GPC

Gen² 2026 | ICLR

An expressive and tractable deep generative model for genetic variation data

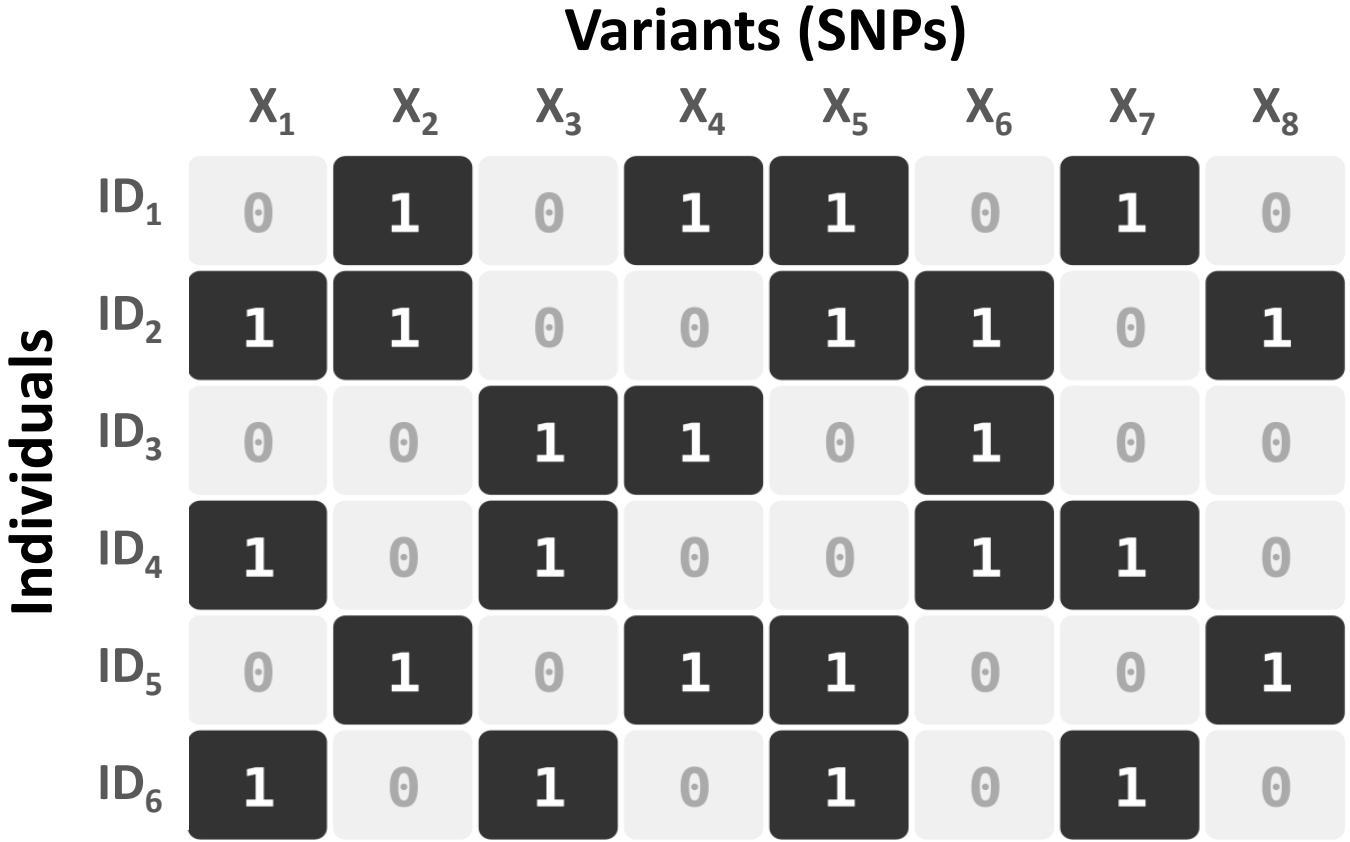
Prateek Anand

Computer Science PhD Student, UCLA

ICLR 2026 Workshop on Generative AI in Genomics

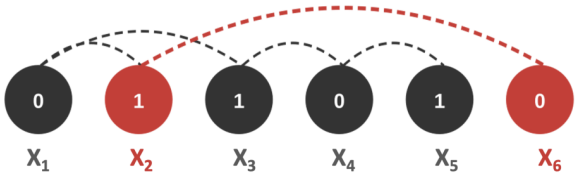
April 27, 2026

What is genetic variation data?



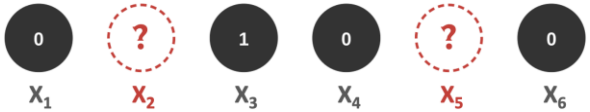
Three key challenges

Generating synthetic data that captures complex correlation



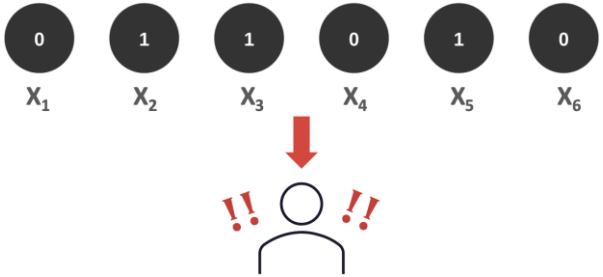
SNPs are correlated across different length scales

