RIGHT FOR THE RIGHT REASONS? TESTING WHETHER CELLULAR PROFILES EXPLAIN GENE–DRUG RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-based cellular profiling records how cells change after a gene is modified or a drug is applied. Comparing these responses can reveal gene–drug relationships and help prioritize candidates for drug discovery and repurposing. However, a high retrieval score does not show why a model recovered a known relationship. The model may use the cellular response, favor genes or drugs with many recorded partners, or exploit the relationship graph, in which genes and drugs are nodes and recorded relationships are edges. We introduce a novel controlled evaluation framework that asks two questions: does a model recover known gene–drug relationships better than randomized relationships, and do cellular features or graph connections support that recovery? The framework combines held-out retrieval with degree-preserving randomized relationships. It separates cellular features, which describe observed cell responses, from graph topology, the pattern of recorded gene–drug connections. We evaluate RxRx3-Core, CPJUMP1, MOTIVE-CRISPR, and a private angiogenesis dataset. CPJUMP1 supports recovery in both directions, RxRx3-Core is stronger from drugs to genes, and the angiogenesis data have no consistently dominant direction across metrics. MOTIVE-CRISPR supports nonlinear feature-based retrieval, while its graph controls expose the same structural cautions. Among the tested models, adding nonlinearity or graph message passing therefore does not consistently produce stronger or more informative alignment. The main contribution is a practical way to determine whether a result is supported by cellular information, graph structure, or both. This provides a clearer basis for selecting gene–drug candidates for experimental follow-up.

1 INTRODUCTION

In image-based cellular profiling, researchers change a gene's activity or treat cells with a drug to understand biological processes and identify possible treatments. Each intervention is called a *perturbation*. Microscopy captures the resulting cellular *phenotype*, including cell shape and internal structures (Figure 1). These images are summarized as numerical profiles, allowing researchers to compare responses across many genes and drugs (Bray et al., 2016; Caicedo et al., 2017).

These comparisons can help connect genes to drugs. If changing a gene and applying a drug produce similar cellular responses, they may affect a shared biological process. Such links can help explain drug action and guide the search for treatments (Rohban et al., 2022). Computational methods use cellular profiles to rank possible gene–drug relationships and test whether known partners receive high ranks (Chandrasekaran et al., 2024; Kraus et al., 2025). Some methods also learn from previously recorded relationships. MOTIVE, for example, compares predictions based on cellular profiles, the network of known relationships, or both (Arevalo et al., 2024).

However, a model can recover known relationships for reasons other than the information in cell images. Some genes and drugs have many more recorded partners than others, so favoring these frequently annotated entities can improve prediction scores (Aiyappa et al., 2025). This creates a difficulty when interpreting performance: the score alone does not show whether the model benefits from cellular responses or from patterns in the recorded relationships. Comparisons of models with and without image information are valuable, but they answer a different question from whether the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

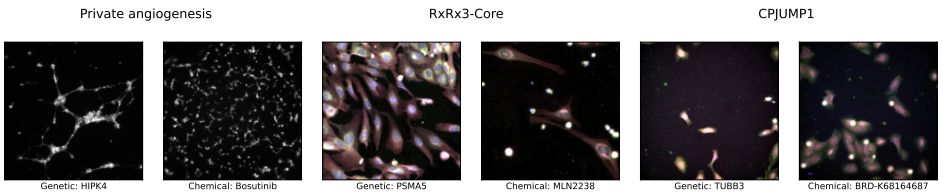

Figure 1: Genetic (left) and chemical (right) perturbations from the three primary datasets. RxRx3-Core and CPJUMP1 show cellular structures; the private angiogenesis assay shows vessel-like networks.

model recovers known relationships better than randomized ones. We need both comparisons to assess alignment, together with matched changes to the graph to examine the contribution of its connections.

To fill this gap, we introduce a novel framework that asks two questions: *Does the model recover known gene–drug relationships better than randomized relationships?* and *How do cellular features and relationship topology contribute to retrieval?* The first provides an operational criterion for alignment; the second examines which information supports it. We apply the framework to RxRx3-Core (Kraus et al., 2025), CPJUMP1 (Chandrasekaran et al., 2024), MOTIVE-CRISPR (Arevalo et al., 2024), and a private angiogenesis dataset. Our aim is to clarify when cellular profiles help recover gene–drug relationships, without treating successful prediction as proof of visual similarity or direct drug targeting. Our contributions are:

- An evaluation framework that assesses gene–drug alignment by comparing held-out retrieval of known relationships with randomized relationships.
- An evaluation of how cellular features and graph connections help or hurt the recovery of gene–drug relationships.
- A study across four datasets and five representations, using linear and nonlinear non-graph models as references and qualitative views to interpret differences in retrieval across settings.

2 Related Work

We review how cell images are represented, how genes and drugs are matched, and how these predictions are evaluated.

2.1 Cellular Profiling and Representations

Cell Painting makes several cellular structures visible through fluorescent staining (Bray et al., 2016). Each image is converted into a numerical feature vector, called a *profile*. Its values describe measured properties such as cell shape, intensity, and texture (Caicedo et al., 2017), or features learned through weak supervision, self-supervision, and masked-image training (Moshkov et al., 2024; Kim et al., 2025; Kraus et al., 2024). Aggregation and normalization also affect the biological relationships recovered from these features (Celik et al., 2024). Our study uses existing image representations and trains downstream alignment models, rather than new image encoders. Reliable profiles are a necessary starting point, but do not guarantee that genetic and chemical responses can be matched.

