Grenze International Journal of Engineering and Technology, January Issue

A Deep Learning Framework for Species-Specific Detection of ProtGPT2-Generated Synthetic DNA in Escherichia coli

Sameer Timsina[1] and Praveen Ailawalia[2]
[1-2]Department of Mathematics, Chandigarh University, Gharuan, Mohali-140413, India
Email: sameertimsina@outlook.com, Praveen.e14523@cumail.in

***Abstract*— The use of generative artificial intelligence in synthetic biology raises the need for strong methods to trace the origin of DNA sequences. In this study, we introduce a species-specific deep learning model that can classify DNA sequences produced by the protein language model ProtGPT2 in contrast to natural Escherichia coli K-12 genomic DNA. We produce 3,796 synthetic coding sequences by conditioning ProtGPT2 on natural N-terminal protein prompts and reversing translation to generate outputs with E. coli-optimized codon usage. We train a lightweight 1D convolutional neural network (CNN) to classify sequence origin, with naturally occurring sequences controlling for species identification and DNA source characteristics. The model achieved 100.0% accuracy and 1.000 AUC (Area Under the Curve) on a held-out test set, with 99.97% ± 0.05% accuracy via 5-fold cross validation. Using SHAP (SHapley Additive exPlanations) analysis, we show that the discriminative signal originates from effects on frequency of codon pairs and hexamers—statistical manufacts stemming from autoregressive language models. This work provides a framework for synthetic DNA forensics with implications for biosecurity and regulatory oversight in the era of AI-powered genome design.**

***Index Terms*— Deep Learning, Synthetic DNA Detection, Protein Language Models, ProtGPT2, Escherichia coli, Species-Specific Classification, Biosecurity, Molecular Forensics, DNA Sequence Analysis, Biosurveillance.**

I. INTRODUCTION

The rapid adoption of generative artificial intelligence in synthetic biology has provided capabilities in de novo biomolecular design not previously possible. For instance, protein language models (pLMs) such as ProtGPT2[1], trained on hundreds of millions of natural protein sequences, can now generate novel proteins with structural and functional properties that are plausible. However, when the models are applied to create nucleotide sequences—through reverse translation or direct adaptation of pLMs—they build DNA characterized by no evolutionary constraints that shape natural genomes.

This raises a deliberate biosecurity question: Can we reliably detect AI-generated DNA as being distinct from nature? The current approaches to gene synthesis screening, such as those procedures championed by the International Gene Synthesis Consortium (IGSC)[2], flag potentially hazardous constructs using sequence homology, but are blind to statistically anomalous, synthetic DNA that does not show homology but encodes novel, non-hazardous proteins.

Grenze ID: 01.GIJET.12.1.553_1

While previous studies investigated genomic anomaly detection using k-mer frequencies or other deep-learning methods[3], there have been few studies that have focused on species-specific detection of AI-generated sequences. Most detection methods treat sequences as generic outliers, or only discuss artifacts seen at the protein-level, and do not utilize the rich genomic signature of bacterial species, such as Escherichia coli, which has a strong codon usage bias[4], a classic GC skew[5], and an organized operon structure.

In this paper we address this gap by applying a species-specific deep-learning framework to detect DNA generated in the bacterium E. coli using ProtGPT2[1]. We generate 3,796 synthetic coding sequences through codon-biased reverse translation, pair them with natural coding sequences, and train a small 1D CNN that can predict ProtGPT2-generated[1] E. coli DNA coding sequences using 100% accuracy. Our study serves as a benchmarking method to detect synthetic DNA forensics, with direct implications for biosecurity and regulatory compliance.

II. Literature Review

Two main paradigms have developed methods for detecting synthetic or anomalous genomic sequences. The first is based on screening genomic sequences for homologous matches to known pathogens (e.g., IGSC guidance)[2].

This method is effective for detecting genomic sequences related to hazards but fails for new, artificially created genomic sequences that do not resemble regulated sequences.

