PIONEER: A VIRTUAL PLATFORM FOR ITERATIVE IMPROVEMENT OF GENOMIC DEEP LEARNING

Alessandro Crnjar, John J. Desmarais, Justin Kinney, & Peter K. Koo
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724, USA
koo@cshl.edu

ABSTRACT

Deep neural networks (DNNs) have improved our ability to predict regulatory activity from DNA sequences, providing valuable insights into gene regulation. However, these models often fail to generalize to sequences underrepresented in their training data, limiting applications like variant effect prediction and de novo sequence design. This limitation reflects a bias toward natural variation across the genome, making DNNs vulnerable to covariate shifts, where test sequences diverge statistically from the training distribution. Here, we introduce PIONEER, a computational platform that simulates functional genomics experiments to systematically benchmark and optimize training data composition through iterative AI-experiment cycles. Using PIONEER, we compare sequence proposal strategies—including active learning and random baselines—evaluating their impact on model generalization across increasing levels of covariate shift. To ensure a fair comparison, we also assess each approach within a fixed experimental budget, accounting for DNA synthesis costs. PIONEER provides a scalable and extensible framework for optimizing training data composition to enhance model generalization, advancing applications in regulatory genomics, synthetic biology, and precision medicine.

1 INTRODUCTION

Understanding the cis-regulatory code—the set of rules by which non-coding DNA sequences regulate gene expression—remains a central challenge in genomics. This code governs the activity of regulatory elements, such as enhancers and promoters, through encoded sequence motifs and their combinatorial interactions. Deciphering these rules is essential for predicting the functional impact of genetic variants, designing synthetic regulatory elements, and exploring the evolution of gene regulation. However, the immense size of sequence space poses a fundamental challenge: for sequences of length L, there are 4^L possible combinations. Even a 200-nucleotide sequence represents an astronomical number of possibilities, and despite advances in high-throughput sequencing technologies, only a tiny fraction of this space can be experimentally probed.

Deep neural networks (DNNs) have emerged as powerful tools for modeling gene regulatory activity (Avsec et al., 2021a; Chen et al., 2022). Trained to take DNA sequences as input and predict the outputs of functional genomics experiments, these models excel at identifying sequence motifs and deciphering their complex syntax (Avsec et al., 2021b; Toneyan et al., 2022; Koo et al., 2020), offering valuable insights into gene regulation. However, because the training data is typically mapped to the reference genome, these models lack exposure to greater genetic diversity. This challenge is particularly evident under covariate shift, where the sequences encountered during model deployment deviate from the distribution of training data (Shimodaira, 2000). Even state-of-the-art models struggle to predict the effects of single nucleotide mutations or population-level variation (Huang et al., 2023; Sasse et al., 2023; Tang et al., 2023) – cases that involve only a modest shift from the training set. The problem becomes even more pronounced in de novo sequence design, which requires generalization to entirely novel regions of sequence space.

Evaluating model generalization under covariate shifts requires datasets that systematically assess performance across varying degrees of distributional shift. However, such datasets remain scarce, limiting rigorous assessment and improvement of genomic models. The lack of standardized benchmarks for optimizing training data further impedes the development of models capable of robust generalization. Iterative exploration of sequence space offers a promising solution, with active learning enabling the targeted enrichment of training datasets to improve predictive accuracy while minimizing experimental costs (Friedman et al., 2025; Jain et al., 2023; Morrow et al., 2024; van Tilborg & Grisoni, 2024; Gorantla et al., 2023; Bailey et al., 2023; Wang et al., 2023). While some studies have explored experiment-in-the-loop strategies (Friedman et al., 2025; Linder & Seelig, 2021), a systematic benchmark comparing these approaches, including advanced deep active learning techniques (Ren et al., 2021), is still lacking. Additionally, the high cost of labor and experimental validation remains a major obstacle to benchmarking and optimizing iterative sequence proposal strategies. Addressing these gaps is essential for developing next-generation predictive models in regulatory genomics, synthetic biology, and precision medicine.

Here, we introduce PIONEER (Platform for Iterative Optimization and Navigation to Enhance Exploration of Regulatory sequences), a virtual platform designed to simulate iterative AI-experiment cycles (Fig. 1). Each cycle in PIONEER consists of three steps: (1) sequence space navigation, (2) *in silico* experimentation, and (3) AI model refinement. In the navigation phase, candidate sequences are sampled from the vast sequence space. The goal is to identify informative sequences that would maximally improve the AI model's performance. In the experimentation phase, these candidate sequences are annotated using an *in silico* oracle (a deep neural network trained on functional genomics data). Finally, the AI refinement phase involves incorporating the newly labeled sequences into the training set and retraining the model to improve its predictive capabilities. By simulating AI-experiment cycles, PIONEER enables systematic exploration of *in silico* sequence-function landscape and facilitates scalable benchmarking of sequence proposal strategies.

2 Active learning for navigating regulatory sequence-function landscapes

A central challenge in training deep neural networks for regulatory genomics is creating training datasets that generalize beyond naturally occurring sequences. Given the astronomical size of sequence space—far exceeding what can be experimentally tested—strategic selection of training examples is essential. Active learning addresses this need by iteratively selecting sequences that are most informative, thereby improving model performance while minimizing the need for extensive labeling. This approach is especially valuable in genomics, where experimental validation is both costly and labor-intensive.

A key component of active learning is the acquisition function, which estimates the potential of a sequence to improve model generalization when added to the training set (Ren et al., 2021). Because directly measuring a sequence's impact is challenging, most methods rely on proxy metrics. For example, uncertainty-based selection uses measures such as entropy (Nguyen et al., 2022; Aggarwal et al., 2014), margin sampling (Scheffer et al., 2001), or least confidence (Aggarwal et al., 2014) to prioritize sequences from regions where the model is least certain. While effective at refining predictions in underexplored areas, this method can result in redundant selections if similar sequences dominate these regions. In contrast, diversity-based selection employs clustering or density estimation to ensure broad coverage of sequence space (Phillips, 2016), though it may miss sequences that would yield the largest improvements in predictive accuracy.

