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

Pin-Jui Huang
Smart Group Solution Corp.

Yu-Hsuan Liao
Smart Group Solution Corp.

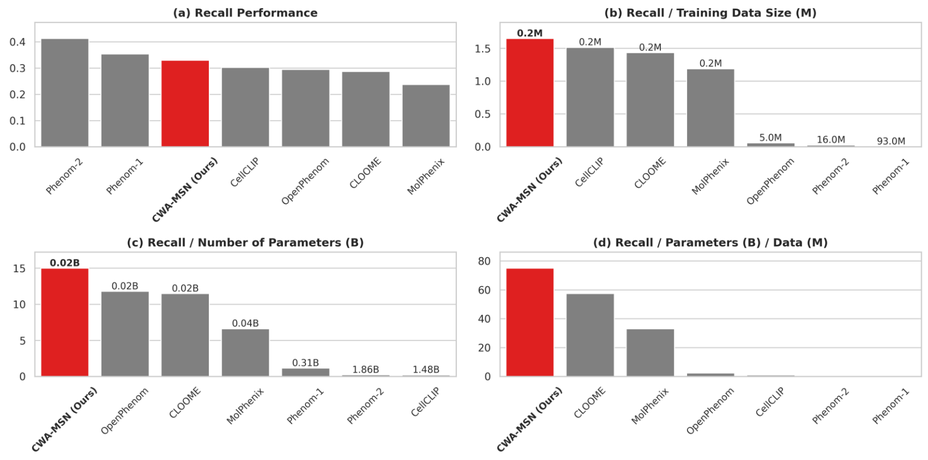
SooHeon Kim
OmixAI Co. Ltd.

NoSeong Park
KAIST

JongBae Park
Kyunghee Univ.
OmixAI Co. Ltd.

DongMyung Shin
OmixAI Co. Ltd.
Oncocross Co. Ltd.

SOTA comparison



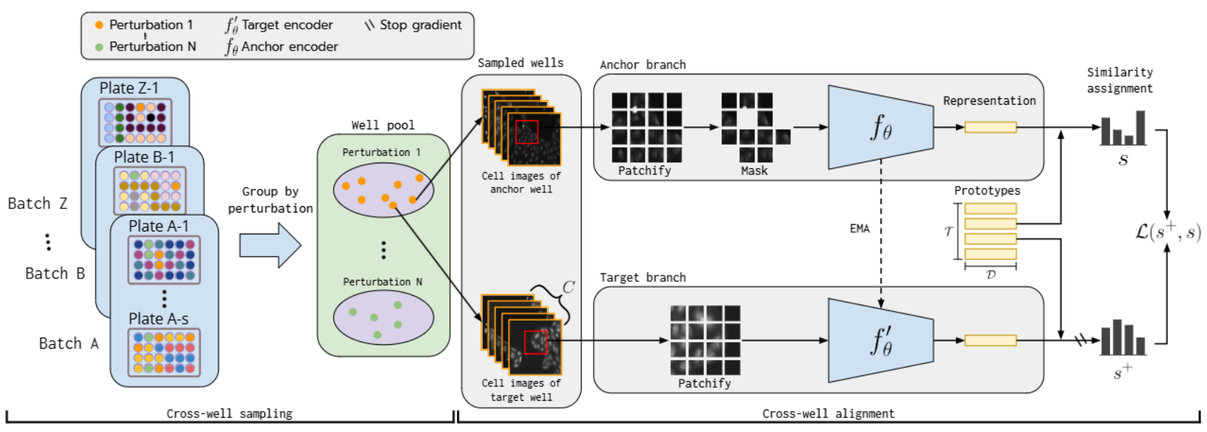
Abstract

Computational models that predict cellular responses to chemical and genetic perturbations can accelerate drug discovery, but learning biologically meaningful and batch-robust representations remains challenging.

