Deep Learning for robust phenotyping of high content cellular images

Yusuf Roohani[1], Ann Hoffman[2], Ryan Musso[2], Nicola Richmond[3]

Author Affiliations:

[1] Platform Technology and Science (PTS) R&D, GSK, Cambridge, MA [2] PTS R&D, GSK, Upper Providence, PA [3] PTS R&D, GSK, Stevenage, UK

Abstract

Image-based high content screening is a well-established method for phenotypic screening.[1] However, tapping the full extent of the high data content of these images demands a new approach to image analysis that is more question-agnostic. Conventional approaches are bound by predefined features, and the processes involved – manually designing, selecting and extracting features – are long, arduous and require a moderate level of expertise with image analysis tools.

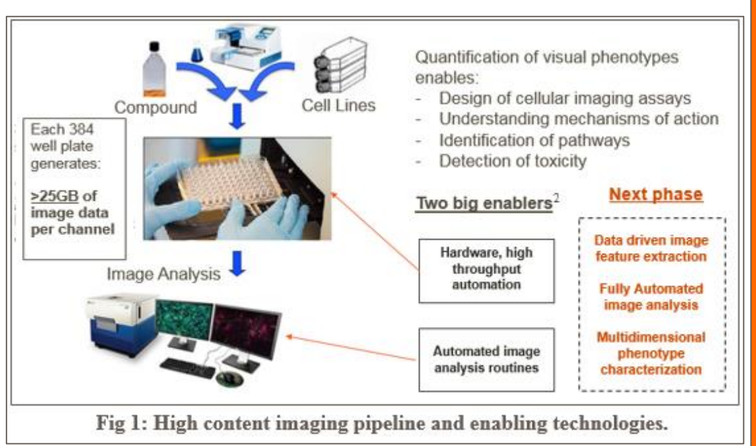


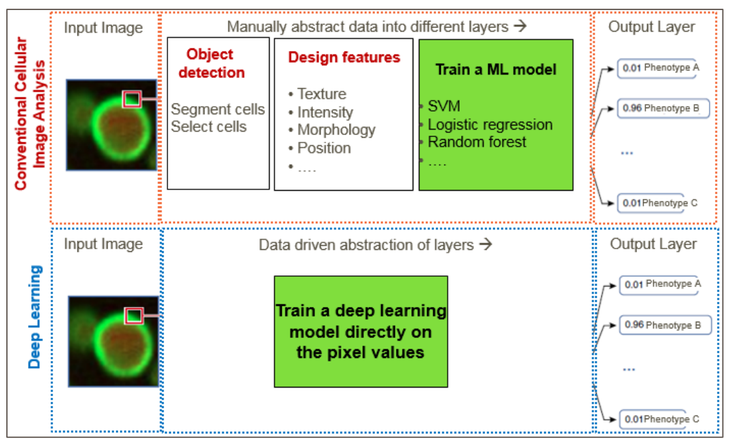
Fig 1: High content imaging pipeline and enabling technologies.

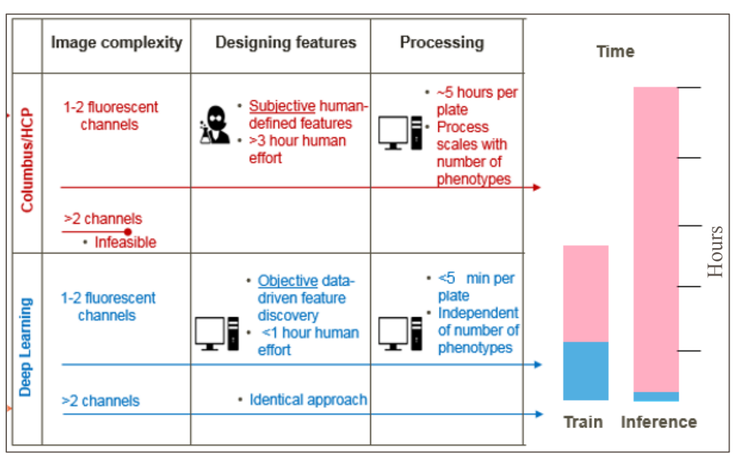
Here, we explore the application of deep convolutional neural networks (CNN) to address these problems. These models are able to automatically extract informative features from raw pixel values thus significantly reducing time and effort while also enabling a richer phenotypic analysis of these images. We then refine this model using unsupervised approaches, including variational autoencoders and generative adversarial networks. The goal here is twofold – to discern informative image features that are not specific to a given problem and to enhance robustness such that the classifier can generalize to parts of the phenotypic space that it would otherwise not have had information about. This poster will describe promising early results for supervised and unsupervised algorithms and will present our approach for industrializing CNN workflows in order to provide immediate benefit for phenotype characterization and assist in interpretation of results.

