# ON THE IMPACT OF EMBEDDING ANISOTROPY IN GENOMIC LANGUAGE MODELS FOR BACTERIAL TAXONOMY

**Mike Leske** *
Department of Computer Sciences
Munster Technological University
Cork, Co.Cork, Ireland
mike.leske@mymtu.ie

**Francesca Bottacini**
Department of Biological Sciences
Munster Technological University
Cork, Co.Cork, Ireland
francesca.bottacini@mtu.ie

**Rafael R. C. Cuadrat**
mBiomics
Munich, Germany
rafael.cuadrat@mbiomics.com

**Haithem Afli**
Department of Computational Sciences
Munster Technological University
Cork, Co.Cork, Ireland
haithem.afli@mtu.ie

**Bruno G. N. Andrade**
Department of Computational Sciences
Munster Technological University
Cork, Co.Cork, Ireland
bruno.andrade@mtu.ie

## ABSTRACT

Genomic language models have emerged as powerful tools for representing DNA sequences, yet the impact of intrinsic properties of pre-trained embeddings, such as anisotropy, on downstream genomic tasks remains underexplored. In this work, we examine the geometric structure of DNABERT-2 embeddings derived from full-length 16S rRNA gene sequences and analyze how anisotropy affects bacterial taxonomic classification. We compare raw embeddings with post-processed representations obtained through a simple whitening transformation and evaluate their performance using distance-based classification across multiple taxonomic ranks. Our results show that DNABERT-2 embeddings exhibit severe anisotropy and that whitening substantially improves isotropy and consistently enhances classification performance, particularly at finer-grained taxonomic levels. These findings highlight the importance of embedding geometry when deploying genomic language models for downstream biological analysis.

## 1 INTRODUCTION

The genome is frequently described as the most sophisticated data storage system in nature (Preuss et al., 2024). The information stored can be described as a natural language, and is composed of a four-letter nucleotide alphabet, and the genome organizes information into gene "words" and functional "phrases," all regulated by a molecular grammar that dictates the flow of life. This linguistic parallel has led to the development of Genomic Language Models (gLMs) such as DNABERT-2 (Zhou et al., 2023) and Evo (Nguyen et al., 2024) to understand and process genomic information. These models learn to capture features ranging from regulatory motifs to evolutionary signals by mapping nucleotide sequences into high-dimensional vector spaces where geometric proximity is intended to mirror biological similarity. However, pre-trained models have intrinsic characteristics

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Funding acknowledgements go at the end of the paper. Please leave out the acknowledgements in the initial submission for the sake of double-blind review process

that can be challenging for DNA analysis, such as anisotropy, a structural bias that remains a significant yet under-addressed issue, even though it is recently gaining some attention (Diera et al., 2024; Huang et al., 2021).

Anisotropy occurs when embeddings collapse into a subspace, or narrow cone, rather than distributing evenly throughout the representation space (Godey et al., 2024). In NLP, this manifests itself as unexpectedly high cosine similarities between unrelated terms; in a genomic context, it leads to a discriminative collapse (Gao et al., 2019). This collapse is not well studied in genomics, but when nearly all vectors point in a similar direction, the margin for separating distinct bacterial classes shrinks, which may make the model hypersensitive to noise and distorting the link between geometric distance and biological relatedness. Regardless, anisotropy can be mitigated by a strategy called whitening, consisting in applying an affine linear transformation that reshapes the space into an isotropic state, while preserving the relative topological relationships between sequences and expanding the range of similarity values (Sasaki et al., 2023).

In this work, we focus on an empirical characterisation of anisotropy in genomic language model embeddings and its impact on downstream bacterial taxonomic classification. Rather than proposing a new representation learning method, our contribution lies in systematically analysing the geometric properties of DNABERT-2 embeddings in a genomic setting and demonstrating that a simple, model-agnostic post-processing step, the whitening operation, can substantially improve isotropy and downstream performance. We view this work as a practical study aimed at informing the use and interpretation of genomic foundation model embeddings.

