BioBO: Biology-informed Bayesian Optimization for Perturbation Design

Yanke Li[*†]
Health Science and Technology
ETH Zurich
yanke.li@hest.ethz.ch

Tianyu Cui[†]
Innovative Medicine
Johnson&Johnson
tcui8@its.jnj.com

Tommaso Mansi
Innovative Medicine
Johnson&Johnson
tmansi@its.jnj.com

Mangal Prakash[†]
Innovative Medicine
Johnson&Johnson
mpraka12@its.jnj.com

Rui Liao[†]
Innovative Medicine
Johnson&Johnson
rliao2@its.jnj.com

Abstract

Efficient design of genomic perturbation experiments is crucial for accelerating drug discovery and therapeutic target identification, yet exhaustive perturbation of the human genome remains infeasible. Bayesian optimization (BO) has recently emerged as a powerful framework for selecting informative interventions, but existing approaches often fail to exploit domain-specific biological prior knowledge. We propose Biology-Informed Bayesian Optimization (BioBO), a method that integrates Bayesian optimization with multimodal gene embeddings and enrichment analysis, a widely used tool for gene prioritization in biology, to enhance surrogate modeling and acquisition strategies. BioBO leverages biologically grounded priors within the πBO framework to balance exploration and exploitation. Through experiments on the established public benchmark, we demonstrate that BioBO improves labeling efficiency, consistently outperforms conventional BO, and identifies top-performing perturbations more effectively. These results highlight the potential of incorporating structured biological knowledge into BO frameworks for more efficient and interpretable genomic experimental design.

1 Introduction

In vitro cellular experimentation with genomic interventions is a critical step in early-stage drug discovery and target prioritization. By perturbing genes and observing cellular responses, researchers can infer gene function and identify potential therapeutic targets [1, 2]. Techniques such as CRISPR-Cas9 [3] knockout screens enable systematic perturbation of individual genes, but they are often resource-intensive and time-consuming. Given the vast number of protein-coding genes in the human genome (approximately 20,000), exhaustively testing all possible perturbations is infeasible [4]. Consequently, strategies that efficiently select the most informative experiments are essential to accelerate drug discovery while minimizing experimental costs.

Bayesian experimental design provides a principled framework for this challenge. In particular, Bayesian optimization (BO) offers a sample-efficient approach to identify genes whose perturbation maximizes desired cellular phenotypes. BO relies on a probabilistic surrogate model, such as a

[*]Summer internship work at Innovative Medicine Johnson&Johnson
[†]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Structured Probabilistic Inference & Generative Modeling (SPIGM).

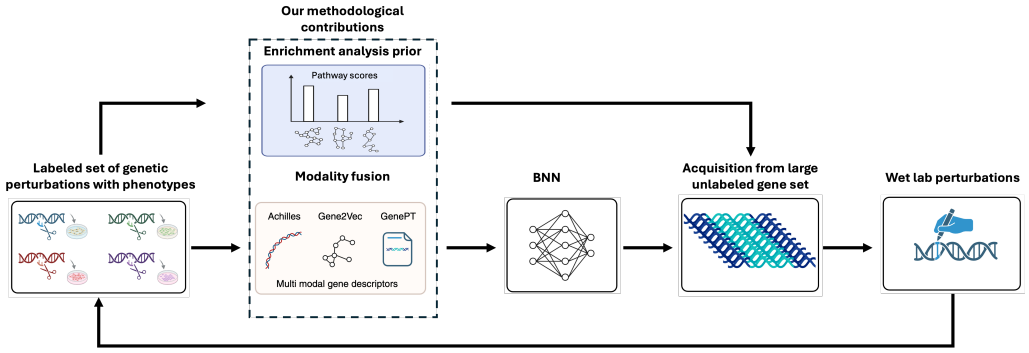

Figure 1: BioBO Pipeline. Our contributions are two-fold: (i). Fusion of gene modalities to improve surrogate modeling; (ii). Enrichment analysis on top of surrogate model predictions to strengthen gene acquisition via incorporating biological information.

Gaussian process [5] or Bayesian neural network [6], to model the response surface, and an acquisition function to balance exploration of uncertain regions with exploitation of promising candidates [7]. While recent works have applied BO to gene perturbation design [8, 9], they typically use generic, uni-modal gene embeddings and do not fully leverage rich biological knowledge, limiting their performance. Integrating multimodal gene embeddings, which capture sequence, functional, and network-based information, can provide more informative representations and improve the efficiency of experimental selection.

Beyond learned embeddings, explicit biological priors can further guide experiment design. For example, gene set enrichment analysis (EA) identifies pathways that are statistically overrepresented among the top-performing genes, providing information on molecular mechanisms and potential high-value targets [10]. However, conventional EA has two key limitations: (i) it lacks granularity, treating all genes within a pathway as equally promising, and (ii) it is purely exploitative, potentially biasing experiments toward well-characterized pathways while neglecting unexplored regions of the genome.

To address these limitations, we propose *Biology-Informed Bayesian Optimization* (BioBO), a framework that integrates multimodal gene embeddings with Bayesian optimization and biological priors, such as enrichment analysis. BioBO helps balancing exploration and exploitation, efficiently guiding experiments toward both well-characterized and underexplored genes. Our contributions are threefold:

1. We introduce multimodal gene embeddings, integrating multiple sources of biological information in the surrogate modeling to improve the designs of BO.

2. We demonstrate that the improvement of BO from multimodal embeddings is mainly from the improvement of the surrogate model on regimes close to optimum rather than on the entire data distribution.

3. We augment the acquisition function in BO using enrichment analysis within the theoretically principled π-BO [11] framework. This approach incorporates prior biological knowledge while maintaining principled exploration–exploitation trade-off and provides interpretable insights into experimental design.

4. We empirically validate BioBO on established public benchmarks, showing that it outperforms conventional BO improves labeling efficiency by 25–40%, and identifies biologically coherent pathways with markedly stronger enrichment signals.

By combining surrogate modeling with biologically informed priors, BioBO enables more efficient, interpretable, and effective experimental designs, ultimately facilitating faster and more targeted discovery in genomic perturbation studies.

