Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation of multiple sclerosis risk loci reveals that genomic oracles fail to generalise sequence effects across host contexts

Selahattin Alperen Uysal

dralperenuysal@gmail.com

Ege University

Research Article

Keywords: regulatory genomics, sequence design, deep learning, Enformer, massively parallel reporter assay, multiple sclerosis

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10850103/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Functional & Integrative Genomics on October 31st, 2026. See the published version at https://doi.org/10.1007/s10142-026-02097-3.

Evaluation of multiple sclerosis risk loci reveals that genomic oracles fail to generalise sequence effects across host contexts

Selahattin Alperen Uysal [1]

[1]Department of Medical Biology and Genetics, Ege University School of Medicine, İzmir, Türkiye.

Contributing authors: dralperenuysal@gmail.com;

Abstract

Sequence-to-function deep-learning oracles are increasingly used to score and select designed regulatory elements, but it is not established that an effect measured for a sequence in one genomic context holds in another. We generated candidate cis-regulatory elements for multiple sclerosis risk loci with a per-cell-type generative model, selected them with a chromatin-accessibility oracle (Enformer), and tested, by direct intervention, what the oracle rewards. Motif density and a specific transcription-factor site were not causal for the oracle's preference; a composition-preserving CpG-content edit was, its sign set by the host locus rather than by the edit itself. Placing the same insert and CpG edit at 24 host loci showed the effect varies with $\boldsymbol{I^2 = 95\%}$ (Cochran's $\boldsymbol{Q = 458.6}$, $\boldsymbol{p = 1.4 \times 10^{-82}}$) and reverses at some loci, no tested host property predicting its direction. The heterogeneity, though not the locus-specific sign reversal, replicated in a second, independently trained oracle (Borzoi). Scored against measured reporter activity from a lentiviral massively parallel reporter assay, the selection objective did not predict measured cell-type specificity, including for a matched positive control. Separately, a sixth-order Markov chain, fitted in seconds, matched or exceeded a trained autoregressive transformer on every generative property required. These results indicate that a sequence-level effect learned or measured by a genomic oracle is not, by default, a property of the sequence alone: it depends on where in the genome it sits, and that dependence is not predictable from cheap host features.

Keywords: regulatory genomics, sequence design, deep learning, Enformer, massively parallel reporter assay, multiple sclerosis

1 Introduction

Sequence-to-function deep-learning models such as Enformer Avšec et al (2021) and Borzoi Linder et al (2025) predict chromatin accessibility, transcription-factor occupancy, and expression directly from DNA sequence at kilobase to megabase scale. Because these oracles are differentiable and cheap to query relative to a wet-lab assay, they are increasingly used not only to interpret existing sequence but to *design* it: a candidate regulatory element is proposed, scored by an oracle, and retained or rejected on the basis of that score Taskiran et al (2024); Gosai et al (2024). This turns the oracle into a stand-in for measurement at the point where a design decision is actually made.

That use rests on a load-bearing assumption that is rarely stated explicitly: a design pipeline that optimises a model score is only as good as the score's behaviour under the kind of sequence perturbation the pipeline performs. If an edit that raises the score at one genomic locus lowers it, or does nothing, at another, then the score is not measuring a property of the inserted sequence alone; instead, it is measuring a property of the sequence in combination with wherever it happens to be evaluated, and a design chosen at one locus gives no guarantee about its behaviour once moved, inserted elsewhere, or integrated into a different construct.

Independent evaluations have already found these models to behave unexpectedly under perturbation. Enformer captures causal determinants of promoters but largely misses the effect of distal enhancers, with the great majority of variation in its predictions for a given sequence set attributable to the promoter alone Karollus et al (2023), and sequence models benchmarked on personal genomes predict the direction of individual variant effects poorly Sasse et al (2023). Both results concern predictions about *existing* sequence. The assumption that matters for design is narrower and has rarely been tested by direct intervention: that an edit which improves the score at one genomic locus does so wherever the edited sequence is placed. To our knowledge this has not been tested across a panel of host loci with the same insert and the same edit held fixed at every locus. Studies that report oracle-guided design typically evaluate the designed sequence at the single locus it was designed for; studies that report genomic-context effects on regulatory activity typically do not hold the causal, oracle-selected edit itself constant across that context. Without both held fixed simultaneously, a locus dependence and an edit effect cannot be distinguished from each other.

Here we test this assumption directly, using multiple sclerosis genetic risk loci as the setting. Multiple sclerosis is well suited to it: its genomic map implicates peripheral immune cells and microglia specifically, placing most risk variants in non-coding regulatory sequence active in identifiable cell types International Multiple Sclerosis Genetics Consortium (2019). We built a generative pipeline using an artificial intelligence architecture (an autoregressive Transformer, designed to learn and generate novel DNA sequences) to propose candidate cis-regulatory elements for three disease-relevant immune and myeloid cell types. We then selected these candidates using a sequence-to-function "oracle" score derived from Enformer, and later verified by Borzoi—both of which are large-scale deep-learning models that act as virtual laboratories, predicting how a DNA sequence will behave biologically across thousands

of cell types. We then asked three questions of the selected sequences: what oracle-rewarded property does selection actually track; is that property causal, established by direct sequence intervention rather than correlation; and does the causal effect of that intervention, measured at one locus, generalise when the identical insert and the identical edit are placed at 24 different host loci. We report that it does not generalise in a predictable way, that a second independently trained oracle reproduces the instability but not its locus-specific direction, and that the selection objective itself does not predict measured reporter activity in an independent reporter assay. The overall design and its checks are summarised in Fig. 1.

2 Materials and Methods

2.1 Chromatin accessibility data and window construction

Peak calls from 51 publicly available ATAC-seq samples spanning two studies were restricted to peaks intersecting multiple sclerosis risk loci in the hg38 assembly: $CD4^+$ T cells and B cells from GSE202087 Guo et al (2022) (MACS narrowPeak Zhang et al (2008)) and central nervous system microglia from GSE307262 Wang et al (2025) (HOMER peaks Heinz et al (2010)). Peaks were centred on the reported summit where the caller provided one and on the peak midpoint otherwise, then resized to a fixed 1000 bp window. This yielded 177,397 peak instances.

Two deduplication steps were applied, both configured in `configs/data_config.yaml`. First, within each cell type, the highest-scoring window was retained and any window starting within 500 bp of an already accepted window was rejected; without this step the same element appears in both the training and validation splits, since every donor contributes a call for the same element. This reduced 177,397 peak instances to 28,026 windows.

Second, windows lying within 500 bp of a window assigned to a *different* cell type were dropped entirely. Before this step, 44.5–57.6% of B-cell and $CD4^+$ windows, and 50.4–53.5% of microglial windows, were shared with another cell type, so near-identical sequences carried conflicting condition labels. Removing them left 10,785 cell-type-exclusive windows ($CD4^+$ T 5,772; B 2,698; microglia 2,315). The retained subset is measurably different in character, with CpG observed/expected falling from 0.388 to 0.265; this is a selection effect and is treated as such throughout.