## 2 METHODS

### 2.1 DATA, AND MODEL ACQUISITION AND PRE-PROCESSING

The Greengenes2 (McDonald et al., 2024) dataset was obtained from its repository for the analysis of the full-length 16S gene. Data was pre-processed to retain only sequences between 1,450 and 1,550 base pairs. A subsequent deduplication step preserved a single entry for each unique sequence-taxonomy pair. This resulted in 293,642 unique sequences spanning 18,301 species from Greengenes2, from which a total of 25,000 sequences were pre-selected (stratified by Family) and used in the classification experiments. DNABERT-2, a model that was pre-trained on multi-species genomic data, was chosen based on previous findings by Leske et al. (2025), retrieved from the Hugging Face hub and used to embed the pre-processed 16S sequences.

### 2.2 NEW - EMBEDDING WHITENING TRANSFORMATION

To resolve the anisotropic effect and restore a more uniform distribution to the DNABERT-2 embeddings, we implement a post-processing PCA Whitening transformation. Given a set of embeddings $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, where $n$ is the number of sequences and $d$ is the hidden dimension, the whitened representations $\widetilde{\boldsymbol{X}}$ are computed as:

$$\widetilde{\boldsymbol{X}} = (\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{W}$$

Here, $\boldsymbol{\mu}$ is the mean embedding vector used to center the distribution. The whitening matrix $\boldsymbol{W}$ is derived from the covariance matrix $\Sigma$ of the centered data. We first compute the eigendecomposition of $\Sigma$:

$$\Sigma = \frac{1}{n-1}(\boldsymbol{X} - \boldsymbol{\mu})^\top(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$$

where $\boldsymbol{U}$ is the orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. The whitening matrix is then defined as:

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}.$$

This transformation rotates the high-dimensional data to decorrelate the embedding dimensions and scales each axis by the inverse of its variance. Geometrically, this transforms the space into a sphere, ensuring that the resulting covariance matrix of $\widetilde{\boldsymbol{X}}$ is the identity matrix $\boldsymbol{I}$. This data transformation prevents a small subset of dominant directions from overshadowing the subtle biological signals necessary for fine-grained taxonomic classification (Figure 1).

Whitening was selected due to its simplicity, computational efficiency, and model-agnostic nature. Unlike learning-based isotropy correction methods, whitening does not require additional training data or task-specific supervision, making it particularly suitable for large-scale genomic embeddings. Prior work in NLP has shown that whitening can effectively mitigate anisotropy while preserving relative neighborhood structure, which is especially important for phylogeny-driven tasks.

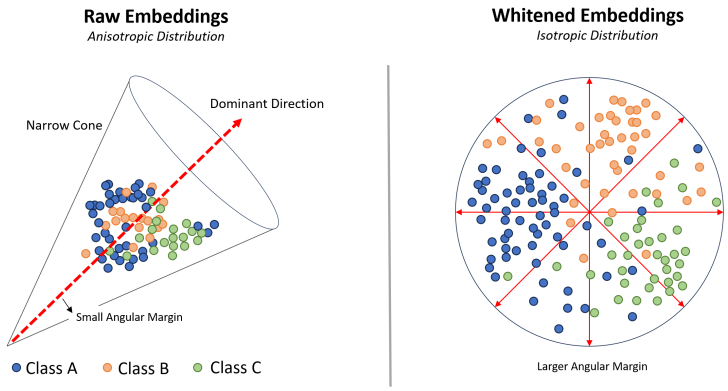


Figure 1: Raw embeddings tend to concentrate along a dominant direction, forming a narrow cone with small angular margins between classes. Whitening decorrelates embedding dimensions and redistributes variance, yielding a more isotropic space in which angular separation between taxonomic groups is increased.

### 2.3 Statistical Analysis

Anisotropy and Isotropy were measured using 3 different methodologies: Pairwise cosine similarity distribution, in which 10,000 random embedding pairs from both baseline (raw) and whitened embeddings were sampled and their cosine similarity measured as:

$$cos(\theta) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|}$$

and assessed by distribution statistics; Isotropy evaluation via Eigenvalue Analysis in which the covariance matrix of centered embeddings is computed and eigendecomposition is performed to obtain the variance along each principal axes, followed by the measurement of embedding dimensionality usage through Participation Ratio as:

$$PR = \frac{(\sum \lambda_i)^2}{\sum {\lambda_i}^2},$$

Normalized Isotropy re-scaling as *PR/dimensions* and the measurement of fraction of variance captured by the k-top eigenvalues; and Cumulative Variance Analysis, where the cumulative variance explained by the first k principal components.

