EFFICIENT CELL PAINTING IMAGE REPRESENTATION LEARNING VIA CROSS-WELL ALIGNED MASKED SIAMESE NETWORK

Pin-Jui Huang*
Smart Group Solution Corp.
jefferyhuang@sgsc.ai

Yu-Hsuan Liao*
Smart Group Solution Corp.
samliao@sgsc.ai

SooHeon Kim
OmixAI Co. Ltd.
shk@omixai.com

NoSeong Park
KAIST
noseong@kaist.ac.kr

JongBae Park
Kyunghee Univ.
OmixAI Co. Ltd.
jbp@khu.ac.kr

DongMyung Shin
OmixAI Co. Ltd.
Oncocross Co. Ltd.
shinsae11@gmail.com

ABSTRACT

Computational models that predict cellular phenotypic responses to chemical and genetic perturbations can accelerate drug discovery by prioritizing therapeutic hypotheses and reducing costly wet-lab iteration. However, extracting biologically meaningful and batch-robust cell painting representations remains challenging. Conventional self-supervised and contrastive learning approaches often require a large-scale model and/or a huge amount of carefully curated data, still struggling with batch effects. We present Cross-Well Aligned Masked Siamese Network (CWA-MSN), a novel representation learning framework that aligns embeddings of cells subjected to the same perturbation across different wells, enforcing semantic consistency despite batch effects. Integrated into a masked siamese architecture, this alignment yields features that capture fine-grained morphology while remaining data- and parameter-efficient. For instance, in a gene-gene relationship retrieval benchmark, CWA-MSN outperforms the state-of-the-art publicly available self-supervised (OpenPhenom) and contrastive learning (CellCLIP) methods, improving the benchmark scores by +29% and +9%, respectively, while training on substantially fewer data (e.g., 0.2M images for CWA-MSN vs. 2.2M images for OpenPhenom) or smaller model size (e.g., 22M parameters for CWA-MSN vs. 1.48B parameters for CellCLIP). Extensive experiments demonstrate that CWA-MSN is a simple and effective way to learn cell image representation, enabling efficient phenotype modeling even under limited data and parameter budgets. The source code for CWA-MSN is available at code link.

1 INTRODUCTION

Computational modeling of cellular responses to perturbations is a promising strategy for drug discovery (Noutahi et al., 2025; Liu et al., 2025; Navidi et al., 2025), predicting therapeutic effects and mechanisms of action (Tanaka et al., 2025), reducing costly wet lab experiments (Bunne et al., 2024; Adduri et al., 2025), and accelerating screening to validation (Stokes et al., 2020). High-content screening (HCS) (Bickle, 2010) enables automated acquisition of cell painting images (Starkuviene & Pepperkok, 2007), generating rich datasets for phenotype-driven modeling (Nierode et al., 2016).

Extracting meaningful representations from these images is challenging. CellProfiler (Stirling et al., 2021) uses handcrafted features to enable discoveries (Boutros et al., 2015; Ariffin, 2023) but is sensitive to batch effects (Arevalo et al., 2024b) and cannot capture complex phenotypic variations (Kim et al., 2025). Self-supervised learning (SSL) (He et al., 2020; Chen et al., 2020; Chen & He, 2021; Caron et al., 2021; He et al., 2022) can learn morphological features (Kraus et al., 2024;

*Equal contribution

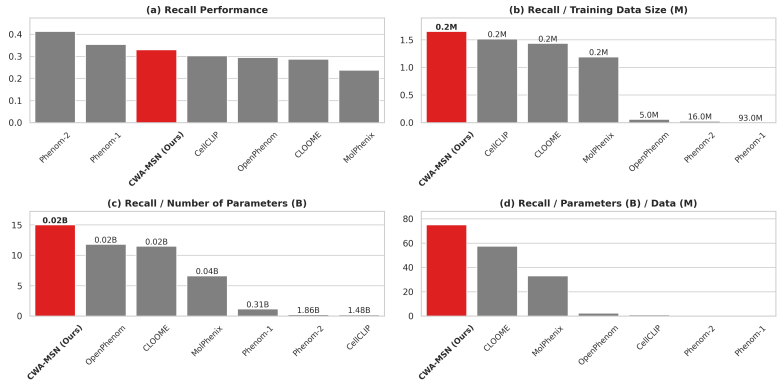

Figure 1: **Comparison of methods based on gene-gene interaction benchmark over multiple efficiency metrics:** (a) benchmark results measured as recall, (b) recall normalized by training data size (per million images), (c) recall normalized by number of parameters (per billion), and (d) recall normalized by the product of both training data size and number of parameters. Our method for each metric plot is highlighted in red. Annotated values indicate either training dataset size (M) or number of parameters (B). Except (a), CWA-MSN is top-performing, showcasing its data- and parameter-efficient learning for cell representation.

