INTERPRETABLE CANCER DRIVER GENE PREDICTION FROM DNA SEQUENCES USING A GENOMIC LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Identifying cancer driver genes, whose mutations confer a selective growth advantage to tumor cells, is critical for understanding tumorigenesis and targeted treatments. However, this task remains challenging due to tumor heterogeneity, context-specific effects, and the limited availability of labeled data (Martínez-Jiménez et al., 2020). Traditional computational methods often rely on handcrafted features, which may not fully capture the complexity of genomic sequences (Malebary & Khan, 2021). Recent advances in genomic language models (gLMs) offer a promising alternative (Benegas et al., 2025) by learning directly from raw DNA sequences, potentially uncovering latent features associated with driver genes.

In this paper, we utilize Caduceus (Schiff et al., 2024), a state-of-the-art gLM, to predict cancer driver genes from DNA sequences. Caduceus is uniquely suited for this task due to its linear scaling, bidirectional context modeling, and support for reverse complementarity (see Appendix for details). We fine-tuned Caduceus on a curated dataset of driver and passenger genes and analyzed the learned representations using post-hoc explainability methods. We aim to test the hypothesis that gLM embeddings capture sequence features that distinguish driver genes from passenger genes. This work could pave the way toward more generalizable and interpretable cancer gene prediction methods applicable to sequencing data from individual patients.

2 METHODS

We fine-tuned Caduceus on a curated dataset of 888 driver genes and 1,528 passenger genes (see Appendix for details). To address class imbalance and sequence length variability, we also extracted all the unique transcript (cDNA) sequences for each gene, resulting in a dataset that include 13,687 driver transcripts and 15,816 passenger transcripts. The dataset was split into 90% training and 10% testing datasets, with five random splits to ensure robust evaluation. We applied majority voting to combine predictions from transcript sequences of the same gene.

To mitigate overfitting, we utilized regularization strategies such as early stopping, L2 regularization, and dropout. We applied principal component analysis (PCA) to visualize learned embeddings and assess context learning. We used SHAP (SHapley Additive exPlanations) to identify sequence features contributing to predictions (Lundberg & Lee, 2017). To further evaluate biological relevance of the identified sequence features, we examined the contributions of somatic point mutations in driver genes to predictions (Tate et al., 2018).

3 RESULTS

The fine-tuned Caduceus model on gene DNA sequences achieved an accuracy of 0.693 ± 0.18 and an F1-score of 0.683 ± 0.02 on the test dataset across five runs. The model showed poorer performance and more overfitting on transcript sequences. While the performance is promising given the complexity of the task, it highlights the need for further optimization to improve its applicability. The PCA plot of learned embeddings reveals partial separation between driver and passenger genes, indicating that the model captures representative sequence features to some extent (Figure 1). However, the overlap between classes suggests that the model could benefit from incorporating additional features or architectural refinements to improve its discrimination power.

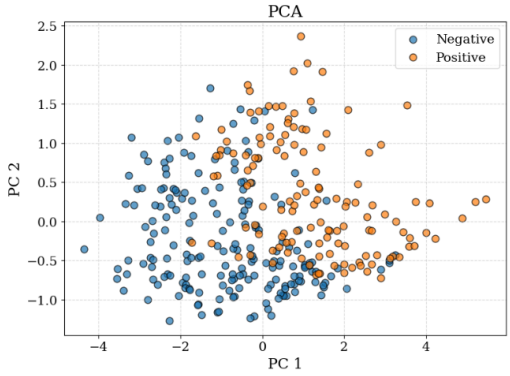

Figure 1: PCA visualization of learned embeddings for driver (orange) and passenger(blue) gene DNA sequences on the test dataset with highest accuracy.

