Active learning of enhancer and silencer regulatory grammar in photoreceptors

Ryan Z. Friedman[1,2], Avinash Ramu[1,2], Sara Lichtarge[1,2], Connie A. Myers[3], David M. Granas[1,2], Maria Gause[3], Joseph C. Corbo[3], Barak A. Cohen[1,2], Michael A. White[1,2,*]

[1]The Edison Family Center for Genome Sciences & Systems Biology, [2]Department of Genetics, [3]Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, MO, 63110.

*Correspondence to: mawhite@wustl.edu

ABSTRACT

Cis-regulatory elements (CREs) direct gene expression in health and disease, and models that can accurately predict their activities from DNA sequences are crucial for biomedicine. Deep learning represents one emerging strategy to model the regulatory grammar that relates CRE sequence to function. However, these models require training data on a scale that exceeds the number of CREs in the genome. We address this problem using active machine learning to iteratively train models on multiple rounds of synthetic DNA sequences assayed in live mammalian retinas. During each round of training the model actively selects sequence perturbations to assay, thereby efficiently generating informative training data. We iteratively trained a model that predicts the activities of sequences containing binding motifs for the photoreceptor transcription factor Cone-rod homeobox (CRX) using an order of magnitude less training data than current approaches. The model's internal confidence estimates of its predictions are reliable guides for designing sequences with high activity. The model correctly identified critical sequence differences between active and inactive sequences with nearly identical transcription factor binding sites, and revealed order and spacing preferences for combinations of motifs. Our results establish active learning as an effective method to train accurate deep learning models of *cis*-regulatory function after exhausting naturally occurring training examples in the genome.

INTRODUCTION

The *cis*-regulatory grammar of a cell type defines the rules for interpreting CREs and their allelic variants by cell type-specific transcription factors (TFs). *Cis*-regulatory grammars describe how TF binding sites in a DNA sequence relate to its *cis*-regulatory activity [1–5]. The effects of TF binding sites depend strongly on their sequence contexts, and CREs composed of similar binding sites often differ widely in their activities [6–23]**.** These effects of sequence context on *cis*-regulatory activity make it difficult to distinguish genuine CREs from non-functional clusters of TF binding sites in the genome, and to accurately predict the effects of non-coding genetic variants. Therefore, a major goal of regulatory genomics is to learn cell-type specific *cis*-regulatory grammars that explain why DNA sequences with similar or identical TF binding sites exhibit a wide range of *cis*-regulatory activities.

Modern machine learning methods, with their power to discover predictive features in high dimensional data, have the potential to discover features of *cis*-regulatory grammars that are missed by conventional enrichment analyses of TF motifs. Several studies report machine learning models trained on large epigenomic datasets [24–31]. These studies attempt to learn the entire *cis*-regulatory grammar of a cell type. A limitation of these approaches is that the genome does not contain enough naturally occuring CREs to provide sufficient training examples to learn the complex interactions that occur among TF motifs. Thus, these approaches typically reveal the TF motifs with the largest independent effects on gene expression, which tend to be the same motifs identified by traditional motif-finding algorithms. In addition, while many of these models accurately predict TF binding and accessible chromatin in the genome, they are trained on indirect proxies of *cis*-regulation, and therefore less accurately predict *cis*-regulatory activity. Models trained on massively parallel reporter gene assays (MPRAs) [19,32–37] do predict *cis*-regulatory activity directly. Still, MPRA studies must also contend with the same fundamental limitation: the number of genomic training examples in any particular cell type is small relative to the scale of the training data typically needed to model the interactions defining *cis*-regulatory grammars [38]. To model the complex interactions among TF motifs, we require efficient strategies to pick informative training examples from the vast space of potential sequences, once the supply of genome-derived sequences has been exhausted [39].

To address this problem, we made two major modifications to the current paradigm for learning *cis*-regulatory grammars. First, we deployed an iterative machine learning approach called active learning, which is here applied to the problem of *cis*-regulation for the first time. We train models of *cis*-regulatory grammar on successive rounds of MPRA experiments that test sequences actively selected for informative sequence perturbations. The key premise of this approach is that the training examples most likely to improve the model in the next round are sequence perturbations whose predicted outcomes are most uncertain under the current model. By prioritizing these perturbations at each iteration, we efficiently sample the space of possible sequences for informative examples, and thereby train accurate machine learning models with less data [40–42]. Active learning has been successfully applied to model metabolic networks [43], optimize cell culture media [44], perform *in silico* drug screens [45–48], improve text and image classifiers [42,49], discover energy-efficient materials [50,51], identify TFs that drive cellular differentiation [52], and select optimal training data for nanopore base calling [53]. However, active learning has not yet been applied to train models of *cis*-regulatory grammars.

Second, instead of attempting to learn the entire *cis*-regulatory grammar of a cell type, we focused on a network of interactions surrounding a single multifunctional TF. This increased our power to detect combinatorial interactions among motifs and allowed us to address why sequences bound by the same TF vary in their activity, ranging from enhancers to silencers. Current models are based on experiments that do not assay silencers. However, our experimental approach enabled us to model the contextual cues that specify whether sequences with the same TF motif are enhancers, silencers, or inactive. Our goal was to capture the detailed interactions that comprise a specific part of a *cis*-regulatory grammar, rather than to perform coarse-grained modeling of the entire grammar.

RESULTS

Iteratively learning the *cis*-regulatory grammar of mammalian rod photoreceptors

A striking example of how a single TF binding motif can specify different activities in different DNA contexts is *cis*-regulation by the retina homeodomain TF Cone-rod homeobox (CRX). [54–61]. In rods, CRX cooperates with other TFs to activate rod-specific genes and repress cone-specific genes [9,57,62–65]. This dual role of CRX as activator and repressor requires context-specific regulatory information in its target CREs. Consistent with this idea, some CRX-bound genomic sequences act as enhancers, while others act as silencers when tested by MPRA [12,15,19,66,67]. Many sequences with CRX motifs exhibit no *cis*-regulatory activity despite being bound by CRX and residing in open chromatin, similar to results reported for TF-bound sequences in cell culture models [16,68–70]. These enhancers, silencers, and inactive sequences tested by MPRA all have one or more copies of the CRX motif, which demonstrates that the *cis*-regulatory effect of a CRX-bound motif depends strongly on its sequence context. Thus, CRX-dependent *cis*-regulation in photoreceptors embodies a problem faced by all *cis*-regulatory grammars: how DNA contexts encode activating or repressing activity.

We sought to learn the sequence features that are responsible for the activity differences among sequences with one or more copies of the CRX motif by training a machine learning model of rod photoreceptor *cis*-regulatory grammar. To train the model, we developed an iterative, active learning approach that leverages the capacities of MPRAs to test large libraries of designed, synthesized DNA sequences with informative perturbations. Because this approach is iterative, it provides a way to continuously improve model performance by generating new rounds of informative training data (**Figure 1a** and Methods). In each round, the model is trained on a cumulative training dataset and its performance is evaluated on an independent test set. The current model is used to actively select new training data by identifying candidate sequences for which the model predictions are least certain, with the expectation that these would be informative training examples if their activities were known. These sequences are then synthesized and assayed, and the results are added to the training data in the next round. The goal of the approach is to iteratively and efficiently sample the search space of possible training data, targeting those parts of the space containing the most informative training examples. Accurate models are thereby trained with less data.

In each round of active learning, we generate millions of candidate sequences by mutagenizing the sequences from the current training data *in silico*. Candidate sequences are then passed through two filters. First, candidates are scored by an open chromatin machine learning model and only candidate sequences predicted to be accessible in photoreceptors are retained. This filter enforces the condition that perturbed sequences should have general sequence properties of photoreceptor regulatory elements, rather than residing in an unrelated region of sequence space. For this filter, we trained a Support Vector Machine (SVM) that predicts whether a sequence will maintain chromatin accessibility in photoreceptors (AUC = 0.868, **Extended Data Figure 1**, Methods). This filter does not depend on MPRA measurements, but only on published measurements of open chromatin in rod photoreceptors [59].

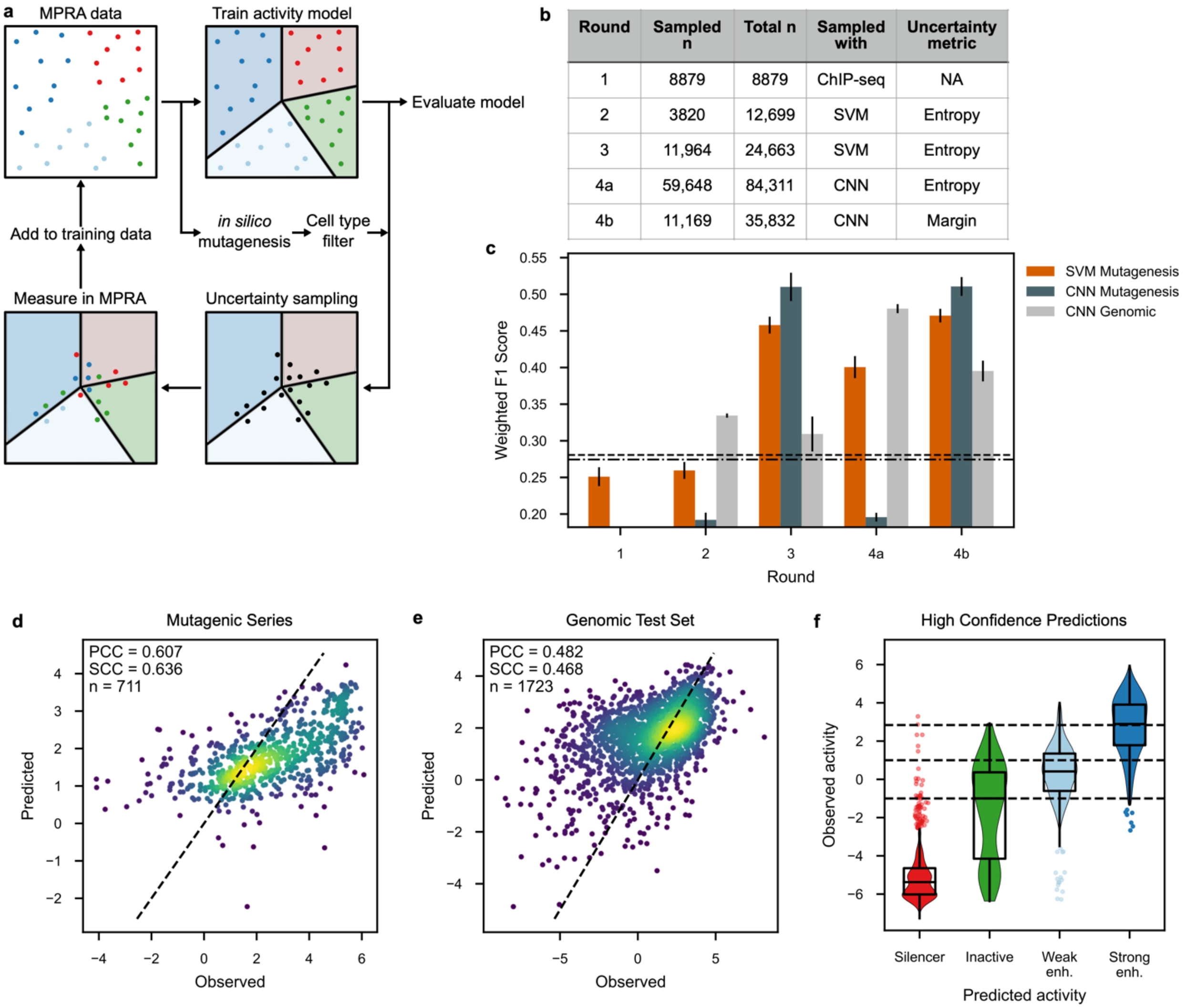
The critical component of active learning is the second filter, which samples candidate sequences whose activities are uncertain under the current model. The premise of uncertainty sampling is that the potential training examples that are most likely to improve the model are those sequences about which the model is most uncertain [42]. To implement this step, candidate sequences that passed the open chromatin SVM filter are scored by the current model of regulatory activity, which assigns a probability mass function of activity to each sequence. These probabilities are used to calculate a metric of model uncertainty (discussed below). We then select sequences with the highest uncertainty to synthesize and test in the next MPRA library. Next, we add the new MPRA data to the existing training data, retrain the model, and reevaluate performance on the test set. Using this approach, we performed three rounds of active learning, beginning with data from a large set of genomic sequences that reside in open chromatin and are bound by CRX in rod photoreceptors (**Figure 1b**).

