MoMIL: Mixture of Multi-Instance Learners for Modeling Multiple Compound Activities in High Content Imaging

Pushpak Pati[1]*, Hsiu-Chi Cheng[2], Steffen Jaensch[1], Walid M. Abdelmoula[1], Krishna Chaitanya[1], Michiel Van Dyck[1], Tomé Albuquerque[1], Samantha Allen[1], Litao Zhang[1], Tommaso Mansi[1], Rui Liao[1], Zhoubing Xu[1]

[1] Janssen R&D, LLC, a Johnson & Johnson Company
[2] ETH Zurich, Switzerland
* ppati4@its.jnj.com

Abstract. High content imaging (HCI) plays a pivotal role in target-directed drug discovery (TDD) by identifying compound activities across tests (or assays) designed for specific therapeutic targets. However, real-world assays often exhibit extreme label sparsity over large compound libraries, making accurate predictions challenging. Recent studies following multi-label learning (MLL) struggle in such scenarios when optimizing a single objective across multiple assays without assay-specific adaptations. To address this, we propose Mixture of Multi-Instance Learners (MoMIL), a multi-task learning (MTL) framework integrating hard-parameter sharing with assay-specific Multiple Instance Learners (MILs), enabling knowledge sharing and task-specific adaptations. Furthermore, we introduce complementary enhancements: HCI-specific foundation models (FMs), an assay selection algorithm, and a label imputation method to boost MoMIL's learning capabilities. We benchmark MoMIL on two extensive HCI datasets, achieving up to ∼6% and ∼8% improvement over state-of-the-art MLL and MTL methods. Moreover, MoMIL shows strong generalization to unseen assays, outperforming assay-specific single-task learning (STL) methods in 11 out of 12 assays.

Keywords: High content imaging · Assay modeling · Multi-task learning · Multiple Instance Learning · Assay selection · Label imputation

1 Introduction

In TDD, assays are conducted to measure how different compounds affect *specific* therapeutic targets, like proteins involved in a disease [14]. HCI is a technique that observes how cells react to these compounds in a *target-agnostic* way [10]. The morphologies from HCI is then used to predict how compounds will interact with *specific* targets, measured via respective assays [9,13,22]. By using HCI to study many compounds across different assays, TDD aims to find important interactions between compounds and targets. However, figuring out these interactions accurately is difficult because of the complex biological effects involved.

Recent advancements in AI/ML have significantly enhanced HCI-based drug discovery [15–17, 21, 23] by automating morphological feature extraction, hit discovery, and identifying mechanisms of action. To predict multiple assays for a compound, recent efforts [9, 13, 22] used MLPs, CNNs, and MLL methods, pre-trained on ImageNet [5]. However, these methods rely on multiple-concentration-response data and multiple measurements per compound [13,22], which are often unavailable at scale during early hit identification. Although single-concentration data have been explored in [9], it assumes a high label density (∼48%). However, real-world assays often exhibit extreme label sparsity, as low as 2-5% [22]. This scarcity arises from cost, time and resource limitations of testing several compounds over various assays. In addition, traditional MLL approaches, optimizing a single objective across all labels, struggle under extreme sparsity, leading to suboptimal performance. This approach does not adequately address diverse patterns and distinct learning strategies necessary for individual assay. Therefore, there is a pressing need to improve TDD by developing robust models that can jointly analyze multiple assays, even when faced with extreme label sparsity.

In this paper, we present **MoMIL**, a MTL framework specifically designed to overcome the challenges of extreme label sparsity in predicting multiple assays from single-concentration data. MoMIL employs a projection backbone with hard-parameter sharing combined with assay-specific Multiple Instance Learning (MIL) modules, effectively enabling knowledge sharing while tailoring models to individual assay, and overcoming the issues of MLL setups. Further, our integration of attention mechanisms within the MILs allows to capture assay-specific heterogeneity from HCI, improving upon standard feature aggregation methods like mean or median pooling [4, 7]. To enhance MTL performance, we implement three key enhancements: (1) enhanced feature extraction through pre-training **FMs on HCI** datasets, outperforming standard ImageNet-based models [16]; (2) an **assay selection algorithm** that identifies relevant auxiliary assays to streamline knowledge transfer in MTL while minimizing noise, redundancy, and model complexity; and (3) an adaptive assay-wise **label imputation method** that boosts model reliability by providing additional high-confidence signals.

We rigorously benchmark MoMIL framework on two extensive HCI datasets from Cell Painting [11] using U2OS and iPSC-derived neurons (iNeurons), focusing on two distinct sets of six assays for TDD. These datasets include an order of magnitude more compounds than those previously studied [7, 9, 13], allowing for comprehensive evaluation across multiple experimental batches. MoMIL achieves notable average **improvements of 9.7%, 6.1%, and 8.1%** compared to traditional [15], state-of-the-art MLL [9], and adapted MTL [26] methods, respectively. Additionally, our results highlight MoMIL's **robust generalization** capabilities, as it outperforms assay-specific STL methods in 11 out of 12 unseen assays, demonstrating its potential to enhance real-world assay modeling.

