Sparse Autoencoders Reveal Interpretable Structure in Small Gene Language Models

Haoxiang Guan[a], **Jiyan He**[a], **Jie Zhang**[b]

[a] *University of Science and Technology of China, Hefei, Anhui, China*

[b] *CFAR and IHPC, Agency for Science, Technology and Research, Singapore,* zhang_jie@cfar.a-star.edu.sg

1. Introduction

Sparse autoencoders (SAEs) have recently emerged as a powerful tool for interpreting the internal representations of large language models (LLMs), revealing latent latent features with semantical meaning [1]. This interpretability has also proven valuable in biological domains: applying SAEs to protein language models uncovered meaningful features related to protein structure and function [2]. More recently, SAEs have been used to analyze genomics-focused models such as Evo 2 [3], identifying interpretable features in gene sequences. However, it remains ***unclear*** whether SAEs can extract meaningful representations from small gene language models, which have fewer parameters and potentially less expressive capacity. To address it, we propose applying SAEs to the activations of a small gene language model. We demonstrate that even small-scale models encode biologically relevant genomic features, such as transcription factor binding motifs, that SAEs can effectively uncover. Our findings suggest that compact gene language models are capable of learning structured genomic representations, and that SAEs offer a scalable approach for interpreting gene models across various model sizes.

2. Methods

To uncover interpretable structures in small gene language models, we trained sparse autoencoders (SAEs) on embeddings derived from HyenaDNA-small-32k [4], a compact gene language model pretrained at single-nucleotide resolution on the human reference genome [5]. The overall pipeline is illustrated in Figure 1. For training, we extracted latent representations from the third layer of HyenaDNA-small-32k using sequences sampled from the human reference genome, each of length 32k nucleotides. To prevent model overfitting to specific genomic contexts, we globally shuffled these activations, ensuring that representations derived from the same input sequence were unlikely to appear together in the same training batch. These processed activations were then used to train SAEs with an expansion factor of 32×, creating feature dictionaries of size 8,192. The learning rate was linearly warmed up over the first 5% of training steps and then fixed at 1e-6, with an L1 penalty of 0.1 and a batch size of 2,048.

To evaluate whether the resulting sparse features correspond to biologically meaningful genomic elements, we annotated chromosome 14 with JASPAR transcription factor binding sites (TFBS) [6], accessed via the UCSC Table Browser [7]. We then applied quality filtering based on motif frequency and p-value thresholds to retain high-confidence annotations. To facilitate a direct comparison between features and annotations, we converted motif-level annotations into nucleotide-level labels, and used an activation threshold of 0.15 to determine whether an SAE feature was activated or not. This allowed us to systematically assess the alignment between SAE features and known regulatory elements with metrics such as precision, recall, and F1 score, consistent with methodologies introduced by InterPLM [2].

3. Results

By applying SAEs to embeddings from HyenaDNA-small-32k, we successfully identified sparse features corresponding to individual nucleotides and biologically relevant transcription factor binding sites (TFBS). As shown in Figure 3-A, nucleotide-specific features exhibit high precision, which indicates that the learned representations are selective for specific nucleotide identities, although recall varies. This is consistent with prior findings from Evo 2 [3]. Figure 2 presents the activation pattern of feature f/357 across a 500 bp segment of human chromosome 14. Notably, the activation peaks consistently align with cytosine positions throughout the sequence, demonstrating that this feature has independently learned to recognize this specific nucleotide.

Beyond nucleotide-level features, we identified sparse dimensions aligned with known transcription factor motifs, as illustrated in Figure 3-B. Notably, these factors have well-established biological roles: MA1596.1 and MA2121.1 belong to the C2H2 zinc finger factor class, which plays crucial roles in gene regulation. MA0052.5 belongs to the MADS-box class, known for its involvement in muscle development, cell proliferation, and differentiation in animals. The relatively high precision of these features indicates that even compact models capture transcription factor binding specificity effectively. Nevertheless, the observed variability in recall highlights the inherent complexity and redundancy of regulatory elements in genomic sequences.

Overall, we demonstrate that small gene language models encode structured and biologically relevant representations, spanning both nucleotide composition and transcription factor binding patterns.

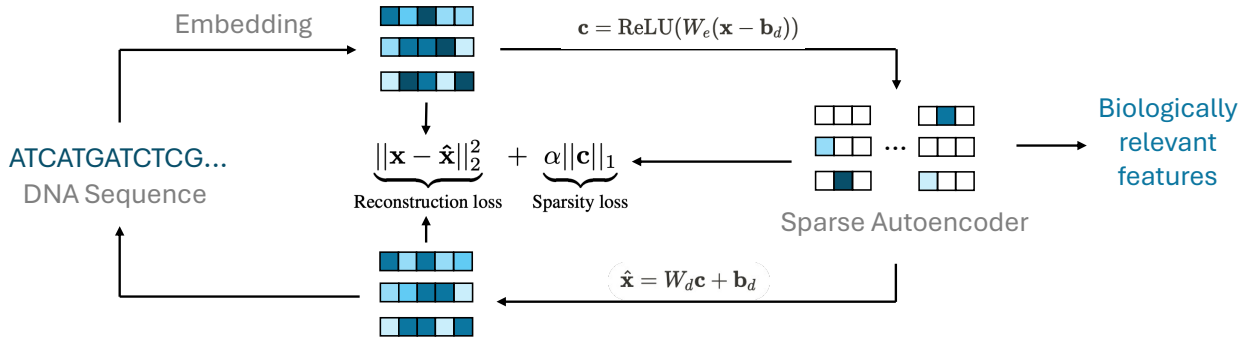

Fig. 1: Overall pipeline for training SAEs on genomes, followed by identifying biologically relevant features.

