31.JULY.2023

SIMON (SIgnal siMulated through diffusiON)

A Novel Diffusion Machine Learning Model to Generate ONT Nanopore Signal Data from Sequencing Data

By Raehash Shah

Bioinformatics Intern - QC Virology - Manager: Christopher Remillard

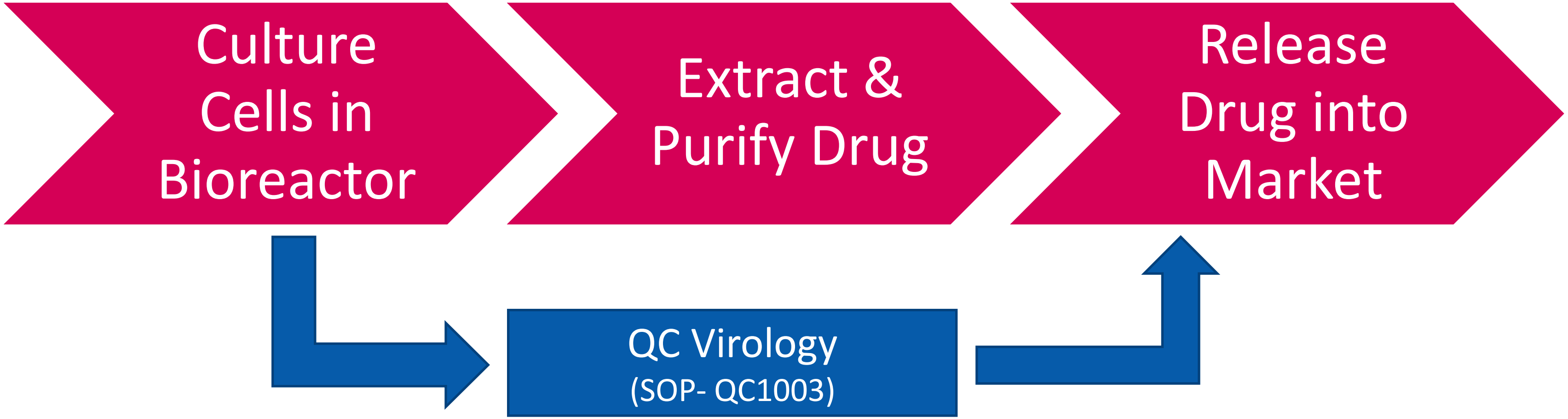
Carnegie Mellon University School of Computer Science