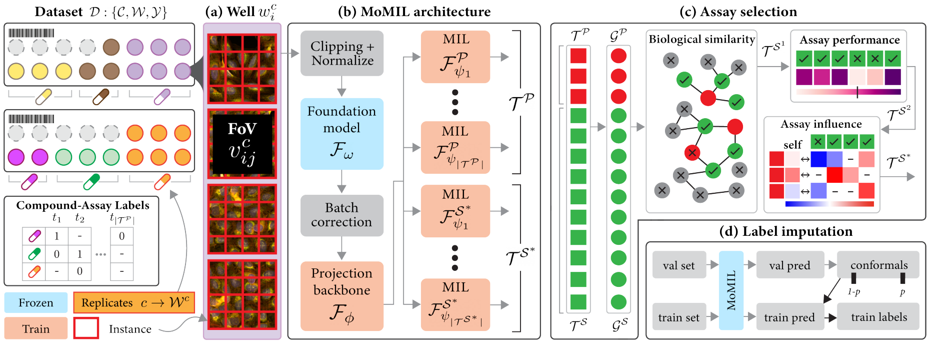
Dataset $\mathcal{D}:\{\mathcal{C},\mathcal{W},\mathcal{Y}\}$
(a) Well w_i^c
(b) MoMIL architecture
(c) Assay selection
Compound-Assay Labels
t_1
t_2
$t_{|\mathcal{T}^\mathcal{P}|}$
1
-
0
0
1
...
-
-
0
-
Frozen
Replicates $c \rightarrow \mathcal{W}^c$
Train
Instance
FoV v_{ij}^c
Clipping + Normalize
Foundation model $\mathcal{F}_\omega$
Batch correction
Projection backbone $\mathcal{F}_\phi$
MIL $\mathcal{F}^\mathcal{P}_{\psi_1}$
MIL $\mathcal{F}^\mathcal{P}_{\psi_{|\mathcal{T}^\mathcal{P}|}}$
MIL $\mathcal{F}^{\mathcal{S}^*}_{\psi_1}$
MIL $\mathcal{F}^{\mathcal{S}^*}_{\psi_{|\mathcal{T}^{\mathcal{S}^*}|}}$
$\mathcal{T}^\mathcal{P}$
$\mathcal{T}^{\mathcal{S}^*}$
$\mathcal{T}^\mathcal{P}$
$\mathcal{G}^\mathcal{P}$
$\mathcal{T}^\mathcal{S}$
$\mathcal{G}^\mathcal{S}$
Biological similarity
$\mathcal{T}^{\mathcal{S}^1}$
Assay performance
$\mathcal{T}^{\mathcal{S}^2}$
Assay influence
self
$\mathcal{T}^{\mathcal{S}^*}$
(d) Label imputation
val set
MoMIL
val pred
conformals
1-p
p
train set
train pred
train labels

Fig. 1: Overview of MoMIL framework. HCI input (a) is fed to multi-task MoMIL model (b) to predict binary activities for the primary assays. (c) and (d) are the assay selection and label imputation algorithms that boost knowledge sharing in MoMIL by selecting relevant secondary assays and augmenting reliable activities.

2 Method

This section presents the MoMIL framework, as shown in Fig.1, which promotes effective knowledge transfer across mixture of MILs for predicting multiple assays in HCI via, (1) MoMIL architecture; (2) an assay selection algorithm based on biological and assay relevance; and (3) an adaptive label imputation algorithm.

2.1 MoMIL architecture

Let dataset $\mathcal{D}$ consists of compounds $\mathcal{C}$ and wells $\mathcal{W}$. Let $c \in \mathcal{C}$ perturbs a set of wells, also called replicates, $\mathcal{W}^c : \{w_i^c\}$. The goal is to use $\mathcal{W}^c$ to predict binary activities $\mathcal{Y}_t^c$ for c across $t \in \mathcal{T}$ assays. To note, $\mathcal{T}$ includes $\mathcal{T}^\mathcal{P}$ primary and $\mathcal{T}^\mathcal{S}$ secondary assays. We aim to maximize the prediction of $\mathcal{T}^\mathcal{P}$ by using relevant knowledge from $\mathcal{T}^\mathcal{S}$. MoMIL addresses the goal as a MTL problem (Fig.1(a)) and predicts $y_{it}^c, \forall t \in \mathcal{T}$ for w_i^c. Final $\mathcal{Y}_t^c$ is derived as the mean of activities $\{y_{it}^c\}$ across all replicates $\{w_i^c\}$. Formally, MoMIL includes: a feature extractor $\mathcal{F}_\omega$, a shared projector $\mathcal{F}_\phi$, and a set of assay-specific heads $\mathcal{F}_\Psi : \{\mathcal{F}_{\psi_t}\}, \forall t$.

