NEURAL INFORMATION
PROCESSING SYSTEMS

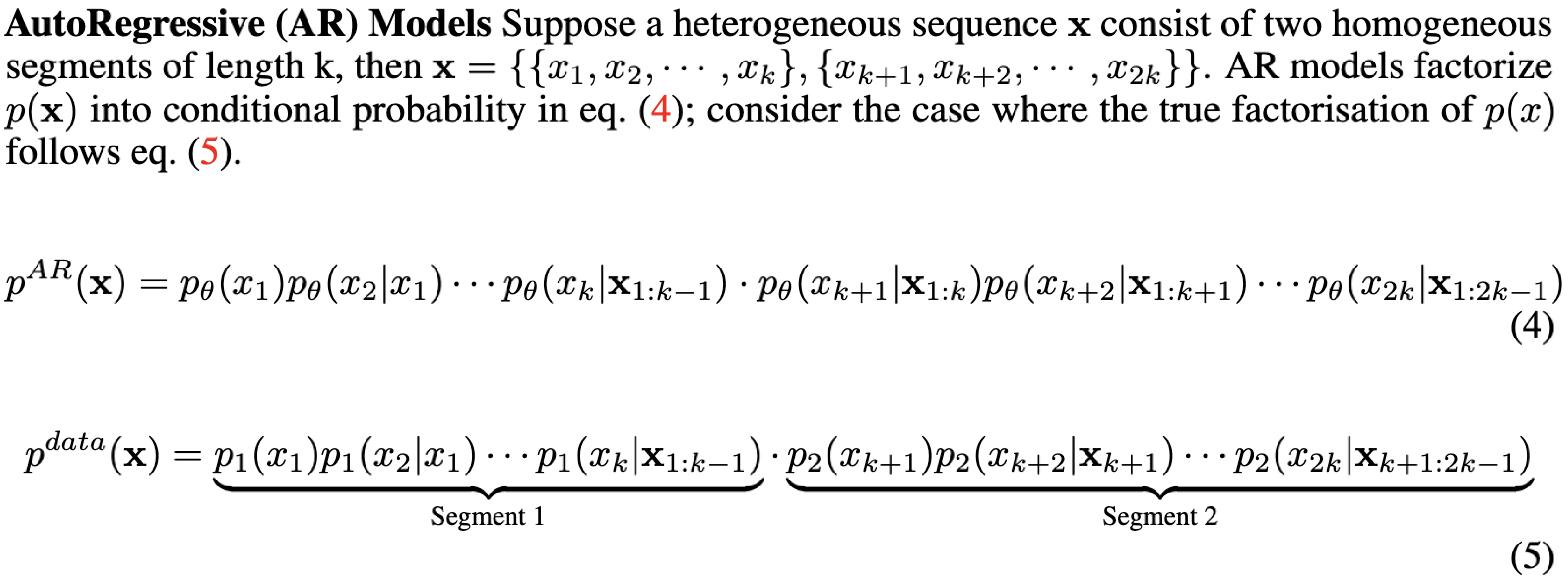
Absorb & Escape: Overcoming Single Model Limitations in Generating Genomic Sequences

Zehui Li[1], Yuhao Ni[1], Guoxuan Xia[1], William Beardall[1], Akashaditya Das[1], Guy-Bart Stan[1], Yiren Zhao[1]

[1]Imperial College London

IMPERIAL

Limitations of Existing Single-Model Approaches in Generating DNA



AutoRegressive (AR) Models Suppose a heterogeneous sequence $\mathbf{x}$ consist of two homogeneous segments of length k, then $\mathbf{x} = \{\{x_1, x_2, \cdots, x_k\}, \{x_{k+1}, x_{k+2}, \cdots, x_{2k}\}\}$. AR models factorize $p(\mathbf{x})$ into conditional probability in eq. (4); consider the case where the true factorisation of $p(x)$ follows eq. (5).

$$p^{AR}(\mathbf{x}) = p_\theta(x_1)p_\theta(x_2|x_1)\cdots p_\theta(x_k|\mathbf{x}_{1:k-1}) \cdot p_\theta(x_{k+1}|\mathbf{x}_{1:k})p_\theta(x_{k+2}|\mathbf{x}_{1:k+1})\cdots p_\theta(x_{2k}|\mathbf{x}_{1:2k-1}) \tag{4}$$

$$p^{data}(\mathbf{x}) = \underbrace{p_1(x_1)p_1(x_2|x_1)\cdots p_1(x_k|\mathbf{x}_{1:k-1})}_{\text{Segment 1}} \cdot \underbrace{p_2(x_{k+1})p_2(x_{k+2}|\mathbf{x}_{k+1})\cdots p_2(x_{2k}|\mathbf{x}_{k+1:2k-1})}_{\text{Segment 2}} \tag{5}$$

Limitations of Existing Single-Model Approaches in Generating DNA

- AR Model may struggle to disassociate the elements of the second segment from the first segment
- Sufficient data is needed for AR model to learn two segments are independent

$$p^{AR}(\mathbf{x}) = p_\theta(x_1)p_\theta(x_2|x_1)\cdots p_\theta(x_k|\mathbf{x}_{1:k-1}) \cdot p_\theta(x_{k+1}|\mathbf{x}_{1:k})p_\theta(x_{k+2}|\mathbf{x}_{1:k+1})\cdots p_\theta(x_{2k}|\mathbf{x}_{1:2k-1}) \tag{4}$$

