



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

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 appendix.

– Scott Warchal, 2018

ACKNOWLEDGEMENTS

Firstly I would like to thank my supervisor, Prof. Neil Carragher, for allowing to be a part of his research group, as well as all the advice and support over the years, opportunities to attend conferences and workshops around the world, and freedom to pursue my own ideas. I would also like to thank everyone in the Carragher groups as well as various others I have shared offices with, it's been a pleasure to work with you all – there has always been good humour and plenty of cake. Special mention goes to Alison Munro, John Dawson and Ash Makda for their help in the lab and answering my many daft questions with seemingly limitless patience. A thankyou to Cancer Research UK for funding this project. Finally I would like to thank my parents and my brother for their support and understanding over the years.

ABSTRACT

High content image-based screening assays utilise cell based models to extract and quantify morphological phenotypes induced by small molecules. The rich datasets produced can be used to identify lead compounds in drug discovery efforts, infer compound mechanism of action, or aid biological understanding with the use of tool compounds. Here I present my work developing and applying high-content image based screens of small molecules across a panel of eight genetically and morphologically distinct breast cancer cell lines.

I implemented machine learning models to predict compound mechanism of action from morphological data and assessed how well these models transfer to unseen cell lines, comparing the use of numeric morphological features extracted using computer vision techniques against more modern convolutional neural networks acting on raw image data.

The application of cell line panels have been widely used in pharmacogenomics in order to compare the sensitivity between genetically distinct cell lines to drug treatments and identify molecular biomarkers that predict response. I applied dimensional reduction techniques and distance metrics to develop a measure of differential morphological response between cell lines to small molecule treatment, which controls for the inherent morphological differences between untreated cell lines.

These methods were then applied to a screen of 13,000 lead-like small molecules across the eight cell lines to identify compounds which produced distinct phenotypic responses between cell lines. Putative hits from a subset of approved compounds were then validated in a three-dimensional tumour spheroid assay to determine the functional effect of these compounds in more complex models, as well as proteomics to determine the responsible pathways.

Using data generated from the compound screen, I carried out work towards integrating knowledge of chemical structures with morphological data to infer mechanistic information of the unannotated compounds, and assess structure activity relationships from cell-based imaging data.

LAY SUMMARY

Drugs act by altering the behaviour of cells, usually by disrupting the internal cellular machinery necessary for normal function, or in the case of diseases, by trying to reverse dysfunctional cellular processes responsible for disease initiation and progression back towards a normal state. Subtle changes in cellular functions can be detected visually through microscopy and fluorescent labels which bind to sub-cellular components such as DNA. Using automated image analysis methods it is possible to analyse these microscope images of cells and create a detailed description of each individual cell, represented as a series of measurements describing various attributes such as the cell's size, location and concentration of various biomolecules, this can be thought of as the cell's "fingerprint". Using these cellular fingerprints it is possible to test drugs in an effort to find those that convert a disease-like fingerprint into a healthy looking one, or to compare the fingerprints produced by unknown drugs to ones produced by molecules whose function is already known.

My work focuses on how to generate and exploit compound fingerprints across a number of different cells which represent different types of breast cancer. A significant challenge in studying distinct cancer cell types is that each cell has its own unique fingerprint regardless of drug treatment, which makes comparisons between cells more difficult. In addition, I investigate how more advanced computational tools alongside this varied dataset can aid predicting how novel compounds work.

CONTENTS

DECLARATION	i
ACKNOWLEDGEMENTS	iii
ABSTRACT	v
LAY SUMMARY	vii
CONTENTS	xi
LIST OF FIGURES	xiii
LIST OF TABLES	xiv
LIST OF ACRONYMS	xv
I 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	2
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	6
1.4.1 Cancer cell line panels	7
1.4.2 Breast cancer	9
1.5 Hypothesis, general aims and thesis structure	9
2 GENERAL METHODS	11
2.1 Cell culture	11
2.1.1 Culturing cells in 96-well plates for imaging	11
2.1.2 Culturing cells in 384-well plates for imaging	11
2.2 Generation of GFP labelled cell lines	12

2.3	Compound handling	12
2.3.1	24 compound validation set	12
2.3.2	Prestwick and BioAscent libraries	14
2.4	Cell painting staining protocol	14
2.5	Imaging	16
2.5.1	ImageXpress	16
2.5.2	Cell painting image capture	16
2.6	Image analysis	16
2.6.1	CellProfiler for 2D image analysis	17
2.7	Data analysis	17
2.7.1	Preprocessing	17
3	CELL MORPHOLOGY CAN BE USED TO PREDICT COMPOUND MECHANISM-OF-ACTION	19
3.1	Introduction	19
3.1.1	Machine learning methods to classify compound MoA	19
3.1.2	Ensemble of decision trees trained on extracted morphological features	20
3.1.3	Convolutional neural networks trained on pixel data	21
3.1.4	Chapter aims	23
3.2	Results	23
3.2.1	CNN predictions are improved using sub-images of just a few cells	23
3.2.2	More complex CNN architectures outperform simpler AlexNet	26
3.2.3	Standardising image intensity does not improve CNN model convergence	27
3.2.4	Decision trees did not benefit from feature transformation via principal component analysis.	28
3.2.5	CNN and ensemble based tree classifiers show equivalent performance at predicting MoA on a single cell-line	29
3.2.6	Additional data from more cell lines does not necessarily improve model performance	29
3.2.7	On the transferrability of classifiers applied to unseen cell lines	29
3.3	Discussion	31
3.4	Methods	34
3.5	Dataset	34
3.5.1	Accuracy	34
3.5.2	Ensemble of decision trees	35
3.5.3	Convolutional neural networks	35
4	MEASURING DISTINCT PHENOTYPIC RESPONSE	39
4.1	Introduction	39
4.1.1	Comparing response to small molecules across a panel of cell lines	39
4.1.2	Quantifying compound response in high content screens	39
4.2	Results	40

4.2.1	Compound titrations produce a phenotypic ‘direction’	40
4.2.2	Difference in phenotypic direction can be used to quantify distinct phenotypes	40
4.2.3	SN38 elicits a distinct phenotypic response between cell lines	42
4.3	Discussion	44
4.4	Methods	47
4.4.1	Data pre-processing	47
4.4.2	Principal component analysis	47
4.4.3	Selecting the number of principal components	47
4.4.4	Centering the data on the negative control	47
4.4.5	Identifying inactive compounds	48
4.4.6	Calculating θ and $\Delta\theta$	48
5	SCREENING APPROVED DRUGS ACROSS 8 BREAST CANCER CELL LINES	49
5.1	Introduction	49
5.1.1	Increasing the complexity of cellular models in drug discovery	49
5.1.2	Proteomics to interrogate hits from high-content screening	50
5.1.3	Screening approved drugs: repurposing old compounds	51
5.1.4	Chapter aims	51
5.2	Results	52
5.2.1	High-content screen of 1280 approved compounds	52
5.2.2	Validation in 2D and 3D apoptotic assays	53
5.2.3	RPPA	58
5.3	Discussion	63
5.4	Methods	68
5.4.1	Imaging and image analysis	68
5.4.2	Compound library	68
5.4.3	Multivariate Z-factor to determine assay quality	68
5.4.4	Identifying hits	69
5.4.5	2D apoptosis assay	69
5.4.6	Spheroids	70
5.4.7	RPPA	70
6	CHEMINFORMATICS AND HIGH-CONTENT IMAGING	75
6.1	Introduction	75
6.1.1	Cheminformatics	75
6.1.2	Structure activity relationships	76
6.1.3	Chemical similarity	76
6.1.4	Application of cheminformatics to high-content screening	77
6.1.5	The BioAscent library	78
6.1.6	Aim of this chapter	78

6.2	Results	79
6.2.1	The BioAscent library contains clusters of phenotypically similar compounds	79
6.2.2	The BioAscent library is chemically diverse	79
6.2.3	There is little evidence that structurally similar molecules produce similar cellular morphologies	80
6.2.4	Identifying the putative MoA of phenotypic hits with ChEMBL structure queries	81
6.2.5	Using phenotypic screening to find “dark chemical matter”	82
6.3	Discussion	82
6.4	Methods	86
6.4.1	Chemical similarity	86
6.4.2	BioAscent library screen	86
6.4.3	Phenotypic similarity	87
6.4.4	ChEMBL structure searches	87
6.4.5	Dark chemical matter	88
6.4.6	Interpro analysis	88
7	CONCLUDING REMARKS	89
7.1	Summary of completed work	89
7.2	Remarks, unanswered questions and new questions	90
8	APPENDIX	107

LIST OF FIGURES

1.1	Single cell aggregation to a median profile	4
1.2	Images of the eight breast cancer cell-lines	8
2.1	Example images of MCF7 cells for each MoA class	15
3.1	Diagram of a simple decision tree	20
3.2	Diagram neural network neuron and activation function.	22
3.3	Representation of a simple ANN	22
3.4	Down-sizing and chopping images for CNN training	24
3.5	Comparison of whole images vs sub-images	25
3.6	Sub image and whole image confusion matrices	25
3.7	Comparison of CNN architectures	26
3.8	Effect of image intensity normalisation on CNN training	28
3.9	Classifying MoA on a single cell-line	30
3.10	The effect of using additional cell-lines during model training	31
3.11	Confusion matrices of classifiers when applied to unseen cell-lines	32
3.12	Multi-GPU distributed training	36
3.13	CNN learning rate and decay	36
4.1	Compound distance in principal component space	40
4.2	PCA compound clustering based on MoA	41
4.3	Two compound titrations highlighted in phenotypic space	41
4.4	Visualisation of $\Delta\theta$ to quantify the difference in phenotypic direction between two compounds.	42
4.5	Visualisation of $\Delta\theta$ to quantify the difference in phenotypic direction between cell lines	43
4.6	Heatmap of $\Delta\theta$ between pairs of cell lines for separate compounds	45
4.7	TCCS workflow	46
5.1	Methods for creating tumour spheroids	50
5.2	Principal components of the Prestwick approved compound library after normalisation and standardisation.	52
5.3	Example images of GFP and DRAQ7 T47D cells imaged with the incucyte	54
5.4	Concentration-response curves for 12 hits from the Prestwick library in a 2D apoptosis assay.	55

5.5	Example images of T47D tumour spheroids.	56
5.6	Concentration-response curves for 12 hits from the Prestwick library.	57
5.7	Hierarchical clustering of RPPA samples.	59
5.8	Heatmap and clustering of RPPA data	60
5.9	Insensitivity of HCC1954 and SKBR3 to niclosamide in 2D	61
5.10	Insensitivity of HCC1954 and SKBR3 to ivermectin in 2D	62
5.11	Sensitivities of cell-lines in 2D to protryptiline	64
5.12	Sensitivities of cell-lines in 3D to protryptiline	65
5.13	Difference in protein expression in cells grown in 3D compared to 2D	66
6.1	Different methods to encode chemical structure	77
6.2	Selecting active compounds based on distance	79
6.3	Morphological clustering of the BioAscent library	80
6.4	Histogram of structural cluster sizes and example of molecules within a cluster	81
6.5	Comparison of structural and phenotypically similar compounds	82
6.6	Interpro target enrichment	83
6.7	BioAscent hits from dark chemical space	83
6.8	Popularity of terms 'bioinformatics' and 'cheminformatics' in the literature	84
6.9	Dendrogram threshold to determine clusters	87

LIST OF TABLES

1.1	Panel of breast cancer cell lines chosen for study	8
2.1	Cell seeding densities of 96 and 384 well plates	11
2.2	Annotated compounds of known MoA	13
2.3	Cell painting reagents and filter wavelengths for imaging.	16
3.1	Number of chopped images for MoA prediction	37
5.1	Z-factor values of assay quality for each cell-line	52
5.2	Number of active compounds in the Prestwick library per cell-line	53
5.3	Table of initial hits from the Prestwick library which produced distinct phenotypic response between cell-lines.	53
5.4	Rank products of top 15 Prestwick hits followed up in triplicate	54
5.5	Bradford standard BSA curve	71
5.6	Antibodies used in RPPA study	73

LIST OF ACRONYMS

2D Two-Dimensional

3D Three-Dimensional

5-HT 5-hydroxytryptamine

ABL Abelson murine leukemia viral oncogene homologue

ANN Artificial Neural Network

BCR Breakpoint Cluster Region

BSA Bovine Serum Albumin

CCLE Cancer Cell Line Encyclopedia

CCM Cerebral Cavernous Malformation

CNN Convolutional Neural Network

DMEM Dulbecco's Modified Eagle Medium

DMSO Dimethyl sulfoxide

ECFP Extended Connectivity FingerPrints

EMA European Medicines Agency

FDA U.S Food and Drug Administration

GDSC Genomics of Drug Sensitivity in Cancer

GFP Green Fluorescent Protein

GPCR G Protein Coupled Receptor

GPU Graphics Processing Unit

HCS High Content Screening

HTS High Throughput Screening

InChI International Chemical Identifier

MCL Markov Clustering Algorithm

MoA Mechanism of Action

MOI Multiplicity Of Infection

mRMR Minimum-Redundancy-Maximum-Relevancy

PBS Phosphate Buffered Saline

PCA Principal Component Analysis

PDD Phenotypic Drug Discovery

QED Quantitative Estimate of Drug-likeness

RGB Red Green Blue

RIPA RadioImmunoPrecipitation Assay

RPPA Rerverse Phase Protein microArray

SAR Structure Activity Relationship

SERT Serotonin Reuptake Transporter

SMILE Simplified Molecular Input Line Entry System

SSRI Selective Serotonin Reuptake Inhibitor

STS Staurosporine

TCCS Theta Comparative Cell Scoring

USR Ultrafast Shape Recognition

USRCAT Ultrafast Shape Recognition with CREDO Atom Types

1 | INTRODUCTION

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

Throughout the last 70 years the economic costs of developing novel drugs has increased dramatically, approaching £1 billion and requiring approximately 10 years from initial concept until regulatory approval. A study by Scannel *et al.*¹ noted that costs approximately double every 9 years, dubbing this observation “Eroom’s law” in a homage to Moore’s law.ⁱ The reasons behind these ever increasing costs are still under debate, although it is clear the issue is multi-faceted. One explanation may be that the low-hanging fruits of drug discovery have already been taken, the most effective traditional remedies and natural bioactive molecules identified, and their active ingredients commercialised. As such, whilst many single gene disorders and eminently druggable oncogene-driven homogeneous tumours have been cured, the more complex diseases and pharmacological targets remain. This pessimism has fed the ever present idea that drug discovery is undergoing a productivity crisis,² and that investments made in early stage research do not translate into actionable pharmacology to develop effective therapies, and 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 pre-defined protein targets. It would be difficult to argue that this target-based high-throughput screening (HTS) approach has not been fruitful, yielding many successful 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 commercial success stories, HTS is not a panacea, with a high attrition rate of lead compounds once they enter clinical trials.³ 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 limited validation of the hypothesised target in the face of complex disease aetiology.⁴

ⁱThe well-known observation that the number of transistors in microprocessors approximately doubles every 2 years.

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 more tractable.⁵

Many concerns related to phenotypic screening are centred on the lack of mechanistic information for a given lead compound. Whilst the lack of a known target presents challenges and 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 the drug is safe and efficacious. Metformin is a first-line therapy for type 2 diabetes and is on the World Health Organisation's list of essential medicines, it decreases liver glucose production and has an insulin sensitising effect on many tissues. Despite approval since 1957 and widespread clinical use, the molecular mechanism of metformin remained unknown for 43 years.⁶ Although knowledge of the molecular target is not necessary to get a drug into the clinic, target deconvolution is still an important part of most 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 instigating a conventional 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.⁷

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,^{8,9} *in vivo* studies in zebrafish¹⁰ and even plants and crops.¹¹

High content screens – screening studies carried out with high content imaging – are particularly useful in phenotypic drug discovery for several reasons. High content imaging provides spatial resolution enabling the use of more complex assays including co-culture and 3D models, which might better represent the biological complexity of disease relative to 2D reductionist models. However, these complex assays often have phenotypes 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 in-depth view into the endpoints which should be measured in a complex assay. A second benefit is the multivariate data generated by high content screening offers a more unbiased method for detecting hits in a phenotypic assay, as predicting which variable to measure beforehand may lead to missed biologically interesting phenotypes. 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

Image analysis is the process in which raw image data from a high-content screen is transformed into measurements which can be used to describe the observed morphology of the biological specimen exposed to a perturbation. 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 segmenting cells and sub-cellular structures into “objects”, and then computing image-based measurements on those objects. Typically each cell within an image is identified by first segmenting nuclei from the background. A number of well-established image thresholding algorithms can be used for segmenting nuclei from background, most automatically calculate an intensity threshold to binarise an image based on histograms of pixel intensities.^{12,13} 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 centre to approximate cell size. There are also less commonly used methods which utilise machine learning based on trained parameters to segment cells,¹⁴ or forgo segmentation entirely to measure morphological features from the raw images.^{15,16}

After cells and sub-subcellular 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.¹⁷

Texture. Measures of patterns of intensities within objects, typically derived from grey level co-occurrence matrices.¹⁸ This can be used to quantify morphologies such as small speckles or stripes within an image. Texture measurements are often computationally expensive and difficult to interpret although can be useful for measuring 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.

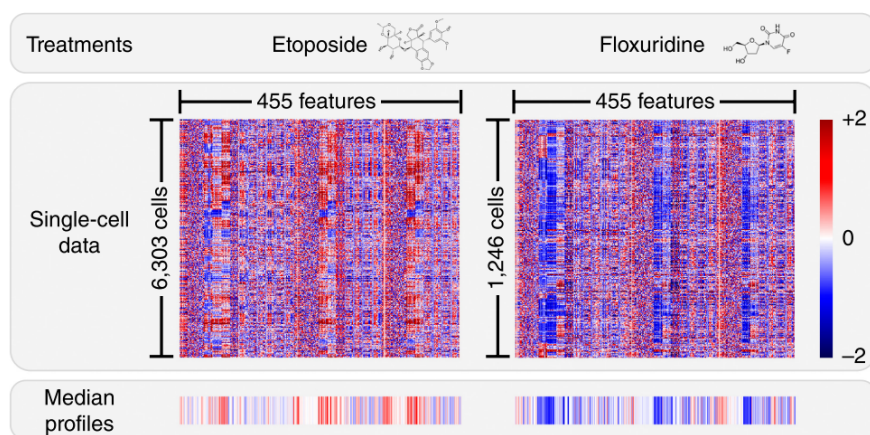


Figure 1.1: Single cell data aggregation to a median profile. Two matrices representing single cell morphology data for a treatment, with columns displaying multiple measured morphological features for each cell represented as a row. (Figure re-used from Caicedo et al. *Nat Methods*, 2017)

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 morphological 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.¹⁹ 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.¹⁹

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 poorly chosen segmentation parameters causing artefacts with otherwise acceptable images 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 for quality, therefore a number of automated 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,²⁰ whereas images containing debris such as dust and fibres contain a large percentage of saturated

pixels. Segmentation errors usually create extreme values for most feature measurements which can be highlighted using typical outlier detection methods such as Hampel filtering²¹ and local outlier factor.²²

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 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 are apparent they can be corrected, the simplest method is to standardise each batch separately, other methods include 2-way ANOVA²³ or canonical correlation analysis.²⁴

Standardisation. When many morphological features are measured from an image, they are unlikely to share the same scale/units or have similar variance – 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 useful to standardise all feature values to be mean centred and have comparable variance. This aids in many downstream data analysis methods which assume standardised feature values.

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”.²⁵ 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. A widely used technique is 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 can be an issue when using PCA, and is why some researchers favour feature selection methods which aim to retain original feature labels whilst 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-redundancy-maximum-relevancy (mRMR) feature selection method which has found use in high-content analyses.²⁶

Following data pre-processing, downstream analysis is typically focused 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 in a biologically relevant cell-based assay in order to identify compounds which produce a favourable phenotype and hits or lead compounds identified from a high throughput biochemical assay are evaluated in a more complex image-based cell assay to determine their quality. These assays typically rely on either a positive control compound which is known to elicit the phenotype of interest in order to optimise and validate that the assay has appropriate signal-to-noise attributes for testing multiple compounds. Or alternatively, a carefully designed assay in which a disease model utilising abnormal patient-derived or genetically engineered cells is used to identify compounds which revert the disease associated phenotype towards a healthy or wild-type phenotype. An example of this is demonstrated by Gibson *et al.*,²⁷ whereby they modelled 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. This is an elegant demonstration that combining good disease models with target agnostic phenotypic screens can effectively yield promising therapeutic candidates without complex bioinformatics techniques.

1.3.4 Image based profiling

In contrast to screening studies which are mainly interested in looking for a defined phenotype, profiling is 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 is 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 neighbouring treatments.

One of the landmark papers of high-content profiling was published in 2004 when Perlman *et al.*²⁸ first demonstrated that morphological profiles between drugs could be clustered according to compound mechanism-of-action using a custom similarity metric and hierarchical clustering. Most studies utilising morphological profiling use unsupervised hierarchical clustering in order to group treatments into bins which produce similar cellular phenotypes,^{29,30} although other clustering algorithms such as graph-based Markov clustering algorithm (MCL),^{31,32} and spanning trees³³ are sometimes used.

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 use in screening studies, giving rise to a number of anti-proliferative and cytotoxic compounds, which are still used in the clinic but 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 increased 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 targeting the BCR-ABL fusion protein in chronic myeloid leukemia. However, despite imatinib's exceptional success, unfortunately 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,⁷ 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,³⁴ 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 targeting microtubule stabilisation and DNA replication dynamics.⁷ 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 to explore novel biological mechanism and thus broader areas of therapeutic target space rather than relying on cellular death as catch-all phenotypic readout.

In addition to high-content imaging screens with cells grown in 2D monolayers, more complex phenotypic models such as 3D tumour spheroids are being increasingly adopted in pre-clinical oncology. 3D tumour spheroids are multi-cellular aggregates thought to better recapitulate environment and biology of real tumours compared to cells grown in 2D monolayers on tissue culture plastic. There is mounting evidence that spheroids offer a more predictive model of *in vivo* compound efficacy than their 2D counterparts,^{35,36,37} this is thought to be caused by the hypoxic environment in the centre of the spheroid, increased cell-cell contact and greater presence of extracellular matrix components which better represents conditions found *in vivo*. Three-dimensional spheroid models lend themselves well to phenotypic and image-based screening projects, with compound efficacy determined through use of fluorescent markers of cell-viability,³⁷ cell-cycle dynamics,³⁸ or by analysis of spheroid morphology which can also incorporate 3D volumetric measurements.³⁹

1.4.1 Cancer cell line panels

Panels of multiple cancer cell lines such as the NCI-60, Cancer Cell Line Encyclopedia (CCLE)⁴⁰ and Genomics of Drug Sensitivity in Cancer (GDSC)⁴¹ have been widely used to facilitate high-throughput screening and increase certainty in hit selection / disease-specificity,^{42,43} and as a research tool to study pharmacogenomics.^{44,45,46} The use of cancer cell line panels can also benefit phenotypic screens by mirroring the heterogeneity found in patient populations, as well as heterogeneous cell populations found in tumours.⁴⁷ Throughout this body of work I have used a panel of eight breast cancer cell lines (table 1.1 and figure 1.2 A), these cell lines were chosen based on a

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:Phosphoinositide-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.

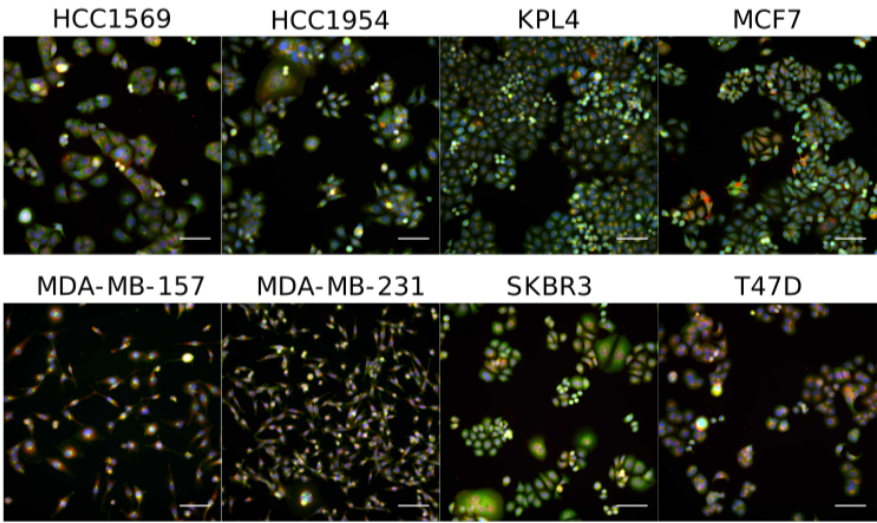


Figure 1.2: Composite image of cell-lines treated with 0.1% DMSO showing distinct morphology between untreated cell-lines. Channels used: Red - MitoTracker DeepRed; Green - Concanavalin A; Blue - Hoechst33342. Scale bars: 100 μ m.

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 morphometric imaging – larger and/or flatter cells allow for better discrimination of sub-cellular features.
5. Distinct morphologies to evaluate the robustness of morphological profiling methods.
6. A collection which represents a range of molecular sub-classes of breast cancer.

I.4.2 Breast cancer

The cell lines used in this work are all immortalised human cancer cell lines originating from breast cancer patients. Breast cancer cell lines were chosen as the disease has been the focus of many years of research resulting in many well characterised cell lines with freely available genomic, proteomic and imaging datasets. Breast cancer is sub-divided into several sub-classes defined by the molecular components which drive disease progression. The three main drivers of breast cancer are oestrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Aberrant signalling in one or more of these pathways is responsible for approximately 80-85% cases of breast cancer. The remaining 15-20% of cases are classified as triple negative (TN). Molecular sub-classes are used clinically to stratify patients based on immunohistochemically stained tumour sections examined by pathologists to inform therapeutic and surgical options. In addition to these simple subtypes, there are alternative and more complex methods of stratifying patients based on histopathological phenotype, response to endocrine and (neo)adjuvant therapy, and copy number alterations.⁴⁸

I.5 Hypothesis, general aims and thesis structure

As has already been discussed, image-based screens can generate large multivariate datasets which differ considerably from those usually found in high-throughput screening environments, the work in this thesis aims to address the hypothesis that informatics tools can be better utilised in the context of high-content screening in cancer drug discovery. This work aims to generate new high-content screening datasets across a panel of breast cancer cell-lines with which to compare, investigate and develop new data-analysis tools to better leverage the data present – as well as how best to combine this high-content imaging data with existing biological and chemical databases to better lead and inform early-stage drug discovery programmes.

The following chapters focus on selected topics from my PhD, some of the work has been previously published (see appendix).

- Chapter 2 contains general methods which are used throughout and apply to multiple chapters.
- Chapter 3 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 4 describes 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 5 describes a high content screen of 1280 approved small molecules in order to identify compounds which produced distinct phenotypic responses between cell lines, functional assays to validate hits and proteomics to investigate potential pathways responsible.

- Chapter 6 describes work towards developing methods which combine cheminformatics of compound chemical structure with high content morphological data in order to infer MoA of unannotated compounds, as well as assess the correlation of chemical similarity and phenotypic similarity.
- Chapter 7 presents concluding remarks about my work and future directions for the field.

2 | GENERAL METHODS

These methods are used throughout the work in this thesis and are listed here to reduce repetition. Each subsequent chapter will have a separate methods section which refers to methods unique to that particular chapter, or how they differ from the general methods described here.

2.1 Cell culture

The cell-lines were all grown in DMEM (#21969-035 gibco) and supplemented with 10% foetal bovine serum and 2 mM L-glutamine, incubated at 37°C, humidified and 5% CO₂.

2.1.1 Culturing cells in 96-well plates for imaging

Cells were seeded at roughly 3,000 cells per well (see table 2.1 for cell-line specific values) into the inner 60 wells of a 96-well optical bottomed imaging plate (#655090 Greiner) in 100 μ L of cell culture media. The outer 36 wells were filled with 100 μ L PBS. Plates were incubated for 24 hours in a tissue culture incubator before the addition of compounds.

2.1.2 Culturing cells in 384-well plates for imaging

Cells were seeded at roughly 1,500 cells per well (table 2.1) into each well of a 384-well optical bottomed imaging plate (#781091 Greiner) in 50 μ L of cell culture media. Plates were incubated for 24 hours in a tissue culture incubator before the addition of compounds.

Cell line	Type of plate	
	96 well	384 well
HCC1569	3000	1500
HCC1954	3000	1500
KPL4	2000	750
MCF7	3000	1500
MDA-231	2000	750
MDA-157	3500	2000
SKBR3	3500	2000
T47D	3000	1500

Table 2.1: Cell seeding densities for 96 and 384 well plates.

2.2 Generation of GFP labelled cell lines

Stable GFP expressing cell lines were created from the eight breast cancer cell lines in order to aid with spheroid image segmentation. Cells were seeded at approximately 35,000 cells per well of a 6-well plate in 3 mL of DMEM and incubated for 24 hours (37°C) to achieve 20% confluence. After 24 hours of incubation, 35 μ L of IncuCyte NuLight Green Lentivirus (#4624 Essen) was added to each well at an MOI of 1 with 1.5 μ L of polybrene (1:2000). Plates were then incubated for an additional 24 hours followed by a media change, and another 24 hour incubation. Media was then changed for selection media consisting of 1 μ g/mL puromycin and complete DMEM, followed by another 24 hour incubation. Following selection of puromycin resistant cells, cells were trypsinised and placed in a T75 tissue culture flask for further growth. GFP labelled cells and parental cell-lines were compared to ensure growth characteristics remained the same. This was achieved by measuring confluence in 6 well plates seeded with 10,000 cells per well and confluence measured with the Incucyte ZOOM. Following successful transduction, GFP labelled cells were maintained in 0.5 μ g/mL puromycin complete DMEM.

2.3 Compound handling

2.3.1 24 compound validation set

Compounds (table 2.2) were diluted in DMSO at a stock concentration of 10 mM. Compounds plates were made in v-bottomed 96-well plates (#3363 Corning), at 1000-fold concentration in 100% DMSO by serial dilutions ranging from 10 mM to 0.3 μ M 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.

Microtubule disruptor The microtubule disrupting compounds (paclitaxel, epothilone B, colchicine, nocodazole, monastrol, ARQ621) all act by disrupting tubulin. Paclitaxel and epothilone B both stabilise microtubules by binding to the α and β tubulin subunits to prevent depolymerisation, this over-stabilisation disrupts normal cellular functions such migration and mitosis which rely on the dynamic nature of the cytoskeleton. Colchicine and Nocodazole act on the same subunits of tubulin but instead cause destabilisation of the tubulin structure. Monastrol and ARQ621 are classed as microtubule disruptors but do not act on tubulin directly, instead they bind to the motor protein Eg5 kinesin which traverses microtubules and plays an important role in mitosis. The microtubule disruptors typically have effects on cell-cycle and large structural changes to cell-shape.

Aurora B inhibitor Barasertib and ZM447439 bind to and inhibit Aurora B kinase, a protein involved in the spindle checkpoint during mitosis. While both aurora B inhibitors and microtubule disruptors can interfere with mitosis, disrupting the spindle checkpoint can result in abnormal cell-division and missegregation of chromosomes during anaphase, whereas microtubule disruptors

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

typically cause arrest of the cell-cycle.

Actin disrupting Cytochalasin D and B are actin both drugs which inhibit actin polymerisation by binding to the F-actin to stop further addition of actin monomers, whereas latrunculin B binds actin monomers to prevent polymerisation. Jaskplakinolide differs in that it stabilises actin formation, although the exact mechanism is not clear. Actin disruptors have much in common with microtubule disruptors as they both exert their effects on the cytoskeleton, although actin disruptors can have direct effects on apoptosis and golgi organisation.

Protein degradation MG132 and lactacystin are both proteasome inhibitors. MG132 blocks the proteolytic activity of the 26S proteasome complex and also inhibits NF κ B activity. Lactacystin inhibits the 20S proteasome, and also NF κ B although to a lesser extent than MG132. ALLN and ALLM are calpain/cysteine protease inhibitors which play important roles in calcium signalling, cell proliferation and apoptosis.

Protein synthesis Emetine and cycloheximide both inhibit protein synthesis. Emetine binds to the 40S ribosomal subunit, whereas cycloheximide binds to the 60S subunit. Inhibition of protein synthesis has a wide range of effects on cellular function

Kinase inhibitor Dasatinib and saracatinib are kinase inhibitors with actions on Src kinase and Bcr-Abl tyrosine kinase, although achieving specificity in kinase inhibitors is difficult due to con-

served ATP-binding domains, so both drugs are reported to hit a number of other kinases.

Statin Lovastatin and simvastatin are statins, they inhibit 3-hydroxy-4-methylglutaryl-coenzyme A reductase (HMG-CoA reductase), a pathway responsible for the production of endogenous cholesterol. Lovastatin and simvastatin have also been shown to inactivate RhoA which in turn can lead to apoptosis and cell-cycle arrest.

DNA damaging agent Camptothecin and SN38 are DNA damaging agents which act by binding and inhibiting topoisomerase I, which prevents DNA unwinding, causing double strand breaks and ultimately cell-death.

2.3.2 Prestwick and BioAscent libraries

The Prestwick approved library and 12K BioAscent libraries were screened in the same run. Compound preparation was performed using the Biomek FX for automated liquid handlingⁱ. Source plates were diluted 1:10 in DMSO from 10 mM to 1 mM. From the 1 mM compound plates 1.5 μ L was transferred to an intermediate plate containing 74.5 μ L of cell culture media for a 1:50 dilution. From the intermediate plate 2.5 μ L was transferred to the assay plate containing cells seeded in a 50 μ L volume for a second dilution of 1:20 after factoring in estimated evaporation. A single intermediate plate was used for 8 assay plates corresponding to the 8 cell-lines. Assay plates were barcoded with cell-line and a sequential number corresponding to the compound source plate. 5 compound plates were screened with 8 cell-lines corresponding to 40 384-well plates each week.

2.4 Cell painting staining protocol

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 availability 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 microscopy setup.^{29,49}

The cell-painting protocol was initially optimised by Gustafsdottir *et al.* for use in the U2OS osteosarcoma 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 morphological changes 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

ⁱA thankyou to Ash Makda (Edinburgh) for helping set up the liquid handling protocol

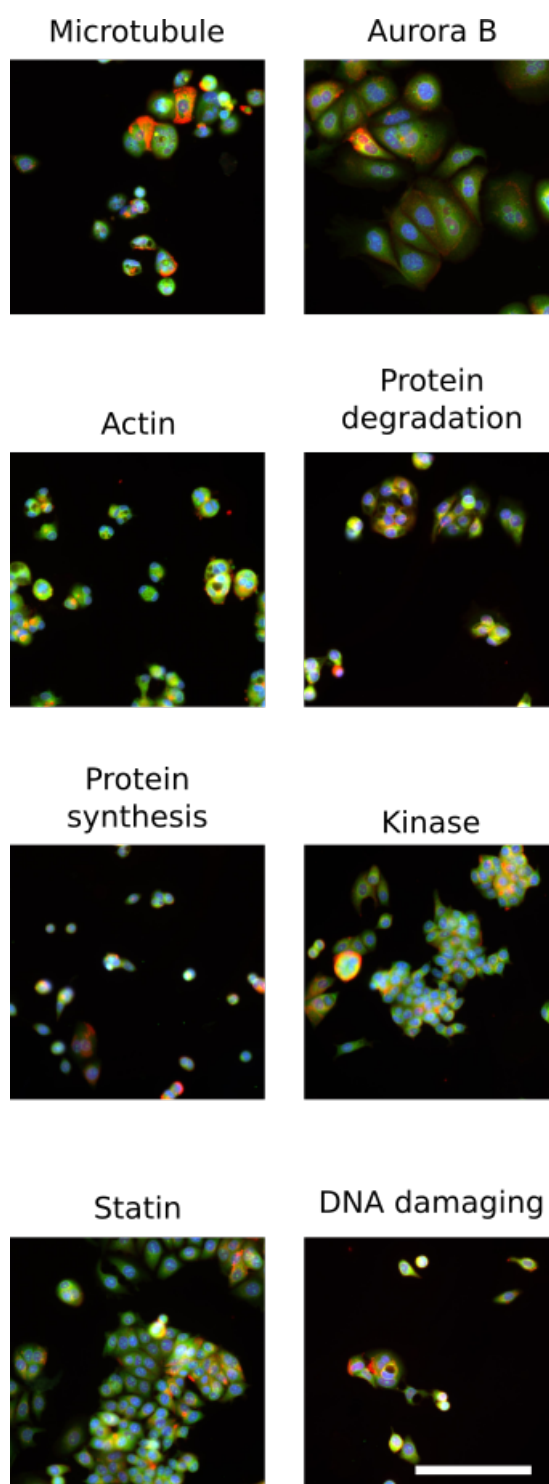


Figure 2.1: Example images of MCF7 cells showing a typical morphology for each MoA class at 1 μM . Microtubule: paclitaxel, actin: cytochalasin D, protein degradation: MG132, protein synthesis: emetine, kinase inhibitor: dasatinib, statin: lovastatin, DNA damaging: SN38. For untreated cells see figure 1.2. Channels used: Red - Mitochondrial Tracker DeepRed; Green - Concanavalin A; Blue - Hoechst33342. Scale bar: 100 μm .

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

Table 2.3: Reagents used in the cell painting protocol and the excitation/emission wavelengths of the filters used in imaging. ex: excitation, em: emission

media this prevented any alterations to the morphology and improved the wheat germ agglutinin staining, although as the MitoTracker stain relies on membrane potential of the 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.

To stain cells in a 96 or 384 well plates, the cells are first fixed by adding an equal volume of 8% paraformaldehyde (#28908 Thermo Scientific) to the existing media resulting in a final paraformaldehyde concentration of 4%, and left to incubate for 30 minutes at room temperature. The plates are then washed with PBS (100 μ L for a 96 well plate, 50 μ L for a 384 well plate) and permeabilised with 0.1% Triton-X100 solution (50 μ L 96-well, 30 μ L 384-well) for 20 minutes at room temperature. A solution of cell painting reagents was made up in 1% bovine serum albumin (BSA) solution (see table 2.3). Cell painting solution was added to plates (30 μ L 96-well, 20 μ L 384-well) and left to incubate for 30 minutes at room temperature in a dark place. Plates were then washed with PBS (100 μ L 96-well, 50 μ L 384) three times, before the final aspiration plates were sealed with a transparent plate seal (#PCR-SP Corning).

2.5 Imaging

2.5.1 ImageXpress

Imaging was carried out on an ImageXpress micro XL (MolecularDevices, USA) a multi-wavelength wide-field fluorescent microscope equipped with a robotic plate loader (Scara4, PAA, UK).

2.5.2 Cell painting image capture

Images were captured in 5 fluorescent channels at 20x magnification, exposure times were kept constant between plates and batches as to not influence intensity values.

2.6 Image analysis

2.6.1 CellProfiler for 2D image analysis

Images were analysed using CellProfiler v2.1.1 to extract morphological features. CellProfiler⁵⁰ 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 this takes considerable memory requirements, and therefore single cell-level data was aggregated to image median profiles. Briefly, cell nuclei were segmented in the Hoechst stained image based on intensity, clumped nuclei were separated based on shape. Nuclei objects were used as seeds to detect and segment cell-bodies in the cytoplasmic stains of the additional channels. Subcellular structures such as nucleoli and Golgi apparatus were segmented and assigned to parent objects (cells). Using these masks marking the boundary of cellular objects, morphological features are measured for multiple image channels returning per object measurements.

2.7 Data analysis

2.7.1 Preprocessing

Out of focus and low-quality images were detected through saturation and focus measurements and removed from the dataset. Image averages of single object (cell) measurements were aggregated by taking the median of each measured feature per image. Feature selection was performed by calculating pair-wise correlations of features and removing one of a pair of features that have correlation greater than 0.9, and removing features with very low ($< 1e^{-5}$) or zero variance. Features were standardised on a plate-by-plate basis by dividing each feature by the median DMSO response for that feature and scaled by a z-score (z) to a zero mean and unit variance by

$$z = \frac{x - \mu}{\sigma} \quad (2.1)$$

where μ is the mean and σ is the standard deviation. This is required as the measurements from CellProfiler use different scales. For example cell area is measured in pixels and typically ranges from a few hundred to several thousand, whereas cell eccentricity is constrained between 0 and 1. Large differences in feature scales causes issues with downstream pre-processing steps such as principal component analysis.

Principal component analysis Principal component analysis was calculated on the standardised CellProfiler features either using ‘prcomp’ in R or ‘sklearn.decomposition.PCA’ in python. Principal components were limited to the minimum number of principal components which accounted for 80% of the variance in the data. This was calculated by obtaining the variance explained by each principal component from the PCA output as a vector v , then calculating the cumulative proportion of variance explained by successive principal components as the cumulative sum of v divided element-wise by the sum of v to yield a new vector w of the same length. Then by finding

the minimum index of w at which the value of $w_i \geq p$ (where p is the portion of variance to be explained) returns the number n , where principal components $[1, \dots, n]$ were used as features.

3

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

3.1 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 which consisted of MCF7 breast cancer cells treated with 113 small molecules grouped into 12 mechanistic classes imaged in three fluorescent channels.⁴⁷ This dataset 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.^{51,52,53,54} 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 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 an unknown compound to the nearest similar annotated compound in a dataset to inform compound MoA, while at the same time being far cheaper than the transcriptional methods such as LINCS L1000.⁵⁵

3.1.1 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 be 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

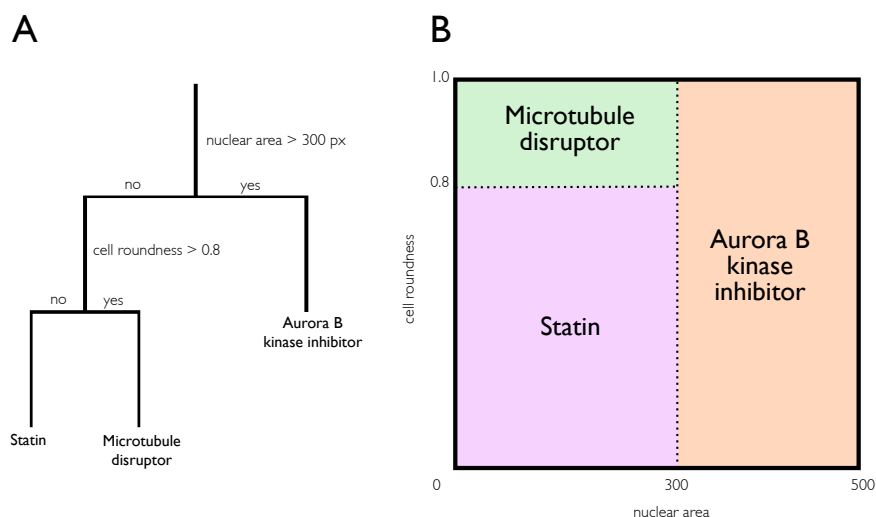


Figure 3.1: (A) An example of a simple mock decision tree to classify compound mechanism of action based on morphological features. (B) Depiction of decision space as divided by the decision tree model. Shaded areas show how new input data will be classified based on the decision rules (dotted lines).