2.2 Genetic–Chemical Matching and Cross-Modal Alignment

One way to connect genes and drugs is to compare their image-derived feature vectors and look for similar cellular responses (Rohban et al., 2022). CPJUMP1 and RxRx3-Core test whether such comparisons recover known gene–drug relationships (Chandrasekaran et al., 2024; Kraus et al., 2025). Learned models can also predict these relationships, although a transformer-based study reports overfitting and difficulty predicting previously unseen targets (Iyer et al., 2025). Other approaches learn by matching cell images to chemical structures, as in CLOOME and MIGA (Sanchez-Fernandez et al., 2023; Zheng et al., 2024), or to text descriptions of perturbations, as in CellCLIP (Lu et al.,

2025). MICON instead learns how a compound changes the cellular appearance represented in an image (Yu et al., 2026). These methods learn from different sources of information, so success in one task does not guarantee better gene–drug matching. For example, MorphoCLIP reports more consistent representations of repeated treatments without reliable improvements in gene–drug matching (Ilyosbekov et al., 2026). Understanding gene–drug prediction therefore requires examining both the cellular features and the known relationships used during training, and our framework represents the first solution addressing this need.

2.3 Graph-Based Relationship Learning

MOTIVE compares morphology-based GNNs with feature-only predictors under random and entity-held-out splits, and includes topology-based controls in its random-split analysis (Arevalo et al., 2024). Its graphs include a bipartite version and versions with additional gene–gene and compound–compound edges. Self-supervision also has precedents in relationship learning: MVGCN initializes biomedical bipartite embeddings through SSL before graph-based link prediction (Fu et al., 2022), while SHGCL-DTI adds a contrastive task to semi-supervised prediction on a heterogeneous network (Yao et al., 2023). More generally, GRACE learns a consistent representation of each node across two versions of a graph, with some connections and input features randomly removed (Zhu et al., 2020). These studies establish precedents for morphology-based graph learning and graph SSL. Models using only the pattern of known gene–drug connections can achieve prediction scores similar to models that also use biological features (Durán et al., 2018). Good graph-based prediction therefore does not by itself show that cellular features helped, while our framework is able to quantify this aspect.

2.4 Evaluating Predictions and the Role of Cellular Features

Evaluation begins by checking whether image-derived features detect treatment effects and give consistent results across repeated treatments (Kalinin et al., 2025). MorphoHELM extends this assessment to differences between experiments and prediction tasks (Hayir et al., 2026). Reliable measurements alone are not enough: predicting relationships for previously unseen drugs or genes is a different test from predicting new links among familiar ones (Pahikkala et al., 2015). Understanding the resulting scores also requires examining what the model uses. GNN studies compare the contribution of node features with that of graph connections (Gomes et al., 2025), while link-prediction research shows that frequently connected nodes can receive an advantage in evaluation (Aiyappa et al., 2025). Curveball provides a reference for this advantage by randomly changing partners while keeping each node's number of connections unchanged (Strona et al., 2014; Carstens et al., 2018). It preserves these counts, not the entire graph structure. Building on this work, we assess both whether gene–drug prediction exceeds this reference and whether the correct cellular features improve prediction. These comparisons answer different questions and in our framework they are interpreted together.

3 Methods

3.1 Problem Formulation

Let $\mathcal{G}$ and $\mathcal{C}$ denote the sets of genes and compounds. Their recorded relationships form a bipartite graph $\mathcal{B} = (\mathcal{G} \cup \mathcal{C}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{G} \times \mathcal{C}$ connects genes to compounds. Repeated applications of each intervention are summarized by a cellular feature vector: $\mathbf{x}_g \in \mathbb{R}^{d_G}$ for gene g and $\mathbf{y}_c \in \mathbb{R}^{d_C}$ for compound c. The modality dimensions may differ, as they do in MOTIVE-CRISPR Arevalo et al. (2024); every model therefore uses modality-specific preprocessing or input projections. An alignment model maps these vectors into a shared embedding space. Given a gene query, retrieval ranks candidate compounds by cosine similarity, aiming to place its recorded partners first; compound-to-gene retrieval reverses these roles. We evaluate the directions separately because their candidate sets and numbers of partners differ. Our operational criterion for alignment is better held-out recovery of known relationships than randomized relationships. This is an evaluation criterion, not a shared training objective. Pairs outside $\mathcal{E}$ are unrecorded, not confirmed negatives, and successful retrieval does not establish visual similarity or causal targeting.

