WHAT TO LEARN AND WHAT TO TRUST: MULTIMODAL LEARNING WITH RELIABLE SUPERVISION FOR BIOACTIVITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cell Painting phenotypes and molecular structures provide complementary views of compound activity, but learning from them is complicated by replicate variation, non-equivalent structure–phenotype relationships, and sparse assay supervision. Existing multimodal approaches typically align paired molecular and morphological observations, implicitly encouraging agreement between the two modality-specific representations. We propose **BioCoReT**, a two-stage framework that addresses **what to learn** from rich multimodal observations and **what to trust** under sparse assay supervision. First, BioCoReT performs contrastive learning across fused views of the same compound. It constructs biology-aware multimodal views from matched controls and Cell Painting replicates, and learns consistency across these compound-level views while preserving modality-specific information. Building on these representations, assay-specific specialists learn task-specific decision boundaries from the available labels and generate candidate predictions for unlabeled compound-task pairs. BioCoReT then evaluates each candidate using complementary evidence from teacher confidence, molecular and morphological topology, and replicate consistency, transferring only reliable task-specific knowledge to a unified multi-task student. BioCoReT consistently outperforms existing methods across three public Cell Painting benchmarks.

1 INTRODUCTION

Cell Painting has emerged as a scalable high-content profiling approach for characterizing compound-induced cellular responses from microscopy images Taylor (2008); Inglese et al. (2007); Blay et al. (2020); Bray et al. (2016). By profiling multiple cellular compartments, Cell Painting captures rich morphological changes in cellular state. Meanwhile, chemical structures provide information about compound identity, functional groups, scaffolds, and substructures Rogers & Hahn (2010); Weininger (1988). Integrating these two modalities Sanchez-Fernandez et al. (2023); Lu et al. (2025); Rao et al. (2025) offers a promising strategy for downstream tasks such as mechanism-of-action (MoA) classes prediction Tian et al. (2023), bioassay activities Moshkov et al. (2023).

Despite this progress, effectively learning from these modalities presents two coupled challenges. The first challenge is **what to learn** from heterogeneous multimodal observations. Existing multimodal approaches directly align paired molecular and Cell Painting observations to learn shared information Lu et al. (2025); Rao et al. (2025). However, this formulation is restrictive in our setting for two reasons. First, a single molecular structure is associated with multiple Cell Painting observations across experimental replicates, which can further be affected by technical variation Arevalo et al. (2024); Scherer (2009); Bray et al. (2016) (Fig. 1A.1). Second, structural and morphological similarities are not always consistent (Fig. 1A.2): structurally similar compounds can induce distinct phenotypes, as in activity cliffs Sanchez et al. (2026); Stumpfe & Bajorath (2012), whereas structurally dissimilar compounds can produce similar cellular responses, as in scaffold hopping Seal et al. (2022). Direct alignment may therefore not fully capture the information across these multimodal observations. The second challenge is **what to trust** under sparse and heterogeneous assay supervision. Assay labels require costly experiments and are available for only a small fraction of compound–task pairs Chandrasekaran et al. (2023); Moshkov et al. (2023), resulting in highly sparse assay label matrices de la Vega de León et al. (2018); Tian et al. (2024) (Fig. 1B.1). Moreover, differ-

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.

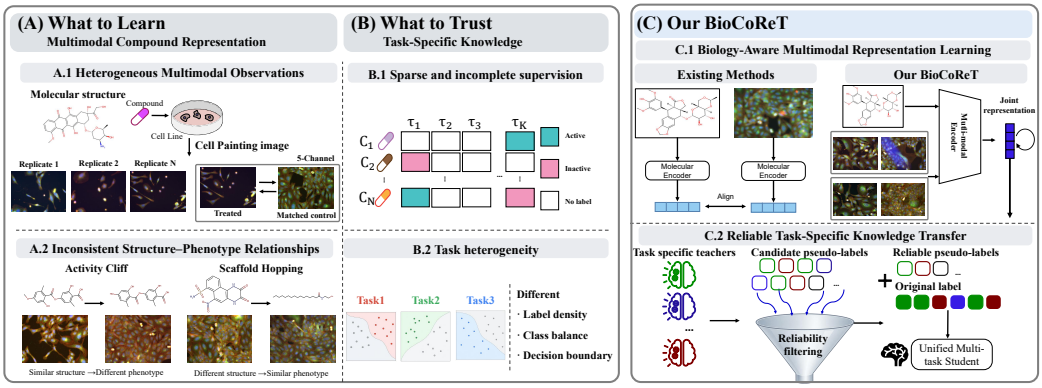

Figure 1: **Motivation and overview of BioCoReT.** (A) **What to Learn:** molecular structure and Cell Painting provide heterogeneous observations with replicate variation (A1), while their biological correspondence is non-equivalent, as illustrated by activity cliffs and scaffold hopping (A2). (B) **What to Trust:** assay supervision is sparse and incomplete (B1), and task heterogeneity in label density, class balance, and decision boundaries leads to uneven prediction reliability (B2). (C) **BioCoReT** addresses these challenges through biology-aware multimodal representation learning (C1) and reliable task-specific knowledge transfer to a unified multi-task student (C2).

ent assays exhibit distinct decision boundaries, label densities, and class balances Ramsundar et al. (2015); de la Vega de León et al. (2018) (Fig. 1B.2; Appendix B.1). Predictions for unobserved compound–task pairs provide a potential way to extend this limited supervision, but their reliability varies across compounds and assays. Using unreliable predictions as additional supervision can introduce noise and impair downstream learning Arazo et al. (2020). The key question is therefore which task-specific predictions are sufficiently reliable to support knowledge transfer.

To address these challenges, we propose **BioCoReT**, a two-stage framework that combines **Bio**logy-aware multimodal **Co**mpound representation learning (what to learn, Fig. 1C.1) with **Re**liable task-specific knowledge **T**ransfer (what to trust, Fig. 1C.2) for assay prediction. The first stage learns a task-agnostic compound representation from molecular structure and Cell Painting observations. BioCoReT does not directly align the two modalities. Instead, it constructs biology-aware multimodal views using matched controls and different Cell Painting replicates. It then learns consistency across fused views of the same compound. This helps preserve complementary information from heterogeneous multimodal observations. Building on these representations, the second stage learns and selectively transfers task-specific knowledge under sparse assay supervision. For each assay, we train specialist task teachers to learn their distinctive decision boundary and generate candidate predictions for unlabeled compound-task pairs. These predictions are further evaluated using complementary evidence from teacher confidence, molecular and morphological topology, and replicate consistency. Reliable pseudo-labels are then combined with assay labels to train a unified student.

Across multiple Cell Painting benchmarks and supervision regimes, BioCoReT consistently improves multimodal prediction, with analyses further showing that its two-stage design learns informative compound representations and enables reliable task-specific knowledge transfer.

2 Related Work

Multimodal learning for phenotypic drug discovery. Cell Painting provides high-content morphological phenotypes of compound perturbations, often with matched controls and multiple experimental replicates Gustafsdottir et al. (2013); Cimini et al. (2023); Caicedo et al. (2017); Arevalo et al. (2024). Together with molecular structure, these observations provide complementary information for compound activity prediction Moshkov et al. (2023); Zheng et al. (2024); Rao et al. (2025); Yu et al. (2026). Existing multimodal approaches either combine the two modalities under downstream supervision or learn representations through cross-modal contrastive alignment Sanchez-Fernandez et al. (2023); Zheng et al. (2024); Rao et al. (2025). The latter typically treats a paired molecular

structure and morphological observation as the contrastive unit and encourages agreement between their modality-specific representations. In our setting, however, one molecule is associated with multiple Cell Painting replicates, while structure–phenotype correspondence is not one-to-one, as discussed in Sec. 1. BioCoReT instead first combines molecular and morphological information into fused multimodal views of the same compound. Different Cell Painting replicates provide biologically meaningful variations of these views, and contrastive learning is performed across the fused views rather than directly between the two modalities.

Learning with incomplete multi-assay supervision. Canonical semi-supervised methods exploit unlabeled samples through teacher–student consistency or confidence-based pseudo-labeling Tarvainen & Valpola (2017); Sohn et al. (2020); Arazo et al. (2020); Wang et al. (2022b). Our setting differs in that supervision is partially observed over a compound–task matrix: the same compound may be labeled for some assays but unlabeled for others, while assays differ in label density, class balance, and decision boundaries Ramsundar et al. (2015); Mayr et al. (2018). Consequently, the reliability of a candidate pseudo-label is both compound- and task-dependent. Existing confidence-based or multimodal semi-supervised methods are not designed to explicitly verify this task-specific reliability using complementary structural, morphological, and replicate evidence. BioCoReT instead trains assay-specific specialists and transfers a missing compound–task prediction only when teacher confidence is supported by molecular topology, morphology topology, and replicate consistency, before integrating the selected knowledge into a unified multi-task student.

3 METHOD

3.1 PROBLEM FORMULATION

For each compound $i \in \{1, \ldots, N\}$, we consider its molecular structure x_i^{mol} and Cell Painting images acquired across multiple experimental runs (referred to as replicates), denoted as $\mathcal{X}_i^{\text{cp}} = \left\{\left(x_{ir}^{\text{cp,tr}}, x_{ir}^{\text{cp,ctrl}}\right)\right\}_{r=1}^{R_i}$, where $x_{ir}^{\text{cp,tr}}$ denotes the treated image, $x_{ir}^{\text{cp,ctrl}}$ denotes the matched untreated control image for replicate r, and R_i is the number of available replicates.(Fig. 2 A.1-2).

We consider a dataset of N compounds $\mathcal{D} = \left\{\left(\mathcal{X}_i^{\text{cp}}, x_i^{\text{mol}}, \mathbf{y}_i\right)\right\}_{i=1}^{N}$,where $\mathbf{y}_i = [y_{i\tau_1}, \ldots, y_{i\tau_K}]$ is a sparse label vector over K binary assay tasks, denoted by $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$.(Fig. 2 A.3). For each compound-task pair (i, τ), where $\tau \in \mathcal{T}$, we define the assay label as $y_{i\tau} \in \{0, 1, \perp\}$, where 0 and 1 indicate inactive and active assay responses, and $\perp$ denotes a missing label. We define the observed and missing compound-task pairs as $\Omega_L = \{(i, \tau) : y_{i\tau} \neq \perp\}$ and $\Omega_U = \{(i, \tau) : y_{i\tau} = \perp\}$.

Given these multimodal inputs, we aim to predict compound activity across multiple assays. Formally, we learn a multimodal predictor $\mathcal{F}\left(\mathcal{X}_i^{\text{cp}}, x_i^{\text{mol}}\right) \to \widehat{\mathbf{y}}_i \in [0, 1]^K$, where $\widehat{y}_{i\tau}$ denotes the predicted probability that compound i is active in assay τ. We address two complementary questions: **what to learn** from abundant multimodal data and **what to trust** under sparse task-specific supervision. Accordingly, our framework follows a two-stage design. The first stage learns a task-agnostic multimodal representation from unlabeled paired data. The second stage trains assay-specific teachers under sparse supervision and transfers reliable predictions to a unified student.

3.2 WHAT TO LEARN: BIOLOGY-AWARE MULTIMODAL REPRESENTATION LEARNING

The first stage addresses what to learn from abundant unlabeled paired Cell Painting and molecular data. We propose control-conditioned encoding to capture compound-induced morphological changes relative to controls(Sec. 3.2.1, Fig. 2B.1). We then introduce biology-aware multimodal contrastive learning to learn consistent fused representations across augmented multimodal views(Sec. 3.2.2, Fig. 2B.2). Finally, we align unimodal representations with the multimodal space to retain information from both modalities (Sec. 3.2.3, Fig. 2B.2).

3.2.1 CONTROL-CONDITIONED REPRESENTATION ENCODING (CCRE)

Cell Painting images capture compound-induced morphological changes together with experimental variation. We propose a control-conditioned encoding that uses the matched control as an

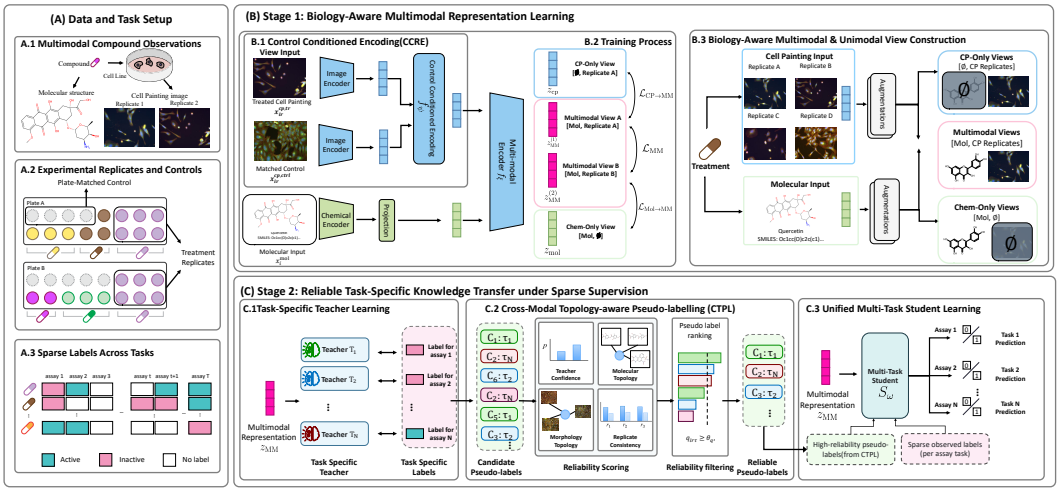

Figure 2: **Overview of BioCoReT.** (A) Data and task setup with multimodal compound observations, experimental replicates and matched controls, and sparse assay labels. (B) Stage 1 learns biology-aware compound representations through control-conditioned Cell Painting encoding and consistency learning across multimodal and unimodal views. (C) Stage 2 trains task-specific teachers, evaluates candidate pseudo-labels with CTPL using complementary evidence, and transfers high-reliability pseudo-labels together with observed labels to a unified multi-task student.

experimental reference. For replicate r of compound i, the treated and control images are encoded using a pretrained image encoder f_ϕ Oquab et al. (2023), yielding $h_{ir}^{\mathrm{cp,tr}} = f_\phi(x_{ir}^{\mathrm{cp,tr}})$ and $h_{ir}^{\mathrm{cp,ctrl}} = f_\phi(x_{ir}^{\mathrm{cp,ctrl}})$. We use a gated residual mechanism to adaptively incorporate control-conditioned changes:

$$z_{ir}^{\mathrm{cp}} = h_{ir}^{\mathrm{cp,tr}} + \alpha\, g\Big(h_{ir}^{\mathrm{cp,tr}}, h_{ir}^{\mathrm{cp,ctrl}}\Big) \odot \varphi\Big(\mathrm{Norm}\left(h_{ir}^{\mathrm{cp,tr}} - h_{ir}^{\mathrm{cp,ctrl}}\right)\Big), \tag{1}$$

where g is a learnable gate controlling the contribution of the treatment-control residual, and α is zero-initialized. In parallel, the molecular structure x_i^{mol} is encoded using a pretrained molecular encoder Chithrananda et al. (2020), yielding the molecular representation z_i^{mol}.