To further support our analyses, the original and whitened cluster structures were evaluated using a Silhouette Score using the following formulae

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where *a(i)* stands for the mean distance to other samples within the same cluster and *b(i)* for the mean distance to samples in the nearest different cluster to measure how well it fits within its own cluster versus how well it would fit in the nearest neighboring cluster. We limited the computation to 50,000 samples per group where this threshold was crossed.

### 2.4 Bacterial taxonomic classification using k Nearest Neighbor (kNN)

Taxonomic classification was performed in all taxonomic levels, from Phylum to Species, using a kNN approach with FAISS (Facebook AI Similarity Search, Douze et al. (2024)) to classify new sequences based on their relative position to the k most similar embeddings in the embedding space. We focus on k-nearest neighbour classification as it directly reflects the geometric structure of the embedding space and is therefore particularly sensitive to anisotropy effects. While other classifiers could be considered, distance-based evaluation allows us to isolate the impact of embedding geometry without introducing additional model-specific inductive biases.

The kNN classification results were evaluated by their accuracy, purity, and Mean Average Precision (MAP) using the 25,000 pre-selected queries, being self-excluded when queried through index exclusion. $k$ values tested were 1, 5, 10, and 20 neighbors. Accuracy@$k$ is the fraction of classifications in which the correct taxonomic label appears among the top-$k$ search results. Similarly Purity@$k$ represents the average fraction of the top-$k$ results that share the correct taxonomy label. For example, a purity of 0.8 at $k$=5 indicates that 4 out of 5 neighbors shared the expected label. To avoid data leakage, we recomputed the whitening components using only non-query embeddings and subsequently transformed all embeddings uniformly.

## 3 Results

### 3.1 The impact of Anisotropy and whitening transformation

The identification of anisotropy of DNABERT-2 model embeddings was performed by using 3 different methods, pairwise cosine similarity distribution, Isotropic metrics through Eigenvalue Analysis, and cumulative variance analysis of the first k principal components.

For raw embeddings, pairwise cosine similarity revealed that nearly all pairs have a mean similarity of $>$0.98 with $\sigma$ = 0.009 (Figure 2). Eigenvalue Analysis showed a participation ratio of 17.5/768 dimensions and a normalized isotropy of 0.02. Considering all 768 available dimensions, the first 10 dimensions held 57% and 120 dimensions 90% of the variance. These results support a case of severe anisotropy, in which all embeddings reside in a narrow cone-shaped subspace and the vectors share a dominant direction.

Conversely, whitened embeddings showed a $\mu$ embed cosine similarity of approximately 0 (0.0002) with $\sigma$ = 0.04, a participation ratio of 562.5 dimensions and a normalized isotropy of 0.73, suggesting that the geometry of embedding space geometry was successfully transformed from a narrow cone to a hypersphere. In addition, the first 10 dimensions held only 5% of variance instead of 57%, showing that the dominance of the first dimensions was eliminated and the variance is distributed across dimensions (Figure 3).

