

RESEARCH

Open Access



Fine-tuning a sentence transformer for DNA

Mpho Mokoatle^{1,2*}, Vukosi Marivate^{1†}, Darlington Mapiye^{3†}, Riana Bornman^{5†} and Vanessa M. Hayes^{4,5,6†}

[†]Vukosi Marivate, Darlington Mapiye, Riana Bornman and Vanessa M. Hayes have contributed equally to this work.

*Correspondence:
Mpho Mokoatle
mphomokoatle64@gmail.com

Full list of author information is available at the end of the article

Abstract

Background Sentence-transformers is a library that provides easy methods for generating embeddings for sentences, paragraphs, and images. Sentiment analysis, retrieval, and clustering are among the applications made possible by the embedding of texts in a vector space where similar texts are located close to one another. This study fine-tunes a sentence transformer model designed for natural language on DNA text and subsequently evaluates it across eight benchmark tasks. The objective is to assess the efficacy of this transformer in comparison to domain-specific DNA transformers, like DNABERT and the Nucleotide transformer.

Results The findings indicated that the refined proposed model generated DNA embeddings that exceeded DNABERT in multiple tasks. However, the proposed model was not superior to the nucleotide transformer in terms of raw classification accuracy. The nucleotide transformer excelled in most tasks; but, this superiority incurred significant computing expenses, rendering it impractical for resource-constrained environments such as low- and middle-income countries (LMICs). The nucleotide transformer also performed worse on retrieval tasks and embedding extraction time. Consequently, the proposed model presents a viable option that balances performance and accuracy.

Keywords Sentence transformers, BERT, DNABERT, SimCSE, The nucleotide transformer

Background

Making machines understand and produce human language is the goal of the branch of artificial intelligence known as natural language processing (NLP). Despite the fact that NLP has been around for a while, its popularity has lately increased due to pre-trained language models. pretrained language models such as Transformers have been pretrained on extensive datasets and fine-tuned on task-specific data. By training these models on large datasets, this allows universal language representations to be learned. These representations are helpful for many downstream NLP tasks, including text summarization, named entity recognition, sentiment analysis, part-of-speech tagging, and many others [1, 2].

Transformer models such as BERT [3], Transformer-XL [4], XLNet [5], and many others have been employed in the field of genomics to resolve numerous DNA-related tasks due to the flood of next-generation sequencing data. For instance, a research publication [6] developed a DNA version of BERT, called DNABERT. DNABERT is a Transformer



model that was implemented in a genomic setting and resolves numerous DNA-related tasks, such as the identification of promoter regions and transcription factor binding sites (TFBS). While DNABERT and other similar language models have demonstrated the application of language models in a DNA context, this study specifically concentrates on refining a Sentence Transformer model for DNA tasks. This work demonstrates that the embeddings generated by a fine-tuned natural language based-model yield outcomes comparable to those produced by larger DNA-based language models. Thus, the hypothesis in this study is that embeddings generated from a natural language-based model, when fine-tuned on DNA sequences, can in certain settings outperform embeddings derived from large language models pretrained exclusively on genomic data. This hypothesis is evaluated through eight DNA benchmark datasets that involve binary and multi-label classification tasks. In doing so, the study seeks to answer whether such embeddings are competitive with more complex domain-specific models, and how simple fine-tuning strategies such as a single epoch of training on limited sequence data affect performance.

Related work

In recent years, individual aspects and interactions of the human DNA and have been studied using various NLP methods. For the purpose of identifying DNA methylation sites, a study [7] presented five Transformer-based methods: BERT, DistilBERT, ALBERT, XLNet, and ELECTRA. The Transformer-based models were trained using an unsupervised method on a dataset that included DNA sequences with information about methylation sites as well as a description of taxonomic lineage corresponding to various organisms. During the fine-tuning step, three methylation sites were used: 6mA, 4mC, and 5hmC. The final predictions of the five models listed above were calculated by averaging the probabilities returned by each model. One of the primary contributions of this work is that the authors demonstrated that a joint utilization of different language models can boost the classification performance of deep learning methods.

In another research paper [8], the authors presented an ALBERT-based architecture called LOGO (Language of Genome) that is pretrained on the unlabelled human reference genome and fine-tuned on downstream sequence labelling task. In this work, the authors contrasted their architectural design to DNABERT, which was pretrained on the unlabeled human reference genome and is likewise built on Transformer. The authors assert that while LOGO has about 1 million parameters and DNABERT has 100 million, LOGO exhibits significantly higher parameter efficiency. Additionally, according to the authors of this study, LOGO has a pretraining speed that is significantly faster than DNABERT's, which takes roughly 25 days.

Another research work [9] trained a model that can recognise DNA-protein bindings using an AWD-LSTM architecture. The model was fine-tuned using task-specific data from a labelled ChIP-seq data set after being trained on the unlabeled human reference genome (hg38). To preprocess the data, the authors used the k -mer based technique, which produces subsequences of length k within a biological sequence. The main drawback of this approach is that during the concatenation process, it creates non-existent subsequences that will never be equal to the original sequence. Moreover, the k -mer approach generates much longer sequences which can be resource intensive. A study [10] also developed a language model that predicts DNA-protein bindings. Instead of

using an AWD-LSTM language model in their model architecture as in [9], the authors used BERT and also applied the k -mer approach for preprocessing of the sequences. While the pretrained model in [9] was based on the unlabelled human reference genome (hg38), in this work, the authors used the 690 unlabeled ChIP-seq data. Similar to studies [9, 10], the authors [11] also selected a k -mer based approach for preprocessing the unlabelled human reference genome model that was later fine-tuned on task-specific data for the prediction of promoters as well as transcription factor binding sites (TFBS). As with [7], these authors also employed a ELECTRA language model as a base model in their architectural model design.