Imputing missing data



Many SNPs are unobserved; need to infer missing genotypes

Generating synthetic data that preserves privacy

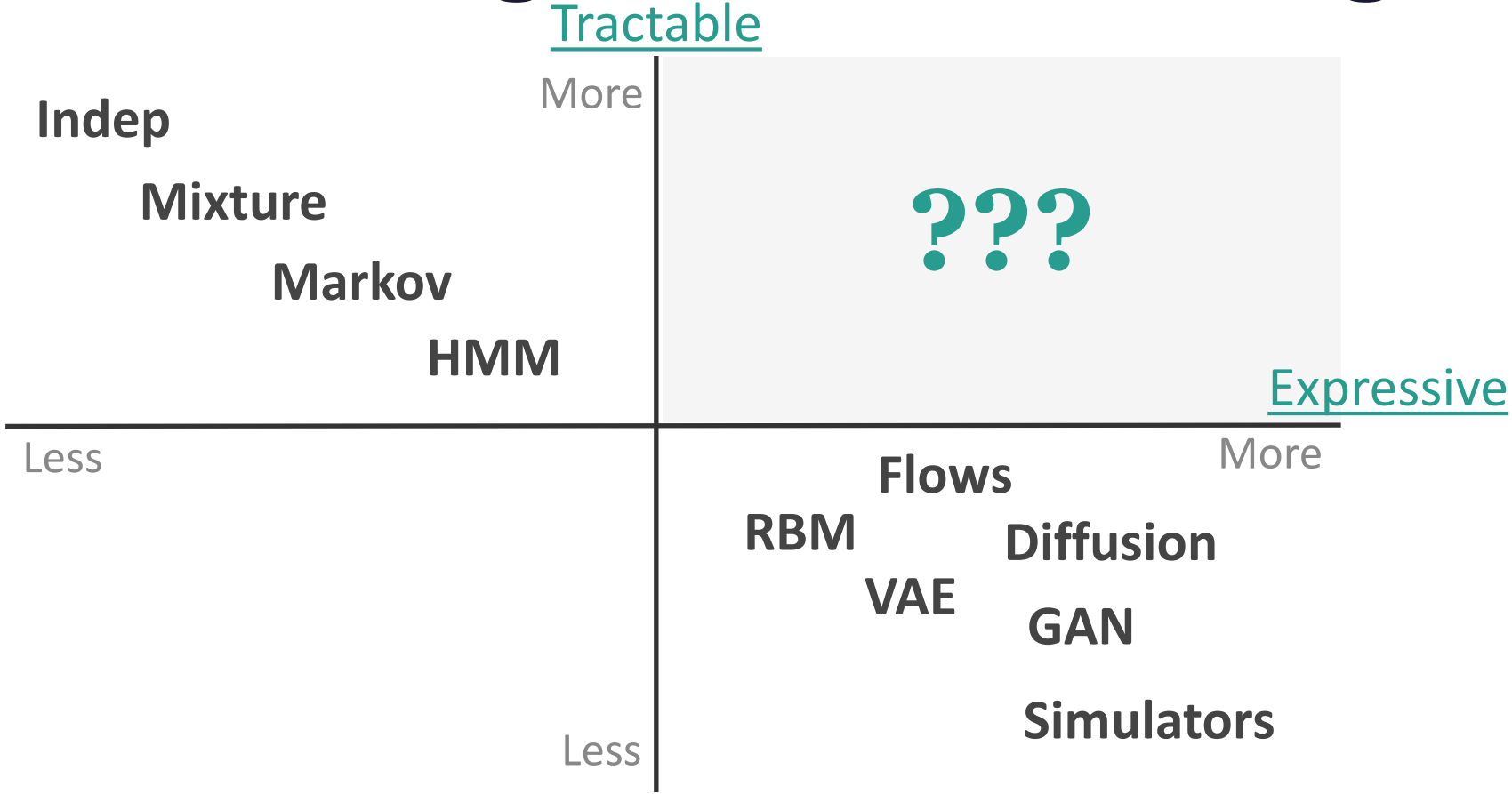


Real genomes can be linked back to individuals

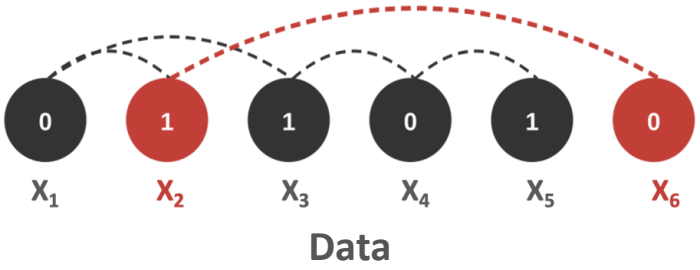
A good model should capture correlation structure, enable imputation, and preserve privacy