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 it was not trained. Most classification 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.

3.1.2 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 3.1). Simple decision trees like those shown in figure 3.1 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.⁵⁶ Bagging⁵⁷ and Boosting⁵⁸ 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 trees, 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.

3.1.3 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,⁵⁹ they have only recently achieved widespread practical use due to improved methods for training⁶⁰ 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 3.2B).

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 3.3). 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.⁶⁰

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 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 convolutional neural networks (CNNs) by starting with many randomly initialised kernels, and updating the kernel values through training in order to best highlight features which prove useful for accu-

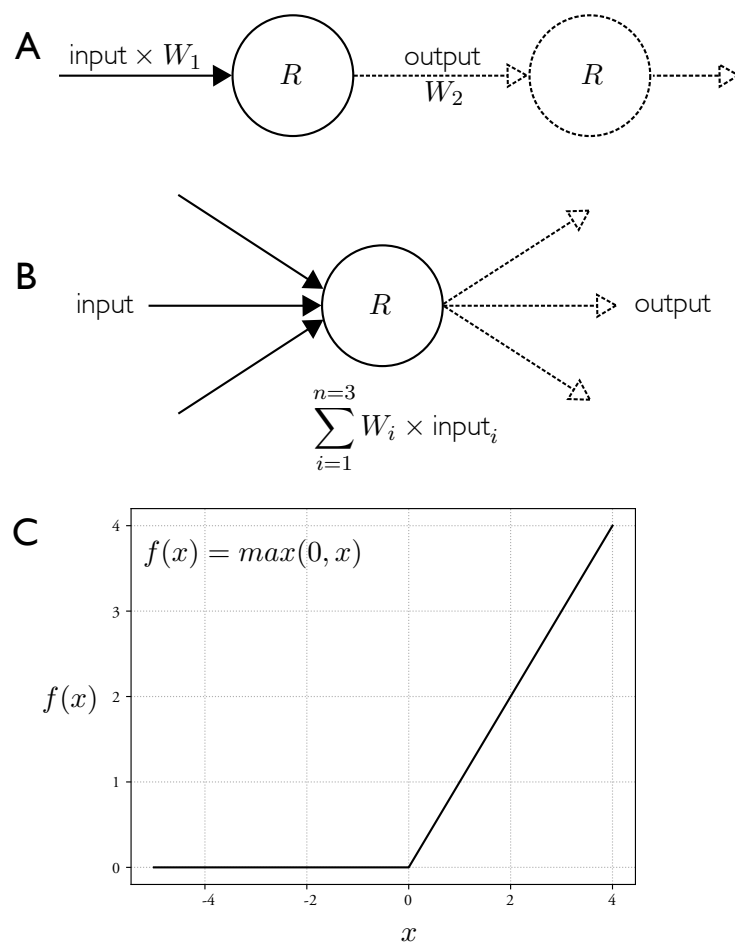


Figure 3.2: (A) A representation of a single connected neuron in an ANN, the input value to the neuron is multiplied by the weight (W_1), before being passed through the activation function R , the output of which is then multiplied by W_2 , and passed as the input to the next neuron. (B) A neuron with 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 then 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 .

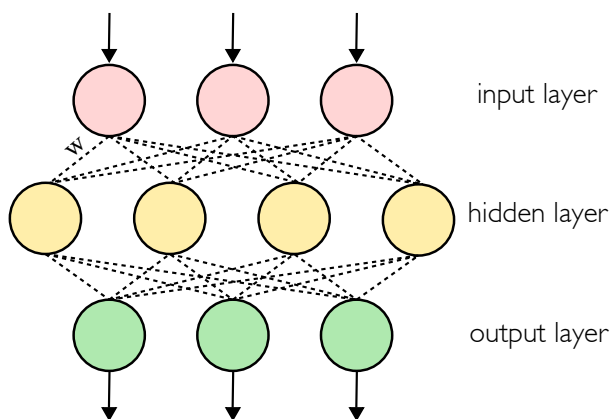


Figure 3.3 Representation of a simple 3-layer ANN with a single fully connected hidden layer, three input neurons and three output neurons. W denotes a weighted connection between an input neuron and a hidden-layer neuron, with all connections between neurons having an associated adjustable weight. A 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.

rately 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 3.3.

3.1.4 Chapter aims

The aims of this chapter are to assess how well machine learning models which predict compound MoA transfer across morphologically distinct cell lines. This is of interest as the ability to predict the MoA of unannotated compounds on a new cell-line with a pre-trained model without the requirement of re-screening an annotated compound library would save time and money. The compound library used in this work consists of 24 annotated compounds with well defined MoAs.

3.2 Results

3.2.1 CNN predictions are improved using sub-images of just a few cells

The images generated by the ImageXpress microscope are 2160×2160 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 smaller images suitable as inputs for CNNs, 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 3.4). 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 has been reported that whole well images are susceptible to over-fitting as the classifier 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 cell 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.

To compare the performance of using either downsized whole images or cropped sub-images, a pair of ResNet18 models were trained using either one of the datasets. It was evident during

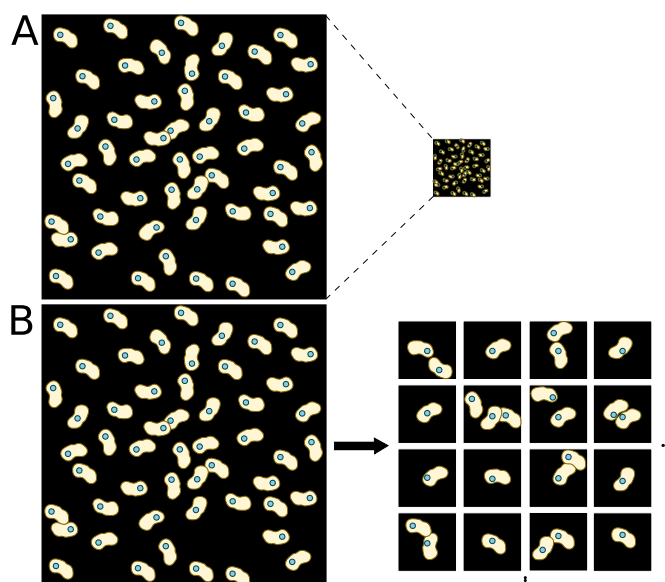


Figure 3.4: Two options for adapting large microscope images to work with the smaller input size of typical CNNs. **(A)** Full-sized images are downsized to the desired dimensions via bi-linear or bi-cubic interpolation. **(B)** Images are chopped into smaller sub-images, cell detection can be carried out beforehand to ensure images contain at least one cell.

training that using sub-images resulted in a higher final validation accuracy (0.847) compared to whole-images (0.778), as well as converging much faster than the whole-image-trained model (figure 3.5). Although downscaled whole images performed surprisingly well given their low resolution of cellular features.

It should be noted that the validation accuracy reported from the sub-image trained model is for classifying individual sub-images. One way to better use these individual sub-image classifications is to predict the parent image class based on a consensus of the predicted classes of the child sub-images. Using this consensus prediction, the sub-image validation classification accuracy increased from 0.847 to 0.912. Looking at confusion matrices calculated for both sub-image and whole images revealed that neither approach had difficulties at predicting a particular MoA class (figure 3.6).

Following these results the rest of the work involving CNNs used sub-images during training and prediction. Whilst sub-images improved model training and classification accuracy, it also introduces more complexity as images have to be pre-processed to identify cells and crop to a bounding box, it also introduces another parameter in terms of image size which has to be considered and optimised. While here I chose 300×300 pixel images corresponding to $97.5 \mu\text{m}^2$, this was chosen pragmatically to fully capture a single cell and a portion of any adjacent cells. This value could be optimised by running several models with differently sized cropped images, although this value is largely dependent on cell line characteristics, magnification and image binning.

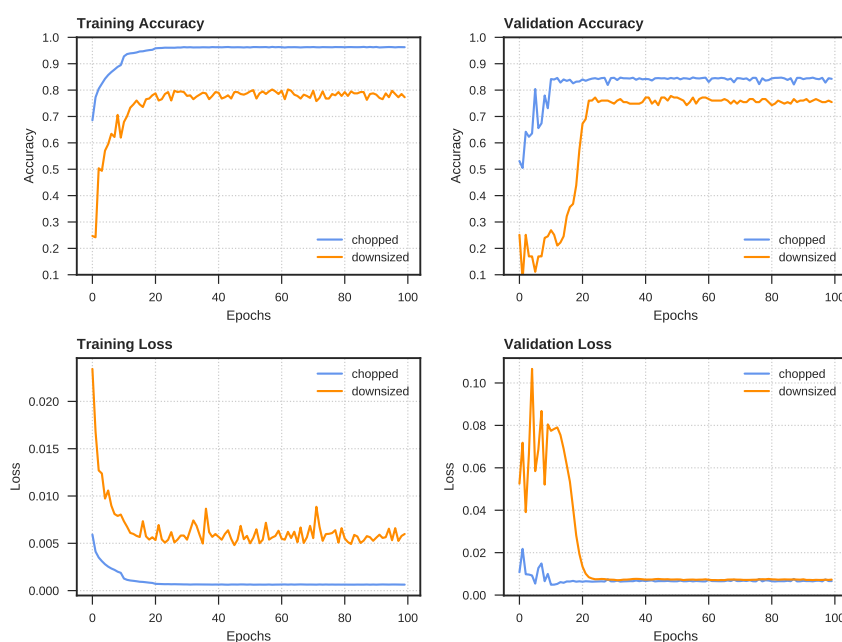


Figure 3.5: Comparison of training ResNet18 model on chopped sub-images vs downsized images from the MDA-MB-231 cell line. Chopped images were 300×300 crops centred on nuclei. Whole images were 2160×2160 images downsized to 300×300 pixels.

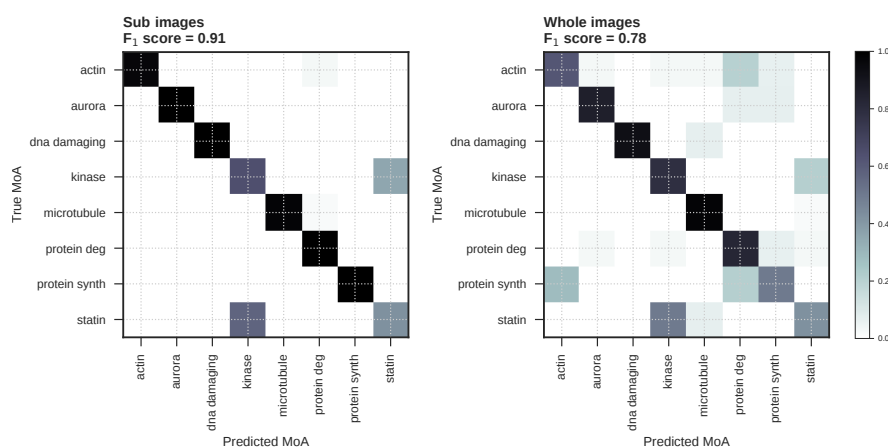


Figure 3.6: Confusion matrices comparing sub-image and whole image classification accuracy on 8 mechanistic classes of compounds with the MDA-MB-231 cell-line.

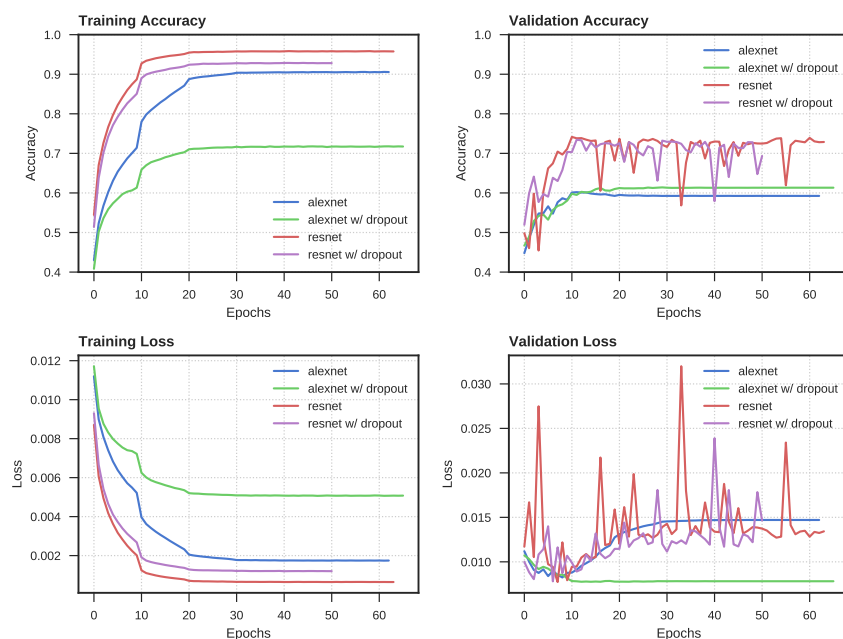


Figure 3.7: Comparison of CNN architectures. A comparison of AlexNet and ResNet18 architectures with and without dropout training and predicting on 5 channel 300×300 pixel images of all eight cell lines. Loss was calculated using cross entropy on 8 mechanistic classes of compounds.

3.2.2 More complex CNN architectures outperform simpler AlexNet

As CNNs can be constructed with a wide variety of architectures, and the field is still rapidly developing, I remained close to well established CNN architectures that are widely used in the field rather than developing my own. However, as most images are digitally 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 two different CNN architectures were tested based on the hypothesis that a deeper, more complex architecture (ResNet18⁶¹) will be capable of learning more subtle features, although more complex models with greater numbers of internal parameters are more prone to overfitting when training data is limited. On the other hand, a more simple model such as AlexNet⁶² which contains fewer convolutional layers will be less able to perform complex transformations 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 and perform better on new test data.

In an effort to reduce over-fitting, both models were evaluated with and without dropout in their dense layers during training. Dropout is a form of regularisation and works by randomly ignoring a fixed proportion of neurons during the training phase, with the theory that this prevents the model becoming too dependent on the output of a particular neuron and leads to more robust features used for classification.

Four models in total were trained on sub-images of all eight cell-lines pooled into a single dataset. The models were ResNet18, ResNet18 with dropout, AlexNet and AlexNet with dropout. During training the two ResNet18 models outperformed the AlexNets in both training and validation accuracy (figure 3.7). AlexNet with dropout layers did outperform the other three approaches when it came to validation loss, as loss did not increase even after many epochs this model demonstrated it is less liable to overfit data. However, the ResNet18 models showed a substantial increase in classification accuracy, and if training is limited to fewer than 10 epochs they do not show worse overfitting compared to the AlexNet models. Additional dropout layers does not seem to reduce ResNet18's liability to overfit beyond 10 epochs, this is not too surprising as the principal behind ResNet18's residual architecture is to limit overfitting, and adding additional dropout to the final fully-connected layers is a crude approach.

3.2.3 Standardising image intensity does not improve CNN model convergence

When training CNN models it is common practice to standardise image intensities. This pre-processing step consists of subtracting the mean of the image (or image batch) from each pixel and dividing the result by the standard deviation. The theory is this reduces training time and helps CNN models converge faster by ensuring the weights calculated during training are all on a similar scale which in turn restrains the gradients used in backpropagation. This pre-processing makes sense in the classical and traditional academic use of CNNs which are often trained on images or photographs from many different sources with inconsistent lighting and colours. However, the images used in this high-content screening dataset are all from a single microscope with a carefully controlled light source, in addition the intensities of the different channels carry a biological information relating to the abundance of different proteins or cellular structures. Therefore I wanted to assess if standardising image intensities per image channel improved model convergence and classification accuracy compared to un-normalised intensity values ⁱ.

Two models based on the ResNet18 architecture were trained on chopped 300×300 pixel images of a pooled dataset of all eight cell lines, one of the models was fed images standardised per channel, the other raw image intensities. After 48 hours of training (54 and 64 epochs for un-normalised and normalised models respectively ⁱⁱ) both models demonstrated identical training curves for training accuracy and loss, while validation accuracy and loss curves showed no striking difference in the performance between the two methods, although the normalisation pre-processing step appears to cause sudden drops in model performance indicated by decreased accuracy and increased loss (figure 3.8).

There is the possibility that training a model on disparate imaging datasets – from either different microscopes with different illumination settings, or different concentrations of reagents – then image standardisation may play a more important role. However, as intensity standardisation did not improve model performance in this case I chose to continue CNN work using un-standardised

ⁱAlthough un-normalised, intensity values were converted from 16 bit unsigned integers (65536 grayscale values) to 8 bit unsigned integers (256 grayscale values). This reduces training time and storage size at the expense of intensity accuracy.

ⁱⁱThe number of epochs per 48 hours does not indicate how fast a model converges, but rather the affect of availability of compute resources used for image loading.

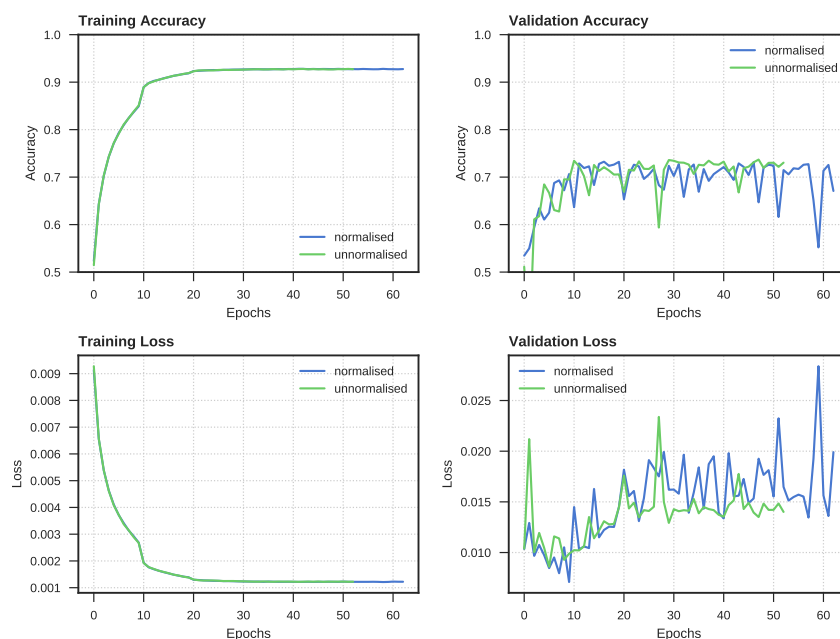


Figure 3.8: Effect of image intensity normalisation on CNN training. ResNet18 models training and predicting on eight pooled cell-lines with and without standardising image intensities per image per channel.

images, as there is an argument that standardisation may remove biologically relevant information for no benefit.

3.2.4 Decision trees did not benefit from feature transformation via principal component analysis.

Many machine learning methods benefit from feature selection or feature transformation. Here I tested if transforming the normalised CellProfiler features into principal components improved prediction accuracy with gradient boosting trees when training and testing on the MDA-MB-231 cell-line. I found that standardisation followed by principal component analysis and using all the principal components resulted in a decreased F_1 score of 0.80 compared to CellProfiler features which produced an F_1 score of 0.83. When the number of features were limited to the minimum number of principal components which explained 90% of the variance in the data, the F_1 score increased to match that of the original CellProfiler features. This meant that 16 principal components produced an equal classification accuracy as using 309 CellProfiler features, while this decreases computational time, it comes at the loss of interpretable morphological feature names, I therefore decided to continue the remaining analyses using the normalised CellProfiler features rather than principal components.

3.2.5 CNN and ensemble based tree classifiers show equivalent performance at predicting MoA on a single cell-line

Recently a number of studies demonstrated that CNN-based classifiers outperformed other existing methods when classifying high-content imaging data for predicting compound MoA.

When both ensemble based tree classifiers and CNNs are trained and tested on a single cell-line with separate training and test datasets they show equivalent performance (figure 3.9) at predicting compound MoA. The MDA-MB-157 cell-line demonstrated particularly poor performance when used with a CNN classifier (55.62%) compared to the average of 88.75% for all the CNN classifiers.

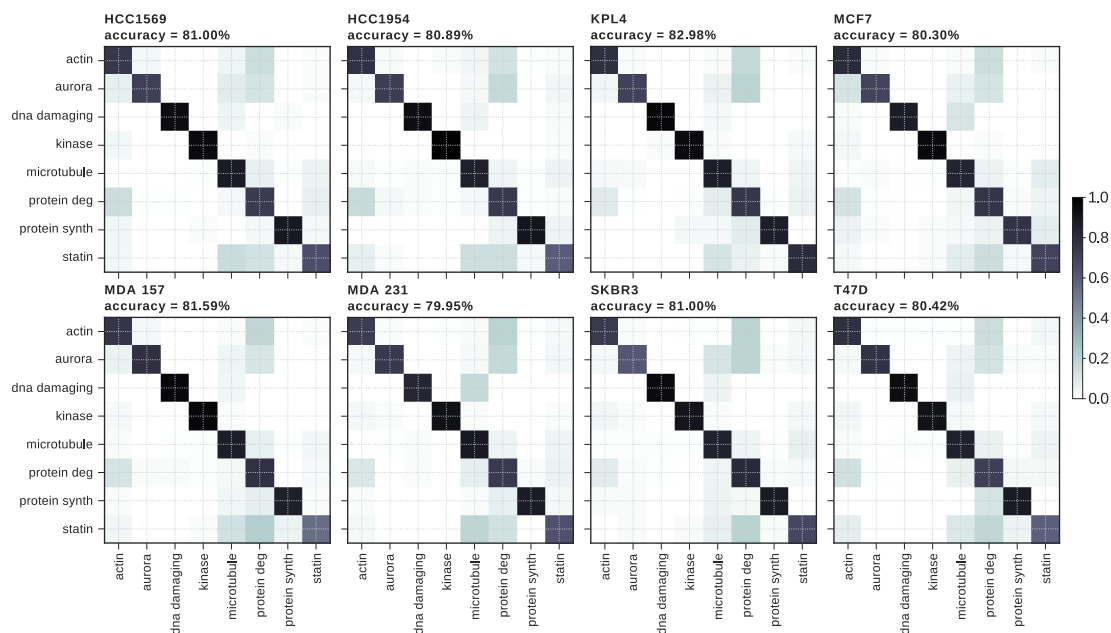
3.2.6 Additional data from more cell lines does not necessarily improve model performance

An adage in machine learning is that more relevant data during training is nearly always beneficial. With this in mind I investigated how training with additional data from morphologically distinct cell lines impacts model performance. I used 30% of the MDA-MB-231 cell line as a test set, and trained multiple tree-based and CNN models with the rest of the MDA-MB-231 and various combinations of increasing numbers of the other cell-lines (figure 3.10). I found that training with the additional data from different cell-lines negatively impacted the performance of CNN models, although interestingly model performance did not further decrease when even more additional cell-lines were used, as one additional cell-line produces similar classification accuracy as with using all 7 cell-lines. The tree-based models generally benefited from training with the additional data, although certain combinations of additional cell-lines did decrease prediction accuracy below that of just training and predicting on the MDA-MB-231 cell-lines. It was not clear which combinations of cell-lines caused this decrease in model performance, and no single cell-line was responsible for the regression in model performance. Owing to the considerable time taken to train the CNN models and the large number of possible combinations of additional cell-lines that could be used, I limited the CNN training to 2,3 and 7 additional cell-lines.

3.2.7 On the transferrability of classifiers applied to unseen cell lines

An important consideration of machine learning models is how well they generalise and transfer to new datasets. To investigate this I trained both tree-based and CNN models on 7 cell-lines and then tested on an unseen 8th cell-line, this was repeated so that all 8 cell-lines were tested as the unseen data. I found that both the tree-based and CNN models suffered a decrease in performance when applied to an unseen cell-line (figure 3.11). The tree-based models averaged 55% accuracy on the unseen cell-lines (figure 3.11 A), compared to 81% accuracy when trained and predicted on the same cell-line (figure 3.9 A). The CNN models suffered an even greater decrease in accuracy when transferred to morphologically distinct cell-lines with an average accuracy of 43% (figure 3.11 B) compared to 78% when trained and predicted on the same cell-line (figure 3.9 B).

A Gradient Boosting Trees



B CNN

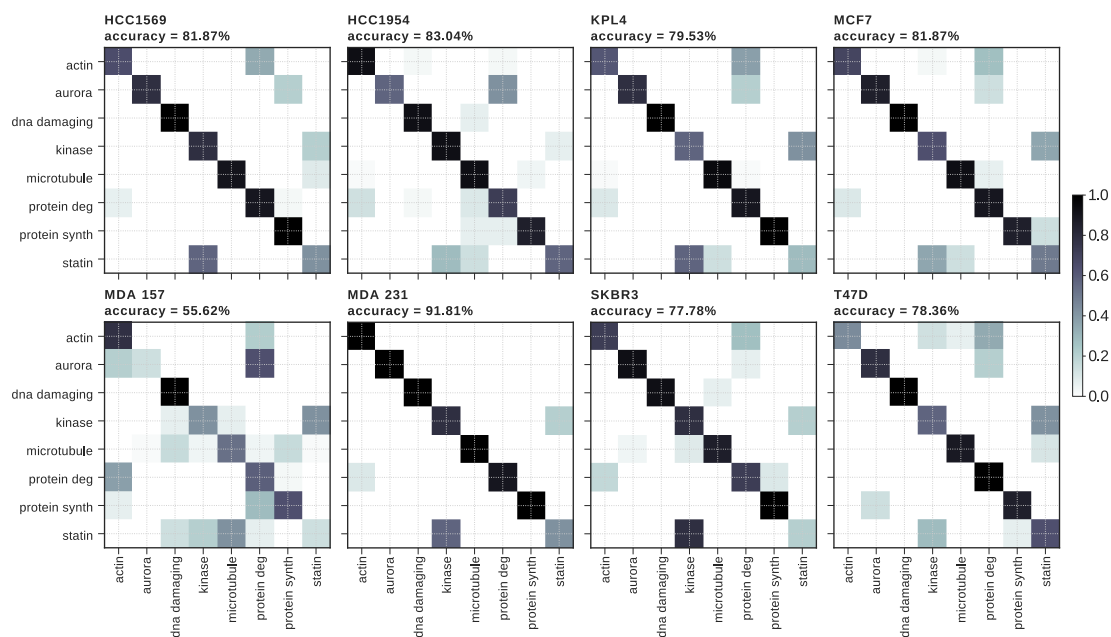


Figure 3.9: Comparison of ensemble based tree classifier and CNN at predicting compound MoA when trained and tested on an individual cell-line. **(A)** Gradient Boosting tree classifier. **(B)** ResNet18 CNN classifier. Accuracy measured as the F_1 score expressed as a percentage.

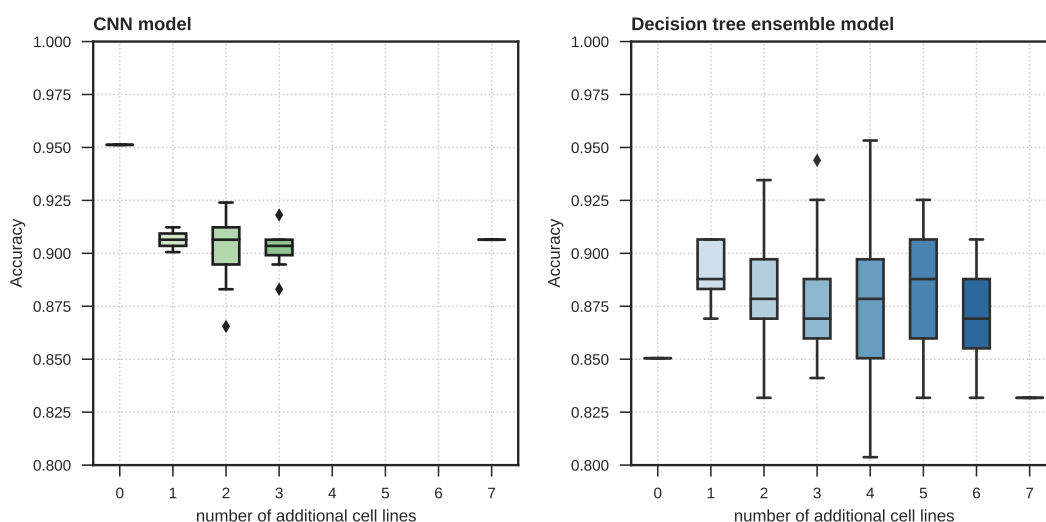


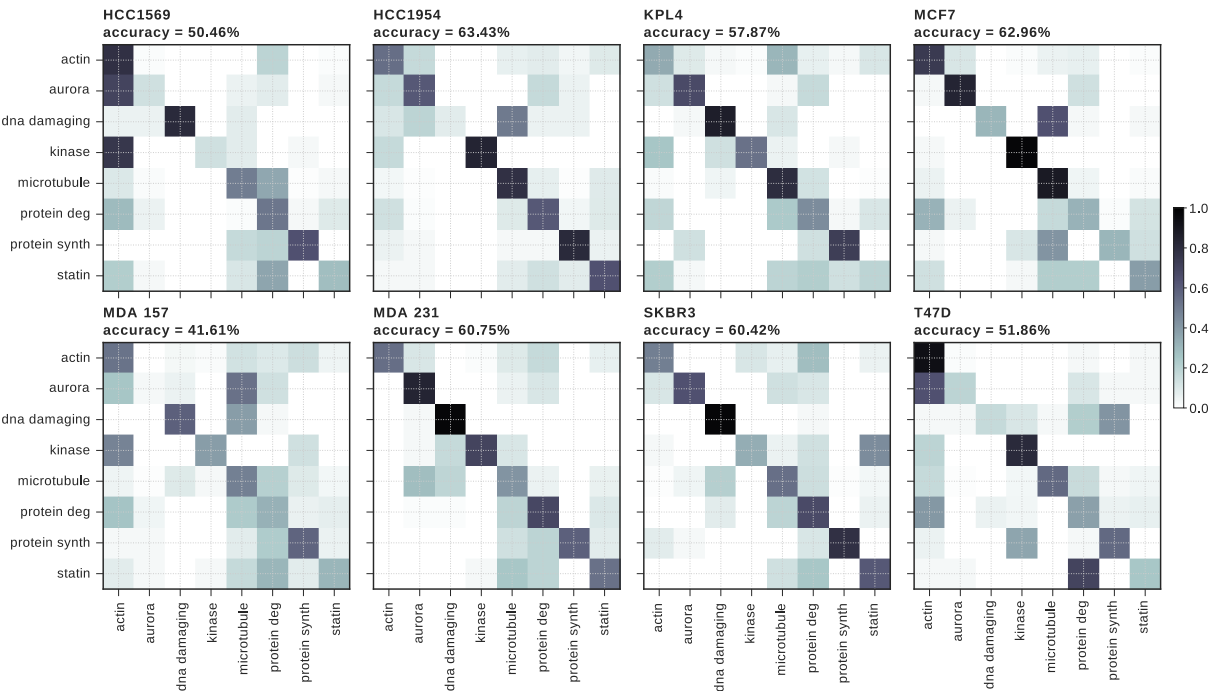
Figure 3.10: The effect of using additional cell-lines during model model training. Models accuracy when tested on a with-held proportion of MDA-MB-231 data. Box-plots show accuracy when trained on different combinations of additional cell-lines and tested on the MDA-MB-231 cell-line. The x-axis indicates the number of additional cell-lines other than MDA-MB-231 used during training, with 0 indicating the baseline model of training and testing on the MDA-MB-231 line.

3.3 Discussion

The main aim of this chapter was to assess to what extent machine learning models which predict compound MoA from high-content imaging data generalise to new morphologically distinct cell-lines. This generalisation would require the recognition of morphological features induced by small-molecules independent of the basal cell-line morphology.

Initial benchmarking studies with the two chosen models (gradient boosting decision trees and CNN) demonstrated roughly equivalent predictive performance when the models were trained with one cell-line and tasked to predict the MoA on withheld data from the same cell-line. This is in contrast to recent publications which have highlighted the ability of CNN-based methods to outperform other more classical approaches.^{54,53} One explanation for why this was not observed here is that I used fairly standard CNN architectures and pre-processing techniques, as my main interest was not absolute predictive performance but rather how well the models generalise. The results demonstrated by Ando *et al.* used novel pre-processing techniques on a carefully curated dataset, and Pawlowski *et al.* relied on fine-tuning CNN models which were pre-trained on the large ImageNet datasets – this was considered as an option, although there is no clear method to use the pre-trained models on my dataset as their weights are specific to the 3 channel architecture. Another option to increase overall predictive performance of the CNN models would be to use a deeper network with more parameters, especially as I already observed that the relatively complex ResNet18 architecture outperformed AlexNet. I decided to limit the complexity of the CNN models at ResNet18, as training the more complex models take considerably more time and computing resources than I had available.

A Gradient Boosting Trees



B CNN

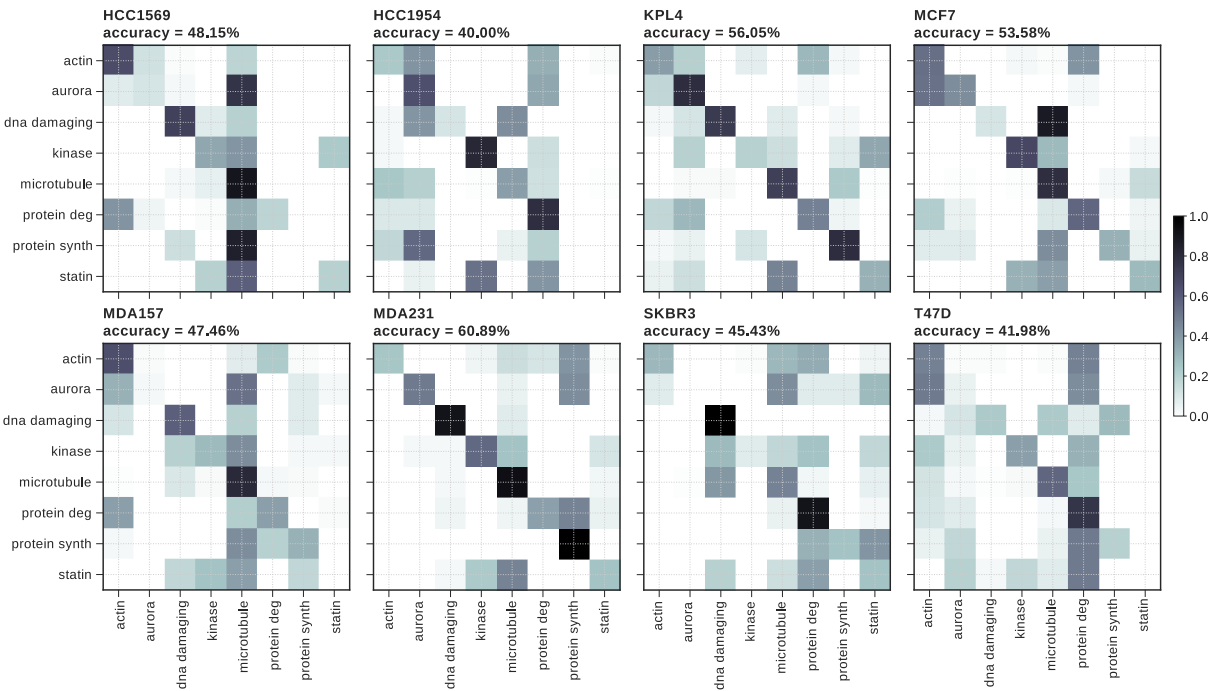


Figure 3.1 | Confusion matrices of classifiers applied to unseen cell-lines. Models were trained on 7 out of the eight cell-lines and tested on the with-held cell-line (named above confusion matrix). **(A)** Models trained with gradient boosting tree classifier. **(B)** Models trained with ResNet18 CNN. Accuracy is F_1 score expressed as a percentage.

The ability, or lack thereof, to generalise classifiers across multiple morphologically distinct cell-lines poses an interesting challenge in the field, a classifier trained on an extensive annotated compound library which demonstrated the ability to predict compound MoA from high content imaging data, either extracted morphological features or from raw images, would be a valuable resource. It is therefore of interest to see if such classifiers can be applied to new datasets of different cell-lines without the need to re-screen and re-train for each cell-line. I found that neither model types generalised particularly well, and suffered from significant decreases in prediction accuracy when applied to any of the unseen cell-lines. The tree-based models trained on extracted morphological features did however generalise slightly better than the CNN model. A likely explanation for this difference is due to the normalisation steps used to pre-process the CellProfiler data for the tree-based model, which divide the feature measurements by the negative control values per micro-titre plate and as each plate contained data from only a single cell-line, this essentially represented the compound effects as changes from the negative control. This normalisation step removes much of the cell-line specific morphology and may account for the increased accuracy of the tree-based models. I do not know of any current methods to apply a similar normalisation procedure to the image inputs of the CNN model in order to remove the cell-line specific morphologies from the images. One possible method to increase the generalisability of CNN models would be to use a greater degree of image augmentation during training. Image augmentation is the fairly common technique of modifying input images during training through distortions of shape and/or colour in order to reduce overfitting,^{63,64} if image augmentation disrupts basal cell-line morphology while largely preserving compound induced changes it may have the potential to produce more transferable classifiers. While difficult to normalise the images directly, it is possible to use CNNs as feature extractors to produce numeric data from images by truncating the CNN before the final classification and using the weights as a feature vector. It should be possible to normalise these weights against the weights of the negative control, as if using CellProfiler features, and use these as inputs to another classifier such as the gradient boosted trees. This combined approach of using the CNN to extract features from an image into numerical output which can be normalised and used with other machine learning tools has already been demonstrated in other domains.⁶⁵

Another method to efficiently adapt a CNN model to a new cell-line would be to use the idea of transfer learning, in which a small set of labelled data from the cell-line is used to fine-tune the later layers in the network. This relies on the idea that the early convolutional layers trained on the original dataset learn to identify features such as speckles and shapes which are applicable across multiple cell-lines, and that freezing these layers during fine-tuning that only the later layers involved in classification are adjusted with a small learning rate.^{53,66} While this has shown promise in adapting existing neural networks to new datasets, it still requires the use of labelled training data from the new application, and so while it may require fewer training examples when given a new cell-line, it still requires re-screening with an annotated compound library.

Overall I have found that current implementations of machine learning models used to classify compound MoA from high-content screening data do not generalise well to new cell-lines. I found that using extracted morphological features from software such as CellProfiler provides a flexible dataset which is easy to manipulate and pre-process when used as input for classification models.

Using raw images on the other hand is not as familiar for most investigators, and CNN models are much more difficult to interpret and understand how modifications to the input data will affect model outcome. Much of the machine learning field is currently focused CNNs and other neural network based methods, and they do offer some useful advantages in the high-content field such as not requiring segmentation or algorithms to measure hand-picked features. Further work investigating how to improve the generalisability of CNN-based classifiers is of great importance to many fields, and I predict that although it is currently a significant limitation, with the rapid development and commercial interests of this field it is likely many new methods will be developed which will aid MoA prediction.

3.4 Methods

3.5 Dataset

The imaging dataset used in this work is from the 24 compound validation set and is described in the general methods chapter (chapter 2). For each compound, data were used from the three highest concentrations (0.1 μ M, 0.3 μ M, 1.0 μ M). This was chosen as at lower concentrations many compounds failed to produce morphological changes that were distinguishable from the DMSO negative control.

3.5.1 Accuracy

Validation accuracy during training was measured using the Jaccard similarity score of the i th samples with true label set y_i and predicted label set \hat{y}_i :

$$J(y_i, \hat{y}_i) = \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|} \quad (3.1)$$

The F_1 score was used post training to determine classification accuracy. The F_1 score is the harmonic mean of both the precision and recall. So given true positives (tp), false positives (fp) and false negatives (fn):

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \quad (3.2)$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (3.3)$$

the F_1 score can be calculated as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (3.4)$$

3.5.2 Ensemble of decision trees

Models were created using scikit-learn (version 0.19) GradientBoostingClassifier in python 3.6.2, with default parameters except for 'n_estimators' which was increased from 100 to 600.

3.5.3 Convolutional neural networks

All code related to neural networks was written in pytorch (version 0.3, python 3.5) and ANN models were trained on nvidia K80 GPUs. The popular ResNet18 architecture was adapted to work with 5 channel numpy arrays rather than RGB colour images. Images stored with 300 by 300 pixel dimensions were downsized to 244 by 244 pixels using scikit-images "resize" function to match the input size required by ResNet and AlexNet. This was carried out by altering the first convolutional layer to accept 5 channels, and in turn increase the size of the input for the first linear layer from 512 to 2048 to account for the increased vector size after flattening the output from the convolutional layers. AlexNet was adapted in a similar way by increasing the number of channels in the first convolutional layer and a corresponding increase in the first linear layer after flattening. When testing the effect of dropout on ResNet18 and AlexNet, dropout layers were added between each of the linear layers, the proportion of dropout was 0.2, apart from the penultimate layer which had a dropout proportion of 0.5.

Data parallelism

As training CNNs is computationally expensive and time consuming, data parallelism was used to share batches of images across multiple GPUs trained in parallel. 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 3.12). This speeds up model training approximately linearly with the number of GPUs and allows use of larger batch sizes.

Training parameters

When testing image intensity standardisation, image intensities were standardised on an individual image and channel basis by taking each image in the form of an array [width × height × channel] and subtracting the mean of each channel from each pixel value in that channel, and dividing the pixel value by the standard deviation of the original channel.

Batch sizes during training were kept at 32 images per GPU. In the case of using GPU arrays then this was multiplied by the number of GPUs. Learning rate was set to $1e^{-3}$ decreasing 10-fold every 10 epochs (figure 3.13). Learning rate decay was used to aid gradient descent and model convergence (see figure 3.13). The optimiser used was ADAM⁶⁷ using the categorical cross entropy loss function for multi-class training.

Image preparation

The number of images per cell-line, test-train phase and MoA after image chopping are shown in table 3.1.

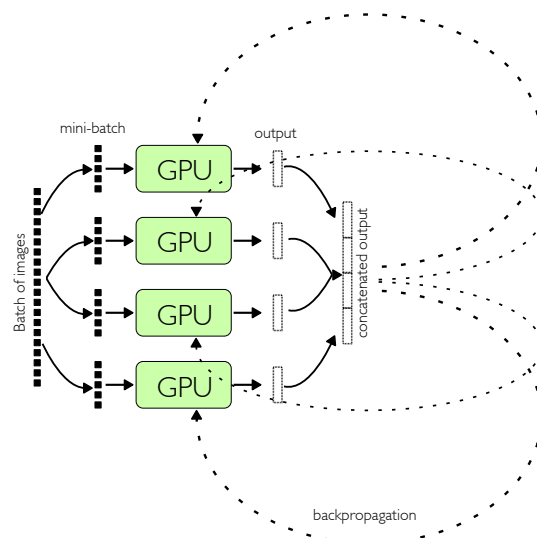


Figure 3.12: Increased training speed by data parallelism. Models are replicated across an array of GPUs, the input batch is split evenly among the devices, with each device processing a portion in parallel. During backpropagation the updated weights for all replicas are averaged and models weights are updated synchronously.