Kenyon-Dean et al., 2024) but requires large models (Dosovitskiy et al., 2020) and curated datasets. Weakly supervised or contrastive methods use proxy labels for data-efficient training (Moshkov et al., 2024; Caicedo et al., 2018; Lu et al., 2025; Sanchez-Fernandez et al., 2023; Fradkin et al., 2024; Bushiri Pwesombo et al., 2025), yet remain sensitive to batch effects.

We introduce the Cross-Well Aligned Masked Siamese Network (CWA-MSN), a novel representation learning framework for cell painting images. Unlike conventional self-supervised methods, CWA-MSN leverages weak perturbation labels to align representations across wells of the same perturbation. This cross-well alignment enforces robust semantic consistency, preserving biologically meaningful relationships instead of confounding batch effects introduced by experiment technical factors. By integrating this alignment into a masked siamese network (Assran et al., 2022), CWA-MSN substantially improves the capture of phenotypic relationships while maintaining high data and parameter efficiency.

Extensive experiments demonstrate that CWA-MSN consistently outperforms existing approaches in biological relationship retrieval tasks, particularly gene–gene and compound–gene associations (Kraus et al., 2025). On gene–gene interaction benchmarks, CWA-MSN surpasses the state-of-the-art self-supervised method OpenPhenom (Kraus et al., 2024) and the weakly supervised CellCLIP (Lu et al., 2025) by 29% and 9%, respectively. These gains are achieved with significantly reduced resources—0.2M versus 2.2M training images for OpenPhenom, or 22M versus 1.48B model parameters for CellCLIP. Fig. 1 illustrates CWA-MSN's advantages in terms of model size, training data, and performance.

2 Related Work

2.1 Self-Supervised Learning for Cell Painting Images

Self-supervised learning (He et al., 2020; 2022; Chen et al., 2020; Chen & He, 2021; Caron et al., 2021) has shown promise for microscopy images, but transferring methods from natural images to HCS data can be challenging. For example, DINO (Caron et al., 2021) relies on augmentations designed for natural images, limiting effectiveness on HCS (Doron et al., 2023; Kim et al., 2025; Kraus et al., 2024). Masked image modeling like MAE (He et al., 2022) reduces augmentation dependency and has successfully retrieved biological relationships (Kraus et al., 2024; Kenyon-Dean et al., 2024). However, these methods require massive compute (256 H100 GPUs (Kenyon-Dean et al., 2024)) and large datasets (93M images (Kenyon-Dean et al., 2024)). We propose a more data- and parameter-efficient approach achieving competitive performance.

2.2 Weakly Supervised and Contrastive Learning for Cell Painting Images

Weakly supervised and contrastive learning (Yu et al., 2025; Bao et al., 2023) leverage proxy labels to train image encoders (Moshkov et al., 2024; Caicedo et al., 2018; Sanchez-Fernandez et al., 2023; Lu et al., 2025; Fradkin et al., 2024; Bushiri Pwesombo et al., 2025). SemiSupCon (Bushiri Pwesombo et al., 2025) aligns replicative features via contrastive learning but ignores cross-well, plate, and batch effects, while CellCLIP (Lu et al., 2025) uses text-based signals. SSLProfiler (Dai et al., 2025) aligns site-level images with DINOv2 but assumes identical perturbations per well, missing cross-plate variation (Moshkov et al., 2024). Our method explicitly aligns wells with the same perturbation across plates with prototype-based objective, reducing confounding from batch effects without proxy labels.

3 Cross-Well Aligned Masked Siamese Network

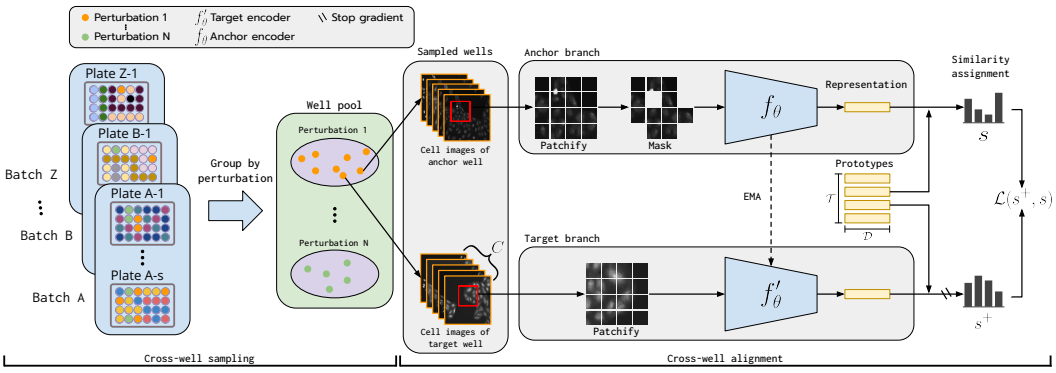

Figure 2: **Overview of CWA-MSN framework:** The framework is composed of two parts, cross-well sampling and cross-well alignment. Cross-well sampling selects cell images under the same perturbation from different wells across batches and plates to serve as an implicit data augmentation strategy. Cross-well alignment utilizes a masked siamese network to align anchor and target well representations by matching their prototype-based similarity distributions.

