Generating Synthetic Genotypes using Diffusion Models

Philip Kenneweg, Raghuram Dandinasivara, Xiao Luo,
Barbara Hammer, Alexander Schönhuth

University Bielefeld

July 21, 2025

Why do we need Synthetic Data?

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Large high quality datasets are becoming available but getting access is a long and costly process (for example UKBB [1])
- High quality synthetic data would allow genomics data to be freely distributed (i.e. Kaggle dataset) without privacy concerns.

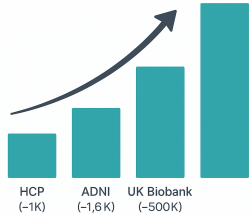
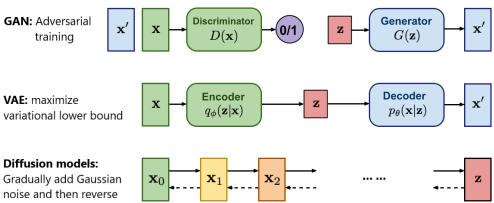


Figure: Increasing availability of large scale genomics datasets.

[1] UK Biobank – cohort of 500 000 participants with extensive health, genetic and imaging data. "About our data" page, UK Biobank. https://www.ukbiobank.ac.uk/enable-your-research/about-our-data.

Possible Generative Models

- Variational Autoencoder
- Generative Adversarial Networks
- Generative Markov Models
- Diffusion Models
- Autoregressive Models (LLM Style)



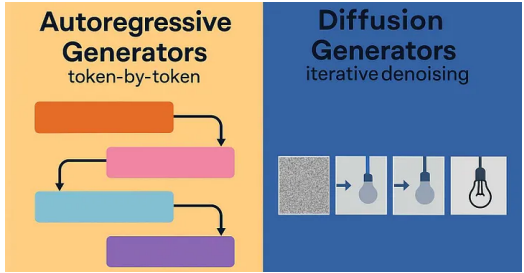
Possible Generative Models

- Variational Autoencoder - Output quality not state of the art, does not scale well, ...
- Generative Adversarial Networks - Unstable, hard to train, ...
- Generative Markov Models - Not very powerfull or general, ...

Possible Generative Models

- Diffusion Models - best suited for fixed length generation with causal interactions in all directions, can produce very high quality output by refining over multiple passes
- Autoregressive Models (LLM Style) - best suited for non-fixed length with causal interactions left to right, produces good output, mostly used for high level tasks

Genome has interactions in all directions, fixed length, high quality needed
-> Diffusion Models



Related Work

UNIVERSITÄT BIELEFELD
Faculty of Technology

Table: An overview of related work on generating synthetic genomes and its differences / similarities in comparison with our work. Our novelties are **highlighted**.

Reference	Model	Data Type	Genome Length	Cond.
DNAGPT (Zhang et al., 2023)	Autoregressive	Base-Pairs	24k BPS	x
HyenaDNA (Nguyen et al., 2023)	Autoregressive	Base-Pairs	10^6 BPS	x
HAPNEST (Wharrie et al., 2023)	LD & Markov	SNPs	1 Chromosome	x
(Perera et al., 2022)	GMMNs	SNPs	1 Chromosome	✓
(Yelmen et al., 2021)	GAN,RBM	SNPs	10k SNPs	x
(Yelmen et al., 2023)	WGAN	SNPs	10k SNPs	x
(Szatkownik et al., 2024)	WGAN	PCA+SNPs	65k SNPs	x
(Ahronoviz and Gronau, 2024)	GAN	SNPs	10k SNPs	✓
(Burnard et al., 2023)	VAE	SNPs	1 Chromosome	x
(Dang et al., 2023)	HCLTs	SNPs	10k SNPs	x
GeneticDiffusion (Ours)	**Diffusion**	PCA+SNPs	**Full Genome**	✓

Pre-Processing

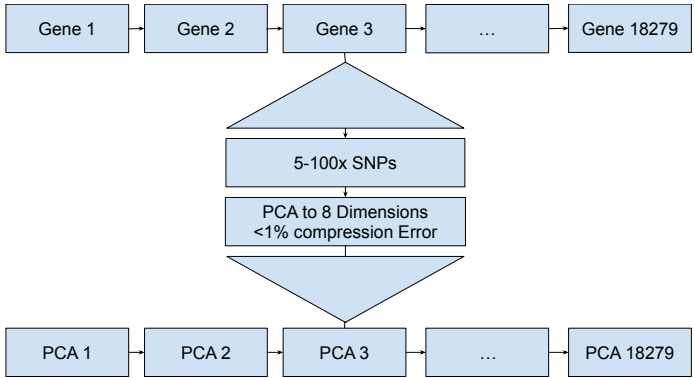
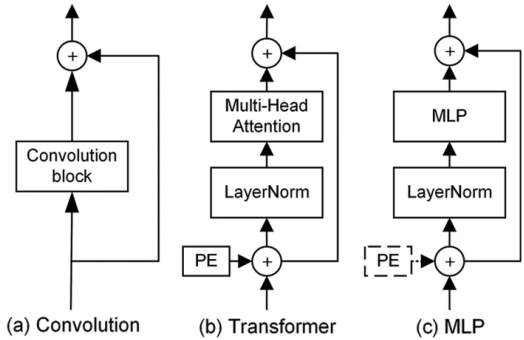