2 Background and Notation

2.1 Notation and problem setup

We consider the task of optimizing a black-box function $f : \mathbb{G} \to \mathbb{R}$, which maps each gene $g \in \mathbb{G}$ represented by the set of integers or one-hot embeddings to a value $f(g) \in \mathbb{R}$ denoting the change of cell phenotype under the gene knockout, across the entire finite gene space $\mathbb{G}$ with $|\mathbb{G}| \approx 20,000$ (i.e., the number of protein-coding genes in human). Similar to [9], we use biologically informed d-dimensional embeddings of genes, $\boldsymbol{X} : \mathbb{G} \to \mathbb{X}$, which maps each gene $g \in \mathbb{G}$ to a corresponding d-dimensional vector $\boldsymbol{X}(g) = \boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^d$ capturing the biological relationships with other genes. Moreover, the gene embeddings $\boldsymbol{X}$ construct a one-to-one mapping from $\mathbb{G}$ and contain the same number of distinct d-dimensional vectors as $\mathbb{G}$, i.e., $|\mathbb{X}| = |\mathbb{G}|$, so we use $f(\boldsymbol{x})$ and $f(g)$ interchangeably where $\boldsymbol{x}$ is the embedding of the gene g. Therefore, we define the optimization problem as follows

$$\boldsymbol{x}^* \in \arg\max_{\boldsymbol{x} \in \mathbb{X}} f(\boldsymbol{x}). \tag{1}$$

In practice, $f(\boldsymbol{x})$ is expensive to evaluate because it requires a CRISPR-Cas9 knockout experiment in the lab, and we would like to maximize $f(\boldsymbol{x})$ in an efficient manner by only evaluating a small number of points from $\mathbb{X}$.

2.2 Bayesian Optimization

Bayesian optimization (BO) [12, 7] is a model-based black-box function optimizer that employs a probabilistic model, e.g., Gaussian process (GP) [5] or Bayesian neural network (BNN) [6], as a surrogate model. Specifically, BO optimizes f from an initial experimental design $\mathcal{D}_0 = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{M}$ and sequentially deciding on one or a batch (with size B) of new designs to label and form the data $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \mathcal{B}_n$ with new labeled dataset $\mathcal{B}_n = \{(\boldsymbol{x}_{n,b}, y_{n,b})\}_{b=1}^{B}$ for the n-th iteration with $n \in \{1, \ldots, N\}$. At each iteration n, BO learns a probabilistic surrogate model $f_n \sim p(f_n|\mathcal{D}_n)$ to approximate the true function f, where $p(f_n|\mathcal{D}_n)$ is the posterior distribution of a GP or BNN given the labeled data. Using the predictive uncertainty from $p(f_n|\mathcal{D}_n)$, BO selects next designs by maximizing an acquisition function (AF), $\alpha_{p(f_n|\mathcal{D}_n)}(\boldsymbol{x})$, across the set of unlabeled data points.

Acquisition functions encapsulate the underlying utilities; therefore, they correspond to the trade-off between exploitation (using the current optimum from the surrogate model) and exploration (considering the uncertainty of the surrogate model). Popular choices of AF include Expected Improvement (EI) [13] and Upper Confidence Bound (UCB) [14]. For instance, EI selects the next point $\boldsymbol{x}$ that maximizes the expected improvement:

$$\alpha^{\text{EI}}_{p(f_n|\mathcal{D}_n)}(\boldsymbol{x}) = \mathbb{E}[|f_n(\boldsymbol{x}) - y_n^*|^+] = Z\sigma_n(\boldsymbol{x})\Phi(Z) + \sigma_n(\boldsymbol{x})\phi(Z), \tag{2}$$

where y_n^* is the best outcome observed so far, $Z = \frac{f_n(\boldsymbol{x}) - \mu_n(\boldsymbol{x})}{\sigma_n(\boldsymbol{x})}$ with $\mu_n(\boldsymbol{x})$ and $\sigma_n(\boldsymbol{x})$ representing the mean and variance of the posterior $p(f_n|\mathcal{D}_n)$ respectively, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and CDF of standard Gaussian distribution. UCB is defined as:

$$\alpha^{\text{UCB}}_{p(f_n|\mathcal{D}_n)}(\boldsymbol{x}) = \mu_n(\boldsymbol{x}) + \beta_n\sigma_n(\boldsymbol{x}), \tag{3}$$

where β_n is the user-specified parameter controlling the exploration–exploitation trade-off. Both EI and UCB provide a myopic strategy for determining informative designs with theoretical guarantees [15, 14]. Other popular myopic acquisition functions include Probability of Improvement (PI) [16] and Thompson Sampling (TS) [17].

In this work, we mainly focus on using BNNs as surrogate models and UCB, EI, and TS as acquisition functions, similar to existing works on perturbation design [8, 9]; however, our work applies to other probabilistic models and myopic acquisition functions as well.

2.3 Enrichment Analysis

Enrichment analysis (EA) or over-representation analysis is a computational approach used to determine whether a set of genes associated with a specific biological process or pathway appears

more often than expected by chance [18, 19, 20]. Specifically, given a background gene set, e.g., all protein-coding human genes $\mathbb{G}$, and a subset $\mathbb{S} \subset \mathbb{G}$ of genes of interest, EA tests whether a pathway i, i.e., a predefined gene set $\mathbb{P}_i \subset \mathbb{G}$, provided by pathway databases, such as hallmark [21], is represented in $\mathbb{S}$ *statistically more frequently* than expected by chance. EA comes with statistical hypothesis tests: under the null hypothesis $\mathcal{H}_0$, that genes in $\mathbb{S}$ are sampled uniformly from $\mathbb{G}$, the probability of observing at least $|\mathbb{S} \cap \mathbb{P}_i|$ overlaps follows the upper tail of the hypergeometric distribution; therefore, we can compute the p-value with

$$p(\mathbb{P}_i) = \sum_{i=|\mathbb{S}\cap\mathbb{P}_i|}^{\min(|\mathbb{P}_i|,|\mathbb{S}|)} \frac{\binom{|\mathbb{P}_i|}{i}\binom{|\mathbb{G}|-|\mathbb{P}_i|}{|\mathbb{S}|-i}}{\binom{|\mathbb{G}|}{|\mathbb{S}|}}, \tag{4}$$

and multiple hypothesis testing across all pathways is controlled via Bonferroni correction [22] or Benjamini–Hochberg FDR [23] to derive the q-value. One can also compute the odds ratio, $o(\mathbb{P}_i)$, from the EA results by constructing the contingency table, and a high $o(\mathbb{P}_i)$ (e.g., $o(\mathbb{P}_i) > 1$) indicates that $\mathbb{P}_i$ is over-represented in $\mathbb{S}$ compared to random. [24] propose to combine the p-value and odds ratio to evaluate the overall representativeness with:

$$c(\mathbb{P}_i) = -o(\mathbb{P}_i) \log p(\mathbb{P}_i). \tag{5}$$

EA has been widely used to design experiments in applications such as target prioritization and biomarker expansion [25, 26, 27, 28, 29]. Intuitively, if several desirable genes have been identified, EA can be applied to discover the pathways enriched by those desirable genes. Therefore, other untested genes in those significantly enriched pathways would construct a good candidate set for the next round of experiments. The significantly enriched pathways serve as a biologically informed prioritization framework for designing experiments, allowing us to target molecular processes where the desirable genes are most likely to be. This approach ensures that experimental interventions are focused on high-value genes within the biological network, thereby increasing the likelihood of eliciting interpretable system-level responses while reducing experimental redundancy.

