On Computing Explanations for Proportion Estimation in Ancient DNA [Experiment, Analysis and Benchmark]

Amit Bergman
Tel Aviv University
Tel Aviv, Israel

Viviane Slon
Tel Aviv University
Tel Aviv, Israel

Daniel Deutch
Tel Aviv University
Tel Aviv, Israel

ABSTRACT

In ancient DNA analysis, a sample from an archaeological site may include a set of DNA sequences originating from multiple species, and the goal is to estimate the proportion of each species. State-of-the-art solutions lack explainability, which often leads to errors or ambiguity when the results are presented to domain experts or fed to downstream analysis tasks. In this paper we provide an explainable solution for the problem and an in-depth analysis of its performance. The solution is based on a simple Maximum Likelihood algorithm equipped with explanations of two flavors, namely attribution using Shapley values and counterfactuals. We highlight the usefulness of explanations in this setting and thoroughly analyze the performance of a suite of algorithms that compute them. Our analysis sheds light on several practical approaches for explainability in this context.

1 INTRODUCTION

This paper focuses on the problem of *proportion estimation* [68, 72, 73] in the context of ancient DNA. DNA fragments can be retrieved from sediments deposited at archaeological sites (e.g., [52, 67, 71]), and a single dataset may contain DNA sequences originating from *multiple organisms*. The goal, then, is to estimate the proportion of different contributors to the sample [5, 14, 24, 71–73]).

We observe that the standard formulation of the problem leads to an inherent ambiguity. As a simple example, an ancient DNA proportion estimation tool may output that the DNA contained in a sample is 50% from *Homo sapiens* and 50% from Neanderthals. Such a result may be interpreted in many different ways, including the following: (1) Neanderthals and *Homo sapiens* lived simultaneously in a single location and both contributed equally to the sample; (2) the DNA sequences are ambiguous, leading to an inability of the model to determine the species of origin; (3) the data has originated from another organism that is equally distant to both Neanderthals and *Homo sapiens*; (4) some small percentage of the data is from *Homo sapiens*, the same small percentage originates from Neanderthals, and the rest is unidentified; or (5) a small number of sequences are strongly correlated to *Homo sapiens*, and there are many sequences that are weakly correlated to Neanderthals (or vice versa). Indeed, state-of-the-art solutions do not allow to easily differentiate between these alternative interpretations (see [75]). Beyond ambiguity, another concern stems from the risk of contamination of the ancient sample with present-day DNA, where some of the sequenced DNA reads can mistakenly be considered as originating from ancient *Homo sapiens*, but in fact come from people handling the sample during or after its excavation [34, 44, 54, 66].

These challenges call for the application of *explanation methods* to the proportion estimation computation. In recent work on explainable computation (and explainable AI in particular), two of the main lines of research focus on *attribution* and *counterfactuals*, which we briefly overview next. The notion of *attribution* is based on assigning values to different parts of the input, reflecting their individual contribution to the computation result. Concretely, the Shapley value [65], originating from game theory, has recently emerged as a measure of choice for attribution in many computational contexts (e.g., [3, 4, 17, 20, 39, 43, 45, 60]). *Counterfactuals* are typically defined as perturbations to the computation input that change its output. The intuition is that if small changes to some parts of the input are sufficient to change the result of a database query or flip the label assigned by an ML model, then these parts of the input are likely to have been crucial for the computation that took place. Counterfactuals complement attribution-based explanations: on one hand, the former allow to capture joint rather than only individual contributions of parts of the input; on the other hand, they typically do not allow to quantify or rank the contribution. See [13, 23, 35, 70] for examples of the extensive research on counterfactual explanations, in different contexts.

In this paper, we follow these lines of work and attach attribution and counterfactual explanations to a simple proportion estimation algorithm. We show that these explanations provide valuable information, that assists the researcher in resolving ambiguities and in identifying sequences potentially stemming from contamination (see the concrete use case example in 6.7). The attribution-based explanation assigns a vector of scores to each DNA sequence, quantifying its influence on the proportion assigned to each species. Continuing the example where the estimated proportions are 50% *Homo sapiens* and 50% Neanderthals, attribution scores reflecting that half of the sequences have a high contribution to the Neanderthal proportion and the other half present a high contribution to the *Homo sapiens* proportion would point the researcher to interpretation (1) out of the above list of interpretations. In contrast, low attribution scores across all sequences and for both labels would rather point to interpretation (2). In addition, if the sample was collected at an archaeological site from a time and place where the presence of *Homo sapiens* is unexpected, high contribution scores to the *Homo sapiens* proportion could lead to a suspicion that these sequences are contamination of the sample with modern-day DNA. For counterfactuals, we propose an intuitive formulation tailored to the problem at hand: a counterfactual explanation is a minimal-size set of DNA sequences whose exclusion from the dataset would change the estimated dominant species, namely the one assigned the highest proportion. When the estimated proportions suggest that the dataset is dominated by one species, but excluding a very small subset of sequences is sufficient to change dominant species, this indicates a lack of stability of the output.

The proportion estimation algorithm that we equip with explanations capabilities is based on the standard approach of *Maximum*

Likelihood, which we briefly overview next (see [38]). The input is an untagged dataset D, a set S of classes, and a reference set R in which each DNA sequence is associated with a single class $s \in S$. In a pre-processing step, the algorithm uses R to learn the conditional probability of each DNA sequence $d \in D$ given a class s, and generates a likelihood function. It then uses these conditional probabilities and likelihood function to predict the estimated classes' proportion in D: treated as probabilities, the chosen proportions are those that maximize the likelihood of observing the dataset.

In computing the explanations, we are inspired by existing methods, but adapt them to our settings, as follows. For attribution, we use the aforementioned notion of Shapley values, that have favorable theoretical properties as outlined in [45, 65]. Computationally, we compare two solutions for estimating Shapley values: one is to simulate the probabilistic process that underlies the definition of Shapley values (see [43]), in a Monte Carlo sampling algorithm. The second solution is based on using the SHAP package [45] as a black-box. Our experimental results show that while both algorithms converge after a similar number of iterations, SHAP is significantly faster, as the latter is able to converge in about 15 minutes whereas the Monte Carlo sampling approach fails to converge in a reasonable time.

Computing the actual Shapley values of individual DNA sequences is not always computationally feasible. In some use cases (see examples in 3 and 6.7), the researcher can derive insights even from explanations that fall short of estimating the exact Shapley values. Two possibly easier tasks are to *rank* the contribution of sequences, where ranking is based on the underlying Shapley values but do not necessarily involve computing them; and to *estimate the sign* of the Shapley values, namely to determine whether a given DNA sequence contributes positively or negatively with respect to the assigned proportion of a label of interest. In both cases, approximating the Shapley values with sufficient precision (which may depend on the actual hidden values) can allow ranking or sign estimation. We show that (1) a simple heuristic based on the parameters of the likelihood function can provide good approximation to the rank and sign extremely fast (*nDCG* score of 0.89 for ranking and a sign accuracy of 0.92); and (2) the SHAP-based solution quickly converges to good rank or sign accuracy (nDCG of 0.92 for ranking and sign accuracy of 0.93 after less than 4 minutes in our experiments).

For counterfactuals, we show a simple greedy algorithm that leverages the attribution ranking solutions mentioned above, and constructs subsets by gradually adding DNA sequences according to their estimated order of Shapley values. Our experiments, on small datasets for which ground truth counterfactual sets may be computed through an exhaustive search, indicate that using Shapley rankings as a basis for counterfactual explanations leads to computation of subsets that are almost as small as the ground truth subsets. For larger datasets where ground truth is not available, we show that the computation converges after about 1K samples of SHAP-based ranking.

The main contributions of our paper are **the adaptation and application of the well-established notions of attribution and counterfactual explanations to the context of proportion estimation in ancient DNA analysis**, as well as **the systematic experimental study of the computation of these explanations, on both synthetic and real datasets.** We use a previously published tool [71] to generate synthetic data mimicking ancient DNA sequences, as well as real ancient DNA data from samples collected at archaeological sites. To our knowledge, the computation of explanations for proportion estimation algorithms has not been studied before (beyond our demonstration paper [9]).

The rest of this paper is organized as follows. In Section 2 we overview the proportion estimation problem in ancient DNA analysis as well as a Maximum Likelihood algorithm for it. In Section 3 we overview the Shapley value and counterfactuals. In Sections 4 and 5 we present algorithms and heuristics for computing explanations based on Shapley values and counterfactuals, respectively. Section 6 includes the results of our experimental study, and a use case that exemplifies the explanation results and possible insights. Section 7 overviews related work and we conclude in Section 8.

2 PROPORTION ESTIMATION VIA MAXIMUM LIKELIHOOD

We recall here the task of proportion estimation in ancient DNA and the standard approach of using Maximum Likelihood for this task [12]. A DNA sequence is a string over the alphabet {A, C, G, T}. We are given an unlabeled dataset D (a set of unlabeled DNA sequences), and a labeled reference set R (a set of labeled sequences). Labels are elements of a set S (a set of biological species). The goal is to output a distribution function $V : S \mapsto [0, 1]$ (where $\sum_{s \in S} V(s) = 1$), capturing the estimated proportion of the different labels in D. When a total order over S is clear from context, we will represent V as a vector of length |S| whose entries represent the proportions of the respective species.

Example 2.1. Consider a dataset D of 100 unlabeled DNA sequences, collected from an archaeological site. The sequences are suspected to be of hominin origin and the label set is thus {*Homo sapiens*, *Neanderthal*, *Denisovan*}. The reference set R is a set of labeled sequences of different hominin individuals (typically, the references in R are full genomes of organisms, and the sequences in D are substantially shorter fragments). Our goal is to estimate what proportion of D has originated from each of the 3 hominin species. A result of the sort $(0.4, 0.5, 0.1)$ may be interpreted as estimated proportions of 40% *Homo sapiens*, 50% *Neanderthal* and to 10% *Denisovan* (this order will serve as the order of species throughout this paper).

A Maximum Likelihood Approach. A simple (yet effective, as we shall show below) approach for proportion estimation, proposed in [12], is based on the notion of *Maximum Likelihood*, as follows.

