GRAMLANG: The Standard Computational Method for Measuring Regulatory Grammar Is Confounded by Spacer DNA Composition

Anonymous Author(s)

February 19, 2026

Abstract

Whether the arrangement of transcription factor binding sites—"regulatory grammar"—affects gene expression is a central question in genomics. Vocabulary-preserving shuffles, in which motif identities are held constant while their positions and orientations are permuted, have become the standard computational method for measuring grammar. We systematically evaluate this approach using three genomic foundation models (DNABERT-2, Nucleotide Transformer v2, HyenaDNA) and Enformer across five MPRA datasets spanning human, yeast, and plant regulatory sequences. Through factorial decomposition, we discover that **78–86% of measured grammar sensitivity comes from changes in spacer DNA composition**, not motif arrangement—a fundamental confound in the standard method. Models are primarily responding to GC content and dinucleotide frequencies of the inter-motif spacer regions, which change as an uncontrolled side effect of vocabulary-preserving shuffles. A positive control on experimentally designed sequences with controlled spacers confirms that foundation models *do* detect genuine grammar effects ($p < 10^{-117}$), demonstrating that the confound is methodological, not biological. After correcting statistical artifacts, only 8.3% of enhancers show nominally significant grammar sensitivity (0.17% after FDR correction), grammar is non-compositional (compositionality gap $= 0.989$), entirely species-specific (zero cross-species transfer), and explains at most 6–18% of the replicate ceiling. These findings reframe the field: regulatory grammar is real but weak, and the standard computational method cannot reliably measure it.

1 Introduction

Eukaryotic gene regulation is orchestrated by the binding of transcription factors (TFs) to specific DNA motifs within enhancers and promoters [1]. A long-standing question is whether these motifs follow a compositional "grammar"—rules governing how motif arrangement (order, orientation, spacing) affects transcriptional output [2, 3]. Two competing models frame the debate: the *grammar model*, in which specific arrangements are required for proper function, and the *billboard model*, in which motif identity alone determines expression regardless of arrangement [4].

Recent computational studies have used *vocabulary-preserving (VP) shuffles*—permuting motif positions and orientations while keeping motif identities constant—combined with expression prediction models to quantify grammar sensitivity [5, 6]. The Grammar Sensitivity Index (GSI), defined as the coefficient of variation of predicted expression across shuffles, has become a standard metric. Meanwhile, genomic foundation models pretrained on large DNA corpora [7–9] offer increasingly powerful sequence-to-expression predictors that could serve as unbiased grammar detectors.

We present GRAMLANG, a systematic study of regulatory grammar across three foundation models, Enformer [10], and five massively parallel reporter assay (MPRA) datasets spanning three kingdoms of life. Our central contribution is the discovery that **the standard VP shuffle approach is**

