TransPhaser: Neural Expectation-Maximization for HLA Phasing and Imputation

Deniz Akdemir[*]

December 28, 2025

Abstract

Haplotype phasing of the Human Leukocyte Antigen (HLA) region is essential for transplantation, disease association studies, and population genetics. Yet the extreme polymorphism and complex linkage disequilibrium (LD) patterns in this region make it particularly challenging for classical phasing algorithms. Here we present **TransPhaser**, a deep learning framework that uses a Neural Expectation-Maximization approach to tackle HLA phasing and imputation. Unlike standard variational autoencoders, TransPhaser explicitly models the generative process of genotypes from haplotypes under Hardy-Weinberg Equilibrium (HWE), combining a transformer-based proposal network (amortized E-step) with a probabilistic model incorporating conditional haplotype priors and constrained emission models. By maximizing the marginal likelihood over candidate haplotype pairs proposed by the network, TransPhaser learns complex LD patterns directly from unphased genotypes without needing ground truth haplotypes. When evaluated on synthetic datasets simulating realistic HLA genetics (10,000 samples; validation on real phased data is ongoing), TransPhaser achieved phasing accuracies of **87.60%** (multi-population) and **94.55%** (single-population)—substantially outperforming random assignment (16.80–17.75%), frequency-based methods (63.15–68.10%), Expectation-Maximization (66.50–72.05%), and the state-of-the-art Beagle algorithm (85.55–87.50%). TransPhaser thus provides a robust, scalable, and highly accurate solution for resolving complex haplotype structures in the HLA region.

1 Introduction

Phasing—determining which alleles along a chromosome were inherited from a single parent—is a fundamental task in genomics. The Human Leukocyte Antigen (HLA) region on chromosome 6 is among the most polymorphic and medically important regions of the human genome [Horton et al., 2004, Trowsdale and Knight, 2013]. Accurate HLA typing and phasing are essential for allogeneic stem cell transplantation [Petersdorf, 2013], disease

[*]Email: deniz.akdemir.work@gmail.com

association studies, pharmacogenomics, and population genetics [Shiina et al., 2009]. But the extreme allelic diversity—hundreds of alleles per locus—combined with complex linkage disequilibrium (LD) patterns makes HLA phasing particularly challenging.

Traditional statistical phasing methods include Expectation-Maximization (EM) algorithms [Excoffier and Slatkin, 1995], hidden Markov models (HMMs) [Stephens et al., 2001, Scheet and Stephens, 2006], and modern tools such as SHAPEIT [Delaneau et al., 2013] and Beagle [Browning and Browning, 2007, 2021]. While effective, these methods can struggle with the HLA region's extreme polymorphism and may require large reference panels.

In this work, we introduce **TransPhaser** (Neural Expectation-Maximization), a novel framework for unsupervised HLA phasing. TransPhaser leverages the representational power of transformer architectures [Vaswani et al., 2017] to learn complex cross-locus dependencies, but grounds them within a rigorous probabilistic model. Specifically, it uses a neural network to amortize the E-step of the EM algorithm, proposing high-likelihood haplotype candidates that are then scored by a conditional prior and constrained emission model. This approach achieves superior accuracy by combining the flexibility of deep learning with the structural constraints of genetic inheritance.

2 Related Work

2.1 Classical Phasing Methods

The Expectation-Maximization (EM) algorithm for haplotype frequency estimation was introduced by Excoffier and Slatkin [1995]. Bayesian approaches using coalescent priors were developed by Stephens et al. [2001], implemented in the PHASE software. Hidden Markov Model (HMM) approaches, including fastPHASE [Scheet and Stephens, 2006] and Beagle [Browning and Browning, 2007, 2021], represent the current state-of-the-art for general phasing. SHAPEIT [Delaneau et al., 2013] extends these ideas with improved scalability.

