CONDITIONAL MONTE CARLO TREE DIFFUSION FOR DESIGNING CELL-TYPE-SPECIFIC AND BIOLOGICALLY FAITHFUL REGULATORY DNA

Animesh Awasthi[1,2], **Raphael Bednarsky**[1,2], **Moritz Schaefer**[1,2] **& Christoph Bock**[1,2*]

[1]Medical University of Vienna, Institute of Artificial Intelligence,
Center for Medical Data Science, Vienna, Austria
[2]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences,
Vienna, Austria

ABSTRACT

Designing regulatory DNA elements with precise cell-type-specific activity is broadly relevant for cell engineering and gene therapy. Deep generative models can generate functional gene-regulatory elements, but existing methods struggle to achieve high specificity against undesired cell types while adhering to the genome's natural regulatory grammar. Here, we introduce DNA-CRAFT, a generative framework that integrates class-conditioned discrete diffusion with Monte Carlo tree search to design cell-type-specific and biologically faithful regulatory elements. We first train a discrete diffusion model on the ENCODE registry of 3.2 million candidate regulatory elements. Second, we condition the model to learn class-specific regulatory grammars of naturally occurring DNA sequences, including enhancers and promoters. Third, we employ conditional Monte Carlo tree guidance, an inference-time alignment algorithm designed to maximize the differential regulatory activity between desired and undesired cell types. By benchmarking DNA-CRAFT on regulatory sequence design tasks for human cell lines and immune cell types, we demonstrate that our model generates sequences with high predicted cell-type-specific activity and biological fidelity, achieving the best trade-offs compared to methods that use diffusion, autoregressive models, and gradient-based optimization.

1 INTRODUCTION

Precise control of gene expression is at the heart of programmable biology, offering transformative potential for gene therapy, cell engineering, and synthetic biology Dunbar et al. (2018); Kitada et al. (2018); Yang et al. (2025); Butterfield et al. (2025). An important task is designing regulatory DNA elements that drive high levels of gene expression in desired cell types while minimizing expression in undesired cell types. For example, future gene therapies for Parkinson's disease may benefit from enhancers that enable highly specific gene expression in the neurons of the relevant brain region (the putamen) while avoiding unwanted expression in other brain areas and cell types. Such DNA-controlled specificity can compensate for insufficiently specific delivery of gene therapies and can enhance both the efficacy and safety profiles of future gene therapies Björklund & Davidsson (2021); Christine et al. (2022); Chen et al. (2023).

Naturally occurring regulatory DNA provides ample evidence that high cell-type-specific gene expression is feasible, mediated by combinations of transcription factor binding sites (TFBSs) as part of the genome's regulatory grammar Li et al. (2023); Mitra et al. (2024). However, the catalog of naturally occurring regulatory elements Moore et al. (2026) is often insufficient for applications in cell engineering and gene therapy, highlighting the need for methods to design synthetic regulatory DNA elements with desired properties. Initial attempts at machine learning-based enhancer design have been very successful de Almeida et al. (2024); Gosai et al. (2024); Taskiran et al. (2024), but

*Correspondence to: Christoph Bock <cbock@cemm.oeaw.ac.at>

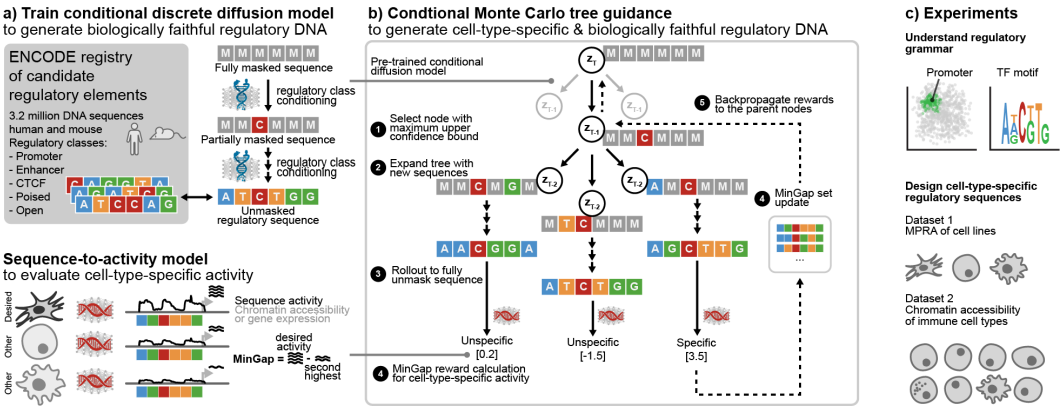

Figure 1: Overview of the DNA-CRAFT Framework. Panel (a) represents the class-conditioned discrete diffusion model trained on the ENCODE registry of naturally occurring regulatory elements. Panel (b) represents conditional Monte Carlo tree guidance with specificity rewards. Panel (c) represents applications of DNA-CRAFT for the design of cell-type-specific regulatory sequences.

they struggle to simultaneously optimize for two critical objectives: (i) achieving high activity in the desired cell type(s) while minimizing activity in a potentially large number of undesired cell types, and (ii) designing DNA sequences that closely resemble naturally occurring regulatory DNA, thereby reducing safety concerns in the context of medical applications DaSilva et al. (2026).

DNA sequence optimization methods Laarhoven & Aarts (1987); Angermueller et al. (2019); Sinai et al. (2020b); Linder & Seelig (2021); Reddy et al. (2024); Jain et al. (2022); Schreiber et al. (2025) utilize sequence-to-activity neural networks Avsec et al. (2021); Linder et al. (2025) to effectively maximize predicted activity; however, they may generate sequences that violate natural regulatory grammar and are prone to converging on local optima Vaishnav et al. (2022). Conversely, deep generative models, such as autoregressive genomic language models Gu et al. (2022); Nguyen et al. (2023); Schiff et al. (2024), excel at generating biologically faithful sequences, but they are difficult to steer towards high cell-type-specific activity without computationally expensive retraining or fine-tuning Lal et al. (2024); Yang et al.; Chen et al. (2025b).

Discrete diffusion language models Campbell et al. (2022); Austin et al. (2023); Lou et al. (2024); Sahoo et al. (2024); Shi et al. (2025) are a compelling choice for regulatory sequence design. These models capture the distribution of natural DNA sequences while overcoming the sequential constraints of autoregressive models through parallel and iterative refinement of long-range dependencies. They can be grounded with biological priors, such as regulatory element annotations (e.g., enhancers and promoters), using classifier-free guidance (CFG) Ho & Salimans (2022); Schiff et al. (2025). While these models ensure adherence to the natural regulatory grammar, the sampling process requires specialized guidance methods to generate sequences with the desired properties. Existing guidance methods, such as inference-time alignment algorithms, typically optimize for a single scalar reward Li et al. (2024); Phillips et al. (2024); Wu et al. (2024); Nisonoff et al. (2025); Wang et al. (2025), while lacking the planning capabilities required to solve the complex task of maximizing activity in desired cell types while minimizing activity in undesired cell types.

Here, we present **DNA-CRAFT** (**DNA C**is-**R**egulatory **A**rchitecture & **F**unction **T**uner), a framework that simultaneously optimizes for cell-type-specific activity and biological faithfulness. We employ conditional Monte Carlo tree diffusion tailored to the characteristics of regulatory DNA. First, to generate **biologically faithful** regulatory elements, we train a discrete diffusion language model (Figure 1a) on the ENCODE registry with over 3.2 million natural regulatory DNA elements in the human and mouse genomes Moore et al. (2026). Second, to generate **regulatory elements of different classes**, we use discrete classifier-free guidance Schiff et al. (2025) to learn the regulatory grammar of promoters, enhancers, and other naturally occurring DNA elements (Figure 1a). Third, to generate DNA sequences with high predicted **cell-type-specific activity**, we adapt Monte Carlo tree guidance Tang et al. (2024) to regulatory element design by introducing two key innovations (Figure 1b). (i) Since different classes of regulatory elements implement characteristic regulatory grammars Friedman et al. (2024), we constrain the tree search to the desired class using conditional diffusion sampling. (ii) To achieve high cell type specificity, we guide the tree search using a reward that explicitly maximizes the differential activity between desired and undesired cell types.

We benchmark DNA-CRAFT on two broadly relevant tasks of regulatory DNA design: (i) Generating cell type specific enhancers for three human cell lines and (ii) generating differentially chromatin-accessible sequences across eight immune cell types. Our results show that DNA-CRAFT effectively navigates these complex design spaces, achieving the best trade-off between predicted specificity and biological faithfulness compared to state-of-the-art methods that use diffusion, autoregressive models, and gradient-based optimization.

2 METHODS

This section outlines the methodology of the DNA-CRAFT framework. First, we use a masked diffusion language model (MDLM) to learn the genome's regulatory grammar Sahoo et al. (2024). Second, we use discrete classifier-free guidance (D-CFG) to train the generative model conditioned on regulatory element classes such as enhancers and promoters Schiff et al. (2025). Third, we employ class-conditioned sampling with Monte Carlo Tree Guidance (MCTG) for cell-type-specific regulatory element design Tang et al. (2024).

Notation. Let V denote the vocabulary size and $\mathcal{V} = \{A, C, G, T, \boldsymbol{m}\}$ denote the token vocabulary, where $\boldsymbol{m}$ is a special mask token used exclusively during the diffusion process. Thus $V = |\mathcal{V}| = 5$. Each token is represented as a one-hot vector in $\mathbb{R}^V$. Discrete variables are denoted by $\boldsymbol{z}_t, \boldsymbol{x} \in \mathcal{V}^L$, where $\boldsymbol{x}$ is a clean DNA sequence (containing only $\{A, C, G, T\}$) and $\boldsymbol{z}_t$ is the partially masked sequence at time t (containing tokens from $\mathcal{V}$, including $\boldsymbol{m}$). We write $\boldsymbol{x} \sim \mathrm{Cat}(\boldsymbol{x}; \boldsymbol{p})$ if $\boldsymbol{x}$ is drawn from a categorical distribution with parameter $\boldsymbol{p} \in \Delta^{V-1}$, the probability simplex. For length L sequences, we write $\boldsymbol{z}_t^{1:L}, \boldsymbol{x}^{1:L} \in \mathcal{V}^L$, with $\boldsymbol{z}_t^\ell, \boldsymbol{x}^\ell$ denoting the ℓth token.

2.1 DISCRETE DIFFUSION MODELS FOR DNA SEQUENCES

We use MDLM to reconstruct clean sequences from fully masked sequences. The diffusion process consists of a forward noising process followed by a reverse denoising process.

Forward noising process. The forward process progressively corrupts a clean, naturally occurring DNA sequence $\boldsymbol{x}$ into a sequence of discrete variables $\boldsymbol{z}_t$ over continuous time $t \in [0, 1]$, where clean tokens (base pairs) transition to the masked token $\boldsymbol{m}$. The marginal probability of a latent state $\boldsymbol{z}_t$ given the clean input DNA sequence $\boldsymbol{x}$ is defined as a categorical distribution $q(\boldsymbol{z}_t|\boldsymbol{x}) = \mathrm{Cat}(\boldsymbol{z}_t; \alpha_t \boldsymbol{x} + (1 - \alpha_t)\boldsymbol{m})$, where $\alpha_t \in [0, 1]$ is a monotonically decreasing noise schedule. Concretely, this means that at time t, each token independently remains as the original nucleotide with probability α_t or is replaced with probability $1 - \alpha_t$ by the absorbing state mask token $\boldsymbol{m}$.

Reverse denoising process. The reverse process involves learning to iteratively unmask a fully masked sequence. Conditioned on the clean DNA sequence $\boldsymbol{x}$, the time-reversal process for time steps $s < t$ is

$$q(\boldsymbol{z}_s|\boldsymbol{z}_t, \boldsymbol{x}) = \begin{cases} \mathrm{Cat}(\boldsymbol{z}_s; \boldsymbol{z}_t), & \boldsymbol{z}_t \neq \boldsymbol{m}, \\ \mathrm{Cat}\Big(\boldsymbol{z}_s; \dfrac{(1-\alpha_s)\boldsymbol{m} + (\alpha_s - \alpha_t)\boldsymbol{x}}{1 - \alpha_t}\Big), & \boldsymbol{z}_t = \boldsymbol{m}. \end{cases} \tag{1}$$

We train a backbone neural network denoted as $\boldsymbol{x}_\theta(\boldsymbol{z}_t, t)$ to approximate the clean DNA sequence $\boldsymbol{x}$ given the noisy input $\boldsymbol{z}_t$. The learned time reversal is effectively $p_\theta(\boldsymbol{z}_s|\boldsymbol{z}_t) = q\big(\boldsymbol{z}_s|\boldsymbol{z}_t, \boldsymbol{x}_\theta(\boldsymbol{z}_t, t)\big)$, with a loss function that simplifies to a weighted average of masked language modeling losses. To generate a DNA sequence of length L, we ancestrally sample from the learned reverse, starting with a completely masked sequence $\boldsymbol{z}_t^{1:L}$ where $t = 1$. Tokens at parallel positions are independently unmasked by discretizing the reverse diffusion process with a finite number of time steps from $t = 1$ to $t = 0$. This corresponds to the factorized reverse transition:

$$p_\theta(\boldsymbol{z}_s^{1:L}|\boldsymbol{z}_t^{1:L}) = \prod_{\ell=1}^{L} q(\boldsymbol{z}_s^\ell|\boldsymbol{z}_t^{1:L}, \boldsymbol{x}_\theta(\boldsymbol{z}_t^{1:L}, t)). \tag{2}$$

Model training. We train this model on the ENCODE registry of regulatory elements, comprising over 3.2 million DNA sequences from the human and mouse genomes (Figure 1a). We use the bidirectional Mamba Gu et al. (2022) state space model backbone for the denoising network $\boldsymbol{x}_\theta$. This architecture is well-suited for regulatory DNA sequences, as it scales linearly with sequence

length while incorporating inductive biases that preserve reverse-complement equivariance Schiff et al. (2024). Training details are in appendix A.2.

