# PETRI: LEARNING UNIFIED CELL EMBEDDINGS FROM UNPAIRED MODALITIES VIA EARLY-FUSION JOINT RECONSTRUCTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Integrating multimodal screening data is challenging because biological signals only partially overlap and cell-level pairing is frequently unavailable. Existing approaches either require pairing or fail to capture both shared and modality-specific information in an end-to-end manner. We present PETRI, an early-fusion transformer that learns a unified cell embedding from unpaired cellular images and gene expression profiles. PETRI groups cells by shared experimental context into multimodal "documents" and performs masked joint reconstruction with cross-modal attention, permitting information sharing while preserving modality-specific capacity. The resulting latent space supports construction of perturbation-level profiles by simple averaging across modalities. Applying sparse autoencoders to the embeddings reveals learned concepts that are biologically meaningful, multimodal, and retain perturbation-specific effects. To support further machine learning research, we release a blinded, matched optical pooled screen (OPS) and Perturb-seq dataset in HepG2 cells.

## 1 INTRODUCTION

A major goal of cell biology is to establish causal links between gene activity and cellular states (Rood et al., 2024). High-throughput perturbation technologies now profile complementary facets of these states at scale. Perturb-seq (Dixit et al., 2016) combines CRISPR-based perturbations with single-cell RNA sequencing to read out transcriptome-wide effects, while Optical Pooled Screening (OPS) (Feldman et al., 2019; Sivanandan et al., 2023; Ramezani et al., 2025) uses cost-effective fluorescence microscopy to capture morphological phenotypes. Together, transcriptomics and imaging provide complementary views of how perturbations reshape cellular state.

The growing availability of large Perturb-seq and OPS datasets motivates methods for multimodal representation learning that capture their shared and modality-specific information. Shared structure can help disentangle true biological signal from technical confounders—such as random gene dropout in RNA-seq or intensity fluctuations in microscopy—that are unlikely to be supported across modalities (Radhakrishnan et al., 2023). At the same time, modality-specific signals contain unique phenotypes and mechanistic clues that, when combined, yield a more complete picture of perturbation effects.

Integrating these modalities is challenging. First, cell-level pairing is unattainable because state-of-the-art assays are destructive and cannot profile the same cells. Second, morphology and gene expression only partially overlap in the biology that they capture (Way et al., 2022), so models must remain robust when signals are disjoint.

To address these challenges, we present PETRI, an early-fusion self-supervised transformer that learns single-cell embeddings from unpaired images and transcriptomes. PETRI groups cells by perturbation into multimodal "documents" and employs self-attention to propagate useful signals across cells and between modalities for better reconstruction. PETRI embeddings represent individual cells from either modality and can be averaged in a modality-agnostic way to form perturbation-level profiles. Sparse autoencoders trained on these embeddings uncover morphological and molecular concepts with cross-modal support and perturbation specificity. To catalyze further research,

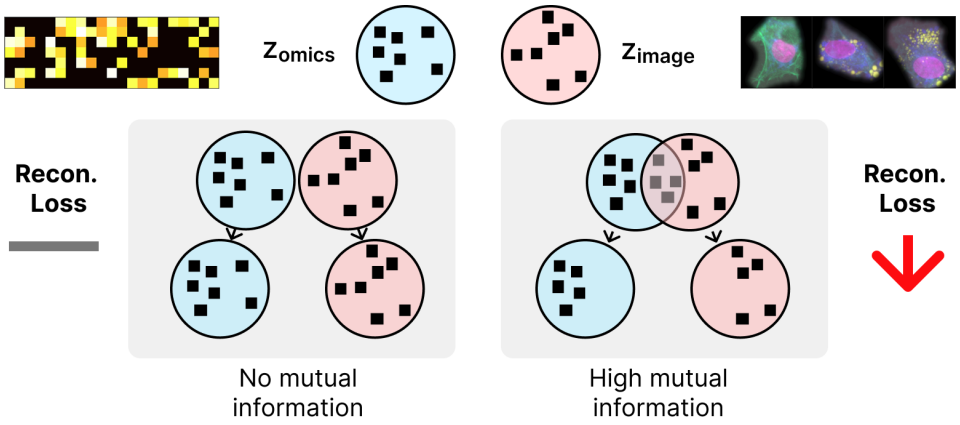


**Figure 1:** Conceptual overview: Joint reconstruction of masked regions promotes multimodal integration. Left: With disjoint modalities, loss matches unimodal decoders. Right: With high mutual information, cross-modal predictions reduce reconstruction loss.

we release a blinded, matched OPS and Perturb-seq dataset in HepG2 cells spanning 569 CRISPR knockouts in four chemical backgrounds.

## 2 Related Work

**Representation learning for cellular images** Image-based profiling aims to convert the rich information in microscopy images into quantitative feature vectors for downstream analysis in drug discovery and functional genomics (Chandrasekaran et al., 2021). Early deep learning approaches used weak supervision, training models to predict the experimental treatment (e.g., drug or genetic perturbation) applied to a cell or group of cells (Caicedo et al., 2018). However, this approach is limited by its core assumption that all perturbations produce a morphological change, which is often not the case. More recent work has shifted to self-supervised learning. These methods, including vision transformer-based self-distillation and masked autoencoders, have demonstrated state-of-the-art performance at inferring known biological relationships from images without relying on experimental labels for training (Doron et al., 2023; Kraus et al., 2024; Pham et al., 2025). Perturbation-aggregated embeddings are also used for predicting a compound's mechanism of action, identifying disease-specific phenotypes, and functional gene annotation (Sivanandan et al., 2023).

**Representation learning for single cell transcriptomics** Generative models have become a cornerstone of single-cell transcriptomics analysis, with pioneering methods like scVI using variational autoencoders to learn a probabilistic latent space that corrects for technical noise and batch effects (Lopez et al., 2018). This paradigm has evolved with the advent of foundation models for biology, which leverage the transformer architecture and pre-training on massive-scale datasets of tens of millions of cells (Cui et al., 2024; Theodoris et al., 2023; Gong et al., 2023; Pearce et al., 2025; Kalfon et al., 2025). Common tasks for these models include automated cell type annotation, integration of datasets from different experiments or technologies, and the prediction of cellular responses to genetic or chemical perturbations.

**Multimodal single cell embeddings** Integrating the many modalities that can be measured from a single cell is a key challenge in modern biology, as highlighted by community-wide efforts such as the NeurIPS 2021 Multimodal Single-Cell Data Integration Challenge (Lance et al., 2022). For datasets with cell-paired modalities, methods focus on learning a joint representation. For instance, deep generative models like MultiVI learn a probabilistic embedding of paired multi-omic data (Ashuach et al., 2023), while contrastive frameworks like scCLIP align paired chromatin accessibility and gene expression profiles (Xiong et al., 2023). This principle also extends to linking imaging with molecular data, where models like OmiCLIP learn to associate histopathology images with their corresponding spatial transcriptomics profiles (Chen et al., 2025b).

However, collecting paired image and transcriptomics data is often expensive or technically infeasible, making methods that can integrate unpaired data crucial. To address this, some approaches aim to align individual cells across modalities; for example, propensity score alignment leverages shared perturbation labels to estimate a matching between cells in different datasets (Xi & Hartford, 2024). Yang et al. (2021) introduce cross-modal autoencoders which use a two-stage learning process, first fitting a variational autoencoder (VAE) to images and then training a VAE for gene expression with a regularization loss that forces the latent spaces to overlap. Other methods operate at the level of cell populations using weaker supervisory signals. CellCLIP learns a shared embedding space between textual descriptions of perturbations and the sets of cell images resulting from them (Lu et al., 2025). Similarly, MultiMIL uses sample-level labels, such as patient disease status, in a multiple-instance learning (MIL) framework to identify the specific cells in different modalities that are characteristic of that label (Litinetskaya et al., 2024).