REGENERON®

Computational Biology Department

QC Virology Overview

Want to test for viruses in the manufacturing of drugs



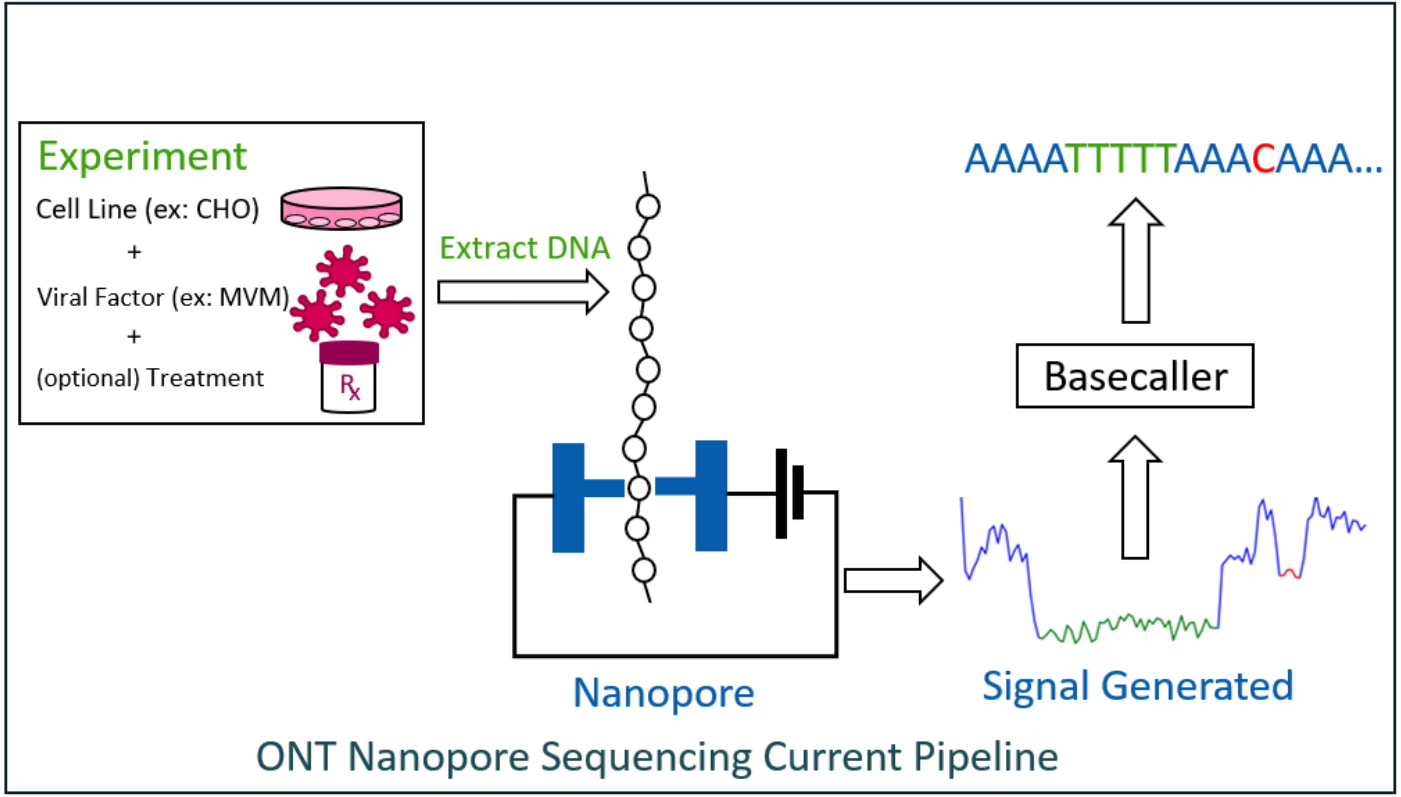
Culture Cells in Bioreactor

Extract & Purify Drug

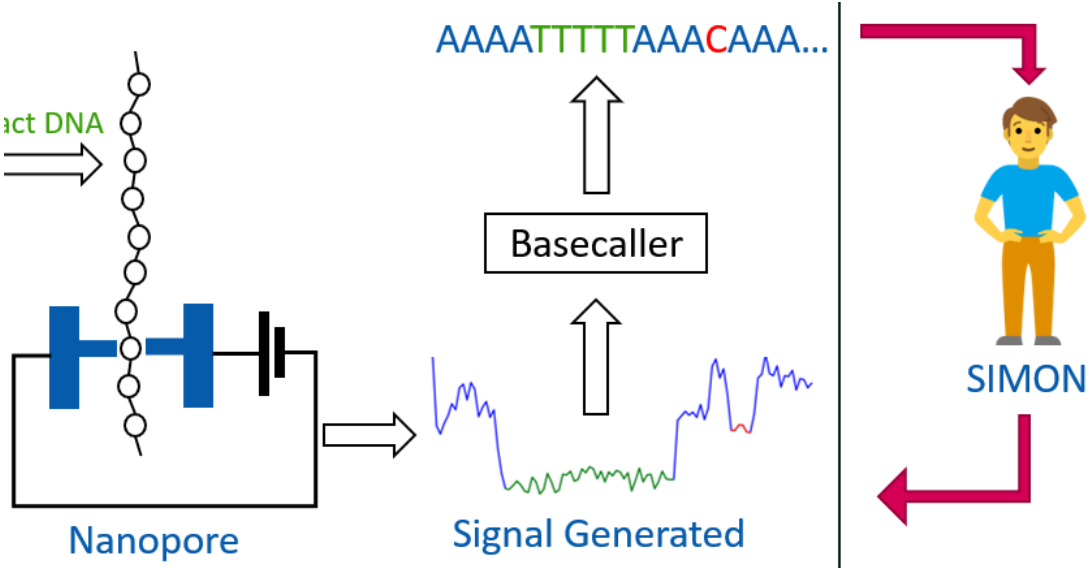
Release Drug into Market

QC Virology (SOP- QC1003)

Identify Potential Viral Species with Sequencing (i.e., Nanopore Sequencing)



Create a Model that Generates Synthetic Signal Data



Goals of the Model is to simulate data

- For the development & testing of computational tools
- For inaccessible viruses
  (i.e., HIV, BSL 3+)

In a **fast, accurate** and **inexpensive** way!

SIMON (Conditional Diffusion Machine Learning Model)

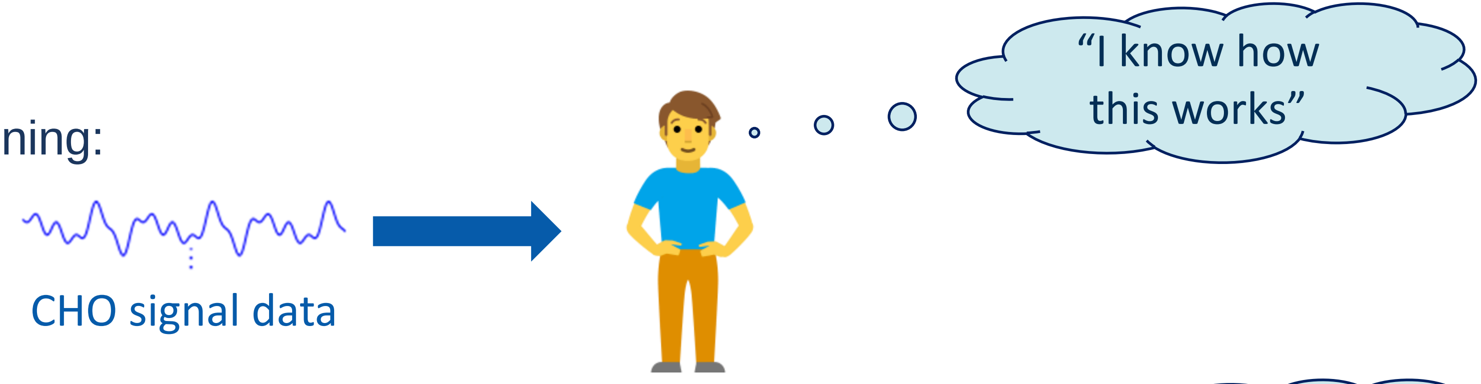
- Training:

CHO signal data

SIMON

SIMON (Conditional Diffusion Machine Learning Model)

- Training:

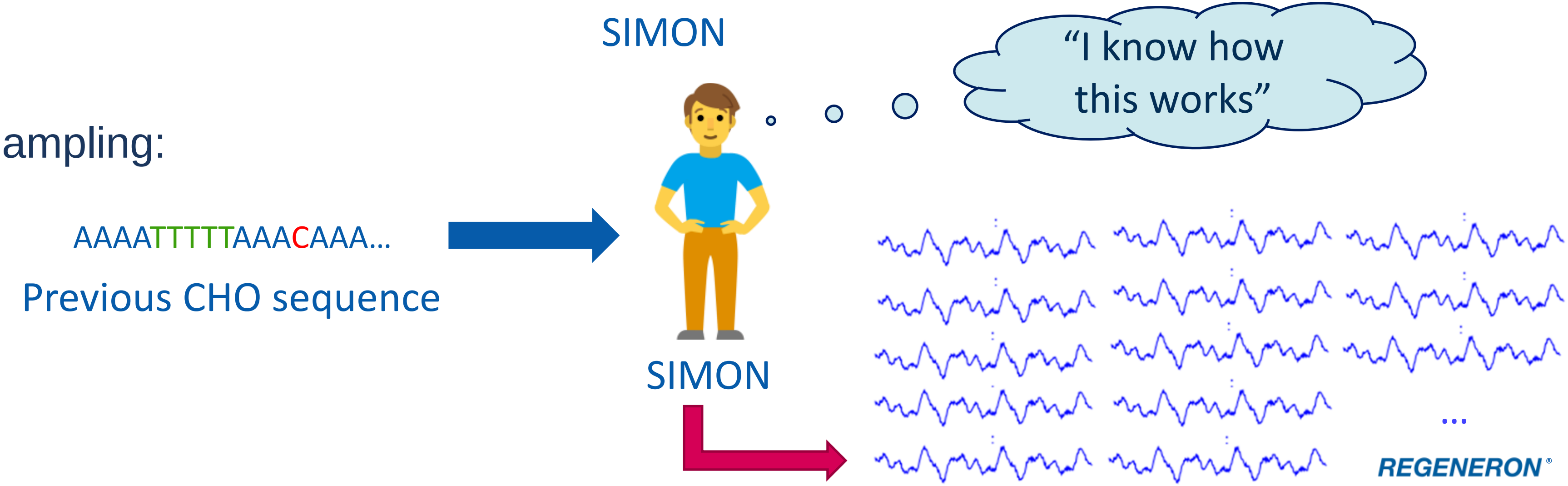


CHO signal data

SIMON

"I know how this works"

- Sampling:



AAAATTTTTAAACAAA...