This section of the work demonstrated the use of multiple language models to address a variety of DNA-related problems (Table 1). In summary, the pretraining approaches often entail exposing the model to a sizable unlabelled corpus in an effort to create a general understanding that is transferable to different DNA tasks. The pretrained models were then routinely fine-tuned using labelled task-specific data for problems such as the prediction of promoter DNA sequences, TFBS, chromatin features, DNA-protein bindings, DNA methylation sites, and so on. Although each of these papers has significantly contributed to how language models can be modified for DNA tasks, this study specifically focuses on fine-tuning a Sentence Transformer model instead of a large language model for DNA tasks and shows that in certain settings, embeddings generated from a natural language-based transformer can outperform embeddings derived from large language models pretrained exclusively on genomic data.

Table 1 A summary of related work

Year and paper	Data	Prediction task(s)	Language model(s)	AUC/Acc
2022 [12]	eCLIP-seq	RNA-protein interactions	DNABERT	78.6
2022 [13]	DNA 6 mA dataset	DNA 6 mA sites	BERT	79.3%
2022 [14]	iDNA-MS, ENCODE data	DNA methylation	BERT	80+%
2022 [11]	GRCh38, EPDnew data, 690 ENCODE CHIP-Seq datasets	Promoter prediction and TFBS	ELECTRA	80-86%
2022 [9]	hg38, ChIP-seq datasets	DNA-protein bindings	AWD-LSTM	97-98%
2022 [8]	Hg19	Promoter prediction, Regulatory interactions between enhancer-promoter, Chromatin features	ALBERT	70+%
2023 [7]	DNA methylation dataset (iDNA-MS), Taxonomic lineage information from Taxonomy databases (NCBI & GTDB).	DNA methylation sites	ERT, DistilBERT	
			ALBERT, XLNet, ELECTRA	74-96%
2023 [10]	690 unlabeled ChIP-seq, Global dataset	DNA-protein binding	BERT	94.7%
2023 [15]	Human reference genome	Regulatory elements, Chromatin profiles		
		Species classification	Hyena operators	80+%

Methods

Sentence transformer

A sentence transformer [16] is a variant of the BERT model, which is based on transformers. These models enable the retrieval of sentence embeddings that possess semantic usefulness. Sentence Transformers were developed to overcome certain limitations introduced by BERT. For instance, in a dataset consisting of 10,000 sentences, identifying the pair with the greatest similarity necessitates performing 49,995,000 inference computations using BERT. On a contemporary V100 GPU, this task necessitates approximately 65 h.

The sentence transformer model applied in this work is SimCSE [17], which uses contrastive learning to generate superior sentence embeddings from both labelled and unlabeled data. There are two types of SimCSE architectures: the supervised SimCSE and the unsupervised SimCSE. With the unsupervised SimCSE, using randomly sampled English Wikipedia sentences, the model is trained to simply predict the input sentence itself by using dropout as noise. First, the input sentence is passed twice to the BERT/ RoBERTa encoder which results in two embeddings (*positive pairs*) with different dropout masks (z, z^i). Then, the other sentences in the same mini-batch are treated as *negative pairs*, and the model is trained to predict the positive one in the batch of negatives. Supervised SimCSE incorporates annotated sentence pairings in contrastive learning using natural language inference (NLI) datasets for sentence embeddings. Contradiction sentence pairings are viewed as negatives, while entailment sentence pairs are viewed as positives.

This study suggests a modification of the original SimCSE model by fine-tuning it on DNA (Fig. 1). To fine-tune the SimCSE model, a SimCSE checkpoint [18] was used and the model was trained on 3000 DNA sequences [19] that have been split into k -mer

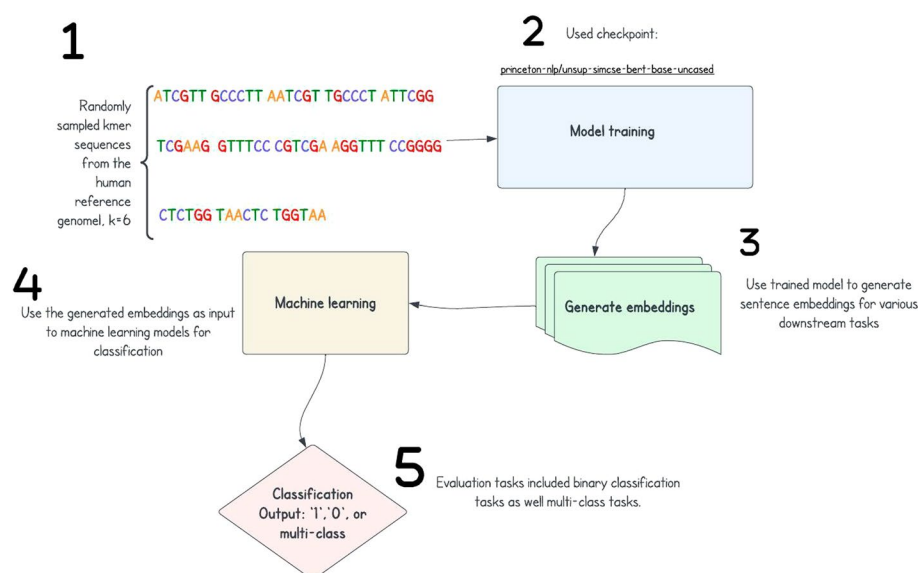


Fig. 1 The proposed model utilizes a pretrained checkpoint of the unsupervised SimCSE model from Hugging Face [18] with the following modified training script [20]. The model was trained on k -mer DNA sequences ($n = 3000$) sampled from the human reference genome for 1 epoch using a batch size of 16 and a maximum sequence length of 312. Next, the fine-tuned model was utilized to create sentence embeddings for DNA tasks. The final step involved using the generated sentence embeddings as input to machine learning algorithms for the classification of the eight DNA tasks described in Table 2

