BEYOND SINGLE VARIANTS: A BENCHMARK FOR REGULATORY EPISTASIS IN GENOMIC SEQUENCE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic sequence models, both supervised sequence-to-function predictors (Enformer, Borzoi, AlphaGenome) and self-supervised genome language models (Nucleotide Transformer, Evo 2), are increasingly used to interpret non-coding variation, yet they are benchmarked almost exclusively on variants tested one at a time against a reference background. Real haplotypes carry many variants per regulatory element and regulatory grammar is non-additive, so the practical question is whether models can compose variant effects. We assemble a benchmark from MPRA diplotype experiments that assay all four haplotypes (REF, A, B, AB) of thousands of variant pairs, together with ˜1.15M single-variant (variant $\times$ cell-type) measurements across three studies, and evaluate five models along a hierarchy of tasks: single-variant regression (T1), interaction regression (T2), non-additive pair detection (T3), and full-versus-additive prediction across 421 phased personal genomes (T4), each under a zero-shot and a frozen-encoder linear-probe readout. On Siraj et al. (2026), models rank confidently measured single-variant effects moderately well (zero-shot Spearman ρ up to 0.42; linear probe up to 0.73) but capture far less of the measured interaction signal ($\rho \leq 0.32$; AUROC ≤ 0.68, against 0.62 for inter-variant distance alone). Linear probes recover substantial single-variant signals from gLM representations that their likelihoods do not expose, yet bring interaction detection at most to the level of the distance baseline. On personal genomes, full-haplotype predictions are no more accurate than additive reconstructions from each model's own marginal effects. Current models' interaction predictions thus add little beyond their marginal effects, a gap that matters for variant prioritization on real haplotypes and for sequence design, and that motivates interaction-aware training data, objectives and benchmarks.

1 INTRODUCTION

More than 90% of trait- and disease-associated variants lie in non-coding DNA (Maurano et al., 2012), motivating models that predict regulatory function directly from sequence. Supervised sequence-to-function (S2F) models such as Enformer (Avsec et al., 2021), Borzoi (Linder et al., 2025) and AlphaGenome (Avsec et al., 2026) predict functional-genomic readouts, while genomic language models (gLMs) such as Nucleotide Transformer (Dalla-Torre et al., 2025) and Evo 2 (Brixi et al., 2026) learn representations by self-supervision. Both families are now used for variant-effect prediction, either from pretrained outputs or after task-specific adaptation.

Regulatory DNA, however, is combinatorial. Transcription-factor binding depends on local context and motif syntax (Avsec et al., 2021b; de Boer et al., 2020), nearby regulatory variants interact non-additively, i.e. exhibit regulatory epistasis (Phillips, 2008; Siraj et al., 2026), and multiple linked functional variants often underlie a single association signal (Abell et al., 2022). Interpreting a genome therefore requires predicting not only individual variant effects but how those effects compose.

Existing benchmarks evaluate variants independently: saturation mutagenesis and CAGI challenges (Kircher et al., 2019; Shigaki et al., 2019), allelic MPRAs (Tewhey et al., 2016) and gLM suites such as BEND (Marin et al., 2024) all score substitutions in a fixed background. Personal-genome studies

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.

offer a natural compositional test, but models explain little inter-individual expression variation and often miss effect direction (Sasse et al., 2023; Huang et al., 2023), and such evaluations cannot isolate compositional failure because differences among individuals also reflect errors in marginal effects, trans regulation and measurement noise. Testing composition directly requires ground truth in which individual and joint variant effects are measured in the same sequence context.

MPRA diplotype experiments provide exactly this. For a variant pair (A, B), all four haplotypes REF, A, B and AB are assayed, and the interaction is measured as

$$\epsilon = y_{AB} - y_A - y_B + y_{REF}. \tag{1}$$

Applying the same contrast to model predictions yields a predicted interaction that can be compared directly with the measured one. Figure 1 summarizes the workflow. We harmonize diplotype measurements from Abell et al. (2022) and Siraj et al. (2026), saturation mutagenesis from Kircher et al. (2019) and phased genomes of 421 GEUVADIS individuals (Fig. 1A); score every haplotype with three S2F models and two gLMs under a zero-shot and a linear-probe readout (Fig. 1A–B); evaluate along a hierarchy from single variants to pairs (Fig. 1C); and finally test whether predictions from complete personal haplotypes beat additive reconstructions of marginal effects (Fig. 1D). Our contributions are:

- **A benchmark for regulatory epistasis.** >4,000 MPRA-tested variant pairs across six cell lines with complete four-haplotype measurements, plus ˜1.15M single-variant measurements, supporting single-variant regression, interaction regression and non-additive pair detection with leakage-free, locus-grouped splits.
- **A systematic evaluation of compositional prediction.** Across three S2F models and two gLMs, marginal effects are predicted far better than measured interactions. Linear probing markedly improves single-variant prediction, including for gLMs whose likelihoods carry little regulatory signal, but yields much smaller gains on interactions, exposing a persistent compositional gap.
- **Extension to natural human genomes.** Across 421 phased GEUVADIS genomes, full-haplotype predictions barely differ from additive reconstructions of each model's own marginal effects, extending the predominantly additive behaviour seen in controlled assays to natural variant combinations.

Together, these results identify variant composition as a distinct, under-evaluated capability of genomic sequence models and motivate compositional benchmarks, interaction-sensitive objectives and diplotype-aware training data.

2 RELATED WORK

Genomic sequence models. Supervised S2F models progressed from convolutional architectures such as DeepSEA and Basenji (Zhou & Troyanskaya, 2015; Kelley et al., 2018) through enhancer-activity models such as DeepSTARR (de Almeida et al., 2022) to long-context models (Enformer, Borzoi, AlphaGenome), with variant effects estimated by contrasting scores between reference and alternative sequences. gLMs (Nucleotide Transformer, Evo 2) learn from sequence alone and score variants via likelihoods or downstream probes. Recent work adapts pretrained representations to new regulatory tasks, e.g. scooby (Hingerl et al., 2025) and Decima (Lal et al., 2026).

Variant-effect evaluation and regulatory epistasis. Benchmarks predominantly score individual substitutions: saturation mutagenesis and CAGI (Kircher et al., 2019; Shigaki et al., 2019), allelic MPRAs (Tewhey et al., 2016), eQTL/fine-mapping evaluations and BEND (Marin et al., 2024). Regulatory sequence is nonetheless non-additive (Avsec et al., 2021b; de Boer et al., 2020), and in silico intervention studies show that models can represent pairwise interactions; whether predicted interactions match experimentally measured ones for naturally occurring pairs has not been tested. Abell et al. (2022) found multiple major allelic effects at 17.7% of tested eQTLs, and Siraj et al. (2026) assayed all four diplotypes of thousands of nearby fine-mapped pairs, finding significant non-additivity in ˜11% of evaluable pairs. We use these measurements not to discover epistasis but to ask whether models quantitatively predict it.

Personal-genome prediction. Sasse et al. (2023) and Huang et al. (2023) showed that current models explain little inter-individual expression variation and frequently invert effect direction, and

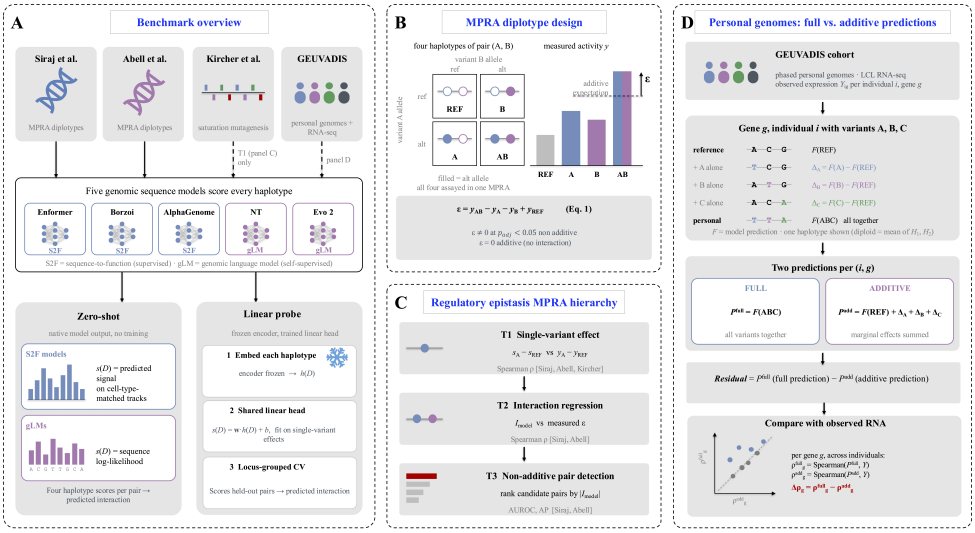


Figure 1: Benchmark workflow. **(A)** Four data sources are scored by five models (three S2F, two gLM) under two readouts: zero-shot (native outputs) and a linear probe on frozen representations. **(B)** Each MPRA variant pair is assayed as four haplotypes, defining the measured interaction (Eq. 1); the same contrast on model scores gives the predicted interaction. **(C)** Task hierarchy: single-variant effect (T1), interaction regression (T2) and non-additive pair detection (T3). **(D)** Personal genomes: for each individual and gene, a full prediction from the complete phased haplotypes is compared with an additive reconstruction that sums the model's own isolated variant effects; g=gfull-gadd quantifies the value of joint inference (T4).