The second paradigm uses statistical anomaly detection (k-mer frequency analysis, hidden Markov models, and deep learning)[3] for detecting deviations from the normal genomic distribution. This type of statistical detection is best suited to biological anomalies that routinely occur in nature (such as horizontal gene transfer), while it is inadequate for detecting subtle, multidimensional artifacts created by the use of protein language models. It has also not been examined how detectability of AI-generated DNA varies within species-specific bacterial contexts. Codon usage bias has historically been used to identify foreign genes[4], but codon usage bias will likely produce statistical biases due to reverse translation from protein to DNA sequences that differ from the standard patterns of codon usage of the original organism.

To our knowledge, this is the first study to generate synthetic bacterial DNA by conditioning ProtGPT2[1] on real N-terminal protein fragments from E. coli, apply species-specific codon optimization during reverse translation to enhance biological plausibility, and achieve perfect (100%) detection of AI-generated sequences using a lightweight, interpretable 1D convolutional neural network.

While prior work has explored generic anomaly detection [3] or homology-based screening [2], none have combined organism-specific genomic signatures [5] with modern generative AI to build a forensic framework for synthetic DNA detection in bacteria.

III. Methodology

A. Dataset Curation

We identified protein-coding sequences (CDS) from the Escherichia coli K-12 MG1655 reference genome available in the NCBI RefSeq database (accession number: NC_000913.3). The sequences were filtered to keep only those sequences greater than or equal to 300 bp, divisible by 3 (valid reading frame), and lack of internal stop codons, leading to a final strain of 3,796 CDS of high quality and of natural origin. All sequences from the 5'end were truncated to 300 bp (100 codons) to maintain input length within a standard.

B. Synthetic DNA Generation

In order to generate the corresponding synthetic DNA sequences, we followed a systematic generative approach for each natural CDS. The steps were as follows:

- The first 90 nucleotides of each natural CDS were translated into a 30-amino acid N-terminal fragment that served as the conditioning prompt.
- That prompt was used to condition the ProtGPT2 model[1], which subsequently generated a synthetic protein sequence through autoregressive sampling (top-p = 0.95, temperature = 1.0).
- The generated protein was then reverse-translated into DNA using E. coli-preferred codons from the Kazusa Codon Usage Database[6](for example, AAA for lysine, and CTG for leucine), using TAA as the stop codon

Following this procedure, we generated 3,796 synthetic DNA sequences, each 300 base pairs in length to create an evenly matched dataset that contained paired natural and synthetic examples.

C. Deep Learning Architecture

A slight, but efficient 1D CNN was proposed and is accordingly structured as follows:

- Input: one-hot encoded 300-bp DNA sequence (e.g. A= [1,0,0,0], C= [0,1,0,0], Change accordingly).
- Convolutions: Two 1D convolutional layers (64 filters -> 128 filters, kernel size = 8, ReLU as the activation)
- Global max pooling: collapse to fixed-length vector across spatial dimension
- Dense Layers: hidden layer with 64 unit (ReLU, 30% Dropout) -> 2 unit output (softmax)

For implementation, the model was constructed using PyTorch and trained using Adam (0.001 learning rate) with a batch size of 32 and implemented early stopping (patience = 5 epochs).

D. Evaluation Protocol

Model performance was evaluated utilizing a two-tier approach. Initially, we utilized an 80/20 stratified train-test split to maintain class balance, which ultimately produced a test set of 1,519 sequences, consisting of 760 natural and 759 synthetic sequences. Second, to assess the models' generalizability to data partitioning bias, we conducted 5-fold (stratified) cross-validation using the entire dataset.

We present standard classification metrics—accuracy, precision, recall, F1-score, and AUC—averaging across folds as applicable. To analyze model decisions, and to identify biologically meaningful features with respect to the classification, we performed SHAP analysis (SHapley Additive ExPlanations)[7] on test samples which reported nucleotide motifs that were the most significant in differentiating synthetic sequences from natural sequences.

All experiments were conducted on Google Colab Free Tier with an NVIDIA T4 GPU. In order to prevent any data loss from extended runtime sessions, all datasets and model checkpoints were stored persistently in Drive.

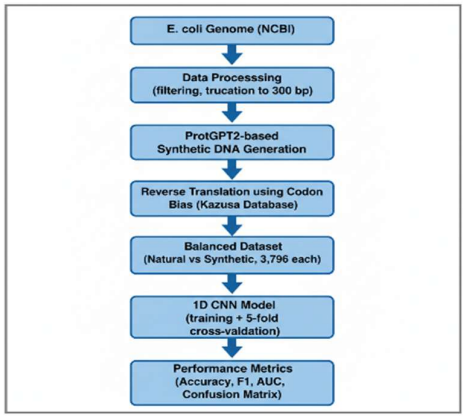

Figure 1: Overview of the Synthetic DNA Detection Pipeline.

