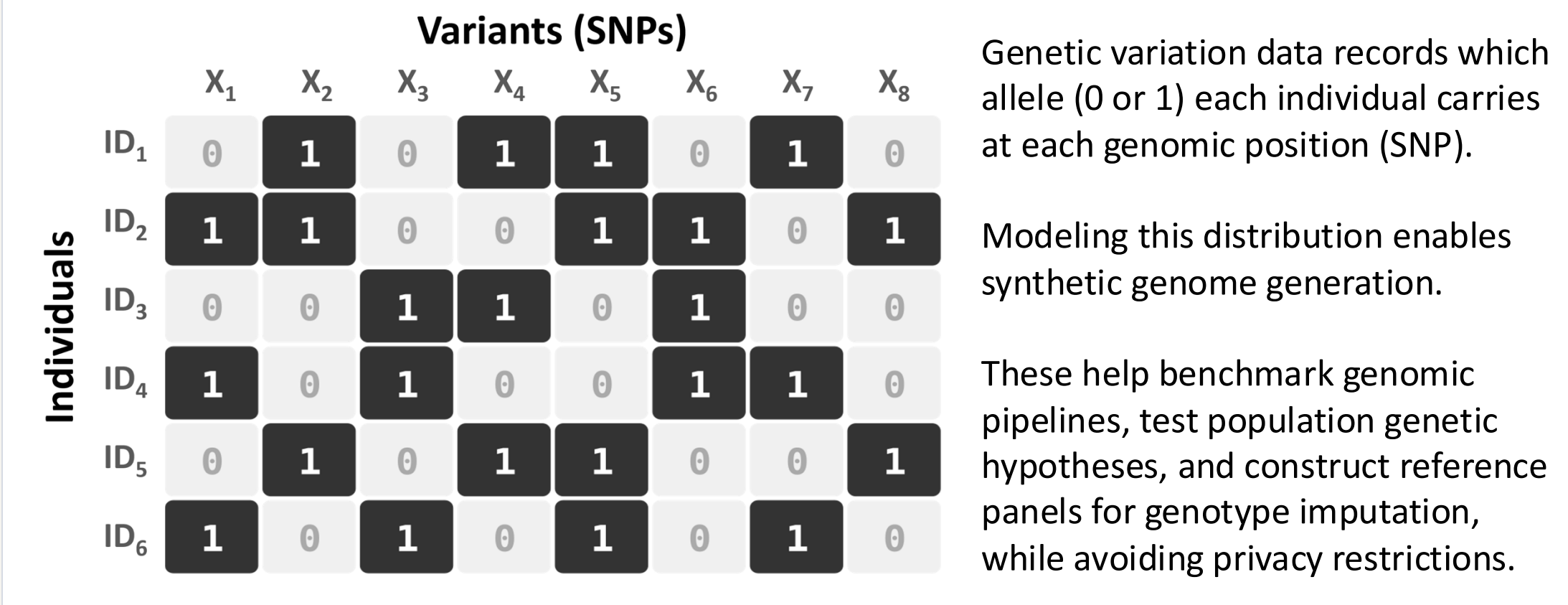


GPC: An expressive and tractable deep generative model for genetic variation data

