

DOES VIRTUAL STAINING FOR HIGH-THROUGHPUT SCREENING GENERALIZE?

A PREPRINT

 **Samuel Tonks**

School of Computer Science,
University of Birmingham,
Birmingham, UK.
sxt118@student.bham.ac.uk

 **Cuong Nguyen**

Artificial Intelligence & Machine Learning,
GSK,
South San Francisco,
California 94080, United States.
cuong.q.nguyen@gsk.com

 **Steve Hood**

GSK Drug Metabolism & Pharmacokinetics,
GSK Medicines Research Centre,
Gunnels Wood Road,
Stevenage, Hertfordshire, SG1 2NY, UK.
steve.r.hood@gsk.com

Ryan Musso

GSK Bioimaging,
1250 S Collegeville Rd,
Collegeville, PA 19426, United States.
ryan.x.musso@gsk.com

Ceridwen Hopely

GSK Bioimaging,
1250 S Collegeville Rd,
Collegeville, PA 19426, United States.
ceridwen.s.hopely@gsk.com

Steve Titus

GSK Bioimaging,
1250 S Collegeville Rd,
Collegeville, PA 19426, United States.
steve.titus@thermofisher.com

 **Minh Doan ***

GSK Bioimaging,
1250 S Collegeville Rd,
Collegeville, PA 19426, United States.
minh.x.doan@gsk.com

 **Iain Styles ***

School of Electronics, Electrical Engineering
and Computer Science,
Queen's University,
Belfast, UK.
i.styles@qub.ac.uk

 **Alexander Krull ***

School of Computer Science,
University of Birmingham,
Birmingham, UK.
a.f.f.krull@bham.ac.uk

July 10, 2024

ABSTRACT

The large volume and variety of imaging data from high-throughput screening (HTS) in the pharmaceutical industry present an excellent resource for training virtual staining models. However, the potential of models trained under one set of experimental conditions to generalize to other conditions remains underexplored. This study systematically investigates whether data from three cell types (lung, ovarian, and breast) and two phenotypes (toxic and non-toxic conditions) commonly found in HTS can effectively train virtual staining models to generalize across three typical HTS distribution shifts: unseen phenotypes, unseen cell types, and the combination of both. Utilizing a dataset of

772,416 paired bright-field, cytoplasm, nuclei, and DNA-damage stain images, we evaluate the generalization capabilities of models across pixel-based, instance-wise, and biological-feature-based levels. Our findings indicate that training virtual nuclei and cytoplasm models on non-toxic condition samples not only generalizes to toxic condition samples but leads to improved performance across all evaluation levels compared to training on toxic condition samples. Generalization to unseen cell types shows variability depending on the cell type; models trained on ovarian or lung cell samples often perform well under other conditions, while those trained on breast cell samples consistently show poor generalization. Generalization to unseen cell types and phenotypes shows good generalization across all levels of evaluation compared to addressing unseen cell types alone. This study represents the first large-scale, data-centric analysis of the generalization capability of virtual staining models trained on diverse HTS datasets, providing valuable strategies for experimental training data generation.

1 Introduction

High-throughput screening (HTS) plays a crucial role in drug discovery by enabling the simultaneous testing of a large number of compounds to assess their effects on cell cultures Szymański et al. [2011]. Fluorescence microscopy is the standard tool in HTS for detecting drug effects on cellular structures Selinummi et al. [2009]. By covalently binding different fluorescent dyes to biomolecules (fluorescent staining), it enables biological structures to be simultaneously revealed by the different emission spectra of the dyes, with each dye captured in a separate image channel Tonks et al. [2023].

Although it is an essential tool in modern biology, conventional fluorescence microscopy has important practical limitations. The staining protocol typically requires the cells to be fixed and permeabilized - a process in which cells are preserved in their biological state, effectively frozen in time - thus limiting the application of this technique to single time-point studies. Furthermore, as the expensive fixation and staining require specialist equipment and the number of fluorescence stains are inherently limited by spectrum saturation, significant interest has been put into label-free microscopy, wherein images of cells are captured without the need for staining Ounkomol et al. [2018]. Label-free microscopy Harrison et al. [2023], Gupta et al. [2022], while cost-effective and scalable, unfortunately, lacks the biological information typically found in the fluorescence stains Pirone et al. [2022].

Recent works have explored the concept of virtual staining to simultaneously leverage the scalability of label-free microscopy and biological information extracted from fluorescence microscopy Cross-Zamirski et al. [2022, 2023], Imboden et al. [2023], Wieslander et al. [2021], Tonks et al. [2023]. Virtual staining is typically framed as a multimodal image-to-image translation (I2I) problem Isola et al. [2017]. In this context, virtual staining models learn to translate unstained microscopy images into the desired labeled images.

While recent works Cross-Zamirski et al. [2022, 2023], Imboden et al. [2023], Wieslander et al. [2021], Tonks et al. [2023] have shown the significant potential of virtual staining, the ability of virtual staining models to generalize to images containing variations not present in the training data remains underexplored. In practice, HTS imaging data is highly diverse, being generated across different imaging systems, experiments, cell types, and phenotypes. This is known as the generalization gap Wagner et al. [2022] and has been identified as a key reason for the lack of reusable virtual staining models limiting its potential impact within large-scale applications. These challenges are analogous to those in DNA sequence modeling Avsec et al. [2021], where models are typically trained on specific cell types and fail to generalize to new cell types. In offline reinforcement learning Mediratta et al. [2023], Cobbe et al. [2019], Levine et al. [2020] where datasets predominantly focus on solving the task in the same environment, limiting the evaluation of generalization to unseen environments. Within the context of virtual staining we intend to bridge this generalization gap by performing a systematic data-centric approach to determine whether virtual staining models trained on specific subsets of data can generalize under common distribution shifts. Our approach has two main benefits. First, it would provide guidance on the best data generation practices to produce highly generalizable models. Second, since scientists need to extract biological insights from virtual stains, there must be a framework to quantify the domain of applicability of virtual staining.

In this work, we explore for the first time the generalizability of virtual staining models under three common HTS data distribution shifts:

- **Task 1:** Generalizing to new phenotypes
- **Task 2:** Generalizing to new cell types
- **Task 3:** Generalizing to new phenotypes & cell types

To investigate these three tasks, we leverage a GSK proprietary dataset of 772,416 images, consisting of bright-field and 3 co-registered widely used fluorescence stains; fluorescein (FITC) for cytoplasm, 6-diamidino-2-phenylindole (DAPI)

