Machine Learning with Applications 24 (2026) 100867

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

# Group-based learning on label-free phase-contrast images across dose and exposure time improves bioactive compound classification

Denis Baručić [a,1], Jarmila Stanková [b,1], Jan Kybic [a,*], Marián Hajdúch [b,c], Petr Džubák [b,c,**]

[a] Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, Prague, 166 27, Czech Republic
[b] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University and University Hospital Olomouc, Hněvotínská 1333/5, 779 00, Olomouc, Czech Republic
[c] Institute of Molecular and Translational Medicine, Czech Advanced Technology and Research Institute, Palacký University Olomouc, Hněvotínská 1333/5, 779 00, Olomouc, Czech Republic

ARTICLE INFO

*Keywords:*
Drug discovery
Compound classification
Digital phase-contrast microscopy
Deep learning

ABSTRACT

High-throughput phenotypic screening enables large-scale analysis of compound effects on cells using microscopy images. Typical approaches to phenotypic screening rely on fluorescence images, which are relatively straightforward to analyze but require fluorescence staining—a time-consuming process that can perturb cells. In this work, we focus instead on digital phase-contrast imaging. This imaging alternative avoids staining and is therefore faster and non-invasive, but it produces visually complex images that are more challenging to analyze. We address classifying 1 280 diverse compounds solely from digital phase-contrast images. To account for a priori unknown drug-specific potency and temporal effects, we propose a method that classifies groups of images acquired at varying compound concentrations and exposure times, rather than processing individual images. Our approach achieves 50% top-1 accuracy on the classification problem of 1 280 compounds. Moreover, we demonstrate that the learned image encoder transfers effectively to related tasks, including mechanism-of-action prediction and control-versus-treated classification. The presented results highlight the potential of combining digital phase-contrast imaging with appropriate machine learning models to enable scalable, cost-effective, and minimally invasive phenotypic screening.

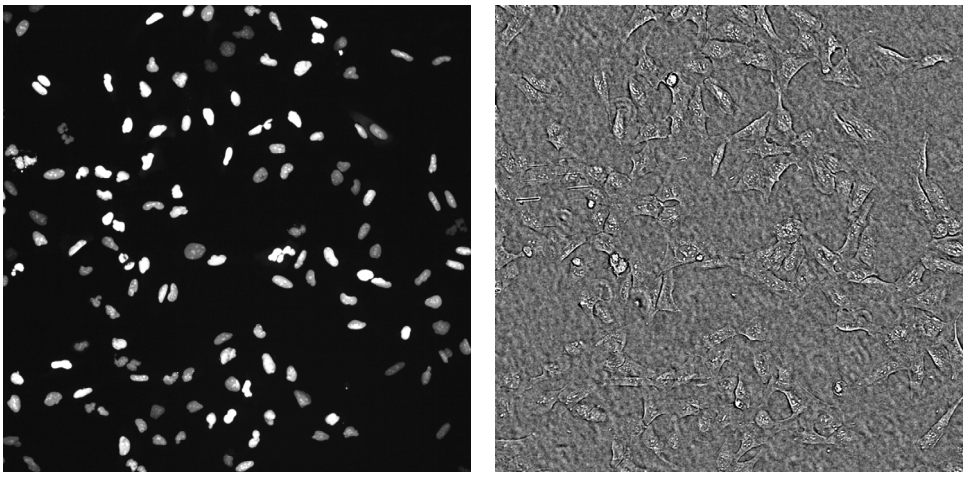
## 1. Introduction

Drug discovery is a process of identifying and developing compounds with therapeutic potential while minimizing side effects (Smith, 2002). A typical pipeline begins with assembling a library of candidate compounds based on prior knowledge and computational modeling. These compounds are then evaluated using *cellular phenotypic screening*, which tests their effects in vitro on standardized cell lines. This step is valuable because it can reveal therapeutic selectivity, mechanisms of action, and potential unintended off-target interactions (Lang et al., 2006; Zheng et al., 2013). In high-throughput image-based screening, thousands of compounds are tested in parallel using automated robotic platforms. Microscopy images captured at defined time points serve as the main output, generating large image datasets, which highlights the need for automated image analysis.

This study addresses the problem of classifying pharmacological compounds, or drugs, directly from digital phase-contrast images of treated cells. In earlier work, we demonstrated the feasibility of compound classification on a small set of four compounds (Baručić et al., 2022), and further compared the performance of phase-contrast and fluorescence imaging within that experimental context. The results indicated that phase-contrast images alone captured sufficient phenotypic information for compound discrimination, motivating a label-free imaging strategy. Here, we extend the previous work to a substantially larger and more diverse compound library of 1280 compounds. A key feature of this work is the exclusive use of *phase-contrast* microscopy (Hecht, 2017; Zernike, 1942), in contrast to most studies that rely on *fluorescence microscopy* (Moshkov et al., 2023; Stossi et al., 2024; Wang et al., 2008). Fluorescence microscopy produces images with clear cell boundaries, which allows for simple cell segmentation and thus

* Corresponding author.
** Corresponding author at: Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University and University Hospital Olomouc, Hněvotínská 1333/5, 779 00, Olomouc, Czech Republic.
*E-mail addresses:* barucden@fel.cvut.cz (D. Baručić), jstankova@ibecbarcelona.eu (J. Stanková), kybic@fel.cvut.cz (J. Kybic), marian.hajduch@upol.cz (M. Hajdúch), petr.dzubak@upol.cz (P. Džubák).
[1] Authors contributed equally.

https://doi.org/10.1016/j.mlwa.2026.100867
Received 18 November 2025; Received in revised form 30 January 2026; Accepted 10 February 2026
Available online 11 February 2026



(a) Fluorescence (b) Digital phase-contrast

**Fig. 1.** Example microscopy images displaying untreated U2OS cells. Brightness and contrast were adjusted for clarity.

facilitates downstream analysis (see Fig. 1(a)). However, it requires fluorescence staining, which increases cost and experimental complexity and potentially perturbs cells. Phase-contrast microscopy is stain-free, faster, cheaper, and thus well-suited for large-scale and longitudinal studies. On the other hand, the cells are more difficult to distinguish due to low contrast, complex textures, and indistinct cell boundaries (see Fig. 1(b)), making the automated analysis more challenging. To our knowledge, this is the first demonstration of large-scale compound classification directly from raw digital phase-contrast images, with our method performing better than recent fluorescence-based approaches (see Section 6).

The second distinguishing feature of this work is the use of groups of images of the same compound across multiple concentrations and exposure times. Previous studies (Baručić et al., 2022; Cox et al., 2020) have highlighted the importance of choosing these parameters carefully. If the dose is too low, no observable effects may appear, whereas excessively high doses can result in cell death, making different compounds difficult to distinguish. Because optimal parameters are drug-specific and unknown in advance, optimizing the concentration for each compound individually would be unsuitable in a high-throughput setting. Our approach instead relies on providing images in several different configurations (concentration and exposure time) and lets the algorithm choose the most informative ones automatically. Conceptually, this is similar to the attention mechanism (Bahdanau et al., 2015) or multiple instance learning (MIL) (Dietterich et al., 1997; Ilse et al., 2018). However, we show that our approach outperforms standard methods on an extensive dataset of cells treated by biologically active compounds (see Section 5.2).

Compound classification can also be regarded as a proxy task for learning suitable features. We demonstrate that intermediate features of our classifier are highly informative and can be used also for related tasks such as mechanism-of-action prediction and treated-versus-control discrimination (Sections 5.5 to 5.7). This represents a step towards the broader goal of predicting the properties of new, unseen compounds in high-throughput phenotypic screening.

## 2. Related work

### 2.1. Cell image analysis

Automatic analysis of microscopy images of cells has numerous applications in fields such as digital pathology or cell biology (Xing et al., 2018). Here, we focus on multiclass image classification, where the goal is to assign a label to each image. In the context of image-based cell profiling (or phenotyping), this involves distinguishing subtle perturbations in population of cells from the same line, caused by exposure to tested compounds. However, similar methods can be used for other classification tasks, such as identifying different cell types, their developmental stage, senescence (cellular aging), or states of biological activity.

Traditional approaches to cell phenotyping rely on hand-crafted features describing individual cells (Caicedo et al., 2017; Li et al., 2020; Prasad & Alizadeh, 2019; Smith et al., 2018), which can be extracted by specialized software such as CellProfiler (Carpenter et al., 2006; Moshkov et al., 2023; Stirling et al., 2021) or CellPhe (Wiggins et al., 2023). These features typically capture shape, morphology, intensity, or texture. Subsequently, the extracted features are passed to a classifier, such as support vector machines (Wang et al., 2008) or random forests (Ko et al., 2011). The advantage of such an approach is processing speed and interpretability, and so it is used even in recent works (Stossi et al., 2024).

In recent years, deep learning has become the dominant approach for cell phenotyping and related image classification tasks (Krentzel et al., 2023; Li et al., 2020; Moen et al., 2019; Pratapa et al., 2021), using standard classification architectures such as ResNet or Inception (Kensert et al., 2019). The image can be processed as a whole, or cells can be segmented first and analyzed individually (Zagajewski et al., 2023), following the classical pipeline. Sometimes, deep learning and classical tools are combined, with deep learning used only for specific steps (Stossi et al., 2024). A multi-scale approach can help capture phenotypic features at multiple spatial resolutions (Godinez et al., 2017). In single-cell imaging, deep learning was used to classify cell types (Meng et al., 2018; Yao et al., 2019) or mechanisms of action (Perakis et al., 2021) from bright-field or fluorescence images.

Approaches that process individual cells or extract classical features critically depend on the segmentation quality of analyzed cells (or their parts). While cell segmentation is relatively straightforward in fluorescence images—either through simple thresholding (Pham, 2007) or more sophisticated methods such as deep networks (Dubach & Dubach, 2024; Pratapa et al., 2021; Stringer et al., 2021)—it remains a challenge in label-free modalities. Consequently, fluorescence imaging continues to dominate large-scale cell profiling studies. It is possible to use generative adversarial networks to synthesize fluorescence images from bright-field or phase-contrast images (Cross-Zamirski et al., 2022; Mertanová et al., 2022). However, while visually convincing, the information content of synthesized images is limited, reducing their utility for downstream tasks.

Direct use of phase-contrast images for cell analysis via deep learning has been explored only in several works, such as detecting cells (Essa & Xie, 2018), determining differentiation (Niioka et al., 2018;

Waisman et al., 2019) or DNA fragmentation (Kikuchi et al., 2024), identifying cancer cells (Kang & Kim, 2025), or detecting senescence (Kusumoto et al., 2021). These studies addressed narrow problems but demonstrated the feasibility of phase-contrast image analysis. To the best of our knowledge, no prior work has attempted large-scale compound classification directly from raw phase-contrast images.

### 2.2. Group classification

The problem of classifying groups of instances has been studied in multiple variants, each characterized by different assumptions about dependencies between the instances in a group and the relationship between the instance and group labels. For example, Samsudin and Bradley (2010) consider group classification where all instances share the same label. We have the same assumption, but we also account for the fact that not all instances are equally informative or relevant. Vural et al. (2006) consider a setting where instances within a group are correlated. We, in contrast, assume conditional independence of the instances in a group given the group label.

According to the taxonomy of group classification problems by Cheplygina et al. (2015), our problem falls into the MI–MI class, where labeled groups—but not labeled instances—are provided for training and unlabeled groups are processed during testing.

