BENCHMARKING THE CAPABILITIES NEEDED TO DESIGN DNA

Anonymous authors
Paper under double-blind review

ABSTRACT

Genomic foundation models are increasingly used to propose DNA sequences and edits. But successful design requires capabilities that current benchmarks do not measure. Here we introduce a benchmark that assesses DNA Design Readiness Across Functional Tasks (DRAFT), which includes sequence priors, functional prediction, biological context, genetic interactions, and selection for a desired response. Across eight tasks built on five experimental datasets, we find that model choice and how model representations are accessed both affect performance. Frozen-representation outputs can match or exceed models' native outputs, and functional post-training substantially improves their predictive capabilities. These outputs can also transfer across experimental assays and genes, suggesting that models capture functional patterns beyond the data used to fit individual predictors. At the same time, strong prediction of average effects can conceal weaker prediction of context-specific effects, and accurate prediction of combined mutation effects does not imply accurate prediction of their interactions. In design tasks, model predictions can reduce the number of candidates needed to find a desired response, but experimental selection reveals substantial room for better choices, particularly for context-specific and multi-cell objectives. Overall, DRAFT shows that genomic models contain useful information for DNA design, while highlighting biological context and genetic interactions as important remaining limitations.

	LEADERBOARD	invalid-redacted-link
	CODE	anonymous.4open.science/r/draft-5709
🤗	DATA	invalid-redacted-link

1 INTRODUCTION

A biologist seeking to change the activity of a regulatory sequence may have hundreds of mutations to choose from. The same mutation can have different effects in different cell types or even when introduced into a slightly different version of the same sequence. A longstanding goal of genomic modeling is to guide experimental interventions and make the search more efficient. The value of models therefore depends on their ability to prioritize the changes most likely to produce the desired response before any experiments are performed.

Genomic models use different approaches to learn the complex and multifaceted patterns encoded in DNA, drawing on both sequence diversity and measurements of biological function. Models such as Evo2 learn from natural sequence variation (Brixi et al., 2026), while GPN-Star uses evolutionary relationships to predict functional constraint (Ye et al., 2026). Other models incorporate functional outcomes rather than relying on natural sequence variation and homology. AlphaGenome, for example, predicts expression, accessibility and splicing from functional genomic data (Avsec et al., 2026), while NTv3 combines sequence learning with such functional supervision (Boshar et al., 2025). Foundation models and task-specific models have guided the design of regulatory sequences with cell-type-targeted activity (Gosai et al., 2024) and, more recently, viable bacteriophage genomes (King et al., 2026). However, the central question remains: which features of sequence and function have these models learned well enough to guide experiments?

Looking only at overall predictive accuracy leaves that question unanswered. A sequence-based score may identify a mutation at an evolutionarily constrained position without any knowledge of how it impacts actual expression in a particular cell type. Accurate predictions for individual mutations do

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.

not establish whether a model captures interactions between mutations or if it can generalize their effects on a different genetic background. Similarly, distinguishing highly active genes from weakly active genes does not indicate which edit will produce a desired change in expression. Each of these cases calls for more precise testing.

Here, we introduce DRAFT *(DNA Design Readiness Across Functional Tasks)*, a benchmark organized around five key capabilities: Prior, Function, Context, Interaction, and Design. Together, our tasks measure what genomic models can predict and whether they can help select DNA sequences that produce a desired response. Table 1 summarizes our selected evaluations.

Related works. Benchmarking efforts such as BEND (Marin et al., 2024), GenBench (Liu et al., 2024), and OmniGenBench (Yang et al., 2025) provide systematic evaluations of genomic representations across downstream tasks. In parallel, model-guided regulatory sequence design has demonstrated cell-type-targeted activity experimentally (Gosai et al., 2024; Yin et al., 2025; Castillo-Hair et al., 2025), and NucleoBench (Shor et al., 2025) evaluates algorithms for searching sequence space using predictive models as scoring functions. DRAFT connects these directions by testing the predictive distinctions that matter for design: resolving context-specific effects, predicting interactions between mutations, and distinguishing candidates with different functional outcomes.

	Fig. 1A	Fig. 2A	Fig. 3B	Fig. 15B	Fig. 4A	Fig. 5A	Fig. 5C	Fig. 6A
	Prior loss bits ↓	Splicing effects ρ ↑	Cell contrast ρ ↑	Tissue ID % ↑	Background interaction ρ ↑	Single-edit rank percentile ↓	Cell selectivity U ↑	Multi-cell gain $\log_2$ ↑
Base composition	1.974	—	—	—	—	—	—	—
Evo2 20B	1.388	0.367	0.041	—	0.050	43.5	-0.350	0.275
CENO 1B	1.549	—	0.036	9.8	0.010	48.9	-0.318	-0.581
GENERATORv2	1.711	0.388	0.066	17.6	0.046	44.7	-0.256	0.234
DNABERT-2	—	0.198	0.076	5.9	0.020	48.6	-0.302	0.233
NTv3 100M pre	1.537	0.349	0.076	7.8	0.010	47.3	-0.207	0.339
NTv3 100M post	1.492	0.502	0.244	51.0	0.139	30.3	0.211	0.144
Native outputs	—	0.168	0.203	27.5	0.152	32.0	0.200	0.344‡
NTv3 650M pre	1.447	0.187	0.010	7.8	0.009	50.4	-0.061	0.368
NTv3 650M post	1.375	0.564	0.256	56.9	0.187	30.0	0.287	0.568
Native outputs	—	0.461	0.239	45.1	0.170	29.3	0.250	0.482‡
AlphaGenome	—	0.650	0.402	64.7	0.269	25.7	0.515	0.833
Native outputs	—	0.621	0.398	68.6	0.260	24.3	0.384	0.707‡
Enformer	—	—	0.282	31.4	0.181	31.5	0.290	0.700‡
k-mer ridge	—	0.318	0.211	7.8	0.032	45.0	-0.097	0.394
Task-trained ResNet	—	0.488	0.020	2.0	-0.001	47.0	-0.215	0.347
SpliceAI	—	0.674	—	—	—	—	—	—
Experimental reference†	—	0.809	0.594	99.0	0.421	11.8	0.962	1.602

Table 1: **The DRAFT benchmark.** Metrics from all experiments collected in one overview table. Referenced figures and their corresponding sections provide detailed information and explorations. Deeper shading indicates better performance within each column. Foundation-model results use embedding-fitted linear heads, except for prior loss (native sequence logits) and Enformer/indented rows (native outputs). For multi-cell design, ‡ marks a linear head fitted to native outputs. †: Experimental references compare donor or replicate subsets; background interaction uses HepG2–K562 transfer. Measurement details are in Appendix A.2.

2 A FRAMEWORK FOR EVALUATING GENOMIC MODELS

DRAFT is centered around five complementary capabilities that connect sequence and functional understanding to genomic design.

Sequence priors: How probable is the DNA? We measure how well a model predicts a masked base x_i from surrounding DNA:

$$p_\theta(x_i \mid \text{available DNA context}) \tag{1}$$

This is computed via the negative log-probability, $-\log_2 p_\theta(x_i \mid \text{available DNA context})$, over all queried positions. We report bits per base where lower values indicate better predictions.

Functional prediction: Can sequence predict a functional outcome? For a reference sequence x and an edit δ, the mutation effect Δy is the change in an experimental measurement between x and $x + \delta$. A predictor f_θ estimates this effect:

$$(x,\ x+\delta) \xrightarrow{f_\theta} \widehat{\Delta y} \tag{2}$$

We test mutation-effect predictions on held-out genes, then evaluate whether the fitted outputs transfer to another assay.

Biological context: Is context-dependent genomic function captured? For a DNA sequence with an experimental measurement y_c measured in several cells or tissues c, we separate its mean across the evaluated contexts $\bar{y}$ from the context-specific deviation κ_c:

$$y_c = \underbrace{\bar{y}}_{\text{mean across contexts}} + \underbrace{\kappa_c}_{\text{context-specific deviation}} \tag{3}$$

We test mutation-effect differences between cells, expression differences between tissues, and whether predicted profiles across held-out genes identify the correct tissue.

Genetic interactions: Does one mutation alter another's effect? Let Δy_1 and Δy_2 denote the effects of two individual mutations, and Δy_{1+2} their combined effect, all relative to a reference sequence in the same context. The interaction $\mathcal{I}_{1,2}$ is the non-additive part of the signal:

$$\Delta y_{1+2} = \underbrace{\Delta y_1 + \Delta y_2}_{\text{additive expectation}} + \underbrace{\mathcal{I}_{1,2}}_{\text{interaction}} \tag{4}$$

We hence score interaction separately since an accurate combined effect can mask poor predictions of interaction.

Genomic design: Which candidate achieves the desired response? Given a target response $\mathbf{y}_{\text{target}}$, the model chooses a candidate $a^\star$ among permitted edits or sequences $\mathcal{C}$ whose predicted response $\widehat{\mathbf{y}}(a)$ best matches the target $\mathbf{y}_{\text{target}}$:

$$a^\star = \arg\min_{a \in \mathcal{C}} \ell(\widehat{\mathbf{y}}(a),\ \mathbf{y}_{\text{target}}) \tag{5}$$

The loss ℓ measures mismatch with the target, using the responses relevant to each task: mutation effects or activity across cells.

Predictions and generalization. Most predictions use linear heads fitted to frozen model sequence embeddings. For mutation effects, the linear head weights w are fitted to the difference between variant and reference embeddings ($\phi_\theta(x)$) with $\widehat{\Delta y} = w^\top[\phi_\theta(x+\delta) - \phi_\theta(x)]$. These predictions and native outputs are reported separately. Detailed descriptions of the output types can be found in Appendix A.2. We compute correlations within each gene or regulatory sequence, then average across the G groups using $\rho_{\text{macro}} = G^{-1}\sum_{g=1}^{G}\rho_g$. This avoids the score being dominated by, e.g., one large gene. For downstream splits, we group related examples by gene, regulatory element, locus or sequence family. In addition to genomic foundation models, our comparisons include simple predictors, specialist models and experimental references where applicable.

3 SEQUENCE PRIORS: MEASURING WHAT MODELS KNOW ABOUT DNA

The models learn useful sequence priors. To compare causal and masked sequence models, we mask out the central base of 512 GRCh38 windows of 1,200bp. For this base reconstruction, NTv3 650M post and Evo2 have the lowest losses among models in Figure 1A. The foundation models outperform simple base composition statistics, but do not seem to have converged on prediction accuracy. For all NTv3 models, we used a fixed 12% additional masking from its training objective, and the final scores averaged the central-base loss over two draws.

Learned sequence knowledge is not always readily accessible. For NTv3 650M post, hiding only the central base gives a loss of 2.199 bits. Masking an additional 12% reduces this to 1.375, the lowest among tested models, although heavier masking eventually worsens performance (Figure 1B).

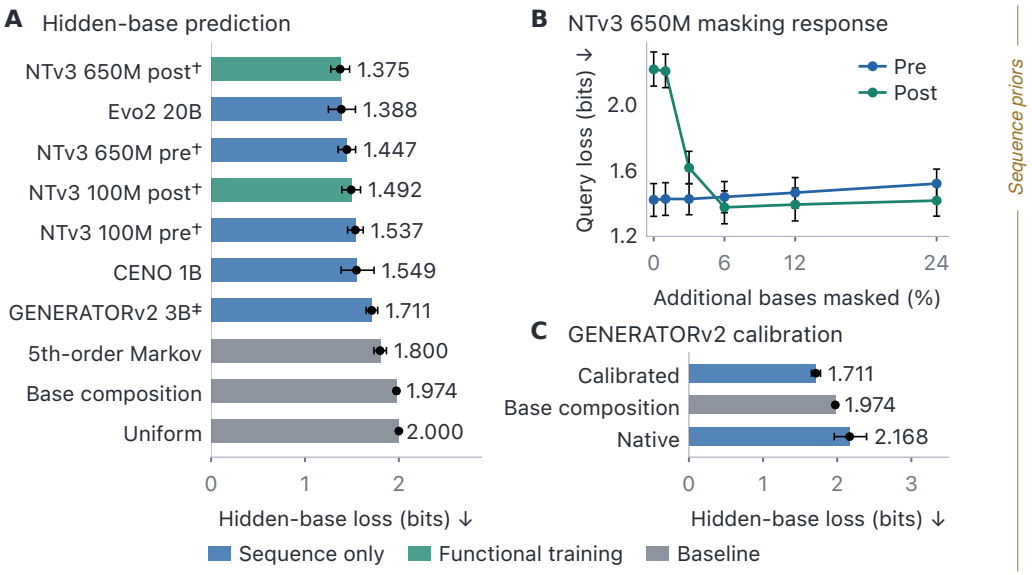

Figure 1: **Sequence-prior performance depends on the evaluation interface.** A: masked loss on 512 windows. NTv3 uses fixed 12% additional masking (†), averaging losses over two draws. GENERATORv2 uses temperature calibration fitted on separate data (‡). B: additional masking substantially improves post-trained NTv3 sequence loss. C: calibration improves GENERATORv2's predicted probabilities without changing base-prediction accuracy. Whiskers show 95% block-bootstrap intervals. Colors denote model/baseline category.

Averaging predicted probabilities across eight masking draws reduces loss further, to 1.346. We tested whether a fixed shift in internal activations could reproduce the benefit of additional masking. We estimated activation differences between additional-mask and single-mask queries, selected the layer and shift strength on separate data, and evaluated the shifted model on single-mask queries from held-out genomic blocks without updating model weights. For NTv3 650M post, accuracy rose from 29% to 56%, close to the pre-trained model's 55%; affine-logit calibration reached 30%. Full results, controls and uncertainty estimates are provided in Appendix B.2.

A simpler adjustment also improves GENERATORv2: temperature calibration lowers loss from 2.168 to 1.711 while leaving accuracy at 50.2% (Figure 1C). Together, these results show that measured sequence-prior performance depends on how the model is queried and its outputs are used.

These gains do not establish that sequence priors capture functional effects or variation between individuals. Evo2 and CENO reconstruct 55.2/62.4% of assay reference bases, yet their allele scores correlate only 0.045/0.029 with reporter effects (Appendix B.5). Our main prior evaluation uses GRCh38; extending strong sequence priors to natural haplotypes remains an important next step, as evident by our results in Appendix B.4.

4 Functional prediction: from sequence to experimental response

Mutation-effect prediction transfers. Next, we ask whether genomic models can predict how mutations affect RNA splicing. OpenSplice measures changes in exon inclusion across 590,104 mutations in 466 genes (Quarantani et al., 2026). On held-out genes, AlphaGenome's embedding-fitted linear head reaches macro-gene Spearman 0.650, compared with 0.621 for its native outputs and 0.564 for NTv3 650M post (Figure 2A). Specialist model SpliceAI reaches 0.674, although native heads retain their historical supervision and therefore differ from readouts fitted under the benchmark's training splits.

NTv3 provides a comparison of the same architecture before and after functional post-training. Post-training improves mutation-effect prediction at both model sizes, and the 650M post-trained readout outperforms both its matched random architecture (0.144) and a task-trained one-hot ResNet (0.488). Label-efficiency, direction and calibration analyses are reported in Appendix C.

