This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Intervention-Aware Multiscale Representation Learning from Imaging Phenomics and Perturbation Transcriptomics

Jiayuan Chen Ruoqi Liu Zishan Gu Ping Zhang*
The Ohio State University
{chen.12930,liu.7324,gu.1090,zhang.10631}@osu.edu

Abstract

Microscopy-based phenotypic profiling is scalable for drug discovery but lacks the mechanistic depth of transcriptomics, which remains costly and scarce. Existing multimodal approaches either use images to support other modalities or naively align representations by sample identity, ignoring cell-type and dose variations in weakly paired data-limiting generalization to unseen interventions. In this paper, we introduce an intervention-aware distillation framework that leverages perturbational transcriptomics to guide image representation learning. A transcriptome-conditioned teacher integrates gene expression and intervention metadata to produce soft distributions over a chemistry-aware codebook organized by drug similarity. The teacher employs a fine-tuned single-cell foundation model to encode cell-type context and disentangle dose effects. An image-only student learns to predict these distributions from microscopy alone, distilling mechanistic knowledge while operating independently at test time. This design emphasizes intervention semantics rather than identity alignment and explicitly handles dose and cell-type mismatches. We provide theoretical guarantees showing that transcriptomic guidance tightens the risk bound for image-based prediction. On Cell Painting and RxRx datasets paired with L1000, our method significantly improves one-shot transfer to unseen interventions and drug-target gene discovery compared to self-supervised and alignment baselines[1].

1. Introduction

Phenotype-based screening [6, 22] has become central to drug discovery [11, 17] and to building virtual cells [4, 28] that predict cellular responses to perturbations. Two complementary readouts dominate this landscape: microscopy imaging [2] provides scalable, low-cost morphological profiles suitable for screening thousands of compounds, while perturbational transcriptomics [25, 28] captures pathway-level mechanistic signatures that reveal how drugs modulate cellular programs. In current practice, image-based models rely primarily on self-supervised learning (SSL) methods such as masked autoencoders (MAE) [12, 14] or DINO [5, 13] to extract transferable visual features. While these approaches learn general-purpose representations, the resulting features remain weakly connected to the biological mechanisms and transcriptomic changes that drugs actually induce [1]. They excel at capturing morphological patterns but lack the mechanistic grounding that transcriptomics provides-the very information needed to understand why a phenotype emerges and to predict responses to novel interventions. This raises a fundamental question: when limited paired image-transcriptome data becomes available, how can we leverage perturbational transcriptomics to guide image representation learning toward mechanistic understanding? Existing image-centric multimodal approaches [20] typically leverage transcriptomic or chemical information to improve compound representations. Some methods [8, 16, 27] align both images and RNA-seq to a shared drug embedding space, while others [10, 23, 29] directly pair drug structure representations with image features, bypassing RNA altogether. While these designs enable cross-modal retrieval, they fundamentally center on drug identity rather than intervention effects. By treating drug identity as the primary supervision signal, they collapse the effect variations introduced by dose and cell type into binary matching objectives-the same compound assayed under different conditions is treated as an equivalent positive pair, discarding the graded response information critical for generalization. Moreover, these approaches primarily aim to learn better drug [16, 23] or transcriptomic representations [1], with images serving mainly as auxiliary signals rather than the target modality for improvement. As a result, the learned image features remain disconnected from the mechanistic pathways that link drug chemistry to transcriptomic response and ultimately to morphological phenotype. Models trained this way struggle to transfer to unseen perturbations and cellular contexts where intervention effects differ from the training distribution [7, 15].

*Corresponding author.
[1] https://github.com/The-Real-JerryChen/BioMicroscopyProfiler

We propose to reframe multimodal learning around intervention semantics rather than sample identity. Drug perturbations follow a structured causal pathway: chemistry determines which pathways are engaged, shaping transcriptomic responses that ultimately manifest in morphological phenotypes. To leverage this structure, we introduce an intervention-aware distillation framework in which perturbational transcriptomics guides image representation learning during training. A transcriptome-conditioned teacher model encodes intervention effects from RNA-seq and metadata, while an image-only student learns to recover these effects from microscopy alone. Two technical designs instantiate this principle. First, we fine-tune a single-cell foundation model (scFM) on perturbation response prediction to encode cell-type context from basal gene expression and intervention-specific information from perturbed transcriptomes and doses. This encoding provides mechanism-aware representations that capture pathway-level effects while accounting for dosage variations across weakly paired samples. Second, to organize supervision by mechanism rather than identity, we construct a chemistry-aware codebook by projecting drug molecular fingerprints into a normalized prototype space. The teacher integrates both RNA-seq and paired images to produce soft distributions over this codebook, while the student learns to predict these distributions from images alone. Transcriptomics thus serves as privileged information during training, shaping image features toward mechanistic understanding, but is not required at test time. We also provide theoretical analysis showing that incorporating transcriptomic signals during training provably tightens the risk bound for image-based intervention prediction.

