BASE COMPOSITION CONFOUNDS CLASSIFIER-FREE GUIDANCE IN DNA DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing short DNA sequences that act as enhancers selectively in a target cell type is a central goal in generative genomics. Diffusion models have recently shown promise for this task, typically using classifier-free guidance to steer generation toward the target cell type. We find that stronger classifier-free guidance can raise predicted chromatin accessibility, a proxy for regulatory activity, while shifting base composition away from that of real sequences and distorting motif frequencies. We introduce composition-neutral guidance (CNG), which subtracts the mean guidance logit update across positions for each base. This removes sequence-wide base preferences from the update without retraining. Applied to BaseDiff, our 21.6M-parameter convolutional discrete diffusion model, CNG reduces composition shifts and improves agreement with real motif frequencies and co-occurrence patterns with little change in predicted accessibility. We compare generators across guidance scales using their accessibility–quality trade-off curves. Across two datasets, BaseDiff sampled with CNG achieves larger trade-off areas than DNA-Diffusion and D3, despite using fewer than one quarter of their parameters.

1 INTRODUCTION

Designing DNA enhancers that regulate gene expression selectively in a chosen cell type is a central goal in synthetic biology (Gosai et al., 2024; Taskiran et al., 2024; de Almeida et al., 2024). Enhancers, $\sim$200–500 base pair (bp) stretches, work by providing binding sites for transcription factors (TFs), proteins that control gene activity. TFs recognize short DNA patterns called motifs. Different cell types express different combinations of TFs, so the same enhancer can be active in one cell type but inactive in another. Enhancer activity depends not only on which motifs are present but also on their combinations and arrangement, often described as regulatory grammar. Diffusion models have recently shown promise for generating sequences with cell-type-specific regulatory patterns (DaSilva et al., 2026; de Almeida et al., 2024). These sequences are often evaluated using predictors of chromatin accessibility, a measure of how readily DNA can be accessed by TFs and a proxy for regulatory activity.

To steer generation toward a target cell type, classifier-free guidance (CFG) amplifies the difference between the model's predictions with and without the cell-type label (Ho & Salimans, 2021). We find that stronger guidance also shifts base composition, the relative frequencies of A, C, G, and T, away from that of real regulatory sequences. Part of the guidance signal reflects differences in base composition between regulatory sequences from different cell types. Amplifying these sequence-wide base preferences can distort motif frequencies even as predicted accessibility increases.

Predicted accessibility alone is therefore not enough to evaluate generated sequences. Methods that directly optimize an activity predictor should not be evaluated solely by the same predictor's score. We therefore assess how well generated sequences reproduce the motif frequencies and co-occurrence patterns of real regulatory DNA, alongside predicted accessibility and cell-type specificity. Since the same guidance scale can produce different accessibility levels across models, we compare their accessibility–quality trade-off curves over a range of scales. We summarize each curve by a trade-off area that rewards both higher motif quality and higher accessibility.

To improve this trade-off, we introduce BaseDiff, a compact discrete diffusion model for cell-type-specific enhancer design. The model operates directly on DNA bases (Austin et al., 2021) and uses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

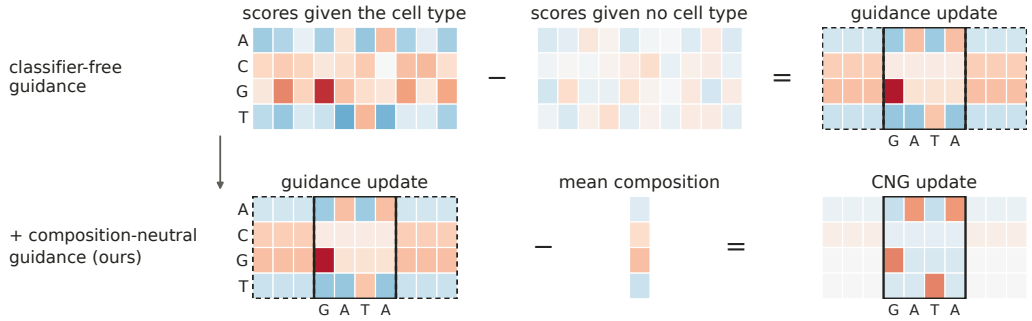

Figure 1: **CNG amplifies cell-type-specific structure while controlling for base composition.** Illustrative values; horizontal is 10 positions, vertical the four bases, red a push towards a base and blue away. A guided diffusion model scores each base at each position twice, once given the target cell type and once given none, and normally amplifies the difference between the two, the guidance update, without modification (top). Only the region in the solid outline holds cell-type structure, here a GATA site; outside it (dashed) the update shows a push towards G and C which is not cell-type-specific. CNG corrects the update instead (bottom), subtracting its mean over positions to reduce unwanted guidance and amplify the region of interest.

a convolutional backbone suited to the local structure of motifs. We propose composition-neutral guidance (CNG), which subtracts the average guidance logit update across positions for each base (Fig. 1). This reduces shifts in base composition and improves motif quality with little change in predicted accessibility. CNG changes only the sampling procedure and requires no retraining. We denote BaseDiff sampled with CNG by BaseDiff-CNG.

Contributions.

1. **A compact convolutional discrete diffusion model for cell-type-specific regulatory DNA** (§3.1). At 21.6M parameters it dominates more of the accessibility–quality plane than every method we compare against, on all three cell types and both quality measures, including alternative backbones, objectives and generative formulations.
2. **Base composition confounds guidance, and can be removed during training or during sampling** (§3.3, 5.2). Composition-neutral guidance (CNG) splits the guidance update into a position-independent composition shift and position-specific motif preferences, and suppresses the former (Fig. 1); matching the forward-process noise to each cell type's composition removes it during training. Both recover a similar share of the quality strong guidance removes. We use CNG, which modifies the sampler alone and improves co-occurrence further. Splitting a guidance update to suppress a global term is established in image diffusion (Lin et al., 2024; Chung et al., 2025; Sadat et al., 2025a;b); we apply it to base composition in discrete logits.
3. **An evaluation metric for guided regulatory-DNA generators** (§4.2, 5.1). Accessibility and quality are computed independently, so raising one does not automatically raise the other, and a model is summarized by the area its guidance curve dominates. Neither measure is sufficient alone: most of the realism range is reachable by copying short-range letter statistics from real sequences, with no grammar at all (App. D).

2 Background and Related Work

Motifs and base composition. A DNA sequence is a string over {A,C,G,T}. A TF does not recognize one word but a family of similar ones, so its preference is summarized as a *motif*: a position weight matrix (PWM) giving, over roughly 6 to 20 positions, the relative preference for each base. Scanning a sequence with a PWM and keeping matches above a significance threshold yields its motif *hits*.

Base composition, the frequency of each letter, indirectly affects which motifs a sequence contains. This is because motifs differ in composition, unrelated to how cell-type-specific they are. The GC-box, a run of Gs and Cs bound by the SP and KLF families, occurs in nearly every cell type's regulatory regions; the AT-rich GATA motif is restricted to red blood cells. Raising GC content adds hits of the first kind and removes hits of the second, so a model can appear to have learned which motifs a cell type uses while having only shifted its letter frequencies.

This is a known problem outside generation. A classifier trained to separate accessible regions from background can reach high accuracy from GC content alone (Krützfeldt et al., 2020), which is why matching those regions on length, GC and repeat content is standard (Ghandi et al., 2014), why composition-reliant models transfer badly (Cochran et al., 2022), and why accessibility models remove the assay's own sequence preference before modeling grammar (Pampari et al., 2025).

Generative design of regulatory DNA. Methods differ in how they represent a distribution over sequences, and each choice has a drawback. GANs (Zrimec et al., 2022) sample quickly but give no likelihood, so one cannot ask how typical a designed sequence is. Autoregressive models (Lal et al., 2024) condition naturally on an activity token but impose a left-to-right order with no biological counterpart: each base is fixed once generated and cannot be revised after later bases are chosen. Continuous diffusion and flow models (Avdeyev et al., 2023; Stark et al., 2024) denoise a relaxation of the one-hot sequence, so intermediate states are not valid DNA and the projection back to discrete bases is an approximation the objective does not model.

Designing against a predictor. A large family instead makes a learned activity model the objective, whether by searching the sequence space against it (Gosai et al., 2024; Taskiran et al., 2024; Reddy et al., 2024) or by folding it into training or sampling as a reward (Chen et al., 2025b; Awasthi et al., 2026; Liu & Ghods, 2026; Chen et al., 2025a). The quantity these methods report is the quantity they optimize, so it cannot tell us whether a generated sequence has the motif grammar that makes a real enhancer work. The predictor-guided optimizer we evaluate reaches comparable accessibility with much lower motif quality.

Discrete diffusion (Austin et al., 2021) corrupts and denoises nucleotides directly. Every intermediate state is a valid sequence, and all positions are refined in parallel, so a base chosen early can still be revised later. We build on discrete diffusion with a convolutional denoiser; Kamb & Ganguli (2025) give some insight into how locality and equivariance in convolutional diffusion drive compositional generalization, which matches the problem of enhancer generation: a collection of short motifs that may appear anywhere in the sequence.

Corrections to classifier-free guidance. CFG (Ho & Salimans, 2021) amplifies the conditional signal by extrapolating from an unconditional prediction, and at high scales the samples degrade. For discrete models the same rule is derived directly in logit space (Nisonoff et al., 2025; Schiff et al., 2025), which is the form we use. The corrections come from image and video generation, where the fixes split the update and suppress one part: rescaling output magnitude (Lin et al., 2024), treating high-guidance failure as leaving the data manifold (Chung et al., 2025), down-weighting the component parallel to the conditional prediction (Sadat et al., 2025a), rescaling the unconditional prediction before the extrapolation (Fan et al., 2025) and guiding the low-frequency band, which contains global structure and color, more conservatively than the rest (Sadat et al., 2025b). A correction that modifies the CFG rule itself needs only the conditional and unconditional predictions; two of these approaches can therefore be applied to DNA as well. We explore this in App. L. In discrete spaces, Dou et al. (2026) obtain guidance from an external reward model where the reward cannot be differentiated with respect to a symbol. To the best of our knowledge, correcting the CFG update has not been explored for DNA design. CNG splits the CFG update as the image methods do; the component it removes is constant across positions, which is base composition.

3 METHOD

BaseDiff generates fixed-length DNA sequences over the alphabet {A,C,G,T}, conditioned on a target cell type c. It is a D3PM (Austin et al., 2021): a forward process corrupts a real sequence and a denoiser learns to undo it, with guidance applied at every reverse step (Fig. 5, App. A.1).

3.1 CONVOLUTIONAL BACKBONE

The denoiser is convolutional. Every convolution spans 7 bp and a dilation cycle lets successive blocks combine those local features across the window; a self-attention block follows every third layer. The network is 21.6M parameters, and the cell label is replaced by a null label with probability 0.1 during training, so one network supplies both predictions guidance requires.

Replacing the convolutions with global attention (DiT-style (Peebles & Xie, 2023)), at comparable and at larger parameter counts, leads to worse performance, while within the convolutional family the width and the attention density make little difference. Layer, optimizer and sampler settings, and these ablations, are in App. A.1 and App. N.

3.2 DIFFUSION PROCESS

Write x_0 for a real sequence of length L over the $K = 4$ bases, each position one-hot, and x_t for its corrupted version at step $t \in \{1, \dots, T\}$ with $T = 128$. BaseDiff uses the *uniform* substitution process, whose per-position transition matrix and cumulative product are

$$\begin{aligned} q(x_t \mid x_{t-1}) &= \mathrm{Cat}\big(x_t;\, x_{t-1}Q_t\big), & Q_t &= (1-\beta_t)I + \tfrac{\beta_t}{K}\mathbf{1}\mathbf{1}^\top, \\ q(x_t \mid x_0) &= \mathrm{Cat}\big(x_t;\, x_0\bar{Q}_t\big), & \bar{Q}_t &= \textstyle\prod_{s\le t} Q_s = (1-\bar{\beta}_t)I + \tfrac{\bar{\beta}_t}{K}\mathbf{1}\mathbf{1}^\top, \end{aligned} \tag{1}$$

where Q_t and $\bar{Q}_t$ are $K \times K$, one row and column per base, $Q_t[i, j]$ is the probability base i becomes base j in one step, β_t and $\bar{\beta}_t$ are the per-step and cumulative corruption rates and the product runs over $s = 1, \dots, t$. We take $\bar{\beta}_t = t/T$, so a position keeps its base with probability $1 - \bar{\beta}_t$ and is otherwise redrawn uniformly from $\{\mathrm{A,C,G,T}\}$. At $t = T$ it is uniform noise.

The denoiser has weights θ and predicts the clean sequence from the corrupted one, the step and the cell label, emitting per-position logits ℓ whose softmax is $p_\theta(x_0 \mid x_t, t, c)$. It is trained by cross-entropy against x_0 on the corrupted positions. Sampling starts from x_T uniform and repeats the posterior step

$$p_\theta(x_{t-1} \mid x_t) \;\propto\; \big(x_t Q_t^\top\big) \odot \big(p_\theta(x_0 \mid x_t, t, c)\, \bar{Q}_{t-1}\big), \tag{2}$$

with $\odot$ the elementwise product, drawing x_{t-1} from it. Guidance modifies ℓ before the softmax at every step.

Real enhancers are not 25% in each base, so the noise could instead be drawn from the training data's pooled base frequencies, or from the target cell type's own base composition. We trained three backbones differing only in that choice. Sampled with the correction of §3.3 they are indistinguishable according to our evaluation, so we use uniform; sampled without it, matching the noise to each cell type's composition removes much of the same shift the correction removes (App. M).

3.3 COMPOSITION-NEUTRAL GUIDANCE

CFG runs the denoiser twice, once with the cell label and once with the null label, giving conditional and unconditional logits ℓ_c and $\ell_\varnothing$, and pushes away from the unconditional ones. For a discrete model this modifies the logits directly, $\ell^{\mathrm{guided}} = \ell_c + w\,(\ell_c - \ell_\varnothing)$, with guidance scale w. The same w places different models at different accessibility levels, which is why single-scale comparisons are confounded (§5.1).

Composition shift. CFG amplifies the direction in which a cell type differs from the average one, $\ell_c - \ell_\varnothing$. Part of that difference is the cell type's TF grammar, which is what we want amplified; part is base composition. If a cell type's regulatory regions are GC-rich relative to average, a large component of $\ell_c - \ell_\varnothing$ is a preference for G and C across all positions, and amplifying it raises GC content throughout the whole sequence. This inadvertently adds hits for GC-rich families and removes them for AT-rich ones, causing realism over the whole motif profile to deteriorate. App. J reports the direction and size of the shift for each of the three benchmark cell types.

