/10.1007/978-1-4020-9060-8_3.

12. Braun, D. R. & Hovers, E. Introduction: Current issues in Oldowan Research. In Interdisciplinary Approaches to the Oldowan (eds Hovers, E. & Braun, D. R.) 1–14 (Springer, 2009).

13. Barsky, D. An overview of some African and Eurasian Oldowan Sites: Evaluation of hominin cognition levels, technological advancement and adaptive skills. In Interdisciplinary Approaches to the Oldowan (eds Hovers, E. & Braun, D. R.) 39–47 (Springer Netherlands, 2009). https://doi.org/10.1007/978-1-4020-9060-8_4.

14. Gowlett, J. A. J. Artefacts of apes, humans, and others: Towards comparative assessment and analysis. J. Hum. Evol. 57, 401–410 (2009).

15. de la Torre, I., Mora, R., Domínguez-Rodrigo, M., de Luque, L. & Alcalá, L. The Oldowan industry of Peninj and its bearing on the reconstruction of the technological skills of Lower Pleistocene hominids. J. Hum. Evol. 44, 203–224 (2003).

16. de la Torre, I. & Mora, R. The transition to the Acheulean in East Africa: An assessment of paradigms and evidence from Olduvai Gorge (Tanzania). J. Archaeol. Method Theory 21, 781–823 (2014).

17. Corbey, R., Jagich, A., Vaesen, K. & Collard, M. The Acheulean handaxe: More like a bird's song than a beatles' tune? The Acheulean Handaxe: More like a bird's song than a beatles' tune?. Evol. Anthropol. Issues News Rev. 25, 6–19 (2016).

18. McNabb, J., Binyon, F. & Hazelwood, L. The large cutting tools from the South African Acheulean and the Question of Social Traditions. Curr. Anthropol. 45, 653–677 (2004).

19. Shea, J. J. Child's play: Reflections on the invisibility of children in the paleolithic record. Evol. Anthropol. Issues News Rev. 15, 212–216 (2006).

20. Wynn, T. & Gowlett, J. The handaxe reconsidered. Evol. Anthropol. Issues News Rev. 27, 21–29 (2018).

21. Kuman, K. Oldowan industrial complex. In Encyclopedia of Global Archaeology (ed. Smith, C.) 5560–5570 (Springer New York, 2014). https://doi.org/10.1007/978-1-4419-0465-2_652.

22. Interdisciplinary Approaches to the Oldowan. (Springer, 2009).

23. Schick, K. D. & Toth, N. An Overview of the Oldowan Industrial Complex: The sites and the nature of their evidence. In The Oldowan: Case studies into the earliest Stone Age (eds Schick, K. D. & Toth, N. P.) 3–42 (Stone Age Institute, 2006).

24. Isaac, G. L. Stages of cultural elaboration in the Pleistocene: Possible archaeological indicators of the development of language capabilities. Ann. N. Y. Acad. Sci. 280, 275–288 (1976).

25. Pettigrew, D. B., Whittaker, J. C., Garnett, J. & Hashman, P. How Atlatl Darts behave: Beveled points and the relevance of controlled experiments. Am. antiq. 80, 590–601 (2015).

26. Eren, M. I. et al. Test, model, and method validation: The role of experimental stone artifact replication in hypothesis-drive archaeology. Ethnoarchaeology 8, 103–136 (2016).

27. Braun, D. R., Tactikos, J. C., Ferraro, J. V., Arnow, S. L. & Harris, J. W. K. Oldowan reduction sequences: Methodological considerations. J. Archaeol. Sci. 35, 2153–2163 (2008).

28. Khreisheh, N. N. The Acquisition of Skill in Early Flaked Stone Technologies: An Experimental Study (University of Exeter, 2013).

29. Moore, M. W. & Perston, Y. Experimental insights into the cognitive significance of early stone tools. PLoS ONE 11, e015880 (2016).

30. Archer, W. & Braun, D. R. Variability in bifacial technology at Elandsfontein, Western cape, South Africa: A geometric morphometric approach. J. Archaeol. Sci. 37, 201–209 (2010).

