Deep generative model of genetic variation data improves imputation accuracy in private populations

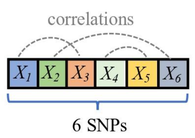
Prateek Anand[1], Anji Liu[2], Meihua Dang[3], Boyang Fu[4], Xinzhu Wei[5], Guy Van den Broeck[1,*], Sriram Sankararaman[1,6,7,*]

*Equal Contribution [1]Department of Computer Science, UCLA [2]Department of Computer Science, National University of Singapore [3]Department of Computer Science, Stanford University
[4]Department of Biomedical Informatics, Harvard Medical School [5]Department of Computational Biology, Cornell University [6]Department of Human Genetics, UCLA [7]Department of Computational Medicine, UCLA

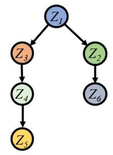
Introduction

- Artificial Genomes (AGs) enable sharing and benchmarking without exposing sensitive genetic data, supporting reproducibility and equity in population genomics
- Coalescent and PAC/HMM approaches are interpretable and tractable but often computationally intensive, and may not fully capture fine-scale genomic dependencies
- GANs, VAEs, and RBMs offer expressiveness but lack tractable inference, reliable likelihood computation, and are challenging to train and tune
- We propose a novel **tractable and expressive deep generative model** of genetic variation based on **Hidden Chow–Liu Trees (HCLTs)** represented as **Probabilistic Circuits (PCs)**, enabling efficient learning and inference

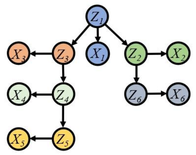
Methods



(a) Genetic data with 6 SNPs



(b) Learned CLT structure captures strong pairwise correlation between SNPs.



(c) Learned HCLT structure given the SNPs in (a).

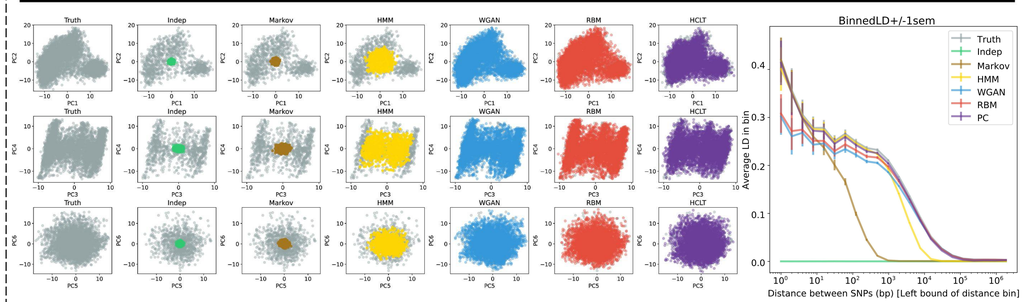
- After constructing the HCLT, we convert to a PC via **PyJuice [1]**
- PCs are <u>efficient</u> to train and contain a <u>single hyperparameter</u>
- PCs can generate samples and quickly query marginals and conditionals → <u>probabilistic convergence and direct imputation</u>

Data

- Phased haplotype data (0/1)
- 1000 Genomes Project Phase 3 (low-coverage)
  - 15:27379578 - 15:29625035 (5008 x 10000)
- UK Biobank
  - 22:29456546 - 22:32665772 (26924 x 9820)
- 1000 Genomes Project Phase 3 (high-coverage)
  - 15:27134431 - 15:29332831 (5008 x 14670)

Reconstructing Global and Local Population Structure

Dataset	Category	INDEP	MARKOV	HMM	PC
1KG	**train LL**	-2386.81	-1806.33	-591.08	**-202.51**
	test LL	-2404.51	-1819.96	-599.88	**-265.06**
	#params	10.00k	39.99k	163774.98k	88473.73k
UKBB	**train LL**	-1642.62	-1360.86	-554.88	**-120.10**
	test LL	-1648.03	-1362.16	-554.38	**-127.75**
	#params	9.82k	39.27k	160825.86k	88850.56k



- <u>Top:</u> Log-likelihood of models that support tractable likelihood inference
- <u>Bottom:</u> Left shows top 6 principal components of ground truth vs. samples, while right shows binned LD as a function of SNP distance

Preserving Privacy of the Training Data

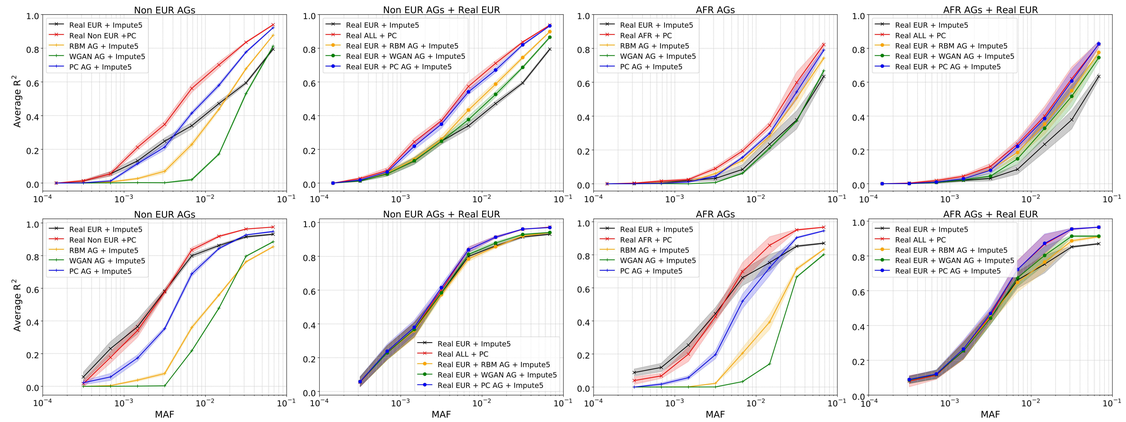
$$\mathcal{AATS} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(d_{TS}(i) > d_{TT}(i)\right) + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(d_{ST}(i) > d_{SS}(i)\right)\right)$$