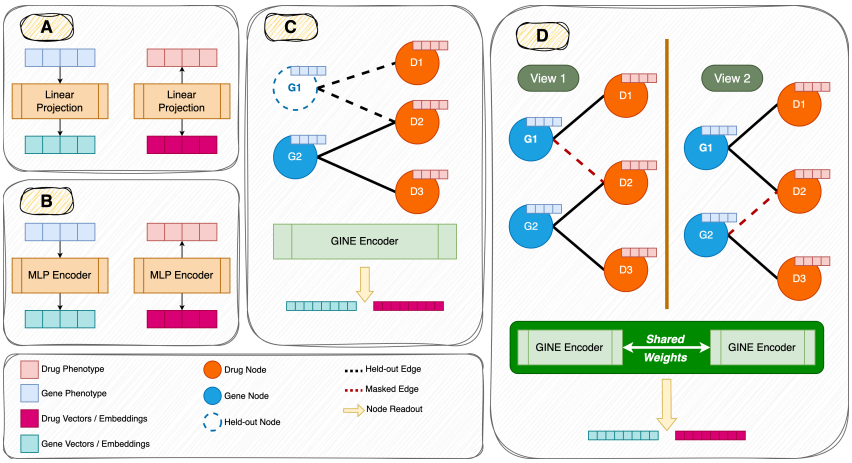

Figure 2: Alignment architectures. (A) CCA learns paired linear projections. (B) The MLP learns supervised gene and drug encoders. (C) Sup-GINE combines message passing with pair supervision. (D) SSL-GINE contrasts each node across two masked graph views. The output vectors are used for cosine retrieval.

3.2 Alignment Architecture

Figure 2 showcases the four alignment modes used in our architecture. Canonical correlation analysis (**CCA**, panel A) learns linear gene and drug projections that maximize correlation between recorded training pairs (Hotelling, 1936). The two-tower multilayer perceptron (**MLP**, panel B) instead uses separate nonlinear encoders, trained with multi-positive InfoNCE to favor recorded partners in both directions (van den Oord et al., 2018). Neither model performs graph message passing.

The graph modes represent genes and drugs as feature-bearing nodes joined by recorded training relationships. GINE extends the graph isomorphism network by including edge features in each neighbor update (Xu et al., 2019; Hu et al., 2020); our edges use one constant feature. Two GINE layers and a linear node readout produce normalized embeddings. **Sup-GINE** (panel C) uses the same supervised pair loss as the MLP. **SSL-GINE** (panel D) creates two views by masking features and removing edges, then applies same-node InfoNCE across views (Zhu et al., 2020). Genes are contrasted only with genes and compounds only with compounds, so gene–compound pairs are not treated as negatives. SSL-GINE uses no pair labels or supervised fine-tuning. Test queries retain their features but have no training edges; in this cold-start setting, only candidate nodes receive information from graph neighbors.

3.3 Evaluation Protocol

Retrieval metrics. AUROC measures how often a recorded partner ranks above an unrecorded candidate. Let A_q be the AUROC for query q and Q the number of queries. We average queries equally, so highly annotated entities do not dominate. Excess AUROC subtracts the mean of B degree-preserving null AUROCs from the observed value:

$$A = Q^{-1} \sum_{q} A_q, \qquad \Delta = A_{\text{obs}} - B^{-1} \sum_{b} A_{\text{null}}^{(b)}. \tag{1}$$

MRR measures the rank of the first partner, while recall@10 measures the fraction of partners in the top ten.

Graph attribution controls. We retrain each graph model with zero or shuffled features to test cellular information, and with disabled messages or rewired edges to test graph connections. Training pairs and held-out labels stay fixed, so paired changes in raw AUROC isolate each intervention. Rewiring preserves every training-node degree but intentionally changes neighborhoods; it therefore tests dependence on the particular graph connections, including structure beyond node degree.

Table 1: Inputs used for retrieval. Gene and drug counts include entities with recorded partners; d_G/d_C gives the gene and compound feature dimensions.

Input	Representation	d_G/d_C	Genes/Drugs	Relations
RxRx3-Core	OpenPhenom	384/384	351/779	2,921
CPJUMP1	CellProfiler	904/904	130/260	260
Angio-Features	Handcrafted	99/99	449/244	5,931
Angio-Embed	OpenPhenom	384/384	449/244	5,931
MOTIVE-CRISPR	CellProfiler	259/737	4,070/3,176	25,810

4 Experiments and Results

We first examine retrieval with linear and nonlinear non-graph models. We then assess the contributions of topology and cellular features through matched changes within each graph model. This separates evidence of alignment from evidence about what supports it.

4.1 Datasets and Training Setup

We use four collections of cell images through five representations (Table 1). RxRx3-Core and CPJUMP1 contain images of cells after gene changes or drug treatment, represented by OpenPhenom embeddings and CellProfiler measurements, respectively. Our private dataset records how these treatments change the networks formed by cells. We call its handcrafted measurements Angio-Features and its OpenPhenom embeddings Angio-Embed; these are two representations of the same dataset, not independent datasets. MOTIVE-CRISPR contributes Cell Painting profiles with different native gene and compound feature dimensions. Retrieval includes only entities with recorded partners. All models use five query-held-out folds, with preprocessing fitted on training entities only. The completed retrieval panel uses five MLP seeds and 1,000 degree-preserving null draws. The graph-attribution analysis applies the same interventions, including three shuffles or rewires per seed, to all five inputs, using 40 paired seeds per input. Training uses an NVIDIA RTX 4090. Values summarize complete out-of-fold predictions within each seed; folds are not treated as independent repetitions.

