CellMINT: Disentangled Multimodal Learning with Biologically Informed Task Interaction for Assay Prediction in Drug Discovery

Anonymous authors
Paper under double-blind review

Abstract

Cell Painting phenotypes are increasingly integrated with chemical perturbation information to support diverse assay prediction tasks in drug discovery. However, existing multimodal approaches often emphasize modality-shared information while underutilizing task-relevant modality-specific signals. Moreover, most multi-task methods for joint assay prediction do not explicitly model interactions among biologically related assay endpoints, limiting effective knowledge sharing across them. In this paper, we propose CellMINT, a multimodal multi-task framework that enhances **Cell** Painting-based prediction through disentangled **M**ultimodal learning and biologically **IN**formed **T**ask interaction. To identify task-relevant multimodal evidence, CellMINT disentangles multimodal features into modality-shared and modality-specific components, enabling each assay endpoint to adaptively aggregate the information most relevant to its prediction. To further exploit complementary information across related tasks, CellMINT constructs a biological prior from assay context to guide selective task interaction among informative endpoints. Experiments on three public Cell Painting benchmarks demonstrate that CellMINT improves average predictive performance and high-performing task coverage over unimodal and multimodal baselines, highlighting its potential for comprehensive assay profiling in drug discovery.

1 Introduction

Cell Painting (CP) has emerged as an important high-content imaging technique for phenotypic drug discovery. By labeling different cellular structures with fluorescent dyes, CP experiments generate multichannel microscopy images that capture perturbation-induced morphological changes (Bray et al., 2016; Chandrasekaran et al., 2021; Seal et al., 2025). These images provide rich, target-agnostic phenotypic readouts for a wide range of biomedical applications, including mechanism-of-action identification (Wong et al., 2023; Tian et al., 2023), bioactivity prediction (Hofmarcher et al., 2019; Moshkov et al., 2023; Fredin Haslum et al., 2024; Pati et al., 2025), and toxicity assessment (Seal et al., 2022; Ewald et al., 2026), supporting compound screening and prioritization in early-stage drug discovery (Sanchez et al., 2026).

Alongside the phenotypic information captured by CP, compound-level molecular information provides a chemical view of the underlying perturbations. These complementary modalities motivate multimodal learning approaches linking compound chemistry to induced morphological responses. Recent studies employ cross-modal contrastive learning to align both modalities in a shared latent space, enabling retrieval between CP phenotypes and candidate compounds (Sanchez-Fernandez et al., 2023; Fradkin et al., 2024; Wang et al., 2024; Yu et al., 2025; Lu et al., 2025). Beyond retrieval, these representations have also shown promise for diverse drug discovery applications (Sanchez-Fernandez et al., 2023). In this work, we focus on predicting diverse biological assay endpoints, namely whether a compound induces specific biological responses measured in separate assay experiments. Accurate in silico prediction of these endpoints from multimodal perturbation representations enables scalable compound profiling and represents an important objective in predictive biology. Recent AI/ML approaches formulate this as a multi-task learning problem, jointly modeling multiple assay endpoints to leverage shared biological information across tasks (Hofmarcher et al., 2019; Moshkov et al., 2023; Fredin Haslum et al., 2024; Pati et al., 2025; 2026).

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.

Despite progress in multimodal and multi-task learning, maximizing the predictive utility of multimodal representations for a variety of biological assays remains challenging. First, different assay endpoints may rely on distinct signals from CP and chemical information. For example, cytotoxicity and mitochondrial endpoints may be more strongly associated with morphological changes captured by CP, whereas enzyme inhibition and target-binding endpoints may rely more on chemistry-specific molecular information (Seal et al., 2022; Moshkov et al., 2023). Unlike many conventional multi-task settings with broadly similar predictive requirements, this heterogeneity needs both comprehensive capture of multimodal information and task-specific utilization of distinct evidence. However, existing cross-modal alignment methods primarily emphasize information shared across modalities, potentially underutilizing modality-specific signals important for individual tasks (Sanchez-Fernandez et al., 2023; Fradkin et al., 2024; Wang et al., 2024; Yu et al., 2025; Lu et al., 2025). Furthermore, how to effectively fuse multimodal information remains an open problem in this domain. Existing approaches often leave redundant and complementary signals entangled in the learned representations, making them difficult for individual tasks to selectively exploit (Tian et al., 2023; Pati et al., 2026). Second, biological assay prediction poses a large-scale multi-task learning problem. Hundreds of heterogeneous assay endpoints may be modeled jointly, while each compound is typically annotated for only a small subset of them. Such sparse supervision makes cross-task information transfer important for robust learning, yet the diversity of assay mechanisms, biological contexts, and experimental readouts makes it difficult to determine which tasks should share information. Existing multi-task models often treat assay endpoints as separate prediction heads supervised only by their observed labels (Hofmarcher et al., 2019; Fredin Haslum et al., 2024), while the few task-interaction approaches are typically restricted to specialized assay domains, such as protein-target activity (Pati et al., 2025; 2026) or mitochondrial toxicity (Herman et al., 2023). How to enable selective knowledge transfer across diverse assay collections therefore remains unresolved.

Motivated by these challenges, we propose **CellMINT**, a multimodal multi-task framework for biological assay prediction that integrates task-adaptive evidence at two complementary levels. First, control-guided multimodal disentanglement preserves phenotype-specific, modality-shared, and chemistry-specific perturbation signals, while task-conditioned queries selectively aggregate them into assay-specific representations. Second, CellMINT constructs a directed task-relation prior from biological and experimental assay context to guide interactions among informative helper tasks, enabling selective cross-task transfer under sparse supervision. Our main contributions are:

- We develop a task-adaptive multimodal learning framework for heterogeneous biological assay prediction that models phenotype-specific, modality-shared, and chemistry-specific information, while enabling each assay endpoint to selectively utilize the evidence most relevant to its prediction.
- We propose a novel biologically informed task-interaction mechanism to enable selective cross-task information transfer and improve joint prediction. To the best of our knowledge, CellMINT is the first task-interaction framework designed for heterogeneous assay collections spanning diverse biological domains.
- We validate CellMINT on three public benchmarks spanning molecular bioactivity, cellular response, toxicity, metabolism, and anti-infective assays. Results demonstrate significant performance gains and broader high-performing task coverage over the baselines, with relative improvements of 5.1%, 3.6%, and 5.4% in mean AUROC and 17.2%, 5.6%, and 44.4% in task coverage at AUROC ≥ 0.7 across the three benchmarks, respectively.

2 Related Works

2.1 Multimodal Learning in Cell Painting

Recent multimodal research has developed various strategies to connect CP images with chemical perturbation information. Most works adopt cross-modal contrastive objectives to align morphological phenotypes with molecular fingerprints (Sanchez-Fernandez et al., 2023; Yu et al., 2025), pretrained chemical representations (Fradkin et al., 2024; Wang et al., 2024), perturbation descriptions (Lu et al., 2025), and chemical perturbation contexts (Zhang et al., 2026). While these alignment methods capture cross-modal correspondence and support retrieval, they primarily emphasize information shared across modalities and may underutilize complementary modality-specific

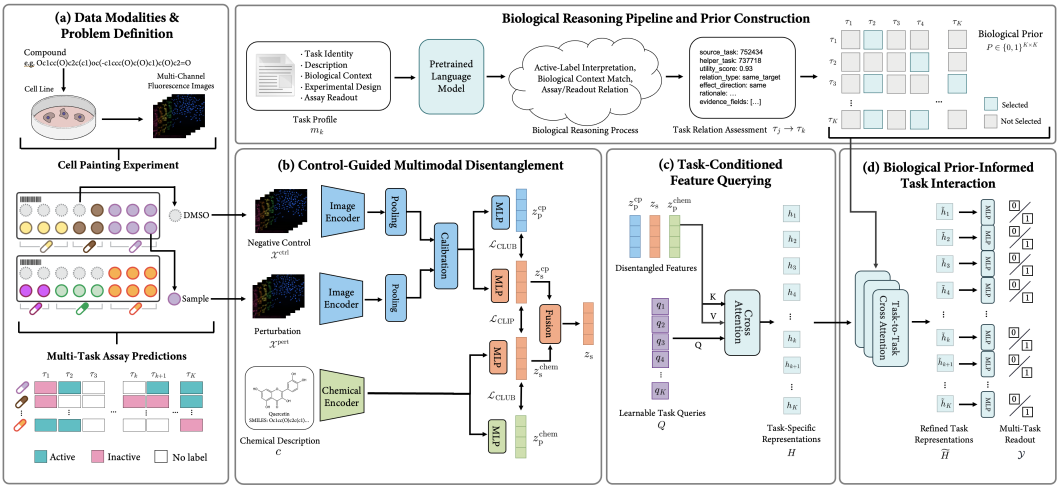

Figure 1: Overview of the proposed CellMINT framework. (a) Perturbed and plate-matched negative-control CP images, together with compound descriptions, are used for sparsely labeled multi-task assay prediction. (b) Control-guided multimodal disentanglement calibrates phenotypic features and models phenotype-specific, modality-shared, and chemistry-specific representations. (c) Learnable task queries aggregate task-relevant multimodal evidence into task-specific representations. (d) A directed biological prior inferred from structured task profiles guides selective task-to-task attention before multi-task readout.

signals valuable for biological assay prediction. Beyond alignment, several studies explore multimodal fusion to integrate phenotypic and chemical representations for downstream prediction (Tian et al., 2023; Dee et al., 2024; Rao et al., 2025; Pati et al., 2026). However, these approaches leave heterogeneous multimodal signals entangled within a single fused representation, making it difficult for diverse biological assays to selectively exploit them. DisentangledSSL explicitly separates modality-shared and modality-specific information through self-supervised representation learning, but its disentanglement is task-agnostic and does not model endpoint-specific utilization of these components (Wang et al., 2025). Related disentanglement strategies have also been explored in other multimodal settings (Du et al., 2025; Yao et al., 2024; Liang et al., 2023). CellMINT builds on this direction by modeling phenotype-specific, modality-shared, and chemistry-specific representations and enabling each assay to selectively query the evidence most relevant to its prediction.

2.2 Multi-Task Prediction in Cell Painting

Multi-task learning provides a natural formulation for CP-based assay prediction, as a perturbation representation can inform multiple biological endpoints. Existing approaches commonly feed the learned representation into a multi-task classifier, with each endpoint supervised by its observed labels (Hofmarcher et al., 2019; Fredin Haslum et al., 2024; Moshkov et al., 2023). However, this formulation may underutilize beneficial cross-task transfer under sparse supervision. Several CP studies exploit task relationships to enhance predictions in specialized assay domains (Herman et al., 2023; Pati et al., 2025; 2026). However, these approaches often rely on auxiliary knowledge or relation structures tailored to a particular assay domain. In contrast, CellMINT infers task relationships from broadly available assay metadata and biological context, enabling selective knowledge transfer across heterogeneous assay collections.