In Multiple Instance Learning (MIL) (Dietterich et al., 1997), a notable representative of the MI–MI class of problems, a group is labeled positive if and only if it contains at least one positive instance, corresponding to a logical disjunction over the hidden instance labels. Three established methods based on deep learning that solve this problem model are mi-Net (Wang et al., 2018), MI-Net (Wang et al., 2018), and Attention-based MIL (AMIL) (Ilse et al., 2018). The first, mi-Net, predicts instance-level scores and aggregates them using differentiable pooling, such as LogSumExp, enabling end-to-end learning. The latter two, MI-Net and AMIL, employ a pooling operation to aggregate the feature vectors extracted from the instances in a group. The aggregated feature vector is then used to classify the group. MI-Net employs a fixed pooling such as mean pooling. AMIL introduces an attention mechanism that learns to assign high weights to key (positive) instances. AMIL is derived from Deep Sets (Zaheer et al., 2017) that describe deep learning models for tasks defined on sets. CytoSummaryNet (van Dijk et al., 2024) uses Deep Sets to aggregate single-cell features for image-based cell profiling.

## 3. Data

This section outlines the data acquisition methods and summarizes the resulting datasets. See Appendix B for a detailed description of the data acquisition. The data and all annotations are published online.[2]

### 3.1. Cell treatments

U2OS cells were seeded in 384-well PhenoPlates and, after 24 h without media exchange, treated with reference compounds from the LOPAC1280 Library of Pharmacologically Active Compounds (Sigma-Aldrich, product L01280) at three concentrations: 50 µM, 5 µM, and 0.5 µM. In parallel, separate plates were treated with JUMP Target compounds (Specs, product CR-12417) at concentrations of 50 µM, 5 µM, and 0.8 µM. All compounds were reconstituted in DMSO, and source plates included wells reserved for DMSO vehicle controls. To minimize positional effects, compounds were dispensed randomly into destination plates containing cells, with randomized layouts differing between replicates.

### 3.2. Digital phase-contrast imaging

Plates were imaged *at two timepoints*, 24 h and 72 h after compound application, to capture early and late responses, using a Yokogawa CV8000 high-throughput spinning disk confocal microscope configured for digital phase contrast. Digital phase-contrast imaging was performed with a 20× objective, acquiring a 3-plane z-stack (1 µm apart with on-the-fly image merging) and six fields of view (FOVs) per well. Under these conditions, a full 384-well plate was imaged in approximately 45 min, generating around 4.5 GB of data. The spatial resolution of each image was 1000×1000 pixels. The data comprise two datasets: *LOPAC* (Section 3.3) and *JUMP Target* (Section 3.4).

### 3.3. LOPAC

The primary dataset consists of cells treated with 1280 compounds from LOPAC1280 (see Fig. 2).

The dataset consists of three equivalent replicates. In each replicate, every compound was applied to one well per concentration, and 768 wells were reserved for DMSO controls. With six FOVs per well, this yielded:

- 3 replicates×1280 wells×3 concentrations×2 time points×6 FOVs = 138 240 images of treated cells,
- 3 replicates × 768 wells × 2 time points × 6 FOVs = 27 648 control images of untreated cells.

#### 3.3.1. Cytotoxic and non-cytotoxic compounds

Cytotoxicity data for all LOPAC1280 compounds was measured by an *MTS metabolic assay* performed on U2OS cells. Compounds were considered *cytotoxic* if they showed measurable activity at 50 µM, and *non-cytotoxic* if no significant activity was observed at that concentration. Of the 1280 compounds, 225 were labeled as cytotoxic and 1055 as non-cytotoxic. All procedures were fully automated using a HighResBio robotics platform (Veselovská et al., 2020).

#### 3.3.2. Mechanism of action and cell viability

We used compound Mechanism of action (MoA) annotations and cell viability profiles derived from the PRISM drug repurposing dataset (Corsello et al., 2020). The MoA annotations were cross-verified using two additional databases (Corsello et al., 2017; Santos et al., 2017). While an MoA annotation in general corresponded to a set of multiple mechanisms of action, the sets for different compounds were either equal or disjoint, allowing us to map each set to a single label.

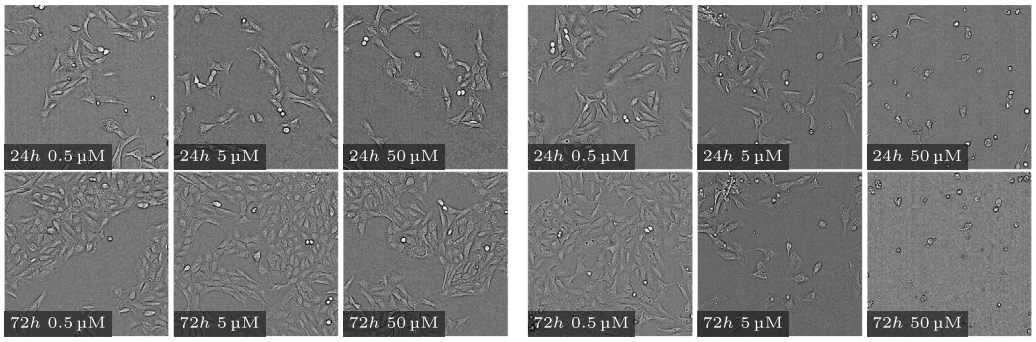
Cytotoxic activity was quantified using the $\log_2$-fold change (LFC) values reported in the dataset. The LFC reflects the relative cell viability after compound treatment compared to controls; a negative LFC indicates reduced cell viability, i.e. inhibitory effect of the compound. Of the 1280 LOPAC compounds, 247 overlapped with the PRISM drug repurposing dataset and had assigned both MoA annotations and LFC values. In total, 138 unique MoA labels were represented among these compounds.

### 3.4. JUMP target

We generated another dataset by treating U2OS cells with 302 JUMP (Joint Undertaking in Morphological Profiling) Target compounds (including DMSO, which should have no effect and serves as control),[3] as defined by the JUMP-Cell Painting Consortium (Ch et al., 2023) (see Fig. 3). In contrast to the LOPAC1280 library, the compounds in the JUMP Target library were curated to modulate a variety of cellular targets, which should in principle make them easier to classify.

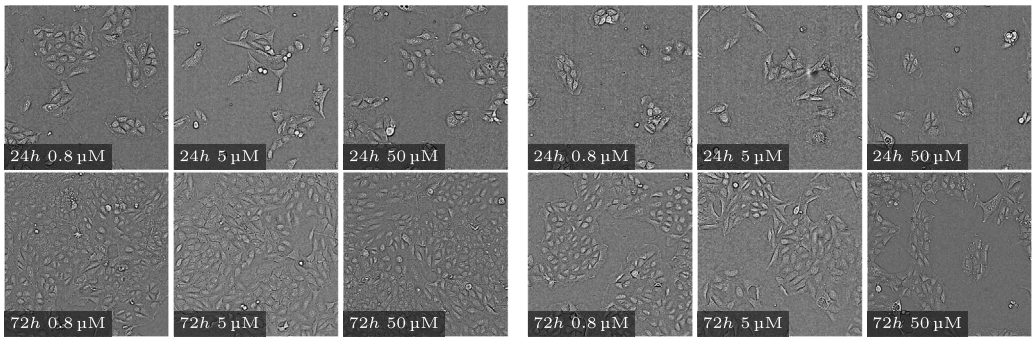
[2] https://lifedata.cz/BIOIMG_0002.

[3] https://github.com/jump-cellpainting/JUMP-Target.




(a) Ebastine (b) Ellipticine

**Fig. 2.** Digital phase-contrast images of two drugs at different configurations from the LOPAC dataset.




(a) Alrestatin (b) Benzamil

**Fig. 3.** Digital phase-contrast images of two drugs at different configurations from the JUMP target dataset.

The dataset contains three independent replicates. In each replicate, 79 wells were reserved for DMSO-treated controls, and 305 wells were treated with active compounds—four compounds (BVT-948, dexamethasone, ME-0328, and thiostrepton) were present in duplicate in the compound library and thus applied to two wells each. Six fields of view (FOVs) were acquired per well at two time points. This experimental setup yielded:

- 3 replicates × 305 wells × 3 concentrations × 2 time points × 6 FOVs = 32 940 images of treated cells,
- 3 replicates × 79 wells × 3 concentrations × 2 time points × 6 FOVs = 8532 control images.

Due to acquisition errors, 102 treated and 30 DMSO images were excluded from analysis.

## 4. Methods

### 4.1. Problem formulation

Let $\mathcal{X}$ be the set of all possible images and $\mathcal{Y} = \{1, \dots, K\}$ the set of all labels (classes or compounds). Instead of classifying individual images (instances), we formulate the task as a group classification. Specifically, we partition the images into disjoint groups, each containing images of cells from one biological replicate treated with the same drug across multiple configurations. Given a group, our construction allows the classifier to focus on the configurations most representative of each drug's effect. In order to maximize the number of labeled groups while still capturing the effects across all configurations, each configuration is included exactly once per group.

As a result, each group $\mathbf{X} \in \mathcal{X}^n$ consists of $n$ images with the same label $y$, where $n$ is the number of configurations. We get 18 groups per compound for the LOPAC dataset, a relatively small number for deep learning methods. Note, however, that our method described below is agnostic to the particular grouping scheme used.

Given a training dataset of labeled groups $\mathcal{T} = \{(\mathbf{X}^j, y^j)\}_{j=1}^m$, the task is to learn to estimate the unknown labels from a similarly defined test dataset. The difference from a standard supervised classification setting is that we can use information from several images to predict the group label. On the other hand, some of the images may not be useful for the classification.

### 4.2. Image-level class probabilities

We start with a standard CNN-based encoder $f_\theta : \mathcal{X} \to \mathbb{R}^D$ with parameters $\theta$ to map an input image $\mathbf{x}$ to a $D$-dimensional feature vector, $f_\theta(\mathbf{x}) \in \mathbb{R}^D$. A classifier $g_\varphi : \mathbb{R}^D \to \mathbb{R}^K$ with parameters $\varphi$ then maps the features to a vector of real scores for every class $y = 1, \dots, K$. The elements of this vector will be denoted $\big[g_\varphi\big(f_\theta(\mathbf{x})\big)\big]_y$.

We apply softmax normalization to turn the scores $g_\varphi\big(f_\theta(\mathbf{x})\big)$ into probabilities,

$$P(y \mid \mathbf{x}) = \frac{\exp\Big(\big[g_\varphi\big(f_\theta(\mathbf{x})\big)\big]_y\Big)}{\sum_{y'=1}^K \exp\Big(\big[g_\varphi\big(f_\theta(\mathbf{x})\big)\big]_{y'}\Big)}. \tag{1}$$

We employ ResNet-18 (He et al., 2016) as the image encoder $f_\theta$ in all experiments. The encoder produces vectors of $D = 512$ elements. The classifier $g_\varphi$ consists of two linear layers with a ReLU activation in between (see Fig. 4).