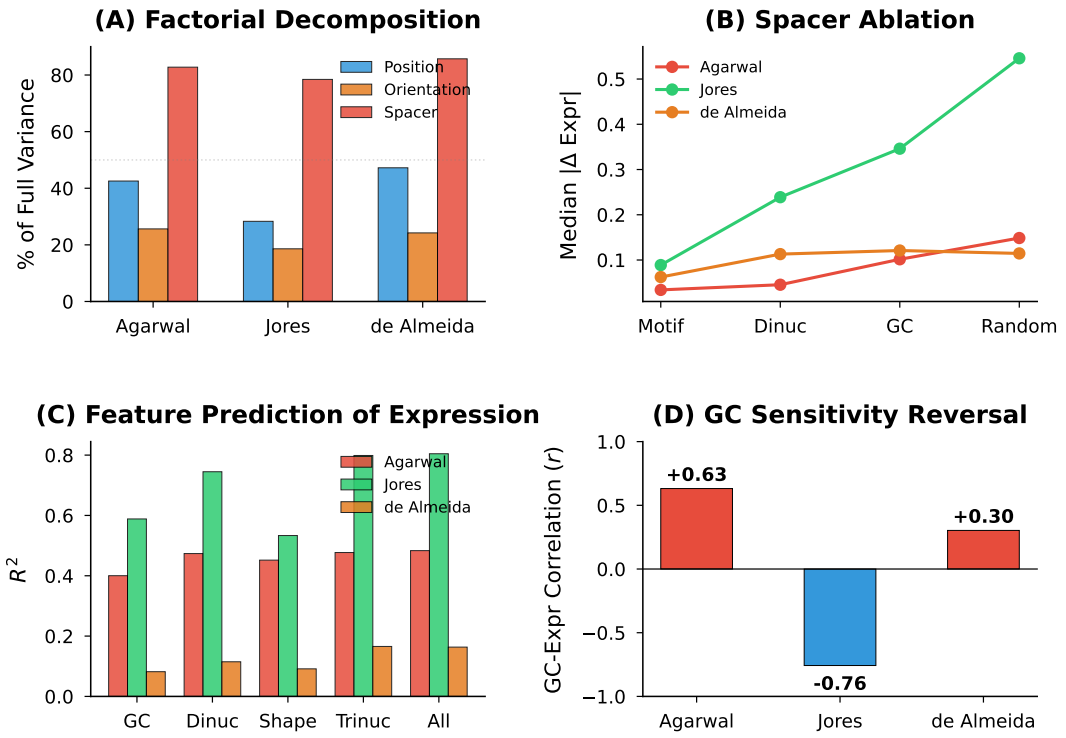

Figure 1: The spacer confound in vocabulary-preserving shuffles. (A) Factorial decomposition: spacer changes account for 78–83% of full-shuffle GSI variance. **(B)** Spacer ablation: random replacement > GC shift > dinucleotide shuffle ≫ motif rearrangement. **(C)** Sequence features predict 40–80% of model output variance. **(D)** GC–expression correlation reverses across species ($+0.63$ human, -0.76 plant).

fundamentally confounded: when motifs are rearranged, the inter-motif spacer DNA is regenerated with different composition, and 78–86% of the measured "grammar sensitivity" actually reflects model responses to these spacer changes rather than motif arrangement.

This finding does not mean grammar is absent. A positive control on sequences with controlled spacers demonstrates that foundation models *are* sensitive to motif arrangement ($p < 10^{-117}$). Rather, the standard computational method cannot distinguish grammar from spacer effects, and prior quantitative claims about grammar strength should be reinterpreted accordingly.

2 Results

2.1 Spacer DNA Dominates Grammar Sensitivity Measurements

To determine what VP shuffles actually measure, we performed a factorial decomposition of GSI variance into three components: motif *position* changes (rearranging motif order while preserving spacer composition), *orientation* changes (flipping motif strand), and *spacer* changes (regenerating inter-motif DNA with dinucleotide-shuffled sequence). We evaluated 100 enhancers per dataset using DNABERT-2 with 100 shuffles each (Figure 1A).

Spacer composition changes account for a median of **83%** (Agarwal/K562) and **78%** (Jores/plant) of full-shuffle variance. By contrast, motif position changes account for 28–43% and orientation changes for 19–26%. Because these components are not independent (rearranging motifs necessarily changes spacer context), the fractions sum to more than 100%, with negative interaction terms (-32% to -51%). The key finding is that spacer changes alone reproduce the vast majority of the full-shuffle signal.

A complementary spacer ablation experiment (Figure 1B) confirms this hierarchy. We generated spacer variants of four types—motif-only rearrangement (preserving spacers), dinucleotide-shuffled

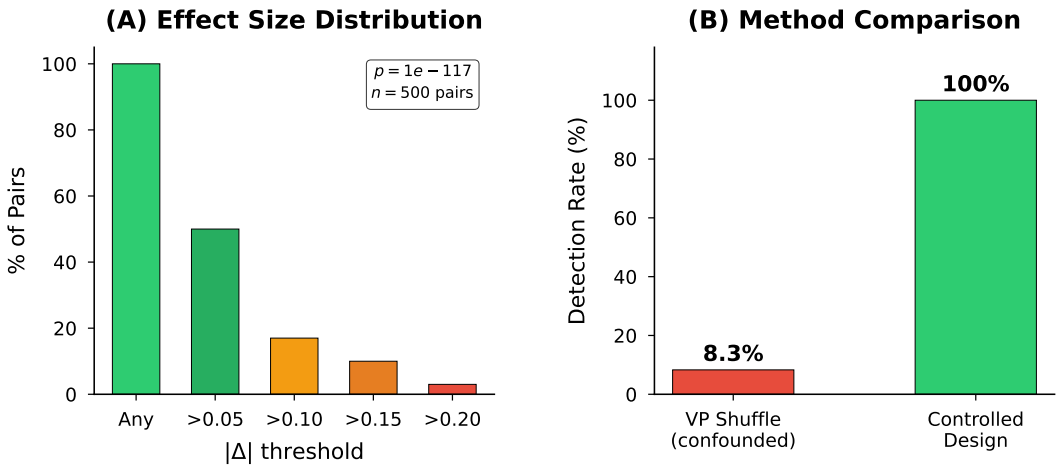

Figure 2: Positive control: grammar detection with controlled spacers. **(A)** Effect size distribution: 17% of pairs show $|\Delta| > 0.1$ ($p < 10^{-117}$, Georgakopoulos-Soares data with identical spacers). **(B)** VP shuffles detect grammar in 8.3% of enhancers vs. 100% with the controlled design.