31. Toth, N. The oldowan reassessed: A close look at early stone artifacts. J. Archaeol. Sci. 12, 101–120 (1985).

32. Putt, S. S. J., Wijeakumar, S. & Spencer, J. P. Prefrontal cortex activation supports the emergence of early stone age toolmakin skill. Neuroimage 199, 57–69 (2019).

33. Putt, S. S., Wijeakumar, S., Franciscus, R. G. & Spencer, J. P. The functional brain networks that underlie Early Stone Age too manufacture. Nat. Hum. Behav. 1, 0102 (2017).

34. Putt, S. S., Woods, A. D. & Franciscus, R. G. The role of verbal interaction during experimental bifacial stone tool manufacture Lithic Technol. 39, 96–112 (2014).

35. Stout, D., Hecht, E., Khreisheh, N., Bradley, B. & Chaminade, T. Cognitive demands of lower Paleolithic toolmaking. PLoS ON 10, e0121804 (2015).

36. Stout, D., Toth, N., Schick, K. & Chaminade, T. Neural correlates of Early Stone Age toolmaking: Technology, language and cognition in human evolution. Philos. Trans. R. Soc. B Biol. Sci. 363, 1939–1949 (2008).

37. Morgan, T. J. H. et al. Experimental evidence for the co-evolution of hominin tool-making teaching and language. Nat. Commun 6, 331–345 (2015).

38. Shipton, C. & Nielsen, M. Before cumulative culture: The evolutionary origins of overimitation and shared intentionality. Hum Nat. 26, 331–345 (2015).

39. Pargeter, J., Khreisheh, N., Shea, J. J. & Stout, D. Knowledge vs. know-how? Dissecting the foundations of stone knapping skill. J Hum. Evol. 145, 102807 (2020).

40. Dibble, H. L. & Režek, Z. Introducing a new experimental design for controlled studies of flake formation: Results for exterio platform angle, platform depth, angle of blow, velocity, and force. J. Archaeol. Sci. 36, 1945–1954 (2009).

41. Režek, Z., Lin, S., Iovita, R. & Dibble, H. L. The relative effects of core surface morphology on flake shape and other attributes. J Archaeol. Sci. 38, 1346–1359 (2011).

42. Magnani, M., Režek, Z., Lin, S. C., Chan, A. & Dibble, H. L. Flake variation in relation to the application of force. J. Archaeol. Sci 46, 37–49 (2014).

43. Lin, S., Režek, Z., Braun, D. & Dibble, H. On the utility and economization of unretouched flakes: The effects of exterior platfor angle and platform depth. Am. Antiq. 78, 724–745 (2013).

44. Bilgen, C., Kopaničáková, A., Krause, R. & Weinberg, K. A phase-field approach to conchoidal fracture. Meccanica 53, 1203–121 (2018).

45. Kopaničáková, A. & Krause, R. A recursive multilevel trust region method with application to fully monolithic phase-field model of brittle fracture. Comput. Methods Appl. Mech. Eng. 360, 112720 (2020).

46. Xiao, L. et al. Neural supersampling for real-time rendering. ACM Trans. Graph. 39, 142:142:1–142:142:12 (2020).

47. Dong, C., Loy, C. C., He, K. & Tang, X. Learning a deep convolutional network for image super-resolution. In Computer Vision ECCV 2014 (eds Fleet, D. et al.) 184–199 (Springer International Publishing, 2014). https://doi.org/10.1007/978-3-319-10593-2_13.

48. Hastie, T., Tibshirani, R. & Friedman, J. H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer 2009).

49. Fernandes de Mello, R. & Antonelli Ponti, M. Machine Learning: A Practical Approach on the Statistical Learning Theory (Springe Nature Springer, 2018).

50. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shif in position. Biol. Cybernet. 36, 193–202 (1980).

