WEAKLY SUPERVISED LATENT VARIABLE INFERENCE OF PROXIMITY BIAS IN CRISPR GENE KNOCKOUTS FROM SINGLE-CELL IMAGES

Aditya Ravuri *†
University of Cambridge

Kristina Ulicna *
Valence Labs

Jana Osea †
University of British Columbia

Konstantin Donhauser †
ETH Zürich

Jason Hartford
Valence Labs, University of Manchester

* Equal contribution. † Work performed while the author was interning at Valence Labs.
Correspondence: aditya.ravuri@cl.cam.ac.uk & kristina@valencelabs.com

ABSTRACT

High-throughput screening enables biologists to study cell perturbations by generating large, high-dimensional datasets, such as gene expression profiles and cell microscopy images. Particularly in CRISPR-Cas9 screens, where gene knockout effects are typically represented using perturbation-specific conditional mean embeddings, these representations can be distorted by off-target effects in which the knockouts impact not only the target gene but also neighboring genes on the same chromosome arm, introducing "proximity bias". To address this, we develop a discrete latent variable inference method that leverages correlations between neighboring perturbations as a weak supervision signal to detect single cells affected by off-target effects. Removing these cells reduces spurious correlations between adjacent gene embeddings, achieving comparable correction performance without relying on additional gene expression data. Moreover, we show that the identified cells exhibit chromosome-arm specificity, reinforcing the validity of our approach and its potential for scaling into a genome-wide proximity bias correction method.

1 INTRODUCTION

High-throughput screening techniques, such as CRISPR-Cas9-based gene knockouts, enable researchers to study gene functions at scale by summarizing the effects of perturbations through high-dimensional embeddings (Chandrasekaran et al., 2023; Fay et al., 2023). However, a pervasive challenge in these experiments is the presence of proximity bias (PB), *i.e.* an unwanted correlation of embedding representations of neighboring (genomically-proximal) genes on the same chromosome arm, caused by off-target effects such as chromosomal truncations (Lazar et al., 2024). These correlations can bias downstream analyses because they make the effect of knocking out genes on the same chromosome arm appear phenotypically similar, even when they produce distinct phenotypes.

Previously, Lazar et al., 2024 showed that we can mitigate the effects of PB by appropriately transforming the aggregated embeddings to remove the bias. However, it is unclear whether such corrections also remove perturbation-specific features, thereby diminishing the biologically-relevant signal present in the samples. The correction proposed by Lazar et al. also relies on information from the gene transcriptional level to subtract an average embedding of the arm-specific unexpressed genes. To get this information, we need additional experiments for each cell type.

To overcome these limitations, we propose a novel filtering approach that directly identifies and removes PB-affected single-cell embeddings before aggregation. Our method frames PB correction as a weakly supervised latent variable inference problem. We train a classifier to distinguish "off-target" cells from those with only on-target effects. The class labels are a latent variable—it is impossible to collect labeled data as even human experts are unable to distinguish proximity biased cells from regular peturbed cells—but we can use the measurable increase in *intra*-arm cosine

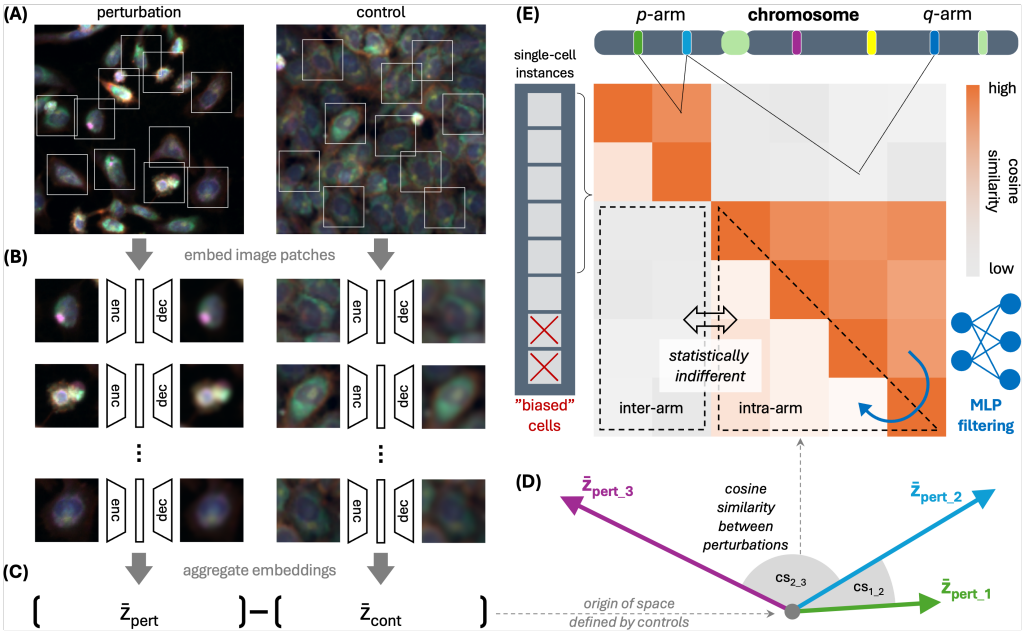

Figure 1: **Graphical abstract** highlighting the image analysis pipeline starting from **(A)** multicell microscopy images of cells under genetic perturbations and corresponding controls, from which **(B)** single-cell patches are cropped and embedded using an image autoencoder. The individual single-cell embeddings are then **(C)** aggregated into representative perturbation embeddings, which are centered relative to control embeddings. **(D)** Cosine similarities between the embeddings capture their relative (dis-)similarities, which are depicted as **(E)** pair-wise similarity matrix between individual genes on a single chromosome. Our intention is to reduce the spuriously high cosine similarities between embeddings of genes co-located on the same chromosome arm (upper triangle) by selectively identifying the "biased" cells and excluding them from the aggregation step. Tossing these off-target cells (leftmost bar) should lower the cosine similarities of gene pairs in an *intra*-arm regions to the levels statistically indistinguishable from the *inter*-arm regions (lower triangle).

similarity caused by PB as a source of weak supervision to supervise the classifier (Figure 1). By selectively excluding PB-affected cells, we ensure that the aggregated embeddings more accurately represents the on-target effects, improving the accuracy of gene function studies by eliminating the need for risky post hoc corrections, while also establishing a foundation for analyzing complex perturbation experiments.

2 Related Work

In many large scale genetic perturbation screens (Chandrasekaran et al., 2023; Fay et al., 2023), biologists knockout genes using the CRISPR-Cas9 gene editing system and measure the resulting effect. The measurements that we focus on in this paper are microscopy images collected using Cell Painting assays (Bray et al., 2016)—essentially images of cells stained with fluorescent dyes to highlight different parts of the cell—but the methods are generic and could be applied to any single-cell assay that is affected by this bias. In these screens, each perturbation is represented as the average of some embedding of the images, where the average is taken uniformly across multiple crops of the images and experimental replicates. There are many potential choices for the embedding function—common choices are hand crafted features (Stirling et al., 2021; Carpenter et al., 2006), the final hidden layer of a classifier (Sypetkowski et al., 2023) or masked autoencoders (He et al., 2021; Kraus et al., 2024; Kenyon-Dean et al., 2024)—but our work does not assume a specific choice of embedding type.