We validate our approach on multiple benchmarks spanning two encoder backbones. On Cell Painting [3, 7] and RxRx [15] datasets paired with L1000 perturbational transcriptomics, we evaluate one-shot transfer to unseen interventions and unsupervised drug-target gene discovery. Compared with well-established self-supervised baselines (MAE, DINO) and contrastive alignment methods, our intervention-aware framework achieves significant improvements across all tasks. In summary, our contributions are:

- We introduce an intervention-aware distillation framework that uses perturbational transcriptomics to guide image representation learning toward mechanistic understanding, with theoretical guarantees showing that transcriptomic guidance tightens the risk bound for image-based intervention prediction.
- We propose a chemistry-aware codebook that structures supervision by drug mechanism, combined with a transcriptome-conditioned teacher that leverages a fine-tuned scFM to encode cell context and intervention effects.
- We demonstrate strong empirical performance on one-shot transfer to unseen interventions and unsupervised drug-target discovery across Cell Painting and RxRx datasets, outperforming self-supervised and alignment baselines.

2. Related Work

2.1. Phenotypic Profiling of Drug Perturbations

Microscopy-based profiling, such as Cell Painting [2], uses multiplexed fluorescent imaging to capture morphological responses to drug perturbations at scale, enabling applications in novel biological relationship discovery [7, 15, 24]. These assays are cost-effective and high-throughput but lack direct mechanistic interpretation. Recent work applies SSL, such as MAE [14] and DINO [13], to extract morphological features from unlabeled images, though these remain weakly connected to biological mechanisms. Perturbational transcriptomics captures genome-wide gene expression changes, providing pathway-level mechanistic insights through datasets like LINCS L1000 [25] and Tahoe-100M [28]. scFMs such as scGPT [11] and Geneformer [26] can be fine-tuned to predict perturbation responses [18], offering rich mechanistic representations but at higher cost and lower throughput. These modalities are complementary, images provide scale, transcriptomics provides mechanism, yet paired datasets remain scarce and often weakly paired [1], with the same drug profiled at different doses, batch, or cell types across modalities.

2.2. Multimodal Learning for Perturbations

Recent work explores molecular representation learning from cellular data for cross-modal retrieval and molecular property prediction. Most methods employ alignment objectives: aligning drug structures with cellular images [23, 27, 29], or integrating gene expression with other omics to provide multi-faceted drug profiles [16, 27]. However, these approaches primarily aim to learn better drug or transcriptomic representations for downstream molecular tasks, such as predicting chemical properties and drug-drug interactions, with images serving mainly as auxiliary signals to enrich compound embeddings rather than as the target modality for improvement. A recent complementary direction distills image foundation models to improve transcriptomics representations [1], but the reverse that using transcriptomics to guide image learning remains underexplored. Critically, existing alignment methods center on drug identity rather than intervention effects, collapsing dose and cell-type variations into binary matches and struggling with weakly paired data. We instead propose intervention-aware distillation that using transcriptomics to guide image representations toward mechanistic understanding.

3. Method

3.1. Problem Statement

In this paper, we study cellular responses to drug interventions observed through two complementary phenotypic read-

outs: cellular imaging I and transcriptomics (RNA-seq) R.

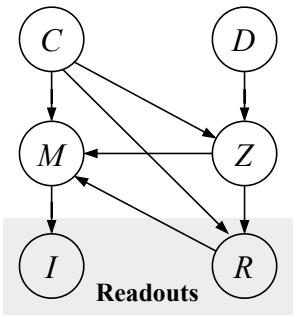


Figure 1. Causal pathway of drug interventions across modalities.

Causal Structure. Figure 1 depicts the multiscale cellular state under a drug intervention. Let D denote the molecular structure of the interventional drug; after exposure, the drug binds to its molecular target(s), producing a target-engagement/signaling state Z. This perturbed state then drives changes at two levels: the transcriptome R and the morphology M. Path $Z \to R$ corresponds to target binding perturbing signaling cascades and transcription-factor activity, inducing gene-expression changes captured by mRNA-seq. Morphology M is shaped both by direct drug-induced protein/biophysical effects via path $Z \to M$ and by indirect, transcription-mediated remodeling in which expression programs adjust protein levels/modifications and reorganize cellular architecture via the path $Z \to R \to M$. Finally, morphology is observed as images I through the optical pipeline that contains staining, imaging, and preprocessing. Cell type C provides the biological context, shaping target engagement Z, basal transcriptomic state R, and intrinsic morphology M, thereby introducing three conditioning paths, $C \to Z$, $C \to R$, and $C \to M$, that modulate the drug-induced cascades.

Scope and Assumptions. We focus on static and do not explicitly model temporal dynamics or multi-cellular interactions in our primary framework. While batch effects are a practical concern in multi-plate imaging and cross-laboratory RNA-seq data, we do not model them explicitly in our causal graph; we discuss mitigation strategies and their relationship to our framework in Appendix.

Task Definition. Given this setup, we target image-based drug discovery: learning image representations that enable drug identification, mechanism clustering, and generalization to unseen compounds and cell types. Formally, we aim to estimate $P(D \mid I, C)$, the posterior distribution over drugs given cellular images and cell type, which captures how morphology encodes intervention-specific signals. At test time, only images I and cell type C are available; transcriptomics R serves as privileged information during training to guide the image encoder toward mechanistic understanding.