We used our active learning framework to train two different four-way classifiers, a modified *k*-mer Support Vector Machine (SVM) and a convolutional neural network (CNN). An advantage of the *k*-mer SVM is that the feature encoding is predetermined, which allows it to be trained on smaller datasets. However, the CNN can flexibly encode higher order features that may not be captured by the SVM. Each model predicts the probability that a given DNA sequence is a strong enhancer, a weak enhancer, an inactive sequence, or a silencer in photoreceptors. The activity thresholds for these classes were based on a prior MPRA study of genomic CREs [19]. To evaluate model performance, we used two independent MPRA test datasets. The first was an exhaustive motif-mutagenesis analysis of 29 strong enhancers derived from the genome (‘mutagenic series’). Our rationale for using this set was that these 711 sequences resemble experiments investigators would perform to study individual enhancers or to evaluate specific genetic variants of interest. The second test set was 1,723 experimentally measured CRX-bound sequences [67] that were not used elsewhere in the active learning pipeline (genomic test set). The two test sets assess model performance in complementary ways. The mutagenic series tests the ability of the models to predict perturbations to TF binding sites. Because this test set is derived from mutagenesis of enhancers, a limitation is that the classes are imbalanced (44% strong enhancers vs 3.7% silencers). The genomic test set reflects the natural ratio of enhancers to silencers among CRX-bound sequences, however it lacks the extensive motif perturbations of the mutagenic series. The performance of the SVM was evaluated only on the mutagenic series because the SVM requires input sequences aligned on a central CRX motif, and sequences in the genomic test set are not aligned (Methods). We utilized the weighted F1 score (wF1) as our performance metric since our models are multi-class classifiers [71].

In the first learning round ('Round 1'), we trained the SVM and CNN on an MPRA dataset derived from photoreceptor accessible chromatin sequences that each contained at least one copy of the CRX motif [19]. Each sequence was 164 bp, centered on a central CRX motif, and placed upstream of the rod-specific *Rhodopsin* promoter, which allowed us to measure both enhancer and silencer activity. The MPRA library included 4,629 genomic sequences along with matched versions in which all CRX motifs were abolished. The MPRA library was assayed in explanted mouse retinas and thus sequences were tested in the full cellular context in which they are normally active. When trained on the Round 1 data, the SVM achieved a wF1 of 0.251 (**Figure 1c** and **Extended Data Figure 2**), where a score of 0.28 represents random guessing on the mutagenic series and a score of one represents perfect four-way classification. The CNN did not generalize beyond the training data, presumably because the Round 1 dataset was too small to effectively train the model (**Extended Data Figure 3**).

The performance of the models in Round 1 demonstrates the limitations of training models using only genome-derived sequences. Because the majority of CRX-bound sequences from the genome are in either the training dataset or the genomic test set, our results show that the genome does not contain enough examples of *bona fide* photoreceptor CREs to train robust models. However, our active learning approach provided a way to generate additional informative training data beyond genomic sequences. We completed three cycles of active learning and took Round 4b as the final round (**Figure 1c**). (Alternate Round 4a is discussed below.) We achieved a CNN performance of wF1 = 0.511 on the mutagenic series and 0.395 on the genomic test set. Thus, our active learning framework provided a route to generating additional informative training data, even after we exhausted all of the CRX-dependent regulatory elements in the genome.

Our discrete multi-class classifiers assign probabilities to their class predictions, but they do not make continuous predictions of regulatory activity. To generate quantitative predictions, we trained a regression CNN on the Round 4b training data. We obtained a Pearson correlation coefficient (PCC) of 0.607 on the mutagenic series (**Figure 1d**). For the genomic test set, our CNN obtained a PCC of 0.482 (**Figure 1e**); although this performance is lower, the test set includes silencers, which current published models of MPRA activity do not predict. Using active learning, we were able to efficiently generate informative training data that spans the full range of enhancer and silencer activities.



Round	Sampled n	Total n	Sampled with	Uncertainty metric
1	8879	8879	ChIP-seq	NA
2	3820	12,699	SVM	Entropy
3	11,964	24,663	SVM	Entropy
4a	59,648	84,311	CNN	Entropy
4b	11,169	35,832	CNN	Margin

Figure 1: Iterative machine learning improves predictions of *cis*-regulatory activity. a, Summary of active learning approach. The colored dots represent sequences measured in MPRAs (dark blue, strong enhancer; light blue, weak enhancer; green, inactive; red, silencer), which are used to train a multi-class classifier (solid lines represent the margins between classifications inferred by the model and shaded areas correspond to the inferred activity classes). A held-out test set is used to evaluate model performance. Candidate training sequences are generated by *in silico* perturbation and passed through the photoreceptor open chromatin filter SVM. Sequences scored as high uncertainty under the current model (black dots) are synthesized, measured by MPRA, and added to the training data for the next round of model fitting. **b**, Summary of the contents and generation of each training data batch. **c**, Iterative improvement of models on the mutagenic and genomic test sets. Horizontal lines represent accuracy expected from random guessing for the mutagenic series (dash) and genomic test set (dash-dot). Error bars denote one standard deviation based on ten-fold cross-validation of the newly added data. **d-e**, Predicted versus observed activity from the final model (Round 4b) for

the mutagenic series **(d)** and genomic test set **(e)**. Diagonal line denotes equality. SCC, Spearman correlation coefficient. **f**, Violin plots of observed activity stratified by predicted activity for high-confidence sequences. Horizontal dashed lines correspond to cutoffs for the activity classes. Enh., enhancer.

High-confidence model predictions are accurate

The above global correlations of model performance are based on both high- and low-confidence predictions. An advantage of our approach is that a measure of uncertainty is assigned to each individual model prediction, and thus high-confidence predictions can be separated from low-confidence predictions. To assess the accuracy of high-confidence CNN predictions, we synthesized and assayed 2,055 sequences whose activity was predicted by the CNN with high confidence. Remarkably, 72% of the high-confidence perturbations were correct (**Figure 1f**). Furthermore, 14.5% of the high confidence predictions were off by one class, indicating that the CNN learned the rank order of classes, even though this information was not explicitly provided. Therefore, the classification models' internal measures of confidence are reliable guides for active learning. In addition, for practical downstream applications, these posterior probabilities allow users to focus on accurate predictions and avoid predictions that are unlikely to be correct.

Uncertainty sampling produces informative training data

We tested whether the improvement in model performance was due to sampling the most uncertain sequences, or whether it was only a consequence of a larger training dataset. We experimentally assayed a set of 6,335 randomly sampled sequences, tested concurrently with 4,438 sequences picked by uncertainty sampling at Round 3. The models performed better when trained on uncertainty sampled training data than when trained on a dataset that included the randomly sampled examples (**Figure 2a**). Since the two groups of sequences were sampled from the same pool, we conclude that uncertainty sampling provides more informative training data than random sampling.

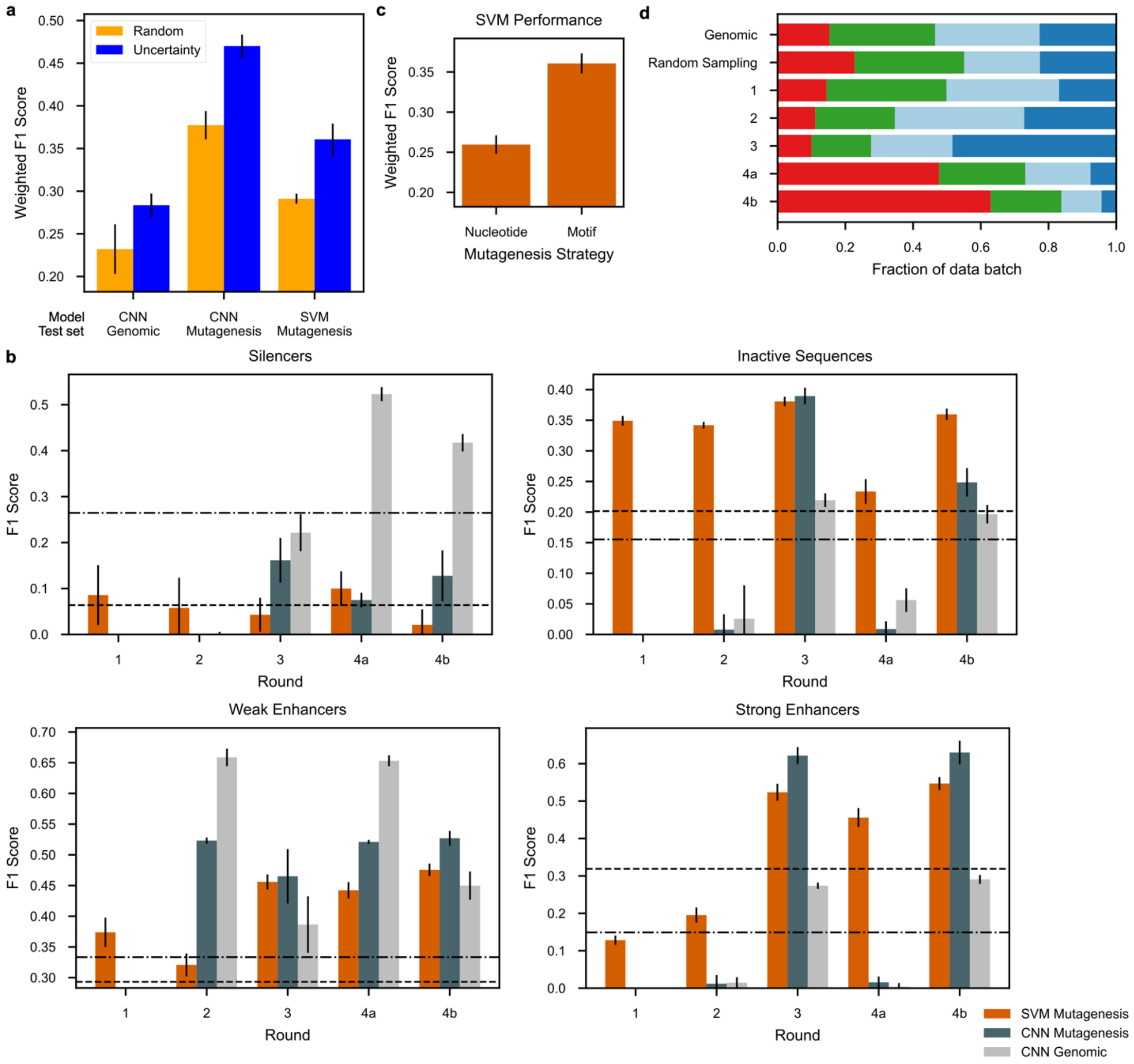


Figure 2: Mutagenesis strategies and sampling techniques influence model improvement. a, Performance of models when trained on new data from random or uncertainty sampling. **b**, Performance of models on each of the four activity classes after each round of learning. Horizontal lines represent chance for the mutagenic series (dash) and genomic test set (dash-dot). **c**, Performance of SVM predictions on the mutagenic series when trained using nucleotide-based mutagenesis (Round 1 and 2) or motif-based mutagenesis (Round 1 and 37% of Round 3 to normalize for dataset size). **d**, Distribution of activity classes in each subset of training data (dark blue, strong enhancer; light blue, weak enhancer; green, inactive; red, silencer). All error bars are the same as in **Figure 1c**.

Effectiveness of active learning depends on sequence perturbation strategy

The choice of perturbation strategy had a large impact on whether active learning could improve model performance. Our initial strategy was to randomly mutate 12% of nucleotides in the

Round 1 sequences and then select new training data with uncertainty sampling (Round 2). However, the resulting training data did not improve model performance (**Figure 1c**). In Round 3 we used a motif-centric perturbation strategy that consisted of adding, subtracting, and rearranging motifs, particularly those previously reported to be enriched in CREs in the retina [19,66]. This perturbation strategy, followed by uncertainty sampling, produced informative training examples that substantially improved model performance when pooled with all prior data (**Figure 1c**). The largest gains in model performance were on predictions for strong and weak enhancers (**Figure 2b**). To directly compare the effectiveness of mutagenesis strategies, we compared SVMs trained using nucleotide-based mutagenesis (Rounds 1+2) or motif-based mutagenesis (Round 1+37% of Round 3 to equalize the dataset size). The model trained with motif-based mutagenesis outperformed the model trained by nucleotide-based mutagenesis (**Figure 2c**). We thus continued to use motif-based perturbations to generate candidates for Round 4 of active learning.