When cells are perturbed using CRISPR-Cas9 guides targeting a specific gene, it is estimated that approximately 5-10% of the time a chromosomal truncation occurs. This truncation biases the embeddings because the associated phenotypes of such cells end up more similar to other gene knock-

outs within the same chromosome arm. If we represent single cells by an embedding $\mathbf{x}$ and assume that $\mathbf{x}$ is sampled independently of other cells, we can model this data generating process as sampling cells from a mixture distribution with two components, where $\approx 95\%$ of the time we get embeddings from cells that react as expected to the perturbation, and otherwise the embeddings are biased by the chromosome-specific proximity bias phenotype that is shared across all genes within a chromosome arm.

Proximity bias detection and correction Proximity bias (PB) has been previously detected in various publicly released datasets, cell lines and CRISPR delivery protocols. These include the *RxRx3* dataset from Recursion (Fay et al., 2023), a CRISPR-Cas9 screen in human unbillical vein epithelial cells (HUVECs), the *cpg0016* dataset from Joint Undertaking in Morphological Profiling-Cell Painting (JUMP-CP) consortium (Chandrasekaran et al., 2023), a CRISPR-Cas9 screen in the U2OS osteosarcoma cell line, and the Cancer Dependency Map (DepMap), consisting of both CRISPR-Cas9 and shRNA screens across hundreds of cancerous cell lines (Tsherniak et al., 2017).

The most effective method to remove largely localised chromosome arm-specific PB signal to date is to perform a *geometric correction*, *i.e.* to adjust the features for each gene knockout by subtracting an estimate of the average features of unexpressed genes from the chromosome arm on which the gene is located (Lazar et al., 2024). While this approach effectively reduces PB, it has notable limitations: it requires (i) access to gene expression data and the expertise to identify expressed and unexpressed genes, which can be costly and challenging to scale for large screens; and (ii) enough samples of unexpressed genes knockouts to form an accurate estimate of their average features. Furthermore, it is unclear whether the geometric correction also leads inadvertent loss of perturbation-specific signal. Instead of post-processing embeddings, in this paper we aim to filter embeddings of biased cells before any aggregation of the embeddings, thereby avoiding any loss of biological signal.

3 METHODS

Our main objectives are to reduce proximity bias—hereafter defined as the average cosine similarity (alternatively, correlations) between all gene embeddings that share the same chromosome arm. In order to remove the proximity bias, we need a method that averages over *only* the unbiased embeddings from our mixture distribution. This is challenging because (1) we do not have labels that distinguish biased and unbiased embeddings, (2) while the bias is measurable (see Figure 2), the morphological changes are extremely subtle, so much so that human experts are unable to visually tell apart the biased and unbiased cells. As a result, we have a latent variable inference problem, but one in which we cannot rely on clustering to separate the latent classes because the effects are too subtle to distinguish from regular morphological variation.

Algorithm 1 Identifying single-cell embeddings to minimize average intra-arm cosine similarity

1: **Input:**
Raw single cell embeddings $\mathbf{X} \in \mathbb{R}^{n \times d}$
Batch-corrected embeddings $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$.
Gene set $\mathcal{G}$ for chromosome N and arm a.
Classifier $f : \mathbb{R}^d \to \{-1, 1\}$, with params θ.
2: **Initialize:** Set initial weights θ_0 and $\mathbf{w} = 1$.
3: **while** not converged **do**
4: Classify embeddings: $\mathbf{w} = f(\mathbf{X})$.
5: Remove proportion of cells under percentile p with lowest scores: $\hat{\mathbf{X}} = \hat{\mathbf{X}} \odot \mathbf{w}$.
6: Group $\hat{\mathbf{X}}$ by genes in $\mathcal{G}$:
$\forall g \in \mathcal{G} : \bar{\mathbf{X}}_g = \frac{1}{|\{i:g_i=g\}|} \sum_{\{i:g_i=g\}} \hat{\mathbf{X}}_i$.
7: Calculate correlation as cosine similarity matrix:
$\mathcal{P}(\mathbf{w}) = \text{cossim}\left(\{\bar{\mathbf{X}}_g : g \in G\}\right)$.
8: Set objective as mean intra-arm cosine similarity:
$\mathcal{L} = \text{mean}(\mathcal{P}(\mathbf{w}))$.
9: Update θ by gradient step:
$\theta \leftarrow \theta - \eta \nabla \mathcal{L}(\theta)$.
10: **end while**
11: **Output:** Optimized classifier f and binary vector $\mathbf{w}$, indicating retained cell embeddings.

Our approach is to leverage weak supervision from the matrices of cosine similarities between aggregated representations of adjacent genes located on a single chromosome (Figure 1). Here, gene representations $\bar{\mathbf{x}}$ are computed across a multi-aggregation step of single-cell embeddings $\mathbf{x}$, which can be modelled as a mixture distribution with two components: *unbiased embeddings*, which constitute the majority of the mixture, and *biased embeddings*, i.e. a small proportion of the embeddings which artificially result in the aggregated embeddings to have higher than expected cosine similarity with

adjacent genes. Our objective is to (i) identify these biased embeddings and (ii) remove them from the aggregation step, thus reducing the similarity (alternatively, correlation) between adjacent gene representations on the chromosome arm level.

We fit a classifier that, given a set of single-cell embeddings for every well, outputs a score for each cell that is used as a proxy for the probability that the cell belongs to the biased distribution. Using these scores as relative well-level cell ranks, we exclude the p percentile of the score distribution from each well, and re-calculate the cosine similarity for a respective gene omitting the embeddings from these cell instances. That is, the classifier is optimized to score each cell such that the subset of lowest scoring cells minimizes the cosine similarity of the embedding to the embedding of the adjacent genes.

Formally, we work with a matrix of embeddings $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$, which contains batch-corrected d-dimensional embeddings of n single cells, each of which is affected by a gene perturbation g. The batch correction step includes a centering and scaling, meaning that the feature-wise means are exactly zero across cells. Our algorithm, detailed in Algorithm 1, uses a classifier $f_\theta : \mathbb{R}^d \to (-1, 1)$, parametrised by θ, constrained such that, in every well, a fraction of p cell embeddings are chosen to be removed (i.e. set to zero). After filtering, we average embeddings by gene, and compute a gene-by-gene cosine-similarity matrix $\mathcal{P}$ on these embeddings, ordered by position on chromosome. See Figures 2, 5 and 7 for an illustration.

