Active learning for efficient discovery of optimal gene combinations in the combinatorial perturbation space

Jason Qin
Neptune Bio
New York, NY
jason@neptune.bio

Hans-Hermann Wessels
Neptune Bio
New York, NY
harm@neptune.bio

Carlos Fernandez-Granda
Center for Data Science
Courant Institute of Mathematical Sciences
New York University
New York, NY
cfgranda@cims.nyu.edu

Yuhan Hao
Neptune Bio
New York, NY
yuhan@neptune.bio

Abstract

The advancement of novel combinatorial CRISPR screening technologies enables the identification of synergistic gene combinations on a large scale. This is crucial for developing novel and effective combination therapies, but the combinatorial space makes exhaustive experimentation infeasible. We introduce NAIAD, an active learning framework that efficiently discovers optimal gene pairs capable of driving cells toward desired cellular phenotypes. NAIAD leverages single-gene perturbation effects and adaptive gene embeddings that scale with the training data size, mitigating overfitting in small-sample learning while capturing complex gene interactions as more data is collected. Evaluated on four CRISPR combinatorial perturbation datasets totaling over 350,000 genetic interactions, NAIAD, trained on small datasets, outperforms existing models by up to 40% relative to the second-best. NAIAD's recommendation system prioritizes gene pairs with the maximum predicted effects, resulting in the highest marginal gain in each AI-experiment round and accelerating discovery with fewer CRISPR experimental iterations. Our NAIAD framework (https://github.com/NeptuneBio/NAIAD) improves the identification of novel, effective gene combinations, enabling more efficient CRISPR library design and offering promising applications in genomics research and therapeutic development.

1 Introduction

Targeting multiple genes through drug combinations or polypharmacology offers a transformative therapeutic approach for developing effective treatments across diverse medical fields, including oncology (Al-Lazikani, Banerji, and Workman 2012; Mokhtari et al. 2017), infectious diseases (Hammond et al. 2022; Shyr et al. 2021), and metabolic disorders (Samms, Coghlan, and Sloop 2020; Jastreboff et al. 2023). Combinatorial gene perturbations can yield additive or synergistic effects, enhancing therapeutic outcomes beyond what is achievable with single-gene targeting ("Rationalizing Combination Therapies" 2017; Hwangbo et al. 2023). One of the most notable successes in transitioning cellular phenotype is the discovery of the Yamanaka factors—a specific combination of four transcription factors capable of reprogramming differentiated cells back to a pluripotent state

AI for New Drug Modalities at NeurIPS 2024.

(Takahashi and Yamanaka 2006). This groundbreaking achievement demonstrates the significant potential within the combinatorial perturbation space to engineer cellular phenotypes.

The critical question now is how to systematically identify additional effective combinatorial perturbations that can transform cells to achieve desired phenotypes. Comprehensively exploring this huge space presents an inherent mathematical challenge due to the exponential growth of the number of possible combinations. With approximately 20,000 protein-coding genes in the human genome, the total number of two-gene combinations approaches 200 million, and four-gene combinations exceeds 6 quadrillion (10^{15}). Experimentally testing all possible combinations is infeasible. Therefore, developing computational models that can predict the most effective gene combinations is essential for the efficient identification of the most potent combinatorial perturbations. However, the combinatorial space makes it infeasible to obtain sufficient data points to train effective models through experiments alone. Active learning frameworks (Eisenstein 2020), such as the AI + Experiment Loop (Rood, Hupalowska, and Regev 2024) offer a promising solution by enabling efficient exploration of this space. In our proposed framework we initially train a model on a small dataset from experiments, which enables it to predict unseen combinatorial perturbations effects across the entire combinatorial space. These predictions guide the design of subsequent CRISPR screening libraries for targeted experiments, thus allowing us to iteratively refine the model and converge on identifying the most effective gene combinations (Figure 1A).

In this work, we focus on the effects of 2-gene combinations and introduce a novel active learning framework, NAIAD, to accelerate the discovery of optimal gene pairs (Figure 1B). Our key contributions are:

(1) A novel combinatorial perturbation model which incorporates adaptive gene embeddings that scale with the training data size, along with an overparametrized representation of single-gene perturbation effects.

(2) Maximum Predicted Effects (MPE)-based recommendation system that suggests gene combinations for subsequent CRISPR library design, facilitating the discovery of synergistic and effective gene combinations.

(3) An AI + Lab active learning framework that effectively identifies optimal gene combinations, significantly reducing the number of experimental iterations needed to achieve robust results

2 Related Work

CRISPR combinatorial perturbation technologies can be broadly classified into two main categories (Norman et al. 2019): single-cell combinatorial perturbation and bulk combinatorial perturbation. Single-cell combinatorial perturbation measures the entire transcriptome, capturing comprehensive gene expression changes in individual cells, but with a limited number of gene combinations. In contrast, bulk combinatorial perturbation focuses on measuring a single phenotype, enabling the investigation of a much broader range of gene combinations.

Predicting combinatorial perturbations has been a significant challenge due to non-linearity of certain gene combinations. Various machine learning approaches have been proposed to address this problem using single-cell combinatorial perturbation data. Variational Autoencoders (VAEs) (Kingma and Welling 2013) have been employed to model genetic and chemical combinatorial perturbations by simultaneously learning embeddings of single perturbations and capturing non-linear interactions in methods such as CPA (Lotfollahi et al. 2023), ComβVAE (Geiger-Schuller et al. 2023), sVAE+ (Lopez et al. 2022), and SAMS-VAE (Bereket and Karaletsos 2023). These approaches facilitate the modeling of complex relationships between genes or compounds within the latent space of embeddings. Methods such as sVAE+ (Lopez et al. 2022) and SAMS-VAE (Bereket and Karaletsos 2023) have been developed to model sparsity in the latent variable intervention effects. By disentangling the perturbation-related sparse latent space, these models effectively identify critical features and interactions within high-dimensional biological data. SALT&PEPER implemented a method to separately learn linear and non-linear effects of gene perturbations (Gaudelet et al. 2024). By using gene-embedding-based autoencoder models, they effectively decomposed the interactions, enabling more interpretable models of gene effects. Recently, several single-cell foundation models, such as scGPT (Cui et al. 2024) and scFoundation (Hao et al. 2024), trained on all publicly available observational data, have demonstrated their ability to predict cellular responses following

perturbations after model fine tuning. Additionally, GEARS leverages graph neural networks (GNNs) (Kipf and Welling 2016) to incorporate prior biological knowledge into the network architecture (Roohani, Huang, and Leskovec 2023). GNNs facilitate the inference of gene-gene interactions by leveraging known pathways and interaction networks, thus enhancing the prediction of combinatorial effects.

However, most of these methods primarily focus on predicting gene expression profiles resulting from given perturbations and do not extend to the prediction of phenotypic outcomes. Also, the number of gene combinations from single-cell combinatorial perturbations dataset is very limited (100-200 gene combinations), so it is difficult to evaluate if those deep learning models are generalizable in the entire combinatorial perturbation space and can outperform linear models (Ahlmann-Eltze, Huber, and Anders 2024).