3.2. Transcriptome-Guided Image Learning

We now establish the theoretical foundation for using transcriptomic guidance to improve image-based intervention prediction. Our goal is to learn a predictor $S_T(D|I, C)$ that estimates the drug (intervention) posterior from images and cell type alone. During training, however, we have access to paired transcriptomes R that capture perturbation-induced gene expression changes: responses that reflect pathway-level mechanisms not directly observable in morphology. The key question is whether and how incorporating transcriptomic information during training provably improve the learned image representations

We formalize this through a risk-bound analysis on the training distribution. Define $\mathcal{L}[q]$ as the expected log-loss of a predictor q that depends only on (I, C), where $\mathcal{P}$ is the joint distribution of observed training data. Consider a richer predictor $T(D|I, R, C)$ that incorporates RNA during training, producing a probability distribution over drugs D. The following proposition shows that training with transcriptomics provably tightens the risk bound.

Proposition 3.1 (Transcriptome-Guided Learning Bound)**.** *Given readout variables I and R conditioned on cell type C, consider a predictor $T : (I, R, C) \mapsto \Delta(\mathcal{D})$ that outputs a distribution over drugs D, where $\Delta(\mathcal{D})$ denotes the probability simplex. Define the marginalized predictor that depends only on (I, C):*

$$S_T(D|I, C) = \mathbb{E}_{R|I,C}[T(D|I, R, C)]$$

Let $\mathcal{L}[q] = \mathbb{E}_{(I,C,D)\sim P}[-\log q(D|I, C)]$ denote the log-loss risk for any predictor q that depends only on (I, C). Then $\mathcal{L}[S_T]$ is upper bounded by:

$$H(D|I, C) + \mathbb{E}_{(I,R,C)}[\mathrm{KL}(P(D|I, R, C)\|T(\cdot|I, R, C))],$$

where $H(D|I, C)$ is the conditional entropy and the second term is the training objective that can be directly optimized.

This bound establishes that fitting the conditional distribution $P(D|I, R, C)$ during training, and marginalizing over R at test time, yields a tighter risk bound on the training distribution than directly fitting $P(D|I, C)$ without RNA guidance. The key insight is that transcriptomic responses provide mechanistic information: conditioning on gene expression allows the teacher T to distinguish drugs that produce similar morphologies but engage different molecular pathways. When this knowledge is distilled into the image student through marginalization, the student learns representations that align with biological mechanisms rather than superficial visual patterns. The complete proof is provided in the Appendix.

The proposition suggests three design principles for our framework: (i) model $P(D|I, R, C)$ in a way that respects the causal structure of interventions, (ii) explicitly condition on cell type and dose to capture context-specific perturbation effects, and (iii) train the image student to approximate the marginalized posterior S_T. Note that the bound concerns the training distribution-it establishes that transcriptomic guidance improves learning on observed data but does not directly

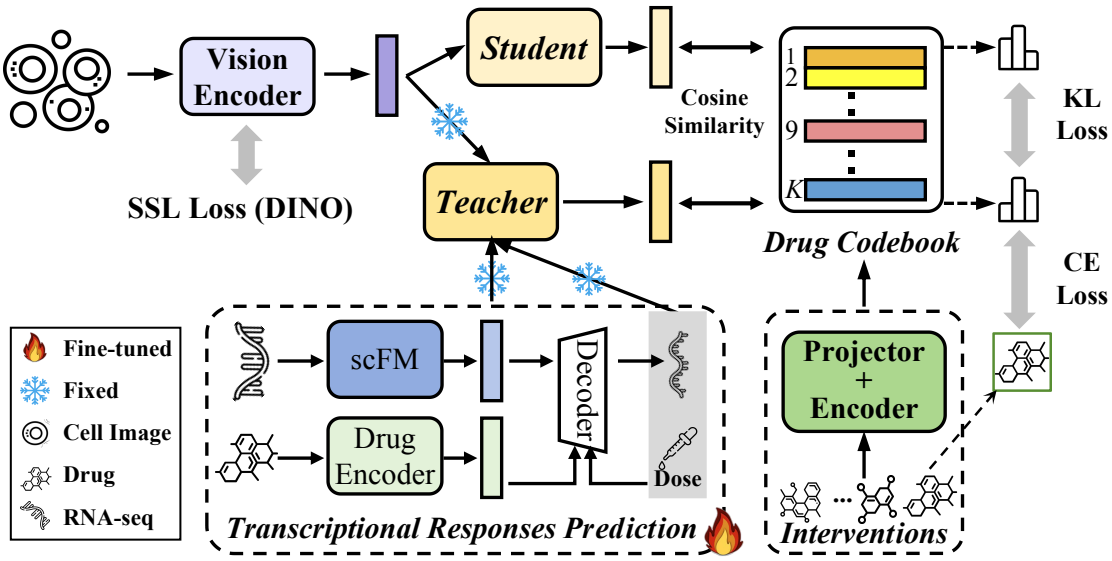


Figure 2. Framework overview. Our approach uses a chemistry-aware codebook and transcriptome-conditioned teacher to guide image student learning through distillation. Only images are required at test time.

guarantee generalization to unseen drugs. However, we hypothesize that the mechanistic features learned through this process will transfer to novel compounds with similar mechanisms, which we validate empirically in Section 4. We instantiate these principles in the following section.