Karollus et al. (2023) showed that they capture promoter-proximal determinants while largely ignoring distal enhancers. These evaluations conflate errors in individual effects, in their composition, and in unmodelled trans or environmental variation. Our MPRA tasks isolate composition; our GEUVADIS task asks whether the same behaviour appears on natural haplotypes.

3 BENCHMARK DEFINITION

The benchmark has four stages (Fig. 1A): (i) assemble ground truth in which single and joint variant effects are measured in the same context; (ii) construct the corresponding haplotype sequences; (iii) score every haplotype with every model under a zero-shot and a linear-probe readout; (iv) form single-variant and four-haplotype contrasts and evaluate tasks T1–T4. This section defines (i) and the tasks; Section 4 gives (ii)–(iii).

3.1 DATASETS AND GROUND TRUTH

Appendix Table A1 summarizes the four data sources; counts below reflect benchmark-level harmonization.

Siraj et al. (2026) assayed ˜304,000 variants in five cell lines (A549, HCT116, HepG2, K562, SK-N-SH) and a targeted screen of 2,522 pairs of fine-mapped variants (PIP > 0.1) within 150 bp of one another in the same cis-regulatory element, with all four haplotypes assayed. T1 uses the reported allelic effect (log2Skew) and expression-modulating-variant (emVar) labels (FDR < 0.1); T2–T3 use the four-haplotype interaction effects. After hg38 harmonization and restriction to units scored by every model under both readouts, we retain 277,312 SNVs (1,111,996 variant $\times$ cell-type measurements; 33,790 emVars in at least one cell line) and 2,164 pairs. Of the 180 significantly non-additive pairs, 178 map to hg38, 149 are SNV–SNV, and 143 have both constituent variants in the single-variant set; these 143 form the T2 set.

Abell et al. (2022) assayed 29,565 variants in lymphoblastoid cells and 2,098 four-haplotype tests of eQTL-associated variant pairs < 75 bp apart (2,038 unique pairs). After harmonization and restriction to units scored by every model, we retain 23,601 variants and 1,673 pairs. T1 uses the allelic log-fold change, with allelic hits defined as BH-adjusted allelic and expression $P < 0.05$ (1,043 hits); T2 uses the published interaction coefficient; and T3 uses the study's non-additivity calls (19 pairs).

Kircher et al. (2019) is an independent single-variant set: saturation mutagenesis of disease-associated regulatory elements; our harmonized subset holds 19,933 measurements (17,663 SNVs) across 14 elements and is used for T1 only.

GEUVADIS. Following Huang et al. (2023), we pair GEUVADIS LCL RNA-seq (Lappalainen et al., 2013) with phased 1000 Genomes genotypes (1000 Genomes Project Consortium, 2015) for 421 individuals, start from the 3,259 genes with a significant cis-eQTL and PEER-normalized expression, and evaluate a fixed, expression-blind subset of 300 genes (75 per common-variant-burden quartile) drawn from the 389 genes that every model can evaluate.

3.2 Task hierarchy

Tasks are ordered by compositional complexity (Fig. 1C–D). T1, single-variant regression: Spearman ρ between predicted and measured marginal effects, on all quality-controlled variants and on each study's significant subset. T2, interaction regression: Spearman ρ between the predicted four-haplotype contrast I_{model} and measured ϵ on each study's high-confidence interaction set. T3, non-additive pair detection: all candidate pairs ranked by $|I_{\mathrm{model}}|$ against the study's non-additivity labels, scored by AUROC and average precision (AP) relative to prevalence. T4, personal-genome compositional gain: for each gene, $\Delta\rho_g = \rho_g^{\mathrm{full}} - \rho_g^{\mathrm{add}}$, the difference in across-individual Spearman correlation with observed expression between the full-haplotype prediction and an additive reconstruction from the same model's isolated variant effects; positive values indicate that joint inference adds information beyond marginal effects.

4 Scores, probes, splits and metrics

4.1 Sequence construction and zero-shot scoring

For each variant pair, we construct the four haplotypes REF, A, B, and AB in a shared genomic context centered on the pair. Model-specific input lengths and pooling windows follow each architecture and are detailed in Appendix B. For sequence-to-function (S2F) models, cell-type-matched accessibility and histone-modification outputs are pooled locally and combined with equal weight across modalities. If S_h denotes the resulting score for haplotype h, we define the marginal and interaction effects as

$$\Delta(v) = S_{\mathrm{alt}} - S_{\mathrm{ref}}, \qquad I_{\mathrm{model}} = S_{11} - S_{10} - S_{01} + S_{00}. \tag{2}$$

For genomic language models (gLMs), S_h is the sequence likelihood score: masked pseudo-log-likelihood for Nucleotide Transformer and autoregressive log-likelihood for Evo 2. All interaction contrasts are computed on the same transformed scale used for each model's marginal scores; exact track sets, transformations, and pooling choices are given in Appendix B.

4.2 Frozen-representation probes and splits

To test whether pretrained representations expose regulatory signal beyond their native readouts, we fit ridge-regression probes on frozen embeddings. Probes are trained only on single-variant effects, mapping $h(A) - h(\mathrm{REF})$ to $y_A - y_{\mathrm{REF}}$. We use five-fold grouped cross-validation, with all variants, haplotypes, overlapping constructs, and assayed pair relationships from the same locus assigned to the same fold; standardization and regularization are selected by group-disjoint inner cross-validation. For held-out pairs, the same marginal-effect probe is applied without retraining:

$$\hat{\epsilon} = f[h(11)] - f[h(10)] - f[h(01)] + f[h(00)]. \tag{3}$$

Thus, probed T2–T3 performance measures whether representation-level non-additivity is aligned with a direction learned from marginal effects and should be interpreted as a lower bound on inter-

action information in the frozen representation. Embedding layers, dimensions, and pooling procedures are provided in Appendix C.

4.3 PERSONAL GENOMES AND EVALUATION

For T4, each individual's two phased haplotypes are scored jointly around each gene. The full prediction P_{ig}^{full} is compared with an additive reconstruction P_{ig}^{add} obtained by summing the same model's isolated single-variant effects; compositional gain is $\Delta\rho_g = \rho_g^{\text{full}} - \rho_g^{\text{add}}$, where correlations are computed across individuals against measured GEUVADIS expression. Full construction details are in Appendix F.

T1 and T2 are evaluated with Spearman correlation; T3 uses AUROC and average precision. To compare marginal and interaction performance directly, Fig. 2B–C uses a matched set consisting of the constituent variants of the same assayed SNV–SNV pairs, evaluated in the same cell type and sequence context. Confidence intervals are 95% locus-level cluster bootstraps with 2,000 resamples. Dataset harmonization, coverage, track-selection robustness, negative subsampling, and distance controls are described in Appendices A–F.

5 RESULTS

Table 1 gives the Siraj headline numbers for every task and both readouts (full results with Abell and Kircher: Appendix Table A2). In this section, we show that models predict single-variant effects far better than interactions, and the gap survives a matched comparison (5.1–5.2); linear probes recover marginal but not compositional information from the same frozen representations (5.2–5.3); and on natural genomes, joint haplotype inference adds no measurable accuracy over the additive sum of each model's own variant effects (5.4).

5.1 THE HIERARCHY GAP: SINGLE-VARIANT SKILL DOES NOT TRANSFER TO INTERACTIONS

Zero-shot, the three S2F models correlate moderately with significant single-variant effects ($\rho = 0.32$–0.42) but weakly with measured interactions ($\rho = 0.07$–0.14), placing them far below the diagonal in Fig. 2A. The gLMs carry no zero-shot single-variant signal, yet Evo 2 has the highest zero-shot interaction ρ of any model (0.19). Because T1 and T2 differ in sample size and selection, we recomputed both quantities on the matched set, i.e. the constituent variants of the very pairs used for T2 (143 pairs), scored in the same cell type and with the same track aggregation (Section 4.5). Linear probes raise single-variant ρ for every model (Fig. 2B), to ˜0.6–0.7 for S2F models and from ≈ 0 to 0.44 (NT) and 0.61 (Evo 2) for the gLMs (Table 1). Probed interaction ρ, by contrast, stays at 0.06–0.32 (Fig. 2C). Borzoi has the strongest probed interaction signal and Evo 2 the strongest zero-shot interaction signal, while AlphaGenome, the best model on marginal effects under both readouts, is the weakest S2F model on zero-shot interactions; model ordering on interactions therefore does not follow ordering on marginal effects. Probing on single-variant effects improves marginal prediction substantially but interaction prediction only modestly: marginal skill is a poor proxy for compositional skill. At the scale of personal genomes, with up to thousands of common variants per gene, zero-shot predictions show little across-individual accuracy for any model (Section 5.4).