Some existing methods (i.e. GEARS (Roohani, Huang, and Leskovec 2023) and CPA (Lotfollahi et al. 2023)), are capable of predicting combinatorial perturbations from single-cell transcriptomic profiles, as well as from single measurements derived from bulk screens (i.e. cell viability). However, these approaches generally assume the availability of sufficient data that would allow for a substantial portion to be used for training. In practice, we are often limited by the number of training samples due to the exponential growth of possible combinations in combinatorial perturbation data. This limitation necessitates having methods that can perform well with minimal data. Active learning frameworks offer a solution by optimizing model performance while using the least amount of training data possible. RECOVER utilized an active learning framework that iteratively selects the most promising drug combinations for testing through the AI + Lap loop (Bertin et al., 2023). RECOVER applies a bilinear operator to create permutation-invariant representations that integrate multiple single perturbation effects, learning the non-linear components of drug combinations. It leverages ensemble models to estimate uncertainty in the predictions through deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2016), guiding the selection of experiments for subsequent rounds based on both predicted effects and associated uncertainties.

3 Methods

In this section, we present the architecture of our model for predicting the effects of combinatorial perturbations, along with our AI + Lab active learning framework (Figure 1A). Our NAIAD framework addresses two key objectives: (1) achieving predictive accuracy with limited training data from the initial experimental round; and (2) implementing a recommendation strategy to select additional gene pairs that maximize information gain, thereby accelerating convergence with fewer AI + experimental iterations. Ultimately, our approach aims to optimize the use of limited experimental resources, reducing the need for exhaustive testing of all possible combinations and efficiently identifying effective gene combinations that drive cells toward desired cellular phenotypes.

3.1 Model design

Let $X_{\text{gene}} \in \mathbb{R}^{k \times p}$ be the learnable gene embedding matrix, where k is the number of genes simultaneously perturbed and p is the dimension of the gene embeddings, which is adapted to the number of training samples. In this work, we focus on the case where $k = 2$. Let Y_i be a scalar value representing the effects of a single-gene (i) perturbation. The value Y_{i+j} denotes the combined effect of the two-gene ($i + j$) perturbation. The target variable Y is a scalar that can represent various biological outcomes, such as cell fitness, marker enrichment levels, or projected gene signatures derived from single-cell transcriptomic data.

Our model is formulated as:

$$Y_{i+j} = \phi([Y_i, Y_j]W_1)A_1^T + f(\phi(W_2 X^i_{gene}), \phi(W_2 X^j_{gene}))A_2^T$$

where ϕ is an activation function (ReLU or GeLU), and W, A are learnable matrix parameters. X^i_{gene} and X^j_{gene} are the row-vectors for gene i and j in X_{gene}. $\phi([Y_i, Y_j]W_1)A_1^T$ models the over-parameterized single-gene effects, and $f(\phi(W_2 X^i_{\text{gene}}), \phi(W_2 X^j_{\text{gene}}))A_2^T$ models the genetic interaction contributions between i and j, where f applies a permutation-invariant function to capture the interactions of genes i and j through their embeddings X_{gene}.

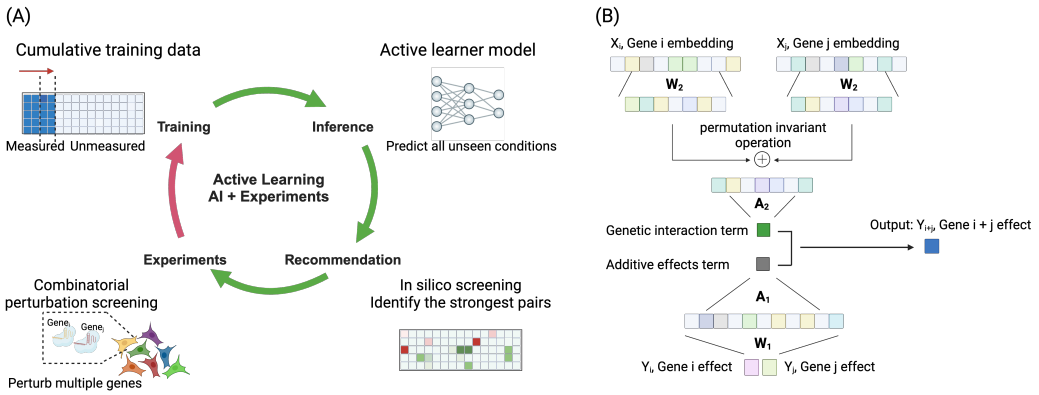

Figure 1: Illustration of active learning framework in CRISPR combinatorial perturbation (A) and our NAIAD model architecture with overparameterized single-gene effects and adaptive gene embedding modules (B).

By explicitly incorporating single-gene effects as input parameters, we condition combinatorial predictions on the known single-gene effects, which substantially enhances the model's performance. The parameter W_1 acts as an over-parameterized encoder, projecting the data from k dimensions to a much higher-dimensional space $\mathbb{R}^m$, where $m \gg k$. The projection of this data into a high-dimensional (though finite) space allows us to capture intricate patterns that are not discernible in the original low-dimensional space.

To model genetic interactions, we sum the gene embeddings of the perturbed gene combinations in the latent space to obtain a single combined embedding. This summation is a permutation-invariant operation, and thus independent of the order of the multiple perturbations. The combined embedding is then passed through an encoder that compresses it into a singular interaction value. This compression reduces the dimensionality of the data while retaining the essential information needed to predict nonlinear interaction of genes. Finally, the compressed representations from both the overparameterized single-gene effect and combined-gene embeddings are used to predict the phenotype. This also allows our model to learn higher-order interactions among gene combination perturbations, leading to improved predictive performance.

We can interpret the model as follows: the overparameterized single-gene perturbation effects tell the model how any pair of genes additively would impact the cellular phenotype, and the gene embeddings provide gene relation information to model the synergistic or buffering interactions between gene pairs.

3.2 Small sample learning as an important initial step for active learning

Unlike most existing models, we propose modeling gene embeddings within a latent space with varying dimensionality tailored to the training-data size. Training large models on small datasets can lead to overfitting. To mitigate this, we initialize the dimension p of the latent space to a small value at the beginning of training. As active learning iteratively incorporates new data, the size of the training dataset continually increases, and we and correspondingly increase p following a predetermined schedule based on the average number of times each gene is seen in the dataset. As the training dataset expands, the increased dimensionality of the latent space enables the model to capture more complex patterns and interactions among genes without overfitting to noisy experimental measurements. By controlling the model complexity based on the available data, this adaptability allows the gene embeddings to effectively leverage both small and large training datasets, optimizing model performance across different data sizes.

3.3 Recommendation system of NAIAD

A critical component of our active learning framework is the recommendation system for selecting gene combinations in subsequent experiment rounds to acquire new data (Figure 1A). We evaluated

an ensemble-based uncertainty estimation, where multiple models with different initializations are aggregated to estimate prediction uncertainty through measures like variance and entropy (Lakshminarayanan, Pritzel, and Blundell 2016). This ensemble approach not only improves predictive performance and enhances interpretability, but also quantifies epistemic uncertainty arising from limited data. We investigated uncertainty calculation by maximizing the likelihood $P(Y|\hat{\mu}, \hat{\sigma})$ under the assumption of a conditional Gaussian distribution (Lahlou et al. 2021) as well. However, we noted that the experimental uncertainty estimated from the ensemble method was more stable. Thus, in this work, we adopted the variance of ensemble predictions as the ensemble-based uncertainty estimator.