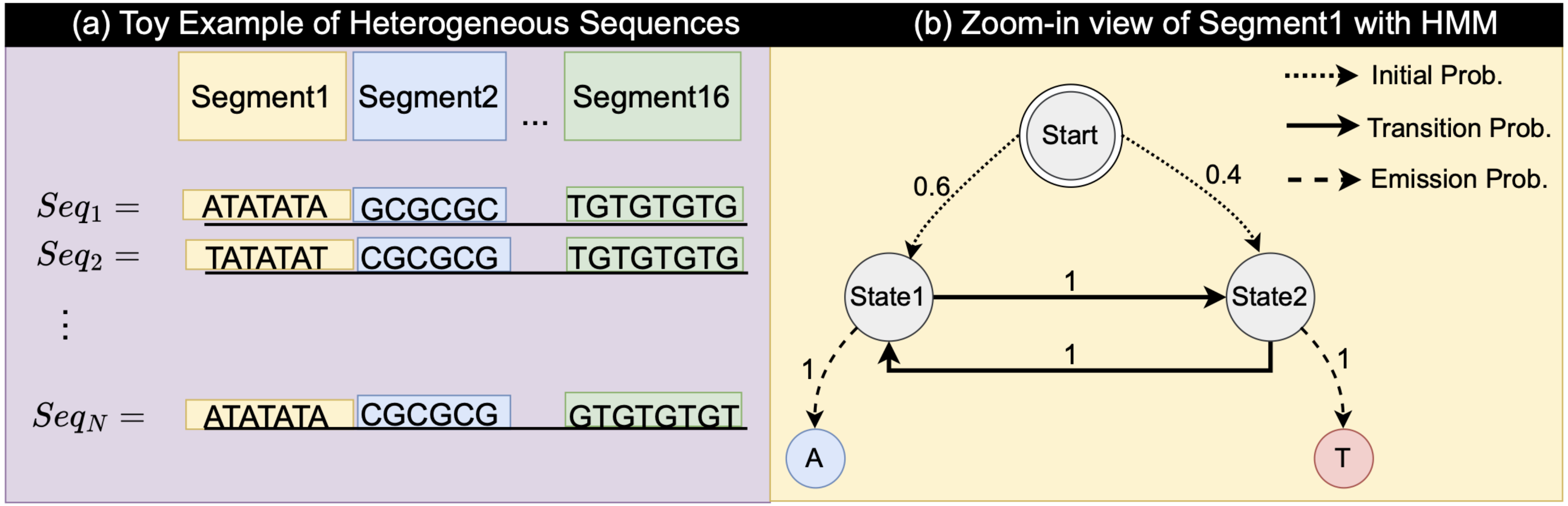
$$p^{data}(\mathbf{x}) = \underbrace{p_1(x_1)p_1(x_2|x_1)\cdots p_1(x_k|\mathbf{x}_{1:k-1})}_{\text{Segment 1}} \cdot \underbrace{p_2(x_{k+1})p_2(x_{k+2}|\mathbf{x}_{k+1})\cdots p_2(x_{2k}|\mathbf{x}_{k+1:2k-1})}_{\text{Segment 2}} \tag{5}$$

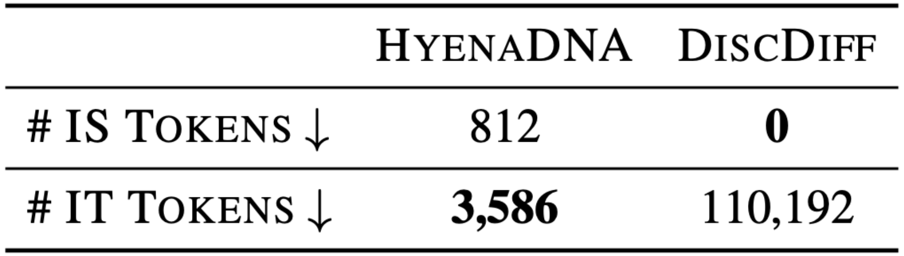
Limitations of Existing Single-Model Approaches in Generating DNA

How about diffusion model?

- DMs estimate the overall probability distribution p(x) without factorization
- However, the removal of the conditional dependence assumption may also decrease the accuracy of generation within each homogeneous segment

Limitations of Existing Single-Model: Toy Example





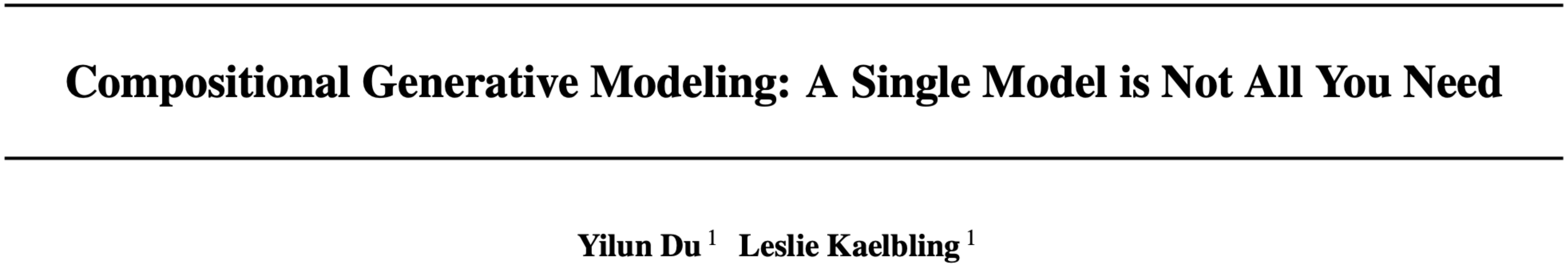
	HyenaDNA	DiscDiff
# IS Tokens ↓	812	**0**
# IT Tokens ↓	**3,586**	110,192

IS Tokens: illegal Start Token

IS Tokens: illegal Transition Token

Number of Incorret Tokens on Synthetic Dataset.