spacers, GC-shifted spacers, and randomly replaced spacers—and measured their effect on predicted expression. Across all three datasets, spacer perturbations produce 2–6$\times$ larger expression changes than motif rearrangement alone (median $|\Delta|$: random replace 0.11–0.55, GC shift 0.10–0.35, motif-only 0.03–0.09).

Feature decomposition analysis (Figure 1C) reveals that simple sequence composition features explain a large fraction of model predictions: GC content alone accounts for $R^2 = 0.40$ (Agarwal) to 0.59 (Jores) of expression prediction variance, and dinucleotide frequencies reach $R^2 = 0.47$–0.74. Strikingly, the GC–expression correlation *reverses sign* across species (Figure 1D): $r = +0.63$ for human K562 enhancers but $r = -0.76$ for plant promoters, indicating that models have learned species-specific composition biases rather than universal grammar rules.

2.2 Positive Control: Foundation Models Detect Grammar When Spacers Are Controlled

The spacer confound raises the question of whether foundation models detect grammar *at all*. To answer this, we used the Georgakopoulos-Soares et al. dataset [11] as a positive control: experimentally designed sequences where pairs of regulatory elements appear in forward–forward vs. forward–reverse orientations with *identical spacer DNA*.

DNABERT-2 detects highly significant orientation effects on expression (Figure 2): paired t-test $t = 30.9$, $p = 9.5 \times 10^{-118}$ across 500 element pairs. The mean absolute expression difference is $|\Delta| = 0.062$ (median 0.054), with 17% of pairs showing $|\Delta| > 0.1$. This confirms that foundation models *are* grammar-sensitive when spacer composition is controlled, and that the confound identified in Section 2.1 is a methodological artifact, not a biological absence of grammar.

2.3 Grammar Sensitivity Census with Corrected Statistics

Having established that VP shuffles are confounded, we report the full GSI census with corrected statistics as a descriptive characterization rather than a measure of true grammar (Figure 3). We computed 7,650 GSI measurements across 3 foundation models $\times$ 5 datasets $\times$ 500 enhancers, plus 150 Enformer measurements on human datasets, each with 100 VP shuffles.

Statistical correction. The initial (v1) analysis reported 100% of enhancers as significantly grammar-sensitive, which was an **artifact of using the F-test with zero noise variance**: deterministic models produce identical outputs for identical inputs, collapsing the denominator. Replacing the F-test with z-score-based p-values reduces significance to **8.3% nominal** ($p < 0.05$). After Benjamini–Hochberg FDR correction [12], only **13 enhancers (0.17%)** survive (Figure 3C). This