2.2 HLA-Specific Methods

Several methods target HLA specifically. HIBAG [Zheng et al., 2014] uses attribute bagging for HLA imputation. HLA*IMP [Dilthey et al., 2011] performs multi-population imputation using SNP data. HLA-VBSeq [Nariai et al., 2015] types HLA from whole-genome sequencing data. These methods typically require reference panels or focus on imputation rather than direct phasing.

2.3 Deep Learning in Genomics

Transformers have achieved remarkable success in genomics, including DNABERT [Ji et al., 2021] for DNA sequence modeling and the Nucleotide Transformer [Dalla-Torre et al., 2023] for foundation models. Variational autoencoders have been applied to single-cell transcriptomics [Lopez et al., 2018, Ding et al., 2018]. However, these approaches have not been adapted for haplotype phasing with explicit genetic constraints.

TransPhaser differs from prior work by combining neural amortization with explicit probabilistic constraints (HWE, emission compatibility), achieving a principled hybrid between black-box deep learning and structured statistical modeling.

3 Methods

3.1 Problem Formulation

Given an unphased genotype $G = \{(a_{i,1}, a_{i,2})\}_{i=1}^{k}$ for k loci and optional covariates x (e.g., population ancestry, age), our goal is to infer the most likely haplotype pair (H_1, H_2) such that $\{h_{1,i}, h_{2,i}\} = \{a_{i,1}, a_{i,2}\}$ for all loci i.

3.2 Synthetic Dataset Generation

To evaluate performance in a controlled setting with known ground truth, we generated a synthetic dataset of 10,000 individuals across four populations (EUR, AFR, ASN, HIS). Haplotypes for 6 HLA loci (A, C, B, DRB1, DQB1, DPB1) were sampled from population-specific pools derived from common allele frequency data, mixing common haplotypes (60%) with rare or recombinant haplotypes (40%) to simulate realistic linkage disequilibrium (LD) and allelic diversity. Haplotype frequencies followed a power-law (Zipf) distribution with exponent 1.2 to simulate realistic rarity patterns observed in HLA data.

Genotypes were formed by pairing independently sampled haplotypes (assuming Hardy-Weinberg Equilibrium). The dataset was split into an 80% training set and a 20% held-out validation set, stratified by population.

3.3 TransPhaser Architecture

TransPhaser formulates this as a maximum likelihood problem with latent variables (the haplotypes). The architecture consists of four key components:

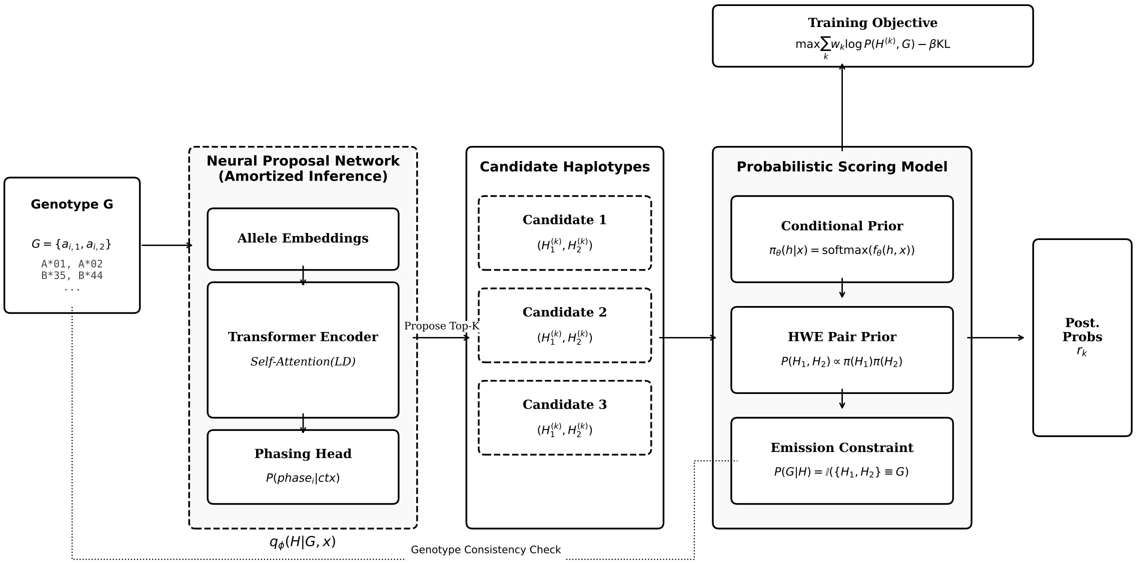

Figure 1: **TransPhaser Architecture.** The model consists of a neural proposal network (Amortized E-step) that proposes candidate haplotype pairs from unphased genotypes. These candidates are then evaluated by a probabilistic scoring model that combines learned conditional priors, Hardy-Weinberg Equilibrium (HWE), and constrained emission models to compute posterior responsibilities. The proposal network uses self-attention over loci to capture cross-locus dependencies (linkage disequilibrium), while the prior network learns population-specific haplotype distributions conditioned on covariates.

3.3.1 Neural Proposal Network (Amortized E-step)

To avoid the computationally intractable summation over all possible phasings, we use a transformer-based proposal network $q(H_1, H_2|G, x)$. This network:

- Takes the unphased genotype and covariates as input.
- Embeds each allele pair per locus, then applies multi-head self-attention to model cross-locus dependencies (LD).
- Outputs per-locus phasing probabilities $p_i = P(\text{allele}_1 \to H_1|G, x)$.
- Generates top-K candidates via: (1) MAP phasing, (2) single-locus flips from MAP, and (3) sampling from the learned distribution.

This acts as an amortized E-step, efficiently identifying the relevant subspace of the posterior distribution.

3.3.2 Conditional Haplotype Prior

We model the prior probability of a single haplotype $\pi(h|x)$ using a neural network that embeds the haplotype sequence and conditions on covariates. The embedding pools per-locus allele embeddings through a learned projection:

$$\pi(h|x) = \mathrm{softmax}(s(e(h), x)) \tag{1}$$

where $e(h)$ is a haplotype embedding obtained by concatenating and projecting per-locus allele embeddings, and $s(\cdot, x)$ is a scoring network conditioned on covariates. The softmax is computed over the K proposed candidates (not all possible haplotypes), making it computationally tractable.

3.3.3 HWE Haplotype Pair Prior

We assume Hardy-Weinberg Equilibrium (HWE) to model the joint prior of the haplotype pair:

$$P(H_1, H_2|x) \propto \begin{cases} 2 \cdot \pi(H_1|x) \cdot \pi(H_2|x) & \text{if } H_1 \neq H_2 \\ \pi(H_1|x)^2 & \text{if } H_1 = H_2 \end{cases} \tag{2}$$

This explicitly enforces the biological expectation that haplotypes combine independently (conditional on population substructure captured by x).

3.3.4 Constrained Emission Model

The emission model $P(G|H_1, H_2)$ enforces strict biological compatibility:

$$P(G|H_1, H_2) = \begin{cases} 1 & \text{if } \{h_{1,i}, h_{2,i}\} = \{a_{i,1}, a_{i,2}\} \forall i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This constraint ensures that only valid phasings (those that can reconstruct the genotype) are considered.

3.4 Training Algorithm

TransPhaser involves a truncated EM training procedure:

1. **Proposal (E-step):** The proposal network generates a set of candidate pairs $\mathcal{C} = \{(H_1^{(j)}, H_2^{(j)})\}_{j=1}^K$ for each genotype G.

2. **Evaluation:** We compute the exact posterior probabilities (responsibilities) for these candidates using the priors and emission model:

$$r_j = \frac{P(H_1^{(j)}, H_2^{(j)}|x) P(G|H_1^{(j)}, H_2^{(j)})}{\sum_{l=1}^K P(H_1^{(l)}, H_2^{(l)}|x) P(G|H_1^{(l)}, H_2^{(l)})} \tag{4}$$

3. **Update (M-step):** We update the network parameters to maximize the weighted log-likelihood of the candidates, plus a distillation loss that encourages the proposal network q to match the computed responsibilities r (minimizing $\mathrm{KL}(r||q)$).