3 Method

3.1 Problem Setup and Framework Overview

In CP, a perturbed well contains cells exposed to a compound, and negative-control wells contain cells exposed only to the vehicle, acting as a reference capturing background variation (Bray

et al., 2016). For a perturbed well n, we denote the image profile as $\mathcal{X}_n^{\text{pert}} := \{x_n^{\text{pert}}\}$. The collective image profile of the control wells, which is on the same plate as n, is denoted as $\mathcal{X}^{\text{ctrl}} := \{\{x^{\text{ctrl}}\}_w\}, w \in \mathcal{W}$. Here, x represents a five-channel fluorescence CP field-of-view, and $\mathcal{W}$ denotes the set of control wells on the same plate. For each $\mathcal{X}_n^{\text{pert}}$, the textual description of the compound is defined as c_n by compiling the compound name and SMILES (Weininger, 1988), following the procedure from (Lu et al., 2025). Next, the set of binary assay endpoints is denoted as $\mathcal{T} = \{\tau_k\}_{k=1}^K$. For n, the activity labels $\mathcal{Y}_n := \{y_{n,k}\}_{k=1}^K$ are observed only for a subset $\mathcal{T}_n^{\text{obs}} \subseteq \mathcal{T}$. The goal is to learn a multi-task classifier that predicts $\{y_{n,k} \mid \tau_k \in \mathcal{T}_n^{\text{obs}}\}$ for a multimodal sample $\mathcal{X}_n := (\mathcal{X}_n^{\text{pert}}, \mathcal{X}^{\text{ctrl}}, c_n)$. For simplicity, we drop the sample index n in the following sections.

To this end, we present **CellMINT**, illustrated in Fig. 1. It addresses two key factors for assay prediction: (1) identifying multimodal evidence relevant to the task, and (2) exploiting complementary information from biologically related tasks. The first factor is addressed by disentangling the multimodal input into phenotype-specific, modality-shared, and chemistry-specific features (Sec.3.2), and adaptive querying by learnable task tokens to derive task-specific representations (Sec.3.3). For the second factor, a biological prior is constructed using a pretrained language model, which enables knowledge transfer among relevant task pairs (Sec. 3.4).

3.2 Control-Guided Multimodal Disentanglement

CP images and compound descriptions provide overlapping yet complementary information about a perturbation. To leverage these heterogeneous signals, we propose a control-guided multimodal disentanglement module to mitigate background variation and transform the multimodal input into phenotype-specific, modality-shared, and chemistry-specific representations, as shown in Fig. 1(b).

Sample-level unimodal representations. For CP, first, each fluorescence channel of x^{pert} is encoded using a pretrained image encoder, and the resulting channel embeddings are averaged. Afterwards, $\mathcal{X}^{\text{pert}}$ level embedding $z_{\text{pert}}^{\text{cp}}$ is obtained using gated attention pooling (Ilse et al., 2018) over the x^{pert} embeddings. The same procedure is applied to encode the plate-matched individual control wells, which are then averaged to obtain $\mathcal{X}^{\text{ctrl}}$ level embedding $z_{\text{ctrl}}^{\text{cp}}$. For compound description c, a pretrained text encoder is used to obtain z^{chem}.

Control-guided calibration. The obtained embedding $z_{\text{pert}}^{\text{cp}}$ mixes two contributions: a baseline effect reflecting the intrinsic cellular state under the given experimental conditions, and a perturbation effect induced by compound exposure. The plate-matched control embedding, $z_{\text{ctrl}}^{\text{cp}}$, can capture only the baseline phenotype—cells assayed under identical conditions but without compound treatment. Because the two embeddings share the same plate-level covariates, taking their difference isolates the compound-induced signal while suppressing technical and baseline variation. Therefore, we calibrate $z_{\text{pert}}^{\text{cp}}$ using $z_{\text{ctrl}}^{\text{cp}}$:

$$z^{\text{cp}} = z_{\text{pert}}^{\text{cp}} - g \cdot z_{\text{ctrl}}^{\text{cp}} \tag{1}$$

where $g > 0$ is a learnable coefficient constrained to be positive, enabling adaptive removal of control signals and more faithful capture of the compound-induced phenotype.

Disentangled representation learning. To capture the heterogeneous information carried by the modalities, we learn modality-shared and modality-specific embeddings from the calibrated image feature z^{cp} and chemical feature z^{chem}. Independent MLPs derive the modality-shared embeddings z_{s}^{cp} and $z_{\text{s}}^{\text{chem}}$, and modality-specific embeddings z_{p}^{cp} and $z_{\text{p}}^{\text{chem}}$.

The modality-shared embeddings are optimized to capture information that is consistently represented across both CP images and chemical descriptions. We enforce cross-modal consistency between them using a symmetric InfoNCE objective (Radford et al., 2021),

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2B} \sum_{b=1}^{B} \left(\ell_{\text{cp}\to\text{chem}}^{(b)} + \ell_{\text{chem}\to\text{cp}}^{(b)} \right) \tag{2}$$

where B is the batch size. $\ell_{\text{cp}\to\text{chem}}^{(b)}$ and $\ell_{\text{chem}\to\text{cp}}^{(b)}$ denote the log-softmax alignment scores for sample $b \in B$, encouraging CP phenotypes and compound descriptions corresponding to the same perturbation to have higher similarity than unmatched in-batch pairs. We then concatenate the modality-shared representations to obtain the shared multimodal embedding $z_{\text{s}} = \text{MLP}\left([z_{\text{s}}^{\text{cp}} || z_{\text{s}}^{\text{chem}}]\right)$.

In contrast, the modality-specific embeddings retain complementary information that is more directly expressed in each modality, such as broad cellular health states in CP images and molecular scaffold information in the chemical description. To encourage separation between modality-shared and modality-specific information, we reduce the statistical dependence between their embeddings within each modality. As direct estimation of mutual information is intractable, we minimize the Contrastive Log-ratio Upper Bound (CLUB) estimates (Cheng et al., 2020) as,

$$\mathcal{L}_{\mathrm{CLUB}} = \widehat{I}_{\mathrm{CLUB}}\left(z_{\mathrm{s}}^{\mathrm{cp}}, z_{\mathrm{p}}^{\mathrm{cp}}\right) + \widehat{I}_{\mathrm{CLUB}}\left(z_{\mathrm{s}}^{\mathrm{chem}}, z_{\mathrm{p}}^{\mathrm{chem}}\right) \tag{3}$$

where $\widehat{I}_{\mathrm{CLUB}}$ is the CLUB-estimated mutual-information upper bound for a pair of embeddings. The resulting disentangled embeddings $\{z_{\mathrm{p}}^{\mathrm{cp}}, z_{\mathrm{s}}, z_{\mathrm{p}}^{\mathrm{chem}}\}$ provide complementary multimodal information for predicting the diverse tasks.

3.3 Task-Conditioned Feature Querying

Different tasks may rely on different combinations of the above disentangled features as they span diverse biological processes and experimental contexts. Therefore, we introduce task-conditioned feature querying in Fig. 1(c), which allows each task to selectively attend to these features and form task-specific representations for predictions.

Learnable task queries. To enable task-relevant retrieval of information, we assign each task $\tau_k \in \mathcal{T}$ a learnable token $q_k \in \mathbb{R}^d$. q_k acts as a query and learns to retrieve information from the disentangled features. We stack all q_k as, $Q = [q_1; q_2; \cdots; q_K] \in \mathbb{R}^{K\times d}$. These task queries are task-level parameters shared across samples. They are randomly initialized and optimized end-to-end under the multi-task prediction objective.

Task-conditioned evidence aggregation. Task queries are used to construct task-specific multimodal representations from the disentangled features. Specifically, we use Q as queries and $Z = [z_{\mathrm{p}}^{\mathrm{cp}}; z_{\mathrm{s}}; z_{\mathrm{p}}^{\mathrm{chem}}]$ as keys and values, to compute $H = \mathrm{CrossAttn}(Q, Z, Z)$. These interactions enable a task to differently weight the modality-specific and modality-shared features, and aggregate information relevant for the task's prediction. The output is presented as $H = [h_1; h_2; \cdots; h_K] \in \mathbb{R}^{K\times d}$, where h_k is the task-specific multimodal representation for τ_k.

3.4 Biological Prior-Informed Task Interaction

Assay endpoints are often biologically related through shared targets, pathways, or toxicity mechanisms, and may therefore benefit from cross-task knowledge transfer, particularly in sparsely labeled settings where tasks receive limited supervision. Hence, we leverage biological knowledge to construct a prior that guides task interactions for augmenting task representations H, as in Fig. 1(d).

Per-task biological profiling. We begin by characterizing the biological meaning of each task τ_k. For each τ_k, we extract metadata from public biochemical and pharmacological resources describing its biological context, the molecular or cellular process being probed, and the experimental readout defining the prediction task. To unify characterization across heterogeneous task domains, we map the collected metadata to a predefined schema of standardized biological and experimental fields, forming a structured task profile m_k for assessing predictive relationships between tasks. Further details on the task-profile schema are provided in Appendix B.4.

Reasoning task relations and prior construction. We define a reasoning pipeline that leverages a pretrained language model to infer predictive relationships among tasks. First, for each task, the model interprets the biological meaning of an *active* label. Second, for each task pair $(\tau_k, \tau_j) \in \mathcal{T} \times \mathcal{T}$ with $k \neq j$, where τ_k denotes a target task and τ_j a candidate helper, the model analyzes the corresponding task profiles (m_k, m_j) to identify biological and experimental relevance, including shared targets or pathways, matched biological contexts, and related experimental protocols or readouts. Based on the relevance, the model evaluates whether the activity signal of τ_j may be informative for predicting τ_k. For each directed relation $j \to k$, the model returns a structured assessment comprising a predictive-utility score $s_{k\leftarrow j} \in [0, 1]$ and supporting evidence. The score is assigned according to a predefined rubric, with higher values indicating greater expected utility of τ_j for predicting τ_k. The prompting protocol and scoring rubric are provided in Appendix B.4.

Finally, we construct a directed task-interaction prior $P \in \{0,1\}^{K\times K}$ from the inferred predictive-utility scores. For each target task τ_k, candidate helpers satisfy $s_{k\leftarrow j} \geq \delta$, where $\tau_j \in \mathcal{T}$ and $j \neq k$. Among these candidates, we retain at most the top-R highest-scoring helper tasks. Formally,

$$P_{k,j} = \begin{cases} 1, & \text{if } \tau_j \in \mathrm{Top}_R^{j\neq k}\left(s_{k\leftarrow j} \mid s_{k\leftarrow j} \geq \delta\right) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Additionally, we follow (Pati et al., 2026) to further sparsify P using the label correlations computed from the training set, as sparsely labeled datasets may provide insufficient joint supervision for some inferred task relations. Full prior-construction details are provided in Appendix B.4.

Prior-informed task interactions. We use the prior P to guide task interactions, allowing each task to adaptively aggregate complementary information from its selected high-utility helpers. For a task τ_k with representation h_k, we collect the helper task representations as $H_k^{\text{prior}} = [h_j \mid P_{k,j} = 1]$. We use h_k as query and H_k^{prior} as keys and values, to augment h_k with task-relevant signals,

$$\widetilde{h}_k = h_k + \mathrm{CrossAttn}\left(h_k, H_k^{\text{prior}}, H_k^{\text{prior}}\right) \tag{5}$$

If τ_k has no selected helper, we set $\widetilde{h}_k = h_k$. With leveraging only prior-selected task representations, information transfer is restricted to biologically relevant helper relations. By preventing noisy transfer between unrelated tasks, this selective design enables a unified task-interaction mechanism for joint prediction across heterogeneous task categories.