3.2.2 Biology-Aware Multimodal Contrastive Learning

Cell Painting contains multiple replicates of the same compound perturbation. We exploit this structure by using different replicates to construct views of the same compound. Meanwhile, molecular structure and morphological phenotype are not in one-to-one correspondence: structurally similar compounds can induce distinct phenotypes (activity cliffs), whereas structurally dissimilar compounds can produce similar phenotypes (scaffold hopping). Therefore, we avoid direct alignment between the two modalities and instead apply contrastive learning to fused multimodal views of the same compound to preserve biologically meaningful information.

For each paired Cell Painting-molecular sample, we construct two independently augmented multimodal views. Both modalities are augmented with Gaussian noise and random feature dropping Chen et al. (2020). For Cell Painting, we additionally use replicate sampling as an extra form of view augmentation. This provides biologically real experimental views of the same compound perturbation beyond synthetic augmentations. Motivated by recent works Dufumier et al. (2025); Liang et al. (2023), the augmented modalities are jointly encoded by a multimodal Transformer f_ξ. With the augmentation operators as $\mathrm{Aug}_{\mathrm{cp}}$ and $\mathrm{Aug}_{\mathrm{mol}}$, each multimodal view is represented as

$$z_{\mathrm{MM},i}^{(k)} = f_\xi\left([\mathrm{Aug}_{\mathrm{cp}}^{(k)}(z_{ir_k}^{\mathrm{cp}}), \mathrm{Aug}_{\mathrm{mol}}^{(k)}(z_i^{\mathrm{mol}})]\right), \qquad k \in \{1,2\}. \tag{2}$$

The two multimodal views of the same compound form a positive pair, while views from different compounds serve as negatives. Using the symmetric InfoNCE estimator $\widehat{I}_{\mathrm{NCE}}(\cdot,\cdot)$ Oord et al.

(2018); He et al. (2020), we define $\mathcal{L}_{\mathrm{MM}} = -\widehat{I}_{\mathrm{NCE}}\left(z_{\mathrm{MM}}^{(1)}, z_{\mathrm{MM}}^{(2)}\right)$. This objective encourages consistent multimodal representations across different views that preserve stable biological information.

3.2.3 Unimodal-to-Multimodal Alignment

Cell Painting exhibits replicate variation, while structurally similar compounds can induce distinct morphological responses (activity cliffs) and structurally dissimilar compounds can produce similar phenotypes (scaffold hopping). Aligning only multimodal views may overlook information from individual modalities. We further align the unimodal representations with their multimodal representations, encouraging the shared space to retain information from both modalities.

We construct Cell Painting-only and molecule-only representations by masking the other modality: $z_{\mathrm{CP},ir} = g_{\mathrm{CP}}\left(f_\xi\left([z_{ir}^{\mathrm{cp}}, \varnothing]\right)\right)$, $z_{\mathrm{Mol},i} = g_{\mathrm{Mol}}\left(f_\xi\left([\varnothing, z_i^{\mathrm{mol}}]\right)\right)$, where $\varnothing$ denotes the missing modality. Each unimodal representation is aligned with the two augmented multimodal views:

$$\mathcal{L}_{\mathrm{CP}\to\mathrm{MM}} = -\frac{1}{2}\sum_{k=1}^{2}\widehat{I}_{\mathrm{NCE}}\left(z_{\mathrm{CP}}, z_{\mathrm{MM}}^{(k)}\right), \mathcal{L}_{\mathrm{Mol}\to\mathrm{MM}} = -\frac{1}{2}\sum_{k=1}^{2}\widehat{I}_{\mathrm{NCE}}\left(z_{\mathrm{Mol}}, z_{\mathrm{MM}}^{(k)}\right). \tag{3}$$

The overall representation-learning objective is $\mathcal{L}_{\mathrm{rep}} = \mathcal{L}_{\mathrm{MM}} + \lambda_{\mathrm{CP}}\mathcal{L}_{\mathrm{CP}\to\mathrm{MM}} + \lambda_{\mathrm{Mol}}\mathcal{L}_{\mathrm{Mol}\to\mathrm{MM}}$, where λ_{CP} and λ_{Mol} control the contributions of the two unimodal alignment terms.

3.3 What to Trust: Reliable Task-Specific Knowledge Transfer under Sparse Supervision

The second stage addresses what to trust under sparse and heterogeneous supervision. We train a task-specific specialist teacher for each assay to learn its decision boundary from observed labels and generate candidate predictions for missing compounds within the same task (Sec. 3.3.1, Fig. 2C.1). We then assess these predictions' reliability using teacher confidence, molecular and morphology topology, and replicate consistency, retaining only reliable predictions as pseudo-labels (Sec. 3.3.2, Fig. 2C.2). Finally, we combine the reliable pseudo-labels with observed labels to train a unified multi-task student that integrates task-specific knowledge across assays (Sec. 3.3.3, Fig. 2C.3)

3.3.1 Task-Specific Teacher Learning

For each assay task $\tau \in \mathcal{T}$, we train a task-specific specialist teacher T_τ using the observed labels of that task. The teacher learns the task-specific decision boundary from the Stage-1 multimodal representations and generates candidate predictions for missing compounds within the same task.

Specifically, for replicate r of compound i, we use the Stage-1 multimodal representation z_{ir}^{MM}. The teacher predicts $p_{ir\tau}^T = T_\tau\left(z_{ir}^{\mathrm{MM}}\right)$, where $p_{ir\tau}^T \in [0,1]$ denotes the predicted probability that compound i is active on assay τ based on replicate r. For each missing compound-task pair $(i,\tau) \in \Omega_U$, the corresponding task-specific teacher generates a candidate pseudo-label for replicate r: $\widetilde{y}_{ir\tau} = 1$ if $p_{ir\tau}^T \geq \theta_+$ and $\widetilde{y}_{ir\tau} = 0$ if $p_{ir\tau}^T \leq \theta_-$; predictions between the two thresholds are discarded. The teacher confidence is defined as $c_{ir\tau} = \max\left(p_{ir\tau}^T, 1 - p_{ir\tau}^T\right)$.

3.3.2 Cross-modal Topology-aware Pseudo-Labeling (CTPL)

Candidate predictions may have different reliability across compounds and assay tasks. We therefore evaluate each candidate using complementary biological evidence.

Topology support. The local neighborhood of a compound provides task-specific evidence for evaluating a candidate prediction. For modality $m \in \{\mathrm{mol}, \mathrm{cp}\}$, let $\mathcal{N}_m(i)$ denote the K nearest neighbors of compound i, constructed using the corresponding molecular or Cell Painting representations. For each replicate-level candidate pseudo-label, we measure whether the observed labels of neighboring compounds support the candidate prediction:

$$s_{ir\tau}^m = \frac{\sum_{j\in\mathcal{N}_m(i)} \mathbb{I}[(j,\tau)\in\Omega_L]\mathbb{I}[y_{j\tau} = \widetilde{y}_{ir\tau}]}{\sum_{j\in\mathcal{N}_m(i)} \mathbb{I}[(j,\tau)\in\Omega_L] + \epsilon}. \tag{4}$$

This yields molecular and morphology support scores $s_{ir\tau}^{\mathrm{mol}}$ and $s_{ir\tau}^{\mathrm{cp}}$, respectively.

Replicate consistency. For compounds with multiple Cell Painting replicates, we further assess whether the teacher predictions are consistent across experimental replicates. Specifically, we define

$$s_{i\tau}^{\mathrm{rep}} = 1 - \frac{2}{R_i(R_i-1)} \sum_{r<r'} \left| p_{ir\tau}^{T} - p_{ir'\tau}^{T} \right|. \tag{5}$$

A high score indicates that the task-specific prediction is stable across different experimental setting of the same compound.

Reliability aggregation. We combine the available reliability signals using a weighted score:

$$q_{ir\tau} = \frac{\alpha c_{ir\tau} + \beta s_{ir\tau}^{\mathrm{mol}} + \gamma s_{ir\tau}^{\mathrm{cp}} + \delta m_i^{\mathrm{rep}} s_{i\tau}^{\mathrm{rep}}}{\alpha + \beta + \gamma + \delta m_i^{\mathrm{rep}}}, \tag{6}$$

where $m_i^{\mathrm{rep}} = \mathbb{I}[R_i > 1]$ indicates whether replicate consistency is available. When multiple replicates are unavailable, the replicate consistency term is excluded from reliability estimation.

Finally, we retain a candidate pseudo-label as reliable supervision when $q_{ir\tau} \geq \theta_q$, where θ_q is the reliability threshold. The selected pseudo-labels form the set $\mathcal{B}_{\mathrm{PL}}$ used to train the unified student.

3.3.3 Unified Multi-Task Student Learning

The task-specific teachers provide reliable supervision for individual assays. We integrate this supervision by training a unified multi-task student S_ω jointly across all assay tasks.

Given the Stage-1 multimodal representation z_{ir}^{MM} of replicate r from compound i, the student predicts $\widehat{\mathbf{y}}_{ir} = S_\omega\left(z_{ir}^{\mathrm{MM}}\right) \in [0,1]^K$, where $\widehat{y}_{ir\tau}$ denotes the prediction for assay τ. For observed compound–task pairs, the supervised objective is $\mathcal{L}_{\mathrm{sup}} = \sum_{(i,\tau)\in\Omega_L} \sum_{r=1}^{R_i} \mathrm{BCE}\left(\widehat{y}_{ir\tau}, y_{i\tau}\right)$.

Let $\mathcal{B}_{\mathrm{PL}}$ denote the reliable pseudo-labels selected by CTPL. We use them as additional supervision weighted by their reliability scores: $\mathcal{L}_{\mathrm{PL}} = \sum_{(i,r,\tau,\widetilde{y}_{ir\tau},q_{ir\tau})\in\mathcal{B}_{\mathrm{PL}}} q_{ir\tau}\,\mathrm{BCE}\left(\widehat{y}_{ir\tau}, \widetilde{y}_{ir\tau}\right)$. The final objective is $\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{PL}}\mathcal{L}_{\mathrm{PL}}$, where λ_{PL} is a hyperparameter.

3.4 Training and Inference

The framework is trained in two stages. In Stage 1, multimodal representations are learned from paired Cell Painting and molecular data using $\mathcal{L}_{\mathrm{rep}}$, without assay supervision. In Stage 2, task-specific teachers are trained on the observed assay labels, CTPL selects reliable pseudo-labels, and the unified multi-task student is trained using both observed labels and selected pseudo-labels. At inference time, only the unified multi-task student is retained for assay prediction.

4 Experiments

We conduct a series of experiments to evaluate BioCoReT. First, we benchmark its predictive performance across Cell Painting datasets with diverse assay tasks and supervision regimes, including controlled reductions in label availability. Second, we evaluate the quality of Stage-1 representations and study how they benefit from increasing amounts of unlabeled multimodal data. Finally, we examine the reliability of task-specific knowledge transfer in Stage 2.

Experimental settings. We evaluate BioCoReT on three public Cell Painting benchmarks, cpg0012 Hofmarcher et al. (2019), cpg0016 Fredin Haslum et al. (2024), and cpg0037 Ewald et al. (2026), covering 209, 29, and 354 assay tasks, respectively. These benchmarks represent different supervision regimes: cpg0012 has many compounds and tasks but sparse labels; cpg0016 also has many compounds but fewer tasks; cpg0037 has fewer compounds but more tasks and denser labels. Detailed dataset statistics are provided in Appendix B.1.

Evaluation. Predictions are aggregated at the compound level before evaluation. We report mean AUROC, AUPRC, and F1 across tasks. Additional evaluation details are provided in Appendix B.6.

