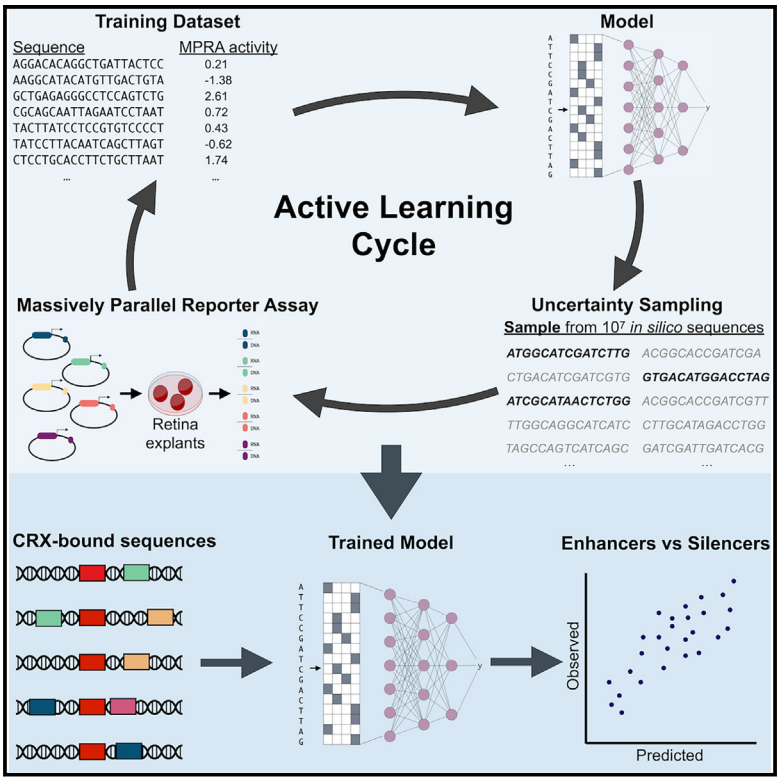
Cell Systems

Article

Active learning of enhancers and silencers in the developing neural retina

Graphical abstract



Highlights

- Transcription factor binding sites activate or repress depending on context
- Genomic examples are insufficient to learn how context affects binding sites
- Active learning iteratively generates informative new training data
- A CNN trained with active learning distinguishes activating and repressing sites

Authors

Ryan Z. Friedman, Avinash Ramu, Sara Lichtarge, ..., Joseph C. Corbo, Barak A. Cohen, Michael A. White

Correspondence

mawhite@wustl.edu

In brief

Friedman et al. introduce an active machine learning workflow to iteratively train deep learning models of regulatory DNA on successive rounds of experimental data. Using active learning with massively parallel reporter assays in the developing mouse retina, they train a convolutional neural network that successfully distinguishes between activating and repressing regulatory DNA elements that are composed of the same transcription factor binding sites.

Friedman et al., 2025, Cell Systems *16*, 101163
January 15, 2025
https://doi.org/10.1016/j.cels.2024.12.004

CellPress

Cell Systems

CellPress
OPEN ACCESS

Article

Active learning of enhancers and silencers in the developing neural retina

Ryan Z. Friedman,[1,2,4] Avinash Ramu,[1,2] Sara Lichtarge,[1,2] Yawei Wu,[1,2] Lloyd Tripp,[1,2] Daniel Lyon,[1,2] Connie A. Myers,[3] David M. Granas,[1,2] Maria Gause,[3] Joseph C. Corbo,[3] Barak A. Cohen,[1,2] and Michael A. White[1,2,5,*]

[1]The Edison Family Center for Genome Sciences & Systems Biology, Saint Louis, MO 63110, USA
[2]Department of Genetics, Saint Louis, MO 63110, USA
[3]Department of Pathology and Immunology, Washington University School of Medicine, Saint Louis, MO 63110, USA
[4]Present address: Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[5]Lead contact
*Correspondence: mawhite@wustl.edu
https://doi.org/10.1016/j.cels.2024.12.004

SUMMARY

Deep learning is a promising strategy for modeling *cis*-regulatory elements. However, models trained on genomic sequences often fail to explain why the same transcription factor can activate or repress transcription in different contexts. To address this limitation, we developed an active learning approach to train models that distinguish between enhancers and silencers composed of binding sites for the photoreceptor transcription factor cone-rod homeobox (CRX). After training the model on nearly all bound CRX sites from the genome, we coupled synthetic biology with uncertainty sampling to generate additional rounds of informative training data. This allowed us to iteratively train models on data from multiple rounds of massively parallel reporter assays. The ability of the resulting models to discriminate between CRX sites with identical sequence but opposite functions establishes active learning as an effective strategy to train models of regulatory DNA. A record of this paper's transparent peer review process is included in the supplemental information.

INTRODUCTION

The contribution of a transcription factor (TF) binding site to the activity of a *cis*-regulatory element (CRE) depends on its local sequence context, including the presence and absence of other TF binding sites.[1–6] Identical TF binding sites can occur in both enhancers and silencers[7–24] and in sequences with no activity at all. As a consequence, standard enrichment analyses of TF binding motifs have limited power to distinguish CREs with opposite activities. Models that can explain why binding sites for the same TF can activate, repress, or have no effect in different contexts would address a major challenge in ongoing efforts to understand the role of the non-coding genome in human health and disease.

Deep learning presents an opportunity to train better models of CREs that accurately predict *cis*-regulatory activity from DNA sequence and that identify critical features of local sequence context. Deep neural networks trained on epigenomic data to predict TF binding and chromatin accessibility from DNA sequence often achieve high accuracy, and they identify important sequence features underlying TF binding.[25–32] However, because TF binding per se is necessary but not sufficient for CRE activity, models that predict TF binding cannot explain differences in the activity of CREs bound by the same TFs. Models trained on direct measurements of CRE activity from massively parallel reporter assays (MPRAs)[20,33–40] can discover sequence features driving differences in CRE activity. However, these models are often less accurate than models trained to predict TF binding, because CRE activity depends on higher-order interactions between bound TFs and their associated co-factors. It is these higher-order interactions that cause identical TF binding sites to activate or repress in different sequence contexts.[2,4–6,41]

A major obstacle to learning higher-order interactions is the limited amount of training data available in the genome. The number of active CREs in the genome is small relative to the scale of training data needed to learn the combinatorial interactions among binding sites.[42,43] As a consequence, current deep learning models of *cis*-regulatory activity often fail to learn how sequence context alters the effects of TF binding motifs, and instead, they typically uncover those TF motifs with large, consistent effects on gene expression. These tend to be the same motifs identified by traditional motif-finding algorithms.

Additional training data beyond what is available in the genome can be generated by MPRAs using synthetic DNA sequences. However, the number of possible synthetic sequences vastly exceeds the number that can be feasibly synthesized and assayed, and most synthetic sequences will be uninformative. There is thus an urgent need for methods to prioritize informative training examples from the space of potential training data, thereby leveraging the capacities of MPRAs and other functional genomics assays to generate large training datasets that are not limited by what the genome alone provides.

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

We address this gap by changing the current paradigm for training models of CREs. We couple active machine learning[44–46] with synthetic biology to iteratively train models on successive rounds of informative MPRA experiments. In contrast to current approaches that rely on a single round of genomic training data, active learning offers a way to iteratively improve models by prioritizing new synthetic data based on their potential to improve the model. Active learning has been successfully applied to model metabolic networks,[47] optimize cell culture media,[48] perform *in silico* drug screens,[49–52] identify TFs that drive cellular differentiation,[53] select optimal training data for nanopore base calling,[54] and to design pooled perturbation screens.[55] Here, we apply active learning to the problem of *cis*-regulation. We use active learning to iteratively train models that learn to distinguish activating, repressing, and inactive TF binding sites in the early post-natal mouse retina, a part of the developing central nervous system that is amenable to electroporation-based, episomal reporter assays.

RESULTS

Active learning applied to enhancers and silencers in the developing retina

The retina-specific TF cone-rod homeobox (CRX) is a striking example of how similar or even identical binding sites for a single TF can have opposite effects in different CREs.[56–62] CRX binds enhancers and silencers,[13,20,63] and its activating and repressing functions are both required for terminal differentiation of photoreceptors.[10,64–74] The CRX motif is pervasive in photoreceptor open chromatin,[60,61] with over 50% of open chromatin regions containing a match to the CRX binding motif. The CRX binding site is thus by far the most enriched cell-type-specific TF binding motif in the most abundant cell type of the mouse retina.[60] Yet, the mere presence of a CRX binding site is by itself insufficient to predict whether a CRX-bound DNA sequence will activate or repress transcription.[13,16,20,61,63,75] When copies of CRX motifs in CRX-bound sequences are abolished, *cis*-regulatory activity can increase, decrease, or have no effect.[13,16,20,61,75,76] Additionally, many CRX-bound sequences exhibit little to no *cis*-regulatory activity in the developing neural retina, similar to results reported for TF-bound sequences in cell culture models.[17,77–80] The case of CRX illustrates a key challenge for models of *cis*-regulation, which is to learn how local sequence context encodes different activities of often identical binding sites for a multi-functional TF.

To address this, we implemented an iterative, active learning strategy to train models on successive rounds of informative MPRA data from synthetic DNA sequences (Figure 1A). The core of the approach is to *actively* select new training examples that are likely to improve the model in the next round. In each round, a pool of candidate synthetic sequences is generated by performing millions of *in silico* perturbations on the sequences in the current training dataset. To ensure the biological relevance of the candidate sequences, the pool is filtered to remove candidates that are not predicted to have sequence properties of open chromatin in photoreceptors (STAR Methods). The pool is then sub-sampled using the current model to prioritize those sequences whose predicted activities are the most uncertain (STAR Methods). This "uncertainty sampling" step is based on the premise that candidate sequences predicted by the model with the least confidence will be those examples most likely to improve the model in the next round.[46] Following uncertainty sampling, the selected sequences are synthesized and assayed by MPRA in explanted mouse retinas, and the new data are added to the training dataset. The model is retrained on the cumulative training data and evaluated on an independent test dataset. Models are thus iteratively optimized to achieve performance above that obtained with only a single round of genomic training data.

To carry out active learning, we used four-way classifiers that predict the probability that a given DNA sequence is a (1) strong enhancer, (2) weak enhancer, (3) inactive sequence, or (4) silencer in photoreceptors. These discrete classifiers facilitated the uncertainty sampling step by enabling straightforward calculations of uncertainty from the probabilities assigned to each activity class. Our initial classifier was a modified *k*-mer support vector machine (SVM), trained on a genomic dataset[20] (training round 1). We used the SVM for two rounds of active learning (rounds 2 and 3). We then switched to a convolutional neural network (CNN) classifier for the last round of active learning (round 4), owing to the greater expressivity of this architecture. After completing active learning using discrete classifiers, we used the final dataset to train a regression CNN. The regression CNN was used to make quantitative predictions of CRE activity and for sequence interpretation.

Model performance more than doubles with active learning

The SVM classifier was initially trained on MPRA data from a library of 8,879 wild-type and mutant genomic sequences, each of which was centered on an intact or mutated CRX motif. After the initial training round (round 1), the SVM performance on an independent test set did not exceed the accuracy expected from random guessing, which shows that the genomic dataset alone was not enough to learn how the sequence context encodes activating, repressing, and inactive CRX motifs (Figure 1B). After two rounds of active learning, the performance of the SVM nearly doubled (Figure 1B, round 3). To test whether the improvement in model performance was due to active learning and not merely to more training data, we compared models trained on data selected by uncertainty sampling versus data selected by random sampling (STAR Methods). Uncertainty sampling resulted in a model that outperformed a model trained on data generated by random sampling (Figure 1C), showing that active learning produces more informative training data and that the improved performance is not merely due to the increased size of the dataset.

