BEYOND ACTIVITY PREDICTION: EVALUATING COMPOSITIONAL REGULATORY GENERALIZATION IN GENOMIC FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models claim that they can recognize regulatory sequences and predict activity, but whether they learn transferable rules governing combinations of regulatory elements remains unclear. We introduce a matched-perturbation evaluation framework that measures regulatory effects and interactions rather than individual sequence properties alone. This distinction is important because regulatory function is inherently combinatorial and context dependent: the effect of a motif, enhancer, or variant can change when other regulatory elements or promoter contexts change. The benchmark spans a hierarchy of increasingly complex capabilities, including single-element perturbations, pairwise interactions, configuration changes, promoter-context transfer, and higher-order combinations involving three to eight regulatory features, covering 14 pretrained model configurations and 59,851 complete combinatorial perturbation sets. Across this hierarchy, we identify a consistent dissociation between activity prediction and compositional regulatory generalization. Models capture sequence-level signals and some local effects, but their ability to predict transferable interactions declines sharply as regulatory relationships become more context dependent. Notably, models that accurately predict construct-level activity fail to recover promoter-dependent interaction changes and do not consistently outperform zero-interaction baselines for higher-order effects. These results reveal that current genomic foundation models learn regulatory activity without reliably capturing the compositional principles that govern how regulatory elements combine and transfer across contexts, motivating perturbation-based evaluation as a necessary complement to standard activity benchmarks.

1 INTRODUCTION

Genomic foundation models (GFMs) have achieved strong performance in recognizing regulatory sequences, predicting regulatory activity, and estimating the effects of sequence variation (Ji et al., 2021; Dalla-Torre et al., 2025; Nguyen et al., 2023; Schiff et al., 2024; Linder et al., 2025). These advances have established large-scale sequence modeling as a powerful approach for learning genomic representations. However, a more fundamental question remains: do these representations capture the principles that govern how regulatory elements interact and combine? Gene regulation is inherently compositional and context dependent. The effect of a regulatory element depends not only on its own sequence, but also on interacting regulatory elements, their configuration, and the surrounding regulatory context (Georgakopoulos-Soares et al., 2023; Martinez-Ara et al., 2024; Waters et al., 2021). Therefore, accurately modeling individual sequences or their activities does not necessarily imply that a model has learned transferable rules governing combinations of regulatory elements (Figure 1(a)).

Existing evaluations of GFMs primarily measure sequence recognition, activity prediction, and variant effects (Zhou et al., 2024; Marin et al., 2024; Patel et al., 2024). While these benchmarks have advanced our understanding of genomic representation learning, they provide limited evaluation of a distinct capability required for biological reasoning: compositional regulatory generalization. In particular, it remains unclear whether learned genomic representations can recover experimentally measured interactions among regulatory elements and whether these relationships remain predic-

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

tive when regulatory configurations or biological contexts change (Koo et al., 2021; Toneyan and Koo, 2024; Seitz et al., 2024). This limitation is closely related to common pretraining objectives: sequence reconstruction and activity prediction encourage models to represent individual sequence properties, but do not explicitly require them to preserve the functional relationships among perturbations that define combinatorial regulation. Thus, a central unresolved question is whether genomic foundation models learn transferable regulatory relationships beyond activity prediction.

To systematically evaluate this missing capability, we introduce a perturbation-based framework for measuring compositional regulatory generalization. Rather than evaluating sequences independently, our framework compares matched experimental perturbations and computes the same functional contrasts from both experimental measurements and model predictions. This design directly tests whether a model captures the effect of a perturbation or interaction, rather than simply predicting the activities of the individual sequences involved. The benchmark spans a hierarchy of increasingly complex regulatory capabilities, including single-element perturbation effects, pairwise interactions, configuration changes, promoter-context transfer, and higher-order combinations involving three to eight regulatory features (Figure 1(b)). Across 14 pretrained model configurations, we evaluate both native model scores and supervised readouts of frozen representations. The benchmark contains 59,851 complete higher-order perturbation sets, with shared sequences and biological backgrounds grouped to prevent information leakage across evaluation splits.

Our results reveal a consistent gap between activity prediction and compositional regulatory generalization (Figure 1(c)). Models capture substantial sequence-level information and some local regulatory effects, but their ability to predict regulatory relationships progressively declines as interactions become more context dependent. In particular, models can accurately predict construct-level activity under held-out combinations and configurations when regulatory contexts are observed, yet fail to transfer interaction effects across unseen promoter contexts. This limitation is not specific to genomic language models: functionally pretrained regulatory models exhibit the same qualitative pattern. The gap further extends to higher-order combinations, where partial agreement in interaction rankings does not consistently translate into accurate prediction of interaction magnitudes. In summary, our work makes three contributions:

- We identify **compositional regulatory generalization** as a distinct capability beyond sequence recognition and activity prediction.
- We introduce a **matched-perturbation benchmark** to systematically evaluate regulatory relationships across local effects, pairwise interactions, contextual transfer, and higher-order combinations.
- We reveal a fundamental **dissociation between activity prediction and transferable regulatory relationship learning** across diverse genomic foundation models.

2 Related Work

Genomic foundation models and regulatory grammar. Genomic foundation models differ substantially in their pretraining signals and inductive biases. DNABERT and Nucleotide Transformer learn genomic sequence representations through masked language modeling objectives (Ji et al., 2021; Dalla-Torre et al., 2025), whereas HyenaDNA and Evo 2 adopt autoregressive sequence modeling to capture long-range genomic dependencies (Nguyen et al., 2023; Brixi et al., 2026). Caduceus incorporates bidirectional modeling and reverse-complement equivariance (Schiff et al., 2024), while PhyloGPN introduces phylogenetic information as an additional constraint for genomic representation learning(Albors et al., 2025).

These objectives provide different inductive biases for learning sequence representations, but they do not by themselves demonstrate that models have recovered functional regulatory relationships among interacting elements. In contrast, functionally supervised models such as BPNet and DeepSTARR have shown that sequence models can learn aspects of motif syntax, enhancer activity, and regulatory grammar from experimental functional measurements (Avsec et al., 2021a; de Almeida et al., 2022).

Genomic benchmarks and combinatorial evaluation. Existing benchmarks cover complementary aspects of genomic modeling. GUE standardizes sequence-classification tasks across species,

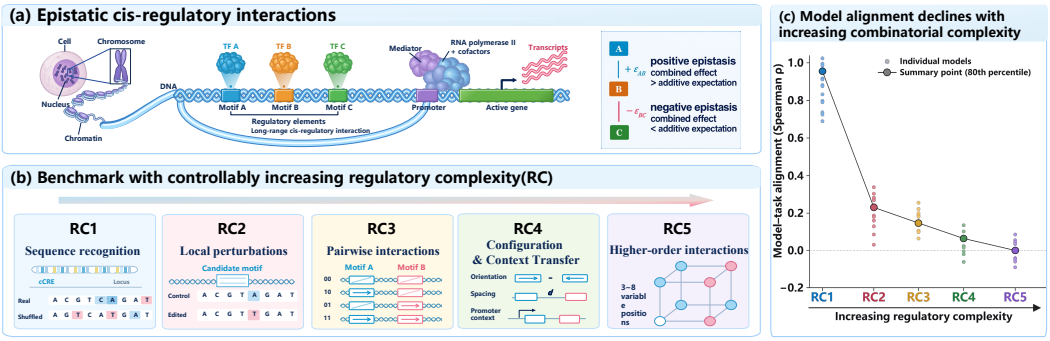


Figure 1: **Evaluating combinatorial regulatory generalization in genomic foundation models.** **(a)** Regulatory effects depend on interactions among elements and their surrounding context, such that accurate activity prediction does not guarantee prediction of combinatorial effects. **(b)** Our benchmark evaluates increasingly complex regulatory relationships, from sequence recognition (RC1) and local effects (RC2) to pairwise interactions (RC3), context dependence (RC4), and higher-order interactions (RC5). **(c)** Model agreement with experimentally measured effects decreases with increasing combinatorial complexity, revealing a gap between activity prediction and transferable regulatory relationships.

BEND emphasizes biologically meaningful annotation and variant prediction, and DART-Eval evaluates regulatory DNA through sequence discrimination, motif sensitivity, activity prediction, and variant effects (Zhou et al., 2024; Marin et al., 2024; Patel et al., 2024). Targeted diagnostics such as the Mechanistic Invariance Test further examine whether model scores follow predefined positional or compositional rules (Cheng and Zhang, 2026). These benchmarks substantially improve sequence-level and rule-based evaluation, but provide less direct evidence about experimentally measured non-additive effects and whether such effects transfer across regulatory backgrounds.

Interaction interpretation and combinatorial generalization. Methods such as GIA, CREME, and SQUID characterize interactions within sequence models by perturbing or decomposing their predictions (Koo et al., 2021; Toneyan and Koo, 2024; Seitz et al., 2024). These approaches provide important tools for probing model-internal regulatory logic, but inferred interactions do not necessarily establish agreement with experimentally measured functional effects. Moreover, non-additivity depends on the measurement scale, reference background, and null model (Tareen et al., 2022; Sailer and Harms, 2017; Poelwijk et al., 2016).

At the same time, strong regulatory prediction does not necessarily require foundation-model pretraining: k-mer methods and task-specific neural networks can achieve competitive performance on sequence-based regulatory tasks (Ghandi et al., 2014; Lee, 2016; Penzar et al., 2023). Generalization also depends critically on what information is withheld at test time. Perturbation-prediction studies such as GEARS distinguish generalization to unseen combinations from generalization to unseen perturbation components (Roohani et al., 2024). For regulatory sequence models, however, whether learned regulatory relationships remain predictive across new combinations, configurations, and regulatory backgrounds remains comparatively underexplored.

3 Matched-Perturbation Evaluation Framework

Our benchmark is designed as a hierarchical evaluation framework for compositional regulatory generalization. It evaluates genomic foundation models along two complementary dimensions: *regulatory capability* and *contextual generalization*. The first dimension defines a hierarchy of increasingly complex regulatory capabilities, progressing from sequence recognition (RC1), to local perturbation effects (RC2), pairwise interactions (RC3), and higher-order interactions among multiple regulatory features (RC5). The second dimension evaluates whether these learned regulatory relationships transfer across changing contexts, including element configuration and promoter context (RC4).

This hierarchy reflects a progression from recognizing regulatory sequences to predicting perturbation responses, modeling interactions among regulatory elements, and ultimately capturing context-dependent combinatorial effects. For each task family, we construct matched experimental comparisons and apply the same functional contrast to experimental measurements and model outputs. Therefore, the benchmark evaluates whether a model recovers the *effect or interaction induced by a perturbation*, rather than merely predicting the activity of individual sequences from which the effect is computed.

We next describe each task family and its corresponding generalization setting. Each task family is instantiated with a specific dataset and generalization protocol:

Sequence recognition (RC1). RC1 asks a basic question: can a model distinguish regulatory DNA from a matched sequence with similar nucleotide composition? We evaluate 489,492 pairs of 350-bp native candidate cis-regulatory elements (cCREs) and composition-matched shuffled controls (ENCODE Project Consortium, 2020). Because this task does not involve an experimentally measured change in regulatory function, it serves only as a reference for sequence recognition. Training and test data are separated by chromosome.

Local perturbation effects (RC2). RC2 asks whether a model responds appropriately when a short region of a regulatory sequence is altered. The dataset contains 1,611 core control–changed sequence pairs, measured in two cell lines, HepG2 and K562. For native pretrained scores, we measure the score change

$$\Delta Z = Z(x_{\text{changed}}) - Z(x_{\text{control}}). \tag{1}$$

We compare this change with the change in a candidate motif-matching score computed from a position weight matrix (PWM) (Stormo, 2000). This PWM comparison is used only as a reference for the direction and relative strength of the local sequence change; it is not an experimentally measured functional effect. For supervised readouts, predicted activity changes are instead compared directly with experimentally measured activity changes. Matched control–changed pairs always remain in the same data split.

Pairwise interactions (RC3). RC3 asks whether a model can predict what happens when two regulatory elements are changed together, beyond the effects of changing either element alone. For each pair, four experimental states are available: neither element changed, only the first changed, only the second changed, and both changed. Their pairwise interaction is defined as

$$\varepsilon_{AB}(f) = f_{11} - f_{10} - f_{01} + f_{00}, \tag{2}$$

where f denotes either experimental activity or the corresponding model output (Poelwijk et al., 2016; Sailer and Harms, 2017). A value different from zero indicates that the joint effect cannot be explained by simply adding the two individual effects (Poelwijk et al., 2016; Sailer and Harms, 2017). We compute exactly the same contrast for experiments and model predictions. All states from the same enhancer are kept together during train–test splitting.

Configuration and regulatory-context dependence (RC4). RC4 asks a different question: even if the regulatory elements themselves are the same, does their relationship remain predictable when their organization or surrounding context changes? RC4A tests changes in regulatory configuration. It contains 6,271 synthetic constructs and 14,509 matched comparisons in which only one factor changes: sequence background, orientation, or spacing between elements (Georgakopoulos-Soares et al., 2023). Comparisons involving the same transcription-factor pair are kept together during evaluation. RC4B tests whether the interaction between the same enhancer components changes predictably when they are placed with different promoters (Martinez-Ara et al., 2024). For each promoter P, we first estimate the activity expected from the two individual enhancers,

$$A_{\text{add}}(P) = A_{EC}(P) + A_{CE}(P) - A_{CC}(P), \tag{3}$$

and compare this expectation with the measured activity when both enhancers are present:

$$I(P) = \log_2 \frac{A_{EE}(P)}{A_{\text{add}}(P)}. \tag{4}$$

We then ask how this interaction changes between two promoters $\Delta_P = I(P_2) - I(P_1)$. The same calculation is applied to predicted activities. Thus, RC4B asks not only whether a model

predicts each construct correctly, but whether it predicts how an enhancer interaction changes when the promoter context changes. We consider two generalization settings. In the first, new enhancer combinations and configurations are held out, but the promoters may have been seen during training; we refer to this as *PairConfig-OOD*. In the second, an entire promoter is unseen during training; we refer to this as *leave-one-promoter-out* (LOPO). The latter therefore provides the stronger test of transfer to a new regulatory context.