Model	Modality		cpg0012 (209 tasks)			cpg0016 (29 tasks)			cpg0037 (354 tasks)		
	CP	Mol	AUROC ↑	AUPRC ↑	F1 ↑	AUROC ↑	AUPRC ↑	F1 ↑	AUROC ↑	AUPRC ↑	F1 ↑
CellProfiler + Head Stirling et al. (2021)	✓		0.705	0.616	0.512	0.663	0.662	0.600	0.600	0.365	0.083
DINO + Head Oquab et al. (2023)	✓		0.713	0.630	0.505	0.608	0.616	0.570	0.604	0.369	0.080
HTI-CNN Hofmarcher et al. (2019)	✓		0.718	0.646	0.489	0.668	0.662	0.574	0.618	0.367	0.059
ECFP4 + Head Rogers & Hahn (2010)		✓	0.662	0.538	0.299	0.707	0.686	0.603	0.600	0.351	0.117
ChemBERTa + Head Chithrananda et al. (2020)		✓	0.626	0.511	0.343	0.671	0.652	0.580	0.646	0.360	0.102
CLOOME + HeadSanchez-Fernandez et al. (2023)	✓	✓	0.696	0.624	0.436	0.591	0.609	0.546	0.593	0.361	0.072
CellCLIP + Head Lu et al. (2025)	✓	✓	0.737	0.685	0.525	0.650	0.649	0.585	0.637	0.373	0.073
MICON + Head Yu et al. (2026)	✓	✓	0.704	0.630	0.475	0.615	0.616	0.540	0.577	0.344	0.063
InfoCORE + Head Wang et al. (2024)	✓	✓	0.720	0.632	0.470	0.628	0.626	0.572	0.632	0.372	0.058
MINERRao et al. (2025)	✓	✓	0.668	0.553	0.372	0.692	0.671	0.598	0.629	0.351	0.049
CP-Chem-MoATian et al. (2023)	✓	✓	0.745	0.684	0.519	0.742	0.729	0.610	0.618	0.373	0.093
Our BioCoReT	✓	✓	**0.776**	**0.705**	**0.570**	**0.752**	**0.731**	**0.655**	**0.664**	**0.398**	**0.159**

Table 1: Main results on three Cell Painting benchmarks. We report mean task-level AUROC, AUPRC, and F1. "+ Head" denotes a common downstream multi-task prediction head applied to the corresponding representation. **Bold** indicates the best result in each column.

4.1 Main Results under Naturally Sparse Assay Supervision

We first evaluate BioCoReT under the original observed assay supervision of each benchmark. All methods are evaluated using the same compound-level data splits and observed compound-task labels. We compare against Cell Painting-only, molecular-only, and multimodal methods. For representation-based baselines, we train the same downstream multi-task prediction head for their representations, while other methods are adapted to the same multi-task prediction setting. Performance is evaluated at the compound level using mean task-level AUROC, AUPRC, and F1.

As shown in Table 1, BioCoReT achieves the best performance across all three benchmarks and evaluation metrics. On cpg0012, it reaches 0.776 AUROC, outperforming the strongest competing method by 0.031, with consistent improvements in AUPRC and F1. On cpg0016, where CP-Chem-MoA already provides a strong multimodal baseline, BioCoReT further improves AUROC to 0.752 and yields a more pronounced gain in F1, from 0.610 to 0.655. On cpg0037, it achieves 0.664 AUROC, exceeding the strongest competing result by 0.018 while also improving AUPRC and F1.

The relative strengths of the baselines vary across datasets: multimodal methods are particularly competitive on cpg0012 and cpg0016, whereas strong molecular and multimodal baselines emerge on different metrics of cpg0037. BioCoReT nevertheless achieves the best performance across all three settings, showing consistent gains over both unimodal and existing multimodal approaches.

4.2 Performance under Reduced Label Supervision

We further evaluate BioCoReT under controlled reductions of assay supervision on cpg0037. We retain 50%, 10%, and 5% of the originally observed compound–task labels. We choose cpg0037 because its relatively dense assay annotations allow meaningful evaluation under progressively reduced supervision. In addition to the supervised baselines, we compare with several semi-supervised methods adapted to the same compound-level prediction setting. All methods use the same data splits and observed-label masks, allowing a controlled comparison under sparse compound–task supervision.

As shown in Table 2, performance generally decreases as fewer assay labels are available, while BioCoReT remains consistently strong. With only 5% of the observed labels, BioCoReT achieves 0.605 AUROC, compared with 0.585 for the strongest semi-supervised baseline. The same overall advantage is maintained at 10%, 50%, and full supervision, with similar trends in AUPRC.

These results show that BioCoReT makes effective use of limited assay supervision and remains robust as the observed compound–task labels become increasingly sparse.

4.3 Contributions of the Two Stages

We examine the individual and combined contributions of Stage 1 and Stage 2 on cpg0012 under 10% and full supervision. "+ Stage 1" uses the proposed representation learning only. "+ Stage 2" applies reliable task-specific knowledge transfer directly to the base representations, without Stage 1. As shown in Fig. 3(a), both stages improve over the base model.

Model	Modality		AUROC / AUPRC ↑			
	CP	Mol	5%	10%	50%	100%
Supervised baselines						
ChemBERTa + Head Chithrananda et al. (2020)		✓	0.590 / 0.325	0.613 / 0.330	0.627 / 0.345	0.646 / 0.360
DINO + Head Oquab et al. (2023)	✓		0.510 / 0.294	0.538 / 0.306	0.583 / 0.343	0.604 / 0.369
CellCLIP + Head Lu et al. (2025)	✓	✓	0.566 / 0.299	0.586 / 0.316	0.619 / 0.333	0.637 / 0.373
Semi-supervised baselines						
Co-training Blum & Mitchell (1998)	✓	✓	0.528 / 0.301	0.577 / 0.326	0.580 / 0.329	0.618 / 0.354
CoMatch Li et al. (2021)	✓		0.585 / 0.332	0.600 / 0.341	0.615 / 0.351	0.629 / 0.360
Multimodal SimMatch Zheng et al. (2022)	✓	✓	0.572 / 0.314	0.572 / 0.319	0.608 / 0.331	0.621 / 0.348
MMatch Wang et al. (2022a)	✓	✓	0.578 / 0.323	0.581 / 0.326	0.594 / 0.331	0.618 / 0.354
CP-SemiSupCon Bushiri Pwesombo et al. (2025)	✓		0.571 / 0.318	0.575 / 0.320	0.603 / 0.369	0.609 / 0.373
STiL Du et al. (2025)	✓	✓	0.572 / 0.326	0.591 / 0.331	0.611 / 0.338	0.630 / 0.367
Our BioCoReT	✓	✓	**0.605 / 0.351**	**0.629 / 0.358**	**0.652 / 0.388**	**0.664 / 0.398**

Table 2: Performance on cpg0037 under different fractions of observed assay labels. Each entry reports AUROC / AUPRC. All methods use the same data splits and observed-label masks. Bold indicates the best performance for each metric at each label fraction.

Labels	Setting	AUROC ↑	AUPRC ↑	F1 ↑
10%	Base	0.673	0.576	0.460
	+ Stage 1	0.716	0.618	0.503
	+ Stage 2	0.703	0.614	0.518
	Full	**0.721**	**0.631**	**0.524**
100%	Base	0.739	0.681	0.517
	+ Stage 1	0.766	0.693	0.559
	+ Stage 2	0.769	0.690	0.569
	Full	**0.776**	**0.705**	**0.570**

Representation	AUROC ↑	AUPRC ↑
Raw CP	0.654	0.568
Raw Mol	0.598	0.484
Raw CP + Mol	0.671	0.577
Stage 1 w/o CCRE	0.711	0.609
Stage 1 w/o $\mathcal{L}_{\text{CP/Mol}\to\text{MM}}$	0.709	0.601
Stage-1 full	**0.716**	**0.618**

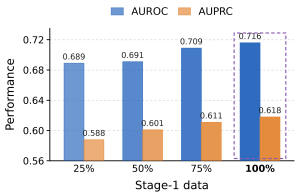

Figure 3: Analysis of the two-stage design on cpg0012. **(a)** Ablation of Stage 1 and Stage 2 using 10% and 100% of the assay labels. **(b)** Frozen linear probing of raw representations and Stage-1 variants. **(c)** Stage-1 representation quality with increasing amounts of unlabeled multimodal data.

Under 10% supervision, Stage 1 improves AUROC from 0.673 to 0.716 and AUPRC from 0.576 to 0.618, indicating that representation learning is particularly beneficial when labeled supervision is limited. Stage 2, in contrast, gives a larger improvement in F1, from 0.460 to 0.518, showing the benefit of transferring reliable task-specific supervision. Under full supervision, both stages again improve over the base model, with Stage 2 providing slightly larger gains in AUROC and F1, while Stage 1 gives slightly higher AUPRC. Combining the two stages achieves the best overall performance under both supervision settings, suggesting that representation learning and reliable task-specific transfer provide complementary rather than redundant benefits.

4.4 Analysis of Stage-1 Representation Learning

We further analyze whether Stage 1 learns informative multimodal representations and which components contribute to its performance. All representations are evaluated using frozen linear probing.

Representation quality and component ablation. As shown in Fig. 3(b), combining the raw Cell Painting and molecular representations improves over either modality alone. It reaches 0.671 AUROC and 0.577 AUPRC. The full Stage 1 representation further improves performance to 0.716 AUROC and 0.618 AUPRC. This indicates that its gains go beyond direct feature combination. Removing either CCRE or the unimodal-to-multimodal alignment objective degrades performance. This shows that both components contribute to the learned representation.

Scaling with unlabeled multimodal data. We then keep the downstream protocol fixed and progressively increase the amount of multimodal data used for Stage-1 pretraining. As shown in Fig. 3(c), performance generally improves with more pretraining data. AUROC increases from 0.689 using 25% of the available compounds to 0.716 using the full set. AUPRC shows a similar improvement. These results show that Stage 1 benefits from scaling unlabeled multimodal data.

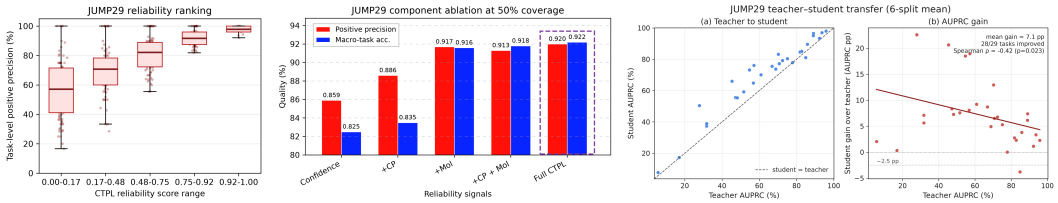

Figure 4: Analysis of reliable task-specific knowledge transfer. **(a)** Task-level positive precision across CTPL reliability-score bins. **(b)** Component ablation of confidence, morphology, molecular, and CTPL reliability signals. **(c)** Comparison between task-specific teachers and the unified student.

4.5 Reliable Task-Specific Knowledge Transfer

We evaluate whether Stage 2 can identify reliable task-specific predictions and transfer this knowledge to the unified multi-task student. To directly assess pseudo-label reliability, we hide a subset of observed assay labels during pseudo-label generation and use them only for evaluation.

Reliability of selected task-specific predictions. Under the default threshold ($\theta_q = 0.8$), CTPL retains 34.9%, 32.0%, and 12.7% of the candidate predictions on cpg0016, cpg0012, and cpg0037, respectively. The corresponding hidden-label accuracies are 96.2%, 97.5%, and 98.1%. These results show that CTPL retains a substantial set of predictions while maintaining high reliability.

Reliability ranking. We next examine whether the CTPL reliability score can distinguish predictions. We group candidate pseudo-labels by their reliability scores and evaluate the hidden-label precision within each group. As shown in Fig. 4(a), task-level positive precision increases with the reliability score, showing that the score provides a meaningful ranking of prediction quality.

Contribution of reliability signals. We compare different reliability signals at a fixed pseudo-label coverage of 50%, so that all methods select the same number of predictions. As shown in Fig. 4(b), teacher confidence alone gives lower selection quality. Adding morphology or molecular topology improves the results. Combining CP and molecular evidence further improves accuracy from 0.916 to 0.918, while positive precision decreases from 0.917 to 0.913. This shows that the two modalities provide complementary but sometimes conflicting evidence. Adding replicate consistency gives the full CTPL model the best overall performance, reaching 0.920 positive precision and 0.922 accuracy.

Specialist-to-generalist transfer. Finally, we evaluate whether the selected task-specific knowledge can be effectively transferred to the unified student. The high hidden-label precision above measures only the selected pseudo-labels. It does not imply that the task-specific teachers are equally accurate over all compounds. As shown in Fig. 4(c), the student outperforms the corresponding teacher on 28 of 29 tasks. The mean AUPRC gain is 7.1 percentage points. The gain is larger for tasks with weaker teachers (Spearman $\rho = -0.42$, $p = 0.023$). These results show that reliable task-specific predictions can be transferred to a unified student and improve task-level prediction.

5 Conclusion

We presented **BioCoReT**, a two-stage semi-supervised multimodal framework for bioactivity prediction under sparse and heterogeneous assay supervision. BioCoReT addresses **what to learn** from rich multimodal observations and **what to trust** under incomplete supervision. The first stage learns biology-aware compound representations from fused molecular and Cell Painting views, while the second stage selectively transfers reliable task-specific predictions to a unified multi-task student. Across three Cell Painting benchmarks, BioCoReT consistently improves bioactivity prediction and remains effective as assay supervision becomes increasingly sparse. Further analyses show that the two stages provide complementary benefits, unlabeled multimodal data improves representation quality, and reliability-aware selection enables effective task-specific knowledge transfer.

AI USE STATEMENT

In this work, generative AI tools were used as an editorial and discussion aid for language editing, translation, manuscript organization, and improving the clarity and presentation of the paper. They were also used to provide limited feedback during the discussion of methodological and experimental choices and the interpretation of experimental results. All AI-assisted suggestions were critically reviewed and revised by the authors. The authors independently verified the technical content, experimental results, and scientific claims, and take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work is a computational study based on existing public Cell Painting and bioassay datasets. It does not involve the recruitment of human participants or the collection of new human or animal experimental data. Some of the analyzed datasets contain measurements derived from human cells, and our study performs only secondary computational analysis of these previously released data. BioCoReT is intended for research in bioactivity prediction and early-stage drug discovery rather than for clinical diagnosis or treatment decisions. Model predictions should therefore be interpreted as computational predictions and supported by appropriate downstream experimental validation when used to draw biological conclusions.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed dataset statistics and input processing in Appendices B.1 and B.2, data-split protocols in Appendix B.3, and baseline adaptations in Appendix B.4. The sparse-supervision and evaluation protocols are described in Appendices B.5 and B.6. Model architectures, optimization settings, and training hyperparameters are provided in Appendix B.7, while Appendix B.8 further documents the repeated split/fold runs and computational setup used in our experiments.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

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, 2024.