The challenge of training f_θ is that we cannot backpropogate through the discrete filtering step in order to minimize the average cosine-similarity of the gene-by-gene cosine-similarity matrix $\mathcal{P}$ with respect to θ. In practice, we avoid this using a soft filtering during training. For every well we sample m cells from which we aim to remove $q = \lfloor m \times p \rfloor$ cells, and average over the rest to compute our filtered embeddings. From the original m cells, we construct a set of m single-cell embeddings $\mathbf{X}$, which a neural network maps to latent scores $\mathbf{Z} \in (0, 1)^{m \times q}$ for every cell. We then calculate a soft-selction vector $\mathbf{w}$ as, $\mathbf{w} = 1 - \min\left[\sum_j \sigma\left(\mathcal{T}\mathbf{Z}_{\cdot,j}\right), 1\right]$, where σ corresponds to the softmax function, applied column-wise (i.e. across cells) and $\mathcal{T}$ is a temperature hyperparameter. This construction leads to a near-discrete vector when the temperature $\mathcal{T}$ is high, and at most q elements being 0, as desired.

We chose this construction over sampling using the Gumbel-softmax trick (the concrete distribution, (Jang et al., 2017; Maddison et al., 2017)) because we found that it led to lower variance gradience which made the objective easier to optimize. Additionally, we found that averaging over relatively few cells (in practice we set m to be 30 cells) helps us get lower variance estimates of $\mathcal{P}$ because we could load larger batches of genes into memory, leading to much more stable optimisation.

Our objective is inspired both from a biological standpoint (as the inter-arm correlations are unexpected if not for unintended effects of using CRISPR), and a probabilistic perspective as, in a Gaussian mixture model (with the mixtures corresponding to cells affected by proximity bias, and those that are not), with centered component means, the average correlation between component means naturally arises as an objective to be minimised (see Appendix A for more detail).

4 Experimental setup

Single-cell embedding dataset We used the curated version of the RxRx3 dataset (Fay et al., 2023), containing $\approx$ 2.2 million images of cells perturbed with single-gene CRISPR knockouts, spanning 15,133 genes located on 22 somatic and 2 sex chromosomes from human unbillical vein endothelial cells (HUVEC). Each well was imaged using a modified *cell painting* protocol (Bray et al., 2016) with six fluorescent staining channels. Each well image in the dataset was segmented using the MAHOTAS software (Coelho, 2012) to detect individual cell nuclei and crop a 64×64 pixel patch around each segmented nuclear centroid (typically $\approx 500 - 700$ cells per well) located at least 256 pixels away from the image edge (Figure 1A). We embedded each single-cell crop using a masked autoencoder (He et al., 2021; Kraus et al., 2024) gigantic model (MAE-G), trained on image crop of 256×256 pixels and finetuned on image crops of variable size, including 64×64 pixels. Each single-cell image instance is represented as a uniform average of token embeddings of dimensionality $d = 1,664$ (Figure 1B).

Gene relationship heatmap To construct a relationship heatmap between individual genes, we first computed a representative embedding for each gene in the dataset by averaging the single-cell embeddings by well, plate and experiment. All single-cell embeddings were centered by PCA on a set of perturbation controls, aligned using center-scaling by the mean and whitened by the standard deviation of the PCA-transformed control representations (Figure 1C) (Celik et al., 2024).

To compute the pair-wise relationships between individual genes in the genome, we first organized the genes sequentially along single chromosomes (gene index, Figure 1E). Cosine similarity was computed to quantify the positive (similar) or negative (opposite) relationship between aggregated embeddings as $\text{CS}(\mathbf{A}, \mathbf{B}) = (\mathbf{A} \cdot \mathbf{B})/(\|\mathbf{A}\|\|\mathbf{B}\|)$, where $\mathbf{A}$ and $\mathbf{B}$ are the representative aggregated embeddings of gene a and gene b, respectively. The similarities are displayed as chromosome heatmaps after a quantile transformation to a normal distribution with $\mu = 0$ and $\sigma = 0.2$ to make the heatmaps visually comparable between different chromosomes as well as between pre- and post-filtering step (Figure 1E).

Labeling of single-cell identities At train time, 30 cells per each gene perturbation well were randomly chosen. This number was selected as a trade-off in accordance to the statistical rule of thumb of the minimum amount of data one needs for hypothesis tests. Aggregating as few as 30 cells was a sufficient amount of data per well for the cosine similarity heatmaps to clearly display expected proximity bias patterns (§ 4). Our main consideration was to fit all wells corresponding to 100-300 genes (which make up an optimizer's batch) into GPU memory for stable optimization, and any fewer risked optimization instability due to high variability of the objective.

At inference time, the trained model classifies all available raw single-cell embeddings per gene perturbation ($\approx 50,000$ single-cell instances per gene). The classifier labels each cell embedding as either "real perturbation" or "proximity bias", where the latter are excluded from the embedding aggregation for heatmap construction (§ 4). We trained a separate proximity bias (PB) classifier on genes located across 22 somatic and 2 sex chromosomes. Based on prior estimates of proximity bias incidence in the RxRx3 dataset (Lazar et al., 2024), with an upper-bound estimate of 15%, we selected two models for subsequent evaluation: a "weak" filtering model, trained to exclude 3 cells out of a batch of 30 (10%), and a "strong" filter, trained to discard 6 out of 30 cells (20%).

5 RESULTS

In this section, we will (i) present a qualitative and quantitative evaluation of our single-cell filtering method in the context of individual chromosomes. To do so, we will (i) show that our filtering approach reduces average correlation (cosine similarity) in *intra*-arm regions compared to the *inter*-arm regions compared to a non-filtering baseline (§ 5.1), (ii) contrast our single-cell approach performance to other proposed corrections strategies (§ 5.2) and (iii) demonstrate that our classifiers learn single-cell characteristics specific to individual chromosomes (§ 5.3).

5.1 CELL FILTERING REDUCES SPURIOUS EMBEDDING CORRELATIONS

Qualitative chromosome heatmap evaluation To assess model performance, we constructed single-chromosome heatmaps (§ 4) representing pairwise cosine similarities between gene knock-

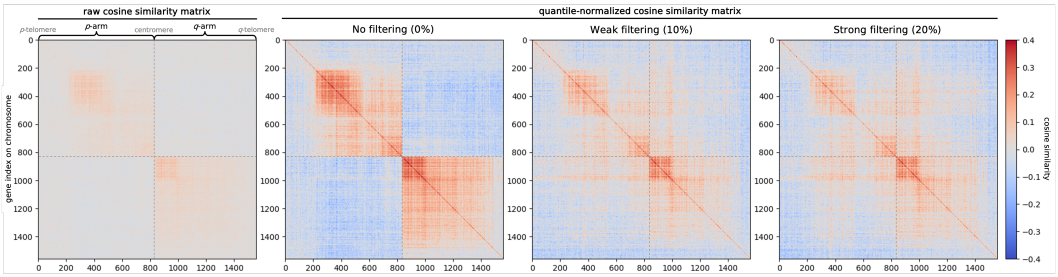

Figure 2: **Visualization of gene cosine similarities** on chromosome 1 as raw and quantile-normalized cosine similarity matrices before and after cell filtering with weak (10% cell toss) and strong (20% cell toss) models.