Higher-order interactions (RC5). RC5 extends the same idea beyond pairs to interactions among three to eight regulatory features. For k features, the experimental data contain all 2^k possible combinations of their two states (Sailer and Harms, 2017). Having the complete set of combinations allows us to separate the genuinely kth-order interaction from effects already explained by individual features and lower-order interactions. The higher-order contrast is

$$\varepsilon^{(k)}(f) = \sum_{s \in \{0,1\}^k} (-1)^{k - \sum_i s_i} f(x_s), \tag{5}$$

where f denotes experimental activity, predicted activity, or a native model score. The full benchmark contains 59,851 complete higher-order perturbation sets spanning interactions among three to eight regulatory features. Because many sets reuse the same sequence or biological parent, related examples are grouped together during train–test splitting rather than treated as independent observations.

4 Experiments

We organize our results around five research questions (RCs): sequence recognition (RC1), local perturbation responses (RC2), pairwise interactions (RC3), contextual transfer (RC4), and higher-order combinations (RC5). These questions progressively evaluate whether genomic foundation models move beyond sequence recognition toward compositional regulatory generalization. Native-score evaluation measures information directly expressed by pretrained objectives, whereas frozen-representation evaluation assesses recoverable functional information through supervised readouts.

We evaluate 14 pretrained configurations, including eleven genomic language/phylogenetic models and three functionally pretrained models (Borzoi, Enformer, and AlphaGenome) (Linder et al., 2025; Avsec et al., 2021b; 2026). Native scores and frozen representations are evaluated separately. Details of checkpoints, input adaptations, and scoring protocols are provided in Appendix B. We compare against matched zero-effect/zero-interaction baselines and task-specific predictors, including k-mer models and CNNs trained from scratch. RC2 localization additionally includes structural controls. All comparisons use task-specific splits and matched valid subsets; protocol details are provided in Appendices B.3–B.5.

4.1 From sequence recognition to local perturbation effects

RC1: Sequence recognition is strong but not specific to pretrained models. Before evaluating compositional regulatory generalization, we first establish whether genomic foundation models capture sequence-level regulatory signatures. RC1 provides a reference capability: distinguishing native regulatory sequences from composition-matched controls without requiring functional perturbation prediction.

On the common zero-shot set of 30,000 real–shuffled pairs, the eleven genomic/phylogenetic models achieve paired accuracies of 67.45%–97.17%, with Caduceus-Ps attaining the highest value (Figure 2(a)). Frozen-representation readouts further improve sequence discrimination on the full 489,492-pair test set, with mean-linear and positional-CNN predictors reaching up to 94.65% and 96.33% paired accuracy, respectively (Figure 2(b,c)). The positional-CNN analysis shows that this performance is consistent across cCRE classes, changed-base fractions, and test chromosomes (Figure 2(d–f)), with only small variation across random seeds (Figure 2(h)).

However, the gain from supervised readouts reflects both representation quality and task-specific fitting. A from-scratch MPRA-LegNet achieves 96.02% accuracy, comparable to pretrained representations (Figure 2(g,i)), showing that strong sequence recognition is not unique to large-scale pretraining.

RC1 therefore confirms that current models encode substantial sequence-level information, but this capability alone does not establish whether they recover functional perturbation effects or regulatory interactions. Complete RC1 results are reported in Appendix D.1, Table D1.

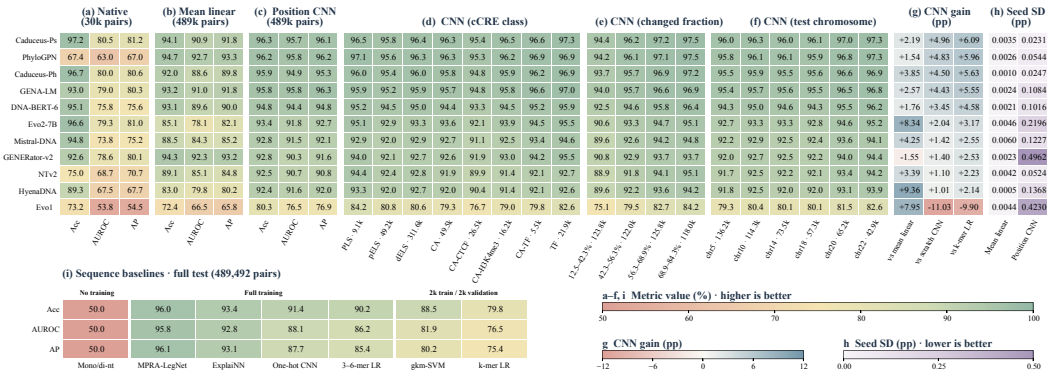


Figure 2: **RC1 sequence recognition and readout comparisons**. (a) Native scoring on the common 30,000-pair subset. (b,c) Supervised mean-linear and positional-CNN readouts on the full 489,492-pair test set. (d–f) Positional-CNN paired accuracy stratified by cCRE class, changed-base fraction, and test chromosome. (g) Paired-accuracy gain of positional CNNs over alternative readouts or sequence baselines. (h) Across-seed standard deviation. (i) Sequence baselines under no-training, full-training, and matched-budget settings.

RC2: Native sequence sensitivity only partially tracks local perturbation effects. Sequence recognition alone does not require models to predict how regulatory function changes after controlled perturbation. We therefore next evaluate whether model representations capture local consequences of sequence changes. Across 159 matched perturbation pairs, native-score changes are compared with changes in a fixed candidate PWM score. Direction-discrimination AUROC ranges from 0.507 to 0.714, whereas signed effect-strength correlations remain weak (Spearman $\rho = 0.006$–0.305; Figure 3(a); Appendix Table D2). Different models lead direction discrimination and effect-strength ranking, indicating that sensitivity to local edits does not yield a consistent representation of perturbation magnitude.

Response localization provides an additional but distinct diagnostic. Caduceus-Ps reaches localization AUROC/AP of 0.933/0.805, yet a center-position prior reaches 0.997/0.978 (Figure 3(a)). Localization therefore indicates where model scores respond, rather than whether models recover the functional role of edited regions. In contrast, supervised readouts from functionally pretrained representations recover modest experimentally measured activity effects. Mean-Ridge readouts achieve correlations of 0.270, 0.287, and 0.315 for Borzoi, Enformer, and AlphaGenome, respectively, exceeding k-mer Ridge (0.101) and a CNN trained from scratch (0.110; Figure 3(b); Appendix Table D3). These results motivate the next question: whether local regulatory information composes into interactions among multiple regulatory elements.

4.2 Pairwise interactions are distinct from single-element preference

RC3: Preference for individual motifs does not reliably predict pairwise non-additivity. Having established that models capture sequence-level information and some local perturbation responses, we next ask whether these signals compose when multiple regulatory elements act together. Pairwise interaction prediction provides the first direct test of compositional regulatory generalization, because it requires recovering effects that cannot be explained by individual element contributions alone.

Across 102 single-motif comparisons, intact-motif preference ranges from 46.1% to 67.6%. However, across 20 complete four-state families, correlations between native-score interactions and experimentally measured pairwise interactions range from -0.398 to 0.386 (Figure 4(a–c); Appendix Table D4). The difference between these capabilities is evident within individual models: Evo2-7B achieves the highest intact-motif preference (67.6%) but an interaction correlation of only 0.150,

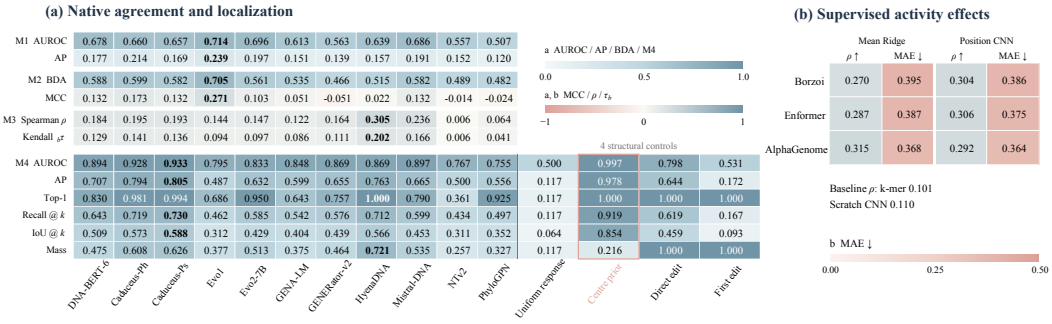


Figure 3: **RC2 local perturbation diagnostics and supervised activity effects**. (a) Native-score agreement across direction discrimination, fixed-threshold direction agreement, signed effect-strength ranking, and response localization for 159 matched perturbation pairs; structural localization controls are shown alongside model results. (b) Experimental activity-effect prediction from frozen Borzoi, Enformer, and AlphaGenome representations using mean-Ridge and positional-CNN readouts; cells report Spearman ρ and MAE, with k-mer Ridge and a from-scratch CNN as sequence baselines.

whereas Caduceus-Ph and Caduceus-Ps achieve the same intact-motif preference (59.8%) but substantially different interaction correlations of 0.386 and -0.141, respectively. Thus, recognizing individual regulatory elements does not imply that models recover the non-additive relationships generated when elements interact.

This limitation also persists under functional supervision. Under enhancer holdout, Borzoi, Enformer, and AlphaGenome achieve pairwise-interaction correlations of -0.191, 0.211, and 0.068, with MAEs of 1.456, 1.370, and 1.382, respectively. All three MAEs exceed the matched zero-interaction baseline of 1.366 (Appendix Table D5). Since pairwise interactions represent residual effects beyond individual regulatory contributions, accurate single-element prediction alone is insufficient to recover these interaction terms.

4.3 Contextual generalization exposes the activity–interaction gap

RC4A: Configuration-dependent effects are only weakly reflected by native scores. Regulatory interactions are not only compositional but also context dependent: the same regulatory elements may behave differently under altered configurations or surrounding backgrounds. We therefore evaluate whether models preserve regulatory relationships when element arrangement changes. Across background, orientation, and spacing perturbations, four-factor macro BDA ranges from 0.491 to 0.525, while macro Spearman correlations range from -0.050 to 0.077 (Figure 4(d,e); Appendix Table D6). Frozen functional representations recover additional signal, but do not provide a consistent advantage over simple task-trained sequence baselines. AlphaGenome mean Ridge achieves $\rho = 0.516$ and MAE$= 0.172$, comparable to k-mer Ridge ($0.517/0.174$), while a one-hot CNN achieves stronger performance ($0.544/0.166$) (Figure 4(f); Appendix Table D3). Standardization sensitivity is reported separately in Appendix C.3.

RC4B: Accurate construct activity does not imply transfer of interaction changes. A stringent test of regulatory generalization is whether learned regulatory relationships remain predictive when the surrounding biological context changes. We therefore evaluate transfer of interaction effects across unseen regulatory backgrounds rather than activity prediction alone. On the legacy RC4B build, ten genomic/phylogenetic frozen readouts achieve high construct-activity prediction under PairConfig-OOD ($R^2 = 0.884$–0.917), where enhancer combinations and configurations are held out but promoters remain represented during training. However, correlations for promoter-dependent interaction changes are only 0.263–0.327, and all models perform worse than the matched zero-change baseline in MAE. Under LOPO, where entire promoters are unseen, interaction-change correlations further decrease to -0.090–0.064 (Figure 4(g,h); Appendix Table D7). Thus, accurate prediction of regulatory states does not imply transfer of the relationships that govern how regulatory elements combine.

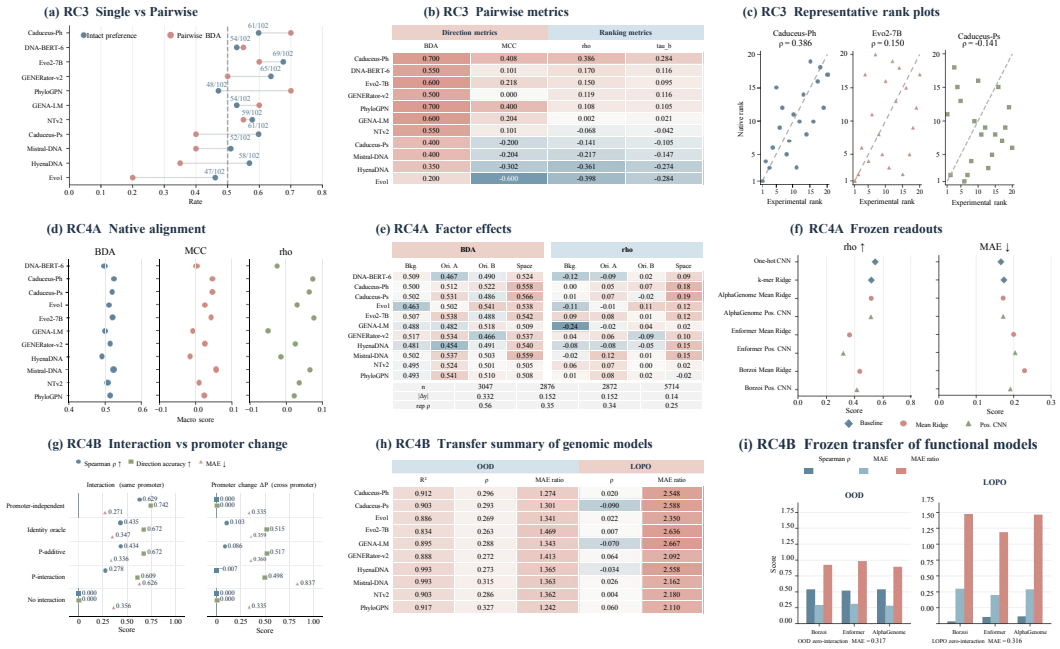


Figure 4: **Pairwise interactions and contextual generalization**. (a–c) RC3 evaluates whether single-element preferences extend to pairwise non-additive interactions. (d–f) RC4A evaluates native and supervised prediction of configuration-dependent regulatory effects across background, orientation, and spacing changes. (g–i) RC4B evaluates whether learned regulatory relationships transfer across promoter contexts.

