BEYOND SINGLE-AXIS DESIGNS: MULTI-OBJECTIVE OPTIMIZATION FOR COMPLEX PERTURBATION ATLASES

Zihe Zheng*, **Soroor Hediyeh Zadeh***, **Jenni Liu & Fabian Theis**
Institute of Computational Biology
Helmholtz Center Munich
Neuherberg, 85764, Germany
{zihe.zheng, soroor.hediyehzadeh, jenni.liu, fabian.theis}@helmholtz-munich.de

ABSTRACT

Single-cell perturbation screens enable detailed profiling of complex cellular responses but experiments remain costly in both time and resources. Bayesian optimization (BO) has been proven effective in data-limited and evaluation-expensive settings, including industrial and molecular design. Recent emergence of large-scale perturbation atlases have opened up the opportunity to inform and guide the design of future experiments. As the experimental design of perturbation atlases becomes more complex, there is a growing need for methods that can identify the most informative axes of variation from the design space of these atlases for the optimal design of experiments that are often inherently multi-objective. In this paper, we extend the existing work on single-axis design spaces to multi-axis complex design spaces, and construct a design space for the Human Cytokine Dictionary (HuCIRA). We introduce Derivative-based Global Sensitivity Measures (DGSM) for single-cell perturbation experimental design and demonstrate that DGSM is an effective strategy for querying the axes of variation relevant to an objective from the HuCIRA perturbation atlas. To demonstrate the practical effectiveness of our framework for representing multi-axis design spaces in large-scale perturbation atlases, we emulate a multi-objective Bayesian optimization (MOBO) experiment using HuCIRA, showing that the selected experimental designs jointly optimize perturbation objectives. We envision that our proposed framework can be used in the design and utilization of current and future perturbation experiments and atlases.

1 INTRODUCTION

Single-cell perturbation screens systematically investigate gene functions and pathways and reveal heterogeneous cellular responses at single-cell resolution. However, due to the infinite combination of cell states, genetic background, perturbagens, dosage and time points, it is unfeasible to exhaustively explore the full space of experimental conditions. The design of perturbation experiments has become an optimization problem under extreme cost constraints.

Large-scale perturbation atlases with single-cell readout have been emerging over recent years Srivatsan et al. (2020)Replogle et al. (2022)Zhang et al. (2025)Oesinghaus et al. (2025). These atlases capture transcriptomic profiles of millions of cells, under hundreds of chemical or genetic perturbations at several time points and dosages, which accelerates the methodological developments for virtual screening, cell state engineering and biological hypothesis generation. Recent perspectives envision the extension of perturbation atlases towards increased context and readout diversity as well as combinatorial perturbations Rood et al. (2024)Dimitrov et al. (2026). In line with this initiative, companies such as TAHOE and Parse Biosciences are partnering with leading research institutes to generate foundational perturbation atlases at unprecedented scale, fueling research and model development efforts toward virtual cells.Tahoe Therapeutics (2026)Parse Biosciences (2026).

*These authors contributed equally to this work.

The generation of large perturbation atlases brings new opportunities and challenges. One realistic consideration in the experimental design is the optimization of the multi-dimensional design space under time and resource constraints. For example, one needs to navigate the trade-off between the breadth and depth of the experiment: determining whether to expand the number of perturbations, or have the perturbation on a wider range of biological contexts (e.g. diverse cell lines). Such decisions have to be made in order to derive a diverse response landscape that best facilitates the inference of cause-effect relationships. Current experimental design tools focus on the iterative selection of genes Li et al. (2025) or small molecules to optimize hit rates with active learning and acquisition functions Mehrjou et al. (2021)Huang et al. (2024), but ignore the selection of biological contexts as a primary goal. Notably, these prior works are concerned with optimizing designs along a single experimental axis.

Once a perturbation atlas is generated, another challenge is to use them as prior knowledge to inform future perturbation screens, for example, to nominate perturbations for cell state engineering. BioDiscoveryAgent Roohani et al. (2024), an LLM-based AI agent was developed to optimize gene selections by leveraging literature search and gene search based on biological databases, but does not explicitly learn from large-scale perturbation atlases. A recent review compares the effect of random or meaningful feedback from the experimental loop to the AI agent, and claims that LLMs depend heavily on priors and do not learn from experimental feedback Gupta et al. (2025). This motivates the need for an experimental design framework that is capable of learning the functional relationships from the perturbation atlases and enable knowledge transfer across contexts to guide smaller experiments.

