IMAGE INFORMATICS APPROACHES TO ADVANCE CANCER DRUG DISCOVERY

Scott J. Warchal

Doctor of Philosophy
The University of Edinburgh
2018

DECLARATION

This thesis presents my own work, and has not been submitted for any other degree or professional qualification. Wherever results were obtained in collaboration with others, I have clearly stated it in the text. Any information derived from the published work of others has been cited in the text, and a complete list of references can be found in the bibliography. Published papers arising from the work described in this thesis can be found in the appendices.

– Scott Warchal, 2018

epigraph here

ACKNOWLEDGEMENTS

Acknowledgements here.

ABSTRACT

Abstract here.

LAY SUMMARY

Lay summary here.

CONTENTS

Declaration i

Acknowledgements v

Abstract vii

Lay summary ix

Contents xi

List of Figures xii

List of Tables xiii

List of Acronyms xv

1 Introduction 1
- 1.1 Eroom's Law: The increasing cost of drug discovery . . . 1
- 1.2 The drug discovery process . . . 1
  - 1.2.1 Target-based screening . . . 1
  - 1.2.2 Phenotypic screening . . . 1
- 1.3 High content imaging . . . 2
  - 1.3.1 Image analysis . . . 3
  - 1.3.2 Data analysis . . . 4
  - 1.3.3 Image based screening . . . 6
  - 1.3.4 Image based profiling . . . 6
- 1.4 Phenotypic screening in cancer drug discovery . . . 7
  - 1.4.1 Cancer cell line panels . . . 7
- 1.5 Thesis structure . . . 8

2 Cell morphology can be used to predict compound mechanism-of-action 9
- 2.1 Introduction . . . 9
- 2.2 Creating an annotated dataset . . . 9
  - 2.2.1 Compounds . . . 10
  - 2.2.2 Cell painting: labelling cellular morphology . . . 10
- 2.3 Machine learning methods to classify compound MoA . . . 11

2.3.1 Comparison between classical and deep-learning machine learning methods 11
2.4 Predicting compound MoA on single cell lines 16
2.4.1 Ensemble of decision trees 16
2.4.2 CNNs 16
2.5 Transferring machine learning models to morphologically distinct cell lines 16
2.5.1 Leave-one-out classification 17
2.5.2 Do additional cell-lines during training increase prediction accuracy? . . . 17
2.6 Discussion 17
2.7 Methods 17
2.7.1 Cell culture 17
2.7.2 Compound handling 17
2.7.3 Staining 17
2.7.4 Imaging 18
2.7.5 Ensemble of decision trees 18
2.7.6 Image analysis: numeric features from images 18
2.7.7 Convolutional neural networks 18

3 MEASURING DISTINCT PHENOTYPIC RESPONSE 19
3.1 Section name 19

4 LARGE COMPOUND SCREEN ACROSS 8 BREAST CANCER CELL LINES 21
4.1 Section name 21

5 CHEMINFORMATICS 23
5.1 Section name 23

6 DISCUSSION AND CONCLUSION 25
6.1 Section name 25

LIST OF FIGURES

1.1 Single cell aggregation to a median profile 4

2.1 Diagram of a simple decision tree 12
2.2 Diagram neural network neuron and activation function. 14
2.3 Representation of a simple ANN 14
2.4 Down-sizing and chopping images for CNN training 16
2.5 Multi-GPU distributed training 18

LIST OF TABLES

1.1 Panel of breast cancer cell lines chosen for study 8

2.1 Annotated compounds of known MoA 10
2.2 Cell painting reagents and filter wavelengths for imaging. 11

LIST OF ACRONYMS

2D	Two-dimensional
3D	Three-dimensional
ABL	Abelson murine leukemia viral oncogene homologue
BCR	Breakpoint cluster region
ANN	Artificial neural network
BSA	Bovine serum albumin
CCLE	Cancer cell line encyclopedia
CCM	Cerebral cavernous malformation
CNN	Convolutional neural networks
DMEM	Dulbecco's modified eagle medium
DMSO	Dimethyl sulfoxide
EMA	European Medicines Agency
FDA	U.S Food and Drug Administration
GPU	Graphics processing unit
HTS	High throughput screening
MCL	Markov clustering algorithm
MOA	Mechanism of action
mRMR	Minimum-redundancy-maximum-relevancy
PBS	Phosphate buffered saline
PCA	Principal component analysis
PDD	Phenotypic drug discovery
RGB	red green blue

SAR Structure activity relationship

TCCS Theta comparative cell scoring

1 INTRODUCTION

1.1 Eroom's Law: The increasing cost of drug discovery

Throughout the last 70 years the cost of developing a new drug has steadily increased. A study by Scannel *et al.* noted the cost to develop a new drug has approximately doubled every 9 years,[1] this observation has been dubbed "Eroom's law", a homage to Moore's law – the well-known observation that the number of transistors in microprocessors approximately doubles every 2 years. The cost of bringing a new drug to market is now approaching £1 billion, taking 10 years from initial concept to approval, the reasons behind this every-increasing cost are still very much in debate, though most agree the issue is multi-faceted. One explanation may be that the low-hanging fruit has been taken, effective traditional remedies have been studied and their active ingredients commercialised, natural products screened, leaving us to tackle the more complex diseases and pharmacological targets. This pessimism has led to the ever present idea that drug discovery is undergoing a productivity crisis, and that the investments made in early stage research do not translate into actionable pharmacology which can be used to develop efficacious therapies and ultimately benefit patients. This belief has led to a renewed interest in alternative drug discovery paradigms.

1.2 The drug discovery process

1.2.1 Target-based screening

Over the past 30 years the majority of drug discovery programmes have seized upon technological advances in robotics and automation to screen ever expansive compound libraries against predefined protein targets. It would be difficult to argue that this target-based high-throughput screening (HTS) approach has not been fruitful, yielding many succesful therapeutics across a range of disease areas, largely attributed to an increased understanding of the genomic basis of many diseases. However, despite numerous clinical and commerical success stories, HTS is not a panacea, with a high attrition rate of lead compounds once they enter clincial trials.[**citation_needed**] A large majority of these clinical trial failures are not due to toxicity, but rather a lack of efficacy which can often be traced back to a poorly hypothesised target in the face of complex disease aetiology.

1.2.2 Phenotypic screening

Phenotypic screening differs from target-based screening in that it does not rely on prior knowledge of a specific target, but instead interrogates a biologically relevant assay to identify compounds which alter the phenotype in a biologically desirable way. This target-agnostic approach can prove useful in diseases with poorly understood mechanisms, or those with no obvious druggable protein targets. Phenotypic screening is not a new approach in small molecule drug discovery, it was the primary method for many decades before the genomics revolution made target hypothesis tractable.[2]

An example of a modern day target-agnostic phenotypic screen is the development of the anti-hepatitis C drug daclatasvir. A model was developed in which human cells were modified to express the hepatitis C virus (HCV) replicon, this was then screened to identify compounds which reduced HCV replication. The target for daclatasvir was later found to be the HCV NS5A protein, which is a novel target and likely not have been suggested in a hypothesis-driven screen as at the time it was thought to have no involvement with HVC.[3,4] This elegant assay has the very useful attribute of modelling the disease whilst simulatenously identifying compounds which may prove to be toxic in human cells.

Many concerns related to phenotypic screening are centered on the lack of mechanistic information for a given lead compound. Whilst the lack of a known target may cause concerns within a commercial drug discovery programme, regulatory bodies such as the Food and Drug Administration (FDA) and European Medicines Agency (EMA) do not require a known target for drug approval – only that it is safe and efficacious. Metformin, a first-line therapy for type 2 diabetes, and is on the World Health Organisation's list of essential medicine, decreases liver glucose production and has a insulin senstitising effect on many tissues. Despite approval since 1957 and widespread clinical use, the molecular mechanism of metformin remained unknown for 43 years.[5] Although knowledge of the molecular target is not necessary to get a drug into the clinic, target deconvolution is still an important part of phenotypic drug discovery programmes – without knowing the protein or proteins a compound is binding to, lead optimisation via structure activity relationship (SAR) studies becomes extremely difficult. In addition, knowledge of the molecular target of a lead compound generated by a phenotypic screen can be used as a basis for starting a more high-throughput hypothesis-driven screen on a novel target, this is why many view phenotypic screening as a complimentary method to target based screening rather than a competing approach or proposed replacement.