**Vision-language models for multimodal representation learning** Recent advances in VLMs have established powerful architectures for multimodal learning. Some models utilize resampling and cross-attention mechanisms to fuse information from interleaved image and text data (Alayrac et al., 2022), while others, like LLaVA, project image features into the word embedding space and process a unified sequence with a standard self-attention mechanism (Liu et al., 2023). Although most often used for visual question answering, the VLM framework can be adapted for representation learning. For example, MoCa is trained to denoise and reconstruct both image and text inputs simultaneously, enabling it to learn effective bidirectional multimodal embeddings from large, unlabeled datasets (Chen et al., 2025a).

**Matched imaging and gene expression datasets** The LINCS dataset (Way et al., 2022) includes a library of 1,327 chemical perturbations with Cell Painting and L1000 readouts. L1000 measures 978 mRNA transcripts from bulk samples, though the authors of the dataset report that it suffers from poor reproducibility of perturbation effects. Perturb-FISH (Binan et al., 2025) has matched Perturb-seq and OPS with MERFISH (Chen et al., 2015) for 35 genetic perturbations. MERFISH provides single cell pairing of morphology and mRNA counts but is limited to a few hundred genes. Relatedly, Perturb-Multi (Saunders et al., 2025) is a unique spatial transcriptomics dataset that genetically perturbs mouse liver cells *in vivo* and records MERFISH and protein staining.

## 3 METHOD

For multimodal representation learning, CLIP (Radford et al., 2021) may seem like a natural starting point. But, contrastive methods like CLIP are ill-suited to our scenario since their effectiveness hinges on strong positive pairs distinguishable from a large number of negatives. Our dataset has limited unique treatments (about 2,200) and the modalities have no explicitly overlapping features.

This motivates a shift in perspective. Instead of treating each modality instance as a distinct item in a pair, we draw inspiration from VLMs that operate on mixed-modality documents. A document, like a webpage for example, contains data that is merely aligned by a common topic – the context increases the odds of finding cross-modality associations. PETRI conceives of perturbations as topics and builds documents from sets of cells. Our central hypothesis is that cellular phenotypes visible in both modalities, and enriched under certain perturbations, will supply mutual information that improves the reconstruction of corrupted data (**Fig. 1**). If the modalities have no mutual information or even contradictory information, we expect cross-modality attention to decrease and for reconstruction loss and representation quality to default to the level of unimodal models.

The document approach is appealing, but introduces a significant technical hurdle: an exploding sequence length. A single cell can be represented by hundreds of image patch tokens or thousands of gene tokens. A set of cells would produce a sequence far too long for standard transformers. Our solution is aggressive token resampling. By distilling the token representation of each cell into a small, fixed number of latent tokens, we can flexibly scale the number of cells per document.

**Fig. 2a** outlines PETRI's architecture, which consists of four steps:

1. Create batches where cells are grouped by perturbation to form sets.

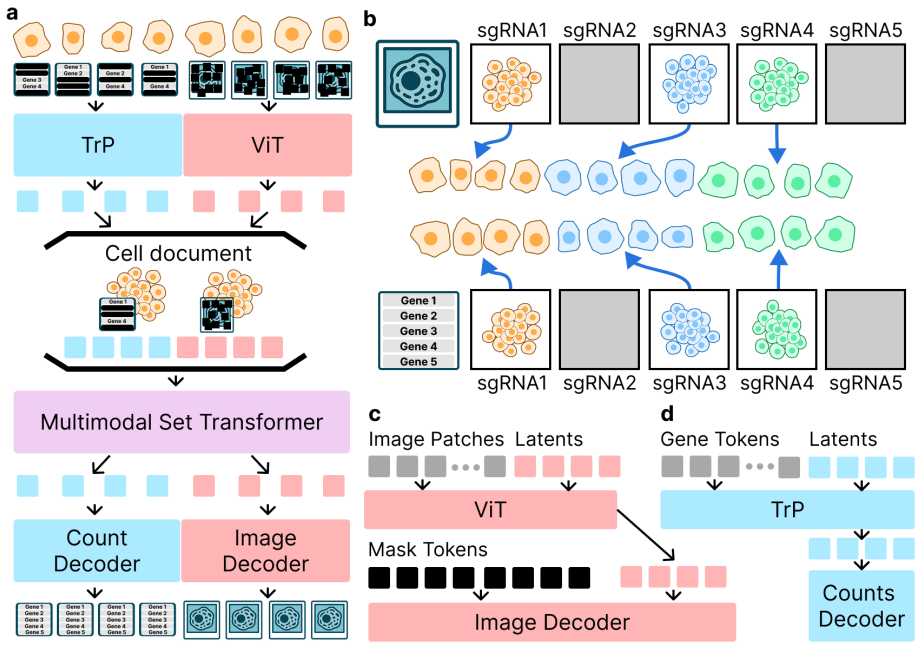


**Figure 2:** The PETRI architecture. **a**, Overview of separate encoders, cell document creation, multimodal set transformer, and separate decoders. **b**, Data grouping and sampling by modality and perturbation. **c**, ViT-based masked autoencoder for images with token resampling. **d**, Perceiver-based masked autoencoder for transcriptomics with token resampling.

2. Convert raw cell data (images or transcriptomes) into tokens, mask a large portion of them, and use a modality-specific encoder to resample the unmasked tokens into a small, fixed-size sequence of latents.
3. Concatenate the latent tokens from cells in the same sets to form multimodal documents and process them with a transformer.
4. Split the document back into individual cell latent tokens and use modality-specific decoders to reconstruct the original masked input from the latent representations.

### 3.1 Batch Construction

For training, we stratify datasets into groups based on their perturbation; e.g., a unique genetic treatment like the guide RNA (sgRNA). From each group, we sample a set of $S$ cells with replacement for each modality. Multiple sets are collated into a mini-batch and dispatched to the appropriate encoder. This process is depicted in **Fig. 2b**.

### 3.2 Tokenization and Per-Cell Resampling

To construct the multimodal documents, each cell must first be encoded into a compact latent representation. This is achieved with modality-specific encoders designed with token resampling. During training, we randomly mask 75% of input tokens (image patches or genes) per cell and remove them from the sequence.

For images, we follow the standard Vision Transformer (ViT) approach of embedding non-overlapping patches with position encodings. To achieve resampling, we concatenate a fixed number of learnable latent tokens, $M$, to the sequence of image patch tokens, $N$, where $M \ll N$. This combined sequence is processed by transformer blocks, and only the $M$ output latent tokens are retained as the cell's image representation (**Fig. 2c**).

For transcriptomics, the input sequence length, corresponding to thousands of genes, makes a standard transformer architecture computationally infeasible. We therefore required an architecture

purpose-built for efficiently processing and resampling extremely long sequences. We adopt the Perceiver (Jaegle et al., 2021), which interleaves cross-attention layers for resampling with self-attention layers over the latents only (**Fig. 2d**). This design directly serves our need for aggressive token resampling, making it a natural architectural choice. Gene expression is tokenized by combining a learned gene embedding with its measured log count via a two-layer MLP.