The correction. Index the L positions of a sequence by p and the bases by $b \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, so the guidance direction $d = \ell_c - \ell_\varnothing \in \mathbb{R}^{L \times K}$ has one logit gap per position and base, and let $\mathcal{M}$ be the positions updated at the current reverse step (all L of them for BaseDiff). We split d into

a position-independent composition shift and position-specific motif preferences, and subtract the former:

$$\mathrm{comp}_b = \frac{1}{|\mathcal{M}|}\sum_{p\in\mathcal{M}} d_{p,b}, \qquad \tilde{d}_{p,b} = d_{p,b} - \mathrm{comp}_b, \qquad \ell^{\mathrm{guided}}_{p,b} = \ell_{c,p,b} + w\,\tilde{d}_{p,b}. \tag{3}$$

The same four numbers are subtracted at every position (Fig. 1), so the correction changes only composition and preserves how the update varies across positions: every base's average push across the sequence is zero. The correction applies to any cell type, adds almost no sampling cost, does nothing at $w = 0$ and modifies the guidance update alone. We write BaseDiff sampled this way as BaseDiff-CNG. App. K compares the mean against two alternatives, one applied after the softmax.

4 Experimental Setup

4.1 Data

We use the cell-type-specific DNase I hypersensitive site (DHS) regions of Meuleman et al. (2020) as assembled by DaSilva et al. (2026): 47,872 sequences of $L = 200$ bp, 11,968 for each of four cell groups. We evaluate on GM12878 (B-lymphocyte), HepG2 (hepatocyte) and K562 (erythroid), among the most heavily assayed human cell lines in ENCODE, a public catalogue of functional genomics experiments (The ENCODE Project Consortium, 2012). The fourth group, hESCT0 (embryonic stem cells), is present during training but not evaluated. The split is by chromosome and fixed: chromosome 1 is the test set, chromosome 2 validation and the remaining 22 training, giving 39,254 training, 4,053 validation and 4,565 test sequences, with per-cell test sets of 1,044, 958 and 1,321. These held-out sets are the real reference for every metric. Every method is run at the same guidance scales, $w \in \{0, 1, 2, 4, 6\}$, generating exactly 5,000 sequences per cell at each scale.

4.2 Metrics

We follow the evaluation of DaSilva et al. (2026), adding motif co-occurrence. *Realism* asks whether each motif appears as often as it does in real sequences. *Motif co-occurrence* asks whether pairs of motifs appear together as often as they do in real sequences, which realism does not capture. *Accessibility* asks how open a generated sequence is predicted to be in its target cell type and how much of that opening is specific to it.

Motif scanning. Both quality measures, realism and motif co-occurrence, are computed from which motifs a sequence contains. We search every sequence for each of the 1,019 vertebrate motifs of JASPAR (Rauluseviciute et al., 2024), scoring each position against that motif's PWM (§2) with MOODS (Korhonen et al., 2009) and keeping matches above a stringent significance threshold; App. A.3 gives the settings. For each sequence i and motif j we record presence (at least one occurrence) or absence as $\mathrm{hit}[i,j] \in \{0,1\}$. The *motif profile* of a set of N sequences is $f_j = \frac{1}{N}\sum_i \mathrm{hit}[i,j]$, the fraction containing motif j. This forms a vector of length 1,019.

Realism. *Realism* is $1 - \mathrm{JS}$, where JS is the Jensen–Shannon distance between a generated set's motif profile f_{gen} and the real test profile f_{real}, both renormalized to sum to 1:

$$\mathrm{JS}(f_{\mathrm{gen}}, f_{\mathrm{real}}) = \sqrt{\tfrac{1}{2}\,\mathrm{KL}(f_{\mathrm{gen}}\|m) + \tfrac{1}{2}\,\mathrm{KL}(f_{\mathrm{real}}\|m)}, \quad m = \tfrac{1}{2}(f_{\mathrm{gen}} + f_{\mathrm{real}}), \tag{4}$$

with KL in bits. It is bounded in $[0, 1]$ and higher is better. Realism is not informative on its own: a shuffle of real sequences preserving every adjacent base pair count reaches 0.794 and an order-5 Markov chain reaches 0.897 (App. D).

Motif co-occurrence. Realism compares motif frequencies one at a time, and a model could match all 1,019 while placing motifs independently, which would not form a proper grammar. For a pair (i, j) we use the lift $P(i \wedge j)/(P(i)P(j))$, which equals 1 under independence, take the mean absolute error in log lift against held-out real sequences, and report $1 - \mathrm{MAE}$ so that, like realism, higher is better.

Pairs come from real data: we keep motifs present in 1–50% of real sequences, drop pairs whose matrices are near-identical, which co-occur because one genomic site matches both, and score the 500 remaining pairs with the highest co-occurrence. App. B gives the filters and a worked example.

Accessibility. We predict chromatin accessibility with Enformer (Avsec et al., 2021), trained to predict experimental measurements along the genome from DNA sequences. Enformer reads 196,608 bp of context, so a sequence cannot be scored in isolation; it must be placed somewhere in the genome. Each sequence is spliced into the center of a fixed window at the cell-type-neutral *TNFSF18* locus (hg38 chr1), replacing the endogenous 200 bp (App. F).

Enformer emits 5,313 tracks over 896 bins of 128 bp; we use the DNase-seq tracks, the assay that measures accessibility, for the three benchmark cell types and average the two bins the centered 200 bp sequence spans. Each sequence i therefore has three raw values, one per cell type, and we write $r_{i,c}$ for the value in cell c. For a set designed for cell c,

$$\text{signal}_{i,c} = \widehat{F}_c(r_{i,c}), \qquad \text{specificity}_{i,c} = \frac{r_{i,c}}{\sum_{c'} r_{i,c'}}, \tag{5}$$

$$\text{accessibility}_{i,c} = \frac{\text{signal}_{i,c} \cdot \text{specificity}_{i,c}}{\max_{\text{real}} \left(\text{signal}_{\cdot,c} \cdot \text{specificity}_{\cdot,c}\right)}, \tag{6}$$

where $\widehat{F}_c$ is the empirical CDF of $r_{\cdot,c}$ over real training enhancers of cell c. signal is how strong the predicted accessibility is, as a percentile within real regulatory DNA. specificity is how much of sequence i's predicted signal falls in the intended cell, which separates a sequence that is open everywhere from one that is cell-type-targeted. A designed enhancer needs both, so accessibility multiplies them. Dividing by the largest product any real training enhancer reaches puts 1 at the most extreme real enhancer, so a score states where a sequence falls relative to real regulatory DNA and can exceed 1. In the experiments below, we report the median of $\text{accessibility}_{i,c}$ over the 5,000 sequences of a set. The ENCODE accessions of the three Enformer tracks are in App. A.3. To check that the comparison does not depend on one predictor, we compute every result a second time with ChromBPNet (Pampari et al., 2025), trained separately for each cell type.

4.3 Baselines

We compare against DNA-Diffusion (DaSilva et al., 2026), Dirichlet flow matching (Stark et al., 2024) and D3 (Sarkar et al., 2026) (App. A.2), as well as alternative backbones, objectives and generative formulations. Three training-free generators have no guidance parameter, and we report them separately (App. D). We also include one method that optimizes a learned activity model. DaSilva et al. (2026) released a Malinois predictor trained on these DHS splits together with the CODA (Gosai et al., 2024) design code, and we take all four refinement rounds their script performs. Each round optimizes against the predictor, finds the motifs it over-produced and penalizes them in the next. Its control parameter is a count of rounds, so its four points can be drawn in the same plane but do not correspond to guidance scales; we summarize it by where those points fall on the two quality measures. Five further reward-optimizing methods are outside the scope of these experiments (App. A.2). We report our model both with and without CNG; every other model is sampled with ordinary classifier-free guidance.

5 Results

5.1 The accessibility–quality trade-off

Raising the guidance scale traces a curve for each model, gaining accessibility and losing motif quality (Fig. 2). Every model can be guided past the predicted accessibility of real enhancers, which is often the goal. We evaluate how much motif structure each sacrifices during the process.

At a fixed guidance scale, a model that responds weakly scores well on quality by staying near its unguided samples. At $w = 1$ the 22.7M transformer of App. N has the highest GM12878 realism of any model we run at an accessibility of 0.418, below real enhancers at 0.611 and half the 0.831 BaseDiff-CNG reaches at the same scale. We therefore summarize each curve by the area of the

Table 1: **Trade-off area on both quality measures.** Computed from the curves of Fig. 2; higher is better on both, best in each column bold and second best underlined. Values are ± a 95% bootstrap in which each replicate resamples all 5,000 sequences at every guidance scale and rebuilds the curve. Every area is measured from the same corner, at zero accessibility and a quality floor of 0.5, below every point of every curve in this table (lowest realism 0.53, lowest co-occurrence 0.65). CODA has no guidance scale, so we omit it here. The last two rows are the same trained network sampled with and without CNG.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
DNA-DIFFUSION (94.6M)	0.357 ± 0.008	0.373 ± 0.008	0.348 ± 0.006	0.389 ± 0.003	0.386 ± 0.006	0.373 ± 0.006
D3 transformer (92.4M)	0.368 ± 0.007	0.373 ± 0.008	0.361 ± 0.007	0.388 ± 0.004	$\underline{0.389 \pm 0.005}$	$\underline{0.374 \pm 0.006}$
Dirichlet flow matching (3.7M)	0.301 ± 0.009	0.265 ± 0.010	0.260 ± 0.007	0.359 ± 0.003	0.310 ± 0.004	0.285 ± 0.005
BaseDiff (21.6M)	$\underline{0.381 \pm 0.007}$	$\underline{0.390 \pm 0.008}$	$\underline{0.367 \pm 0.007}$	$\underline{0.402 \pm 0.003}$	$\mathbf{0.398 \pm 0.004}$	0.364 ± 0.005
BaseDiff-CNG (21.6M)	$\mathbf{0.388 \pm 0.006}$	$\mathbf{0.396 \pm 0.008}$	$\mathbf{0.392 \pm 0.006}$	$\mathbf{0.405 \pm 0.004}$	$\mathbf{0.398 \pm 0.004}$	$\mathbf{0.423 \pm 0.005}$

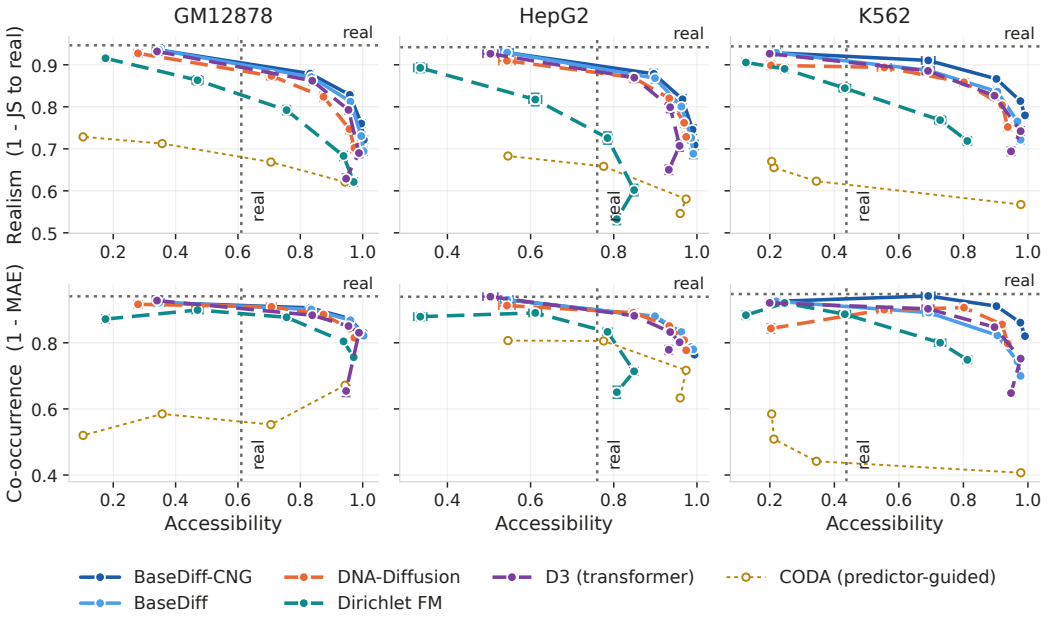

Figure 2: **Accessibility against motif realism and motif co-occurrence.** Horizontal: accessibility from Enformer, an external model independent of both quality measures, as the median over a set of 5,000 sequences. *Top:* realism. *Bottom:* co-occurrence; higher is better in both. Each curve has five points, the guidance scales $w \in \{0, 1, 2, 4, 6\}$ from left to right, with 95% bootstrap intervals in both directions. CODA has no guidance scale, so its four points are its refinement rounds, running right to left. Dotted lines mark real held-out enhancers, scored exactly as a generated set is.

plane it dominates (Table 1), the hypervolume of a Pareto front (Zitzler & Thiele, 1999), which a model enlarges by having more quality at a given accessibility or reaching higher accessibility.

BaseDiff-CNG has the larger area on both measures and all three cell types; we report the difference between two areas ± 1.96 standard deviations of that difference across bootstrap replicates. Over the next best of the other methods on each cell type, the realism margins are $+0.020 \pm 0.005$ ($+5.5\%$), $+0.023 \pm 0.006$ ($+6.1\%$) and $+0.031 \pm 0.006$ ($+8.6\%$), and the co-occurrence margins $+0.016 \pm 0.005$ ($+4.1\%$), $+0.009 \pm 0.006$ ($+2.4\%$) and $+0.050 \pm 0.008$ ($+13.3\%$). BaseDiff is also ahead of alternative backbones, objectives and generative formulations on both measures and all three cell types except HepG2 co-occurrence (App. N). The ordering is mostly unchanged when the area is restricted to the accessibility range every model covers (App. O), when ChromBPNet replaces Enformer (App. E), when every sequence is spliced into a second genomic locus (App. F) and on a second dataset, a lentiMPRA reporter library (App. G).