1.3 High content imaging

High content imaging is a technique utilising high-throughput microscopes and automated image analysis, commonly used in phenotypic screening as a method for gathering multivariate datasets from images of biological specimens and has proven useful in a wide variety of phenotypic assays, ranging from 2D mammalian cells[cite_HCA_cell_papers], *in vivo* studies in zebrafish[6] and even plants and crops.[7]

High content screens – screening studies carried out with high content imaging – are particularly useful in phenotypic drug discovery for a few reasons. The first is that they allow the use of more complicated assays, which might better represent the biological complexity than simple reductionist models. However, these complex assays often have readouts which are more difficult to quantify, which a single univariate readout may fail to accurately recapitulate, therefore the multivariate datasets produced by high content screening enables a more detailed representation of a complex assay endpoint. A second benefit is that the multivariate nature of the data generated by high content screening may offer a more unbiased method for detecting hits in a phenotypic assay, as any single variable chosen beforehand may miss biologically interesting phenotypes which were not predicted.

With the advent of more complex datasets generated from high-content imaging, the process of image-analysis and computational methods for data processing has given rise to the term "high-content analysis".

1.3.1 Image analysis

After images are generated in a high-content screen, image analysis is the process in which raw image data is transformed into measurements which can be used the describe the observed morphology of the biological specimen exposed to a perturbagen. Here I will focus on cell-based assays for small-molecule screening, though the same methods apply for most other assays (spheroids/organoids etc) and perturbagens (siRNA, CRISPR etc).

The standard approach to extracting numerical features from cell morphologies is through segmentating cells and sub-cellular structures into "objects", and then computing image-based measurements on those objects. Typically each cell within image is identified by first segmenting nuclei from the background. A number of well-established image thresholding algorithms can be used for segmentating nuclei from background, most automatically calculate an intensity threshold to binarise an image based on histograms of pixel pixel intensities.[8,9] The segmented nuclei can then be used as seeds to detect cell boundaries, either through edge detection in a channel containing a cytoplasmic marker, or more crudely by expanding a number of pixels from the nuclei center to approximate cell size. There are also less commonly used methods which utilise machine learning to segment cells,[10] or forgo segmentation entirely.[11,12]

After cells and sub-subcelluar objects have been segmented morphological characteristics are measured for each object, these measurements can cover a wide variety of morphologies depending on the aims of the assay, although can be grouped into 4 main classes:

- **Shape.** Calculated on the properties of the object masks, e.g. area, perimeter, eccentricity. Shape features are commonly used as they are interpretable, robust, and quick to calculate.

- **Intensity.** These features are based on the pixel intensity values within the object boundaries. They can be calculated for multiple channels and include measurements such as average intensity, integrated intensity, and radial distribution of intensity values. Great care has to be taken when using intensity values as they are susceptible to batch effects and microscope artefacts such as vignetting.[13]

- **Texture.** Measures of patterns of intensities within objects, typically measured using grey level co-occurance matrices.[14] This can be used to quantify morphologies such as small speckles or stripes from labelled actin filaments. Texture measurements are often computationally

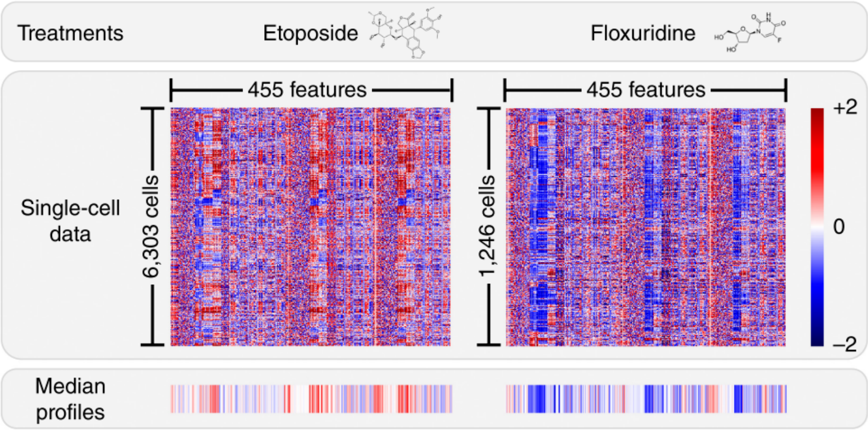

Figure 1.1: Single cell data aggregation to a median profile. Two matrices representing single cell morphology data for a treatment, with columns displaying multiple measured morphological features for each cell represented as a row. *(Figure re-used from Caicedo et al. Nat Methods, 2017)*

expensive and difficult to interpret, although can be used to quantify subtle morphological changes.

- **Spatial context.** These are typically relationships between objects, such as the number of neighbouring cells or nuclei, percentage of a cell boundary in contact with neighbouring cells. This class can also include the simple measure of cell or nuclei count within a field of view.

1.3.2 Data analysis

Measuring morphological features produces an $m \times n$ dataset per object class, where m is the number of objects and n is the number of morpholgical features measure for that object. Commonly single object level data is aggregated to population level, where the population can be a field of view, microtitre-well, or treatment level (see figure 1.1); with the most popular aggregation method being a simple median average.[15] Once the object-level data has been aggregated to a common population level such as per well data, the features from each object class can be combined into a dataset represented by a single $p \times q$ matrix, where p is the number of wells (or other level of aggregation), and q is the total number of combined features from all object classes. It is then useful to view each row of this matrix as a feature vector, or morphological profile which summarises the morphology induced by a treatment.

There are a number of fairly standard data pre-processing steps involved in high content analysis, consisting of: quality-control checks and outlier removal, batch correction, normalisation, standardising feature values, and dimensional reduction or feature selection.[15]

Quality control. Errors are usually introduced at the imaging or segmentation phase of high-content assays, either through poor image quality caused by out-of-focus wells or debris, or otherwise acceptable images causing segmentation artefacts and subsequent outlier morphological features. As assays often generate thousands if not millions of images it is not practical to manually check each image and segmentation mask, therefore a number of methods have been developed to flag potential image artefacts and extreme feature values.

Image artefacts can be detected through measures of image intensity, as out-of-focus images tend to have shallow intensity gradients across the image and lose high-frequency intensity changes.[16] Extreme feature values caused by segmentation errors fall into the well-researched field of outlier detection, with readily available methods such as Hampel filtering[17] or local outlier factor.[18]

Batch correction. Batch effects are accumulations of multiple sources of technical variation such as equipment, liquid-handling error, reagents and environmental conditions which can influence measurements and mislead researchers, and are particularly prevalent in high-throughput experiments. They are normally identified by identifying visually through boxplots of features, with plates or weeks on the x-axis, or through comparing correlations, within plates, between plates of the same batch and across batches. If batch effects need to corrected the simplist method is to standardise each batch separately, other methods include 2-way ANOVA[19] or canonical correlation analysis.[20]

Standardisation. When many morphological features are measured from an image, they are unlikely to be in the same scale or have similar variation – e.g. cell-area measured in pixels which may range from zero to several thousand and cell-eccentricity which is constrained between zero and one. It is therefore common to standardize all feature values to be mean centered and have comparable variance.