Active learning can be applied to individual data points or in batches (Ren et al., 2021; Settles & Craven, 2008; Du et al., 2017; Zhan et al., 2021). Batch-mode active learning selects multiple samples per iteration, which is particularly well-suited for genomic experiments, which are typically conducted in batches. By evaluating the collective informativeness of a group of sequences, batch selection aims to balance high information content with sufficient diversity. However, designing effective batches poses its own challenges—if batches are poorly constructed, they may over-sample from narrow regions of sequence space, ultimately limiting model generalization.

Traditional active learning methods assume a fixed, finite pool of unlabeled data, allowing models to screen and rank all available samples (Ren et al., 2021). In genomics, however, the enormous sequence space precludes exhaustive screening. Instead, genomic active learning must first explore

sequence space through generation or sampling, and then apply acquisition functions that integrate both exploration and selection. The challenge is not only to identify the best sequences to label but also to design exploration strategies that optimize both search efficiency and model generalization.

PIONEER OVERVIEW

PIONEER is a PyTorch-based computational framework designed to systematically benchmark sequence proposal strategies through iterative AI-experiment cycles. Each iteration consists of generating or selecting candidate sequences, applying an acquisition function to prioritize the most informative ones, and retraining the AI model with newly labeled data. This continuous feedback loop, wherein new sequences are labeled by an *in silico* oracle (typically a deep learning model trained on functional genomics data) and incorporated into the training set, progressively refines model predictions and improves generalization across diverse regulatory sequences. Designed to be extensible, PIONEER currently implements a specific set of strategies but can readily incorporate additional sequence generation and selection methods. The framework adheres to FAIR software practices, ensuring accessibility, interoperability, and reproducibility for broad community adoption.

Sampling sequence space. To efficiently explore sequence space, PIONEER implements multiple sequence proposal strategies, including random sequence generation, partial random mutagenesis, and uncertainty-guided mutagenesis (UGM). Random sequence generation has been successfully used for dataset enrichment, such as in yeast promoter studies (Rafi et al., 2024), while partial random mutagenesis introduces mutations at a fixed rate (1–25%), facilitating controlled exploration of local sequence neighborhoods. Unlike these approaches, which introduce mutations randomly, UGM actively selects mutations that maximize a model's predictive uncertainty. UGM operates as a sequence optimization process, where uncertainty estimates guide mutation selection. A certain number of mutations, given by the mutation rate, is chosen based on its predicted effect (via uncertainty backpropagation to the inputs) on increasing uncertainty. By targeting regions where the model is least confident, UGM systematically explores unexplored sequence space while refining predictions.

PIONEER supports multiple approaches for estimating predictive uncertainty. Deep Ensembles (Lakshminarayanan et al., 2017) aggregate predictions from independently trained models to quantify epistemic uncertainty, while Monte Carlo (MC) dropout (Gal & Ghahramani, 2016) approximates uncertainty by performing multiple stochastic forward passes with dropout active during inference and computing the prediction variance. MC dropout offers a computationally efficient alternative to Deep Ensembles, achieving comparable performance on held-out in-distribution data (Appendix Fig. 6).

Acquisition function. Once candidate sequences are generated, an acquisition function determines which ones will be included in the next training batch. PIONEER implements three acquisition strategies: random sampling, uncertainty maximization, and largest cluster maximum distance (LCMD) (Holzmüller et al., 2023). Uncertainty maximization ranks sequences by their predictive uncertainty and selects the highest-ranked ones, though evaluating each sequence independently may lead to redundant choices. To promote diversity, LCMD—developed for regression tasks (Holzmüller et al., 2023)—balances informativeness, diversity, and representativeness when selecting batches. Moreover, PIONEER's extensible framework allows seamless integration of other batch-selection methods, such as BatchBALD (Kirsch et al., 2019) and BADGE (Ash et al., 2020), offering flexibility to meet diverse active learning objectives.

3 EXPERIMENTAL OVERVIEW

We applied PIONEER to benchmark sequence proposal strategies across three regulatory genomics tasks, using two distinct biological systems: lentiMPRA in human K562 cells and STARR-seq in Drosophila S2 cells. These datasets capture different aspects of gene regulation, with lentiMPRA measuring chromatin-integrated cis-regulatory activity and STARR-seq assessing enhancer activity in an episomal context. Evaluating PIONEER in these settings allowed us to assess its effectiveness in improving model generalization across diverse regulatory landscapes.

Fly developmental enhancer activity with STARR-seq. The *Drosophila melanogaster* STARR-seq dataset comprises 249 bp enhancer sequences measured under two promoter contexts (developmental and housekeeping) (de Almeida et al., 2022). To simplify analysis, we treated each promoter independently and modeled developmental enhancer activity as a single-task regression. The dataset was partitioned into 402,296 training, 41,186 validation, and 40,570 test sequences.

Human regulatory sequences with lentiMPRA. The K562 lentiMPRA dataset consists of 230 bp cis-regulatory sequences, each associated with a scalar activity measurement (Agarwal et al., 2023). To prevent data leakage, forward and reverse complement sequences were assigned to the same split. The dataset was divided into 180,564 training, 22,570 validation, and 22,571 test sequences.

Models. For STARR-seq, we employed DeepSTARR (de Almeida et al., 2022), a convolutional network optimized for enhancer activity prediction. For lentiMPRA, we used LegNet (Penzar et al., 2023), an EfficientNetV2-inspired convolutional model that achieved state-of-the-art performance on this dataset. To enable uncertainty quantification with MC Dropout, we modified LegNet by adding dropout layers after each 1D convolution (0.1 probability), after 1D max pooling (0.1 probability), and after the final dense layer (0.5 probability). Both models were trained using Adam (learning rate 0.001, weight decay 1e-6) with a batch size of 100, a maximum of 100 epochs, early stopping (patience 10 epochs), and a learning rate decay factor of 0.2 (patience 5 epochs).

Oracles. An *in silico* oracle was generated by training an ensemble of five independent models on the full training set using EvoAug (Lee et al., 2023), an evolution-inspired data augmentation strategy that improved performance relative to standard training (see Appendix Table 1). The ensemble's average predictions were used to provide functional labels for proposed sequences.