Our results suggest that training models on datasets with more sequences from the extremes of the activity distribution (strong enhancers or silencers) leads to better model performance. This may explain why motif-centric perturbations produce more informative training examples. The nucleotide mutagenesis strategy (Round 2) generated a dataset in which silencers and strong enhancers occurred at roughly the same frequencies as genomic sequences, while motif-centric perturbations (Round 3) nearly doubled the fraction of strong enhancers (**Figure 2d**). Critically, uncertainty sampling is still required in conjunction with motif-centric perturbations to produce training data enriched for active sequences. The set of randomly sampled sequences, which were also generated by motif-centric mutagenesis, includes a larger fraction of weak enhancers and inactive sequences (54% of training examples) compared with the Round 3 dataset selected by uncertainty sampling (42% of training examples, **Figure 2d**). Notably all rounds of active learning that used motif-centric perturbations with active learning produced datasets that were highly enriched in strong enhancers (Round 3) or silencers (Rounds 4a and 4b) relative to genomic sequences (**Figure 2d**).

Effect of the uncertainty metric

We found that the metric of uncertainty used for the uncertainty sampling step affects the effectiveness of active learning. In Rounds 2 and 3 we used entropy sampling, which uses Shannon entropy as the uncertainty metric. Entropy reaches its maximum when the model is equally uncertain about all four activity classes (strong enhancer, weak enhancer, inactive, and silencer), and thus entropy sampling selects candidate sequences for which the model has no strong predictions. Entropy sampling at Round 3 improved the global performance, but at the per-class level, the models were much better at predicting strong and weak enhancers compared to silencers (**Figure 2b**). This is likely due to the large fraction of strong enhancers in the Round 3 training data (**Figure 2d**). We reasoned that enriching for silencers in the next round might improve the model.

Since the activities of candidate sequence are unknown at the time of selection, we tried two different uncertainty sampling strategies to increase the number of silencers in the training data. The first strategy used entropy sampling once more, while increasing the sample size five-fold by testing 59,648 high-entropy perturbations (Round 4a). The second strategy used margin uncertainty as the metric, which is highest when the two most likely activity classes have similar probabilities. We used margin sampling to select 11,169 candidate sequences for which silencers were one of the two most likely classes (Round 4b). Both strategies produced training datasets with a higher fraction of silencers (**Figure 2d**). However, the CNN trained on the entropy-sampled training data (Round 4a) led to overfitting (**Extended Data Figure 3**), which caused a substantial drop in the performance of global and per-class predictions (**Figures 1c and 2b**). The CNN predictions on the genomic test set appeared to improve both globally and for silencers, but at the cost of catastrophically forgetting [72] strong enhancers (**Figures 1c and 2b**). In contrast, the CNN trained on the margin sampling dataset (Round 4b) was not overfit (**Extended Data Figure 3**). Although this model did not improve its performance on the mutagenic series, it did improve its performance on the genomic test set both globally and for silencers, without any catastrophic forgetting of enhancers (**Figures 1c and 2b**). While entropy sampling and margin sampling both yielded more silencers in Round 4, margin sampling targeted silencers without losing information about the other classes of sequences, and it provided information about silencers with fivefold less training data than entropy sampling. Our results suggest that entropy sampling works well in early rounds of active learning when a model is relatively naive, but that margin sampling works well in later rounds when it is necessary to improve model performance on certain classes of predictions.

In silico **analysis reveals global grammar**

We sought to understand the global features of rod photoreceptor *cis*-regulatory grammar learned by the trained regression CNN. We used an *in silico* perturbation analysis [73] to quantify the effect of motifs for CRX and other TFs on predicted activity. We first generated a set of 4,658 background elements from shuffled genomic sequences. We then injected one or more binding sites for CRX and other TFs, and used the regression CNN to predict the activities of both the background sequences and the motif-containing sequences. By comparing the distributions of CNN activity for background sequences and motif-containing sequences, we quantified the global importance of CRX binding sites. This analysis revealed that CRX binding sites have both general and context-specific effects on activity (**Figure 3a** and **Extended Data Figure 4**). The addition of increasing numbers of CRX sites caused a general trend towards silencers, a trend which we have observed experimentally in both genomic and synthetic CREs [12,15,19,74]. However, the distribution of predicted effects also broadens as CRX sites are added, which shows that CRX sites can either increase or decrease CRE activity depending on the sequence context. For example, the predicted effect of four CRX motifs is positively correlated with the background GC content (PCC=0.387) even though GC content itself is not predicted to influence activity (**Figure 3b** and **Extended Data Figure 5**). The effect of GC content on the activity of sequences with CRX sites predicted by the model is an effect that we have observed experimentally [12]. These results suggest that even basic sequence properties, such as GC content, can have context-dependent effects on TF motifs.

We similarly quantified the predicted effects of seven other TF motifs involved in photoreceptor regulatory activity (**Figure 3a** and **Extended Data Figure 4**). The model predicts that GFI1 motifs have a repressive effect in sequences that also include a CRX site. This is consistent with our experimental finding that GFI1 motifs are enriched in genomic silencers [19]. NRL and RORB motifs are both predicted to increase CRE activity in the presence of a CRX motif, with the RORB motif leading to larger gains in activity with increasing copy number (**Figure 3a**). The strong predicted effect of RORB motifs in CRX-containing CREs is consistent with prior experimental observations of both genomic and synthetic CREs [66,74]. We further corroborated these predictions by examining our experimentally measured genomic sequences. We compared the measured activities of genomic sequences containing exactly one CRX motif and either NRL or RORB motifs (and no detectable instances of the other five motifs in our set). Consistent with the models' predictions, genomic sequences with RORB motifs had higher activities than those with NRL motifs (**Figure 3c**). We also examined an independent experimental dataset in which we abolished all NRL and RORB motifs in a set of genomic strong enhancers. Again, consistent with the model's predictions, mutating either NRL or RORB sites decreased expression, with RORB having larger effects (**Figures 3d and e**). These results show that the model accurately learned important quantitative effects of photoreceptor TF binding sites.

We next explored what the model learned about the effects of motif order and spacing. We systematically varied *in silico* the positions of one CRX and one NRL motif in a set of >4,600 background sequences, for a total of >100,000,000 unique predictions. The predicted activating effects of both CRX and NRL motifs increase as the motifs are moved closer to the basal promoter (**Figure 3f**). The *in silico* analysis also reveals a predicted synergistic effect of CRX and NRL at certain spacings (evident in the diagonal stripes in **Figure 3f**). In contrast to the predicted activating effects of CRX with NRL, the model predicts that sequences with GFI1 and CRX motifs are largely repressive (**Figure 3g**). However, this repressive effect is eliminated when the CRX motif is within ~65 bp of the basal promoter and the GFI1 motif is in a more distal position. However, when the GFI1 motif is moved close to the CRX motif, the sequences are predicted to be repressive, suggesting that the model infers that GFI1 acts through short-range repression [75]. This analysis demonstrates how the trained model can be used to generate hypotheses about *cis*-regulatory grammar.

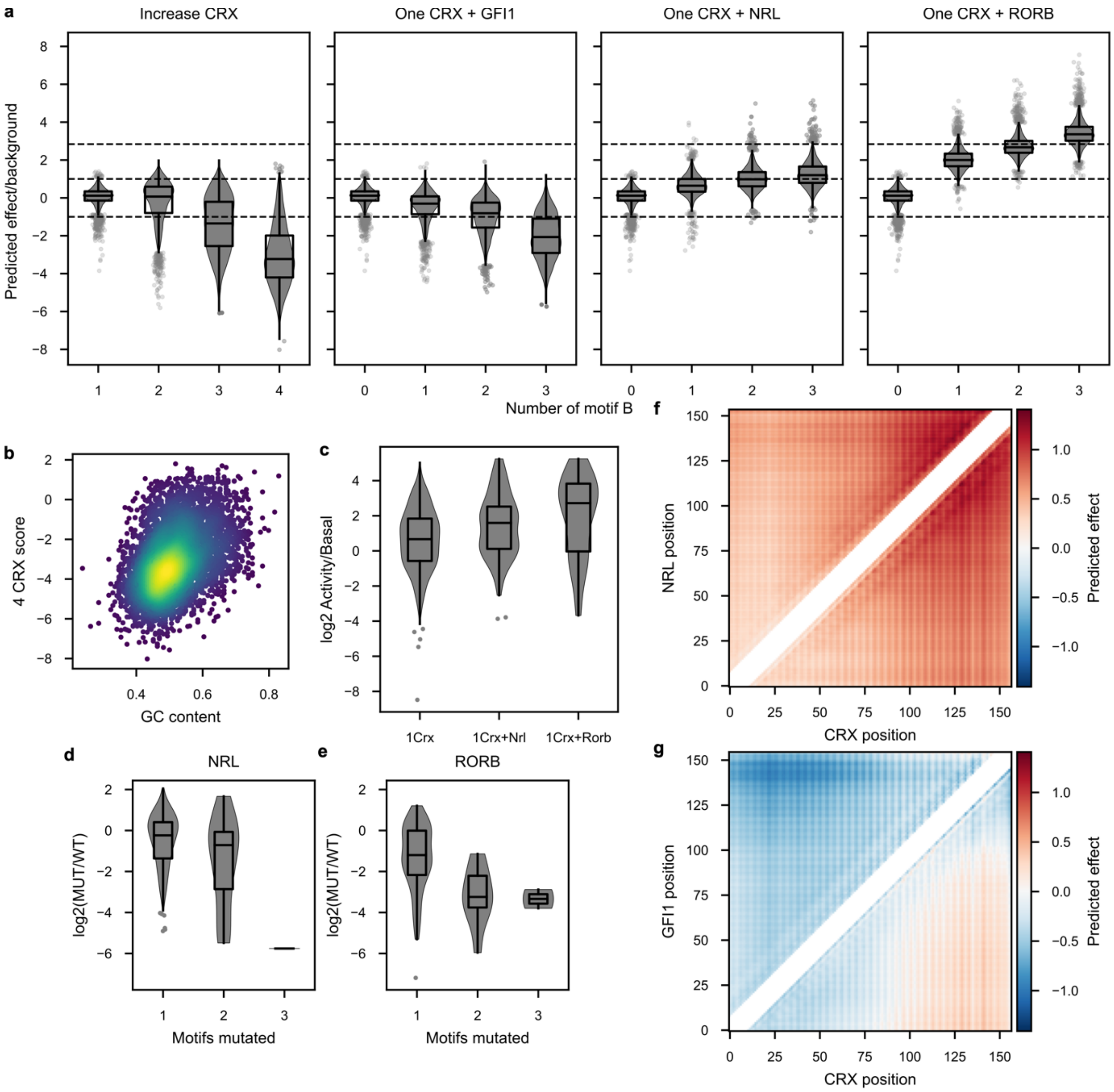


Figure 3: Regression CNN reveals global grammar. a, Predicted effect of motifs on regulatory activity. Leftmost panel shows the effect of 1-4 CRX motifs, subsequent panels show the effect of 0-3 motifs for other TFs (motif B) in the context of one CRX motif. Horizontal dashed lines denote activity class boundaries. **b**, Directional effect of 4 CRX motifs is correlated with background GC content. Each dot represents a different background sequence. Warmer colors denote higher point density. **c**, Activity distributions of genomic sequences with exactly 1 CRX motif and either NRL or RORB motifs. **d-e**, Effects of scrambling all NRL **(d)** or RORB **(e)** motifs in genomic strong enhancers, stratified by the number of motifs scrambled. **f-g**, Predicted effects of a CRX motif and an NRL **(f)** or GFI1 **(g)** motif at all possible positions. Each pixel corresponds to the mean predicted effect of the two motifs in >4,600 different background sequences. White diagonal denotes excluded arrangements where the motifs would have overlapped. The basal promoter is at position 164. Colorbar is the same for both heatmaps.

