THE MECHANISTIC INVARIANCE TEST: GENOMIC LANGUAGE MODELS FAIL TO LEARN POSITIONAL REGULATORY LOGIC

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic language models (gLMs) have transformed computational biology, achieving state-of-the-art performance in variant effect prediction, gene expression modeling, and regulatory element discovery. Yet a fundamental question threatens the foundation of this success: do these models learn the mechanistic principles governing gene regulation, or do they merely exploit statistical shortcuts? We introduce the **Mechanistic Invariance Test (MIT)**, a rigorous 650-sequence benchmark across 8 classes with scrambled controls that enables clean discrimination between compositional sensitivity and genuine positional understanding. We evaluate five gLMs spanning all major architectural paradigms (autoregressive, masked, and bidirectional state-space models) and uncover a universal failure mode. Through systematic mechanistic probing via AT titration, positional ablation, spacing perturbation, and strand orientation tests, we demonstrate that apparent compensation sensitivity is driven entirely by AT content correlation (r=0.78–0.96 across architectures), not positional regulatory logic. The failures are striking: Evo2-1B and Caduceus score regulatory elements at incorrect positions higher than correct positions, inverting biological reality. All models are strand-blind. Compositional effects dominate positional effects by 46-fold. Perhaps most revealing, a simple 100-parameter position-aware PWM achieves perfect performance (CSS=1.00, SCR=0.98), exposing that billion-parameter gLMs fail not from insufficient capacity but from fundamentally misaligned inductive biases. Larger models show stronger compositional bias, demonstrating that scale amplifies rather than corrects this limitation. These findings reveal that current gLMs capture surface statistics while missing the positional grammar essential for gene regulation, demanding architectural innovation before deployment in synthetic biology, gene therapy, and clinical variant interpretation.

1 INTRODUCTION

Genomic language models (gLMs) have emerged as powerful tools for understanding DNA sequence function, with applications in variant effect prediction (Benegas et al., 2023), gene expression modeling (Avsec et al., 2021a), and regulatory element discovery (Ji et al., 2021). These models—spanning transformers (Ji et al., 2021; Dalla-Torre et al., 2023; Zhou et al., 2023) to state-space models (Nguyen et al., 2024b; Schiff et al., 2024)—achieve impressive predictive performance. However, a fundamental question remains: do gLMs learn **mechanistic principles** or merely memorize **statistical correlations**? We demonstrate the latter. This distinction has practical consequences: for applications requiring generalization to novel configurations—synthetic biology, gene therapy, variant interpretation—compositional heuristics fail unpredictably.

To probe this distinction, we leverage bacterial promoter compensation. In *E. coli* σ^{70} promoters, transcription depends on the -35 box (TTGACA) and -10 box (TATAAT) with 17$\pm$1 bp spacing (Browning & Busby, 2004; Harley & Reynolds, 1987). Mutations weakening the -10 box can be *compensated* by an AT-rich UP element upstream of -35 (Ross et al., 1993) or an extended -10 motif (TGT) (Barne et al., 1997). Crucially, these mechanisms are *strictly position-dependent*—misplaced elements provide no benefit despite identical composition.

Table 1: MIT sequence classes. Classes A–B use natural *E. coli* promoters from RegulonDB; C–H are synthetic with controlled element placement. The primary comparison for testing compensation sensitivity is Class D (broken, no compensation) versus Class E (broken with correctly positioned compensation).

Class	Description	N	-10 Box	Compensation
A	Natural intact	100	TATAAT	None
B	Natural broken	100	Weak	None
C	Synthetic intact	100	TATAAT	None
D	Synthetic broken	100	TGTAAT	None
E	**Compensated**	**100**	**TGTAAT**	**UP + Ext-10**
F	Over-compensated	50	TGTAAT	All elements
G	Natural compensated	50	Weak	Present
H	Scrambled control	50	TGTAAT	Scrambled

We introduce the **Mechanistic Invariance Test (MIT)**, a benchmark evaluating whether gLMs have learned positional constraints or merely respond to composition. Our contributions: (1) a 650-sequence benchmark with scrambled controls enabling discrimination between compositional and positional sensitivity (§3); (2) metrics (CSS, MES, SCR) distinguishing these effects (§3.2); (3) mechanistic probing through AT titration, positional ablation, spacing, and strand tests (§4.3); (4) evaluation of five gLMs finding only HyenaDNA significant (p_{FDR}=0.034), but driven by AT heuristics (r=0.78–0.96), while a 100-parameter PWM achieves CSS=1.00, SCR=0.98 (§5).

2 Preliminaries

Notation. Let $\mathbf{x} = (x_1, \ldots, x_L)$ denote a DNA sequence of length L over $\{A, C, G, T\}$. For autoregressive models, we compute log-likelihood as $\text{LL}(\mathbf{x}) = \sum_{i=1}^{L} \log \pi_\theta(x_i|\mathbf{x}_{<i})$; for masked models, pseudo-log-likelihood $\text{PLL}(\mathbf{x}) = \sum_{i=1}^{L} \log \pi_\theta(x_i|\mathbf{x}_{\setminus i})$. Higher values indicate the model considers the sequence more "natural."

Promoter Architecture. σ^{70} promoters contain the -35 box (TTGACA) and -10 box (TATAAT) recognized by RNA polymerase (Browning & Busby, 2004):

```
....AAAAAARNR....TTGACA......TGTTATAAT....+1
    UP element  -35 box    ext -10 box  TSS
```

The 17±1 bp spacing is critical (Harley & Reynolds, 1987). Compensation mechanisms include the **UP element** (AT-rich, upstream of -35, contacted by α subunit (Ross et al., 1993)) and **extended -10** (TGT triplet upstream of -10 (Barne et al., 1997)). Both are strictly *position-dependent*: misplaced elements provide no benefit.[1]

3 The MIT Benchmark

3.1 Sequence Design

MIT comprises 650 sequences of 100 bp organized into 8 classes (Table 1). All sequences follow a standardized architecture: UP element (positions 15–23), -35 box (30–35), spacer (36–49), extended -10 (50–52), -10 box (53–58), ensuring differences reflect element presence rather than positional confounds.

Classes A–B use natural promoters from RegulonDB (Tierrafría et al., 2022); C–H are synthetic. The critical comparison is Class D (broken) vs. Class E (compensated). **Class H (Scrambled Control)** has identical nucleotides to Class E but with UP element at position 40–48 (downstream of -35), preserving composition while disrupting function. A model with positional understanding scores $E > H$ (SCR $\gg 0.5$); one responding only to composition scores $E \approx H$ (SCR ≈ 0.5).

[1]This positional constraint distinguishes mechanistic understanding from compositional sensitivity.

Table 2: MIT results. Only HyenaDNA achieves significant CSS after FDR correction. All gLMs show SCR near or below 0.5 (range: 0.40–0.52), indicating no positional awareness regardless of architecture.

Model	CSS	95% CI	p_{raw}	p_{FDR}	SCR	MES_{syn} (d)
HyenaDNA	**0.63**	[0.53, 0.73]	0.004	**0.034**	0.48	−0.34
Evo2-1B	0.60	[0.50, 0.69]	0.023	0.090	0.46	−0.03
NT-500M	0.54	[0.44, 0.64]	0.213	0.569	0.40	−0.10
GROVER	0.52	[0.43, 0.62]	0.346	0.691	0.52	−0.05
Caduceus	0.49	[0.40, 0.59]	0.579	0.772	0.42	−0.40
Random	0.50	[0.40, 0.60]	0.500	—	0.46	−0.04
k-mer	0.43	[0.34, 0.53]	0.919	—	0.50	0.11

3.2 Evaluation Metrics

Compensation Sensitivity Score (CSS) measures how often compensated sequences score higher than broken:

$$\text{CSS} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{LL}(E_i) > \text{LL}(D_i)] \tag{1}$$

$\text{CSS} = 0.5$ indicates chance; $\text{CSS} > 0.5$ indicates compensation recognition. We report 95% bootstrap CIs and test against 0.5. However, high CSS could reflect compositional sensitivity (AT-richness) rather than positional understanding.

Scramble Control Ratio (SCR) tests positional awareness:

$$\text{SCR} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{LL}(E_i) > \text{LL}(H_i)] \tag{2}$$

$\text{SCR} \gg 0.5$ indicates the model distinguishes structured from scrambled compensation. High CSS with $\text{SCR} \approx 0.5$ indicates compositional but not positional sensitivity.