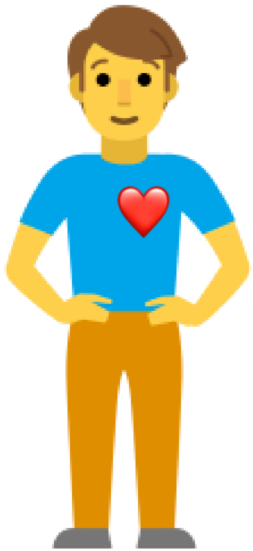
Previous CHO sequence

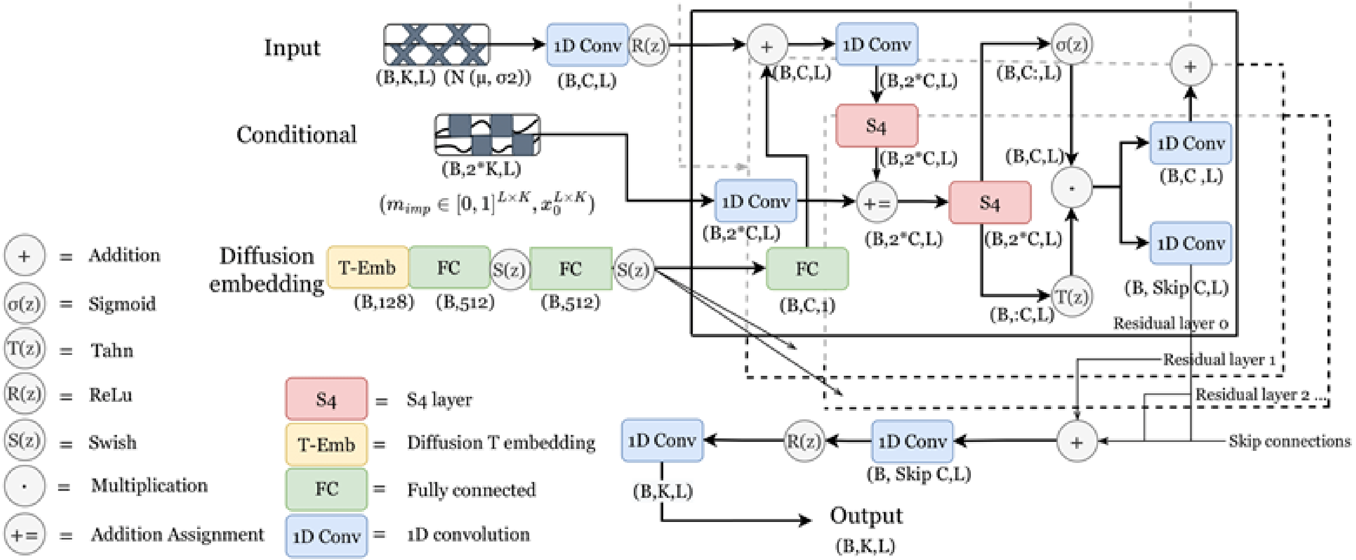
SIMON

"I know how this works"

...

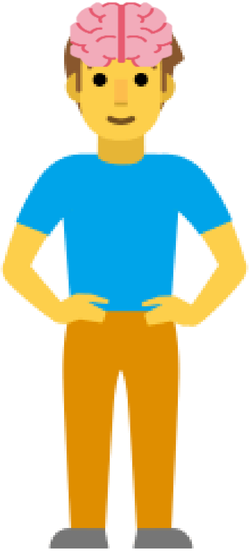
In reality, SIMON is much more complicated than that…