To address these challenges, we leveraged the Bayesian Optimization (BO) framework for perturbation experimental design. Thanks to recent methodological advances in high-dimensional BO Papenmeier et al. (2025) and multi-objective acquisition functions Ament et al. (2023), BO has become well-suited for modeling the perturbation design space as well as optimizing solutions that balance exploration and exploitation. The BO framework also allows the navigation of trade-offs among design axes. In particular Belakaria et al. (2024) proposed a derivative-based global sensitivity measurement in BO, that specifically measures the sensitivity of the optimization objective with respect to each design axis.

In this paper, we present a BO framework that scales to single-cell perturbation atlases, accommodates complex design spaces and multi-objective optimizations, and enables sensitivity analysis with respect to each experimental axis. Our contributions are as follows:

- We extend the existing single-axis design space formulations to multiple-axes, complex designs.
- We apply this formulation to model a complex design space for the Human Cytokine Dictionary.
- We provide a framework that identifies objective-specific axes of variation from perturbation atlases to guide future experimental design via derivative-based sensitivity analysis.
- We show that selections by MOBO based on the formulated design space are jointly optimal for a set of perturbation objectives.

2 METHODS

2.1 DATASET

We used the recently published Human Cytokine Dictionary (HuCIRA) Oesinghaus et al. (2025) for our experiments. HuCIRA contains single-cell transcriptomics of 10 million human peripheral blood mononuclear cells (PBMC) from 12 donors, under stimulation of 90 cytokines and PBS control condition.

2.2 REPRESENTATION OF THE DESIGN SPACE

There are three design axes of variation in the HuCIRA dataset: Donor, cell type and cytokine. Each design axis is represented by biologically or chemically meaningful embeddings. We trained an scVI Lopez et al. (2018) model on the single-cell RNA-seq dataset to obtain a 10-dimensional

representation of each cell. Cells under the PBS control condition were aggregated per donor to get the donor representation and aggregated per cell type to get the cell type representation. We obtained the cytokine embeddings in HuCIRA from the CellFlow Klein et al. (2025) package, where the cytokines were embedded using ESM2 Zhang et al. (2024) protein language model. Then, the principal component analysis (PCA) was applied to reduce the dimensionality to 10. Each donor-cell type-cytokine combination in the dataset was generated by concatenation of the individual embeddings, resulting in a 30-dimension vector, representing the context and stimulation performed in the experiment.

2.3 Optimization objectives

In the design of single-cell perturbation experiments, the experimenter often has multiple conflicting or correlating objectives that one wishes to optimize. One possible objective is to maximize the effect magnitude of the perturbation condition, which can be characterized by a distance measurement from the perturbed to the control cells in the perturbation context Peidli et al. (2024). Biological pathways and cell state signatures are among the other desired optimization goals when designing perturbation experiments. Multi-objective Bayesian optimization (MOBO) provides a reasonable framework for perturbation experimental design, where achieving a trade-off between biological objectives is more desired than optimization of all objectives at the same time (Figure 1).

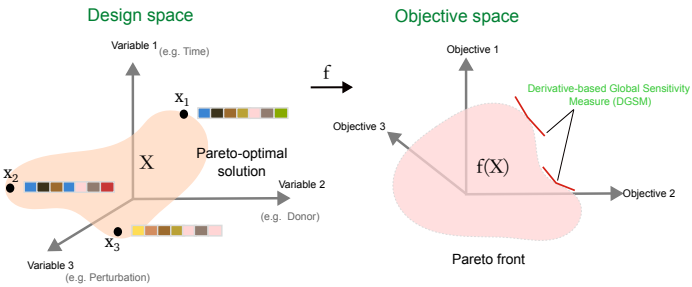

Figure 1: Schematic representation of Multi-Objective Bayesian Optimization (MOBO) and Derivative-based Global Sensitivity Measure (DGSM).

To design objectives for the perturbation design space of HuCIRA, we used a variant of the response magnitude value from the paper analysis pipeline Oesinghaus et al. (2025). For each donor and cell type, we first compute absolute-$\log_2$-fold-change of gene expression relative to PBS control, and then downweight small effects. The values are then summed across genes and taken the square root to obtain the final score, which we denote as `abs_sum`.