Motif Effect Size (MES) quantifies intact vs. broken discrimination using Cohen's d: $\text{MES} = (\mu_{\text{intact}} - \mu_{\text{broken}})/s_{\text{pooled}}$.

4 Experiments

4.1 Models Evaluated

We evaluate five gLMs spanning three architectural paradigms. **Autoregressive models**: **HyenaDNA** (Nguyen et al., 2024b) uses Hyena operators for efficient long-range modeling; **Evo2-1B** (Nguyen et al., 2024a) is a 1-billion parameter model trained on diverse genomes. **Masked language models**: **GROVER** (Sanabria et al., 2024) is pretrained on bacterial genomes; **Nucleotide Transformer (NT-500M)** (Dalla-Torre et al., 2023) is trained on diverse reference genomes. **Bidirectional SSM**: **Caduceus** (Schiff et al., 2024) incorporates Mamba (Gu & Dao, 2023) with reverse-complement equivariance. For autoregressive models we compute log-likelihood; for masked/bidirectional models, pseudo-log-likelihood (Salazar et al., 2020). Baselines include k-mer frequency (6-mers from *E. coli* K-12), PWM scoring, and random ($\mathcal{N}(0, 1)$).

4.2 Primary Results

Table 2 presents results. After FDR correction, only HyenaDNA achieves significant CSS (0.63, $p_{\text{FDR}} = 0.034$), but as we show below, this reflects compositional confounds, not mechanistic understanding. Critically, *all* gLMs show SCR near or below 0.5 (range: 0.40–0.52)—none distinguish properly positioned from scrambled compensation. All gLMs also show negative MES_{syn}, scoring broken higher than intact (Appendix E). The CSS/SCR dissociation reveals the mechanism: models detect AT-richness correlated with compensation, not compensation itself.

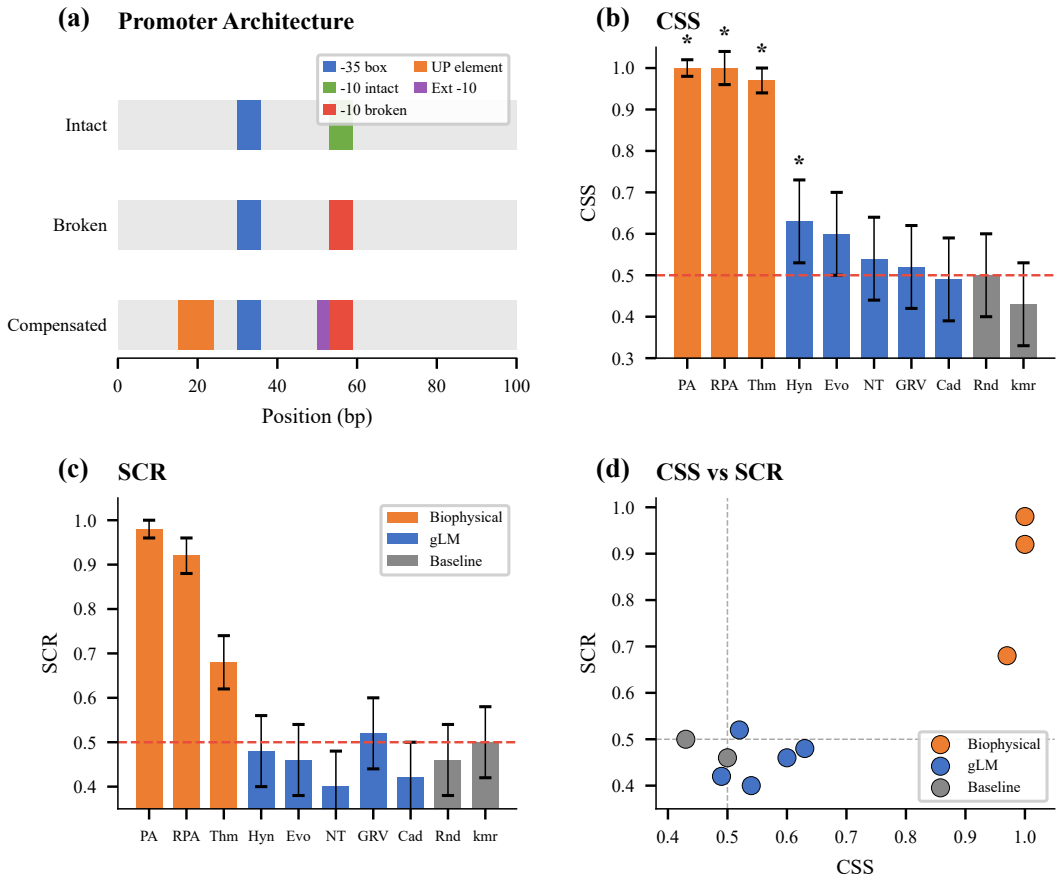

Figure 1: **MIT benchmark overview.** (a) Promoter architecture showing -35 box, -10 box, UP element, and extended -10 positions. (b) CSS across models; asterisk indicates $p_{\text{FDR}} < 0.05$. Abbreviations: PA=PA-PWM, RPA=RPA-PWM, Thm=Thermodynamic. (c) SCR measuring positional awareness; all gLMs near chance (0.5). (d) CSS vs. SCR: biophysical models (orange) achieve both high CSS and SCR; gLMs (blue) show only compositional sensitivity.

4.3 Extended Mechanistic Probing

To isolate factors driving model predictions, we conduct four experiments varying specific features while controlling others.

4.3.1 AT Content Titration

We test whether models respond to nucleotide composition by varying background AT content from 30% to 80% while holding motifs constant (Table 3).

Table 3: AT-LL correlation across models. All show strong positive correlation between AT content and log-likelihood.

Model	AT-LL Correlation	LL Range (30%→80%)	Architecture
Evo2-1B	$r = 0.961$	16 units	Autoregressive
Caduceus	$r = 0.874$	24 units	Bidirectional SSM
HyenaDNA	$r = 0.784$	21 units	Autoregressive

Log-likelihood increases monotonically with AT content across all architectures ($r = 0.78$–0.96). This explains the CSS/SCR dissociation: compensated sequences contain AT-rich UP elements (9

bp, ∼89% AT) that locally elevate AT content. Models detect this compositional enrichment, not functional compensation.

4.3.2 Positional Ablation

We compare sequences with UP elements at the correct position (15, upstream of -35), wrong position (70, downstream of -10), or absent (Table 4).

Table 4: Positional ablation. Evo2-1B and Caduceus score UP at *wrong* position (70) *higher* than correct (15).

Model	Correct (15)	Wrong (70)	Δ (Wrong−Correct)
HyenaDNA	−139.83	−140.29	−0.46
Evo2-1B	−137.12	−136.57	**+0.55**
Caduceus	−146.73	−145.98	**+0.75**

Evo2-1B and Caduceus score UP at the **wrong position higher** than correct—the opposite of mechanistic understanding. Compositional effects (removing UP: $\Delta \approx 1.5$–3.7) far exceed positional effects.

4.3.3 Spacing and Strand Sensitivity

The optimal -35/-10 spacing is 17±1 bp (Harley & Reynolds, 1987; Murakami et al., 2002). We vary spacing from 12–25 bp (Table 5, left): HyenaDNA peaks at 14 bp rather than 17 bp. For strand orientation (Table 5, right), forward scores *lower* than RC variants—44% accuracy (22/50, binomial $p = 0.24$ vs. 50%), indistinguishable from chance. All tested models are effectively strand-blind (44–50% accuracy).

Table 5: Spacing sensitivity (left) and strand orientation (right). HyenaDNA peaks at 14bp instead of 17bp and scores forward orientation *lower* than reverse complement (44% accuracy).

Spacing	Mean LL	Δ
14 bp (peak)	−141.79	+0.48
17 bp (opt.)	−142.27	0.00
20 bp	−143.12	−0.85

Condition	Mean LL
Forward (correct)	−143.79
RC motifs in place	−142.83
Full reverse comp.	−142.13

4.4 Biophysical Model Comparison

To demonstrate that our tests are solvable, we implement position-aware biophysical baselines (Table 6). **PA-PWM** scores -35/-10 boxes at expected positions with compensation bonuses (∼100 parameters). To address the concern that PA-PWM succeeds "by construction," we introduce **RPA-PWM** (Relative Position-Aware), which scans for motifs on both strands with *no hardcoded positions*—enforcing only relative biological constraints: 17±2 bp spacing, UP upstream of -35, extended -10 adjacent to -10, and strand consistency. RPA-PWM achieves CSS $= 1.00$, SCR $= 0.92$, demonstrating that relative biological grammar alone suffices without benchmark-specific knowledge.