Current models struggle to address these while also being **tractable**

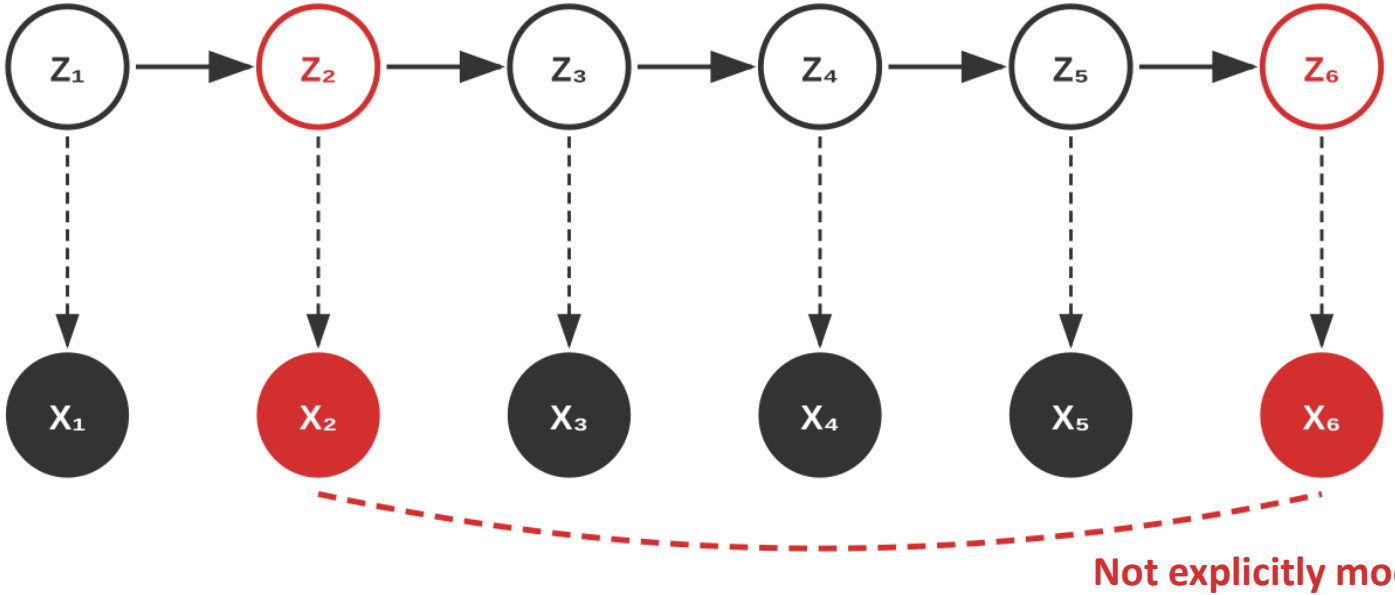
Tradeoffs in generative modeling



Hidden markov models

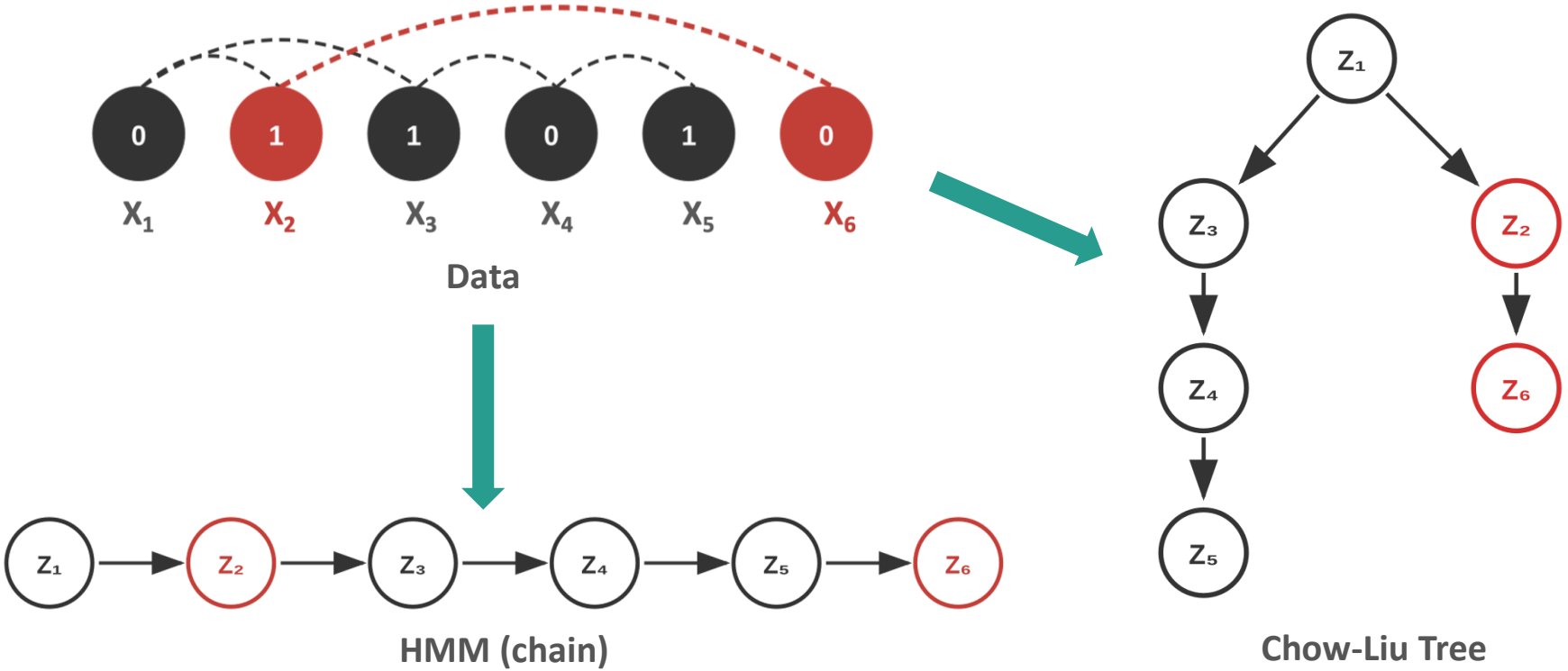


HMM



Chain structure forces information between distant SNPs through all intermediates