5.2 SINGLE-VARIANT EFFECTS (T1)

Figure 3 compares the two readouts on the significant variants of all three MPRA datasets. Zero-shot, S2F models capture a substantial part of the measured variation in the Siraj cell lines ($|\rho| = 0.34$–0.55 for the best S2F model and track set per cell line, selected on the evaluation data; fixed-track-set results, which lead to the same conclusions, are in Appendix D), with AlphaGenome best in every cell line; Enformer and Borzoi do best with accessibility tracks, and adding histone or transcriptional tracks brings no gain or reduces performance. Sequence-only likelihood scores from NT and Evo 2 are largely uninformative on the MPRA datasets ($|\rho| < 0.1$ on Siraj and Abell).

Linear probing improves every model. On Siraj pooled across cell types the probes reach $|\rho| = 0.73$ (AlphaGenome), 0.68 (Enformer), 0.66 (Borzoi), 0.61 (Evo 2) and 0.45 (NT), with the same rank-

Table 1: Main results on Siraj et al. (2026) across T1 single-variant regression, T2 interaction regression, T3 non-additive pair detection, and T4 personal-genome compositional gain. Zero-shot and frozen-representation linear-probe results are shown on the common evaluation set; 1/Distance is the inverse inter-variant-distance baseline. Best value per row and readout in bold. Full task definitions, sample sizes, scoring details, and additional metrics are provided in Sections 3–4 and the Appendix.

	Zero-shot						Linear probing					
Task / metric	Enformer	Borzoi	AlphaGenome	NT	Evo 2	1/Distance	Enformer	Borzoi	AlphaGenome	NT	Evo 2	1/Distance
T1 ρ, all variants	0.066	0.091	**0.101**	−0.036	−0.031	–	0.224	0.217	**0.250**	0.159	0.206	–
T1 ρ, emVar subset	0.322	0.383	**0.416**	−0.044	−0.026	–	0.683	0.657	**0.728**	0.445	0.614	–
T2 interaction ρ	0.113	0.141	0.072	0.000	**0.185**	0.145	0.152	**0.315**	0.179	0.063	0.187	−0.221
T3 AUROC	0.623	0.651	**0.653**	0.584	0.501	0.619	0.676	**0.683**	0.681	0.589	0.615	0.582
T3 AP	0.194	0.197	**0.232**	0.136	0.081	0.135	0.231	0.261	**0.314**	0.150	0.227	0.125
T4 median $\Delta\rho_g$	−0.002	**0.002**	0.001	0.000	−0.001	–	–	–	–	–	–	–
T4 % genes $\Delta\rho_g > 0$	47.3	**53.7**	51.7	51.0	48.7	–	–	–	–	–	–	–

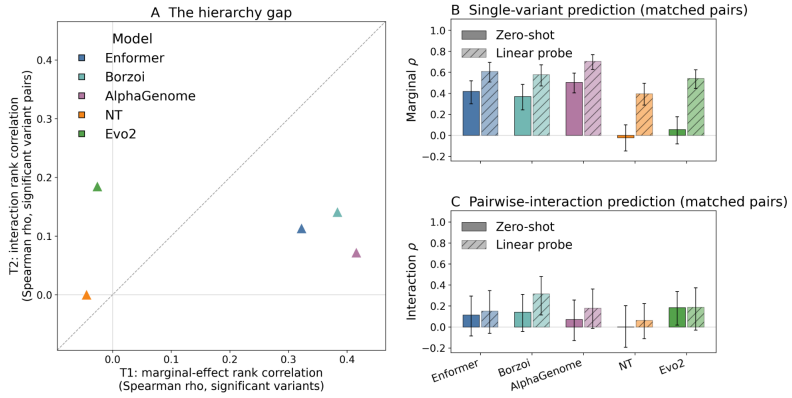


Figure 2: The hierarchy gap. **(A)** Zero-shot Spearman ρ on significant Siraj single variants (T1, x-axis) versus significant interaction pairs (T2, y-axis); the diagonal marks equal skill. **(B, C)** Single-variant (B) and pairwise-interaction (C) ρ on the same matched set of SNV–SNV pairs, zero-shot (solid) and linear probe (hatched). Error bars: 95% locus-level cluster-bootstrap intervals (2,000 resamples).

ing in every cell line; the largest gains come from the gLMs, whose embeddings encode variant-effect information that their likelihoods do not reveal. Kircher, a dense per-element scan evaluated element-held-out, is the exception for S2F models: their zero-shot scores are already strong ($|\rho| = 0.63$–0.72) and probing lowers them (AlphaGenome 0.72→0.62, Enformer 0.63→0.49, Borzoi 0.68→0.45; Table A2), plausibly because an element-held-out probe must generalize across few elements, whereas the pretrained heads already encode cell-type-specific activity; gLMs again gain substantially (NT 0.10→0.39; Evo 2 0.18→0.46). On Abell, correlations within significant variants are uniformly weak ($|\rho| \leq 0.09$ zero-shot, ≤ 0.15 probed), reflecting the narrow dynamic range of confidently active variants rather than absence of signal: S2F models still enrich allelic hits roughly fourfold among their top 1% of predictions (Appendix Fig. D1).

5.3 Interactions: regression (T2) and detection (T3)

Regression. On the 143 high-confidence Siraj pairs, zero-shot interaction ρ ranges from 0.00 (NT) to 0.19 (Evo 2); among S2F models it is 0.07 (AlphaGenome) to 0.14 (Borzoi) (Table 1; Appendix Fig. E1). NT's four-haplotype likelihood contrasts are uncorrelated with ϵ, whereas Evo 2's show a weak positive association ($\rho = 0.19$, 95% CI 0.02–0.34). Probing raises ρ for Borzoi (0.14→0.32), AlphaGenome (0.07→0.18) and Enformer (0.11→0.15), increase it slightly for NT (0.00→0.06), and leaves Evo 2 unchanged (0.19). On Abell, no model exceeds $|\rho| = 0.06$ under either readout (Appendix Table A2).

Detection. The applied question is whether a model can find interacting pairs among all candidates. Ranking Siraj's 2,164 candidate pairs by $|I_{\text{model}}|$, AlphaGenome attains AUROC 0.653 and

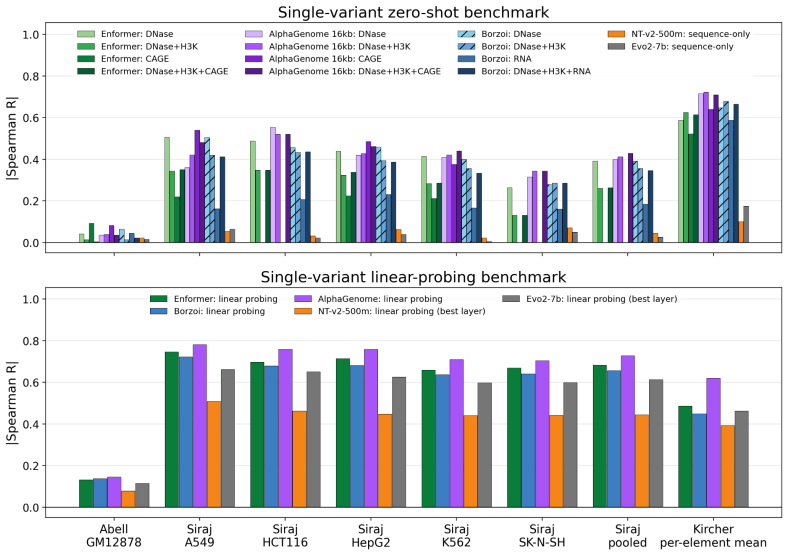


Figure 3: Single-variant benchmark on significant MPRA variants (T1). Top: zero-shot |Spearman ρ| for Enformer, AlphaGenome and Borzoi across track sets, and sequence-only NT-v2 and Evo 2-7b. Bottom: linear probes on frozen representations (best prespecified layer for gLMs). Siraj "pooled" aggregates cell types; Kircher reports mean per-element correlation across shared elements. Significance follows each study's definition.

AP 0.232 against a prevalence of 0.066, about 3.5× over random but far from usable (precision 31% at 20% recall). Borzoi (0.651) and Enformer (0.623) follow, NT reaches 0.584, and Evo 2 is at chance (0.501). A model-free distance baseline (1/distance) reaches AUROC 0.619, on par with Enformer (0.623) and above both gLMs. Probing raises AP for every model (e.g., Evo 2 0.08→0.23, AlphaGenome 0.23→0.31) and compresses the spread among S2F models, but the best AUROC remains 0.68 (Appendix Fig. E2). On Abell, discrimination is weak and unstable across readouts (zero-shot AUROC 0.44–0.63; Evo 2 0.44→0.61 and Borzoi 0.63→0.56 under probing), consistent with an underpowered near-null control rather than a model failure (Appendix E). With only 19 positives this set has limited power, so we treat Abell as a secondary, low-power dataset rather than an independent replication. Negative-class subsampling at multiple ratios leaves every model's AUROC and the Siraj ordering unchanged (Appendix Fig. E3), so the Siraj signal is robust to prevalence, and the apparent Abell AP gains under subsampling reflect prevalence rather than discrimination.

