Membership Inference Attacks on Embedding Models

Huong Tran
University of North Carolina at Charlotte
Charlotte, NC, USA
htran55@charlotte.edu

Liyue Fan
University of North Carolina at Charlotte
Charlotte, NC, USA
liyue.fan@charlotte.edu

ABSTRACT

Machine learning models trained on sensitive genomic data can inadvertently leak information about their training sequences. In this work, we investigate membership inference attacks (MIAs) against DNABERT-2, a transformer-based embedding model for DNA sequences. We conducted three complementary attack pipelines: (1) a sliding-window method that computes local cosine-similarity profiles over fixed-size token contexts; (2) a sentence-level segmentation approach that aggregates similarities between adjacent non-overlapping windows; and (3) a whole-sequence embedding comparison using both mean and max pooling strategies across large-scale perturbation studies. We evaluate these attacks on sets of up to 5000 member and 100,000 perturbed non-member sequences under various mutation rates. Our results showed that higher mutation rates yield clearer separation between member and non-member similarity distributions, with max-pooled embeddings providing the greater sensitivity to small sequence changes. T-SNE visualizations further reveal tight clustering of non-member representations versus the broader spread of member points. Finally, we discussed the trade-offs between context size and score variance, and outline future work on learned similarity metrics and cross-model evaluations with alternative genomic transformers. These findings underscored the privacy risks inherent in high-capacity genomic embedding models and motivated the development of robust defenses..

CCS Concepts

• **Security & Privacy → Privacy preserving protocols**
• **Security & Privacy → Security in Machine Learning**
• **Applied Computing → Bioinformatics.**

Keywords

Membership inference attacks; DNABERT-2; embeddings, transformer model; privacy leakage; cosine similarity; sliding window; sentence-level.

1. INTRODUCTION

Membership Inference Attacks (MIAs) have emerged as a critical concern in evaluating the privacy risks associated with machine learning models, particularly with respect to training data leakage. These attacks were designed to determine whether specific data instances were used during model training, posing significant privacy risks for sensitive datasets. While previous research primarily concentrated on supervised learning scenarios, recent advancements in embedding models necessitate exploration into privacy risks associated with embeddings, especially in bioinformatics contexts where genomic data sensitivity is paramount.

In this study, we implemented and evaluated membership inference attacks targeting DNABert-2, a foundational embedding model trained specifically for genomic sequences. We investigated two primary attack methodologies: word-level and sentence-level embedding MIAs. To generate realistic non-member data, we sample sequences from the DNABert-2 pre-training dataset and systematically perturb nucleotides within these sequences at varying mutation rates. By adjusting perturbation frequency, we ensure generated sequences do not overlap with training data, thus providing a robust baseline for evaluating membership inference.

For word-level embeddings, we compute similarity scores across sliding windows within both member and perturbed non-member sequences, analyzing statistical patterns such as mean and standard deviation. Conversely, for sentence-level embeddings, we segment genomic sequences into non-overlapping windows, deriving sentence embeddings through mean pooling. Similarities between adjacent embeddings are then computed, averaged, and analyzed to infer membership status. These approaches extend and adapt methodologies discussed in prior embedding-based privacy research, notably in the context of textual data.

Our work provides valuable insights into the susceptibility of genomic embedding models like DNABert-2 to privacy breaches through membership inference. This investigation not only underscores the critical necessity of privacy-preserving strategies in bioinformatics but also contributes practical methodologies and benchmarks to guide future research and policy development in embedding model security.

2. BACKGROUND

Embedding models map discrete inputs—such as words or genomic k-mers—to continuous vector spaces, preserving semantic or biological context while enabling downstream tasks like classification or retrieval. However, embedding vectors may also leak sensitive information about their inputs. Privacy of machine learning models encompasses several dimensions, including inference-time input privacy (e.g., inverting embeddings) and training-data privacy, which addresses what information about the training set is exposed during or after training [1].

Membership inference attacks evaluate training-data privacy by determining whether a given input was part of the model’s training set. Classic MIAs exploit models’ tendency to assign higher confidence or lower loss to training members compared to non-members [2]. In embedding models, exact loss computation is often infeasible due to sampling-based training objectives; thus, prior work proposes thresholding attacks based on embedding similarity scores rather than loss values. As described in Section 5 of Song and Raghunathan [1], "To decide the membership for a window of words C, … the adversary computes a set of similarity scores Δ and uses the averaged score as the decision metric."