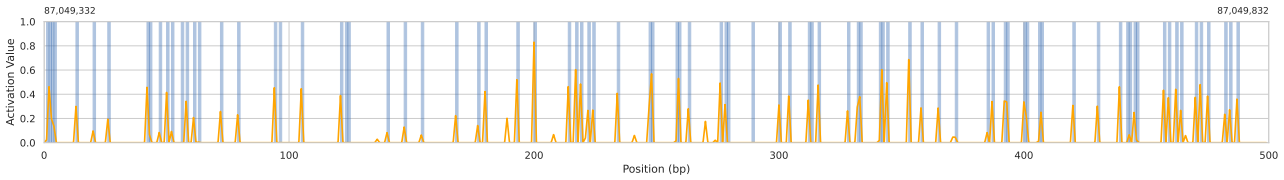

Fig. 2: Activation pattern of feature f/357 across a 500 bp segment of human chromosome 14. Orange peaks represent activation values, while blue bars indicate cytosine positions starting at position 87,049,332, revealing a strong correlation between the feature and this specific nucleotide.

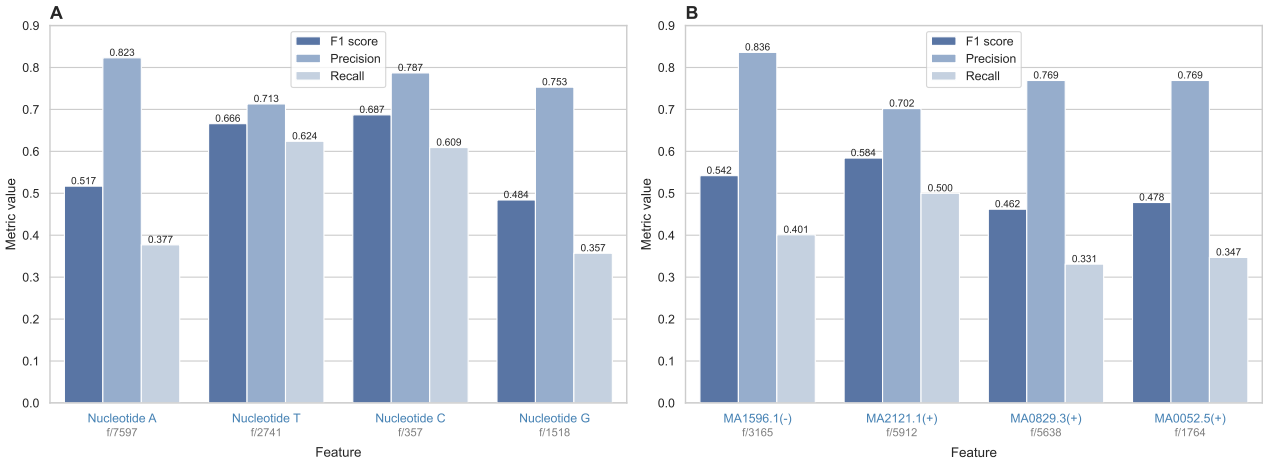

Fig. 3: Sparse autoencoders reveal interpretable nucleotide and transcription factor binding site (TFBS) features in HyenaDNA-small-32k. (**A**) Performance metrics for sparse features corresponding to individual nucleotides (A, T, C, G). (**B**) Metrics for sparse features associated with known TFBSs from JASPAR database [6]. Strand specificity was indicated by the +/-.

4. Conclusion

Our study demonstrates that sparse autoencoders (SAEs) can extract biologically meaningful representations from small gene language models, revealing structured features at both the nucleotide and regulatory element levels. By applying SAEs to embeddings from HyenaDNA-small-32k, we identified sparse dimensions corresponding to individual nucleotides as well as transcription factor binding motifs, highlighting the ability of compact models to capture essential genomic features. Future research could extend this approach to other genomic contexts, such as non-coding regions or species-specific variations, and explore how SAEs could aid model refinement and interpretability across different architectures. Additionally, SAEs could be applied to other modalities of biological models and data, such as single-cell gene expression and multi-omics datasets, to uncover interpretable representations across diverse biological systems.

References

[1] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[2] Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, 2025.

[3] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025.

[4] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. 2023.

[5] Homo sapiens genome assembly GRCh38 — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/. [Accessed 12-03-2025].

[6] Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen, Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, 11 2023.

[7] Donna Karolchik, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The ucsc table browser data retrieval tool. *Nucleic Acids Research*, 32:D493–D496, 01 2004.