Solution to Single Molde Limitations: Model Composition

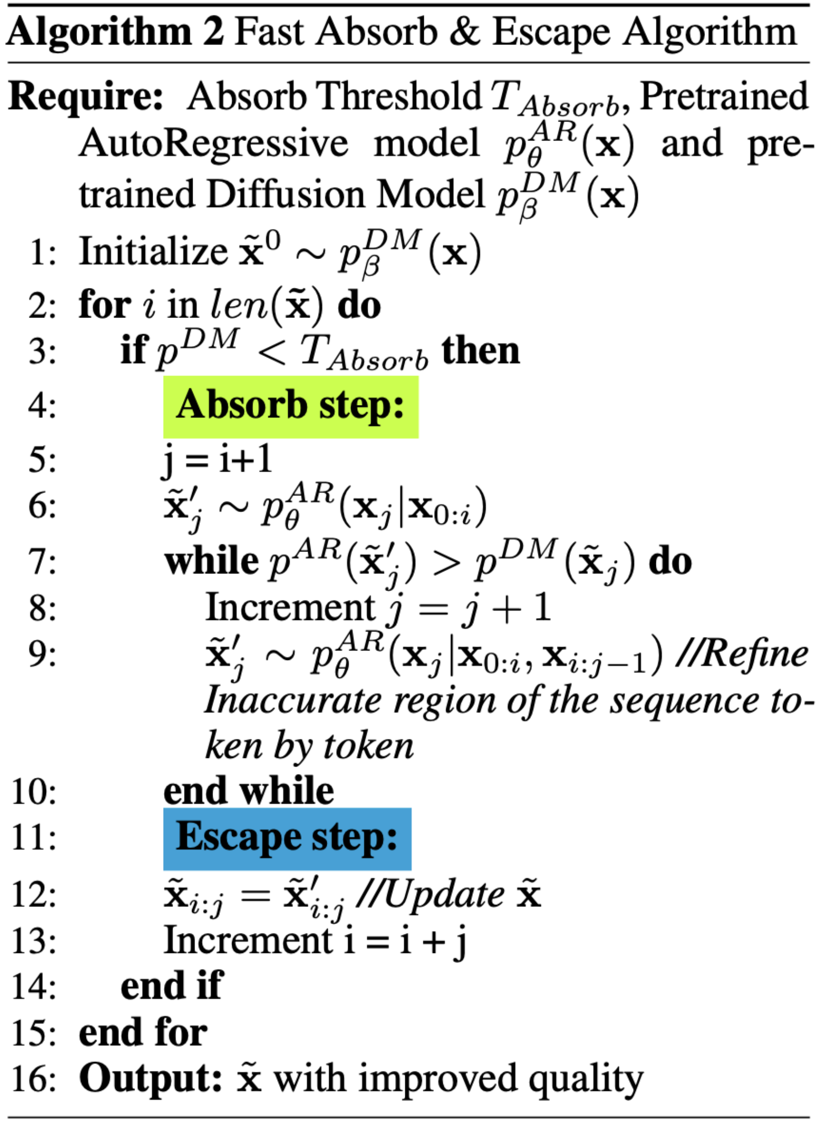


Compositional Generative Modeling: A Single Model is Not All You Need

Yilun Du[1] **Leslie Kaelbling**[1]

