Neural posterior estimation for population genetics

Jiseon Min[1,♢], Yuxin Ning[2,♢], Nathaniel S. Pope[1,♢], Franz Baumdicker[2,3,♡], and Andrew D. Kern[1,♡]

[1]Institute of Ecology and Evolution, University of Oregon, Eugene, OR
[2]Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Germany
[3]Big Data Analytics in Bioinformatics, Justus-Liebig University, Gießen, Germany
[♢]Indicates shared first author
[♡]Indicates shared corresponding author

December 1, 2025

Abstract

Simulation-based inference methods are increasingly being used in population genetics due to their flexibility and ability to be applied in settings where likelihood-based methods are intractable. Perhaps the best known such method is Approximate Bayesian Computation (ABC); however, its popularity is offset by its shortcomings which include computational expense and an unfortunate inability to fit models to high-dimensional summaries of the data. An alternative approach that solves these issues is supervised machine learning (ML); however, ML methods generally do not yield Bayesian uncertainty estimates of the quantities they predict. Here, we apply a recently introduced method, neural posterior estimation (NPE), that combines the best facets of ABC and supervised ML by training a neural network to estimate the posterior distribution of a population genetics model. We first compare neural posterior estimation with other inference methods for a variety of population genetic tasks, and show that neural posterior estimators yield posterior distributions with high accuracy and efficiency. We compare learned posterior distributions given raw genotypes and various summary statistics as input data. Additionally, we apply neural posterior estimation for demographic inference for simple and more complex models to highlight its application. Finally, we provide a user friendly workflow that enables others to perform neural posterior estimation to their own genetic data.

Introduction

Population genetic inference typically proceeds via statistical estimation using a few well-described models (i.e., the Wright-Fisher model or its genealogical dual, the coalescent).

Generally speaking, the key step is to describe how some summary of the data (e.g., the site frequency spectrum) depends on the parameters of a given evolutionary model. A familiar example is the software package `∂a∂i` (Gutenkunst et al., 2009), which uses a diffusion approximation to the Wright-Fisher model to estimate the population parameters that best fit an observed site frequency spectrum. The fitting occurs through maximum likelihood estimation, wherein the probability of the data (in this case, the site frequency spectrum) is optimized with respect to the parameters. Maximum likelihood estimation is a fundamental approach to statistical inference which allows uncertainty quantification via Fisher's information matrix (or its generalization, the Godambe information matrix)—the variance-covariance matrix that describes how much information the data holds for each parameter of the model— through the assumption of asymptotic normality of the estimators. Likelihood estimation has a long history in population genetics, and accordingly, there are many methods that use this approach, including for demographic inference (Adams and Hudson, 2004; Gutenkunst et al., 2009; Jouganous et al., 2017; Ragsdale and Gravel, 2019), estimation of recombination rates (Hudson, 2001; McVean et al., 2002; Chan et al., 2012; Spence and Song, 2019), and estimation of selection parameters (Bustamante et al., 2001, 2003; Nielsen et al., 2005) among other uses. While likelihood-based methods can be accurate, they are also limiting—model complexity and/or realism is often sacrificed when we depend on analytical results for inference.

An alternative approach relies on simulation to estimate the probability of observing certain features of a population genetic dataset and then to perform inference accordingly. Perhaps the best-known such approach in population genetics is Approximate Bayesian Computation (ABC; Beaumont et al. (2002)). Using ABC, a dataset is summarized by an informative set of features (e.g., heterozygosity, the number of haplotypes, etc.) and simulations of a given evolutionary scenario are performed using prior distributions of the parameters of interest. Each simulation is compared to the observed dataset, and either a rejection step or a regression step (or both) is used to estimate quantities from the posterior distribution of the parameters. ABC at this point has been well-studied and widely-applied (e.g., Thornton and Andolfatto, 2006; Laurent et al., 2011; Smith and Flaxman, 2020, among others); however, its popularity is offset by its shortcomings which include computational expense and an unfortunate inability to fit models to high-dimensional summaries of the data (Prangle, 2015). At the same time, ABC highlights two substantial advantages of working with carefully chosen summary statistics in population genetics: they can be made interpretable in terms of classical theory and under some conditions they provide a principled, model-based route to inference with formal asymptotic guarantees. In what follows, we will use ABC primarily as a conceptual reference point for simulation-based Bayesian inference, rather than as a direct competitor that we benchmark against in detail.

Recently the field of population genetics has moved towards supervised ML and deep learning approaches which have been able to overcome many of the potential drawbacks of ABC (reviewed in Schrider and Kern, 2018; Huang et al., 2024). Through leveraging classical ML approaches, such as random forests and boosting, the associated "curse of dimensionality" with ABC feature selection can be side-stepped, allowing for a large number of potentially correlated summaries to be used without issue. Some early applications of this approach include methods for selective and demographic inference (Lin et al., 2011; Ronen et al., 2013; Schrider and Kern, 2015, 2016; Schrider et al., 2018), as well as for recombination rate

estimation (Lin et al., 2013). These ML-based approaches have the additional benefit of amortization—once a model is sufficiently trained, no further computational investment is needed for doing estimation again and again on new data points. While that is so, classic ML methods don't save the developer from the hard work of feature engineering—just as in ABC, one needs to proffer an appropriately informative set of summaries to the method for learning to occur.

To solve this problem, increasingly, population genetics has used so-called "end-to-end" learning as implemented by modern approaches to deep learning (DL). By using artificial neural network (ANN) architectures capable of automated feature extraction (Chan et al., 2018; Flagel et al., 2019; Adrion et al., 2020a), such as convolutional and recurrent neural networks (CNNs and RNNs respectively; reviewed in Huang et al. (2024)), a feature set needs not be hand-crafted, but instead *learned* by the method itself. While less interpretable than summary statistic based methods, this approach allows ANNs to automatically extract informative features from the data without human supervision. Indeed, this effort has been quite successful in population genetics over the last five or so years, and its applications have blossomed (e.g. Gower et al. (2021); Sanchez et al. (2021); Torada et al. (2019); López-Cortés et al. (2020)).

However, ABC outshines current deep learning approaches in population genetics in one crucial respect: an explicit quantification of the uncertainty surrounding parameter estimates in the form of a posterior distribution. Recent progress in training deep neural networks to produce posterior densities rather than point estimates promises to fuse the efficiency of DL for simulation-based inference with the principled probabilistic logic of Bayesian statistics (Papamakarios et al., 2019; Greenberg et al., 2019; Hermans et al., 2020). These neural posterior estimation (NPE) methods combine deep neural networks and with flexible density estimators (e.g., normalizing flows) to learn a function that maps an input (such as a genotype array or a vector of summary statistics) directly to an approximation of the posterior distribution. Conceptually, NPE can be viewed as an amortized, learned analogue of the regression-adjustment step in ABC, but with far fewer user-chosen tuning knobs.

A key advantage of NPE in practice is that it "just plugs in" on top of whatever representation of the data is convenient. Given an expressive enough network, the same machinery can operate on hand-crafted summary statistics or on features extracted by a separate deep network, without needing to specify distance functions, tolerance thresholds, kernel shapes, or regression corrections as in ABC. This substantially reduces the amount of problem-specific hand-holding required from the user: once a simulator is in place and a prior is defined, NPE can be trained in a largely standardized way across very different population genetic applications. The trade-off is that, unlike some formulations of ABC, current NPE methods generally do not come with the same strong asymptotic guarantees of consistency for arbitrary choices of architecture and training scheme, and are instead justified empirically and by approximation arguments.

Importantly, we view the use of summary statistics as a central and enduring use case for NPE, rather than merely a stepping stone toward fully end-to-end models. When the summaries are interpretable and grounded in classical population genetic theory (e.g., site frequency spectra, LD statistics, etc.), NPE inherits those interpretability benefits while providing a fast, amortized posterior approximation. Because NPE learns a parametric approximation to the posterior, once trained it can evaluate posteriors for new datasets

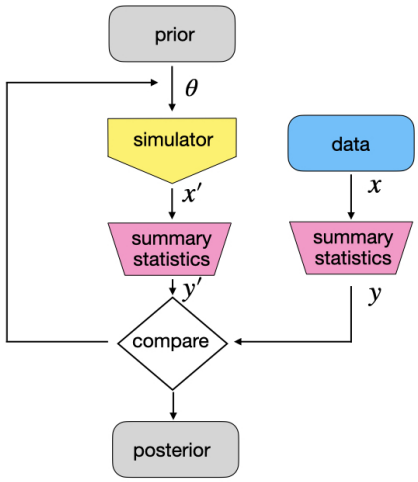
in milliseconds, making it particularly well-suited to settings where a small number of theoretically motivated summaries are available but inference must be repeated many times (e.g., across genomic windows, bootstrap replicates, or multiple populations).

In this study, we demonstrate that this approach—neural posterior estimation—offers incredible promise for the population genetics community by allowing the "best of all worlds" for simulation-based inference: interpretable and theoretically grounded summary statistics when desired, end-to-end feature extraction when needed, and in both cases rapid, amortized posterior evaluation with well-calibrated uncertainty. We illustrate how NPE can be used with both classical summaries and neural feature extractors (e.g., CNNs) that ingest genotype matrices directly, and how these components can be expressively composed so that feature extraction networks summarize the data before posteriors are estimated. In what follows, we demonstrate the utility of NPE in a number of population genetic settings, including recombination rate estimation and inference of demographic parameters: we show that learned posteriors are both rapidly generated, well-calibrated, and in some cases more accurate than those produced by currently available techniques.

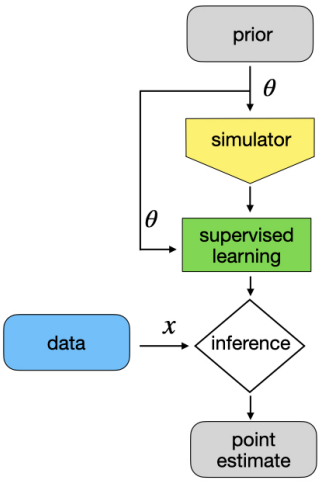
Methods

To date, the most popular approaches to simulation-based inference in population genetics are Approximate Bayesian Computation (ABC) and supervised machine learning approaches. A general schematic of these workflows is show in Figures 1a and 1b. Both of these approaches share a common strategy of defining a prior distribution $p(\theta)$ on the parameters of a model θ, and then using a simulator to generate data x' conditional upon a draw from the prior; e.g. simulating from $p(x', \theta) = p(x' \mid \theta)p(\theta)$. However, from that point in the procedure the two approaches diverge. ABC generates draws from the posterior $p(\theta \mid x)$ via approximate rejection sampling: by first reducing the simulated data x' to a summary y', and accepting the corresponding draw of θ if $\phi(y, y') < \epsilon$ (where ϕ is some measure of discordance between simulated and observed summary statistics, and ϵ is an arbitrary threshold). In practice the ABC rejection step may involve post-processing by regression or a more complex model to reduce bias and increase statistical efficiency. In contrast, the supervised machine learning approach aims to train a heavily parameterized model—e.g., an artificial neural network (ANN)—on the simulated, unreduced data to predict the value of θ which generated it. Through training of the ANN, the model can automatically create informative features from the raw data, in a process that is generally called automated feature extraction, rather than depend on the user's choice of summaries to produce y'. The trained model applied to the observed data results in a point estimate of θ (e.g. Adrion et al., 2020a).

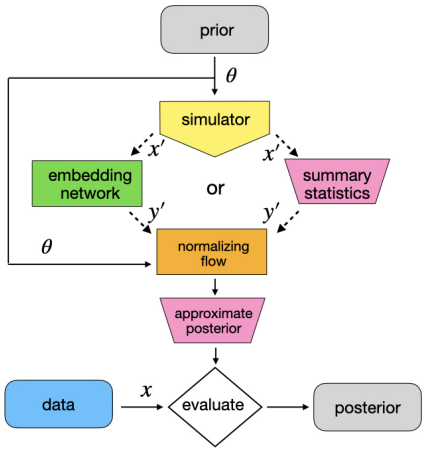
Here we apply a recently developed method for simulation-based inference of the posterior distribution of the parameters given the data, neural posterior estimation (NPE; Figure 1c that combines elements of ABC and machine learning). NPE also uses samples from $p(x', \theta)$, but feeds these through a particular form of deep neural network designed for density estimation known as a *normalizing flow* that approximates this joint distribution. After sufficient training, the NPE model is conditioned on the observed data (or summaries thereof) to directly approximate the posterior $p(\theta \mid x)$. Like the supervised machine learning approach, NPE can operate on the original data rather than on summary statistics, potentially using



(a) Approximate Bayesian Computation



(b) Supervised Machine Learning Inference



(c) Neural Posterior Inference

Figure 1: Schematic workflows comparing strategies of simulation-based inference in population genetics. Common workflows currently are Approximate Bayesian Computation (ABC; Fig. 1a) and Supervised Machine learning (Fig. 1b). Here we apply neural posterior inference (Fig. 1c) for inference in population genetics.

an *embedding network* of arbitrary complexity prepended to the normalizing flow—however, the output is a distribution over plausible values of θ rather than a point estimate.