The model identifies causal differences between similar CREs

A major challenge in genomics is to understand why some sequences with similar motif content show large differences in *cis*-regulatory activity. We identified two genomic sequences whose motif content is nearly identical. Each sequence contains one CRX and one NRL site. When measured by MPRA, one sequence was inactive (**Figure 4a**) while the other was a strong enhancer (**Figure 4b**). We identified the contextual differences that underlie the activity differences between these two sequences by using the regression CNN to predict the contribution of each nucleotide to the activities of both sequences [76]. Strikingly, the NRL motif was only predicted to contribute to activity in the strong enhancer, even though the inactive sequence contains the exact same NRL motif sequence (**Figures 4a and b**). This indicates that the CNN learned how the activities of two identical NRL sites are modified by local sequence context.

This local sequence context appeared to be positions scored as important by the CNN and which lie adjacent to the NRL site in the strong enhancer (**Figure 4b**). The sequence in this position does not match any of the eight TF motifs that we previously found to be globally enriched in photoreceptor CREs [19], and thus it was not identified in our motif scan. We used this sequence to search the HOCOMOCO database [77] and identified a near-optimal motif for the NR1 TF family (**Extended Data Figure 6**), which includes TFs reported to complex with CRX, NRL, and NR2E3, another photoreceptor TF [58,78]. This motif belongs to a clade of nuclear receptor DNA binding domains that is distinct from that of more well-characterized rod TFs RORB and NR2E3. The NR1-family motif is not sufficiently enriched among all strong enhancers to be detected by traditional motif-finding methods. This shows that the model learned important features of the photoreceptor *cis*-regulatory grammar that do not reach statistical significance at a global level, but which nevertheless play important roles in specific CREs.

We attempted to experimentally validate the model's predictions by performing perturbation analyses on these two sequences via MPRA (Methods). Scrambling either the CRX or the NRL motif alone in the strong enhancer reduced activity, while scrambling both motifs together nearly abolished activity (**Figure 4c**). This confirms that both the CRX and NRL sites contribute to activity in the strong enhancer. Scrambling the NR1-family site largely abolished activity, confirming the model prediction that this sequence is important for activity. Scrambling other regions of the strong enhancer had little or no effect, again consistent with the model's prediction that these intervening sequences do not contribute to activity. We then determined the sufficiency of sequence elements by testing whether the inactive sequence could be made active by swapping in the CRX and NRL motifs, or by swapping in the spacer regions between these motifs. The spacer between the CRX and NRL sites that includes the NR1-family site conferred strong activity when swapped into the inactive sequence, while swapping in other spacer regions did not (**Figures 4d**). The CNN predicted that the NR1-family motif caused the observed gain in activity via both an independent effect of the NR1-family site as well as through synergy with the CRX motif, but not the NRL motif (**Figure 4e** and **Extended Data Figure 7a**). Without the NR1-family site, neither changing the position of the NRL motif nor the sequence of the CRX motif was sufficient to convert the inactive sequence to an active one (**Figures 4d**). We conclude that the CNN identified a new sequence feature, the NR1-family site, that provides the necessary and sufficient context to discriminate between similar clusters of motifs.

The results from the model suggest that the activity of the NRL motif is affected by its proximity to the NR1-family motif. The CNN predicts that in the strong enhancer, when the NRL motif is adjacent to the NR1-family motif, NRL has an activating effect, while in the inactive sequence the isolated NRL motif is inert (**Figures 4a and b**). The CNN also predicts that the NRL motif remains inert when the NR1-family motif is inserted in a distal position (**Figure 4e**). We experimentally tested the effect of the position of the NRL motif by sliding it along the strong enhancer sequence, while using the model to predict the importance of the NRL motif as it was moved (**Figure 4f**). Moving the NRL motif caused a loss of activity only when it disrupted the NR1-family motif, except at position 128, where wild type activity was restored despite disruption of the original NR1-family site (**Figure 4f, top**). The model interpretation of these results suggests that the NRL motif has a very low predicted contribution when it is not adjacent to the NR1-family site, while the NR1-family motif's predicted contribution remains high as NRL is moved away (**Figure 4f, bottom**). When the NRL motif disrupts the NR1-family motif, both motifs have low predicted contributions, except when the NRL motif is at position 128. We found that placing the NRL motif at position 128 created a cryptic NR1:NRL motif (**Extended Data Figure 7b**). This cryptic motif restores wild-type activity and is also predicted to be important by the model. These results suggest a regulatory grammar that governs the activities of the NR1-family and NRL motifs. In the strong enhancer, if the NRL motif is moved out of context, then the motif is decommissioned, but the enhancer can tolerate this loss because of the independent contribution of the NR1-family motif. But if the NRL site moves to a location that disrupts the NR1-family motif, then the enhancer can no longer function. We conclude that the strong enhancer is composed of both a modular (NR1) and a context-dependent (NRL) motif. So long as these constraints are satisfied, the same activity can be achieved with multiple combinations and arrangements of sites, providing a flexible grammar for constructing and editing an enhancer.

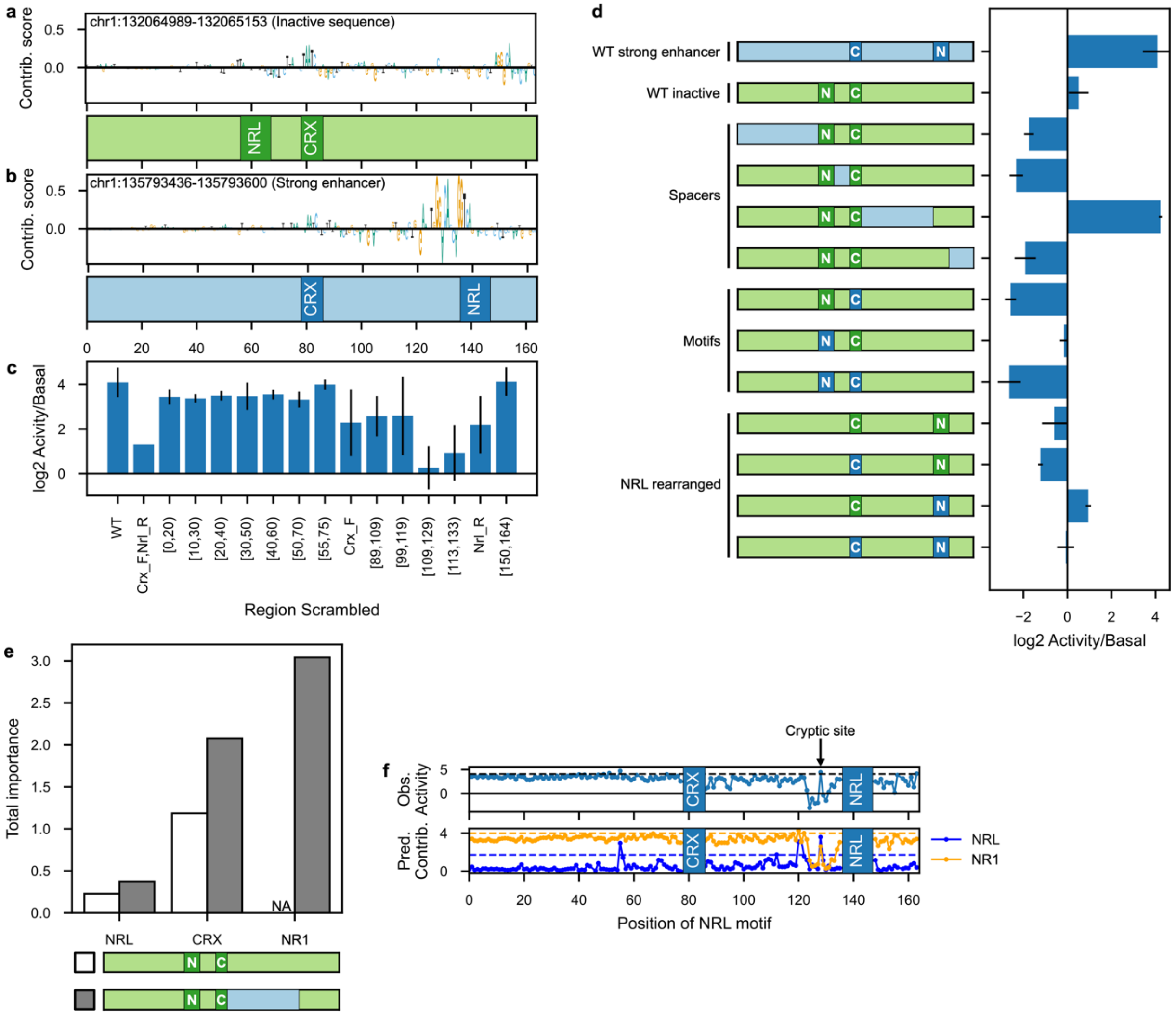


Figure 4: Model discovers novel sequence feature necessary and sufficient for activity. a-b, Relative nucleotide contribution scores and locations of known motifs for an inactive sequence **(a)** and a strong enhancer **(b)**. **c**, Effect of scrambling different parts of the strong enhancer. Error bars show standard deviation of 3-5 independent scrambles, or 4 independent replicates of wild-type. **d**, Effect of swapping strong enhancer regions into the inactive sequence. Cartoons show the designed construct with colors matching (**a and b**). Spacers were swapped into their complementary coordinates. Error bars show standard deviation of one sequence across 3 replicates. C, CRX motif; N, NRL motif. **e**, Total predicted importance of the NRL, CRX, and NR1-family motif in the inactive sequence before and after swapping in region 86-136 from the strong enhancer. **f**, Observed effect of moving NRL motif within the strong enhancer (top) and predicted contribution of the NRL and NR1-family motifs to activity (bottom). Horizontal dashed lines represent the activity (top) and motif contribution (bottom) of the wild-type sequence.

DISCUSSION

We have presented an active machine learning framework for modeling *cis*-regulatory grammar. With four rounds of MPRAs in living mouse retinas, we arrived at a model that makes accurate predictions, recapitulates prior knowledge, and reveals novel sequence features necessary and sufficient for enhancer activity in photoreceptors. We discovered a nuclear receptor-type motif that is necessary and sufficient for enhancer activity. The model also revealed that, in some cases, the activities of NRL sites depend on an adjacent NR1-family motif. We previously found that motifs for other nuclear hormone receptors are enriched in CRX-bound enhancers [19,59,66], but the NR1-family motif belongs to a distinct clade of DNA binding domains. Independent experimental evidence supports a role for NR1-family TFs in photoreceptor-specific gene regulation. For example, NR1D1 can interact with CRX, NRL, and NR2E3, and can transactivate a subset of photoreceptor promoters *in vitro* [78,79]. In a prior analysis, NR1-family motifs were not detected as enriched among photoreceptor CREs [19], however the model was able to identify a functional occurrence of this motif. This shows that the model has the power to learn features of the *cis*-regulatory grammar that are missed by conventional statistical tests for significant enrichment. Taken together, our results suggest that active learning will be a potent approach for unraveling the interactions among TFs that control cell type-specific gene regulation.

Machine learning is often portrayed as an unbiased, automated approach, but we emphasize that success requires critical “human in the loop” decisions [40]. The metric of uncertainty, perturbation strategy, size of the training data, and complexity of the machine learning model are all variables that require tuning based on experience and intuition. For example, modeling silencers required introducing margin sampling in the later rounds, which biased the training data towards silencers. Thus, success often depends on introducing bias at key points.

The active learning framework, which biases training data towards the most uncertain predictions, allowed us to train models using an order of magnitude less training data than current approaches, while modeling both enhancers and silencers. Complementary approaches rely on hundreds of millions of random sequences [34,37] or hundreds of thousands of genomic sequences [33,80] measured in a single, large-scale screen. In contrast, we measured only tens of thousands of sequences, but spread over iterative rounds of screening. We suggest that active learning may be a more efficient approach because it generates training data that is enriched for positive examples of highly active sequences, whereas genomic sequences and random sequences are enriched for negative examples of inactive sequences. The iterative approach allows active learning to home in on training data that can rapidly improve a model.

Active learning also solves the problem of where to find training data once examples from the genome have been exhausted. *Cis*-regulatory grammars are complex enough that the genome does not contain enough examples of active CREs from which to learn complete models. Active learning circumvents this problem by using informative perturbations to natural sequences as training data. Active learning will continue to generate highly informative rounds of training data, even after every naturally occurring element in the genome has been measured. Large libraries of random DNA may complement active learning [34,35,37,81,82]. Models trained on random DNA, or fragmented genomic DNA, attempt to learn the entire regulatory grammar of a cell at once. In contrast, our approach gained power by focusing only on CRX-dependent CREs. Even focusing on this subset of the photoreceptor regulatory network required several rounds of active learning, which highlights the complexity of *cis*-regulatory grammars. Given that informative training data will be sparse in random DNA, and that most random DNA has weak or no regulatory activity [34,37], the size of random DNA libraries required to provide training data to create complete *cis*-regulatory grammars may be experimentally intractable. However, the results from random library screens may serve as good starting points from which to perform rounds of active learning. Active learning provides a framework for focusing on small sets of training data with large effects on model performance.

