# Agentic Supervision of Iterative Assay Design in High-Throughput Regulatory Genomics

**Arya S. Rao** [1 2] **Rodrigo I. Castro** [3] **Sager J. Gosai** [4] **Kenneth Bryan Hsu** [1] **Yasha Ektefaie** [1]
**Sangeeta Bhatia** [1 5 6 7 8] **Steven K. Reilly** [9] **Ryan Tewhey** [3] **Pardis C. Sabeti** [1 6 10 11]

## Abstract

Designing maximally informative experiments remains a central challenge in genomics. In regulatory genomics, massively parallel reporter assays (MPRAs) enable large-scale functional measurements but require carefully constructed sequence libraries to be maximally informative. We introduce an agent-based framework for MPRA library design in which a large language model autonomously constructs and improves MPRA libraries designed with the goal of training predictive models of regulatory activity. Candidate libraries are evaluated using a high-fidelity surrogate of MPRA measurements, enabling rapid closed-loop optimization under realistic experimental constraints in a simulated wet-lab/dry-lab loop. Across independent runs, agent-designed libraries outperform a set of human-designed strategies and yield consistent improvements in predictive performance. In addition, the agent recovers interpretable design principles. These results suggest that AI agents can contribute meaningfully to experimental design in regulatory genomics and provide a reusable *in silico* benchmark for evaluating AI co-scientist workflows.

[1]The Broad Institute of MIT and Harvard; Cambridge, MA 02142, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. [3]The Jackson Laboratory, Bar Harbor, ME 04609, USA. [4]Sutter Hill Ventures, Palo Alto, CA 94304, USA. [5]David H. Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA 02139, USA. [6]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA. [7]The Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, MA 02115, USA. [8]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA. [9]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA. [10]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [11]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. Correspondence to: Arya S. Rao <arao@broadinstitute.org>, Pardis C. Sabeti <pardis@broadinstitute.org>.

*Proceedings of the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

## 1. Introduction

A central question for AI in biology is whether agents can assume part of the supervisory role that human researchers play in experimental science. As AI scientists take on increasingly integrated roles in the scientific workflow (Lu et al., 2026; Musslick et al., 2024), we need methods to evaluate not just their abilities as analysts, but as experimentalists.

Massively parallel reporter assays (MPRAs) offer an optimal and well-defined setting for this evaluation. MPRAs have enabled large-scale functional measurement of hundreds of thousands of regulatory sequences, and while they were historically primarily used to test specific hypotheses (Kircher et al., 2019; Tewhey et al., 2016), they are increasingly also used to generate training data for predictive sequence-to-activity models (Gosai et al., 2024; Siraj et al., 2026; De Winter et al., 2025). As the field aims to predict regulatory activity in harder-to-assay conditions such as primary cell types where experimental budget is limited, library composition becomes a consequential modeling decision. This is exactly the kind of experimental judgment we should be testing in AI scientists.

We therefore ask whether a large language model (LLM) agent can drive large-scale experimental design in a simulated wet-lab/dry-lab loop (Figure 1). We introduce an agent-based framework for experimental design in which an LLM iteratively constructs and improves MPRA libraries designed with the goal of training predictive models of regulatory activity. Candidate libraries are evaluated using a high-fidelity surrogate model of wet-lab measurement that provides sequence-level activity labels for training of predictive models. The agent updates its strategy using only the resulting performance metrics, while operating in an open-ended computational environment with access to the internet and the ability to search for and download resources at will.

This work makes three primary contributions. First, we outline large-scale experimental design as an emerging capacity for frontier LLM agents. Second, we formulate MPRA library construction as an agentic experimental design problem, where the objective is to generate training data for

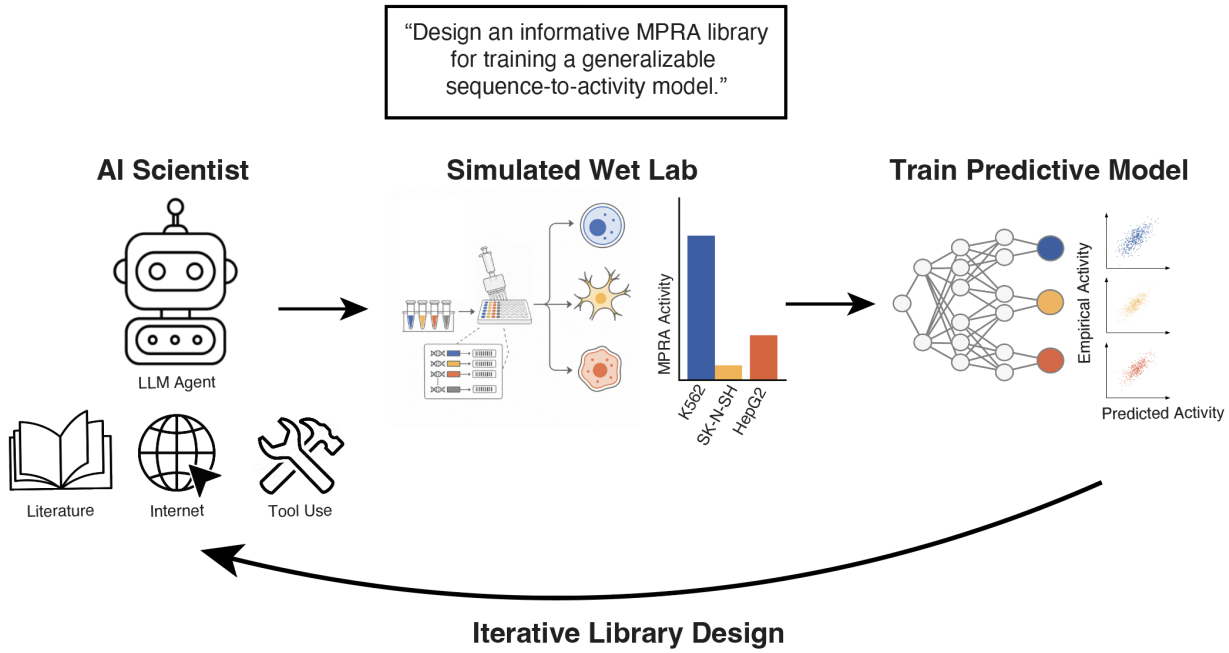


*Figure 1.* At each round, an LLM agent proposes a library of 50,000 DNA sequences, which are labeled by a high-fidelity surrogate model (simulated wet lab), used to train a fresh sequence-to-activity model, and evaluated against held-out test sets. The agent receives the returned performance metrics and proposes the next library.

generalizable sequence-to-activity models. We introduce **MPRAgent**, a reproducible *in silico* sandbox that preserves the iterative structure of wet-lab design, allowing many candidate libraries to be evaluated rapidly. Third, we show that language-model agents can exceed strong expert-designed baselines and recover interpretable design principles for regulatory genomics. Together, these results support a role for AI agents as contributors to experimental design, not only as tools for downstream computational analysis.