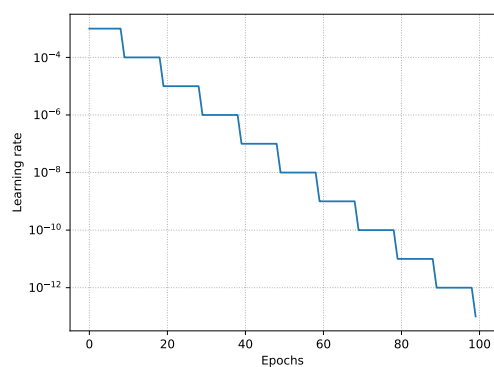


Figure 3.13: Learning rate and decay for training CNN models, initialised at $1e^{-3}$ and reduced 10-fold every 10 epochs.

Number of images

		actin	aurora	DNA damaging	kinase	microtubule	protein deg	protein synth	statin
HCC1569	test	1691	776	617	704	2804	2287	595	1100
	train	3780	2020	1491	1773	6656	4791	1752	2901
HCC1954	test	2612	1502	665	923	3234	3436	1200	1963
	train	6298	4760	1478	2940	6370	8312	2774	4015
KPL4	test	10588	5491	1279	3542	6474	16453	4711	4716
	train	22620	9914	3592	10770	12053	35911	9780	12951
MCF7	test	4591	2157	1075	2217	6128	6166	1694	3109
	train	10588	4804	2255	5858	12193	15093	4113	6818
MDA-157	test	627	435	222	274	913	1127	413	458
	train	1498	843	511	842	1333	2343	950	1010
MDA-231	test	4335	1688	1364	2094	5520	6253	1712	1993
	train	8628	5249	3148	4392	14369	13253	4589	6525
SKBR3	test	2543	1626	571	1722	2618	3681	1406	1884
	train	6183	3578	1635	4065	4468	8114	3209	4458
T47D	test	1402	1393	884	830	4526	2952	1070	1227
	train	4969	3235	2190	2009	8516	6616	2855	3273

Table 3.1: Number of images per cell-line after image chopping and test-train split.

Encoding

Images were originally stored as 16 bit tif files which were stored as 8-bit unsigned integers after chopping. While training, a custom data-loader transformed the 8-bit integers into floating point arrays to be used as input in pytorch. If images were intensity normalised then they were centered on zero by subtracting the mean and dividing by the standard deviation of each channel.

Image chopping

To chop the images for the CNN models I detected nuclei locations in the Hoechst stained image using a simple difference of Gaussian blob detection algorithm (`skimage.feature.blob_dog`, threshold of 0.1) to detect the centre points of bright objects in the form of x,y co-ordinates. These x,y co-ordinates were then used to calculate a 300 by 300 pixel bounding box to which the parent image was cropped in all 5 fluorescent channels. 300 pixels was decided on by assessing a number of different bounding box sizes and choosing the one that robustly captured enough of the image to contain the complete cell as well as a portion of neighbouring cells. For each cell-location, it was determined if the bounding box would be contained within the confines of the parent image, if the cell was located near the edge of the image and the bounding box would extend beyond the image border, then the x,y co-ordinates were adjusted so that the bounding box would be contained within the image borders.

Once the bounding box co-ordinates had been calculated for all cells within an image, the image was chopped into n sub-images, where n is the number of detected cells, and these sub-images were saved as individual 5 channel numpy arrays recording the parent image in the filename, which were used directly as input in pytorch. The image chopping code was released as a python package ⁱⁱⁱ.

ⁱⁱⁱwww.github.com/swarchal/NN_cell

4

MEASURING DISTINCT PHENOTYPIC RESPONSE

Note: this chapter is based on previously published work: "Development of the Theta Comparative Cell Scoring Method to Quantify Diverse Phenotypic Responses Between Distinct Cell Types", S Warchal, J Dawson, N.O Carragher. *ASSAY and Drug Development Technologies*, pages 395-406, 7:14, 2016. and "High-Dimensional Profiling: The Theta Comparative Cell Scoring Method", *Phenotypic Screening. Methods in Molecular Biology* 1787, 171-181.

4.1 Introduction

4.1.1 Comparing response to small molecules across a panel of cell lines

Comparative analysis of cell line panels treated with compounds are routinely used in pharmacogenomic studies and drug sensitivity profiling. These studies often use large numbers of cell lines and simple measures of compound response such as growth inhibition or cell death, allowing researchers to interrogate sensitivity of various small molecule therapies in a number of genomic backgrounds representing different diseases, disease-subtypes or patient populations.

Using high-content imaging methods with cell line panels enables more complex cellular readouts than cell death, creating a more detailed characterisation of compound effect. However, in order to apply multiparametric high-content data to pharmacogenomic studies, there needs to be a robust – and ideally univariate – measure of compound response to correlate drug sensitivity with genomic or proteomic datasets.

4.1.2 Quantifying compound response in high content screens

A simple but effective method to quantify the magnitude of compound response from multiparametric data is to calculate the distance from the negative control to the compound induced phenotype in feature space. This idea was first demonstrated by Tanaka *et al.* using PCA to reduce the dimensionality of a high content screening dataset to 3 principal components, and taking the distance from the centroid of the negative control replicates to the compound co-ordinates.⁶⁸ The distance from the negative control in PCA space is an effective metric for detecting phenotypically active compounds. In addition, distance measurements can be repeated for multiple concentrations of a compound to produce a concentration – phenotypic-distance response curve (see figure 4.1) and EC₅₀ values. However, one issue in calculating the distance-from-negative-control metric of compound activity is that it disregards much of the information relating to the position in feature space, as depicted in figure 4.1, two compounds may have similar distances yet those distances may be produced by very different morphological changes. In order to discern between two such compounds there needs to be a measure of directionality.

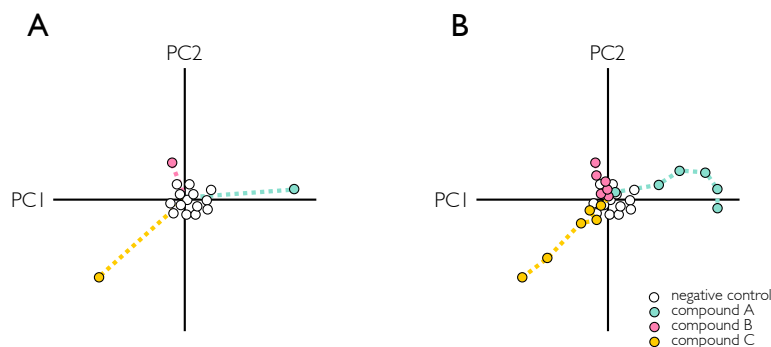


Figure 4.1: Diagram illustrating measuring magnitude of compound response by distance from the negative control centroid in principal component space. **(A)** Phenotypic distance to three different compounds. Compound A and C show phenotypic activity as they are distanced from the negative control cluster, whereas compound B shows little activity. Note that compound A and compound C have similar distances from the negative control centroid, yet have very different values in principal component space. **(B)** A titration series for each of the three compounds, showing how increasing concentrations of compounds A and C show increasing distance from the negative control, whereas weakly active compound B does not increase in distance. PC1: principal component 1. PC2: principal component 2.

4.2 Results

4.2.1 Compound titrations produce a phenotypic ‘direction’

Visualising high-content imaging data from compound screens in principal component space produces a representation of the overall structure of the dataset.

Using a dataset of morphological features produced by 24 compounds representing 9 mechanistic classes, plotting the first 2 principal components of this data reveals that compounds with the same MoA tend to cluster with one another (figure 4.2).

Plotting multiple concentrations of a compound in 2D PCA space allows us to visualise how an active compound becomes further away from the negative control with increasing concentrations. Figure 4.3 shows two compounds highlighted from the same data as in figure 4.2, we can see as compound concentration increases morphologies become increasingly distant from the untreated negative control cluster positioned centrally in the axes, with the two compounds producing opposite directions. Mirroring the differences in direction, the morphologies produced by barasertib and cycloheximide are also very different from one another, with barasertib – an Aurora B kinase – inhibitor producing large irregular nuclei, and cycloheximide creating small bright nuclei. This direction in PCA space can be thought of as a phenotypic direction, which can be measured and quantified independent from potency as measured by distance from the negative control cluster centroid.

4.2.2 Difference in phenotypic direction can be used to quantify distinct phenotypes

Using phenotypic direction in addition to distance from the negative control it is now possible to distinguish between equally phenotypically potent compounds with distinct morphological effects. By calculating an angle (θ) between phenotypic directions with cosine dissimilarity, a univariate

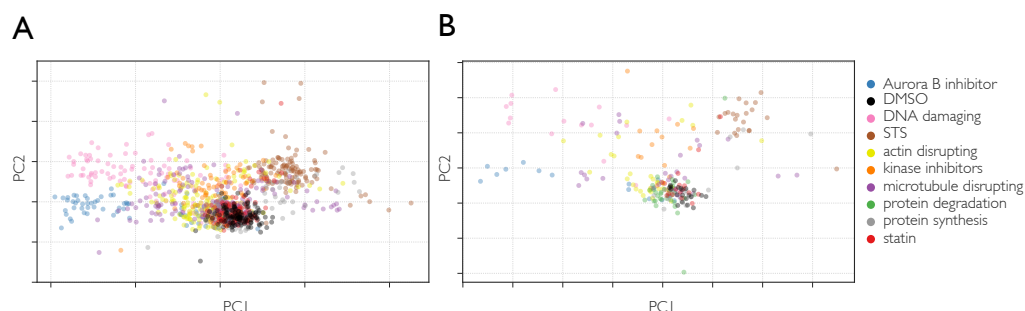


Figure 4.2: MoA clustering of compounds based on PCA of their morphological features. Principal components calculated from morphological features of 24 compounds grouped into 8 mechanistic classes. **(A)** Principal components calculated from an image average of individual cell measurements. **(B)** Each point represents a well average from individual cell measurements as each well contains 9 image sites. STS: staurosporine. DMSO: dimethyl sulphoxide

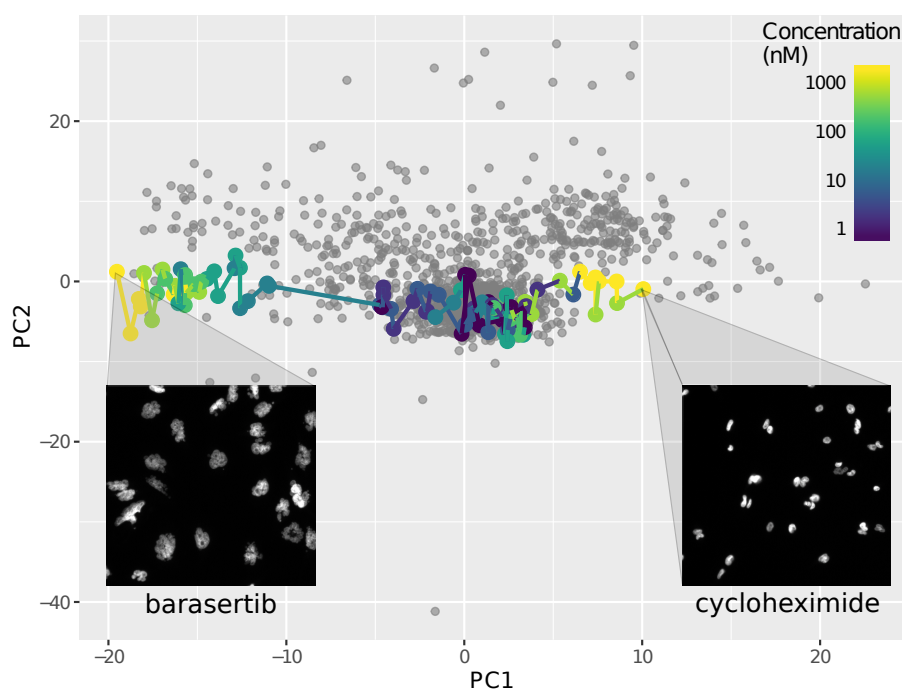


Figure 4.3: Principal components of Cellprofiler features calculated from a 24 compound high content screen in MDA-MB-231 cells. Barasertib (left) and cycloheximide (right) titrations are highlighted to show two active compounds with distinct phenotypes heading in different directions in phenotypic space with increasing concentration. Images shown next to points are from the Hoechst stain labelling nuclei morphology produced by 1 μ M of each compound.

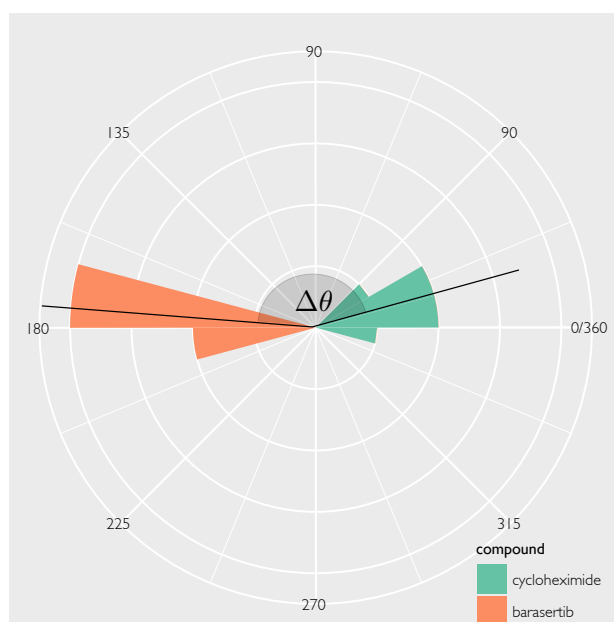


Figure 4.4: Visualisation of $\Delta\theta$ to quantify the difference in phenotypic direction between two compounds. Histograms in polar co-ordinates show the θ values of treatments against a fixed reference vector, with $\Delta\theta$ calculated as the difference between the average θ (black lines) of each compound.

value can be used to quantify phenotypic distance between either different compounds, or cell-lines treated with the same compound to detect distinct phenotypic response. By calculating θ against a fixed reference vector, the difference in θ ($\Delta\theta$) between two treatments can be quantified and visualised in polar co-ordinates as histograms or rose plots (figure 4.4). Compounds with the same phenotypic direction will have a small $\Delta\theta$ and compounds with dissimilar phenotypes having a large $\Delta\theta$, when expressed in degrees the values are constrained between 0° and 180° .

Although the data in figures 4.3 & 4.2 show the negative control points clustered near the median (0, 0) in principal component co-ordinates this is not guaranteed and should not be relied upon, so it is necessary to translate the principal components co-ordinates so that the negative control centroid is positioned over the median. In addition, inactive compounds will be positioned in close proximity to the negative control points and the calculated θ values will be misleading, therefore removing inactive compounds based on distance from the negative control is an important pre-processing step.

4.2.3 SN38 elicits a distinct phenotypic response between cell lines

Instead of calculating $\Delta\theta$ between compounds it is also possible to calculate $\Delta\theta$ between cell lines for a given compound. To identify and quantify differential phenotypic responses between cell-lines, $\Delta\theta$ was calculated between pairs of 8 breast cancer cell lines treated with 24 small molecules at three concentrations ($0.1 \mu\text{M}$, $0.3 \mu\text{M}$, $1 \mu\text{M}$). 21 out of the 24 compounds were found to be sufficiently active across the 8 cell lines to proceed, and the difference in phenotypic direction was calculated for all pairs of cell lines for each compound. Figure 4.6 shows a heatmap of the calculated

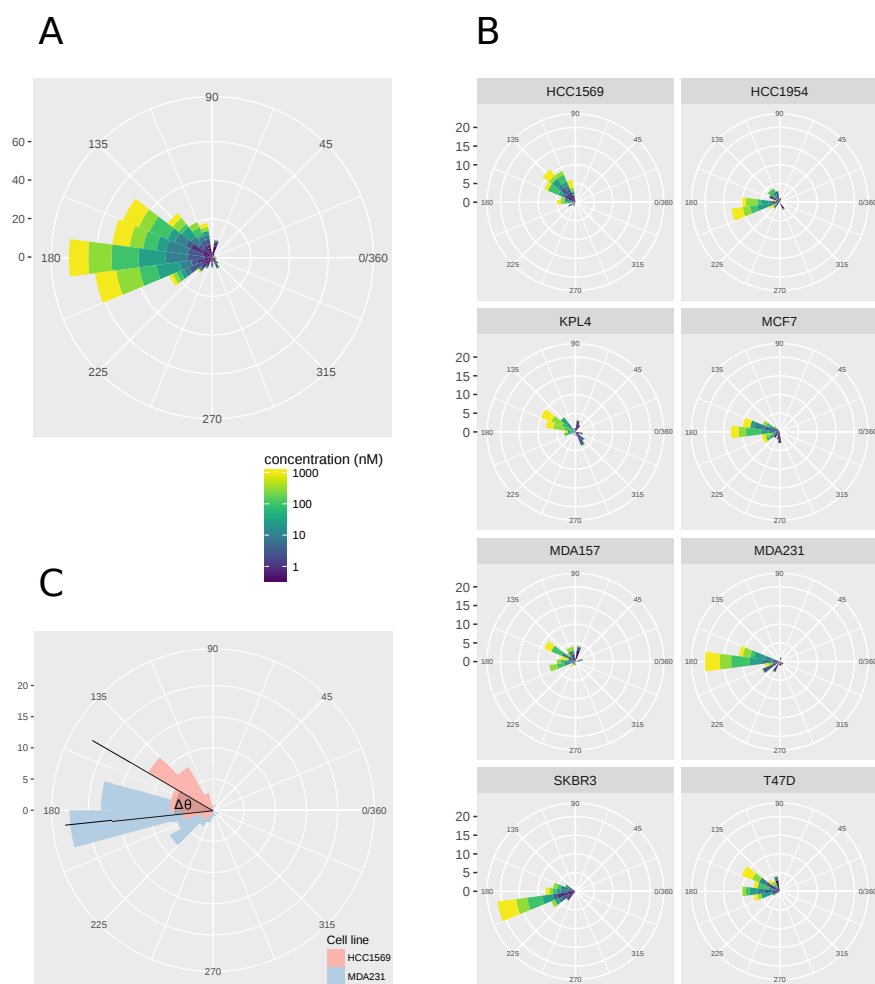


Figure 4.5: Visualisation of $\Delta\theta$ to quantify the difference in phenotypic response between cell lines when treated with barasertib. **(A)** Circular histogram of θ values of barasertib calculated for eight cell lines. **(B)** Phenotypic direction of cell lines treated with barasertib stratified by cell line. **(C)** Representation of $\Delta\theta$ for the difference between HCC1569 and MDA-MB-231 cell lines. Note that in this case $\Delta\theta$ is relatively small.

$\Delta\theta$ values. Some compounds such as the Aurora B inhibitors (ZM447439 and barasertib) showed very little difference in phenotypic response between the breast cancer cell lines, whereas compounds such as the topoisomerase I inhibitor SN38 demonstrated a single cell-line (KPL4) having a distinct response compared to the 7 others. Particularly striking is the difference between the MCF7 and KPL4 cell lines with a $\Delta\theta$ of 179° , indicated near opposite phenotypic responses between the pair of cell lines to the topoisomerase I inhibitor (figure 4.6).

4.3 Discussion

A number of methods exist to classify drug MoA and profile drug response in the context of high-content imaging studies, most of these have only been applied to a single cell type. The method described in this chapter, named as theta comparative cell scoring (TCCS), was developed to provide a pragmatic way to perform comparative high-content imaging studies across genetically and morphologically distinct cell lines. TCCS should be viewed as an extension to the common distance-in-PCA approach taking directionality into consideration in addition to distance from controls. The benefits of TCCS over previous methods are as follows: (1) the use of distance from the negative control to remove inactive compounds as one of the first steps prevents spurious differences that would be present in measures such as correlation or simple cosine similarity; (2) The comparison of each data point to a common reference vector enables visualisation of a phenotypic direction.

When comparing compound response between cell lines the most critical step, regardless of subsequent methods, is to account for the inherent morphological differences between untreated cell lines. Without this normalisation step morphologically distinct cell lines are not directly comparable as their large scale morphological differences will mask any difference in morphological response to a compound.

The TCCS method removes compounds which are deemed to be inactive if they are not sufficiently distant from the negative control (see figure 4.7). While this increases the robustness of the calculation by removing spurious differences in direction, it also introduces a new problem when compounds show large differences in potency between cell lines. This would result in the removal of such compounds from the analysis despite producing a genuine, and potentially biologically interesting, differential response between cell lines. This can be rectified by identifying these compounds when computing compound distances from the negative control in principal component space – any compounds that show large differences in this distance between cell lines can be flagged for further analysis before removal.

When using high-content imaging data with a lot of morphological feature measurements, using the first two principal components as depicted in this chapter may only account for a small proportion of variation in the data. This may lead to potentially missing interesting differences which are only evident in later principal components. Fortunately, as part of the TCCS algorithm the cosine similarity equation uses the dot product of the two vectors reducing any two equal length vectors to a single number, enabling the use of 3 or more principal components. Therefore the proportion of variance to keep in the data can be specified beforehand, and the dimensionality of the data reduced in a way to suit the statistical properties of different datasets.

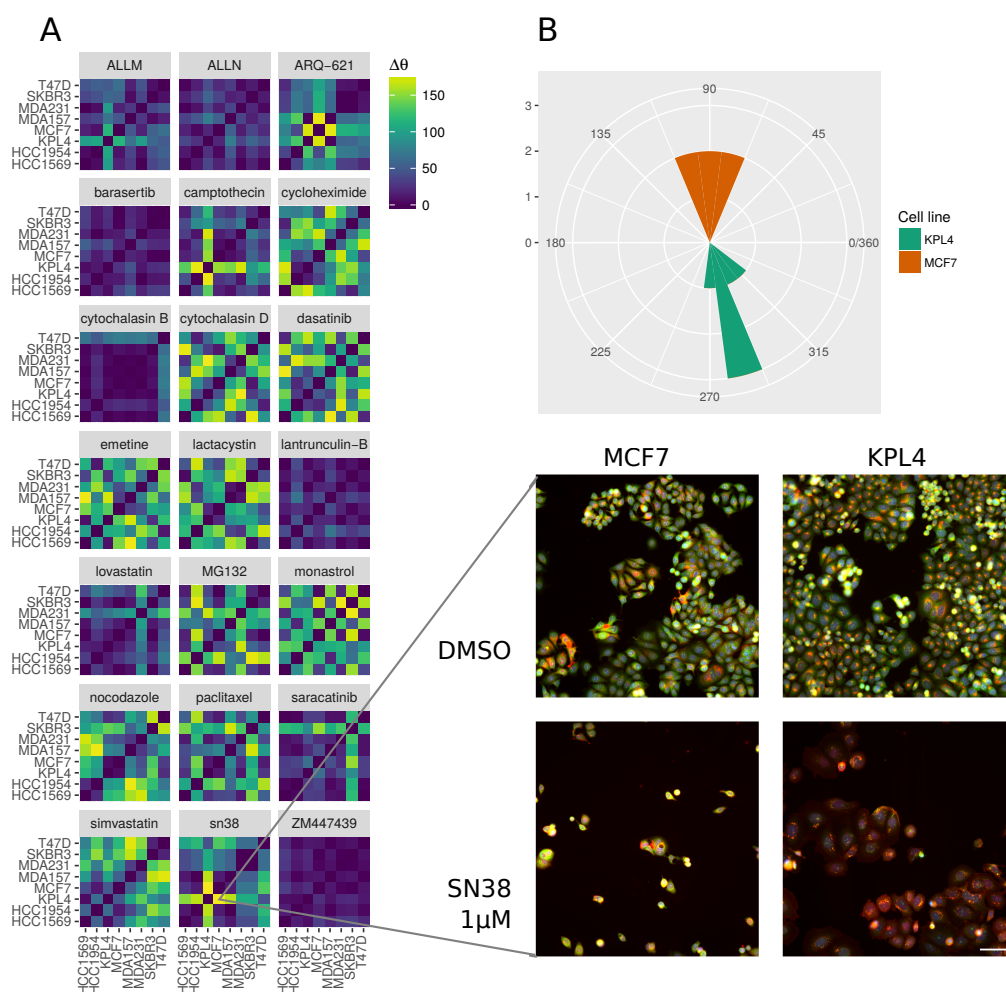


Figure 4.6: Heatmap of $\Delta\theta$ values between pairs of cell lines for 21 compounds which demonstrated phenotypic activity in all eight cell-lines. **(A)** $\Delta\theta$ calculated between pairs of cell lines treated with 21 compounds at (0.1 μ M, 0.3 μ M, 1 μ M). Images show differential response between KPL4 and MCF7 cell lines treated with 1 μ M SN38. MCF7 cells are observed to decrease in cell area with bright staining for the endoplasmic reticulum, whereas KPL4 cells produce a 'fried egg' morphology with large spread cells and weak endoplasmic reticulum staining. Channels used are as follows: Red - MitoTracker DeepRed (mitochondria); Green - Concanavalin A (endoplasmic reticulum); Blue - Hoechst33342 (nuclei). Scale bar: 100 μ m. **(B)** Histogram of θ values calculated for MCF7 and KPL4 cells treated with 1 μ M SN38. $\Delta\theta = 179^\circ$

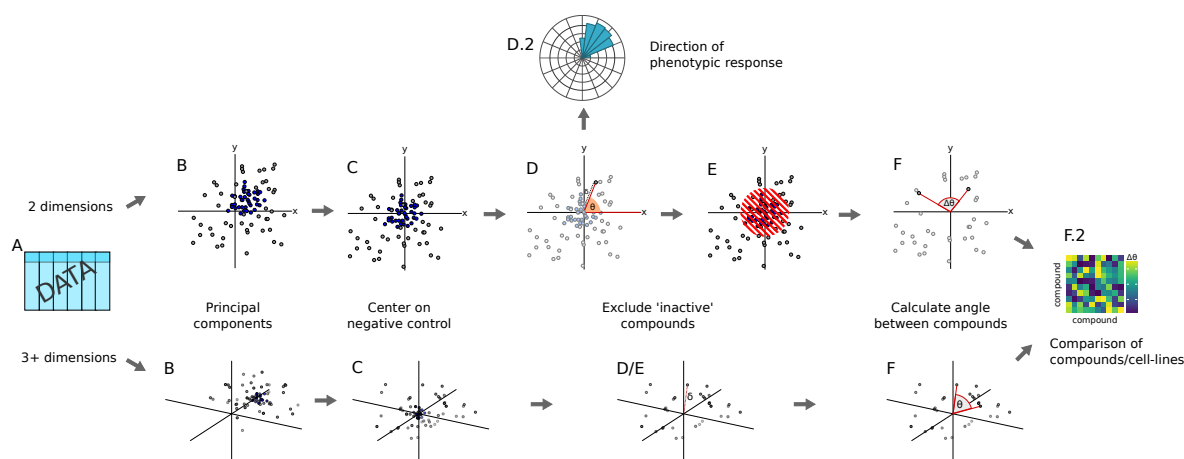


Figure 4.7: Theta comparative cell scoring (TCCS) workflow. **(A)** Normalised and standardised numerical data. **(B)** Principal component analysis, negative control values coloured in blue. **(C)** Centering of principal component values to the negative control centroid. **(D)** Calculation of distance from the origin to each data point, an activity cutoff is derived from the standard deviation of the distance to the negative control values. **(D.2)** In two-dimensional space, a directional histogram can be created by the angle of each vector against a reference vector. **(E)** Inactive compounds excluded based on distance from the origin. **(F)** Determining the angle between compounds/cell-lines. **(F.2)** Visualisation or clustering of compounds based on θ values.

An interesting prospect of ‘phenotypic direction’ is relating directions back to combinations of morphological features to provide more interpretability to the results. This is possible with PCA by using the feature loadings describe the contributions of original features used to construct each principal component. However, as PCA uses arbitrary positive and negative weights for these feature loadings, other dimensional reductions techniques might be better suited for generating more interpretable results. One example is non-negative matrix factorisation which would return only positive weights for the morphological features, making the contribution of morphological features to the phenotypic direction more interpretable.

Multiple concentrations are not often used in high throughput cell based screening assays despite providing useful information to detect off-target effects as well as reducing false negatives by screening at incorrect concentrations. A potential improvement of the TCCS method is to incorporate data from compound titrations as in figure 4.3 and fitting a linear model to the data points providing information relating to goodness of fit. This could potentially be used to identify compounds with off-target effects at higher concentrations if they do not fit a linear model well which indicates the data points going off at a tangent at higher concentrations towards phenotypic space indicative of cell death (e.g figure 4.1 B compound A).

In conclusion, the TCCS method presents an alternative to (dis)similarity measures such as correlation and cosine distance with important prior steps to account for peculiarities in high-content screening data, enabling high-content screening studies for quantifying distinct phenotypic response between morphologically diverse cell types.

4.4 Methods

4.4.1 Data pre-processing

Tabular data from Cellprofiler measuring 309 morphological features for each cell was aggregated to an image median. To remove batch effects and to remove inherent cell-line specific morphologies data was normalised by dividing each morphological feature by the median negative control value for that feature per plate. Each feature was then standardised to a mean of zero and unit variance on the pooled data.

4.4.2 Principal component analysis

Principal components were calculated using the `prcomp` function in R v3.2, with no centering or scaling as this was performed manually beforehand.

4.4.3 Selecting the number of principal components

The number of principal components to used in the analysis can be determined by specifying beforehand the proportion of variance in the data that should be kept, and then finding the minimum number of principal components that account for that proportion of variance in the dataset.

E.g in R:

```
1 threshold = 0.8
2 pca_output = prcomp(data, center = FALSE, scale = FALSE)
3 pc_variance = pca_output$stdev^2
4 cumulative_prop_variance = cumsum(pc_variance) / sum(pc_variance)
5 n_components = min(which(cumulative_prop_variance >= threshold))
```

where `data` is numeric dataframe of morphological features.

4.4.4 Centering the data on the negative control

In order to centre the principal component data so that the mediod of the negative control was positioned on the origin, the median value for each feature column for the negative control data was calculated. Then finding how much this differs from the origin for each feature, all principal component values were adjusted by this difference.

1. Calculate the median value m for each principal component for the negative control data (medioids).
2. Subtract each medioid from 0 in order to find the difference from the origin to δm_i , where i is the i^{th} principal component.
3. Add δm_i to each value in the i^{th} principal component.

For example in R, given a dataframe `data` containing a metadata column "`compound_name`" of compound names, with "DMSO" as a negative control, and `feature_cols` as a list of non-metadata column names:

```

1 | mediods = apply(data[data[, "compound_name"] == "DMSO"], 2, median)
2 | delta_m = 0 - mediods #  $\delta m$ 
3 | for (i in seq_along(feature_cols)) {
4 |   feature = feature_cols[i]
5 |   # feature_columni := feature_columni +  $\delta m_i$ 
6 |   data[, feature] = data[, feature] + delta_m[i]
7 | }

```

4.4.5 Identifying inactive compounds

Inactive compounds were identified by determining a minimum cut-off distance to the negative control centroid in principal component space. This was calculated by first finding the l_1 norm from each compound at all concentrations to the negative control centroid. The standard deviation of all these distances was calculated and any compound which was within 2 standard deviations of the negative control centroid at 1 μ M was deemed inactive, if a compound was found to be inactive in any one of the eight cell lines it was removed from the analysis.

4.4.6 Calculating θ and $\Delta\theta$

θ was calculated by taking cosine dissimilarity between two vectors (u and v) in principal component space and converting into degrees.

$$\theta = \cos^{-1} \left(\frac{u \cdot v}{||u|| ||v||} \right) \cdot \frac{180}{\pi} \quad (4.1)$$

When v is a common fixed reference vector, $\Delta\theta = |\theta_i - \theta_j|$ where θ_i and θ_j are theta values for 2 vectors. As opposite phenotypic directions are at 180° , $\Delta\theta$ values greater than 180° should be thought as converging towards similar phenotypes. Therefore $\Delta\theta$ values were constrained to a maximum value of 180° by subtracting any value greater than 180° from 360, or written as:

$$\theta = \begin{cases} 360 - \theta & \text{if } \theta > 180 \\ \theta & \text{otherwise} \end{cases} \quad (4.2)$$

5

SCREENING APPROVED DRUGS ACROSS 8 BREAST CANCER CELL LINES

5.1 Introduction

5.1.1 Increasing the complexity of cellular models in drug discovery

Immortalised human cell-lines are a widely used model to study cell biology and human disease. Recently there has been an increasing focus on the relevance of cells grown *in vitro* on tissue culture plastic in 2D monolayers and how these extremely artificial conditions compromise the predictive power of cellular models by their influence on cellular signalling pathways and response to external stimuli. This has triggered a number of studies suggesting further development and application of complex cellular models with the aim of better recapitulating the environment found *in vivo*⁶⁹. There is a wide range of 3D cellular models which have been developed for a number of different assays and physiological systems, although here the focus will be limited to tumour spheroids, which are 3D aggregates of one or more tumour-related cell-types which can range in size from 100 μm to a few mm. The assumption of tumour spheroids is that densely packed aggregate of cells with a gradient of nutrients, pH and metabolic waste from the outer edge of the spheroid to the hypoxic core characterised by dormant cells and poor drug penetration better resembles *in vivo* solid tumour micro-environment.^{70,71} There are a number of methods to produce tumour spheroids, the choice is largely a compromise of complexity versus scalability and reproducibility. The simplest method is through the use of low attachment U-bottomed plates and centrifugation of a cell-suspension to pellet cells together (figure 5.1 A), this leads to the formation of single uniformly sized spheroids in each well. A similar method is the hanging drop, which uses suspended drops of cell-suspension to create an environment for the cells to aggregate together in a single spheroid⁷² (figure 5.1 B). The hanging drop method has the advantage that while custom plates have been developed it does not necessarily require specialised consumables, although without the use of custom plates it is more labour intensive and there is an upper bound of the size of spheroids which can be created due to too large a droplet overcoming surface tension. A third method for generation of tumour spheroids is through the use of micro-patterned multi-well plates which provide a structure thought to aid cell-motility and aggregation. These have the advantage of ease-of-use as they do not require any additional steps, although the main drawback is the inconsistency of spheroid size, number, location, as well as requiring the use of expensive plates.

Despite the rapid adoption of 3D cellular models there is a lack of definitive evidence for their benefit over the more simple 2D models, in turn there are a number of additional issues which have

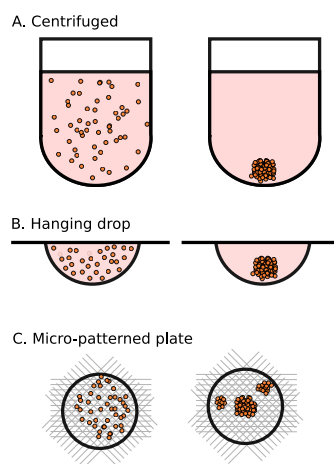


Figure 5.1: Methods for creating tumour spheroids. **(A)** Centrifugation in ultra-low attachment U-bottomed plates. **(B)** Hanging drop method. **(C)** Spheroid aggregation in micro-patterned plates.

to be addressed when using 3D tumour spheroids in an image-based assay. Cells located within the centre of the spheroid are often difficult to image and in turn segment due to limited penetration depth of light sources and poor labelling of fluorescent reagents. A number of commercially available high-throughput confocal microscopes are available which go some way to countering this issue, although to obtain adequate optical quality for single-cell segmentation usually requires chemical clearing methods. Assuming sufficient clarity of fluorescently labelled cells throughout the depth of a spheroid there is a choice of using the 3D data for segmentation and analysis, or to project the 3D structure onto a 2D plane using maximum projection or another similar algorithm and process the image using standard 2D segmentation and image analysis tools. While 3D data does offer a greater number of measurements through volumetric analysis and therefore a greater amount of morphological information, many researchers still opt for 2D image analysis of tumour spheroids owing to familiarity, the reduced computational resources required for storage and analysis, and the greater availability of established software tools.

5.1.2 Proteomics to interrogate hits from high-content screening

Interrogating hits found from target-agnostic phenotypic screens is often viewed as an important step to gain mechanistic information as well as generating hypotheses for new targets and disease aetiology. Methods such as thermostability shift assays, microarrays, RNAseq, whole-genome CRISPR knockouts and quantitative mass spec all have various strengths and weaknesses which make them appropriate for certain experimental questions. However, these methods are limited by either their focus on a single protein or by their limited throughput – mainly due to high costs – allowing analysis of only a small number of samples.

RPPA (Reverse Phase Protein microArray) is a miniaturised high-throughput antibody-based method for measuring abundance of total protein or translationally modified epitopes across a large number of samples. Protein lysates are spotted onto a solid substrate in multiple arrays which are then individually probed in parallel with mono-specific antibodies conjugated to a fluorophore,

protein abundance is then measured by comparing fluorescent signal against a dilution series of a known standard.⁷³ One of the main benefits of RPPA over other methods is the ability to process a large number of samples in parallel, which can be used to profile a number of treatment conditions, time-points, or concentrations. In contrast, the main limitation of RPPA is the reliance on high quality specific antibodies, which confines the detectable proteins and epitopes to those with well-validated and commercially available antibodies. RPPA has a number of advantages over other proteomic techniques such as high-sensitivity, high sample capacity and low sample consumption which makes it a well suited tool to investigate hits resulting from a target-agnostic high-content screen.

5.1.3 Screening approved drugs: repurposing old compounds

Repurposing an existing drug to treat a new disease or indication is an attractive strategy for cost-effective drug discovery programmes. Existing drugs have already been through pre-clinical and clinical safety studies, clinical trials and regulatory approval for their original indication and so the path from *in vitro* and *in vivo* screening to clinical use can be expedited and development costs greatly reduced. These advantages have resulted in a number of pharmaceutical companies investing time and resources into looking for new opportunities to reposition their existing compounds while still under-patent, as well as a number of new biotech companies hoping to re-patent old drugs under a new method of use.

Using knowledge of an existing drugs mechanism on known targets to treat other diseases which share similar molecular targets is a simple strategy for drug repurposing. One example is duloxetine, a serotonin and adrenergic reuptake inhibitor originally developed for the treatment of depression, which was later repositioned as an anti-incontinence therapy.⁷⁴ Serotonin and noradrenaline while well known for their effects on mood and behaviour, also produce an excitatory effect in smooth muscle neurons which can lead to an improvement in bladder control. This was noted by Eli Lilly who now have approval to market duloxetine as both an anti-depressant as well as the first approved urinary incontinence medication. Another approach to drug repurposing is to take advantage of so-called “off-target” effects. The non-specific binding to other protein targets can be leveraged for unrelated diseases; an example of this is itraconazole, a broad spectrum anti-biotic which has been found to act through the Hedgehog pathway as a potential anti-cancer therapy.⁷⁵

5.1.4 Chapter aims

The aim of this work was to screen a library of approved small molecules across a panel of eight breast cancer cell-lines to identify compounds which cause a distinct phenotypic response in one or more cell-lines. Then to investigate these compounds using functional 2D and 3D tumour spheroid assays of cell death and viability and to highlight potential pathways responsible for their selective breast cancer cell-line response using RPPA to measure the abundance of 60 proteins representative of canonical cell survival and cell proliferation signalling pathways.

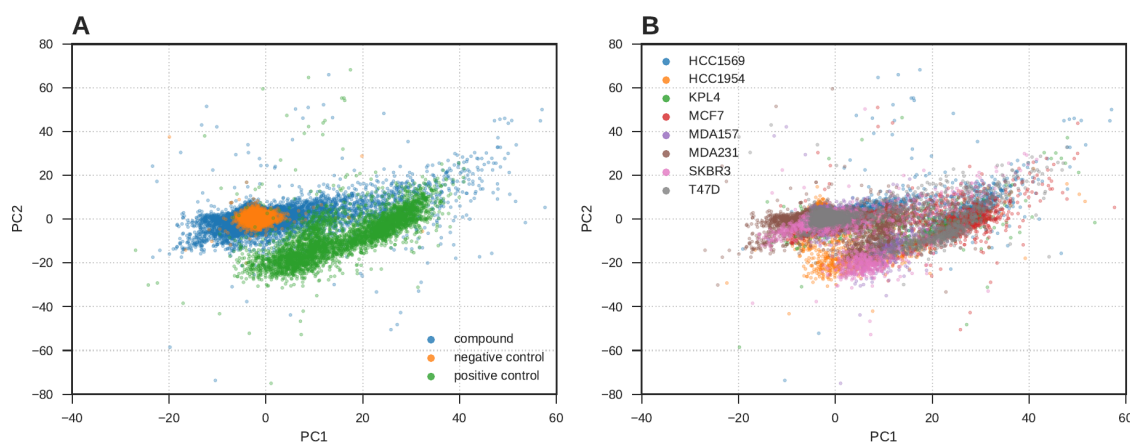


Figure 5.2: Principal component analysis of the Prestwick approved compound library after normalisation and feature standardisation. Points represent individual images. **(A)** Data points colour coded drug treatment, positive control (300 nM staurosporine) or negative control (0.1 % DMSO). **(B)** Data points colour coded by cell-line.

Cell line	Z-factor
HCC1569	0.72
HCC1954	0.77
KPL4	0.84
MCF7	0.77
MDA-MB-157	0.78
MDA-MB-231	0.74
SKBR3	0.79
T47D	0.79

Table 5.1: Multivariate Z-factor values of assay quality showing separation between the positive and negative control per cell-line.

5.2 Results

5.2.1 High-content screen of 1280 approved compounds

The Prestwick library of 1280 approved compounds was used in a high-content image based screen at a single 1 μ M concentration across all eight breast cancer cell-lines (table 1.1). Multiple morphological features were quantified from the images using Cellprofiler image analysis software and aggregated to an image median, and normalised to the plate negative control values and standardised. Plotting the first two principal components of this data revealed a clear separation between the positive (300 mM staurosporine) and negative control (0.1 % DMSO) (figure 5.2 A) with a multivariate Z-factor⁷⁶ of 0.6 for the pooled cell-lines, and greater than 0.7 for individual cell-lines (table 5.1) demonstrating a robust screening assay. In addition, the data from the morphologically distinct cell-lines was mixed and not separately clustered (figure 5.2 B), indicating that the normalisation step successfully removed basal cell-line morphologies enabling comparison of phenotypic response between morphologically distinct cell-lines.

For each phenotypically active compound in the Prestwick library the difference in response between pairs of cell-lines was measured using the TCCS method (chapter 4), and compound-cell-

Cell line	# active compounds
HCC1569	283
HCC1954	182
KPL4	236
MCF7	287
MDA-MB-157	96
MDA-MB-231	352
SKBR3	218
T47D	327

Table 5.2: Number of active compounds in the Prestwick library per cell-line. Compounds were defined as phenotypically active by calculating the l_1 norm distance from the negative control centroid in the principal components of the morphological features.

