A Genomic Language Model for Zero-Shot Prediction of Promoter Indel Effects

Courtney A. Shearer [1] **Felix Teufel** [1 2 3] **Rose Orenbuch** [1] **Christian J. Steinmetz** [4] **Daniel Ritter** [1 5] **Erik Xie** [6] **Artem Gazizov** [1] **Aviv Spinner** [1] **Jonathan Frazer** [7] **Mafalda Dias** [7] **Pascal Notin** [1 8] **Debora S. Marks** [1 9]

Abstract

Disease-associated genetic variants occur extensively across the human genome, predominantly in noncoding regions like promoters. While crucial for understanding disease mechanisms, current methods struggle to predict effects of insertions and deletions (indels) that can disrupt gene expression. We present LOL-EVE (Language Of Life for Evolutionary Variant Effects), a conditional autoregressive transformer trained on 13.6 million mammalian promoter sequences. By leveraging evolutionary patterns and genetic context, LOL-EVE enables zero-shot prediction of indel effects in human promoters. We introduce three new benchmarks for promoter indel prediction: ultra rare variant prioritization, causal eQTL identification, and transcription factor binding site disruption analysis. LOL-EVE's dominate performance across these tasks suggests the potential of region-specific genomic language models for identifying causal non-coding variants in disease studies.

1. Introduction

DNA, the molecular language of life, has evolved for over 4 billion years under constant evolutionary pressure. Evolution through natural selection represents countless experiments continuously refining the genomic code to maximize organismal fitness. A fundamental challenge in computational biology is learning the mapping from genomic state to organism state—genotype to phenotype. Utilizing evolutionary sequences for unsupervised phenotype predictions is valuable as it enables assessment of mutational impacts without requiring prior knowledge of mechanisms or experimental validation. While there has been progress in predicting how protein variants affect phenotype (Frazer et al., 2021; Hopf et al., 2017; Orenbuch et al., 2023; Su et al., 2024; Notin et al., 2022; 2023), methods for predicting variant effects in non-coding regions remain underdeveloped.

Current approaches to variant effect prediction (VEP) in non-coding regions focus primarily on single nucleotide variants (SNVs) due to their ease of detection in whole-genome sequencing (Mullaney et al., 2010; Jiang et al., 2015). However, insertions and deletions (indels) represent an important but understudied source of genetic variation (Li et al., 2023). Individual SNVs typically have relatively low probability of large organismal effects, especially in non-coding regions, due to biological redundancy and smaller effect sizes (Kircher et al., 2014; Short et al., 2018; Zhu et al., 2017). Yet substantial heritability exists in promoter regions, suggesting that larger variants beyond SNVs likely drive these effects (Gazal et al., 2017; Finucane et al., 2015).

Furthermore, many methods have relied on expression or chromatin accessibility, which can be highly informative in specific biological contexts (Smedley et al., 2016), yet is often difficult and sometimes impossible to gather. As such, models that generalize to unseen variants and make accurate predictions in a zero-shot capacity, without requiring additional experimental data, provide tremendous value.

We hypothesize that expanding the scope of VEP to include indels, particularly in promoter regions, will lead to the discovery of variants with larger phenotypic effects (Zheng et al., 2024; Chiang et al., 2017). This approach will potentially identify overlooked sources of genetic variation with significant phenotypic impacts, contributing to a deeper understanding of rare and undiagnosed diseases.

2. Benchmarks

- We construct and open source, PromoterZoo, a dataset of **13.6 million promoter sequences** comprising almost 20 thousand 1kb promoter region sequences from

[1]Harvard Medical School, Boston, USA [2]Novo Nordisk A/S, Bagsværd, Denmark [3]University of Copenhagen, Copenhagen, Denmark [4]Queen Mary University of London, London, UK [5]Cornell University, Ithaca, NY, USA [6]Massachusetts Institute of Technology, Cambridge, MA, USA [7]Centre for Genomic Regulation, Barcelona, Spain [8]University of Oxford, Oxford, UK [9]Broad Institute, Cambridge, MA, USA. Correspondence to: Pascal Notin <$pascal_notin@hms.harvard.edu$>, Debora S. Marks <$debbie@hms.harvard.edu$>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

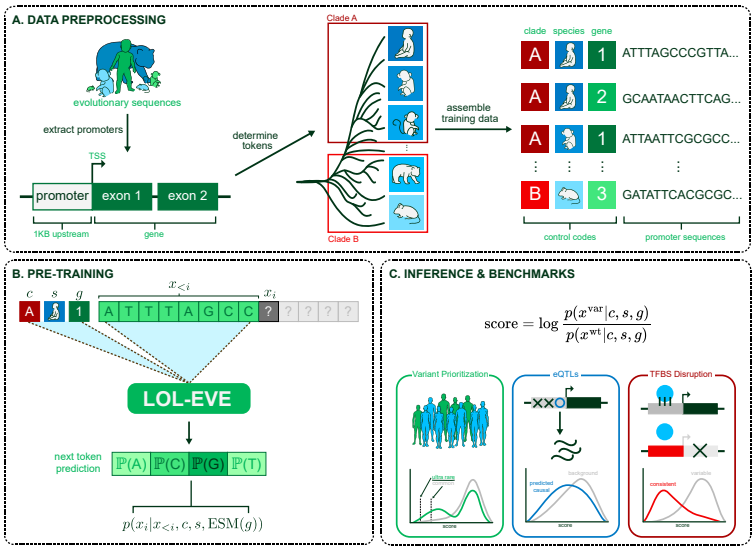

Figure 1: **LOL-EVE overview** A. Data preprocessing: Promoter sequences extracted from evolutionary sequences across mammals, grouped into clades and tokenized with control codes. B. Pre-training: Next-token prediction conditioned on sequence context and control codes. C. Inference & Benchmarks: Variant prioritization, eQTLs, and TFBS disruption tasks.

447 species across mammalian evolution identified in the Zoonomia project (Christmas et al., 2023) (§ A.1);

- We develop LOL-EVE, **a 235 million parameter conditional generative model of promoter evolution** for predicting variant effects (§ A);
- We introduce **three benchmarks** designed for zero-shot indel variant effect prediction in promoter regions, encompassing ultra rare indel detection, causal variant prioritization and TF binding site disruption (§ 2).

3. LOL-EVE

LOL-EVE learns a generative model over full promoter nucleotide sequences, conditioning its predictions on the promoter's most proximal gene, species, and clade (Figure 1). We implement this using a decoder-only transformer following the CTRL framework (Keskar et al., 2019), which allows the model to capture both broad evolutionary patterns and species-specific variations in regulatory elements.

Unlike LMs that use k-mer tokenization schemes (Dalla-Torre et al., 2023; Zhou et al., 2024), LOL-EVE tokenizes promoter sequences at base pair resolution. This design choice enables accurate handling of insertions and deletions without causing tokenization shifts in the remainder of the sequence—a critical capability for indel VEP.

To encode evolutionary context, we use three types of conditioning: (1) the most proximal gene g is encoded using mean-pooled ESM2 embeddings (Lin et al., 2023) projected to the embedding dimension, (2) species s and (3) clade c are encoded using learned embeddings. This allows LOL-EVE to model the autoregressive conditional distribution

$$p(x|c,s,g) = \frac{1}{L}\sum_{i=1}^{L} \log p(x_i|x_{<i},c,s,\mathrm{ESM}(g)) \quad (1)$$

We developed an adaptive local position embedding that treats control codes and genomic sequences differently, using absolute positions for control tokens and relative positions that reset at sequence start for genomic content. To prevent overfitting, we apply control tag dropout and strand-aware length dropout during training (see Appendix A.2).

At inference, we score variants with the log-likelihood ratio

$$\text{score} = \log \frac{p(x^{\text{var}}|c,s,g)}{p(x^{\text{wt}}|c,s,g)}, \quad (2)$$

which captures how likely the variant sequence is compared to wildtype given evolutionary patterns learned during pre-training. Training data, detailed hyperparameters, and additional implementation details are provided in Appendix A.

Given the lack of established benchmarks for promoter indel VEP, we developed a benchmark collection. To ensure rigorous comparisons, we maintain methodological consistency across all models, using standardized scoring approaches and identical evaluation pipelines without task-specific training. Details on scoring methodologies and potential data leakage in supervised models are provided in Appendix C.1.

3.1. Ultra Rare Variant Prioritization

Rationale Ultra rare variants (MAF < 0.0001) (Wang et al., 2021) are more likely to be functionally important

or disease-causing compared to common variants (MAF $>$ 0.001). Models that effectively identify deleterious variants should assign their most extreme predictions to variants in these ultra rare frequency ranges.

Task We evaluate how strongly models prioritize ultra rare variants (MAF $<$ 0.00001) versus common variants (MAF $>$ 0.001) by comparing their scores at each percentile cutoff. Specifically, for each percentile we take the ratio

$$\frac{\text{score}_{p,\text{ultra}}}{\text{score}_{p,\text{common}}}, \tag{3}$$

($>$ 1 means stronger ultra-rare signal), and we report the mean of these ratios stratified by indel-length category (Small $\leq$ 2 bp, Medium 3–10 bp, Large 11–100 bp).