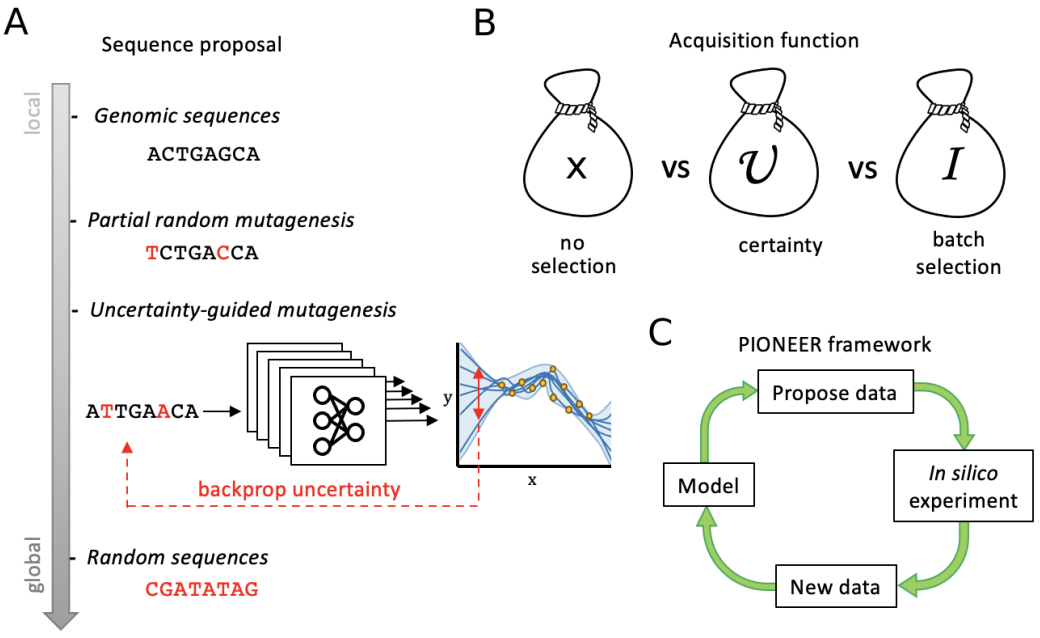

Figure 1: A) Schematic of the different types of sequence proposal strategies stratified by increasing order of covariate shift (from "local" to "global") with respect to a reference genome. B) The three different types of acquisition funciton available in PIONEER: no selection, uncertainty-based selection, or batch selection. C) Schematic representation of the PIONEER workflow: a model can be used to propose new data (e.g. by looking at the uncertainty of possible sequences), these get annotated through an *in silico* experiment (e.g., using an *in silico* oracle), and finally the new data can be leverage for a new cycle of model training.

PIONEER settings. To evaluate how different sequence proposal strategies affect model generalization, we employed an iterative active learning framework. First, we randomly downsampled the original training set to 20,000 sequences to simulate a small-data regime. Each active learning cycle comprised:

1. **Candidate Sequence Generation**: We generated 100,000 candidate sequences using one of four sampling strategies: Random, Partial Random Mutagenesis (5% mutation rate), Uncertainty-Guided Mutagenesis (UGM) (5% mutation rate), or All, which included an equal mixture of random, mutagenized, and UGM-generated sequences. UGM uses the standard deviation on five different inferences using MC Dropout as a measure of uncertainty. The 5% mutation rate was chosen based on a benchmark against 10% and 25% (Appendix Fig. 6). Additionally, drawing more sequences from the original dataset was included as a control to simulate an (unrealistic) scenario where additional genomic sequences that have similar properties as the original training distribution could be sampled.
2. **Sequence Selection**: 20,000 sequences were selected from the candidate pool using one of three acquisition strategies: No Selection (20,000 sequences generated initially), Uncertainty-Based Selection (selecting sequences with the highest predictive uncertainty), and Batch Selection with LCMD (optimizing for informativeness, diversity, and representativeness).
3. **Labeling and Model Update**: The selected sequences were labeled using an *in silico* oracle and added to the training set.
4. **Retraining**: The model was trained from scratch using the updated training set, and the process was repeated.

Each combination of sampling and acquisition strategy was tested over five active learning cycles, and model performance was evaluated at each cycle to measure generalization under covariate shifts (see Evaluation below).

Evaluation. To assess how different sequence proposal strategies impact model generalization, we evaluated performance under four levels of covariate shift (Fig. 2A), each simulating a distinct downstream application:

- **No shift (in-distribution generalization)**: Models were evaluated on a held-out test set drawn from the same distribution as the training data, assessing generalization to new sequences with similar evolutionary constraints. This reflects a common real-world scenario where newly encountered sequences belong to the same genomic context as the training data.
- **Small shift (near-distribution variation)**: To test robustness to minor genetic variation, we applied a single round of 5% partial random mutagenesis to test sequences, repeated five times with different random seeds and aggregated into the final test set. This simulates small-scale sequence perturbations relevant to variant effect prediction and population-level genetic diversity.
- **Large shift (low activity, out-of-distribution random sequences)**: To evaluate generalization beyond naturally occurring sequences, we tested models on a set of entirely synthetic random sequences, which lacked evolutionary constraints. The *in silico* oracle predicted these sequences to have low regulatory activity (Fig. 2B), representing a scenario where models encounter functionally inactive sequences that do not resemble training examples.
- **Large shift (high activity, optimized sequences)**: To assess a model's ability to extrapolate to de novo sequence design, we applied an iterative *in silico* evolution procedure to the same randomly generated sequences from the previous setting Vaishnav et al. (2022). At each step, a single nucleotide mutation was introduced, selecting mutations that maximized predicted activity. This setting mimics sequence design tasks, where the goal is to evolve highly active regulatory sequences from a random sequence (Fig. 2B).