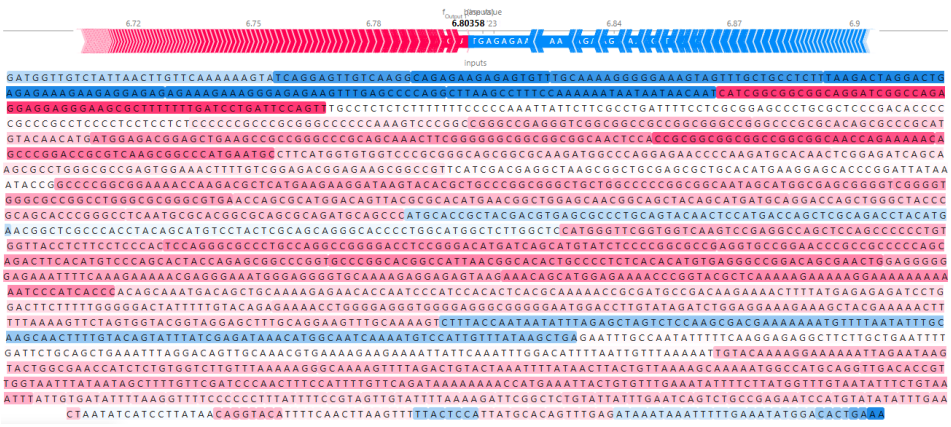

Figure 2: SHAP force plot for the prediction of cancer driver gene SOX2. The highlighted regions indicate subsequences that the model deems important for its predictions. Red regions contribute positively to the prediction, whereas blue regions contribute negatively. The direction plot at the top indicates the overall influence of each region on the prediction, though its interpretability is limited for large sequences. Color intensity reflects the strength of the contribution, with darker shades indicating greater influence.

The SHAP force plots allow the identification and analysis of specific regions in driver genes that are critical for model predictions (Figure 2). For SOX2, a well-established cancer driver gene across multiple cancer types and a promising drug target, approximately 76% of its somatic point mutations were positively associated with the model's prediction of it as a driver gene. For 50 short driver genes ranging from 1,576 to 5,539 base pairs, including SOX2, the mean positive contribution ratio was 0.61 (standard deviation 0.13), indicating a general tendency of more somatic point mutations contributing to the prediction of driver genes (see Appendix for details).

4 Conclusion

In this work, we leveraged Caduceus, a cutting-edge gLM, to predict cancer driver genes from DNA sequences. The model demonstrated moderate accuracy and provided interpretable insights into sequence features related to gene fitness. However, challenges such as overfitting and partial class separation indicate room for improvement. Future work will focus on enhancing the generalizability of the model, further exploring the biological relevance of the learned representations, incorporating additional biological context for predictions, and extending the approach to patient-specific DNA sequencing data.

MEANINGFULNESS STATEMENT

Cancer driver genes are usually detected as positively selected genes with high fitness. We hypothesize that subsequences within these genes carry signals of positive selection that can be learned by genomic language models (gLMs). In this work, we fine-tuned Caduceus, a high-performing long-range gLM, to predict cancer driver genes from DNA sequences. Post-hoc interpretations of our fine-tuned model helped to explain important sequence features associated with gene fitness such as known somatic mutations in driver genes. Our approach generates meaningful representations of DNA sequences related to cancer driver genes and provides a framework toward interpretable cancer driver gene prediction.

REFERENCES

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S Song. Genomic language models: opportunities and challenges. *Trends in Genetics*, pp. D941–D947, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Ben Kinnersley, Amit Sud, Andrew Everall, Alex J Cornish, Daniel Chubb, Richard Culliford, Andreas J Gruber, Adrian Lärkeryd, Costas Mitsopoulos, David Wedge, et al. Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. *Nature Genetics*, pp. 1868–1877, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Sharaf J Malebary and Yaser Daanial Khan. Evaluating machine learning methodologies for identification of cancer driver genes. *Scientific reports*, 11(1):12281, 2021.

Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, et al. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572, 2020.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018. doi: 10.1093/nar/gky1015.

A APPENDIX

The source code and datasets used in this study will be released upon publication.