51. Aggarwal, C. C. Neural Networks and Deep Learning: A Textbook (Springer, 2018).

52. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification in 2015 IEEE International Conference on Computer Vision (ICCV), 1026–1034 (IEEE, 2015). https://doi.org/10.1109/ICCV.2015.123.

53. Schwarting, W., Alonso-Mora, J. & Rus, D. Planning and decision-making for autonomous vehicles. Annu. Rev. Control Robot Auton. Syst. 1, 187–210 (2018).

54. He, X. et al. Neural collaborative filtering. in Proceedings of the 26th International Conference on World Wide Web 173–182 (International World Wide Web Conferences Steering Committee, 2017). https://doi.org/10.1145/3038912.3052569.

55. El-Dahshan, E.-S.A., Mohsen, H. M., Revett, K. & Salem, A.-B.M. Computer-aided diagnosis of human brain tumor through MRI A survey and a new algorithm. Expert Syst. Appl. 41, 5526–5545 (2014).

56. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Imag Computing and Computer-Assisted Intervention – MICCAI 2015 (eds. Navab, N., Hornegger, J. et al.), 234–241 https://doi.org/10 1007/978-3-319-24574-4_28 (Springer International Publishing, 2015).

57. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation In 2016 Fourth International Conference on 3D Vision (3DV), 565–571 (IEEE, 2016). https://doi.org/10.1109/3DV.2016.79.

58. Wang, W., Huang, Q., You, S., Yang, C. & Neumann, U. Shape inpainting using 3D generative adversarial network and recurren convolutional networks. in 2017 IEEE International Conference on Computer Vision (ICCV), 2317–2325 (IEEE, 2017). https://doi org/10.1109/ICCV.2017.252.

59. Lin, S. C., Rezek, Z. & Dibble, H. L. Experimental design and experimental inference in stone artifact archaeology. J. Archaeol Method Theory 25, 663–688 (2018).

60. Dogandžić, T. et al. The results of lithic experiments performed on glass cores are applicable to other raw materials. Archaeol Anthropol. Sci. 12, 44 (2020).

61. Archer, W. et al. A geometric morphometric relationship predicts stone flake shape and size variability. Archaeol. Anthropol. Sci 10, 1991–2003 (2017).

62. Van Rossum, G. & Drake, F. L. Python 3 Reference Manual (CreateSpace, 2009).

63. Zhou, Q. PyMesh/PyMesh (PyMesh Development Team, 2020).

64. Zhou, Q.-Y., Park, J. & Koltun, V. Open3D: A modern library for 3D data processing. arXiv preprint arXiv:1801.09847 (2018).

65. Oliphant, T. E. A guide to NumPy Vol. 1 (Trelgol Publishing USA, 2006).

66. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632.

67. TensorFlow. Pix2Pix | TensorFlow Core. TensorFlow Core https://www.tensorflow.org/tutorials/generative/pix2pix (2020).

68. Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv preprint arXiv:1603 04467 (2016).

69. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).

Acknowledgements

We would like to thank Isabel Valera and Mehdi Sajjadi for many helpful suggestions and useful advice, Korinna Allhoff for the continuous support given throughout the project, Will Archer for the help and advice provided during the earlier stages of the project, and Mark W. Moore for the many fruitful discussions on this project. CT thanks Cesare de Filippo and Chris Nolan for theoretical debates related to this project. SPM thanks Aylar Abdolahzadeh, Tamara Dogandžić and Li Li for their help in scanning glass cores. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 714658). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

All authors made substantial contributions to the conception of the work. SPM created the software for the data generation. JDOF created all other software. JDOF

performed the data analysis. JDOF wrote the main manuscript text and prepared all figures. All authors reviewed and contributed substantial revisions to the manuscript.

Competing interests

The authors declare no competing interests.

Appendix 2:
Orellana Figueroa, et al. (*in press*)

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.

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.

A Computer-Based Alternative

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.

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.

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.

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.

