Highlights

Characterization of drug effects on cell cultures from phase-contrast microscopy images

Denis Baručić,Sumit Kaushik,Jan Kybic,Jarmila Stanková,Petr Džubák,Marián Hajdúch

- Three topoisomerase inhibitors can be distinguished from phase-contrast microscopy images.
- Phase-contrast images are as good for classification as fluorescence images.
- Classification accuracy is better on images taken 72h after treatment than 24h.
- Classification is better when substance-specific concentrations are used.

Characterization of drug effects on cell cultures from phase-contrast microscopy images

Denis Baručić[a], Sumit Kaushik[a], Jan Kybic[a,*], Jarmila Stanková[b], Petr Džubák[b] and Marián Hajdúch[b]

[a]Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[b]Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czech Republic

ARTICLE INFO

Keywords:
Deep learning
Phase-contrast images
Drug discovery
Anti-cancer drugs
Convolutional neural networks

ABSTRACT

In this work, we classify chemotherapeutic agents (topoisomerase inhibitors) based on their effect on U-2 OS cells.We use phase-contrast microscopy images, which are faster and easier to obtain than fluorescence images and support live cell imaging.We use a convolutional neural network (CNN) trained end-to-end directly on the input images without requiring for manual segmentations or any other auxiliary data. Our method can distinguish between tested cytotoxic drugs with an accuracy of 98%, provided that their mechanism of action differs, outperforming previous work. The results are even better when substance-specific concentrations are used. We show the benefit of sharing the extracted features over all classes (drugs). Finally, a 2D visualization of these features reveals clusters, which correspond well to known class labels, suggesting the possible use of our methodology for drug discovery application in analyzing new, unseen drugs.

1. Introduction

Drug discovery aims to search for effective treatment of diseases with minimum side effects. Machine learning methods [4] can significantly reduce the time, effort and costs involved. Here we shall focus on one step of this process — cellular phenotypic screening, where the effect of a large set of potential candidate chemical compounds is evaluated on standard target cell lines [20, 38]. The goal is to examine as many combinations of chemicals and cell lines as fast as possible. The combinations are contained in so-called 'wells', and a single array may contain hundreds of them. The fastest and least invasive way of evaluating the state of the cells is automatic microscopy imaging. The whole process can be robotized, producing a vast number of microscopic images of cells in particular wells at various time points. Therefore, automated image analysis techniques are necessary to achieve the desired high throughput.

Existing methods are usually based on fluorescence microscopy images (Fig. 1 and 2), which provide very clear images (especially of cell nuclei) that are straightforward to segment and evaluate automatically, using relatively simple methods. However, fluorescence imaging requires additional labeling by fluorescent dye or protein, which increases the cost and processing time and could affect the cellular morphology and the final analytical output [15]. Furthermore, only a limited combination of dyes can be used simultaneously.

As an alternative to fluorescence imaging, we use phase-contrast microscopy images (Fig. 1, 2), which do not damage the cells and can be acquired much easier and faster. However, these images are more challenging to segment (see Fig. 3) and analyze because of the intricate cell appearance and frequent imaging artifacts.

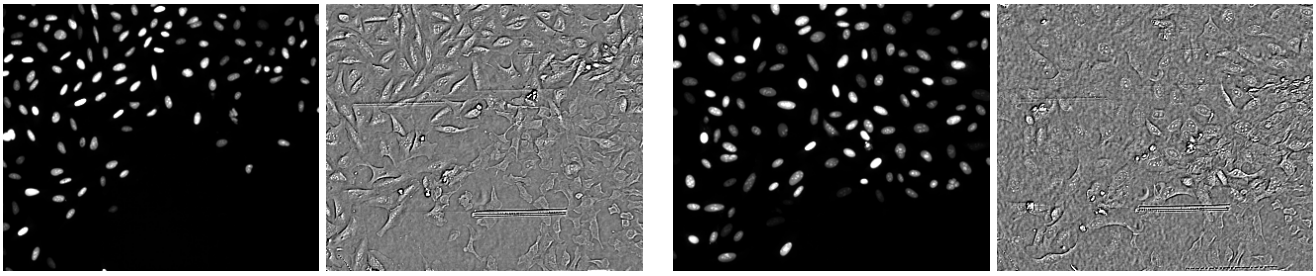
In our previous work [24, 25], we have shown that it is possible to distinguish the effect of several chemical compounds on cell culture from phase-contrast images. However, our procedure was complicated. It used simultaneously acquired geometrically aligned fluorescence and phase-contrast images to learn to 'translate' phase-contrast images to binary segmentations obtained from fluorescence images. These segmentations were then analyzed using classical geometric shape features.

Here, in contrast to the previous work, we analyze the phase-contrast images directly using convolutional neural networks (CNNs), avoiding the necessity of acquiring the paired fluorescence and phase-contrast images and the limitation of considering only the shape of the segmented nuclei. We also show the benefit of sharing the features by formulating the task as a multiclass classification instead of solving independent binary problems separately for each class (chemical compound). This approach leads to a much-improved classification accuracy compared to the work of Mertanova et al. [24, 25]. Moreover, we show visually that clusters of the extracted features correspond to the mechanism of action of the chemical compounds tested. This hints at the generalization ability of these features, which could be used as image-based fingerprints [37].

