universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Performance benchmarking of methods for detecting genetic introgression from unsampled lineages

verfasst von | submitted by

Josef Anna Leopold Hackl BA BA BSc BSc BA MA MSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 877
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Genetik und Entwicklungsbiologie
Betreut von | Supervisor:	Ass.-Prof. Dr. Martin Kuhlwilm

Abstract

Genetic admixture and introgression, i.e. the exchange of genetic information between different species, is a common phenomenon in the history of species, including human evolution. Of special interest is 'ghost introgression', i.e. gene flow events where the source population of the introgressed material is unknown. This might be common, especially for populations where data is scarce or the source lineages are extinct. In this thesis, computational methods for detecting introgressed fragments, in particular ghost introgression, are evaluated. The applied tools use different statistical and computational methods, e.g. summary statistics are computed and machine learning approaches like logistic regression or neural networks are applied. A central part was the implementation of a pipeline to compare their performance systematically on data simulated under different demographic models with distinct features, e.g. gene flow from Neanderthals to modern humans or from an extinct population into bonobos. Furthermore, two Machine Learning (ML) approaches, based on Logistic Regression and on a Deep Learning (DL) architecture using convolutional neural networks (CNNs), have been fully reimplemented so that they can be applied more flexible on data of different demographic models. Their functionality was extended and the performance in the detection of introgressed fragments compared to the other tools.

Kurzfassung

Admixture und Introgression, der Austausch genetischen Materials zwischen verschiedenen Spezies, ist ein Phänomen, das in der Geschichte der Evolution, einschließlich der Entwicklung von Homo sapiens, eine bedeutsame Rolle spielt. Von besonderem Interesse ist „ghost introgression“, womit solche Fälle, bei denen die Quelle des genetischen Materials unbekannt ist, bezeichnet werden. Dies kann insbesondere bei Arten vorkommen, für die nur wenige Daten vorliegen oder die ausgestorben sind, ohne ausreichend fossile Evidenz zu hinterlassen. In dieser Arbeit werden informatische Methoden zur Erkennung introgredierter Fragmente, insbesondere eben der "ghost regression", evaluiert. Die eingesetzten Tools nutzen unterschiedliche statistische und informatische Methoden, so werden Methoden des maschinellen Lernens wie logistische Regression oder neuronale Netzwerke angewendet. Ein zentraler Teil ist die Implementierung einer Pipeline zum systematischen Vergleich der Leistung verschiedener Tools anhand von Daten, die unter verschiedenen demografischen Modellen mit unterschiedlichen Merkmalen simuliert wurden, z. B. Genfluss vom Neandertaler zum modernen Menschen oder von einer ausgestorbenen Population zum Bonobo. Darüber hinaus wurden zwei Tools, die auf logistischer Regression und auf einer Deep Learning (DL)-Architektur unter Verwendung von Convolutional Neural Networks (CNNs) basieren, vollständig neu implementiert, sodass sie flexibler auf Daten verschiedener demografischer Modelle angewendet werden können. Der Funktionsumfang wurde erweitert und ihre Leistung in der Detektion von introgredierten Fragmenten mit derjenigen der anderen Tools verglichen.

Contents

Abstract i

Kurzfassung iii

1 Introduction 1

2 Theoretical foundations 3
- 2.1 Admixture / Introgression . . . 3
  - 2.1.1 Introgression events in human evolution . . . 4
  - 2.1.2 Biological and functional relevance . . . 4
  - 2.1.3 Research on ghost introgression . . . 4
- 2.2 Statistical methods for detection of ghost introgression . . . 7
- 2.3 Machine Learning methods for introgression detection . . . 8

3 Methods 11
- 3.1 Tools for detecting Ghost introgression . . . 11
  - 3.1.1 Sstar . . . 11
  - 3.1.2 SPrime . . . 11
  - 3.1.3 HMMix . . . 12
  - 3.1.4 ArchIE . . . 13
  - 3.1.5 IntroUNET . . . 13
- 3.2 Models . . . 15
  - 3.2.1 ArchIE model . . . 15
  - 3.2.2 Human-Neanderthal model . . . 15
  - 3.2.3 African-Ghost model . . . 17
  - 3.2.4 Weakly Structured Stem model . . . 17
  - 3.2.5 Bonobo-Ghost model . . . 17
  - 3.2.6 Gorilla-Ghost model . . . 19

4 Implementation details 21
- 4.1 Modifications of the general workflow . . . 21
  - 4.1.1 Snakemake pipelines . . . 21
  - 4.1.2 Extraction of migration events . . . 22
- 4.2 Integration of the latest version of HMMix . . . 23
- 4.3 Reimplementation of ArchIE . . . 23
- 4.4 Reimplementation of IntroUNET . . . 23
- 4.5 Evaluation of predictions . . . 24

5 Results **25**
5.1 ArchIE results . . . 25
5.1.1 ArchIE feature importance . . . 25
5.1.2 Reliability of results . . . 31
5.1.3 Different classifiers . . . 34
5.2 IntroUNET results . . . 36
5.2.1 Variations of IntroUNET . . . 37
5.2.2 IntroUNET performance under model misspecification . . . 38
5.2.3 Performance on different individual sizes . . . 38
5.2.4 Performance on unphased data . . . 40
5.2.5 Performance for the African WSS models with and without gene flow 41

6 Discussion **47**
6.1 ArchIE results . . . 47
6.2 IntroUNET results . . . 48

7 Conclusion and Outlook **51**

Bibliography **53**

1 Introduction

Genetic admixture and introgression, *i.e.* the exchange of genetic information between different species, is a common phenomenon in the history of species. A remarkable example of archaic introgression is the gene flow from Neanderthals and Denisovans to modern humans. The consequences as well as the evolutionary impact of genetic admixture are complex and only partly understood. Contrariwise, the detection of introgression deserts, i.e. regions of the genome depleted of introgressed material, could hint to functional incompatibilities and thus indicate speciation factors.

Of special interest is 'ghost introgression', *i.e.* gene flow events where the source population of the introgressed material is unknown [1]. This might be a common phenomenon, especially for populations where data is scarce or the source lineages are extinct.

The robust and reliable identification of introgressed fragments is still an area under development, in particular for ghost introgression. Several measures to detect ghost introgression have been proposed, *e.g.* the S* metric which analyses patterns of linkage disequilibrium [2, 3].The currently available computational tools are based on different methods from statistics and machine learning, *e.g.* Hidden Markov Models [4], logistic classification [5] or neural networks [6], relying on features like the individual frequency spectrum or the S* metric.

In this Master thesis, some tools for determining introgressed fragments from ghost admixture were evaluated. A central part was the implementation of a pipeline to compare their performance systematically on data simulated under different demographic models with distinct features, *e.g.* gene flow from Neanderthals to modern humans or from an extinct population into bonobos. Furthermore, two Machine Learning (ML) approaches, based on Logistic Regression and on a Deep Learning (DL) architecture using convolutional neural networks (CNNs), have been fully reimplemented so that they can be applied more flexible on data of different demographic models. There functionality was extended and compared to the other tools.

2 Theoretical foundations

2.1 Admixture / Introgression

Admixture is the process through which two or more previously isolated populations begin to interbreed, resulting in a mixture of their genetic materials. This phenomenon common in nature is a fundamental aspect of evolutionary biology and plays a significant role in shaping the genetic diversity of populations. However, detecting admixture events, especially those involving 'ghost' populations (extinct or unsampled populations), is particularly challenging. It is also difficult to distinguish introgression from other evolutionary processes, in particular incomplete lineage sorting (ILS), *i.e.* the imperfect segregation of all alleles into all lineages which causes that the evolutionary history of individual genomic fragments does not match the species tree [7].

Introgression is the movement of genetic material from one species or population into another through hybridization and repeated backcrossing with individuals from one of the parent populations. After two populations interbreed, their hybrid offspring continue to reproduce with one of the original populations, leading to the gradual incorporation of genetic material from one population into another. A typical admixture event can be visualized as the mixing of genetic fragments from different populations into a single genome (fig. 2.1) [8], where typically, these fragments are longer than expected for ILS given the divergence time of these populations. The population which obtains genetic material can be denoted as target population, the donor represents the source population.

These fragments, however, are not always easily identifiable, particularly when one of the source populations is a ghost population. Ghost introgression refers to the transfer of genetic material from an unsampled, extinct or unknown, (and hence termed 'ghost') population into the gene pool of a living species or population. There is no direct evidence of the existence of the ghost population through fossils or other physical remains, *i.e.* gene flow from unsampled lineages. The existence of this unidentified or unsequenced population is inferred solely from genetic data. Since there is no physical evidence (*e.g.* fossils) of this population, *i.e.* the source of the donor material is unknown, but its genetic traces can be found in the DNA of modern species. Such ghost admixture can leave subtle signatures in the genetic material of extant populations that are difficult to detect, requiring sophisticated statistical and computational approaches.

Various statistical methods have been invented to detect admixture, *e.g.* D-statistics which compares patterns of shared alleles between populations. However, finding introgressed fragments from a ghost population within the individual genomes is particularly challenging: Many methods, in particular classical statistics-based ones, rely on knowledge about the putative source population for inferring admixture, whereas methods

for detecting ghost introgression must not require an archaic reference genome for the identification of archaic genetic material.

2.1.1 Introgression events in human evolution

A well-known example of introgression is the contribution of other hominin lineages to modern humans. Genetic analysis of modern human populations and the recovering of ancient DNA from extinct hominins has revealed traces of genetic exchange between humans and archaic hominins, like Neanderthals [9–11] and Denisovans [12]. It is estimated that Neanderthals have contributed approximately 2% to non-African genomes, Denisovan introgression has been observed in Oceanian populations where up to 5% of the genome can be derived from Denisovans.

2.1.2 Biological and functional relevance

Generally, introgression has functional implications in individuals of the receiving population, in humans as well as other clades [13]. This is also true for Neanderthal and Densiovan introgression in human evolution [14]. For example, there is evidence that Neandertal introgression contributed to risk factors for severe COVID-19 [15] as well as protection against it [16]. Furthermore, introgression likely shaped circadian traits [17] and Denisovan introgression might affect immune-related processes like gene regulation in Papuans [18].

In some cases, introgression can lead to beneficial adaptions which are enforced by positive selection in the recipient population; this phenomenon is called adaptive introgression. Also in human evolution, there are functional implications due to the presence of admixture [13]. A striking example is the Denisovan-*EPAS1* haplotype in Tibetans which regulates the response to hypoxia and is advantageous in high altitude [19, 20]. Similarly, there is evidence that Neanderthal introgression was beneficial for adaption to new environments, *e.g.* providing adaptions due to changed climate conditions, pathogens and UV exposure levels, although also deleterious effects have been investigated [21].

On the other hand, introgression deserts, *i.e.* regions depleted of signs of introgression can indicate genetic material which was of special importance in human evolution [22]. Such introgression deserts can point to crucial factors of speciation, for instance in case of the *FOXP2* gene [23] or depletion of introgressed fragments on the X-chromosome which can be found in humans as well as in bonobos and gorillas [24, 25].

For the detection of admixture, adaptive introgression poses additional challenges. In particular, introgression itself can lead to false positives of positive selection and it is difficult to disentangle these processes [26].

2.1.3 Research on ghost introgression

In recent years, due to the availability of high-quality genome assemblies and improved statistical and computational methods, new examples of introgression were identified; in particular, further evidence for ghost introgression events has appeared, suggesting that

it had important influence on evolutionary processes like adaption and speciation and also calling for a revision of the classical image of a tree of life which seems to be much more reticulate and interconnected than previously thought. Therefore, it is imperative to develop and evaluate methods for taking into account ghost introgression [1]. Ignoring possible ghost introgression events can lead to false demographic assumptions [27, 28]. Recently, it was reported that overlooked ghost lineages can significantly bias the results of introgression detection methods and thus led to erroneous evolutionary trees [27] and its occurrence has important implications for species tree reconstruction [7, 29].

Besides the genetic material of Neanderthal and Denisovan origin, genetic studies have also suggested that other unknown hominin species may have contributed to the diversity of the human genome, even though no fossil record of these 'ghost' species has been found [30–34]. Furthermore, genetic exchange with unknown populations has been suggested for other animals, in particular great apes. Ancient gene flow between bonobos and an unknown lineage is reported in [24]. This admixture event likely occurred several hundred thousand years ago, contributing up to 4.8% of the genome from this 'ghost' ape population. In gorilla evolution, there is evidence of admixture from an archaic ghost lineage in the common ancestor of eastern gorillas, in contrast to western gorillas. This lineage diverged over 3 million years ago, and up to 3% of the genome in eastern gorillas shows signs of this ancient introgression, which likely happened more than 40,000 years ago [29]. In both cases, these admixture events might have had functional consequences concerning behaviour, immunity and physiology; for example, the genetic exchange between gorillas might had influence on bitter taste perception [29].

In the past few years, signs of ghost introgression have been found in multiple species, *e.g.* in plants (Cycas revoluta, [35]) and ricefishes [36]. Given this widespread evidence for ghost introgression, it is crucial to devise methods to precisely locate introgressed fragments within individual genomes.

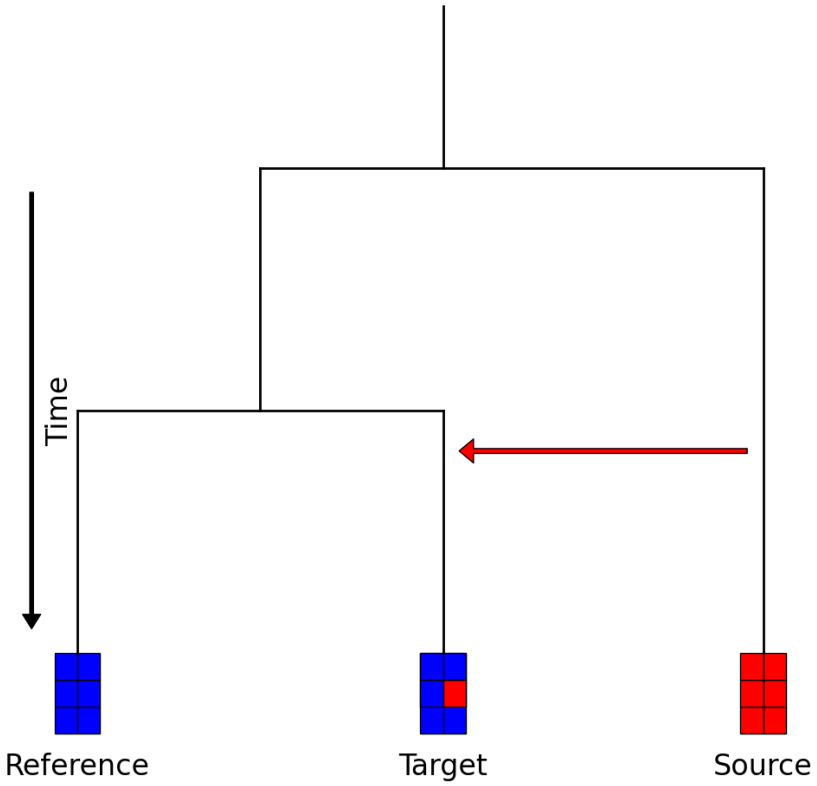

Figure 2.1: Introgression between different lineages: Exchange of genetic material between source and target population. The reference population is the population without introgressed fragments, the target population is the population that obtains introgressed fragments from the source population, and the source population is the population that donates introgressed fragments to the target population (shown in red). In case of an unknown source population, the source population is termed ghost population.

2.2 Statistical methods for detection of ghost introgression

Since ghost introgression describes the lack of knowledge of the archaic population which exchanged genetic material with the target population of interest, methods for detecting ghost introgression must not require an archaic reference genome for the identification of archaic genetic material. A classical statistical method to scan genetic material for introgression without need for knowing about the source population is the S* statistic.

It was proposed in [2] (even before the first draft of the Neanderthal genome was published) to detect signatures of linkage disequilibrium, *i.e.* the discrepancy between the probability to observe two alleles at two loci together and the probability to observe them independently, $D_{AB} = p_{AB} - p_A p_B$. To determine whether the level of linkage disequilibrium within the target population provides significant evidence for introgression, a comparison with a reference population without introgression is necessary.

Since the calculation of the S* statistic does not incorporate information from the source population, it is applicable for the task of detecting ghost introgression.

S* is defined as

$$S^* = S(I) = \max_{J \subset \{1,\dots,n\}} S(J) \tag{2}$$

where

$$S(i_1, \dots, i_k) = \sum_j S(i_j, i_{j+1}) \tag{3}$$

. i_j denotes the individual alleles, the index corresponds to the ordering of the positions along the chromosome. I is the subset of SNPs which maximize the score. The distance between two SNPs is computed as the differences in the genotypes. If the distance is above a certain threshold it is set to -infinity, if below, to a negative number (-10000 is used in [2]). In case of a distance of zero, the score is set to the distance in bp between the pair of SNPs plus a positive number (5000).

