UNDERSTANDING DNA DISCRETE DIFFUSION FOR ENGINEERING REGULATORY DNA SEQUENCES

Anirban Sarkar*
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

Yijie Kang
Stony Brook University

Nirali Somia
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

Peter K Koo*
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory

ABSTRACT

Engineering regulatory DNA sequences with precise activity levels remains a major challenge in medicine and biotechnology due to the vast combinatorial space of possible sequences and the complex regulatory grammars governing gene expression. DNA discrete diffusion (D3) has emerged as a promising approach for learning these distributions and generating biologically relevant sequences, yet several key aspects of its capabilities remain unexplored. Here we systematically investigate D3's performance in biologically relevant, understudied scenarios. First, we demonstrate that D3 maintains robust performance even with limited training data, highlighting its practical utility in real-world applications where data is scarce. Second, we extend D3's conditional generation capabilities for categorical data, employing classifier-free guidance to improve the quality and specificity of generated sequences. Third, we analyze sequence trajectories during the diffusion process, providing insights into how discrete diffusion navigates the sequence-function landscape. Together, these findings expand our understanding of D3's strengths and limitations, while introducing new methodological advances for engineering functional regulatory DNA sequences.

1 INTRODUCTION

Gene expression is tightly regulated by non-coding DNA sequences that encode complex rules for transcription factor (TF) binding. These cis-regulatory elements are essential for cellular function, development, and environmental adaptation, but mutations can rewire their activity, leading to disease. Precisely engineering these elements to control gene expression holds great promise for therapeutic and biotechnological applications. However, this remains challenging due to the vast sequence space and the complex, context-dependent nature of cis-regulatory grammars.

Recent approaches to regulatory DNA design rely on treating supervised models as in silico oracles, where candidate sequences are scored based on predicted activity (Taskiran et al., 2024; de Almeida et al., 2024; Gosai et al., 2023). However, supervised oracles are constrained by their training distribution and often struggle to generalize under covariate shifts, a well-documented limitation of genomic deep learning (Sasse et al., 2023; Huang et al., 2023; Tang et al., 2023). Furthermore, sequence optimization methods—such as in silico evolution (Vaishnav et al., 2022), simulated annealing (Gosai et al., 2023), and gradient-based approaches (Linder & Seelig, 2020)—are fundamentally local search strategies. These methods navigate a restricted subset of sequence space, making incremental modifications rather than discovering truly novel regulatory elements, often leading to suboptimal solutions.

Deep generative models offer a more powerful approach by learning the full distribution of functional sequences, rather than treating design as an optimization problem over a fixed oracle. Generative models have demonstrated success in protein design, where strong sequence-level evolutionary

*Send correspondence to: asarkar@cshl.edu & koo@cshl.edu

constraints provide clear inductive biases. However, applying these methods to DNA sequence design has proven far more challenging. Unlike proteins, which have structured grammars and conserved domains, regulatory DNA is governed by weak, context-dependent constraints and contains sparse functional elements embedded within vast non-informative regions. This sparsity makes genomic language models and other reconstruction-based approaches (Nguyen et al., 2023; Dalla-Torre et al., 2023; Schiff et al., 2024; Gu & Dao, 2023) poorly suited for regulatory sequence design, as they rely on objective functions that reconstruct entire sequences, leading to memorization rather than generalization.

Diffusion-based generative models offer a compelling alternative. Unlike genomic language models, diffusion models iteratively refine sequences through a structured denoising process, making them better suited for capturing sparse regulatory signals. These methods have recently been introduced for DNA sequence generation, demonstrating early promise in models such as Dirichlet Diffusion Score Model (DDSM) (Avdeyev et al., 2023), DiscDiff (Li et al., 2024), DNA-Diffusion (Ferreira DaSilva et al., 2024), DNA Flow Matching (DFM) (Stark et al., 2024), and DNA Discrete Diffusion (D3) (Sarkar et al., 2024).

Among these, D3's discrete approach is particularly suited for regulatory sequence design, where functional motifs are sparsely distributed within large non-informative regions. By directly modeling nucleotide transitions, D3 refines key motifs without being overwhelmed by surrounding noise. In contrast, models like DiscDiff and DFM, which reconstruct sequences from compressed latent spaces or continuous vector fields, risk introducing distortions during reconstruction. We hypothesize that D3 mitigates these limitations by selectively mutating nucleotides, mirroring how evolution fine-tunes functional elements while tolerating neutral drift elsewhere.