2.2 Transcription factor motif analysis

Motifs were scanned with position-specific scoring matrices derived from JASPAR2024 CORE Rauluseviciute et al (2024) (*Homo sapiens*) via `pyjaspar`, on both strands, at a relative score threshold of 0.80, where the threshold is placed on the matrix's own 0–1 normalised score range so that it is comparable between matrices of different information content.

Two vocabularies were used and are reported separately. A 16-matrix panel, specified from lineage biology before any scan was run, comprises four shared immune factors (NFKB1, STAT3, IRF4, RUNX1) and twelve lineage-defining factors (TCF7, LEF1, GATA3, TBX21, ETS1 for $CD4^+$ T; PAX5, EBF1, SPIB, POU2F2 for B;

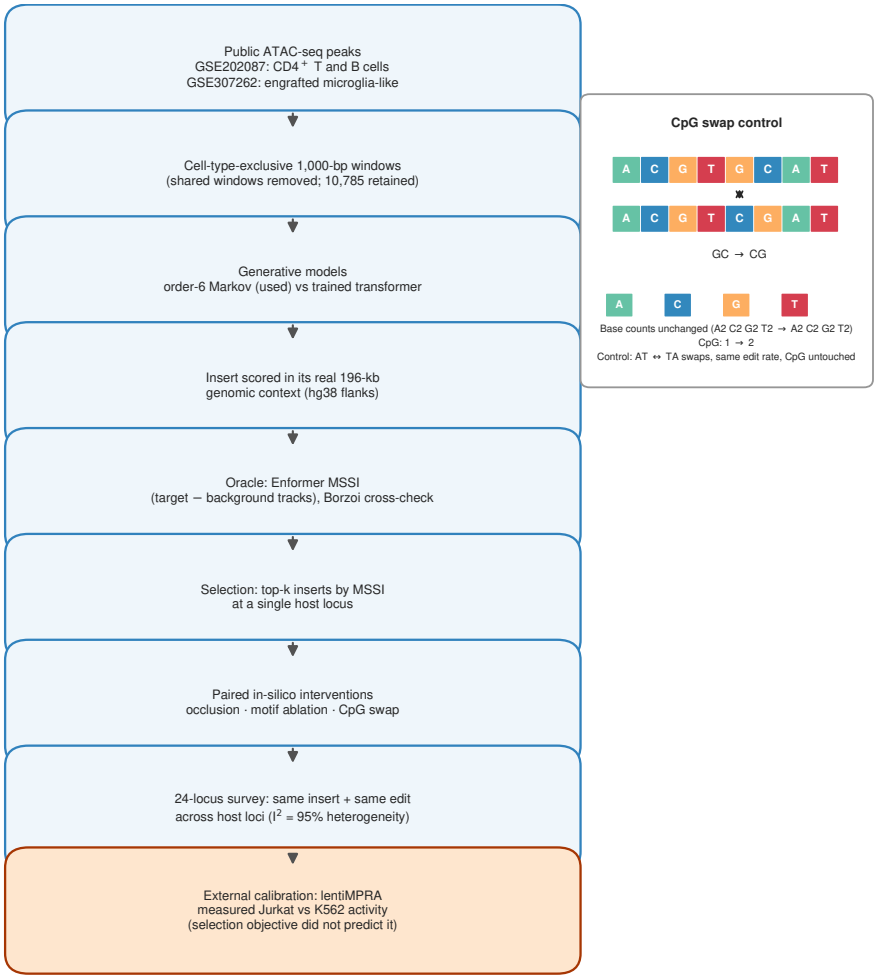


Fig. 1 Study design. Public ATAC-seq peaks from three immune cell types are reduced to cell-type-exclusive 1,000-bp windows; candidate inserts are drawn by generative models (an order-6 Markov chain, with a trained transformer audited alongside) and each insert is scored inside its real 196-kb genomic context by a sequence-to-function oracle (Enformer; cross-checked with Borzoi). Selected inserts are then audited by paired in-silico interventions — occlusion, motif ablation, and a composition-preserving CpG swap (inset: reversing a `GC` dinucleotide to `CG` changes the CpG count while leaving every base count unchanged; the control applies `AT`/`TA` reversals at the same rate) — by a 24-locus survey that holds insert and edit fixed while varying the host locus, and by calibration against measured lentiMPRA reporter activity

SPI1, IRF8, CEBPA for microglia and myeloid). POU2AF1 and SALL1 were omitted because neither binds DNA directly and JASPAR holds no matrix for them. Separately, the entire collection of 720 human CORE matrices was used wherever an

unbiased vocabulary was required; as we show, conclusions about motif density differ between the two, and the panel result is the one that does not survive.

2.3 Generative models

An order-k Markov chain with add-one smoothing and backoff to shorter contexts and finally to the uniform distribution served as the baseline generator (`src/evaluation/markov_baseline.py`), fitted per cell type on that cell type's exclusive windows. Order 6 was used throughout unless stated. Because such a chain reproduces $(k+1)$-mer frequencies by construction, any property of generated sequence that the chain also reproduces cannot be evidence that a learned model captured anything beyond a transition table; this is the standard against which the neural models are judged.

To evaluate neural network capacity on this task, a conditional sequence transformer was implemented and trained from scratch. To make the architecture accessible to readers outside computer science: the model was designed as an autoregressive decoder (conceptually similar to the architecture of generative text models like GPT, but scaled down and adapted for a 4-letter DNA alphabet). Rather than processing the entire sequence at once, it learns to read the DNA string left-to-right and iteratively predicts the next single nucleotide (A, C, G, or T) based on the sequence history it has generated so far. To ensure the synthetic sequences possessed the regulatory grammar of the desired cell type, the model was mathematically conditioned using a "cross-attention" mechanism. This mechanism acts as a constant biological prompt, instructing the network during generation (e.g., "generate a $CD4^+$ T cell enhancer").

A conditional variational autoencoder was also trained for comparison on the same data splits. Held-out negative log-likelihood is reported in nats per base, against the $\ln 4 = 1.3863$ ceiling that corresponds to having learned nothing (i.e., uniform random guessing).

2.4 In-silico oracles

Two independently trained sequence-to-function models were used. Enformer Avšec et al (2021) (`EleutherAI/enformer-official-rough`) takes a 196,608 bp context and returns 5,313 human tracks; Borzoi Linder et al (2025) (`johahi/borzoi-replicate-0`) takes 524,288 bp and returns 7,611. Both are wrapped by a shared base class (`src/evaluation/track_oracle.py`) so that the only model-specific code is weight loading and the axis convention of the forward pass.

We summarise a prediction by a cell-type specificity index,

$$\mathrm{MSSI} = \frac{1}{|T|}\sum_{t\in T}\bar{y}_t - \frac{1}{|B|}\sum_{b\in B}\bar{y}_b, \tag{1}$$

where $\bar{y}$ is a track's mean over output bins, T is the set of target tracks for the cell type in question and B a fixed set of non-immune background tracks. Both track sets are enumerated explicitly in `configs/model_config.yaml`; they are never chosen at runtime.