Data GnomAD V4.0 variants are used (Chen et al., 2024).

3.2. Causal eQTL Prioritization

Rationale An expression quantitative trait locus (eQTL) is a variant associated with a change in gene expression. Fine-mapping methods such as SuSiE (Wang et al., 2020) assign each indel a posterior inclusion probability (PIP) reflecting its likelihood of being causal. We focus on *cis*-eQTLs—indels within promoter regions whose eGene is proximal.

Task Given two sets of promoter indels—putatively causal ($PIP > 0.99$) and background ($PIP < 0.01$)—models should assign larger absolute effect scores to the causal group. We assess discrimination by AUROC and AUPRC normalized by the causal-variant fraction.

Data We retrieved fine-mapped cis-eQTL indels from the eQTL Catalogue (Kerimov et al., 2021), filtered to promoters whose eGene matches the variant's nearest gene. Applying PIP thresholds of 0.99 and 0.01. For the cumulative-slippage analysis (see Section C.6), we compute running-mean AUROC and normalized AUPRC at slippage cutoffs of 25 bp, 50 bp, 100 bp, 200 bp, and $>$200 bp (Table A5).

3.3. TFBS Disruption

Rationale Transcription factors (TFs) are essential regulators of gene expression, binding to specific DNA sequences in promoter regions to control transcriptional activity. Disruptions to TF binding sites (TFBS) can impact gene regulation, with the severity depending on the evolutionary constraint and expression characteristics of the target gene. We hypothesize that variants disrupting TFBS should be most deleterious in genes that are both evolutionarily constrained and consistently expressed across tissues, as these genes are typically intolerant to regulatory perturbations (Wolf et al., 2023).

Task We evaluate whether models correctly predict that TFBS disruptions are more deleterious in genes with high evolutionary constraint and low expression variability compared to genes with low constraint and high variability. Performance is measured as delta accuracy across transcription factors using balanced sampling with the following setup:

Let $\mathcal{H}$ be the set of *high-constraint/low-variability* genes and $\mathcal{L}$ the set of *low-constraint/high-variability* genes (see Sec. C.5). For each transcription factor $t = 1, \ldots, T$, let $\text{Score}_t(\mathcal{G})$ be the model's mean disruption score across genes in set $\mathcal{G}$. We define

$$\Delta\text{Acc} := \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\big(\text{Score}_t(\mathcal{H}) < \text{Score}_t(\mathcal{L})\big) \; - \; 0.5, \tag{4}$$

where $\mathbf{1}(\cdot)$ is the indicator function. A positive ΔAcc means the model assigns lower disruption scores to the high-constraint set $\mathcal{H}$ more often than expected by chance.

Data Genes are categorized using mammalian evolutionary rates (OrthoDB) and expression variability (GTEx CV). TFBS disruptions are identified using JASPAR CORE TFs and position-specific scoring matrices. Complete methodology and data processing details are in Appendix C.5.

4. Results

We benchmark LOL-EVE against a diverse set of unsupervised DNA Language Models described in B, supervised predictors (CADD (Kircher et al., 2014) and Enformer (Avsec et al., 2021)), and conservation metrics (PhyloP). CADD integrates diverse genomic annotations to predict deleteriousness, while Enformer is trained to predict functional genomic signals. For LMs that make multiple checkpoints available, we focus our discussion on the best performing checkpoint in each experiment, with remaining checkpoints evaluated in section C.7. Scoring detail for each model can be found in B as well.

4.1. Ultra Rare Variant Prioritization

LOL-EVE delivers superior performance overall, achieving the highest enrichment for medium indels (2.150 ± 0.051) and ranking a close second for both small (1.482 ± 0.032) and large (1.956 ± 0.118) categories. GPN-Promoter, a masked-language transformer trained on promoter sequences, excels on small indels (2.297 ± 0.051) where its local-context objective is most effective. CADD again leads for large indels (2.055 ± 0.039), but this likely reflects data leakage from its use of allele-frequency annotations rather than true zero-shot generalization (Table A3). Generic LMs like NT-2.5b-multi and Caduceus-ph achieve only moderate ratios ($\approx$ 1.02–1.26), while DNABERT-2 and speciesLM remain near baseline.

Table 1: Mean-ratio and standard error for all models across indel length categories and percentiles (1%,2.5%,5%,10%).Best checkpoints are `medium-450k` (HyenaDNA), `ph-131k` (Caduceus), and `2.5B-1000G` (NT). See all models in Table A4. Variants/Genes per threshold are shown in in Table A5.

	Small (1-2bp)		Medium (3–10bp)		Large (11–100bp)	
Model	**Mean Ratio**	**Std. Error**	**Mean Ratio**	**Std. Error**	**Mean Ratio**	**Std. Error**
CADD	1.863	0.145	1.675	0.035	**2.055**	0.039
GPN-Promoter	**2.297**	0.051	1.912	0.031	1.456	0.072
Evo2	0.822	0.041	0.757	0.014	1.043	0.039
speciesLM	1.220	0.017	0.993	0.078	1.124	0.051
DNABERT-2	1.069	0.012	0.974	0.017	1.050	0.004
Caduceus-ph	1.028	0.045	1.116	0.074	1.253	0.172
NT-2.5b-multi	1.022	0.014	1.053	0.014	1.261	0.026
HyenaDNA-tiny	1.323	0.028	1.400	0.015	1.361	0.014
LOL-EVE (ours)	1.482	0.032	**2.150**	0.051	1.956	0.118

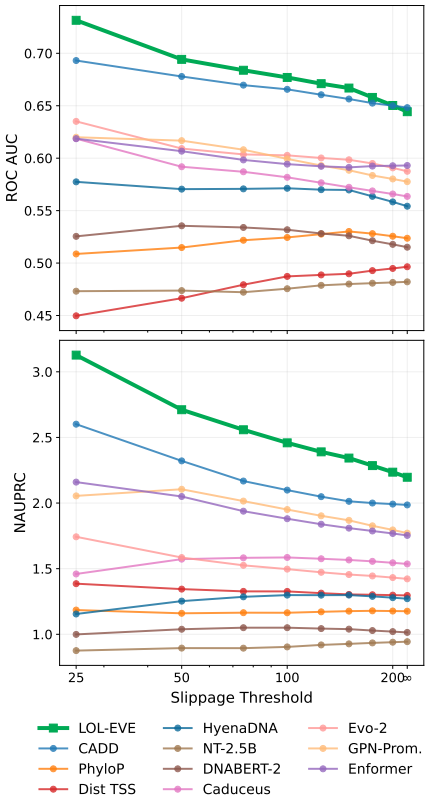

Figure 2: Cumulative causal-eQTL prioritization performance across slippage thresholds (log scale). Top: running mean ROC AUC; Bottom: running mean normalized AUPRC (AUPRC / baseline). Full models in A4

4.2. Causal eQTL Prioritization

Figure 2 shows cumulative ROC AUC and normalized AUPRC for causal versus background cis-eQTL indels as we include variants within increasing slippage cutoffs (C.6). LOL-EVE leads at every threshold—peaking near 0.73 ROC AUC and 3.1 × baseline AUPRC at 25 bp—and sustains the strongest separation even at large distances. CADD and Enformer follow closely, while generic LMs (e.g., NT-2.5B, DNABERT-2) achieve more modest gains. Importantly, LOL-EVE also generalizes to SNPs (Table A7), performing on par with the MLM model GPN-Promoter.

4.3. TFBS Disruption

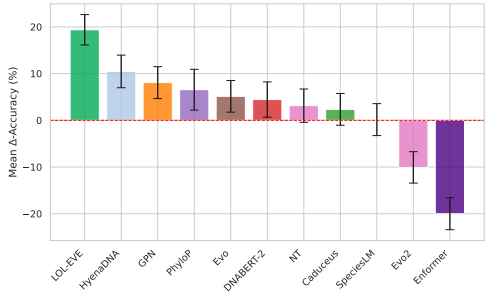

Figure 3: Mean delta-accuracy in TFBS disruption ($\pm SE$) for high-constraint/low-variability versus low-constraint/high-variability genes. Full results in A5, with gene counts per threshold in A6.

Figure 3 shows that LOL-EVE most accurately distinguishes TFBS disruptions in high-constraint, consistently expressed genes from those in low-constraint, variably expressed genes. For the greatest proportion of transcription factors, LOL-EVE correctly assigns lower disruption scores to the high-constraint set, reflecting their greater sensitivity to binding-site loss. This aligns with the expectation that variably expressed genes tolerate TFBS disruptions more readily than consistently expressed ones. By better capturing these differential sensitivities, LOL-EVE demonstrates superior predictive power for promoter variant impact.

5. Conclusion

Across three zero-shot benchmarks, ultra-rare indel enrichment, causal eQTL prioritization, and TFBS disruption, LOL-EVE consistently outperforms other unsupervised and conservation-based methods, demonstrating its ability to predict promoter indel effects without task-specific training. By modeling mammalian promoter evolution and local sequence context, LOL-EVE captures regulatory constraints that supervised predictors may overlook. In contrast, CADD and Enformer, while powerful, depend on population frequencies or cell-type–specific data that can introduce circularity or limit generalization to promoter indels. That said, LOL-EVE does not uniformly dominate every scenario—its gains are smaller on some variant classes and it can be outperformed by supervised methods when abundant labeled data are available—underscoring opportunities for future hybrid approaches. As the first model specifically designed for promoter indels, LOL-EVE may prove especially useful for variant prioritization in disease studies.