Dimensional reduction and feature selection. As with any high-dimensional data a large number of features can cause issues with analysis and interpretation, this is commonly known as the "curse of dimensionality". Another issue is that many of the measured features may not contribute information, either as they have little or no variation between samples, or are redundant due to high correlation with existing features. Dimensional reduction and feature selection methods are both commonly used in other biological fields such as genomics and proteomics, and are now routinely used in high-content imaging analysis. The most commonly used technique in principal component analysis (PCA), which is an unsupervised approach to maximise variation through a linear combination of orthogonal features. PCA can be used to reduce the number of features by selecting a subset of principal components which explain a specified proportion of variance in the data. Loss of interpretability of one of the issues caused by using PCA, and why some researchers favour feature selection methods which aim to retain feature labels, while still reducing dimensionality by removing uninformative features. Many of the feature selection methods are supervised, which may not fit in with unbiased analyses, although Peng *et al.* developed an unsupervised minimum-redudancy-maximum-relevancy (mRMR) feature selection method which has found use in high-content analyses.[21]

Following data pre-processing, downstream analysis is typically focussed on one of two tasks: identifying hit compounds in a screen, or comparing the similarity of morphology profiles created by treatments – both of which use distance as a metric, either comparing hits against a negative control, or treatments against one another respectively.

1.3.3 Image based screening

Phenotypic and image based screens can be used in traditional drug discovery roles whereby a compound library is screened a biologically relevant cell-based assay in order to identify compounds which produce a favourable phenotype. These assays rely on either a positive control compound which is known to elicit the phenotype of interest, or a carefully designed assay in which a disease model is used to identify compounds which push the disease associated phenotype towards a healthy one. An example of this is demonstrated by Gibson *et al.*,[22] whereby they modeled cerebral cavernous malformation (CCM) using siRNA knockdown of the *CCM2* gene in human primary cells, and screened small molecules to identify candidates which rescued the siRNA induced phenotype using fluorescent markers of the nucleus, actin filaments, and VE-cadherin cell-cell junctions. Candidate compounds were then validated in an *in vivo* mouse model, which lead to the ongoing pre-clinical development of 4-Hydroxy-TEMPO as a novel therapeutic for CCM.

1.3.4 Image based profiling

In contrast to screening studies which are mainly interested in looking for a defined phenotype, profiling studies are used to create phenotypic "fingerprints" of perturbagens analogous to transcriptional profiles, which can be used for clustering, inference and prediction. One of the main uses of phenotypic profiling to compare the similarity of morphological profiles, allowing clustering and machine learning methods to build rules in order to classify new or blinded treatments according to similar annotated neigbours.

One of the landmark papers of high-content imaging and profiling was published in 2004 when Perlman *et al.*[23] first demonstrated that morphological profiles between drugs could be clustered according to compound mechanism-of-action using a custom similarity metric and hierarcical clustering. Most studies utilising morphological profiling use unsupervised hierarchical clustering in order to group treatments into bins which produce similar cellular phenotypes,[24,25] though other clustering algorithms such as graph-based Markov clustering algorithm (MCL),[26,27] and spanning trees[28] are commonly used.

An important component of profiling studies is the metric to compare similarity or distance between treatment profiles. The choice of the distance or similarity metric is largely dependent on the dimensionality of the data and the intended downstream applications, although the distiction between distance and similarity is not important, as many similarity metrics and be converted to distances with $1 -$ similarity. A comparative study of similarity measures in high-content fingerprints found that rank-based nonlinear Spearman and Kendall correlations performed the best when comparing full-length [i] fingerprints.[29]

---

[i] Full-length meaning complete morphological profiles that have not been subjected to dimensionality reduction or feature selection methods.

1.4 Phenotypic screening in cancer drug discovery

Cancer drug discovery programmes of past decades seized upon uncontrolled proliferation as a clinically relevant phenotype to screen against, giving rise to a number of anti-proliferative and cytotoxic compounds, often renowned for their severe side-effects. Many modern day oncology drug discovery programmes still retain anti-proliferation as a key predictor for pre-clinical success, although our ever increasing understanding of cancer's molecular underpinnings has driven many oncology programmes towards a more target-directed approach. The prototypical success story of target-driven drug discovery in oncology is imatinib, a tyrosine kinase inhibitor targetting the BCR-ABL fusion protein in chronic myeloid leukemia. However, despite imatinib's exceptional success, it is unfortunately very much an exception – in most cases targeting a single driver in a complex signalling network results in compensatory signalling, activation of redundant pathways and unpredicted feedback mechanisms, all of which diminish efficacy *in vivo*.

In a review of 48 small molecule drugs approved for use in oncology between 1999 and 2013, 31/48 were discovered through target based screens, whereas 17/48 were based on leads from target-agnostic phenotypic screens,[30] of those compounds discovered through target directed screening programmes the vast majority (75%) were kinase inhibitors. However phenotypically derived compounds did not live up to the hypothesis that target-agnostic screening should be more likely to identify compounds with novel MoAs,[31] with only 5/17 being first in class molecules. An explanation for this sparsity of novel mechanisms is that phenotypic assays which use cytotoxicity readouts are likely to find low-hanging fruit such as targetting microtubule stabilisation and DNA replication dynamics.[30] One option to combat this narrow attention on a select few targets – caused by either hypothesis-driven or simplistic phenotypic screens – is to utilise the more detailed mechanistic information offered by high-content imaging rather than relying on cellular death as catch-all phenotypic readout.

In addition to 2D high-content imaging screens, more complex phenotypic models such as 3D tumour spheroids are becoming increasingly prevelent in preclinical oncology, much of this stems from the hypothesis that they better recapitulate the microenviroment found in tumours.

1.4.1 Cancer cell line panels

Panels of multiple cancer cell lines such as the NCI-60 and Cancer Cell Line Encyclopedia (CCLE) have been widely used; initially to facilitate high-throughput screening and increase certainty in hit selection/disease-specificity[32,33], and later on as a research tool to study pharmacogenomics.[34,35,36] The use of cancer cell line panels can also benefit phenotypic screens by mirroring the heterogeneity found in patient populations, as well as heterogenous cell populations found in tumours[37]. Throughout this body of work I have used a panel of eight breast cancer cell lines (table 1.1), these cell lines were chosen based on a number of criteria: *(1)* Relatively fast growth to allow compound screening to be performed in weekly batches. *(2)* Adherent to tissue culture plastic to enable 2D imaging. *(3)* Form a monolayer when grown in 2D – overlapping cells cause difficulties for most image segmentation methods. *(4)* Amenable for imaging – larger and/or flatter cells allow for better discrimination of sub-cellular features. *(5)* Distinct morphologies to evaluate the robustness of

Cell line	Molecular subclass	Mutational status PTEN	PI3K
MCF7	ER	WT	E545K
T47D	ER	WT	H1047R
MDA-MB-231	TN	WT	WT
MDA-MB-157	TN	WT	WT
HCC1569	HER2	WT	WT
SKBR3	HER2	WT	WT
HCC1954	HER2	⋆	H1047R
KPL4	HER2	⋆	H1047R

Table 1.1: Panel of breast cancer cell lines chosen for study. PI3K:Phosphoinsitide-3-kinase, PTEN:Phosphatase and tensin homolog, ER:Estrogen receptor, TN:triple-negative, HER2:human epidermal growth factor, WT:wild-type, ⋆:lack of consensus regarding the mutational status.

morphological profiling methods. *(6)* A collection which represents a range of molecular subclasses of breast cancer.

1.5 Thesis structure

The following chapters focus on selected topics from 4 years of work, some of which has been published (see appendix). Chapter 2 is an analysis of machine learning methods to classify compound MoA from high content imaging data, with a focus on how well classifiers transfer across to new data from morphologically distinct cell lines. Chapter 3 is the development and application of a novel analytical method to detect and quantify differential phenotypic responses between morphologically distinct cell lines when treated with small molecules. Chapter 4 is a high content screen of 13,000 small molecules in order to identify compound that produced distinct phenotypic responses between cell lines, functional assays to validate hits and proteomics to investigate potential pathways responsible. Chapter 5 is a work towards combining cheminformatics with high content morphological data in order to infer MoA of unannotated compounds.