tokens of size 6 using the training scripts [20]. The model was trained for 1 epoch using a batch size of 16 and a maximum sequence length of 312. After training, the model was evaluated by generating sentence embeddings for eight classifications that are described below.

DNA-based embedding techniques

Nucleotide transformer

The nucleotide transformer (NT) [21] is a foundational transformer-based language model that uses Masked Language Modeling (MLM). This model learns to predict masked nucleotides represented as 6-mer tokens similar to BERT's training style and has been pretrained on unannotated genomic data. The model consists of four varying model sizes (500 million parameters, 2.5 billion parameters, and 50 million-500 million parameter sizes) constructed from different datasets such as the human reference genome, 3202 diverse human genomes, and 850 genomes of different species. The models were evaluated on eighteen genomic datasets which include splice site prediction tasks, promoter tasks, enhancer tasks, and etc. In this study, embeddings extracted from the proposed fine-tuned SimCSE model were compared against those derived from the NT model *InstaDeepAI/nucleotide-transformer-500 m-human-ref*. This particular variant was selected as it represents one of the smaller NT models, making it computationally feasible within the available resources for this work.

DNABERT

DNABERT [6] is a transformer model derived from BERT. Similar to the NT, DNABERT employs a MLM training objective to predict masked k -mer DNA tokens. DNABERT possesses various model versions, including $k = 3$, $k = 4$, $k = 5$, and $k = 6$. Each model is trained with a fixed k -mer size, resulting in distinct vocabularies and embeddings. DNABERT has undergone pretraining on the human reference genome, where sequences from the genome were segmented into overlapping k -mers and subsequently processed using the MLM objective function. DNABERT comprises 12 transformer layers and 12 attention heads. Analogous to NT, one of the evaluative tasks for the model involved the prediction of promoters, enhancers, and splice sites. DNABERT and NT both adjust parameter size, but NT explored much larger models. These two models are regarded as pioneers in DNA language modeling.

In this work, embeddings derived from the [CLS] token representation of DNABERT-6 were compared against those produced by the proposed fine-tuned SimCSE model.

Evaluation tasks

Task 1 (T1): APC gene, Task 8 (T8)TP53 gene: detection of colorectal cancer cases

The University of Pretoria EBIT Research Ethics Committee in South Africa (EBIT/139/2020) approved the analysis of FASTA exon DNA sequences from the *APC* and *TP53* genes that were collected from 95 colorectal cancer patients and matched-normal samples from previous work [22]. The samples were taken from a biobank of fresh tumor tissue and blood and collected with patient informed consent. This was approved by the South Eastern Sydney Local Health District Human Research Ethics Committee (approval number H00/022 and 00113), and all participants provided written informed consent. For pre-processing the sequences in preparation for machine learning, a python

script was created that removed IDs from the FASTA files, as well as divided the DNA strings into k -mers tokens of size 6. The target class consisted of a binary value: “1” for colorectal cancer cases, and “0” for colorectal cancer controls.

Task 2 (T2): The prediction of the Gleason grade group

The EBIT Research Ethics Committee at the University of Pretoria, South Africa, granted ethical approval (Ethics Reference No: 43/ 2010; 11 August 2020) for the utilization of blood *BRCA 1* DNA sequences from twelve patients with histopathological ISUP-Grade Group of 1 (representing low-risk prostate cancer) and 5 (representing high-risk prostate cancer). The patients were enrolled and provided consent in accordance with the approval obtained from the University of Pretoria Faculty of Health Sciences Research Ethics Committee (43/2010) in South Africa and the DNA sequencing was conducted with approval from the St. Vincent’s Hospital Human Research Ethics Committee (HREC) SVH/15/227 in Sydney, Australia. For data pre-processing such as alignment and conversion of FASTQ to FASTA files, the BWA-MEM aligner and samtools were used. Next, an in-house python script was used for removing sequence IDs, splitting the DNA sequences into k -mers of size 6, and converting the DNA sequences.

Task 3 (T3): Detection of human TATA

300-bp-long TATA and non-TATA DNA sequences were acquired from the source [23]. Next, the long DNA sequences were partitioned into k -mer tokens of size 6. For the protein-based model, an in-house python script was employed to translate DNA sequences into protein sequences which were then split into non-overlapping tokens of size 3. The independent variable employed in this study was the ‘seq_ori’, whereas the dependent variable, represented as a binary value, was ‘TATA’.