In addition to sampling gene pairs with high uncertainty, we incorporate maximum-predicted-effect (MPE) and residual-based sampling strategies to diversify our experimental selection. The MPE sampling focuses on gene pairs with strong predicted effects, identifying combinations that may yield substantial biological insights or therapeutic benefits. Additionally, to balance the exploitation of known MPE areas with the exploration of uncertain regions, we also combined ensemble prediction uncertainty with residual-based sampling. Residual-based sampling targets areas where the model's predictions deviate most strongly from a linear model baseline, allowing exploration of complex interactions that the model has not yet captured and helping to uncover gene interactions that might be missed by linear models. Combining residual-based sampling with uncertainty estimation corresponds to the Upper Confidence Bound (UCB) sampling method used in RECOVER.

4 Experiments

4.1 Datasets

We evaluated our models across four bulk combinatorial CRISPR perturbation screening datasets with cell viability measurements from two cell types (Norman et al. 2019; Simpson et al. 2023; Horlbeck et al. 2018). We treat each gene combination as one sample. Detailed descriptions of these datasets are provided in Appendix A.

4.2 Data splitting

We split the data differently depending on the experiment and dataset used. In our downsampling experiments in Section 5.2, we used [100, 200, 350, 500, 750, 1000, 1250, and 1500] samples during training for the Norman dataset (6,328 combinations), and [100, 500, 1000, 2000, 3000, 4000, 5000, 6000] samples for training on the Simpson (147,658 combinations) and Horlbeck datasets (100,576 combinations for K562; 95,703 combinations for Jurkat T), along with 10% for validation and 30% for testing for each dataset.

To mimic an active learning scenario, in Section 5.3 we started with 100 samples from the Norman dataset in the first round, and incrementally included an additional 100 samples each active learning round, for a total of 4 rounds. For the Simpson and two Horlbeck datasets, we began with 500 samples in the first round, and incrementally included 500 more each round, for a total of 4 rounds. The data for the first round is selected uniformly, and the incremental data added in each subsequent round is selected via an acquisition function (see Appendix C for description of different acquisition functions used). The iterative data selection process allowed us to progressively improve the models by incorporating more data based on the active learning strategy.

5 Results

5.1 Small sample learning and adaptive gene embeddings

One of the primary challenges in active learning is that it typically begins with a small training dataset (Chandra et al. 2020). When the amount of training data is limited, linear models consistently outperform deep learning models. We observed similar patterns in our case particularly when the training data size is below a certain threshold (e.g., 10% in the (Norman et al. 2019) dataset, corresponding to 20 observations per gene during training). As the training data size increases beyond this threshold, the Multi-Layer Perceptron (MLP) begins to surpass the linear model in performance (Figure 2).

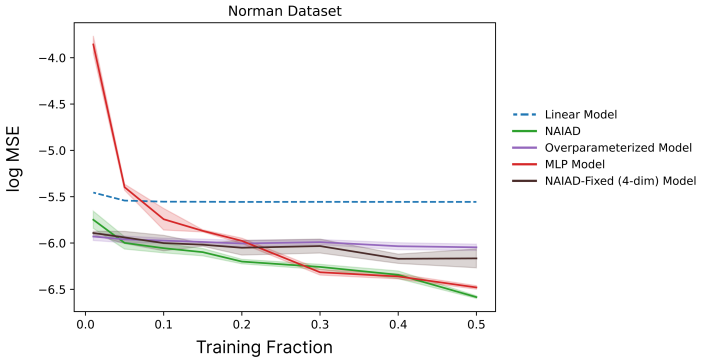

Figure 2: Performance on test data of different gene embedding settings in NAIAD using the Norman dataset (4429 training combinations) across varying training data sizes, reported as $\log$(Mean Square Error). The models without gene embeddings or with low-dimensional embeddings perform well with small training data but do not improve as more data are added. In contrast, the MLP model with larger gene embeddings outperforms these models when the training data exceeds 30%. The adaptive embedding approach achieves the best performance across all training data sizes.

We investigated different gene embedding configurations: an over-parameterized single-gene effect model without embeddings, NAIAD with fixed 4-dimensional gene embeddings, and NAIAD with training-data-size adaptive embeddings ranging from 2 to 128 dimensions. The results suggest that adaptive embeddings consistently achieve superior performance in most cases, regardless of the training data size (Figure 2) in the Norman dataset. This adaptability allows the model to effectively leverage the strengths of both embedding representations and combining single-gene effects. Of note, we observed that the compressed scalar values from the single-gene components strongly correlate with the linear-model predicted results, demonstrating that the NAIAD model leverages the strong baseline performance of linear models. Moreover, the correlation between the scalar values from the gene embedding components and the linear residuals becomes stronger as the training data size increases (Appendix Figure 5).

5.2 Benchmark analysis in small sample learning

We benchmarked NAIAD against the linear model, MLP, GEARS, and RECOVER on the previously-described four bulk combinatorial perturbation datasets that cover 6,328; 147,658; 100,576; and 95,703 gene combinations (detailed information of benchmark models is described in Appendix B). NAIAD consistently outperformed all other models, particularly in situations with a limited number of observed gene pairs, as measured by $\log$(Mean Square Error), Pearson correlation coefficient, and true positive rate from the held-out test data (Figure 3, Appendix Figure 6).

Due to variations in the total number of gene pairs measured across these datasets, we also analyzed model performance based on the average frequency of each gene's occurrence among the training data combinations. Gene occurrence was approximated as Gene Occurrence $= \frac{2N}{M}$, where N is the number of gene combinations in the training set (the factor of 2 assumes a symmetrical screen), and M is the number of unique genes covered in the screen. We observed that when each gene was seen on average four times, NAIAD's performance was consistently the best—over 40% better than the second-best model on average across the four datasets (Table 1) based on root mean square error (RMSE). As the frequency of gene occurrence increased, the performance difference among the models gradually decreased (Table 1, Appendix Table 4). When each gene appeared 20 times within different gene combinations in the training dataset, all models achieved comparable performance levels (Table 1). This suggests that as more data becomes available, the gene embeddings learned from different deep learning models can all largely capture the genetic interactions that dominate performance. Notably, because gene occurrence depends on the number of unique genes included in the screen, achieving a high frequency of observations per gene becomes increasingly challenging when the aim is to cover a wide range of genes. For example, screening 20,000 genes across

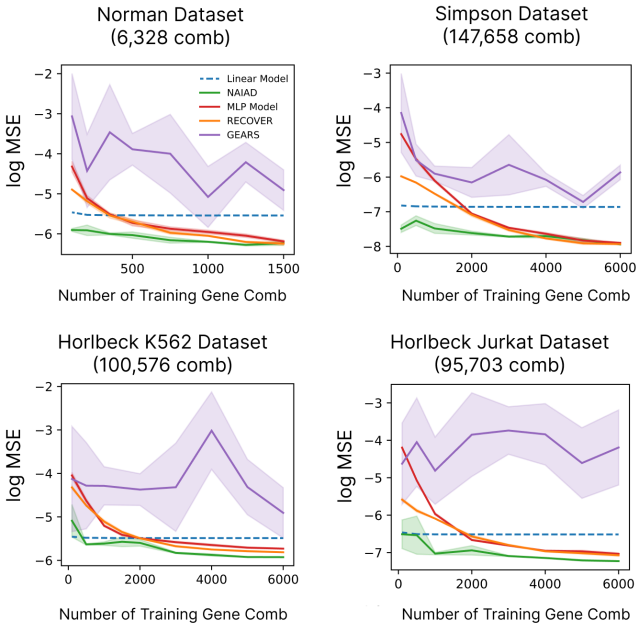

Figure 3: Benchmark analysis comparing the NAIAD model with GEARS and RECOVER models, evaluated using test data $\log(\text{MSE})$ across different numbers of gene combinations in training data. Error bars are SE across three cross-fold replicates.