### 3.3 Multimodal Set Transformer (MST)

The modality-specific encoders output a fixed number of latent tokens per cell, resulting in a tensor of shape $(G \times S, L, D)$, where $G$ is the number of groups in the batch, $S$ is the set size, $L$ is the number of latent tokens, and $D$ is their dimensionality.

To form the cell documents, we reshape this tensor to $(G, S \times L, D)$ and concatenate the representations from both modalities along the token dimension to form a batch of unified sequences with shape $(G, 2 \times S \times L, D)$. These sequences are then processed through a series of standard transformer blocks, allowing for cross-modal and cross-cell attention. Afterwards, we split the sequence by modality and reconstitute the original $(G \times S, L, D)$ shapes for decoding.

### 3.4 Decoders and Loss

The final step is to reconstruct the original inputs from the processed latent tokens, enforcing that the latents capture comprehensive information about each cell.

The image decoder is adapted from Masked Autoencoders (MAEs) (He et al., 2022). Since our latent tokens are not tied to specific patch locations, we concatenate them with a full sequence of learnable mask tokens. The decoder is trained to reconstruct the original masked image patches from this combined sequence. The loss is the mean squared error (MSE) between the reconstructed and original pixel values of the masked patches only.

For the transcriptomics decoder, we mean-pool the latent tokens for each cell and pass them through a three-layer MLP that outputs a value for each gene. When raw counts are available, we apply a softmax over the gene dimension and use the negative log-likelihood of a negative binomial distribution as the loss. If reconstructing log-normalized counts instead, we use an MSE loss. Analogous to the image modality, the loss is calculated exclusively on masked-out genes.

### 3.5 Evaluation

We evaluate two metrics on aggregated embeddings from genetic treatment metadata.

**Guide Consistency.** In CRISPR screens, multiple guide RNAs (sgRNAs) are designed to target the same gene for editing, and thus these guides should induce similar phenotypic effects. To assess whether the models' representations are consistent with this prior knowledge, we compute cosine similarities of mean guide embeddings within each target gene and compare them to an empirical null distribution of similarities with the same cardinality between unrelated sgRNAs. The metric we report is the fraction of target genes in the screen with guides that have a statistically significant ($p < 0.05$) similarity after multiple testing correction.

**StringDB edge classification.** Introduced by Sivanandan et al. (2023), this metric uses physically interacting gene-gene pairs collected from the StringDB database as ground truth labels in a zero-shot classification task. Pairwise cosine similarities are computed for aggregated target gene embeddings. The similarities are treated as pseudo classifier probabilities and we evaluate the true positive rate (TPR) at 5% false positive rate (FPR) from the ROC curve. We expect this metric to be challenging because StringDB is not cell or phenotype specific and many single gene perturbations are expected to have weak or no effect.

Before aggregation for both metrics we robust center scale embeddings relative to a per replicate control and apply PCA and whitening without dimensionality reduction. Justification for and results without this preprocessing are in **A.5**. When working with multimodal embeddings, we aggregate modalities separately and average the results.

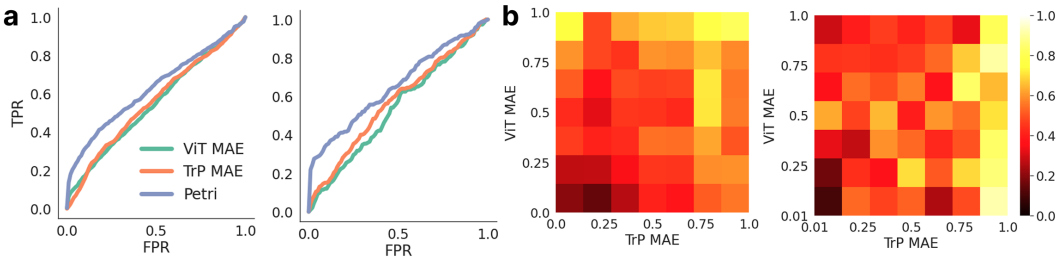


**Figure 3:** StringDB network edge metric analysis. The left panel of both sub-figures corresponds to the HepG2 dataset and the right to Perturb-Multi. **a**, ROC curves for PETRI and unimodal models (HepG2: PETRI AUC=0.628, ViT MAE AUC=0.549, TrP MAE AUC=0.556; Perturb-Multi: PETRI AUC=0.655, ViT MAE AUC=0.537, TrP MAE AUC=0.571). **b**, 2D histograms showing the pseudo classifier probabilities for unimodal models against the probabilities in PETRI.

## 4 Experiments

We evaluate PETRI on two datasets and compare against CLIP and unimodal baselines.

**HepG2: Matched Perturbations** Dataset consists of matched OPS and Perturb-seq in HepG2 cells in four different chemical background. Cells were imaged with fluorescence and label-free microscopy, while a separate but matched population was profiled with whole transcriptome Flex sequencing. We included cells that received exactly one sgRNA from a CRISPR knockout library of 569, with four unique sgRNAs per target gene. Cells were grouped first by chemical background and then by sgRNA as in **Fig. 2b**; 16 cells per modality were sampled from each group to form a document (see **A.1**). 8 latent tokens were used per cell such that documents contained $16 \times 2 \times 8 = 256$ tokens. The total dataset has 2M cells (0.9M images, 1.1M transcriptomics).

**Perturb-Multi (Saunders et al., 2025): Matched Cells** Dataset consists of spatial transcriptomics acquired from a single section of mouse liver tissue, including paired single-cell MERFISH measurements of 209 mRNAs and fluorescence images of 18 stained proteins. The dataset contains cells that received a sgRNA from a CRISPR knockout library of 203, with two unique sgRNAs per target gene. Cells were grouped by unique cell ID such that all groups contained a single cell; the set size was therefore 1 per modality and sampling was not required. Again, 8 latent tokens were used per cell, giving a document length of $1 \times 2 \times 8 = 16$ tokens.

To validate our hypothesis that joint reconstruction is robust even when the multimodal sets are not mutually informative, we additionally trained PETRI models on permuted data, where cells were randomly assigned to groups.

Unless otherwise stated, PETRI cell embeddings were extracted directly from the output of the modality-specific encoders, i.e., before cell document creation and the MST. The MST's cross-modality attention encourages the upstream encoders to create tokens that are well aligned and compatible. As tokens approach the decoders, they become specialized for the specific reconstruction task of that training step. The information sharing that happens in the MST is a critical driver for multimodal integration, even if the layers themselves do not produce the best embeddings for downstream tasks. More generally, SSL methods commonly benefit from removing layers (Bordes et al., 2022). Crucially, this choice also makes it possible to use the trained model for embedding individual cell images or gene expression profiles in unmatched screening data.

