Direct Prediction of Gene Expression with Promoter-0

Authors: Siqi Zhao, Jake Wintermute, Valentin Zulkower, and Alyssa Morrow

Contributors: Justin Gardin, Michal Jastrzebski, Dana Merrick, Joshua Moller, Lood van Niekerk, Porfirio Quintero Cadena, Seth Ritter, Hao Shen, and Ankit Gupta

INTRODUCTION

Programmable gene expression is essential for the design of many engineered biological systems. Applications in gene therapy and biologics manufacturing, among many others, depend on our ability to express a target gene in the right cell type at the right level. Today, much of the biotech industry depends on a small number of widely used legacy promoters, many of which have never been systematically optimized for purpose, that offer little sequence variety and often fall short of desired tissue specificity.

The emergence of large DNA foundational models presents an opportunity to revolutionize promoter design. Here, we describe Promoter-0, an AI framework capable of modeling tunable and tissue-specific promoters. Our approach builds on Borzoi, a sequence-based machine learning model that learns to predict RNA-seq coverage from DNA sequence [1]. Using Ginkgo's high-throughput screening platform, we collected tens of thousands of data points to validate and expand this framework.

Promoter-0 can predict promoter activity across diverse cell and tissue types without requiring additional model fine-tuning. Remarkably, zero-shot predictions from Promoter-0 perform comparably to standard models trained with labeled data in some settings. To the best of our knowledge, this represents the first demonstration of direct prediction of context-specific expression of a synthetic expression cassette, an important practical milestone in promoter design.

A simple, direct-prediction tool for gene expression has the potential to streamline many DNA design tasks. We envision two broad applications of Promoter-0.

1) Rational promoter design. Using Promoter-0 can allow engineers to select promoters that are more likely to achieve a desired expression level in a desired cell type. We show our model's ability to do so with commonly used promoters.

2) Iterative promoter optimization. For high-performance applications requiring multiple rounds of design-build-test, Promoter-0 can generate a balanced and diverse library of candidate promoters and enable more informative iterations. We demonstrate the ability of Promoter-0 to predict the activity of diverse sets of candidate promoters.

Access Promoter-0 is available through the Ginkgo Model API. You can read additional documentation or follow this Google Colab notebook for a demonstration of usage.

A PROMPT ENGINEERING STRATEGY TO CAPTURE PROMOTER ACTIVITY IN CONTEXT

The Borzoi model released by Calico Labs in 2023 predicts genomic readouts (gene expression level, chromatin accessibility, transcription factor binding strength, etc) in both human donor tissues and cell lines. We hypothesized that Borzoi's learned representations of DNA sequences in different tissue contexts would be similarly useful for predicting the activity of engineered promoter sequences.

But before we could use Borzoi as a model for promoter activity, we had to adapt the promoter DNA sequences to match Borzoi's input requirements. Borzoi is trained with very long sequence inputs, 524 kilobases, to

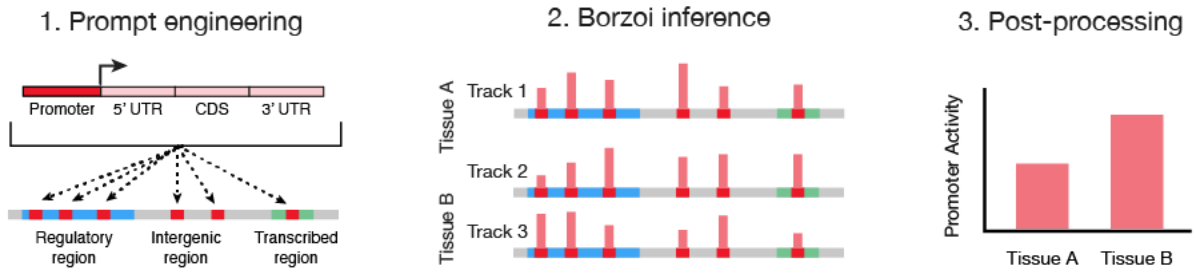

FIGURE 1. Overview of Promoter-0, a Borzoi-based promoter activity prediction framework. During prompt engineering, a short-expression cassette, including the promoter of interest, is embedded in 100-1000 random sites in a much larger DNA sequence. During Borzoi inference, the Borzoi model is applied to the longer sequence to predict various biochemical features, including promoter activity. In post-processing, the predicted promoter activities for each copy of the embedded cassette are averaged to produce a context-independent predicted activity.

capture extensive genomic context. We were interested in much shorter constructs: a typical promoter + payload combination is only about 1000-3000 bases long.