We scored the perturbed cells with AUCell Aibar et al. (2017) by the expression of Interferon-stimulated genes (ISG) Schoggins & Rice (2011). In addition, we obtained cellular disease state signatures of multiple sclerosis (MS) from a single-cell transcriptomics dataset Schafflick et al. (2020) through differential expression and scored the perturbed cells by differentially expressed genes in B cell, NK cell and CD8 T cell. The scores are then aggregated for each perturbation condition, and used as optimization objectives.

2.4 Derivative-based global sensitivity analysis

Let $X = [\mathbf{x}_1, \cdots \mathbf{x}_t]$ be a set of observed inputs in the input space, $\mathcal{X}$. Let $Y = [y_1, \cdots, y_t]$ be the function evaluations at those inputs, that is $y = f(x) + \epsilon$, where f is black-box. Let $\mathcal{D} = \{X, Y\}$ be the full observed data. We learn a surrogate model of f on $\mathcal{D}$. A Gaussian Process (GP) model is often used as a surrogate of the black-box function, since it is differentiable when using a twice differentiable kernel function and a differentiable mean function.

Derivative-based global sensitivity measures (DGSM) are defined as the integral over the input space of a function of the derivative of the black-box function (Kucherenko & Iooss (2017)). DGSM quantify the sensitivity of f to each input dimension $x_i, i = 1, \cdots, d$. Denote DGSM as $S(f, \mathcal{X})$. Given a surrogate model $\hat{f}$, we estimate S as the sensitivity of the surrogate, that is $\hat{S}(f, \mathcal{X}|\mathcal{D}) = S(\hat{f}, \mathcal{X})$. Specifically, we use the squared-DGSM:

$$S_{\text{sq}}(\hat{f}, \mathcal{X})_i = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} \left(\frac{\partial \hat{f}(\mathbf{x})}{\partial x_i} \right)^2 d\mathbf{x}$$

We use the results from Williams & Rasmussen (2006) who prove that the derivative of a GP is a GP to evaluate the gradient at $\mathbf{x}_*$. Given the observed data $\mathcal{D}$, the gradient at $\mathbf{x}_*$ has a multivariate normal distribution: $\nabla f(\mathbf{x}_*)|\mathcal{D} \sim \mathcal{N}(\mu'_*, \Sigma'_*)$, where

$$\mu'_* = \nabla m_{\mathbf{x}_*} + \nabla_{\mathbf{x}_*} \mathcal{K}_{\mathbf{x}_*, X} K_{\mathcal{D}}^{-1} (Y - m_X) \tag{1}$$

$$\Sigma'_* = \nabla^2_{\mathbf{x}_*} \mathcal{K}_{\mathbf{x}_*, \mathbf{x}_*} - \nabla_{\mathbf{x}_*} \mathcal{K}_{\mathbf{x}_*, X} K_{\mathcal{D}}^{-1} \nabla_{\mathbf{x}_*} \mathcal{K}_{X, \mathbf{x}_*} \tag{2}$$

In Bayesian optimization, DGSMs have been used as acquisition functions to acquire new samples from input design space Belakaria et al. (2024). We leverage DGSM to measure the sensitivity of each optimization objective to each design axis in context of perturbation experimental design.

2.5 Hypervolume-based and information-theoretic-based acquisition functions for selection of perturbation designs

To evaluate the validity of our framework for modeling the multi-axis design space of single-cell perturbation atlases, we emulated a MOBO experiment on the design space that we constructed for HuCIRA using parallel Noisy Expected Hypervolume Improvement (qNEHVI) Daulton et al. (2021) and parallel multi-objective Lower Bound Max-value Entropy Search (qLB-JES) Tu et al. (2022) acquisition functions. For a batch of q points $\mathcal{X}_{\text{cand}} = \{\mathbf{x_i}\}_{i=1}^{q}$, qNEHVI is defined as:

$$\hat{\alpha}_{\text{qNEHVI}}(\mathcal{X}_{\text{cand}}) = \frac{1}{N} \sum_{t=1}^{N} \text{HVI}(\tilde{\mathbf{f}}_t(\mathcal{X}_{\text{cand}})|\mathcal{P}_t),$$

where HVI is the hypervolume improvement in posterior distribution over function values at $\mathcal{X}_{\text{cand}}$, and $\mathcal{P}_t$ is the Pareto front over the previously evaluated points under the sampled function $\tilde{\mathbf{f}}_t \sim p(\mathbf{f}|\mathcal{D}_n)$ for $t = 1, \cdots, N$ samples from the posterior, using the $\mathcal{D}_n = \{\mathbf{x}_i, y_i, (\Sigma_i)\}_{i=1}^{n}$ points from the input space.