Multi-task classification. We feed the augmented task representation $\widetilde{h}_k$ into a task-specific MLP predictor to predict the label y_k for task τ_k. Provided labels only for a subset $\mathcal{T}^{\text{obs}} \subseteq \mathcal{T}$ for each sample, we compute the masked binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ using only the observed labels. The multi-task classifier is optimized end-to-end with loss weights γ_{CLIP} and γ_{CLUB},

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \gamma_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \gamma_{\text{CLUB}}\mathcal{L}_{\text{CLUB}}. \tag{6}$$

4 Experiment and Results

4.1 Experimental Setup

Datasets. Experiments are conducted on three public CP benchmarks covering different data sources, assay endpoints, and cell types. **cpg0012** (Hofmarcher et al., 2019) links ChEMBL activity labels to Cell Painting profiles from (Bray et al., 2017), comprising 209 tasks for 10,574 compounds profiled in U2OS cells. **cpg0016** (Fredin Haslum et al., 2024) is built from the JUMP-CP dataset (Chandrasekaran et al., 2023), covering 29 ChEMBL tasks over 10,660 U2OS-imaged compounds. **cpg0037** (Ewald et al., 2026) is a toxicity-related benchmark including 1,085 compounds profiled in primary human hepatocytes and annotated with ToxCast activity labels. We convert its original single-task evaluation protocol into a unified multi-task benchmark with 354 tasks. All benchmarks are sparsely labeled, with further details provided in Appendix C.1.

Baselines. We benchmark CellMINT against diverse unimodal and multimodal baselines, adapting all methods to the same multi-task setting for fair comparison. For CP-only baselines, we evaluate CellProfiler (Stirling et al., 2021) and DINO (Oquab et al., 2023) features with multi-task learning, alongside the end-to-end image model HTI-CNN (Hofmarcher et al., 2019). For chemical-only methods, we assess ECFP4 molecular fingerprints (Rogers & Hahn, 2010) and ChemBERTa features (Chithrananda et al., 2020) with prediction heads. Multimodal baselines include the cross-modal alignment methods CLOOME (Sanchez-Fernandez et al., 2023), CellCLIP (Lu et al., 2025), MICON (Yu et al., 2025), and InfoCORE (Wang et al., 2024), as well as the multimodal fusion methods MINER (Rao et al., 2025) and CP-Chem-MoA (Tian et al., 2023).

Evaluation protocol and metrics. We follow the predefined compound-level splits and evaluation protocols, aggregating predictions at the compound level before metric computation (see Appendix C.3). For each approach, we report mean AUROC, AUPRC, and macro F1, averaged across tasks within each dataset. In addition, we compute the high-performing task coverage at AUROC thresholds of 0.7, 0.8, and 0.9, following (Hofmarcher et al., 2019; Pati et al., 2026).

Modality	Methods	cpg0012 (209 tasks)			cpg0016 (29 tasks)			cpg0037 (354 tasks)		
		$\overline{\text{AUROC}}$	$\overline{\text{AUPRC}}$	$\overline{\text{Macro F1}}$	$\overline{\text{AUROC}}$	$\overline{\text{AUPRC}}$	$\overline{\text{Macro F1}}$	$\overline{\text{AUROC}}$	$\overline{\text{AUPRC}}$	$\overline{\text{Macro F1}}$
Image	CellProfiler + Head	0.705	0.616	0.512	0.663	0.662	0.600	0.600	0.365	0.083
	DINO + Head	0.713	0.630	0.505	0.608	0.616	0.570	0.604	0.369	0.080
	HTI-CNN	0.718	0.646	0.489	0.668	0.662	0.574	0.618	0.367	0.059
Chemical	ECFP4 + Head	0.662	0.538	0.299	0.707	0.686	0.603	0.600	0.351	<u>0.117</u>
	ChemBERTa + Head	0.626	0.511	0.343	0.671	0.652	0.580	<u>0.646</u>	0.360	0.102
Multimodal	CLOOME + Head	0.696	0.624	0.436	0.591	0.609	0.546	0.593	0.361	0.072
	CellCLIP + Head	0.737	<u>0.685</u>	<u>0.525</u>	0.650	0.649	0.585	0.637	0.373	0.073
	MICON + Head	0.704	0.630	0.475	0.615	0.616	0.540	0.577	0.344	0.063
	InfoCORE + Head	0.720	0.632	0.470	0.628	0.626	0.572	0.632	0.372	0.058
	MINER	0.668	0.553	0.372	<u>0.692</u>	0.671	0.598	0.629	0.351	0.049
	CP-Chem-MoA	<u>0.745</u>	0.684	0.519	<u>0.742</u>	<u>0.729</u>	<u>0.610</u>	0.618	<u>0.373</u>	0.093
	CellMINT	**0.783**†	**0.704**†	**0.568**†	**0.769**†	**0.736**	**0.669**†	**0.681**†	**0.404**†	**0.139**†

Table 1: $\overline{\text{Mean}}$ task-level performance across assay endpoints on three Cell Painting benchmarks. Higher values indicate better performance. **Bold** and <u>Underline</u> denote the best and the second best results in each column, respectively. † indicates a statistically significant improvement over the second best baseline, computed by a paired two-sided t-test ($p < 0.05$).

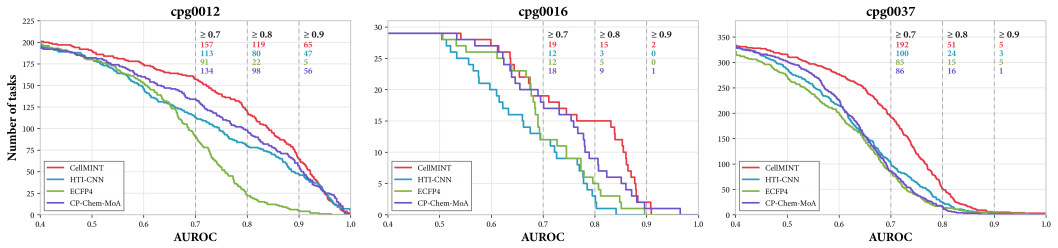

Figure 2: Inverse cumulative histograms across three Cell Painting benchmarks. The Y-axis denotes the number of tasks exceeding the AUROC threshold on the X-axis. The task counts at AUROC thresholds 0.7, 0.8, and 0.9 are reported, with higher counts indicating better performance.

Implementation details. To extract sample-level features, we use pretrained DINOv2 (Oquab et al., 2023) as the image encoder and BERT (Devlin et al., 2019) as the chemical description encoder. Feature and task representations use a hidden dimension of $d = 256$. The task-conditioned feature querying module uses a 3-layer, 8-head cross-attention block. We use OpenAI GPT-5.5 (OpenAI, 2026) to reason over structured task profiles and construct the biological prior. Prior-informed task interaction is implemented with a single cross-attention layer. Loss weights are set to $\gamma_{\text{CLIP}} = 0.15$ and $\gamma_{\text{CLUB}} = 0.01$. Further details are in Appendices B and C.4.

4.2 Assay Endpoints Modeling Performance

We compare CellMINT against diverse baselines in Table 1. Across all three benchmarks, CellMINT consistently achieves the best overall performance, with mean AUROC of 0.783 on cpg0012, 0.769 on cpg0016, and 0.681 on cpg0037. Compared to unimodal baselines, CellMINT improves performance by leveraging task-relevant complementary information from morphological phenotypes and chemical descriptions. Among multimodal baselines, CellMINT outperforms cross-modal alignment methods such as CLOOME, CellCLIP, MICON, and InfoCORE, which primarily emphasize modality-shared information, highlighting the value of capturing both modality-shared and modality-specific signals for biological assay predictions. CellMINT also surpasses multimodal fusion baselines, MINER and CP-Chem-MoA, demonstrating the benefits of multimodal disentanglement for task-conditioned feature querying and selective knowledge transfer across related tasks.

To further examine performance across individual tasks, Fig. 2 shows inverse cumulative histograms of task-level AUROC for CellMINT and representative unimodal and multimodal baselines. Across the three datasets, CellMINT generally achieves higher task coverage across the AUROC spectrum. At an AUROC threshold of 0.7, CellMINT reaches 157 of 209 tasks (75.1%) on cpg0012, 19 of 29 tasks (65.5%) on cpg0016, and 192 of 354 tasks (54.2%) on cpg0037, with advantages maintained

	Method	AUROC	AUPRC	Macro F1
Task-adaptive multimodal learning (w/o task interaction)	Baseline (Concat)	0.740	0.641	0.481
	+ Control-guided calibration	0.748	0.667	0.506
	+ Disentanglement	0.753	0.678	0.519
	+ Task-conditioned query	0.767	0.678	0.543
Task-interaction strategies	No interaction	0.767	0.678	0.543
	Unrestricted	0.747	0.685	0.502
	Random	0.759	0.669	0.541
	Language-model prior only	0.777	0.703	0.541
	Label-corr. prior only	0.771	0.695	0.548
	Language-model + Label-corr. (CellMINT)	**0.783**	**0.704**	**0.568**

Table 2: Progressive component addition and task-interaction ablation on cpg0012. The first block progressively introduces control-guided calibration, disentangled representation learning, and task-conditioned feature querying. The second block compares different task-interaction strategies building on the final model configuration from the first block.

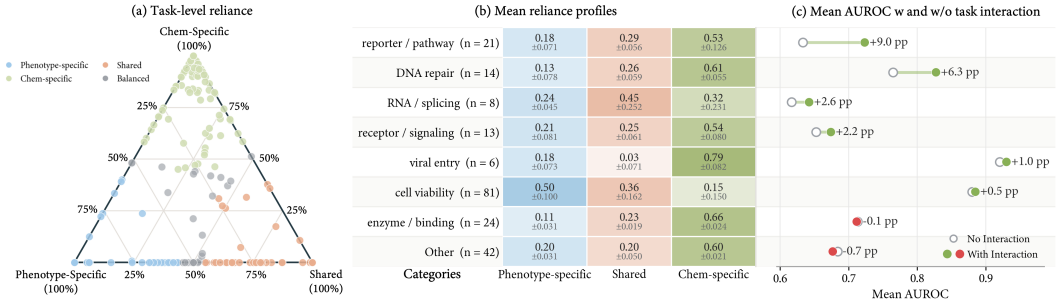

Figure 3: Task-level analysis on cpg0012. (a) Task-level reliance on phenotype-specific, modality-shared, and chemistry-specific features across 209 tasks. (b) Mean reliance profiles by assay category, reported as mean $\pm$ std. across splits; n denotes the number of assays per category. (c) Percentage change in mean AUROC w/ and w/o task interaction for the assay categories.

at the more stringent thresholds of 0.8 and 0.9. These results suggest that the gains are broadly distributed across tasks rather than being driven by a small subset. Overall, CellMINT improves both average predictive performance and the coverage of well-predicted tasks across heterogeneous CP benchmarks, supporting its effectiveness for biological assay prediction.

4.3 Ablation Study

Task-adaptive multimodal learning. To validate the design of learning disentangled multimodal representations (Sec. 3.2) and enabling task-conditioned query (Sec. 3.3), we progressively extend a feature-concatenation baseline ($z^{\text{cp}}_{\text{pert}}$, z^{chem}), as shown in the first block of Table 2. Control calibration improves AUROC from 0.740 to 0.748, indicating the benefit of using matched negative controls to reduce background variation. Disentangled representation learning further raises AUROC to 0.753 by explicitly modeling modality-shared and modality-specific information rather than leaving heterogeneous signals entangled under direct concatenation. Finally, task-conditioned feature querying enhances predictive performance to the AUROC of 0.767, highlighting the contribution of allowing each task to selectively aggregate the most relevant multimodal evidence.