Neural Posterior Estimation Overview

Neural posterior estimation (NPE) provides a flexible way to learn the posterior distribution $p(\theta \mid x)$ directly from simulated pairs (θ, x) using deep neural networks designed for density

estimation. Whereas ABC relies on approximate rejection sampling and supervised machine learning produces point estimates, NPE learns an explicit, differentiable mapping that transforms a complex posterior distribution over θ into a simple latent probability distribution and vice versa. This is accomplished using *conditional normalizing flows*—invertible neural networks that implement learned changes of variables conditioned on the data.

The central idea is to introduce a latent variable z with a tractable density (typically a standard multivariate Gaussian) and to learn an invertible, differentiable transformation

$$z = f_\phi(\theta; x), \qquad \theta = f_\phi^{-1}(z; x),$$

parameterized by network weights ϕ. Because f_ϕ is invertible, the density of θ conditional on x can be evaluated using the change-of-variables formula applied to the density of z. In practice, normalizing flows accomplish this by composing a sequence of simple, invertible transformations whose Jacobians are efficiently computable. Through training on simulated (θ, x) pairs, the network learns how the posterior distribution $p(\theta \mid x)$ varies with the data.

A key advantage of NPE in population genetic applications is that the input x is not restricted to hand-crafted summary statistics. Instead, x may consist of raw genetic data, image-like encodings of alignments, or summaries produced automatically by an *embedding network* ψ. Such embedding networks—CNNs, RNNs, transformers, or other architectures appropriate to the data modality—learn informative low-dimensional representations from the raw data. These learned features are then provided to the normalizing flow, which maps them to an approximate posterior over θ. This compositional structure allows NPE to implement *end-to-end learning*: feature extraction and posterior estimation are learned jointly within the same model.

Training proceeds by optimizing the parameters of both the embedding network and the normalizing flow using stochastic gradient descent. All transformations are differentiable, enabling the use of automatic differentiation to compute gradients without resorting to MCMC, rejection sampling, or other sampling-based approximations. Once trained, the model provides amortized inference: for any new dataset x, a single evaluation of $f_\phi^{-1}(z; x)$ produces a full approximate posterior distribution over the parameters. Figure 2 illustrates this inference workflow.

NPE Workflow

We have developed a `snakemake` pipeline targeted towards posterior estimation from genotype data, using arbitrary embedding networks. Two constraints on simulation-based inference using genome-wide population genetic data is that (a) genome-scale simulations are very costly to produce, and (b) the "raw" data (genotype arrays) are too large to fit in GPU memory. Hence, we take an amortized approach, where an NPE model (embedding network attached to normalizing flow) is trained on shorter simulated sequences then applied over the observed data in windows, to generate a sequence of posteriors. Local posteriors may be sufficient for the task at hand (e.g. recombination rate prediction, sweep detection) or may be averaged (or otherwise combined) to produce global estimates (e.g. for demographic inference). The workflow prepares training and testing datasets by simulating tree sequences with `msprime` (Baumdicker et al., 2022), performs preprocessing (as appropriate; see below),

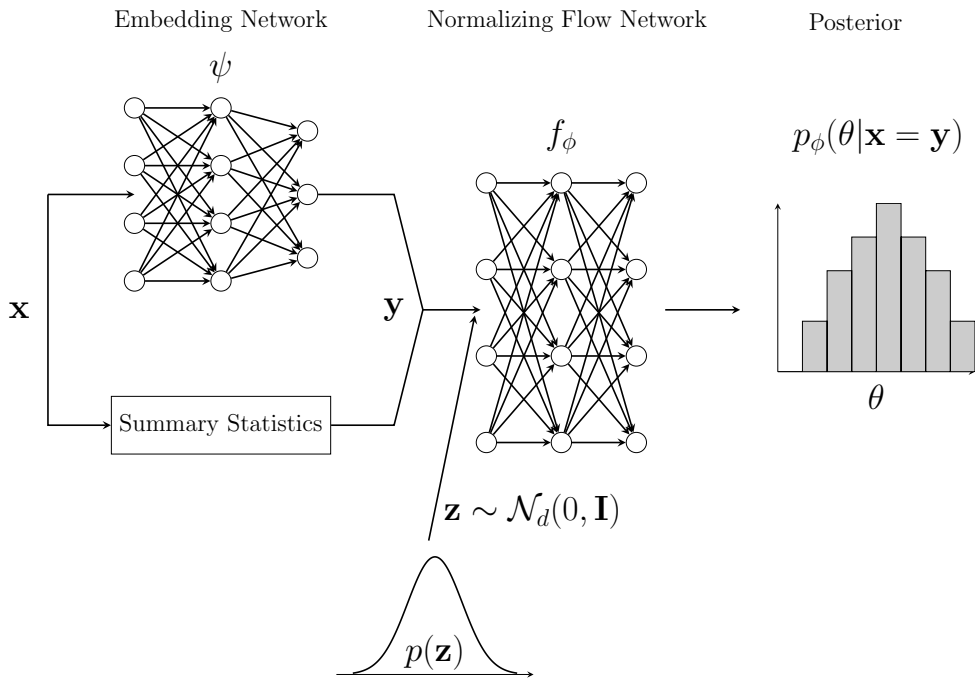


Figure 2: A schematic diagram of the inference step (a forward pass) in neural posterior estimation. Simulated or empirical population genetic data in the form of a genotype matrix, $\mathbf{x}$, is either passed to an embedding network, ψ, (e.g. a CNN or RNN) or optionally summarized into user engineered statistics (e.g., π, Tajima's D, etc.). The collection of summaries from the embedding network / summarization procedure, $\mathbf{y}$, is then passed to the normalizing flow network, f_ϕ, along with a Gaussian latent variable, $\mathbf{z}$. Through training, either by jointly or separately adjusting the parameters of the embedding and normalizing flow networks, the posterior distribution of the parameters, $p_\phi(\theta|\mathbf{x} = \mathbf{y})$, is learned.

trains NPE estimators, estimates model parameters across genomic intervals, and subsequently summarizes results from estimation. Our workflow and all associated code can be found at https://github.com/kr-colab/popgen-npe. We use normalizing flow architectures from the SBI python package Tejero-Cantero et al. (2020), and use a modular design such that arbitrary neural networks may be used to generate embeddings. We note that for some applications, a fully amortized approach may be statistically inefficient (in the sense that training examples simulated from the prior are too sparsely scattered across parameter space); in which case sequential refinement of the posterior estimator via importance sampling may be necessary (Tejero-Cantero et al., 2020).

Population Genetic Tasks

Throughout this manuscript we apply NPE to a variety of common population genetic tasks, highlighting the benefits of NPE and contrasting its performance to established methods both from simulation-based methods and from probabilistic methods when appropriate. Many of the methods are shared among tasks and outlined in separate sections below.

Coalescent Simulation

All simulation for training and testing was performed using `msprime` (Baumdicker et al., 2022). Specific parameterizations were task dependent and details are given in the subsequent sections. For estimation of demographic parameters, demographic models is either written in `demes` (Gower et al., 2022) from scratch or adapted from a model from `stdpopsim` catalog (Adrion et al., 2020b; Lauterbur et al., 2023), by letting a subset of the parameters (i.e. those targeted in estimation) be replaced by a prior sample.

Embedding Networks and Summary Statistics

Throughout the manuscript we tried a number of different embedding network architectures, referred to as ψ in Figure 2, for automated feature extraction from raw genotype data and in places contrast performance with hand-designed collections of summary statistics.

For recombination rate estimation in Section *Recombination rate estimation*, we set ψ to the recurrent GRU architecture of Adrion et al. (2020a) reimplemented in `PyTorch` to be compatible with our NPE workflow. This network takes as input the genotype matrix as well as a vector of positions encoding the locations of SNPs. The genotype matrix is passed through a bi-directional GRU layer, followed by a series of dense and dropout layers. The position input vector is separately passed through a dense layer, and then concatenated with the dropout layer output from the genotype branch of the network.

For inference of demographic parameters in Section *Inference of Population Bottleneck Parameters*, we used an exchangeable convolutional neural network (CNN) architecture modeled after Chan et al. (2018) and similar to the discriminator network implemented in `pg-gan` (Wang et al., 2021). Genotype and relative SNPs position information for input to these networks were prepared using the feature extractors in `dinf` (Gower et al. (2023)).

For inference of effective population size histories in Section *Inference of Population Bottleneck Parameters*, we again contrasted different choices of embedding neural networks with a summary statistic approach. Three DNN architectures were compared: 1) an RNN, 2) the exchangeable CNN architecture from Chan et al. (2018), and 3) a bespoke DNN architecture called `SPIDNA` from Sanchez et al. (2020). `SPIDNA` is designed specifically for genome sequence data, and incorporates haplotype permutation invariance while capturing long range dependencies between SNPs and variation in SNP density among regions. Due to its sophisticated design, training NPE with embedding `SPIDNA` demands the most computing power. In addition we examine performance in NPE inference when using either the SFS or LD statistics alone, for larger sequence lengths where end-to-end learning (e.g. from the raw genotypes) is infeasible.

For inference of effective population size histories in Section *Priors for Historical Population Size Inference*, we used a combination of linkage disequilibrium (LD) and site frequency spectrum (SFS) summary statistics as input features for neural posterior estimation. Simulated genomic regions of 2 Mb were divided into non-overlapping 1 Mb windows, and summary statistics were computed for each window and then averaged across windows. For LD, we computed the squared correlation coefficient (r^2) between pairs of SNPs using the Rogers-Huff estimator (Rogers and Huff, 2009). SNP pairs were subsampled at gap sizes following powers of two, and r^2 values were binned into 19 logarithmically-spaced distance classes ranging from

10^2 bp to the window length. The mean r^2 was computed for each distance bin. For the SFS, we computed the polarized (unfolded) site frequency spectrum for each window using `tskit`, excluding monomorphic site classes. With 25 diploid individuals (50 haplotypes), this yielded 49 SFS bins after excluding the zero and fixed classes. The final feature vector concatenated the 19 mean r^2 values with the 49 mean SFS values, resulting in 68 summary statistics per simulation that were passed directly to the neural density estimator via a summary statistics embedding layer.

Comparisons to `moments`

In Section *Inference of Population Bottleneck Parameters* we directly compare point and uncertainty estimates derived from NPE to those from the popular composite likelihood approach `moments` (Jouganous et al., 2017). As `moments` takes the SFS as input, we compare its performance to NPE by using the normalized SFS alone as the summary statistics for input to the normalizing flow (see above). Optimization of `moments` was performed by repeating optimization (with the BFGS algorithm) from 100 random starting points in parameter space, and keeping the top optimization result. Uncertainty estimation using `moments` was performed using the Godambe information matrix routine with `eps` = 0.05.

Comparisons to Approximate Bayesian Computation

In Section *Inference of Population Bottleneck Parameters* we additionally compare NPE estimators to estimates based on approximate Bayesian computation (ABC). In this comparison we utilized the same set of training simulations used to train the NPE, but performed ABC rejection sampling (*c.f.,* Pritchard et al., 1999), where the top 1% quantile of closest simulations was retained as the posterior sample. We used three summary statistics for the ABC as our summary vector: π, the number of segregating sites, and Tajima's D. The joint posterior distribution is visualized using a kernel density estimate of the accepted simulation parameter values.

Application to *Drosophila melanogaster*

We analyzed whole-genome sequence data from 19 haploid genomes (derived from inbred lines), including 9 sampled from France (FR) and 10 from Cameroon (CO), which were taken from the published DPGP2 collection (Pool et al., 2012). Prior to analysis, we filtered out genomic regions in which any sample contained missing data, ensuring that all downstream inference was based on complete genotype calls. Demographic inference was conducted under a modified Li and Stephan out-of-Africa model (Li and Stephan, 2006), which incorporates isolation with migration and exponential growth in the European lineage. This model includes seven free parameters: the ancestral population size, the effective sizes of the African and European populations following the split, the contemporary European effective size, the time of population divergence, and asymmetric migration rates between the European and African populations. For all parameters, we use uniform priors on a log scale ($10^{3.5}$ to $10^{6.5}$ for population sizes; 10^3 to 10^6 for the split time; and 10^{-8} to 10^{-3} for migration rates).

We employed neural posterior estimation (NPE) to infer posterior distributions of these demographic parameters, using a recurrent neural network embedding trained on 300 Kb

non-overlapping windows across chromosome arm 2L (each window containing approximately 6,000 SNPs). In each simulation, the population-scaled mutation rate θ was fitted assuming a per-base mutation rate of 0.45×10^{-8}, which accounts for the observed rate of missing data and corresponds to a correction from the baseline rate of 0.54×10^{-8} reported by Schrider *et al.* (2013). This rescaling ensures that simulated nucleotide divergence between African and European samples matches the observed value, and is analogous to the standard practice of fitting θ in site-frequency spectrum–based inference. The recombination rate was treated as a nuisance parameter and integrated out under a uniform prior ranging from 0.5×10^{-8} to 3×10^{-8} per base per generation. Input genotype features were unpolarized, with the minor allele at each site taken as the "derived" state.