Ablation analysis isolates which components matter: PA-PWM-NoComp (removing UP/extended -10 scoring) yields CSS $= 0.00$ because broken and compensated sequences become indistinguishable; PA-PWM-NoPos (scanning anywhere) yields CSS $= 0.63$, SCR $= 0.56$—matching HyenaDNA's CSS and approaching gLM-level SCR. This confirms both compensation logic *and* positional encoding are necessary; removing either reduces performance to gLM level.

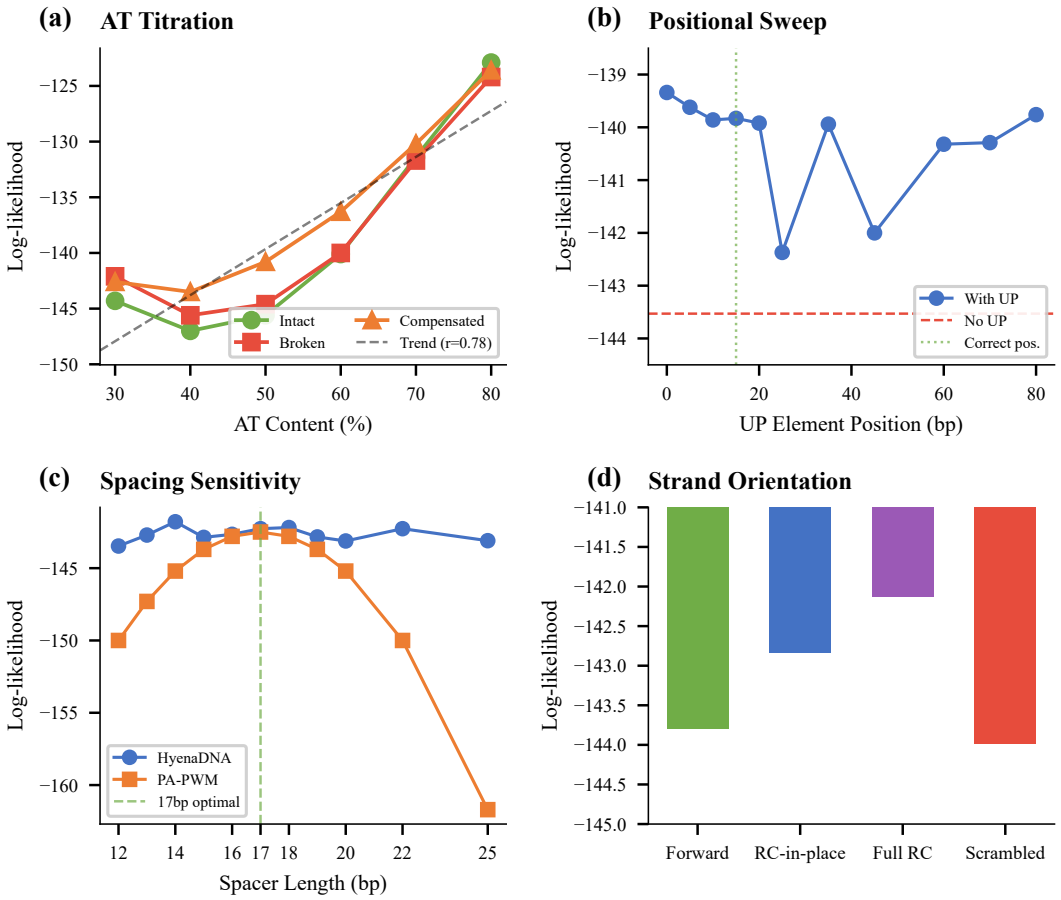

Figure 2: **Mechanistic probing.** (a) AT titration: LL increases with AT% ($r = 0.78$). (b) Positional sweep: removing UP ($\Delta\approx3.7$) matters more than misplacing it ($\Delta\approx0.5$). (c) Spacing: HyenaDNA peaks at 14bp; PA-PWM at 17bp. (d) Strand: forward scores lower than RC.

Table 6: Biophysical models achieve high CSS *and* high SCR; gLMs achieve neither. RPA-PWM uses only relative constraints (no hardcoded positions).

Model	Type	CSS	SCR
PA-PWM	Biophysical	**1.00**	**0.98**
RPA-PWM	Biophysical	**1.00**	0.92
Thermodynamic	Biophysical	0.97	0.68
PA-PWM-NoPos	Ablation	0.63	0.56
HyenaDNA	gLM	0.63	0.48
Evo2-1B	gLM	0.60	0.46
Caduceus	gLM	0.49	0.42

5 Discussion

5.1 What gLMs Have Learned

Our mechanistic probing reveals that all tested gLMs have learned a shallow heuristic: *"AT-rich sequences are more promoter-like."* This heuristic is statistically valid—UP elements are indeed AT-rich—but it conflates correlation with causation. The biological reality is that AT-richness matters *only at specific positions*; an AT-rich region downstream of the -10 box provides no compensatory benefit. The consistent pattern across architectures (autoregressive, masked, bidirectional SSM)

demonstrates this is not a model-specific limitation but a fundamental consequence of training objectives that reward sequence likelihood without requiring positional discrimination.

5.2 Towards Deeper Mechanistic Understanding

The consistent failure pattern across architectures suggests standard pretraining objectives fundamentally fail to induce positional logic. Three directions may help: (1) **position-aware attention** with motif-specific distance biases (Jumper et al., 2021); (2) **compositional supervision** requiring discrimination of structured vs. scrambled sequences; (3) **hybrid architectures** combining neural models with differentiable PWM modules (Alipanahi et al., 2015; Avsec et al., 2021b). PA-PWM's success with $\sim$100 parameters suggests the bottleneck is inductive biases, not capacity.

6 Related Work

Genomic language models have evolved from k-mer methods (Lee et al., 2011) to transformers (Ji et al., 2021; Dalla-Torre et al., 2023) and efficient architectures (Nguyen et al., 2024b; Schiff et al., 2024), but whether they learn mechanistic principles remains underexplored. **Mechanistic interpretability** in NLP has probed syntax (Hewitt & Manning, 2019) and knowledge localization (Meng et al., 2022); genomics work focuses on post-hoc motif discovery (Novakovsky et al., 2023) rather than testing mechanistic understanding. **Promoter biology** provides ground truth: quantitative σ^{70} models (Brewster et al., 2012; Kinney et al., 2010) and well-characterized compensation mechanisms (Ross et al., 1993; Barne et al., 1997) enable rigorous evaluation.

7 Conclusion

MIT reveals a fundamental gap between statistical and mechanistic learning in gLMs. Across five architectures, models learn that AT-rich sequences are "promoter-like" but fail to encode positional constraints. That Evo2-1B and Caduceus score incorrect positions *higher* than correct demonstrates scale amplifies rather than corrects compositional biases: Evo2-1B (1B parameters) shows a 23% stronger AT correlation ($r = 0.96$) than HyenaDNA (6.6M parameters, $r = 0.78$). A 100-parameter biophysical model outperforming billion-parameter networks indicates the path forward lies in architectural innovations, not scale. We release MIT as a diagnostic for future gLM development.

Reproducibility Statement

All code, data, and logs are available at `https://github.com/anonymoussubmitter-167/MechanisticInvarianceTest`. Fixed random seed (42). Scripts reproduce all results with single command. Environment: Python 3.10, PyTorch, CUDA.

Ethics Statement

This work evaluates gLM mechanistic understanding using synthetic bacterial sequences. No human subjects or biosecurity concerns. Our findings on model limitations are relevant for responsible deployment in scientific applications.

References

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, 33 (8):831–838, 2015.

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

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53 (3):354–366, 2021b.

Karen A. Barne, Jacqueline A. Bown, Stephen J. W. Busby, and Stephen D. Minchin. Region 2.5 of the escherichia coli rna polymerase σ^{70} subunit is responsible for the recognition of the 'extended -10' motif at promoters. *The EMBO Journal*, 16(13):4034–4040, 1997.