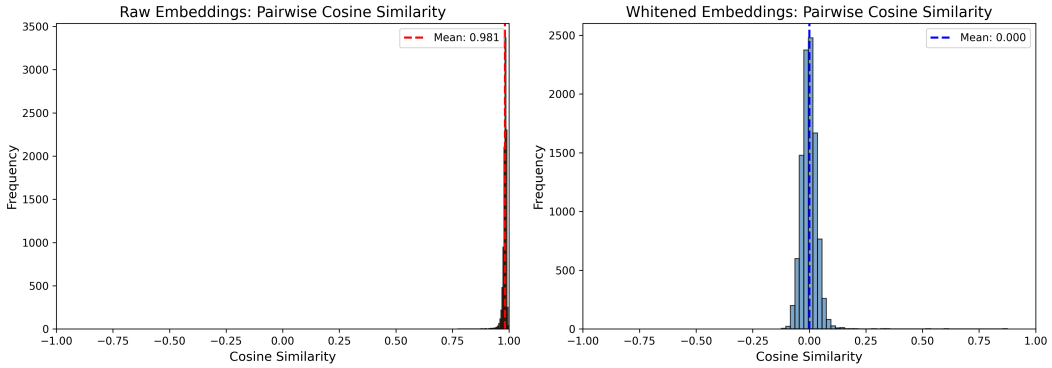


Figure 2: Pairwise Cosine Similarity distribution of Raw and Whitened DNABERT-2 embeddings.

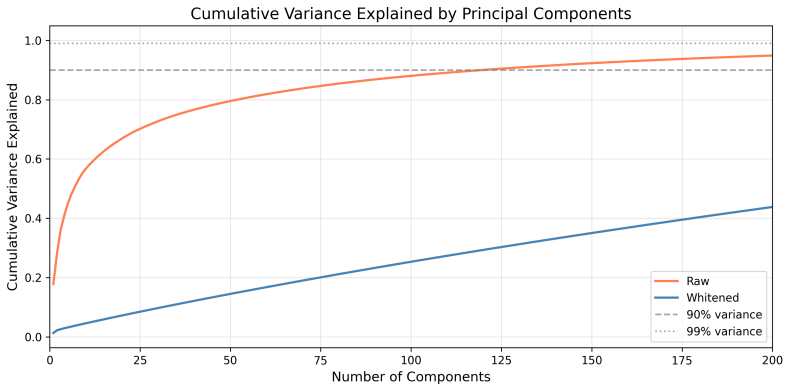


Figure 3: Cumulative explained variance across the 768 embedding's eigenvalues. The first dimensions of raw embeddings contained more information, with 57% of the variance being contained by the first 10 dimensions, conversely, the variance was spread across more dimensions for whitened embeddings.

### 3.2 Whitening on bacterial taxonomic hierarchy and classification

Performance of raw and whitening embeddings were evaluated on 123 unique Phyla, 307 classes, 859 orders, 1,854 families, 6,995 genera, and 18,391 species, corresponding to the entirety of Greengenes2 dataset. Accuracy, Purity, MAP metrics and associated improvement by taxonomy level for raw and whitened embeddings can be found in Table 1.

## 4 Discussion

Anisotropy is a common issue in pre-trained language models like BERT and GPT (Godey et al., 2024) and the gLM Evo (Wu et al., 2025), and is characterized by embeddings being clustered within a narrow, cone-shaped geometry instead of a hypersphere. This problem can cause vectors to share a dominant direction, which can lead to high average cosine similarity regardless of their semantic relationship; however, its impact in gLMs is understudied.

### 4.1 Anisotropy heavily affects 16s rRNA gene embedding space

The 16s rRNA gene is a widely used evolutionary marker and conserved gene. Its eight hypervariable regions are used to identify and cluster different microorganisms into the established taxonomy hierarchy for classification purposes (Woese, 1987; López-Aladid et al., 2023). Due to these properties, we hypothesized that DNABERT-2 embeddings created by using these genes would be heavily anisotropic.

Altogether, our results give support to this hypothesis and indicate that the raw 16s embeddings reside in a narrow subspace, and its small discriminative power has the potential to hinder genomic analysis. In addition, whitening transformed the geometry of the space into a hypersphere, an isotropic state, and increased the distance between taxonomic groups and distributed the variance across a larger number of dimensions, allowing us to make use of almost the entire embedding information.