1.1. Related work

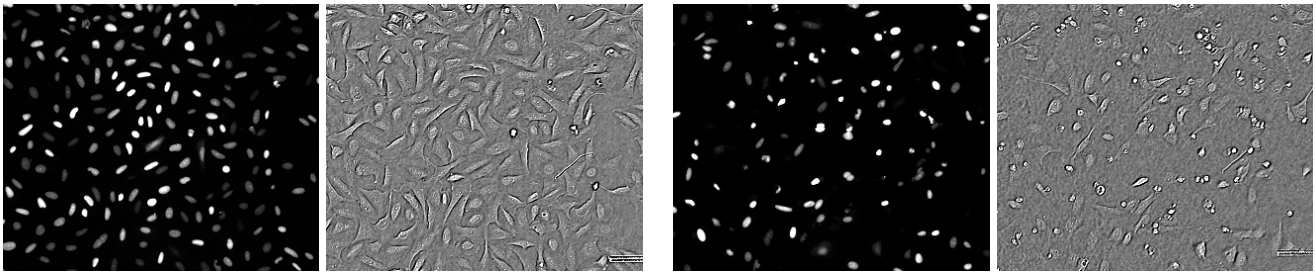
Image-based high-throughput screening for drug discovery has become an established and frequently used technique described in multiple review articles [2, 33, 35]. Classical approaches typically start by segmenting individual cells and evaluating especially the shape features, which describe the cell morphology, a part of the cell phenotype. Cell shape is known to be related to the cell type, state, and other relevant

*Corresponding author
barucden@fel.cvut.cz (D. Baručić); kaushsum@fel.cvut.cz (S. Kaushik); kybic@fel.cvut.cz (J. Kybic)
ORCID(s): 0000-0003-0428-3354 (D. Baručić); 0000-0002-4146-291X (S. Kaushik); 0000-0002-9363-4947 (J. Kybic)



24 hours 72 hours

Figure 1: Example of two pairs of the corresponding fluorescence (left) and phase-contrast (right) images taken 24 hours and 72 hours after being treated with Topotecan.



24 hours 72 hours

Figure 2: Example of two pairs of the corresponding fluorescence (left) and phase-contrast (right) images taken 24 hours and 72 hours after being treated with Etoposide.

properties, such as metastatic capacity [30]. The final step involves machine learning for feature-based classification or clustering. Such pipelines can be implemented using open software such as ImageJ/Fiji [34], Icy [7], CellProfiler [3], and EBImage [28]. Later on, deep learning methods [13] appeared, combining feature extraction and classification for single-cell analysis [8]. An independent segmentation step [12] can be avoided by processing the input images directly [29, 31]. These methods are usually based on well-known neural network architectures from computer vision for image segmentation (e.g., U-Net [32]) and classification (e.g., ResNet [14]). The networks are adapted to microscopic images, for example by adding color normalization [5] and the multi-scale approach [9] to capture both short and long-range patterns.

Note that this work addresses the task of classifying the whole sample (slide or well) into one class. It is also possible to classify individual cells in the image [42], which is outside of this article's scope.

The high-throughput screening for drug discovery described in the aforementioned publications works mainly with fluorescence images [2, 11], which is also the majority modality in dataset repositories such as the Broad Bioimage Benchmark Collection [22] used for performance evaluation. Phase-contrast images are much rarer in high-throughput applications because of their more complicated

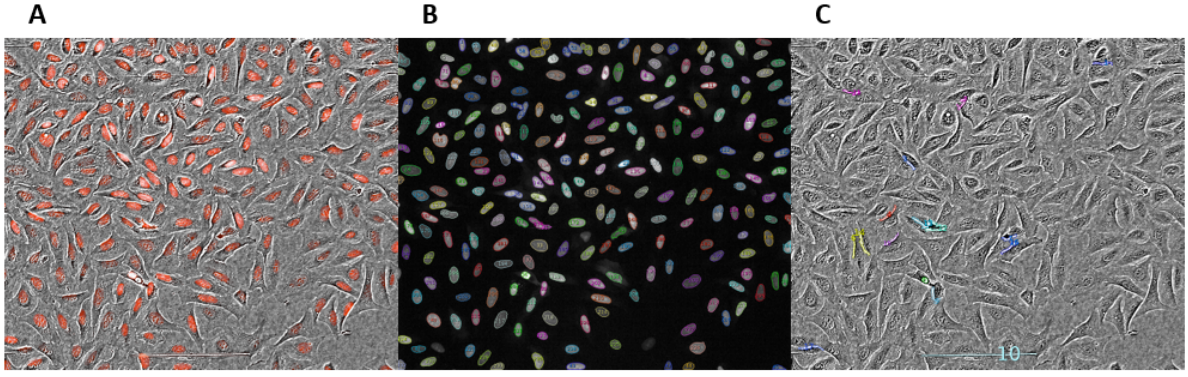


Figure 3: Fluorescence image (in red) overlaid over a phase-contrast image (A). The fluorescence image can be segmented by standard tools, such as the Columbus software by PerkinElmer (B). The same method fails when applied to phase-contrast images (C). In (B,C), segmented objects are individually colored and overlaid over the grayscale input image.

analysis. Nevertheless, there are methods to perform some processing steps, for example segmentation of individual cells [1], mitosis detection [16], classification [26] and segmentation [27] of different cell types, or morphology classification of individual cells [40]. However, we are not aware of any work where phase-contrast cell culture images similar in appearance to ours would be analyzed with the goal of drug discovery.

In our previous work [24, 25], we have taken the roundabout way of learning to transform the phase-contrast images using the pix2pix model [18] to look similar to fluorescence images, which are then straightforward to segment using a U-Net [32]. Shape-based features were extracted from the segmented images and fed to a support vector machine (SVM) [6] for binary classification. The approach required paired fluorescence and phase-contrast images.