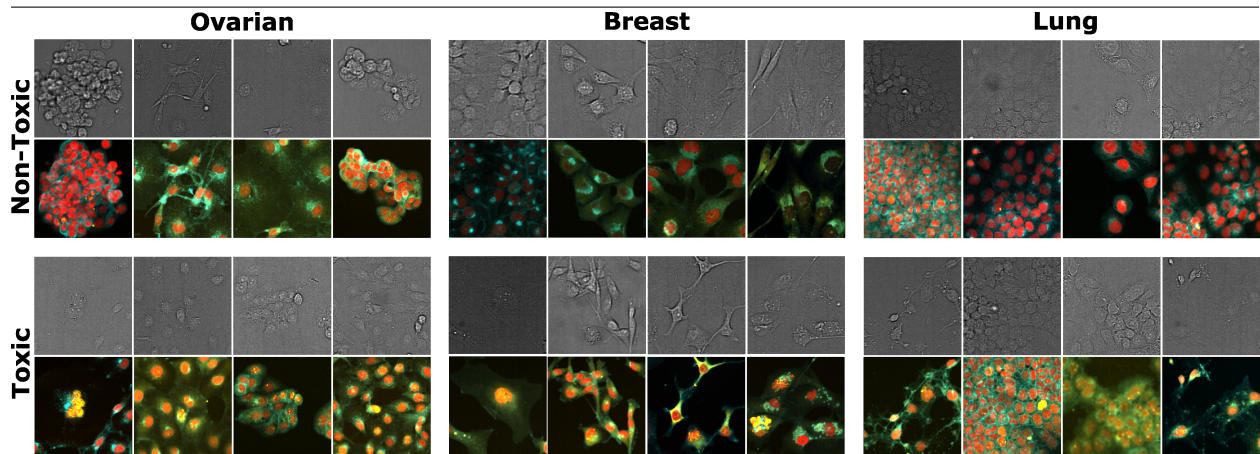


Figure 1: GSK HTS dataset comprised of three different cell types; ovarian, breast, and lung and two phenotypes; non-toxic and toxic. Each 2x4 is comprised of 4 randomly selected bright-field and fluorescence stain image pairs (shown as a composite image) for each cell type and phenotype. The composite image shows the nuclei stain (DAPI) in red, the cytoplasm stain (FITC) in cyan, and the DNA-damage stain (Cy5) in yellow. Within the dataset, we observe variability within the toxic and non-toxic samples of each cell type as well as anatomical differences between the different cell types. We explore the generalization performance across all three virtual staining tasks for three common HTS data distribution shifts; generalizing to new phenotypes, generalizing to new cell types, and both combined.

for nuclei detection and Cyanine (Cy5) for DNA-damage with 3 cell types (ovarian, lung, breast) and 2 phenotypes that correspond to the conditions under which the cells have been cultured (toxic, non-toxic). We define the wells that contain DMSO (negative control) and the three lowest levels of concentration for 10 GSK prospective compounds to be non-toxic and the three highest concentrations of each compound as well as etoposide and starausporine (positive controls) to be toxic. Example bright-field and fluorescence stain images of non-toxic and toxic sample conditions for each cell type are shown in Figure 1.

A total of 54 models were trained; one model for each of the 2 phenotypes \times 3 cell types \times 3 stains \times 3 initializations. A total of 243 individual inference runs were completed (54 for task 1, 54 for task 2 and 135 for task 3) each corresponding to the different train and test combinations. We evaluate outputs using three levels of evaluation; pixel-based, instance-based, and biological-feature-based. As the performance in absolute terms of our method has been reported Tonks et al. [2023] and because the focus of this work is generalization we report the difference in performance relative to the baseline model trained on the same biological variation found in the training set.

For task 1, we find training only on non-toxic samples generalizes well but can also surprisingly lead to improved performance over training on toxic samples, for virtual nuclei and virtual cytoplasm, even when evaluated on toxic samples. This suggests a new potential strategy for model training that could have broad implications for improving the robustness of virtual staining models when access to phenotype specific data is limited. For task 2, we find generalizing to unseen cell types to be a mixed result with different generalization outcomes depending on the level of evaluation. These findings underscore the critical challenge of developing models that perform consistently well across diverse conditions.

For task 3, we observe for virtual nuclei and cytoplasm improvements in pixel-based peak-signal-to-noise ratio (PSNR) Faragallah et al. [2020] for the majority of experiments. Specifically, when trained on ovarian non-toxic, testing on lung toxic and when trained on lung non-toxic test on ovarian toxic we observe good generalization performance across all levels of evaluation.

The performance on task 3, which can be viewed as a combination of tasks 1 and 2, combines the qualitative characteristics of the two simpler tasks. We observe a similar positive effect from training on non-toxic samples as was seen in task 1, and a similarly mixed picture as in task 2 when generalizing to an unseen cell type with particular improvements in PSNR and normalized mean-absolute-error (N-MAE) but a mixture of negative and positive effects on performance across all levels of evaluation. Consistently across all three tasks, images of breast cells were the hardest to generalize from and generalize to.

Despite good generalization performance across all tasks and several evaluation metrics, virtual DNA-damage predictions are less accurate compared to virtual nuclei and virtual cytoplasm when compared to the fluorescence stain aligning with previous findings Tonks et al. [2023]. Overall, we believe this work provides robust insight into potential strategies for generating HTS training data for generalizable virtual staining models.

2 Related Work

Virtual staining is a specific formulation of multimodal image-to-image translation (I2I) Isola et al. [2017], a machine learning technique in which we want to train a model to translate one modality - label-free brightfield images - to another - fluorescence images. Virtual staining using I2I has been widely explored using approaches based on a regression-loss Ounkomol et al. [2018], auto-regressive models Christiansen [2018], generative adversarial network (GANs) Tonks et al. [2023], Upadhyay et al. [2021], Cross-Zamirski et al. [2022] and diffusion-based approaches Cross-Zamirski et al. [2023]. Across these methods when tested on samples similar to those in the training sets, virtual staining predictions have been shown to consistently and reliably produce high quality images at the pixel-level Ounkomol et al. [2018], Christiansen [2018], Upadhyay et al. [2021], Cross-Zamirski et al. [2022] that preserve the majority of biological information found in fluorescence images Tonks et al. [2023], Cross-Zamirski et al. [2023]. In order to determine the usability of virtual staining at scale, these systems need to be able to generalize under various data distribution shifts

Despite improvements in virtual staining very few models Van der Laak et al. [2021], Echle et al. [2021] are known to have been integrated into routine clinical workflows. A recent review Wagner et al. [2022] of 161 peer-reviewed computational pathology articles identified a core reason being the generalization gap; models failing to maintain performance for unseen data with a shifted distribution. Nevertheless, in real-world applications, encountering unseen data is very common. For example in high-throughput screenings (HTS) Szymański et al. [2011] a single imaging machine, among many in a lab, generates imaging data for large numbers of compounds at different concentration levels tested on cell cultures of different cell types (lung or breast) composed of different cell lines (H1299, HCC827).

Early virtual staining works Ounkomol et al. [2018] showed models trained on images of hiPSC cells did not perform as well when generalizing to HEK-293, cardiomyocytes and HT-1080 cells. Although gross image features were visually comparable, morphological detail improved when the model was trained on data of the same cell type. Meanwhile, other virtual staining works Christiansen [2018] showed when trained separately on images containing cortical and motor neuron cells generalizing to a single well of a breast cell line (MDE-MD-231), sourced from a new laboratory and new transmitted-light technology led to an increase in pixel-level performance. Similarly, Cross-Zamirski et al. Cross-Zamirski et al. [2023] trained virtual staining models utilizing a diverse set of 290 cell type and phenotypes from the JUMP Cell Painting Dataset Chandrasekaran [2023] also testing on a single hold out plate, providing limited exploration of performance under domain shift. The question of whether virtual staining models can generalise at scale remains underexplored.