Let $\alpha : S \mapsto [0, 1]$ be a function capturing the prior probability of each species. Let $A_{d,s}$ be the probability of observing a sequence d given that its species is s. Assuming independence among data items, the likelihood of the observed sequences D is then:

$$L_R(D, \alpha) = \prod_{d \in D} \sum_{s \in S} \alpha_s \cdot A_{d,s} \tag{1}$$

The goal is then (a) to estimate the $A_{d,s}$ values and (b) given these values, to find α that maximizes $L_R(D, \alpha)$. Note that the likelihood function is further parameterized by the reference set R which in turn dictates the probability of sequences.

Algorithm. We next discuss both parts of the computation, which are schematically given in Algorithm 1. The input to the algorithm consists of the dataset D, the labeled references set R and the labels (species) set S. An additional input is a *substitution matrix* (such as BLOSUM [27] or PAM [18]), which is encoding the probability of a nucleotide i changing to nucleotide j over evolutionary time.

Computing individual probabilities. This step (implemented by the procedure *ComputeDataProb*) estimates the probability of observing a sequence d in an organism of species s. The procedure first aligns each DNA sequence $d \in D$ to each labeled DNA reference $r \in R$ (line 6 of Algorithm 1). In this context, "alignment" means searching for a sub-sequence r_d of r that is most similar to d. Then, using r_d and the substitution matrix M, the procedure calculates $P[r_d]$, namely the estimated likelihood of observing d in a genome of an organism that is closely related to an organism with genome r (line 7). In lines 8-9, for each species $s \in S$ and sequence $d \in D$, the procedure then estimates $A_{d,s}$ (the probability of observing the sequence d in an organism from species s), as the average of the $P[r_d]$ values for references labeled with s (the set of these references is denoted R_s).

Maximizing the Likelihood. The goal of this part is to find the parameters α that maximize the likelihood function $L_R(D, \alpha)$; these will be outputted as an estimation of the proportions (line 13). This step is performed using a variant of gradient descent [36, 62]: we initialize multiple arbitrary starting points (configurable in the solution and defaults to 5); then, in every iteration, we find the index for which a change is most beneficial for increasing the value of the likelihood function, and make a small change accordingly (configurable in the solution and defaults to 0.005).

Algorithm 1: Maximum Likelihood algorithm for proportion estimation in a genomic dataset

Data: Dataset D, label set S, labeled set of references R, substitution matrix M
Result: A distribution function α^* estimating proportions

1 A = `ComputeDataProb`(D, R, M)
2 **return** `MaximizeLikelihood`(A, D)
3 **Procedure** `ComputeDataProb`(D, R, M)
4 **foreach** $d \in D$ **do**
5 **foreach** $r \in R$ **do**
6 $r_d = align(d, r)$
7 $P[r_d] = \prod_{t=1}^{|d|} M_{(r_d[t], d[t], t)}$
8 **foreach** $s \in S$ **do**
9 $A_{d,s} = \frac{\sum_{r \in R_s} P[r_d]}{|R_s|}$
10 **return** all $A_{d,s}$ values
11 **Procedure** `MaximizeLikelihood`(A,D)
12 $L(D, \alpha) := \prod_{d \in D} \sum_{s \in S} \alpha_s \cdot A_{d,s}$
13 **return** $argmax_\alpha(L(D, \alpha))$ s.t. $\sum_i \alpha_i = 1, \forall i. \alpha_i \geq 0$

3 ATTRIBUTION AND COUNTERFACTUALS

We next recall two forms of explanations that are commonly used in the literature, namely *attribution* based on Shapley values and *counterfactuals*, and provide the concrete definitions that adapt these notions to our setting.

3.1 Attribution via Shapley Values

In a nutshell, attribution is a function that assigns a score to each data item, reflecting its contribution to some computation result. In recent years, a commonly used attribution approach (e.g., [10, 41, 42, 50, 61, 69]) is based on Shapley values [64], originating in Game Theory and defined as follows. Given a set N of players and a game value function v, the Shapley value of a player i, is defined as:

$$Shap(i, v, N) = \frac{1}{|N|} \sum_{G \subseteq N \setminus \{i\}} [v(G \cup \{i\}) - v(G)] \cdot |G|! \cdot (|N| - |G| - 1)! \quad (2)$$

To use Shapley values for attributions, the game value function is instantiated with some computation (e.g. ML classification in [45], a query in [20], image classification in [33], etc.) that one wishes to explain, and the players stand for different elements of the input (e.g. features, database tuples, medical information, pixels of an image). In our setting, the players are DNA sequences from the dataset D, and the value function is implemented via the Maximum Likelihood Algorithm (Algorithm 1). A subtlety here is that the algorithm involves randomization (in the choice of starting points for gradient descent), whereas a value function needs to be deterministic. A technical solution could be to fix the random seeds of the algorithm to some arbitrary value. We show below that, since the algorithm picks multiple starting points and takes the maximum over all obtained values, the choice of random seeds has only a very marginal effect on the algorithm's result in our setting, so that in practice, the algorithm behaves as a deterministic process.

Example 3.1. Continuing our running example, a vector of Shapley values for a sequence $d \in D$ may have the form $(-0.2, -0.1, 0.6)$. Intuitively, this means that d has negatively influenced the estimated proportion of *Homo sapiens* and (somewhat less so) of Neanderthals, and positively influenced the estimated proportion of Denisovans. This also informs the researcher that the specific sequence d in the mixed dataset is more likely to originate from a Denisovan than from either of the two other species.

Computational Problems. We will focus on the following computational problems with respect to attribution.

(1) (Approximate) *Computation* of Shapley values.
(2) *Ranking* of sequences with respect to their underlying Shapley values, and thus their perceived impact, without necessarily computing the values themselves.
(3) *Sign*: determining whether the Shapley value of each sequence with respect to a given label is positive or negative.

We next exemplify the usefulness of each variant.

Example 3.2 (Shapley values computation). As mentioned in the Introduction, the outputted proportions may be ambiguous and the Shapley values may be helpful in resolving ambiguities. For instance, if estimated proportions are similar across species, Shapley values of sequences can inform on whether similar percentages of the sequences support the proportion of each label. In another example, where one species is found to be dominant in the dataset, Shapley values can help determine whether this is caused by only a small amount of DNA sequences with a very strong influence (very high Shapley values), or by the domination of the species across

all data (similar Shapley values across all sequences). In practice, DNA sequences with outlying Shapley value with respect to some label may be subject to further investigation to test the possibility that these sequences are affected by damage, sequencing errors, or that they originate from contamination of present-day humans (see Introduction and 6.7).

We next exemplify how ranking and sign estimation may be useful in the absence of a high-quality estimation of the Shapley values themselves.

Example 3.3 (Shapley values ranking and sign). It may be of interest to isolate DNA sequences with high Shapley values for further inspection. The choice of which sequences to inspect further may be based on ranking, even in the absence of the actual values: for instance, given a "budget" of inspecting k sequences, one may choose the top-k with respect to a species of interest (also see [20] for a study of a similar ranking problem in the context of attribution in query evaluation, and [15] for a feature selection algorithm that only requires the order of the Shapley values).

The estimation of the sign of the Shapley value can also be a useful proxy for determining influence, as it reflects the direction in which a sequence affects the output with respect to each species. Thus, these signs can be utilized to filter the data and obtain subsets that influence the output in a similar manner. For instance, if a researcher is conducting a study on Neanderthal populations, they may want to filter out any DNA sequence from the dataset that is *not* a Neanderthal, which would generally correspond to sequences with a negative influence on the Neanderthal proportion. Another example could be of a sample collected from a time and place for which it would be extremely surprising to encounter any *Homo sapiens* DNA. In such a dataset, any sequence with a positive Shapley value with respect to *Homo sapiens* proportion could be suspected to originate from contamination by modern-day human DNA. These sequences can be isolated for further analysis in order to better assess whether they stem from contamination (for example, using deamination analysis - see [59, 66]).

Example 3.4. For a visual demonstration of the usage of Shapley values in the context of ancient DNA, we generated a dataset from genomes that their true label is known (see below on the dataset generation process), and hence we expect the Shapley value of a data point to align with the label of the genome from which it was generated. For example, if we generate a DNA sequence from a known genome of a Neanderthal, we expect its Shapley value to be positive with respect to the *Neanderthal* label, and negative with respect to other labels. We generated a dataset with 50 Neanderthal, 50 *Homo sapiens* and 50 Denisovan DNA sequences and estimated the Shapley values using a Shapley estimation algorithm the solution provides (see below). The output proportion estimation was, as expected, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Figure 1 displays the raw estimation of the Shapley values for the Neanderthal proportion. The estimations generally align with the label from which the data point was generated (marked in purple in the plot), and provide a visual insight on the characteristics of the input data D. For example, the researcher can observe that there is an abundance of sequences that contribute to increase the proportion assigned to the Neanderthal label in the first 50 sequences, from which one can deduce which sequences in D are the leading contributors to the proportions determined.

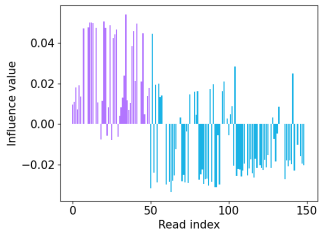


Figure 1: Estimated Shapley values with respect to Neanderthal proportion in a dataset comprising 150 sequences, of which the first 50 were generated from a Neanderthal genome and the last 100 from other hominins. Sequences that are expected to have positive Shapley values according to the label from which they were generated are marked in purple.

3.2 Counterfactuals

Given an algorithm A and an input x such that $A(x) = y$, a *counterfactual explanation* is a modified instance x', based on a typically small perturbation, for which $A(x') \neq y$. Intuitively, modifications to x that lead to a different output are indicative of the parts of the data that were most significant in the original computation. The above definition is commonly used in both Machine Learning models where the output is a label, and in database queries where the output is a relation (e.g., [51], [47]). For a numeric algorithm such as Algorithm 1, any arbitrary change to the input is expected to change the output. We next introduce a refined notion of counterfactuals that takes into account our particular use case.