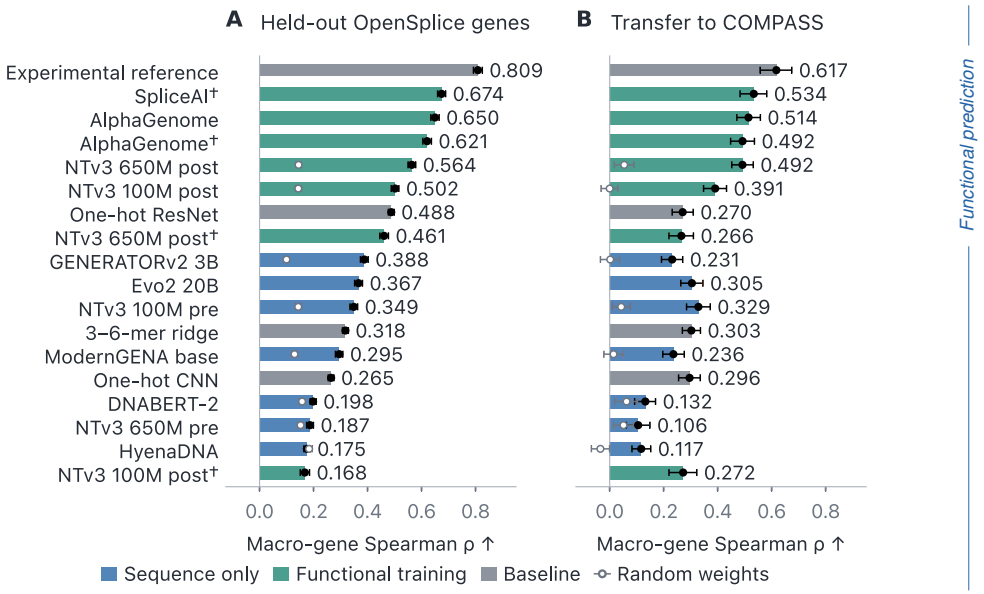

Figure 2: **Functional post-training supports transfer across assays and genes.** A: mutation-induced changes in exon inclusion on OpenSplice, with complete genes held out of readout fitting (466 genes). B: transfer of OpenSplice-trained predictors to 4,618 COMPASS variants in 88 genes excluded from source training, without COMPASS fitting or model selection. Scores average within-gene Spearman correlations. Experimental references compare repeat measurements; they are not ceilings. Native NTv3 (†) uses splice-boundary annotation probabilities; native heads retain historical supervision. Hollow gray points show random weights; whiskers are 95% gene-bootstrap intervals (Appendix C).

We then test whether the OpenSplice-trained readouts transfer to a different assay. Readouts trained on OpenSplice are evaluated on 4,618 COMPASS mutations across 88 genes excluded from the training set, without using COMPASS labels for fitting or selection (Koplik et al., 2025). Both AlphaGenome and NTv3 650M post retain substantial predictive accuracy for mutation effects averaged across cell lines, outperforming the k-mer baseline fitted on the OpenSplice training set (Figure 2B). Their performance is similar to that of native SpliceAI and AlphaGenome. These results suggest that the readouts capture functional relationships that remain useful beyond the assay and genes used for fitting.

Editing efficiency depends on genomic context. The K562 TRIP library tests editing efficiency across genomic locations while holding the reporter and editing program fixed (Mathis et al., 2025). Both post-trained NTv3 models improve over their pre-trained counterparts, while their differences from AlphaGenome are not statistically resolved (Appendix C.1).

5 Biological context: how does function depend on the context

Mutation effects differ between cells. Predicting a mutation's average effect does not establish whether a model knows where that effect is strongest. The Siraj assay measures the same mutations in HepG2 and K562, allowing us to evaluate both their mean effect and the difference between cells (Siraj et al., 2026). Across 128 regulatory elements, AlphaGenome's native, cell-matched accessibility outputs reach correlation 0.532 for the mean effect and 0.398 for the cell difference, without fitting to Siraj labels (Figure 3A/B). NTv3 650M post's fitted readouts reach 0.492 and 0.256, respectively. Both models predict the mean more accurately than the difference, with a larger gap for NTv3.

The difficulty extends beyond this cell pair. On COMPASS, AlphaGenome reaches correlation 0.479 for the average splice effect but -0.018 for deviations between cells; replicate agreement for these deviations is 0.151 (Appendix D). Thus, accurate average effects can coexist with little ability to resolve cell dependence.

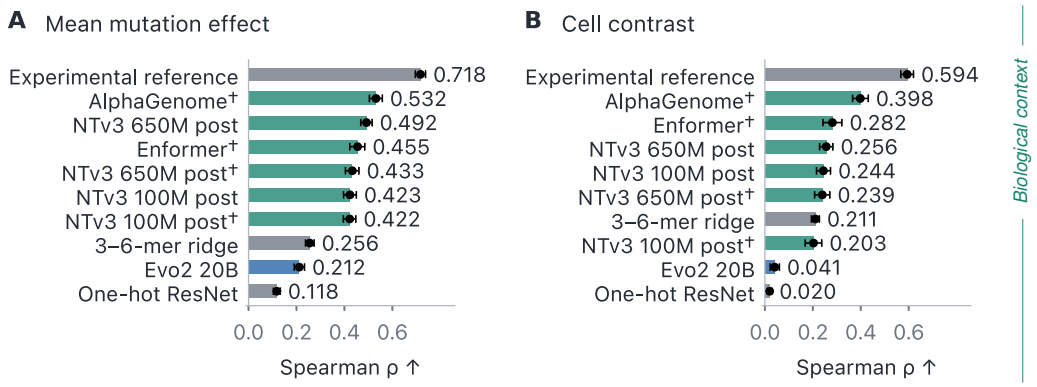

Figure 3: **Predicting cell-specific effects.** A/B: mean mutation effect and HepG2–K562 difference across 128 Siraj regulatory elements. Experimental references compare groups of repeat measurements (127 elements); they are not ceilings. (†) mark native outputs; unmarked NTv3 entries use fitted readouts. Whiskers show available 95% element-bootstrap intervals.

Expression profiles distinguish tissues. We also evaluate how well models distinguish expression profiles across tissues, using 5,782 held-out genes measured in 51 GTEx tissues (GTEx Consortium, 2020). For each gene and query tissue, we compare expression with the tissue showing the largest measured difference. Using native outputs, AlphaGenome reaches correlation 0.775 on these within-gene contrasts, compared with 0.594 for NTv3 650M post (Figure 15A). Because the comparison tissue is chosen using measured expression, this tests prediction of observed contrasts rather than prospective tissue selection.

A second test asks whether a predicted expression profile across genes identifies the correct tissue among the 51 candidates. AlphaGenome achieves 68.6% accuracy, compared with 45.1% for NTv3 650M post and 7.8% for k-mer ridge (Figure 15B). Chance accuracy is 2.0%, while expression profiles from independent donor halves agree at 99.0%, averaging both directions. The models therefore recover substantial tissue-specific information, but a sizeable gap to the experimental reference remains. These evaluations hold out genes within represented tissues rather than testing transfer to unseen tissues (Appendix D.1).

6 Genetic interactions: predicting departures from additivity

A mutation's effect depends on its genetic background. The context assessed above is on the cell level, but a different genetic background can also alter mutation effects. The Siraj assay measures the same edit on paired allelic backgrounds, allowing us to isolate this interaction. Across 68,026 paired mutations in 114 regulatory elements, AlphaGenome reaches interaction correlation 0.260, compared with 0.187 for NTv3 650M post's fitted readout and 0.181 for Enformer (Figure 4A). Predictions are most accurate for mutations within 10 bp of one another and become less accurate with increasing separation (Figure 4C).

Combined-effect accuracy can conceal weak interaction prediction. COMPASS tests non-additive interactions splicing effects by measuring single and double mutations relative to the same reference. On its 110-gene panel, NTv3 650M post reaches correlation 0.409 for single-mutation effects and 0.500 for double-mutation effects, but only 0.046 for the interaction remaining after subtracting the two single effects (Figure 4B). Native splice predictors show the same pattern, and replicate correlation for the interaction measurements is 0.171.

To assess how much interaction prediction contributes to overall accuracy, we remove the predicted interaction and retain only the additive effects. NTv3's double-mutation correlation changes only from 0.500 to 0.493. Similarly, removing AlphaGenome's predicted background dependence on Siraj changes overall correlation from 0.528 to 0.521, a small but statistically supported decrease of 0.0070 (95% interval: 0.0014–0.0120; Appendix E.2). This indicates that a model can predict combined effects accurately while recovering little of the interaction between mutations.

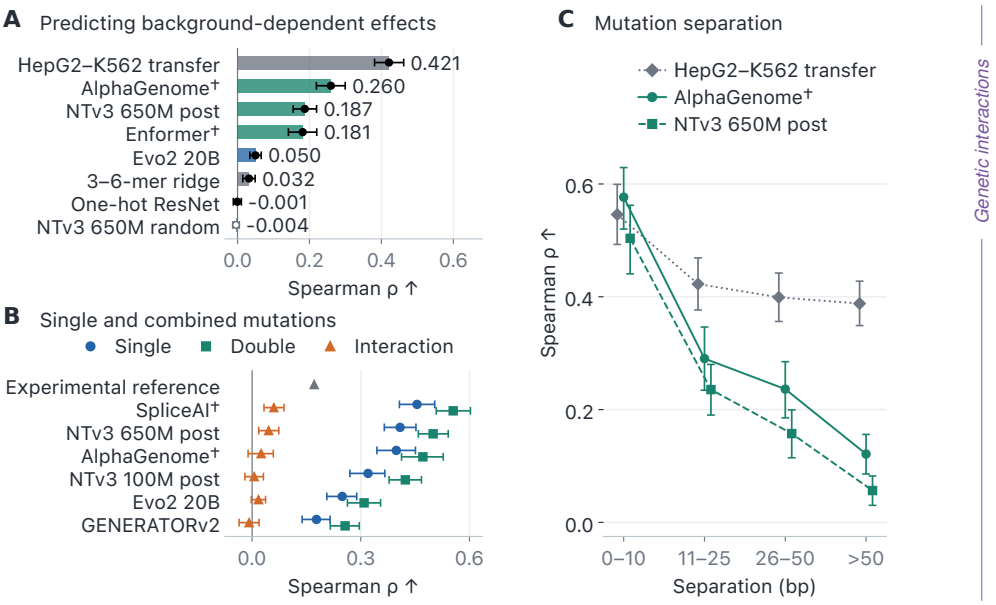

Figure 4: **Accurate combined effects can conceal weak interaction prediction.** A: difference in mutation effects between two backgrounds differing at one base. B: single, double and non-additive mutation effects in 110 COMPASS genes. C: Interaction prediction by distance between mutations. HepG2–K562 transfer compares measured interactions across cells (A/C). The experimental reference compares repeat measurements of interactions (B). Neither is a ceiling. (†) mark native outputs, and whiskers show 95% bootstrap intervals over regulatory elements (A/C) or genes (B).

7 GENOMIC DESIGN: SELECTING SEQUENCES FOR A DESIRED RESPONSE

We finally ask whether model predictions help choose DNA with a desired response. These tasks evaluate the measured outcomes of model-selected candidates, progressing from matching a single response to controlling expression across multiple cell types. All choices are made within experimentally assayed libraries, with candidate measurements withheld from selection.

Matching a target response. Can a model find an alternative mutation that reproduces the effect of a reference edit? In the Siraj assay, each element contains a natural variant and a measured change in activity, which we define as the target response. Excluding that position leaves 578–597 candidate substitutions per element. A model predicts an effect for every candidate and the reference edit, then ranks candidates by how closely their predicted effects match its own reference prediction. AlphaGenome scores effects as changes in its cell-matched accessibility outputs; NTv3 uses linear heads fitted on held-out elements. All candidates' measured effects are withheld until scoring.

Across 331 targets in 110 elements, AlphaGenome's first choice has a mean measured residual percentile of 24.3, compared with 30.0 for NTv3 650M post and 50 for uniform selection (Figure 5A; lower is better). Among the top ten choices, at least one mutation matches the reference response within assay uncertainty for 62% of targets with AlphaGenome and 59% with NTv3, compared with 38% under uniform selection (B). These results show that model rankings can reduce the number of candidates needed to find a matching response.

Increasing expression selectively. We also test a more challenging task, choosing edits which increase expression in one cell type while minimally changing expression in another. We evaluate 167,958 substitutions across 280 parental backgrounds at 127 Siraj loci, with HepG2 and K562 alternately serving as the target cell. Each model chooses one substitution or no edit by maximizing predicted utility $U_{\text{cell}} = t - \lambda|o|$, where t and o are expression changes relative to the unmutated parent in the target and other cell. Six fixed penalties describe the tradeoff between the two responses (Figure 5C). Complete loci were held out of readout fitting and native-output calibration.

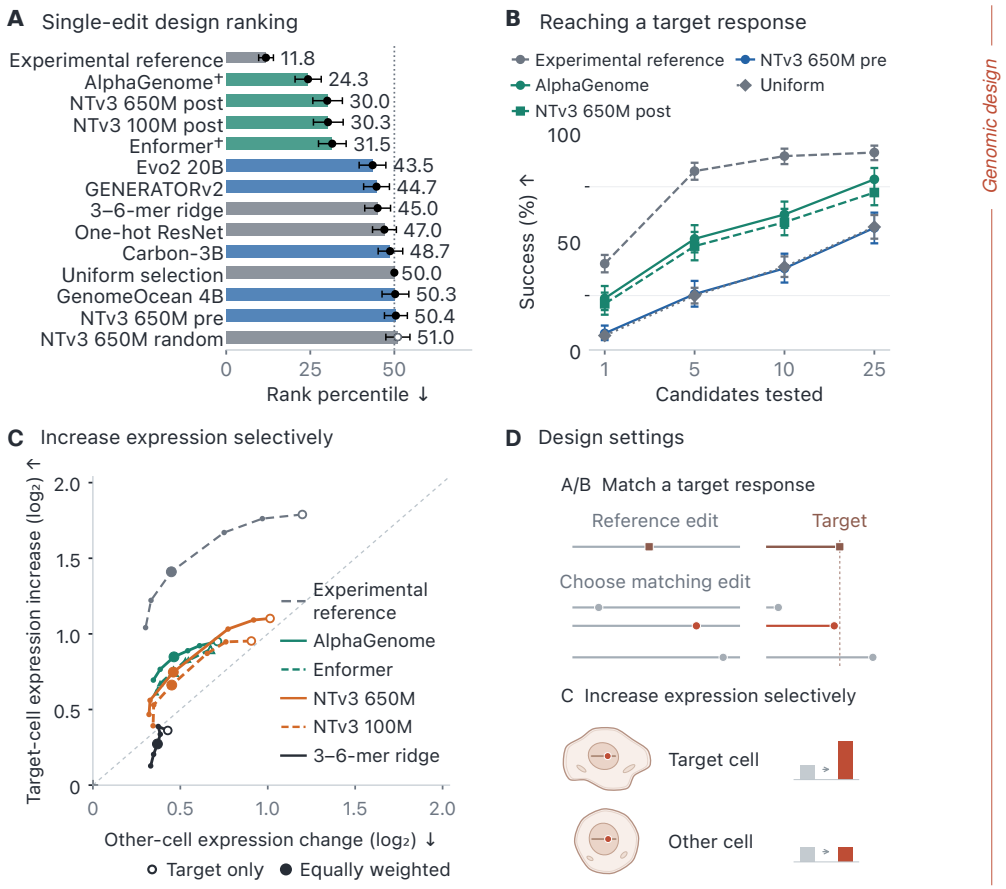

Figure 5: **Selecting edits for a desired response and cell specificity.** A/B: rank of the first selected edit and success at matching a reference-edit response as testing budget increases (331 targets, 110 regulatory elements). C: target-cell expression increase versus absolute other-cell change across 127 loci, varying the penalty on unwanted effects. Open markers ignore the other cell ($\lambda = 0$), and filled markers penalize it equally ($\lambda = 1$). Experimental references in A–C select edits on one group of repeat measurements and evaluate on the other. Models may outperform these references. D: task settings. (†) mark native outputs in A, and whiskers show 95% element-bootstrap intervals.

At equal weighting ($\lambda = 1$), calibrated native AlphaGenome achieves measured utility 0.384, compared with 0.287 for NTv3 650M post, 0.211 for NTv3 100M post and zero for no edit. All three also improve over controls matched for edit probability and predicted target strength; AlphaGenome's gain is 0.086 (simultaneous 95% interval: 0.030–0.142). This indicates that considering the other cell is beneficial, though equal measured target responses are not established. An embedding-fitted linear head for AlphaGenome reaches utility 0.515 under the same selection rule (Table 1).

To test whether the library contains better selectable edits, we select using one group of repeat measurements and evaluate on the other, then reverse the roles. On these same evaluation subsets, experimental selection exceeds native AlphaGenome's utility by 0.603, with an advantage at 108 of 127 loci. This reference demonstrates reproducible opportunities for better selection within the library; it is not an upper limit on model performance (Appendix F.3). A related test of selectivity between allelic backgrounds is reported in Appendix F.2.