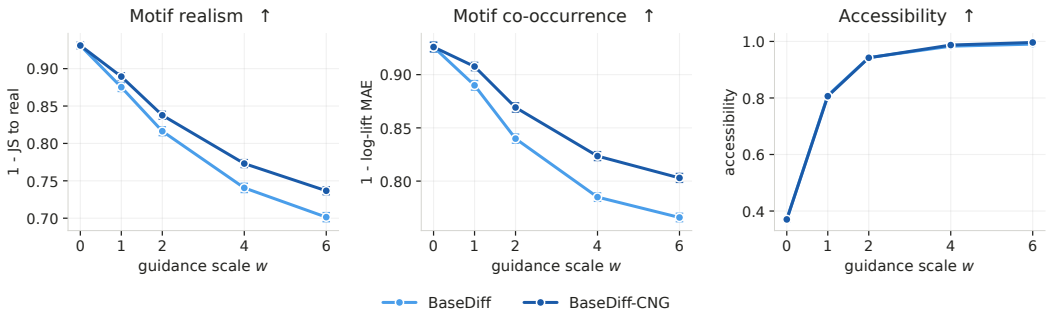

Figure 3: **Motif realism, co-occurrence and accessibility against the guidance scale.** One trained network sampled two ways, so the curves can be compared vertically at equal guidance. Each point is one set of 5,000 sequences, accessibility is the median over that set and curves are means over cells, with 95% bootstrap intervals over sequences.

Since raising w amplifies the conditional signal, one might expect higher levels to raise accessibility. Not every method behaves that way: Dirichlet flow matching loses accessibility between $w = 4$ and $w = 6$ on HepG2 alongside lower quality on both measures, so its curve folds back in Fig. 2. D3 does the same on all three cell types at the highest level. The area counts only the best value a curve reaches, which is generous to those models. Over the range we evaluate, BaseDiff-CNG, BaseDiff and DNA-DIFFUSION keep rising in accessibility, making the effect of w predictable. Generated sequences also resemble the training split and one another more closely as w rises, with and without CNG (App. I). We do not go beyond $w = 6$, where accessibility already reaches the most extreme real enhancers.

5.2 THE EFFECT OF CNG

Raising the guidance scale pulls base composition away from its real level in every guided model. By $w = 6$ BaseDiff has moved GC content by up to 7.9 percentage points away from the held-out value, DNA-DIFFUSION by up to 8.4, D3 by up to 11.6 and Dirichlet flow matching by up to 17.8. With CNG no cell moves more than 1.8 percentage points at any scale (App. J).

Averaged over cells, the correction raises motif realism at every scale (Fig. 3), by $+0.014$ ($+1.6\%$) at $w = 1$ up to $+0.035$ ($+5.0\%$) at $w = 6$, and co-occurrence by $+0.018$ ($+2.0\%$) up to $+0.038$ ($+4.9\%$), recovering roughly a third of the pairwise structure strong guidance destroys against an unguided 0.926. Accessibility is nearly unchanged. The gain is largest on K562, by $+0.059$ ($+16.3\%$) on co-occurrence and $+0.025$ ($+6.9\%$) on realism in Table 1. This makes sense, since K562 is also the cell where base composition accounts for more of the per-motif error than in the other two.

The shift can also be removed during training, by drawing the forward process noise from each cell's own base composition instead of uniformly. Under plain CFG that model performs about the same as BaseDiff-CNG on both quality measures, the exception being K562 co-occurrence, where it is lower by 0.027 (7.0%) (App. M). We use CNG for two reasons. It modifies the guidance update alone, so it can be applied to other guided discrete sequence models without retraining. It also improves co-occurrence further: on top of the composition-matched model it adds $+0.024$ ($+6.1\%$) on K562.

5.3 ON-TARGET AND OFF-TARGET ACCESSIBILITY

A designed enhancer has to be active in its target cell type and inactive elsewhere. Enformer scores all three cell types from one inserted sequence, so we take two numbers per set: the median prediction in the cell type its sequences were designed for and the median in the other two (Fig. 4).

Guidance raises the first in every model and leaves the second relatively unchanged, apart from a dip between $w = 4$ and $w = 6$ in Dirichlet flow matching on HepG2 and in D3 on all three cell types, as seen in Fig. 2. BaseDiff-CNG is the highest curve at every guidance scale and cell type, the one exception being unguided HepG2, where DNA-DIFFUSION matches it. By $w = 6$ its prediction

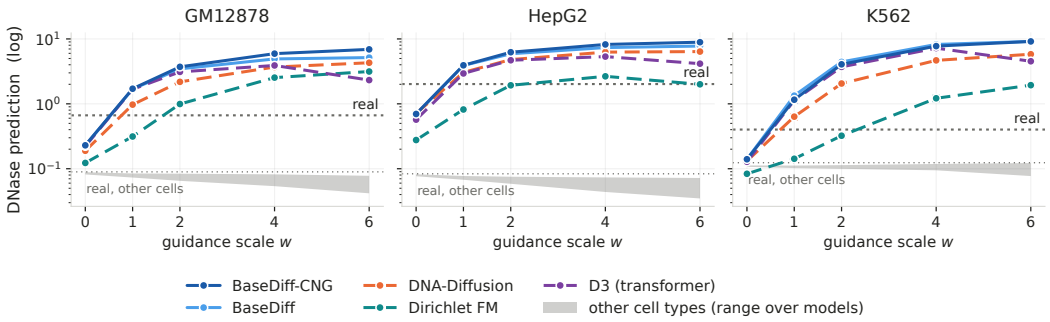

Figure 4: **On-target and off-target DNase predictions against the guidance scale.** Raw Enformer DNase prediction on a log scale; App. H gives the same predictions per model and scale. Each solid curve is the median over a set of 5,000 sequences, scored in the cell type they were designed for, with a 95% bootstrap interval. The shaded band is those same sequences scored in the other two cell types, running from the lowest such median to the highest across models. The upper dotted line is real training enhancers in that cell type and the lower one is the same enhancers in the other two.

reaches 6.9, 8.9 and 9.2 against 4.3, 6.4 and 5.8 for the next best of the other methods, while its off-target prediction falls from 0.08–0.11 to 0.07–0.10.

6 Discussion

A compact convolutional discrete diffusion model, sampled with composition-neutral guidance, dominates more of the accessibility–quality plane than every other method we compare against, on all three cell types and both quality measures, at 21.6M parameters and against models up to four times that size. Raising the guidance scale moves GC content by up to 7.9 percentage points away from its real level in our model and further in every other model we evaluate. Removing that component retains more realism and co-occurrence at the same predicted accessibility. It can be removed at sampling time, by CNG, or during training, by drawing the forward process noise from each cell type's own base composition; both recover most of the same quality, and we prefer CNG because it modifies the sampler alone. Predicted accessibility keeps rising while motif content degrades, so we summarize each model over its whole guidance curve.

Despite the advancements we have demonstrated here, several limitations remain. First, we test one form of conditioning, a categorical cell-type label, on two datasets: a chromatin accessibility atlas and a reporter library (App. G). Reporter data also records a continuous activity for each sequence (Gosai et al., 2024), which we do not condition on, and whether the correction helps there is untested. Second, the conclusions depend on how we measure quality: realism and co-occurrence both come from one JASPAR scan, so a different motif collection would move both. We vary the accessibility predictor and the genomic locus it scores in, and the ordering holds under both. Most importantly, these sequences have not been tested in cells: every accessibility value is a prediction and every quality value describes motif content, so confirming that they work needs a reporter assay (Muerdter et al., 2015; Llimos et al., 2022). Such an experiment would also directly test whether sequences matching the motif content of real regulatory DNA are the ones more likely to function.

Reproducibility Statement

We release the code for BaseDiff at [redacted]. App. A gives the architecture, the training schedule and the sampling settings for BaseDiff and for every variant we train, and App. A.3 the settings for each measurement, including the motif scan and the ENCODE accessions of the tracks, with the genomic window each accessibility predictor reads in App. F and the source, subsample and retraining of the second dataset in App. G. Where a choice has an established convention we follow it: the data, splits and quality metrics are those of DaSilva et al. (2026), and each baseline is run from its authors' released code with default hyperparameters where possible (App. A.2).

AI USE STATEMENT

We used generative AI tools to write and edit code during development, and to write the analysis and plotting scripts. We also used them to help debug the released code of other methods that did not initially run on this dataset without modification, such as the D3 source changes documented in App. A.2. We have not used generative AI tools to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to propose or refine hypotheses, to design the models or the evaluation protocol, or to clean or reformat the datasets. Additionally, we used generative AI tools to draft and edit the text, to search for background information and to improve readability.

We have reviewed all AI-assisted work. No result here was produced by a generative AI tool: all results come from running each method on real data, and every AI-written or AI-edited script was read and run by the authors, whose reading of the results is their own. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Vikram Agarwal et al. Massively parallel characterization of transcriptional regulatory elements. *Nature*, 2025.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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 (ICML)*, 2023.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18:1196–1203, 2021.

Animesh Awasthi, Raphael Bednarsky, Moritz Schaefer, and Christoph Bock. Conditional Monte Carlo tree diffusion for designing cell-type-specific and biologically faithful regulatory DNA, 2026.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design. *arXiv preprint arXiv:2505.07086*, 2025a.

Xingyu Chen, Shihao Ma, Runsheng Lin, Jiecong Lin, and Bo Wang. Ctrl-DNA: Controllable cell-type-specific regulatory DNA design via constrained RL, 2025b.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Kelly Cochran, Divyanshi Srivastava, Avanti Shrikumar, Akshay Balsubramani, Ross C. Hardison, Anshul Kundaje, and Shaun Mahony. Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. *Genome Research*, 32(3):512–523, 2022. doi: 10.1101/gr.275394.121.

Lucas Ferreira DaSilva, Simon Senan, Judith F. Kribelbauer-Swietek, et al. Designing synthetic regulatory elements using the generative AI framework DNA-Diffusion. *Nature Genetics*, 58: 180–194, 2026. doi: 10.1038/s41588-025-02441-6.

Bernardo P. de Almeida et al. Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. *Nature*, 626:207–211, 2024.

Hongkun Dou, Zike Chen, Fengji Li, Hongjue Li, and Yue Deng. Plug-and-play guidance for discrete diffusion models via gradient-informed logit correction. In *International Conference on Machine Learning (ICML)*, 2026.

Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. CFG-Zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Computational Biology*, 10 (7):e1003711, 2014. doi: 10.1371/journal.pcbi.1003711.

Sager J. Gosai et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, 634:1211–1220, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *International Conference on Machine Learning (ICML)*, 2025.

W. James Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002.

Janne Korhonen, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. MOODS: Fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, 25(23):3181–3182, 2009.

Louisa-Marie Krützfeldt, Max Schubach, and Martin Kircher. The impact of different negative training data on regulatory sequence predictions. *PLOS ONE*, 15(12):e0237412, 2020. doi: 10.1371/journal.pone.0237412.

Avantika Lal, David Garfield, Tommaso Biancalani, and Gokcen Eraslan. Designing realistic regulatory DNA with autoregressive language models. *Genome Research*, 34:1411–1420, 2024. doi: 10.1101/gr.279142.124.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Jonathan Liu and Kia Ghods. Continuous diffusion transformers for designing synthetic regulatory elements, 2026.

G. Llimos et al. A leukemia-protective germline variant mediates chromatin module formation via transcription factor nucleation. *Nature Communications*, 13:2042, 2022. doi: 10.1038/s41467-022-29625-6.

Wouter Meuleman et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, 584:244–251, 2020.

Felix Muerdter, Łukasz M. Boryń, and Cosmas D. Arnold. STARR-seq—principles and applications. *Genomics*, 106(3):145–150, 2015. doi: 10.1016/j.ygeno.2015.06.001.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. In *International Conference on Learning Representations (ICLR)*, 2025.

Anusri Pampari et al. Chrombpnet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv*, 2025. doi: 10.1101/2024.12.25.630221.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Ieva Rauluseviciute et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182, 2024.

Aniketh Janardhan Reddy et al. Designing cell-type-specific promoter sequences using conservative model-based optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 37, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M. Weber. Guidance in the frequency domain enables high-fidelity sampling at low CFG scales. *arXiv preprint arXiv:2506.19713*, 2025b.

Anirban Sarkar, Alejandra Duran, Yiyang Yu, Da-Wei Lin, Yijie Kang, Nirali Somia, Pablo Mantilla, Jessica Zhou, Masayuki Nagai, Ziqi Tang, Kaarina Hanington, Kenneth Chang, and Peter K. Koo. Designing DNA with tunable regulatory activity using discrete diffusion. *bioRxiv*, 2026. doi: 10.1101/2024.05.23.595630.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. In *International Conference on Machine Learning (ICML)*, 2024.

Ibrahim I. Taskiran et al. Cell-type-directed design of synthetic enhancers. *Nature*, 626:212–220, 2024.

The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

Eckart Zitzler and Lothar Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

Jan Zrimec et al. Controlling gene expression with deep generative design of regulatory DNA. *Nature Communications*, 13:5099, 2022.

A HYPERPARAMETERS AND SAMPLING

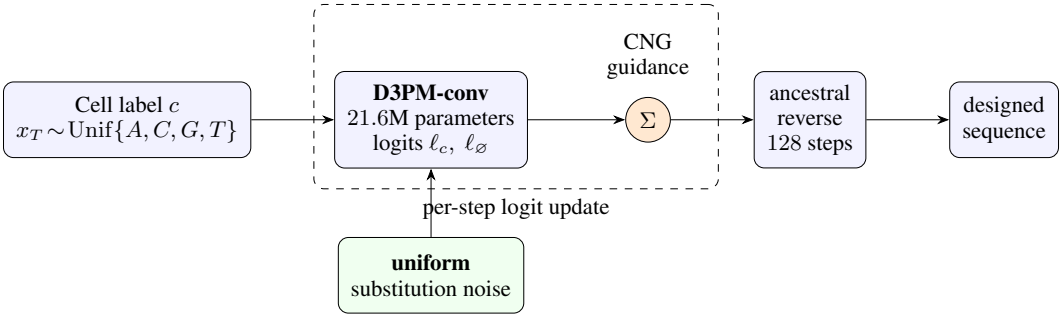

Figure 5: **BaseDiff-CNG pipeline.** A convolutional D3PM backbone is trained under a uniform substitution forward process. At sampling time the reverse process runs ancestrally from a uniform x_T and classifier-free guidance is applied with the composition shift removed (CNG).

A.1 BASEDIFF AND RELATED VARIANTS

The settings below are shared by BaseDiff, the forward-process variants of App. M and the BaseDiff variants of App. N. Only the component under test differs.