NT genomic benchmark datasets

The following datasets were adopted from the NT revised genomic benchmark dataset [24]:

- *Enhancers Task 4 (T4)*: this dataset consist of human enhancer elements that were downloaded from ENCODE’s SCREEN database. Distal and proximal enhancers were combined and enhancers were divided into tissue-specific and tissue-invariant.
- *Promoter_all Task 5 (T5)*: this dataset consists of promoter and non-promoter DNA sequences. The human promoter sequences were downloaded from the Eukaryotic Promoter Database, spanning 49kb upstream and 10bp downstream of transcriptions start sites. This dataset resulted in promoter regions which were TATA-Box promoters (positive examples) and sequences not overlapping promoters as negative examples.
- *H3K4me3 Task 6 (T6)*: this dataset comprised H3K4Me3 histone mark ChIP-seq data obtained from the ENCODE project. Genomic sequences of 1 kb in length that overlapped with identified peaks were designated as positive examples, whereas 1 kb sequences without peak overlap were designated as negative examples.
- *Splice_sites_all Task 7 (T7)*: this a three-class classification task including donor sites, acceptor sites, and non-splice sites. Each sample is a 400-nucleotide sequence centered on one of these positions, with balanced distributions across the classes.

Table 2 Dataset description

Task	Total sequences(n)	Number of labels
T1	15,000	2
T2	15,000	2
T3	564	2
T4	30,000	2
T5	30,000	2
T6	17,468	2
T7	30,000	3
T8	30,000	2

Table 3 A brief description of the machine-learning algorithms that were applied to evaluate the usefulness of the generated sentence embeddings

Algorithm description	
Logistic regression (LR)	Logistic regression is a statistical model that predicts binary classification tasks by modeling the log-odds of the event as a linear combination of predictor variables. LR applies the sigmoid function to confine predictions between 0 and 1 [25]
LightGBM (LGBM)	LightGBM is a gradient boosting ensemble technique that is based on decision trees. As with other decision tree-based methods, LightGBM can be applied to both classification and regression problems [26]
Random Forest (RF)	Random forest is a machine learning algorithm that consolidates the output of multiple decision trees to generate a single outcome. Its widespread use is motivated by its adaptability and usability and can solve both classification and regression issues [27]
XGBoost (XGB)	Extreme Gradient Boosting, popularly referred to as XGBoost, is a distributed, scalable gradient-boosted decision tree that offers parallel tree boosting. It is classified as a top machine-learning library for regression, classification, and ranking tasks [28]

The dataset was constructed from human genomic sequences. Table 2 provides the data description of each dataset.

Machine learning classifiers

After obtaining the embedding representations from the three embedding methods listed above, they were applied as input to machine learning classifiers which are described (Table 3).

The machine learning algorithms were trained using stratified *K*-fold cross-validation (*k* = 5), and the reported accuracy and F1 scores represent the mean values across folds along with their corresponding 95% confidence intervals.

Embedding extraction time

The computational efficiency of each model was measured through embedding extraction time where a set of DNA sequences were tokenized and processed in batches of size 32 using the model’s tokenizer and forward pass. The pooled representation was extracted for each DNA sequence, and the total wall clock time required to encode the entire data set was recorded. This measure reflects the runtime cost of generating embeddings for downstream tasks. Lower time means faster embedding generation, which is important for scalability and real-time applications (Fig. 2).

Retrieval benchmark

The quality of the learned DNA sequence embeddings were assessed using a retrieval benchmark using FAISS [29]. All embeddings were L2-normalized and indexed with two approaches: (1) a flat index for exact nearest neighbor search, and (2) a Hierarchical