### 4.2 Whitened embeddings improved Taxonomic classification

In addition, our analyses demonstrate that the geometric regularization provided by the whitening improved downstream classification performance. The raw DNABERT-2 embeddings already successfully captured phylogenetic signals in an anisotropic state with >92% species-level accuracy, which is further supported by Leske et al. (2025), in which the classification task with DNABERT-2 embeddings achieved a classification accuracy on par with the state-of-the-art alignment-based tools, however, the whitening transformation demonstrated consistent additional gains in Accuracy, Mean

Table 1: Accuracy, Purity, and Mean Average Precision (MAP) for Raw and Whitened embeddings across different taxonomic levels and top-k.

| | | Accuracy | | | Purity | | | MAP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Level | k | Raw | White | Δ | Raw | White | Δ | Raw | White | Δ |
| Phylum | 1 | 0.9987 | 0.9990 | +0.03% | 0.9987 | 0.9990 | +0.03% | 0.9987 | 0.9990 | +0.03% |
| | 5 | 0.9990 | 0.9993 | +0.03% | 0.9966 | 0.9978 | +0.12% | 0.9964 | 0.9976 | +0.12% |
| | 10 | 0.9992 | 0.9995 | +0.03% | 0.9950 | 0.9971 | +0.21% | 0.9945 | 0.9967 | +0.22% |
| | 20 | 0.9994 | 0.9996 | +0.02% | 0.9921 | 0.9959 | +0.38% | 0.9912 | 0.9953 | +0.40% |
| Class | 1 | 0.9986 | 0.9989 | +0.03% | 0.9986 | 0.9989 | +0.03% | 0.9986 | 0.9989 | +0.03% |
| | 5 | 0.9990 | 0.9993 | +0.03% | 0.9963 | 0.9976 | +0.13% | 0.9961 | 0.9974 | +0.13% |
| | 10 | 0.9992 | 0.9995 | +0.04% | 0.9945 | 0.9968 | +0.24% | 0.9940 | 0.9964 | +0.24% |
| | 20 | 0.9994 | 0.9996 | +0.02% | 0.9914 | 0.9958 | +0.44% | 0.9905 | 0.9951 | +0.46% |
| Order | 1 | 0.9972 | 0.9981 | +0.10% | 0.9972 | 0.9981 | +0.10% | 0.9972 | 0.9981 | +0.10% |
| | 5 | 0.9982 | 0.9987 | +0.05% | 0.9923 | 0.9957 | +0.34% | 0.9917 | 0.9954 | +0.36% |
| | 10 | 0.9985 | 0.9990 | +0.06% | 0.9880 | 0.9941 | +0.61% | 0.9869 | 0.9934 | +0.65% |
| | 20 | 0.9988 | 0.9992 | +0.05% | 0.9819 | 0.9921 | +1.02% | 0.9800 | 0.9908 | +1.08% |
| Family | 1 | 0.9961 | 0.9978 | +0.17% | 0.9961 | 0.9978 | +0.17% | 0.9961 | 0.9978 | +0.17% |
| | 5 | 0.9974 | 0.9985 | +0.12% | 0.9878 | 0.9941 | +0.63% | 0.9870 | 0.9936 | +0.66% |
| | 10 | 0.9978 | 0.9987 | +0.09% | 0.9812 | 0.9918 | +1.06% | 0.9796 | 0.9908 | +1.12% |
| | 20 | 0.9982 | 0.9990 | +0.08% | 0.9715 | 0.9887 | +1.72% | 0.9684 | 0.9868 | +1.83% |
| Genus | 1 | 0.9885 | 0.9929 | +0.44% | 0.9885 | 0.9929 | +0.44% | 0.9885 | 0.9929 | +0.44% |
| | 5 | 0.9942 | 0.9971 | +0.29% | 0.9659 | 0.9812 | +1.52% | 0.9628 | 0.9788 | +1.61% |
| | 10 | 0.9956 | 0.9978 | +0.23% | 0.9506 | 0.9739 | +2.33% | 0.9441 | 0.9690 | +2.49% |
| | 20 | 0.9964 | 0.9985 | +0.22% | 0.9308 | 0.9649 | +3.40% | 0.9195 | 0.9562 | +3.66% |
| Species | 1 | 0.9254 | 0.9432 | +1.79% | 0.9254 | 0.9432 | +1.79% | 0.9254 | 0.9432 | +1.79% |
| | 5 | 0.9696 | 0.9833 | +1.36% | 0.8950 | 0.9238 | +2.88% | 0.8729 | 0.9037 | +3.08% |
| | 10 | 0.9784 | 0.9901 | +1.17% | 0.8810 | 0.9172 | +3.63% | 0.8453 | 0.8845 | +3.92% |
| | 20 | 0.9850 | 0.9943 | +0.93% | 0.8685 | 0.9124 | +4.39% | 0.8162 | 0.8647 | +4.85% |