The qLB-JES acquisition function is defined as follow:

$$\hat{\alpha}_{\text{qLB-JES}}(\mathcal{X}_{\text{cand}}|\mathcal{D}_n) = H[p(y_{\text{cand}}|\mathbf{x}, \mathcal{D}_n)] - \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{q} h((\mathbb{X}_t^*, \mathbb{Y}_t^*); \mathbf{x}^i, \mathcal{D}_n),$$

where $(\mathbb{X}_t^*, \mathbb{Y}_t^*)$ are the Pareto-optimal input-output pairs at the tth iteration, h is the conditional entropy estimate and

$$H[p(y_{\text{cand}}|\mathbf{x}, \mathcal{D}_n)] = \frac{M}{2} \log(2\pi e) + \frac{1}{2} \sum_{m=1}^{M} \log \det(\Sigma_n^{(m)}(\mathcal{X}_{\text{cand}}, \mathcal{X}_{\text{cand}}) + \text{diag}(\sigma^{(m)}(\mathcal{X}_{\text{cand}})))$$

is the initial entropy and M is the total number of objectives. We later examine how the designs selected by these conventional acquisition functions provide insight into the validity of the design space construction.

3 Experiments and Results

3.1 Sensitivity Analysis

The HuCIRA experimental design dataset consists of 11,677 different combinations of donor, cell type, and cytokine stimulation, along with the computed value for each objective. Since the response magnitude (`abs_sum`) is left-skewed, a log transformation was applied to the values. Additionally, we performed min-max normalization of the design dimensions and standardization of the objectives. We randomly split the dataset into 10 chunks, where each of the chunk is split to train (80%)

and test (20%) set to fit a single-task GP regression model for each of the objectives. After fitting the model, we evaluated the DGSM per design dimension and summed the values per design axis to represent the objective-specific sensitivity.

We find that the response magnitude of the cytokine stimulation compared to PBS control is the most sensitive to the variation of cytokines, with much larger DGSM than cell type and donor (Figure 2 a), indicating that more cytokines need to be included in the experiment if a spectrum of diverse response magnitude is the goal.

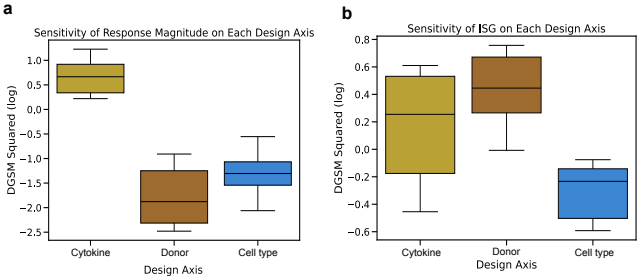

Figure 2: Log of DGSM-squared summed over the design dimensions of each design axis across 10 chunks. **a** Sensitivity of response magnitude **b** Sensitivity of ISG expression score

Similar analysis results were highlighted in HuCIRA that response magnitudes vary among cytokines while being consistent between donors. It was also shown that one group of donors in the dataset have a high baseline expression of interferon-stimulated genes (ISG), although the response to cytokine is mostly consistent with the non-interferon group. In line with the findings, we observed a high DGSM of ISG on the donor axis, which suggests the inclusion of more donors for a more variable ISG expression readout (Figure 2 b).

3.2 Using HuCIRA to Select Perturbations Recapitulating Multiple Sclerosis Cell Type–Specific Signatures

Nominating perturbations that optimize specific cellular state signatures or pathway activities from single-cell perturbational atlases is a realistic and compelling objective, particularly for cell state engineering applications. Disease signatures are often cell type–specific and may comprise multiple distinct states, motivating a multi-objective Bayesian optimization framework. Multiple sclerosis (MS) is an autoimmune disease characterized by diverse, cell type–specific transcriptomic changes. Here, we use MS as a case study to demonstrate how a Bayesian optimization loop can iteratively select experimental designs—namely cytokine–cell type–donor triplets— from HuCIRA that produce responses that are transcriptionally similar to MS-associated disease cell states.