3.3. Transcriptome-Guided Distillation Framework

We now present our transcriptome-guided distillation framework, which operationalizes the theoretical insights from Section 3.2. Figure 2 illustrates the overall architecture. Our framework consists of three key components. First, we construct a chemistry-aware codebook that organizes the drug space into K prototypes based on molecular representations, providing a structured target space where similar drugs are clustered. Second, we design a transcriptome-conditioned teacher that integrates paired images I, RNA-seq R, and intervention metadata (cell type C, doses δ_I, δ_R) to produce soft distributions over the codebook, capturing mechanistic signals from transcriptomics. Third, an image-only student learns to predict the same distributions from images alone, distilling the teacher's multimodal knowledge. We now detail each component.

3.3.1. Chemistry-aware Codebook

We construct a codebook that organizes drugs by molecular similarity. Given the set of K training drugs, we extract a molecular representation for each drug. These representations form the rows of a drug feature matrix, which we project into the same embedding space as the teacher and student encoders using a lightweight MLP followed by ℓ_2 normalization. This yields a prototype matrix $V \in \mathbb{R}^{K \times d}$, where each row $\mathbf{v}_k$ represents a drug prototype on the unit hypersphere that are fully learnable parameters updated through backpropagation during training. The chemistry-aware structure allows models to predict soft distributions over prototypes rather than hard drug identities, facilitating transfer to novel compounds through chemical similarity. While the prototypes adapt to encode mechanistic similarity as the teacher learns to predict codebook distributions from both images and transcriptomics.

3.3.2. Transcriptome-conditioned Teacher

The teacher model T_θ takes paired multi-modal observations and produces distributions over the codebook that reflect intervention-specific mechanistic signatures. A key design principle is that the teacher does not directly encode the drug's molecular structure; instead, it must infer pathway-level perturbation effects from the observed cellular responses (morphology and transcriptomics), and supervision from the codebook provides the link between inferred mechanisms and drug chemistry. This avoids shortcut learning where the teacher trivially matches drug identity.

To obtain robust encodings of cell type and dose-dependent transcriptomic effects, we employ a scFM [11], fine-tuned on perturbation response prediction. During fine-tuning, the model takes as input a cell's basal gene expression R_{basal}, a drug representation, and a dose value δ_r, and predicts the post-perturbation expression profile. Specifically, the scFM encodes R_{basal} and combines it with drug and dose embeddings to form a latent state, which is then decoded to reconstruct the perturbed transcriptome. After fine-tuning, we use the trained scFM encoder to extract two types of representations: (i) a cell-type encoding $\mathbf{h}_C$ from basal gene expression, providing stable biological context across perturbations; and (ii) a dose-conditioned transcriptome encoding that disentangles dose information from the observed RNA-

seq measurements. The scFM's training objective ensures these encodings capture interpretable perturbation-response structure. Implementation details are provided in Appendix.

The teacher then encodes three sources of information:

- **Image with dose**: We encode the cellular image I and then detach dose-specific information δ_I from the image features using a FiLM-style conditioning mechanism [19], producing a dose-aware image representation $\mathbf{h}_I$.
- **RNA with dose**: We encode the observed perturbed transcriptome R along with its associated dose δ_R, yielding $\mathbf{h}_R$. The dose disentanglement allows this encoding to separately represent the perturbation magnitude.
- **Cell type**: We use the basal gene expression encoding $\mathbf{h}_C$ from the scFM to represent cell-type-specific context.

Critically, by encoding image dose δ_I and RNA dose δ_R separately within their respective modality features, the teacher can handle weakly paired data where doses differ across modalities. Rather than treating mismatched doses as identical through binary label matching, the teacher learns to reason about dose-dependent effects explicitly, producing appropriate soft targets that reflect the actual dosage contexts.

For the codebook prediction, the teacher fuses the three encoded features through concat and a projection network f_t:

$$\mathbf{h}_t = f_t([\mathbf{h}_I||\mathbf{h}_R||\mathbf{h}_C]) \in \mathbb{R}^d.$$

It then computes similarity scores with all codebook prototypes, and produces a soft distribution over the codebook via temperature-scaled softmax:

$$P_t = \text{Softmax}\left(\frac{V \cdot \mathbf{h}_t}{\tau \cdot |\mathbf{h}_t|_2}\right) \in \mathbb{R}^K$$

where $V \in \mathbb{R}^{K \times d}$ is drug codebook and τ controls distribution sharpness. The teacher is trained with cross-entropy loss $\mathcal{L}_{teacher}$ against true drug labels, establishing the connection between observed cellular responses and drug chemistry through the codebook. The learned soft distributions, which encode fine-grained mechanistic similarities beyond hard labels, serve as targets for student distillation.

3.3.3. Image Student and Distillation Objective

The student learns to predict codebook distributions from images alone, without explicit cell type input. This design reflects the deployment scenario where only microscopy is available, and is justified by the fact that cell type information is inherently encoded in morphological features. The student uses the shared image encoder from the teacher, followed by a projection head to get the feature $\mathbf{h}_s$, producing codebook distributions P_s from images alone. We train the student to match the teacher's distributions via KL divergence:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{(I,R,C,\delta)}\left[\text{KL}(P_t|P_s\right].$$

Our framework is compatible with any self-supervised method. We apply $\mathcal{L}$ssl (e.g., DINO, MAE) to all images, paired and unpaired, leveraging abundant imaging data while learning general morphological features. The complete objective is:

$$\mathcal{L} = \mathcal{L}_{\text{distill}} + \alpha \cdot \mathcal{L}_{\text{teacher}} + \beta \cdot \mathcal{L}_{\text{ssl}}.$$

Our framework design operationalizes Proposition 3.1 by enabling the student to learn mechanistic features through two complementary mechanisms: the chemistry-aware codebook provides transfer to novel compounds via chemical similarity, while the transcriptome-conditioned teacher captures pathway-level perturbation effects that generalize beyond training drugs. By explicitly handling dose variations and avoiding identity shortcuts, the framework learns transferable representations from limited paired data. We empirically validate this generalization in Section 4.