Architecture. Bases are embedded into $d = 320$ channels and added to a learned positional embedding. The conditioning vector $c \in \mathbb{R}^{320}$ is the sum of a sinusoidal timestep embedding passed through a Linear–SiLU–Linear MLP and a learned cell embedding whose table has one row per cell type plus a null row for the unconditional pass. The network is 12 residual blocks. Each block applies GroupNorm (8 groups), GELU, a 1D convolution of width 7, FiLM modulation (Perez et al., 2018) by c (zero-initialized, so a block begins unconditioned and learns to modulate), then GroupNorm, GELU and a second width-7 convolution, and adds the result to its input. Block i uses dilation $[1, 2, 4, 8, 16]_{\,i \bmod 5}$. After every third block a self-attention block is inserted (pre-LayerNorm, 8 heads, dense over all positions), giving four attention blocks over the 12. A final GroupNorm and a width-1 convolution produce the logits, giving 21.6M parameters. Attention is not necessary: with none at all, and with one after every block, the trade-off area of Table 1 moves by less than the run-to-run spread of a single configuration, and attention in every block adds 3.3M parameters and results in a worse validation loss. Convolution width 3 matches width 7 on every cell type and both measures at half the parameters, and width 15 is behind everywhere at 41.3M.

Training. AdamW, learning rate 3×10^{-4}, weight decay 10^{-5}, batch size 256, 500 warmup steps followed by cosine decay, gradient-norm clipping at 1.0, and an exponential moving average of the weights with decay 0.9999 used for all validation and sampling. The cell label is replaced by the null row with probability 0.1 during training. All runs are 1000 epochs, and the reported model is the checkpoint with the best validation loss.

Forward process. $T = 128$ for training and for sampling, and positions are corrupted independently. Training draws $t \sim \mathrm{Uniform}\{1, \ldots, T\}$ and normalizes the cross-entropy of §3.2 per sequence.

Sampling. Every model is sampled over the same guidance grid at 5,000 sequences per cell. BaseDiff-CNG adds the correction of §3.3; every variant is sampled with plain classifier-free guidance, and at $w = 0$ the two samplers are the same.

Autoregressive regularization. The autoregressive model overfits badly under the settings above, reaching training loss 0.80 against validation 1.43, and the 1000-epoch and 150-epoch cosine schedules reach the same best validation loss. Dropout (0.1) and weight decay (0.1), with nothing else changed, moved it from 1.337 to 1.296. We did not search over those values, and we did not tune BaseDiff against the same grid.

A.2 BASELINE SETTINGS

Each was run from its authors' released code, with their weights where they released them, at the same guidance scales and with 5,000 sequences per cell. Below, we list the changes we made to run the methods on our dataset.

D3 (Sarkar et al., 2026): the transformer variant, which the authors report as the stronger of their two, trained with their released configuration, 300 epochs of AdamW at 3×10^{-4} with 2500 warmup steps, EMA decay 0.9999 and their 0.1 class-dropout rate. D3 needed the most setup: all six of their datasets condition through a continuous signal path, with no categorical class and no classifier-free guidance, and their transformer includes the guidance code but never calls it, with the `label_embed` call commented out in the forward pass. We un-commented it, set `use_label_embed` and re-initialized that embedding table at std 0.02.

Dirichlet flow matching (Stark et al., 2024): their hyperparameters, including the 0.3 conditioning-dropout rate, and their convolutional backbone, which the authors report as the stronger of the two. Training runs 1000 epochs ($\approx$ 154k steps), matching the optimizer-step budget our own model receives. Checkpoint selection is their default, best validation perplexity, evaluated here on the full validation split. Their guidance scale g and ours are related by $g = 1 + w$, since both of their classifier-free paths interpolate $(1 - g)$ uncond $+ g$ cond, so $g = 1$ is the pure conditional model.

CODA (Gosai et al., 2024): the Malinois predictor trained on these DHS splits and the design code, both released by DaSilva et al. (2026), with their energy, optimizer and per-cell penalization strengths (0.3, 0.3, 0.25), taking all four refinement rounds their script performs.

Reward-optimizing methods. Awasthi et al. (2026); Liu & Ghods (2026) have no public implementation, and the reward in Liu & Ghods (2026) is the genomic predictor our accessibility measure also uses. Chen et al. (2025b;a) release code, but both guide a generator with reward models built for their own tasks, so running either here means training new ones and choosing an objective for this dataset. Taskiran et al. (2024) was adapted to these DHS splits by DaSilva et al. (2026) and would need no such setup, but it designs a sequence by testing every single-base change at every position over many rounds, which takes several GPU-days at the sample size used here, and CODA already covers this family.

A.3 MEASUREMENT SETTINGS

Enformer and ChromBPNet tracks. Accessibility uses one DNase track per cell type, chosen by ENCODE accession: ENCSR000EMT for GM12878 (track 69), ENCSR149XIL for HepG2 (track 234) and ENCSR000EOT for K562 (track 121). These are the most recent released DNase experiment for each line, and the three ChromBPNet also has models for, so both predictors use the same assays. App. F gives the genomic window each is scored in.

Motif scanning. The scan of §4.2 converts each JASPAR matrix to a natural-log-odds PWM against a uniform background (0.25 per base) with a pseudocount of 10^{-4}, and scans with MOODS (Korhonen et al., 2009) at $p < 10^{-4}$, taking for each matrix the lower of its forward and reverse-complement thresholds so both strands are scanned at the same stringency. A sequence records a hit if it contains at least one match on either strand.

B MOTIF CO-OCCURRENCE

For a motif pair (A, B) over a set of sequences, let $P(A)$ and $P(B)$ be the fractions of sequences containing each motif and $P(A \wedge B)$ the fraction containing both. The *lift* is $P(A \wedge B)/(P(A)P(B))$, which equals 1 under independence whatever $P(A)$ and $P(B)$ are. We use $\log$ lift, so that 0 means independent. Raw pair frequency mostly reflects how common each motif is on its own, which is what realism already measures; dividing by the independent expectation leaves only the dependence.

As a worked example, in real GM12878 training data IRF9 appears in 27.2% of sequences and STAT1::STAT2 in 39.6%, so under independence both would appear together in 10.8%. They co-occur in 22.1%, a lift of $2.06\times$ ($\log$ lift 0.72). IRF9 and the STAT1–STAT2 dimer bind interferon-stimulated response elements together as one complex, so their two motifs mark one bound site. BaseDiff produces $2.07\times$ on the same pair, an error of 0.005.

Selecting pairs. The pair list is built from the real data of the respective cell type, by three filters whose per-cell counts are in Table 2.

1. Keep motifs present in 1–50% of real training sequences. Below 1% a motif appears in roughly 50 of $5{,}000$ sequences and pairs of them cannot be estimated; above 50% nearly everything co-occurs with everything and lift has no signal.
2. Drop *redundant* pairs. Two PWMs from one TF family can be near-identical, so a single genomic site matches both and registers as perfect co-occurrence. We drop a pair if the Jaccard overlap of its two hit sets exceeds 0.5, where Jaccard is (sequences containing both)/(sequences containing either), or if the two motif names share a TF-family root including a three-letter stem, so that FOXC2 and FOXD2 count as one family.
3. Sort what remains by real co-occurrence frequency and keep the top 500.

The most extreme redundant pair in GM12878 is Arnt with MAX, two transcription factors whose JASPAR matrices both describe the E-box `CACGTG`: they appear in exactly the same 352 sequences, a Jaccard of 1.000 and an apparent lift of $28\times$, because one genomic site matches both matrices. Left in, such pairs crowd out the top-500 list and are reproduced by anything with roughly correct base composition: on the unfiltered set a shuffle that preserves adjacent base pairs reaches a $\log$ lift correlation of 0.94 against real data.

Table 2: **Pairs removed by each co-occurrence filter.** How many pairs each filter removes, per cell type. JASPAR CORE-vertebrate has $1{,}019$ motifs, giving $518{,}671$ possible pairs before any filtering.

	GM12878	HepG2	K562
motifs present in 1–50% of real sequences	979	973	966
motif pairs	478,731	472,878	466,095
dropped, Jaccard > 0.5	1,597	1,566	1,546
dropped, shared TF-family root	14,276	14,300	14,104
(dropped by both, counted once)	579	586	574
remaining pairs	463,437	457,598	451,019
scored: top 500 by real co-occurrence	500	500	500

Scoring. The reference $\log$ lifts are computed on the real *test* split for those 500 pairs. A generated set is scored by computing $\log$ lift for the same 500 pairs and taking the mean absolute difference from the reference.

C ACROSS-RUN VARIABILITY

Every interval in the paper resamples the sequences of one trained model. We also trained BaseDiff three times from independent initializations and scored all three at the same guidance scales (Table 3). Run 1 is the model reported in every other result. Realism moves by an SD of 0.001–0.003, motif co-occurrence by 0.002–0.004, and the area by 0.001 on the realism and 0.002–0.004 on co-occurrence. The weakest of the three runs still has a larger area than every other method in Table 1 on both measures and all three cell types.

Table 3: **Metrics and trade-off areas over three independent training runs.** Run 1 is the model behind every other result; its two area rows reproduce the BaseDiff-CNG row of Table 1. The first two metrics are unguided samples. SD is taken across the three runs.

	GM12878			HepG2			K562			SD across runs		
Metric	1	2	3	1	2	3	1	2	3	GM	Hep	K562
Realism ↑	0.9353	0.9342	0.9331	0.9291	0.9339	0.9292	0.9285	0.9280	0.9303	0.0011	0.0027	0.0012
Motif co-occurrence ↑	0.9215	0.9183	0.9211	0.9302	0.9274	0.9215	0.9253	0.9311	0.9310	0.0017	0.0044	0.0033
Area, realism ↑	0.3879	0.3851	0.3872	0.3956	0.3947	0.3945	0.3924	0.3899	0.3909	0.0015	0.0006	0.0012
Area, co-occ. ↑	0.4053	0.3970	0.4004	0.3982	0.3958	0.3932	0.4234	0.4202	0.4192	0.0042	0.0025	0.0022

D TRAINING-FREE REFERENCE GENERATORS

Realism compares motif frequencies, so a set can score well on it by reproducing the local statistics of real DNA without actually producing relevant motifs. We score three generators that have no model of grammar and need no training. *Independent per-position* draws each base independently from the real per-cell base frequencies. *Dinucleotide shuffle* permutes each real enhancer while preserving its exact dinucleotide counts. *Order-5 Markov* is a fifth-order Markov chain fit to the real training split. Each is scored exactly as a generated set is (Table 4).

Most of the realism range is reachable this way. The shuffle reaches 0.794 on GM12878 and the Markov chain 0.897, against 0.946 for real held-out enhancers. The Markov value is higher than any model we benchmark reaches once guided to the accessibility of real enhancers, which is visible in Fig. 2: no curve rises above it to the right of the real-enhancer line. Realism alone therefore cannot separate a model that has learned grammar from one that has learned short-range statistics; only the last few hundredths distinguish them.

Accessibility separates them, and none of the three has any: theirs spans 0.031 to 0.316 against 0.437 to 0.760 for real enhancers. A generator can maintain most of the realism while producing nothing an external predictor recognizes as regulatory, which is why §5.1 compares quality at matched accessibility. These generators also help set the threshold for the co-occurrence pair filter: on the unfiltered pair set the shuffle reaches a $\log$ lift correlation of 0.94 against real data (App. B).

Table 4: **Realism, accessibility and co-occurrence for the training-free reference generators.** Realism ($1 - \mathrm{JS}$, higher better), accessibility (higher better) and motif co-occurrence ($1 - \mathrm{MAE}$, higher better), each $\pm$ a 95% sequence bootstrap, at $N = 5{,}000$ per cell.

Cell	Metric	independent per-position	dinucleotide shuffle	order-5 Markov	real training data
GM12878	Realism ↑	0.687 ± 0.011	0.794 ± 0.009	0.897 ± 0.007	0.946 ± 0.006
	Accessibility ↑	0.045 ± 0.001	0.086 ± 0.002	0.172 ± 0.005	0.611 ± 0.027
	Co-occurrence ↑	0.433 ± 0.044	0.637 ± 0.023	0.868 ± 0.012	0.941 ± 0.004
HepG2	Realism ↑	0.693 ± 0.012	0.799 ± 0.010	0.884 ± 0.008	0.942 ± 0.007
	Accessibility ↑	0.031 ± 0.001	0.088 ± 0.004	0.316 ± 0.012	0.760 ± 0.017
	Co-occurrence ↑	0.567 ± 0.036	0.759 ± 0.021	0.859 ± 0.013	0.939 ± 0.004
K562	Realism ↑	0.684 ± 0.011	0.800 ± 0.009	0.900 ± 0.008	0.944 ± 0.007
	Accessibility ↑	0.168 ± 0.003	0.084 ± 0.002	0.162 ± 0.005	0.437 ± 0.035
	Co-occurrence ↑	0.455 ± 0.038	0.686 ± 0.021	0.889 ± 0.009	0.947 ± 0.005

E A SECOND ACCESSIBILITY PREDICTOR

Table 1 uses Enformer to measure accessibility. We repeat the DHS comparison with ChromBPNet (Pampari et al., 2025) to check that the result is not specific to one model. It is also the predictor the second dataset uses throughout (App. G). Enformer predicts many tracks at once; ChromBPNet is one model per cell type, and we use the three released for the same DNase experiments, at fold 0 of the five chromosome splits ENCODE publishes. Each takes 2,114 bp and predicts DNase coverage over the central 1,000. It returns that prediction in two parts: 1,000 logits whose softmax distributes reads over those positions, and one scalar for the log of the total reads the window receives, so predicted coverage at a base is the product of the two. We splice the sequence into the same locus Enformer uses and average the 200 positions it occupies.

Enformer's three tracks come from one model. In contrast, ChromBPNet's three were trained separately at different sequencing depths, and a ratio across them would partly measure depth. Following DaSilva et al. (2026), who report the same confounder, we quantile normalize the three models against each other. Each prediction is replaced by the mean of the three predictions sharing its rank across the scored set, and signal, specificity and accessibility then follow §4.2 on the normalized values.

The ordering holds and the margins are wider (Fig. 6, Table 5). Over the next best of the other methods on each cell type, the realism margins are $+0.026 \pm 0.005$ ($+7.7\%$), $+0.037 \pm 0.006$ ($+11.4\%$) and $+0.046 \pm 0.006$ ($+13.8\%$), and the co-occurrence margins $+0.022 \pm 0.004$ ($+5.8\%$), $+0.019 \pm 0.006$ ($+5.5\%$) and $+0.067 \pm 0.007$ ($+19.3\%$). CNG adds $+0.011 \pm 0.005$, $+0.015 \pm 0.005$ and $+0.030 \pm 0.005$ of realism area over plain CFG.

