Promoter Sequence Generation using Homology Prompting

Erik Xie [*1,2] **Courtney Shearer** [*2] **Ruben Weitzman** [*2] **Pascal Notin** [2] **Debora Marks** [2]

Abstract

Promoters are critical regulatory elements that control gene expression and harbor disease-associated variants. We present PROSE (PROmoter SEt transformer), a generative model that learns from evolutionary relationships across mammalian species without requiring sequence alignments. PROSE adapts set transformer architecture to process families of homologous promoters, capturing patterns of conservation and variation that define functional regulatory elements. Trained on 13.6 million promoter sequences from 447 mammalian species, PROSE generates human promoters that accurately reproduce characteristic motifs while maintaining appropriate nucleotide distributions and achieving strong Sei regulatory activity scores. Unlike single-sequence baselines that overfit to repetitive patterns, PROSE produces diverse, biologically plausible sequences by leveraging evolutionary context. Our homology-based prompting approach outperforms single sequence models and demonstrates the value of incorporating cross-species information for genomic sequence design.

1. Introduction

Homologous sequences across species encode the accumulated knowledge from billions of years of natural selection. This information has been utilized extensively in protein language modeling (Rao et al., 2021; Notin et al., 2022; Orenbuch et al., 2023; Su et al., 2024), but its effective usage in DNA language modeling remains in its infancy. While current DNA language models train across evolution and can implicitly learn selection patterns, none explicitly condition generation on sets of homologous sequences.

We propose evolutionary prompting for promoter sequences, using sets of evolutionarily related promoter elements across species to learn the underlying distribution. Promoters are particularly valuable targets for such modeling as they regulate gene expression and harbor many disease-associated variants in humans (Maurano et al., 2012; Albert & Kruglyak, 2015). Learning these promoter sequence constraints through generation represents a first step toward understanding the impact of non-coding variation in human disease.

Promoters present unique challenges for computational modeling due to their evolutionary dynamics—they evolve more rapidly than coding sequences while maintaining essential functional constraints across species, often in ways difficult to capture through traditional sequence alignments (Benegas et al., 2023). Alignment-based methods may struggle with promoters where functional elements exhibit positional flexibility or undergo compensatory mutations while maintaining regulatory function, as transcription factor binding site orientation and order are major drivers of gene regulatory activity (Georgakopoulos-Soares et al., 2023).

Previous autoregressive approaches like LOL-EVE (Shearer et al., 2024) have made progress in modeling promoters but are limited to single-sequence representations that miss evolutionary context. Recent DNA sequence design advances include diffusion models like DDSM (Avdeyev et al., 2023) that enable tunable regulatory activity, yet these methods still process sequences independently. Meanwhile, homology-based approaches have shown remarkable success in protein engineering, with set transformer architectures like PoET demonstrating state-of-the-art performance for predicting complex mutation effects using only sequences (Truong Jr & Bepler, 2023), and CloneBO efficiently guiding protein optimization by leveraging evolutionary sequence families (Amin et al., 2024). Building on advances in set transformers for protein families (Truong Jr & Bepler, 2023), we introduce PROSE (PROmoter SEt transformer), a generative model that learns directly from sets of homologous promoters without requiring sequence alignment. By processing collections of evolutionarily related promoters across species, PROSE captures patterns of conservation and variation that define functional regulatory elements. This homology-based approach enables the design of human promoter sequences that respect evolutionary constraints, with potential applications in understanding

[*]Equal contribution [1]Department of EECS, Massachusetts Institute of Technology, MA, USA [2]Systems Biology, Harvard Medical School, MA, USA. Correspondence to: Erik Xie <ejxie@mit.edu>, Debora Marks < debbie@hms.harvard.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

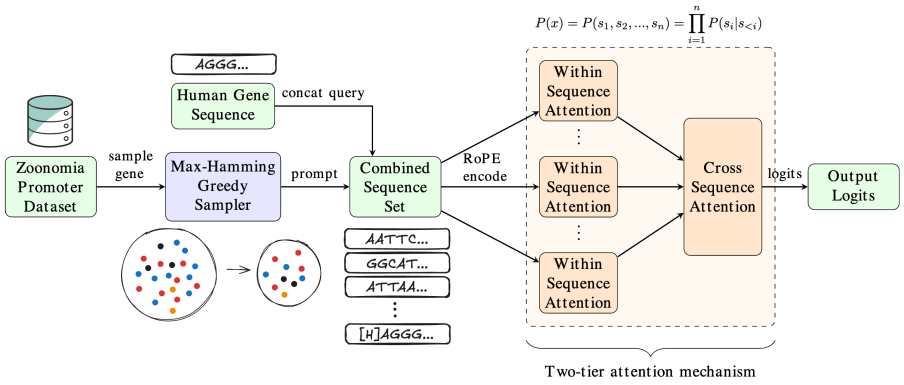


Figure 1. Architecture of PROSE. The highlighted encoder (and RoPE) is adapted from PoET.

non-coding variants in disease and designing synthetic regulatory elements.