Table 2: Silhouette scores for Raw and Whitened embeddings. The silhouettes for Phylum and Classes are below 0, which means that some individuals are breaching the boundaries between groups, while higher levels, such as genus and species are more defined for both embeddings, but more evidently on whitened embeddings.

| | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|
| Raw | -0.029 | -0.100 | 0.063 | 0.088 | 0.094 | 0.347 |
| Whitened | -0.026 | -0.031 | 0.029 | 0.081 | 0.207 | 0.407 |
| Δ | 0.003 | 0.069 | -0.024 | -0.007 | 0.113 | 0.060 |

Average Precision (MAP), and Purity across all taxonomic ranks (see Table 1), with an accuracy gain of 1.79% for species and 0.44% for genus level.

This performance improvement can be explained by the space geometry itself (Diera et al., 2024). In the raw anisotropic cone, the angular margins between distinct but closely related species are artificially compressed, making nearest-neighbor searches sensitive to noise and dominant directional biases. Besides, while this is true for finer-grained levels, such as Genus and Species, where boundaries are biologically sharper, negative silhouette scores observed at higher taxonomic ranks (e.g., Phylum and Class) are likely reflect inherent biological overlap and evolutionary continuity between broad taxa rather than solely embedding deficiencies.

Whitening expanded these margins, ensuring that even subtle genetic differences translate into meaningful geometric distances, and did it without compromising the hierarchical taxonomy structure and is particularly effective when subtle genetic differences must be resolved.

## 5 Conclusion

In this work, we investigated the impact of anisotropy in genomic language model embeddings and its implications for bacterial taxonomic classification. Focusing on DNABERT-2 embeddings derived from full-length 16S rRNA gene sequences, we provided a systematic analysis of embedding geometry using multiple isotropy diagnostics and showed that raw embeddings exhibit severe representation collapse.

Building on this analysis, we demonstrated that a simple, model-agnostic whitening transformation effectively restores isotropy by redistributing variance across embedding dimensions while preserving neighborhood structure. Across multiple taxonomic ranks, from Phylum to Species, whitened embeddings consistently improved distance-based classification performance, yielding higher accuracy, purity, and mean average precision, with particularly notable gains at finer-grained taxonomic levels.

Our results confirm that embedding geometry plays a critical role in downstream genomic analysis and that addressing anisotropy can lead to measurable and reliable improvements without modifying the underlying language model or training procedure. In general, this work highlights whitening as a lightweight yet effective post-processing strategy and underscores the importance of isotropy-aware evaluation when deploying genomic language model embeddings for computational biology tasks.

## 6 Limitations

This study focuses on DNABERT-2 embeddings and full-length 16S rRNA sequences; therefore, the findings may not directly generalize to other genomic language models, shorter reads, or non-ribosomal genes. In addition, downstream evaluation is limited to distance-based kNN classification, which was chosen to directly probe embedding geometry. Future work will explore whether similar isotropy effects and improvements hold for other genomic foundation models, alternative downstream tasks, and learning-based classifiers, as well as investigate isotropy-aware training objectives tailored to genomic data.

## 7 Ethical Considerations

The research presented in this work focuses on the analysis of genomic sequence representations and does not involve human subjects, personally identifiable information, or clinical decision-making. All experiments were conducted using publicly available microbial genomic data from the Greengenes2 database, which aggregates bacterial and archaeal sequences for research purposes.

The proposed whitening procedure is a post-processing transformation applied to model embeddings and does not alter the underlying biological sequences or taxonomic annotations. As such, it does not introduce new biological interpretations or predictions beyond those already implied by the original embeddings. Instead, it aims to improve the geometric properties of representation spaces to enable more reliable downstream analysis.