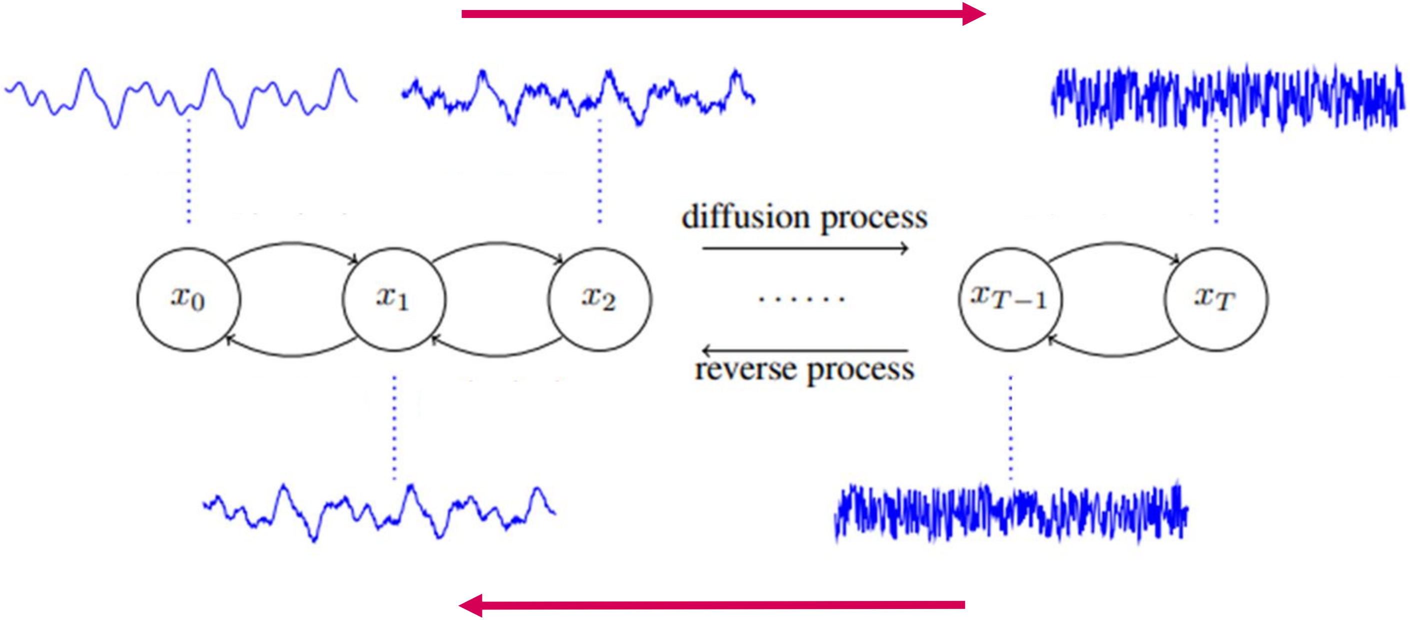
SIMON's architecture

In reality, SIMON is much more complicated than that…



SIMON’s behavior

Training

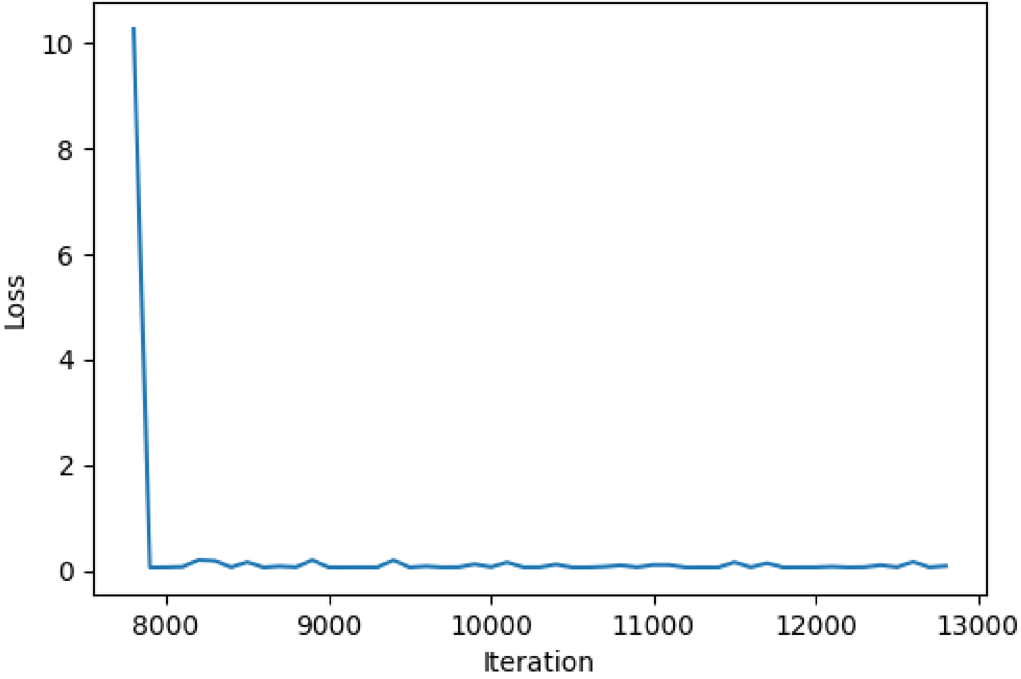


Sampling

Loss Function: $L = \boldsymbol{min}_{\theta} \mathbb{E}_{x_0 \sim D,\, \epsilon \sim N(0,1),\, t \sim U(1,T)} ||\epsilon - \epsilon_{\theta}(\sqrt{\alpha_t} x_0 + (1-\alpha_t)\epsilon, t)||_2^2$

So, I’m sure many of you are wondering…how well did SIMON do?