Task-to-task interaction. To evaluate the biological prior-informed task interaction module (Sec. 3.4), we investigate two questions: whether cross-task information transfer can improve assay prediction, and how different task interaction strategies influence the effectiveness of such transfer. Using the final model configuration from the first block in Table 2 as the no-interaction baseline, we compare five task-interaction strategies. Unrestricted interaction allows each task to attend to all other tasks, whereas random interaction preserves the sparsity of the prior but randomly reas-

signs helper-task connections. We further evaluate interaction guided only by the language-model-derived prior, interaction guided only by training-label correlations, and their combination used in CellMINT. As shown in the second block of Table 2, the language-model prior alone improves AUROC from 0.767 to 0.777, while the label-correlation prior alone reaches an AUROC of 0.771. Combining both sources of task relations further improves performance to 0.783 AUROC. In contrast, unrestricted and random interactions underperform with AUROCs of 0.747 and 0.759, respectively. These results indicate that task interaction is most beneficial when information is exchanged among biologically relevant tasks, and that combining the language-model prior with label-correlation refinement enables more effective selection of helper tasks and cross-task knowledge transfer.

4.4 Task-Level Analysis

Task reliance on disentangled features. To investigate how individual endpoints draw on different components of the disentangled representation, we conduct a component-level ablation analysis on the cpg0012 benchmark. Inspired by prior work on modality contribution analysis (Frank et al., 2021; Gat et al., 2021), we compare the per-task AUROC of CellMINT with three ablated variants obtained by removing the modality-shared, phenotype-specific, or chemistry-specific component. For each endpoint, we quantify its reliance on each component by the corresponding AUROC drop upon removal, clipped at zero and normalized across the three ablations. Figure 3(a) visualizes the resulting task-level reliance profiles in a ternary plot. The 209 endpoints are broadly distributed across the simplex rather than concentrating in a single region, suggesting that different assays exploit distinct combinations of the disentangled features. To reveal broader category-level patterns, we further average the normalized profiles within each task category in Fig. 3(b). Cell viability assays show greater reliance on phenotype-specific features, consistent with viability-related effects being directly reflected in cellular morphology. Meanwhile, RNA/splicing assays exhibit greater sensitivity to the removal of modality-shared features, possibly reflecting splicing-related signals represented in both compound chemistry and cellular phenotypes. By contrast, enzyme/binding and receptor/signaling assays place greater weight on chemistry-specific features, consistent with their closer association with molecular structure and compound–target activity. These category-level differences may suggest that CellMINT models biologically meaningful heterogeneity in how different assays use modality-shared and modality-specific information.

Impact of task interaction across assays. We examine how different tasks benefit from the biological prior-informed task interaction. For each task in the cpg0012 benchmark, we compute the AUROC difference between CellMINT and its no-interaction variant, and then aggregate these differences according to the assay categories used in Fig. 3(b). As shown in Fig. 3(c), task interaction enhances assay predictions in heterogeneous categories. The largest gains are observed for reporter/pathway and DNA repair tasks, with average AUROC improvements of 9.0 and 6.3 percentage points, respectively. These two categories contain dense, mechanistically coherent helper tasks that provide relevant signals across assays. In contrast, many tasks in the enzyme/binding and Other categories are relatively independent and have limited biological relatedness to the remaining endpoints, making cross-task transfer less effective. Overall, these results indicate that biological prior-informed task interaction is beneficial when sufficient cross-task relationships are available.

5 Conclusion

In this paper, we present CellMINT, a multimodal multi-task framework for Cell Painting-based biological assay prediction. By combining control-guided multimodal disentanglement with task-conditioned feature querying, CellMINT enables each assay endpoint to selectively aggregate phenotype-specific, modality-shared, and chemistry-specific evidence. A biological prior-informed task interaction module further enhances learning through selective information transfer from related helper tasks. Across three Cell Painting benchmarks, CellMINT consistently improves average predictive performance and achieves broader high-performing task coverage than unimodal and multimodal baselines, demonstrating its robustness and applicability across diverse assays. For future work, we will investigate the biological interpretability of task-specific feature reliance and task interaction patterns. Building on CellMINT's ability to model heterogeneous assay types, we will also explore joint learning across multiple Cell Painting benchmarks by unifying their assay endpoints and datasets, toward a comprehensive model for biological assay profiling in drug discovery.

AI USE STATEMENT

In this work, we used generative AI as part of the proposed methodology. Specifically, OpenAI GPT-5.5 (OpenAI, 2026) was used to infer directed predictive relationships between assay endpoints from structured biological and experimental metadata, producing the predictive-utility scores used to construct the biological task-interaction prior described in Sec. 3.4 and Appendix B.4. Anthropic Claude Opus 5 (Anthropic, 2026) was additionally used to repeat this inference procedure for the cross-model reproducibility analysis in Appendix D.3. Generative AI tools were additionally used for manuscript editing and software code editing. All AI-assisted code and text were reviewed and verified by the authors. No other uses requiring disclosure were made. We take responsibility for the final content of this work, including text, claims, results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sections 3 and 4.1 describe the overall methodology and experimental setup. Appendix B provides detailed descriptions of the data preprocessing, model architecture, and biological prior construction. Appendix C documents the datasets and data partitions, evaluation protocols, hyperparameter settings, training procedures, and implementation environment. Appendix D provides detailed benchmark results across predefined splits or folds, hyperparameter sensitivity analyses, and additional analyses of the biological task-relation inference. Upon acceptance, we will publicly release the source code, preprocessing and training scripts, and implementation instructions required to reproduce the reported experiments.

REFERENCES

Anthropic. What's new in claude opus 5, 2026. URL `https://platform.claude.com/docs/en/about-claude/models/whats-new-opus-5`. Accessed: 2026-07-31.

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, and others. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience*, 6(12):giw014, 2017.

Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature reviews drug discovery*, 20(2):145–159, 2021.

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

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. 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.

William Dee, Ines Sequeira, Anna Lobley, and Gregory Slabaugh. Cell-vision fusion: A Swin transformer-based approach for predicting kinase inhibitor mechanism of action from Cell Painting data. *Iscience*, 27(8), 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of*

the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186, 2019.

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 Computer Vision and Pattern Recognition Conference*, pp. 15549–15559, 2025.

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

Philip Fradkin, Puria Azadi, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How molecules impact cells: Unlocking contrastive phenomolecular retrieval. *Advances in Neural Information Processing Systems*, 37:110667–110701, 2024.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9847–9857, 2021.

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.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

Dorota Herman, Maciej M Kanduła, Lorena GA Freitas, Caressa van Dongen, Thanh Le Van, Natalie Mesens, Steffen Jaensch, Emmanuel Gustin, Liesbeth Micholt, Charles-Hugues Lardeau, and others. Leveraging cell painting images to expand the applicability domain and actively improve deep learning quantitative structure–activity relationship models. *Chemical Research in Toxicology*, 36(7):1028–1036, 2023.

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.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Bo Li, Bob Zhang, Chengyang Zhang, Minghao Zhou, Weiliang Huang, Shihang Wang, Qing Wang, Mengran Li, Yong Zhang, and Qianqian Song. PhenoProfiler: advancing phenotypic learning for image-based drug discovery. *Nature Communications*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:124505–124537, 2025.

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.

OpenAI. Introducing GPT-5.5, April 2026. URL `https://openai.com/index/introducing-gpt-5-5/`. Accessed: 2026-07-10.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and others. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pushpak Pati, Hsiu-Chi Cheng, Steffen Jaensch, Walid M Abdelmoula, Krishna Chaitanya, Michiel Van Dyck, Tomé Albuquerque, Samantha Allen, Litao Zhang, Tommaso Mansi, and others. MoMIL: Mixture of Multi-instance Learners for Modeling Multiple Compound Activities in High Content Imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 371–381. Springer, 2025.

Pushpak Pati, Bo Li, Abbas Rayabat Khan, Tomé Albuquerque, Steffen Jaensch, Amina Mollaysa, Walid M Abdelmoula, Samantha J Allen, Joke Reumers, Helai P Mohammad, and others. BiG-MINT: Biologically-guided Hierarchical Multimodal Integration for Modeling Multiple Compound Activities in Drug Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6982–6993, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, 9(1):225, January 2026. ISSN 2399-3642. doi: 10.1038/s42003-025-09500-y. URL https://doi.org/10.1038/s42003-025-09500-y.

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.

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.

Srijit Seal, Maria-Anna Trapotsi, Ola Spjuth, Shantanu Singh, Jordi Carreras-Puigvert, Nigel Greene, Andreas Bender, and Anne E Carpenter. Cell Painting: a decade of discovery and innovation in cellular imaging. *Nature methods*, 22(2):254–268, 2025.

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.

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.

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.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An Information Criterion for Controlled Disentanglement of Multimodal Data. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 3151–3180,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/08fe50bf209c57eecf0804f9f9ed639f-Paper-Conference.pdf.

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.

Erin Weisbart, Ankur Kumar, John Arevalo, Anne E Carpenter, Beth A Cimini, and Shantanu Singh. Cell Painting Gallery: an open resource for image-based profiling. *Nature methods*, 21(10):1775–1777, 2024.

Daniel R Wong, David J Logan, Santosh Hariharan, Robert Stanton, Djork-Arné Clevert, and Andrew Kiruluta. Deep representation learning determines drug mechanism of action from cell painting images. *Digital Discovery*, 2(5):1354–1367, 2023.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16416–16424, 2024.

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

Yuxin Zhang, Yiyao Li, Ping Shu Ho, Simon See, Zhenqin Wu, and Kevin Kin-Man Tsia. CP-Agent: Context-Aware Multimodal Reasoning for Cellular Morphological Profiling under Chemical Perturbations. In *The Fourteenth International Conference on Learning Representations*, 2026.

A NOTATIONS

Table 3 summarizes the main notations used in the paper.

Notation	Description
$\mathcal{X}_n^{\text{pert}}$	Cell Painting image profile of the perturbed well corresponding to sample n.
$\mathcal{X}^{\text{ctrl}}$	Collective image profile of the plate-matched negative-control wells.
c_n	Textual compound description of sample n, constructed from its compound name and SMILES.
$\mathcal{X}_n$	Multimodal input of sample n, defined as $(\mathcal{X}_n^{\text{pert}}, \mathcal{X}^{\text{ctrl}}, c_n)$.
$\mathcal{T} = \{\tau_k\}_{k=1}^K$	Set of K binary assay prediction tasks.
$\mathcal{T}_n^{\text{obs}}$	Subset of tasks with observed labels for sample n.
$y_{n,k} \in \{0, 1\}$	Binary activity label of sample n for task τ_k, observed when $\tau_k \in \mathcal{T}_n^{\text{obs}}$.
$z_{\text{pert}}^{\text{cp}}$	Cell Painting representation of the perturbed well.
$z_{\text{ctrl}}^{\text{cp}}$	Aggregated representation of the plate-matched negative-control wells.
z^{cp}	Control-calibrated Cell Painting representation.
z^{chem}	Chemical representation obtained from the compound description.
$z_{\text{s}}^{\text{cp}}, z_{\text{s}}^{\text{chem}}$	Modality-shared representations derived from the Cell Painting and chemical modalities, respectively.
$z_{\text{p}}^{\text{cp}}, z_{\text{p}}^{\text{chem}}$	Modality-specific representations derived from the Cell Painting and chemical modalities, respectively.
z_{s}	Fused modality-shared multimodal representation.
q_k	Learnable query token associated with task τ_k.
h_k	Task-specific multimodal representation for task τ_k.
m_k	Structured biological and experimental profile of task τ_k.
$s_{k \leftarrow j}$	Predictive-utility score of candidate helper task τ_j for target task τ_k.
$P_{k,j}$	Binary task-interaction prior indicating whether τ_j is selected as a helper for τ_k.
$\widetilde{h}_k$	Task representation augmented using prior-selected helper tasks.

Table 3: Main notations in the paper. The index n is omitted.