Feature extractor $\mathcal{F}_\omega$: This module featurizes w_i^c into $\mathbf{w}_i^c$. w_i^c is imaged as a set of field-of-views (FoVs) $\{v_{ij}^c\}$. First, v_{ij}^c is divided into instances of shape $\mathbb{R}^{p \times q \times ch}$, with p, q as the spatial dimensions and ch as the image channels. Next, $\mathcal{F}_\omega$ encodes each instance into $\mathbb{R}^D$, followed by batch correction. $\mathcal{F}_\omega$ is a Vision Transformer (ViT) pretrained using self-supervised learning and instance images from an exclusive dataset $\mathcal{D}^*$, s.t., $\mathcal{D} \cap \mathcal{D}^* = \varnothing$. Note, $\mathcal{D}^*$ and $\mathcal{D}$ contain compounds with scaffold-level overlap, as extreme annotation sparsity prevents feasible scaffold-level splitting while ensuring adequate class-wise annotations across assays. Finally, $\mathbf{w}_i^c : \mathcal{F}_\omega(w_i^c) \in \mathbb{R}^{N_i^c \times D}$, where N_i^c is the total number of instances in w_i^c across all $\{v_{ij}^c\}$ FoVs.

Projection backbone $\mathcal{F}_\phi$: This module extracts generalized representations of compound-assay interactions through hard-parameter sharing, *i.e.*, sharing ϕ

across $\mathcal{T}$. Hard-parameter sharing in MoMIL facilitates three benefits: (1) enables to learn robust assay-agnostic patterns by leveraging the biological relatedness among the assay-corresponding targets, detailed in Sec.2.2; (2) stronger regularization, which is crucial for reducing the risk of overfitting under sparse labels; (3) compared to soft-parameter sharing approaches, it incurs less learning complexity, computational cost, and risk of overfitting with limited labels. Formally, $\mathcal{F}_\phi(\mathbf{w}_i^c) : \mathbb{R}^{N_i^c \times D} \rightarrow \mathbb{R}^{N_i^c \times d}$, where $\mathcal{F}_\phi$ is an MLP and $d << D$.

Assay-specific $\mathcal{F}_\Psi$: This module includes assay-specific heads $\{\mathcal{F}_{\psi_t}\}, \forall t \in \mathcal{T}$ to optimize individual compound-assay interaction. $\mathcal{F}_{\psi_t}(\mathcal{F}_\phi(\mathbf{w}_i^c)) : \mathbb{R}^{N_i^c \times d} \rightarrow \mathbb{R}^1$ optimally aggregates the encoded instance features in w_i^c to predict y_{it}^c. $\mathcal{F}_{\psi_t}$ is a MIL, that learns permutation-invariant instance-level attention weights, uses the attention-scaled instance features to derive a well representation, and maps the well representation to the assay label. Formally, $\mathcal{F}_{\psi_t}$ is defined as,

$$\mathcal{F}_{\psi_t}\big(\mathcal{F}_\phi(\mathbf{w}_i^c)\big) = \sigma\left(\sum_{i=1}^{N_i^c} \mathcal{F}_{\alpha_t}\Big\{\mathcal{F}_{\beta_t}\big(\mathcal{F}_{\gamma_t}(\mathcal{F}_\phi(\mathbf{w}_i^c))\big) \times \mathcal{F}_{\gamma_t}(\mathcal{F}_\phi(\mathbf{w}_i^c))\Big\}\right)$$

where, $\mathcal{F}_{\gamma_t}, \mathcal{F}_{\beta_t}$, and $\mathcal{F}_{\alpha_t}$ are MLP projector, attention module with softmax activation, and MLP classifier, respectively, and σ is the sigmoid activation.

Optimization objective: $\mathcal{F}_\phi$ and $\mathcal{F}_\Psi$ are trained in an end-to-end manner by optimizing well-level multi-task binary cross-entropy loss, given as,

$$\mathcal{L} = -\sum_{c=1}^{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{W}|}\sum_{t\in\mathcal{T}} y_{it}^c \times \log\Big(\mathcal{F}_{\psi_t}\big(\mathcal{F}_\phi(\mathbf{w}_i^c)\big)\Big) + (1-y_{it}^c) \times \log\Big(1-\mathcal{F}_{\psi_t}\big(\mathcal{F}_\phi(\mathbf{w}_i^c)\big)\Big)$$

2.2 Assay selection algorithm

To enhance the prediction of $\mathcal{T}^{\mathcal{P}}$ leveraging knowledge from $\mathcal{T}^{\mathcal{S}}$, we propose a selection algorithm that identifies a subset $\mathcal{T}^{\mathcal{S}^*} \subset \mathcal{T}^{\mathcal{S}}$ (Fig. 1(c)). It minimizes negative knowledge transfer from irrelevant assays while improving representation learning, model complexity, generalization, and addressing label scarcity. Assays are prioritized based on: (1) **Biological similarity**: assays from $\mathcal{T}^{\mathcal{S}}$ sharing pathways or interactions with $\mathcal{T}^{\mathcal{P}}$ to provide relevant biological knowledge, (2) **Assay performance**: assays with high uni-assay performance to provide high-quality features, (3) **Assay influence**: assays positively influencing $\mathcal{T}^{\mathcal{P}}$ performance. These criteria are applied hierarchically for computation efficiency.