Vincent Blay, Bhairavi Tolani, Sunita P Ho, and Michelle R Arkin. High-throughput screening: today's biochemical and cell-based approaches. *Drug discovery today*, 25(10):1807–1821, 2020.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

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.

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

Nathan Brown. Identifying and representing scaffolds. *Scaffold Hopping in Medicinal Chemistry*, pp. 1–14, 2013.

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, January 2025. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.4c00835.

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.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, pp. 2023–03, 2023.

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.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Beth A Cimini, Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Byrne, Sakshi Garg, Nasim Jamali, David J Logan, et al. Optimizing the cell painting assay for image-based profiling. *Nature protocols*, 18(7):1981–2013, 2023.

Antonio de la Vega de León, Beining Chen, and Valerie J Gillet. Effect of missing data on multitask prediction methods. *Journal of cheminformatics*, 10(1):26, 2018.

Siyi Du, Xinzhe Luo, Declan P O'Regan, and Chen Qin. Stil: Semi-supervised tabular-image learning for comprehensive task-relevant information exploration in multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15549–15559, 2025.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Jessica D Ewald, Katherine L Titterton, Alex Bäuerle, Alex Beatson, Daniil A Boiko, Ángel A Cabrera, Jaime Cheah, Beth A Cimini, Bram L Gorissen, Joshua Harrill, et al. Cell painting for cytotoxicity and mode-of-action analysis in primary human hepatocytes. *Cell Systems*, 17(5), 2026.

Johan Fredin Haslum, Charles-Hugues Lardeau, Johan Karlsson, Riku Turkki, Karl-Johan Leuchowius, Kevin Smith, and Erik Müllers. Cell painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. *Nature communications*, 15(1):3470, 2024.

Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman A Carrel, Todd R Golub, Stuart L Schreiber, et al. Multiplex cytological profiling assay to measure diverse cellular states. *PloS one*, 8(12):e80999, 2013.

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.

Markus Hofmarcher, Elisabeth Rumetshofer, Djork-Arne Clevert, Sepp Hochreiter, and Gunter Klambauer. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of chemical information and modeling*, 59(3):1163–1171, 2019.

James Inglese, Ronald L Johnson, Anton Simeonov, Menghang Xia, Wei Zheng, Christopher P Austin, and Douglas S Auld. High-throughput screening assays for the identification of chemical probes. *Nature chemical biology*, 3(8):466–479, 2007.

Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9455–9464. IEEE, 2021.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

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

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

Nikita Moshkov, Tim Becker, Kevin Yang, Peter Horvath, Vlado Dancik, Bridget K Wagner, Paul A Clemons, Shantanu Singh, Anne E Carpenter, and Juan C Caicedo. Predicting compound activity from phenotypic profiles and chemical structures. *Nature communications*, 14(1):1967, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Jiahua Rao, Hanjing Lin, Leyu Chen, Jiancong Xie, Shuangjia Zheng, and Yuedong Yang. Multimodal contrastive learning with negative sampling calibration for phenotypic drug discovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30752–30762, 2025.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Maxime Sanchez, Nicolas Bourriez, Ihab Bendidi, Ethan Cohen, Ivan Svatko, Elaine Del Nery, Hamza Tajmouati, Guillaume Bollot, Laurence Calzone, and Auguste Genovesio. Large scale compound selection guided by cell painting reveals activity cliffs and functional relationships. *Communications Biology*, 2026.

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.

Andreas Scherer. *Batch effects and noise in microarray experiments: sources and solutions*, volume 868. Wiley Online Library, 2009.

Srijit Seal, Jordi Carreras-Puigvert, Maria-Anna Trapotsi, Hongbin Yang, Ola Spjuth, and Andreas Bender. Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. *Communications Biology*, 5(1):858, 2022.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608, 2020.

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.

Dagmar Stumpfe and Jurgen Bajorath. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, 55(7):2932–2942, 2012.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

D Lansing Taylor. *High content screening*, volume 356. Springer Science & Business Media, 2008.

Guangyan Tian, Philip J Harrison, Akshai P Sreenivasan, Jordi Carreras-Puigvert, and Ola Spjuth. Combining molecular and cell painting image data for mechanism of action prediction. *Artificial Intelligence in the Life Sciences*, 3:100060, 2023.

Tingzhong Tian, Shuya Li, Ziting Zhang, Lin Chen, Ziheng Zou, Dan Zhao, and Jianyang Zeng. Benchmarking compound activity prediction for real-world drug discovery applications. *Communications Chemistry*, 7(1):127, 2024.

Chenyu Wang, Sharut Gupta, Caroline Uhler, and Tommi Jaakkola. Removing biases from molecular representations via information maximization. In *International Conference on Learning Representations*, volume 2024, pp. 25810–25835, 2024.

Xiaoli Wang, Liyong Fu, Yudong Zhang, Yongli Wang, and Zechao Li. Mmatch: Semi-supervised discriminative representation learning for multi-view classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6425–6436, 2022a.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022b.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

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.

Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14451–14461. IEEE, 2022.

Shuangjia Zheng, Jiahua Rao, Jixian Zhang, Lianyu Zhou, Jiancong Xie, Ethan Cohen, Wei Lu, Chengtao Li, and Yuedong Yang. Cross-modal graph contrastive learning with cellular images. *Advanced Science*, 11(32):2404845, August 2024. ISSN 2198-3844, 2198-3844. doi: 10.1002/advs.202404845.

A NOTATION

Table 3 summarizes the main notation used throughout the paper.

Symbol	Description
i, τ, r	Indices of compounds, assay tasks, and Cell Painting replicates, respectively.
N, K, R_i	Number of compounds, assay tasks, and available replicates for compound i, respectively.
x_i^{mol}	Molecular structure of compound i.
$\mathcal{X}_i^{\mathrm{cp}}$	Collection of treated and matched-control Cell Painting images of compound i.
$y_{i\tau}$	Assay label of compound i for task τ, where $y_{i\tau} \in \{0, 1, \perp\}$.
$\mathbf{y}_i$	Sparse assay-label vector of compound i across all assay tasks.
Ω_L, Ω_U	Sets of observed and missing compound–task pairs, respectively.
z_{ir}^{cp}	Control-conditioned Cell Painting representation of replicate r of compound i.
z_i^{mol}	Molecular representation of compound i.
$z_{\mathrm{MM},i}^{(k)}$	Multimodal representation of augmented view k of compound i.
z_{ir}^{MM}	Stage-1 multimodal representation for replicate r of compound i.
T_τ	Task-specific teacher for assay task τ.
$p_{ir\tau}^T$	Teacher-predicted probability for compound i, replicate r, and task τ.
$\widetilde{y}_{ir\tau}$	Candidate pseudo-label generated by the task-specific teacher.
$c_{ir\tau}$	Teacher confidence associated with candidate pseudo-label $\widetilde{y}_{ir\tau}$.
$\mathcal{N}_m(i)$	Nearest neighbors of compound i in modality $m \in \{\mathrm{mol}, \mathrm{cp}\}$.
$s_{ir\tau}^{\mathrm{mol}}, s_{ir\tau}^{\mathrm{cp}}$	Molecular- and morphology-topology support scores for candidate pseudo-label $\widetilde{y}_{ir\tau}$.
$s_{i\tau}^{\mathrm{rep}}$	Replicate-consistency score for compound i and task τ.
$q_{ir\tau}$	CTPL reliability score for candidate pseudo-label $\widetilde{y}_{ir\tau}$.
$\mathcal{B}_{\mathrm{PL}}$	Set of pseudo-labels retained by CTPL.
S_ω	Unified multi-task student.
$\widehat{\mathbf{y}}_{ir}$	Student prediction vector across all assay tasks for replicate r of compound i.

Table 3: Summary of the main notation used in the proposed framework.

B EXPERIMENTAL SETUP AND IMPLEMENTATION

B.1 DATASET STATISTICS

We evaluate BioCoReT on three public Cell Painting benchmarks, cpg0012, cpg0016, and cpg0037. Although all three datasets combine morphological phenotypes with compound-level bioactivity annotations, they originate from different experimental studies and were constructed into prediction benchmarks through different annotation pipelines. Consequently, they differ substantially in compound scale, assay composition, supervision density, class distribution, and replicate availability. Table 4 summarizes the processed benchmarks used in our experiments, while Fig. 5 provides a more detailed comparison of their supervision and replicate distributions.

cpg0012. cpg0012 originates from the large-scale Cell Painting study of Bray et al. (2017), which profiled more than 30,000 small-molecule treatments in U2OS cells using high-content fluorescence microscopy. Hofmarcher et al. (2019) subsequently linked compounds from this imaging collection to public bioactivity annotations in ChEMBL, constructing a benchmark for predicting biological assay outcomes directly from cellular morphology. Binary activity labels were derived from available compound–assay measurements, resulting in 10,574 compounds associated with 209 assay tasks.

The resulting label matrix is highly incomplete because individual compounds were experimentally characterized in only a subset of these assays. In our processed benchmark, only 2.51% of all possible compound–task pairs are observed. As shown in Fig. 5(c), most compounds therefore have labels for only a small fraction of the available tasks. cpg0012 consequently provides a large-scale and highly multi-task setting, while simultaneously representing the most extreme supervision sparsity among the three datasets.

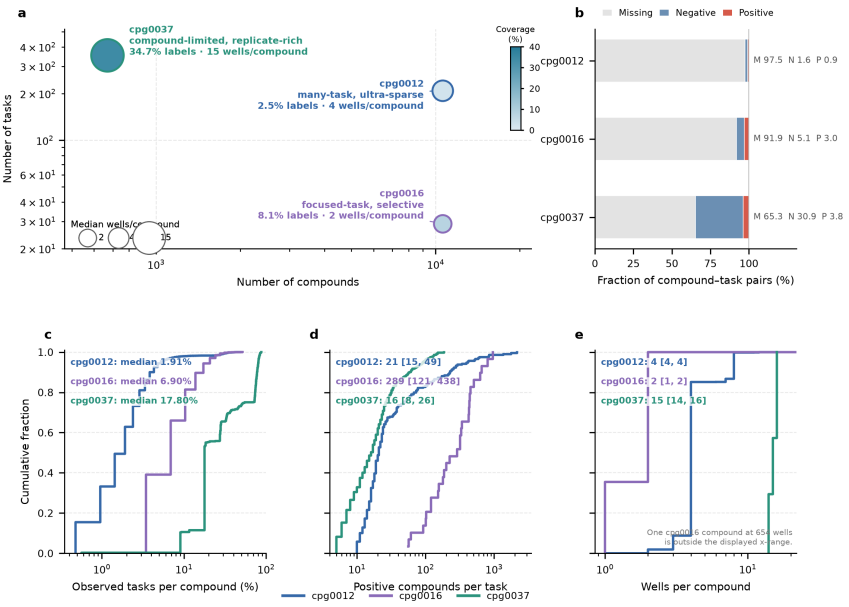

Figure 5: **Detailed statistics of the three Cell Painting benchmarks.** **(a)** Dataset scale in terms of the number of compounds and assay tasks; bubble size denotes the median number of Cell Painting wells per compound and color denotes label coverage. **(b)** Fractions of missing, negative, and positive compound–task pairs. **(c)** Cumulative distribution of the fraction of observed assay tasks per compound. **(d)** Cumulative distribution of the number of positive compounds per assay task. **(e)** Cumulative distribution of the number of Cell Painting wells per compound. Together, the three datasets span markedly different regimes of compound scale, task diversity, supervision sparsity, class distribution, and replicate availability.

Table 4: Statistics of the processed Cell Painting benchmarks used in our experiments. Label coverage denotes the fraction of observed compound–task labels among all possible compound–task pairs.

Dataset	# Compounds	# Tasks	Label Coverage	Supervision Regime
cpg0012	10,574	209	2.51%	Large-scale, ultra-sparse
cpg0016	10,588	29	8.10%	Few-task, selectively measured
cpg0037	670	354	34.68%	Compound-limited, task-rich

cpg0016. cpg0016 is constructed from the JUMP Cell Painting resource Chandrasekaran et al. (2023), a substantially larger collaborative profiling effort in which chemical and genetic perturbations were imaged in U2OS cells across multiple experimental sources. The full JUMP-CP compound collection contains more than 100,000 chemical perturbations, but only a subset has sufficiently matched public bioactivity information for downstream assay prediction.

Following . Fredin Haslum et al. (2024), compounds in JUMP-CP were cross-referenced against ChEMBL activity records. To reduce variation arising from different imaging sources, the public benchmark was constructed primarily from JUMP source 11, which provided the largest overlap with ChEMBL. ChEMBL activity annotations were converted into binary active and inactive labels, and assays without sufficient numbers of both classes were excluded. This procedure produced a public benchmark with 29 assay endpoints. After our preprocessing and multimodal data matching, the benchmark used in our experiments contains 10,588 compounds and has an overall label coverage of 8.10%.

Compared with cpg0012, cpg0016 therefore contains a similar number of compounds but substantially fewer downstream tasks. Its available labels are concentrated on this smaller collection of assay endpoints, yielding a moderately denser but still incomplete compound–task matrix. It provides a complementary evaluation setting in which task diversity is limited while bioactivity supervision remains selectively observed across compounds.

cpg0037. cpg0037 originates from the toxicity-oriented Cell Painting study of Ewald et al. (2026). In contrast to cpg0012 and cpg0016, which use the U2OS cell line, this dataset profiles primary human hepatocytes (PHH), providing a more directly toxicity-related cellular context. In the original study, 1,085 compounds spanning pharmaceuticals, pesticides, and industrial chemicals were profiled across eight concentrations using Cell Painting together with conventional cytotoxicity measurements. The resulting morphological phenotypes were further associated with curated ToxCast assay readouts covering a broad range of cytotoxicity, cell-based, and cell-free endpoints.