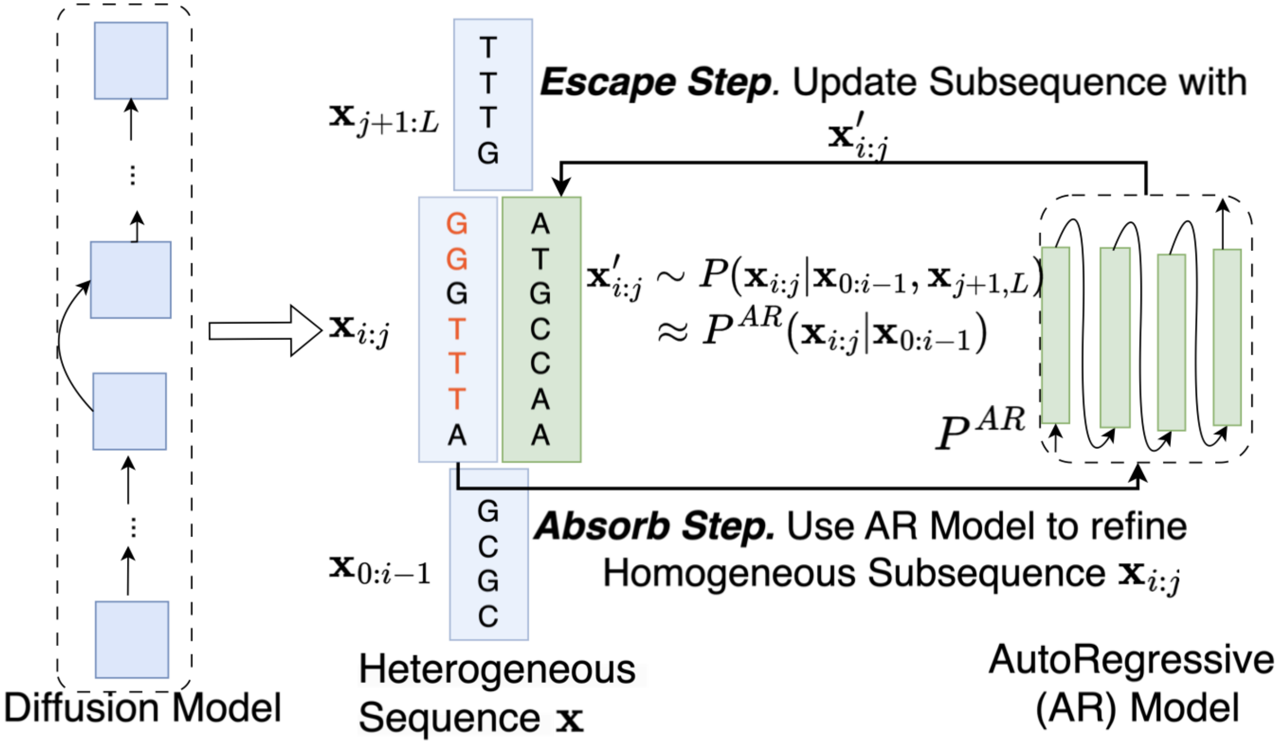
But Energy Based Model is Slow …

Solution to Single Molde Limitations: Model Composition

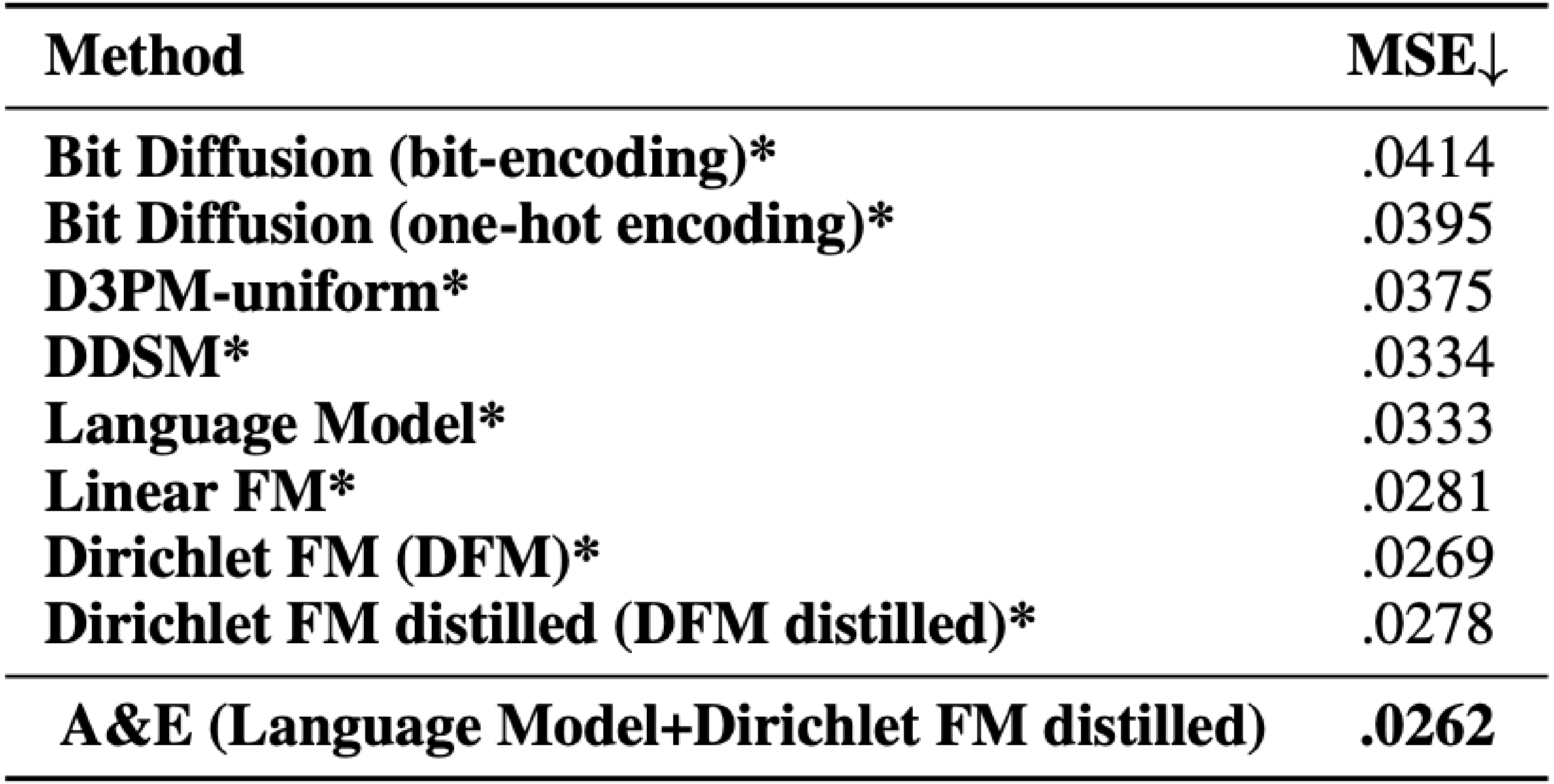


Algorithm 2 Fast Absorb & Escape Algorithm

```
Require: Absorb Threshold $T_{Absorb}$, Pretrained AutoRegressive model $p_\theta^{AR}(\mathbf{x})$ and pretrained Diffusion Model $p_\beta^{DM}(\mathbf{x})$
Initialize $\tilde{\mathbf{x}}^0 \sim p_\beta^{DM}(\mathbf{x})$
for $i$ in $len(\tilde{\mathbf{x}})$ do
    if $p^{DM} < T_{Absorb}$ then
        Absorb step:
        j = i+1
        $\tilde{\mathbf{x}}'_j \sim p_\theta^{AR}(\mathbf{x}_j|\mathbf{x}_{0:i})$
        while $p^{AR}(\tilde{\mathbf{x}}'_j) > p^{DM}(\tilde{\mathbf{x}}_j)$ do
            Increment $j = j + 1$
            $\tilde{\mathbf{x}}'_j \sim p_\theta^{AR}(\mathbf{x}_j|\mathbf{x}_{0:i}, \mathbf{x}_{i:j-1})$ //Refine Inaccurate region of the sequence token by token
        end while
        Escape step:
        $\tilde{\mathbf{x}}_{i:j} = \tilde{\mathbf{x}}'_{i:j}$ //Update $\tilde{\mathbf{x}}$
        Increment i = i + j
    end if
end for
Output: $\tilde{\mathbf{x}}$ with improved quality
```