Dataset	RxRx3	CPG-Pilot	CPG-12
Image	61,690	93,696	916,721
Paired RNA	3,682	1,883	22,066
# Drugs	1,662	302	30,340
# Overlaps	235	85	6,989
# Eval Intervention	736	260	121

Table 1. Dataset statistics and imaging-transcriptomics pairing. Numbers indicate image samples, RNA profiles, and compound overlap across modalities.

4. Experiments

We evaluate our framework on image-based drug discovery tasks to assess whether transcriptomic guidance improves (i) generalization to novel interventions and (ii) recovery of mechanistic drug-target relationships. We conduct experiments on three large-scale Cell Painting datasets paired with perturbational transcriptomics (L1000), spanning multiple cell lines and thousands of compounds. We compare against self-supervised baselines (MAE, DINO) and alignment-based methods (CLIP-style contrastive learning) to demonstrate the benefit of our approach.

4.1. Datasets

Imaging readouts. We use three Cell Painting datasets representing different scales and perturbation types: Cell Painting and RxRx.

1. **JUMP-Pilot (CPG-Pilot) [7]** profiles A549 and U2OS cells under both chemical and genetic perturbations with standardized Cell Painting protocols. The joint coverage of compounds and CRISPR perturbations enables evaluation of mechanism-level relationships and chemical-genetic concordance.
2. **CPG0012 (CPG-12) [3]** is a large-scale collection of ∼30,000 bioactive compounds in U2OS cells with consistent re-profiling. Its scale and diversity support eval-

Dataset	RxRx3				CPG-Pilot				CPG-12	
Backbone	ViT		CA-ViT		ViT		CA-ViT		ViT	
Metric	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
MAE	1.54	6.79	1.00	3.26	2.95	12.31	2.05	8.97	4.68	11.02
DINO	4.82	14.86	4.62	13.32	11.36	30.82	13.97	33.46	36.36	60.01
CL (D)	<u>5.16</u>	<u>15.63</u>	<u>5.93</u>	<u>16.40</u>	<u>14.61</u>	<u>36.66</u>	<u>15.17</u>	<u>38.46</u>	<u>40.04</u>	<u>65.07</u>
CL (R)	4.93	14.27	5.25	13.90	13.97	30.64	14.87	35.51	30.57	60.89
TIDE	**5.62**	**16.86**	**6.04**	**17.27**	**16.01**	**39.12**	**17.80**	**40.74**	**42.09**	**71.07**

Table 2. One-shot transfer to unseen interventions. We report Top-1 and Top-5 accuracy (%) for transfer to unseen genetic perturbations (CPG-Pilot, RxRx3) and unseen compounds (CPG-12). Best results highlighted are in bold, and second-best results are underlined

uation of cross-compound generalization and large-scale drug discovery applications.

3. **RxRx3-core (RxRx3) [15]** is a research-friendly subset of RxRx3 with six-channel imaging in HUVEC cells, covering both drug perturbations and CRISPR knockouts. The shared indexing across chemical and genetic conditions facilitates drug-gene association analyses.

Transcriptomic readouts. We pair imaging data with L1000 [25], a high-throughput perturbational transcriptomics platform that measures 978 landmark genes. L1000 provides bulk-cell gene expression profiles, matching the population-level nature of our imaging readouts. We extract L1000 profiles for compounds in the same cell lines as our imaging datasets, enabling transcriptome-guided image learning. To leverage L1000 data with our framework, we fine-tune a single-cell foundation model (scGPT [11]) on perturbation response prediction using L1000 measurements. Although scGPT is pretrained on single-cell atlases, we adapt it to bulk-cell L1000 data and training the model to predict perturbed expression given basal profiles, drug representations, and doses. This fine-tuned scFM provides the cell-type and dose encoders used in our teacher model (Section 3.3.2).

Table 1 summarizes the pairing between imaging and transcriptomic data. Importantly, the pairing is weakly paired: images and RNA-seq measurements share the same cell type and compound but often differ in dose, reflecting realistic data collection constraints. This weak pairing is a key challenge our framework addresses through explicit dose conditioning.

4.2. Tasks and Evaluation Metrics

We evaluate on two complementary tasks that assess generalization and mechanistic understanding:

4.2.1. Task 1: One-shot transfer to unseen intervention

This task evaluates whether learned image representations generalize to novel perturbations in a low-data regime. We follow a metric-based few-shot learning protocol: given a pretrained and frozen image encoder, we initialize a cosine similarity-based classifier from 1-shot support examples (one image per unseen intervention class), then optimize only the classification head on query images while keeping the encoder frozen. Features and class prototypes are L2-normalized, and predictions use temperature-scaled cosine similarity.