Despite the growing promise of diffusion-based DNA design, several key questions remain: How do these models perform in data-limited settings, which characterize most biological applications? Can they reliably generate functional sequences that capture cell-type-specific regulatory grammars? And how do the internal dynamics of the diffusion process shape the sequences they generate?

In this study, we focus on D3 to address these gaps. We investigate its capabilities in small-data regimes, extend its conditional generation to accommodate cell-type specificity, and analyze its diffusion process to gain insights into how sequence transformations unfold during generation. These efforts not only enhance D3's utility for designing functional regulatory DNA but also advance our broader understanding of diffusion-based approaches in genomics. Our main contributions include:

- Demonstrating D3's efficacy in the small data regime and introducing strategies to enhance learning efficiency.
- Extending D3 for conditional sequence generation, enabling cell-type-specific regulatory sequence design.
- Providing insights into the diffusion process through sequence trajectory analysis.

2 Efficient Generation in the Small Data Regime

Motivation. Biological applications often operate in data-limited settings, where the underlying biological phenomena or experimental constraints—such as cost, time, and resource availability—restrict dataset size and diversity. Understanding how D3 performs under these conditions is essential for its practical application.

Experiments. To assess D3's generative capabilities in low-data scenarios, we systematically downsampled the training set of a fly enhancer dataset, where D3 previously outperformed DFM (de Almeida et al., 2022). Specifically, we used a STARR-seq dataset that quantifies enhancer activity in Drosophila S2 cells (de Almeida et al., 2022), originally consisting of 402,278 training sequences, 40,570 validation sequences, and 41,186 test sequences. Each sequence was measured under two independent conditions, capturing enhancer activity when driving a promoter from either a housekeeping gene or a developmental gene. To simulate data-limited conditions, the training set was reduced to 10%, 25%, 50%, and 75% of its original size, while validation and test sets remained unchanged for consistency. Models were trained on each downsampled dataset, and performance was evaluated based on sequences generated to match activities from the test set.

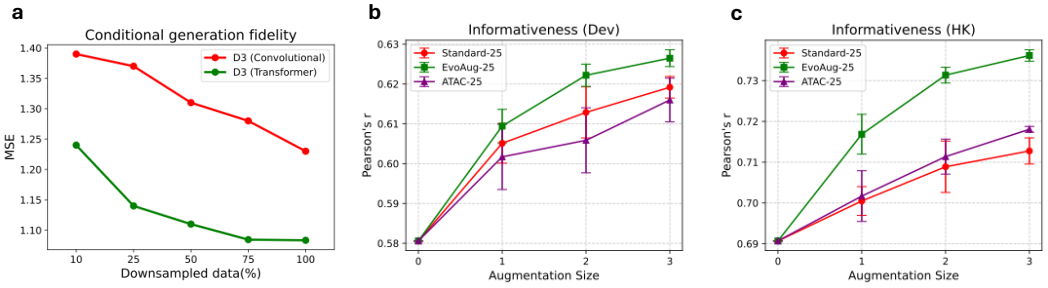

Figure 1: (**a**) Performance evaluation of D3 models in small data regime. We train with both convolution and transformer architecture, for each downsampled version, and evaluate on conditional generation fidelity averaged over developmental (Dev) and housekeeping (HK) enhancer task activities. (**b,c**) Informativeness analysis of D3 generated sequences with different training strategies. Downsampled STARR-seq dataset (25%) are used to train D3 models with (EvoAug and ATAC-seq) and without fine-tuning. Each of these trained models are then tasked to generate 3 sets of sequences conditioned on the enhancer activities of that downsampled 25% dataset. These generated sequences are augmented progressively to the original 25% dataset to create different sized dataset versions for each D3 model. Performance of all these trained prediction models through Pearson correlation coefficient for Dev and HK enhancer task activities are presented. Augmentation size refers to the sets of additional generated sequences. Each model training is repeated 5 times leading to error bars for every result in both these plots.

Model architecture. Our transformer model architecture and training procedure followed the specifications outlined in the original D3 study (Sarkar et al., 2024). We also considered the same convolutional architecture, employed by DFM (Stark et al., 2024) and DDSM (Avdeyev et al., 2023), and later followed by D3 (Sarkar et al., 2024) for consistent comparison. No modifications were made to the hyperparameter details as of D3.