2

CELL MORPHOLOGY CAN BE USED TO PREDICT COMPOUND MECHANISM-OF-ACTION

2.1 Introduction

Cellular morphology is influenced by multiple intrinsic and extrinsic factors acting on a cell, and striking changes in morphology are observed when cells are exposed to biologically active small molecules. This compound-induced alteration in morphology is a manifestation of various perturbed cellular processes, and we can hypothesise that compounds with similar MoA which act upon the same signalling pathways will produce comparable phenotypes, and that cell morphology can, in turn, be used to predict compound MoA.

In 2010 Caie *et al.* generated, as part of a larger study, an image dataset consisting of MCF7 breast cancer cells treated with 113 small molecules grouped into 12 mechanistic classes, these cells were then fixed, labelled and imaged in three fluorescent channels[37]. This dataset (also known as BBBC021) has become widely used as a benchmark in the field for MoA classification tasks, with multiple publications using the images to compare machine learning and data pre-processing approaches[38,39,40,41]. Whilst this is important work, it has led to the situation whereby the vast majority of studies in this field have based their work on a single dataset generated with a one cell-line.

One of the issues associated with phenotypic screening when used in a drug discovery setting is target deconvolution. Once a compound has been identified which results in a desirable phenotype in a disease-relevant assay it is common to want to know which molecular pathways the hit compound is acting upon. While target deconvolution is a complex and difficult task, image-based morphological profiling represents one option similar to transcriptional profiling that can match and unknown compound to the nearest similar annotated compound in a dataset, while at the same time being far cheaper than the transcriptional methods such as LINCS1000[42].

2.2 Creating an annotated dataset

As part of a preliminary study, a dataset similar to that of Caie *et al.*'s was generated consisting of 24 compounds grouped into 8 mechanistic classes screened across the panel of 8 breast cancer cell-lines (see table 1.1)

Compound	MoA class	Supplier	Catalog no.
Paclitaxel	Microtubule disrupting	Sigma	T7402
Epothilone B	Microtubule disrupting	Selleckchem	S1364
Colchicine	Microtubule disrupting	Sigma	C9754
Nocodazole	Microtubule disrupting	Sigma	M1404
Monastrol	Microtubule disrupting	Sigma	M1404
ARQ621	Microtubule disrupting	Selleckchem	S7355
Barasertib	Aurora B inhibitor	Selleckchem	S1147
ZM447439	Aurora B inhibitor	Selleckchem	S1103
Cytochalasin D	Actin disrupting	Sigma	C8273
Cytochalasin B	Actin disrupting	Sigma	C6762
Jaskplakinolide	Actin disrupting	Tocris	2792
Latrunculin B	Actin disrupting	Sigma	L5288
MG132	Protein degradation	Selleckchem	S2619
Lactacystin	Protein degradation	Tocris	2267
ALLN	Protein degradation	Sigma	A6165
ALLM	Protein degradation	Sigma	A6060
Emetine	Protein synthesis	Sigma	E2375
Cycloheximide	Protein synthesis	Sigma	1810
Dasatinib	Kinase inhibitor	Selleckchem	S1021
Saracatinib	Kinase inhibitor	Selleckchem	S1006
Lovastatin	Statin	Sigma	PHR1285
Simvastatin	Statin	Sigma	PHR1438
Camptothecin	DNA damaging agent	Selleckchem	S1288
SN38	DNA damaging agent	Selleckchem	S4908

Table 2.1: Annotated compounds and their associated mechanism-of-action label used in the classification tasks.

2.2.1 Compounds

The 24 compounds see table 2.1 were chosen based on previous knowledge of their biological activity and wide range of morphological responses, most of the compounds feature in Caie *et al.*s original dataset.

2.2.2 Cell painting: labelling cellular morphology

In order to capture a broad view of morphological changes within a cell using fluorescent microscopy, a choice has to be made which cellular structures to label. This choice is limited by the availablity of the fluorescent filter sets fitted to the microscope, reagent costs, and the scalability of the protocol when used in a large screen. Fortunately, this problem was already addressed by another group who published a protocol – named “cell painting” – for labelling 7 cellular structures, using 6 non-antibody stains imaged in the same 5 fluorescent channels available with our miscoscopy setup.[24,43]

The cell-painting protocol was initially optimised by Gustafsdottir *et al.* for use in the U2OS oesteosarcoma cell line, and briefly tested in a few other commonly used cell-lines. However, when tested on the panel of 8 breast cancer cell lines, the staining protocol was observed to induce morphological changes on certain cell lines, in the absence of compounds. It was found that changing the media, and adding the MitoTracker DeepRed stain to live MDA-MB-231 cells produced a rounded morphology, which was not observed in the other cell lines. As any morpholgical changes

Stain	Labeled Structure	Wavelength (ex/em [nm])	Concentration	Catalog no.; Supplier
Hoechst 33342	Nuclei	387/447 ±20	2 μg/mL	#H1399; Mol. Probes
SYTO14	Nucleoli	531/593 ±20	3 μM	#S7576; Invitrogen
Phalloidin 594	F-actin	562/624 ±20	0.85 U/mL	#A12381; Invitrogen
Wheat germ agglutinin 594	Golgi and plasma membrane	562/624 ±20	8 μg/mL	#W11262; Invitrogen
Concanavalin A 488	Endoplasmic reticulum	462/520 ±20	11 μ/mL	#C11252; Invitrogen
MitoTracker DeepRed	Mitochondria	628/692 ±20	0.6 μM	#M22426; Invitrogen

Table 2.2: Reagents used in the cell painting protocol and the excitation/emission wavelengths of the filters used in imaging. ex: excitation, em: emission

introduced by the staining protocol would mask those caused by small-molecules, the protocol was adapted by removing the media change step, and moving the addition of wheat germ agglutinin and MitoTracker DeepRed until after fixation. As the cells were now fixed immediately in their existing media this prevented any alterations to the morphology and improved the wheat germ agglutinin staining, although as the MitoTracker stain relies on membrane potential in mitochondria, the selectivity of the MitoTracker stain was reduced when used on fixed cells, though it still produced selective enough labelling to capture large changes in mitochondrial morphology.

2.3 Machine learning methods to classify compound MoA

Predicting compound MoA from phenotypic data is a classification task. This type of machine learning problem is well researched, and there are several models appropriate for our labelled data. As the raw data is in the form of images, it can approached as an image classification task, a problem in the field receiving lots of attention due to recent theoretical and technological breakthroughs. Whereas a more classical approach would be to extract morphological information from the images, generating a multivariate dataset from the images, and training a classifier on these morphological features.

To develop and validate a machine learning model the dataset has to be split into training, validation and test sets. This is because overfitting is a common problem in machine learning, whereby the model is trained and accurately predicts labels on one dataset, but performs poorly when applied to new data on which is was not trained. Most classifiction models will overfit to some degree, typically performing better on the training dataset than any other subsequent examples, but the challenge is to limit this overfitting, and also to ensure that the data used to report accuracy measures has not been used in any way to train or validate the model.

2.3.1 Comparison between classical and deep-learning machine learning methods

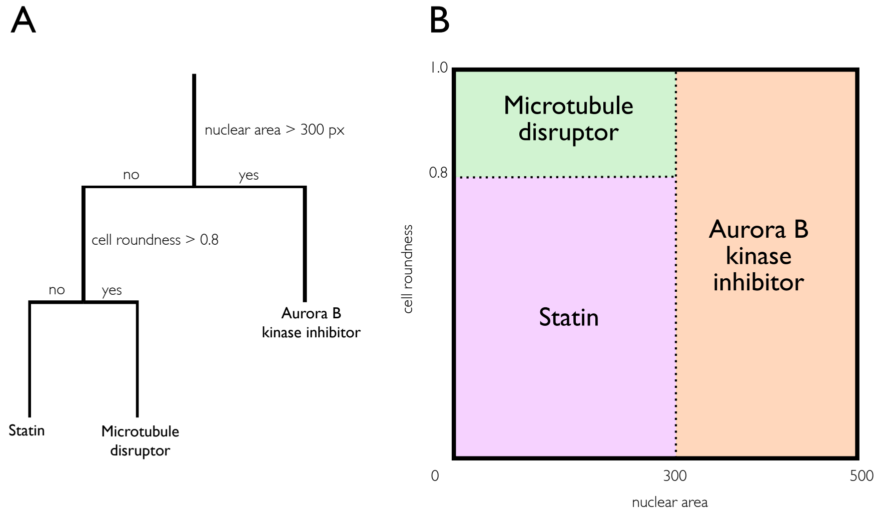

Figure 2.1: **(A)** An example of a simple mock decision tree to classify compound mechanism of action based on morphological features. **(B)** Depiction of decision space as divided by the decision tree model. Shaded areas show how new input data will be classified based on the decision rules (dotted lines).