Functionally pretrained models are evaluated on the corrected-v2 RC4B build and are therefore not directly comparable with genomic/phylogenetic models. On their common valid subsets, Borzoi, Enformer, and AlphaGenome improve MAE over the zero-change baseline under PairConfig-OOD, but this advantage disappears under LOPO: correlations decrease to 0.032–0.111 and MAEs increase to 0.454–0.545, above the zero-change MAE of 0.316 (Figure 4(i); Appendix Table D8). These results indicate that functional pretraining can improve in-context prediction but does not guarantee transfer of regulatory relationships across biological contexts.

4.4 Higher-order combinations show non-transferable gains

RC5: Higher-order interaction prediction remains inconsistent across sources. Finally, we extend evaluation beyond pairwise interactions to combinations of multiple regulatory features. Higher-order effects are quantified using alternating contrasts that remove lower-order contributions (Figure 5(a)). The primary motif/module panel contains 721 complete combinatorial cubes, including King third- and fourth-order interactions, e2MPRA third-order interactions, and two small QeFS third-order conditions. Native-score alignment remains weak and source dependent. For King, correlations range from -0.230 to 0.025 for $k = 3$ and from -0.255 to 0.320 for $k = 4$, while e2MPRA $k = 3$ ranges from -0.100 to 0.126 (Figure 5(d,e); Appendix Table D9). No model is consistently strongest across sources or interaction orders. Functionally pretrained models show similar limitations in interaction-magnitude prediction. None of Borzoi, Enformer, or AlphaGenome improves over the zero-interaction baseline for King $k = 3$ or e2MPRA $k = 3$; for King $k = 4$, Borzoi and AlphaGenome achieve only small point-estimate reductions (MAE ratios 0.992 and 0.994, respectively) (Figure 5(b,f); Appendix Table D10). Extensions to additional sources and orders show the same heterogeneous pattern (Figure 5(c)). The two QeFS conditions contain only three and five cubes and are interpreted descriptively.

Across expanded non-N8 strata, both better- and worse-than-zero estimates occur. For example, the large Yu fourth-order sets contain 25,886 overlapping contrasts per condition, yet native correlations remain near zero and all three functional models have MAE ratios above one. We therefore inter-

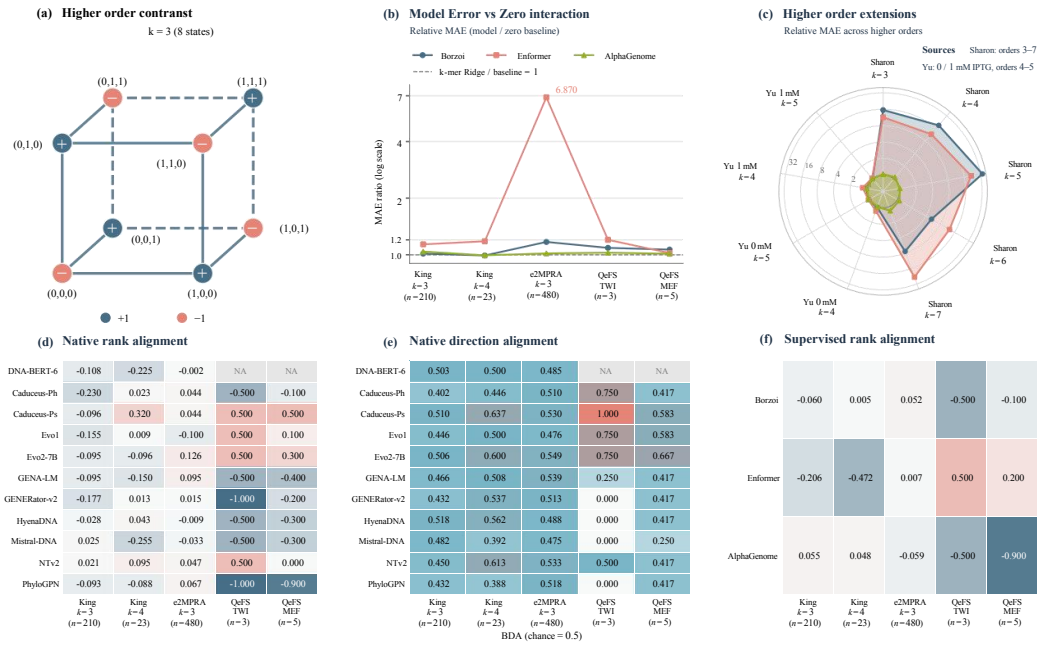


Figure 5: **Higher-order regulatory interactions across sources and model classes.** (a) Complete $k = 3$ factorial contrast: eight states enter with alternating signs to isolate the third-order interaction. (b) Supervised MAE relative to the zero-interaction baseline on the primary motif/module panel; values below 1 indicate improvement. (c) Relative MAE for higher orders in Sharon and Yu (logarithmic radial scale). (d,e) Native-score Spearman correlation and parent-macro BDA across King, e2MPRA, and QeFS strata; chance BDA is 0.5. (f) Supervised Spearman correlation for Borzoi, Enformer, and AlphaGenome.

pret RC5 at the source and background level rather than constructing an overall ranking (Appendix Tables D11 and D12).

Finally, N8 extends evaluation to nucleotide-level interactions from third to eighth order. Across twelve test-library–order strata, mean native Spearman correlations range from -0.078 to 0.044 and mean BDA from 0.467 to 0.527. Under cross-library supervision, Borzoi, Enformer, and AlphaGenome achieve mean correlations of 0.029, 0.096, and 0.143, but MAE-to-zero ratios of 1.077, 1.097, and 1.007, respectively (Appendix Figure D1). Together, these results indicate that higher-order signals may appear in individual strata, but current representations do not provide consistent transferable advantages across sources, contexts, and interaction orders.

5 Discussion and Conclusion

Our results reveal a clear separation between activity prediction and compositional regulatory generalization. Genomic foundation models capture sequence-level information and some local effects, but these abilities do not consistently transfer to predicting regulatory interactions across contexts. Strong activity prediction can therefore coexist with poor prediction of interaction changes under unseen backgrounds, indicating that compositional generalization is a distinct capability beyond standard activity prediction. Matched perturbation benchmarks are thus required to evaluate functional changes, interactions, and contextual transfer. Although our conclusions are limited by the evaluated systems and measurement variability, they highlight a gap between learning genomic sequence patterns and transferable regulatory principles. Future models may require objectives that explicitly capture perturbation responses, interactions, and context-dependent regulation.

REFERENCES

Carlos Albors, Jianan Canal Li, Gonzalo Benegas, Chengzhong Ye, and Yun S. Song. A Phylogenetic Approach to Genomic Language Modeling. arXiv preprint arXiv:2503.03773v1, 2025. URL https://arxiv.org/abs/2503.03773v1. Version 1.

Ž. Avsec et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53:354–366, 2021a. doi: 10.1038/s41588-021-00782-6. URL https://doi.org/10.1038/s41588-021-00782-6.

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, 2021b. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x.

Žiga Avsec et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649: 1206–1218, 2026. doi: 10.1038/s41586-025-10014-0. URL https://www.nature.com/articles/s41586-025-10014-0.

G. Brixi et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652: 1349–1361, 2026. doi: 10.1038/s41586-026-10176-5. URL https://doi.org/10.1038/s41586-026-10176-5.

Bryan Cheng and Jasper Zhang. The Mechanistic Invariance Test: Genomic Language Models Fail to Learn Positional Regulatory Logic. arXiv preprint arXiv:2604.06549v1, 2026. URL https://arxiv.org/abs/2604.06549v1. Version 1.

H. Dalla-Torre et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22:287–297, 2025. doi: 10.1038/s41592-024-02523-z. URL https://doi.org/10.1038/s41592-024-02523-z.

B. P. de Almeida, F. Reiter, M. Pagani, and A. Stark. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, 54:613–624, 2022. doi: 10.1038/s41588-022-01048-5. URL https://doi.org/10.1038/s41588-022-01048-5.

ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in human and mouse genomes. *Nature*, 583:699–710, 2020. doi: 10.1038/s41586-020-2493-4. URL https://doi.org/10.1038/s41586-020-2493-4.

V. Fishman et al. GENA-LM: a family of open-source foundational DNA language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310, 2025. doi: 10.1093/nar/gkae1310. URL https://doi.org/10.1093/nar/gkae1310.

I. Georgakopoulos-Soares, C. Deng, V. Agarwal, C. S. Y. Chan, J. Zhao, F. Inoue, and N. Ahituv. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature Communications*, 14:2333, 2023. doi: 10.1038/s41467-023-37960-5. URL https://doi.org/10.1038/s41467-023-37960-5.

M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLOS Computational Biology*, 10(7):e1003711, 2014. doi: 10.1371/journal.pcbi.1003711. URL https://doi.org/10.1371/journal.pcbi.1003711.

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. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002. doi: 10.1101/gr.229102. URL https://doi.org/10.1101/gr.229102.

P. Kheradpour et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5):800–811, 2013. doi: 10.1101/gr.144899.112. URL https://doi.org/10.1101/gr.144899.112.

Dana M. King, Clarice Kit Yee Hong, James L. Shepherdson, David M. Granas, Brett B. Maricque, and Barak A. Cohen. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *eLife*, 9:e41279, 2020. doi: 10.7554/eLife.41279. URL https://doi.org/10.7554/eLife.41279.

P. K. Koo, A. Majdandzic, M. Ploenzke, P. Anand, and S. B. Paul. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLOS Computational Biology*, 17:e1008925, 2021. doi: 10.1371/journal.pcbi.1008925. URL https://doi.org/10.1371/journal.pcbi.1008925.

D. Lee. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics*, 32(14):2196–2198, 2016. doi: 10.1093/bioinformatics/btw142. URL https://doi.org/10.1093/bioinformatics/btw142.

Qiuyi Li, Zhihao Zhan, Shikun Feng, Yiheng Zhu, Yuan He, Wei Wu, Zhenghang Shi, Shengjie Wang, Zongyong Hu, Zhao Yang, Jiaoyang Li, Jian Tang, Haiguang Liu, and Tao Qin. GENERator-v2: Reconciling Coarse Tokenization with Single-Nucleotide Resolution in Genomic Language Modeling. *bioRxiv*, page 2026.01.27.702015, 2026. doi: 10.64898/2026.01.27.702015. URL https://doi.org/10.64898/2026.01.27.702015. Preprint.

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

Frederikke Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/429e7b31625a8b7839f9e4d6e2aa9bb9-Abstract-Conference.html.

M. Martinez-Ara, F. Comoglio, and B. van Steensel. Large-scale analysis of the integration of enhancer-enhancer signals by promoters. *eLife*, 12:RP91994, 2024. doi: 10.7554/eLife.91994.3. URL https://doi.org/10.7554/eLife.91994.3. Version of Record, published 28 October 2024.

E. Nguyen, M. Poli, M. G. Durrant, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024. doi: 10.1126/science.ado9336. URL https://doi.org/10.1126/science.ado9336.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Ré, and Stephen Baccus. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. doi: 10.52202/075280-1872. URL https://papers.nips.cc/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.

G. Novakovsky, O. Fornes, M. Saraswat, et al. ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biology*, 24:154, 2023. doi: 10.1186/s13059-023-02985-y. URL https://doi.org/10.1186/s13059-023-02985-y.

Aman Patel, Arpita Singhal, Austin Wang, Anusri Pampari, Maya Kasowski, and Anshul Kundaje. DART-Eval: A Comprehensive DNA Language Model Evaluation Benchmark on Regulatory DNA. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/71998bfc3217ffe1cca1ee084dfadadd-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

D. Penzar, D. Nogina, et al. LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics*, 39(8):btad457, 2023. doi: 10.1093/bioinformatics/btad457. URL https://doi.org/10.1093/bioinformatics/btad457.

Frank J. Poelwijk, Vinod Krishna, and Rama Ranganathan. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLOS Computational Biology*, 12(6):e1004771, 2016. doi: 10.1371/journal.pcbi.1004771. URL https://doi.org/10.1371/journal.pcbi.1004771.

I. Rauluseviciute, R. Riudavets-Puig, R. Blanc-Mathieu, et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1): D174–D182, 2024. doi: 10.1093/nar/gkad1059. URL https://doi.org/10.1093/nar/gkad1059.

Franziska Reiter, Bernardo P. de Almeida, and Alexander Stark. Enhancers display constrained sequence flexibility and context-specific modulation of motif function. *Genome Research*, 33(3): 346–358, 2023. doi: 10.1101/gr.277246.122. URL https://doi.org/10.1101/gr.277246.122.

Y. Roohani, K. Huang, and J. Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Zachary R. Sailer and Michael J. Harms. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics*, 205:1079–1088, 2017. doi: 10.1534/genetics.116.195214. URL https://doi.org/10.1534/genetics.116.195214.

Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43632–43648. PMLR, 2024. URL https://proceedings.mlr.press/v235/schiff24a.html.

Evan E. Seitz, David M. McCandlish, Justin B. Kinney, and Peter K. Koo. Interpreting cis-regulatory mechanisms from genomic deep neural networks using surrogate models. *Nature Machine Intelligence*, 6:701–713, 2024. doi: 10.1038/s42256-024-00851-5. URL https://doi.org/10.1038/s42256-024-00851-5.

Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012. doi: 10.1038/nbt.2205. URL https://doi.org/10.1038/nbt.2205.

Gary D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 2000.

Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, et al. MAVE-NN: learning genotype–phenotype maps from multiplex assays of variant effect. *Genome Biology*, 23:98, 2022. doi: 10.1186/s13059-022-02661-7. URL https://doi.org/10.1186/s13059-022-02661-7.

Shushan Toneyan and Peter K. Koo. Interpreting cis-regulatory interactions from large-scale deep neural networks. *Nature Genetics*, 56:2517–2527, 2024. doi: 10.1038/s41588-024-01923-3. URL https://doi.org/10.1038/s41588-024-01923-3.