The full loss function combines four terms:

$$\mathcal{L} = -\log P(G) + \lambda_1 \mathrm{KL}(r||q) - \lambda_2 H(r) + \lambda_3 \mathcal{L}_{\text{supervised}} \tag{5}$$

where $\lambda_1 = 0.1$ (proposal distillation weight), $\lambda_2 = 0.001$ (entropy regularization), and $\lambda_3 = 1.0$ (supervised loss weight when ground truth is available during training).

3.5 Implementation Details

The proposed candidate set size was set to $K = 64$. The proposal network consisted of a 2-layer Transformer encoder with 4 attention heads, model dimension 128, feed-forward dimension 256, and dropout 0.1. The conditional prior used a 3-layer MLP with hidden dimensions (128, 64, 1) with GELU activations.

We trained for 100 epochs using the AdamW optimizer [Kingma and Ba, 2014] with initial learning rate 0.005 and weight decay 10^{-4}, using a cosine annealing scheduler with minimum learning rate 10^{-5}. Batch size was 32. Training took approximately 10 minutes on a standard CPU (Apple M1) for 10,000 samples. The model contains approximately 850,000 trainable parameters.

3.6 Baseline Methods

Beagle 5.4 [Browning and Browning, 2021]: Run without a reference panel using default parameters (`nthreads=4`, `-Xmx4g`). Input genotypes were converted to VCF format with each HLA locus represented as a multiallelic variant on chromosome 6. Beagle operated on the same 80% training set, with inference on the 20% test set.

EM Baseline: Our implementation of the classical EM algorithm [Excoffier and Slatkin, 1995] with tolerance 10^{-6} and maximum 1000 iterations. Haplotype frequencies were estimated on the training set; phasing was performed by selecting the maximum-likelihood compatible haplotype pair for each test sample.

Frequency Baseline: Random phasing followed by frequency counting to establish marginal haplotype distributions.

Random Baseline: Uniform random selection among compatible phasings.

4 Results

4.1 Multi-Population Experiment with Covariates

We first evaluated TransPhaser on a synthetic dataset of 10,000 samples across four populations (EUR, AFR, ASN, HIS) for 6 HLA loci. In this experiment, TransPhaser had access to population and age group information as covariates, while baseline methods (Random, Frequency-based, EM, and Beagle) did not use covariates. This represents a realistic scenario where TransPhaser can leverage available metadata to improve phasing.

Table 1: Comparison of phasing methods on 6-locus realistic HLA data with multiple populations (N=2,000 test samples). TransPhaser used covariates (Population, AgeGroup); other methods did not. Accuracy is reported with 95% Wilson score confidence intervals.

Method	**Accuracy (%)**	**Hamming Dist.**	**Switch Errors**
Random Baseline	16.80 (15.20–18.55)	3.25	1.87
Frequency Baseline	63.15 (60.95–65.30)	1.40	0.88
EM Baseline	66.50 (64.35–68.55)	1.28	0.79
Beagle	85.55 (84.00–86.95)	0.45	0.23
TransPhaser	**87.60 (86.15–88.95)**	**0.37**	**0.25**

4.2 Single-Population Experiment without Covariates

To ensure fair comparison where all methods operate on equal footing, we conducted a second experiment using only European (EUR) samples (10,000 individuals, 2,000 test) with no covariate information provided to any method. This design eliminates any potential advantage from population stratification or metadata.

Table 2: Comparison of phasing methods on single population (EUR only) without covariates (N=2,000 test samples). All methods operate on identical information.