### 4.1 Perturbation profiles from PETRI embeddings recapitulate known biology

In this section, we benchmark PETRI for the task of aggregating multimodal data into a holistic perturbation embedding. As unimodal baselines, we tested strong pre-trained models, scGPT (Cui et al., 2024) and DINOv2 (Oquab et al., 2023), and modality-specific MAEs (TrP MAE and ViT MAE). For the HepG2 dataset only, we also experimented with variants of PETRI that operate on unimodal cell documents, which we designate PETRI Image and PETRI Omics. **Table 1** summarizes our findings. PETRI was substantially better on both datasets and evaluation metrics. The only

**Table 1:** Evaluation of PETRI against unimodal and multimodal baselines (GC=Guide Consistency; StringDB=StringDB edge classification). CLIP* denotes two different models: using single cell pairs for Perturb-Multi and using mean aggregated perturbation-level pairs for HepG2 (see **A.3**). Values are median of $N = 3$ for HepG2 and $N = 1$ for Perturb-multi.

| Modality | Method | Perturb-Multi | | HepG2 | |
|---|---|---|---|---|---|
| | | GC | StringDB | GC | StringDB |
| Transcriptomics | PCA | 0.030 | 0.073 | **0.304** | 0.150 |
| | scGPT - Pretrained | 0.059 | 0.078 | 0.048 | 0.057 |
| | TrP MAE | 0.059 | 0.109 | 0.195 | 0.144 |
| | PETRI Omics | - | - | 0.209 | 0.167 |
| Imaging | DINOv2 - Pretrained | 0.015 | 0.068 | 0.008 | 0.062 |
| | ViT MAE | 0.000 | 0.094 | 0.067 | 0.116 |
| | PETRI Image | - | - | 0.031 | 0.086 |
| Multimodal | TrP + ViT MAE Concat. | 0.000 | 0.099 | 0.155 | 0.153 |
| | TrP + ViT MAE Mean Cos. | 0.035 | 0.068 | 0.178 | 0.163 |
| | TrP + ViT MAE Max Cos. | 0.000 | 0.094 | 0.169 | 0.219 |
| | Cross-modal AE | - | - | 0.032 | 0.100 |
| | CLIP* | 0.000 | 0.057 | 0.051 | 0.174 |
| | PETRI Permuted Data | 0.163 | 0.260 | 0.274 | **0.255** |
| | PETRI | **0.208** | **0.260** | 0.278 | 0.242 |

exception was PCA on gene expression for the HepG2 dataset, which showed slightly higher guide consistency but a much lower StringDB score. ROC curves for PETRI against unimodal MAEs show better detection of StringDB edges at all false positive rates (**Fig. 3a**).

Next, we considered simple late-fusion of unimodal perturbation profiles with two methods: (1) Mean or max aggregation of the existing cosine similarity matrices and (2) concatenation of unimodal embeddings and computation of new similarity matrices. Notably, max aggregation gave StringDB scores closer to PETRI for the HepG2 dataset, though guide consistency showed no increase over TrP MAE. Additional late-fusion strategies are evaluated in **A.7**.

Cross-modal autoencoders (AE) were considered as another late-fusion method, designed expressly for unpaired data like the HepG2 dataset. Using default hyperparameters for images and gene expression, we found that the latent space alignment enforced by the loss function prevented the accurate reconstruction of RNA-seq counts. Results from this model were on par with unimodal image baselines, suggesting that gene expression was not effectively fused into the frozen image-only latent space.

CLIP, our early-fusion baseline, performed worse than late-fusion, especially for Perturb-Multi. We included Perturb-Multi because we expected its cell-level pairs of protein-stained images and mRNA counts would be more appropriate for CLIP. We theorize that the two modalities may not have a close enough relationship for contrastive learning. As a test, we trained a ViT to regress the mRNA counts directly from protein images. On a held out validation set we found that 80% of mRNAs were predicted with $r^2 < 0.20$ (mean $r^2 = 0.117$; see **A.8**). To adapt CLIP for the HepG2 dataset, we used sets of perturbed cells, as in PETRI, and mean aggregated their profiles before computing the contrastive loss. This is effectively the architecture of CellCLIP without pre-trained and frozen encoders. Embeddings from this model performed worse than simple late-fusion baselines.

To establish that PETRI is incorporating information from both modalities into its aggregated perturbation profiles, we plotted the pseudo classifier probabilities against unimodal models (**Fig. 3b**). Probabilities for PETRI were highest when they were also high in both modalities. They remained high, but decreased, where the modalities disagreed. The correspondence is stronger for HepG2 than Perturb-Multi, which appears to favor gene expression over images. Overall, this shows that PETRI learns associations between gene pairs even when that association is only strongly visible in one.

A final striking finding was that training PETRI on the permuted variants of the datasets gave broadly similar results (somewhat worse on Perturb-Multi and better on HepG2). Importantly, note that the metrics evaluated here measure the ability to aggregate unimodal information into a perturbation pro-

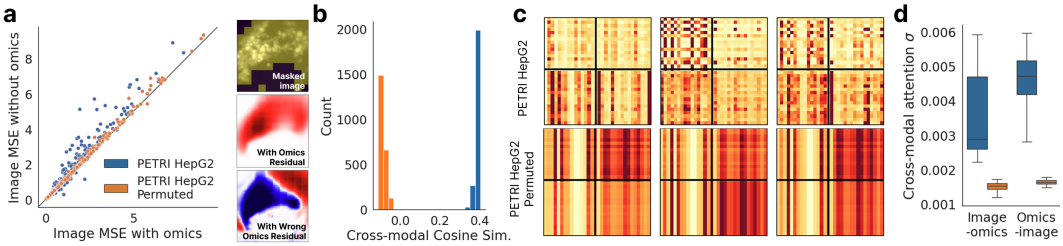


**Figure 4:** PETRI cross-modality information usage. **a**, Left: BODIPY reconstruction loss with vs. without access to perturbation-matched omics. Right: Example masked image of BODIPY channel (top) and residual maps between the image-only reconstruction and with matched vs. random omics (bottom two; red positive, blue negative). **b**, Cross-modality cosine similarities of treatment-aggregated embeddings. **c**, Representative attention heads; black lines separate image (top-left) and omics (bottom-right) token blocks; off-diagonals indicate cross-modal attention. **d**, Comparison of the mean of standard deviations, $\sigma$, over rows in the upper right (image-omics) and lower left quadrants (omics-image) of attention maps, from all attention heads in the MST layers.

file. Training on permuted data discourages cross-modality learning, but unimodal learning should be unaffected. The finding that PETRI is robust to permuted data is a positive result. It seems that by virtue of existing in the same space, unimodal PETRI embeddings can more effectively be aggregated into a multimodal profile with simple averaging than more sophisticated late-fusion methods on embeddings from separately trained models. The following sections dig deeper and show that, with correctly matched data, PETRI actually learns cross-modality relationships with biological relevance that are useful beyond aggregate perturbation profiles.

### 4.2 PETRI INTERNALLY MAKES CROSS-MODALITY PREDICTIONS

To validate that PETRI uses cross-modality information, we performed an ablation analysis. We focused on the reconstruction of the BODIPY channel, which stains for lipid droplets, in the HepG2 dataset. This choice was motivated by our screening library that includes control perturbations known to regulate lipid droplet size, quantity, and distribution. Using a simple intensity threshold, we selectively masked out image patches containing droplets and evaluated the MSE loss for those patches. We calculated loss when cell documents only had latent tokens from the masked image and when given access to latent tokens from a set of unmasked gene expression profiles.

Providing gene expression data from the same control perturbation decreased the reconstruction loss of BODIPY relative to the image only case (**Fig. 4a**). The effect was sporadic across cells, but when present the MSE decrease was sizable – the same was not true for the model trained on permuted data. Inspecting an image where including gene expression had a large effect, we saw an increase in predicted BODIPY intensity in the masked-out region. Providing gene expression from a different control perturbation known to decrease lipid droplets, we saw a decrease in BODIPY intensity instead. This conclusively shows that PETRI image reconstructions are influenced by transcriptomics data, though it required targeted analysis to detect (see **A.11**).

