LAST CHANCE LIT

JOSEPH BOYD

Contents

14. Dürr, Oliver, and Beate Sick. "Single-cell phenotype classification using deep convolutional neural networks." Journal of biomolecular screening 21.9 (2016): 998-1003. 11
15. Kandaswamy, Chetak, et al. "High-content analysis of breast cancer using single-cell deep transfer learning." Journal of biomolecular screening 21.3 (2016): 252-259. 11
16. Perlman, Zachary E., et al. "Multidimensional drug profiling by automated microscopy." Science 306.5699 (2004): 1194-1198. 12
17. Singh, Shantanu, Anne E. Carpenter, and Auguste Genovesio. "Increasing the content of high-content screening: an overview." Journal of biomolecular screening 19.5 (2014): 640-650. 12
18. Adams, Cynthia L., et al. "[24]-Compound Classification Using Image-Based Cellular Phenotypes." Methods in enzymology 414 (2006): 440-468. 13

1. Neumann, Beate, et al. "Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes." Nature 464.7289 (2010): 721-727.

This study is motivated by the need to understand the connections between the 21000 genes in the human genome and some of the basic cellular functions, in pariticular, mitosis. For this purpose, a highly developed screening platform was used to analyse the phenotypic profiles of cell populations subject to *RNA interference*. This interference is achieved with small interference RNA (siRNA), synthetic RNA molecules that reduce gene expression mRNA by a significant amount (by an average of 87% in the study).

The data consisted of two days worth of time-lapse fluorescence microscopy, with the green fluorescent protein (GFP) tagging the core histone 2B, a protein within chromosomal histones. Hence, cell nuclei alone were recorded. The usual pipeline was used: segmentation, feature extraction (200 features), and classification of cell nuclei into 16 phenotype classes by an SVM trained on 3000 annotated samples and achieving 87% accuracy. As time-lapse microscopy, (not an "end-point" assay), mitotic states are more easily detected. Figure 1B shows interestingly how dominant the interphase class is (82.2%). For each phenotypic class, the proportion was tracked over time and compared with a negative control. The maximum difference in proportions over time was used as an index. Potential hits were identified to be siRNA yielding indices in at least four phenotypic classes below a certain threshold. Apart from the identification of hits, several analyses were conducted: event order maps were used to track the progression of phenotypes to discover the chronology and causality of each. Hierarchical clustering was performed on phenotypic profiles.

The study produced a publicly available dataset known as the MitoCheck dataset[1].

[1] http://www.mitocheck.org/

2. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

The Visual Geometry Group (VGG) at Oxford produced a streamlined CNN achieving greater depths and taking second place in the ImageNet 2014 competition. They focussed on a simple 3×3 kernel in every layer, with blocks of convolutional layers separated with max pooling, and the architecture ending with affine layers. In effect, it is just like a deeper AlexNet, but with 3×3 convolutional layers used everywhere. As they note, though the effective receptive field of a single 7×7 layers and three 3×3 layers are the same, the former has $49C^2$ parameters for C activation maps, whereas the latter has only $27C^2$. As they note,

> [VGG Net[2] architecture] did not depart from the classical ConvNet architecture of LeCun" [unlike GoogLeNet with its "inception" layers] (page 8)

and, as such, can be see as a bridge between AlexNet and ResNet.

3. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

ResNet represents an innovation on classical CNNs to address the "degradation" problem of training very deep networks. This phenomenon is distinct from overfitting as it refers to a degradation of *training error* when more and more layers are added. VGG showed that depth is usually more important than width (larger kernels), and ResNet is a successful attempt to reformulate CNNs to achieve extreme depths. The authors hypothesise that the way to achieve superior generalisation in deep neural nets is to make small changes to the feature representation over many layers rather than a dramatic transformation of inputs: "multiple nonlinear layers can asymptotically approximate complicated functions" (page 3). Therefore, they introduce residual learning where, after each two layers[3] of convolutions, the outputs are computed as,

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x},$$

where $\mathcal{F} = W_2\sigma(W_1\mathbf{x})$. They observe that it is much easier for a network to find $||\mathbf{F}||_2 \approx \mathbf{0}$ (i.e. the residual) than to find $\mathcal{F}$ that creates $\hat{\mathbf{x}} \approx \mathbf{x}$, that is, an effective identity mapping, which is what the network needs for good generalisation. Despite the benefits, adding the input comes at negligible cost to compute, and does not affect backpropagation, as it is a constant for future layers.