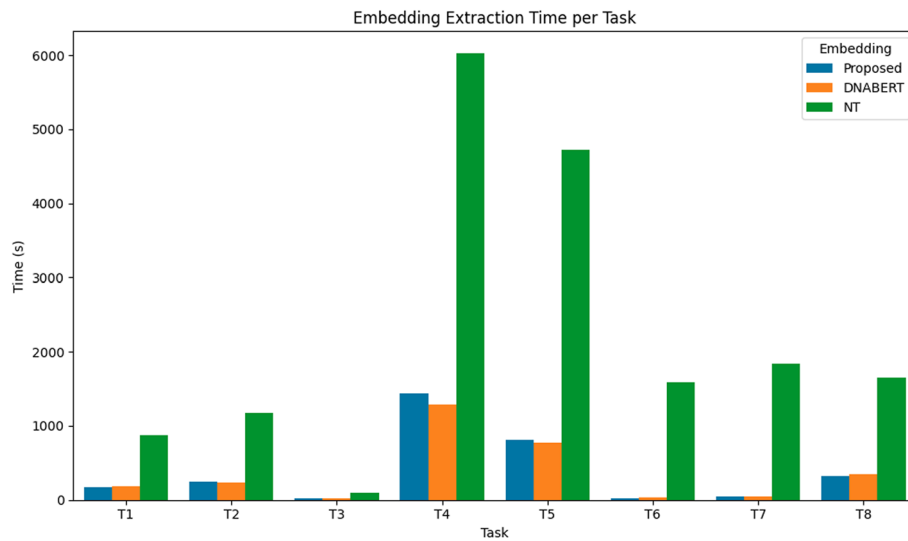


Fig. 2 Embedding extraction time by model

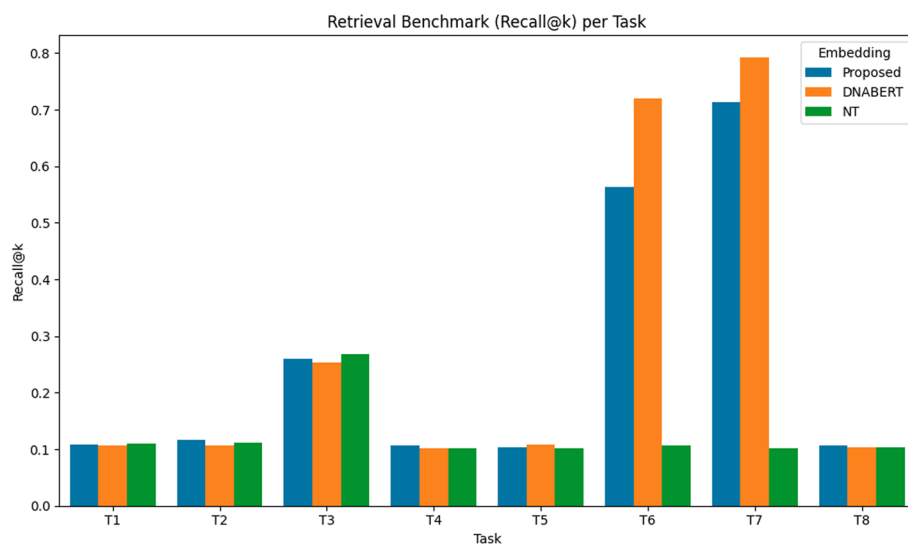


Fig. 3 Retrieval benchmark

Navigable Small World (HNSW) graph [30] for approximate nearest neighbor search. A subset of 100 embeddings were used as queries against the full index and for each query, the top-k ($k = 10$) most similar embeddings were retrieved. Retrieval performance was quantified using Recall@k, defined as the proportion of retrieved neighbors that belonged to the query set of 100. Higher recall means embeddings better capture semantic similarity, so neighbors are more relevant (Fig. 3).

Results

In this section, the classification results, embedding extraction time, and performance on retrieval tasks are presented (Tables 4, 5, Figs. 2, 3).

The first noticeable observation is that the embeddings of the proposed model perform reasonably better than the embeddings of DNABERT. Specifically, the embeddings of the proposed model surpassed DNABERT in T1 (highest average accuracy of 65%,

Table 4 Accuracy scores (with 95% confidence intervals) across datasets T1–T8 for each model and embedding method

Model	Embed.	T1	T2	T3	T4	T5	T6	T7	T8
LR	Proposed	<u>0.65 ± 0.01</u>	<u>0.67 ± 0.0</u>	<u>0.85 ± 0.01</u>	<u>0.64 ± 0.01</u>	<u>0.80 ± 0.0</u>	<u>0.49 ± 0.0</u>	<u>0.33 ± 0.0</u>	<u>0.70 ± 0.01</u>
	DNABERT	0.62 ± 0.01	0.65 ± 0.0	0.84 ± 0.04	0.69 ± 0.01	0.85 ± 0.01	0.49 ± 0.0	0.33 ± 0.0	0.60 ± 0.01
	NT	0.66 ± 0.0	0.67 ± 0.0	0.84 ± 0.01	0.73 ± 0.0	0.85 ± 0.01	0.81 ± 0.0	0.62 ± 0.01	0.99 ± 0.0
LGBM	Proposed	<u>0.64 ± 0.01</u>	<u>0.66 ± 0.0</u>	<u>0.90 ± 0.02</u>	<u>0.61 ± 0.01</u>	<u>0.78 ± 0.0</u>	<u>0.49 ± 0.0</u>	<u>0.33 ± 0.0</u>	<u>0.81 ± 0.01</u>
	DNABERT	0.62 ± 0.01	0.65 ± 0.01	0.90 ± 0.02	0.65 ± 0.01	0.83 ± 0.0	0.49 ± 0.0	0.33 ± 0.0	0.75 ± 0.01
	NT	0.63 ± 0.01	0.66 ± 0.0	0.91 ± 0.02	0.72 ± 0.0	0.85 ± 0.0	0.80 ± 0.0	0.59 ± 0.01	0.97 ± 0.0
XGB	Proposed	<u>0.60 ± 0.01</u>	<u>0.62 ± 0.0</u>	<u>0.90 ± 0.02</u>	<u>0.60 ± 0.0</u>	<u>0.77 ± 0.0</u>	<u>0.49 ± 0.0</u>	<u>0.33 ± 0.0</u>	<u>0.85 ± 0.01</u>
	DNABERT	0.59 ± 0.01	0.62 ± 0.01	0.90 ± 0.01	0.64 ± 0.01	0.82 ± 0.01	0.49 ± 0.0	0.33 ± 0.0	0.79 ± 0.01
	NT	0.61 ± 0.01	0.64 ± 0.0	0.90 ± 0.02	0.89 ± 0.03	0.85 ± 0.01	0.81 ± 0.01	0.60 ± 0.01	0.98 ± 0.0
RF	Proposed	<u>0.61 ± 0.0</u>	<u>0.66 ± 0.01</u>	<u>0.90 ± 0.02</u>	<u>0.61 ± 0.01</u>	<u>0.77 ± 0.0</u>	<u>0.49 ± 0.0</u>	<u>0.33 ± 0.0</u>	<u>0.86 ± 0.0</u>
	DNABERT	0.60 ± 0.0	0.66 ± 0.01	0.90 ± 0.02	0.63 ± 0.01	0.82 ± 0.0	0.49 ± 0.0	0.33 ± 0.0	0.81 ± 0.01
	NT	0.62 ± 0.01	0.67 ± 0.01	0.90 ± 0.01	0.71 ± 0.01	0.85 ± 0.0	0.79 ± 0.0	0.55 ± 0.01	0.97 ± 0.0

Best results per column are in bold, while scores for the Proposed model are underlined