### 4.3. Aggregation

To aggregate the class probabilities (1) within each group, we take their product and normalize the results, obtaining estimated group label

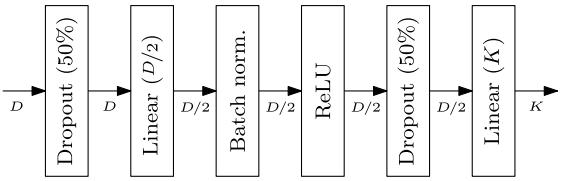



**Fig. 4.** Architecture of classifier $g_\varphi$. The labels below the arrows indicate the dimension of vectors passed between the respective layers. In our experiments, $D = 512$ and $K$ depends on the particular classification task.

probabilities $Q$,

$$Q(y \mid \mathbf{X}) = \frac{\prod_{\mathbf{x}\in\mathbf{X}} P(y \mid \mathbf{x})}{\sum_{y'=1}^{K} \prod_{\mathbf{x}\in\mathbf{X}} P(y' \mid \mathbf{x})}, \tag{2}$$

which can be simplified using (1) to a softmax activation of the sum of scores $g_\varphi$ in a group,

$$Q(y \mid \mathbf{X}) = \frac{\exp\left(\sum_{\mathbf{x}\in\mathbf{X}} \left[g_\varphi\left(f_\theta(\mathbf{x})\right)\right]_y\right)}{\sum_{y'=1}^{K} \exp\left(\sum_{\mathbf{x}\in\mathbf{X}} \left[g_\varphi\left(f_\theta(\mathbf{x})\right)\right]_{y'}\right)}. \tag{3}$$

We argue that the product-based aggregation is even more powerful than the alternative, attention-based aggregation (Ilse et al., 2018). The image-based probabilities $P(\cdot \mid \mathbf{x})$ may be uniform if the image $\mathbf{x}$ does not carry any useful information for the classification. This would correspond to a low weight in the attention formulation. However, if the evidence is sufficiently strong, $P(\cdot \mid \mathbf{x})$ can also rule out one or more classes (if $P = 0$) or force one of the classes (if $P = 1$).

### 4.4. Learning and inference

Given annotated training data $\mathcal{T} = \left\{(\mathbf{X}^j, y^j)\right\}_{j=1}^{m}$, we learn the parameters $(\theta, \varphi)$ of the encoder $f_\theta$ and the classifier $g_\varphi$ by maximizing the likelihood of $\mathcal{T}$, which is equivalent to minimizing the logarithmic loss $L$:

$$L = -\frac{1}{m} \sum_{j=1}^{m} \log Q(y^j \mid \mathbf{X}^j). \tag{4}$$

At inference time, to predict the class of a given group $\mathbf{X}$, we predict the most probable label according to $Q$,

$$\tilde{y} = \arg\max_y Q(y \mid \mathbf{X}). \tag{5}$$

This is equivalent to

$$\tilde{y} \in \arg\max_y \sum_{\mathbf{x}\in\mathbf{X}} \left[g_\varphi\left(f_\theta(\mathbf{x})\right)\right]_y, \tag{6}$$

since the normalizing term in Eq. (3) is constant with respect to $y$.

### 4.5. Related aggregation strategies

The proposed model of $Q(y \mid \mathbf{X})$ (2) resembles *Product of Experts* (PoE) (Cao & Fleet, 2014; Hinton, 2002). Like PoE, our approach combines $n$ predictions via a product of probabilities (2) to form an overall distribution. However, unlike PoE, which uses $n$ models—called experts—to produce $n$ distributions from a single input, we employ a single model that accepts $n$ distinct inputs and predicts a separate distribution for each (see Fig. 5).

The main contribution of PoE is that each expert can focus on modeling a different aspect of the target distribution. It uses relatively simple models, the combination of which is more powerful than each constituent model. In contrast, our goal is to integrate information from multiple input images using a common, relatively complex classifier.

The well-known Naive Bayes (NB) classifier also aggregates predictions from multiple sources via a product. The difference is that it uses the product to model the likelihood $P(\mathbf{X} \mid y)$, instead of the posterior probability $P(y \mid \mathbf{X})$. Moreover, the individual probability models $P(\mathbf{x}_i \mid y)$ are usually simple and unsuitable to classify images.

## 5. Results

Section 5.2 evaluates our method's performance on the main task of compound classification using the primary dataset. Next, Section 5.3 analyzes the impact of compound cytotoxicity on the classification performance. Section 5.4 shows the benefits of our group-based formulation. Section 5.5 evaluates mechanism-of-action prediction, and Section 5.6 test treated-versus-control recognition. Section 5.7 evaluates compound prediction on a secondary dataset containing a curated set of compounds with observable effects on the cells. The last three experiments demonstrate the feature encoder's generalization ability. Finally, we evaluate the classification performance on perturbed inputs to test the robustness of the proposed aggregation strategy (Section 5.8).

### 5.1. Training and testing

We trained all models for 350 epochs. We used the AdamW optimizer (Loshchilov & Hutter, 2019) with weight decay fixed to $10^{-4}$ and learning rate initialized to $10^{-4}$ and adjusted via the cosine annealing with warm restarts (Loshchilov & Hutter, 2017) with periods of 50, 100, and 200 epochs. To prevent overfitting, we augmented the training images via flipping, affine transforms, contrast and brightness adjustments, blurring, and adding noise.

To obtain quantitative experiment results, we performed a 6-fold cross-validation. In each cross-validation run, we partitioned the image groups into disjoint training (5/6 of the total) and testing (1/6) subsets using stratified sampling to ensure that the class distribution was identical in the training and testing subsets. A validation subset was not needed since all hyperparameters were fixed.

The evaluation was always performed on the group level (not the image level). We computed standard evaluation metrics for classification tasks (see Appendix A).

All experiments were performed on a computer with an AMD EPYC 7543 (32 cores, 3.7 GHz), a single NVIDIA Tesla A100 (40 GB), and 180 GB of memory. Training one model using our data took approximately three days, with a batch of 24 image groups taking about 800 ms. Our implementation of the proposed method is available online.[4]

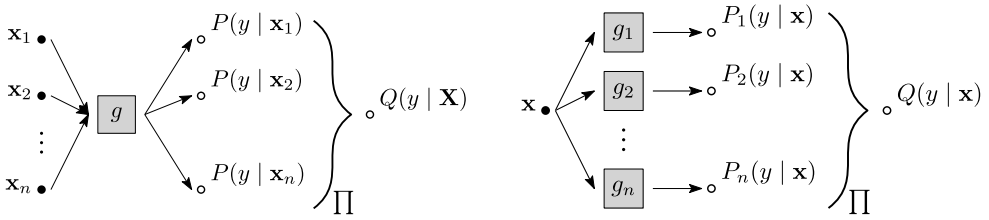
### 5.2. Compound classification

We tested the group classification performance on the LOPAC dataset (Section 3.3). Recall that there are 1280 classes (compounds), each group consists of 6 images taken in 6 different configurations (3 concentrations and 2 exposure times). There are 15 groups per class in each training splits and 3 groups per class in each test split. The groups are created according to the formulation in Section 4.1. Given a group of images, the objective is to predict the applied drug.

We compare the performance of the proposed method with a baseline classification method, which considers every image as an independent instance, and three existing group classification approaches: AMIL (Ilse et al., 2018), MI-Net (Wang et al., 2018), and mi-Net (Wang et al., 2018) (see Section 2.2). We use mean as the pooling operation in MI-Net. The baseline method is a naive group classifier trained on individual images (not groups) using a cross-entropy loss. During inference, predictions were aggregated over the instances within each group using the proposed strategy (5), ensuring a fair comparison with the group-based methods. Note that the baseline and proposed methods are equivalent during inference but differ in training. All alternative methods use the same architecture for the networks $f$ and $g$ as our group-based method (Section 4).

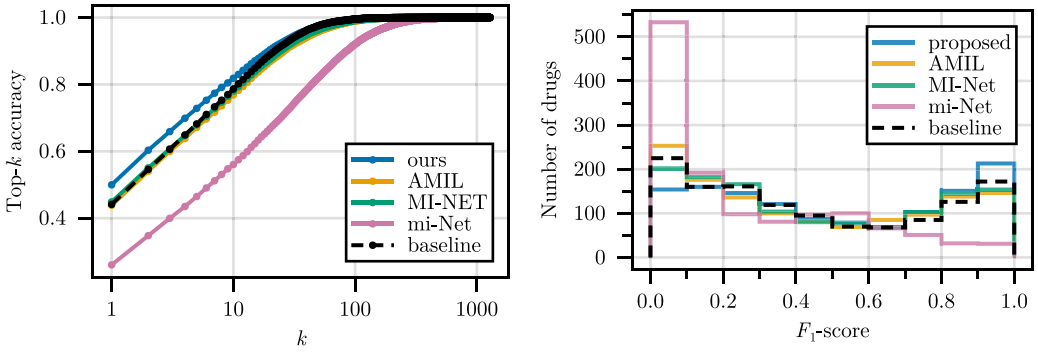
Table 1 shows the top-$k$ accuracy, i.e., the probability that the correct class appears among the top $k \in \{1, 5, 10\}$ candidates determined by a given method. Fig. 6(a) shows a graph of top-$k$ accuracy for

[4] https://github.com/barucden/comp-class-groups.




(a) Proposed (b) Product of Experts

**Fig. 5.** Schematic comparison of two related approaches to modeling aggregate distributions.




(a) Top-$k$ accuracy plot. (b) Histogram of $F_1$-scores.

**Fig. 6.** Performance of compound classification based on the results collected from all six cross-validation runs.

**Table 1**
Top-$k$ accuracy [%] of compound classification achieved by the tested methods. The bottom part shows results achieved by the baseline method when processing only one concentration and both exposure times. The values represent the mean and standard deviation computed over six cross-validation runs.

| Approach | Top-1 | Top-5 | Top-10 |
|---|---|---|---|
| Proposed | 49.98 ± 0.96 | 72.84 ± 0.53 | 81.84 ± 0.46 |
| AMIL | 43.84 ± 1.00 | 67.17 ± 0.57 | 76.85 ± 0.39 |
| MI-Net | 43.11 ± 0.97 | 67.96 ± 1.15 | 78.02 ± 1.07 |
| mi-Net | 26.04 ± 0.42 | 46.48 ± 0.47 | 56.04 ± 0.46 |
| Baseline | 44.12 ± 0.76 | 68.18 ± 0.75 | 78.64 ± 0.74 |
| Baseline (0.5 μM) | 8.04 ± 0.69 | 23.47 ± 0.89 | 33.69 ± 0.90 |
| Baseline (5 μM) | 13.43 ± 0.59 | 30.43 ± 0.90 | 41.00 ± 0.54 |
| Baseline (50 μM) | 36.68 ± 0.60 | 57.37 ± 0.81 | 66.44 ± 0.58 |

$k = 1, \ldots, 1280$. The proposed method consistently and significantly outperformed the other tested methods. AMIL, MI-Net, and the baseline method performed comparably, whereas mi-Net's performance was the lowest. In particular, the proposed method's accuracy ($k = 1$) was almost 50%.

Some compounds were easier to classify than others. We report the histogram of the achieved $F_1$ scores for each class (see Appendix A) in Fig. 6(b). The proposed method accurately classified ($F_1 > 0.9$) over 200 compounds and failed to recognize ($F_1 < 0.1$) approximately 150 compounds. Compounds varied widely in difficulty, but our approach classified more compounds accurately than alternatives.