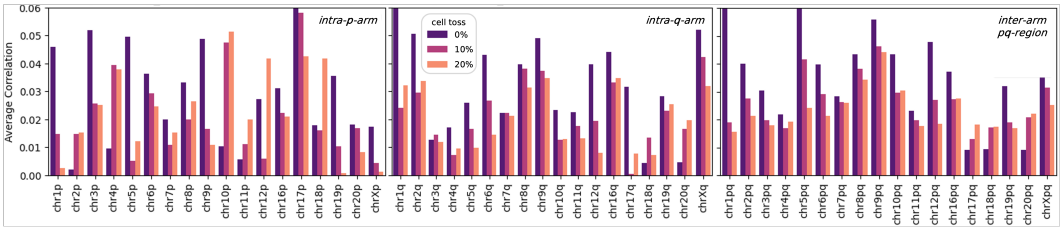

Figure 3: **Average correlation reduction** across *intra*- as well as *inter*-arm regions (not trained for) in all chromosomes with at least 10 genes on both arms, pre- and post-filtering with our weak and strong models.

outs before and after filtering (Figure 2). While *inter*-arm gene pairs are not inherently dissimilar, the *intra*-arm gene pairs exhibit a higher similarities consistent across both arms.

This relative contrast is further amplified by the quantile transformation of the similarity matrix, which enhances the visualization of the proximity bias effect. Here, genes located on the same chromosome arm exhibit consistently higher similarity to one another compared to those on opposite arms (Figure 2). These elevated similarity values exceed those expected by chance, despite the absence of functional or morphological justification, as adjacent genes do not universally share functional similarity. Importantly, our visualizations demonstrate that the proximity bias effect remains detectable even when image resolution is reduced to the single-cell level.

Furthermore, a qualitative evaluation suggests that both the 10% and 20% filtering models effectively mitigate proximity bias. This reduction appears to be driven by a decrease in *intra*-arm correlations, thereby enabling previously dissimilar *inter*-arm gene pairs to exhibit greater relative similarity. Consequently, this redistribution of similarity values leads to an apparent increase in *inter*-arm similarity within the heatmap visualizations (Figure 2).

Quantitative reduction of embedding correlations To quantitatively assess this observation, we compared the average correlations within *intra*-arm regions (*i.e.*, within the *p*- and *q*-arms) to those in the *inter*-arm region (*i.e.*, between *p*- and *q*-arms) across all chromosomes with two arms (Figure 3). Our results indicate that at least one of the filtering models reduces the average correlation in 14 (77%) and 16 (88%) out of 18 *intra p*- and *q*-arm regions, respectively, compared to the non-filtered baseline. Strikingly, 15 (83%) out of 18 chromosome *pq*-regions show a consistent correlation drop, despite that reducing *inter*-arm cosine similarities was not included in the training objective (§ 3).

While these findings demonstrate a measurable reduction in proximity bias, it remains unclear whether the average correlation can be further minimized. Variability in arm-specific cosine similarity across chromosome heatmaps may arise due to differences in chromosome arm lengths or the specific filtering model applied. Additionally, average correlation may not represent the most consistent or optimal evaluation metric (as discussed in § 6), suggesting that further refinements are necessary to fully mitigate the bias.

5.2 Method benchmarking to other correction strategies

Encouraged by the success of our single-cell filtering method in mitigating spurious *intra*-arm correlations, we compared its performance against previously reported proximity bias correction strategies (Lazar et al., 2024). This comparison is inherently challenging due to methodological differences: prior work used larger (256×256 pixel) image crops containing multiple cells and a distinct embedding model for feature extraction (Sypetkowski et al., 2023). To enable a fair evaluation, we contrasted our filtering approach with two geometric chromosome-arm corrections, all performed on a single-cell image embedding level: (i) a *naïve* method, subtracting an average embedding of all genes on the chromosome arm from each gene representation, and (ii) an *expression-based* method adapted from (Lazar et al., 2024), subtracting an average embedding derived solely from *unexpressed* genes (zFPKM < -3.0 in bulk RNA-seq of untreated, control wells).

All three approaches successfully reduce *intra*-arm correlations compared to the unfiltered case (Figure 4). However, while their numerical performance appears similar for chromosome 1, closer

heatmap examination reveals key differences, which underscores the importance of visual inspection alongside quantitative metrics.

The *naïve* approach reduces *p*-arm similarity but introduces artificial similarity in previously dissimilar regions, particularly near the *p*-telomere (Figure 4). This effect arises because subtracting a universal embedding, irrespective of gene position, inadvertently increases adjacency-based similarity. The *targeted* approach mitigates this issue to some extent but still induces artificial correlations in the chr1p telomeric region and fails to fully resolve residual similarities in chr1q.

While our strategy does not completely eliminate the chr1p central square, it reduces its prominence more effectively than the alternative approaches while avoiding the unintended introduction of artificial correlations elsewhere. This suggests that our method provides the most robust correction among the tested strategies, without the need to obtain expensive gene expression profiles in the studied cell line. However, a more comprehensive statistical analysis may be required for a definitive and rigorous comparison.

5.3 Cell filtering beyond chromosome-specific contexts

To determine whether our filtering models effectively exclude proximally biased cells rather than simply removing imaging artifacts or debris-like objects, we evaluated their ability to generalize across different chromosome contexts. Specifically, we tested whether a model trained to reduce *intra*-arm correlations on one chromosome could achieve similar results when applied to another. This approach is based on the premise that each chromosome arm consists of a distinct set of genes, and knocking out this specific set of genes would produce arm-specific morphological phenotypes. If a model successfully corrects bias only on the chromosome it was trained on but fails on others, it suggests that the filtering process is correctly targeting proximity-biased cells. Conversely, if bias is reduced across multiple chromosomes, this would indicate that the excluded cells share genome-wide phenotypic characteristics, which would undermine the interpretability of our approach.

We focused on the two chromosomes with the highest gene counts in our dataset—chromosome 1 and 17. While models applied to their training chromosomes effectively reduced the spurious correlations, exchanging them had the opposite effect—correlations not only failed to decrease but increased compared to pre-filtering step, suggesting that the model exchange exacerbates the undesired similarity pattern (Figure 5). This validates our approach to proximity bias correction targeted on a chromosome- or even arm-specific approach, as embedding feature distributions differ between chromosomes. These findings provide indirect evidence that our models correctly identify and filter proximity-biased cells, aligning with prior observations (Lazar et al., 2024).

6 Discussion and conclusions

This work demonstrates how we can use the bias that arises from off-target effects of CRISPR-Cas9 as a source of weak supervision to infer the latent cell state (i.e. on- or off-target effect). We use the inferred state to reduce spurious correlations between single-gene perturbations, such as the ones arising from proximity bias in chromosome-specific contexts. We found that the approach reduced

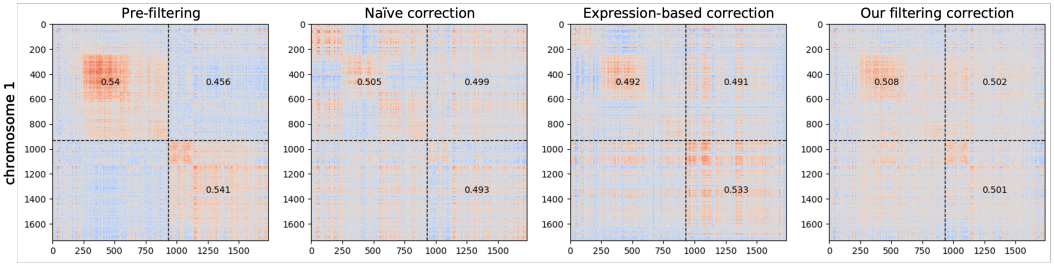

Figure 4: **Comparison of correction approaches** on chromosome 1 of the pre-filtering baseline to the *(naïve)* and expression-guided corrections and our filtering approach which visually most accurately lowers the correlation signal with minimal introduction of artificial patterns into the heatmap.