A trivial solution to generalization would be to train virtual staining models using samples from each domain shift, but in practice training this amount of models is computationally very expensive and not scalable. Alternatively, domain adaptation methods for I2I such as DAI2I Murez et al. [2018] and OST Luo et al. [2020] have been shown to improve out-of-distribution performance on natural image datasets. None of the aforementioned approaches to bridging the generalization gap have focused on whether certain image domains, when used as training sets, are better able to learn domain-invariant features compared to others. In this paper we perform the first large-scale systematic analysis of whether virtual staining models trained on specific subsets of HTS data can generalize under three common distribution shifts.

3 Experiments & Results

We first discuss the general points about our dataset, training, inference and evaluation procedures and then the three generalization tasks.

Dataset: Our experiments are based on a pool of 772,416 individual images generated as part of a GSK HTS study. The data comprises 98 16x24 well plates, with each plate containing 384 wells containing a combination of dimethyl sulfoxide (DMSO) (negative control) as it has a relatively low order of systemic toxicity Wilson et al. [1965], 10 GSK candidate compounds with 6 levels of toxicity from low to high and known apoptosis (programmed cell death) inducing compounds etoposide Kobayashi and Ratain [1994], and starausporine Qiao et al. [1996] with high orders of toxicity (positive controls). All plates of one cell type have a fixed layout of compounds and controls across the wells.

We define the wells that contain DMSO and the three lowest levels of toxicity for each compound to be non-toxic and the three highest concentrations of each compound as well as the etoposide and starausporine to be toxic. Every well consists of 9 fields of view each containing a bright-field and three co-registered fluorescent stains; fluorescein (FITC) for cytoplasm, 6-diamidino-2-phenylindole (DAPI) for nuclei detection and Cyanine (Cy5) for DNA-damage. Each cell type was represented by six different cell lines. For each cell type and stain, 27,000 bright-field and fluorescence image pairs were sampled from toxic and non-toxic labeled wells separately, with 21,000 image pairs used for training and 6,000 image pairs used for validation. Random samples for each cell type and labeled wells are shown in Figure 1. In addition, for each cell type three plates, excluded from any training or validation sets, were used to sample toxic and

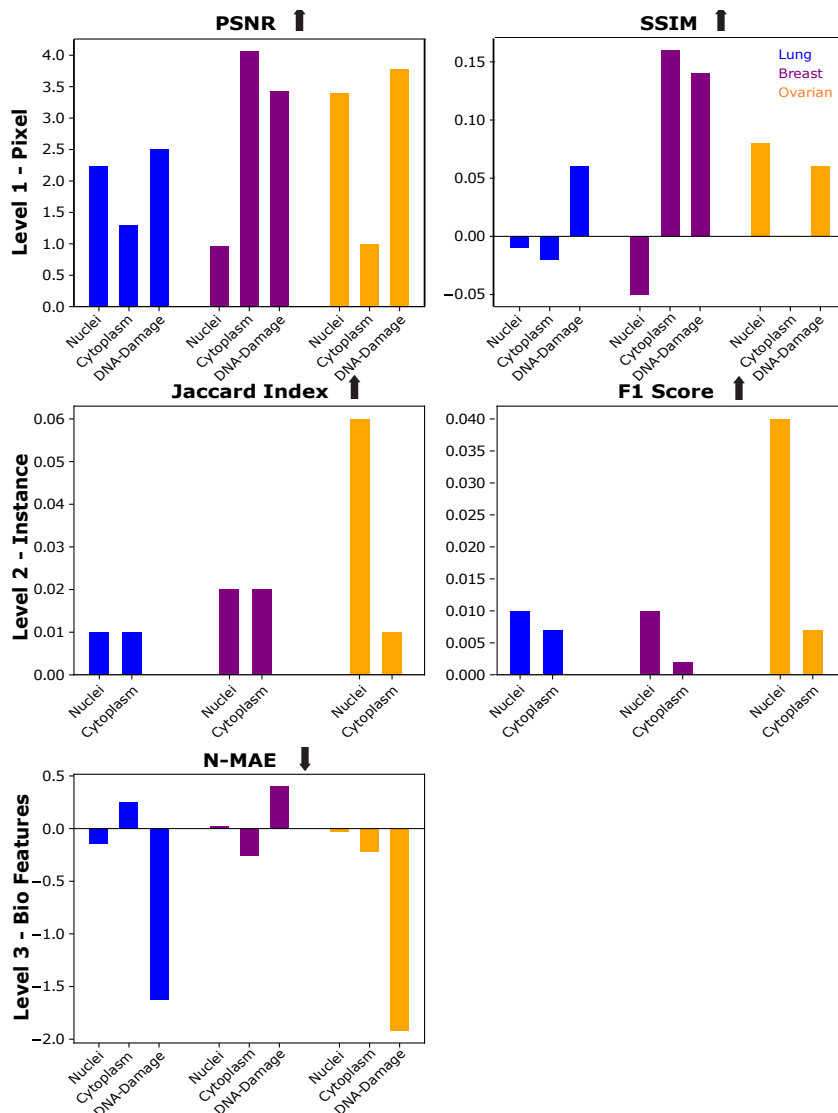


Figure 2: **Generalization performance of virtual staining models to an unseen phenotype across three levels of evaluation.** Each chart represents the results for that metric, within the chart all virtual stain channels are shown separately and grouped by cell type. Each bar shows the average difference between the virtual stain models trained on non-toxic and the baseline virtual stain models trained on toxic samples. For all three cell types and virtual stain tasks, the PSNR, Jaccard Index, and F1 Score results reveal improved performance from training on non-toxic samples compared to training on toxic samples. Consistently across all metrics, training on ovarian non-toxic leads to improved performance when generalizing to images of ovarian toxic cells.

non-toxic test sets. The test sets for both lung and ovarian toxic were 5,562 and for non-toxic were 4,806 while breast toxic was 3,510 and non-toxic was 6,856.

Training & Inference: Using the non-toxic and toxic samples for each cell type, three models were independently trained with different random initialization of weights to translate from bright-field to each fluorescence stain. In the following, all performance metrics are computed as the mean performance over all three initializations. This resulted in a total of 54 models being trained; one model for each of the 2 phenotypes \times 3 cell types \times 3 channels \times 3 initializations. This work used the Pix2PixHD Wang et al. [2018] architecture with the same hyperparameters as Tonks et al. [2023]. Each model was trained for a maximum of 200 epochs using early stopping. A total of 243 individual inference runs were completed each corresponding to the different train and test combinations. For task 1, we have 27 trained models (\times 3 cell types \times 3 stains \times 3 initializations) for each model we run inference on the non-toxic and

toxic test sets producing 54 in total. For task 2, we have 27 trained models and run inference on the two different cell type test sets producing 54 runs in total. Finally for task 3, we have 9 trained models for each cell type and phenotype (3 stains \times 3 initializations) for the toxic phenotype models we run inference on all three toxic cell type test sets and for the non-toxic only the two toxic test sets with different cell types. This leads to 45 inference runs per cell type totalling 135 plus the 108 from tasks 1 and 2 giving a total 243 runs. Each was evaluated at three levels; pixel-based, instance-based and biological-feature-based.