6. Impact Statement

This paper introduces LOL-EVE, a genomic language model designed for the prediction of promoter indel effects. By enabling more accurate identification of non-coding variants linked to gene regulation and disease, LOL-EVE has the potential to contribute to improvements in genetic research, variant interpretation, and clinical genomics. While the model's insights may aid in diagnosing rare and undiagnosed diseases, its predictions should be used in conjunction with experimental validation to avoid misinterpretation of genetic risk factors. Ethical considerations include the responsible use of genomic AI models to prevent potential biases or misapplications in clinical decision-making. As with any AI-driven approach, care must be taken to ensure equitable benefits across populations and to prevent misuse in genetic profiling. Nonetheless, this work primarily seeks to enhance computational methods for studying genome evolution and variant effects, with no foreseeable direct societal harm.

References

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x.

Benegas, G., Batra, S. S., and Song, Y. S. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023.

Benegas, G., Albors, C., Aw, A. J., Ye, C., and Song, Y. S. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, pp. 1–6, 2025a.

Benegas, G., Eraslan, G., and Song, Y. S. Benchmarking DNA sequence models for causal regulatory variant prediction in human genetics. *bioRxivorg*, pp. 2025.02.11.637758, March 2025b.

Benegas, G., Ye, C., Albors, C., Li, J. C., and Song, Y. S. Genomic language models: opportunities and challenges. *Trends in Genetics*, 2025c.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N. K., Adams, E., Baccus, S. A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A. X., Mehta, R., Mofrad, M. R. K., Ng, M. Y., Pannu, J., Re, C., Schmok, J. C., St. John, J., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A. B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D. P., Goodarzi, H., Hsu, P. D., and Hie, B. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, pp. 2025.02.18.638918, February 2025.

Chen, K. M., Wong, A. K., Troyanskaya, O. G., and Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.*, 54(7):940–949, July 2022.

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993):92–100, 2024.

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Pratt, T., Ziyatdinov, A., Maller, F. E., Ronning, C., et al. The impact of structural variation on human gene expression. *Nature Genetics*, 49(5):692–699, 2017.

Christmas, M. J., Kaplow, I. M., Genereux, D. P., Dong, M. X., Hughes, G. M., Li, X., Sullivan, P. F., Hindle, A. G., Andrews, G., Armstrong, J. C., Bianchi, M., Breit, A. M., Diekhans, M., Fanter, C., Foley, N. M., Goodman, D. B., Goodman, L., Keough, K. C., Kirilenko, B., Kowalczyk, A., Lawless, C., Lind, A. L., Meadows, J. R. S., Moreira, L. R., Redlich, R. W., Ryan, L., Swofford, R., Valenzuela, A., Wagner, F., Wallerman, O., Brown, A. R., Damas, J., Fan, K., Gatesy, J., Grimshaw, J., Johnson, J., Kozyrev, S. V., Lawler, A. J., Marinescu, V. D., Morrill, K. M., Osmanski, A., Paulat, N. S., Phan, B. N., Reilly, S. K., Schäffer, D. E., Steiner, C., Supple, M. A., Wilder,

A. P., Wirthlin, M. E., Xue, J. R., Zoonomia Consortium§, Birren, B. W., Gazal, S., Hubley, R. M., Koepfli, K.-P., Marques-Bonet, T., Meyer, W. K., Nweeia, M., Sabeti, P. C., Shapiro, B., Smit, A. F. A., Springer, M. S., Teeling, E. C., Weng, Z., Hiller, M., Levesque, D. L., Lewin, H. A., Murphy, W. J., Navarro, A., Paten, B., Pollard, K. S., Ray, D. A., Ruf, I., Ryder, O. A., Pfenning, A. R., Lindblad-Toh, K., and Karlsson, E. K. Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, 380(6643):eabn3943, April 2023.

Dalla-Torre, H., Gonzalez, L., Revilla, J. M., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., Sirelkhatim, H., Richard, G., Skwark, M., Beguir, K., Lopez, M., and Pierrot, T. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023. doi: 10.1101/2023.01.11.523679. URL https://www.biorxiv.org/content/early/2023/01/15/2023.01.11.523679.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015.

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., and Mathelier, A. Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 48(D1): D87–D92, 2020.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Structure-aware protein embedding using deep learning. *bioRxiv*, 2021.

Gankin, D., Alexander Karollus, Martin Grosshauser, Kristian Klemon, Johannes Hingerl, and Julien Gagneur. Species-aware DNA language modeling. *bioRxiv*, pp. 2023.01.26.525670, January 2023. doi: 10.1101/2023.01.26.525670. URL http://biorxiv.org/content/early/2023/01/27/2023.01.26.525670.abstract.

Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10): 1421–1427, 2017.

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., and Borgwardt, K. M. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.*, 36(5):513–523, May 2015.

Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, 29(10): 1341–1342, May 2013.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology*, 35(2):128–135, 2017.

Huang, C., Shuai, R. W., Baokar, P., Chung, R., Rastogi, R., Kathail, P., and Ioannidis, N. M. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55(12):2056–2059, 2023. doi: 10.1038/s41588-023-01574-w.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

Jiang, H., Ling, Y., Stella, A., Zhang, M. C., Narzisi, G., Hahn, W., Zody, M. C., Schatz, M. C., and Iossifov, I. Indel variant analysis of short-read sequencing data with scalpel. *Nature protocols*, 10(5):723–733, 2015.

Karollus, A., Hingerl, J., Gankin, D., Grosshauser, M., Klemon, K., and Gagneur, J. Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol.*, 25(1):83, April 2024.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, 28(5):739–750, May 2018.

Kerimov, N., Hayhurst, J., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samovici, I., McCarthy, D. J., Breschi, A., Zhang, X., et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics*, 53(9):1290–1299, 2021.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. CTRL: A conditional transformer language model for controllable generation. *arXiv [cs.CL]*, September 2019.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014.

Kuderna, L. F. K., Ulirsch, J. C., Rashid, S., Ameen, M., Sundaram, L., Hickey, G., Cox, A. J., Gao, H., Kumar, A., Aguet, F., Christmas, M. J., Clawson, H., Haeussler, M., Janiak, M. C., Kuhlwilm, M., Orkin, J. D., Bataillon, T., Manu, S., Valenzuela, A., Bergman, J., Rouselle, M., Silva, F. E., Agueda, L., Blanc, J., Gut, M., de Vries, D., Goodhead, I., Harris, R. A., Raveendran, M., Jensen, A., Chuma, I. S., Horvath, J. E., Hvilsom, C., Juan, D., Frandsen, P., Schraiber, J. G., de Melo, F. R., Bertuol, F., Byrne, H., Sampaio, I., Farias, I., Valsecchi, J., Messias, M., da Silva, M. N. F., Trivedi, M., Rossi, R., Hrbek, T., Andriaholinirina, N., Rabarivola, C. J., Zaramody, A., Jolly, C. J., Phillips-Conroy, J., Wilkerson, G., Abee, C., Simmons, J. H., Fernandez-Duque, E., Kanthaswamy, S., Shiferaw, F., Wu, D., Zhou, L., Shao, Y., Zhang, G., Keyyu, J. D., Knauf, S., Le, M. D., Lizano, E., Merker, S., Navarro, A., Nadler, T., Khor, C. C., Lee, J., Tan, P., Lim, W. K., Kitchener, A. C., Zinner, D., Gut, I., Melin, A. D., Guschanski, K., Schierup, M. H., Beck, R. M. D., Karakikes, I., Wang, K. C., Umapathy, G., Roos, C., Boubli, J. P., Siepel, A., Kundaje, A., Paten, B., Lindblad-Toh, K., Rogers, J., Marques Bonet, T., and Farh, K. K.-H. Identification of constrained sequence elements across 239 primate genomes. *Nature*, November 2023.

Levy, B., Xu, Z., Zhao, L., Kremling, K., Altman, R., Wong, P., and Tanner, C. FloraBERT: cross-species transfer learning withattention-based neural networks for geneexpression prediction. preprint, In Review, August 2022. URL https://www.researchsquare.com/article/rs-1927200/v1.

Li, S., Consortium, U. B. W.-G. S., Carss, K. J., Halldorsson, B. V., and Cortes, A. Whole-genome sequencing of half-a-million uk biobank participants. *medRxiv*, pp. 2023–12, 2023.

Li, Z., Subasri, V., Stan, G.-B., Zhao, Y., and Wang, B. Gv-rep: A large-scale dataset for genetic variant representation learning, 2024. URL https://arxiv.org/abs/2407.16940.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Livesey, B. J. and Marsh, J. A. Variant effect predictor correlation with functional assays is reflective of clinical classification performance. *bioRxiv*, May 2024.

Marin, F. I., Teufel, F., Horlacher, M., Madsen, D., Pultz, D., Winther, O., and Boomsma, W. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uKB4cFNQFg.