Section 5.1 of the same work extends this attack to sliding windows of word embeddings, computing $\Delta = \{\delta(v_{w_0}, v_{w_i}\}|\forall\, w_i\ \in C\{w_0\}\}$ where δ is cosine similarity, and classifying as member if the mean exceeds a threshold. Section 5.2 further generalizes the attack to sentence embeddings, using pairwise similarities $\delta = (\Phi(x_a), \Phi(x_b))$ to infer membership of sentence pairs and aggregated contexts. This foundational methodology informs our adaptation to genomic sequence embeddings, motivating our strategy to leverage similarity-based thresholds in both word- and sentence-level MIAs on DNABERT-2 embeddings.

Several studies have explored membership inference beyond simple thresholding. Shokri et al. [3] introduced shadow-model techniques to train adversarial classifiers on auxiliary data, improving attack accuracy but requiring substantial adversarial resources. More recent works investigate reference-free attacks that compare model outputs to synthetic neighbors [4] and adapt to privacy defenses such as differential privacy [5]. We reference these methods to contextualize our contributions: by focusing on similarity-based MIAs without auxiliary models, we prioritize attack efficiency and applicability to large embedding models where resource constraints preclude complex adversarial training.

3. METHOD

Our methodology implements two similarity-based membership inference attacks adapted to DNABERT-2 embeddings: one targeting word-level (k-mer) embeddings and one targeting sentence-level (windowed) embeddings. We built upon Song and Raghunathan's thresholding attack framework [1] while tailoring data preprocessing, perturbation strategies, and statistical analyses for genomic data.

3.1 Word-Level Embedding Attack

3.1.1 Data Preparation

We conduct two complementary experiments to evaluate membership inference under varying dataset scales and perturbation regimes:

(a) 1000-sequence experiment: We sampled N_1 = 1000 member sequences from DNABERT-2's pretraining corpus. For each sequence s, we generated a perturbed non-member sequence s' by mutating a fraction f ∈ {0.1, 0.2, …,0.5} of nucleotide positions, selecting replacement nucleotides uniformly from the three alternatives. Perturbation repeated until s' is not in training set, ensuring one-to-one correspondence between member and non-member sequences.

(b) 5000-sequence robustness experiment: To examine embedding robustness, we sampled N_2 = 5000 member sequences and applied smaller perturbation rates f ∈ {0.002, 0.005, 0.01, 0.02, …, 0.1, 0.2, …, 0.5}. For each original sequence s, we derived one corresponding perturbed sequence s' per f, verifying non-membership as above. We then obtain sequence embeddings via mean pooling over token vectors for both s and s', and compute cosine similarity $\delta(\Phi(s_i),\ \Phi(s'_i)) = \frac{\Phi(s_i) * \Phi(s\prime_i)}{\|\Phi(s_i)\|\ \|\Phi(s\prime_i)\|}$.

For each perturbation rate, this procedure produces N_2 = 5000 similarity scores. We recorded results in the format (train_index, similarity_f) and plot kernel density estimates (KDEs) of these distributions, expecting similarity curves to shift toward lower values as f increases.[1]

3.1.2 Word-Level Embeddings

First, each sequence was tokenized using DNABERT-2's AutoTokenizer, and token embeddings were retrieved directly from the model's pretrained embedding layer. We then chose a fixed window length and slided this window one token at a time along the sequence, ensuring that each window has a well-defined "center" token. Within each window, we computed cosine similarities between the embedding of the center token and the embeddings of all other tokens in that window. These similarities were averaged to yield a single score for that window. Repeating this process for every valid window position produced a vector of average-similarity values for the entire sequence.

Our extractor operated exclusively on token embeddings rather than deeper hidden-state representations, substantially reducing memory usage and simplifying computation. We experimented with multiple window lengths $l = \{5, 15, 25, 35, 45\}$ and observed the results through KDE plots. We also identified common token windows in both member and non-member sets and compared their average similarity scores to quantify discriminative analysis.

3.1.3 Sentence-Level Embeddings

To extend membership inference beyond local token contexts, we treated each DNA sequence as a "paragraph" and generated a series of "sentences" by partitioning its token list into non-overlapping windows of at least 25 tokens. Concretely, for a sequence of length L, we slide a fixed-length frame in strides of 125 tokens, yielding $L/125$ segments. Each segment's token indices were decoded back into an A/C/G/T string and re-tokenized for input to DNABERT-2. We then obtained a sentence embedding for each segment via mean pooling over its token embeddings.