3.1 Problem Statement

HCS experiments produce hierarchically structured cellular imaging data. A batch corresponds to a set of plates (e.g., Batch A in Fig. 2) processed under uniform experimental conditions, and each plate contains multiple wells (e.g., 96 or 384) with replicates of cells under a specific perturbation (e.g., six wells per perturbation).

Batch effects from instrumentation, imaging, sample preparation, and technical noise can obscure true perturbation signals. We address this with a data-efficient approach using cross-well alignment and masked Siamese network (MSN) learning, which we select over alternatives such as DINOv2 because its prototype-aligned objective provides a stable signal with minimal reliance, avoiding distortion of subtle perturbation-dependent features in microscopy.

3.2 Cross-Well Sampling

In CWA-MSN, cross-well images of the same perturbation across plates and batches are used as implicit data augmentation. Let $P = \{p_1, \ldots, p_N\}$ denote N perturbations, each associated with a set of wells:

$$W_i = \{w_1^{(i)}, \ldots, w_{M_i}^{(i)}\}, \quad w \in \mathbb{R}^{C \times H \times W},$$

where M_i is the number of wells for perturbation p_i, C is the number of channels, and $H \times W$ is the spatial dimension.

For sampling, a perturbation $p \in P$ is randomly selected, and two distinct wells under p are chosen:

$$w_a^p, w_t^p \in W_p, \quad w_a^p \neq w_t^p,$$

where w_a^p and w_t^p are the anchor and target wells, potentially from the same or different plates/batches (Fig. 2).

3.3 Cross-Well Alignment via Masked Siamese Network

CWA-MSN combines cross-well sampling with a masked siamese network (Assran et al., 2022). Unlike MAE, which reconstructs masked regions of a single image, MSN aligns representations of two images (here, cross-well pairs) using masked and unmasked views via prototype-based learning.

For the anchor well w_a^p, V_a augmented views are generated using random crops and flips:

$$\mathbf{X}_a^{(p)} \in \mathbb{R}^{V_a \times C \times H \times W}.$$

The target well w_t^p is represented by a single augmented view:

$$\mathbf{X}_t^{(p)} \in \mathbb{R}^{1 \times C \times H \times W}.$$

A mini-batch is formed by stacking anchor and target views for perturbations $P_B \subset P$ with $|P_B| = B$:

$$\mathbf{X}_a = \{\mathbf{X}_a^{(p)}\}_{p \in P_B} \in \mathbb{R}^{B \times V_a \times C \times H \times W}, \quad \mathbf{X}_t = \{\mathbf{X}_t^{(p)}\}_{p \in P_B} \in \mathbb{R}^{B \times 1 \times C \times H \times W}.$$

The anchor view $\mathbf{X}_a$ is patchified and masked with ratio α, while $\mathbf{X}_t$ is only patchified. Embeddings are computed via anchor and target encoders:

$$z = f_\theta(\mathbf{X}_a) \in \mathbb{R}^{B \times V_a \times \mathcal{D}}, \quad z^+ = f'_\theta(\mathbf{X}_t) \in \mathbb{R}^{B \times 1 \times \mathcal{D}}, \tag{1}$$

with representation dimension $\mathcal{D}$.

With a set of prototype embeddings $\mathrm{O} \in \mathbb{R}^{\mathcal{T} \times \mathcal{D}}$, where $\mathcal{T}$ is the number of prototypes, we compute the similarity assignment scores as:

$$s = \mathrm{sim}(\mathrm{O}, z) \in \mathbb{R}^{B \times V_a \times \mathcal{T}}, \quad s^+ = \mathrm{sim}(\mathrm{O}, z^+) \in \mathbb{R}^{B \times 1 \times \mathcal{T}}. \tag{2}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes the normalized cosine similarity.

The model is trained to align anchor and target similarity distributions with an auxiliary mean-entropy term to prevent collapse:

$$\mathcal{L}(s^+, s) = \lambda_1 \, CE(s^+, s) + \lambda_2 \frac{1}{B\mathcal{T}} \sum_{j=1}^{B} \sum_{m=1}^{\mathcal{T}} s_{j,m}, \tag{3}$$

where $CE(\cdot, \cdot)$ is cross-entropy and λ_1, λ_2 balance the terms. The target encoder f'_θ is updated via exponential moving average (EMA) of f_θ (Fig. 2).