2. Data

We are using the same dataset as in [24, 25]: The U-2 OS cell line, derived from human osteosarcoma (American Tissue Culture Collection), was transduced by fluorescent mCherry-NLS (nuclear localization signal, cat. n. 0023VCT, Vectalis-TaKaRa, Japan). The cells were seeded at a density of 1500 cells per well and treated with three topoisomerase inhibitors, Topotecan, Daunorubicin, and Etoposide, at a final concentration 0.5 μM. The treated and control cells were imaged at 20× magnification and sampled at five locations per well and two time points — 24 and 72 hours after the treatment. The images are of size 2560 × 2160 pixels. The original dataset also contained images taken before the treatment. However, we have decided not to use these images to reduce the over-representation of cells with no treatment.

Fluorescence images were acquired in parallel with the digital phase-contrast images. We use them here only for comparison with existing methods.

There are five different image classes. *Topotecan, Daunorubicin*, and *Etoposide* are each applied to 8 wells. The remaining two classes are controls: 8 wells with a 0.05% solution of *DMSO* (Dimethyl Sulfoxide) and 12 wells with no treatment. In total, we have 440 phase-contrast images from 44 wells. Excluding controls, there are 120 images from 24 wells with active treatment for each time point. Finally, skipping images containing less than three cells—mostly due to failed acquisition—leads to class sizes summarized in Table 1.

Table 1
The number of images per class.

Class	Fluorescence	Phase-contrast
Topotecan	79	64
Daunorubicin	78	67
Etoposide	80	71
DMSO	79	74
No treatment	119	116
Total	435	392

3. Method

We formulate the task as a standard multiclass image classification: given an input phase-contrast image, a convolutional neural network (CNN) assigns it to one of the $n = 5$ classes defined above.

3.1. Tiling

The input images are too big and cannot be fed directly to the CNN due to the limited GPU memory. Instead, we uniformly divide each image into $m = 16$ partially overlapping tiles of size 1024×864 pixels (see Sec. 4.1 for experiments with other tile sizes).

Each tile i is processed separately by the CNN, and the resulting tile-wise class probabilities p_k^i are aggregated by averaging to obtain image-wise probabilities

$$\bar{p}_k = \frac{1}{m} \sum_{i=1}^{m} p_k^i. \tag{1}$$

Maximization over the class index leads to the final image-wise prediction

$$k^* = \arg\max_k \bar{p}_k. \tag{2}$$

See Sec. 4.2 for an experimental comparison of this approach with (i) maximum-based aggregation and (ii) no aggregation, considering all tiles independently and sharing the same image label.

3.2. Network architecture

We use the ResNet18 [14] network, known to perform well on a number of tasks. It uses residual blocks with skip connections for regularization and to combat the vanishing gradient problem. In our case, we have used the smaller, 18-layer version due to the limited size of our dataset. The final layer has $n = 5$ outputs, one for each class, to which we apply the softmax transformation [10] to obtain class probability estimates p_k for each class k. Since our batch size is small (5, due to GPU memory limitation), we use instance normalization [19, 41] instead of the more standard batch normalization [17].

Apart from the multiclass approach, we have also trained binary classifiers ($n = 2$) for each class pair (Sec. 4.2 and 4.3). The binary classifiers were either trained in two ways: either from scratch or by fine-tuning the final layer of a pre-trained multiclass classifier.

3.3. Training and evaluation

The network is trained by minimizing the cross-entropy loss using the Adam optimizer. We use class weights inversely proportional to class sizes to account for class imbalances. The initial learning rate $\lambda = 10^{-4}$, determined by the cyclical method [39], is reduced 10× during training whenever the validation loss stops improving. For augmentation, we apply random horizontal and vertical flipping and color adjustments. The best model with respect to the validation loss is used for the final evaluation on the test set.

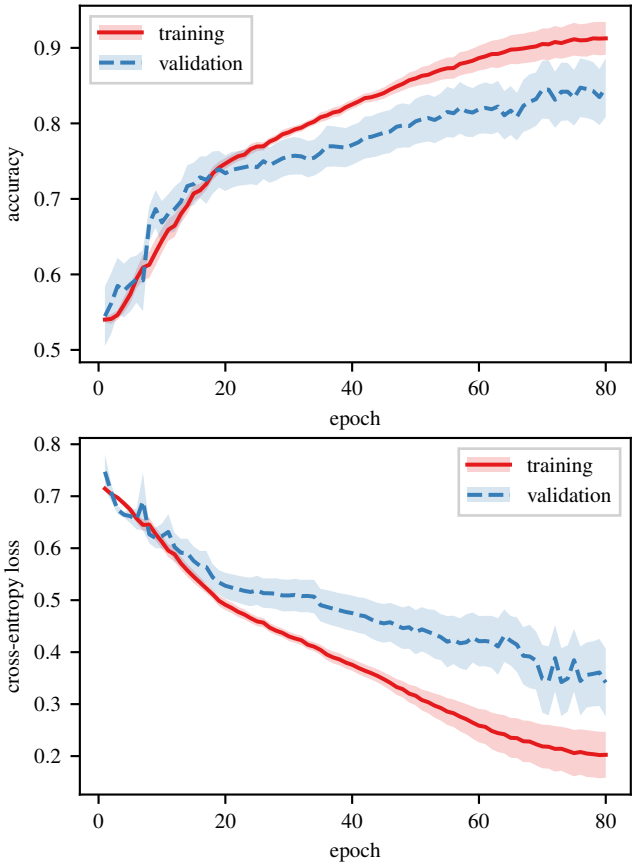


Figure 4: The evolution of the training and validation accuracy and loss during training for the *Etoposide* vs. *Topotecan* classification on phase-contrast images. Mean (solid line) and standard error (shaded region) are shown.