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

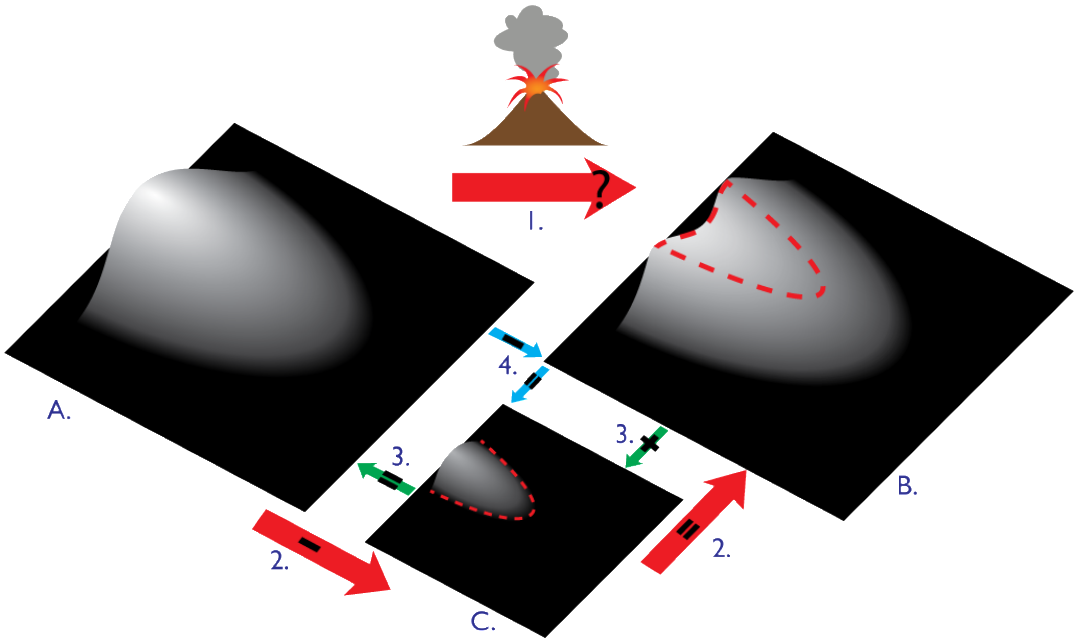

Figure 1. Diagram showing an overview of the Krakatoa model data concept. A: Original volcano heightmap. B: Post-eruption volcano heightmap. C: Heightmap of the volume of material removed from the original volcano during eruption. 1: Process of eruption turns *A* into *B*, the latter is what we wish 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.

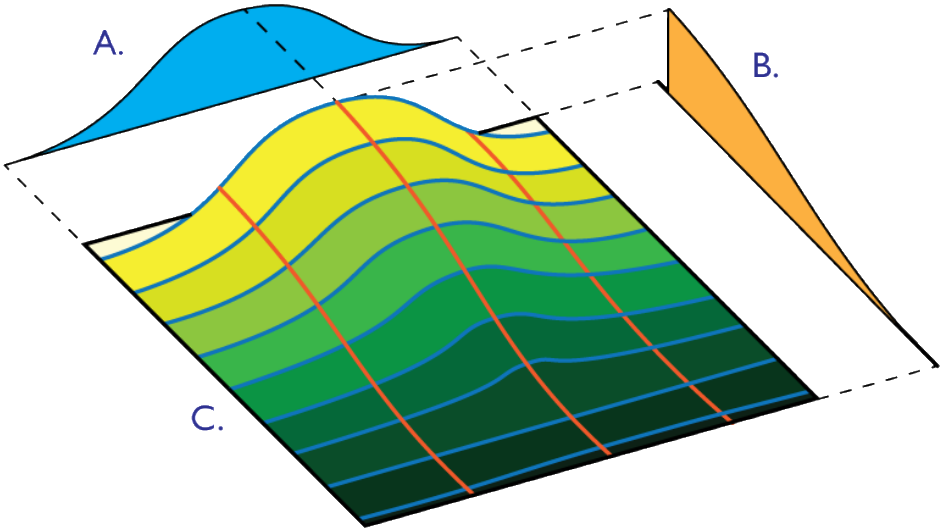

Figure 2. Diagram depicting the generation of the heightmaps through the use of two probability distributions. A: Normal distribution used for the shape of the x-axis. B: Non-central chi-squared distribution used for the shape of the y-axis. C: Combination of both distributions in two dimensions, generating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