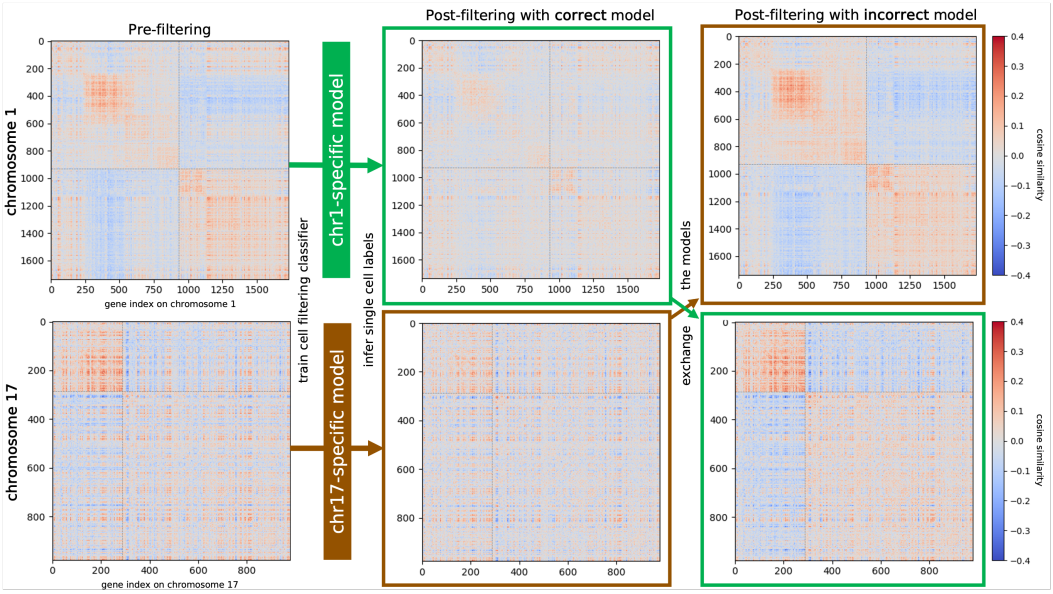

Figure 5: **Reduction of cell embedding correlations requires chromosome-specific models.** Heatmaps for gene perturbations on chromosomes 1 and 17 are shown before filtering (left), after filtering with a chromosome-specific model (middle), and after filtering with a model trained on the other chromosome (right). Exchanging models fails to reduce and amplifies proximity bias, emphasizing the need for arm-specific models.

the average *intra*-arm correlations for each chromosome and reduced some of the artifacts that arise from naïve or expression-based corrections, though performance varied across the chromosomes. While promising, our analysis is currently limited to comparisons on a per-chromosome basis; we are currently working on scaling the approach to a genome-wide study which will require us to correct for any systemic differences between the per-chromosome distribution of cosine similarities.

Likewise, reducing average correlations does not necessarily guarantee that our approach preserves known biological relationships as effectively as the baseline. To evaluate this, measuring the recall of gene pairs with established interactions from public benchmarking databases of gene-gene relationships would strengthen our argument that the arm correlation reduction also leads to successful proximity bias reduction. Additionally, implementing these benchmarks is needed not only to compare pre- and post-filtering results within the single-cell approach but also to assess how it performs relative to the previously introduced multi-cell approaches (Lazar et al., 2024; Kraus et al., 2024).

From an interpretability standpoint, our method provides single-cell labels indicating whether an embedding represents a true perturbation, enabling a detailed characterization of the subpopulation of proximally biased cells. This could be performed in the image space, facilitating feature dissection in the latent embedding. Additionally, integrating these latent labels with hand-crafted feature sets from CellProfiler (Carpenter et al., 2006), designed by human experts to enhance image explainability, could further illuminate the characteristics and origins of proximity bias. This, in turn, may lead to more targeted strategies for its detection and mitigation.

In summary, this work introduces a novel approach to address proximity bias at the single-cell level, being the first to utilize weakly supervised latent variable inference to filter off-target single-cell instances from the image embedding space. Given the scarcity of label-free methods for disentangling biological heterogeneity, such as separating cell population mixtures, we envision extending our approach to more complex scenarios, including dual- or multi-gene CRISPR-Cas9 knockouts and gene-compound combination screens, with potential implications for therapeutic applications.

ACKNOWLEDGEMENTS

We would like to thank Saber Saberian, Nathan H. Lazar, Imran S. Haque, and Berton Earnshaw for their invaluable feedback and thoughtful insights on shaping this manuscript.

MEANINGFULNESS STATEMENT

To understand the effect of genetic perturbations in phenomics screens, we must convert images to numerical representations that are suitable for downstream processing. Often, these are aggregated featurizations of a set of images, potentially transformed in accordance with causal principles. It has been noticed in previous work that such numerical representations of what a gene deletion does, within CRISPR-based phenomics screens, are confounded by a mechanism thought to be chromosomal truncation. We offer a method to correct for these effects, paving a path to clearer representations of genetic deletions, and hence cell biology, a core component of understanding life.

REFERENCES

Mark A. Bray, Shantanu Singh, Han Han, Claire T. Davis, Brittany Borgeson, Christopher Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Craig C. Gibson, and Anne E. Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, Sep 2016. doi: 10.1038/nprot.2016.105. Epub 2016 Aug 25.

Anne E. Carpenter, Thouis R. Jones, Michael R. Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A. Guertin, Joo Han Chang, Robert A. Lindquist, Jason Moffat, Polina Golland, and David M. Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, Oct 2006. ISSN 1474-760X. doi: 10.1186/gb-2006-7-10-r100. URL https://doi.org/10.1186/gb-2006-7-10-r100.