Table 5 F1-scores (with 95% confidence intervals) across datasets T1–T10 for each model and embedding method

Model	Embed.	T1	T2	T3	T4	T5	T6	T7	T8
LR	Proposed	0.78 ± 0.0	0.80 ± 0.01	<u>0.20 ± 0.05</u>	<u>0.64 ± 0.01</u>	<u>0.79 ± 0.0</u>	<u>0.13 ± 0.0</u>	<u>0.16 ± 0.0</u>	<u>0.70 ± 0.01</u>
	DNABERT	0.75 ± 0.01	0.78 ± 0.0	0.47 ± 0.09	0.69 ± 0.01	0.84 ± 0.01	0.13 ± 0.0	0.16 ± 0.0	0.59 ± 0.01
	NT	0.56 ± 0.01	0.54 ± 0.0	0.78 ± 0.01	0.73 ± 0.0	0.85 ± 0.01	0.81 ± 0.0	0.62 ± 0.01	0.99 ± 0.0
LGBM	Proposed	<u>0.76 ± 0.01</u>	<u>0.79 ± 0.0</u>	<u>0.60 ± 0.11</u>	<u>0.63 ± 0.01</u>	<u>0.77 ± 0.0</u>	<u>0.47 ± 0.0</u>	<u>0.26 ± 0.0</u>	<u>0.82 ± 0.0</u>
	DNABERT	0.74 ± 0.0	0.78 ± 0.0	0.60 ± 0.08	0.66 ± 0.01	0.82 ± 0.01	0.47 ± 0.0	0.26 ± 0.0	0.75 ± 0.01
	NT	0.59 ± 0.01	0.56 ± 0.0	0.89 ± 0.02	0.72 ± 0.01	0.85 ± 0.0	0.80 ± 0.0	0.59 ± 0.01	0.97 ± 0.0
XGB	Proposed	<u>0.72 ± 0.01</u>	<u>0.75 ± 0.0</u>	<u>0.59 ± 0.08</u>	<u>0.60 ± 0.0</u>	<u>0.76 ± 0.0</u>	<u>0.47 ± 0.0</u>	<u>0.26 ± 0.0</u>	<u>0.85 ± 0.01</u>
	DNABERT	0.71 ± 0.01	0.75 ± 0.01	0.58 ± 0.05	0.64 ± 0.01	0.82 ± 0.01	0.47 ± 0.0	0.26 ± 0.0	0.79 ± 0.01
	NT	0.59 ± 0.01	0.57 ± 0.01	0.72 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.81 ± 0.01	0.60 ± 0.01	0.9893 ± 0.0
RF	Proposed	<u>0.73 ± 0.0</u>	<u>0.79 ± 0.0</u>	<u>0.58 ± 0.08</u>	<u>0.61 ± 0.01</u>	<u>0.75 ± 0.0</u>	<u>0.53 ± 0.0</u>	<u>0.24 ± 0.0</u>	<u>0.86 ± 0.0</u>
	DNABERT	0.72 ± 0.0	0.79 ± 0.0	0.59 ± 0.09	0.63 ± 0.01	0.80 ± 0.01	0.53 ± 0.0	0.24 ± 0.0	0.82 ± 0.01
	NT	0.59 ± 0.01	0.56 ± 0.01	0.89 ± 0.02	0.71 ± 0.01	0.84 ± 0.0	0.79 ± 0.0	0.55 ± 0.01	0.97 ± 0.0

Best results per column are in bold, while scores for the Proposed model are underlined

average F1 score of 78% from the LR classifier), T2 (highest average accuracy of 67%, average F1 score of 80% from the LR classifier), and T8 (highest average accuracy of 87%, average F1 score of 86% from the RF model). While the performance was the same from both models on tasks T3, T6, and T7. Embeddings of DNABERT only outperformed the embeddings of the proposed on two tasks only: T4 and T5. This result highlights a critical finding which indicate that embedding methods pretrained on natural language and fine-tuned on a small set of DNA sequences can, under certain conditions, match or even surpass the performance of encoders trained directly on DNA (such as DNABERT). This suggests that large-scale domain-specific pretraining is not always a prerequisite for competitive performance in genomics tasks.

When comparing embeddings from the nucleotide transformer (NT) against both DNABERT and the proposed model, NT embeddings consistently outperformed the others across all tasks T1–T8. However, this superior accuracy comes at a significant computational cost. As shown in Fig. 2, the NT model required the longest embedding extraction times, far exceeding both DNABERT and the proposed model. Furthermore, Fig. 3 demonstrates that NT embeddings performed worst on retrieval benchmark tasks, limiting their applicability for fast, large-scale genomic information retrieval.

This trade-off between accuracy and computational efficiency underscores the practical importance of the proposed model. While the NT achieves the highest accuracy overall, its computational demands make it less feasible for settings where resources are limited or rapid inference is required. By contrast, the proposed fine-tuned SimCSE-based transformer offers a substantially lower computational footprint, faster embedding extraction, and competitive accuracy relative to DNABERT. Although it does not surpass the NT in raw classification accuracy, it provides a balanced alternative that enables wider accessibility of embedding methods for genomic applications.

