Internship Report

Understanding of Aims of the Project

During my project, the overall aim of my group's project was to advance the field of synthetic DNA sequence generation and genomics by developing a novel approach using a latent diffusion machine learning model to generate DNA sequences. Specifically, I worked on the UNET subcomponent of the overall model and on the collection of promoter sequence data from the EPDnew website. This research is important as it contributes to advances in our ability to design functional DNA sequences, particularly promoter sequences.

Summary of the Work Undertaken

Methodologies and Techniques

I worked extensively in Python, specifically using the pytorch package, a very popular tool for working with machine learning models, and during the data collection and processing part of the project I mainly used pandas dataframes (within Python). We used the capabilities of continuous diffusion models to generate discrete data. I also contributed to the construction of sequence similarity validation metrics by writing and compiling code for the Levenshtein distance metric using numba as a tool to evaluate the quality of DNA sequence generation.

Dataset Curation

One of my main tasks during my internship was the meticulous curation and creation of a cross-species dataset comprising 150,000 unique promoter gene sequences from 15 different species. Using the European Promoter Database (EPD) and the High-Throughput EPD (HT-EPD), I carefully selected sequences ranging from -1024 to +1023 base pairs relative to the transcription start site to ensure the inclusion of important promoter elements. The dataset, which is current as of August 2023, underwent thorough data collection, normalisation and integration processes, resulting in a well-structured resource containing 159,125 different promoter sequences associated with 130,014 individual genes. Each entry in this dataset has been associated with experimental expression levels and transcription start sites, making it a fundamental resource for the advancement of generative genomic modelling and research.

UNet Implementation

In my role in UNet construction, I was instrumental in integrating the standard diffusion model into the latent space. This involved training the transformation function and minimising the differences between the score matching function and the gradient of the probability density function. To achieve this, we used a UNet architecture, specifically from StableDiffusion, to capture important features and generate high quality samples. Our UNet model consisted of four down and four up blocks, each containing eight sequential ResNet

blocks and a single block for channel adjustments. I also implemented cross-attention layers in the third down and second up blocks to improve information capture. Within each ResNet block, I ensured the inclusion of critical elements such as normalisation, swish non-linearity, upsampling, two convolutional layers and time-embedding projection. In addition, attention layers with eight 64-dimensional attention heads were integrated into the third down and second up blocks, using a scaled dot product attention mechanism. To optimise the performance of the model for genomic generative modelling, I carefully defined its specifications, including input and output dimensions, channel dimensions, and the number of steps in the forward process (N = 1000).

Description of Results/Outcome of the Project and How This Will Be Taken Forward

Our research produced significant results. The model demonstrated its ability to generate synthetic DNA sequences that closely resembled real DNA sequences in terms of motif distribution, latent embedding distribution (FReD) and chromatin profiles. This highlights the potential of DiscDiff to advance the generation of synthetic DNA sequences.

We have now also submitted our work to the NeurIPS 2023 AI4Science Workshop, where it can be scrutinised by the wider scientific community, and hopefully contribute to wider discoveries in the field

Consideration of the Impact of the Work/Results

Our research is not only important in the context of synthetic DNA sequence generation but also in the broader field of genomics and biotechnology. It has the potential to benefit biologists, geneticists, and researchers working on DNA-related projects.

Discussion of Subject-Specific and Transferable Skills Gained

During my internship, I acquired a range of subject-specific skills, including expertise in machine learning, genomics, and data analysis. I also developed transferable skills such as project management and effective communication. These skills have significantly contributed to my personal and professional growth.

My internship experience has influenced my future career plans and goals. I am now more determined to pursue a career in computational biology and machine learning, where I can continue to make a meaningful impact on scientific research and innovation.

Conclusion

In conclusion, my internship focused on advancing synthetic DNA sequence generation using a latent diffusion machine learning model, particularly within the UNet component, and

involved curating a cross-species dataset. Our work is crucial for engineering functional DNA sequences, especially promoters.

During the internship, I extensively used Python and PyTorch for data processing and employed continuous diffusion models while introducing the FReD metric for quality evaluation. The curated dataset, containing 150,000 promoter-gene sequences from 15 species, is a valuable resource.

In UNet construction, I integrated the diffusion model, achieving high-quality sample generation. Our research has shown promising results, and the introduction of the FReD metric adds quantitative rigor. Our commitment to open-source the code will facilitate future research.

This work impacts genomics and biotechnology, benefiting professionals in various fields. Personally, I've gained valuable skills in machine learning and genomics, shaping my career aspirations in computational biology and machine learning for scientific research and innovation.