A designed insert is placed at the centre of the model's receptive field and the flanks are taken from the reference genome at the host locus. Synthetic or zero-padded flanks are refused by the implementation, because the prediction is dominated by the surrounding sequence and made-up padding yields a score that reflects the padding rather than the insert.

2.5 Sequence interventions

Every intervention below is paired with a control that applies the same kind and amount of edit without the property under test, and every variant of a given sequence is scored at the same host locus, so that locus and insert identity are held constant and only the edited property varies.

Occlusion.

Successive non-overlapping 100 bp tiles were replaced with random DNA drawn at uniform base composition (`scripts/occlusion_scan.py`), one variant per tile, which localises an effect if one exists.

Motif ablation.

All sites for a named factor, or a randomly drawn subset of sites across all factors at a specified dose, were shuffled in place (`scripts/motif_ablation.py`). Shuffling rather than replacing preserves local base composition. Controls shuffle an equal number of windows of equal width placed at positions carrying no site for the targeted factor, so that a change in score cannot be attributed merely to some DNA having been scrambled.

CpG manipulation.

Reversing a `GC` dinucleotide to `CG` adds one CpG and reversing `CG` to `GC` removes one (`scripts/cpg_swap.py`). Either edit leaves the count of every individual base exactly unchanged, so GC content and all mononucleotide frequencies are identical between a sequence and its variants; the implementation aborts if they are not. Controls apply the same number of `AT`/`TA` reversals, which is the same class of edit at the same rate without touching CpG.

Multi-locus survey.

To separate host effects from insert effects, one fixed set of inserts, each carrying one fixed edit, was placed at many host loci in a single scoring run (`scripts/locus_survey.py`). Because the inserts do not vary, any difference in the measured effect between loci is attributable to the host. Host loci were drawn one per stratum of the peak-score range.

2.6 External calibration against measured reporter activity

Measured regulatory activity was taken from ENCODE Luo et al (2020) lentiMPRA experiments Agarwal et al (2025) of the Tewhey laboratory that share a single 60,000-element library (ENCSR604ZOZ): Jurkat, a T-lymphocyte line (ENCSR247RPW),

and K562, with HepG2 and SK-N-SH available on a common subset. Per-element activity is the reported $\log_2$ fold change of RNA over input DNA; oligo sequences were taken from the library's own reference FASTA rather than reconstructed from coordinates. We define measured cell-type specificity as the difference in $\log_2$ fold change between Jurkat and K562 for the same element.

Elements were scored by the oracle at a single fixed host locus, for the reason given in Results: the oracle's score is dominated by the host locus, so scoring elements at their own native loci would rank loci rather than elements. A fixed locus also matches what an episomal reporter measures, since the assayed construct carries no genomic context of its own. The scored set comprises an unbiased random subset, which is the only part used for correlation estimates, plus both extremes of measured specificity, which provide power for the group contrast and are reported separately.

2.7 Statistics

Group differences were assessed by two-sided permutation tests on the difference in means; paired perturbation effects were assessed by sign-flip permutation, both with 10,000 permutations unless stated. Where many statistics were tested simultaneously the number of permutations was raised so that the attainable p-value floor, $1/(n_{\text{perm}} + 1)$, lies below the correction threshold, and Benjamini–Hochberg false discovery rate Benjamini and Hochberg (1995) is reported alongside Bonferroni. Between-locus heterogeneity was quantified by Cochran's Q and I^2 Higgins and Thompson (2002) on the per-locus effect estimates and their standard errors. All randomness takes an explicit seed.

3 Results

3.1 Cell-type-specific chromatin signal requires removing shared windows

We assembled 10,785 open-chromatin windows exclusive to one of three cell types ($CD4^+$ T cells, 5,772; B cells, 2,698; microglia, 2,315) from 51 GEO samples across two studies (GSE202087 for $CD4^+$ T and B cells Guo et al (2022); GSE307262 for microglia Wang et al (2025)), after two rounds of deduplication: first within each cell type (177,397 raw windows collapsed to 28,026 unique windows), then across cell types, removing any window whose interval overlapped an open window from another cell type. On the pooled, non-exclusive window set, a 16-factor lineage-specific transcription-factor motif panel (four factors shared across the three immune lineages, twelve lineage-defining factors specific to one) showed no cell-type-discriminating signal: lineage-defining motifs were not enriched over the other two cell types' windows. Restricting to the 10,785 cell-type-exclusive windows changed this: the myeloid factor SPI1 was enriched 1.88-fold in its own windows against 0.96– and 1.24-fold enrichment in the other two cell types (Fig. 2b; $p = 1 \times 10^{-4}$, permutation test), a 6.58-fold enrichment over GC-content-matched random sequence. CpG observed/expected dropped from 0.388 in the pooled windows to 0.265 in the exclusive set (Fig. 2a), consistent with the exclusive windows being enriched for distal enhancers rather than CpG-island

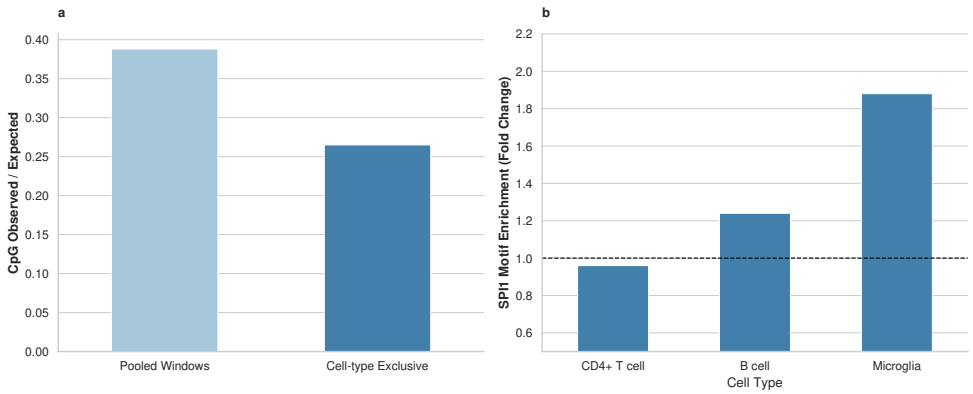


Fig. 2 Effect of restricting to cell-type-exclusive open chromatin windows. (a) Removing windows shared across immune cell types decreases the overall CpG observed/expected ratio, indicating a shift towards distal enhancers. **(b)** SPI1 motif enrichment in the cell-type-exclusive window set: the myeloid factor is enriched 1.88-fold in microglia windows against 0.96- and 1.24-fold in the $CD4^+$ T-cell and B-cell windows (dashed line: no enrichment). The pooled window set carried no cell-type-discriminating signal, so no pooled bars are plotted

promoters. Because this enrichment appeared only after removing windows shared between cell types, all subsequent analyses use the cell-type-exclusive window sets; the exclusivity restriction was a post-hoc analysis choice made to recover this signal, not a decision made in advance of looking at the data.