2.2 Classifier-Free Guidance for Class-Conditioned Diffusion Models

D-CFG constructs a class-conditioned MDLM by training on natural regulatory DNA elements and paired regulatory classes. This is achieved by training a conditional denoising diffusion network $p_\theta(\boldsymbol{z}_s^{1:L}|\boldsymbol{y}, \boldsymbol{z}_t^{1:L})$ alongside an unconditional one $p_\theta(\boldsymbol{z}_s^{1:L}|\boldsymbol{z}_t^{1:L})$, where $\boldsymbol{y} \in \{1, ..., K\}$ represents one of K regulatory element classes. At inference, given a guidance scale γ, we sample sequences of length L from the guided distribution as follows:

$$p_\theta^\gamma(\boldsymbol{z}_s^{1:L}|\boldsymbol{z}_t^{1:L}, \boldsymbol{y}) = \prod_{\ell=1}^{L} \frac{1}{Z^{(\ell)}} p_\theta(\boldsymbol{z}_s^\ell|\boldsymbol{y}, \boldsymbol{z}_t^{1:L})^\gamma p_\theta(\boldsymbol{z}_s^\ell|\boldsymbol{z}_t^{1:L})^{1-\gamma} \tag{3}$$

where $Z^{(\ell)} = \sum_{\boldsymbol{z}_s'} p_\theta(\boldsymbol{z}_s' \mid \boldsymbol{z}_t^{1:L}, y)^\gamma p_\theta(\boldsymbol{z}_s' \mid \boldsymbol{z}_t^{1:L})^{(1-\gamma)}$ is the per-token normalization constant.

Model training. We trained this model on regulatory sequences and their corresponding classes from the ENCODE registry (Figure 1a). We consolidated the more detailed ENCODE annotations into five classes: (i) Enhancers, combining proximal and distal enhancer-like sequences; (ii) Promoters; (iii) CTCF-bound elements, also known as insulators; (iv) Open chromatin, combining generic and TF-bound chromatin accessible sequences; and (v) Poised elements, combining TF-bound inaccessible sequences and H3K4me3 marked accessible sequences. (Appendix A.2)

2.3 Conditional Monte Carlo Tree Guidance for Cell-Type-Specific Regulatory Sequence Design

MCTG for regulatory sequences frames the iterative denoising process (Equation 2) as a Monte Carlo tree search. To improve cell-type-specific regulatory sequence design, we introduce two key adaptations. First, we explicitly constrain the tree search to the desired regulatory element class using our trained class-conditional model (Equation 3). Second, we use a specificity reward to steer the tree search and explicitly maximize the differential activity between desired and undesired cell types. The algorithm iterates through five steps: selection, expansion, rollout, reward calculation, and backpropagation, which are tailored to cell-type-specific regulatory sequence design (Figure 1b).

Search State. A node in the search tree represents a full-sequence latent state $\boldsymbol{z}_t^{1:L}$ at diffusion time t. Each node tracks a visit count $N_{\text{visit}}(\boldsymbol{z}_t)$ and a cumulative reward $W(\boldsymbol{z}_t)$. The search proceeds for N_{iter} iterations to identify branches with high rewards.

Initialization. We initialize the search tree with a root node $\boldsymbol{z}_1^{1:L}$, representing a fully masked sequence of length L. We specify a conditioning label $\boldsymbol{y}$ for the regulatory element class and a guidance scale γ that remains fixed throughout the search. We also initialize an empty set $\mathcal{G}^*$ to store the best sequences encountered during the search.

Step 1: Selection. At each iteration, we traverse the tree from the root to a leaf node by recursively selecting the child node that maximizes the Upper Confidence Bound (UCB). For a parent node $\boldsymbol{z}_t$ and a child $\boldsymbol{z}_s$, the selection score $U(\boldsymbol{z}_t, \boldsymbol{z}_s)$ is:

$$U(\boldsymbol{z}_t, \boldsymbol{z}_s) = \frac{W(\boldsymbol{z}_s)}{N_{\text{visit}}(\boldsymbol{z}_s)} + c_{\text{expl}} \cdot \frac{\sqrt{N_{\text{visit}}(\boldsymbol{z}_t)}}{1 + N_{\text{visit}}(\boldsymbol{z}_s)}, \tag{4}$$

where $W(\boldsymbol{z}_s)$ is the scalar cumulative reward of the child, initialized at 0 for unvisited child nodes. $N_{\text{visit}}(\cdot)$ denotes the visit count, and c_{expl} is a hyperparameter balancing exploration and exploitation.

Step 2: Expansion. Upon reaching a leaf node $\boldsymbol{z}_t$, we expand it by sampling M children. Unlike standard MCTG, which expands using the unconditional prior (Equation 2), we explicitly bias branching towards the target regulatory class $\boldsymbol{y}$ using the conditional reverse transition (Equation 3). We utilize the Gumbel-Max trick on the conditional probabilities $p_\theta^\gamma(\cdot|\boldsymbol{y})$ to sample children. For the i-th child, the latent state $\boldsymbol{z}_{s,i}$ is obtained via

$$\boldsymbol{z}_{s,i}^{1:L} \sim \text{Cat}\left(\text{softmax}\left(\log p_\theta^\gamma(\boldsymbol{z}_s|\boldsymbol{z}_t, \boldsymbol{y}) + \boldsymbol{G}_i\right)\right) \tag{5}$$

where $G_i \sim \text{Gumbel}(0, 1)$ are i.i.d. noise samples. This ensures that the expanded branches are not only diverse but also valid transitions under the specific grammar of the desired regulatory element class (e.g., enhancers).

Step 3: Rollout. To evaluate the activity of a newly expanded i-th child node $\boldsymbol{z}_{s,i}$, we estimate its terminal value. In contrast to greedily sampling tokens from the learned unconditional prior, we employ conditional ancestral sampling with a fixed guidance scale γ (Equation 3). Starting from time s, we iteratively sample until we reach the clean sequence $\boldsymbol{x_i}$. This conditional rollout ensures that the estimated sequence $\boldsymbol{x_i}$ is a valid member of the target class.

Step 4: Reward Calculation and Set Maintenance. The activity of sequence $\boldsymbol{x_i}$ is evaluated to calculate rewards. Standard MCTG typically uses Pareto dominance to maintain a frontier of non-dominated solutions. To enhance cell type specificity, we maximize the margin between desired and undesired cell types rather than simply selecting dominant sequences. We employ the **MinGap Score** Gosai et al. (2024) to explicitly optimize differential activity. Let $\mathcal{C}$ be the set of evaluated cell types, and $\boldsymbol{s}(\boldsymbol{x}_i) \in \mathbb{R}^{|\mathcal{C}|}$ be the activity vector predicted by a sequence-to-activity model. We partition $\mathcal{C}$ into desired ($\mathcal{C}^+$) and undesired ($\mathcal{C}^-$) subsets. Specificity $g(\boldsymbol{x}_i)$ is computed as the difference between the mean activity of the desired cell types and the maximum activity of the undesired cell types.

$$g(\boldsymbol{x}_i) = \underbrace{\frac{1}{|\mathcal{C}^+|}\sum_{c\in\mathcal{C}^+} s_c(\boldsymbol{x}_i)}_{\text{Mean Activity (Desired)}} - \underbrace{\max_{c\in\mathcal{C}^-} s_c(\boldsymbol{x}_i)}_{\text{Max Activity (Undesired)}} . \tag{6}$$

To encourage exploration of diverse, high-reward regions of the sequence space, we maintain a MinGap Set ($\mathcal{G}^*$), a bounded archive with a capacity of N_{max} that contains the highest-specificity sequences discovered so far. The sequence $\boldsymbol{x}_i$ is admitted to $\mathcal{G}^*$ if it improves the quality of the set (i.e., $g(\boldsymbol{x}_i) > \min_{\boldsymbol{x}\in\mathcal{G}^*} g(\boldsymbol{x})$). If the set exceeds capacity, the lowest scoring sequence is removed. This dynamic set serves as a baseline for calculating a relative reward $r(\boldsymbol{x}_i)$, which is highest for outstanding sequences but still provides a reward for non-optimal sequences to allow exploration beyond the current optimum in sequence space.

$$r(\boldsymbol{x}_i) = \frac{1}{|\mathcal{G}^*|}\sum_{\boldsymbol{x}^*\in\mathcal{G}^*} \mathbb{I}\big[g(\boldsymbol{x}_i) \geq g(\boldsymbol{x}^*)\big]. \tag{7}$$

Step 5: Backpropagation. The computed reward $r(\boldsymbol{x}_i)$ is backpropagated up the tree. For every node along the path from the expanded leaf to the root, we update the cumulative statistics to inform future selection steps.

$$W(\boldsymbol{z}_{\text{parent}}) \leftarrow W(\boldsymbol{z}_{\text{parent}}) + r(\boldsymbol{x}_i) \tag{8}$$

$$N_{\text{visit}}(\boldsymbol{z}_{\text{parent}}) \leftarrow N_{\text{visit}}(\boldsymbol{z}_{\text{parent}}) + 1. \tag{9}$$

After N_{iter} iterations, the method outputs the sequences in $\mathcal{G}^*$ as the final design candidates.

In summary, DNA-CRAFT combines the generative fidelity of class-conditioned discrete diffusion with the directed exploration of specificity-driven tree search, providing a framework for designing regulatory elements that adhere to natural grammar while achieving high cell-type specificity.

3 Experiments

3.1 Evaluation of the Generative Backbone

We first assess the fidelity of our discrete diffusion model in capturing the natural distribution of regulatory DNA.

Experimental Setup. To validate our architectural choice, we benchmarked the test-set perplexity (PPL) of the bidirectional Mamba (DiMamba) backbone against DNA convolutional neural network (DNA-Conv) and DNA diffusion transformer (DDiT) architectures, adapted from Stark et al. (2024); Sarkar et al. (2025); Sahoo et al. (2024), of comparable parameter sizes, trained on the same ENCODE registry dataset for 50,000 steps. Perplexity measures

Table 1: Evaluation of backbone architectures trained on the ENCODE registry. Shown are test-set perplexities (PPL; ↓), 3-mer Pearson correlation (↑), and JASPAR motif Spearman correlation (↑) relative to natural sequences.

Backbone	Test PPL ↓	3-mer Corr. ↑	Motif Corr. ↑
DNA-Conv	3.527	0.923	0.897
DDiT	3.510	0.873	0.860
DiMamba	**3.497**	**0.970**	**0.969**

the model's uncertainty in predicting unseen data, where lower values indicate better generalization. To further verify biological fidelity, we sampled 2,048 sequences from each backbone and compared them against a random subset of 2,048 sequences from the held-out test set. We evaluated low-level statistical alignment using the Pearson correlation of 3-mer counts and global regulatory grammar adherence using the Spearman correlation of TFBS frequency distributions with the JASPAR 2024 core vertebrate database Rauluseviciute et al. (2024) and FIMO Bailey et al. (2015).

Results. As shown in Table 1, the DiMamba backbone achieves the lowest perplexity, indicating superior generalization to natural sequences. Furthermore, the sequences exhibited a strong correlation with natural DNA 3-mer distributions ($r = 0.97$) and accurately recapitulated global TF motif frequencies ($r = 0.97$). This confirms that the base model learns realistic regulatory syntax, even without explicit conditioning. In summary, the DiMamba backbone-based MDLM provides a strong foundation for regulatory sequence generation, effectively capturing the regulatory grammar of natural DNA sequences.

3.2 Evaluation of Class-Conditioned Regulatory Grammar

Having established the fidelity of our discrete diffusion model, we next assessed whether our D-CFG-based class-conditioned prior effectively resolves the regulatory grammars associated with different regulatory element classes. We hypothesize that sequences generated under class-specific conditioning should exhibit differential enrichment of known TFBS consistent with biological priors, even without explicit supervision on motif position weight matrices.