## 2. Related Work

Recent agentic research systems have mostly been demonstrated in computational settings, where hypotheses can be tested by running code and evaluated with immediate quantitative feedback. Frameworks such as autoresearch (Karpathy, Andrej), AutoRA (Musslick et al., 2024), and The AI Scientist (Lu et al., 2026) automate parts of the research loop—including experiment selection, code execution, and iterative refinement—but their primary domain remains computational research questions. Rather than optimizing a fixed training script or proposing purely computational follow-up experiments, we task an agent to design an MPRA sequence library whose value is measured by the informativeness of the resulting assay data for downstream predictive model evaluation as a proxy for wet-lab validation. To our knowledge, this represents a substantial departure from the previous paradigm of agents performing parameter optimization to one where agents are responsible for optimizing experimental design.

Prior closed-loop biological design approaches have typically relied on Bayesian optimization (Belanger et al., 2019), active learning (Yang et al., 2025), or directed evolution (Sellés Vidal et al., 2023) rather than open-ended language model agents to generate candidates for iterative experimentation. These methods are usually framed around optimizing individual sequences or variants for a target property, with the next round strongly conditioned on prior results—limiting the diversity of designs the loop can explore. Active learning has also been applied to improve experimental outcomes in MPRA assays for specific regulatory elements (Friedman et al., 2025; Yin et al., 2025). Most closely related to our work, BioDiscoveryAgent (Roohani et al., 2024) uses an LLM agent in a closed loop to select genes for perturbation experiments.

To our knowledge, this is the first application of LLM agents to large-scale biological assay design. Where prior closed-loop approaches select candidates from a fixed and enumerable set, our agent must compose an entire experimental library of 50,000 sequences per round from a combinatorially vast space of $4^{200}$ possible sequences, a qualitatively different regime. This reframes the problem from local candidate optimization to experimental design over a large and open-ended search space.

Finally, our work builds on the substantial literature on MPRA library construction and model-guided regulatory sequence design. MPRA has been used for high-throughput dissection of regulatory grammar, including saturation mutagenesis of promoters and enhancers and functional testing of disease- and GWAS-associated noncoding variants (Kircher et al., 2019; Hua, 2023). Library construction across these efforts is human-designed, broadly sampling from regulatory element databases, sequence perturbations, or randomly drawn sequences. Resources such as MPRAbase and MPRAVarDB (Zhao et al., 2023; Jin et al., 2024) illustrate how broadly these design patterns are now used across cell types and applications. More recently, MPRA-trained neural models have been used prospectively to engineer synthetic cis-regulatory elements; in particular, Malinois (Gosai et al., 2024) was introduced as a sequence-to-activity model trained on MPRA data and used to design cell-type-specific CREs, and has since been widely adopted as an oracle model and benchmark for predicting CRE behavior, including in the commercial sector (Butts et al., 2025; DaSilva et al., 2026; Han et al., 2025; Nazaretyan et al., 2025; Reddy et al., 2024; Origin Bio, 2026; Exonic, 2026). We build on this model-guided design paradigm by using Malinois as the surrogate model inside an iterative agentic loop, optimizing the composition of the library itself rather than the activity of individual sequences.

## 3. Methods

### 3.1. Task

We evaluate whether an LLM agent can iteratively design MPRA libraries that improve downstream predictive models of regulatory activity. At each round $t$, the agent writes a program that produces a library $L_t$ of 50,000 DNA sequences (200 bp, alphabet $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$). The purpose of the library is to generate maximally informative data for training predictive sequence-to-activity models across any arbitrary cell type.

To enable rapid iteration over library designs, we simulate the MPRA measurement step with Malinois, a convolutional neural network-based predictor of episomal MPRA activity. Malinois was trained on 776,474 reporter sequences measured in K562, HepG2, and SK-N-SH, and achieves Pearson $r = 0.88$ against empirical measurements on held-out chromosomes. Malinois has been widely adopted as an *in silico* surrogate for MPRA activity, both for variant effect prediction and as a fitness function for generative regulatory sequence design. Given a 200 bp sequence, Malinois returns predicted regulatory activity for each cell type; these predictions serve as labels for each proposed library. The agent therefore faces the same design problem as in a physical MPRA (which sequences to include in order to train the most informative downstream model) but the measurement step runs *in silico*.

Each submitted library is labeled using Malinois and then used to train a fresh sequence-to-activity model. All models share a fixed architecture and hyperparameters (identical to Malinois), are initialized from random weights, and trained exclusively on the 50,000 library sequences and their oracle-predicted activity labels.

For a library $L_t$, let $f_{L_t}$ denote the model trained on that library and its assigned labels. Let $x_{\text{test}}$ denote the held-out evaluation sequences and $y^c_{\text{test}}$ the corresponding empirical MPRA measurements for cell type $c$. Performance is defined as:

$$\text{score}(L_t) = \frac{1}{3} \sum_c r\big(f_{L_t}(x_{\text{test}}),\ y^c_{\text{test}}\big), \tag{1}$$

the Pearson correlation $r$ between the trained model's predictions $f_{L_t}(x_{\text{test}})$ and the held-out labels $y^c_{\text{test}}$, averaged over the three cell types.

This task is sequential: after each round, the agent observes the evaluation results and proposes the next library. The labeling, training, and evaluation pipeline is sealed and contained in a `prepare.py` file; the agent is instructed to treat `prepare.py` as a wet-lab collaborator. It is able to submit proposed libraries and receive the evaluation metrics for models trained on those libraries, but cannot inspect any component of the process.

The agent receives a natural-language task specification via `instructions.md` describing the task at hand and the rules of interaction. The specification frames K562, HepG2, and SK-N-SH as measurement constraints rather than the ultimate biological target: before each experiment, the agent must justify why its proposed library should remain informative for models evaluated in unmeasured regulatory contexts. Importantly, the agent is instructed not only to propose new libraries for training, but to develop and iterate upon a theory for what makes a training library informative for this task.

### 3.2. Agent

We deploy Claude Opus 4.7 (Anthropic), accessed via the Vertex AI API, in a sandboxed Linux environment hosted on an NVIDIA DGX Spark, with shell access, Python, internet access, and a writable filesystem. We structure the run as an autonomous experimental loop with persistent state. Each run lasts 30 rounds. Before each round, the agent rereads its prior results and written notes, states the hypothesis motivating the next library, searches for relevant external evidence, and writes a planning entry. It then implements the next library selection strategy in code, submits the resulting libraries for evaluation via `prepare.py`, records the returned metrics, and updates its interpretation before proceeding to the next round.