Metrics. Performance was evaluated using conditional generation fidelity (Sarkar et al., 2024), measured by the mean squared error (MSE) between predicted activities for matched sets of natural and generated sequences using DeepSTARR (de Almeida et al., 2022) – a supervised model trained on the full STARR-seq dataset.

Results. Our experiments show that D3 maintains strong performance even with significantly reduced training data for both convolution and transformer architectures(Figure 1(a)). However, the transformer model displays greater resilience when trained on smaller datasets. With just 25% of the original training data, D3 performs comparably to models trained on the full dataset measured by our evaluation metric. However, performance declines sharply when trained on only 10% of the data. These findings suggest that D3 is highly data-efficient, generalizing well even under extreme data limitations—down to as little as 10% of the original dataset. Further benchmarking across more datasets and baseline diffusion and language models is needed to establish whether these findings provide a generalizable solution, that we plan to explore as future research.

3 D3-GENERATED SYNTHETIC SEQUENCES IMPROVE SUPERVISED MODELING

Motivation. A key question in evaluating generative models is whether the generated sequences capture meaningful biological mechanisms. If D3 effectively learns underlying cis-regulatory grammars, its synthetic sequences should provide complementary value when integrated with real data.

Experiment. Using the transformer-based D3 model trained on the 25% downsampled training set, we generated synthetic sequences conditioned on training set activities. These sequences were used to augment the training set by factors of 2x, 3x, and 4x. We then trained a DeepSTARR model from scratch on these augmented datasets and evaluated its performance using the Pearson correlation coefficient for both enhancer tasks on the original test set.

Methods. The transformer-based D3 trained on 25% downsampled training set was used as a baseline. To enhance performance in small-data settings, we explored transfer learning and data augmentation, both of which expose the model to broader cis-regulatory sequence patterns that may overlap with the fly enhancer dataset.

Transfer learning. We leveraged an ATAC-seq dataset from Drosophila embryos—previously used for pretraining supervised models for fly enhancer design(de Almeida et al., 2024)—as a pretraining dataset for D3. The dataset consists of 1,001-bp sequences with binary accessibility labels across 18 tissue pseudo-bulk ATAC-seq profiles. For training, we used the fold09 dataset, with the training set (471,536 sequences) drawn from all chromosomes except chrX, while validation and test sets were from chrX. During pretraining, D3 used the binary labels as conditional inputs. Fine-tuning was then performed on the downsampled STARR-seq dataset, with the learning rate reduced from 1e-4 to 1e-5.

Data augmentations. We applied EvoAug, a 2-stage evolution-inspired augmentation strategy that has demonstrated significant performance gains in supervised learning. During pretraining, EvoAug introduces random translocations, insertions, deletions, inversions, and mutations (using default parameters) to increase sequence diversity. While these transformations may alter sequence function, the assumption—aligned with the billboard model of gene regulation—is that key regulatory features are largely preserved. In the fine-tuning stage, the model was trained on the original dataset, refining its understanding of cis-regulatory grammar and motif interactions. This two-step process—first learning generalizable motifs through augmentation, then fine-tuning on unperturbed data—has been shown to improve performance in supervised tasks. Here, we assess whether EvoAug similarly enhances D3 generated synthetic sequences to improve supervised modeling in small-data regimes. A learning rate of 3e-4 was used during pretraining, followed by a reduction to 1e-5 during fine-tuning.

Results. Supervised models like DeepSTARR struggle in small-data regimes, as they rely on large datasets to learn complex cis-regulatory mechanisms. However, augmenting the training set with D3-generated sequences consistently improved DeepSTARR's predictive performance on both enhancer tasks (Fig. 1(b,c)). These results suggest that D3 has learned meaningful cis-regulatory features, enabling it to generate sequences that provide complementary information to real data, ultimately benefiting supervised models. Performance gains were observed across all augmentation levels (2x, 3x, and 4x the original dataset size), demonstrating that D3-generated sequences enhance regulatory signal learning in data-limited settings. Among the different D3 training strategies, EvoAug-trained D3 models produced the most substantial and consistent improvements, suggesting that EvoAug enhances the diversity and informativeness of generated sequences. In contrast, sequences generated by D3 models trained solely on 25% of the real data yielded more modest improvements, while ATAC-seq pretraining provided only limited additional benefit.