Efficient calculation of the S* score can be done using dynamic programming. Given

$$S_j^* = \max_{J \subseteq \{1,\dots,j-1\}} S(\{J \cup \{j\}\}), \tag{2.1}$$

, S_{j+1} can be computed using the recursion:

$$S_{j+1}^* = \max_{k=1,\dots,j} [S_k^* + S(k, j+1)] \tag{2.2}$$

In [37] and [38], slightly different versions of the S* statistic are proposed. [38] investigates Neanderthal and Denisovan introgression in Melanesian individuals. Whereas in the previous versions a group of (potential) target individuals was used for the calculation, only one is used here, giving rise to the following scoring function:

$$\mathrm{S(J)} = \sum_{j \in J} \begin{cases} -10000 & d(j, j+1) > 0 \\ 5000 + bp(j, j+1) & d(j, j+1) = 0 \\ 0 & j = \max(J) \end{cases} \tag{2.3}$$

, with $d(j, j+1)$ denoting the genotype distance between two adjacent variants and $bp(j, j+1)$ the distance in base pairs.

2.3 Machine Learning methods for introgression detection

Available tools can be roughly divided into two classes: Either summary statistics are used or features are extracted from raw data, *e.g.* genotype matrices in which the genetic variation across different samples at specific positions on the chromosome is encoded. However, also hybrid forms are possible: Summary statistics are calculated and subsequently used for creation of a feature vector which serves as input for a ML classifier. (For a more detailed description of the applied tools see the next chapter.)

Previous work has indeed shown that ML methods are very suitable for admixture detection [39]. However, classical methods struggle with ghost admixture scenarios, complex demographic histories and small sample sizes.

Given the size and potentially complex patterns of the involved data, the application of Deep Learning approaches seems very promising to study phenomena like admixture (in [40], one finds a comprehensive review of DL in population genetics). Taking into account the achievements in other fields, it is plausible that approaches that rely on DL succeed in detecting subtle admixture events that traditional methods may miss and significantly increase the accuracy in determining introgressed material.

In contrast to classical statistical approaches, many of the ML approaches, and in particular DL architectures which take raw data, *e.g.* genotype matrices, as input, do not or only modestly rely on the background knowledge and the assumptions from population genetics theory. Instead of calculating summary statistics, automatic feature extraction gathers information which is predictive for the given task.

A very popular DL architecture are Convolutional Neural Networks (CNNs). CNNs have shown great success in population genetics tasks, amongst others in introgression [41], where they were shown to accurately predict introgressed loci and outperform summary statistics-based ML approaches [41]. CNNs were also applied to discern adaptive introgression [42].

CNNs have been widely used in image classification and segmentation tasks; their classical use case is in the processing of image-like input, *i.e.* grid-/matrix-shaped data. The core part of a CNN is the convolutional layer. In the convolutional operation, a filter/kernel slides over the input and performs a mathematical operation (usually element-wise multiplication and summation) at each position, thereby creating a feature map. Subsequent pooling operations reduce the size of the intermediate representations. Common Pooling operations are max-pooling (the maximum value is taken for each

window) and mean-pooling (the mean value is taken). In the training process, the learnable weights of the filters are updated.

In contrast to tools which are tailored to a specific model or use case, *e.g.* SPrime (see the next chapter for a more detailed description of the tools) which is optimized for Neandertal introgression, DL methods can be assumed to work for a broad range of demographic models and parameters. On the other hand, robustness to model misspecifications is crucial since the parameters of the demographical model will not exactly be met. Potential pitfalls common for such architectures, *e.g.* overfitting to the demographic model of the training data, should therefore also be explored.

3 Methods

3.1 Tools for detecting Ghost introgression

For benchmarking, the following tools were applied: Sstar, SPrime, HMMix, ArchIE and IntroUNET. In case of the last two tools, a large part of the code was reimplemented in Python.

In [3], a comparative benchmark including Sstar, HMMix, SPrime was presented. It was shown that the performance of the tools heavily depends on the data; for example, SPrime performs well only in the Neanderthal and Denisovan introgression scenarios and Sstar outperforms HMMix in the case of small sample sizes.

3.1.1 Sstar

A recent implementation of the S* statistic can be found in [3]. Simulations under a demographic model without introgression are used for fitting a generalized additive model. This model is used to predict the expected S* scores which allows to assess the significance of the S* scores obtained for the input data. No source genome is necessary; however, if available, it can be used to calculate the source match rate between an individual from the target population and one from the source population.

The Sstar-package directly implements the S*-approach described in the previous chapter. Similar to the version in [38], the algorithm only processes the information from one target individual. Variants which feature fixed, derived alleles in the reference and the target population are removed. Simulations under a demographic model without introgression are used for fitting a generalized additive model. This model is used to predict the expected S* scores which allows to assess the significance of the S* scores obtained for the input data. No source genome is necessary; however, if available, it can be used to calculate the source match rate between an individual from the target population and one from the source population.

3.1.2 SPrime

SPrime is a similar tool which does not need source genomes and relies on the S*-approach [43]. It implements an adapted version of the S* statistic which allow for the inclusion of multiple individuals of the target population and alleles of low frequency in the outgroup. The parameters of the involved scoring function have been optimized for detecting introgression from Neanderthals and Denisovans in modern humans; previous benchmarks [3] have shown that for other models its performance is significantly worse.

As in the case of the S*-statistics,

$$S^*_{j+1} = \max\left\{0, \max_{k=1,2,\ldots,j}\left(S^*_k + T(v_k, v_{j+1})\right)\right\} \tag{3.1}$$

is recursively computed. However, the scoring function $T(v_1, v_2)$ differs from the original formulation.

There are several modifications: In the original S* statistics, only private alleles are counted, whereas SPrime also includes alleles which have a low frequency in the outgroup. Furthermore, SPrime considers local differences for mutation and recombination rates for computing the score.

This leads to the following scoring formula for a pair of alleles:

$$6000w\frac{1-\exp\left(-\frac{0.01}{m}\right)}{1-\exp(-1)} - 25000\frac{D}{n}. \tag{3.2}$$

, where m denotes an estimate of the local rate of mutation in centiMorgan per meiosis, w is a weight factor taking into account the frequency of alleles uncommon in the outgroup. Considering alleles v_1 and v_2, the number of copies of the alleles in individual i is given by X_{i1} and X_{i12}, respectively. For a single introgressed haplotype, it is expected that $X_{i1} = X_{i2}$, but in case of recombination events between introgressed and non-introgressed haplotypes, it will differ. This difference is given by $D = \sum_i |X_{i1} - X_{i2}|$. n indicates the minimum number of haplotypes found for both alleles, $n = \min(\sum_i X_{i1}, \sum_i X_{i2})$.

The parameters of this scoring function have been found using simulated data and the Altai Neanderthal genome as reference. Values above the threshold (which is 150000 per default) are assumed to consist of introgressed material.

SPrime is optimized for detecting introgression from Neanderthals and Denisovans in modern humans [44] and relies on an out-of-Africa Neanderthal-admixture model [43]; for non-human models, it performs rather poorly [3]. Furthermore, introgressed fragments have to be sufficiently long to be detected.

3.1.3 HMMix

Another tool for detecting archaic introgression from Neanderthals and Denisovans is HMMix [4].

The approach is based on a Hidden Markov-model with two states, Archaic and Ingroup. Genomic segments containing introgression are expected to have a high density of variants which are missing in the population which did not obtain introgressed material. Only private variants are maintained, *i.e.* variants which are only observed in the ingroup, but not the outgroup, *i.e.* the population without introgression. This facilitates the discrimination between introgressed and non-introgressed segments. The Baum-Welch algorithm is applied to find the transition parameters.

For application of the model, the genome is partitioned into short windows (per default 1000 bp) and the number of the private variants within the window is counted.

In principle, this method does not make assumptions on the demographic history, nor does it assume specific numbers of individuals to be available in the reference or target

panels. Hence, it could be applied to any population complex; however, according to [3], it performs well only for a sufficiently large sample size.

3.1.4 ArchIE

ARCHaic Introgression Explorer (ArchIE) can be seen as a hybrid between classical statistical-based approaches and ML since the tool combines several population genetic summary statistics to infer archaic local ancestry without the need for a reference genome. ArchIE is based on a logistic regression model that predicts the probability of archaic ancestry for each window along an individual genome. In [5], it is claimed that the method outperforms the application of S*-statistics.

In the first step, a feature vector which serves as input for the logistic classifier is created. The features encompass the individual frequency spectra, the number of private SNPs, i.e. variants which are not observed in the reference population, the pairwise differences between the haplotypes of the two genotype matrices, diverse summary statistics (mean, variance, skew, kurtosis) of this distance vector, and a variation of the S*-score. The full set of features adds up to a feature vector of more than 200 entries.

There are only three features which take the reference into account: The minimum distance to the reference, the number of private variants and, somewhat more indirectly, the S*-score (because only the variants which are not present in the reference are selected). Except the frequency spectra and the pairwise differences, which depend on the number of sampled individuals, all features consist only of one entry in the feature vector. In the following, these features are called fixed-size features.

The default window size (i.e. the feature vector which acts as input for the logistic regression classifier) is 50k, the sliding window stepsize 10k. Per default, a threshold of 30 and 70 % is applied; vectors below 30 % introgressed content are labeled as non-introgressed, above 70 % as introgressed and all vectors in between are considered ambiguous and are discarded from the training set. For inference, it is averaged over the predictions for all windows.

3.1.5 IntroUNET

IntroUNET [6] is an example for the usage of modern Deep Learning approaches; it is able to detect admixture by recognizing patterns in genetic data given as genotype matrices.

It consists of a U-Net which is a convolutional neural network (CNN) architecture designed for image segmentation; in particular, it is a sort of fully convolutional network since it consists only of convolutional layers and contains no additional dense layers. It comprises of multiple convolution layers and upsampling operations which create representations of different resolution. These representations are combined so that information at different scales can be extracted. In the first half of the network, four (given the standard settings) pooling operations reduce the height and width of the input matrix by a factor of 2, and in the second half upsampling operations restore the original size. This structure also requires that the input is a multiple of 16. Finally, one value per entry is predicted.

Therefore, in contrast to the methods mentioned before, IntroUNET allows predictions on a single nucleotide polymorphism (SNP)-level, therefore enabling a much finer granularity, whereas the already mentioned approaches which rely on summary statistics give predictions for the windows these statistics are calculated for.

Two genotype matrices, containing a fixed number of SNPs, are provided for training, one for all reference and one for the target individuals. Furthermore, an 'introgression' matrix consisting of 0 and 1 has to be provided for training: Each row indicates for each of the SNPs for an individual whether it is in an introgressed region (1) or not (0). For inference, the genotype matrices are given as input and an introgression matrix is predicted.

In the original IntroUNET publication, the simulation data is provided via simulations of a modified version of the coalescent simulator (see chapter 'implementation' for some more details on coalescent simulations) ms [45]. Finally, a HDF-file (Hierarchical Data Format, this format allows an efficient storage of large amounts of data) is created using these arrays. Depending on the parameters, one can determine one population as having introgression or choose bi-directional introgression (in this case, there are two introgression matrices stored in the HDF-file , one for each population). For the models investigated in this study, only the uni-directional case was of interest.

As noted above, CNNs have shown state-of-the-art performance in pattern recognition. However, in the context of genetic studies, this requires that such patterns are discernible in the input representation. For this reason, a technique called seriation is crucial for the performance of IntroUNET [6]: The genotype matrices are sorted in such a way that more similar haplotypes, *i.e.* sharing more variants, are stored in adjacent rows. The maximized similarity between adjacent rows leads to a pattern which can be detected by the convolution operation. One of the populations is sorted by employing the seriation algorithm, and the other one is matched with the seriated one using the Kuhn-Munkres algorithm for linear matching so that also the similarity between the two channels is maximized. In the IntroUNET workflow, the seriation part is done during the creation of the HDF-files.

By default, a CNN window consists of 192 polymorphisms and 112 individuals. If the number of reference and target individuals does not agree with this number, they have to be randomly upsampled to fit into the input matrix.

For predictions, it is averaged over all CNN output windows. One can apply either usual averaging or techniques like Gaussian smoothing, in which a weighted average corresponding to a Normal distribution is calculated, i.e. the values in the middle of a window get higher weight than at the edges. According to [6], this is preferable, since predictions at the edge of a polymorphism window are less accurate and should not contribute so much to the final value.

For IntroUNET, for individual sizes strongly differing from the nref/ntgt=50/50 case, some modifications are necessary (see also chapter implementation). On the one hand, the input dimensions are not arbitrary since they have to be multiples of 16 to ensure that all upscaling operations can be performed. To satisfy this condition for random numbers of individuals, random upsampling is performed. For larger numbers of individuals,

the dimensions of the layers have to be increased, whereas for smaller numbers it is advantageous to round up to the next multiple of 16 because it would be inefficient or even detrimental to populate the input data with replicates of the same individuals.

3.2 Models

A variety of demographic models were implemented using the DEMES specification [46]. In table 3.1, the values for mutation and recombination rate as well as admixture proportion which were used with the specific models is indicated.

Model	Mutation Rate	Recombination Rate	Admixture Proportion
ArcHIE	1.25E-08	1.00E-08	0.02
Human-Neanderthal	1.29E-08	1.00E-08	0.0225
African-Ghost	1.20E-08	1.30E-08	0.02
AfricanWSS-Ghost	1.20E-08	1.00E-08	0.02
Bonobo-Ghost	1.20E-08	7.00E-09	0.02
Gorilla-Ghost	1.235E-08	9.40E-09	0.0247

Table 3.1: Mutation, Recombination and Admixture Proportion for applied Models

3.2.1 ArchIE model

This model 3.1 was used for evalution of the logistic regression-based ArchIE tool described in the previous chapter [5] and describes Human-Neanderthal admixture.

It should be noted that in the model simulated via ms in [5] an additional bottleneck within the ghost population is present 3.2; however, this bottleneck is not mentioned in the description of the model. For the analyses, two versions of the model were set up (see fig 3.1 and 3.2).

Furthermore, there is another population present, obviously representing Denisovans. Since this population is not participating in any introgression events, it was simply discarded for the analyses presented in this thesis.

In the original ArchIE Model, there is also a fourth population representing Denisovans. However, this population is not involved in any migration or introgression events and thus can be safely ignored.

3.2.2 Human-Neanderthal model

This three-population model 3.3 including an African, a European and a Neanderthal population describes Neanderthal gene flow into Europeans. The Yoruba population is used as outgroup. In contrast to the ArchIE model, it does not assume constant population sizes for the target population 3.3.

It was proposed in [42], in which genomatnn, a CNN approach for detecting adaptive introgression, is presented.