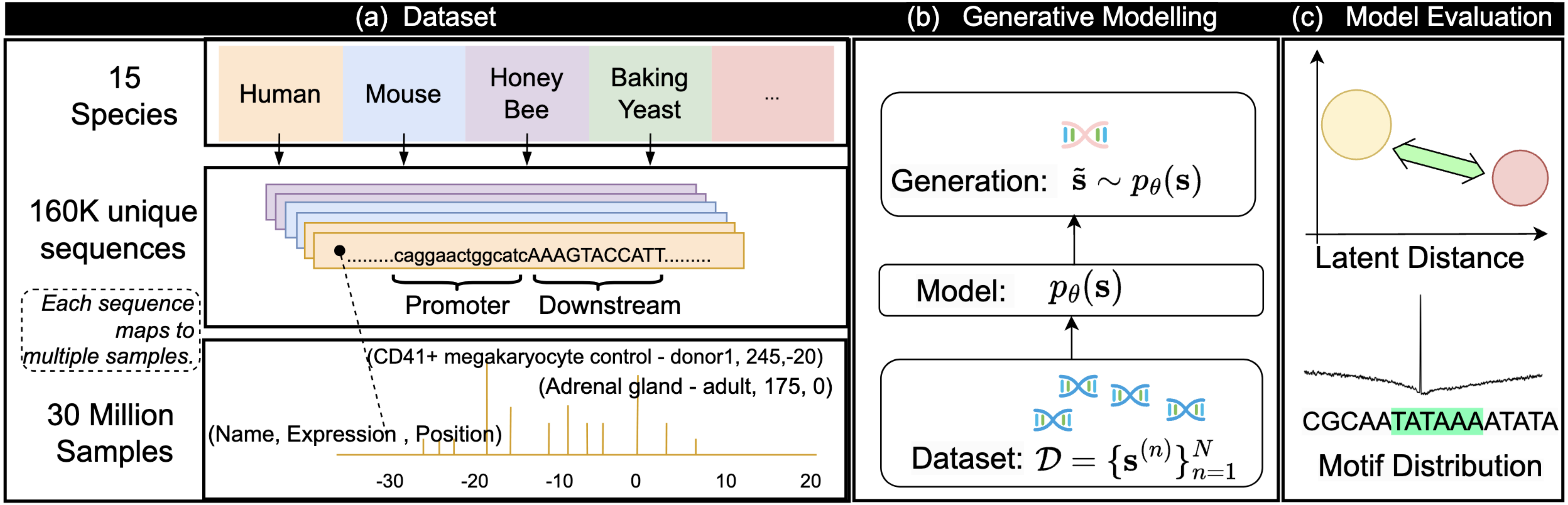

Diffusion Model
$\mathbf{x}_{j+1:L}$
T
T
T
G
Escape Step. Update Subsequence with $\mathbf{x}'_{i:j}$
$\mathbf{x}_{i:j}$
G
G
G
T
T
T
A
A
T
G
C
C
A
A
$\mathbf{x}'_{i:j} \sim P(\mathbf{x}_{i:j}|\mathbf{x}_{0:i-1}, \mathbf{x}_{j+1,L})$
$\approx P^{AR}(\mathbf{x}_{i:j}|\mathbf{x}_{0:i-1})$
P^{AR}
$\mathbf{x}_{0:i-1}$
G
C
G
C
Absorb Step. Use AR Model to refine Homogeneous Subsequence $\mathbf{x}_{i:j}$
Heterogeneous Sequence $\mathbf{x}$
AutoRegressive (AR) Model