Compound	Usage / MoA
Amodiaquine	Anti-malarial.
Cisapride	5-HT ₄ agonist
Dilazep	Vasodilator. Adenosine reuptake inhibitor
Fluvoxamine	Anti-depressant. SSRI
Ivermectin	Anti-helminthic. GluCl agonist
Niclosamide	Anti-helminthic
Paroxetine	Anti-depressant. SSRI
Pirenperone	5-HT _{2A} antagonist
Podophyllotoxin	Microtubule destabiliser
Protriptyline	Tricyclic anti-depressant. NA, SERT
Triflupromazine	Antipsychotic. D ₁ , D ₂ antagonist
Zalcitabine	nucleoside reverse transcriptase inhibitor

Table 5.3: Table of hits selected from the Prestwick library which produced distinct phenotypic responses between cell-lines. SERT: serotonin reuptake transporter, SSRI: selective serotonin reuptake inhibitor, 5-HT: 5-hydroxytryptamine, D_{1/2} dopamine receptor.

line-pairs were ranked in terms of decreasing $\Delta\theta$ values. Compounds which demonstrated distinct phenotypic response were triaged by replication studies to confirm activity and selecting those with interesting MoAs for further studies as well as removing several microtubule disruptors. Twelve hits were selected for further study (table 5.3) based on phenotypic activity as determined with the l_1 norm from the negative control in principal component space, and rank by their ability to induce distinct phenotypic responses between the cell-lines measured with the TCCS method. Selected compounds were repeated in triplicate, and ranked by $\Delta\theta$ value for each replication, and used to calculate a rank product to rank overall distinct phenotypic effects between cell-lines (table 5.4 shows the top 15 compound-cell-line pairs).

5.2.2 Validation in 2D and 3D apoptotic assays

2D

Using GFP expressing cell-lines with DRAQ7 as a marker of cell-death I performed concentration-response experiments with the 12 selected compounds. Viable cells expressed nuclear GFP which was used as a simple readout of cell number, although the DRAQ7 apoptotic marker – which

Cell line A	Cell line B	Compound	Rank product	% False positive
HCC1569	MDA231	Podophyllotoxin	7.01	$3.88e^{-4}$
KPL4	MDA231	Podophyllotoxin	9.06	$4.66e^{-4}$
MDA231	SKBR3	Podophyllotoxin	12.11	$8.54e^{-4}$
HCC1569	HCC1954	Fluvoxamine	12.22	$6.31e^{-4}$
HCC1569	SKBR3	Ivermectin	13.16	$6.57e^{-4}$
HCC1569	HCC1954	Triflupromazine	13.66	$6.28e^{-4}$
HCC1954	MCF7	Ivermectin	16.54	$9.52e^{-4}$
HCC1954	SKBR3	Protriptyline	22.04	$1.90e^{-3}$
HCC1569	MDA231	Cisapride	22.10	$1.73e^{-3}$
HCC1954	SKBR3	Fluvoxamine	23.37	$1.81e^{-3}$
HCC1954	T47D	Triflupromazine	23.49	$1.68e^{-3}$
MDA231	T47D	Cisapride	25.34	$1.88e^{-3}$
HCC1569	MDA157	Zalcitabine	25.96	$1.88e^{-3}$
HCC1954	T47D	Protriptyline	26.14	$1.76e^{-3}$

Table 5.4: Table showing the 12 selected Prestwick compounds repeated in triplicate, which were ranked by decreasing difference in phenotypic response between cell-lines and used to calculate a rank product. Table shows the top 15 out of 266 compound-cell-line pairs when ranked by increasing rank product. MDA231: MDA-MB-231. MDA157: MDA-MB-157.

fluoresces when bound to DNA but does not penetrate intact cell membranes – did not provide robust or consistent data, as DRAQ7 positive apoptotic cells fluoresced only briefly before detaching from the bottom of the well and drifting out of the plane of focus (figure 5.3). Therefore cell count using the GFP labelled nuclei was instead used as the readout in the concentration response experiment.

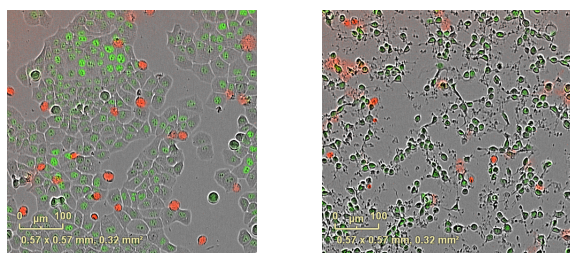


Figure 5.3 Representative cropped images from the incucyte. GFP-labelled (green) T47D cells with DRAQ7 apoptotic marker (red) and phase contrast image (grey). Whole images from which these are cropped measure 2.15 mm^2 . **(Left)** 0.1% DMSO negative control cells. **(Right)** $0.3 \mu\text{M}$ staurosporine positive control.

Using 8 semi-log concentrations ranging from 0.3 nM to $1 \mu\text{M}$ and GFP cell-count normalised to the DMSO negative control as a measure of cell-viability, concentration response curves were plotted for the 12 compounds and 8 cell-lines at the 72 hour time point (figure 5.4). Despite a selection criteria aiming to limit overtly cytotoxic compounds, 11 out of the 12 compounds demonstrated some form of concentration dependent reduction in cell-count in at least one of the cell-lines. Zalcitabine was an exception with very little reduction in cell count, although at $1 \mu\text{M}$ concentration there was some evidence of reduced cell-count in MDA-MB-157 and HCC1569 cell-lines. The HCC1569 cell-line demonstrated the greatest sensitivity to the majority of the tested compounds, especially to protriptyline which at $1 \mu\text{M}$ effectively killed all cells whilst largely unaffected the 7 other breast cancer cell-lines. Podophyllotoxin proved to be especially potent, with a relative cell-count below 50% at the lowest tested concentration of 0.3 nM in 5 of the cell-lines.

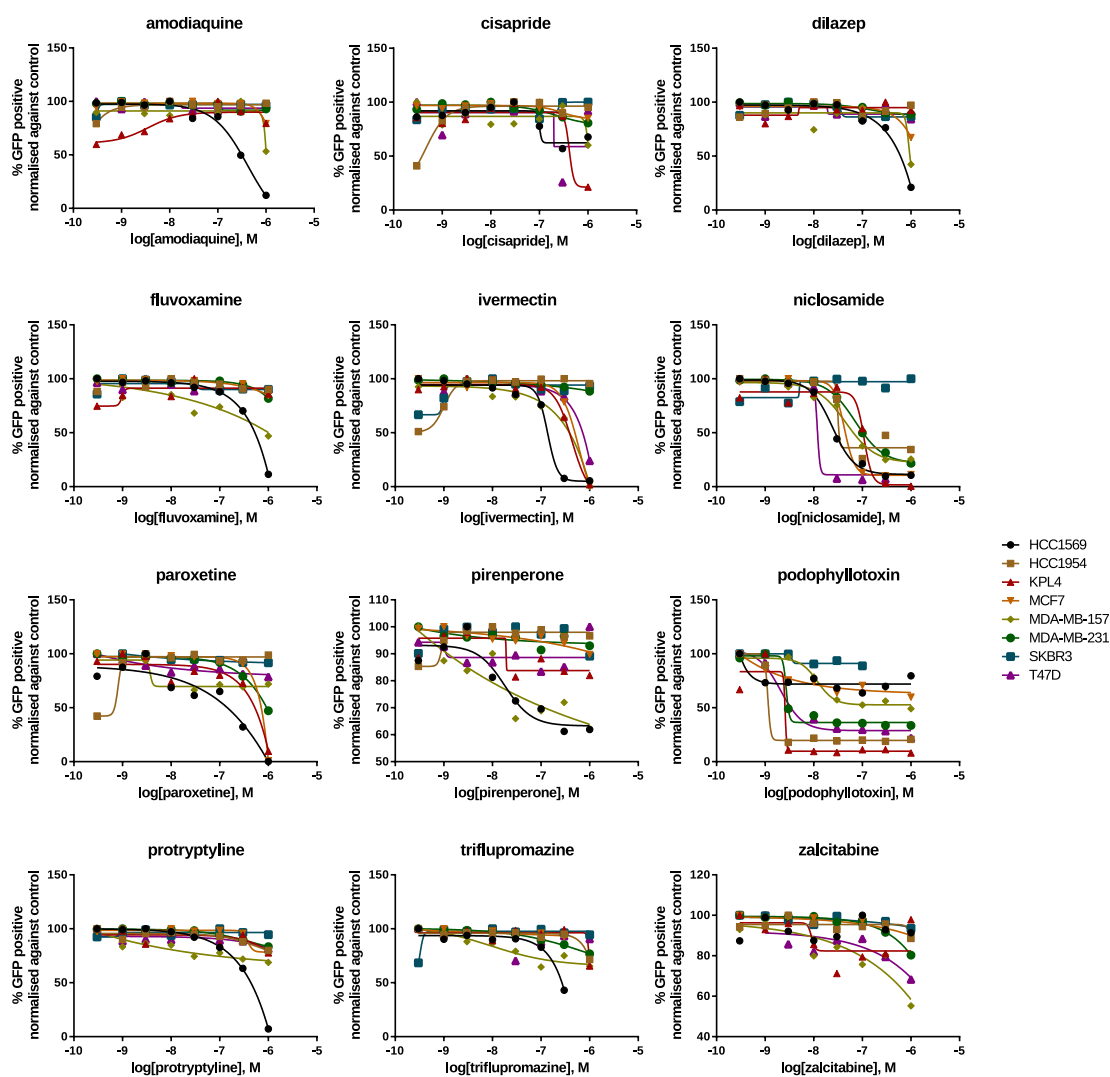


Figure 5.4: Concentration-response curves for 12 hits from the Prestwick library. Compounds were used in a 2D viability assay measuring cell count expressed as the percentage of the DMSO control after 72 hours.

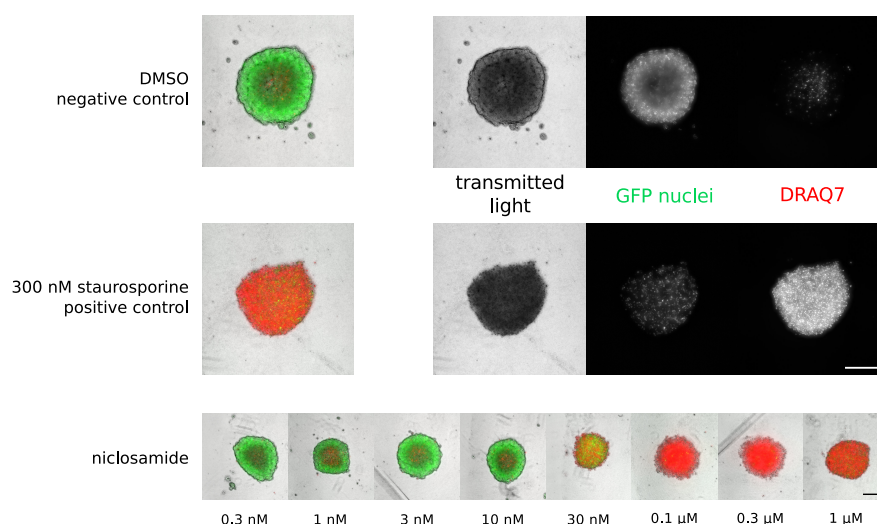


Figure 5.5: Example images of T47D tumour spheroids as imaged on the ImageXpress using transmitted light (grey), FITC (green) and CY5 (red) filters to visualise spheroid morphology. Scale bar = 300 μm

3D spheroid models

A more complex 3D tumour spheroid assay was used to determine the functional effects of the 12 selected compounds in a more physiologically relevant environment. All eight cell-lines were found to form consistent spheroids using the aggregation through centrifugation method and spheroid sizes with diameters around 500 μm . Mirroring the 2D assay, GFP labelled cell-lines were used alongside DRAQ7 to measure cell viability and cell death, although in the case of spheroids the DRAQ7 staining was more consistent as DRAQ7-positive apoptotic cells remained aggregated in the spheroid in the focal plane (figure 5.5). Spheroid area proved to be a poor readout for cell viability as cytotoxic treatment caused the spherical structure to collapse and disaggregate and so when imaged in 2D from above spheroid collapse results in an increase in measured spheroid area. When increasing doses of cytotoxic compound were tested on spheroid this resulted in a paradoxical increase in spheroid area, even in the GFP channel as some residual GFP staining remained despite cell-death. A measure of integrated intensity however proved a more intuitive readout of cell death within 3D spheroids and a comparison between the GFP and DRAQ7 channels revealed that the integrated intensity of GFP was more consistent and produced more robust concentration responses with staurosporine.

Concentration response studies in 3D tumour spheroids with the 12 hits from the Prestwick library at 0.3 nM to 1 μM in semi-log concentrations after 72 hours revealed a decrease in sensitivity to the compounds compared to results obtained in the 2D assay (figure 5.6). Of the compounds which produced a concentration dependent response in 3D, most only elicited a decrease in GFP integrated intensity at the maximum 1 μM concentration tested. The increased sensitivity of the HCC1569 cell-line compared to the others tested was not observed in 3D. Of the 12 compounds tested podophyllotoxin and niclosamide produced robust sigmoidal concentration response curves, although not in all of the cell-lines.

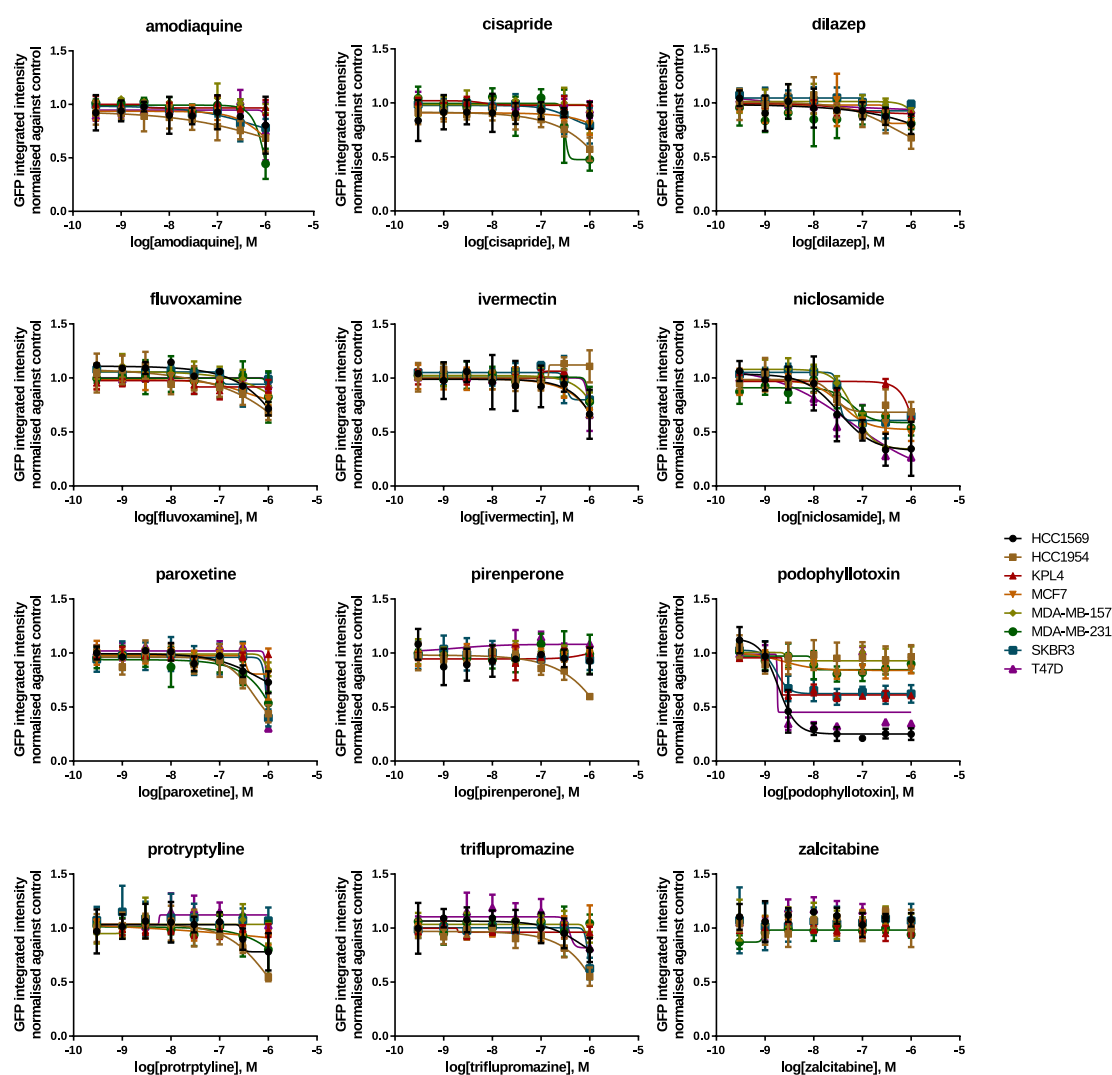


Figure 5.6: Concentration-response curves for 12 hits from the Prestwick library. Compounds were used in a 3D tumour spheroid viability assay measuring integrated intensity of GFP-labelled nuclei after 72 hours of compound treatment normalised against the negative control.

5.2.3 RPPA

Three compounds (ivermectin, protryptiline and niclosamide) were selected from the initial hit list based on their reproducibility and distinct response between cell-lines in both the morphological and viability assays. These compounds were used in a proteomic study, in which the 8 breast cancer cell-lines were grown in 2D or 3D culture conditions and treated with 100 nM of compound for 72 hours, after which cells were lysed and RPPA was used to measure the abundance of 60 proteins and phosphoproteins. Of the 60 proteins and phosphoproteins analysed 13 were discounted due to poor quality data such as low-signal to noise, poor spot morphology of samples printed on the RPPA chip and non-homogeneous or non-specific binding of antibodies to sample and/or chip – leaving 47 measurements per sample.

Cell-line and growth environment has a greater effect on protein expression than compound treatment

These results show that the 3 active compounds (ivermectin, niclosamide and protryptiline) selected for RPPA analysis produce similar pathway response within in each cell-line. However, compound induced pathway response diverge between distinct cell-lines and between 2D and 3D cell culture conditions. Using the readout from the 47 measured epitopes, the 64 samples displayed obvious clustering according to cell-line in hierarchical clustering and projecting the data into 2 dimensions (figure 5.7 A,B,C). Within the cell-line clusters there were distinct sub-clusters showing clear separation in the protein expression profiles of the two environmental conditions (figure 5.7 D). Interestingly, growth environment (2D versus 3D culture conditions) produced a more distinct change in protein expression profile than compound treatment at 100 nM (figure 5.7 E).

Resistant cell lines treated with niclosamide or ivermectin show decreased expression of E-cadherin

From the 2D concentration response studies it appeared that the cell-lines HCC1954 and SKBR3 both showed resistance to the anti-helminthic drugs niclosamide (figure 5.9 A) and ivermectin (figures ?? A). By aggregating the sensitive (HCC1569, KPL4, MCF7, MDA-MB-157, MDA-MB-231, T47D) and insensitive (HCC1954 and SKBR3) RPPA data, it was possible to look at changes in protein expression shared among the differentially responding cell-lines. Using data normalised to the DMSO control for each cell-line, and averaging the sensitive or insensitive cell-lines together revealed that both niclosamide and ivermectin treatment caused a distinct reduction of E-Cadherin in resistant cell-lines. Cyclin D1 was another protein which was upregulated in the resistant cell-lines in response to both anti-helminthic treatments, which was not mirrored in sensitive cell-lines (figure 5.9 % 5.10 C&D).

Protryptiline shows different resistant cell-lines in 2D and 3D

In the 2D concentration response assay HCC1569 demonstrated a singular sensitivity to the tri-cyclic antidepressant protryptiline (figure 5.11 A). However, repeating the concentration response assay in 3D tumour spheroids ameliorated this sensitivity, and instead another cell-line – HCC1954

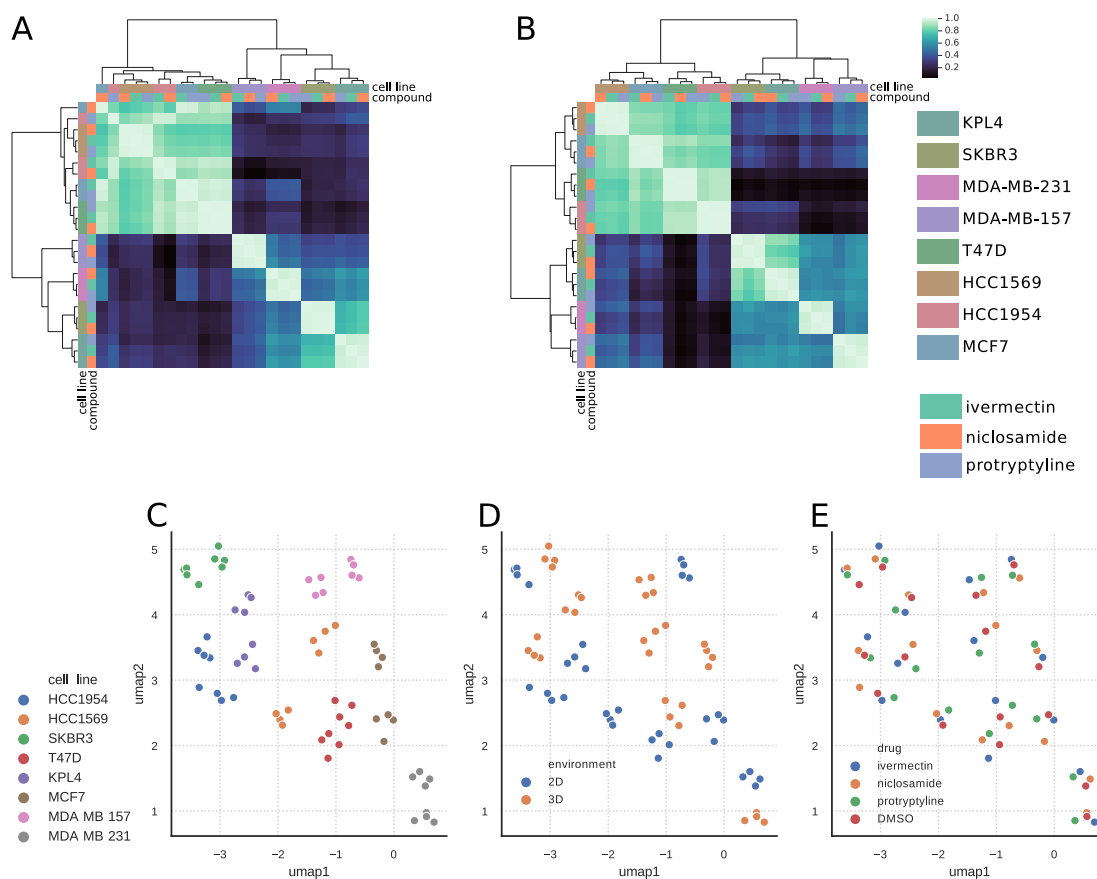


Figure 5.7: Globally normalised abundance of 64 samples, consisting of 8 cell-lines, 4 treatments and 2 growth conditions, measuring abundance of 47 proteins and phosphoproteins with RPPA. **(A)** Hierarchical clustering of protein samples from cells grown in 2D culture. **(B)** Hierarchical clustering of protein samples from cells grown in 3D spheroids. **(C)** Embedding of protein samples colour coded by cell-line. **(D)** Embedding of protein samples colour coded by environment conditions, either 2D culture of 3D tumour spheroids. **(E)** Embedding of protein samples colour coded by drug treatment.

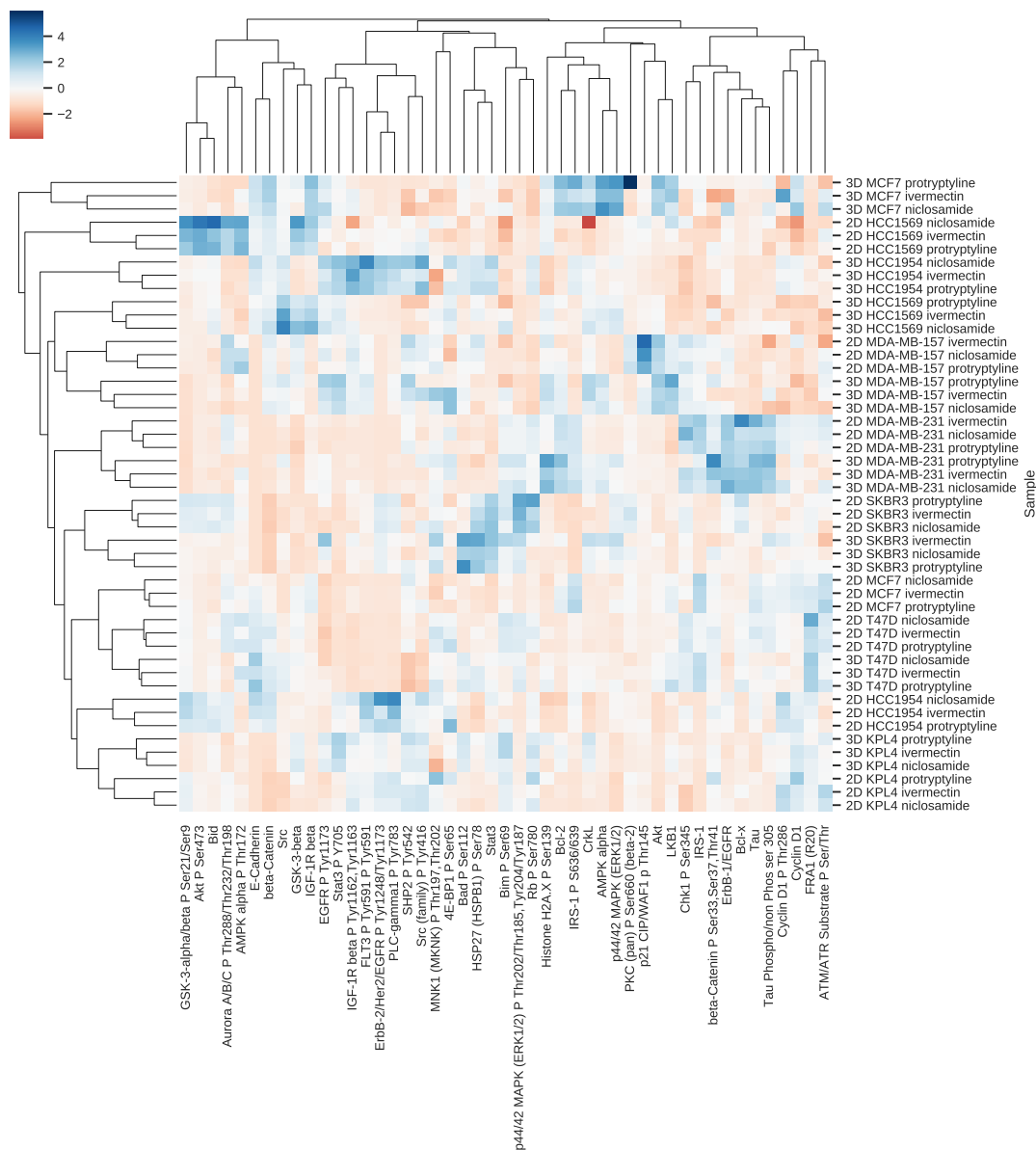


Figure 5.8: Heatmap of hierarchical clustering of globally normalised protein expression across cell-lines, growth environments and compound treatments.

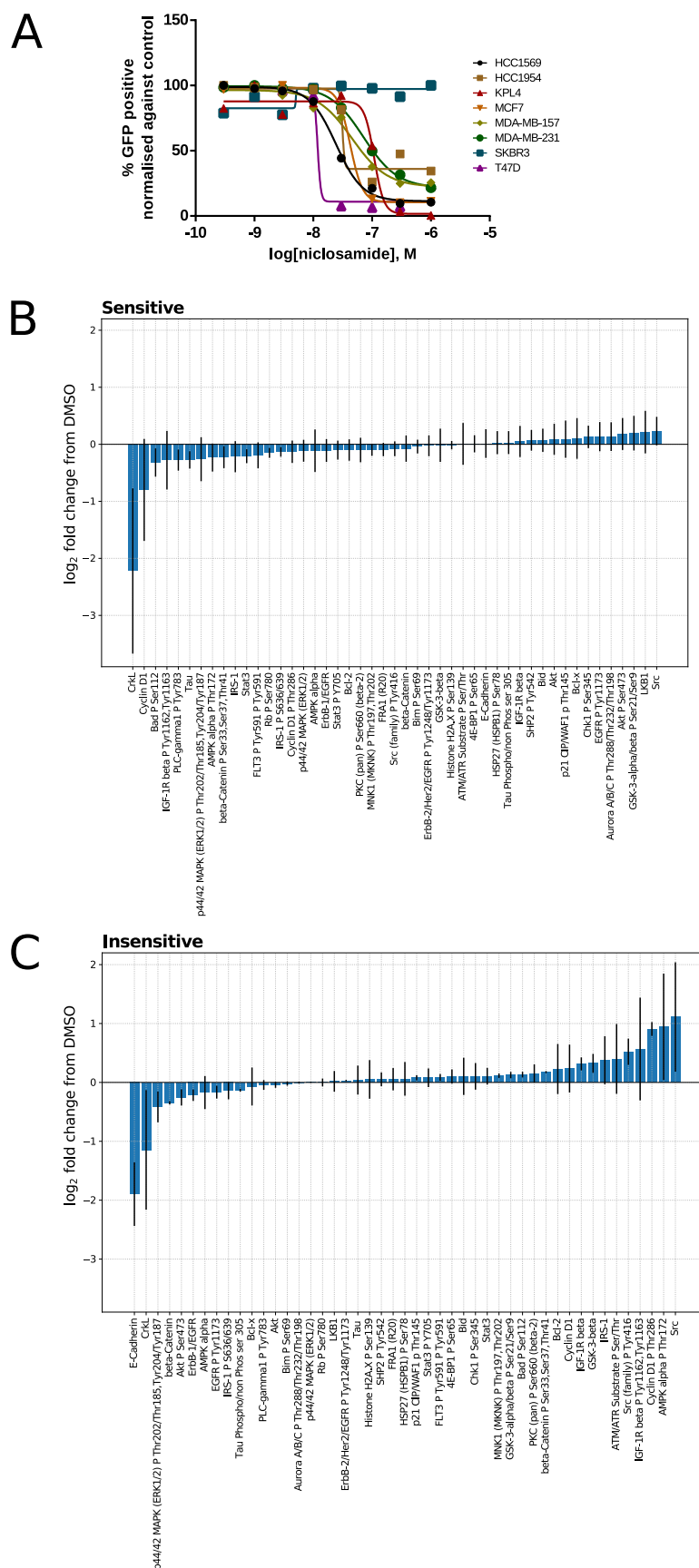


Figure 5.9: Insensitivity of HCC1954 and SKBR3 to niclosamide in 2D. **(A)** Concentration response curve of normalised cell-count of 8 cell-lines treated with niclosamide (as shown in figure 5.4). **(B&C)** Mean change in protein abundance of cells grown in 2D treated with 100 nM drug compared to DMSO treated cells, averaged over multiple cell-lines. Y-axis indicates \log_2 fold change from DMSO treated cells, error bars indicate \pm standard deviation.

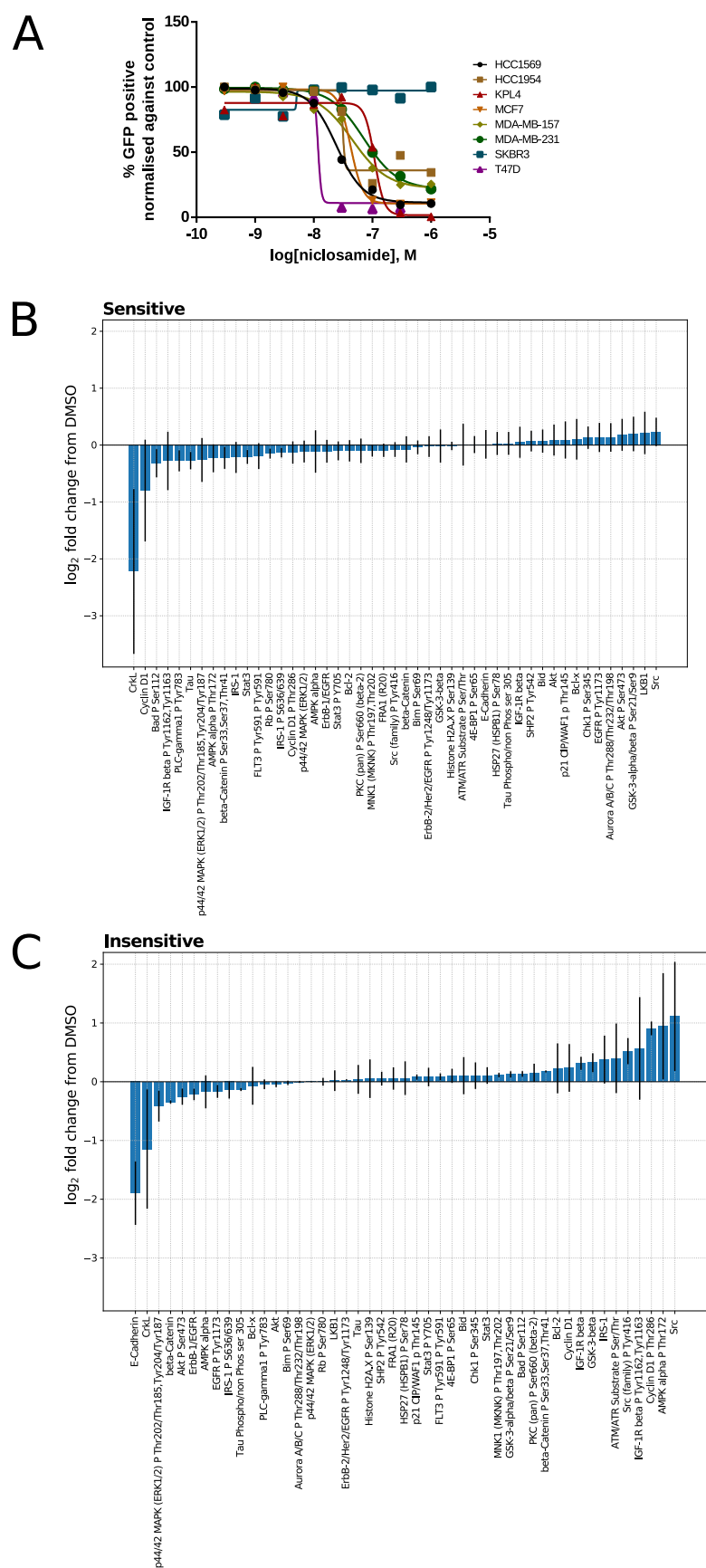


Figure 5.10: Insensitivity of HCC1954 and SKBR3 to ivermectin in 2D. **(A)** Concentration response curve of normalised cell-count of 8 cell-lines treated with ivermectin (as shown in figure 5.4). **(B&C)** Mean change in protein abundance of cells grown in 2D treated with 100 nM drug compared to DMSO treated cells, averaged over multiple cell-lines. Y-axis indicates \log_2 fold change from DMSO treated cells, error bars indicate \pm standard deviation.

– showed a slight response to protryptiline which was not seen in the 7 other breast cancer cell-lines (figure 5.12 A). RPPA data indicated that in 2D the sensitive HCC1569 cell-line had decreased levels of IGF- β and MAPK phosphorylation compared to the resistant cell-lines (figure 5.11 B & C), while in 3D the sensitive HCC1954 cell-line had a similar decrease in IGF- β although a contradictory 2-fold increase in phosphorylated MAPK (figure 5.12 B).

Cells cultured in tumour spheroids have increased Src and decreased Aurora kinase phosphorylation

Cells grown in 2D on tissue culture plastic are subjected to different environmental stimuli to those grown in 3D, with cell-to-cell adhesions, presence of extracellular matrix and differences in oxygen and nutrient supply all effecting intracellular signalling and therefore protein expression and response to external stimuli. Using the RPPA data from the negative control treated samples grown in 2D and 3D revealed a number of proteins which had altered expression dependent on the environmental conditions. In 3D conditions the phosphorylated form of Aurora A/B/C had a two-fold decrease over cells grown in 2D, while Src kinase and AMPK alpha showed a 2 and 4-fold increase respectively in cells grown in 3D compared to 2D (figure 5.13). The decrease in phosphorylated aurora kinase in 3D spheroids may be indicative of the cell-cycle arrest commonly seen in cells located towards the centre of the spheroid.⁷⁷ For certain proteins such as AMPK alpha there is a large variation between the cell-lines and as there is only a single sample per condition it is not possible to determine if this is an inherent difference in the response to environmental conditions between the cell-lines or simply noise within the data. The elevation of Src kinase in 3D spheroid cultures (figure 5.13) is potentially interesting in relation to the general decrease in drug sensitivity of hit compounds observed in 3D spheroid and 2D assays (figures 5.4 & 5.6).

5.3 Discussion

A number of approved compounds were identified with a high-content screen which resulted in distinct phenotypic responses between breast cancer cell-lines. Following validation in 2D and 3D models of cell proliferation and survival, two anti-helminthic compounds (ivermectin and niclosamide) and an antidepressant (protryptiline) were selected for further investigation of their anti-cancer MoA. Proteomic analysis of cells grown in 2D and 3D tumour spheroids treated with these compounds revealed that both anti-helminthic treatments caused a reduction in E-cadherin levels in resistant cell-lines which was not observed in the sensitive cell-lines. Reduction in E-cadherin expression levels has previously been associated with epithelial-mesenchymal transition during tumour progression and the onset of more cancer stem-cell like phenotypic associated with drug resistance.^{78,79} Cyclin D1 was another protein upregulated in the resistant cell-lines in response to both anti-helminthic treatments. Several studies have implicated elevated Cyclin D1 with drug resistance in part through a dual role in promoting cell proliferation and inhibition of drug-induced apoptosis.⁸⁰ Previous transcriptomic studies have also shown that Cyclin D1 over-expression in tumours alters the expression of genes controlling cell metabolism and disrupt REDOX balance by producing reactive oxygen species and oxidative stress signalling pathways which influence drug

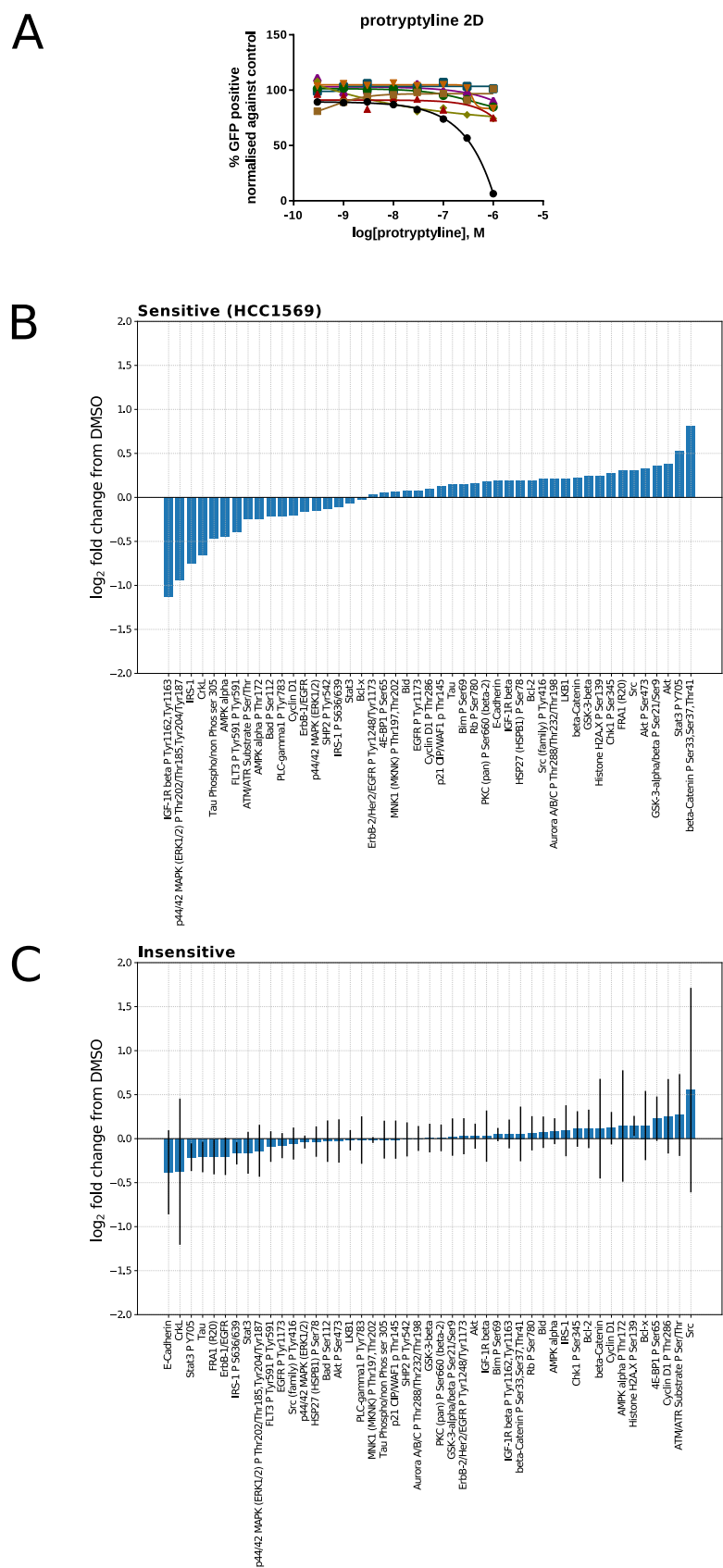


Figure 5.11: Sensitivities of cell-lines in 2D assays of cell viability, and corresponding changes in protein levels of sensitive and resistant groups. **(A)** Concentration response curve of integrated intensity of GFP expressing cell-lines treated with protryptiline in 2D cell-culture. **(B-C)** Mean changes in protein abundance of cells treated with 100 nM protryptiline. Y-axis indicates \log_2 fold-change from DMSO treated cells, error bars indicate \pm standard deviation. **(B)** RPPA data of the sensitive HCC1569 cell-line grown in 2D cell culture treated with protryptiline. **(C)** RPPA data of the 7 resistant cell-lines grown in 2D cell-culture treated with protryptiline.