3.2 A fitted sixth-order Markov chain matches or exceeds a trained transformer

We fitted two generators per cell type on the exclusive windows: a genomic Transformer trained by gradient descent, and an order-k Markov chain fitted by conditional frequency counting, which requires no training step and no GPU. On held-out sequence negative log-likelihood, the transformer reached the lowest value overall (1.2209 nats/base against a 1.3863 nats/base uniform-random baseline), with the order-6 Markov chain close behind at 1.3097 nats/base (Fig. 3a; Table 1). Composition realism (dinucleotide and higher-order k-mer frequencies) was already largely reproduced by an order-1 Markov chain, and motif-profile fidelity against the 16-factor panel favoured the transformer only in $CD4^+$ T cells, with the Markov chain matching or exceeding it in B cells and microglia. On the property most directly tied to Section 3.1 – reproducing the SPI1 enrichment specific to microglia – the order-6 Markov chain reproduced a 1.48-fold enrichment against a real-data value of 1.63-fold ($p = 2 \times 10^{-4}$), whereas the trained transformer reproduced only a 1.17-fold enrichment (Fig. 3b; Table 1) that did not differ from the GC-matched null ($p = 0.20$). Given this, we used the order-6 Markov chain, not the transformer, as the generator supplying candidate sequences for the selection experiments below; order 6 conditions on the six preceding bases, so it reproduces all 7-mer frequencies of its training set without the sparsity of an order-8 chain's 65,536 contexts estimated from only ∼2,300 microglia windows.

Table 1 Generative model performance on held-out data and cell-type signal recovery. The autoregressive transformer achieved lower negative log-likelihood (NLL) but failed to reproduce the biological SPI1 enrichment found in real microglia enhancers. The order-6 Markov chain captured the signal despite higher NLL.

Model / Source	Held-out NLL (nats/base)	SPI1 Enrichment	p-value
Real sequences	—	1.63×	10^{-4}
Markov chain (order 6)	1.3097	1.48×	2×10^{-4}
Autoregressive Transformer	1.2209	1.17×	0.20
Markov chain (order 3)	1.3345	1.06×	0.62
GC-matched random	—	1.04×	0.93

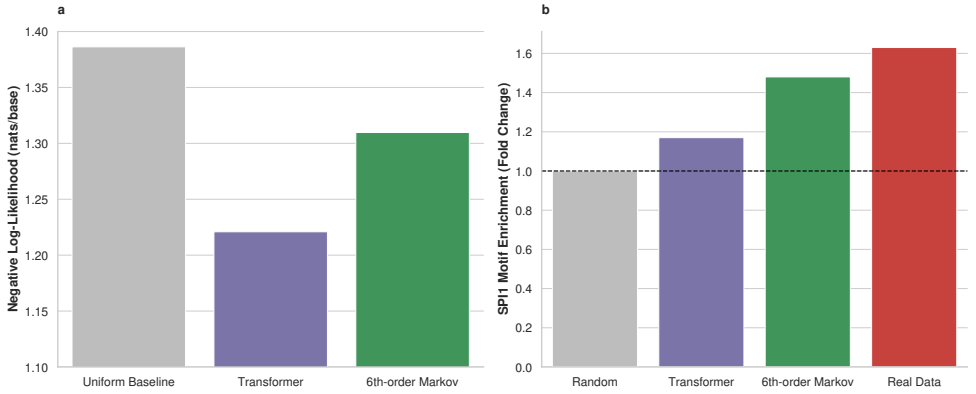


Fig. 3 A simple Markov chain matches or outperforms a trained transformer. (a) Negative log-likelihood (nats/base) on held-out sequences. The transformer performs best, but the 6th-order Markov chain remains highly competitive. **(b)** Recovery of microglia-specific SPI1 motif enrichment. Despite having no learned parameters, the simple Markov chain (1.48-fold) successfully mimics the real data (1.63-fold) much better than the trained transformer (1.17-fold), demonstrating that model capacity was not the bottleneck for this generative task

3.3 The oracle score is dominated by host genomic context, not by insert identity

We scored candidate 1000 bp inserts placed inside a fixed 196,608 bp host window using Enformer (5,313 output tracks; target and background tracks for each cell type enumerated in advance in `configs/model_config.yaml`). The same 331 inserts (real, order-6 Markov and GC-matched random sequence) were scored at each of the five candidate host loci: the standard deviation of the resulting oracle score (cell-type-specific signal index, MSSI) across host loci (0.92, standard deviation of the five locus means) was 36 times larger than its standard deviation across different inserts placed at the same locus (Fig. 4; 0.025, pooled within-locus standard deviation). A paired comparison over 100 host loci additionally showed real sequence and order-6 Markov-generated sequence both scoring above GC-matched random sequence by a small but reproducible margin (real−random: +0.0338, $p = 1 \times 10^{-4}$; Markov−random: +0.0238, $p = 4 \times 10^{-4}$), and the two were not distinguishable from each other. The oracle

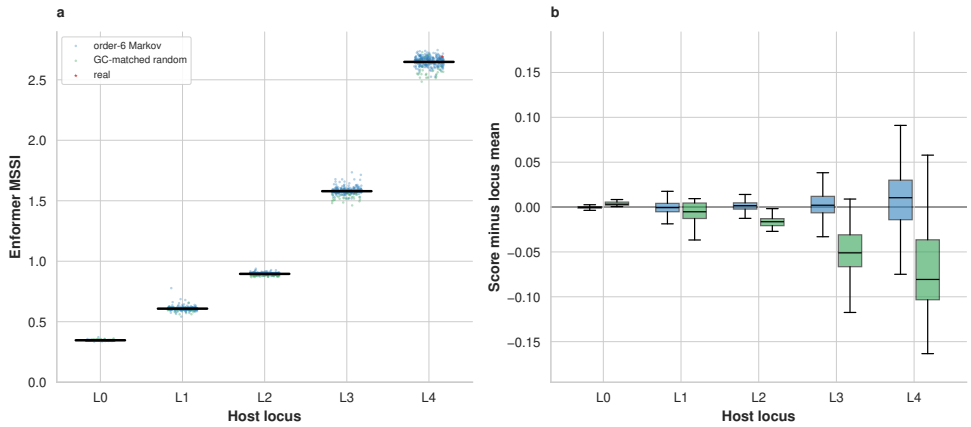


Fig. 4 Dominance of host genomic context over insert sequence. (a) Enformer MSSI for the same 331 inserts, coloured by insert type (order-6 Markov, GC-matched random, or the single real sequence), scored at each of the five candidate host loci; black lines mark locus means. **(b)** The same scores with each locus mean subtracted, split by insert type and drawn on its own scale (note the axis range is roughly 20-fold narrower than in a). The standard deviation across host-locus means (0.92) is 36 times the pooled within-locus standard deviation (0.025); in variance terms the host locus outweighs the insert by more than three orders of magnitude. Within that much smaller within-locus range, GC-matched random inserts trend below order-6 Markov inserts at the two highest-scoring loci (L3, L4), a secondary, insert-type-linked pattern that is itself locus-dependent and an order of magnitude smaller than the host-locus effect

score is therefore comparable across inserts placed at the same host locus, but is not comparable across loci: an MSSI value is informative about relative insert ranking at a fixed locus, not about the absolute regulatory potential of a sequence independent of where it is evaluated.