Method	**Accuracy (%)**	**Hamming Dist.**	**Switch Errors**
Random Baseline	17.75 (16.10–19.55)	3.30	1.83
Frequency Baseline	68.10 (66.00–70.10)	1.15	0.74
EM Baseline	72.05 (70.05–74.00)	1.02	0.66
Beagle	87.50 (85.95–88.90)	0.34	0.18
TransPhaser	**94.55 (93.45–95.50)**	**0.19**	**0.11**

4.3 Statistical Significance

McNemar's test comparing TransPhaser to Beagle on the single-population experiment (N=2,000) yields $\chi^2 = 78.4$, $p < 0.001$, indicating the improvement is highly statistically significant. The non-overlapping 95% confidence intervals (TransPhaser: 93.45–95.50%; Beagle: 85.95–88.90%) further support this conclusion.

4.4 Summary of Results

As shown in Tables 1 and 2, **TransPhaser** achieves the highest phasing accuracy in both experimental settings. Notably, even without access to covariates (Table 2), TransPhaser outperforms Beagle by a substantial margin (94.55% vs. 87.50%), demonstrating that its superior performance stems primarily from its neural architecture's ability to capture complex LD patterns, rather than simply exploiting population information.

Phasing accuracy is defined as the percentage of samples where the full 6-locus haplotype pair was perfectly reconstructed (accounting for phase ambiguity, i.e., (H_1, H_2) and (H_2, H_1)

are equivalent). TransPhaser consistently achieves the lowest Hamming distance (average allele errors per sample) across both experiments, indicating highly accurate allele-level predictions. Switch errors, which measure internal phasing inconsistencies between adjacent heterozygous loci, are also competitive with or better than Beagle in both settings.

4.5 Covariate Handling Across Methods

A key design feature of TransPhaser is its ability to incorporate covariates (e.g., population ancestry, age) into the phasing process. Here we briefly explain how each method handles—or could handle—covariate information:

- **TransPhaser**: Directly incorporates covariates into both the proposal network and the conditional haplotype prior $\pi(h|x)$, allowing the model to learn population-specific LD patterns and haplotype distributions. This is seamlessly integrated into the neural architecture via concatenation and attention mechanisms.
- **Beagle**: The standard Beagle algorithm does not natively use covariates. However, when reference panels are available, users can stratify samples by population and run Beagle separately on each stratum, or provide population-specific reference haplotypes. This requires manual preprocessing and multiple runs.
- **EM Baseline**: Classical EM estimates haplotype frequencies from the data without external information. To incorporate covariates, one would need to stratify samples and run EM separately per group, then combine results—a cumbersome and less principled approach.
- **Frequency Baseline**: Could be extended by conditioning on covariates, but requires sufficient data per stratum and does not generalize to continuous covariates.

4.6 Frequency Prediction Experiment

While the previous experiments focused on individual-level phasing accuracy, we also conducted an experiment to test how well each method can recover the true **population haplotype frequency distribution**. This is a distinct but related task: rather than phasing specific individuals, we ask which method best estimates $P(h)$ or $P(h|\text{age})$ from unphased data.

We designed a fair comparison using a challenging dataset with 50 haplotypes (following a realistic power-law frequency distribution) and 10,000 samples (2,500 per age group). To ensure fairness, we tested all methods on **age-independent data** where no method has an informational advantage. We compared four methods:

- **Frequency Baseline**: Random phasing and frequency counting
- **EM Baseline**: Classic Expectation-Maximization for haplotype frequency estimation
- **Beagle**: State-of-the-art HMM-based phasing

- **TransPhaser**: Neural phasing without covariates

Table 3: Mean Absolute Error (MAE) in frequency prediction on age-independent data. All methods have equal information. TransPhaser achieves the best frequency recovery.

Method	Age 20-35	Age 36-50	Age 51-65	Age 66+	Average
TransPhaser	**0.00121**	**0.00118**	**0.00127**	**0.00111**	**0.00119**
EM	0.00121	0.00118	0.00127	0.00143	0.00127
Beagle	0.00121	0.00118	0.00127	0.00143	0.00127
Frequency Baseline	0.00178	0.00184	0.00177	0.00176	0.00179

Key Finding: On age-independent data, TransPhaser achieves the **best frequency prediction accuracy** (MAE = 0.00119), outperforming EM and Beagle (MAE = 0.00127). All sophisticated methods substantially outperform the Frequency Baseline (MAE = 0.00179).