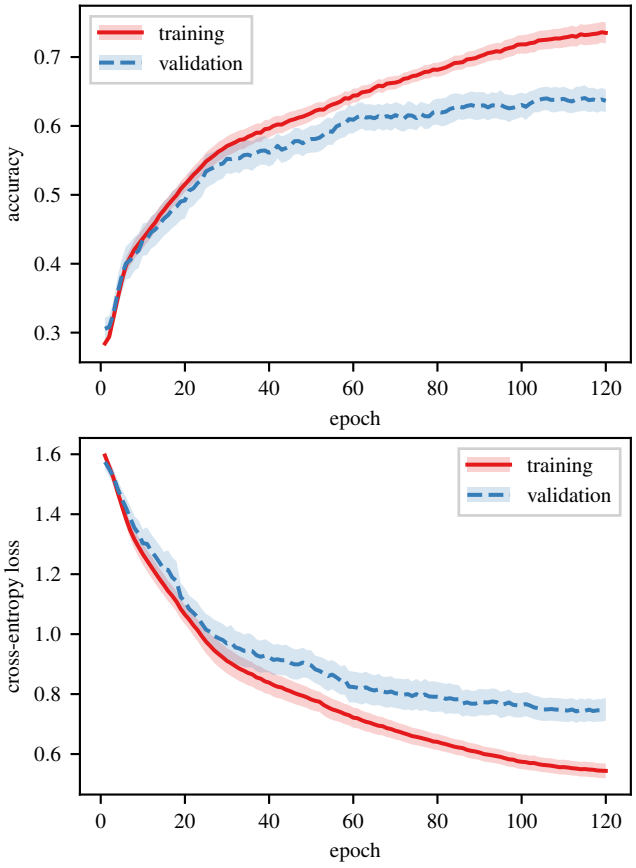


Figure 5: The evolution of the training and validation 5 class accuracy and loss during training for phase-contrast images. Mean (solid line) and standard error (shaded region) is shown.

4. Experiments

All experiments were conducted using 10-fold stratified cross-validation. In each fold, 10% of the training data is used for validation, e.g., for learning rate adaptation. Average results and standard errors are reported. The number of training epochs for the binary and multiclass classification was set to 80 and 120, respectively, which seems sufficient for convergence (see Fig. 4 and 5).

We first experimentally justify the choice of the tiling parameters (see Section 3.1) and compare our deep learning approach with previous work on fluorescence images in Sections 4.1 and 4.2. The main experiments on supervised classification of phase-contrast images are in Sections 4.3 and 4.4. Furthermore, we analyze the effect of the time delay between drug application and image acquisition in Section 4.5 and the dependency on the dataset size in Section 4.6. We visualize the features extracted by our network in Sec. 4.7. Finally, we illustrate the effect of equalizing the effects of the substances by using substance-specific concentrations in Section 4.8. See Appendix A for a summary of the used statistical measures.

4.1. Tiling effect

For the binary classification of phase-contrast images of *Topotecan* and *Etoposide*, we have tried tile sizes 256×216, 512×432, 768×648, and 1024×864. The classification accuracies were 0.76 ± 0.17, 0.88 ± 0.18, 0.90 ± 0.15, and 0.96 ± 0.07, respectively, increasing monotonously with the tile size. Moreover, greater tile size leads to faster processing. This justifies our choice of 1024×864. Larger sizes are not feasible due to GPU memory limitations.

4.2. Fluorescence images

We first apply our CNN to the binary classification task of distinguishing between a pair of classes from fluorescence images to enable a direct comparison with earlier work [24, 25]. We consider all ten possible pairs of our five classes. In Table 2, note that distinguishing *Topotecan* from *Daunorubicin* and *DMSO* from the *No treatment* class (last two rows) seems to be much more difficult than for the remaining pairs. This holds for all methods and both types of input data. We call these pairs *ambiguous* and do not consider them in the observations below. See Section 5 for more discussion.

We compare the average and maximum aggregation operators (see Sec. 3.1) with the tile-wise accuracy, i.e., no aggregation. We see that average aggregation outperforms

Table 2
Pairwise class accuracies for average and maximum tile aggregations and no-aggregation (tile-wise accuracy) for fluorescence images. The last column (SVM) shows accuracies obtained by the previous method [24, 25]. Cross-validation means and standard errors are reported, and a bold font denotes the best results for a particular class pair. Ambiguous class pairs are denoted by a star (⋆).

		Image-wise		
Classes	Tile-wise	Average	Maximum	SVM [25]
Topotecan vs. Etoposide	0.96 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.85 ± 0.04	0.85
Topotecan vs. DMSO	0.92 ± 0.02	0.97 ± 0.02	0.75 ± 0.06	**0.98**
Topotecan vs. No treatment	0.94 ± 0.02	0.94 ± 0.03	0.94 ± 0.03	**0.96**
Daunorubicin vs. Etoposide	0.97 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.82 ± 0.05	0.95
Daunorubicin vs. DMSO	0.92 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.80 ± 0.05	**1.00**
Daunorubicin vs. No treatment	0.95 ± 0.02	0.98 ± 0.02	0.89 ± 0.09	**1.00**
Etoposide vs. DMSO	0.97 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.84 ± 0.03	0.97
Etoposide vs. No treatment	0.99 ± 0.03	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.03	0.98
DMSO vs No treatment⋆	0.61 ± 0.04	0.62 ± 0.04	$\mathbf{0.66 \pm 0.05}$	0.64
Topotecan vs. Daunorubicin⋆	0.62 ± 0.04	0.42 ± 0.04	0.37 ± 0.04	**0.52**
Average (all)	0.89 ± 0.04	$\mathbf{0.89 \pm 0.06}$	0.79 ± 0.05	$\mathbf{0.89 \pm 0.05}$
Average (unambiguous)	0.95 ± 0.02	$\mathbf{0.99 \pm 0.01}$	0.86 ± 0.02	0.96 ± 0.02

Table 3
Pairwise class accuracies on phase-contrast images, with the binary classifiers obtained by training from scratch or by fine-tuning a pre-trained multiclass classifier. We show both the tile-wise and image-wise accuracies (after aggregation). The last column (SVM) shows accuracies obtained by the previous method [24, 25]. Cross-validation means and standard errors are reported, and a bold font denotes the best results for a particular class pair. Ambiguous class pairs are denoted by a star (⋆).