3.4 Selecting inserts by oracle score at one locus transfers to other loci

At a single selection locus we generated a pool of order-6-Markov candidate inserts per cell type, scored them by MSSI, and defined the selected set as the top-scoring inserts. Moving this identical selected set to four additional, previously unseen host loci and rescoring, the selected set retained a positive score advantage over an unselected control set at the new loci (+0.0351 pooled across the four transfer loci, $p = 1 \times 10^{-4}$), retaining 85% of the score advantage measured at the original selection locus. Selection therefore identifies something about the inserts that is not purely an artefact of the locus at which selection was performed, even though (Section 3.3) the oracle score itself is dominated by locus identity (Fig. 5).

3.5 What the oracle rewards: motif density and a candidate TF site are not causal

We tested whether the selection advantage was explained by motif density or by an individual transcription-factor site. Restricted to the 16-factor lineage panel, selected

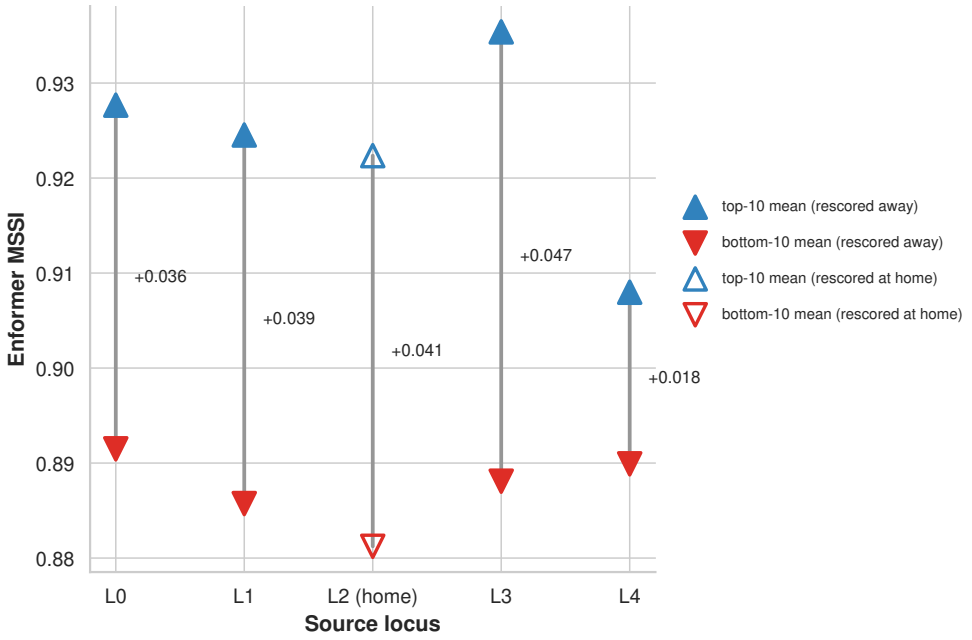


Fig. 5 Selection advantage transfers to unseen loci. The top-10 and bottom-10 candidates from each of the five selection loci were rescored at a locus where they had never been selected (L2 rescored at its own locus, hollow markers). Top-10 means exceed bottom-10 means at all five loci; pooled across the four away loci the advantage is +0.0351 ($p = 1 \times 10^{-4}$, positive at 4/4 loci), 85% of the +0.0411 advantage at the home locus

inserts had higher total motif hit counts (correlational contrast +3.28, $p = 3 \times 10^{-4}$) and higher RUNX1 hit counts specifically (+1.19, $p = 2 \times 10^{-4}$), both surviving Bonferroni correction across the panel. These correlations were not causal on intervention. An occlusion scan across the insert found no single tile whose removal accounted for the score (Fig. 6b); a direct RUNX1 ablation (removing predicted RUNX1 sites) left the score unchanged ($p = 0.33$); and a motif-density dose-response ablation that removed 3, 10, or 30 predicted sites (Fig. 6a) showed no effect at 3 or 30 removed sites and only a marginal effect at 10. Repeating the correlational test against the full JASPAR2024 CORE human collection Rauluseviciute et al (2024) (720 matrices, rather than the 16-factor lineage panel) reversed the sign of the total-hits correlation (−36.4, $p = 0.35$, no longer significant) and showed the direction of individual-matrix correlations tracked each matrix's own GC content ($r = +0.437$; Fig. 7) rather than any specific transcription-factor identity, indicating the 16-factor panel result in this subsection was an artefact of which factors happened to be included in that smaller panel rather than a motif-grammar signal. Repeating the density-ablation interventions at two of the 24 survey loci from Section 3.7 (below) reproduced this pattern only partially: density removal was causal with a monotonic dose-response at one locus (chr11, $q = 0.042$) and null at the other (chr13), so even the negative motif-density result is itself locus-dependent (Table 2).

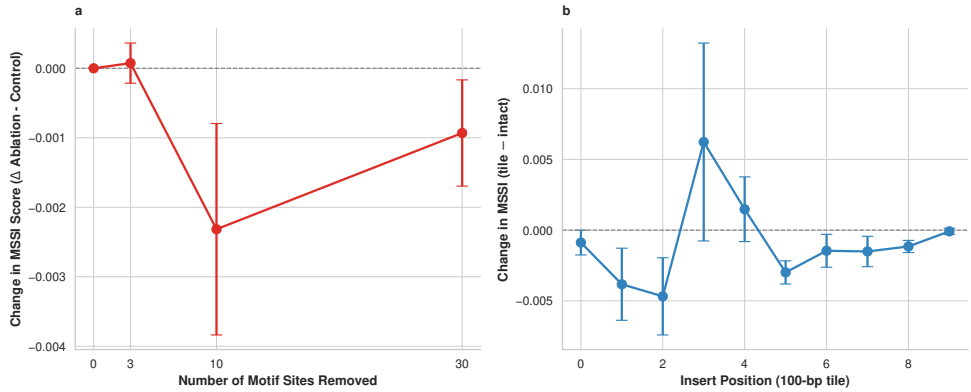


Fig. 6 Motif ablations at the selection locus. (a) Dose-response ablation removing 3, 10 and 30 predicted transcription-factor binding sites: near-zero changes in Enformer MSSI, with only a marginal effect at the 10-site dose. **(b)** Occlusion scan: each successive 100 bp tile of the insert replaced with random DNA; no single tile accounts for the score (mean ± SE over 10 candidates). Both nulls are themselves locus-dependent: repeating the density ablation at two survey loci produced a monotonic dose-response at one (chr11, $q = 0.042$) and a null at the other (chr13), so motif-density insensitivity is not a general property of the oracle (Table 2)

Table 2 Summary of tested sequence mechanisms and their locus-dependence. Mechanisms were evaluated using strictly controlled *in silico* interventions. Motif density was causal at only one of three tested loci, while CpG content showed a consistent causal effect whose sign was entirely determined by the host locus context.