These scenarios capture a spectrum of real-world applications, from modeling natural variation (no shift, small shift) to designing high-activity regulatory elements (large shift, high activity). Although

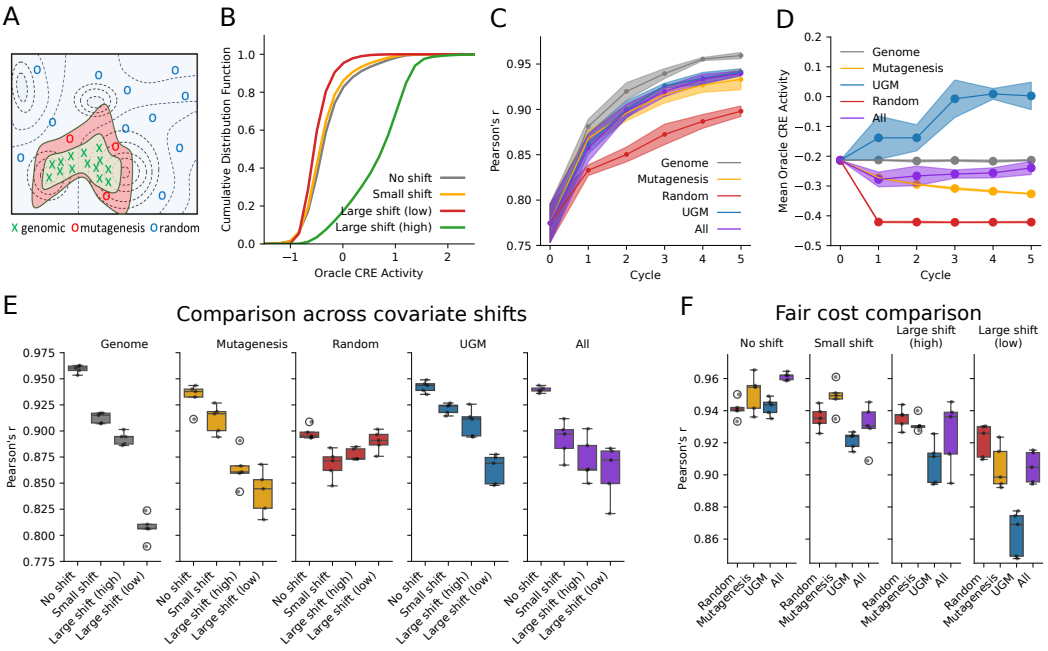

Figure 2: Performance comparison for K562 leniMPRA data. A) Schematic of sequence-function landscape highlighting covariate shifts: genomic sequences (green crosses), partially mutagenized sequences (red circles), and random sequences (blue circles). B) Cumulative distribution function of the oracle CRE activity for the four test sets: in-distribution genome sequences (no shift), partially mutagenized sequences (small shift), random sequences (large shift, low activity), and random sequences evolved for high activity (large shift, high activity). C) Performance as a function of cycles of the PIONEER pipeline for different sequence proposal methods. D) Average oracle-predicted activity at each cycle for different sequence proposal methods. D) Performance for different sequence proposal methods after 5 AI-experiment cycles for the different test sets. E) Performance for different sequence proposal methods after 5 AI-experiment cycles for the different test sets in the context of a fair-price comparison.

the large shift (low activity) setting is not directly applicable, it provides insights into how models handle functionally inactive sequences and defines the limits of extrapolation beyond training data.

4 Generalization Across Covariate Shifts

We found that iteratively augmenting the training set with new sequences consistently improved performance on held-out genomic test sets, regardless of the sequence proposal strategy, although gains eventually plateaued (Fig. 2C, Appendix Fig. 3C). The rate of improvement varied with the sequence proposal method employed. For instance, models supplemented with additional genome-derived sequences achieved the highest accuracy on genomic test sets, confirming that aligning training and test distributions optimizes in-distribution performance. However, this approach alone struggled to generalize beyond naturally occurring sequences, limiting its utility for de novo sequence design (Fig. 2E, Appendix Fig. 3E).

For tasks involving sparse mutations, UGM produced the greatest performance gains. Models trained with this method outperformed alternatives in predicting the effects of small sequence variations (Fig. 2E, Appendix Fig. 3E), demonstrating that selecting sequences based on predictive uncertainty effectively identifies functionally relevant mutations. This highlights the utility of active learning in applications such as personalized medicine and evolutionary studies, where accurate predictions of small sequence variations are critical.

For generalization to synthetic regulatory elements, UGM again yielded the strongest improvements on out-of-distribution sequences. Although random sequence generation introduced the greatest

diversity, it did not consistently enhance functional predictions. Notably, random sequences performed best under conditions where most sequences exhibited low regulatory activity—a scenario misaligned with design objectives favoring high-activity elements (Fig. 2E and Appendix Fig. 3D). In contrast, UGM preferentially selected sequences with higher predicted activity, contributing to improved generalization for high-activity design.

We further benchmarked acquisition functions to determine their impact on sequence selection. Across all covariate shifts, acquisition strategies had only modest effects (Appendix Figs. 4 and 5), suggesting that the sequence proposal method is the primary driver of improved generalization. Within the mixed proposal strategy ("All"), batch selection via LCMD consistently prioritized UGM over other methods (Appendix Fig. 7). Early in training, random sequences were frequently selected for their novelty; however, as the model gained more knowledge through iterative refinement, partial random mutagenesis became increasingly informative, indicating that the most valuable sequences shift from being entirely novel to introducing subtle, functionally meaningful perturbations.

Taken together, these results likely reflect the model's evolving sensitivity to fine-grained sequence variations. Early in training, random sequences broaden the model's exposure to diverse patterns; as learning progresses and regulatory motifs become better defined, sequences from partial mutagenesis more effectively refine predictions. In other words, indiscriminate dataset expansion is insufficient for robust generalization—data augmentation must be adaptive, balancing broad exploration with targeted refinement. Our findings indicate that while genome-derived sequences optimize in-distribution performance, uncertainty-guided mutagenesis enables effective extrapolation to novel regulatory sequences. These insights provide a framework for optimizing training sets in de novo sequence design, where dynamic sequence selection is critical for achieving robust model generalization.

5 Cost-aware AI-experiment cycles

Experimental costs impose substantial constraints on large-scale data acquisition. To evaluate sequence proposal strategies under realistic budgetary constraints, we designed cost-equivalent experiments in which the total expense of generating and assaying sequences was held constant. Recognizing that synthesizing random or mutagenized sequences is considerably cheaper than producing computationally optimized sequences, we adjusted the number of sequences per cycle accordingly.