The specific unseen interventions differ by dataset characteristics. On CPG-Pilot and RxRx3, we evaluate transfer to unseen genetic perturbations (CRISPR knockouts), as these datasets contain limited chemical diversity but extensive genetic screens measured with the same imaging protocol. On CPG-12, which contains more than 30,000 diverse compounds, we evaluate transfer to unseen compounds by holding out a subset of drugs. Both settings assess the core capability of generalizing features to novel interventions.

4.2.2. Task 2: Unsupervised drug-target gene discovery

This task evaluates whether image representations capture mechanistic relationships by testing drug-target gene association without explicit supervision. Following Kraus et al. [15], we compute embeddings for compound-treated images and gene knockout images, score all drug-gene pairs by cosine similarity, and evaluate against curated ground-truth associations derived from each dataset's benchmark annotations, which is further filtered for high-confidence links. We report results averaged over 100 random seeds.

On RxRx3, compounds appear at multiple doses and may target multiple genes. We compute per-dose drug embeddings (averaged over replicate images), score associations at each dose, and report dose-averaged Average Precision (AP) and AUC. On CPG-Pilot, compounds are profiled at a single dose with a single primary target gene; we report AP and Hit@k. This task directly measures whether the learned representations reflect biological mechanisms, a key advantage of transcriptomic guidance.

4.3. Baselines and Implementation

We refer to our framework as TIDE (Transcriptome-Informed Distillation for image Encoding) and compare against two families of methods:

SSL Methods: MAE [12] and DINO [5], widely adopted for cellular imaging. These methods learn from images alone without transcriptomic or drug information.

Backbone	ViT				CA-ViT			
Dataset	RxRx3		CPG-Pilot		RxRx3		CPG-Pilot	
Metric	AP	AUC	AP	Hit@5	AP	AUC	AP	Hit@5
MAE	0.256	0.538	0.079	0.103	0.245	0.518	0.097	0.124
DINO	0.300	0.633	0.109	0.135	0.252	0.545	0.111	0.154
CL (D)	0.252	0.537	0.099	0.127	0.255	0.540	0.103	0.131
CL (R)	0.247	0.532	0.105	0.134	0.250	0.540	0.105	0.135
TIDE	**0.317**	**0.640**	**0.119**	**0.149**	**0.301**	**0.612**	**0.120**	**0.171**

Table 3. Unsupervised drug-target gene discovery. We report AP and retrieval metrics without explicit supervision on drug-target associations. Best results in bold.

Alignment-based contrastive learning: We implement CLIP-style approaches that align image embeddings to auxiliary modalities: CL(D): Following recent phenomolecular retrieval work [10, 23], aligns images to drug molecular representations; CL(R): Aligns images and transcriptomic readouts based on shared drug perturbations [16, 27]. To ensure fair comparison, we evaluate hybrid variants that combine contrastive alignment with DINO self-supervision, which typically improve performance.

Vision Encoder. All methods use the same backbone architectures for fair comparison. We evaluate two vision transformers: a standard ViT [9] and a Channel-Agnostic ViT (CA-ViT) [14], a cellular imaging-specific architecture that achieves comparable performance to ViT. Model sizes are matched to dataset scale: ViT-Small for CPG-Pilot and RxRx3, ViT-Base for CPG-12 (CA-ViT-Base is omitted on CPG-12 due to computational cost).

All methods train for 400 epochs with matched augmentations and optimizers within each method family. We select checkpoints via k-nearest neighbor (k=20) evaluation on a held-out validation split (10 images per perturbation class). For Task 1, we split interventions (not samples) into train/val/test to ensure true generalization evaluation. For Task 2, we use all available data to compute embeddings and evaluate on curated drug-target associations. Detailed hyperparameters and split information are provided in Appendix.

4.4. Main Results

We design experiments to answer two research questions that validate TIDE's core contributions:

- **RQ1**: Does transcriptomic guidance improve generalization to unseen interventions?
- **RQ2:** Do learned image representations capture biologically meaningful mechanisms?

Table 2 presents Top-1 and Top-5 accuracy on one-shot classification. All results are averaged over 50 runs. TIDE achieves the best performance across all settings, substantially outperforming all baselines. CL(D) attains the second-best results among baselines, which together with TIDE's performance validates the importance of incorporating perturbation information for model training. CL(R) performs slightly better than SSL baselines but falls short of drug-based alignment, as the limited availability of paired transcriptomic data constrains its training signal compared to CL(D)'s ability to align with molecular representations for all drugs. Among SSL methods, DINO substantially outperforms MAE across all datasets. Although training for the same number of epochs, DINO's use of local and global crop augmentations results in higher computational cost per epoch, effectively providing richer training signal. Comparing backbones, ViT and CA-ViT show comparable overall performance on RxRx3 and CPG-Pilot across most methods. However, with TIDE CA-ViT demonstrates superior results, as its channel-agnostic patchification design achieves better representational capacity under sufficient training. TIDE's consistent improvements across datasets and backbones demonstrate that transcriptomic guidance enables image models to learn mechanistically grounded features that generalize to novel interventions.