Next, we computed cosine similarities between every pair of adjacent sentence embeddings—i.e., between segment 1 and 2, segment 2 and 3, and so on—and average these similarity scores to produce a single per-sequence metric. The resulting table records, for each original sequence, which serves as the decision statistic for membership inference.

As a baseline, we also computed a "whole-sequence" embedding by mean pooling over all tokens in the original sequence (without segmentation) and measured cosine similarity between member and perturbed non-member embeddings. This comparison

[1] The DNABERT-2 implementation used in this study is publicly available on GitHub https://github.com/MAGICS-LAB/DNABERT_2.

revealed whether preserving broader context yields stronger membership signals than the segmented approach.

These sentence-level methods followed the aggregate-level attack paradigm described by Zheng et al. [1], demonstrating how embedding similarity at coarser granularity can leak training membership.

4. RESULTS

This section presents the outcomes of our sliding-window and sentence-level membership inference experiments. We begin by summarizing the datasets, parameter settings, computation platforms, and software libraries used (Section 4.1), then report detailed findings for each attack variant (Sections 4.2–4.3).

4.1 Experimental Setup

Datasets:

- Sliding-window MIA: 1000 member sequences and 1000 non-member sequences obtained from perturbations at two rates (10 % and 50 %) as described in Section 3.1.
- Sentence-level MIA: 1000 original "paragraph" sequences, each paired with one perturbed sequence at f ∈ {0.1, 0.5}; additionally, a 5000-sequence set evaluated at finer perturbation rates f ∈ {0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2, 0.5}.

Parameter Settings:

- Sliding windows: lengths of 5, 15, 25, 35, and 45 tokens.
- Sentence windows: fixed block size of 125 tokens; whole-sequence baseline also evaluated.
- Similarity metric: cosine similarity on DNABERT-2 (mean-pooled and max-pooled embeddings only in the sentence-level embeddings).

Platforms:

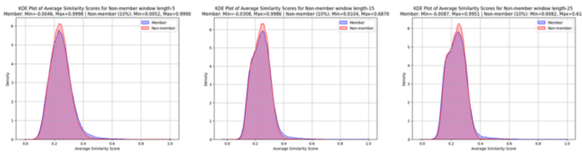
- Ubuntu 20.04 server
- Python 3.9, PyTorch 1.11, Transformers 4.19, scikit-learn 1.0.

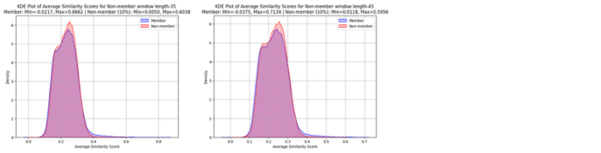
Libraries & Tools:

- Tokenization & Embeddings: Hugging Face Transformers (zhihan1996/DNABERT-2-117M)
- Similarity Computation: sklearn.metrics.pairwise.cosine_similarity
- Data Handling: pandas, NumPy
- Visualization: Matplotlib

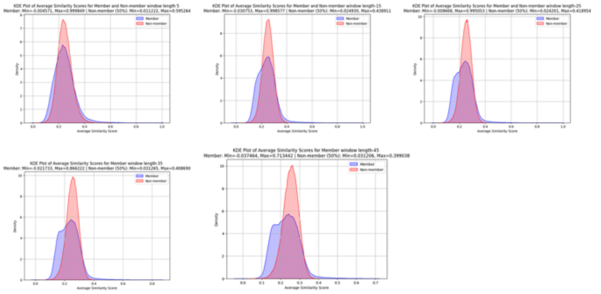
4.2 Word-Level Attacks Results

We first assessed how well local similarity profiles distinguish member from non-member windows. The bigger window length is, the wider distribution spread outs. The results mostly overlapped at f = 0.1, while it tends to separate more at f = 0.5. Figure 1 shows KDE plots for similarity scores of different window embeddings at two perturbation rates.





a) f = 0.1



b) f = 0.5

Figure 1: KDE Plots for window lengths of $l \in \{5, 15, 25, 35, 45\}$ at f ∈ {0.1, 0.5}