Our results show that machine learning approaches benefit from training data that spans the widest possible dynamic range. Our decision to employ a cell type-specific basal promoter in our MPRA experiments allowed us to reliably measure both enhancers and silencers, which improved power relative to approaches that only measure enhancer activity. Future approaches should include experimental designs that reproducibly measure the full range of possible regulatory activities, regardless of whether the training data consists of random, genomic, or mutated DNAs.

Our work highlights the risks of relying too heavily on one technique for uncertainty sampling. In Round 4, both entropy sampling and margin sampling successfully produced training datasets biased toward silencers, but entropy sampling, which had also been used in the prior rounds, resulted in catastrophic forgetting [72]. Future work can protect against catastrophic forgetting by employing diversity criteria [83,84] and utilizing multiple sampling techniques in each round, including ensemble-based sampling [85,86]. Every round should also include some perturbations from random sampling to test whether uncertainty sampling is working as expected. With respect to the strategy used to generate perturbations from which to sample, recent advances in generative modeling and gradient-based design could provide a complementary approach to generating training data [28,87,88].

We implemented active learning in the context of discrete classifications (silencer, inactive, weak enhancer, and strong enhancer). Using these discrete classifications provided a straightforward way to implement uncertainty sampling using concepts from information theory. However, it will be desirable to implement active learning in a regression setting by taking advantage of sampling strategies that rely on Gaussian and neural processes [45,89,90].

METHODS

Library design

All MPRA libraries were created using oligonucleotide (oligo) synthesis from Agilent Technologies™ as previously described [19,91]. Every MPRA library contained 164 bp test sequences marked with unique 9 bp barcodes following the scheme: 5’ priming sequence, EcoRI, library sequence, SpeI, filler sequence, SphI, CRE barcode (cBC), NotI, 3’ priming sequence (**Extended Data Figure 8**). The filler sequence is used to ensure all oligos are 230 nt for synthesis and is subsequently eliminated during cloning. In addition to this common design scheme, all libraries contained several groups of constant sequences: (1) a construct for the basal promoter alone, which is present with multiple redundant barcodes, (2) a set of 150 scrambled genomic sequences (3) 20 genomic sequences that span the full dynamic range of the assay. The MPRA libraries are described in **Supplementary Tables 1-3**.

Library cloning

We created MPRA libraries from oligo pools using two-step cloning as described [19,91] with the following modifications. Oligos were amplified through multiple PCR reactions (New England Biolabs [NEB] Q5 High-Fidelity 2X Master Mix, cat. #M0515, see **Supplementary Table 4** for primer sequences), purified from an agarose gel, digested with EcoRI-HF and NotI-HF (NEB), and then cloned into the EagI and EcoRI sites of pJK03 (AddGene #173,490) in multiple ligation reactions (NEB T4 ligase). Ligation products were transformed into either 5-alpha or 10-beta electrocompetent cells (NEB) and grown in liquid LB-Amp cultures. Plasmid pools were digested with SphI-HF and SpeI-HF and treated with Antarctic phosphatase or Quick CIP (NEB), then ligated to reporter gene inserts in multiple reactions (NEB T4 ligase). Ligation products were transformed into electrocompetent cells and grown in liquid culture from which we prepared plasmid DNA.

We next cloned the *Rho* basal promoter into the plasmid library in between the test sequence and its cognate barcode. Basal promoter inserts were prepared by amplifying the *Rho* basal promoter and *DsRed* from the plasmid pJK01 (AddGene #173,489) using the forward primer MO566 and reverse primers that add 9bp multiplexing barcodes (mBC, **Supplementary Tables 1 and 2**), purified from an agarose gel, and digested with NheI-HF and SphI-HF (NEB). Adding mBCs to the reporter gene allows us to test larger libraries by amplifying sublibraries with different primer sets and cloning each sublibrary in parallel with a unique mBC. When necessary, we could then mix sublibraries (**Supplementary Table 1**) for parallel analyses. Barcode complexity was always verified by sequencing the final library on the Illumina MiniSeq™ platform.

MPRA in mouse retinal explants

Animal procedures were performed in accordance with a Washington University in St. Louis Institutional Animal Care and Use Committee approved vertebrate animals protocol. CD-1 IGS mice were obtained from Charles River Laboratory. Retinas from newborn (P0) mice were dissected and electroporated as described [62]. The sex of the mice could not be determined at the P0 stage. Retinas were dissected in serum-free medium (SFM; 1:1 Dulbecco's Modified Eagle Medium (DMEM):Ham's F12 (Gibco, 11330-032), 100 units per ml penicillin and 100 μg ml−1 streptomycin (Gibco, 15140122), 2 mM GlutaMax (Gibco, 35050-061) and 2 μg ml−1 insulin (Sigma, I6634) from surrounding sclera and soft tissue leaving the lens in place. Retinas were then transferred to an electroporation chamber (model BTX453 Microslide chamber, BTX Harvard Apparatus modified as described [92]) containing 0.5 μg μl−1 of MPRA library. Five retinas were pooled for each biological replicate and at least three replicates were performed for each library (**Supplementary Table 1**). Five square pulses (30 V) of 50-ms duration with 950-ms intervals were applied using a pulse generator (model ECM 830, BTX Harvard Apparatus). Electroporated retinas were removed from the electroporation chamber and allowed to recover in SFM for several minutes before being transferred to the same medium supplemented with 5% fetal calf serum (Gibco, 26140-079). The retinas were then placed (lens side down) on polycarbonate filters (Whatman, 0.2 μm pore size 110,606) and cultured at 37 °C in SFM supplemented with 5% fetal calf serum. After eight days of culture, we harvested the retinas in TRIzol, homogenized tissues with a sterile needle, and extracted RNA following the manufacturer's protocol. RNA was treated with TURBO DNase and then reverse transcribed with SuperScript IV First Strand Synthesis following the manufacturer’s protocol. Barcodes were amplified from cDNA and plasmid pools with Q5 using primers BC_CRX_Nested_F and BC_CRX_R for 25 cycles. We performed 2 PCR reactions per cDNA sample and 1-2 reactions per plasmid pool. PCRs from the same sample were then pooled and purified. Custom sequencing adapters were added with two rounds of PCR with Q5. The final libraries were sequenced on the Illumina NextSeq or NovaSeq platform.

Data processing

Sequencing reads were filtered for reads that contained both the cBC and the mBC (when utilized) in the correct sequence context. One sequencing library (**Supplementary Table 1**) contained a systematic error that led to N’s in positions 5, 16, 21, 27, 29, and 34. Position 5 was in a sequencing adapter, positions 16 and 21 were in constant regions, and positions 21, 27, and 34 were in the mBC region. Since the mBC could only be one of two 9 bp sequences with a Hamming distance of 8, we could still unambiguously assign mBC-cBC pairs if there are no other errors in the read outside of these 6 positions.

We removed cBCs with fewer than 50 counts in the plasmid pool or with a coefficient of variation above 0.8 across cDNA samples. Each sample was then normalized by sequencing depth; sublibraries that were cloned separately but co-electroporated were processed separately to account for any cloning batch effects. cDNA barcodes were normalized to plasmid barcode abundances; then, barcodes corresponding to the same CRE were averaged together to obtain an activity score for each CRE in each replicate. These activity scores were normalized to within-sample basal activity and then averaged together to obtain a final activity score. In cases where basal recovery was poor (median RNA/DNA barcode ratio below 0.05 in any one replicate), we calculated a pseudobasal activity using the cBCs corresponding to scrambled sequences. We took this as a reasonable approximation of basal activity because the average activity of scrambled sequences in our initial library is no different from basal. All within-library quality control summary statistics are in **Supplementary Table 5**. Activity scores were discretized into four classes using the cutoffs we previously reported [19].

Filter SVM for open chromatin

We fit a *k*-mer SVM to classify rod photoreceptor ATAC-seq peaks [59] from the rest of the mouse genome. We used the central 300 bp of all 39,265 rod ATAC-seq peaks as positives and selected an equal number of GC-matched sequences from the mm10 genome using the script `nullseq_generate.py` from the gkmSVM package. We fit this model using LS-GKM [93] with parameters `-l 10 -k 6 -d 3` and five-fold cross-validation. We assessed model performance using the Area Under the Receiver Operating Characteristic (AUROC, **Extended Data Figure 1**). In each round of *in silico* mutagenesis, we selected perturbations that had an LS-GKM score of 1 or greater; this score corresponds to a sequence that is beyond the maximum-margin hyperplane. Empirically, this removed at least half of all perturbations in each cycle.

SVM classifiers of MPRA activity

For our SVM classifier of MPRA activity, we implemented a version of the *k*-mer kernel that accounts for *k*-mer position. Since all sequences in our initial training data were centered on a high-quality CRX motif, *k*-mer position effectively reflects the distance of a *k*-mer from a CRX motif. As Giguère *et al.* show [94,95], the Generic String Kernel for any two strings $\mathbf{y}, \mathbf{y}'$ is:

$$GS(\mathbf{y}, \mathbf{y}'; n, \sigma_p, \sigma_c) = \sum_{k=1}^{n} \sum_{i=0}^{|\mathbf{y}|-k} \sum_{j=0}^{|\mathbf{y}'|-k} \exp\left(-\frac{(i-j)^2}{2\sigma_p^2}\right) \exp\left(-\frac{\|\psi^k(y_{i+1}, \ldots, y_{i+k}) - \psi^k(y'_{j+1}, \ldots, y'_{j+k})\|^2}{2\sigma_c^2}\right)$$

where n controls the maximum length of k-mers, σ_p controls the weight of k-mer position, σ_c controls the weight of k-mer similarity, and ψ^k is a k-mer encoding function. (We use k wherever Giguère *et al.* used ℓ.)

When $\sigma_c = 0$, the second $\exp$ becomes an indicator function that is only true if the two *k*-mers are identical. If we further remove the first summation from all $1, \ldots, n$ possible *k*-mers and only consider *k*-mers of length n, we can rewrite the above kernel function as:

$$GS(\mathbf{y}, \mathbf{y}'; n, \sigma_p) = \sum_{i=0}^{|\mathbf{y}|-n} \sum_{j=0}^{|\mathbf{y}'|-n} \exp\left(-\frac{(i-j)^2}{2\sigma_p^2}\right) \mathbf{1}((y_{i+1}, \ldots, y_{i+n}) = (y'_{j+1}, \ldots, y'_{j+n}))$$

When $\sigma_p = \infty$, the first $\exp$ becomes constant and we recover the *k*-mer kernel [96,97]. We extended Giguère *et al.*'s implementation for peptide sequences to allow for DNA *k*-mers; using this implementation, we pre-computed the Gram matrix for all available data. Then, we used the `SVC` class from scikit-learn [98] with `probability = True` to fit our multi-class classifier. We performed a grid search over the hyperparameters $k \in [6, 8]$ and $\sigma_p \in [0, 3, 10, 20, 50, \infty]$ using five-fold cross-validation on our initial training data. We selected $k = 6, \sigma_p = 10$ based on the AUROC and used these hyperparameters for all future modeling.

CNN classifiers of MPRA activity

Our CNN was designed as a multi-class classifier that uses one-hot encoded 164-bp long DNA sequence (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) to predict its activity in retinal MPRAs. Our model architecture consists of two convolutional layers, a max-pooling layer, a third convolutional layer, and a second max-pooling layer; this is followed by a single fully connected layer, and finally a four-node output layer with log soft-max activation, which corresponds to the log-probability the sequence belongs to each of four classes. Every convolutional and fully connected layer is followed by batch normalization, Leaky ReLU activation, and dropout regularization. We implemented the model in Pytorch [99] and used Selene [100] to train the model with Stochastic Gradient Descent (learning rate = 0.0001, momentum = 0.9, weight decay = 10^{-6}), negative log likelihood as a loss function, and a batch size of 64. We fit the model for 500 epochs using the default learning rate scheduler in Selene and kept the model with the lowest loss on a held-out validation set. We manually adjusted hyperparameters and model architectures to yield best performance on the validation set when using Round 3 as training data. We used these hyperparameters for all other datasets.