To evaluate model fit, we performed posterior predictive checks by generating new sequence data for each window from the posterior distribution using `msprime`, and then calculated summary statistics (π, F_{ST}, Tajima's D). Posterior predictive distributions closely matched the observed statistics, indicating that the model adequately captured the data (Fig. 11). Moreover, systematic genomic variation in demographic parameter estimates was reflected in the posterior distributions across windows (Fig. S7). Finally, we averaged posteriors across windows to obtain global demographic parameter estimates, which are summarized in Fig. 10.

Results

Recombination Rate Estimation

Before tackling high-dimensional problems, we start with a 1-dimensional problem—inferring a uniform recombination rate. Our particular goal here is to show that NPE derived posteriors are well-calibrated and computationally efficient.

Previously, Adrion et al. (2020a) introduced a recurrent neural network (`ReLERNN`) that learns recombination rates from genotype data. The `ReLERNN` architecture consists of a bidirectional recurrent (GRU) layer that takes raw genotypes as input, followed by a series of fully-connected layers which end in a scalar output (the predicted recombination rate). Adrion et al. (2020a) showed that parametric bootstrapping provides well-calibrated uncertainty estimates: given a predicted recombination rate, genotype data is simulated via `msprime` (Baumdicker et al., 2022) and the trained `ReLERNN` model is applied to each synthetic dataset, generating a distribution of bootstrap replicates from which confidence intervals may be readily calculated. However, when the trained network is used across a large number of genomic windows (as would be the typical application, to build a recombination rate map) this procedure becomes expensive as a new set of simulations must be generated for every window. An alternative approach based on NPE is appealing as the cost of genotype simulation is paid up front.

For the current comparison we reimplemented the `ReLERNN` architecture within our pipeline. For training, we used coalescent simulations of ten diploids over a sequence of length 1Mb with an effective population size of 10^4 and a variable per-base mutation rate uniformly drawn from $[0.66 \times 10^{-8}, 1.33 \times 10^{-8}]$. For each simulation, the recombination rate r was sampled from a uniform $[0, 10^{-8}]$. We trained `ReLERNN` with 20,000 simulations with mean squared error loss against true r, and applied the parametric bootstrap procedure described above to

get confidence intervals for a test set of 2,000 examples (using 1,000 bootstrap replicates per example).

To generate a comparable set of posteriors with NPE, we first froze the trained `ReLERNN` network without dropout and removed the last fully-connected layer (that collapses a latent space of size 64 to a scalar output). With this "decapitated" pre-trained network, we generated 64-dimensional summary statistics for each input sequence. We used these learned **embeddings** of the genetic data to train an NPE model, then generated posterior distributions for the same test set that was used for parametric bootstrapping. Finally, we compared the NPE and bootstrapped uncertainty estimates by measuring confidence/credibility interval coverage (Figure 3) and concentration (Figure 4) across a range of interval widths. We found that both methods produced similarly well-calibrated intervals; however, the NPE procedure was far less computationally expensive, as it does not require generating novel simulations per prediction.

These results show that NPE can be used to learn posteriors for recombination rates with comparable calibration but large computational savings in comparison to parametric bootstrapping. We next move on to a simple two-dimensional problem: inferring the timing and magnitude of a bottleneck event.

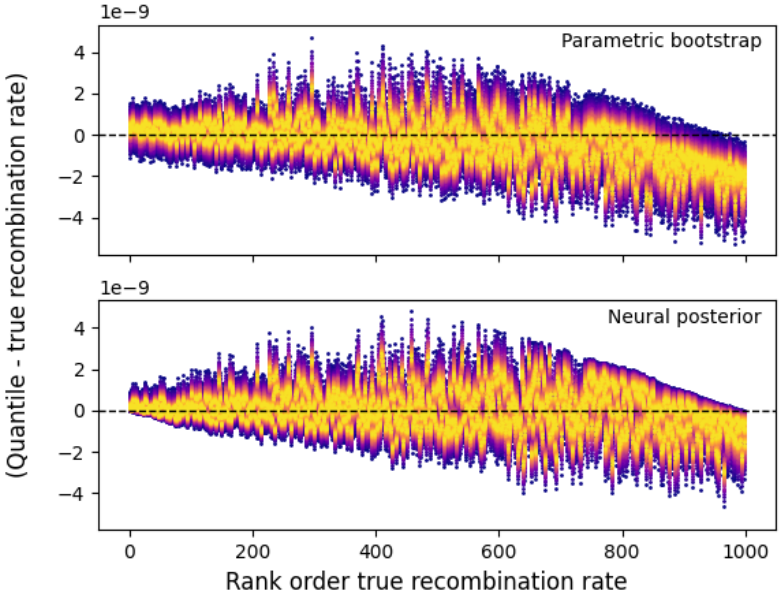


Figure 3: **Posterior intervals for recombination rate estimation**

Inference of Population Bottleneck Parameters

Our next task is to infer the timing and the intensity of a bottleneck event. For this task we use *Arabidopsis thaliana* as an example, building upon the the South Middle Atlas African demographic model from `stdpopsim` catalog (Huber et al. (2018); Adrion et al. (2020b); Lauterbur et al. (2023)).

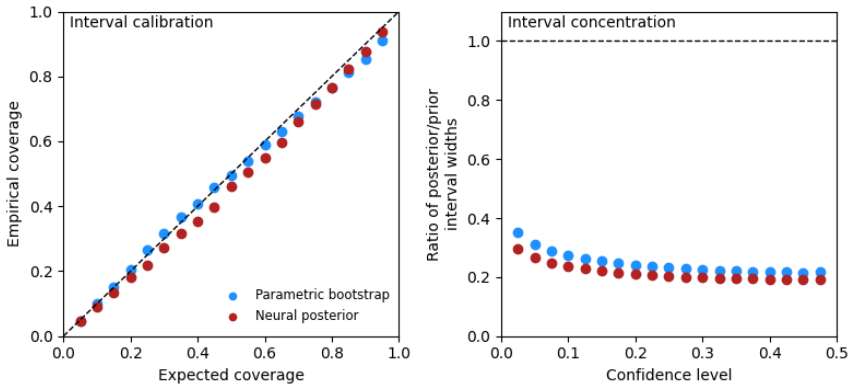


Figure 4: **Posterior coverage for recombination rate estimation**

Training simulations consisted of 10Mb of simulated data from chromosome 1 using the recombination map of Salomé et al. (2012). For our training set for NPE, we simulated 20,000 tree sequences with T (bottleneck time in generations from present divided by twice the ancestral population size) and ν (nu; the ratio between the current population and the ancestral population sizes) sampled from uniform distributions $\text{Unif}(0, 1.5)$ and $\text{Unif}(0, 1)$ respectively.

As a baseline comparison we used the `moments` package (Jouganous et al. (2017)) to infer the bottleneck parameters from the same simulated data. `moments` uses a composite likelihood approach to infer the bottleneck parameters from the site-frequency spectrum (SFS) of a sample and utilizes the Godambe information matrix to estimate the uncertainty of the parameter estimates (which is asymptotically correct, despite the model mis-specification intrinsic in composite likelihood). Throughout this section, we compare the performance of `moments` to NPE-derived estimates and posteriors using three different embedding approaches: 1) using the SFS as input as a form of summary statistics (i.e. without using the embedding network), 2) using a 1D convolutional neural network (CNN) to embed raw genotypes generated through simulation, and 3) using a recurrent neural network (RNN). We note that using the SFS as input to NPE amounts to a direct comparison to `moments` as the same information is available to both methods. As an additional point of comparison for this task we did ABC rejection sampling of our NPE training simulations, using a simple set of summary statistics known to be informative for population bottlenecks (π, the number of segregating sites, and Tajima's D). We retained the closest 1% of simulations as our posterior sample.

An important starting point is to compare the `moments` derived likelihood surface to the NPE-derived posteriors. In Figure 5 we show comparisons of the NPE posteriors to the `moments` likelihood surfaces and the ABC-derived posterior distribution for three independent points in parameter space (given on rows). Time of the bottleneck (T) is shown on the y-axis and the intensity of the bottleneck (ν) is shown on the x-axis. True parameter values are indicated by a red dot. For the NPE posteriors, we show the results from three different embedding approaches labeled npe-cnn, npe-rnn, and npe-sfs, respectively. As can be seen from the figure, the likelihood and posterior densities of T and ν show that the parameters are correlated in a non-linear fashion. This feature is problematic from the standpoint of

likelihood-based quantification of uncertainty, as canonical confidence intervals based on Fisher information, including Godambe information, are not appropriate as they assume a Gaussian form. Qualitatively, all methods contain the true parameter values within their respective areas of high density. However, there is some variation among the NPE embedding approaches.

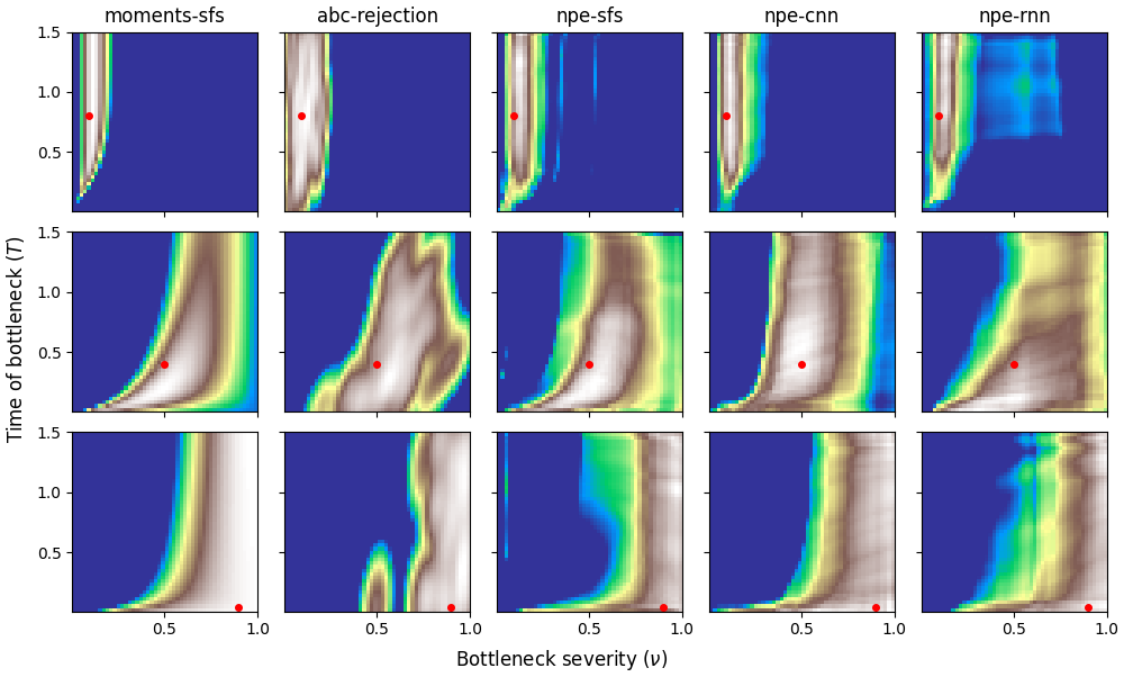


Figure 5: **Comparison of NPE posteriors to ABC and `moments` likelihood surfaces.** Here we show comparisons of the NPE posteriors to the ABC-derived posteriors and the `moments` likelihood surfaces for three independent points in parameter space (given on rows). Time of the bottleneck (T) is shown on the y-axis and the intensity of the bottleneck (ν) is shown on the x-axis. True parameter values are indicated by a red dot. For the NPE posteriors we show the results from three different embedding approaches labelled npe-cnn, npe-rnn, and npe-sfs. Colors indicate the likelihood/posterior density.

We can further characterize the performance of the NPE approach by looking at the point estimates and their associated uncertainty. In Figure 6 we show scatterplots of the NPE-derived point estimates compared to the `moments` and ABC point estimates for each of the three embedding approaches. As the NPE approach yields a distribution of point estimates, we show the mean of the posterior distribution as a point estimate, corresponding to the Bayes estimate. For each estimated we also show the estimated mean squared error (MSE) loss across the parameter space for each parameter (labeled MSE). As can be seen from the figure, the NPE-derived point estimates each yield a lower loss than the `moments` point estimates and equivalent to that from ABC, indicating that the NPE approach is able to accurately estimate the true parameter values. A caveat here is that our simulated data is generally "small", as we are only using 10Mb of simulated data, as we would expect `moments` to better with additional data. As an additional view of this comparison, Figure S1 shows boxplots of MSE loss for the NPE-derived point estimates compared to the `moments` and

ABC point estimates across parameter space for ν and T (left and right panels, respectively). Again, the general trend here is that the NPE-derived point estimates yield a slightly lower loss than the `moments` point estimates.