3.4 Implementation Details

We use ViT-S/16 for both anchor and target encoders and train for 100 epochs with batch size 64 using AdamW. The initial learning rate is 0.0002 with a 15-epoch warm-up and cosine decay; weight decay follows a cosine schedule from 0.04 to 0.4. We set $\mathcal{T} = 1024$ prototypes (see Appendix A.1 for ablation), representation dimension $\mathcal{D} = 256$, loss weights $\lambda_1 = \lambda_2 = 1$, and anchor masking ratio $\alpha = 0.15$. Anchor views follow Assran et al. (2022) with one random and ten focal crops ($V_a = 11$). The target encoder is updated via EMA with momentum from 0.996 to 1.0.

4 Experiments

4.1 Training Data of CWA-MSN

For the development of CWA-MSN, we utilized Bray dataset (Bray et al., 2016) which encompasses five-channel cell painting images perturbed by diverse small-molecules. First, we applied the pre-processing pipeline described in Sanchez-Fernandez et al. (2023). Then, following CellCLIP (Lu et al., 2025), we selected 70% of the total data for training, including 198,609 cell images with 7,401 distinct perturbations. Note that the size of the training data (that is, 0.2 M) is much smaller than that of recent self-supervised methods (from 5M to 93M; see Fig. 1).

4.2 BENCHMARKS

Gene-Gene Interaction Benchmark RxRx3-core (Kraus et al., 2025) is a curated benchmark dataset to evaluate zero-shot performance of a cell painting image encoder, circumventing the limitations of existing benchmarks (Chandrasekaran et al., 2024; Arevalo et al., 2024a) such as small perturbation coverage and biased well positions. The dataset consists of 1,335,606 images perturbed by 736 gene knockouts and 1,674 small-molecules.

In the RxRx3-core gene–gene interaction benchmark (Celik et al., 2024), models are evaluated by computing pairwise cosine similarities between features of all gene–gene pairs (e.g., MTOR–TSC2), selecting the top and bottom 5% similarities, and comparing these predicted positive and negative interactions against curated databases, including Reactome, HuMAP, SIGNOR, StringDB, and CORUM (Giurgiu et al., 2019; Drew et al., 2017; Gillespie et al., 2022; Szklarczyk et al., 2021). Performance is measured using recall (discovered / known interactions) for each database.

In Section 5.1, we compare the proposed CWA-MSN with handcrafted features (CellProfiler (Stirling et al., 2021)), weakly supervised methods (SupCon (Khosla et al., 2020), MolPhenix (Fradkin et al., 2024), CLOOME (Sanchez-Fernandez et al., 2023), CellCLIP (Lu et al., 2025)), contrastive learning (SimCLR (Chen et al., 2020)), and self-supervised approaches (OpenPhenom, Phenom-1 (Kraus et al., 2024), and Phenom-2 (Kenyon-Dean et al., 2024)). We additionally report ViT/S-16 baselines trained without HCS data (ViT-ImageNet) and with perturbation-label supervision (ViT-WSL). Results are summarized in Table 1, along with training data size and parameter counts to assess data and parameter efficiency; FLOPs analysis is provided in Appendix A.2.

Compound-Gene Interaction Benchmark The RxRx3-core compound–gene interaction benchmark evaluates a model's ability to associate gene knockouts with small-molecule perturbations by computing cosine similarity between their embeddings (Celik et al., 2024). For each compound, known target genes are ranked against random genes, and performance is measured using AUC and average precision (AP). Results are reported as the mean and standard deviation of AUC and AP across compounds, together with z-scores relative to a random baseline. Ground-truth compound–gene associations are curated from PubChem, Guide to Pharmacology, WIPO, D3R, BindingDB, US Patents, and ChEMBL (Liu et al., 2007; Zdrazil et al., 2024; Harding et al., 2024).

In Section 5.2, we compare CWA-MSN against CellProfiler, CellCLIP, OpenPhenom, Phenom-1, and Phenom-2, as well as ViT/S-16 baselines trained on ImageNet-1K and the Bray dataset. We exclude CLOOME and MolPhenix due to unavailable source code and reported metrics.

4.3 VALIDATION OF CWA-MSN

Batch-Effects Probing Batch effects in image-based profiling are commonly quantified by measuring how well technical metadata (plate, batch, acquisition day) can be recovered from embeddings. In Section 5.3, we followed this standard practice by probing the recoverability of plate identity via linear classifiers and kNN (k=5) with 5-fold cross-validation. We evaluated on the full RxRx3-core dataset as well as a variant with all negative controls removed.

Single-Well vs. Cross-Well Alignment One of the key innovations in CWA-MSN is to utilize cross-well images as implicit data augmentation for training. In Section 5.4, we validated the effect of this cross-well sampling strategy, by changing it to a conventional single-well sampling method. We performed the comparison between single-well and cross-well based on the gene-gene interaction benchmark, using the Bray dataset for training.

Masked Siamese Network vs. Masked Autoencoder We checked whether a masked siamese network has indeed benefits over the popular alternative, masked autoencoder. Specifically, we adopted CropMAE (Eymaël et al., 2024), which uses pairs of cropped images as anchor and target, but optimizes masked reconstruction instead of prototype alignment.