We performed the initial rounds of active learning with the SVM because a CNN trained on the genomic dataset alone did not generalize to the validation data, likely because the training dataset was too small (Figure S1A). After generating more training data in rounds 2 and 3, we successfully trained a CNN four-way classifier. We used this CNN for the final round of active learning (round 4) because this architecture can flexibly encode higher-order features that may not be captured by the SVM. When we trained the CNN on the round 4 dataset, performance increased by 53% relative to the round 3 dataset (Figure 1D). We tested the effect of uncertainty sampling versus

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

CellPress
OPEN ACCESS

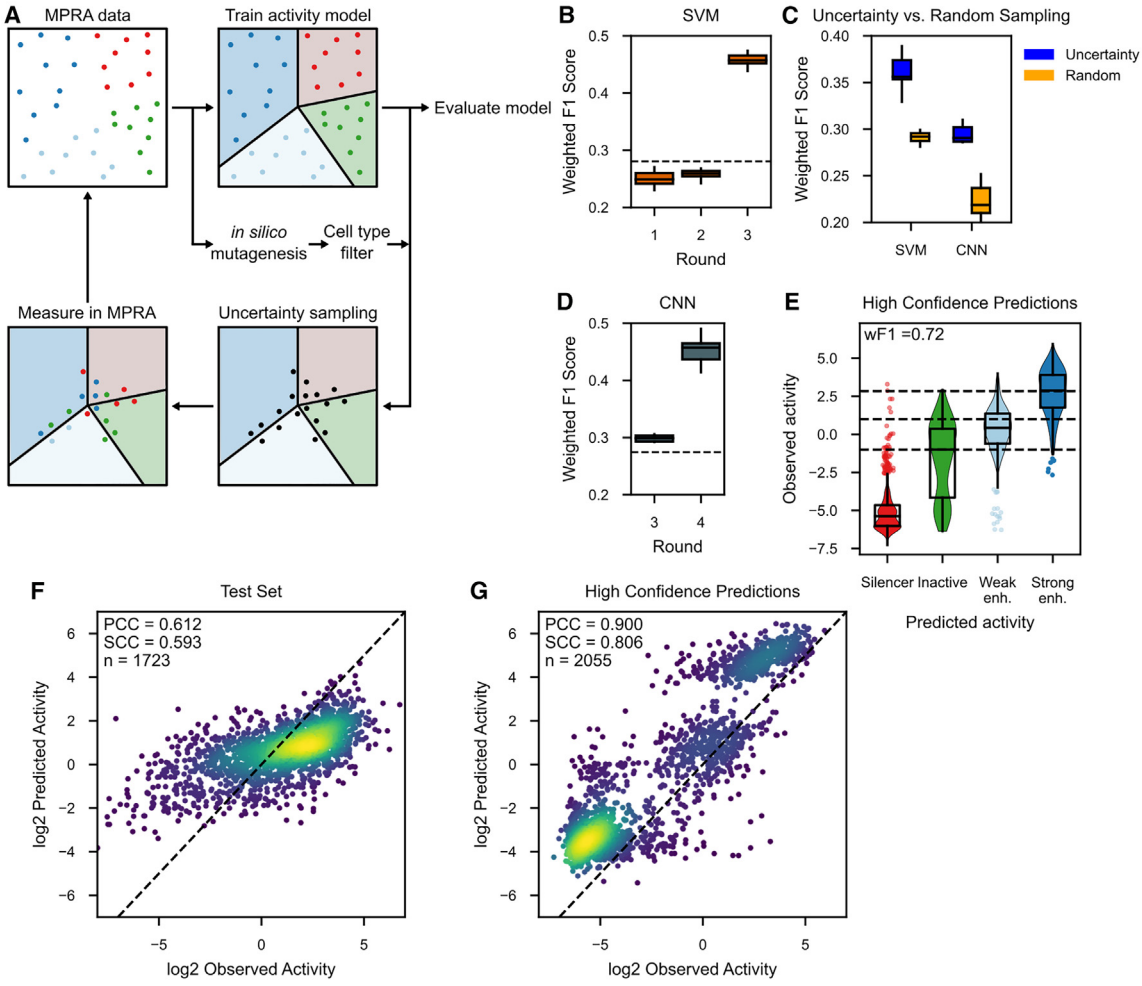


Figure 1. Iterative machine learning improves predictions of *cis*-regulatory activity

(A) Summary of active learning approach. 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). After generating a filtered pool of candidate sequences, those predicted with high uncertainty under the current model (black dots closest to the margins) are synthesized, measured by MPRA, and added to the training data for the next round of model fitting.

(B) Iterative improvement of the SVM classifier over two rounds of active learning. Horizontal line represents accuracy expected from random guessing. Boxplots show the performances of 10-fold cross-validation of the newly added data.

(C) Performance of SVM and CNN classifiers when trained on new data obtained by either random or uncertainty sampling that were added to the training data from rounds 1 and 2. Boxplots show performances of 10-fold cross-validation.

(D) Iterative improvement of the CNN classifier performance over one round of active learning. Horizontal dashed line and boxplots are as in (B).

(E) Observed activity of synthetic sequences (n = 2,055) predicted with high confidence, stratified by predicted activity class. Horizontal dashed lines correspond to cutoffs for the activity classes. Enh., enhancer; wF1, weighted F1 score.

(F and G) Observed activity versus activity predicted by the regression CNN for (F) the held-out test set of CRX-bound genomic sequences (n = 1,723) and (G) the high-confidence sequences (n = 2,055) of (E). Diagonal line indicates x = y.

PCC, Pearson correlation coefficient; SCC, Spearman correlation coefficient. Warmer colors denote higher point density.

See also Figure S1.

random sampling on the performance of the CNN classifier, and again we found that a model trained on data selected by active learning outperformed a model trained on data that included randomly sampled sequences (Figure 1C).

The global performance measures reported above are based on all model predictions, regardless of the uncertainty of those predictions. However, a key advantage of our approach is that there is an uncertainty estimate for each prediction, and thus high-confidence predictions can be separated from low-confidence predictions. To test the accuracy of the uncertainty estimate, we synthesized and assayed 2,055 new synthetic sequences whose activity was predicted by the CNN classifier with high confidence (STAR Methods). We found that 72% of these sequences were predicted correctly at round 3 (Figure 1E) versus only 29% of the test set predictions, which included both low- and high-confidence predictions. This shows that the uncertainty estimate successfully captures model confidence and that it is effective for *de novo* enhancer design by

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

focusing attention on the predictions that are most likely to be correct.

After three rounds of active learning, we used the cumulative dataset to train a regression CNN, since a model that makes quantitative predictions of CRE activity is more useful for interpreting the role of individual TF binding sites. The regression CNN achieved a Pearson correlation coefficient (PCC) of 0.61 on an independent test set of genomic CRX-bound sequences[76] (n = 1,723, Figure 1F; STAR Methods). Performance was even higher for the high-confidence dataset (PCC = 0.90, Figure 1G). A regression CNN trained only on the original genomic data (round 1) had much lower performance (PCC = 0.29, Figure S1B). This lower performance is close to the limit of what can be achieved by training with genomic data alone, because nearly 80% of CRX-bound sequences are in the round 1 dataset or the test set. Our results show that with active learning, models can be improved after exhausting genomic training examples.

Active learning is influenced by *in silico* sequence generation strategy and measures of uncertainty

While implementing active learning, we tested alternative strategies for two key steps in the active learning cycle: *in silico* sequence generation and uncertainty sampling. To generate new sequences in the first active learning round (round 2), we performed uniform, random *in silico* mutagenesis of the original genomic sequences. This led to only a slight performance improvement in the SVM (Figure 1B), and we hypothesized that random mutagenesis likely produced many uninformative perturbations of small effect. For round 3, we implemented a motif-centric perturbation strategy, reasoning that adding, subtracting, or moving motifs known to be important in photoreceptors would generate more informative training sequences (STAR Methods). We observed a much greater increase in model performance in round 3, suggesting that the motif-perturbation strategy was more effective (Figure 1B). In a direct comparison using equally sized training datasets, we found that the motif-centric perturbation strategy was much more effective at generating informative training data (Figure S1C). Because the motif-perturbation strategy was effective, we also used it to generate sequences in round 4.

As the measure of model uncertainty, we used Shannon entropy in rounds 2 and 3. When training the CNN classifier at round 3, we observed that the model performed somewhat better on strong enhancers (F1 score = 0.28) than it did on silencers (F1 score = 0.17, Figure S1D). We hypothesized that a sampling strategy targeting potential silencers might improve performance. Thus, in round 4, we tested a second uncertainty sampling strategy, margin sampling, in parallel with entropy sampling. Shannon entropy reaches its maximum value when the classifier model is equally uncertain about all four activity classes (strong enhancer, weak enhancer, inactive, and silencer). Thus, entropy sampling selects candidate sequences for which the model makes no strong predictions. In contrast, margin sampling prioritizes examples for which the two most likely activity classes have similar probabilities, while allowing probabilities for the other two activity classes to be low (STAR Methods).

Using margin sampling, we targeted silencers by prioritizing sequences for which "silencer" was one of the two most probable classes. In parallel, we also generated additional data using entropy sampling. With further entropy sampling (round 4a), we observed substantial improvement in the classification of silencers, but this came at the expense of nearly all predictive power for strong enhancers (Figure S1D). This suggests that a third cycle of entropy sampling caused an episode of "catastrophic forgetting" in which a model forgets previously learned information when learning new information.[81] With margin sampling (round 4b), the CNN improved its performance on silencers while maintaining its performance on strong enhancers (Figure S1D). These results suggest that entropy sampling works well in early rounds of active learning when a model is relatively naive, but margin sampling works well in later rounds when it is necessary to improve model performance on certain classes of predictions.

The model learned to distinguish CRX motifs with different effects

After generating three rounds of new data using active learning with the classifiers, we used the cumulative dataset to train the regression CNN (Figure 1F), and the regression model was subsequently used to analyze CRX motifs in enhancers and silencers. In previous work, we found that a 4A>C mutation in the CRX motif abolishes binding[82] and tends to cause loss of activation in enhancers and a loss of repression in silencers.[20,76] We therefore asked whether the regression CNN learned to distinguish activating from repressing CRX sites, as defined by the change in CRE activity when the CRX motif is mutated by 4A>C. To generate model predictions for specific CRX motifs, we used the CNN to assign importance scores to the CRX motifs in every genomic sequence of the test set used to evaluate the regression CNN (STAR Methods; Figures S2A and S2B). We interpreted positive importance scores as predicting activating CRX motifs and negative importance scores as predicting repressing motifs. Consistent with these predictions, sequences with CRX motifs assigned positive importance typically lost activity when the motifs were mutated, while the opposite was true for sequences with CRX motifs assigned negative importance (Figure 2A). The model also correctly predicted non-functional CRX sites. CREs with CRX importance scores near zero exhibited only small changes in activity when CRX sites were mutated, despite those motifs being high-scoring matches for the CRX position weight matrix. Thus, using training data generated by active learning, the model learned to distinguish among activating, repressing, and inactive CRX motifs. These effects could not have been discovered by standard motif analyses because CRX motifs with different effects all match the CRX position weight matrix equally well, and their sequences are often identical.[13,16,20]

We next examined whether the model learned the known repressive effect of homotypic clusters of CRX sites. We previously showed that synthetic and genomic sequences with multiple copies of the CRX motif are often repressive.[16,20,83] Using an *in silico* perturbation analysis,[84] we quantified the predicted effect of increasing the number of CRX motifs from 1 to 4 in a set of 4,658 randomly generated background elements (Figures 2B and S2C). The model correctly predicted the repressive effect of increasing the number of CRX sites in a sequence, matching the results of previous experiments showing that most synthetic sequences with four CRX sites are repressive.[16,83]