Figure: Overview of the Pre-processing pipeline. Genes, which consist of between 5 to 100 SNPs are each processed by a custom PCA. This is done independently for each Gene.

Model Architecture

Possible options:

- Transformer - No spatial bias, popular, needs high amount of training data
- U-Net CNN - Local spatial bias, few parameters
- U-Net MLP - Overfits easily, hard to train
- Combinations are also possible



Model Architecture

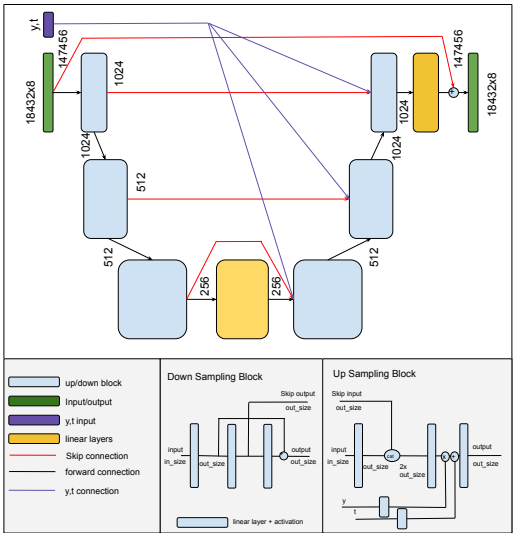


Figure: A structural overview of the architecture of the MLP diffusion model.

Evaluation

We evaluate on 2 different classification tasks:

- ALS classification
- 1KG region classification

Recovery Rate compares the performance of the syn model a_s vs the true model a_r:

$$R(a_r, a_s) = \frac{a_s}{a_r} \tag{1}$$

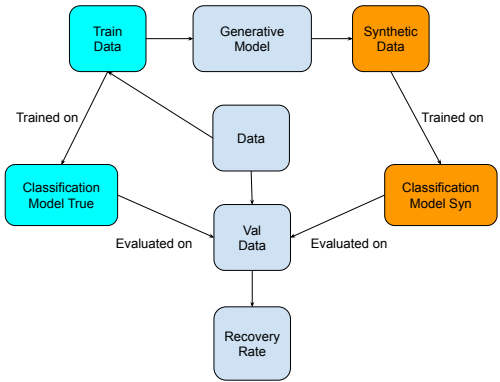
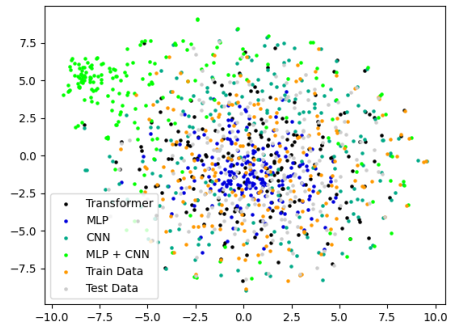


Figure: A diagram of the evaluation pipeline.

Evaluation

Additional metric:
Nearest Neighbour Adversarial Accuracy (Yale et al., 2019) - relies on distances of generated samples to original samples.

Problem:
Distances/neighbourhood in high dimensional space are not inherently meaningful (Beyer et al., 1999).



Experiments

We evaluated 4 different generative architectures:

- MLP
- CNN
- MLP + CNN
- Transformer

The metrics we used:

- Recovery Rate
- Nearest Neighbour Adversarial Accuracy
- Accuracy improvement with partial synthetic data

Results Recovery Rate

Table: Recovery rates on a hold out test set of true genotypes after training different ALS or 1KG population classifiers (MLP, Transformer or CNN) on different synthetically generated data types (generated by: MLP, Transformer, CNN, MLP + CNN). The best synthetic data for each classifier type is marked in **bold**.

Classifier	CNN	MLP	MLP+CNN	Transformer
		ALS data		
MLP	71.51	**96.69**	94.26	73.77
Transformer	66.06	93.44	**93.89**	69.30
CNN	69.88	91.72	**93.17**	68.72
		1KG data		
MLP	15.58	65.80	**93.02**	13.28
Transformer	16.23	62.99	**84.98**	8.38
CNN	19.52	56.57	**77.54**	21.21
Average (all)	43.17	78.06	**89.48**	42.56

Results NNAA

Table: Result of Nearest Neighbour Adversarial Accuracy for generated datasets on the ALS and 1KG data; For AA the closer to 0.5 the better, For Privacy Loss the closer to 0 the better; best performance is **highlighted**.