They present results on the ImageNet 2015 challenge where they won first prize, and also results on their theory that ResNets make smaller contributions per layer. They also

[2] A name attributed afterwards

[3] Larger groups are also tried, but if it were every layer, it would just be a linear mapping (prior to the activation function)

run ResNet on CIFAR-10 with up to 1202 (!) layers. They finally discuss ResNet's use as the CNN in object detection frameworks (*Faster R-CNN*).

4. Ljosa, Vebjorn, et al. "Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment." Journal of biomolecular screening 18.10 (2013): 1321-1329.

This is a survey of different approaches to image-based profiling, that is, characterising cell populations by some index as a means of predicting the mechanism of action[4] (MOA) of a particular compound. As they state,

> Image-based screens for particular cellular phenotypes are a proven technology contributing to the emergence of high-content screening as an effective drug- and target-discovery strategy [...] may help reduce the high levels of late-stage project attrition associated with target-directed drug-discovery strategies (page 2)
>
> Profiling cell-based phenotypes is the next challenge for quantitative microscopy. The principle of phenotypic profiling is to summarize multiparametric, feature-based analysis of cellular phenotypes of each sample so that similarities between profiles reflect similarities between samples (page 2)
>
> Profiling is well established for biological readouts such as transcript expression and proteomics. Comparatively, image-based profiling comes at a much lower cost, can be scaled to medium and high throughput with relative ease, and provides single- cell resolution. (page 2)

Thus, the authors extract the "core profiling methods" from five methodologies that create "per-sample" profiles from "per-cell" measurements. They test them on the MCF-7 cell line on 96-well plates, treated with 113 compounds at 8 concentrations in triplicate ($\implies$ 29 plates at least) and marked with fluorescent tags. The resulting images were analysed with CellProfiler[5] producing 453 features for each cell. The methods were used to create profiles of a certain size, and the element-wise median over the triplicate was used. The prediction of mechanism of action was by nearest neighbour (1-NN) in cosine distance,

$$d(p, q) = 1 - \cos\theta_{p,q}$$

from a pool of labelled compounds. The five methods are each quite distinct. The most interesting is perhaps the "normal to SVM hyperplanes". Here the profile is the normal vector of the SVM hyperplane separating the treated sample cells and a negative control (DMSO). Dimensionality reduction was done by training the SVM on all D measurements, then $D-1, D-2, \ldots, 1$, with the dimensionality chosen to be the one maximising accuracy. This scheme is known as SVM recursive feature elimination (SVMRFE). Despite this and

[4]Mechanism of action is a pharmacological term referring to the biochemical reaction between the drug and the organism. The biochemical reaction usually makes reference to an enzyme or receptor protein.

[5]http://cellprofiler.org/

other sophisticated approaches, the best approach was simply to take the mean of each of the 453 features over the whole population, with a predictive accuracy of 88%.

5. Myers, Gene. "Why bioimage informatics matters." Nature methods 9.7 (2012): 659.

In this short article from 2012, prominent researcher Gene Myers characterises bioimage informatics as an emerging field of analysis offering insights into cell structure and behaviour lost in classical -omics data, giving examples thereof. He paints the bioimage informatician as either a computer vision practitioner seeking new problems, or biologists embracing the world of data science. He contends bioimage informatics matters for two reasons: because of the production of large image data sets and because of the trend in bioinformatics to answer questions pertaining to spatial properties of cells.

> From my perspective, it is very reminiscent of the state of bioinformatics in the early 1980s: the exciting, somewhat chaotic free-for-all that is potentially the birth of something new. (page 1)

6. Ganin, Yaroslav, and Victor Lempitsky. "Unsupervised domain adaptation by backpropagation." International Conference on Machine Learning. 2015.