Biological similarity between two assays $t^p \in \mathcal{T}^{\mathcal{P}}$ and $t^s \in \mathcal{T}^{\mathcal{S}}$ is measured by the association between the corresponding targets $g^p \in \mathcal{G}^{\mathcal{P}}$ and $g^s \in \mathcal{G}^{\mathcal{S}}$. The association between g^p and g^s is derived from STRING [24] by incorporating phylogenetic co-occurrence, homology co-expression, experimentally determined interactions, and text mining. The selected assays $\forall t^p$ form $\mathcal{T}^{\mathcal{S}^1} \subset \mathcal{T}^{\mathcal{S}}$.

Next, assay performances are measured by training individual MIL models $\forall t \in \mathcal{T}^{\mathcal{S}^1}$. Specifically, we use the $\mathcal{F}_\omega$ and $\mathcal{F}_{\psi_t}$ from MoMIL. We exclude $\mathcal{F}_\phi$ and use $\mathcal{F}_{\gamma_t}$ inside $\mathcal{F}_{\psi_t}$ to project $\mathbf{w}_i^c$. A subset $\mathcal{T}^{\mathcal{S}^2} \subset \mathcal{T}^{\mathcal{S}^1}$ is selected by applying a threshold $\mathbf{th_{perf}}$ on the uni-assay validation-set performances.

Influence of an assay $t \in \mathcal{T}^{\mathcal{S}^2}$ on $t^p \in \mathcal{T}^{\mathcal{P}}$ is assessed via the transfer learning performance of uni-assay $\mathcal{F}_{\psi_t}$ on t^p. First, we use $\mathcal{F}_\omega$ and the trained $\mathcal{F}_{\psi_t}$ from above to infer the well embeddings for t^p's train and validation sets. Then we train a Logistic Regression model to compute the influence score as the predictive performance on the validation set. Influence scores are computed for all the associated assays in $\mathcal{T}^{\mathcal{S}^2}$ on t^p, and assays exceeding a threshold ($\mathbf{th_{inf}}$% of t^p's self-influence) are selected. $\mathcal{T}^{\mathcal{S}^*}$ is the collection of the influential assays $\forall t^p$.

Our algorithm is flexible and scalable to large sets of secondary assays, computationally efficient, and facilitates interpretability in selection. Additionally, compared to reinforcement learning [25, 27] and meta-learning methods [8], it is more label-efficient, less prone to overfitting, and simpler to implement.

2.3 Adaptive label imputation algorithm

Given the extreme label sparsity, we propose a conformal-based multi-label imputation algorithm to transfer knowledge across assay labels for $\mathcal{T}^* : \mathcal{T}^{\mathcal{P}} \cup \mathcal{T}^{\mathcal{S}^*}$. We leverage conformal prediction [1] to estimate empirical confidence thresholds ("conformals") on the validation set, and apply them on the train set to effectively expand training labels. First, we train the initial MoMIL model, compute probabilities $\mathcal{P}_t^{\text{val}}, \forall t \in \mathcal{T}^*$ on the validation set, and identify correctly classified wells $\mathcal{W}_t^{\text{val}}$. Next, for each assay $t \in \mathcal{T}^*$, we determine labeling thresholds $\mathbf{th}_t^{\text{pos}}$ and $\mathbf{th}_t^{\text{neg}}$ as the $\mathbf{p}^{\text{th}}$ and $(1-\mathbf{p})^{\text{th}}$ percentiles on $\mathcal{P}_t^{\text{val}}$ over $\mathcal{W}_t^{\text{val}}$. Unlabeled training wells with probabilities exceeding $\mathbf{th}_t^{\text{pos}}$ or below $\mathbf{th}_t^{\text{neg}}$ are assigned positive and negative labels, respectively. To prevent $\mathcal{T}^{\mathcal{S}^*}$ from dominating $\mathcal{T}^{\mathcal{P}}$, we enforce stricter thresholds as the $(\mathbf{p}+\alpha)^{\text{th}}$ and $(1-\mathbf{p}-\alpha)^{\text{th}}$ percentiles for $\mathcal{T}^{\mathcal{S}^*}$. The imputed labels are then incorporated into subsequent MoMIL training iterations, iteratively refining model predictions through self-training.

The algorithm offers a statistically grounded method for imputing labels, ensuring confident label assignment and minimizing noise propagation. Further, the assay-specific adaptive thresholds promote balanced learning across assays.

3 Experiments

Datasets: We evaluated the MoMIL framework on two large in-house datasets acquired using Cell Painting on U2OS and iNeuron cell lines at 10μM and 20μM concentrations, respectively, and 24-hour incubation. Different cell components were labeled using fluorescent dyes and acquired 16-bit 5-channel fluorescence images with a Yokogawa CellVoyager 8000 confocal HCI reader at 20× magnification [12]. Both datasets were split at compound-level to define $\mathcal{D}^*$ for training FMs and $\mathcal{D}$ for evaluating MoMIL, presented in Tab.1. Both U2OS and iNeuron datasets contain over an order of magnitude more compounds for assay prediction compared to recent studies [7, 9, 13], enabling robust analysis. The datasets also contain multiple experimental batches, enabling robust assessment of batch variations in HCI and ensuring reliable, reproducible assay modeling [2].