In Section 5.5, we tested CropMAE with single-well and cross-well settings, comparing their performance with that of CWA-MSN. To examine the performance and training efficiency together, we reported not only the gene-gene interaction benchmarks but also the training time on Bray dataset.

Masking vs. No Masking Given that CWA-MSN is trained with weak supervision from perturbation labels, we performed an ablation study to evaluate the contribution of the asymmetric masking mechanism. In Section 5.6, we compared our masked-reconstruction scheme against a no-masking variant. The two training settings are identical except for the masking rate, which is set to zero for the no-masking model.

5 RESULTS

5.1 GENE-GENE INTERACTION BENCHMARK

Table 1: Gene–gene interaction benchmark results of different methods. *: Values from Lu et al. (2025). **: Not publicly available. N.A.: Not available.

Training Dataset	# Images	# Perturb.	Parameters	Method	CORUM ↑	hu.MAP ↑	Reactome ↑	StringDB ↑
-	-	-	-	Random	.107	.111	.107	.115
ImageNet-1K	1M	-	22M	ViT-ImageNet	.342	.420	.144	.305
-	-	-	-	CellProfiler	.361	.444	.160	.330
Bray *et al.*	0.2M	>7K	22M	SupCon	.242	.271	.123	.224
Bray *et al.*	0.2M	>7K	22M	ViT-WSL	.249	.290	.148	.242
Bray *et al.*	0.2M	>7K	36M	MolPhenix*	.262	.306	.142	.241
Bray *et al.*	0.2M	>7K	25M	CLOOME*	.328	.406	.135	.278
Bray *et al.*	0.2M	>7K	1,477M	CellCLIP	.354	.416	.145	.307
Bray *et al.*	0.2M	>7K	22M	SimCLR	.256	.290	.137	.239
RxRx3+cpg0016	>10M	>116K	25M	OpenPhenom	.300	.352	**.158**	.281
RPI-93M	93M	∼4M	307M	Phenom-1**	.395	.482	.188	.349
PP-16M	16M	N.A.	1,860M	Phenom-2**	.486	.553	.197	.415
Bray *et al.*	0.2M	>7K	22M	**CWA-MSN (Ours)**	**.386**	**.447**	**.158**	**.327**

As shown in Table 1, CWA-MSN outperformed all handcrafted, weakly supervised, contrastive learning methods on the benchmark gene-gene interaction, except a few large-scale private models (i.e., Phenom-1 and Phenom-2). In particular, it surpassed the SOTA weakly supervised contrastive learning method, CellCLIP, with significant performance gaps (e.g., CORUM: .354 for CellCLIP vs. .386 for CWA-MSN). Additional analysis in Section 5.3 further verifies that our performance gains indeed stem from batch-effect mitigation. Considering that the same Bray dataset was used for CellCLIP and CWA-MSN training, these results demonstrate the superior parameter efficiency of CWA-MSN with a much smaller model size (1,477M for CellCLIP vs. 22M for CWA-MSN).

Furthermore, CWA-MSN outperformed OpenPhenom, which is publicly available SOTA self-supervised method, in most of the retrieval tasks (CORUM: .300 vs. .386, hu.MAP: .352 vs. .447, and StringDB: .281 vs. .327)). The results indicate better data efficiency for CWA-MSN compared to OpenPhenom, even with the large gap between the number of training images (>10M for OpenPhenom vs. 0.2M for CWA-MSN). Additionally, its superior computational efficiency in terms of FLOPs is reported in Appendix A.2.

The benchmark results for Phenom-1 and Phenom-2 are in fact better than those for CWA-MSN. However, CWA-MSN has significantly superior data and parameter efficiency over Phenom-1 and Phenom-2 (e.g., see (b), (c) and (d) in Fig. 1). Also, it should be noted that the training data (RPI-93M and PP-16M) and source codes of Phenom-1 and Phenom-2 are not publicly available.

5.2 COMPOUND-GENE INTERACTION BENCHMARK

Fig. 3 shows the graphs of the compound-gene interaction benchmark for each method, reporting the AUC-ROC and AP values over the concentration. In general, the graphs of CWA-MSN and OpenPhenom are competing with each other as the top performing method. For example, CWA-MSN consistently outperforms all other methods in the AUC-ROC graph within a range of 0.25 μ Mol up to the maximum concentration, whereas OpenPhenom dominates in the other range of concentrations (see Fig. 3).