Experimental Setup. We generated 2,048 DNA sequences for each of the five regulatory element classes using conditional sampling with a guidance scale of $\gamma = 3.0$. To assess motif enrichment, we scanned the sequences against the JASPAR 2024 core vertebrate database using the Analysis of Motif Enrichment (AME) tool from the MEME suite Bailey et al. (2009). Enrichment was calculated relative to a control set of shuffled sequences that preserved the dinucleotide frequencies of the generated set, ensuring that the results reflect regulatory grammar rather than general sequence composition. The statistical significance of motif enrichment was determined using Fisher's exact test, and the resulting p-values were transformed into Z-scores to standardize enrichment strength across motifs. To facilitate interpretation, we aggregated individual transcription factors into broad biological categories such as "Promoter Binding Factors" and computed the mean Z-score for each category.

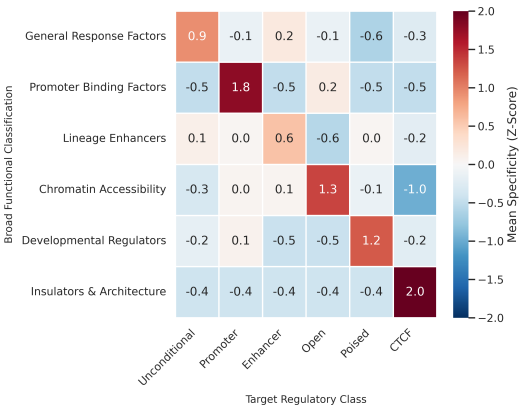

Figure 2: Motif enrichment Z-scores of generated regulatory DNA sequences, grouped by regulatory element classes (columns) and aggregated across broad biological categories (rows).

Results. Our analysis yields three key observations. First, DNA-CRAFT generates sequences rich in biologically relevant motifs, indicated by high positive Z-scores across all conditioned classes (Figure 2). Second, we observe distinct motif enrichment signatures for each class, confirming that the model has learned valid conditional distributions rather than generic regulatory signals. Third, these signatures align with biological knowledge; for instance, sequences conditioned on the Promoter class strongly enrich for promoter binding factors such as Nuclear Factor Y (NF-Y) and ETS-family proteins, whereas the Enhancer class is dominated by lineage-determining factors (e.g., SPIC, AP-2) that control cell-type identity. Similarly, the CTCF class enriches for CTCF, validating the model's ability to capture specific structural regulatory grammar. Detailed TFBS analysis is provided in appendix A.3. In summary, the class-conditioned diffusion model successfully learns and reproduces the motif compositions that define diverse regulatory elements, providing a biologically grounded foundation for targeted sequence design.

Table 2: Comparison of methods to design cell-type-specific enhancer across three human cell lines. Shown are the MinGap score, motif Spearman correlation, 3-mer Pearson correlation and the diversity. Values are reported as mean (std) over 3 independent runs.

Cell Line	Metric	SMC	CG	TDS	DRAKES	D3	Ledidi	Ctrl-DNA	DNA-CRAFT
HepG2	MinGap Eval ↑	1.614 (1.665)	-0.226 (0.096)	0.404 (0.569)	-1.401 (0.054)	0.046 (0.026)	5.771 (0.053)	**7.786 (0.070)**	4.346 (0.050)
	Motif Corr. ↑	0.554 (0.049)	0.860 (0.009)	0.397 (0.096)	0.057 (0.013)	0.869 (0.011)	0.584 (0.025)	0.629 (0.045)	**0.921 (0.006)**
	3-mer Corr. ↑	0.808 (0.102)	0.968 (0.003)	0.744 (0.098)	-0.361 (0.012)	0.975 (0.001)	0.755 (0.013)	0.494 (0.028)	**0.980 (0.009)**
	Diversity ↑	0.828 (0.432)	1.976 (0.002)	0.956 (0.096)	1.864 (0.002)	1.976 (0.004)	**1.981 (0.001)**	1.897 (0.026)	1.979 (0.000)
K562	MinGap Eval ↑	4.124 (0.893)	-0.003 (0.046)	1.622 (1.611)	-0.202 (0.067)	0.178 (0.066)	7.662 (0.154)	**9.067 (0.170)**	5.686 (0.043)
	Motif Corr. ↑	0.454 (0.025)	0.849 (0.026)	0.511 (0.130)	0.143 (0.024)	0.861 (0.041)	0.647 (0.039)	0.634 (0.084)	**0.933 (0.010)**
	3-mer Corr. ↑	0.659 (0.133)	0.940 (0.010)	0.647 (0.198)	-0.354 (0.007)	0.964 (0.018)	0.689 (0.022)	0.413 (0.058)	**0.976 (0.000)**
	Diversity ↑	0.309 (0.112)	1.977 (0.001)	0.637 (0.523)	1.958 (0.003)	1.977 (0.003)	1.980 (0.001)	1.896 (0.021)	**1.981 (0.001)**
SK-N-SH	MinGap Eval ↑	0.556 (0.146)	-0.278 (0.006)	0.186 (0.332)	0.094 (0.046)	-0.007 (0.006)	3.026 (0.222)	**3.720 (0.179)**	3.230 (0.022)
	Motif Corr. ↑	0.519 (0.155)	0.855 (0.026)	0.476 (0.092)	0.226 (0.017)	0.836 (0.009)	0.380 (0.043)	0.477 (0.037)	**0.881 (0.031)**
	3-mer Corr. ↑	0.775 (0.035)	0.949 (0.007)	0.719 (0.030)	-0.382 (0.001)	0.931 (0.011)	0.366 (0.019)	0.201 (0.172)	**0.969 (0.007)**
	Diversity ↑	1.269 (0.108)	1.976 (0.002)	0.918 (0.211)	1.826 (0.001)	1.969 (0.001)	**1.981 (0.002)**	1.855 (0.091)	1.976 (0.002)

3.3 Human Cell Line Specific Enhancer Design

We benchmarked DNA-CRAFT on the task of designing regulatory sequences specific to three human cell lines: HepG2 (liver), K562 (leukemia), and SK-N-SH (neuroblastoma).

Experimental Setup. We utilized a dataset of approximately 700,000 sequences with massively parallel reporter assay (MPRA) activity measurements across all three cell lines Gosai et al. (2024). We employed a split-model validation strategy following Wang et al. (2025). We randomly split the MPRA dataset into two halves. We fine-tuned the Enformer model Avsec et al. (2021) on the first half to create a `Design-Model`. We fine-tuned a separate instance of Enformer on the second half to serve as the `Evaluation-Model`. This prevents the generator from exploiting the reward model. Results were further confirmed using independent evaluation models (Appendix A.4).

We compared DNA-CRAFT (conditioned on "Enhancer", $\gamma = 3.0$) with state-of-the-art regulatory sequence design methods using the same `Design-Model`. We included inference-time alignment algorithms (SMC, TDS, CG) adapted to use the MinGap score, an RL-based diffusion fine-tuning method (DRAKES), CFG-based diffusion sampling (D3), constrained RL-based fine-tuning of autoregressive models (Ctrl-DNA), and gradient-based optimization (Ledidi). Implementation details are given in appendix A.4. Performance was assessed with four metrics: (i) **MinGap Score.** We computed the MinGap score (Equation 6) to measure differential activity in the desired cell type using the `Evaluation-Model`. (ii) **Motif Correlation.** For each desired cell type, we selected the top 99.9th percentile of real sequences ranked by their MinGap score of true MPRA activity. We scanned these top sequences with FIMO Bailey et al. (2015) and the JASPAR 2024 core vertebrate database to compute TFBS frequency distribution. This serves as the reference for evaluating the biological fidelity of our designed sequences. We compute the Spearman correlation between TFBS frequencies in the generated sequences and the reference. High correlations imply adherence to natural regulatory grammar. (iii) **3-mer Correlation.** We computed the Pearson correlation of 3-mer counts between the generated and top reference sequences, capturing sequence composition. (iv) **Diversity.** We calculate the mean per-position Shannon entropy across the generated batch. This metric quantifies the variability of nucleotides at each position, serving as a check against mode collapse.

Results. The benchmarking results highlight a trade-off between maximizing specificity and biological fidelity. In Table 2, we observe that DNA-CRAFT achieves higher predicted specificity scores compared to other diffusion-based alignment methods (SMC, TDS, CG) and fine-tuning approaches (DRAKES) across all cell lines. We note that DRAKES maximizes activity in the desired cell-type without explicitly minimizing background activity, resulting in low differential scores despite high overall activity. While methods like Ledidi and Ctrl-DNA optimize for the highest predicted differential activity across cell lines, they exhibit a reduction in motif and 3-mer correlations, suggesting a deviation from the natural regulatory grammar. DNA-CRAFT effectively navigates this landscape, achieving high predicted specificity while maintaining biological faithfulness (Figure 7).

3.4 Designing Immune Cell-State Specific Sequences

As a complementary test of specificity, we evaluated DNA-CRAFT's ability to differentiate between related cell types by designing sequences with differential chromatin accessibility across eight immune cell types: $CD8^+$ T cells, $CD4^+$ T cells, natural killer (NK) T cells, naive T cells, memory B cells, plasma B cells, macrophages, and mast cells.

Experimental Setup. We used a fine-tuned Enformer model on single-cell ATAC-seq profiles of immune cells from the CATlas dataset Zhang et al. (2021) as the prediction model Lal et al. (2025). The objective was to maximize chromatin accessibility in $CD8^+$ and $CD4^+$ T cells while minimizing it in other cell types, including B cells and naive T cells. We benchmarked DNA-CRAFT against methods that support plug-and-play inference (SMC, TDS, CG, Ledidi), excluding fine-tuning-based methods due to the computational cost of adapting to this multi-class setting. Evaluation followed the metrics defined in section 3.3, utilizing natural differentially accessible regions specific to $CD8^+$ / $CD4^+$ T cells as our reference for assessing biological fidelity.

Table 3: Benchmarking of methods to design T-cell specific sequences. Shown is the mean chromatin accessibility in $CD8^+$ and $CD4^+$ T-cell (desired cell types), MinGap differential accessibility, motif Spearman correlation, and 3-mer Pearson correlation. Values are reported as mean (std) over 3 independent runs.

Metric	SMC	CG	TDS	Ledidi	DNA-CRAFT
Mean T cell accessibility. ↑	0.011 (0.008)	0.065 (0.011)	0.029 (0.007)	0.348 (0.018)	**0.512 (0.015)**
MinGap accessibility. ↑	-0.029 (0.010)	-0.062 (0.020)	-0.046 (0.036)	-0.063 (0.003)	**0.123 (0.010)**
Motif Corr. ↑	0.385 (0.133)	0.898 (0.043)	0.670 (0.130)	0.384 (0.061)	**0.928 (0.011)**
3-mer Corr. ↑	0.790 (0.087)	**0.979 (0.009)**	0.817 (0.009)	0.482 (0.038)	0.967 (0.010)
Diversity ↑	1.278 (0.141)	1.977 (0.002)	1.393 (0.243)	**1.983 (0.001)**	1.978 (0.002)

Results. As shown in Table 3, DNA-CRAFT is the only method that achieves a positive MinGap score, successfully decoupling chromatin accessibility in $CD8^+$ / $CD4^+$ T cells from closely related cell types. While Ledidi achieves high mean predicted accessibility in the desired cell types, we observe low specificity due to insufficient suppression in the undesired cell types. Conversely, inference-time alignment methods (SMC, CG, TDS) struggle to shift the distribution towards the desired cell types. These results suggest that DNA-CRAFT effectively optimizes specificity while maintaining high biological fidelity.

To better understand how it navigates this fitness landscape, we tracked the optimization trajectory of the sequences discovered along the MinGap set during a single tree search (Figure 3). This plot tracks the predicted chromatin accessibility of all cell types across successive discovery steps. Initially, the search yields sequences with low-to-moderate activity and minimal specificity. As the tree search progresses, we observe a sharp rise in chromatin accessibility for the desired cell types, and this activity gain did not come at the cost of specificity. In summary, these results demonstrate that DNA-CRAFT effectively navigates the complex landscape of regulatory activity across multiple cell types, generating regulatory elements of the desired class that achieve highly cell-type-specific activity without sacrificing biological fidelity.

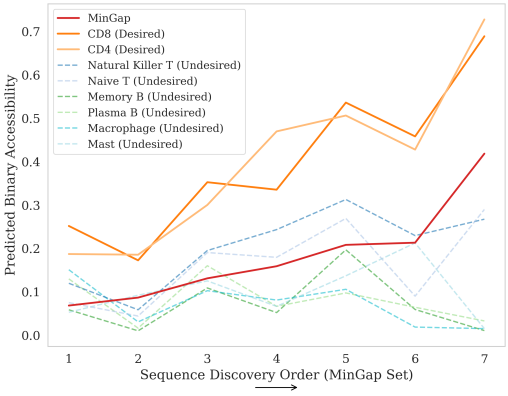

Figure 3: The trajectory of predicted accessibility of sequences discovered along the MinGap Set during a single tree search.