Evaluation: In level 1, we evaluate performance using two established pixel-level metrics (SSIM Wang et al. [2004], PSNR Faragallah et al. [2020]).

In level 2 we evaluate the instance segmentation quality that can be obtained from virtually stained nuclei (DAPI) and cytoplasm (FITC) channels. We leverage Cellpose Stringer et al. [2021] to generate instance masks for the fluorescence and virtual staining channels. To obtain high-quality segmentations we apply a gamma correction of 0.3 before feeding the images to Cellpose. The value of 0.3 was validated by manually checking a subset of random nuclei and cytoplasm fluorescence images and the Cellpose segmentation masks. We then compare the masks generated from the fluorescence and virtual staining using common segmentation metrics; Jaccard Index Jaccard [1912] for instance-wise pixel area quality and F1 score Sasaki [2007] for object detection. To compute the F1 score, we use a Jaccard Index Jaccard [1912] threshold of 0.7 in alignment with previous nuclei detection works Caicedo [2019].

In level 3, we utilise a Cellprofiler Carpenter [2006] pipeline that takes the intensity image and Cellpose masks as input to compute instance-wise scores for a collection of 209 morphological, intensity and textural, among other features frequently used in HTS. The types of features are similar to those used in Tonks et al. [2023] but by extracting these from Cellprofiler Carpenter [2006] we can compute values for each instance instead of a single mean value over all instances in an image Using a random forest based feature selection method similar to Saabas 2014 Saabas [2014] we extract the twenty most informative features identified on each of the fluorescence test sets and compare the results between the fluorescence channels and those obtained from virtual staining. The mean absolute error (MAE) between the fluorescence and virtual staining feature scores is computed over each instance, building on previous works Tonks et al. [2023]. To enable the comparison and interpretation of MAE across different feature ranges the resulting MAE for each feature is normalized (N-MAE) by the standard deviation of the fluorescence feature score. Finally, across all three levels of evaluation, we report the difference between the mean score obtained for each of the trained models and the baseline model trained on images sampled from the same distribution as the test set. This enables us to compare the change in performance across the different cell type and phenotype combinations relative to each tasks virtual staining baseline.

3.1 Task 1 - Generalization to new phenotype

We begin by exploring for each cell type, how virtual staining models trained on images containing samples of one phenotype; non-toxic, perform on images containing samples of a different phenotype; toxic. We report the difference in performance across the three levels of evaluation between the virtual staining models trained on non-toxic samples and the virtual staining models trained on toxic samples of the same cell type.

Training on non-toxic vs toxic improves performance: Across all three levels of evaluation, all three cell types, and all three virtual staining tasks shown in Figure 2, we find when testing on toxic samples training on non-toxic samples leads to improved results as measured by several metrics compared to training on toxic samples of the same cell type. In particular, for all virtual staining tasks we see consistently improved performance in PSNR, Jaccard Index, and F1 score across all cell types as well as the majority of virtual staining tasks showing improved performances in SSIM and N-MAE. However, for the task of generalising to lung toxic when training on lung non-toxic for virtual cytoplasm we see consistently improved performance in levels 1 and 2 but worsening results in level 3. We believe this is in part due to the heterogeneity of cell expression that occurs in non-toxic healthy cells providing an increase in the diversity of cell states to train generalizable virtual staining models on. In contrast, the toxic cells are induced into specific homogeneous cell states leading to potentially less diversity in training data producing less generalizable models. We can see some evidence of this in Figure 1 where despite non-toxic conditions we observe small amounts of naturally occurring DNA-damage signal in all non-toxic cell types images shown.

Ovarian non-toxic samples see the largest improvement: Across all measurements and virtual staining channels, when generalizing to ovarian toxic samples, training on ovarian non-toxic samples leads to improved performance compared to training on ovarian toxic samples. Upon visual inspection of Figure 3 for the virtual nuclei and virtual cytoplasm stains both the baseline and non-toxic models have produced predictions that replicate the general shape and intensity of a large number of cells seen in the fluorescence.

The prediction of the ovarian virtual nuclei trained on non-toxic is almost indistinguishable from the prediction of the virtual nuclei trained on toxic. However, upon closer inspection, as shown in yellow for certain nuclei the intensity profile and shape of the non-toxic nuclei more closely resembles that found in the fluorescence. Similar findings are