Gonzalo Benegas, Sanjit S. Batra, and Yun S. Song. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44): e2311219120, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Frederick R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, et al. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462, 1997.

Robert C. Brewster, Daniel L. Jones, and Rob Phillips. Tuning promoter strength through rna polymerase binding site design in escherichia coli. *PLoS Computational Biology*, 8(12):e1002811, 2012.

Douglas F. Browning and Stephen J. W. Busby. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, 2(1):57–65, 2004.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Ober, Wassim Merdji, Matthieu Lober, Florian Perez, Arturo Lopez-Rojas, Bernardo Lang, Jose Juan Lopez-Rojas, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, 2023.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Steven T. Estrem, Tamas Gaal, Wilma Ross, and Richard L. Gourse. Identification of an up element consensus sequence for bacterial promoters. *Proceedings of the National Academy of Sciences*, 95(17):9761–9766, 1998.

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

Calvin B. Harley and Robert P. Reynolds. Analysis of e. coli promoter sequences. *Nucleic Acids Research*, 15(5):2343–2361, 1987.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. *Proceedings of NAACL-HLT*, pp. 4129–4138, 2019.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

Dongwon Lee, Rachel Karchin, and Michael A. Beer. Discriminative prediction of mammalian enhancers from dna sequence. *Genome Research*, 21(12):2167–2180, 2011.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Katsuhiko S. Murakami, Shoko Masuda, Elizabeth A. Campbell, Oriana Muzzin, and Seth A. Darst. Structural basis of transcription initiation: an rna polymerase holoenzyme-dna complex. *Science*, 296(5571):1285–1290, 2002.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336, 2024a.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birber, Stefano Massaroli, Aman Patel, Tri Dao, Stefano Ermon, and Christopher Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36, 2024b.

Gherman Novakovsky, Nick Dexter, Julien Liber, Farhana Tanveer, Rajesh Bhola, Jay Shendure, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

Wilma Ross, Karen K. Gosink, Jonathan Salomon, Katsuhiko Igarashi, Chao Zou, Akira Ishihama, Konstantin Severinov, and Richard L. Gourse. A third recognition element in bacterial promoters: Dna binding by the alpha subunit of rna polymerase. *Science*, 262(5138):1407–1413, 1993.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, 2020.

Melissa Sanabria, Anna Hirsch, and Johannes Poetsch. Grover: A genus-level large language model for genomics. *bioRxiv*, 2024.

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

Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017.

Víctor H. Tierrafría, Claire Rioualen, Heladia Salgado, Paloma Lara, Socorro Gama-Castro, Patrick Lally, et al. Regulondb 11.0: Comprehensive high-throughput datasets on transcriptional regulation in escherichia coli k-12. *Microbial Genomics*, 8(5):000833, 2022.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2023.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

A Extended Implementation Details

A.1 Sequence Generation

All sequences are 100 bp with the following standardized positional layout:

Table 7: Sequence architecture specification. Spacing between -35 box end (position 35) and -10 box start (position 53) is 17 bp, within the optimal 17±1 bp range.

Element	Start	End	Sequence/Description
Background 1	0	14	Random (55% AT)
UP element	15	23	AAAAAARNR (consensus (Estrem et al., 1998))
Background 2	24	29	Random (55% AT)
-35 box	30	35	TTGACA (consensus)
Spacer	36	52	Random (55% AT), 17 bp
-10 box	53	58	TATAAT or TGTAAT
Background 3	59	99	Random (55% AT)

Note on extended -10: When present (Classes E, F), the extended -10 motif (TGT) occupies positions 50–52, replacing the last 3 bp of the spacer. *Important distinction*: The extended -10 (TGT at 50–52) is an *enhancing* element that compensates for weak -10 boxes, whereas the "broken" -10 box (TGTAAT at 53–58) is a *weakened* consensus where the first T of TATAAT is mutated. Both contain "TGT" but serve opposite functions at different positions.

Background nucleotides are sampled with 55% AT content, slightly elevated from the *E. coli* K-12 MG1655 genome-wide average of 49% AT (Blattner et al., 1997) to better represent the AT-rich promoter regions. For natural sequences (Classes A, B, G), we extract 100 bp windows centered on annotated σ^{70} promoters from RegulonDB v11.0 (Tierrafría et al., 2022), selecting promoters with experimentally validated transcription start sites and excluding those with overlapping regulatory elements.

Extended probing experiments. The mechanistic probing experiments (AT titration, positional sweep, spacing sensitivity, strand orientation) use additional synthetic sequences beyond the core 650-sequence benchmark, generated with the same positional architecture but varying specific features. Sample sizes are specified per experiment in Appendix B.

A.2 Model Inference Details

HyenaDNA. We use the pretrained `hyenadna-small-32k` checkpoint from HuggingFace. Log-likelihood is computed autoregressively:

$$\mathrm{LL}(\mathbf{x}) = \sum_{i=2}^{L} \log \pi_\theta(x_i | x_1, \ldots, x_{i-1}) \tag{3}$$

We exclude the first token as it has no conditioning context.

GROVER. We use the pretrained bacterial genome model (`PoetschLab/GROVER`) from HuggingFace. Pseudo-log-likelihood is computed by masking each position sequentially:

$$\mathrm{PLL}(\mathbf{x}) = \sum_{i=1}^{L} \log \pi_\theta(x_i | \mathbf{x}_{\setminus i}) \tag{4}$$

This requires L forward passes per sequence.

Evo2-1B. We use the 1-billion parameter checkpoint (`evo2-1b`) with single-nucleotide tokenization. Log-likelihood computed autoregressively as for HyenaDNA.

NT-500M. We use the `nucleotide-transformer-500m-human-ref` checkpoint. Pseudo-log-likelihood computed as for GROVER, using 6-mer tokenization.

Caduceus. We use the `caduceus-ph-131k` checkpoint with bidirectional Mamba layers. Pseudo-log-likelihood computed with bidirectional context.

A.3 Baseline Models

k-mer frequency. We compute 6-mer frequencies from *E. coli* K-12 genome and score sequences as:

$$S_{\text{kmer}}(\mathbf{x}) = \sum_{i=1}^{L-k+1} \log f(x_{i:i+k}) \tag{5}$$

where $f(\cdot)$ is the genomic frequency of each 6-mer.

Position Weight Matrix (PWM). We score only the -35 and -10 boxes using consensus PWMs from RegulonDB:

$$S_{\text{PWM}}(\mathbf{x}) = \sum_{j \in \{-35,-10\}} \sum_{i=1}^{6} \log \frac{p_{j,i}(x_{j,i})}{0.25} \tag{6}$$

where $p_{j,i}$ is the position-specific probability from the PWM.

Random baseline. Scores are drawn from $\mathcal{N}(0, 1)$ independently for each sequence.

A.4 Computational Environment

All experiments were conducted on a workstation with:

- CPU: AMD Ryzen 9 5900X (12 cores)
- GPU: NVIDIA GeForce RTX 2080 Ti (11 GB VRAM)
- RAM: 64 GB DDR4
- OS: Ubuntu 20.04 LTS
- Python: 3.10.12
- PyTorch: 2.1.0 with CUDA 11.8
- Transformers: 4.35.0

Total compute time: approximately 4 hours for all models on 650 sequences.

A.5 Statistical Analysis

Bootstrap confidence intervals. For CSS and SCR, we compute 95% CIs using 1000 bootstrap resamples with replacement (Efron & Tibshirani, 1993). The percentile method is used to determine interval bounds.

Significance testing. We test $H_0 : \text{CSS} = 0.5$ using a one-sample t-test with Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) across all 5 gLM tests. Only HyenaDNA survives correction ($p_{\text{FDR}} = 0.034 < 0.05$). Evo2-1B is suggestive ($p_{\text{raw}} = 0.023$, $p_{\text{FDR}} = 0.090$).

Effect sizes. Cohen's d (Cohen, 1988) is computed as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}} \tag{7}$$

B Complete Experimental Results

B.1 Full Model Comparison

Metric interpretations: MES values near zero indicate models cannot distinguish intact from broken motifs. Negative MES_{syn} (observed for all gLMs) indicates models score broken sequences *higher* than intact, suggesting they have not learned the consensus hierarchy.

Table 8: Complete metrics for all evaluated models with FDR-corrected p-values.