4 Conclusion & Discussion

We introduced DNA-CRAFT, a generative framework for designing regulatory DNA elements that maximize predicted cell-type specificity while adhering to the natural regulatory grammars of the genome. By integrating discrete diffusion, class-specific guidance, and conditional Monte Carlo tree search, DNA-CRAFT generates biologically faithful and functionally active DNA sequences, as demonstrated in benchmarks across human cell lines and immune cell types.

Despite its promising results, this study has certain limitations. First, our method relies on sequence-to-activity models and is therefore limited by the performance and biases of the surrogate model. Second, while inference-time alignment avoids retraining, the computational cost of Monte Carlo

tree search constrains large-scale library design comprising millions of sequences. Third, while we demonstrate high biological fidelity, experimental validation is ultimately needed to confirm the activity of the sequences.

Future work will explore the use of pre-trained DNA foundation model backbones, such as Nucleotide Transformer v3 Boshar et al. (2025), to enhance learned representations. We also plan to extend conditioning to diverse tissues and species, enabling an improved understanding of tissue-specific regulatory grammars and evolutionary constraints. Finally, using sequence-to-activity models, we intend to apply DNA-CRAFT to clinically relevant cell types and experimentally validate its designs.

REFERENCES

Vikram Agarwal, Fumitaka Inoue, Max Schubach, Dmitry Penzar, Beth K. Martin, Pyaree Mohan Dash, Pia Keukeleire, Zicong Zhang, Ajuni Sohota, Jingjing Zhao, Ilias Georgakopoulos-Soares, William S. Noble, Galip Gürkan Yardımcı, Ivan V. Kulakovskiy, Martin Kircher, Jay Shendure, and Nadav Ahituv. Massively parallel characterization of transcriptional regulatory elements. *Nature*, 639(8054):411–420, March 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08430-9. URL https://www.nature.com/articles/s41586-024-08430-9. Publisher: Nature Publishing Group.

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. September 2019. URL https://openreview.net/forum?fileGuid=3xgr169o12oUrbxS&id=HklxbgBKvr&ref=https%3A%2F%2Fgithubhelp.com.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces, February 2023. URL http://arxiv.org/abs/2107.03006. arXiv:2107.03006 [cs].

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.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18 (10):1196–1203, 2021.

Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl_2):W202–W208, July 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp335. URL https://doi.org/10.1093/nar/gkp335.

Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.

Tomas Björklund and Marcus Davidsson. Next-Generation Gene Therapy for Parkinson's Disease Using Engineered Viral Vectors. *Journal of Parkinson's Disease*, 11(s2):S209–S217, June 2021. ISSN 1877-7171. doi: 10.3233/JPD-212674. URL https://journals.sagepub.com/action/showAbstract. Publisher: SAGE Publications.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence-function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963. URL https://www.biorxiv.org/content/early/2025/12/25/2025.12.22.695963.

Gabriel L. Butterfield, Samuel J. Reisman, Nahid Iglesias, and Charles A. Gersbach. Gene regulation technologies for gene and cell therapy. *Molecular Therapy*, 33(5):2104–2122, May 2025. ISSN

1525-0016. doi: 10.1016/j.ymthe.2025.04.004. URL https://www.sciencedirect.com/science/article/pii/S1525001625002783.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A Continuous Time Framework for Discrete Denoising Models, October 2022. URL http://arxiv.org/abs/2205.14987. arXiv:2205.14987 [stat].

Zubing Cao, Timothy S. Carey, Avishek Ganguly, Catherine A. Wilson, Soumen Paul, and Jason G. Knott. Transcription factor AP-2 induces early Cdx2 expression and represses HIPPO signaling to specify the trophectoderm lineage. *Development*, 142(9):1606–1615, May 2015. ISSN 0950-1991. doi: 10.1242/dev.120238. URL https://doi.org/10.1242/dev.120238.

Viorica Chelban, Nisha Patel, Jana Vandrovcova, M. Natalia Zanetti, David S. Lynch, Mina Ryten, Juan A. Botía, Oscar Bello, Eloise Tribollet, Stephanie Efthymiou, Indran Davagnanam, SYNAPSE Study Group, Fahad A. Bashiri, Nicholas W. Wood, James E. Rothman, Fowzan S. Alkuraya, and Henry Houlden. Mutations in NKX6-2 Cause Progressive Spastic Ataxia and Hypomyelination. *American Journal of Human Genetics*, 100(6):969–977, June 2017. ISSN 1537-6605. doi: 10.1016/j.ajhg.2017.05.009.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design, 2025a. URL https://arxiv.org/abs/2505.07086.

Xingyu Chen, Shihao Ma, Runsheng Lin, Jiecong Lin, and Bo Wang. Ctrl-DNA: Controllable Cell-Type-Specific Regulatory DNA Design via Constrained RL, May 2025b. URL http://arxiv.org/abs/2505.20578. arXiv:2505.20578 [cs].

Yefei Chen, Zexuan Hong, Jingyi Wang, Kunlin Liu, Jing Liu, Jianbang Lin, Shijing Feng, Tianhui Zhang, Liang Shan, Taian Liu, Pinyue Guo, Yunping Lin, Tian Li, Qian Chen, Xiaodan Jiang, Anan Li, Xiang Li, Yuantao Li, Jonathan J. Wilde, Jin Bao, Ji Dai, and Zhonghua Lu. Circuit-specific gene therapy reverses core symptoms in a primate Parkinson's disease model. *Cell*, 186(24):5394–5410.e18, November 2023. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2023.10.004. URL https://www.cell.com/cell/abstract/S0092-8674(23)01086-3. Publisher: Elsevier.

Chadwick W. Christine, R. Mark Richardson, Amber D. Van Laar, Marin E. Thompson, Elisabeth M. Fine, Omar S. Khwaja, Chunming Li, Grace S. Liang, Andreas Meier, Eiry W. Roberts, Madeline L. Pfau, Josh R. Rodman, Krystof S. Bankiewicz, and Paul S. Larson. Safety of AADC Gene Therapy for Moderately Advanced Parkinson Disease. *Neurology*, 98(1):e40–e50, January 2022. doi: 10.1212/WNL.0000000000012952. URL https://www.neurology.org/doi/10.1212/WNL.0000000000012952. Publisher: Wolters Kluwer.

Peter N. Cockerill. NFAT Is Well Placed to Direct Both Enhancer Looping and Domain-Wide Models of Enhancer Function. *Science Signaling*, 1(13):pe15–pe15, April 2008. doi: 10.1126/stke.113pe15. URL https://www.science.org/doi/10.1126/stke.113pe15. Publisher: American Association for the Advancement of Science.

Timothy Dahlem, Scott Cho, Gerald J. Spangrude, Janis J. Weis, and John H. Weis. Overexpression of Snai3 suppresses lymphoid- and enhances myeloid-cell differentiation. *European Journal of Immunology*, 42(4):1038–1043, 2012. ISSN 1521-4141. doi: 10.1002/eji.201142193. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eji.201142193. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eji.201142193.

Lucas Ferreira DaSilva, Simon Senan, Judith F. Kribelbauer-Swietek, Zain Munir Patel, Lithin Karmel Louis, Aniketh Janardhan Reddy, Sameer Gabbita, Jonathan D. Rosen, Zach Nussbaum, César Miguel Valdez Córdova, Aaron Wenteler, Noah Weber, Tin M. Tunjic, Martino Mansoldo, Talha Ahmad Khan, Gue-Ho Hwang, Vincent Gardeux, David T. Humphreys, Cameron Smith, Matei Bejan, Peter Bromley, Will Connell, Bart Deplancke, Michael I. Love, Emily S. Wong, Wouter Meuleman, and Luca Pinello. Designing synthetic regulatory elements using the generative AI framework DNA-Diffusion. *Nature Genetics*, 58(1):180–194, January 2026. ISSN 1546-1718. doi: 10.1038/s41588-025-02441-6. URL https://www.nature.com/articles/s41588-025-02441-6. Publisher: Nature Publishing Group.

Bernardo P. de Almeida, Christoph Schaub, Michaela Pagani, Stefano Secchia, Eileen E. M. Furlong, and Alexander Stark. Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. *Nature*, 626(7997):207–211, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06905-9. URL https://www.nature.com/articles/s41586-023-06905-9. Publisher: Nature Publishing Group.

Bieke Decaesteker, Geertrui Denecker, Christophe Van Neste, Emmy M. Dolman, Wouter Van Loocke, Moritz Gartlgruber, Carolina Nunes, Fanny De Vloed, Pauline Depuydt, Karen Verboom, Dries Rombaut, Siebe Loontiens, Jolien De Wyn, Waleed M. Kholosy, Bianca Koopmans, Anke H. W. Essing, Carl Herrmann, Daniel Dreidax, Kaat Durinck, Dieter Deforce, Filip Van Nieuwerburgh, Anton Henssen, Rogier Versteeg, Valentina Boeva, Gudrun Schleiermacher, Johan van Nes, Pieter Mestdagh, Suzanne Vanhauwaert, Johannes H. Schulte, Frank Westermann, Jan J. Molenaar, Katleen De Preter, and Frank Speleman. TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets. *Nature Communications*, 9(1):4866, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06699-9. URL https://www.nature.com/articles/s41467-018-06699-9. Publisher: Nature Publishing Group.

Cynthia E. Dunbar, Katherine A. High, J. Keith Joung, Donald B. Kohn, Keiya Ozawa, and Michel Sadelain. Gene therapy comes of age. *Science*, 359(6372):eaan4672, January 2018. doi: 10.1126/science.aan4672. URL https://www.science.org/doi/10.1126/science.aan4672. Publisher: American Association for the Advancement of Science.

Meyer J. Friedman, Tobias Wagner, Haram Lee, Michael G. Rosenfeld, and Soohwan Oh. Enhancer–promoter specificity in gene transcription: molecular mechanisms and disease associations. *Experimental & Molecular Medicine*, 56(4):772–787, April 2024. ISSN 2092-6413. doi: 10.1038/s12276-024-01233-y. URL https://www.nature.com/articles/s12276-024-01233-y. Publisher: Nature Publishing Group.

Sager J Gosai, Rodrigo I Castro, Natalia Fuentes, John C Butts, Kousuke Mouri, Michael Alasoadura, Susan Kales, Thanh Thanh L Nguyen, Ramil R Noche, Arya S Rao, et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, pp. 1–10, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces, August 2022. URL http://arxiv.org/abs/2111.00396. arXiv:2111.00396 [cs].

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Caixia Hou and Oleg V Tsodikov. Structure and cooperative formation of a FLI1 filament on contiguous GGAA DNA sites. *Nucleic Acids Research*, 53(6):gkaf205, April 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf205. URL https://doi.org/10.1093/nar/gkaf205.

Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.

Fumiki Katsuoka, Hozumi Motohashi, Ko Onodera, Naruyoshi Suwabe, James Douglas Engel, and Masayuki Yamamoto. One enhancer mediates mafK transcriptional activation in both hematopoietic and cardiac muscle cells. *The EMBO Journal*, 19(12):2980–2991, June 2000. ISSN 1460-2075. doi: 10.1093/emboj/19.12.2980. URL https://doi.org/10.1093/emboj/19.12.2980.

W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12 (6):996–1006, June 2002. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.229102. URL http://genome.cshlp.org/content/12/6/996. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Tasuku Kitada, Breanna DiAndreth, Brian Teague, and Ron Weiss. Programming gene and engineered-cell therapies with synthetic biology. *Science*, 359(6376):eaad1067, 2018. doi: 10.1126/science.aad1067. URL https://www.science.org/doi/abs/10.1126/science.aad1067.

P. J. van Laarhoven and E. H. Aarts. *Simulated Annealing: Theory and Applications*. Springer Science & Business Media, June 1987. ISBN 978-90-277-2513-4.

Avantika Lal, David Garfield, Tommaso Biancalani, and Gokcen Eraslan. Designing realistic regulatory dna with autoregressive language models. *Genome Research*, 34(9):1411–1420, 2024.

Avantika Lal, Laura Gunsalus, Surag Nair, Tommaso Biancalani, and Gokcen Eraslan. gReLU: a comprehensive framework for DNA sequence modeling and design. *Nature Methods*, 22 (11):2253–2257, November 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02868-z. URL https://www.nature.com/articles/s41592-025-02868-z. Publisher: Nature Publishing Group.

Anne-Sophie Laramée, Hannah Raczkowski, Peng Shao, Carolina Batista, Devanshi Shukla, Li Xu, S. M. Mansour Haeryfar, Yodit Tesfagiorgis, Steven Kerfoot, and Rodney DeKoter. Opposing Roles for the Related ETS-Family Transcription Factors Spi-B and Spi-C in Regulating B Cell Differentiation and Function. *Frontiers in Immunology*, 11, May 2020. ISSN 1664-3224. doi: 10.3389/fimmu.2020.00841. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2020.00841/full. Publisher: Frontiers.