Potential risks associated with this work are primarily related to the misuse or overinterpretation of embedding-based classification results. Although improved isotropy can enhance distance-based discrimination, embedding similarity should not be treated as definitive evidence of biological function, pathogenicity, or evolutionary causality. We emphasize that the methods presented are intended as analytical tools to support exploratory genomic analysis, rather than as replacements for established biological validation or clinical pipelines.

Finally, by highlighting anisotropy as an inherent property of genomic language model embeddings, this work contributes to more transparent and responsible use of foundation models in computational biology. Improving awareness of representation biases and their impact on downstream tasks aligns with broader efforts toward trustworthy and interpretable AI systems in the life sciences.

## FUNDING

This research was supported by the Horizon Europe project GenDAI (Grant Agreement ID: 101182801) and the ADAPT Taighde Éireann – Research Ireland Centre at Munster Technological University. The ADAPT Centre for Digital Media Technology is funded by Taighde Éireann – Research Ireland through the Research Centres Programme and co-funded by the European Regional Development Fund (ERDF) under Grant 13/RC/2106_P2.

## REFERENCES

Andor Diera, Lukas Galke, and Ansgar Scherp. Isotropy matters: Soft-ZCA whitening of embeddings for semantic code search. *arXiv [cs.CL]*, November 2024.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv [cs.LG]*, January 2024.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv [cs.CL]*, July 2019.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35–48, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

Mike Leske, Jamie A FitzGerald, Keith Coughlan, Francesca Bottacini, Haithem Afli, and Bruno Gabriel Nascimento Andrade. Alignment-free bacterial taxonomy classification with genomic language models. *bioRxiv*, June 2025.

Ruben López-Aladid, Laia Fernández-Barat, Victoria Alcaraz-Serrano, Leticia Bueno-Freire, Nil Vázquez, Roque Pastor-Ibáñez, Andrea Palomeque, Patricia Oscanoa, and Antoni Torres. Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. *Sci. Rep.*, 13(1):3974, March 2023.

Daniel McDonald, Yueyu Jiang, Metin Balaban, Kalen Cantrell, Qiyun Zhu, Antonio Gonzalez, James T Morton, Giorgia Nicolaou, Donovan H Parks, Søren M Karst, Mads Albertsen, Philip Hugenholtz, Todd DeSantis, Se Jin Song, Andrew Bartko, Aki S Havulinna, Pekka Jousilahti, Susan Cheng, Michael Inouye, Teemu Niiranen, Mohit Jain, Veikko Salomaa, Leo Lahti, Siavash Mirarab, and Rob Knight. Greengenes2 unifies microbial data in a single reference tree. *Nat. Biotechnol.*, 42(5):715–718, May 2024.

Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, Jeremy Sullivan, Madelena Y Ng, Ashley

Lewis, Aaron Lou, Stefano Ermon, Stephen A Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D Hsu, and Brian L Hie. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, November 2024.

Inbal Preuss, Michael Rosenberg, Zohar Yakhini, and Leon Anavy. Efficient DNA-based data storage using shortmer combinatorial encoding. *Sci. Rep.*, 14(1):7731, April 2024.

Shota Sasaki, Benjamin Heinzerling, Jun Suzuki, and Kentaro Inui. Examining the effect of whitening on static and contextualized word embeddings. *Inf. Process. Manag.*, 60(3):103272, May 2023.

C R Woese. Bacterial evolution. *Microbiol. Rev.*, 51(2):221–271, 1987.

Weimin Wu, Zhihan Zhou, Robert Riley, Mohammad Abdulqader, Xuefeng Song, Satria Kautsar, Rob Egan, Steven Hofmeyr, Gefei Liu, Shira Goldhaber-Gordon, Mutian Yu, Harrison Ho, Yaping Liu, Andrei Stecca Steindorff, Fengchen Liu, Feng Chen, Rachael Morgan-Kiss, Lizhen Shi, Han Liu, and Zhong Wang. Uncovering the genomic manifold via scalable learning from the global microbiome. *bioRxiv*, 2025. doi: 10.1101/2025.01.30.635558. URL https://www.biorxiv.org/content/early/2025/12/09/2025.01.30.635558.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv [q-bio.GN]*, June 2023.