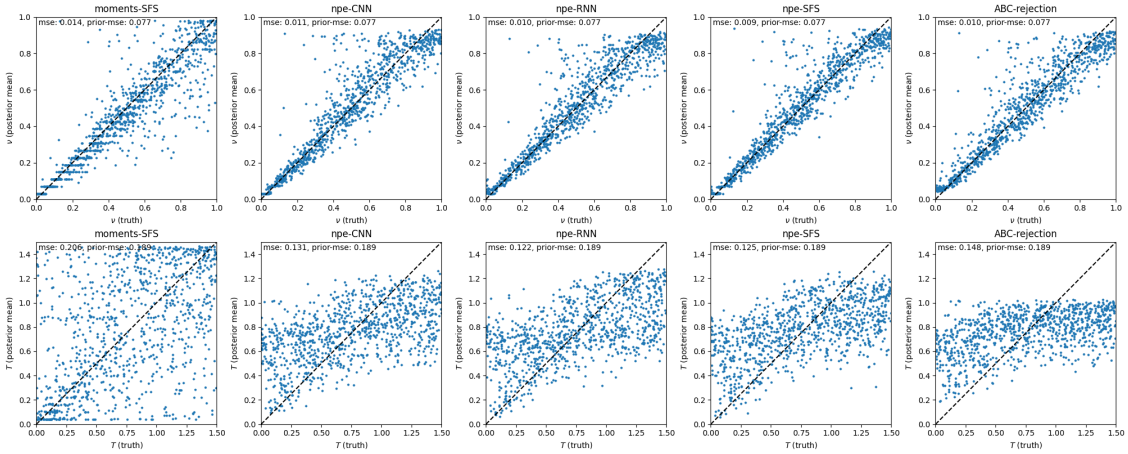


Figure 6: **Comparison of NPE point estimates to `moments` and ABC point estimates.** Here we show comparisons of NPE-derived point estimates to the `moments` and ABC point estimates across parameter space for T and ν (given on rows). Columns show the different embedding approaches: npe-cnn, npe-rnn, and npe-sfs. As the NPE approach yields a distribution of point estimates, we show the mean of the posterior distribution as a point estimate, corresponding to the Bayes estimate. Each facet additionally shows the estimated loss across the parameter space for each estimator (labeled MSE) as well as the prior loss implied by the uniform prior (labeled prior-mse).

Ultimately our goal is to quantify uncertainty in the point estimates, rather than point estimation itself. To this end, we can compare the coverage of the estimated posterior across our embedding approaches to the coverage of Godambe information matrix-derived confidence intervals for `moments` as well as that from the ABC-derived posterior. In Figure 7 we show the coverage of the NPE-derived posteriors across our embedding approaches to the coverage of Godambe information matrix (GIM)-derived confidence intervals for `moments`. There are two major takeaways from this figure. First, the NPE-derived posteriors display excellent coverage for both ν and T. As with our recombination rate results, this is a strong indication that the NPE-derived posteriors are well-calibrated. Second, the `moments`-derived confidence intervals are less well-calibrated for T, in particular as T gets larger. As pointed to above in looking at the posterior surfaces, this is a results of the non-linear relationship between T and ν; GIM-derived confidence intervals are not able to capture the full complexity of the likelihood surface, and as a result the confidence intervals are too narrow.

As the NPE and ABC procedure depend on a prior distribution, and the `moments` approach is based on bounded optimization, an additional comparison of posterior concentration is warranted. Ideally, any Bayesian approach should yield a posterior distribution that is more concentrated around the true parameter values than the prior distribution. In Figure S2 we show the ratio of the interval widths of the prior and posterior distributions across for ν and T for each of the three embedding approaches as well as to ABC a comparison point. We compare this to `moments` by using as a prior the bounds on the optimization and

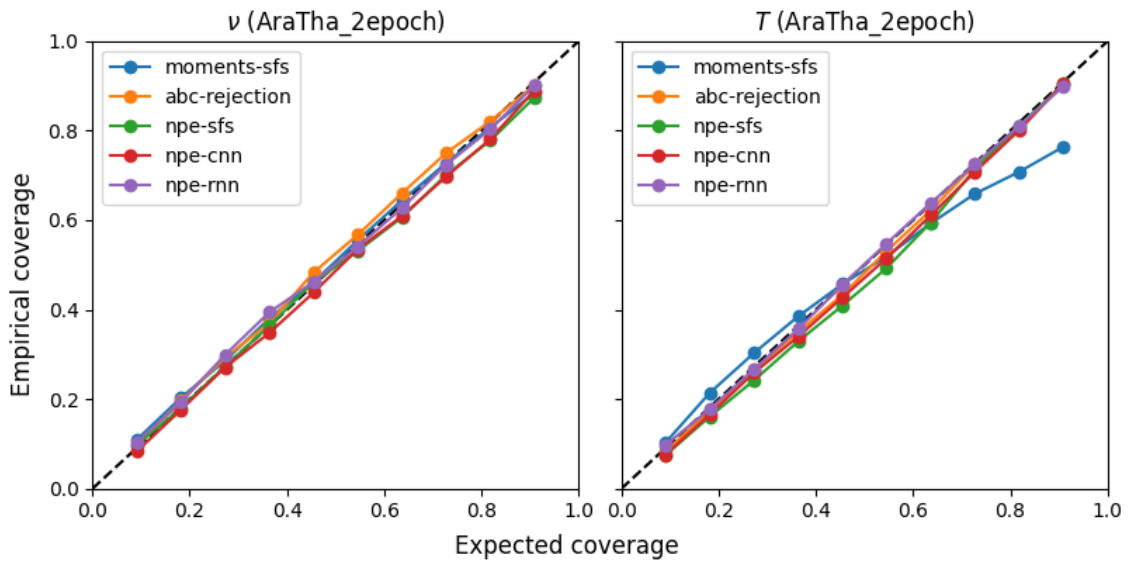


Figure 7: Comparison of NPE and ABC posterior coverage to `moments` confidence intervals. Here we show comparisons of the empirical and expected coverage of the NPE-derived and ABC-derived posteriors across our embedding approaches to the coverage of Godambe information matrix-derived confidence intervals for `moments`. The left and right panels show the coverage for ν and T, respectively. Labels are as in Figure 5.

considering as the posterior the interval width of the Godambe information matrix-derived confidence intervals. As can be seen from the figure, the NPE-derived posteriors are heavily concentrated and show a large reduction in the interval width compared to the prior for both parameters. Indeed for our choice of summary statistics, the NPE posterior is even more heavily concentrated than is the ABC-derived coverage. Again this is an indication that the NPE procedure is doing a good job of learning the posterior distribution.

Inference of Historical Population Sizes in a One-Population-Model

Another task that we were keen on exploring is the inference of effective population size over time. This is a common task in population genetics, and it is often used to infer the population size history of a species modulo caveats about the relationship between the effective population size and the census population size. While specialized methods for this task exist (e.g., Li and Durbin, 2011; Schiffels and Durbin, 2014; Terhorst et al., 2017), we were interested in seeing if a neural posterior estimator could be used to infer the effective population size over time. Indeed, others have explored the use of Bayesian methods for inference of effective population size over time both with and without neural networks (Boitard et al., 2016; Sanchez et al., 2020).

To train the neural posterior estimator, we simulated 40,000 tree sequences with `msprime`. For each of the 40,000 tree sequences, we simulated histories with 15 population size changes for samples of size $N = 20$ individuals. Population size changes occurred on a fixed grid of 16 log-spaced time points from 10^2 to 10^5 with $\Delta \log_{10} \approx 0.2$. Population sizes for each of the 15 time epochs were sampled independently from a log-uniform distribution with

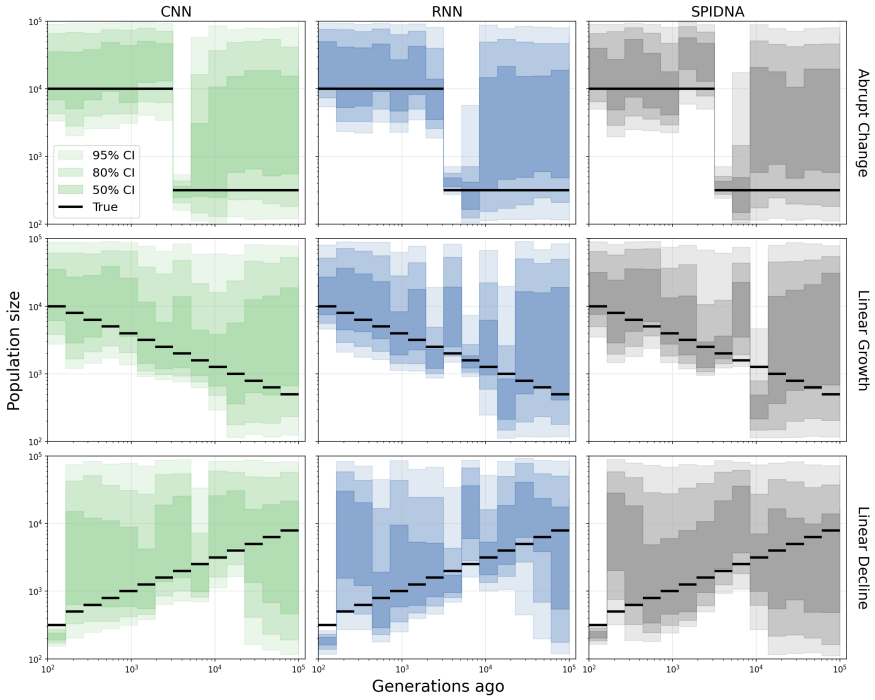


Figure 8: Benchmark of the neural posterior estimator on the task of inferring the effective population size over time using three different embedding networks. The three scenarios shown on rows are: 1) an abrupt change in population size, 2) population growth, and 3) population decline. The three embedding networks shown on columns are: a CNN embedding network, a RNN embedding network, and the SPIDNA embedding network from (Sanchez et al., 2020). For each plot we show the true population size history (black line) and the inferred credible intervals (shaded regions) for the 15 time epochs.

$\log_{10}(N_e) \sim \text{Uniform}(2.0, 5.0)$, corresponding to effective population sizes ranging from 100 to 100,000 individuals. Training of the neural posterior estimator then followed as in the previous sections. We compared three different embedding networks: a CNN embedding network (as above), a RNN embedding network, and finally the SPIDNA embedding network from (Sanchez et al., 2020). We chose to include the SPIDNA embedding network as it was specifically designed for embedding population genetics data for the task of population size inference.

We benchmarked the performance of the neural posterior estimator for inferring effective population size over time under three demographic scenarios: 1) an abrupt change in population size, 2) a scenario of a power-law population growth with exponent 1/2, and 3) a scenario of power-law population decline with exponent 1/2. In each case, the trained neural posterior estimator was used to infer the effective population size over time. Figure 8 shows the results of this benchmark for each of our three embedding networks. Generally, the neural posterior estimator is able to recover the true population size history relatively well, with each of the three embedding networks containing the true population size within the 50% credible interval for the vast majority of the time epochs. Generally performance declines moving back in time, as the amount of information available decreases. This can be seen

more clearly in Figure S3 where we have added a line showing the percentage of remaining lineages at each time point within the test simulation.

In Figure S4 we show the normalized credible interval (CI) width for each of the parameters estimated by the NPE for each of the scenarios and each of the embedding networks. Note that we estimated 15 population size parameters as well as the recombination rate. Normalized credible interval width was calculated as the 95% credible interval width divided by the geometric mean for population size parameters and by the posterior mean for the recombination rate parameter. The geometric mean was used for population sizes as the prior is log-uniformly distributed, where as the posterior mean for the recombination rate as its prior was drawn from a linear space. We see that there is some variation in the inferred CI width among embedding networks that differs across the considered demographic scenarios. There is a general tendency for CIs to grow wider for population size parameters that describe more ancient times, which we again interpret as the models reflecting less information as time moves towards the past.

Finally, in Figure S5 we show mean relative errors for aggregated point estimates (posterior means) of population size and recombination parameters across test scenarios and embedding networks. As with CIs, mean relative errors use geometric means for population size parameters and the simple posterior mean for the recombination rate. Across scenarios the three embedding networks are roughly comparable in their accuracy for population size parameters, with the RNN producing the most accuracy population size estimates for two of the three scenarios, and the CNN performing best in the case of the linear decline. For recombination rate we observe similar variation, where there is no clear universally best embedding network; although for the linear growth scenario, the CNN embedding yields a considerably better recombination rate estimate than do the other networks.

Priors for Historical Population Size Inference

Prior specification plays a crucial role in Bayesian inference. To explore the impact of prior choice, we introduced a second prior for training the neural posterior estimator. As an alternative to the uniform prior, where each historical population size is sampled independently, we used a structured dependent prior. This alternative prior is designed to generate more contiguous and realistic population size histories by inducing auto-correlation between population sizes in adjacent time intervals; i.e., in the dependent prior, each population size depends on its preceding value, such that changes in population size between neighboring time windows are constrained to be at most a factor of 10 in either direction.

Both the independent and dependent priors were used to separately simulate training data corresponding to 21 population size parameters from log-spaced time windows extending to 130,000 generations ago. As input for the NPE we used, instead of features extracted by an embedding net, the site frequency spectrum and linkage disequilibrium summary statistics calculated for each simulated dataset. We then analyzed the performance of neural posterior estimators trained on the summary statistics generated by the uniform prior compared to the dependent prior by inferring the effective population size histories under six exemplary demographic scenarios.