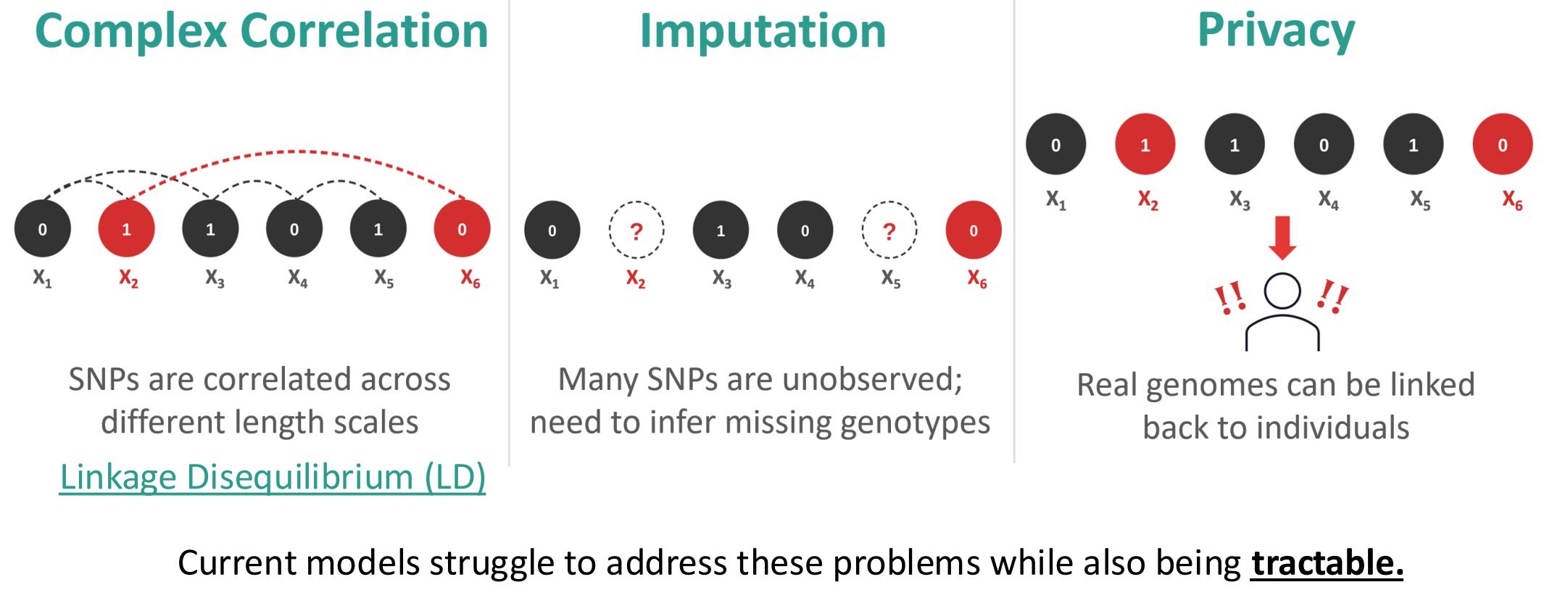
Prateek Anand¹, Anji Liu², Meihua Dang³, Boyang Fu⁴, Xinzhu Wei⁵, Guy Van den Broeck^{1,*}, Sriram Sankararaman^{1,6,7,*}

¹Dept. of Computer Science, UCLA ²Dept. of Computer Science, NUS ³Dept. of Computer Science, Stanford ⁴Dept. of Biomedical Informatics, Harvard Medical School
⁵Dept. of Computational Biology, Cornell ⁶Dept. of Human Genetics, UCLA ⁷Dept. of Computational Medicine, UCLA **Joint supervision*

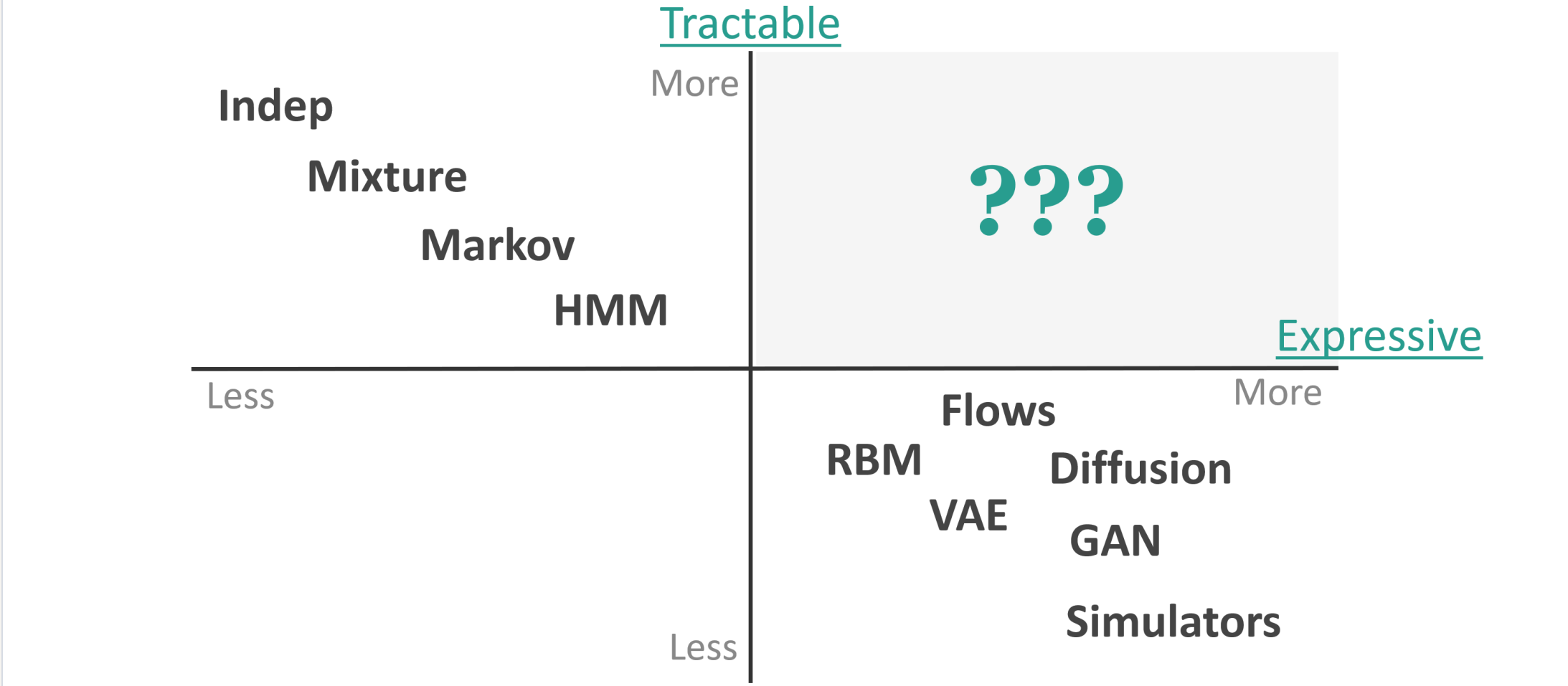
1 What is genetic variation data?