2. Approach

2.1. Training Data

We collect a training dataset across mammalian species from the Zoonomia project (Christmas et al., 2023), which provides full genome assemblies of a large number of species. We adapt a comparative genomic approach to identify promoter sequences from the 447 mammalian species included in the dataset, which share the most homology with humans (see subsection A.1).

2.2. Methods

PROSE adapts the architecture of PoET (Truong Jr & Bepler, 2023), designed for processing sets of sequences. This adoption is crucial for handling homologous promoter sets, as their collective evolutionary identity should be independent of any specific presentation order. To achieve this permutation invariance at the set level, PROSE employs Rotary Positional Encoding (RoPE) applied *within* each sequence, ensuring positional information is relative and local. A two-tiered attention mechanism captures evolutionary context: first, self-attention *within* each sequence models internal dependencies, and second, attention *across* sequence representations identifies shared motifs and higher-order relationships among homologs.

Our Zoonomia dataset comprises per-gene families of homologous promoter sequences. Each family typically includes a human promoter sequence and sequences from various other mammalian species. Due to the large size of these families, we employ sampling strategies to construct smaller, representative input sets for efficient training.

For a given gene, let S_{mam} denote the set of its non-human mammalian promoter sequences, and let t_{hum} be its corresponding human promoter sequence. Each training input consists of a **query sequence** (q) and a **homology set** ($H_S = \{h_1, h_2, \ldots, h_{k-1}\}$), which contains $k-1$ homologous sequences.

We explore various strategies for constructing the model input by varying the selection methods for the homology set H_S and the query q.

Homology Set Selection: The $k-1$ sequences for H_S are chosen from S_{mam} using one of two methods:

Random Sampler (-Rand): H_S is formed by randomly selecting $k-1$ sequences from S_{mam} without replacement.

Max-Hamming Greedy Sampler (-Greedy): This sampler aims to maximize diversity within H_S. It initializes with a single sequence h_0 randomly sampled from S_{mam}. Subsequently, sequences are iteratively added to H_S from $S_{mam} \setminus H_S$. In each step, the sequence $h_i \in S_{mam} \setminus H_S$ that maximizes the average Hamming distance to sequences already in H_S is chosen:

$$h_i = \text{argmax}_{x \in S_{mam} \setminus H_S} \left(\frac{1}{|H_S|} \sum_{h' \in H_S} \text{Hamming}(x, h') \right)$$

until $|H_S| = k-1$. This greedy sampling strategy maintains diversity of evolutionary context within a computationally tractable subset size (Rao et al., 2021).

Query Selection: Given a selected homology set H_S, the query sequence q is chosen according to one of the following strategies:

Human-Prioritized Query (-Human-): With probability $P_{human} = 0.3$, the human sequence t_{hum} is chosen as the query ($q = t_{hum}$). Otherwise (with probability 0.7), q is randomly sampled from $S_{mam} \setminus H_S$. No special token is used.

Conditional Human Prompt (-Prompt-): With probability P_{prompt}, the human sequence t_{hum} is chosen as the query ($q = t_{hum}$), and a conditioning token `[H]` is associated with this input. Otherwise, q is randomly sampled from $S_{mam} \setminus$

Table 1. Performance Metrics and Model Augmentations. Left to Right: Model Category; Mean Sei; Delta Frequencies of Nucleotide, TATA Boxes, GC Islands, and CCAAT Boxes; Fréchet Inception Distance

Type	Model	Sei ↑	ATCG ↓ (10^{-2})	TATA ↓ (10^{-4})	GC ↓ (10^{-4})	CAT ↓ (10^{-4})	FID ↓
PROSE	30%-Human-Greedy	**9.76**	1.67	1.81	3.18	1.70	**0.0039**
	30%-Prompt-Greedy	8.87	1.72	1.67	4.45	1.89	0.1062
	30%-Human-Rand	6.21	1.04	2.29	6.03	2.52	0.0081
	100%-Human-Greedy	9.43	**0.93**	**1.39**	**2.86**	**1.43**	0.0593
Baseline	Single Sequence	**21.39**	9.62	3.20	23.5	5.17	0.0621

H_S without the `[H]` token. The inclusion of the `[H]` token aims to guide model learning of specific characteristics of human promoters in relation to their homologs.