Comparison with a distance-only baseline. Nearby variant pairs are more often called non-additive in Siraj, although measured interaction magnitude itself does not depend on distance (Appendix E). To test whether models add pair-specific information beyond proximity, we use inverse inter-variant distance as a baseline under both readouts (Table 1). Zero-shot, pairs are ranked by 1/distance directly. Under probing, a one-feature linear (T2) or logistic (T3) model on 1/distance is fit with the same grouped five-fold cross-validation as the model probes. The zero-shot baseline is competitive with every zero-shot model. On interaction regression it exceeds Enformer, AlphaGenome and NT ($\rho = 0.145$ vs 0.113, 0.072 and 0.000), equals Borzoi (0.141), and is exceeded only by Evo 2 (0.185), a difference well within one standard error at $n = 143$ (SE of $\rho \approx 0.08$). On detection it matches Enformer (AUROC 0.619 vs 0.623) and trails AlphaGenome and Borzoi by only ˜0.03 (0.653, 0.651). The fitted baseline is weaker than its unfitted counterpart (T2 $\rho = -0.221$, AUROC 0.582, AP 0.125). The extreme skew of 1/distance (30 of the 143 pairs lie ≤ 4 bp apart) makes the out-of-fold fit unstable, and its sign reverses even though the rank association is positive. Every probed model exceeds the fitted baseline on all three metrics, but the more informative comparison is with the unfitted ranking. Every probed model except NT exceeds its T2 ρ of 0.145, Borzoi by about two standard errors (0.315), and probed S2F models exceed its AUROC of 0.619 by ˜0.06. Models do outperform distance in early precision: all S2F models under both readouts and all probed models attain higher AP than either baseline (e.g., probed AlphaGenome 0.314 vs

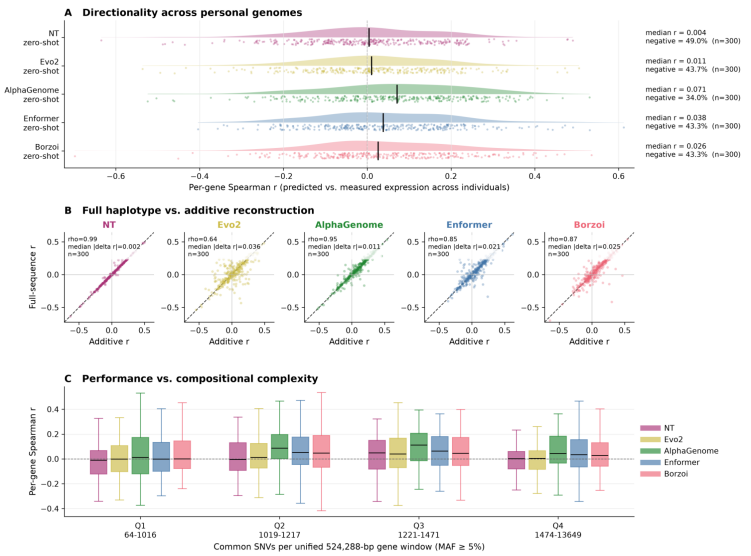


Figure 4: Personal-genome prediction (T4, GEUVADIS, 421 individuals). **(a)** Distribution of per-gene Spearman r between zero-shot full-haplotype predictions and measured expression across individuals; bars mark medians and the annotation gives the fraction of negatively correlated genes. **(b)** Per-gene r from the full-haplotype prediction versus the additive reconstruction of the same model's isolated variant effects; the dashed line is identity. **(c)** Per-gene r by quartile of common-SNV burden in a common 524,288-bp TSS window (MAF $\geq 5\%$). No probe was fit on evaluation outcomes.

0.135 unfitted and 0.125 fitted). Zero-shot interaction predictions therefore carry little pair-specific information beyond proximity; what probing adds is concentrated among the top-ranked pairs.

Anatomy of the gap. Better marginal prediction does not translate into better interaction prediction. On Siraj, zero-shot marginal performance on confident variants rises in the order NT $\approx$ Evo 2 $<$ Enformer $<$ Borzoi $<$ AlphaGenome ($\rho = 0.32 \rightarrow 0.42$ across S2F models), but interaction ρ is small for all models (0.00–0.19) and does not follow this order: AlphaGenome, the strongest marginal predictor, is the weakest S2F model on interactions (0.07). The gap is not explained by distance. Controlling for distance leaves interaction ρ essentially unchanged (≤ 0.02 for every model; Appendix E). Four of five models also predict interaction magnitudes that decay steeply with distance, although measured magnitudes do not (Appendix Fig. E4), a proximity prior without the pair-specific logic. What signal exists is concentrated at the very top of the ranking: current models can nominate a handful of candidate interacting pairs for follow-up but cannot capture the interaction landscape.

5.4 Personal genomes: full haplotypes are no better than additive reconstructions (T4)

If models compose variants, predictions from complete personal haplotypes should outperform the additive reconstruction built from the same model's isolated effects. Figure 4 shows that they do not. Across 300 genes, per-gene Spearman correlations between predicted and observed expression across 421 individuals are low for every model (median $\rho = 0.004$ for NT, 0.011 for Evo 2, 0.026 for Borzoi, 0.038 for Enformer and 0.071 for AlphaGenome), and a large fraction of genes are predicted with the wrong sign (34–49% negative; Fig. 4a). The directionality problem reported by Sasse et al. (2023) and Huang et al. (2023) has therefore not been resolved by scale: it persists from earlier architectures such as Basenji and Enformer into state-of-the-art S2F models and gLMs, including AlphaGenome and Evo 2.

More importantly, full-haplotype and additive correlations are nearly identical gene by gene (Fig. 4b). They agree with $\rho = 0.99$ (NT), 0.95 (AlphaGenome), 0.87 (Borzoi) and 0.85 (Enformer), with median $|\Delta\rho_g|$ between 0.002 and 0.025. Only Evo 2 shows appreciable scatter ($\rho = 0.64$, median $|\Delta\rho_g| = 0.036$), yet its full prediction is no more accurate than its additive one. Median

$\Delta\rho_g$ is within ± 0.002 of zero for all five models, and the fraction of genes with $\Delta\rho_g > 0$ is 47–54% (Appendix Table F1). Finally, if joint inference mattered, it should matter most where haplotypes carry many variants. Stratifying genes by common-variant burden shows no such effect: |median $\Delta\rho_g$| is at most 0.007 in every quartile for every model (Fig. 4c). Per-gene accuracy itself shows no monotonic trend with burden (Spearman 0.02–0.06, all $p > 0.3$); it is near zero in the least variant-dense quartile, peaks in the third, and is lower again in the most variant-dense quartile, with AlphaGenome highest in each quartile. On natural haplotypes, accounting for variant combinations therefore adds no measurable accuracy over each model's additive reconstruction. From the models' point of view, natural variant combinations behave as sums of single-variant effects.

6 DISCUSSION AND LIMITATIONS

Why the gap is expected, and why it matters. Nothing in the training of current genomic sequence models rewards composition. S2F models are trained on functional readouts of essentially one reference genome and never observe two haplotypes of the same element differing at two positions, so they receive no direct supervision for how variant effects combine. gLMs observe cross-species and cross-individual variation but are trained on likelihood, an observable that carries little regulatory-interaction signal. Probes trained on marginal effects did not recover strong interaction prediction: a linear head trained only on marginal effects recovers substantial single-variant information from every frozen representation, yet applied to the four haplotypes of a pair it recovers little interaction information, so the limitation lies in what the representations encode, not merely in the zero-shot readout. Yet the intended downstream uses of these models, causal-variant prioritization on real haplotypes, rare-variant interpretation and regulatory sequence design, are precisely the settings where composition matters; a model that is additive by construction silently degrades to the additive baseline in all of them, as T4 shows on natural genomes.

Implications for next-generation models. Our results point in three directions. (i) Training data with haplotype contrast: Personal genomes paired with individual-level functional readouts, or engineered multi-variant assays, supply the supervision that single-reference training lacks. While diplotype MPRAs are currently small in scale, emerging high-throughput platforms like STAG-seq by Liu et al. (2025), which pairs single-cell genotyping with transcriptomics to explicitly capture monoallelic and biallelic variant effects, demonstrate that generating this combinatorial training data at scale is now feasible.(ii) Interaction-aware objectives: training or fine-tuning on Eq. 1-type contrasts, i.e. predicting the effect of a variant set relative to its additive expectation, turns epistasis from an emergent hope into a supervised target. (iii) Compositional evaluation as a standard: single-variant benchmarks reward what models already do; T2–T4-style tasks in standard suites would expose the gap during development rather than after deployment.