Hypothesis	**Intervention / Test**	**Result**
Specific factor (RUNX1)	Site shuffling vs matched control (3 loci)	Null at all three
Bulk motif density	Dose-response 3/10/30 (3 loci)	Causal at chr11; null at chr4 and chr13
Localised element	100 bp occlusion tiles (3 loci)	Null everywhere; distributed effect
Oracle's sensitivity	Gradient × input attribution	Weak; rank test $p = 0.18$
CpG content	**Dinucleotide swap (3 loci)**	**Causal ($p = 10^{-4}$), locus-dependent sign**

3.6 CpG content is a causal axis, and its sign depends on the host locus

At the sequence level, CpG observed/expected was the single strongest discriminator between selected and unselected inserts (selected 0.364 vs. unselected 0.535, $p = 1 \times 10^{-4}$), exceeding every motif-based statistic tested. This correlation, however, reversed sign across loci: at one locus (chr4) CpG o/e correlated positively with score ($r = +0.480$), while at the other four candidate loci it correlated negatively ($r = -0.122$ to -0.590). We tested CpG content directly with a composition-preserving intervention that reverses `GC` and `CG` dinucleotides in place, changing CpG dinucleotide count while holding every individual base count in the sequence identical. Applied at three loci, this edit produced a significant locus × edit interaction ($p = 1 \times 10^{-4}$): the edit's effect on score was not just different in magnitude but opposite in sign depending on the host locus, and at two of the three loci the CpG edit alone accounted for 44–47% of the

Fig. 7 With an unbiased motif vocabulary, the density contrast is compositional. Per-matrix selected-minus-rejected hit-count differences across all 720 JASPAR2024 CORE human matrices versus each matrix's own GC content (fraction of positions whose consensus base is G or C). The per-matrix direction tracks matrix GC ($r = +0.437$, $p < 1 \times 10^{-32}$); among the 135 matrices significant at BH-FDR $q < 0.05$ (coloured), enriched matrices average 0.615 GC and depleted matrices 0.294 GC, so the apparent density signal of the 16-factor panel is a property of that panel, not of the sequences

total selection advantage measured in Section 3.4. CpG content is therefore a causal determinant of the oracle score, but its causal sign is a property of the combination of insert and host locus, not of the insert alone.

3.7 A 24-locus survey: heterogeneity is real, substantial, and not explained by any tested host property

To characterise this locus dependence directly, we placed an identical set of inserts and applied the identical composition-preserving CpG edit at 24 host loci genome-wide. The effect of the edit on oracle score varied far beyond what sampling noise predicts (Fig. 8a; $I^2 = 95\%$, Cochran's $Q = 458.6$, $p = 1.4 \times 10^{-82}$): 12 of the 24 loci showed an individually significant edit effect, 11 of those negative and one positive, and the remaining 12 were not individually distinguishable from no effect. We then tested 13 host locus properties as candidate predictors of the edit's direction and magnitude, spanning sequence composition measured from 1 kb windows up to Enformer's full 196,608 bp receptive field, and predicted chromatin state at the locus; none reached

significance ($p > 0.06$ for all 13). We report that the sign of an edit's effect on this oracle is not explained by any of the host features we were able to test cheaply, from local composition to genome-scale context (Table 3).

Table 3 Host locus properties tested as predictors of the CpG edit's effect sign. The effect of the identical sequence edit was measured at 24 independent $CD4^+$ T cell loci. None of the 13 tested host properties significantly correlated with the effect size after controlling for the baseline score, demonstrating that the locus-dependent sign reversal is not predicted by local composition or predicted chromatin state.

Host Property	**Spearman ρ (relative)**	***p*-value**
Window CpG o/e (1 kb)	+0.149	0.49
Context CpG o/e (10 kb)	+0.060	0.78
Context CpG o/e (50 kb)	−0.074	0.73
Context CpG o/e (196 kb)	−0.085	0.69
Context GC (10 kb)	−0.302	0.15
Baseline MSSI	−0.026	0.90
Peak score	−0.286	0.18
Predicted chromatin tracks (6 features)	≤ 0.27	≥ 0.17

3.8 Cross-oracle replication: instability replicates, the specific sign reversal does not

We repeated the locus-dominance and heterogeneity analyses using Borzoi, a second, independently trained sequence-to-function oracle with a different architecture, a larger 524,288 bp receptive field, and (for microglia specifically) a genuine microglia accessibility track rather than the $CD14^+$ monocyte proxy Enformer's human head provides. Enformer and Borzoi MSSI scores for the same inserts correlated at $r = 0.628$, and the selection-transfer effect from Section 3.4 replicated in Borzoi (+0.0027, $p = 1 \times 10^{-4}$). The locus-dependent heterogeneity from Section 3.7 also replicated: Borzoi showed substantial heterogeneity of its own across the same 24 loci ($I^2 = 65\%$, $p = 5.4 \times 10^{-6}$), and the magnitude of the per-locus edit effect correlated across the two oracles (Fig. 8b, $\rho = 0.502$, $p = 0.013$). The one specific sign reversal highlighted in Section 3.6 did not replicate: at that locus, Borzoi's edit effect was indistinguishable from zero ($q = 0.93$). Cross-oracle agreement on the existence and approximate magnitude of locus dependence, without agreement on which loci reverse sign, is consistent with locus dependence being a real property of the underlying sequence-function relationship that both oracles partially capture, while the precise per-locus direction remains oracle-specific and should not be over-interpreted from a single model.

3.9 The selection objective does not predict measured reporter activity

We calibrated the selection objective against an independent, orthogonal measurement: lentiMPRA reporter activity for 54,865 elements from the ENCODE library

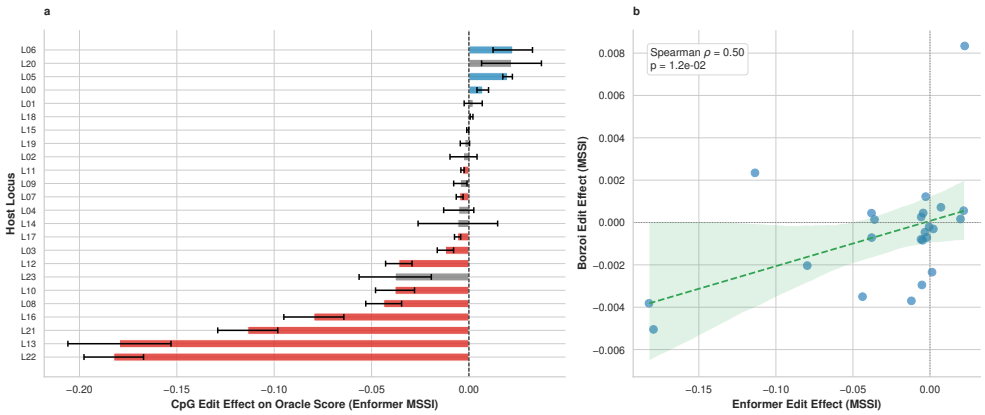


Fig. 8 Heterogeneity of edit effect across genomic loci and cross-oracle replication. (a) The effect size of a composition-preserving CpG edit on the Enformer oracle score (MSSI) at 24 different host loci. The effect varies drastically beyond sampling noise, with some loci showing a significant positive effect (blue) and others a significant negative effect (red). Error bars represent standard errors. **(b)** Cross-oracle replication of the edit effect magnitude between Enformer and Borzoi. While the precise per-locus direction does not perfectly match at every locus, the magnitude of the locus dependence correlates significantly between the two independent oracles