Returning to the hypothesis that embeddings before the Multimodal Set Transformer (MST) would be aligned across modalities, we measured the cosine similarity of unimodal perturbation profiles created from the PETRI modality-specific encoder outputs. Whereas, the model trained on correctly matched data had a clear and positive cosine similarity, the model trained on permuted data showed near orthogonality (**Fig. 4b**). This orthogonality would not hurt performance on the evaluation metrics in the previous section, but does point to a bifurcation of the modalities in the latent space. Further evidence of this is suggested by visualizing the attention heads in the MST. Training on correctly matched data uniquely showed patterns of non-trivial and statistically significant ($p < 0.001$) cross-modality attention (**Fig. 4c,d**).

### 4.3 IDENTIFICATION OF MULTIMODAL CELLULAR PHENOTYPES

Having established that PETRI uses cross-modality information for reconstruction, we probed the concepts it learns to encode. To disentangle shared and modality-specific structure in PETRI embeddings, we trained BatchTop$K$ sparse autoencoders (SAEs) with 15,360 dimensions and $K = 500$.

We examined all SAE dimensions to test whether grouping cells by perturbation during training encouraged more salient perturbation-specific concepts than uninformed baselines. We performed a differential analysis of SAE activations for perturbations versus the negative control, controlling false discovery rate with Benjamini–Hochberg. PETRI had substantially more perturbations with at least one differentially activated dimension than unimodal models or the model trained on permuted data (**Fig. 5a**). Intuitively, cells within a document vary in orientation, intensity, and cell cycle, but a perturbation-enriched phenotype provides context that reduces uncertainty (e.g., the appropriate BODIPY intensity to predict), nudging the model toward concepts that distinguish perturbations. While supervised training also aligns concepts to perturbations, it implicitly assumes each perturbation has a unique and appreciable effect, which is often violated in CRISPR screens.

We then searched for concepts shared across modalities, defining "multimodal" dimensions as those activated in 10–90% of cells in both imaging *and* transcriptomics. PETRI produced 298 such dimensions, compared with 0 for the permuted model and 1 for CLIP (**Fig. 5b**); these results speak to PETRI's ability to align modalities in the latent space, leading to the existence of multimodal concepts. If these concepts truly reflect biology common to both modalities, they should be less sensitive to modality-specific technical artifacts. We tested this by training logistic regression models to predict OPS well identity using either the 298 multimodal SAE dimensions or 298 randomly selected ones. Classifiers using the multimodal dimensions were significantly less accurate ($p < 0.001$; **Fig. 5c**), consistent with reduced encoding of undesirable well-specific technical factors.

Finally, we inspected the multimodal SAE dimensions for interpretability. For each dimension, we selected and compared images and transcriptomic profiles with zero activation against those with activation above the 99th percentile and asked whether there were statistically significant differences between these sets of images and transcriptomic profiles, respectively. For images, we ran differential analysis on handcrafted features of fluorescence intensity and nuclear and lipid droplet morphology. For transcriptomics, we performed differential expression analysis and computed prerank enrichment against Gene Ontology (GO) terms. In total, 127 of the 298 dimensions showed significant differences for at least one image feature and enrichment in at least one GO term.

To further interrogate these differential PETRI SAE dimensions for relevant biology, we searched by keywords for SAE dimensions that showed enrichment in terms related to phenotypes we know should be present in our dataset: cell cycle, lipid metabolism, and mitochondrial activity. Corresponding images for those dimensions revealed interpretable biological concepts including DNA replication (correlating with nucleus shape and DAPI intensity), cholesterol homeostasis (correlating with lipid droplet size and quantity), and aerobic respiration (correlating with mitochondrial fusion and network structure) (**Fig. 5d**).

These results are significant, demonstrating a scalable, self-supervised method for linking molecular states to morphological phenotypes at single-cell resolution. This enables researchers to confidently prioritize morphological changes that are validated by a corresponding molecular signature, ensuring they are both real and biologically relevant. SAEs trained on unimodal embeddings cannot be used for this purpose without first finding correspondences between separately learned concepts.

## 5 CONCLUSION

A key insight is that joint reconstruction over context-grouped data can induce meaningful multimodal alignment without explicit cross-modal losses. PETRI demonstrates this, opening a path to unify historically separate screening modalities. At the same time, our results highlight limits of prevailing proxies (e.g., guide consistency, protein–protein interaction prediction): useful as benchmarks, but insufficient to capture the biologically meaningful structure revealed by our downstream analyses. Instead, there is a need for task-grounded evaluation frameworks tailored to multimodal phenotypic screening and therapeutic discovery, with metrics that directly assess biological utility.

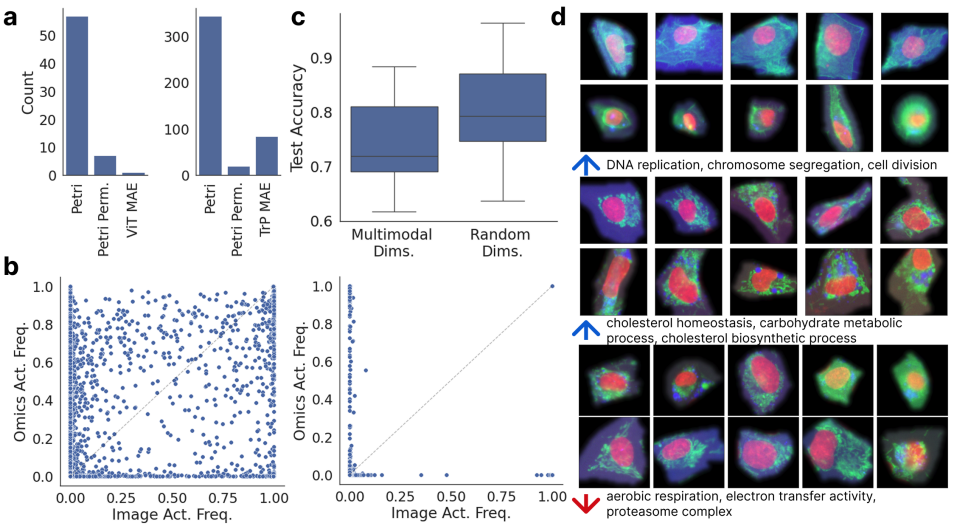


**Figure 5:** SAE analysis on HepG2. **a**, Number of perturbations with $\geq 1$ differentially activated SAE dimension in images (left) or omics (right). **b**, Fraction of cells activating each SAE dimension in images vs. omics (left) and on permuted data (right). **c**, Test-accuracy distributions across classifiers for a batch–prediction task using multimodal SAE dimensions vs. randomly sampled dimensions (20 random draws per classifier). **d**, Representative images: top, non-activating samples; bottom, 99th-percentile activations. Arrows annotate the top three GO terms (pre-ranked enrichment; direction indicated). Channels: DAPI; BODIPY; Phalloidin (top); MitoProbe (bottom two).

Several other important questions remain: How closely matched must experimental contexts be to enable cross-modality learning? Is it possible to incorporate other biological priors? For instance, instead of documents that focus on a single perturbation, what if we centered them on protein complexes or pathways? Our framework provides a foundation for exploring these questions and suggests that strategic contextual grouping could serve as a mechanism for easily incorporating domain knowledge into representation learning. Our release of the unique HepG2 dataset should enable such future ML development. A matched imaging and omics dataset of this scale and quality has not previously been made publicly available.

Although PETRI is designed for images and gene expression, the core idea of using joint reconstruction from context-aligned cellular documents could be adapted to other modalities. As diverse omics technologies continue to proliferate, approaches that can integrate complementary views of cellular state without requiring perfect experimental alignment will become increasingly valuable for advancing our understanding of cellular biology and accelerating therapeutic discovery.