This paper develops a new approach to domain adaptation in feed-forward neural nets, where a large labelled source set is available and an unlabelled (or partly labelled) target set is available[6]. The approach is compatible with any feed forward network. It involves training a classifier for the source data set, and a classifier to distinguish between domains, as part of the same loss function. The parameter set, $\theta = \{\theta_f, \theta_y, \theta_d\}$ consists of the weights for the shared deep feature layers, θ_f, the weights of the classifier layers, θ_y, and the weights of the domain discriminator, θ_d. The domain discriminator loss is scaled by a negative tuning parameter, λ. Hence, minimising the full loss function for the deep features θ_f *maximises* the loss of the discriminator. Thus, deep features are chosen that are invariant between the domains, as they are chosen to be maximally indistinguishable between the domains. Thus, features that are both useful for classification and invariant to the domain shift are discovered. Simultaneously, the parameters θ_d are chosen to minimise the loss. Clearly, if they also maximised the loss, there would be no onus on the deep features to be invariant. Formally, the loss function is $\mathbb{E}(\theta_f, \theta_y, \theta_d) = \mathcal{L}_y(\theta_f, \theta_y) - \lambda\mathcal{L}_d(\theta_f, \theta_d)$. Then,

[6]Note that the difference between domain adaptation and transductive transfer seems very slight: in this study (domain adaptation), features are learned with respect to both domains so that they will be useful to both. Nevertheless, in the end, a single classifier is learned. In an example of transductive transfer, a transfer function is learned over the parameters of source classifiers, and used to create target classifiers directly, based only on the (unlabelled) target domain. The fact that the source and target models are distinct seems to be the only difference.

$$\begin{aligned}
\hat{\theta}_f &= \min_{\theta_f} \mathbb{E}(\theta_f, \hat{\theta}_y, \hat{\theta}_d) \\
&= \min_{\theta_f} \mathcal{L}_y + \lambda \min_{\theta_f} \{-\mathcal{L}_y\} = \min_{\theta_f} \mathcal{L}_y + \lambda \max_{\theta_f} \{\mathcal{L}_y\} \\
\hat{\theta}_y &= \min_{\theta_y} \mathbb{E}(\hat{\theta}_f, \theta_y, \hat{\theta}_d) = \min_{\theta_y} \mathcal{L}_y \\
\hat{\theta}_d &= \max_{\theta_d} \mathbb{E}(\hat{\theta}_f, \hat{\theta}_y, \theta_d) = \max_{\theta_d} \{-\mathcal{L}_d\} = \min_{\theta_d} \mathcal{L}_d
\end{aligned}$$

These lead to the set of rules for optimisation,

$$\theta_f \leftarrow \theta_f - \mu \left(\frac{\partial \mathcal{L}_y^i}{\partial \theta_f} - \lambda \frac{\partial \mathcal{L}_d^i}{\partial \theta_f} \right)$$

$$\theta_y \leftarrow \theta_f - \mu \frac{\partial \mathcal{L}_y^i}{\partial \theta_y}$$

$$\theta_d \leftarrow \theta_f - \mu \frac{\partial \mathcal{L}_d^i}{\partial \theta_d}$$

However, in practice SGD solvers are cannot both minimise and maximise a loss function. Therefore, the authors introduce a *gradient reversal layer*, defined by,

$$R_\lambda(\mathbf{x}) = \mathbf{x}$$

$$\frac{\partial R_\lambda}{\partial \mathbf{x}} = -\lambda \mathbf{I}$$

noting that the derivative is not "compatible", to reformulate the loss function such that the gradients are correctly calculated by any standard SGD solver as a matter of course. This amounts to a computational hack. While it is indeed true that a deep learning framework such as TensorFlow allows the user to specify the gradient arbitrarily, it is not clear why the user could not just specify the update rules manually instead. The paper concludes with the presentation of results and t-SNE visualisation of how the features distributions overlap after domain adaptation.

7. Zhang, Ji-Hu, Thomas DY Chung, and Kevin R. Oldenburg. "A simple statistical parameter for use in evaluation and validation of high throughput screening assays." Journal of biomolecular screening 4.2 (1999): 67-73.