Limitations. Our scores are one defensible choice among several; Appendix D shows that marginal-task conclusions are robust to fixed track sets, and linear probes give a lower bound on what a readout can recover from frozen representations; nonlinear or interaction-supervised readouts, and fine-tuning, could recover more. The high-confidence interaction sets are small (143 Siraj pairs; 19 Abell pairs), which limits the precision of T2 and makes Abell a low-power secondary dataset rather than an independent replication. We do not yet report an experimental noise ceiling for ϵ; replicate reliability would convert "models capture little of the measured interaction signal" into "models capture x% of the explainable signal" and is the highest-value addition to the benchmark. AlphaGenome was evaluated at 16 kb for MPRA tasks; longer contexts may matter for distal pairs, though most measured pairs are proximal. Finally, T4 inherits the known limits of personal-genome prediction (low per-gene correlations, trans and environmental variance), which is why it is framed as a within-model contrast rather than an absolute accuracy benchmark. Three further caveats bear on interpretation. First, the magnitude of ϵ depends on measurement scale, and part of the measured non-additivity may reflect a global nonlinearity of the reporter readout (global epistasis) rather than pair-specific regulatory logic; separating the two, for example by fitting an additive-latent model with a monotone link on single-variant data, is an important extension. Second, MPRAs measure reporter activity of short oligos, whereas S2F models are scored in genomic context through accessibility and histone readouts, so the context and the readout differ between measurement and prediction. Third, we do not yet include MPRA-trained models or supervised reference models trained on diplotype data, which would anchor the benchmark between the zero-shot models and the noise ceiling.

REPRODUCIBILITY STATEMENT

All predictions derive from publicly available models applied to published sequences; no model was retrained. The supplementary materials include harmonized ground-truth and prediction tables, per-figure backing tables, track-selection manifests, analysis and figure scripts, and QC summaries documenting joint coverage and harmonization (Appendix A).

ETHICS STATEMENT

This work uses only published, de-identified MPRA measurements, public GEUVADIS/1000 Genomes data and public model weights; no human-subject data were generated. Improved interpretation of non-coding variation carries standard dual-use considerations for genetic privacy, which a benchmark of published loci does not materially change.

AI USE STATEMENT

Generative AI tools assisted with analysis and plotting code and with editing for length. All AI-assisted code was reviewed and its outputs verified against source data tables by the authors, and all quantitative claims were checked against the analysis tables. The authors take full responsibility for the content.

REFERENCES

1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526, 68–74.

Abell, N. S., DeGorter, M. K., Gloudemans, M. J. et al. (2022). Multiple causal variants underlie genetic associations in humans. *Science*, 375(6586), 1247–1254.

Avsec, Ž., Agarwal, V., Visentin, D. et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18, 1196–1203.

Avsec, Ž., Weilert, M., Shrikumar, A. et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53, 354–366.

Avsec, Ž., Latysheva, N., Cheng, J. et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649, 1206–1218.

Brixi, G., Durrant, M. G., Ku, J. et al. (2026). Genome modelling and design across all domains of life with Evo 2. *Nature*, 652, 1349–1361.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J. et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22, 287–297.

de Almeida, B. P., Reiter, F., Pagani, M., Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, 54, 613–624.

de Boer, C. G., Vaishnav, E. D., Sadeh, R. et al. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, 38, 56–65.

Hingerl, J. C., Martens, L. D., Karollus, A. et al. (2025). scooby: modeling multimodal genomic profiles from DNA sequence at single-cell resolution. *Nature Methods*, 22, 2275–2285.

Huang, C., Shuai, R. W., Baokar, P. et al. (2023). Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55, 2056–2059.

Karollus, A., Mauermeier, T., Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24, 56.

Kelley, D. R., Reshef, Y. A., Bileschi, M. et al. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5), 739–750.

Kircher, M., Xiong, C., Martin, B. et al. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, 10, 3583.

Lal, A., Karollus, A., Gunsalus, L. et al. (2026). Decoding sequence determinants of gene expression in diverse cellular and disease states. *Nature Methods*, 23, 1138–1151.

Lappalainen, T., Sammeth, M., Friedländer, M. et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501, 506–511.

Linder, J., Srivastava, D., Yuan, H., Agarwal, V., Kelley, D. R. (2025). Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nature Genetics*, 57, 949–961.

Liu, K., Rasmussen A., Han W., et al. (2025). Sensitive Transcriptomics and Genotyping reveals function of genetic variants in immunity, bioRxiv, doi:10.1101/2025.03.12.642879.

Marin, F. I., Teufel, F., Horlacher, M. et al. (2024). BEND: Benchmarking DNA language models on biologically meaningful tasks. ICLR 2024.

Maurano, M. T., Humbert, R., Rynes, E. et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190–1195.

Phillips, P. C. (2008). Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9, 855–867.

Sasse, A., Ng, B., Spiro, A. E. et al. (2023). Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics*, 55, 2060–2064.

Shigaki, D., Adato, O., Adhikari, A. N. et al. (2019). Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Human Mutation*, 40(9), 1280–1291.

Siraj, L., Castro, R. I., Dewey, H. B. et al. (2026). Functional dissection of complex trait variants at single-nucleotide resolution. *Nature*, `https://doi.org/10.1038/s41586-026-10121-6`.

Tewhey, R., Kotliar, D., Park, D. S. et al. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165(6), 1519–1529.

Zhou, J., Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12, 931–934.

A BENCHMARK DATA, HARMONIZATION, COVERAGE AND FULL RESULTS

Harmonization. Predictions and ground truth are joined on variant/pair identifier and cell type after normalizing cell-type labels between sources (e.g., HEPG2→HepG2, SKNSH→SK-N-SH). Normalizing these labels restores approximately half of the Siraj pairwise rows that would otherwise be lost during joining.

Significance labels. Siraj significance uses the study's precomputed emVar flag; because the accompanying raw FDR column can contain spuriously small values for small-effect rows, we do not threshold that column directly. Abell uses the study's allelic-hit definition (FDR < 0.05 on both the expression and allelic tests in at least one pool) and the published non-additive pair flags.

Interaction ground truth. T2 uses each study's reported ϵ. For Siraj analyses requiring ϵ outside the reported high-confidence subset (e.g., Appendix E), we compute the activity-derived contrast $\epsilon = y_{\text{altalt}} - y_{\text{altref}} - y_{\text{refalt}} + y_{\text{refref}}$. This derived quantity is used only for supplementary analyses and not for T2. In Abell, the reported ϵ equals $y_{\text{joint}} - y_{\text{var1}} - y_{\text{var2}}$ on variant-effect activities (REF baseline zero by construction).

Coverage. Every figure and table is computed on fixed common sets: units scored by all five models under both readouts (track models also require a modality-balanced score), with ground truth, and for pairs, both constituent variants in the single-variant set. Of 9,988 Siraj candidate pair rows, 8,698 have track-model predictions under the track/cell-type mappings used here and all 9,988 have gLM predictions; 8,311 rows (2,164 unique pairs; 143 of the 178 official non-additive pairs) enter the common set. For Abell, 1,673 of 2,038 pairs enter (all 19 non-additive pairs). The single-variant sets contain 277,312 Siraj variants (1,111,996 variant × cell-type values), 23,601 Abell variants and 12,466 Kircher SNVs.

Personal genomes. Cohort, expression normalization, TSS windows and genotype-to-haplotype construction follow Huang et al. (2023). The 300-gene evaluation subset was selected blind to expression and model performance: the 389 genes evaluable by all five models were stratified into common-variant-burden quartiles, and 75 per quartile were taken by a deterministic hash rank.

Table A1: Benchmark datasets. Siraj and Abell counts reflect benchmark-level harmonization and may differ from the original libraries; model-specific coverage is described above.

Dataset	Assay / cohort	Benchmark units	High-confidence subset	Tasks
Siraj et al. (2026)	MPRA, 5 cell types	277,312 variants (1,111,996 variant × cell-type); 2,164 pairs	33,790 emVars (57,908 variant × cell-type); 143 interaction pairs	T1–T3
Abell et al. (2022)	MPRA, GM12878	23,601 variants; 1,673 pairs	1,043 allelic hits; 19 non-additive pairs	T1–T3
Kircher et al. (2019)	Saturation MPRA, 14 elements	12,466 SNVs (10 elements)	5,332 significant variants (FDR < 0.05)	T1
GEUVADIS	Phased 1000G genotypes + LCL RNA-seq	421 individuals; 3,259 eQTL genes	300-gene evaluation subset	T4

Table A2: Full results across Siraj, Abell and Kircher. T1: Spearman ρ for all variants / the significant subset. T2: Spearman ρ against the measured interaction effect ϵ. T3: AUROC and average precision (AP) for detecting published non-additive pairs. The 1/distance baseline has no sequence readout; Kircher has no pair task.