### Reproducibility Statement

Code and the blinded HepG2 dataset will be made public. All results aside from StringDB metrics on the HepG2 dataset will be reproducible from the provided code and data.

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. *arXiv [cs.CV]*, April 2022.

Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods*, 20(8):1222–1231, August 2023.

Loc Binan, Aiping Jiang, Serwah A Danquah, Vera Valakh, Brooke Simonton, Jon Bezney, Robert T Manguso, Kathleen B Yates, Ralda Nehme, Brian Cleary, and Samouil L Farhi. Simultaneous CRISPR screening and spatial transcriptomics reveal intracellular, intercellular, and functional transcriptional circuits. *Cell*, 188(8):2141–2158.e18, April 2025.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *arXiv [cs.LG]*, June 2022.

Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of single-cell feature embeddings. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018:9309–9318, June 2018.

Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.*, 20(2):145–159, February 2021.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. MoCa: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. *arXiv [cs.CV]*, June 2025a.

Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233): aaa6090, April 2015.

Weiqing Chen, Pengzhi Zhang, Tu N Tran, Yiwei Xiao, Shengyu Li, Vrutant V Shah, Hao Cheng, Kristopher W Brannan, Keith Youker, Li Lai, Longhou Fang, Yu Yang, Nhat-Tu Le, Jun-Ichi Abe, Shu-Hsia Chen, Qin Ma, Ken Chen, Qianqian Song, John P Cooke, and Guangyu Wang. A visual-omics foundation model to bridge histopathology with spatial transcriptomics. *Nat. Methods*, 22 (7):1568–1582, July 2025b.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods*, 21(8):1470–1480, August 2024.

Andor Diera, Lukas Galke, and Ansgar Scherp. Isotropy matters: Soft-ZCA whitening of embeddings for semantic code search. *arXiv [cs.CL]*, November 2024.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxivorg*, June 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv [cs.CV]*, October 2020.

David Feldman, Avtar Singh, Jonathan L Schmid-Burgk, Rebecca J Carlson, Anja Mezger, Anthony J Garrity, Feng Zhang, and Paul C Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799.e17, October 2019.

Jing Gong, Minsheng Hao, Xingyi Cheng, Xin Zeng, Chiming Liu, Jianzhu Ma, Xuegong Zhang, Taifeng Wang, and Le Song. xTrimoGene: An efficient and scalable representation learner for single-cell RNA-seq data. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664. PMLR, July 2021.

Jérémie Kalfon, Jules Samaran, Gabriel Peyré, and Laura Cantini. scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nat. Commun.*, 16(1):3607, April 2025.

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 Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11757–11768. IEEE, June 2024.

Christopher Lance, Malte D Luecken, Daniel B Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, Qiao Liu, Nikolay Russkikh, Gleb Ryazantsev, Uwe Ohler, NeurIPS 2021 Multimodal data integration competition participants, Angela Oliveira Pisco, Jonathan Bloom, Smita Krishnaswamy, and Fabian J Theis. Multimodal single cell data integration challenge: Results and lessons learned. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 162–176. PMLR, July 2022.

Anastasia Litinetskaya, Maiia Shulman, Soroor Hediyeh-zadeh, Amir Ali Moinfar, Fabiola Curion, Artur Szałata, Alireza Omidi, Mohammad Lotfollahi, and Fabian J Theis. Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. *bioRxiv*, pp. 2024.07.29.605625, July 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023.

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

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP – learning perturbation effects in cell painting via text-guided contrastive learning. *arXiv [cs.LG]*, June 2025.

M Oquab, Timothée Darcet, Théo Moutakanni, Huy Q Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, H Jégou, J Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, April 2023.

James D Pearce, Sara E Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li, Stephen R Quake, and Theofanis Karaletsos. A cross-species generative cell atlas across 1.5 billion years of evolution: The TranscriptFormer single-cell model. *bioRxiv*, pp. 2025.04.25.650731, April 2025.

Chau Pham, Juan C Caicedo, and Bryan A Plummer. ChA-MAEViT: Unifying channel-aware masked autoencoders and multi-channel vision transformers for improved cross-channel learning. *arXiv [cs.CV]*, October 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv [cs.CV]*, February 2021.

Adityanarayanan Radhakrishnan, Sam F Friedman, Shaan Khurshid, Kenney Ng, Puneet Batra, Steven A Lubitz, Anthony A Philippakis, and Caroline Uhler. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nat. Commun.*, 14(1):2436, April 2023.

Meraj Ramezani, Erin Weisbart, Julia Bauman, Avtar Singh, John Yong, Maria Lozada, Gregory P Way, Sanam L Kavari, Celeste Diaz, Eddy Leardini, et al. A genome-wide atlas of human cell morphology. *Nature methods*, 22(3):621–633, 2025.

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, August 2024.

Shota Sasaki, Benjamin Heinzerling, Jun Suzuki, and Kentaro Inui. Examining the effect of whitening on static and contextualized word embeddings. *Inf. Process. Manag.*, 60(3):103272, May 2023.

Reuben A Saunders, William E Allen, Xingjie Pan, Jaspreet Sandhu, Jiaqi Lu, Thomas K Lau, Karina Smolyar, Zuri A Sullivan, Catherine Dulac, Jonathan S Weissman, and Xiaowei Zhuang. Perturb-multimodal: A platform for pooled genetic screens with imaging and sequencing in intact mammalian tissue. *Cell*, 188(17):4790–4809.e22, August 2025.

Srinivasan Sivanandan, Bobby Leitmann, Eric Lubeck, Mohammad Muneeb Sultan, Panagiotis Stanitsas, Navpreet Ranu, Alexis Ewer, Jordan E Mancuso, Zachary F Phillips, Albert Kim, et al. A pooled cell painting crispr screening platform enables de novo inference of gene function by self-supervised deep learning. *bioRxiv*, pp. 2023–08, 2023.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, and Patrick T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

Gregory P Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C Caicedo, Beth A Cimini, Kyle Karhohs, David J Logan, Mohammad H Rohban, Maria Kost-Alimova, Kate Hartland, Michael Bornholdt, Srinivas Niranj Chandrasekaran, Marzieh Haghighi, Erin Weisbart, Shantanu Singh, Aravind Subramanian, and Anne E Carpenter. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Syst.*, 13(11):911–923.e9, November 2022.

Johnny Xi and Jason S Hartford. Propensity score alignment of unpaired multimodal data. *Neural Inf Process Syst*, abs/2404.01595:141103–141128, April 2024.

Lei Xiong, Tianlong Chen, and Manolis Kellis. scCLIP: Multi-modal single-cell contrastive learning integration pre-training. In *NeurIPS 2023 AI for Science Workshop*, October 2023.

Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, G V Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.*, 12(1):31, January 2021.

# A Appendix

## A.1 HepG2 Dataset

Briefly, a pool of constitutive Cas9-expressing hepatocellular carcinoma cells (HepG2s) received approximately one sgRNA per cell from a CRISPR knockout library of 569 genes, with four unique sgRNAs per target gene. Cells were expanded as a uniform pool to maximize parity between imaging and transcriptomic populations, then treated with one of four chemical treatments. Cells were imaged with fluorescence at 20X magnification by staining the cells with a variation of CellPaint, which included HOESCHT, PHALLOIDIN, BODIPY, and MITOPROBE to label DNA, actin, neutral lipids, and mitochondrial rRNA, respectively. Cells were additionally imaged via quantitative phase imaging and brightfield. Imaging was followed by sgRNA amplification and in-situ sequencing. A parallel population was used for the creation of the transcriptomic dataset, which was generated using Flex technology, with probes reading out both the transcription, as well as the gRNA sequences, within each cell.