In each active learning cycle, we scaled up the number of random and mutagenized sequences relative to uncertainty-guided mutagenesis (UGM):

- Random and mutagenized sequences: 100,000 sequences were added per cycle.
- UGM-selected sequences: 20,000 sequences per cycle.
- Mixed-sequence strategy ("All"): 25,000 random sequences, 25,000 mutagenized sequences, and 10,000 UGM-generated sequences, totaling 60,000 sequences – comparable in sequencing cost to the previous two methods.

This design allowed a direct comparison between larger datasets of less curated sequences and smaller, actively optimized datasets.

Surprisingly, models trained on higher-quantity but lower-curation datasets often generalized better than those trained on fewer, computationally designed sequences (Fig. 2F and Appendix Fig. 3F). These results suggest that in resource-limited settings, prioritizing data quantity over extensive computational optimization can be a more effective strategy for improving model performance.

The scaling factor of five between targeted mutagenesis and random/mutagenized sequences represents a conservative lower bound, as random sequences can be generated in much larger quantities per experimental batch. Even at this threshold, increasing the volume of random sequences outperformed a smaller number of UGM-selected sequences. Mixed-sequence strategies, which combine random, mutagenized, and UGM-selected sequences, offered a practical balance between cost-efficiency and predictive accuracy.

DISCUSSION

This study highlights key trade-offs between sequence proposal strategies and their impact on model generalization. Genome-derived sequences excel at in-distribution tasks but fail to extrapolate beyond naturally occurring regulatory elements. Mutagenesis enables generalization to nearby sequence variants, making it valuable for applications such as variant effect prediction. Random sequence generation introduces the greatest sequence diversity and is most effective for exploring novel sequence space, but it shifts the predicted activity distribution, often biasing models toward lower-activity sequences. These findings emphasize the need to align training data composition with the specific objectives of a given modeling task. At the same time, they demonstrate that hybrid approaches, which integrate multiple sequence proposal strategies, can improve generalization more effectively than any single method alone.

PIONEER provides a systematic framework for evaluating these trade-offs, enabling iterative AI-experiment cycles to assess how different sequence proposal strategies influence generalization. Our results illustrate that different applications require distinct data augmentation strategies: genome-derived sequences optimize in-distribution performance but fail to support generalization beyond known sequences, while uncertainty-guided mutagenesis maintains alignment with the training distribution while promoting exploration of functionally relevant sequence variants. These insights suggest that future iterations of PIONEER could explore adaptive sequence proposal strategies that dynamically balance random sampling, uncertainty-based selection, and targeted mutagenesis based on task-specific objectives.

Beyond comparing sequence proposal strategies, this study underscores the influence of dataset size on generalization. In cost-aware AI-experiment cycles—where total sequencing costs were held constant—larger datasets of randomly generated or mutagenized sequences often outperformed smaller datasets enriched with computationally optimized sequences. This suggests that in resource-limited settings, prioritizing data volume—even with less computationally "informative" sequences—can provide greater benefits than relying exclusively on guided sequence selection. However, the trade-off between data quantity and quality is context-dependent, and future studies should explore where diminishing returns set in for different regulatory genomics applications.

Designed as an extensible and FAIR (findable, accessible, interoperable, and reproducible) framework, PIONEER allows researchers to incorporate additional sequence proposal methods, alternative in silico oracles, and novel acquisition functions. While this study focused on regulatory genomics, the framework is broadly applicable to other sequence-based modeling challenges, including synthetic biology and protein engineering. The structure of the sequence-function landscape—whether smooth or rugged—varies depending on the choice of oracle and dataset, shaping the performance of different sequence proposal strategies. By systematically evaluating these factors, PIONEER provides a scalable platform for investigating how training data composition affects model generalization across diverse biological and synthetic sequence landscapes.

Moving forward, PIONEER can be expanded in several directions. First, integrating population-based optimization methods could further refine sequence proposal strategies by incorporating adaptive search mechanisms (Angermueller et al., 2020). Second, future studies could explore hybrid sequence proposal strategies that combine exploration (random sequence generation) and exploitation (targeted mutagenesis or uncertainty-based selection) within a single AI-experiment cycle. Finally, PIONEER offers a testbed for developing improved active learning strategies that explicitly balance informativeness, diversity, and cost-effectiveness. Addressing these challenges will enhance the next generation of AI-driven models for genomics, synthetic biology, and precision medicine.

DATA AVAILABILITY

Training data for lentiMPRA and STARR-seq and all model weights are available on Zenodo at: `https://doi.org/10.5281/zenodo.15045788`.

CODE AVAILABILITY

PIONEER is installable via `pip` (PyPI: `https://pypi.org/project/pioneer-nn/`) and an open-source version is available via GitHub: `https://github.com/p-koo/pioneer-nn`. A code repository for reproducibility of the analysis can be found via GitHub: `https://github.com/alescrnjar/PIONEER_reproducibility`.).

ACKNOWLEDGMENTS

This work was supported in part by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory, NIH grants R01HG012131 (PKK, JBK, and AC), R01GM149921 (PKK and AC), R35GM133777 (JJD, JBK) and R01HG011787 (JJD, JBK). Computations were performed using equipment supported by NIH grant S10OD028632.

REFERENCES

Vikram Agarwal, Fumitaka Inoue, Max Schubach, Beth K. Martin, Pyaree Mohan Dash, Zicong Zhang, Ajuni Sohota, William Stafford Noble, Galip Gürkan Yardimci, Martin Kircher, Jay Shendure, and Nadav Ahituv. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. *bioRxiv*, 2023. doi: 10.1101/2023.03.05.531189. URL https://www.biorxiv.org/content/early/2023/03/06/2023.03.05.531189.

Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and Philip S. Yu. Active Learning: A Survey. In *Data Classification*. Chapman and Hall/CRC, 2014. ISBN 978-0-429-10263-9.

Christof Angermueller, David Belanger, Andreea Gane, Zelda Mariet, David Dohan, Kevin Murphy, Lucy Colwell, and D Sculley. Population-based black-box optimization for biological sequence design. 2020.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds, 2020. URL https://arxiv.org/abs/1906.03671.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, October 2021a.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53 (3):354–366, Mar 2021b. ISSN 1546-1718. doi: 10.1038/s41588-021-00782-6. URL https://doi.org/10.1038/s41588-021-00782-6.