Definition 3.5. Given a dataset D and a proportion estimation algorithm A, let $H(D, A)$ be the index of the maximal value in the output vector of A when executed on D (for some set R of references; in case of ties in the outputted proportions, $H(D, A)$ is the smallest index attaining the maximal value). We say that a subset $D' \subseteq D$ is a *dominance-changing subset* if $H(D, A) \neq H(D \backslash D', A)$.

Last, we define $CF(D, A)$ as a subset $D' \subseteq D$ that is a dominance-changing subset, and is also the minimum-size subset [1] with this property. Namely, for every subset $D'' \subseteq D$ that is dominance-changing with respect to A and D, it holds that $|D'| \leq |D''|$.

Intuitively, $H(D, A)$ is the label to which A assigns the highest proportion when executed on D. Then, a dominance-changing subset is a subset D' of D whose removal changes the label assigned the highest proportion.

Example 3.6. In our running example, if the output of a proportion estimation algorithm A for a given dataset D is (0.6,0.4,0), then $H(D, A) = 0$ (the first index includes the maximal value). If there are two DNA sequences $d_1, d_2 \in D$ whose removal from D would cause A to output the vector $(0, 0, 1)$ instead of $(0.6, 0.4, 0)$, then the set $\{d_1, d_2\}$ is a dominance changing subset with respect to A and D (changed the dominant label from *Homo sapiens* to Denisovan). If this set is the smallest set with this property, then the counterfactual explanation would be the set $\{d_1, d_2\}$.

This flavor of counterfactual explanations is useful both to identify high-impact sequences as well as to assess the output stability:

[1] An alternative definition would focus on minimal subsets that are not necessarily of minimum size, yet the size of the set is of crucial interest in the settings of ancient DNA, as if we encounter a small subset that is sufficient to change the dominance in the dataset, it reflects on the abundance of the original dominant label in the dataset.

if the minimal-size dominance changing set is large, then the algorithm is intuitively stable with respect to errors in the data (a significant modification needs to occur in the data to change the estimation), and otherwise it is unstable. To exemplify this idea, we generated 2 datasets and ran the counterfactual algorithm our solution provides (see below) on them: D_1 consists of 20 sequences generated from a *Homo sapiens*, and of 40 non-informative sequences (sequences from regions in the genome that are identical across all hominin species), and D_2 consists of 40 sequences from *Homo sapiens* and 20 non-informative sequences. Though the proportion estimation is the same for both datasets ($(1, 0, 0)$, i.e. full dominance of *Homo sapiens*, as expected), the counterfactual explanation for D_1 is of size 11, whereas for D_2 it is of size 30. This conveys to the researcher that the output on D_2 is more stable while also showing the subsets of data that could cause changes.

4 COMPUTATION OF SHAPLEY VALUES

Due to the exponential nature of the Shapley value definition, a common practice is to approximate the values rather than attempt an exact computation (see [1, 10, 19, 22, 33, 45, 49]). We follow this line of work and show inexact solutions that either approximate the Shapley values or allow to rank or estimate the sign of DNA sequences based on their underlying Shapley values, without actually computing them. We start by providing two estimation methods: one that is based on direct Monte Carlo sampling, and another that is adapting and using the SHAP tool, commonly used for Shapley values computation in the context of explaining Machine Learning predictions.

4.1 Approximation via Monte Carlo Sampling

Shapley values have a natural interpretation as a stochastic process (see [42] for a detailed exposition): pick at random an order over the players and have players join the game according to that order. Upon joining, each player is given a reward which is computed as their marginal contribution, with respect to the current subset of players, to the game value function. The Shapley value of each player is their expected payoff in this stochastic process.

Our solution provides a Monte Carlo algorithm that directly samples from this distribution, and works as follows. As a preprocessing step, it runs the `ComputeDataProb` function of Algorithm 1 on the reference set R and the full dataset D, to obtain the $A_{d,s}$ values. In each sample, we then obtain subsets of D by repeatedly adding DNA sequences (the order of their choice is uniformly drawn) and observing the marginal effect of each added sequence on the result of executing the `MaximizeLikelihood` function of Algorithm 1. These effects are averaged over all samples to output the approximated Shapley value for each data point (recall that the result of `MaximizeLikelihood` is a vector of length $|S|$; the above computation is performed separately for each coordinate).

4.2 Using SHAP

As we will show below, though the Monte Carlo method is commonly used and directly samples from the Shapley distribution, it is in practice often too slow. We next describe an alternative algorithm based on SHAP [46], a commonly used tool for Shapley value computation in Machine Learning.

Adapting SHAP to our settings requires several modifications. In particular, recall that the definition of Shapley values involves computation of the game value function – namely, execution of the algorithm that is explained – on multiple subsets of its input set. In the context of Machine Learning models, running the model on a subset of its features is infeasible, and the implementation of SHAP is instead using a background dataset, to sample values for features it leaves out at each step (see [78] for implications of the size of the background dataset). In our setting, we do not use a background dataset but can rather use SHAP in a manner that directly executes Maximum Likelihood on a subset. Concretely, we rewrite the formula in line 12 of Algorithm 1, so that we plug-in a subset G instead of the entire dataset D, and thus calculate what would be the output of the algorithm on G. Thus, when we consider multiple subsets, we only need to execute the (computationally costly) `ComputeDataProb` step once, and reuse the $A_{d,s}$ values that it computes in multiple invocations of `MaximizeLikelihood` (A, G). This is possible since the values of $A_{d,s}$ are independent of the current sample that the algorithm is being executed upon.

4.3 Ranking and Sign Estimation

We next propose a simple and (as we shall experimentally show) effective heuristic solution for estimating the order of Shapley values, without computing the actual values. The solution uses the $A_{d,s}$ values of the likelihood function as a proxy for the Shapley value of d w.r.t. s, namely instead of ranking based on the (unknown) Shapley values, it ranks based on $A_{d,s}$. Similarly, to estimate the sign of the Shapley value of a data point $d \in D$ with respect to a label $s \in S$, we output the sign of $A_{d,s} - AVG_{i \in S}(A_{d,i})$ (if the value is greater than the average, we output "positive", otherwise we output "negative").

This solution is based on the following observation (details deferred to the full version in [8] for lack of space): for $S = \{s_1, s_2\}$ and $A_{d,s_1} > A_{d,s_2}$ for some sequence $d \in D$, then the Shapley value of d with respect to s_1 is positive, and with respect to s_2 is negative. While a generalization to cases where $| S | > 2$ does not necessarily hold (a counterexample appears in [8]), there is an intuitive correlation between the relative magnitudes of $A_{d,s}$ and the relative magnitudes of Shapley values. Indeed, we will experimentally show that the simple algorithms for estimating the order and sign of the Shapley values using the $A_{d,s}$ values generally achieve good ranking and sign accuracy.

5 COMPUTATION OF COUNTERFACTUALS

For counterfactual explanations, we propose a simple (yet novel, to our knowledge) greedy algorithm whose details appear in Algorithm 2. The idea is to leverage attribution values, which are indicators of the individual contribution of sequences in D, to guide the search for a minimal dominance-changing subset. The algorithm consequently uses a *ranking oracle*, that ranks the data items with respect to their contribution to the dominant label. The ranking oracle may be implemented either based on estimating the actual Shapley values or via $A_{d,s}$ values (see previous section). The algorithm starts with the full set D and gradually removes sequences in descending order with respect to the provided ranking. After every removal of a sequence, the algorithm re-evaluates the proportion

estimation algorithm A on the remaining sequences, and checks if $H(D, A)$ (i.e., the dominant label) has changed. It halts upon finding a subset whose removal changes the label, outputting it as a dominance-changing subset.

Algorithm 2: Algorithm to estimate counterfactuals

Data: Dataset D, dominant label l, ranking oracle O, proportion estimation algorithm A
Result: A subset of D

Use O to compute an ordered list $[d_1, ..., d_n]$ with respect to contribution to l
$D' = \emptyset, i = 0$
while $H(D \backslash D', A) == l \ \ \& \ \ i < |D|$ **do**
    $D' = D' \cup d_i$
    $i = i + 1$
return D'

Note that, when A is Algorithm 1 we can again leverage the optimization allowing us to execute `ComputeDataProb` only once, and then execute the `MaximizeLikelihood` at most $|D|$ times .

Example 5.1. As an example, if on a dataset D the output of a proportion estimation algorithm A is (0.6,0.4,0), then the dominant label outputted by the algorithm is *Homo sapiens*. The counterfactual algorithm first uses one of the algorithms mentioned in Section 4 as a ranking oracle, to attain an estimation on the order of data points in D, with respect to their Shapley value corresponding to the *Homo sapiens* label. Then the algorithm starts to greedily remove data points that have the highest influence on the *Homo sapiens* proportion. After removing D_1 for example, the algorithm reevaluates the explained algorithm, to check what is the dominant proportion on $D \backslash \{D_1\}$, and so on, until the output of A has another dominant label. For example, it might be the case that after removing D_1, D_{54}, D_{98}, we will run A on D without them and get the output (0.2,0.6,0.2). This will cause the algorithm to output the set $\{D_1, D_{54}, D_{98}\}$ as an estimation to the minimal dominance-changing subset. If the algorithm does not find a dataset that changes the dominant label, it will eventually remove all the data points from D and return D itself as the minimal dominance-changing subset.

6 EXPERIMENTAL STUDY

We implemented all algorithms detailed in this paper in python, using the following tools: *joblib* for parallel computation, *SHAP* for attribution explanations and visualizations, *Jupyter Widgets* for a GUI that interacts with the solution (see our demonstration paper [9]), *biopython* for alignments and parsing of DNA data, downloaded reference genomes from the *NCBI database* ([63, 76]), and simulated ancient DNA data using a script published in [71]. We made the code publicly available, including the GUI, execution examples, and usage instructions, in a GitHub repository [8]. Experiments were executed on a personal computer with the following specs: Windows 11, 11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz, 2995 Mhz, 4 Cores, 8 Logical Processors and 32 GB RAM.