### 5.3. Effect of cytotoxicity

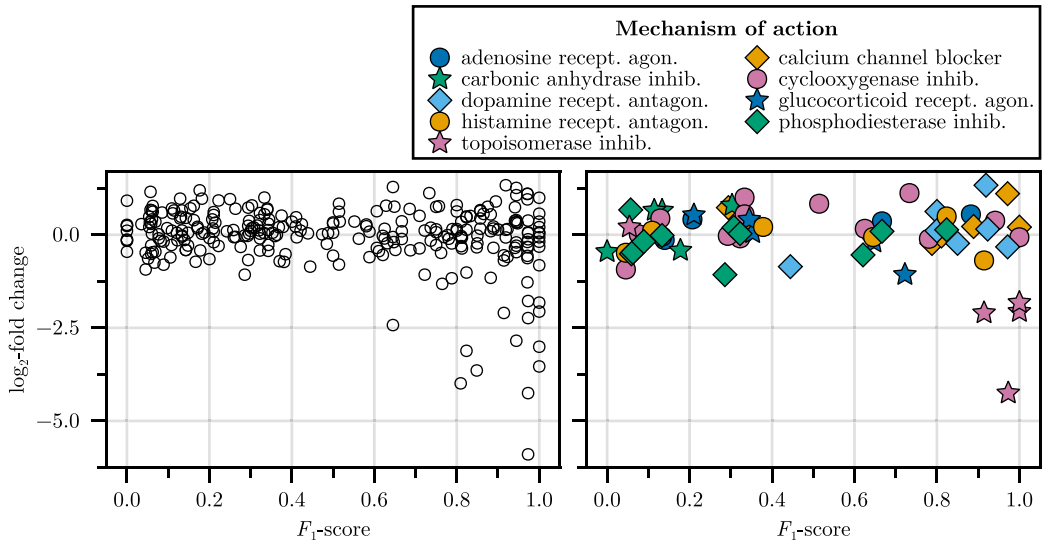
To investigate why some compounds were easier to classify, we plotted the $\log_2$-fold change (LFC) values against the $F_1$-score for the subset of 247 compounds from the LOPAC dataset with available LFC measurements (Section 3.3.2). Compounds with a strongly negative LFC (i.e. high inhibitory effect) were classified accurately (see Fig. 7a). On the other hand, a LFC around zero corresponded to both good and bad predictions. To explore the misclassifications in more detail, we show also the compounds' mechanism-of-action (MoA) information (see Fig. 7b) to see whether specific mechanisms could explain the variation in classification performance among non-cytotoxic compounds (LFC $\approx$ 0). Interestingly, calcium channel blockers and dopamine receptor antagonists tended to cluster in regions with higher $F_1$ scores, suggesting that certain MoAs may be easier to detect even when their effects on proliferation are minimal. A few exceptions—two calcium channel blockers (felodipine and levetiracetam, which do not target L-type channels) and the dopamine receptor antagonist amisulpride, a benzamide derivative with a distinct pharmacological profile—likely reflect differences in selectivity and target engagement. Other MoA classes were distributed more broadly across the $F_1$-LFC space.

We further analyzed the classification accuracy separately on the subsets of cytotoxic and non-cytotoxic compounds (Section 3.3.1). Technically, this is done by zeroing the predicted probabilities of all compounds not belonging to the subset of interest, without retraining the network.

While the accuracy on the non-cytotoxic subset (45.50%) was comparable to that on the whole set of compounds (49.98%), the accuracy on the cytotoxic subset (86.74%) was substantially higher (see Table 2). This result indicates that recognizing the cytotoxic compounds is—as expected—easier, but recognizing the non-cytotoxic compounds is also possible to some extent. Our proposed method again outperformed others.

### 5.4. Configuration relevance

We evaluated the baseline method on individual compound concentrations. Specifically, we used groups of images corresponding to the given concentration and both exposure times (24 and 72 h). Higher concentrations resulted in higher classification accuracy (see Table 1). In particular, the lowest concentration (0.5 μM) yielded an accuracy of only 8%, whereas the highest concentration (50 μM) achieved 37%. On the other hand, using all concentrations together worked even




(a) Compounds (b) Compounds by mechanism of action

**Fig. 7.** Dependency between the achieved $F_1$-scores and the corresponding $\log_2$-fold change values for (a) 247 compounds and (b) a subset of compounds associated with selected mechanisms of action.

**Table 2**
Top-$k$ accuracy [%] of compound classification achieved by the tested methods on the subset of cytotoxic and non-cytotoxic compounds. The bottom part shows results achieved by the baseline method when processing only one concentration and both exposure times. The values represent the mean and standard deviation computed over six cross-validation runs.

| Approach | Cytotoxic (225 drugs) | | | Non-cytotoxic (1055 drugs) | | |
|---|---|---|---|---|---|---|
| | Top-1 | Top-5 | Top-10 | Top-1 | Top-5 | Top-10 |
| Proposed | 86.74 ± 0.78 | 97.70 ± 0.35 | 99.31 ± 0.35 | 45.50 ± 0.92 | 70.16 ± 0.71 | 80.26 ± 0.52 |
| AMIL | 82.96 ± 1.10 | 96.99 ± 0.48 | 99.21 ± 0.22 | 39.02 ± 0.99 | 63.74 ± 0.48 | 74.49 ± 0.36 |
| MI-Net | 83.80 ± 0.57 | 97.48 ± 0.51 | 99.53 ± 0.22 | 40.18 ± 1.14 | 64.56 ± 1.38 | 75.85 ± 1.21 |
| mi-Net | 63.38 ± 1.94 | 87.14 ± 1.93 | 93.75 ± 0.82 | 22.65 ± 0.45 | 42.27 ± 0.57 | 52.31 ± 0.93 |
| Baseline | 83.63 ± 1.31 | 97.31 ± 0.40 | 99.31 ± 0.29 | 39.25 ± 0.65 | 65.34 ± 0.53 | 76.99 ± 0.80 |
| Baseline (0.5 μM) | 27.41 ± 2.33 | 57.85 ± 1.59 | 74.35 ± 1.45 | 7.82 ± 0.49 | 24.47 ± 0.96 | 35.35 ± 0.72 |
| Baseline (5 μM) | 46.17 ± 1.34 | 77.51 ± 1.20 | 88.22 ± 0.41 | 11.05 ± 0.62 | 27.83 ± 0.73 | 38.67 ± 0.62 |
| Baseline (50 μM) | 73.98 ± 1.05 | 93.19 ± 0.66 | 96.81 ± 0.44 | 32.45 ± 0.74 | 53.41 ± 0.85 | 63.26 ± 0.84 |

better in the baseline method and even this result was outperformed by our group method. This indicates that there is no configuration that would be uniformly better for all compounds. This trend held true even when considering cytotoxic and non-cytotoxic compounds separately (see Table 2).

Fig. 8 illustrates that the optimal configuration (concentration and exposure time) to identify a particular compound varies. We collected the image-level predictions (1) and used them to compute the $F_1$-scores for each concentration separately. We observe that there are, for example, compounds that are much better recognized at low concentration than at high concentration and vice versa. Moreover, we see that all three concentrations and both times worked best for at least some compounds, i.e., all concentrations are useful. Note also that the group classification performance is always better than the performance at the best single configuration.

Similarly, Table 3 evaluates each configuration in terms of the relative number of compounds best recognized at that configuration according to the $F_1$-score. We see that most compounds (63%) were best recognized at the highest concentration of 50 μM. In terms of exposure time, both 24 and 72 h were optimal for approximately half of the compounds.

**Table 3**
Proportion of compounds best recognized at each configuration according to $F_1$-score.

| Time [h] | Concentration [μM] | | | Total |
|---|---|---|---|---|
| | 0.5 | 5 | 50 | |
| 24 | 7 % | 7 % | 29 % | 42 % |
| 72 | 10 % | 14 % | 35 % | 58 % |
| Total | 17 % | 20 % | 63 % | 100 % |

### 5.5. Mechanism of action

In drug development, the mechanism of action (MoA), i.e., how the compound influences the cell, is often more relevant than the identity of an individual compound. MoA classification should also generalize better to unseen compounds. For example, the goal might be to discover new compounds that act through the same MoA as existing compounds but with fewer side effects. We only report the results of the proposed method, which performed best in the previous experiments. We used a subset of the LOPAC dataset containing the 247 compounds for which the MoA annotations were available (see Section 3.3.2). Please note that the classes are highly imbalanced (see Fig. 9).

We kept the encoder $f_\theta$, trained on the compound prediction task (Section 5.2), frozen. The classifier $g$ was retrained from scratch, once with $K = 247$ outputs to classify the compounds and once with $K = 138$ outputs for MoA classification. We used cross-validation with

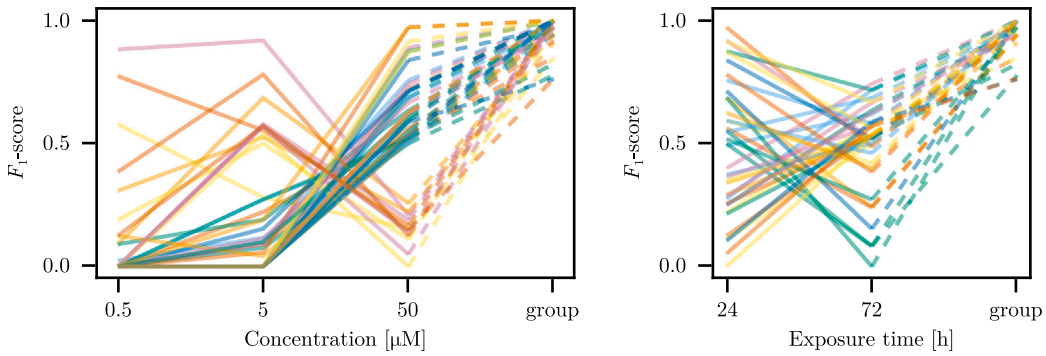



**Fig. 8.** $F_1$-scores obtained by the proposed method for individual compounds at different configurations (using Eq. (1) for prediction) or by considering whole groups (using Eq. (2)). Each line corresponds to a single drug. When plotting the results for a concentration, we always plot the highest $F_1$-score associated with that concentration and any exposure time; similarly for exposure times. For visual clarity, we plot only the compounds where $F_1 \geq 0.5$ at any configuration and where the difference between the plotted $F_1$-scores is at least 0.1.

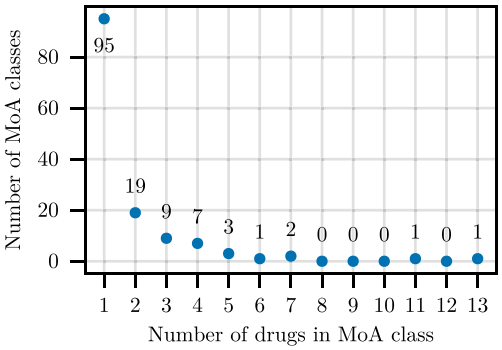



**Fig. 9.** Summary of the numbers of compounds belonging to a MoA class. Out of the total of 138 MoA classes, 95 contain just a single drug. On the other hand, only two MoA classes contain more than seven compounds.

identical splits as in the compound prediction task, guaranteeing that the encoder had not been exposed to the test data during training.

The MoA classifier achieved an accuracy of 60.01% ± 1.34%. Interestingly, predicting compounds on the same data resulted in a higher accuracy of 68.40% ± 1.64%, despite the larger number of classes (247 compound classes vs. 138 MoA classes). This suggests that predicting MoA is more challenging than predicting individual compounds, possibly because the mapping between compounds and MoAs is often not clearly defined and majority of small molecules are known to have multiple effects (Lin et al., 2019). Nevertheless, the MoA performance demonstrates the usefulness and generalizability of the encoder.

### 5.6. Treatment recognition

In this experiment, we addressed the binary classification task of predicting whether a group of images was drug-treated or consisted entirely of untreated controls. Similarly to the previous experiment, we employed a frozen encoder $f_\theta$ from the compound classification task (see Section 5.5) and only trained a classifier $g_\varphi$ (with $K = 2$). The model achieved a very high testing accuracy of 91.66% ± 0.61% with the area under the ROC and Precision–Recall curves of 95.46% and 99.06%, respectively (see Fig. 10).

As in the compound prediction task (see Section 5.2), we wanted to see whether the classification results depended on the tested compound. To this end, we computed the number of false negatives (FNs) for each compound, i.e., the number of groups associated with the compound that were predicted as controls. Since the number of groups per compound is constant (18), FN is complementary to the number of true positives (TP), i.e. TP + FN = 18 for each class.