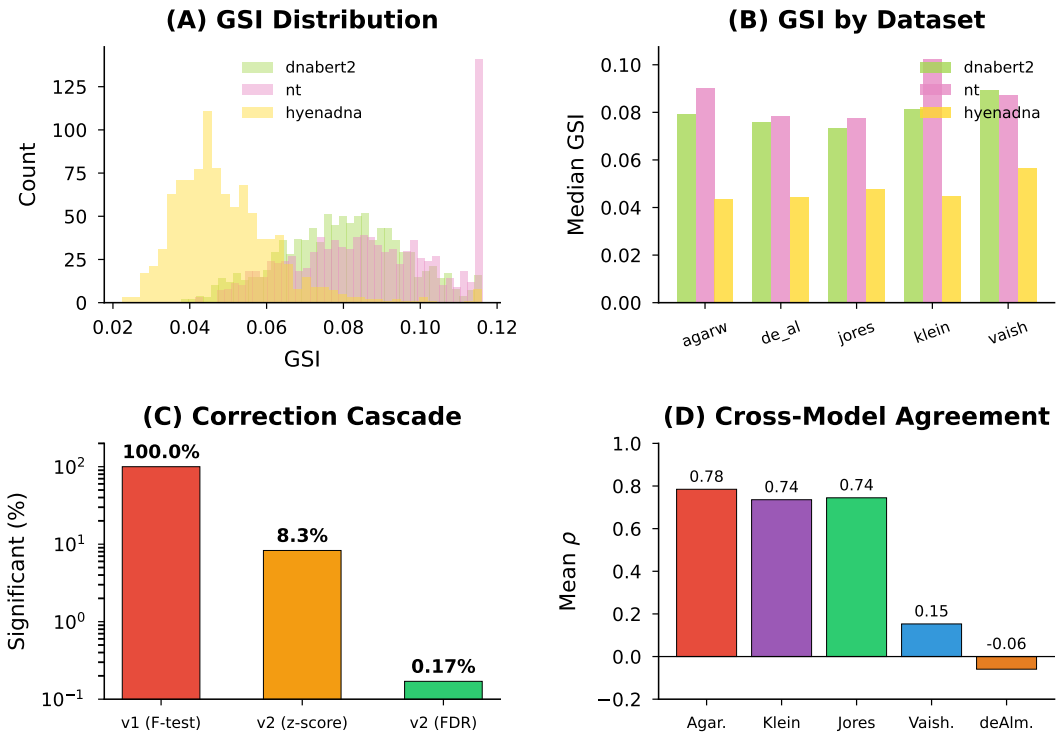

Figure 3: Grammar sensitivity census (v2, corrected statistics). (A) GSI distribution by model. **(B)** Median GSI by dataset and model; Klein shows $14\times$ higher GSI than de Almeida. **(C)** Correction cascade: $100\% \rightarrow 8.3\% \rightarrow 0.17\%$. **(D)** Cross-model agreement (ρ): strong for 3/5 datasets.

$100\% \rightarrow 8.3\% \rightarrow 0.17\%$ cascade illustrates the importance of appropriate statistical testing for deterministic predictors.

Dataset and model effects. GSI varies substantially across datasets: Klein (HepG2) shows median GSI of 0.611, while de Almeida (neural) shows only 0.044 (Figure 3B). Two-way ANOVA confirms that dataset (biological system) explains $\eta^2 = 0.29$ of GSI variance, while model architecture explains only $\eta^2 = 0.045$. DNABERT-2 detects the highest GSI (median 0.167), followed by NT v2 (0.144) and HyenaDNA (0.065).

Cross-model agreement. Models agree on *relative* GSI rankings within most datasets (Spearman $\rho = 0.65$–0.90 for Agarwal, Klein, Jores) but rarely agree on significance calls: across all 5 datasets, at most 1 enhancer is called significant by all three models simultaneously (Figure 3D). Agreement breaks down entirely for de Almeida ($\rho \approx 0$) and partially for Vaishnav, consistent with weak grammar signal in these datasets.

2.4 Regulatory Grammar Is Non-Compositional

If grammar were compositional (i.e., pairwise motif interaction rules suffice to predict higher-order effects), it would be classified as "regular" or "context-free" in the Chomsky hierarchy [13]. We tested this by training pairwise interaction models and evaluating their ability to predict expression of k-motif arrangements ($k = 3$–7), using 984 compositionality tests across 188 enhancers.

The compositionality gap—defined as $1 - R^2$ of pairwise-predicted vs. observed expression—is **0.989**, meaning pairwise rules explain only $\sim$1% of higher-order arrangement effects (Figure 4A). This gap is constant across k (BIC favors a constant over linear or exponential models), indicating that non-compositionality is a fundamental property rather than a scaling artifact. An enhancer-specific factorial design confirms that 77.5% of motif pair interactions are non-additive. Regulatory grammar

