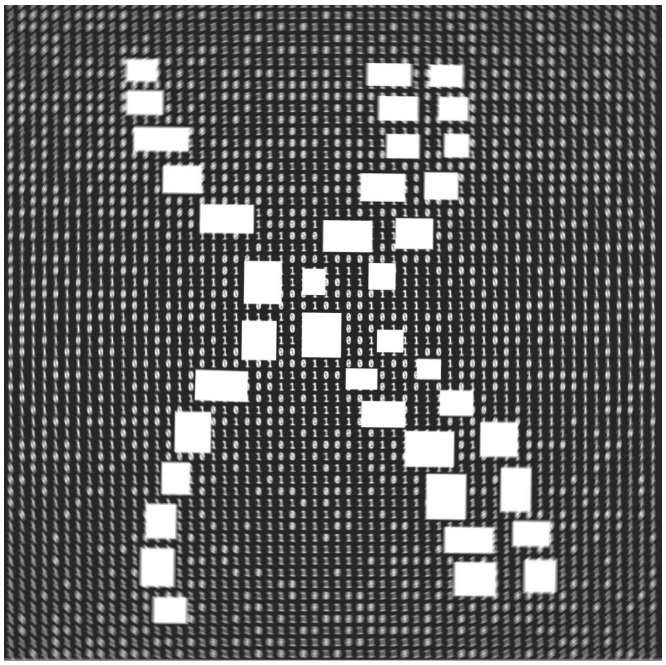
Machine learning generates realistic genomes for imaginary humans

5 February 2021

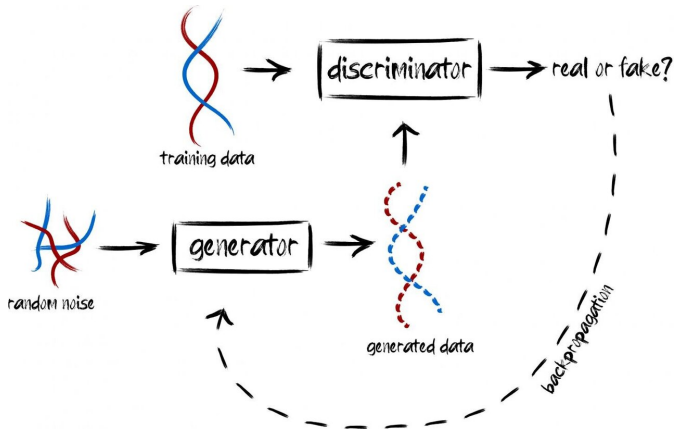


A chromosome emerges from random digital noise. Credit: Burak Yelmen

Machines, thanks to novel algorithms and advances in computer technology, can now learn complex models and even generate high-quality synthetic data such as photo-realistic images or even resumes of imaginary humans. A study recently published in the international journal *PLOS Genetics* uses machine learning to mine existing biobanks and generate chunks of human genomes which do not belong to real humans but have the characteristics of real genomes.

"Existing genomic databases are an invaluable resource for biomedical research, but they are either not publicly accessible or shielded behind long and exhausting application procedures due to valid ethical concerns. This creates a major scientific barrier for researchers. Machine-generated genomes, or artificial genomes as we call them, can help us overcome the issue within a safe ethical framework," said Burak Yelmen, first author of the study and Junior Research Fellow of Modern Population Genetics at the University of Tartu.

The pluridisciplinary team performed multiple analyses to assess the quality of the generated genomes compared to real ones. "Surprisingly, these genomes emerging from random noise mimic the complexities that we can observe within real human populations and, for most properties, they are not distinguishable from other genomes from the biobank we used to train our algorithm, except for one detail: they do not belong to any gene donor," said Dr. Luca Pagani, one of the senior authors of the study and a Mobilitas Pluss fellow.



A generator machine shapes random noise while a discriminator machine tests the generated data against a database of available real data. Once the process is complete, the algorithm will generate artificial data that looks like the real one, but is actually completely new. Credit: Yelmen et al. 2021

The study additionally involves the assessment of the proximity of artificial genomes to real genomes

to test whether the privacy of the original samples is preserved. "Although detecting privacy leaks among thousands of genomes could appear as looking for a needle in a haystack, combining multiple statistical measures allowed us to check all models carefully. Excitingly, the detailed exploration of complex leakage patterns can lead to improvements in generative model evaluation and design, and will fuel back the machine learning field," said Dr. Flora Jay, the coordinator of the study and CNRS researcher in the Interdisciplinary computer science laboratory (LRI/LISN, Université Paris-Saclay, French National Centre for Scientific Research).

All in all, machine learning approaches had provided faces, biographies and multiple other features to a handful of imaginary humans: now we know more about their biology. These imaginary humans with realistic genomes could serve as proxies for all the real genomes which are not publicly available or require long application procedures or collaborations, hence removing an important accessibility barrier in genomic research, in particular for underrepresented populations.

More information: Burak Yelmen et al, Creating artificial human genomes using generative neural networks, *PLOS Genetics* (2021). DOI: 10.1371/journal.pgen.1009303