For the vast majority of compounds (1133 out of 1280), there were two or fewer FNs (see Fig. 11). Only for 15 compounds the model produced more than five FNs, with a maximum of nine.

### 5.7. Classification of JUMP target compounds

In the final experiment, we evaluated compound classification on the JUMP Target dataset (Section 3.4), which is distinct from LOPAC. We reused the encoder trained on the LOPAC dataset and only trained a new classifier with $K = 302$ classes from scratch. The classifier achieved a top-1 accuracy of 68.23% ± 0.65%, notably higher than the accuracy obtained on the whole LOPAC dataset (∼50%). This result is consistent with expectations, as the compounds in the JUMP Target dataset were selected to modulate diverse cellular targets.

#### 5.7.1. Target connectivity analysis

As a preliminary step toward understanding why classification accuracy on the JUMP Target dataset plateaued at approximately 68%, we manually analyzed compound pairs that were frequently misclassified as each other. We examined their annotated protein targets using the STRING database of known and predicted protein–protein interactions (version 12.0 (Szklarczyk et al., 2023)). Target information was retrieved from the dataset annotations and queried in STRING using a high-confidence interaction threshold (0.7). As illustrative examples, we present four misclassified compound pairs in Fig. 12.

In these cases, the predicted protein interaction networks showed that the annotated targets corresponding to the misclassified compound pairs were interlinked and connected to common broader cellular processes through multiple interactions. This suggests that some JUMP Target compounds target proteins involved in multiple cellular signaling pathways and processes. Such target pleiotropy could contribute to overlapping phenotypic outcomes, thereby reducing separability between compounds and limiting classification performance.

To assess whether this observation extends beyond anecdotal examples, we calculated STRING network distance across all compound pairs (see Appendix A.1 for details). Using the co-expression and experimental interaction evidence channels, we compared the distribution of network distances for misclassified pairs against those for all pairs (see Fig. 13). In both channels, misclassified pairs showed a higher tendency towards close proximity in the network. However, the difference with respect to the experimental interaction is small and the co-expression information is only available for a small number of proteins.

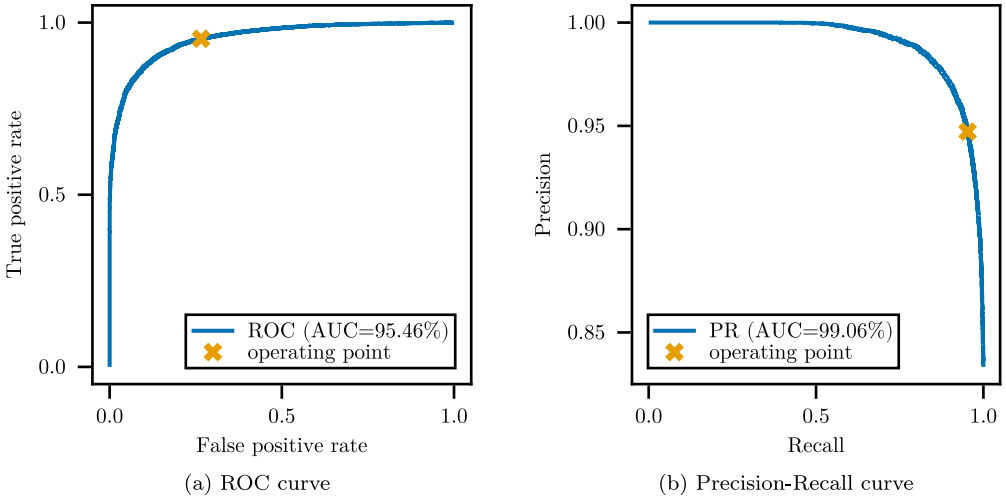
Our results indicate that target network proximity may contribute to classification ambiguity, but additional factors beyond direct target connectivity are likely required to fully explain the performance limits observed on the JUMP Target dataset.

**Table 4**
Classification accuracy [%] for groups with perturbed images using either Gaussian blur or noise.

| Approach | Original | Gauss. blur (kernel size) | | | Gauss. noise (std. dev.) | | |
|---|---|---|---|---|---|---|---|
| | | 3 | 5 | 11 | 0.1 | 0.2 | 0.5 |
| Proposed | 50.0 ± 1.0 | 49.8 ± 0.7 | 49.5 ± 0.6 | 33.6 ± 1.6 | 49.8 ± 0.9 | 48.7 ± 1.0 | 33.2 ± 0.4 |
| AMIL | 43.8 ± 1.0 | 43.7 ± 1.0 | 43.5 ± 1.1 | 30.1 ± 1.5 | 43.7 ± 1.0 | 43.0 ± 0.9 | 30.0 ± 0.5 |
| MI-Net | 43.1 ± 1.0 | 42.8 ± 0.9 | 42.7 ± 0.8 | 28.5 ± 0.9 | 42.9 ± 1.0 | 42.1 ± 0.8 | 29.0 ± 0.8 |
| mi-Net | 26.0 ± 0.4 | 25.9 ± 0.5 | 25.8 ± 0.4 | 20.3 ± 1.1 | 25.8 ± 0.5 | 25.7 ± 0.5 | 19.3 ± 0.6 |

**Table 5**
Classification accuracy [%] for groups with different numbers of inserted (duplicate) images.

| Approach | Original | Number of inserted images | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Proposed | 50.0 ± 1.0 | 48.4 ± 0.5 | 44.5 ± 0.5 | 40.7 ± 0.6 | 36.9 ± 0.4 | 33.5 ± 0.5 | 30.9 ± 0.3 |
| AMIL | 43.8 ± 1.0 | 41.1 ± 0.8 | 34.9 ± 0.6 | 29.5 ± 0.8 | 25.1 ± 0.7 | 21.8 ± 1.0 | 19.3 ± 0.7 |
| MI-Net | 43.1 ± 1.0 | 40.8 ± 1.1 | 36.2 ± 0.8 | 31.1 ± 0.8 | 26.9 ± 0.7 | 23.5 ± 0.8 | 20.9 ± 0.6 |
| mi-Net | 26.0 ± 0.4 | 26.0 ± 0.4 | 25.9 ± 0.5 | 25.9 ± 0.4 | 25.9 ± 0.4 | 25.9 ± 0.4 | 25.8 ± 0.4 |




(a) ROC curve (b) Precision-Recall curve

**Fig. 10.** ROC and precision–recall curves for the binary treatment recognition, where the positive and negative class is *treated* and *control*, respectively. The operating point corresponds to the probability threshold of 1/2.

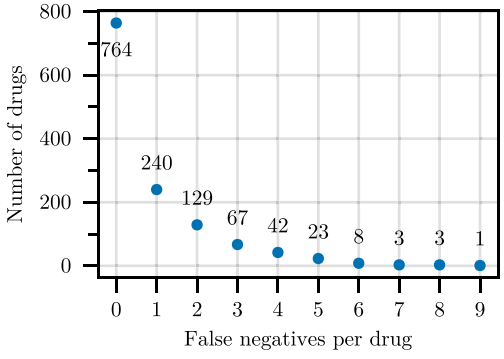



**Fig. 11.** Histogram of false negatives per compound for the treatment recognition. Each compound has 18 examples in total.

### 5.8. Robustness

We evaluated the robustness of the tested methods to input perturbations. Each input image was perturbed either by applying Gaussian blur or by adding Gaussian noise. For Gaussian blur, we considered three kernel sizes (3, 5, and 11 pixels), while Gaussian noise was applied with one of three standard deviations (0.1, 0.2, and 0.5). The standard deviation of pixel intensities prior to perturbation was approximately 1. All models were evaluated on the compound classification task described in Section 5.2 without fine-tuning for the applied perturbations.

As expected, classification accuracy decreased with increasing perturbation strength for all methods (see Table 4). The relative performance degradation was comparable across methods, with the exception of mi-Net, which exhibited low accuracy already on the unperturbed inputs.

Next, following a similar experimental setting, we evaluated robustness to repeated inclusion of a low-information image within each group. Beginning with the original groups of six images, we progressively increased the group size by duplicating a single image, corresponding to the least informative configuration 24 h and 0.5 μM (see Table 3), inserting one duplicate at a time for up to six insertions (i.e., doubling the group size).

As the number of duplicated images increased, the classification performance of all evaluated methods deteriorated (see Table 5). Nevertheless, the proposed method consistently achieved the highest classification accuracy across all group sizes and exhibited the smallest relative performance degradation compared to its original accuracy.

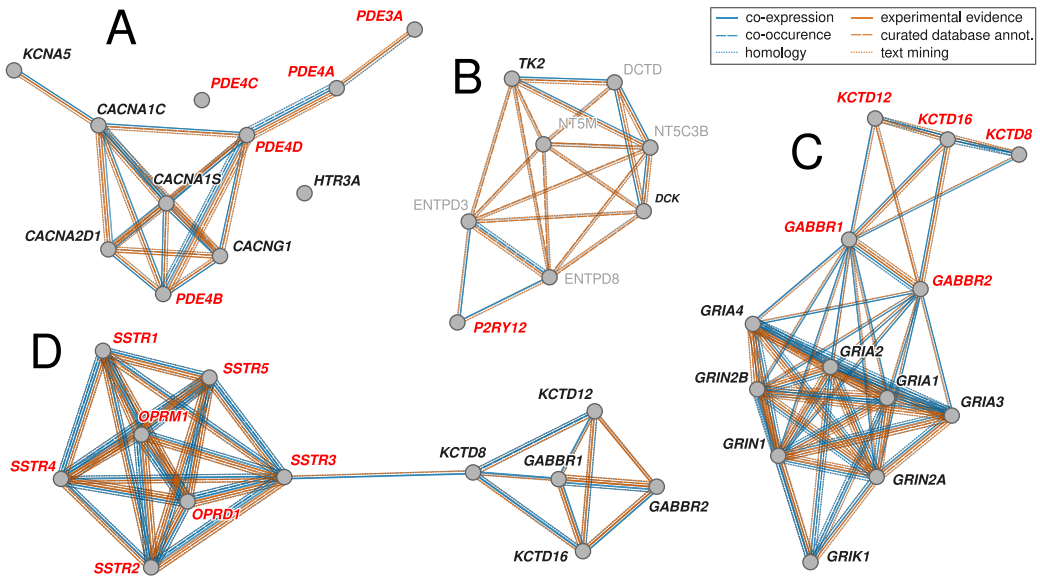
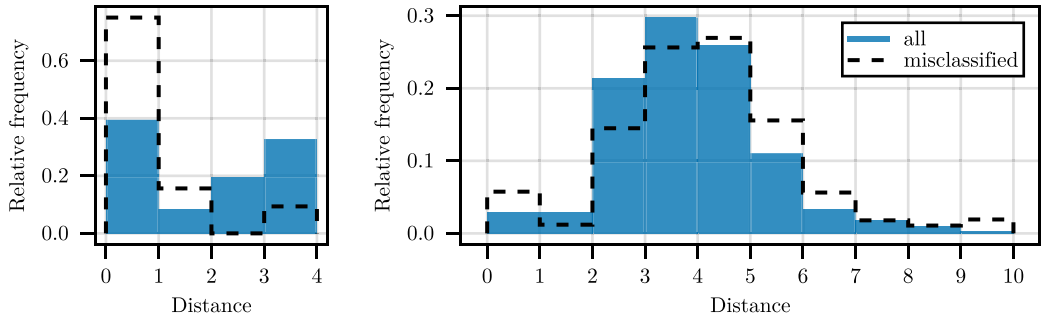



**Fig. 12.** STRING protein–protein interaction networks for representative misclassified JUMP Target compounds. Networks show annotated targets for four compound pairs: (A) *Diltiazem — Ro-20-1724*, (B) *Torcitabine — AZD1283*, (C) *NBQX — Baclofen*, and (D) *Saclofen — Somatostatin*. Targets for the first and second compound in a pair are highlighted in black and red, respectively. In each case, the targets form interconnected clusters. Edges represent protein–protein associations supported by different evidence channels in STRING.