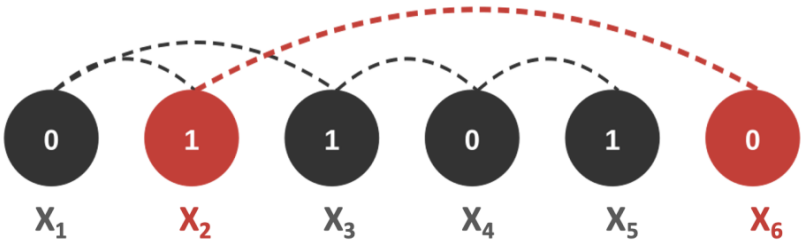
From chains to trees



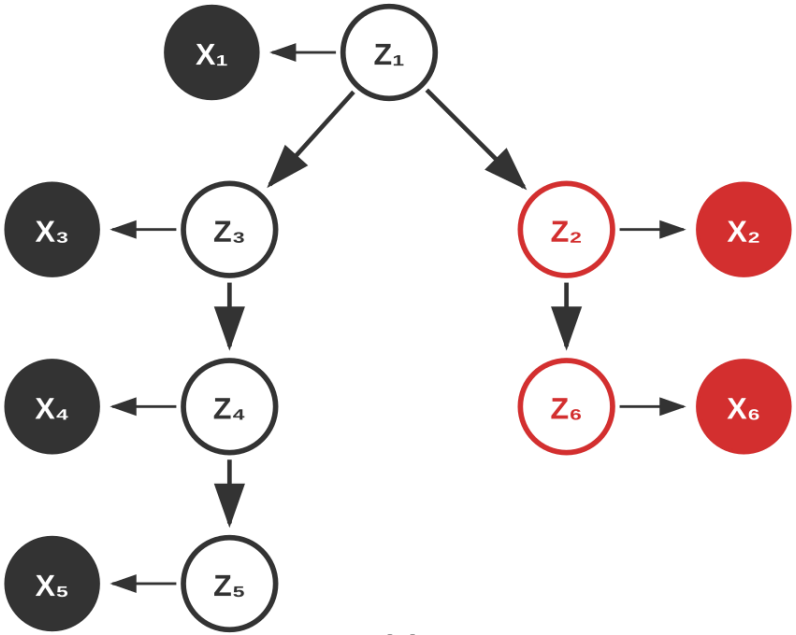
Tree structure places correlated SNPs close together regardless of genomic position

Chow and Liu IEEE 1968

Genetic Probabilistic Circuits (GPC)



Data



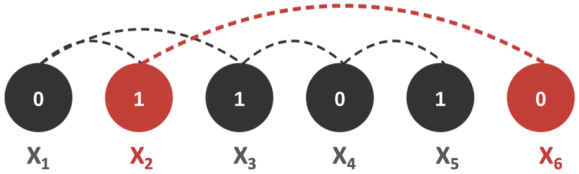
Hidden Chow-Liu Tree

Compile HCLT → **probabilistic circuit** → exact likelihoods, efficient training and inference

Vergari et al. AAAI 2020; Liu et al. ICML 2024; for detailed PC specification

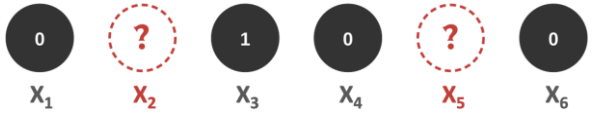
Three key challenges

Generating synthetic data that captures complex correlation



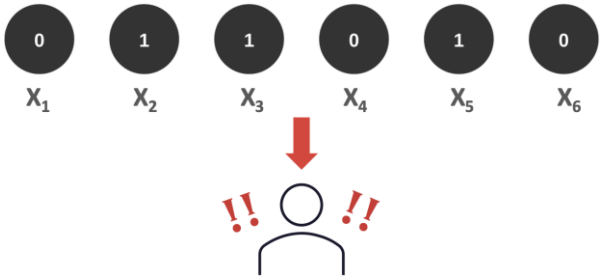
SNPs are correlated across different length scales