To bridge this gap, we took inspiration from TRIP (Thousands of Reporters Integrated in Parallel) experiments [2]. In TRIP experiments, identical promoters are randomly integrated throughout a host genome. By measuring the activity of the same construct across many different random insertions, TRIP experiments effectively average out the effect of any particular insertion site.

Promoter-0 uses a prompt engineering framework to similarly embed short expression cassettes into long stretches of genomic DNA (Fig. 1). It randomly selects 100 - 1000 insertion sites, runs Borzoi inference, and calculates an average activity prediction for the cell types of interest.

TESTING PROMOTER-0 AGAINST CLINICALLY RELEVANT PROMOTER MEASUREMENTS

We evaluated Promoter-0's predictive capabilities for 20 clinically relevant promoters in four selected cell lines: HepG2, Hek293T, A549, and iPSC-derived cardiomyocytes (heart). The test data was generated using MPRA (Massively Parallel Reporter Assays) onboarded on Ginkgo's high-throughput screening platform. The model robustly predicted the activity of the 20 promoters across different cell lines (Average Spearman ρ: 0.77, Average Pearson r: 0.72, Fig. 2).

Given that Borzoi was trained using genomic data, we considered that it might be less effective for synthetic or viral promoters. We labeled each promoter with its origin, and to our surprise, we observed no strong bias against viral or synthetic promoters.

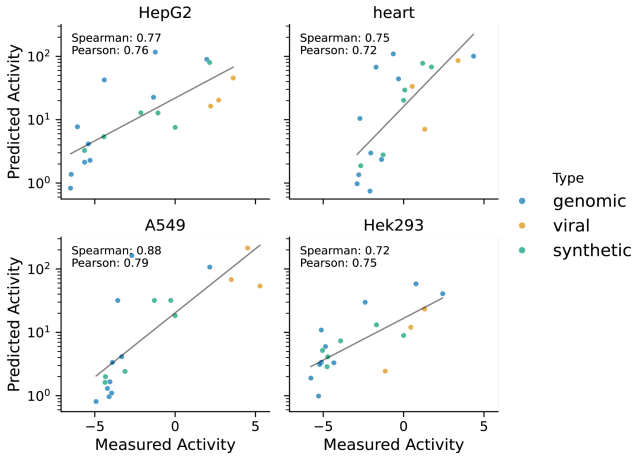

FIGURE 2. Promoter-0 enables the prediction of commonly used promoter activities in multiple cell types. Scatterplot showing the predicted promoter activity (processed from Borzoi DNase-seq tracks) vs. measured MPRA activity for 20 clinically relevant promoters in HepG2, iPSC-derived cardiomyocytes (heart), A549, and Hek293 cells.

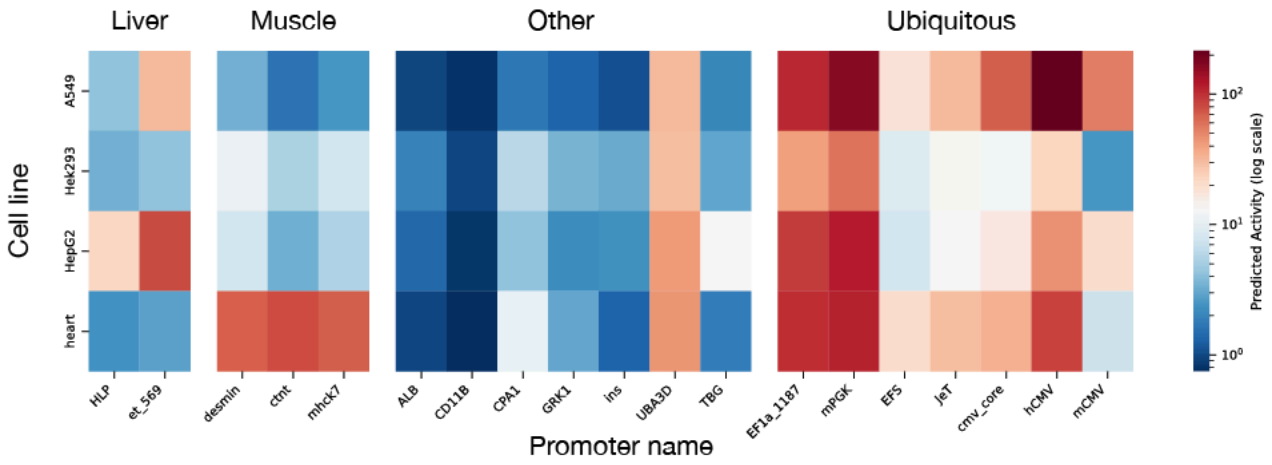

FIGURE 3. The Borzoi-based framework accurately predicts known cell type-specific promoter activity. Heatmap showing the predicted promoter activities (processed from Borzoi DNase-seq tracks) for commonly used promoters grouped by their intended target cell lines.