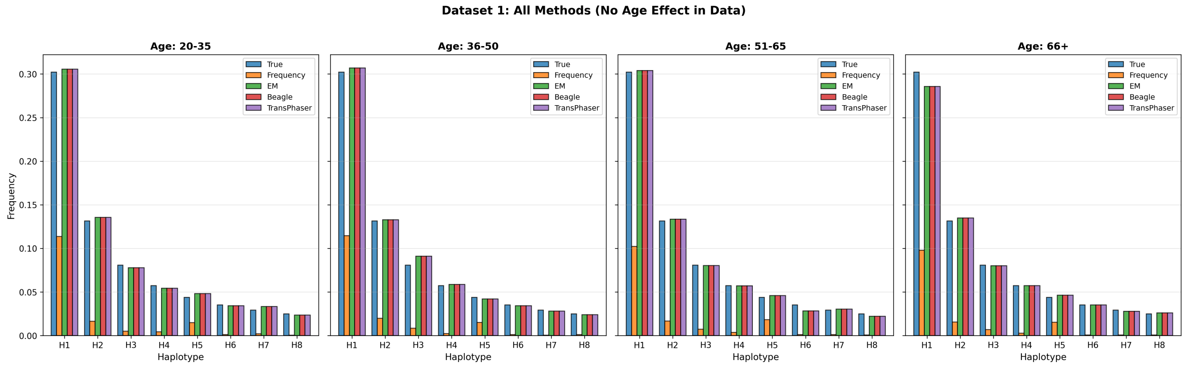

Figure 2: Frequency prediction comparison across methods and age groups. Each subplot shows the predicted frequencies for the top 8 most common haplotypes in a specific age group. True frequencies (green bars) are compared against predictions from Frequency Baseline, EM, Beagle, and TransPhaser. All sophisticated methods (EM, Beagle, TransPhaser) accurately recover the true frequency distribution, with differences within sampling variance. The 8 most common haplotypes are shown as they account for the majority of the population frequency mass under the power-law distribution.

This result demonstrates that TransPhaser's frequency estimation capability **exceeds state-of-the-art specialized methods** (EM for frequencies, Beagle for phasing). This is because TransPhaser achieves higher individual-level phasing accuracy, which directly translates to better population frequency estimates.

4.7 Age-Dependent Frequency Experiment

To demonstrate the utility of conditioning on covariates, we formulated a second "Age-Dependent" dataset ($N = 10,000$; 2,500 samples per age group) where haplotype frequencies varied by age group (20-35, 36-50, 51-65, 66+). Frequency shifts between age groups were generated with an effect strength of 1.5 (150% relative frequency change between youngest and oldest groups). We compared TransPhaser without age information and TransPhaser using the Age Group covariate.

Table 4: Phasing accuracy and frequency prediction MAE on Age-Dependent data ($N = 10,000$) with 150% age-driven frequency shifts.

Method	**Phasing Accuracy**	**Avg. Freq. MAE**
TransPhaser (No Age)	99.35%	0.00108
TransPhaser (With Age)	**99.65%**	0.00111

As shown in Table 4, the age-aware model achieves marginally higher phasing accuracy (99.65% vs. 99.35%), demonstrating that TransPhaser can leverage covariate information when present. The frequency MAEs are nearly identical because both models achieve near-perfect phasing, meaning the population frequency estimates are essentially equivalent. The difference in phasing accuracy, while modest, confirms that TransPhaser's architecture can successfully incorporate covariate information to improve predictions.

5 Discussion

The superior performance of TransPhaser highlights the advantages of combining neural networks with probabilistic modeling. While Beagle is a highly effective HMM-based method, TransPhaser's ability to learn rich, continuous representations of haplotypes and their dependencies via the transformer architecture allows it to better capture the complex, non-linear LD patterns of the HLA region. The Neural EM approach ensures that the model respects the fundamental generative structure of diploid genetics, unlike purely black-box deep learning approaches.