The generation process is modeled autoregressively. The probability of observing the k biological sequences $X_{seqs} = (x_1, \ldots, x_k)$ is given by:

$$P(x|H, X) = P(H) \cdot \prod_{i=1}^{|X|} P(x_i|x_{<i}) \quad (1)$$

$$= P(H) \cdot \prod_{i=1}^{|X|} \prod_{j=1}^{\text{len}(x_i)} P(x_{i,j}|x_{<i}, x_{i,<j}) \quad (2)$$

where H refers to the optional presence of the `[H]` token.

For each query selection strategy (-Human-, -Prompt-), we test performance when paired with homology set selection methods (-Rand and -Greedy). For all approaches involving homology sets, each gene is sampled with probability $P_{gene} \propto |S_{mam}|$ during training to ensure larger families are adequately represented.

We also evaluate a baseline model that does not explicitly leverage homology information. In this setup, the input consists of a single sequence, randomly sampled from $S_{mam} \cup \{t_{hum}\}$, and generated autoregressively without any homology context.

3. Results

We experiment with the controllability of query sequences during training of PROSE, focusing on the ability to generate human promoter sequences due to their relevance in the genetic basis of human diseases. We evaluate results in three aspects:

1. The frequencies of common motifs that characterize human promoter sequences: TATA, GC box, and CCAAT boxes (Avdeyev et al., 2023). We report the mean delta between motif frequencies of designed and ground truth promoters.
2. The frequency of nucleotides, where most promoters are rich in GC base pairs (Avdeyev et al., 2023). We report the mean delta in the same fashion.
3. The mean score outputted by Sei (Chen et al., 2022), a foundational model trained on human sequences that scores promoter activities. Higher scores typically correspond to higher promoter activity.

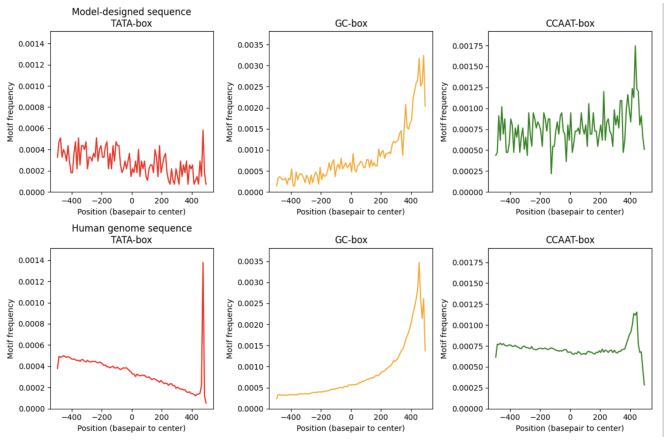
3.1. Homology Prompting Outperforms Single Sequence Models

In Table 1, metrics are reported for PROSE trained with 30% human queries, 100% human queries, 30% human queries with special human token (`[H]`), random sampling instead of greedy, and a baseline single sequence input model. Although the baseline model has a high Sei score, the model scores poorly on all other metrics. Upon inspecting generated sequences, the single sequence baseline model lacks diversity and generates mostly GC repetitions (Figure 2), which is likely picked up by Sei as an overwhelmingly strong signal (as in Table 1). Since the score far surpasses the natural human promoter set that the model is trained on, this indicates the model failed to generalize to functional promoters.

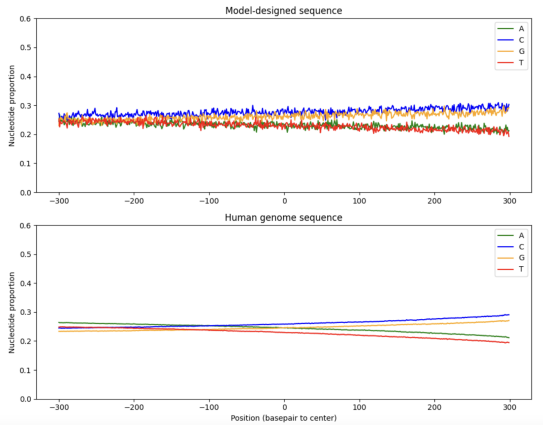
Apart from Sei score, the PROSE-based models achieve better performance than the baseline model, with Greedy 30% and 100% models achieving the best overall performance. Prompting with an additional token shows no benefit for generating more human-like sequences. We therefore choose PROSE with 30% Human-Greedy for the rest of the generative results.