6.1 Experimental Goals and Methodology

Our experiments aim at analyzing the performance of the solutions described throughout this paper. We in particular study the following:

- *Quality and stability of the proportion estimation algorithm*: we examine the performance of the simple proportion estimation algorithm that we use (Algorithm 1), compared to a state-of-the-art solution, over datasets for which the ground truth is known (see Section 6.2). Quality is measured using the standard measure of KL-divergence (see [2, 40]). We also measure the stability of the algorithm, namely to what extent do multiple runs with different random seeds return consistent results.
- *Quality of attribution-based explanations*: we measure different facets of quality, in terms of accuracy of the outputted Shapley values ($L1$ and maximal observed error), of produced ranking (Normalized Discounted Cumulative Gain [31, 32, 74]), and of the sign estimation (accuracy). These measures are based on the ground truth, for small datasets where exact Shapley values may be computed using a brute-force algorithm. For large datasets, where the ground truth is unattainable, we employ an intrinsic measurement, where we compare to the "estimated ground truth", attained by running SHAP for a very large number of iterations.
- *Quality of counterfactual-based explanations*: we compare the size of computed subsets to the ground truth minimal dominance-changing subset, for small datasets where the ground truth can be obtained via brute-force computation. For larger datasets, where the ground truth is out of reach, we study the correlation between the quality of the ranking oracle (specifically, the number of samples given as input to SHAP), and the quality of obtained counterfactuals. While this is not a direct measure of quality, a high correlation would indicate the usefulness of the approach. We show the results for all proposed choices of ranking oracles.
- *Execution times* of all algorithms mentioned in this paper.
- *Use case* that shows a running example of the explanations and of the insights that they facilitate.

Based on the experimental results, we establish guidelines and recommendations to researchers on which solutions are most suitable in different use cases.

6.2 Datasets

Unlike the case in many other data analytics settings, typical datasets studied in ancient sedimentary DNA are small in size. This is because ancient sedimentary samples often allow the retrieval of only a small number of DNA fragments. For example, in [67], 5 out of 15 samples containing ancient hominin mtDNA could not be confidently assigned to a hominin group of origin. On average, the unidentifiable datasets contained fewer sequences (231 DNA sequences, range: 53-629) than the identifiable ones (average 612 sequences, range: 64-1,487). A similar problem arose in [71], where again on average, datasets that contained traces of ancient hominin mtDNA and yet could not be attributed confidently to a specific hominin group of origin were smaller than the identifiable ones (average of 238 DNA sequences [range: 46-1,088] versus 2,938 sequences [range: 73-90,751], respectively) [2]. In fact, the typically

[2]While some of these ambiguities could in principle be resolved by generating additional sequencing data, such a course of action requires both the availability of leftover

small size of these datasets is a main source of challenge for analysis in general and for proportion estimation in particular. This is because the probability of encountering a genomic region that differentiates between species drops as the number of available DNA sequences decreases (especially when the species are close in the evolutionary tree). Computationally, the exponential nature of the problems that we study here (recall that the definitions of both Shapley values and Counterfactuals involve enumeration over all subsets) renders them highly challenging even for small datasets.

Our experiments are executed over the following datasets:

- **SYNTHETIC**: Simulated datasets were generated as follows. We used the ancient dataset generator python tool published in [71] (`ChunkGenome.py` script, with the default substitution matrix provided in the package, deamination pattern of 3, minimal sequence size of 35, maximal sequence size of 100, and all other properties set at the default values of the package) to generate 5000 synthetic ancient DNA sequences of length 35-100 from each of the following mitochondrial genomes freely-available online (downloaded from the NCBI database [63, 76]): 1) *Homo sapiens*: "Chinese" AF346973; 2) Neanderthal: "Goyet" KX198085; and 3) Denisovan: "Denisova 2" KX663333. Then, to generate a `SYNTHETIC` dataset of size X with proportions (a, b, c), we randomly sampled $a \cdot X$ sequences from the generated *Homo sapiens* sequences, $b \cdot X$ Neanderthal sequences and $c \cdot X$ Denisovan sequences. Note that simulated data was generated based on genomes that are not in the reference set R fed to the algorithm. For `SYNTHETIC` datasets we have the ground truth proportions (based on the label of the genome from which each sequence was generated), but generally not the ground truth Shapley values/counterfactuals since the exponential computation time is out of reach. For some of the experiments where we still want to compare with the ground truth explanations, we used small synthetic datasets for which brute-force exhaustive search for ground truth explanations was feasible. Hereafter, we use `SYNTHETIC-X-Y` to denote a dataset that was generated using the `SYNTHETIC` method and includes X genomic datasets of size Y. For example, `SYNTHETIC-3-12` is composed of 3 `SYNTHETIC` datasets, with 12 DNA sequences in each dataset. Additionally, `SYNTHETIC-X-[Y,Z]` stands for X `SYNTHETIC` subsets of sizes in the range $[Y, Z]$.
- **REAL**: These datasets consist of DNA sequences that were sequenced from ancient samples collected in archaeological sites. To mimic ancient sedimentary DNA, which often contains DNA from multiple individuals, we mixed DNA sequences from two different ancient DNA libraries, in equal proportions, as follows: One `REAL` dataset (denoted `REAL 1`) contains 50 mitochondrial DNA sequences from an ancient *Homo sapiens* tooth, the "Fumane 2" individual (library L5184/L5202 from [7]) and 50 sequences from a Neanderthal bone, the "Hohlenstein-Stadel" individual (library L5291 from [55]). The second `REAL` dataset (denoted `REAL 2`) contains 50 sequences from an ancient *Homo sapiens* tooth, the "Fumane 2" individual (as described above), and 50 sequences from a sediment sample collected at Denisova Cave, which was found to be rich in Neanderthal mitochondrial DNA (library D5276 from [67]). Note that as these are real datasets, they are affected by all issues common to ancient DNA data, namely, they contain short and damaged sequences, affected by contamination with present-day *Homo sapiens* DNA. To generate a `REAL` dataset of size N, we randomly selected a dataset of the two mentioned above, and generated a random subset of size N of that dataset. Then, `REAL-X-Y` is a dataset composed of X `REAL` datasets, each including Y DNA sequences. Each `REAL` dataset contains 100 DNA sequences, hence `REAL-2-100` is the full dataset.

Both the `REAL` datasets and the script to generate the `SYNTHETIC` datasets from references are available alongside the code in [8].

material and the investment of further time, effort and resources, and is thus not always feasible.

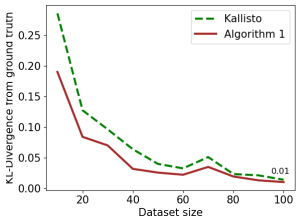


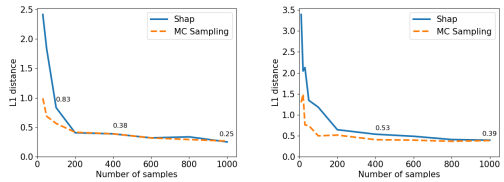
Figure 2: Comparison of Algorithm 1 and Kallisto on `SYNTHETIC-500-[10-100]`

6.3 Quality of the Maximum Likelihood Algorithm

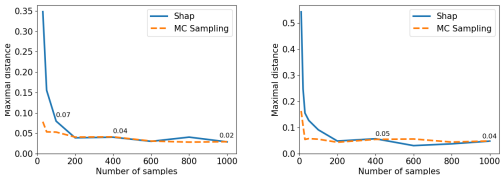
We first measured the accuracy of the proportion estimation algorithm (Algorithm 1), compared to Kallisto ([12]). The latter is a program originally developed for quantifying abundances in RNA sequencing data, which has recently been used as state-of-the-art in classification of ancient sedimentary DNA datasets ([71]). The comparison was performed based on `SYNTHETIC` datasets for which, as explained above, we can sample different datasets with desired ground truth proportions. The accuracy metric is the KL-Divergence ([2, 40]) between the output vector of each algorithm and the ground truth proportion vector. The measurements are made for `SYNTHETIC-500-[10,100]`, where for each dataset we computed the KL-divergence between the output vectors and the ground truth. Both algorithms were executed using the default substitution matrix published in [71], and a reference set of 18 mitochondrial genomes - 6 Neanderthals, 6 *Homo sapiens*, and 6 Denisovans (all different from the references that were used to generate the data).

Figure 2 depicts the results, showing that for datasets of size up to 100, Algorithm 1 consistently performs similarly or better than the Kallisto method (as reflected in similar or lower KL divergence from the ground truth). For larger datasets, the task becomes easier for both algorithms, both of which attain almost perfect accuracy (KL-Divergence of 0.01 when running on datasets of size 100). As we shall report below, this accuracy comes at a computational cost: Kallisto is substantially faster than Algorithm 1.

Stability of the algorithm. We further examined the effect of randomness on the Algorithm 1, to show that in practice it behaves as a deterministic process. We generated 50 datasets (comprised of `REAL-25-30` and `SYNTHETIC-25-30`), and executed the maximization part of Algorithm 1 100 times on each dataset with different

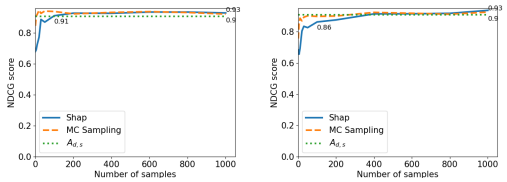


(a) SYNTHETIC-3-12 (b) REAL-3-12



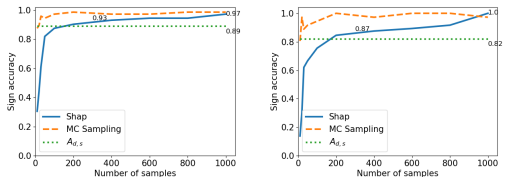
(c) SYNTHETIC-3-12 (d) REAL-3-12

Figure 3: Comparison to ground truth using L1 (top figures) and maximal distance (bottom).