Table 1: Root mean square error (RMSE) of five models on test data when each gene is approximately observed in training data 4 and 20 times. Error is SE across three cross-fold replicates.

		Dataset RMSE ($\times 10^{-2}$)			
Gene Frequency	**Model**	Norman	Simpson	Horlbeck K562	Horlbeck Jurkat
4	Linear	6.2 (1.7)	3.3 (0.3)	6.4 (0.9)	3.9 (0.6)
	MLP	7.7 (0.9)	4.3 (0.9)	7.8 (2.5)	5.5 (2.1)
	GEARS	16.6 (19.8)	5.4 (3.4)	13.0 (11.6)	13.5 (16.1)
	RECOVER	7.1 (2.8)	3.9 (0.4)	7.9 (2.0)	5.0 (0.7)
	NAIAD	**5.1 (1.8)**	**2.2 (0.1)**	**6.1 (1.9)**	**3.0 (0.6)**
20	Linear	6.1 (1.1)	3.3 (0.2)	6.4 (0.9)	3.8 (0.6)
	MLP	5.0 (1.4)	2.0 (0.3)	5.9 (0.1)	3.0 (0.4)
	GEARS	10.7 (12.1)	3.5 (2.0)	14.0 (14.0)	20.7 (24.0)
	RECOVER	**4.7 (0.5)**	**1.9 (0.4)**	5.6 (1.0)	3.0 (0.4)
	NAIAD	4.7 (0.1)	1.9 (0.2)	**5.4 (0.6)**	**2.8 (0.6)**

the genome to obtain an average of 20 observations per gene would require measuring 200,000 combinations in the initial training dataset. This represents a substantial experimental cost and even exceeds the size of the current largest screen involving over 145,000 combinations (Simpson et al. 2023).

These results demonstrate the robustness of NAIAD in handling both small and large training datasets, making it a valuable tool for exploring vast combinatorial spaces. This highlights its potential for discovery of effective gene combinations in settings with constrained experimental resources.

5.3 Effectiveness of maximum predicted effect sampling in identifying effective gene pairs

The second challenge in active learning frameworks is the design of the recommendation system or acquisition function. We explored multiple acquisition functions—including uncertainty sampling,

Maximum Predicted Effects (MPE) sampling, and Upper Confidence Bound (UCB) sampling from RECOVER (Bertin et al. 2023), which combines residual and uncertainty sampling—to evaluate their impact on overall performance (Figure 4). To simulate the active learning process using current publicly available symmetrical screening data, we started with the same uniformly sampled gene pairs for all acquisition functions and trained a baseline ensemble of NAIAD models. We then used the trained model ensemble to infer unseen combinations across the entire combinatorial space and applied different acquisition functions on the corresponding ensemble metric (e.g. MPE or ensemble uncertainty) to select gene pairs for measurement as the additional training data for the next round. We repeated the sampling and retraining process across four iterations.

We found that MPE sampling outperforms other sampling methods by identifying a higher fraction of the globally strongest gene pairs (Figure 4). The advantage of MPE sampling becomes larger with each additional iteration, even though all approaches show improved performance over iterations. By the fourth iteration, across the four datasets, we were able to uncover over twice as many strong perturbations using MPE sampling compared to uniform sampling, and nearly 1.5 times as many as UCB, the second-best method (Table 2). Specifically, the MPE method identified approximately 150 out of the top 200 strongest gene pairs in three of the datasets, achieving the highest marginal gain in each dataset (Table 3). Although MPE sampling exhibited worse overall performance based on the test data MSE (Appendix Figure 7), this discrepancy is likely due to the distribution of the sampled training data. The MPE sampling method skews the distribution toward strong-effect gene pairs, leading to more accurate predictions for these pairs but less accuracy across the entire dataset, thus affecting the overall MSE (Appendix Figure 8).

The primary goal of NAIAD is not to accurately predict all gene pair interactions but to effectively select the strongest gene pairs that induce significant phenotypic changes. Therefore, the superior performance of MPE sampling in identifying potent gene combinations aligns well with our objectives. By prioritizing the discovery of those strongest gene pairs, we can accelerate the identification of gene combinations that are most relevant for therapeutic development.

6 Discussion and conclusion

In this work, we introduced NAIAD, an active learning framework designed to efficiently identify effective gene pairs from the huge combinatorial perturbation space by utilizing single-gene perturbation effects, adaptive gene embeddings, and an MPE acquisition function. Our framework leverages the principles of Bayesian optimization (Frazier 2018), employing sequential experimentation and learning to enable effective decision-making with limited data. Whereas traditional Bayesian optimization employ a fixed surrogate model, our framework adapts to learn different surrogate functions at varying sizes of training data. This adaptability enables NAIAD to adjust its modeling complexity based on the available data, effectively bridging the gap between linear models and deep learning techniques.

Table 2: Number of top 200 gene combinations correctly discovered at round 4

	Norman	Simpson	Horlbeck K562	Horlbeck Jurkat
Uniform	92	66	39	79
MPE	**144**	**144**	**94**	**161**
UCB	97	111	55	117
Uncertainty	92	74	45	84

Table 3: Marginal gain of correctly discovered top 200 gene combinations at round 4

	Norman	Simpson	Horlbeck K562	Horlbeck Jurkat
Uniform	1	7.5	0.5	2.5
MPE	**14**	**27**	**14.25**	**23**
UCB	2.25	18.75	4.5	12
Uncertainty	1	9.5	2	3.75