Although EA is a well-established, biologically informed experimental design framework, it contains two major shortcomings:

1. Lack of granularity: EA can prioritize pathways; however, all untested genes in the same pathway are equally likely. This can still construct a huge pool if the significantly enriched pathway is large.

2. Lack of exploration: EA-based experimental design is a pure exploitation process and has potential bias toward known biology. The significantly enriched pathway would be more significant by selecting more genes from it, and non-significant pathways will never be explored.

In this work, we propose a principled approach to combine the BO-based and EA-based experimental design framework to equip BO with extensive domain information in biology from EA and equip EA with granularity and exploration from BO.

3 Method: Biology-Informed Bayesian Optimization

3.1 Surrogate Modelling with Multimodal Gene Representations

We first improve the BO experimental design by improving the surrogate modeling. Specifically, we propose to use multi-modal gene embeddings rather than the uni-modal embeddings used in the existing gene perturbation design literature [8, 9]. We consider the following two extra gene embeddings that are effective in many gene-level tasks [30, 31]:

1. Gene2vec [32], $\boldsymbol{x}^{\text{g2v}}$: gene embeddings encode gene-gene relations defined in gene ontology [33] learned with self-supervised learning;

2. GenePT [31], $\boldsymbol{x}^{\text{GenePT}}$: ChatGPT embeddings of genes based on the literature.

We use Bayesian neural networks (BNN) as surrogate models, and we concatenate the original gene embedding $\boldsymbol{x}$ with the gene embeddings from the above-mentioned modalities as the input of a BNN, i.e., $f([\boldsymbol{x}, \boldsymbol{x}^{\text{g2v}}, \boldsymbol{x}^{\text{GenePT}}])$.

3.2 Augmented Acquisition Function with Enrichment Analysis

Incorporating prior knowledge into the BO framework has been widely discussed in the literature. We mainly focus on πBO [11], a principled generalization of the acquisition function to incorporate prior beliefs about the location of the optimum in the form of probability distributions $\pi(\boldsymbol{x})$. Specifically, for acquisition function $\alpha_{p(f_n|\mathcal{D}_n)}(\boldsymbol{x})$, the corresponding augmented acquisition function is:

$$\pi\alpha_{p(f_n|\mathcal{D}_n)}(\boldsymbol{x}) = \alpha_{p(f_n|\mathcal{D}_n)}(\boldsymbol{x})\pi_n(\boldsymbol{x})^{\frac{\beta}{L_n}}, \tag{6}$$

where β is a hyperparameter set by the user, reflecting their confidence in $\pi_n(\boldsymbol{x})$, and L_n is the number of labeled data so far. This reflects the intuition that, as the optimization progresses, we should increasingly trust the surrogate model over the prior, as BO will likely have enough data to reach the optimum confidently. This also comes with theoretical properties discussed in the next section.

In this work, we propose to augment the acquisition function with the prioritization results from enrichment analysis within the πBO framework. At each iteration n, we rank labeled genes according to their labels (i.e., change of phenotype under the gene knockout). We consider the top-k (e.g., top-10%) genes as the genes of interest, i.e., $\mathbb{S}_n$, and use enrichment analysis [24] to find top enriched pathways, ranked by the combined score defined in Eq.5. If one unlabeled gene is within the top pathway, we increase the probability of selecting the gene in the acquisition function. Specifically, we define the probability of selecting an unlabeled gene $\boldsymbol{x}$ as follows:

$$s_n(\boldsymbol{x}) = \text{logit}(\frac{1}{U_n}) + \frac{1}{t}\mathbf{agg}_{\{\mathbb{P}_i|\boldsymbol{x}\in\mathbb{P}_i, p_n(\mathbb{P}_i)<0.05\}}[c_n(\mathbb{P}_i)], \;\; \pi_n(\boldsymbol{x}) = \frac{e^{s_n(\boldsymbol{x})}}{\sum_{\boldsymbol{x}} e^{s_n(\boldsymbol{x})}}, \tag{7}$$

where U_n is the number of unlabeled genes at iteration n and $\mathbf{agg}[\cdot]$ is a set aggregation operation that summarizes the combined score $c_n(\cdot)$ at iteration n across all significant pathways (with p-value $p_n(\mathbb{P}_i) < 0.05$) that contains the unlabeled gene $\boldsymbol{x}$ and we use mean operation in practice. The hyperparameter temperature t controls the level of information that we keep from the enrichment analysis. When $t = \infty$, $\pi(\boldsymbol{x})$ reduces to a uniform distribution.

3.2.1 Theoretical properties

BioBO comes with the same *no-harm guarantee* as the original πBO [11], because of the decaying effect of the prior in Eq.6. When paired with the EI acquisition function, we can prove that the loss, $\mathcal{L}_n(\text{BioEI}_n)$, to the optimum at iteration n of the BioEI strategy, i.e., using the EI AF in Eq.6, can be bounded by the loss of the corresponding EI strategy, $\mathcal{L}_n(\text{EI}_n)$, using the Theorem 1 of [11] as following:

$$\mathcal{L}_n(\text{BioEI}_n) \leq C_{\pi,n}\mathcal{L}_n(\text{EI}_n), \;\; C_{\pi,n} = \left(\frac{\max_{\boldsymbol{x}} \pi_n(\boldsymbol{x})}{\min_{\boldsymbol{x}} \pi_n(\boldsymbol{x})}\right)^{\frac{\beta}{L_n}}. \tag{8}$$

Therefore, we have the *no-harm guarantee* that the loss of the BioEI strategy is asymptotically equal to the loss of the EI strategy:

$$\mathcal{L}_n(\text{BioEI}_n) \sim \mathcal{L}_n(\text{EI}_n), \tag{9}$$

which indicates that BioEI is robust against errors and biases from the enrichment analysis.

4 Experiments

4.1 GeneDisco Datasets

Datasets We use two large-scale genome-wide CRISPR assays, the log fold change of Interferon-γ (IFN-γ) and Interleukin-2 (IL-2) production in primary human T cells [34], from the GeneDisco dataset [8]. We also use the Achilles (dependency score of genetic intervention across cancer cell lines) [35] gene descriptor, i.e., gene embeddings $\boldsymbol{X}$, from GeneDisco. GeneDisco also includes two other different descriptors of genes, CCLE (quantitative proteomics information from cancer cell lines) [36] and STRING (protein-protein interactions) [37]. However, only Achilles is informative to predict the cell phenotypes, as shown in [8, 9]; therefore, we focus on the Achilles gene embedding from GeneDisco. We included another two embeddings: Gene2vec and GenePT, introduced in Section 3.1.