(a) SYNTHETIC-3-12 (b) REAL-3-12

Figure 4: Ranking score (nDCG) for small datasets



(a) SYNTHETIC-3-12 (b) REAL-3-12

Figure 5: Sign accuracy for small datasets

random seeds in each execution (the maximization part is the non-deterministic part of the algorithm). For each dataset, we compared the 100 proportion estimation vectors that were generated, and computed the $L1$ distance between any pair. The maximal distance between any two outputs was 0.01, the 50th percentile was 10^{-16}, the 95th percentile was 10^{-15} and the average was 0.0004.

6.4 Quality of Attribution-based Explanations

We next describe experiments that measure the quality of the proposed algorithms for Shapley value computation.

Accuracy of value estimation with respect to the ground truth. We start by comparing the accuracy of the Monte Carlo (MC) sampling and of the SHAP-based algorithm for SYNTHETIC-3-12 and REAL-3-12. We computed ground truth Shapley values for these datasets using the exponential time algorithm that directly uses the Shapley value formula. Our measures are (1) $L1$ distance and (2) the maximal observed difference over all pairs of data points and labels, both as a function of the sample size that the algorithm uses (for sample sizes [10, 20, 30, 50, 100, 200, 400, 600, 800, 1000]). For MC, the sample size is the number of random permutations; while for SHAP, it is the parameter *nsamples* given as input to SHAP.

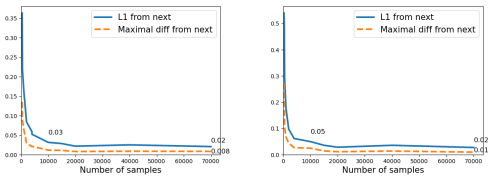
Figure 3 shows that both algorithms converge to good accuracy on both datasets (maximal distance of 0.04 and 0.05 and L1 of 0.38 and 0.53 for SYNTHETIC and REAL, respectively) after 400 samples, following which they exhibit a steady yet slower improvement. While both the MC sampling and the SHAP-based algorithm behave similarly in terms of accuracy as a function of the number of samples, we will show below that MC is significantly slower in terms of execution time per sample.

Ranking Accuracy with respect to ground truth. We further measured the accuracy of the *ranking* on SYNTHETIC-3-12 and REAL-3-12, as produced by the MC sampling and SHAP-based algorithm, as well as the simple ranking heuristics based on $A_{d,s}$ values described in 4.3. For each label, we compared the ranking of sequences according to their contribution as estimated by each of the three algorithms (for sample sizes [1, 5, 10, 20, 30, 50, 100, 200, 400, 600, 800, 1000]), to the ranking based on the ground truth values. Our accuracy metric measurement is the Normalized Discounted Cumulative Gain (nDCG, a standard measure for comparing ranked lists, see [31, 32, 74]), as follows. We define the gain of a ranking at position i as $gain(i) := \frac{|D|-|rank(i)-desired(i)|}{log_2(i+1)}$ (when $|D|$ is the number of elements in the ranking, in this case, 12) and then $DCG = \sum_{i \in D} gain(i)$, and $nDCG = \frac{DCG}{iDCG}$ where $iDCG$ is the gain assigned to the optimal (ideal) ranking. The score we assign to an algorithm is the average nDCG score of the algorithm across labels.

As shown in figure 4, the simple $A_{d,s}$-based heuristic achieves an $nDCG$ score of 0.9 on both REAL and SYNTHETIC datasets, and both the MC and SHAP-based algorithms attain a ranking score higher than 0.9 after about 100 samples for the SYNTHETIC datasets and about 400 samples for the REAL datasets.

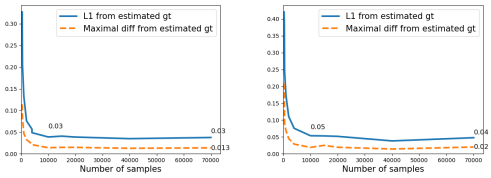
Sign accuracy with respect to ground truth. We next measure the "sign accuracy" for the SYNTHETIC-3-12 and REAL-3-12 datasets. For each sample size in [1, 5, 10, 20, 30, 50, 100, 200, 400, 600, 800, 1000] and for each sign estimation algorithm (SHAP, MC, $A_{d,s}$), we counted the percentage of values in which the sign of the ground truth Shapley value agrees with the sign of the output estimation. Figure 5 shows that both SHAP and MC achieve high sign accuracy of 0.9 already after 200 samples for SYNTHETIC and 0.87 after 400 samples for REAL. The simple heuristic based on $A_{d,s}$ values attains sign accuracy of 0.89 on SYNTHETIC and 0.82 on REAL.

Convergence and estimated accuracy of computed Shapley values for larger datasets. For larger datasets, the MC method is too slow to execute with reasonably sufficient number of samples (as we show in the execution time experiments below). Hence, we focus hereafter on the SHAP and $A_{d,s}$ methods. For SYNTHETIC-X-100 and REAL-X-100 datasets, obtaining the ground truth Shapley values is beyond reach. Instead, we first test the intrinsic behaviour of the SHAP-based algorithm: we ran the algorithm with increasing number of samples and measured the change in the output matrix of Shapley values between each sample size to the one that follows (the examined sample sizes were [200, 400, 500, 1000, 2000, 4000,



(a) SYNTHETIC-3-100

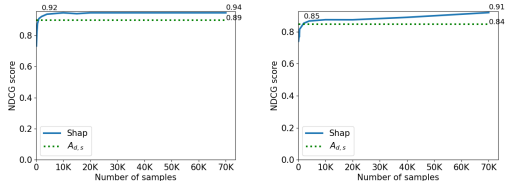
(b) REAL-2-100



(c) SYNTHETIC-3-100

(d) REAL-2-100

Figure 6: Comparison to next sample size (top figures) and to estimated ground truth (bottom)



(a) SYNTHETIC-3-100

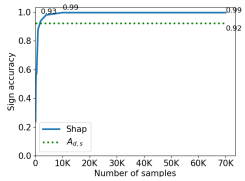
(b) REAL-2-100

Figure 7: Ranking score compared to estimated ground truth

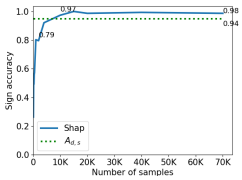
10000, 15000, 20000, 40000, 70000, 100000]). Figures 6a and 6b show the results on SYNTHETIC-3-100 and REAL-2-100, respectively. After 10000 samples, SHAP converges: changes are less than 0.03 and 0.05 in $L1$ and less than 0.01 in terms of maximal distance for SYNTHETIC-3-100 and REAL-2-100, respectively.

We have further conducted an experiment where we used the result of the SHAP-based algorithm with 100K samples as the estimated ground truth, and compared the $L1$ and maximal distance from this estimated ground truth, for each of the results obtained for smaller number of samples. Figures 6c and 6d show the results on SYNTHETIC-3-100 and REAL-2-100, respectively. Again after 10000 samples, SHAP converges to $L1$ distance of 0.03 for SYNTHETIC-3-100 (and 0.05 for REAL-2-100) from the execution result of SHAP with 100K samples, and to maximal distance ≤ 0.02 for SYNTHETIC-3-100 (and 0.02 for REAL-2-100).

Ranking accuracy for larger datasets. We again use SHAP with 100K samples as an estimation for the ground truth, and compare the rankings obtained by the SHAP-based algorithm with fewer samples (sample sizes are [100, 200, 300, 400, 500, 1K, 2K, 4K, 10K, 15K, 20K, 40K, 70K]) and by the $A_{d,s}$-based algorithm to the ranking that is based on the estimated ground truth. Figures 7a and 7b show the average nDCG score over SYNTHETIC-3-100 and REAL-2-100, respectively. The figure shows that the $A_{d,s}$-based algorithm attains a ranking score of 0.89 on SYNTHETIC-3-100 and 0.84 on REAL-2-100 and that the SHAP algorithm with 2K samples already achieves a ranking score of 0.92 on SYNTHETIC and 0.85 on REAL.



(a) SYNTHETIC-3-100



(b) REAL-2-100

Figure 8: Sign accuracy compared to estimated ground truth

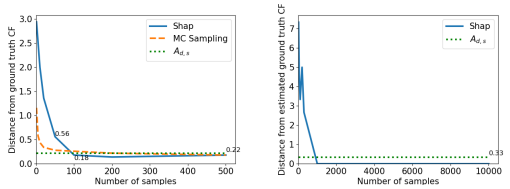
Sign accuracy for larger datasets. We used, as before, the result of the SHAP-based algorithm with 100K samples as an estimated ground truth, and calculated the sign accuracy with respect to these values. Sample size for SHAP are [100, 200, 300, 400, 500, 1K, 2K, 4K, 10K, 15K, 20K, 40K, 70K]. Figures 8a and 8b show that the $A_{d,s}$-based algorithm already attains sign accuracy of 0.92 on SYNTHETIC-3-100 and 0.94 on REAL-2-100, while SHAP achieves sign accuracy of 0.93 after 2K samples and 0.99 after 10K samples for SYNTHETIC, and an accuracy of 0.79 after 2K samples and 0.97 after 10K samples for REAL.

6.5 Quality of Counterfactual Explanation

To empirically test the accuracy of the counterfactual algorithm (Algorithm 2), we assembled a SYNTHETIC-100-[10,15] dataset, with different proportions of *Homo sapiens*, Neanderthal, and Denisovan sequences. We calculated the ground truth counterfactual (minimal dominance-changing subset) through an exhaustive brute-force search over all subsets.

We then executed Algorithm 2 using each of the three options for the ranking oracle (Shapley value estimations from section 4), and compared the output to the ground truth counterfactuals, using different sample sizes ([1, 5, 10, 20, 50, 100, 200, 500]). Figure 9a shows the average distance from the ground truth (defined as $|CF| - |GT|$, when CF is the subset that the algorithm outputs, and GT is the ground truth subset), averaged over 100 runs, as a function of the number of samples used by the ranking oracle. The results show that on these small datasets, all ranking oracles converge to an average distance of approximately 0.22 from the ground truth counterfactual. The sampling-based oracles yield this accuracy already when using 100 samples.