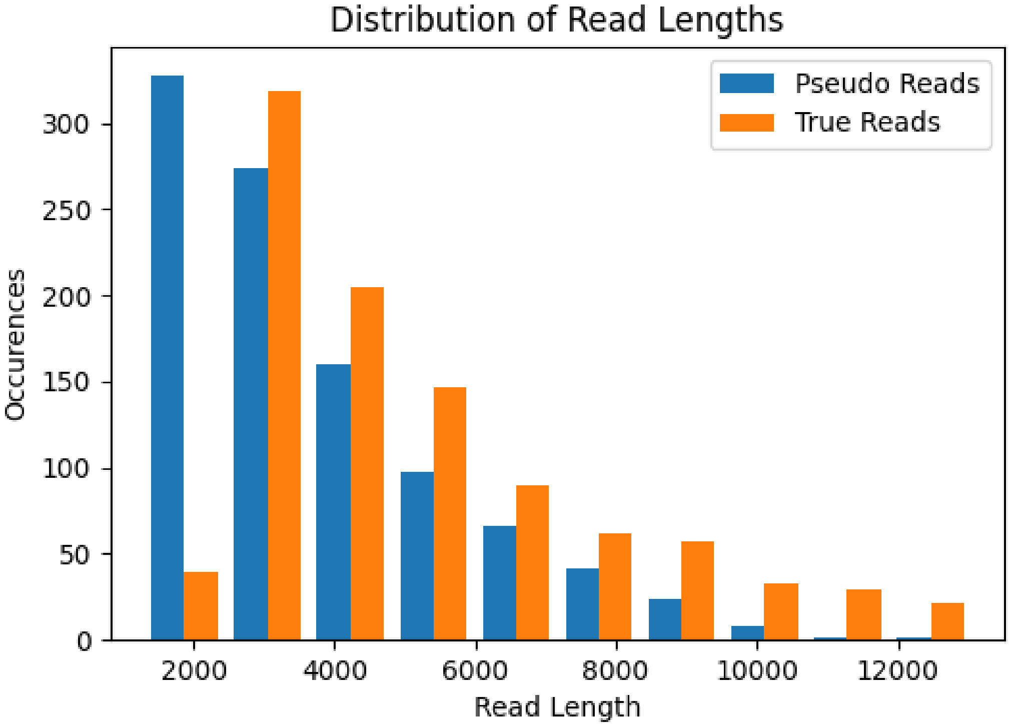
SIMON has undergone training



Final Loss Value of 0.088 -> 91.2% Accuracy

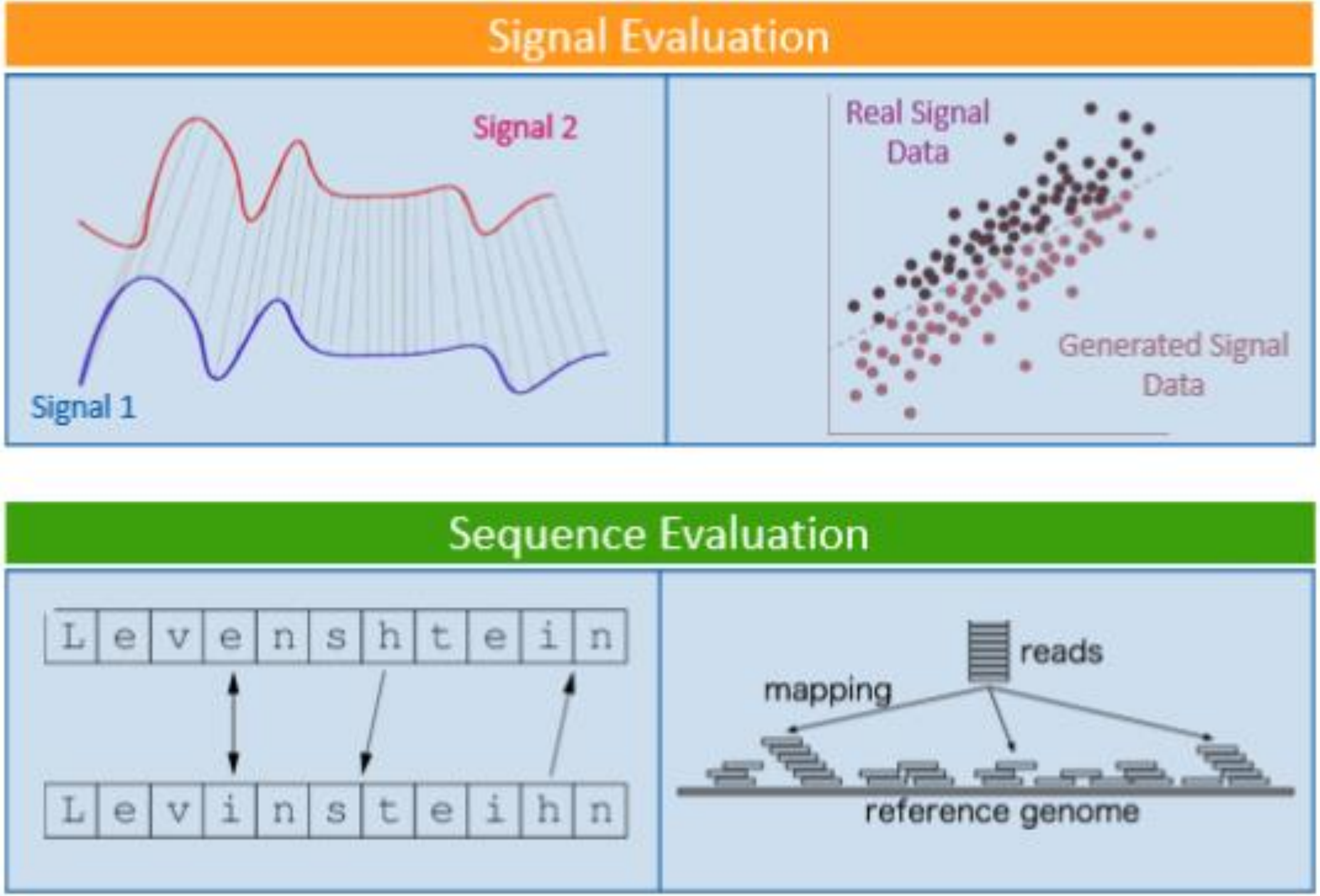
So, I’m sure many of you are wondering…how well did SIMON do?