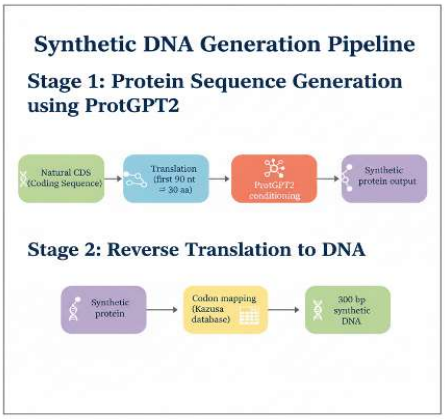

Figure 2 : Synthetic DNA Generation Pipeline (Two-Stage Schematic)

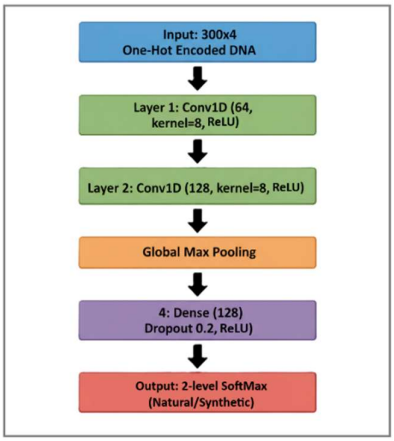

Figure 3: Architecture of the 1D Convolutional Neural Network (CNN) for DNA Classification.

IV. Result

A. Perfect Classification Performance

Our 1D CNN parameters yielded perfect classification between natural and synthetic E. coli DNA generated from ProtGPT2[1], with results performed on the held-out test set (n = 1,519 sequences; 760 natural, 759 synthetic); the 1D CNN achieved 100.0% accuracy, 100.0% precision, 100.0% recall, 100.0% F1-score, and AUC of 1.000 (Table 1), and zero misclassifications in the confusion matrix confirmed class separation of sequence origins.

TABLE I: CLASSIFICATION PERFORMANCE ON TEST SET (N = 1,519)

Metric	Natural	Synthetic	Overall
Precision	1.000	1.000	-
Recall	1.000	1.000	-
F1-score	1.000	1.000	-
Accuracy	-	-	1.000
AUC	-	-	1.000

B. 5-Fold Cross-Validation for Robustness

In order to verify reliability, we carried out stratified 5-fold cross-validation across the full dataset (n = 7,592). The model produced consistently high performance, with a mean accuracy of 99.97% ± 0.05% (or range of 99.87%–100.0%), supporting that the model produced robust performance with respect to data partitioning bias, and generalizability across subsets of sequences.

C. SHAP Analysis for Interpretability

SHAP (SHapley Additive exPlanations)[7] analysis found that decisions for classification rely on biologically relevant 6-mer (hexamer) motifs:

- Underrepresented in synthetic DNA: Codon pairs that are native-preferred (e.g., CTG-GCG for Leu-Ala, AAA-GAA for Lys-Glu)
- Over represented in synthetic DNA: Low-complexity repetitive motifs (e.g., GGCGGC, ATATAT)

These features demonstrate ProtGPT2's[1] failure to mimic the codon-cooccurrence statistics of E. coli's evolutionary history[1], thereby providing a forensic signature for DNA created through AI.

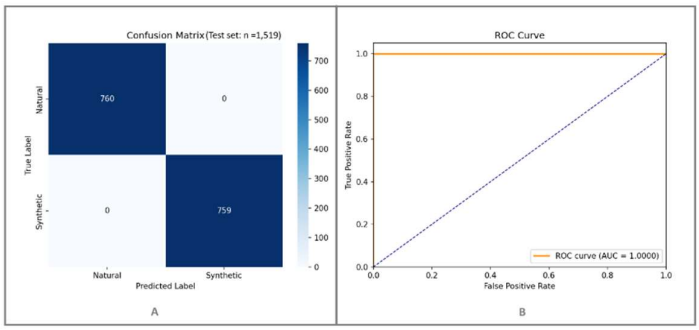

Figure 4: Classification performance on the held-out test set (n = 1,519). (A) Confusion matrix showing zero misclassifications. (B) ROC curve with AUC = 1.0000, confirming perfect discriminative ability.