3.2. Generated Sequences Capture Known Promoter Motifs

Figure 2 illustrates the occurrences of motifs and nucleotide frequencies on a human promoter and PROSE designed promoter on a validation chromosome for PROSE with 30% Huma-Greedy. PROSE-generated sequences faithfully reproduce characteristic motifs such as TATA boxes, GC islands, and CCAAT boxes while maintaining appropriate nucleotide distributions. The model accurately represents positional preferences and frequency patterns of these critical regulatory elements, with patterns closely matching



(a) TATA, GC, CCAAT motif frequencies



(b) ATCG nucleotide frequencies

Figure 2. Biological properties of PROSE (Human 30%) designed sequences on validation chromosome 19 (top) and the human promoter set (bottom). All frequencies were computed using a moving window of 600 base-pairs.

those observed in natural human promoters. The nucleotide composition analysis further confirms that PROSE captures the characteristic GC-rich nature of promoter sequences. These features are essential for proper promoter function, providing binding sites for transcription factors and establishing structural properties required for transcription initiation.

3.3. PROSE generates diverse promoter sequences that remain functional

In Figure 3, we compute the Hamming distance from each PROSE designed sequence to the set of human sequences, and visualize its correlation with the Sei score of promoter activity. The model produces promoters with diversity from human sequences, as shown by the Hamming distances, while retaining functional aspects via Sei scores. Sequences more similar to the human set have higher activity scores, as expected. In addition to the 30% Human-Greedy, we looked at the 100% Human-Greedy in 3. 100% Human-Greedy model generates sequences that are further away from the training distribution compared to the the 30% model, FID 0.0593 vs 0.0039, as well as the hamming distance to human sequences. We hypothesize this is because the 100% model, trained exclusively on human promoter queries, lacks sufficient query diversity and consequently overfits to the human promoter distribution, leading to mode collapse and reduced generalization capability.

4. Conclusion & Future Directions

With homology prompting, PROSE designs biologically consistent and diverse promoter sequences by effectively capturing evolutionary context through a mixture of human and non-human queries. Utilizing the PoET architecture, PROSE is able to leverage a mixture of human and non-

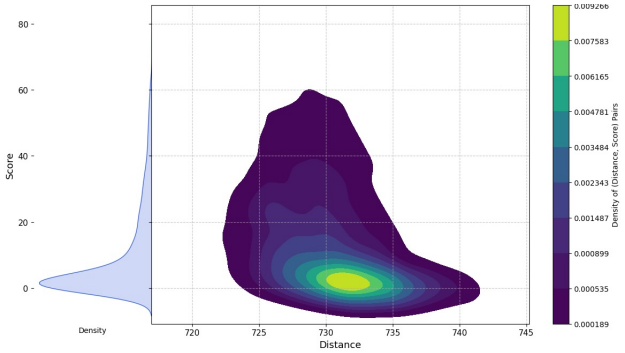


Figure 3. Correlation between SEI score and Hamming distance of PROSE designed promoters to human promoters. Left: PROSE (30% Human-Greedy) score distribution

human queries to learn additional evolutionary context, complementing performance with fitting on human queries only. The model's superior performance over single-sequence baselines demonstrates that it has internalized the selective pressures governing promoter homology, suggesting potential applications beyond sequence generation. Similar to LOL-EVE's approach for predicting indel effects, PROSE's learned representations could be adapted to score regulatory variants, prioritize mutations in genome-wide association studies, or guide therapeutic design targeting regulatory elements.

Future improvements could incorporate phylogenetic information into sampling strategies to create more evolutionarily representative subsets and develop more sophisticated human-specific prompting control. The convergence of generative modeling and variant effect prediction represents a promising direction for understanding genetic disease mechanisms, with PROSE demonstrating that evolutionary homology provides a powerful framework for both generating and evaluating regulatory sequences.