For larger datasets, the ground truth counterfactual subset is unattainable. Thus, we generated a SYNTHETIC-3-100 dataset, with proportions of $(0.5, 0.4, 0.1)$, to control the approximate size of the counterfactual set (hence, we did not use REAL here, for which we cannot control the proportions). We ran the counterfactual algorithm, using SHAP and $A_{d,s}$ as ranking oracles (with sample sizes [20, 50, 100, 200, 300, 1000, 2000, 5000, 10000]), and compared the results to the best (smallest) attained subset in each iteration. Figure 9b shows the average distance from the best attained counterfactual over the execution. Note that in all the executions, using SHAP with ≥ 1000 samples has already yielded the best attained subset. In two out of the three executions, using the $A_{d,s}$-based ranking was optimal as well. In the case where using the $A_{d,s}$-based ranking was non-optimal, the size of the outputted set was only larger by 1 than that of the optimal set (hence the average distance of 0.33 for $A_{d,s}$).



(a) SYNTHETIC-100-[10,15] (b) SYNTHETIC-2-100

Figure 9: 9a shows average distance from ground truth counterfactual as a function of the number of samples on SYNTHETIC-100-[10,15], and 9b shows the average distance from the best attained counterfactual on SYNTHETIC-2-100 dataset. The curves correspond to different ranking oracles used by the algorithm.

As expected, as the number of samples for SHAP increases, the counterfactual algorithm is able to find better (i.e. smaller) subsets.

6.6 Execution Times

We will now report on the results of the execution time experiments we conducted on the algorithms mentioned in this paper. The results for the two datasets (SYNTHETIC and REAL) were very similar. We thus report results only for a single dataset: for attribution, we report the results for the REAL dataset, while for counterfactuals we report the results for the SYNTHETIC dataset so that we can control the ground truth proportions.

Figure 10a shows the execution time of the proportion estimation algorithm on SYNTHETIC-5-[10,100] as a function of the dataset size. Compared to the Kallisto method, the algorithm is substantially slower (around 4 minutes for a dataset of size 100, where Kallisto executes in less than 3 seconds). Since the process of ancient sedimentary DNA analysis can be quite lengthy (see [67, 71]), proportion estimation may anyway be executed offline.

Figure 10b shows the execution time of the SHAP and MC algorithms on REAL-5-12. Note that even on a dataset of size 12, the execution time of the MC method approaches 9 minutes for 1000 samples.

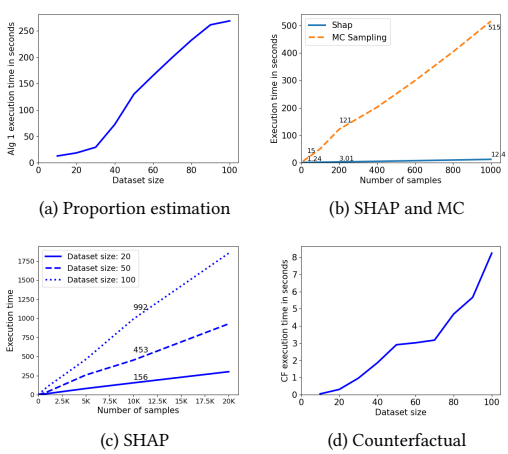
Figure 10c shows the execution time of SHAP on REAL-5-20, REAL-5-50 and REAL-2-100. Note that for the same number of samples, the execution time is higher when the dataset is larger (as SHAP might sample larger subsets to evaluate the algorithm upon). The execution time of SHAP for 10K samples on a dataset of size 100 is approximately 16 minutes.

Figure 10d shows the execution time of the counterfactual algorithm on SYNTHETIC-5-[10,100]. The figure displays the overhead of the algorithm, not including the execution time of the ranking oracle, which has already been studied above. The figure shows that given a Shapley values ranking estimation, the overhead of generating counterfactual explanations is only a few seconds.

6.7 Use Case

We now exemplify the usefulness and actionability of the explanations with a full running example on the two REAL datasets, including insights that can be obtained from the generated explanations ([8] hosts a Jupyter Notebook for reproducing these results).

We first ran Algorithm 1 on REAL 1 and REAL 2, and got the output of (0.8,0.175,0.025) and (0.625,0.375,0), respectively, consistent with



(a) Proportion estimation (b) SHAP and MC

(c) SHAP (d) Counterfactual

Figure 10: 10a shows execution time of Algorithm 1 on SYNTHETIC-5-[10-100] (the Kallisto method terminates in less than 3 seconds for a dataset of size 100 and is thus omitted from the plot). 10b shows execution time of the attribution explanation algorithms on REAL-5-12 (the $A_{d,s}$-based heuristics terminates in split seconds and is thus omitted from the plot). 10c shows execution time of the SHAP-based explanation algorithm on REAL-5-20, REAL-5-50, REAL-2-100. 10d shows execution time of the counterfactual algorithm (not including the ranking oracle execution time) on SYNTHETIC-5-[10,100].

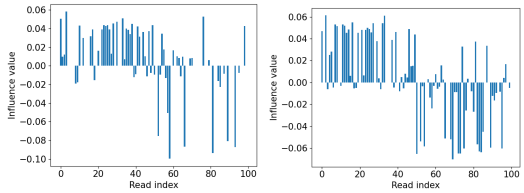
the contribution from both *Homo sapiens* and Neanderthals in both datasets.

To help interpret the output proportions, we executed the SHAP attribution explanation with 10K samples for each dataset (execution time - 14 minutes for each dataset). Figure 11 shows the raw outputted Shapley value estimations. As expected, almost all sequences have a small absolute value contribution to the proportion of Denisovans. Furthermore, in both datasets, the first 50 sequences tend to contribute positively to the *Homo sapiens* proportion, and negatively to the Neanderthal proportion; while the latter 50 sequences tend to contribute positively to the Neanderthal proportion and negatively to the *Homo sapiens* proportion.

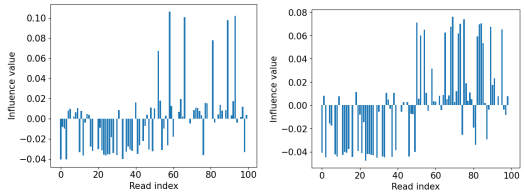
We then executed the counterfactual algorithm (Algorithm 2) using the SHAP output as the ranking oracle (execution time was approximately 10 seconds for each dataset, after already having the SHAP values). The counterfactual estimation for the minimal dominance changing subset of dataset REAL 1 was that to change the dominance species from *Homo sapiens* to Neanderthal, we will need to remove 20 sequences: [0, 3, 10, 17, 22, 23, 24, 25, 26, 28, 30, 33, 35, 36, 38, 41, 47, 49, 76, 98]; while on dataset REAL 2, only 10 sequences would suffice: [2, 7, 8, 11, 12, 16, 24, 25, 27, 32, 33]. This conveys to the researcher that the output on REAL 1 is more stable, as a larger change is required for it to change.

We then analyzed each library separately.

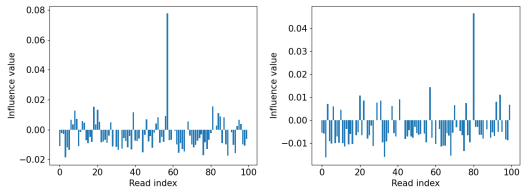
- On a dataset that consists of the sequences from the *Homo sapiens* library (same in both datasets), the output of Algorithm 1 was (1,0,0).
- On a dataset that consists of the Neanderthal sequences from REAL 1, the output of Algorithm 1 was (0.275, 0.675, 0.05).
- On a dataset that consists of the Neanderthal sequences from REAL 2 the output of Algorithm 1 was (0,1,0).



(a) SHAP *Homo sapiens* REAL 1 (b) SHAP *Homo sapiens* REAL 2



(c) SHAP Neanderthals REAL 1 (d) SHAP Neanderthals REAL 2



(e) SHAP Denisovans REAL 1 (f) SHAP Denisovans REAL 2

Figure 11: Raw SHAP values for REAL 1 (left) and REAL 2 (right)

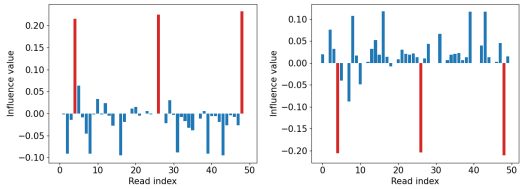
To further analyze the Neanderthal sample from REAL 1 (which has an unexpected 0.275 contribution from *Homo sapiens*), we executed the Shapley estimation using SHAP with 10K samples (execution time - 7 minutes). Our suspicion is that the 0.275 contribution of *Homo sapiens* sequences to this sample is due to contamination by present-day *Homo sapiens* DNA. Figure 12 shows the results, and displays that there are three sequences in the dataset that are outliers in their positive contribution to the *Homo sapiens* proportion and negative contribution to the Neanderthal proportion (the indexes of these sequences are 4, 26, 48).

As we suspected that these three sequences are due to contamination, we filtered them out of the dataset, and executed Algorithm 1 again. This time, the output was (0, 0.95, 0.05), i.e., as expected, these three sequences were responsible for the proportion that was assigned to *Homo sapiens*. Note that 3 out of 50 sequences roughly coincides with the extent of present-day contamination (3.3-5.4%) previously estimated using other methods on the full dataset generated from this DNA library (7,045 sequences, [55]).

We then looked at these three DNA sequences and their alignment to the reference dataset:

- Sequence 4 (ACCACCATCCTCCGTGAAATCAATATCCCGCA): Aligns perfectly to all *Homo sapiens* and all Denisovan references, and has one mismatch compared to all Neanderthals.
- Sequence 26 (ACTACGATAGCCCTTATGAAACTTAAGGGTCGA): Aligns perfectly to all *Homo sapiens*, and has one mismatch compared to all Neanderthals and all Denisovans.
- Sequence 48 (ACCCACATCCCTTTCTCCATAAAATTCTTCT-TAGTAGCTATT): Has 3-5 mismatches compared to all sequences.

A plausible conclusion from this output is that sequences 4 and 26 originate from contamination with *Homo sapiens* DNA; whereas



(a) SHAP *Homo sapiens* (b) SHAP Neanderthals

Figure 12: Raw SHAP values for the subset of 50 Neanderthal sequences in REAL 1 (Influence on *Homo sapiens* proportions on the left, and on Neanderthal proportions on the right; Denisovan values omitted for space consideration). Outlier values discussed in the text are highlighted in red.