Results: transcription profile conditioned promoter sequence design

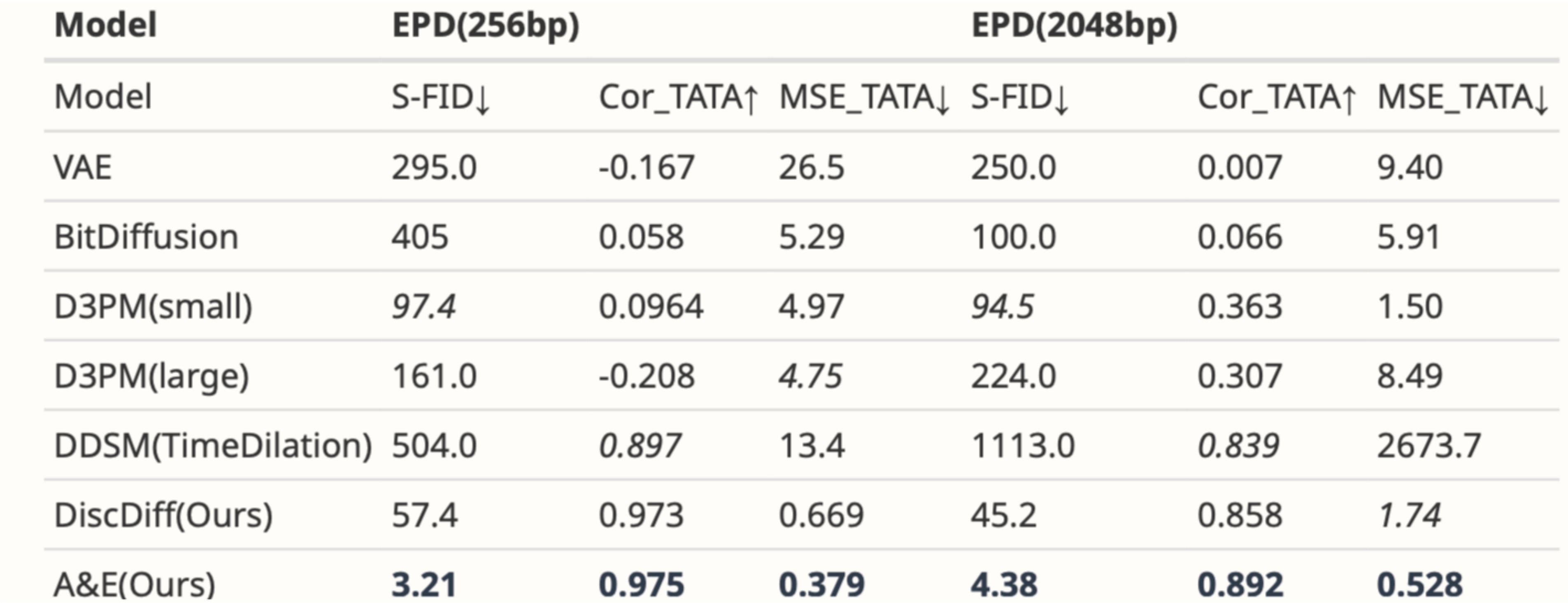


Method	MSE↓
Bit Diffusion (bit-encoding)*	.0414
Bit Diffusion (one-hot encoding)*	.0395
D3PM-uniform*	.0375
DDSM*	.0334
Language Model*	.0333
Linear FM*	.0281
Dirichlet FM (DFM)*	.0269
Dirichlet FM distilled (DFM distilled)*	.0278
A&E (Language Model+Dirichlet FM distilled)	**.0262**

Multi-species Promoter Generation

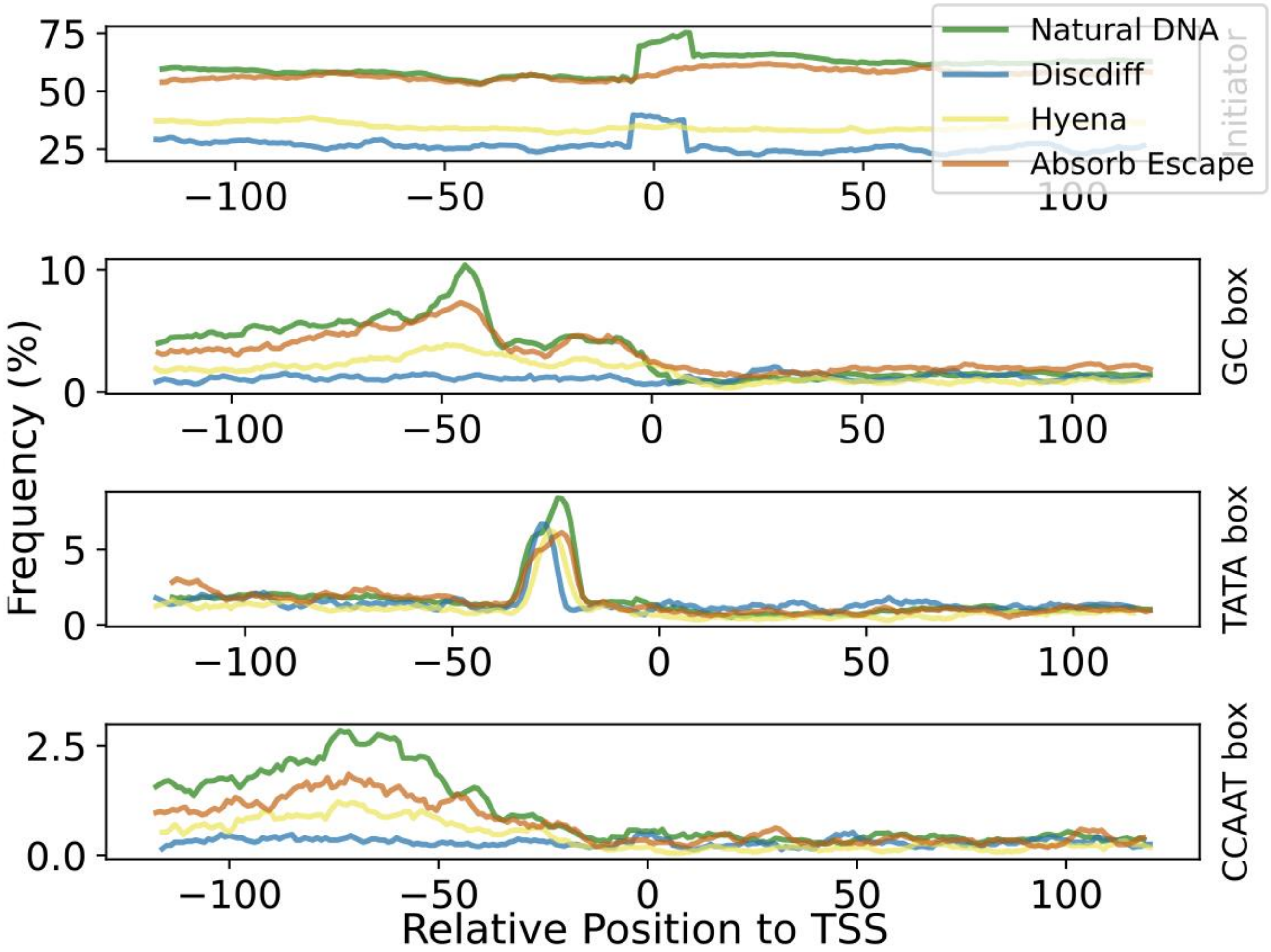

(a) Dataset
(b) Generative Modelling
(c) Model Evaluation
15 Species
Human
Mouse
Honey Bee
Baking Yeast
...
160K unique sequences
.........caggaactggcatcAAAGTACCATT.........
Promoter
Downstream
Each sequence maps to multiple samples.
30 Million Samples
(CD41+ megakaryocyte control - donor1, 245,-20)
(Adrenal gland - adult, 175, 0)
(Name, Expression , Position)
-30
-20
-10
0
10
20
Generation: $\tilde{\mathbf{s}} \sim p_\theta(\mathbf{s})$
Model: $p_\theta(\mathbf{s})$
Dataset: $\mathcal{D} = \{\mathbf{s}^{(n)}\}_{n=1}^{N}$
Latent Distance
CGCAATATAAAATATA
Motif Distribution