ENCSR604ZOZ, assayed in Jurkat cells (ENCSR247RPW) and K562 cells, of which 2,000 elements were scored with the same Enformer-based objective used for selection. The objective did not predict measured cell-type specificity between Jurkat and K562 (Fig. 9b; $\rho = -0.046$, $p = 0.14$), and elements at the extremes of the predicted-specificity distribution showed reversed measured specificity relative to prediction (Fig. 9c; $p = 8 \times 10^{-8}$). As a positive control, we checked whether the same objective, scored against Enformer's DNase:K562 tracks, correlated with K562's own measured reporter activity; it did not (Fig. 9a), and the correlation was in the wrong direction ($\rho = -0.208$, $p = 3 \times 10^{-11}$), strengthening to $\rho = +0.203$ only when recomputed against a deliberately mismatched CD4$^+$ T cell track index after adjusting for GC content. This calibration failure does not by itself establish that the oracle's chromatin-accessibility predictions are wrong; it does establish that the specific selection objective used in this study, scored at a single host locus, is not currently validated against an independent measurement of reporter activity, and this is the most serious limitation of the selection pipeline reported here (Section 4).

4 Discussion

The central result of this study is that a sequence-level effect learned by a genomic deep-learning oracle, and confirmed causal by direct intervention at the locus it was discovered at, does not generalise across genomic context in a predictable way. Placing an identical insert carrying an identical, composition-preserving edit at 24 different host loci produced an effect that varied almost fivefold beyond sampling noise ($I^2 = 95\%$) and reversed sign at some loci. Furthermore, none of the 13 host properties we tested, including sequence composition measured across scales from 1 kb

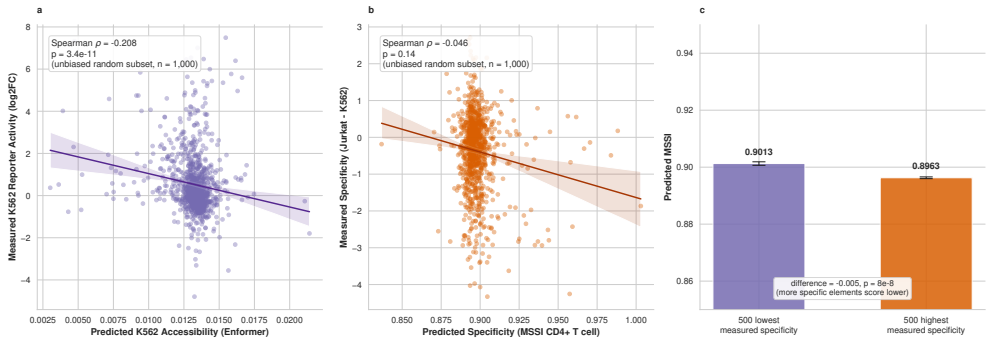


Fig. 9 Calibration of the selection objective against measured reporter activity. (a) Positive control checking the correlation between predicted K562 accessibility (Enformer) and measured K562 reporter activity ($\log_2$FC). The correlation is weak and goes in the wrong direction. **(b)** Measured cell-type specificity (Jurkat - K562) plotted against the predicted specificity (MSSI). The objective failed to predict the measured specificity in the independent episomal reporter assay. **(c)** The 500 elements with the highest measured T-cell specificity score lower on the predicted MSSI (0.8963) than the 500 lowest (0.9013; difference -0.005, $p = 8 \times 10^{-8}$): the extremes of measured specificity reverse the predicted ordering

to the oracle's entire 196 kb receptive field and predicted chromatin state, predicted which way it would go (Section 3.7). This matters directly for the growing practice of using sequence-to-function oracles to score and select designed regulatory elements (Section 1): a design chosen and validated *in silico* at one locus carries no guarantee, from this evidence, about its behaviour once moved or inserted elsewhere.

Two results in combination make this more than a single-locus curiosity. First, the causal edit itself was not a plausible artefact: CpG content, not motif density and not a specific transcription-factor site, was the property that survived direct intervention (Sections 3.5–3.6), and it did so consistently enough to account for close to half the total selection advantage at two of the three loci where it was tested directly. Second, the instability itself, though not the specific per-locus direction, replicated in an independently trained oracle with a different architecture and a larger receptive field (Section 3.8). Because Enformer and Borzoi were trained on overlapping public functional-genomics data, this cross-oracle agreement is necessary but not sufficient evidence that the heterogeneity reflects biology rather than a training-data bias shared by both models; it does, however, rule out heterogeneity being an artefact specific to one architecture or one set of weights.

The failure to predict measured lentiMPRA reporter activity (Section 3.9) is the most serious limitation of the selection pipeline as reported here, and we report it rather than omit it. A single-locus episomal reporter assay in transformed cell lines and native genomic chromatin accessibility are different quantities, so this result does not show that Enformer's accessibility predictions are wrong; the anti-correlation on the K562 positive control, however, shows that our specific selection objective, as scored at the single host locus used for design, is not currently validated against any independent measurement of regulatory activity. Anyone using an oracle score of this kind to make design decisions should treat it as an unvalidated, locus-specific ranking

signal rather than as a measurement of a portable regulatory property, until it has been checked against reporter data at more than one locus.

A secondary finding is that a sixth-order Markov chain, fitted by counting in seconds and requiring no GPU, matched or exceeded a trained autoregressive transformer on every generative property this task required, including the one cell-type-specific enrichment (SPI1 in microglia) that mattered for downstream selection (Section 3.2). This serves as a useful negative result for anyone building a similar pipeline: model capacity was not the bottleneck here, and a much simpler, non-neural baseline should be tried before a trained generator is assumed necessary.