	From scratch		Fine-tuning		
Classes	Tile-wise	Image-wise	Tile-wise	Image-wise	SVM [25]
Topotecan vs. Etoposide	0.79 ± 0.06	0.93 ± 0.03	0.82 ± 0.04	$\mathbf{0.96 \pm 0.02}$	0.78
Topotecan vs. DMSO	0.89 ± 0.03	0.94 ± 0.03	0.93 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.79
Topotecan vs. No treatment	0.94 ± 0.01	0.99 ± 0.01	0.95 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.77
Daunorubicin vs. Etoposide	0.88 ± 0.02	0.95 ± 0.02	0.84 ± 0.03	$\mathbf{0.97 \pm 0.01}$	0.79
Daunorubicin vs. DMSO	0.88 ± 0.02	0.97 ± 0.02	0.91 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.91
Daunorubicin vs. No treatment	0.94 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.95 ± 0.01	0.99 ± 0.01	0.95
Etoposide vs. DMSO	0.87 ± 0.03	0.94 ± 0.04	0.93 ± 0.01	$\mathbf{0.97 \pm 0.02}$	0.72
Etoposide vs. No treatment	0.98 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.95 ± 0.01	0.99 ± 0.01	0.77
DMSO vs. No treatment⋆	0.61 ± 0.04	0.63 ± 0.05	0.46 ± 0.04	$\mathbf{0.65 \pm 0.03}$	0.54
Topotecan vs. Daunorubicin⋆	0.60 ± 0.06	0.53 ± 0.08	0.46 ± 0.04	$\mathbf{0.58 \pm 0.06}$	0.41
Average (all)	0.84 ± 0.04	0.89 ± 0.05	0.82 ± 0.06	$\mathbf{0.91 \pm 0.05}$	0.74 ± 0.05
Average (unambiguous)	0.90 ± 0.02	0.97 ± 0.02	0.91 ± 0.02	$\mathbf{0.99 \pm 0.01}$	0.81 ± 0.03

maximum aggregation and no aggregation. We shall therefore use average aggregation for all the remaining experiments.

The rightmost column of Table 2 shows the result of the previous method from [24, 25] using hand-crafted shape features and an SVM classifier. We see that the new CNN method either outperforms the earlier SVM method or matches its performance.

4.3. Phase-contrast images

We repeated the binary classification experiment from Section 4.2 with phase-contrast images (see Table 3). We see that fine-tuning a pre-trained multiclass classifier (see Section 3.2) is almost always better (and never significantly worse) than training each classifier independently from scratch. This is probably because the images in all classes are similar, so high-quality features learned on a large dataset work well for all pairs of classes. Another advantage of sharing the features is speed, as fine-tuning is faster than repeated tuning-from-scratch by a factor of at least 6.

4.4. Multiclass classification

Our main result is the multiclass classification method for phase-contrast images. Table 4a shows the confusion matrix (for the sum of all folds). The ambiguous classes identified earlier (*Topotecan* vs. *Daunorubicin* and *DMSO* vs. *No treatment*) are often confused, leading to the average classification accuracy of only 70%. We, therefore, also report results where the ambiguous class pairs are considered together (Table 4b). Then the classification is almost

Table 4
Confusion matrices for the multiclass classification of phase-contrast images. We provide (a) a confusion matrix for the five-class classification, where the average prediction accuracy is 70%, and (b) a reduced confusion matrix, where the ambiguous classes are joined together, leading to the average prediction accuracy of 98%. The matrices were obtained as a sum of matrices from all folds.

(a) Five classes

	Prediction				
True class	Topot.	DMSO	Daun.	Etop.	No treat.
Topot.	16	0	15	0	0
DMSO	1	4	0	0	28
Daun.	10	0	25	0	0
Etop.	1	0	0	31	1
No treat.	0	1	0	0	59
Recall [%]	51.6	12.1	71.4	93.9	98.3
Precision [%]	57.1	80.0	62.5	100.0	67.0

(b) Reduced

	Prediction		
True class	Daun. + Topot.	Etop.	No treat. + DMSO
Daun.+ Topot.	66	0	0
Etop.	1	31	1
No treat.+ DMSO	1	0	92
Recall [%]	100.0	93.9	98.9
Precision [%]	97.1	100.0	98.9

Table 5
Confusion matrices for the multiclass classification of fluorescence images. We provide (a) a confusion matrix for the five-class classification, where the average prediction accuracy is 65%, and (b) a reduced confusion matrix, where the ambiguous classes are joined together, leading to the average prediction accuracy of 99%. The matrices were obtained as a sum of matrices from all folds.

(a) Five classes

	Prediction				
True class	Topot.	DMSO	Daun.	Etop.	No treat.
Topot.	21	0	19	0	0
DMSO	0	4	1	0	32
Daun.	14	0	14	0	0
Etop.	0	0	0	33	0
No treat.	1	0	0	0	51
Recall [%]	52.5	10.8	50.0	100.0	98.1
Precision [%]	58.3	100.0	41.2	100.0	61.4

(b) Reduced

	Prediction		
True class	Daun. + Topot.	Etop.	No treat. + DMSO
Daun.+ Topot.	68	0	0
Etop.	0	33	0
No treat.+ DMSO	2	0	87
Recall [%]	100.0	100.0	98.0
Precision [%]	97.0	100.0	100.0

perfect, with an average classification accuracy of 98%. On fluorescence images, the results are similar (Tables 5a and 5b).