We propose Cross-Well Aligned Masked Siamese Network (CWA-MSN), a representation learning framework that aligns cells exposed to the same perturbation across wells, reducing batch effects while preserving morphological detail. Integrated with a masked siamese architecture, CWA-MSN is both data- and parameter-efficient.

It outperforms state-of-the-art methods on gene-gene/compound-gene retrieval benchmarks, achieving higher accuracy with fewer images and smaller models. Our results demonstrate a simple, scalable approach for robust phenotype modeling under limited data conditions.

Overview



Method

To mitigate pervasive technical correlation, namely batch effect in high-content screening, CWA-MSN leverages a novel cross-well sampling strategy as implicit data augmentation to achieve efficient cell representation learning.

For a given biological perturbation p, two distinct wells w_a^p and w_t^p undergoes perturbation p are randomly sampled from different plates or batches. Both well will sample one image that serves as the anchor view and target view for alignment. Both view are then processed by the Masked Siamese Network (MSN) to extract cell representation embeddings. Specifically, the anchor view X_a from w_a^p is heavily masked (e.g., ratio $\alpha = 0.15$) and passed through an online encoder to yield embeddings z, while the target view X_t from w_t^p remains unmasked and is processed by a momentum encoder to yield embedding z^+.

Rather than reconstructing missing pixels, the network performs prototype-based alignment. We compute similarity assignment scores between these embeddings and a set of learnable prototypes and regularize the distribution of both views to achieve consistency regularization of cross-well alignment.

By enforcing semantic consistency between differently masked, identically perturbed cells from entirely different wells, **CWA-MSN ignores low-level technical artifacts and forces the model to capture perturbation-relevant morphology efficiently.**

Datasets

- Training Data (Bray Dataset):
  Consists of 5-channel cell painting images capturing small-molecule perturbations.
  Highly Data-Efficient: Trained on a minimal subset of just 198,609 images (0.2M) covering 7,401 perturbations.
  Contrast: Drastically smaller than OpenPhenom, Phenom-1, Phenom2 that require 5M to 93M+ images.

- Evaluation Benchmark (RxRx3-core):
  A rigorously curated, unbiased dataset used for zero-shot evaluation.
  Scale: Contains 1.33M images perturbed by 736 gene knockouts and 1,674 small molecules with corresponding gene/compound interaction label.
  Tasks: Assessed on Gene-Gene and Compound-Gene interaction retrieval.
  Validation: Model predictions are strictly validated against established biological databases (e.g., CORUM, HuMAP, Reactome, StringDB, ChEMBL).

Experiment results

- **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**

- **COMPOUND-GENE INTERACTION BENCHMARK**

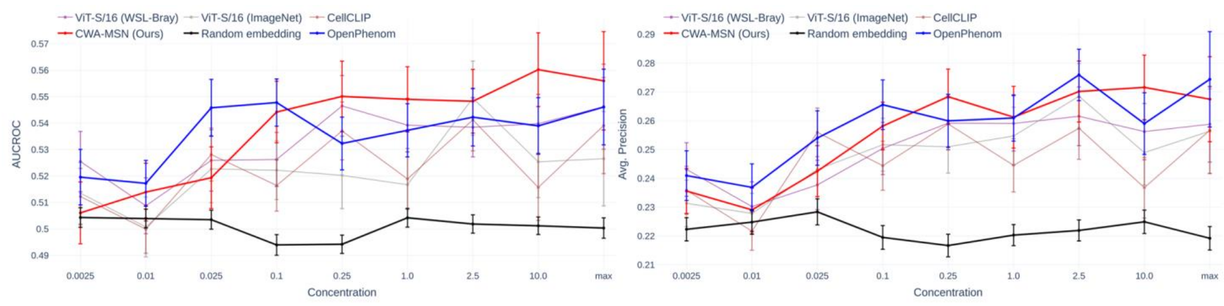


Figure 3: Compound-gene interaction benchmark graphs. AUC-ROC and AP values are reported over concentration.

- **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%**

- **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**

Contacts