Table 3 presents performance on unsupervised drug-target gene discovery. Our model significantly outperforms all baselines on both datasets, although overall performance remains modest across all methods. As reported in [7], cross-modal perturbation pairing signals are inherently weak, only slightly above random. Most baseline results on RxRx3 are only marginally better than random, underscoring the task's difficulty and demonstrating that our model learns mechanistic features beyond visual supervision alone. Unlike the one-shot transfer task, CL(D) shows no clear advantage over DINO or CL(R) in this setting. This reflects the nature of drug-target discovery that while molecular structure alignment helps CL(D) learn drug identity and chemical similarity, identifying biological targets requires pathway-level mechanistic understanding that chemical fingerprints do not directly encode. In contrast, even limited transcriptomic signals or morphological patterns learned through augmentation-rich SSL may capture aspects of cellular response more relevant to target identification, explaining their comparable performance to CL(D) despite the differences observed in Task 1. The substantial improvement TIDE achieves on this challenging task, where baselines barely exceed random performance, confirms that our framework successfully distills pathway-level mechanistic knowledge from transcriptomics into image representations, enabling identification of biologically meaningful drug-target relationships from morphology alone.

4.5. Ablation Studies and Analysis

We conduct ablation studies to dissect the contributions of TIDE's design components and validate key hypotheses.

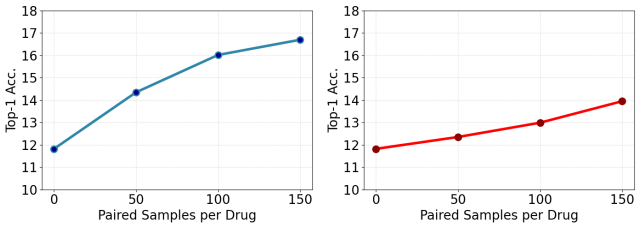
4.5.1. Effect of Transcriptomic Data

To validate whether richer transcriptomic guidance improves performance, we separately evaluate TIDE on A549 and U2OS cell lines in CPG-Pilot. These two cell lines have different overlap with L1000 gene expression profiles: A549

Task	one-shot transfer				drug-target discovery			
Cell Line	U2OS		A549		U2OS		A549	
Metric	Top-1	Top-5	Top-1	Top-5	AP	Hit@5	AP	Hit@5
DINO	12.8	28.9	12.1	28.2	0.084	0.111	0.133	0.159
TIDE	15.9	36.9	16.3	39.8	0.092	0.120	0.146	0.178

Table 4. Performance on CPG-Pilot separately for U2OS and A549.

has more paired compounds than U2OS. By comparing performance across these two settings, we can assess whether increased transcriptomic coverage leads to better mechanistic learning. Table 4 presents one-shot transfer accuracy and drug-target discovery performance for both cell lines on ViT. Examining the DINO baseline performance, we observe that both cell lines achieve comparable results on one-shot transfer, with U2OS even slightly outperforming A549. However, after training with TIDE, A549 that has more paired transcriptomic data, achieves better performance and larger improvements over its DINO baseline. On the drug-target discovery task, while A549 demonstrates advantages on this task even with DINO, it shows more improvements with TIDE compared to U2OS. These findings validate the importance of transcriptomic information: when sufficient paired data is available, TIDE can more effectively distill mechanistic knowledge from transcriptomics, leading to stronger gains in both generalization and mechanistic understanding.



(a) TIDE with fine-tuned scFM (b) TIDE with pretrained scFM

Figure 3. Effect of paired sample size and scFM fine-tuning on one-shot transfer performance of TIDE.

4.5.2. Sample Efficiency and Effect of scFM Fine-tuning

To understand how the quantity of paired data per drug affects performance and validate the necessity of scFM fine-tuning on perturbation response prediction, we vary the number of paired image-transcriptome samples per drug from 0 to 150 on CPG-Pilot. At 0 paired samples, TIDE reduce to DINO. We compare TIDE using a fine-tuned scFM encoder versus one with pretrained-only weights. Figure 3 presents the results. With fine-tuned scFM, TIDE shows substantial performance improvement as paired samples increase. This demonstrates that the fine-tuned encoder effectively extracts dose-specific and cell-context information from transcriptomic measurements, and more paired samples provide richer supervision for distillation. In contrast, TIDE using pretrained-only scFM shows only marginal improvement, despite also benefiting from the additional training images that accompany more paired samples. The minimal gain isolates the effect of increased image data alone and confirms that

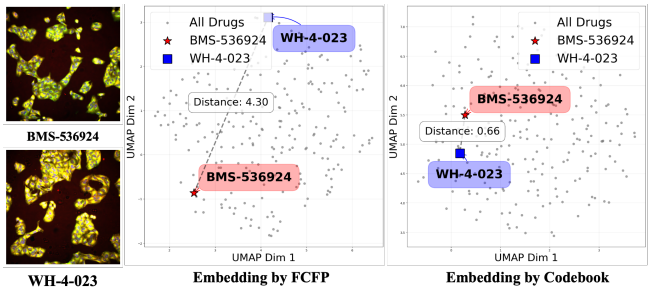


Figure 4. Left: Microscopy images of A549 cells treated with BMS-536924 and WH-4-023. Right: Embedding visualization by FCFP fingerprint v.s. learned codebook embedding

without fine-tuning on perturbation response prediction, the scFM encoder cannot effectively disentangle dose effects or encode cell-type-specific basal states.