C. T. Waters, S. S. Gisselbrecht, Y. A. Sytnikova, T. M. Cafarelli, D. E. Hill, and M. L. Bulyk. Quantitative-enhancer-FACS-seq (QeFS) reveals epistatic interactions among motifs within transcriptional enhancers in developing Drosophila tissue. *Genome Biology*, 22:348, 2021. doi: 10.1186/s13059-021-02574-x. URL https://doi.org/10.1186/s13059-021-02574-x.

Sean Whalen, Fumitaka Inoue, Hane Ryu, et al. Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron*, 111(6):857–873.e8, 2023. doi: 10.1016/j.neuron.2022.12.026. URL https://doi.org/10.1016/j.neuron.2022.12.026.

Timothy C. Yu, Winnie L. Liu, Marcia S. Brinck, Jessica E. Davis, Jeremy Shek, Grace Bower, Tal Einav, Kimberly D. Insigne, Rob Phillips, Sriram Kosuri, and Guillaume Urtecho. Multiplexed characterization of rationally designed promoter architectures deconstructs combinatorial logic for IPTG-inducible systems. *Nature Communications*, 12:325, 2021. doi: 10.1038/s41467-020-20094-3. URL https://doi.org/10.1038/s41467-020-20094-3.

Zicong Zhang, Ilias Georgakopoulos-Soares, Guillaume Bourque, et al. Simultaneous epigenomic profiling and regulatory activity measurement using e2MPRA. *Nature Communications*, 17:1724, 2026. doi: 10.1038/s41467-026-68422-3. URL https://doi.org/10.1038/s41467-026-68422-3.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b633e7052970b8f5aa1a69164d99e9e8-Abstract-Conference.html.

APPENDIX

This appendix provides additional details supporting the construction, reproducibility, and interpretation of the matched-perturbation benchmark and the corresponding model evaluations. It supplements the main text with data provenance and label construction, task-specific generalization protocols, model and readout configurations, reliability and sensitivity analyses, and complete numerical results. Appendix A documents datasets, label construction, and generalization splits; Appendix B specifies model configurations and evaluation protocols; Appendix C reports reliability and sensitivity analyses; and Appendix D provides additional numerical results.

A DATASET CONSTRUCTION AND GENERALIZATION SPLITS

A.1 DATA SOURCES AND STATISTICAL UNITS

Data are organized into three levels: raw source files, construct tables, and matched-relation tables. Raw files retain source provenance and dataset version; construct tables store the evaluated sequences together with measurement scale, cellular condition, and quality-control fields; matched-relation tables link constructs through stable IDs and encode the controlled perturbation, interaction, or context change used for evaluation. This separation allows label construction and matched contrasts to be audited independently of model predictions. Pairing, state encoding, and label recomputation are performed without using model outputs. Test labels are not used to select edit windows, impute missing controls, alter contrast direction, or redefine evaluation groups. Because multiple matched relations can share the same constructs or biological background, the number of evaluated relations should not be interpreted as the number of independent biological samples. Task-specific splitting and resampling therefore operate on the corresponding biological grouping units rather than on individual relation rows whenever required. Tables A1 and A2 summarize data sources, evaluation populations, and grouping units; RC5 source coverage is detailed separately in Table A4.

Table A1: Primary data sources and recovered inputs.

Task	Publication / accession	Data used
RC1	(Patel et al., 2024; ENCODE Project Consortium, 2020); DART-Eval Task 1; syn60581046	Official `data.h5`; ENCODE `ENCFF420VPZ` cCREs and GRCh38 no-alt reference.
RC2	(Kheradpour et al., 2013); GSE33367	Six RNA/DNA count tables (GSM825352–357), 145 bp sequences; candidate motifs from JASPAR (Rauluseviciute et al., 2024).

Continued on next page

Task	Publication / accession	Data used
RC3	(Waters et al., 2021); GSE149906	QeFS Tables S3/S4 and filtered activity; GSE149908 for tag/clone validation.
RC4A	(Georgakopoulos-Soares et al., 2023); PRJNA854975	`Library_MPRA_TFBSs.txt` and three insert-level replicate files.
RC4B	(Martinez-Ara et al., 2024); GSE240586; OSF yuc4e	Supplementary files 1–3; EEPCombinations; `activity_all`; and mm10 component sequences (Kent et al., 2002).

Table A2: Evaluation populations and grouping units.

Comparison and cube counts are not independent biological-replicate counts. RC4B versions and prediction-validity subsets are kept separate.

Task	Retained sequences / constructs	Evaluation relations	Holdout unit
RC1	2,348,786 pairs, each 350 bp	489,492 test pairs; 30,000-pair common native subset	Chromosomes; paired controls remain together
RC2	1,739 pairs, each 145 bp; 1,611 core	159 core test pairs; 174 expanded test pairs	Reconstructed pair parent, not author-defined enhancer family
RC3	95 constructs; 81 with valid activity	102 motif edges; 20 four-state families from 11 enhancers	Enhancer; nested leave-one-out
RC4A	6,271 constructs with ≥ 2 valid replicates; 6,263 with three	14,509 comparisons; 153 unordered TF pairs	Unordered TF pair; five folds
RC4B	110,180 constructs in each build	273,736 legacy / 210,173 v2 biologically valid comparisons	Eight-fold LOPO or five-fold PairConfig-OOD
RC5	55,934 unique DNA sequences	59,851 complete cubes; 38 strata; 62 leakage groups	Shared DNA, reverse complements, and parent families; N8 cross-library transfer

A.2 RC1: Paired Shuffled Controls

RC1 is a QC-filtered reuse of DART-Eval Task 1 (Patel et al., 2024), rather than a newly constructed recognition dataset. DART-Eval constructs 350-bp windows centered on ENCODE candidate cis-regulatory elements (cCREs) (ENCODE Project Consortium, 2020) and shuffles only the portion of each window overlapping the target cCRE, while preserving dinucleotide composition and the surrounding sequence context. Each real–control pair also retains the same recorded reverse-complement transformation. The resulting controls should therefore not be interpreted as uniform full-window shuffles or as experimentally validated nonfunctional sequences. We retain the original DART-Eval pair-generation procedure without redefining the controls.

The official seqs, ctrls, and idxs arrays preserve paired region indices and the original train/validation/test splits. After removing 68 pairs with invalid sequence encoding (40/10/18 from the train/validation/test sets), 1,691,618/167,676/489,492 pairs remain; we refer to this QC-filtered export as DART1_QC_tutorial_v1. All retained real–control pairs are nonidentical and have matched mono- and dinucleotide counts, although higher-order sequence statistics may still differ and remain available for discrimination. Native-score results on the common 30,000-pair subset and on the full test set are reported separately. These composition checks do not establish cross-split homology isolation and do not exclude possible overlap with model pretraining data.

A.3 RC2: Pair Reconstruction and Local-Effect Labels

RC2 uses the human enhancer MPRA measurements of Kheradpour et al. in HepG2 and K562. Six count tables provide two mRNA measurements per cell type and two plasmid-DNA measurements, joined by sequence ID and barcode.

Reconstructed pairs and limitations. Because the complete author-supplied design mapping was unavailable, pairs were recovered from 5,418 experimentally measured sequence records. Candidates have equal lengths, a minimal edit interval of 3–20 bp, and exact identity outside that interval. Hamming-nearest neighbors are ordered deterministically by distance, edit interval, and ID; one-to-one matching prevents sequence reuse across pairs. This yields 1,739 pairs: 1,611 composition-preserving, scramble-like pairs in the core set and 128 additional local-substitution candidates. These reconstructed pairs are not asserted to reproduce the authors' enhancer-family mapping.

Activity and contrast direction. For sequence i, cell type c, and replicate r, RNA and DNA counts are summed over the shared barcode set $B_{i,c,r}$:

$$y_{i,c,r} = \log_2 \frac{1 + \sum_{b \in B_{i,c,r}} R_{i,c,r,b}}{1 + \sum_{b \in B_{i,c,r}} D_{i,r,b}}.$$

The label Δ_c is the mean of $y_{\text{changed},c,r} - y_{\text{control},c,r}$ over shared valid replicates. This pseudocount, aggregation, and replicate matching define the reconstructed labels; the original authors' complete sequencing-depth normalization is not reproduced. `control/changed` is assigned by lexicographic source-ID order, independently of the measured effect, and does not automatically mean WT/mutant.

Candidate PWMs. GRCh38 matching and motif scans add provenance and candidate annotations rather than experimental validation. A fixed PWM, interval, and strand are used for both members; negative-strand intervals are reverse-complemented, and sites are not reselected separately after perturbation. The exact matrices are HNF1: MA0046.3; HNF4: MA0114.4; FOXA: MA0148.4; GATA: MA0035.5; NFE2L2: MA0150.2; GFI1: MA0038.2; and ZFP161: UN0482.1. Position counts C are converted to log-odds with a 0.5 pseudocount per base and a uniform background:

$$w_{b,t} = \log_2 \frac{(C_{b,t} + 0.5)/(\sum_{b'} C_{b',t} + 2)}{0.25}, \qquad W(x) = \sum_t w_{x_t,t}.$$

Only windows spanning the edit interval are scanned on either strand. Scores are normalized by each matrix's theoretical minimum and maximum; the highest candidate is retained, with relative score ≥ 0.80 marking a high-score annotation. Of all 1,739 pairs, 469 meet this threshold and 1,270 remain unresolved. Reference support is exact for 1,607 pairs; the other 132 are non-exact, while 17 pairs have ambiguous independent locus hits. These annotations do not establish recovery of the original WT designs.

Primary native evaluation uses the 159 core test pairs once each, with candidate intervals of 12–22 bp (39 high-score and 120 low-score candidates). It compares $\Delta Z = Z(x_{\text{changed}}) - Z(x_{\text{control}})$ with the fixed ΔW, which has 18 positive and 141 negative values. The 15 additional test substitutions are excluded. Functional readouts instead train on core activity labels and evaluate activity changes in the same 159 parents separately in HepG2 and K562, followed by an equal-cell average. PWM alignment and experimental-effect prediction are distinct endpoints.

A.4 RC3: QeFS State Mapping and Pairwise Labels

RC3 uses QeFS in the fixed Drosophila embryo condition `TWICRC2PLUS`. Sequences come from Table S3; activity is `avgUMIcount` in `TWICRC2PLUS_fullSummaryFiltered`, cross-checked against Table S4. The original-scale `standardError` and tag/day records are retained. The central enhancer is extracted from the verified structure "8 nt barcode + GCTAGC + enhancer + 8 nt reverse-complement barcode", so barcodes are not model inputs.

Mutation suffixes map D/S/Z/1 to Deaf1/Schlank/ZIPIC/CG12236-PB; GEO's 7 is normalized to Z only within suffixes. Each series assigns its three motifs in that fixed order to A/B/C. State 1 denotes intact and 0 perturbed. Of 95 constructs, 81 have valid activity. Strict QC retains 20 complete four-state families from 11 enhancers; they reuse 58 distinct constructs, up to three times each. The 102 single-motif edges come from 12 enhancers. The third motif, where present, is fixed within a family.

$$\varepsilon_{AB}^{\text{exp}} = y_{11} - y_{10} - y_{01} + y_{00}, \quad y = \log_2 A, \qquad \varepsilon_{AB}^{Z} = \bar{S}_{11} - \bar{S}_{10} - \bar{S}_{01} + \bar{S}_{00}.$$

Here $\bar{S}$ is native score per biological base. Supervised readouts first predict activity in each state and then apply the same contrast. Missing activities are not imputed. Expected tags, QC-passing

tags, and tag-by-day rows are different measurement levels. The original Table S4 construct-level Conover–Iman tests are not tests of these four-state contrasts, and original-scale SE does not determine covariance between shared log-scale differences. Consequently, native interaction statistics are descriptive.

A.5 RC4A: Configuration Matching

The synthetic lentiMPRA library contains 18 TFBS classes, two neutral backgrounds, and three HepG2 measurements. We retain designs with one copy of each of two distinct TFBSs and no third inserted TFBS. The actual 230 bp tested sequences are used; the description "200 bp neutral background" is not an input-cropping rule. Insert IDs are joined to the three replicate tables through the source `seq_N` convention rather than row-wise concatenation.

A valid replicate requires source ratio > 0 and at least three tags. Activity is the mean of replicate-level $\log_2(\text{ratio})$. Of 6,472 parsed designs, 6,271 have at least two valid replicates and 6,263 have three. Activity readouts use the former set; both endpoints of a matched comparison require three replicates. Comparisons change only one configuration attribute, retain one fixed direction, and use $y_2 - y_1$. Palindromic CREB1 pairs are excluded from both orientation tracks.

Formal evaluation contains 14,509 comparisons: background (3,047), orientation A (2,876), orientation B (2,872), and spacing (5,714). A/B denote source position-dependent identities; grouping uses 153 unordered TF pairs rather than 306 ordered source names. Canonical identity normalization recovers 2,998 order controls, but these were used only for data auditing, not model evaluation. The reported results are therefore four-factor, not five-factor, assessments.

A.6 RC4B: Component Recovery, Labels, and Dataset Versions

RC4B uses the authors' normalized `activity_all` measurements and three biological replicates from enhancer–enhancer–promoter reporters in mouse embryonic stem cells. BED coordinates, control FASTA, and strand annotations recover 166 orientation-specific components from UCSC mm10. Connecting fragments and complete reporter assemblies cannot be uniquely recovered. Results therefore use component-level supervised inputs; native whole-reporter zero-shot scoring is N/A.

CC, EC, CE, and EE denote control–control, position-specific single-enhancer, and joint-enhancer configurations. Lower-order support records are aggregated by median within the same promoter and relevant source element/position. Joint configurations retain element order and orientation; support aggregation follows the source definitions and is not asserted to match every joint construct's orientation individually. For promoter P,

$$A_{\text{add}}(P) = A_{E_1CP} + A_{CE_2P} - A_{CCP}, \quad I(P) = \log_2 \frac{A_{E_1E_2P}}{A_{\text{add}}(P)}, \quad \Delta_P = I(P_2) - I(P_1).$$

Both observed and additive activities must be finite and positive. Adding on the original activity scale before taking a log-ratio differs from the QeFS log-scale finite-difference null. The multiplicative reference is auxiliary and does not replace this primary additive reference.