Mullaney, J. M., Mills, R. E., Pittard, W. S., and Devine, S. E. Small insertions and deletions (indels) in human genomes. *Human molecular genetics*, 19(R2):R131–R136, 2010.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., Ermon, S., Ré, C., and Baccus, S. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 43177–43201. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/86ab6927ee4ae9bde4247793c46797c7-Paper-Conference.pdf.

Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., Ng, M. Y., Lewis, A., Patel, A., Lou, A., Ermon, S., Baccus, S. A., Hernandez-Boussard, T., Ré, C., Hsu, P. D., and Hie, B. L. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, pp. 2024.02.27.582234, March 2024.

Notin, P., Niekerk, L. V., Kollasch, A. W., Ritter, D., Gal, Y., and Marks, D. S. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022. URL https://openreview.net/forum?id=l7Oo9DcLmR1.

Notin, P., Kollasch, A. W., Ritter, D., Niekerk, L. V., Paul, S., Spinner, H., Rollins, N. J., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., and Marks, D. S. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=URoZHqAohf.

Orenbuch, R., Kollasch, A. W., Spinner, H. D., Shearer, C. A., Hopf, T. A., Franceschi, D., Dias, M., Frazer, J., and Marks, D. S. Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. *Medrxiv*, 2023.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. Detection of nonneutral substitution rates on mam-

malian phylogenies. *Genome research*, 20(1):110–121, 2010.

Sasse, A., Ng, B., Spiro, A. E., Tasaki, S., Bennett, D. A., Gaiteri, C., De Jager, P. L., Chikina, M., and Mostafavi, S. Benchmarking of deep neural networks for predicting personal gene expression from dna sequence highlights shortcomings. *Nature Genetics*, 55(12):2060–2064, 2023. doi: 10.1038/s41588-023-01524-6.

Schiff, Y., Kao, C. H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43632–43648. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

Schubach, M., Maass, T., Nazaretyan, L., Röner, S., and Kircher, M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.*, 52(D1):D1143–D1154, January 2024.

Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., Wright, C. F., Firth, H. V., FitzPatrick, D. R., Barrett, J. C., et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, 555 (7698):611–616, 2018.

Smedley, D., Schubach, M., Jacobsen, J. O., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *American Journal of Human Genetics*, 99(3):595–606, 2016.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6MRm3G4NiU.

Vilov, S. and Heinig, M. Investigating the performance of foundation models on human 3'utr sequences. *bioRxiv*, pp. 2024–02, 2024.

Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S. V. V., Mackay, A., Muthas, D., Hühn, M., Monkley, S., Olsson, H., AstraZeneca Genomics Initiative, Wasilewski, S., Smith, K. R., March, R., Platt, A., Haefliger, C., and Petrovski, S. Rare variant contribution to human disease in 281,104 UK biobank exomes. *Nature*, 597(7877):527–532, September 2021.

Wang, Y., Pritchard, J. K., and Stephens, M. Simple new approaches to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82 (5):1273–1300, 2020.

Wittkopp, P. J. and Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.*, 13(1):59–69, December 2011.

Wolf, S., Melo, D., Garske, K. M., Pallares, L. F., Lea, A. J., and Ayroles, J. F. Characterizing the landscape of gene expression variance in humans. *PLoS genetics*, 19(7): e1010833, 2023.

Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I. M., et al. Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics*, pp. 1–11, 2024.

Zhou, J., Theesfeld, C., Yao, K., Chen, K., Wong, A., and Troyanskaya, O. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 15(8):541–548, 2018.

Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R. V., and Liu, H. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oMLQB4EZE1.

Zhu, X., Li, M., Pan, H., Bao, X., Zhang, J., and Wu, X. Whole-genome sequencing in a family with twin boys with autism and intellectual disability suggests multiallelic inheritance. *Molecular autism*, 8(1):39, 2017.

Appendix

A. Model details

A.1. Training Data Construction

Promoters and other regulatory regions generally evolve faster than protein-coding sequences, as regulatory changes can often be more easily tolerated than changes to protein structure and function (Wittkopp & Kalay, 2011). To capture these evolutionarily relevant regulatory signals, particularly those that have evolved recently, we focused on training data from mammals. We curated a promoter dataset across 447 diverse species from the Zoonomia project (Christmas et al., 2023; Kuderna et al., 2023).

Transcription Start Site (TSS) annotations, which are often used to infer promoter regions, are not readily available for most species in our dataset due to several factors. Many of the 447 species lack comprehensive genome annotations, particularly for regulatory regions like promoters. Even in well-annotated species, TSS and promoter definitions can vary significantly across different databases and research groups. To address this, we employed a comparative genomics approach to identify putative promoter regions, leveraging sequence similarity to the first exon of 19,254 protein-coding genes from the NCBI RefSeq human genome annotation (assembly GRCh38.p14, annotation release 109). This strategy allowed us to consistently infer promoter regions across species by aligning known human exonic regions to homologous exons in other species, then extracting sequences upstream of the start of the first exon (which we define as the putative TSS). It's important to note that no genome has "promoter annotations" as such; rather, we use these inferred TSS positions and their upstream sequences as proxies for promoter regions. Importantly, in the human annotations we utilized, the 5'UTR often overlaps with the annotations for exon 1, which influences our definition of putative promoter regions across species.

Using the HAL toolkit (Hickey et al., 2013), we performed a liftover of these exon coordinates to each species in the Zoonomia project. For each species, exons were retained if their length was at least 50% of the length of the corresponding human exon. This threshold ensured that conserved regions were captured while excluding regions where the alignment is unreliable.

To define promoter regions, we extracted the 1,000 base pairs upstream of each exon start, accounting for the strand orientation of the gene. If the upstream region overlapped with the neighboring gene body, we shortened the promoter region to avoid misclassifying coding regions or intergenic space as promoters. This conservative approach minimized the risk of including non-promoter sequences but may exclude more distal regulatory elements, a potential caveat of the 1,000 bp window approach. Additionally, in cases where promoter regions from neighboring genes were within 100 base pairs of each other, we merged the coordinates. This merging process ensured that promoter regions were not artificially fragmented due to closely spaced genes.

Including reverse complements, this resulted in a dataset of 13.6 million sequences. We employed a chromosome-wise split for development, with chromosome 19 used for validation. Promoters from non-human species were assigned to the respective set based on the chromosome of the human gene used for liftover, thereby ensuring that all instances of a gene are placed in the same partition and no gene information leakage between the training and validation set.

A.1.1. Training Data Validation

To gain further insight into the validity of the upstream 1,000 bp approach, we scored all extracted sequences using the Sei promoter score (Chen et al., 2022), which is trained on functional genomics data from humans. Despite Sei being human-based, we found that the promoter scores generalize well across species, showing strong conservation of regulatory elements in many mammalian species. Notably, promoters from species closely related to humans, such as other primates, tend to have higher Sei scores, indicating similar promoter activity, while more distant species still retain significant functional signal, suggesting that core regulatory sequences are preserved across mammals (A1). Further we assessed how the Sei score distributions for 3 groups: Human (CoDing Sequence) CDS regions, Human promoters, and our training data compare in (A2). Our training data promoter distribution aligns more closely with the raw Human promoters than the Human CDS regions.

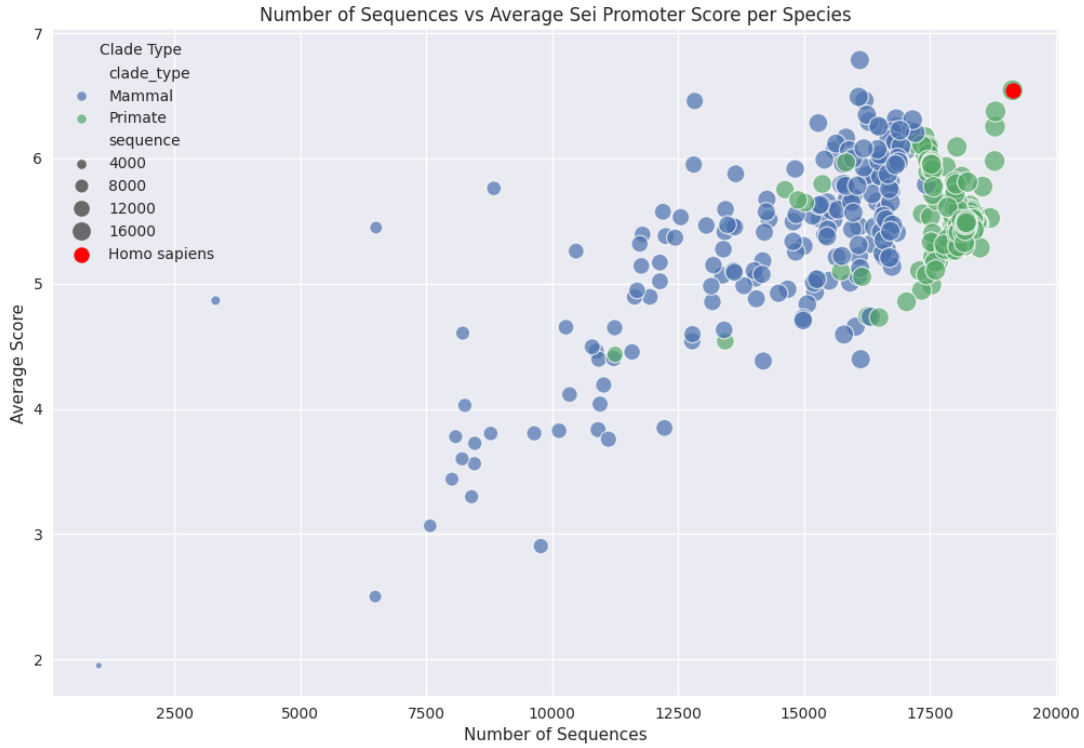

Figure A1: Average Promoter Sei scores plotted against the number of promoter sequences gathered for model training from the comparative genomics analysis conducted with the HAL suite. Clade types are specified by color and the red dot represents Homo sapiens. The maximum number of sequences per species is 19,254. Point sizes reflect the number of sequences.