(a) Co-expression (b) Experimental

**Fig. 13.** Histogram of distances in STRING network according to two channels, *co-expression* and *experimental*, between all compound pairs and misclassified compound pairs.

Specifically, with six inserted images, the classification accuracy of the proposed method retained 61% of its original value, whereas AMIL and MI-Net retained only 44% and 49%, respectively.

## 6. Discussion

The primary focus of this work was compound classification directly from digital phase-contrast microscopy images. We demonstrated that a model trained specifically for this task could be repurposed for various related applications, including mechanism-of-action prediction (Section 5.5), treated-versus-control recognition (Section 5.6), and classification of a completely independent set of compounds (Section 5.7). This transferability demonstrates that the learned vector representations capture biologically meaningful features rather than dataset-specific artifacts. Such an observation is encouraging for the long-term goal of analyzing the effects of previously unseen compounds, which could ultimately revolutionize visual *in vitro* drug screening.

Classification difficulty varied across compounds (Section 5.2). To understand the factors influencing classification difficulty, we compared performance with cell viability data from the PRISM assay (Corsello et al., 2020). Compounds with substantial negative effects on cell viability or proliferation (low LFC values) tended to be classified more accurately (Section 5.3). To investigate misclassifications among compounds that were not cytotoxic, we analyzed their MoA classes (Fig. 7(b). L-type calcium channel blockers and dopamine receptor antagonists clustered in regions with higher F1 scores, indicating that these MoAs are more readily recognized even in the absence of strong proliferation effects. This is consistent with the observation that L-type calcium channel blockers, even when non-cytotoxic—or often even pro-proliferative—induce morphological changes, including cytoskeletal remodeling, altered cell shape, and organelle reorganization, through modulation of intracellular calcium signaling (Jacquemet et al., 2016; Timar et al., 1992). Such phenotypic alterations can be captured by a CNN-based analysis, explaining why these non-cytotoxic compounds can still be classified accurately. Dopamine receptor antagonists, which are also highly distinguishable in our analysis, are known to exert diverse effects on cancer cells (Chiang et al., 2022; Consoli et al., 2025; Park et al., 2025), potentially giving rise to heterogeneous but still detectable phenotypic signatures. Together, these results suggest that distinct morphological signatures associated with a compound's MoA may contribute to classification performance, beyond simple effects on proliferation. Nevertheless, cytotoxicity emerged as

a key factor: prediction accuracy on a subset of 225 cytotoxic compounds (87% Top-1, 98% Top-5) was substantially higher than for non-cytotoxic compounds and approached the accuracy reported in prior work conducted in a simpler setting with only four compounds and drug-specific configurations (98% Top-1) (Baručić et al., 2022). Accurate recognition of cytotoxic compounds is consistent with the strong morphological disruptions they induce, such as proliferation arrest, cell shrinkage, detachment, fragmentation, or loss of confluency. In contrast, non-cytotoxic compounds often act through subtler mechanisms that may not generate distinct morphological signatures, e.g., signaling modulation or metabolic interference.

Distinguishing treated from untreated samples performed very well, with an accuracy of 92% (Section 5.6). For context, this exceeds reported single-cell treatment recognition accuracies on three fluorescence datasets (94%, 85%, and 81%; average 87%) (Wiggins et al., 2023). Such a predictor is particularly useful for primary phenotypic screening, where the main question is whether a compound has any detectable effect on cells.

Mechanism-of-action (MoA) classification was more challenging than compound identity prediction (Section 5.5). Besides technical challenges, the biology itself introduces ambiguity: many compounds act pleiotropically on multiple pathways and targets, and thus MoA labels are often based on dominant or historical definitions rather than strict molecular exclusivity. For example, *5-fluorouracil* is annotated as a *thymidylate synthase inhibitor*, but it also interferes with RNA and DNA metabolism through incorporation into nucleic acids and interactions with other proteins (Longley et al., 2003). Such multifunctional activity leads to heterogeneous phenotypic outcomes that cannot be mapped cleanly onto a single MoA. Conversely, compounds with distinct molecular targets may converge on shared outcomes such as apoptosis or cell cycle arrest. Nevertheless, the achieved MoA classification accuracy (60%) exceeded that of a recent study (45%) based on fluorescence images (Stossi et al., 2024). Kensert et al. (2019) reported a prediction accuracy of 95%–97%. However, their study was limited to only twelve MoAs exhibiting distinct *a priori* known visual features in the processed fluorescence images. Cox et al. (2020) also predicted MoAs from fluorescence images but restricted their analysis to active compounds. Despite these restrictions, they could accurately recognize only about half of the MoAs (41 out of 83).

Our STRING-based inspection of frequently misclassified compound pairs (Section 5.7.1) suggests that some JUMP Target compounds act on proteins within interconnected signaling and functional modules. In the examples in Fig. 12, the targets of confused compounds were linked by high-confidence protein–protein interactions and shared cellular processes, consistent with pathway-level or pleiotropic targeting producing overlapping phenotypes and reduced separability. Extending beyond these examples, the network-distance analysis across all compound pairs showed that misclassified pairs are, on average, more closely positioned in STRING than arbitrary pairs, although the difference between the distributions was small. This indicates that target network proximity contributes to classification ambiguity but is not a dominant or sufficient explanation for the ≈68% performance plateau. Together, these findings suggest that while pathway connectivity and shared biological context shape confusion patterns, more research is required to explain the remaining performance limits on the JUMP Target dataset.

We confirmed a previous hypothesis (Baručić et al., 2022) that no single optimal concentration or exposure time exists for identifying all compounds. Both too high and too low compound concentrations can obscure phenotypic signatures. Similarly, some compounds act only after prolonged exposure, while others act rapidly. Our results show that grouping across multiple concentrations and exposure times consistently outperforms single-concentration or single-configuration analysis, which is in agreement with fluorescence-based findings on 15 reporter cell lines (Cox et al., 2020). This reflects the biological reality that drug-induced phenotypes emerge progressively rather than at discrete thresholds. Future studies should examine whether expanding the range of experimental conditions, e.g., more concentrations, exposure times, or using multiple cell lines of different histogenetic origin and bearing different genetic alterations, could further improve classification performance.

From a machine learning perspective, our contribution is a simple yet effective group-based classification method, which yielded consistent improvements over existing, more complex alternatives. Interestingly, even the baseline method matched the performance of these more complex alternatives. Since the baseline method employed the same product-based aggregation for inference as our approach, this result highlights the suitability of the product-based aggregation for the problem at hand. Despite the limited number of training examples per class (15 groups per compound), our method was able to achieve high performance. The robustness of the proposed method to input perturbations (blurring and noise addition) is comparable to that of the alternative methods. Sensitivity to such perturbations seems independent of the aggregation strategy and is likely attributable to the feature encoder. In contrast, the proposed product-based aggregation strategy proves beneficial when handling groups containing a high proportion of low-information images.

Finally, our study underscores the broader advantages of digital phase-contrast imaging. Biologically, the method is non-invasive, preserves cell physiology, and enables longitudinal imaging of the same treated cells — a capability not offered by endpoint fluorescence assays. It also offers significant economy and efficiency gains: a full 384-well plate can be imaged in about 45 min, compared with three to five hours reported for fluorescence-based campaigns (Stossi et al., 2024). This reduction in acquisition time directly alleviates one of the primary bottlenecks in high-throughput screening. We emphasize that comparisons to fluorescence-based methods in this work are indirect and rely on published results rather than experiments conducted under identical conditions.

Despite strong performance, challenges remain for future work. Dataset diversity was limited to two libraries, and further studies on larger and more heterogeneous compound collections are needed to confirm generalizability. While the learned representations are biologically meaningful, their interpretability is limited. Future work should explore integrating phase-contrast imaging with complementary modalities (e.g., transcriptomics or proteomics), expanding the range of experimental conditions, and developing interpretable ML approaches to link features more directly to mechanisms.

## 7. Conclusion

We present a deep learning-based method for identifying compounds applied to cultured cells from digital phase-contrast microscopy images. The method achieves high classification accuracy on more than 1000 compounds from a large library, demonstrating its suitability for high-throughput phenotypic drug screening.

A key innovation of our approach is the use of image groups, where each group combines multiple concentrations and exposure times of a given compound. This formulation reflects the biological reality that compound responses are both concentration- and time-dependent.

Beyond compound identity, the learned feature encoder generalizes to related tasks, including MoA prediction and treated-versus-control classification, as well as classification of an independent compound library. This transferability indicates that the model captures biologically meaningful features rather than dataset-specific artifacts.

Despite relying solely on label-free digital phase-contrast imaging, our approach achieved performance that matched or exceeded the performance reported by prior fluorescence-based methods across several tasks. These results highlight the potential of digital phase-contrast imaging, when paired with appropriate machine learning models, as a viable, scalable, cost-effective, and non-invasive alternative for phenotypic screening.

## CRediT authorship contribution statement

**Denis Baručić:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Jarmila Stanková:** Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Jan Kybic:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Marián Hajdúch:** Conceptualization, Funding acquisition, Interpretation, Supervision, Writing – review & editing. **Petr Džubák:** Conceptualization, Funding acquisition, Supervision, Resources, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by infrastructural projects *CZ-OPENSCREEN* (LM2023052) and *EATRIS-CZ* (LM2023053); project *National Institute for Cancer Research* (Program EXCELES, Project No. LX22NPO5102) funded by the European Union; *Next Generation EU* from the Ministry of Education, Youth and Sports of the Czech Republic (MEYS); the *SALVAGE* project (CZ.02.01.01/00/22_008/0004644) supported by OP JAK, with co-financing from the EU and the State Budget; and project *Research Center for Informatics* (CZ.02.1.01/0.0/0.0/16_019/0000765) funded by OP VVV. The authors thank Soňa Gurská for performing the MTS cytotoxicity assays, which provided the cytotoxicity annotations used in this study.

Denis Baručić would like to thank Vít Filo for his valuable input during the development of the proposed method.