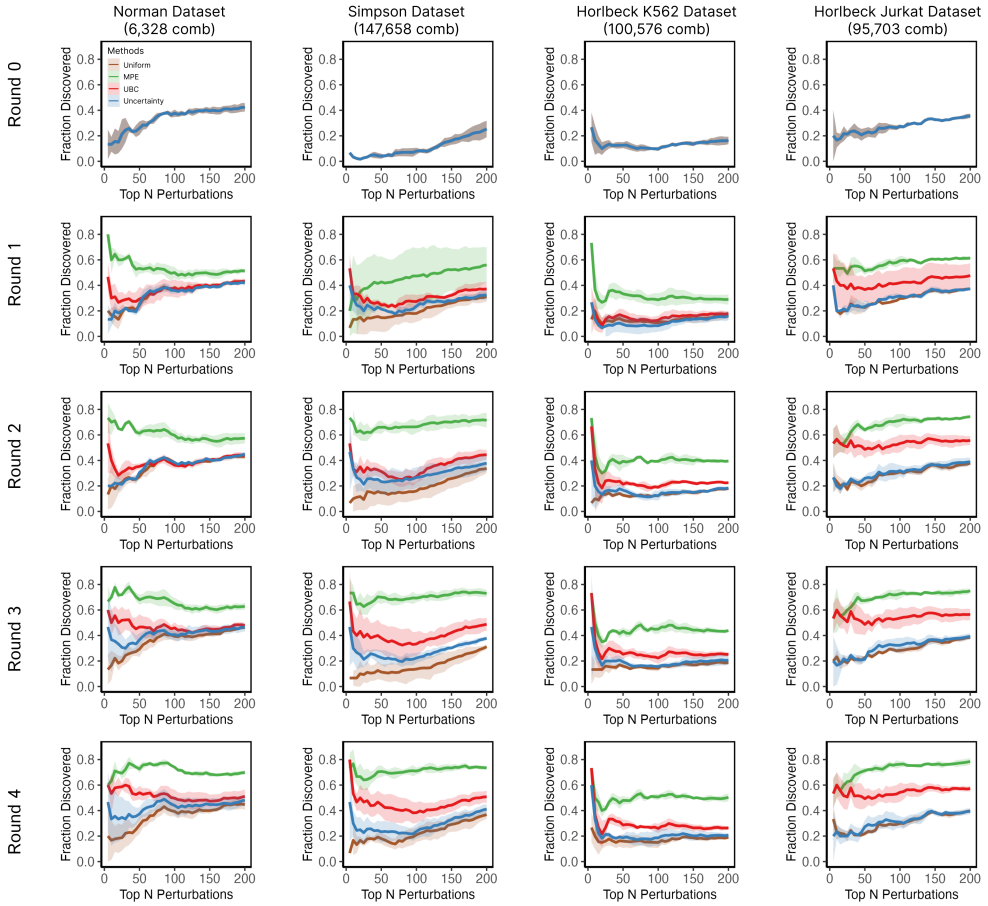

Figure 4: Comparison of different acquisition functions evaluated by top N prediction accuracy for the top N perturbations across four iteration rounds (see Appendix D for full description of accuracy metric). Error bars are SE from three cross-fold replicates.

By incorporating adaptive gene embeddings, NAIAD mitigates overfitting in small-sample training and captures complex interactions as more data becomes available. The MPE acquisition function further enhances the model's efficiency by prioritizing gene pairs with significant predicted effects, accelerating the discovery process with fewer experimental iterations. As a result, NAIAD outperforms existing approaches, particularly in scenarios with limited training data when using the MPE acquisition function. We demonstrate its effectiveness in small-sample learning and its high discovery rate for identifying the top N strongest combinations across the entire search space.

Despite its advantages, there are areas where the NAIAD model could be further enhanced. Currently, NAIAD does not support predicting gene expression profiles resulting from single-cell combinatorial perturbation data. However, when projecting single-cell expression data into one relevant phenotype value, the NAIAD framework can still be adapted to predict that specific phenotype. Another limitation is that our model assumes information of each gene's individual effect on the phenotype. For unseen individual genes, the current NAIAD model cannot predict the effects of combinations involving these genes, especially when both genes in a pair are unseen. Incorporating properly pre-trained gene embeddings could potentially allow us to predict such unseen situations (Cui et al. 2024). By incorporating pre-trained gene embeddings, we can obtain prior knowledge of the similarities between unseen and known genes within relevant latent spaces. Assuming that these gene relationships are conserved across different domains, the model can leverage this information to infer the single-gene effects of unseen genes, even to predict the outcomes of combinations involving

two previously unseen genes. Additionally, refining the acquisition score function to be a learnable component—such as employing a monotonic submodular regularization (Alieva et al. 2020; Wei, Iyer, and Bilmes 2015; Golovin and Krause 2010)—could enable the model to adaptively prioritize experiments that maximize information gain, rather than relying on a fixed heuristic.

Our model also holds significant potential for higher-order combinatorial perturbation data. With the advancement of combinatorial CRISPR technologies, higher-order gene combination datasets are becoming increasingly common (Tieu et al. 2024; Hsiung et al. 2024). NAIAD is theoretically well-suited and can be easily adapted to accommodate interactions beyond gene pairs. Handling higher-order CRISPR combinatorial perturbations would allow for the exploration of more complex genetic interactions, potentially leading to a more effective induction of desired cellular phenotypes.

7 Acknowledgements

This work would not have been possible without the support of the entire Neptune Bio team.

References

[1] Ahlmann-Eltze, Constantin, Wolfgang Huber, and Simon Anders. 2024. "Deep Learning-Based Predictions of Gene Perturbation Effects Do Not yet Outperform Simple Linear Methods." *bioRxiv*. https://doi.org/10.1101/2024.09.16.613342.

[2] Alieva, Ayya, Aiden Aceves, Jialin Song, Stephen Mayo, Yisong Yue, and Yuxin Chen. 2020. "Learning to Make Decisions via Submodular Regularization." In *International Conference on Learning Representations*. https://openreview.net/pdf?id=ac288vnG_7U.

[3] Al-Lazikani, Bissan, Udai Banerji, and Paul Workman. 2012. "Combinatorial Drug Therapy for Cancer in the Post-Genomic Era." *Nature Biotechnology*, 30 (7): 679–92.

[4] Bereket, Michael, and Theofanis Karaletsos. 2023. "Modelling Cellular Perturbations with the Sparse Additive Mechanism Shift Variational Autoencoder." http://arxiv.org/abs/2311.02794.

[5] Bertin, P., J. Rector-Brooks, D. Sharma, T. Gaudelet, A. Anighoro, T. Gross, F. Martínez-Peña, et al. 2023. "RECOVER Identifies Synergistic Drug Combinations in Vitro through Sequential Model Optimization." *Cell Reports Methods*, 3 (10). https://doi.org/10.1016/j.crmeth.2023.100599.

[6] Chandra, Akshay L., Sai Vikas Desai, Chaitanya Devaguptapu, and Vineeth N. Balasubramanian. 2020. "On Initial Pools for Deep Active Learning." http://arxiv.org/abs/2011.14696.

[7] Cui, Haotian, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. "scGPT: Toward Building a Foundation Model for Single-Cell Multi-Omics Using Generative AI." *Nature Methods*, 21 (8): 1470–80.

[8] Eisenstein, Michael. 2020. "Active Machine Learning Helps Drug Hunters Tackle Biology." *Nature Biotechnology*, 38 (5): 512–14.

[9] Frazier, Peter I. 2018. "A Tutorial on Bayesian Optimization." http://arxiv.org/abs/1807.02811.

[10] Gaudelet, Thomas, Alice Del Vecchio, Eli M. Carrami, Juliana Cudini, Chantriolnt-Andreas Kapourani, Caroline Uhler, and Lindsay Edwards. 2024. "Season Combinatorial Intervention Predictions with Salt & Peper." http://arxiv.org/abs/2404.16907.