Ensemble of decision trees trained on extracted morphological features

A decision tree is a very simple method that can be used for both regression and classification. The method works by repeatedly dividing the decision space using binary rules on the feature values until a terminal node containing a classification label is reached (figure 2.1). Simple decision trees like those shown in figure 2.1 perform relatively poorly on all but the simplest of classification problems. However, by aggregating many decision trees and their predictions we can create more accurate and robust models in a practice known as ensemble learning.[44] Bagging[45] and Boosting[46] are two popular methods for constructing ensembles of decision trees. As combining the output of several decision trees is useful only if there is a disagreement among them, these two methods both attempt to solve the same problem of generating a set of correct decision tress, that still disagree with one another as much as possible on incorrect predictions.

Decision tree methods work best with multivariate tabular data, with well defined features describing each observation, this is in contrast to image data which consists of 2D arrays of pixel intensities. Therefore, in order to train such a model, cellular morphology needs to be quantified by measuring cellular features. This is a common task with multiple software packages available, which follow two main steps: (1) Segment objects from the background. Objects may be sub-cellular structures or whole-cell masks (2) Measure various attributes from the object, this is typically based on size, shape and intensity. Cellprofiler[47] was chosen primarily due to the high configurability and the permissive license enabling large-scale distributed processing on compute clusters in order to reduce the image analysis time. The images captured on the ImageXpress were analysed using Cellprofiler, quantifying approximately 400 morphological features. The datasets produced by the Cellprofiler analysis contained morphological measurements on an individual cell level. Although we can train a model on single cell data we are not interesting in classifying morphologies of single cells, but rather classifying an image or a collection of images that represent a compound treatment, this therefore allows several approaches to structuring the training data:

1. Train and test on median profiles.
2. Train on single cell data, test on image or well median profiles.
3. Train on single cell data, test on single cell data and classify the parent image as the most commonly predicted class of cell in that image.
4. Train on median profiles of boostrapped single cell samples within an image, and test on median profiles.

Convolutional neural networks trained on pixel data

Artificial neural networks (ANNs) are becoming increasingly common in a wide range of machine learning tasks. Although many of the theories underpinning ANNs are decades old,[48] they have only recently achieved widespread practical use due to improved methods for training[49] and the availability of more computing power allowing the use of more complex models. ANNs are (very) loosely inspired by the structure of biological brains, with interconnected neurons passing signals through layers onto subsequent neurons forming a chain with the output of one neuron becoming the input for the next neuron. In between neurons, the signals can be altered by multiplying the value by a weight (W), it is through adjusting these individual weights that ANNs optimise their performance for a particular task, similar to how long-term potentiation is used to strengthen synaptic connections in biological brains. When a signal reaches a neuron, it is combined via a weighted sum with all the other inputs from other connected neurons and passed through an activation function. This activation function – similar to an action potential in neurons – determines the output of the neuron for the given aggregated input, which is then passed as new inputs onto subsequent neurons and so on, however, in contrast to an all-or-nothing output of an action potential there are several types of activation functions used in ANNs, most of which have a graded output (figure 2.2B).

The neurons in an ANN are typically arranged in several layers: an input layer; one or more hidden layers; and a final output layer (figure 2.3). With each layer, the network transforms the data into a new representation, through training the network these representations make the data easier to classify. In the final layer, the data is ultimately represented in a way which makes a single output neuron activate more strongly than the other neurons in that layer, and so the data is ultimately transformed into a single value – the index of the active neuron which corresponds to a particular class. A new ANN is initialised with random weights, to train a neural network these weights are adjusted by feeding in labelled data and adjusting weights in order to minimise classification errors through a process known as backpropagation.[49]

The convolution aspect of convolutional neural networks plays an important role when working with image data. Two-dimensional convolutions are widely used in image processing – blurring, sharpening and edge detection are all common operations which use this operation. They work by mapping a kernel – a smaller matrix of values – across a larger matrix, thereby using information

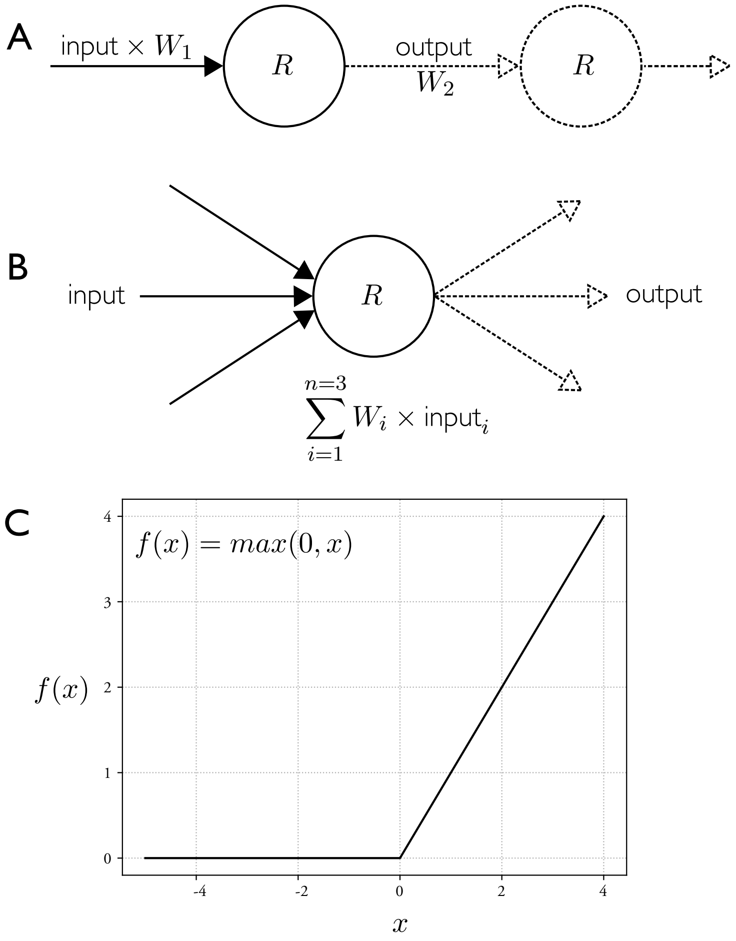

Figure 2.2: **(A)** A representation of a single connected neuron in an ANN, the input value to the neuron is multiplied by the weight (W_1), before being passed through the activation function R, the output of which is then multiplied by W_2, and passed as the input to the next neuron. **(B)** A neuron with multiple inputs and outputs, typical of those in a hidden layer. The activation function acts on the weighted sum of all inputs, and returns a single output value which is when directed to all connected neurons in the next layer. Where W_i is the weight of input_i. **(C)** A common activation function also known as a rectifier, in this example a rectified linear unit (ReLU), in the inputs (x) is transformed and passed as output. So $f(x)$ can be viewed as the output for a given value of x.