Michael Bailey, Saeed Moayedpour, Ruijiang Li, Alejandro Corrochano-Navarro, Alexander Kötter, Lorenzo Kogler-Anele, Saleh Riahi, Christoph Grebner, Gerhard Hessler, Hans Matter, Marc Bianciotto, Pablo Mas, Ziv Bar-Joseph, and Sven Jager. Deep batch active learning for drug discovery. *bioRxiv*, 2023. doi: 10.1101/2023.07.26.550653. URL https://www.biorxiv.org/content/early/2023/10/26/2023.07.26.550653.

Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7):940–949, Jul 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01102-2. URL https://doi.org/10.1038/s41588-022-01102-2.

Bernardo P. de Almeida, Franziska Reiter, Michaela Pagani, and Alexander Stark. Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, 54(5):613–624, May 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01048-5. URL https://doi.org/10.1038/s41588-022-01048-5.

Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, Wei Liu, Jialie Shen, and Dacheng Tao. Exploring representativeness and informativeness for active learning. *IEEE Transactions on Cybernetics*, 47(1):14–26, 2017. doi: 10.1109/TCYB.2015.2496974.

Ryan Z. Friedman, Avinash Ramu, Sara Lichtarge, Yawei Wu, Lloyd Tripp, Daniel Lyon, Connie A. Myers, David M. Granas, Maria Gause, Joseph C. Corbo, Barak A. Cohen, and Michael A. White. Active learning of enhancers and silencers in the developing neural retina. *Cell Systems*, 16(1): 101163, 2025. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2024.12.004. URL https://www.sciencedirect.com/science/article/pii/S2405471224003685.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. URL https://arxiv.org/abs/1506.02142.

Rohan Gorantla, Alžbeta Kubincová, Benjamin Suutari, Benjamin P. Cossins, and Antonia S. J. S. Mey. Benchmarking active learning protocols for ligand binding affinity prediction. *bioRxiv*, 2023. doi: 10.1101/2023.11.24.568570. URL https://www.biorxiv.org/content/early/2023/11/24/2023.11.24.568570.

David Holzmüller, Viktor Zaverkin, Johannes Kästner, and Ingo Steinwart. A framework and benchmark for deep batch active learning for regression, 2023. URL https://arxiv.org/abs/2203.09410.

Connie Huang, Richard W Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail, and Nilah M Ioannidis. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55(12):2056–2059, December 2023.

Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with gflownets, 2023. URL https://arxiv.org/abs/2203.04115.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, 2019. URL https://arxiv.org/abs/1906.08158.

Peter K Koo, Antonio Majdandzic, Matthew Ploenzke, Praveen Anand, and Steffan B Paul. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. 2020. doi: 10.1101/2020.09.08.288068. URL https://doi.org/10.1101/2020.09.08.288068.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Nicholas Keone Lee, Ziqi Tang, Shushan Toneyan, and Peter K. Koo. Evoaug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations. *Genome Biology*, 24(1):105, May 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02941-w. URL https://doi.org/10.1186/s13059-023-02941-w.

Johannes Linder and Georg Seelig. Fast activation maximization for molecular sequence design. *BMC Bioinformatics*, 22:1–20, 12 2021. ISSN 14712105. doi: 10.1186/S12859-021-04437-5/FIGURES/4. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04437-5.

Alyssa K. Morrow, Ashley Thornal, Elise Duboscq Flynn, Emily Hoelzli, Meimei Shan, Görkem Garipler, Rory Kirchner, Aniketh Janardhan Reddy, Sophia Tabchouri, Ankit Gupta, Jean-Baptiste Michel, and Uri Laserson. Ml-driven design of 3' utrs for mrna stability. *bioRxiv*, 2024. doi: 10.1101/2024.10.07.616676. URL https://www.biorxiv.org/content/early/2024/10/07/2024.10.07.616676.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, Jan 2022. ISSN 1573-0565. doi: 10.1007/s10994-021-06003-9. URL https://doi.org/10.1007/s10994-021-06003-9.

Dmitry Penzar, Daria Nogina, Elizaveta Noskova, Arsenii Zinkevich, Georgy Meshcheryakov, Andrey Lando, Abdul Muntakim Rafi, Carl de Boer, and Ivan V Kulakovskiy. LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics*, 39(8):btad457, 07 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad457. URL https://doi.org/10.1093/bioinformatics/btad457.

Jeff M. Phillips. Coresets and sketches, 2016. URL https://arxiv.org/abs/1601.00617.

Abdul Muntakim Rafi, Daria Nogina, Dmitry Penzar, Dohoon Lee, Danyeong Lee, Nayeon Kim, Sangyeup Kim, Dohyeon Kim, Yeojin Shin, Il-Youp Kwak, Georgy Meshcheryakov, Andrey Lando, Arsenii Zinkevich, Byeong-Chan Kim, Juhyun Lee, Taein Kang, Eeshit Dhaval Vaishnav, Payman Yadollahpour, , Sun Kim, Jake Albrecht, Aviv Regev, Wuming Gong, Ivan V. Kulakovskiy, Pablo Meyer, and Carl de Boer. Evaluation and optimization of sequence-based gene regulatory deep learning models. *bioRxiv*, 2024. doi: 10.1101/2023.04.26.538471. URL https://www.biorxiv.org/content/early/2024/02/17/2023.04.26.538471.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning, 2021.

Alexander Sasse, Bernard Ng, Anna E Spiro, Shinya Tasaki, David A Bennett, Christopher Gaiteri, Philip L De Jager, Maria Chikina, and Sara Mostafavi. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics*, 55(12):2060–2064, December 2023.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *Advances in Intelligent Data Analysis, 4th International Conference, IDA 2001, Cascais, Portugal, September 13-15, 2001, Proceedings*, volume 2189 of *Lecture Notes in Computer Science*, pp. 309–319. Springer, 2001.

Burr Settles and Mark W. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2008. URL https://api.semanticscholar.org/CorpusID:8197231.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Ziqi Tang, Shushan Toneyan, and Peter K. Koo. Current approaches to genomic deep learning struggle to fully capture human genetic variation. *Nature Genetics*, 55(12):2021–2022, Dec 2023. ISSN 1546-1718. doi: 10.1038/s41588-023-01517-5. URL https://doi.org/10.1038/s41588-023-01517-5.