These findings indicate that D3 can generate functionally relevant cis-regulatory sequences that help supervised models overcome data limitations, improving their ability to capture meaningful sequence-function relationships. They also highlight the potential of data augmentation and transfer learning in enhancing the sequence quality generated by D3 in the small data regime. A major advantage of EvoAug is that it eliminates the need for a carefully curated pretraining dataset that closely matches the target task, which is often difficult or impossible to obtain. This makes EvoAug a promising approach for further exploration, as it could provide performance improvements even in cases where suitable pretraining data is unavailable. These methods underscore D3's potential as a synthetic data generation tool for advancing predictive modeling in regulatory genomics.

4 Evaluation of Conditional Generation Capabilities

Motivation. Generating cell-type-specific enhancer sequences is essential for understanding regulatory mechanisms and designing functional DNA sequences. While D3 has demonstrated conditional generation capabilities, its evaluation has been limited to a small set of enhancer tasks or cell types. It remains unclear whether D3 can generalize across a diverse range of cell types and accurately model enhancer sequence distributions.

Experiment. We evaluate D3's ability to generate enhancer sequences conditioned on cell types using a fly brain enhancer dataset (Janssens et al., 2022). This dataset contains 104,000 enhancer sequences (500-bp), each labeled with one of 81 cell types, determined from ATAC-seq data (Buenrostro et al., 2013). The base architecture for D3 on this dataset was the same 20 layer 1D convolution architecture to train D3 as DFM (Stark et al., 2024). We trained D3 for 500,000 steps with batch size of 128 on a single NVIDIA H100 GPU, using FBD (see below) on the validation set for early stopping.

Metric. We use Fréchet Biological Distance (FBD) (Stark et al., 2024) to evaluate how well generated sequences match real enhancer distributions. FBD measures the similarity between the feature space representations of real and generated sequences, computed using a pretrained classifier. Lower FBD values indicate better alignment between synthetic and real enhancer sequences. Following (Stark et al., 2024), we use the same 5-layer 1D convolutional classifier to compute FBD. We calculate FBD between the data and the generative model's distribution, where the 10.4k test sequences from the fly brain enhancer dataset are treated as the real data distribution. Generated sequences are sampled using different generative models under the same cell-type conditions as the test sequences to ensure a direct comparison.

Method. D3 learns a neural network $s_\theta(x, t)$ that models density ratios through score entropy loss (Lou et al., 2023; Sarkar et al., 2024), enabling the reverse diffusion process. The forward diffusion step perturbs sequences over time, and the model learns to reverse this process by estimating probability changes at each step. To enable categorical conditioning, we introduce Classifier-Free Guidance (CFG), where D3 is trained with and without conditioning labels to ensure robustness. During sampling, the conditional and unconditional score functions, $s_\theta(x, t, y)$ and $s_\theta(x, t, \phi)$, are combined as: $s_\theta^{CFG}(x, t, y) = \gamma s_\theta(x, t, y) + (1 - \gamma) s_\theta(x, t, \phi)$, where $\gamma > 1$ amplifies the conditioning effect, improving sampling quality. We follow the standard D3 sampling procedure using the adjusted score function to generate sequences for a given category y.

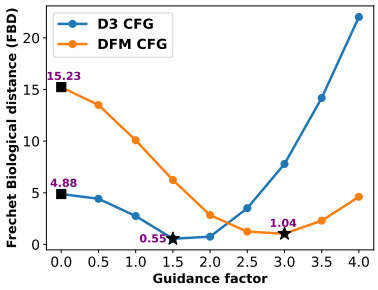

Figure 2: Performance comparison of DFM and D3 for cell type conditional enhancer design with varying CFG levels where cell types are sampled based on empirical frequency. ■ - Unconditional enhancer generation (for guidance factor 0), ⋆ - Best CFG guided generation.

Results. We compare D3 with DFM, as DFM has been shown to outperform other methods such as random sequence benchmarks, language models, and linear flow matching (Stark et al., 2024). We first evaluate by calculating the FBD between the models' generated sequences and the unconditional data distribution. D3 outperforms DFM by a large margin, which is shown by black squares in Fig. 2. Applying CFG to D3 further improves performance, reducing FBD to as low as 0.55. Notably, the performance of both methods varies with the guidance factor, reaching optimal levels before degrading at higher values. D3 achieves its best FBD (0.55) at guidance level 1.5, whereas DFM attains its lowest FBD (1.0) at level 3 (marked by black stars in Figure 2).