4.3 Sentence-Level Attack Results

Next, we evaluated membership inference at the sentence (segment) granularity at two perturbation rates. From observation KDE plots, the results are similar between member and non-member at both f values. However, with the t-SNE visualization, embeddings of member and non-member seems to distribute in different ways. While member embeddings (blue dots) still spread out, non-member ones start to compact into certain spots. Figure 2 presents KDE plots for cosine similarity scores of sentence embeddings for 1000 sampled datasets at two perturbation rates. Figure 3 displays T-SNE visualization of mean pooling between 1000-member dataset and 1000-non-member datasets at two perturbation rates

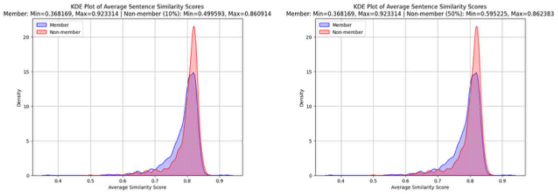


Figure 2: KDE plots for 8 non-overlapping "sentences" for each sequence in 1000-sample

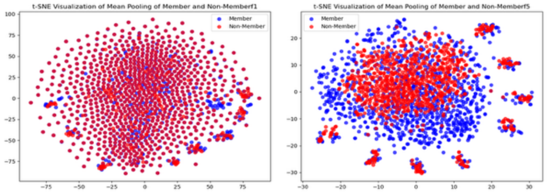


Figure 3: T-SNE visualization of embeddings between both 1000 sampled member and non-member datasets

We also evaluated the robustness of DNABERT-2 on 5000 sampled datasets with various perturbation rates. For each training sequence *s,* we generated a corresponding *s'* according to f. Figure 4 represents KDE plots for cosine similarity scores on 5000 sampled datasets with multiple f values over mean pooling.

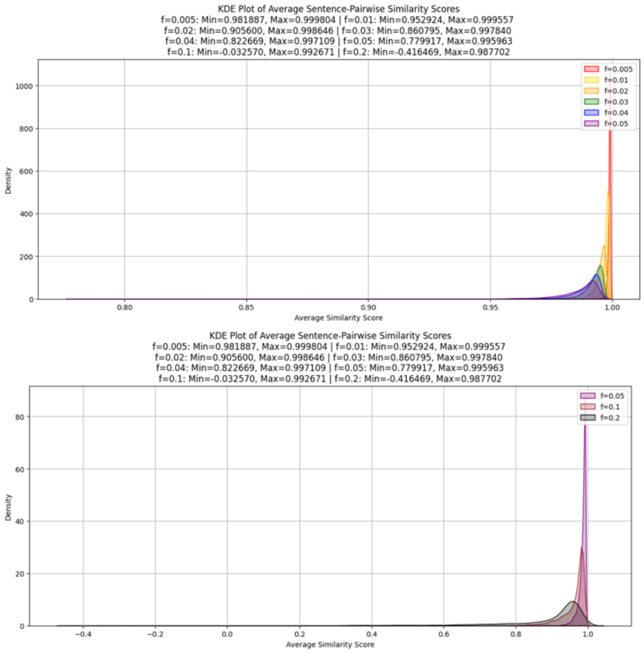
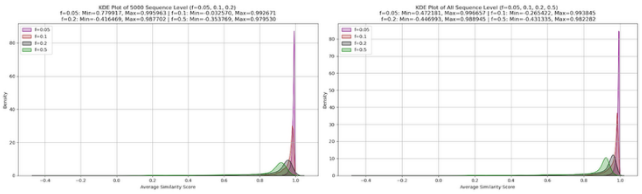


Figure 4: KDE plots for various perturbation of 5000 sampled datasets

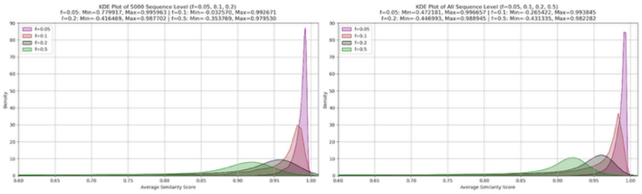
In addition to the experiments, we conducted a large-scale robustness study using a 5 000-sequence member set paired with 20 non-member variants per sequence at perturbation rates f ∈ {0.05, 0.10, 0.20, 0.50}. From the experiments with various perturbation rates, DNA-BERT2 seemed to be more sensitive with the mutation from these rates, so we decided to focus on them to quantify the previous work results. For each original sequence, we generated 20 perturbed version at each of the perturbation rates (totaling 100 000 non-member sequences) and computed whole-sequence embeddings under two pooling strategies:

- **Mean pooling**, averaging all token embeddings.
- **Max pooling**, taking the maximum activation per embedding dimension.