The test set included promoters with known cell-type specific activity profiles: three muscle/cardio-specific promoters, two liver-specific promoters, six ubiquitous promoters, and promoters targeting other cell contexts. Promoter-0 successfully recapitulated these established tissue specificities (Fig. 3). Muscle-specific and liver-specific promoters had higher predicted expression in the correct contexts. In contrast, promoters targeting other tissues are largely inactive in these cell lines.

BENCHMARKING PROMOTER-0 WITH MASSIVELY PARALLEL REPORTER ASSAY (MPRA) DATA

Next, we assessed the framework's capacity to predict promoter activity in large MPRA datasets. MPRA experiments provide 10^3 – 10^5 parallel measurements of gene expression in specific cellular contexts. These datasets traditionally serve two purposes: building mechanistic models of gene regulation and powering ML-guided optimization of regulatory sequences [3,4]. However, a significant limitation persists with MPRAs: large-scale experimental data generation is limited to immortalized cell lines or highly restricted in vivo contexts. Our framework addresses this limitation by making use of genomic training data from a variety of cell lines, primary cells, and human tissues.

To compare Promoter-0's performance with that of models trained on MPRA datasets, we curated ~1.4 M sequence-regulatory activity pairs in HepG2 cells from public and internal sources. To minimize potential data leakage, we clustered the sequences using mmseq2 [5].

Following clustering, we partitioned the dataset into training, validation, and test subsets using an 80:10:10 split. We used the hyenaDNA [6] model architecture with a multilayer perceptron (MLP) output layer to establish the trained model benchmarks and initialize the training with random and pre-trained weights.

We trained two versions of hyenaDNA: one with random weights and another pre-trained on the human genome. Both versions were then finetuned using MPRA data. We used the trained models to predict the held-out test set (10% of the total dataset, 140K sequences).

Comparing the MPRA-trained models with Promoter-0, we found Promoter-0 to underperform only slightly (Table 1).

Model	Spearman ρ	Pearson *r*
HyenaDNA (random initialization)	**0.47**	**0.50**
HyenaDNA (pre-trained)	0.44	0.48
Promoter-0	0.31	0.41

Table 1. Promoter-0 predicts MPRA data comparably to HyenaDNA models trained with MPRA datasets. Correlations are between model-predicted activities and promoter measurements MPRA datasets. HyenaDNA models were initialized with random weights or pre-trained on human genomic DNA, then finetuned using 80% of the MPRA dataset. Promoter-0 and HyenaDNA inference were performed on the same test dataset.

We also benchmarked against recent work by Tang and Koo (2024) [7] using genomic language models to predict regulatory element activities. That study trained several foundational models and a CNN using a dataset containing ~120K regulatory elements measured in HepG2 and K562 cells [8].

Following the evaluation procedure established by Tang and Koo, we randomly selected ~12K sequences for testing with Promoter-0. Our model performed comparably to Tang and Koo's HyenaDNA MLP model but was outperformed by their CNN (Table 2). In this case, the superior performance of the CNN could be due to the random split of the training and test data sets. In contrast, Promoter-0 had no previous exposure to similar sequences' regulatory activity.

Cell Line	Method	Pearson *r*
K562	HyenaDNA-MLP	0.46
	CNN	**0.71**
	Promoter-0	0.5
HepG2	HyenaDNA-MLP	0.36
	CNN	**0.65**
	Promoter-0	0.37

Table 2. The Borzoi-based direct-prediction framework outperforms some supervised training settings on a Lenti-MPRA dataset. Correlations for HyenaDNA with an MLP layer and custom CNN were from Tang and Koo[5]. Correlations for Promoter-0 were calculated using a random subset of 10% of the lendiMPRA data from Agarwal *et al*. [6].

PREDICTING PROMOTER ACTIVITY IN DATASETS FROM REAL R&D PROJECTS

Next, we evaluated Promoter-0 on smaller MPRA datasets generated using specific design strategies. Data of this format most resembles promoter engineering campaigns in real-world R&D projects. Iterative promoter design campaigns often begin with a set of candidate natural genomic promoters, systematically measuring and varying their sequences to achieve a desired gene expression target.