4.5.3. Analysis and Visualization

To understand what TIDE's learned codebook encodes, we visualize drug representations in chemical fingerprint space versus learned codebook space. Figure 4 presents a case study of two compounds, BMS-536924 and WH-4-023, profiled on A549 cells in CPG-Pilot. While the morphological images (left) show complex cellular responses with no obvious shared visual patterns to the human, the two drugs share a common mechanism of action: both are ATP-competitive tyrosine kinase inhibitors that modulate downstream signaling pathways. The middle panel shows UMAP projection of all CPG-Pilot drugs based on FCFP fingerprints [21]. BMS-536924 and WH-4-023 are distant in chemical space, reflecting their different molecular structures, while in the learned codebook embedding space the two drugs are positioned much closer, indicating that the codebook has learned to group them by their shared mechanistic properties despite structural differences. This demonstrates that through end-to-end training with transcriptomic guidance, the codebook moves beyond chemical similarity to capture biological mechanism similarity, organizing drugs by how they perturb cellular pathways rather than merely by molecular structure.

5. Conclusion

We presented a transcriptome-guided distillation framework that uses perturbational RNA-seq as privileged information to guide image representation learning for drug discovery. By structuring supervision through a chemistry-aware codebook and explicitly conditioning on dose and cell type, our approach handles weakly paired data and avoids identity-based shortcuts that limit existing alignment methods. We provided theoretical analysis showing that transcriptomic guidance tightens the risk bound for image-based prediction, and demonstrated improvements over SSL and contrastive baselines on one-shot transfer to unseen interventions and unsupervised drug-target gene discovery.

Acknowledgments

This work was funded in part by the National Science Foundation under award number IIS-2145625 and by the National Institutes of Health under awards number R01AI188576 and R01CA301579.

References

[1] Ihab Bendidi, Yassir El Mesbahi, Alisandra K Denton, Karush Suri, Kian Kenyon-Dean, Auguste Genovesio, and Emmanuel Noutahi. A cross modal knowledge distillation & data augmentation recipe for improving transcriptomics representations through morphological features. *arXiv preprint arXiv:2505.21317*, 2025. 1, 2

[2] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher 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, 2016. 1, 2

[3] Mark-Anthony Bray, Sigrun M Gustafsdottir, Mohammad H Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L Sokolnicki, Joshua A Bittker, Nicole E Bodycombe, Vlado Dančík, Thomas P Hasaka, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *Gigascience*, 6(12):giw014, 2017. 2, 5

[4] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 6

[6] Safiye Celik, Jan-Christian Huetter, Sandra Melo, Nathan Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta Fay, Berton Earnshaw, and Imran S Haque. Biological cartography: Building and benchmarking representations of life. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022. 1

[7] Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024. 1, 2, 5, 7

[8] Jiayuan Chen, Thai-Hoang Pham, Yuanlong Wang, and Ping Zhang. Integrating biological knowledge for robust microscopy image profiling on de novo cell lines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22846–22856, 2025. 1

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[10] Philip Fradkin, Puria Azadi Moghadam, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How molecules impact cells: Unlocking contrastive phenomolecular retrieval. *Advances in Neural Information Processing Systems*, 37:110667–110701, 2024. 1, 7

[11] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature methods*, 21(8):1481–1491, 2024. 1, 2, 4, 6

[12] 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*, pages 16000–16009, 2022. 1, 6

[13] Kian Kenyon-Dean, Zitong Jerry Wang, John Urbanik, Konstantin Donhauser, Jason Hartford, Saber Saberian, Nil Sahin, Ihab Bendidi, Safiye Celik, Marta Fay, et al. Vitally consistent: Scaling biological representation learning for cell microscopy. *arXiv preprint arXiv:2411.02572*, 2024. 1, 2

[14] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768, 2024. 1, 2, 7

[15] Oren Kraus, Federico Comitani, John Urbanik, Kian Kenyon-Dean, Lakshmanan Arumugam, Saber Saberian, Cas Wognum, Safiye Celik, and Imran S Haque. Rxrx3-core: Benchmarking drug-target interactions in high-content microscopy. *arXiv preprint arXiv:2503.20158*, 2025. 1, 2, 6

[16] Gang Liu, Srijit Seal, John Arevalo, Zhenwen Liang, Anne E Carpenter, Meng Jiang, and Shantanu Singh. Learning molecular representation in a cell. *ArXiv*, pages arXiv–2406, 2024. 1, 2, 7

[17] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *Nature communications*, 15(1):1594, 2024. 1

[18] Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3): 531–540, 2024. 2

[19] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[20] Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14): 2559–2575, 2022. 1

[21] David Rogers and Mathew Hahn. Extended-connectivity fin-

gerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010. 8

[22] Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024. 1

[23] Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023. 1, 2, 7

[24] 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*, pages 2023–08, 2023. 2

[25] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017. 1, 2, 6

[26] 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, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. 2

[27] Chenyu Wang, Sharut Gupta, Caroline Uhler, and Tommi Jaakkola. Removing biases from molecular representations via information maximization. *arXiv preprint arXiv:2312.00718*, 2023. 1, 2, 7

[28] Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *BioRxiv*, pages 2025–02, 2025. 1, 2

[29] Shuangjia Zheng, Jiahua Rao, Jixian Zhang, Lianyu Zhou, Jiancong Xie, Ethan Cohen, Wei Lu, Chengtao Li, and Yuedong Yang. Cross-modal graph contrastive learning with cellular images. *Advanced Science*, 11(32):2404845, 2024. 1, 2