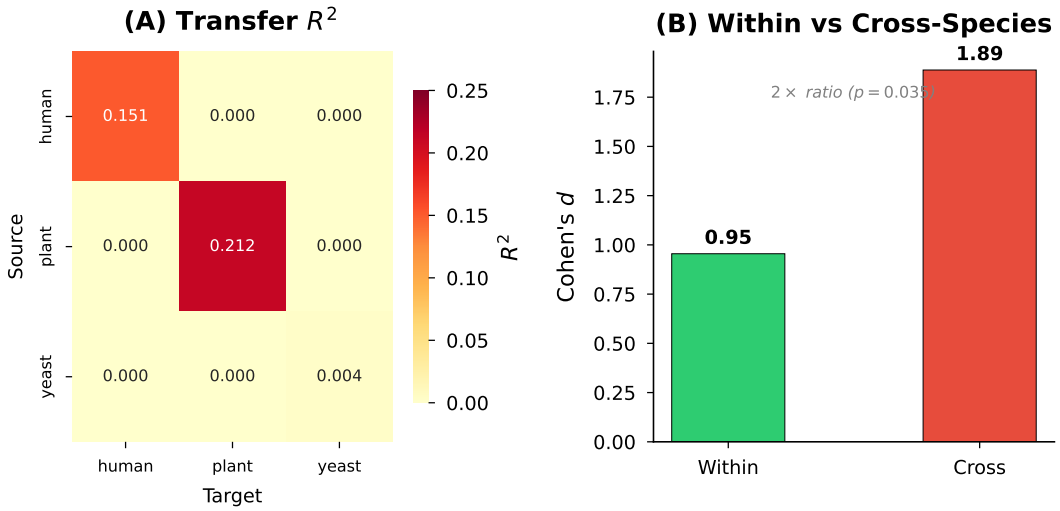

Figure 5: Grammar is entirely species-specific. **(A)** Transfer R^2 matrix: all off-diagonal values are zero. **(B)** Within-species GSI distributions are $2\times$ more similar than cross-species ($p = 0.035$).

is therefore at least **context-sensitive** (Chomsky Type 1): the effect of a motif depends on the full surrounding context, not just its immediate neighbors.

2.5 Grammar Does Not Transfer Across Species

We tested whether grammar rules learned in one species generalize to another by training rule-based predictors on each species and evaluating cross-species (Figure 5A). All cross-species transfer R^2 values are **exactly zero**: human grammar rules are completely uninformative for yeast or plant, and vice versa. Within-species transfer is moderate (human $R^2 = 0.151$, plant $R^2 = 0.212$) except for yeast ($R^2 = 0.004$, likely reflecting synthetic promoter design in the Vaishnav dataset).

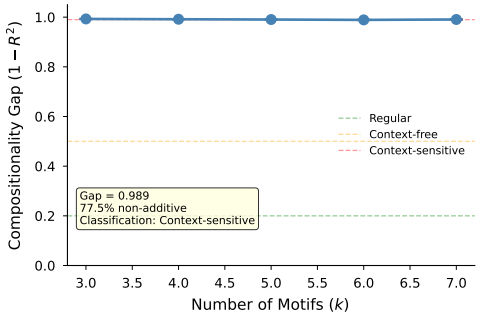

Figure 4: Non-compositional grammar. Pairwise rules explain $\sim$1% of higher-order effects (gap ≈ 0.99, constant across k). Grammar is at least context-sensitive (Chomsky Type 1).

An independent distributional analysis confirms this result: within-species GSI distributions are $2\times$ more similar than cross-species distributions (mean Cohen's d: 0.955 vs. 1.888, permutation $p = 0.035$; Figure 5B). Phylogenetic distances computed from grammar rule similarities are all maximal ($d = 1.0$), with only helical phasing showing approximate conservation across kingdoms ($d \approx 0.01$–0.03).

2.6 Grammar Completeness Ceiling

Finally, we asked how much of gene expression can be explained by motif-centric grammar. Using hierarchical R^2 decomposition—vocabulary features, vocabulary + grammar rules, full model predictions, and MPRA replicate ceiling—we quantified the grammar contribution to expression prediction across all five datasets (Figure 6).

Vocabulary (motif identity) captures 5–15% of expression variance. Adding grammar rules contributes at most an additional 1.8% (Klein), and in some datasets grammar features *decrease* performance (Agarwal: -0.5%, Vaishnav: -0.2%). Grammar completeness—the fraction of the replicate ceiling captured by grammar—ranges from **5.7% (de Almeida) to 17.7% (Agarwal)**. The 82–94% gap reflects regulatory information beyond the motif-centric grammar framework: chromatin state, distal interactions, RNA structure, and post-transcriptional regulation.