We also evaluate D3's ability to generate sequences conditioned on specific cell types. For this, we calculate FBD between generated sequences and real sequences from the test set belonging to a given cell type. Sequences are generated under the same conditioning labels as their real counterparts, ensuring a direct comparison. Additionally, we assess how well the generated sequences are classified as their intended cell types using the pretrained classifier from (Stark et al., 2024). We focus on three representative cell types (2, 68, and 16) used in (Stark et al., 2024) and evaluate performance across different guidance factors. D3 consistently achieves lower FBD and higher classification probabilities than DFM across all tested cell types (Figure 3 in Appendix). These results demonstrate the effectiveness of D3 in generating cell-type-specific sequences, with classifier-free guidance further enhancing performance. D3 significantly outperforms DFM in both unguided and guided scenarios, highlighting its potential for precise regulatory sequence generation.

5 INVESTIGATING DIFFUSION PROCESS THROUGH SEQUENCE EVOLUTION

Motivation. D3 has demonstrated strong potential for designing regulatory sequences, particularly in small data regimes and class-conditional generation. However, the underlying evolutionary pathways that guide sequence transformation during sampling remain unclear. Investigating these trajectories provides key insights into how sequences traverse the diffusion landscape, shedding light on the internal mechanisms of the model and its biological implications. By analyzing intermediate sequences throughout the sampling process, we aim to understand how the model navigates sequence space, whether it efficiently optimizes functional properties, and whether there are inherent inefficiencies in its generative trajectory.

Experiment. Using a transformer-based D3 model trained on the STARR-seq dataset, we examined sequences throughout the generative sampling process. To assess their functional and mechanistic properties, we utilized the DeepSTARR model (de Almeida et al., 2022) as an oracle to predict sequence activity and applied GradientSHAP from Captum (Kokhlikyan et al., 2020) with gradient correction (Majdandzic et al., 2023) to infer underlying cis-regulatory mechanisms. Sequence logos were plotted using Logomaker (Tareen & Kinney, 2020). This analysis allowed us to investigate how the model modifies sequences across sampling trajectories, revealing whether sequence evolution follows an efficient optimization pathway or encounters unnecessary deviations.

Results. To analyze how D3 modifies regulatory sequences during sampling, we tracked sequence activity at each step of the reverse diffusion process. We found that D3 consistently drives sequences toward low-activity states before optimizing them to the desired final condition. This was observed even when starting from sequences with already high activity in the target condition. For example, when evolving sequences from high-high Dev/HK activity to the same high-high state, D3 initially reduced activity before restoring it to the final target. This inefficiency suggests that the model does not retain beneficial sequence features but instead follows a generic reconstruction pathway, treating all sequences as if they originated from noise. This behavior was consistent across various initial-target conditions (Fig. 4).

To better understand this generative trajectory, we analyzed mutational dynamics across the sampling process. Sequences exhibited widespread mutations in the early steps, resembling a nearly complete mutagenesis of the original sequence, followed by more targeted refinements in later steps (Fig. 5). This suggests that early diffusion steps primarily erase existing sequence features, replacing them with background-like sequences before motifs are gradually reintroduced. Attribution analyses further confirmed this pattern. Using GradientSHAP to visualize regulatory motif structures during sampling, we observed that known motifs were lost in the early steps. Motif-like patterns only started re-emerging around step 65, with increasingly refined structures appearing by step 132 (Fig. 6). This pattern suggests that D3 does not operate by iteratively improving functional sequences but instead destroys and then reconstructs them.

To quantify this behavior across a broader set of sequences, we computed mutation fractions at each sampling step across 40k sequences. Mutation rates were highest in the initial steps and gradually declined around step 50, stabilizing near zero (Fig. 7). This supports the hypothesis that early steps aim to generate a generic sequence background before refining motifs.