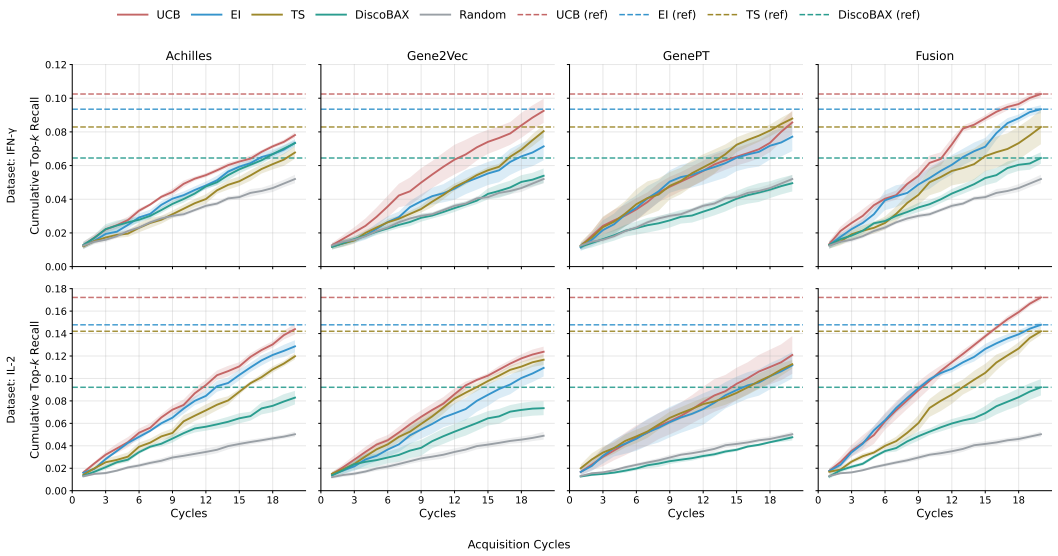

Figure 2: Performance across single modalities (Achilles, Gene2Vec, GenePT) and their Fusion on IFN-γ (top) and IL-2 (bottom). Columns: Achilles, Gene2Vec, GenePT, Fusion. Curves show UCB, EI, TS, DiscoBAX, and Random; Row-wise dashed lines indicate the Fusion value at the final cycle (20) for UCB/EI/TS/DiscoBAX to aid comparison.

Metrics for BO We use cumulative top-k recall to measure the ability of a method to identify the top interventions as those in the top percentile of the experimentally measured phenotypes following [9].

Metrics for surrogate model We measure the performance of the surrogate model on a separate test set using RMSE (Root Mean Squared Error) and ECE (Expected Calibration Error). Moreover, we calculate decile RMSE and decile ECE, i.e., RMSE and ECE on datapoints with top 10% highest phenotype value, to evaluate the performance of the surrogate model near the optimum.

Baselines In terms of surrogate modeling, we use the BNN defined in [9], using Achilles, Gene2Vec, GenePT, and the fusion of these three modalities. We use UCB, EI, TS, DiscoBAX as acquisition functions, as well as augmented acquisition functions, BioUCB, BioEI, and BioTS, with enrichment analysis using Gene Ontology (GO) [33] and Hallmark (HM) [21] pathways.

4.2 Exploring the effects of multi-modal gene features

Here, we study the effects of fusing multi-modal gene information in the surrogate model. Figure 2 shows the cumulative top-k recall of different acquisition functions at each cycle of the experimental design. We observe that all BO acquisition functions are better than random, especially the UCB acquisition function, and BO saves the labeling efforts 25%-75% compared with random, which indicates the benefits of BO in experimental design. Moreover, we observe that using surrogate models with multiple modalities is always better than using single-modal surrogate models, with labeling effort saving ranging from 4% to 40%. The best-performing model is using the fusion of three modalities with UCB. We also observe that DiscoBAX [9] is worse than existing standard acquisition functions most of the case. Therefore, we remove DiscoBAX in the following experiments.

In Table 1, we answer the question of why using fused gene representation in the surrogate model can improve BO. Intuitively, using a more expressive gene embedding can improve the prediction accuracy of the surrogate model as well as the uncertainty quantification, which will lead to better Bayesian optimization. We find that fusion does not decrease model RMSE and ECE. In fact, models with different features have similar RMSE, and the model with Gene2Vec achieves the lowest ECE; Therefore, **a lower RMSE or ECE does not always lead to a better BO**, which is consistent to the conclusions in [38]. Moreover, we observe that fusion improves the model accuracy and uncertainty quantification **near optimum** as shown by the decile RMSE and decile ECE when using almost all

Table 1: **Performance metrics with standard error of each acquisition function and each feature setup with the IL-2 phenotype**. We observe that although fusion does not improve the model prediction and uncertainty quantification globally (i.e., RMSE, ECE), it improves on data points that are close to the optimum (Decile RMSE and Decile ECE), which leads to better Bayesian optimization results (top-k recall). The best performance (with the smallest standard error) of each AF is bold.

AF	Feature	Top-k recall	Decile RMSE	RMSE	Decile ECE	ECE
EI	Fusion	**0.149 (0.002)**	**0.420 (0.008)**	0.233 (0.002)	**0.320 (0.005)**	0.098 (0.002)
EI	Achilles	0.128 (0.003)	0.424 (0.010)	0.233 (0.002)	0.325 (0.003)	0.101 (0.003)
EI	Gene2Vec	0.115 (0.002)	0.471 (0.006)	0.232 (0.001)	0.337 (0.001)	**0.091 (0.001)**
EI	GenePT	0.107 (0.005)	0.451 (0.008)	**0.229 (0.001)**	0.333 (0.002)	0.093 (0.002)
TS	Fusion	**0.142 (0.001)**	**0.425 (0.011)**	0.234 (0.002)	**0.315 (0.007)**	0.097 (0.003)
TS	Achilles	0.117 (0.001)	0.435 (0.009)	**0.233 (0.002)**	0.327 (0.004)	0.096 (0.002)
TS	Gene2Vec	0.119 (0.002)	0.469 (0.006)	**0.233 (0.002)**	0.336 (0.001)	**0.091 (0.001)**
TS	GenePT	0.133 (0.001)	0.458 (0.009)	0.233 (0.003)	0.328 (0.003)	0.095 (0.002)
UCB	Fusion	**0.174 (0.001)**	0.420 (0.008)	**0.233 (0.002)**	**0.323 (0.004)**	0.098 (0.002)
UCB	Achilles	0.142 (0.003)	**0.418 (0.010)**	0.236 (0.002)	0.325 (0.004)	0.105 (0.004)
UCB	Gene2Vec	0.126 (0.000)	0.471 (0.007)	**0.233 (0.002)**	0.337 (0.001)	**0.092 (0.001)**
UCB	GenePT	0.118 (0.011)	0.463 (0.009)	**0.233 (0.002)**	0.328 (0.004)	0.095 (0.002)
AVG	Fusion	**0.155 (0.001)**	**0.421 (0.009)**	0.233 (0.002)	**0.319 (0.005)**	0.098 (0.003)
AVG	Achilles	0.129 (0.001)	0.426 (0.010)	0.234 (0.002)	0.326 (0.004)	0.101 (0.003)
AVG	Gene2Vec	0.120 (0.002)	0.470 (0.006)	0.233 (0.001)	0.336 (0.001)	**0.091 (0.001)**
AVG	GenePT	0.120 (0.006)	0.457 (0.009)	**0.232 (0.002)**	0.330 (0.003)	0.094 (0.002)