B METHOD

B.1 DATA PREPROCESSING

Raw Cell Painting images and associated metadata were obtained from the Cell Painting Gallery (Weisbart et al., 2024), and compound and assay annotations were linked to the images using plate, well, and field-of-view identifiers. We applied a common preprocessing procedure following (Hofmarcher et al., 2019). Briefly, each fluorescence channel was loaded from its original single-channel 16-bit TIFF image. For each field of view and channel independently, the lower boundary of the brightest 0.0028% of pixels was used as the upper intensity threshold. Values above this threshold were clipped, and the intensity range from zero to this threshold was linearly rescaled to $[0, 255]$, yielding an 8-bit image and reducing the influence of extreme-intensity artifacts. The five channel images corresponding to the same field of view were then arranged in a dataset-specific canonical order and stacked to form a five-channel image of shape $H \times W \times 5$, yielding each x^{pert} or x^{ctrl}. Field of views with missing or unreadable channels, non-two-dimensional images, or inconsistent channel dimensions were excluded. Each valid field-of-view image was stored as a `uint8` array in an NPZ container, while its plate, well, and field-of-view identifiers were retained for downstream linkage to compound annotations, assay labels, and plate-matched negative controls.

B.2 CONTROL-GUIDED MULTIMODAL DISENTANGLEMENT

Phenotypic feature extraction. Each of the five fluorescence channels in an individual field of view x^{pert} was independently encoded using the pretrained DINOv2 image encoder (Oquab et al., 2023), and the resulting channel embeddings were averaged to obtain a single representation v_i^{pert} for the i-th field of view. To aggregate multiple fields of view in $\mathcal{X}^{\text{pert}}$, we employed gated attention pooling (Ilse et al., 2018). Specifically, a learnable gated scoring function assigns a scalar relevance score to each field-of-view representation. The scoring function consists of two parallel learnable linear projections followed by $\tanh$ and sigmoid activations, respectively; their outputs are combined by element-wise multiplication and mapped to a scalar score through another learnable linear layer. A softmax across all fields of view converts the resulting scores into normalized, input-dependent attention weights α_i^{pert}. The perturbed-well representation is then obtained as

$$z_{\text{pert}}^{\text{cp}} = \sum_{i=1}^{|\mathcal{X}^{\text{pert}}|} \alpha_i^{\text{pert}} v_i^{\text{pert}}. \tag{7}$$

Thus, fields of view containing more informative phenotypic responses can contribute more strongly to the well-level representation. The same channel encoding and gated attention pooling procedure was applied separately to each plate-matched negative-control well in $\mathcal{X}^{\text{ctrl}}$, and the resulting control-well representations were averaged to obtain $z_{\text{ctrl}}^{\text{cp}}$.

Chemical feature extraction. For each sample n, the corresponding chemical description c_n was constructed from its compound metadata using a fixed template:

```
<compound name>.  SMILES: <canonical SMILES>.
```

To avoid potential task-label leakage, the chemical input contained no assay metadata, activity labels, or other task-specific information. Each description was tokenized using the `bert-base-cased` tokenizer, padded or truncated to a maximum sequence length of 128 tokens, and encoded using pretrained BERT (Devlin et al., 2019) to obtain z^{chem}.

Control-guided calibration. The perturbed Cell Painting representation contains both the compound-induced phenotype and baseline variation arising from plate-specific experimental conditions. Because the plate-matched negative controls capture this baseline variation without compound treatment, they provide a natural reference for isolating the perturbation effect. We therefore followed (Li et al., 2026) to calibrate the perturbed representation as

$$z^{\text{cp}} = z_{\text{pert}}^{\text{cp}} - g \cdot z_{\text{ctrl}}^{\text{cp}}. \tag{8}$$

To keep the subtraction coefficient positive and prevent excessive removal of phenotypic information, we did not optimize g directly. Instead, we optimized an unconstrained scalar $\hat{g}$ and transformed it using a sigmoid function:

$$g = \sigma(\hat{g}). \tag{9}$$

Because $\sigma(\hat{g}) \in (0, 1)$, this parameterization constrains the subtraction coefficient to $0 < g < 1$. The model then learned the appropriate subtraction strength during training.

Disentangled representation learning. To capture the heterogeneous information carried by the two modalities, we disentangled the calibrated phenotypic representation z^{cp} and chemical representation z^{chem} into modality-shared and modality-specific components. Following (Du et al., 2025), four independent projection MLPs were used to obtain the modality-shared representations z_s^{cp} and z_s^{chem}, and the modality-specific representations z_p^{cp} and z_p^{chem}, respectively. Each projection MLP consisted of two linear layers separated by a ReLU activation.

To encourage the modality-shared representations to capture information consistently expressed across the phenotypic and chemical modalities, we aligned z_s^{cp} and z_s^{chem} using a symmetric InfoNCE objective (Radford et al., 2021). For a mini-batch of size B, the pairwise cosine similarity between the phenotypic representation of sample b and the chemical representation of sample j was defined as

$$\text{sim}_{b,j} = \frac{\left(z_s^{\text{cp},(b)}\right)^\top z_s^{\text{chem},(j)}}{\left\|z_s^{\text{cp},(b)}\right\|_2 \left\|z_s^{\text{chem},(j)}\right\|_2}. \tag{10}$$

Based on these similarities, we computed bidirectional log-softmax alignment scores by using the phenotypic representations to retrieve their corresponding chemical representations and vice versa. The symmetric contrastive objective was defined as

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2B}\sum_{b=1}^{B}\left(\ell^{(b)}_{\text{cp}\to\text{chem}} + \ell^{(b)}_{\text{chem}\to\text{cp}}\right), \tag{11}$$

where $\ell^{(b)}_{\text{cp}\to\text{chem}}$ and $\ell^{(b)}_{\text{chem}\to\text{cp}}$ denote the directional log-softmax alignment scores computed from $\text{sim}_{b,j}$ with temperature $\tau = 0.07$. The objective increases the relative similarity between phenotypic and chemical representations corresponding to the same perturbation while reducing their similarity to unmatched in-batch representations. The two modality-shared representations were subsequently concatenated and mapped through a MLP to obtain the shared multimodal representation z_s.

While the InfoNCE objective encourages the modality-shared representations to capture signals consistently expressed across the phenotypic and chemical modalities, the modality-specific representations are designed to retain complementary, task-relevant information unique to each modality. To encourage z_p^{cp} and z_p^{chem} to capture information distinct from the modality-shared components, we reduced the statistical dependence between each shared–specific representation pair using CLUB (Cheng et al., 2020). Specifically, we instantiated two independent CLUB estimators, one for each modality $r \in \{\text{cp}, \text{chem}\}$. Each estimator approximates the conditional distribution of z_s^r given z_p^r as

$$q_{\theta_r}\left(z_s^r \mid z_p^r\right) = \mathcal{N}\left(z_s^r; \mu_r\left(z_p^r\right), I\right), \tag{12}$$

where μ_r was parameterized by a two-layer MLP, and I denotes the identity matrix. For a mini-batch of size B, the CLUB estimate was computed by contrasting the conditional log-likelihood of matched shared–specific representations with the average conditional log-likelihood over all in-batch pairings:

$$\widehat{I}_{\text{CLUB}}\left(z_s^r, z_p^r\right) = \frac{1}{B}\sum_{b=1}^{B}\log q_{\theta_r}\left(z_s^{r,(b)} \mid z_p^{r,(b)}\right) - \frac{1}{B^2}\sum_{b=1}^{B}\sum_{j=1}^{B}\log q_{\theta_r}\left(z_s^{r,(j)} \mid z_p^{r,(b)}\right). \tag{13}$$

The conditional estimators were learned from matched shared–specific representation pairs during joint model training. The CLUB estimates for the phenotypic and chemical modalities were summed to obtain $\mathcal{L}_{\text{CLUB}}$. Together with the task-prediction objective, this regularization encourages the modality-specific representations to preserve complementary signals relevant to downstream tasks while discouraging redundancy with their modality-shared counterparts. The resulting representations $\{z_p^{\text{cp}}, z_s, z_p^{\text{chem}}\}$ were passed to the task-conditioned feature-querying module.

B.3 Task-Conditioned Feature Querying

For each sample, we organized the three disentangled representations as a set of feature tokens,

$$Z = \left[z_p^{\text{cp}}; z_s; z_p^{\text{chem}}\right] \in \mathbb{R}^{3\times d}, \tag{14}$$

corresponding to phenotype-specific, modality-shared, and chemistry-specific information, respectively. Each prediction task τ_k was associated with a learnable query $q_k \in \mathbb{R}^d$. The query was shared across samples, whereas the feature tokens were sample-specific, allowing the same task to retrieve different multimodal evidence for different inputs.

Let Z_t denote the t-th feature token in Z, such that $Z_1 = z_p^{\text{cp}}$, $Z_2 = z_s$, and $Z_3 = z_p^{\text{chem}}$. For attention head m, the attention weight assigned by task query q_k to feature token Z_t was computed as

$$\alpha_{k,t}^{(m)} = \text{softmax}_t\left(\frac{(q_k W_m^Q)(Z_t W_m^K)^\top}{\sqrt{d_h}}\right). \tag{15}$$

The corresponding head output was computed as

$$o_k^{(m)} = \sum_{t=1}^{3}\alpha_{k,t}^{(m)} Z_t W_m^V. \tag{16}$$

Here, W_m^Q, W_m^K, and $W_m^V \in \mathbb{R}^{d \times d_h}$ are the query, key, and value projection matrices for head m, and $d_h = d/M$ is the dimension of each head. The outputs of the M heads were concatenated and transformed using an output projection $W^O \in \mathbb{R}^{Md_h \times d}$.

We applied this task-to-feature attention in L stacked, pre-normalized querying blocks. Each block comprised multi-head cross-attention and a position-wise feed-forward network, with a residual connection around each sublayer. The evolving task representations served as queries, while the three disentangled feature tokens served as keys and values. After the final block, layer normalization produced

$$H = [h_1; h_2; \ldots; h_K] \in \mathbb{R}^{K \times d}, \tag{17}$$

where h_k denotes the task-specific representation for τ_k.

In our implementation, we used $d = 256$, $M = 8$, and $L = 3$, giving a per-head dimension of $d_h = 32$. The feed-forward hidden dimension was 1,024, and dropout was set to 0.3. Task queries were initialized from $\mathcal{N}(0, 0.02^2)$ and optimized jointly with the remaining model parameters. The resulting task-specific representations aggregated the multimodal evidence most relevant to each assay endpoint and were subsequently passed to the prior-informed task-interaction module, which selectively incorporated complementary information from biologically related helper tasks.

B.4 Biological Prior-Informed Task Interaction

Per-task biological profiling. For each task τ_k, we collected its assay description and available structured metadata from the associated biochemical or pharmacological records. The collected information was converted into a shared schema covering the biological process, molecular target, modulation type, biological context, assay design, and experimental readout, forming the task profile m_k. Unavailable fields were left unspecified, and no activity labels or label-derived statistics were included.

For example, the profile of cpg0012 task `809086` was converted into the following structured profile:

`assay_id`: 809086
`task_id`: 809086
`description`: PUBCHEM_BIOASSAY: NCI human tumor cell line growth inhibition assay. Data for the SW-620 Colon cell line. (Class of assay: confirmatory)
`assay_type`: functional
`assay_category`: confirmatory
`assay_organism`:
`assay_strain`:
`assay_tissue`:
`assay_cell_type`: SW-620
`tid`: 81331
`relationship_type`: non-molecular target assigned
`chembl_id`: CHEMBL1964072
`bao_format`: BAO_0000219
`src_assay_id`: 81
`cell_id`: 305
`tissue_id`:

The structured task profiles were subsequently used to infer the biological and experimental meaning of each task.