Bum-Kyu Lee, Yu jin Jang, Mijeong Kim, Lucy LeBlanc, Catherine Rhee, Jiwoon Lee, Samuel Beck, Wenwen Shen, and Jonghwan Kim. Super-enhancer-guided mapping of regulatory networks controlling mouse trophoblast stem cells. *Nature Communications*, 10(1):4749, October 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12720-6. URL https://www.nature.com/articles/s41467-019-12720-6. Publisher: Nature Publishing Group.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding, October 2024. URL http://arxiv.org/abs/2408.08252. arXiv:2408.08252 [cs].

Yang Eric Li, Sebastian Preissl, Michael Miller, Nicholas D. Johnson, Zihan Wang, Henry Jiao, Chenxu Zhu, Zhaoning Wang, Yang Xie, Olivier Poirion, Colin Kern, Antonio Pinto-Duarte, Wei Tian, Kimberly Siletti, Nora Emerson, Julia Osteen, Jacinta Lucero, Lin Lin, Qian Yang, Quan Zhu, Nathan Zemke, Sarah Espinoza, Anna Marie Yanny, Julie Nyhus, Nick Dee, Tamara Casper, Nadiya Shapovalova, Daniel Hirschstein, Rebecca D. Hodge, Sten Linnarsson, Trygve Bakken, Boaz Levi, C. Dirk Keene, Jingbo Shang, Ed Lein, Allen Wang, M. Margarita Behrens, Joseph R. Ecker, and Bing Ren. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science*, 382(6667):eadf7044, 2023. doi: 10.1126/science.adf7044. URL https://www.science.org/doi/abs/10.1126/science.adf7044.

Johannes Linder and Georg Seelig. Fast activation maximization for molecular sequence design. *BMC bioinformatics*, 22(1):510, October 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04437-5.

Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R. Kelley. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57(4):949–961, April 2025. ISSN 1546-1718. doi: 10.1038/s41588-024-02053-6. URL https://www.nature.com/articles/s41588-024-02053-6. Publisher: Nature Publishing Group.

Wolfgang Link and Bibiana I. Ferreira. FOXO Transcription Factors: A Brief Overview. In Wolfgang Link (ed.), *FOXO Transcription Factors: Methods and Protocols*, pp. 1–8. Springer US, New York, NY, 2025. ISBN 978-1-07-164217-7. doi: 10.1007/978-1-0716-4217-7_1. URL https://doi.org/10.1007/978-1-0716-4217-7_1.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Giulia Lunazzi, Maria Buxadé, Marta Riera-Borrull, Laura Higuera, Sarah Bonnin, Hector Huerga Encabo, Silvia Gaggero, Diana Reyes-Garau, Carlos Company, Luca Cozzuto, Julia Ponomarenko, José Aramburu, and Cristina López-Rodríguez. NFAT5 Amplifies Antipathogen Responses by Enhancing Chromatin Accessibility, H3K27 Demethylation, and Transcription Factor Recruitment. *Journal of Immunology*, 206(11):2652–2667, June 2021. ISSN 1550-6606. doi: 10.4049/jimmunol.2000624. Place: Baltimore, Md. : 1950.

Raleigh E. Malik and Simon J. Rhodes. The role of DNA methylation in regulation of the murine Lhx3 gene. *Gene*, 534(2):272–281, January 2014. ISSN 1879-0038. doi: 10.1016/j.gene.2013.10.045.

Sneha Mitra, Rohan Malik, Wilfred Wong, Afsana Rahman, Alexander J. Hartemink, Yuri Pritykin, Kushal K. Dey, and Christina S. Leslie. Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nature Genetics*, 56 (4):627–636, April 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01689-8. URL https://www.nature.com/articles/s41588-024-01689-8. Publisher: Nature Publishing Group.

Young-Mee Moon, Seon-Yeong Lee, Seung-Ki Kwok, Seung Hoon Lee, Deokhoon Kim, Woo Kyung Kim, Yang-Mi Her, Hea-Jin Son, Eun-Kyung Kim, Jun-Geol Ryu, Hyeon-Beom Seo, Jeong-Eun Kwon, Sue-Yun Hwang, Jeehee Youn, Rho H. Seong, Dae-Myung Jue, Sung-Hwan Park, Ho-Youn Kim, Sung-Min Ahn, and Mi-La Cho. The Fos-Related Antigen 1–JUNB/Activator Protein 1 Transcription Complex, a Downstream Target of Signal Transducer and Activator of Transcription 3, Induces T Helper 17 Differentiation and Promotes Experimental Autoimmune Arthritis. *Frontiers in Immunology*, 8, December 2017. ISSN 1664-3224. doi: 10.3389/fimmu.2017.01793. URL https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2017.01793/full. Publisher: Frontiers.

Jill E. Moore, Henry E. Pratt, Kaili Fan, Nishigandha Phalke, Jonathan Fisher, Shaimae I. Elhajjajy, Gregory Andrews, Mingshi Gao, Nicole Shedd, Yu Fu, Matthew C. Lacadie, Jair Meza, Mansi Khandpekar, Mohit Ganna, Eva Choudhury, Ross Swofford, Huong Phan, Christian C. Ramirez, Maxwell Campbell, Mary Likhite, Nina P. Farrell, Annika K. Weimer, Anusri Pampari, Vivekanandan Ramalingam, Fairlie Reese, Beatrice Borsari, Xuezhu Yu, Eve Wattenberg, Marina Ruiz-Romero, Milad Razavi-Mohseni, Jinrui Xu, Timur Galeev, Andres Colubri, Michael A. Beer, Roderic Guigó, Mark B. Gerstein, Jesse M. Engreitz, Mats Ljungman, Timothy E. Reddy, Michael P. Snyder, Charles B. Epstein, Elizabeth Gaskell, Bradley E. Bernstein, Diane E. Dickel, Axel Visel, Len A. Pennacchio, Ali Mortazavi, Anshul Kundaje, and Zhiping Weng. An expanded registry of candidate cis-regulatory elements. *Nature*, pp. 1–10, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09909-9. URL https://www.nature.com/articles/s41586-025-09909-9. Publisher: Nature Publishing Group.

Hozumi Motohashi, Fumiki Katsuoka, Jordan A. Shavit, James Douglas Engel, and Masayuki Yamamoto. Positive or Negative MARE-Dependent Transcriptional Regulation Is Determined by the Abundance of Small Maf Proteins. *Cell*, 103(6):865–876, December 2000. ISSN 0092-8674, 1097-4172. doi: 10.1016/S0092-8674(00)00190-2. URL https://www.cell.com/cell/abstract/S0092-8674(00)00190-2. Publisher: Elsevier.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

Maria Nilsson, Karin Dahlman-Wright, Charlotta Karelmo, Jan-rAke Gustafsson, and Knut R. Steffensen. Elk1 and SRF transcription factors convey basal transcription and mediate glucose response via their binding sites in the human LXRB gene promoter. *Nucleic Acids Research*, 35(14):4858–4868, July 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm492. URL https://doi.org/10.1093/nar/gkm492.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking Guidance for Discrete State-Space Diffusion and Flow Models, March 2025. URL http://arxiv.org/abs/2406.01572. arXiv:2406.01572 [cs].

Andrew J. Oldfield, Telmo Henriques, Dhirendra Kumar, Adam B. Burkholder, Senthilkumar Cinghu, Damien Paulet, Brian D. Bennett, Pengyi Yang, Benjamin S. Scruggs, Christopher A. Lavender, Eric Rivals, Karen Adelman, and Raja Jothi. NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nature Communications*, 10(1):3072, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10905-7. URL https://www.nature.com/articles/s41467-019-10905-7. Publisher: Nature Publishing Group.

Erin K. O'Shea, Rheba Rutkowski, and Peter S. Kim. Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. *Cell*, 68(4):699–708, February 1992. ISSN 0092-8674, 1097-4172. doi: 10.1016/0092-8674(92)90145-3. URL https://www.cell.com/cell/abstract/0092-8674(92)90145-3. Publisher: Elsevier.

Angus Phillips, Hai-Dang Dau, Michael John Hutchinson, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Particle Denoising Diffusion Sampler, June 2024. URL http://arxiv.org/abs/2402.06320. arXiv:2402.06320 [stat].

Julia R. Pon and Marco A. Marra. MEF2 transcription factors: developmental regulators and emerging cancer genes. *Oncotarget*, 7(3):2297–2312, October 2015. ISSN 1949-2553. doi: 10.18632/oncotarget.6223. URL https://www.oncotarget.com/article/6223/text/. Publisher: Impact Journals.

Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen, Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W Wasserman, Franccois Parcy, and Anthony Mathelier. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, January 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1059. URL https://doi.org/10.1093/nar/gkad1059.

Aniketh Janardhan Reddy, Xinyang Geng, Michael H. Herschl, Sathvik Kolli, Aviral Kumar, Patrick D. Hsu, Sergey Levine, and Nilah M. Ioannidis. Designing Cell-Type-Specific Promoter Sequences Using Conservative Model-Based Optimization, June 2024. URL https://www.biorxiv.org/content/10.1101/2024.06.23.600232v1. Pages: 2024.06.23.600232 Section: New Results.

Gang Ren, Wenfei Jin, Kairong Cui, Joseph Rodrigez, Gangqing Hu, Zhiying Zhang, Daniel R. Larson, and Keji Zhao. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Molecular Cell*, 67(6):1049–1058.e6, September 2017. ISSN 1097-2765. doi: 10.1016/j.molcel.2017.08.026. URL https://www.cell.com/molecular-cell/abstract/S1097-2765(17)30624-X. Publisher: Elsevier.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and Effective Masked Diffusion Language Models, November 2024. URL http://arxiv.org/abs/2406.07524. arXiv:2406.07524 [cs].

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

Anirban Sarkar, Yijie Kang, Nirali Somia, Pablo Mantilla Puccetti, Jessica Zhou, Masayuki Nagai, Ziqi Tang, Chris Zhao, and Peter K. Koo. Designing DNA With Tunable Regulatory Activity Using Score-Entropy Discrete Diffusion, May 2025. URL https://www.biorxiv.org/content/10.1101/2024.05.23.595630v2. Pages: 2024.05.23.595630 Section: New Results.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling, June 2024. URL http://arxiv.org/abs/2403.03234. arXiv:2403.03234 [q-bio].

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple

Guidance Mechanisms for Discrete Diffusion Models, May 2025. URL http://arxiv.org/abs/2412.10193. arXiv:2412.10193 [cs].

Jacob Schreiber, Franziska Katharina Lorbeer, Monika Heinzl, Franziska Reiter, Baptiste Rafanel, Yang Young Lu, Alexander Stark, and William Stafford Noble. Programmatic design and editing of cis-regulatory elements, December 2025. URL https://www.biorxiv.org/content/10.1101/2025.04.22.650035v2. ISSN: 2692-8205 Pages: 2025.04.22.650035 Section: New Results.

Yoshiyuki Seki. PRDM14 Is a Unique Epigenetic Regulator Stabilizing Transcriptional Networks for Pluripotency. *Frontiers in Cell and Developmental Biology*, 6, February 2018. ISSN 2296-634X. doi: 10.3389/fcell.2018.00012. URL https://www.frontiersin.org/journals/cell-and-developmental-biology/articles/10.3389/fcell.2018.00012/full. Publisher: Frontiers.

Conglin Shi, Liuting Chen, Hui Pi, Henglu Cui, Chenyang Fan, Fangzheng Tan, Xuanhao Qu, Rong Sun, Fengbo Zhao, Yihua Song, Yuanyuan Wu, Miaomiao Chen, Wenkai Ni, Lishuai Qu, Renfang Mao, and Yihui Fan. Identifying a locus in super-enhancer and its resident NFE2L1/MAFG as transcriptional factors that drive PD-L1 expression and immune evasion. *Oncogenesis*, 12 (1):56, November 2023. ISSN 2157-9024. doi: 10.1038/s41389-023-00500-3. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10662283/.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and Generalized Masked Diffusion for Discrete Data, January 2025. URL http://arxiv.org/abs/2406.04329. arXiv:2406.04329 [cs].

Javier E Sierra-Pagan, Nikita Dsouza, Satyabrata Das, Thijs A Larson, Jacob R Sorensen, Xiao Ma, Patricia Stan, Erik J Wanberg, Xiaozhong Shi, Mary G Garry, Wuming Gong, and Daniel J Garry. FOXK1 regulates Wnt signalling to promote cardiogenesis. *Cardiovascular Research*, 119(8):1728–1739, June 2023. ISSN 0008-6363. doi: 10.1093/cvr/cvad054. URL https://doi.org/10.1093/cvr/cvad054.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D Kelsic. Adalead: A simple and robust adaptive greedy search algorithm for sequence design. *arXiv preprint arXiv:2010.02141*, 2020a.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric D. Kelsic. AdaLead: A simple and robust adaptive greedy search algorithm for sequence design, October 2020b. URL http://arxiv.org/abs/2010.02141. arXiv:2010.02141 [cs].