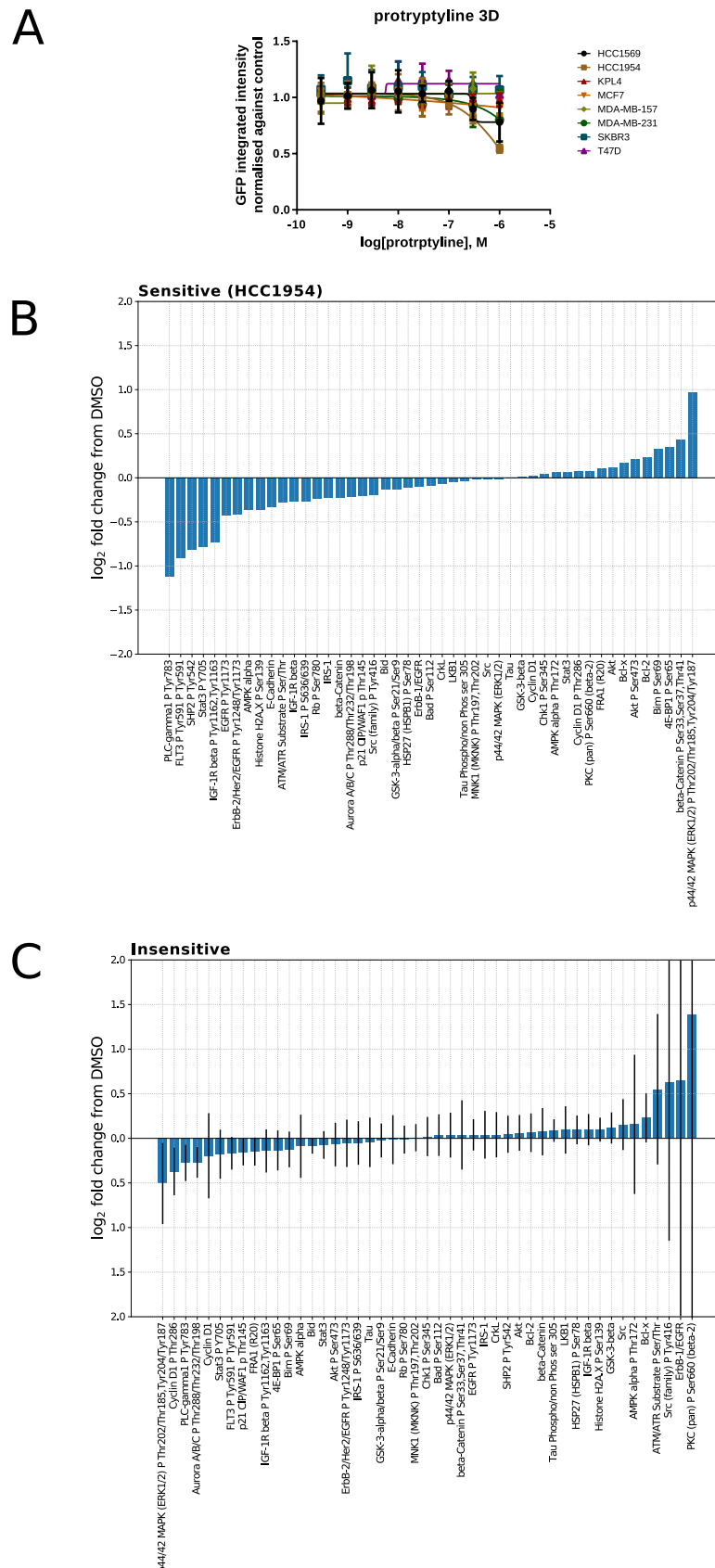


Figure 5.12: Sensitivities of cell-lines in 3D assays of cell viability, and corresponding changes in protein levels of sensitive and resistant groups. **(A)** Concentration response curve of integrated intensity of GFP expressing cell-lines treated with protryptiline in 3D tumour spheroids. **(B-C)** Mean changes in protein abundance of cells treated with 100 nM protryptiline. Y-axis indicates \log_2 fold-change from DMSO treated cells, error bars indicate \pm standard deviation. **(B)** RPPA data of the sensitive HCC1954 cell-line grown in 3D spheroids treated with protryptiline. **(C)** RPPA data of the 7 resistant cell-lines grown in 3D spheroids treated with protryptiline.

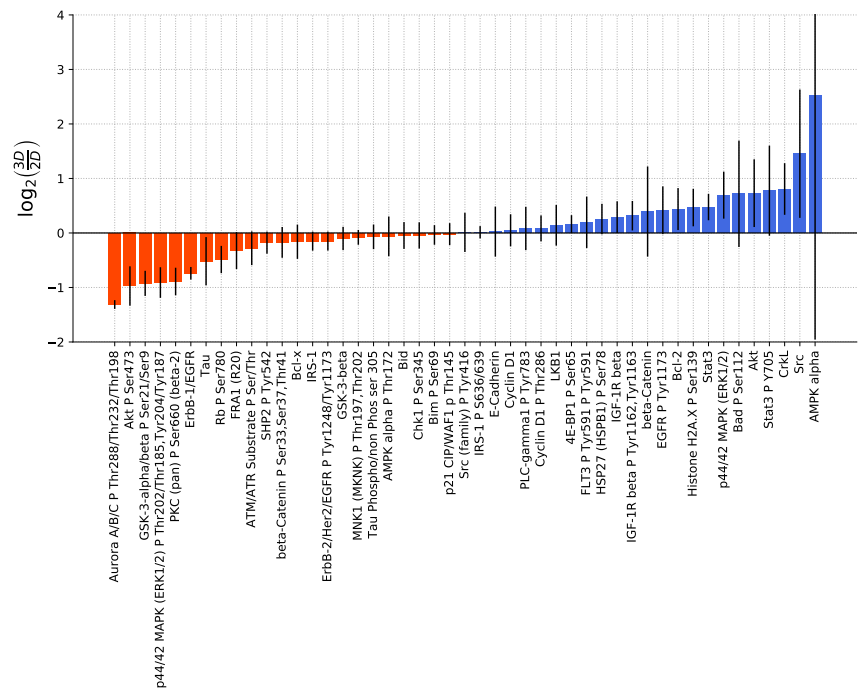


Figure 5.13: Comparison of protein expression between cells grown in 3D or 2D environments. The median difference from combined data from 8 cell-lines treated with 0.1 % DMSO. Difference is represented as the \log_2 fold change of 3D expression divided by 2D expression. Proteins with increased expression in 3D shown in blue, those with decreased expression in 3D relative to 2D shown in orange/red. Error bars indicate \pm median absolute deviation of the 8 cell-lines.

sensitivity.⁸¹ It was also found that culturing tumour cells in complex 3D environments resulted in a number of changes in protein levels of cell-cycle regulators, which is in agreement with existing studies which have found reduced proliferation and cell-cycle arrest of many cells within the core of a spheroid.⁷⁷ The reduced proliferation of cells when cultured in tumour spheroids may also explain the reduced sensitivity of cell-lines in response to all test compounds when compared to those grown in 2D (figure 5.4 and 5.6), as many cytotoxic compounds act through cell-cycle checkpoints or microtubule dynamics during cell-division. In addition, RPPA analysis further revealed elevated levels of the non-receptor tyrosine kinase Src in 3D spheroid cultures which is associated with cancer cell survival signalling and a common pathway of drug resistance in breast cancer and other tumour types.⁸²

The approved compounds identified in this work which produce distinct morphological effects between different cell-lines may be a consequence of altered signalling pathways, differences in expression of target receptors or one of many other biologically interesting possibilities. However, a simple explanation would be that the morphological differences observed are actually differences in sensitivity caused by multi-drug resistance efflux pumps. Elevated expression of ATP binding cassette (ABC) drug efflux transporters are found in many types of cancer, and are suggested to contribute towards chemo-resistance in breast cancer.⁸³ Considering certain cell-lines such as HCC1569 were commonly more sensitive to a number of the tested compounds (figure 5.4), it would be worthwhile to determine if expression of multi-drug efflux pumps correlated with cell-line sensitivity and may explain a number of the distinct responses observed.

The majority of the hits found in this phenotypic screen (table 5.3) have also been proposed as potential candidates for repurposing. Amodiaquine originally developed as a selective anti-malarial treatment has been reported to have anti-adipogenic properties.⁸⁴ Multiple studies have proposed repurposing the anti-helminthic drugs ivermectin and niclosamide as potential anti-cancer treatments,^{85,86} cisapride as a treatment for Chagas disease,⁸⁷ dilazep to aid HIV treatment,⁸⁸ fluvoxamine as an inhibitor of glioblastoma invasion,⁸⁹ protryptiline as a treatment for osteosarcoma,⁹⁰ paroxetine as a neuroprotective agent^{91,92} and podophyllotoxin and triflupromazine as anti-cancer treatments.^{86,93} A potential issue with the Prestwick chemical library screen performed in this study is the concentration at which the compounds are screened at. The choice of which concentration to screen a compound library at is an open problem in the field, typically compound libraries consist of lead-like molecules which have not been optimised for potency – which is not the case for many of the highly potent compounds found in the approved Prestwick library. It may be a possibility that screening at 1 μ M and discarding compounds which caused considerable toxicity has removed many potent and selective hits from the analysis, and screening at a lower concentration may have yielded a considerably different selection of hit compounds. Similarly, performing RPPA analysis of pathway effects across concentration-response and time-series studies for each compound may reveal further insights into the pathways underpinning the anti-cancer activity than has been revealed by the single (100 nM) experiment performed in this project.

The functional assays used in this work are using cell-count (and integrated intensity of nuclei staining in the case of 3D models) as a surrogate measurement for cytotoxicity or cell-proliferation. This functional readout is not ideal as the compounds chosen from the high-content screen were

selected based preferentially on morphology measurements and their limited cytotoxicity. The originally proposed functional assay was to test compound effect on the cell's ability to migrate through an extracellular matrix. After trialling both 2D scratch wound assays through collagen and Matrigel substrates as well as 3D cell invasion assay from tumour spheroids embedded in extracellular matrix, I found that only the MDA-MB-231 cell-line was capable of migrating through collagen or Matrigel. So while it would have been possible to confirm hits from the high-content screen in a single cell-line, it would not provide a comparison of functional activity between the cell-lines.

In summary the use of more complex cellular models to follow up hits resulting from a multiparametric high-content screen does offer an opportunity to gain an understanding of the functional affects of altered cellular morphology, however the choice of functional assay needs to be carefully considered to ensure it is relevant and robust. Ideally a morphological change in simple 2D assay that is predictive of a functional response in a more complex and disease relevant 3D cell model offers an opportunity to combine large chemical libraries with more predictive and biologically relevant assays without the considerable cost burden associated with screening complex cell models at scale.

5.4 Methods

5.4.1 Imaging and image analysis

Cells were stained following the cell painting protocol, imaged with the ImageXpress and images were analysed with cellprofiler as previously described in the general methods (chapter 2).

5.4.2 Compound library

The compound library used was the Prestwick chemical library of 1280 off-patent small molecules, 95% of which are approved drugs (FDA, EMA or others) stored as 10 mM stocks in DMSO. For screening the library was assayed at a 1 μ M final concentration.

5.4.3 Multivariate Z-factor to determine assay quality

A multivariate Z-factor as defined by Kümmel *et al.*⁷⁶ is a multi-variate adaptation of the original Z-factor,⁹⁴ which is a measure of assay robustness in high-throughput screening. The original measure is univariate, defined as:

$$\text{Z-factor} = 1 - \frac{3(\sigma_p + \sigma_n)}{|\mu_p - \mu_n|} \quad (5.1)$$

where σ_p and σ_n are the standard deviations of the positive and negative control, and μ_p and μ_n the means of the positive and negative control. A Z-factor of greater than 0.5 shows a very clear separation of positive and negative control, and is interpreted as an ideal assay. The adaption of Kümmel *et al.* uses linear discriminant analysis to find a combination of features to best separate the positive and negative controls, and calculates the Z-factor on the first linear discriminant.

5.4.4 Identifying hits

Hits were identified as those compounds which caused distinct phenotypic responses between cell-lines, excluding compounds which demonstrated significant toxicity. This was first carried out by screening the entire 1280 compound library across the panel of eight cell-lines at 1 μM concentration in 384-well optical bottomed plates (see chapter 2). Following data pre-processing, distinct phenotypic responses between active compounds were calculated using the TCCS method (chapter 4). An initial hit list was created by ranking compounds and cell-line pairs by decreasing $\Delta\theta$. Compounds were triaged by removing those with less interesting mechanistic properties such as microtubule disruptors leaving 14 hits. From these 14 hits, 2 were not easily available due to lack of a commercial supplier (pinaverium bromide) or being a controlled substance (3,4-dimethoxyphenethylamine). The remaining 10 compounds were re-screened in triplicate at 8 semi-log concentrations ranging from 0.3 nM to 1 μM across the eight-cell lines to confirm a concentration-response relationship using the l_1 norm from the negative control as a measure of compound response. Of those compounds that validated with a robust concentration dependent response $\Delta\theta$ values were calculated between all pairs of cell-line for each replicate dataset, and ranked by order of decreasing $\Delta\theta$, so that compounds-cell-line-pairs with a more distinct phenotypic response received a lower rank. A rank product⁹⁵ was calculated from the replicates and compound-cell-line-pairs were sorted by increasing rank product. Compounds that demonstrated repeatability and significantly low rank-products were carried onto more complex 2D and 3D apoptosis assays.

Rank product

A rank product algorithm with permutation-based significance testing was implemented in python. Given a dataset of k replicates and n ranks, an n by k matrix with each row representing ranks from $[1..n]$. For $[1..p]$ where p is the number of permutations, the ranks in each row are shuffled and the geometric mean calculated for each column. This yields a p by n matrix of permuted rank products which are used to count how many times the observed rank products from the replicated data is smaller than or equal to the permuted rank products, giving a value c . The averaged expected value E is then calculated as $E = p/c$ which is then used to calculate the percentage of false positives as E divided by the rank of compound-cell-line-pair ordered by increasing rank-product value.

5.4.5 2D apoptosis assay

GFP-labelled cell-lines were seeded into the inner 60 wells of a flat-bottomed tissue culture treated 96-well plate (#655180 Greiner), with approximately 10,000 cells per 90 μL of DMEM media, with the addition of 10 μL of 10% DRAQ7 apoptotic marker (#DR710HC biostatus), for a final DRAQ7 concentration of 3 μM . Assay plates were then incubated at 37°C, 5% CO_2 incubator for 24 hours before addition of compounds. Compounds were diluted in DMSO at 1000X concentration and assay plates were treated with the use of an intermediate plate as described in chapter 2. Compound concentrations ranged from 0.3 nM to 1 μM with 0.1 % DMSO as a negative control and 300 nM staurosporine as a positive control. After compound treatments assay plates

were then incubated for an additional 72 hours in the presence of compounds imaging with the Incucyte ZOOM, imaging 3 sites per well in phase, red, and green channels every 3 hours. Using the Incucyte ZOOM software, cells were counted in both the red and green channels for each site and timepoint and exported as csv files. Data was merged with compound name and concentration data and filtered to select just the 72 hour time point. Using the cell-count from the GFP channel, the count for each well then was expressed as a percentage of the median DMSO values per plate. Concentration response-curves were fitted in GraphPad Prism (version 5) using a 4-parameter non-linear curve fit with least-squares.

5.4.6 Spheroids

Creating spheroids

Spheroids were created by seeding approximately 10,000 GFP-expressing cells per well in 50 μ L of media into each well of a 96-well ultra low attachment U-bottomed plate (#7007 Corning). A solution containing 4% growth-factor reduced Matrigel (#35623 Corning) and 2% DRAQ7 apoptotic stain (#DR710HC biostatus) was made in cold media, and 50 μ L per well was added to the existing cell suspension, for a final Matrigel concentration of 2% and 1% DRAQ7. Plates were then centrifuged for 10 minutes at 1000X G and 4° with brake speed reduced to pellet down the cells in the centre of each well. After centrifugation plates were placed in a tissue culture incubator for 24 hours before addition of compounds. Compounds from a 1000x source plate were diluted 1:50 by transferring 3 μ L from the source plate to an intermediate plate containing 150 μ L of media. From the intermediate plate 5 μ L were transferred to the spheroid assay plate containing 100 μ L for a final dilution of 1:1000 and a DMSO concentration of 0.1%. Following compound addition spheroid plates were incubated for an additional 72 hours.

Imaging spheroids

Spheroids were imaged on the ImageXpress using the 4X objective lens in 3 channels (transmitted light, GFP and CY5). Images were captured by first detecting the well-bottom in the centre of the U-bottomed well with a laser-based autofocus and offsetting by the well thickness, then capturing images in a z-stack at 8 focal planes spaced at 50 μ m intervals for a total range of 350 μ m. Z-stacks of the GFP and CY5 fluorescent channels were collapsed into a single image per channel using a maximum intensity projection, while the z-stack of transmitted light images were transformed using a minimum intensity projection.

5.4.7 RPPA

Protein extraction

2D cells. Protein extraction from 2D cells was performed by first seeding approximately 50,000 cells per well of a 6-well plates in 3 mL of media followed by incubation in a tissue culture incubator for 24 hours. Compound addition was performed by diluting compound stocks in DMSO 1:50 in media to an intermediate plate, followed by 1:20 from the intermediate plate to the assay plate

for a 1000-fold dilution and 0.1% DMSO. Assay plates were then incubated for an additional 72 hours, after which wells were washed with 1 mL of room temperature PBS followed by addition of 100 μ L of room temperature CLB1 (Zeptosens, Bayer) lysis buffer. Cells and lysis buffer were then scraped into 1.5 mL eppendorf tube and incubated at room temperature for 30 minutes with frequent vortexing. After 30 minutes of incubation lysis solution was centrifuged for 10 minutes at 13,000X G at room temperature and the supernatant was transferred into new 1.5 mL eppendorf tubes.

Spheroids. Protein extraction from spheroids was performed by first growing spheroids in 96-well plates following the same protocol as for imaging. 20 spheroids per treatment group were extracted with a pipette into a 1.5 mL eppendorf tube. Pipette tips were widened by cutting with scissors. The spheroids were then centrifuged for 30 seconds at 13,000X G at room temperature to pellet at the bottom of the tube, media was removed with a pipette and replaced with room temperature PBS. Spheroids were pelleted again, PBS removed and replaced with 75 μ L of room temperature CLB1 lysis buffer. The spheroid lysis buffer mixture was incubated at room temperature for 30 minutes with frequent vortexing to break up cell aggregates. Following incubation the lysis solution was centrifuged for 10 minutes at 13,000X G at room temperature, and supernatant extracted into a new 1.5 mL eppendorf tube.

Determining protein concentration. Protein concentration was determined with a Bradford assay, using a standard curve of known BSA concentrations and the addition of CLB1 lysis buffer to control for the lysis buffer concentration of the samples. A curve of known BSA concentrations was created using 2 mg/mL BSA protein standard (#23209 Thermo Scientific) diluted in PBS, with a 1:20 concentration of lysis buffer (see table 5.5). Samples were diluted 1:20 in PBS by adding 2.5 μ L of sample to 47.5 μ L of PBS and mixed with a vortex. 10 μ L of diluted samples and standard were added to each well of a flat-bottomed 96-well plate, followed by 240 μ L of room temperature Coomassie Plus Protein Assay (#1856210 Thermo Scientific) and incubated at room temperature for 10 minutes. Plates were then read with a microplate reader (BIORAD iMark) at a wavelength of 595 nm. The protein concentrations of samples were calculated from a linear model of the BSA standard curve. All protein samples were normalised to 1 mg/mL by dilution in CLB1 lysis buffer.

BSA final concentration (mg/mL)	BSA 2 mg/mL (μ L)	PBS (μ L)	Lysis Buffer (μ L)
0	0	95	5
0.05	2.5	92.5	5
0.1	5	90	5
0.15	7.5	87.5	5
0.2	10	85	5
0.3	15	80	5
0.4	20	75	5
0.6	30	65	5

Table 5.5: Volumes for the BSA standard curve. Lysis buffer was CLB1, the same as used for the sample preparation.

Zeptosens RPPA platform

The RPPA study was performed on a Zeptosens platform by the Protein and Antibody Microarray facility at the Edinburgh Cancer Research UK Centre.ⁱ

With the concentration-normalised protein lysates a final 4-fold concentration series of; 0.2; 0.15; 0.1 and 0.75 mg/mL in spotting buffer CSBL1 (Zeptosens-Bayer) was created. The diluted concentration series of each sample was printed onto hydrophobic Zeptosens protein microarray chips (ZeptoChip™, Zeptosens-Bayer) under environmentally controlled conditions (constant 50% humidity and 14°C temperature) using a non-contact printer (Nanoplotter 2.1e, GeSiM). A single 400 pL droplet of each lysate concentration was deposited onto the Zeptosens chip. A reference grid of Alexa Fluor 647 conjugated BSA was spotted onto each sub-array, each sample concentration series was spotted in between reference columns. After array printing, the arrays were blocked with an aerosol of BSA solution using a custom designed nebuliser device (Zepto-FOG™, Zeptosens-Bayer) for 1.5 h to prevent non-specific antibody binding. The protein array chips were subsequently washed in double deionised water and dried prior to performing a dual antibody immunoassay comprising of a 16 h incubation of primary antibodies (table 5.6) followed by 2.5 h incubation with secondary Alexa Fluor 647 conjugated antibody detection reagent (anti-rabbit or anti-mouse 647 Fab, Invitrogen). Following secondary antibody incubation and a final wash step in BSA solution, the immunostained arrays were imaged using the ZeptoREADER instrument (Zeptosens-Bayer). For each-sub-array, five separate images were acquired using different exposure times ranging from 0.5-10 s. Microarray images representing the longest exposure without saturation of fluorescent signal detection were automatically selected for analysis using the ZeptoView™ 3.1 software. A weighted linear fit through the 4-fold concentration series was used to calculate the relative fluorescence intensity value for each sample replicate. Local normalisation of sample signal to the reference BSA grid was used to compensate for any intra- or inter-array/chip variation. Global normalisation was performed using Tukey's median polish.

Hierarchical clustering of globally normalised RPPA data

Figure 5.7 Globally normalised RPPA data was subset into two separate datasets consisting of either samples grown in 2D or 3D, and negative control samples were removed. A correlation distance matrix was calculated between rows (samples) of the two datasets, and used to calculate a hierarchical clustering of the data using “`scipy.cluster.hierarchy.linkage`” with average linkage and euclidean distance.

Figure 5.8 Globally normalised RPPA data was used in the form of a matrix with rows as samples and columns as proteins. A heatmap was created with “`seaborn.clustermap`” using a Euclidean distance metric, and z-scoring the columns (antibodies). Distance between clusters for hierarchical clustering was calculated with the average linkage method in `scipy`.

ⁱA thank you to Alison Munro (University of Edinburgh) for running the 64 samples on the Zeptosens RPPA platform.

Antibody	Supplier	Cat. #	Type	Pathway, Function
ATM/ATR Substrate P Ser/Thr	CST	2851	rabbit	Cell Cycle Control, DNA Repair
Aurora A/B/C P Thr288/Thr232/Thr198	CST	2914	rabbit	Cell Cycle
Bad P Ser112	CST	9291	rabbit	Apoptosis, Akt Signaling
Crkl P Tyr207	CST	3181	rabbit	Adaptor Proteins
FLT3 P Tyr591 P Tyr591	CST	3461	rabbit	Receptors, Tyrosine Kinases, Cytokine Receptor
HSP27 (HSPB1) P Ser78	CST	2405	rabbit	Chaperones, MAPK Signaling, Stress pathway
MNK1 (MKNK) P Thr197, Thr202	CST	2111	rabbit	MAPK Signaling, Translational Control
PKC (pan) P Ser660 (beta-2)	CST	9371	rabbit	Calcium, cAMP, Lipid Signaling, PKC Signaling
IGF-1R beta P Tyr1162, Tyr1163	Invitrogen	44-804G	rabbit	Metabolism, Receptors, Tyrosine Kinases
ErbB-1/EGFR	CST	2232	rabbit	Akt & MAPK Signaling, Receptors, Tyrosine Kinases
ErbB-2/Her2/EGFR P Tyr1248/Tyr1173	CST	2244	rabbit	Akt & MAPK Signaling, Receptors, Tyrosine Kinases
EGFR P Tyr1173	CST	4407	rabbit	Akt & MAPK Signaling, Receptors, Tyrosine Kinases
p44/42 MAPK (ERK1/2)	CST	9102	rabbit	MAPK Signaling
p44/42 MAPK (ERK1/2) P Thr202/Thr185...	CST	4370	rabbit	MAPK Signaling
Src	CST	2109	rabbit	ErbB Signaling, VEGF Signaling, Adhesion
Akt	CST	9272	rabbit	Akt Signaling, Metabolism
Akt P Ser473	CST	4060	rabbit	Akt Signaling, Metabolism
Chk1 P Ser345	CST	2348	rabbit	Cell Cycle Control
c-Myc	CST	5605	rabbit	MAPK Signaling, Transcription Factors
E-Cadherin	CST	3195	rabbit	Adhesion
Rb	Abcam/Epitomics	ab113074	rabbit	Apoptosis, Cell Cycle Control
4E-BP1 P Ser65	CST	9451	rabbit	Metabolism, Translational Control, mTOR signal...
beta-Catenin	CST	9562	rabbit	Wnt Signaling
beta-Catenin P Ser33, Ser37, Thr41	CST	9561	rabbit	Wnt Signaling
Cyclin D1	CST	2926	mouseIgG2a	Cell Cycle Control
LKB1	CST	3047	rabbit	mTOR Signaling
GSK-3-alpha/beta P Ser21/Ser9	CST	9331	rabbit	Akt Signaling, Metabolism, Wnt Signaling, Hedge...
p53 P Ser15	CST	9284	rabbit	Apoptosis, Cell Cycle Control
p21 CIP/WAF1	CST	2946	mouseIgG2a	Cell Cycle Control
PLC-gamma1 P Tyr783	CST	2821	rabbit	Calcium, cAMP, Lipid Signaling
c-Myc P Thr58, Ser62	Epitomics	1203-1	rabbit	MAPK Signaling, Transcription Factors
Rb P Ser780	CST	9307	rabbit	Apoptosis, Cell Cycle Control
Src (family) P Tyr416	CST	2101	rabbit	ErbB Signaling, VEGF Signaling, Adhesion
Smad2/3 P Ser465/Ser423, Ser467/Ser425	CST	8828	rabbit	cell growth, apoptosis, morphogenesis, develop...
Smad1/5 P Ser463/Ser465	CST	9516	rabbit	cell growth, apoptosis, morphogenesis, develop...
Cyclin D1 P Thr286	CST	3300	rabbit	Cell Cycle Control
AMPK alpha	CST	2532	rabbit	Metabolism
AMPK alpha P Thr172	CST	2535	rabbit	Metabolism
Bcl-2	Epitomics	1017-1	rabbit	Apoptosis
Bid	Abcam/Epitomics	ab32060	rabbit	Apoptosis
Bim P Ser69	CST	4585	rabbit	Apoptosis
p53	CST	9282	rabbit	Apoptosis, Cell Cycle Control
IRS-1	CST	2382	rabbit	Metabolism, Insulin Signaling
GSK-3-beta	CST	9315	rabbit	Akt Signaling, Metabolism, Wnt Signaling, Hedge...
Crkl	CST	3182	mouseIgG1	Adaptor Proteins
HSP27 (HSPB1)	CST	2402	mouseIgG1	Chaperones, MAPK Signaling, Stress pathway
PKC-alpha	Beckton Dickinson	610108	mouseIgG2b	Calcium, cAMP, Lipid Signaling, PKC Signaling
IRS-1 P S636/639	CST	2388	rabbit	Metabolism, Insulin Signaling
PLC-gamma1	CST	2822	rabbit	Calcium, cAMP, Lipid Signaling, PKC Signaling
SHP2 P Tyr542	CST	3751	rabbit	Tyrosine Phosphatases
Tau	Abcam/Epitomics	ab32057	rabbit	Neuroscience, Alzheimer
Stat3	CST	12640	rabbit	Cytokine Signaling, Jak/Stat Signaling
Tau Phospho/non Phos ser 305	Epitomics	2368-1	rabbit	Neuroscience
IGF-1R beta	CST	3027	rabbit	Insulin Signaling, Metabolism, Receptors, Tyro...
Akt P Ser473	CST	9271	rabbit	Akt Signaling, Metabolism
Stat3 P Y705	CST	9131	rabbit	Cytokine Signaling, Jak/Stat Signaling
Histone H2A.X P Ser139	Millipore (Upstate)	05-636	mouseIgG1	cell cycle, DNA Damage repair
beta-Tubulin	Abcam	ab6046	rabbit	Housekeeping, Cytoskeleton
p21 CIP/WAF1 p Thr145	Santa Cruz	sc-20220-R	rabbit	Cell Cycle Control
FRA1 (R20)	Santa Cruz	sc-605	rabbit	Transcription Factors

Table 5.6: Antibodies used in the RPPA study. CST: Cell Signaling Technologies.

Two-dimensional projections of globally normalised RPPA data

To embed proteomic data into two dimensions in order to visualise local structure within the data, each column relating to measurements of a single (phospho)protein was standardised to a mean of zero and unit variance and used as input to the Uniform Manifold Approximation and Projection (UMAP) algorithm in python.ⁱⁱ UMAP projects high-dimensional datasets into a lower-dimensional sub-space (in this case two dimensions for visualisation) by attempting to model the data as a locally connected manifold.⁹⁶ The UMAP algorithm was used with the following non-default parameters: number of neighbours set to 20 and minimum distance to $1e^{-5}$.

ⁱⁱ<https://github.com/lmcinnes/umap>

6

CHEMINFORMATICS AND HIGH-CONTENT IMAGING

6.1 Introduction

6.1.1 Cheminformatics

The term “cheminformatics” was first coined in 1998⁹⁷ although the use of computers to interact with chemical data predates this by many years with early systems used to index, search and catalogue databases of chemical compounds.⁹⁸ Most of the early work in this field was concerned with efficient means to search chemical databases for similar molecules or molecules containing certain sub-structures. This early work developed a number of important methods to generate, represent and compare chemical structures in a time of limited computational power, as a by-product these methods are very efficient and are still used today as the size of chemical databases has grown alongside computational power.

It was later on that researchers attempted to correlate biological activity and physiochemical parameters with structure activity relationships (SAR), this was partly due to the advancement of statistical techniques which gave rise to new tools such as multiple linear regression. One of the first quantitative SAR (QSAR) studies was carried out by Hansch and Fujita, in which they found the lipophilicity of a molecule correlated strongly with biological activity.⁹⁹ Since then the QSAR field has advanced to include many more parameters and is now a key part in most empirical drug discovery efforts.

Another use of cheminformatics in drug discovery is the analysis and design of compound screening libraries. In industrial high-throughput screening a full-deck compound library typically contains several million small molecules, screening this entire library is a costly endeavour, even for pharmaceutical companies, and therefore a lot of research has been carried out in how to maximise the value and information gained from screening large compound collections. One of the ways compound libraries can be optimised is by covering a large range of chemical space as possible. A compound library that contains many extremely similar molecules may be useful in certain specific circumstances, but in most cases this is viewed as a redundancy and a library which covers the same chemical space with fewer compounds would reduce costs. Alternatively, a compound collection of equal size which contains more diverse chemistry may lead to a more varied selection of lead candidates.¹⁰⁰ The concept of chemical space in compound collections can also be used to identify potential blind-spots or bias in drug discovery libraries, which are areas of chemical space with potential biological potential that are not covered by an existing library, in contrast to areas

of chemical space which are well covered by a compound collection but have historically failed to show biological activity, termed “dark chemical matter”.¹⁰¹

6.1.2 Structure activity relationships

A structure activity relationship is the link between a chemical's structure and its effect in a biological system, which underpins much of the medicinal chemistry field. The underlying premise of SAR is that compounds with similar structures and physiochemical properties have similar biological effects by virtue of binding to the same or similar targets. This idea is commonly applied during lead optimisation whereby a candidate molecule is iteratively modified in order to optimise parameters such as specificity and affinity, all the while ensuring that these modifications do not disrupt binding to the desired target, leading to the identification and determination of functional groups which are required for target engagement and biological activity.

Relating changes to a compound's structure to biological activity is relatively straightforward if compound activity can be represented as a single variable such as binding affinity or EC₅₀, applying quantitative SAR (QSAR) to multiparametric data such as that found with high-content imaging is not as well defined.

6.1.3 Chemical similarity

The premise of QSAR is “*similar molecules* have similar biological effects” presenting the challenge of how to measure similarity between chemical structures. Chemical structures can be represented in a number of different formats and we typically think of the skeletal 2D graphical representation (figure 6.1 A) when considering complex organic molecules which have to be interpreted by chemists. Computers however require a different format to efficiently store and parse chemical structure data. SMILES (simplified molecular input line entry system) and InChIs (international chemical identifier) are two formats which encode chemical structures as short character strings representing atoms as human readable characters (such as CH for carbon and hydrogen) with other symbols to represent branches and stereochemistry (figure 6.1 B&C). These relatively simple formats sometimes suffer from ambiguity, in which a single encoding could represent several molecules, or a single molecule could be represented by multiple valid encodings. A less ambiguous but also less human-readable file format is SDF (structure data file) or Molfile, which encode chemical structures as a table of x, y, z co-ordinates and bonds for each atom (figure 6.1 D).

Given these encodings of chemical structure and the task to calculate similarity (or distance) between molecules, the most direct and simple method is to calculate distance based on the string encodings (usually SMILE format), such as hamming distance or longest-common-substring divided by total length between two SMILE strings.¹⁰² However, these naive methods suffer from a number of drawbacks, mainly stemming from the ambiguity and variability of SMILE encodings which limit their widespread use in chemical similarity calculations. A more nuanced approach to measuring chemical similarity is to first calculate compound fingerprints such as daylight or extended connectivity fingerprints (ECFP)¹⁰³ which are abstract representations of molecules in the form of fixed-length binary arrays – generated from local patterns in the molecule such as the

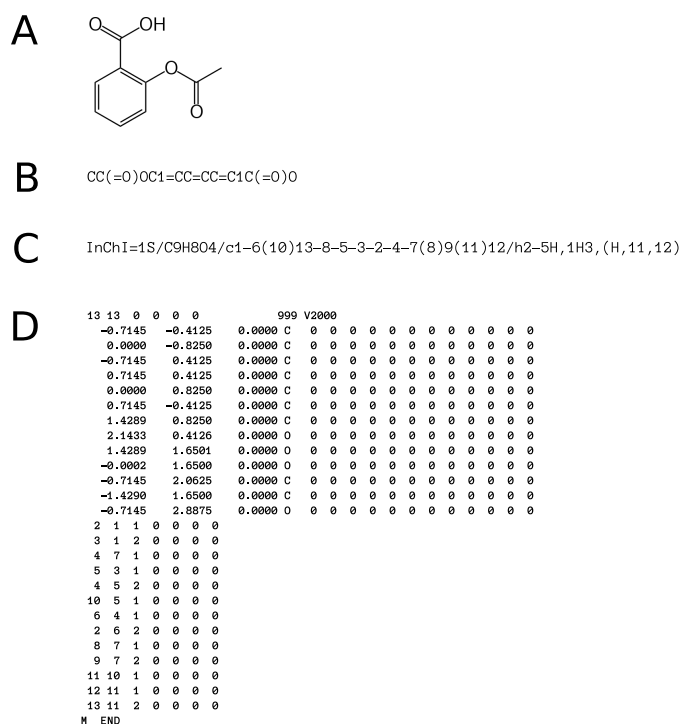


Figure 6.1 Different methods to encode the chemical structure of a molecule (aspirin). **(A)** A 2D skeletal graphical representation commonly used by chemists. **(B)** SMILE format, a concise relatively human readable format encoding atoms as characters. **(C)** InChI format, another commonly used string format which is less human readable but contains more details to reduce ambiguity. **(D)** SDF / Mol format. A tabular format which lists the coordinates of atoms in 3 dimensions along with bonds and distances.

identity of neighbouring atoms (where neighbouring is extended to several bonds away). The distance between compound fingerprints can then be found using one of a variety of distance metrics. To compare the binary compound fingerprints the most commonly used metric is Tanimoto similarity (T_s) and distanceⁱ (T_d), where T_s is defined as the ratio of common elements between two equal length fingerprints divided by the length of either fingerprint, and $T_d = -\log_2(T_s)$. Another approach to molecular fingerprinting is to summarise the 3D shape of a molecule. Ultrafast shape recognition (USR) was developed and used for *in silico* drug screening to efficiently describe molecular shape in 12 measurements. USR however is optimised for computational efficiency at the expense of detailed information and is agnostic to the atom types contained in the molecule. This drawback led to an extension of USR (USRCAT - USR with CREDO atom types) which was later developed for users to search the protein data bank and describes a molecule's 3D shape and constituent atoms.¹⁰⁴ Recently a number of studies have leveraged advances in the machine learning field to generate alternative chemical fingerprints using neural networks.^{105,106,107,108} The idea behind these methods is that deep neural networks are able to learn appropriate representations of the input data in order to maximise performance in a certain task. They typically represent chemical structures as un-directional graphs of atoms, and apply convolutional techniques – which have proven themselves in image-related tasks – to the graph structures to generate molecular fingerprints which can be used in downstream machine learning and cheminformatics work.

6.1.4 Application of cheminformatics to high-content screening

Much of the work in cheminformatics is carried out in industrial rather than academic laboratories, coupled with the relatively immature field of high-content imaging has resulted in a sparsity of

ⁱWhilst not a distance in the strict mathematical sense it is commonly referred to as a distance metric.

published research in the application of cheminformatics to high-content imaging and screening.

One of the earliest papers which combined cheminformatics with image-based screening was by Young *et al.*³⁰ who screened a library of 6,547 compounds in HeLa cells and extracted 30 morphological features regarding nuclei morphology. They used factor analysis and hierarchical clustering to group their compound library into 7 clusters describing similar nuclear morphologies, and created matrices of phenotypic similarity with cosine similarity of phenotypic features and compound similarities with Tanimoto coefficients of ECFP features. They then found a correlation between the rank ordering of phenotypic similarities and compound similarities, as well as identified instances of “activity-cliffs” when two structurally similar compounds demonstrated very different phenotypic activities which matched up to known SAR studies on the two compounds.

A second study by Wawer *et al.*¹⁰⁹ incorporated high-content morphological profiling to construct compound libraries based on the diversity of biological response as opposed to diversity of chemical space. Using a library of 31,000 compounds, they performed a image-based screen and selected a subset of compounds which produced a diverse range of bioactivities defined with cell morphology. They then compared this subset to a second subset generated by maximising diversity of chemical space, and investigated the performance of each subset of compounds in a wide range of previously performed cell-based screens. They found that subsetting compounds based on morphological diversity resulted in an increased performance compared to compounds chosen on chemical diversity or compounds chosen at random.

Another study published by the same group developed a method for SAR with high-dimensional profiling data, assessing both high-content imaging and gene expression profiling datasets. They used pattern mining techniques originally developed in advertising and marketing to find frequently linked sub-structures with certain biological activities.¹¹⁰

6.1.5 The BioAscent library

The BioAscent compound library consists of a 12,000 compound subset of a larger 125,000 chemical diversity library. The library was designed to include compounds with drug-like properties such as adherence to Lipinski’s rule of 5 and avoiding known pan-assay interference compounds (PAINs). The bioascent collection has been found to contain a considerable proportion of molecules which are likely to be kinase-interacting (27%) and GPCR-interacting (20%) according to computational models of chemical structure performed by the vendor.

6.1.6 Aim of this chapter

This chapter is based on work using the BioAscent compound library which is supplied with detailed structural information of each of the 12,000 compounds. My aim was to incorporate this chemical information with existing public datasets and my own high-content imaging data in a way to aid target convolution as well as investigate the link between chemical structure structure activity relationship (SAR) applied to cellular morphology as an indicator of compound activity.

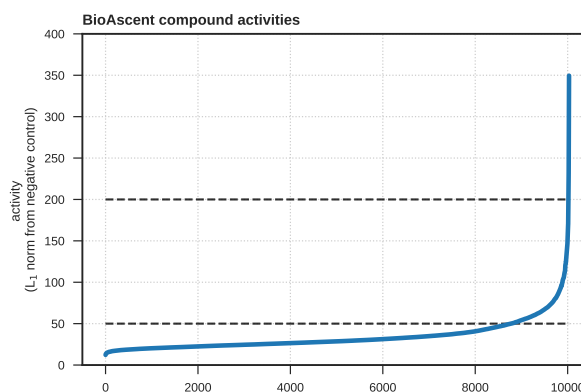


Figure 6.2: Selection of active BioAscent compounds based on the l_1 norm distance from the DMSO negative control centroid in PCA space. Lower and upper bounds of the selected compounds are indicated by dashed lines. In total 1244 compounds were selected.

6.2 Results

6.2.1 The BioAscent library contains clusters of phenotypically similar compounds

In order to compare the phenotypic profiles produced by compounds in the BioAscent library, active compounds were selected based on the l_1 norm distance from the negative control centroid (figure 6.2). As many of the compounds were cytotoxic and produced images containing only a few cells which do not produce robust morphological measurements, an activity window was used to exclude cytotoxic compounds.

Hierarchical clustering of morphological profiles produced by these phenotypically active compounds showed that despite the chemical diversity of the BioAscent library, the active compounds formed distinct clusters of compounds which produced similar cellular morphologies (figure 6.3 A). To confirm the validity of the clustering, the hierarchical labels were compared with clusters found in an unsupervised algorithm. The morphological profiles were embedded into 2-dimensional space using the t-SNE algorithm¹¹¹ which aims to preserve local structure within the data and reveals clusters of similar points in an unsupervised manner. When these points were coloured by the cluster labels identified by hierarchical clustering they appeared to match up with the t-SNE embedding (figure 6.3 B).

6.2.2 The BioAscent library is chemically diverse

The BioAscent library is marketed as chemically diverse, yet I still wanted to see to what extent and if there are clusters of chemically similar compounds such as those based around a common scaffold. All 12,000 BioAscent compounds were converted into molecular fingerprints to produce a Tanimoto distance matrix between all pairs of compound fingerprints, this was then clustered using agglomerative hierarchical clustering. As could be predicted, the heatmaps and dendrograms did not reveal any large clusters of structurally similar compounds in the 12,000 compound library. This chemical diversity continued when the compounds were filtered to only contain the pheno-

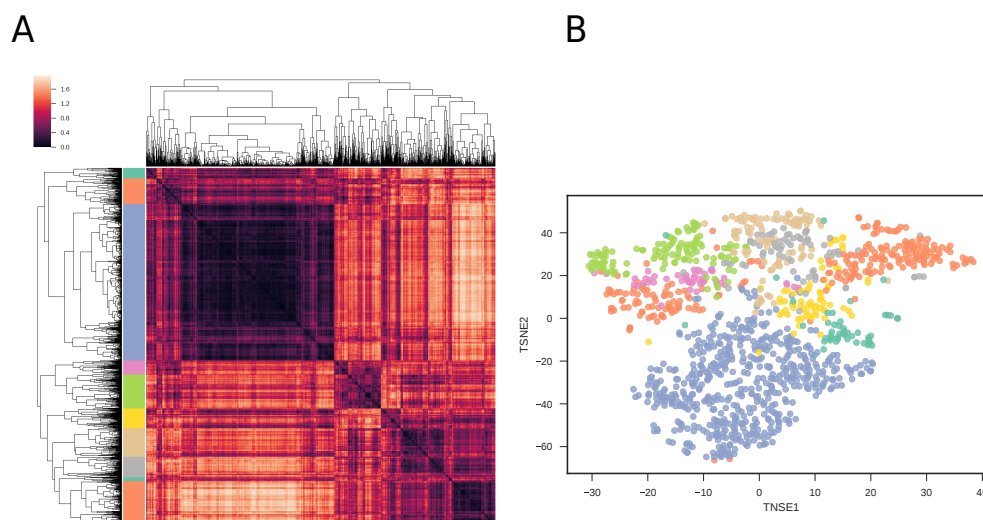


Figure 6.3: Morphological clustering of active compounds within the BioAscent library. **(A)** Hierarchical clustering of the 1244 active BioAscent compounds based on a distance matrix of principal components. Clusters formed by cutting the produced dendrogram. **(B)** Unsupervised t-SNE clustering of active BioAscent compounds based on principal components of morphological features. Points are colour coded with cluster labels derived from the hierarchical clustering.

typically active molecules. The use of more novel compound fingerprinting techniques such as USRCAT¹⁰⁴ and autoencoded features¹¹² did not increase the degree of clustering.