Dataset	Usage	#compounds	#batches	#plates	#wells	#FoVs	#assays
U2OS	FM train	69,019	338	505	123,907	495,531	-
iNeuron	FM train	73,101	48	308	95,908	862,757	-
U2OS	MoMIL eval	41,843	314	477	84,030	336,075	200
iNeuron	MoMIL eval	42,808	48	627	61,239	550,907	200

Table 1: HCI dataset statistics for pre-training FMs and evaluating MoMIL.

FM training: U2OS and iNeuron FoVs of sizes 970×970 and 1938×1938 were resized to 960×960, channel-wise intensity clipped at < 0.01 and > 99.9 percentiles, and min-max normalized. Then, we trained ViT-B/16 [6] model with DINOv2 [3] and DINO [20] for U2OS and iNeuron, respectively. Empirically, we observed inferior representation quality with DINOv2 for iNeuron, likely due to the inadequacy in reconstructing fine neurite structures. The FMs were trained for 400 epochs with 480×480 global and 128×128 local crop sizes. For effective knowledge distillation, crops were selected from high-intensity nuclei areas, and augmented using flips, rotations, blur and color jitter. For inference, FoVs were divided into 480×480 instances with stride 240×240, featurized by the FMs, and plate-wise batch corrected using robust z-scoring over DMSO statistics.

MoMIL evaluation: $\mathcal{D}$ was split into 5-folds at compound-level based on chemical similarities [12] to evaluate model generalization to compounds unseen during training. Compounds-to-assays label matrix had a sparse fill rate of **4.3%** in U2OS and **4.2%** in iNeuron, with 18.3% and 17.1% for $\mathcal{T}^{\mathcal{P}}$. Assay labels were assigned to well-replicates for model training. $\mathcal{F}_{\psi_t}$ includes DSMIL [18] following its success in [7]. Model hyperparameters include, $\mathcal{F}_{\phi}$ hidden-dim $\{64, 128\}$, $\mathcal{F}_{\psi_t}$ hidden-dim $\{32, 64, 128\}$ with 0.5 dropout and ReLU. MoMIL was trained for 100 epochs with a warm-up of 20 epochs, AdamW optimizer with learning rates $\{1e^{-4}, 5e^{-5}, 1e^{-5}\}$ & $5e^{-4}$ weight decay, cosine annealing, & early-stopping with patience of 10 epochs. $\mathbf{th_{perf}}$, $\mathbf{th_{inf}}$, $\mathbf{p}$ and α were set to 70%, 0.9, 90% and 5%. Well-level mean ROC-AUC over $\mathcal{T}^{\mathcal{P}}$ on the val-set was used for model selection. We followed a 3-1-1 train-val-test strategy by permuting the test-set over the 5-folds. Finally, the well predictions over $\mathcal{T}^{\mathcal{P}}$ on the test-sets were combined and averaged across well-replicates to derive compound-level assay estimates.

Baselines: We benchmarked MoMIL against four relevant baselines selected per recent studies and their scalability to large number of assays. The baselines were optimized similarly to MoMIL in the evaluation setup.

•**Acapella → MLP** [12]: The most widely used method in HCI [21]. Cells were segmented, featurized by Acapella (commercial version of CellProfiler [19]), and mean-pooled over the well to get well features, and followed by MLP classifiers.

• **FM → Mean → MLP** [15]: Instance features per-FoV were extracted using a FM, mean-pooled per well, and followed by MLP classifiers.

• **FM → Multi-label-MIL** [9]: We adapted the original MLL method by replacing transfer learning on ResNet50 with FM → MIL to ensure a fair comparison. The MIL classifier was adapted for multi-label predictions and optimized per [9].