Selecting multi-cell expression programs. The final task is to select for activity in several cell types while avoiding activity elsewhere. From the twelve-cell Atlas assay (Castillo-Hair et al., 2025), we use sixteen requests: eight with two cells ON and ten OFF, and eight with three ON and nine OFF. For each request, a model selects ten enhancers from the same catalogue of 2,411 measured 145-nt sequences. Selection maximizes the predicted counterpart of

$$U_{\text{multi}}(x) = \min_{c \in \text{ON}} y(x, c) - \max\left(0, \max_{c \in \text{OFF}} y(x, c)\right), \tag{6}$$

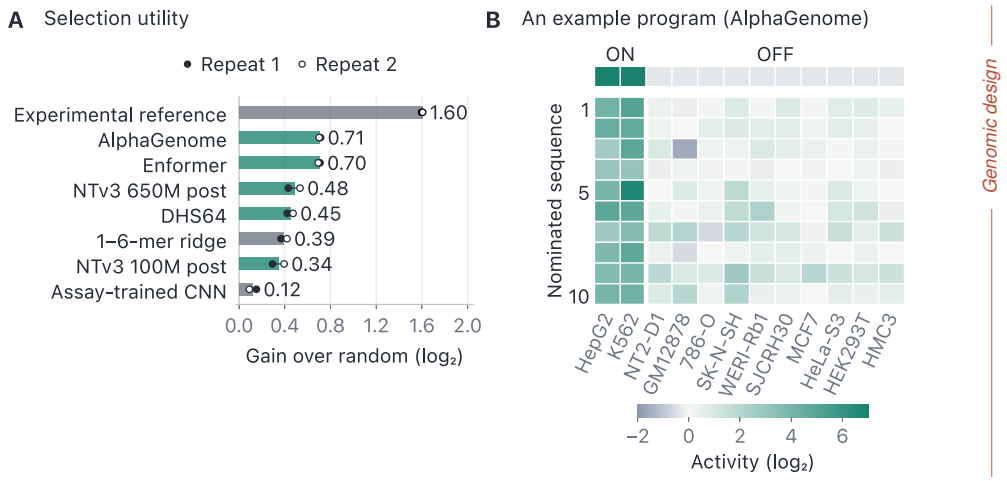

Figure 6: **Selecting DNA with activity in desired cells and little activity elsewhere.** Models choose ten enhancers per request, targeting activity in two or three cell types. A: gain over intent-matched random selection across 16 requests. Each sequence is scored by its weakest requested activity, minus its strongest activity elsewhere. Dots show the two repeat measurements. The experimental reference selects sequences using one and evaluates them using the other, then swaps their roles, which is not a performance ceiling. B: measured activities for a comparatively successful HepG2/K562 request using AlphaGenome.

where y is negative-control-normalized $\log_2$ reporter activity. This rewards activity in every requested cell while penalizing the strongest unwanted response. Readouts are fitted on 5,246 separate source constructs, excluding sequence families shared with the candidate catalogue. All twelve cell types are represented in the source data.

Linear heads fitted to AlphaGenome and Enformer native outputs select sequences scoring 0.707 and 0.700 over random selection from sequences originally designed for the same task. The corresponding gains are 0.482 for NTv3 650M post and 0.344 for NTv3 100M post (Figure 6A). For all four pipelines, considering OFF-cell activity improves selection over ranking by ON-cell activity alone in every program and both repeat measurements (Appendix F.4).

Despite these gains, mean absolute utility remains negative: -0.222 for AlphaGenome and -0.228 for Enformer. An embedding-fitted linear head for AlphaGenome improves the gain over random to 0.833, but its absolute utility also remains negative (-0.096; Table 1). By comparison, selecting on one repeat measurement and evaluating on the other yields mean utility $+0.673$. Thus, the library contains sequences that better satisfy the joint expression objective than current model selections.

8 Conclusions

Genomic models recover functional information that transfers across assays and genes, and helps select DNA with desired responses. Functional training largely expands these capabilities: predictions from embedding-fitted linear heads can match or exceed native outputs, suggesting that its benefits extend beyond the trained prediction heads. Yet accurate overall predictions can conceal weak prediction for cell types and genetic interactions, which greatly impact what sequences are worth investigating. DRAFT separates these capabilities and tests each against measured outcomes.

Tasks with least reproducible experiments are often the more difficult tasks for the models, meaning performance should be interpreted in light of an assay's reliability. The chosen experimental references are somewhat naive (for example, transferring mutation effects across cell types), yet they still beat current models. The linear heads may also depend on the embedding sizes of the models, and models with different input context lengths may have an effect on task performance. Despite these limits, DRAFT shows where genomic models may help narrow hundreds of candidate mutations to a testable few, and where further progress is needed.

AI USE STATEMENT

Generative AI tools were used in evaluating models on HPC clusters, preparing code, organizing the tasks, and preparing the manuscript and supporting materials. The authors reviewed AI-assisted work and take full responsibility for the final manuscript, including its methods, results and conclusions.

ETHICS STATEMENT

This work evaluates machine learning models using existing functional-genomics experiments, reference sequences and population-genomic resources. As human genomic data can be identifying, it was and should always be used in accordance with their applicable access, consent and redistribution conditions. Performance on the genomic contexts represented here should not be assumed to generalize uniformly to other populations or biological settings. Models supporting genomic design may also have dual-use applications. Downstream experimental use therefore requires appropriate biological validation, biosafety review and oversight.

REPRODUCIBILITY STATEMENT

We provide the evaluation code, fixed biological splits, dataset and model provenance, and machine-readable results underlying the reported figures and tables. The appendices describe task-specific preprocessing, model interfaces, statistical procedures, and evaluation protocols in detail. Dataset checksums and model revisions are recorded in the accompanying code and experiment metadata, where applicable.

REFERENCES

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R. Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 649(8099):1206–1218, 2026. doi: 10.1038/s41586-025-10014-0.

Gonzalo Benegas, Carlos Albors, Alan J. Aw, Chengzhong Ye, and Yun S. Song. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 43:1960–1965, 2025. doi: 10.1038/s41587-024-02511-w.

Sam Boshar, Benjamin Evans, Ziqi Tang, Armand Picard, Yanis Adel, Franziska K. Lorbeer, Chandana Rajesh, Tristan Karch, Shawn Sidbon, David Emms, Javier Mendoza-Revilla, Fatimah Al-Ani, Evan Seitz, Yair Schiff, Yohan Bornachot, Ariana Hernandez, Marie Lopez, Alexandre Laterre, Karim Beguir, Peter Koo, Volodymyr Kuleshov, Alexander Stark, Bernardo P. de Almeida, and Thomas Pierrot. A foundational model for joint sequence–function multi-species modeling at scale for long-range genomic prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.22.695963.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, et al. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, 2026. doi: 10.1038/s41586-026-10176-5.

Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, Wayne E. Clarke, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, 2022. doi: 10.1016/j.cell.2022.08.004.

Sebastian M. Castillo-Hair, Christopher H. Yin, Leah VandenBosch, Timothy J. Cherry, Wouter Meuleman, and Georg Seelig. Programming human cell type-specific gene expression via an atlas of AI-designed enhancers. *bioRxiv*, 2025. doi: 10.1101/2025.09.30.679565.

Ilias Georgakopoulos-Soares, Chengyu Deng, Vikram Agarwal, Candace S. Y. Chan, Jingjing Zhao, Fumitaka Inoue, and Nadav Ahituv. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nature Communications*, 14(1):2333, 2023. doi: 10.1038/s41467-023-37960-5.

Sager J. Gosai, Rodrigo I. Castro, Natalia Fuentes, John C. Butts, Kousuke Mouri, Michael Alasoadura, Susan Kales, Thanh Thanh L. Nguyen, Ramil R. Noche, Arya S. Rao, et al. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, 634(8036):1211–1220, 2024. doi: 10.1038/s41586-024-08070-z.

GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

Samuel H. King, Claudia L. Driscoll, David B. Li, Daniel Guo, Aditi T. Merchant, Garyk Brixi, Max E. Wilkinson, and Brian L. Hie. Generative design of bacteriophages with genome language models. *Science*, 393(6811):eaec2657, 2026. doi: 10.1126/science.aec2657.

Samantha E. Koplik, Angela M. Yu, Madelyn R. Shelby, Gabriel C. Fonseca, Charles M. Roco, Yue Zhang, Nicholas Bogard, Alex K. Sabo, Alexander B. Rosenberg, Johannes Linder, and Georg Seelig. Massively parallel assay of human splice variants reveals cis-regulatory drivers of disease-associated and cell type-specific splicing regulation. *bioRxiv*, 2025. doi: 10.1101/2025.10.12.681955.

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*, 2026. doi: 10.64898/2026.01.27.702015.

Zicheng Liu, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z. Li. GenBench: A benchmarking suite for systematic evaluation of genomic foundation models. arXiv preprint arXiv:2406.01627, 2024. URL https://arxiv.org/abs/2406.01627.

Mingqian Ma, Yucheng Wu, Xin Chen, Feifei Jiang, Peijun Lin, Dongxin Ye, Yidi Sun, Yijing Zhang, Tianqiong Shi, Yu Zhao, Wanli Ouyang, Bowen Zhou, Lei Bai, and Yuchen Ren. CENO: A genome-scale world model for evolutionary sequence interpretation and programmable regulatory design. *bioRxiv*, 2026. doi: 10.64898/2026.07.28.741284.

Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uKB4cFNQFg.

Nicolas Mathis, Ahmed Allam, András Tálas, Lucas Kissling, Elena Benvenuto, Lukas Schmidheini, Ruben Schep, Tanav Damodharan, Zsolt Balázs, Sharan Janjuha, Eleonora I. Ioannidi, Desirée Böck, Bas van Steensel, Michael Krauthammer, and Gerald Schwank. Machine learning prediction of prime editing efficiency across diverse chromatin contexts. *Nature Biotechnology*, 43:712–719, 2025. doi: 10.1038/s41587-024-02268-2. Published online 21 June 2024.

Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010. doi: 10.1101/gr.097857.109.

Gioia Quarantani, Joseph Clarke, Mike Thompson, Fei Sang, Juan Valcárcel, and Ben Lehner. OpenSplice: The impact of half a million mutations on the alternative splicing of 600 human exons. *bioRxiv*, 2026. doi: 10.64898/2026.05.22.727141. URL https://www.biorxiv.org/content/10.64898/2026.05.22.727141v1. Version 1, posted 23 May 2026.

Joel Shor, Erik Strand, and Cory Y. McLean. NucleoBench: A large-scale benchmark of neural nucleic acid design algorithms. *bioRxiv*, 2025. doi: 10.1101/2025.06.20.660785. URL https://www.biorxiv.org/content/10.1101/2025.06.20.660785v3. Version 3, 26 August 2025.

Layla Siraj, Rodrigo I. Castro, Hannah B. Dewey, Susan Kales, John C. Butts, Thanh Thanh L. Nguyen, Masahiro Kanai, et al. Functional dissection of complex trait variants at single-nucleotide resolution. *Nature*, 2026. doi: 10.1038/s41586-026-10121-6.

Dylan J. Taylor, Surya B. Chhetri, Michael G. Tassia, Arjun Biddanda, Stephanie M. Yan, Genevieve L. Wojcik, Alexis Battle, and Rajiv C. McCoy. Sources of gene expression variation in a globally diverse human cohort. *Nature*, 632(8023):122–130, 2024. doi: 10.1038/s41586-024-07708-2.

Heng Yang, Jack Cole, Yuan Li, Renzhi Chen, Geyong Min, and Ke Li. OmniGenBench: A modular platform for reproducible genomic foundation models benchmarking. arXiv preprint arXiv:2505.14402, 2025. URL https://arxiv.org/abs/2505.14402.

Chengzhong Ye, Gonzalo Benegas, Carlos Albors, Jianan Canal Li, Sebastian Prillo, Peter D Fields, Brian Clarke, and Yun S Song. Predicting genome-wide functional constraints with GPN-Star. *Nature*, 2026. doi: 10.1038/s41586-026-11005-5.

Christopher Yin, Sebastian Castillo-Hair, Gun Woo Byeon, Peter Bromley, Wouter Meuleman, and Georg Seelig. Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity. *Cell Systems*, 16(7):101302, 2025. doi: 10.1016/j.cels.2025.101302.

Appendix Contents

A Evaluation details . . . 14

A.1 Data split overview . . . 14

A.2 Predictions and readouts . . . 15

B Sequence priors: Sequence probability and reconstruction . . . 16

B.1 Sequence and evolutionary priors on the same bases . . . 18

B.2 How post-training changes access to the sequence prior . . . 19

B.3 Probability calibration . . . 21

B.4 Sequence prediction on natural haplotypes . . . 21

B.5 From sequence score to functional response . . . 21

C Functional prediction: Variant-effect prediction and label efficiency . . . 22

C.1 Editing efficiency from genomic context . . . 23

D Biological context: Cell, tissue and personal-sequence effects . . . 24

D.1 Tissue-expression contrasts and identification . . . 25

D.2 Personal DNA at familiar and unseen loci . . . 25

E Genetic interactions: Interaction and unseen regulatory arrangements . . . 26

E.1 COMPASS predictions . . . 26

E.2 Removing predicted interaction while preserving main effects . . . 26

E.3 Unseen motif arrangements . . . 27

F Genomic design: Target matching and selective expression . . . 28

F.1 Matching a target response . . . 28

F.2 Selective mutation choice with target-strength matching . . . 28

F.3 Cell-selective expression: cohort and measurement . . . 30

F.4 Selecting enhancers for multi-cell expression programs . . . 31

G Statistical comparisons and reference results . . . 33

G.1 Inference and comparison families . . . 33

G.2 Selected prediction results . . . 33

G.3 Selected design results . . . 34

A EVALUATION DETAILS

A.1 DATA SPLIT OVERVIEW

Table 2 summarizes the data partitions, their use across experiments, and the corresponding appendix methods.

Table 2: **Data partitions and experiment locations.** Counts denote sequence windows unless otherwise stated. Background colors identify the experiment categories in Table 1; shading has no numerical meaning. Appendix references distinguish experiment uses from partition-construction methods. Study-specific holdouts do not establish exclusion from foundation-model pretraining.

Partition / panel	Fitting, selection and evaluation	Main-paper uses	Appendix uses	Ref.
Prior benchmark	1,024 fitting; 512 evaluation. Markov tuning: 768/256 within fitting, then refit on all 1,024.	Prediction and calibration, Fig. 1	Calibration (B.3); mask placement and state exchange (B.2). Donor analysis forms 256 pairs from the same 512 windows.	p. 16.
Activation shift	768 estimation; 128 selection; 1,536 confirmation windows from 48 blocks.	Activation shift, Section 3	Shift, calibration, direction controls and ensemble comparison (B.2)	p. 20.
Natural haplotypes	Evaluation only: 512 SNP sites, two carrier sequences per site. Calibration uses the prior benchmark fitting set.	Haplotype prediction, Section 3	Carried alleles and background sensitivity (B.4); calibration evaluation (B.3)	p. 21.
OpenSplice folds	Five gene-held-out folds. Label-efficiency analyses use nested training budgets.	Mutation effects, Fig. 2A	Model comparisons, label efficiency and error checks (C)	p. 22.
Strict assay transfer	OpenSplice source fitting; COMPASS evaluation on 88 genes. Shared genes and overlapping sequences excluded.	Cross-assay transfer, Fig. 2B	Transfer evaluation (C)	p. 22.
Editing-context folds	Five outer chromosome-group folds. Tuning sensitivity uses three inner chromosome folds.	Editing efficiency, Section 4	Editing-efficiency prediction and tuning sensitivity (C.1)	p. 23.
Siraj readout folds	Element-grouped fitting and evaluation; related allelic backgrounds kept together.	Cell contrasts, Fig. 3	Context effects (D)	p. 24, p. 26.
		Background interactions, Fig. 4	Interaction analysis and ablation (E.2)	
GTEx genes	12,146 training; 1,350 validation; 5,782 test genes. Both analyses use the same test genes.	Tissue contrasts and identification, Section 5	Both tissue analyses (D.1)	p. 25.
Personal-sequence folds	Two settings: person-held-out; person- and local-locus-held-out.	Appendix only	Prediction at familiar and unseen loci (D.2)	p. 25.
COMPASS interactions	OpenSplice-trained readouts; COMPASS interaction evaluation cohort. Ablation reuses fixed predictions.	Interaction prediction, Fig. 4	Prediction (E.1); ablation (E.2)	p. 26.
Siraj adaptation	Locus-grouped training, validation and test; 68 training and 23 test loci.	Appendix only	Conditional-response training (E.2)	p. 27.
Motif constructs	Train on singles/pairs; evaluate triplets from held-out motif multisets.	Motif arrangements, Section 6	Triplet prediction (E.3)	p. 27.
Siraj response matching	Response-matching evaluation cohort; fitted readouts hold out complete elements.	Target matching, Fig. 5	Response-matching evaluation (F.1)	p. 28.