Figure 9 illustrates the reconstructed effective population size histories for these scenarios. Both neural posterior estimators successfully capture the general demographic trends, with

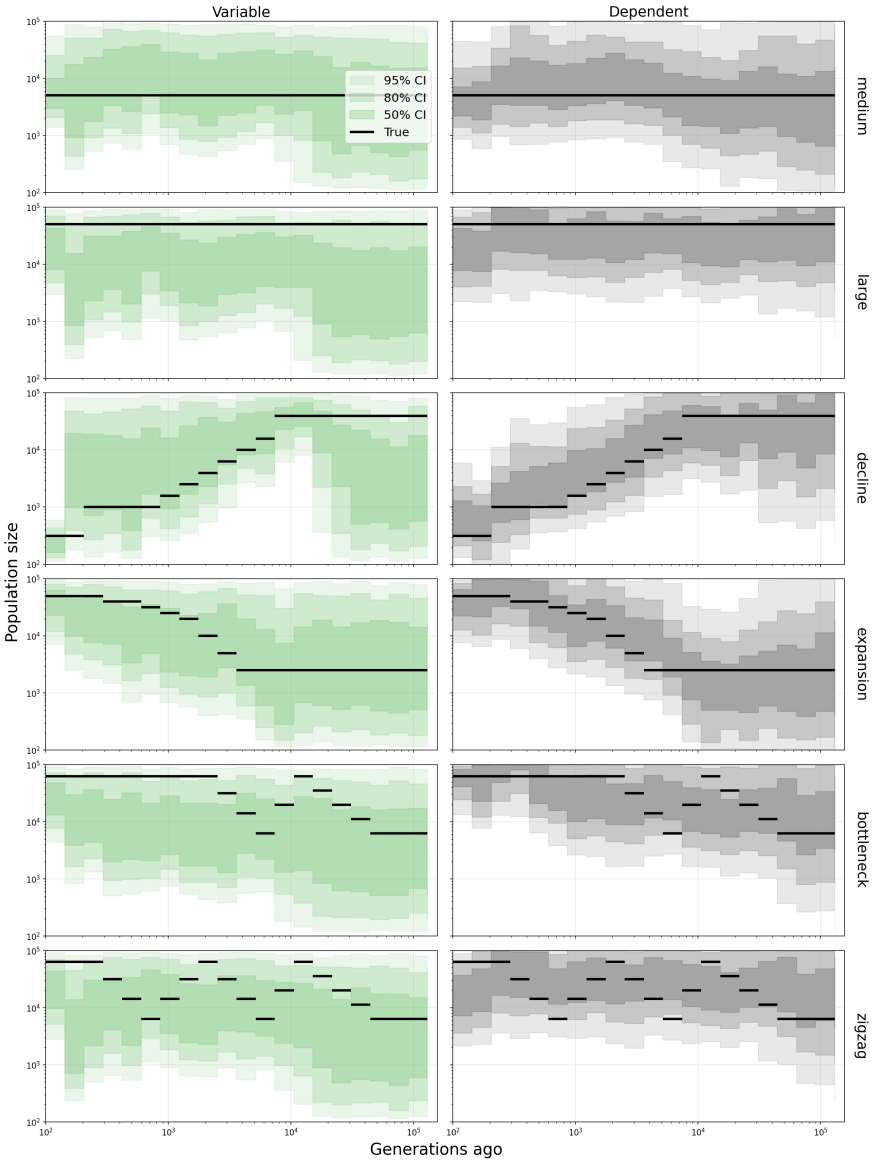


Figure 9: Inference of effective population sizes using variable and dependent prior under different scenarios. Scenarios shown in rows are: "Medium" and "Large" constant population sizes. "Decline" and "Expansion" where the population sizes slowly drop and increase. "Bottleneck" and "Zigzag" where population sizes fluctuate over time.

true population sizes largely falling within the 50% credible intervals. However, the choice of prior had a substantial impact on the inference quality and certainty.

In the "large" scenario, the neural posterior estimator trained on the dependent prior achieved a higher accuracy, when the one using uniform prior simply "missed" the ground truth in a large part of all time windows. We hypothesize that the uniform prior struggles to generate enough demographies with constant large population sizes, as every single historical population size window is drawn independently and uniformly, setting the population size more frequently to intermediate and low values. Therefore, scenarios close to the "large"

scenario are less likely generated by the independent prior training data. In contrast, such scenarios occur more likely in the training data generated from the dependent prior, which improves the performance of NPE for these scenarios. In addition, the dependent prior also shows a relatively smooth and confident reconstruction of the "decline" and "expansion" scenarios.

Just as any other population genetic feature, the summary statistics used here provide less information for more ancient times (particularly beyond the MRCA), and consequently the neural posterior estimator trained on the log-uniform prior quickly reverts to the prior distribution in the ancient time windows. Meanwhile, the contiguous structure of the dependent prior leads the neural posterior estimator to project more constrained estimates backward in time. For the complex "bottleneck" and "zigzag" scenarios, both priors resulted in NPEs with reduced accuracy and confidence, struggling to fully resolve the population size fluctuations.

Similarly to the previous example, we quantify these performance differences using the mean relative errors in Figure S6. Consistent with the qualitative results, the neural posterior estimator trained on the dependent prior demonstrates lower relative error for effective population sizes estimation across all scenarios except for "zigzag". In contrast, for the estimation of the recombination rate, the neural posterior estimator trained on the uniform prior consistently achieves higher accuracy. This might be caused by the highly variable effective population size histories sampled under the uniform prior, where information from neighboring time windows provides less information such that the neural posterior estimator might be encouraged to more robustly disentangle the genomic signals of recombination from those of population size.

Drosophila melanogaster out-of-Africa model

As an example application to empirical data, we used our NPE pipeline to fit a joint demographic model for samples of African and European *Drosophila melanogaster*. In particular, we analyzed inbred samples from France and Cameroon, collected as part of the DPGP2 project (Pool et al., 2012), and fitted a modification of the Li and Stephan (2006) model. This model was originally used to represent the out-of-Africa expansion of *Drosophila*, and incorporates divergence from an ancestral population, migration, and exponential growth in the European lineage. Our modified model has seven free parameters: the ancestral population size, the effective sizes of the African and European populations following the split, the contemporary European effective size, the divergence time between populations, and asymmetric migration rates between the European and African populations. As described in the Methods section, we used an RNN embedding network trained on 300 Kb non-overlapping windows across chromosome 2L (we focus on 2L purely out of convenience and as a proof of principle).

In Figure 10 (Table S8), we show averaged posterior estimates of the seven demographic model parameters, with marginal distributions shown on the diagonal, and joint distributions shown in off-diagonal plots. Generally, all parameters show relatively narrow credible intervals, and distributions of population size estimates are particularly sharp, with the one exception of the founding size of the French population. Our population size estimates are in good agreement with those of Li and Stephan (2006), although we have used a different chromosome

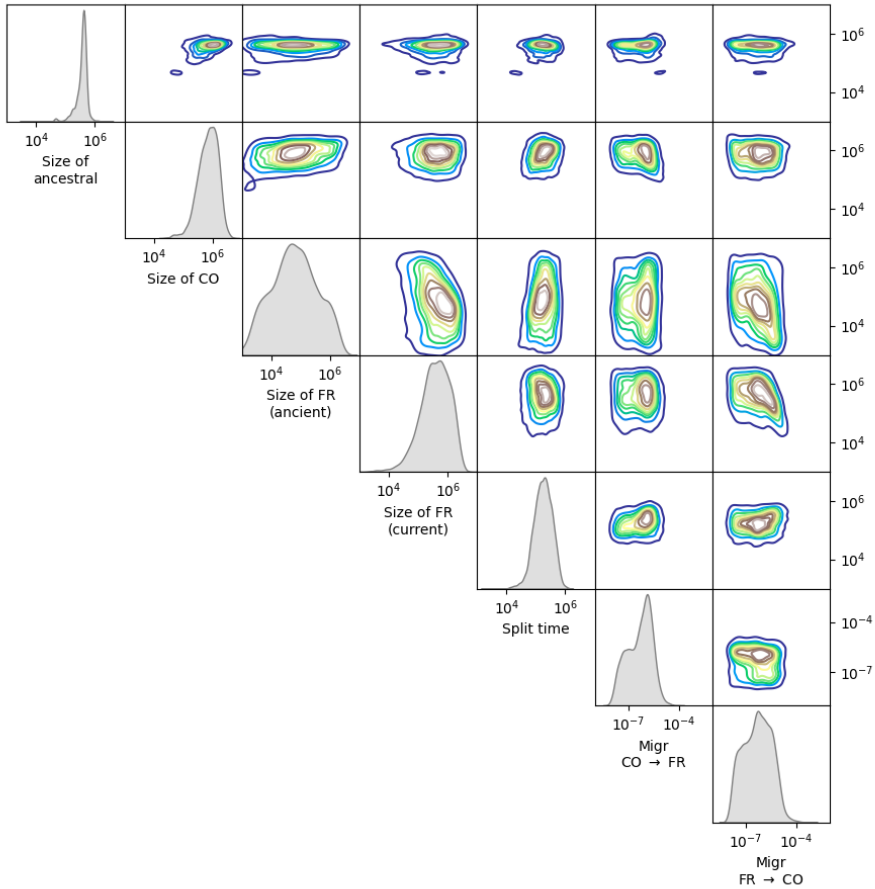


Figure 10: Joint posteriors for an *Drosophila melanogaster* out-of-Africa demographic model, fit via a neural posterior estimator with an RNN embedding network to 10 haploids from Africa (Cameroon) and 9 from Europe (France). Posteriors were estimated from unpolarised genotypes for 300Kb windows across 2L (see Fig S7, Fig 11), and then averaged to produce the contours shown here. Migration rates are backwards-in-time.

arm (the initial study used chromosome X) and different population samples (i.e., Li and Stephan used flies from the Netherlands and Zimbabwe). We estimate the current French effective population size is approximately 1×10^6 and our estimate of the effective population size in Cameroon is appears to be a bit larger than that. We don't find evidence in our estimates of as strong a bottleneck in the founding of the French population as was found in Li and Stephan (2006)— our estimate (posterior median) is on the order of $N_e \sim 82,000$ which is considerably larger than the estimate of $N_e = 2,200$ reported in Li and Stephan. However the 95% credibility interval overlaps with the confidence interval on founding European N_e presented in that paper, and indeed our uncertainty of that parameter is larger than all other parameters examined. Our estimate of the split time between populations is on the

order of $t_{div} \sim 180,000$ generations ago, and is thus quite comparable to estimates from Li and Stephan, who estimated divergence time between Netherlands and Zimbabwe to be approximately 158,000 generations ago. Examination of the joint posterior distributions reveals little posterior correlation among the majority of parameters, with most joint posteriors showing generally uncorrelated, elliptical shapes. A notable exception is the joint posterior between the current French effective population size and the migration parameter which determines the rate of influx (forward-in-time) from Cameroon to France, which show some negative correlation (Pearson correlation of $r = -0.4$, pooling posterior samples over windows). This makes sense from the perspective of diversity—higher levels of contemporary genetic variation in France could be the result of larger effective population size or greater migration from Africa.

As we have estimated demographic parameters on a window-by-window basis, we can also examine how our estimates vary systematically across chromosome 2L (Fig S7). A few features are worth noting here. First, we see increased variation in effective population size parameters towards the distal (telomeric) and proximal (centromeric) regions of 2L. This is most likely due to the increased importance of linked selection in determining patterns of polymorphism in regions of low recombination, in the *Drosophila* genome (Aguade et al., 1989; Begun and Aquadro, 1992, e.g.,). We also see an interesting 'blip' in the estimates of split time and migration in the second-closest window to the centromere. We can only speculate at this point that this may be pointing to local genomic processes, such as an adaptive allele, but it perhaps echoes some of the unique dynamics found in centromere-proximal regions from humans (Langley et al., 2019, and C. Langley *pers. comm.* regarding flies).

To verify that the per-window variability in demographic parameter estimates is accurate, we performed posterior predictive simulations for each window, in which we drew samples from the per-window posterior, conducted coalescent simulations, and then summarized those simulations using a collection of summary statistics not used in estimation. The results of our posterior predictive checks are shown in Figure 11. As can be seen clearly, our posterior predictive simulations produce per-window variation in observed summary statistics that very closely matches the empirical summaries of the data. This gives us confidence that our NPE procedure is working well and providing meaningful fits to the data.

Discussion

In this study, we have demonstrated that neural posterior estimation (NPE) provides a powerful and flexible framework for population genetic inference that combines the principled uncertainty quantification of Bayesian methods with the computational efficiency and automatic feature extraction capabilities of deep learning. Our results across multiple inference tasks—recombination rate estimation, demographic parameter inference, and effective population size reconstruction—consistently show that NPE produces well-calibrated posterior distributions from both summary statistics and learned neural representations of raw genomic data.