			CNN	MLP	MLP + CNN	Transformer
ALS data	test data	AA_{truth}	0.735	0.255	**0.485**	0.92
		AA_{syn}	0.68	1.0	0.93	**0.66**
	train data	AA_{truth}	0.81	0.005	**0.405**	0.93
		AA_{syn}	**0.67**	1.0	0.92	0.69
		Privacy Loss	0.0325	0.125	0.0475	**0.02**
1KG data	test data	AA_{truth}	0.76	0.0	**0.63**	0.345
		AA_{syn}	0.995	1.0	0.94	**0.92**
	train data	AA_{truth}	0.765	0.0	**0.385**	0.285
		AA_{syn}	1.0	0.99	**0.74**	0.82
		Privacy Loss	0.05	**-0.005**	-0.2225	0.08

Partial Synthetic and Real

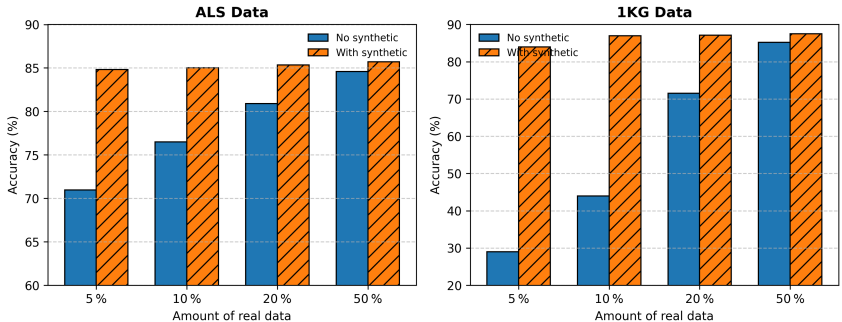


Figure: Accuracy improvements by integration of synthetic data for best performing classification architecture.

Conclusion

- Synthetic data does not just copy real data, while close to real data distribution (see NNAA)
- Recovery Rate indicates high fidelity of synthetic data.

Next Steps:

- Bigger dataset (for example UK Biobank)
- Multimodal data
- Publish synthetic dataset

Questions and Discussion

Feel free to ask anything.

Bibliography

Ahronoviz, S. and Gronau, I. (2024).
Genome-ac-gan: Enhancing synthetic genotype generation through auxiliary classification.
bioRxiv, pages 2024–02.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999).
When is "nearest neighbor" meaningful?
In *International conference on database theory*, pages 217–235. Springer.

Burnard, C., Mancheron, A., and Ritchie, W. J. (2023).
Generating realistic artificial human genomes using adversarial autoencoders.
bioRxiv, pages 2023–12.

Dang, M., Liu, A., Wei, X., Sankararaman, S., and Van den Broeck, G. (2023).
Tractable and expressive generative models of genetic variation data.
bioRxiv.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., and Ré, C. (2023).
Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution.

Perera, M., Montserrat, D. M., Barrabés, M., Geleta, M., Giró-i Nieto, X., and Ioannidis, A. G. (2022).
Generative moment matching networks for genotype simulation.
In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1379–1383. IEEE.

Szatkownik, A., Furtlehner, C., Charpiat, G., Yelmen, B., and Jay, F. (2024).
Towards creating longer genetic sequences with gans: Generation in principal component space.
In *Machine Learning in Computational Biology*, pages 110–122. PMLR.

Wharrie, S., Yang, Z., Raj, V., Monti, R., Gupta, R., Wang, Y., Martin, A., O'Connor, L. J., Kaski, S., Marttinen, P., Palamara, P. F., Lippert, C., and Ganna, A. (2023).
HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes.
Bioinformatics, 39(9):btad535.

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., and Bennett, K. P. (2019).
Privacy preserving synthetic health data.
In *ESANN 2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

Yelmen, B., Decelle, A., Boulos, L. L., Szatkownik, A., Furtlehner, C., Charpiat, G., and Jay, F. (2023).
Deep convolutional and conditional neural networks for large-scale genomic data generation.
PLoS Computational Biology, 19(10):e1011584.

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., and Jay, F. (2021).
Creating artificial human genomes using generative neural networks.
PLoS genetics, 17(2):e1009303.

Zhang, D., Zhang, W., He, B., Zhang, J., Qin, C., and Yao, J. (2023).
Dnagpt: a generalized pretrained tool for multiple dna sequence analysis tasks.
bioRxiv, pages 2023–07.