Model	Readout	T1 ρ (all / sig.)	T2 ρ	T3 AUROC	T3 AP
Siraj et al. (2026)					
Enformer	Zero-shot	0.066 / 0.322	0.113	0.623	0.194
Borzoi	Zero-shot	0.091 / 0.383	0.141	0.651	0.197
AlphaGenome	Zero-shot	0.101 / 0.416	0.072	0.653	0.232
NT	Zero-shot	−0.036 / −0.044	0.000	0.584	0.136
Evo 2	Zero-shot	−0.031 / −0.026	0.185	0.501	0.081
Enformer	Probe	0.224 / 0.683	0.152	0.676	0.231
Borzoi	Probe	0.217 / 0.657	0.315	0.683	0.261
AlphaGenome	Probe	0.250 / 0.728	0.179	0.681	0.314
NT	Probe	0.159 / 0.445	0.063	0.589	0.150
Evo 2	Probe	0.206 / 0.614	0.187	0.615	0.227
1/Distance	—	–	0.145	0.619	0.135
Abell et al. (2022)					
Enformer	Zero-shot	−0.001 / 0.008	0.021	0.547	0.015
Borzoi	Zero-shot	0.012 / 0.027	−0.031	0.632	0.022
AlphaGenome	Zero-shot	0.015 / 0.038	0.020	0.514	0.015
NT	Zero-shot	−0.004 / 0.022	−0.022	0.548	0.028
Evo 2	Zero-shot	−0.006 / 0.016	−0.010	0.435	0.029
Enformer	Probe	0.046 / 0.131	−0.055	0.567	0.014
Borzoi	Probe	0.040 / 0.137	−0.020	0.558	0.015
AlphaGenome	Probe	0.045 / 0.146	−0.011	0.513	0.014
NT	Probe	0.029 / 0.078	0.015	0.530	0.014
Evo 2	Probe	0.046 / 0.115	0.009	0.608	0.018
1/Distance	—	–	−0.013	0.482	0.012
Kircher et al. (2019)					
Enformer	Zero-shot	0.458 / 0.639	–	–	–
Borzoi	Zero-shot	0.508 / 0.678	–	–	–
AlphaGenome	Zero-shot	0.541 / 0.720	–	–	–
NT	Zero-shot	0.043 / 0.100	–	–	–
Evo 2	Zero-shot	0.090 / 0.176	–	–	–
Enformer	Probe	0.330 / 0.486	–	–	–
Borzoi	Probe	0.297 / 0.448	–	–	–
AlphaGenome	Probe	0.445 / 0.619	–	–	–
NT	Probe	0.254 / 0.393	–	–	–
Evo 2	Probe	0.307 / 0.464	–	–	–

B SEQUENCE CONSTRUCTION AND ZERO-SHOT SCORING

Haplotype construction. For each variant or pair we insert the alternative allele(s) into GRCh38 and extract a window centred on the variant or the pair midpoint, so the four haplotypes REF, A, B, AB share identical flanks (Fig. 1B). Windows follow each model's native input: 196,608 bp (Enformer), 524,288 bp (Borzoi), 16,384 bp (AlphaGenome), 12 kb (Nucleotide Transformer) and 8,192 bp (Evo 2). For T4 we build each individual's two phased haplotypes (hg19, matching Huang et al., 2023) over a TSS-centred window (Borzoi 524,288; Enformer 393,216; AlphaGenome 131,072; Evo 2 16,384; NT 12,282 bp). Primary cross-model analyses use SNVs and SNV pairs scored by all models. No model is fine-tuned.

Zero-shot S2F scores. For each model m and assayed cell type c we select cell-type-matched accessibility (DNase/ATAC, set $A_{m,c}$) and histone-modification tracks ($H_{m,c}$); exact track manifests are provided in the supplementary materials. Predictions are pooled around the locus (Enformer: central 3 bins, $\log(1+x)$; Borzoi: central 33 bins, $\log(1+x)$; AlphaGenome: central 501 bp, or 2,001 bp for histone outputs, $\log_2(1+x)$). With $S_{m,h,t}$ the pooled signal for haplotype h on track t, the marginal and interaction contrasts are

$$\Delta_{m,t}(v) = S_{m,\mathrm{alt},t} - S_{m,\mathrm{ref},t}, \qquad I_{m,t} = S_{m,11,t} - S_{m,10,t} - S_{m,01,t} + S_{m,00,t}. \tag{4}$$

where 0/1 denote reference/alternate alleles at the two positions. Because modalities differ in track count, contrasts are averaged within modality and then equally weighted across modalities, e.g. $M_m(v,c) = \frac{1}{2}[\mathrm{mean}_A \Delta_{m,t} + \mathrm{mean}_H \Delta_{m,t}]$, and likewise for $I_m(c)$. Transcriptional outputs are excluded from the primary score because no RNA/CAGE modality is available uniformly across models and cell types (robustness to track sets: Appendix D). Interaction contrasts are computed on each model's log-transformed outputs and compared with interactions on the log-scale MPRA activities reported by each study. As the magnitude of non-additivity depends on measurement scale, T2 is defined on these native log scales.

Zero-shot gLM scores. Nucleotide Transformer and Evo 2 assign a sequence score $s(x)$: for NT, the masked 6-mer pseudo-log-likelihood averaged over both strands; for Evo 2, the mean per-position autoregressive log-likelihood over the 8,192-bp window. The single-variant score is $\Delta s = s(x_{\mathrm{alt}}) - s(x_{\mathrm{ref}})$ and the interaction score is Eq. 1 applied to s: $I_{\mathrm{model}} = s(x_{AB}) - s(x_A) - s(x_B) + s(x_{REF})$.

C FROZEN-REPRESENTATION PROBES, SPLITS AND MATCHED EVALUATION

Linear probes and leakage-free splits. For each haplotype D we extract $h(D)$ over a prespecified 200-bp window: for S2F models a fixed pre-head embedding (AlphaGenome 1-bp post-decoder, 1,536-d; Enformer 128-bp trunk, 3,072-d; Borzoi 32-bp trunk, 1,920-d; pooled by base-pair overlap), and for gLMs mean-pooled hidden states from six prespecified layers (full sweeps in Appendix C; no test-set layer selection). Probes are trained only on single-variant effects, per dataset and per Siraj cell type, mapping $\Delta h_A = h(A) - h(REF)$ to $y_A - y_{REF}$.

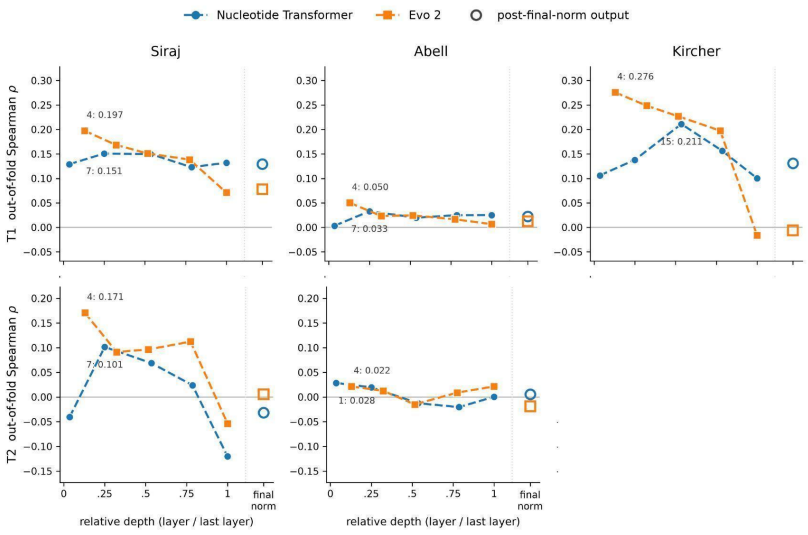
Evaluation uses five-fold grouped outer cross-validation in which groups are connected components of locus/construct identifiers, assayed-pair relationships and overlapping oligos, so both members of a pair, and every haplotype of a construct, fall in the same fold; standardization and the ridge penalty are chosen by group-disjoint inner cross-validation. Out-of-fold predictions give T1. For T2–T3 the same probe is applied without retraining to the four haplotypes of each held-out pair (pair windows: 200 bp at the midpoint for gLMs; 99 bp upstream of the left variant to 100 bp downstream of the right for S2F models), and the predicted interaction is

$$\hat{\epsilon} = f[h(11)] - f[h(10)] - f[h(01)] + f[h(00)]. \tag{5}$$

T2 compares $\hat{\epsilon}$ with measured ϵ; for T3, each coordinate pair is scored by its maximum $|\hat{\epsilon}|$ across cell types. For a linear probe $f(h) = w^T h + b$, this reduces to $\hat{\epsilon} = w^T[h(11) - h(10) - h(01) + h(00)]$. T2–T3 under probing therefore tests whether the non-additivity present in a frozen representation is aligned with the direction that predicts marginal effects. This is a lower bound on the interaction information the representation contains; information in orthogonal or nonlinear directions is not assessed.

C.1 LAYER SELECTION

Zero-shot likelihood scoring conflates two hypotheses: that the representation lacks regulatory-interaction information, or that the likelihood readout fails to expose it. The probes in the main text separate them. For each gLM we evaluated six prespecified layers (NT: embedding layer-norm, layers 1, 7, 15, 22 and the final layer; Evo 2: blocks 4, 10 and four further prespecified blocks); the probe values in Table A2 use layers selected by inner cross-validation, never on the test fold.