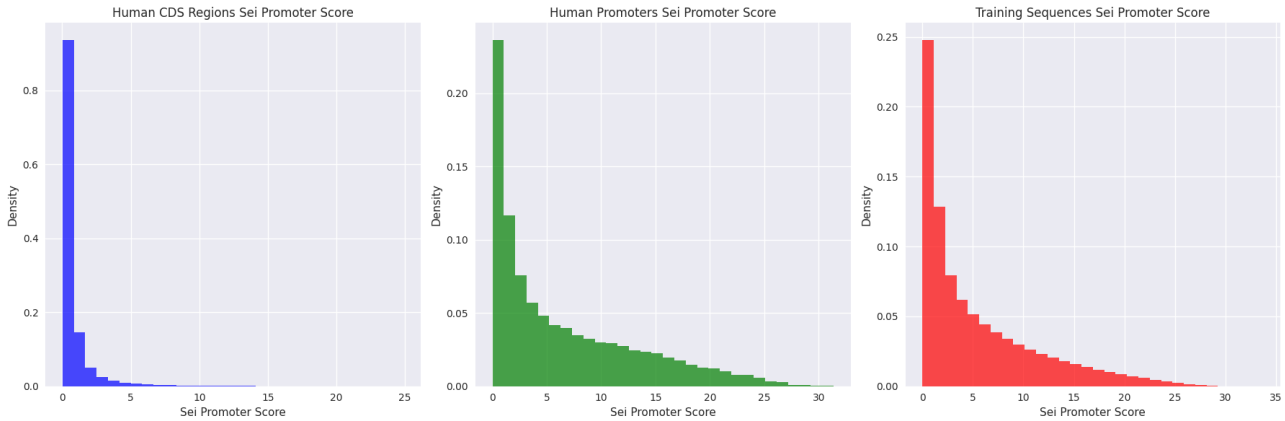

Figure A2: Average Promoter Sei scores were plotted for Human CDS regions, Human promoter regions, and all of the promoter data used gathered for training.

A.2. Training Strategies and Augmentation

To prevent overfitting and improve model generalization, we apply several data augmentation strategies during training:

A.2.1. CONTROL TAG DROPOUT

We apply control tag dropout to encourage the model to learn representations that are robust to the presence of control tags and mitigate sequence memorization:

$$\begin{aligned}\mathcal{L}(D) = -\sum_{k=1}^{|D|} \log p_\theta(x_i^k \mid x_{<i}^k, c^k, s^k, g^k, \\ m^k \odot [c^k, s^k, g^k]) \\ \text{where } m^k \sim \text{Bernoulli}(p)\end{aligned} \tag{5}$$

A.2.2. STRAND-AWARE LENGTH DROPOUT

To account for the inherent directionality of DNA sequences, we implement a strand-aware length dropout mechanism. For sequences on the forward strand (d = 1), tokens are shifted leftward after dropping out l tokens from the right end, maintaining causal attention over the remaining sequence. For reverse strand sequences (d = -1), tokens are simply dropped from the right end without shifting, preserving the natural 5' to 3' processing order. In both cases, dropped tokens are replaced with padding tokens that are ignored in self-attention layers, and the maximum dropout length is capped at 90% of the sequence length to ensure sufficient context is retained for prediction.

$$\begin{aligned}\mathcal{L}(D) = -\sum_{k=1}^{|D|} \log p_\theta(x_i^k \mid x_{<i}^k, c^k, s^k, g^k, d^k) \\ \text{where } l^k \sim \text{Uniform}(0, 0.9|x^k|) \\ \text{and } d^k \in \{-1, 1\} \text{ indicates strand direction}\end{aligned} \tag{6}$$

A.2.3. ADAPTIVE LOCAL POSITION EMBEDDING

To better capture the distinct roles of control codes and genomic sequences, we developed an adaptive local position embedding scheme defined as:

$$\mathbf{p}_i = \begin{cases} \mathbf{p}_i^{\text{ctrl}} & \text{if } i \in [0, 3] \text{ (control tokens)} \\ \mathbf{p}_{i-j_{\text{SOS}}}^{\text{seq}} & \text{if } i \geq j_{\text{SOS}} \text{ (sequence tokens)} \end{cases}$$

where $\mathbf{p}_i^{\text{ctrl}} \in \mathbb{R}^d$ are absolute position embeddings for control tokens and $\mathbf{p}_{i-j_{\text{SOS}}}^{\text{seq}} \in \mathbb{R}^d$ are relative position embeddings that reset at the sequence start token position j_{SOS}. This adaptive approach allows the model to maintain structural understanding of control codes while enabling biologically meaningful positional representations for genomic sequences.

A.2.4. HYPERPARAMETERS

Table A1: The hyperparameters of the LOL-EVE model.

Hyperparameter	Value
Dimension	768
Layers	12
Heads	12
Feedforward dimension	8192
Learning rate	$1e^{-5}$
Batch size	32
Steps	150,00

B. Baseline details

Methods for modeling genomic sequences can be broadly classified as alignment-free or alignment-based for functional constraints, activity predictors, and meta-predictors.

Alignment-free methods A growing number of unsupervised language models (LMs) for eukaryotic genomic DNA have been proposed, including DNABERT (Ji et al., 2021; Zhou et al., 2024), Nucleotide Transformer (Dalla-Torre et al., 2023), HyenaDNA (Nguyen et al., 2023), and Caduceus (Schiff et al., 2024). While they have some differences in their architectures, training objectives, and training data, these models are all fully unsupervised and trained only on genome-wide data (Benegas et al., 2025c). While LMs have shown utility in some downstream prediction tasks, their performance in variant effect prediction varies. Independent benchmarks have revealed that models trained on genome-wide data learn different aspects of the genome to varying extents, sometimes focusing on splice site patterns and other times on regulatory elements, in ways that are difficult to anticipate (Marin et al., 2024; Li et al., 2024).

An alternative approach involves specialized LMs trained on local genomic regions, such as plant promoters or fungal 5' and 3' regions (Levy et al., 2022; Gankin et al., 2023). These models reliably capture regulatory motifs and learn embeddings useful for downstream tasks. Recently, Vilov & Heinig (2024) proposed and evaluated several 3'UTR-specific language models for the human genome. Their study showed that these region-specific models often outperformed genome-wide models and even conservation-based approaches like PhyloP (Pollard et al., 2010) on various tasks, including variant effect prediction.

Alignment-based methods Multiple sequence alignments (MSAs) offer a powerful approach to understanding natural sequence variation, enabling the identification of potentially non-neutral mutations with likely functional consequences. PhyloP is an MSA-based statistical method that assigns a conservation score to each position in a sequence and compares observed substitutions to those expected under a neutral evolution model. GPN-MSA (Benegas et al., 2025a), a more recent development, combines whole-genome alignments with a genomic LM approach. Trained to reconstruct masked nucleotides given an MSA as input, GPN-MSA has shown improvement in SNV effect prediction compared to PhyloP. However, a major limitation of alignment-based approaches is their treatment of positions individually, which doesn't naturally generalize to indel variants.

Activity Predictors & Meta Predictors An alternative approach to unsupervised modeling of sequences involves training supervised models on measurements of sequence activity. These models often use data from high-throughput functional genomics experiments that measure various aspects of genomic function, such as expression initiation or epigenetic modifications. Models like Enformer (Avsec et al., 2021) have demonstrated an understanding of factors contributing to gene expression in different cell types. However, recent studies by Sasse et al. (Sasse et al., 2023) and Huang et al. (Huang et al., 2023) have shown that the performance of sequence-to-activity models such as DeepSEA (Zhou et al., 2018), Basenji2 (Kelley et al., 2018), and Enformer (Avsec et al., 2021) in explaining expression variation between individuals due to cis-regulatory genetic variants remains limited. Another widely used method, CADD (Combined Annotation Dependent Depletion), integrates numerous genomic annotations into a single deleteriousness score(Schubach et al., 2024). However, (Grimm et al., 2015) and (Livesey & Marsh, 2024) have demonstrated, comparative evaluations of meta predictors like CADD are complicated by circularity issues in their training and testing datasets leading to data leakage. As such, their

performance is likely inflated due to circularity. These findings underscore the need for zero shot methods to overcome these limitations and enhance our understanding of genetic variant effects in humans.