.

sequence 48 could either contain errors (due to sequencing or damage), and/or could originate from a distant biological species not represented in our reference dataset. Note that we could not infer these insights using solely the outputted proportions of Algorithm 1 - the proportions merely contain the number 0.275 for the *Homo sapiens* label, but the explanation values pointed us to the three sequences in the data that are responsible for this estimation.

To evaluate the ranking accuracy of SHAP, we ran SHAP with only 100 samples (3 seconds), compared to with 1000 samples (33 seconds). The ranking of the three outlier sequences with respect to their contribution to the *Homo sapiens* proportion was 1,4,5 after 100 samples, and 1,2,3 after 1000 samples. Their ranking using the $A_{d,s}$ values was 1,4,5. This suggests that even the faster solutions implemented here would have been sufficient to pinpoint these sequences as outliers for further analysis.

6.8 Summary of Main Experimental Findings

We next summarize the main conclusions from our experimental study.

- **Maximum Likelihood is a simple and explainable approach for estimating proportions in ancient DNA data.** Its accuracy is on par with state-of-the-art solutions to the problem, and it has the advantage of facilitating explainability. A downside is that it is significantly slower (see figure 10a: execution time exceeds 4 minutes for a dataset of size 100, compared to a few seconds for Kallisto); yet its performance is sufficient to render it applicable for the offline processing of ancient DNA datasets.
- For computing Shapley values **to explain the results of a Maximum Likelihood solution** for proportion estimation, **SHAP is an effective solution**. It converges in approximately 15 minutes on a dataset of size 100 (see figures 6, 10c) and is much faster than the alternative Monte Carlo approximation scheme.
- **Faster alternative solutions to computing the actual Shapley values are available**, including the *ranking* of DNA sequences based on their contribution and *determining whether they positively or negatively contribute* to a species estimated proportion. Specifically, using SHAP to estimate the ranking of a dataset of size 100 converges in 4 minutes (see figures 7 and 10c), and a similar time is sufficient to converge to a good sign accuracy (see figures 8 and 10c). Note that with reasonable sacrifice in accuracy of sign or ranking, one can also choose the linear algorithm proposed in 4.3, which is based on the $A_{d,s}$ values of

the likelihood function, and which terminates in fractions of a second (see figures 7 and 8).
- **A simple greedy algorithm for finding counterfactual explanations** by traversing DNA sequences in descending order of individual contributions and adjoining them in subsets **achieves high accuracy** on small datasets in which the ground truth can be computed. For larger datasets, the ground truth counterfactual is not available, but we have shown a correlation between the quality of the ranking oracle and the quality of obtained counterfactuals: specifically, increasing the number of samples allowed in the execution of the SHAP-based oracle yields better counterfactuals. Since the fast $A_{d,s}$-heuristics already yield high-quality rankings, and the overhead of the greedy counterfactual search is only a few seconds (figure 10d), they combine to an algorithm that plausibly yields high-quality counterfactuals (figure 9b).

7 RELATED WORK

Proportion Estimation. The problem of estimating the proportions of different labels in a dataset (sometimes referred to as the problem of "quantification", see [21]) is a natural question that arises in many domains. For instance, it is relevant in estimating the fraction of positive statements in a corpus of texts ([14]), the prevalence of different political views in blog posts ([28]), the proportion of medical reports or medical images with some pathology (see [5, 24]), as a preliminary for use in further machine learning methods ([79], [57]), or as in our use case, to estimate the proportion of different biological species in a genomic dataset ([30]).

The most naive method to solve the problem is by "classify and count", i.e - train a classifier to label each data point, and count the proportion of each label. However. this method is not leveraging the fact that estimating how common a phenomenon is in the data is easier than labeling each data point, and that data point labels might be dependent on each other. There are a plethora of algorithms devised to the problem (see for example [21, 24, 58]), but a simple and standard solution is using a Maximum Likelihood estimator to address the problem (see [38, 56]).

DNA Analysis. In the context of genetic analyses, the estimation of mixtures is of relevance to any setting where the data generated may originate from multiple individuals and/or multiple species. Typical examples include forensic cases (e.g., [16, 26]), the analysis of environmental DNA (e.g., [11, 52]), and the study of ancient DNA. For ancient DNA, mixture estimations are relevant both for the evaluation of contamination by modern and/or exogenous DNA in datasets generated from ancient specimens (e.g., [54]), and, in the case of ancient environmental DNA, to distinguish between multiple ancient organisms that left only minute traces of their DNA in the studied samples (e.g., [52, 67]). We focused on the latter case study here, taking into account that ancient sedimentary DNA datasets can contain low amounts of short and highly-degraded DNA fragments from multiple donors. Notably, existing algorithms to tease apart different donors in mixtures of DNA that have been used for ancient environmental DNA analyses (e.g., [12, 29, 30, 37, 53, 67, 77]) can perform well in many cases; yet, they typically require far more data than what can be retrieved from particularly degraded samples. In addition, to our knowledge, none provide a framework of explainability to investigate their output, hindering the extent of inferences that can be made from particularly challenging datasets.

Explainability. There is a large body of work on algorithmic solutions to explain computation results, from database queries [25, 48] to Machine Learning models [45], with a recent focus on the latter (in the context of *explainable AI*). We have followed in this paper two of the most prominent approaches for explainability, namely attribution via Shapley values [45, 65] and counterfactual explanations [48, 70], and implemented them in the context of proportion estimation. Other measures for attribution have been proposed, including Banzhaf values [6] and causality-based measures. Studying these in the context of proportion estimation is an interesting avenue for future work.

An implementation of a preliminary version of our solution was demonstrated in a conference. The short paper accompanying the demonstration ([9]) focuses on defining and demonstrating the explanations, and solving User Interface design issues (that are out of the scope for the present paper). In contrast, the extensive empirical analysis present here is absent from [9].

8 CONCLUSIONS

We presented in this paper an analysis of methods that compute explanations for proportion estimation in ancient DNA datasets, implemented via Maximum Likelihood. The two notions of explanations that we focused on are (1) attribution, namely assigning scores to individual DNA sequences, based on Shapley values, that capture how they influence the estimated proportion of each species, and (2) counterfactuals, namely subsets of small size that change the estimation made by the algorithm of what is the most dominant species in the dataset. We have demonstrated, through use cases, the usefulness of explanations in this context.

We have further carried out the first (to our knowledge) experimental study conducted in the context of explainability for proportion estimation algorithms, using both synthetic and real data. While we have focused here on the particular application domain of ancient sedimentary DNA, proportion estimation and Maximum Likelihood are prevalent in many other areas of research (see discussion of related work above). Most notions discussed here, as well as the experimental design, are applicable to other settings, and could guide further experimental investigations in these domains.

Our work has focused on explainability with respect to the input dataset of unlabeled DNA sequences. In future work, we plan to study attribution and counterfactual explanations for the set of references used to label the sequences in the data. This could help researchers to better pinpoint the population of origin for the individuals that contributed DNA to the data. For example, by highlighting which of the Neanderthal genomes in the reference dataset drives an identification of Neanderthal sequences in a given sample, researchers could learn which group or population of Neanderthals were the most likely contributors to it. We also plan to design and deploy further optimizations to boost the computation of explanations in this setting.

REFERENCES

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* 298 (2021), 103502.

[2] Sajia Akhter, Ramy K Aziz, Mona T Kashef, Eslam S Ibrahim, Barbara Bailey, and Robert A Edwards. 2017. Kullback Leibler divergence in complete bacterial and phage genomes. *PeerJ* 5 (2017), e4026.

[3] Marcelo Arenas, Pablo Barceló, Leopoldo Bertossi, and Mikaël Monet. 2021. The tractability of SHAP-score-based explanations over deterministic and decomposable Boolean circuits. In *Proceedings of AAAI*. https://arxiv.org/abs/2007.14045

[4] Marcelo Arenas, Pablo Barceló, Leopoldo E Bertossi, and Mikaël Monet. 2023. On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results. *J. Mach. Learn. Res.* 24, 63 (2023), 1–58.

[5] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2013. Variable-constraint classification and quantification of radiology reports under the ACR Index. *Expert systems with applications* 40, 9 (2013), 3441–3449.

[6] J.F. Banzhaf. 1965. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review* 19, 2 (1965), 317–343.

[7] Stefano Benazzi, Viviane Slon, Sahra Talamo, Fabio Negrino, Marco Peresani, Shara E Bailey, Susanna Sawyer, Daniele Panetta, Guiseppe Vicino, Elisabetta Starnini, et al. 2015. The makers of the Protoaurignacian and implications for Neandertal extinction. *Science* 348, 6236 (2015), 793–796.

[8] Amit Bergman. 2023. GitHub repo. https://github.com/Amitbergman/ExplainableProportionEstimationForDNAAnalysis.

[9] Amit Bergman, Viviane Slon, and Daniel Deutch. 2022. exML: An Explainable Maximum Likelihood Tool for Proportion Estimation in DNA Data. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4818–4822.

[10] Leopoldo Bertossi, Benny Kimelfeld, Ester Livshits, and Mikaël Monet. 2023. The Shapley Value in Database Management. *ACM Sigmod Record* 52, 2 (2023), 6–17.

[11] Kristine Bohmann, Alice Evans, M Thomas P Gilbert, Gary R Carvalho, Simon Creer, Michael Knapp, W Yu Douglas, and Mark De Bruyn. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in ecology & evolution* 29, 6 (2014), 358–367.

[12] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* 34, 5 (2016), 525–527.

[13] Ruth MJ Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.. In *IJCAI*. 6276–6282.

[14] Dallas Card and Noah A Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1636–1646.

[15] Shay Cohen, Eytan Ruppin, and Gideon Dror. 2005. Feature selection based on the shapley value. *other words* 1 (2005), 98Eqr.

[16] Camila Costa, Carolina Figueiredo, António Amorim, Lourdes Prieto, Sandra Costa, Paulo Miguel Ferreira, and Nádia Pinto. 2022. Statistical analysis tools of mixture DNA samples: When the same software provides different results. *Forensic Science International: Genetics Supplement Series* 8 (2022), 37–39.