Shushan Toneyan, Ziqi Tang, and Peter K. Koo. Evaluating deep learning for predicting epigenomic profiles. *Nature Machine Intelligence*, 4(12):1088–1100, Dec 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00570-9. URL https://doi.org/10.1038/s42256-022-00570-9.

Eeshit Dhaval Vaishnav, Carl G de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A Thompson, Joshua Z Levin, Francisco A Cubillos, and Aviv Regev. The evolution, evolvability and engineering of gene regulatory DNA. *Nature*, 603(7901):455–463, March 2022.

Derek van Tilborg and Francesca Grisoni. Traversing chemical space with active deep learning for low-data drug discovery. *Nature Computational Science*, 4(10):786–796, Oct 2024. ISSN 2662-8457. doi: 10.1038/s43588-024-00697-2. URL https://doi.org/10.1038/s43588-024-00697-2.

Jingtao Wang, Gregory Fonseca, and Jun Ding. scsemiprofiler: Advancing large-scale single-cell studies through semi-profiling with deep generative models and active learning. *bioRxiv*, 2023. doi: 10.1101/2023.11.20.567929. URL https://www.biorxiv.org/content/early/2023/11/21/2023.11.20.567929.

Xueying Zhan, Huan Liu, Qing Li, and Antoni B. Chan. A comparative survey: Benchmarking for pool-based active learning. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4679–4686. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/634. URL https://doi.org/10.24963/ijcai.2021/634. Survey Track.

A APPENDIX

LegNet, K562	DeepSTARR, dev. enh.
0.7917	0.7044
0.7881	0.7072
0.7891	0.7059
0.7892	0.7068
0.7952 0.7068	

Table 1: Oracle Pearson's r for the five oracles for LegNet (in K562 cells) and for DeepSTARR (fly developmental enhancers).

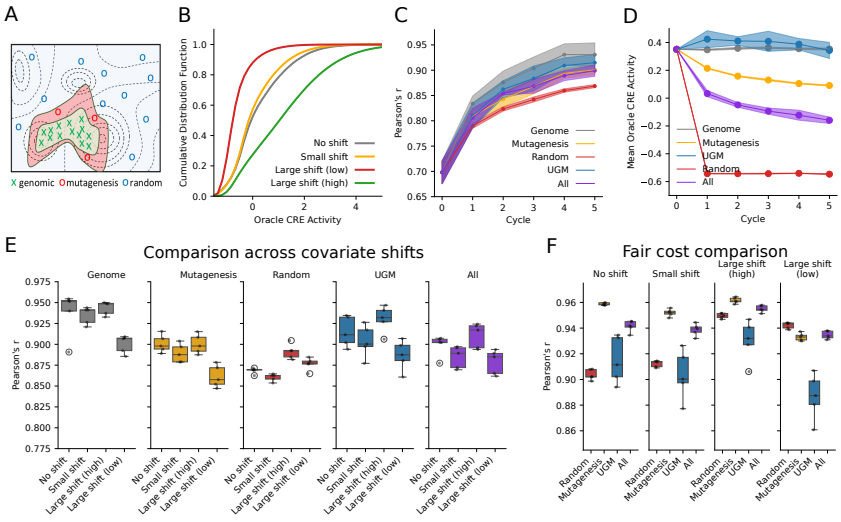

Figure 3: Performance comparison for Drosophila S2 developmental enhancer STARR-seq data. A) Schematic of sequence-function landscape highlighting covariate shifts: genomic sequences (green crosses), partially mutagenized sequences (red circles), and random sequences (blue circles). B) Cumulative distribution function of the oracle CRE activity for the four test sets: in-distribution genome sequences (no shift), partially mutagenized sequences (small shift), random sequences (large shift, low activity), and random sequences evolved for high activity (large shift, high activity). C) Performance as a function of cycles of the PIONEER pipeline for different sequence proposal methods. D) Average oracle-predicted activity at each cycle for different sequence proposal methods. D) Performance for different sequence proposal methods after 5 AI-experiment cycles for the different test sets. E) Performance for different sequence proposal methods after 5 AI-experiment cycles for the different test sets in the context of a fair-price comparison.

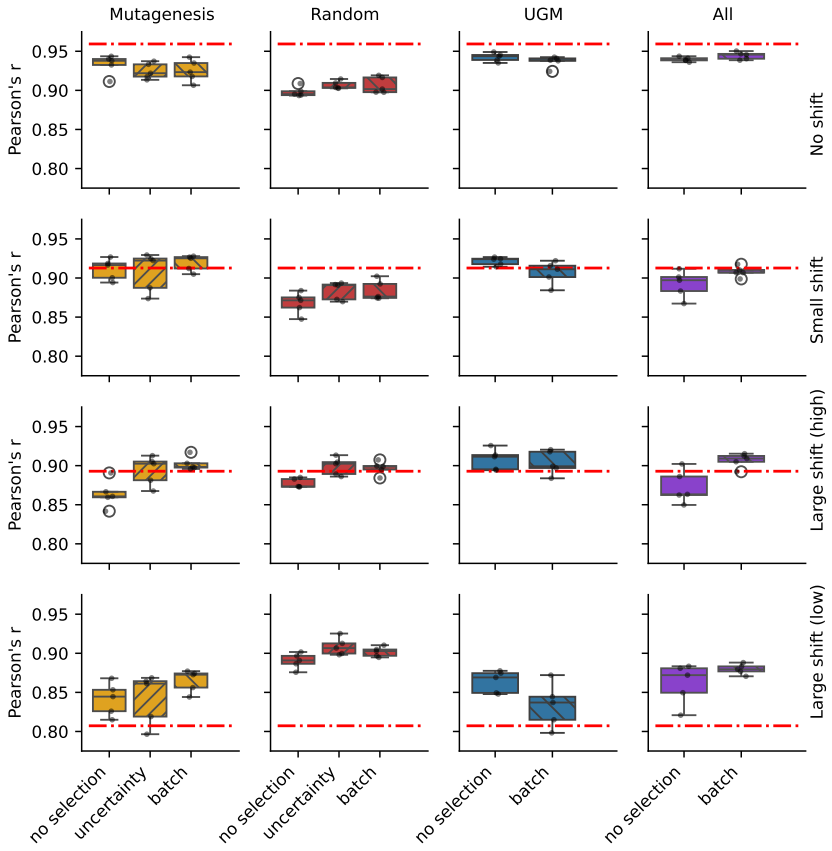

Figure 4: Performance comparison for the last cycle for different sequence proposal methods (partial mutagenesis, random, UGM, and All) with different acquisition functions (no selection, uncertainy-based selection, batch selection) for different covariate shifts (no shift, small shift, large shift with high activity, large shift with low activity) in human K562 cells. Dashed red lines represent the mean for the genome sequence proposal method.