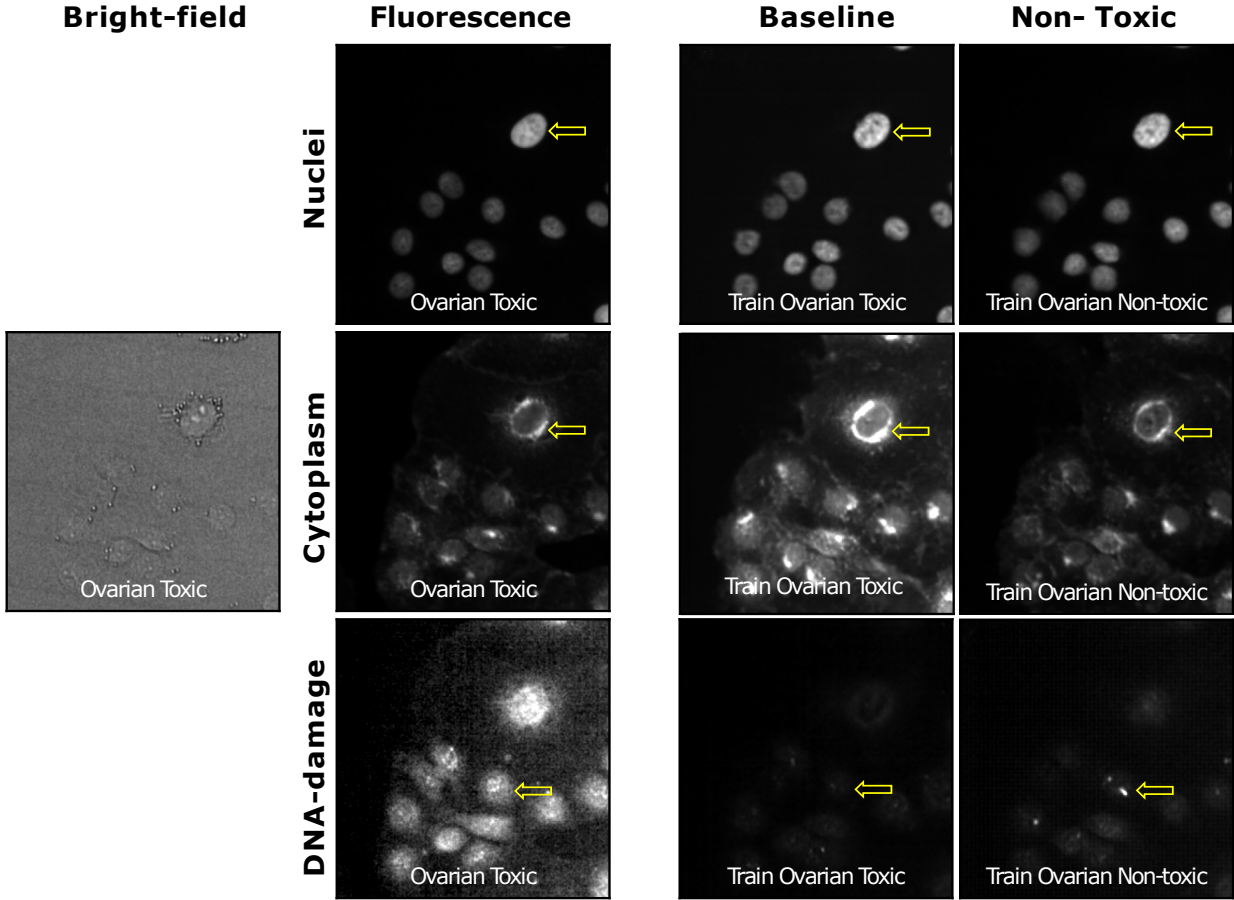


Figure 3: **Qualitative results for the task of generalizing to an unseen phenotype; ovarian toxic from ovarian non-toxic.** Randomly selected bright-field and paired fluorescence for nuclei, cytoplasm and DNA-damage stains alongside the virtual staining predictions from each of the virtual stain models trained on ovarian toxic samples and the virtual stain models trained on ovarian non-toxic samples. We observe the general shape of nuclei and cells are reproduced well relative to the baseline and fluorescence stain. The DNA-damage spots are considerably different from the fluorescence stain for both the baseline and model trained on non-toxic samples. Examples for all three virtual stain tasks are shown by yellow arrows.

highlighted for the cytoplasm intensity profile of individual cells. In contrast, the virtual DNA-damage predictions for both trained models are very similar to each other, with the model trained on non-toxic showing a small number of DNA-damage spots but they both display considerable losses of information compared to the fluorescence stain.

Figure 4 shows the N-MAE obtained for the 20 most informative biological features for the virtual staining models trained on ovarian toxic and ovarian non-toxic samples. Across all virtual staining tasks, we observe a reduction in the average N-MAE when training on images of ovarian non-toxic. Consistently, we observe this result is not driven by an outlier, but by reductions in N-MAE across a diverse set of biological features.

These findings support the previously shown results that virtual nuclei and cytoplasm models trained on non-toxic samples of specific cell types such as ovarian when generalizing to toxic samples of the same cell type can learn meaningful and diverse biological feature representations that more closely align with the fluorescence stains.

3.2 Task 2 - Generalization to new cell type

Having explored the generalization to an unseen phenotype, in this section, we focus on the second task of generating virtually stained samples of bright-field images of a different cell type to the images the virtual staining models were trained on.

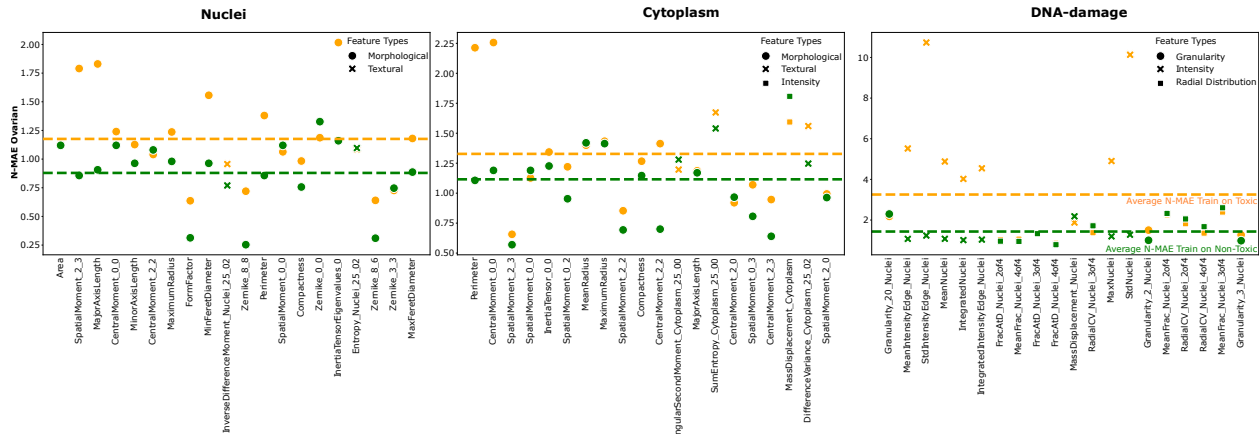


Figure 4: N-MAE values of the virtual stain models trained on ovarian non-toxic and ovarian toxic samples for the 20 CellProfiler features identified on the fluorescence stain. The green line shows the average N-MAE for the virtual stain models trained on ovarian non-toxic samples and the yellow line shows the average N-MAE for the virtual stain models trained on ovarian toxic samples. Across a diverse set of features, training on ovarian non-toxic leads to a biological feature representation that more closely aligns with that found in the fluorescence stains compared to training on ovarian toxic.

Generalizing to new cell types is complex: Upon visual inspection of Figure 5, we observe that for the virtual nuclei and virtual cytoplasm task the model trained on ovarian cells can reproduce the general shape of the lung nuclei and cytoplasm compared to the baseline model trained on images of lung cells. Although gross image features can be visually seen to be comparable, performance for morphological detail improved when the model was trained on data of the same cell type as also seen in previous works Christiansen [2018], Ounkomol et al. [2018]. The model trained on breast is unable to reproduce several nuclei and cytoplasm that can be seen in the fluorescence stain. Figure 6a shows the different train and test cell type combinations explored for this generalization task and the mean scores for the virtual nuclei and cytoplasm models for each level of evaluation.

We observe some high-level patterns of consistently good and bad generalization performance across the different train and test combinations for levels 1 and 2 of our evaluation pipeline.

For all metrics in levels 1 and 2, the virtual nuclei and virtual cytoplasm models trained on images of ovarian cells consistently outperform the corresponding virtual staining models trained on images of breast cells when testing on images of lung cells (shown in row 1) and, in some cases leads to a very small positive effect compared to the baseline model.

However, when evaluating the biological features in level 3 we observe a substantial increase in N-MAE relative to the in distribution baseline model, highlighting the importance of evaluating virtual staining in the biological feature space as opposed to using only common pixel-level and instance-level metrics. Interestingly, despite virtual models trained on images of ovarian cells performing well on images of lung cells Figure 6a shows this is not true for the majority of metrics for generalizing to images of breast cells, providing further evidence that the generalization to multiple different cell types is complex and that generalization performance to one cell type does not necessarily mean the same generalization performance to a different cell type.

In contrast to the virtual staining models trained on images of ovarian cells, for all three levels of evaluation, the virtual nuclei and cytoplasm models trained on images of breast cells produce scores that show the largest difference in performance compared to the baselines when testing on images of ovarian and lung cells.