The agent maintains two forms of memory. First, an append-only lab notebook records hypotheses, plans, results, and theory updates before and after each experiment. Second, a `skills/` directory stores editable notes on reusable datasets, procedures, code patterns, and empirical observations. Together with the generated code, submitted sequences, and returned metrics, these files provide an auditable record of the agent's experimental trajectory.

Isolation of the evaluation pipeline is enforced through filesystem constraints: `prepare.py` and all pipeline dependencies, including the Malinois model weights and evaluation set sequences, reside outside the agent's writable working directory and are not readable by the agent process. The agent can submit a library file and receive back aggregate metrics, but cannot inspect any component of the labeling, training, or evaluation pipeline. The agent is additionally instructed to treat `prepare.py` as a wet-lab collaborator—a black-box process that accepts sequence libraries and returns performance scores—and to not attempt to reverse-engineer or replicate it.

We evaluate two information conditions using independent 30-round runs. In the **blind** condition, the agent receives only `instructions.md` and the black-box submission interface `prepare.py`. In the **informed** condition, the agent additionally receives `strategies.md`, which contains a set of human-designed libraries, together with their evaluation results. We describe these expert-designed baselines below.

### 3.3. Human-designed baselines

We compare the agent against 14 human-designed library strategies produced by a domain expert in regulatory genomics. These strategies were evaluated with the same labeling, training, and evaluation pipeline. Each strategy defines a sampling procedure rather than a fixed library. For each strategy, we generated five independent 50,000-sequence libraries using different random seeds, so that replicates differed only in the specific sequences sampled under the same design rule. The 14 strategies are drawn from four sequence source families, each selected to cover a different aspect of human regulatory biology, and vary in how sequences are sampled within and across sources.

**DHS index.** The first source is the DNase I Hypersensitive Sites (DHS) index, a genome-wide catalog of approximately 3.6 million sites where chromatin is physically accessible to regulatory proteins, mapped across 733 human biosamples (Meuleman et al., 2020). Accessible chromatin is a strong marker of active or poised regulatory function: most enhancers, promoters, and insulator elements reside in DHS regions. Meuleman et al. applied non-negative matrix factorization (NMF) to the accessibility profiles of these sites across biosamples, identifying 16 components ("topics") that capture distinct tissue-level regulatory programs—for example, one topic groups elements accessible primarily in immune cells, another in hepatocytes. The DHS pool is heavily imbalanced: a few ubiquitous topics (e.g., constitutively accessible promoter-proximal elements) contain the majority of sites, while topics capturing tissue-restricted regulatory programs are much smaller. To ensure that rare regulatory programs are represented in the library, we test three sampling schemes: topic-weighted sampling, where elements are drawn with probability proportional to their NMF topic loadings, upweighting underrepresented topics; topic-stratified sampling, where sequences are assigned to their maximum topic and an equal number of sequences is drawn from each of the 16 topics ($n_k = n/16$ for each topic $k$), forcing uniform representation regardless of topic size; and uniform random sampling, where every element is equally likely.

**Sei sequence classes.** The second source is Sei, a deep learning model that assigns functional annotations to genomic positions based on chromatin and transcription factor binding profiles (Chen et al., 2022). Sei classifies the genome into regulatory sequence classes spanning enhancers, promoters, CTCF (an insulator protein) binding sites, polycomb-repressed regions, transcribed regions, and other functional categories, covering approximately 3 million genomic regions. Where DHS captures accessibility—whether a region is physically open—Sei captures function: what kind of regulatory role a region plays. As with DHS topics, the class distribution is highly imbalanced; a few common classes (e.g., heterochromatin, quiescent regions) contain far more sequence than rare functional classes (e.g., strong tissue-specific enhancers). We test two sampling schemes: class-proportional, where regions are drawn with probability proportional to their class frequency, and class-balanced, where an equal number of sequences is drawn from each of the 40 classes ($n_j = n/40$ for each class $j$).

**Synthetic random DNA.** The third source is fully synthetic sequence. Each of the 50,000 sequences is generated by drawing each of its 200 nucleotide positions independently from a uniform distribution over $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. These sequences help establish how much a model can learn from pure sequence diversity alone.

**Published MPRA sequences.** The fourth source is the Malinois training set itself. Sequences are sampled uniformly at random from the training set. We test two label conditions: one where sampled sequences are relabeled by the Malinois oracle, and one where they retain their original empirical MPRA measurements. This pair isolates the effect of label noise at fixed sequence composition.

*Table 1.* Human-designed baseline strategies. All libraries contain 50,000 sequences and are evaluated across 5 random seeds.

| Strategy | Description |
|---|---|
| *Single-source libraries* | |
| **dhs_topic** | DHS elements, sampled proportional to NMF topic loading |
| **dhs_random** | DHS elements, sampled uniformly at random |
| **dhs_stratified** | DHS elements, equal allocation across 16 NMF topics |
| **sei_class** | Sei regions, sampled proportional to class frequency |
| **sei_random** | Sei regions, sampled uniformly at random |
| **synth_oracle** | Each nucleotide drawn independently and uniformly at random from {A,C,G,T} |
| **mpra_oracle** | Malinois training set sampled uniformly at random; oracle labels |
| **mpra_real** | Malinois training set sampled uniformly at random; empirical labels |
| *Two-source mixtures (50/50)* | |
| **dhs_sei** | DHS topic-weighted + Sei class-proportional |
| **dhs_synth** | DHS topic-weighted + synthetic |
| **dhs_strat_sei** | DHS topic-stratified + Sei class-balanced |
| **sei_synth** | Sei class-balanced + synthetic |
| *Three-source mixtures (equal thirds)* | |
| **dhs_sei_synth** | DHS topic-weighted + Sei class-proportional + synthetic |
| **dhs_strat_sei_synth** | DHS topic-stratified + Sei class-balanced + synthetic |

The 14 strategies consist of single-source libraries under each sampling scheme and mixtures combining two or three sources at fixed ratios. Strategies vary both in source composition and in within-source sampling—for example, a DHS+Sei mixture can pair topic-weighted DHS with class-proportional Sei, or topic-stratified DHS with class-balanced Sei. Full definitions are given in Table 1.

### 3.4. Evaluation

Because our goal is to design libraries that train broadly useful regulatory models, evaluation is not restricted to a single held-out test set. Each trained model is evaluated on 14 held-out test sets derived from 9 source distributions Table S1. Five sets use empirical MPRA measurements as labels; nine use Malinois-generated labels. The test sets include held-out reporter elements from the original Gosai et al. MPRA experiment, UK Biobank and GTEx variant sequences, DNase I hypersensitive sites from the DHS index, Sei regulatory regions, random genomic windows, and fully synthetic random sequences. All genomic evaluation sets are drawn from chromosomes excluded from Malinois training (7, 13, 19, 21, X). The empirical evaluation sets guard against oracle-specific biases: a library that exploits regularities in Malinois predictions without capturing real regulatory biology will score poorly on empirical labels. The oracle-labeled sets provide additional coverage of sequence distributions not represented in the experimental data.