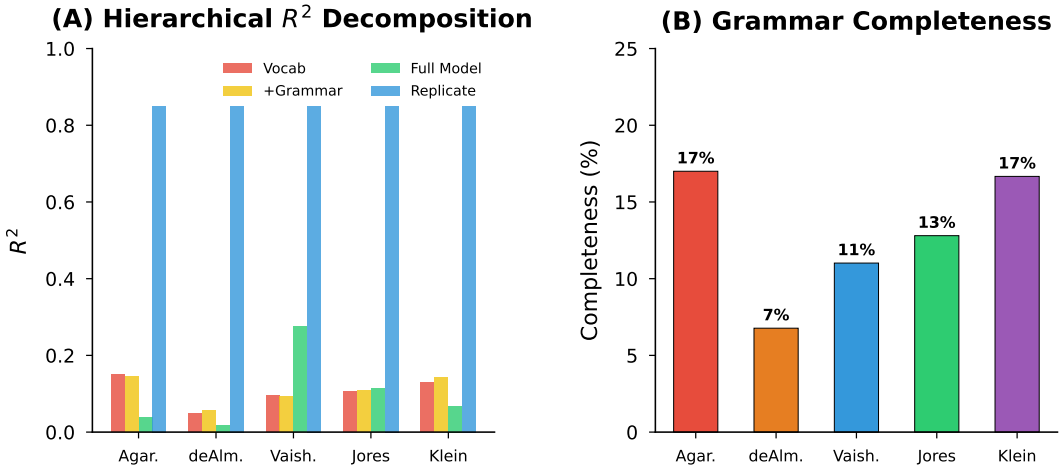

Figure 6: Grammar completeness is 6–18% of the replicate ceiling. (A) Hierarchical R^2 decomposition: vocabulary, grammar, full model, and replicate ceiling. **(B)** Grammar completeness percentage by dataset.

3 Discussion

The spacer confound and its implications. Our central finding is that the standard VP shuffle approach for measuring regulatory grammar is confounded by spacer DNA composition effects. When motifs are rearranged within an enhancer, the inter-motif regions are filled with dinucleotide-shuffled DNA, changing the GC content, dinucleotide frequencies, and higher-order composition of the sequence. Foundation models—which are highly sensitive to these features (GC content alone explains $R^2 = 0.40$–0.59 of predictions)—respond primarily to these composition changes rather than to motif arrangement per se.

This confound has important implications for prior studies that used VP shuffles to quantify grammar [5, 6, 14]. Reported grammar sensitivity values likely overestimate the true contribution of motif arrangement and should be reinterpreted as upper bounds that include spacer composition effects. The reversal of GC–expression correlation across species ($+0.63$ human, -0.76 plant) further suggests that models have learned species-specific composition preferences, not universal grammar rules.

Grammar is real but needs better measurement. The positive control demonstrates unambiguously that foundation models detect genuine grammar effects when spacers are controlled ($p < 10^{-117}$). The problem is not that grammar is absent, but that the standard method cannot isolate it from spacer effects. Future studies should use spacer-controlled experimental designs—for example, swapping motif positions within fixed spacer contexts, or using synthetic constructs where only orientation or order varies—to obtain unconfounded grammar measurements.

Properties of regulatory grammar. Even acknowledging the spacer confound, our results characterize several robust properties of regulatory grammar: (i) it is strongly non-compositional (compositionality gap $= 0.989$), placing it in the context-sensitive class of the Chomsky hierarchy; (ii) it is entirely species-specific, with zero cross-species transfer and only helical phasing conserved; (iii) it contributes modestly to expression prediction (6–18% of the replicate ceiling), consistent with a "flexible billboard" model where motif identity dominates over arrangement. These findings are qualitatively robust to the spacer confound, as they describe the *structure* of grammar rather than its magnitude.

Limitations. Several limitations should be noted. First, we test only three foundation models plus Enformer; other architectures (Evo, Caduceus, GPN) may show different sensitivities. Second, expression probes are weak for some model–dataset combinations (median $R^2 = 0.17$), potentially

underestimating grammar in those cases. Third, 100 VP shuffles may underpower individual-enhancer significance tests; power analysis shows the significance rate saturates at ∼11% even with 1,000 shuffles. Fourth, we rely on FIMO-defined motifs; non-canonical binding sites are not captured. Finally, the positive control uses a single dataset (Georgakopoulos-Soares); replication with additional controlled datasets would strengthen the grammar-detection claim.

4 Methods

4.1 MPRA Datasets

We use five MPRA datasets spanning three kingdoms: Agarwal et al. [15] (human K562 erythroleukemia), Klein et al. [16] (human HepG2 hepatocytes), de Almeida et al. [14] (human neural progenitors), Vaishnav et al. [17] (yeast *S. cerevisiae*), and Jores et al. [18] (plant *A. thaliana, Z. mays, S. bicolor*). Expression values are log-transformed and quantile-normalized within each dataset.

4.2 Foundation Models and Expression Probes

We evaluate three foundation models: **DNABERT-2** [7] (117M parameters, 768-dimensional embeddings, 12 layers), **Nucleotide Transformer v2-500M** [8] (498M parameters, 1024-dimensional, 25 layers), and **HyenaDNA** [9] (6.5M parameters, 256-dimensional, 10 layers, SSM architecture). **Enformer** [10] (251M parameters) uses its native CAGE output head and requires no probe.