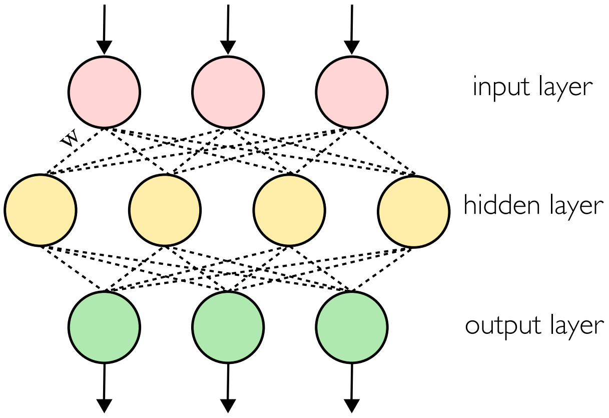

Figure 2.3 Representation of a simple 3-layer ANN with a single fully connected hidden layer, three input neurons and three output neurons. W denotes a weighted connection between an input neuron and a hidden-layer neuron, with all connections between neurons having an associated adjustable weight. A network such as this would take a vector of three numbers as input, and would be capable of predicting three classes from the output layer of three neurons depending on the activation strengths of the neurons in the final output layer.

from a small region of pixels in their transformation of each individual pixel. This lends itself well to ANNs, as a pixel value in isolation is less informative than a pixel value in the context of the neighbouring values. Depending on the size and the values within the kernel, the transformations highlight different features within an image. Two dimensional convolutions are used in ANNs by starting with many randomly initialised kernels, and updating the kernel values through training in order to best highlight features which prove useful for accurately predicting classes. Using a single convolutional layer highlights simple features in an image such as edges and speckles, by combining several convolutional layers more complex features are highlighted through combinations of these simple features. These convolved images are then flattened into a one-dimensional vector which is used as an input in a fully connected ANN such as that depicted in figure 2.3.

As CNNs can be constructed with a wide variety of architectures, and the field is still rapidly developing, I remained closed to well established architectures in the literature. However, as most images are digitially represented in three colour channels (red, green, blue (RGB)), the vast majority of CNN models are constructed in a way that input is restricted to three colour channels, therefore it is necessary to adapt these architectures to work with the differently shaped inputs and additional parameters generated by the 5 channel images generated with the ImageXpress.

The images generated by the ImageXpress microscope with zero binning are 2160 by 2160 pixel tiff files, with a bit-depth of 16, whilst these image properties are common in microscopy, they are extreme for current CNN implementations. Most image classification tasks involving CNN's use 8-bit images in the region of 300 by 300 pixels, relatively small images are used as the convolutional layers of deep CNN's generate many thousands of matrices, and using smaller input images drastically reduces the computing resources and time required to train such classifiers.

This presents the issue of how to reduce the 2160×2160 images into small images, one option is to downscale the entire image using bi-linear or bi-cubic interpolation, while a second option is to chop the original image up into smaller sub-images (figure 2.4). Downsizing the original image by simple scaling has a few potential problems which make it unsuitable for this particular task: many of the finer-grained cell morphologies such as mitochondria and endoplasmic reticulum distribution will be lost due to the reduction in image resolution; in addition, it was found that whole well images are susceptible to over-fitting as the classfier learned biologically irrelevant features such as the locations of cells within an image, which although should be random might have some spurious association with particular class labels. When chopping images into sub-images the most simple and commonly used method is to chop each image into an evenly spaced grid, whilst this is unbiased and easy to implement, it has the downside of potentially returning many images that do not contain any cells. A more nuanced approach is to first detect the x,y co-ordinates of each in the image, and creating a 300×300 bounding-box around the centre of each cell. This method returns an image per cell, negating the issue of empty images; it does however require detecting cell locations and handling cells located next to the image border.

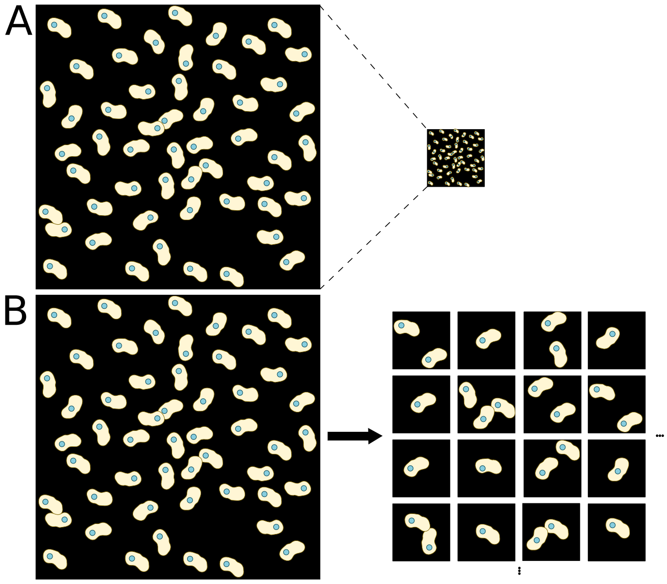

Figure 2.4: Two options for adapting large microscope images to work with the smaller input size of typical CNNs. **(A)** Full-sized images are downsized to the desired dimensions via bilinear or bicubic interpolation. **(B)** Images are chopped into smaller sub-images, cell detection can be carried out beforehand to ensure images contain at least one cell.

2.4 Predicting compound MoA on single cell lines

The first step is to determine a baseline of how well the predictive models perform when trained and tested on the same cell line, this also acts as a platform with which to test and optimise model architecture and hyperparameters.

The two different CNN architectures were tested based on the hypothesis that a deeper, more complex architecture (ResNet18[50]) will be capable of learning more subtle features, although more complex models with greater numbers of interal parameters are more prone to overfitting when training data is limited. On the other hand, a more simple model such as AlexNet[51] which contains fewer convolutional layers will be less able to perform complex tranformations of the data, and therefore theoretically limit the subtle features which can be extracted and learned from an image. While this might theoretically reduce accuracy, in the absence of large amount of training data it may reduce overfitting due to the fewer number of parameters.

2.4.1 Ensemble of decision trees

2.4.2 CNNs

2.5 Transferring machine learning models to morphologically distinct cell lines

TODO: How well machine learning models generalise across cell-lines.

2.5.1 Leave-one-out classification

TODO.

2.5.2 Do additional cell-lines during training increase prediction accuracy?

TODO: When predicting MoA for a given cell-line, does adding additional training examples from other morphologically distinct cell-lines improve classification accuracy?

2.6 Discussion

TODO.

2.7 Methods

2.7.1 Cell culture

The cell-lines were all grown in DMEM (#CATNO MANUFACTURER) supplemented with 10% foetal bovine serum and 2 mM L-glutamine, incubated at 37°C, 5% CO_2. 2,500 cells were seeded into the inner 60 wells of optical bottomed 96-well plates (#165305 ThermoFisher) in 100 μL of media, whilst outer wells were filled with 100 μL of PBS.

After seeding with cells, assay plates were incubated for 24 hours prior to the addition of compounds.

2.7.2 Compound handling

Compounds (table 2.1) were diluted in DMSO at a stock concentration of 10 mM. Compounds plates were made in v-bottomed 96-well plates (#CATNO MANUFACTURER), at 1000-fold concentration in 100% DMSO by serial dilutions ranging from 10 mM to 0.3 μM in semi-log concentrations. Compounds were added to assay plates containing cells after 24 hours of incubation by first making a 1:50 dilution in media to create an intermediate plate, followed by a 1:20 dilution from intermediate plate to the assay plate, with an overall dilution of 1:1000 from the stock compound plate to the assay plate.

2.7.3 Staining

After 48 hours in the presence of compounds, assay plates were fixed by adding equal volume (100 μL) of 8% paraformaldehyde (#CATNO MANUFACTURER) to the existing media, resulting in a final paraformaldehyde concentration of 4%, and left to incubate at room temperature for 30 minutes. Wells were then washed with 100 μL of PBS and permeabilised with a 0.1% Triton-X100 solution for 20 minutes at room temperature. A cell-staining solution was made up in 1% bovine serum albumin (BSA) solution (see table 2.2). Wells were then washed again with 100 μL of PBS followed by addition of 30 μL staining solution. Plates were then incubated in the dark for 30 minutes at room temperature, washed 3 times with 100 μL of PBS. Before the final aspiration, plates were then sealed (#CATNO MANUFACTURER) and imaged.