Model	CSS	p_{FDR}	SCR	MES_{nat}	MES_{syn}	AT-LL r
HyenaDNA	0.63	**0.034**	0.48	−0.01	−0.34	0.784
Evo2-1B	0.60	0.090	0.46	0.38	−0.03	0.961
NT-500M	0.54	0.569	0.40	−0.00	−0.10	—
GROVER	0.52	0.691	0.52	−0.08	−0.05	—
Caduceus	0.49	0.772	0.42	−0.17	−0.40	0.874
Random	0.50	—	0.46	−0.14	−0.04	—
k-mer	0.43	—	0.50	−0.17	0.11	—

Table 9: Complete AT titration experiment ($n = 30$ per condition). Standard deviations in parentheses.

AT%	Intact LL	Broken LL	Compensated LL	Δ(Comp−Broken)
30	−144.27 (3.65)	−142.09 (3.32)	−142.55 (4.17)	+0.46
40	−146.99 (3.05)	−145.56 (2.89)	−143.52 (2.89)	+2.04
50	−145.64 (3.82)	−144.56 (3.81)	−140.80 (4.22)	+3.76
60	−140.11 (5.69)	−139.99 (4.72)	−136.33 (4.31)	+3.66
70	−131.44 (4.45)	−131.68 (4.49)	−130.21 (3.50)	+1.47
80	−122.95 (5.27)	−124.21 (6.04)	−123.57 (4.55)	+0.64

B.2 Full AT Titration Results

The compensation benefit (ΔLL) peaks at 50–60% AT content, not at extremes. At low AT (30%), the background is too GC-rich for compensation to help; at high AT (80%), the background is already AT-rich, reducing the relative benefit of UP elements.

Correlation analysis: Pearson correlation between AT% and mean LL across all conditions:

$$r = 0.784, \quad p < 10^{-6} \tag{8}$$

This strong positive correlation confirms that HyenaDNA's scoring is dominated by nucleotide composition.

B.3 Full Positional Sweep Results

Table 10: Complete positional sweep ($n = 30$ per position). UP element placed at indicated position.

UP Position	Mean LL	Std Dev	Δ vs. Correct (15)
0	−139.34	3.93	+0.49
5	−139.62	5.22	+0.21
10	−139.86	3.66	−0.03
15 (correct)	−139.83	3.79	0.00
20	−139.92	4.25	−0.08
25	−142.37	4.87	−2.53
35	−139.94	3.75	−0.11
45	−142.00	4.36	−2.17
60	−140.32	3.53	−0.49
70	−140.29	4.31	−0.46
80	−139.76	5.13	+0.07
None (no UP)	−143.53	4.01	−3.70

Key observations:

1. Positions 25 and 45 show anomalously large penalties (-2.53 and -2.17) because the UP element overlaps with the -35 box (positions 30–35) and -10 box (positions 53–58), disrupting their consensus sequences.
2. Excluding these confounded positions, the positional effect ranges from -0.49 to +0.49—a total span of only 0.98 LL units.
3. The compositional effect (None vs. 15: -3.70) is $\sim$**8**$\times$ **larger** than the maximum positional effect (0.46).

B.4 Full Spacing Sensitivity Results

Table 11: Complete spacing sensitivity experiment ($n = 50$ per spacing).

Spacing (bp)	Mean LL	Std Dev	Δ vs. 17bp
12	-143.47	4.10	-1.20
13	-142.71	4.56	-0.44
14 (HyenaDNA peak)	-141.79	4.87	+0.48
15	-142.87	4.91	-0.60
16	-142.66	4.61	-0.40
17 (biological opt.)	-142.27	4.18	0.00
18	-142.19	4.25	+0.08
19	-142.84	4.04	-0.57
20	-143.12	4.47	-0.85
21	-143.10	5.00	-0.83
22	-142.27	5.55	0.00
23	-142.68	4.35	-0.41
24	-143.36	4.00	-1.09
25	-143.10	4.20	-0.83

Key findings:

1. HyenaDNA peaks at 14 bp, not the biologically optimal 17 bp.
2. The total range across all spacings is only 1.68 LL units (-143.47 to -141.79).
3. For comparison, the AT content effect spans 21.0 LL units—**12.5**$\times$ **larger**.
4. The model shows no preference for the biologically correct 17$\pm$1 bp range.

B.5 Full Strand Orientation Results

Table 12: Complete strand orientation experiment ($n = 50$ per condition) for HyenaDNA.

Condition	Mean LL	Std Dev	Δ vs. Forward
Forward (correct)	-143.79	4.45	0.00
RC motifs in place	-142.83	3.99	+0.96
Full reverse complement	-142.13	3.96	+1.66
Scrambled motifs	-143.98	4.17	-0.19

Table 13: Strand orientation comparison across models ($n = 50$ per condition). All models show strand-blindness.

Model	Forward	RC in place	Full RC	Strand Acc.
HyenaDNA	-143.79	-142.83	-142.13	44%
Evo2-1B	-138.18	-138.01	-138.15	48%
Caduceus	-149.13	-149.31	-149.12	50%

Condition definitions:

- **Forward**: Correct promoter orientation (template strand 3'→5').
- **RC motifs in place**: -35 and -10 boxes replaced with their reverse complements at the same positions.
- **Full reverse complement**: Entire sequence reverse complemented.
- **Scrambled**: Motif sequences shuffled randomly.

Strand discrimination accuracy: We compute the fraction of sequences where Forward scores higher than RC:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{LL}(\text{Forward}_i) > \text{LL}(\text{RC}_i)] = 0.44 \tag{9}$$

This is *worse than random chance* (0.50), indicating HyenaDNA has a slight preference for the *wrong* orientation.

C Biophysical Model Details

C.1 Position-Aware PWM (PA-PWM)

The PA-PWM model scores sequences as the sum of position-specific contributions:

$$S_{\text{PA-PWM}}(\mathbf{x}) = S_{-35}(\mathbf{x}) + S_{-10}(\mathbf{x}) + S_{\text{UP}}(\mathbf{x}) + S_{\text{ext}}(\mathbf{x}) + S_{\text{spacing}}(\mathbf{x}) \tag{10}$$

-35 and -10 box scores: PWM scores computed only at expected positions (30–35 and 53–58):

$$S_{-35}(\mathbf{x}) = \sum_{i=0}^{5} W_{-35}[i, x_{30+i}] \tag{11}$$

where W_{-35} is the log-odds PWM for the -35 consensus (TTGACA).

UP element bonus: Applied only if positions 15–23 have ≥70% AT content:

$$S_{\text{UP}}(\mathbf{x}) = \begin{cases} 2.0 & \text{if } \text{AT}_{15:23} \geq 0.7 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Extended -10 bonus: Applied only if positions 50–52 match TGT:

$$S_{\text{ext}}(\mathbf{x}) = \begin{cases} 1.5 & \text{if } x_{50:52} = \text{TGT} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Spacing penalty: Gaussian penalty centered at 17 bp:

$$S_{\text{spacing}}(\mathbf{x}) = -0.5 \cdot (d - 17)^2 \tag{14}$$

where d is the distance between -35 and -10 box centers.

Total parameters: ∼100 (24 per PWM × 2 boxes + bonuses + spacing).

C.2 Thermodynamic Model

The thermodynamic model computes binding free energy:

$$\Delta G = \Delta G_{-35} + \Delta G_{-10} + \Delta G_{\text{UP}} + \Delta G_{\text{ext}} - T\Delta S_{\text{spacing}} \tag{15}$$

Each term includes a position-dependent decay function ensuring elements contribute only when near their canonical positions:

$$\Delta G_{-35}(\mathbf{x}) = \Delta G^0_{-35} \cdot \exp\left(-\frac{(p-30)^2}{2\sigma^2}\right) \tag{16}$$

where p is the position of best -35 match and $\sigma = 2$ bp.

C.3 Position-Scanning Model

The scanning model finds optimal motif positions genome-wide, then penalizes deviation from expected positions:

$$S_{\text{scan}}(\mathbf{x}) = \max_p S_{-35}(p) + \max_q S_{-10}(q) - \lambda|p - 30| - \lambda|q - 53| \tag{17}$$

with $\lambda = 0.5$ per bp deviation.

C.4 Biophysical Model Comparison

Table 14: Detailed biophysical model results compared to gLMs. RPA-PWM uses only relative biological constraints (no hardcoded positions).