The most striking advantage of NPE over existing methods is its ability to learn complex, non-linear posterior distributions directly from data. Unlike composite likelihood approaches such as `moments`, which rely on Gaussian approximations through asymptotic variance-

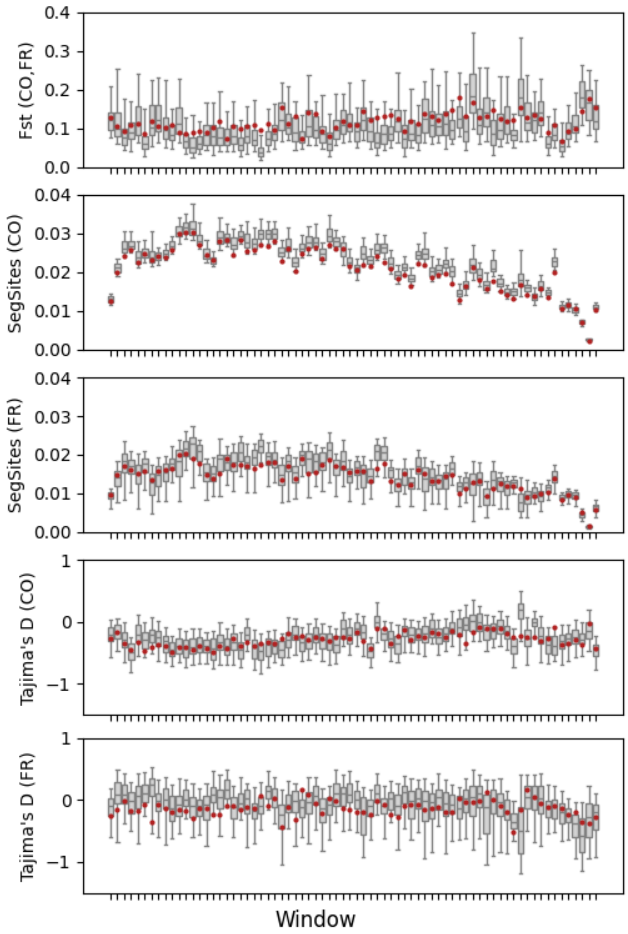


Figure 11: Posterior predictive checks for a *Drosophila melanogaster* out-of-Africa demographic model, fit via a neural posterior estimator with an RNN embedding network in 300Kb windows across 2L. For each window, demographic parameters were sampled from the posterior, genotype data was simulated with `msprime`, and posterior summary statistics (boxplots) plotted alongside to observed values (red points). The demographic parameter estimates in each window are shown in Fig S7.

covariance matrices, NPE captures the full shape of the posterior distribution. This was particularly evident in our bottleneck inference results (Figure 5), where the non-linear correlation between bottleneck timing (T) and intensity (ν) was accurately represented by NPE but led to miscalibrated confidence intervals when using Godambe information matrix approximations. This finding has important implications for demographic inference, where parameter correlations are common and often non-linear due to the complex relationship between demographic events and genetic patterns.

Compared to maximum likelihood based approaches, NPE offers dramatic computational savings through amortization. Once trained, an NPE model can generate posterior distributions for new observations almost instantaneously, whereas likelihood based-inference requires new optimization runs for each new obseration. Our recombination rate estimation results exemplify this advantage in a comparison to a previous method developed by our group

(Adrion et al., 2020a): while parametric bootstrapping required 1,000 new simulations per genomic window to generate confidence intervals, the trained NPE model could produce equally well-calibrated credible intervals without any additional simulation. For genome-wide analyses involving thousands of windows, this represents a computational speedup of several orders of magnitude.

Furthermore, NPE's flexibility in handling different data representations — from hand-crafted summary statistics to raw genomic data processed through embedding networks — allows researchers to choose the most appropriate representation for their specific problem. For instance, our results show that SFS summaries can be used effectively in conjunction with NPE, and thus can produce richer posterior representations than likelihood-based inference alone; however, incorporating linkage disequilibrium information through learned representations substantially improves inference accuracy and posterior concentration. This flexibility extends beyond current methods: as new network architectures are developed for genomic data, they can be readily incorporated into the NPE framework as embedding networks.

Our comparison of different data representations reveals important insights about information content in population genetic data. The superior performance of methods incorporating linkage information over SFS-alone approaches underscores that single-site summaries, while mathematically convenient, discard substantial information about population history. This finding aligns with recent theoretical work showing that linkage patterns contain independent information about demographic history, particularly for recent events and complex scenarios involving multiple populations (e.g., Jouganous et al., 2017).

The success of end-to-end learning approaches using CNNs and RNNs demonstrates that neural networks can automatically discover informative features from raw genetic data, potentially capturing patterns that would be difficult to encode in traditional summary statistics. However, our results also show that carefully chosen summary statistics can perform competitively with or potentially even outperform end-to-end approaches in some scenarios. Moreover, we can leverage theoretical results to improve the performance of machine learning methods based on predefined summary statistics (Burger et al., 2022) and interpret feature importance. This suggests that the optimal choice between summary statistics and end-to-end learning may be problem-specific, depending on factors such as the dimensionality of the parameter space, the complexity of the demographic model, existing theoretical insights, computational costs for summary statistics, and the amount of available training data.

Challenges and Limitations

Despite these advantages, several challenges remain in applying NPE to population genetic problems. First, the quality of posterior estimates depends critically on the realism of the simulations used for training. While we used neutral coalescent simulations throughout this study, real genomic data is shaped by both demographic history and natural selection. Incorporating selection into the training simulations, while computationally more demanding, will be essential for applying NPE to empirical data where the effects of background selection and selective sweeps cannot be ignored. This challenge of model misspecification, of course, touches not just simulation-based methods of inference, as we present here, but also traditional,

statistical methods which make assumptions about the distribution from which data is drawn.

Second, NPE requires choosing an appropriate prior distribution for the parameters of interest. While this is a general challenge in Bayesian inference, it is particularly acute for NPE because the method learns the posterior-to-prior ratio. Poorly chosen priors can lead to inefficient learning or biased posteriors. We primarily used uniform and log-uniform priors, which represents a relatively uninformative choice. However, we also showed that more complex and informative priors that incorporate biological knowledge can improve the performance (Figure 9).

Third, scaling NPE to high-dimensional parameter spaces remains challenging. While we successfully inferred up to 22 parameters (21 population sizes plus recombination rate), many population genetic models of interest involve substantially more parameters. Recent advances in normalizing flow architectures and training procedures may help address this limitation, but further research is needed to determine the practical limits of NPE for complex models.

Implications for Population Genomics

The ability of NPE to provide rapid, well-calibrated uncertainty estimates has important implications for population genomic studies. In conservation genetics, where demographic inference informs management decisions, having reliable confidence intervals is crucial for risk assessment. NPE's ability to quantify uncertainty while accounting for parameter correlations provides more reliable inference than methods assuming Gaussian posteriors.

For human population genomics, where complex models involving multiple populations, admixture events, and migration are the norm, NPE offers a path toward more realistic inference. The computational efficiency of amortized inference makes it feasible to analyze genomic data from thousands of individuals, potentially revealing fine-scale population structure and history that would be computationally prohibitive with ABC or composite likelihood methods.

Moreover, NPE's modular structure—separating feature extraction (embedding networks) from density estimation (normalizing flows)—aligns well with the modular nature of population genetic models and simulators. Different embedding networks can be designed for different data types (SNPs, structural variants, microsatellites), while the same normalizing flow architecture can be used for posterior estimation. This modularity facilitates the development of specialized tools for different inference problems while maintaining a common statistical framework.

Future Directions

Several promising directions emerge from this work. First, extending NPE to handle missing data and uncertainty in data preprocessing (such as genotype-calling errors) would increase its applicability to real datasets. Second, developing embedding networks specifically designed for population genetic data, perhaps incorporating known properties of the coalescent process, could improve both accuracy and interpretability.

The combination of NPE with model selection and model averaging procedures represents another important direction. While the technology for model selection through Bayes factors is well-established in the context of NPE, more sophisticated approaches could allow

simultaneous inference of model structure and parameters (Schröder and Macke, 2024). This would be particularly valuable for scenarios where the demographic model itself is uncertain.

Finally, the integration of NPE with functional genomic data could enable joint inference of demographic history and selection, which is a particularly challenging and important goal (Boyko et al., 2008; Kim et al., 2017; Johri et al., 2020, 2021). By incorporating information about gene function, expression, and regulation into the embedding networks, it may be possible to disentangle the effects of demography and selection more effectively than current methods allow.

In conclusion, neural posterior estimation represents a significant methodological advance for population genetics, combining the best features of likelihood-based, ABC, and machine learning approaches. As genomic datasets continue to grow in size and complexity, methods like NPE that can efficiently extract information while providing principled uncertainty quantification will become increasingly important. The framework presented here provides a foundation for future developments in simulation-based inference, with the potential to enable more realistic and informative analysis of population genomic data.

Acknowledgments

We thank Dan Schrider for helpful comments on the manuscript. We thank Jakob Macke, Cornelius Schröder, and members of the Kern-Ralph colab for discussion and input on this project. J.M., N.S.P, and A.D.K were supported in part by NIH grants R01HG010774, R01HG012473, and R35GM148253. F.B. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645, and EXC 2124 – Project number 390838134.

References

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLOS Genetics*, 5(10):1–11, 10 2009. doi: 10.1371/journal.pgen.1000695. URL https://doi.org/10.1371/journal.pgen.1000695.

Alison M Adams and Richard R Hudson. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3):1699–1712, 2004.

Julien Jouganous, Will Long, Aaron P Ragsdale, and Simon Gravel. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, 206(3):1549–1567, 2017.

Aaron P Ragsdale and Simon Gravel. Models of archaic admixture and recent history from two-locus statistics. *PLoS genetics*, 15(6):e1008204, 2019.

Richard R Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4): 1805–1817, 2001.

Gil McVean, Philip Awadalla, and Paul Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–1241, 2002.

Andrew H Chan, Paul A Jenkins, and Yun S Song. Genome-wide fine-scale recombination rate variation in drosophila melanogaster. *PLoS genetics*, 8(12):e1003090, 2012.

Jeffrey P Spence and Yun S Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10): eaaw9206, 2019.

Carlos D Bustamante, John Wakeley, Stanley Sawyer, and Daniel L Hartl. Directional selection and the site-frequency spectrum. *Genetics*, 159(4):1779–1788, 2001.

Carlos D Bustamante, Rasmus Nielsen, and Daniel L Hartl. Maximum likelihood and bayesian methods for estimating the distribution of selective effects among classes of mutations using dna polymorphism data. *Theoretical population biology*, 63(2):91–103, 2003.

Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic scans for selective sweeps using snp data. *Genome research*, 15(11):1566–1575, 2005.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–35, Dec 2002.

Kevin Thornton and Peter Andolfatto. Approximate bayesian inference reveals evidence for a recent, severe bottleneck in a netherlands population of drosophila melanogaster. *Genetics*, 172(3):1607–1619, 2006.

Stefan JY Laurent, Annegret Werzner, Laurent Excoffier, and Wolfgang Stephan. Approximate bayesian analysis of drosophila melanogaster polymorphism data reveals a recent colonization of southeast asia. *Molecular biology and evolution*, 28(7):2041–2051, 2011.

Chris CR Smith and Samuel M Flaxman. Leveraging whole genome sequencing data for demographic inference with approximate bayesian computation. *Molecular ecology resources*, 20(1):125–139, 2020.

Dennis Prangle. Summary statistics in approximate bayesian computation, 2015.

Daniel R Schrider and Andrew D Kern. Supervised machine learning for population genetics: A new paradigm. *Trends Genet*, Jan 2018. doi: 10.1016/j.tig.2017.12.005.

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.

Kao Lin, Haipeng Li, Christian Schlotterer, and Andreas Futschik. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. *Genetics*, 187(1):229–244, 2011.

Roy Ronen, Nitin Udpa, Eran Halperin, and Vineet Bafna. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–93, Sep 2013. doi: 10.1534/genetics.113.152587.

Daniel R Schrider and Andrew D Kern. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. *Genome biology and evolution*, 7(12):3511–3528, 2015.

Daniel R Schrider and Andrew D Kern. S/hic: Robust identification of soft and hard sweeps using machine learning. *PLoS Genet*, 12(3):e1005928, Mar 2016. doi: 10.1371/journal.pgen.1005928.

Daniel R Schrider, Julien Ayroles, Daniel R Matute, and Andrew D Kern. Supervised machine learning reveals introgressed loci in the genomes of drosophila simulans and d. sechellia. *PLoS genetics*, 14(4):e1007341, 2018.

Kao Lin, Andreas Futschik, and Haipeng Li. A fast estimate for the population recombination rate based on regression. *Genetics*, 194(2):473–484, 2013.

Jeffrey Chan, Valerio Perrone, Jeffrey Spence, Paul Jenkins, Sara Mathieson, and Yun Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems*, 31, 2018.

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.

Jeffrey R Adrion, Jared G Galloway, and Andrew D Kern. Predicting the landscape of recombination using deep learning. *Molecular biology and evolution*, 37(6):1790–1808, 2020a.