Dataset	Metric	RBM	WGAN	PC
1KG	AA_{TRUTH} (Train)	0.9561	0.8103	**0.7185**
	AA_{TRUTH} (Test)	0.9928	0.7764	**0.7680**
	AA_{SYN} (Train)	0.0024	0.7356	**0.4225**
	AA_{SYN} (Test)	0.0276	0.7847	**0.5304**
UKBB	AA_{TRUTH} (Train)	0.9954	0.9674	**0.9204**
	AA_{TRUTH} (Test)	0.9962	0.9688	**0.9198**
	AA_{SYN} (Train)	0.0064	0.7768	**0.5324**
	AA_{SYN} (Test)	0.0160	0.7582	**0.4630**

- **Nearest Neighbor Adversarial Accuracy [2]**
- T = Truth, S = Synthetic
- **AA_{TRUTH}:** <u>Low</u> → synthetic mimics real too closely(overfit, low privacy). <u>High</u> → synthetic is far from real (high privacy, lower utility).
- **AA_{SYN}:** <u>Low</u> → synthetic copies real samples (low privacy). <u>High</u> → synthetic has its own structure (high privacy, may reduce utility if too different).
- **Ideal:** Both around **0.5** → realistic synthetic data that balances **privacy and utility**
- Only report components since average can hide poor performance

Improving Imputation Accuracy in Private Populations

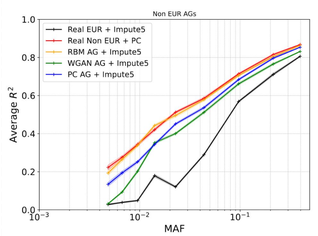
- Imputation of one SNP at a time using AGs as reference panel in Impute5 pipeline (or combined with real EUR data)
- Black line represents Impute5 only using real EUR data as reference



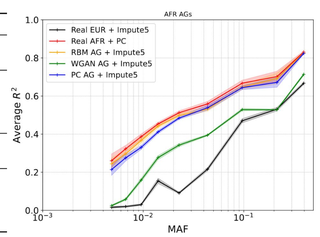
1000 Genomes low-coverage

UK Biobank

- Imputation of all SNPs not present on InfiniumOmni2-5-8v1-4 A1 array using 1000 Genomes high-coverage as training
- Out of 14670 SNPs, 1302 are present on the array, and remaining are imputed simultaneously (91% missingness)



Method	Mean R^2 [95% CI]
ALL SNPs	
Real + Impute5	0.3108 [0.3078, 0.3138]
Real + PC	**0.5294 [0.5273, 0.5315]**
RBM AG + Impute5	0.5215 [0.5188, 0.5242]
WGAN AG + Impute5	0.4289 [0.4257, 0.4321]
PC AG + Impute5	0.4725 [0.4696, 0.4754]
Low-freq (MAF $< 1\%$)	
Real + Impute5	0.0341 [0.0312, 0.0370]
Real + PC	**0.2776 [0.2727, 0.2826]**
RBM AG + Impute5	0.2598 [0.2528, 0.2667]
WGAN AG + Impute5	0.1039 [0.0998, 0.1079]
PC AG + Impute5	0.1902 [0.1855, 0.1948]



Method	Mean R^2 [95% CI]
ALL SNPs	
Real + Impute5	0.2447 [0.2392, 0.2501]
Real + PC	**0.5224 [0.5131, 0.5317]**
RBM AG + Impute5	0.5055 [0.4950, 0.5159]
WGAN AG + Impute5	0.3370 [0.3326, 0.3414]
PC AG + Impute5	0.4889 [0.4809, 0.4969]
Low-freq (MAF $< 1\%$)	
Real + Impute5	0.0186 [0.0138, 0.0235]
Real + PC	**0.3239 [0.3051, 0.3427]**
RBM AG + Impute5	0.3007 [0.2816, 0.3198]
WGAN AG + Impute5	0.0759 [0.0712, 0.0805]
PC AG + Impute5	0.2735 [0.2577, 0.2894]

Conclusion

- We present a novel **deep generative model** for genetics based on **PCs**
- **Fast & tractable:** Easy to train and use at scale
- **Realistic:** Matches key genetic patterns, with more accurate LD
- **Privacy-preserving:** Balances similarity and separation effectively
- **Boosts imputation:** Improves accuracy for underrepresented groups
- **Direct inference:** Enables imputation and likelihood evaluation
- **Enables access:** Supports open, equitable genomic research

References

[1] Anji Liu, Kareem Ahmed, and Guy Van den Broeck. Scaling tractable probabilistic circuits: A systems perspective. In Proceedings of the 41th International Conference on Machine Learning (ICML), jul 2024.

[2] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. Neurocomputing, 416:244–255, 2020.