Rather than looking at large-scale clustering of many thousands of compounds with hierarchical clustering, I tried the Butina clustering method to identify small collections of structurally similar compounds. This method does not return similarity measures, but rather groups compounds into bins of similar compounds.¹¹³ After removing clusters which contained fewer than 3 compounds, this left 96 clusters, with the largest cluster containing 20 compounds and 58% of the clusters containing only 3 compounds (figure 6.4).

6.2.3 There is little evidence that structurally similar molecules produce similar cellular morphologies

Following the premise of SAR, structurally similar molecules are likely to share a common target, therefore activating the same or similar signalling pathways and producing similar cellular morphologies. I investigated to what extent structurally similar molecules in the BioAscent library produce similar cellular morphologies, and also how structurally similar are compounds which were shown to produce similar phenotypes. Using the phenotypic clusters as defined in fig. 6.3, I compared the structural similarity between compounds within these phenotypic clusters compared to a null distribution of pairs of compounds picked at random. I found that compounds within phenotypic clusters were very slightly more structurally similar than compounds in the null distribution (figure 6.5 A, $p = 1.81 \times 10^{-15}$, $D = 0.011$, 2-sample Kolmogorov-Smirnov test). In addition, I approach the problem from the opposite direction and investigated the phenotypic

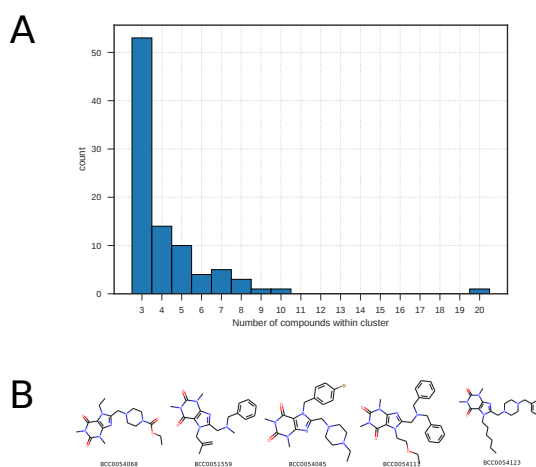


Figure 6.4: (A) Histogram of number of compounds within structurally similar clusters, with most clusters only containing 3 molecules. (B) An example of one of the structurally similar clusters as found with the Butina clustering algorithm.

similarity within clusters of structurally similar molecules as found with the Butina clustering algorithm, compared to the phenotypic similarity between compounds picked at random from the pooled compound list of those contained within Butina clusters. I again found that structurally similar molecules are more likely to produce similar cellular morphologies than compounds picked at random (figure 6.5 B, $p = 0.037$, $D = 0.018$, 2-sample Kolmogorov-Smirnov test).

Another approach is to see how well the distance matrix of phenotypic profiles correlates with the distance matrix of chemical structures. Using Mantel's test of correlation between two distance matrices,¹¹⁴ I found no significant correlation between the phenotypic and structural distance matrices for the active 1244 compound subset ($r = 0.02$, $p = 0.116$).

6.2.4 Identifying the putative MoA of phenotypic hits with ChEMBL structure queries

Another way to utilise the chemical structure data available with the BioAscent library is through querying publicly available databases such as ChEMBL for exact compounds matches or structurally similar compounds. This returns large amounts of data from a variety of assays in which the compound or a structural analogue was screened against a number of targets with information relating to EC/IC₅₀ values, binding affinities etc. I investigated if this historical dataset could be used to suggest putative MoAs of hits from target agnostic phenotypic screening assays.

For this I used the compounds within the 10 phenotypic clusters (figure 6.3), and for each cluster queried ChEMBL based on a structure similarity search to identify records for either the query compound, or structural analogues. Then using these compounds identifying which human proteins they have been screened against, and filtering these protein based on EC/IC₅₀ values. This returns a list of Uniprot accession codes which were used with interpro¹¹⁵ to test for enrichment of protein regions compared to a background.

Eight out of the ten phenotypic clusters returned at least one significantly enriched target with fold-enrichment ranging between 1.5 and 10. The most significantly enriched target in 6/8 of the

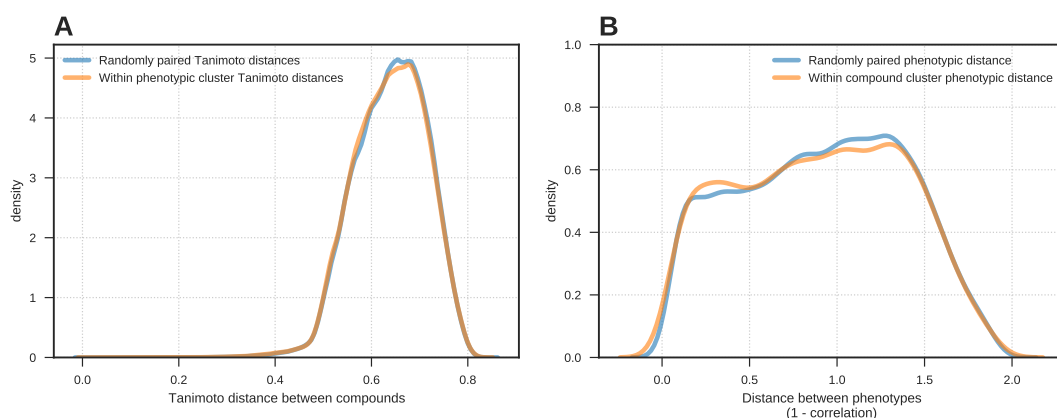


Figure 6.5: (A) Tanimoto distance between compounds from within phenotypic clusters (as found in fig. 6.3) and between randomly paired active compounds. ($p = 1.81 \times 10^{-15}$, $D = 0.011$, 2-sample Kolmogorov-Smirnov test) (B) Phenotypic distance between compounds from within structurally similar clusters and between randomly paired phenotypic profiles. ($p = 0.037$, $D = 0.018$, 2-sample Kolmogorov-Smirnov test)

clusters was related to protein kinases, whereas the remaining two were rhodopsin-like GPCRs and adrenergic receptors.

6.2.5 Using phenotypic screening to find “dark chemical matter”

An area of interest in drug discovery is finding new pharmacologically active compounds which occupy new areas of chemical space.¹⁰¹ One way to incorporate the phenotypically active hits from the BioAscent library is to query historical screening databases by structural similarity. To do this I took the list of 1244 phenotypically active BioAscent compounds and performed a structural similarity search on the ChEMBL database to look for those BioAscent compounds which have a large Tanimoto distance from all compounds deposited in the database.

From the 1244 active BioAscent compounds 59 (4.7%) were found to have no structurally similar analogues in the ChEMBL database (figure 6.7). To assess if these 59 compounds contained undesirable physiochemical properties which would limit their inclusion in screening libraries and explain their absence from historic screening databases I used a quantitative estimate of drug-likeness (QED),¹¹⁶ to compare the 59 compounds from ‘dark chemical space’ to the 1244 active BioAscent compounds. The QED metric did not reveal any significant differences in desirable physiochemical properties between the two groups ($\text{QED}_{\text{dark compounds}} = 0.57$, $\text{QED}_{\text{all active}} = 0.60$, 2 sample t-test $t = 0.85$, $p = 0.39$).

6.3 Discussion

Cheminformatics as a field is largely overshadowed by bioinformatics in terms of academic interests and publications (figure 6.8), it has however arguably had a greater positive impact on the design and identification of new small molecule therapeutics. As high-content screening becomes more

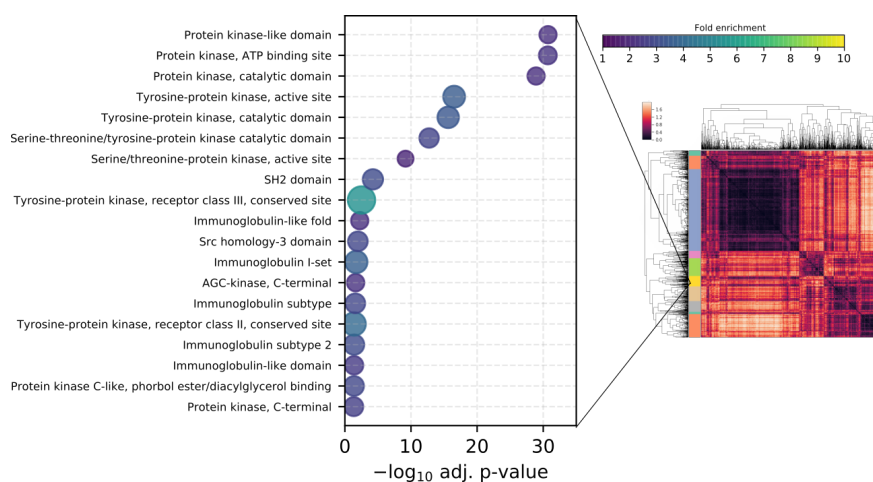


Figure 6.6: Enriched interpro targets found within a phenotypic cluster of the BioAscent library when compared to a background of all active BioAscent compounds.

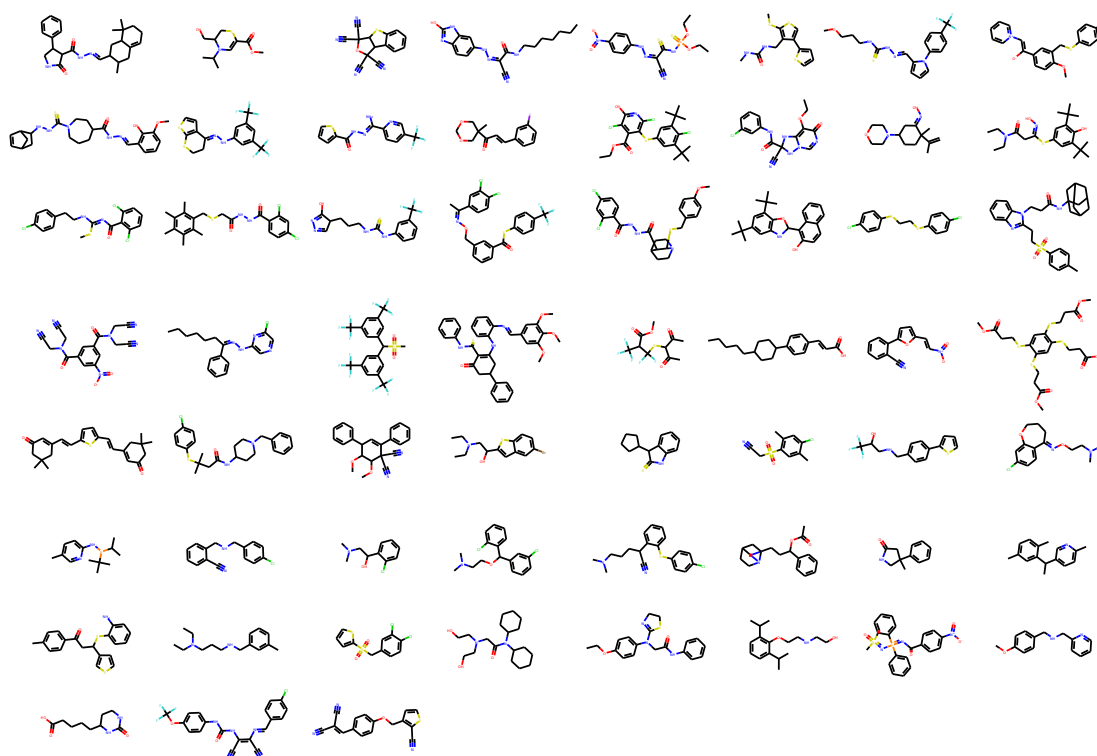


Figure 6.7: 59 phenotypically active BioAscent compounds with no close structural analogues in the ChEMBL database.

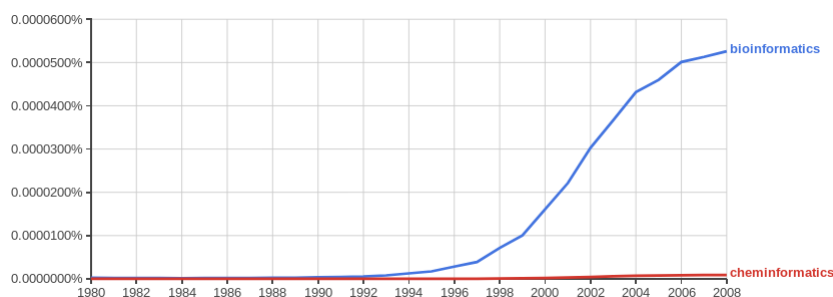


Figure 6.8: Popularity of the terms ‘bioinformatics’ and ‘cheminformatics’ in the published literature as found using Google’s Ngram viewer between 1980 and 2010. y-axis represents the cumulative percentage of literature containing the term, x-axis represents the year.

and more prevalent in drug discovery incorporation of the fields will become more increasingly likely. I therefore aimed to investigate methods in which cheminformatics analyses can aid high-content and phenotypic screening, and also the other way round: how high-content screening and morphological profiling can inform cheminformatics.

From the global analysis of the BioAscent compound library I failed to find any evidence of clusters of structural similarity, which is not surprising when using a compound library specifically designed to maximise structural diversity. I did however find smaller regions of the BioAscent library consisting of a handful of structurally similar compounds using the Butina clustering algorithm. The choice of using the BioAscent compound library – rather than one of many other alternatives – was made by what was available to me at the time, as large compound collections are a precious resource in academia. In hindsight, a chemical diversity library may not have been the ideal compound collection to use for a study relying heavily chemical similarity measures, and a compound library which consists of clusters of structurally similar molecules may have resulted in different conclusions regarding chemical similarity and phenotypic similarity.

My hypothesis that structurally similar compounds should produce similar morphological changes, and therefore compounds that cause similar phenotypes should be structurally similar on average did not yield particularly striking results. While I found compounds within phenotypic clusters had lower Tanimoto distances than compounds paired at random, and the opposite: that compounds within structurally similar clusters as found with the Butina algorithm were more phenotypically similar (figure 6.5), despite statistical significance the effect size was small, in globally assessing correlation between the two distance matrices showed no significant correlation. This result is largely in agreement with that of Young *et al.* who found a “modest” correlation of 0.0074 between between rank-ordered pairs of compounds for phenotypic similarity and structural similarity.³⁰ One possible explanation for these low effect sizes could be due to largely uncorrelated data with small regions of high correlation. I feel that a more fine-grained analysis with a carefully constructed compound collection would be better suited for this task, and could result in stronger evidence for the association between chemical structure and phenotype. Another consideration to explain the largely uncorrelated data are activity-cliffs – where a small change to a molecule’s structure can result in large differences in biological activity. There is no doubt that a small change to overall

chemical structure can inhibit binding of a small molecule to a target receptor, although this brings into question the usefulness of chemical similarity measures, and if many of these activity-cliffs are artefacts caused by poorly measured ‘similar’ compounds, which we may see change as more nuanced chemical similarity measures are developed.

The availability of large public datasets which can be queried with chemical identifiers such as SMILE strings is a great resource with a number of potential applications. The ChEMBL database contains information for 2.2 million compounds, and the results from over a million assays and 12,000 targets. In my efforts to incorporate this rich dataset with the results of the high-content screen, I encountered issues associated with a dataset constructed from many heterogeneous sources, such as lack of information describing the assay, and no consistent system to label the type of assay to allow filtering of less relevant assay types. The idea was to find existing data from assays which used the 12,000 compound BioAscent library, however none of the data sources used the exact BioAscent compound library, but rather there were compounds within the BioAscent library that are shared in other compound collections, and so the data returned by exactly matching the BioAscent compounds was too sparse for further analysis. I therefore relaxed the searching criteria, and searched instead for compounds with a Tanimoto similarity greater than 0.9 which resulted in an adequate number of results but added an additional layer of assumptions. The enriched protein sequences found for the compounds (or similar compounds) in each phenotypic cluster consisted predominantly of protein kinase regions (see figure 6.6 for an example of one cluster). While this did serve as a nice sanity check, in that 20% of the BioAscent compounds are predicted to be kinase-interacting, it was not particularly interesting for hypothesis generation. In addition I would warn against putting too much faith in the hypothesised protein targets: the protein targets were filtered using single concentration regardless of the assay type. It is easy to envisage that a concentration which is selective to a particular protein in a cell-based assay would not be stringent enough when used as a cutoff in an *in vitro* protein binding assay. Another source of uncertainty is the use of tools such as DAVID and InterPro to predict enriched protein regions, these rely on heuristics and combining another set of heterogeneous datasets which in turn have their own errors and biases.

The concept of dark chemical matter was introduced by Wasserman and colleagues from Novartis to describe compounds in their high-throughput screening library which have failed to show biological activity in any screening assay, yet through gene-expression studies demonstrated the potential for biological activity in future screens.¹⁰¹ These compounds offer interesting starting points for drug discovery as their lack of activity in historically target-driven screens may mean they have the potential to act through novel mechanisms of action. A target agnostic approach coupled with unbiased detection of subtle biological activity positions high-content imaging as a useful tool to identify dark chemical matter in compound collections. As I did not have access to historical records of the BioAscent’s performance in a wide range of assays, I instead used the records in the ChEMBL database. From the 1244 active BioAscent compounds, 59 were structurally distinct from any listed in the ChEMBL records (figure 6.7). There is also the possibility that there may be more dark chemical matter in the BioAscent library, as I did not investigate the bio-activity of the structurally similar records in the ChEMBL database, and that many of those which returned structural analogues may not have shown activity in previous assays. As the BioAscent library has

been designed around drug-like molecules, and a measure of drug-likeness did not reveal any undesirable physiochemical properties of these dark chemical matter the reason behind their exclusion in previous screening assays remains unclear.

Overall, incorporating cheminformatics and high-content screening presents an interesting opportunity for drug discovery by combining the well-defined and annotated cheminformatics field with the rich datasets high-content imaging can provide. In this chapter I have shown that high-content screening data can be combined with existing datasets to aid interpretation using chemical structure as a common linker to retrieve data for either the same compound or similar compounds, as well as demonstrating the use of high-content screens to identify interesting areas of chemical space for the development of novel therapeutics.

6.4 Methods

6.4.1 Chemical similarity

Compound structural information was provided in the form of .sdf files by the supplier. To create daylight-like compound fingerprints the RDKit library was used to convert .sdf entries into an RDKit's implementation of the daylight fingerprint using the `'rdkit.Chem.Fingerprints.FingerprintMols'` function with default parameters.

USRCAT features of the BioAscent library were generously calculated and supplied by Dr. Steven Shave (Edinburgh).

Latent representations of chemical structure features were calculated using a molecular autoencoder pre-trained on the ChEMBL22 datasetⁱⁱ, based on the work published by Gomez-Bombarelli *et al.*¹¹² using one-hot encoded SMILE strings of the molecules.

To compute the distance between RDKit daylight fingerprints the Tanimoto distance was used, in the case of USRCAT and autoencoded features I used the Euclidean distance. Hierarchical clustering was performed on the distance matrix using the complete linkage method and euclidean distance. To define clusters from the calculated dendrogram, a threshold was defined as 70% of the maximum linkage distance. Butina clustering was implemented using RDKit with Tanimoto distances calculated from daylight fingerprints, with a cutoff value of 0.2.

Mantel's test for comparing two distance matrices was implemented with scikit-bio's implementation using Pearson's correlation coefficient and 999 permutations for significance testing. The distance matrices used were standardised Euclidean distance for the morphological profiles and standardised Tanimoto distances of the daylight fingerprints for compound structure profiles.

6.4.2 BioAscent library screen

The morphological data used in this chapter is from the MCF7 cell-line stained using the cell-painting protocol, imaged with the ImageXpress and morphological features calculated using Cell-profiler.

ⁱⁱwww.github.com/cxhernandez/molencoder

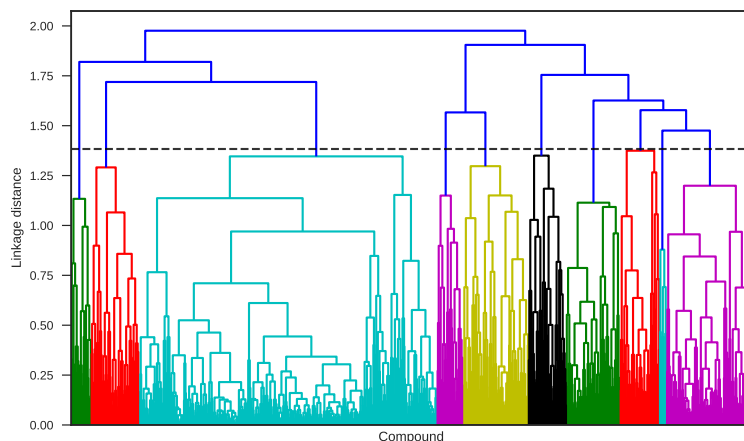


Figure 6.9: Dendrogram thresholding to determine the number of phenotypic clusters in the active BioAscent compounds. Dashed line indicates cutoff of 70% of the maximum linkage distance, resulting in 10 clusters.

Compound activity window

Data was normalised to plate-based controls and features standardised, then transformed with PCA to the minimum number of principal components which accounted for 80% of the variance in the data. l_1 norm distances were calculated from the DMSO negative control centroid in PCA space. The lower bound of the activity window was defined visually using a plot of ranked l_1 distances. The upper bound was chosen based on images containing at least 10 cells and visual assessment of images produced by higher l_1 distances ensuring images did not consist entirely of dying cell (small, rounded and bright cytoplasmic staining).

6.4.3 Phenotypic similarity

Clustering of morphological profiles was carried out by first calculating a correlation matrix between between all pairs of active compound morphologies. Hierarchical clustering was performed on the correlation matrix using the complete linkage method and euclidean distance. To define clusters from the calculated dendrogram, a threshold was defined as 70% of the maximum linkage distance which produced 10 clusters (figure 6.9)

t-SNE clustering was performed using sklearn's 'manifold.TSNE' implementation using the Barnes-Hut approximation with the default parameters.

6.4.4 ChEMBL structure searches

To programmatically query the ChEMBL database I used the python ChEMBL webresource client.

ⁱⁱⁱ In order to identify records for similar compounds I first queried structures based on SMILE strings of the BioAscent compounds with a filter to return only compounds with a Tanimoto sim-

ⁱⁱⁱ www.github.com/chembl/chembl_webresource_client

ilarity of 0.9, recording the similar compounds as ChEMBL identifiers. Then in a second query using the ChEMBL identifiers, I searched for historical screening results against human protein targets and returned a list in the form of Uniprot accession codes. As this returned a list of all protein targets which had been screened against, I filtered this list to protein targets with an assay EC/IC₅₀ value less than 1 μ M. This was repeated for each cluster of BioAscent compounds returning a list of Uniprot accession codes for each cluster.

6.4.5 Dark chemical matter

To search for active compounds in the BioAscent library which are structurally distinct from any compounds in the ChEMBL database I queried the ChEMBL webresource with the 1244 active BioAscent compounds, returning compounds within 70% similarity, which is equivalent of compounds within 0.3 Tanimoto distance (this is the minimum similarity value allowed when using ChEMBL's API). Any BioAscent compound that failed to return any structurally similar ChEMBL record was listed as a 'dark SMILE'.^{iv} QED values of drug-likeness were computed using RDKit Chem.QED.qed function with default parameters on molecules computed from the supplied sdf file.

6.4.6 Interpro analysis

Interpro analysis was carried out using DAVID 6.8.¹¹⁷ DAVID was chosen despite more up-to-date alternatives, as DAVID allows uploading a custom background list of genes or proteins. Therefore I created a background list of protein targets by repeating the Uniprot lookup as before but with a list of all 12,000 BioAscent compounds, which was used as a background for each cluster analysis with DAVID. Significantly enriched interpro targets were selected based on a Benjamini-Hochberg corrected p-value with an α of 0.05.

^{iv}A thanks to Michał Nowotka from the EMBL-EBI for his help making changes to the ChEMBL servers and API to enable such time-intensive queries.

7 | CONCLUDING REMARKS

7.1 Summary of completed work

This work describes my research into the development and application of high-content image-based screening methods in the context of cancer drug discovery. The first results chapter builds on published literature to further investigate how compound induced morphological changes measured with high-content imaging can be used to describe and inform the compound mechanism-of-action across a panel of genetically distinct breast cancer cell lines. A comparison of two machine learning approaches revealed that, in fairly naive implementations, there is not a large difference in predictive performance between tree-based ensemble classifiers trained on extracted morphological measurements and CNN classifiers trained on pixel values. There is however a difference in how well these two types of models can generalise to new data from new or unseen cell-lines, as the extracted morphological measurements used with the tree-based classifier can be more easily normalised which in turn affects how the addition of data during training from morphologically distinct cell-lines impacts model performance. In chapter 4 I described the development of a measure of morphological dissimilarity. Inspired by a talk given by Simon Gordonovⁱ, I thought to extend the idea of measuring distance from a negative control in principal component space of morphological features to incorporate the idea and quantification of phenotypic direction. This method was then applied in chapter 5 with a small molecule screen of approved drugs across the eight breast cancer cell-lines to identify compounds that produced distinct phenotypic responses between the cell-lines. These compounds were then further investigated in 2D and 3D tumour spheroid assays of cell-viability, and proteomics performed the levels and activation state of common cancer cell growth and survival signalling pathways in cells treated with compounds grown in 2D and 3D environments. The final results chapter is my effort to incorporate data from small molecule high-content screening studies with data relating to chemical structure. The original premise behind this work was the hypothesis that structurally similar molecules are likely to produce similar morphological changes in cells. I found evidence of correlation between chemical similarity and phenotypic similarity, although the effect size was extremely small. Perhaps more interesting was the use of chemical structure data and existing chemical databases to generate hypotheses towards mechanistic understanding of hits found with target-agnostic screens, as well as high-content screens to identify compounds from interesting and rarely explored areas of chemical space.

During this work I spent a considerable amount of time developing software tools, either to

ⁱA talk at an SBI² meeting describing phenotypic directions in principal components which has since been published¹¹⁸

implement new ideas or to streamline repetitive workflows which are commonplace in screening assays. One of the biggest challenges was the time taken to analyse the millions of images generated by compound screens across a panel of cell-lines. Whilst microscope vendors typically have software to automatically analyse and quantify images once acquired, in my case the software was limited by licences to a single desktop and unable to analyse the 12,000 compound screen in a reasonable amount of time. Image analysis tools such as Cellprofiler, EBIImage and HCS-analyser with permissive licences allow the analysis of thousands of images in parallel using distributed processing across compute clusters. The most useful tool I developed was used to link the ImageXpress images to Cellprofiler running on the University's high-performance compute clusterⁱⁱ, enabling the analysis of a 5 million image dataset in roughly 24 hours, which would have otherwise taken months using the vendor supplied image analysis software.

7.2 Remarks, unanswered questions and new questions

High-content analysis and screening is an evolving field and has not yet reached a consensus on established workflows or best practices, with numerous labs developing their own image analysis software and data handling pipelines in isolation. Recently there has been some effort to coordinate sharing methods and ideas between groups to converge on a standardised workflow.ⁱⁱⁱ While this is an important step, high-content analysis has the enviable position of being at the crossroads of computer vision, multivariate analysis and machine learning, all of which are rapidly developing fields in their own right. Therefore, despite efforts to reach an agreement on some form of standardised workflow, there is the conflicting temptation for researchers to adopt the latest tools and techniques in their analyses.

With the rapid development of new machine learning tools, particularly in computer vision, I envisage that the field will adopt these technologies where they show increased performance over hand-crafted algorithms in areas such as segmentation,^{119,120} feature extraction^{121,122} and image classification.¹²³ However, the use of “classical” extracted morphological features from images such as cell area or nuclei intensity offer a huge advantage in their simplicity, interpretability and the ability to investigate specific biological questions or image-analysis tasks.

With the increasing ability to generate large multivariate datasets from high-content screening, perhaps a more pertinent area of research is how best to leverage this data to improve our understanding of biological processes and find new and efficacious drugs for patients. The interpretation of large multivariate datasets in biology is not unique to high-content imaging and is a task shared in common with most -omics technologies, with the only difference is that high-content imaging is usually cheaper than its -omics counterparts per sample – and as a result sample sizes are typically much larger. This commonality between technologies will hopefully lead to the development of new methods which are applicable to drug screening studies, which have historically relied upon univariate measures and statistical assumptions that do not necessarily hold true with more complex

ⁱⁱ<https://github.com/carragherlab/cptools2>

ⁱⁱⁱ<http://cytodata.org/>

<https://github.com/shntnu/cytomining-hackathon-wiki/wiki>

datasets.

In my opinion, the “profiling” of perturbations such as small molecule treatments or gene knock-outs, while certainly possible with high-content imaging, may benefit more from the standardised measured features such as L100h0 gene expression profiles of the connectivity map^{124,125} which allow far easier comparisons and meta-analyses of disparate datasets in lieu of increased costs and lower throughput. However, the low-cost and high-throughput of high-content screening is ideally suited for drug discovery projects using complex disease-relevant models which require multivariate measurements in order to accurately capture and quantify their complexity.

To conclude, I have presented work relating to a number of varied aspects of image informatics and high-content screening, these contributions are part of a rapidly developing field with many remaining questions and unverified assumptions. As the field of biology progresses towards generating ever larger and more complex datasets there needs to be a similar progression in the research and development of data analysis methods to gain more understanding from the data we generate. It is my hope that the evolution of new biological and analytical methods which enable in-depth profiling of compound mechanism-of-action and target biology across more complex *in vitro* models of disease will better lead early stage drug discovery programmes towards increased clinical success rates.

BIBLIOGRAPHY

- [1] Jack W. Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. “Diagnosing the decline in pharmaceutical R&D efficiency”. *Nature Reviews Drug Discovery* 11.3 (2012), pp. 191–200.
- [2] Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. “The productivity crisis in pharmaceutical R&D”. *Nature Reviews Drug Discovery* 10.6 (2011), pp. 428–438.
- [3] Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace, and Alex Weir. “An analysis of the attrition of drug candidates from four major pharmaceutical companies”. *Nature Reviews Drug Discovery* 14.7 (2015), pp. 475–486.
- [4] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. “Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells”. *Nucleic Acids Research* 41.D1 (2013), pp. D955–61.
- [5] 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.
- [6] Ripudaman S. Hundal, Martin Krssak, Sylvie Dufour, Didier Laurent, Vincent Lebon, Visvanathan Chandramouli, Silvio E. Inzucchi, William C. Schumann, Kitt F. Petersen, Bernard R. Landau, and Gerald I. Shulman. “Mechanism by which metformin reduces glucose production in type 2 diabetes”. *Diabetes* 49.12 (2000), pp. 2063–2069.
- [7] 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.
- [8] Susan E. Leggett, Jea Yun Sim, Jonathan E. Rubins, Zachary J. Neronha, Evelyn Kendall Williams, and Ian Y. Wong. “Morphological single cell profiling of the epithelial-mesenchymal transition”. *Integrative Biology (United Kingdom)* 8.11 (2016), pp. 1133–1144.
- [9] Yoshikuni Tabata, Norio Murai, Takeo Sasaki, Sachie Taniguchi, Shuichi Suzuki, Kazuto Yamazaki, and Masashi Ito. “Multiparametric phenotypic screening system for profiling

- bioactive compounds using human fetal hippocampal neural stem/progenitor cells". *Journal of Biomolecular Screening* 20.9 (2015), pp. 1074–1083.
- [10] 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.
 - [11] Dijun Chen, Kerstin Neumann, Svetlana 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.
 - [12] N Otsu. "A Threshold Selection Method from Gray-Level Histogram". *IEEE Transactions on Systems, Man and Cybernetics* 9.1 (1979), pp. 62–66.
 - [13] 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.
 - [14] Christoph Sommer, Christoph Straehle, Ullrich Kothe, and Fred A. Hamprecht. "Ilastik: Interactive learning and segmentation toolkit". *Proceedings - International Symposium on Biomedical Imaging* 1 (2011), pp. 230–233.
 - [15] Satwik Rajaram, Benjamin Pavie, Lani F. Wu, and Steven J. Altschuler. "PhenoRipper: Software for rapidly profiling microscopy images". *Nature Methods* 9.7 (2012), pp. 635–637.
 - [16] 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.
 - [17] Daniel B. Goldman. "Vignette and exposure calibration and compensation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.12 (2010), pp. 2276–2288.
 - [18] Robert M. Haralick, Its'hak Dinstein, and K. Shanmugam. "Textural Features for Image Classification". *IEEE Transactions on Systems, Man and Cybernetics* SMC-3.6 (1973), pp. 610–621.
 - [19] 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.
 - [20] 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.

- [21] Frank R. Hampel. “The influence curve and its role in robust estimation”. *Journal of the American Statistical Association* 69.346 (1974), pp. 383–393.
- [22] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. “Lof”. *ACM SIGMOD Record* 29.2 (2000), pp. 93–104.
- [23] 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.
- [24] 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.
- [25] Richard E. Bellman. *Adaptive Control Processes - A Guided Tour*. Princeton University Press, 1961, p. 255.
- [26] Hanchuan Peng, Fuhui Long, and Chris Ding. “Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1226–1238.
- [27] 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.
- [28] Zachary E. Perlman, Michael D. Slack, Yan Feng, Timothy J. Mitchison, Lani F. Wu, and Steven J. Altschuler. “Multidimensional drug profiling by automated microscopy”. *Science* 306.5699 (2004), pp. 1194–1198. arXiv: [1311.1716](#).
- [29] Cellular States, Sigrun M. Gustafsdottir, Vebjorn Ljosa, Katherine L. Sokolnicki, J. Anthony Wilson, Deepika Walpita, Melissa M. Kemp, Kathleen Petri Seiler, Hyman A. Carrel, Todd R. Golu, Stuart L. Schreiber, Paul A. Clemons, Anne E. Carpenter, and Alykhan F. Shamji. “Multiplex cytological profiling assay to measure diverse”. *PLoS ONE* 8.12 (2013), e80999.
- [30] 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.
- [31] 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), pp. 415–427.
- [32] Stijn Van Dongen. “Graph Clustering Via a Discrete Uncoupling Process”. *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 121–141.

- [33] 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.
- [34] David C. Swinney and Jason Anthony. “How were new medicines discovered?” *Nature Reviews Drug Discovery* 10.7 (2011), pp. 507–519.
- [35] M. Pickl and C. H. Ries. “Comparison of 3D and 2D tumor models reveals enhanced HER2 activation in 3D associated with an increased response to trastuzumab”. *Oncogene* 28.3 (2009), pp. 461–468.
- [36] Susan Breslin and Lorraine O’Driscoll. “Three-dimensional cell culture: The missing link in drug discovery”. *Drug Discovery Today* 18.5-6 (2013), pp. 240–249.
- [37] Carrie J. Lovitt, Todd B. Shelper, and Vicky M. Avery. “Miniaturized Three-Dimensional Cancer Model for Drug Evaluation”. *ASSAY and Drug Development Technologies* 11.7 (2013), pp. 435–448.
- [38] Jennifer Laurent, Céline Frongia, Martine Cazales, Odile Mondesert, Bernard Ducommun, and Valérie Lobjois. “Multicellular tumor spheroid models to explore cell cycle checkpoints in 3D”. *BMC Cancer* 13 (2013).
- [39] Yongyang Huang, Shunqiang Wang, Qiongyu Guo, Sarah Kessel, Ian Rubinoff, Leo Li Ying Chan, Peter Li, Yaling Liu, Jean Qiu, and Chao Zhou. “Optical coherence tomography detects necrotic regions and volumetrically quantifies multicellular tumor spheroids”. *Cancer Research* 77.21 (2017), pp. 6011–6020.
- [40] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie De Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi A. Garraway. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. *Nature* 483.7391 (2012), pp. 603–607.
- [41] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J. Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A. Smith, I. Richard Thompson, Sridhar Ramaswamy, P. Andrew Futreal, Daniel A. Haber, Michael R. Stratton, Cyril Benes, Ultan McDermott, and Mathew J. Garnett. “Genomics of Drug Sensitivity in Cancer (GDSC): A

- resource for therapeutic biomarker discovery in cancer cells". *Nucleic Acids Research* 41.D1 (2013), pp. 955–961.
- [42] 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 Multidrug-resistant Phenotype of Disease-oriented Panels of Human Tumor Cell Lines Used for Anticancer Drug Screening". *Special Topics in Drug Discovery*. Vol. 52. InTech, 1992, pp. 3029–3034.
- [43] Robert H. Shoemaker. "The NCI60 human tumour cell line anticancer drug screen". *Nature Reviews Cancer* 6.10 (2006), pp. 813–823.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] P. D. Caie, R. E. Walls, A. Ingleston-Orme, S. Daya, T. Houslay, R. Eagle, M. E. Roberts, and N. O. Carragher. "High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells". *Molecular Cancer Therapeutics* 9.6 (2010), pp. 1913–1926.
- [48] Andrew H. Sims, Anthony Howell, Sacha J. Howell, and Robert B. Clarke. "Origins of breast cancer subtypes and therapeutic implications". *Nature Clinical Practice Oncology* 4.9 (2007), pp. 516–525.
- [49] 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 protocols* 11.9 (2016), pp. 1757–1774.

- [50] Phuong Doan, Anzhelika Karjalainen, Jerome G. Chandraseelan, Ossi Sandberg, Olli Yli-Harja, Tomi Rosholm, Robert Franzen, Nuno R. Candeias, and Meenakshisundaram Kandhavelu. “Synthesis and biological screening for cytotoxic activity of N-substituted indolines and morpholines”. *European Journal of Medicinal Chemistry* 120.10 (2016), pp. 296–303. arXiv: [arXiv:1201.3109v1](#).
- [51] Vebjorn Ljosa, Peter D. Caie, Rob Ter Horst, Katherine L. Sokolnicki, Emma L. Jenkins, Sandeep Daya, Mark E. Roberts, Thouis R. Jones, Shantanu Singh, Auguste Genovesio, Paul A. Clemons, Neil O. Carragher, and Anne 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.
- [52] S. Singh, M. A. Bray, T. R. Jones, and A. E. Carpenter. “Pipeline for illumination correction of images for high-throughput microscopy”. *Journal of Microscopy* 256.3 (2014), pp. 231–236.
- [53] Nick Pawlowski, Juan C Caicedo, Shantanu Singh, Anne E Carpenter, and Amos Storkey. “Automating Morphological Profiling with Generic Deep Convolutional Networks”. *bioRxiv* (2016), pp. 4–8.
- [54] D. Michael Ando, Cory McLean, and Marc Berndl. “Improving Phenotypic Measurements in High-Content Imaging Screens”. *bioRxiv* (2017), p. 161422.
- [55] Qiaonan Duan, Corey Flynn, Mario Niepel, Marc Hafner, Jeremy L. Muhlich, Nicolas F. Fernandez, Andrew D. Rouillard, Christopher M. Tan, Edward Y. Chen, Todd R. Golub, Peter K. Sorger, Aravind Subramanian, and Avi Ma’Ayan. “LINCS Canvas Browser: Interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures”. *Nucleic Acids Research* 42.W1 (2014), W449–W460.
- [56] David Opitz and Richard Maclin. “Popular Ensemble Methods: An Empirical Study”. *Journal of Artificial Intelligence Research* 11 (1999), pp. 169–198. arXiv: [1106.0257](#).
- [57] Leo Breiman. “Bagging predictors”. *Machine Learning* 24.2 (1996), pp. 123–140.
- [58] Y Freund and R E Schapire. “Experiments with a new boosting algorithm”. *Machine Learning: Proc. of the 13th Int. Conf.* 1996, pp. 148–156.
- [59] F. Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain”. *Psychological Review* 65.6 (1958), pp. 386–408.
- [60] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. *Nature* 323.6088 (1986), pp. 533–536.
- [61] Randall C. O’Reilly, Dean Wyatte, Seth Herd, Brian Mingus, and David J. Jilk. “Recurrent processing during object recognition”. *Frontiers in Psychology* 4.APR (2013).
- [62] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2323.
- [63] Luis Perez and Jason Wang. “The Effectiveness of Data Augmentation in Image Classification using Deep Learning”. *ArXiv* (2017). arXiv: [1712.04621](#).

- [64] Marcus D. Bloice, Christof Stocker, and Andreas Holzinger. “Augmentor: An Image Augmentation Library for Machine Learning”. *ArXiv* (2017), pp. 1–5. arXiv: [1708.04680](#).
- [65] Garcia Gasulla Dario, Ferran Parés, Armand Vilalta, Jonatan Moreno, Eduard Ayguade, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. “On the behavior of convolutional nets for feature extraction”. *Journal of Artificial Intelligence Research* 61 (2018), pp. 563–592. arXiv: [1703.01127](#).
- [66] Alexander Kensert, Philip J Harrison, and Ola Spjuth. “Transfer learning with deep convolutional neural networks for classifying cellular morphological changes”. *bioRxiv* (2018), p. 345728.
- [67] Daehyun Lee and Kyungsik Myung. “Read my lips, login to the virtual world”. *2017 IEEE International Conference on Consumer Electronics, ICCE 2017* (2017), pp. 434–435.
- [68] Masahiro Tanaka, Raynard Bateman, Daniel Rauh, Eugeni Vaisberg, Shyam Ramachandani, Chao Zhang, Kirk C. Hansen, Alma L. Burlingame, Jay K. Trautman, Kevan M. Shokat, and Cynthia L. Adams. “An unbiased cell morphology-based screen for new, biologically active small molecules”. *PLoS Biology* 3.5 (2005), pp. 0764–0776.
- [69] Peter Horvath, Nathalie Aulner, Marc Bickle, Anthony M. Davies, Elaine Del Nery, Daniel Ebner, Maria C. Montoya, Päivi Östling, Vilja Pietiäinen, Leo S. Price, Spencer L. Shorte, Gerardo Turcatti, Carina Von Schantz, and Neil O. Carragher. “Screening out irrelevant cell-based models of disease”. *Nature Reviews Drug Discovery* 15.11 (2016), pp. 751–769.
- [70] Richard Herrmann, Walid Fayad, Stephan Schwarz, Maria Berndtsson, and Stig Linder. “Screening for compounds that induce apoptosis of cancer cells grown as multicellular spheroids”. *Journal of Biomolecular Screening* 13.1 (2008), pp. 1–8.
- [71] Isabelle Dufau, Céline Frongia, Flavie Sicard, Laure Dedieu, Pierre Cordelier, Frédéric Ausseil, Bernard Ducommun, and Annie Valette. “Multicellular tumor spheroid model to evaluate spatio-temporal dynamics effect of chemotherapeutics: Application to the gemcitabine/CHK1 inhibitor combination in pancreatic cancer”. *BMC Cancer* 12 (2012), pp. 1–11.
- [72] Jens M. Kelm, Nicholas E. Timmins, Catherine J. Brown, Martin Fussenegger, and Lars K. Nielsen. “Method for generation of homogeneous multicellular tumor spheroids applicable to a wide variety of cell types”. *Biotechnology and Bioengineering* 83.2 (2003), pp. 173–180.
- [73] Rehan Akbani, Karl-Friedrich Becker, Neil Carragher, Ted Goldstein, Leanne de Koning, Ulrike Korf, Lance Liotta, Gordon B. Mills, Satoshi S. Nishizuka, Michael Pawlak, Emanuel F. Petricoin, Harvey B. Pollard, Bryan Serrels, and Jingchun Zhu. “Realizing the Promise of Reverse Phase Protein Arrays for Clinical, Translational, and Basic Research: A Workshop Report”. *Molecular & Cellular Proteomics* 13.7 (2014), pp. 1625–1643.
- [74] Ted T. Ashburn and Karl B. Thor. “Drug repositioning: Identifying and developing new uses for existing drugs”. *Nature Reviews Drug Discovery* 3.8 (2004), pp. 673–683.