Seda Celik, Jens-Christian Hütter, Silvia M. Carlos, Nicholas H. Lazar, Ritika Mohan, Claire Tillinghast, et al. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *PLOS Computational Biology*, 20(10):e1012463, 2024. doi: 10.1371/journal.pcbi.1012463. URL https://doi.org/10.1371/journal.pcbi.1012463.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D. Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D. Boyd, Laurent Brino, Patrick J. Byrne, Hugo Ceulemans, Carolyn Ch'ng, Beth A. Cimini, Djork-Arne Clevert, Nicole Deflaux, John G Doench, Thierry Dorval, Regis Doyonnas, Vincenza Dragone, Ola Engkvist, Patrick W. Faloon, Briana Fritchman, Florian Fuchs, Sakshi Garg, Tamara J. Gilbert, David Glazer, David Gnutt, Amy Goodale, Jeremy Grignard, Judith Guenther, Yu Han, Zahra Hanifehlou, Santosh Hariharan, Desiree Hernandez, Shane R Horman, Gisela Hormel, Michael Huntley, Ilknur Icke, Makiyo Iida, Christina B. Jacob, Steffen Jaensch, Jawahar Khetan, Maria Kost-Alimova, Tomasz Krawiec, Daniel Kuhn, Charles-Hugues Lardeau, Amanda Lembke, Francis Lin, Kevin D. Little, Kenneth R. Lofstrom, Sofia Lotfi, David J. Logan, Yi Luo, Franck Madoux, Paula A. Marin Zapata, Brittany A. Marion, Glynn Martin, Nicola Jane McCarthy, Lewis Mervin, Lisa Miller, Haseeb Mohamed, Tiziana Monteverde, Elizabeth Mouchet, Barbara Nicke, Arnaud Ogier, Anne-Laure Ong, Marc Osterland, Magdalena Otrocka, Pieter J. Peeters, James Pilling, Stefan Prechtl, Chen Qian, Krzysztof Rataj, David E Root, Sylvie K. Sakata, Simon Scrace, Hajime Shimizu, David Simon, Peter Sommer, Craig Spruiell, Iffat Sumia, Susanne E Swalley, Hiroki Terauchi, Amandine Thibaudeau, Amy Unruh, Jelle Van de Waeter, Michiel Van Dyck, Carlo van Staden, Michał Warchoł, Erin Weisbart, Amélie Weiss, Nicolas Wiest-Daessle, Guy Williams, Shan Yu, Bolek Zapiec, Marek Żyła, Shantanu Singh, and Anne E. Carpenter. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023. doi: 10.1101/2023.03.23.534023. URL https://www.biorxiv.org/content/early/2023/03/24/2023.03.23.534023.

Luís Pedro Coelho. Mahotas: Open source software for scriptable computer vision. *CoRR*, abs/1211.4907, 2012. URL http://arxiv.org/abs/1211.4907.

Marta M. Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, Jordan Christensen, Berton A. Earnshaw, Imran S. Haque, and Ben Mabey. Rxrx3: Phenomics map of biology. *bioRxiv*, 2023. doi: 10.1101/2023.02.07.527350. URL https://www.biorxiv.org/content/early/2023/02/08/2023.02.07.527350.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Kian Kenyon-Dean, Zitong Jerry Wang, John Urbanik, Konstantin Donhauser, Jason Hartford, Saber Saberian, Nil Sahin, Ihab Bendidi, Safiye Celik, Marta Fay, Juan Sebastian Rodriguez Vera, Imran S Haque, and Oren Kraus. Vitally consistent: Scaling biological representation learning for cell microscopy, 2024. URL https://arxiv.org/abs/2411.02572.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. pp. 11757–11768, 06 2024. doi: 10.1109/CVPR52733.2024.01117.

Nathan H. Lazar, Safiye Celik, Lu Chen, Marta M. Fay, Jonathan C. Irish, James Jensen, Conor A. Tillinghast, John Urbanik, William P. Bone, Christopher C. Gibson, and Imran S. Haque. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by crispr–cas9 editing. *Nature Genetics*, 56(7):1482–1493, Jul 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01758-y. URL https://doi.org/10.1038/s41588-024-01758-y.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables, 2017. URL https://arxiv.org/abs/1611.00712.

David R. Stirling, Madison J. Swain-Bowden, Alice M. Lucas, Anne E. Carpenter, Beth A. Cimini, and Allen Goodman. Cellprofiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, 22(1):433, Sep 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04344-9. URL https://doi.org/10.1186/s12859-021-04344-9.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4285–4294, June 2023.

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, Barbara A. Weir, Gregory Kryukov, Glenn S. Cowley, Stanley Gill, William F. Harrington, Sasha Pantel, John M. Krill-Burger, Robin M. Meyers, Levi Ali, Amy Goodale, Yenarae Lee, Guozhi Jiang, Jessica Hsiao, William F.J. Gerath, Sara Howell, Erin Merkel, Mahmoud Ghandi, Levi A. Garraway, David E. Root, Todd R. Golub, Jesse S. Boehm, and William C. Hahn. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2017.06.010. URL https://www.sciencedirect.com/science/article/pii/S0092867417306517.

A A PROBABILISTIC INTERPRETATION OF OUR OBJECTIVE

This section shows that assuming a Gaussian mixture model with a centered Gaussian prior over the component means, we recover an objective that penalizes the average covariance between component means.

MODEL

Consider a setting with two genes being perturbed, a and b, located on the same chromosome arm. We assume that genetic perturbations to single cells have an additive impact on a gene's representation w.r.t. control cells, measured by $\boldsymbol{\mu}^a, \boldsymbol{\mu}^b \in \mathbb{R}^d$. With some probability, a cell is affected by a mechanism thought to be chromosomal truncation (i.e. if a cell is "proximally biased"), the effect of which is denoted by $\boldsymbol{\mu}^p$. We assume a multivariate Gaussian distribution on $\mathbf{M}$, and we assume that embeddings are centered,

$$\mathbf{M} = \begin{bmatrix} \boldsymbol{\mu}^{a\top} \\ \boldsymbol{\mu}^{b\top} \\ \boldsymbol{\mu}^{p\top} \end{bmatrix} \sim \mathcal{MN}\left(\mathbf{0}, (1+\epsilon)\mathbf{I} - \frac{1}{n_g}\mathbf{O}, \mathbf{I}_d\right).$$

The choice of covariance is just a centering matrix $\mathbf{H}$ with jitter ϵ added along the diagonal. The covariance naturally arises if we zero-center the embeddings by construction; assume a non-zeroed normal matrix $\mathbf{M}'$, centering $\mathbf{M}'$ as $\mathbf{M} = \mathbf{H}\mathbf{M}'\mathbf{H}$ leads to the row-covariance $\mathbf{H}\mathbf{I}_{n_g}\mathbf{H} = \mathbf{H}$ (as $\mathbf{H}$ is idempotent).

If the i-th cell of perturbation k is proximally biased, we represent it with $\mathbf{w}_i^k = 1$, and zero otherwise. The prior distribution on the vector $\mathbf{w} = [\mathbf{w}^a \quad \mathbf{w}^b]^T$ is such that, for every well (which contains n' cells), exactly $\lfloor p_w n' \rfloor$ cells are sampled uniformly. This ensures that the marginal probability $\mathbb{P}(\mathbf{w}_i^k = 1)$ is p_w.