Any HTS assay is subject to random variation as well as systematic measurement error. A typical validation step in HTS is to assure statistics measured between controls and test samples can be distinguished with confidence. The *Z-factor* expresses the ratio of the separation band $|\mu_c - \mu_s| + 3\sigma_c + 3\sigma_s$ (that is, the distance between the opposing tails of

the distributions at three standard deviations) and the dynamic range, $|\mu_c - \mu_s|$ between the control c and the test sample s. Thus, the Z factor,

$$Z = 1 - \frac{3\sigma_c + 3\sigma_s}{|\mu_c - \mu_s|}$$

The maximum (ideal) value is therefore 1. In the field, it is an accepted fact that anything below 0 is considered as a readout lacking robustness. The related Z'-factor calculates the same statistic for the positive and negative controls. The positive and negative controls should represent the high and low of the range of measurements of a property of interest. The Z'-factor thus refers to the overall quality of the assay.

8. Loo, Lit-Hsin, Lani F. Wu, and Steven J. Altschuler. "Image-based multivariate profiling of drug responses from single cells." Nature methods 4.5 (2007): 445-453.

In this monumental paper, the authors introduce a thorough framework for phenotypic profiling. The framework centers around using a linear SVM to compare test compounds at (13) different dosage levels with a negative control (DMSO) in a large assay. The normal[7] to the separating hyperplane (the SVM weight vector) is used to characterise the test compound-dosage combination. 100 compounds are tested at 13 different levels, giving a *titration*[8] series for each. The negative control is dimethyl sulfoxide (DMSO) alone (noting that all tests use a DMSO substrate). The cells are segmented using the watershed transformation[9] and several hundred features are extracted per cell. These include Haralick and Zerneke features[10]. The dimensionality is reduced using the technique SVMRFE (recursive feature elimination), which is a backward elimination wrapper method for feature selection (Guyon et alii). The profiles are clustered (titration clustering) with multiple clusters indicating *multiphasic* responses at different dosage levels. The cluster

[7]A reminder on normal vectors: the normal vector of a function $z = f(x, y)$ is $\mathbf{n} = [f_x, f_y, -1]^T$. This is equivalently the gradient of an auxiliary function, $g(x, y, z)$ with *level curve* defined by the relation $f(x, y) - z = 0$. Recall the gradient, ∇f of a function is always normal to the local level curve or contour of that function. This may be seen from the definition of *directional derivative*, defined as $\nabla_{\mathbf{v}} f(\mathbf{x}) = \lim_{h\to 0}(f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x}))/h$. This quantity is 0 for a choice of $\mathbf{v}$ to be tangential to the level curve, since then $f(\mathbf{x} + h\mathbf{v}) = f(\mathbf{x})$. Furthermore, $\nabla_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v}$, and therefore the gradient is normal to the tangent vector of the level curve. N.B. this is most easily observed visually.

[8]A titration is an experimental process for determining the concentration of a certain substance by incrementally adding a reactant.

[9]The watershed transformation is a technique for image segmentation. It works by "flooding" the topological landscape of the (inverted) image. Where catchment basins intersect, one marks *watershed lines* that correspond to the segmentation contours.

[10]Haralick features derive from the co-occurrence matrix of an image. The co-occurrence matrix is a matrix whose elements count the number of times pairs of pixel intensities occur in the image at a given offset. The elements are thus indexed by the pixel values themselves, and hence the co-occurrence matrix is square with dimension the dynamic range of the image. Haralick features somehow derive from this and are used to measure texture. Zernike features are features based on the Zernicke polynomials that describe shape.

constituents are aggregated to form a representative dosage range profile or *d-profile*. The profiles are then used for:

> Typical applications of high-throughput image- based assays, such as drug screening, phenotypic change detection and category prediction(page 3)

The study encompasses a rich assortment of analyses and heuristic decisions that are too numerous to list. Two interesting techniques are multidimensional scaling (MDS)[11] for visually comparing the profile clusters and kernel density estimation[12] (KDE) in the resampling procedure. Note that this study predates the development of CellProfiler, and much of the software was custom made by the authors.

9. Swinney, David C., and Jason Anthony. "How were new medicines discovered?." Nature reviews Drug discovery 10.7 (2011): 507-519.