For our experiments, we reorganize the available compound–assay annotations into a unified sparse multi-task prediction benchmark. After preprocessing, the resulting dataset contains 670 compounds and 354 assay tasks, with 34.68% of all possible compound–task pairs observed. Although cpg0037 is substantially smaller in compound scale than cpg0012 and cpg0016, it provides a considerably larger number of assay tasks relative to its size and much denser supervision.

cpg0037 also differs markedly in its experimental replication structure. As shown in Fig. 5(e), compounds have a median of 15 Cell Painting wells, compared with 4 and 2 wells for cpg0012 and cpg0016, respectively. It therefore provides a compound-limited but task-rich and replicate-rich setting, allowing us to examine the framework under a substantially different balance between compound scale, task supervision, and repeated morphological observations.

Comparison of supervision regimes. The three benchmarks therefore differ not only in scale, but also in how their supervision was originally collected and assembled. cpg0012 combines a large morphological profiling collection with sparse ChEMBL-derived assay annotations, resulting in an extremely sparse multi-task benchmark. cpg0016 starts from the much larger JUMP-CP resource but retains a relatively small set of assay endpoints with sufficient matched ChEMBL annotations, resulting in a selectively measured setting. cpg0037 instead originates from a toxicity-focused experimental study and provides substantially denser ToxCast supervision together with richer experimental replication. These complementary settings allow us to evaluate BioCoReT across substantially different balances of compound scale, task diversity, label sparsity, and replicate availability.

B.2 Input Processing

Each compound is represented by its molecular structure and the available Cell Painting observations collected across experimental replicates. For each replicate, we retain the treated Cell Painting image together with its matched untreated control. Assay annotations are represented as a sparse compound–task label matrix.

Cell Painting observations. For compound i, we denote the available Cell Painting observations as

$$X_i^{\mathrm{cp}} = \left\{ \left(x_{ir}^{\mathrm{cp,tr}}, x_{ir}^{\mathrm{cp,ctrl}} \right) \right\}_{r=1}^{R_i},$$

where $x_{ir}^{\mathrm{cp,tr}}$ and $x_{ir}^{\mathrm{cp,ctrl}}$ denote the treated image and its matched untreated control for replicate r, respectively, and R_i is the number of available replicates.

Replicates are retained as separate observations rather than being averaged before representation learning. This preserves experimentally observed variation across runs and allows Stage 1 to construct different biologically valid views of the same compound. For downstream prediction, replicate-level predictions belonging to the same compound are aggregated at the compound level before evaluation.

Matched controls. Each treated Cell Painting observation is paired with its corresponding matched control from the same experimental context. BioCoReT uses this pairing in the control-conditioned representation encoding described in Sec. 3.2.1, where the matched control serves as a reference for characterizing compound-induced morphological changes. We retain the original treated–control correspondence throughout training.

Molecular structure. Each compound is additionally associated with a molecular input x_i^{mol} derived from its chemical structure. The same molecular input is shared across all Cell Painting replicates of the same compound. The molecular encoder used by BioCoReT is described in the main method, while the molecular representations used by individual baselines are detailed in Sec. B.4.

Assay annotations. For N compounds and K assay tasks $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$, each compound is associated with a sparse label vector

$$\mathbf{y}_i = [y_{i\tau_1}, \ldots, y_{i\tau_K}], \qquad y_{i\tau} \in \{0, 1, \perp\},$$

where 0 and 1 denote inactive and active assay responses, respectively, and $\perp$ denotes an unobserved compound–task pair. Accordingly, we define

$$\Omega_L = \{(i, \tau) : y_{i\tau} \neq \perp\}, \qquad \Omega_U = \{(i, \tau) : y_{i\tau} = \perp\}.$$

Only entries in Ω_L are used as observed assay supervision. Entries in Ω_U are treated as unobserved and are never interpreted as negative labels.

B.3 Data Splits

We follow the original compound-level evaluation protocol of each benchmark. Specifically, cpg0012 provides three predefined data splits, cpg0016 is evaluated using six-fold cross-validation, and cpg0037 is evaluated using five-fold cross-validation. For each split or fold, all Cell Painting observations, experimental replicates, matched controls, molecular inputs, and assay annotations associated with the same compound are assigned to the same partition. This prevents information from the same compound from appearing across the training, validation, and test sets.

The same benchmark-specific splits or folds are used across all comparable methods. Model selection and hyperparameter tuning are performed using only the corresponding training and validation data, while the held-out test partition is used only for final evaluation.

For the reduced-label experiments, the compound-level split remains unchanged. We only subsample the observed assay labels within the training partition while retaining all available multimodal compound observations. Thus, reducing the label fraction changes the amount of task-specific supervision without changing the compounds available for representation learning.

The hidden-label evaluation used to assess pseudo-label reliability follows a separate protocol. A subset of originally observed training labels is temporarily hidden from the teacher and CTPL pipeline and used only to evaluate the reliability of the selected pseudo-labels. These hidden labels are not used for teacher training, pseudo-label selection, or student supervision.

No held-out test labels are used for representation learning, teacher training, pseudo-label construction, model selection, or hyperparameter tuning.

B.4 Baselines

We compare BioCoReT against Cell Painting-only, molecular-only, and multimodal approaches. Whenever an original method is formulated for a different downstream task, we retain its representation learning or fusion strategy and adapt the final prediction module to the same multi-task assay prediction setting used by BioCoReT. All comparable methods use the same compound-level data splits.

B.4.1 Cell Painting-only baselines

CellProfiler. We use CellProfiler Stirling et al. (2021) morphology features as a classical Cell Painting representation. The resulting profiles are used as input to a multi-task prediction model trained only on the observed assay annotations.

DINO. We evaluate image representations obtained from DINO/DINOv2-style self-supervised visual encoders Oquab et al. (2023). The extracted image features are aggregated into Cell Painting representations and used with the same multi-task prediction protocol.

HTI-CNN. HTI-CNN Hofmarcher et al. (2019) is included as an end-to-end Cell Painting image baseline. We adapt its output layer to the corresponding set of assay tasks for each dataset.

B.4.2 MOLECULAR-ONLY BASELINES

ECFP4. We compute extended-connectivity fingerprints with radius two (ECFP4) Rogers & Hahn (2010) for each compound and train a multi-task predictor directly on the resulting molecular fingerprints.

ChemBERTa. We use ChemBERTa Chithrananda et al. (2020) to obtain learned molecular representations from chemical structure and train the same multi-task prediction head on top of these features.

B.4.3 MULTIMODAL BASELINES

CLOOME. CLOOME Sanchez-Fernandez et al. (2023) learns aligned chemical and Cell Painting representations through cross-modal contrastive learning. We use its multimodal representation under the same downstream assay prediction protocol.

CellCLIP. CellCLIP Lu et al. (2025) learns correspondences between molecular and cellular representations using a CLIP-style cross-modal objective. We adapt its learned representations to the same multi-task setting.

MICON. MICON Yu et al. (2026) is included as an additional multimodal representation learning baseline combining molecular and Cell Painting information. Its downstream prediction module is adapted to the assay tasks considered in each benchmark.

InfoCORE. InfoCORE Wang et al. (2024) is evaluated as a representation learning baseline designed to improve biological representations in the presence of technical variation. We use its learned representations with the common downstream prediction protocol.

MINER. MINER Rao et al. (2025) represents a direct multimodal fusion strategy for integrating chemical and Cell Painting information. We adapt its prediction head to our multi-task assay setting.

CP-Chem-MoA. CP-Chem-MoA Tian et al. (2023) combines molecular and Cell Painting representations for downstream prediction. Although originally developed for mechanism-of-action prediction, we retain its multimodal fusion strategy and replace the original classification head with the common multi-task assay prediction head.

For all adapted baselines, we preserve the core representation or fusion mechanism of the original method while changing only the task-specific output layer when necessary. This ensures that performance differences primarily reflect how the methods represent and integrate the two modalities rather than differences in downstream task formulation.

B.4.4 SEMI-SUPERVISED BASELINES

For the reduced-label experiments, we additionally compare BioCoReT with representative semi-supervised learning methods. These methods are evaluated using the same compound-level data splits and the same observed-label masks as BioCoReT. Because the supervision in our setting is sparse at the compound–task level, we adapt each method to multi-task assay prediction while preserving its core semi-supervised learning strategy.

Co-training. Co-training Blum & Mitchell (1998) is a classical multi-view semi-supervised learning method that iteratively uses predictions from one view to provide additional supervision for another. We treat molecular structure and Cell Painting morphology as the two complementary views and adapt the method to multi-task assay prediction under the same sparse compound–task supervision.

CoMatch. CoMatch Li et al. (2021) combines pseudo-labeling with representation consistency to exploit unlabeled samples. We adapt its semi-supervised objective to the assay prediction setting and apply it using the same observed and unobserved compound–task pairs used by the other methods.

Multimodal SimMatch. SimMatch Zheng et al. (2022) combines semantic pseudo-labeling with instance-level similarity consistency. For our experiments, we extend the method to the multimodal setting by using the available molecular and Cell Painting representations while retaining its original pseudo-labeling and similarity-consistency principles.

MMatch. MMatch Wang et al. (2022a) is a multimodal semi-supervised learning method that exploits complementary information across modalities to improve pseudo-label learning. We apply it to molecular and Cell Painting inputs and adapt the prediction head to the same multi-task assay setting.

CP-SemiSupCon. CP-SemiSupCon Bushiri Pwesombo et al. (2025) is a semi-supervised contrastive learning method developed specifically for Cell Painting-based bioactivity prediction. We evaluate it as a Cell Painting-only semi-supervised baseline using the same compound-level splits and reduced-label masks.

STiL. STiL Du et al. (2025) is a semi-supervised multimodal learning approach that transfers information between labeled and unlabeled samples using cross-modal supervision. We adapt its prediction module to the multi-task assay setting while preserving its original semi-supervised learning mechanism.

For all adapted baselines, we preserve the core representation-learning, fusion, or semi-supervised learning mechanism of the original method and modify only the downstream prediction component when necessary. All methods are evaluated using the same benchmark-specific data splits, observed-label masks, and compound-level evaluation protocol. This provides a controlled comparison in which differences are primarily attributable to the representation, multimodal integration, or semi-supervised learning strategy rather than differences in data partitioning or evaluation.

B.5 Sparse-Supervision Protocol

To evaluate performance under increasingly limited assay supervision, we construct controlled reduced-label settings by subsampling the originally observed assay annotations in the training partition while keeping the multimodal compound data unchanged.

Let Ω_L^{train} denote the set of observed compound–task pairs in the original training data. For a label fraction ρ, we sample a subset

$$\Omega_L^{(\rho)} \subseteq \Omega_L^{\text{train}},$$

and expose only the labels in $\Omega_L^{(\rho)}$ to the supervised components of the model. The remaining originally observed training entries are treated as unobserved during training.

In the reduced-label experiments, we consider

$$\rho \in \{5\%, 10\%, 50\%, 100\%\},$$

where 100% corresponds to using all originally observed training labels.

Reducing the label fraction changes only the amount of assay supervision. All compounds and their available molecular structures, Cell Painting observations, matched controls, and experimental replicates remain available during training. Thus, the protocol isolates the effect of task-specific label availability from the amount of multimodal compound data available to the model.

We conduct the controlled reduced-label experiments on cpg0037. Compared with the other benchmarks, cpg0037 contains substantially denser assay annotations, allowing the observed labels to be progressively subsampled while retaining sufficient supervision for meaningful evaluation.

Validation and test annotations are not subsampled. Within each experimental run, the same sampled training-label mask is used across all comparable supervised and semi-supervised methods. This ensures that differences between methods are not caused by differences in which assay annotations are available during training.

B.6 Evaluation Protocol

All methods are evaluated at the compound level. When a compound contains multiple Cell Painting replicates, the corresponding replicate-level predictions are first aggregated to obtain a single compound-level prediction for each assay task. All evaluation metrics are then computed using these compound-level predictions.

For each assay task $\tau \in \mathcal{T}$, we independently compute AUROC, AUPRC, and F1. The reported result is the macro average across tasks for which the corresponding metric is well defined:

$$\text{Metric}_{\text{macro}} = \frac{1}{|\mathcal{T}_{\text{valid}}|} \sum_{\tau \in \mathcal{T}_{\text{valid}}} \text{Metric}_\tau,$$

where $\mathcal{T}_{\text{valid}}$ denotes the set of valid assay tasks for that metric on the corresponding evaluation split.

AUROC. AUROC evaluates the ability to rank active compounds above inactive compounds within each assay task. Because it is threshold independent, we use it as one of the primary metrics for comparing predictive performance across tasks. Tasks containing only one class in the evaluation split are excluded from AUROC computation.

AUPRC. AUPRC summarizes the precision–recall trade-off for the active class. It is particularly informative for assay tasks with substantial class imbalance, where the number of active compounds can be much smaller than the number of inactive compounds.

F1. F1 evaluates thresholded binary predictions by balancing precision and recall for the active class. For each task, F1 is computed at the compound level and then macro-averaged across valid assay tasks using the same task-level averaging protocol as AUROC and AUPRC. For F1, binary predictions are obtained using a fixed threshold of 0.5 for all methods and assay tasks.

B.7 Implementation Details

Stage-1 representation learning. We use frozen pretrained visual and molecular encoders. For cpg0012, Cell Painting images are represented using DINOv2-Giant (ViT-g/14) features, with 17 tokens per channel (16 spatial tokens and one global token), each of dimension 1536. For cpg0016 and cpg0037, we use 768-dimensional DINO profiles. Molecular structures are represented using a frozen ChemBERTa encoder Chithrananda et al. (2020), with a maximum sequence length of 128.

The trainable multimodal encoder is a four-layer pre-norm Transformer with hidden dimension 256, eight attention heads, an FFN dimension of 1024, and dropout of 0.30. The contrastive projection dimension is 128. During training, we construct different views using replicate sampling, feature masking/dropout, and Gaussian perturbations with $\sigma = 0.02$. We use a contrastive temperature of $\tau = 0.07$ and set $\lambda_{\text{CP}} = \lambda_{\text{Mol}} = 0.05$.