- [75] Rachel Pounds, Sarah Leonard, Christopher Dawson, and Sean Kehoe. “Repurposing itraconazole for the treatment of cancer (review)”. *Oncology Letters* 14.3 (2017), pp. 2587–2597.
- [76] A. Kummel, H. Gubler, P. Gehin, M. Beibel, D. Gabriel, and C. N. Parker. “Integration of Multiple Readouts into the Z’ Factor for Assay Quality Assessment”. *Journal of Biomolecular Screening* 15.1 (2010), pp. 95–101.
- [77] Jennifer Laurent, Céline Frongia, Martine Cazales, Odile Mondesert, Bernard Ducommun, and Valérie Lobjois. “Multicellular tumor spheroid models to explore cell cycle checkpoints in 3D”. *BMC Cancer* 13 (2013).
- [78] Motoko Izumiya, Ayano Kabashima, Hajime Higuchi, Toru Igarashi, Gen Sakai, Hideko Iizuka, Shoko Nakamura, Masayuki Adachi, Yasuo Hamamoto, Shinsuke Funakoshi, Hiromasa Takaishi, and Toshifumi Hibi. “Chemoresistance is associated with cancer stem cell-like properties and epithelial-to-mesenchymal transition in pancreatic cancer cells”. *Anticancer Research* 32.9 (2012), pp. 3847–3853.
- [79] M. Farmakovskaya, N. Khromova, V. Rybko, V. Dugina, B. Kopnin, and P. Kopnin. “E-Cadherin repression increases amount of cancer stem cells in human A549 lung adenocarcinoma and stimulates tumor growth”. *Cell Cycle* 15.8 (2016), pp. 1084–1092.
- [80] Hector Biliran, Yong Wang, Sanjeev Banerjee, Haiming Xu, Henry Heng, Archana Thakur, Aliccia Bollig, Fazlul H. Sarkar, and Joshua D. Liao. “Overexpression of cyclin D1 promotes tumor cell growth and confers resistance to cisplatin-mediated apoptosis in an elastase-myc transgene-expressing pancreatic tumor cell line”. *Clinical Cancer Research* 11.16 (2005), pp. 6075–6086.
- [81] Sophie Bustany, Jérôme Bourgeais, Guergana Tchakarska, Simon Body, Olivier Hérault, Fabrice Gouilleux, and Brigitte Sola. “Cyclin D1 unbalances the redox status controlling cell adhesion, migration, and drug resistance in myeloma cells.” *Oncotarget* 7.29 (2016), pp. 1–11.
- [82] Siyuan Zhang, Wen Chien Huang, Ping Li, Hua Guo, Say Bee Poh, Samuel W. Brady, Yan Xiong, Ling Ming Tseng, Shau Hsuan Li, Zhaoxi Ding, Aysegul A. Sahin, Francisco J. Esteva, Gabriel N. Hortobagyi, and Dihua Yu. “Combating trastuzumab resistance by targeting SRC, a common node downstream of multiple resistance pathways”. *Nature Medicine* 17.4 (2011), pp. 461–469.
- [83] M Tien Kuo. “Roles of multidrug resistance genes in breast cancer chemoresistance.” *Advances in experimental medicine and biology* 608 (2007), pp. 23–30.
- [84] Tae Hee Kim, Hyo Kyeong Kim, and Eun Sook Hwang. “Novel anti-adipogenic activity of anti-malarial amodiaquine through suppression of PPAR γ activity”. *Archives of Pharmacal Research* 40.11 (2017), pp. 1336–1343.
- [85] Mandy Juarez, Alejandro Schcolnik-Cabrera, and Alfonso Dueñas-Gonzalez. “The multi-targeted drug ivermectin: from an antiparasitic agent to a repositioned cancer drug.” *American journal of cancer research* 8.2 (2018), pp. 317–331.

- [86] Gao Rong Wu, Bing Xu, Yu Qin Yang, Xin Yu Zhang, Kang Fang, Tao Ma, Hui Wang, Nan Nan Xue, Meng Chen, Wen Bo Guo, Xiao Hui Jia, Peng Long Wang, and Hai Min Lei. "Synthesis and biological evaluation of podophyllotoxin derivatives as selective antitumor agents". *European Journal of Medicinal Chemistry* 155 (2018), pp. 183–196.
- [87] R. C. Dietrich, L. N. Alberca, M. D. Ruiz, P. H. Palestro, C. Carrillo, A. Talevi, and L. Gavernet. "Identification of cisapride as new inhibitor of putrescine uptake in *Trypanosoma cruzi* by combined ligand- and structure-based virtual screening". *European Journal of Medicinal Chemistry* 149 (2018), pp. 22–29.
- [88] Hanxian Zeng, Sijie Liu, Pengfei Wang, Xiyang Qu, Haiyan Ji, Xiaohui Wang, Xiaoli Zhu, Zhishuo Song, Xinyi Yang, Zhongjun Ma, and Huanzhang Zhu. "Dilazep synergistically reactivates latent HIV-1 in latently infected cells". *Molecular Biology Reports* 41.11 (2014), pp. 7697–7704.
- [89] Keiichiro Hayashi, Hiroyuki Michiue, Hiroshi Yamada, Katsuyoshi Takata, Hiroki Nakayama, Fan Yan Wei, Atsushi Fujimura, Hiroshi Tazawa, Akira Asai, Naohisa Ogo, Hiroyuki Miyachi, Tei Ichi Nishiki, Kazuhito Tomizawa, Kohji Takei, and Hideki Matsui. "Fluvoxamine, an anti-depressant, inhibits human glioblastoma invasion by disrupting actin polymerization". *Scientific Reports* 6.July 2015 (2016), pp. 1–12.
- [90] C.-K. Su, C.-T. Chou, K.-L. Lin, W.-Z. Liang, J.-S. Cheng, H.-T. Chang, I.-S. Chen, T. Lu, C.-C. Kuo, C.-C. Yu, P. Shieh, D.-H. Kuo, F.-A. Chen, and C.-R. Jan. "Effect of protriptyline on $[Ca^{2+}]_i$ and viability in MG63 human osteosarcoma cells". *Toxicology Mechanisms and Methods* 26.8 (2016).
- [91] Joseph P. Steiner, Muznabanu Bachani, Brett Wolfson-Stofko, Myoung Hwa Lee, Tonguang Wang, Guanhan Li, Wenxue Li, David Strayer, Norman J. Haughey, and Avindra Nath. "Interaction of Paroxetine with Mitochondrial Proteins Mediates Neuroprotection". *Neurotherapeutics* 12.1 (2015), pp. 200–216.
- [92] Kelly A. Meulendyke, Suzanne E. Queen, Elizabeth L. Engle, Erin N. Shirk, Jiayang Liu, Joseph P. Steiner, Avindra Nath, Patrick M. Tarwater, David R. Graham, Joseph L. Mankowski, and M. Christine Zink. "Combination fluconazole/paroxetine treatment is neuroprotective despite ongoing neuroinflammation and viral replication in an SIV model of HIV neurological disease". *Journal of NeuroVirology* 20.6 (2014), pp. 591–602.
- [93] V A Rigas, H Van Vunakis, and L Levine. "The effect of phenothiazines and their metabolites on prostaglandin production by rat basophilic leukemia cells in culture." *Prostaglandins and medicine* 7.2 (1981), pp. 183–93.
- [94] Ji Hu Zhang, Thomas D Y 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), pp. 67–73.
- [95] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. "Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments". *FEBS Letters* 573.1-3 (2004), pp. 83–92.

- [96] Leland McInnes and John Healy. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. *Arxiv* (2018), pp. 1–18. arXiv: [1802.03426](#).
- [97] Frank K. Brown. “Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery.” *Annual Reports in Medicinal Chemistry* 33.C (1998), pp. 375–384.
- [98] Louis C. Ray and Russell A. Kirsch. “Finding chemical records by digital computers”. *Science* 126.3278 (1957), pp. 814–819.
- [99] Corwin Hansch, Peyton P. Maloney, Toshio Fujita, and Robert M. Muir. “Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients”. *Nature* 194.4824 (1962), pp. 178–180.
- [100] P. A. Clemons, J. A. Wilson, V. Dancik, S. Muller, H. A. Carrinski, B. K. Wagner, A. N. Koehler, and S. L. Schreiber. “Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections”. *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 6817–6822.
- [101] Anne Mai Wassermann, Eugen Lounkine, Dominic Hoepfner, Gaelle Le Goff, Frederick J. King, Christian Studer, John M. Peltier, Melissa L. Grippo, Vivian Prindle, Jianshi Tao, Ansgar Schuffenhauer, Iain M. Wallace, Shanni Chen, Philipp Krastel, Amanda Cobos-Correa, Christian N. Parker, John W. Davies, and Meir Glick. “Dark chemical matter as a promising starting point for drug lead discovery”. *Nature Chemical Biology* 11.12 (2015), pp. 958–966.
- [102] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. “A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction”. *BMC Bioinformatics* 17.1 (2016), pp. 1–11.
- [103] David Rogers and Mathew Hahn. “Extended-connectivity fingerprints”. *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754.
- [104] Adrian M. Schreyer and Tom Blundell. “USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints”. *Journal of Cheminformatics* 4.11 (2012), p. 1.
- [105] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. “Molecular graph convolutions: moving beyond fingerprints”. *Journal of Computer-Aided Molecular Design* 30.8 (2016), pp. 595–608. arXiv: [1603.00856](#).
- [106] Evan N. Feinberg, Debnil Sur, Brooke E. Husic, Doris Mai, Yang Li, Jianyi Yang, Bharath Ramsundar, and Vijay S. Pande. “Spatial Graph Convolutions for Drug Discovery”. *ArXiv* (2018), pp. 1–14. arXiv: [1803.04465](#).
- [107] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. “Drug Similarity Integration Through Attentive Multi-view Graph Auto-Encoders”. *ArXiv* (2018). arXiv: [1804.10850](#).
- [108] Shengchao Liu, Thevaa Chandereng, and Yingyu Liang. “N-Gram Graph, A Novel Molecule Representation”. *ArXiv* (2018). arXiv: [1806.09206](#).

- [109] M. J. Wawer, K. Li, S. M. Gustafsdottir, V. Ljosa, N. E. Bodycombe, M. A. Marton, K. L. Sokolnicki, M.-A. Bray, M. M. Kemp, E. Winchester, B. Taylor, G. B. Grant, C. S.-Y. Hon, J. R. Duvall, J. A. Wilson, J. A. Bittker, V. Dan ik, R. Narayan, A. Subramanian, W. Winckler, T. R. Golub, A. E. Carpenter, A. F. Shamji, S. L. Schreiber, and P. A. Clemons. "Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling". *Proceedings of the National Academy of Sciences* 111.30 (2014), pp. 10911–10916.
- [110] Mathias J. Wawer, David E. Jaramillo, Vlado Dančík, Daniel M. Fass, Stephen J. Haggarty, Alykhan F. Shamji, Bridget K. Wagner, Stuart L. Schreiber, and Paul A. Clemons. "Automated structure-activity relationship mining: Connecting chemical structure to biological profiles". *Journal of Biomolecular Screening* 19.5 (2014), pp. 738–748.
- [111] L J P Van Der Maaten and G E Hinton. "Visualizing high-dimensional data using t-sne". *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.
- [112] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". *ACS Central Science* 4.2 (2018), pp. 268–276. arXiv: [1610.02415](https://arxiv.org/abs/1610.02415).
- [113] Darko Butina. "Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets". *Journal of Chemical Information and Computer Sciences* 39.4 (1999), pp. 747–750.
- [114] Nathan Mantel. "The Detection of Disease Clustering and a Generalized Regression Approach". *Cancer Research* 27.2 (1967), pp. 209–220.
- [115] Robert D. Finn, Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin Yu Chang, Zsuzsanna Dosztanyi, Sara El-Gebali, Matthew Fraser, Julian Gough, David Haft, Gemma L. Holliday, Hongzhan Huang, Xiaosong Huang, Ivica Letunic, Rodrigo Lopez, Shennan Lu, Aron Marchler-Bauer, Huaiyu Mi, Jaina Mistry, Darren A. Natale, Marco Necci, Gift Nuka, Christine A. Orengo, Youngmi Park, Sebastien Pesseat, Damiano Piovesan, Simon C. Potter, Neil D. Rawlings, Nicole Redaschi, Lorna Richardson, Catherine Rivoire, Amaia Sangrador-Vegas, Christian Sigrist, Ian Sillitoe, Ben Smithers, Silvano Squizzato, Granger Sutton, Narmada Thanki, Paul D. Thomas, Silvio C.E. Tosatto, Cathy H. Wu, Ioannis Xenarios, Lai Su Yeh, Siew Yit Young, and Alex L. Mitchell. "InterPro in 2017-beyond protein family and domain annotations". *Nucleic Acids Research* 45.D1 (2017), pp. D190–D199.
- [116] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. "Quantifying the chemical beauty of drugs". *Nature Chemistry* 4.2 (2012), pp. 90–98.

- [117] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. *Nature Protocols* 4.1 (2009), pp. 44–57.
- [118] Simon Gordonov, Mun Kyung Hwang, Alan Wells, Frank B. Gertler, Douglas A. Lauffenburger, and Mark Bathe. “Time series modeling of live-cell shape dynamics for image-based phenotypic profiling”. *Integrative Biology (United Kingdom)* 8.1 (2016), pp. 73–90. arXiv: [15334406](#).
- [119] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. *ArXiv* (2015), pp. 1–8. arXiv: [1505.04597](#).
- [120] Sajith Kecheril Sadanandan, Johan Karlsson, and Carolina Wählby. “Spheroid Segmentation Using Multiscale Deep Adversarial Networks”. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017* 2018-Janua (2018), pp. 36–41.
- [121] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. “How transferable are features in deep neural networks?” *ArXiv* 27 (2014). arXiv: [1411.1792](#).
- [122] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. “CNN features off-the-shelf: An astounding baseline for recognition”. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2014), pp. 512–519. arXiv: [1403.6382](#).
- [123] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge”. *International Journal of Computer Vision* 115.3 (2015), pp. 211–252. arXiv: [1409.0575](#).
- [124] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott a Armstrong, Stephen J Haggarty, Paul a Clemons, Ru Wei, and Steven a Carr. “The Connectivity Map : Using”. *Science* 313.September (2006), pp. 1929–1935.
- [125] Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. “A Next Generation Connec-

tivity Map: L1000 Platform and the First 1,000,000 Profiles". *Cell* 171.6 (2017), 1437–1452.e17. arXiv: [/biorxiv.org/content/early/2017/05/10/136168](https://doi.org/10.1101/136168) [http:].

This appendix contains journal articles and a book chapter published during my PhD.

1. “Next Generation Phenotypic Screening”; S.J Warchal, A Unciti-Broceta, N.O Carragher; Future Medicinal Chemistry; 2016; DOI:10.4155/fmc-2016-0025
2. “Development of the Theta Comparative Cell Scoring Method to Quantify Diverse Phenotypic Responses Between Distinct Cell Types”; S.J Warchal, J.C Dawson, N.O Carragher; Assay and Drug Development Technologies; 2016; DOI:10.1089/adt.2016.730
3. “Data-analysis strategies for image-based cell profiling”; J.C Caicedo, S Cooper, F. Heigwer, S.J Warchal, P Qui, C Molnar, A Vasilevich, J.D Barry, H.S Bansal, O Kraus, M Wawer, L Paavolainen, M.D Herrmann, M Rohban, J Hung, H Hennig, J Concannon, I Smith, P Clemons, S Singh, P Rees, P Horvath, R Linington, A.E Carpenter; Nature Methods; 2017; doi:10.1038/nmeth.4397
4. “High-Dimensional Profiling: The Theta Comparative Cell Scoring Method”; S.J Warchal, J.C Dawson, N.O Carragher; Phenotypic Screening: Methods and Protocols, Methods in Molecular Biology; 2018; DOI:10.1007/978-1-4939-7847-2_13

Publication submitted but not included in the appendix:

1. “Evaluation of Machine Learning Classifiers to Predict Compound Mechanism of Action when Transferred Across Distinct Cell-Lines”; S.J Warchal, J.C Dawson, N.O Carragher; SLAS Discovery; **under review**

Next-generation phenotypic screening

Phenotypic drug discovery (PDD) strategies are defined by screening and selection of hit or lead compounds based on quantifiable phenotypic endpoints without prior knowledge of the drug target. We outline the challenges associated with traditional phenotypic screening strategies and propose solutions and new opportunities to be gained by adopting modern PDD technologies. We highlight both historical and recent examples of approved drugs and new drug candidates discovered by modern phenotypic screening. Finally, we offer a prospective view of a new era of PDD underpinned by a wealth of technology advances in the areas of *in vitro* model development, high-content imaging and image informatics, mechanism-of-action profiling and target deconvolution.

First draft submitted: 29 January 2016; Accepted for publication: 4 April 2016; Published online: 30 June 2016

Keywords: • high content • image informatics • mechanism of action • phenotypic screening • target deconvolution • transcriptomics • pathway profiling

Background

During the past 5 years, the drug discovery field has witnessed a re-emerging interest in phenotypic drug discovery (PDD) strategies and increased research activity in phenotypic assay development and screening. PDD describes the screening and selection of hit or lead compounds based on quantifiable phenotypic endpoints from cell-based assays or model organisms without prior knowledge of the drug target. The renewed interest in phenotypic screening may be attributed to several factors including: the demand to identify high-value novel drug targets to feed contemporary target-directed drug discovery (TDD) capabilities and commercial drug discovery pipelines; high attrition rates in late stage clinical development and an overall decrease in pharmaceutical R&D productivity, while not directly attributed to limitations of TDD, nevertheless, correlate with the widespread adoption of the TDD operating model in favor of PDD strate-

gies [1–5]; significant duplication of effort and focus upon a relatively small number of well-characterized targets across industrial and academic drug discovery groups; urgent unmet medical need in complex human conditions such as heterogeneous solid cancers and neurodegeneration, where target biology is poorly understood; recent retrospective analysis of all drugs approved by the US FDA since 1999 indicating significant success rates in development of novel, first-in-class, small-molecule drugs by PDD approaches [6–8].

While the three recent retrospective studies of drug approval rates present discrepancies in the number of drug approvals attributed to PDD and TDD strategies, primarily because of differences in terminology, disease area focus and period of analysis, all three studies demonstrate that PDD approaches are providing a significant contribution to clinical approval rates of first-in-class drugs [6–8]. This recent clinical success of PDD is con-

Scott J Warchal¹,
Asier Unciti-Broceta¹
& Neil O Carragher^{*,1}

¹Cancer Research UK Edinburgh Centre,
Institute of Genetics & Molecular
Medicine, University of Edinburgh, Crewe
Road South, Edinburgh EH4 2XR, UK

*Author for correspondence:

Tel.: +44 0131 651 8671

n.carragher@ed.ac.uk

FUTURE
SCIENCE

part of

fsg

sidered by many as remarkable given the relatively low investments in PDD in comparison to TDD by translational funding bodies in academia and industry over the past three decades. PDD, however, does not represent a new drug discovery strategy and was indeed the preferred drug discovery approach prior to increased understanding of human disease at the genetic level and the emergence of molecular biology techniques, which advanced elegant molecular pharmacology studies and high-throughput screening of specific targets [2]. Traditional PDD approaches have utilized a variety of biological model systems such as *in vivo* physiological and behavioral models, *ex vivo* tissue-based assays and basic *in vitro* cellular assays to guide drug development. While many drugs successfully used in the clinic today were discovered using such early PDD approaches, traditional PDD methods were laborious and did not provide ample mechanistic information and thus tend to favor the discovery of less selective agents including cytotoxics rather than novel classes of targeted therapies. Given the duration of time between early-stage drug discovery and clinical approval, many of the example first-in-class medicines attributed to PDD described in the three recent retrospective articles [6–8] are somewhat historical. They utilize, by modern standards, rudimentary phenotypic assays, and thus such retrospective analysis may indeed underestimate the true value of modern phenotypic screening strategies, with regard to identifying novel targets and translation into clinical success.

A significant driving force behind the resurgence of PDD may be attributed to substantial technology developments across several inter-related areas, which advance the PDD paradigm. Such advances include more sophisticated cell-based and small model organism-based automated screening platforms [9–11]. Advances in the

development of more complex and more disease-relevant phenotypic assays incorporating: multicellular co-cultures, 3D models, patient-derived primary and Induced Pluripotent Stem Cell (iPSC) models, including gene-edited and isogenic controls, which recapitulate key disease driver mutations, are all well-placed to advance the molecular and pathophysiological relevance of phenotypic screening assays. Improvements in cell-based assay technologies are further complemented by advances in target deconvolution strategies including affinity mass spectrometry, cellular thermal shift assays and cDNA expression microarray technologies among others (Box 1) [12–14]. Also the development of new methodologies, which enable profiling drug mechanism-of-action (MOA) in complex biological samples at genomic, proteomic and phenotypic levels at scale [15–17], supports informed mechanistic classification and triaging of phenotypic hits to assist further target deconvolution or progress preclinical development of phenotypic hits in the absence of target knowledge.

In this article, we attempt to address some common misconceptions and challenges associated with phenotypic screening. We highlight both historical and recent success stories of approved drugs and new drug candidates discovered by PDD. We describe some of the challenges and pitfalls of poorly designed phenotypic screening and target deconvolution strategies and how these may be resolved by the application of new technologies. We place specific emphasis upon the evolution of new gene transcription, pathway profiling and multiparametric high-content screening technologies, which support more advanced phenotypic screening and MOA studies. We provide specific examples and discuss the advantages and limitations of each new approach. Finally, we conclude by discussing how the

Box 1. Target deconvolution methods

Chemical proteomics

- Affinity chromatography and mass spectrometry [12,18–20]
- Quantitative proteomic and silac labeling [21]
- Thermostability shift assays: *in vitro* and in cells [13,22]

Expression cloning

- Phage display [23]
- Yeast three-hybrid assays [24,25]
- cDNA cell microarray [26]

Genetic-based screens

- Yeast deletion collections [27]
- Haploinsufficiency profiling [28]
- Resistance screens combined with Next-Generation Sequencing (NGS) profiling of resistant clones [29]
- Small model organism knock-out (KO) and genetic mutant collections (under development in Zebrafish, *Caenorhabditis elegans* and *Drosophila*)
- Modifier screens: si/shRNA or CRISPR-Cas9 library screens to identify modulators of small-molecule activity [30,31]
- Activity-based protein profiling [32]

This is a nonexhaustive list of target deconvolution methods and reflects some of the most common approaches selected by the authors.

combination of new technology developments such as more advanced primary and induced pluripotent stem cell (iPSC) culture techniques, gene editing, high-throughput gene transcription, pathway profiling and multiparametric high-content screening technologies are well-placed to advance phenotypic screening toward increased success across multiple disease areas.

Historical examples of drugs discovered by phenotypic screening

Comprehensive discussion of approved drugs originating from PDD strategies have been reviewed previously [6–8,33]. In this article, we highlight specific examples of PDD drugs currently used in the clinics, which challenge conservative views on the necessity for target deconvolution to progress candidate drugs through clinical development. We further describe new lead compounds and candidate drugs discovered by more modern phenotypic screening strategies, which guide chemical design toward specific MOA and which can integrate with ligand-based drug design and TDD strategies to develop highly potent and selective lead compounds and drug candidates.

Metformin (Figure 1, structure 1) belongs to the biguanide class of compounds and represents the first-line standard-of-care therapy for Type 2 diabetes by virtue of its confirmed physiological effects upon decreased glucose production by the liver. Approved in Europe in 1957, metformin has been used for decades as a safe and efficacious medicine to manage the morbidity and mortality associated with Type 2 diabetes and represents a core component of new drug combination therapies for diabetes [34,35]. It is, however, only recently that the MOA by which metformin regulates glucose levels has been revealed. In mouse hepatocytes, metformin leads to the accumulation of AMP and related nucleotides, which inhibit adenylate cyclase, reduces levels of cyclic AMP and PKA activity, abrogates phosphorylation of downstream protein targets of PKA and blocks glucagon-dependent glucose output from hepatocytes [36]. These new insights into metformin MOA will most likely pave the way to development of novel antidiabetic drugs.

Further examples of approved drugs derived from compound library screening in phenotypic models were the molecular target of the drug is not known include, daptomycin (Figure 1, structure 2), a naturally occurring antibiotic targeting cell membrane function of Gram-positive bacteria to treat systemic and life-threatening infections [37]. Pemirolast (Figure 1, structure 3) is an antiallergic drug therapy that is proposed to work through suppression of mast cell degranulation, histamine release and eosinophil activation, although precise target mechanism remains to be confirmed [38]. Rufinamide (Figure 1, structure 4) is a tri-

azole derivative used as an anticonvulsant/antiepileptic medication to treat several seizure disorders including Lennox–Gastaut syndrome [39]. The specific molecular target or targets of rufinamide remain to be established.

These examples, described in Figure 1, serve to highlight that if target deconvolution was always a prerequisite for drug development, valuable treatments for such serious human disorders would not have been developed and would not progress the next generation of therapies for many of the most serious and life-threatening conditions.

Sirolimus also known as rapamycin, (Figure 1, structure 5) is a macrolide derived from the bacterium, *Streptomyces hygroscopicus* and discovered through PDD to possess immunosuppressive, antifungal and anticancer properties [40–42]. The National Cancer Institute (NCI) Developmental Therapeutics Program demonstrated that rapamycin inhibited cell growth in panels of tumor cell lines [43]. Subsequent mechanistic studies indicated that the MOA was mediated through inhibition of a serine/threonine protein kinase critical to cell growth, proliferation and survival, the subsequently named mammalian target of rapamycin (mTOR) [44]. Inhibition is mediated through forming a complex between rapamycin bound FKBP12 with mTORC1 [45]. It is worth noting that while rapamycin has been clinically approved, it violates the Lipinski rule of 5 defining optimal lead and drug like properties and thus may never have been developed through a conventional small molecule drug discovery program. The development of rapamycin through PDD and subsequent understanding of mTOR signaling and the target of rapamycin within preclinical and clinical settings provided important target validation data to support the development of several rapamycin analogs known as rapalogs and second generation ATP-competitive mTOR kinase inhibitors targeting mTOR catalytic activity associated with both mTORC1 and mTORC2 complexes [46,47].

Recent examples of phenotypic screening outcomes

We recently reported application of an iterative process consisting of ligand-based design and phenotypic screening of focused chemical libraries to develop novel antiproliferative inhibitors. The strategy employs promiscuous kinase inhibitors as templates to design high-quality small-molecule collections to facilitate the concurrent search for enhanced physicochemical properties and novel pharmacological features. Using this method, target deconvolution of identified hits and leads is largely simplified (for example, focused kinome screening), thereby assisting subsequent lead optimization campaigns [47]. The application of this strategy resulted in the discovery of the first subnanomolar SRC

inhibitor with 1000-fold selectivity over ABL [48] and highly potent dual mTORC1 and mTORC2 inhibitors (eCF309 – Figure 2, structure 6) with high selectivity over other family kinases [47]. A further example of a highly selective kinase inhibitor derived from a phenotypic screen is the allosteric inhibitor of MEK, Trametinib (Figure 2, structure 7), which was initially identified by screening for increased mRNA expression of the cyclin-dependent kinase inhibitor p15 and cell proliferation [49].

Modern high-content phenotypic screening assays, which quantify specific functional endpoints, can also be used to identify compounds with precise target MOA such as the identification of novel Eg5 kinesin inhibitors which induce the monopolar and monaster phenotype [50]. Similar approaches were used to discover and confirm the MOA of second generation Eg5 kinesin inhibitors (AZD4877), which have progressed into clinical development (Figure 2, structure 8) [51,52]. Olesoxime (Figure 2, structure 9) was originally discovered by performing a screen of 40,000 small mol-

ecules in an *in vitro* cell-based assay to identify compounds capable of preventing motor neuron cell death in the absence of trophic support [53].

The historical exemplars of approved drugs discovered by PDD chosen (metformin, daptomycin, pemirolast and rufinamide, Figure 1, structures 1–4) highlight that clinically useful and safe medicines can be developed without precise knowledge of the target mechanism. Examples of approved drugs discovered by PDD also serve to highlight that compound structures, which lie out with conventional characteristics of drug likeness can be proved effective in patients (rapamycin, Figure 1, structure 5). Further recent examples of approved drugs discovered by phenotypic screening demonstrate that highly selective targeted therapies can be discovered by phenotypic screening. These examples include drugs targeting ubiquitously expressed regulators of critical cellular functions such as protein synthesis (mTOR inhibitors, eCF309, Figure 2, structure 6), MAPK/ERK signaling (trametinib, Figure 2, structure 7), Eg5 kinesin and mitotic spindle assembly

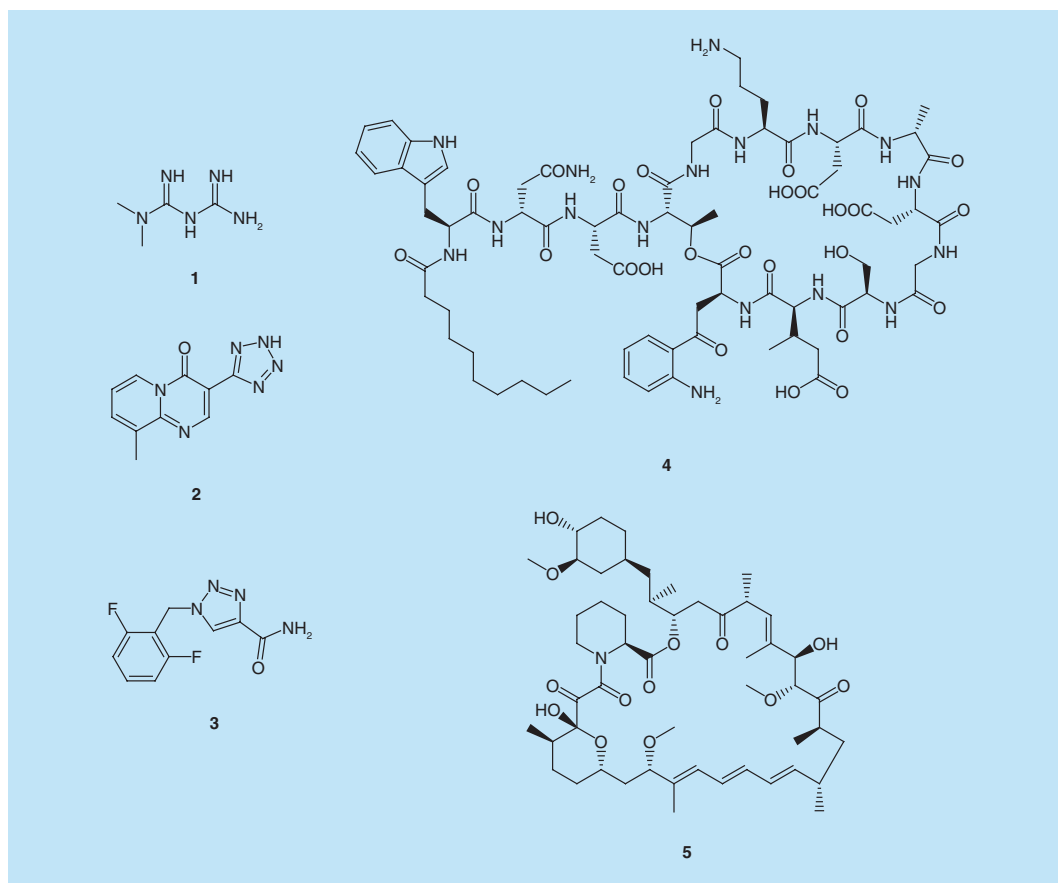


Figure 1. Compound structures of historical examples of drugs discovered by phenotypic drug discovery.
1: Metformin; 2: Daptomycin; 3: Pemirolast; 4: Rufinamide; 5: Rapamycin/silormycin.

(AZD4877, **Figure 2**, structure 8) and mitochondrial function (olesoxime, **Figure 2**, structure 9). Such target classes are unlikely to be prioritized by current drug target review or translational funding committees using conventional target selection criteria to support investment in novel therapeutic targets. Thus, PDD approaches have resulted in the development of many clinically valuable drugs, which would not be developed by TDD programs. The recent examples, which we highlight in **Figures 1** and **2** represent a small number of drugs and drug candidates discovered by PDD. For more comprehensive listings of drugs approved by PDD, we direct readers to three recent review articles providing in depth description on the origins of drugs discovered by PDD [6,8,33].

While the overwhelming development of modern targeted therapies has been derived from TDD approaches, these recent examples highlight how advanced phenotypic screening can efficiently direct structure–activity relationships (SAR) and identify novel chemotypes with high potency and selectivity. The above examples further highlight how PDD and TDD approaches complement one another and how new opportunities for combining PDD and TDD strategies are supported by more advanced phenotypic screening, MOA profiling and target deconvolution technologies.

Pitfalls of poorly designed phenotypic screens & black box assays

The phrase ‘phenotypic screening’ is a broadly used term to describe the extraction of quantifiable read-outs of biological relevance from any cell-, tissue- or organism-based system suitable for medium- to high-throughput chemical or functional genomic screening in a target agnostic fashion. Phenotypic screens can range from simplistic 2D cell line viability or reporter assays/pathway screens to more complex multicellular, 3D and multiparametric assays. Phenotypic screening is applied in both the industrial and academic research settings to support functional genomic studies, discover novel candidate drugs and/or useful chemical probes and pharmacological tools for further exploring biology. Thus, analysis and debate on success and challenges of phenotypic screening and target deconvolution strategies must be placed into appropriate context of the value and information provided by the primary phenotypic screen. Traditional single endpoint cell viability and reporter-based cellular assays provide limited information of drug MOA and thus limited opportunity to triage and precisely direct further development of phenotypic hits prior to target deconvolution. Such traditional phenotypic assays, which provide limited mechanistic data, so called ‘black box’ assays may

amplify phenotypic screening challenges and common pitfalls such as, preferential selection of cytotoxic compounds, pan-assay interference compounds or PAINS and sharp activity cliffs, which confound SAR studies. Such pitfalls can largely be avoided by development of information-rich phenotypic screening assays such as multicellular co-culture assays to discriminate phenotypic effect between distinct cell types or multiparametric high-content phenotypic profiling assays, which provide more informative insights into cellular pharmacology. Such high-content assays can classify MOA based upon specific cell targeting or by phenotypic fingerprint similarity with compounds of known MOA and target binding [54]. High-content screening in co-culture assays incorporating target and nontarget cell types may help guide hit selection and chemical design away from toxicity toward enhanced efficacy and novel target space within a single primary high-throughput phenotypic screen [55]. While many ‘black box’ phenotypic assays represented the state-of-the-art at the time of their development and have had many notable successes in supporting the development of novel drugs, including many of the examples described in **Figures 1** and **2**. In contrast to modern high-content phenotypic assays, ‘black box’ assays provide limited opportunity to design screens, which guide selection of hits and leads toward increased therapeutic index and novel phenotypic and target space.

Several review articles, editorials and commentators also suggest that phenotypic screening may help reduce high attrition rates observed during late-stage clinical development specifically the high failure rate observed during Phase II clinical trials resulting from lack of efficacy [4,56,57]. However, the ability of a phenotypic screen to reduce attrition from poor efficacy is directly related to the ability of the primary phenotypic screening assays and any secondary phenotypic assays used for hit selection to predict clinical outcomes. For many complex diseases it is unlikely that the primary screen will recapitulate the full complexity of human disease. Thus, phenotypic screening assays must be developed that ask specific clinical questions or recapitulate key segments of disease pathophysiology to inform subsequent decision-making and effectively guide the next stages of preclinical development and validation. This approach is supported by recent advances in cell-based assay methodology and technologies. Examples in the oncology area include techniques for culturing glioma progenitor cells representing the cancer stem cell niche [58], 3D tumor and fibroblast co-culture organotypic assays, which recapitulate the dense fibrosis and poor drug penetration of poorly vascularized tumors [59] and 3D tumor spheroid cultures, which recapitulate the hypoxic and host cell stromal microenvironment

of many tumors [60]. Screening phenotypic hits across suites of such assays raises the bar with regard to early assessment of the clinical relevance of hit and lead compounds, and also informs subsequent preclinical and clinical development strategies. Ongoing advances in iPSc, gene editing and microfluidic technologies support the development of more physiologically relevant assays across disease areas further advancing more robust approaches to prioritizing phenotypic hits.

Challenges in target deconvolution

An emerging simplistic view of phenotypic screening is that it is an effective strategy for identification of new therapeutic targets from physiological-based models to feed TDD. However, as discussed above, it is unlikely that a primary phenotypic screening assay by itself will predict clinical efficacy and it is also unlikely that initial chemical hits from a large chemical library phenotypic screen will have sufficient potency or selectivity to support rapid and robust target deconvolution. Caution should, therefore, be taken to ensure that poorly designed phenotypic screening and target deconvolution strategies do not create expensive new drug discovery bottlenecks in target deconvolution and further investment of significant chemistry resources on poorly validated targets.

As discussed below, target deconvolution is a challeng-

ing and expensive endeavor with limited success rates, we therefore propose that the pathway from phenotypic screening to target deconvolution should not directly follow one another. Rather phenotypic hits should be carefully triaged through increasingly more complex and disease relevant secondary phenotypic assays to build further confidence in the translational potential of the phenotypic hit. Further panel screening across *in vitro* toxicity assays, physiologically relevant assay formats and phenotypic profiling against reference compound libraries will help select the most novel and desirable compounds for target deconvolution. Further experimental medicine studies, including transcriptomic and proteomic analysis, across genetically defined cell assays to prioritize biomarker and drug combination strategies support subsequent chemical optimization using specific pathway reporter assays, target deconvolution and preclinical development. This more in-depth biological investigation will then shift the PDD bottleneck from target deconvolution toward increased disease relevance, novelty, safety and hopefully improved efficacy and drug discovery productivity.

Target deconvolution

Following careful triaging and selection of high-value lead compounds identified by PDD, a number of dis-

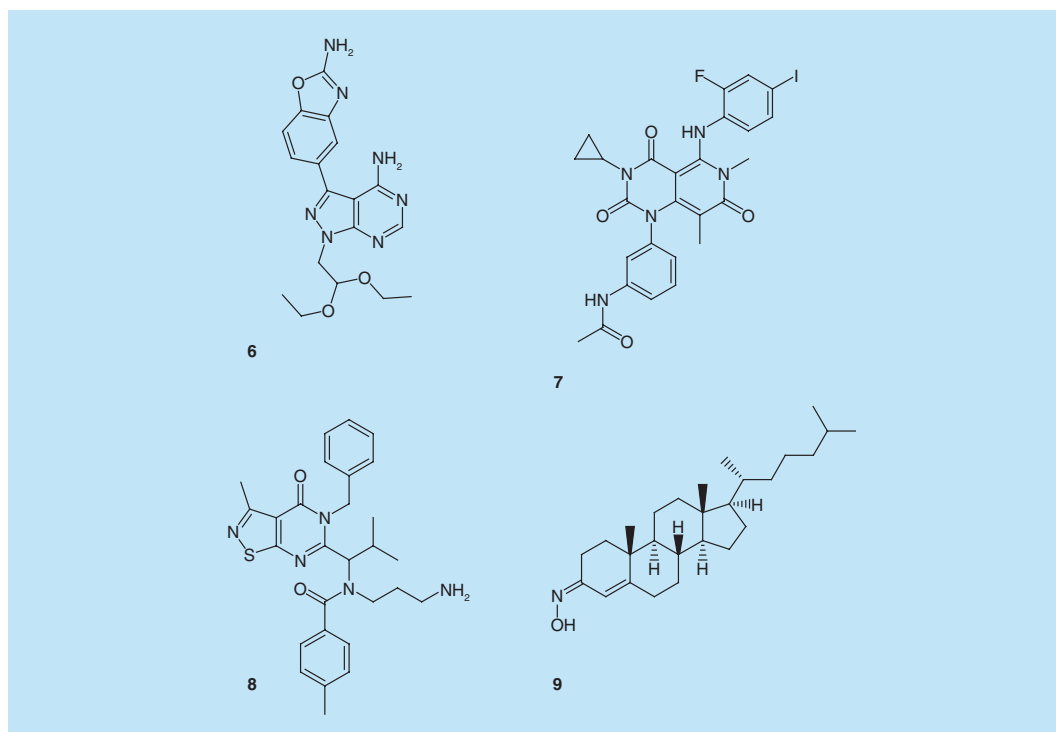


Figure 2. Compound structures from recent examples of phenotypic drug discovery and modern phenotypic screening. 6: eCF309; 7: Trametinib; 8: AZD4877; 9: Olesoxime.

tinct and complementary target deconvolution strategies can be employed (Box 1). Many early target deconvolution studies relied heavily upon affinity-based chemical proteomics approaches which have demonstrated success in identifying targets for a variety of novel inhibitors including hedgehog pathway, bromodomain and N-acetyltransferase inhibitors derived from phenotypic screens [18–20]. However, despite such examples of clear target deconvolution success stories, affinity-based chemical proteomics are often hampered by nonspecific binding of proteins [61–63]. The recent development of publicly available databases characterizing nonspecific protein binding contaminants associated with affinity-based proteomics methods attempts to address the background noise associated with affinity-based proteomics [62]. Competition assays where parent compound is preincubated with cell lysates prior to adding the conjugated affinity capture compound have been developed to determine nonspecific binding to affinity capture reagents and can be combined with databases describing common background contaminant profiles to increase confidence in identifying specific drug-target interactions [61,62]. Such approaches, however, do not completely resolve the issue of nonspecific binding, and affinity-based chemical proteomics is limited to providing lists of potential target binders rather than conclusive evidence of which target is responsible for the phenotypic response, thus further target confirmation studies are required. Several new and complementary target deconvolution strategies are rapidly emerging (Box 1), although no target identification method provides conclusive evidence of which target is responsible for the complete pharmacological profile of a compound. The application of distinct target deconvolution methods (Box 1) combined with other MOA profiling tools may provide strong corroborative evidence to prioritize target hypothesis, which may be responsible for phenotypic response. However, validation of target hypothesis will only be confirmed through further biochemical and cell pharmacology studies. Established and emerging target deconvolution strategies have been reviewed in depth elsewhere [64,65] and so will not be covered in further detail here; however, we do highlight the latest trends in target deconvolution strategies in Box 1.