Two noninterchangeable dataset builds. The genomic/phylogenetic readouts use the legacy build, while the three functional models use corrected v2. Legacy maps the source RR/ER/RE/EE categories to CC/EC/CE/EE. Version v2 instead identifies controls from the 20 identities in Supplementary file 2, checks `Controls` suffixes, normalizes component identities/orientations/configuration groups, and recomputes support terms and labels. The differences are not successive filters on one dataset; comparison IDs and labels are not interchangeable.

Table A3: RC4B dataset builds and their biological QC populations.

Quantity	Legacy	Corrected v2
Constructs	110,180	110,180
CC / EC / CE / EE		
CC	142	2,584
EC	4,972	17,613
CE	5,724	20,777
EE / candidate families	99,342	69,206
Candidate cross-promoter comparisons	303,313	218,197
Valid families	94,241	67,877
Biologically valid comparisons	273,736	210,173
Complete eight-promoter profiles	5,112	5,366

Legacy component recovery preserves the original sequence-missingness flags as a separate track: 94,494 families have finite support and 94,241 satisfy the positive-ratio criterion. In v2, family QC requires complete component mapping, valid joint activity and all three support estimates, and finite positive observed/additive values. Cross-promoter comparisons hold identity, order, orientation, and cellular condition fixed. Primary QC does not add the authors' single-support count threshold.

Prediction-validity subsets. After biological QC, each v2 split scheme takes the intersection of finite valid predictions across the three functional models: 207,266 PairConfig-OOD comparisons (98.62%; 9,816 groups) and 193,402 LOPO comparisons (92.02%; 9,753 groups). These subsets are constructed separately, not sequentially, and the zero-change baseline is recomputed on each. Legacy readouts use model- and split-specific valid subsets with matched zero baselines. The two model categories cannot be combined into a common-data performance ranking. Invalid predicted positive-ratio terms remain undefined; no positive clipping or pseudocount is applied.

A.7 RC5: Complete States, Sources, and Higher-Order Labels

RC5 contains 59,851 complete third- to eighth-order cubes, 55,934 unique DNA sequences, 38 source–condition–order strata, and 62 leakage groups. Non-N8 sources contribute 59,179 cubes, 5,397 DNA sequences, 26 strata, and 60 groups; N8 contributes 672 cubes, 50,537 sequences, 12 strata, and two groups. The primary motif/module panel contains 721 cubes: King (210 third-order and 23 fourth-order), e2MPRA (480 third-order), and QeFS (three TWI and five MEF third-order).

Table A4: RC5 source coverage and observation scales.

Counts describe complete measured state sets, not independent backgrounds. Source names retain links from the supplied provenance records.

Source	Cubes	Scale / background scope
King et al., 2020	233	Linear basal-normalized RNA/DNA; shared-component/parent groups.
Zhang et al., 2026	480	Declared $\log_2$ scale; two templates.
Waters et al., 2021	8	$\log_2$ activity; five CREs across TWI/MEF.
Sharon et al., 2012	1,402	Declared $\log_2$ scale; two connected shared-sequence groups.
Yu et al., 2021	57,044	Author-normalized RNA expression; two architectures per IPTG condition.
Whalen et al., 2023	12	Declared $\log_2$ scale; one HAR per order within each condition.
Reiter et al., 2023	672	Control-normalized $\log_2$ RNA/DNA UMI; two parent enhancers.

For all 2^k measured states in one condition, experimental and native-score contrasts are

$$\varepsilon^{(k)} = \sum_{s\in\{0,1\}^k} (-1)^{k-\sum_i s_i} y_s, \qquad \varepsilon_Z^{(k)} = \sum_{s\in\{0,1\}^k} (-1)^{k-\sum_i s_i} \frac{S(x_s)}{L_{\text{biological}}}.$$

Supervised predictions replace y_s with $\hat{y}_s$ from one outer-fold predictor for every cube state. States 0/1 may encode orientation, element identity, motif perturbation, or nucleotide allele, not uniformly motif absence/presence. Source scales are fixed independently of model performance. Float64 accumulation removes lower-order terms on the declared observation scale, without establishing molecular physical synergy.

QeFS RC5 labels first average the collections shared by all states within each week and then weight weeks equally. Aggregates are not additional replicates. RC5 readouts are refitted under these labels rather than reusing RC3 pooled-label predictions.

N8 nucleotide extension. Complete 249/250 bp Drosophila S2 sequences are used. A prespecified random rule selects 16 complete cubes per window/order across seven 8 bp windows in two parent enhancers. The mix2/mix3 libraries contain 64/48 cubes per order. Labels average negative-control-normalized log2 RNA/DNA UMI values across two biological replicates. Every state requires at least five DNA UMIs and one RNA UMI in each replicate, plus finite author-reported DESeq2 log2FC; exploratory windows are excluded. N8 extends coverage through eighth order, but its interactions are among nucleotide positions.

A.8 Quality Control and Fixed Generalization Splits

QC checks ID uniqueness, sequence validity, relation foreign keys, state completeness, duplicate comparisons, and independent label recomputation. These establish internal consistency, not complete recovery of experimental designs or universal homology isolation. Shared states that belong to the held-out unit are never split independently.

RC1 preserves the original chromosome split: chr6/chr21 for validation; chr5/chr10/chr14/chr18/chr20/chr22 for test; the remaining designated chromosomes for training. RC2 fixes deterministically ordered pair indices modulo 10: remainder 0 is test, 1 validation, and all others training. Expanded splits contain 1,391/174/174 pairs; core splits contain 1,291/161/159. RC2 parent IDs represent reconstructed neighboring-sequence pairs. Forward/reverse-complement full-length ungapped comparisons find no cross-split pair at $\geq 80\%$ identity (maximum 92/145); this excludes neither shifted/local/indel homology nor pretraining overlap.

RC3 uses 12 outer enhancer holdouts and inner-loop readout selection on the remaining enhancers. RC4A uses five fixed folds over 153 unordered TF pairs. RC4B separately uses eight-fold leave-one-promoter-out (LOPO) and five-fold element-combination–configuration holdout (PairConfig-OOD); joint and required support terms use the same outer-fold predictor.

RC5 groups shared DNA, reverse complements, and parent families across conditions/orders. Readouts are fitted within source–condition: King uses fixed five-fold group splits, and other identifiable non-N8 sources use leave-one-group-out. N8 uses mix2→mix3 and mix3→mix2 transfer across libraries/enhancers. Stratum-specific group counts do not equal training-condition group counts. More relations sharing backgrounds do not create more independent biological replicates.

B Models and Evaluation Protocols

B.1 Model Checkpoints and Evaluation Coverage

Fourteen pretrained configurations are evaluated: eleven genomic language/phylogenetic models and three functional-genomics models. Table B1 records the exact checkpoints used in our experiments. The genomic language and phylogenetic model families include DNABERT, Nucleotide Transformer, GENA-LM, Caduceus, HyenaDNA, GENErator-v2, Evo 1, Evo 2, and PhyloGPN (Ji et al., 2021; Dalla-Torre et al., 2025; Fishman et al., 2025; Schiff et al., 2024; Nguyen et al., 2023; Li et al., 2026; Nguyen et al., 2024; Brixi et al., 2026; Albors et al., 2025). Mistral-DNA is evaluated using the publicly released Mistral-DNA-v1-422M-hg38 model card and checkpoint.

Table B1: Checkpoints, pretraining categories, and input constraints.

MLM: masked language modeling; AR: autoregressive; BPE: byte-pair encoding. Nominal context denotes the reported checkpoint context or receptive field, not the actual input length used in each benchmark task. Public checkpoint names link to their model records where available; citations identify the corresponding model publications.

Model / checkpoint	Category	Input / context or representation
DNA-BERT-6 (Ji et al., 2021)	MLM	Overlapping 6-mers; 512 tokens; approximately 89.2M parameters
NTv2-500M (Dalla-Torre et al., 2025)	MLM	Nonoverlapping 6-mers; 2,048 tokens
GENA-LM BigBird-base-T2T (Fishman et al., 2025)	MLM	BPE; 4,096 tokens; approximately 114M parameters
HyenaDNA-large-1M (Nguyen et al., 2023)	AR	Single bases; approximately 1 Mb context; approximately 6.6M parameters
Mistral-DNA-v1-422M-hg38	AR	BPE; configured for 512 tokens; see context-length discrepancy discussed below
GENERator-v2-eukaryote-1.2B-base (Li et al., 2026)	AR	Nonoverlapping 6-mers; 16,384 tokens
Caduceus-Ps-131K (Schiff et al., 2024)	MLM	Single bases; 131,072 bp; reverse-complement parameter sharing
Caduceus-Ph-131K (Schiff et al., 2024)	MLM	Single bases; 131,072 bp; reverse-complement augmentation
Evo1-131K-base (Nguyen et al., 2024)	AR	Single bases; 131,072 bp; 7B parameters
Evo2-7B (`evo2_7b`) (Brixi et al., 2026)	AR	Single bases; approximately 1 Mb context; not the 8K-base checkpoint
PhyloGPN (Albors et al., 2025)	Phylogenetic	Single bases; 481 bp receptive field per position; approximately 83.2M parameters
Borzoi, human replicate 0 (Linder et al., 2025)	Functional	Pre-assay-head representations; 32 bp output spacing
Enformer, official Sonnet (Avsec et al., 2021b)	Functional	Pre-assay-head representations; 128 bp output spacing
AlphaGenome, official all-folds (Avsec et al., 2026)	Functional	Pre-assay-head representations; 1 bp output spacing

Checkpoint qualifications. Evo1's pretraining domain is prokaryotes and associated phages and plasmids; its eukaryotic evaluation is marked as cross-pretraining-domain transfer (‡). GENERator-v2 length adaptation is marked †. Mistral-DNA's local configuration specifies 512 positions, whereas its model card describes 10 kb pretraining; these records are not reconciled and do not validate 10 kb support here. Caduceus-PS structural RC equivariance differs from PH's RC augmentation; downstream RC ensembling remains task-specific. PhyloGPN's receptive field is not a maximum total sequence length, and downstream inputs are individual sequences, not multispecies alignments.

Coverage. All eleven genomic/phylogenetic models have RC1 common-subset and RC2/RC3/RC4A native results. RC1 additionally includes five full-test AR scorers, eleven mean-linear readouts, and ten positional CNN readouts; functional-model RC1 results are unavailable. RC4B has ten genomic/phylogenetic readouts, excluding DNA-BERT-6, and three functional readouts. RC5 retains every supported model–stratum result; DNA-BERT-6's complete QeFS inputs exceed its verified context and are N/A. These coverage differences are not filled using other tasks.

B.2 Input Adaptation and Native Scoring

Native evaluation freezes all weights and uses no target-activity labels to fit a head, reverse score direction, or select RC ensembling. Higher scores always denote greater native sequence preference. Native scores are not compared numerically across tokenizations/objectives, and sums of PLL terms need not be normalized joint probabilities. RC2 uses whole-sequence sums, RC3 per-biological-

base scores, RC4A its established construct-scoring convention, and RC5 actual-biological-length normalization excluding technical padding.

Table B2: Native scoring rules for the 145 bp RC2 inputs.

Model	Scoring / adaptation
DNA-BERT-6	Sum of block-masked log probabilities for 140 overlapping 6-mers; target and all overlapping tokens masked together.
Caduceus-Ph / Ps	Sum of independently masked log probabilities for 145 bases, retaining the full native vocabulary.
GENA-LM	Sum of masked probabilities for biological BPE tokens; exclude special/padding tokens.
NTv2	Independently masked probabilities for 24 hexamers and one terminal base; BF16 forward and FP32 log-softmax.
Evo1	Sum of 145 AR target probabilities conditioned on BOS.
Evo2-7B / HyenaDNA	Sum of AR probabilities for bases 2–145 without BOS; first base is context only. Evo2 uses BF16 logits and FP32 log-softmax.
GENERator-v2	Add five technical As; marginalize the artificial portion of the first hexamer. Biological intervals are $[0, 1)$ and 24 subsequent six-base intervals through $[139, 145)$.
Mistral-DNA	Shifted AR probabilities for biological BPE targets conditioned on CLS/BOS; exclude special tokens.
PhyloGPN	Sum of F81 stationary log probabilities at all actual bases; each sequence state receives its own forward pass with the central base visible.

PhyloGPN full-state extension. For F81 log-rate outputs $\theta_i(a; x)$,

$$\pi_i(a; x) = \frac{\exp \theta_i(a; x)}{\sum_b \exp \theta_i(b; x)}, \qquad Z_{\mathrm{F81}}(x) = \sum_{i=0}^{L-1} \log \pi_i(x_i; x).$$

RC1 common-subset and RC2 use sums; RC3/RC4A/RC5 use Z_{F81}/L. Unlike the official variant LLR evaluated in a fixed reference context, every state here recomputes its context, with the central base visible. This is neither masked PLL nor AR joint likelihood. The mapping was fixed before inspecting task metrics. Inputs retain orientation and 240 official – tokens per side; the first L convolution outputs align to actual DNA, with padding excluded. No truncation, fabricated genomic flanks, or RC averaging is used. The backbone uses FP32 with TF32 disabled; log-softmax and summation use FP64. Original checks cover normalization, fixed-reference LLR identity, parameterization equivalence, batch-padding invariance, and score summation.

Full sequences in RC5. Local windows or crops do not replace unsupported full sequences. DNA-BERT-6 supports overlapping-hexamer inputs of at most 515 bp; its 40 QeFS sequences (924–1,073 bp) are `unsupported_context`. PhyloGPN scores every full sequence despite its local receptive field. RC5 Evo2-7B uses the audited EOD/BOS rule, distinct from RC2's no-BOS rule; DNA-BERT-6 uses FP32 and retains the extra output-bias application in the released custom MLM head. These task-specific implementations are not label-based calibration.

Functional-model inputs. Actual DNA is centered in the full model input; unknown flanking positions have all four channels set to 0.25. Only representations within the actual inserted interval are pooled. Repeating coarse bins at base resolution does not increase independent spatial resolution. Short-sequence adaptation, unknown flanks, and cross-species experimental domains limit the interpretation of these readouts.