We then measured cosine similarity between each member embedding and its 20 non-member counterparts, recording the average similarity for each pooling method and each f-value. The distribution in both datasets has similar patterns. Figure 5 displays KDE plots for comparisons of mean pooling between 5000 sampled datasets and concatenation of their 20 versions in four perturbation rates.

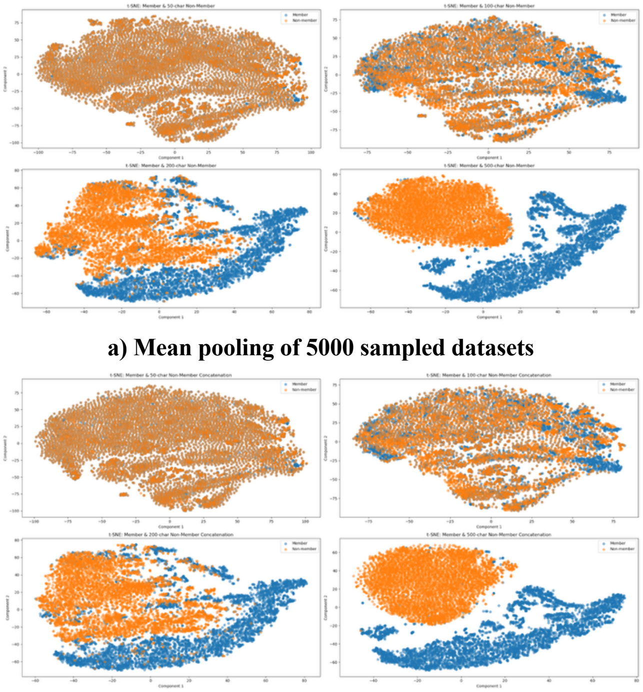


a) Overall looks



b) Focused on the peaks for better analysis

Figure 5: KDE plots for mean pooling between 5000 sampled dataset vs 20-version of each set in four f values.



a) Mean pooling of 5000 sampled datasets

b) Mean pooling of 20 versions of each dataset

Figure 6: T-SNE plots displayed the mean pooling distribution in 5000 sampled datasets and 20-version concatenation

In the t-SNE embedding visualization (Figure 6), non-member points form tight clusters around a few distinct regions, whereas member points are more broadly dispersed across the projection space, indicating greater variability in their learned representations.

In the max-pooling comparison (Figure 7), the non-member similarity curves shift progressively lower as the perturbation rate f increases, confirming that DNABERT-2's max pooled representations somewhat degrade with mutations. The higher density distributions at low f values underscores the heightened sensitivity of max-pooled features to even slight sequence changes. However, these results did not validate both pooling as an indicator of membership leakage in DNABERT-2 since the similarity scores are still high.

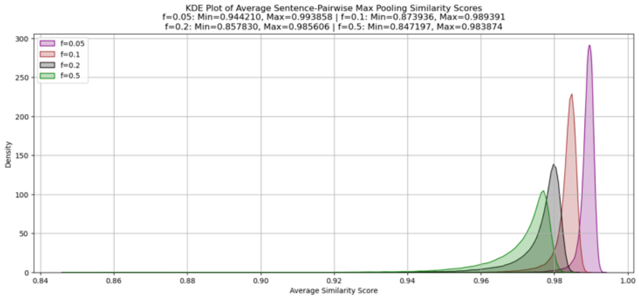


Figure 7: Max pooling of 20 version concatenation

Additionally, the t-SNE projections of max-pooled embeddings (Figure 8) reveal a consistently tight, well-defined cluster for member sequences (blue) across all non-member lengths, whereas non-member points (orange) form a sprawling, anisotropic cloud. At shorter perturbation lengths (50–100 chars), the non-member cloud elongates primarily along one axis, indicating that small mutations induce directional shifts in the embedding space. As the non-member sequence length increases (200–500 chars), the cloud becomes both denser and more isotropic, suggesting that larger context windows introduce greater embedding variability but also a convergence toward a central region. These patterns demonstrate that max-pooled embeddings can robustly separate member and non-member representations, with non-member mutations manifesting as characteristic deformations in the low-dimensional manifold.