[11] Geiger-Schuller, Kathryn, Basak Eraslan, Olena Kuksenko, Kushal K. Dey, Karthik A. Jagadeesh, Pratiksha I. Thakore, Ozge Karayel, et al. 2023. "Systematically Characterizing the Roles of E3-Ligase Family Members in Inflammatory Responses with Massively Parallel Perturb-Seq." *bioRxiv*. https://doi.org/10.1101/2023.01.23.525198.

[12] Golovin, Daniel, and Andreas Krause. 2010. "Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization." http://arxiv.org/abs/1003.3967.

[13] Hammond, Jennifer, Heidi Leister-Tebbe, Annie Gardner, Paula Abreu, Weihang Bao, Wayne Wisemandle, Marylynn Baniecki, et al. 2022. "Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19." *The New England Journal of Medicine*, 386 (15): 1397–1408.

[14] Hao, Minsheng, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. "Large-Scale Foundation Model on Single-Cell Transcriptomics." *Nature Methods*, 21 (8): 1481–91.

[15] Horlbeck, Max A., Albert Xu, Min Wang, Neal K. Bennett, Chong Y. Park, Derek Bogdanoff, Britt Adamson, et al. 2018. "Mapping the Genetic Landscape of Human Cells." *Cell*, 174 (4): 953–67.e22.

[16] Hsiung, C. C-S, C. M. Wilson, N. A. Sambold, R. Dai, Q. Chen, N. Teyssier, S. Misiukiewicz, et al. 2024. "Engineered CRISPR-Cas12a for Higher-Order Combinatorial Chromatin Perturbations." *Nature Biotechnology*, May, 1–15.

[17] Hwangbo, Haeun, Sarah C. Patterson, Andy Dai, Deborah Plana, and Adam C. Palmer. 2023. "Additivity Predicts the Efficacy of Most Approved Combination Therapies for Advanced Cancer." *Nature Cancer*, 4 (12): 1693–1704.

[18] Jastreboff, Ania M., Lee M. Kaplan, Juan P. Frías, Qiwei Wu, Yu Du, Sirel Gurbuz, Tamer Coskun, Axel Haupt, Zvonko Milicevic, and Mark L. Hartman. 2023. "Triple–Hormone-Receptor Agonist Retatrutide for Obesity — A Phase 2 Trial." *The New England Journal of Medicine*, August. `https://doi.org/10.1056/NEJMoa2301972`.

[19] Kingma, Diederik P., and Max Welling. 2013. "Auto-Encoding Variational Bayes." `http://arxiv.org/abs/1312.6114`.

[20] Kipf, Thomas N., and Max Welling. 2016. "Semi-Supervised Classification with Graph Convolutional Networks." `http://arxiv.org/abs/1609.02907`.

[21] Lahlou, Salem, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2021. "DEUP: Direct Epistemic Uncertainty Prediction." `http://arxiv.org/abs/2102.08501`.

[22] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2016. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." `http://arxiv.org/abs/1612.01474`.

[23] Lopez, Romain, Nataša Tagasovska, Stephen Ra, Kyunghyn Cho, Jonathan K. Pritchard, and Aviv Regev. 2022. "Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling." `http://arxiv.org/abs/2211.03553`.

[24] Lotfollahi, Mohammad, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, et al. 2023. "Predicting Cellular Responses to Complex Perturbations in High-throughput Screens." *Molecular Systems Biology*, May. `https://doi.org/10.15252/msb.202211517`.

[25] Mokhtari, Reza Bayat, Tina S. Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. 2017. "Combination Therapy in Combating Cancer." *Oncotarget*, 8 (23): 38022.

[26] Norman, Thomas M., Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. 2019. "Exploring Genetic Interaction Manifolds Constructed from Rich Single-Cell Phenotypes." *Science*, 365 (6455): 786–93.

[27] "Rationalizing Combination Therapies." 2017. *Nature Medicine*, 23 (10): 1113–1113.

[28] Rood, Jennifer E., Anna Hupalowska, and Aviv Regev. 2024. "Toward a Foundation Model of Causal Cell and Tissue Biology with a Perturbation Cell and Tissue Atlas." *Cell*, 187 (17): 4520–45.

[29] Roohani, Yusuf, Kexin Huang, and Jure Leskovec. 2023. "Predicting Transcriptional Outcomes of Novel Multigene Perturbations with GEARS." *Nature Biotechnology*, 42 (6): 927–35.

[30] Samms, Ricardo J., Matthew P. Coghlan, and Kyle W. Sloop. 2020. "How May GIP Enhance the Therapeutic Efficacy of GLP-1?" *Trends in Endocrinology and Metabolism: TEM*, 31 (6): 410–21.

[31] Shyr, Zeenat A., Yu-Shan Cheng, Donald C. Lo, and Wei Zheng. 2021. "Drug Combination Therapy for Emerging Viral Diseases." *Drug Discovery Today*, 26 (10): 2367.

[32] Simpson, Danny, Jia Ling, Yangwode Jing, and Britt Adamson. 2023. "Mapping the Genetic Interaction Network of PARP Inhibitor Response." *bioRxiv*, August. `https://doi.org/10.1101/2023.08.19.553986`.

[33] Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell*, 126 (4): 663–76.

[34] Tieu, Victor, Elena Sotillo, Jeremy R. Bjelajac, Crystal Chen, Meena Malipatlolla, Justin A. Guerrero, Peng Xu, et al. 2024. "A Versatile CRISPR-Cas13d Platform for Multiplexed Transcriptomic Regulation and Metabolic Engineering in Primary Human T Cells." *Cell*, 187 (5): 1278–95.e20.

[35] Wei, Kai, Rishabh Iyer, and Jeff Bilmes. 2015. "Submodularity in Data Subset Selection and Active Learning." In *International Conference on Machine Learning*, 1954–63. PMLR.

A Dataset summary

We utilized four bulk combinatorial perturbation datasets in our study:

Combinatorial CRISPRa on K562 cells (Norman et al. 2019): This dataset was generated using combinatorial CRISPR activation (CRISPRa), involving 112 genes and 6,328 unique gene combinations.

Large-scale combinatorial CRISPRi on K562 cells (Simpson et al. 2023): This dataset was generated using CRISPR interference (CRISPRi), involving 543 genes and 147,658 unique gene combinations.

Combinatorial CRISPRi on K562 cells (Horlbeck et al. 2018): This dataset was generated using combinatorial CRISPRi, involving 448 genes and 100,576 unique gene combinations.

Combinatorial CRISPRi on Jurkat T Cells (Horlbeck et al. 2018): This dataset was generated using combinatorial CRISPRi, involving 437 genes and 95,703 unique gene combinations.

For the bulk cell viability datasets, we calculated the log-fold change in cell viability for each gene combination compared to negative control treatments. This normalization allows for the quantification of the effect size of each perturbation on cell survival. For a deeper explanation of calculating cell viability (also referred to as γ), we refer the reader to (Simpson et al. 2023).

B Benchmark models

Linear model: We employed a simple linear regression model using the effects of two single-gene perturbations as independent variables to predict their joint effect. This model involved three parameters: two coefficients for the individual gene effects and an intercept term.

Multi-layer perceptron (MLP): We implemented an MLP model where each gene was represented by a 128-dimensional embedding. For each two-gene perturbation, we created a joint embedding for the two genes by summing their embeddings along each dimension and passing it through a MLP layer that projects from 128 dimensions down to a single dimension corresponding to the phenotype value. The MLP captured non-linear interactions between gene embeddings.