Scott Smemo, Luciene C. Campos, Ivan P. Moskowitz, José E. Krieger, Alexandre C. Pereira, and Marcelo A. Nobrega. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Human Molecular Genetics*, 21(14):3255–3263, July 2012. ISSN 1460-2083. doi: 10.1093/hmg/dds165.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design, 2024. URL https://arxiv.org/abs/2402.05841.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. PepTune: De Novo Generation of Therapeutic Peptides with Multi-Objective-Guided Discrete Diffusion, December 2024. URL https://arxiv.org/abs/2412.17780v4.

Ibrahim I. Taskiran, Katina I. Spanier, Hannah Dickmänken, Niklas Kempynck, Alexandra Pančíková, Eren Can Ekcsi, Gert Hulselmans, Joy N. Ismail, Koen Theunis, Roel Vandepoel, Valerie Christiaens, David Mauduit, and Stein Aerts. Cell-type-directed design of synthetic enhancers. *Nature*, 626(7997):212–220, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06936-2. URL https://www.nature.com/articles/s41586-023-06936-2. Publisher: Nature Publishing Group.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review, 2024. URL https://arxiv.org/abs/2407.13734.

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review, 2025. URL https://arxiv.org/abs/2501.09685.

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.

Romain Villot, Audrey Poirier, Inan Bakan, Karine Boulay, Erlinda Fernández, Romain Devillers, Luciano Gama-Braga, Laura Tribouillard, Andréanne Gagné, Éma Duchesne, Danielle Caron, Jean-Sébastien Bérubé, Jean-Christophe Bérubé, Yan Coulombe, Michèle Orain, Yves Gélinas, Stéphane Gobeil, Yohan Bossé, Jean-Yves Masson, Sabine Elowe, Steve Bilodeau, Venkata Manem, Philippe Joubert, Frédérick A. Mallette, and Mathieu Laplante. ZNF768 links oncogenic RAS to cellular senescence. *Nature Communications*, 12(1):4841, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24932-w. URL https://www.nature.com/articles/s41467-021-24932-w. Publisher: Nature Publishing Group.

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Tommaso Biancalani, Avantika Lal, Tommi Jaakkola, Sergey Levine, Hanchen Wang, and Aviv Regev. Fine-Tuning Discrete Diffusion Models via Reward Optimization with Applications to DNA and Protein Design, March 2025. URL http://arxiv.org/abs/2410.13643. arXiv:2410.13643 [cs].

Lu Wang, Tianhui Liu, Linjie Xu, Ya Gao, Yonglong Wei, Caiwen Duan, Guo-Qiang Chen, Shuo Lin, Roger Patient, Bo Zhang, Dengli Hong, and Feng Liu. Fev regulates hematopoietic stem cell development via ERK signaling. *Blood*, 122(3):367–375, July 2013. ISSN 0006-4971. doi: 10.1182/blood-2012-10-462655. URL https://doi.org/10.1182/blood-2012-10-462655.

Xin Wang, Anjun Jiao, Lina Sun, Wenhua Li, Biao Yang, Yanhong Su, Renyi Ding, Cangang Zhang, Haiyan Liu, Xiaofeng Yang, Chenming Sun, and Baojun Zhang. Zinc finger protein Zfp335 controls early T-cell development and survival through -selection-dependent and -independent mechanisms. *eLife*, 11:e75508, February 2022. ISSN 2050-084X. doi: 10.7554/eLife.75508. URL https://doi.org/10.7554/eLife.75508. Publisher: eLife Sciences Publications, Ltd.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and Asymptotically Exact Conditional Sampling in Diffusion Models, November 2024. URL http://arxiv.org/abs/2306.17775. arXiv:2306.17775 [stat].

Jian Xu and Wei Du. HES6: an emerging player in human hematopoiesis. *Haematologica*, 109 (11):3466–3468, May 2024. ISSN 1592-8721. doi: 10.3324/haematol.2024.285426. URL https://haematologica.org/article/view/haematol.2024.285426.

Carol H. Yan, Martin Levesque, Suzanne Claxton, Randy L. Johnson, and Siew-Lan Ang. Lmx1a and Lmx1b Function Cooperatively to Regulate Proliferation, Specification, and Differentiation of Midbrain Dopaminergic Progenitors. *Journal of Neuroscience*, 31(35):12413–12425, August 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1077-11.2011. URL https://www.jneurosci.org/content/31/35/12413. Publisher: Society for Neuroscience Section: Articles.

Jinshou Yang, Feihan Zhou, Xiyuan Luo, Yuan Fang, Xing Wang, Xiaohong Liu, Ruiling Xiao, Decheng Jiang, Yuemeng Tang, Gang Yang, Lei You, and Yupei Zhao. Enhancer reprogramming: critical roles in cancer and promising therapeutic strategies. *Cell Death Discovery*, 11(1):84, March 2025. ISSN 2058-7716. doi: 10.1038/s41420-025-02366-3. URL https://www.nature.com/articles/s41420-025-02366-3. Publisher: Nature Publishing Group.

Zhao Yang, Bing Su, Chuan Cao, and Ji-Rong Wen. Regulatory dna sequence design with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.

Hongyong Zhang, Zechen Li, Yanmei Zhu, Wencong Lyu, Wenlu Wei, Haochen Wang, Shuangjie Tian, Wei Yue, Jiajing Zhong, Qing-Yuan Sun, and Yiting Guan. Fosl2 facilitates chromatin accessibility to determine developmental events during follicular maturation. *Nature Communications*, 16(1):8955, October 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-64009-6. URL https://www.nature.com/articles/s41467-025-64009-6. Publisher: Nature Publishing Group.

Kai Zhang, James D. Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B. Poirion, Yunjiang Qiu, Yang E. Li, Kyle J. Gaulton, Allen Wang, Sebastian Preissl, and Bing Ren. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24):5985–6001.e19, November 2021. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2021.10.024. URL https://www.cell.com/cell/abstract/S0092-8674(21)01279-4. Publisher: Elsevier.

A SUPPLEMENTARY MATERIAL AND APPENDIX

The appendix consists of five sections. The first section (Appendix A.1) is an extension of the introduction, discussing work related to DNA-CRAFT. The second section (Appendix A.2) covers the training details of DNA-CRAFT's class-conditioned diffusion model. The third section (Appendix A.3) includes supplementary data for the classifier-free guidance related experiments. The fourth section (Appendix A.4) covers details of the regulatory sequence design tasks, and the last section (Appendix A.5) consists of ablation studies conducted for DNA-CRAFT.

A.1 RELATED WORK

Machine learning methods for regulatory element design rely on sequence-to-activity neural networks. These models are trained on large-scale genomic data to predict gene expression or chromatin accessibility from DNA sequences Avsec et al. (2021); Linder et al. (2025). By generalizing to unseen sequences, these models enable the design of regulatory elements with the desired activity. Existing approaches generally fall into two distinct categories: DNA sequence optimization methods and generative AI-based methods.

DNA Sequence Optimization Methods. Sequence optimization methods treat regulatory DNA design as a search problem. These approaches utilize a pre-trained sequence-to-activity model as a reward function to guide the search toward sequences with high activity. Classical greedy approaches, including simulated annealing Laarhoven & Aarts (1987), *in silico* mutagenesis, AdaLead Sinai et al. (2020a), and gradient-based algorithms Schreiber et al. (2025), start the process with random or known sequences and iteratively mutate them to maximize predicted activity. More recently, reinforcement learning (RL) techniques, such as DyNA-PPO Angermueller et al. (2019) and GFlowNets Jain et al. (2022), have been used to navigate this sequence search space more effectively. Because these methods focus on maximizing a single scalar reward, they may generate sequences that violate the regulatory grammar of natural DNA, and they are prone to converging on local optima Vaishnav et al. (2022).

Generative AI for Regulatory Sequence Design. Generative AI approaches learn the underlying distribution of natural DNA sequences. Recent methods utilize autoregressive genomic language models (LMs) Nguyen et al. (2023); Schiff et al. (2024); Lal et al. (2024) and discrete diffusion models Avdeyev et al. (2023); Sarkar et al. (2024); DaSilva et al. (2026) to capture long-range sequence patterns with high fidelity. For example, the masked diffusion language model (MDLM) achieved strong performance on genomic benchmarks Sahoo et al. (2024). To use such generative priors for optimizing cell-type-specific regulatory activity, three alternative guidance mechanisms exist. **Classifier-free guidance** builds class-conditioned discrete diffusion models Schiff et al. (2025). In the context of regulatory sequence design, these models can be conditioned on sequences with high activity in desired cell types and then used to sample active sequences. **RL-based fine-tuning** updates the weights of the generative model to maximize a reward function Uehara et al. (2024). Methods such as Ctrl-DNA use genomic LMs with constrained RL for cell-type-specific designs Chen et al. (2025b). DRAKES backpropagates the reward gradients using the Gumbel-Softmax trick to fine-tune a discrete diffusion model for designing highly active enhancers Wang et al. (2025). While effective, fine-tuning is computationally expensive and requires updating the model parameters for every new design objective. **Inference-time alignment** methods steer the sampling process of a frozen diffusion model using external guidance, avoiding the cost of retraining Uehara et al. (2025). Techniques like soft value-based decoding (SVDD) Li et al. (2024), sequential monte carlo (SMC) Phillips et al. (2024), twisted diffusion sampling (TDS) Wu et al. (2024), and classifier guidance (CG) Nisonoff et al. (2025) use reward-weighted resampling or auxiliary gradients computed by external sequence-to-activity models to steer the generative process. These methods typically optimize for high activity in one cell type, which often leads to high background activity in undesired cell types. Multi-objective inference-time alignment methods could address this and have been used successfully in other application areas, including peptide design Tang et al. (2024); Chen et al. (2025a).

A.2 DNA-CRAFT DIFFUSION MODEL TRAINING DETAILS

Dataset Curation. We trained DNA-CRAFT using the ENCODE Registry of candidate cis-Regulatory Elements V4 Moore et al. (2026). To capture cross-species regulatory grammar and maximize biological diversity, we integrated data from both human (hg38) and mouse (mm10) genomes. The final dataset comprises 2,348,854 annotated regions across 1,888 human cell types and 926,843 annotated regions across 366 mouse cell types.

Data Pre-Processing. DNA sequences corresponding to the regulatory regions were extracted from their respective reference genomes. All sequences were centered and padded to a fixed length of 350 base pairs (bp) using a distinct `[PAD]` token. To accommodate variable effective lengths, we implemented a masking mechanism within both the model backbone and the diffusion loss function, ensuring that padding tokens are excluded from the generative process. We applied reverse complement augmentation by taking either the forward or reverse complement strand with equal probability during training.

Class Consolidation. The ENCODE V4 Registry categorizes regulatory elements based on biochemical signatures derived from DNase hypersensitivity, histone modifications, and CTCF binding. The original classification schema distinguishes between:

- **Promoter-like (PLS):** Accessible regions proximal to transcription start sites (TSS) enriched for H3K4me3.
- **Enhancer-like (ELS):** Accessible regions enriched for H3K27ac, further stratified into proximal (pELS) and distal (dELS) based on TSS proximity.
- **CA-H3K4me3:** Accessible regions with H3K4me3 enrichment but low H3K27ac, often indicative of poised or primed regulatory states.
- **CA-CTCF:** Accessible regions enriched for CTCF binding sites with low histone acetylation, often indicating insulators.
- **CA-TF, CA, and TF:** Elements defined primarily by chromatin accessibility or transcription factor binding, lacking canonical histone marks.

We consolidated these fine-grained categories into five broad functional classes for DNA-CRAFT conditioning (Table 4).

Table 4: Mapping of ENCODE V4 cCRE classifications to DNA-CRAFT conditioning labels.

DNA-CRAFT Class	Original ENCODE Class
Promoter	PLS
Enhancer	dELS, pELS
CTCF	CA-CTCF
Poised	CA-H3K4me3, TF
Open Chromatin	CA, CA-TF

Cross-Species Train-Test-Validation Split. To ensure generalization and prevent overfitting due to sequence homology, we constructed our data splits using a graph-based clustering approach adapted from the Enformer training protocol Avsec et al. (2021).

We constructed an undirected graph $G = (V, E)$ in which the vertices V represent approximately 3.2 million processed regulatory elements. Edges E were defined to capture both orthology and sequence similarity:

1. **Homology Edges:** An edge connects a human cCRE and a mouse cCRE if they share a sequence alignment of > 100 bp, determined via the hg38-mm10 syntenic nets Kent et al. (2002).
2. **Overlap Edges:** Within a single species, an edge connects any two cCREs that share sequence similarity of more than 100 bp.