The agent sees only anonymous identifiers and aggregate metrics for these evaluation sets. It does not observe their identities, sequences, labels, or genomic coordinates. The primary metric is empirical performance on the Malinois held-out test set, shown to the agent only as `eval_01`; the agent is also encouraged to maintain strong performance across all evaluation sets.

We release the full **MPRAgent** sandbox, including the evaluation pipeline, baseline strategy implementations, and agent scaffolding, at https://anonymous.4open.science/r/MPRAgent/.

## 4. Results

### 4.1. Human-designed baselines reveal substantial variation in library informativeness

Library composition strongly determines downstream model quality. Across 14 expert-designed strategies at a fixed budget of 50,000 sequences, performance on the primary evaluation metric (`eval_01`, Pearson $r$ against empirical MPRA measurements) ranged from $0.603$ for **mpra_real**—which samples sequences from a prior MPRA experiment with empirical labels—to $0.723$ for **dhs_topic**, which samples DHS elements proportional to NMF topic weights (Figure 2). This $0.121$-point spread substantially exceeded the variation observed across five random seeds within any single strategy, indicating that library composition rather than sampling stochasticity drives performance differences. Strategy rankings were broadly consistent across all 14 evaluation sets, indicating that the effect of library composition is not specific to `eval_01` alone.

The dominant axis of variation is regulatory sequence content. DHS-based strategies occupy the top five positions on `eval_01`, while strategies relying primarily on Sei chromatin state regions (**sei_class** $= 0.659$; **sei_random** $= 0.645$) or on existing MPRA sequences with empirical labels (**mpra_real** $= 0.603$) occupy the bottom. Within DHS-based strategies, mixing in additional sequence sources produces modest and inconsistent changes relative to topic-weighted DHS sampling alone, suggesting that annotation quality is a stronger lever than source diversity at this library size.

Performance improves substantially with library size across all strategies (Figure S1), with the steepest gains between 10,000 and 150,000 sequences and diminishing returns thereafter. Relative rankings are largely preserved across library sizes, justifying 50,000 sequences as the primary compari-

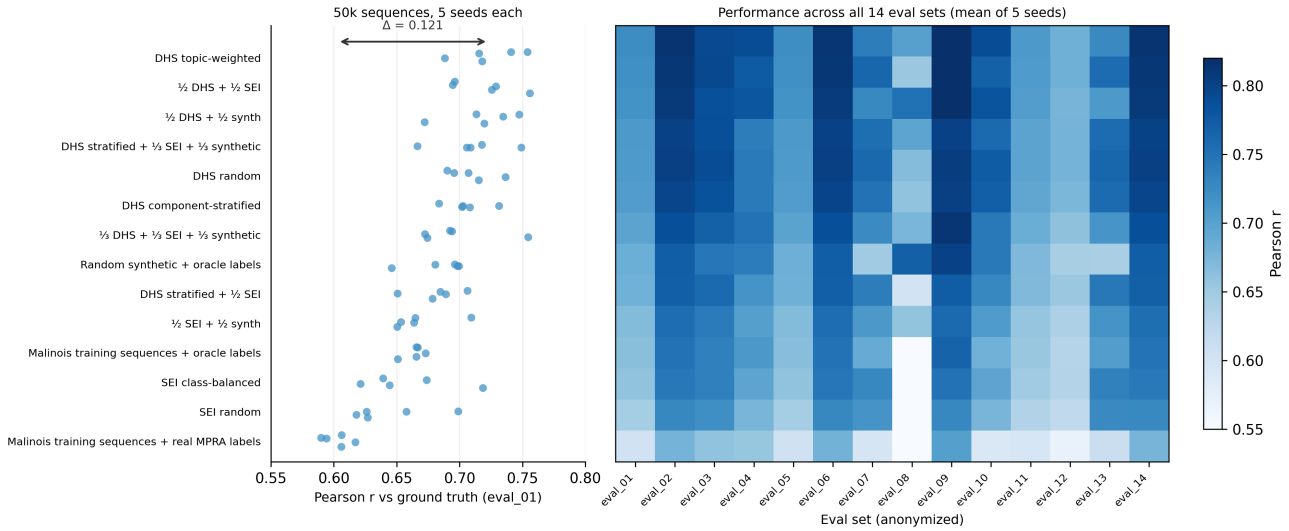


*Figure 2.* Library composition strongly determines downstream model quality. *(Left)* `eval_01` Pearson $r$ for all 14 human-designed strategies at 50,000 sequences (5 seeds each), spanning a range of $\Delta = 0.121$. *(Right)* Performance across all 14 anonymized evaluation sets (mean of 5 seeds), showing that strategy rankings are broadly consistent beyond the primary metric.

son point: composition differences are largest in absolute terms in this regime, and the best human design at 50,000 sequences (**dhs_topic**, `eval_01` $= 0.723$) is far from saturated, reaching $r \approx 0.845$ at 300,000 sequences. The agent comparison therefore operates in a regime where design choices have meaningful consequences.

### 4.2. Agent-designed libraries outperform expert designs

All four agent runs produced libraries that exceeded the best human-designed strategy on the primary evaluation metric within the 30-round budget (Figure 3). The strongest blind run reached an `eval_01` Pearson $r$ of $0.748$, improving over the best human baseline by $0.025$. The strongest informed run reached $r = 0.760$, improving over the best human baseline by $0.037$. These gains were achieved by novel library designs outside the space of expert-evaluated strategies.

The improvements were broadly distributed across the evaluation suite rather than being driven by a single test set (Figure 4). For each of the 14 evaluation sets, we computed the best score achieved by any human-designed strategy as a per-set ceiling. The best designs from blind run 2 and both informed runs exceeded this ceiling on 13 of 14 evaluation sets; blind run 1 exceeded it on 11 of 14. For the top informed design, improvements over the per-evaluation human best ranged from $+0.016$ on `eval_09` to $+0.047$ on `eval_03`, with a mean improvement of $+0.035$ across the 13 sets on which it outperformed the human ceiling. Agent-designed libraries thus improved general downstream model performance rather than selectively optimizing the primary metric.

The single exception was `eval_08`, on which all four agent designs fell below the human best of $r = 0.770$, held by **synth_oracle**. Working from anonymous evaluation metrics alone, both blind and informed agents identified `eval_08` as rewarding coverage of random sequence space rather than regulatory content, and explicitly deprioritized it in favor of stronger performance across the remaining 13 evaluation sets. This tradeoff is not unique to the agents: every DHS-based human strategy also underperforms **synth_oracle** on `eval_08`, with the best DHS-containing design reaching only $r = 0.752$ by mixing in 50% synthetic content. Libraries enriched for curated regulatory elements, whether designed by humans or agents, consistently sacrifice `eval_08` performance for broader regulatory generalization.