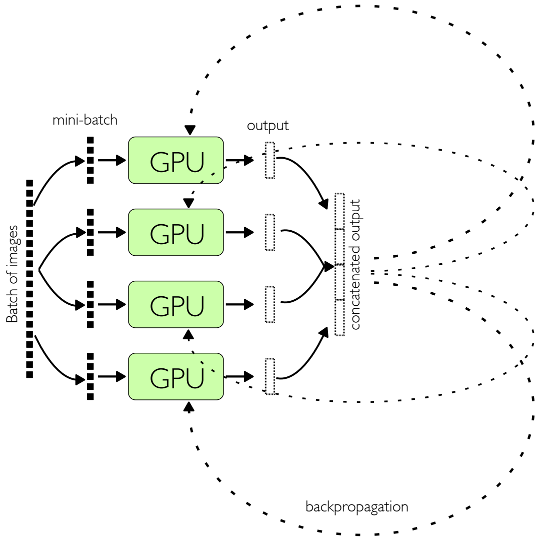

Figure 2.5: Increased training speed by data parallelism. Models are replicated across an array of GPUs, the input batch is split evenly among the devices, with each device processing a portion in parallel. During backpropagation the updated weights for all replicas are averaged and models weights are updated synchronously.

2.7.4 Imaging

Imaging was carried out on a multi-wavelength wide-field fluorescent microscope (ImageXpress micro XL, MolecularDevices, USA) with a robotic plate loader (Scara4, PAA, UK). Images were captured in 5 fluorescent channels (see table 2.2) at 20x magnification, exposure times were kept constant between plates and batches, as to not influence intensity values used in subsequent analyses.

2.7.5 Ensemble of decision trees

Models were created using scikit-learn version 0.19 in python 3.6.2.

2.7.6 Image analysis: numeric features from images

TODO

2.7.7 Convolutional neural networks

All code related to neural networks was written in pytorch (version 0.3) for python 3.5, and all ANN models were trained on Nvidia K80 GPUs. As training CNNs is computationally expensive and time consuming, data parallelism was leveraged to share batches of images across multiple GPUs. This technique replicates the CNN model on each device, which processes a portion of the input data, the updated weights for all devices are then averaged and model replicates are updated synchronously after each batch (figure 2.5). This speeds up model training approximately linearly with the number of GPUs and allows use of larger batch sizes.

Image preparation

TODO

3 MEASURING DISTINCT PHENOTYPIC RESPONSE

3.1 Section name

Random text before.

text text text here:

```
# example code highlighting
dist <- function(x) {
    return(sqrt(x %*% x))
}
```

Example text inbetween.

```
# some comment
def new_function(x):
    """docstring"""
    total = 0
    for i in x:
        total += 1
    return total
```

4 LARGE COMPOUND SCREEN ACROSS 8 BREAST CANCER CELL LINES

4.1 Section name

5 CHEMINFORMATICS

5.1 Section name

6 DISCUSSION AND CONCLUSION

6.1 Section name

BIBLIOGRAPHY

[1] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. "Diagnosing the decline in pharmaceutical R&D efficiency". *Nature Reviews Drug Discovery* 11.March (2012), pp. 191–200.

[2] Wei Zheng, Natasha Thorne, and John C. McKew. "Phenotypic screens as a renewed approach for drug discovery". *Drug Discovery Today* 18.21-22 (2013), pp. 1067–1073. arXiv: `15334406`.

[3] Makonen Belema and Nicholas A. Meanwell. "Discovery of daclatasvir, a pan-genotypic hepatitis C virus NS5A replication complex inhibitor with potent clinical effect". *Journal of Medicinal Chemistry* 57.12 (2014), pp. 5057–5071.

[4] James H. Nettles, Richard A. Stanton, Joshua Broyde, Franck Amblard, Hongwang Zhang, Longhu Zhou, Junxing Shi, Tamara R. McBrayer, Tony Whitaker, Steven J. Coats, James J. Kohler, and Raymond F. Schinazi. "Asymmetric binding to NS5A by daclatasvir (BMS-790052) and analogs suggests two novel modes of HCV inhibition". *Journal of Medicinal Chemistry* 57.23 (2014), pp. 10031–10043.

[5] Gerald I. Shulman Ripudaman S. Hundal, Martin Krssak, Sylvie Dufour, Didier Laurent, Vincent Lebon, Visnathan Chandramouli, Silvio E. Inzucchi, William C. Schumann, Kitt F. Petersen, Bernard R. Landau. "Mechanism by which metformin reduces glucose production in type 2 diabetes". *Diabetes* 49.12 (2000), pp. 2063–2069. arXiv: `NIHMS150003`.

[6] C Geoffrey Burns, David J. Milan, Eric J. Grande, Wolfgang Rottbauer, Calum A MacRae, and Mark C. Fishman. "High-throughput assay for small molecules that modulate zebrafish embryonic heart rate". *Nature Chemical Biology* 1.5 (2005), pp. 263–264.

[7] Dijun Chen, Kerstin Neumann, Swetlana Friedel, Benjamin Kilian, Ming Chen, Thomas Altmann, and Christian Klukas. "Dissecting the Phenotypic Components of Crop Plant Growth and Drought Responses Based on High-Throughput Image Analysis". *The Plant Cell Online* 26.12 (2014), pp. 4636–4655.

[8] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". *IEEE Transactions on systems, man and cybernetics* 20.1 (1979), pp. 62–66.

[9] Krishnan Padmanabhan, William F. Eddy, and Justin C. Crowley. "A novel algorithm for optimal image thresholding of biological data". *Journal of Neuroscience Methods* 193.2 (2010), pp. 380–384.

[10] Christoph Sommer, Christoph Straehle, K Ullrich, and Fred a Hamprecht. "ILASTIK : Interactive learning and segmentation toolkit". *Eighth IEEE International Symposium on Biomedical Imaging (ISBI)* 1 (2011), pp. 230–233.

[11] Satwik Rajaram, Benjamin Pavie, Lani F Wu, and Steven J Altschuler. "PhenoRipper: software for rapidly profiling microscopy images." *Nature methods* 9.7 (2012), pp. 635–7.

[12] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D. Mark Eckley, and Ilya G. Goldberg. "WND-CHARM: Multi-purpose image classification using compound image transforms". *Pattern Recognition Letters* 29.11 (2008), pp. 1684–1693. arXiv: `NIHMS150003`.

[13] Daniel B. Goldman. "Vignette and exposure calibration and compensation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.12 (2010), pp. 2276–2288.

[14] Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. "Textural Features for Image Classification". *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (1973), pp. 610–621.

[15] Juan C. Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, Joseph D. Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D. Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A. Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G. Linington, and Anne E. Carpenter. "Data-analysis strategies for image-based cell profiling". *Nature Methods* 14.9 (2017), pp. 849–863.

[16] Mark-Anthony Bray, Adam N. Fraser, Thomas P. Hasaka, and Anne E. Carpenter. "Workflow and Metrics for Image Quality Control in Large-Scale High-Content Screens". *Journal of Biomolecular Screening* 17.2 (2012), pp. 266–274.

[17] Frank R. Hampel. "The influence curve and its role in robust estimation". *Journal of the American Statistical Association* 69.346 (1974), pp. 383–393.

[18] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. "LOF: Identifying Density-Based Local Outliers". *Proceedings of the 2000 Acm Sigmod International Conference on Management of Data* (2000), pp. 1–12.

[19] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. "Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses". *Biostatistics* 17.1 (2016), pp. 29–39.

[20] Saman Vaisipour. "Detecting, correcting, and preventing the batch effects in multi-site data, with a focus on gene expression Microarrays". PhD thesis. University of Alberta, 2014, pp. 1–175.

[21] Hanchuan Peng, Fuhui Long, and Chris Ding. "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy". *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1226–1238. arXiv: `f`.