We implemented an iterative selection BO loop to select experimental conditions that maximize cell-type specific disease signatures in a non-dominated way to identify cytokine perturbations that may regulate the disease mechanism. The selection of the designs was made using the qNEHVI and qLB-JES acquisition functions introduced earlier. We compared the selections made by these two acquisition functions to random selection as the baseline.

A GP regression model was trained to predict each objective i.e. MS-disease cell state scores in B cells, NK cells and CD8 T cells. The initial fit was performed on 20 randomly selected training samples. In each iteration, 10 new samples are selected from the unseen pool either randomly or by optimizing the qNEHVI or qLB-JES acquisition function, which are then included in the next model fitting. We ran 10 iterations for each selection strategy, accumulating to 100 selected samples in the end.

The model performance is evaluated at the end of each iteration by the Hypervolume (HV) and Mean Squared Error (MSE) on all unseen samples for the specific selection strategy. Hypervolume is defined as the space dominated by the current set of samples, including the newly selected samples in the current iteration and all samples seen before by the strategy.

We observed a steady increase of HV across all selection strategies in the first three iterations, reflecting the contribution of newly selected, diverse samples to the HV improvement observed even under the random selection strategy (Figure 3a). From the fourth iteration onward, the HV for the random strategy plateaued, whereas both qNEHVI and qLB-JES continued to expand the

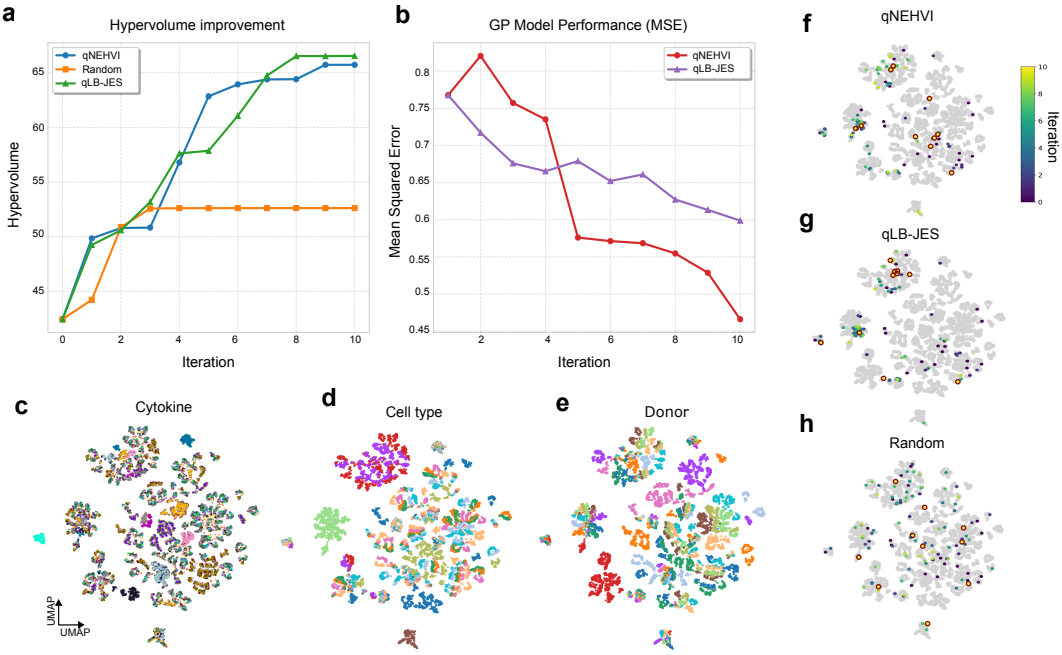

Figure 3: Results of the iterative selection BO loop. **a** Improvement of hypervolume over 10 iterations for qNEHVI, qLB-JES acquisition functions and random selection. **b** Mean Squared Error of the GP regression model over 10 iterations for qNEHVI, qLB-JES acquisition functions. **c-e** UMAP of the design space colored by different cytokines, cell types and donors. **f-h** UMAP of the design space, where the designs selected by qNEHVI, qLB-JES and random are colored. The spectrum shows the iteration in which the design is selected. Designs selected in the last round are circled in red.