B.1. Autoregressive models

Autoregressive LMs assign scores to sequences s using their log likelihood

$$p(s) = \frac{1}{n}\sum_{i=1}^{n} \log p(s_i|s_{<i}). \tag{7}$$

HyenaDNA HyenaDNA uses base pair tokenization. For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$ and exclude special tokens. We ignore the final `EOS` position when taking the mean over the sequence.

Evo1 Evo1 from (Nguyen et al., 2024) For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$ and exclude special tokens. We do not apply any masking.

Evo2 Evo2 from (Brixi et al., 2025) 7b version. For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$ and exclude special tokens. We do not apply any masking.

B.2. Masked language models

For computational efficiency, we evaluate bidirectional masked LMs using their pseudo log likelihood,

$$p(s) = \frac{1}{n}\sum_{i=1}^{n} \log p(s_i|s). \tag{8}$$

Caduceus Caduceus uses base pair tokenization. For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$ and exclude special tokens. We do not apply any masking.

Nucleotide Transformer Nucleotide Transformer uses 6-mer tokenization. For computing the cross entropy, we subset the logits and labels to the dimensions of the 6-mer and five trailing single-base tokens and exclude special tokens. We do not apply any masking.

DNABERT-2 DNABERT-2 uses byte pair tokenization. For computing the cross entropy, we subset the logits and labels to the dimensions of the BPE tokens and the `[UNK]` token which represents N. Remaining special tokens are excluded. We do not apply any masking.

BEND - GPN The original GPN model (Benegas et al., 2023) was only trained on *Brassicales* species and is not applicable to the human genome. We instead evaluate a human GPN-based model ("Dilated ResNet") that is included in the BEND benchmark (Marin et al., 2024). For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$. We do not apply any masking.

Promoter-GPN The original Promoter-GPN model from (Benegas et al., 2025b) For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$. We do not apply any masking.

Species-LM Species-LM from (Karollus et al., 2024) metazoa version. For computing the cross entropy, we subset the logits and labels to the dimensions of actual nucleotides $x \in \{A, G, C, T, N\}$ and exclude special tokens. We do not apply any masking.

B.3. Alignment-based approaches

PhyloP As they are based on an MSA, PhyloP scores are not naturally amenable to indel variants, as a change in sequence length by insertion or deletion cannot be modeled by column-wise scores. We follow gnomAD's approach to computing

PhyloP scores: For any indel, the PhyloP score of the position in the reference genome at which the indel occurs is used for the indel as a whole. Note that this inherently does not consider the actual sequence consequence of the indel - it only reflects the conservation of the position at which the indel occurs.

B.4. Activity predictors

Enformer We run Enformer following the official notebook[1]. For each variant, we compute the mean difference over the sequence between the wild type and variant sequence using all human output tracks. We report the max channel to capture the largest change between the wildtype and the variant sequence.

$$S_{\text{max}}(s) = \max_{c \in C} \left(\frac{1}{L} \sum_{i=1}^{L} [p_c(s_i^{\text{alt}}) - p_c(s_i^{\text{ref}})] \right) \tag{9}$$

where:

C is the set of all human output tracks in Enformer L is the length of the output sequence $p_c(\cdot)$ is the Enformer prediction for track c s_i^{ref} and s_i^{alt} are the reference and alternate sequences at position i

Our Enformer evaluation computes the mean difference between wild-type and variant sequences across all human output tracks, taking the maximum as the final score. This methodology captures the maximum regulatory impact across all potential regulatory mechanisms and cell types, which is particularly important for promoter indels that may affect multiple regulatory processes simultaneously and manifest differently across diverse cellular contexts. This approach provides a more holistic assessment than the GTEx-focused SLDP regression used in the original Enformer paper, analogous to organism-scale models that consider effects across all tissue types rather than focusing on single expression outputs.

B.5. Meta Predictors

CADD Combined Annotation Dependent Depletion (CADD)(Kircher et al., 2014) provides a deleteriousness score across the whole genome by integrating genomic annotations and functional information, including in-silico predictions from other models. It is one of the first models to provide predictions for all single-nucleotide variants and short indels and is therefore frequently used by the community, particularly the clinical community. Of particular relevance for this work, CADD trains on population data (gnomAD frequencies), expression data (ENCODE RNAseq and epigenetic markers), transcription factor binding site annotations (ChIP transcription factor binding sites), and clinical annotations (indirectly, through training on PolyPhen2, which was itself directly trained on ClinVar labels). More information about exact features trained on can be found here: CADD features.

[1] https://github.com/google-deepmind/deepmind-research/blob/master/enformer/enformer-usage.ipynb

C. Extended Benchmark Details

C.1. Benchmark Implementation Details

C.1.1. Scoring Methodologies

To ensure fair comparisons across all models, we implement standardized scoring approaches detailed below. All models are evaluated without task-specific training or fine-tuning, though some supervised models may have been exposed to task-relevant data during their original training.

C.1.2. Data Leakage Assessment

While most models operate in a true zero-shot capacity, some supervised models in our evaluation have been previously exposed to task-relevant data during their training. Table A3 shows potential data leakage that occurs in supervised models for each benchmark. CADD was not used for the TFBS benchmarks due to lack of coverage.

Table A2: Training Data Leakage for Benchmark Tasks

Model	Ultra Rare Variant	Causal eQTL	TFBS Disruption
LOL-EVE	-	-	-
CADD	Population frequencies, ClinVar	ENCODE, RNA-seq	N/A
Enformer	-	RNA-seq	ChIP-seq, RNA-seq
DNABERT-2	-	-	-
NT	-	-	-
HyenaDNA	-	-	-
PhyloP	-	-	-

C.2. Ultra Rare Variant Prioritization Details

For each length category:

1. **Length bins & weights.** Partition indel lengths into 10 logarithmically spaced bins and compute the empirical bin weights w_i.

2. **Percentiles.** For each bin i and percentile $p \in \{1, 2.5, 5, 10\}\%$, compute

$$\tau_{i,p}^{(\mathcal{U})} = \text{the } p\text{th percentile of scores for } \{j : \mathrm{MAF}_j < 10^{-5},\, \ell_j \in \text{bin } i\},$$

$$\tau_{i,p}^{(\mathcal{C})} = \text{the } p\text{th percentile of scores for } \{j : \mathrm{MAF}_j \geq 10^{-3},\, \ell_j \in \text{bin } i\}.$$

3. **Safe ratio.**

$$r_{i,p} = \max\Big(1,\ \frac{\tau_{i,p}^{(\mathcal{U})}}{\tau_{i,p}^{(\mathcal{C})}}\Big).$$

4. **Weighted mean per percentile.**

$$R_p = \sum_{i=1}^{10} w_i\, r_{i,p}\,.$$

5. **Aggregate.** Report

$$\bar{R} = \frac{1}{P}\sum_{p} R_p \quad \text{with} \quad \mathrm{SE} = \sqrt{\frac{\mathrm{Var}(R_p)}{P}},$$

where $P = 4$ is the number of percentiles.

C.3. Causal eQTL Prioritization Details

C.4. Running-Mean Metric Computation

For each slippage cutoff s, we restrict to all indels with distance $\leq s$. Within that subset we compute

$$\mathrm{ROC}_s = \mathrm{AUROC}\big(\{|\hat{e}_j|\}\big), \quad \mathrm{nAUPRC}_s = \frac{\mathrm{AUPRC}(\{|\hat{e}_j|\})}{\text{baseline AUPRC}},$$

where $\hat{e}_j$ is the model's effect-score for variant j. Plotting ROC_s and nAUPRC_s against s (log-spaced) yields the cumulative performance curves in Fig. 2.

C.5. TFBS Disruption Detailed Methodology

C.5.1. Gene Stratification

We classified genes into two extreme groups using (1) evolutionary constraint—amino-acid substitution rates inferred from OrthoDB mammalian orthologs—and (2) expression variability—CV of GTEx median-TPM across tissues. "High-constraint/low-variability" genes occupy the bottom percentile in both metrics; "low-constraint/high-variability" genes occupy the top percentile. We tested robustness at 20–40% cutoffs.

C.5.2. TFBS Disruption Scoring

We sourced human TF motifs from JASPAR CORE (Fornes et al., 2020) and retained TFs with median $TPM > 1$ in ≥ 30 GTEx tissues. Promoter sequences were scanned with PSSMs ($threshold > 0.8$) to identify binding sites; *in silico* deletions were generated, and a site was deemed "disrupted" if its post-deletion PSSM score fell below 0.8.

C.5.3. Balanced Comparison & Statistics

For each TF, we (a) randomly sampled equal numbers of genes from each category, (b) computed disruption scores for their TFBSs, and (c) assessed separation via point-biserial correlation. P values were FDR-corrected across TFs. Finally, we report "delta accuracy" as the fraction of TFs for which high-constraint/low-variability genes scored lower (more deleterious) minus 50% baseline.

C.6. Slippage Calculation Methodology

C.6.1. Rationale

DNA slippage events during replication can lead to insertions and deletions, particularly in regions with repetitive sequences or secondary structures. Understanding the relationship between model predictions and slippage propensity provides insight into whether models are learning biologically relevant mutational mechanisms versus purely statistical patterns.