The observed control-centered representations of single cells from the phenomics experiments are represented by $\mathbf{X}^a, \mathbf{X}^b \in \mathbb{R}^{n\times d}$. Assume that,

$$\mathbf{X}^a \mid \mathbf{w}^a, \boldsymbol{\mu}^a, \boldsymbol{\mu}^p \sim \mathcal{MN}\left((\mathbf{1}-\mathbf{w}^a)\otimes \boldsymbol{\mu}^{a\top} + \mathbf{w}^a \otimes \boldsymbol{\mu}^{p\top}, \sigma_m^2\mathbf{I}_n, \mathbf{I}_d\right),$$
$$\mathbf{X}^b \mid \mathbf{w}^b, \boldsymbol{\mu}^b, \boldsymbol{\mu}^p \sim \mathcal{MN}\left((\mathbf{1}-\mathbf{w}^b)\otimes \boldsymbol{\mu}^{b\top} + \mathbf{w}^b \otimes \boldsymbol{\mu}^{p\top}, \sigma_m^2\mathbf{I}_n, \mathbf{I}_d\right).$$

INFERENCE: THE OBJECTIVE

The posterior factorises as,

$$p\big(\boldsymbol{\mu}^a, \boldsymbol{\mu}^b, \boldsymbol{\mu}^p, \mathbf{w} \big| \mathbf{X}^a, \mathbf{X}^b\big) \propto p\big(\mathbf{X}\big|\mathbf{M}, \mathbf{w}\big)p\big(\mathbf{M}\big)p(\mathbf{w}).$$

We use a variational approximation,

$$q(\mathbf{w}|\mathbf{X}) = \delta(f_\theta(\mathbf{X})) \equiv \delta(\tilde{\mathbf{w}}),$$

where f_θ is a neural network parameterised such that its support matches that of the prior (i.e. a fixed number of cells per well are identified as proximally biased). The implied evidence lower bound (ELBO) is,

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}_{q(\mathbf{w},\mathbf{M}|\mathbf{X})}\left(\log p\big(\mathbf{X}\big|\mathbf{M}, \mathbf{w}\big)p(\mathbf{M})\right) - \mathrm{KL}(q(\mathbf{w})\|p(\mathbf{w}) \\ &= \log p(\mathbf{X}|\mathbf{M}, \tilde{\mathbf{w}}) + \log p(\mathbf{M}) + c.\end{aligned}$$

We use block coordinate ascent for inference (i.e. CAVI/EM).

INFERENCE: STEP 1

First, we maximise the objective w.r.t. $\mathbf{M}$, resulting approximately in the intuitive maximum likelihood estimator. W.l.o.g., consider the partial derivative of $\mathcal{L}(\theta)$ w.r.t. a specific component $\mu_k^a \in \mathbb{R}$, and let $\tilde{\mathbf{X}^a} = \mathbf{X}^a - (\mathbf{1}-\mathbf{w}^a)\otimes\boldsymbol{\mu}^{a\top} - \mathbf{w}^a\otimes\boldsymbol{\mu}^{p\top}$. First, note that,

$$p\big(\mathbf{X}\big|\mathbf{M}, \mathbf{w}\big) = p\big(\mathbf{X}^a\big|\boldsymbol{\mu}^a, \boldsymbol{\mu}^p, \mathbf{w}^a\big) \cdot p\big(\mathbf{X}^b\big|\boldsymbol{\mu}^b, \boldsymbol{\mu}^p, \mathbf{w}^b\big),$$

and that,

$$\begin{aligned}
\log p(\mathbf{M}) &= -\frac{1}{2}\text{tr}\left(\mathbf{M}\mathbf{M}^\top\left((1+\epsilon)\mathbf{I} - \frac{1}{n_g}\mathbf{O}\right)^{-1}\right) + c \\
&= -\frac{1}{2}\text{tr}\left(\mathbf{M}\mathbf{M}^\top\left(\frac{1}{1+\epsilon}\mathbf{I} + \frac{1}{\epsilon n_g(1+\epsilon)}\mathbf{O}\right)\right) + c \qquad \text{Sherman-Morrison} \\
&\approx -\frac{1}{2}\sum_{k_a}\|\boldsymbol{\mu}^{k_a}\|^2 - \frac{1}{2n_g\epsilon}\sum_{k_a,k_b}\boldsymbol{\mu}^{k_a\top}\boldsymbol{\mu}^{k_b}.
\end{aligned}$$

The derivative of the ELBO w.r.t. a component of the mean is,

$$\begin{aligned}
\frac{\partial}{\partial\mu_k^a}\mathcal{L}(\theta) &= \frac{\partial}{\partial\mu_k^a}\left[\log p(\mathbf{M}) + \log p\big(\mathbf{X}^a\big|\boldsymbol{\mu}^a, \boldsymbol{\mu}^p, \mathbf{w}^a\big)\right] \\
&= \frac{\partial}{\partial\mu_k^a}\left[\log p(\mathbf{M}) - \frac{1}{2\sigma_m^2}\text{tr}\Big[\tilde{\mathbf{X}}^a\tilde{\mathbf{X}}^{a\top}\Big] - \frac{nd}{2}\log(\sigma_m^2)\right], \\
&\propto \frac{\partial}{\partial\mu_k^a}\left[\frac{1}{\sigma_m^2}\sum_{i=1}^{n}\Big\|\mathbf{X}_i^a - \big[(1-w_i^a)\boldsymbol{\mu}^a + w_i^a\boldsymbol{\mu}^p\big]\Big\|^2\right] + \mu_k^{a\,2} + \frac{2\sum_m\mu_k^a\mu_k^m}{n_g\epsilon}.
\end{aligned}$$

Setting this derivative to 0 for the maximum-likelihood solution leads to,

$$\begin{aligned}
&2\mu_k^a + \frac{2(2\mu_k^a + \mu_k^b + \mu_k^p)}{n_g\epsilon} - \frac{1}{\sigma_m^2}\sum_{i=1}^{n}2\big[X_{ik}^a - (1-w_i)\mu_k^a - w_i\mu_k^p\big]\big[1-w_i\big] = 0, \\
\Rightarrow&\mu_k^a + \frac{\mu_k^a}{n_g\epsilon} - \frac{1}{\sigma_m^2}\sum_{i=1}^{n}\Big[X_{ik}^a(1-w_i) - (1-w_i)\mu_k^a\Big] = 0, \\
\Rightarrow&\mu_k^a + \frac{\mu_k^a}{n_g\epsilon} + \frac{n(1-p_w)\mu_k^a}{\sigma_m^2} - \frac{1}{\sigma_m^2}\sum_{i=1}^{n}\Big[X_{ik}^a(1-w_i)\Big] = 0, \\
\Rightarrow&\hat{\mu}_k^a = \frac{\sum_{i=1}^{n}\Big[X_{ik}^a(1-w_i)\Big]}{\left(\sigma_m^2 + \frac{\sigma_m^2}{n_g\epsilon} + n(1-p_w)\right)} \overset{n\to\infty}{\approx} \frac{\sum_{i=1}^{n}\Big[X_{ik}^a(1-w_i)\Big]}{n(1-p_w)} \equiv \bar{\mathbf{X}}_{:k}^{a,\not p}.
\end{aligned}$$

We see that the solution for the mean embedding of a genetic perturbation is simply the average embedding corresponding to non-proximally biased cells known to have that perturbation induced.