Imputing missing data



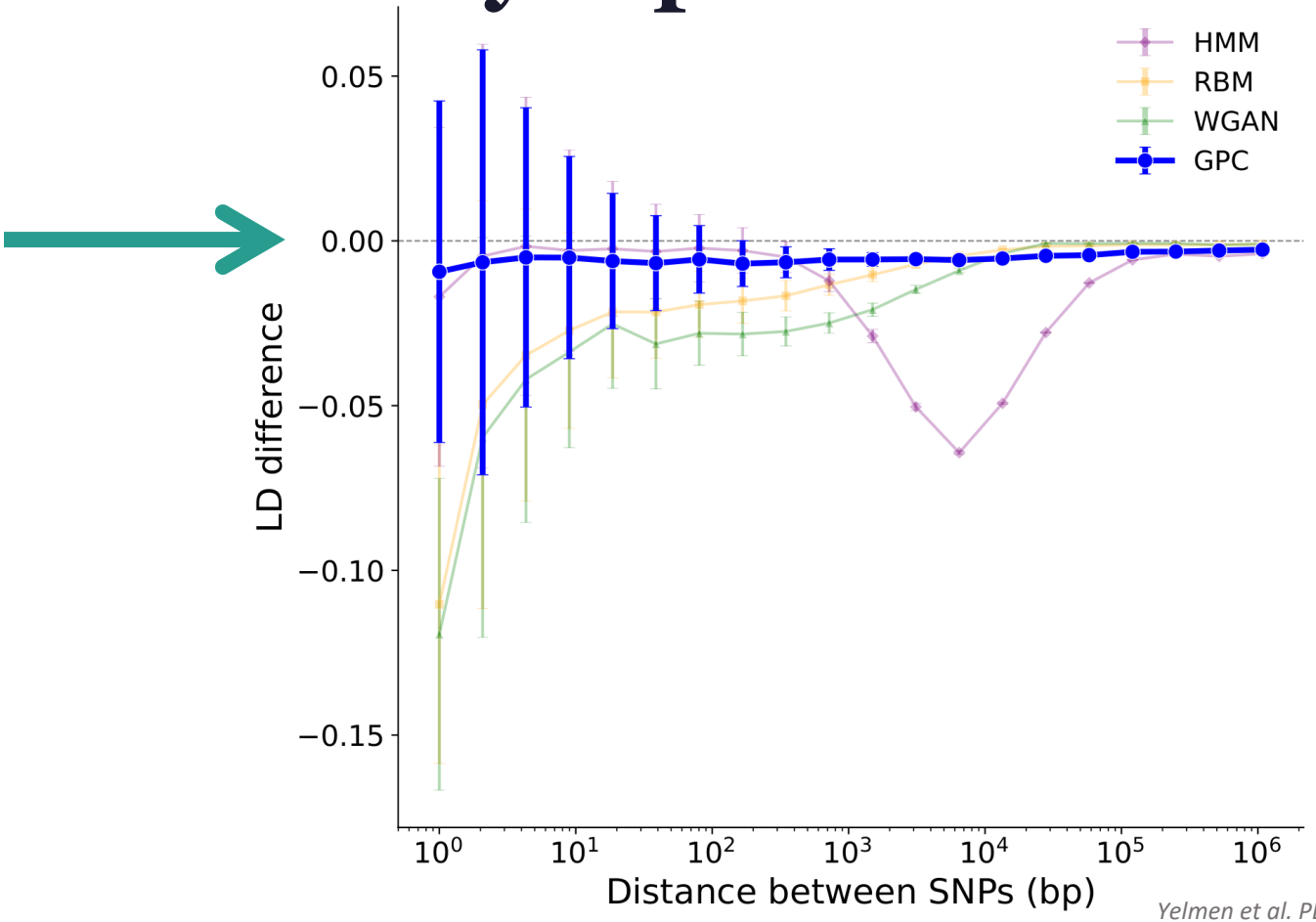
Many SNPs are unobserved; need to infer missing genotypes

Generating synthetic data that preserves privacy



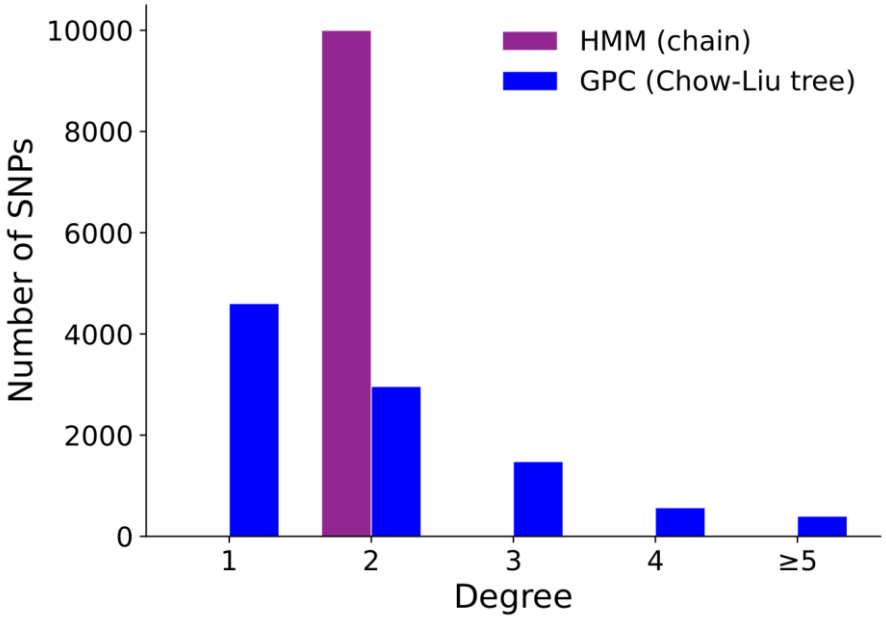
Real genomes can be linked back to individuals