4.5. Time delay effect

To investigate the effect of the delay between drug application and image acquisition, we divided the dataset into two subsets, acquired 24h and 72h after application. We examined different combinations of these subsets for training and testing.

Table 6a shows that for the 5 class formulation, the 24h images seem about as difficult to classify as 72h images, with accuracies 61% and 63%, respectively. Interestingly, the two subsets seem quite different, as we observe a significant drop in accuracy when training on one and testing on the other subset (24h vs. 72h). On the other hand, aggregating the ambiguous classes leads to almost perfect classification results for the 72h data in the 3 class formulation (Table 6b), with a weaker performance on the 24h subset. We can conclude that given more time, the effects of the drugs are more substantial.[1] We have repeated the same experiments also for fluorescence images with qualitatively similar results (Tables 7a and 7b).

4.6. Dataset size impact

To understand the effect of the training data size on the generalization ability, we evaluated the multiclass classification accuracy using 20 $\sim$ 100% of the original data for training. We found that, for the 5-class formulation, the test image accuracy already gets saturated at 60% of the data (Figure. 6). However, this is probably due to the ambiguous classes, since for the 3-class problem, the accuracy continues to improve with more data, albeit slowly.

4.7. Feature visualization

This experiment illustrates the potential usefulness of the features extracted by our CNN to analyze unseen data. We took the 512-dimensional feature vectors from the penultimate layer of the multiclass CNN classifier trained on the image tiles from one of the cross-validation folds. We employed

[1] The slight difference between reported accuracies in Tables 6a and 4a is caused by averaging accuracies over runs instead of averaging the counts.

Table 6
Multiclass classification accuracy as a function of the time between drug application and image acquisition for phase-contrast images for (a) 5 classes and (b) 3 aggregated classes.

(a) Five class

	Tested on		
Trained on	24h	72h	all
24h	0.61 ± 0.06	0.39 ± 0.08	0.51 ± 0.05
72h	0.36 ± 0.04	0.63 ± 0.04	0.48 ± 0.03
all	0.72 ± 0.05	0.74 ± 0.04	0.73 ± 0.03

(b) Reduced

	Tested on		
Trained on	24h	72h	all
24h	0.88 ± 0.06	0.71 ± 0.06	0.81 ± 0.05
72h	0.57 ± 0.05	0.98 ± 0.02	0.76 ± 0.03
all	0.95 ± 0.03	1.00 ± 0.00	0.96 ± 0.03

Table 7
Multiclass classification accuracy as a function of the time between drug application and image acquisition for fluorescence images for (a) 5 classes and (b) 3 aggregated classes.

(a) Five class

	Tested on		
Trained on	24h	72h	all
24h	0.55 ± 0.05	0.50 ± 0.04	0.53 ± 0.04
72h	0.42 ± 0.06	0.64 ± 0.04	0.55 ± 0.05
all	0.64 ± 0.07	0.53 ± 0.05	0.61 ± 0.04

(b) Reduced

	Tested on		
Trained on	24h	72h	all
24h	0.92 ± 0.04	0.93 ± 0.03	0.92 ± 0.03
72h	0.58 ± 0.05	0.98 ± 0.01	0.78 ± 0.04
all	0.98 ± 0.02	0.99 ± 0.01	0.99 ± 0.01

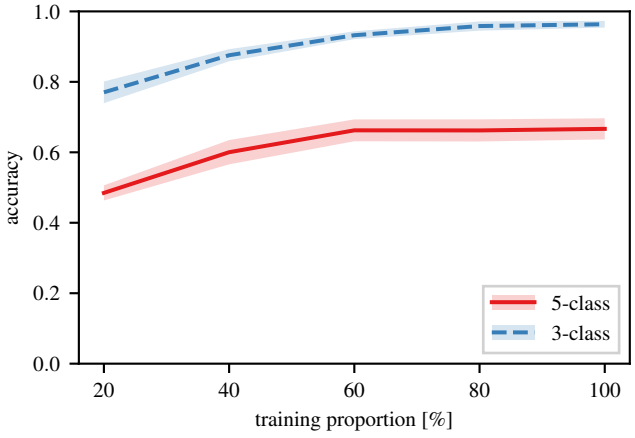


Figure 6: Dependency of the multiclass image classification accuracy on the training dataset size. We show a curve for the 5-class problem and also the 3-class formulation with the two ambiguous class pairs merged into two classes. Mean (solid line) and standard error (shaded region) are shown.

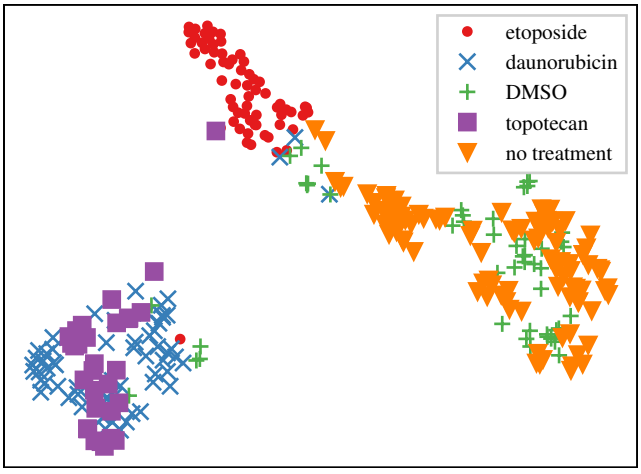


Figure 7: The extracted features from the penultimate layer of the classification CNN visualized using the t-SNE mapping algorithm.

the t-SNE [23] dimension reduction method to visualize the features corresponding to unseen test image tiles as two-dimensional vectors. Three distinct clusters emerge (Fig. 7), with the ambiguous classes belonging to the same clusters. This indicates that the features characterize well the drugs' mechanism of action.