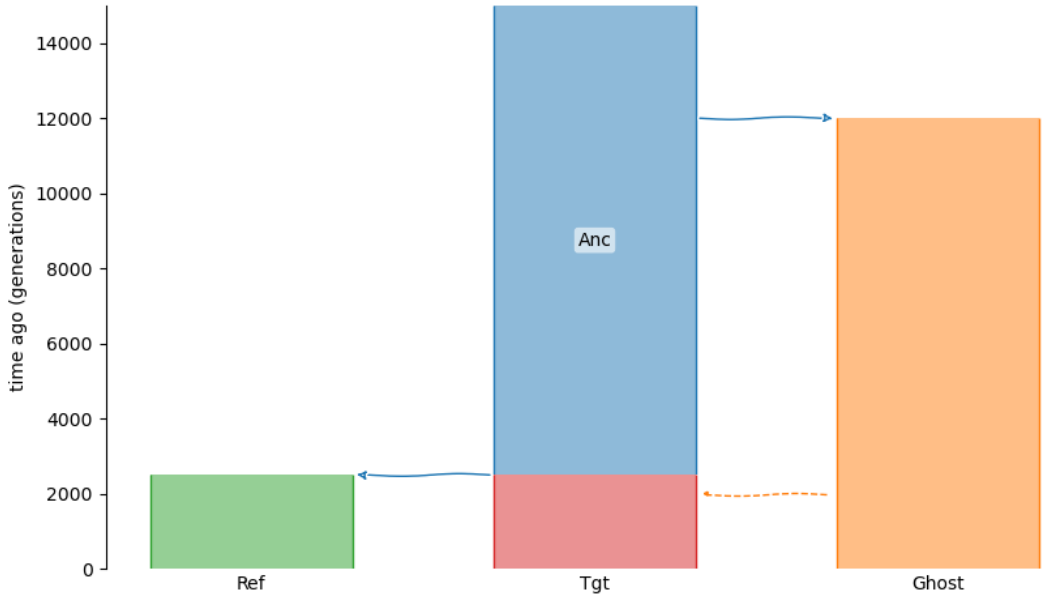

Figure 3.1: ArchIE model without bottleneck

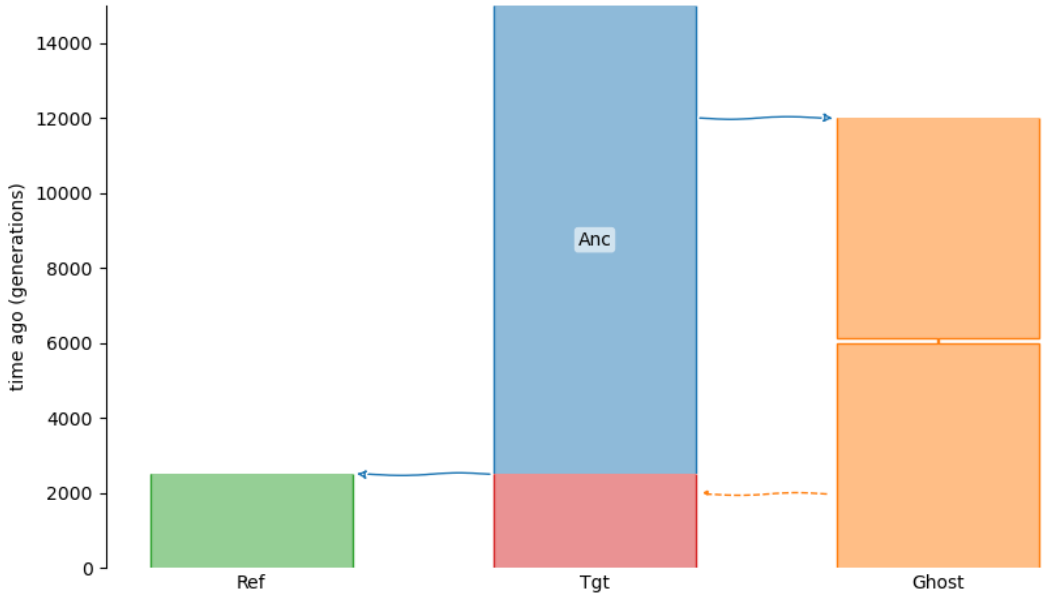

Figure 3.2: ArchIE model

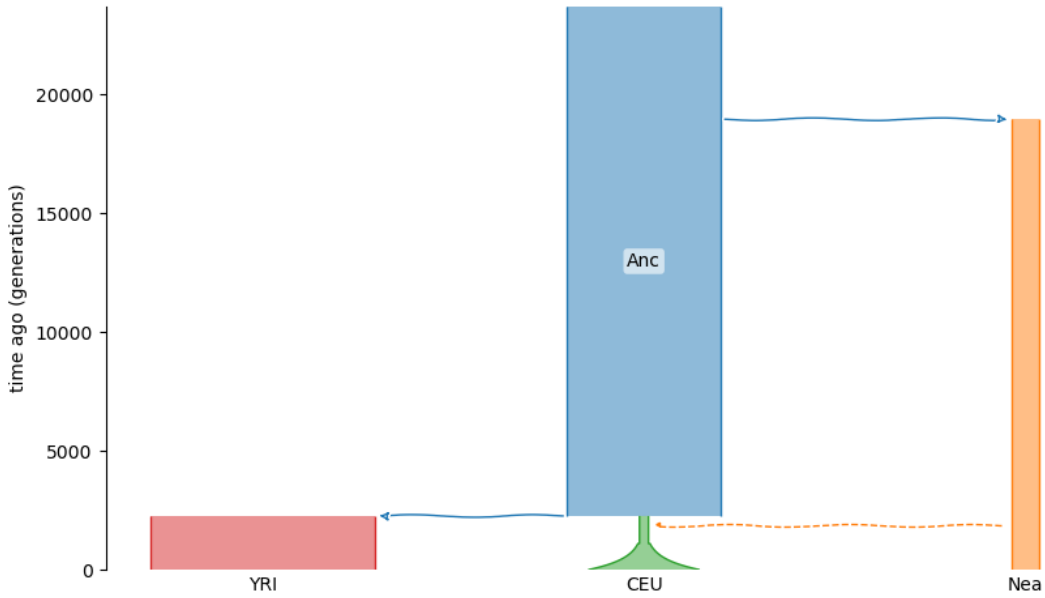

Figure 3.3: Human-Neanderthal model

3.2.3 African-Ghost model

In [32], the genetic contribution of archaic hominins to the variation in recent African populations is studied. ArchIE was applied to detect archaic ghost ancestry in these populations.

The authors report findings of ghost admixture in four West African populations. It is estimated that these populations obtained between 2 and 19% of their genetic material from an archaic population before the split of Neanderthals and modern humans occurred.

In the simulations done for the analyses in this report using this model, an admixture of 2% is assumed.

3.2.4 Weakly Structured Stem model

In [47], a weakly structured stem model for archaic populations in Africa is proposed3.5. In contrast to models of archaic hominin admixture, 1–4% of genetic diversity in modern humans is attributed to genetic drift between these stem populations.

In the model involving ghost introgression applied in the next chapter, exchange between the unknown population and the Mende population in Sierra Leone is modeled.

3.2.5 Bonobo-Ghost model

The Bonobo-Ghost model 3.6 represents introgression of an extinct ape lineage into bonobos [24] which happened more than 1 million years ago. In [24] two of the tools

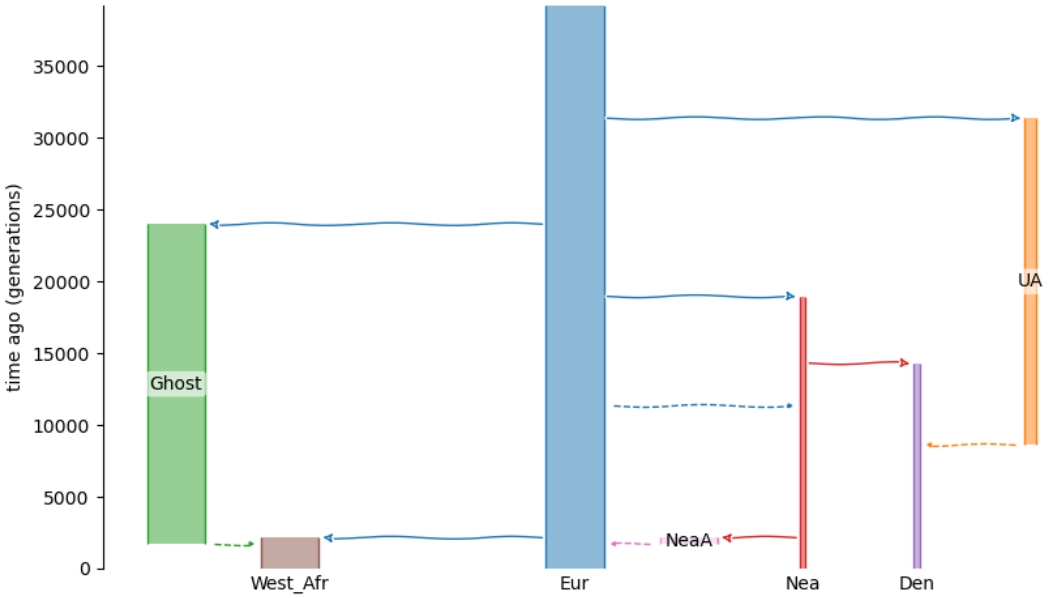

Figure 3.4: African-Ghost Model

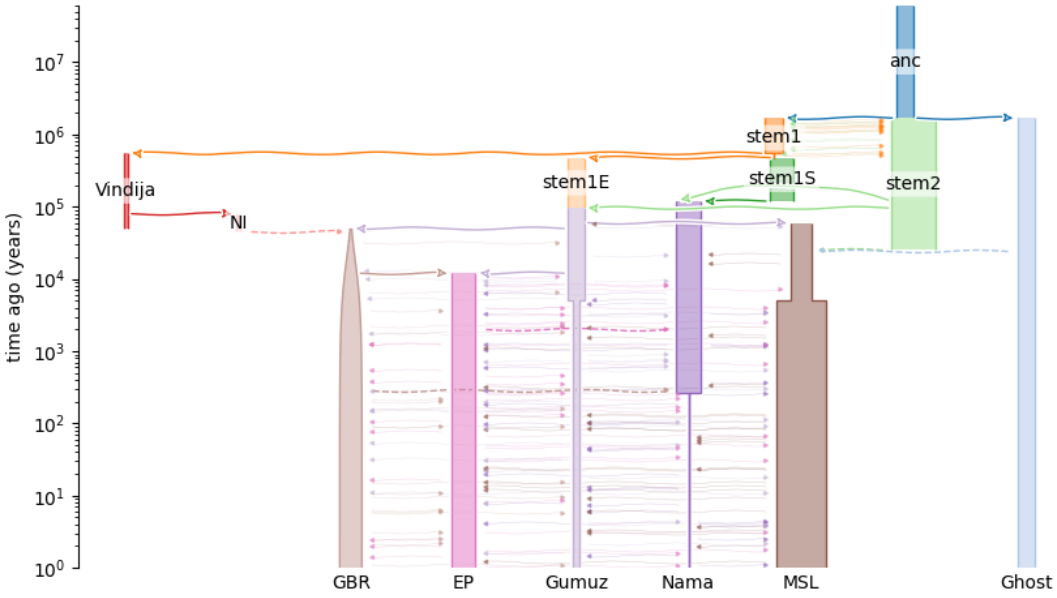

Figure 3.5: African-WSS Model

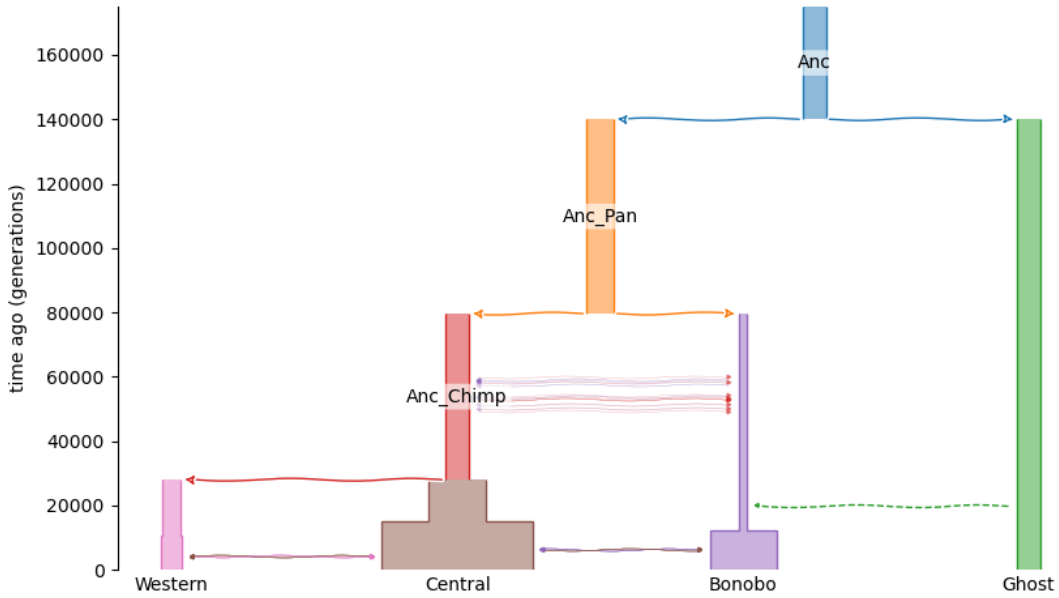

Figure 3.6: Bonobo-Ghost model

mentioned above, Skov HMM / HMMix and sstar, were applied for investigating admixture in bonobos.

3.2.6 Gorilla-Ghost model

Furthermore, a model of ghost introgression in Gorillas 3.7 was set up [25]. For this model, no new results are presented; however, its structure led to necessary changes in the simulation and introgression extraction code (see next chapter).

In [29], evidence for admixture from a ghost lineage into the common ancestor of eastern gorillas, but not western gorillas, which occurred before the mountain and eastern lowland gorillas split approx. 10,000 years ago, was reported. Approximate Bayesian computation with neural networks was applied for modeling of demographic models. This approach revealed signatures of admixture from an archaic ghost lineage into the common ancestor of eastern gorillas. Up to 3% of the genome of eastern gorillas is introgressed from an archaic lineage that diverged more than 3 million years ago from the common ancestor of all extant gorillas. For detection of introgressed windows, also SkovHMM (an older version of HMMix) and an implemention of the S*-statistic have been applied. [25]

The full DEMES model exhibits a highly intricate structure because it also involves the two extant subspecies of the eastern gorilla, the eastern lowland gorilla (ELG) and the mountain gorilla (MG). as well as the two extant subspecies of the western gorilla, the western lowland gorilla (WLG) and the Cross River gorilla (CR).

Depending on the introgression event of interest (*e.g.* Ghost introgression into the

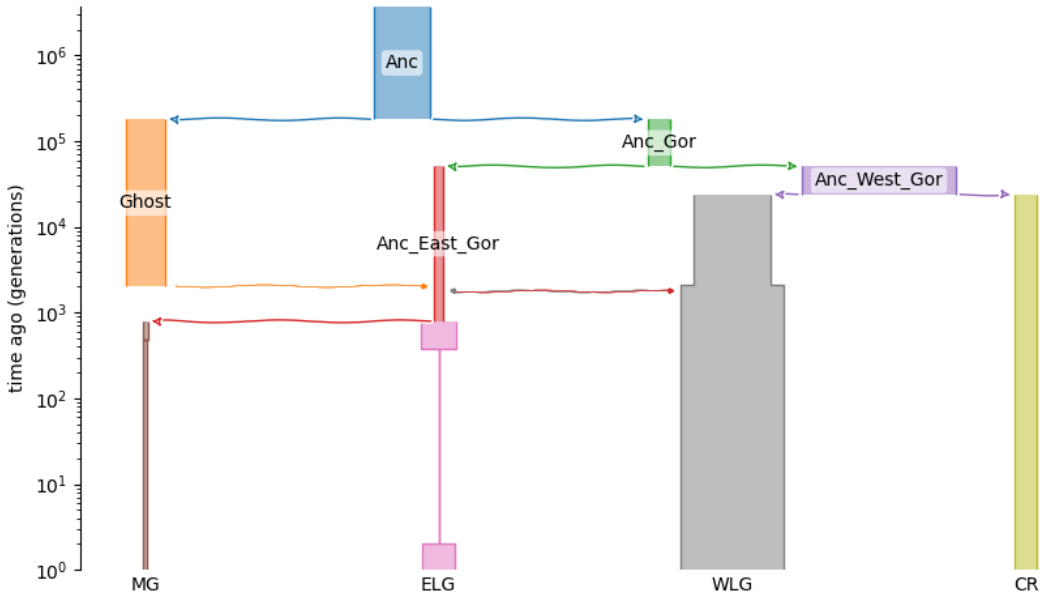

Figure 3.7: Gorilla-Ghost Model

Ancient Eastern Gorilla), there are many migration / introgression events and even full branches (in the aforementioned case the split of the Ancient Western Gorilla into WLG and CR) which can be discarded. If one is interested in introgression concerning the ancient eastern gorilla, there is also a further difficulty compared to the other outlined models: Whereas in the other models introgression in currently existing populations is described, here the ancient eastern gorilla received genetic contributions from the ghost lineage.

4 Implementation details

Logistic Regression (and also additional classifiers) for ArchIE were implemented using the python-packages scikit-learn and statsmodels.

For the neural network architectures, Pytorch was used [48] (which is also used for the original IntroUNET implementation). In the IntroUNET implementation, central functions from the original implementation were adopted, but heavily modified to allow application in a more flexible and customized way.

The code was modified so that it works with different mutation models (in addition to the Jukes-Cantor model also with a binary mutation model distinguishing only between two states instead of the four bases of the Jukes-Cantor model).

Simulations for training and test data are created using msprime, a coalescent simulator based on tskit [49]. Coalescent simulators are founded in coalescent theory (Kingman 1982) which traces back the evolutionary history of lineages until the point of time in which they coalesce in their most recent common ancestor (MRCA). Therefore, in contrast to forward simulators, coalescent models trace the genealogies of the sampled individuals from the present to the past. The focus is on when and how the genetic lineages of individuals in the sample coalesce (i.e., share a common ancestor), so only the genealogies for the sampled individuals are simulated, ignoring the rest of the population. For tree sequence processing - which is in particular necessary for the determination of the introgressed regions within the simulated data, tskit is used [50].

For many ML-algorithms, imbalanced data, i.e. one of the classes or labels of the input data are much rarer than others, can lead to heavily impaired performance and prolonged training time. The introgression tasks contain such imbalance data, because usually the regions of introgressed material are only a very small fraction of the complete genetic material. In [6], it was claimed that the results from the ArchIE paper could only be replicated by using balanced data (which was not confirmed in the original publication). Thus for ArchIE also models with balanced data have been trained; for obtaining the required amount of data, it was necessary to simulate until enough feature vectors with sufficient introgressed material had been simulated.

4.1 Modifications of the general workflow

4.1.1 Snakemake pipelines

For all the tools presented, Snakemake pipelines [51] were implemented. In some cases, previous versions from [3] were modified. In particular, it was paid attention that there are no hardcoded variables (*e.g.* for the number of individuals or mutation and recombination

rates), so that the pipelines can be adapted to new models without modifying the code of the involved rules.

The input for training and testing of the tools consists of simulations with msprime. Variant Call Format (VCF) is supported as input. This is a standardized text file format which stores the variation between gene sequences, for instance SNPs. This format allows a very straightforward creation of genotype matrices, as for IntroUNET, and the calculation of features of interest which is necessary for tools like ArchIE.

Additionally, for the simulations also tree sequences are created. Tree sequences do not directly store all SNPs, but a succinct tree sequence which represents the ancestral relationships between a set of DNA sequences. For many tasks, this representation is more suitable, *e.g.* it requires much less memory. Tree sequences are based on fundamental biological principles of inheritance, DNA duplication, and recombination (due to recombination events, i.e. exchange of material between different molecules of DNA, nucleotides from the same DNA sample have different paths of genetic inheritance).

The efficient processing of tree sequences is possible using packages like tskit. For the benchmarking workflow, it is necessary to extract the introgressed features from the tree representation and map it onto the SNPs (see below).

4.1.2 Extraction of migration events

After the creation of simulation data, the introgressed tracts must be extracted so that this information about the ground truth can be used in the training process and for evaluation. In particular, in the Gorilla-Ghost model, these additional migration and mutation events lead to a tremendous computational overflow.