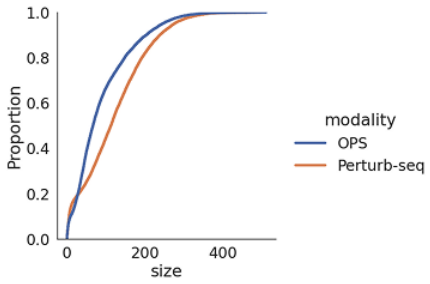


**Figure S1:** ECDF of cell counts per perturbation group (sgRNA)

Due to fitness effects, the number of barcoded cells that received each sgRNA varies considerably **(Figure S1)**. 19% of perturbation groups had fewer than 16 cells in one or the other modality, which motivated the choice of 16 as the default set size for PETRI. With larger set sizes and sampling with replacement, the possibility of leakage increases, e.g., the same cell with different masking patterns can appear in the same set.

## A.2 PETRI Implementation

**Unimodal encoders and Multimodal Set Transformer (MST)** The image encoder is a standard ViT-Base model (Dosovitskiy et al., 2020) with 85M parameters. The gene expression encoder is a Perceiver with four cross-attention layers and 12 self-attention layers, where each cross-attention layer is followed by three self-attention layers. The total parameter count is 131M, of which 16M correspond to learnable gene embeddings. To match ViT-Base, the embedding dimension is 768. After the unimodal encoders, the MST is a four layer transformer comprised of ViT-style self-attention blocks with a total of 14M parameters.

**Decoders** The image decoder architecture follows Masked Autoencoders and using a ViT with embedding dimension of 512 and eight transformer layers amounting to 26M parameters. The gene expression decoder is a simple three-layer MLP with hidden dimension of 128 and final output dimension of either 18082 for HepG2 or 209 for Perturb-Multi. It consists of 2M parameters.

The total parameter count is 259M with roughly equal numbers of parameters for each modality. However, the gene expression-related parameters are almost entirely in the encoder whereas a substantial fraction of the image-related parameters are in the decoder.

## A.3 CLIP Baseline Implementation

To adapt CLIP to our setting, we used the same architecture as PETRI but removed the decoders and Multimodal Set Transformer. For Perturb-Multi, with cell pairs, we used a batch size of 4096 and the standard CLIP loss over positive and negative pairs. For HepG2, with perturbation-level

matching, we used sets of 16 cells sampled from the same perturbation as positive pairs. Before computing the loss, we mean-aggregated the embeddings of all cells in the set. The batch size was 4096 cells per modality; however, after aggregation there were $4096 \div 16 = 256$ pairs over which to compute the CLIP loss.

### A.4 Model Training

PETRI and baseline models were all trained for 100 epochs. For the HepG2 dataset, every epoch included exactly 30 samples of cells drawn from each perturbation group. This balanced sampling was consistent for all models that we trained, regardless of whether they required sets of cells or not. For Perturb-Multi, with cell-level pairs, this sampling was unnecessary and an epoch included each cell just once. In total, HepG2 models trained for 250K iterations and Perturb-Multi models trained for 145K iterations. Training took 2 days on a single node with 8 H100 GPUs and a batch size of 1536 masked images and/or expression profiles.

**Image augmentations and patching** Perturb-Multi and HepG2 images were augmented with random vertical and horizontal flips, 360 degree rotations and 5% translations. To account for cell sizes, we used center crops of 64 pixels for Perturb-Multi and 80 pixels for HepG2. ViT patch size was 8 pixels for both.

**Loss weighting** For the HepG2 dataset, we used a weight of 1.0 for the image MSE and $1 \times 10^{-4}$ for the negative binomial loss. For Perturb-Multi, we used a weight of 1.0 for the image MSE and 0.01 for the mRNA count MSE.

### A.5 Embedding postprocessing and evaluation

We preprocessed single cell embeddings before aggregating them into perturbation profiles and computing guide consistency and StringDB network edge metrics. For HepG2 image embeddings, we performed robust center scaling (RCC), i.e., we subtracted the median embedding of the intergenic control cells in each OPS well and divided by the interquartile range. Standardizing by per replicate controls is a common method for reducing batch effects introduced by natural well-to-well variability in culture media or fluorescence intensities and empirically works well for ViT embeddings (Kraus et al., 2024). For HepG2 gene expression embeddings, we performed RCC using the global intergenic controls. Cells from the four chemical backgrounds were preprocessed independently and the evaluation metrics for each background were averaged.

Since Perturb-Multi cells all come from a single batch, we used global statistics from the non-targeting control. PCA and whitening promote isotropy and enhance semantic search for language embeddings (Diera et al., 2024; Sasaki et al., 2023), with similar benefits for perturbation embeddings (Kraus et al., 2024).

**Table S1** shows that StringDB metric evaluation on the raw embeddings gives no better than random performance (0.05) on HepG2.

**Table S1:** Evaluation of unimodal and multimodal embeddings without preprocessing on the HepG2 dataset for the StringDB edge classification metric.

| Modality | Method | HepG2 StringDB |
|---|---|---|
| Transcriptomics | PCA | 0.028 |
| | scGPT - Pretrained | 0.043 |
| | TrP MAE | 0.037 |
| | PETRI Omics | 0.040 |
| Imaging | DINOv2 - Pretrained | 0.050 |
| | ViT MAE | 0.040 |
| | PETRI Image | 0.043 |
| Multimodal | CLIP* | 0.040 |
| | PETRI Permuted Data | 0.045 |
| | PETRI | 0.031 |

### A.6 Inference Procedure and Intermediate Layer Embeddings

We chose to use PETRI embeddings from the outputs of the unimodal encoders because they can be computed without access to multimodal groups of cells. However, we also considered using cell embeddings from the intermediate layers of the Multimodal Set Transformer. This required a more complex inference procedure.

**Deterministic multimodal set inference.** Each perturbation group was first shuffled to mitigate potential batch effects, then wrap-around padded to ensure the total number of cells in the group was divisible by the set size. The padded indices were then partitioned into non-overlapping sets of fixed size, creating a deterministic enumeration. The embeddings for duplicated cells were averaged such that the final inference result had no duplicates. When the number of available sets differed between modalities for the same perturbation, the modality with fewer sets was cyclically repeated to match the longer modality. This guarantees that every cell appears at least once in each modality while maintaining proper multimodal alignment between sets.

**Table S2:** Evaluation of PETRI embeddings from intermediate layers on the HepG2 dataset (GC=Guide Consistency; StringDB=StringDB edge classification).

| Layer | HepG2 | |
|---|---|---|
| | GC | StringDB |
| Before MST (default) | **0.260** | **0.242** |
| After MST Layer 1 | 0.202 | 0.200 |
| After MST Layer 2 | 0.211 | 0.197 |
| After MST Layer 3 | 0.220 | 0.190 |
| After MST Layer 4 | 0.221 | 0.186 |

**Table S2** summarizes the results and shows that the embeddings taken directly from the unimodal encoders give the best guide consistency and StringDB scores on the HepG2 dataset.