Construction of validation set

We created a validation set for the CNN by randomly sampling 10% of our original genomic sequences and then adding all perturbations derived from those sequences in Rounds 1-3. Similarly, all perturbations derived from our test set (described below) were removed from training and validation datasets for all machine learning models. To assess model performance, we performed ten-fold cross-validation on the newly added data while holding any previous data constant. This strategy ensures that the variation in model performance is only due to variation within the new data.

In silico mutagenesis for candidate perturbations

In each round, we generated a new pool of candidate sequences by perturbing sequences from the current round of training data. We defined multiple operations and performed combinations of these operations many times to generate multiple candidate sequences from each training sequence. In Round 2, the possible operations were: (1) randomly mutagenize ~12% of the positions (the exact number of positions was chosen by sampling from a Poisson distribution), (2) insert, delete, or move a random *k*-mer, (3) create a chimera with another randomly selected sequences, and (4) randomly rearrange blocks within the sequence. We performed each of these operations multiple times. In Round 3, we implemented motif-centric perturbations based on our reference list of 8 TFs (CRX, GFI1, MAZ, MEF2D, NeuroD1, NRL, RORB, and RAX) [19], plus ELF1 and TBX20. We computed the predicted occupancy with μ=9 for these TFs, defined spacer sites as positions with total predicted occupancy below 0.5, and then randomly selected one of the following operations: (1) sample from one of the position weight matrices and insert that motif into a random spacer region, (2) select an occupied site and scramble it to an unoccupied state, (3) select an occupied site and replace it with a motif sampled from a different position weight matrix, (4) select an occupied site and swap it with a length-matched spacer region. We generated additional candidate sequences by systematically scrambled tiles of spacer regions. For Round 4 we excluded ELF1 and TBX20 from the motif pool due to their very low representation in the genomic sequences.

Active learning

Our machine learning models take a 164-bp sequence $\mathbf{x}$ as input and predicts $p(y_i|\mathbf{x})$, the probability that the sequence belongs in the i-th activity bin, $i = 1, \ldots, 4$. Our objective is to determine which perturbations are the most uncertain to the model. Entropy uncertainty is quantified with Shannon entropy, $S(\mathbf{x}) = -\sum_i p(y_i|\mathbf{x}) \log_2 p(y_i|\mathbf{x})$. Margin uncertainty is defined as $M(\mathbf{x}) = 1 - [p(\hat{y}|\mathbf{x}) - p(\widehat{y^*}|\mathbf{x})]$, the difference in probability between the two most likely outcomes, $\hat{y}$ and $\widehat{y^*}$. When $\hat{y}, \widehat{y^*}$ are equally likely the term in brackets is zero, so the complement represents the uncertainty.

To provide intuition for the difference between Shannon entropy and margin uncertainty, consider three cases where $\hat{y}, \widehat{y^*}$ are equally likely. In the first case, a model outputs $[0.25, 0.25, 0.25, 0.25]$, so $S = 2$ and $M = 1$. In the second case, a model outputs $[0.33, 0.33, 0.33, 0]$; once again $M = 1$, but $S = 1.6$. In the third case, the output is $[0.5, 0.5, 0, 0]$ and M is still 1, but $S = 1$. Thus, as the two most likely classes become more distinguishable from the remaining classes, the entropy can drop without a change in margin uncertainty.

In Rounds 2 and 3, we calculated probabilities with our SVM from the previous round. In Round 2, we sampled 4800 perturbations with high entropy uncertainty. In Round 3, we sampled the 13,986 perturbations with the highest entropy uncertainty. We also randomly sampled 6584 perturbations and 25 perturbations with high probability for each of the 4 classes (100 sequences total).

For Round 4, we calculated probabilities with our CNN trained in Round 3. In Round 4a, we sampled 96,190 perturbations with the highest entropy uncertainty. In Round 4b, we sampled 18,000 perturbations with entropy uncertainty below 1.8 and high margin uncertainty for silencers. Of these, 6000 were on the silencer side of the strong enhancer vs. silencer margin, 6000 were on the strong enhancer side of the strong enhancer vs. silencer margin, and 6000 were on the silencer vs. inactive margin. Very few perturbations were on the silencer vs. weak enhancer margin, so we did not sample on this margin. We chose an entropy uncertainty cutoff of 1.8 because this approximately corresponds to probabilities of $[0.32, 0.32, 0.32, 0.05]$, which represents cases where one outcome is unlikely but the rest are equally likely. Finally, we sampled 500 perturbations with the highest probability of being a strong enhancer, 500 high-probability weak enhancers, 500 high-probability inactive sequences, and 1500 high-probability silencers.

For all data batches, the number of sequences sampled is larger than what is listed in the main text because not all sequences were recovered upon sequencing.

Evaluating model performance on test datasets

The genomic test set is from a parallel study [67] and contains 1723 CRX ChIP-seq peaks not tested in any previous study. We normalized activity scores to the basal *Rho* promoter and grouped strong and weak silencers into one silencer category, but otherwise did not re-process the data.

The mutagenic series is based on 29 CRX ChIP-seq peaks that are strong enhancers in our original Round 1 data [19]; 17 of these become weak enhancers when all CRX motifs are mutated, while 12 remain strong enhancers. For each sequence, we computed the predicted occupancy for our reference list of 8 TFs. Then we selected all possible combinations of motifs and scrambled each motif and 3 bp of flanking sequence until its predicted occupancy was below 0.01. We tested these 727 sequences in the same library as Round 3 (**Supplementary Table 1**).

We utilized the F1 score as our evaluation metric. For any class i, the F1 score is the harmonic mean between the precision and recall, or equivalently, $F_1^{(i)} = \frac{2TP_i}{2TP_i + FP_i + FN_i}$ where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. The weighted F1 score is $F_1 = \sum_i \frac{N_i}{N} F_1^{(i)}$ where N is the total number of examples and N_i is the number of examples belonging to class i.

Additional perturbation datasets

To find an inactive sequence whose motif content was similar to the strong enhancer in **Figure 4a**, we selected previously tested inactive genomic sequences with the same number and identity of motifs, and selected the sequence with the highest 6-mer similarity. We partitioned the sequences into non-overlapping blocks based on the motif positions in the inactive and strong enhancer sequences (coordinates in **Supplementary Table 6**). Then we swapped blocks individually and in combinations of either motif blocks or spacer blocks (but not both); we also moved a motif internally from its native position to its corresponding position in the other sequence before swapping blocks **(Supplementary Table 7**). Last, we moved the NRL motif in the strong enhancer to every other position that did not overlap with the CRX motif (**Supplementary Table 8**).

To test the effects of NRL, NeuroD1, RORB, and MAZ motifs, we scrambled all instances of these motifs in all strong enhancers in the Round 1 genomic library. Motifs were scrambled either individually or in combination with mutating all CRX motifs via point mutation. We assayed these sequences in the same library as Round 3 (**Supplementary Table 1**).

Regression model

To perform model interpretation, we trained a new regression CNN that predicts log2 MPRA activity directly from sequence. We updated our architecture to include residual skip connections, dilated convolutions, and first-layer exponential activation [101]. This architecture consists of three “convolution blocks,” a fully connected layer, and a final single-output node. A convolutional block consists of a convolutional layer, multiple dilated convolutional layers with residual skip connections, and then max-pooling. We trained the model at Round 4b and used the same validation set to tune hyperparameters. We trained the model using the Adam optimizer (learning rate = 0.0003, weight decay = 10^{-6}), mean squared error as a loss function, and a batch size of 128. We fit the model for 50 epochs with early stopping (patience = 10, metric = Spearman) and a custom learning rate scheduler (patience = 3, decay = 0.2). Our final model was selected by training 20 initializations and selecting the one with the highest PCC on the validation set.

Motif analysis

All motif analyses were performed using our predicted occupancy framework and μ=9 [102]. At this value, a motif with a relative K_D = 3% of the consensus site has 50% probability of being occupied. To identify individual motifs, we compute the predicted occupancy of a TF and identify the positions where the predicted occupancy is at least 0.5. To identify the total number of motifs for a TF, we sum the predicted occupancy across every position of the sequence. Cartoon sequences in **Figure 4** were generated using the predicted occupancies of our reference list of 8 TFs.

In silico global importance analysis

We used Global Importance Analysis [73] to predict the global effect of specific sequence features on MPRA activity. We generated a background distribution by dinucleotide shuffling each of our 4658 genomic sequences and predicting their activity with our regression model. Then, we injected a fixed sequence feature in a fixed location of every sequence in our background distribution, predicted their activity with the same model, and subtracted the predictions for the background distribution. The result is the predicted log2 fold change of a sequence feature on MPRA activity, and when averaged across all sequences, represents the global importance of that feature.

Nucleotide contribution scores

We predicted the contribution of each nucleotide to regulatory activity by calculating a sequence's saliency map from the regression CNN followed by the Majdandzic correction method [76]. This method has been shown to reduce noise in feature attribution maps.

Statistics and data visualization

All statistical analyses and data visualization were performed in Python with Numpy [103], Scipy [104], Pandas [105], Matplotlib [106], and Logomaker [107]. All correlations were calculated using the functions `scipy.stats.pearsonr` and `scipy.stats.spearmanr`. In all box plots, the line denotes the median, the box represents the interquartile range (25th to 75th percentile), and whiskers extend to 1.5x the interquartile range. Violin plots cover the same range as box plots, with any outliers shown as translucent dots.

Ethics

This study was performed in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. All of the animals were handled according to protocol A-3381-01 approved by the Institutional Animal Care and Use Committee of Washington University in St. Louis. Euthanasia of mice was performed according to the recommendations of the American Veterinary Medical Association Guidelines on Euthanasia. Appropriate measures were taken to minimize pain and discomfort to the animals during experimental procedures.

Data availability

Sequence data that support the findings of this study have been deposited with the NCBI Gene Expression Omnibus (GEO) with the primary accession codes GSE165812 (Round 1 library) and GSE (number pending) (https://www.ncbi.nlm.nih.gov/geo/). Processed MPRA data files used in this study are available at [https://github.com/barakcohenlab/CRX-Active-Learning].

Code availability

Code used to process all MPRA data, to train machine learning models, and to reproduce the results of this study are available at [https://github.com/barakcohenlab/CRX-Active-Learning]. Our implementation of the Generic String Kernel is available at [https://github.com/barakcohenlab/preimage]. Our fork of the Selene package with support for custom early stopping and learning rate decay is available at [https://github.com/rfriedman22/selene].

ACKNOWLEDGEMENTS

We thank Roman Garnett, Peter Koo, and Kathleen Chen for machine learning and software help; members of the Cohen laboratory for helpful discussions and critical feedback on the manuscript; Jessica Hoisington-Lopez and MariaLynn Crosby in the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing; and Brian Koebbe and Eric Martin for computing cluster support. This work was supported by National Institutes of Health grants R01 GM121755 to M.A.W.; R01 GM092910 to B.A.C.; R01 EY030075, HL149961, and MH122451 to J.C.C.; and F31 HG011431 to R.Z.F.

AUTHOR CONTRIBUTIONS

R.Z.F. conceived the project. R.Z.F., M.A.W., and B.A.C. designed the overall project. R.Z.F. performed all data processing and quality control. R.Z.F. and S.L. implemented the machine learning methods. R.Z.F., A.R., and S.L. performed computational analyses. R.Z.F. and D.G. cloned libraries and prepared samples for sequencing. C.A.M. and M.G. performed retinal dissections and electroporations. J.C.C., B.A.C., and M.A.W. supervised the project. R.Z.F., J.C.C., B.A.C., and M.A.W. prepared the manuscript with input from all authors.

COMPETING INTERESTS

B.A.C. is on the scientific advisory board of Patch Biosciences. The remaining authors declare no competing interests.