Impact Statement

This paper introduces PROSE, a generative model designed for creating biologically plausible promoter sequences by leveraging evolutionary relationships across mammalian species. By enabling the design of synthetic regulatory elements that capture natural evolutionary constraints, PROSE has the potential to contribute to advances in synthetic biology, gene therapy design, and our understanding of regulatory genomics. While the model's ability to generate human-like promoter sequences may aid in developing therapeutic interventions and studying disease mechanisms, generated sequences should be thoroughly validated experimentally before any clinical applications to ensure safety and efficacy. Ethical considerations include the responsible development of synthetic regulatory elements to prevent potential misuse in genetic engineering applications that could have unintended consequences. As with any AI-driven approach for biological sequence design, care must be taken to ensure that applications benefit diverse populations equitably and that generated sequences are not used for harmful purposes such as creating dangerous biological agents. The model's training on evolutionary data from multiple species raises considerations about biodiversity conservation and the responsible use of genomic resources. Nonetheless, this work primarily seeks to enhance computational methods for understanding promoter evolution and regulatory sequence design, with the goal of advancing basic biological research and potentially contributing to beneficial medical applications.

References

Albert, F. W. and Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

Amin, A. N., Gruver, N., Kuang, Y., Li, L., Elliott, H., McCarter, C., Raghu, A., Greenside, P., and Wilson, A. G. Bayesian optimization of antibodies informed by a generative model of evolving sequences, 2024. URL https://arxiv.org/abs/2412.07763.

Avdeyev, P., Shi, C., Tan, Y., Dudnyk, K., and Zhou, J. Dirichlet Diffusion Score Model for biological sequence generation. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2305.10699.

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.

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.

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.

Faltings, F., Stark, H., Jaakkola, T., and Barzilay, R. Protein fid: Improved evaluation of protein structure generative models, 2025. URL https://arxiv.org/abs/2505.08041.

Georgakopoulos-Soares, I., Deng, C., Agarwal, V., Chan, C. S., Zhao, J., Inoue, F., and Ahituv, N. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature communications*, 14 (1):2333, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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, March 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt128. URL http://dx.doi.org/10.1093/bioinformatics/btt128.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

Notin, P., Dias, M., Frazer, J., Hurtado, J. M., and others. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *International*, 2022.

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. *bioRxiv*, November 2023.

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. Msa transformer, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Shearer, C. A., Notin, P., Teufel, F., Orenbuch, R., Ritter, D., Spinner, A., Xie, E., Frazer, J., Dias, M., and Marks, D. S. LOL-EVE: Predicting promoter variant effects from evolutionary sequences, 2024. URL https://openreview.net/forum?id=LQRglYZ2Ri.

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.

Truong Jr, T. and Bepler, T. Poet: A generative model of protein families as sequences-of-sequences. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 77379–77415. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=1CJ8D7P8RZ.

A. Appendix

A.1. Dataset

Typically, promoter regions are identified using Transcription Start Site (TSS) annotations. However, these are not readily available for many species, which have not been extensively annotated by researchers. To address this, we identify putative promoters using a genomic lift-over of the exon starting positions in a standard human genome, which is annotated, to the 477 mammalian species in the dataset in Zoonomia. The lift-over algorithm is part of the HAL toolkit (Hickey et al., 2013) and aligned the exon positions in humans to other species by sequence similarity. 1000 base pairs upstream the exon starting position were taken as the promoter sequence for each gene. We filter our dataset by retaining only promoters that have corresponding exons that are at least 50% in length of their corresponding human exons to exclude unreliable alignment. In total, we collect a dataset of 13.6 million promoter sequences. The Zoonomia training dataset has a mean Sei score of 5.487, and the human subset has a mean score of 6.542.

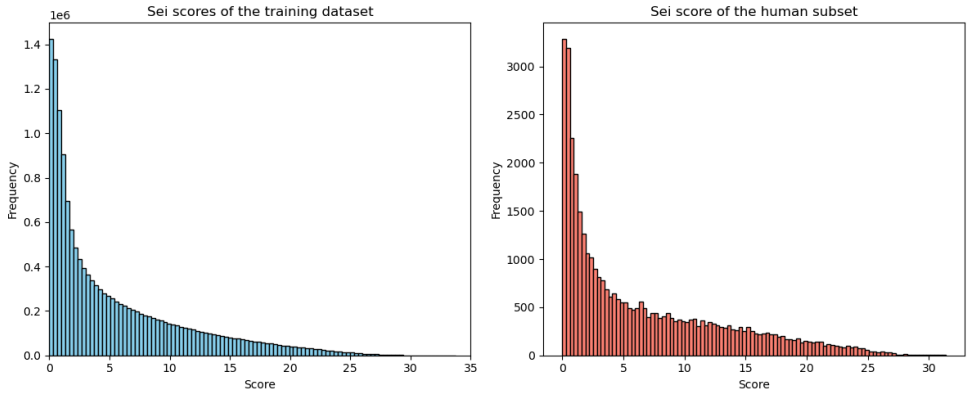


Figure 1. Sei score distributions of 447 species promoters (left) and human promoters (right)