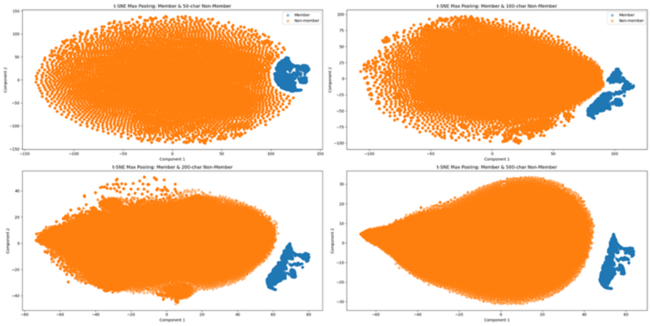


Figure 8: T-SNE visualization of max pooling between member and 20-version concatenation non-member datasets

5. CONCLUSIONS

We have demonstrated that similarity-based membership inference attacks can exploit embedding behaviors in DNA-BERT2. In our sliding-window experiments (Figure 1), larger window sizes produced wider score distributions, and member–non-member separation emerged more clearly at higher perturbation rates (f = 0.5). At the sentence level (Figures 2–3), adjacent-segment similarities alone offered limited discrimination, but t-SNE projections revealed distinct clustering of non-member embeddings versus the broader spread of member points. Extending to a 5000-sequence robustness study (Figure 4), we confirmed that whole-sequence mean-pooling similarities decline monotonically with increasing mutation rates. Finally, comparing mean vs. max pooling over 100,000 perturbed variants (Figures 5–7) showed that both mean and max pooling were not sensitive to perturbation. Even though max pooling amplifies sensitivity to small sequence changes, high absolute similarity values limit easy thresholding. T-SNE visualizations of max-pooled embeddings (Figure 8) further highlighted characteristic clustering patterns for non-members, suggesting that manifold structure may offer stronger leakage signals than raw similarity scores.

Through these evaluations, we accomplished:

1. A lightweight sliding-window MIA that aligns with theoretical attacks [1] but requires only pretrained token embeddings and high perturbation rates.
2. A sentence-level pipeline that quantifies aggregate embedding leakage at multiple granularities.
3. A large-scale perturbation study illustrating DNABERT-2's varying robustness across mutation rates and pooling strategies.

Had we to revisit this work, we would integrate a learned similarity metric—rather than cosine thresholds—to maximize discrimination between member and non-member distributions. We also plan to explore the accuracy of DNA-BERT2 model on downstream tasks. How will the model perform when it is not sensitive to perturbations?

Future work:

- □ **Downstream robustness assessment.** Identify one or more representative downstream tasks (e.g., promoter prediction or species classification) and measure model accuracy on perturbed inputs. This will reveal whether embedding insensitivity to mutation correlates with maintained task performance.
- □ **Cross-model evaluation.** Apply the sliding-window and sentence-level MIA pipelines to alternative genomic embedding models—such as the Nucleotide Transformer—to compare their leakage profiles and mutation sensitivity. Investigate DNABERT-S which was developed based on DNABERT-2.
- □ **Defense benchmarking.** Evaluate embedding-level defenses (e.g., noise regularization or adversarial fine-tuning) by measuring both MIA success rates and downstream task accuracy, seeking configurations that minimize privacy leakage without degrading utility
- □ **Learned Similarity Metrics.** Replace static cosine thresholds with trainable projection layers or metric networks that adaptively weight embedding dimensions to amplify the separable features uncovered by max pooling and t-SNE clustering. Such learned metrics may overcome the high-similarity "ceiling" observed at low perturbation rates.

6. REFERENCES

[1] D. Zheng *et al.* "Information Leakage in Embedding Models," *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2020. DOI:10.1145/3372297.3417270.

[2] L. Yin, Y. He, X. Chen, R. Lawrence, A. Nguyen, B. Ramprasad, and M. Zhao. "DNABERT-2: Efficient Foundation Model and Benchmark for Multi-Species Genome." *OpenReview*, 2024. https://openreview.net/forum?id=oMLQB4EZE1.