		Method	$\lvert\mathcal{T}\rvert$	$\mathbf{t}_1^U$	$\mathbf{t}_2^U$	$\mathbf{t}_3^U$	$\mathbf{t}_4^U$	$\mathbf{t}_5^U$	$\mathbf{t}_6^U$	**Average**
U2OS	**STL**	Acapella → MLP [12]	1	60.8	62.3	62.7	61.2	70.2	65.8	63.8
		FM → Mean → MLP [15]	1	61.7	61.5	68.6	62.6	71.0	67.1	65.4
		MoMIL	1	65.4	62.6	65.5	67.3	74.0	71.6	67.7
	MTL	Acapella → MLP [12]	6	64.0	61.1	59.9	59.7	73.7	69.7	64.7
		FM → Mean → MLP [15]	6	61.5	58.8	62.0	67.0	74.0	71.4	65.8
		FM → Multi-label-MIL [9]	6	68.1	66.8	64.8	71.1	71.2	71.8	68.9
		MoMIL	6	65.8	65.1	67.5	69.8	72.2	70.6	68.5
		FM → MIL + AMTL [26]	200	65.2	61.4	63.0	70.8	75.2	71.3	67.6
		FM → Multi-label-MIL [9]	200	67.8	66.3	66.2	71.7	74.0	72.2	69.7
		MoMIL	200	67.6	65.8	69.0	72.7	77.2	73.4	70.9
		MoMIL + AS (bio. sim.)	78	68.7	67.2	69.7	72.0	77.8	74.3	71.6
		MoMIL + AS (assay perf.)	53	67.9	66.8	69.4	71.4	76.0	74.4	71.0
		MoMIL + AS	28	67.7	67.4	71.7	76.3	78.0	75.2	72.7
		MoMIL + AS + LI	28	$68.3^{\dagger}$	69.1^{*}	74.6^{*}	76.4^{*}	$77.4^{\dagger}$	75.7^{*}	73.6
		Method	$\lvert\mathcal{T}\rvert$	$\mathbf{t}_1^N$	$\mathbf{t}_2^N$	$\mathbf{t}_3^N$	$\mathbf{t}_4^N$	$\mathbf{t}_5^N$	$\mathbf{t}_6^N$	**Average**
iNeuron	**STL**	Acapella → MLP [12]	1	71.7	59.5	62.5	69.3	57.3	67.9	64.7
		FM → Mean → MLP [15]	1	69.1	63.2	54.4	69.9	62.9	68.4	64.6
		MoMIL	1	71.4	65.8	58.2	71.7	64.6	68.3	66.7
	MTL	Acapella → MLP [12]	6	69.7	61.5	58.5	67.8	64.0	69.1	65.1
		FM → Mean → MLP [15]	6	71.4	64.5	58.2	69.4	61.7	68.1	65.6
		FM → Multi-label-MIL [9]	6	66.4	63.8	59.9	66.4	56.5	68.6	63.6
		MoMIL	6	67.1	64.6	60.3	73.3	65.7	68.1	66.5
		FM → MIL + AMTL [26]	200	68.9	64.2	58.8	70.8	63.1	69.1	65.8
		FM → Multi-label-MIL [9]	200	64.3	66.0	60.1	71.7	63.6	70.7	66.1
		MoMIL	200	70.8	67.7	60.8	71.9	64.6	71.7	67.9
		MoMIL + AS (bio. sim.)	89	70.1	65.7	62.5	72.6	64.5	73.0	68.1
		MoMIL + AS (assay perf.)	52	71.2	66.2	62.3	72.7	62.2	71.7	67.8
		MoMIL + AS	32	72.1	67.4	62.8	74.6	65.9	72.5	69.2
		MoMIL + AS + LI	32	71.8^{*}	$69.2^{\dagger}$	$68.2^{\dagger}$	$75.6^{\ddagger}$	$65.5^{\dagger}$	72.9^{*}	70.5

Table 2: Benchmarking average and per-assay AUC of STL and MTL across U2OS and iNeuron. AS: assay selection. LI: label imputation. Our framework & its ablations are in **bold** and underline. 1st & 2nd best AUCs in yellow and blue. *, †, ‡ denote p-values < 0.001, < 0.05 & < 0.1 from a paired one-sided bootstrap test between MoMIL + AS + LI and FM → Multi-label-MIL at assay-level.

- **FM → MIL + AMTL** [26]: The method optimized achievement-based multi-task (AMTL) loss to modulate training speed, where achievement per-assay was defined as the ratio of current to STL ROC-AUC, and multi-assay loss was defined as the weighted geometric mean of individual assay losses.

4 Results and Discussions

Average AUC over $\mathcal{T}^{\mathcal{P}}$ and per-assay AUC for MoMIL and competing baselines are presented in Tab. 2, across STL and MTL setups on both U2OS and iNeuron.

MoMIL consistently outperformed mean-pooling baselines, confirming MIL's ability to better capture instance-level variations. Also, FM features consistently outperformed Acapella features. On average, MoMIL achieved 3.4% higher AUC