We observe that models trained on images of ovarian or lung cells do not generalize well to images of breast cells. This suggests that images of breast cells are not as good as lung or ovarian for training virtual staining models that generalize well to unseen cell types and they are also difficult to generalize to. Finally, for the virtual nuclei and virtual cytoplasm results, we observe a lack of symmetry in performance across the different cell types for each of the measurements. The large majority of results do not show that performance when training on one cell type and testing on another is very similar if the train and test cell types are interchanged.

For the virtual DNA-damage models (not shown in Figure 6a), similar to virtual nuclei and cytoplasm, training on images of breast cells and testing on images of ovarian and lung cells consistently leads to bad generalization across all metrics. Meanwhile, training on images of lung cells when testing on images of ovarian and breast cells leads to small

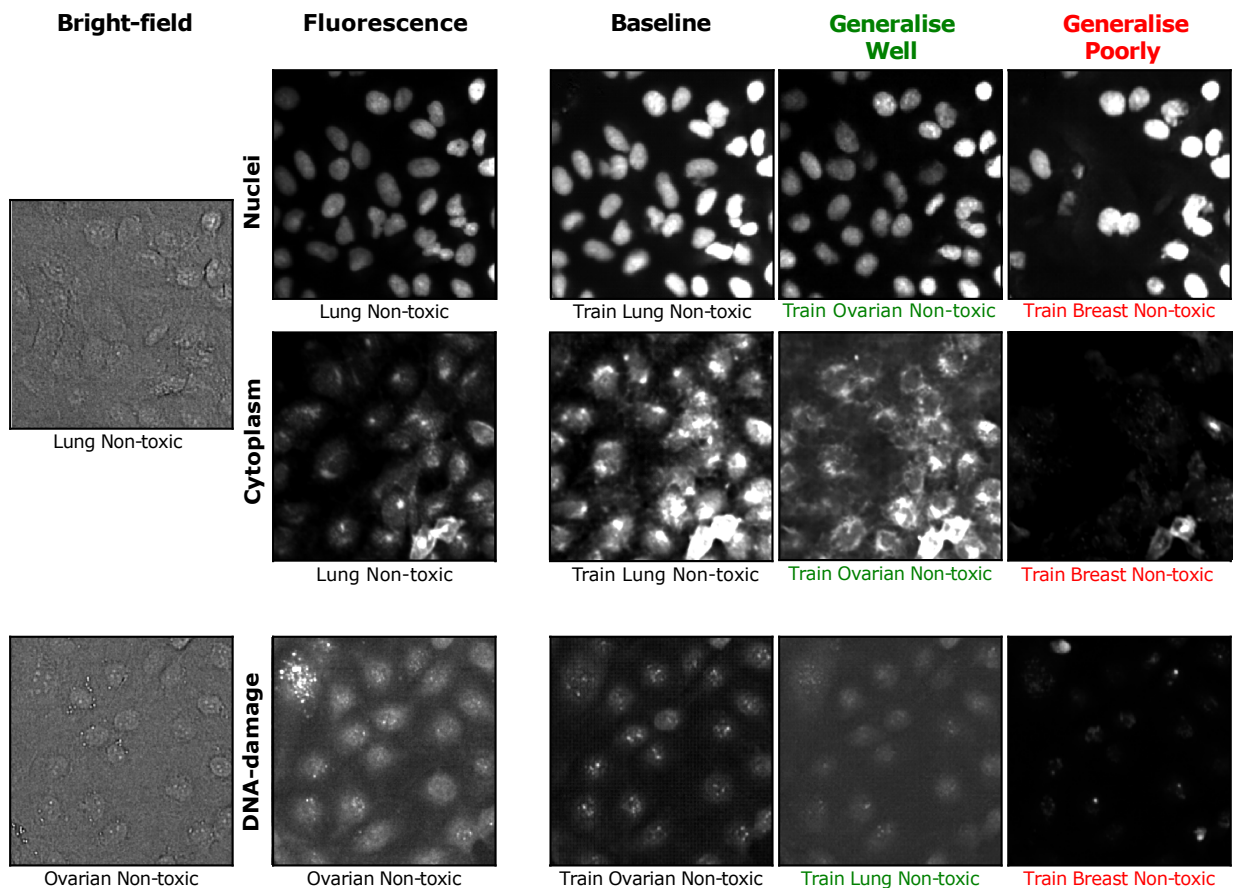


Figure 5: **Qualitative results for the task of generalizing to unseen cell types.** Randomly selected bright-field and fluorescence for nuclei, cytoplasm and DNA-Damage stains with the virtual stain from the baseline model and the virtual stain models that generalize well and generalize poorly. Both the virtual nuclei and virtual cytoplasm trained on images of ovarian non-toxic cells can reproduce the general shape of the lung cells and in some cases the intensity profile well relative to the baseline model trained on images of lung cells. Meanwhile, the models trained on images of breast cells show a considerable number of nuclei and cytoplasm missing as well as incorrect morphology. For the virtual DNA-damage although the model trained on lung does well relative to the model trained on breast there are still very clear differences in intensity and DNA-damage spot locations between all three virtual stain predictions and the fluorescence stain.

positive effects across PSNR and SSIM as well as a very small increase in N-MAE.

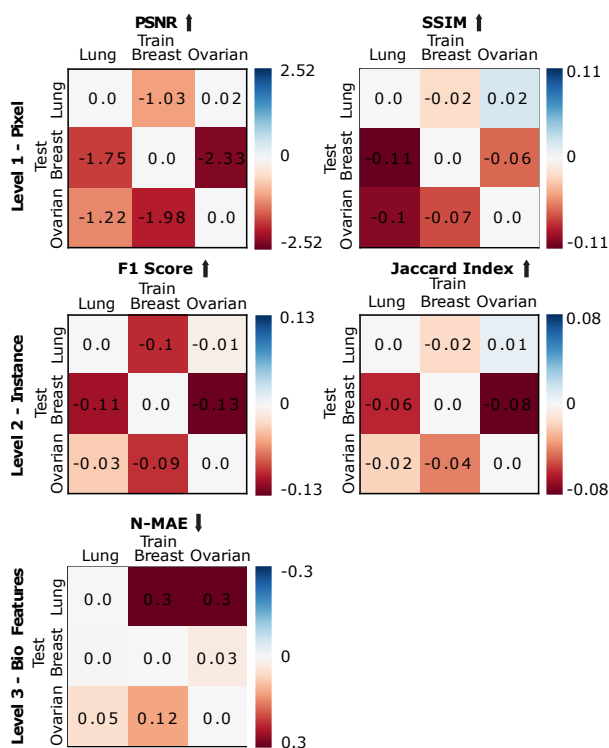
Examining the individual test set image scores for each cell type generalization task across all three virtual staining models, for all metrics we find that the differences are consistent across multiple images and not driven by a small number of outliers.