Stage 1 is trained for 20 epochs using AdamW. The batch size is 24 for cpg0012 and 128 for cpg0016 and cpg0037.

Task-specific teachers. For each assay task, the specialist teacher uses an MLP head $256 \to 128 \to 1$ with LayerNorm, GELU, and dropout of 0.10. The multimodal backbone is initialized from the Stage-1 checkpoint and fine-tuned jointly with the task-specific head using positive-weighted BCE.

We use learning rates of 2×10^{-5} for the multimodal backbone and 3×10^{-4} for the teacher head, with weight decay 10^{-4}. Teachers are trained for 40 epochs with a batch size of 256. Candidate positive and negative pseudo-labels are generated using

$$\theta_+ = 0.8, \qquad \theta_- = 0.2,$$

respectively. We use four MC-dropout passes when estimating teacher predictions.

CTPL. For topology-based reliability estimation, we use $K = 20$ nearest neighbors. Morphological similarity is measured using cosine similarity, while molecular similarity is computed using

Tanimoto similarity between molecular fingerprints. Teacher confidence, molecular topology, morphology topology, and replicate consistency are equally weighted in the reliability score. Unless otherwise specified, we retain pseudo-labels using $\theta_q = 0.8$.

To limit imbalance in the number of selected pseudo-labels across tasks, the maximum number retained for task τ is

$$\min(100, \lceil 2N_\tau^+ \rceil),$$

where N_τ^+ denotes the number of observed positive training examples for that task.

Unified multi-task student. The unified student uses the same four-layer multimodal Transformer architecture as Stage 1, followed by a $256 \to K$ multi-task classification head. The model is initialized from the Stage-1 checkpoint and trained using the observed-label objective together with the reliability-weighted pseudo-label objective, with $\lambda_{\mathrm{PL}} = 0.10$.

We use learning rates of 2×10^{-5} for the backbone, 5×10^{-5} for the final Transformer block, and 2×10^{-4} for the classification head. The student is trained for 30 epochs using AdamW with weight decay 10^{-3} and cosine annealing.

B.8 Reproducibility Details

All methods are evaluated using the same benchmark-specific compound-level splits or folds described in Sec. B.3. Specifically, cpg0012 uses three compound-disjoint predefined splits, cpg0016 uses six compound-disjoint folds, and cpg0037 uses five compound-disjoint folds. All replicates and associated molecular inputs from the same compound remain within the same partition.

For the two-stage ablation, we evaluate each configuration on all three cpg0012 splits and all five cpg0037 folds, resulting in eight split/fold runs per configuration across the two datasets. The four configurations therefore comprise 32 runs in total. The same data partitions and, where applicable, the same reduced-label masks are used across comparable methods.

Training is performed using bfloat16 automatic mixed precision with gradient clipping at a maximum norm of 1.0. Experiments are run on NVIDIA RTX A6000 GPUs; our compute server contains six RTX A6000 GPUs and 251 GB of system memory.

C Biological and Experimental Background

This section provides additional biological and experimental background for Cell Painting-based compound profiling and motivates several properties of the data considered in this work. We first describe how Cell Painting observations are generated and organized across wells, plates, experimental runs, and controls. We then discuss two characteristic structure–phenotype relationships, activity cliffs and scaffold hopping, which illustrate why molecular structure and cellular morphology provide complementary but non-equivalent views of compound biology.

C.1 Cell Painting for Morphological Profiling

Cell Painting is a high-content imaging assay designed to characterize perturbation-induced changes in cellular morphology Bray et al. (2016); Caicedo et al. (2017). In a typical experiment, cells are cultured in multi-well plates and exposed to chemical or genetic perturbations. Following perturbation, cells are stained with a panel of fluorescent dyes that label multiple cellular compartments and are subsequently imaged using high-throughput fluorescence microscopy. The standard Cell Painting protocol uses six fluorescent dyes imaged across five channels to capture morphological information from several cellular components, including the nucleus, cytoplasm, mitochondria, endoplasmic reticulum, nucleoli, and cytoskeleton-related structures Bray et al. (2016).

The resulting microscopy images provide a high-dimensional description of cellular state after perturbation. Classical Cell Painting pipelines use image-analysis software such as CellProfiler to segment individual cells and quantify morphological properties, including intensity, texture, shape, size, and spatial organization. More recent approaches instead learn image representations directly using deep neural networks. Regardless of the representation method, the central goal is the same: to obtain a quantitative description of the cellular phenotype induced by a perturbation.

Importantly, a Cell Painting observation should not be interpreted as a measurement of compound identity alone. The observed phenotype reflects the interaction between the perturbation, the biological system, and the experimental environment. Consequently, observations generated for the same compound can exhibit variation even when the underlying chemical structure is unchanged.

C.2 Wells, Plates, Controls, and Experimental Replicates

Cell Painting experiments are organized hierarchically. Individual cell populations are cultured in *wells*, multiple wells are arranged on an experimental *plate*, and multiple plates may be processed within an experimental run or batch. A well is typically associated with a particular perturbation and experimental condition, while the same perturbation may be repeated across multiple wells, plates, or experimental runs.

Control wells are included to provide references for experimental variation. Depending on the experimental protocol, these may include vehicle-treated, untreated, or other negative and positive controls. Controls are particularly important in high-throughput imaging because measurements can vary across plates and experimental runs even in the absence of a meaningful biological perturbation. Comparing treated observations with controls acquired under matched experimental conditions therefore helps distinguish perturbation-induced changes from background experimental variation.

The same compound can also be profiled multiple times, resulting in multiple experimental replicates. Replicates are intended to capture the same underlying perturbation but are not expected to produce identical images or representations. Variability may arise from differences in cell state, staining, imaging, plate position, environmental conditions, or other experimental factors. Accordingly, replicate observations contain both shared biological information and replicate-specific variation.

In this work, we use the term *replicate* to refer to repeated Cell Painting observations associated with the same compound across experimental runs. We retain these observations separately rather than averaging them before representation learning. This distinction is important because replicate variation itself provides information about which aspects of a compound-induced phenotype remain stable across experimental conditions.

C.3 Batch Effects in High-Content Imaging

High-throughput biological imaging is particularly susceptible to systematic technical variation, commonly referred to as *batch effects* Caicedo et al. (2017); Arevalo et al. (2024). Batch effects can arise from many sources, including differences in cell culture conditions, reagent preparation, staining efficiency, microscope configuration, illumination, acquisition time, plate position, laboratory, and data-generation date. Such factors may alter image appearance or derived representations without corresponding to a change in the underlying biological perturbation.

This issue becomes especially relevant in large-scale Cell Painting datasets, which may contain observations collected across many plates, batches, instruments, or data-generating sites. For example, the JUMP Cell Painting project intentionally collected perturbations across multiple data-generating centers and experimental contexts, introducing substantial technical diversity into the dataset Chandrasekaran et al. (2023). Reference and control plates are consequently used to facilitate comparison across batches and experimental sources.

Batch effects pose a particular challenge for representation learning. A model may inadvertently encode plate- or batch-specific information instead of perturbation-induced biology, causing observations from the same compound to appear different simply because they were acquired in different experimental contexts. Conversely, observations from unrelated compounds measured within the same batch may become spuriously similar. Distinguishing stable compound-associated information from experimental variation is therefore a central consideration when learning representations from Cell Painting data.

Matched controls provide one source of information for addressing this problem. Because treated and control observations share much of their experimental context, their comparison can help characterize morphological changes associated with the perturbation rather than absolute image appearance. Experimental replicates provide a complementary source of information: biological signals

that remain consistent across different replicates are more likely to represent stable properties of the compound perturbation than signals that appear only in one experimental observation.

C.4 Molecular Structure and Cellular Phenotype

Molecular structure and Cell Painting morphology describe different aspects of a compound. Chemical structure specifies atoms, bonds, functional groups, scaffolds, and other structural properties of the molecule. Cell Painting, in contrast, provides an indirect observation of the cellular state resulting from the interaction between the compound and a biological system. The two modalities are therefore related through compound biology, but their relationship is neither deterministic nor one-to-one.

Several factors contribute to this non-equivalence. Small structural modifications can alter target affinity, selectivity, permeability, metabolism, or other properties and consequently produce substantially different biological responses. Conversely, chemically distinct compounds can perturb the same target, pathway, or cellular process and therefore induce similar downstream phenotypes. The relationship between structural similarity and phenotypic similarity should thus not be assumed to be monotonic.

Two concepts from medicinal chemistry provide intuitive examples of this relationship: activity cliffs and scaffold hopping.

C.4.1 Activity Cliffs

An *activity cliff* describes a pair or group of compounds that are structurally highly similar but exhibit substantially different biological activities Stumpfe & Bajorath (2012). Activity cliffs are a well-established phenomenon in structure–activity relationship analysis and demonstrate that small structural changes can produce disproportionately large changes in biological response.

Although activity cliffs are conventionally defined with respect to measured biological activity, the same principle is relevant when considering morphological phenotypes. Two structurally similar compounds may differ in potency, selectivity, intracellular exposure, or pathway engagement and can consequently induce different cellular states. High molecular similarity therefore does not guarantee high Cell Painting similarity.

From a multimodal learning perspective, this implies that forcing representations of molecular structure and morphology to agree too strongly can be restrictive. Structural proximity alone is not sufficient evidence that two compounds should occupy nearby positions in a phenotype-sensitive representation space.

C.4.2 Scaffold Hopping

Scaffold hopping describes the identification of biologically active compounds with substantially different molecular core structures while retaining related functional or pharmacological properties Brown (2013). It is widely used in medicinal chemistry to discover chemically distinct compounds that preserve a desired biological activity.

The concept provides the complementary example to activity cliffs. Two compounds may have low global structural similarity yet interact with the same molecular target, engage related pathways, or ultimately produce similar cellular responses. In Cell Painting, such compounds may therefore exhibit similar morphological phenotypes despite belonging to different chemical scaffolds.

Scaffold hopping illustrates why phenotypic similarity should not necessarily be interpreted as structural similarity. It also highlights the value of Cell Painting as a complementary observation: morphological profiling can reveal biological relationships between compounds that are difficult to infer from chemical structure alone.

C.5 Implications for Multimodal Compound Learning

The experimental and biological properties above lead to two important characteristics of the multimodal setting considered in this work.

First, the correspondence between a molecular structure and Cell Painting is not a simple one-to-one pairing. A single molecular structure may be associated with multiple Cell Painting observations across experimental replicates, and these observations may contain both stable biological responses and experiment-specific variation. Matched controls and replicate structure therefore provide useful information for distinguishing compound-induced effects from experimental variation.

Second, similarity relationships need not agree across modalities. Activity cliffs demonstrate that structurally similar compounds can exhibit different biological responses, whereas scaffold hopping demonstrates that structurally dissimilar compounds can retain related biological functions and phenotypes. Molecular and morphological observations should therefore be viewed as complementary but non-equivalent descriptions of compound biology.

These properties motivate the design of Stage 1 in BioCoReT. Rather than assuming that each molecular representation should directly match a single morphological representation, BioCoReT constructs multimodal views from different experimental replicates and learns consistency between fused views of the same compound. Matched controls are used to condition morphological encoding, while the multimodal objective allows molecular and morphological information to contribute jointly without requiring the two modalities to become identical. The mathematical implications of this design are further discussed in Appendix D.1.

D Additional Mathematical Analysis of BioCoReT

D.1 Additional Analysis of What to Learn

We provide additional analysis of the Stage-1 objective from three complementary perspectives: matched-control conditioning, consistency across experimental replicates, and information preservation in multimodal contrastive learning. These analyses characterize properties of the proposed objective beyond the motivation presented in the main text.

Matched-control conditioning reduces shared experimental nuisance. Consider a simplified additive decomposition of a treated Cell Painting representation,

$$h_{ir}^{\mathrm{cp,tr}} = s_i + b_{ir} + \epsilon_{ir}^{\mathrm{tr}},$$

where s_i denotes the compound-associated biological component, b_{ir} denotes nuisance variation shared within the matched experimental context, and $\epsilon_{ir}^{\mathrm{tr}}$ denotes observation-specific variation. The corresponding matched control can be written as

$$h_{ir}^{\mathrm{cp,ctrl}} = b_{ir} + \epsilon_{ir}^{\mathrm{ctrl}}.$$

The treatment–control residual therefore becomes

$$h_{ir}^{\mathrm{cp,tr}} - h_{ir}^{\mathrm{cp,ctrl}} = s_i + \epsilon_{ir}^{\mathrm{tr}} - \epsilon_{ir}^{\mathrm{ctrl}},$$

such that the shared nuisance term b_{ir} cancels under this simplified model.

BioCoReT does not directly replace the treated representation with this residual. Instead, its gated residual formulation adaptively incorporates the control-conditioned difference into the treated representation. The analysis above therefore provides an interpretation of matched-control conditioning: when treated and control observations share experimental nuisance, their difference reduces sensitivity to this shared variation while retaining the compound-associated response.

Cross-replicate consistency suppresses replicate-specific variation. We next consider two experimental replicates of the same compound. Let the multimodal representation associated with replicate r be locally approximated as

$$z_{ir}^{\mathrm{MM}} = u_i + A\epsilon_{ir},$$

where u_i represents compound-associated information that is stable across replicates, while ϵ_{ir} represents replicate-specific variation. For two replicates r and r' of the same compound,

$$z_{ir}^{\mathrm{MM}} - z_{ir'}^{\mathrm{MM}} = A\left(\epsilon_{ir} - \epsilon_{ir'}\right).$$

The replicate-invariant component u_i disappears from the difference, whereas variation specific to individual experimental observations remains.

