Computational Molecular Biology and Bioinformatics

cpDistiller

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

October, 2025

Triple-effect correction for Cell Painting data

Cell Painting (CP) generates high-throughput cell-stained imaging data, which provides unique morphological insights for biological research. The CP data suffer from three types of technical effects (termed as triple-effect). These are batch effects and gradient-influenced row/ column effects (well-position effects).

- **Batch effect:** This is caused by systematic, non-biological variations in the data (e.g., a different plate, or microscope, or imaging session) that arise from differences in experimental conditions across batches.
- **Gradient-influenced row effect:** These are spatial biases within a single plate caused by variation in measured features depending on the row position of a well.
- **Gradient-influenced column effect:** These are spatial biases within a single plate caused by variation in measured features depending on the column position of a well.

What is cpDistiller?

The cpDistiller is a triple-effect correction method applicable to CP data [1]. It uses a pre-trained segmentation model coupled with a semi-supervised Gaussian mixture variational autoencoder employing contrastive and domain-adversarial learning.

The assays used for generating CP data contain multiple fluorescent dyes (or stains) that *p*aint different cellular components, allowing to measure how various treatments (e.g., drugs, genetic perturbations, or environmental changes) affect cell structure and organization.

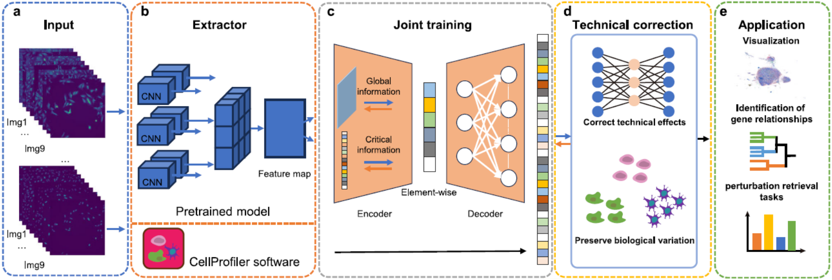
Each well in the CP assays contains nine 1080 x 1080 sub-images, captured across 5 channels: Mitochondria (Mito), Nucleus (DNA), Nucleoli and Cytoplasmic RNA (RNA), Endoplasmic Reticulum (ER), and Actin Cytoskeleton, Golgi and Plasmamembrane (AGP).

Overview of cpDistiller

The cpDistiller consists of three main modules as follows:

1. The extractor module.
2. The joint training module.
3. The technical correction module.

The cpDistiller framework



cpDistiller – The extractor

The extraction process is modeled as a segmentation task by transformig raw images into low-dimensional representations using an end-to-end process. For this, Mesmer [2], pre-trained on the TissueNet datasets, is chosen as the base model.

To align with Mesmer's specifications for pixel density and reduce computational resources, a tiling operation is applied. This adjusts the stride and overlap ratio to crop each large subimage into multiple small 256 × 256 images.

After preprocessing with Mesmer, the tiling operation is reversed to reconstruct the large feature map, followed by a 2D max pooling operation, and flattening to produce the embedding for each sub-image. The embeddings of the nine sub-images are sequentially concatenated to form a single overall embedding for each well.

cpDistiller – The joint training module

The joint training module integrates the features from CellProfiler and cpDistiller-extractor. To reduce potential noise and redundancy in the high-dimensional cpDistiller features, average pooling is used to reduce the feature dimensionality and smooth out irrelevant variation, resulting in the feature matrix $\mathbf{E}_{\mathbf{pooled}}$ for further processing. Further two approaches are employed to extract valuable information. In the first, we approximate principal component analysis (PCA) using a linear layer to obtain the low-dimensional representation for each well as follows:

$$\mathbf{Y}_1 = PCA(\mathbf{E}_{\mathbf{pooled}}).$$

In the second, $\mathbf{E}_{\mathbf{pooled}}$ is reshaped back into 2D feature maps and passed through an attention module to get the global information as follows:

$$\mathbf{Y}_2 = \mathrm{Reshape}(\mathbf{E}_{\mathbf{pooled}}) \odot \mathrm{Sigmoid}(\mathrm{Cov1d}(\mathrm{AvgPool}(\mathrm{Reshape}(\mathbf{E}_{\mathbf{pooled}})))).$$