C.6.2. Slippage Score Calculation

We implement a computational approach to estimate slippage propensity for each indel variant based on local sequence context and repetitive elements.

Repeat Detection Algorithm For each variant, we extract a 20 base pair window centered on the variant position and analyze it for repetitive elements using the following approach:

1. **Homopolymer Run Detection**: We identify consecutive runs of identical nucleotides with a minimum length of 3 bases. Each homopolymer run contributes to the slippage score with a weight proportional to the square of its length.
2. **Short Tandem Repeat Detection**: We systematically search for dinucleotide, trinucleotide, and tetranucleotide repeats by:
    - Scanning the sequence with sliding windows of size 2, 3, and 4 nucleotides
    - Counting consecutive occurrences of each repeat unit
    - Requiring a minimum of 3 repeat units for classification as a tandem repeat

3. **Variant-Repeat Matching**: For each detected repeat, we check whether:
    - The deleted sequence (for deletions) matches or contains the repeat unit
    - The inserted sequence (for insertions) matches or contains the repeat unit
    - The variant position falls within the boundaries of a repeat region

Slippage Score Computation The final slippage score combines contributions from all detected repeats:

$$\text{Slippage Score} = \sum_{\text{homopolymers}} L^2 + \sum_{\text{STRs}} (C \times U)^{1.5} \times W$$

where:

- L = length of homopolymer run
- C = count of repeat units in short tandem repeat (STR)
- U = length of repeat unit
- W = weight factor: 0.8 for dinucleotides, 0.6 for trinucleotides, 0.5 for tetranucleotides

This scoring scheme gives higher weights to homopolymer runs and progressively lower weights to longer repeat units, reflecting the relative propensity for slippage in different repeat contexts.

Implementation Details

- Window size: 20 base pairs centered on variant position
- Minimum repeat threshold: 3 consecutive units
- Repeat unit sizes analyzed: 1-4 nucleotides
- Variants are classified as slippage-prone if they occur within or match any detected repeat region

This methodology allows us to quantitatively assess whether model predictions correlate with known mechanisms of indel formation, helping to distinguish between models that learn genuine biological constraints versus those that primarily capture mutational biases.

C.7. Extended Results

Table A3: Training Data Leakage for Benchmark Tasks

Model	Ultra Rare Variant	Causal eQTL	TFBS Disruption
LOL-EVE	-	-	-
CADD	Population frequencies, ClinVar	ENCODE, RNA-seq	N/A
Enformer	-	RNA-seq	ChIP-seq, RNA-seq
DNABERT-2	-	-	-
NT	-	-	-
HyenaDNA	-	-	-
PhyloP	-	-	-
GPN-Promoter	-	-	-
Caduceus	-	-	-
speciesLM	-	-	-
Evo1/2	-	-	-

C.7.1. Ultra Rare Variant Prioritization

	Small (1-2bp)		Medium (3–10bp)		Large (11–100bp)	
Model	**Mean Ratio**	**Std. Error**	**Mean Ratio**	**Std. Error**	**Mean Ratio**	**Std. Error**
CADD	1.863	0.145	1.675	0.035	**2.055**	0.039
LOL-EVE	1.482	0.032	**2.150**	0.051	1.956	0.118
GPN-Promoter	**2.297**	0.051	1.912	0.031	1.456	0.072
HyenaDNA-tiny	1.323	0.028	1.400	0.015	1.361	0.014
HyenaDNA-small	1.283	0.016	1.361	0.012	1.214	0.024
HyenaDNA-medium-160k	1.373	0.011	1.431	0.025	1.214	0.029
HyenaDNA-medium-450k	1.335	0.025	1.369	0.031	1.222	0.026
HyenaDNA-large	1.345	0.018	1.352	0.022	1.173	0.025
GPN	1.198	0.020	1.283	0.078	1.281	0.058
NT-v2-500m	1.131	0.024	1.180	0.028	1.272	0.035
NT-500m	1.059	0.021	1.147	0.005	1.170	0.026
NT-2.5b-multi	1.022	0.014	1.053	0.014	1.261	0.026
NT-2.5b-1000g	1.013	0.017	1.093	0.031	1.206	0.033
Caduceus-ps	0.943	0.018	1.062	0.046	1.258	0.119
Caduceus-ph	1.028	0.045	1.116	0.074	1.253	0.172
DNABERT-2	1.069	0.012	0.974	0.017	1.050	0.004
PhyloP	1.067	0.037	1.044	0.012	1.203	0.025
speciesLM	1.220	0.017	0.993	0.078	1.124	0.051
Evo1	1.262	0.033	1.380	0.114	1.071	0.040
Evo2	0.822	0.041	0.757	0.014	1.043	0.039
Enformer	1.000	0.000	1.000	0.000	1.000	0.000
GC Content	1.005	0.001	1.036	0.011	0.893	0.013
Distance TSS	1.000	0.000	0.897	0.004	0.793	0.118

Table A4: Mean-ratio and standard error for all models across indel length categories and percentiles

	Ultra-rare Var			Common Var			Ultra-rare Gene			Common Gene		
Percentile	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
1.0%	13.0	7.2	4.0	3.0	3.0	2.0	11.8	6.7	3.4	2.8	2.8	1.6
2.5%	31.1	16.0	9.0	8.0	6.0	3.0	28.1	14.8	7.9	7.2	5.3	2.2
5.0%	61.0	32.0	17.2	15.0	11.0	6.0	54.2	29.4	14.0	13.2	9.4	4.0
10.0%	121.0	63.0	34.0	29.0	21.0	11.0	105.2	57.0	27.9	24.9	17.7	6.1

Table A5: Average counts per model of variants (Var) and genes (Gene), stratified by rarity (ultra-rare vs common), percentile, and indel size. No model reported zero counts in any of these splits.

C.7.2. Causal eQTL Prioritization

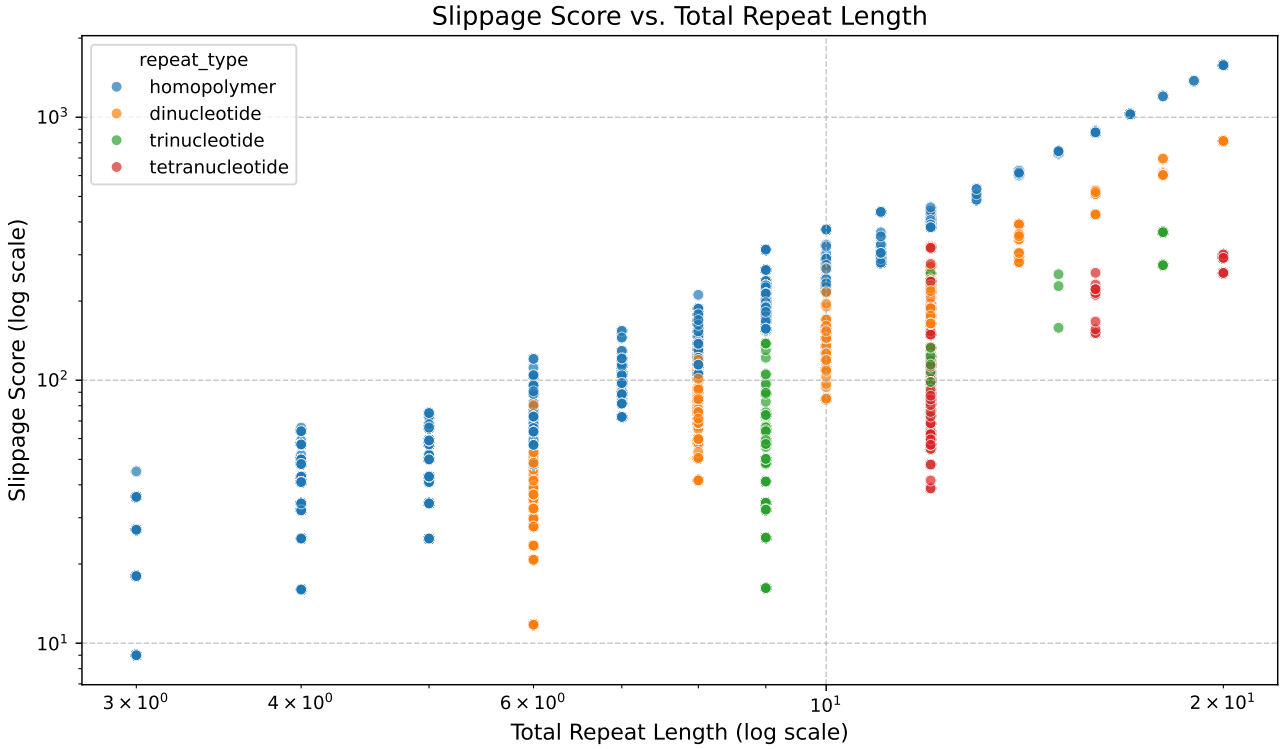

Figure A3: The slippage score assigned to different repeat types.