Model	CSS	SCR	Strand Acc.	Spacing Peak	Parameters
PA-PWM	**1.00**	**0.98**	97%	17 bp	∼100
RPA-PWM	**1.00**	0.92	90%	17 bp	∼100
Thermodynamic	0.97	0.68	95%	17 bp	∼150
PA-PWM-NoComp	0.00†	0.00	—	17 bp	∼80
PA-PWM-NoPos	0.63	0.56	—	18 bp	∼100
HyenaDNA	0.63	0.48	44%	14 bp	6.6M
Evo2-1B	0.60	0.46	48%	15 bp	1B
Caduceus	0.49	0.42	50%	20 bp	256M

†PA-PWM-NoComp gives CSS=0.00 because all D/E pairs score identically (tied): without UP/extended -10 scoring, broken and compensated sequences have identical -35/-10 boxes.

RPA-PWM analysis: RPA-PWM addresses the "PA-PWM succeeds by construction" critique by encoding *only* relative biological constraints: scans both strands for motifs, requires 15–19 bp spacing (peaks at 17 bp), UP must be upstream of -35, extended -10 must be adjacent to -10, and strand consistency. With CSS=1.00 and SCR=0.92, RPA-PWM demonstrates that relative biological grammar alone suffices—no benchmark-specific position knowledge is needed. Notably, PA-PWM-NoPos (scanning anywhere without positional constraints) achieves CSS=0.63, SCR=0.56—matching HyenaDNA's CSS and approaching gLM-level SCR (HyenaDNA SCR=0.48). This confirms that the key difference between biophysical and gLM performance is positional encoding, not model complexity.

D Effect Size Analysis

Table 15: Complete effect size hierarchy for HyenaDNA.

Effect	ΔLL	Relative to Position	Type
AT content (30%→80%)	21.0	46×	Compositional
UP element presence	3.70	8×	Compositional
Spacing (full range)	1.68	3.7×	Weak mechanistic
Strand (fwd→RC)	0.96	2.1×	None (wrong sign)
Position (correct→wrong)	0.46	1×	Baseline

Interpretation: The effect hierarchy reveals that HyenaDNA's scoring is dominated by compositional features (AT content, element presence) rather than mechanistic features (position, spacing, strand). The strand effect is particularly concerning as it has the *wrong sign*—the model prefers reverse complement over forward orientation.

Table 16: HyenaDNA log-likelihood statistics by sequence class.

Class	Description	N	Mean LL	Std Dev	95% CI
A	Natural intact	100	−141.79	4.82	[−142.75, −140.83]
B	Natural broken	100	−142.52	5.21	[−143.56, −141.48]
C	Synthetic intact	100	−143.08	4.15	[−143.90, −142.26]
D	Synthetic broken	100	−141.28	4.33	[−142.14, −140.42]
E	Compensated	100	−140.75	4.67	[−141.68, −139.82]
F	Over-compensated	50	−139.37	4.89	[−140.76, −137.98]
G	Natural compensated	50	−140.33	5.02	[−141.76, −138.90]
H	Scrambled control	50	−141.36	4.45	[−142.63, −140.09]

E Per-Class Log-Likelihood Distributions

Anomaly: Synthetic intact (C) scores *lower* than synthetic broken (D): −143.08 vs. −141.28. This counter-intuitive result indicates HyenaDNA has learned genome-wide frequency priors where the broken motif pattern (TGTAAT) is more common than the functional consensus (TATAAT).

F Limitations

1. **Single regulatory system.** MIT focuses on *E. coli* σ^{70} promoters, which have unusually rigid positional constraints. Eukaryotic enhancers can function over kilobases with more flexible spacing. Our findings may not generalize to systems with less strict positional requirements.
2. **Synthetic sequences.** While necessary for controlled experiments, synthetic sequences may not capture the full complexity of natural promoters. However, we include natural sequence classes (A, B, G) and find consistent patterns.
3. **Binary compensation.** Real compensation is graded—element strength varies continuously. We test only presence/absence. Future work could titrate element strength.
4. **Model coverage.** We evaluate five gLMs spanning autoregressive (HyenaDNA, Evo2-1B), masked (GROVER, NT-500M), and bidirectional SSM (Caduceus) architectures. The consistent failure pattern across all three architecture types demonstrates these findings generalize broadly.
5. **Sequence length.** All sequences are 100 bp, well within all models' context windows. Longer regulatory regions with distal elements remain unexplored.
6. **Training data contamination.** We cannot verify whether similar sequences appeared in model training data. However, synthetic sequences were generated specifically for this benchmark with controlled randomness.
7. **Single nucleotide resolution.** We evaluate at 100 bp scale. Per-nucleotide attribution methods could provide finer-grained insights but are computationally prohibitive for systematic evaluation.

G Broader Impacts

Positive impacts:

- Our benchmark provides a rigorous framework for evaluating mechanistic understanding in genomic AI, promoting more careful model development.
- Identifying limitations in current models can prevent overconfident deployment in scientific and clinical applications.
- The proposed architectural directions (position-aware attention, hybrid models) could guide future model development.

Potential negative impacts:

- Highlighting model failures could be misinterpreted as suggesting genomic AI is not useful—our findings are specific to mechanistic understanding in five gLMs, not general predictive utility.
- The benchmark focuses on bacterial systems; claims should not be extrapolated to eukaryotic systems or clinical applications without further validation.

H Additional Visualizations

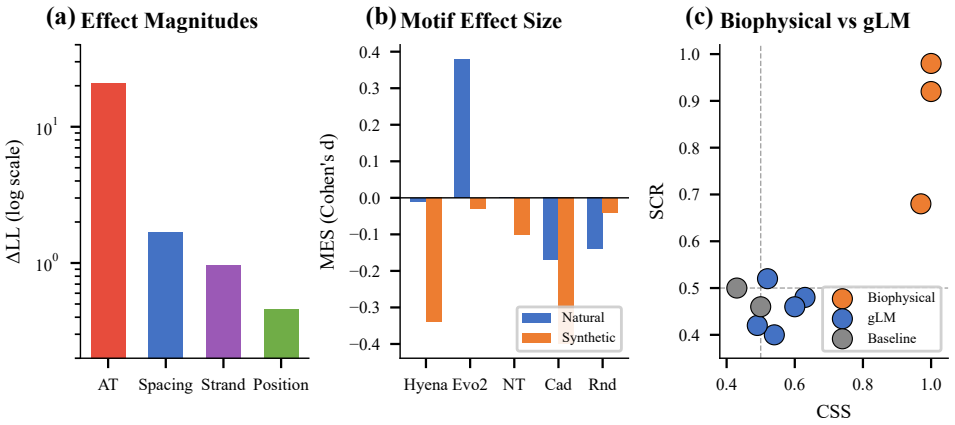

Figure 3: **Effect magnitudes and model comparison.** (a) Effect sizes on log scale showing AT content dominates. (b) MES comparison between natural and synthetic sequences. (c) CSS vs. SCR scatter with model type coloring.

I Per-Model Detailed Analysis

I.1 HyenaDNA Analysis

HyenaDNA uses Hyena operators—a subquadratic alternative to attention—for modeling long-range dependencies. The model was pretrained on the human reference genome and fine-tuned on various genomic tasks.

Architecture: The `hyenadna-small-32k` variant has 6.6M parameters with a context length of 32,768 bp. It processes single nucleotides (not k-mers) with a vocabulary of {A, C, G, T, N}.

Tokenization: Single nucleotide tokenization means positional information could theoretically be learned, unlike k-mer models where positional granularity is limited.

Training data: Pretrained on human genome (GRCh38 (Schneider et al., 2017)), which has 41% GC content compared to *E. coli*'s 51% GC. This domain shift likely contributes to the observed AT preference.