4.2 Non-Graph Alignment Across Datasets

Table 2 shows how well CCA and MLP rank recorded partners, and how much of that result exceeds randomized relationships. In **CPJUMP1**, both models recover partners in both directions: excess AUROC ranges from 0.111 to 0.145. The two models therefore agree that the profiles contain information beyond the degree-preserving null. In **RxRx3-Core**, the direction matters. MLP excess AUROC is 0.068 when drugs are queries, compared with 0.017 when genes are queries; recall@10 likewise falls from 0.247 to 0.082. In **Angiogenesis**, early partner ranks look strong for both representations. Angio-Embed MLP, for example, reaches 0.479 gene-to-drug recall@10 and 0.920 MRR. But its excess AUROC is only 0.005, and the corresponding Angio-Features value is 0.006. Those good ranks provide much less evidence of recovery beyond randomized relationships. In **MOTIVE-CRISPR**, MLP has excess AUROC of 0.056–0.059 across the two directions, compared with 0.007–0.016 for CCA. MLP therefore detects a signal beyond the null even though few recorded partners appear in the first ten ranks. Across the inputs, early ranks and excess AUROC give different, complementary views of alignment.

4.3 Contribution of Relationship Topology

Table 3 tests graph structure in two ways: we remove neighbor messages, then separately rewire the training edges while preserving node degrees. When **removing messages** on CPJUMP1, Sup-GINE gene-to-drug AUROC rises from 0.533 to 0.664, while SSL-GINE falls from 0.523 to 0.501. The same opposing pattern appears in the other direction. Removing messages also raises SSL-GINE scores on RxRx3-Core. On MOTIVE-CRISPR, it raises SSL-GINE from below chance to near chance in both directions, but lowers Sup-GINE in both directions. Sup-GINE also gets worse on Angio-Features when genes are queries. Thus, neighbor messages help some models and hurt others. When **changing the connections**, rewiring raises Sup-GINE scores on CPJUMP1 in both directions, but lowers both models on Angio-Features when genes are queries. This test retains message passing;

Table 2: CCA and MLP retrieval measured by excess AUROC, recall@10, and MRR. MLP values are mean ± SD over five seeds.

Input	Dir.	Excess AUROC		Recall@10		MRR	
		CCA	MLP	CCA	MLP	CCA	MLP
RxRx3-Core	G→D	.028	.017±.005	.032	.082±.003	.073	.221±.008
	D→G	.057	.068±.004	.057	.247±.004	.085	.403±.002
CPJUMP1	G→D	.111	.124±.010	.062	.080±.005	.049	.066±.011
	D→G	.130	.145±.008	.131	.171±.008	.067	.079±.003
Angio-Features	G→D	.005	.006±.001	.083	.418±.009	.214	.920±.002
	D→G	−.002	.032±.006	.041	.229±.008	.219	.739±.011
Angio-Embed	G→D	−.009	.005±.002	.105	.479±.017	.249	.920±.003
	D→G	−.005	.015±.004	.053	.236±.006	.235	.755±.010
MOTIVE-CRISPR	G→D	.007	.056±.002	.003	.048±.002	.012	.077±.003
	D→G	.016	.059±.004	.004	.031±.002	.013	.055±.002

Table 3: Raw AUROC under the original graph, disabled neighbor messages, and degree-preserving rewired edges. Values are mean ± SD over 40 seeds for every input; rewired results average three graphs per seed.

Input	Dir.	Sup-GINE			SSL-GINE		
		Original	No messages	Rewired	Original	No messages	Rewired
RxRx3-Core	G→D	.542±.011	.550±.009	**.553±.009**	.439±.007	**.504±.009**	.431±.007
	D→G	**.701±.006**	.675±.011	.701±.006	.411±.009	**.505±.008**	.389±.007
CPJUMP1	G→D	.533±.016	**.664±.010**	.632±.013	**.523±.016**	.501±.020	.501±.017
	D→G	.558±.016	**.662±.010**	.636±.009	**.530±.014**	.506±.017	.504±.018
Angio-Features	G→D	**.722±.041**	.690±.022	.519±.046	**.503±.030**	.498±.006	.380±.037
	D→G	.651±.042	.641±.014	**.653±.035**	.427±.023	**.499±.006**	.317±.020
Angio-Embed	G→D	.683±.032	**.735±.029**	.493±.033	.423±.026	**.495±.006**	.308±.019
	D→G	**.649±.040**	.626±.017	.647±.033	.422±.021	**.499±.009**	.289±.010
MOTIVE-CRISPR	G→D	.738±.022	.671±.003	**.780±.012**	.342±.004	**.509±.003**	.338±.004
	D→G	.749±.005	.555±.004	**.761±.003**	.407±.005	**.504±.004**	.395±.004

it changes which nodes exchange information. For Angio, however, the rewired graphs keep about 74% of their original edges, so the intervention is limited. Taken together, these controls show whether scores depend on messages or specific edges, but no single graph effect holds across all inputs.