Therefore, a more efficient version has been implemented which creates a simplified tree only involving relevant migration events and supports parallelization of the tree processing. Since the tskit package does currently not allow simplification of tree-objects containing migrations, a copy of the original tree is created and all migrations are removed from this object. Afterwards, simplification can be applied and, using the information from the original tree, only the relevant individuals belonging to the populations of interest and involved in migration events, are retained. For demographic models like the Gorilla-Ghost model, this simplification process leads to a tremendous reduction of individuals and migration events which have to be taken into account.

Batches with indices of all remaining objects are created, either looping over all trees or all migration events. The optimal batch size is determined by taking into account the number of indices and the available cpus.

The remaining trees are shared across processes using the python multiprocessing-package. Batches with indices of all remaining objects are created, either looping over all trees or all migration events, so that no copy of the full tree object is necessary and memory usage remains almost constant. Finally, the output is written to a shared list.

4.2 Integration of the latest version of HMMix

The latest released version of HMMix has major changes and a completely overhauled interface compared to the version used in the performance benchmark in [3]. Therefore, the snakemake pipelines for evaluation of HMMix had to be fully reworked. In particular, the new workflow allows automatic identification of the population sizes given in the msprime simulations without need for hardcoded variables.

Furthermore, the original code was modified so that it works with the Jukes-Cantor-Model as well as with the BinaryMutation-Model.

4.3 Reimplementation of ArchIE

For ArchIE, the functionality of the original implementation has been extended by allowing a custom selection of features and the application of regularization for logistic regression as well as the application of other classifiers.

Instead of R as in the original implementation, all parts - feature extraction, training and inference - are done in Python.

It should be noted in advance that the features which size depends on the sampled individuals do not contribute much to performance. In particular, the pairwise differences are problematic since their implementation probably is not really appropriate for logistic regression. Therefore, the new implementation also allows a very flexible selection of individual feature combinations.

4.4 Reimplementation of IntroUNET

The code for the reimplementation of IntroUNET can be found on GitHub (https://github.com/jalhackl/introunet).

Issues with the original IntroUNET implementation

Inspecting the original IntroUNET code, some crucial problems for replicating the results - or even application on new data - have been found. In particular, the random number generator is not correctly implemented, giving rise to exact copies of the input data when the simulations are created in batches.

Furthermore, information about the fourth population which is simulated via ms, but does not participate in any introgression pattern, is not fully eradicated from the training and test data. Those additional polymorphisms are potentially biasing the results since they can act as introgression pattern.

Extensions of IntroUNET

The original IntroUNET-implementation relies on ms and slim simulations (concerning ms, a modified version is used), whereas the reimplementation directly processes vcf-files. This corresponds to the workflow of the other tools and thus enables direct comparison.

Furthermore, analogous to the ArchIE reimplementation, improvements of the program structure have been made which, amongst others, allow direct parallelization of all steps.

Furthermore, some problems with the official IntroUNET code have been observed, and the processing of the random seed generator is flawed. The code for the neural network architecture was adapted in such a way that it works for input data of different size.

For smaller numbers of individuals, convolutional layers can be removed, whereas in the opposite case larger dimensions can be allocated. In each case, the number of individuals is upsampled accordingly to fit in the input layer with as few upsampled individuals as possible.

The corresponding functions in the reimplementation of IntroUNET create exactly these three arrays from vcf-/bed-file input, which is provided by the msprime simulations. Afterwards, the output is passed to the IntroUNET-functions which create the HDF-files.

The IntroUNET reimplementaion also allows the handling of phased and unphased data. In the latter case, some addtional changes were necessary: For unphased data, the variant arrays are coded as 0 (homozygous non-derived / ancestral), 1 (heterozygous) and 2 (homozygous derived).

For input data of different size, parameters for selecting the correct input dimensions have been added so that for larger input the U-Net is enriched with additional layers and for smaller input only the necessary size is chosen.

4.5 Evaluation of predictions

The metrics primarily used for evaluation of the predictions are precision and recall, given by

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.1}$$

and

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{4.2}$$

Precision-Recall curves plot precision against recall for various thresholds.

Since the introgression tasks are heavily imbalanced, i.e. the positive class is very rare, precision and recall are much more insightful and reliable than receiver operating characteristic (ROC) curves, plotting the true positive rate against the false positive rate, or accuracy values. In some cases, additionally confusion matrices are provided. An optimal cut-off was determined using the F1-Score given by

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

.

The evaluation functions for computation of Precision-Recall curves had to be modified so that they also output correct results for small chromosomes for which it can happen that there is no introgression present or detected (whereas in the previous analysis using the Snakemake files only chromosomes with 200MB were evaluated so that no problems with undefined precision or recall values occurred).

5 Results

In a first step, the results obtained by [3] were successfully replicated (see fig. 5.1) for the Human-Neanderthal model. Test data of the same size (100 replicates of 200MB chromosomes) were also used for runs with the ArchIE model (fig. 5.1).

It seems that the ArchIE model is much more challenging for these tools than the Human-Neanderthal model. All tools perform rather poorly; in general, the performance with 50 reference individuals on the ArchIE model is worse than for 10 reference individuals on the Human-Neanderthal model.

5.1 ArchIE results

In this section, plots without additional specification always show the results for the ArchIE model. Except for fig 5.2, in which 100 replicates of 200MB chromosomes are assessed, the evaluation was done on 1000 replicates of 1MB chromosomes.

Fig. 5.2 shows that ArchIE performs comparatively well on both demographic models.

On the ArchIE model, ArchIE even performs better than all other tools at all cut-offs using 50 reference individuals. Similarly, on the Human-Neanderthal model, 10 reference individuals are sufficient to get a performance only moderately worse than the other tools with 50 reference individuals. Using 50 reference individuals, ArchIE outperforms all other tools (see fig. 5.2).

Interestingly, runs with balanced data do not improve performance (see fig. 5.3).

5.1.1 ArchIE feature importance

For a more refined assessment of the importance of individual features, ArchIE models were trained with various combinations of features.

Similar to the results in [5], the minimum distance to the reference and the number of private SNPs are crucial features. Indeed, for most settings, the number of private SNPs is the most important feature by far. For most models, the S*-feature was only of modest, but significant importance.

The weights associated with the features were analyzed. For this purpose, three versions of the coefficients were computed: the raw coefficients, the standardized coefficients in which the variances of the variables are scaled to one and thus its value provides information about magnitude and direction of their influence on the outcome, and the logarithm of the absolute value of these standardized coefficients which can be directly interpreted as the magnitude of the feature importance. For the standard ArchIE model with nref/ntgt=50/50, fig. 5.4 shows raw coefficients for all features for two different

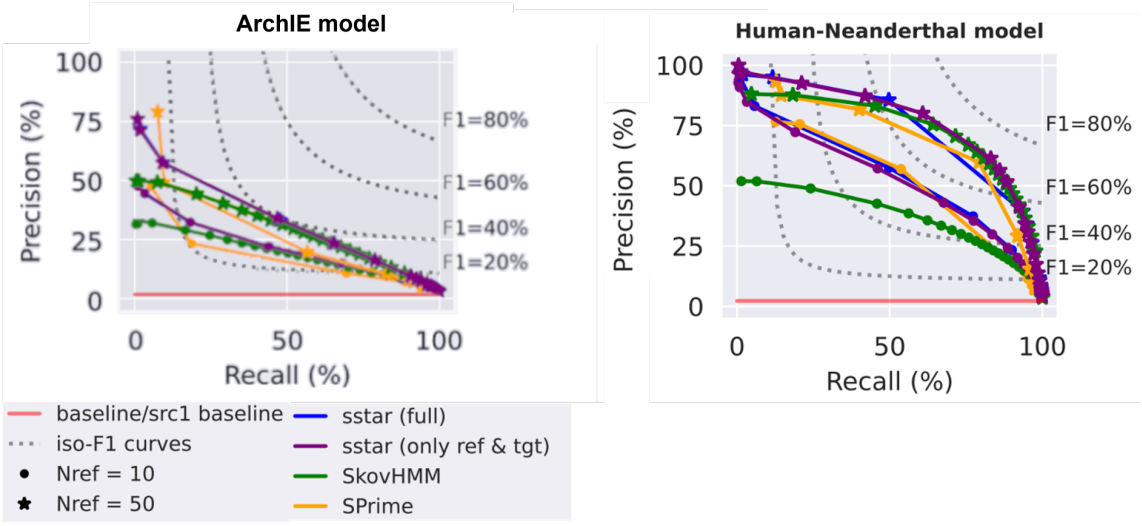

Figure 5.1: Precision-recall curves for the performance of various models on the ArchIE model (left) and the Human-Neanderthal model (right): sstar, SkovHMM/HMMix (SkovHMM indicates an older version of HMMix) and Sprime. Nref indicates the number of individuals for which reference genomes were available, the iso-F1 curves connect points with the same F1 score.

trained models, fig. 5.4 shows the logarithm of the absolute value of the standardized coefficients for the same models. The weights of the fixed-size features are relatively stable, whereas the weights of the frequency spectra and the pairwise differences between both runs show only agreement regarding the general pattern.

In 5.6, the feature importance on the ArchIE model with and without L1 regularization is compared (standardized feature coefficients). The parameter C tunes the strength of the L1 penalty, a lower value means stronger regularization. One sees that most of the pairwise distances are set to zero, in particular, the middle region is thinned out; however, there is, except for the first few distances, no clear pattern which are maintained and which are eliminated, but it seems rather to be random. In particular, for stronger regularization (C=0.1), the sparsity of the coefficients for the pairwise differences is obvious; however, even for such strong regularization, no impairment in performance was detected. Similarly, for the frequency spectra, there is a clear pattern, but again the importance of individual features is not stable. On the other hand, the fixed-size features show a very stable importance also in the case when regularization is applied.

5.7 shows the normalized feature importance for the fixed-size features only on the ArchIE model.

Fig. 5.8 shows the drop of performance when also further fixed-size features are removed. The results of this ablation study reveal that the number of private SNPs and the minimum distance to the reference are indeed the most important features. The number of private SNPs is the feature which causes the largest drop in performance when

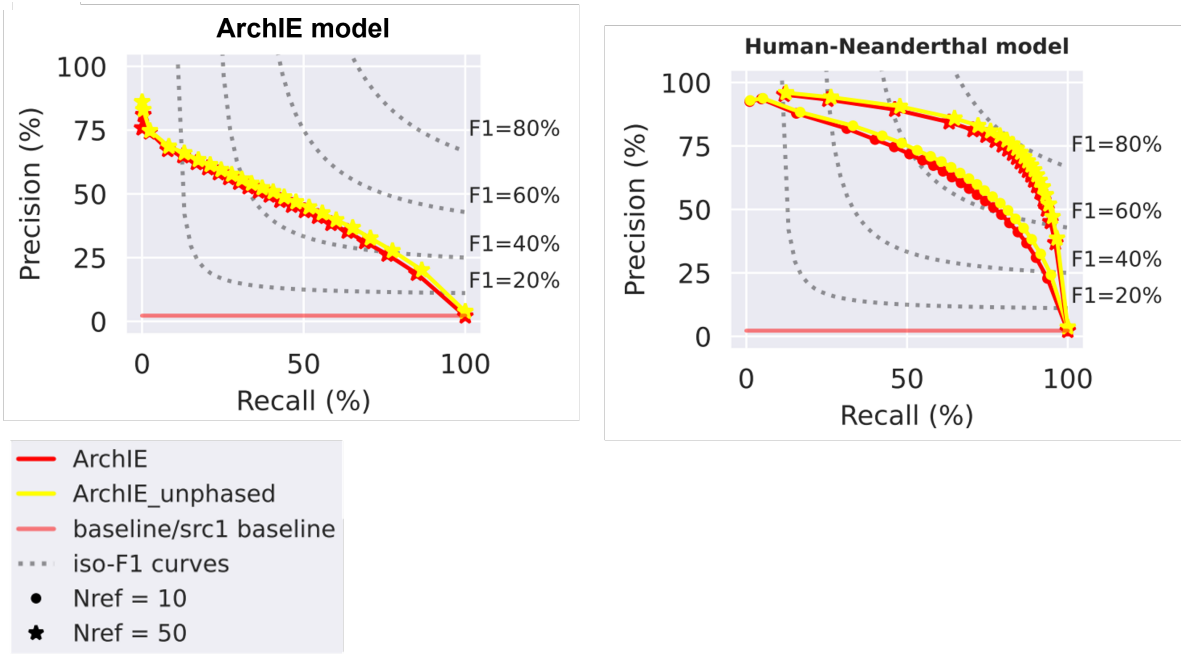

Figure 5.2: Precision-recall curves for the performance of ArchIE on the ArchIE model (left) and the Human-Neanderthal model (right): ArchIE (red) indicates the usual evaluation of introgressed tracts on each haplotype, ArchIE_unphased (yellow) shows the performance when phased haplotypes are merged for evaluation (i.e. if only one haplotype contains introgression at the specific position and ArchIE classifies one of both haplotypes as introgressed, the prediction is counted as correct). Nref indicates the number of individuals for which reference genomes were available, the iso-F1 curves connect points with the same F1 score.

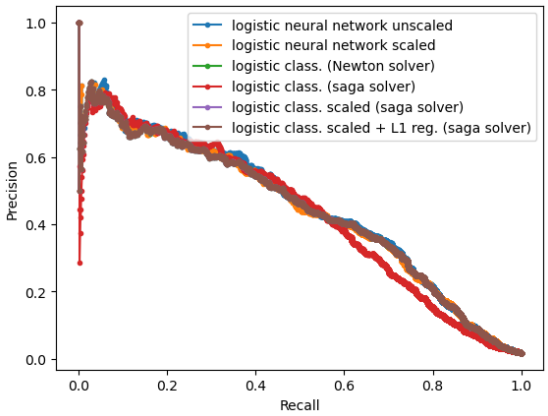

Figure 5.3: Results for the ArchIE model with standard settings (nref/ntgt=50/50) for different solvers

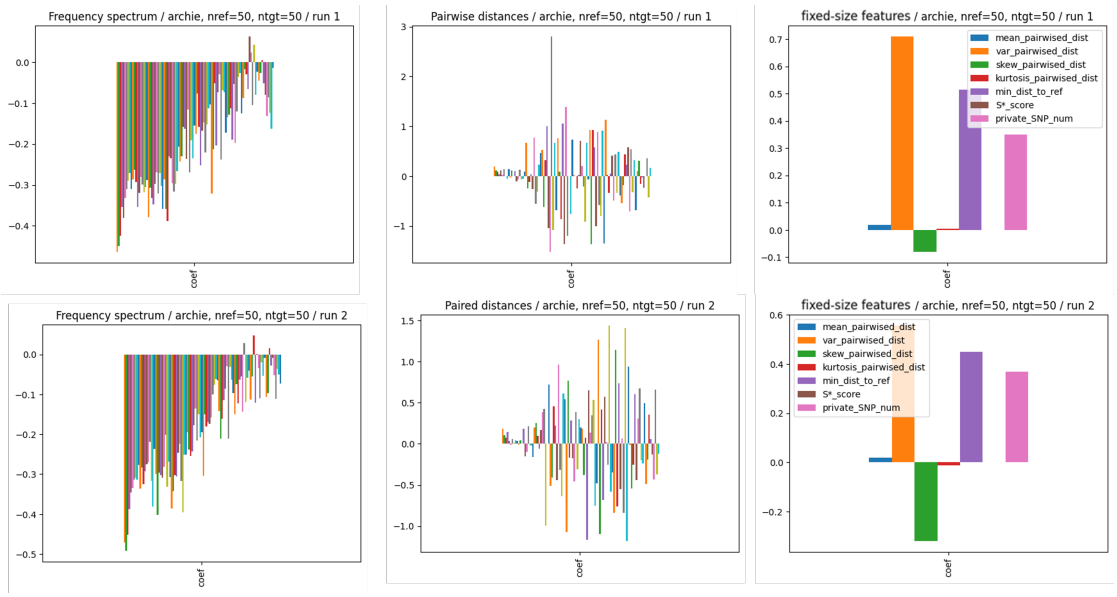

Figure 5.4: Raw coefficients for two trained models with the ArchIE model

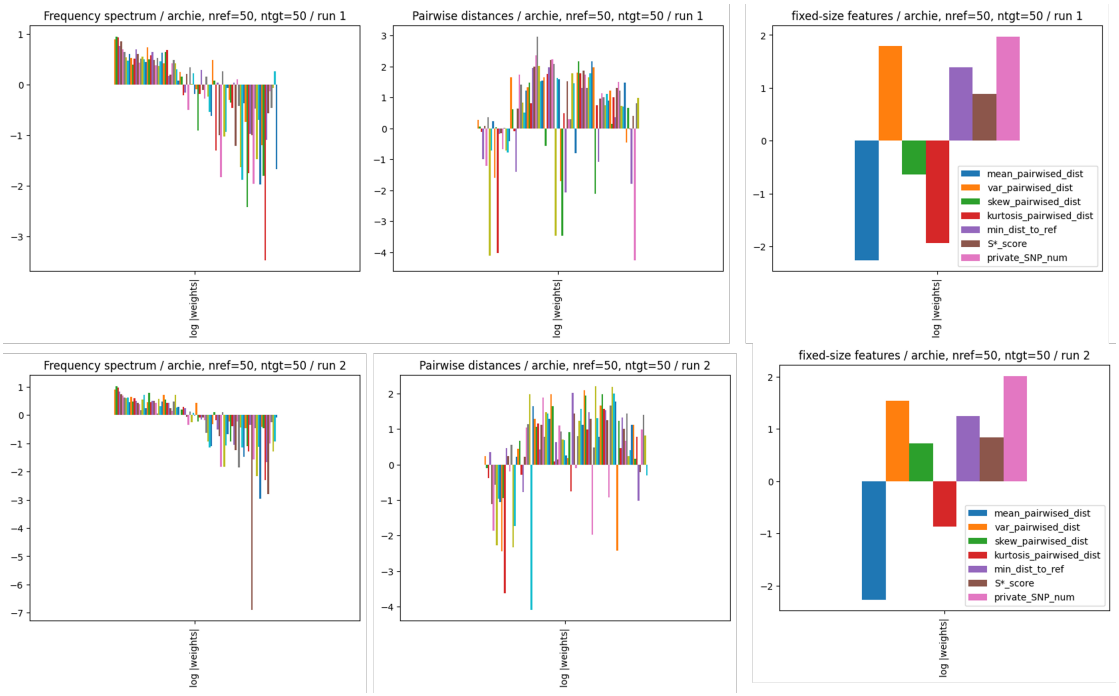

Figure 5.5: Logarithm of the absolute value of these standardized coefficients for two trained models with the ArchIE model