Table 2 (continued)

Partition / panel	Fitting, selection and evaluation	Main-paper uses	Appendix uses	Ref.
Siraj allele selectivity	Strength-control calibration excludes evaluation loci. Repeat subsets provide measurement checks.	Allele-selective edits, Section 7	Selective mutation choice and controls (F.2)	Cohort: p. 28; calibration: p. 29.
Siraj cell selectivity	Five locus folds for native calibration; fitted readouts supply held-out predictions.	Cell-selective edits, Fig. 5C	Cell-selective evaluation and controls (F.3)	p. 30.
Atlas families	4,223 training / 1,023 validation sequences; refit on all 5,246. Candidate families excluded from fitting.	Multi-cell selection, Section 7	Expression-program selection (F.4)	p. 31.

A.2 Predictions and readouts

Native outputs and fitted linear heads. We evaluate models using native outputs and embedding-fitted linear heads. Native outputs are produced by the models' existing output heads. Depending on the model, these include self-supervised sequence logits, supervised predictions of functional measurements or genomic annotations. Embedding-fitted linear heads are trained on sequence embeddings while the original model is kept frozen. For multi-cell design, we also fit linear heads to native outputs; these are identified separately. Any task-specific calibration of native outputs is described with the task.

The use of native outputs and fitted linear heads varies depending on the task. Below are examples given for how we obtain variant-effect scores with the different approaches.

To use native outputs, where sequence output logits are available, we score a variant in causal models as its change in log sequence probability. For masked models, the score is calculated as the alternate-versus-reference log odds:

$$\begin{aligned} s_\theta^{\text{causal}}(\delta \mid x) &= \log p_\theta(x+\delta) - \log p_\theta(x), \\ s_\theta^{\text{masked}}(\delta \mid x) &= \log q_\theta(a_{\text{alt}} \mid x_{\setminus i}) - \log q_\theta(a_{\text{ref}} \mid x_{\setminus i}). \end{aligned} \tag{7}$$

To use native functional outputs for models with supervised output heads (AlphaGenome, Enformer, post-trained NTv3), we measure the difference in predicted assay quantity c under the sequence variant. Where c, e.g. DNase accessibility or another functional output, the predicted effect of a variant is the difference between the outputs on the variant and reference sequences, $f_\theta(x+\delta, c) - f_\theta(x, c)$.

To measure variant effect using an embedding-fitted linear head, we determine the difference between frozen embeddings. With $\phi_\theta(x)$ representing the frozen pooled representation of a sequence and Δ_i as the measured effect of variant δ_i, we fit ridge regression on the representation change:

$$w^* = \arg\min_w \sum_{i \in \text{train}} \left(\Delta_i - w^\top \Delta\phi_\theta(\delta_i \mid x_i)\right)^2 + \lambda \|w\|_2^2, \qquad \Delta\phi_\theta(\delta_i \mid x_i) = \phi_\theta(x_i+\delta_i) - \phi_\theta(x_i). \tag{8}$$

Feature standardization is estimated on the training data only. Projections used to reduce feature dimensionality are random, seeded and label-independent, and the ridge penalty λ is either fixed in advance or chosen on validation data. Per-task settings are provided in their corresponding sections.

Embedding-fitted linear heads evaluated on COMPASS are fit only on OpenSplice.

NTv3 outputs. Because NTv3 produces multiple native outputs, we define here the native outputs used for each figure. Post-trained NTv3 models produce nucleotide logits, functional tracks and genome annotations (Boshar et al., 2025). For example, native-output results use nucleotide logits for the sequence-prior task, predicted splice-boundary probabilities for splicing (Figure 2), and cell-matched DNase outputs for cell contrasts (Figure 3). These evaluations use the existing output heads. NTv3 also uses embedding-fitted linear heads and, for multi-cell design, linear heads fitted to native outputs, as specified for each task.

The pre-trained checkpoints(NTV3 100M/650M pre) produce nucleotide logits only, and have no functional heads.

AlphaGenome fitted readouts. AlphaGenome's supported input is larger than the benchmark's sequences, so padding is used. We take frozen 1-bp representations from the final model layer, mean-pool over the inserted bases, project to 512 features, and fit ridge with $\alpha = 10$. Siraj and OpenSplice/COMPASS inserts use 16,384-bp N-padded inputs without genomic or vector flanks. We hold out complete elements for Siraj, and the transfer to COMPASS excludes sequence-overlapping target events and removes the 23 shared genes from both datasets. GTEx uses 131-kb inputs, training-only scaling, and validation selects $\alpha = 100$. The other design readouts follow Appendices F.3 and F.4. Native and fitted results remain separate.

Experimental references. We use experimental references to investigate how well the measurements themselves predict or select outcomes. The references should not be considered performance ceilings, as it is possible for a model to outperform noisy measurements. In the reporter assays, activity is measured from RNA counts normalized by DNA counts.

The split is dependent on the assay:

- For OpenSplice, we compare the average of two variant PSI repeats with the third, averaging all three rotations. Effects share the reported wild-type estimate.
- For Siraj, background interactions use HepG2–K562 transfer. Other references compare measurements averaged over repeats 1–2 with those averaged over repeats 3–5. Effects are reconstructed separately within each subset, while model scores use the published pooled estimates (minus cell-selective design, which uses physical-parent contrasts throughout).
- For GTEx, tissue identification averages both donor-half directions (99.0%), and its uncertainty was not re-estimated.
- For Atlas, each of the two RNA measurements is evaluated against the other.

Several cohort details apply to the Siraj-derived references. Cell-effect references cover 127 of the 128 elements, as one lacks count coverage. Ranking and target-response references retain all 331 targets in 110 elements and average reciprocal selections. Success curves retain the published error-based tolerances and request eligibility. Selection procedures and candidate sets otherwise match the corresponding tasks.

Uncertainty and multiplicity. Bootstrapping resamples over the units the corresponding task splits on (ex, genes, regulatory elements, or loci), rather than individual measurements, so that measurements sharing a unit stay correlated. When a pretrained model is compared with its architecture-matched random pair, both are evaluated on the same resample. Simultaneous confidence intervals use the family named by the figure or table, and experiments with several training seeds jointly resample seeds and biological units. The primary rank metric gives every biological group equal weight,

$$\rho_{\mathcal{G}}(\hat{y}, y) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathrm{Spearman}(\hat{y}_g, y_g). \tag{9}$$

B SEQUENCE PRIORS: SEQUENCE PROBABILITY AND RECONSTRUCTION

Sampling and controls. We sample 512 autosomal GRCh38 windows from separate 1-Mb blocks, excluding the 2,048 blocks used for earlier evaluation or baseline fitting. We call the 1,024 windows of that earlier panel the fitting windows; they are disjoint from the 512 benchmark windows and are never evaluated here. Each window contains 1,200 A/C/G/T bases; positions and strands are sampled without assay annotations or model scores. These are new benchmark blocks, but GRCh38 exposure during pretraining is expected.

Uniform DNA costs two bits per base. Composition and Markov controls fit the original 1,024 fitting windows without crossing window boundaries. The tuned Markov model uses order ten and smoothing strength 64, chosen on a 768/256 split and then refitted on all 1,024 fitting windows.

Confidence intervals use 4,000 paired resamples of complete genomic blocks. Repeated queries do not count as independent observations. Full settings, seeds and checkpoint identifiers are preserved with the released code.

Sequence probability. For each causal model we report

$$H_\theta = -\frac{1}{600} \log_2 p_\theta(x_{601:1200} \mid x_{1:600}). \tag{10}$$

Evo2 20B and CENO use base tokens (Brixi et al., 2026; Ma et al., 2026). Their probabilities retain the full vocabulary, including non-DNA mass. GENERATORv2 uses its released factorized nucleotide interface (Li et al., 2026): each sixmer is represented by six base marginals,

$$p_\theta(x_b \mid x_{<b}) = \prod_{j=1}^{6} \sum_{u:u_j=x_{b,j}} \mathrm{softmax}(z_\theta(x_{<b}))_u, \tag{11}$$

The softmax covers all 4,096 DNA sixmers. This gives a normalized block distribution, distinct from the raw joint sixmer score. The 600-bp boundary falls between blocks. We checked this implementation against the released scorer and report loss per base, not per token.

Repeat-stratified sequence prediction. GRCh38 soft-masked annotations divide the 512 windows into 26 with no annotated repeats, 228 with some but less than half their bases annotated, and 258 with at least half annotated. The coding gain over tuned Markov is largest in repeat-rich windows, but remains positive in every group for all three causal models (Figure 7).

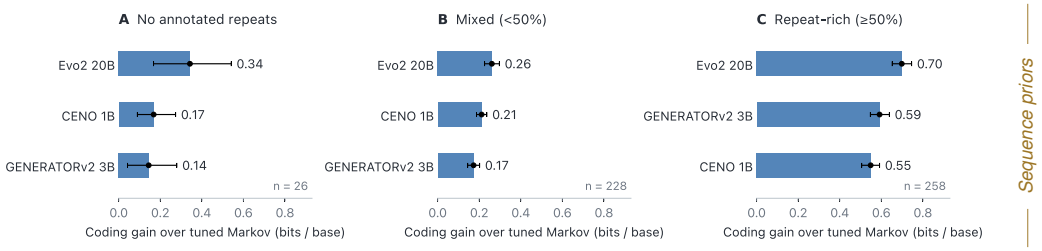

Figure 7: **Sequence prediction beyond local composition.** Reduction in continuation coding cost relative to the frozen tuned Markov control, on all three repeat strata. Larger is better. Black intervals are paired 95% block-bootstrap intervals; counts give windows.

Base reconstruction. Position 601 is hidden in every window. For a causal model, replace it by each candidate a and score the entire resulting sequence $x^{(a)}$, including all predictions after the changed position:

$$q_\theta(a \mid x_{\setminus i}) = \frac{p_\theta(x^{(a)})}{\sum_{b\in\{\mathrm{A,C,G,T}\}} p_\theta(x^{(b)})}. \tag{12}$$

NTv3 supplies normalized A/C/G/T probabilities at a masked position (Boshar et al., 2025). We evaluate its original single-mask query and a fixed probability readout that averages eight queries:

$$\bar{q}_\theta(a \mid x_{\setminus i}) = \frac{1}{8} \sum_{k=1}^{8} q_\theta(a \mid \mathrm{mask}_{M_k \cup \{i\}}(x)). \tag{13}$$

Each M_k masks 12% of the other positions. Seeds depend on window coordinates and draw index, never the hidden base. Both NTv3 sizes and training stages use the same rule, averaging probabilities rather than logits. Twelve percent is the expected MASK fraction under NTv3's training corruption, not a density chosen on these test windows. Post-trained checkpoints retain human-species conditioning.

We report base log loss $-\log_2 q(x_i \mid x_{\setminus i})$ and accuracy, sharing credit for exact ties. These four-base probabilities do not define a joint sequence likelihood. Enformer, AlphaGenome and the released GPN-Star scores enter the functional comparisons rather than this nucleotide-probability test.

Masking changes the recovered prior. Longer inputs and a swapped pre-training head did not recover post-trained sequence prediction in the preliminary audit; additional masking did. On the final 512 blocks, the fixed readout reduces NTv3 650M post loss by 0.853 bits (95% paired interval 0.723–0.990). Its loss relative to the pre-trained model changes from $+0.796$ bits with one mask to -0.073 bits with the same denoising readout. Both model sizes show this query sensitivity.

B.1 Sequence and evolutionary priors on the same bases

Matched sites. We compare all models at 22,779 positions in the reference backgrounds of 114 non-haplotype Siraj regulatory sequences, spanning 109 genomic blocks. Each site has all three substitutions and agrees with the uniquely lifted hg38 reference. Allele log odds and a zero reference score define a four-base distribution. Causal models score the full candidate sequence; NTv3 uses native or fixed denoising queries. Strand averaging uses the normalized geometric mean.

GPN-MSA conditions on a multispecies alignment (Benegas et al., 2025); we reconstruct its four-base distribution from the released log odds. The hg19 100-way phyloP track measures conservation or acceleration (Pollard et al., 2010). Both have alignment information absent from sequence-only models, so their agreement is not independent validation. GPN-Star's calibrated scores enter rank comparisons only, since they are not normalized probabilities.

Conservation and assay response. We correlate phyloP with observed-base gain over tuned Markov, $\log_2 q_\theta(x_i \mid r) - \log_2 q_{\text{Markov}}(x_i \mid r)$, within each regulatory sequence. We also correlate allele scores with GPN-MSA and measured mutation effects within each sequence and cell, then average over cells and sequences. These comparisons use the same 136,674 assay records. Intervals use 4,000 paired genomic-block resamples and give each regulatory sequence equal weight.

Evo2's gain over Markov has a small positive association with phyloP ($\rho = 0.062$, interval 0.038–0.086), which is not shared by every model (Figure 8B). Its allele odds agree more with GPN-MSA ($\rho = 0.296$) than with measured expression response ($\rho = 0.047$). Agreement with evolutionary preference therefore need not imply prediction of an assay's signed response.

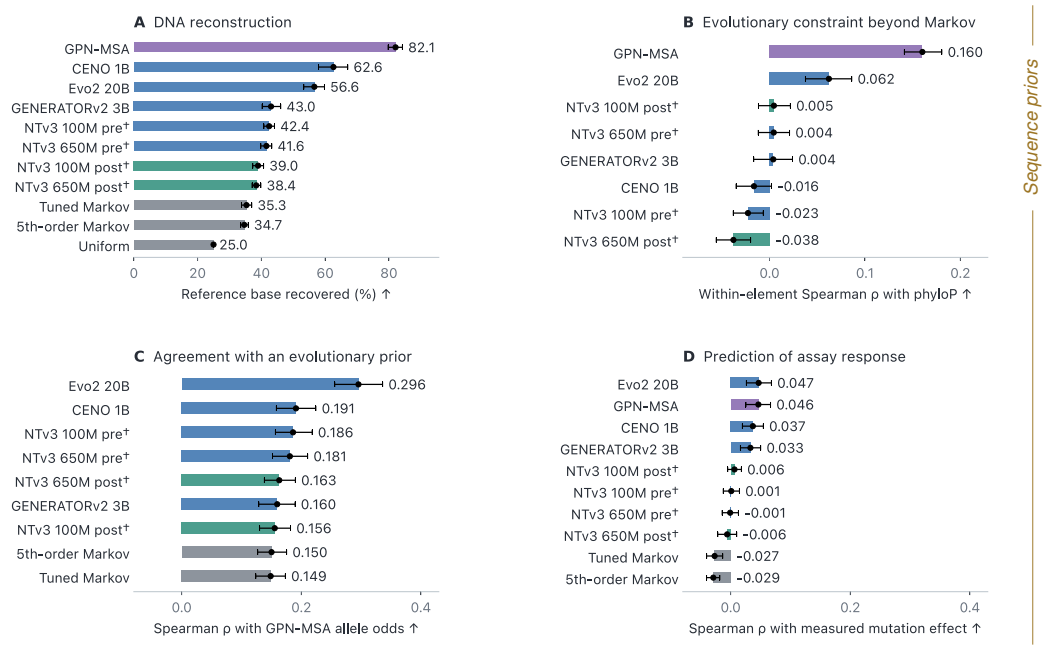

Figure 8: **Sequence probability, evolutionary preference and assay response.** All panels use 22,779 matched reference positions in 114 Siraj elements. *A:* base reconstruction. *B:* within-element association of phyloP with observed-base log-probability gain over tuned Markov. *C–D:* the same allele scores compared with GPN-MSA and measured response, on matching scales. NTv3 (†) uses eight-query denoising; purple denotes the alignment-conditioned GPN-MSA comparator. Intervals bootstrap genomic blocks (95%).