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

CellPress
OPEN ACCESS

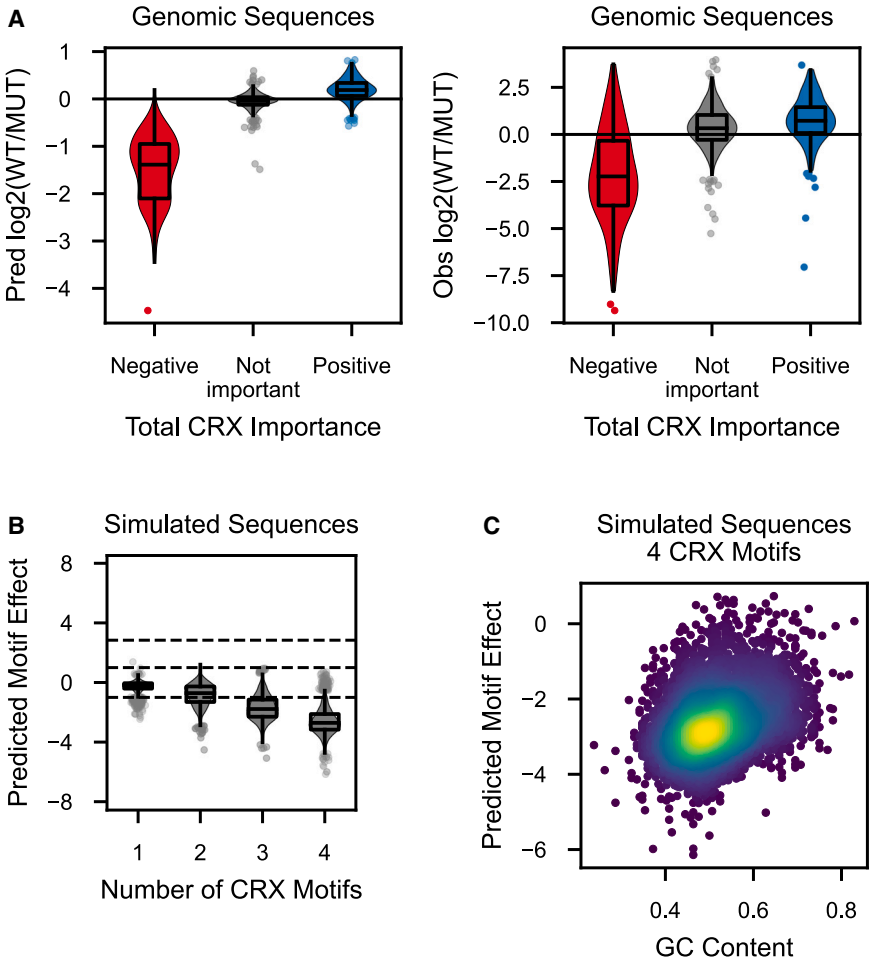


Figure 2. Regression CNN discriminates among activating, inactive, and repressive CRX motifs

(A) Predicted (left) and observed (right) effects of mutating CRX motifs in genomic CRX-bound sequences, stratified by CRX importance scores. n = 173 negative importance, 388 not important, 173 positive importance. WT, wild-type CRX motifs; MUT, mutated CRX motifs; Pred, predicted; Obs, observed.

(B) Predicted repressive effect of increasing numbers of CRX motifs in 4,658 simulated sequences. Horizontal dashed lines denote activity class boundaries.

(C) Predicted effect of background GC nucleotide content on the activating or repressive effect of adding four CRX motifs to a sequence (n = 4,658 simulated sequences). Each dot represents a different background sequence. Warmer colors denote higher point density.

See also Figure S2.

Finally, the model captured the positive influence of high guanine and cytosine (GC) sequence context on the activities of CRX sites. The importance assigned by the model to four CRX sites in the *in silico* analysis increased as the surrounding GC content increased (Figure 2C, PCC = 0.33), even though GC content itself is not predicted to independently influence activity (Figure S2D). Among genomic CRX chromatin immunoprecipitation sequencing (ChIP-seq) peaks with varying numbers of CRX sites,[13] the Pearson correlation between measured activity and GC content is 0.23. Taken together, these results show that the model successfully learned the sequence context that distinguishes activating CRX sites in enhancers and repressive CRX sites in silencers, a crucial feature of *cis*-regulation in developing photoreceptors.

The model learned the effects of other TF motifs enriched in photoreceptor CREs

Sequences in the training dataset all contain CRX motifs, but most sequences also contain motifs for other cell-type-specific TFs that interact with CRX. We examined whether the model learned to distinguish functionally distinct instances of motifs for these additional TFs. We focused on motifs for six lineage-specific TFs whose binding sites are enriched in CRX-bound CREs.[20] To test the model, we used a dataset of CRX and non-CRX motif mutations in a set of 29 strong enhancers (STAR Methods). We compared the model-assigned importance scores for each non-CRX motif against the measured effect of mutating that individual motif by scrambling it. Motif-level importance scores were highly correlated with the measured effects for motif mutations for these six additional TFs (PCC = 0.684, Figure 3A). Sites for the retinoid-related orphan receptor beta (RORB) in particular were assigned a wide range of importance scores by the model, and they exhibited a corresponding range of effect sizes when mutated. To further examine the predicted effects of RORB motifs, we increased our sample size by considering not only the single RORB motif mutations of the 29 strong enhancers but also RORB motif mutants made in the presence of mutations of other motifs (n = 258). We found that relative affinities of the RORB motifs corresponded with the importance scores assigned by the model and that mutations of higher affinity sites were predicted by the model to have larger effects. The predicted effects of the motif mutations correlated with the measured effects (PCC = 0.677, Figure 3B), demonstrating that the model learned to distinguish high- and low-affinity RORB motifs.

The model learned the relative effect sizes of motifs for RORB and neural retina leucine zipper (NRL) on the activity of sequences containing a CRX motif. An *in silico* perturbation analysis predicted that adding RORB motifs to a sequence with one central CRX motif has a stronger positive effect on *cis*-regulatory activity than the addition of NRL motifs (Figures 3C and S3A). These predictions are consistent with experimental data showing that the loss of RORB motifs generally causes a greater reduction in activity than the loss of NRL motifs (Figure 3D). The model also correctly predicted the repressive effects of sites for the transcriptional repressor growth factor independent 1 (GFI1) (Figure S3B), consistent with our previous finding that GFI1 sites

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

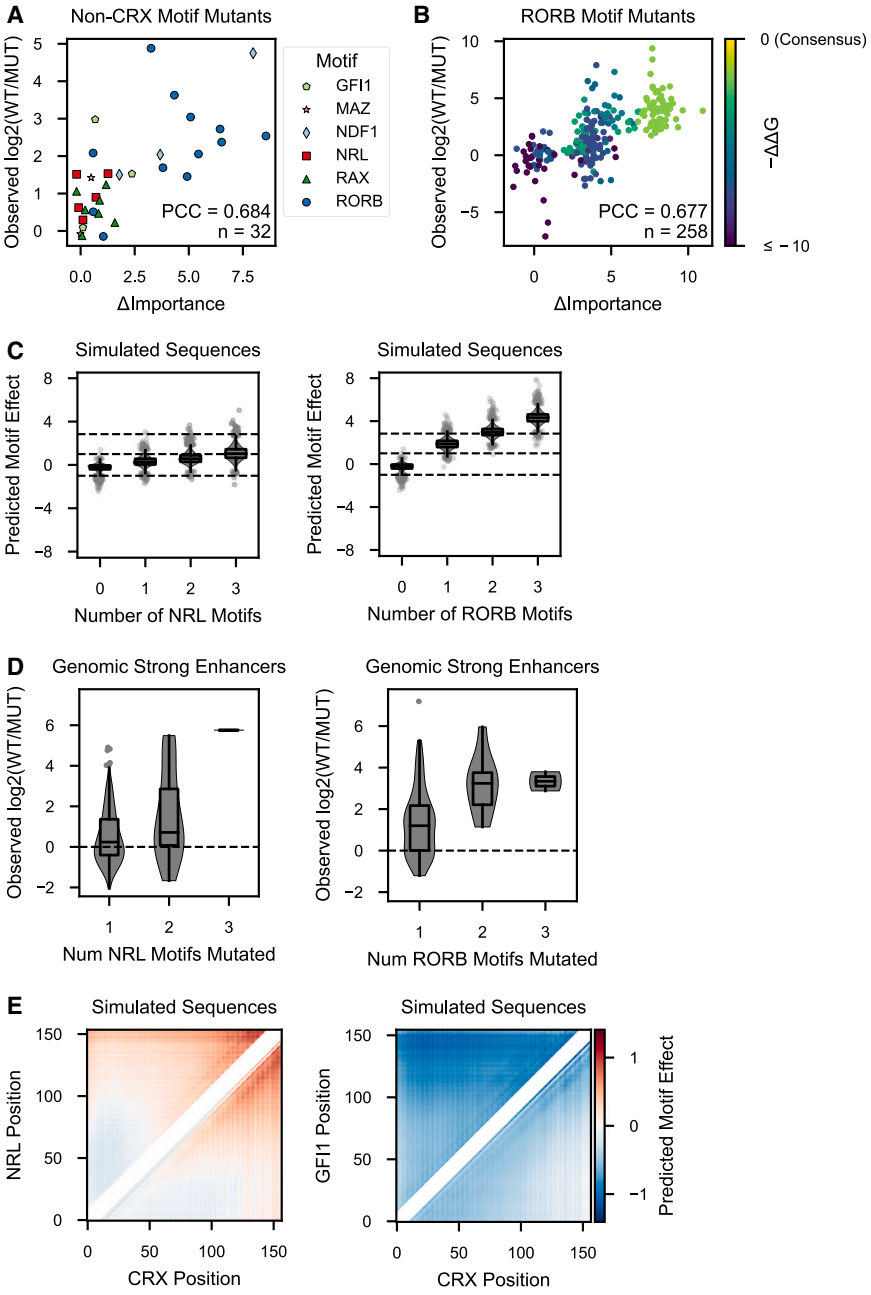


Figure 3. Regression CNN learned the effects of other TF motifs

(A) Change in predicted importance of non-CRX motifs versus observed effect when motifs are individually mutated in a set of strong enhancers (n = 32 single motif mutants). WT, wild-type motifs; MUT, mutated motifs.

(B) Change in predicted RORB motif importance versus observed effect motif mutants in a set of strong enhancers (n = 258 mutant sequences). Predicted motif importance corresponds with motif affinity, indicated by the color map representing the motif affinity relative to the consensus binding sequence, in arbitrary units of relative Gibbs free energy (ΔΔG).

(C) Predicted effect of adding NRL (left) or RORB (right) motifs to simulated sequences containing one CRX motif (n = 4,658). Horizontal dashed lines denote activity class boundaries. Zero on the x axis denotes the effect of one CRX motif.

(D) Observed effects of scrambling all motifs for NRL (left, n = 193 one motif, 25 two motifs, 1 three motifs) or RORB (right, n = 178 one motif, 9 two motifs, 1 three motifs) in a set of genomic strong enhancers,[20] stratified by the number of motifs scrambled. Dashed line indicates no effect.

(E) Predicted effects of a CRX motif and an NRL (left) or GFI1 (right) motif at all possible positions. Each pixel corresponds to the mean predicted effect of the two motifs in 4,658 different background sequences. White diagonal denotes excluded arrangements where the motifs would have overlapped. The basal promoter is at position 164. Color bar indicates mean predicted motif effects across background sequences in arbitrary units.

See also Figure S3.

are enriched in genomic CRX-bound silencers.[20] The results above show that the model learned the relative average contributions of several different cell-type-specific TF binding motifs. Explicit information about TF binding motifs was not provided as part of the model training but was instead learned by the model by iteratively training on informative data consisting only of DNA sequence and measured activity. These results validate our active learning strategy as an effective method to generate the informative training data to learn the complex interactions between TF binding motifs that determine cell-type-specific *cis*-regulatory activity.