A.2. Training Details

PROSE is trained with 201M parameters on 2 NVIDIA L40s GPUs for 2 days. We withhold promoters on chromosome 19 for validation and testing. At training time, each set of promoters is reversed with probability $P = 0.5$ to improve model robustness (Truong Jr & Bepler, 2023). We use a max set length of 16384 (including query), and Adafactor optimizer with default parameters and a learning rate of 1e-3.

A.3. Additional Metrics

We report the FID (Fréchet Inception Distance) metric of validation promoters generated by PROSE. FID has been used to measure diversity of images created by generative models with images in reference datasets (Heusel et al., 2017). Recent works have adopted it to protein generation (Faltings et al., 2025). We compute it based on mean and covariance of k-mer frequency and sequence properties of PROSE generated and ground truth human sequences. Low FID represents less deviation from the ground truth distribution. We use 3 for k-mer size, and consider these properties: sequence length, GC content, and dinucleotide frequencies.

Table A1. FID scores and Model Augmentations

Model	FID ↓
30%	**0.0039**
30%[H]	0.1062
30%[R]	0.0081
100%	0.0593
Baseline	0.0621

PROSE with 30% human query achieves the lowest deviation from the human promoter distribution, achieving a balance

between diversity and satisfaction of biological constraints.

A.4. Baseline Models

We examine the baseline model's behavior with the same metrics and provide comparison against human promoters. The plot confirms our reasoning about the high Sei score of the single sequence model, which potentially highlights a pitfall of the Sei foundation model in validation of promoter designs.

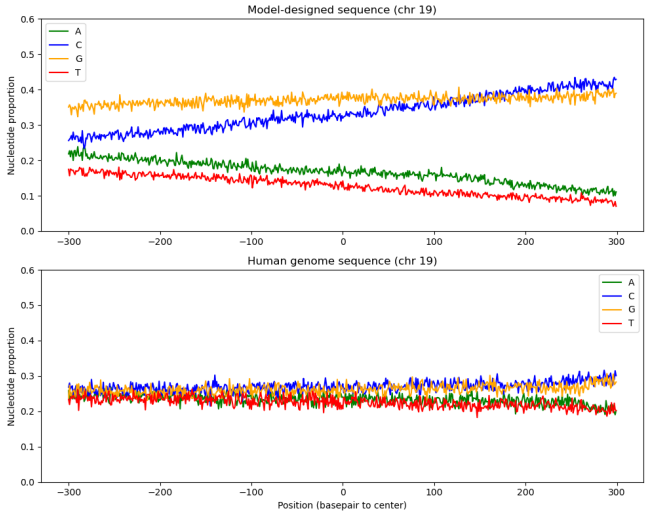


Figure 2. GC-overfit patterns seen without homology prompting. Top: baseline designed promoters (chromosome 19), Bottom: ground truth validation (chromosome 19) human promoters

We also examine the correlation between PROSE generated sequences and their Sei scores when queries are 100% human sequence. We observe that sequences further from the human set have lower Sei scores compared to PROSE with 30%, which was included in the main section.

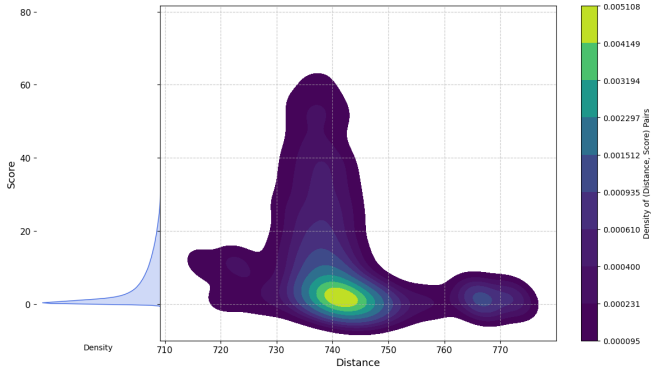


Figure 3. Sei score vs. Hamming Distance for PROSE (100% Q-HumanP-Greedy). Left: training set Sei Score distribution.