INFERENCE: STEP 2

In this step, we optimise the ELBO w.r.t. θ, i.e. the parameters associated with the latent variables $\mathbf{w}$. In particular, we will argue that this step leads to minimisation of the average covariance between the average gene embeddings.

Our first argument is that the optimisation of the objective $\mathcal{L}$ is dominated by the maximisation of $\log p(\hat{\mathbf{M}})$. A sketch is as follows. The first term of the ELBO without constants (w.r.t. θ) is,

$$\begin{aligned}
\mathcal{T}_1^a &\equiv -\frac{1}{2\sigma_m^2}\sum_{i=1}^{n}\left\|\mathbf{X}_i^a - (1-w_i^a)\boldsymbol{\mu}^a - w_i^a\boldsymbol{\mu}^p\right\|^2 \\
&= -\frac{1}{2\sigma_m^2}\sum_{i=1}^{n}\sum_{k=1}^{d}\left(X_{ik}^a - (1-w_i^a)\mu_k^a - w_i^a\mu_k^p\right)^2 \\
&= -\frac{1}{2\sigma_m^2}\sum_{i=1}^{n}\sum_{k=1}^{d}(1-w_i^a)\mu_k^{a\,2} + w_i^a\mu_k^{p\,2} - 2X_{ik}^a(1-w_i^a)\mu_k^a - 2X_{ik}^a w_i^a\mu_k^p + c \\
&= \frac{1}{2\sigma_m^2}\sum_{k=1}^{d} n(1-p_w)(\bar{\mathbf{X}}_{:k}^{a,\not p})^2 + np_w\bar{\mathbf{X}}_{:k}^{a,p}\frac{\bar{\mathbf{X}}_{:k}^{a,p} + \bar{\mathbf{X}}_{:k}^{b,p}}{2} + c.
\end{aligned}$$

Optimisation of this term w.r.t. θ (and therefore $\mathbf{w}$) should just lead to a balancing dynamic, as the relabelling of a cell would inversely affect $\bar{\mathbf{X}}_{:k}^{a,\not p}$ and $\bar{\mathbf{X}}_{:k}^{a,p}$.

Therefore, the optimisation of the ELBO w.r.t θ should just force $\log p(\hat{\mathbf{M}})$ upwards, which from the previous inference step is known to be inversely proportional to the average covariance between rows of $\hat{\mathbf{M}}$,

$$\log p(\hat{\mathbf{M}}) = -\frac{1}{2n_g\epsilon}\sum_{k_a,k_b}\boldsymbol{\mu}^{k_a\top}\boldsymbol{\mu}^{k_b}(1+\epsilon n_g\delta_{k_ak_b}) \qquad \square$$

A sense-check, illustrated in figure 6 verifies that average covariances are reduced by such an optimisation. We hypothesize that for our data, this proxy objective may be a better objective than the likelihood, as our data distributions are verifiably non-normal.

B ALTERNATIVE METHODS FOR PROXIMITY BIAS DETECTION

Assuming that the phenotypes of PB cells are identifiable and differ by chromosome arm, one can fit a chromosome-arm classifier, and, with enough data (number of cells, genes per arm, etc.), classify with high confidence cells affected by PB. However, if we fit a logistic regression, whose uncertainties are known to be well calibrated, and remove such cells, we see that PB is not necessarily reduced, as illustrated in Figure 7. Moreover, PB does not appear in gene-by-gene confusion matrices corresponding to classification models fit to identify which gene perturbation (the "class") has affected an embedding representing a cell.

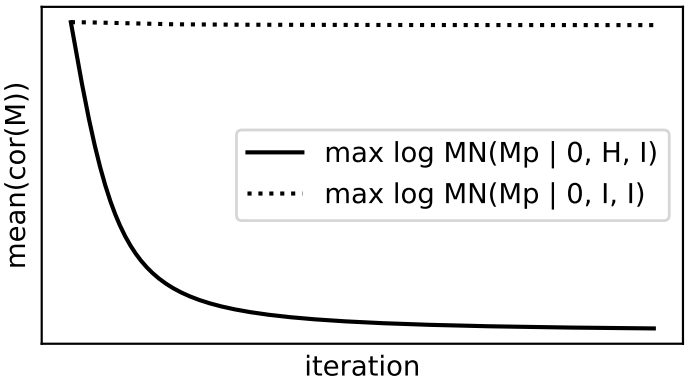

Figure 6: **A sense check** that confirms that the maximization of $\log \mathcal{MN}(\mathbf{M}'|\mathbf{0}, \mathbf{H}, \mathbf{I})$ w.r.t. $\boldsymbol{\theta}$ leads to a minimization of average off-diagonal covariance between rows of $\mathbf{M}'$, as opposed to $\log \mathcal{MN}(\mathbf{M}'|\mathbf{0}, \mathbf{I}, \mathbf{I})$, where $\mathbf{M}' = \mathbf{M} + p_w \tanh(\boldsymbol{\theta})$, and $\mathbf{M}$ is such that $\text{cor}(\mathbf{M}_i, \mathbf{M}_j) = 0.2 + 0.8\delta_{ij}$.

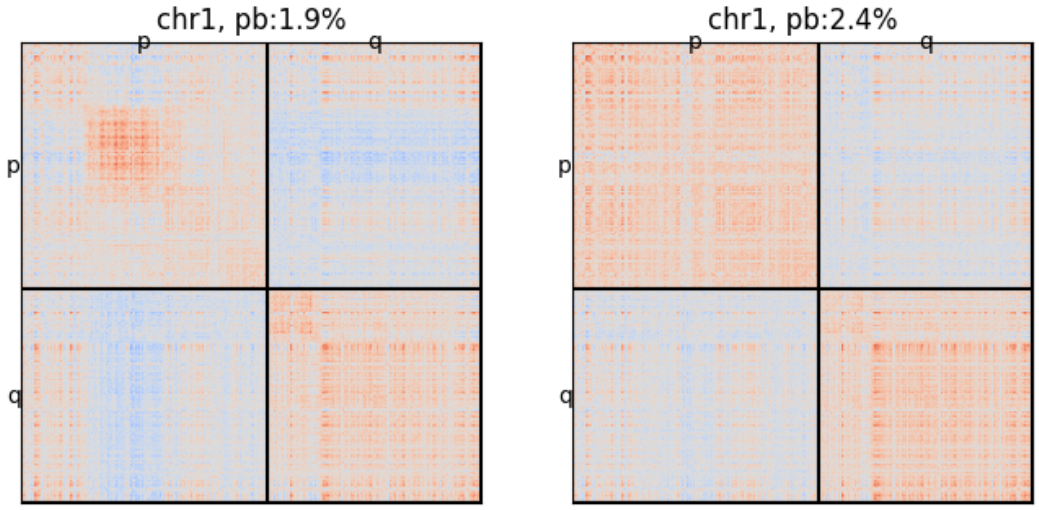

Figure 7: **Confidence-based classification correction does not reduce proximity bias.** Comparison of raw **(left)** and filtered **(right)** similarity maps of chromosome 1, where the filtered map was generated by dropping cells based on confidence scores from a chromosome arm classifier.