As shown in Table 2, if we closely investigate the z-scores of each method at the maximum concentration, OpenPhenom achieved the highest z-scores in both the AP and AUC-ROC metrics (3.89 and 3.16) compared to the second-best z-scores of CWA-MSN (3.55 and 2.88). Although the z-scores of CWA-MSN are slightly lower than those of OpenPhenom, these two methods possibly

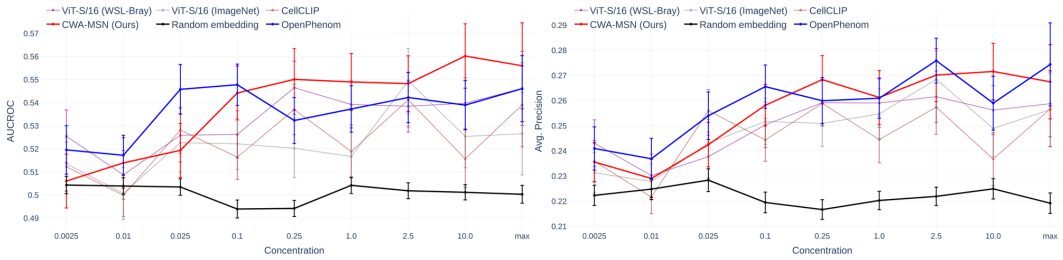

Figure 3: Compound-gene interaction benchmark graphs. AUC-ROC and AP values are reported over concentration.

Table 2: Compound-gene interaction benchmark results at the maximum concentration level. The best performance is in **bold**, and second best is in underline. *: Not publicly available. **: Evaluated using open models.

	AP			AUC-ROC		
Method	**Mean**	**Std.**	**Z-score↑**	**Mean**	**Std.**	**Z-score↑**
Phenom-2*	.307	.015	6.04	-	-	-
Phenom-1*	.290	.017	4.35	-	-	-
OpenPhenom**	.274	.017	**3.89**	.546	.014	**3.16**
ViT-WSL	.259	.013	3.37	.546	.016	2.79
CellProfiler	.276	.018	3.34	-	-	-
CellCLIP**	.257	.015	2.81	.539	.018	2.12
ViT-ImageNet	.256	.015	2.75	.527	.018	1.46
Random	.214	.003	0.00	.500	.004	0.00
CWA-MSN (ours)	.267	.015	3.55	.556	.019	2.88

have complementary strengths. For example, the std. of AP is slightly lower in CWA-MSN (i.e., better feature consistency among known relationships), whereas the mean AP is marginally higher in OpenPhenom (i.e., better capturing known relationships on average).

Most importantly, we want to highlight that this competitive performance of CWA-MSN relative to OpenPhenom was achieved despite training in a significantly (over $50\times$) smaller data size.

5.3 Batch-Effects Probing

Table 3: Five-fold cross-validation results for predicting plate identity from learned embeddings (in macro-F1 scores).

Embedding	**Linear**		**KNN**	
	Full↓	**No Ctrl↓**	**Full↓**	**No Ctrl↓**
OpenPhenom	27.07% ± 0.55%	28.32% ± 0.37%	26.83% ± 0.15%	27.23% ± 0.46%
CWA-MSN (Ours)	**13.22% ± 0.64%**	**13.99% ± 0.85%**	**13.32% ± 0.22%**	**13.34% ± 0.33%**

As summarized in Table 3, across all probing strategies, CWA-MSN yields significantly lower plate-predictability (approximately half) than OpenPhenom, indicating substantially weaker entanglement with batch-specific artifacts. These results prove that the cross-well alignment strategy effectively suppresses technical variation while preserving morphological signal.

5.4 Single-Well vs. Cross-Well Alignment

As summarized in Table 4, when we tested the effect of single-well and cross-well sampling strategies combined with a masked simease network, we observed significant performance gaps between the two models. Concretely, compared to the single-well alignment (Single-Well-MSN in Table

Table 4: Gene-gene interaction benchmark results between single-well and cross-well masked simease networks. The best performance is highlighted in **bold**.

Model	**CORUM**	**hu.MAP**	**Reactome**	**StringDB**
# relationships	*1,209*	*958*	*569*	*1,737*
Random	.107	.111	.107	.115
Single-Well-MSN	.281	.330	.130	.261
CWA-MSN (Ours)	**.386**	**.447**	**.158**	**.327**

4), the cross-well alignment (CWA-MSN in Table 4) largely improves recall in all gene-gene association databases. These findings show that cross-well sampling consistently outperforms the single-well counterpart in biological relationship retrieval.

5.5 Masked Siamese Network vs. Masked Autoencoder

Table 5: Gene–gene interaction benchmark comparison of CWA-MSN and CropMAE. The best performance per metric is highlighted in **bold**.

Training Time (GPU hours)	**Model**	**CORUM**	**hu.MAP**	**Reactome**	**StringDB**
	# relationships	*1,209*	*958*	*569*	*1,737*
-	Random	.107	.111	.107	.115
109	CropMAE-Single	.338	.408	.137	.303
14	CropMAE-Cross	.348	.443	.135	.309
<9	CWA-MSN (Ours)	**.386**	**.447**	**.158**	**.327**

Table 5 shows that CWA-MSN consistently surpasses CropMAE (Eymaël et al., 2024) with either single-well or cross-well settings in gene-gene relationship retrieval tasks. In detail, compared to CropMAE with cross-well sampling (i.e., CropMAE-Cross), CWA-MSN achieved higher recall in all gene-gene interaction databases with the minimum training time . The results indicate that applying cross-well alignment strategy to a masked siamese network (prototype-based learning) is a more effective combination than to a masked autoencoder (reconstruction-based learning) in terms of performance and training cost. Interestingly, applying the proposed cross-well sampling strategy to CropMAE alone substantially reduced training cost while also improving benchmark performance (see CropMAE-Single vs. CropMAE-Cross in Table 5).