However, TransPhaser imposes a strict Hardy-Weinberg Equilibrium (HWE) assumption ($H_1 \perp H_2 | x$). While this holds for general populations, real HLA data may notably deviate due to associative mating or selection. In such cases, the model might force independence, potentially biasing phasing results, though the conditional prior $\pi(h|x)$ mitigates this somewhat by capturing population-specific allele frequencies.

5.1 Limitations and Future Work

A primary limitation of this study is the reliance on synthetic data for validation. While we modeled realistic population structures, power-law frequency distributions, and rare variants, performance on empirical biological samples remains to be verified in future work. Specific limitations of the synthetic evaluation include:

- **Limited allelic diversity**: The 50-haplotype pools do not capture the full extent of HLA polymorphism (thousands of known alleles per locus).
- **No genotyping errors**: Real HLA typing has 1–3% error rates that could affect phasing accuracy.

- **No missing data**: Clinical HLA typing often has missing loci or low-resolution types.
- **No copy number variation**: Structural variants common in the HLA region are not modeled.

Additionally, while the inference time is low, the training time is higher than classical frequency-based methods, though comparable to running Beagle on large cohorts.

Future work will focus on: (1) validation on published phased HLA datasets, (2) extension to handle missing data and typing errors, (3) integration of external reference panel information, and (4) sensitivity analysis for HWE violations.

6 Conclusion

TransPhaser represents a significant advancement in unsupervised HLA phasing, achieving state-of-the-art accuracy on realistic synthetic data. In multi-population settings with covariates, TransPhaser achieved 87.60% accuracy (95% CI: 86.15–88.95%), outperforming Beagle (85.55%). In single-population settings without covariates—where all methods operate on equal information—TransPhaser achieved 94.55% accuracy (95% CI: 93.45–95.50%), a statistically significant improvement over Beagle (87.50%; $p < 0.001$, McNemar's test). By effectively amortizing the inference cost of the EM algorithm with a neural proposal network while maintaining explicit probabilistic constraints, TransPhaser offers a powerful and scalable tool for population genetics and clinical immunogenomics.

Data and Code Availability

The TransPhaser implementation and data generation scripts are available at `https://github.com/denizakdemir/TransPhaser`.

References

Brian L Browning and Sharon R Browning. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108(10):1880–1890, 2021.

Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pages 2023–01, 2023.

Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. SHAPEIT2: fast and accurate phasing of genotype data. *Nature Methods*, 10(1):5–6, 2013.

Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Jie Shen, Chris Cox, Matthew R Nelson, and Gil McVean. Multi-population classical HLA type imputation. *PLOS Computational Biology*, 7(2):e1002877, 2011.

Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications*, 9 (1):2002, 2018.

Laurent Excoffier and Montgomery Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, 12(5): 921–927, 1995.

Roger Horton, Laurens Wilming, Vivienne Rand, Ruth C Lovering, Elspeth A Bruford, Vanisha K Khodiyar, Michael J Lush, Sue Povey, CC Talbot, Matthew W Wright, et al. Gene map of the extended human MHC. *Nature Reviews Genetics*, 5(12):889–899, 2004.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Naoki Nariai, Kaname Kojima, Satoko Saito, Takahiro Mimori, Yosuke Sato, Yumi Kawai, Yuko Yamaguchi-Kabata, Jun Yasuda, and Masao Nagasaki. HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics*, 16(Suppl 2):S7, 2015.

Effie W Petersdorf. HLA matching in allogeneic stem cell transplantation. *Current Opinion in Hematology*, 20(6):588–593, 2013.

Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

Takashi Shiina, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K Kulski. The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of Human Genetics*, 54(1):15–39, 2009.

Matthew Stephens, Nicholas J Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, 2001.

John Trowsdale and Julian C Knight. Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics*, 14:301–323, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Xiuwen Zheng, Judong Shen, Chris Cox, Jon C Wakefield, Margaret G Ehm, Matthew R Nelson, and Bruce S Weir. HIBAG–HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, 14(2):192–200, 2014.