	Slippage Threshold												
PIP Threshold	**10**	**20**	**30**	**40**	**50**	**75**	**100**	**150**	**200**	**300**	**400**	**500**	**inf**
Background Variant Counts													
0.001	9	20	22	24	25	29	31	35	37	42	45	47	56
0.01	315	506	654	774	856	1081	1252	1563	1796	2254	2616	2932	4320
0.05	827	1306	1681	1951	2158	2681	3055	3705	4173	5083	5777	6375	8794
Causal Variant Counts													
0.001	9	20	22	24	25	29	31	35	37	42	45	47	56
0.01	315	506	654	774	856	1081	1252	1563	1796	2254	2616	2932	4320
0.05	827	1306	1681	1951	2158	2681	3055	3705	4173	5083	5777	6375	8794
Background Gene Counts													
0.001	9	19	21	23	24	28	30	34	36	41	44	46	55
0.01	298	472	601	701	768	948	1074	1299	1460	1779	2015	2211	3019
0.05	766	1189	1501	1712	1862	2245	2515	2966	3281	3882	4333	4719	6075
Causal Gene Counts													
0.001	9	19	21	23	24	28	30	34	36	41	44	46	55
0.01	298	472	601	701	768	948	1074	1299	1460	1779	2015	2211	3019
0.05	766	1189	1501	1712	1862	2245	2515	2966	3281	3882	4333	4719	6075

Table A6: Full breakdown of gene and variant counts per pip threshold and slippage threshold.

C.7.3. TFBS Disruption

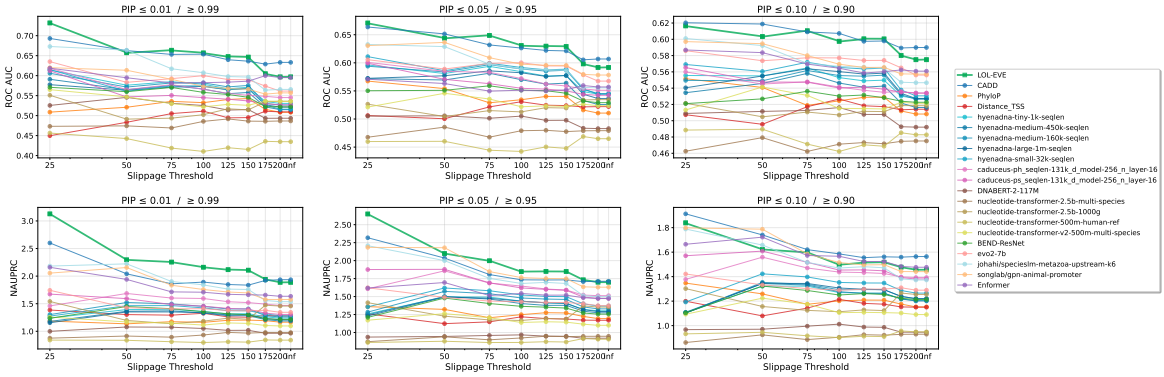

Figure A4: Cumulative causal-eQTL performance curves (running-mean AUROC and normalized AUPRC) as a function of slippage cutoff (log scale).

Table A7: SNP prioritization: AUPRC, normalized AUPRC, and ROC AUC for each model.

Model	**AUPRC**	**NAUPRC**	**ROC AUC**
CADD	0.110	1.920	0.600
Enformer	0.101	1.754	0.593
PhyloP	0.077	1.334	0.549
LOL-EVE	0.069	1.196	0.517
GPN-Promoter	0.062	1.078	0.515
NT-v2-50m-multi	0.062	1.072	0.518
NT-v2-250m-multi	0.061	1.057	0.521
DNABERT-2	0.059	1.035	0.512
NT-v2-500m-multi	0.059	1.031	0.514
NT-2.5b-multi	0.059	1.021	0.508
SpeciesLM	0.055	0.950	0.499
evo2-7b	0.054	0.948	0.487
HyenaDNA-medium-450k	0.054	0.944	0.500
Caduceus-ps	0.054	0.936	0.492
Caduceus-ph	0.053	0.923	0.488
HyenaDNA-large-1m	0.053	0.917	0.487
HyenaDNA-medium-160k	0.052	0.909	0.484
HyenaDNA-small-32k	0.052	0.900	0.480
HyenaDNA-tiny-1k	0.052	0.900	0.481

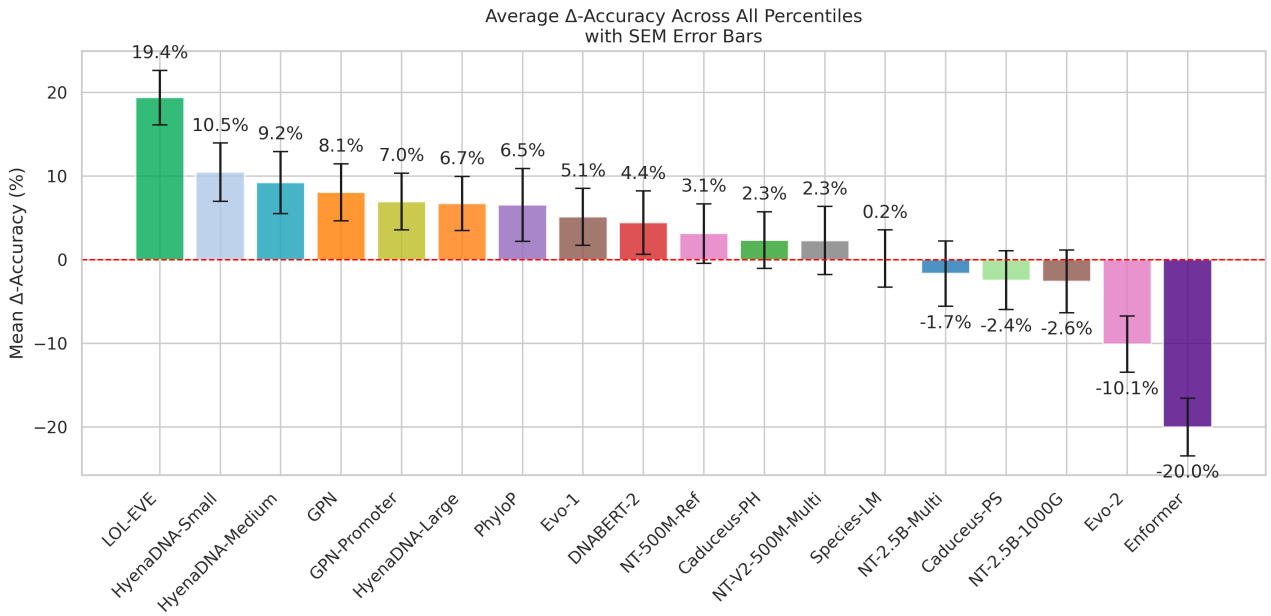

Figure A5: All models show for TFBS task.

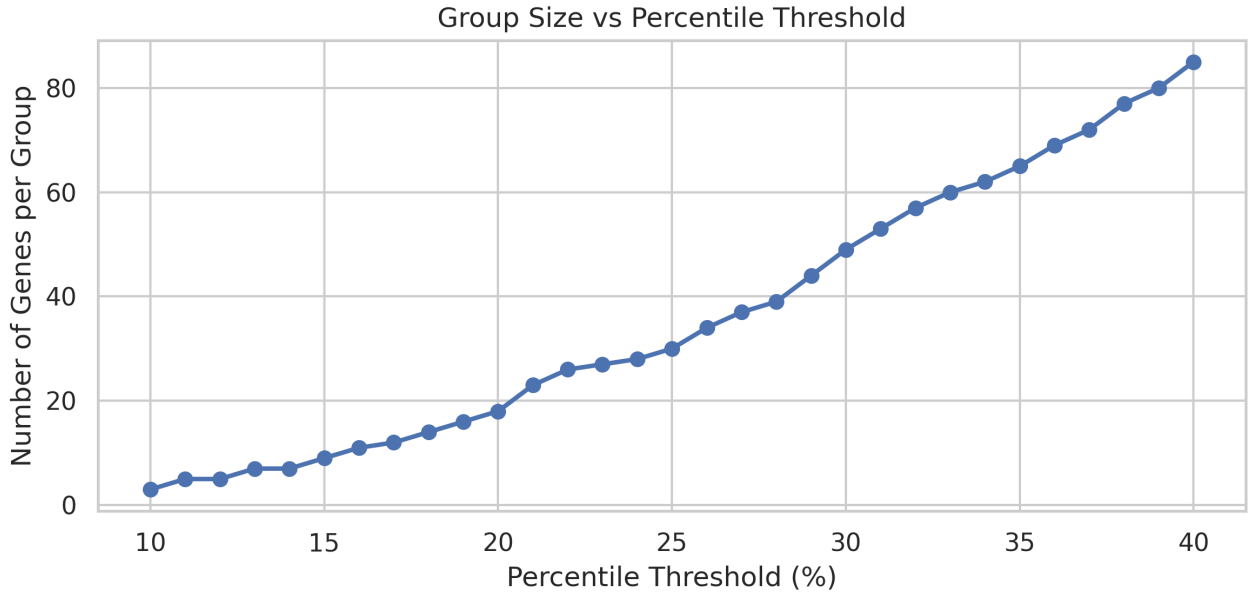

Figure A6: Cumulative gains of Genes per percentile threshold.