### 4.3. Agents discover interpretable design principles for MPRA library design

All four agents identified distinct, interpretable design principles for what makes a training library informative for generalizing sequence-to-activity models. Across 120 total experiments, agents explored a broad range of design axes: annotation source selection, class-level sampling proportions, mixing ratios, sequence-composition stratification, augmentation strategies, and cross-species sequence inclusion, among others. These experiments collectively mapped the design space in ways that informed the principles the agents ultimately declared.

The blind and informed agents took qualitatively different approaches. Both blind runs began from random sequences and worked upward, independently converging on cCRE class composition as the primary lever—one through class

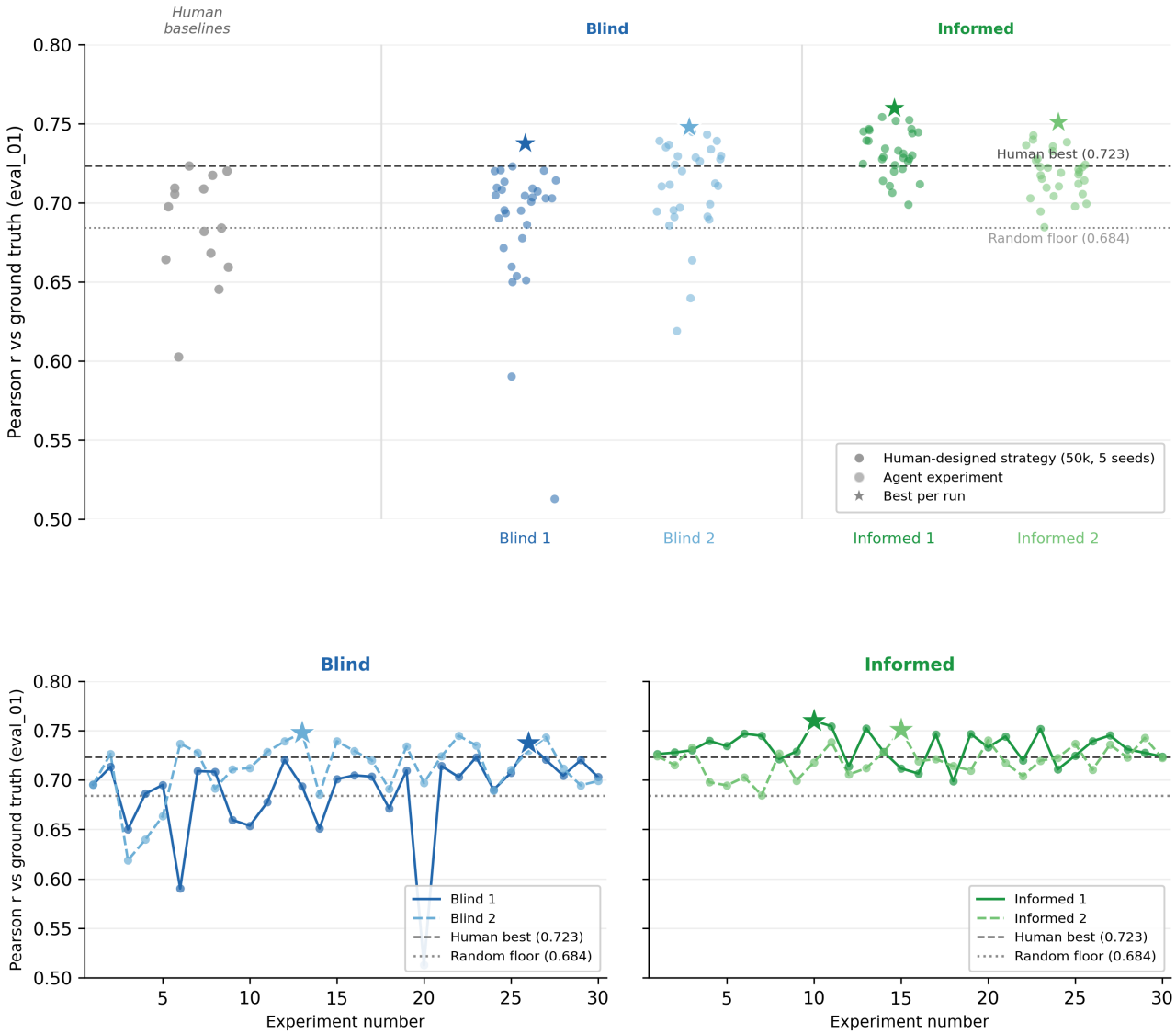


*Figure 3.* Agent performance on `eval_01` across 30 experimental rounds. *(Top)* Distribution of `eval_01` scores for all agent experiments alongside human-designed strategies; stars indicate the best design per run. *(Bottom)* Per-round learning curves for blind and informed runs. The human best (dashed) and random floor (dotted) are shown for reference. All four runs exceed the human best within the 30-round budget.

combination (proximal enhancer-like elements with CA-H3K4me3, `eval_01` $= 0.738$), the other through rare-class upweighting (`eval_01` $= 0.748$). The solutions are distinct but reflect the same underlying finding: structural diversity across cCRE classes is more informative than depth within any single class.

The informed agents entered with knowledge of which DHS-based strategies had already been evaluated and how they performed. Rather than refining within that space, both chose orthogonal directions. Informed run 1 moved toward annotation systems and sequence sources absent from the human baselines entirely, discovering that cross-species genomic sequences improve generalization. From anonymous evaluation metrics alone, the agent recovered an evolutionary-distance pattern: chicken regulatory sequences performed best ($r = 0.760$) followed by zebrafish ($r = 0.754$), mouse ($r = 0.747$), and platypus ($r = 0.739$). This result suggests that evolutionary distance can provide a useful axis for selecting out-of-distribution regulatory patterns that improve model generalization. Informed run 2 stayed within the DHS index but asked a different question about how to sample it, finding that sequence-composition stratification by GC content is an independent diversity axis unaddressed by any human baseline, contributing the