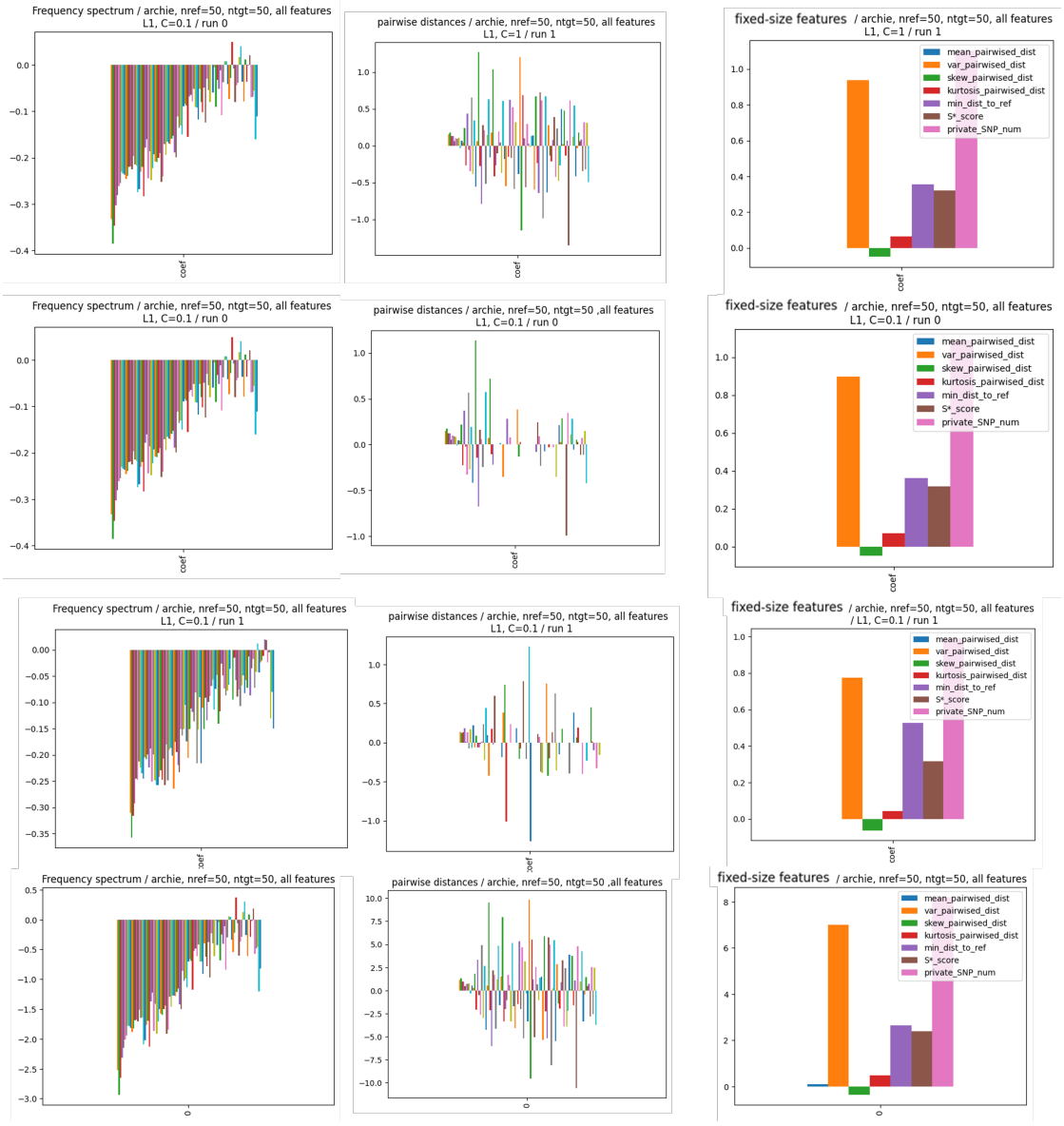

Figure 5.6: Feature importance for ArchIE on the ArchIE model with different regularizations; First row: L1 regularization, C=1; two and third row: two independent runs with L1 regularization, C=0.1; fourth row: no regularization; first column: frequency spectra; second column: pairwise differences; third column: fixed-size features; the parameter C tunes the strength of the L1 penalty, a lower value means stronger regularization.

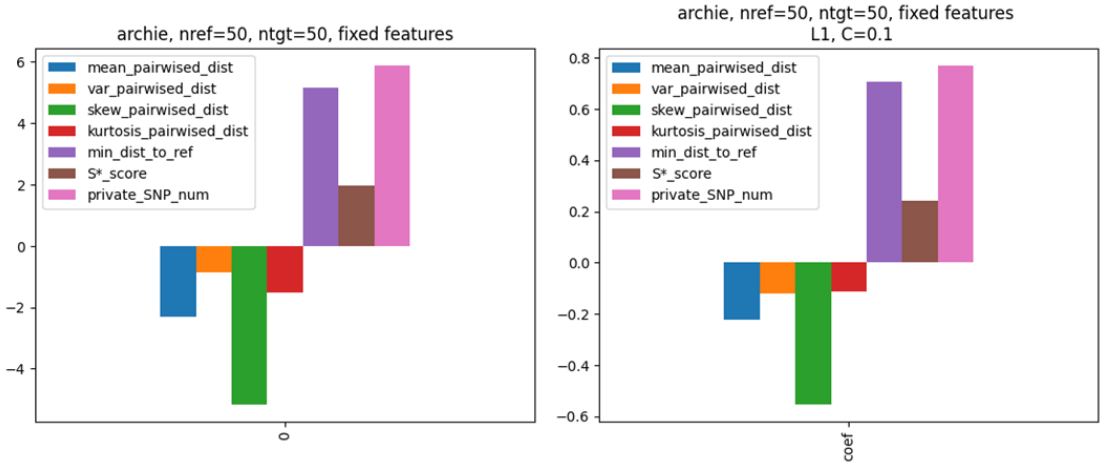

Figure 5.7: Feature importance for ArchIE with fixed-size features only; left: no regularization, right: L1 regularization, C=0.1

it is the only feature removed from the fixed-size feature set.

The S*-score has a smaller, but nonetheless significant impact; in particular, when it is removed additionally to the number or private SNPs, the performance decrease is clearly visible. In accordance to expectations, the feature importance of the pairwise distances is very small and the magnitude of the assigned weights does not exhibit a clear pattern, but rather fluctuates randomly. 5.9 shows the difference in performance with and without the pairwise distances.

There are visible differences between the demographic models, - and due to the collinearity of features also between runs on data from the same model -, however the ranking of the most important features is relatively stable. For a low number of target individuals, the collinearity of the data is even more problematic because, e.g., the mean of the paired differences equals the second pairwise difference divided by two and the other other pairwise difference equals zero (since there are only two haplotypes in this case). In such cases, the distribution of weight is arbitrary and will differ between trained models.

Fig. 5.10 also shows the feature importance for simulations with only one target individual for the ArchIE and the Human-Neanderthal model. There are differences, but again, the most important features differ only slightly.

5.1.2 Reliability of results

Fig. 5.11 shows the results for 31 trained models with all data, 5.12 for 45 models with fixed-size features only. In both cases, the results are relatively robust.

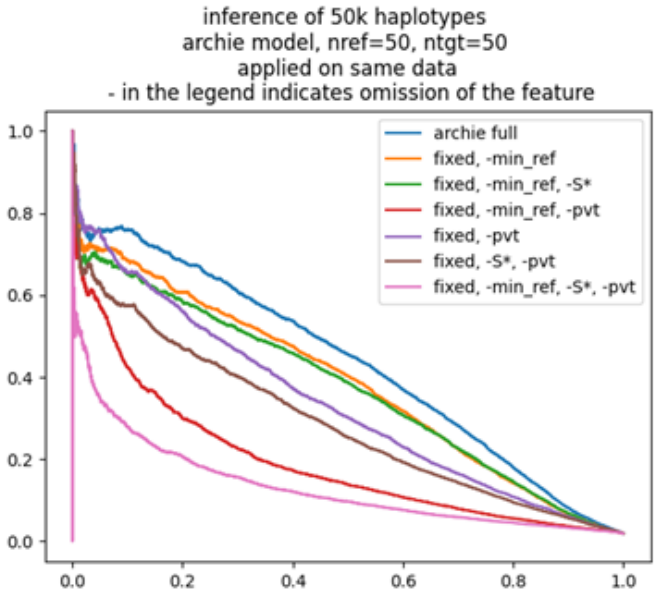

Figure 5.8: Ablation study of the ArchIE features: full feature set (blue) and subsets in which the features of dynamic size and additional fixed-size features are removed; x-axis: precision, y-axis: recall

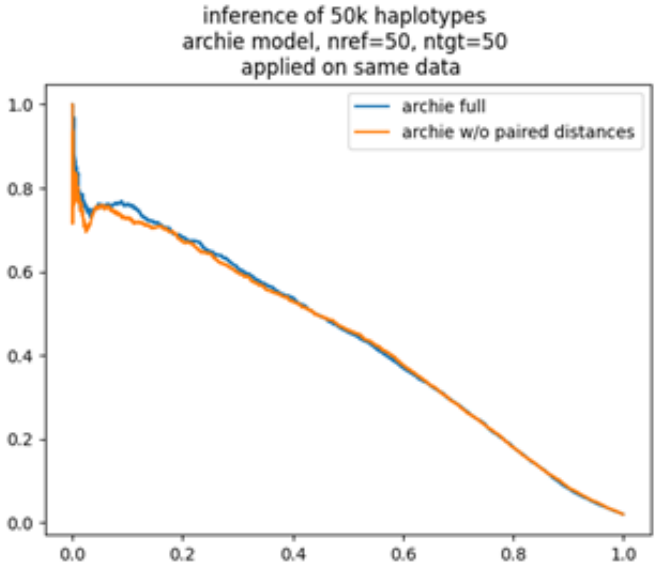

Figure 5.9: ArchIE performance with full feature set and with pairwise distances removed; x-axis: precision, y-axis: recall

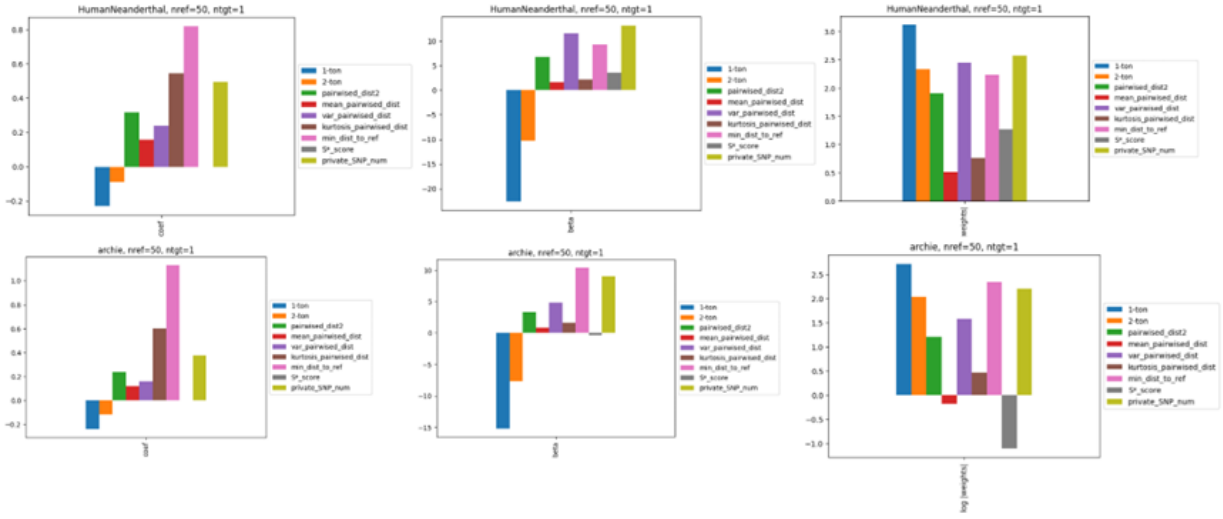

Figure 5.10: Upper row: Human-Neanderthal model, lower row: ArchIE model; First column: raw coefficients, second row: standardized coefficients, third row: logarithm of the absolute values of the standardized coefficients

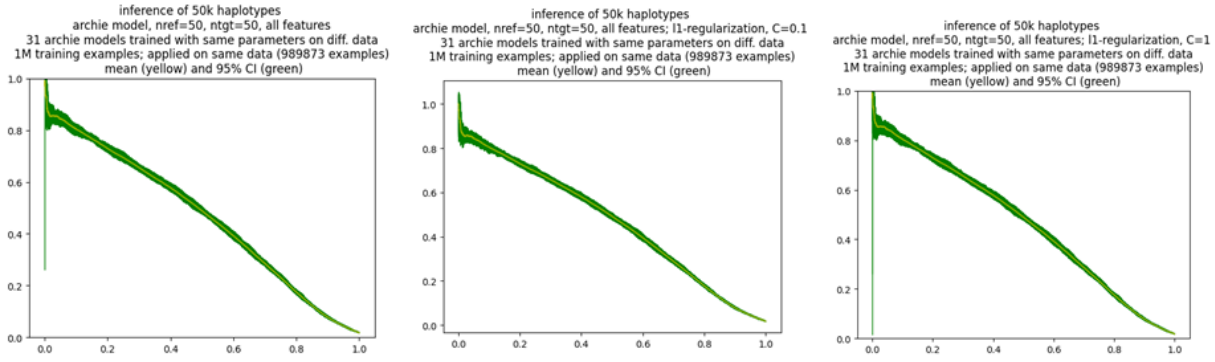

Figure 5.11: Results for 31 trained ArchIE models with all features; x-axis: Recall, y-axis: Precision

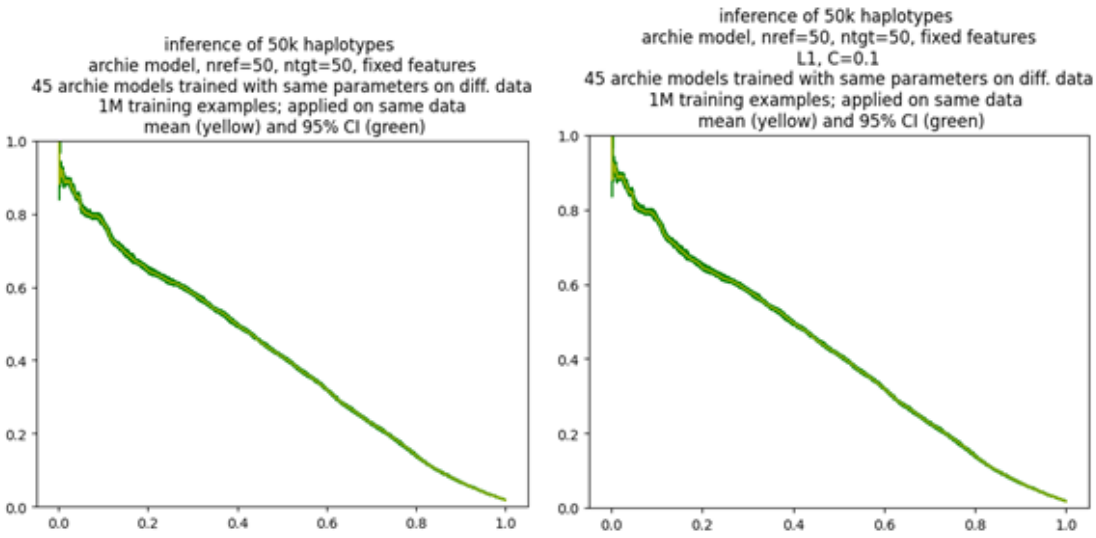

Figure 5.12: Results for 45 trained ArchIE models with fixed features only; x-axis: Recall, y-axis: Precision

Balanced data

Interestingly, for no classifier it is possible to get better results using balanced data. Fig. 5.13 shows results for ArchIE with unbalanced data and two rounds with balanced data.

However, now also scaling seems to be without any impact – with and without scaling, all classifiers have more or less the same performance.

A quite plausible explanation for this rather unexpected behaviour is that one million training instances also provide more than sufficient examples of the positive class, notwithstanding its rarity.