Generated 20,000 Signal Datapoints

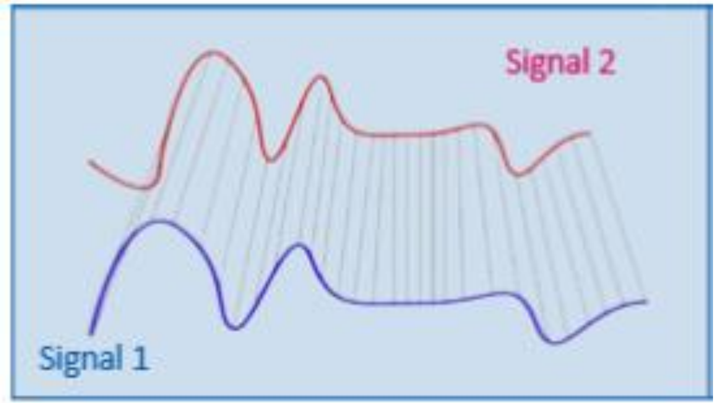


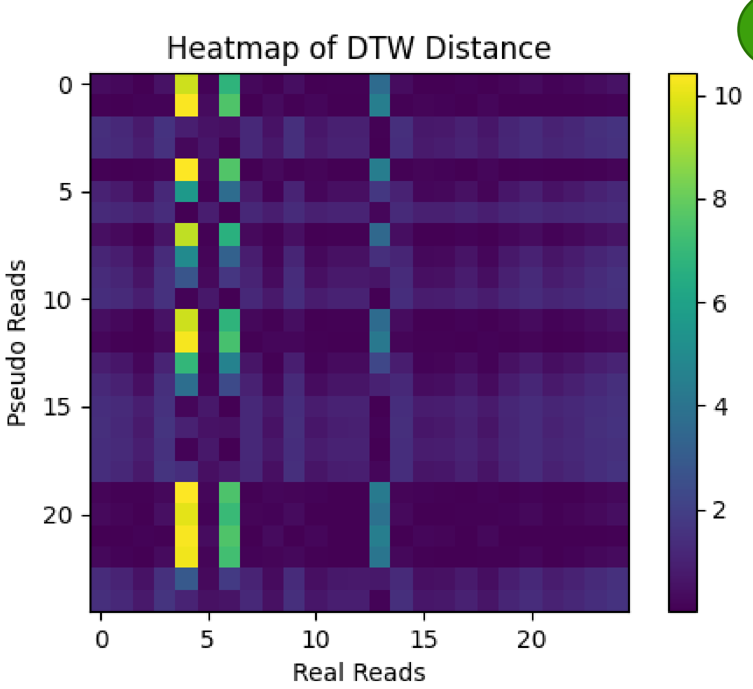
Note that SIMON fit the best distribution but this led to SIMON significantly overestimating the number of “shorter” reads which is where SIMON can be improved.

How to evaluate these signals generated by SIMON?



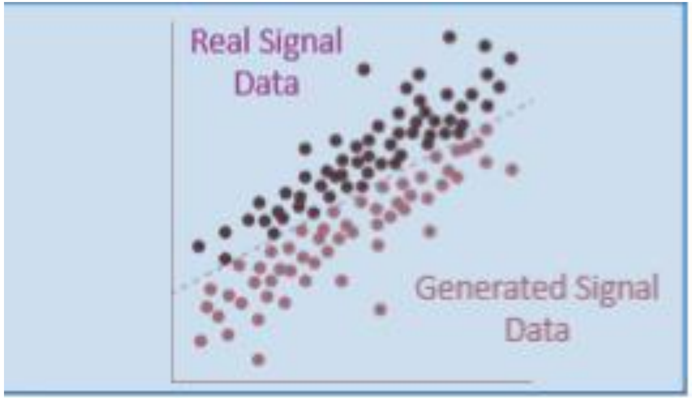
How to evaluate these signals generated by SIMON?

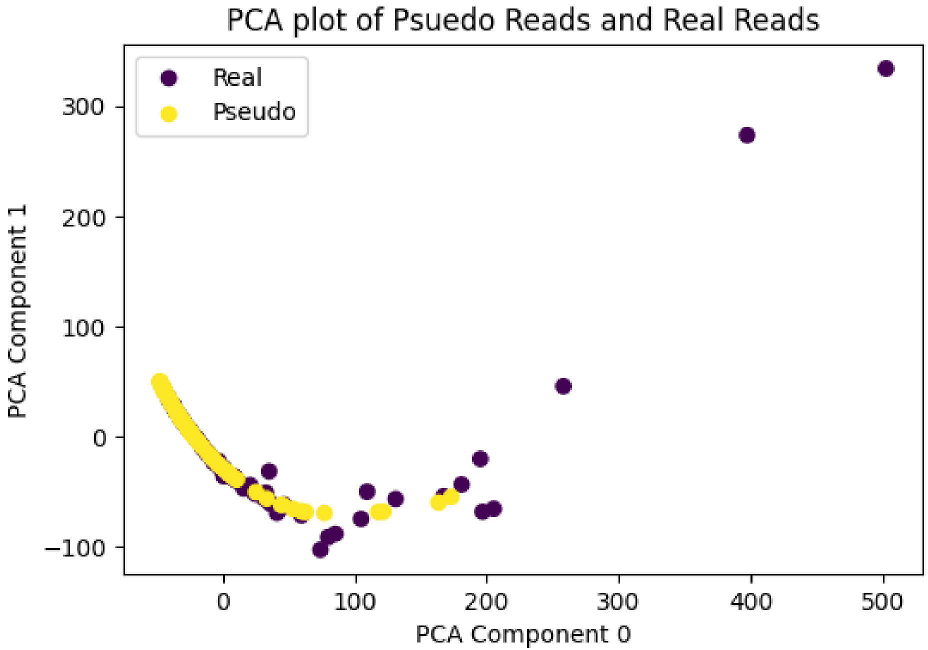




This is the normalized DTW distance, so each value represents the number of waveform pattern adjustments necessary to align the two signal values. Note the larger DTW distance may not simply indicate more adjustments necessary but rather that one real read is longer than others.

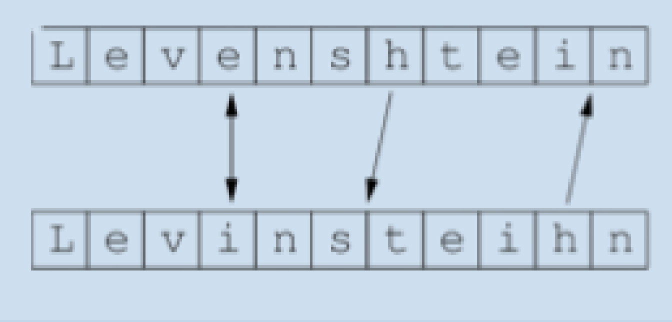
How to evaluate these signals generated by SIMON?