C.2 MATCHED-SET CONSTRUCTION AND UNCERTAINTY

For the T1–T2 comparison in Figure 2B–C, we use a matched set of single-variant and interaction measurements.

Metrics, matched sets and uncertainty. T1 and T2 differ in sample size, selection and target, so a raw T1–T2 difference could reflect scale rather than composition. Our headline gap (Fig. 2B–C) is

therefore computed on a matched set: the constituent variants of the same assayed SNV–SNV pairs, in the same cell type and oligo, with the same track aggregation as the interaction score (Appendix C gives the construction and a worked example). Confidence intervals are 95% locus-level cluster bootstraps (2,000 resamples). Because non-additive pairs are rare, AUROC is the headline detection statistic, and AP is reported against prevalence (sensitivity to negative subsampling: Appendix E). Harmonization, coverage and quality control are detailed in Appendix A.

D SINGLE-VARIANT ANALYSES (T1)

Robustness to score definitions. Re-running T1 with fixed assay-specific track sets (DNase; DNase+H3K; CAGE or RNA; their union; no per-variant selection) leaves all qualitative conclusions unchanged: on Siraj's significant variants per-cell-type ρ ranges from 0.13 to 0.55 across models and track sets with AlphaGenome strongest (Fig. 3, top), and Abell remains near zero under every definition. For interactions, median versus mean aggregation of predicted ϵ across tracks changes no conclusion.

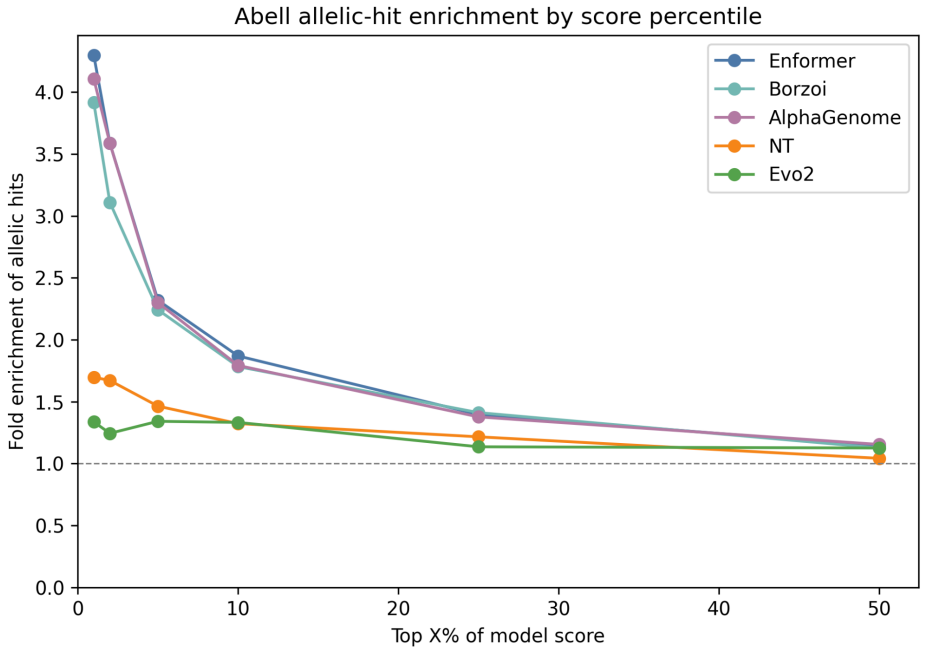


Figure D1: Abell-style variant prioritization across five models. Ground truth is Abell et al.'s allelic-MPRA-hit definition (BH-adjusted allelic $p < 0.05$ and expression $p < 0.05$; 1,043 hits among 23,601 variants). Fold enrichment of allelic hits among the top X% of each model's score distribution (X = 50, 25, 10, 5, 2, 1%) relative to that model's background rate; dashed line marks no enrichment.

E PAIRWISE INTERACTION ANALYSES (T2–T3)

E.1 INTERACTION REGRESSION AND DETECTION CURVES

Abell provides a secondary evaluation on its 19 labeled non-additive pairs; full results are reported in Appendix Table A2.

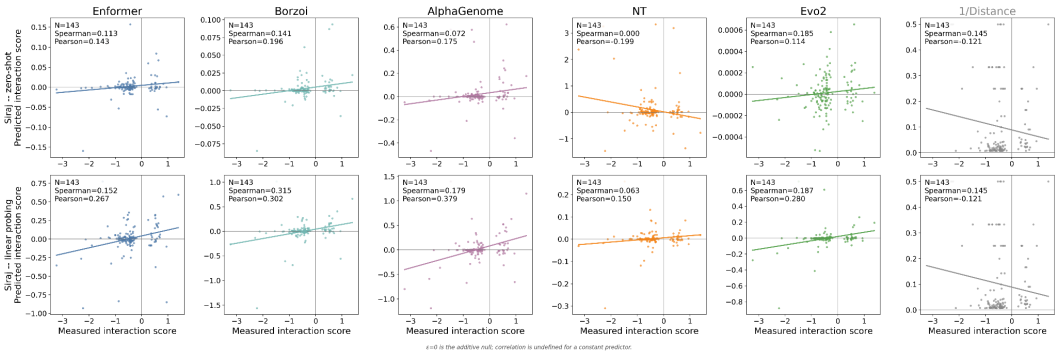


Figure E1: Measured versus predicted pairwise interaction effects (Siraj, 143 high-confidence pairs). Top: zero-shot; bottom: linear probe. Lines are least-squares fits; Spearman and Pearson correlations are annotated. Note the differing y-axis scales: gLM zero-shot contrasts are orders of magnitude smaller than S2F contrasts.

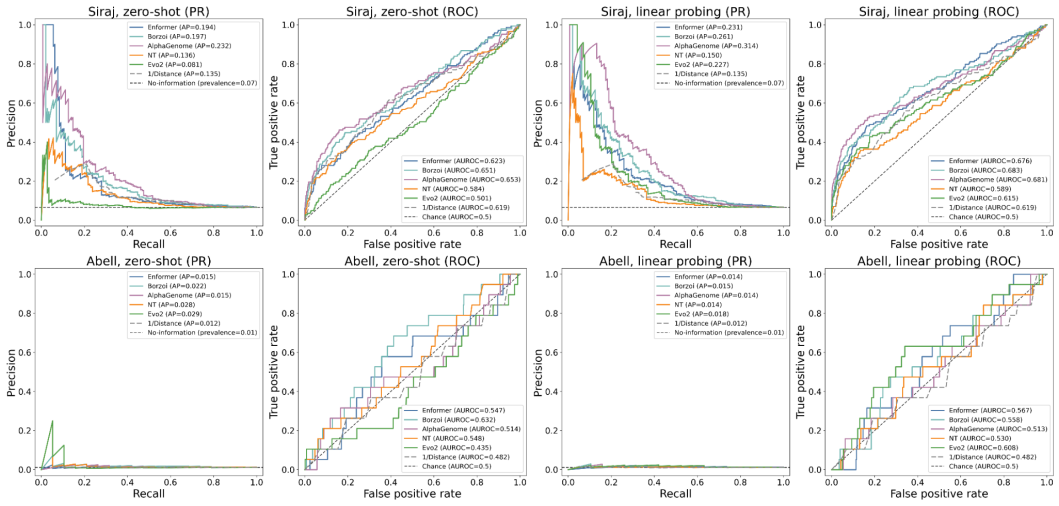


Figure E2: Non-additive pair detection (T3). Precision–recall and ROC curves for ranking the studies' non-additive pairs by |predicted interaction|, zero-shot (left pair) and linear probe (right pair). Top: Siraj (2,164 candidate pairs, 143 positives). Bottom: Abell (1,673 pairs, 19 positives). All candidate pairs are evaluated at natural prevalence (no subsampling), with the same pairs and labels for every model and both readouts. The model-free 1/Distance baseline is shown in grey. Dashed lines mark prevalence (PR) and chance (ROC). Abell discrimination is weak and unstable across readouts (AUROC 0.44–0.63).

E.2 PREVALENCE SENSITIVITY OF DETECTION METRICS

Both detection tasks have rare positive classes (Siraj 6.6%, 143 of 2,164 pairs; Abell 1.1%, 19 of 1,673), and average precision rises mechanically as negatives are subsampled regardless of real

signal. We recomputed PR and ROC for the zero-shot scores at negative-retention fractions from 100% down to 1% (two orders of magnitude; 50 random draws per fraction). Siraj: AUROC is stable at every fraction (S2F models 0.62–0.66; NT ≈ 0.58; Evo 2 ≈ 0.50) with a constant model ordering (AlphaGenome $>$ Borzoi $>$ Enformer $>$ NT $>$ Evo 2), while AP rises from 0.08–0.23 to 0.89–0.94 as expected, i.e. real but modest discriminative signal reported differently by the two metrics. Abell: AUROC is likewise flat across fractions for every model, so its AP gains under subsampling (from ≤ 0.03 to 0.55–0.67) are a prevalence artefact. Discrimination is near chance for AlphaGenome (0.50–0.52) and Evo 2 (0.43–0.44) and weak for Enformer and NT (≈ 0.55), with Borzoi the only clearly above-chance model (0.62–0.64); with 19 positives, all intervals are wide. AUROC is therefore the headline detection statistic in the main text.