While the global performance of the model shows that it has not yet learned all the properties of CREs in developing photoreceptors, the results above indicate that the model accomplished the key goal of learning to distinguish the different effect motifs for multiple cell-type-specific TFs. This suggests that the model in its current state can serve as a hypothesis generator about different features of photoreceptor CREs. We generated hypotheses about motif order and spacing by conducting two *in silico* perturbation analyses that systematically varied the positions of a CRX motif and either an NRL motif or a GFI1 motif (STAR Methods). This analysis predicts that the activity of CRX and NRL motifs increases as they are moved closer to the basal promoter, as well as a synergistic effect between CRX and NRL at certain spacings (Figure 3E, left). The model predicts that sequences with GFI1 and CRX motifs are largely repressive, except when the CRX motif is within ~65 bp of the basal promoter, and the GFI1 motif is placed in a more distal position (Figure 3E, right). However, when the GFI1 motif is placed close to a promoter-proximal CRX motif, the sequence activity is predicted to again be repressive, suggesting that the model infers that GFI1 acts through short-range repression.[85] This analysis shows how the model can be used to

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

generate hypotheses about complex, higher-order interactions between TF binding sites, which would be difficult to identify in the absence of a guiding model.

The model identifies discriminative features between CREs with similar TF binding sites

Many TF-bound DNA sequences that reside in open chromatin lack *cis*-regulatory activity, despite having occurrences of binding motifs that are similar or identical to motif occurrences in active CREs.[13,17,77,79,86,87] Accurate models of CREs should be able to identify sequence features that are necessary and sufficient to discriminate between inactive and active sequences with similar or identical TF binding sites. We used the regression CNN to identify functional differences between an active and an inactive photoreceptor-accessible chromatin sequence. Each sequence had one copy of the CRX motif and one copy of a motif for NRL, a rod-specific TF (Figure 4A). While the CRX motifs differed slightly between the two sequences, the NRL motifs were identical. There were no additional motifs known to be enriched in CRX-bound strong enhancers[20] present in either sequence.

Using the regression CNN to predict the importance of each nucleotide in these sequences, we found significant differences between the inactive and the active sequences (Figure 4A). Only the NRL site in the strong enhancer was scored as important, while an identical NRL site in the inactive sequence was scored as not important. In the strong enhancer, the NRL motif was flanked by high-importance nucleotides that form a near-optimal match to a nuclear receptor 1 (NR1)-family motif (Figure 4B). This motif was not detected as globally enriched among CRX-bound strong enhancers in a prior motif enrichment analysis.[20] While NR1-family TFs include factors that are known to interact with CRX, NRL, and other cell-type-specific TFs in photoreceptors,[59,88] the identified motif is recognized by a clade of nuclear receptor DNA-binding domains that is distinct from that of more well-characterized, photoreceptor-specific nuclear receptor TFs. We tested the model's predictions of important positions in the strong enhancer using an MPRA-based perturbation analysis (Figure 4C). Scrambling the CRX or NRL motifs each decreased activity, and scrambling both sites together abolished nearly all activity, showing that both the CRX and NRL motifs are necessary for the function of the strong enhancer. Scrambling the NR1-family motif also led to a near-total loss of activity, while perturbations to any other region of the enhancer had little or no effect on activity. These results show that all three motifs contribute to the activity of the strong enhancer, as predicted by the model.

We next tested whether the NR1-family motif was sufficient to activate the inactive sequence. We swapped the sub-sequence containing the NR1 motif into the inactive sequence, producing a chimeric sequence that was nearly as active as the original strong enhancer (Figure 4D). Other chimeric sequences that did not include the region with the NR1-family motif did not lead to activation. Together with the results above, these experiments show that the NR1 motif is both necessary for activity in the strong enhancer and sufficient to confer activity on the inactive sequence. The model thus discovered a functionally important motif that had failed to reach statistical significance in a prior global enrichment analysis.

Finally, we found an interaction between the NRL and NR1 motifs in the strong enhancer. When the NRL motif was moved away from the NR1 motif, the NRL motif lost its predicted importance (Figure S4A). Interestingly, when the NRL motif was placed at position 128, it disrupted the original NR1-family motif, but it created a cryptic NR1:NRL dimer motif to which the model assigned high importance (Figures S4A and S4B). Corresponding experimental perturbations bear out the model predictions, showing a loss of activity when NRL is moved out of its native position, while activity was restored upon creation of the cryptic NR1:NRL site (Figure S4A). Thus, the model learned to identify an interaction between two non-CRX motifs.

We examined a second active/inactive sequence pair, each containing a central CRX motif and a motif for the nuclear receptor RORB (Figure 4E). In the inactive sequence, the model predicted that neither CRX nor RORB motif was important. Notably, in the strong enhancer, the model predicted that the CRX motif was not important, but the RORB motif was. The model also predicted that the RORB motif would remain active when positioned 3′ of the CRX motif but that it would lose activity when positioned 5′ of the CRX motif. These predictions were confirmed by experimental motif perturbations (Figure S4C) and by placing the RORB motif in different positions of the strong enhancer (Figure 4F). Furthermore, tests of chimeric sequences showed that swapping the strong enhancer RORB motif into the inactive sequence was sufficient to confer high activity only if the RORB motif was placed 3′ of the CRX motif (Figure 4G). These results show that the model correctly identified a general positional requirement of the RORB motif in this strong enhancer. Taken together, these analyses of active/inactive sequence pairs confirm that the model learned critical features of sequence context that distinguish motifs for the same TF that have different effects.

Compared with random sampling, uncertainty sampling oversamples active training examples

During the course of active learning, we found that sequences selected by uncertainty sampling were enriched for active sequences (i.e., enhancers and silencers) relative to genomic sequences or synthetic sequences picked by random sampling (Figure 5A). This was unexpected, because the activities of the *in silico*-generated sequences are not known when they are sampled. In the first two active learning rounds (rounds 2 and 3), uncertainty sampling selected sequences that, upon measurement, were enriched for weak and strong enhancers. Notably, about half of the new sequences in round 3 were strong enhancers, which was more than double their frequency in the randomly sampled set. In round 4, silencers were the most oversampled activity class. This was true for both margin sampling (round 4b), which targeted sequences that might be silencers, as well as for entropy sampling (round 4a), which sampled sequences based on model uncertainty across all four activity classes. These results show that uncertainty sampling oversamples active sequences as a matter of course, even though the sequence activities are not known when they are picked.

To further explore this phenomenon, we performed retrospective simulations using a published genome-wide MPRA dataset for enhancer activity in K562 cells.[89] We trained a binary classifier CNN to distinguish the most active enhancers (top 20%)

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

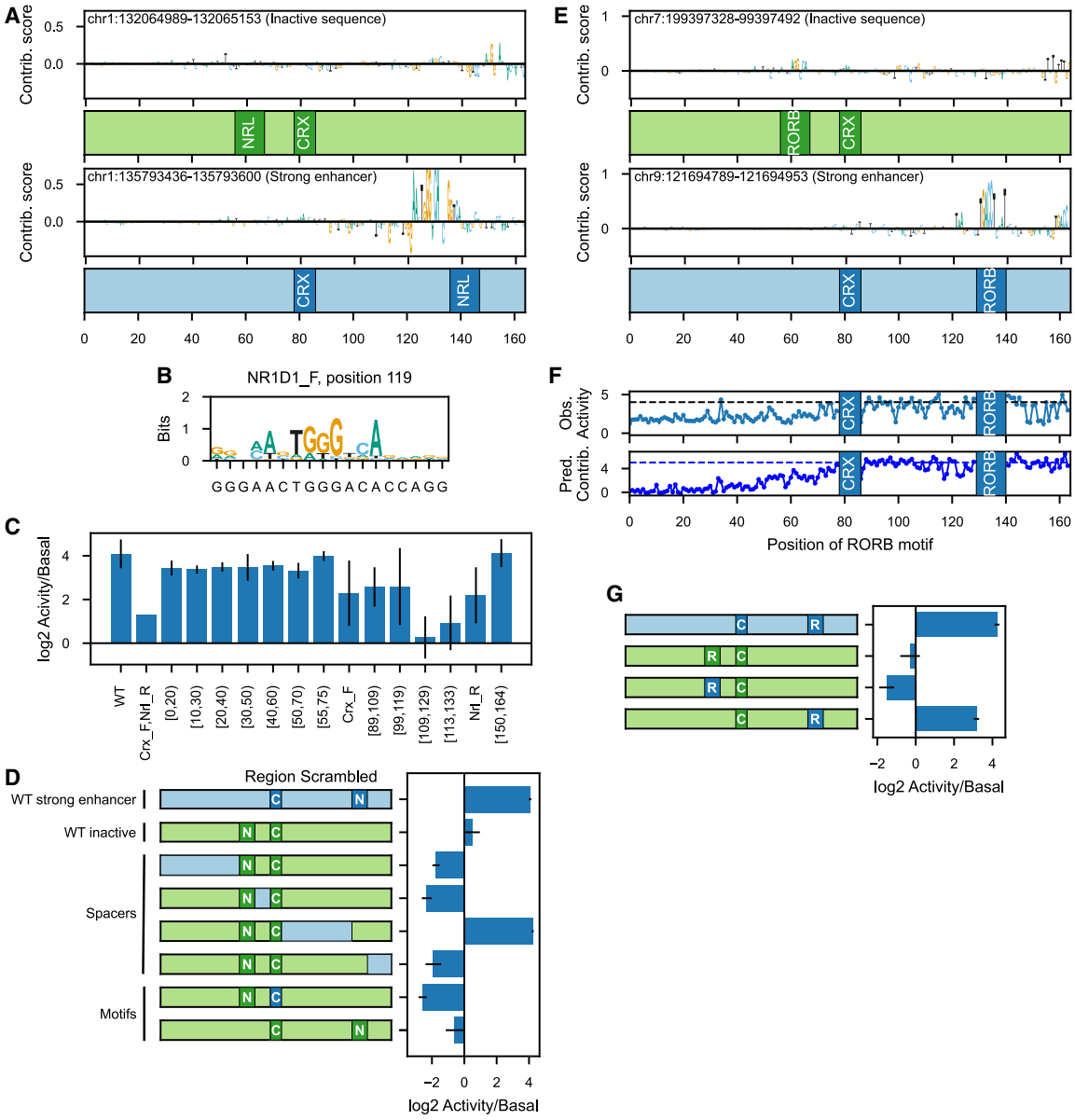


Figure 4. Regression CNN identifies discriminative sequence features between active and inactive sequences

(A) Relative nucleotide contribution scores and locations of known motifs for an inactive sequence (top) and a strong enhancer (bottom).

(B) NR1-family motif match in the strong enhancer. X-tick labels show the motif sequence in the enhancer. The position and orientation of the motif is indicated in the title.

(C) Observed effect of scrambling different components of the strong enhancer. Error bars show standard deviation of 3–5 independent scrambles, or 4 independent replicates of wild type.

(D) Effect on activity of swapping strong enhancer regions into the inactive sequence. Cartoons show the chimeric constructs with colors matching (A). Error bars show standard deviation of one sequence across 3 replicates. C, CRX motif; N, NRL motif.

(E) Relative nucleotide contribution scores and locations of known motifs for a second inactive sequence (top) and strong enhancer (bottom).

(F) Observed effect on activity of moving the RORB motif within the strong enhancer (top) and predicted contribution of the RORB motif to activity (bottom). Horizontal dashed lines represent the activity (top) and motif contribution (bottom) of the wild-type sequence.

(G) Effect on activity of swapping the strong enhancer RORB motif into the inactive sequence at different positions. Error bars show standard deviation of one sequence across 3 replicates. C, CRX motif; R, RORB motif.

See also Figure S4.