It also seems plausible that the diversity of non-introgressed feature vectors is much higher than those of introgressed ones (which, of course, depends on the models, hence one cannot assume that this result can be generalized), so that only a smaller number of positive class examples is necessary to learn its feature patterns and thus an unbalanced training set with an exceeding number of negative, non-introgressed examples is indeed beneficial.

For fixed-size features only, the results for balanced data are also rather worse.

5.1.3 Different classifiers

It was also tested whether the replacement of logistic regression by a different classifier impacts predictive power, e.g. Extra-Trees [52] or a neural network with more than one dense layer (which is, given a sigmoid function is used as activation function, equivalent to logistic regression). Additionally, a DNN architecture in which the pairwise distances were preprocessed by some additional layers, was applied. In general, the differences in

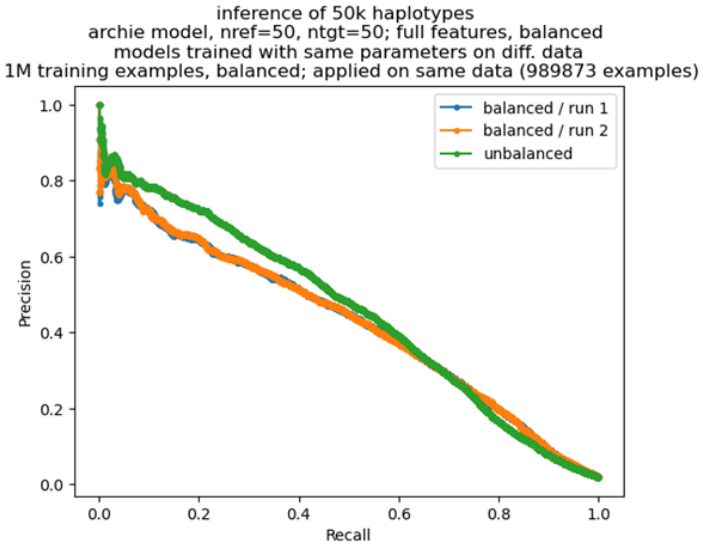

Figure 5.13: ArchIE performance with balanced data

performance are very small (see fig.5.14).

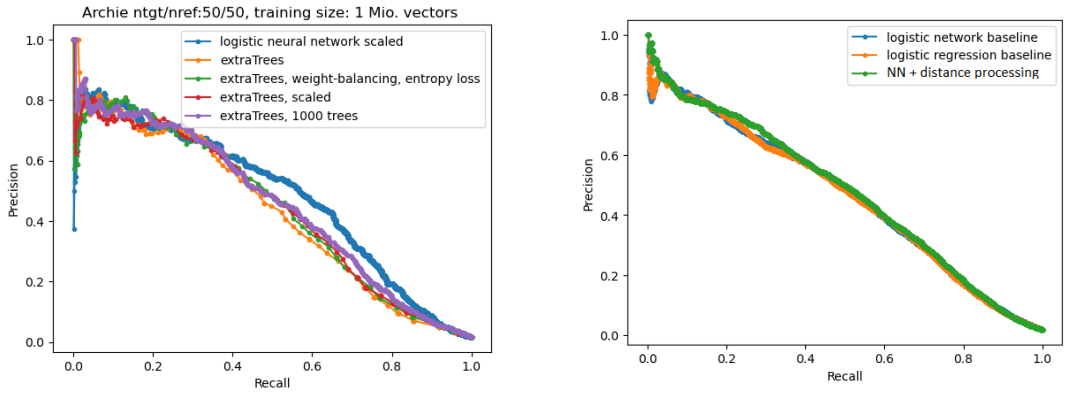

Figure 5.14: Comparison of classifier performance on the ArchIE model: Left: Extra-Trees classifier with different parameters and logistic regression; right: standard logistic regression (orange), logistic regression implemented as neural network with one layer (blue), and neural network with extra layers (green).

5.2 IntroUNET results

IntroUNET was evaluated with the settings used in the original publication, i.e. a number of reference and target individuals of 50 (however, in section 'Performance on different individual sizes' results for different numbers of individuals can be found); hence the performance cannot be directly compared to that of the other tools shown above.

Moreover, one should keep in mind that the Precision-Recall curves are created differently since IntroUNET works on a SNP-resolution - and also the Precision-Recall curves contain information about the predictions for each SNP -, whereas the other tools work using windowing and the evaluation extracts introgressed tracts. For an accurate comparison, one would have to construct continuous introgressed regions from the SNP-results or extract the predicted values for each SNP from the introgressed tracts.

As in the original publication, early stopping was used with a patience of 10, and the maximum number of epochs was set to 100. This maximum number was never reached in the simulations presented here, but in some cases training stopped earlier due to time limits.

In [6], it was tested which distance metric works best for seriation. It turned out that the Dice coefficient metric and Euclidean distance work well. As it is also the default setting in the original implementation, for all experiments the Euclidean distance was employed. In principle, one could apply seriation on the reference as well as on the target individuals. For the results shown, it was always applied on the reference individual matrix.

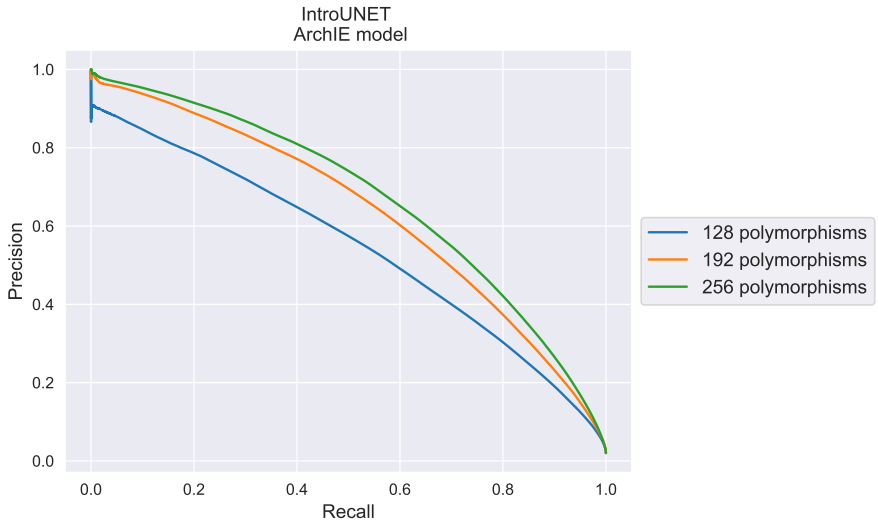

Figure 5.15: IntroUNET results on the ArchIE model (default settings, nref/ntgt=50/50) for windows with different numbers of polymorphisms

5.2.1 Variations of IntroUNET

The default polymorphism window length is 192. Fig. 5.15 shows the influence of the window size on the ArchIE model. Whereas between 128 and 192 polymorphisms there is a remarkable difference, for 256 polymorphisms the performance still increases, but to a much smaller amount. On the other hand, also training time increases for 256 polymorphisms. The time per epoch increases about 30%, but it seems that the training would take much longer than it arrives at the early stopping criterion. However, probably an en earlier stopping of training harms performance not so much, because after a few epochs loss does not change much anymore. (It also should be noted that training time depends on many factors and seems also to differ significantly for the demographic models.)

The default introUNET model does not incorporate any positional information, neither absolute (at which position at the chromosome is the SNP located) nor relative (the distance between two adjacent SNPs).

To account for the importance of relative position, two channels are added: One contains the difference to the previous SNP, the other to the next SNP. (For the first and last SNP of a window, the respective previous or next value is set to zero.) It should be noted that this obviously introduces redundant information; however it is a very straightforward procedure to incorporate the relative positional information fully in a way it is processable by CNNs (whereas for instance, the absolute position is not usable for the convolutional operations capturing local information).

Fig. 5.17 and 5.16 show the impact of the relative position encoding for the Human-Neanderthal and the Bonobo-Ghost model, respectively. For the Bonobo-Ghost model, there is a rather larger performance improvement, for the Human-Neanderthal model it

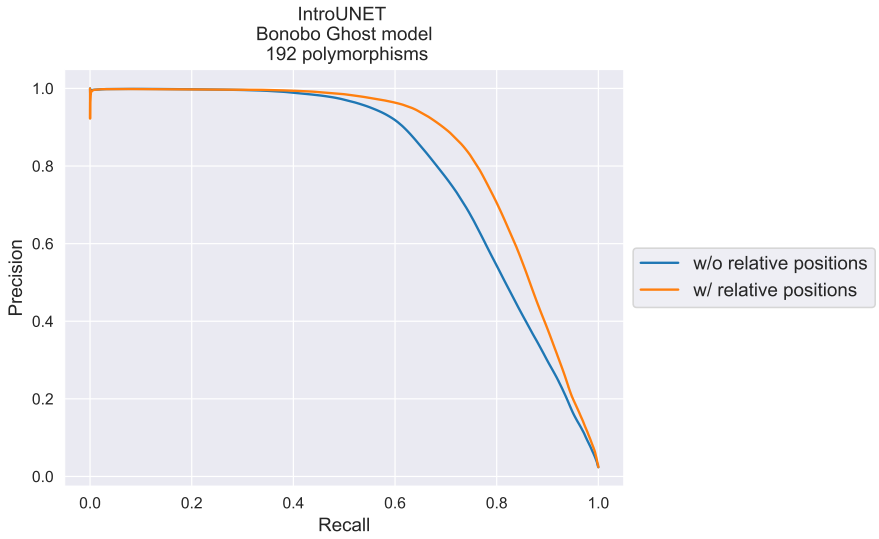

Figure 5.16: IntroUNET results on the Bonobo-Ghost model (default settings, nref/ntgt=50/50, window size: 192 polymorphisms) with and without relative position incorporation

is considerably smaller. Runs for the ArchIE model have shown no improvement.

Some other variants to increase performance have been explored. Other ways of incorporating the positional information did not work: For example, the output of the IntroUNET convolutional layers for each individual, *i.e.* each row of the original output matrix, was augmented with the corresponding positional vector and fed to a LSTM. This did not alter performance; perhaps a fine-tuning of the parameters would be necessary, but it is also reasonable that the providing of the positional information together with the genotype matrix data (in this setting, the positional information is part of the pattern to be detected by the convolutional operations) is simply better suited.

5.2.2 IntroUNET performance under model misspecification

The trained IntroUNET models were applied on the test data from another demographic model to analyze its performance under model misspecification (see 5.18). One sees that for both investigated combinations (train: Human-Neanderthal / infer: ArchIE, train: ArchIE, infer: Human-Neanderthal) the performance drops heavily.

5.2.3 Performance on different individual sizes

Performance on the ArchIE model for reduced numbers of individuals was tested by reducing either the numer of reference or the number of target individuals to two. As also for the default number of individuals, the remaining rows of the input matrix are filled with copies. For a standard input row size of 128, this means that 124 copies of

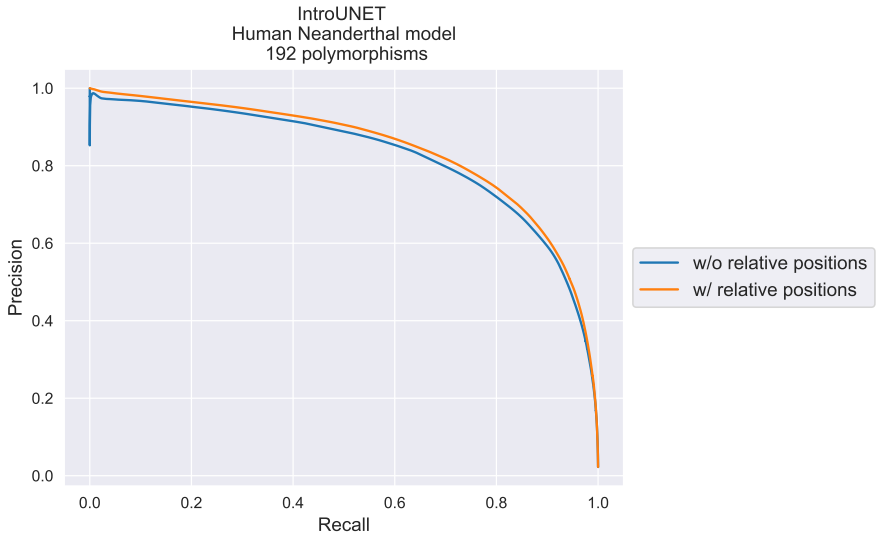

Figure 5.17: IntroUNET results on the Human-Neanderthal model (default settings, nref/ntgt=50/50, window size: 192 polymorphisms) with and without relative position incorporation

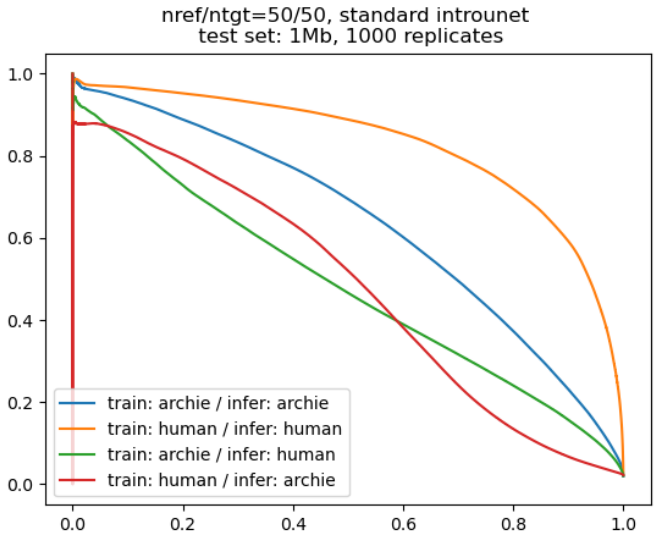

Figure 5.18: IntroUNET performance for misspecified models

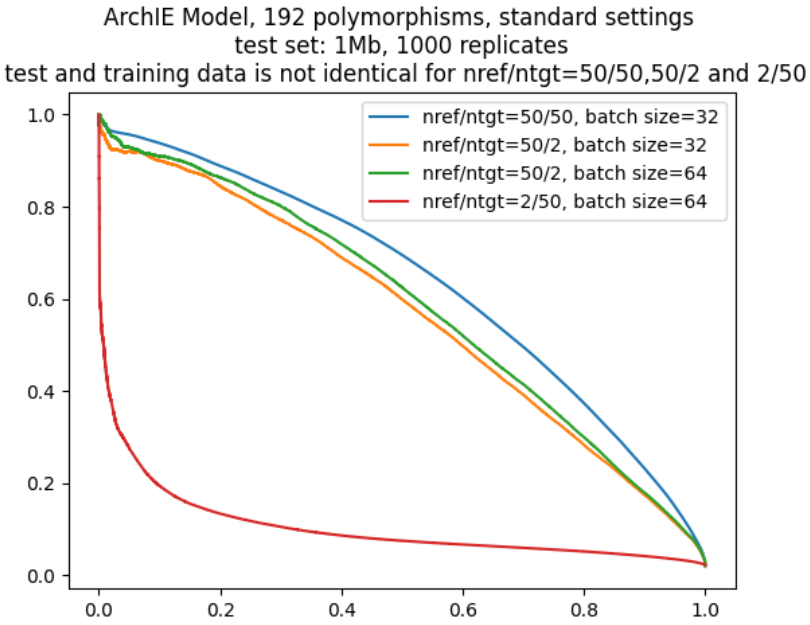

Figure 5.19: IntroUNET performance on the ArchIE model for simulations with reduced number of reference / target individuals. For nref/ntgt=50/2 two runs, one with a batch size of 32, the other with 64, are shown.

the four haplotypes are necessary, whereas in the default case due to upsampling only 28 copies are generated. This suggests that the emerging pattern is highly different and likely leads to significant changes in performance.

For a low number of target individuals, performance is still acceptable, although a drop in performance is visible. In contrast, for only two reference individuals, IntroUNET loses all of its predictive power (see fig. 5.19).

Furthermore, the effect of a larger number of individuals, which also requires increase in CNN dimensions, was investigated. The row dimension was doubled to satisfy the constraint that it has to be a number divisible by four.

In 5.20, the impact of a bigger number of individuals is visualized. In case of the African-Ghost model, the performance with 50 reference and 50 target individuals is already very good and the performance improvement with 99 and 108 individuals, respectively, is small. However, in case of the more difficult ArchIE model, a moderate, but considerably larger, improvement is observed.