$\mathcal{T}^{\mathcal{P}}$	$\mathcal{T}^{\mathcal{P}}_{\text{rel}}$	**5-shot** (Lin. Prob.)		**10-shot** (Lin. Prob.)		**All** (Lin. Prob.)		**All** (MIL)
		STL	**MTL**	**STL**	**MTL**	**STL**	**MTL**	**STL***
$\mathbf{t}_{\mathbf{1}}^{U}$	$\mathbf{t}_{\mathbf{1\text{-}1}}^{U}$	65.3±4.4	69.6±4.3	66.8±4.4	70.3±2.6	66.9	71.0	69.1
$\mathbf{t}_{\mathbf{2}}^{U}$	$\mathbf{t}_{\mathbf{2\text{-}1}}^{U}$	53.2±4.4	57.0±4.0	54.6±2.8	58.5±5.6	55.1	59.6	56.8
	$\mathbf{t}_{\mathbf{2\text{-}2}}^{U}$	58.3±8.4	60.2±6.8	62.8±3.3	63.8±4.7	75.5	75.8	74.5
	$\mathbf{t}_{\mathbf{2\text{-}3}}^{U}$	61.6±4.8	69.4±2.2	64.6±5.0	70.8±1.9	67.6	71.7	68.2
	$\mathbf{t}_{\mathbf{2\text{-}4}}^{U}$	59.1±6.0	68.2±6.8	56.9±8.2	77.3±7.2	76.1	84.2	80.5
	$\mathbf{t}_{\mathbf{2\text{-}5}}^{U}$	63.1±7.0	67.9±4.5	64.2±3.1	68.3±4.1	66.7	71.2	67.4
	$\mathbf{t}_{\mathbf{2\text{-}6}}^{U}$	53.6±5.5	53.9±5.3	55.7±5.7	57.9±4.6	59.0	59.2	63.2
$\mathbf{t}_{\mathbf{3}}^{U}$	$\mathbf{t}_{\mathbf{3\text{-}1}}^{U}$	52.2±5.3	61.9±3.8	52.9±2.4	62.6±3.4	56.0	64.3	60.4
	$\mathbf{t}_{\mathbf{3\text{-}2}}^{U}$	52.1±3.5	59.4±3.7	53.6±4.0	61.8±2.1	59.1	62.7	57.5
$\mathbf{t}_{\mathbf{4}}^{U}$	$\mathbf{t}_{\mathbf{4\text{-}1}}^{U}$	65.3±5.5	72.6±5.7	68.0±4.7	73.1±2.6	70.1	72.8	64.9
	$\mathbf{t}_{\mathbf{4\text{-}2}}^{U}$	60.9±7.5	69.2±4.5	61.9±6.3	73.8±3.5	67.3	76.0	59.1
$\mathbf{t}_{\mathbf{5}}^{U}$	$\mathbf{t}_{\mathbf{5\text{-}1}}^{U}$	88.2±2.0	93.5±0.7	88.6±1.6	93.8±0.7	90.1	94.8	93.3
	$\mathbf{t}_{\mathbf{5\text{-}2}}^{U}$	77.2±1.7	83.4±1.2	78.1±1.4	83.7±0.6	81.3	84.4	84.1
$\mathbf{t}_{\mathbf{6}}^{U}$	$\mathbf{t}_{\mathbf{6\text{-}1}}^{U}$	61.0±3.7	65.6±2.9	62.3±2.7	66.0±4.0	63.7	68.9	66.2
	$\mathbf{t}_{\mathbf{6\text{-}2}}^{U}$	65.8±3.1	73.9±1.6	67.3±3.7	74.4±2.9	73.2	78.3	74.9

Table 3: Few-shot generalizability assessment of pre-trained STL & MTL models on primary assays to unseen U2OS assays. AUC surpassing **STL*** are in yellow.

than FM → Mean → MLP and 4.6% higher than Acapella → MLP in STL, with similar gains of 2.8% and 4.1%, respectively, in MTL. In comparing STL and MTL MoMILs, MTL with $|\mathcal{T}| = 6$ performed $\geq$ STL, while including secondary assays ($|\mathcal{T}| = 200$) significantly improved overall AUC by 3.3%. While auxiliary assays offer benefits, simply including more assays can complicate learning, leading to conflicting gradients and convergence issues.

We observe that MoMIL consistently outperformed the AMTL and MLL baselines by overall AUC of 4.1% and 2.2%, respectively. The under-performance of AMTL and MLL can be attributed to the complexity of MTL and weaker knowledge sharing, respectively. Assay selection (AS) and label imputation (LI) further enhanced MoMIL. AS identified 22/194 and 26/194 relevant assays for U2OS and iNeuron, respectively, and improved MoMIL($|\mathcal{T}| = 200$) by 2.2%. AS outperformed the standalone biological-similarity and assay-performance based selection by 1.6% and 2.3% in overall AUC. LI produced overall gains of 0.8% and 1.6% in STL and MTL MoMILs, indicating its efficacy. LI performed better in MTL, with a robust classifier better supporting the imputation. In summary, MoMIL + AS + LI resulted in the best performance, achieving overall gains of 11.0%, 9.7%, 8.0%, and 6.1% over Acapella→MLP, FM→Mean→MLP, FM → MIL + AMTL, and FM → Multi-label-MIL($|\mathcal{T}| = 200$), respectively.

We evaluated the generalizability of MoMIL(STL) and MoMIL+AS+LI (MTL) to unseen assays in U2OS, Tab. 3. For each $t \in \mathcal{T}^{\mathcal{P}}$, we selected unseen $\mathcal{T}^{\mathcal{P}}_{\text{rel}_t}$ sharing the same target but different assay protocols, and inferred their well features using $\mathcal{F}_{\psi_t}$. Next, we performed linear probing via Logistic Regression for different few-shot setups $\forall t^{'} \in \mathcal{T}^{\mathcal{P}}_{\text{rel}_t}$. 5-, 10-, and All-shot refer to using 5, 10, and all positive and negative compounds per-fold, respectively. We compared the compound-level AUC against MoMIL (STL) trained specifically for $t^{'}$, denoted as STL*. Results show that MTL consistently generalizes better than STL. MTL outperformed STL* in 8/12 assays for 5- and 10-shot, and 11/12 assays for All-shot. These highlight the practicality of our framework for real-world assay modeling, achieving high performance even with limited assay activity data.