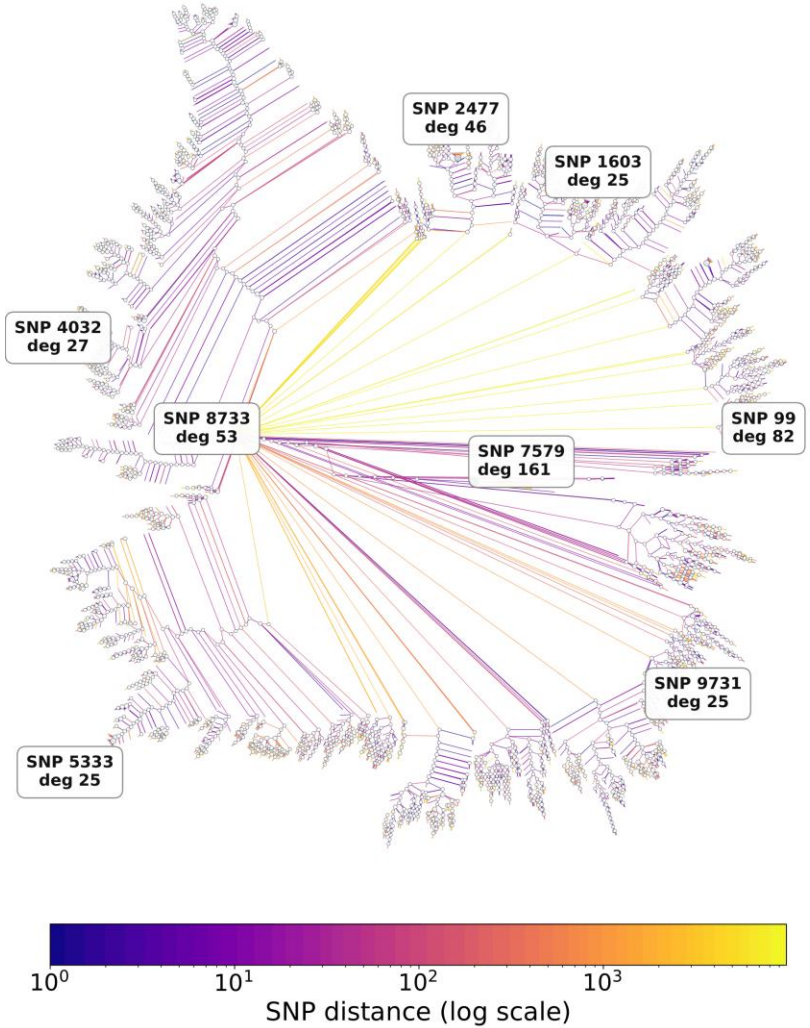
GPC closely reproduces correlation



Yelmen et al. PLoS Comp Bio 2023 (plotting scripts adapted)

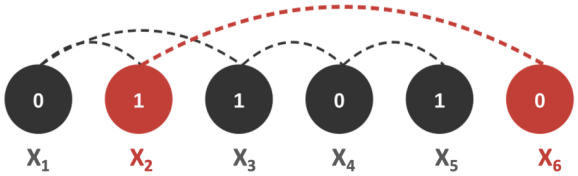
GPC tree structure reflects complex correlation





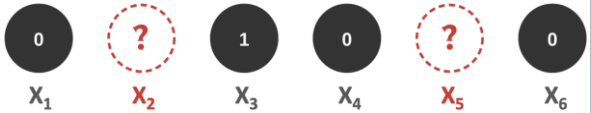
Three key challenges

Generating synthetic data that captures complex correlation



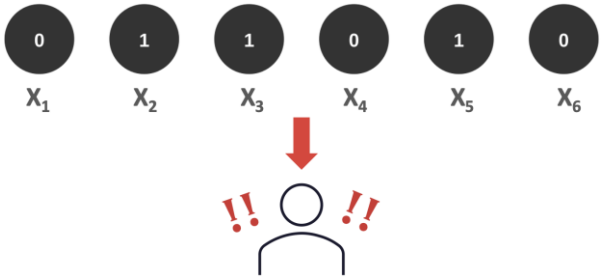
SNPs are correlated across different length scales

Imputing missing data



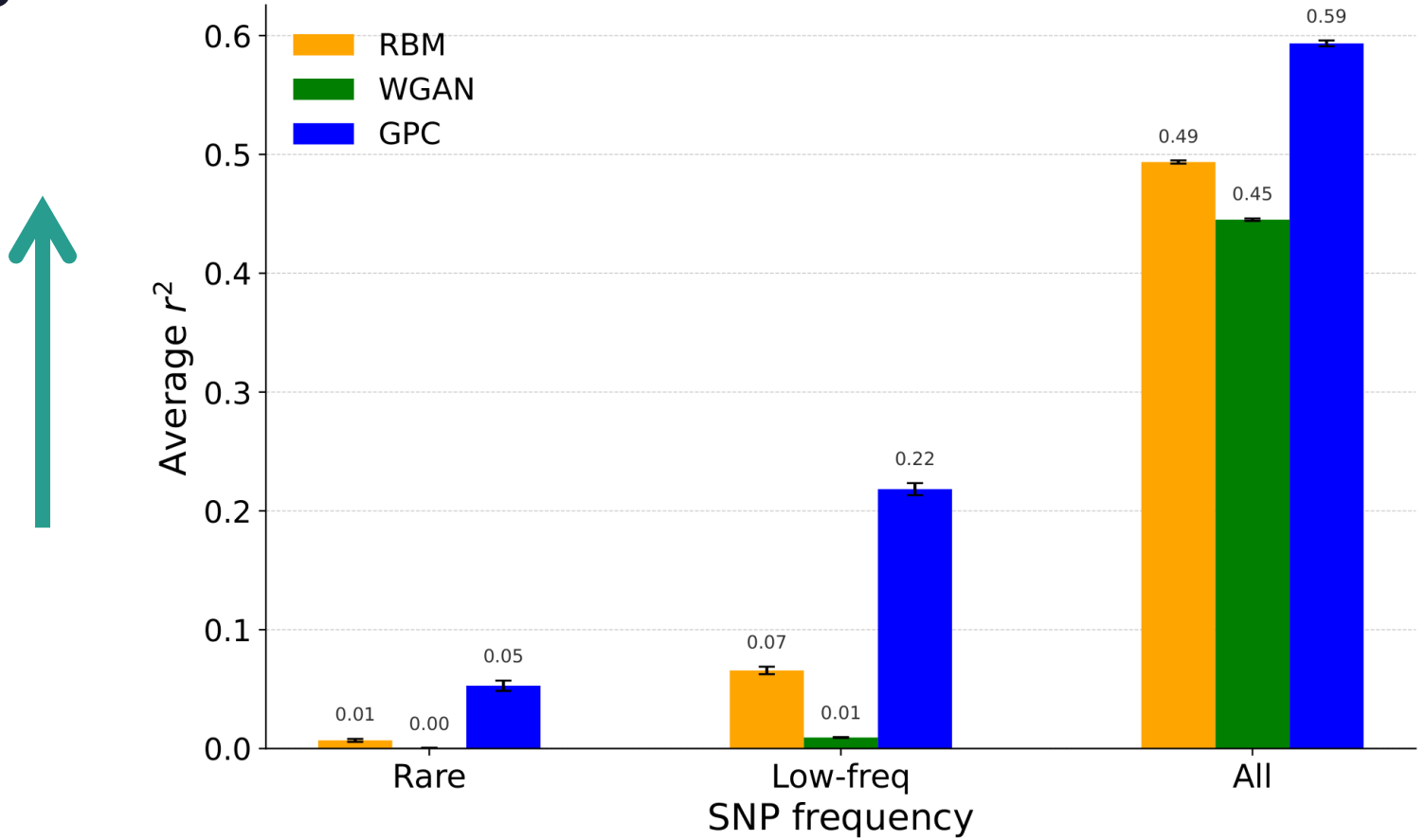
Many SNPs are unobserved; need to infer missing genotypes