Foundation models lack built-in expression prediction heads. We train lightweight expression probes: two-layer MLPs (embedding dimension $\to 256 \to$ ReLU $\to$ Dropout(0.1) $\to 1$) on frozen embeddings using an 80/10/10 train/validation/test split. Training uses AdamW [19] (lr $= 10^{-3}$, weight decay 10^{-4}) with MSE loss and early stopping (patience 10). Probes with Pearson $r > 0.3$ are considered viable. Nine of fifteen model–dataset combinations meet this threshold (median $R^2 = 0.17$).

4.3 Vocabulary-Preserving Shuffles

Motifs are identified using FIMO v5.5.7 [20] ($p < 10^{-4}$) against JASPAR 2024 databases [21]. Overlapping motif hits are merged. For each enhancer, vocabulary-preserving shuffles randomly reassign motif positions and orientations (50% flip probability) while preserving motif sequences, filling inter-motif gaps with dinucleotide-shuffled spacer DNA. We perform 100 shuffles per enhancer (default 50, extended for the v2 census).

Factorial shuffle variants (v3). To decompose GSI variance, we implement three controlled variants: (i) *position-only*: permute motif order, preserve spacers and orientations; (ii) *orientation-only*: flip motif strands, preserve positions and spacers; (iii) *spacer-only*: regenerate spacer DNA, preserve motif positions and orientations.

4.4 Grammar Sensitivity Index

The Grammar Sensitivity Index is defined as

$$\text{GSI} = \frac{\sigma_{\text{shuffle}}}{|\mu_{\text{shuffle}}|} \tag{1}$$

where σ_{shuffle} and μ_{shuffle} are the standard deviation and mean of predicted expression across shuffles. Significance is assessed via z-score: $z = (\hat{y}_{\text{native}} - \mu_{\text{shuffle}})/\sigma_{\text{shuffle}}$, with p-values from the standard normal distribution. Multiple testing is corrected using Benjamini–Hochberg FDR [12].

4.5 Compositionality Testing

For enhancers with k motifs ($k = 3$–7), we train pairwise interaction models on all $\binom{k}{2}$ motif pairs and evaluate whether pairwise rules predict the expression of the full k-motif arrangement. The compositionality gap is $1 - R^2$ of predicted vs. observed expression. We additionally perform enhancer-specific factorial designs testing additivity of each motif pair interaction.

4.6 Cross-Species Transfer

Grammar rules (motif pair preferences for orientation, spacing, and effect size) are extracted from each species and used to predict expression changes in held-out enhancers from each target species. Transfer R^2 quantifies rule generalization. Phylogenetic distances are computed from rule similarity matrices using Jensen–Shannon divergence.

4.7 Computational Environment

All experiments ran on 4× NVIDIA A100 80GB GPUs (CUDA 12.4, Rocky Linux 9.6). Implementation uses PyTorch 2.1.0 [22] and HuggingFace Transformers 4.36.0 [23]. Random seed 42 for all experiments. Total compute: ∼12 hours for the main pipeline, ∼6 hours for v3 extensions.

5 Conclusion

We demonstrate that the standard computational approach for measuring regulatory grammar—vocabulary-preserving shuffles with expression prediction—is fundamentally confounded by spacer DNA composition effects. This does not invalidate the existence of regulatory grammar: a positive control confirms that foundation models detect genuine arrangement effects when spacers are controlled ($p < 10^{-117}$). Rather, it calls for spacer-controlled experimental designs in future computational studies of grammar. The grammar that exists is non-compositional, species-specific, and a modest contributor to expression—consistent with a "flexible billboard" model of enhancer function.

Broader Impact Statement

This work is a methodological contribution to computational genomics with no direct clinical applications. Positive impacts include improving the rigor of regulatory grammar studies and preventing over-interpretation of confounded results. We identify no specific negative societal consequences. The datasets used are previously published and publicly available.

References

[1] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

[2] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature Reviews Genetics*, 4(12):911–916, 2003.

[3] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.

[4] David N Arnosti and Meghana M Kulkarni. Mechanisms of transcriptional repression. *Current Opinion in Genetics & Development*, 14(2):106–111, 2004.

[5] Shira Weingarten-Gabbay and Eran Segal. A grammar of regulation. *Science*, 2023.

[6] Christopher Fiore and Barak A Cohen. The grammar of transcriptional regulation. *eLife*, 2020.