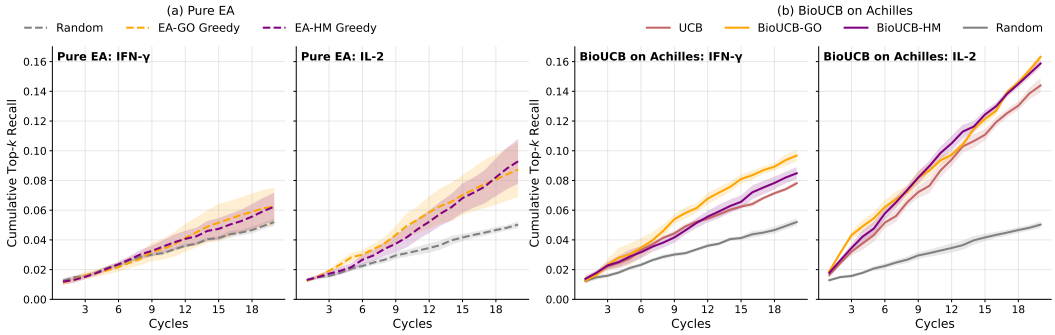

Figure 3: Pure EA vs. BioUCB on Achilles. Figure (A) (dashed): Pure EA on IFN-γ and IL-2 (Random, EA-GO Greedy, EA-HM Greedy). Figure (B) (solid): BioUCB on Achilles for IFN-γ and IL-2 (UCB, BioUCB-GO, BioUCB-HM, Random).

acquisition functions and also on average (AVG). The Spearman correlation of Top-k recall is -0.676 (p-value: 0.004) with decile RMSE and -0.765 (p-value: 0.0005) with decile ECE. Therefore, the improvement over local (decile) RMSE and ECE close to optimum explains the improvements BO achieves with modality fusion.

4.3 Exploring the effects of enrichment analysis

Here we study the benefits of combining enrichment analysis with Bayesian optimization using the proposed BioBO framework in design experiments. First, we analyze if the prior distribution, Eq.7, constructed from results of enrichment analysis is beneficial in experimental design, i.e., using a model-free approach. We select genes with the highest prior probabilities in Eq.7. Figure 3 (a) shows that using both gene ontology and hallmark as the pathway database can improve the design, compared with random.

Next, we combine the enrichment analysis prior with the acquisition function in BO, i.e., the model-based BioBO approach. We observe that adding the enrichment analysis prior can improve the labeling efficiency over BO with the corresponding acquisition function without the prior. Specifically, the enrichment analysis prior improves the labeling efficiency of UCB by 25% with Achillie gene

Table 2: **Cumulative top-k recall with standard error of each acquisition function on different datasets**. We observe that BioBO achieves the best performance on 19/24 different settings, and BioUCB-HM with surrogate function using fused features achieves the best performance for both IFN-γ and IL-2. The best performance (with the smallest standard error) is bold.

Phenotype: IFN-γ	Fusion	Achilles	GenePT	Gene2Vec
EI	0.095 (0.001)	0.072 (0.001)	**0.084 (0.004)**	0.079 (0.006)
BioEI-GO (ours)	0.095 (0.000)	0.072 (0.000)	0.082 (0.005)	0.075 (0.004)
BioEI-HM (ours)	**0.096 (0.001)**	**0.076 (0.001)**	0.062 (0.007)	**0.084 (0.002)**
TS	0.090 (0.001)	0.071 (0.001)	0.093 (0.002)	0.078 (0.002)
BioTS-GO (ours)	**0.095 (0.001)**	**0.095 (0.000)**	0.092 (0.004)	**0.083 (0.005)**
BioTS-HM (ours)	0.088 (0.001)	0.068 (0.005)	**0.093 (0.005)**	0.074 (0.004)
UCB	0.102 (0.001)	0.077 (0.001)	0.078 (0.004)	0.096 (0.005)
BioUCB-GO (ours)	0.102 (0.001)	**0.100 (0.002)**	0.084 (0.005)	0.098 (0.002)
BioUCB-HM (ours)	**0.110 (0.001)**	0.079 (0.003)	**0.102 (0.001)**	**0.103 (0.004)**
Random	0.050 (0.001)	0.050 (0.001)	0.050 (0.001)	0.050 (0.001)
Phenotype: IL-2	Fusion	Achilles	GenePT	Gene2Vec
EI	0.148 (0.002)	0.128 (0.003)	0.107 (0.005)	0.115 (0.002)
BioEI-GO (ours)	0.147 (0.003)	0.128 (0.003)	0.107 (0.005)	0.115 (0.002)
BioEI-HM (ours)	**0.149 (0.002)**	0.128 (0.003)	0.107 (0.005)	0.115 (0.002)
TS	0.142 (0.001)	0.117 (0.001)	**0.133 (0.014)**	0.119 (0.002)
BioTS-GO (ours)	0.147 (0.003)	**0.124 (0.002)**	0.098 (0.011)	**0.119 (0.001)**
BioTS-HM (ours)	**0.148 (0.002)**	0.123 (0.004)	0.106 (0.013)	0.119 (0.002)
UCB	0.174 (0.001)	0.143 (0.003)	0.118 (0.011)	0.125 (0.000)
BioUCB-GO (ours)	0.169 (0.001)	0.148 (0.001)	0.098 (0.008)	**0.128 (0.002)**
BioUCB-HM (ours)	**0.178 (0.001)**	**0.151 (0.001)**	**0.122 (0.012)**	0.125 (0.000)
Random	0.049 (0.001)	0.048 (0.001)	0.049 (0.001)	0.046 (0.002)

embedding on optimizing IFN-γ. We show the cumulative top-k recall of all experiments in Table 2, where we observe that the prior from enrichment analysis can improve the original acquisition function most of the time (19/24 cases), with the best performance achieved by BioUCB using hallmark with fused gene embeddings in both IFN-γ and IL-2.