REFERENCES

1. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).
2. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

3. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).
4. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell* **56**, 575–587 (2021).
5. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**, 373–392 (2023).
6. Barolo, S. & Posakony, J. W. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes and Development* **16**, 1167–1181 (2002).
7. Alexandre, C. & Vincent, J.-P. Requirements for transcriptional repression and activation by Engrailed in Drosophila embryos. *Development* **130**, 729–739 (2003).
8. Iype, T. *et al.* The transcriptional repressor Nkx6.1 also functions as a deoxyribonucleic acid context-dependent transcriptional activator during pancreatic beta-cell differentiation: evidence for feedback activation of the nkx6.1 gene by Nkx6.1. *Mol. Endocrinol.* **18**, 1363–1375 (2004).
9. Peng, G. H., Ahmad, O., Ahmad, F., Liu, J. & Chen, S. The photoreceptor-specific nuclear receptor Nr2e3 interacts with Crx and exerts opposing effects on the transcription of rod versus cone genes. *Hum. Mol. Genet.* **14**, 747–764 (2005).
10. Martínez-Montañés, F., Rienzo, A., Poveda-Huertes, D., Pascual-Ahuir, A. & Proft, M. Activator and repressor functions of the Mot3 transcription factor in the osmostress response of Saccharomyces cerevisiae. *Eukaryot. Cell* **12**, 636–647 (2013).
11. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45**, 1021–1028 (2013).
12. White, M. A., Myers, C. A., Corbo, J. C. & Cohen, B. A. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11952–11957 (2013).

13. Stampfel, G. *et al.* Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**, 147–151 (2015).

14. Rister, J. *et al.* Single-base pair differences in a shared motif determine differential Rhodopsin expression. *Science* **350**, 1258–1261 (2015).

15. White, M. A. *et al.* A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep.* **17**, 1247–1254 (2016).

16. Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1291–E1300 (2017).

17. Carleton, J. B., Berrett, K. C. & Gertz, J. Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor α-Bound Enhancers. *Cell Syst* **5**, 333–344.e5 (2017).

18. King, D. M. *et al.* Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *Elife* **9**, e41279 (2020).

19. Friedman, R. Z. *et al.* Information content differentiates enhancers from silencers in mouse photoreceptors. *Elife* **10**, (2021).

20. Tokuhiro, S. & Satou, Y. Cis-regulatory code for determining the action of Foxd as both an activator and a repressor in ascidian embryos. *Dev. Biol.* **476**, 11–17 (2021).

21. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254–263 (2020).

22. Gisselbrecht, S. S. *et al.* Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Mol. Cell* **77**, 324–337 (2020).

23. Junion, G. *et al.* A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473–486 (2012).

24. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).

25. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

26. Atak, Z. K. *et al.* Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. *Genome Res.* **31**, 1082–1096 (2021).

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

28. Taskiran, I. I., Spanier, K. I., Christiaens, V., Mauduit, D. & Aerts, S. Cell type directed design of synthetic enhancers. *bioRxiv* 2022.07.26.501466 (2022) doi:10.1101/2022.07.26.501466.

29. Cofer, E. M. *et al.* Modeling transcriptional regulation of model species with deep learning. *Genome Res.* **31**, 1097–1105 (2021).

30. VandenBosch, L. S. *et al.* Machine learning prediction of non-coding variant impact in human retinal cis-regulatory elements. *Transl. Vis. Sci. Technol.* **11**, 16 (2022).

31. González-Blas, C. B. *et al.* Enhancer grammar of liver cell types and hepatocyte zonation states. *bioRxiv* 2022.12.08.519575 (2022) doi:10.1101/2022.12.08.519575.

32. Movva, R. *et al.* Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One* **14**, e0218073 (2019).

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

34. Sahu, B. *et al.* Sequence determinants of human gene regulatory elements. *Nat. Genet.* **54**, 283–294 (2022).

35. Penzar, D. *et al.* LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics* **39**, btad457 (2023).

36. Linder, J., Koplik, S. E., Kundaje, A. & Seelig, G. Deciphering the impact of genetic

variation on human polyadenylation using APARENT2. *Genome Biol.* **23**, 232 (2022).

37. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).
38. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
39. de Boer, C. G. & Taipale, J. Hold out the genome: A roadmap to solving the cis-regulatory code. *bioRxiv* 2023.04.20.537701 (2023) doi:10.1101/2023.04.20.537701.
40. Monarch, R. M. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. (Simon and Schuster, 2021).
41. Settles, B. *Active Learning*. vol. 18 (Morgan & Claypool Publishers, 2012).
42. Lewis, D. D. & Gale, W. A. A Sequential Algorithm for Training Text Classifiers. in *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)* 3–12 (1994).
43. King, R. D. *et al.* Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252 (2004).
44. Kanda, G. N. *et al.* Robotic search for optimal cell culture in regenerative medicine. *Elife* **11**, (2022).
45. Hie, B., Bryson, B. D. & Berger, B. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Syst.* **11**, 461–477.e9 (2020).
46. Garnett, R., Gärtner, T., Vogt, M. & Bajorath, J. Introducing the 'active search' method for iterative virtual screening. *J. Comput. Aided Mol. Des.* **29**, 305–314 (2015).
47. Oglic, D. *et al.* Active Search for Computer-aided Drug Design. *Mol. Inform.* **37**, 1700130 (2018).
48. Warmuth, M. K. *et al.* Active learning with support vector machines in the drug discovery process. in *Journal of Chemical Information and Computer Sciences* vol. 43 667–673 (2003).
49. Miao, Z. *et al.* Iterative human and automated identification of wildlife images. *Nature*

Machine Intelligence **3**, 885–895 (2021).

50. Liu, Y. *et al.* Experimental discovery of structure–property relationships in ferroelectric materials via active learning. *Nature Machine Intelligence* **4**, 341–350 (2022).
51. Zhong, M. *et al.* Accelerated discovery of CO2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
52. Singh, R. *et al.* Prioritizing transcription factor perturbations from single-cell transcriptomics. *bioRxiv* 2022.06.27.497786 (2023) doi:10.1101/2022.06.27.497786.
53. Guan, X., Li, Z., Zhou, Y., Shao, W. & Zhang, D. Active learning for efficient analysis of high-throughput nanopore data. *Bioinformatics* **39**, (2023).
54. Furukawa, T., Morrow, E. M. & Cepko, C. L. Crx, a Novel otx-like Homeobox Gene, Shows Photoreceptor-Specific Expression and Regulates Photoreceptor Differentiation. *Cell* **91**, 531–541 (1997).
55. Chen, S. *et al.* Crx, a Novel Otx-like Paired-Homeodomain Protein, Binds to and Transactivates Photoreceptor Cell-Specific Genes. *Neuron* **19**, 1017–1030 (1997).
56. Freund, C. L. *et al.* Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell* **91**, 543–553 (1997).
57. Swaroop, A., Kim, D. & Forrest, D. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. *Nat. Rev. Neurosci.* **11**, 563–576 (2010).
58. Hennig, A. K., Peng, G.-H. & Chen, S. Regulation of photoreceptor gene expression by Crx-associated transcription factor network. *Brain Res.* **1192**, 114–133 (2008).
59. Hughes, A. E. O., Enright, J. M., Myers, C. A., Shen, S. Q. & Corbo, J. C. Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors. *Sci. Rep.* **7**, 43184 (2017).
60. Murphy, D. P., Hughes, A. E., Lawrence, K. A., Myers, C. A. & Corbo, J. C. Cis-regulatory basis of sister cell type divergence in the vertebrate retina. *Elife* **8**, e48216 (2019).

61. Swain, P. K. *et al.* Mutations in the cone-rod homeobox gene are associated with the cone-rod dystrophy photoreceptor degeneration. *Neuron* **19**, 1329–1336 (1997).

62. Hsiau, T. H.-C. *et al.* The Cis-regulatory logic of the mammalian photoreceptor transcriptional network. *PLoS One* **2**, e643 (2007).

63. Campla, C. K. *et al.* Targeted deletion of an NRL- and CRX-regulated alternative promoter specifically silences FERM and PDZ domain containing 1 (Frmpd1) in rod photoreceptors. *Hum. Mol. Genet.* **28**, 804–817 (2019).

64. Oh, E. C. T. *et al.* Transformation of cone precursors to functional rod photoreceptors by bZIP transcription factor NRL. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1679–1684 (2007).

65. Ruzycki, P. A., Tran, N. M., Kolesnikov, A. V., Kefalov, V. J. & Chen, S. Graded gene expression changes determine phenotype severity in mouse models of CRX-associated retinopathies. *Genome Biol.* **16**, 171 (2015).

66. Hughes, A. E. O., Myers, C. A. & Corbo, J. C. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. *Genome Res.* **28**, 1520–1531 (2018).

67. Shepherdson, J. L. *et al.* Pathogenic variants in CRX have distinct cis-regulatory effects on enhancers and silencers in photoreceptors. *bioRxiv* 2023.05.27.542576 (2023) doi:10.1101/2023.05.27.542576.

68. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).

69. Chaudhari, H. G. & Cohen, B. A. Local sequence features that influence AP-1 cis-regulatory activity. *Genome Res.* **28**, 171–181 (2018).

70. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

71. Powers, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2**, 37–63 (2011).

72. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3521–3526 (2017).

73. Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P. & Paul, S. B. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Comput. Biol.* **17**, e1008925 (2021).

74. Loell, K. J. *et al.* Transcription factor interactions explain the context-dependent activity of CRX binding sites. *bioRxiv* 2023.03.05.531194 (2023) doi:10.1101/2023.03.05.531194.

75. Sayal, R., Dresch, J. M., Pushel, I., Taylor, B. R. & Arnosti, D. N. Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early Drosophila embryo. *Elife* **5**, (2016).

76. Majdandzic, A., Rajesh, C. & Koo, P. K. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biol.* **24**, 109 (2023).

77. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

78. Cheng, H. *et al.* Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. *Hum. Mol. Genet.* **13**, 1563–1575 (2004).

79. Mollema, N. J. *et al.* Nuclear receptor Rev-erb alpha (Nr1d1) functions in concert with Nr2e3 to regulate transcriptional networks in the retina. *PLoS One* **6**, e17494 (2011).

80. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv* 2023.03.05.531189 (2023).

81. Cuperus, J. T. *et al.* Deep learning of the regulatory grammar of yeast 5’ untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).

82. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455–463 (2022).

83. Nguyen, Q. & Garnett, R. Nonmyopic Multiclass Active Search for Diverse Discovery. *arXiv*

[cs.LG] (2022).

84. Nguyen, H. T. & Smeulders, A. Active learning using pre-clustering. in *Proceedings of the twenty-first international conference on Machine learning* 79 (Association for Computing Machinery, 2004).
85. Dagan, I. & Engelson, S. P. Committee-Based Sampling For Training Probabilistic Classifiers. in *Proceedings of the Twelfth International Conference on Machine Learning* (eds. Prieditis, A. & Russell, S.) 150–157 (Morgan Kaufmann, 1995).
86. Siddhant, A. & Lipton, Z. C. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 2904–2909 (Association for Computational Linguistics, 2018).
87. Linder, J. & Seelig, G. Fast activation maximization for molecular sequence design. *BMC Bioinformatics* **22**, 510 (2021).
88. Linder, J., Bogard, N., Rosenberg, A. B. & Seelig, G. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. *Cell Syst* **11**, 49–62.e16 (2020).
89. Garnelo, M. *et al.* Neural Processes. *arXiv [cs.LG]* (2018).
90. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*. vol. 14 (MIT Press, 2005).
91. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19498–19503 (2012).
92. Montana, C. L., Myers, C. A. & Corbo, J. C. Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. *J. Vis. Exp.* (2011) doi:10.3791/2821.
93. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198

(2016).

94. Giguère, S., Marchand, M., Laviolette, F., Drouin, A. & Corbeil, J. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC Bioinformatics* **14**, 82 (2013).

95. Giguère, S., Rolland, A., Laviolette, F. & Marchand, M. Algorithms for the hard pre-image problem of string kernels and the general problem of string prediction. *Proceedings of the 32nd International Conference on Machine Learning* 2021–2029 (2015).

96. Leslie, C., Eskin, E. & Noble, W. S. The spectrum kernel: A string kernel for SVM protein classification. in *Proceedings of the Pacific Symposium on Biocomputing, 2002* 564–575 (2001).

97. Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **21**, 2167–2180 (2011).

98. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

99. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *arXiv [cs.LG]* (2019).

100. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a PyTorch-based deep learning library for sequence-level data. *Nat. Methods* **16**, 315–318 (2019).

101. Koo, P. K. & Ploenzke, M. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nat Mach Intell* **3**, 258–266 (2021).