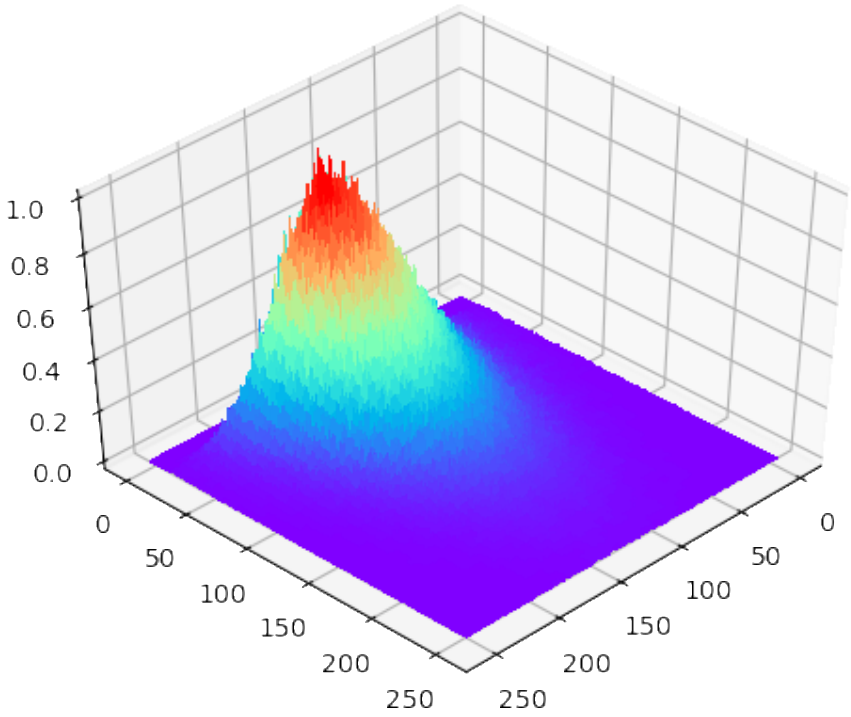

Figure 3. Example of a normalized *volcano* heightmap. Note that the maximum height of the *volcano* was set during generation to a random value – in this case, slightly above 0.9.