The authors indicate that despite an increase in funding in drug screening, the annual number of FDA-approved[13] innovative drugs (termed "first-in-class") does not rise accordingly. They attribute this to the predominance of target-based approaches in drug discovery:

> Since the dawn of the genomics era in the 1990s, the main focus of drug discovery has been on drug targets (page 1)

Target-based approaches encompass a principled approach to drug discovery: identify a target protein that plays a role in a disease, and design a treatment to alter this protein. RNA interference is one tool involved in the identification of targets, whereby targeted mRNA is down regulated and the effects (transcriptomic, phenotypic, etc.) noted. On the other hand, these hypotheses may be based on spurious or incomplete evidence: often there is not just one smoking gun gene behind disease pathogenesis[14], rather a combination of genes. The authors indicate that the marginalisation of phenotypic screens by the pharmaceutical industry has resulted in the stagnation of innovation drug treatments, with a high *attrition rate* in the process.

> A strength of the phenotypic approach is that the assays do not require prior understanding of the molecular mechanism of action (MMOA), and activity in such assays might be translated into therapeutic impact in a given disease state more effectively than in target-based assays, which are often more artificial. A disadvantage of phenotypic screening approaches is the

[11]Multidimensional scaling is a technique for visualising the similarities of data points in lower dimensions. It is thus similar to t-SNE or Sammon mapping.

[12]Kernel density estimation or *Parzen window* density estimation is a non-parametric technique for inferring the probability density function generated a set of sample observations. It is done by choosing a kernel (probability distribution) that is placed at each observation. The estimate is the sum of all these overlapping functions. It is thus similar to a histogram. The choice of kernel and bandwidth (variance) parameter is made by the implementer.

[13]FDA is the Food and Drug Administration, a wing of the US government responsible for regulating food safety, and drug sales.

[14]Either the biological mechanisms responsible for or the development of a disease.

> challenge of optimizing the molecular properties of candidate drugs without the design parameters provided by prior knowledge of the MMOA. An additional challenge is to effectively incorporate new screening technologies into phenotypic screening approaches, which is important for addressing the traditional limitation of some of these assays: a considerably lower throughput than target-based assays. (page 2)

In summary, it is perhaps more effective to try out many likely compounds and see what happens than to try to identify a particular one. Where target-based approaches are more effective, however, is in the creation of *follower drugs*, third party equivalents to first-in-class drugs. Note the *molecular mechanism of action* (MMOA) refers to the specific molecular interaction between the drug and the target. Mechanism of action (MOA) refers to a physiological response (e.g. anti-inflammatory).

The study looks at 257 drugs published between 1999 and 2008. The findings are that, despite the preeminence of target-based approaches, the most common mode of first-in-class drug discovery is phenotypic screening. This is most true of infectious and central nervous system diseases. Cancer treatments are most frequently discovered by biologics[15], which also predominate for diseases of the immune system. Target-based approaches succeed for the discovery of half of the follower drugs. The authors postulate the high attrition rate of target-based approaches to be due to the failing of one or more of three hypotheses that must all succeed for a discovery to be made:

> The first hypothesis, which also applies to other discovery approaches, is that activity in the preclinical screens that are used to select a drug candidate will translate effectively into clinically meaningful activity in patients. The other two hypotheses are that the target that is selected is important in human disease and that the MMOA of drug candidates at the target in question is one that is capable of achieving the desired biological response. (page 9)

The two main weaknesses of phenotypic screens are identified as being: the necessity of finding the MMOA *a posteriori* and; the lower throughput bottleneck.

This paper also uses a lot of useful terminology: preclinical strategies, potential drug candidates, target-based, phenotypic screens, modification of natural substances, biologic-based approaches, molecular mechanism of action (MMOA), pharmacological response, allosteric or orthosteric.

10. Orlov, Nikita, et al. "WND-CHARM: Multi-purpose image classification using compound image transforms." Pattern recognition letters 29.11 (2008): 1684-1693.

Weighted neighbor distances using a compound hierarchy of algorithms representing morphology (WND-CHARM) is a classification framework designed to be versatile over many problems. It works by extracting a large number of image features (> 1000) before

[15]Biologics are genetically-engineered proteins that target the immune system.

weighting them using Fischer scores (soft thresholding). The features with lowest weights are then discarded (hard thresholding). Test samples are then classified with a 1-NN with a particular distance measure. This they call weighted neighbour distances (WND).

> The advent of high content screening (HCS) where the goal is to search through tens of thousands of images for a specific target morphology requires a flexible classification tool that allows any morphology to be used as a target.