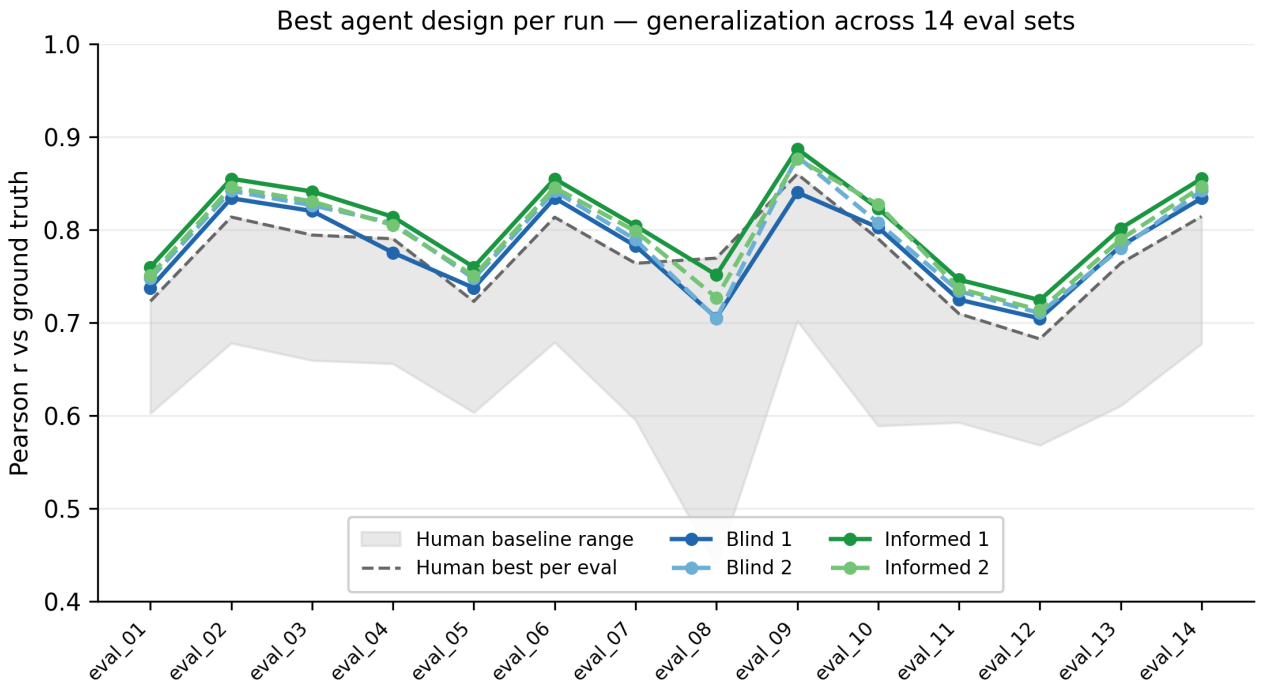


*Figure 4.* Generalization of the best agent design per run across all 14 evaluation sets. The gray band spans the full range of human-designed baseline performance; the dashed line marks the per-set human best. Agent designs exceed the human best on 13 of 14 evaluation sets, with `eval_08` (random synthetic sequences) as the consistent exception.

largest single-experiment gain in the run ($+0.013$, reaching `eval_01` $= 0.751$). Partial maps of the design space appear to act as scaffolds for agentic discovery, allowing agents to move beyond known high-performing strategies toward new and complementary design axes.

Without being instructed to do so, each agent adopted a common two-phase experimental structure: open-ended exploration followed by systematic ablation. Once a strong design was identified, agents held the recipe fixed and varied one component at a time to establish which elements were load-bearing, producing a mechanistic account of why the design works rather than simply confirming that it does. Across all four runs, the strategies that emerged—cCRE class combination, rare-class upweighting, cross-species augmentation, and GC-stratified dual-axis DHS sampling—provide not just high-performing libraries but falsifiable claims about the structure of the regulatory sequence space most informative for training generalizable sequence-to-activity models from MPRA data.

## 5. Discussion

Designing maximally informative experiments is one of the core tasks of biological research, yet it remains largely unstudied as an optimization problem within the realm of AI for science. We show that a language model agent can take on this role in a realistic setting, improving on expert-designed MPRA libraries within a fixed budget and identifying design principles that were not present in the human baseline set. **MPRAgent** makes this capacity measurable and reproducible, and we release it as a resource for evaluating agentic experimental design more broadly.

The principles the agents discovered are each interpretable and directly actionable for future library design, and their experimental trajectories are informative about the design space of MPRA library design itself. Rare cCRE class upweighting was independently recovered by both blind runs from random initialization, suggesting that its informativeness advantage is strong enough to emerge from black-box feedback alone. The cross-species result points to evolutionary distance as a useful axis for selecting out-of-distribution regulatory sequence variation, especially because it was recovered from anonymous metrics without access to biological annotation of the evaluation sets. GC stratification corrects a compositional bias in annotation-based sampling that no human-designed strategy in our baseline set had addressed. The two informed runs found different principles despite starting from the same baselines, suggesting that the design space has multiple qualitatively different near-optimal peaks, supporting a broader principle that deploying multiple independent agents on a given task may uncover diverse solutions.

Several caveats are worth noting. The task we study, designing a library that trains models generalizing to unseen cell types, is inherently harder to evaluate than optimizing model performance on a fixed labeled set. We address this by testing trained models with 14 diverse evaluation sets. Nevertheless, the suite is finite and the oracle was trained on three cell types, so libraries that score well here may not fully capture the regulatory grammar relevant to all possible deployment contexts. Wet-lab validation of the winning designs across novel cell types remains the definitive test. Second, in this task, we hold model architecture fixed: all designs are evaluated by training the same model class from scratch. Whether optimal library composition changes under different architectures, or whether architecture and training data should be co-optimized, is an open question this study does not address. Lastly, the human-designed baselines were constructed as a fixed comparison point, not as the result of iterative refinement; the agents' advantage lies not in discovering principles that are in principle inaccessible to human experts, but in recovering them systematically from black-box feedback within a constrained budget.

As autonomous AI co-scientists become more capable, they must be able to intelligently design wet-lab experiments. Tools for evaluating this capacity are therefore increasingly important. By improving on expert-designed MPRA libraries and recovering interpretable design principles from limited black-box feedback, the agents studied here suggest an early form of experimental judgment that will be essential for more autonomous scientific systems.

## Software and Data

All code and data associated with this project are provided at: https://anonymous.4open.science/r/MPRAgent/

## Impact Statement

This paper presents work whose goal is to advance the capacity of AI agents to perform experimental design in biological research. Improved methods for designing informative assays could accelerate the development of regulatory sequence models with relevance to gene therapy and precision medicine. The agent operates in a sandboxed computational environment and does not direct synthesis or deployment of genetic material; sequences proposed remain subject to standard institutional oversight. The in silico benchmark we release provides a reproducible evaluation resource for the broader AI for science community.

## References

Belanger, D., Vora, S., Mariet, Z., Deshpande, R., Dohan, D., Angermueller, C., Murphy, K., Chapelle, O., and Colwell, L. Biological sequences design using batched Bayesian optimization. December 2019. Accessed: 2026-4-29.

Butts, J., Rong, S., Gosai, S., Castro, R., Noon, M., Adeniran, K., Ghosh, R., Sabeti, P., Tewhey, R., and Reilly, S. Identifying non-coding variant effects at scale via machine learning models of cis-regulatory reporter assays. *bioRxiv*, pp. 2025.04.16.648420, April 2025.