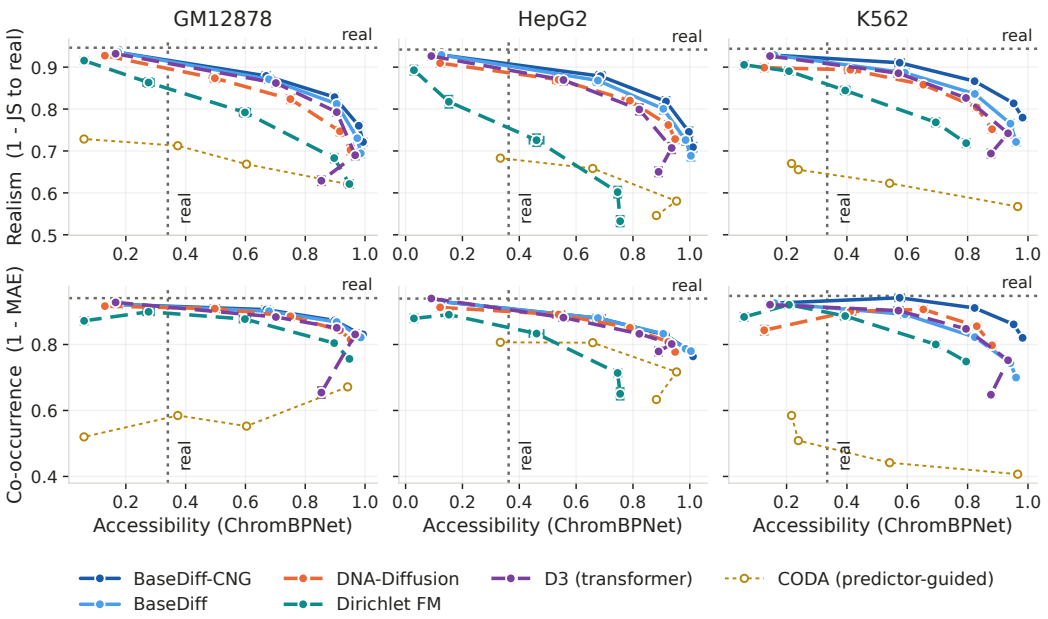

Figure 6: **The accessibility–quality trade-off with ChromBPNet in place of Enformer.** Horizontal: accessibility from three per-cell ChromBPNet models in place of Enformer, plotted as the median over the set. Five points per curve at the guidance scales $w \in \{0, 1, 2, 4, 6\}$ with 95% bootstrap intervals in both directions, and CODA's four points are its refinement rounds. The dotted lines mark real enhancers, their median accessibility vertically and their realism or co-occurrence horizontally.

Table 5: **Trade-off area with ChromBPNet in place of Enformer.** Computed exactly as Table 1 and grouped the same way, from the curves of Fig. 6. Higher is better; the best value in each column is bold and the second best underlined. App. O repeats the table over the accessibility range every model covers.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
DNA-DIFFUSION (94.6M)	0.323 ± 0.008	0.325 ± 0.009	0.316 ± 0.006	0.370 ± 0.003	$\underline{0.349 \pm 0.004}$	0.343 ± 0.005
D3 transformer (92.4M)	0.337 ± 0.008	0.314 ± 0.010	0.333 ± 0.007	0.368 ± 0.004	0.340 ± 0.005	$\underline{0.345 \pm 0.005}$
Dirichlet flow matching (3.7M)	0.258 ± 0.009	0.149 ± 0.010	0.249 ± 0.007	0.336 ± 0.003	0.225 ± 0.005	0.275 ± 0.005
BaseDiff (21.6M)	$\underline{0.352 \pm 0.008}$	$\underline{0.347 \pm 0.011}$	$\underline{0.349 \pm 0.007}$	$\underline{0.385 \pm 0.003}$	$\mathbf{0.368 \pm 0.005}$	0.344 ± 0.005
BaseDiff-CNG (21.6M)	$\mathbf{0.363 \pm 0.007}$	$\mathbf{0.362 \pm 0.009}$	$\mathbf{0.379 \pm 0.006}$	$\mathbf{0.391 \pm 0.003}$	$\mathbf{0.368 \pm 0.005}$	$\mathbf{0.412 \pm 0.005}$

F A SECOND LOCUS

Enformer reads 196,608 bp of context, so a 200 bp sequence cannot be scored on its own: it has to be placed somewhere in the genome, and the prediction can be affected by the locus as well as the inserted sequence. A locus that is itself differentially accessible across the three cell types would contribute signal the generated sequence did not produce.

Every result above uses one window: the 196,608 bp of hg38 centered on chr1:173,050,941, the GENCODE transcription start site of *TNFSF18* (DaSilva et al., 2026), with the two output bins the insert spans, 447 and 448, averaged. *TNFSF18* carries no pre-existing cell-type-specific accessibility in these three lines. ChromBPNet takes a 2,114 bp window, so it is given the center of the same locus (App. E).

The results replicate in an additional locus. We repeat the DHS comparison at chr11:61,685,072, a site chosen by the same criterion: with nothing spliced in, its DNase prediction across the three cell types is 0.059, 0.082 and 0.089, against 0.099, 0.072 and 0.064 at *TNFSF18*, so neither one favors a cell type. Every method and guidance scale of Fig. 2 is scored again there. BaseDiff-CNG again has the largest area on all three cell types and both measures (Table 6). Its margins over the next best other method are $+0.026 \pm 0.005$, $+0.022 \pm 0.005$ and $+0.038 \pm 0.005$ on realism and $+0.020 \pm 0.005$, $+0.007 \pm 0.006$ and $+0.050 \pm 0.008$ on co-occurrence.

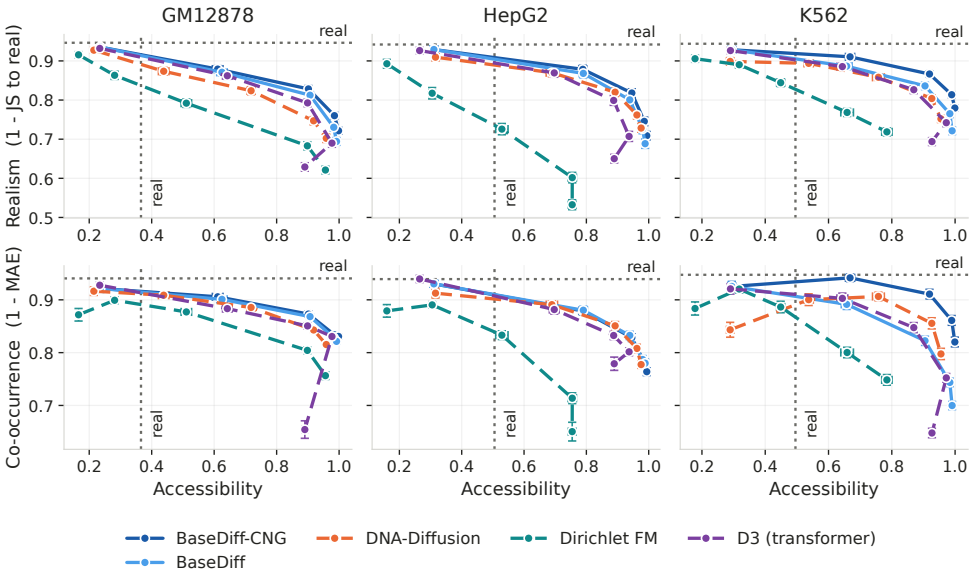

Figure 7: **The accessibility–quality trade-off at the second locus.** Every sequence of Fig. 2 spliced into chr11:61,685,072 in place of *TNFSF18* and scored again, plotted as in Fig. 2. *Top:* realism. *Bottom:* co-occurrence. Five points per curve at $w \in \{0, 1, 2, 4, 6\}$, accessibility the median over a set of 5,000 sequences. Dotted lines mark real enhancers.

Table 6: **Trade-off area at the second locus.** Computed exactly as Table 1, with every sequence spliced into chr11:61,685,072 in place of *TNFSF18*. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
DNA-DIFFUSION (94.6M)	0.324 ± 0.008	0.354 ± 0.008	0.351 ± 0.006	0.370 ± 0.003	$\underline{0.373 \pm 0.004}$	$\underline{0.376 \pm 0.006}$
Dirichlet flow matching (3.7M)	0.256 ± 0.008	0.182 ± 0.010	0.256 ± 0.007	0.332 ± 0.003	0.242 ± 0.004	0.278 ± 0.006
D3 transformer (92.4M)	0.339 ± 0.007	0.340 ± 0.009	0.359 ± 0.006	0.372 ± 0.004	0.360 ± 0.005	0.369 ± 0.005
BaseDiff (21.6M)	$\underline{0.354 \pm 0.007}$	$\underline{0.365 \pm 0.009}$	$\underline{0.372 \pm 0.007}$	$\underline{0.387 \pm 0.003}$	$\mathbf{0.380 \pm 0.004}$	0.368 ± 0.005
BaseDiff-CNG (21.6M)	$\mathbf{0.365 \pm 0.007}$	$\mathbf{0.376 \pm 0.009}$	$\mathbf{0.396 \pm 0.006}$	$\mathbf{0.392 \pm 0.003}$	$\mathbf{0.380 \pm 0.004}$	$\mathbf{0.427 \pm 0.005}$

G A SECOND DATASET: LENTIMPRA

lentiMPRA measures how strongly a short DNA sequence drives transcription after a lentivirus inserts it into the genome. We use the human release of Agarwal et al. (2025) as repackaged in the HuggingFace dataset `zhangtaolab/human_lentiMPRA`, which covers K562, HepG2 and WTC11 (an induced pluripotent stem cell line) at the same 200 bp length as the DHS data. Each sequence also has a measured activity value in each cell. We ignore this value and instead just use the cell type as the label, making it the same task as before. Conditioning a generator on a continuous activity is a related task this work does not address.

The released training split is imbalanced, with 184,450, 123,029 and 54,684 sequences for K562, HepG2 and WTC11. We subsample it to 11,968 per cell, so that it matches the training size of the DHS dataset. The released validation and test splits are kept whole, 45,272 and 45,271 sequences, and the test split is the real reference for every metric. Every model is retrained on this training split with its own setup and its checkpoint chosen on validation, at the same guidance scales and 5,000 sequences per cell per scale as on DHS. Both quality measures are the ones of §4.2, over the same 1,019 motifs, with the real references taken from this dataset's own splits.

Enformer has no DNase track for WTC11 or for any other induced pluripotent line among its 5,313 human tracks, so we measure accessibility with ChromBPNet instead (App. E), one released model per cell, splicing each generated sequence into the same *TNFSF18* locus used previously. We compare against the D3 transformer, the strongest other method on DHS, and against DNA-DIFFUSION, retrained here on the same split. We tried to train Dirichlet flow matching on this data, but did not get reasonable performance: accessibility decreases with higher guidance on the K562 sequences it generates.

We observe lower maximum accessibility values on two of the three cell types here, K562 and WTC11, than on DHS, where every model ends near the top of the range on all three cells. One possibility is the selection criteria. The DHS sequences are peaks accessible in one cell group and not the others, while the lentiMPRA sequences are candidates tested for activity in one cell, with no requirement that they be inactive in the rest: the median share of the three-cell ChromBPNet signal falling in a sequence's own cell is about 0.70 for real DHS peaks and between 0.39 and 0.49 here, against 0.333 for no specificity at all. D3, BaseDiff and BaseDiff-CNG still pass the real level on every cell, but how far past it they go varies by method and by cell type: BaseDiff-CNG reaches furthest on K562 and D3 on WTC11. We therefore report the area over the accessibility window the methods share. DNA-DIFFUSION reaches much lower accessibility here, at most 0.103 on K562 against a real level of 0.212, so on that cell it shares no range with the others; on HepG2 and WTC11 it overlaps them and is compared over the range all four share.

BaseDiff-CNG has the largest area on all three cell types and both measures. Over D3 the margins are $+0.010 \pm 0.003$, $+0.020 \pm 0.004$ and $+0.002 \pm 0.001$ on realism and $+0.020 \pm 0.004$, $+0.052 \pm 0.006$ and $+0.010 \pm 0.001$ on co-occurrence. Over plain CFG they are $+0.008 \pm 0.002$, $+0.011 \pm 0.003$ and $+0.001 \pm 0.001$ on realism and $+0.011 \pm 0.003$, $+0.046 \pm 0.006$ and $+0.007 \pm 0.001$ on co-occurrence.

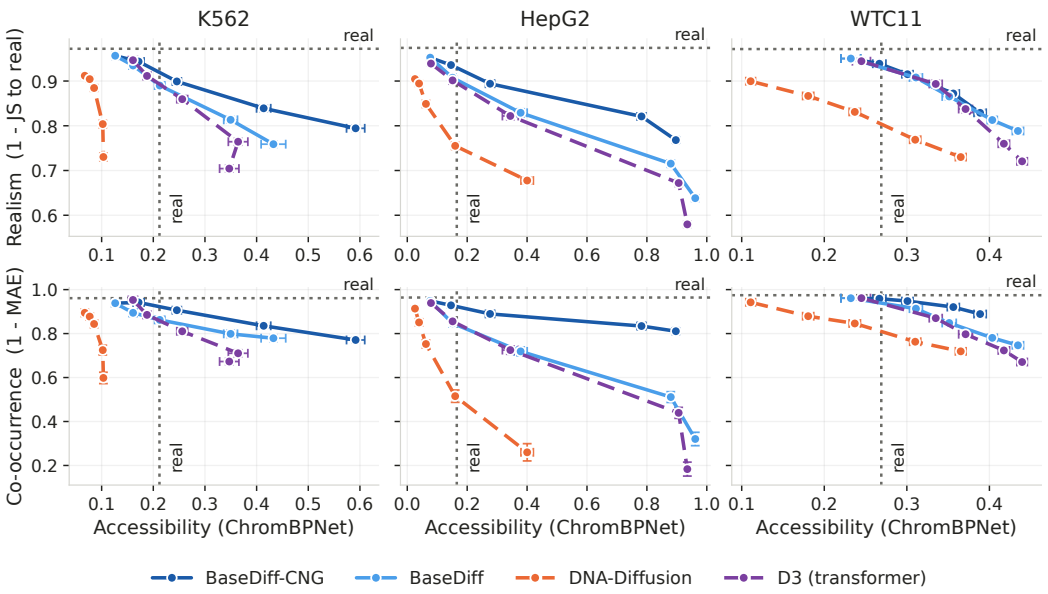

Figure 8: **The accessibility–quality trade-off on lentiMPRA.** Accessibility from ChromBPNet, plotted as in Fig. 2. *Top:* realism. *Bottom:* co-occurrence. Five points per curve at $w \in \{0, 1, 2, 4, 6\}$, accessibility the median over a set of 5,000 sequences. Dotted lines mark real sequences.