HV, eventually reaching a plateau, suggesting convergence toward the edge of the underlying Pareto front. Notice that qLB-JES reached this plateau earlier than qNEHVI, consistent with its design to progressively reduce uncertainty in Pareto front estimation. The iterative selection of samples has also benefited the performance of the GP regression model (Figure 3b). For both qNEHVI and qLB-JES, the MSE of the GP model continues to drop even after the HV has plateaued (Figure 3a-b), indicating that the selected samples remained informative and contributed to more accurate model fitting. This highlights that the BO loop not only selects the jointly optimal experimental designs, but also enriches the training dataset for the GP surrogate model.

Next, we examined the list of designs selected by each strategy, by standardizing and projecting the design space to a UMAP, which shows distinct cluster of cytokines, cell types and donors, as well as their combinations (Figure 3c-e, Figure S1).

While the designs nominated by random selections were scattered in the design space (Figure 3h), selections by qNEHVI and qLB-JES formed localized clusters (Figure 3f-g), visualizing the underlying Pareto front. We observe that qLB-JES selections stays in known cluster in later iterations, while qNEHVI continues to explore new spaces throughout the iterations.

Moreover, the selected designs shows disease relevancy. For example, IFN-omega stimulation of natrual killer (NKT) cells and mucosal-associated invariant T (MAIT) cells were selected in the last iteration of qNEHVI to be op-

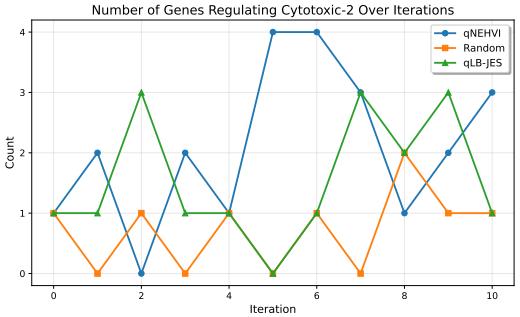

Figure 4: Number of cytokines known to regulate the Cytotoxic-2 program in MS disease in selected designs at each iteration. Color represents selection made by an acquisition function.

timized solutions that capture MS disease signatures (Figure 3f). Concordantly, IFN-omega was found to regulate a Cytotoxic-2 cytokine-induced immune program (CIP)Oesinghaus et al. (2025), which is enriched in MS disease samples.

We additionally quantified the count of cytokines that are known to regulate the Cytotoxic-2 program in the designs selected by each acquisition function at each iteration (Figure 4). We found that designs selected at random mostly contained no or one of the 6 known cytokines involved in the Cytotoxic-2 program, whereas the designs selected by qNEHVI contained more of the known cytokines. In particular, qNEHVI-selected designs contained the highest number of known cytokines in the last iteration.

4 DISCUSSION

In this work, we proposed a strategy for representing the multi-axis design space of single-cell perturbation atlases. Using this formulation, we constructed a complex design space for the Human Cytokine Dictionary and computed perturbation response–specific objectives, including response magnitude, enrichment of biological process, and cellular state signatures. We show that DGSM quantifies the sensitivity of perturbation response magnitude and ISG expression with respect to each design axis, enabling assessment of trade-offs among axes to optimize response diversity. These sensitivity measures further guide experimental design by indicating which axes warrant increased variance in future experiments.

In the second experiment, we demonstrate that a multi-objective BO loop can iteratively select designs that jointly optimize diverse disease-associated cell states while improving the performance of the GP surrogate model. We found that qNEHVI and qLB-JES acquisition functions both select designs with balanced, high objective values and show consistent improvement over the iterations toward the underlying Pareto front. In the future, we aim to extend the multi-objective BO loop to identify optimal designs in additional scenarios (e.g. reversion of disease signature).

There are several limitations to our work. First, training GP models is memory intensive, particularly for DGSM computation, where incorporating derivative information substantially increases the size and complexity of the covariance matrix. Therefore, we had to resort to chunking the design space, which may compromise the preciseness of the sensitivity quantification. Despite this potential compromise, sensitivity analysis reproduced the original findings of the Human Cytokine Atlas, supporting DGSM as a reliable method for identifying informative axes in single-cell perturbational atlases for objective-specific experimental design.

Furthermore, in our Bayesian optimization frameworks, we have only explored GP models as surrogates, which are limited in generative capability and output flexibility. In particular, correlated objectives, which are common in biology, were not explicitly modeled in our framework but can be incorporated with a multi-task GP. Future work could also leverage Bayesian Neural networks (BNN) as an alternative surrogate model, as they have been proven to be effective in perturbation experimental design Li et al. (2025).