Chen, K. M., Wong, A. K., Troyanskaya, O. G., and Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.*, 54(7):940–949, July 2022.

DaSilva, L. F., Senan, S., Kribelbauer-Swietek, J. F., Patel, Z. M., Louis, L. K., Reddy, A. J., Gabbita, S., Rosen, J. D., Nussbaum, Z., Córdova, C. M. V., Wenteler, A., Weber, N., Tunjic, T. M., Mansoldo, M., Khan, T. A., Hwang, G.-H., Gardeux, V., Humphreys, D. T., Smith, C., Bejan, M., Bromley, P., Connell, W., Deplancke, B., Love, M. I., Wong, E. S., Meuleman, W., and Pinello, L. Designing synthetic regulatory elements using the generative AI framework DNA-diffusion. *Nat. Genet.*, 58 (1):180–194, January 2026.

De Winter, S., Konstantakos, V., and Aerts, S. Modelling and design of transcriptional enhancers. *Nat. Rev. Bioeng.*, 3(5):374–389, February 2025.

Exonic. Exonic – AI for precision gene therapy. https://exonic.ai/, 2026. Accessed: 2026-4-29.

Friedman, R. Z., Ramu, A., Lichtarge, S., Wu, Y., Tripp, L., Lyon, D., Myers, C. A., Granas, D. M., Gause, M., Corbo, J. C., Cohen, B. A., and White, M. A. Active learning of enhancers and silencers in the developing neural retina. *Cell Syst.*, 16(1):101163, January 2025.

Gosai, S. J., Castro, R. I., Fuentes, N., Butts, J. C., Mouri, K., Alasoadura, M., Kales, S., Nguyen, T. T. L., Noche, R. R., Rao, A. S., Joy, M. T., Sabeti, P. C., Reilly, S. K., and Tewhey, R. Machine-guided design of cell-type-targeting cis-regulatory elements. *Nature*, 634(8036):1211–1220, October 2024.

Han, X., Wu, H., Wang, X., Liu, D., Fu, Y., Yang, L., Wang, R., Zhang, P., Wang, J., Ma, L., Mao, J., Zhou, L., Wang, S., Zhang, X., Jiang, M., Wang, X., Wen, G., Jia, D., and Guo, G. Modeling the vertebrate regulatory sequence landscape by UUATAC-seq and deep learning. *Cell*, 188 (19):5343–5362.e29, September 2025.

Hua, H. From GWAS to single-cell MPRA. *Nat. Methods*, 20(3):349, March 2023.

Jin, W., Xia, Y., Nizomov, J., Liu, Y., Li, Z., Lu, Q., and Chen, L. MPRAVarDB: an online database and web server for exploring regulatory effects of genetic variants. *Bioinformatics*, 40(10):btae578, October 2024.

Karpathy, Andrej. autoresearch: AI agents running research on single-GPU nanochat training automatically.

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Costello, J. F., Shendure, J., and Ahituv, N. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.*, 10(1):3583, August 2019.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D., and Clune, J. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, March 2026.

Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Sandstrom, R., Vierstra, J., Kaul, R., and Stamatoyannopoulos, J. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, 584(7820):244–251, August 2020.

Musslick, S., Andrew, B., Williams, C. C., Hewson, J. T. S., Li, S., Marinescu, I., Dubova, M., Dang, G. T., Strittmatter, Y., and Holland, J. G. AutoRA: Automated research assistant for closed-loop empirical research. *J. Open Source Softw.*, 9(104):6839, December 2024.

Nazaretyan, L., Rentzsch, P., and Kircher, M. varCADD: large sets of standing genetic variation enable genome-wide pathogenicity prediction. *Genome Med.*, 17(1):84, August 2025.

Origin Bio. 10,000 AI-designed regulatory DNA sequences, open for research. https://origin.bio/blogs/switch/, 2026. Accessed: 2026-4-29.

Reddy, A. J., Herschl, M. H., Geng, X., Kolli, S., Lu, A. X., Kumar, A., Hsu, P. D., Levine, S., and Ioannidis, N. M. Strategies for effectively modelling promoter-driven gene expression using transfer learning. *bioRxivorg*, pp. 2023.02.24.529941, May 2024.

Roohani, Y., Lee, A., Huang, Q., Vora, J., Steinhart, Z., Huang, K., Marson, A., Liang, P., and Leskovec, J. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. *arXiv*, 2405.17631, 2024. URL https://arxiv.org/abs/2405.17631.

Sellés Vidal, L., Isalan, M., Heap, J. T., and Ledesma-Amaro, R. A primer to directed evolution: current methodologies and future directions. *RSC Chem. Biol.*, 4 (4):271–291, April 2023.

Siraj, L., Castro, R. I., Dewey, H. B., Kales, S., Butts, J. C., Nguyen, T. T. L., Kanai, M., Berenzy, D., Mouri, K., Wang, Q. S., Fiziev, P. P., Tsuo, K., McCaw, Z. R., Gosai, S. J., Aguet, F., Cui, R., Kassam, I., McRae, J., Vockley, C. M., Lareau, C. A., Abramov, S., Boystov, A., Vierstra, J., Okada, Y., Gusev, A., Jones, T. R., Lander, E. S., Sabeti, P. C., Finucane, H. K., Reilly, S. K., Ulirsch, J. C., and Tewhey, R. Functional dissection of complex trait variants at single-nucleotide resolution. *Nature*, February 2026.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165(6):1519–1529, June 2016.

Yang, J., Lal, R. G., Bowden, J. C., Astudillo, R., Hameedi, M. A., Kaur, S., Hill, M., Yue, Y., and Arnold, F. H. Active learning-assisted directed evolution. *Nat. Commun.*, 16(1):714, January 2025.

Yin, C., Castillo-Hair, S., Byeon, G. W., Bromley, P., Meuleman, W., and Seelig, G. Iterative deep learning design of human enhancers exploits condensed sequence grammar to achieve cell-type specificity. *Cell Syst.*, 16(7):101302, July 2025.

Zhao, J., Baltoumas, F. A., Konnaris, M. A., Mouratidis, I., Liu, Z., Sims, J., Agarwal, V., Pavlopoulos, G. A., Georgakopoulos-Soares, I., and Ahituv, N. MPRAbase: A massively parallel reporter assay database. *bioRxivorg*, pp. 2023.11.19.567742, November 2023.