4.8. IC_{50} concentrations

The response to different active substances varies both in the appearance of the affected cells as well as in their number. We want to focus on appearance since the concentration will be unknown in real use cases. To eliminate the differences in the effect size, we have repeated the data acquisition not with equal concentrations but with substance-specific IC_{50}[2] concentrations of the active substances, chosen so that 50% of the cells are affected. In particular, we used: Topotecan (1.24μM), Daunorubicin (0.18μM), and Etoposide (5.88μM). We then repeated the multiclass classification experiment on phase-contrast images (Section 4.4). Very interestingly, the classification results have much improved from 70% to 87.5% in the 5 class case — compare Tables 8a and 4a. The formerly ambiguous Topotecan and Daunorubicin can be distinguished in this case. Aggregating DMSO

[2] half maximal inhibitory concentrations

Table 8
Confusion matrices for the multiclass classification of phase-contrast images with biologically-active drug concentrations: (a) a confusion matrix for the five-class classification with the average prediction accuracy of 87.5%, and (b) a reduced confusion matrix, where *DMSO* and *No treatment* are joined together, leading to the average prediction accuracy of 97.7%.

(a) Five class

	Prediction				
True class	Topot.	DMSO	Daun.	Etop.	No treat.
Topot.	44	0	1	0	0
DMSO	0	22	0	0	12
Daun.	0	0	39	1	0
Etop.	1	1	0	37	0
No treat.	0	6	0	0	12
Recall [%]	97.8	64.7	97.5	94.9	66.7
Precision [%]	97.8	75.9	97.5	97.4	50.0

(b) Reduced

	Prediction			
True class	Topot.	DMSO + No treat.	Daun.	Etop.
Topot.	44	0	1	0
DMSO + No treat.	0	52	0	0
Daun.	0	0	39	1
Etop.	1	1	0	37
Recall [%]	97.8	100.0	97.5	94.9
Precision [%]	97.8	98.1	97.5	97.4

and *no treatment* classes (which are truly ambiguous) leads to an almost perfect classification accuracy of 98% for this 4 class problem (Table 8b), the same as we had previously for the 3 class problem (Table 4b). See the next section for more discussion.

5. Discussion and conclusions

We created a method capable of distinguishing the effect of several cytotoxic compounds on a cell line population from phase-contrast microscopy images instead of the more commonly used fluorescence images. This paves the way to a much simpler and faster high-throughput screening for new potential drugs. Moreover, we could visually observe a meaningful separation between classes in the feature space. This is very promising for the future task of clustering yet unseen drugs according to their mechanism of action, which we believe will be one of the primary use cases of this methodology. Finally, we saw that the drug effects are easier to distinguish after 72 hours than after 24 hours.

Our method can improve the speed and accuracy of the cellular micro-array screening, potentially leading to improved efficiency of the drug discovery process and, thus, to better clinical outcomes in the long term. Of course, we need to be aware that high-throughput cellular screening under standardized conditions is only one of the many steps in the drug discovery process.

It is well known that the cell response can be different *in vivo* than *in vitro*, vary for different cell lines, and be influenced by multiple genes. This is not a problem for our use case where the conditions are well controlled, and the cell lines are identical. However, they would need to be addressed for this technique to be used for predicting the effect in more general situations with more confounding factors. The principal limitation of our study is the relatively small dataset size — we plan to extend it to a much larger dataset with more active substances, various mechanisms of action and possibly more cell line types and other confounding factors, which should improve the robustness of the classification.

The preliminary results in Section 4.8 suggest that substance concentration is one of the confounding factors that have an essential effect on classification accuracy. In particular, by optimizing the concentrations, it was possible to distinguish Topotecan and Daunorubicin, which could not be discerned in the fixed concentration dataset, possibly because they are both topoisomerase inhibitors [21] and their mechanism of action is similar. Although the concentration will be unknown in real use cases, it should be possible to analyze images at several dilution levels and automatically choose the most relevant ones, complicating the experiment and increasing the number of necessary acquisitions. We also cannot distinguish between DMSO and *no treatment,* which is not surprising since DMSO is not supposed to affect the cells.

In theory, our method could be used directly on histological images in the clinical setting, similar to [36], although the network would have to be retrained. The principal difficulty would be obtaining a sufficiently large and comprehensive database.

Acknowledgments

This work was supported by the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics”; the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/170/OHK3/3T/13; the Czech Ministry of Education, Youth and Sports projects CZ-OPENSCREEN (LM2018130) and EATRIS-CZ (LM2018133); the European Regional Development Fund project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868); and the project National Institute for Cancer Research (Programme EXCELES, No. LX22NPO5102) funded by the European Union – Next Generation EU.

A. Statistical evaluation measures

The classification performance is evaluated via *accuracy*, the proportion of correct classifications,

$$\text{accuracy} = \frac{\left|\text{correctly classified instances}\right|}{\left|\text{all instances}\right|}.$$

For each class i separately, we also report *recall* (also known as sensitivity),

$$\text{recall} = \frac{\left|\text{correctly classified instances of } i\right|}{\left|\text{all instances of } i\right|},$$

and *precision* (or positive predictive value),

$$\text{precision} = \frac{\left|\text{correctly classified instances of } i\right|}{\left|\text{instances classified as } i\right|}.$$

We do not use specificity, which we consider less relevant in the multiclass setting.