5.2.4 Performance on unphased data

For the default IntroUNET architecture with 192 polymorphisms and 50 reference and target individuals, also the performance on unphased data was evaluated on the ArchIE model. In this case, the number of rows in the raw input data is halved, because the two haplotypes of each individual are merged; thus the amount of input is considerably

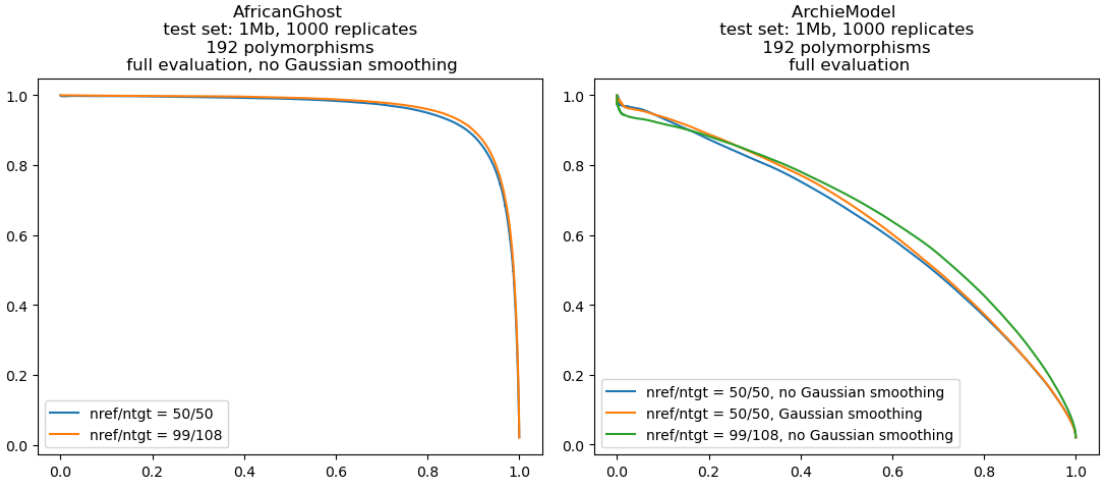

Figure 5.20: IntroUNET performance on the African-Ghost (left) and the ArchIE model (right) for different numbers of reference and target individuals (50/50 and 99/108). Additionally, the positive, but very small effect of Gaussian smoothing is shown for the ArchIE model with nref/ntgt=50/50.

smaller. Nonetheless, IntroUNET seems to maintain much of its performance (see fig. 5.21).

5.2.5 Performance for the African WSS models with and without gene flow

Given the reasonable performance of IntroUNET, it was evaluated whether this method correctly infers introgressed fragments in modern human populations from Africa, which has been proposed before [5]. However, a competing model for the evolutionary history of these populations exists, which is a weakly structured stem (WSS) model, as proposed in [47]. Hence, IntroUNET was also trained with the WSS-model, for which results are presented in this section.

For prediction, several replicates of 200MB chromosomes with and without gene flow were used. For a detailed interpretation of the results, multiple visualization were created: Precision-recall curves and a histogram of the probabilities predicted per class (i.e. the output of the final sigmoid layer of the network) as well as confusion matrices at two fixed cut-offs (0.5 and 0.99) and at the cut-off with the best F1-score. In the case of no gene flow, precision-recall curves and the optimal cut-off are omitted because these values are undefined in the absence of true positives.

The 'probability' distribution peaks between the two classes are clearly distinguishable and almost all predicted probabilities are near 1 or below 0.2 (see fig. 5.22). In the case of data without gene flow, most data points get a very low score, i.e. are definitely predicted as not introgressed (see fig. 5.23). One should note, however, that there is always also a small bar at 1. Even for a cut-off of 0.99 there are false positives; however, given the

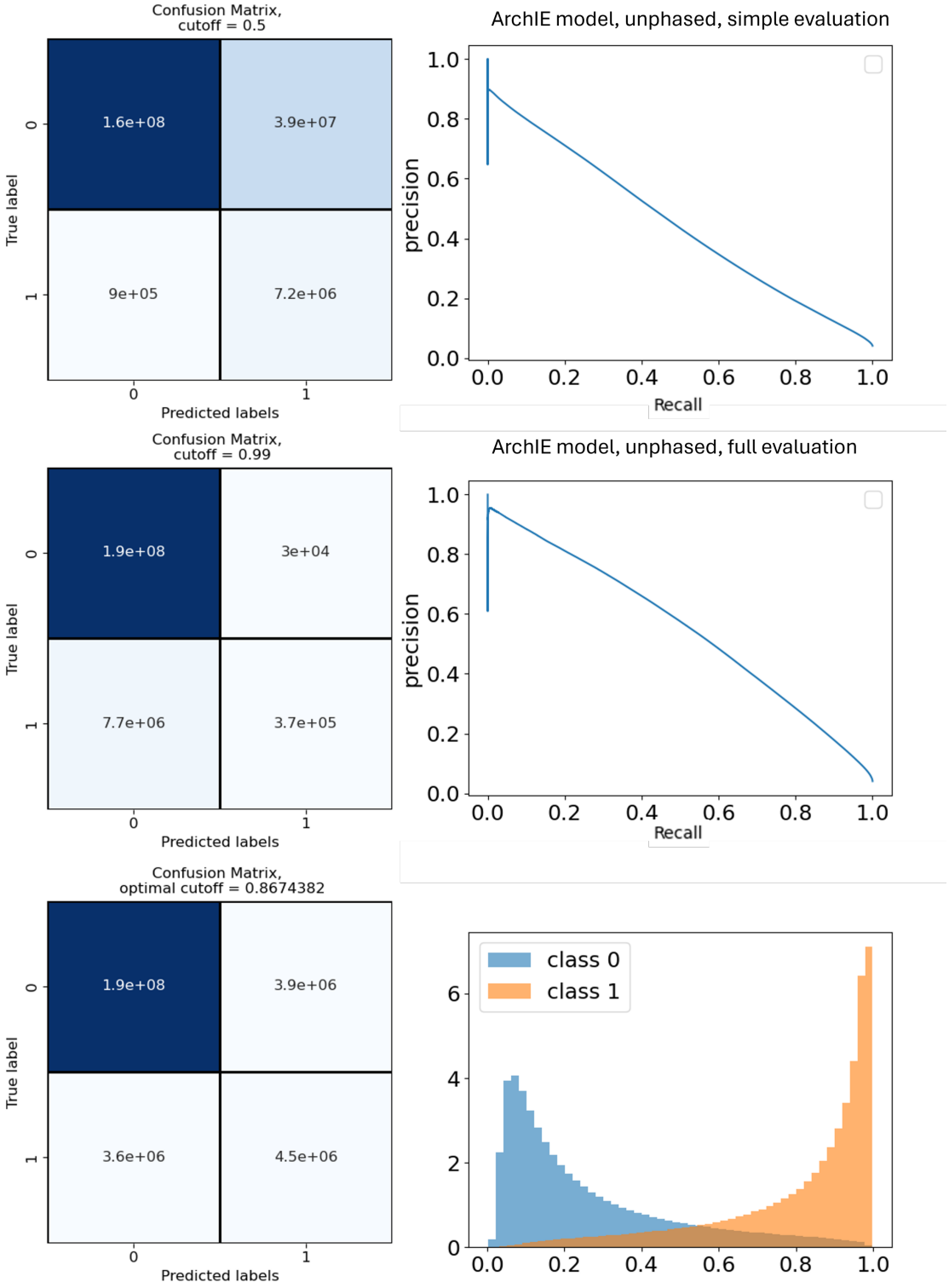

Figure 5.21: IntroUNET performance on the ArchIE model with unphased data, 192 polymorphisms, nref/ntgt=50/50; left: confusion matrices at 0.5 (upper row), 0.99 (middle row) and at the optimal cut-off according to F1-score; right: upper row: Precision-Recall curve for evaluation of each window; middle row: Precision-Recall curve for full evaluation of all windows after averaging; lower row: histogram of the predictions for the full evaluation (class 0: not introgressed, class 1: introgressed)

total amount of evaluated variants, their number is negligible. It is however interesting, that in the data without introgression some variants get such a high value, although most are correctly predicted as not introgressed, i.e. get a very low value.

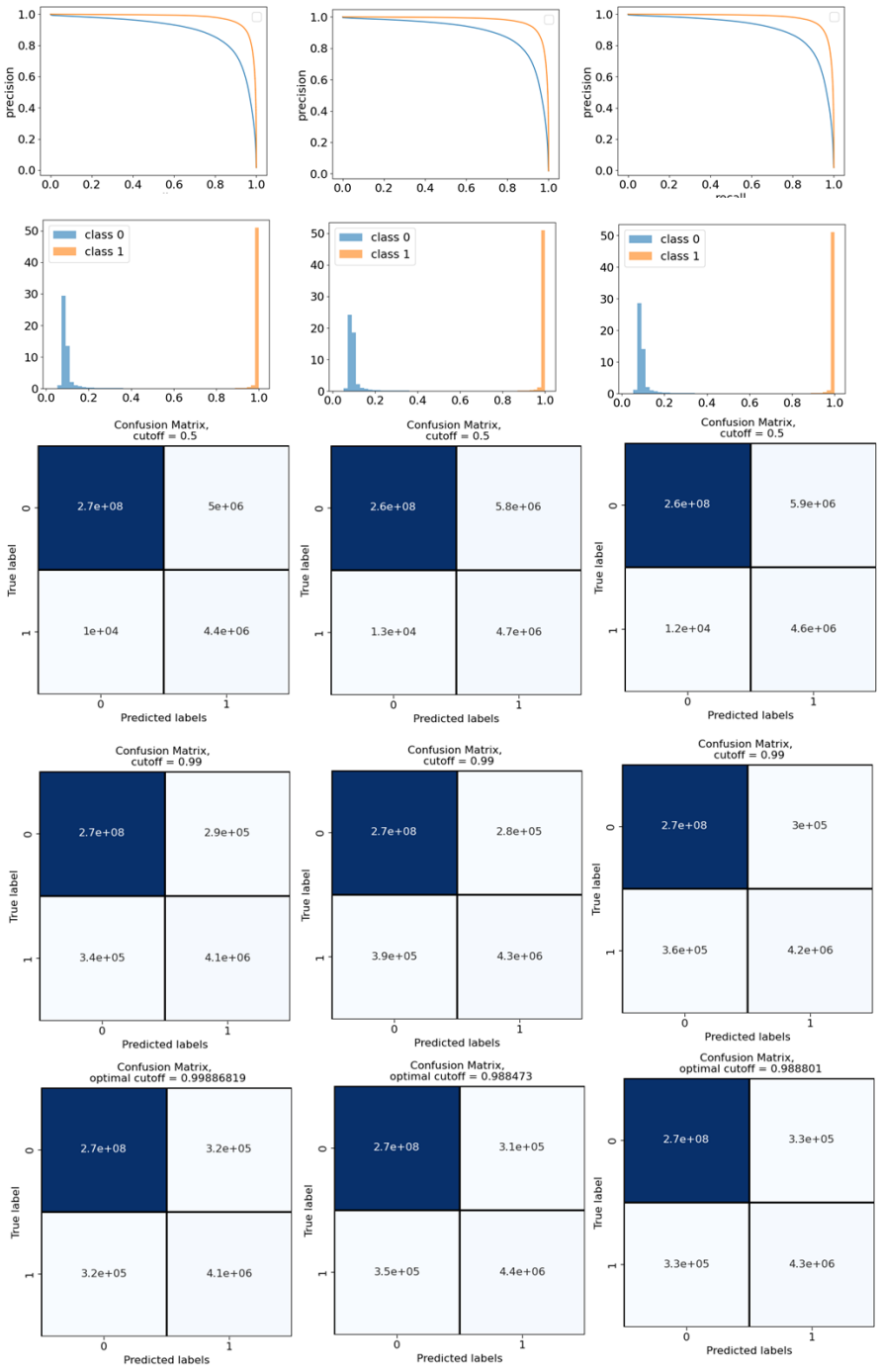

Figure 5.22: Results for three 200MB test sets of the African WSS model. Each column shows results for one test set. First row: Precision recall curves; Second row: histogram of the probability distributions for both classes; Third row: Confusion matrix for cut-off=0.5; Fourth row: Confusion matrix for cut-off=0.99; Fifth row: Confusion matrix for the optimal cut-off in terms of F1 score

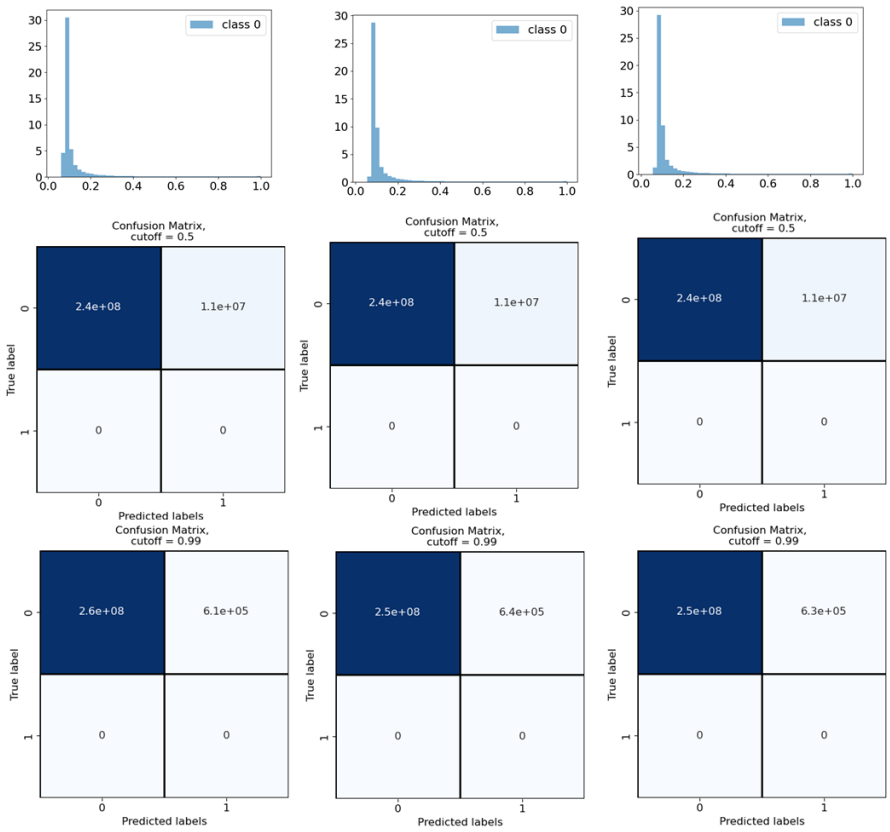

Figure 5.23: Results for three 200MB test sets of the African WSS model without gene flow. Each column shows results for one test set. First row: histogram of the probability distributions (since there are no true positives, there is only one class); Second row: Confusion matrix for cut-off=0.5; Third row: Confusion matrix for cut-off=0.99

6 Discussion

In the following, the central results of the benchmarking are critically discussed.

In general, both tools, ArchIE and IntroUNET, show a good performance on all datasets (Human-Neanderthal and ArchIE model for ArchIE, for IntroUNET also the Bonobo-Ghost model) compared to Sstar, HMMix and SPrime.

It should be noted again that the results for IntroUNET cannot be directly compared to the other tools since IntroUNET gives a prediction for each SNP, whereas for the other tools, *e.g.* ArchIE, predictions are averages over sliding windows of a specific length. For these tools, the stepsize is a lower bound for the resolution of the predictions (however, usually a sliding window is used and so one can average for each SNP over all windows which cover its position).

A more refined analysis could take this into account by either extracting all SNPs which are located within a predicted introgression tract or estimate continuous tracts from the IntroUNET results.

ArchIE performs best for a relatively large number of target individuals, but also in the case of only one target individual, it outperforms the other tools on both datasets. Notably, the performance for 10 reference individuals is similar to the other tools with 50 reference individuals. Similarly, IntroUNET benefits from additional target individuals, but for the ArchIE model also shows decent performance even for only two individuals.

6.1 ArchIE results

For ArchIE, a predominance of a few features has been observed, suggesting that removal of most features will not harm performance. This has been confirmed by applying L1-regularization which leads to a removal of most features (*i.e.* its weight is set to zero).

As the results have shown, in some cases, due to their dubious information content, omission of features can be recommended. This is especially true of the features with size dependent on the number of sampled individuals, which for models with a larger number of individuals account for the bigger part of the feature vector length. The random fluctuations between runs with the same model and consequentially their removal by regularization suggests that these are not really informative features. On the other hand, for models with a low number of individuals / haplotypes (in the most extreme, however not uncommon, case only one or two) some of the features become either redundant (*e.g.*, the mean of the distances is equal to the one and only pairwise distance) or cannot be computed (*e.g.*, kurtosis and skew).

A more sparse implementation would also make the application of ArchIE much more flexible.

In contrast to the statement in [6], that for replicating the ArchIE results balanced data was necessary, no necessity of balancing the training data could be found. On the contrary, balanced data slightly impaired overall performance. However, it cannot be ruled out that for smaller training data sizes, the difference between balanced and imbalanced data could be more pronounced and one could, using balanced data, get along with less training data. At least for the ArchIE model, the usage of balanced datasets cannot be recommended, in particular when the immense computational resources necessary due to the rarity of the introgressed class are considered.

6.2 IntroUNET results

The incorporation of relative position information as additional CNN channels provided very interesting results. Depending on the demographic model, it improved results remarkably (Bonobo-Ghost model), slightly, but stable (Human-Neanderthal model) or not at all (ArchIE model).

One the one hand, this result suggests that positional information is especially important for discerning more archaic introgression events. Alternatively, it could be that this specific, rather simple incorporation of the position information, works only for some models, whereas for models like the ArchIE model other ways of harnessing this information have to be considered.

A more thorough investigation of the parameters - *i.e.* which demographic models benefit most from this information - could be very insightful.

The runs with a small number of reference and target individuals have shown that for accurate predictions the pattern emerging due to the reference individuals is crucial. Additional target individuals give rise to better predictions, but even for only two, the results are convincing. The pattern detectable by the CNN is partly generated by the seriation process which at first orders the reference individuals so that similar ones are adjacent to another and subsequently assigns the individuals from the other population applying the Kuhn-Munkres algorithm. In [6], it is recommended to apply seriation to the population with higher diversity and match the other one. For cases in which one population has a significant lower number of sampled individuals, its diversity is dramatically diminished due to upsampling. Thus it could be that for a low number of reference individuals, seriation should be applied on the target population for improving performance (otherwise, multiple copies of the same haplotype are arranged next to each other, which certainly does not generate a meaningful pattern).