Generating synthetic data that preserves privacy



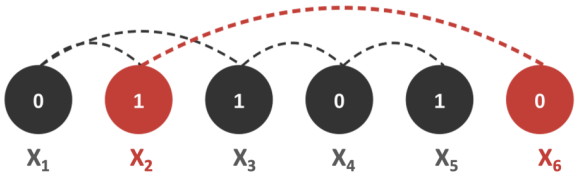
Real genomes can be linked back to individuals

GPC achieves best imputation among generative baselines



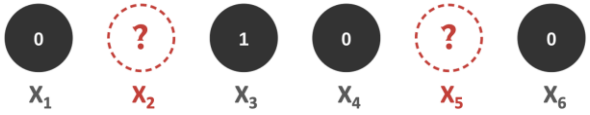
Three key challenges

Generating synthetic data that captures complex correlation



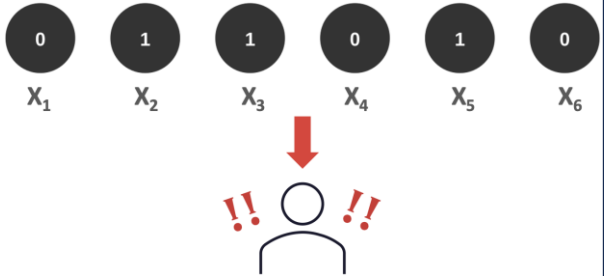
SNPs are correlated across different length scales