(page 1)

WND-CHARM is compared to tailored models on various datasets, including the **HeLa dataset** where it performs favourably.

11. Uhlmann, Virginie, Shantanu Singh, and Anne E. Carpenter. "CP-CHARM: segmentation-free image classification made accessible." BMC bioinformatics 17.1 (2016): 51.

> Segmentation-free image classification using whole-image features has already been widely used in computer vision, especially for image databases [1, 2]. It is however less popular for bioimage analysis, where segmentation remains the most common paradigm.

This is a confusing point to make. Segmentation and whole/full image classification were never a focus of WND-CHARM. Even for the biological datasets, they were all pre-segmented anyway (the HeLA dataset consists of crops of cells). Therefore, CP-CHARM is also reinterpreting the purpose of WND-CHARM. CP-CHARM is an adaptation of WND-CHARM to comprise more conventional technologies. It replaces the custom feature extraction with CellProfiler (this is what CP stands for), feature weighting is replaced with PCA, WND is replaced with LDA, and 75 : 25 cross validation is replaced with 10-fold cross validation.

12. Haney, Steven A., et al. "High-content screening moves to the front of the line." Drug discovery today 11.19 (2006): 889-894.

In this older paper, the authors describe how, with the advancement of relevant technologies, high content screening (HCS) has moved to the early stages of preclinical screening[16], and *hit-to-lead* stages. The authors identify the strengths of HCS are in its precision: multivariate per-cell analysis. Assays are further *multiplexed* with multiple fluorescent probes (usually up to 3 or 4).

13. Shay, Jerry W., and Woodring E. Wright. "Hayflick, his limit, and cellular ageing." Nature reviews Molecular cell biology 1.1 (2000): 72-76.

This short article traces the history of the *Hayflick limit*, the upper limit on the number of possible cell divisions. This is due to the eventual loss of the *telomeres* at the ends of

[16]Preclinical refers to experiments done before *clinical trials*, that is, tests involving humans.

the DNA strands which finally causes senescence, a cessation of mitosis. This originates with American medical scientist Leonard Hayflick (1928-), who defied the earlier assertion of eminent scientist Alexis Carrel:

> The largest fact to have come from tissue culture in the last fifty years is that cells inherently capable of multi- plying will do so indefinitely if supplied with the right milieu in vitro. (page 2)

Cancel cells therefore need to maintain telomeres to continue metastasis.

14. Dürr, Oliver, and Beate Sick. "Single-cell phenotype classification using deep convolutional neural networks." Journal of biomolecular screening 21.9 (2016): 998-1003.

In this simple study, the authors pit a CNN against shallow models trained on CellProfiler-extracted features for classification of 40,000 images from the "cell painting" assay, BBBC022v1. This is the U2OS cell line–human bone epithelial cells. The classification task is to identify the MOA (MMOA?) of a drug from four classes (including the null–DMSO) given a cropped cell nucleus. **Note this is different from classifying the cell phenotype itself.** The dataset is imbalanced, with a preponderance of DMSO training samples. The CNN outperforms (by a dubious amount) the best of the three shallow models–LDA. The CNN architecture resembles VGG net, albeit with an input size of $(5 \times 72 \times 72)$ (a channel for each of the five stains). It would probably play more to the strengths of a CNN to classify phenotype itself.

15. Kandaswamy, Chetak, et al. "High-content analysis of breast cancer using single-cell deep transfer learning." Journal of biomolecular screening 21.3 (2016): 252-259.

Refers to high content analysis assays and high-throughput image-based assays interchangeably. Here, the BBBC021 dataset is used, the same as in Ljosa et al. (2013). They divide the 12 MOA class dataset into two (6 MOA classes each). They train stacked autoencoders (which they call autoassociators for some reason) on the 453-dimensional data (extracted with CellProfiler N.B. this is not pixel data). This unsupervised learning learns a succession of hidden layers, each compressing the last as an autoencoder. The learned weights are then used to initialise a deep architecture with a (6-way) logistic output that is fine-tuned on the labeled data. This alone has good performance on per-cell MOA classification. They try transferring layers from each domain to the other ($P_1 \to P_2$ and $P_2 \to P_1$) at four different depths. The results are varied and generally unconvincing. In fact, this contrived transfer framework feels rather wanton. Their comparison with an SVM (20% accuracy??) is bizarre. They further contrast their work with the profiling techniques for MOA prediction, but without clarifying how their method makes a population-level prediction of MOA. Presumably by using the mode prediction.