References

[1] Bensch, R., Ronneberger, O., 2015. Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs, in: 2015 IEEE 12th International Symposium on Biomedical Imaging, IEEE. pp. 1220–1223.

[2] Boutros, M., et al., 2015. Microscopy-based high-content screening. Cell 163, 1314–1325.

[3] Carpenter, A.E., et al., 2006. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome biology 7, 1–11.

[4] Carracedo-Reboredo, P., et al., 2021. A review on machine learning approaches and trends in drug discovery. Computational and Structural Biotechnology Journal 19, 4538.

[5] Ciompi, F., et al., 2017. The importance of stain normalization in colorectal tissue classification with convolutional networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging, IEEE. pp. 160–163.

[6] Cortes, C., Vapnik, V., 1995. Support vector networks. Machine Learning 20, 273–297.

[7] De Chaumont, F., et al., 2012. Icy: an open bioimage informatics platform for extended reproducible research. Nature methods 9, 690–696.

[8] Dürr, O., Sick, B., 2016. Single-cell phenotype classification using deep convolutional neural networks. Journal of biomolecular screening 21, 998–1003.

[9] Godinez, W.J., et al., 2017. A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics 33, 2010–2019.

[10] Goodfellow, I.J., et al., 2016. Deep Learning. MIT Press.

[11] Götte, M., Gabriel, D., 2011. Image-based high-content screening in drug discovery. Drug discovery and development–present and future. Croatia: InTech , 339–361.

[12] Guerriero, M.L., et al., 2020. Delivering robust candidates to the drug pipeline through computational analysis of arrayed CRISPR screens. SLAS Discovery 25, 646–654.

[13] Gupta, A., et al., 2019. Deep learning in image cytometry: a review. Cytometry Part A 95, 366–380.

[14] He, K.o., 2016. Deep residual learning for image recognition, in: Computer vision and pattern recognition, pp. 770–778.

[15] Hickey, S., et al., 2022. Fluorescence microscopy—an outline of hardware, biological handling, and fluorophore considerations. Cells 11, 35.

[16] Huh, S., et al., 2010. Automated mitosis detection of stem cell populations in phase-contrast microscopy images. IEEE transactions on medical imaging 30, 586–596.

[17] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, PMLR. pp. 448–456.

[18] Isola, P., et al., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

[19] Kolarik, M., et al., 2020. Comparing normalization methods for limited batch size segmentation neural networks, in: 2020 43rd International Conference on Telecommunications and Signal Processing, IEEE. pp. 677–680.

[20] Land, P., et al., 2006. Cellular imaging in drug discovery. Nature Reviews Drug Discovery 5, 343–356.

[21] Liang, X., et al., 2019. A comprehensive review of topoisomerase inhibitors as anticancer agents in the past decade. European journal of medicinal chemistry 171, 129–168.

[22] Ljosa, V., et al., 2012. Annotated high-throughput microscopy image sets for validation. Nature methods 9, 637–637.

[23] van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research 9.

[24] Mertanová, H., et al., 2022. Learning to segment cell nuclei in phase-contrast microscopy from fluorescence images for drug discovery, in: Medical Imaging 2022: Image Processing, SPIE. pp. 688–694.

[25] Mertanová, H., 2021. Cell segmentation in microscopy using a reference modality. Diploma Thesis 2021, Czech Technical University in Prague .

[26] Niioka, H., et al., 2018. Classification of C2C12 cells at differentiation by convolutional neural network of deep learning using phase contrast images. Human cell 31, 87–93.

[27] Pan, J., et al., 2009. Learning to detect different types of cells under phase contrast microscopy. Microscopic Image Analysis with Applications in Biology 2009.

[28] Pau, G., et al., 2010. EBImage—an R package for image processing with applications to cellular phenotypes. Bioinformatics 26, 979–981.

[29] Pawlowski, N., 2016. Towards Image-Based Morphological Profiling using Deep Learning Techniques. Ph.D. thesis. University of Edinburgh.

[30] Prasad, A., Alizadeh, E., 2019. Cell form and function: interpreting and controlling the shape of adherent cells. Trends in biotechnology 37, 347–357.

[31] Pratapa, A., et al., 2021. Image-based cell phenotyping with deep learning. Current opinion in chemical biology 65, 9–17.

[32] Ronneberger, O., et al., 2015. U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

[33] Scheeder, C., et al., 2018. Machine learning and image-based profiling in drug discovery. Current Opinion in Systems Biology 10, 43–52. Pharmacology and drug discovery.

[34] Schindelin, J., et al., 2012. Fiji: an open-source platform for biological-image analysis. Nature methods 9, 676–682.

[35] Shariff, A., et al., 2010. Automated image analysis for high-content screening and analysis. Journal of biomolecular screening 15, 726–734.

[36] Shih-Chiang, H., Chi-Chung, C., et al., 2022. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. Nature communications 3347, 1–14.

[37] Simm, J., et al., 2018. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. Cell Chemical Biology 25, 611–618.

[38] Smith, A., 2002. Screening for drug discovery: the leading question. Nature 418, 453–455.

[39] Smith, L.N., 2017. Cyclical learning rates for training neural networks, in: 2017 IEEE winter conference on applications of computer vision, IEEE. pp. 464–472.

[40] Theriault, D.H., et al., 2012. Cell morphology classification and clutter mitigation in phase-contrast microscopy images using machine learning. Machine Vision and Applications 23, 659–673.

[41] Ulyanov, D., et al., 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, in: Computer Vision and Pattern Recognition, pp. 6924–6932.

[42] Yanagisawa, K., et al., 2020. Convolutional neural network can recognize drug resistance of single cancer cells. International journal of molecular sciences 21, 3166.