[17] Susan Davidson, Daniel Deutch, Nave Frost, Benny Kimelfeld, Omer Koren, and Mikaël Monet. 2022. ShapGraph: An Holistic View of Explanations through Provenance Graphs and Shapley Values. In *Proceedings of the 2022 International Conference on Management of Data*. 2373–2376.

[18] M Dayhoff, R Schwartz, and B Orcutt. 1978. 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure* 5 (1978), 345–352.

[19] Daniel Deutch, Nave Frost, Amir Gilad, and Oren Sheffer. 2021. Explanations for data repair through shapley values. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 362–371.

[20] Daniel Deutch, Nave Frost, Benny Kimelfeld, and Mikaël Monet. 2022. Computing the Shapley value of facts in query answering. In *Proceedings of the 2022 International Conference on Management of Data*. 1570–1583.

[21] Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9, 4 (2015), 1–27.

[22] Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. 2008. A linear approximation method for the Shapley value. *Artificial Intelligence* 172, 14 (2008), 1673–1699.

[23] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 196–204.

[24] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. 2013. Class distribution estimation based on the Hellinger distance. *Information Sciences* 218 (2013), 146–164.

[25] Todd J Green and Val Tannen. 2017. The semiring framework for database provenance. In *Proceedings of PODS*. 93–99. https://dl.acm.org/doi/10.1145/3034786.3056125

[26] H Haned, T Egeland, D Pontier, L Pene, and P Gill. 2011. Estimating drop-out probabilities in forensic DNA samples: a simulation approach to evaluate different models. *Forensic Science International: Genetics* 5, 5 (2011), 525–531.

[27] Steven Henikoff and Jorja G Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 22 (1992), 10915–10919.

[28] Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54, 1 (2010), 229–247.

[29] Ron Hübler, Felix M Key, Christina Warinner, Kirsten I Bos, Johannes Krause, and Alexander Herbig. 2019. HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology* 20, 1 (2019), 1–13.

[30] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. 2007. MEGAN analysis of metagenomic data. *Genome research* 17, 3 (2007), 377–386.

[31] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[32] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[33] Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. 2021. Fastshap: Real-time shapley value estimation. In *International Conference on Learning Representations*.

[34] Hákon Jónsson, Aurélien Ginolhac, Mikkel Schubert, Philip LF Johnson, and Ludovic Orlando. 2013. mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29, 13 (2013), 1682–1684.

[35] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *Comput. Surveys* 55, 5 (2022), 1–29.

[36] Nikhil Ketkar and Nikhil Ketkar. 2017. Stochastic gradient descent. *Deep learning with Python: A hands-on introduction* (2017), 113–132.

[37] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research* 26, 12 (2016), 1721–1729.

[38] Youngseok Kim, Peter Carbonetto, Matthew Stephens, and Mihai Anitescu. 2020. A fast algorithm for maximum likelihood estimation of mixture proportions using sequential quadratic programming. *Journal of Computational and Graphical Statistics* 29, 2 (2020), 261–273.

[39] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2021. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 652–663.

[40] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[41] Yongchan Kwon and James Y Zou. 2022. WeightedSHAP: analyzing and improving Shapley based feature attributions. *Advances in Neural Information Processing Systems* 35 (2022), 34363–34376.

[42] Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, and Moshe Sebag. 2020. The Shapley value of tuples in query answering. In *ICDT*, Vol. 155. Schloss Dagstuhl, 20:1–20:19. https://arxiv.org/abs/1904.08679

[43] Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, and Moshe Sebag. 2021. The Shapley Value of Tuples in Query Answering. *Log. Methods Comput. Sci.* 17, 3 (2021). https://doi.org/10.46298/lmcs-17(3:22)2021

[44] Bastien Llamas, Guido Valverde, Lars Fehren-Schmitz, Laura S Weyrich, Alan Cooper, and Wolfgang Haak. 2017. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Science & Technology of Archaeological Research* 3, 1 (2017), 1–14.

[45] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[46] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[47] Alexandra Meliou, Wolfgang Gatterbauer, Joseph Y Halpern, Christoph Koch, Katherine F Moore, and Dan Suciu. 2010. Causality in databases. *IEEE Data Engineering Bulletin* 33, ARTICLE (2010), 59–67.

[48] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2010. The complexity of causality and responsibility for query answers and non-answers. *Proceedings of the VLDB Endowment* (2010).

[49] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. 2022. Sampling permutations for shapley value estimation. *The Journal of Machine Learning Research* 23, 1 (2022), 2082–2127.

[50] Nicholas Moehle, Stephen Boyd, and Andrew Ang. 2021. Portfolio performance attribution via Shapley value. *arXiv preprint arXiv:2102.05799* (2021).

[51] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[52] Mikkel Winther Pedersen, Søren Overballe-Petersen, Luca Ermini, Clio Der Sarkissian, James Haile, Micaela Hellstrom, Johan Spens, Philip Francis Thomsen, Kristine Bohmann, Enrico Cappellini, et al. 2015. Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370, 1660 (2015), 20130383.

[53] Mikkel W Pedersen, Anthony Ruter, Charles Schweger, Harvey Friebe, Richard A Staff, Kristian K Kjeldsen, Marie LZ Mendoza, Alwynne B Beaudoin, Cynthia Zutter, Nicolaj K Larsen, et al. 2016. Postglacial viability and colonization in North America's ice-free corridor. *Nature* 537, 7618 (2016), 45–49.

[54] Stéphane Peyrégne and Kay Prüfer. 2020. Present-Day DNA Contamination in Ancient DNA Datasets. *BioEssays* 42, 9 (2020), 2000081.

[55] Stéphane Peyrégne, Viviane Slon, Fabrizio Mafessoni, Cesare De Filippo, Mateja Hajdinjak, Sarah Nagel, Birgit Nickel, Elena Essel, Adeline Le Cabec, Kurt Wehrberger, et al. 2019. Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Science advances* 5, 6 (2019), eaaw5873.

[56] Pere Puigbò, Alexander E Lobkovsky, David M Kristensen, Yuri I Wolf, and Eugene V Koonin. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC biology* 12 (2014), 1–19.

[57] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. 2008. Estimating labels from label proportions. In *Proceedings of the 25th international conference on Machine learning*. 776–783.

[58] Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. 2016. Mixture Proportion Estimation via Kernel Embeddings of Distributions. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 2052–2060. https://proceedings.mlr.press/v48/ramaswamy16.html

[59] Gabriel Renaud, Viviane Slon, Ana T Duggan, and Janet Kelso. 2015. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome biology* 16, 1 (2015), 1–18.

[60] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[61] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The shapley value in machine learning. *arXiv preprint arXiv:2202.05594* (2022).

[62] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).

[63] Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Conrad L Schoch, Stephen T Sherry, and Ilene Karsch-Mizrachi. 2022. GenBank. *Nucleic acids research* 50, D1 (2022), D161.

[64] LS Shapley. 1953. QUOTA SOLUTIONS OP n-PERSON GAMES1. *Edited by Emil Artin and Marston Morse* (1953), 343.

[65] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317. http://www.library.fa.ru/files/Roth2.pdf#page=39

[66] Pontus Skoglund, Bernd H Northoff, Michael V Shunkov, Anatoli P Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences* 111, 6 (2014), 2229–2234.

[67] Viviane Slon, Charlotte Hopfe, Clemens L Weiß, Fabrizio Mafessoni, Marco De La Rasilla, Carles Lalueza-Fox, Antonio Rosas, Marie Soressi, Monika V Knul, Rebecca Miller, et al. 2017. Neandertal and Denisovan DNA from Pleistocene sediments. *Science* 356, 6338 (2017), 605–608.

[68] Caitlin M Stewart, Prachi D Kothari, Florent Mouliere, Richard Mair, Saira Somnay, Ryma Benayed, Ahmet Zehir, Britta Weigelt, Sarah-Jane Dawson, Maria E Arcila, et al. 2018. The value of cell-free DNA for molecular pathology. *The Journal of pathology* 244, 5 (2018), 616–627.

[69] Mukund Sundararajan and Amir Najmi. 2020. The many Shapley values for model explanation. In *International conference on machine learning*. PMLR, 9269–9278.

[70] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual explanations and algorithmic recourses for machine learning: a review. *arXiv preprint arXiv:2010.10596* (2020).

[71] Benjamin Vernot, Elena I Zavala, Asier Gómez-Olivencia, Zenobia Jacobs, Viviane Slon, Fabrizio Mafessoni, Frédéric Romagné, Alice Pearson, Martin Petr, Nohemi Sala, et al. 2021. Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. *Science* 372, 6542 (2021), eabf1667.

[72] Samuel H Vohr, Rachel Gordon, Jordan M Eizenga, Henry A Erlich, Cassandra D Calloway, and Richard E Green. 2017. A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. *Forensic Science International: Genetics* 30 (2017), 93–105.

[73] Jinliang Wang. 2003. Maximum-Likelihood Estimation of Admixture Proportions From Genetic Data. *Genetics* 164, 2 (06 2003), 747–765. https://doi.org/10.1093/genetics/164.2.747 arXiv:https://academic.oup.com/genetics/article-pdf/164/2/747/42052529/genetics0747.pdf

[74] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of COLT*, Vol. 8. 6. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.680.490&rep=rep1&type=pdf

[75] David S Watson. 2021. Interpretable machine learning for genomics. *Human genetics* (2021), 1–15.

[76] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. 2007. Database resources of the national center for biotechnology information. *Nucleic acids research* 35, suppl_1 (2007), D5–D12.

[77] Derrick E Wood and Steven L Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15, 3 (2014), 1–12.

[78] Han Yuan, Mingxuan Liu, Lican Kang, Chenkui Miao, and Ying Wu. 2022. An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. *arXiv preprint arXiv:2204.11351* (2022).

[79] Daniel Zeiberg, Shantanu Jain, and Predrag Radivojac. 2020. Fast nonparametric estimation of class proportions in the positive-unlabeled classification setting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6729–6736.