6 CONCLUSION

Our study establishes DNA Discrete Diffusion as a robust framework for regulatory DNA sequence design, demonstrating strong performance in data-limited settings and effective conditional sequence generation. By leveraging a structured diffusion process, D3 generalizes cis-regulatory grammars beyond the constraints of supervised models, which often require large datasets and can struggle with spurious correlations. We show that D3 excels in generating functional regulatory sequences and enhancing predictive modeling by providing informative synthetic data. Its ability to integrate data augmentation and transfer learning further extends its utility, making it well-suited for applications where experimental data is scarce. The successful application of classifier-free guidance in D3 also underscores its flexibility in cell-type-specific sequence generation. While D3 significantly advances generative modeling for regulatory genomics, its mutation trajectory could be optimized to improve sampling efficiency. Early diffusion steps often transition sequences into

low-density regions before refinement, which could be mitigated with better-guided transitions. Future work should focus on refining diffusion dynamics to enhance efficiency and further improve the interpretability of generated sequences. Overall, D3 provides a scalable and biologically grounded approach to sequence design, expanding the capabilities of generative models in genomics. Future directions include experimental validation, continued refinement of sampling strategies, and further exploration of its potential for regulatory sequence engineering and synthetic biology applications.

REFERENCES

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pp. 1276–1301. PMLR, 2023.

Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.

Bernardo P de Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature genetics*, 54(5):613–624, 2022.

Bernardo P de Almeida, Christoph Schaub, Michaela Pagani, Stefano Secchia, Eileen EM Furlong, and Alexander Stark. Targeted design of synthetic enhancers for selected tissues in the drosophila embryo. *Nature*, 626(7997):207–211, 2024.

Lucas Ferreira DaSilva, Simon Senan, Zain Munir Patel, Aniketh Janardhan Reddy, Sameer Gabbita, Zach Nussbaum, Cesar Miguel Valdez Cordova, Aaron Wenteler, Noah Weber, Tin M Tunjic, et al. Dna-diffusion: Leveraging generative models for controlling chromatin accessibility and gene expression via synthetic regulatory elements. *bioRxiv*, pp. 2024–02, 2024.

Sager J Gosai, Rodrigo I Castro, Natalia Fuentes, John C Butts, Susan Kales, Ramil R Noche, Kousuke Mouri, Pardis C Sabeti, Steven K Reilly, and Ryan Tewhey. Machine-guided design of synthetic cell type-specific cis-regulatory elements. *bioRxiv*, 2023.

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

Connie Huang, Richard W Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail, and Nilah M Ioannidis. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55(12):2056–2059, 2023.

Jasper Janssens, Sara Aibar, Ibrahim Ihsan Taskiran, Joy N Ismail, Alicia Estacio Gomez, Gabriel Aughey, Katina I Spanier, Florian V De Rop, Carmen Bravo Gonzalez-Blas, Marc Dionne, et al. Decoding gene regulation in the fly brain. *Nature*, 601(7894):630–636, 2022.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Zehui Li, Yuhao Ni, William AV Beardall, Guoxuan Xia, Akashaditya Das, Guy-Bart Stan, and Yiren Zhao. Discdiff: Latent diffusion model for dna sequence generation. *arXiv preprint arXiv:2402.06079*, 2024.

Johannes Linder and Georg Seelig. Fast differentiable dna and protein sequence optimization for molecular design. *arXiv preprint arXiv:2005.11275*, 2020.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion language modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Antonio Majdandzic, Chandana Rajesh, and Peter K Koo. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biology*, 24(1):109, 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*, 2023.

Anirban Sarkar, Ziqi Tang, Chris Zhao, and Peter Koo. Designing dna with tunable regulatory activity using discrete diffusion. *bioRxiv*, pp. 2024–05, 2024.

Alexander Sasse, Bernard Ng, Anna E Spiro, Shinya Tasaki, David A Bennett, Christopher Gaiteri, Philip L De Jager, Maria Chikina, and Sara Mostafavi. Benchmarking of deep neural networks for predicting personal gene expression from dna sequence highlights shortcomings. *Nature Genetics*, 55(12):2060–2064, 2023.

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

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

Ziqi Tang, Shushan Toneyan, and Peter K Koo. Current approaches to genomic deep learning struggle to fully capture human genetic variation. *Nature Genetics*, 55(12):2021–2022, 2023.

Ammar Tareen and Justin B Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2020.

Ibrahim I Taskiran, Katina I Spanier, Hannah Dickmänken, Niklas Kempynck, Alexandra Pančíková, Eren Can Ekşi, Gert Hulselmans, Joy N Ismail, Koen Theunis, Roel Vandepoel, et al. Cell-type-directed design of synthetic enhancers. *Nature*, 626(7997):212–220, 2024.

Eeshit Dhaval Vaishnav, Carl G de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A Thompson, Joshua Z Levin, Francisco A Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory dna. *Nature*, 603(7901):455–463, 2022.