Assuming that replicate-specific variations are conditionally zero-mean and independent with covariance Σ_ϵ, we further obtain

$$\mathbb{E}\left[\left\|z_{ir}^{\mathrm{MM}} - z_{ir'}^{\mathrm{MM}}\right\|_2^2 \,\middle|\, i\right] = 2\,\mathrm{tr}\left(A\Sigma_\epsilon A^\top\right).$$

Thus, reducing disagreement between independently sampled replicate views directly penalizes sensitivity to replicate-specific variation. At the same time, the contrastive objective separates views from different compounds, preventing the trivial solution in which all representations collapse to the same point.

Multimodal contrastive learning preserves complementary information. The fused-view contrastive objective can also be interpreted from the information-theoretic perspective developed in CoMM Dufumier et al. (2025). For two modalities, task-relevant multimodal information can be conceptually decomposed into redundant, modality-specific, and synergistic components:

$$I\left(X_{\mathrm{CP}}, X_{\mathrm{Mol}}; Y\right) = R + U_{\mathrm{CP}} + U_{\mathrm{Mol}} + S,$$

where R denotes information redundantly available from both modalities, U_{CP} and U_{Mol} denote information uniquely available from Cell Painting and molecular structure, respectively, and S denotes information available only through their joint observation.

Under the assumptions considered in CoMM, contrastive learning between appropriately augmented *multimodal* representations can preserve not only redundant information, but also modality-specific and synergistic task-relevant information Dufumier et al. (2025). This provides an additional interpretation of our objective

$$\mathcal{L}_{\mathrm{MM}} = -\widehat{I}_{\mathrm{NCE}}\left(z_{\mathrm{MM}}^{(1)}, z_{\mathrm{MM}}^{(2)}\right):$$

because contrast is performed after the two modalities have been jointly encoded, the learned representation can retain complementary information from both modalities rather than being restricted to their common information.

The auxiliary objectives

$$\mathcal{L}_{\mathrm{CP}\to\mathrm{MM}} \quad \text{and} \quad \mathcal{L}_{\mathrm{Mol}\to\mathrm{MM}}$$

provide a complementary role. They encourage information identifiable from each individual modality to remain accessible from the multimodal representation. Accordingly, the complete Stage-1 objective

$$\mathcal{L}_{\mathrm{rep}} = \mathcal{L}_{\mathrm{MM}} + \lambda_{\mathrm{CP}}\mathcal{L}_{\mathrm{CP}\to\mathrm{MM}} + \lambda_{\mathrm{Mol}}\mathcal{L}_{\mathrm{Mol}\to\mathrm{MM}}$$

combines consistency across multimodal observations with explicit preservation of information accessible from Cell Painting and molecular structure.

Summary. The three components therefore provide complementary effects: matched-control conditioning reduces nuisance shared within the experimental context, cross-replicate consistency reduces sensitivity to replicate-specific variation, and fused multimodal contrastive learning supports the preservation of redundant, modality-specific, and synergistic information. Together, these properties provide an additional interpretation of Stage 1 beyond the design motivation presented in the main text.

D.2 Additional Analysis of What to Trust

We provide additional analysis of Stage 2 from three perspectives: aggregation of complementary reliability evidence, reliability-based pseudo-label selection, and the influence of selected pseudo-labels on student optimization. These analyses characterize properties of the proposed transfer mechanism beyond the motivation described in the main text.

Reliability aggregation combines complementary evidence. For a candidate pseudo-label $\tilde{y}_{ir\tau}$, CTPL computes

$$q_{ir\tau} = \frac{\alpha c_{ir\tau} + \beta s^{\mathrm{mol}}_{ir\tau} + \gamma s^{\mathrm{cp}}_{ir\tau} + \delta m^{\mathrm{rep}}_i s^{\mathrm{rep}}_{i\tau}}{\alpha + \beta + \gamma + \delta m^{\mathrm{rep}}_i},$$

where the four terms correspond to teacher confidence, molecular topology, morphology topology, and replicate consistency, respectively. Assuming non-negative weights and component scores in $[0, 1]$, the resulting reliability score also satisfies

$$0 \le q_{ir\tau} \le 1.$$

Importantly, the indicator m^{rep}_i removes the replicate term from both the numerator and denominator when multiple replicates are not available. Thus, missing replicate information does not automatically decrease the reliability score; instead, the score is normalized over the evidence that is available for the corresponding compound.

The weighted-average form also means that no individual signal is required to determine reliability by itself. A candidate receives a high score only when the available evidence is collectively supportive, allowing disagreement between different sources of evidence to reduce its reliability.

Reliability thresholding induces a precision–coverage trade-off. Let

$$C_{ir\tau} = \mathbb{I}\left[\tilde{y}_{ir\tau} = y_{i\tau}\right]$$

indicate whether a candidate pseudo-label is correct, and define

$$\eta(q) = \Pr(C_{ir\tau} = 1 \mid q_{ir\tau} = q).$$

Suppose that the reliability score is rank-consistent with pseudo-label correctness, such that $\eta(q)$ is non-decreasing in q. The expected precision of pseudo-labels retained at threshold θ is then

$$\Pi(\theta) = \Pr\left(C_{ir\tau} = 1 \mid q_{ir\tau} \ge \theta\right) = \mathbb{E}\left[\eta(q_{ir\tau}) \mid q_{ir\tau} \ge \theta\right].$$

Under the monotonicity assumption, increasing θ preferentially removes lower-reliability candidates, so $\Pi(\theta)$ is non-decreasing with the selection threshold.

Conversely, the retained pseudo-label coverage

$$\Gamma(\theta) = \Pr(q_{ir\tau} \ge \theta)$$

is non-increasing in θ. CTPL therefore induces an explicit precision–coverage trade-off: a stricter reliability threshold retains fewer candidate predictions but favors those with stronger supporting evidence. The hidden-label experiments in the main text empirically evaluate this ranking behavior.

Reliability weighting controls pseudo-label influence. For a selected pseudo-label, the student uses the weighted binary cross-entropy term

$$\ell^{\mathrm{PL}}_{ir\tau} = \lambda_{\mathrm{PL}} q_{ir\tau}\, \mathrm{BCE}\left(\hat{y}_{ir\tau}, \tilde{y}_{ir\tau}\right).$$

Let $a_{ir\tau}$ denote the corresponding student logit, such that $\hat{y}_{ir\tau} = \sigma(a_{ir\tau})$. The gradient of the pseudo-label loss with respect to the logit is

$$\frac{\partial \ell^{\mathrm{PL}}_{ir\tau}}{\partial a_{ir\tau}} = \lambda_{\mathrm{PL}} q_{ir\tau}\left(\hat{y}_{ir\tau} - \tilde{y}_{ir\tau}\right).$$

Thus, the reliability score affects Stage 2 in two distinct ways. First, the threshold $q_{ir\tau} \ge \theta_q$ determines whether a candidate prediction is transferred at all. Second, among selected pseudo-labels, $q_{ir\tau}$ directly scales their contribution to the student update. Consequently, pseudo-label transfer is controlled through both *selection* and *optimization weight*, rather than treating all teacher predictions as equally informative.

Model	Modality		cpg0012 (209 tasks)		
	CP	Mol	AUROC ↑	AUPRC ↑	F1 ↑
CellProfiler + Head Stirling et al. (2021)	✓		0.705 ± 0.022	0.616 ± 0.034	0.512 ± 0.017
DINO + Head Oquab et al. (2023)	✓		0.713 ± 0.020	0.630 ± 0.041	0.505 ± 0.007
HTI-CNN Hofmarcher et al. (2019)	✓		0.718 ± 0.010	0.646 ± 0.012	0.489 ± 0.026
ECFP4 + Head Rogers & Hahn (2010)		✓	0.662 ± 0.051	0.538 ± 0.051	0.299 ± 0.007
ChemBERTa + Head Chithrananda et al. (2020)		✓	0.626 ± 0.019	0.511 ± 0.014	0.343 ± 0.035
CLOOME + Head Sanchez-Fernandez et al. (2023)	✓	✓	0.696 ± 0.009	0.624 ± 0.021	0.436 ± 0.067
CellCLIP + Head Lu et al. (2025)	✓	✓	0.737 ± 0.016	0.685 ± 0.014	0.525 ± 0.024
MICON + Head Yu et al. (2026)	✓	✓	0.704 ± 0.010	0.630 ± 0.032	0.475 ± 0.013
InfoCORE + Head Wang et al. (2024)	✓	✓	0.720 ± 0.009	0.632 ± 0.025	0.470 ± 0.035
MINER Rao et al. (2025)	✓	✓	0.668 ± 0.031	0.553 ± 0.057	0.372 ± 0.048
CP-Chem-MoA Tian et al. (2023)	✓	✓	0.745 ± 0.014	0.684 ± 0.030	0.519 ± 0.023
Our BioCoReT	✓	✓	$\mathbf{0.776 \pm 0.010}$	$\mathbf{0.705 \pm 0.019}$	$\mathbf{0.570 \pm 0.018}$

Table 5: Complete results on the cpg0012 benchmark. We report mean task-level AUROC, AUPRC, and F1 together with the sample standard deviation across the three predefined compound-level splits. "+ Head" denotes a common downstream multi-task prediction head applied to the corresponding representation. **Bold** indicates the best result in each column.

Specialist-to-generalist transfer. Each specialist teacher T_τ is trained only using the observed labels of assay τ and therefore learns a task-specific prediction function. Let $\mathcal{B}^{\tau}_{\mathrm{PL}}$ denote the reliable pseudo-labels selected from teacher T_τ. The unified student objective can be written task-wise as

$$\mathcal{L}_{\mathrm{student}} = \sum_{\tau \in \mathcal{T}} \left[\sum_{(i,\tau) \in \Omega_L} \sum_{r=1}^{R_i} \mathrm{BCE}\left(\hat{y}_{ir\tau}, y_{i\tau}\right) + \lambda_{\mathrm{PL}} \sum_{(i,r,\tau) \in \mathcal{B}^{\tau}_{\mathrm{PL}}} q_{ir\tau}\, \mathrm{BCE}\left(\hat{y}_{ir\tau}, \tilde{y}_{ir\tau}\right) \right].$$

This formulation preserves task specificity during pseudo-label generation, while allowing the selected supervision from all assays to jointly optimize a shared multi-task student. Stage 2 can therefore be interpreted as transferring selectively validated knowledge from multiple task specialists into a single generalist model.

Summary. CTPL separates reliable knowledge transfer into three operations: complementary evidence is first aggregated into a bounded reliability score, thresholding controls which task-specific predictions are transferred, and reliability weighting controls their influence during student optimization. This provides an additional mathematical interpretation of the what to trust stage beyond its empirical motivation.

E Detailed Experimental Results

E.1 Detailed Benchmark Results

We provide detailed benchmark results for cpg0012, cpg0016, and cpg0037. In addition to the mean performance reported in the main text, we report the corresponding sample standard deviation across the predefined compound-level splits or folds. Specifically, cpg0012 is evaluated over three predefined splits, cpg0016 over six folds, and cpg0037 over five folds. For each method, we report the mean $\pm$ standard deviation of the split/fold-level macro AUROC, AUPRC, and F1. All methods are evaluated using the same partitions within each benchmark.

E.2 Detailed Results under Reduced Label Supervision

Table 8 reports the detailed reduced-label results on cpg0037 using 5%, 10%, and 50% of the observed training labels. We report mean $\pm$ sample standard deviation across the five predefined compound-level folds. All methods use the same data splits and observed-label masks.

E.3 Additional Two-Stage Ablation

We further examine the individual and combined contributions of Stage 1 and Stage 2 on cpg0012 and cpg0037 under both 10% and full observed supervision. The cpg0012 results correspond to the ablation reported in the main text, while cpg0037 provides an additional evaluation under a distinct supervision regime, as shown in Table 9.

Model	Modality		cpg0016 (29 tasks)		
	CP	Mol	AUROC ↑	AUPRC ↑	F1 ↑
CellProfiler + Head Stirling et al. (2021)	✓		0.663 ± 0.023	0.662 ± 0.025	0.600 ± 0.022
DINO + Head Oquab et al. (2023)	✓		0.608 ± 0.020	0.616 ± 0.023	0.570 ± 0.021
HTI-CNN Hofmarcher et al. (2019)	✓		0.668 ± 0.029	0.662 ± 0.028	0.574 ± 0.041
ECFP4 + Head Rogers & Hahn (2010)		✓	0.707 ± 0.014	0.686 ± 0.020	0.603 ± 0.017
ChemBERTa + Head Chithrananda et al. (2020)		✓	0.671 ± 0.008	0.652 ± 0.024	0.580 ± 0.020
CLOOME + Head Sanchez-Fernandez et al. (2023)	✓	✓	0.591 ± 0.018	0.609 ± 0.023	0.546 ± 0.028
CellCLIP + Head Lu et al. (2025)	✓	✓	0.650 ± 0.017	0.649 ± 0.025	0.585 ± 0.017
MICON + Head Yu et al. (2026)	✓	✓	0.615 ± 0.018	0.616 ± 0.019	0.540 ± 0.009
InfoCORE + Head Wang et al. (2024)	✓	✓	0.628 ± 0.027	0.626 ± 0.028	0.572 ± 0.016
MINER Rao et al. (2025)	✓	✓	0.692 ± 0.018	0.671 ± 0.033	0.598 ± 0.027
CP-Chem-MoA Tian et al. (2023)	✓	✓	0.742 ± 0.020	0.729 ± 0.020	0.610 ± 0.018
Our BioCoReT	✓	✓	$\mathbf{0.752 \pm 0.017}$	$\mathbf{0.731 \pm 0.026}$	$\mathbf{0.655 \pm 0.024}$

Table 6: Complete results on the cpg0016 benchmark. We report mean task-level AUROC, AUPRC, and F1 together with the sample standard deviation across the six predefined compound-level folds. "+ Head" denotes a common downstream multi-task prediction head applied to the corresponding representation. **Bold** indicates the best result in each column.