5.6 Masking vs. No Masking

Table 6: Ablation study results of CWA-MSN with and without asymmetric masking strategy.

Model	**CORUM**	**hu.MAP**	**Reactome**	**StringDB**
No Masking	0.354	0.423	0.147	0.320
CWA-MSN (Ours)	**0.386**	**0.447**	**0.158**	**0.327**

Table 6 reports the performance of CWA-MSN compared to its variant without asymmetric masking strategy. These results indicate that asymmetric masking provides a clear benefit even in the presence of perturbation supervision. By masking only the anchor view, the model is forced to rely on perturbation-relevant morphological cues rather than low-level artifacts, improving robustness under the noisy and batch-variable conditions of cell-painting data.

6 Conclusion

In conclusion, we present CWA-MSN, a simple and effective framework for representation learning of cell painting images, which can extract phenotypic changes according to chemical and genetic

perturbations with high data and parameter efficiency. By aligning embeddings of identically perturbed cells across wells using a masked siamese architecture, CWA-MSN mitigates batch effects while preserving fine-grained morphology. This yields biologically meaningful features that improve relationship retrieval across gene–gene and compound–gene, surpassing state-of-the-art public self-supervised and contrastive baselines, even under limited data and parameter budgets.

7 DISCLOSURE

MEANINGFULNESS STATEMENT

A meaningful representation of life captures the true phenotypic state of cells, reflecting how they respond to genetic or chemical perturbations while disentangling technical noise and batch effects. Our work advances this goal by using cross-well alignment and masked siamese networks to learn robust, data-efficient embeddings from cell painting images. These representations preserve biologically relevant variations, enabling more accurate modeling of cellular behavior and accelerating discovery of mechanisms and therapeutic effects.

LLM USAGE

We used large language models (ChatGPT and Claude) to assist with code design and manuscript editing. All outputs were reviewed and validated by the authors, who take full responsibility for the accuracy and originality of this work.

REFERENCES

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, pp. 2025–06, 2025.

John Arevalo, Ellen Su, Anne E. Carpenter, and Shantanu Singh. Motive: A drug-target interaction graph for inductive link prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 140320–140333. Curran Associates, Inc., 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fdb3fa770c2e0ecbb4b7dc7083ef5be9-Paper-Datasets_and_Benchmarks_Track.pdf.

John Arevalo, Ellen Su, Jessica D Ewald, Robert Van Dijk, Anne E Carpenter, and Shantanu Singh. Evaluating batch correction methods for image-based cell profiling. *Nature Communications*, 15 (1):6516, 2024b.

Nur Syamimi Ariffin. The cellprofiler pipeline analysis of cell migration. *Acta Histochemica*, 125 (7):152074, 2023.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: an image is worth 1 x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023.

Marc Bickle. The beautiful cell: high-content screening in drug discovery. *Analytical and bioanalytical chemistry*, 398(1):219–226, 2010.

Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-based high-content screening. *Cell*, 163(6):1314–1325, 2015.

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.

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.

David Bushiri Pwesombo, Carsten Beese, Christopher Schmied, and Han Sun. Semisupervised contrastive learning for bioactivity prediction using cell painting image data. *Journal of Chemical Information and Modeling*, 65(2):528–543, 2025.

Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of single-cell feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9309–9318, 2018.

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*, pp. 9650–9660, 2021.

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.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Siran Dai, Qianqian Xu, Peisong Wen, Yang Liu, and Qingming Huang. Self-supervised representation learning with local aggregation for image-based profiling, 2025. URL `https://arxiv.org/abs/2506.14265`.

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. *bioRxiv*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kevin Drew, Chanjae Lee, Ryan L Huizar, Fan Tu, Blake Borgeson, Claire D McWhite, Yun Ma, John B Wallingford, and Edward M Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, 13(6): 932, 2017.

Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with siamese cropped masked autoencoders. In *European Conference on Computer Vision*, pp. 348–366. Springer, 2024.

Philip Fradkin, Puria Azadi Moghadam, Karush Suri, Frederik Wenkel, Maciej Sypetkowski, and Dominique Beaini. Molphenix: A multimodal foundation model for phenomolecular retrieval. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*, 2024. URL `https://openreview.net/forum?id=elA8hwvYAm`.

Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47(D1):D559–D563, 2019.