5 Conclusion

In summary, this paper presents MoMIL, a multi-task learning framework that addresses the challenges of predicting multiple assays in TDD under extreme label sparsity. By integrating assay-specific MIL heads, pre-trained FMs on HCI data, the assay selection and label imputation algorithms, MoMIL outperforms state-of-the-art MLL- and MTL-based compound activity modeling methods. On top of MoMIL's success, assay-specific adaptations can be further optimized. Future work will explore advanced assay selection strategies, such as algorithm-based selection of relevant assays, and refined imputation techniques to further enhance model robustness and generalization in drug discovery applications.

6 Disclosure of Interests[1]

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)
2. Arevalo, J., Su, E., Ewald, J.D., van Dijk, R., Carpenter, A.E., Singh, S.: Evaluating batch correction methods for image-based cell profiling. Nature Communications **15**(1), 6516 (2024)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
4. Crowell, H.L., Soneson, C., Germain, P.L., Calini, D., Collin, L., Raposo, C., Malhotra, D., Robinson, M.D.: Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nature communications **11**(1), 6077 (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
7. Engelmann, J.P., Palma, A., Tomczak, J.M., Theis, F., Casale, F.P.: Mixed models with multiple instance learning. pp. 3664–3672 (2024)

[1] The authors have no competing interests to declare that are relevant to the content of this article.

8. Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., Finn, C.: Efficiently identifying task groupings for multi-task learning. Advances in Neural Information Processing Systems **34**, 27503–27516 (2021)
9. Fredin Haslum, J., Lardeau, C.H., Karlsson, J., Turkki, R., Leuchowius, K.J., Smith, K., Müllers, E.: Cell painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. Nature Communications **15**(1), 3470 (2024)
10. Giuliano, K.A., Haskins, J.R., Taylor, D.L.: Advances in high content screening for drug discovery. Assay and drug development technologies **1**(4), 565–577 (2003)
11. Gustafsdottir, S.M., Ljosa, V., Sokolnicki, K.L., Anthony Wilson, J., Walpita, D., Kemp, M.M., Petri Seiler, K., Carrel, H.A., Golub, T.R., Schreiber, S.L., et al.: Multiplex cytological profiling assay to measure diverse cellular states. PloS one **8**(12), e80999 (2013)
12. Herman, D., Kandúła, M.M., Freitas, L.G., van Dongen, C., Le Van, T., Mesens, N., Jaensch, S., Gustin, E., Micholt, L., Lardeau, C.H., et al.: 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)
13. Hofmarcher, M., Rumetshofer, E., Clevert, D.A., Hochreiter, S., Klambauer, G.: Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. Journal of Chemical Information and Modeling **59**(3), 1163–1171 (2019)
14. Hughes, J.P., Rees, S., Kalindjian, S.B., Philpott, K.L.: Principles of early drug discovery. British journal of pharmacology **162**(6), 1239–1249 (2011)
15. Kim, V., Adaloglou, N., Osterland, M., Morelli, F.M., Halawa, M., König, T., Gnutt, D., Zapata, P.A.M.: Self-supervision advances morphological profiling by unlocking powerful image representations. bioRxiv (2023)
16. Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al.: Masked autoencoders for microscopy are scalable learners of cellular biology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11757–11768 (2024)
17. Krentzel, D., Shorte, S.L., Zimmer, C.: Deep learning in image-based phenotypic drug discovery. Trends in Cell Biology **33**(7), 538–554 (2023)
18. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2021)
19. McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B.A., Karhohs, K.W., Doan, M., Ding, L., Rafelski, S.M., Thirstrup, D., et al.: Cellprofiler 3.0: Next-generation image processing for biology. PLoS biology **16**(7), e2005970 (2018)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
21. Seal, S., Trapotsi, M.A., Spjuth, O., Singh, S., Carreras-Puigvert, J., Greene, N., Bender, A., Carpenter, A.E.: Cell painting: a decade of discovery and innovation in cellular imaging. Nature methods pp. 1–15 (2024)
22. Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J.K., Gustin, E., Chupakhin, V., Chong, Y.T., Vialard, J., Buijnsters, P., et al.: Repurposing high-

throughput image assays enables biological activity prediction for drug discovery. Cell chemical biology **25**(5), 611–618 (2018)

23. Sivanandan, S., Leitmann, B., Lubeck, E., Sultan, M.M., Stanitsas, P., Ranu, N., Ewer, A., Mancuso, J.E., Phillips, Z.F., Kim, A., et al.: A pooled cell painting crispr screening platform enables de novo inference of gene function by self-supervised deep learning. bioRxiv pp. 2023–08 (2023)
24. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al.: The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic acids research **51**, D638–D646 (2023)
25. Yu, Y., Yang, T., Lv, Y., Zheng, Y., Hao, J.: T3s: Improving multi-task reinforcement learning with task-specific feature selector and scheduler. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2023)
26. Yun, H., Cho, H.: Achievement-based training progress balancing for multi-task learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16935–16944 (2023)
27. Zhang, G., Jain, A., Hwang, I., Sun, S.H., Lim, J.J.: Efficient multi-task reinforcement learning via selective behavior sharing. arXiv preprint arXiv:2302.00671 (2023)