We computed the connected components of G to define independent sequence clusters. These clusters were randomly partitioned into training (90%), validation (5%), and test (5%) sets, ensuring a strict separation of homologous sequences across splits.

Hyperparameters and Models. We trained the model on the processed ENCODE dataset using the AdamW optimizer. To evaluate the impact of conditional generation, we trained two variants of DNA-CRAFT for 100 epochs each:

1. **Unconditional Model:** Trained without classifier-free guidance (CFG) using a global batch size of 1,024.
2. **Conditional Model:** Trained with CFG (dropout $p = 0.1$) using a larger global batch size of 4,096 to ensure the representation of all 5 regulatory classes within every batch update.

Model hyperparameters are provided in Table 5.

Table 5: DNA-CRAFT Hyperparameters. The model utilizes a bidirectional DiMamba backbone.

Hyperparameter	**Value**
Backbone Architecture (DiMamba)	
Model Parameters	1.93 Million
Sequence Length (L)	350
Hidden Dimension (d_{model})	128
Mamba Blocks	10
Bidirectional Strategy	Addition (Tied weights)
Dropout	0.1
Diffusion Process	
Noise Schedule	Cosine
Time Conditioning	True
Optimization	
Optimizer	AdamW
Learning Rate	1×10^{-3}
Training Epochs	100
Batch Size	1,024 (Uncond.) / 4,096 (Cond.)
Conditioning (CFG)	
Condition Type	Regulatory Element Class
Number of Classes	5
Condition Dropout	0.1
Conditioning Dim	128

Computational Infrastructure. All experiments were conducted on NVIDIA H100 GPUs (80GB VRAM). The validation loss trajectories for both model variants are visualized in Figure 4.

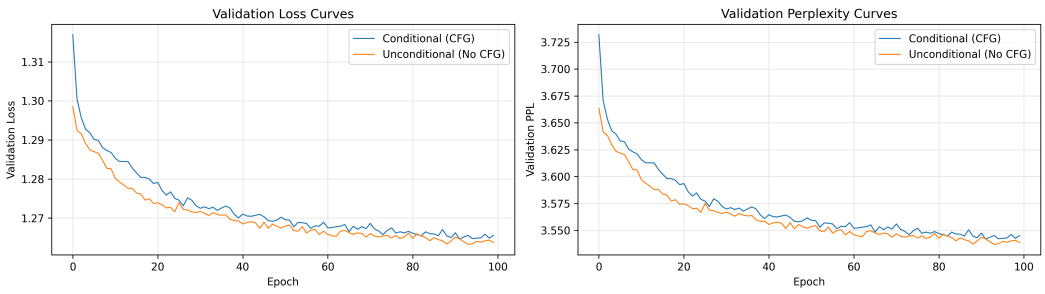

Figure 4: Validation loss curves for Unconditional and Conditional DNA-CRAFT models over 100 epochs.

A.3 Class-Conditioned Sampling and Motif Analysis

Latent Space Organization. To understand how class-specific sequence features are encoded within the model's representations, we analyzed the latent space of the generated sequences. We extracted latent representations from the final layer of the pre-trained DiMamba backbone for all test-set sequences. The embeddings were projected into two dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE) with a perplexity of 30 and a cosine distance metric. As shown in Figure 5, the model learns to separate sequences into clusters corresponding to their conditioning labels.

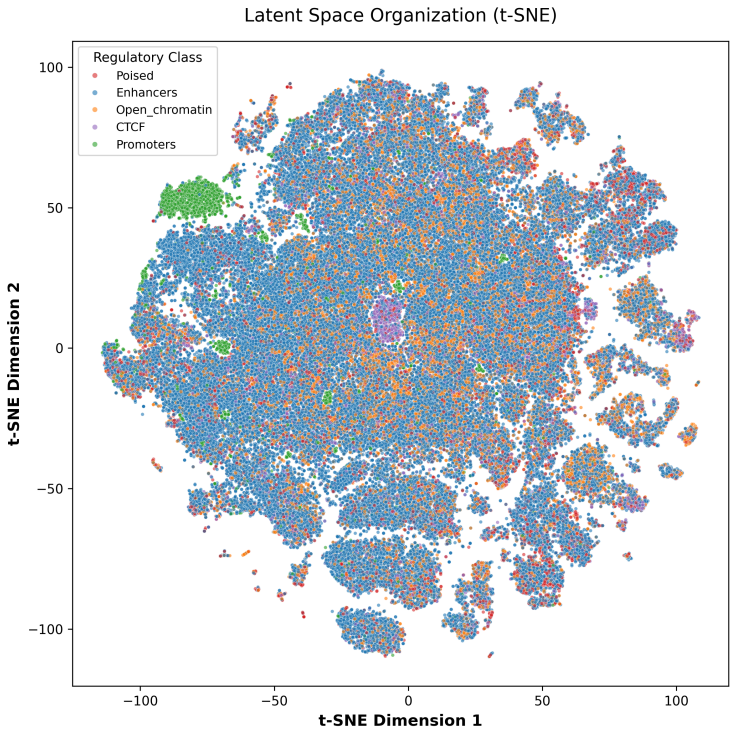

Figure 5: Latent space organization of test set sequence embeddings. t-SNE visualization of the final-layer representations, colored by the conditioning regulatory label.

Detailed Motif Enrichment Analysis. To validate the biological relevance of the class-conditioned generated sequences, we performed a motif enrichment analysis. Figure 6 displays a heatmap of the Z-scores for the top 5 most specific motifs per class. The recovery of these specific factors aligns with known biological priors. Table 6 lists the top 5 enriched TFs for each class. To provide biological context, Table 7 maps these factors to their broad functional categories based on prior literature.

A.4 Regulatory Sequence Design Benchmarks

Benchmarking Model Implementation Details. For all benchmark comparisons, we generated 128 sequences per experimental run. All experiments were repeated using independent random seeds to ensure statistical robustness. To ensure a fair comparison, all methods utilized the exact same `Design-Model` for optimization or guidance.

1. **Ledidi:** We initialized the optimization with 128 random DNA sequences. We implemented a custom differentiable wrapper to compute MinGap scores from `Design-Model`'s output and backpropagate gradients directly to the input sequence representation. Optimization was conducted independently for each sequence for a maximum of 20,000 steps to ensure convergence.

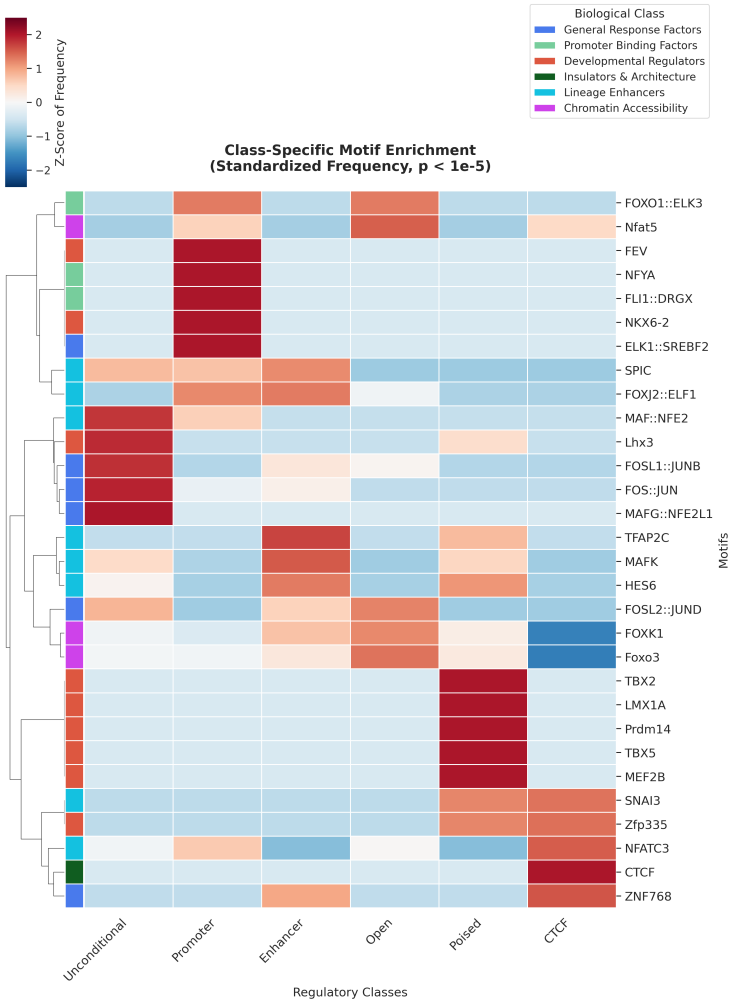

Figure 6: Heatmap of Z-scores for the top 5 most enriched motifs (rows) and their corresponding regulatory sequence class (columns).

Table 6: Top 5 Enriched Transcription Factors per Regulatory Class. Factors are ranked by enrichment Z-score.

Regulatory Class	Top Enriched Motifs
CTCF	CTCF, ZNF768, NFATC3, Zfp335, SNAI3
Enhancer	TFAP2C, MAFK, HES6, FOXJ2::ELF1, SPIC
Open Chromatin	Nfat5, Foxo3, FOXO1::ELK3, FOSL2::JUND, FOXK1
Poised	LMX1A, TBX5, Prdm14, MEF2B, TBX2
Promoter	NKX6-2, FEV, NFYA, FLI1::DRGX, ELK1::SREBF2
Unconditional	MAFG::NFE2L1, FOS::JUN, Lhx3, FOSL1::JUNB, MAF::NFE2

2. **Classifier Guidance (CG):** We used DNA-CRAFT's unconditional base diffusion model, trained on the entire ENCODE registry of regulatory elements without further fine-tuning for any experiments. We performed 128 parallel diffusion inference steps using our pre-trained unconditional diffusion model. Gradients were computed via the `Design-Model` wrapper to guide the sampling trajectory. We utilized a guidance scale of $\gamma = 1000$.

Table 7: Biological Functional Categories of Enriched Transcription Factors. Classifications are derived from established biological literature Motohashi et al. (2000); Shi et al. (2023); O'Shea et al. (1992); Malik & Rhodes (2014); Moon et al. (2017); Villot et al. (2021); Decaesteker et al. (2018); Pon & Marra (2015); Seki (2018); Zhang et al. (2025); Xu & Du (2024); Cao et al. (2015); Wang et al. (2013); Chelban et al. (2017); Oldfield et al. (2019); Hou & Tsodikov (2025); Nilsson et al. (2007); Katsuoka et al. (2000); Lee et al. (2019); Laramée et al. (2020); Lunazzi et al. (2021); Link & Ferreira (2025); Sierra-Pagan et al. (2023); Yan et al. (2011); Smemo et al. (2012); Ren et al. (2017); Cockerill (2008); Wang et al. (2022); Dahlem et al. (2012).

Biological Category	**Transcription Factors**
Chromatin Accessibility	Nfat5, Foxo3, FOXO1::ELK3, FOSL2::JUND, FOXK1
Developmental Regulators	Lhx3, NKX6-2, FEV, LMX1A, TBX5, Prdm14, MEF2B, TBX2, Zfp335
General Response Factors	MAFG::NFE2L1, FOS::JUN, FOSL1::JUNB, ZNF768
Insulators & Architecture	CTCF
Lineage Enhancers	MAF::NFE2, TFAP2C, MAFK, HES6, FOXJ2::ELF1, SPIC, NFATC3, SNAI3
Promoter Binding Factors	NFYA, FLI1::DRGX, ELK1::SREBF2

3. **Sequential Monte Carlo (SMC):** We used the same unconditional base diffusion model and MinGap wrapper as the CG method. The sampling process tracked 128 particles and applied a resampling parameter of $\alpha = 0.5$.
4. **Twisted Diffusion Sampling (TDS):** We followed the same setup as CG and SMC for this baseline. We applied a guidance scale of $\gamma = 1000$ and a resampling parameter of $\alpha = 0.5$.
5. **D3 (Discrete Denoising Diffusion):** We trained the discrete diffusion model utilizing the transformer backbone (2M parameters) for a 100 epochs on the MPRA dataset ($\sim$700,000 sequences) with the cell type activity as classes. We generated 128 sequences per cell type with conditional sampling ($\gamma = 4.0$).
6. **Ctrl-DNA:** We followed the original protocol and hyperparameters specified for each target cell type. We fine-tuned three separate models (one for each cell line: HepG2, K562, SK-N-SH) for 100 optimization steps. The top 128 sequences from each fine-tuning run were selected for evaluation. We note that Ctrl-DNA's base autoregressive model Nguyen et al. (2023) is trained on the entire human reference genome, which is a substantially larger and more diverse training corpus than the ENCODE registry of 3.2 million cCREs used to train DNA-CRAFT's diffusion backbone. Additionally, the TFBS regularization term was excluded, as the specific motif lists were not available in the public repository.
7. **DRAKES:** We followed the protocol described in the original publication. For the HepG2 cell line, we utilized the provided pre-trained checkpoint. Similarly, we fine-tuned two separate models for K562 and SK-N-SH. We generated sequences using the respective fine-tuned models. We note that DRAKES fine-tunes its diffusion model using reward gradients computed over the full MPRA dataset ($\sim$700,000 sequences), which encompasses the data used to train both our `Design-Model` and `Evaluation-Model`. This gives DRAKES implicit access to the evaluation distribution during its fine-tuning process.
8. **DNA-CRAFT (Ours):** We employ DNA-CRAFT's class-conditioned base diffusion model, which is trained on the ENCODE dataset without further fine-tuning for all experiments. Candidate sequences were generated using conditional Monte Carlo tree guidance. We selected the final candidate from the MinGap set $\mathcal{G}^*$ at the end of the tree search. Table 8 details the specific inference configuration.