IntroUNET performed well on all demographic models (ArchIE model, Human-Neanderthal and Bonobo-Ghost). Without exception, evaluation of the trained model on test data from a different demographic model caused a tremendous drop in performance. This is in contrast to ArchIE which has been shown to be quite robust even in the case of major model misspecifications in which not only parameters like mutation rate, but the demographic model itself is modified.

Likely, this is due to the ability of DL architectures to better fit to the data, causing also the increased danger of overfitting. The superior ability of interpolation goes along

with dubious results in extrapolation, i.e. the prediction of test instances which are from a distribution differing from the training data.

Given that it is dealt with a binary classification problem, it learns to distinguish between the negative (0) and positive (1) class; however, instances from a different demographic model might appear significantly different from both classes (i.e. different from all instances the model has seen during training) and thus for such instances no stable classification is made. Rather than setting fixed parameters for this, one could provide training data within the full range of the estimation.

In contrast, the summary statistics of the ArchIE feature vector focus more on general statistics which are presumably more robust in case of deviations of the demographic model. However, it is debatable whether prediction outside the training space is desirable. A possible solution (which does not only improve benchmark results on synthetic data, but could also strengthen reliability in realistic scenarios) would be to use simulation data which stems from different models or parameter settings. This would be particularly advantageous because the demographic models (like all models derived from empirical data) are inherently subject to uncertainty. For example, in [32] a range from 2-19 % for the admixture proportion is estimated; using such approach, one could use simulations within the full range as input.

Regarding the improved predictions for longer polymorphism windows, it has to be mentioned that the predictions at the edges of the windows are probably worse [6] (this is even more likely for the new version with relative position information since at the margins the values are currently set to zero). It remains to be tested whether the longer windows with 256 polymorphisms, which are very expensive in terms of GPU computing time, are worthwhile or simply discarding of the predictions for the edges would yield a similar performance boost.

Similar to other tools (*e.g.* also for ArchIE the minimum distance to the reference is one of the most important models), introUNET relies crucially on information about the reference. As expected, for only two reference individuals, it fails completely. It would be necessary to investigate this relation in more detail (*e.g.* how many individuals are at least necessary that the precision-recall curves become convex and is there saturation, *i.e.* a number above which predictions do not or only scarcely improve) and also study its dependence on specific demographic models.

7 Conclusion and Outlook

The results show that ML and DL approaches like ArchIE and IntroUNET provide good performance on various demographic models. The reimplementations of ArchIE and IntroUNET take vcf-files as input and can therefore easily be applied to msprime simulations based on custom demographic models provided in the DEMES format. This provides a flexibility which can be used for studying introgression at a much broader range.

For IntroUNET, a more nuanced analysis of its capability to handle unphased data would be advisable. For example, for unphased data, one could also implement a multi-class classification problem in which three cases, homozygous and heterozygous introgressed as well as not introgressed, are discriminated.

In particular, the dependence of the various parameters (divergence time, population sizes) should be investigated in even more detail to discover which settings are especially challenging for such methods.

Furthermore, the application of other neural network architectures seems promising. In particular, the IntroUNET architecture leans itself to a graph version. GNNs can be seen as generalizations of CNNs; in contrast to the latter, they work directly on data received as graph representation, and do not rely on grid-like data with regular connections [40]. This allows to process heterogeneous input, for example samples of different length or containing missing data. In the context of genotype matrices, their higher flexibility could be harnessed to overcome some limitations of CNNs, for instance by representing positional information as edge attribute or by replacing the need for seriation with appropriate connections between the graph nodes which represent variants or tree sequences.

Furthermore, it would be interesting to frame the problem rather as an anomaly detection than of a classification problem; for this purpose, also appropriate DL architectures like autoencoders could be applied.

Transformers and other DL architectures incorporating attention mechanisms are also promising, given that long-range interactions could be of interest for very complex models. Given their capacity, it is also desirable to deploy such architectures in adaptive introgression tasks. The extension to simulations which deviate from the neutral model and include phenomena like adaptive introgression would be insightful, extending predictions to more realistic scenarios.

Bibliography

[1] Jente Ottenburghs. Ghost introgression: Spooky gene flow in the distant past. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 42(6):e2000012, 2020.

[2] Vincent Plagnol and Jeffrey D. Wall. Possible ancestral structure in human populations. *PLoS genetics*, 2(7):e105, 2006.

[3] Xin Huang, Patricia Kruisz, and Martin Kuhlwilm. sstar: A python package for detecting archaic introgression from population genetic data with s. *Molecular biology and evolution*, 39(11), 2022.

[4] Laurits Skov, Ruoyun Hui, Vladimir Shchur, Asger Hobolth, Aylwyn Scally, Mikkel Heide Schierup, and Richard Durbin. Detecting archaic introgression using an unadmixed outgroup. *PLoS genetics*, 14(9):e1007641, 2018.

[5] Arun Durvasula and Sriram Sankararaman. A statistical model for reference-free inference of archaic local ancestry. *PLoS genetics*, 15(5):e1008175, 2019.

[6] Dylan D. Ray, Lex Flagel, and Daniel R. Schrider. Introunet: identifying introgressed alleles via semantic segmentation. *bioRxiv : the preprint server for biology*, 2023.

[7] Mark S. Hibbins and Matthew W. Hahn. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2), 2022.

[8] Richard G. Harrison and Erica L. Larson. Hybridization, introgression, and the nature of species boundaries. *The Journal of heredity*, 105 Suppl 1:795–809, 2014.

[9] Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de La Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo. A draft

sequence of the neandertal genome. *Science (New York, N.Y.)*, 328(5979):710–722, 2010.

[10] Martin Kuhlwilm, Ilan Gronau, Melissa J. Hubisz, Cesare de Filippo, Javier Prado-Martinez, Martin Kircher, Qiaomei Fu, Hernán A. Burbano, Carles Lalueza-Fox, Marco de La Rasilla, Antonio Rosas, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Tomas Marques-Bonet, Aida M. Andrés, Bence Viola, Svante Pääbo, Matthias Meyer, Adam Siepel, and Sergi Castellano. Ancient gene flow from early modern humans into eastern neanderthals. *Nature*, 530(7591):429–433, 2016.

[11] Kay Prüfer, Cesare de Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pinghsun Hsieh, Stéphane Peyrégne, David Reher, Charlotte Hopfe, Sarah Nagel, Tomislav Maricic, Qiaomei Fu, Christoph Theunert, Rebekah Rogers, Pontus Skoglund, Manjusha Chintalapati, Michael Dannemann, Bradley J. Nelson, Felix M. Key, Pavao Rudan, Željko Kućan, Ivan Gušić, Liubov V. Golovanova, Vladimir B. Doronichev, Nick Patterson, David Reich, Evan E. Eichler, Montgomery Slatkin, Mikkel H. Schierup, Aida M. Andrés, Janet Kelso, Matthias Meyer, and Svante Pääbo. A high-coverage neandertal genome from vindija cave in croatia. *Science (New York, N.Y.)*, 358(6363):655–658, 2017.

[12] David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053–1060, 2010.

[13] Claudia Fontsere, Marc de Manuel, Tomas Marques-Bonet, and Martin Kuhlwilm. Admixture in mammals and how to understand its functional implications: On the abundance of gene flow in mammalian species, its impact on the genome, and roads into a functional understanding. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 41(12):e1900123, 2019.

[14] Dora Koller, Frank R. Wendt, Gita A. Pathak, Antonella de Lillo, Flavio de Angelis, Brenda Cabrera-Mendoza, Serena Tucci, and Renato Polimanti. Denisovan and neanderthal archaic introgression differentially impacted the genetics of complex traits in modern populations. *BMC biology*, 20(1):249, 2022.

[15] Hugo Zeberg and Svante Pääbo. The major genetic risk factor for severe covid-19 is inherited from neanderthals. *Nature*, 587(7835):610–612, 2020.

[16] Hugo Zeberg and Svante Pääbo. A genomic region associated with protection against severe covid-19 is inherited from neandertals. *Proceedings of the National Academy of Sciences of the United States of America*, 118(9), 2021.

[17] Keila Velazquez-Arcelay, Laura L. Colbran, Evonne McArthur, Colin M. Brand, David C. Rinker, Justin K. Siemann, Douglas G. McMahon, and John A. Capra. Archaic introgression shaped human circadian traits. *Genome biology and evolution*, 15(12), 2023.

[18] Davide M. Vespasiani, Guy S. Jacobs, Laura E. Cook, Nicolas Brucato, Matthew Leavesley, Christopher Kinipi, François-Xavier Ricaut, Murray P. Cox, and Irene Gallego Romero. Denisovan introgression has shaped the immune system of present-day papuans. *PLoS genetics*, 18(12):e1010470, 2022.

[19] Wei-Ping Zhang, Ya-Mei Ding, Yu Cao, Pan Li, Yang Yang, Xiao-Xu Pang, Wei-Ning Bai, and Da-Yong Zhang. *Uncovering Ghost Introgression Through Genomic Analysis of a Distinct East Asian Hickory Species*. 2023.

[20] Emilia Huerta-Sánchez, Xin Jin, Asan, Zhuoma Bianba, Benjamin M. Peter, Nicolas Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixiang Ni, Bo Wang, Xiaohua Ou, Huasang, Jiangbai Luosang, Zha Xi Ping Cuo, Kui Li, Guoyi Gao, Ye Yin, Wei Wang, Xiuqing Zhang, Xun Xu, Huanming Yang, Yingrui Li, Jian Wang, Jun Wang, and Rasmus Nielsen. Altitude adaptation in tibetans caused by introgression of denisovan-like dna. *Nature*, 512(7513):194–197, 2014.

[21] Patrick F. Reilly, Audrey Tjahjadi, Samantha L. Miller, Joshua M. Akey, and Serena Tucci. The contribution of neanderthal introgression to modern human traits. *Current biology : CB*, 32(18):R970–R983, 2022.

[22] Raül Buisan, Juan Moriano, Alejandro Andirkó, and Cedric Boeckx. A brain region-specific expression profile for genes within large introgression deserts and under positive selection in homo sapiens. *Frontiers in cell and developmental biology*, 10:824740, 2022.

[23] Martin Kuhlwilm. The evolution of foxp2 in the light of admixture. *Current Opinion in Behavioral Sciences*, 21:120–126, 2018.

[24] Martin Kuhlwilm, Sojung Han, Vitor C. Sousa, Laurent Excoffier, and Tomas Marques-Bonet. Ancient admixture from an extinct ape lineage into bonobos. *Nature ecology & evolution*, 3(6):957–965, 2019.

[25] Harvinder Pawar, Aigerim Rymbekova, Sebastian Cuadros-Espinoza, Xin Huang, Marc de Manuel, Tom van der Valk, Irene Lobon, Marina Alvarez-Estape, Marc Haber, Olga Dolgova, Sojung Han, Paula Esteller-Cucala, David Juan, Qasim Ayub, Ruben Bautista, Joanna L. Kelley, Omar E. Cornejo, Oscar Lao, Aida M. Andrés, Katerina Guschanski, Benard Ssebide, Mike Cranfield, Chris Tyler-Smith, Yali Xue, Javier Prado-Martinez, Tomas Marques-Bonet, and Martin Kuhlwilm. Ghost admixture in eastern gorillas. *Nature ecology & evolution*, 7(9):1503–1514, 2023.

[26] Fernando Racimo, Davide Marnetto, and Emilia Huerta-Sánchez. Signatures of archaic adaptive introgression in present-day human populations. *Molecular biology and evolution*, 34(2):296–317, 2017.

[27] Théo Tricou, Eric Tannier, and Damien M. de Vienne. Ghost lineages highly influence the interpretation of introgression tests. *Systematic biology*, 71(5):1147–1158, 2022.

[28] Théo Tricou, Eric Tannier, and Damien M. de Vienne. Ghost lineages can invalidate or even reverse findings regarding gene flow. *PLoS biology*, 20(9):e3001776, 2022.

[29] Xiao-Xu Pang and Da-Yong Zhang. Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. *Systematic biology*, 72(1):35–49, 2023.

[30] Mayukh Mondal, Jaume Bertranpetit, and Oscar Lao. Approximate bayesian computation with deep learning supports a third archaic introgression in asia and oceania. *Nature communications*, 10(1):246, 2019.

[31] Belen Lorente-Galdos, Oscar Lao, Gerard Serra-Vidal, Gabriel Santpere, Lukas F. K. Kuderna, Lara R. Arauna, Karima Fadhlaoui-Zid, Ville N. Pimenoff, Himla Soodyall, Pierre Zalloua, Tomas Marques-Bonet, and David Comas. Whole-genome sequence analysis of a pan african set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-saharan populations. *Genome biology*, 20(1):77, 2019.

[32] Arun Durvasula and Sriram Sankararaman. Recovering signals of ghost archaic introgression in african populations. *Science advances*, 6(7):eaax5097, 2020.

[33] Shaohua Fan, Jeffrey P. Spence, Yuanqing Feng, Matthew E. B. Hansen, Jonathan Terhorst, Marcia H. Beltrame, Alessia Ranciaro, Jibril Hirbo, William Beggs, Neil Thomas, Thomas Nyambo, Sununguko Wata Mpoloka, Gaonyadiwe George Mokone, Alfred Njamnshi, Charles Folkunang, Dawit Wolde Meskel, Gurja Belay, Yun S. Song, and Sarah A. Tishkoff. Whole-genome sequencing reveals a complex african population demographic history and signatures of local adaptation. *Cell*, 186(5):923–939.e14, 2023.

[34] Aaron Pfennig, Lindsay N. Petersen, Paidamoyo Kachambwa, and Joseph Lachance. Evolutionary genetics and admixture in african populations. *Genome biology and evolution*, 15(4), 2023.

[35] Jui-Tse Chang, Koh Nakamura, Chien-Ti Chao, Min-Xin Luo, and Pei-Chun Liao. Ghost introgression facilitates genomic divergence of a sympatric cryptic lineage in cycas revoluta. *Ecology and evolution*, 13(8):e10435, 2023.

[36] Kazunori Yamahira, Hirozumi Kobayashi, Ryo Kakioka, Javier Montenegro, Kawilarang W. A. Masengi, Noboru Okuda, Atsushi J. Nagano, Rieko Tanaka, Kiyoshi Naruse, Shoji Tatsumoto, Yasuhiro Go, Satoshi Ansai, and Junko Kusumi. Ghost introgression in ricefishes of the genus adrianichthys in an ancient wallacean lake. *Journal of evolutionary biology*, 36(10):1484–1493, 2023.

[37] Benjamin Vernot and Joshua M. Akey. Resurrecting surviving neandertal lineages from modern human genomes. *Science (New York, N.Y.)*, 343(6174):1017–1021, 2014.

[38] Benjamin Vernot, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, Laura B. Scheinfeldt, David A. Merriwether, George Koki, Jonathan S. Friedlaender, Jon Wakefield, Svante Pääbo, and Joshua M. Akey. Excavating neandertal and denisovan dna from the genomes of melanesian individuals. *Science (New York, N.Y.)*, 352(6282):235–239, 2016.

[39] Sriram Sankararaman. Methods for detecting introgressed archaic sequences. *Current opinion in genetics & development*, 62:85–90, 2020.

[40] Xin Huang, Aigerim Rymbekova, Olga Dolgova, Oscar Lao, and Martin Kuhlwilm. Harnessing deep learning for population genetic inference. *Nature reviews. Genetics*, 25(1):61–78, 2024.

[41] Lex Flagel, Yaniv Brandvain, and Daniel R. Schrider. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular biology and evolution*, 36(2):220–238, 2019.

[42] Graham Gower, Pablo Iáñez Picazo, Matteo Fumagalli, and Fernando Racimo. Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife*, 10, 2021.

[43] Sharon R. Browning, Brian L. Browning, Ying Zhou, Serena Tucci, and Joshua M. Akey. Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell*, 173(1):53–61.e9, 2018.

[44] Ying Zhou and Sharon R. Browning. Protocol for detecting introgressed archaic variants with sprime. *STAR protocols*, 2(2):100550, 2021.

[45] Richard R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–338, 2002.

[46] Graham Gower, Aaron P. Ragsdale, Gertjan Bisschop, Ryan N. Gutenkunst, Matthew Hartfield, Ekaterina Noskova, Stephan Schiffels, Travis J. Struck, Jerome Kelleher, and Kevin R. Thornton. Demes: a standard format for demographic models. *Genetics*, 222(3), 2022.

[47] Aaron P. Ragsdale, Timothy D. Weaver, Elizabeth G. Atkinson, Eileen Hoal, Marlo Möller, Brenna M. Henn, and Simon Gravel. *A weakly structured stem for human origins in Africa.* 2022.

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban

Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library.

[49] Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P. Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E. Castedo Ellerman, Jared G. Galloway, Ariella L. Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W. Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S. Pope, Consuelo D. Quinto-Cortés, Murillo F. Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W. Wohns, Yan Wong, Simon Gravel, Andrew D. Kern, Jere Koskela, Peter L. Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3), 2022.

[50] Jerome Kelleher, Kevin R. Thornton, Jaime Ashander, and Peter L. Ralph. Efficient pedigree recording for fast population genetics simulation. *PLoS computational biology*, 14(11):e1006581, 2018.

[51] Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28(19):2520–2522, 2012.

[52] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.