Third, we focused on implementing and comparing two acquisition functions in this work, while many other multi-objective acquisition functions remain unexplored. We envision future work on benchmarking and optimizing various acquisition functions, including development of specialized functions, for multi-objective optimization tasks with multi-axis design spaces for perturbation experiment.

Lastly, perturbation responses are highly context-specific. As an increasing number of perturbation atlases become available, there is a growing need to systematically aggregate prior knowledge and to leverage it to provide context-specific guidance for the design of new experiments.

REFERENCES

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts,

et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11): 1083–1086, 2017.

Sebastian Ament, Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Unexpected improvements to expected improvement for bayesian optimization. *Advances in Neural Information Processing Systems*, 36:20577–20612, 2023.

Syrine Belakaria, Benjamin Letham, Janardhan R Doppa, Barbara Engelhardt, Stefano Ermon, and Eytan Bakshy. Active learning for derivative-based global sensitivity analysis with gaussian processes. *Advances in Neural Information Processing Systems*, 37:53887–53917, 2024.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in neural information processing systems*, 34:2187–2200, 2021.

Daniel Dimitrov, Stefan Schrod, Martin Rohbeck, and Oliver Stegle. Interpretation, extrapolation and perturbation of single cells. *Nature Reviews Genetics*, pp. 1–22, 2026.

Rushil Gupta, Jason Hartford, and Bang Liu. Llms for bayesian optimization in scientific domains: Are we there yet? *arXiv preprint arXiv:2509.21403*, 2025.

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*, pp. 17–37. Springer, 2024.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, pp. 2025–04, 2025.

Sergey Kucherenko and Bertrand Iooss. Derivative-based global sensitivity measures. In *Handbook of uncertainty quantification*, pp. 1241–1263. Springer, 2017.

Yanke Li, Tianyu Cui, Tommaso Mansi, Mangal Prakash, and Rui Liao. Biobo: Biology-informed bayesian optimization for perturbation design. *arXiv preprint arXiv:2509.19988*, 2025.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Arash Mehrjou, Ashkan Soleymani, Andrew Jesson, Pascal Notin, Yarin Gal, Stefan Bauer, and Patrick Schwab. Genedisco: A benchmark for experimental design in drug discovery. *arXiv preprint arXiv:2110.11875*, 2021.

Lukas Oesinghaus, Sören Becker, Larsen Vornholz, Efthymia Papalexi, Joey Pangallo, Amir Ali Moinfar, Jenni Liu, Alyssa La Fleur, Maiia Shulman, Simone Marrujo, Bryan Hariadi, Crina Curca, Alexa Suyama, Maria Nigos, Oliver Sanderson, Hoai Nguyen, Vuong K. Tran, Ajay A. Sapre, Olivia Kaplan, Sarah Schroeder, Alec Salvino, Guillermo Gallareta-Olivares, Ryan Koehler, Gary Geiss, Alexander B. Rosenberg, Charles M. Roco, Georg Seelig, and Fabian J. Theis. A single-cell cytokine dictionary of human peripheral blood. *bioRxiv*, 2025. doi: 10.64898/2025.12.12.693897. URL https://www.biorxiv.org/content/early/2025/12/15/2025.12.12.693897.

Leonard Papenmeier, Matthias Poloczek, and Luigi Nardi. Understanding high-dimensional bayesian optimization. *arXiv preprint arXiv:2502.09198*, 2025.

Parse Biosciences. Parse biosciences and graph therapeutics partner to build large functional immune perturbation atlas, jan 2026. URL https://www.parsebiosciences.com/news/parse-biosciences-and-graph-therapeutics-partner-to-build-large-functional-immune- Accessed: 2026-02-01.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024.

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. *arXiv preprint arXiv:2405.17631*, 2024.

David Schafflick, Chenling A Xu, Maike Hartlehnert, Michael Cole, Andreas Schulte-Mecklenbeck, Tobias Lautwein, Jolien Wolbert, Michael Heming, Sven G Meuth, Tanja Kuhlmann, et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature communications*, 11(1):247, 2020.