Task-relation reasoning pipeline. Given the structured task profiles, we built a biological reasoning pipeline using a pretrained language model to assess the directed predictive utility of every ordered pair of tasks. Specifically, each task τ_k was considered once as the target task, while all remaining $K - 1$ tasks were provided as potential candidate helpers. For each candidate helper τ_j, the model assessed whether its active/inactive signal could provide useful evidence for predicting the active/inactive signal of τ_k. Thus, each language-model call produced $K - 1$ directed assessments of the form $\tau_j \to \tau_k$.

Within this biological reasoning pipeline, the model was instructed to first interpret the operational meaning of an active label for each assay and then assess task relevance based on shared targets, pathways, biological contexts, assay protocols, and experimental readouts. Based on the inferred biological and experimental relevance, the model then evaluated whether the activity signal of each candidate helper could be informative for predicting the target task. We queried GPT-5.5 (OpenAI, 2026) with high reasoning effort using the structured task profiles. Specifically, the system prompt was as follows:

> You are constructing a directed, metadata-only task-interaction prior for multi-task bioactivity prediction. Given one TARGET assay and a set of candidate HELPER assays, assess whether the active/inactive signal of each HELPER would be useful for predicting the active/inactive signal of the TARGET.
>
> First interpret what an active label operationally represents in each assay, such as target inhibition, pathway modulation, growth suppression, cytotoxicity, pathogen inhibition, or assay interference. Then compare the TARGET and HELPER profiles for shared targets or pathways, matched biological or cellular contexts, related assay designs or readouts, common phenotypic liabilities, and explicit counter-screen relationships. Evaluate predictive utility rather than broad topic or textual similarity.
>
> Use only the supplied assay descriptions and structured metadata.
>
> Assign a `utility_score` in $[0, 1]$:
>
> - 0.85–1.00: direct duplicate, same campaign replicate, same target/action/readout, paired measurement, or explicit paired counter-screen.
> - 0.65–0.84: strong metadata candidate from clearly shared operational label meaning, endpoint, liability, pathway, organism/cell context, assay mechanism, or close target family.
> - 0.45–0.64: plausible candidate worth testing; related phenotype, broad cell-health/toxicity/proliferation liability, related target family, similar assay construct, or concrete readout/artifact mechanism.
> - 0.25–0.44: weak but nontrivial candidate; broad biological/readout relation or incomplete metadata that may still be useful after label-covariance filtering.
> - 0.00–0.24: no meaningful metadata relation.
>
> For every HELPER, return exactly one directed assessment containing:
>
> - `target_task`
> - `helper_task`
> - `utility_score`
> - `label_effect_direction`
> - `relation_type`
> - `reason`
> - `evidence_fields`
>
> Use same, opposite, mixed, uncertain, or none for `label_effect_direction`. Use one of same_assay, same_target, same_readout, same_cell_context, same_assay_format, same_family, counter_screen, assay_artifact, topic_only, or unrelated for `relation_type`. The reason should briefly cite concrete metadata evidence and explain why the HELPER signal should or should not transfer to the TARGET. Do not output anything beyond the required structured assessments.

The system prompt was followed by a user message requesting exactly one assessment for every candidate helper:

> Score every candidate helper task for prediction utility toward the target task. Return exactly one edge per helper candidate. In every returned edge, `target_task` must be exactly the `target_task_id` string, and `helper_task` must be exactly one of the `candidate_helper_task_ids` strings.

The accompanying payload contained the target-task identifier and profile, the identifiers and profiles of all candidate helpers, and the required number of returned edges:

> `target_task_id`: target task identifier
> `target_task_metadata`: target task profile

`candidate_helper_task_ids`: identifiers of the other $K-1$ tasks
`candidate_helper_tasks`: profiles of the other $K-1$ tasks
`required_edge_count`: $K-1$

The pipeline therefore produces a structured and biologically grounded assessment of each directed target–helper relationship. The response contained the target and helper identifiers, a predictive-utility score, the expected label-effect direction, a categorical relation type, a short reason, and the task profile fields supporting the assessment. For example, when cpg0012 task `809086` was the target task, the language model returned the following assessment for candidate helper `809322`:

`target_task`: 809086
`helper_task`: 809322
`utility_score`: 0.93
`label_effect_direction`: same
`relation_type`: same_assay
`reason`: HCT-116 is a colon carcinoma NCI growth-inhibition assay with the same operational endpoint as SW-620. It is a strong same-tissue panel helper.
`evidence_fields`: description, summary, assay_cell_type, bao_format

This assessment defines a directed candidate relation `809322` $\rightarrow$ `809086`: activity in the HCT-116 colon carcinoma growth-inhibition assay was judged potentially informative for predicting activity in the SW-620 growth-inhibition assay. The resulting predictive-utility scores across all directed candidate relations were then used to construct the biological prior for guiding selective task-to-task information transfer.

Prior construction. Given the predictive-utility scores, we constructed a directed biological prior P. To retain the most biologically plausible target–helper relations, we selected a utility-score threshold δ to exclude weakly supported candidates and limited the number of helpers retained for each target task. Specifically, for each target task τ_k, candidate helpers satisfying $s_{k\leftarrow j} \geq \delta$ were ranked by their predictive-utility scores, and at most the top-R highest-scoring helpers were retained:

$$P_{k,j} = \begin{cases} 1, & \text{if } \tau_j \in \mathrm{Top}_R^{j\neq k}\left(s_{k\leftarrow j} \mid s_{k\leftarrow j} \geq \delta\right) \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

This enables selective information transfer among biologically relevant tasks. To further restrict information sharing to the most strongly supported task relations, we sparsified P following the strategy of BiGMINT (Pati et al., 2026). Using only the labels from the corresponding training split, we retained existing relations with label correlations above 0.1 and kept at most the R_{corr} helpers with the highest correlations for each target task. Validation and test labels were not used during this process. This promotes information sharing among the most relevant task pairs while mitigating noisy transfer from weakly supported connections.

We visualize the predictive-utility score matrix and the resulting biological prior for cpg0012 as an example (Fig. 4). The concentration of retained edges in high-scoring regions shows that the predictive-utility scores guide the selection of biologically relevant helper tasks for information transfer.

Prior-informed task interactions. For each sample, the task-conditioned feature-querying module produces the task representations $H = [h_1, \ldots, h_K] \in \mathbb{R}^{K\times d}$. For each target task τ_k, we collect the representations of its prior-selected helpers as

$$H_k^{\mathrm{prior}} = [h_j \mid j \neq k,\ P_{k,j} = 1]\,. \tag{19}$$

The prior P is fixed for all training data and determines which tasks may interact, whereas the attention weights are computed adaptively from the task representations of each input sample.

We apply a single multi-head cross-attention layer using h_k as the query and H_k^{prior} as the keys and values. For attention head m, the attention weights over the selected helpers are

$$\beta_k^{(m)} = \mathrm{softmax}\left(\frac{(h_k U_m^Q)(H_k^{\mathrm{prior}} U_m^K)^\top}{\sqrt{d_h}}\right). \tag{20}$$

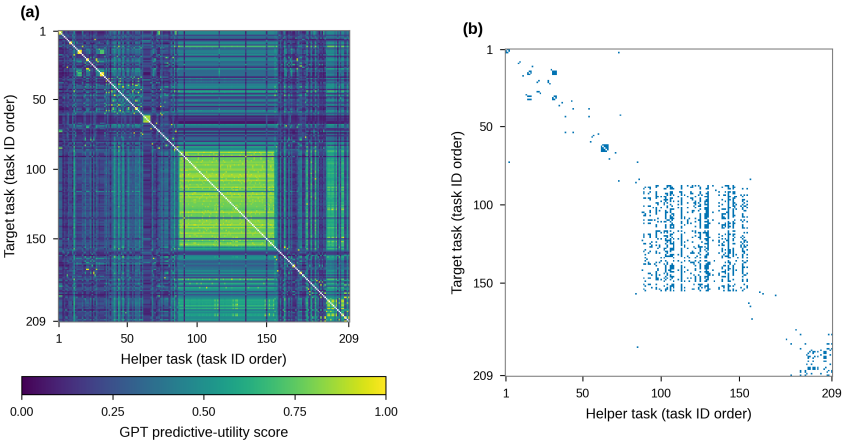

Figure 4: Visualization of task-prior construction on the cpg0012 benchmark. (a) Predictive-utility scores for all directed pairs of distinct tasks, with target tasks as rows and helper tasks as columns. (b) Final prior P obtained after processing. Blue entries denote retained task relations. All 209 tasks are ordered numerically by task identifier.

The corresponding helper summary is

$$u_k^{(m)} = \beta_k^{(m)} \left(H_k^{\text{prior}} U_m^V \right). \tag{21}$$

Here, $U_m^Q, U_m^K, U_m^V \in \mathbb{R}^{d \times d_h}$ are the interaction-specific projection matrices, and d_h is the dimensionality of each attention head. The vector $\beta_k^{(m)}$ contains one attention weight for each representation in H_k^{prior}.

The outputs of all attention heads are concatenated and projected:

$$u_k = \text{Concat}\left(u_k^{(1)}, \ldots, u_k^{(M)} \right) U^O, \tag{22}$$

where $U^O \in \mathbb{R}^{M d_h \times d}$ is the output projection. The helper information is then incorporated through a residual update:

$$\widetilde{h}_k = h_k + u_k. \tag{23}$$

Because the keys and values contain only the representations selected by P, tasks not included in H_k^{prior} cannot contribute to the target representation. If no helper is retained for τ_k, we set $\widetilde{h}_k = h_k$.

The prior-informed task-interaction module comprised $L_{\text{prior}} = 1$ cross-attention layer with $M = 8$ attention heads, representation dimension $d = 256$, head dimension $d_h = 32$, and dropout rate 0.3. Finally, the enhanced task representation was passed to the task-specific MLP classifier f_k:

$$\widehat{y}_k = \sigma\Big(f_k(\widetilde{h}_k) \Big). \tag{24}$$

C Experimental Setup

C.1 Datasets and Benchmarks

cpg0012 benchmark. The cpg0012 benchmark evaluates sparse multi-task bioactivity prediction from Cell Painting phenotypes. It was introduced in (Hofmarcher et al., 2019) by linking ChEMBL activity annotations to the Cell Painting dataset from (Bray et al., 2017). After matching compounds to ChEMBL and retaining assays with at least 10 active and 10 inactive compounds, the benchmark comprises 10,574 compounds and 209 binary assay endpoints. The resulting compound-by-endpoint label matrix has an observed-label coverage of 2.93%. Based on the released assay descriptions and metadata, the endpoints span diverse functional families, including reporter and pathway activity, DNA repair, RNA processing and splicing, receptor and signaling activity, viral entry, cell viability

and growth, and enzyme activity and molecular binding, alongside other assays. We use the released compound-by-endpoint label matrix and the three predefined compound-level splits. Each split contains 7,401 training, 1,058 validation, and 2,115 test compounds, corresponding to a 70%/10%/20% partition. Compounds are mutually exclusive across the three partitions, and all images and replicate wells associated with a compound remain in the same partition.