4.4 Contribution of Cellular Features

Table 4 tests cellular profiles while keeping the graph fixed. We either reassign profiles among entities of the same type or replace them with zeros. When **reassigning profiles**, SSL-GINE loses AUROC in both directions on RxRx3-Core and Angio-Features, and from drugs to genes on Angio-Embed. Correct profile identity therefore matters in these settings. Sup-GINE behaves differently on the two Angio representations: reassignment improves its AUROC. On MOTIVE-CRISPR, reassignment instead lowers Sup-GINE AUROC in both directions, while SSL-GINE changes little from genes to drugs. A high graph score does not always mean that the model uses the right cellular profile. When **removing profiles** entirely, Sup-GINE improves in every input and direction. Its gene-to-drug AUROC rises from 0.533 to 0.899 on CPJUMP1 and from 0.738 to 0.835 on MOTIVE-CRISPR. These gains cannot come from information in the removed profiles. On CPJUMP1, every drug has one recorded gene partner. Holding out that gene leaves the drug isolated in the training graph, which lets the model rank such drugs without their cell images. By contrast, zero features reduce SSL-GINE AUROC on RxRx3-Core, both Angio representations, and MOTIVE-CRISPR; CPJUMP1 is the exception. Overall, a score that improves without features can reflect the graph rather than a better match between cellular phenotypes.

Table 4: Raw AUROC with correct, zero, or reassigned cellular features. Values are mean ± SD over 40 seeds for every input; shuffled results average three feature assignments per seed.

Input	Dir.	Sup-GINE			SSL-GINE		
		Original	Zero	Shuffled	Original	Zero	Shuffled
RxRx3-Core	G→D	.542±.011	**.673±.008**	.513±.007	**.439±.007**	.388±.004	.430±.005
	D→G	.701±.006	**.726±.014**	.634±.008	**.411±.009**	.295±.007	.403±.006
CPJUMP1	G→D	.533±.016	**.899±.025**	.510±.011	.523±.016	**.734±.141**	.501±.009
	D→G	.558±.016	**.838±.008**	.498±.010	.530±.014	**.679±.123**	.501±.011
Angio-Features	G→D	.722±.041	**.937±.013**	.747±.023	**.503±.030**	.299±.072	.452±.019
	D→G	.651±.042	**.913±.010**	.748±.028	**.427±.023**	.194±.033	.398±.017
Angio-Embed	G→D	.683±.032	**.937±.014**	.714±.021	.423±.026	.298±.057	**.425±.016**
	D→G	.649±.040	**.914±.009**	.679±.023	**.422±.021**	.193±.029	.377±.017
MOTIVE-CRISPR	G→D	.738±.022	**.835±.011**	.716±.024	.342±.004	.231±.013	**.342±.003**
	D→G	.749±.005	**.789±.013**	.713±.006	**.407±.005**	.264±.010	.396±.003

5 Discussion

The quantitative results reveal a gap between ranking recorded partners and establishing phenotype alignment. CPJUMP1 shows clear recovery beyond the degree-preserving null, whereas Angio can rank partners early with little excess AUROC. The graph controls add another caution: scores can rise when cellular features or neighbor messages are removed. Our goal is to learn when cellular responses support a gene–drug relationship, rather than merely predict that a relationship was recorded. Retrieval metrics test the latter, so we examine the embeddings and cell images to ask whether the numerical results correspond to visible phenotype similarity.

Figure 3 compares the input profiles with the learned representations. RxRx3-Core and Angio-Embed show gene–drug separation before alignment, while CPJUMP1 has some overlap. Learning changes these layouts, but overlap alone does not establish a shared phenotype. CPJUMP1's MLP and Sup-GINE projections mix many gene and drug points, yet Sup-GINE improves when its cellular features are removed. Conversely, the Angio MLP and Sup-GINE projections remain separated by modality. The RxRx3-Core and MOTIVE-CRISPR MLP projections also show separation despite positive excess AUROC in selected directions. The visual comparison thus sharpens the quantitative result: a model can rank recorded partners better than the null without bringing gene and drug responses into common neighborhoods. For Angio, this separation coexists with high early retrieval and little excess beyond randomized relationships.

Figure 4 compares each selected drug with a recorded gene partner within the same assay. High AUROC and early partner rank do not consistently accompany similar-looking cells. In CPJUMP1, the pair with AUROC 1.000 and rank 1 differs in cell density and arrangement. In RxRx3-Core, the low-AUROC pair (0.006, rank 349) has more similar overall cell coverage than the high-AUROC pair (1.000, rank 1), whose drug field is visibly sparser than its gene field. The highest-AUROC MOTIVE-CRISPR pair (1.000, rank 2) also differs in coverage: its drug field has larger cell-free regions than the gene field. For both Angio representations, the highest-AUROC drug fields appear denser and more continuous than their gene partners; vessel-like structures themselves are expected in this assay. Rank and AUROC can also disagree: the highest-AUROC Angio-Features query (0.904) has its first recorded partner at rank 44, while its median-AUROC query (0.579) has one at rank 2. Across these within-assay examples, visible phenotype differences do not shrink steadily as AUROC rises or rank improves. The fields are illustrative rather than a profile-level similarity test, but they show that neither ranking metric measures phenotype resemblance.