### A.7 Late-fusion methods

We tested three late-fusion methods, all of which make use of aggregated embeddings from ViT MAE and TrP MAE after RCC and whitening:

1. **Concatenation:** Imaging and transcriptomics each had 768 dimension embeddings, which we concatenated to produce 1,536 dimensions. With these we computed pairwise cosine similarities between target gene or sgRNA profiles, depending on the metric.
2. **CCA:** We fit CCA with 30 components to find a shared space between embeddings from the two modalities. Embeddings from both modalities were projected into this space and we either concatenated or averaged them directly.
3. **Cosine similarity matrix aggregation:** Instead of working with the unimodal embeddings, we directly aggregated the cosine similarity matrices from each modality by min, max, mean, or median.

**Table S3** summarizes the results.

### A.8 Perturb-Multi ViT Regression

We trained a ViT-Base with the same settings as those in **A.4**. The class token was processed with a two-layer MLP to regress the 209 normalized mRNA counts in Perturb-Multi, loss was evaluated with MSE. **Fig. S2** shows the distribution of $r^2$ values for the mRNAs evaluated on a randomly chosen held out test set containing 20% of total cells.

### A.9 Permuted Dataset Evaluation

We observed that the performance on guide consistency and StringDB network edge metrics was roughly equivalent between PETRI models trained on the correctly grouped and matched versions of

**Table S3:** TrP + ViT MAE late-fusion results (GC=Guide Consistency; StringDB=StringDB edge classification).

| Method | Perturb-Multi | | HepG2 | |
|---|---|---|---|---|
| | GC | StringDB | GC | StringDB |
| TrP + ViT MAE Concat. | 0.000 | 0.099 | 0.155 | 0.153 |
| TrP + ViT MAE CCA Concat. | 0.000 | **0.104** | 0.019 | 0.079 |
| TrP + ViT MAE CCA Mean. | 0.000 | 0.099 | 0.020 | 0.079 |
| TrP + ViT MAE Min Cos. | 0.000 | 0.094 | 0.115 | 0.116 |
| TrP + ViT MAE Mean Cos. | 0.035 | 0.068 | **0.178** | 0.163 |
| TrP + ViT MAE Median Cos. | 0.035 | 0.068 | 0.178 | 0.163 |
| TrP + ViT MAE Max Cos. | **0.040** | 0.094 | 0.169 | **0.219** |

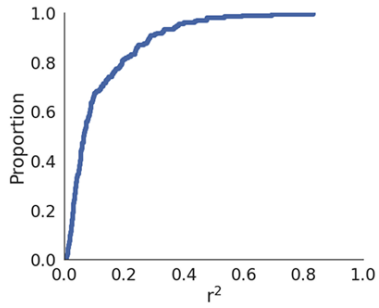


**Figure S2:** ECDF of $r^2$ scores for predictions on each of the 209 mRNA counts in Perturb-Multi.

the benchmark datasets versus their permuted versions. **Figure S3** shows the ROC curves and Venn diagrams for StringDB metrics. The results suggest that training on correct or permuted datasets leads to identification of broadly similar gene pairs.

### A.10 Document length and latent token count

The resampling mechanism allows us to trade-off the number of latent tokens for the number of cells in a document. In the main experiments, we used 8 tokens per cell and a set size of 16. At least for the available evaluation metrics, we did not observe a clear pattern of improvement or degradation from adjusting these parameters (**Figure S4**) and absolute differences were small. With smaller set sizes we might expect to see less cross-modality learning, especially if perturbations induce heterogeneous responses.

### A.11 Reconstruction loss ablation analysis

**Fig. 4a** shows the effect of a targeted ablation analysis on image reconstruction loss. At the aggregate level, we did not notice clear differences in reconstruction loss for images or gene expression in the HepG2 dataset when removing access to the other modality. However, we do observe that image reconstruction loss is lower overall than the model trained on permuted data **Figure S5a,b**.

For Perturb-Multi, on the other hand, there is a statistically significant ($p \ll 0.05$) increase in loss when access to the other modality is removed. This effect disappears for the model trained on permuted data. Again, supporting the notion that the model learns to ignore irrelevant multimodal information.

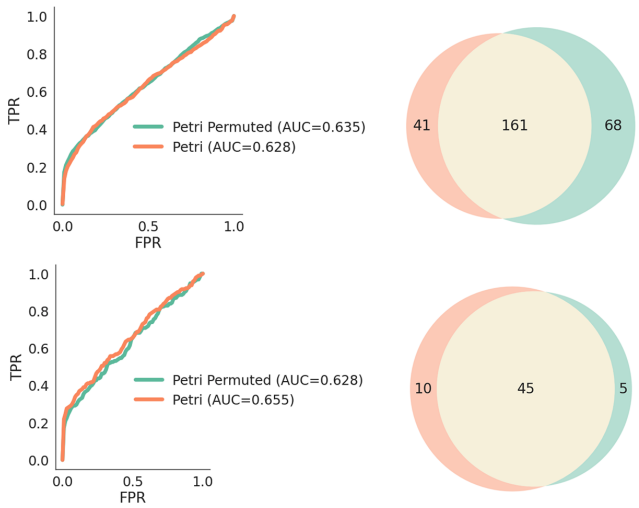


**Figure S3:** ROC curves and Venn diagrams for TP StringDB gene pair detections. **Top:** HepG2. **Bottom:** Perturb-Multi

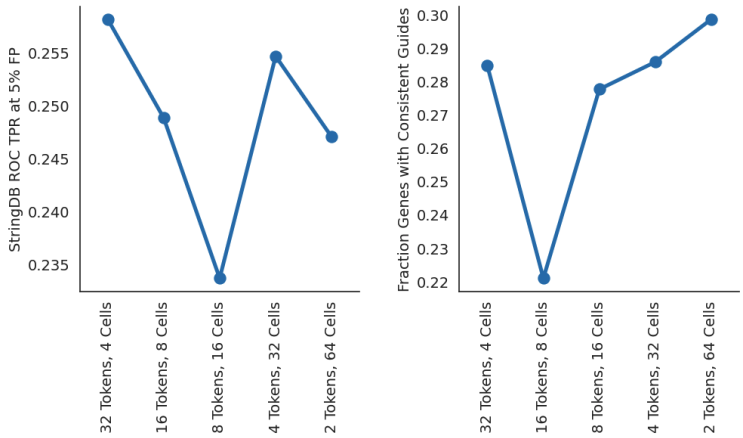


**Figure S4: Left:** StringDB metric on HepG2 with different combinations of latent tokens and set sizes. The listed number of cells is per modality. **Right:** Same for the guide consistency metric.

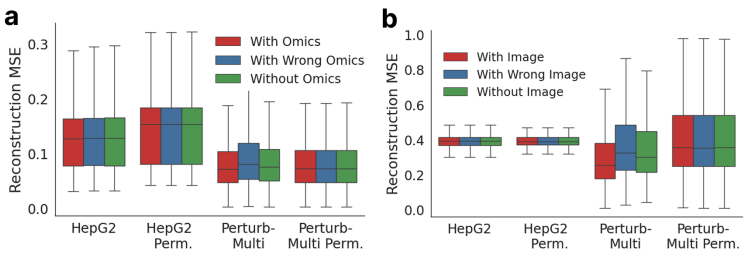


**Figure S5: a.** Image reconstruction loss across all channels when randomly masking 75% of patches; includes no access to omics, access to correctly paired omics, and access to random omics. **b.** Average reconstruction loss across all genes when randomly masking 75% of gene tokens; includes no access to images, access to correctly paired images, and access to random images.