cpg0016 benchmark. The cpg0016 benchmark corresponds to the sparse bioactivity prediction task introduced in (Fredin Haslum et al., 2024). It combines Cell Painting data from the compound subset of the JUMP-CP dataset (Chandrasekaran et al., 2023) with binary activity labels derived from ChEMBL. The original study retained compounds from source 11, which provided the largest overlap with compounds having sufficient ChEMBL activity data, while excluding other sources to reduce variation in imaging settings. After retaining assays with at least 50 active and 50 inactive compounds, the published benchmark comprises 29 assay endpoints and 10,660 compounds. The resulting label matrix has an observed-label coverage of 7.91%. Based on the released ChEMBL assay metadata, the endpoints span biochemical target activity, gene regulation and pathway activity, cytotoxicity and genotoxicity, and anti-infective activity. We use the released preprocessing pipeline to generate six structure-cluster folds and evaluate six rotations, assigning four folds to training, one to validation, and one to testing in each rotation. This yields approximately 7,059 compounds for training and 1,764–1,765 compounds for each of validation and testing, while keeping structurally clustered compounds and all replicate profiles within a single partition.

cpg0037 benchmark. The cpg0037 benchmark evaluates whether Cell Painting phenotypes predict toxicity-related ToxCast activities. It is derived from the Cell Painting dataset by (Ewald et al., 2026), in which 1,085 compounds were profiled in primary human hepatocytes at eight concentrations with two replicates. The released study (Ewald et al., 2026) evaluated individual cytotoxicity, cell-based, and cell-free ToxCast endpoints using separate single-task models under compound-level five-fold cross-validation. To support joint multi-task learning, we retain compounds with available profiles, select the labels for the 354 officially evaluated ToxCast endpoints, and organize these labels into a common compound-by-endpoint matrix with an observed-label coverage of 34.75%. The resulting benchmark contains 10,139 well-level profiles and 354 endpoints, comprising 34 cytotoxicity, 267 cell-based, and 53 cell-free tasks. We construct five balanced compound-level folds shared by all endpoints. In each rotation, three folds are used for training, one for validation, and one for testing. All wells associated with the same compound are assigned to the same fold, preventing compound-level leakage.

C.2 Baselines

To benchmark CellMINT, we reproduce a diverse collection of unimodal and multimodal baselines across the three Cell Painting benchmarks. For all baselines that produce feature representations, we train an MLP classifier on top of the resulting representations. For end-to-end baselines, we retain their original architectures while adapting their output layers for multi-task assay prediction. All models are trained under the same sparse multi-task setting using identical data splits and available assay labels. Missing assay annotations are excluded from the training objective using masked binary cross-entropy.

CellProfiler (Stirling et al., 2021). CellProfiler is a classical image-analysis pipeline that characterizes cellular morphology using hand-crafted measurements extracted from segmented cells and subcellular compartments. These measurements capture properties such as object shape, size, intensity, texture, spatial organization, and compartment-level morphology, providing an interpretable representation of cellular phenotypes without requiring a learned image encoder.

DINOv2 (Oquab et al., 2023). DINO is a self-supervised vision foundation model that learns transferable visual representations from large-scale image collections without task-specific supervision. Its Vision Transformer architecture produces general-purpose image embeddings that can be transferred to downstream visual recognition tasks. We use DINO as a morphology-based representation method for extracting phenotypic features from Cell Painting images.

HTI-CNN (Hofmarcher et al., 2019). HTI-CNN is a supervised convolutional framework for learning biologically informative representations directly from high-throughput microscopy images. Unlike methods based on precomputed morphological profiles or frozen image features, HTI-CNN jointly learns image representations and downstream predictions from raw multi-channel cellular images, thereby serving as an end-to-end morphology-based baseline.

ECFP4 (Rogers & Hahn, 2010). ECFP4 is a circular molecular fingerprint that represents a compound through the local chemical environments surrounding its constituent atoms. Atom-centred neighborhoods are iteratively expanded to a fixed radius and mapped to a binary fingerprint, yielding a compact representation of molecular substructures. ECFP4 provides a widely used chemical-only baseline that does not require representation learning.

ChemBERTa (Chithrananda et al., 2020). ChemBERTa is a Transformer-based molecular representation model pretrained on large collections of SMILES strings using self-supervised language-modelling objectives. By treating molecular structures as token sequences, ChemBERTa learns contextual compound representations that capture recurring chemical patterns and structural relationships. It provides a learned chemical-only representation baseline complementary to fingerprint-based methods.

CLOOME (Sanchez-Fernandez et al., 2023). CLOOME is a cross-modal representation-learning framework that extends CLOOB-style contrastive learning to paired microscopy images and molecular structures. It uses modality-specific image and molecule encoders to map matched cellular phenotypes and chemical structures into a shared embedding space. The learned representations support cross-modal retrieval and capture associations between molecular structure and induced cellular morphology.

CellCLIP (Lu et al., 2025). CellCLIP is a text-guided contrastive learning framework for high-content Cell Painting data. It represents perturbations using natural-language descriptions and aligns the resulting text embeddings with microscopy image embeddings in a shared representation space. By incorporating language-based perturbation semantics during pretraining, CellCLIP aims to learn transferable image representations that connect cellular morphology with biologically meaningful perturbation information.

MICON (Yu et al., 2025). MICON is a multimodal representation-learning framework that incorporates molecular structures into the learning of microscopy representations. Rather than treating images and compounds solely as paired modalities, MICON models chemical compounds as treatments that induce transformations in cellular phenotypes. This treatment-aware formulation encourages the learned image representations to capture phenotypic changes associated with molecular perturbations.

InfoCORE (Wang et al., 2024). InfoCORE is an information-theoretic framework for learning drug representations while reducing the influence of experimental confounders. It maximizes conditional mutual information between molecular and phenotypic representations and adaptively reweights training samples to suppress associations driven by technical factors such as experimental batch variation. InfoCORE therefore provides a confounder-aware multimodal representation-learning baseline for large-scale Cell Painting data.

MINER (Rao et al., 2025). MINER is a multimodal learning framework for phenotypic drug discovery that combines unimodal denoising with calibrated cross-modal contrastive learning. It separates informative features from noise-related components within each modality and calibrates negative samples to reduce the influence of unreliable cross-modal pairs. For downstream prediction, MINER generates representations for missing modalities and employs a router-based Mixture-of-Experts module to dynamically fuse multimodal features and pass them to the final prediction head.

CP-Chem-MoA (Tian et al., 2023). CP-Chem-MoA is a multimodal framework that combines molecular structure with Cell Painting morphology for mechanism-of-action prediction. It separately encodes chemical and image-derived information and concatenates the resulting modality-specific representations before passing them through a shared prediction network. CP-Chem-MoA

therefore represents feature-level multimodal fusion, in which molecular and phenotypic information are integrated before the final prediction layer.

C.3 Evaluation Protocols

We follow the dataset-specific compound-level splits described above and compute all evaluation metrics at the compound level. For methods that produce well-level outputs, we first average the logits across all test wells associated with the same compound and then apply the sigmoid function to obtain one prediction per compound and endpoint. Such aggregation avoids treating multiple wells from the same labeled compound as independent test samples, which would disproportionately weight compounds with more available profiles. Additionally, evaluating predictions at the compound level reflects the intended bioactivity prediction setting, where the goal is to determine whether a compound is active in an assay by integrating evidence across its available Cell Painting profiles (Seal et al., 2025). This per-compound aggregation strategy is consistent with prior Cell Painting studies (Fredin Haslum et al., 2024; Pati et al., 2025).

During evaluation, missing assay labels are masked, and endpoints containing only one observed class in a given test split or fold are excluded from metric computation. We compute AUROC, AUPRC, and F1 separately for each endpoint using the aggregated compound-level predictions, with F1 calculated using a fixed probability threshold of 0.5. Within each split or fold, we report the unweighted mean across valid endpoints and then average these split- or fold-level metrics to obtain the final dataset-level results. High-performing task coverage is calculated from the split- or fold-mean endpoint AUROCs at thresholds of 0.7, 0.8, and 0.9.

C.4 Implementation Details

Hyperparameter settings. CellMINT used a hidden dimension of $d = 256$, $L = 3$ stacked task-conditioned feature-querying blocks, and $L_{\text{prior}} = 1$ cross-attention layer in the biological prior-informed task-interaction module. For biological-prior construction, we used a utility-score threshold of $\delta = 0.65$ to select biologically relevant helper tasks. To determine the most important tasks for information transfer, we set $R = 48$ and $R_{\text{corr}} = 16$ for cpg0012 and cpg0037. Because cpg0016 contains only 29 tasks, we used smaller values of $R = 12$ and $R_{\text{corr}} = 4$. The CLIP and CLUB loss weights were set to $\gamma_{\text{CLIP}} = 0.15$ and $\gamma_{\text{CLUB}} = 0.01$.

Training details. We trained CellMINT using AdamW for up to 30 epochs with a batch size of 128 and a learning rate of 2×10^{-4} for all three benchmarks. The learning rate was linearly increased to 2×10^{-4} during the first 5% of the planned optimization steps and subsequently decayed following a cosine schedule. We used a weight decay of 10^{-3} and gradient clipping with a maximum norm of 1.0. The checkpoint with the highest validation AUROC was selected, with early stopping applied after eight epochs without improvement. Experiments were conducted on four NVIDIA RTX 6000 Ada Generation GPUs with 48 GB of memory, with each split or fold trained on a single GPU. CellMINT was implemented using Python 3.11, PyTorch 2.9, and Transformers 4.57.

D Experimental Results

D.1 Benchmark Performance

Tables 4–6 provide the detailed results on the three Cell Painting benchmarks. Task-averaged AUROC, AUPRC, and Macro F1 are reported as the mean $\pm$ sample standard deviation across the predefined splits or folds, while AUROC coverage reports the number of assays whose split- or fold-mean AUROC reaches 0.7, 0.8, or 0.9. On cpg0012, CellMINT achieves the best task-averaged performance, with an AUROC of 0.783 ± 0.010, an AUPRC of 0.704 ± 0.019, and a Macro F1 of 0.568 ± 0.014. It also provides the highest coverage at the 0.7 and 0.8 thresholds and remains competitive at 0.9. On cpg0016, CellMINT similarly obtains the best task-averaged results, reaching 0.769 ± 0.018 AUROC, 0.736 ± 0.025 AUPRC, and 0.669 ± 0.030 Macro F1, together with the highest coverage at all three thresholds. On cpg0037, CellMINT achieves 0.681 ± 0.024 AUROC, 0.404 ± 0.016 AUPRC, and 0.139 ± 0.020 Macro F1, while substantially increasing the number of assays reaching the 0.7 and 0.8 AUROC thresholds. Across the three benchmarks, the gener-

Modality	Methods	Task-averaged performance			AUROC coverage		
		AUROC	AUPRC	Macro F1	≥ 0.7	≥ 0.8	≥ 0.9
Image	CellProfiler + Head	0.705 ± 0.022	0.616 ± 0.034	0.512 ± 0.017	111	81	27
	DINO + Head	0.713 ± 0.020	0.630 ± 0.041	0.505 ± 0.007	117	86	44
	HTI-CNN	0.718 ± 0.010	0.646 ± 0.012	0.489 ± 0.026	113	80	47
Chemical	ECFP4 + Head	0.662 ± 0.051	0.538 ± 0.051	0.299 ± 0.007	91	22	5
	ChemBERTa + Head	0.626 ± 0.019	0.511 ± 0.014	0.343 ± 0.035	71	23	2
Multimodal	CLOOME + Head	0.696 ± 0.009	0.624 ± 0.021	0.436 ± 0.067	97	74	56
	CellCLIP + Head	0.737 ± 0.016	0.685 ± 0.014	0.525 ± 0.024	117	90	**66**
	MICON + Head	0.704 ± 0.010	0.630 ± 0.032	0.475 ± 0.013	99	81	57
	InfoCORE + Head	0.720 ± 0.009	0.632 ± 0.025	0.470 ± 0.035	121	94	55
	MINER	0.668 ± 0.031	0.553 ± 0.057	0.372 ± 0.048	105	41	2
	CP-Chem-MoA	0.745 ± 0.014	0.684 ± 0.030	0.519 ± 0.023	134	98	56
	CellMINT	$\mathbf{0.783 \pm 0.010}$	$\mathbf{0.704 \pm 0.019}$	$\mathbf{0.568 \pm 0.014}$	**157**	**119**	65

Table 4: Performance comparison on the cpg0012 benchmark. AUROC, AUPRC, and Macro F1 are reported as the mean $\pm$ sample standard deviation across three predefined compound-level splits. Coverage columns report the number of tasks whose split-mean AUROC reaches each cutoff; tasks with undefined AUROC are not counted. Higher values are better, and boldface denotes the best available result in each column.