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

CellPress
OPEN ACCESS

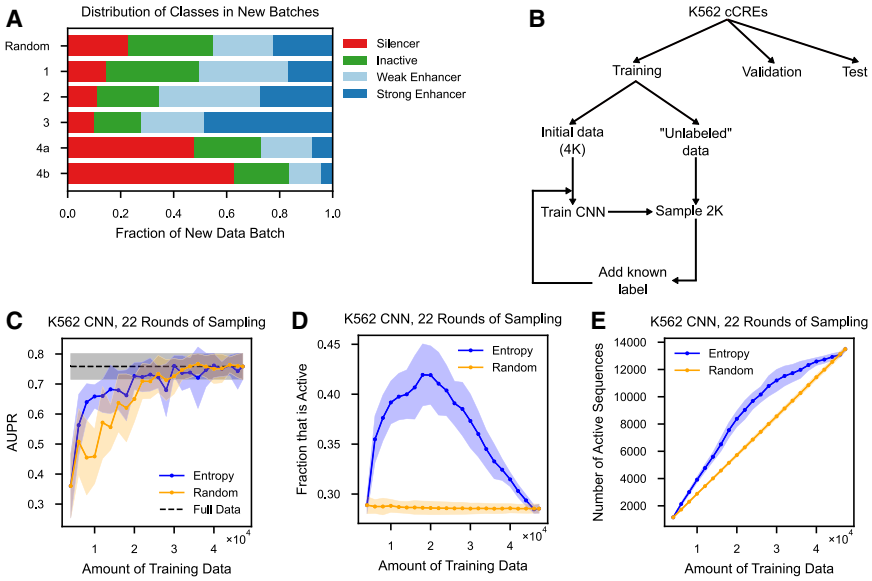


Figure 5. Entropy sampling oversamples active sequences

(A) Distribution of activity classes of sampled sequences in each round of random or uncertainty sampling. Round 1 is the initial genomic training dataset.

(B) Schematic of entropy sampling benchmarking analysis using K562 candidate CREs (cCREs). At each round, the validation chromosomes are used to monitor CNN training, and the test chromosomes are used to evaluate final performance.

(C–E) Comparison of multiple rounds of entropy versus random sampling for (C) CNN performance, (D) fraction of cumulative training data that are active sequences, and (E) total number of active sequences sampled. Lines denote the mean from 10-fold cross-validation, shaded areas denote one standard deviation, and black denotes CNN performance with the full dataset.

from inactive sequences (bottom 50%). There were no silencers in this dataset. After holding out validation and test datasets, we randomly assigned 4,000 sequences as initial training data out of ~46,000 training sequences. The remaining training sequences were treated as "unlabeled" data (Figure 5B). We then used entropy sampling or random sampling to pick batches of 2,000 at a time until all available data were exhausted.

Entropy sampling was more efficient and outperformed random sampling in the early rounds of training, when there still was a large pool of unlabeled examples available. After five rounds of entropy sampling, model performance doubled, and performance approached the upper bound after using only ~40% of the full training dataset (Figure 5C). As was observed with active learning in the mouse retina, entropy sampling oversampled unlabeled active sequences, until those sequences were exhausted from the pool of potential examples (Figures 5D and 5E). The performance of the model improved most in rounds where entropy sampling produced a dataset enriched for high-activity sequences. These results confirm that a dataset enriched for active sequences is more informative for model training. In these benchmarking experiments with an existing dataset, the pool of potential training examples was limited, and eventually high-activity sequences were exhausted. However, in real-world scenarios, the training dataset can be indefinitely increased using synthetic DNA sequences and functional genomics technologies such as MPRAs. Thus, informative, active sequences can be added to the training data until the desired level of model performance is achieved.

DISCUSSION

We have presented an active machine learning framework to learn the sequence context that causes different occurrences of motifs for the same TF to activate, repress, or show no activity. We showed that active learning more than doubled the performance of models trained on a single round of genomic data alone and that a model trained by active learning makes accurate predictions, recapitulates prior knowledge, and reveals novel sequence features necessary and sufficient for enhancer activity. Critically, the model distinguishes between binding sites with identical sequences but opposite functions, thus achieving the major goal of this study. Our work demonstrates how active learning leverages the capacities of DNA synthesis and functional genomic assays to generate successive rounds of informative training data. With this approach, we overcome a critical limitation of existing genomic datasets, which is that they include too few natural training examples, and these are not sufficient to learn the complex, higher-order interactions that distinguish activating and repressing binding sites for the same TF. Given the continually decreasing cost of DNA synthesis and the ever-growing capacities of functional assays, active machine learning has broad potential across a range of applications, including large-scale perturbation studies of CREs and model-guided design of synthetic regulatory DNA elements.

Our results indicate that active learning may be a more efficient approach because it generates training data that are enriched for positive examples of highly active sequences, whereas genomic sequences and random sequences contain a larger fraction of negative examples of inactive sequences (Figure 5A). We found that uncertainty sampling identifies unlabeled candidate sequences which, upon measurement, are more likely to fall into the strongest activity classes (strong enhancers and silencers). This suggests that the candidate sequences that are predicted by the model with the least confidence contain functionally relevant patterns of sequence elements, and as a result these sequences are more likely to be active when measured. Such sequences may contrast with low-information, inactive sequences that are easily learned by the model, and thus they are not prioritized by uncertainty sampling. If this is true, then the complex, higher-order interactions between TF binding sites that define CREs will not be learned from training data with a large fraction of inactive genomic or random sequences. In fact, the problem of limited genomic training examples becomes

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

even more acute if training data with a high fraction of active sequences are necessary to fully model CREs, because only a minority of candidate CREs are active when tested by functional assays.[13,17,77,79] Complementary training approaches currently rely on hundreds of millions of random sequences[35,38] or hundreds of thousands of genomic sequences[34,89,90] measured in a single, large-scale screen. Our work suggests that many of the sequences in these datasets may be low-information training examples and that iteratively training models on smaller but more informative training data will be more effective. This conclusion is supported by a recent study showing that a deep learning model trained on a small but highly active MPRA dataset performed nearly as well as a model trained on a 10× larger MPRA dataset composed of less active genomic sequences.[91]

The most dramatic examples of TF binding sites with opposite effects are sites in enhancers and silencers bound by the same TF.[16,17,92–95] Loss of the CRX binding motif in enhancers causes a decrease in CRE activity, while loss of the same motif in silencers causes an increase in activity. Models that correctly predict the direction of effect of motif mutations have learned the local contextual patterns responsible for the differences in motif effects. However, current deep learning models of gene expression often fail to predict the direction of effect of non-coding variants and thus have not captured the effects of local sequence context.[96–98] This may in part be due to a lack of silencers in most MPRA training datasets, because MPRAs are often not designed to capture silencer effects. An advantage of our MPRA in mouse retinal explants is that we measure both enhancer and silencer activity, thereby generating training data that allow the model to learn the contextual features distinguishing binding sites in enhancers and silencers. Non-coding genetic variants often change activity in a direction that is not expected,[99] highlighting the importance of creating training data that can disambiguate between positive and negative effects of the same motif.

Uncertainty sampling is the crucial step of active learning. In this work, we used discrete classifiers to facilitate uncertainty sampling by using classifier-generated probabilities to calculate uncertainty. However, models that make quantitative predictions of *cis*-regulatory activity are often more useful, and an improvement on this work would be to implement active learning in a regression setting by taking advantage of sampling strategies that rely on Gaussian and neural processes.[49,100–102] Our work also highlights the risks of relying too heavily on one sampling technique. In the initial rounds of active learning, we used entropy sampling, which generated enhancer-biased datasets (Figure 5A) that led to improved predictions of enhancers. In round 4, both entropy sampling and margin sampling produced training datasets biased toward silencers, but entropy sampling resulted in catastrophic forgetting of strong enhancers (Figure S1D),[81] while margin sampling resulted in continued improvements to the model. Future work can protect against catastrophic forgetting by employing diversity criteria[103,104] and using multiple sampling techniques in each round, including ensemble-based sampling.[105,106] Because active learning can generate imbalanced training data, training strategies to mitigate the effects of imbalanced data, such as oversampling the minority class,[107] may improve performance.

When generating candidate sequences *in silico*, we found that accounting for prior knowledge of motifs produced more informative perturbations than random mutagenesis. However, such an approach is biased against discovering new sequence features. Recent advances in generative modeling,[29,91] gradient-based design,[108,109] transfer learning,[110] and evolutionary-inspired data augmentation[111] could provide a complementary approach to generating training data. Active learning is a broadly applicable and effective strategy for learning the impact of sequence context on TF binding sites, one that can take advantage of these ongoing developments in deep learning.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Michael White (mawhite@wustl.edu).

Materials availability

The MPRA vector used in this study has been deposited to AddGene, plasmid #173489. All other unique/stable reagents are available from the lead contact with a completed materials transfer agreement.

Data and code availability

- All sequencing data have been deposited with the NCBI Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) with the primary accession codes GEO: GSE165812 (round 1 library), GEO: GSE230090 (test set library), and GEO: GSE241353 (this study) and are publicly available. Processed data files are also available in Data S2.
- All original code is available a public GitHub repository at https://github.com/barakcohenlab/CRX-Active-Learning, with an archived release deposited at Zenodo: https://doi.org/10.5281/zenodo.4263463. Our implementation of the Generic String Kernel is publicly available at Zenodo: https://doi.org/10.5281/zenodo.13948443. Our fork of the Selene package with support for custom early stopping and learning rate decay is publicly available at Zenodo: https://doi.org/10.5281/zenodo.13948445.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

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., S.L., Y.W., L.T., and D.L. performed computational analyses. R.Z.F. and D.M.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.

DECLARATION OF INTERESTS

B.A.C. is on the scientific advisory board of Patch Biosciences.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL DETAILS
  - Mouse retina explants
- METHOD DETAILS
  - MPRA Library design
  - Plasmid library cloning
  - Electroporation of mouse retinal explants
  - RNA extraction
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - MPRA data processing
  - SVM classifiers of MPRA activity
  - CNN classifiers of MPRA activity
  - Evaluating model performance on test datasets
  - In silico candidate sequence generation
  - Filtering candidate perturbations
  - Active learning
  - Selection of sequences predicted with high confidence
  - Additional perturbation datasets
  - Regression model
  - Motif analysis
  - In silico global importance analysis
  - Nucleotide contribution scores
  - Analysis of MPRA data from K562 cells
  - Statistics and data visualization

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2024.12.004.

Received: March 26, 2024
Revised: October 17, 2024
Accepted: December 6, 2024
Published: January 7, 2025