3.3 Task 3 - Generalization to new phenotypes & new cell types

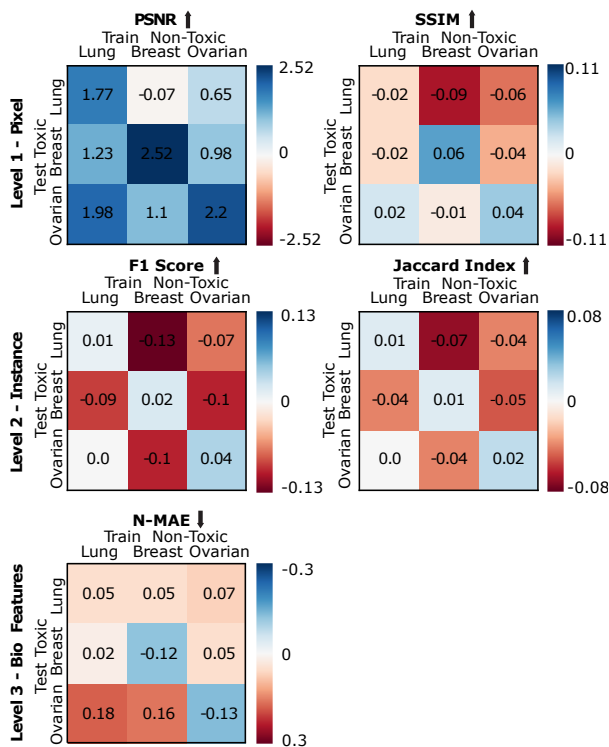
In the final section, we combine the two previous distribution shifts and explore how virtual staining models trained on images of cells in non-toxic conditions of one cell type generalize to cells of a different cell type in toxic conditions.

We report the difference in performance across the three levels of evaluation between the virtual staining models trained on non-toxic samples and the virtual staining models trained on toxic samples of the same cell type.

Generally good generalization performance In a similar approach to Figure 6a, Figure 6b shows the different non-toxic training and toxic test cell type image set combinations explored for this generalization task and the corresponding mean scores for the virtual nuclei and virtual cytoplasm models for each of the three levels of evaluation.



(a) **Task 2. Generalization to unseen cell types.** The values along the diagonal represent the baseline and are therefore 0.0. Models trained on breast cell images consistently generalize poorly to other cell types and breast cell images appear hard to generalize too. Levels 1 and 2 show ovarian can generalize well to lung for pixel and instance-level metrics but not when evaluating the biological features. Good generalization to one cell type does not imply the same for other cell types.



(b) **Task 3. Generalization to unseen cell types and phenotype.** For each metric, the layout of train/test and the conditional formatting scales are the same as Figure 6a. The values along the diagonal represent the results from generalizing to an unseen phenotype and are kept in for completion. We observe similar patterns of performance when training and testing on images of breast cells to the generalization to unseen cell types as well as considerable improvements in PSNR and N-MAE for several train and test combinations.

Figure 6: **Generalization performance of virtual nuclei and virtual cytoplasm models.** Both (a) and (b) have 5 heatmaps with each showing the average results for each metric within the 3 levels of evaluation. Each heatmap shows the cell type train and test set combinations. The metric values are the differences between the models and the baseline model. For each metric, conditional formatting is centered at 0 and scaled to the worst performance across tasks 2 and 3 (to allow for comparison) and its negative, with red indicating bad generalization performance changes and blue positive.

We observe similar differences between the performance results within levels 1 and 2 against level 3 found in task 1 when training on non-toxic lung cells and testing on toxic ovarian showing improvements in pixel-wise and instance-wise metrics relative to the baseline but increases in N-MAE when evaluating the biological feature representation.

In general, we see considerable improvements across all metrics compared to those shown in task 2 (See Figure 6a) which supports our findings from the phenotype generalization task. In particular, we observe an increase in positive PSNR results and a reduction in the highest N-MAE from 0.3 to 0.18. However, in some cases, such as when testing on ovarian toxic we observe an increase in N-MAE. For levels 1 and 2 we observe similar reduced generalization performance when training and testing on images of breast cells.

4 Discussion & Conclusion

We have investigated the generalization performance of virtual nuclei, virtual cytoplasm and virtual DNA-damage models for three common HTS generalization tasks. Firstly, generalizing to an unseen phenotype. Secondly, generalizing to unseen cell types. Finally, generalizing to an unseen phenotype and cell types combined. Performance has been evaluated using metrics at the pixel level, instance level and biological feature level.

For the first generalization task, for both virtual nuclei and virtual cytoplasm, training on non-toxic samples leads to both good generalization and improved performance relative to training on toxic samples across all levels of evaluation. In particular, when testing on toxic ovarian samples, training on non-toxic ovarian samples produced predictions that achieved higher pixel-level quality, more accurate instance-wise translation and scores for the majority of biological features that more closely match the scores found in the fluorescence channels compared to training on toxic ovarian samples.

Previous work on the generation of The JUMP Cell Painting Dataset Chandrasekaran [2023] suggests it is necessary to have training sets with a high diversity of phenotypes to effectively train machine learning models. However, these results demonstrate that for this specific HTS dataset and the explored virtual staining tasks, training on non-toxic samples alone can lead to both good generalization and also performance improvements when measured using a variety of evaluation metrics. As such, we believe that for virtual nuclei and virtual cytoplasm staining of toxic samples training on widely available non-toxic samples is a viable alternative to training on toxic samples. These findings could lead to a reduction in the number of cell- and phenotype-specific models needed to efficiently utilize virtual staining for diverse HTS datasets.

For the second generalization task, we find generalizing to unseen cell types is complex, we identify for pixel-wise and instance-based metrics, training on ovarian and testing on lung led to good generalization performance for the virtual nuclei and virtual cytoplasm. However, when evaluating at the biological feature level we observed relatively high differences in N-MAE values. These results reveal that determining the correct evaluation metric for a chosen virtual staining application is important to effectively evaluate whether certain cell type-specific training sets can better generalize compared to other cell types. Additionally, we find that good generalization to one unseen cell type does not necessarily mean the same can be expected for another cell type. In contrast, we observe that training on images of breast cells consistently leads to bad generalization performance across all levels of evaluation. We also found that models trained on other cell types were poor at generalizing to images of breast cells. Further analysis into the three non-toxic cell type data sets revealed after a manual inspection of a random subset of 5,000 images from each cell type that breast cells are distributed more sparsely than the other cell types (Figure 1). The average number of cells in the inspected breast images was less than half of that in ovarian and a third of that in lung. This corresponds to the anatomical function of breast cells, which require more fat in their surroundings Ellis and Mahadevan [2013], reducing the number of cells in any image as shown in Figure 1.

On the other hand endocrine signaling Cooper [2000] of ovarian cells requires they be closely packed together, and the vital role of lung cells to exchange gas, in principle requires more cells. We believe that these structural differences are responsible for the poor generalization performance of images of breast cells. An additional possible explanation could be that the lower density of breast cells corresponds to a smaller volume of training data for the virtual staining models.