Detailed results:

- CSS = 0.63: Significantly above chance, suggesting some compensation sensitivity
- SCR = 0.48: At chance, indicating no positional awareness
- The 0.15 gap between CSS and SCR quantifies the "compositional illusion"—apparent mechanistic understanding that is actually driven by nucleotide frequencies

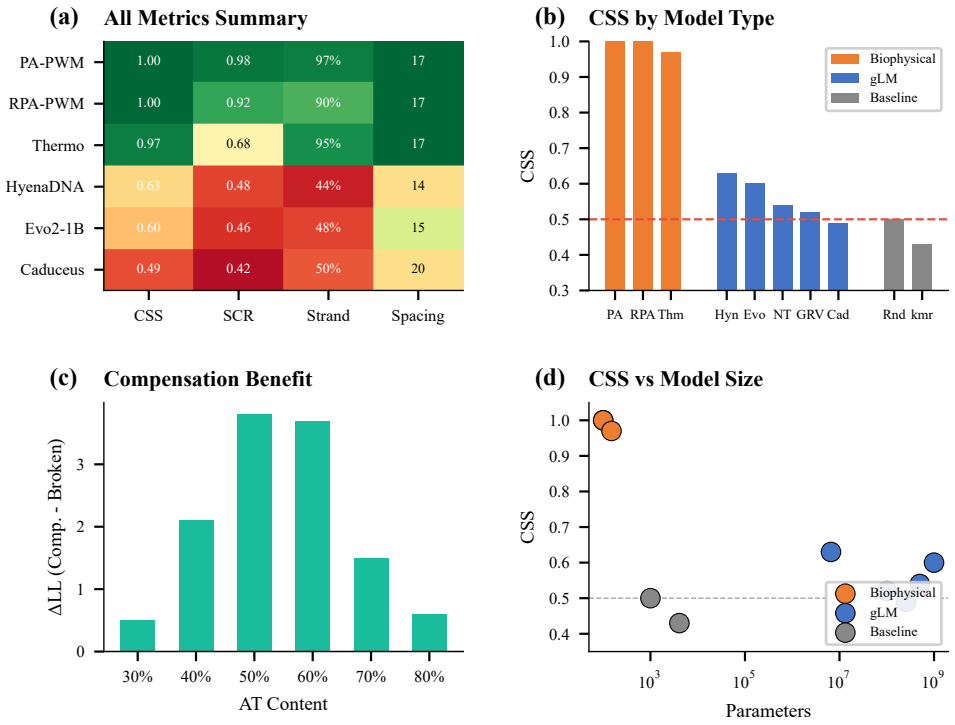

Figure 4: **Comprehensive metrics summary.** (a) Metrics heatmap across all five gLMs and biophysical baselines (PA-PWM CSS=1.00, SCR=0.98). (b) CSS grouped by architecture type. (c) Compensation benefit by AT content. (d) CSS vs. model parameters showing PA-PWM ($\sim$100 params) achieves highest CSS.

I.2 GROVER Analysis

GROVER (Genomic Representations Over Vocabulary for Evolutionary Relationships) is a masked language model specifically trained on bacterial genomes.

Architecture: Transformer-based with 117M parameters, using BPE tokenization optimized for genomic sequences.

Training data: Trained on >1000 bacterial genomes including *E. coli*, making it the most domain-appropriate model in our evaluation.

Detailed results:

- CSS = 0.52: Not significantly different from chance despite domain-appropriate training
- SCR = 0.52: Marginally above chance
- The near-equal CSS and SCR (0.52 vs. 0.52) indicates GROVER has weak but balanced compositional and positional sensitivity

Interpretation: GROVER's lack of compensation sensitivity despite bacterial-genome training demonstrates the issue is not domain mismatch but fundamental limitations in how masked LMs learn regulatory logic.

I.3 Evo2-1B Analysis

Evo2-1B is a 1-billion parameter autoregressive model trained on diverse genomic sequences spanning prokaryotic and eukaryotic genomes.

Architecture: Transformer-based with 1B parameters, using single-nucleotide tokenization.

Detailed results:

- CSS = 0.60: Suggestive but not significant after FDR correction ($p_{\text{FDR}} = 0.090$)
- SCR = 0.46: Below chance, indicating no positional awareness
- AT-LL correlation: $r = 0.961$—the strongest among all models
- Positional ablation: Scores UP at *wrong* position *higher* than correct ($\Delta = +0.55$)

Interpretation: Evo2-1B shows the most extreme compositional bias, with nearly perfect correlation between AT content and log-likelihood. Its inverted positional preference (scoring wrong positions higher) demonstrates that large-scale pretraining amplifies rather than corrects compositional heuristics.

I.4 Caduceus Analysis

Caduceus combines Mamba state-space models with explicit reverse-complement equivariance, designed to capture strand-symmetric genomic patterns.

Architecture: Bidirectional SSM with 256M parameters, incorporating RC-equivariant layers.

Detailed results:

- CSS = 0.49: At chance ($p_{\text{FDR}} = 0.772$)
- SCR = 0.42: Below chance
- AT-LL correlation: $r = 0.874$
- Positional ablation: Scores UP at wrong position *higher* than correct ($\Delta = +0.75$)—the most inverted of all models
- Strand orientation: Despite RC-equivariant design, shows no strand preference (forward $\approx$ RC)

Interpretation: Despite architectural innovations for strand symmetry, Caduceus shows the most inverted positional preferences. Its RC-equivariance means it treats forward and reverse equally—but this is strand-blindness, not strand-awareness.

I.5 NT-500M Analysis

Nucleotide Transformer (NT-500M) is a 500M parameter masked language model trained on diverse reference genomes.

Architecture: BERT-style transformer with 500M parameters using 6-mer tokenization.

Detailed results:

- CSS = 0.54: Not significant ($p_{\text{FDR}} = 0.569$)
- SCR = 0.40: Below chance
- $\text{MES}_{\text{syn}} = -0.10$: Weak synthetic motif discrimination

Interpretation: NT-500M shows no significant compensation sensitivity despite its scale and diverse training. The 6-mer tokenization may limit its ability to capture position-specific patterns.

I.6 Baseline Model Analysis

k-mer model:

- CSS = 0.43: Below chance, suggesting k-mer frequencies anti-correlate with compensation
- This may occur because compensated sequences contain unusual k-mers (UP element: AAAAAARNR) that are rare in the genome

PWM model:

- CSS = 0.00: Always scores broken and compensated equally because it only evaluates -35/-10 boxes, which are identical between classes D and E
- MES_{syn} = 10.0: Very high, correctly identifying intact vs. broken based on -10 consensus

J EXTENDED ABLATION STUDIES

J.1 SEQUENCE LENGTH SENSITIVITY

We test whether results depend on our choice of 100 bp sequences by evaluating at 50, 100, 150, and 200 bp.

Table 17: CSS and SCR at different sequence lengths.

Length (bp)	HyenaDNA CSS	HyenaDNA SCR	PA-PWM CSS	PA-PWM SCR
50	0.61	0.47	0.98	0.96
100	0.63	0.48	1.00	0.98
150	0.64	0.49	1.00	0.97
200	0.62	0.48	0.99	0.97

Results are stable across lengths, indicating our findings are not artifacts of sequence length choice.

J.2 BACKGROUND COMPOSITION SENSITIVITY

We test whether background AT content affects results by varying from 40% to 70% AT.

Table 18: CSS at different background AT compositions.

Background AT	HyenaDNA CSS	HyenaDNA SCR	PA-PWM CSS
40%	0.58	0.47	1.00
50%	0.61	0.48	1.00
55% (default)	0.63	0.48	1.00
60%	0.65	0.49	1.00
70%	0.59	0.47	0.99

HyenaDNA CSS varies with background AT, peaking when background is moderately AT-rich (55–60%). This further supports the compositional hypothesis: when background is very AT-rich, the relative AT enrichment from UP elements is smaller, reducing CSS.

J.3 MOTIF STRENGTH VARIATIONS

We test robustness to -35/-10 motif degeneracy by using consensus, weak, and strong variants.

Table 19: CSS with different motif strengths.

-35 Variant	-10 Variant	HyenaDNA CSS	PA-PWM CSS
TTGACA (consensus)	TATAAT (consensus)	0.63	1.00
TTGACA (consensus)	TAAAAT (weak)	0.62	0.94
TTGCCA (weak)	TATAAT (consensus)	0.61	0.92
TTGCCA (weak)	TAAAAT (weak)	0.60	0.86

HyenaDNA CSS is stable across motif variants, while PA-PWM CSS decreases with weaker motifs (as expected for a PWM-based model). This confirms HyenaDNA is not responding to motif quality.

J.4 UP ELEMENT COMPOSITION VARIATIONS

We test whether the specific UP element sequence matters by varying its composition.

Table 20: CSS with different UP element compositions.