Table 7: **Trade-off area on lentiMPRA.** Computed over the accessibility window the methods of that cell share, $[0.161, 0.364]$ on K562, $[0.079, 0.400]$ on HepG2 and $[0.245, 0.365]$ on WTC11. DNA-DIFFUSION shares no range with the others on K562, so it has no entry there and does not enter that cell's window. Areas use the corner and the quality floor of 0.5 of Table 1; four points here fall below that floor, all HepG2 co-occurrence at high guidance and the lowest 0.183, and contribute nothing. Higher is better; the best value in each column is bold and the second best underlined. Values are $\pm$ a 95% bootstrap.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	K562	HepG2	WTC11	K562	HepG2	WTC11
DNA-DIFFUSION (94.6M)	–	0.063 ± 0.003	0.030 ± 0.001	–	0.001 ± 0.002	0.029 ± 0.001
D3 transformer (92.4M)	0.064 ± 0.003	0.101 ± 0.004	$\underline{0.046 \pm 0.001}$	0.054 ± 0.003	0.069 ± 0.006	0.042 ± 0.001
BaseDiff (21.6M)	$\underline{0.067 \pm 0.002}$	$\underline{0.109 \pm 0.003}$	$\underline{0.046 \pm 0.001}$	$\underline{0.064 \pm 0.002}$	$\underline{0.075 \pm 0.006}$	$\underline{0.045 \pm 0.001}$
BaseDiff-CNG (21.6M)	$\mathbf{0.075 \pm 0.001}$	$\mathbf{0.120 \pm 0.002}$	$\mathbf{0.047 \pm 0.001}$	$\mathbf{0.075 \pm 0.002}$	$\mathbf{0.121 \pm 0.003}$	$\mathbf{0.052 \pm 0.001}$

H ENFORMER/CHROMBPNET RAW OUTPUT VALUES

Fig. 9 shows the raw Enformer predictions the DHS accessibility of §4.2 is built from, and Fig. 10 the normalized ChromBPNet predictions behind App. E.

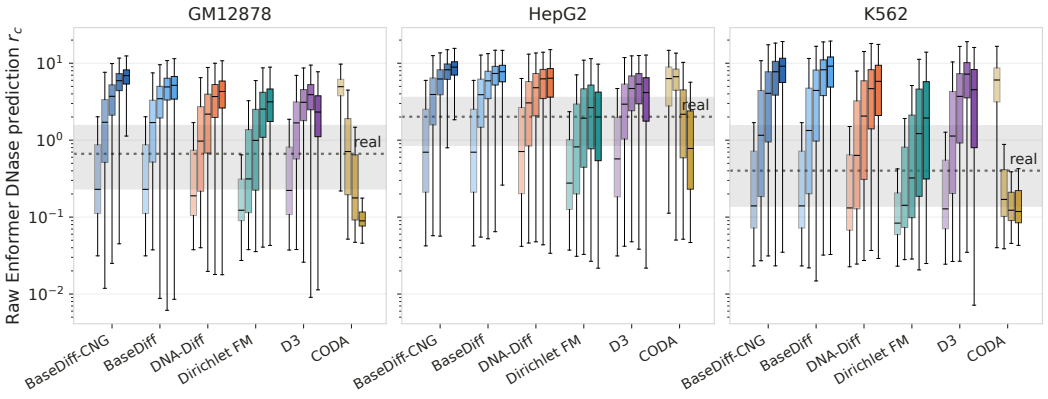

Figure 9: **Enformer DNase predictions across cell types.** Raw per-sequence values, before the normalization of §4.2. One group per model, five boxes each running $w = 0, 1, 2, 4, 6$ from light to dark, at 5,000 sequences per cell. Boxes are quartiles with whiskers at $1.5\times$ the interquartile range, outliers omitted, on a log scale. The shaded band is the interquartile range of the real training sequences and the dotted line their median; every model starts below it, and all but Dirichlet flow matching cross it by $w = 2$ and end well above. CODA has no guidance scale, so its four boxes are its four refinement rounds and darker means further into the optimization.

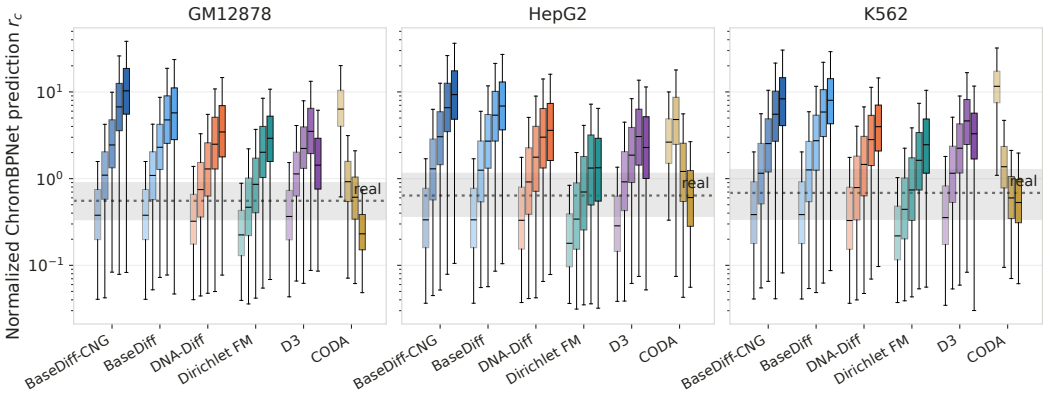

Figure 10: **Quantile-normalized ChromBPNet DNase predictions across cell types.** Per-sequence values after the normalization of App. E and before that of §4.2. One group per model, five boxes each running $w = 0, 1, 2, 4, 6$ from light to dark, at 5,000 sequences per cell. Boxes are quartiles with whiskers at $1.5\times$ the interquartile range, outliers omitted, on a log scale. The shaded band is the interquartile range of the real training sequences and the dotted line their median; unguided models start below it, cross by $w = 2$, and end well above. CODA has no guidance scale, so its four boxes are its four refinement rounds and darker means further into the optimization.

I NOVELTY AND DIVERSITY

A generator can score well on motif statistics by copying training sequences or by repeating itself, so we measure both as guidance rises (Fig. 11). Following DaSilva et al. (2026) we align sequences with BLAT (Kent, 2002) at their settings, `-q=dna -tileSize=11 -stepSize=5 -minMatch=1 -repMatch=2253 -minIdentity=100 -minScore=20 -noHead`, and count a match when some alignment reports at least 20 matched bases. *Training overlap* aligns a generated set against the training split. *Self-matching* aligns a set against itself, discards alignments whose query and target names agree, and reports the fraction matching another sequence in the same set. Each point is the mean over the three cell types at $N = 5{,}000$.

Unguided, both fall below real data on both measures: 6.8% training overlap against the 11.2% that real held-out enhancers reach, and 4.8% self-matching against the 6.7% of the training split. Guidance raises both. Training overlap peaks at 12.8–13.4% at $w = 4$, above the held-out level, and self-matching reaches 33.2–36.0% at $w = 6$. CNG is below plain CFG on self-matching at every guided scale, by 0.3, 0.6, 2.1 and 2.7 percentage points at $w = 1, 2, 4, 6$, and the two are indistinguishable on training overlap. The rise with guidance is expected, since stronger conditioning concentrates the distribution the sampler draws from.

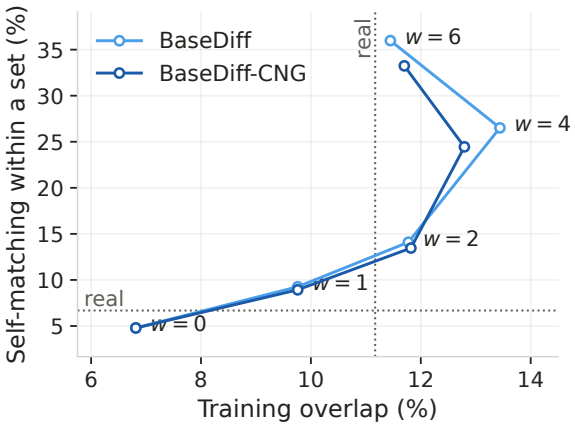

Figure 11: **Novelty against diversity as guidance rises.** Each curve is one sampler over the guidance grid $w \in \{0, 1, 2, 4, 6\}$. Horizontal: the fraction of a generated set that BLAT aligns to the training split at $\geq$20 bp. Vertical: the fraction matching another sequence in the same set. Both are means over the three cell types at $N = 5{,}000$. Dotted lines mark real data, held-out enhancers against the training split (11.2%) and the training split against itself (6.7%).

J BASE COMPOSITION AND PER-MOTIF ERROR UNDER GUIDANCE

Guidance pulls base composition away from the real level in whichever direction the target cell differs from the average, for every other method as well as for ours (Fig. 12); CNG holds the curve at the real level throughout. It shows up as a GC effect because these three cell types differ mainly along GC: GM12878 and HepG2 fall below the three-cell average of 47.1% and K562 above it.

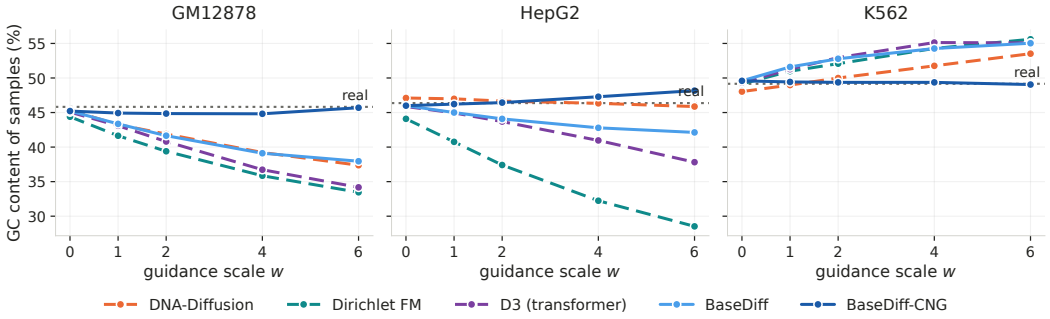

Figure 12: **GC content of generated samples against the guidance scale.** GC content of generated samples against the guidance scale, with the dotted line at the GC content of real held-out enhancers. Uncorrected guidance moves composition away from real in whichever direction the cell type differs from the average, downward on GM12878 and HepG2, upward on GC-rich K562. The other methods move the same way and further, by up to 8.4, 11.6 and 17.8 percentage points at $w = 6$ for DNA-DIFFUSION, D3 and Dirichlet flow matching against 7.9 for BaseDiff under plain CFG. With CNG (dark blue) the curve stays close to the real level throughout.

After guidance, the part of each motif's profile error predictable from base composition is gone (Table 8). Unguided samples show no slope on motif GC content. Plain CFG introduces a large one, negative on the two cells whose composition is below the three-cell average and positive on K562, the one above it. CNG returns the slope to roughly its unguided value, and the share of error variance composition explains falls from 0.07–0.13 to 0.003–0.011. The mean absolute error falls with it, by 9%, 11% and 24%. Realism rises by +0.027 (+3.9%), +0.020 (+2.9%) and +0.058 (+8.0%), and accessibility by +0.001 (+0.1%), +0.003 (+0.3%) and +0.014 (+1.4%). Co-occurrence moves with realism on GM12878 and K562 and falls 0.016 (−2.1%) on HepG2, the one cell whose areas in Table 1 also overlap.

Table 8: **Per-motif error against motif GC content, and the quality metrics, at $w = 6$.** A set's error on motif j is $e_j = f_j^{\text{gen}} - f_j^{\text{real}}$ against the real held-out profile, with f_j the motif profile of §4.2. Left: an ordinary least-squares fit of e_j against the GC content of motif j itself, the fraction of positions in its matrix whose highest-scoring base is C or G, one point per motif over all 1,019. Right: the mean of $|e_j|$ over the same motifs, and the three quality metrics. All on the same three sets of 5,000 sequences per cell, guided at $w = 6$.

		error vs. motif GC			quality		
Cell	Sampling	slope	R^2	mean |error|	realism	co-occ.	access.
GM12878	unguided	−0.006	0.019	0.0067	0.935	0.921	0.348
	plain CFG	−0.088	0.072	0.0373	0.694	0.821	1.003
	+CNG	+0.017	0.003	0.0341	0.722	0.831	1.004
HepG2	unguided	−0.005	0.010	0.0077	0.929	0.930	0.544
	plain CFG	−0.108	0.088	0.0433	0.689	0.780	0.991
	+CNG	−0.035	0.011	0.0387	0.709	0.764	0.994
K562	unguided	0.000	0.000	0.0072	0.928	0.925	0.220
	plain CFG	+0.112	0.133	0.0331	0.721	0.700	0.977
	+CNG	+0.021	0.007	0.0251	0.780	0.820	0.991

K ALTERNATIVE GUIDANCE CORRECTIONS

Eq. 3 subtracts comp_b, the mean of the guidance direction $d = \ell_c - \ell_\varnothing$ over the updated positions $\mathcal{M}$, from every position. That is the optimal constant shift in logit space, but the quantity we care about is motif content after 128 reverse steps, and the map from a logit shift to that is not linear due to the softmax. We therefore tested two alternative estimators of the shift against it, both on the same trained network at $w = 4$ with 5,000 sequences per cell. The first replaces the mean of d with its per-base *median*. The second works in *probability space*. Softmaxing the logits at a position gives the four base probabilities there, and averaging those over positions gives the base composition a set of logits predicts. Write $\rho^\star$ for that composition taken from ℓ_c, which is the target; $\ell^{(k)}$ for the logits after k rounds of correction, starting from the guided logits of Eq. 3; and $\rho^{(k)}$ for the composition of $\ell^{(k)}$:

$$\rho^\star_b = \frac{1}{|\mathcal{M}|}\sum_{p\in\mathcal{M}} \mathrm{softmax}(\ell_{c,p})_b, \quad \rho^{(k)}_b = \frac{1}{|\mathcal{M}|}\sum_{p\in\mathcal{M}} \mathrm{softmax}\big(\ell^{(k)}_p\big)_b, \quad \ell^{(k+1)}_{p,b} = \ell^{(k)}_{p,b} + \log\frac{\rho^\star_b}{\rho^{(k)}_b}, \tag{7}$$

with $\ell^{(0)} = \ell^{\mathrm{guided}}$, $k = 0, \ldots, 39$ and $\ell^{(k)}_p$ the four logits at position p. Each round adds one number per base to every position, as the mean does, but unlike the mean and median corrections, which are complete in one step, this one is iterated, since the softmax renormalizes each position separately. The number of rounds, 40, is arbitrary; by the last one $\rho^{(k)}$ and $\rho^\star$ are close to matching. The iteration is repeated at every one of the T reverse steps. Eq. 3 zeroes the average push each base receives in logit space; Eq. 7 holds the predicted composition at the value ℓ_c predicts, which the mean does only approximately.