For the final generalization task, we see the same challenges when training on and generalizing to images of breast cells for the virtual nuclei and virtual cytoplasm. However, in general, across all levels of evaluation and the explored celltype and phenotype combinations we find that training on non-toxic is preferable even when additionally generalizing to toxic samples of an unseen cell type (See Figure 6b compared to Figure 6a). We saw particular improvements in PSNR and N-MAE and relatively similar performance for SSIM, F1 Score and Jaccard Index.

Despite what appears to be good virtual DNA-damage generalization performance across certain evaluation metrics for all three generalization tasks, when examining the absolute values the results are poor compared to virtual nuclei and virtual cytoplasm affirming the findings from previous work Tonks et al. [2023]. This becomes clear qualitatively in Figure 3 and Figure 5 and quantitatively in Figure 4 where the N-MAE values are noticeably larger than in the other two channels.

Future work should explore the generalization performance to a broader selection of unseen cell types, and investigate further the issues with producing accurate virtual DNA-damage stains.

References

- Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.
- Juan C et al. Caicedo. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature Methods*, 16(12):1247–1253, 2019.
- Anne E et al. Carpenter. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7:1–11, 2006.
- Srinivas Niranj et al. Chandrasekaran. Jump cell painting dataset: Morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, Jan 2023. URL <https://www.biorxiv.org/content/10.1101/2023.03.23.534023v2>.
- Eric M et al. Christiansen. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803, 2018.
- Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pages 1282–1289. PMLR, 2019.
- Geoffrey M Cooper. The cell: A molecular approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. *Signaling Molecules and Their Receptors.*, 2000. URL <https://www.ncbi.nlm.nih.gov/books/NBK9924/>.
- Jan Oscar Cross-Zamirski, Elizabeth Mouchet, Guy Williams, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Label-free prediction of cell painting from brightfield images. *Scientific reports*, 12(1):10001, 2022.
- Jan Oscar Cross-Zamirski, Praveen Anand, Guy Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels. *arXiv preprint arXiv:2303.08863*, 2023.
- Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696, 2021.
- Harold Ellis and Vishy Mahadevan. Anatomy and physiology of the breast. *Surgery (Oxford)*, 31(1):11–14, 2013.
- Osama S Faragallah, Heba El-Hoseny, Walid El-Shafai, Wael Abd El-Rahman, Hala S El-Sayed, El-Sayed M El-Rabaie, Fathi E Abd El-Samie, and Gamal GN Geweid. A comprehensive survey analysis for present solutions of medical image fusion and future directions. *IEEE Access*, 9:11358–11371, 2020.
- Ankit Gupta, Philip J Harrison, Håkan Wieslander, Jonne Rietdijk, Jordi Carreras Puigvert, Polina Georgiev, Carolina Wählby, Ola Spjuth, and Ida-Maria Sintorn. Is brightfield all you need for mechanism of action prediction? *bioRxiv*, pages 2022–10, 2022.
- Philip John Harrison, Ankit Gupta, Jonne Rietdijk, Håkan Wieslander, Jordi Carreras-Puigvert, Polina Georgiev, Carolina Wählby, Ola Spjuth, and Ida-Maria Sintorn. Evaluating the utility of brightfield image data for mechanism of action prediction. *PLOS Computational Biology*, 19(7):e1011323, 2023.
- Sara Imboden, Xuanqing Liu, Marie C Payne, Cho-Jui Hsieh, and Neil YC Lin. Trustworthy in silico cell labeling via ensemble-based image translation. *Biophysical Reports*, 3(4), 2023.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.
- Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New Phytologist*, 11(2):37–50, 1912.
- K Kobayashi and M J Ratain. Pharmacodynamics and long-term toxicity of etoposide. *Cancer Chemotherapy and Pharmacology*, 34 Suppl:S64–8, 1994.
- Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
- Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. *Advances in neural information processing systems*, 33:20612–20623, 2020.
- Ishita Mediratta, Qingfei You, Minqi Jiang, and Roberta Raileanu. The generalization gap in offline reinforcement learning. *arXiv preprint arXiv:2312.05742*, 2023.
- Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4500–4509, 2018.

-
- Chawin Ounkomol, Sharmishta Seshamani, Mary M Maleckar, Forrest Collman, and Gregory R Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature Methods*, 15(11): 917–920, 2018.
- Daniele Pirone, Joowon Lim, Francesco Merola, Lisa Miccio, Martina Mugnano, Vittorio Bianco, Flora Cimmino, Feliciano Visconte, Annalaura Montella, Mario Capasso, et al. Stain-free identification of cell nuclei using tomographic phase microscopy in flow cytometry. *Nature photonics*, 16(12):851–859, 2022.
- L Qiao, M Koutsos, L L Tsai, V Kozoni, J Guzman, S J Shiff, and B Rigas. Staurosporine inhibits the proliferation, alters the cell cycle distribution and induces apoptosis in ht-29 human colon adenocarcinoma cells. *Cancer letters*, 107(1):83–9, Oct 1996.
- Ando Saabas. Selecting good features - part iii: Random forests, 2014. URL <https://blog.datadive.net/selecting-good-features-part-iii-random-forests/>. Data Science Central Blog.
- Yutaka et al. Sasaki. The truth of the f-measure. *Teaching, Tutorial Materials, Version: 26th October*, 1(5):1–5, 2007.
- Jyrki Selinummi, Pekka Ruusuvuori, Irina Podolsky, Adrian Ozinsky, Elizabeth Gold, Olli Yli-Harja, Alan Aderem, and Ilya Shmulevich. Bright field microscopy as an alternative to whole cell fluorescence in automated analysis of macrophage images. *PLoS One*, 4(10):e7497, 2009.
- Carsten Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.
- Paweł Szymański, Magdalena Markowicz, and Elżbieta Mikiciuk-Olasik. Adaptation of high-throughput screening in drug discovery—toxicological screening tests. *International journal of molecular sciences*, 13(1):427–452, 2011.
- Samuel Tonks, Chih Hsu, Steve Hood, Ryan Musso, Ceriden Hopely, Minh Doan, Erin Edwards, Alexander Krull, and Iain Styles. Evaluation of virtual staining for high-throughput screenings. In *20th IEEE International Symposium on Biomedical Imaging*. IEEE, 2023.
- Uddeshya Upadhyay, Yanbei Chen, Tobias Hepp, Sergios Gatidis, and Zeynep Akata. Uncertainty-guided progressive gans for medical image translation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 614–624. Springer, 2021.
- Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.
- Sophia J Wagner, Christian Matek, Sayedali Shetab Boushehri, Melanie Boxberg, Lorenz Lamm, Ario Sadafi, Dominik JE Waibel, Carsten Marr, and Tingying Peng. Make deep learning algorithms in computational pathology more reproducible and reusable. *Nature Medicine*, 28(9):1744–1746, 2022.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- Håkan Wieslander, Ankit Gupta, Ebba Bergman, Erik Hallström, and Philip John Harrison. Learning to see colours: Biologically relevant virtual staining for adipocyte cell images. *Plos one*, 16(10):e0258546, 2021.
- J E Wilson, D E Brown, and E K. Timmens. A toxicologic study of dimethyl sulfoxide. *Toxicology and Applied Pharmacology*, 7:104–12, Jan 1965.