Discussion

BERT is a freely available language model based on Transformers. By establishing context through surrounding text, this language model is designed to assist computers in deciphering meaning in ambiguous text and emanates from several smart concepts that have recently surfaced in the NLP community, such as semi-supervised sequence learning (e.g., [31]), ELMo [32], ULMFiT [33] the OpenAI transformer [34], and the Transformer [35]. The BERT framework was pretrained using text from the English Wikipedia, and the Brown Corpus and can be fine-tuned on downstream tasks. Two distinct but related natural language processing tasks—MLM and Next Sentence Prediction (NSP)—were used to pre-train BERT. In the MLM task, BERT masks 15% of words in the input and asks the model to predict the missing word. Using NSP, given two sentences (A and B), the model is trained to predict the likelihood that sentence B belongs after sentence A. BERT can be used for single/pair sentence classification, sentiment analysis, question answering tasks, single sentence tagging tasks. Moreover, BERT can also be used to obtain contextualized sentence embeddings.

When it comes to analyzing the human DNA, an extension of BERT was introduced called DNABERT [6]. DNABERT was pretrained on the human reference genome as opposed to English Wikipedia and the Brown Corpus. DNABERT uses the same training procedure as BERT, but it does not include the NSP task found in the original BERT model. Other modifications found in DNABERT includes adjusting the sequence length

and forcing the model to predict contiguous k tokens adapting to DNA scenario. As with BERT, DNABERT can also be used to generate sentence embeddings as well as perform classification tasks at DNA-level.

As with DNABERT, the Nucleotide Transformer (NT) is built on the same underlying architecture as BERT, and has been adapted for genomic data. It uses a transformer encoder trained on massive amounts of DNA sequences with MLM, allowing it to capture long-range dependencies in genomic data. Compared to DNABERT, which uses k -mer tokenization and a BERT-base backbone, the NT scales up to larger model sizes and more diverse training corpora, making it more expressive but also far more computationally demanding. Although DNABERT and NT and similar DNA language models are useful in genomic research, finding the most similar sentence pair in these kinds of language models can require a lot of inference computations thus the introduction of Sentence Transformers.

In this study, we showed how an existing Sentence Transformer model, SimCSE, can be fine-tuned for DNA. By fine-tuning this model on a small sample size of the human reference genome, this work demonstrated that credible outcomes can still be obtained without extensive pretraining as with DNABERT and NT. This is especially crucial for LMICs, where access to high-performance computing equipment is generally limited. Large models like the NT require a lot of memory, storage, and computing power, which isn't often available in these kinds of settings. The proposed model demonstrates that effective genomic embeddings can be obtained without the prohibitive costs of training or inference on massive architectures. The proposed approach offers a more affordable, sustainable, and scalable solution for genomic research in resource-constrained environments.

Conclusion

This study fine-tuned a Sentence Transformer model derived from a natural language transformer—SimCSE. The results indicated that the embeddings derived from the fine-tuned model outperformed those of DNABERT across multiple tasks. Nonetheless, regarding raw classification accuracy, the nucleotide transformer model exceeded the proposed model in multiple tasks. However, this advantage resulted in substantial computing costs, making it unfeasible for resource-limited settings like LMICs. Furthermore, although the nucleotide transformer provided superior classification performance, it exhibited the poorest performance in retrieval tasks and embedding extraction duration.

Acknowledgements

The authors would like to thank the DAC for MCO colorectal cancer genomics at The University of New South Wales, for providing the data used in the study. The authors would also like to thank Prof. Jason Wong, for facilitating the data access requests and approvals. The authors also express gratitude to the South African Prostate Cancer Society (SAPCS) for sourcing and maintaining the data used in this work. The authors acknowledge the Centre for High Performance Computing (CHPC), South Africa, for providing computational resources to this research project.

Author contributions

M.M. devised the work and authored the primary draft. V.M. and D.M. jointly supervised and verified the findings of the experiments documented in the study. R.B. and V.M.H. offered their specialised guidance on the subject matter and provided one of the datasets used in this work.

Funding

Not applicable.

Data availability

The benchmark datasets can be accessed here [23, 24]. For the other tasks (T1 and T2), the data can be accessed at the host database (The European Genome-phenome Archive at the European Bioinformatics Institute, accession number: EGAD00001004582 Data access). We share the DNA-based model on Hugging Face [36].

Declarations

Ethics approval and consent to participate

T1 dataset: Ethics approval for T1 data was granted by the University of Pretoria EBIT Research Ethics Committee (EBIT/139/2020), together with the South Eastern Sydney Local Health District Human Research Ethics Committee (approval number H00/022 and 00113), and all participants provided written informed consent. T2 dataset: The EBIT Research Ethics Committee at the University of Pretoria, South Africa, granted ethical approval (Ethics Reference No: 43/2010; 11 August 2020) for the utilization of blood BRCA 1 DNA sequences from twelve patients with histopathological ISUP-Grade Group of 1 (representing low-risk prostate cancer) and 5 (representing high-risk prostate cancer). The patients were enrolled and provided consent in accordance with the approval obtained from the University of Pretoria Faculty of Health Sciences Research Ethics Committee (43/2010) in South Africa and the DNA sequencing was conducted with approval from the St. Vincent's Hospital Human Research Ethics Committee (HREC) SVH/15/227 in Sydney, Australia. T3 dataset: This dataset was obtained from ENCODE and other public genome annotation repositories. [23]. Repositories already handle the ethics approvals and informed consent at the stage of data collection. T4–T7 datasets: These datasets were drawn from public biology/genomics repositories (e.g. ENCODE, human genomes) [24]. Repositories already handle the ethics approvals and informed consent at the stage of data collection. Accordingly, all datasets used in this study were obtained and utilized in full compliance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interest

The authors declare no competing interest.