1. Automating Analytics

Conventional image analyses typically start by object segmentation followed by extraction of relevant features and, optionally, the training of a machine learning based classifier.[1] Using deep learning we have an entirely data driven approach where the model extracts the most relevant features from the data itself.

Shaded boxes indicate components that can learn from data

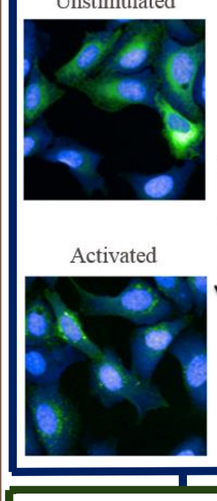




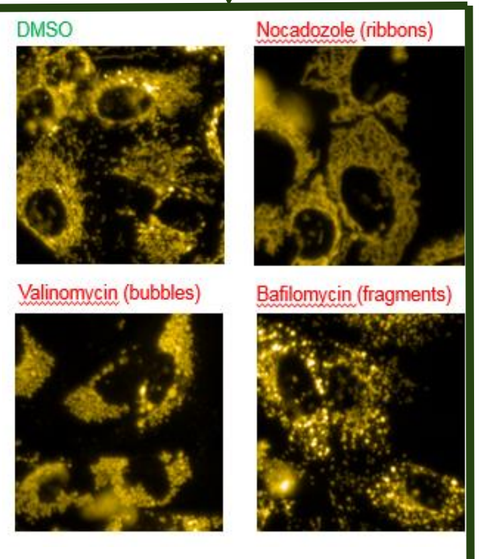
2. Industrializing Deep Learning Workflows

Run assay and read image data

A **Goal:** Automate an image analysis workflow for detecting translocation of a fluorescent protein from the cytoplasm to peri nuclear localization.



Dataset: Cells transduced with eGFP-protein via bacmam and activated using additional bacmam added as dsDNA stimulus Activation defined as puncta formation and translocation to peri-nuclear localization
Training set: Single 384 well plate - 2048x2048 pixel images (40x, 4 fields per plate)



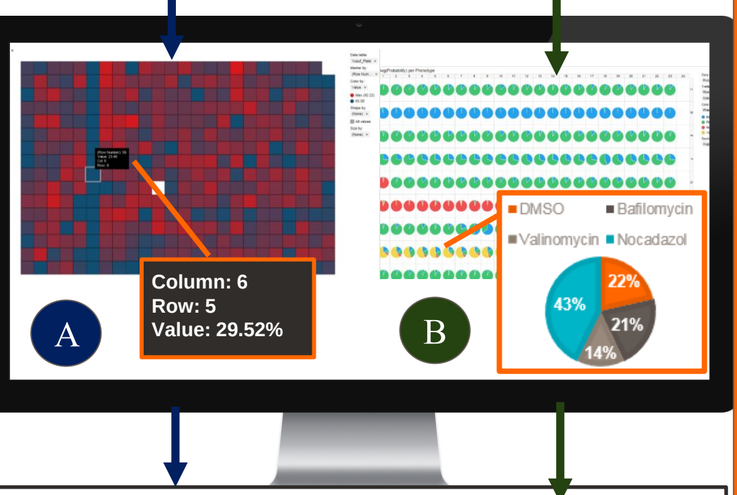
Dataset: Cells treated with 3 known mitotoxic compounds or DMSO
Training set: Single 384 well plate - 12 fields per plate (40x) - 1024x1024 pixels

Goal: Apply CNNs to two channel cellular images with fluorescently labelled mitochondria treated with a set of compounds expressing a range of mitotoxicity. The model is trained to generate a phenotypic fingerprint within a space described by these known mitotoxic phenotypes. **B**

A Deep Learning based Image Analytics **B**

Simply provide a control plate to automatically train a model

Use a model trained on a library of phenotypes to perform phenotypic fingerprinting



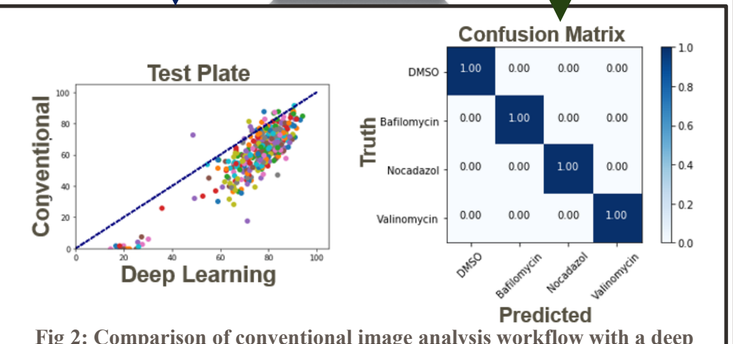


Fig 2: Comparison of conventional image analysis workflow with a deep learning based workflow from a model training perspective