GEARS: For the GEARS model, we adhered to the original settings as specified in their supplementary materials (Roohani, Huang, and Leskovec 2023) for cell viability prediction. This included their specific network architecture and parameters used in their tutorial. For bulk-combinatorial datasets lacking single-cell experiments expected for GEARS, we generated synthetic Pertub-seq datasets of normalized gene expression matrices, with a separate Gaussian $\mathcal{N}(0, 1)$ used for sampling the expression of each gene.

RECOVER adaptation: RECOVER was originally developed for drug perturbations involving small molecule embeddings and bilinear operations to combine drugs (Bertin et al. 2023). The models were trained to predict Bliss synergy scores, which capture the non-linear components of phenotypic outcomes, instead of directly modeling the phenotypic outcomes themselves. We adapted RECOVER for gene perturbations by incorporating a bilinear projection module to combine 128-dimensional gene embeddings. This adaptation allowed us to model gene-gene interactions using the same principles applied to drug combinations in the original RECOVER framework. We train the model to predict the gene-equivalent of Bliss scores, which is the difference between measured viability and the product of the viability of single-gene perturbations. We then convert the gene-Bliss-score predictions back to overall cell viability predictions by re-adding the product of the viability of the single-gene perturbations.

C Active learning sampling strategies

In our active learning framework, we employ several sampling strategies to select the most informative experiments for subsequent rounds. Below, we provide the mathematical formulations for these strategies.

Uncertainty-based sampling

We use an ensemble of models $\{M_i\}_{i=1}^{N}$, each initialized randomly. For a candidate gene combination x, each model provides a prediction $y_i = M_i(x)$. We estimate the prediction uncertainty using the variance of the ensemble predictions:

$$\text{Var}(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

where $\bar{y}$ is the mean prediction across the ensemble:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

The acquisition function for uncertainty-based sampling is defined as:

$$a_{\text{uncertainty}}(x) = \text{Var}(y).$$

We select the candidate combinations with the highest $a_{\text{uncertainty}}(x)$ values, as they represent samples where the model is most uncertain and additional data could significantly improve the model.

Maximum predicted effect (MPE) sampling

This strategy targets gene combinations predicted to have strong effects. The acquisition function is:

$$a_{\text{MPE}}(x) = \hat{y},$$

where $\hat{y}$ is the mean prediction as defined above. We select candidates with the highest absolute predicted effects $\hat{y}$, prioritizing experiments likely to yield substantial biological insights or therapeutic benefits.

Residual-based sampling

To identify areas where nonlinear interactions are significant and the model may not have fully captured them, we compute the residual between predictions from a non-linear model and a linear model. Let $N(x)$ be the prediction from the non-linear model and $L(x)$ be the prediction from a linear approximation. The residual is calculated as:

$$r(x) = |N(x) - L(x)|.$$

The acquisition function for residual-based sampling is:

$$a_{\text{residual}}(x) = r(x).$$

We select candidate combinations with the highest residuals $a_{\text{residual}}(x)$, focusing on samples where the non-linear effects are most pronounced and the model's predictions differ significantly from linear expectations.

Upper Confidence Bound (UCB) sampling

We perform UCB sampling following the strategy provided by RECOVER (Bertin et al. 2023):

$$a_{\text{UCB}}(x) = a_{\text{residual}}(x) + \kappa * a_{\text{uncertainty}}(x)$$

Where κ is a hyperparameter set to $\kappa = 1$ based on the recommendation from the RECOVER paper.

By combining these sampling strategies, we aim to efficiently explore the combinatorial space of gene interactions. The goal is to maximize information gain with each experimental round, ultimately using a minimal number of experiments to identify effective gene combinations that can induce desired cellular phenotypes.

D Model evaluation metrics

We use several metrics to evaluate the performance of our models.

MSE: Mean Square Error

$$\frac{1}{N_{\text{pairs}}}\sum_{i,j\in\text{genes}}(y_{i+j}^{\text{pred}} - y_{i+j}^{\text{truth}})^2$$

Fraction Discovered: Through active learning, we want to identify how many of the top P strong perturbations have been identified by our model of interest. Let M be the total number of rounds of active learning, and let $\mathbf{N} = [n_0, n_1, \ldots, n_M]$ be the list that defines how many samples n_i are used for each active learning round $i = 0, 1, \ldots, M$.

In each round of learning, we train on the n_i unmasked samples chosen depending on the acquisition function. After training, we make predictions for the entire unseen dataset. We concatenate the predictions on the unseen data with the unmasked measured values from the seen data, and call this combined set of predictions and measurements X_i for each round i.

Let $X_{i,P}$ be the set of top P values in the set X_i. Let $Y_{i,P}$ be the set of top P measured (ground truth) values across all samples in round i. Let $P_i = |X_{i,P} \cap Y_{i,P}|$ be the number of times a perturbation from the top P predictions is also in the top P targets for round i. Fraction Discovered is then defined as $\frac{P_i}{P}$.

Marginal Gain: The marginal gain at round 4 is $\frac{P_4(top=200)-P_0(top=200)}{4}$, where $P_4(top = 200)$ is the number of top-200 perturbations correctly identified in Round 4, and $P_0(top = 200)$ is the number of top-200 perturbations correctly identified in Round 0.

TPR: True positive rate

Let X_{targets} be the set of top N measured combinations, and X_{preds} be the set of top N predicted combinations. We find the number of matches $N_{\text{match}} = |X_{\text{targets}} \cap X_{\text{preds}}|$, and calculate TPR as $\frac{N_{\text{match}}}{N}$.

E Hardware Configuration

All model training was done on a single Paperspace A100-80G server with 100GB of RAM.

F Hyperparameter Selection

For all training of the NAIAD and MLP models, we use learning_rate $= 10^{-2}$, and a batch_size $= 1024$. We also used a linear rate scheduler with 10% of training steps used for warm up, and weight_decay $= 0$.

To identify these optimal hyperparameters, we testing hyperparameters across the following ranges:

n_epoch: $[50, 100, 200, 500, 1000, 2000]$

batch_size: $[512, 1024, 2048, 4096]$

learning_rate: $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$

d_embed: $[2, 4, 8, 16, 32, 64, 128, 256]$

d_single-gene: $[8, 16, 32, 64, 128, 256, 512]$

weight_decay: $[0, 10^{-4}, 10^{-3}]$

For training the RECOVER model, we use a learning_rate $= 10^{-1}$. For GEARS, we keep all hyperparameters assigned by default from the package.

F.1 Adaptive Model Embedding Size Selection

For the NAIAD model, we choose the embedding size hyperparameter based roughly on the number of times each gene was seen in the training set, following the schedule shown in Table 5 . We determined these values using the idea that model size depends on the number of times each individual gene is seen during training, and that these properties hold across all datasets.