4.4 Computational efficiency of BioBO

The runtime per iteration of BioBO is comparable to existing BO methods. We report detailed runtimes in Appendix D. The choice of 20 acquisition cycles (selecting 400 genes with 20 genes per cycle) follows exactly the experimental protocol established in [8, 9], ensuring comparability. The total of 400 perturbations selected by 20 iterations corresponds to less than 5% of the typical gene pool, aligning with realistic experimental budgets in high-throughput CRISPR screens [8, 9]. Thus, BioBO is fast from a practical standpoint and identifies high-value perturbations more efficiently compared to baseline methods.

5 Conclusion

We introduce BioBO, a biology-informed BO framework for perturbation design, combining standard BO with multimodal gene representations and enrichment analysis to guide experimental prioritization. Our theoretical analysis establishes a no-harm guarantee when integrating biological priors from enrichment analysis, ensuring robustness to noisy or biased pathway information. Empirical results on the GeneDisco datasets demonstrate substantial gains in sample efficiency, with BioBO outperforming traditional BO methods and enrichment-only strategies. By fusing principled optimization with domain-specific biological insights, BioBO enables more efficient discovery of high-value perturbations, reducing experimental costs. Looking forward, this approach provides a foundation for integrating broader biological knowledge sources—such as single-cell profiles and literature-derived embeddings—into experimental design frameworks, paving the way for faster and more targeted advances in genomics and therapeutic discovery.

References

[1] Yau-Tuen Chan, Yuanjun Lu, Junyu Wu, Cheng Zhang, Hor-Yue Tan, Zhao-xiang Bian, Ning Wang, and Yibin Feng. Crispr-cas9 library screening approach for anti-cancer drug discovery: overview and perspectives. *Theranostics*, 12(7):3329, 2022.

[2] Christoph Bock, Paul Datlinger, Florence Chardon, Matthew A Coelho, Matthew B Dong, Keith A Lawson, Tian Lu, Laetitia Maroc, Thomas M Norman, Bicna Song, et al. High-content crispr screening. *Nature Reviews Methods Primers*, 2(1):8, 2022.

[3] Fuguo Jiang and Jennifer A Doudna. Crispr–cas9 structures and mechanisms. *Annual review of biophysics*, 46:505–529, 2017.

[4] Federico Abascal, David Juan, Irwin Jungreis, Laura Martinez, Maria Rigau, Jose Manuel Rodriguez, Jesus Vazquez, and Michael L Tress. Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic acids research*, 46(14):7070–7084, 2018.

[5] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[6] Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[7] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[8] Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. Genedisco: A benchmark for experimental design in drug discovery. 2021.

[9] Clare Lyle, Arash Mehrjou, Pascal Notin, Andrew Jesson, Stefan Bauer, Yarin Gal, and Patrick Schwab. Discobax: Discovery of optimal intervention sets in genomic experiment design. In *International Conference on Machine Learning*, pages 23170–23189. PMLR, 2023.

[10] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[11] Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. π bo: Augmenting acquisition functions with user beliefs for bayesian optimization. *arXiv preprint arXiv:2204.11051*, 2022.

[12] Jonas Mockus. The application of bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117, 1998.

[13] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[14] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, number 118, pages 1015–1022. PMLR, 2010.

[15] Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

[16] Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

[17] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[18] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. Go:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[19] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.

[20] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[21] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[22] Winston Haynes. Bonferroni correction. In *Encyclopedia of systems biology*, pages 154–154. Springer, 2013.

[23] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[24] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013.

[25] Samuel Katz, Jian Song, Kyle P Webb, Nicolas W Lounsbury, Clare E Bryant, and Iain DC Fraser. Signal: A web-based iterative analysis platform integrating pathway and network approaches optimizes hit selection from genome-scale assays. *Cell systems*, 12(4):338–352, 2021.

[26] Yueshan Zhao, Min Zhang, and Da Yang. Bioinformatics approaches to analyzing crispr screen data: from dropout screens to single-cell crispr screens. *Quantitative Biology*, 10(4):307–320, 2022.

[27] Weiwei Dai, Fengting Wu, Natalie McMyn, Bicna Song, Victoria E Walker-Sperling, Joseph Varriale, Hao Zhang, Dan H Barouch, Janet D Siliciano, Wei Li, et al. Genome-wide crispr screens identify combinations of candidate latency reversing agents for targeting the latent hiv-1 reservoir. *Science translational medicine*, 14(667):eabh3351, 2022.

[28] Azucena Ramos, Catherine E Koch, Yunpeng Liu-Lupo, Riley D Hellinger, Taeyoon Kyung, Keene L Abbott, Julia Fröse, Daniel Goulet, Khloe S Gordon, Keith P Eidell, et al. Leukemia-intrinsic determinants of car-t response revealed by iterative in vivo genome-wide crispr screening. *Nature Communications*, 14(1):8048, 2023.

[29] Adriana Ordóñez, David Ron, and Heather P Harding. Protocol for iterative enrichment of integrated sgrnas via derivative crispr-cas9 libraries from genomic dna of sorted fixed cells. *STAR protocols*, 5(4):103493, 2024.

[30] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[31] Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, 9(4):483–493, 2025.

[32] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(Suppl 1):82, 2019.

[33] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[34] Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W Freimer, Vinh Q Nguyen, Franziska Blaeschke, and Alexander Marson. Crispr activation and interference screens in primary human t cells decode cytokine regulation. *bioRxiv*, pages 2021–05, 2021.

[35] Joshua M Dempster, Jordan Rossen, Mariya Kazachkova, Joshua Pan, Guillaume Kugener, David E Root, and Aviad Tsherniak. Extracting biological insights from the project achilles genome-scale crispr screens in cancer cell lines. *BioRxiv*, page 720243, 2019.

[36] David P Nusinow, John Szpyt, Mahmoud Ghandi, Christopher M Rose, E Robert McDonald, Marian Kalocsay, Judit Jané-Valbuena, Ellen Gelfand, Devin K Schweppe, Mark Jedrychowski, et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, 180(2):387–402, 2020.

[37] Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

[38] Jonathan Foldager, Mikkel Jordahn, Lars K Hansen, and Michael R Andersen. On the role of model uncertainties in bayesian optimisation. In *Uncertainty in Artificial Intelligence*, pages 592–601. PMLR, 2023.

[39] Ihor Neporozhnii, Julien Roy, Emmanuel Bengio, and Jason Hartford. Efficient biological data acquisition through inference set design. 2025.