Author details

¹Department of Computer Science, University of Pretoria, Pretoria, South Africa

²NGEI, CSIR, Pretoria, South Africa

³AstraZeneca, London, UK

⁴Ancestry and Health Genomics Laboratory, Charles Perkins Centre, School of Medical Sciences, The University of Sydney, Sydney, Australia

⁵School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa

⁶Manchester Cancer Research Centre, University of Manchester, Manchester, United Kingdom

Received: 5 February 2024 / Accepted: 1 October 2025

Published online: 29 October 2025

References

1. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv.* 2021;56:1–40.
2. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. *Engineering.* 2022;25:51–65.
3. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding; 2018. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
4. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-xl: attentive language models beyond a fixed-length context; 2019. arXiv preprint [arXiv:1901.02860](https://arxiv.org/abs/1901.02860).
5. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. *Adv Neural Inf Process Syst.* 2019;32:64.
6. Ji Y, Zhou Z, Liu H, Davuluri RV. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics.* 2021;37(15):2112–20.
7. Zeng W, Gautam A, Huson DH. Mula-methyl-multiple transformer-based language models for accurate dna methylation prediction; 2023. *bioRxiv.* 2023:01.
8. Yang M, Huang L, Huang H, Tang H, Zhang N, Yang H, et al. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res.* 2022;50(14):81–81.
9. He Y, Zhang Q, Wang S, Chen Z, Cui Z, Guo Z-H, et al. Predicting the sequence specificities of dna-binding proteins by dna fine-tuned language model with decaying learning rates. *IEEE/ACM Trans Comput Biol Bioinf.* 2022;20(1):616–24.
10. Luo H, Shan W, Chen C, Ding P, Luo L. Improving language model of human genome for dna-protein binding prediction based on task-specific pre-training. *Interdiscip Sci Comput Life Sci.* 2023;15(1):32–43.
11. An W, Guo Y, Bian Y, Ma H, Yang J, Li C, Huang J. Modna: motif-oriented pre-training for dna language model. In: *Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics*, 2022. pp. 1–5.
12. Yamada K, Hamada M. Prediction of rna-protein interactions using a nucleotide language model. *Bioinform Adv.* 2022;2(1):023.
13. Le NQK, Ho Q-T. Deep transformers and convolutional neural network in identifying dna n6-methyladenine sites in cross-species genomes. *Methods.* 2022;204:199–206.
14. Jin J, Yu Y, Wang R, Zeng X, Pang C, Jiang Y, et al. Idna-abf: multi-scale deep biological language learning model for the interpretable prediction of dna methylations. *Genome Biol.* 2022;23(1):1–23.

15. Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, Patel A, Rabideau C, Massaroli S, Bengio Y, et al. Hyenadna: long-range genomic sequence modeling at single nucleotide resolution; 2023. arXiv preprint [arXiv:2306.15794](https://arxiv.org/abs/2306.15794).
16. Reimers N, Gurevych I. Sentence-bert: sentence embeddings using siamese bert-networks; 2019. arXiv preprint [arXiv:1908.10084](https://arxiv.org/abs/1908.10084).
17. Gao T, Yao X, Chen D. SimCSE: simple contrastive learning of sentence embeddings; 2022.
18. Princeton-nlp: model card for unsup-simcse-bert-base-uncasedl. <https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased>.
19. Ji Y, Zhou Z, Liu H, Davuluri RV. https://raw.githubusercontent.com/jerryji1993/DNABERT/master/examples/sample_data/pre/6_3k.txt
20. Princeton-nlp: SimCSE: simple contrastive learning of sentence embeddings. https://github.com/princeton-nlp/SimCSE/blob/main/run_unsup_example.sh.
21. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nat Methods*. 2025;22(2):287–97.
22. Poulos RC, Perera D, Packham D, Shah A, Janitz C, Pimanda JE, et al. Scarcity of recurrent regulatory driver mutations in colorectal cancer revealed by targeted deep sequencing. *JNCI Cancer Spectrum*. 2019;3(2):012.
23. Ji Y, Zhou Z, Liu H, Davuluri RV. Supplementary data. <https://academic.oup.com/bioinformatics/article/37/15/21112/6128680?login=false#supplementary-data>.
24. InstaDeepAI: nucleotide transformer downstream tasks (Revised). https://huggingface.co/datasets/InstaDeepAI/nucleotide_transformer_downstream_tasks_revised. Accessed 18 Aug 2025. https://huggingface.co/datasets/InstaDeepAI/nucleotide_transformer_downstream_tasks_revised.
25. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Stat Methodol*. 1958;20(2):215–32.
26. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inform Process Syst*. 2017;8:30.
27. Parmar A, Katariya R, Patel V. A review on random forest: an ensemble classifier. In: *International conference on intelligent data communication technologies and internet of things (ICICI)* 2018. Springer; 2019. pp. 758–763.
28. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining*; 2016. pp. 785–794.
29. Johnson J, Douze M, Jégou H. Billion-scale similarity search with gpus. In: *IEEE transactions on big data*. IEEE; 2019.
30. Malkov YA, Yashunin DA. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans Pattern Anal Mach Intell*. 2018;42(4):824–36.
31. Dai AM, Le QV. Semi-supervised sequence learning. *Adv Neural Inform Process Syst*. 2015;3:28.
32. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations; 2018.
33. Howard J, Ruder S. Universal language model fine-tuning for text classification; 2018.
34. Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training; 2018.
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inform Process Syst*. 2017;6:30.
36. <https://huggingface.co/dsfsi/simcse-dna>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.