## Appendix A. Evaluation metrics

Consider a sample of $m$ labeled groups $\{\mathbf{X}^j, y^j\}_{j=1}^m$, where $y^j$ is the true label. Let $T_k^j$ denotes a sequence of predicted labels, with $T_1^j = \tilde{y}^j$ being the most likely label, $T_2^j$ the second most likely label, etc. The notation $[\![\cdot]\!]$ is the Iverson bracket. For a binary problem, we assume $y = 0$ for the negative class and $y = 1$ for the positive class.

| Problem | Metric | Formula |
|---|---|---|
| General | Accuracy (Top-1 accuracy) | $\mathrm{Acc} = \frac{1}{m}\sum_{j=1}^{m}[\![y^j = \tilde{y}^j]\!]$ |
| | Top-$k$ accuracy | $\mathrm{Acc}_k = \frac{1}{m}\sum_{j=1}^{m}[\![y^j \in T_k^j]\!]$ |
| Binary | Precision | $\mathrm{Prec} = \frac{\sum_{j=1}^{m}[\![y^j = 1 \wedge \tilde{y}^j = 1]\!]}{\sum_{j=1}^{m}[\![\tilde{y}^j = 1]\!]}$ |
| | Recall (True positive rate) | $\mathrm{Rec} = \frac{\sum_{j=1}^{m}[\![y^j = 1 \wedge \tilde{y}^j = 1]\!]}{\sum_{j=1}^{m}[\![y^j = 1]\!]}$ |
| | False positive rate | $\mathrm{FPR} = \frac{\sum_{j=1}^{m}[\![y^j = 0 \wedge \tilde{y}^j = 1]\!]}{\sum_{j=1}^{m}[\![y^j = 0]\!]}$ |
| Multiclass | Precision (for class $l$) | $\mathrm{Prec}(l) = \frac{\sum_{j=1}^{m}[\![y^j = l \wedge \tilde{y}^j = l]\!]}{\sum_{j=1}^{m}[\![\tilde{y}^j = l]\!]}$ |
| | Recall (for class $l$) | $\mathrm{Rec}(l) = \frac{\sum_{j=1}^{m}[\![y^j = l \wedge \tilde{y}^j = l]\!]}{\sum_{j=1}^{m}[\![y^j = l]\!]}$ |
| | $F_1$-Score (for class $l$) | $F_1(l) = \frac{2 \cdot \mathrm{Prec}(l) \cdot \mathrm{Rec}(l)}{\mathrm{Prec}(l) + \mathrm{Rec}(l)}$ |

### A.1. STRING network distance

Each JUMP Target compound $k$ is associated with a set of target proteins $P_k$. We computed the distance between two compounds $k$ and $l$ as

$$\min_{p_k \in P_k} \min_{p_l \in P_l} d(p_k, p_l),$$

where $d(p_k, p_l)$ is the number of edges between proteins $p_k$ and $p_l$ in the STRING network. We considered a high-confidence interaction threshold (>700/1000) and employed the `stringdb` library for Python to query the STRING network. Compound pairs for which no connecting path was present in the STRING network were excluded from the analysis.

## Appendix B. Data preparation

### B.1. Cell culture and treatments

The U2OS cell line was obtained from ATCC and cultured in McCoy's 5 A Medium supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin. Cells were routinely screened for mycoplasma contamination via qPCR. For experiments, cells were seeded in 384-well PhenoPlate™ (Revvity, cat 6057328) at a density of 1000 cells per well in 30 µL and allowed to settle at room temperature for 30 min before being transferred to a rotating incubator at 37 °C with 5% $CO_2$. After 24 h, without media exchange, cells were treated with the reference compounds at three concentrations: 50 µM, 5 µM, and 0.5 µM. All compounds were reconstituted in DMSO at stock concentrations of 10 mM or 1 mM. The first and last two columns of each source plate were reserved for DMSO controls. Compounds and DMSO were dispensed at the desired concentrations in randomized order using a Labcyte Echo 550 acoustic liquid handler to minimize positional effects. Reference compounds were sourced from the LOPAC1280 library (LO1280; Sigma-Aldrich). Each experimental batch also included a plate treated with JUMP Target compounds purchased from Spec with product number CR-12417. While the JUMP consortium recommends a standard concentration of 5 µM, compounds were additionally screened at 50 µM and 0.8 µM to allow direct comparison with the LOPAC dataset. As with the LOPAC plates, the JUMP source plate was reconstituted in DMSO at a stock concentration of 10 mM and included DMSO controls distributed across the plate and compounds were dispensed in randomized order.

### B.2. Digital phase-contrast imaging

Plates were imaged at two timepoints, 24 h and 72 h, to capture early and late responses, using a Yokogawa CV8000 high-throughput spinning disk confocal microscope configured for digital phase contrast. Imaging was conducted with a $20 \times W/1.0$ objective, acquiring a short z-stack (3 z-planes, 1 µm apart) with on-the-fly image merging. Six fields of view (FOVs) were captured per well. Under these settings, a full 384-well plate could be imaged in approximately 45 min, depending on exposure time and the number of FOVs, generating a total file size of approximately 4.5 GB per plate. The spatial resolution of each image was 1000 × 1000 pixels. All imaging steps were automated using a HighResBio robotics platform, which included a fully integrated plate-loading system synchronized with the Yokogawa CV8000 and a SteriStore incubator.

### B.3. Fluorescence imaging

Following the acquisition of digital phase-contrast images at the 72-hour treatment timepoint, the same plates were subsequently stained with the Hoechst 33342 nucleic acid stain (Thermo Fisher, cat. no. 62249) at a final concentration of 2 μM for 20 min. Stained plates were then imaged using a Yokogawa CV8000 high-content imaging system with a 10× objective. A short z-stack (2 z-planes, 1 μm apart) was acquired with on-the-fly image merging. The Hoechst signal was excited with a 405 nm laser and detected using a 445/45 nm emission filter. Four fields of view (FOVs) were captured, each covering a different quadrant of the well. Under these settings, a full 384-well plate could be imaged in approximately 20 min, depending on exposure time and number of FOVs, with each plate generating approximately 11.5 GB of image data. All imaging steps were automated using a HighResBio robotics platform, which included a fully integrated plate-loading system synchronized with the Microplate Sample Processor (BioTek Precision XS), the Yokogawa CV8000, and a SteriStore incubator. Fluorescent image data were imported into Signals Image Artist (Revvity) for analysis, including quantification of nuclear count, size, and intensity, which served as internal quality control metrics for assessing cell health and consistency across plates and batches.

### B.4. Cytotoxicity MTS assay

Cytotoxicity assays were performed using the MTS tetrazolium salt,[5] which serves to quantify the number of metabolically active cells in the culture. Briefly, cells were seeded into 384-well clear-bottom Corning plates using a MultiDrop Combi dispenser (Thermo Fisher Scientific) at a density of 1000 cells per well in 30 μL of medium. The following day, cells were treated with test compounds at a final concentration of 50 μM, along with vehicle control (DMSO) and positive cytotoxic controls—Actinomycin D (2.67 μM) and Mitomycin C (100 μM). Compound addition was performed using an Echo 555 acoustic liquid handler (Labcyte, USA). The same compound source plates were used as in the imaging experiments. After 72 h of compound exposure, 5 μL of MTS reagent (Promega) was added to each well, and plates were incubated for additional 1–4 h. Absorbance was then measured at 490 nm using an EnVision multimode plate reader (PerkinElmer). All experiments were performed in technical duplicates with at least three independent biological replicates. This was a single-concentration primary screen; therefore, $IC_{50}$ values were not determined. Instead, compounds were considered cytotoxic if they showed measurable activity compared to control at 50 μM and non-cytotoxic if no significant activity was observed at that concentration.

## Data availability

All data and analysis code are openly available in our repository and on GitHub. Persistent open links to both are provided in the manuscript.

## References


Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd international conference on learning representations* (pp. 1–15). http://dx.doi.org/10.48550/arXiv.1409.0473.

Baručić, D., Kaushik, S., Kybic, J., Stanková, J., Džubák, P., & Hajdúch, M. (2022). Characterization of drug effects on cell cultures from phase-contrast microscopy images. *Computers in Biology and Medicine*, *151*, Article 106171. http://dx.doi.org/10.1016/j.compbiomed.2022.106171.

Caicedo, J. C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A. S., Barry, J. D., Bansal, H. S., Kraus, O., et al. (2017). Data-analysis strategies for image-based cell profiling. *Nature Methods*, *14*(9), 849–863. http://dx.doi.org/10.1038/nmeth.4397.

Cao, Y., & Fleet, D. J. (2014). Generalized product of experts for automatic and principled fusion of Gaussian process predictions. http://dx.doi.org/10.48550/arXiv.1410.7827, arXiv preprint arXiv:1410.7827.

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, *7*(10), R100. http://dx.doi.org/10.1186/gb-2006-7-10-r100.

Chandrasekaran, S. N., Ackerman, J., Alix, E., Ando, D. M., Arevalo, J., Bennion, M., Boisseau, N., Borowa, A., Boyd, J. D., Brino, L., et al. (2023). JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, http://dx.doi.org/10.1101/2023.03.23.534023.

Cheplygina, V., Tax, D. M., & Loog, M. (2015). On classification with bags, groups and sets. *Pattern Recognition Letters*, *59*, 11–17. http://dx.doi.org/10.1016/j.patrec.2015.03.008.

Chiang, J.-Y., Lei, F.-J., Chang, H.-J., Wei, S.-T., Wang, C.-C., Huang, Y.-C., Wang, H.-L., Chuang, C.-F., Hu, S.-Y., & Hsieh, C.-H. (2022). Haloperidol instigates endometrial carcinogenesis and cancer progression by the NF-κb/CSF-1 signaling cascade. *Cancers*, *14*(13), 3089. http://dx.doi.org/10.3390/cancers14133089.

Consoli, V., Sorrenti, V., Saccone, S., Federico, C., Bruno, F., Cristaldi, D. A., Pittalà, V., Intagliata, S., & Vanella, L. (2025). Haloperidol drug repurposing unveils ferroptosis involvement in breast cancer cells. *Scientific Reports*, *15*(1), 26948. http://dx.doi.org/10.1038/s41598-025-12645-9.

Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. *Nature Medicine*, *23*(4), 405–408. http://dx.doi.org/10.1038/nm.4306.

Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., Humeidi, R., Peck, D., Wu, X., Tang, A. A., et al. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, *1*(2), 235–248. http://dx.doi.org/10.1038/s43018-019-0018-6.

Cox, M. J., Jaensch, S., Van de Waeter, J., Cougnaud, L., Seynaeve, D., Benalla, S., Koo, S. J., Van Den Wyngaert, I., Neefs, J.-M., Malkov, D., et al. (2020). Tales of 1,008 small molecules: phenomic profiling through live-cell imaging in a panel of reporter cell lines. *Scientific Reports*, *10*(1), 13262. http://dx.doi.org/10.1038/s41598-020-69354-8.

Cross-Zamirski, J. O., Mouchet, E., Williams, G., Schönlieb, C.-B., Turkki, R., & Wang, Y. (2022). Label-free prediction of cell painting from brightfield images. *Scientific Reports*, *12*(1), 10001. http://dx.doi.org/10.1038/s41598-022-12914-x.

Dietterich, T. G., Lathrop, R. H., & Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89*(1–2), 31–71. http://dx.doi.org/10.1016/S0004-3702(96)00034-3.

van Dijk, R., Arevalo, J., Babadi, M., Carpenter, A. E., & Singh, S. (2024). Capturing cell heterogeneity in representations of cell populations for image-based profiling using contrastive learning. *PLoS Computational Biology*, *20*(11), Article e1012547. http://dx.doi.org/10.1371/journal.pcbi.1012547.

Dubach, R. A., & Dubach, J. M. (2024). Autocorrelation analysis of a phenotypic screen reveals hidden drug activity. *Scientific Reports*, *14*(1), 10046. http://dx.doi.org/10.1038/s41598-024-60654-x.

Essa, E., & Xie, X. (2018). Phase contrast cell detection using multilevel classification. *International Journal for Numerical Methods in Biomedical Engineering*, *34*(2), Article e2916. http://dx.doi.org/10.1002/cnm.2916.

Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., & Zhang, X. (2017). A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, *33*(13), 2010–2019. http://dx.doi.org/10.1093/bioinformatics/btx069.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

Hecht, E. (2017). Optics, (5). (pp. 646–650). Pearson Education.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–1800. http://dx.doi.org/10.1162/089976602760128018.

Ilse, M., Tomczak, J., & Welling, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of machine learning research: vol. 80, Proceedings of the 35th international conference on machine learning* (pp. 2127–2136). PMLR.