B.2 How post-training changes access to the sequence prior

Mask placement. To follow up the query dependence in Figure 1B, we test both NTv3 sizes and training stages on all 512 windows (B). Each query hides the central base and corrupts exactly 144 other positions, using two fixed random draws. Uniform masks span the window; nearby masks fall within 128 bases of the target; distant masks leave 256 bases on either side untouched. Distant scattered and compact masks have equal counts on each side and mean distances matched within half a base. A further control replaces the same distant bases with other nucleotides, without adding MASK tokens.

Distant scattered masks give lower loss than compact masks for 650M post (1.794 versus 2.167 bits; paired difference 0.373, 95% interval 0.283–0.467). Thus the number of hidden bases and their mean distance do not explain the difference. Base replacement does not reproduce the benefit. The 100M post model behaves similarly, while pre-trained checkpoints change much less (Figure 9A–B).

Exchanging internal states. NTv3's decoder combines the final transformer representation with convolutional skip features (Boshar et al., 2025). We exchange complete states after the final transformer block between single-mask and corrupted inputs, preserving the recipient's skips. We also exchange only the skips. Exchanging identical inputs changes log probabilities by at most 1.6×10^{-5}. For 650M post, uniform masking improves loss by 0.821 bits; exchanging only the deep state reproduces 0.812 bits. Exchanging only skips changes loss by $+0.014$ bits, with an interval crossing zero. Distant masks and the 100M model show the same pattern. The effect follows the deep state.

Unrelated donor sequences. We next take the deep state from an unrelated sequence in the same GC stratum. The recipient's single-mask input and skips stay fixed. Comparing single-mask and distant-mask states from the same donor isolates the effect of masking that donor. We use 256 randomly matched pairs, with no window reused in either role, and bootstrap the pairs.

Masking the unrelated donor improves loss by 0.246 bits for 650M post (95% interval 0.147–0.348) and 0.612 for 100M post (0.325–0.885). Pre-trained donor controls show no supported improvement. Part of the masking response therefore transfers between sequences. NTv3's training recipe uses clean inputs for functional losses and corrupted inputs for language modeling (Boshar et al., 2025), so mask patterns could serve as a task cue.

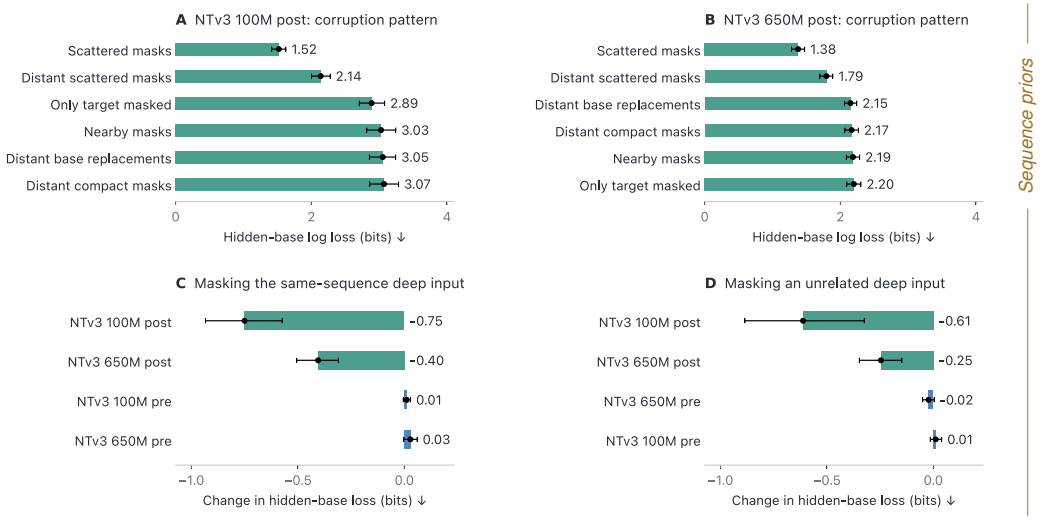

Figure 9: **The retained prior depends on the pattern of missing DNA.** *A–B:* 144 additional corruptions with varied placement or token type; distant conditions preserve 512 neighboring bases. *C–D:* loss change when the deep state comes from a masked input, keeping single-mask skips fixed. C uses the same sequence; D compares masked and single-mask states from the same unrelated, GC-matched donor. Blue denotes pre-trained and teal post-trained checkpoints. Intervals use 4,000 paired resamples: 512 genomic blocks in A–C and 256 disjoint pairs in D, after averaging mask draws.

Fixed internal activation shift. To test whether a fixed change in an internal activation could reproduce the loss reduction shown by additional masking, for single-mask queries, we did the following: For each of the 768 training blocks, one 1,200-base sequence was selected. For each model in Table 3, two queries were produced per sequence: one masking the central target base and one masking the target base as well as 144 uniformly sampled positions from the remaining 1,199 bases. Three activation shift vectors are produced per model by averaging the difference between activations of the two queries over the 768 sequences, and all latent positions, from the outputs of the first, middle and last transformer blocks:

$$\boldsymbol{\Delta}_\ell = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{t=1}^{L}\left[\mathbf{h}_{\ell,t}\big(x_n^{\text{additional}}\big) - \mathbf{h}_{\ell,t}\big(x_n^{\text{single}}\big)\right]. \tag{14}$$

Here $\mathbf{h}_{\ell,t}(x) \in \mathbb{R}^d$ is the output activation at latent position t of transformer block ℓ, L is the number of latent positions, and $N = 768$ is the number of sequences. Averaging over both indices yields a single d-dimensional vector per block, which is why it can be broadcast to every latent position at test time.

The choice of vector and multiplier was made on a separate 128-window selection set. Each vector was tested at its own block by adding it to the original activation with multipliers 0.5, 1, 1.5 and 2, using single-mask queries. The first block with multiplier 1.5 gave the best result and was fixed for all subsequent evaluation. With the vector and multiplier fixed, we tested on 1,536 windows from 48 unused 1-Mb blocks, with balanced strands and target base positions between 256 and 943. The activation shift vector is added at every latent position, with model weights frozen and only the target base masked. Results are shown in Table 3.

To test whether the benefit depended on the direction of the shift, we compared with a negated vector and three random vectors with the same Euclidean norm and multiplier, applied at the same block output. Activation shift was also compared to temperature scaling and affine calibration of the four native logits, fitted on this experiment's own 768 estimation windows rather than the 1,024 fitting windows of Appendix B.3, with five-fold selection of affine regularization. All activation experiments use FP32 with TF32 disabled; zero-shift controls are exact.

Activation shift substantially improved the accuracy in the post trained models, approximately matching the MLM-only pretrained models, which remained almost unchanged across all conditions. Temperature scaling and affine-logit calibration do not reproduce these gains. Negating or randomizing the shift produced small, mixed changes compared to the unmodified model. The accuracy advantage of activation shift over affine calibration was 20.5 percentage points (simultaneous 95% interval: 15.7–25.5) for 100M post and 25.7 points (20.9–30.6) for 650M post. These intervals use 20,000 shared block resamples and Bonferroni correction across 28 contrasts.

Condition	NTv3 100M pre	NTv3 650M pre	NTv3 100M post	NTv3 650M post
Unmodified	51.3	55.3	28.8	29.1
Temperature scaling	51.3	55.3	28.8	29.1
Affine calibration	52.0	55.5	30.7	30.2
Activation shift	51.4	55.6	51.2	55.9
Negated shift	51.7	55.0	29.9	28.9
Random shift 1	51.6	55.8	28.9	29.8
Random shift 2	51.3	55.3	27.0	29.6
Random shift 3	51.3	55.1	30.0	29.8
Eight-draw ensemble	51.5	55.1	54.4	58.4

Table 3: **Masked-base accuracy under changes to the first transformer block output activation and output logits.** The reported accuracy (%) is measured on the same 1,536 windows from 48 held-out genomic blocks. All conditions except the eight-query ensemble use single-mask queries. The eight-draw ensemble averages nucleotide probabilities across queries with the same target base masked but different additional random masks. Values are point estimates.

B.3 PROBABILITY CALIBRATION

We fit one positive temperature per native checkpoint on the original 1,024 fitting windows, minimizing hidden-base log loss:

$$q_T(a \mid r) = \frac{q(a \mid r)^{1/T}}{\sum_b q(b \mid r)^{1/T}}. \tag{15}$$

We search $0.1 \leq T \leq 100$; no fit reaches a boundary. Temperatures are then frozen for the 512-window test and natural-haplotype panel. The first panel had informed the earlier diagnosis; natural haplotypes were assembled after freezing calibration. Temperature scaling changes confidence but preserves base and allele-score rankings. It does not alter reported causal sequence likelihoods.

On the balanced natural-haplotype panel below, calibration reduces GENERATORv2 loss from 3.735 to 2.091 bits (paired change -1.644, 95% interval -1.961 to -1.363). Switching NTv3 650M post from a single masked query to the eight-query readout reduces loss from 2.337 to 2.225 bits and raises accuracy from 26.3% to 36.3%. Neither beats the two-bit uniform control under balanced sampling.

B.4 SEQUENCE PREDICTION ON NATURAL HAPLOTYPES

We select 512 nonoverlapping chromosome-22 SNP windows from the phased 1000 Genomes callset (Byrska-Bishop et al., 2022), using its 2,504 unrelated participants. There are 128 sites in each alternate-frequency range: 1–5%, 20–40%, 60–80% and 95–99%. Each site supplies one REF-carrying and one ALT-carrying chromosome, plus the GRCh38 sequence. We reconstruct all called variants in each window, including indels. Selection uses sequence and genotype validity, not model scores.

Predicting the carried allele. We evaluate all four bases in each carrier's own surrounding sequence, giving the two carriers equal weight within a site and then averaging sites. All primary readouts have loss above uniform (Figure 10A). This is a balanced sample of known polymorphic sites, not average prediction across personal genomes. Reweighting each sampled pair by its cohort allele frequency changes the result: calibrated GENERATORv2 reaches 1.971 bits.

Changing the surrounding haplotype. For each site, we subtract the ALT/REF log odds in the REF-carrier background from those in the ALT-carrier background. Positive values favor the allele carried by the new background. The 116 pairs with identical hidden flanks give exactly zero, as required: identical inputs are scored once before assigning carrier labels. On the 396 distinct pairs, Evo2 is near zero, while NTv3 650M post shows a small positive response (0.042 bits, 95% interval 0.004–0.093; Figure 10B).

Target positions vary across six token phases; we also check matched phases and windows without indels. GENERATORv2's negative response varies under these checks, so we do not interpret it as a stable preference for the wrong haplotype. NTv3 uses fixed masks tied to homologous positions. Intervals resample 39 genomic blocks, keeping each block's sites together.

B.5 FROM SEQUENCE SCORE TO FUNCTIONAL RESPONSE

We score all 170,292 Siraj substitutions in HepG2 and K562. Causal models use full-sequence alternate-minus-reference log probabilities, averaged over both strands. NTv3 uses masked allele log odds with either one mask or the fixed eight-query readout. All three substitutions at a site share the same masks.

To measure base reconstruction, we normalize the three alternate log odds and a zero reference score into a four-base distribution. This uses 56,742 sites with all three substitutions; 35 incomplete sites are excluded. Strand averaging uses the normalized geometric mean. Reconstruction and response correlations give each of the 128 regulatory sequences equal weight.

Evo2 and CENO recover 55.2% and 62.4% of reference bases on sequences where their signed response correlations are only 0.045 and 0.029. PhyloP, GPN-MSA and calibrated GPN-Star-V also predict signed response weakly ($\rho = 0.047$, 0.046 and 0.062), whereas a fitted NTv3 650M post readout reaches 0.429 (Figure 11).

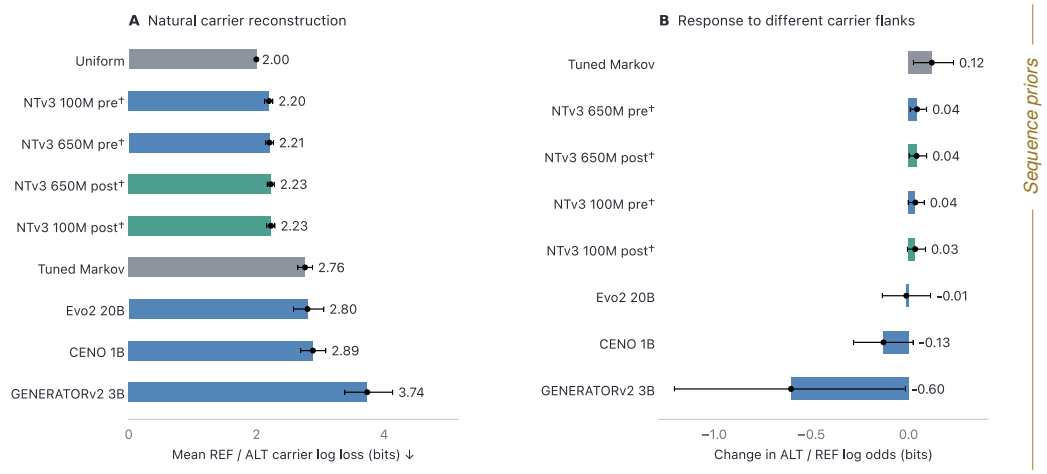

Figure 10: **Predicting alleles in natural haplotypes.** *A:* base log loss on 512 balanced REF/ALT carrier pairs at polymorphic sites; uniform prediction costs two bits. *B:* change in ALT/REF log odds between the carriers' backgrounds, for 396 pairs with distinct hidden flanks. Positive values favor the allele carried by the changed background. NTv3 (†) uses eight-query denoising; causal models use native, uncalibrated probabilities. Intervals bootstrap the 39 genomic blocks (95%).

Evolutionary priors do better on unsigned splice disruption. Across 306,343 OpenSplice SNVs in 466 genes (Quarantani et al., 2026), GPN-MSA, PhyloP and GPN-Star-V's calibrated magnitude reach 0.226, 0.198 and 0.247, respectively; four released splice heads reach 0.386–0.476.

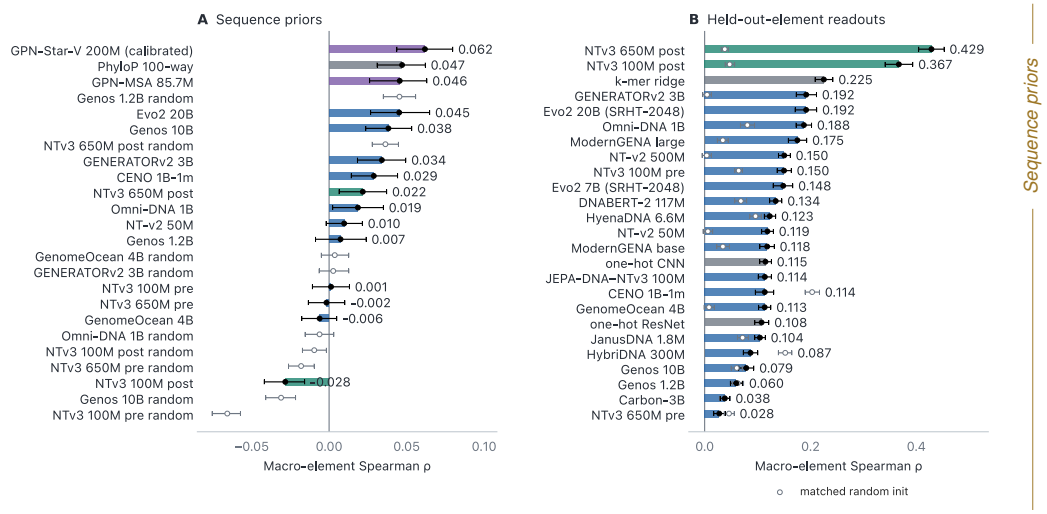

Figure 11: **Sequence score and assay response.** Siraj objective-native scores and calibrated alignment priors (left) compared with held-out-element readouts and matched random networks (right). Evo2 20B uses full-sequence log odds; NTv3 rows use single-mask queries. Black intervals bootstrap complete elements (95%); hollow points and gray intervals denote random networks.