6 Limitations

Our study provides controlled evidence about gene–drug retrieval, but its conclusions remain limited by the available relationship annotations. Current biomedical knowledge contains only a subset of the possible gene–drug relationships. Identifying missing relationships is itself a goal of drug discovery and repurposing, so an unrecorded pair cannot be treated as biologically false. This is a limitation of the available evidence rather than a failure of our protocol. We address it by treating unrecorded pairs

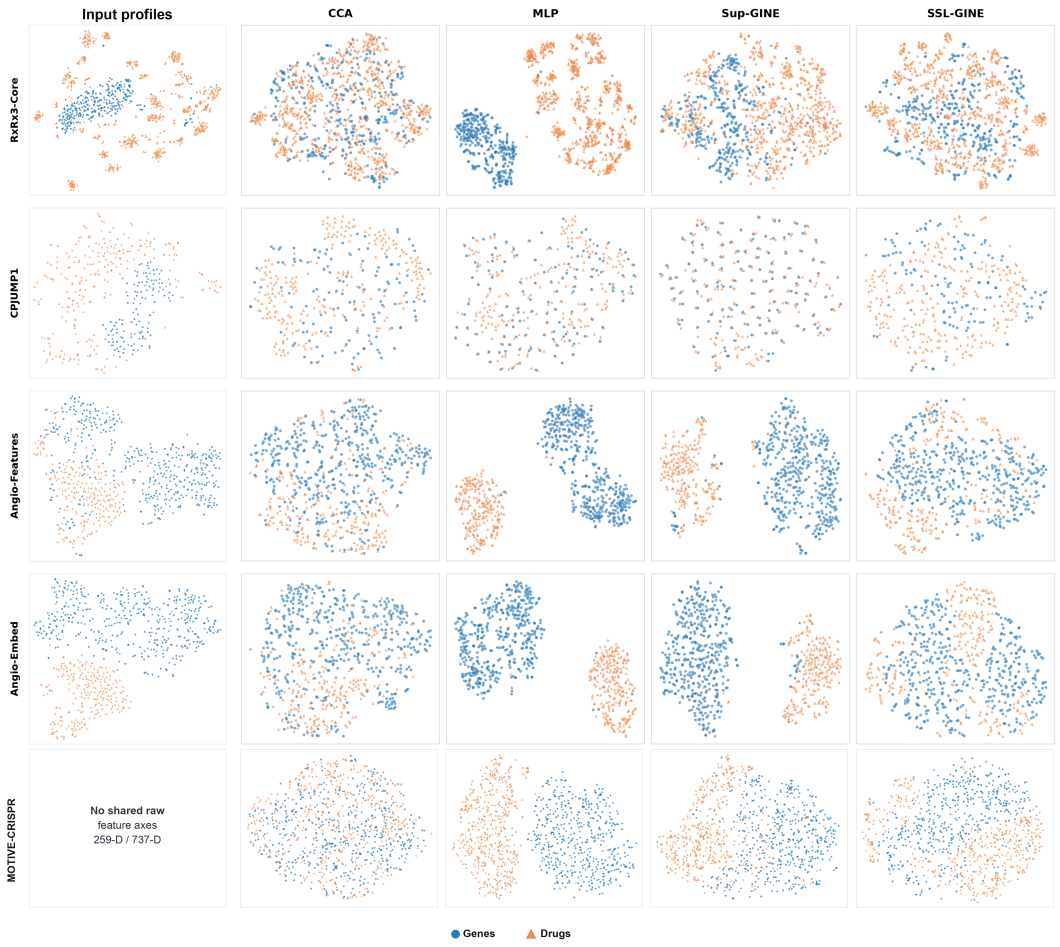

Figure 3: Input profiles and learned gene–drug embeddings across five benchmark inputs. Rows show inputs; columns show the profiles before alignment followed by CCA, MLP, Sup-GINE, and SSL-GINE. Blue circles are genes and orange triangles are drugs. Each panel has its own t-SNE projection. Input profiles are projected jointly only when gene and drug features share axes.

as unknown candidates and by evaluating every model and randomized reference against the same documented relationships.

7 Conclusion

In summary, we introduce a controlled framework for evaluating gene–drug phenotype alignment. It asks whether cellular profiles recover known relationships better than randomized relationships and whether cellular features or graph connections support that recovery. We address the first question through query-held-out retrieval and degree-preserving nulls, and the second through matched changes to features and message-passing edges. CPJUMP1 shows clear bidirectional recovery, RxRx3-Core is stronger from drugs to genes, and Angio shows no consistently dominant direction across metrics. On MOTIVE-CRISPR, MLP retrieves recorded partners better than randomized relationships in both directions. Yet Sup-GINE scores higher when its cellular features are removed. The graph experiments further show that features and connections can help, hurt, or leave performance unchanged depending on the dataset and training objective. Our contributions are a controlled criterion for phenotype alignment, a matched analysis of feature and topology effects, and a dataset-specific comparison supported by embedding and image examples.