16. Perlman, Zachary E., et al. "Multidimensional drug profiling by automated microscopy." Science 306.5699 (2004): 1194-1198.

This is a short yet important paper. It studies 100 compounds with different known toxicity or therapeutic mechanisms in cancer, at different dosage levels. It is seminal foremostly for the following reason:

> To date [2004], drug effects have been broadly profiled with transcript analysis, proteomics, and measurement of cell line dependence of toxicity [...] Here, we suggest that large sets of unbiased measurements might serve as high-dimensional cytological profiles analogous to transcriptional profiles. We present a method that is hypothesis-free [...] (page 2)

although they only use 93 features. They are an obvious inspiration for, among others, the Loo paper.

> [...] profiling should be performed as a function of drug concentration.(page 2)

Their profile consist of a K-S statistic comparing the test cells to the control cells, for each feature. This gives a vector of K-S statistics that they further normalise to give their titration-invariant similarity score (TISS) (though this is not derived in the body).

They interestingly note that a histogram of total DNA content ought to be bimodal, given the interplay of cell cycle phases. Furthermore,

> G_2 and M populations may be distinguished by 2D display of total DNA signal against nuclear area. (page 2)

It also helpfully uses interesting technical terms: perturbations, systems level, micromolar to picomolar, "cultured in 384-well plates to near confluence, treated with drugs for 20 hours, fixed, and stained with fluorescent probes for various cell components and processes.", multiplexing, control population, arrested state, specificity and affinity, substrate.

17. Singh, Shantanu, Anne E. Carpenter, and Auguste Genovesio. "Increasing the content of high-content screening: an overview." Journal of biomolecular screening 19.5 (2014): 640-650.

This review remarks on the fact that although HCS articles become steadily more numerous from 2000 to 2012, the number of features used in the studies does not appear to increase, or at least far more slowly. Thus:

> [...] there is a strong tendency for HCS assays to typically be, in truth, quite low content. (page 2)

To demonstrate this, 118 papers were sampled fairly according to three search term strategies and their contents scrutinised. When visualised, their is only a weak trend towards higher content in HCS studies, despite a major rise in the amount of HCS research done. One interesting explanation for the low content in HCS studies is the adoption of the Z'-factor as an assay quality measure, whereas this was originally intended as an HTS property. There ensues an interesting discussion about why the Z'-factor may be

unsuitable for HCS. It is is problematic because measurements are collected at the cell level, and are then aggregated over phenotypic sub-populations. The lower throughput HCS assays may further mean controls are not replicated enough to meaningfully compare the distributions of their readouts. The authors admonish the use of the Z'-factor in the conclusion, though without concluding on a suitable replacement. The latter section summarise the five profiling methods compared in Ljosa et al. (2013).

> We predict that advanced data analysis methods that enable full multiparametric data to be harvested for entire cell populations will enable HCS to finally reach its potential. (page 2)

18. Adams, Cynthia L., et al. "[24]-Compound Classification Using Image-Based Cellular Phenotypes." Methods in enzymology 414 (2006): 440-468.

The very first sentence of the abstract is the fundamental assumption on which all else is predicated:

> Compounds with similar target specificities and modes of inhibition cause similar cellular phenotypes. (page 1)

The aim of the study is to classify compounds into 12 classes of mechanism of action, which they do to an accuracy of about 90% (45/51). This is motivated by the benefits of detecting unintended compounds in the early, *in vitro* stages of screening, prior to clinical trials. This has the added ethical motivation of reducing the number of trials with unknown effects carried out on animals. They again make the case for cytoskeletal biology and morphological-based screens, pointing out the phenomena that evade observation at the genetic and proteomic levels.

For this purpose, the authors, who were part of a firm called Cytokinetics, developed a system called Cytometrix Technologies that may or my not still exist. The Cytometrix system comprises all the stages of high content screening: seeding of cells, application of fluorescent markers, microscopy, and image analysis software. The aim is to compute a phenotypic signature for each compound across different incubation periods, drug concentrations, and genotypes (cell lines).