REFERENCES

1. Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. *94*, 890–898. https://doi.org/10.1002/jcb.20352.
2. Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626. https://doi.org/10.1038/nrg3207.
3. Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-changing landscapes: Transcriptional enhancers in development and evolution. Cell *167*, 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018.
4. Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81. https://doi.org/10.1016/j.gde.2016.12.007.
5. Jindal, G.A., and Farley, E.K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev. Cell *56*, 575–587. https://doi.org/10.1016/j.devcel.2021.02.016.
6. Kim, S., and Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. Mol. Cell *83*, 373–392. https://doi.org/10.1016/j.molcel.2022.12.032.
7. Barolo, S., and Posakony, J.W. (2002). Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. Genes Dev. *16*, 1167–1181. https://doi.org/10.1101/gad.976502.
8. Alexandre, C., and Vincent, J.-P. (2003). Requirements for transcriptional repression and activation by Engrailed in Drosophila embryos. Development *130*, 729–739. https://doi.org/10.1242/dev.00286.
9. Iype, T., Taylor, D.G., Ziesmann, S.M., Garmey, J.C., Watada, H., and Mirmira, R.G. (2004). 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. https://doi.org/10.1210/me.2004-0006.
10. Peng, G.H., Ahmad, O., Ahmad, F., Liu, J., and Chen, S. (2005). 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. https://doi.org/10.1093/hmg/ddi070.
11. Martínez-Montañés, F., Rienzo, A., Poveda-Huertes, D., Pascual-Ahuir, A., and Proft, M. (2013). Activator and repressor functions of the Mot3 transcription factor in the osmostress response of Saccharomyces cerevisiae. Eukaryot. Cell *12*, 636–647. https://doi.org/10.1128/EC.00037-13.
12. Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028. https://doi.org/10.1038/ng.2713.
13. White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. USA *110*, 11952–11957. https://doi.org/10.1073/pnas.1307449110.
14. Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147–151. https://doi.org/10.1038/nature15545.
15. Rister, J., Razzaq, A., Boodram, P., Desai, N., Tsanis, C., Chen, H., Jukam, D., and Desplan, C. (2015). Single-base pair differences in a shared motif determine differential Rhodopsin expression. Science *350*, 1258–1261. https://doi.org/10.1126/science.aab3417.
16. White, M.A., Kwasnieski, J.C., Myers, C.A., Shen, S.Q., Corbo, J.C., and Cohen, B.A. (2016). A simple grammar defines activating and repressing cis-regulatory elements in photoreceptors. Cell Rep. *17*, 1247–1254. https://doi.org/10.1016/j.celrep.2016.09.066.
17. Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. USA *114*, E1291–E1300. https://doi.org/10.1073/pnas.1621150114.
18. Carleton, J.B., Berrett, K.C., and Gertz, J. (2017). Multiplex enhancer interference reveals collaborative control of gene regulation by estrogen receptor α-bound enhancers. Cell Syst. *5*, 333–344.e5. https://doi.org/10.1016/j.cels.2017.08.011.
19. King, D.M., Hong, C.K.Y., Shepherdson, J.L., Granas, D.M., Maricque, B.B., and Cohen, B.A. (2020). Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. eLife *9*, e41279. https://doi.org/10.7554/eLife.41279.
20. Friedman, R.Z., Granas, D.M., Myers, C.A., Corbo, J.C., Cohen, B.A., and White, M.A. (2021). Information content differentiates enhancers from silencers in mouse photoreceptors. eLife *10*, e67403. https://doi.org/10.7554/eLife.67403.
21. Tokuhiro, S., and Satou, Y. (2021). Cis-regulatory code for determining the action of Foxd as both an activator and a repressor in ascidian embryos. Dev. Biol. *476*, 11–17. https://doi.org/10.1016/j.ydbio.2021.03.010.
22. Pang, B., and Snyder, M.P. (2020). Systematic identification of silencers in human cells. Nat. Genet. *52*, 254–263. https://doi.org/10.1038/s41588-020-0578-5.
23. Gisselbrecht, S.S., Palagi, A., Kurland, J.V., Rogers, J.M., Ozadam, H., Zhan, Y., Dekker, J., and Bulyk, M.L. (2020). Transcriptional silencers in

Drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. Mol. Cell *77*, 324–337.e8. https://doi.org/10.1016/j.molcel.2019.10.004.

24. Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., and Furlong, E.E.M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell *148*, 473–486. https://doi.org/10.1016/j.cell.2012.01.030.

25. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366. https://doi.org/10.1038/s41588-021-00782-6.

26. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

27. Atak, Z.K., Taskiran, I.I., Demeulemeester, J., Flerin, C., Mauduit, D., Minnoye, L., Hulselmans, G., Christiaens, V., Ghanem, G.-E., Wouters, J., et al. (2021). Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. Genome Res. *31*, 1082–1096. https://doi.org/10.1101/gr.260851.120.

28. Chen, K.M., Wong, A.K., Troyanskaya, O.G., and Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. Nat. Genet. *54*, 940–949. https://doi.org/10.1038/s41588-022-01102-2.

29. Taskiran, I.I., Spanier, K.I., Dickmänken, H., Kempynck, N., Pančíková, A., Ekşi, E.C., Hulselmans, G., Ismail, J.N., Theunis, K., Vandepoel, R., et al. (2024). Cell-type-directed design of synthetic enhancers. Nature *626*, 212–220. https://doi.org/10.1038/s41586-023-06936-2.

30. Cofer, E.M., Raimundo, J., Tadych, A., Yamazaki, Y., Wong, A.K., Theesfeld, C.L., Levine, M.S., and Troyanskaya, O.G. (2021). Modeling transcriptional regulation of model species with deep learning. Genome Res. *31*, 1097–1105. https://doi.org/10.1101/gr.266171.120.

31. VandenBosch, L.S., Luu, K., Timms, A.E., Challam, S., Wu, Y., Lee, A.Y., and Cherry, T.J. (2022). Machine learning prediction of non-coding variant impact in human retinal cis-regulatory elements. Transl. Vis. Sci. Technol. *11*, 16. https://doi.org/10.1167/tvst.11.4.16.

32. Bravo González-Blas, C., Matetovici, I., Hillen, H., Taskiran, I.I., Vandepoel, R., Christiaens, V., Sansores-García, L., Verboven, E., Hulselmans, G., Poovathingal, S., et al. (2024). Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation. Nat. Cell Biol. *26*, 153–167. https://doi.org/10.1038/s41556-023-01316-4.

33. Movva, R., Greenside, P., Marinov, G.K., Nair, S., Shrikumar, A., and Kundaje, A. (2019). Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. PLoS One *14*, e0218073. https://doi.org/10.1371/journal.pone.0218073.

34. de Almeida, B.P., Reiter, F., Pagani, M., and Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat. Genet. *54*, 613–624. https://doi.org/10.1038/s41588-022-01048-5.

35. Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., Kaasinen, E., Lidschreiber, K., Lidschreiber, M., Daub, C.O., et al. (2022). Sequence determinants of human gene regulatory elements. Nat. Genet. *54*, 283–294. https://doi.org/10.1038/s41588-021-01009-4.

36. Penzar, D., Nogina, D., Noskova, E., Zinkevich, A., Meshcheryakov, G., Lando, A., Rafi, A.M., de Boer, C., and Kulakovskiy, I.V. (2023). LegNet: a best-in-class deep learning model for short DNA regulatory regions. Bioinformatics *39*, btad457. https://doi.org/10.1093/bioinformatics/btad457.

37. Linder, J., Koplik, S.E., Kundaje, A., and Seelig, G. (2022). Deciphering the impact of genetic variation on human polyadenylation using APARENT2. Genome Biol. *23*, 232. https://doi.org/10.1186/s13059-022-02799-4.

38. de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., and Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. *38*, 56–65. https://doi.org/10.1038/s41587-019-0315-8.

39. Vaishnav, E.D., de Boer, C.G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D.A., Levin, J.Z., Cubillos, F.A., and Regev, A. (2022). The evolution, evolvability and engineering of gene regulatory DNA. Nature *603*, 455–463. https://doi.org/10.1038/s41586-022-04506-6.

40. Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S., and Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. Genome Res. *27*, 2015–2024. https://doi.org/10.1101/gr.224964.117.

41. Kim, Y.J., Rhee, K., Liu, J., Jeammet, S., Turner, M.A., Small, S.J., and Garcia, H.G. (2022). Predictive modeling reveals that higher-order cooperativity drives transcriptional repression in a synthetic developmental enhancer. eLife *11*, e73395. https://doi.org/10.7554/eLife.73395.

42. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18. https://doi.org/10.1038/s41588-018-0295-5.

43. de Boer, C.G., and Taipale, J. (2024). Hold out the genome: a roadmap to solving the cis-regulatory code. Nature *625*, 41–50. https://doi.org/10.1038/s41586-023-06661-w.

44. Monarch, R.M. (2021). Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI (Simon and Schuster).

45. Settles, B. (2012). Active Learning (Springer International Publishing).

46. Lewis, D.D., and Gale, W.A. (1994). A sequential algorithm for training text classifiers. In SIGIR '94, B.W. Croft and C.J. van Rijsbergen, eds. (Springer), pp. 3–12. https://doi.org/10.1007/978-1-4471-2099-5_1.

47. King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature *427*, 247–252. https://doi.org/10.1038/nature02236.

48. Kanda, G.N., Tsuzuki, T., Terada, M., Sakai, N., Motozawa, N., Masuda, T., Nishida, M., Watanabe, C.T., Higashi, T., Horiguchi, S.A., et al. (2022). Robotic search for optimal cell culture in regenerative medicine. eLife *11*, e77007. https://doi.org/10.7554/eLife.77007.

49. Hie, B., Bryson, B.D., and Berger, B. (2020). Leveraging uncertainty in machine learning accelerates biological discovery and design. Cell Syst. *11*, 461–477.e9. https://doi.org/10.1016/j.cels.2020.09.007.

50. Garnett, R., Gärtner, T., Vogt, M., and Bajorath, J. (2015). Introducing the "active search" method for iterative virtual screening. J. Comput. Aided Mol. Des. *29*, 305–314. https://doi.org/10.1007/s10822-015-9832-9.

51. Oglic, D., Oatley, S.A., Macdonald, S.J.F., Mcinally, T., Garnett, R., Hirst, J.D., and Gärtner, T. (2018). Active search for computer-aided drug design. Mol. Inform. *37*, 1700130. https://doi.org/10.1002/minf.201700130.

52. Warmuth, M.K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., and Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. J. Chem. Inf. Comput. Sci. *43*, 667–673. https://doi.org/10.1021/ci025620t.

53. Singh, R., Li, J.S.S., Tattikota, S.G., Liu, Y., Xu, J., Hu, Y., Perrimon, N., and Berger, B. (2023). Prioritizing transcription factor perturbations from single-cell transcriptomics. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.27.497786.

54. Guan, X., Li, Z., Zhou, Y., Shao, W., and Zhang, D. (2023). Active learning for efficient analysis of high-throughput nanopore data. Bioinformatics *39*, btac764. https://doi.org/10.1093/bioinformatics/btac764.

55. Huang, K., Lopez, R., Hütter, J.-C., Kudo, T., Rios, A., and Regev, A. (2023). Sequential optimal experimental design of perturbation screens guided by multi-modal priors. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.12.571389.

56. Furukawa, T., Morrow, E.M., and Cepko, C.L. (1997). Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and

regulates photoreceptor differentiation. Cell *91*, 531–541. https://doi.org/10.1016/S0092-8674(00)80439-0.