Imputing missing data



Many SNPs are unobserved; need to infer missing genotypes

Generating synthetic data that preserves privacy

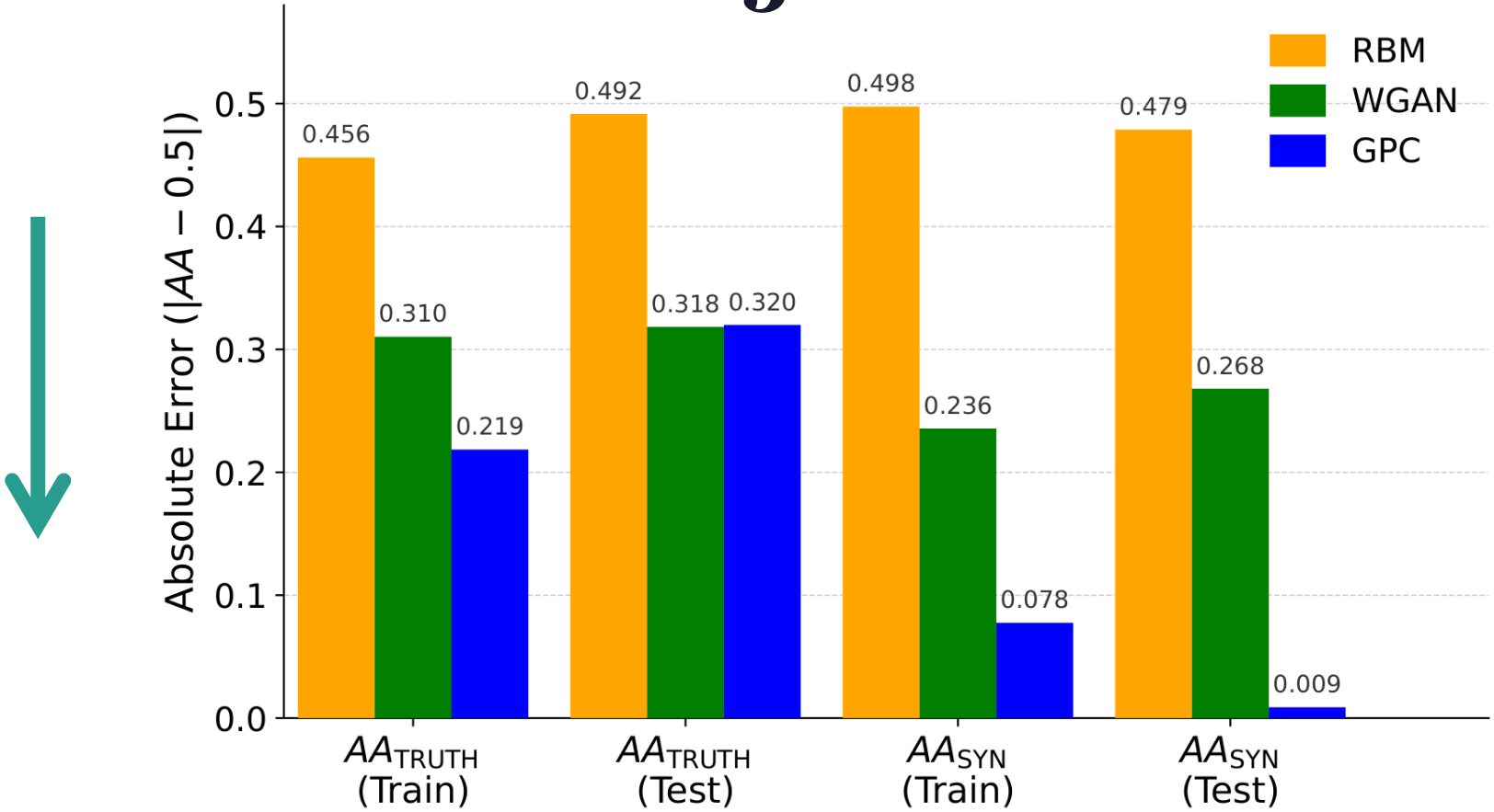


Real genomes can be linked back to individuals

Adversarial Accuracy (AA)
Based on nearest neighbors

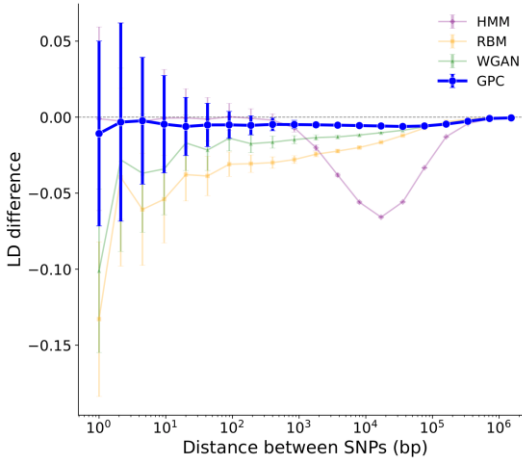
Yale et al. 2019

GPC balances privacy and utility: closest to ideal values of 0.5



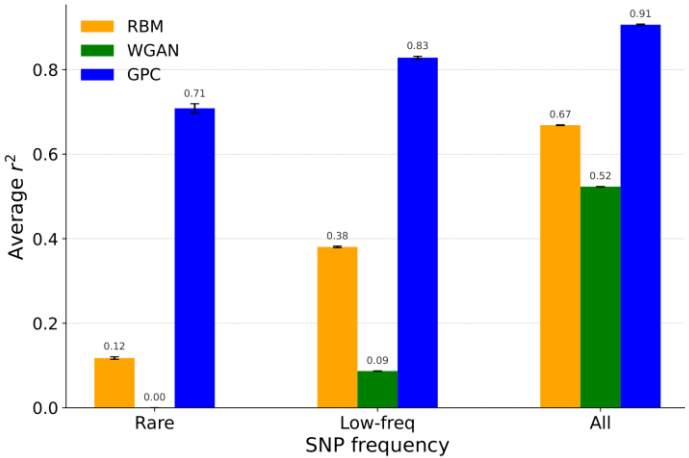
All findings replicate in UK Biobank

Generating synthetic data that captures complex correlation

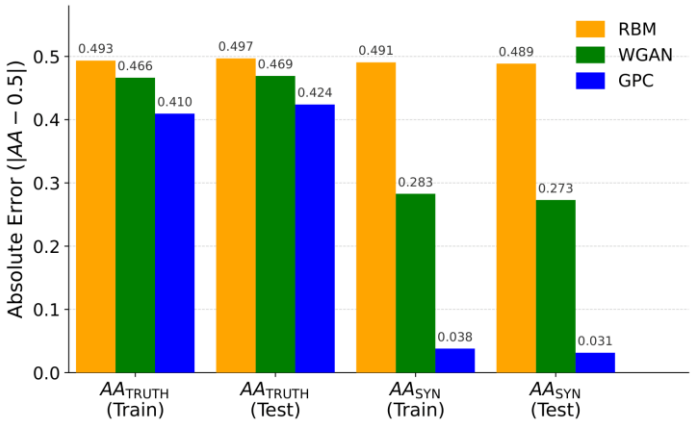


Close to 0

Imputing missing data



Generating synthetic data that preserves privacy



GPC takeaways

1. Expressive, tractable, and efficient: tree structure captures LD at all scales with exact inference

2. Best imputation accuracy among generative models

3. Better privacy-utility tradeoff for synthetic data than existing methods

Acknowledgments

Anji Liu

NUS

Meihua Dang

Stanford

Boyang Fu

Harvard

Xinzhu Wei

Cornell

Guy Van den Broeck*

UCLA

Sriram Sankararaman*

UCLA

Sriram Lab

UCLA

* Joint supervision

prateek@cs.ucla.edu | https://prateekanand2.github.io/ | github.com/sriramlab/GPC

Thank you!

Preprint

GitHub

Come find me at **Poster Session 1 Board 26!**