Model	Modality		cpg0037 (354 tasks)		
	CP	Mol	AUROC ↑	AUPRC ↑	F1 ↑
CellProfiler + Head Stirling et al. (2021)	✓		0.600 ± 0.034	0.365 ± 0.039	0.083 ± 0.016
DINO + Head Oquab et al. (2023)	✓		0.604 ± 0.024	0.369 ± 0.028	0.080 ± 0.013
HTI-CNN Hofmarcher et al. (2019)	✓		0.618 ± 0.024	0.367 ± 0.016	0.059 ± 0.005
ECFP4 + Head Rogers & Hahn (2010)		✓	0.600 ± 0.049	0.351 ± 0.031	0.117 ± 0.033
ChemBERTa + Head Chithrananda et al. (2020)		✓	0.646 ± 0.031	0.360 ± 0.037	0.102 ± 0.014
CLOOME + Head Sanchez-Fernandez et al. (2023)	✓	✓	0.593 ± 0.033	0.361 ± 0.033	0.072 ± 0.007
CellCLIP + Head Lu et al. (2025)	✓	✓	0.637 ± 0.030	0.373 ± 0.033	0.073 ± 0.010
MICON + Head Yu et al. (2026)	✓	✓	0.577 ± 0.022	0.344 ± 0.027	0.063 ± 0.009
InfoCORE + Head Wang et al. (2024)	✓	✓	0.632 ± 0.022	0.372 ± 0.040	0.058 ± 0.006
MINER Rao et al. (2025)	✓	✓	0.629 ± 0.024	0.351 ± 0.030	0.049 ± 0.008
CP-Chem-MoA Tian et al. (2023)	✓	✓	0.618 ± 0.049	0.373 ± 0.049	0.093 ± 0.023
Our BioCoReT	✓	✓	$\mathbf{0.664 \pm 0.028}$	$\mathbf{0.398 \pm 0.035}$	$\mathbf{0.159 \pm 0.016}$

Table 7: Complete results on the cpg0037 benchmark. We report mean task-level AUROC, AUPRC, and F1 together with the sample standard deviation across the five predefined compound-level folds. "+ Head" denotes a common downstream multi-task prediction head applied to the corresponding representation. **Bold** indicates the best result in each column.

Across both datasets, Stage 1 and Stage 2 provide complementary improvements. Stage 1 consistently yields larger gains in ranking-based metrics, particularly AUROC and AUPRC, whereas Stage 2 contributes more strongly to F1. For example, under 10% supervision on cpg0012, Stage 1 improves AUROC and AUPRC over the base model by 4.36 and 4.20 percentage points, respectively, while Stage 2 produces a larger F1 improvement of 5.83 percentage points. A similar pattern is observed on cpg0037, where Stage 1 improves AUROC by 5.28 points under 10% supervision, whereas Stage 2 improves F1 by 4.92 points.

The two stages remain complementary when full supervision is available. On cpg0037, Stage 1 improves AUROC from 0.6125 to 0.6591, while Stage 2 raises F1 from 0.0910 to 0.1579. Combining both stages achieves 0.6638 AUROC, 0.3983 AUPRC, and 0.1589 F1, giving the strongest overall performance. The same trend holds on cpg0012, where the full model reaches 0.7760 AUROC, 0.7050 AUPRC, and 0.5700 F1.

These results suggest that the two stages improve different aspects of the prediction problem: Stage 1 primarily strengthens the underlying compound representation and ranking quality, while Stage 2 provides complementary task-specific supervision that is particularly beneficial for thresholded classification performance. Their combination therefore yields more balanced gains across AUROC, AUPRC, and F1 than either stage alone.

E.4 CTPL Threshold Analysis

Table 10 examines the sensitivity of CTPL to the reliability threshold θ_q. Across all three benchmarks, decreasing θ_q consistently increases pseudo-label coverage while reducing hidden-label accuracy, revealing a clear trade-off between the quantity and reliability of transferred supervision.

At the default threshold $\theta_q = 0.8$, CTPL retains 34.9%, 32.0%, and 12.7% of candidate pseudo-labels on cpg0016, cpg0012, and cpg0037, respectively, while maintaining hidden-label accuracies

Model	Modality		AUROC / AUPRC ↑		
	CP	Mol	5%	10%	50%
Supervised baselines					
ChemBERTa + Head Chithrananda et al. (2020)		✓	0.590 ± 0.051 / 0.325 ± 0.047	0.613 ± 0.035 / 0.330 ± 0.042	0.627 ± 0.034 / 0.345 ± 0.039
DINO + Head Oquab et al. (2023)	✓		0.510 ± 0.034 / 0.294 ± 0.028	0.538 ± 0.052 / 0.306 ± 0.034	0.583 ± 0.032 / 0.343 ± 0.019
CellCLIP + Head Lu et al. (2025)	✓	✓	0.566 ± 0.047 / 0.299 ± 0.027	0.586 ± 0.022 / 0.316 ± 0.036	0.619 ± 0.023 / 0.333 ± 0.027
Semi-supervised baselines					
Co-training Blum & Mitchell (1998)	✓	✓	0.528 ± 0.034 / 0.301 ± 0.020	0.577 ± 0.024 / 0.326 ± 0.025	0.580 ± 0.045 / 0.329 ± 0.053
CoMatch Li et al. (2021)	✓		0.585 ± 0.052 / 0.332 ± 0.042	0.600 ± 0.024 / 0.341 ± 0.022	0.615 ± 0.041 / 0.351 ± 0.046
Multimodal SimMatch Zheng et al. (2022)	✓	✓	0.572 ± 0.028 / 0.314 ± 0.019	0.572 ± 0.025 / 0.319 ± 0.017	0.608 ± 0.046 / 0.331 ± 0.042
MMatch Wang et al. (2022a)	✓	✓	0.578 ± 0.007 / 0.323 ± 0.017	0.581 ± 0.017 / 0.326 ± 0.019	0.594 ± 0.012 / 0.331 ± 0.016
CP-SemiSupCon Bushiri Pwesombo et al. (2025)	✓		0.571 ± 0.042 / 0.318 ± 0.027	0.575 ± 0.037 / 0.320 ± 0.017	0.603 ± 0.042 / 0.369 ± 0.050
STiL Du et al. (2025)	✓	✓	0.572 ± 0.017 / 0.326 ± 0.016	0.591 ± 0.023 / 0.331 ± 0.026	0.611 ± 0.041 / 0.338 ± 0.047
Our BioCoReT	✓	✓	$\mathbf{0.605 \pm 0.017}$ / $\mathbf{0.351 \pm 0.026}$	$\mathbf{0.629 \pm 0.023}$ / $\mathbf{0.358 \pm 0.025}$	$\mathbf{0.652 \pm 0.026}$ / $\mathbf{0.388 \pm 0.041}$

Table 8: Detailed performance on cpg0037 under reduced assay supervision. Each entry reports mean ± sample standard deviation for AUROC / AUPRC across the five predefined compound-level folds. All methods use the same data splits and observed-label masks. Bold indicates the best performance for each metric at each label fraction.

Setting	cpg0012 (209 tasks)						cpg0037 (354 tasks)					
	10% observed labels			100% observed labels			10% observed labels			100% observed labels		
	AUROC ↑	AUPRC ↑	F1 ↑	AUROC ↑	AUPRC ↑	F1 ↑	AUROC ↑	AUPRC ↑	F1 ↑	AUROC ↑	AUPRC ↑	F1 ↑
Base	0.67	0.58	0.46	0.74	0.68	0.52	0.58	0.31	0.10	0.61	0.37	0.09
+ Stage 1	0.72	0.62	0.50	0.77	0.69	0.56	0.63	0.35	0.12	0.66	0.39	0.11
+ Stage 2	0.70	0.61	0.52	0.77	0.69	0.57	0.60	0.32	0.15	0.65	0.39	0.16
Full	**0.72**	**0.63**	**0.52**	**0.78**	**0.71**	**0.57**	**0.63**	**0.36**	**0.15**	**0.66**	**0.40**	**0.16**

Table 9: Ablation of Stage 1 and Stage 2 on cpg0012 and cpg0037 under 10% and 100% observed assay supervision. Results are rounded to two decimal places. Boldface denotes the best performance within each dataset and supervision setting based on the unrounded mean values.

of 96.2%, 97.5%, and 98.1%. In contrast, retaining all candidate predictions ($\theta_q = 0$) reduces accuracy to 78.9%, 80.8%, and 79.0%, respectively.

These results indicate that the CTPL score provides a meaningful ranking of pseudo-label reliability rather than merely acting as a binary selection criterion. They also support the use of $\theta_q = 0.8$ as a conservative operating point that substantially filters candidate predictions while maintaining high selection accuracy.

E.5 Biological Success and Failure Cases

We further examine representative compound–assay pairs to illustrate how different sources of reliability evidence interact in individual predictions. Rather than focusing only on aggregate pseudo-label accuracy, these examples show when multimodal and replicate evidence supports a teacher prediction and when disagreement between the evidence sources prevents an unreliable prediction from being transferred.

Representative success case. We first consider compound `FUWUEFKEXZQKKA-UHFFFAOYSA-N` for task 190 (assay 845231), corresponding to a RORγ antagonist/inhibition assay. The task-specific teacher correctly predicts the compound response. Importantly, this prediction is simultaneously supported by the local molecular neighborhood and the Cell Painting neighborhood, while the predictions across available experimental replicates are also highly consistent.

The three independent sources of supporting evidence therefore agree with the teacher prediction, resulting in a high CTPL reliability score and selection of the corresponding pseudo-label. This example illustrates the intended behavior of CTPL when structural, morphological, and replicate-level evidence provide mutually consistent support for a task-specific prediction.

Failure case: conflicting molecular and morphological evidence. We next consider `JUMP_1003996` for task 18 (assay 752590). The specialist teacher produces an incorrect negative prediction with very high confidence. However, the two topology-based signals do not provide consistent support: the Cell Painting neighborhood strongly supports the teacher prediction, whereas the molecular neighborhood provides substantially weaker support.

θ_q	cpg0016 Coverage / Hidden-label Acc.	cpg0012 Coverage / Hidden-label Acc.	cpg0037 Coverage / Hidden-label Acc.
0.99	5.0% / 98.7%	7.4% / 99.4%	1.6% / 99.0%
0.95	13.1% / 98.1%	15.2% / 98.3%	4.3% / 98.8%
0.90	22.6% / 97.4%	22.8% / 98.2%	7.2% / 98.5%
0.85	29.2% / 96.7%	28.0% / 97.8%	9.9% / 98.3%
0.80	34.9% / 96.2%	32.0% / 97.5%	12.7% / 98.1%
0.75	39.7% / 95.3%	34.9% / 97.3%	15.7% / 98.1%
0.70	44.0% / 94.3%	37.9% / 96.8%	18.9% / 98.0%
0.65	47.9% / 93.4%	41.1% / 96.5%	22.3% / 97.9%
0.60	52.0% / 92.2%	44.3% / 95.8%	26.2% / 97.8%
0.55	55.5% / 91.3%	47.7% / 95.3%	30.3% / 97.5%
0.50	58.8% / 90.4%	51.5% / 94.9%	34.8% / 97.3%

Table 10: Effect of the CTPL reliability threshold θ_q on pseudo-label selection. Each entry reports pseudo-label coverage and hidden-label accuracy. Lowering the threshold increases the number of selected pseudo-labels while gradually reducing their reliability. The default threshold $\theta_q = 0.8$ used in the main experiments is highlighted in gray.

Consequently, high teacher confidence alone is insufficient to make the candidate reliable. The disagreement between molecular and morphological topology lowers the CTPL score below the selection criterion, and the erroneous pseudo-label is therefore rejected. This case demonstrates how multimodal evidence can prevent an overconfident teacher prediction from being transferred directly to the student.

Failure case: replicate inconsistency. A complementary failure mode is observed for `JUMP_878566` on task 11 (assay 737287). Here, the teacher produces an incorrect positive prediction with high confidence, and both molecular and morphological neighborhoods provide moderate support. In contrast, the prediction is substantially less consistent across Cell Painting replicates.

The low replicate-consistency signal reduces the overall CTPL reliability score, causing the candidate pseudo-label to be rejected. This example shows that agreement in molecular and morphology topology is not sufficient when the corresponding phenotype-level prediction is unstable across independent experimental observations.

Interpretation. Together, these examples illustrate that the four CTPL signals play complementary roles. A prediction supported consistently by teacher confidence, molecular topology, morphology topology, and experimental replicates can be retained with high reliability. Conversely, disagreement between modalities or instability across replicates can counteract an otherwise confident teacher prediction. The reliability mechanism therefore acts as an evidence aggregation step rather than a confidence-only pseudo-labeling rule.

F Applications and Limitations

Application scope. BioCoReT is designed for multimodal compound learning with sparse assay supervision. It is particularly suitable when molecular structures and Cell Painting observations are available for the same compounds. The setting may contain many unobserved compound–task pairs. Matched controls and experimental replicates can provide additional biological evidence.

The framework is applicable to multi-task bioactivity prediction and phenotypic drug discovery. It can also support assay prioritization by predicting unobserved compound–task responses. More broadly, the same formulation may be useful for other perturbation profiling settings with complementary modalities and sparse task-specific annotations. Such extensions, however, are not evaluated in this work.

Limitations. BioCoReT benefits from matched controls and multiple Cell Painting replicates. When these are unavailable, the corresponding reliability signals cannot be fully exploited.

Stage 2 also depends on the quality of the task-specific teachers. Extremely sparse assay tasks may provide insufficient supervision for learning reliable decision boundaries. This can limit the quality of candidate pseudo-labels.

The topology-based reliability signals depend on the quality of local molecular and morphological neighborhoods. These neighborhoods may be unreliable for difficult or out-of-distribution compounds. Combining multiple signals reduces this dependence, but does not eliminate it.

Finally, our experiments focus on three Cell Painting benchmarks. The current formulation does not explicitly model the mechanism of missing compound–task labels. Its generalization to other cell types, imaging protocols, assay collections, and perturbation modalities remains to be further evaluated.