Several limitations bound how far these conclusions extend. The cell-type-specific signal that motivated the exclusive-window restriction (Section 3.1) was recovered post hoc, after the pooled analysis returned nothing, and should be read as exploratory rather than confirmatory. The microglia arm of this study used Enformer's $CD14^+$ monocyte tracks as a proxy, because Enformer's human head has no microglia track; microglia are yolk-sac-derived and transcriptionally distinct from monocytes Ginhoux et al (2010), so any microglia-specific number from Enformer is a myeloid-lineage proxy, not a microglia measurement, and the one motif-arrangement effect we found in microglia (Section 3.5) is confounded by the fact that the underlying GSE307262 microglia are $CD34^+$-derived and engrafted rather than primary tissue. The 24-locus survey, while far larger than the three-locus pilot it followed, samples 24 of the many thousands of plausible host loci in the genome, and its host-feature predictors were chosen for being cheap to compute rather than for being an exhaustive account of what could distinguish a sign-reversing locus from a null one; a host feature that does predict the direction may still exist and simply was not among the 13 tested. No sequence in this study was tested in a wet-lab reporter assay generated for this project; the lentiMPRA comparison (Section 3.9) is the only calibration against measured activity, and it used a pre-existing library assayed at a different locus, in different cell types, from the ones this pipeline was designed for. The chromatin accessibility data (GSE202087, GSE307262) are live GEO deposits rather than a frozen snapshot, and GSE307262 in particular is preprint-linked and may be reprocessed by its depositors after the date reported here; an independent, later re-run of the released Nextflow/Docker pipeline against the same accessions reproduced window counts within 0.1% of those reported in Section 3.1 and reproduced every qualitative conclusion in Sections 3.5–3.7 (CpG causality and its sign-dependence on host locus, the absence of a motif-density effect, and the between-locus heterogeneity itself, all at comparable significance and effect size), so point estimates reported below should be read as accurate to that tolerance rather than as exactly reproducible bit-for-bit from the cited accessions alone. Finally, every result here concerns a research pipeline applied to genetic risk loci for multiple sclerosis for the purpose of studying oracle behaviour under intervention; nothing in this study establishes or is intended to establish therapeutic, diagnostic, or other clinical relevance for any of the sequences, loci, or scores reported.

Together, these results argue for a specific, low-cost addition to any pipeline that uses a sequence-to-function oracle to select designed regulatory sequence: test the selected effect by intervention at more than one host locus before trusting it, and

calibrate the selection objective against an independent measurement wherever one is available, before that objective is used to make a design decision.

Data availability. All chromatin accessibility data are public: GSE202087 Guo et al (2022) and GSE307262 Wang et al (2025). Measured reporter activity is from ENCODE experiments ENCSR247RPW (Jurkat) and the associated K562, HepG2 and SK-N-SH experiments sharing library ENCSR604ZOZ; the element sequences are ENCODE file ENCFF701QCG. Model weights are the public checkpoints named in the Materials and Methods. Both GEO series are live deposits, not a static snapshot bundled with this manuscript; see the Discussion limitations paragraph for what that means for exact reproducibility of the reported point estimates.

Code availability. Analysis code, configuration files and unit tests are at github.com/dralperenuysal/ms-enhancer; every threshold, track index and accession used here is declared in `configs/*.yaml` rather than in code. The full pipeline (locus retrieval, dataset construction, generation, oracle scoring, and every intervention in Sections 3.5–3.7) is also packaged as a Nextflow workflow (`main.nf`) with a prebuilt Docker image, dralperenuysal/ms-enhancer on Docker Hub (Singularity/Apptainer also supported by building from the same image), runnable end to end with a single command: `nextflow run dralperenuysal/ms-enhancer -profile docker`; see `docs/REPRODUCING.md` in the repository above. The same pipeline runs against any GEO accession in the same narrowPeak/HOMER peak format by substituting the accession in the command above; the repository also bundles a small synthetic example dataset in that identical format (`data/example_synthetic/`) so the pipeline mechanics can be exercised offline, without live GEO access, before pointing it at real data.

Acknowledgements. Computational resources for this study, including GPU acceleration, were provisioned through Google Cloud.

Software and figure preparation. All figures were prepared in Python with matplotlib, seaborn, NumPy, pandas and SciPy, developed in Visual Studio Code.

Funding. No specific grant from any funding agency in the public, commercial, or not-for-profit sectors was received for this work.

Author contributions. S.A.U. conceived and designed the study, performed all computational analyses, interpreted the results, and wrote the manuscript.

Competing interests. The author declares no competing interests.

Ethics approval and consent to participate. This study used only publicly available, previously published datasets (GEO, ENCODE, GWAS Catalog) and involved no human or animal participants, so no ethics approval was required. The work complies with the current laws of Türkiye, the country in which it was performed.

Consent for publication. Not applicable.

Use of AI and AI-assisted technologies. The large language model DeepSeek was used for copywriting support and for editorial revisions of the manuscript (wording and clarity). All content was reviewed and approved by the author, who takes full responsibility for the final version of the text.

References

Agarwal V, Inoue F, Schubach M, et al (2025) Massively parallel characterization of transcriptional regulatory elements. Nature 639(8054):411–420. https://doi.org/10.1038/s41586-024-08430-9

Avšec Ž, Agarwal V, Visentin D, et al (2021) Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods 18(10):1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Ginhoux F, Greter M, Leboeuf M, et al (2010) Fate mapping analysis reveals that adult microglia derive from primitive macrophages. Science 330(6005):841–845. https://doi.org/10.1126/science.1194637

Gosai SJ, Castro RI, Fuentes N, et al (2024) Machine-guided design of cell-type-targeting cis-regulatory elements. Nature 634(8036):1211–1220. https://doi.org/10.1038/s41586-024-08070-z

Guo MH, Sama P, LaBarre BA, et al (2022) Dissection of multiple sclerosis genetics identifies B and CD4+ T cells as driver cell subsets. Genome Biology 23:127. https://doi.org/10.1186/s13059-022-02694-y

Heinz S, Benner C, Spann N, et al (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell 38(4):576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Higgins JPT, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21(11):1539–1558. https://doi.org/10.1002/sim.1186

International Multiple Sclerosis Genetics Consortium (2019) Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science 365(6460):eaav7188. https://doi.org/10.1126/science.aav7188

Karollus A, Mauermeier T, Gagneur J (2023) Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biology 24:56. https://doi.org/10.1186/s13059-023-02899-9

Linder J, Srivastava D, Yuan H, et al (2025) Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. Nature Genetics 57:949–961. https://doi.org/10.1038/s41588-024-02053-6

Luo Y, Hitz BC, Gabdank I, et al (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Research 48(D1):D882–D889. https://doi.org/10.1093/nar/gkz1062

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

Sasse A, Ng B, Spiro AE, et al (2023) Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. Nature Genetics 55:2060–2064. https://doi.org/10.1038/s41588-023-01524-6

Taskiran II, Spanier KI, Dickmänken H, et al (2024) Cell-type-directed design of synthetic enhancers. Nature 626(7997):212–220. https://doi.org/10.1038/s41586-023-06936-2

Wang J, Warden A, Xia B, et al (2025) Global human myeloid replacement with peripheral progenitors induces interferonopathy and neurodegeneration. Research Square https://doi.org/10.21203/rs.3.rs-7762822/v1, preprint, not peer reviewed. Article ID rs.3.rs-7762822, version 1

Zhang Y, Liu T, Meyer CA, et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biology 9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137