# A. Supplementary Material

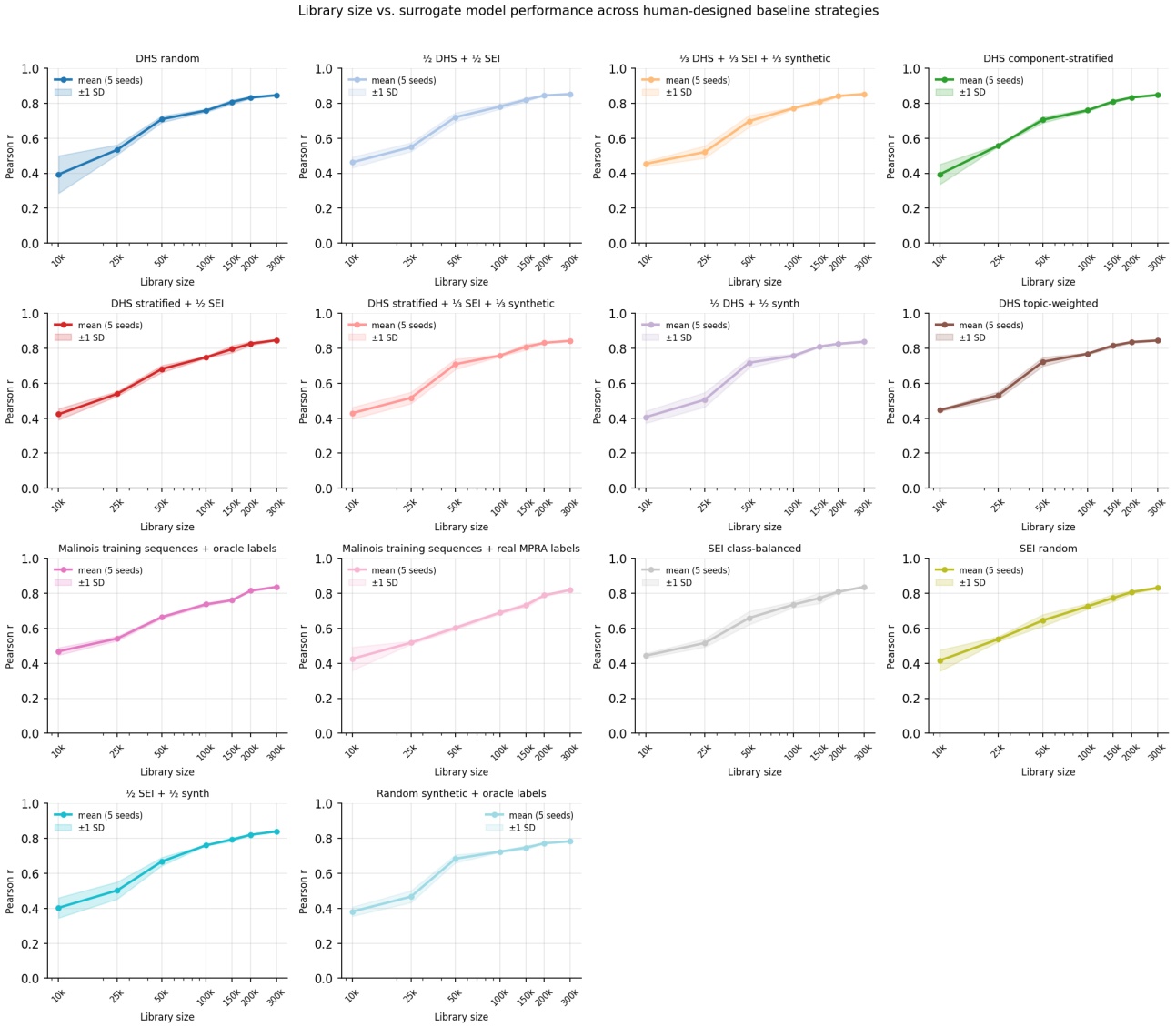


*Figure S1.* Downstream model performance as a function of library size for all 14 human-designed strategies. Each curve shows mean Pearson $r$ on `eval_01` across 5 random seeds ($\pm 1$ SD shaded). Rankings are largely preserved across library sizes; the steepest gains occur between 10,000 and 150,000 sequences with diminishing returns thereafter.

*Table S1.* Evaluation sets used to score trained sequence-to-activity models. All genomic sets are drawn from chromosomes excluded from Malinois training (chr7, 13, 19, 21, X). Empirical sets use wet-lab MPRA measurements as labels; oracle sets use Malinois-predicted activity as labels.

| Eval | Name | Source | $n$ |
|---|---|---|---|
| `eval_01` | chr7_13_gt | Held-out sequences from Gosai et al. (2024) episomal MPRA (Table S2), chr7 and chr13, standard error $< 1.0$; empirical labels | 60,055 |
| `eval_02` | ukbb_gtex_gt_oracle | UK Biobank and GTEx variant sequences from `eval_05`, relabeled by Malinois; oracle labels | 59,084 |
| `eval_03` | ukbb_gtex_one_oracle | UK Biobank and GTEx variants, chr7 and chr13, deduplicated to one allele per locus, no standard error filter; relabeled by Malinois; oracle labels | 30,505 |
| `eval_04` | chr19_21_X_gt | Held-out sequences from Gosai et al. (2024) episomal MPRA (Table S2), chr19, chr21, and chrX, standard error $< 1.0$; empirical labels | 56,340 |
| `eval_05` | ukbb_gtex_gt | Fine-mapped UK Biobank and GTEx cis-eQTL variants from Gosai et al. (2024) Table S2, chr7 and chr13, standard error $< 1.0$; empirical labels | 59,084 |
| `eval_06` | ukbb_gtex_both_oracle | UK Biobank and GTEx variants, chr7 and chr13, reference and alternate alleles retained, no standard error filter; relabeled by Malinois; oracle labels | 62,966 |
| `eval_07` | sei_chr7_13 | Genomic regions from the Sei sequence class annotation (Chen et al., 2022), chr7 and chr13, regions $\geq 200$ bp, midpoint-centered; oracle labels | 20,000 |
| `eval_08` | synthetic | Uniform random sequences over $\{A, C, G, T\}$, 200 bp, drawn independently per position (seed 7777); oracle labels | 20,000 |
| `eval_09` | chr19_21_X_gt_oracle | Sequences from `eval_04`, relabeled by Malinois; oracle labels | 56,340 |
| `eval_10` | dhs_chr7_13 | DNase I hypersensitive sites from the DHS index (Meuleman et al., 2020) (hg38), chr7 and chr13, 200 bp centered on summit; oracle labels | 20,000 |
| `eval_11` | ukbb_gtex_both | UK Biobank and GTEx variants, chr7 and chr13, reference and alternate alleles retained, no standard error filter; empirical labels | 62,966 |
| `eval_12` | ukbb_gtex_one | UK Biobank and GTEx variants, chr7 and chr13, deduplicated to one allele per locus, no standard error filter; empirical labels | 30,505 |
| `eval_13` | genomic_chr7_13 | Random 200 bp windows tiling chr7 and chr13 (hg38, non-N sequence only); oracle labels | 20,000 |
| `eval_14` | chr7_13_gt_oracle | Sequences from `eval_01`, relabeled by Malinois; oracle labels | 60,055 |