Modality	Methods	Task-averaged performance			AUROC coverage		
		AUROC	AUPRC	Macro F1	≥ 0.7	≥ 0.8	≥ 0.9
Image	CellProfiler + Head	0.663 ± 0.023	0.662 ± 0.025	0.600 ± 0.022	13	4	0
	DINO + Head	0.608 ± 0.020	0.616 ± 0.023	0.570 ± 0.021	5	1	0
	HTI-CNN	0.668 ± 0.029	0.662 ± 0.028	0.574 ± 0.041	12	3	0
Chemical	ECFP4 + Head	0.707 ± 0.014	0.686 ± 0.020	0.603 ± 0.017	12	5	0
	ChemBERTa + Head	0.671 ± 0.008	0.652 ± 0.024	0.580 ± 0.020	12	1	0
Multimodal	CLOOME + Head	0.591 ± 0.018	0.609 ± 0.023	0.546 ± 0.028	5	1	0
	CellCLIP + Head	0.650 ± 0.017	0.649 ± 0.025	0.585 ± 0.017	10	1	0
	MICON + Head	0.615 ± 0.018	0.616 ± 0.019	0.540 ± 0.009	6	0	0
	InfoCORE + Head	0.628 ± 0.027	0.626 ± 0.028	0.572 ± 0.016	8	1	0
	MINER	0.692 ± 0.018	0.671 ± 0.033	0.598 ± 0.027	14	4	0
	CP-Chem-MoA	0.742 ± 0.020	0.729 ± 0.020	0.610 ± 0.018	18	9	1
	CellMINT	$\mathbf{0.769 \pm 0.018}$	$\mathbf{0.736 \pm 0.025}$	$\mathbf{0.669 \pm 0.030}$	**19**	**15**	**2**

Table 5: Performance comparison on the cpg0016 benchmark. AUROC, AUPRC, and Macro F1 are reported as the mean $\pm$ sample standard deviation across the six predefined folds. Coverage columns report the number of tasks whose fold-mean AUROC reaches each cutoff; tasks with undefined AUROC are not counted. Higher values are better, and boldface denotes the best available result in each column.

ally modest standard deviations indicate consistent performance across data partitions. Overall, CellMINT improves both average predictive performance and high-performing task coverage across diverse assay collections.

D.2 Hyperparameter Sensitivity Analysis

We evaluated the sensitivity of CellMINT to key hyperparameters on cpg0012, following a one-factor-at-a-time protocol. Specifically, we examined the weights of the contrastive alignment and mutual-information regularization objectives, γ_{CLIP} and γ_{CLUB}, as well as the numbers of task-conditioned feature-querying blocks L and biological prior-informed cross-attention layers L_{prior}. For the loss weights, $0.5\times$ and $2\times$ denote scaling the default value by factors of 0.5 and 2. For depth hyperparameters L and L_{prior}, we considered values from 1 to 3. Results are reported as the compound-level mean $\pm$ standard deviation across three splits.

As shown in Table 7, CellMINT remains generally stable across the tested configurations. Scaling either loss weight produces only modest variations in AUROC, AUPRC, and Macro F1, indicating that the model does not rely on a narrowly tuned balance between the contrastive and disentangle-

Modality	Methods	Task-averaged performance			AUROC coverage		
		AUROC	AUPRC	Macro F1	≥ 0.7	≥ 0.8	≥ 0.9
Image	CellProfiler + Head	0.600 ± 0.034	0.365 ± 0.039	0.083 ± 0.016	83	14	2
	DINO + Head	0.604 ± 0.024	0.369 ± 0.028	0.080 ± 0.013	90	14	4
	HTI-CNN	0.618 ± 0.024	0.367 ± 0.016	0.059 ± 0.005	100	24	3
Chemical	ECFP4 + Head	0.600 ± 0.049	0.351 ± 0.031	0.117 ± 0.033	85	15	5
	ChemBERTa + Head	0.646 ± 0.031	0.360 ± 0.037	0.102 ± 0.014	133	26	6
Multimodal	CLOOME + Head	0.593 ± 0.033	0.361 ± 0.033	0.072 ± 0.007	83	18	6
	CellCLIP + Head	0.637 ± 0.030	0.373 ± 0.033	0.073 ± 0.010	132	33	1
	MICON + Head	0.577 ± 0.022	0.344 ± 0.027	0.063 ± 0.009	60	13	1
	InfoCORE + Head	0.632 ± 0.022	0.372 ± 0.040	0.058 ± 0.006	107	32	**7**
	MINER	0.629 ± 0.024	0.351 ± 0.030	0.049 ± 0.008	105	24	1
	CP-Chem-MoA	0.618 ± 0.049	0.373 ± 0.049	0.093 ± 0.023	86	16	1
	CellMINT	$\mathbf{0.681 \pm 0.024}$	$\mathbf{0.404 \pm 0.016}$	$\mathbf{0.139 \pm 0.020}$	**192**	**51**	5

Table 6: Performance comparison on the cpg0037 benchmark. AUROC, AUPRC, and Macro F1 are reported as the mean $\pm$ sample standard deviation across the five predefined folds. Coverage columns report the number of tasks whose fold-mean AUROC reaches each cutoff. Tasks with undefined AUROC are not counted. Higher values are better, and boldface denotes the best available result in each column.

Hyperparameter	Setting	AUROC	AUPRC	Macro F1
γ_{CLIP}	0.075 (0.5×)	0.782 ± 0.011	0.697 ± 0.036	0.560 ± 0.014
	0.30 (2×)	0.775 ± 0.013	0.709 ± 0.010	0.547 ± 0.025
γ_{CLUB}	0.005 (0.5×)	0.782 ± 0.013	0.702 ± 0.023	0.561 ± 0.015
	0.02 (2×)	0.768 ± 0.027	0.690 ± 0.038	0.535 ± 0.075
L	1	0.751 ± 0.007	0.699 ± 0.018	0.510 ± 0.037
	2	0.778 ± 0.010	0.711 ± 0.007	0.553 ± 0.026
L_{prior}	2	0.777 ± 0.013	0.706 ± 0.007	0.568 ± 0.019
	3	0.780 ± 0.010	0.715 ± 0.008	0.560 ± 0.020
Default	–	0.783 ± 0.010	0.704 ± 0.019	0.568 ± 0.014

Table 7: Hyperparameter sensitivity of CellMINT on cpg0012 (209 tasks). Each non-default row changes only the indicated hyperparameter. Results are reported as the compound-level mean $\pm$ standard deviation across three splits.

ment objectives. Reducing L leads to a more noticeable performance drop, supporting the use of multiple querying blocks to progressively extract task-relevant multimodal information. Increasing L_{prior} provides no consistent improvement, suggesting that a single prior-informed cross-attention layer is sufficient to facilitate information transfer among biologically related tasks. Overall, the default configuration performs strongly across the three metrics and is used consistently in all main experiments.

D.3 Predictive-Utility Inference Across Language Models

To assess whether CellMINT's biological task-relation reasoning is reproducible across pretrained language models, we repeated our biological reasoning pipeline on the cpg0012 benchmark using Claude Opus 5 (Anthropic, 2026). For all 209 assay endpoints, GPT-5.5 and Claude Opus 5 used the same structured task profiles, prompt, and predictive-utility scoring rubric. Figure 5 compares the predictive-utility matrices produced by the two models and shows highly similar global patterns and block structures across directed task relations. Across 43,472 directed non-self task pairs, the inferred scores achieved a Pearson correlation of $r = 0.925$. The result suggests that CellMINT's biological reasoning pipeline generates reproducible task-relation structures across different language-model families and consistently identifies biologically meaningful relationships encoded in task profiles.

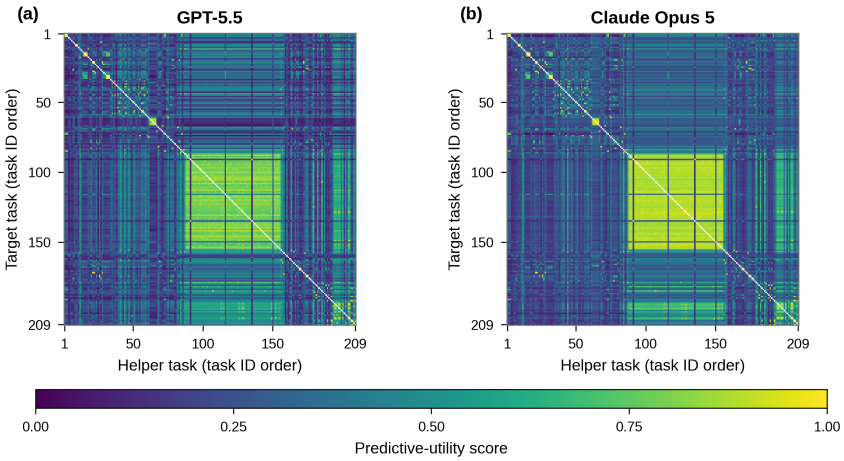

Figure 5: Predictive-utility score matrices generated by the biological reasoning pipeline for the cpg0012 benchmark using (a) GPT-5.5 and (b) Claude Opus 5. Rows denote target tasks and columns denote candidate helper tasks among the 209 assay endpoints. Tasks are ordered numerically by task identifier.

E Discussion

CellMINT provides a unified framework for modeling two important forms of heterogeneity in biological assay prediction: different endpoints may rely on different combinations of phenotype-specific, modality-shared, and chemistry-specific evidence, and may benefit from information provided by biologically related tasks. By combining control-guided multimodal disentanglement, task-conditioned feature querying, and biologically informed task interaction, CellMINT consistently improves both average predictive performance and high-performing task coverage across three heterogeneous Cell Painting benchmarks.

Nevertheless, several limitations remain and suggest directions for future development. In particular, constructing the task-interaction prior depends on the availability of assay metadata, making biologically relevant helper tasks harder to identify when metadata are incomplete or uninformative. Moreover, the benefit of task interaction may vary across endpoints, with biologically well-connected tasks receiving greater gains from knowledge transfer than relatively isolated tasks. While these isolated tasks may still benefit from task-adaptive integration of multimodal evidence, they have fewer opportunities to exploit complementary information from related tasks. From a practical perspective, the current framework also assumes that both Cell Painting and chemical modalities are available for each sample, motivating extensions that can operate under missing-modality settings. Furthermore, the present implementation considers only two modalities, while incorporating additional sources, such as transcriptomic profiles or experimental-context information, may provide a more comprehensive characterization of compound perturbations. Addressing these directions may further improve the flexibility and applicability of CellMINT across broader multimodal and multi-task assay prediction settings.