We do not observe any noticeable difference between the approaches (Table 9), except on K562 co-occurrence, where the mean is ahead of the median by 0.012 ± 0.013 and of the probability space variant by 0.025 ± 0.013. The mean appears to remove most of the composition error on its own, and it is the simplest of the three: the probability space variant needs 40 iterations at every reverse step to reach a similar correction.

Table 9: **Alternative ways of estimating the composition shift, at $w = 4$.** One trained network, one guidance scale, 5,000 sequences per cell; the estimator is the only variable. Higher is better on both measures. Values are $\pm$ a 95% sequence bootstrap Plain CFG is shown for scale.

	GM12878		HepG2		K562	
Estimator	realism	co-occ.	realism	co-occ.	realism	co-occ.
per-base mean (used)	0.760 ± 0.008	0.834 ± 0.005	0.745 ± 0.011	0.782 ± 0.006	0.814 ± 0.008	0.861 ± 0.009
per-base median	0.760 ± 0.008	0.830 ± 0.006	0.739 ± 0.011	0.771 ± 0.007	0.816 ± 0.007	0.849 ± 0.009
probability space	0.764 ± 0.008	0.836 ± 0.006	0.741 ± 0.011	0.778 ± 0.007	0.811 ± 0.008	0.836 ± 0.009
no correction (plain CFG)	0.730 ± 0.010	0.828 ± 0.005	0.726 ± 0.012	0.787 ± 0.007	0.765 ± 0.010	0.744 ± 0.008

All three corrections modify the same trained network (BaseDiff), so they reach the same accessibility at a given guidance scale and realism and co-occurrence can be compared at one scale. ChromBPNet puts the median accessibility of the mean at $w = 4$ at 0.979, 0.997 and 0.952 on GM12878, HepG2 and K562, with the median and probability space variants within 0.003 of it on every cell.

L CORRECTIONS TO CLASSIFIER-FREE GUIDANCE FROM IMAGE GENERATION

The corrections of §2 were mainly developed for image and video generation. Two of them only use the conditional and unconditional predictions, so they can be applied to categorical logits: APG (Sadat et al., 2025a), which removes the component of the update lying along the conditional prediction, and CFG-Zero⋆ (Fan et al., 2025), which rescales the unconditional prediction before the extrapolation. We apply both to BaseDiff and generate on the same guidance grid as Table 1 with their recommended settings: for APG the parallel component removed ($\eta = 0$), reverse momentum ($\beta = -0.5$) and a rescaling threshold set by their rule that it be comparable to the norm of the update ($r = 3$, against a norm of 1.4–5.6 here). The other methods do not apply: Chung et al. (2025) reformulates the DDIM step, Sadat et al. (2025b) needs a frequency decomposition and Dou et al. (2026) guides with an external reward model that would have to be trained for this task. Lin et al. (2024) rescales the magnitude of the guided prediction, which on logits changes only the sampling temperature.

CNG has the larger area on five of the six comparisons against APG measured with Enformer (Fig. 13, Table 10) and on all six measured with ChromBPNet (Fig. 14, Table 11). The Enformer exception is HepG2 co-occurrence, where all four sampling methods lie within 0.005 of each other. The margins are widest on K562, the benchmark cell whose base composition differs most from the three-cell average (App. J).

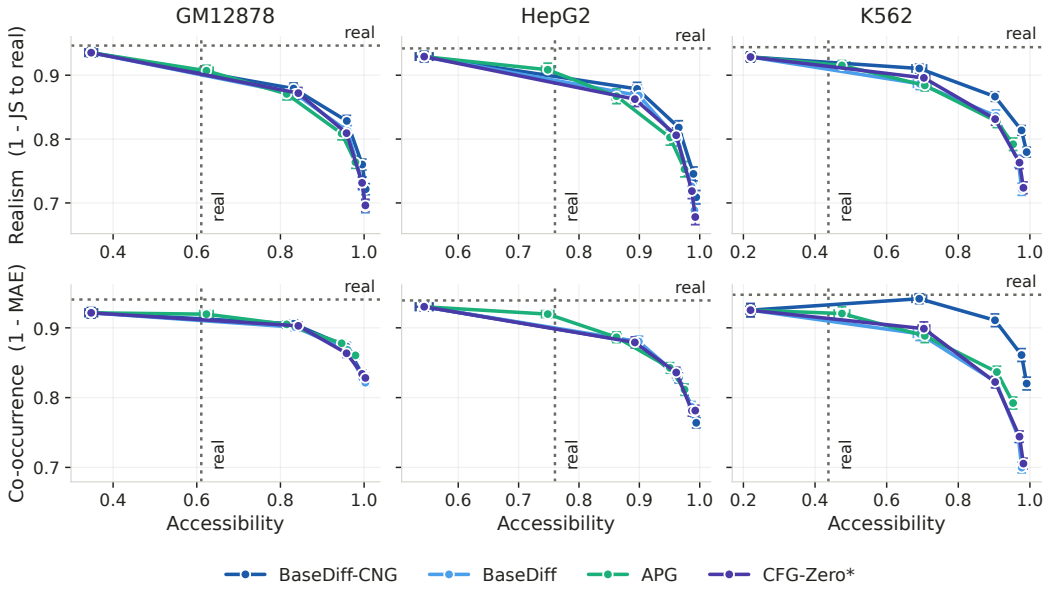

Figure 13: **The accessibility–quality trade-off for BaseDiff under each correction.** Drawn as in Fig. 2. All four curves are BaseDiff, sampled with plain CFG, with APG, with CFG-Zero⋆ and with CNG. Five points per curve at $w \in \{0, 1, 2, 4, 6\}$, accessibility the median over the set.

Table 10: **Trade-off area for each correction.** Computed exactly as Table 1, the same trained network sampled four ways. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
BaseDiff (21.6M)	0.381 ± 0.007	0.390 ± 0.008	0.367 ± 0.007	0.402 ± 0.003	$\underline{0.398 \pm 0.004}$	0.364 ± 0.005
CFG-Zero⋆	0.381 ± 0.007	0.387 ± 0.008	$\underline{0.372 \pm 0.007}$	$\underline{0.403 \pm 0.004}$	$\underline{0.398 \pm 0.005}$	0.370 ± 0.006
APG	$\underline{0.384 \pm 0.007}$	$\underline{0.392 \pm 0.008}$	0.368 ± 0.006	0.401 ± 0.003	$\mathbf{0.402 \pm 0.004}$	$\underline{0.372 \pm 0.005}$
BaseDiff-CNG (21.6M)	$\mathbf{0.388 \pm 0.006}$	$\mathbf{0.396 \pm 0.008}$	$\mathbf{0.392 \pm 0.006}$	$\mathbf{0.405 \pm 0.004}$	$\underline{0.398 \pm 0.004}$	$\mathbf{0.423 \pm 0.005}$

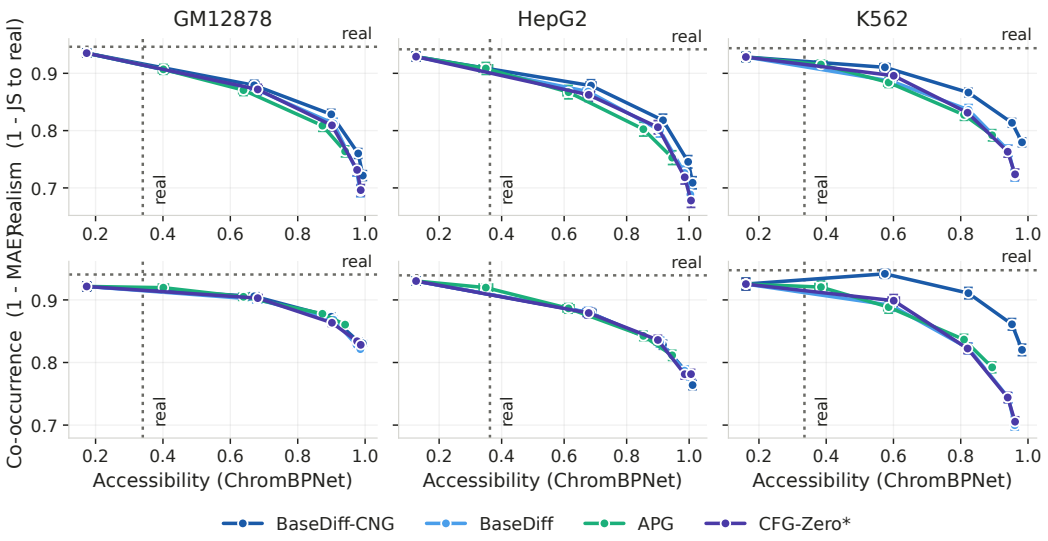

Figure 14: **Fig. 13 with ChromBPNet in place of Enformer.** Accessibility from three per-cell ChromBPNet models, as in App. E. Five points per curve at $w \in \{0, 1, 2, 4, 6\}$, accessibility the median over the set.

Table 11: **Table 10 recomputed with ChromBPNet.** Accessibility from ChromBPNet in place of Enformer, as in App. E. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
BaseDiff (21.6M)	$\underline{0.352 \pm 0.008}$	$\underline{0.347 \pm 0.011}$	0.349 ± 0.007	$\underline{0.386 \pm 0.003}$	$\mathbf{0.368 \pm 0.005}$	0.344 ± 0.005
CFG-Zero*	$\underline{0.352 \pm 0.008}$	0.344 ± 0.011	$\underline{0.352 \pm 0.007}$	$\underline{0.386 \pm 0.003}$	$\mathbf{0.368 \pm 0.005}$	$\underline{0.348 \pm 0.005}$
APG	0.346 ± 0.007	0.338 ± 0.010	0.337 ± 0.006	0.378 ± 0.003	$\underline{0.361 \pm 0.003}$	0.340 ± 0.004
BaseDiff-CNG (21.6M)	$\mathbf{0.363 \pm 0.007}$	$\mathbf{0.362 \pm 0.010}$	$\mathbf{0.379 \pm 0.006}$	$\mathbf{0.391 \pm 0.003}$	$\mathbf{0.368 \pm 0.005}$	$\mathbf{0.412 \pm 0.005}$

M THE CHOICE OF FORWARD PROCESS

Three backbones were trained identically in architecture (conv, width 320, depth 12, 21.6M), data, optimizer, schedule and budget, differing only in that target: each cell's own base composition π_c; a single global composition π; and uniform. All three share one sampler, with uniform the $\pi = 1/4$ case. Each is run at $w \in \{0, 1, 2, 4, 6\}$ and sampled twice, with CNG and with plain CFG (Fig. 15, Table 12). With CNG the three are indistinguishable, so we use the simplest. With plain CFG they are not: π_c matches the corrected models except on K562 co-occurrence, where CNG is ahead, while uniform and global π are lower on both measures.

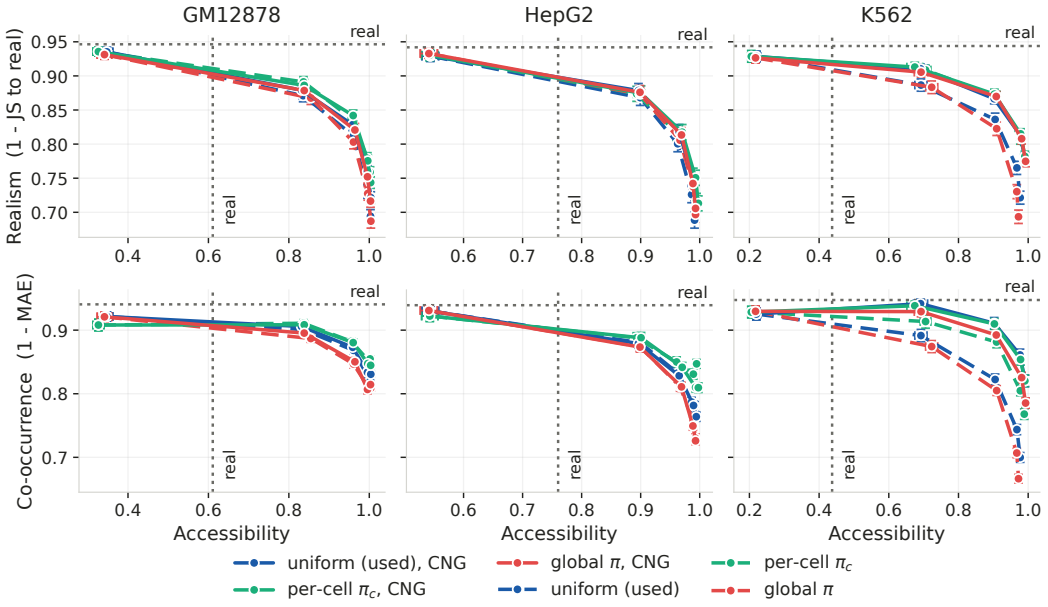

Figure 15: **The accessibility–quality trade-off for three forward processes.** Three otherwise-identical backbones differing only in what the diffusion destroys toward, each drawn twice: with CNG (solid) and with plain CFG (dashed), as in Fig. 2, five points per curve at $w \in \{0, 1, 2, 4, 6\}$. Accessibility is the median over the set and the dotted lines mark real held-out enhancers, both as in Fig. 2.