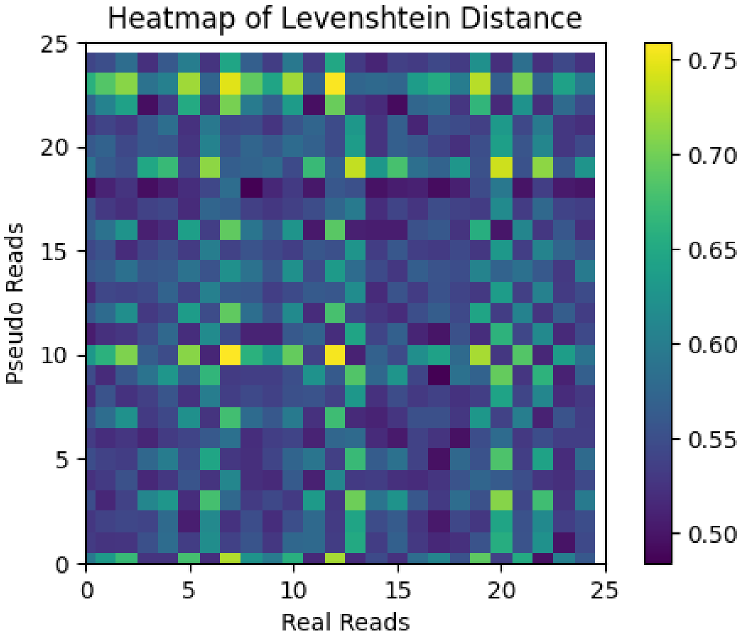




Binary Classifier had an accuracy of 46% in separating these reads which is good since a trained model does worse than randomly assigning values suggesting that the data appears to be really similar to each other. In addition, as seen by the PCA plot, the real and generated reads are quite indistinguishable, however there is still some room for improvement to be made here as the generated reads don't completely follow the real sequencing reads

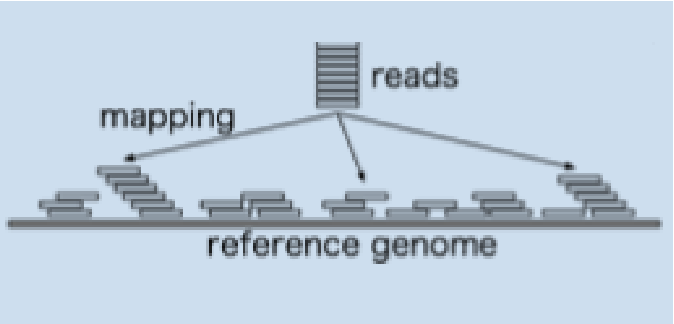
How to evaluate these signals generated by SIMON?

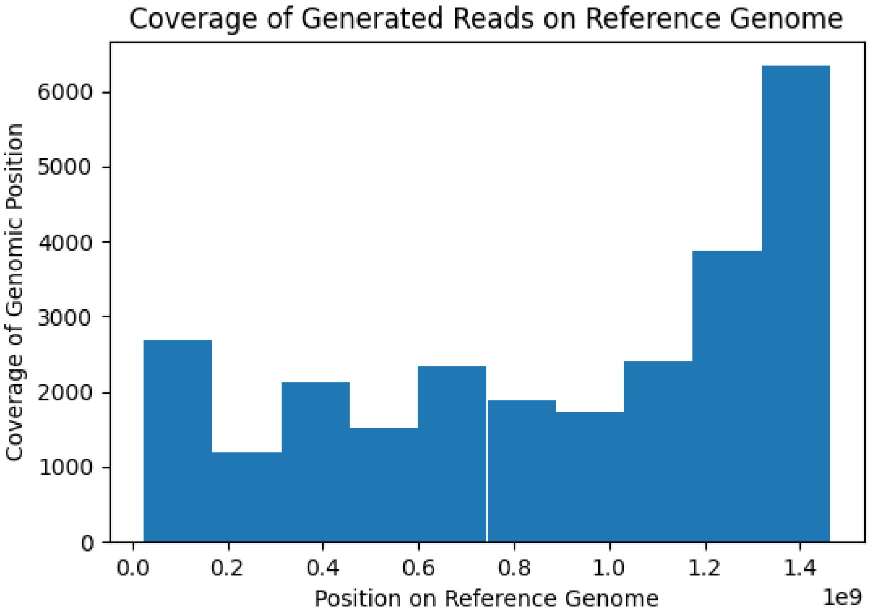




The number of mismatches/indels normalized by the longer read were plotted here to show how many operations needed to be made to match the shorter read on average per base pair. As seen in the heatmap, it seems to 2 operations every 3 base pairs which is quite high and could be problematic in isolation. However, a mismatch doesn't necessarily mean that the read isn't a representative of the genome so we should use this metric more as a supplement with other evaluation metrics

How to evaluate these signals generated by SIMON?





This figure shows how well the reads mapped to the reference genome. As we can see there is a quite a nice spread across the reference genome and the coverage is quite high as well suggesting that the generated reads are quite representative of the reference genome and therefore accurate.

Future Directions

Compare
Simulated Data

Improve SIMON’s
Accuracy

Additional Applications
of SIMON

Thank you!!!

Any questions?