Graham Gower, Pablo Iáñez Picazo, Matteo Fumagalli, and Fernando Racimo. Detecting adaptive introgression in human evolution using convolutional neural networks. *Elife*, 10:e64669, 2021.

Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2021.

Luis Torada, Lucrezia Lorenzon, Alice Beddis, Ulas Isildak, Linda Pattini, Sara Mathieson, and Matteo Fumagalli. Imagene: a convolutional neural network to quantify natural selection from genomic data. *BMC bioinformatics*, 20:1–12, 2019.

Xaviera Alejandra López-Cortés, Felipe Matamala, Carlos Maldonado, Freddy Mora-Poblete, and Carlos Alberto Scapim. A deep learning approach to population structure inference in inbred lines of maize. *Frontiers in Genetics*, 11:543459, 2020.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 837–848. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/papamakarios19a.html.

David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference, 2019.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators, 2020.

Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 2022.

Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. sbi: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505. URL https://doi.org/10.21105/joss.02505.

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):iyac131, 2022.

Jeffrey R Adrion, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, et al. A community-maintained standard library of population genetic models. *elife*, 9:e54967, 2020b.

M Elise Lauterbur, Maria Izabel A Cavassim, Ariella L Gladstein, Graham Gower, Nathaniel S Pope, Georgia Tsambos, Jeffrey Adrion, Saurabh Belsare, Arjun Biddanda, Victoria Caudill, et al. Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *Elife*, 12:RP84874, 2023.

Zhanpeng Wang, Jiaping Wang, Michael Kourakos, Nhung Hoang, Hyong Hark Lee, Iain Mathieson, and Sara Mathieson. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, 21(8):2689–2705, 2021. doi: https://doi.org/10.1111/1755-0998.13386. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13386.

Graham Gower, Pablo Iáñez Picazo, Finn Lindgren, and Fernando Racimo. Inference of population genetics parameters using discriminator neural networks: an adversarial monte carlo approach. *bioRxiv*, 2023. doi: 10.1101/2023.04.27.538386. URL https://www.biorxiv.org/content/early/2023/04/29/2023.04.27.538386.

Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2020. doi: 10.1111/1755-0998.13224. URL https://doi.org/10.1111/1755-0998.13224.

Alan R Rogers and Chad Huff. Linkage disequilibrium between loci with unknown phase. *Genetics*, 182(3):839–844, 2009.

Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

John E Pool, Russell B Corbett-Detig, Ryuichi P Sugino, Kristian A Stevens, Charis M Cardeno, Marc W Crepeau, Pablo Duchen, JJ Emerson, Perot Saelao, David J Begun, et al. Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. *PLoS genetics*, 8(12):e1003080, 2012.

Haipeng Li and Wolfgang Stephan. Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS genetics*, 2(10):e166, 2006.

Daniel R Schrider, David Houle, Michael Lynch, and Matthew W Hahn. Rates and genomic consequences of spontaneous mutational events in drosophila melanogaster. *Genetics*, 194 (4):937–954, 2013.

C. D. Huber, A. Durvasula, A. M. Hancock, et al. Gene expression drives the evolution of dominance. *Nature Communications*, 9:2750, 2018. doi: 10.1038/s41467-018-05281-7.

P. Salomé, K. Bomblies, J. Fitz, et al. The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*, 108:447–455, 2012. doi: 10.1038/hdy.2011.95.

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, 49(2): 303–309, 2017.

Simon Boitard, Willy Rodríguez, Flora Jay, Stefano Mona, and Frédéric Austerlitz. Inferring population size history from large samples of genome-wide molecular data: An approximate bayesian computation approach. *PLoS Genetics*, 12(3):e1005877, 2016. doi: 10.1371/journal.pgen.1005877.

Montserrat Aguade, Naohiko Miyashita, and Charles H Langley. Reduced variation in the yellow-achaete-scute region in natural populations of drosophila melanogaster. *Genetics*, 122(3):607, 1989.

David J Begun and Charles F Aquadro. Levels of naturally occurring dna polymorphism correlate with recombination rates in d. melanogaster. *Nature*, 356(6369):519–520, 1992.

Sasha A Langley, Karen H Miga, Gary H Karpen, and Charles H Langley. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic dna. *elife*, 8: e42989, 2019.

Klara Elisabeth Burger, Peter Pfaffelhuber, and Franz Baumdicker. Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown. *PLOS Computational Biology*, 18(8):e1010407, 2022. doi: 10.1371/journal.pcbi.1010407.

Cornelius Schröder and Jakob H. Macke. Simultaneous identification of models and parameters of scientific simulators. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43895–43927. PMLR, 2024.

Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, 4(5):e1000083, 2008.

Bernard Y Kim, Christian D Huber, and Kirk E Lohmueller. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, 206 (1):345–361, 2017.

Parul Johri, Brian Charlesworth, and Jeffrey D Jensen. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192, 2020.

Parul Johri, Kellen Riall, Hannes Becher, Laurent Excoffier, Brian Charlesworth, and Jeffrey D Jensen. The impact of purifying and background selection on the inference of population history: problems and prospects. *Molecular biology and evolution*, 38(7):2986–3003, 2021.

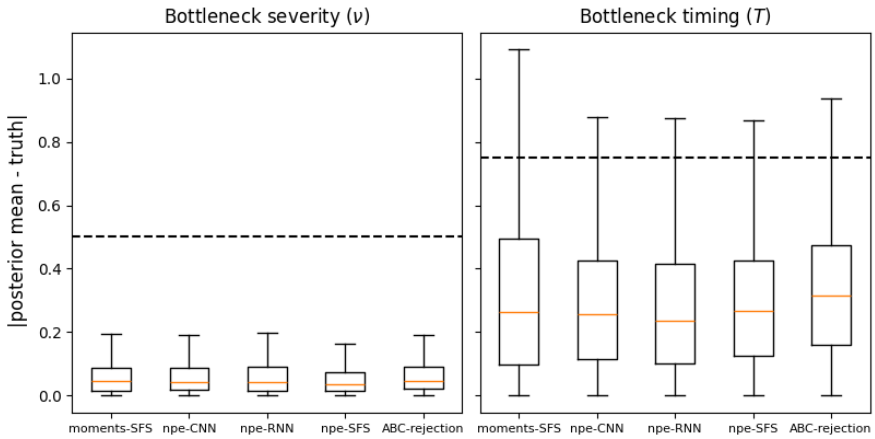


Figure S1: **Comparison of NPE point estimates to `moments` point estimates.** Here we show boxplot comparisons of NPE-derived point estimates to the `moments` point estimates across parameter space for ν and T (left and right panels, respectively). Shown are comparisons of MSE loss across the parameter space for each estimator.

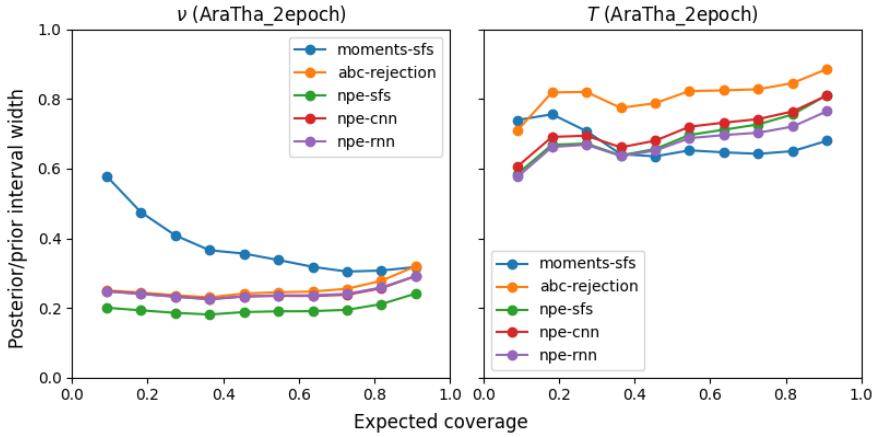


Figure S2: **Comparison of NPE posterior concentration to prior.** Here we show comparisons of the prior and posterior distributions for nu and T for each of the three embedding approaches. Additionally we show estimated concentration for the `moments` estimates where the prior in this case represents the bounds on the optimization.

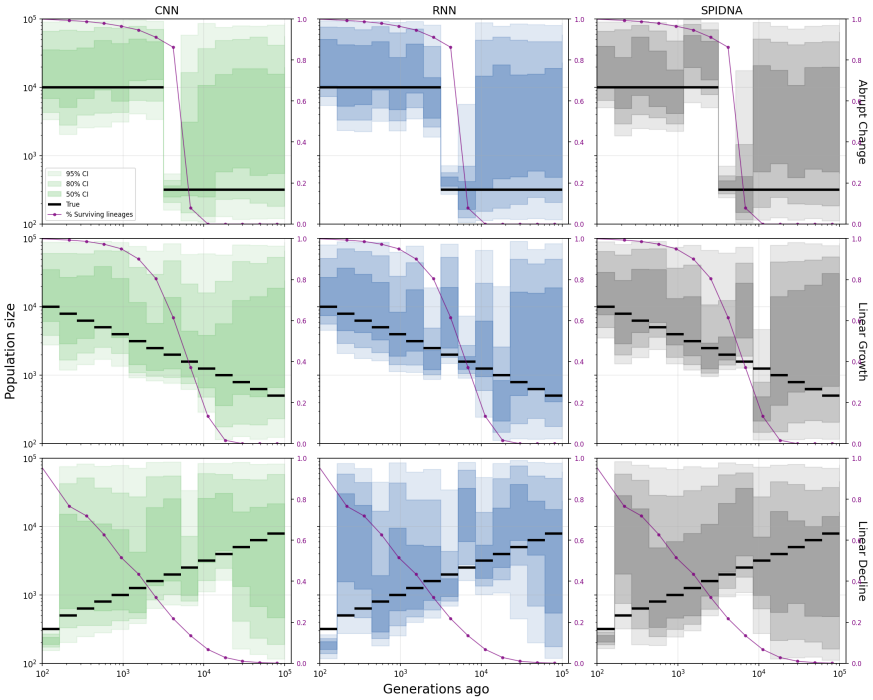
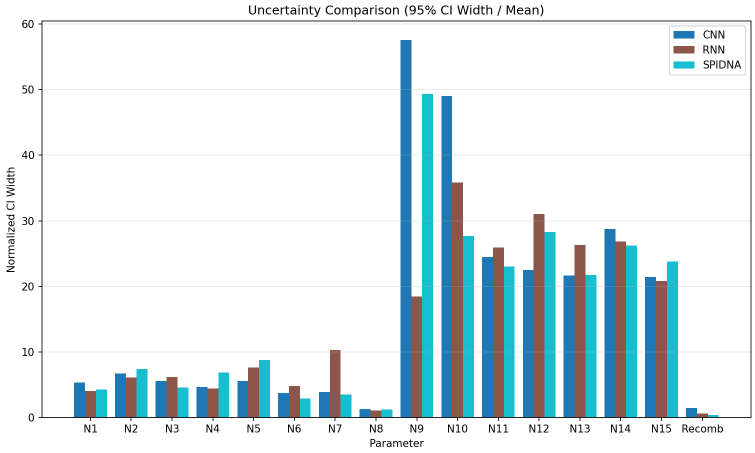
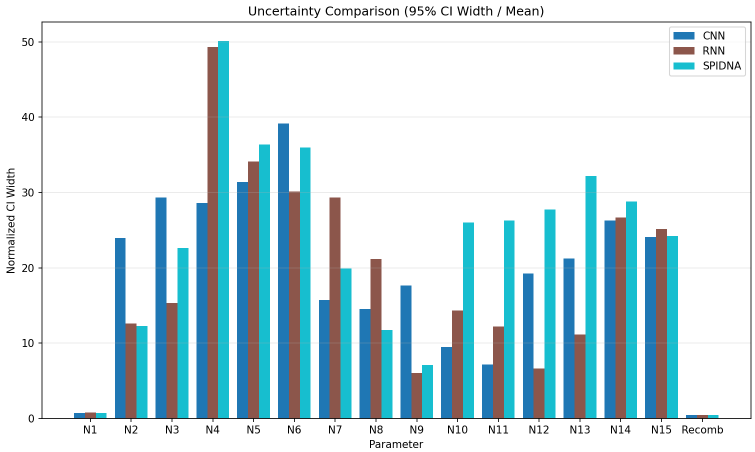


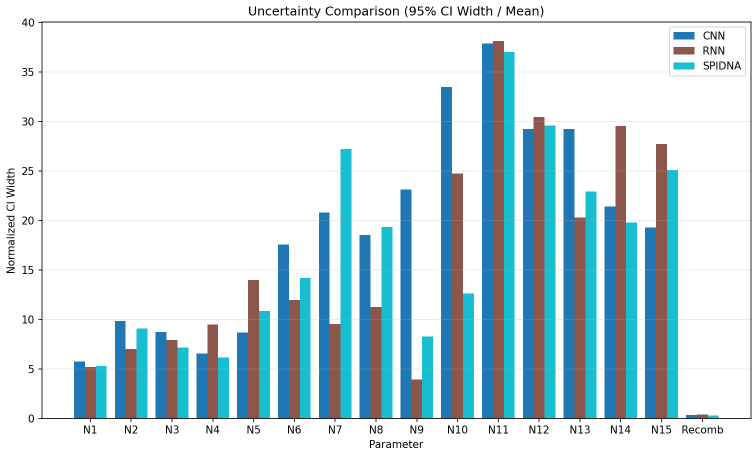
Figure S3: Benchmark of the neural posterior estimator on the task of inferring the effective population size over time using three different embedding networks. The three scenarios shown on rows are: 1) an abrupt change in population size, 2) a scenario of linear population growth, and 3) a scenario of exponential population growth. The three embedding networks shown on columns are: a CNN embedding network, a RNN embedding network, and the SPIDNA embedding network from (Sanchez et al., 2020). For each plot we show the true population size history (black line) and the inferred credible intervals (shaded regions) for the 15 time epochs. The purple line in each figure represents the percentage of remaining lineages at each time point within the test simulation.