Table 12: **Trade-off area for three forward processes.** Computed as Table 1, the same three trained networks sampled two ways. In the upper group every interval overlaps every other on every cell type and both measures. In the lower group per-cell π_c matches the corrected models while uniform and global π are lower.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Forward process	GM12878	HepG2	K562	GM12878	HepG2	K562
Sampled with CNG						
uniform (used)	0.388 ± 0.006	0.396 ± 0.008	0.392 ± 0.006	0.405 ± 0.004	0.398 ± 0.004	0.423 ± 0.005
per-cell π_c	0.393 ± 0.006	0.396 ± 0.008	0.394 ± 0.006	0.404 ± 0.004	0.400 ± 0.005	0.420 ± 0.006
global π	0.385 ± 0.007	0.396 ± 0.007	0.391 ± 0.006	0.396 ± 0.004	0.394 ± 0.005	0.409 ± 0.005
Sampled with plain CFG						
uniform (used)	0.381 ± 0.007	0.390 ± 0.008	0.367 ± 0.007	0.402 ± 0.003	0.398 ± 0.004	0.364 ± 0.005
per-cell π_c	0.395 ± 0.006	0.394 ± 0.008	0.392 ± 0.006	0.406 ± 0.004	0.401 ± 0.005	0.396 ± 0.006
global π	0.378 ± 0.008	0.395 ± 0.007	0.361 ± 0.007	0.392 ± 0.004	0.395 ± 0.005	0.352 ± 0.006

N VARIANTS OF BASEDIFF

Each variant changes one component of BaseDiff and is otherwise trained and sampled identically. Each is evaluated under plain CFG against our own plain-CFG curve (Fig. 16, Table 13).

We find that the backbone has more impact than the training objective. Swapping the objective for score entropy, with backbone, data, schedule and parameter count fixed, moves the realism area by $+0.010$, $+0.007$ and -0.006, small and inconsistent in sign. Swapping the convolutional backbone for a DiT transformer lowers the area by more and always in the same direction, by $+0.031$ to $+0.082$ depending on cell type and size. Replacing the substitution process with a masked one, which uses its own decoder, falls between the two: $+0.009$, $+0.005$ and $+0.003$ on realism. Retraining one configuration moves this area by about 0.001 (App. C), so the backbone differences are tens of times larger than retraining noise and the objective differences are close enough to it that their exact sizes mean little.

BaseDiff-CNG is ahead of every variant on both measures and all three cell types except HepG2 co-occurrence, where masked diffusion is ahead by less than its interval. Both DiT transformers lose accessibility between $w = 4$ and $w = 6$ on HepG2 and the autoregressive model does the same on K562, in each case alongside lower quality on both measures; the score-entropy objective and masked diffusion raise accessibility monotonically on every cell type. Fig. 17 and Table 14 repeat both with ChromBPNet in place of Enformer.

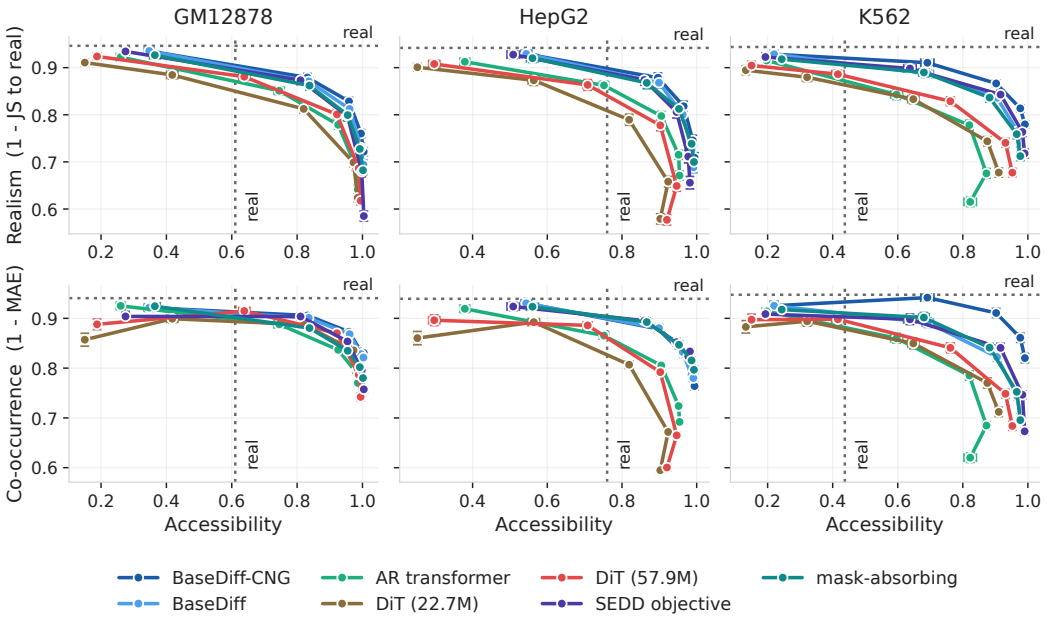

Figure 16: **The accessibility–quality trade-off for BaseDiff variants.** Drawn as in Fig. 2, with the other methods replaced by the variants and every curve sampled under plain CFG except BaseDiff-CNG. Five points per curve at $w \in \{0, 1, 2, 4, 6\}$, accessibility the median over the set.

Table 13: **Trade-off area for BaseDiff variants.** Computed exactly as Table 1, with the other methods replaced by the variants. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
AR transformer (21.4M)	0.342 ± 0.007	0.347 ± 0.006	0.290 ± 0.008	0.377 ± 0.004	0.353 ± 0.005	0.299 ± 0.005
DiT transformer (22.7M)	0.322 ± 0.007	0.308 ± 0.008	0.294 ± 0.007	0.378 ± 0.004	0.318 ± 0.005	0.310 ± 0.006
DiT transformer (57.9M)	0.350 ± 0.007	0.331 ± 0.008	0.322 ± 0.007	0.390 ± 0.004	0.340 ± 0.005	0.329 ± 0.006
SEDD objective (21.6M)	0.370 ± 0.008	0.383 ± 0.008	$\underline{0.373 \pm 0.006}$	0.393 ± 0.004	0.397 ± 0.005	0.368 ± 0.006
mask-absorbing (21.6M)	0.372 ± 0.008	0.384 ± 0.009	0.364 ± 0.007	0.388 ± 0.004	$\mathbf{0.400 \pm 0.005}$	$\underline{0.369 \pm 0.005}$
BaseDiff (21.6M)	$\underline{0.381 \pm 0.007}$	$\underline{0.390 \pm 0.008}$	0.367 ± 0.007	$\underline{0.402 \pm 0.003}$	$\underline{0.398 \pm 0.004}$	0.364 ± 0.005
BaseDiff-CNG (21.6M)	$\mathbf{0.388 \pm 0.006}$	$\mathbf{0.396 \pm 0.008}$	$\mathbf{0.392 \pm 0.006}$	$\mathbf{0.405 \pm 0.004}$	$\underline{0.398 \pm 0.004}$	$\mathbf{0.423 \pm 0.005}$

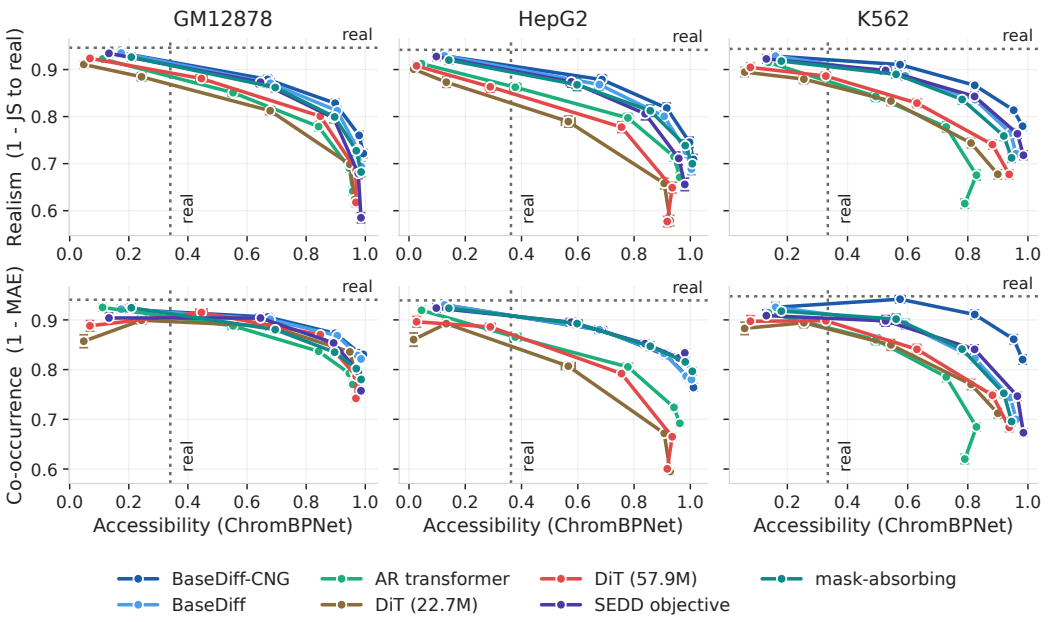

Figure 17: **The variants' trade-off with ChromBPNet in place of Enformer.** Fig. 16 with accessibility from ChromBPNet in place of Enformer, as in App. E. Five points per curve at $w \in \{0, 1, 2, 4, 6\}$, accessibility the median over the set.

Table 14: **Table 13 recomputed with ChromBPNet.** Accessibility from ChromBPNet in place of Enformer, as in App. E. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
AR transformer (21.4M)	0.318 ± 0.007	0.294 ± 0.008	0.306 ± 0.006	0.355 ± 0.004	0.287 ± 0.004	0.339 ± 0.005
DiT transformer (22.7M)	0.311 ± 0.007	0.240 ± 0.010	0.311 ± 0.006	0.370 ± 0.004	0.249 ± 0.005	0.361 ± 0.005
DiT transformer (57.9M)	0.335 ± 0.007	0.267 ± 0.010	0.333 ± 0.006	0.374 ± 0.004	0.276 ± 0.004	$\underline{0.380 \pm 0.004}$
SEDD objective (21.6M)	0.349 ± 0.007	0.336 ± 0.010	$\underline{0.362 \pm 0.006}$	0.377 ± 0.004	$\underline{0.371 \pm 0.004}$	0.371 ± 0.006
mask-absorbing (21.6M)	0.345 ± 0.008	0.343 ± 0.011	0.339 ± 0.007	0.368 ± 0.004	$\mathbf{0.376 \pm 0.004}$	0.343 ± 0.005
BaseDiff (21.6M)	$\underline{0.352 \pm 0.008}$	$\underline{0.347 \pm 0.011}$	0.349 ± 0.007	$\underline{0.385 \pm 0.003}$	0.368 ± 0.005	0.344 ± 0.005
BaseDiff-CNG (21.6M)	$\mathbf{0.363 \pm 0.007}$	$\mathbf{0.362 \pm 0.009}$	$\mathbf{0.379 \pm 0.006}$	$\mathbf{0.391 \pm 0.003}$	0.368 ± 0.005	$\mathbf{0.412 \pm 0.005}$

O TRADE-OFF AREAS OVER THE SHARED ACCESSIBILITY RANGE

Table 15 repeats Table 1 and Table 16 repeats Table 5, each over the accessibility range every model covers.

Table 15: **Table 1 over the accessibility range every model covers.** The area is measured over the accessibility range every model covers, from the highest of their starting points to the lowest of their end points: GM12878 $[0.36, 0.97]$, HepG2 $[0.56, 0.85]$, K562 $[0.24, 0.81]$. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
DNA-DIFFUSION (94.6M)	0.205 ± 0.006	$\underline{0.107 \pm 0.003}$	$\underline{0.214 \pm 0.004}$	0.237 ± 0.002	$\mathbf{0.113 \pm 0.002}$	$\underline{0.231 \pm 0.004}$
D3 transformer (92.4M)	0.209 ± 0.005	$\underline{0.107 \pm 0.003}$	0.212 ± 0.005	0.228 ± 0.003	$\underline{0.110 \pm 0.002}$	0.222 ± 0.005
Dirichlet flow matching (3.7M)	0.159 ± 0.006	0.062 ± 0.004	0.163 ± 0.006	0.214 ± 0.002	0.091 ± 0.002	0.183 ± 0.004
BaseDiff (21.6M)	$\underline{0.216 \pm 0.005}$	0.106 ± 0.003	$\underline{0.214 \pm 0.005}$	$\underline{0.238 \pm 0.002}$	$\underline{0.110 \pm 0.002}$	0.214 ± 0.004
BaseDiff-CNG (21.6M)	$\mathbf{0.222 \pm 0.005}$	$\mathbf{0.109 \pm 0.003}$	$\mathbf{0.228 \pm 0.004}$	$\mathbf{0.241 \pm 0.003}$	0.109 ± 0.002	$\mathbf{0.248 \pm 0.003}$

Table 16: **Table 5 over the accessibility range every model covers.** The window is GM12878 $[0.21, 0.95]$, HepG2 $[0.14, 0.75]$, K562 $[0.18, 0.80]$. Higher is better; the best value in each column is bold and the second best underlined.

	Accessibility–realism tradeoff			Accessibility–co-occurrence tradeoff		
Model	GM12878	HepG2	K562	GM12878	HepG2	K562
DNA-DIFFUSION (94.6M)	0.237 ± 0.007	0.216 ± 0.006	0.221 ± 0.004	0.282 ± 0.002	$\mathbf{0.231 \pm 0.004}$	$\underline{0.243 \pm 0.004}$
D3 transformer (92.4M)	0.246 ± 0.006	0.212 ± 0.007	0.224 ± 0.005	0.274 ± 0.003	0.224 ± 0.004	0.235 ± 0.004
Dirichlet flow matching (3.7M)	0.179 ± 0.007	0.102 ± 0.009	0.178 ± 0.006	0.252 ± 0.003	0.169 ± 0.004	0.199 ± 0.004
BaseDiff (21.6M)	$\underline{0.255 \pm 0.007}$	$\underline{0.220 \pm 0.007}$	$\underline{0.228 \pm 0.005}$	$\underline{0.286 \pm 0.003}$	$\underline{0.229 \pm 0.004}$	0.227 ± 0.004
BaseDiff-CNG (21.6M)	$\mathbf{0.263 \pm 0.006}$	$\mathbf{0.228 \pm 0.006}$	$\mathbf{0.243 \pm 0.004}$	$\mathbf{0.289 \pm 0.003}$	$\underline{0.229 \pm 0.004}$	$\mathbf{0.265 \pm 0.003}$