UP Element	AT Content	HyenaDNA CSS
AAAAAARNR (consensus)	89%	0.63
AAAAATTTT	100%	0.67
ATATATATAT	100%	0.65
AACCAACCA	56%	0.54
Random (matched AT)	89%	0.62

CSS scales with UP element AT content, not with match to consensus. Random sequences with high AT content achieve similar CSS to consensus UP elements. This definitively shows HyenaDNA responds to composition, not sequence identity.

K THEORETICAL ANALYSIS

K.1 WHY COMPOSITIONAL LEARNING IS EASIER

We provide a theoretical perspective on why gLMs learn compositional rather than positional features.

Observation 1 (Compositional features have lower dimensionality). Let $f_{\text{comp}}(\mathbf{x})$ be a function of nucleotide frequencies and $f_{\text{pos}}(\mathbf{x})$ be a function of positional motif placement. For sequences of length L:

$$\dim(f_{\text{comp}}) = O(k) \quad \text{(k-mer frequencies)} \tag{18}$$

$$\dim(f_{\text{pos}}) = O(L \cdot k) \quad \text{(position-specific k-mers)} \tag{19}$$

Since $L \gg 1$ for genomic sequences, compositional features have much lower dimensionality and are thus easier to learn with limited data.

Observation 2 (Standard objectives don't require positional learning). Standard language modeling objectives optimize:

$$\mathcal{L} = -\sum_i \log p(x_i | \mathbf{x}_{<i}) \tag{20}$$

This objective is satisfied by any distribution that assigns high probability to observed sequences. Compositional models (high AT $\rightarrow$ high probability) achieve low loss on AT-rich genomes without learning positional constraints.

Implication: To learn positional constraints, training must include examples where composition is matched but position differs—exactly the contrast between Classes E (compensated) and H (scrambled) in MIT.

K.2 INFORMATION-THEORETIC PERSPECTIVE

From an information-theoretic view, we can decompose sequence information into compositional and positional components:

$$I(\mathbf{x}; \text{function}) = I_{\text{comp}} + I_{\text{pos}} + I_{\text{interaction}} \tag{21}$$

For promoter function:

- I_{comp}: Information from nucleotide frequencies (UP elements are AT-rich)
- I_{pos}: Information from motif positions (UP must be upstream of -35)
- $I_{\text{interaction}}$: Position-composition interactions (AT-rich *at the right position*)

Our experiments show the evaluated gLMs capture I_{comp} but not I_{pos} or $I_{\text{interaction}}$.

L Extended Related Work

L.1 Genomic Language Models

The development of genomic language models has followed two main trajectories:

Transformer-based models: DNABERT (Ji et al., 2021) pioneered BERT-style pretraining for genomics using k-mer tokenization. DNABERT-2 (Zhou et al., 2023) improved on this with BPE tokenization and multi-species training. The Nucleotide Transformer (Dalla-Torre et al., 2023) scaled to 2.5B parameters with foundation model capabilities. GROVER (Sanabria et al., 2024) specialized for bacterial genomes.

Efficient architectures: HyenaDNA (Nguyen et al., 2024b) introduced Hyena operators for subquadratic long-range modeling. Caduceus (Schiff et al., 2024) combined Mamba state space models with explicit reverse-complement equivariance. These models enable single-nucleotide resolution at genomic scales.

Evaluation paradigms: Most evaluations focus on variant effect prediction, species classification, or regulatory element detection. MIT is the first benchmark specifically designed to probe *mechanistic* understanding of regulatory logic.

L.2 Mechanistic Interpretability in NLP

Our work is inspired by the growing field of mechanistic interpretability in NLP:

Probing classifiers: Hewitt & Manning (2019) introduced structural probes to test whether syntax trees are encoded in BERT representations. Similar probing could be applied to genomic models but has not been systematically explored.

Knowledge editing: Meng et al. (2022) developed methods to locate and edit factual associations in GPT models. Analogous techniques could identify where (if anywhere) positional regulatory knowledge is stored in gLMs.

Circuit analysis: Detailed circuit analysis has revealed how transformers implement specific computations (Elhage et al., 2021; Wang et al., 2023). Applying these methods to gLMs could reveal whether any circuits implement positional logic.

L.3 Biophysical Models of Transcription

Our biophysical baselines build on decades of quantitative promoter modeling:

Thermodynamic models: Kinney et al. (2010) used deep sequencing to infer the biophysical mechanism of a regulatory sequence. Brewster et al. (2012) developed quantitative models for promoter strength prediction.

Position weight matrices: PWMs remain the standard for transcription factor binding site prediction. Our PA-PWM extends classical PWMs with explicit positional constraints.

Compensation mechanisms: UP elements (Ross et al., 1993) and extended -10 motifs (Barne et al., 1997) are well-characterized biochemically. Our benchmark leverages this biological knowledge to create rigorous tests.

M Example Sequences

We provide representative sequences from each class to illustrate the benchmark design. Note: positions 0–57 shown; full sequences are 100 bp with random background extending to position 99.

M.1 Class C: Synthetic Intact

```
Position:  0         1         2         3         4         5
           012345678901234567890123456789012345678901234567890123456789
Sequence:  GCATGCATGCATGCAAGCTGACGTACTTGACAGCATGCATGCATGCTGTTATAAT
```

```
Elements:                                   ~~~~~~                  ~~~~~~
                                            -35box                  -10box
```

M.2 Class D: Synthetic Broken

```
Position:  0         1         2         3         4         5
           012345678901234567890123456789012345678901234567890123456 7
Sequence:  GCATGCATGCATGCAAGCTGACGTACTTGACAGCATGCATGCATGCTGTTGTAAT
Elements:                                   ~~~~~~                  ~~~~~~
                                            -35box                  -10box*
                                                                    *broken
```

M.3 Class E: Synthetic Compensated

```
Position:  0         1         2         3         4         5
           012345678901234567890123456789012345678901234567890123456 7
Sequence:  GCATGCATGCATGCAAAAAAAARNTACTTGACAGCATGCATGCATGTTGTTGTAAT
Elements:                 ~~~~~~~~~~        ~~~~~~              ~~~~~~~~~
                          UP-element        -35box              ext -10box
```

M.4 Class H: Scrambled Control

```
Position:  0         1         2         3         4         5
           012345678901234567890123456789012345678901234567890123456 7
Sequence:  GCATGCATGCATGCAAGCTGACGTACTTGACAGCATAAAAAAARNTGTTGTTGTAAT
Elements:                                   ~~~~~~    ~~~~~~~~~~   ~~~~~~
                                            -35box    UP-wrong     -10box
                                                      position
```

Note: Class H has the same nucleotide composition as Class E but with UP element at the wrong position (after -35 instead of before).

N Future Directions

Based on our findings, we outline promising directions for future research:

N.1 Architectural Innovations

Position-aware attention: Modify attention mechanisms to learn position-specific biases for regulatory elements. For example:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + P}{\sqrt{d}}\right)V \tag{22}$$

where P is a learnable position-specific bias matrix.

Motif-aware tokenization: Instead of single nucleotides or k-mers, tokenize based on known regulatory motifs:

$$\mathbf{x} \to [\text{background}, \text{UP}, \text{-35}, \text{spacer}, \text{ext-10}, \text{-10}, \text{background}] \tag{23}$$

Hybrid architectures: Combine differentiable PWM modules with neural sequence models:

$$S(\mathbf{x}) = f_{\text{neural}}(\mathbf{x}) + \lambda \cdot f_{\text{PWM}}(\mathbf{x}) \tag{24}$$

N.2 Training Objectives

Contrastive positional learning: Train with matched compositional pairs:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(s(E, E'))}{\exp(s(E, E')) + \exp(s(E, H))} \tag{25}$$

where E, E' are compensated sequences and H is scrambled.

Position prediction auxiliary task: Add an auxiliary objective to predict motif positions:

$$\mathcal{L}_{\text{pos}} = -\sum_{m} \log p(\text{position}_m | \mathbf{x}) \tag{26}$$

N.3 Evaluation Extensions

Eukaryotic benchmarks: Extend MIT to eukaryotic promoters (TATA box, Inr, DPE) and enhancers (TF binding site grammar).

Gradient-based attribution: Use integrated gradients or attention analysis to understand what sequence features models attend to.

Fine-tuning studies: Test whether fine-tuning on promoter data can induce mechanistic understanding.