We focused on high-quality MPRA datasets that investigated short core promoters (Hong & Cohen, 2022 [9]), mid-sized genomic promoters (Ginkgo-internal project data), and synthetically designed promoters using transcription factor binding motifs (Ginkgo-internal data). Promoter-0 performed consistently at this practical DNA design task (Table 3).

PREDICTING THE EFFECTS OF SEQUENCE VARIATION AT SINGLE-BASE RESOLUTION

Lastly, we used Promoter-0 to analyze saturation mutagenesis datasets in which each base pair of a regulatory sequence is systematically mutated to all three alternative nucleotides to create a comprehensive map of sequence-function relationships. These represent a particularly stringent test because they require the Borzoi-based model to determine the effect of a single base pair change across more than 500 kb of sequence context.

Dataset	Spearman *ρ*	Pearson *r*	# of datapoints	Length	Cell line
Hong *et al*. [7]	0.53	0.57	670	80	K562
Genomic (internal)	0.71	0.58	2000	230	HepG2
Synthetic (internal)	0.64	0.64	1000	200	HepG2

Table 3. Promoter-0 performance on smaller datasets resembling real promoter design R&D projects. Correlations were calculated for all three MPRA datasets between the measured promoter activity and the predicted promoter activity using DNase-seq tracks.

We focused on two clinically relevant promoters measured in the HepG2 cell line: the coagulation factor IX (F9) promoter, crucial for blood clotting disorders, and the low-density lipoprotein receptor (LDLR) promoter, essential for cholesterol metabolism [10].

To establish a performance baseline, we used the HyenaDNA model trained with 1.4 M MPRA data points to predict the variant effects. Promoter-0 showed similar performance and superior generalizability compared to the supervised approach (Table 4).

Dataset	Method	ρ	r
F9	HyenaDNA	0.27	**0.4**
	Promoter-0	**0.31**	0.35
LDLR	HyenaDNA	0.22	0.29
	Promoter-0	**0.48**	**0.53**

Table 4. Performance evaluation of Promoter-0 on saturation mutagenesis datasets. The correlation coefficients (Spearman's ρ and Pearson's r) were calculated using mutation scanning experiments of the F9 and LDLR promoters. The predicted activity measurements were the DNase-seq tracks from HepG2 cells.

CONCLUSIONS

Genomic foundation models like Borzoi can learn key features of gene regulation across large segments (>500 kb) of genomic DNA. The results from Promoter-0 show that the Borzoi model can be effectively generalized to predict the activity of much smaller engineered promoter sequences (1000-3000 bp) independently from their larger genomic context.

Promoter-0 offers a direct route for estimating promoter activity, even without laboratory measurements. In some cases, the direct prediction may allow biotech R&D teams to identify promoters with desired activity profiles and use them immediately in engineered biological systems. However, for many demanding biotech R&D applications, it is unlikely the direct predictions of Promoter-0 will be precise enough to enable purely *in silico* DNA designs.

We anticipate that a more common use case for generative AI in promoter engineering will be to provide a well-balanced library of candidates for wet lab characterization and subsequent AI-driven iterative design improvement.

Promoter-0 can offer more efficient and informative DNA designs, allowing R&D teams to iterate toward high-performance promoters more quickly. To learn more, read the documentation for Promoter-0 or sign up for the API at models.ginkgobioworks.ai.

ACKNOWLEDGEMENTS

Thanks to the Ginkgo Foundry team for generating the large datasets required for testing and validation.

REFERENCES

1. Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. bioRxiv 2023.08.30.555582; doi: https://doi.org/10.1101/2023.08.30.555582

2. Akhtar, W. et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. Cell 154, 914–927 (2013).

3. Gosai, S. J. et al. Machine-guided design of cell-type-targeting cis-regulatory elements. Nature 634, 1211–1220 (2024).

4. Yin, C. et al. Iterative deep learning-design of human enhancers exploits condensed sequence grammar to achieve cell type-specificity. bioRxiv (2024) doi:10.1101/2024.06.14.599076.

5. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1026–1028 (2017).

6. Nguyen, E. et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. arXiv [cs.LG](2023).

7. Tang, Z. & Koo, P. K. Evaluating the representational power of pre-trained DNA language models for regulatory genomics. bioRxiv (2024) doi:10.1101/2024.02.29.582810.

8. Agarwal, V. et al. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. bioRxiv (2023) doi:10.1101/2023.03.05.531189.

9. Hong, C. K. Y. & Cohen, B. A. Genomic environments scale the activities of diverse core promoters. Genome Res. 32, 85–96 (2022).

10. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun 10, 1–15 (2019).