B.3 Frozen Readouts, Baselines, and Training Selection

A frozen backbone does not imply zero-shot evaluation. Functional models compare mean-representation Ridge and positional CNN readouts in RC2/RC4A, use mean Ridge in RC3, and use component representations plus second-order product terms in RC4B. Genomic/phylogenetic RC4B

results are also supervised readouts, not native scores. Primary supervised predictions replace the state activities defined in Appendix A, followed by the same experimental contrasts. Standardization, PCA, and hyperparameter selection exclude test partitions. Within a task, models share split, readout type, and tuning budget; different representation dimensions can still produce different parameter counts.

RC1 implementation and baselines. NTv2 retains the terminal two bases, giving 60 DNA tokens; after excluding CLS, mean pooling is weighted by base coverage and positional features expand to 350 bases. GENERator-v2 uses a four-A prefix (354 bp, 59 hexamers), marginalizes 256 first-token prefixes for native scoring, and removes the four artificial positions from final-normalization features (350×2048). PhyloGPN retains 350×960 aligned features after official padding. These three linear heads fit standardization on training data only; their positional CNN has 32/64 channels with kernel widths 7/5, GELU, and mean/max pooling. Seeds are 20260914/20260915/20260916; training lasts at most four epochs, selected by full-validation binary cross-entropy. NTv2/PhyloGPN use BF16 backbones and FP32 heads; GENERator uses FP32 with TF32. Evo2's linear head uses mean final-normalization features.

RC1 baselines are 3–6-mer logistic regression, a from-scratch one-hot CNN, MPRA-LegNet (Penzar et al., 2023), ExplaiNN (Novakovsky et al., 2023), and budget-limited gkm-SVM (Ghandi et al., 2014; Lee, 2016). MPRA-LegNet/ExplaiNN load no externally supervised weights and use three seeds with at most four epochs and full-validation selection. gkm-SVM and its matched k-mer LR use the same 2,000 training and 2,000 validation pairs per seed but the entire test set. This differs from the full-training baseline budget. Supervised summaries average seed-specific metrics rather than ensemble predictions.

RC5 supervised protocol. Activity readouts combine orders within each source–condition. N8 has only one parent per library, so within-condition supervision is N/A and reported results use the two cross-library directions. Training-only standardization uses SD floor 10^{-4}; groups receive equal weights. Normalized Ridge penalty is fixed to 1 ($\alpha = n_{\text{train}}$ for sum loss), with an unpenalized intercept and no test-label tuning. Non-N8 and N8 use the group splits in Appendix A.8; QeFS is refitted under its shared-collection/equal-week labels.

RC5 local sequence baselines use concatenated 1/2/3-mer composition under the same grouping and Ridge budget. The nonlinear composition supplement uses at most 128 training-only RBF landmarks, fixed $\gamma = 1/4$, $\alpha = n_{\text{train}}$, and seed 20260923. These are sequence-statistics baselines, not biological lower-order models. Observed corners of test cubes are not primary prediction inputs.

Information conditions. RC2/RC3/RC4A use task-appropriate k-mer Ridge, from-scratch CNN, or constant-zero interaction controls. RC4B and RC5 zero baselines match each model's valid evaluation population and label scale. The three functional models have no randomly initialized, same-architecture backbone controls: architecture, pretraining domain, representation resolution, and readout can all affect comparisons, so differences from simple baselines do not isolate pretraining gains.

B.4 Metrics and Aggregation Rules

RC1 evaluates paired source recognition; RC2 native scores are compared with candidate PWM changes; supervised RC2 and functional RC3–RC5 use experimental contrasts. Native values are not calibrated experimental activities and do not receive MAE in activity units.

Recognition and native effect metrics. RC1 paired accuracy gives ties half credit; AUROC/AP use real sequences as positives and assess cross-sequence ranking. Common-subset and full-test populations stay separate. For RC2, M1 reports AUROC/AP of ΔZ against the sign of ΔW, with constant-score AP $18/159$ and AUROC 0.5. AP uses threshold-merged ties, not trapezoidal PR integration. M2 uses the fixed threshold $\Delta Z > 0$ and excludes exact-zero predictions/references for directional metrics. Balanced direction accuracy (BDA) is the mean of positive- and negative-class recall; MCC is undefined for a zero denominator. M3 Spearman and Kendall τ_b rank signed effects, retain finite zeros, and handle ties; they do not measure calibrated magnitude. RC3 intact-motif preference counts strict score increases along 102 edges; ties are not successes.

Cell and configuration averages. RC2 supervised metrics weight the two cells equally. RC4A computes each metric separately for background, orientation A, orientation B, and spacing, then takes an arithmetic mean over factors. Macro Spearman is not a pooled correlation. Evo1's RC4A direction statistics exclude four exact-zero predictions, while rank statistics retain them; other evaluable genomic models retain all comparisons.

RC4B matched error. For the same valid comparison set J,

$$\frac{\mathrm{MAE}_{\mathrm{model}}}{\mathrm{MAE}_0} = \frac{\sum_{j\in J} |\widehat{\Delta}_{P,j} - \Delta_{P,j}|}{\sum_{j\in J} |\Delta_{P,j}|}.$$

Construct-activity R^2 and interaction contrasts are separate endpoints. Legacy model-specific sets and v2 common sets follow Appendix A.6, without cross-build ranking.

RC5 zeros, dependence, and parent weights. Observed/predicted contrasts with magnitude $\leq 10^{-10}$ are fixed numerical zeros, not experimentally nonsignificant effects. Hit rate excludes true zeros, averages within parent and then equally across parents; prediction ties count as errors. BDA similarly uses parent-macro recall for each true-sign class and requires both classes. MCC uses nonzero true/predicted signs, reports valid coverage and ties, and is N/A for zero denominator. Spearman requires at least three observations and variation in both variables. Pooled correlations of shared-state differences are descriptive; parent-macro correlations include only defined parents.

Errors are averaged within parent and then equally across the valid parent set P:

$$\mathrm{MAE}_{\mathrm{parent}} = \frac{1}{|P|}\sum_{p\in P}\frac{1}{n_p}\sum_{j\in p}|\hat{\varepsilon}_j - \varepsilon_j|, \qquad R_{\mathrm{MAE}} = \frac{\mathrm{MAE}_{\mathrm{parent}}(\hat{\varepsilon},\varepsilon)}{\mathrm{MAE}_{\mathrm{parent}}(0,\varepsilon)}.$$

Ratios below one indicate lower point-estimate error, not significant improvement. Source–condition–order strata remain separate; raw errors and ratios are not pooled into an overall leaderboard. N8 additionally averages metrics equally over its 12 test-library–order strata; these are not pooled correlations or independent replicates. Undefined values remain N/A, not zero.

B.5 Uncertainty and Statistical Interpretation

RC1 uses 1,000 paired and 1 Mb block bootstrap resamples; the reported intervals prioritize blocks. Seed variation describes retraining separately. RC2 native AUROC/BDA use 500 class-stratified percentile-bootstrap resamples, and Spearman uses 500 paired resamples with ranks recomputed. No uncomputed intervals are added to AP, MCC, or Kendall estimates. RC2 functional differences use 10,000 parent-cluster resamples, sampling both cells jointly for each parent.

Genomic-model RC3/RC4A and updated RC5 results are descriptive point estimates. Functional-model differences resample RC3 enhancers (5,000), RC4A TF pairs (500), or RC4B pair groups (500). These intervals condition on fixed held-out predictions, omit full retraining variability, and are not adjusted for multiple comparisons. Shared-state relations are not treated as independent replicates.

C Reliability and Robustness Analyses

Data diagnostics use complete eligible experimental populations, not samples chosen by model scores or rankings. Species, measurement scale, sample size, and holdout scheme differ across tasks, so these diagnostics do not provide a controlled cross-task sample-efficiency comparison.

C.1 Experimental Repeatability and Shared Backgrounds

RC2. Replicate-effect Spearman is 0.558 in HepG2 and 0.443 in K562, with sign agreement 67.2% and 65.5% (Figure C1). Across cells, effect Spearman is 0.126 and sign agreement 53.9%; both biological context and reconstructed-label noise may contribute. These measurements describe reproducibility, not denoised performance ceilings.

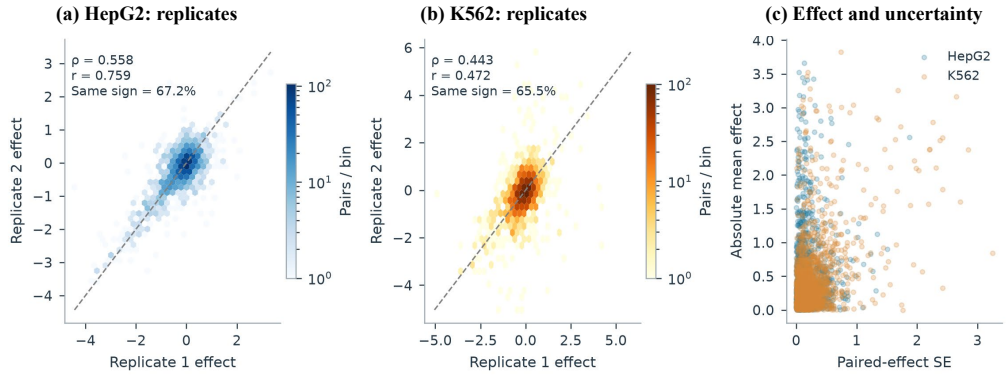


Figure C1: RC2 replicate paired effects and measurement variability. Panels (a) and (b) compare two experimental replicates in each cell type using all reconstructed pairs. Panel (c) compares paired-effect SE with absolute mean effect on the reconstructed scale. Replicate consistency does not establish a corrected model-performance ceiling.

RC4A. For replicate differences $\delta_r = y_{2r} - y_{1r}$, paired SE is $\mathrm{SD}(\delta_1, \delta_2, \delta_3)/\sqrt{3}$, retaining endpoint covariance within the same replicate. It differs from independent endpoint-SE propagation; the mean-effect label is unchanged. Background changes have larger effects and greater replicate agreement than spacing. Relations reuse constructs up to five times. Three-replicate sign agreement and $|\delta| > 2\mathrm{SE}$ are descriptive, not multiplicity-corrected significance tests.

Table C1: RC4A factor-level effect magnitudes and repeatability.

A/B follow source naming; effects/SE are in log2 units. Comparison counts are given in Appendix A.5.

Factor	Median $\lvert\delta\rvert$	Median paired SE	Replicate sign agreement	Replicate ρ range
Spacing	0.140	0.129	37.0%	0.203–0.310
Orientation A	0.152	0.126	40.4%	0.312–0.389
Orientation B	0.152	0.127	41.7%	0.295–0.372
Background	0.332	0.130	64.7%	0.496–0.595

RC5. The ready flag denotes complete traceable states passing measurement QC, not significant higher-order effects. Inter-replicate higher-order Spearman ranges are 0.712–0.743 for third-order King, 0.615–0.840 for fourth-order King, and 0.086–0.175 for third-order e2MPRA. N8's 12 strata range from -0.129 to 0.237, with sign agreement 43.8%–60.4% (Figure D1e). Low model correlation in weakly reproducible strata does not by itself establish absence of higher-order information.

King's 46 groups share a synthetic architecture; e2MPRA has two templates; QeFS has five CREs across two conditions; Sharon has two connected shared-sequence groups. Yu's two architectures per condition supply different interaction orders, and Whalen has one HAR per order. Cube counts therefore do not support independent-and-identically-distributed significance inference.

C.2 RC4B Promoter Dependence in the Legacy Build

Figure C2 describes only legacy data, not the v2 functional-model population. Among 15,893 pair–orientation configurations, 5,112 have all eight promoters and 1,381 only one eligible promoter. On the complete profiles, cross-promoter interaction Spearman ranges from 0.305 to 0.824 (median 0.624). Across all 273,736 valid comparisons, median absolute promoter-associated change is 0.248, and 4.21% exceed one log2 unit. Complete-profile statistics apply to that selected set. Each family can enter up to seven promoter differences, with additional sharing of support terms and promoter baselines.



Figure C2: Legacy RC4B coverage and promoter dependence; not a diagnostic of the corrected-v2 evaluation set. Panel (a): candidate/valid families; (b): interaction distributions; (c): correlations on the same 5,112 complete profiles; (d): eligible promoters per configuration; (e): comparison reuse; (f): absolute context-change distribution. Plotted ranges do not change the full-data statistics. Shared families and support terms are not independent replicates.

C.3 RC4A Standardization Sensitivity

Table C2 retains the original primary settings and distinguishes the status of SD-floor sensitivity analyses. Borzoi's SD-floor change is post hoc, whereas Enformer/AlphaGenome prespecified it before their respective RC4A training. Primary results are not selected by test performance. Both AlphaGenome readouts have paired correlation/MAE difference intervals against k-mer Ridge containing zero; its mean Ridge versus the one-hot CNN has $\Delta\rho = -0.028$ $[-0.042, -0.012]$ and $\Delta\text{MAE} = +0.0062$ $[+0.0038, +0.0085]$.

Table C2: Standardization sensitivity of RC4A mean Ridge.

Model	Standardization	Analysis status	ρ	MAE	RMSE
Borzoi	Original standardization	Primary protocol	0.435	0.229	1.659
Borzoi	SD floor 10^{-4}	Post hoc sensitivity	0.463	0.183	0.238
Enformer	Original standardization	Primary protocol	0.363	0.200	0.265
Enformer	SD floor 10^{-4}	Prespecified sensitivity	0.415	0.193	0.252
AlphaGenome	Original standardization	Primary protocol	0.516	0.172	0.221
AlphaGenome	SD floor 10^{-4}	Prespecified sensitivity	0.516	0.172	0.222

D Additional Numerical Results

Evaluation populations, supervision conditions, and observation scales remain separate. The tables retain all models in each reported panel, including point estimates below the zero baseline; no cross-source overall ranking is constructed.