Jacquemet, G., Baghirov, H., Georgiadou, M., Sihto, H., Peuhu, E., Cettour-Janet, P., He, T., Perälä, M., Kronqvist, P., Joensuu, H., et al. (2016). L-type calcium channels regulate filopodia stability and cancer cell invasion downstream of integrin signalling. *Nature Communications*, *7*(1), 13297. http://dx.doi.org/10.1038/ncomms13297.

Kang, M., & Kim, J. (2025). Cancer cell classification based on morphological features of 3D phase contrast microscopy using deep neural network. *IEEE Access*, http://dx.doi.org/10.1109/ACCESS.2025.3547711.

Kensert, A., Harrison, P. J., & Spjuth, O. (2019). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS Discovery*, *24*(4), 466–475. http://dx.doi.org/10.1177/2472555218818756.


[5] 3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium.

Kikuchi, Y., Okuhashi, Y., Ishihata, H., Kashiba, M., & Sasaki, S. (2024). AI-based apoptosis cell classification using phase-contrast images of K562 cells. *Anticancer Research*, *44*(3), 935–939. http://dx.doi.org/10.21873/anticanres.16888.

Ko, B., Gim, J., & Nam, J. (2011). Cell image classification based on ensemble features and random forest. *Electronics Letters*, *47*(11), 638–639. http://dx.doi.org/10.1049/el.2011.0831.

Krentzel, D., Shorte, S., & Zimmer, C. (2023). Deep learning in image-based phenotypic drug discovery. *Trends in Cell Biology*, *33*(7), 538–554. http://dx.doi.org/10.1016/j.tcb.2022.11.011.

Kusumoto, D., Seki, T., Sawada, H., Kunitomi, A., Katsuki, T., Kimura, M., Ito, S., Komuro, J., Hashimoto, H., Fukuda, K., & Yuasa, S. (2021). Anti-senescent drug screening by deep learning-based morphology senescence scoring. *Nature Communications*, *12*(1), 257. http://dx.doi.org/10.1038/s41467-020-20213-0.

Lang, P., Yeow, K., Nichols, A., & Scheer, A. (2006). Cellular imaging in drug discovery. *Nature Reviews Drug Discovery*, *5*(4), 343–356. http://dx.doi.org/10.1038/nrd2008.

Li, Y., Di, J., Wang, K., Wang, S., & Zhao, J. (2020). Classification of cell morphology with quantitative phase microscopy and machine learning. *Optics Express*, *28*(16), 23916–23927. http://dx.doi.org/10.1364/OE.397029.

Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. L., Sausville, E. L., Lukow, D. A., Liu, L., Chait, A. R., et al. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, *11*(509), Article eaaw8412. http://dx.doi.org/10.1126/scitranslmed.aaw8412.

Longley, D. B., Harkin, D. P., & Johnston, P. G. (2003). 5-fluorouracil: mechanisms of action and clinical strategies. *Nature Reviews Cancer*, *3*(5), 330–338. http://dx.doi.org/10.1038/nrc1074.

Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International conference on learning representations* (pp. 1–16).

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations* (pp. 1–18).

Meng, N., Lam, E. Y., Tsia, K. K., & So, H. K.-H. (2018). Large-scale multi-class image-based cell classification with deep learning. *IEEE Journal of Biomedical and Health Informatics*, *23*(5), 2091–2098. http://dx.doi.org/10.1109/JBHI.2018.2878878.

Mertanová, H., Kybic, J., Stanková, J., Džubák, P., & Hajdúch, M. (2022). Learning to segment cell nuclei in phase-contrast microscopy from fluorescence images for drug discovery. *12032*, In *Medical imaging 2022: image processing* (pp. 688–694). SPIE, http://dx.doi.org/10.1117/12.2607500.

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., & Van Valen, D. (2019). Deep learning for cellular image analysis. *Nature Methods*, *16*(12), 1233–1246. http://dx.doi.org/10.1038/s41592-019-0403-1.

Moshkov, N., Becker, T., Yang, K., Horvath, P., Dancik, V., Wagner, B. K., Clemons, P. A., Singh, S., Carpenter, A. E., & Caicedo, J. C. (2023). Predicting compound activity from phenotypic profiles and chemical structures. *Nature Communications*, *14*(1), 1967. http://dx.doi.org/10.1038/s41467-023-37570-1.

Niioka, H., Asatani, S., Yoshimura, A., Ohigashi, H., Tagawa, S., & Miyake, J. (2018). Classification of C2C12 cells at differentiation by convolutional neural network of deep learning using phase contrast images. *Human Cell*, *31*, 87–93. http://dx.doi.org/10.1007/s13577-017-0191-9.

Park, J., Jang, H. J., Jung, W. K., Kang, D. Y., Gong, Y. L., Kim, H.-J., Kang, J. S., Yang, J. W., Byun, Y., & Park, S.-K. (2025). Suppression of dopamine receptor 2 inhibits the formation of human prostate cancer PC-3-derived cancer stem cell-like cells through AMPK inhibition. *Oncology Letters*, *29*(3), 142. http://dx.doi.org/10.3892/ol.2025.14888.

Perakis, A., Gorji, A., Jain, S., Chaitanya, K., Rizza, S., & Konukoglu, E. (2021). Contrastive learning of single-cell phenotypic representations for treatment classification. In *International workshop on machine learning in medical imaging* (pp. 565–575). Springer, http://dx.doi.org/10.1007/978-3-030-87589-3_58.

Pham, T. D. (2007). Geo-thresholding for segmentation of fluorescent microscopic cell images. In *International conference on mass data analysis of images and signals in medicine, Biotechnology, and chemistry* (pp. 15–26). Springer, http://dx.doi.org/10.1007/978-3-540-76300-0_2.

Prasad, A., & Alizadeh, E. (2019). Cell form and function: interpreting and controlling the shape of adherent cells. *Trends in Biotechnology*, *37*(4), 347–357. http://dx.doi.org/10.1016/j.tibtech.2018.09.007.

Pratapa, A., Doron, M., & Caicedo, J. C. (2021). Image-based cell phenotyping with deep learning. *Current Opinion in Chemical Biology*, *65*, 9–17. http://dx.doi.org/10.1016/j.cbpa.2021.04.001.

Samsudin, N. A., & Bradley, A. P. (2010). Nearest neighbour group-based classification. *Pattern Recognition*, *43*(10), 3458–3467. http://dx.doi.org/10.1016/j.patcog.2010.05.010.

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., et al. (2017). A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, *16*(1), 19–34. http://dx.doi.org/10.1038/nrd.2016.230.

Smith, A. (2002). Screening for drug discovery: the leading question. *Nature*, *418*(6896), 453–455. http://dx.doi.org/10.1038/418453a.

Smith, K., Piccinini, F., Balassa, T., Koos, K., Danka, T., Azizpour, H., & Horvath, P. (2018). Phenotypic image analysis software tools for exploring and understanding big image data from cell-based assays. *Cell Systems*, *6*(6), 636–653. http://dx.doi.org/10.1016/j.cels.2018.06.001.

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., & Goodman, A. (2021). CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics*, *22*, 1–11. http://dx.doi.org/10.1186/s12859-021-04344-9.

Stossi, F., et al. (2024). SPACe: an open-source, single-cell analysis of Cell Painting data. *Nature Communications*, *15*(1), 10170. http://dx.doi.org/10.1038/s41467-024-54264-4.

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, *18*(1), 100–106. http://dx.doi.org/10.1038/s41592-020-01018-x.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, *51*(D1), D638–D646. http://dx.doi.org/10.1093/nar/gkac1000.

Timar, J., Chopra, H., Rong, X., Hatfield, J. S., Fligiel, S. E. G., Onoda, J. M., Taylor, J. D., & Honn, K. V. (1992). Calcium channel blocker treatment of tumor cells induces alterations in the cytoskeleton, mobility of the integrin $\alpha$IIb$\beta$3 and tumor-cell-induced platelet aggregation. *Journal of Cancer Research and Clinical Oncology*, *118*(6), 425–434. http://dx.doi.org/10.1007/BF01629425.

Veselovská, L., Kudlová, N., Gurská, S., Lišková, B., Medvedíková, M., Hodek, O., Tloušťová, E., Milisavljevic, N., Tichý, M., Perlíková, P., Mertlíková-Kaiserová, H., Trylčová, J., Pohl, R., Klepetářová, B., Džubák, P., Hajdúch, M., & Hocek, M. (2020). Synthesis and cytotoxic and antiviral activity profiling of All-Four isomeric series of Pyrido-Fused 7-Deazapurine ribonucleosides. *Chemistry – A European Journal*, *26*(57), 13002–13015. http://dx.doi.org/10.1002/chem.202001124.

Vural, V., Fung, G., Krishnapuram, B., Dy, J., & Rao, B. (2006). Batch classification with applications in computer aided diagnosis. In *Machine learning: ECML 2006* (pp. 449–460). Springer, http://dx.doi.org/10.1007/11871842_43.

Waisman, A., La Greca, A., Möbbs, A. M., Scarafía, M. A., Velazque, N. L. S., Neiman, G., Moro, L. N., Luzzani, C., Sevlever, G. E., Guberman, A. S., et al. (2019). Deep learning neural networks highly predict very early onset of pluripotent stem cell differentiation. *Stem Cell Reports*, *12*(4), 845–859. http://dx.doi.org/10.1016/j.stemcr.2019.02.004.

Wang, X., Yan, Y., Tang, P., Bai, X., & Liu, W. (2018). Revisiting multiple instance neural networks. *Pattern Recognition*, *74*, 15–24. http://dx.doi.org/10.1016/j.patcog.2017.08.026.

Wang, M., Zhou, X., Li, F., Huckins, J., King, R. W., & Wong, S. T. (2008). Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. *Bioinformatics*, *24*(1), 94–101. http://dx.doi.org/10.1093/bioinformatics/btm530.

Wiggins, L., Lord, A., Murphy, K. L., Lacy, S. E., O'Toole, P. J., Brackenbury, W. J., & Wilson, J. (2023). The CellPhe toolkit for cell phenotyping using time-lapse imaging and pattern recognition. *Nature Communications*, *14*(1), 1854. http://dx.doi.org/10.1038/s41467-023-37447-3.

Xing, F., Xie, Y., Su, H., Liu, F., & Yang, L. (2018). Deep learning in microscopy image analysis: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(10), 4550–4568. http://dx.doi.org/10.1109/TNNLS.2017.2766168.

Yao, K., Rochman, N. D., & Sun, S. X. (2019). Cell type classification and unsupervised morphological phenotyping from low-resolution images using deep learning. *Scientific Reports*, *9*(1), 13467. http://dx.doi.org/10.1038/s41598-019-50010-9.

Zagajewski, A., Turner, P., Feehily, C., Al-Jabri, A., Williams, P., Spiteri, C., Fagerberg, P. H., Spjuth, O., Bågeman, D., Nygård, F., Svensson, E., & Vrouwe, E. (2023). Deep learning and single-cell phenotyping for rapid antimicrobial susceptibility detection in Escherichia coli. *Communications Biology*, *6*(1), 1164. http://dx.doi.org/10.1038/s42003-023-05524-4.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., & Smola, A. J. (2017). Deep sets. *Advances in Neural Information Processing Systems*, *30*.

Zernike, F. (1942). Phase contrast, a new method for the microscopic observation of transparent objects part II. *Physica*, *9*(10), 974–986. http://dx.doi.org/10.1016/S0031-8914(42)80079-8.

Zheng, W., Thorne, N., & McKew, J. C. (2013). Phenotypic screens as a renewed approach for drug discovery. *Drug Discovery Today*, *18*(21), 1067–1073. http://dx.doi.org/10.1016/j.drudis.2013.07.001.