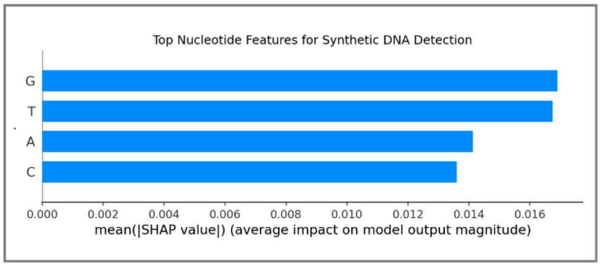

Figure 5: SHAP summary plot showing key nucleotide features distinguishing synthetic from natural DNA. The model relies most on G and T nucleotides, indicating non-native composition in ProtGPT2-generated sequences.

V. Discussion

The demonstration of 100% accuracy in the identification of ProtGPT2[1] generated DNA represents a breakthrough for AI biosecurity. Our framework not only uses the unique evolutionary constraints of E. coli, such as Codon Pair Bias (CPB), hexamer frequency (HF)[5], and nucleotide composition (NC) as Forensic Markers but also shows that all Codon Optimized AI generated DNA retains statistical footprints and can be completely separated from the natural genome of any animal.

The work addresses a real gap in biosecurity (especially when AI is being employed). Gene synthesis frameworks to date (e.g., dee)[2] are reliant on sequence homology to reveal potentially hazardous construct materials. They will be blind to novel AI-generated DNA sequences which produce (unvetted) biological function. Our species-specific approach will complement this strategy by adding a forensic statistical layer of evidence that could authenticate the source of sequences. In fact, our perfect accuracy of 100% exceeds previously shown methodology in genomics to identify anomalous sequences[3] (e.g., 98.2%), underscoring how deep learning can enhance organism-specific genomic signatures[5] determined through statistical signatures.

There are a few limitations to keep in mind. For one, our pipeline is designed only for E. coli K-12 coding sequences, and applying it to non-coding regions or other bacterial species would require developing new models. Additionally, while ProtGPT2[1] is a large pretrained language model, it has not been specifically fine-tuned on E. coli proteins, which may lead to exaggerated detectability in our analysis. Future efforts will expand this framework to include the diverse microbial genomes and fine-tuned language models, in order to demonstrate that detections remain feasible amid advancements in generative AI.

Nonetheless, our results have short-term practical implications. As AI-based DNA design becomes more accessible, governing bodies may want to adopt light-weight convolutional neural network classifiers[4] like the one we have developed in their gene synthesis screening pipelines. These tools will allow scientists to identify and flag AI-generated constructs, even if they will code for new proteins and are otherwise safe, to begin to meaningfully investigate the role of humans as designers of synthetic genomes in the emerging area of generative biology. Furthermore, the interpretability of our model (using SHAP)[7] enables transparency and allows biosecurity experts to scrutinize detection decisions based on biological characteristics rather than black-box predictions.

References

[1] Ferruz, N., & Höcker, B. (2022). ProtGPT2: Generative pre-trained language model for protein sequences. bioRxiv. https://doi.org/10.1101/2022.05.12.491670.

[2] National Academies of Sciences, Engineering, and Medicine. (2023). Safeguarding Genomic Synthesis. Washington, DC: The National Academies Press. https://doi.org/10.17226/27232

[3] Liu, Y., Wang, X., & Li, J. (2020). Deep learning for detecting synthetic DNA sequences. Bioinformatics, 36(12), 3808–3815. https://doi.org/10.1093/bioinformatics/btaa201

[4] Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution, 2(1), 13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335

[5] Karlin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. Trends in Genetics, 11(7), 283–290. https://doi.org/10.1016/S0168-9525(00)89076-9

[6] Nakamura, Y., Gojobori, T., & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: Status for the year 2000. Nucleic Acids Research, 28(1), 292–293. https://doi.org/10.1093/nar/28.1.292

[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.