D.1 RC1: Recognition and Readout Comparisons

High RC1 performance establishes source discrimination, not perturbation-function or interaction prediction. Higher-order local sequence statistics and from-scratch models can also provide strong recognition. Positional-readout gains combine representation, readout capacity, and supervised fitting and cannot isolate the effect of pretraining.

Table D1: RC1 paired recognition by evaluation population and readout.

A–E retain the original separate settings. Paired accuracy gives ties half credit; brackets are 95% 1 Mb block-bootstrap intervals. AUROC/AP are auxiliary cross-sequence metrics. Supervised entries average three seed-specific metrics, not ensembles; seed SD and block intervals measure different variability. †: length adaptation; ‡: cross-pretraining-domain transfer. Common/full test sets and unequal training budgets are not ranked jointly.

Table D1A. Common zero-shot test set (30,000 pairs)

Model / baseline	Score / readout	Paired accuracy [95% CI]	AUROC	AP
DNA-BERT-6	Z-blockPLL-common	0.9510 [0.9483, 0.9538]	0.7579	0.7562
NTv2	Z-PLL-common	0.7500 [0.7443, 0.7558]	0.6873	0.7066
GENA-LM	Z-PLL-common	0.9305 [0.9272, 0.9337]	0.7898	0.8025
HyenaDNA	Z-AR-common	0.8930 [0.8892, 0.8964]	0.6750	0.6769
Mistral-DNA	Z-AR-common	0.9484 [0.9457, 0.9511]	0.7384	0.7520
GENERator-v2†	Z-AR-common	0.9257 [0.9222, 0.9296]	0.7862	0.8008
Caduceus-Ps	Z-PLL-common	0.9717 [0.9698, 0.9737]	0.8054	0.8121
Caduceus-Ph	Z-PLL-common	0.9669 [0.9647, 0.9691]	0.7997	0.8059
Evo1‡	Z-AR-common	0.7324 [0.7267, 0.7378]	0.5383	0.5448
Evo2-7B	Z-AR-common	0.9657 [0.9632, 0.9681]	0.7930	0.8098
PhyloGPN	Z-F81-common	0.6745 [0.6682, 0.6806]	0.6299	0.6704

Table D1B. Full zero-shot test set (489,492 pairs)

Model / baseline	Score / readout	Paired accuracy [95% CI]	AUROC	AP
HyenaDNA	Z-AR-full	0.8911 [0.8896, 0.8926]	0.6739	0.6759
Mistral-DNA	Z-AR-full	0.9454 [0.9439, 0.9468]	0.7365	0.7504
GENERator-v2†	Z-AR-full	0.9255 [0.9233, 0.9276]	0.7847	0.7996
Evo1‡	Z-AR-full	0.7298 [0.7269, 0.7326]	0.5378	0.5440
Evo2-7B	Z-AR-full	0.9641 [0.9626, 0.9655]	0.7914	0.8086

Table D1C. Frozen mean linear readout (all 489,492 pairs)

Model / baseline	Paired accuracy [95% CI]	AUROC	AP	Seed SD (percentage points)
DNA-BERT-6	0.9306 [0.9289, 0.9322]	0.8960	0.8998	0.0021
NTv2	0.8907 [0.8889, 0.8925]	0.8507	0.8477	0.0042
GENA-LM	0.9322 [0.9309, 0.9336]	0.9099	0.9183	0.0024
HyenaDNA	0.8301 [0.8283, 0.8318]	0.7979	0.8019	0.0005
Mistral-DNA	0.8853 [0.8835, 0.8870]	0.8428	0.8522	0.0060
GENERator-v2[†]	0.9431 [0.9416, 0.9444]	0.9233	0.9316	0.0023
Caduceus-Ps	0.9413 [0.9399, 0.9426]	0.9095	0.9181	0.0035
Caduceus-Ph	0.9202 [0.9185, 0.9219]	0.8864	0.8982	0.0010
Evo1[‡]	0.7238 [0.7211, 0.7263]	0.6653	0.6580	0.0044
Evo2-7B	0.8507 [0.8478, 0.8537]	0.7808	0.8208	0.0046
PhyloGPN	0.9465 [0.9453, 0.9477]	0.9269	0.9328	0.0026

Table D1D. Frozen positional CNN readout (all 489,492 pairs)

Model / baseline	Paired accuracy [95% CI]	AUROC	AP	Seed SD (percentage points)
NTv2	0.9246 [0.9229, 0.9262]	0.9068	0.9084	0.0524
GENA-LM	0.9579 [0.9568, 0.9589]	0.9580	0.9629	0.1084
HyenaDNA	0.9237 [0.9224, 0.9251]	0.9160	0.9200	0.1368
Mistral-DNA	0.9278 [0.9264, 0.9292]	0.9148	0.9212	0.1227
GENERator-v2[†]	0.9276 [0.9259, 0.9291]	0.9030	0.9159	0.4962
Caduceus-Ps	0.9633 [0.9623, 0.9642]	0.9570	0.9609	0.0231
Caduceus-Ph	0.9587 [0.9576, 0.9597]	0.9487	0.9531	0.0247
Evo1[‡]	0.8034 [0.8011, 0.8055]	0.7651	0.7685	0.4230
Evo2-7B	0.9340 [0.9325, 0.9356]	0.9182	0.9274	0.2196
PhyloGPN	0.9619 [0.9610, 0.9629]	0.9579	0.9618	0.0544

Table D1E. Independent supervised baselines and composition controls (all 489,492 test pairs)

Model / baseline	Score / readout	Paired accuracy [95% CI]	AUROC	AP
Mono-/dinucleotide control	Composition control (no training)	0.5000 [0.5000, 0.5000]	0.5000	0.5000
3–6-mer LR	Full training set; linear classifier	0.9023 [0.9009, 0.9039]	0.8620	0.8537
One-hot CNN from scratch	Full training set; CNN from scratch	0.9136 [0.9121, 0.9151]	0.8809	0.8771
MPRA-LegNet from scratch	Full training set; neural network from scratch	0.9602 [0.9592, 0.9611]	0.9582	0.9606
ExplaiNN from scratch	Full training set; neural network from scratch	0.9344 [0.9331, 0.9357]	0.9283	0.9314
gkm-SVM (2,000 training pairs)	2,000 training pairs; 2,000 validation pairs	0.8854 [0.8839, 0.8870]	0.8187	0.8021
k-mer LR (matched 2,000 training pairs)	2,000 training pairs; 2,000 validation pairs	0.7980 [0.7965, 0.7994]	0.7651	0.7536

D.2 RC2: Native Effects and Functional Readouts

Native PWM correspondence and experimentally supervised activity changes use different references and are not ranked together. On the same 159 parents, all paired $\Delta\rho$ intervals between functional models with the same readout include zero; model ordering by correlation is therefore not established.

Table D2: RC2 native-score discrimination, direction, and strength ordering.

The 159 matched pairs contain 18 positive and 141 negative candidate PWM changes. AP denotes average precision; BDA, balanced direction accuracy; MCC, Matthews correlation coefficient. PhyloGPN uses the prespecified F81 stationary-probability extension.

A. Direction discrimination and fixed-threshold agreement

Model	M1 AUROC [95% CI]	M1 AP	M2 BDA [95% CI]	M2 MCC
DNA-BERT-6	0.678 [0.562, 0.783]	0.177	0.588 [0.485, 0.703]	0.132
Caduceus-Ph	0.660 [0.525, 0.780]	0.214	0.599 [0.502, 0.710]	0.173
Caduceus-Ps	0.657 [0.539, 0.756]	0.169	0.582 [0.488, 0.691]	0.132
Evo1	0.714 [0.566, 0.841]	0.239	0.705 [0.588, 0.818]	0.271
Evo2-7B	0.696 [0.562, 0.810]	0.197	0.561 [0.467, 0.671]	0.103
GENA-LM	0.613 [0.496, 0.727]	0.151	0.535 [0.408, 0.650]	0.051
GENERator-v2	0.563 [0.428, 0.698]	0.139	0.466 [0.386, 0.552]	-0.051
HyenaDNA	0.639 [0.524, 0.747]	0.157	0.515 [0.421, 0.622]	0.022
Mistral-DNA	0.686 [0.561, 0.792]	0.191	0.582 [0.480, 0.693]	0.132
NTv2	0.557 [0.436, 0.669]	0.152	0.489 [0.378, 0.611]	-0.014
PhyloGPN	0.507 [0.362, 0.638]	0.120	0.482 [0.365, 0.602]	-0.024

B. Signed strength ordering

Model	M3 ρ [95% CI]	M3 τ_b
DNA-BERT-6	0.184 [0.041, 0.330]	0.129
Caduceus-Ph	0.195 [0.041, 0.358]	0.141
Caduceus-Ps	0.193 [0.047, 0.346]	0.136
Evo1	0.144 [-0.022, 0.286]	0.094
Evo2-7B	0.147 [0.004, 0.302]	0.097
GENA-LM	0.122 [-0.019, 0.272]	0.086
GENERator-v2	0.164 [0.016, 0.309]	0.111
HyenaDNA	0.305 [0.163, 0.434]	0.202
Mistral-DNA	0.236 [0.080, 0.372]	0.166
NTv2	0.006 [-0.146, 0.168]	0.006
PhyloGPN	0.064 [-0.094, 0.223]	0.041

Table D3: Functional frozen-readout effects in RC2 and RC4A.

Cells show effect Spearman / MAE. RC2 uses equal cell weights and RC4A equal factor weights; RC4A retains original standardization. MAE is not comparable across tasks.

Task	Readout	Borzoi	Enformer	AlphaGenome
RC2	Mean Ridge	0.270 / 0.395	0.287 / 0.387	0.315 / 0.368
RC2	Positional CNN	0.304 / 0.386	0.306 / 0.375	0.292 / 0.364
RC4A	Mean Ridge	0.435 / 0.229	0.363 / 0.200	0.516 / 0.172
RC4A	Positional CNN	0.415 / 0.191	0.319 / 0.204	0.514 / 0.172

D.3 RC3: Pairwise Interaction Prediction

Table D4: RC3 single-motif preference and pairwise native-score alignment.

Single-motif preference uses 102 edges from 12 enhancers; interactions use 20 complete families from 11 enhancers. Shared states make these point estimates descriptive.

Model	Intact count	Intact rate	Pairwise BDA	Pairwise MCC	Pairwise ρ	Pairwise τ_b
DNA-BERT-6	54/102	0.529	0.550	0.101	0.170	0.116
Caduceus-Ph	61/102	0.598	0.700	0.408	0.386	0.284
Caduceus-Ps	61/102	0.598	0.400	-0.200	-0.141	-0.105
Evo1	47/102	0.461	0.200	-0.600	-0.398	-0.284
Evo2-7B	69/102	0.676	0.600	0.218	0.150	0.095
GENA-LM	54/102	0.529	0.600	0.204	0.002	0.021
GENERator-v2	65/102	0.637	0.500	0.000	0.119	0.116
HyenaDNA	58/102	0.569	0.350	-0.302	-0.361	-0.274
Mistral-DNA	52/102	0.510	0.400	-0.204	-0.217	-0.147
NTv2	59/102	0.578	0.550	0.101	-0.068	-0.042
PhyloGPN	48/102	0.471	0.700	0.400	0.108	0.105

Table D5: RC3 mean-Ridge readouts under enhancer holdout.

All three correlation intervals include zero, and no MAE point estimate improves on zero interaction. Directional hits alone do not establish magnitude accuracy.

Model	Effect ρ	MAE
Borzoi	-0.191	1.456
Enformer	0.211	1.370
AlphaGenome	0.068	1.382
Zero interaction	N/A	1.366

D.4 RC4A: Configuration Alignment

Table D6: RC4A factor-specific direction and macro agreement.

Factor columns report BDA; macro metrics weight the four factors equally over 14,509 comparisons. Functional readouts and standardization sensitivity are in Tables D3 and C2.

Model	Back-ground	Ori. A	Ori. B	Spacing	Macro BDA	Macro MCC	Macro ρ
DNA-BERT-6	0.509	0.467	0.490	0.524	0.497	0.002	-0.025
Caduceus-Ph	0.500	0.512	0.522	0.558	0.523	0.046	0.074
Caduceus-Ps	0.502	0.531	0.486	0.566	0.521	0.046	0.064
Evo1	0.463	0.502	0.541	0.538	0.511	0.026	0.030
Evo2-7B	0.507	0.538	0.488	0.542	0.519	0.043	0.077
GENA-LM	0.488	0.482	0.518	0.509	0.499	-0.008	-0.050
GENERator-v2	0.517	0.534	0.466	0.537	0.514	0.026	0.025
HyenaDNA	0.481	0.454	0.491	0.540	0.491	-0.016	-0.016
Mistral-DNA	0.502	0.537	0.503	0.559	0.525	0.056	0.066
NTv2	0.495	0.524	0.501	0.505	0.506	0.010	0.036
PhyloGPN	0.493	0.541	0.510	0.508	0.513	0.025	0.022

D.5 RC4B: Activity Prediction and Promoter Transfer

Tables D7 and D8 use different dataset builds (Appendix A.6). Legacy genomic-model activity R^2 is reported only for PairConfig-OOD; every listed genomic-model MAE exceeds its matched zero-change reference in both settings. On corrected v2, functional models have lower MAE point estimates than zero change for PairConfig-OOD, but all exceed it under LOPO. These findings support protocol-specific transfer limitations, not a common-sample ranking across model categories.

Table D7: RC4B genomic/phylogenetic frozen readouts on the legacy build.
The biological-QC population is 273,736 comparisons. Each MAE/zero ratio uses the same model- and split-specific valid set. DNA-BERT-6 has no RC4B readout.

Model	PairConfig-OOD activity R^2	PairConfig-OOD Δ_P ρ	PairConfig-OOD MAE / zero	LOPO Δ_P ρ	LOPO MAE / zero
Caduceus-Ph	0.912	0.296	1.274	0.020	2.548
Caduceus-Ps	0.903	0.293	1.301	-0.090	2.588
Evo1	0.886	0.269	1.341	0.022	2.350
Evo2-7B	0.884	0.263	1.469	-0.007	2.636
GENA-LM	0.895	0.288	1.343	-0.070	2.667
GENERator-v2	0.888	0.272	1.413	0.064	2.092
HyenaDNA	0.891	0.273	1.365	-0.034	2.535
Mistral-DNA	0.898	0.315	1.335	0.026	2.162
NTv2	0.903	0.286	1.362	0.004	2.180
PhyloGPN	0.917	0.327	1.242	0.060	2.110

Table D8: RC4B functional readouts on corrected-v2 common valid subsets.