[40] Muhammad Arslan Masood, Samuel Kaski, and Tianyu Cui. Molecular property prediction using pretrained-bert and bayesian active learning: a data-efficient approach to drug design. *Journal of Cheminformatics*, 17(1):58, 2025.

[41] Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10):1066–1075, 2023.

[42] Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In *International Conference on Research in Computational Molecular Biology*, pages 17–37. Springer, 2024.

[43] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[44] Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *International conference on artificial intelligence and statistics*, pages 7331–7348. PMLR, 2023.

[45] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[46] Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, and Andrew Gordon Wilson. Accelerating bayesian optimization for biological sequence design with denoising autoencoders. In *International Conference on Machine Learning*, pages 20459–20478. PMLR, 2022.

[47] Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in Neural Information Processing Systems*, 36:12489–12517, 2023.

[48] Siddharth Ramchandran, Manuel Haussmann, and Harri Lähdesmäki. High-dimensional bayesian optimisation with gaussian process prior variational autoencoders. In *International Conference on Learning Representations*, 2025.

[49] Aldo Pacchiano, Drausin Wulsin, Robert A Barton, and Luis Voloch. Neural design for genetic perturbation experiments. 2023.

[50] Yongju Lee, Dyke Ferber, Jennifer E Rood, Aviv Regev, and Jakob Nikolas Kather. How ai agents will change cancer research and oncology. *Nature Cancer*, 5(12):1765–1767, 2024.

[51] Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. 2025.

[52] Minsheng Hao, Yongju Lee, Hanchen Wang, Gabriele Scalia, and Aviv Regev. Perturboagent: A self-planning agent for boosting sequential perturb-seq experiments. *bioRxiv*, pages 2025–05, 2025.

[53] Petrus Mikkola, Milica Todorović, Jari Järvi, Patrick Rinke, and Samuel Kaski. Projective preferential bayesian optimization. In *International Conference on Machine Learning*, pages 6884–6892. PMLR, 2020.

[54] Masaki Adachi, Brady Planden, David A Howey, Michael A Osborne, Sebastian Orbell, Natalia Ares, Krikamol Muandet, and Siu Lun Chau. Looping in the human collaborative and explainable bayesian optimization. *arXiv preprint arXiv:2310.17273*, 2023.

[55] José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *International conference on machine learning*, pages 1699–1707. PMLR, 2015.

[56] Masaki Adachi, Satoshi Hayakawa, Martin Jørgensen, Harald Oberhauser, and Michael A Osborne. Fast bayesian inference with batch bayesian quadrature via kernel recombination. *Advances in Neural Information Processing Systems*, 35:16533–16547, 2022.

[57] Artur Souza, Luigi Nardi, Leonardo B Oliveira, Kunle Olukotun, Marius Lindauer, and Frank Hutter. Bayesian optimization with a prior for the optimum. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 265–296. Springer, 2021.

[58] Abdoulatif Cissé, Xenophon Evangelopoulos, Sam Carruthers, Vladimir V Gusev, and Andrew I Cooper. Hypbo: Accelerating black-box scientific experiments using experts' hypotheses. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3881–3889, 2024.

A Related Work

Experimental design in drug discovery Many drug discovery and design applications have been using experimental design to speed up the process. Active learning, a framework that finds the most informative unlabeled datapoints to label for improving the model, has been applied to molecular property prediction [39, 40], Perturb-seq [41, 42], and genomics CRISPR assays [8]. Active learning uses the information gain of the probabilistic surrogate model to guide the selection, such as BALD [43] and EPIG [44]; therefore, it is an exploration-only process. On the other hand, Bayesian optimization trades off between exploration and exploitation to query the most informative unlabeled datapoints to the optimum. BO has been applied to bio-sequence optimization by combining with deep generative models, including small-molecular and protein sequences [45, 46, 47, 48], as well as on genomics CRISPR assays [49, 9]. Recently, large language model (LLM) based agents have shown great potential in experimental design by leveraging the rich background knowledge and reasoning capabilities [50, 51], and enrichment analysis has been shown to be an important tool in the multi-agent system [52].

Exploiting external knowledge in BO Incorporating external knowledge in BO has recently been studied extensively. External knowledge can be elicited from the feedback of human experts through preference learning and used in BO [53, 54] when the explicit knowledge is challenging to obtain. However, when the external knowledge on the input space over the potential candidates is ready, it can be either treated as a constraint [55, 56] or a prior belief [57, 58, 11], and our BioBO fits within this framework.

B Data description

GeneDisco contains three different embeddings: Achilles, String, and CCLE, which are available for 17,655, 17,972, and 11,943 genes. We also consider two gene embeddings: Gene2Vec and GenePT, which are available for 23,940 and 61,287 genes. In order to remove the effect of the different missingness level of each gene embedding, we use the 10,556 genes that have all five embeddings. We consider two phenotypes in GeneDisco, IFN-γ and IL-2, which are available for 18,416 genes. Therefore, we consider an intersection of genes with all modalities and two phenotypes, which contains 10,467 genes.

C Experimental details

Device details. All experiments were run on Debian GNU/Linux 10 (buster) with Python 3.10.16, PyTorch 2.6.0, and CUDA 12.8. Training and inference used two NVIDIA L4 GPUs (each with 24 GB VRAM). The host machine had an AMD EPYC 7R13 processor with 192 hardware threads and 80 GB of system memory. Computations used 64-bit floating-point precision where required by the Bayesian layers.

Hyperparameters. Unless noted, the BNN surrogate used a *2*-layer MLP with hidden width *64* and ReLU activations, Bayesian weights with Gaussian priors (mean 0, std. 1), and observation noise std. 0.5. We optimized with Adam (learning rate $\eta = 0.001$, weight decay $\lambda = 0.0001$) for up to 200 epochs with early stopping (patience 30) on a 10% validation split; batch size was 256. Monte Carlo inference used 100 stochastic forward passes per acquisition. For modality fusion we concatenated L2-normalized embeddings (Achilles, Gene2Vec, GenePT; and where used, CCLE/STRING). Acquisition functions followed standard definitions for UCB (trade-off κ= 1), EI ($\xi = 0$), and TS; biology-informed variants added enrichment weights from GO or Hallmark with temperature coefficient α = 0.1 and β = 1 (IFN-γ), α = 0.01 and β = 0.1 (IL-2). Hyperparameters were selected on the validation split and kept fixed across cycles.

Reproducibility and error bars. For every dataset–modality–acquisition setting we ran **three** independent random seeds (`1,1000,2000`). Plotted curves report the mean across seeds; shaded bands show $\pm$ s.e.m. (standard error of the mean). Final-cycle bar plots likewise report mean $\pm$ s.e.m.