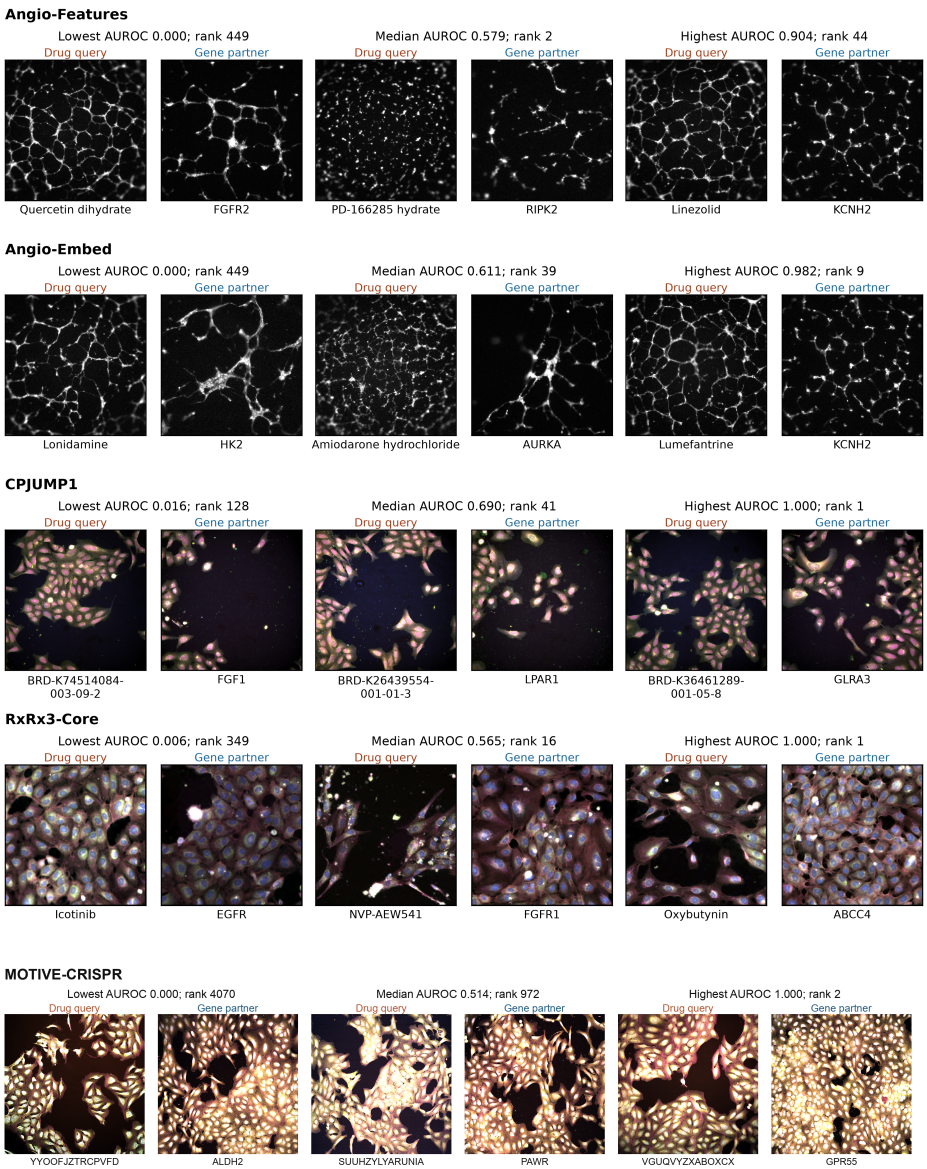

Figure 4: Cell images for selected held-out CCA drug-to-gene queries. Within each input row, the three pairs are the lowest-, median-, and highest-AUROC queries. Each pair shows the drug query on the left and its highest-ranked recorded gene partner on the right. The heading gives the query's AUROC and that gene's rank among candidate genes. AUROC considers all recorded partners; the rank refers only to the displayed partner. Each image is one field selected independently of appearance, not the aggregated profile.

Future work will incorporate this framework into an evidence-based candidate-retrieval pipeline for drug discovery and repurposing. The pipeline would prioritize high-ranked, undocumented gene–drug pairs when their retrieval is supported by cellular features and exceeds the randomized reference, then test these candidates experimentally. Together, these results show why gene–drug retrieval should be interpreted through both its performance and its source, rather than through a raw score alone.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, retrieve and discover related work, and support research ideation and execution. The authors checked cited papers against their original sources, reviewed and tested AI-assisted code, verified reported results against saved experiment outputs, and revised the final text. The authors take responsibility for all claims, analyses, code, and other content in this paper.

ETHICS STATEMENT

This study uses previously collected images and derived cellular representations and does not involve human participants or personal data. Public datasets remain subject to their original licenses, and the private angiogenesis data will not be redistributed without permission. Retrieved gene–drug relationships are research hypotheses, not evidence of direct targeting, causality, safety, or clinical benefit. Experimental and biological validation is required before using any candidate in drug-development or clinical decisions.

REPRODUCIBILITY STATEMENT

The supplementary material documents the data sources, preprocessing, relationship construction, evaluation splits, models, losses, hyperparameters, seeds, and statistical tests. Public inputs can be obtained from their cited sources; the private angiogenesis data are described in detail but remain subject to their access conditions. The complete code and processed embeddings will be made publicly available.

REFERENCES

Rachith Aiyappa, Xin Wang, Munjung Kim, Ozgur Can Seckin, Yong-Yeol Ahn, and Sadamori Kojaku. Implicit degree bias in the link prediction task. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

John Arevalo, Ellen Su, Anne E. Carpenter, and Shantanu Singh. MOTIVE: A drug-target interaction graph for inductive link prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fdb3fa770c2e0ecbb4b7dc7083ef5be9-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chad T. Davis, Bo 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. doi: 10.1038/nprot.2016.105.

Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature Methods*, 14(9):849–863, 2017. doi: 10.1038/nmeth.4397.

Corrie Jacobien Carstens, Annabell Berger, and Giovanni Strona. A unifying framework for fast randomization of ecological networks with fixed (node) degrees. *MethodsX*, 5:773–780, 2018. doi: 10.1016/j.mex.2018.06.018.