PairConfig-OOD and LOPO use 207,266 and 193,402 comparisons, respectively, from the 210,173 biological-QC set; subsets are constructed independently. All rows within a split share the same comparisons.

Split	Model	ρ	MAE	RMSE
PairConfig-OOD	Borzoi	0.538	0.292	0.448
PairConfig-OOD	Enformer	0.516	0.312	0.508
PairConfig-OOD	AlphaGenome	0.538	0.285	0.434
PairConfig-OOD	Zero change	N/A	0.317	0.435
LOPO	Borzoi	0.032	0.545	0.821
LOPO	Enformer	0.101	0.454	0.650
LOPO	AlphaGenome	0.111	0.543	0.786
LOPO	Zero change	N/A	0.316	0.434

Under LOPO, each functional model's paired MAE-difference interval relative to zero change is above zero. Under PairConfig-OOD, AlphaGenome minus Borzoi/Enformer has $\Delta\mathrm{MAE} = -0.0072$ $[-0.0106, -0.0037]$ / -0.0278 $[-0.0321, -0.0236]$. These fixed-prediction intervals do not include retraining variability.

D.6 RC5: Primary Panel and Complete Non-N8 Strata

All 26 non-N8 strata appear below; the 12 N8 strata are shown separately in Figure D1. Cube, parent, and group count complete contrasts, parent families, and leakage groups within a stratum; summing groups across orders/conditions double-counts shared groups. Native ranges are descriptive pooled-Spearman minima/maxima over defined model results, not a pooled cross-source score. Single-cube strata have N/A correlations. Unsupported DNA-BERT-6 QeFS inputs are excluded from native ranges.

Third-order King and e2MPRA have no functional-model MAE below the matched zero baseline; fourth-order King has small point-estimate advantages for Borzoi/AlphaGenome. Other strata include both improvements and degradations. These heterogeneous results should be interpreted with the repeatability and background limitations in Appendix C1, not as uniformly absent higher-order information.

Table D9: RC5 primary motif/module panel: native scores.

Cells show pooled Spearman / parent-macro BDA; cube counts are 210, 23, 480, 3, and 5. Prediction ties count as direction errors. PhyloGPN uses full-state F81 score per actual base.

Model	King $k = 3$	King $k = 4$	e2MPRA $k = 3$	QeFS TWI $k = 3$	QeFS MEF $k = 3$
DNA-BERT-6	-0.108 / 0.503	-0.225 / 0.500	-0.002 / 0.485	N/A	N/A
Caduceus-Ph	-0.230 / 0.402	0.023 / 0.446	0.044 / 0.510	-0.500 / 0.750	-0.100 / 0.417
Caduceus-Ps	-0.096 / 0.510	0.320 / 0.637	0.044 / 0.530	0.500 / 1.000	0.500 / 0.583
Evo1	-0.155 / 0.446	0.009 / 0.500	-0.100 / 0.476	0.500 / 0.750	0.100 / 0.583
Evo2-7B	-0.095 / 0.506	-0.096 / 0.600	0.126 / 0.549	0.500 / 0.750	0.300 / 0.667
GENA-LM	-0.095 / 0.466	-0.150 / 0.508	0.095 / 0.539	-0.500 / 0.250	-0.400 / 0.417
GENERator-v2	-0.177 / 0.432	0.013 / 0.537	0.015 / 0.513	-1.000 / 0.000	-0.200 / 0.417
HyenaDNA	-0.028 / 0.518	0.043 / 0.562	-0.009 / 0.488	-0.500 / 0.000	-0.300 / 0.417
Mistral-DNA	0.025 / 0.482	-0.255 / 0.392	-0.033 / 0.475	-0.500 / 0.000	-0.300 / 0.250
NTv2	0.021 / 0.450	0.095 / 0.613	0.047 / 0.533	0.500 / 0.500	0.000 / 0.417
PhyloGPN	-0.093 / 0.432	-0.088 / 0.388	0.067 / 0.518	-1.000 / 0.000	-0.900 / 0.417

Table D10: RC5 primary motif/module panel: supervised frozen readouts.

Cells show pooled Spearman / parent-macro MAE divided by matched zero-interaction MAE. Ratios below one are point-estimate, not significance, improvements; readouts are refitted by source and condition.

Model	King $k = 3$	King $k = 4$	e2MPRA $k = 3$	QeFS TWI $k = 3$	QeFS MEF $k = 3$
Borzoi	-0.060 / 1.015	0.005 / 0.992	0.052 / 1.171	-0.500 / 1.089	-0.100 / 1.066
Enformer	-0.206 / 1.137	-0.472 / 1.180	0.007 / 6.870	0.500 / 1.202	0.200 / 1.022
AlphaGenome	0.055 / 1.041	0.048 / 0.994	-0.059 / 1.018	-0.500 / 1.027	-0.900 / 1.014

Table D11: Sizes and native correspondence of all 26 non-N8 RC5 strata.

Source / condition	k	Cube	Parent / group	Native ρ range
e2MPRA / HepG2	3	480	2/2	-0.100–0.126
Whalen HAR / chimp Pt2A N2	3	1	1/1	N/A
Whalen HAR / chimp Pt2A N2	4	1	1/1	N/A
Whalen HAR / chimp Pt2A N2	5	1	1/1	N/A
Whalen HAR / chimp Pt2A N3	3	1	1/1	N/A
Whalen HAR / chimp Pt2A N3	4	1	1/1	N/A
Whalen HAR / chimp Pt2A N3	5	1	1/1	N/A
Whalen HAR / human WTc N2	3	1	1/1	N/A
Whalen HAR / human WTc N2	4	1	1/1	N/A
Whalen HAR / human WTc N2	5	1	1/1	N/A
Whalen HAR / human WTc N3	3	1	1/1	N/A
Whalen HAR / human WTc N3	4	1	1/1	N/A
Whalen HAR / human WTc N3	5	1	1/1	N/A
King / RW4 mESC	3	210	46/46	-0.230–0.025
King / RW4 mESC	4	23	23/23	-0.255–0.320
QeFS / MEF	3	5	5/5	-0.900–0.500
QeFS / TWI	3	3	3/3	-1.000–0.500
Sharon / yeast SC Gal	3	913	4/2	-0.147–0.156
Sharon / yeast SC Gal	4	379	3/2	-0.208–0.136
Sharon / yeast SC Gal	5	95	3/2	-0.318–0.327
Sharon / yeast SC Gal	6	14	1/1	-0.578–0.626
Sharon / yeast SC Gal	7	1	1/1	N/A
Yu / 0 mM IPTG	4	25,886	1/1	-0.082–0.052
Yu / 0 mM IPTG	5	2,636	1/1	-0.126–0.088

Continued on next page

Source / condition	k	Cube	Parent / group	Native ρ range
Yu / 1 mM IPTG	4	25,886	1/1	-0.078–0.042
Yu / 1 mM IPTG	5	2,636	1/1	-0.171–0.082

Table D12: Parent-macro MAE / matched zero-interaction MAE for all non-N8 RC5 strata.

Source / condition	k	Borzoi	Enformer	AlphaGenome	k-mer Ridge
e2MPRA / HepG2	3	1.171	6.870	1.018	1.000
Whalen HAR / chimp Pt2A N2	3	1.406	0.629	0.922	1.000
Whalen HAR / chimp Pt2A N2	4	1.024	0.894	1.013	1.000
Whalen HAR / chimp Pt2A N2	5	1.028	0.459	0.134	1.000
Whalen HAR / chimp Pt2A N3	3	1.300	1.035	0.751	1.000
Whalen HAR / chimp Pt2A N3	4	1.038	0.802	1.047	1.000
Whalen HAR / chimp Pt2A N3	5	1.041	0.771	0.976	1.000
Whalen HAR / human WTc N2	3	1.262	0.827	0.906	1.000
Whalen HAR / human WTc N2	4	1.034	0.687	1.061	1.000
Whalen HAR / human WTc N2	5	0.990	0.546	0.919	1.000
Whalen HAR / human WTc N3	3	1.225	0.881	0.890	1.000
Whalen HAR / human WTc N3	4	1.082	0.577	1.074	1.000
Whalen HAR / human WTc N3	5	1.095	3.125	0.385	1.000
King / RW4 mESC	3	1.015	1.137	1.041	1.000
King / RW4 mESC	4	0.992	1.180	0.994	1.000
QeFS / MEF	3	1.066	1.022	1.014	1.000
QeFS / TWI	3	1.089	1.202	1.027	1.000
Sharon / yeast SC Gal	3	15.507	11.396	1.032	1.000
Sharon / yeast SC Gal	4	18.969	11.646	1.088	1.000
Sharon / yeast SC Gal	5	34.567	21.343	1.025	1.000
Sharon / yeast SC Gal	6	5.237	12.509	1.102	1.000
Sharon / yeast SC Gal	7	7.436	23.539	1.201	1.000
Yu / 0 mM IPTG	4	1.032	1.210	1.005	1.000
Yu / 0 mM IPTG	5	1.003	1.021	1.005	1.000
Yu / 1 mM IPTG	4	1.037	1.196	1.007	1.000
Yu / 1 mM IPTG	5	1.003	1.031	1.007	1.000

D.7 N8: Cross-Library Transfer

The eleven native models' means across 12 test-library–order strata range from -0.078 to 0.044 for Spearman and 0.467 to 0.527 for BDA. Functional-model mean Spearman/BDA/MAE-ratio is 0.029/0.495/1.077 for Borzoi, 0.096/0.527/1.097 for Enformer, and 0.143/0.555/1.007 for AlphaGenome. Ratios are below one in 2/12, 2/12, and 5/12 strata, respectively, using unrounded values; AlphaGenome's seventh-order mix3 ratio is approximately 0.999865 despite displaying as 1.000.

These are equal-stratum summaries, not pooled correlations or independent replicates. N/A within-condition supervision is not mixed with completed cross-library results. Weak experimental-repeat agreement limits conclusions from low model correlation alone. Sequence baselines, undefined correlations, and both better/worse-than-zero point estimates are retained in the figure.

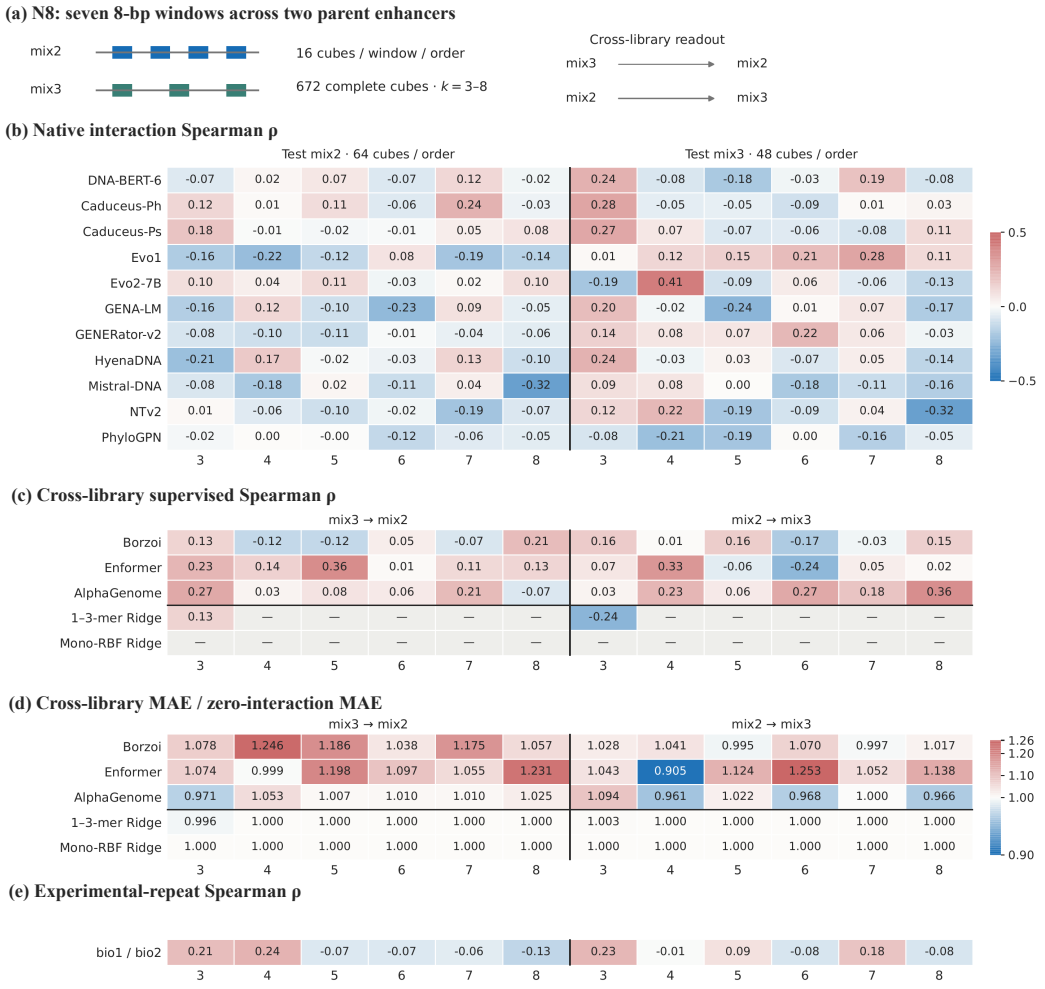


Figure D1: N8 nucleotide interactions and cross-library prediction. Panel (a): perturbation/transfer design; (b): native Spearman; (c) and (d): cross-library supervised Spearman and matched MAE ratios; (e): experimental-repeat Spearman. Test mix2 and mix3 are opposite transfer directions. Dashes indicate undefined constant-prediction correlations. Rounded ratios are not used to decide whether error is below the zero baseline.