(a) Abrupt change

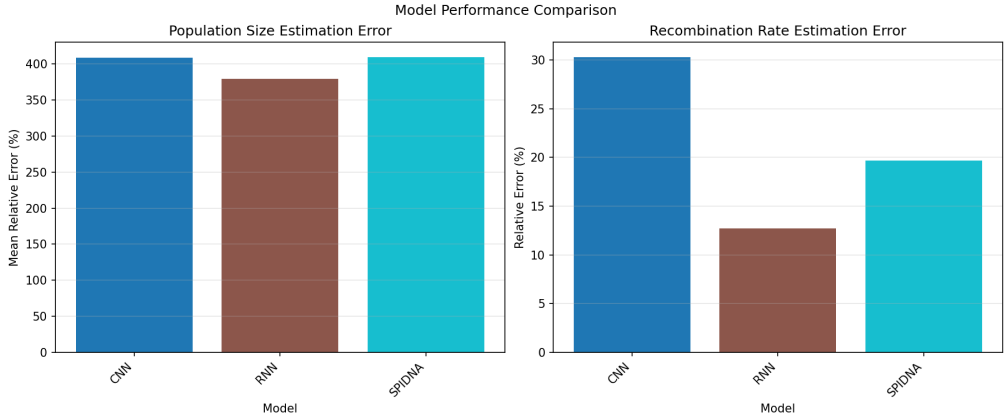


(b) Linear growth

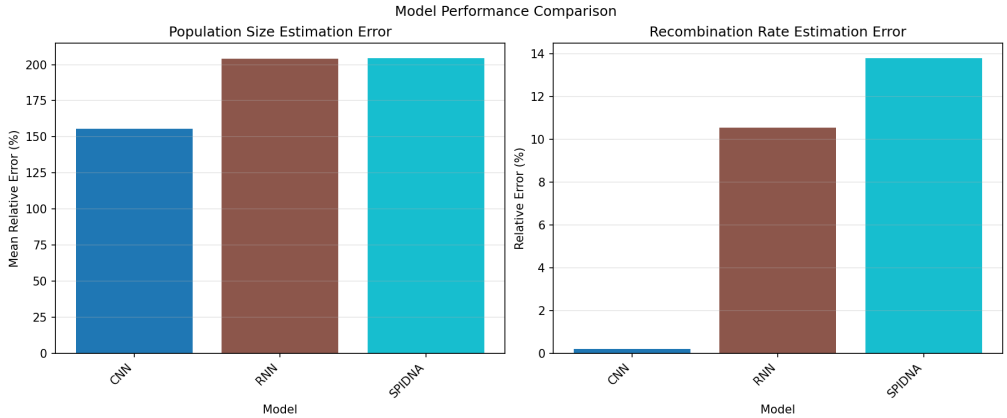


(c) Linear decline

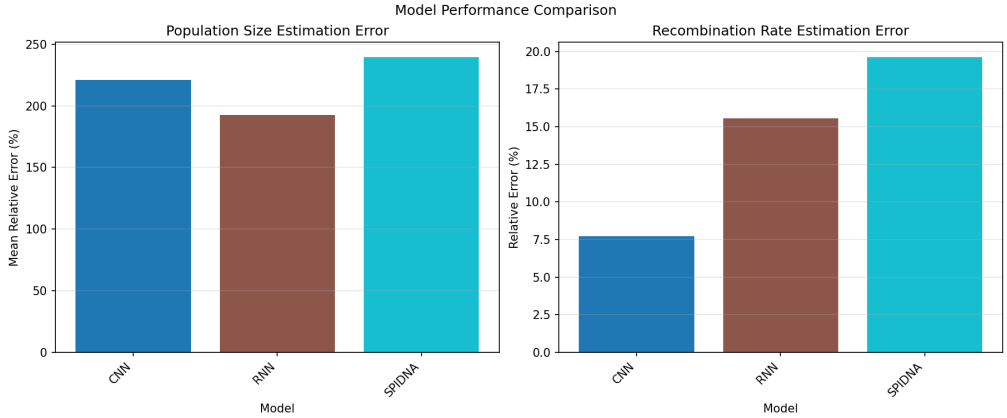
Figure S4: Comparison of credible interval width across the three variable population size scenarios for each estimated parameter. Normalized credible interval width was calculated as the 95% credible interval width divided by the geometric mean for population size parameters and by the posterior mean for the recombination rate parameter.



(a) Abrupt change

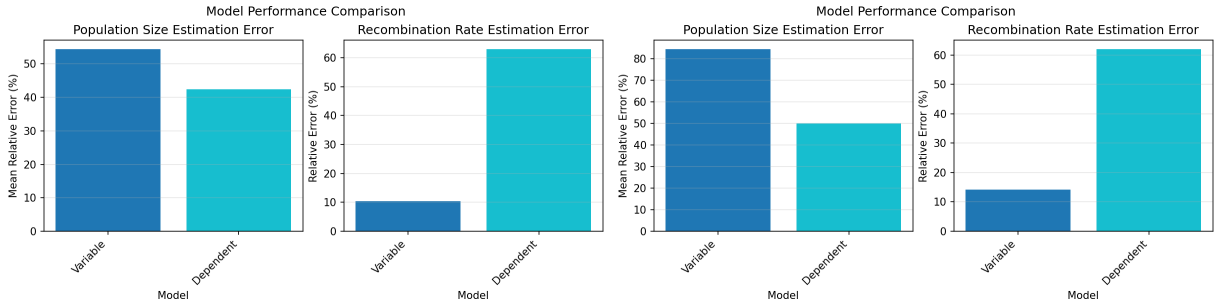


(b) Linear growth



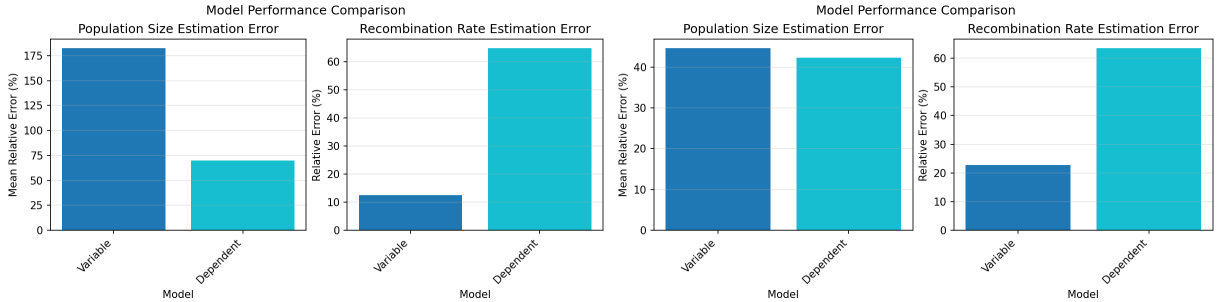
(c) Linear decline

Figure S5: Comparison of error in point estimation across the three variable population size scenarios. Left panel shows aggregated error across population size parameters, the right panel shows error for the recombination rate



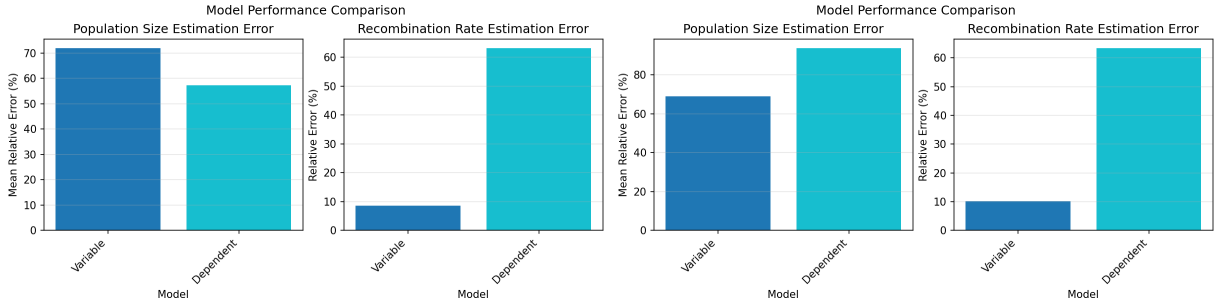
(a) Medium

(b) Large



(c) Decline

(d) Expansion



(e) Bottleneck

(f) Zigzag

Figure S6: Comparison of error in point estimation across six population size scenarios. Left panel shows aggregated error across population size parameters, the right panel shows error for the recombination rate

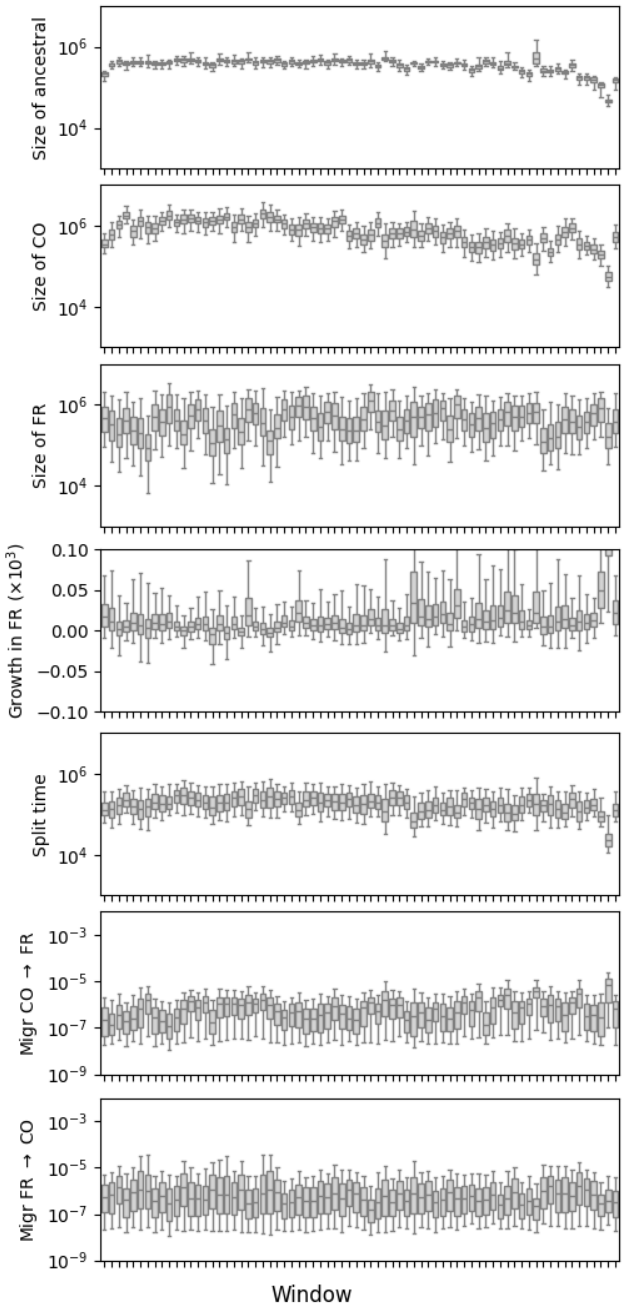


Figure S7: Marginal posterior distributions of parameters for a *Drosophila melanogaster* out-of-Africa demographic model, fit via a neural posterior estimator with an RNN embedding network in 300Kb windows across 2L. Migration rates are backwards-in-time.

			Quantiles			
	Mean	Median	0.025	0.25	0.75	0.975
Size of ancestral	380,867	389,288	109,605	305,789	451,407	632,108
Size of CO	888,294	740,976	125,023	416,628	1,218,365	2,420,950
Size of FR (ancient)	229,219	59,621	2,242	16,218	211,696	1,589,905
Size of FR (current)	619,429	395,504	27,203	172,051	863,027	2,283,723
Split time	216,028	176,029	40,075	107,993	285,626	593,438
Migr CO $\rightarrow$ FR	1.35e-06	6.04e-07	1.77e-08	1.17e-07	1.60e-06	6.81e-06
Migr FR $\rightarrow$ CO	2.56e-06	5.16e-07	1.67e-08	1.02e-07	2.11e-06	1.37e-05

Figure S8: Posterior summaries for parameters in the *Drosophila melanogaster* out-of-Africa demographic model, fit via a RNN embedding network.