C Functional prediction: Variant-effect prediction and label efficiency

Model comparisons. Figure 2 compares the same 18 model/readout configurations in both panels. SpliceAI and AlphaGenome use published minigene scores on all 590,104 OpenSplice interventions in 466 genes, reaching macro-gene correlations of 0.674 and 0.621. These native heads retain historical supervision. The 3–6-mer ridge baseline ($\alpha = 10$) uses the same five gene folds as the frozen readouts and reaches 0.318. Strict COMPASS transfer excludes sequence-overlapping target events and removes the 23 shared genes from both datasets. CNN/ResNet models fit substitutions from 398 source genes and use another 45 to select the checkpoint. Training settings are given in

the released configuration files. COMPASS labels enter only final scoring. CNN and ResNet reach 0.296 and 0.270 on 4,618 variants in 88 genes. Effects use clipped-logit targets; intervals use 10,000 whole-gene resamples.

For a target correlation τ, label savings in Figure 12 are

$$L_\theta(\tau) = \log_{10} \frac{n_{\text{baseline}}(\tau)}{n_\theta(\tau)}. \tag{16}$$

We interpolate only between observed log-budgets and censor models that never reach τ. At $\tau = 0.20$, NTv3 650M post uses about 10.6 labels versus 20,477 for the one-hot ResNet, $L = 3.29$ (roughly $2{,}000\times$). All nine seed pairings retain a greater-than-three-order advantage over the prespecified target range.

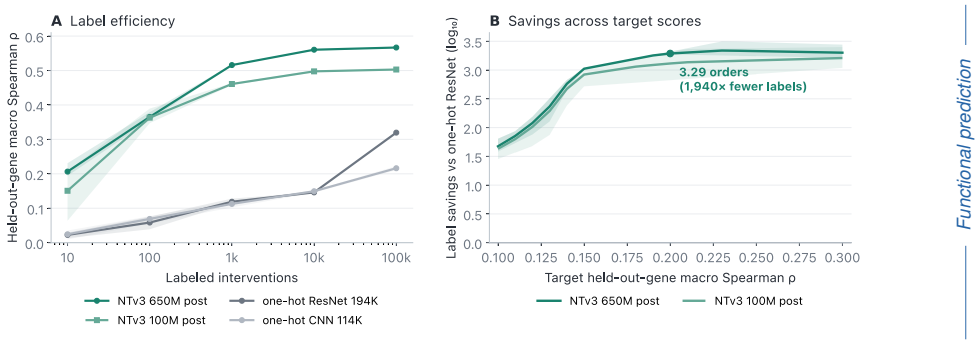

Figure 12: **OpenSplice label efficiency.** Nested budgets use identical sampled records across models and controls. No curve is extrapolated beyond a measured budget.

Direction and magnitude checks. NTv3 650M post correctly predicts the direction of 74.1% of assay-resolved effects ($|\Delta\text{PSI}| > 1.96\sigma$), versus 69.2% for the one-hot ResNet. Across all interventions, their noise-adjusted RMSEs are 16.55 and 17.27 percentage points. Metrics are averaged across genes. NTv3's mean within-gene calibration slope is 0.98 $[0.95, 1.01]$, without test-label recalibration. Full results remain in the archive.

C.1 Editing efficiency from genomic context

Experimental setting. We use the K562 TRIP integration library from PRIDICT2 (Mathis et al., 2025) to test whether genomic surroundings predict editing efficiency for a fixed program. Its 1,202 barcode records span 1,196 genomic coordinates on 23 chromosomes and share a 640-bp reporter insert. Prime editing uses one fixed guide; ABE8e-SpG, BE4max and SpCas9 share another. The four endpoints measure genomic-context variation in one library, not alternative guides or independent cohorts. Excluding missing outcomes leaves 1,182 prime-editing, 1,169 adenine-editing, 1,194 cytosine-editing and 1,196 Cas9 records.

Sequence representations and readouts. Every primary model receives the same 8,192 native reference bases surrounding the integration, without the reporter. We concatenate representations pooled over the whole window and central 1,024 bases. The frozen backbones include NTv3 100M/650M pre and post, four architecture-matched random networks, AlphaGenome, Evo2 20B and GENERATORv2 3B; 3–6-mer counts provide a baseline. AlphaGenome receives symmetric `N`-padding to its supported 16-kb input and uses the same downstream readout. A separate configuration receives 16 kb of native sequence with the same pooling regions. Both are reported without selecting the higher score.

Each endpoint uses five outer chromosome-group folds, training-fold feature scaling and ridge with fixed $\alpha = 1000$. No measured Cas9 efficiency, chromatin measurement or reporter-program covariate enters the predictor. Held-out chromosomes contribute no assay labels to fitting; their sequences may have appeared in backbone pretraining. For every model and endpoint, a nested sensitivity selects $\alpha \in \{1, 10, \ldots, 10^9\}$ by pooled mean squared error in three inner chromosome folds.

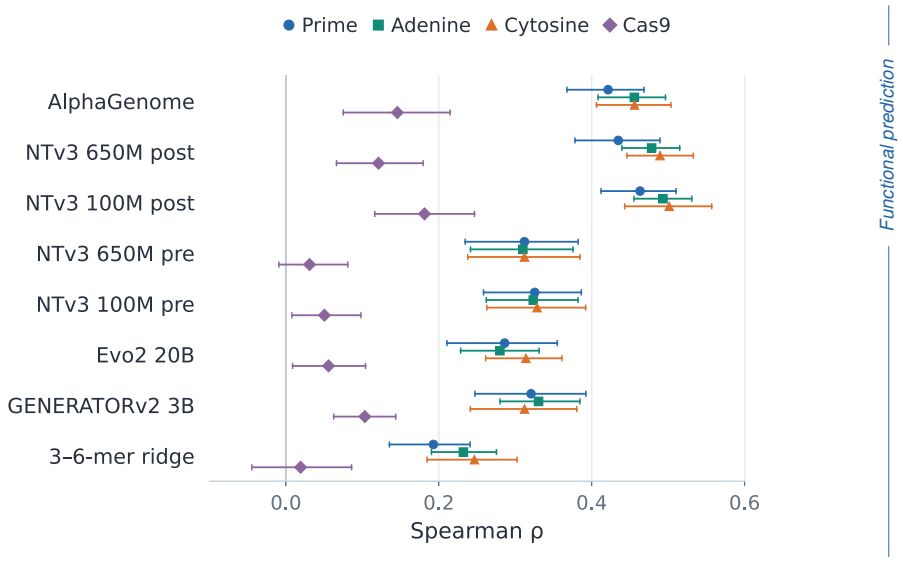

Figure 13: **Editing efficiency from genomic context.** Spearman correlations from five chromosome-held-out folds for four editor systems in the shared K562 integration library. Models receive matched 8-kb reference sequence and fitted ridge readouts; AlphaGenome uses symmetric padding to its supported 16-kb input. Whiskers are pointwise 95% whole-chromosome bootstrap intervals, not simultaneous comparisons between models. Random controls, AlphaGenome with 16 kb of unpadded sequence and the regularization sensitivity remain in the full results.

Metrics and uncertainty. Figure 13 reports Spearman correlation over pooled held-out predictions. Intervals use 50,000 shared whole-chromosome resamples; paired claims retain the simultaneous family of 312 comparisons across 13 configurations and four endpoints. Resampling conditions on fixed fits and omits refitting variation and unknown cell-clone dependence.

Results. Both post-trained NTv3 sizes outperform their pre-trained counterparts, Evo2 and GENERATOR on prime and base editing under both readouts. AlphaGenome and post-trained NTv3 are not statistically ordered; NTv3's size ordering reverses under nested regularization. Cas9 is harder. Correlations evaluate efficiency ranking.

D BIOLOGICAL CONTEXT: CELL, TISSUE AND PERSONAL-SEQUENCE EFFECTS

We test context dependence beyond the HepG2–K562 variant-effect contrast. On COMPASS (Figure 14), AlphaGenome with a 2,001-bp scoring window reaches 0.479 for average effects but -0.018 for cell-specific deviations; replicate agreement for the latter is 0.151. Predicting the average splice effect therefore does not imply predicting its cell-specific component.

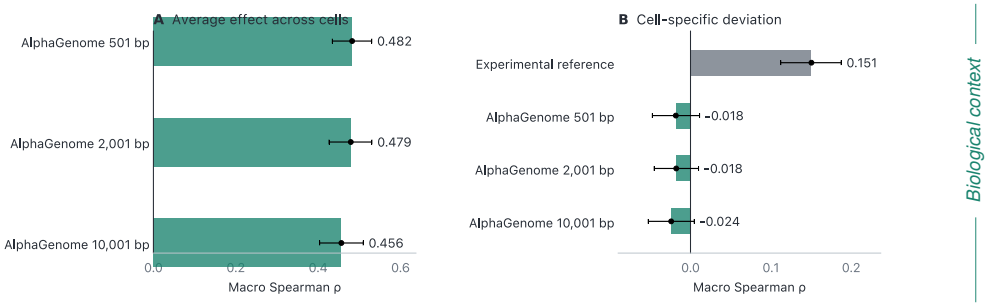

Figure 14: **Average effect prediction does not guarantee cell-specific effect prediction.** Average splice effects use pooled measurements; cell-specific deviations use repeat 2. The experimental reference compares repeat measurements. Whiskers show 95% bootstrap intervals over genes.

D.1 TISSUE-EXPRESSION CONTRASTS AND IDENTIFICATION

The tissue comparison in Figure 15 uses 5,782 common test genes across 51 tissues and $\log(1+\mathrm{TPM})$ expression. For each gene and query tissue, the contrast partner maximizes the absolute observed expression difference. The same partners apply to every model and the split-half reference. This tests outcome-defined contrasts, not prospective tissue selection. Tissue identification compares predicted and measured expression profiles among the 51 candidates, giving fractional credit for ties; chance is $1/51$. Intervals use 10,000 paired whole-gene resamples, with reranking in each resample. The reference predicts donor half B from half A using the same full-data contrast partners; its uncertainty conditions on the saved donor split. Native RNA outputs retain their original scales and historical supervision. Fitted tissue readouts use frozen mean-pooled representations, a fixed 512-dimensional projection and ridge regression selected on validation genes. The added models use 131,072-base windows, except GENERATORv2, which uses a centered 98,298-base crop to respect its token limit. Fits use 12,146 training and 1,350 validation genes, disjoint from the common test set.

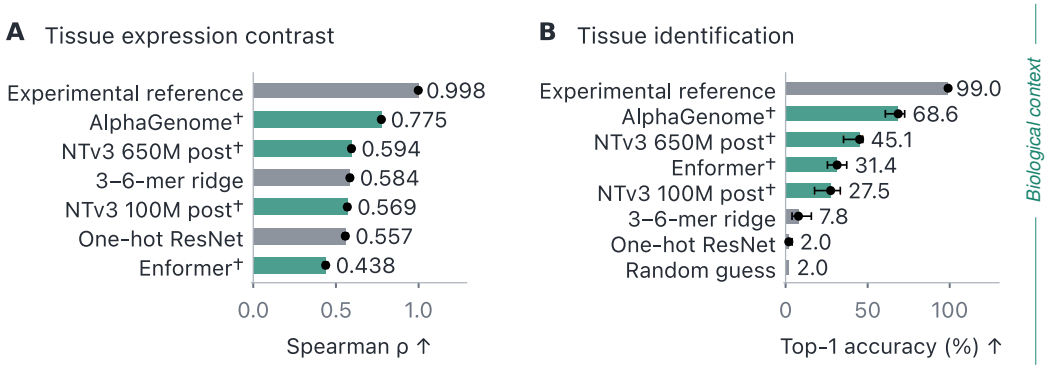

Figure 15: **Predicting tissue expression contrasts and identifying tissues.** A: within-gene GTEx expression contrasts, choosing the comparison tissue by the largest absolute measured difference. B: tissue identification from expression profiles; chance is $1/51$. Both use 5,782 held-out genes and 51 represented tissues, not unseen tissues (GTEx Consortium, 2020). Experimental references compare donor groups; they are not ceilings. (†) mark native outputs. Whiskers show available 95% gene-bootstrap intervals.

D.2 PERSONAL DNA AT FAMILIAR AND UNSEEN LOCI

We compare expression predictions from a phased personal genome x_p and a matched control x_p^{ctrl} under the same model and holdout:

$$\text{personal sequence gain} = \rho_{\mathcal{G}}(\widehat{y}(x_p), y_p) - \rho_{\mathcal{G}}(\widehat{y}(x_p^{\mathrm{ctrl}}), y_p). \tag{17}$$

MAGE tests whether the correct cis genotype improves expression prediction for a new person (Taylor et al., 2024). We reconstruct two haplotypes for 729 people and 557 chromosome-22 genes. Overlapping 8-kb windows are clustered before splitting, so strict folds withhold both person and complete local locus.

At genes represented in training, a positional one-hot readout benefits from the correct phased genotype over a population- and fold-matched whole-genotype shuffle. The gain rises from 0.036 at 1 kb to 0.062 at 8 kb (Figure 16, left). When person *and* locus are unseen, every interval crosses zero and the 8-kb gain is -0.001 (right).

The MAGE analysis covers phased chromosome-22 SNVs and total expression, not allele-specific expression, indels, structural variation or distal regulation. Both source-identified sample-swap identities are excluded.

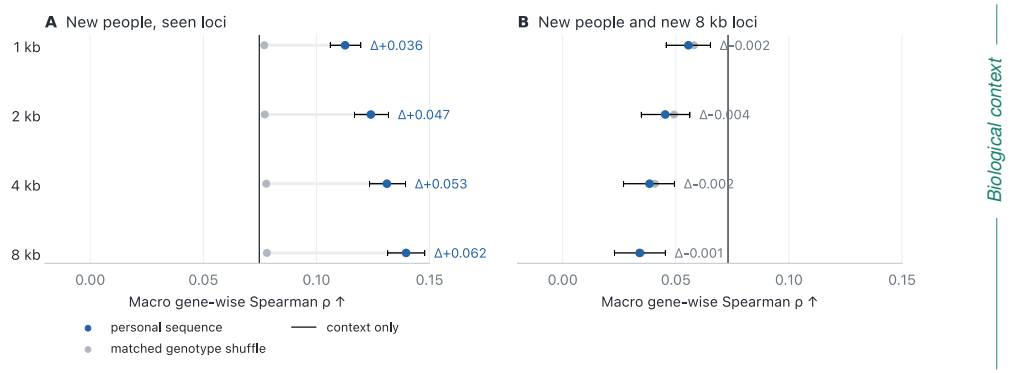

Figure 16: **Cis genotype helps at represented loci.** Correct phased sequence is compared with a population- and fold-matched genotype shuffle. Wider windows help at represented loci (left), but not when the 8-kb locus is also held out (right). Intervals bootstrap genes.

E GENETIC INTERACTIONS: INTERACTION AND UNSEEN REGULATORY ARRANGEMENTS

E.1 COMPASS PREDICTIONS.

Transferred component predictions use the OpenSplice substitution readouts (ridge $\alpha = 10$, SRHT-512 projection), excluding exact sequence overlaps from the target panel. We subtract the fitted wild-type self-delta before forming the interaction. The comparison includes 15,035 single-mutant and 48,054 double-mutant cell measurements in 110 genes; intervals use 10,000 gene resamples. No COMPASS outcomes enter fitting or selection. Native heads use their released outputs, so the comparison includes differences in supervision.

E.2 REMOVING PREDICTED INTERACTION WHILE PRESERVING MAIN EFFECTS

We remove predicted interactions from the COMPASS and Siraj panels to test how much they affect overall effect scores. Predictions stay fixed and no new parameters are fitted.