Results: Unconditional Generation



Model	**EPD(256bp)**			**EPD(2048bp)**		
Model	S-FID↓	Cor_TATA↑	MSE_TATA↓	S-FID↓	Cor_TATA↑	MSE_TATA↓
VAE	295.0	-0.167	26.5	250.0	0.007	9.40
BitDiffusion	405	0.058	5.29	100.0	0.066	5.91
D3PM(small)	*97.4*	0.0964	4.97	*94.5*	0.363	1.50
D3PM(large)	161.0	-0.208	*4.75*	224.0	0.307	8.49
DDSM(TimeDilation)	504.0	*0.897*	13.4	1113.0	*0.839*	2673.7
DiscDiff(Ours)	57.4	0.973	0.669	45.2	0.858	*1.74*
A&E(Ours)	**3.21**	**0.975**	**0.379**	**4.38**	**0.892**	**0.528**

Results: Species-wise Conditional Generation (Motif Distribution)



Results: Species-wise Conditional Generation (Gene Integration)

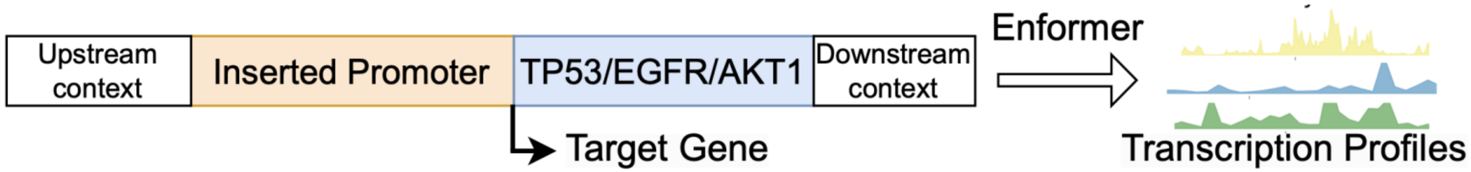
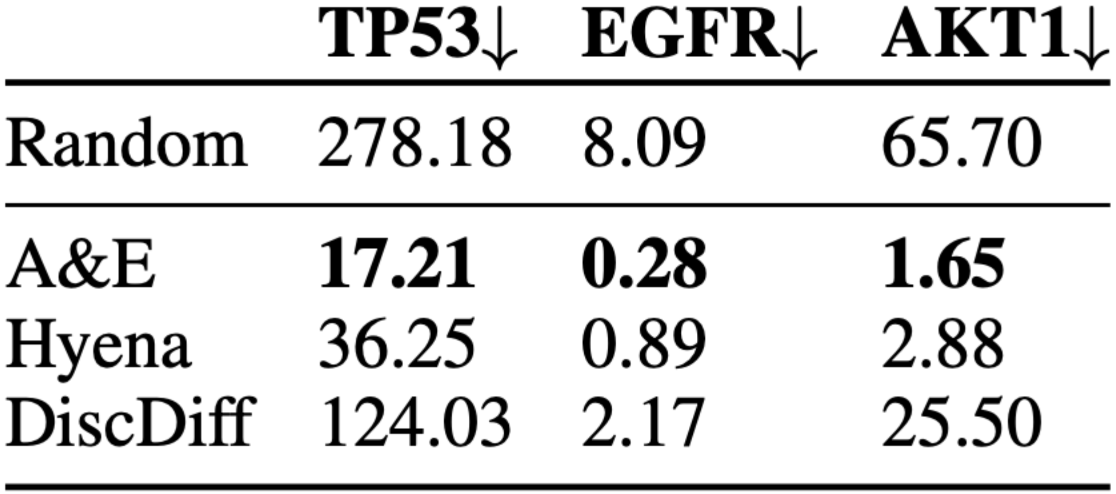


Figure 5: **Evaluation of Generated Promoters for gene regulation through Genome Integration**



	TP53↓	**EGFR↓**	**AKT1↓**
Random	278.18	8.09	65.70
A&E	**17.21**	**0.28**	**1.65**
Hyena	36.25	0.89	2.88
DiscDiff	124.03	2.17	25.50

IMPERIAL

Absorb & Escape: Overcoming Single Model Limitations in Generating Genomic Sequences

Zehui Li[1], Yuhao Ni[1], Guoxuan Xia[1], William Beardall[1], Akashaditya Das[1], Guy-Bart Stan[1], Yiren Zhao[1]

[1]Imperial College London

NEURAL INFORMATION PROCESSING SYSTEMS

1. Motivation

AutoRegressive (AR) Models and Diffusion Models (DMs) both have their limitations.

- **AR Models:** *Sufficient data* is needed for AR model to learn independence in the data
- **DMs:** DMs are less competent than AR models for discrete data generation

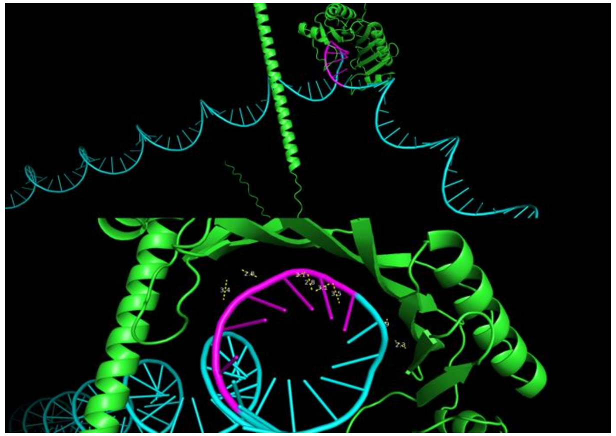
2. Contribution

Our contribution is three-fold:

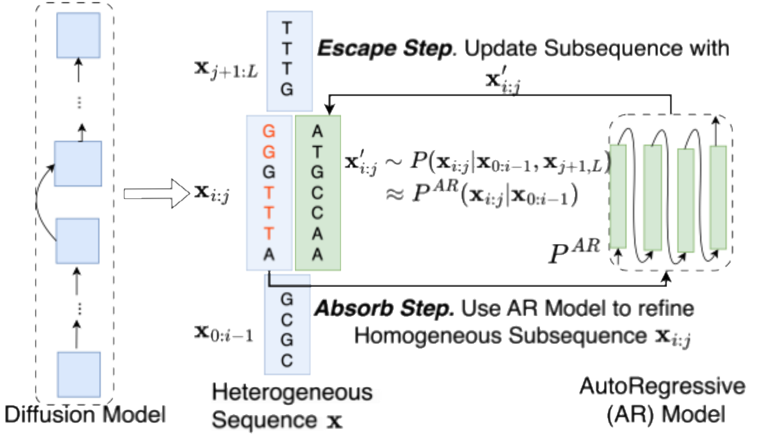
a) Study the properties of AR models and DMs in DNA sequence generation

b) Introduce **Absorb & Escape (A&E)**: a novel approach for DNA generation combining the strengths of AR models and DMs.

c) Demonstrate Fast A&E's superior performance across 15 species.



3. Method


Diffusion Model
$\mathbf{x}_{j+1:L}$
T T T G
$\mathbf{x}_{i:j}$
G G G T T T A
A T G C C A A
$\mathbf{x}_{0:i-1}$
G C G C
Escape Step. Update Subsequence with $\mathbf{x}'_{i:j}$
$\mathbf{x}'_{i:j} \sim P(\mathbf{x}_{i:j}|\mathbf{x}_{0:i-1}, \mathbf{x}_{j+1,L}) \approx P^{AR}(\mathbf{x}_{i:j}|\mathbf{x}_{0:i-1})$
P^{AR}
Absorb Step. Use AR Model to refine Homogeneous Subsequence $\mathbf{x}_{i:j}$
Heterogeneous Sequence $\mathbf{x}$
AutoRegressive (AR) Model

Algorithm 2 Fast Absorb & Escape Algorithm

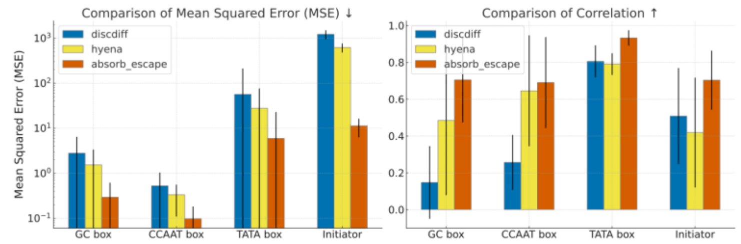
```
Require: Absorb Threshold $T_{Absorb}$, Pretrained AutoRegressive model $p_\theta^{AR}(\mathbf{x})$ and pre-trained Diffusion Model $p_\beta^{DM}(\mathbf{x})$
Initialize $\tilde{\mathbf{x}}^0 \sim p_\beta^{DM}(\mathbf{x})$
for $i$ in $len(\tilde{\mathbf{x}})$ do
    if $p^{DM} < T_{Absorb}$ then
        Absorb step:
        j = i+1
        $\tilde{\mathbf{x}}'_j \sim p_\theta^{AR}(\mathbf{x}_j|\mathbf{x}_{0:i})$
        while $p^{AR}(\tilde{\mathbf{x}}'_j) > p^{DM}(\tilde{\mathbf{x}}_j)$ do
            Increment $j = j + 1$
            $\tilde{\mathbf{x}}'_j \sim p_\theta^{AR}(\mathbf{x}_j|\mathbf{x}_{0:i}, \mathbf{x}_{i:j-1})$ //Refine Inaccurate region of the sequence token by token
        end while
        Escape step:
        $\tilde{\mathbf{x}}_{i:j} = \tilde{\mathbf{x}}'_{i:j}$ //Update $\tilde{\mathbf{x}}$
        Increment i = i + j
    end if
end for
Output: $\tilde{\mathbf{x}}$ with improved quality
```

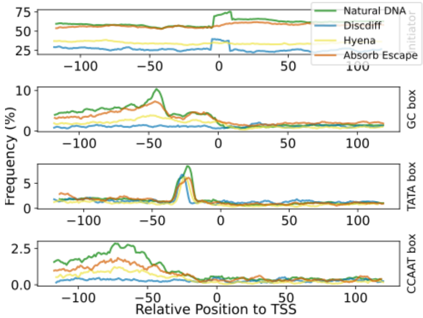
4. Results

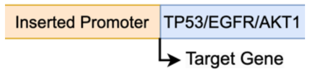
Evaluation of transcription profile conditioned promoter sequence design.

Method	MSE↓
Bit Diffusion (bit-encoding)*	.0414
Bit Diffusion (one-hot encoding)*	.0395
D3PM-uniform*	.0375
DDSM*	.0334
Language Model*	.0333
Linear FM*	.0281
Dirichlet FM (DFM)*	.0269
Dirichlet FM distilled (DFM distilled)*	.0278
A&E (Language Model+Dirichlet FM distilled)	**.0262**

Multi-species Promoter Generation







Sum of Squared Errors (SSE) of Transcription Profiles between Real and Generated Sequences

	TP53↓	EGFR↓	AKT1↓
Random	278.18	8.09	65.70
A&E	**17.21**	**0.28**	**1.65**
Hyena	36.25	0.89	2.88
DiscDiff	124.03	2.17	25.50

References

Avdeyev (2023) Dirichlet diffusion score model for biological sequence generation. In International Conference on Machine Learning (pp. 1276-1301). PMLR.

Stark, Hannes (2024) "Dirichlet flow matching with applications to dna sequence design." In International Conference on Machine Learning

Paper (Arxiv)   Code (Github)