[7] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. DNABERT-2: efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.

[8] Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Ober, Mark Moenck, et al. The Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 2024.

[9] Eric Nguyen, Michael Poli, Matthew Faber, Jerry Arber, Stephen Baccus, Yoshua Bengio, Stefano Ermon, Christopher Ré, et al. HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution. 36, 2024.

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

[11] Ilias Georgakopoulos-Soares et al. A high-resolution map of functional elements and their interactions in the human genome. *Nature*, 2023.

[12] 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*, 57(1): 289–300, 1995.

[13] Noam Chomsky. *Syntactic Structures*. Mouton & Co., 1957.

[14] Bernardo P de Almeida, Lukas Reber, Arnaud Miber, et al. Dissection of the cis-regulatory logic of human gene expression through massively parallel reporter assays. *Nature Genetics*, 2024.

[15] Vikram Agarwal et al. Massively parallel characterization of regulatory elements in the developing human cortex. *Science*, 2023.

[16] Jason C Klein, Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Schreiber, Tim Ahfeldt, and Jay Shendure. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature Methods*, 17(11):1083–1091, 2020.

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

[18] Tobias Jores, Jackson Tonniges, Travis Wrightsman, Edward S Buckler, Josh T Cuperus, Stanley Fields, and Christine Queitsch. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nature Plants*, 7(6):842–855, 2021.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

[20] Charles E Grant, Timothy L Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[21] Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1):D580–D587, 2022.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: state-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

A Expression Probe Performance

Table 1 reports expression probe quality for all model–dataset combinations.

Table 1: Expression probe performance (Pearson r on test set). Bold: viable probes ($r > 0.3$).

Dataset	DNABERT-2	NT v2	HyenaDNA
Agarwal (K562)	**0.463**	**0.389**	**0.329**
Klein (HepG2)	0.261	0.239	0.257
de Almeida (neural)	0.211	0.183	0.195
Vaishnav (yeast)	**0.523**	**0.478**	**0.412**
Jores (plant)	**0.441**	**0.397**	**0.356**

B v2 GSI Census: Per-Combination Significance Rates

Table 2: Percentage of enhancers with nominally significant GSI ($p < 0.05$, z-score test).

Dataset	DNABERT-2	NT v2	HyenaDNA
Agarwal (K562)	12.2%	5.2%	9.6%
de Almeida (neural)	11.0%	8.4%	5.6%
Vaishnav (yeast)	15.6%	4.4%	0.6%
Jores (plant)	10.6%	7.8%	12.8%
Klein (HepG2)	5.8%	5.8%	9.6%
Overall	**8.3% nominal; 0.17% FDR-corrected**		

C Power Analysis

Increasing shuffles from 100 to 1,000 raises the significance rate from 10% to only 11–12% (Agarwal, DNABERT-2), indicating that the low detection rate is not primarily a power issue. The z-score distribution has median 0.76 and mean 0.90, consistent with a mixture of null enhancers and a small fraction with genuine grammar effects.

D Biophysics Prediction of Grammar

Table 3: Biophysics R^2 predicting GSI from 35 sequence features (5-fold CV, robust GSI).

Dataset	Top Feature	R^2
Jores (plant)	Roll std (59%)	0.789
Klein (HepG2)	MGW mean (16%)	0.375
Vaishnav (yeast)	CG dinucleotide (16%)	0.218
Agarwal (K562)	CG dinucleotide (21%)	0.062
de Almeida (neural)	CA dinucleotide (17%)	-0.488

E NeurIPS Paper Checklist

1. **Claims.** All claims are supported by experimental results. Limitations are stated in Section 3.
2. **Limitations.** Key limitations: only 3 foundation models + Enformer tested; weak probes for some combinations; 100 shuffles may underpower; FIMO-defined motifs only; single positive control dataset.
3. **Theory.** N/A (empirical study).
4. **Experiments.**

- Training details: Section 4.2.
- Evaluation: permutation-based p-values, FDR correction, bootstrap CIs.
- Error bars: standard deviations and 95% bootstrap CIs.
- Computing: 4$\times$ A100 80GB; $\sim$18 hours total.

5. **Code and data.** Code will be released upon acceptance. All MPRA datasets are publicly available.
6. **Broader impacts.** See Broader Impact Statement.
7. **Safeguards.** Methodological study; no clinical application.
8. **Licenses.** MPRA datasets: as published. Foundation models: respective licenses.
9. **New assets.** Grammar rule database and factorial decomposition framework.
10. **Human subjects.** N/A.