John W Schoggins and Charles M Rice. Interferon-stimulated genes and their antiviral effector functions. *Current opinion in virology*, 1(6):519–525, 2011.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Tahoe Therapeutics. Tahoe therapeutics, arc institute, and biohub partner to generate the largest perturbation dataset for virtual cell models, jan 2026. URL `https://www.tahoebio.ai/news/tahoe-arc-and-biohub-partnership`. Accessed: 2026-02-01.

Ben Tu, Axel Gandy, Nikolas Kantas, and Behrang Shafei. Joint entropy search for multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 35:9922–9938, 2022.

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

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pp. 2025–02, 2025.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

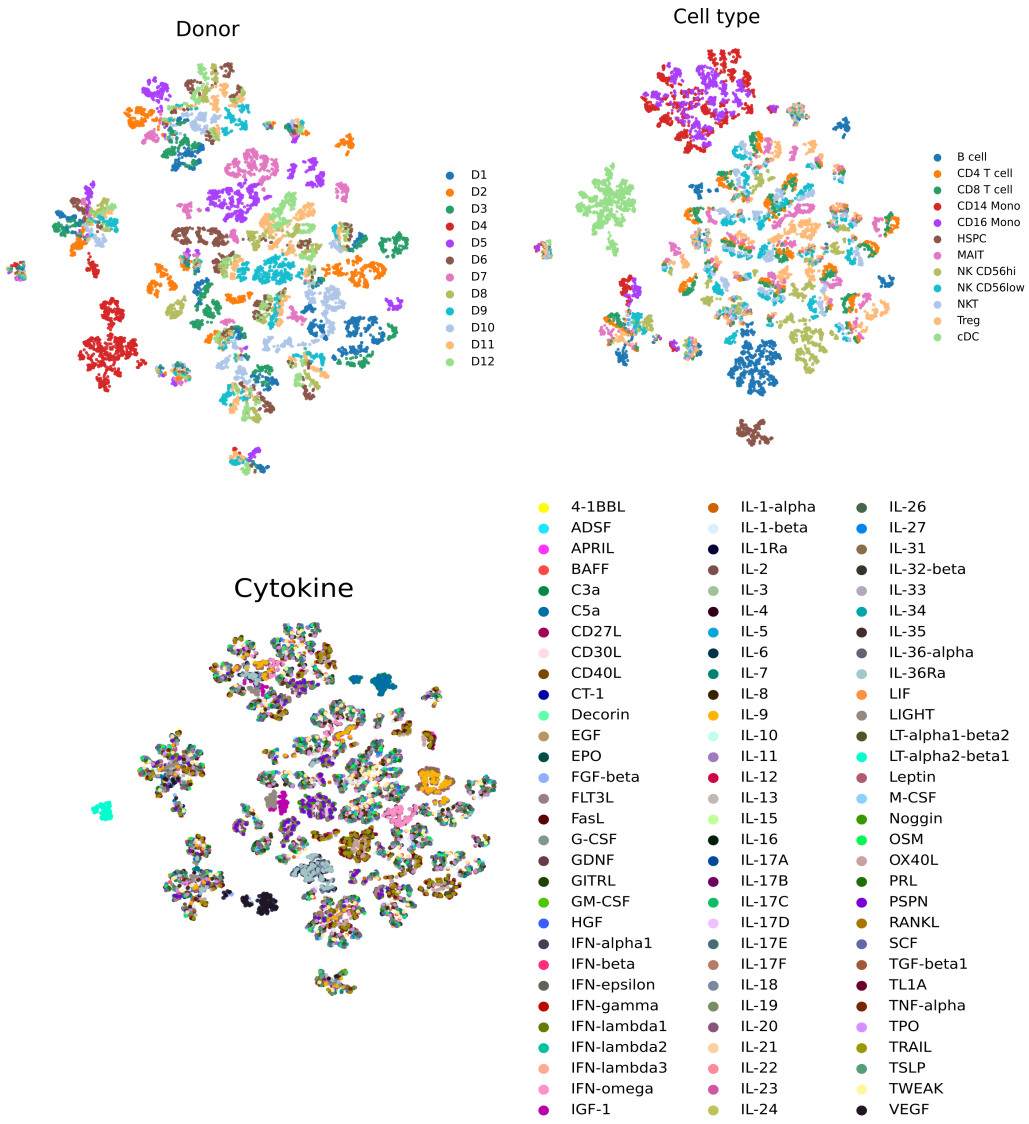

Figure S1: UMAP of the embedding-based, multi-axis design space