Two exact ablations. On COMPASS, replace the predicted double-mutant effect by the sum of predicted single-mutant effects, $\widehat{\Delta}^{\mathrm{abl}}(\delta+\epsilon) = \widehat{\Delta}(\delta) + \widehat{\Delta}(\epsilon)$. This sets $\widehat{\mathcal{I}}$ to zero while preserving both single effects. We use the same 48,054 double-mutant–cell observations from 110 genes as Figure 4. Native AlphaGenome/SpliceAI outputs and the OpenSplice-trained readouts retain their original construction and scales.

Siraj measures a query mutation on two allelic backgrounds. Replacing its two predicted effects by their mean, $\widehat{\Delta}^{\mathrm{abl}}_A = \widehat{\Delta}^{\mathrm{abl}}_R = (\widehat{\Delta}_A + \widehat{\Delta}_R)/2$, preserves the average effect and removes background dependence. We score 68,026 paired mutations in 114 elements as 136,052 query–background observations, averaging each measured effect across HepG2 and K562. Frozen readouts hold out complete elements containing both backgrounds; native outputs use no Siraj outcomes. The two assays test different conditional responses.

Uncertainty and limits. We use 20,000 paired biological-unit resamples and Bonferroni correction across 22 contrasts: Spearman and RMSE for all eleven model–assay comparisons. Correlations use 104–106 COMPASS genes with defined values under both conditions and all 114 Siraj elements; RMSE uses all units. Computing COMPASS correlations within gene and cell before averaging cells preserves each rank contrast's direction. Native Siraj accessibility scores have different units from the assay: their RMSE contrasts remain archived and in the correction family, but are not interpreted as biological model-choice evidence.

The largest absolute rank change is 0.0134; all simultaneous rank intervals lie inside $[-0.019, 0.031]$. Retaining interaction provides some small benefits (Table 4) and reduces COMPASS RMSE for SpliceAI and NTv3 100M post. COMPASS interaction replicate agreement is 0.171. In Siraj,

Predictor	Full ρ	Ablated ρ	Difference	Simultaneous 95% interval
COMPASS: double-mutant effect				
AlphaGenome	0.471	0.462	$+0.0098$	$[-0.0059, \ 0.0245]$
SpliceAI	0.555	0.559	-0.0039	$[-0.0181, \ 0.0044]$
NTv3 100M post	0.423	0.418	$+0.0050$	$[-0.0109, \ 0.0219]$
NTv3 650M post	0.500	0.493	$+0.0070$	$[\ 0.0014, \ 0.0144]$
GENERATORv2	0.257	0.243	$+0.0134$	$[-0.0066, \ 0.0307]$
Siraj: query effect on paired backgrounds				
AlphaGenome	0.528	0.521	$+0.0070$	$[\ 0.0014, \ 0.0120]$
Enformer	0.450	0.451	-0.0006	$[-0.0064, \ 0.0046]$
NTv3 100M post	0.417	0.417	-0.00003	$[-0.0057, \ 0.0051]$
NTv3 650M post	0.487	0.481	$+0.0055$	$[\ 0.0009, \ 0.0099]$
GENERATORv2	0.208	0.211	-0.0039	$[-0.0077, -0.0001]$
Evo2 20B	0.206	0.209	-0.0027	$[-0.0068, \ 0.0013]$

Table 4: Overall effect correlation after removing the tested interaction. Differences are full minus ablated macro-unit Spearman, computed before rounding; positive values favor retaining interaction. All eleven model–assay comparisons are shown. Intervals condition on the frozen predictions and resample complete genes or elements. They use the 22-comparison family described below.

AlphaGenome's interaction correlation is 0.260, yet removing that response changes its overall effect correlation by just 0.007.

Conditional-response training. We tested whether adapting NTv3 improves interaction prediction under matched supervision. Complete loci stay together across training, validation and test. These retrospective experiments are separate from transfer without COMPASS labels.

NTv3 100M post's final decoder block was updated using mutations measured on both allelic backgrounds at 68 Siraj training loci. Frozen and adapted models use the same construct labels; all three seeds are reported. On 23 test loci, ordinary adaptation raises effect correlation from 0.381 to 0.446 (paired gain 0.065, interval 0.005–0.125), but interaction changes from 0.053 to 0.038 (difference -0.016, interval -0.040–0.007). Selecting models for validation interaction accuracy while constraining effect accuracy also gives no established interaction gain.

E.3 Unseen motif arrangements

The Georgakopoulos-Soares MPRA trains on single- and double-site constructs and tests 77,895 triplets from 1,136 held-out motif multisets (Georgakopoulos-Soares et al., 2023) (Figure 17). Small one-hot networks reach $0.549/0.547$ on absolute activity, statistically tied with the leading foundation model at 0.556 and close to the 0.603 replicate reference. After removing additive motif-count predictions from outcomes and predictions, the small networks lead. This residual tests unseen arrangements beyond motif counts; unlike the COMPASS interaction score, it does not compare measured reference, single-variant and double-variant sequences.

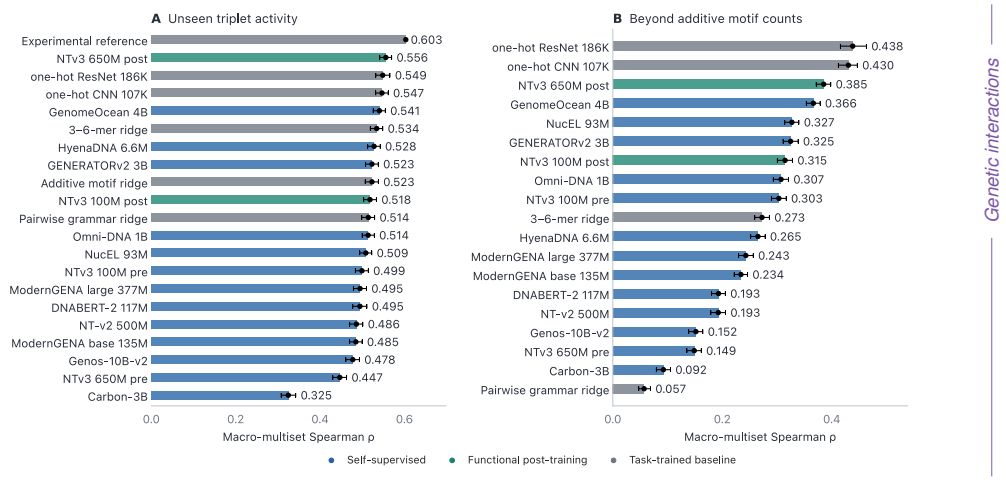

Figure 17: **Unseen triplet arrangements do not require pretraining at full supervision.** Models train on singles and pairs; complete motif multisets are held out. The residual panel removes an additive motif-count prediction from both measured and predicted activity. The experimental reference in A compares repeat measurements.

F GENOMIC DESIGN: TARGET MATCHING AND SELECTIVE EXPRESSION

F.1 MATCHING A TARGET RESPONSE

The direct swap sets each model's target; its position is excluded from the 578–597 candidates per held-out element. AlphaGenome uses released exact-cell accessibility, and NTv3 uses a complete-element out-of-fold Siraj readout. The 331 targets in 110 elements have swap effects exceeding twice their standard error. Equation 18 defines target matching. AlphaGenome and NTv3 650M post rank their first choice at candidate percentiles 24.3 and 30.0, compared with 50 for exact uniform selection. At budget ten their success rates are 0.62 and 0.59, versus 0.38. These within-assay choices use already measured candidates.

Response-matching criterion. For a budget of B measurements, TopB_m contains the candidates ranked by model m without held-out outcomes. A measured candidate is accepted when its effect matches the direct-swap target within combined assay uncertainty:

$$S_B = \mathbf{1}\left\{ \min_{\eta \in \mathrm{TopB}_m} \frac{|\Delta(\eta \mid x, c) - \Delta(\delta_0 \mid x, c)|}{\sqrt{\mathrm{SE}(\eta)^2 + \mathrm{SE}(\delta_0)^2}} \le 1 \right\}. \tag{18}$$

F.2 SELECTIVE MUTATION CHOICE WITH TARGET-STRENGTH MATCHING

On the 114-locus Siraj panel, AlphaGenome's selective policy reduces the other-background response by about 35% in both replicate subsets while preserving mean target response (Figure 18). The control matches attempts and target strength using calibration outside evaluation loci. Adjusted target-difference intervals lie within ± 0.05 log2 units; utility gains are 0.1961 $[0.1390, 0.2577]$ and 0.1857 $[0.1314, 0.2428]$. Smaller panels do not establish target equivalence.

For B backgrounds, a target t and desired direction $d \in \{-1, +1\}$, selective mutation utility is

$$U(\epsilon) = d\,\Delta_t(\epsilon) - \frac{\lambda}{B-1} \sum_{b \neq t} |\Delta_b(\epsilon)|, \qquad U(\text{no edit}) = 0. \tag{19}$$

We fix $\lambda = 1$ and report target and absolute other-background effects separately.

Candidates and decision units. The three panels contain 114 paired-background loci, 11 additional paired-background loci and 13 loci with four backgrounds. Candidates are substitutions within ten bases of the natural variant positions, excluding those positions. Each cell, target background and desired direction defines a decision slot. The policy maximizes predicted Equation 19, abstaining

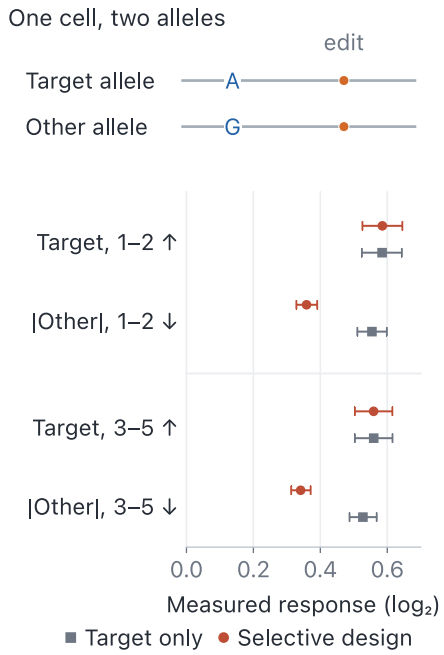

Figure 18: **One cell, two alleles.** AlphaGenome's selected substitutions on the 114-locus Siraj panel. Target and absolute other-background responses are shown for measurement groups 1–2 and 3–5. The target-only control matches attempts and calibrates target strength outside evaluation loci. Whiskers are nominal 95% locus-bootstrap intervals; the equivalence conclusions use multiplicity-adjusted intervals. The same substitution is evaluated on both allelic backgrounds.

at maxima of at most 10^{-10} and weighting tied positive maxima equally within absolute tolerance 10^{-10}. We average decisions within loci, then loci equally; abstentions contribute zero.

Native accessibility predictions. The released AlphaGenome all-folds checkpoint receives each experimental haplotype in its hg38 context with the query substitution. Inputs contain 131,072 bases, adjusting the right flank for shortened haplotypes. For background b, query ϵ and cell c, the score is

$$\widehat{\Delta}_b(\epsilon, c) = \frac{1}{2} \sum_{m \in \{\text{ATAC,DNase}\}} \frac{1}{\mu_{mc}} \sum_{i \in W_{501}(\epsilon)} \left[f_{mc,i}(x_b + \epsilon) - f_{mc,i}(x_b) \right], \tag{20}$$

Here W_{501} is the centered 501-base window and μ_{mc} is the track's released `nonzero_mean`. Each cell and modality has one exact matching track. Scores remain in normalized accessibility units; no Siraj outcomes fit the predictor or a conversion to reporter units.

Replicate-specific reporter responses. We estimate responses separately for subsets A (repeats 1–2) and B (3–5), summing barcode counts for each physical oligonucleotide. With RNA counts R_{icr}, plasmid counts D_{ir} and replicate subset S, define

$$\begin{aligned} z_{ic}^{(S)} &= \frac{1}{|S|} \sum_{r \in S} \log_2 \frac{R_{icr} + 0.5}{D_{ir} + 0.5}, \\ \widetilde{\Delta}_{ic}^{(S)} &= z_{ic}^{(S)} - \frac{1}{|J_{bc}^{(S)}|} \sum_{j \in J_{bc}^{(S)}} z_{jc}^{(S)}, \end{aligned} \tag{21}$$

Here b is i's background and $J_{bc}^{(S)}$ contains its library IDs with z strictly between that background's first and third quartiles. The pool includes all saturation-library records and unmutated m0 controls. Utility uses $\widetilde{\Delta}$. The subsets share library construction and barcodes, so they test within-study robustness.

A control for fewer and weaker attempts. Within each locus and cell, the control allocates the selective policy's K attempts to slots with the largest positive maximum target score, sharing weight at ties. Attempt counts match, but background/direction slots may differ. Four policies sample uniformly from positive target scores in $(0, 25\%]$, $(25\%, 50\%]$, $(50\%, 75\%]$ or $(75\%, 100\%]$ of the slot's maximum; empty bands use positive candidates nearest the midpoint. A fifth chooses the target maximum, sharing ties equally. Utilities use exact policy expectations.

A mixture matches the selective policy's mean *measured target* response by interpolating between bracketing policy means on non-evaluation loci; unbracketed means are flagged unsupported. It uses neither other-background outcomes nor candidate-specific other-background predictions. Five deterministic locus folds cross-fit the 114-locus comparison; smaller panels use the corresponding subset's full 114-locus calibration. This control therefore uses target-assay supervision.

Each of 50,000 locus-bootstrap draws refits calibration outside the evaluation fold, with separate development/evaluation resampling for smaller panels. A and B share draws. The 18-contrast family covers three panels, two subsets and three differences: selective-minus-control utility and target response, and control-minus-selective absolute-other response. Target equivalence requires the adjusted interval strictly inside ± 0.05 $\log_2$ units and concerns mean response, not every locus. All calibration fits are supported. The smaller panels' full results remain in the archive; their contrasts remain in the correction family.

F.3 Cell-selective expression: cohort and measurement

The Siraj HepG2/K562 saturation assay (Siraj et al., 2026) supplies 167,958 substitutions, 280 parental backgrounds and 127 loci. One locus with four indel-containing backgrounds lacks prepared physical-count coverage, excluding 2,334 substitutions. No activity, agreement or model-score filter is added. Each parent and requested target cell defines an opportunity. Both cells serve as target, and all measured substitutions, including the natural-allele substitution, compete with no edit. Unlike response matching, this task permits direct reversions.

For sequence s, cell c and repeat r, define

$$z_{s,c,r} = \log_2 \frac{\mathrm{RNA}_{s,c,r} + 0.5}{\mathrm{DNA}_{s,r} + 0.5}, \qquad \Delta_c(s) = \frac{1}{5}\sum_{r=1}^{5}(z_{s,c,r} - z_{m0,c,r}), \tag{22}$$

where $m0$ is the physical unmutated parent. Effects average log ratios against that parent; counts are not pooled, and no saturation-library baseline is used. DNA counts are shared between cells. Subsets A (repeats 1–2) and B (3–5) recompute their own parent contrasts and test robustness within one library, not independent studies.

Predicting effects. Figure 5C compares AlphaGenome and Enformer accessibility, NTv3 100M/650M post readouts, 3–6-mer ridge and the task-trained one-hot ResNet. AlphaGenome and Enformer effects are aggregated over 501-bp and 1,024-bp output windows, respectively. To calibrate native accessibility, each cell and fold receives a nonnegative through-zero slope:

$$b = \max\left(0, \frac{\sum_i w_i a_i \Delta_i}{\sum_i w_i a_i^2}\right). \tag{23}$$

Here a_i is the raw native effect; a zero denominator gives slope zero. Five deterministic folds keep complete loci together. Fitting excludes evaluation-fold loci and weights training loci equally, then backgrounds within loci and candidates within backgrounds equally. These native outputs are assay-calibrated, not zero-shot.