cpDistiller – The joint training module

To combine critical and global information in the low-dimensional space, we apply element-wise addition as the encoding process for $\mathbf{E}_{\mathbf{pooled}}$. The latent representations for $\mathbf{E}_{\mathbf{pooled}}$ are computed as follows:

$$z_e = \mathbf{Y}_1 \oplus (W_1 \mathbf{Y}_2 + b_1).$$

Finally, we combine the CellProfiler-based features with the cpDistiller-extractor-based features transformed by the attention mechanism-based encoder-decoder architecture and feed the combined features into the technical correction module for further refinement.

cpDistiller – The technical correction module

It consists of three modules as listed below.

1. Gaussian mixture variational autoencoder (GMVAE) as the core component,
2. Gradient reversal module, and
3. Contrastive learning module.

cpDistiller – The technical correction module

Given the feature matrix $\mathbf{X}_p \in R^{n\times l_p}$ integrated by the joint training module, where n denotes the number of wells and l_p denotes the feature dimensions, the Gaussian mixture variational autoencoder (GMVAE) takes the input $\mathbf{X}_p$ to obtain low-dimensional representations $\mathbf{Z}_p$.

cpDistiller – The technical correction module

In the gradient reversal module, Within the low-dimensional space ($\mathbf{Z}_p$) derived from the GMVAE, discriminators are employed to identify the source of each well's representation $\mathbf{z}$, specifically the batch, row, and column labels of the corresponding well. To implement adversarial learning similar to generative adversarial networks (GANs), different batches, rows, and columns can be treated as distinct domains. We then employ domain-adversarial learning, specifically the gradient reversal layer (GRL), to remove triple effects across these domains.

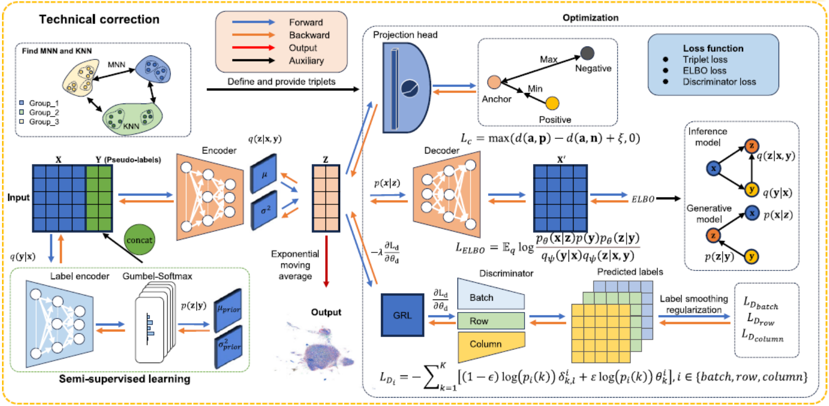
Here, discriminators are denoted as: D_{batch}, D_{row} and D_{column}, which are used to predict the batch, row, and column labels of $\mathbf{z}$.

cpDistiller – The technical correction module

In the contrastive learning module, we first establish nearest neighbor relationships by utilizing k-nearest neighbors (KNN) intra technical effects and mutual nearest neighbors (MNN) inter technical effects, based on the features from CellProfiler and cpDistiller-extractor, using cosine distance as the similarity metric.

This approach is applied to batch, row, and column effects, respectively. The intersection of the nearest neighbors is used to construct the adjacency matrix that captures nearest neighbor relationships between wells. We then leverage the relationships identified across multiple technical effects to form triplets and use triplet loss to restore more accurate nearest-neighbor relationships for each well.

The technical correction module



References

1. Yan, C., Zhang, Y., Feng, J., Hua, H., Ruan, Z., Li, Z., Li, S., Yan, C., Li, P., Liu, J. and Chen, S., Triple-effect correction for Cell Painting data with contrastive and domain-adversarial learning. Nature Communications, 16(1):6886, 2025.
2. Greenwald, N.F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C.C., McIntosh, B.J., Leow, K.X., Schwartz, M.S. and Pavelchek, C., Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. Nature Biotechnology, 40(4):555-565, 2022.