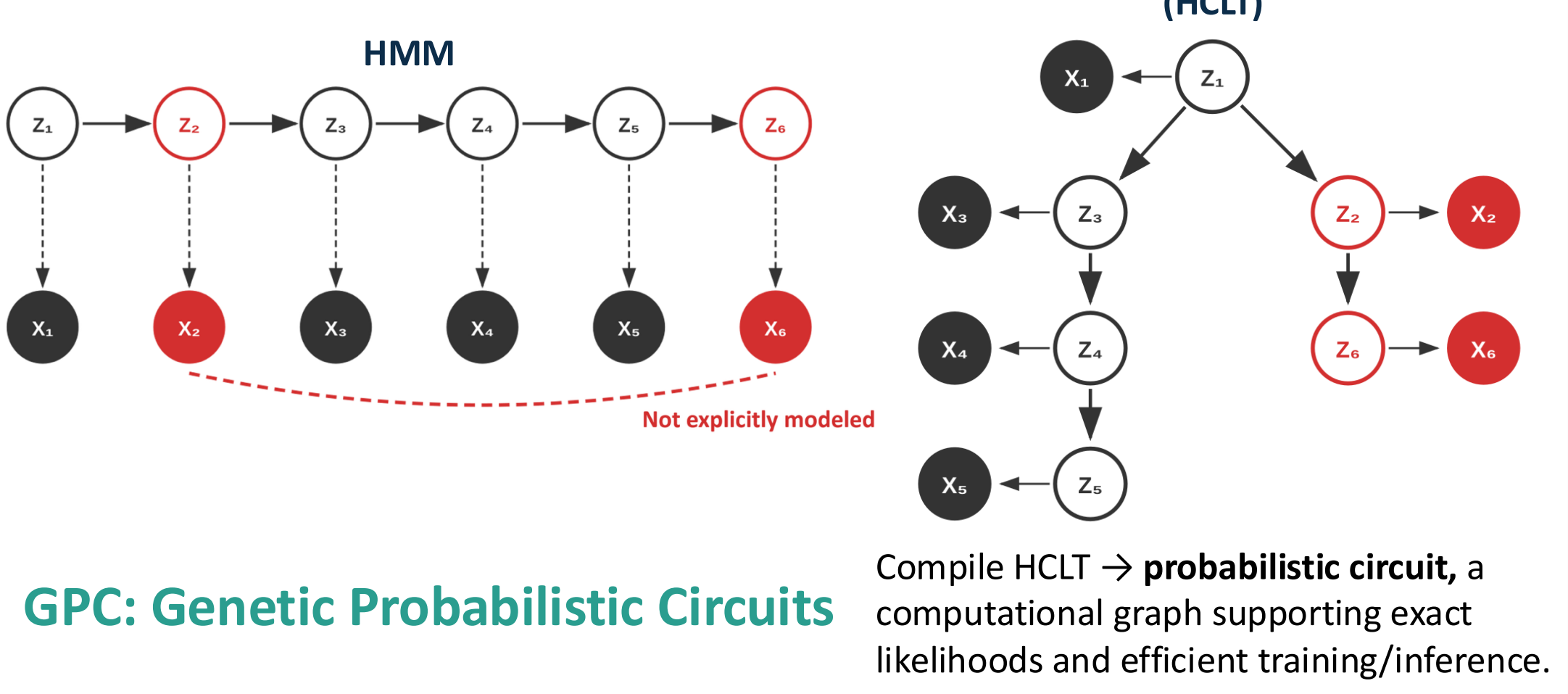
2 Three key challenges



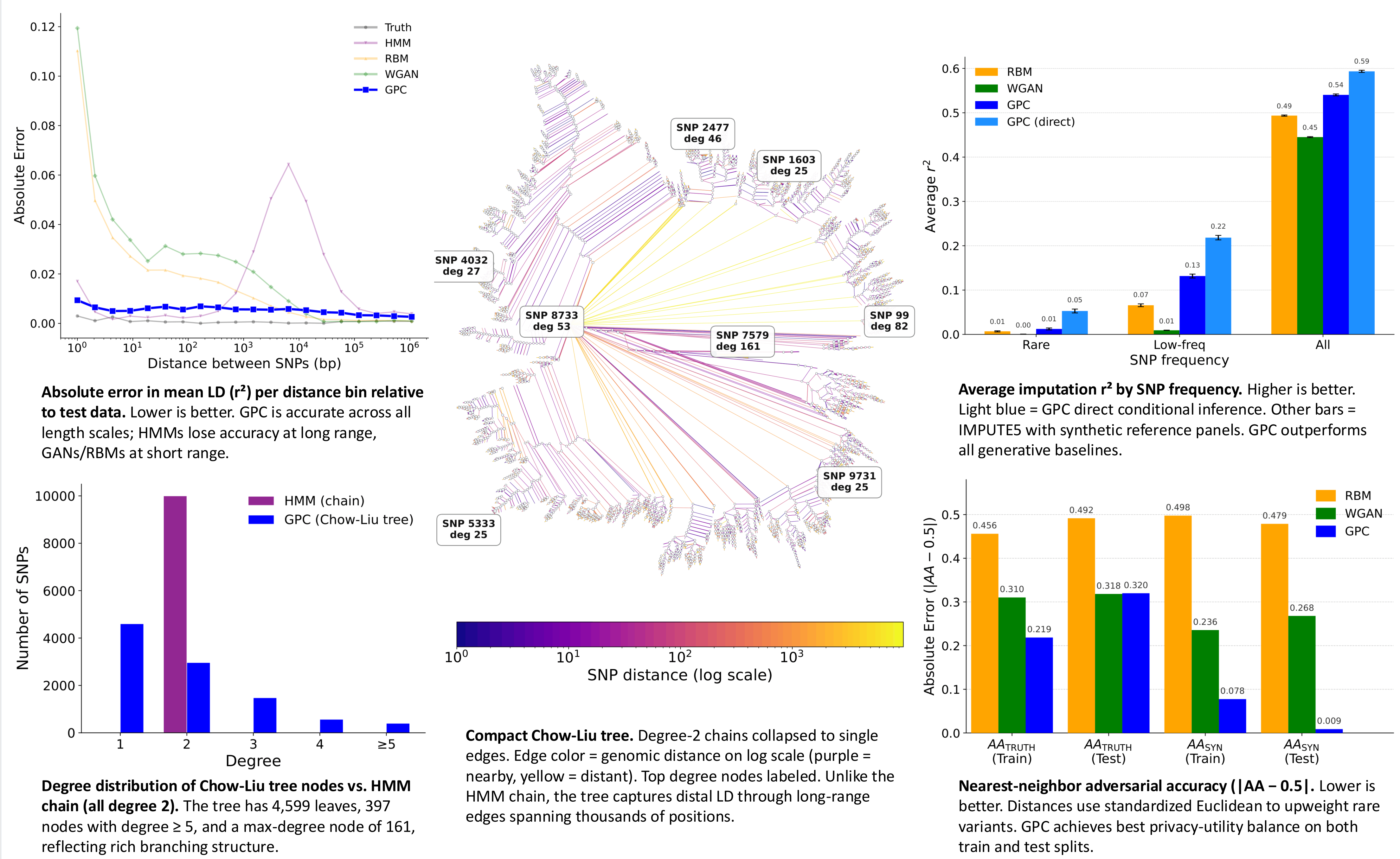
3 Tradeoffs in generative modeling



4 From chains to trees



5 Results



These figures highlight results from the 1000 Genomes Project. Results replicate in several genomic regions chosen from the UK Biobank as well.

GPC takeaways

- Expressive, tractable, and efficient:** Hidden Chow-Liu tree captures LD with exact inference in a probabilistic circuit.
- Best imputation:** Outperforms all generative baselines on imputation accuracy.
- Better privacy:** Superior privacy-utility tradeoff for synthetic genomes.

References
Chow and Liu IEEE 1968; Vergari et al. AAAI 2020; Liu et al. ICML 2024; Yelmen et al. PLoS Comp Bio 2023; Yale et al. 2019

Links



Preprint



GitHub