Linear regression output (X-axis): Percentage of responders per well
Deep learning output (Y-axis): Probability of well showing activation

Confusion matrix showing output from deep learning model

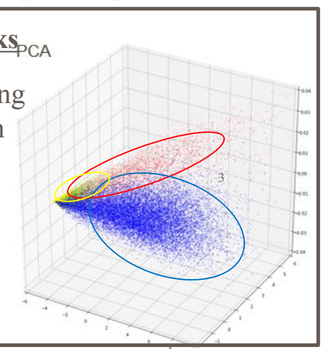
MODEL OUTPUT ACCURACY

3. Increasing Phenotypic Information

The phenotypic fingerprint can be conceived of as a vector quantifying similarity to known phenotypes. This is achieved along dimensions (features) identified by the model as being most discriminative. These features may not necessarily have obvious biological meaning but can, in most cases, be visualized through sampling different clusters or using saliency maps. Below, we use low dimensional embeddings to visualize the three different mitotoxic phenotypes and control.

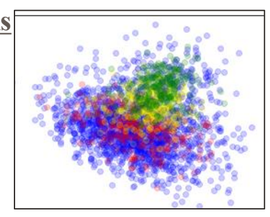
2a. Convolutional Neural Networks

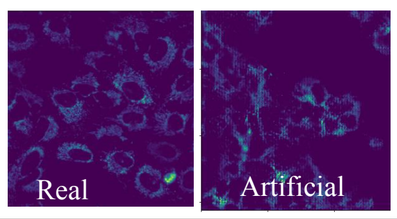
This is a standard supervised learning approach where the model is shown the four different phenotypes and trained to identify different classes. With a sufficiently large library of phenotypes, a supervised approach could prove robust for this task.

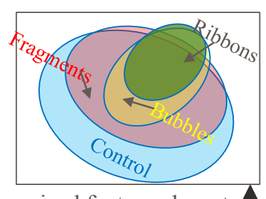


2b. Generative Adversarial Networks

Here, we apply adversarial training using only the control dataset. In this technique two separate models train against each other – the 'generator' generates new samples, while the 'discriminator' learns to detect artificial examples.[4]



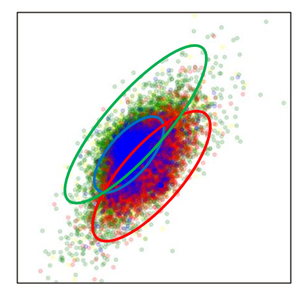


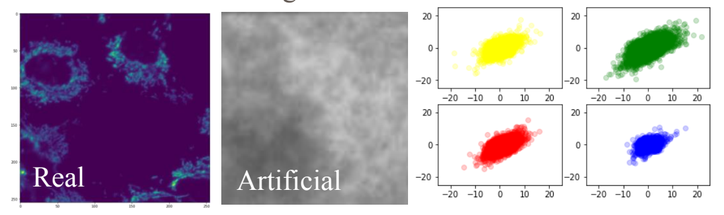


Unsupervised features learnt by the discriminator using only unlabeled DMSO images

2c. Variational Autoencoders

VAEs are a type of autoencoder that impose additional constraints on the distribution used to model the data. It has also been found to produce more distinct classes than regular autoencoders.[5] We trained the VAE on DMSO images alone.





DMSO Bafilomycin Valinomycin Nocadazol

4. Conclusions and Ongoing Work

Ongoing Work

Infrastructure: New pipelines and data storage solutions.
Generalization: Using unsupervised approaches to ensure robustness to novel phenotypes with large datasets
Representation: Interpreting model outputs and providing biological reasoning

Summary

We applied deep learning as a means of automating routine analytic procedures performed by screening groups. The solution we aim to industrialize would maintain accuracy while significantly reducing time and human resource required.
We are also actively exploring deep learning as a means of expanding the actionable phenotypic information available to the scientist. For this effort, we have looked at unsupervised techniques as a means of attaining a clearer distinction between phenotypic classes while also enhancing robustness. So far, using a supervised algorithm on a large enough library of phenotypes appears to be the optimal approach.

References:
[1] Godinez, W. J., et al.. (2017). *Bioinformatics.* [2] Abraham, V. C. et al. (2004). Trends in biotechnology, 22(1), 15-22. [3] Goodfellow, I., et. al (2016). *Deep learning*. MIT Press, [4] Goodfellow, Ian, et al. NIPS (2014), [5] Kingma, Diederik P. arXiv:1306.0733 (2013)

Created by: Digital Media Services