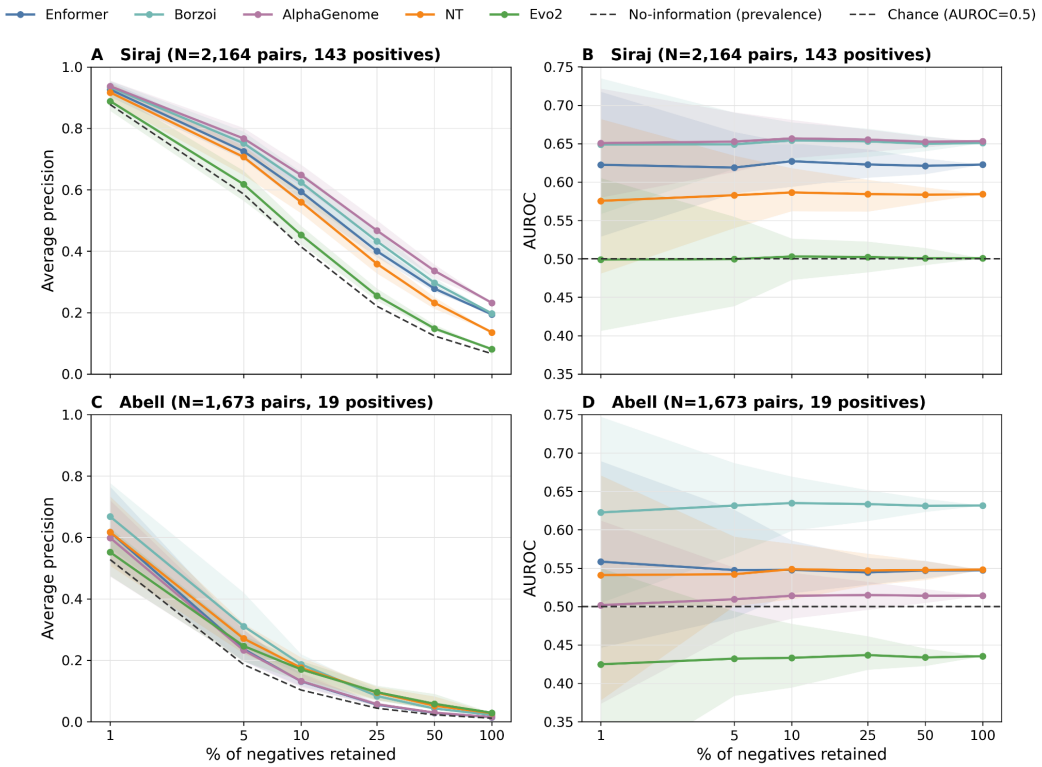


Figure E3: Detection performance across negative-subsampling levels. Average precision (A, C) and AUROC (B, D) for ranking non-additive pairs by zero-shot $|I_{\text{model}}|$ as negatives are subsampled from 100% to 1% of the candidate pool. Top: Siraj (2,164 pairs; 143 positives). Bottom: Abell (1,673 pairs; 19 positives). Lines show means over 50 subsamples and shaded bands indicate 95% intervals; dashed lines mark prevalence for AP and 0.5 for AUROC. AP increases as prevalence rises, whereas AUROC remains stable, indicating that AP changes under subsampling primarily reflect class prevalence rather than improved discrimination.

E.3 Distance dependence of interaction magnitude

Measured interaction magnitude shows no dependence on inter-variant distance in either study: median $|\epsilon|$ is flat across distance quintiles (Siraj $\rho = 0.02$, $p = 0.26$; Abell $\rho = 0.01$, $p = 0.55$; Siraj ϵ is activity-recoded for pairs without an official estimate). Models nevertheless predict it: zero-shot $|I_{\text{model}}|$ decays steeply with distance for Enformer, Borzoi, AlphaGenome and NT ($\rho = -0.27$ to -0.48, $p < 10^{-6}$ in both studies; 1.6–3.7$\times$ the median at the shortest distances), whereas Evo 2 is nearly flat. The models thus encode a proximity prior on interaction strength that the measurements do not support.

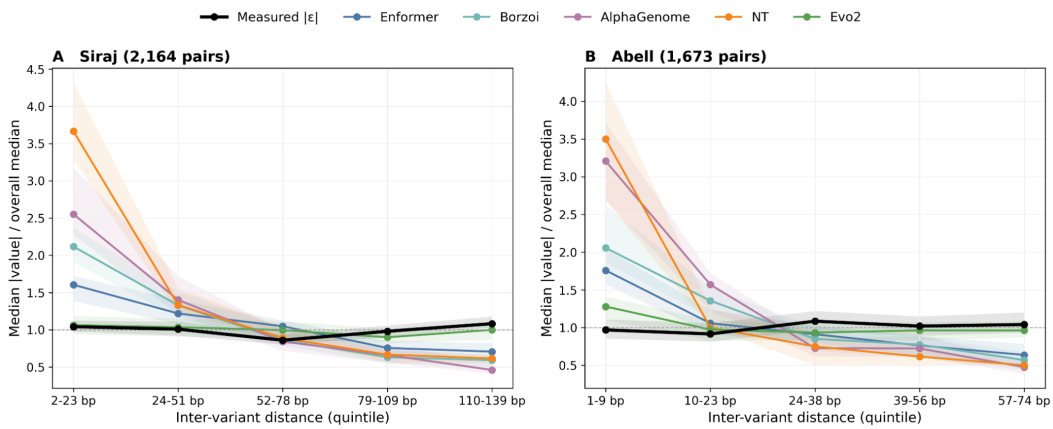


Figure E4: Interaction magnitude versus inter-variant distance. Median measured $|\epsilon|$ (black) and zero-shot predicted $|I_{\text{model}}|$ (colours) per distance quintile for Siraj (A; 2,164 pairs) and Abell (B; 1,673 pairs), using the common-coverage pair sets. Siraj cell-type rows are collapsed to one value per pair by maximum magnitude, and track models use the modality-balanced score. Each curve is divided by its own overall median so that trends, not scales, are compared across models; shaded bands are bootstrap 95% CIs (1,000 resamples of pairs within each bin). Measured magnitude is flat with distance, whereas predicted magnitude decays steeply for four of the five models.

F PERSONAL-GENOME ANALYSES (T4)

Personal-genome scoring. For models with a scalar gene score $S_g(x)$ (GM12878 CAGE pooled at the TSS for S2F models; mean log-likelihood for Evo 2), the full prediction for individual i averages the two phased haplotypes, and the additive reconstruction sums the model's own isolated effects weighted by dosage $d_{igv} \in \{0, 1, 2\}$:

$$\begin{aligned} P_{ig}^{\text{full}} &= \frac{1}{2}\left[S_g(x_g^{\text{ref}} \oplus V_{ig}^{(1)}) + S_g(x_g^{\text{ref}} \oplus V_{ig}^{(2)})\right], \\ P_{ig}^{\text{add}} &= P_g^{\text{ref}} + \frac{1}{2}\sum_v d_{igv}\delta_{gv}, \\ \delta_{gv} &= S_g(x_g^{\text{ref}} \oplus \{v\}) - S_g(x_g^{\text{ref}}). \end{aligned} \tag{6}$$

This preserves each model's marginal effects while removing all interactions among variants. NT lacks a tractable whole-sequence likelihood, so isolated effects are masked-token log-likelihood ratios in the reference context and joint scores are recomputed with all other haplotype variants retained. Per gene, ρ_g^{full} and ρ_g^{add} are Spearman correlations with observed expression across individuals, and $\Delta\rho_g$ is their difference. Variant burden is counted in a common 524,288-bp TSS window (biallelic SNVs, MAF $\geq 5\%$), used only for stratification. No parameter is fit to the GEUVADIS expression.

F.1 SUMMARY STATISTICS

Table F1: Full versus additive prediction on GEUVADIS (300-gene subset). Median per-gene Spearman r for the full-haplotype and additive predictions, median Δr, fraction of genes with $\Delta r > 0$, and the across-gene correlation between per-gene ρ_{full} and ρ_{add}.

Model	Genes	Median r_{full}	Median r_{add}	Median Δr	Genes $\Delta r > 0$	r(full, add)
Enformer	300	0.038	0.044	−0.002	142 (47.3%)	0.852
Borzoi	300	0.026	0.036	0.002	161 (53.7%)	0.872
AlphaGenome	300	0.071	0.076	0.001	155 (51.7%)	0.951
NT	300	0.004	0.003	0.000	153 (51.0%)	0.992
Evo 2	300	0.011	0.008	−0.001	146 (48.7%)	0.642