57. Chen, S., Wang, Q.L., Nie, Z., Sun, H., Lennon, G., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., and Zack, D.J. (1997). Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. Neuron *19*, 1017–1030. https://doi.org/10.1016/S0896-6273(00)80394-3.
58. Freund, C.L., Gregory-Evans, C.Y., Furukawa, T., Papaioannou, M., Looser, J., Ploder, L., Bellingham, J., Ng, D., Herbrick, J.A.S., Duncan, A., et al. (1997). Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. Cell *91*, 543–553. https://doi.org/10.1016/S0092-8674(00)80440-7.
59. Hennig, A.K., Peng, G.-H., and Chen, S. (2008). Regulation of photoreceptor gene expression by Crx-associated transcription factor network. Brain Res. *1192*, 114–133. https://doi.org/10.1016/J.BRAINRES.2007.06.036.
60. Hughes, A.E.O., Enright, J.M., Myers, C.A., Shen, S.Q., and Corbo, J.C. (2017). Cell type-specific epigenomic analysis reveals a uniquely closed chromatin architecture in mouse rod photoreceptors. Sci. Rep. *7*, 43184. https://doi.org/10.1038/srep43184.
61. Murphy, D.P., Hughes, A.E., Lawrence, K.A., Myers, C.A., and Corbo, J.C. (2019). Cis-regulatory basis of sister cell type divergence in the vertebrate retina. eLife *8*, e48216. https://doi.org/10.7554/eLife.48216.
62. Swain, P.K., Chen, S., Wang, Q.L., Affatigato, L.M., Coats, C.L., Brady, K.D., Fishman, G.A., Jacobson, S.G., Swaroop, A., Stone, E., et al. (1997). Mutations in the cone-rod homeobox gene are associated with the cone-rod dystrophy photoreceptor degeneration. Neuron *19*, 1329–1336. https://doi.org/10.1016/s0896-6273(00)80423-7.
63. Corbo, J.C., Lawrence, K.A., Karlstetter, M., Myers, C.A., Abdelaziz, M., Dirkes, W., Weigelt, K., Seifert, M., Benes, V., Fritsche, L.G., et al. (2010). CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. Genome Res. *20*, 1512–1525. https://doi.org/10.1101/gr.109405.110.
64. Hsiau, T.H.-C., Diaconu, C., Myers, C.A., Lee, J., Cepko, C.L., and Corbo, J.C. (2007). The Cis-regulatory logic of the mammalian photoreceptor transcriptional network. PLoS One *2*, e643. https://doi.org/10.1371/journal.pone.0000643.
65. Campla, C.K., Mast, H., Dong, L., Lei, J., Halford, S., Sekaran, S., and Swaroop, A. (2019). 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. https://doi.org/10.1093/hmg/ddy388.
66. Oh, E.C.T., Khan, N., Novelli, E., Khanna, H., Strettoi, E., and Swaroop, A. (2007). Transformation of cone precursors to functional rod photoreceptors by bZIP transcription factor NRL. Proc. Natl. Acad. Sci. USA *104*, 1679–1684. https://doi.org/10.1073/pnas.0605934104.
67. Ruzycki, P.A., Tran, N.M., Kefalov, V.J., Kolesnikov, A.V., and Chen, S. (2015). Graded gene expression changes determine phenotype severity in mouse models of CRX-associated retinopathies. Genome Biol. *16*, 171. https://doi.org/10.1186/s13059-015-0732-z.
68. Wang, S., Sengel, C., Emerson, M.M., and Cepko, C.L. (2014). A gene regulatory network controls the binary fate decision of rod and bipolar cells in the vertebrate retina. Dev. Cell *30*, 513–527. https://doi.org/10.1016/j.devcel.2014.07.018.
69. Montana, C.L., Lawrence, K.A., Williams, N.L., Tran, N.M., Peng, G.-H., Chen, S., and Corbo, J.C. (2011). Transcriptional regulation of neural retina leucine zipper (Nrl), a photoreceptor cell fate determinant. J. Biol. Chem. *286*, 36921–36931. https://doi.org/10.1074/jbc.M111.279026.
70. Swaroop, A., Wang, Q.L., Wu, W., Cook, J., Coats, C., Xu, S., Chen, S., Zack, D.J., and Sieving, P.A. (1999). Leber congenital amaurosis caused by a homozygous mutation (R90W) in the homeodomain of the retinal transcription factor CRX: direct evidence for the involvement of CRX in the development of photoreceptor function. Hum. Mol. Genet. *8*, 299–305. https://doi.org/10.1093/hmg/8.2.299.
71. Swaroop, A., Kim, D., and Forrest, D. (2010). Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. Nat. Rev. Neurosci. *11*, 563–576. https://doi.org/10.1038/nrn2880.
72. Nishida, A., Furukawa, A., Koike, C., Tano, Y., Aizawa, S., Matsuo, I., and Furukawa, T. (2003). Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. Nat. Neurosci. *6*, 1255–1263. https://doi.org/10.1038/nn1155.
73. Koike, C., Nishida, A., Ueno, S., Saito, H., Sanuki, R., Sato, S., Furukawa, A., Aizawa, S., Matsuo, I., Suzuki, N., et al. (2007). Functional roles of Otx2 transcription factor in postnatal mouse retinal development. Mol. Cell. Biol. *27*, 8318–8329. https://doi.org/10.1128/MCB.01209-07.
74. Mitton, K.P., Swain, P.K., Chen, S., Xu, S., Zack, D.J., and Swaroop, A. (2000). The leucine zipper of NRL interacts with the CRX homeodomain. A possible mechanism of transcriptional synergy in rhodopsin regulation. J. Biol. Chem. *275*, 29794–29799. https://doi.org/10.1074/jbc.M003658200.
75. Hughes, A.E.O., Myers, C.A., and Corbo, J.C. (2018). A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. Genome Res. *28*, 1520–1531. https://doi.org/10.1101/gr.231886.117.
76. Shepherdson, J.L., Friedman, R.Z., Zheng, Y., Sun, C., Oh, I.Y., Granas, D.M., Cohen, B.A., Chen, S., and White, M.A. (2024). Pathogenic variants in CRX have distinct *cis*-regulatory effects on enhancers and silencers in photoreceptors. Genome Res. *34*, 243–255. https://doi.org/10.1101/gr.278133.123.
77. Kwasnieski, J.C., Fiore, C., Chaudhari, H.G., and Cohen, B.A. (2014). High-throughput functional testing of ENCODE segmentation predictions. Genome Res. *24*, 1595–1602. https://doi.org/10.1101/gr.173518.114.
78. Chaudhari, H.G., and Cohen, B.A. (2018). Local sequence features that influence AP-1 cis-regulatory activity. Genome Res. *28*, 171–181. https://doi.org/10.1101/gr.226530.117.
79. Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al.; ENCODE Project Consortium (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.
80. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811. https://doi.org/10.1101/gr.144899.112.
81. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. USA *114*, 3521–3526. https://doi.org/10.1073/pnas.1611835114.
82. Lee, J., Myers, C.A., Williams, N., Abdelaziz, M., and Corbo, J.C. (2010). Quantitative fine-tuning of photoreceptor cis-regulatory elements through affinity modulation of transcription factor binding sites. Gene Ther. *17*, 1390–1399. https://doi.org/10.1038/gt.2010.77.
83. Loell, K.J., Friedman, R.Z., Myers, C.A., Corbo, J.C., Cohen, B.A., and White, M.A. (2024). Transcription factor interactions explain the context-dependent activity of CRX binding sites. PLoS Comput. Biol. *20*, e1011802. https://doi.org/10.1371/journal.pcbi.1011802.
84. Koo, P.K., Majdandzic, A., Ploenzke, M., Anand, P., and Paul, S.B. (2021). Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. PLoS Comput. Biol. *17*, e1008925. https://doi.org/10.1371/journal.pcbi.1008925.
85. Sayal, R., Dresch, J.M., Pushel, I., Taylor, B.R., and Arnosti, D.N. (2016). Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early Drosophila embryo. eLife *5*, e08445. https://doi.org/10.7554/eLife.08445.
86. Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive

regions in the Drosophila blastoderm. PLoS Biol. *6*, e27. https://doi.org/10.1371/journal.pbio.0060027.

87. Kok, K., Ay, A., Li, L.M., and Arnosti, D.N. (2015). Genome-wide errant targeting by Hairy. eLife *4*, e06394. https://doi.org/10.7554/eLife.06394.

88. Cheng, H., Khanna, H., Oh, E.C.T., Hicks, D., Mitton, K.P., and Swaroop, A. (2004). Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. Hum. Mol. Genet. *13*, 1563–1575. https://doi.org/10.1093/hmg/ddh173.

89. Agarwal, V., Inoue, F., Schubach, M., Martin, B.K., Dash, P.M., Zhang, Z., Sohota, A., Noble, W.S., Yardimci, G.G., Kircher, M., et al. (2023). Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.05.531189.

90. Gosai, S.J., Castro, R.I., Fuentes, N., Butts, J.C., Kales, S., Noche, R.R., Mouri, K., Sabeti, P.C., Reilly, S.K., and Tewhey, R. (2023). Machine-guided design of synthetic cell type-specific cis-regulatory elements. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.08.552077.

91. Yin, C., Hair, S.C., Byeon, G.W., Bromley, P., Meuleman, W., and Seelig, G. (2024). Iterative deep learning-design of human enhancers exploits condensed sequence grammar to achieve cell type-specificity. Preprint at bioRxiv. https://doi.org/10.1101/2024.06.14.599076.

92. Grass, J.A., Boyer, M.E., Pal, S., Wu, J., Weiss, M.J., and Bresnick, E.H. (2003). GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. Proc. Natl. Acad. Sci. USA *100*, 8811–8816. https://doi.org/10.1073/pnas.1432147100.

93. Majello, B., De Luca, P., and Lania, L. (1997). Sp3 is a bifunctional transcription regulator with modular independent activation and repression domains. J. Biol. Chem. *272*, 4021–4026. https://doi.org/10.1074/jbc.272.7.4021.

94. Sloan, J., Hakenjos, J.P., Gebert, M., Ermakova, O., Gumiero, A., Stier, G., Wild, K., Sinning, I., and Lohmann, J.U. (2020). Structural basis for the complex DNA binding behavior of the plant stem cell regulator WUSCHEL. Nat. Commun. *11*, 2223. https://doi.org/10.1038/s41467-020-16024-y.

95. Robbe, Z.L., Shi, W., Wasson, L.K., Scialdone, A.P., Wilczewski, C.M., Sheng, X., Hepperla, A.J., Akerberg, B.N., Pu, W.T., Cristea, I.M., et al. (2022). CHD4 is recruited by GATA4 and NKX2-5 to repress noncardiac gene programs in the developing heart. Genes Dev. *36*, 468–482. https://doi.org/10.1101/gad.349154.121.

96. Tang, Z., Toneyan, S., and Koo, P.K. (2023). Current approaches to genomic deep learning struggle to fully capture human genetic variation. Nat. Genet. *55*, 2021–2022. https://doi.org/10.1038/s41588-023-01517-5.

97. Huang, C., Shuai, R.W., Baokar, P., Chung, R., Rastogi, R., Kathail, P., and Ioannidis, N.M. (2023). Personal transcriptome variation is poorly explained by current genomic deep learning models. Nat. Genet. *55*, 2056–2059. https://doi.org/10.1038/s41588-023-01574-w.

98. Sasse, A., Ng, B., Spiro, A.E., Tasaki, S., Bennett, D.A., Gaiteri, C., De Jager, P.L., Chikina, M., and Mostafavi, S. (2023). Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. Nat. Genet. *55*, 2060–2064. https://doi.org/10.1038/s41588-023-01524-6.

99. Yanchus, C., Drucker, K.L., Kollmeyer, T.M., Tsai, R., Winick-Ng, W., Liang, M., Malik, A., Pawling, J., De Lorenzo, S.B., Ali, A., et al. (2022). A noncoding single-nucleotide polymorphism at 8q24 drives *IDH1*-mutant glioma formation. Science *378*, 68–78. https://doi.org/10.1126/science.abj2890.

100. Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D.J., Ali Eslami, S.M., and Teh, Y.W. (2018). Neural processes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1807.01622.

101. Rasmussen, C.E., and Williams, C.K.I. (2005). Gaussian Processes for Machine Learning (MIT Press).

102. Sluijterman, L., Cator, E., and Heskes, T. (2023). Optimal training of mean variance estimation neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2302.08875.

103. Nguyen, Q., and Garnett, R. (2022). Nonmyopic multiclass active search with diminishing returns for diverse discovery. Preprint at arXiv. https://doi.org/10.48550/arXiv.2202.03593.

104. Nguyen, H.T., and Smeulders, A. (2004). Active learning using pre-clustering. In Proceedings of the Twenty-First International Conference on Machine Learning ICML '04 (ACM Press), p. 79. https://doi.org/10.1145/1015330.1015349.

105. Dagan, I., and Engelson, S.P. (1995). Committee-based sampling for training probabilistic classifiers. In Proceedings of the Twelfth International Conference on Machine Learning (Morgan Kaufmann), pp. 150–157. https://doi.org/10.1016/B978-1-55860-377-6.50027-X.

106. Siddhant, A., and Lipton, Z.C. (2018). 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 (Association for Computational Linguistics), pp. 2904–2909. https://doi.org/10.18653/v1/D18-1318.

107. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. *16*, 321–357. https://doi.org/10.1613/jair.953.

108. Linder, J., Bogard, N., Rosenberg, A.B., and Seelig, G. (2020). A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. Cell Syst. *11*, 49–62.e16. https://doi.org/10.1016/j.cels.2020.05.007.

109. Linder, J., and Seelig, G. (2021). Fast activation maximization for molecular sequence design. BMC Bioinformatics *22*, 510. https://doi.org/10.1186/s12859-021-04437-5.

110. de Almeida, B.P., Schaub, C., Pagani, M., Secchia, S., Furlong, E.E.M., and Stark, A. (2024). Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. Nature *626*, 207–211. https://doi.org/10.1038/s41586-023-06905-9.

111. Lee, N.K., Tang, Z., Toneyan, S., and Koo, P.K. (2023). EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. Genome Biol. *24*, 105. https://doi.org/10.1186/s13059-023-02941-w.

112. Tareen, A., and Kinney, J.B. (2020). Logomaker: beautiful sequence logos in Python. Bioinformatics *36*, 2272–2274. https://doi.org/10.1093/bioinformatics/btz921.

113. Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc. Natl. Acad. Sci. USA *109*, 19498–19503. https://doi.org/10.1073/pnas.1210678109.

114. Montana, C.L., Myers, C.A., and Corbo, J.C. (2011). Quantifying the activity of cis-regulatory elements in the mouse retina by explant electroporation. J. Vis. Exp. *52*, e2821. https://doi.org/10.3791/2821.

115. Giguère, S., Marchand, M., Laviolette, F., Drouin, A., and Corbeil, J. (2013). Learning a peptide-protein binding affinity predictor with kernel ridge regression. BMC Bioinformatics *14*, 82. https://doi.org/10.1186/1471-2105-14-82.

116. Giguère, S., Rolland, A., Laviolette, F., and Marchand, M. (2015). Algorithms for the hard pre-image problem of string kernels and the general problem of string prediction. In Proceedings of the 32nd International Conference on Machine Learning Proceedings of Machine Learning Research (PMLR), *37*, pp. 2021–2029.

117. Leslie, C., Eskin, E., and Noble, W.S. (2002). The spectrum kernel: A string kernel for SVM protein classification. Pac. Symp. Biocomput. 564–575. https://doi.org/10.1142/9789812799623_0053.

118. Lee, D., Karchin, R., and Beer, M.A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. *21*, 2167–2180. https://doi.org/10.1101/gr.121905.111.

119. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).

Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490.

120. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.01703.
121. Chen, K.M., Cofer, E.M., Zhou, J., and Troyanskaya, O.G. (2019). Selene: a PyTorch-based deep learning library for sequence-level data. Nat. Methods *16*, 315–318. https://doi.org/10.1101/438291.
122. Lee, D. (2016). LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics *32*, 2196–2198. https://doi.org/10.1093/bioinformatics/btw142.
123. Koo, P.K., and Ploenzke, M. (2021). Improving representations of genomic sequence motifs in convolutional networks with exponential activations. Nat. Mach. Intell. *3*, 258–266. https://doi.org/10.1038/s42256-020-00291-x.
124. Zhao, Y., Granas, D., and Stormo, G.D. (2009). Inferring binding energies from selected binding sites. PLoS Comput. Biol. *5*, e1000590. https://doi.org/10.1371/journal.pcbi.1000590.
125. Majdandzic, A., Rajesh, C., and Koo, P.K. (2023). Correcting gradient-based interpretations of deep neural networks for genomics. Genome Biol. *24*, 109. https://doi.org/10.1186/s13059-023-02956-3.
126. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
127. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
128. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.
129. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
E. coli 10-beta electrocompetent cells	NEB	Cat#C3020K
Chemicals, peptides, and recombinant proteins		
Q5 High-Fidelity 2X Master Mix	NEB	Cat#M0492S
EcoRI-HF restriction enzyme	NEB	Cat#R3101S
NotI-HF restriction enzyme	NEB	Cat#R3189S
SphI-HF restriction enzyme	NEB	Cat#R3182S
SpeI-HF restriction enzyme	NEB	Cat#R3133S
NheI-HF restriction enzyme	NEB	Cat#R3131S
TRIzol	ThermoFisher	Cat#15596026
Critical commercial assays		
Monarch PCR Cleanup Kit	NEB	Cat# 1030S
Monarch DNA Gel Extration Kit	NEB	Cat#T1020L
TURBO DNA-free kit	Invitrogen	Cat#AM1907
Superscript IV first strand synthesis kit	Invitrogen	Cat#18091050
Zymo Pure II Plasmid Maxiprep Kit	Zymo	Cat#D4203
Deposited data		
MPRA assay of CRX-bound genomic sequences (Round 1 training data)	Friedman et al.[20]	GSE165812
MPRA assay of active learning datasets	This study	GSE241353
MPRA assay of CRX-bound sequences (test dataset)	Shepherdson et al.[76]	GSE230090
MPRA assay in K562 cells	Agarwal et al.[89]	https://doi.org/10.1101/2023.03.05.531189
Experimental models: Organisms/strains		
M. musculus: strain background CD-1	Charles River	Strain code 022
Oligonucleotides		
CRE sequences for MPRA libraries 2-4	Agilent	Listed in Data S1
Primers	IDT	Listed in Table S2
Recombinant DNA		
pJK01_Rhominprox_DsRed	AddGene	Plasmid #137489
pJK03_Rho_basal_DsRed	AddGene	Plasmid # 173490
Software and algorithms		
Numpy	https://numpy.org/	https://doi.org/10.1038/s41586-020-2649-2
Scipy	https://scipy.org/	https://doi.org/10.1038/s41592-019-0686-2
Pandas	https://pandas.pydata.org/	https://doi.org/10.5281/zenodo.3509134
Matplotlib	https://matplotlib.org/	https://doi.org/10.1109/MCSE.2007.55
Logomaker	Tareen and Kinney[112]	https://doi.org/10.1093/bioinformatics/btz921
Data processing and model code	This study	https://doi.org/10.5281/zenodo.4263463

EXPERIMENTAL MODEL DETAILS

Mouse retina explants

CD-1 mice were obtained from Charles River Laboratory. Retinas from newborn (P0) mice were dissected and electroporated.[64] The sex of the mice could not be determined at the P0 stage and mice were thus not selected by sex. 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)

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

from surrounding sclera and soft tissue leaving the lens in place. 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.

METHOD DETAILS

MPRA Library design

All MPRA libraries were obtained using custom oligonucleotide (oligo) synthesis from Agilent Technologies™.[20,113] 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. 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 Table S1 and Data S1.

Plasmid library cloning

We created MPRA libraries from oligo pools using two-step cloning as described[20,113] with the following modifications. Oligos were amplified through multiple PCR reactions (New England Biolabs [NEB] Q5 High-Fidelity 2X Master Mix, cat. #M0515, see Table S2 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, Table S2), 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 (Table S1) for parallel analyses. Barcode complexity was always verified by sequencing the final library on the Illumina MiniSeq™ platform.

Electroporation of mouse retinal explants

Dissected retinas were transferred to an electroporation chamber (model BTX453 Microslide chamber, BTX Harvard Apparatus[114]) 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 (Table S1). 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 for eight days.

RNA extraction

Retinas were harvested in TRIzol, homogenized with a sterile needle, and RNA was extracted 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

MPRA 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 (Table S1) 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.

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

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. Activity scores were discretized into four classes using the cutoffs we previously reported.[20] All activity scores and statistics are in Data S2.

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,[115,116] the Generic String Kernel for any two strings y, y' is:

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

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:

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

When $\sigma_p = \infty$, the first *exp* becomes constant and we recover the *k*-mer kernel.[117,118] 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[119] 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 Area Under the Receiver Operating Characteristic (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[120] and used Selene[121] 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 did not deploy minority oversampling or any other training-based method for learning on imbalanced datasets. for 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.

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.

Evaluating model performance on test datasets

For each round and for each model, 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. We evaluated model performance on two independent MPRA test datasets that represent different prediction tasks.

The SVM test set is an exhaustive motif perturbation analysis of 29 CRX ChIP-seq peaks that are strong enhancers in our original Round 1 data[20]; 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 711

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

sequences in the same library as Round 3 (Table S1). This test set assesses the ability of the model to predict the effects of perturbations to TF binding sites in highly active enhancers. A limitation of this test set is that the classes are imbalanced (44% strong enhancers vs 3.7% silencers).

The CNN test set includes 1,723 CRX ChIP-seq peaks from a parallel study[76] that were not used elsewhere in the active learning pipeline. This test set reflects the natural ratio of enhancers to silencers among wild-type CRX-bound sequences (8% strong enhancers, 22% silencers). These sequences were not centered on CRX motifs so we could not use it to evaluate the SVM because the kernel function considers absolute *k*-mer positions, rather than relative positions, and thus requires input sequences centered on a CRX motif as a reference point. We normalized the published 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.

We utilized the F1 score as our evaluation metric for both the SVM and the CNN classifier. 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.

In silico candidate sequence generation

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.

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),[20] 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 scrambling 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.

Filtering candidate perturbations

To filter out any candidate sequences that lack the general properties of photoreceptor CREs, we trained a *k*-mer SVM to classify rod photoreceptor ATAC-seq peaks[60] 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[122] with parameters -l 10 -k 6 -d 3 and five-fold cross-validation. We assessed model performance using the AUROC (Figure S5). In each round of *in silico* mutagenesis, we kept 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.

Active learning

Our machine learning models take a 164-bp sequence as input and predicts $p(y_i|x)$, the probability that the sequence belongs in the -th activity bin, $i = 1,\ldots,4$. Our objective is to determine which perturbations are the most uncertain to the model. In Rounds 2, 3, and 4a, we used entropy uncertainty, which is quantified with Shannon entropy, $S(x) = -\sum_i p(y_i|x)\log_2 p(y_i|x)$ and reaches its maximum when a classifier assigns equal probabilities of belonging to each class. In Round 4b, we used margin uncertainty, which is defined as $M(x) = 1 - [p(\hat{y}|x) - p(\widehat{y^*}|x)]$, 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

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

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. 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 in the MPRA experiments.

Selection of sequences predicted with high confidence

During Round 4, we ranked *in silico*-generated sequences by their most probable class as predicted by the CNN classifier. We selected 500 sequences with the highest probability of being a strong enhancer, 500 high-probability weak enhancers, 500 high-probability inactive sequences, and 1500 high-probability silencers.

Additional perturbation datasets

To find inactive sequences whose motif content was similar to the strong enhancers in Figure 4, 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. 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. Last, we moved the non-CRX motif in the strong enhancer to every other position that did not overlap with the CRX motif.

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 (Table S1).

Regression model

To perform model interpretation, we trained a new regression CNN that predicts $\log_2$ MPRA activity directly from sequence. We updated our architecture to include residual skip connections, dilated convolutions, and first-layer exponential activation.[123] 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 surrounded by a residual skip connection, and then max-pooling. We trained the model after Round 4 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[16,124] and μ=9. 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.

In silico global importance analysis

We used Global Importance Analysis[84] 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 $\log_2$ 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.[125] This method has been shown to reduce noise in feature attribution maps. Motif importances scores were obtained by summing across all nucleotides that overlap predicted occupancy hits.

Analysis of MPRA data from K562 cells

We downloaded an existing large-scale K562 dataset.[89] Data were obtained from Supplementary Tables 3 and 4 from that work. We selected sequences that were observed in multiple replicates, had a sample coefficient of variation less than or equal to 0.75, belonged to the "putative enhancer" category, and were in the positive strand orientation. Among these sequences, we defined the top 20% as positives, the bottom 50% as negatives, and removed sequences in the 50-80th percentile.

Please cite this article in press as: Friedman et al., Active learning of enhancers and silencers in the developing neural retina, Cell Systems (2024), https://doi.org/10.1016/j.cels.2024.12.004

We split the data into 10 folds based on chromosomal origin so that each fold contained approximately 10% of the data. The folds are:

Fold	Chromosomes
1	chr1
2	chr2, chr14
3	chr4, chr7
4	chr3, chr15
5	chr5, chr19, chr21
6	chr6, chrX, chrY
7	chr8, chr9
8	chr10, chr11
9	chr12, chr13, chr16
10	chr17, chr20, chr22

We used the same CNN architecture as our regression model, but using 230 bp input and adding a sigmoid activation function to the output to convert it into a binary classifier. We trained all models using the Adam optimizer (learning rate = 0.0001, weight decay = 10^{-6}), binary cross-entropy as a loss function, a batch size of 128, and 100 epochs with early stopping (patience = 15, metric = AUPR) and the default learning rate scheduler in Selene. We kept the model with the lowest loss on the held-out validation set for evaluating on the test set and performing additional rounds of sampling.

Statistics and data visualization

All statistical analyses and data visualization were performed in Python with Numpy,[126] Scipy,[127] Pandas,[128] Matplotlib,[129] and Logomaker.[112] 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.