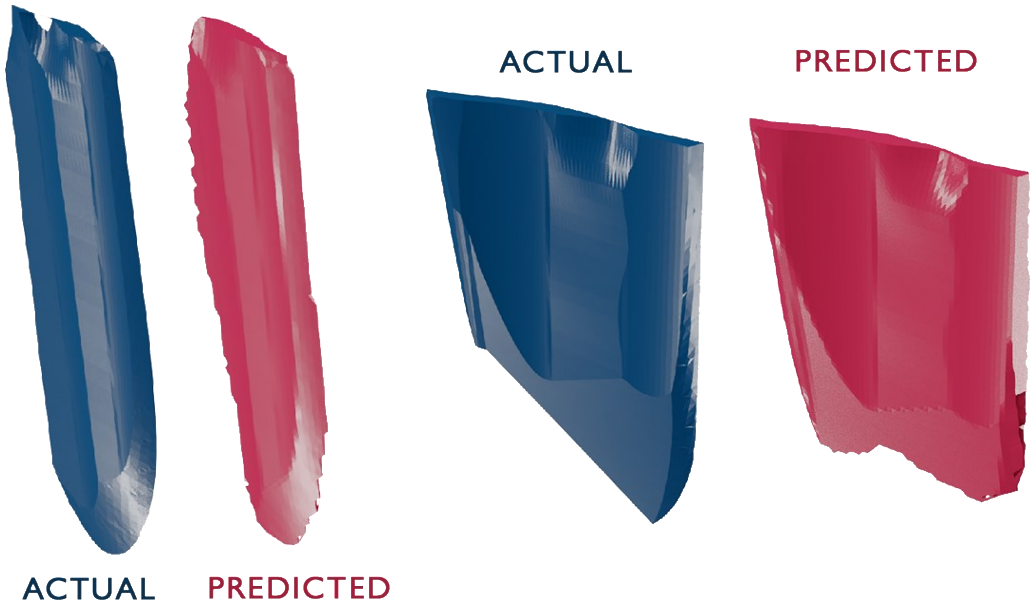

Figure 4. Side-by-side comparison of two predicted flakes with their counterparts in the testing dataset.

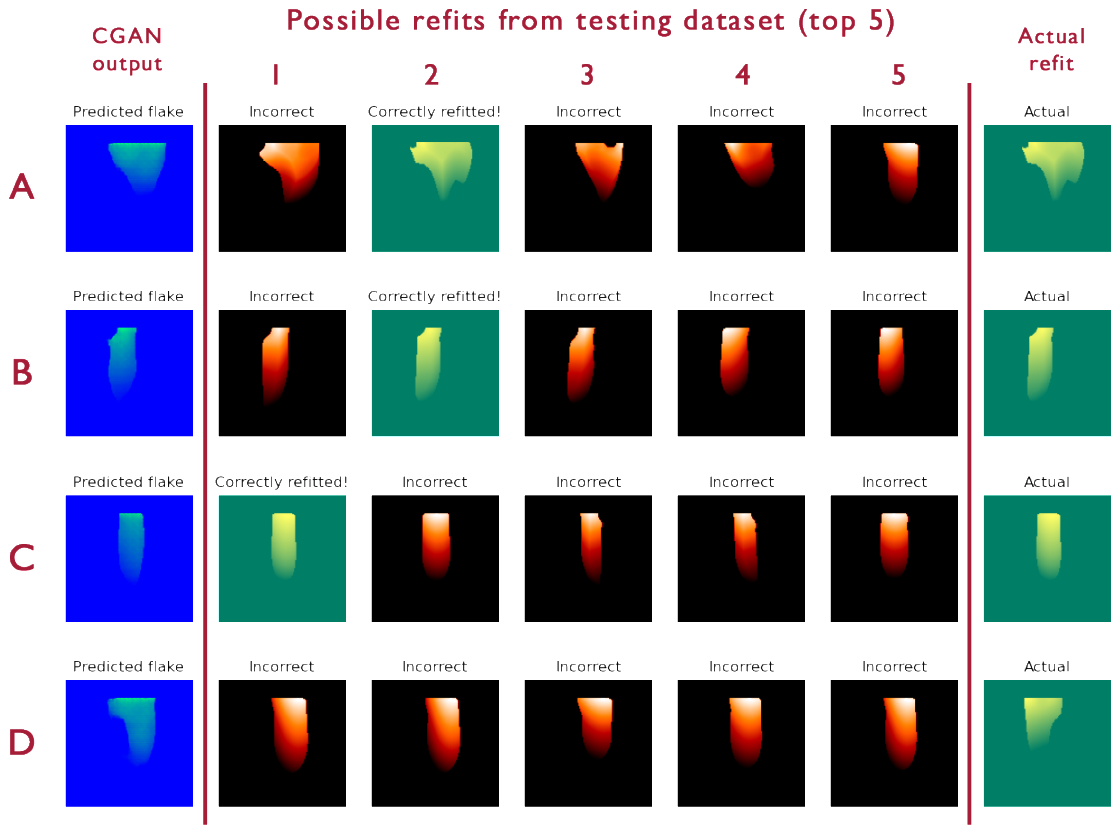

Figure 5. Visualization of the output of the virtual refitter prototype. On the left (*CGAN output*) is the predicted shape of the flake. In the middle are shown the top five possible refits selected by the program. On the right (*Actual refit*) is the actual refitting flake being sought. The refitter 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.