Extended Benchmark Results We extended our evaluation to include cross-architecture and cross-study validation metrics. To validate the robustness and biological plausibility of our designs beyond the training distribution, we employed external models from Lal et al. (2025) to perform cross-model and cross-study evaluations:

Table 8: DNA-CRAFT Inference Parameters. Settings for the MCTS-guided diffusion sampling.

Parameter	Value
Diffusion Sampling	
Sampling Steps	64
Classifier-Free Guidance	
Guidance Scale (γ)	3.0
Conditioning Class	Enhancer
Monte Carlo Tree Guidance	
Total Iterations	64
Number of Children	128
Exploration Coefficient	0.5
N_{max}	64

- **Complete Dataset Validation:** We evaluated the predicted activity of the generated sequences using a model trained on the full MPRA dataset Gosai et al. (2024), ensuring that the performance was not an artifact of the dataset split used during optimization.
- **Cross-Study Validation:** We utilized models trained on an independent MPRA study involving HepG2, K562, and induced pluripotent stem cells (WTC11) by Agarwal et al. (2025). We analyzed sequences designed for the shared HepG2 and K562 cell lines to verify their generalizability across different experiments. MinGap scores for the shared cell lines were calculated considering all three cell lines, despite not actively optimizing for WTC11 during the design process. Since this study did not include SK-N-SH cells, we excluded them from the analysis.
- **Cross-Modality Validation:** We predicted binary chromatin accessibility using an independent classifier and calculated the proportion of designed sequences with a predicted accessibility probability > 0.5.

Table 9 presents the comprehensive performance across all metrics. In figure 7, we visualize the trade-off between cell-type-specific activity and biological fidelity.

Table 9: Comparison of methods to design enhancer specific across three human cell lines. Shown are various MinGap scores from three different studies, motif correlation, 3-mer correlation, fraction accessibility and diversity. Values are reported as mean (std) over 3 independent runs.

Cell Line	Metric	SMC	CG	TDS	DRAKES	Ledidi	Ctrl-DNA	DNA-CRAFT
HepG2	MinGap Eval ↑	1.614 (1.665)	-0.226 (0.096)	0.404 (0.569)	-1.401 (0.054)	5.771 (0.053)	**7.786 (0.070)**	4.346 (0.050)
	MinGap (Full MPRA) ↑	1.472 (1.473)	-0.235 (0.074)	0.482 (0.673)	-1.351 (0.036)	6.484 (0.065)	**8.572 (0.186)**	4.501 (0.053)
	MinGap (Agarwal) ↑	0.359 (0.571)	-0.189 (0.017)	-0.161 (0.103)	-0.847 (0.068)	2.057 (0.043)	**3.232 (0.179)**	1.500 (0.059)
	Motif Corr. ↑	0.554 (0.049)	0.860 (0.009)	0.397 (0.096)	0.057 (0.013)	0.584 (0.025)	0.629 (0.045)	**0.921 (0.006)**
	3-mer Corr. ↑	0.808 (0.102)	0.968 (0.003)	0.744 (0.098)	-0.361 (0.012)	0.755 (0.013)	0.494 (0.028)	**0.980 (0.009)**
	Fraction Acc. ↑	0.281 (0.474)	0.016 (0.008)	0.141 (0.141)	0.940 (0.012)	0.914 (0.021)	**1.000 (0.000)**	0.966 (0.020)
	Diversity ↑	0.828 (0.432)	1.976 (0.002)	0.956 (0.096)	1.864 (0.002)	**1.981 (0.001)**	1.897 (0.026)	1.979 (0.000)
K562	MinGap Eval ↑	4.124 (0.893)	-0.003 (0.046)	1.622 (1.611)	-0.202 (0.067)	7.662 (0.154)	**9.067 (0.170)**	5.686 (0.043)
	MinGap (Full MPRA) ↑	4.166 (1.050)	-0.031 (0.041)	1.523 (1.536)	-0.170 (0.048)	8.395 (0.163)	**9.874 (0.212)**	5.831 (0.059)
	MinGap (Agarwal) ↑	1.144 (0.551)	0.140 (0.047)	0.328 (0.334)	0.187 (0.013)	2.636 (0.027)	**3.020 (0.251)**	1.648 (0.045)
	Motif Corr. ↑	0.454 (0.025)	0.849 (0.026)	0.511 (0.130)	0.143 (0.024)	0.647 (0.039)	0.634 (0.084)	**0.933 (0.010)**
	3-mer Corr. ↑	0.659 (0.133)	0.940 (0.010)	0.647 (0.198)	-0.354 (0.007)	0.689 (0.022)	0.413 (0.058)	**0.976 (0.000)**
	Fraction Acc. ↑	0.451 (0.393)	0.021 (0.012)	0.380 (0.541)	0.526 (0.023)	0.971 (0.012)	**1.000 (0.000)**	0.930 (0.008)
	Diversity ↑	0.309 (0.112)	1.977 (0.001)	0.637 (0.523)	1.958 (0.003)	1.980 (0.001)	1.896 (0.021)	**1.981 (0.001)**
SK-N-SH	MinGap Eval ↑	0.556 (0.146)	-0.278 (0.006)	0.186 (0.332)	0.094 (0.046)	3.026 (0.222)	**3.720 (0.179)**	3.230 (0.022)
	MinGap (Full MPRA) ↑	0.647 (0.197)	-0.261 (0.013)	0.167 (0.263)	0.026 (0.067)	3.587 (0.262)	3.656 (0.265)	**3.698 (0.041)**
	MinGap (Agarwal) ↑	-	-	-	-	-	-	-
	Motif Corr. ↑	0.519 (0.155)	0.855 (0.026)	0.476 (0.092)	0.226 (0.017)	0.380 (0.043)	0.477 (0.037)	**0.881 (0.031)**
	3-mer Corr. ↑	0.775 (0.035)	0.949 (0.007)	0.719 (0.030)	-0.382 (0.001)	0.366 (0.019)	0.201 (0.172)	**0.969 (0.007)**
	Fraction Acc. ↑	0.049 (0.066)	0.016 (0.014)	0.039 (0.034)	0.820 (0.031)	**0.836 (0.039)**	0.638 (0.424)	0.747 (0.027)
	Diversity ↑	1.269 (0.108)	1.976 (0.002)	0.918 (0.211)	1.826 (0.001)	**1.981 (0.002)**	1.855 (0.091)	1.976 (0.002)

Sequence-to-activity Model Training Details. Both the `Design-Model` and `Evaluation-Model` were fine-tuned based on the Enformer architecture, adapted for output tasks corresponding to the target cell lines. Models were trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a global batch size of 512.

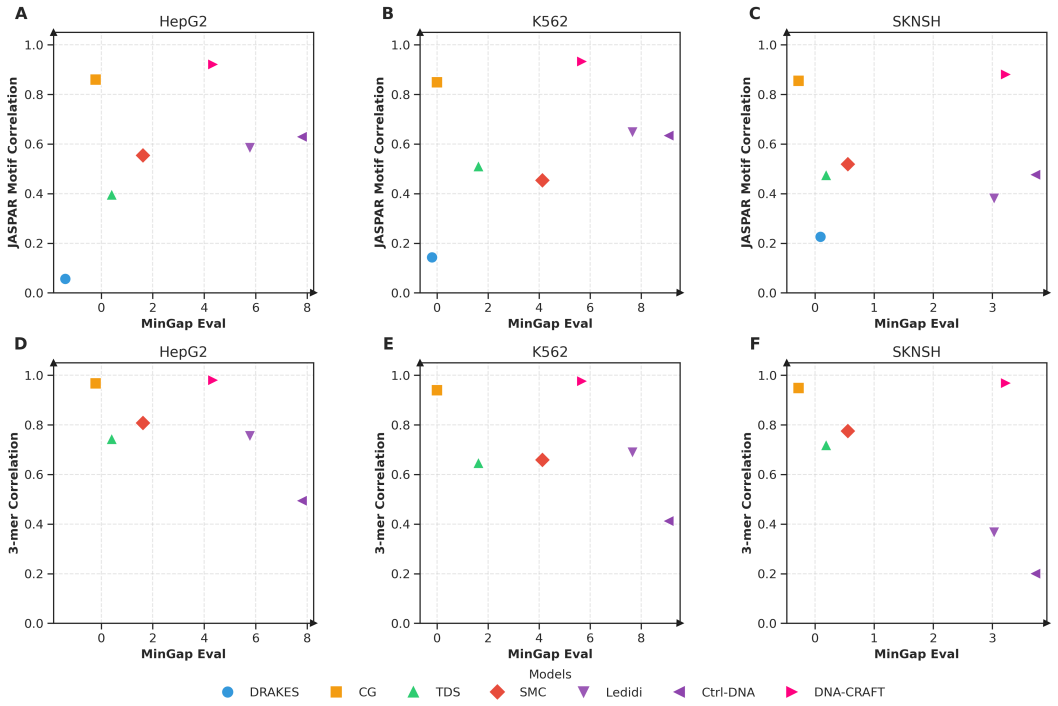

Figure 7: Trade-off between cell-type specificity and biological fidelity. Performance of DNA-CRAFT compared to baselines for HepG2, K562, and SK-N-SH cell lines. The x-axis represents the MinGap score, serving as a proxy for cell-type-specific activity. The y-axis represents biological fidelity, measured by JASPAR Motif Correlation and 3-mer Correlation relative to top specific natural enhancers.

A.5 Ablation Studies

Table 10: Ablation study on MinGap reward and class guidance. We compare variants of DNA-CRAFT with randomly generated and unguided sampled baselines. Results are generated for **SK-N-SH** cell-line-specific sequences. Values are reported as mean over sequences generated from a single run.

Method	SK-N-SH ↑	HepG2 ↓	K562 ↓	MinGap ↑	Motif Corr. ↑	3-mer Corr. ↑	Diversity ↑
Baselines							
Random	0.770	0.787	0.893	-0.122	0.325	0.230	1.981
Unguided Sampling	0.536	0.656	0.694	-0.158	0.846	0.916	**1.982**
MinGap Reward Ablation							
DNA-CRAFT-Pareto	1.058	-0.063	-0.256	1.121	0.854	0.911	1.980
Class Guidance Ablation							
DNA-CRAFT (Uncond., $\gamma = 0$)	4.885	1.685	1.095	3.200	0.877	0.965	1.974
DNA-CRAFT (Promoter, $\gamma = 3$)	**6.452**	4.047	3.993	2.405	0.447	0.092	1.920
DNA-CRAFT (Enhancer, $\gamma = 3$)	5.063	1.723	1.065	**3.340**	**0.906**	**0.975**	1.977

MinGap reward for Monte Carlo Tree Guidance. MCTG with a MinGap reward seeks to balance high activity in desired cell types with low activity in undesired cell types. To validate this, we tested random and diffusion-sampled sequences without tree guidance as a baseline. We also tested MCTG as implemented in PepTune Tang et al. (2024) using Pareto front optimization, treating the desired cell type activity as a maximization objective and undesired activities as minimization objectives (DNA-CRAFT-Pareto). Table 10 shows that standard sampling fails to consistently generate sequences with high specificity, highlighting the need for tree guidance. Replacing the MinGap set with the Pareto front yielded unspecific sequences since Pareto optimization retains sequences that excel in any one of the objectives, irrespective of their performance in other objectives.

Regulatory Sequence Class Guidance. Among all classes of regulatory elements, enhancers exhibit the highest degree of cell-type specificity, functioning as the key drivers of cell-type-specific gene expression programs Friedman et al. (2024). Hence, we tested whether conditioning the generative

process towards enhancer like sequences could achieve higher specificity. We generated sequences unconditionally ($\gamma = 0$) and conditioned on the "Enhancer" or "Promoter" classes with $\gamma = 3$, respectively. Table 10 shows that the "Enhancer" class conditioning indeed improved specificity, as well as motif correlation and diversity compared to unconditional and "Promoter"-conditional generation.

In summary, the ablation study suggests that using both the MinGap specificity score and class-conditioned sampling, as incorporated in DNA-CRAFT, improves the biological fidelity and predicted cell-type specificity of the designed sequences.