A.1 MODEL OVERVIEW

Caduceus is a gLM designed for modeling long-range genomic sequences (Schiff et al., 2024). Built upon the Mamba architecture (Gu & Dao, 2023), Caduceus introduced bi-directionality and reverse-complement (RC) equivariance, enabling it to capture context from both upstream and downstream genomic regions while maintaining consistency. This makes Caduceus particularly well-suited for tasks requiring long-range dependencies, such as predicting the effects of genetic variants and identifying regulatory elements. The RC-equivariant architecture also eliminates the need for explicit RC data augmentation during training, further enhancing its robustness and generalizability.

Caduceus has demonstrated strong performance across a variety of downstream genomic tasks. Notably, it outperforms larger transformer-based models on tasks requiring long-range context, such as predicting the effects of mutations located far from transcription start sites. Its efficiency in handling sequences of up to 131kb base pairs, without the quadratic scaling issues of attention-based models, also makes it a practical choice for detecting cancer driver genes, which typically range in size from a few hundred to hundreds of thousand base pairs.

A.2 Dataset Details

We obtained 888 protein-coding driver genes from IntOGen (version 2024-06-18, 633 genes) (Martínez-Jiménez et al., 2020) and COSMIC Cancer Gene Census (version 101, 255 genes) (Tate et al., 2018) for training. By excluding newly discovered driver genes, including 78 genes in a recent study (Kinnersley et al., 2024), we obtained 1,528 passenger genes from 1,743 genes previously employed to evaluate various machine learning methods for cancer driver gene prediction (Malebary & Khan, 2021). We extracted DNA sequences of these genes based on their genomic locations on BSgenome.Hsapiens.UCSC.hg38 and transcript sequences from Ensembl (version 113).

The driver gene sequences range from 490 to 2,473,539 base pairs, whereas the passenger gene sequences range from 630 to 2,173,324 base pairs. Transcript sequences are much shorter, with driver gene transcripts ranging from 60 to 46,191 base pairs and passenger gene transcripts ranging from 19 to 24,020 base pairs.

A.3 Mutation Analysis

Using SHAP, we analyzed 50 short genes out of 114 driver genes in the test dataset where the highest prediction accuracy was achieved on gene DNA sequences. We obtained 11,483 somatic point mutations within these genes from COSMIC Cancer Mutation Census (Tate et al., 2018) and evaluated whether these mutations contributed positively or negatively to model predictions. We define the positive contribution ratio for a gene as the percentage of its mutations contributing to the prediction of it being a cancer driver gene.

The distribution of positive contribution ratios across genes slightly skews toward larger values (Figure 3), suggesting that a high proportion of driver mutations were captured by the model. Among these genes, MYD88 exhibited the highest positive contribution ratio, with 96.3% of its mutations classified as positive. In contrast, TNFRSF17 had the lowest ratio at 15.4%, suggesting that its mutations were more frequently linked to negative contributions. Further analysis will be needed to better understand the heterogeneity among these genes.

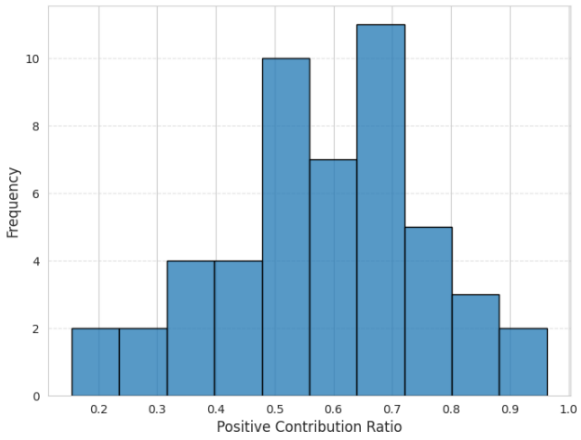

Figure 3: Distribution of positive contribution ratios across 50 cancer driver genes. The x-axis represents the positive contribution ratio, indicating the proportion of a gene's mutations that positively contributed to the model's prediction of driver genes. The y-axis represents the frequency of genes.