[22] Christopher C. Gibson, Weiquan Zhu, Chadwick T. Davis, Jay A. Bowman-Kirigin, Aubrey C. Chan, Jing Ling, Ashley E. Walker, Luca Goitre, Simona Delle Monache, Saverio Francesco Retta, Yan Ting E. Shiu, Allie H. Grossmann, Kirk R. Thomas, Anthony J. Donato, Lisa A. Lesniewski, Kevin J. Whitehead, and Dean Y. Li. “Strategy for identifying repurposed drugs for the treatment of cerebral cavernous malformation”. *Circulation* 131.3 (2015), pp. 289–299. arXiv: `15334406`.

[23] ZE Perlman, MD Slack, Y Feng, and TJ Mitchison. “Multidimensional drug profiling by automated microscopy”. *Science* 306 (2004), pp. 1194–1199.

[24] Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman a Carrel, Todd R Golub, Stuart L Schreiber, Paul a Clemons, Anne E Carpenter, and Alykhan F Shamji. “Multiplex cytological profiling assay to measure diverse cellular states.” *PloS one* 8.12 (2013), e80999.

[25] Daniel W Young, Andreas Bender, Jonathan Hoyt, Elizabeth McWhinnie, Gung-Wei Chirn, Charles Y Tao, John a Tallarico, Mark Labow, Jeremy L Jenkins, Timothy J Mitchison, and Yan Feng. “Integrating high-content screening and ligand-target prediction to identify mechanism of action.” *Nature chemical biology* 4.1 (2008), pp. 59–68.

[26] Felix Reisen, Amelie Sauty de Chalon, Martin Pfeifer, Xian Zhang, Daniela Gabriel, and Paul Selzer. “Linking Phenotypes and Modes of Action Through High-Content Screen Fingerprints”. *ASSAY and Drug Development Technologies* 13.7 (2015), p. 150810081821009.

[27] Stijn Van Dongen. “Graph Clustering Via a Discrete Uncoupling Process”. *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 121–141.

[28] Peng Qiu, Erin F. Simonds, Sean C. Bendall, Kenneth D. Gibbs, Robert V. Bruggner, Michael D. Linderman, Karen Sachs, Garry P. Nolan, and Sylvia K. Plevritis. “Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE”. *Nature Biotechnology* 29.10 (2011), pp. 886–893.

[29] Felix Reisen, Xian Zhang, Daniela Gabriel, and Paul Selzer. “Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery.” *Journal of biomolecular screening* 18.10 (2013), pp. 1284–97.

[30] John G Moffat, Joachim Rudolph, and David Bailey. “Phenotypic screening in cancer drug discovery - past, present and future.” *Nature reviews. Drug discovery* 13.8 (2014), pp. 588–602.

[31] David C Swinney and Jason Anthony. “How were new medicines discovered?” *Nature reviews. Drug discovery* 10.7 (2011), pp. 507–19.

[32] Lin Wu, Anne M Smythe, Sherman F Stinson, Leslie a Mullendore, Anne Monks, Dominic a Scudiero, Kenneth D Paull, Antonis D Koutsoukos, Lawrence V Rubinstein, Michael R Boyd, and Robert H Shoemaker. “Multidrug-resistant Phenotype of Disease-oriented Panels of Human Tumor Cell Lines Used for Anticancer Drug Screening”. *Cancer Research* 52 (1992), pp. 3029–3034.

[33] R H Shoemaker. “The NCI60 human tumour cell line anticancer drug screen”. *Nature Rev.* 6.10 (2006), pp. 813–823.

[34] Laura M Heiser, Anguraj Sadanandam, Wen-lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Frances Tong, Nora Bayani, Zhi Hu, Jessica I Billig, Andrea Dueregger, Sophia Lewis, Lakshmi Jakkula, James E Korkola, Steffen Durinck, François Pepin, Yinghui Guan, Elizabeth Purdom, Pierre Neuvial, Henrik Bengtsson, Kenneth W Wood, Peter G Smith, Lyubomir T Vassilev, Bryan T Hennessy, Joel Greshock, Kurtis E Bachman, Mary Ann, John W Park, Laurence J Marton, Denise M Wolf, Eric A Collisson, Richard M Neve, Gordon B Mills, Terence P Speed, Heidi S Feiler, Richard F Wooster, David Haussler, Joshua M Stuart, Joe W Gray, and Paul T Spellman. “Subtype and pathway specific responses to anticancer compounds in breast cancer”. *Proceedings of the National Academy of Sciences* 109.8 (2012), pp. 2724–2729.

[35] Ogan D. Abaan, Eric C. Polley, Sean R. Davis, Yuelin J. Zhu, Sven Bilke, Robert L. Walker, Marbin Pineda, Yevgeniy Gindin, Yuan Jiang, William C. Reinhold, Susan L. Holbeck, Richard M. Simon, James H. Doroshow, Yves Pommier, and Paul S. Meltzer. “The exomes of the NCI-60 panel: A genomic resource for cancer biology and systems pharmacology”. *Cancer Research* 73.14 (2013), pp. 4372–4382. arXiv: `15334406`.

[36] Samira Jaeger, Miquel Duran-Frigola, and Patrick Aloy. “Drug sensitivity in cancer cell lines is not tissue-specific”. *Molecular Cancer* 14.1 (2015), pp. 1–4.

[37] Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. “High-content phenotypic profiling of drug response signatures across distinct cancer cells.” *Molecular cancer therapeutics* 9.6 (2010), pp. 1913–26.

[38] V. Ljosa, P. D. Caie, R. ter Horst, K. L. Sokolnicki, E. L. Jenkins, S. Daya, M. E. Roberts, T. R. Jones, S. Singh, A. Genovesio, P. A. Clemons, N. O. Carragher, and A. E. Carpenter. “Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment”. *Journal of Biomolecular Screening* 18.10 (2013), pp. 1321–1329.

[39] S Singh, M Bray, T R Jones, and A E Carpenter. “Pipeline for illumination correction of images for high-throughput”. *Journal of Microscopy* 256.3 (2014), pp. 231–236.

[40] Nick Pawlowski, Juan C Caicedo, Shantanu Singh, Anne E Carpenter, and Amos Storkey. “Automating Morphological Profiling with Generic Deep Convolutional Networks”. *bioRxiv* (2016).

[41] D Michael Ando, Cory Y Mclean, and Marc Berndl. “Improving Phenotypic Measurements in High-Content Imaging Screens”. *bioRxiv* (2017).

[42] Qiaonan Duan, Corey Flynn, Mario Niepel, Marc Hafner, Jeremy L Muhlich, F Fernandez, Andrew D Rouillard, Christopher M Tan, Edward Y Chen, R Golub, Peter K Sorger, Aravind Subramanian, and Avi Ma. “LINCS Canvas Browser : interactive web app to query , browse

and interrogate LINCS L1000 gene expression signatures”. *Nucleic acids research* 42.W1 (2014), W449–W460.

[43] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. “Cell Painting , a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes”. *Nature Methods* 11.9 (2016), pp. 1757–1774.

[44] David W Opitz and Richard Maclin. “Popular Ensemble Methods: An Empirical Study”. *J. Artif. Intell. Res. (JAIR)* 11 (1999), pp. 169–198.

[45] Leo Breiman. “Bagging predictors”. *Machine Learning* 24.2 (1996), pp. 123–140.

[46] Yoav Freund and Robert E. Schapire. “Experiments with a New Boosting Algorithm”. *ICML’96 Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. 1996, pp. 148–156.

[47] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David a Guertin, Joo Han Chang, Robert a Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. “CellProfiler: image analysis software for identifying and quantifying cell phenotypes.” *Genome biology* 7.10 (2006), R100.

[48] F Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological Review* 65.6 (1958), pp. 386–408. arXiv: `arXiv:1112.6209`.

[49] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. *Nature* 323.6088 (1986), pp. 533–536.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. *ArXiv* (2015). arXiv: `1512.03385`.

[51] Alex Krizhevsky, Ilya Sutskever, and Hinton Geoffrey E. “ImageNet Classification with Deep Convolutional Neural Networks”. *Advances in Neural Information Processing Systems 25 (NIPS2012)* (2012), pp. 1–9. arXiv: `1102.0183`.