Fitted predictions combine the predicted mean effect across cells with the predicted cell contrast: HepG2 is the mean plus half the HepG2-minus-K562 contrast, and K562 is the mean minus half that contrast. Readouts fitted to the original assay estimator are evaluated against physical-parent effects without recalibration. Supervision, context and readout training differ, so comparisons concern these pipelines rather than architecture alone.

Choosing and evaluating edits. For fixed $\lambda \in \{0, 0.25, 0.5, 1, 2, 4\}$, nominate the substitution maximizing $\widehat{\Delta}_{\mathrm{target}} - \lambda|\widehat{\Delta}_{\mathrm{other}}|$. Abstain at maxima of at most 10^{-10}; otherwise share weight equally among ties within absolute tolerance 10^{-10} and zero relative tolerance. The primary setting is $\lambda = 1$, chosen without measured outcomes. Evaluate target effect T, absolute other-cell effect O and $U = T - O$. Take absolute values before averaging ties, backgrounds or cells. Average opportunities within loci, then the 127 loci equally. No edit contributes zero; uniform choice uses the exact expectation over eligible substitutions.

Prediction-matched strength control. Weaker edits can reduce desired and unwanted effects together. The control therefore uses the same attempted opportunities and matches predicted target response within each one. It mixes six target-only policies: uniform choice in relative predicted-strength bands $(0, .25]$, $(.25, .5]$, $(.5, .75]$, $(.75, 1]$, minimum positive strength and maximum strength. Empty bands use candidates nearest the midpoint. Interpolation between adjacent component means matches the selective policy's expected predicted target effect. Other-cell scores do not rank candidates within components; measured responses do not construct the mixture. Matching predictions does not establish equal measured effects.

Uncertainty and results. The 24-contrast family tests six models on utility versus no edit, utility versus matched control, target-effect difference and other-cell reduction. We use 50,000 shared locus-bootstrap draws and Bonferroni tails $0.025/24$. Fits, slopes and nominations remain fixed; intervals omit training and calibration-fitting variation. Mean target equivalence requires the adjusted interval inside ± 0.05 $\log_2$; no pipeline establishes it. AlphaGenome, both post-trained NTv3 readouts and Enformer achieve positive utility; AlphaGenome and both NTv3 readouts also improve on their strength controls. Enformer's control gain is inconclusive: 0.055 $[-0.011, 0.120]$. These within-model comparisons do not establish a between-model ranking. NTv3 650M pre attempts only 10.24% of opportunities, explaining its near-zero utility.

Experimental reference. Select using subset A's measurements and evaluate in B, then reverse, retaining the same candidate sets, penalties, no-edit option and strength control. Both directions must have positive utility and gains over target-only and matched-strength controls; averaged utility and target-only gain must also have positive lower pointwise 95% locus-bootstrap bounds. These requirements hold: mean utility is 0.962 $[0.829, 1.092]$, target-only gain 0.371 $[0.297, 0.448]$, and matched-control gain 0.292 $[0.239, 0.347]$. On the same subsets, reference-minus-AlphaGenome utility is 0.603, positive at 108 loci, negative at 18 and tied at one. The reference uses measured candidate responses and shares oligos, parents and library construction across subsets. It shows that selective edits can be found across repeat subsets. It is not a noise ceiling or independent replication.

Protocol. This evaluation is retrospective. The protocol, cohort and interfaces were fixed before native-slope fitting and evaluation; nominations were fixed after calibration and before scoring.

F.4 Selecting enhancers for multi-cell expression programs

Assay, candidates and source split. We use Atlas reporter measurements (Castillo-Hair et al., 2025) (Zenodo record 17410822). The sixteen requests are its eight double-target and eight triple-target Fast SeqProp (FSP) programs, without outcome filtering. Every request receives all 2,411 distinct 145-nt sequences from the double/triple FSP and genomic categories, regardless of original target. Exact DNA duplicates with identical counts are collapsed. The twelve cell lines are NT2_D1, GM12878, 786_O, SKNSH, WERI_Rb1, SJCRH30, HepG2, K562, MCF7, HeLaS3, HEK293 and HMC3; HEK293 denotes the authors' HEK293T key. Retina is excluded because it has only one RNA measurement.

Calibration uses single-target FSP, Deep Exploration Network, genomic and tunable-FSP constructs. Families are connected components linking shared background/DHS identifiers and sequences within global Levenshtein distance 29 on either strand. Excluding families that touch candidates leaves 5,246 source sequences. A deterministic family-hash split assigns 4,223 to training and 1,023 to validation; refitting uses all 5,246.

Measured activity. For each cell and repeat $r \in \{1, 2\}$, activity is

$$y_{icr} = \log_2 \frac{R_{icr} + 0.5}{D_{icr} + 0.5} - \text{median}_{j \in \mathcal{N}} \log_2 \frac{R_{jcr} + 0.5}{D_{jcr} + 0.5},$$

using UMI-deduplicated counts and the published 35 negative controls $\mathcal{N}$, which had undergone pooled outlier removal from forty. DNA preparations follow Supplementary Table 8. We retain the source DNA QC (> 20 total reads in every preparation) without an additional RNA outcome filter. The two RNA measurements come from separate transfections; DNA repeats are technical sequencing replicates of shared plasmid preparations. Normalization is separate for each cell and repeat.

Model predictions. Inserts in the published orientation receive the 16-nt left and 6-nt right cloning junctions, with Ns outside. Inputs are 16,384 bases for AlphaGenome, 196,608 for Enformer and 1,024 for NTv3 post; they represent short inserts, not complete reporters or genomic loci. We use all 4,418 AlphaGenome human ATAC, DNase, CAGE, RNA-seq and histone/TF ChIP tracks, 5,313 Enformer human outputs and 7,362 NTv3 human BigWig outputs. Released nonlinear heads precede insert averaging, with 128-bp bins weighted by overlap. NTv3 uses its official 384-position crop. Pooled outputs receive $\log(1 + x)$, without reverse-complement ensembling or projection. These native outputs require assay calibration. AlphaGenome hidden-feature readouts use the same calibration and candidate catalogue.

Calibration and controls. Source targets average both normalized repeats. Training statistics standardize features; dimensions with standard deviation at most 10^{-7} are removed, and features are divided by the square root of retained width. Ridge minimizes sample-mean summed squared error plus $\lambda\|W\|_F^2$, leaving the intercept unpenalized. Mean twelve-cell validation MSE selects among seventeen penalties $10^{-8}, 10^{-7.5}, \ldots, 1$; scaling and coefficients are then refitted on all source rows. The same procedure calibrates the authors' 64 DHS predictions and strand-invariant 1–6-mer frequencies.

The CNN has three convolutional layers and twelve outputs. Source validation selects the learning rate and stopping epoch for each of three seeds. Models are then refitted on all source sequences and their predictions averaged. Candidate measurements are used only for evaluation; architecture and training settings are supplied with the code.

Selection and evaluation. Predictors rank candidates by Equation 6, selecting ten with ascending-ID ties and no abstention. The same nominations are scored in both repeat measurements, averaged within program and then equally across sixteen programs. Controls are each predictor's minimum-ON-only ranking, uniform catalogue selection and uniform selection within the author-intended target cohort; random utilities are exact expectations. Experimental selection ranks with one repeat measurement and scores the other, reciprocally; these evaluations are dependent. Figure 6A subtracts intent-matched random utility per program/repeat; markers show repeats, not confidence intervals.

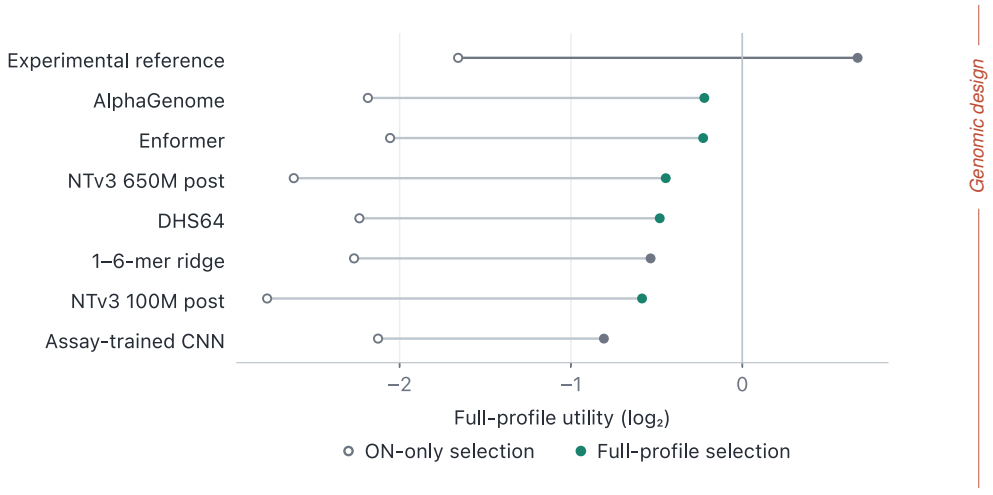

Figure 19: **Considering unwanted activity improves sequence selection.** The same full-profile utility evaluates target-only and full-profile selections, averaged over ten nominations, sixteen requests and both repeat measurements. Open markers rank by the weakest requested activity alone; filled markers also penalize the strongest unwanted activity. The experimental reference selects on the other repeat measurement, reciprocally. Connecting lines are policy differences, not uncertainty intervals.

Figure 6B shows the ten AlphaGenome nominations for HepG2/K562 in predicted order, with profiles averaged across both repeats. This comparatively successful request illustrates the setting, not aggregate performance. Figure 19 compares full-profile and ON-only selection.

G STATISTICAL COMPARISONS AND REFERENCE RESULTS

G.1 INFERENCE AND COMPARISON FAMILIES

Each evaluation uses its stated population. Multiplicity adjustment covers its declared comparison family, not the full sequence of retrospective analyses. Bootstrap draws are shared across models; each task states its resampling unit and correction family. Intervals crossing zero are inconclusive; equivalence requires the entire adjusted interval within the stated margin.

G.2 SELECTED PREDICTION RESULTS

The following tables give uncertainty for the prediction endpoints selected in Table 1; design endpoints follow in Appendix G.3. Native and fitted interfaces retain the main-table definitions.

Model	Prior loss (bits)	Splicing effects ρ
Base composition	1.974 [1.953, 1.997]	—
Evo2 20B	1.388 [1.250, 1.538]	0.367 [0.352, 0.381]
CENO 1B	1.549 [1.385, 1.736]	—
GENERATORv2	1.711 [1.652, 1.775]	0.388 [0.374, 0.402]
DNABERT-2	—	0.198 [0.188, 0.210]
NTv3 100M pre	1.537 [1.454, 1.620]	0.349 [0.335, 0.363]
NTv3 100M post	1.492 [1.393, 1.592]	0.502 [0.488, 0.516]
NTv3 100M post (native)	—	0.168 [0.151, 0.185]
NTv3 650M pre	1.447 [1.352, 1.539]	0.187 [0.174, 0.199]
NTv3 650M post	1.375 [1.275, 1.475]	0.564 [0.550, 0.578]
NTv3 650M post (native)	—	0.461 [0.445, 0.477]
AlphaGenome (native outputs)	—	0.621 [0.605, 0.637]
AlphaGenome (embedding-fitted)	—	0.650 [0.634, 0.665]
SpliceAI	—	0.674 [0.659, 0.689]
k-mer ridge	—	0.318 [0.307, 0.329]
Task-trained ResNet	—	0.488 [0.477, 0.498]

Table 5: **Selected sequence-prior and splicing-effect endpoints.** Estimates and marginal 95% biological-unit bootstrap intervals for the interfaces in Table 1. Prior uses 512 genomic blocks; OpenSplice uses 466 held-out genes. Intervals are not paired model-comparison tests. Native and fitted interfaces retain their main-table definitions; missing values are not zero.

Model	Cell contrast ρ	Interaction ρ
Evo2 20B	0.041 [0.022, 0.060]	0.050 [0.034, 0.066]
CENO 1B	0.036 [0.018, 0.053]	0.010 [-0.007, 0.027]
GENERATORv2	0.066 [0.050, 0.083]	0.046 [0.030, 0.062]
DNABERT-2	0.076 [0.066, 0.087]	0.020 [0.007, 0.032]
NTv3 100M pre	0.076 [0.063, 0.089]	0.010 [-0.000, 0.021]
NTv3 100M post	0.244 [0.215, 0.273]	0.139 [0.110, 0.168]
NTv3 100M post (native)	0.203 [0.168, 0.237]	0.152 [0.123, 0.182]
NTv3 650M pre	0.010 [-0.001, 0.023]	0.009 [-0.004, 0.021]
NTv3 650M post	0.256 [0.229, 0.283]	0.187 [0.155, 0.220]
NTv3 650M post (native)	0.239 [0.207, 0.271]	0.170 [0.138, 0.201]
AlphaGenome (native outputs)	0.398 [0.366, 0.431]	0.260 [0.219, 0.299]
AlphaGenome (embedding-fitted)	0.402 [0.371, 0.433]	0.269 [0.232, 0.304]
Enformer	0.282 [0.242, 0.321]	0.181 [0.141, 0.220]
k-mer ridge	0.211 [0.194, 0.228]	0.032 [0.015, 0.049]
Task-trained ResNet	0.020 [0.011, 0.029]	-0.001 [-0.012, 0.011]

Table 6: **Selected cell-contrast and interaction endpoints.** Estimates and marginal 95% biological-unit bootstrap intervals for the interfaces in Table 1. Cell contrast and interaction use 128 and 114 elements. Intervals are not paired model-comparison tests. Native and fitted interfaces retain their main-table definitions; missing values are not zero.

G.3 SELECTED DESIGN RESULTS

Design metrics. Single-edit design ranking reports measured residual percentile $100(r-1)/(N-1)$, using average ranks for ties among N eligible edits, excluding direct reversions. Average within and then across 110 elements; uniform selection gives 50. Cell-selective and multi-cell metrics follow Appendices F.3 and F.4. Overlapping programs are not independent replicates; raw utilities across tasks should not be averaged.

Model	Single-edit rank percentile [95% interval]	Cell-selective U [95% interval]	Multi-cell gain
Evo2 20B	43.5 [39.5, 47.5]	-0.350 [-0.418, -0.286]^{p}	0.275
CENO 1B	48.9 [44.9, 52.7]	-0.318 [-0.392, -0.248]^{p}	-0.581
GENERATORv2	44.7 [40.9, 48.5]	-0.256 [-0.326, -0.190]^{p}	0.234
DNABERT-2	48.6 [45.0, 52.2]	-0.302 [-0.377, -0.230]^{p}	0.233
NTv3 100M pre	47.3 [43.8, 50.8]	-0.207 [-0.265, -0.151]^{p}	0.339
NTv3 100M post	30.3 [25.9, 34.8]	0.211 [0.073, 0.349]	0.144
NTv3 100M post (native)	32.0 [27.9, 36.2]	0.200 [0.101, 0.302]^{p}	0.344^{c}
NTv3 650M pre	50.4 [47.1, 53.9]	-0.061 [-0.094, -0.034]^{p}	0.368
NTv3 650M post	30.0 [25.8, 34.5]	0.287 [0.144, 0.429]	0.568
NTv3 650M post (native)	29.3 [25.5, 33.3]	0.250 [0.162, 0.341]^{p}	0.482^{c}
AlphaGenome (native outputs)	24.3 [20.5, 28.3]	0.384 [0.237, 0.543]	0.707^{c}
AlphaGenome (embedding-fitted)	25.7 [21.3, 30.3]	0.515 [0.422, 0.612]^{p}	0.833
Enformer	31.5 [27.4, 35.7]	0.290 [0.134, 0.446]	0.700^{c}
k-mer ridge	45.0 [41.2, 48.9]	-0.097 [-0.237, 0.041]	0.394
Task-trained ResNet	47.0 [43.5, 50.6]	-0.215 [-0.322, -0.125]	0.347

Table 7: Selected design endpoints. Single-edit ranking intervals are pointwise element-bootstrap intervals. Cell-selective intervals use the original simultaneous 24-contrast family, except entries marked p: pointwise intervals for additional pipelines outside that family. All condition on fixed fits and nominations. No interval is attached to the 16 overlapping Atlas programs; c marks linear heads fitted to native outputs.