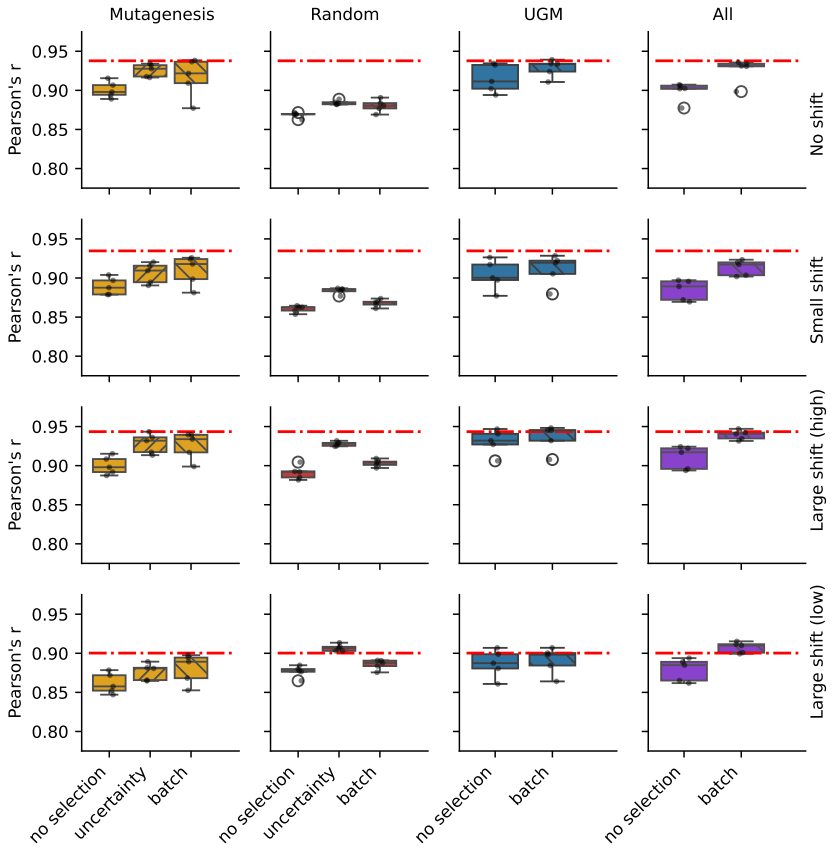

Figure 5: DNN performance, as Pearsons's r, for the last PIONEER cycle for different sequence proposal methods (partial mutagenesis, random, UGM, and All) with different acquisition functions (no selection, uncertainy-based selection, batch selection) for different covariate shifts (no shift, small shift, large shift with high activity, large shift with low activity) in fly developmental enhancers. Dashed red lines represent the mean for the genome sequence proposal method.

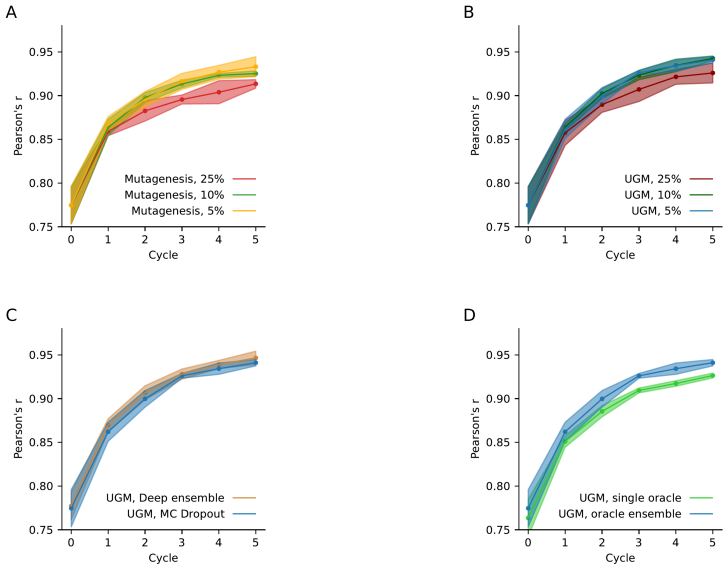

Figure 6: Performance comparison of different hyperparameter choices for human K562 cells. A) Partial random mutagenesis at different mutation rates (5%, 10%, 25%). B) UGM at different mutation rates (5%, 10%, 25%). C) UGM as modeled with two different uncertainties: standard deviation from a Deep Ensemble, or from Monte Carlo Dropout. D) UGM as modeled with single oracle or with an oracle ensemble.

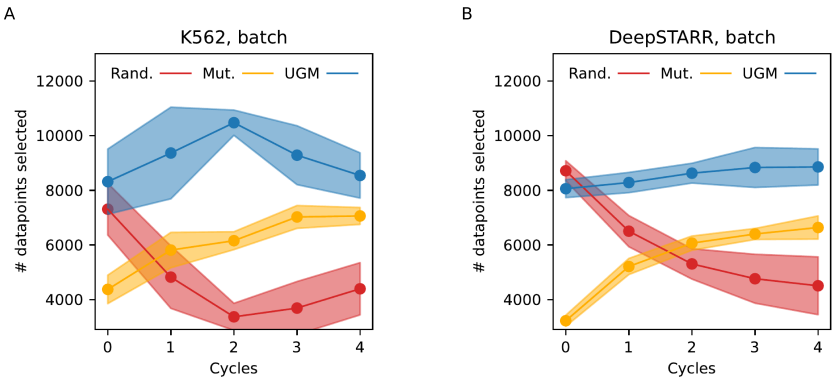

Figure 7: Selection comparison given by LCMD based on "All + batch" in (A) K562 cells and (B) fly developmental enhancers.