D Runtime Comparison

We report average runtime per iteration (evaluating 20 genes per cycle) for BioBO and baseline BO methods over all datasets. All experiments were run on a standard GPU (NVIDIA A10). The table includes variants with and without multimodal fusion and enrichment analysis (EA) to show the computational overhead introduced by these components. While multimodal fusion and EA slightly increase runtime compared to single-modality models, the additional cost remains modest and practical for typical high-throughput CRISPR experiments.

Method	Avg Runtime per BO Cycle (s)
UCB (Achilles)	8.55
UCB (Fusion)	10.50
BioUCB-HM (Fusion)	12.45
EI (Achilles)	7.64
EI (Fusion)	12.57
BioEI-HM (Fusion)	13.05
TS (Achilles)	6.95
TS (Fusion)	12.18
BioTS-HM (Fusion)	12.87

Table 3: Runtime per iteration for BioBO and baseline BO methods, averaged over datasets. Variants with multimodal fusion and/or enrichment analysis (EA) are included to show the overhead of these components.

E Sensitivity to the top-k% Threshold in Enrichment Analysis

We evaluate the effect of different top-k% thresholds used for enrichment analysis, varying k from 5% to 50% on the IFN-γ dataset (Achilles features). As shown in Table 4, BioBO remains robust for k between 5–20%, exhibiting only minor performance variation. Larger thresholds (30–50%) dilute the enrichment signal by including a broader, noisier set of genes, leading to slightly reduced BO performance. We use k = 10% as a practical default.

Top-k%	5%	10%	15%	20%	30%	50%
BioEI-GO	0.090 ± 0.001	0.085 ± 0.006	0.084 ± 0.009	0.085 ± 0.010	0.074 ± 0.002	0.071 ± 0.001

Table 4: Sensitivity of BioBO to top-k% used for enrichment analysis. Performance shown as cumulative top-k recall.

F Supplementary Experimental Results

Across both datasets and all four representations (Achilles, Gene2Vec, GenePT, Fusion), biology-informed variants (BioUCB/BioEI/BioTS) generally match or exceed their base counterparts (UCB/EI/TS), with the most consistent gains on Fusion. Improvements are most evident in early–mid cycles (better sample efficiency) and narrow later as methods converge. UCB remains a strong base policy; GO and Hallmark offer broadly comparable lifts. The Random baseline is consistently inferior.

CCLE and STRING yield substantially lower absolute recall and a reduced dynamic range versus the main representations, indicating these are comparatively poor features; this is why they are excluded from the main paper and reported here for completeness. Even so, biology-informed variants provide modest, consistent gains over their bases—particularly at smaller budgets.

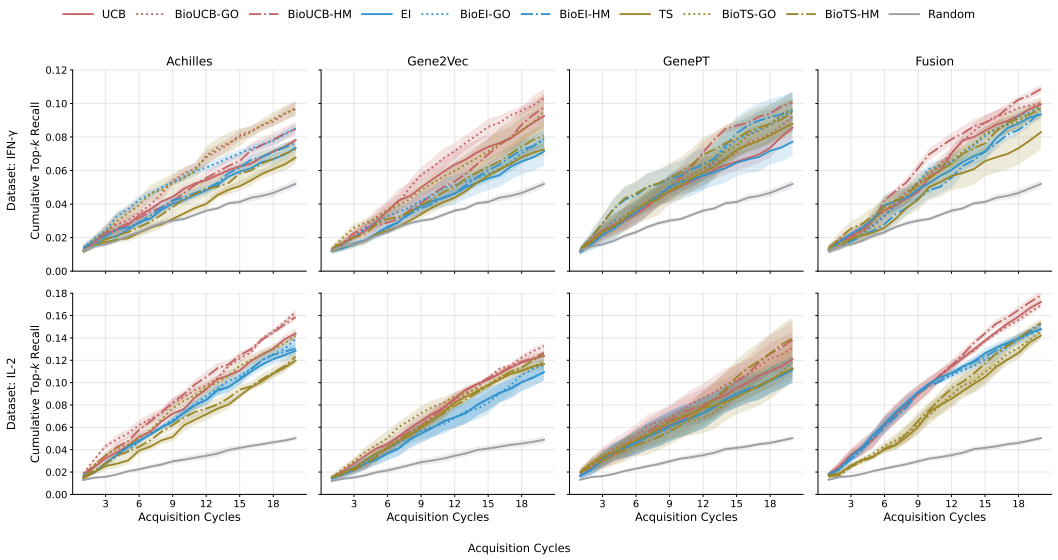

Figure 4: Base vs. BioBO across modalities and datasets. Rows: IFN-γ (top), IL-2 (bottom). Columns: Achilles, Gene2Vec, GenePT, Fusion. Solid lines show base acquisitions **UCB/EI/TS**; dotted lines show GO-informed variants (**BioUCB-GO/BioEI-GO/BioTS-GO**); dash–dot lines show Hallmark-informed variants (**BioUCB-HM/BioEI-HM/BioTS-HM**). Shaded ribbons denote mean $\pm$ s.e.m. over replicates.

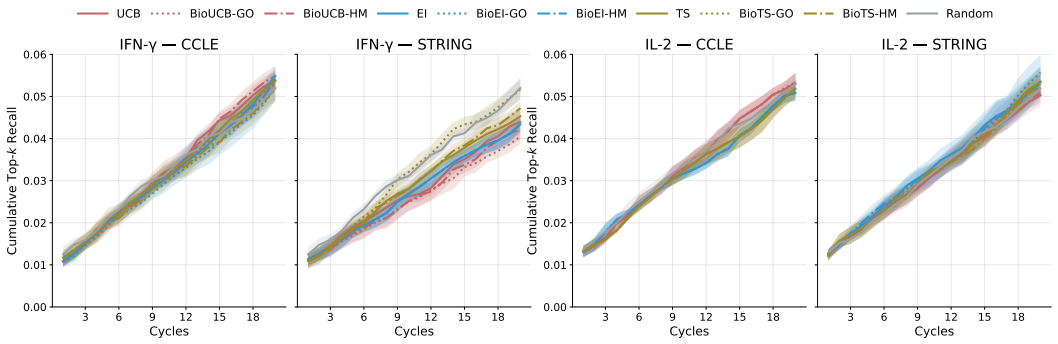

Figure 5: CCLE and STRING modalities across datasets. Panels (left$\rightarrow$right): IFN-γ—CCLE, IFN-γ—STRING, IL-2—CCLE, IL-2—STRING. Curves show base acquisitions **UCB/EI/TS** (solid), biology-informed variants **BioUCB/BioEI/BioTS** with **GO** (dotted) and **HM** (dash–dot) in the same family color, plus **Random** (gray). Shaded ribbons denote mean $\pm$ s.e.m.