> Changes in cell cycle, cell shape, cytoskeleton organisation, and protein trafficking and machinery in response to compounds often results in concomitant cellular morphological changes. (page 2)

They argue that a smaller number of fluorescent markers over a greater number of cell lines will give a more robust characterisation, unlike in Perlman et al. (2004), which measured many markers on a single cell line. The following subsections detail the protocol: the cell lines used; the cell plating–24h incubation at body temperature; preparation of compounds–triplicates of each compound at 8 concentrations including positive controls, and except negative control DMSO (single concentration); compound addition; immunocytochemistry–pipetting, fixing (formaldehyde), staining, washing; and microscopy.

They go on to describe the features measured, which include various shape-, texture-, and intensity-related measurements. These features are averaged by the strata of cell cycle phase: interphase, mitotic, preanaphase, as well as the proportions of each cell cycle phase. In this way, approach encompasses our own, though the classification of cycle phase is not based on supervised learning; rather on uni- or bivariate classification.

These population- and sub-population-level attributes are normalised (the features are of course on different scales) as standard deviations to the negative control attributes. Various comparisons can be made. One is in comparing attributes of each cell line w.r.t. each treatment (Figure 3). Another mode of comparison is in dose-response curves (titration series).

To classify compounds, they take the normalised attributes for each cell line and concatenate them as a row (32 attributes $\times$ 6 cell lines). Rows corresponding to different compound-concentration combinations form a matrix. PCA is performed on this matrix to reduce the compound-concentration signatures to 2-3 dimensions. These are visualised in Figure 5.

A concentration-independent measure of comparing drugs is created as the integral of the angle between dose responses of two compounds over a range of common dosage levels. The integral is necessarily discrete as we only have a small number of discrete dosage level measurements. Their strategy is then to use the signature data (reduced dimensionality) to cluster (Figure 6) and classify the data. Classification involves firstly identifying an subset of attributes, and then to predict the MOA using a nearest-centroids classifier based square distance between signatures. This they do to a high degree of accuracy (88%).

Their conclusions are varied, the main findings being that 1) measuring fewer markers over multiple cell lines is a viable approach to quantifying phenotypic differences 2) phenotypes can change significantly before cell mortality kicks in.

At this point, it really seems the Adams approach is not well represented in the Ljosa survey (although they do note that), even though it performs well. There, the 459 CellProfiler features are simply averaged and used as a profile directly. Adams takes a well-defined set of per-cell features, uses a primitive classifier identify cell cycle sub-populations and stratify, then normalise w.r.t. negative controls, then identify an optimal subset with LOOCV on a nearest-centroids classifier. If we run with the simple approach of averaging, consider the following argument:

We denote our phenotypic profile as $\mathbf{p} \in \mathbb{R}^K$ for K phenotype classes. Observe that the manually defined phenotype ontology acts as a latent variable behind the distribution of features. When it comes to the straightforward approach of phenotype characterisation in Ljosa et al. (2013) (inspired by Adams et al. (2004)), the phenotype profile $\mathbf{p}' \in \mathbb{R}^D$ for D features is given as $\mathbf{p}' = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$ for the N observations in the data set. Yet, under the hypothesis of a latent phenotype variable,

$$\mathbf{p}' \to \sum_{k=1}^{K} p_k \mathbf{f}_k$$

as $N \to \infty$, where the p_k are the proportions of each of the K phenotypes as measured by our approach, and the $\mathbf{f}_k$ are the average feature vector for each phenotype–the archetypal phenotypes, as it were. As such, $\mathbf{p}' \to \mathbf{Fp}$ as $N \to \infty$, with $\mathbf{F} \in \mathbb{R}^{D \times K}$ the matrix of $\mathbf{f}_k$. Thus, the Adams profile is approximately equivalent to our own projected into a high dimensional linear space. It therefore seems unlikely that cosine distances calculated by the former approach will be as meaningful as with our own.

The difference between our profiling approach and the surveyed once is that our require manual intervention (annotation). The other techniques are indices taken from already available information. Even the approach of Loo et al. (2007) that relies on an SVM, it is a binary classifier trained to distinguish sample populations from negative controls, and the training data is by definition already annotated. A learning approach that circumvents this problem would need to be based on unsupervised or transfer learning techniques.