Simon D Harding, Jane F Armstrong, Elena Faccenda, Christopher Southan, Stephen PH Alexander, Anthony P Davenport, Michael Spedding, and Jamie A Davies. The iuphar/bps guide to pharmacology in 2024. *Nucleic acids research*, 52(D1):D1438–D1449, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

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(1):4876, 2025.

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*, pp. 11757–11768, 2024.

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.

Gang Liu, Srijit Seal, John Arevalo, Zhenwen Liang, Anne E Carpenter, Meng Jiang, and Shantanu Singh. Learning molecular representation in a cell. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BbZy8nI1si.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

Mingyu Lu, Ethan Weinberger, Chanwoo Kim, and Su-In Lee. Cellclip–learning perturbation effects in cell painting via text-guided contrastive learning. *arXiv preprint arXiv:2506.06290*, 2025.

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.

Zeinab Navidi, Jun Ma, Esteban Miglietta, Le Liu, Anne E Carpenter, Beth A Cimini, Benjamin Haibe-Kains, and BO WANG. Morphodiff: Cellular morphology painting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PstM8YfhvI.

Gregory Nierode, Paul S Kwon, Jonathan S Dordick, and Seok-Joon Kwon. Cell-based assay design for high-content screening of drug candidates. *Journal of microbiology and biotechnology*, 26(2): 213, 2016.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, et al. Virtual cells: Predict, explain, discover. *arXiv preprint arXiv:2505.14613*, 2025.

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.

V Starkuviene and R Pepperkok. The potential of high-content high-throughput microscopy in drug discovery. *British journal of pharmacology*, 152(1):62–71, 2007.

David R Stirling, Madison J Swain-Bowden, Alice M Lucas, Anne E Carpenter, Beth A Cimini, and Allen Goodman. Cellprofiler 4: improvements in speed, utility and usability. *BMC bioinformatics*, 22(1):433, 2021.

Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

Damian Szklarczyk, Annika L Gable, Katerina C Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.

Tatsuya Tanaka, Toshiaki Katayama, and Takeshi Imai. Predicting the effects of drugs and unveiling their mechanisms of action using an interpretable pharmacodynamic mechanism knowledge graph (ipm-kg). *Computers in Biology and Medicine*, 184:109419, 2025. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2024.109419. URL https://www.sciencedirect.com/science/article/pii/S001048252401504X.

Yemin Yu, Neil Tenenholtz, Lester Mackey, Ying Wei, David Alvarez-Melis, Ava P Amini, and Alex X Lu. Causal integration of chemical structures improves representations of microscopy images for morphological profiling. *arXiv preprint arXiv:2504.09544*, 2025.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

A APPENDIX

This supplement presents extended methodological details, ablations, and quantitative analyzes in support of the main text. These experiments provide further evidence for the optimal number of prototypes(A.1) and computational efficiency(A.2).

A.1 PROTOTYPE NUMBER OPTIMIZATION

As prototype alignment plays a key role in the training of CWA-MSN, it is important to find an optimal number of prototypes that can effectively capture biological relationships between cellular images. Therefore, we optimized the number by changing the number of prototypes (256, 512, 1024, and 2048) and measuring the performance based on the gene-gene interaction benchmark.

Table 7: Optimization results for the number of prototypes in CWA-MSN based on gene-gene interaction prediction. The best performance for each metric is highlighted in **bold**.

Number of Prototypes	**CORUM**	**hu.MAP**	**Reactome**	**StringDB**
# relationships	*1,209*	*958*	*569*	*1,737*
256	.372	.433	.132	.321
512	.344	.401	.151	.311
1,024	**.386**	**.447**	**.158**	**.327**
2,048	.369	.438	.141	.314

The optimization results for the number of prototypes is summarized in Table 7. When we changed the number from 256 to 2,048, the best performance was achieved at the number equal to 1,024. We potentially concluded that this is a point that balances the redundancy and diversity of prototypes.

A.2 COMPUTATIONAL EFFICIENCY (FLOPS ANALYSIS)

In the main text, we show that CWA-MSN matches or surpasses prior methods under constrained data and parameter budgets, indicating better data and parameter efficiency. To further assess computational efficiency, we compare training FLOPs.

Table 8: Computational efficiency (in GFLOPs) and gene–gene interaction retrieval performance on RxRx3-core.

Method	**GFLOPs**	**#Params (M)**	**CORUM**	**hu.MAP**	**Reactome**	**StringDB**
ViT-WSL	8.79	22	0.249	0.290	0.148	0.242
CellCLIP	339.17	1,477	0.354	0.416	0.145	0.307
OpenPhenom	104.68	25	0.300	0.352	0.158	0.281
CWA-MSN (Ours)	**23.66**	**22**	**0.386**	**0.447**	**0.158**	**0.327**

CWA-MSN achieves state-of-the-art biological performance while maintaining a computation footprint far below large-scale alternatives. These results show the superior computational efficiency of CWA-MSN compared to the other methods.