G Supplemental figures and tables

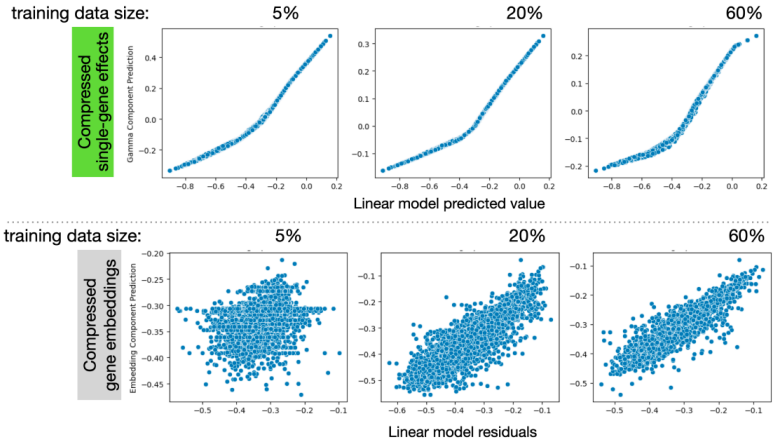

Figure 5: Evaluation of compressed single-gene effects and gene embeddings shows a strong correlation between the compressed single-gene effects and the values predicted by the linear model. As the training data increases, the correlation between gene embeddings and the residuals of the linear model predictions gradually becomes stronger.

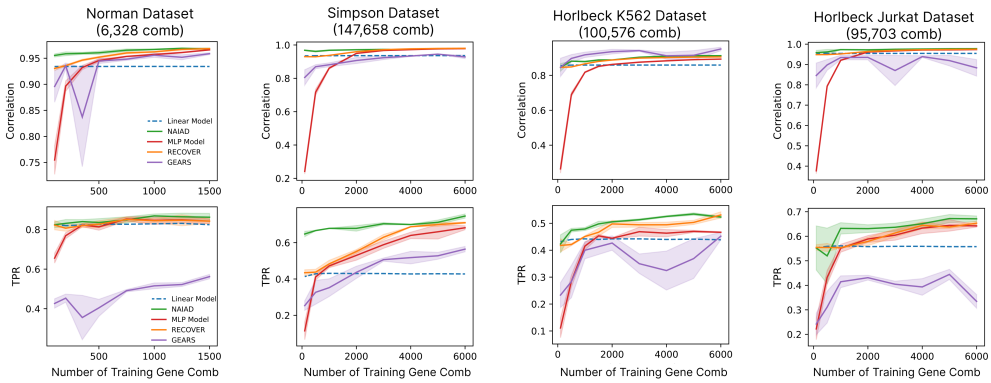

Figure 6: Benchmark analysis comparing the NAIAD model with GEARS and RECOVER models, evaluated using test data correlation and true positive rate (for identifying the top 200 perturbations of a 30% held-out test set) across different numbers of gene combinations in training data. Error bars are SE across three cross-fold replicates.

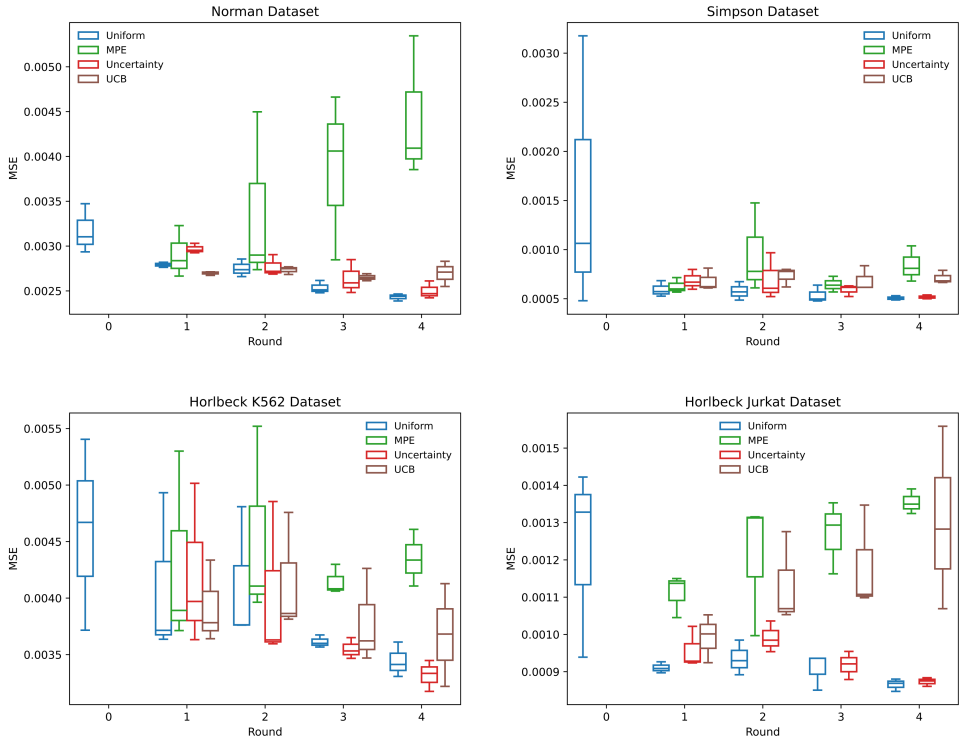

Figure 7: Comparison of different acquisition functions evaluated by MSE across four iteration rounds. Error bars are SE across three cross-fold replicates.

Figure 8: Comparison of training sample distribution changes across different iterations using four acquisition functions. The X-axis represents the linear predicted values, while the Y-axis shows the actual measured values. The plot highlights gene pairs belonging to the top 200 gene pairs with the strongest effects in the entire dataset.

Table 4: RMSE of five models on test data when each gene is approximately observed in training data 10 times. Error is SE across three cross-fold replicates.

		Dataset RMSE ($\times 10^{-2}$)			
Gene Frequency	**Model**	Norman	Simpson	Horlbeck K562	Horlbeck Jurkat
10	Linear	6.1 (1.3)	3.3 (0.2)	6.4 (0.9)	3.8 (0.6)
	MLP	5.7 (1.3)	2.4 (0.4)	6.3 (1.1)	3.5 (1.0)
	GEARS	15.5 (12.8)	8.7 (10.3)	12.0 (9.6)	24.2 (28.5)
	RECOVER	5.9 (1.0)	2.5 (0.2)	6.1 (0.2)	3.7 (1.0)
	NAIAD	**4.7 (1.2)**	**2.0 (0.2)**	**5.7 (0.8)**	**3.0 (0.8)**

Table 5: Training and model hyperparameters for each data set and data split

Dataset	Train Epochs	Avg. Times All Genes Seen	Approx. Training Dataset Size	Embed Dim	Single-Gene Dim
Norman	500	2	100	2	64
		4	200	4	64
		10	500	16	64
		20	1000	16	64
		30	1500	32	64
		40	2000	64	64
		60	3000	64	64
		80	4000	128	64
		100+	5000+	128	64
Simpson	200	2	500	2	256
		4	1000	4	256
		10	5000	16	256
		20	10000	16	256
		30	15000	32	256
		40	20000	64	256
		60	30000	64	256
		80	40000	128	256
		100+	50000+	128	256
Horlbeck	200	2	500	4	256
		4	1000	16	256
		10	5000	32	256
		20	10000	32	256
		30	15000	64	256
		40	20000	64	256
		60	30000	64	256
		80	40000	128	256
		100+	50000+	128	256