Safiye Celik, Jan-Christian Hütter, Sandra Melo Carlos, Nathan H. Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta M. Fay, Berton A. Earnshaw, and Imran S. Haque. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *PLOS Computational Biology*, 20(10):e1012463, 2024. doi: 10.1371/journal.pcbi.1012463.

Srinivas Niranj Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G. Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, Alexandr A. Kalinin, John Arevalo, Marzieh Haghighi, Juan C. Caicedo, Daniel Kuhn, Desiree Hernandez, James Berstler, Hamdah Shafqat-Abbasi, David E. Root, Susanne E. Swalley, Sakshi Garg, Shantanu Singh, and Anne E. Carpenter. Three million images and morphological profiles of cells

treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114–1121, 2024. doi: 10.1038/s41592-024-02241-6.

Claudio Durán, Simone Daminelli, Josephine M. Thomas, V. Joachim Haupt, Michael Schroeder, and Carlo Vittorio Cannistraci. Pioneering topological methods for network-based drug–target prediction by exploiting a brain-network self-organization theory. *Briefings in Bioinformatics*, 19 (6):1183–1202, 2018. doi: 10.1093/bib/bbx041.

Haitao Fu, Feng Huang, Xuan Liu, Yang Qiu, and Wen Zhang. MVGCN: data integration through multi-view graph convolutional network for predicting links in biomedical bipartite networks. *Bioinformatics*, 38(2):426–434, 2022. doi: 10.1093/bioinformatics/btab651.

Diana Gomes, Ann Nowe, and Peter Vrancx. Understanding feature/structure interplay in graph neural networks. In *Proceedings of the Third Learning on Graphs Conference*, volume 269 of *Proceedings of Machine Learning Research*, pp. 17:1–17:15, 2025. URL https://proceedings.mlr.press/v269/gomes25a.html.

Emre Hayir, Lorin Crawford, and Alex X. Lu. MorphoHELM: A comprehensive benchmark for evaluating representations for microscopy-based morphology assays. *arXiv preprint arXiv:2605.15383*, 2026.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

Sukhrobbek Ilyosbekov, Shubham Gajjar, and Rongfei Jin. MorphoCLIP: Text-supervised contrastive learning for perturbation matching in cell painting images. *arXiv preprint arXiv:2608.22690*, 2026. URL https://arxiv.org/abs/2608.22690.

Niveditha S. Iyer, Daniel J. Michael, S-Y Gordon Chi, John Arevalo, Srinivas Niranj Chandrasekaran, Anne E. Carpenter, Pranav Rajpurkar, and Shantanu Singh. Cell morphological representations of genes enhance prediction of drug targets. *bioRxiv*, 2025. doi: 10.1101/2024.06.08.598076. URL https://www.biorxiv.org/content/10.1101/2024.06.08.598076v2. Preprint, version 2, revised March 11, 2025; first posted in 2024.

Alexandr A. Kalinin, John Arevalo, Erik Serrano, et al. A versatile information retrieval framework for evaluating profile strength and similarity. *Nature Communications*, 16:5181, 2025. doi: 10.1038/s41467-025-60306-2.

Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M. Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A. Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *Scientific Reports*, 15:4876, 2025. doi: 10.1038/s41598-025-88825-4.

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.

Oren Kraus 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*, 2024.

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. CellCLIP: Learning perturbation effects in cell painting via text-guided contrastive learning. *Advances in Neural Information Processing Systems*, 2025.

Nikita Moshkov, Michael Bornholdt, Santiago Benoit, et al. Learning representations for image-based profiling of perturbations. *Nature Communications*, 15:1594, 2024. doi: 10.1038/s41467-024-45999-1.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, 16(2):325–337, 2015. doi: 10.1093/bib/bbu010.

Mohammad H. Rohban, Alice M. Fuller, Chui-Kian Tan, et al. Virtual screening for small-molecule pathway regulators by image-profile matching. *Cell Systems*, 13(9):724–736.e9, 2022. doi: 10.1016/j.cels.2022.08.003.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14:7339, 2023. doi: 10.1038/s41467-023-42328-w.

Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5:4114, 2014. doi: 10.1038/ncomms5114.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Kainan Yao, Xiaowen Wang, Wannian Li, Hongming Zhu, Yizhi Jiang, Yulong Li, Tongxuan Tian, Zhaoyi Yang, Qi Liu, and Qin Liu. Semi-supervised heterogeneous graph contrastive learning for drug–target interaction prediction. *Computers in Biology and Medicine*, 163:107199, 2023. doi: 10.1016/j.compbiomed.2023.107199.

Yemin Yu, Emre Hayir, Neil Tenenholtz, Lester Mackey, Ying Wei, David Alvarez-Melis, Ava P. Amini, and Alex X. Lu. Integrating chemical structures as treatments improves representations of microscopy images for morphological profiling. *PLOS Computational Biology*, 22(7):e1014483, 2026. doi: 10.1371/journal.pcbi.1014483.

Shuangjia Zheng, Jiahua Rao, Jixian Zhang, et al. Cross-modal graph contrastive learning with cellular images. *Advanced Science*, 11(32):2404845, 2024. doi: 10.1002/advs.202404845.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020. URL https://arxiv.org/abs/2006.04131.