102. Zhao, Y., Granas, D. & Stormo, G. D. Inferring binding energies from selected binding sites. *PLoS Comput. Biol.* **5**, e1000590 (2009).

103. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

104. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

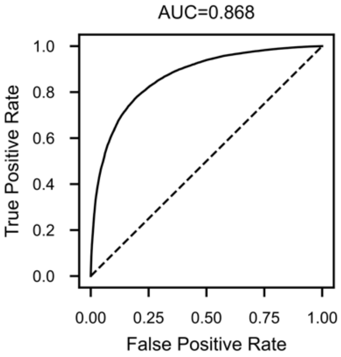
105. McKinney, W. Data structures for statistical computing in python. in *Proceedings of the 9th*

Python in Science Conference vol. 445 51–56 (Austin, TX, 2010).

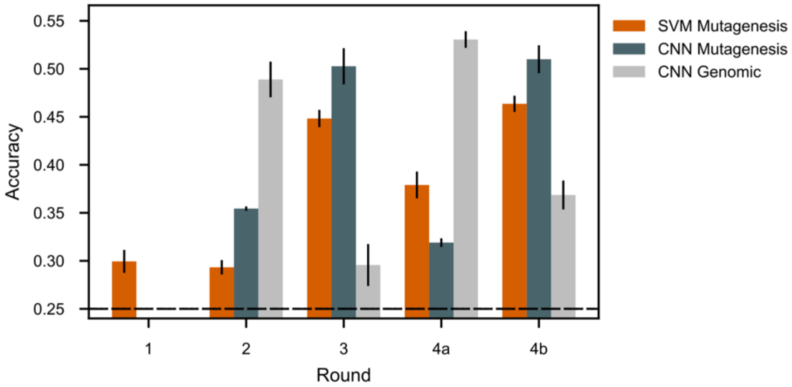
106. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

107. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

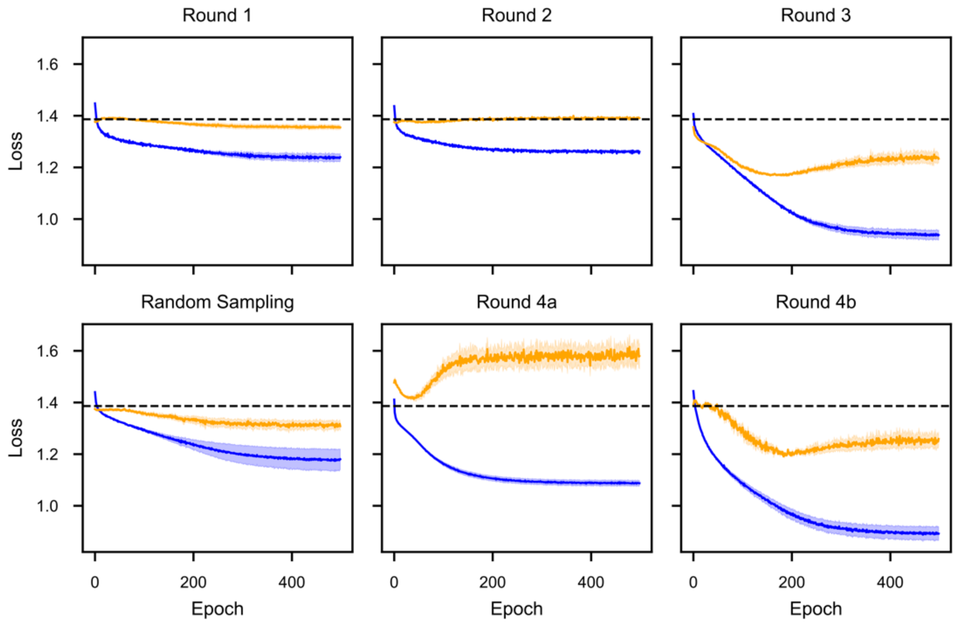
EXTENDED DATA FIGURES



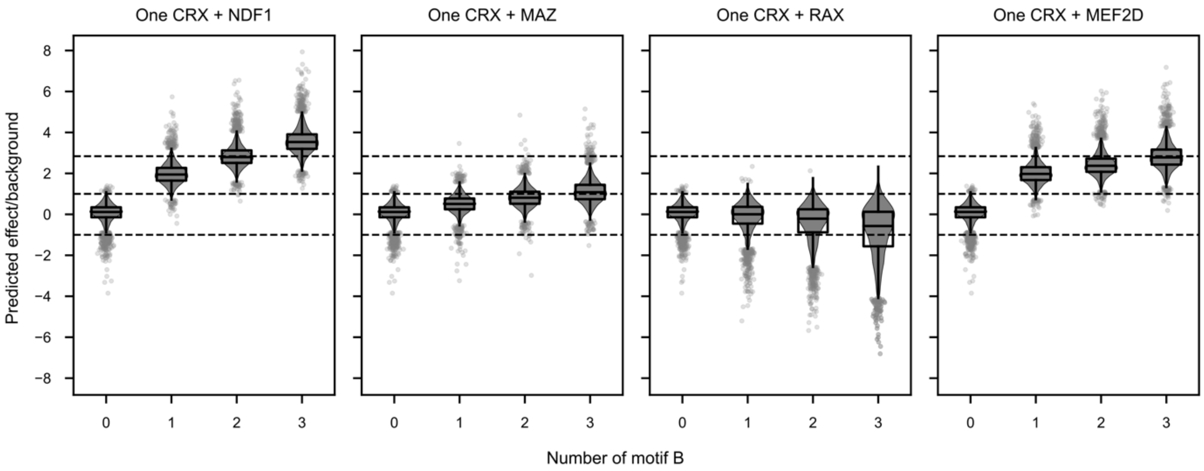
Extended Data Figure 1: Receiver operating characteristic (ROC) curve of the filter SVM. Performance is based on five-fold cross-validation.



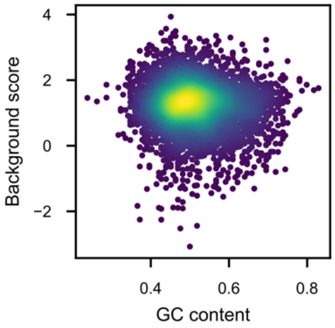
Extended Data Figure 2: Additional performance evaluation of classifiers. Accuracy is equivalent to the micro F1 score. Horizontal dashed line represents chance for both test sets. Error bars denote one standard deviation based on ten-fold cross-validation of the newly added data.



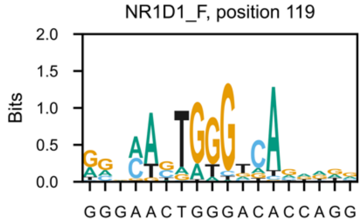
Extended Data Figure 3: CNN classifier training history for each round of learning. Blue line denotes training data, orange line denotes validation data, horizontal dashed line denotes chance. Shaded areas denote 1 standard deviation based on ten-fold cross-validation on the newly added data. The final model parameters used for each round were selected from the epoch where the validation loss is minimized.



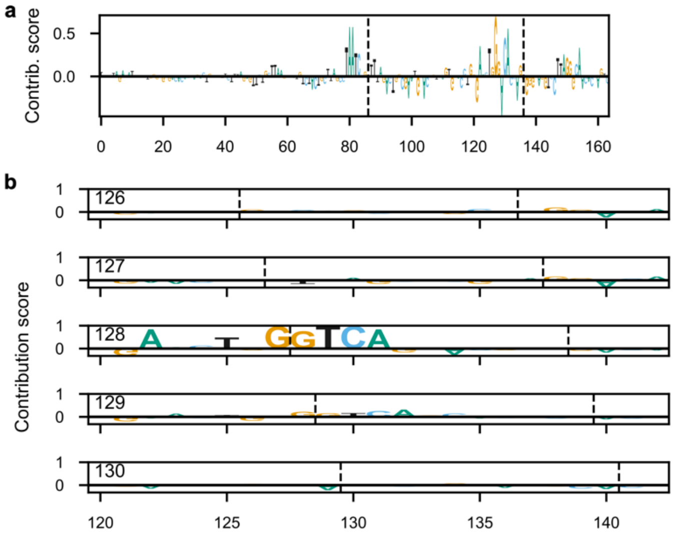
Extended Data Figure 4: Predicted effects of additional photoreceptor motifs. Predicted effect of motifs conditioned on the presence of a CRX motif (as in **Figure 3a**). Panels show the effect of 0-3 motifs of another TF in the context of one CRX motif. Horizontal dashed lines denote activity class boundaries.



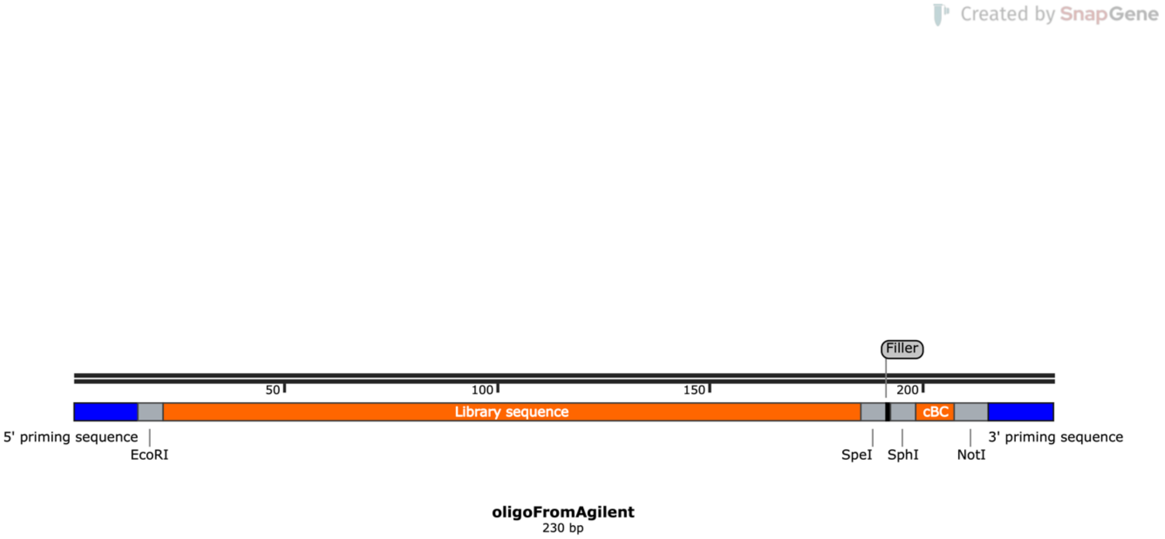
Extended Data Figure 5: Predicted activity of background distribution vs. GC content. GC content alone without CRX sites has no global effect on predicted activity. Each dot represents a different background sequence. Warmer colors denote higher point density.



Extended Data Figure 6: NR1-family motif match in the strong enhancer. The motif for NR1D1 is used as a representative motif of the NR1 TF family. X-tick labels indicate the sequence that matches the motif. The position and orientation of the motif is indicated in the title.



Extended Data Figure 7: Nucleotide contribution scores for sequence perturbations in a strong enhancer and inactive sequence. a, Relative nucleotide contribution scores when positions 86-136 from the strong enhancer (containing the NR1-family site) are swapped into the inactive sequence. **b**, Nucleotide contribution scores of the strong enhancer when the NRL motif is moved to positions ranging from 126-130. The cryptic NR1:NRL site is formed when NRL is moved to position 128. Vertical dashes denote the breakpoints of the swap or the boundaries of the sliding NRL sequence.



Extended Data Figure 8: Snapgene schematic of MPRA library oligo design.

SUPPLEMENTARY TABLES

Supplementary Table 1: Library design experimental metadata.

Supplementary Table 2: mBC metadata.

Supplementary Table 3: Oligo amplification primer pair metadata.

Supplementary Table 4: Primers used in this study.

Supplementary Table 5: Within-library quality control statistics.

Supplementary Table 6: Block coordinates for the strong enhancer and inactive sequence pair.

Supplementary Table 7: Annotations for swapping of blocks. Uppercase letter denotes the block originates from the strong enhancer twin, lowercase denotes the inactive twin. A block letter followed by an ‘x’ indicates the region is scrambled, with the exact coordinates indicated by the numbers. All coordinates are on the zero-based interval [start, stop). Note that scrambles corresponding to motifs stretch 3bp into the flanks of adjacent blocks.

Supplementary Table 8: Annotations for experiment sliding the NRL motif along the strong enhancer.