Mechanism-of-action profiling

A critical success factor in any drug discovery project is the understanding of candidate drug MOA within complex and physiologically relevant biological settings. Several new technologies enable rapid MOA profiling in complex cell models at genetic, proteomic and phenotypic levels at scale. Such higher throughput MOA profiling can facilitate the selection of appropri-

ate phenotypic hits to take forward the further pre-clinical development, identify new assay endpoints and biomarkers to support early hit-to-lead chemical optimization, provide corroborative evidence for target deconvolution studies and support further pre-clinical development and translation toward clinical studies with or without conclusive target identification. Recent advances in MOA profiling technologies include: high-throughput gene transcription profiling, pathway profiling at the post-translational level and high-throughput phenotypic imaging and image informatics [15–17,66]. The latest developments and application of these approaches in PDD are described in further detail in the following sections.

High-throughput gene transcription profiling

Gene transcription-based profiling approaches using whole genome expression arrays provide a comprehensive overview of gene activity in biological samples. Common applications of gene expression arrays include genome-wide differential expression studies, disease classification and drug MOA analysis. The concept of using gene transcription profiling to elucidate drug MOA and deconvolve therapeutic targets was first applied by Hughes *et al.* who created a compendium of 300 yeast deletion strains and associated transcription profiles [27]. By correlating similarity of transcription profiles from drug-treated cells with those derived from each individual yeast deletion they identified the C-8 sterol isomerase, ERG2 as the target for the anesthetic Dyclonine [27]. To progress a more systematic comparative bioinformatics analysis of gene expression profiles, the Connectivity Map concept and public repository of transcription profiles was developed [14]. Connectivity Map combines a catalog of gene expression profiles from large panels of compound perturbed samples with computational and statistical methods to support similarity profiling of gene expression patterns to infer compound MOA [14,67]. As a proof-of-concept study, connectivity map gene expression profiling was applied to identify the MOA of the compound, Gedunin, identified as a hit from a screen for androgen receptor inhibitors. Gene transcription profiles of LNCaP prostate cancer cells treated for 6 h with Gedunin were used to query the Connectivity Map database, which identified high similarity with multiple HSP90 inhibitors; subsequent studies further support Gedunin as an inhibitor of HSP90 function [14]. The Connectivity Map approach has proven particularly useful for discovering the MOA of natural products from traditional remedies. A recent example used Connectivity Map to identify the MOA of Berberine, an isoquinoline alkaloid used in traditional Chinese

herbal medicine and which has demonstrated anticancer properties in phenotypic assays [68]. Transcription profiles of HepG2 cells treated with Berberine for 4 h demonstrated similarity with gene expression profiles of the protein synthesis inhibitor cycloheximide as well as several mTOR and HSP90 inhibitors. Subsequent cellular pharmacology studies demonstrated that Berberine inhibits protein synthesis, Akt activity but not mTOR activity and induces AMPK-mediated endoplasmic reticulum stress and autophagy [68]. Therefore, in this case, the initial application of Connectivity Map and gene transcription similarity profiling identified mechanistically similar compounds with known target activities to guide subsequent studies to further elucidate the MOA of Berberine.

Recent technical advances in gene expression profiling include the development of higher throughput and low-cost gene-expression methods such as the L1000™ platform. L1000™ Expression Profiling is based upon the rapid quantification of a reduced number of landmark transcripts in 384-well plate format and a computational model to infer expression across the genome [15]. The L1000™ technology underpins the Library of Integrated Cellular Signatures (LINCS) NIH program, which funds the generation of perturbed gene expression profiles across multiple cell and perturbation types supporting drug MOA profiling at scale [15,69,70]. While gene transcription profiling has proven effective in elucidation of compound MOA, success is dependent upon the use of appropriate biological assays where the relevant target pathway for any given compound is activated. A further dependency is the cross referencing to a comprehensive and well-annotated reference set of compound signatures also generated under appropriate biological context. An alternative approach to inferring MOA from gene expression signatures is the comparison of drug sensitivity/phenotypic response across large panel of cells with their basal gene expression profiles. A recent study used correlation-based analyses to associate the sensitivity of 481 compounds tested across 860 human cancer cell lines with the basal gene expression profile of each cell line [69]. The study included 115 small molecules of unknown mechanism with the aim of identifying novel targets for these compounds; correlation analysis was focused on single-transcript correlation outliers to prioritize potential target hypothesis [69]. Cancer cell sensitivity to the compounds BRD5468 and ML239 correlated with high expression of the monoglyceride lipase MGLL and the fatty acid desaturase FADS2, respectively [69]. Treatment with the MGLL inhibitor, JZL184 or shRNA knockdown of MGLL attenuated the cytotoxicity of BRD5468 and FADS2 knockdown and cotreatment with the selective FADS2 inhibitor

SC-26196 reduced ML239 cytotoxicity [69]. These studies demonstrate that correlation of drug sensitivity profiles with basal gene expression patterns across large cell panels can reveal specific target hypothesis. An advantage of correlating transcription profiles across large panels of cells is the ability to distinguish between distinct transcript correlations with drug sensitivity from coregulated transcripts thereby prioritizing the most likely targets. However, limitations of this approach include the prerequisite for compounds that display distinct sensitivity across cell panels, which also display differential gene expression patterns and confounding correlation with mechanisms of metabolism or indirect regulators of compound sensitivity. Indeed the analysis by Rees *et al.* demonstrated that for 57% of the compounds tested, no significant correlation with any target could be detected [69]. Despite the recent advances in transcription-based profiling technologies, the costs associated with such analysis limit high-throughput application to larger compound sets and dose–response and temporal studies. Transcription-based profiling may also only reveal the downstream effects of compound exposure rather than the direct therapeutic targets.

Pathway profiling across panels of primary cell-based assays

Profiling compound response at the post-translational pathway level across panels of primary cells and pathway targets has also demonstrated success in determining drug MOA, confirming selectivity, identifying toxicity liabilities and guiding SAR [71]. For example the BioMAP® – Human Primary Cell Phenotypic Profiling Services provided by DiscoverRX consists of panels of primary human cell-based assay systems, a database of reference compound profiles, and computational data mining and analysis tools to support drug MOA analysis [72]. The comparison of BioMAP profiles from testing of two p38MAPK inhibitors, PD169316 and SB203580 revealed activity features unique to SB203580 including inhibition of VCAM-1, E-selectin, IL-8 and P-selectin expression [71]. To further explore the structural determinants of the unique activities of SB203580, the BioMAP profiles of several well-studied p38MAPK inhibitors and SB203580 analogs were generated for comparison. These studies reveal that many of the unique activities of SB203580 represent secondary off-target activities independent of catalytic activity [71]. The BioMAP approach is applicable to large numbers of compounds tested across dose–response and time–series studies supporting precise SAR studies upon pathway responses. While the assays and core pathways tested represent highly sensitive readouts for multiple biological mechanisms,

the biological space covered will not be appropriate for elucidating the MOA of all molecules.

Reverse phase protein microarray

Reverse phase protein microarray (RPPA) represents a highly sensitive and quantitative high-throughput antibody-based proteomics methodology for measuring abundance of multiple proteins and phospho-proteins across large sample sets [73]. Key applications of RPPA include, dynamic pathway profiling at the post-translational network level following chemical or genetic perturbation, screening modulators of key pathway markers and protein biomarker discovery in clinical and preclinical studies [73–75]. Recent advances in RPPA technology include more sophisticated sample handling, quality control, better quality affinity reagents and optical detection, including planar waveguide detection systems providing femtomole to zeptomole sensitivity in protein analyte detection in formats suitable for medium-throughput applications [76]. The development of ultrasensitive RPPA facilitates large-scale multiplex analysis of multiple post-translational markers across small samples from *in vitro*, preclinical or clinical biopsies. Thus, RPPA technology is particularly suited to proteomic analysis of miniaturized assay formats of a few thousand cells from an individual well of a microtiter plate and microfluidic devices. Similar to the BioMAP and gene-expression approaches, multiple pathways can be monitored across large sets of assay panels and RPPA profiles compared with reference compound can help predict MOA and triage common/nonnovel pathway inhibitors or highly promiscuous pathway inhibitors with toxic liabilities. Retrospective analysis of esophageal adenocarcinoma patients who were also under treatment with metformin (Figure 1, structure 1) for diabetes demonstrated a better response to chemoradiation therapy compared with patients who were not receiving metformin [77]. However, the MOA of metformin in esophageal cancer was unknown. RPPA analysis applied to esophageal cancer cells treated with metformin revealed inhibition of PI3K/mTOR signaling pathway, which correlated with reduced cell growth and increased apoptosis [78]. In a similar approach, a Danish study comparing recurrence rates for breast cancer between Simvastatin users and nonusers demonstrated a significant reduction in recurrence rates in the statin users [79]. RPPA analysis of triple-negative breast cancer cell lines following Simvastatin treatment demonstrated decreased phosphorylation of FOXO3a. Subsequent knockdown of FOXO3a attenuated the effect of Simvastatin on mammosphere formation and migration [80]. Corilagin has recently been identified as a major active component in a well-known herbal medicine (*Phyllanthus niruri* L.) with antitumor activity although the antitumor mechanism

has not been clearly defined. RPPA analysis of a panel of ovarian cancer cell lines treated with Corilagin demonstrated inhibited activation of canonical Smad and non-canonical ERK/AKT pathways, which correlated with inhibition of TGF- β secretion and TGF- β pathway activation [81]. Similar to correlation of drug sensitivity across cell panels with basal gene expression profiles, drug sensitivity across cell panels have also been correlated with basal protein levels and pathway activation states by RPPA to identify both MOA and mechanism-of-resistance [82,83]. Correlation of sensitivity of a panel of small-cell lung cancer lines treated with the PARP inhibitor BMN 673 with RPPA analysis indicated the compound sensitivity is associated with elevated baseline expression levels of several DNA repair proteins [83]. This study identified a novel 'DNA repair score' consisting of a group of 17 DNA repair proteins, which predict sensitivity to BMN 673 [83]. Small-cell lung cancer insensitivity or resistance to BMN 673 correlated with baseline activation of the PI3K/mTOR pathway identifying a potential drug combination hypothesis [83]. While the majority of exemplar studies describing RPPA applications in drug MOA analysis have been applied to late-stage drug candidates or approved drugs, many of which have come from target-directed drug discovery, the success of this approach indicates that it will also be a useful method for uncovering the MOA of hits and lead compounds derived from phenotypic screens. A significant advantage of antibody-based proteomic profiling approaches is that they can help identify translatable pharmacodynamic or predictive biomarker reagents to guide appropriate preclinical proof-of-concept studies and clinical development strategies of drug candidates with or without conclusive target deconvolution.

High-content image-based multiparametric phenotypic profiling

Advances in automated microscopic image acquisition and image analysis tools enable the extraction of functional phenotypic endpoints from complex assay formats including 3D and co-culture models. Integration of high-throughput imaging assays with new image informatics resources enable high-throughput phenotypic profiling and classification of MOA across multiple assays, dose–response and time–series studies. We outline below the development in high-content imaging and image informatics methods and the new opportunities they present to phenotypic screening.

Evolution of high-content imaging & image informatics methods applicable to phenotypic screening

The rapid development of automated microscope platforms has enabled the ability to generate tens of thou-

sands of images a day on a single platform supporting medium- and high-throughput image-based phenotypic screening. With image analysis software capable of extracting several hundred measurements per cell from these images, researchers can detect and quantify subtle phenotypic changes that would otherwise be missed with the naked eye or from a single endpoint assay. These developments have stimulated a new field of biological profiling in cell-based assay systems called high-content analysis [9,84]. However, due to the high-dimensional nature of the high-content datasets, tried-and-tested methods to determine hits and guide SAR developed in TDD are no longer applicable. This means new methods for hit selection and triaging are required, and with the parallel developments in machine learning and other quantitative fields there are many options open to researchers.

Image-based phenotypic measurements can be recorded on two levels: an average of whole-well/-cell population measurements or individual cell measurements. Whole-well measurements are less computationally intensive and easier to obtain and can prove useful when individual cell segmentation is not feasible. Measurements taken from individual cells can be much more detailed, such as individual cell areas or number of organelles per cell. However, individual cell measurements generate large datasets that can become unwieldy and difficult to analyze without significant computing power and data handling pipelines. Therefore, many image-based phenotypic assays use well or population averages of data obtained from individual cell measurements, describing the mean or median cell within each image. While this reduces the amount of data, and allows for more simple analyses, calculating a population average removes any information about heterogeneity or possible phenotypic subpopulations. In instances of two equal sized subpopulations, a well average phenotypic measurement may be a representative of few cells within that image and thus does not accurately record phenotypic response across the cell population. A method to quantify cellular heterogeneity within cell populations has recently been suggested based on three simple statistic procedures: percentage of outliers; the Kolmogorov–Smirnov (KS) test of normality; and quadratic entropy. This method can then be used to classify a cellular population according to the type of heterogeneity observed [85].

The development of such cellular subpopulation analysis methods is important as the origins behind heterogeneity within clonal populations are not well-understood and the diverse response to therapeutics can be a driver underlying clonal selection, a well-known contributor to the evasion of anticancer therapeutics observed in many tumors. New methods calculating

heterogeneity and the impact of pharmacological intervention upon heterogeneous cell populations are thus especially relevant to anticipating therapeutic response and monitoring evolution of the disease in response to treatment within complex tumor microenvironments. Several studies have also reported that the expression of specific transcription factors associated with stem cell pluripotency are expressed in a heterogeneous fashion in embryonic stem cell cultures. For example, approximately 80% of embryonic stem cells express Nanog, while 10–20% do not [86]. Stem cell heterogeneity and conversion between distinct pluripotent or differentiated stem cell fates also impact upon therapeutic areas dependent upon endogenous stem cell differentiation and reprogramming such as tissue regeneration and repair. The evolution of image-based methods monitoring cell heterogeneity and classification of subpopulation responses at the single-cell level support the development of more complex and clinically relevant heterogeneous and multicellular models for automated cell-based screening. However, the challenge remains in how to distill such complex multiparametric data to enable key decision-making. Advances in the fields of multivariate statistics and machine learning offer potential solutions.

Development of multivariate high-content methods to predict compound MOA

In 2004, Perlman *et al.* published a landmark paper describing the use of compound ‘fingerprints’ derived from phenotypic measurements. It was shown that compounds with known similar MOA exhibited similar phenotypic fingerprints [17] and this could be used to predict the MOA of unknown compounds by their similarity to that of known compounds. In order to create the compound fingerprints a modified KS test was developed to compare the distribution of every measurement against the same measurement for the negative control, producing a list of numbers for each compound [87]. These vectors were aligned to other compound vectors in order to maximize correlation to account for differences in potency across ranges of concentrations. The pairwise Euclidean distance was calculated to create a similarity matrix between all the tested compounds; following hierarchical clustering, compounds with similar MOA were found closely aligned to one another. This was the first published demonstration that image-based phenotypic information proved descriptive enough to discern compounds from one another [17]. Further development on multiparametric phenotypic assays combined with different compound profiling methods utilizing multivariate statistics, machine learning and artificial neural networks have steadily evolved [88–92]. In a recent study, 2725

compounds were profiled in a multiparametric high-content assay measuring phenotypic effects upon the nucleus, cytoplasm, endoplasmic reticulum, golgi and cytoskeleton of the U2OS osteosarcoma cell line [54]. The high-content phenotypic fingerprints subsequently generated were used to cluster mechanistically similar compounds using the Markov Clustering Algorithm and then each compound cluster was analyzed for enrichment of individual targets and gene sets to facilitate MOA analysis [54]. Individual target annotations for compounds were obtained from public and commercial drug target databases such as, ChEMBL, Drugbank, GVK (GOSTAR), Integrity and Metabase. Gene set enrichments were obtained from the following databases: Biosystems, Metabase, Integrity, Metabase pathway-derived gene sets (Metabase noodles) and Gene Go Ontologies [54]. Two compounds, 6-[6-(diethylaminopyridin-3-yl)-*N*-[4-(4-morpholinyl)phenyl]-9*H*-purin-2-amine and Silmitasertib clustered with each other and a collection of other compounds inducing similar phenotypic response. In contrast to the majority of compounds in this cluster, which were associated with gene sets enriched in PI3K/Akt/mTOR, the previously described Jak3 inhibitor, 6-[6-(diethylaminopyridin-3-yl)-*N*-[4-(4-morpholinyl)phenyl]-9*H*-purin-2-amine and the Casein kinase II inhibitor, Silmitasertib had not previously been associated with direct inhibition of PI3K/AKT/mTOR pathway targets. Subsequent biochemical analysis revealed 6-[6-(diethylaminopyridin-3-yl)-*N*-[4-(4-morpholinyl)phenyl]-9*H*-purin-2-amine inhibited 3-phosphoinositide-dependent protein kinase 1 (PDK1), a component of PI3K/AKT/mTOR signaling and Silmitasertib inhibited mTOR and PI3K- α with IC₅₀ of 390 nM and 461 nM, respectively [54]. These studies demonstrate that novel compound–target associations can be identified from image-based multiparametric high-content profiling. In contrast to transcription or post-translational pathway profiling methods (BioMAP and RPPA) previously discussed, multiparametric high-content profiling assays can run in high-throughput across arrayed whole genome screens, large chemical libraries and compound profiling studies incorporating dose response and time series if necessary.

Integrating phenotypes & SAR to predict MOA

The study of SAR by the generation and screening of compounds with similar chemical structures, is one of the fundamental methods used by medicinal chemists to determine which structural motifs are required for inducing a biological effect on a particular protein, cell or organism. In principle, compounds with analogous chemical structures often bind to the same or similar

protein targets, a principle that is used to develop derivatives with improved drug metabolism and pharmacokinetic (DMPK) properties, and as Perlman *et al.* demonstrated, compounds with similar MOA produce similar phenotypes. Young *et al.* then filled the gap in this reasoning by investigating if compounds with comparable chemical structures produce similar phenotypes [87]. They screened HeLa cells with a small molecule compound library and performed factor analysis on 36 features to produce a fingerprint for each compound, with which a pair-wise similarity matrix was created by the cosine distance between phenotypic fingerprints. In order to determine the similarity between chemical structures, they defined the molecular structure through radial atom neighbors and a structure similarity matrix was constructed through Tanimoto distances between the compounds. The two similarity matrices, one for phenotype and one for chemical structure, were clustered by phenotypic similarity, which revealed distinct phenotypic clusters that matched up to distinct groups of structurally similar compounds [87]. Performing SAR studies in cell-based phenotypic assays is significantly challenged by the fact that effects of compound modulation upon phenotypic activity are multifactorial, influenced not only by target engagement, but also cellular permeability (cLogP/D-mediated), subcellular distribution, cell transport mechanisms, membrane interactions and off-target activities. These issues increase the likelihood of obtaining sharp activity cliffs, which hinder directional SAR studies contributing to more complex and lengthy medicinal chemistry programs. In the study by Young *et al.*, it was found that small changes in chemical structure were associated with large phenotypic differences indicating that sharp chemical activity cliffs are retained in information-rich high-content screening data [87]. Further biological investigation into the distinct multiparametric profiles obtained between chemically similar analogs may reveal the underlying causes of activity cliffs. For example, multiparametric high-content analysis can help diagnose if loss of activity is a consequence of reduced potency on a specific target mechanism, impaired distribution and influence within specific subcellular compartments or completely distinct MOA indicative of new off-target activities. Thus, high-content analysis supports a more in-depth cellular pharmacology approach to guiding subsequent chemical library design from initial phenotypic hits.

High-content imaging quality control/assay standards

To gain meaningful results from any screen, including phenotypic-guided SAR, assay quality control is critical. For many, the lack of reliable metrics to determine assay robustness, such as the z-factor in high-throughput

biochemical screens [93] is a deterrent for widespread adoption of multiparametric high-content methods in industrial drug discovery. Despite criticism of the inappropriate use of the z-factor in high-content studies by many groups, there is still not a universally accepted replacement. Any metric suggested has to address three primary concerns: ability to work with multivariate data; assay independence; and ease of implementation and interpretation. Many attempts to develop such a method have used the z-factor as their basis [94,95], although none have addressed all the issues or gained widespread adoption. The same principles apply for identifying phenotypic endpoints to guide SAR. Feature extraction and selection methods can reduce the data to a single value analogous to an IC_{50} to guide chemical design toward specific areas of phenotypic space. Successful implement of such phenotypic-guided SAR is, however, critically dependent upon reproducible assay formats, appropriate feature extraction and selection methods and deep biological insight to ensure phenotypic features guide chemical design toward desired outcomes. The use of open datasets, such as the Broad Bioimage Benchmark Collection [96,97] enables researchers to compare image analysis and informatics methods on a common collection of annotated images, allowing iterative improvement of methods through collaboration and the replication of results. Further collaborative initiatives to develop common standards for HCS and image analysis methods

will promote further adoption and stimulate cross-collaboration between both academic and industrial groups to advance the field of high-content image-based phenotypic screening and profiling.

Conclusion

To date, the majority of phenotypic screening assays that have been implemented in chemical or si/shRNA library screening campaigns and examples of compounds and drugs developed through PDD strategies have used simple biological models and assay readouts. Historical success of PDD combined with acceptance of a significant contribution to recent drug approvals has stimulated renewed interest in PDD strategies. In this article, we describe limitations of traditional PDD approaches and highlight solutions and new opportunities for PDD presented by recent advances in assay development and image-based screening technology. With new advances in precise gene editing technologies such as CRISPR-cas9, primary patient-derived cell culture, iPSc differentiation combined with multiparametric high-content phenotypic profiling, all advance the applications of phenotypic screening under more relevant and well-defined biological contexts. We propose that further development and adoption of new phenotypic assay technologies are well-placed to advance a new era of next-generation phenotypic screening contributing to both PDD and TDD success rates.

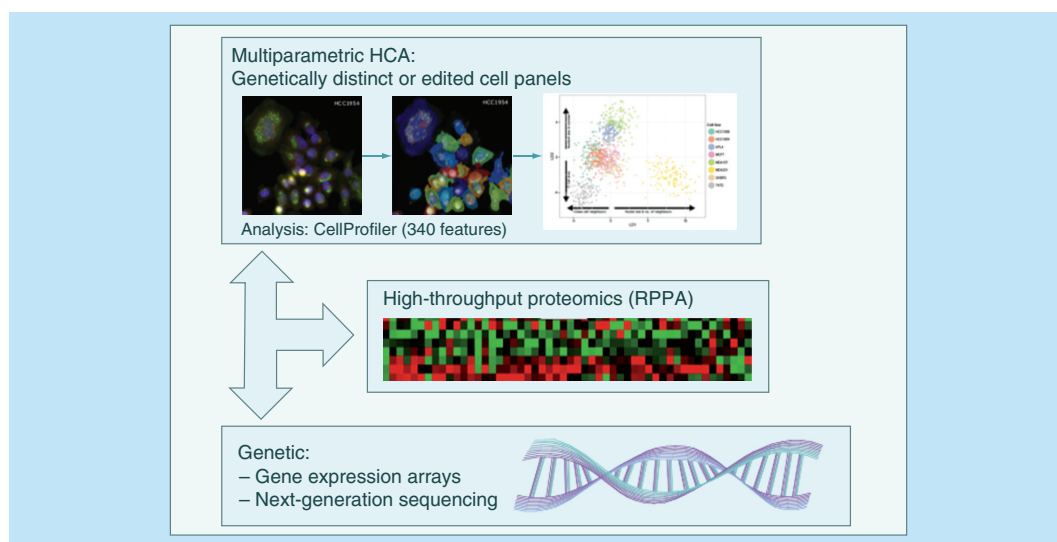


Figure 3. Integration of high-content multiparametric phenotypic profiles with proteomic and genetic datasets. Representative displays of multiparametric cell morphometry analysis using optimized cell staining and CellProfiler image-analysis protocols followed by correlation of distinct drug-induced phenotypes between cells with molecular data at post-translational and genetic levels. Application of these methods support understanding of drug mechanism of action; identification of resistance mechanisms/pathways to guide biomarker discovery; novel drug combination hypotheses and high-throughput pharmacogenomics incorporating more complex phenotypes across disease areas and across advanced multicellular or 3D models. RPPA: Reverse phase protein microarray.

Future perspective

Despite over 10 years of research carried out with high-content phenotypic screening, the majority of studies have focused their efforts on a small selection of established cell lines, picked primarily due to amenable cell culture propagation and imaging properties rather than relevance to human disease. The reasons behind this are understandable, as cost and speed represent important criteria in medium- to high-throughput screening projects. Thus, selection of cell lines, which can be rapidly bulked up and accurately and reliably segmented into 2D cell culture assays using readily available image analysis methods, are attractive. An important advantage of image-based high-content screening over other screening platforms is the ability to extract functional endpoints from more complex *in vitro* assays, which extend beyond simple 2D cultures and may include 3D multicellular tissue models and small model organism screens, which exploit more complex biology. The development and adoption of more complex *in vitro* assays may benefit PDD in several ways:

- Application of assays which more accurately represent disease pathophysiology thus contributing to improved translation and clinical success rates;
- Identify novel target space including unbiased identification of novel target classes, which are not currently being pursued by drug discovery groups;
- Identify targets with more relevant functional validation, increasing confidence in target hypothesis to justify subsequent TDD investments;

- Recapitulate intact autocrine, paracrine and juxtacrine pathway signaling networks supporting discovery and development of novel multitargeted therapies and combination approaches.

The primary goal of PDD is to identify small molecules that beneficially modify a disease-associated phenotype, selecting a single cell line to model the disease can, however, prove risky. As demonstrated in cystic fibrosis disease models, there is little overlap between compounds that show efficacy in correcting the CFTR trafficking defect when the mutant CFTR protein is expressed across multiple cell lines [98]. This should lead us to question how well we place our trust in conclusions drawn from an experiment modeled in a single cell line. Application of high-content screening across genetically distinct primary cells or precise CRISPR-cas9 gene-edited cell panels can help elucidation of drug MOA by linking phenotype to genotype and also stimulate biomarker and drug combinations studies (Figure 3) [69,99].

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background

Historical examples of drugs discovered by phenotypic screening

- Approved drugs with unknown target mechanism.
- Recent examples of modern phenotypic screening outcomes, including development of highly potent and selective targeted agents.
- Pitfalls of poorly designed phenotypic screens/black-box assays.
- Challenges in target deconvolution.
- New approaches in target deconvolution.
- New approaches in mechanism-of-action determination (genomic profiling, proteomics and high-content analysis).

Evolution of high-content imaging & image informatics methods applicable to phenotypic screening

- Early multiparametric high-content methods for compound classification.
- More advanced image analysis and image informatics, including integration of multiparametric phenotypic fingerprints with chemical similarity.
- Current limitations in high-content analysis and high-throughput image informatics.

Future perspective

- New disease models, incorporating 3D assays, induced pluripotent stem cell and gene editing technologies.
- New opportunities for application of phenotypic screening across genetically distinct/gene-edited cell panels linking phenotype to genotype to support high-throughput genomics and personalized healthcare strategies in new disease areas.

References

Papers of special note have been highlighted as:

• of interest; •• of considerable interest

- 1 Sams-Dodd F. Target-based drug discovery: is something wrong? *Drug Discov. Today* 10(2), 139–147 (2005).
- 2 Lee JA, Berg EL. Neoclassic drug discovery: the case for lead generation using phenotypic and functional approaches. *J. Biomol. Screen* 18(10), 1143–1155 (2013).
- 3 Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 11(3), 191–200 (2012).
- 4 Carragher NO, Brunton VG, Frame MC. Combining imaging and pathway profiling: an alternative approach to cancer drug discovery. *Drug Discov. Today* 17(5–6), 203–214 (2012).
- 5 Carragher NO. Advancing high content analysis towards improving clinical efficacy. *European Pharmaceutical Review* (2011).
www.europeanpharmaceuticalreview.com/5648
- 6 Eder J, Sedrani R, Wiesmann C. The discovery of first-in-class drugs: origins and evolution. *Nat. Rev. Drug Discov.* 13(8), 577–587 (2014).
- 7 Swinney DC, Anthony J. How were new medicines discovered? *Nat. Rev. Drug Discov.* 10(7), 507–519 (2011).
- 8 Moffat JG, Rudolph J, Bailey D. Phenotypic screening in cancer drug discovery – past, present and future. *Nat. Rev. Drug Discov.* 13(8), 588–602 (2014).
- 9 Bickle M. The beautiful cell: high-content screening in drug discovery. *Anal. Bioanal. Chem.* 398(1), 219–226 (2010).
- 10 Chang TY, Pardo-Martin C, Allalou A, Wahlby C, Yanik MF. Fully automated cellular-resolution vertebrate screening platform with parallel animal processing. *Lab Chip* 12(4), 711–716 (2012).
- 11 Isherwood B, Timpson P, Mcghee EJ *et al.* Live cell *in vitro* and *in vivo* imaging applications: accelerating drug discovery. *Pharmaceutics* 3(2), 141–170 (2011).
- 12 Rix U, Superti-Furga G. Target profiling of small molecules by chemical proteomics. *Nat. Chem. Biol.* 5(9), 616–624 (2009).
- 13 Martinez Molina D, Nordlund P. The cellular thermal shift assay: a novel biophysical assay for *in situ* drug target engagement and mechanistic biomarker studies. *Annu. Rev. Pharmacol. Toxicol.* 56 141–161 (2016).
- 14 Lamb J, Crawford ED, Peck D *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313(5795), 1929–1935 (2006).
- **Development of the Connectivity Map database to correlate transcription profiles from compound perturbed samples with transcription profiles from other compounds and disease to support mechanism-of-action analysis and disease positioning.**
- 15 Liu C, Su J, Yang F, Wei K, Ma J, Zhou X. Compound signature detection on LINCS L1000 big data. *Mol. Biosyst.* 11(3), 714–722 (2015).
- 16 Pawlak M, Schick E, Bopp MA, Schneider MJ, Oroszlan P, Ehrat M. Zeptosens' protein microarrays: a novel high performance microarray platform for low abundance protein analysis. *Proteomics* 2(4), 383–393 (2002).
- 17 Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. *Science* 306(5699), 1194–1198 (2004).
- **First demonstration that multiparametric high-content profiling can discriminate and predict compound mechanism-of-action.**
- 18 Lee J, Wu X, Pasca Di Magliano M *et al.* A small-molecule antagonist of the hedgehog signaling pathway. *Chembiochem* 8(16), 1916–1919 (2007).
- 19 Filippakopoulos P, Qi J, Picaud S *et al.* Selective inhibition of BET bromodomains. *Nature* 468(7327), 1067–1073 (2010).
- 20 Larrieu D, Britton S, Demir M, Rodriguez R, Jackson SP. Chemical inhibition of NAT10 corrects defects of laminopathic cells. *Science* 344(6183), 527–532 (2014).
- 21 Ong SE, Schenone M, Margolin AA *et al.* Identifying the proteins to which small-molecule probes and drugs bind in cells. *Proc. Natl Acad. Sci. USA* 106(12), 4617–4622 (2009).
- 22 Jafari R, Almqvist H, Axelsson H *et al.* The cellular thermal shift assay for evaluating drug target interactions in cells. *Nat. Protoc.* 9(9), 2100–2122 (2014).
- 23 Geuijen CA, Bijl N, Smit RC *et al.* A proteomic approach to tumour target identification using phage display, affinity purification and mass spectrometry. *Eur. J. Cancer* 41(1), 178–187 (2005).
- 24 Moser S, Johnsson K. Yeast three-hybrid screening for identifying anti-tuberculosis drug targets. *Chembiochem* 14(17), 2239–2242 (2013).
- 25 Chidley C, Haruki H, Pedersen MG, Muller E, Johnsson K. A yeast-based screen reveals that sulfasalazine inhibits tetrahydrobiopterin biosynthesis. *Nat. Chem. Biol.* 7(6), 375–383 (2011).
- 26 Sandercock AM, Rust S, Guillard S *et al.* Identification of anti-tumour biologics using primary tumour models, 3-D phenotypic screening and image-based multi-parametric profiling. *Mol. Cancer* 14, 147 (2015).
- 27 Hughes TR, Marton MJ, Jones AR *et al.* Functional discovery via a compendium of expression profiles. *Cell* 102(1), 109–126 (2000).
- 28 Giaever G, Shoemaker DD, Jones TW *et al.* Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* 21(3), 278–283 (1999).
- 29 Wacker SA, Houghtaling BR, Elemento O, Kapoor TM. Using transcriptome sequencing to identify mechanisms of drug action and resistance. *Nat. Chem. Biol.* 8(3), 235–237 (2012).
- 30 Burgess DJ, Doles J, Zender L *et al.* Topoisomerase levels determine chemotherapy response *in vitro* and *in vivo*. *Proc. Natl Acad. Sci. USA* 105(26), 9053–9058 (2008).
- 31 Heynen-Genel S, Pache L, Chanda SK, Rosen J. Functional genomic and high-content screening for target discovery and deconvolution. *Expert Opin. Drug Discov.* 7(10), 955–968 (2012).
- 32 Dominguez E, Galmozzi A, Chang JW *et al.* Integrated phenotypic and activity-based profiling links Ces3 to obesity and diabetes. *Nat. Chem. Biol.* 10(2), 113–121 (2014).

- 33 Swinney DC. The contribution of mechanistic understanding to phenotypic screening for first-in-class medicines. *J. Biomol. Screen.* 18(10), 1186–1192 (2013).
- 34 Nathan DM, Buse JB, Davidson MB *et al.* Medical management of hyperglycemia in Type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* 32(1), 193–203 (2009).
- 35 Ross S, Thamer C, Cescutti J, Meinicke T, Woerle HJ, Broedl UC. Efficacy and safety of empagliflozin twice daily versus once daily in patients with Type 2 diabetes inadequately controlled on metformin: a 16-week, randomized, placebo-controlled trial. *Diabetes Obes. Metab.* 17(7), 699–702 (2015).
- 36 Miller RA, Chu Q, Xie J, Foretz M, Viollet B, Birnbaum MJ. Biguanides suppress hepatic glucagon signalling by decreasing production of cyclic AMP. *Nature* 494(7436), 256–260 (2013).
- 37 Pogliano J, Pogliano N, Silverman JA. Daptomycin-mediated reorganization of membrane architecture causes mislocalization of essential cell division proteins. *J. Bacteriol.* 194(17), 4494–4504 (2012).
- 38 Kawashima T, Iwamoto I, Nakagawa N, Tomioka H, Yoshida S. Inhibitory effect of pemirolast, a novel antiallergic drug, on leukotriene C₄ and granule protein release from human eosinophils. *Int. Arch. Allergy. Immunol.* 103(4), 405–409 (1994).
- 39 Hakimian S, Cheng-Hakimian A, Anderson GD, Miller JW. Rufinamide: a new anti-epileptic medication. *Expert Opin. Pharmacother.* 8(12), 1931–1940 (2007).
- 40 Vezina C, Kudelski A, Sehgal SN. Rapamycin (AY-22, 989), a new antifungal antibiotic. I. Taxonomy of the producing streptomycete and isolation of the active principle. *J. Antibiot. (Tokyo)* 28(10), 721–726 (1975).
- 41 Martel RR, Klicius J, Galet S. Inhibition of the immune response by rapamycin, a new antifungal antibiotic. *Can. J. Physiol. Pharmacol.* 55(1), 48–51 (1977).
- 42 Houchens DP, Ovejera AA, Riblet SM, Slagel DE. Human brain tumor xenografts in nude mice as a chemotherapy model. *Eur. J. Cancer Clin. Oncol.* 19(6), 799–805 (1983).
- 43 Seto B. Rapamycin and mTOR: a serendipitous discovery and implications for breast cancer. *Clin. Transl. Med.* 1(1), 29 (2012).
- 44 Brown EJ, Albers MW, Shin TB *et al.* A mammalian protein targeted by G1-arresting rapamycin-receptor complex. *Nature* 369(6483), 756–758 (1994).
- 45 Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1(4), 337–341 (2004).
- 46 Pike KG, Malagu K, Hummersone MG *et al.* Optimization of potent and selective dual mTORC1 and mTORC2 inhibitors: the discovery of AZD8055 and AZD2014. *Bioorg. Med. Chem. Lett.* 23(5), 1212–1216 (2013).
- 47 Fraser C, Carragher NO, Unciti-Broceta A. eCF309: a potent, selective, cell-permeable mTOR inhibitor. *Med. Chem. Comm.* 7, 471–477 (2016).
- 48 Fraser C, Dawson JC, Dowling R *et al.* Rapid discovery and structure–activity relationships of pyrazolopyrimidines that potently suppress breast cancer cell growth via SRC kinase inhibition with exceptional selectivity over ABL kinase. *J. Med. Chem.* 59(10), 4697–4710 (2016).
- 49 Yoshida T, Kakegawa J, Yamaguchi T *et al.* Identification and characterization of a novel chemotype MEK inhibitor able to alter the phosphorylation state of MEK1/2. *Oncotarget* 3(12), 1533–1545 (2012).
- 50 Mayer TU, Kapoor TM, Haggarty SJ, King RW, Schreiber SL, Mitchison TJ. Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. *Science* 286(5441), 971–974 (1999).
- 51 Theoclitou ME, Aquila B, Block MH *et al.* Discovery of (+)-N-(3-aminopropyl)-N-[1-(5-benzyl-3-methyl-4-oxo-[1,2]thiazolo[5,4-d]pyrimidin-6-yl)-2-methylpropyl]-4-methylbenzamide (AZD4877), a kinesin spindle protein inhibitor and potential anticancer agent. *J. Med. Chem.* 54(19), 6734–6750 (2011).
- 52 Gercitano JF, Stephenson JJ, Lewis NL *et al.* A Phase I trial of the kinesin spindle protein (Eg5) inhibitor AZD4877 in patients with solid and lymphoid malignancies. *Invest. New Drugs* 31(2), 355–362 (2013).
- 53 Bordet T, Buisson B, Michaud M *et al.* Identification and characterization of cholest-4-en-3-one, oxime (TRO19622), a novel drug candidate for amyotrophic lateral sclerosis. *J. Pharmacol. Exp. Ther.* 322(2), 709–720 (2007).
- 54 Reisen F, Sauty De Chalon A, Pfeifer M, Zhang X, Gabriel D, Selzer P. Linking phenotypes and modes of action through high-content screen fingerprints. *Assay Drug Dev. Technol.* 13(7), 415–427 (2015).
- **Identification of target hypothesis by correlating multiparametric phenotypic profiling with gene enrichment datasets of phenotypically similar compounds.**
- 55 Isherwood BJ, Walls RE, Roberts ME *et al.* High-content analysis to leverage a robust phenotypic profiling approach to vascular modulation. *J. Biomol. Screen* 18(10), 1246–1259 (2013).
- 56 Vincent F, Loria P, Pregel M *et al.* Developing predictive assays: the phenotypic screening “rule of 3”. *Sci. Transl. Med.* 7(293), 293ps215 (2015).
- 57 Zheng W, Thorne N, Mckew JC. Phenotypic screens as a renewed approach for drug discovery. *Drug Discov. Today* 18(21–22), 1067–1073 (2013).
- 58 Pollard SM, Yoshikawa K, Clarke ID *et al.* Glioma stem cell lines expanded in adherent culture have tumor-specific phenotypes and are suitable for chemical and genetic screens. *Cell Stem Cell* 4(6), 568–580 (2009).
- 59 Nobis M, Mcghee EJ, Morton JP *et al.* Intravital FLIM-FRET imaging reveals dasatinib-induced spatial control of src in pancreatic cancer. *Cancer Res.* 73(15), 4674–4686 (2013).
- 60 Hirschhaeuser F, Menne H, Dittfeld C, West J, Mueller-Klieser W, Kunz-Schughart LA. Multicellular tumor spheroids: an underestimated tool is catching up again. *J. Biotechnol.* 148(1), 3–15 (2010).

- 61 Tang H, Duggan S, Richardson PL, Marin V, Warder SE, McLoughlin SM. Target identification of compounds from a cell viability phenotypic screen using a bead/lysate-based affinity capture platform. *J. Biomol. Screen.* 21(2), 201–211 (2016).
- 62 Mellacheruvu D, Wright Z, Couzens AL *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* 10(8), 730–736 (2013).
- 63 Trinkle-Mulcahy L, Boulton S, Lam YW *et al.* Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J. Cell Biol.* 183(2), 223–239 (2008).
- 64 Lee J, Bogoy M. Target deconvolution techniques in modern phenotypic profiling. *Curr. Opin. Chem. Biol.* 17(1), 118–126 (2013).
- 65 Cong F, Cheung AK, Huang SM. Chemical genetics-based target identification in drug discovery. *Annu. Rev. Pharmacol. Toxicol.* 52 57–78 (2012).
- 66 Kunkel EJ, Dea M, Ebens A *et al.* An integrative biology approach for analysis of drug action in models of human vascular inflammation. *FASEB J.* 18(11), 1279–1281 (2004).
- 67 Broad Institute. Cancer Program Resource Gateway. Connectivity Map. www.broadinstitute.org/software/cprg/?q=node/12
- 68 Lee KH, Lo HL, Tang WC, Hsiao HH, Yang PM. A gene expression signature-based approach reveals the mechanisms of action of the Chinese herbal medicine berberine. *Sci. Rep.* 4, 6394 (2014).
- 69 Rees MG, Seashore-Ludlow B, Cheah JH *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* 12(2), 109–116 (2016).
- 70 NIH LINCS Program. www.lincsproject.org/
- 71 Kunkel EJ, Plavec I, Nguyen D *et al.* Rapid structure-activity and selectivity analysis of kinase inhibitors by BioMAP analysis in complex human primary cell-based models. *Assay. Drug Dev. Technol.* 2(4), 431–441 (2004).
- 72 Berg EL, Yang J, Polokoff MA. Building predictive models for mechanism-of-action classification from phenotypic assay data sets. *J. Biomol. Screen.* 18(10), 1260–1269 (2013).
- 73 Charboneau L, Tory H, Chen T *et al.* Utility of reverse phase protein arrays: applications to signalling pathways and human body arrays. *BriefFunct. Genomic Proteomic* 1(3), 305–315 (2002).
- 74 Lee MJ, Ye AS, Gardino AK *et al.* Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* 149(4), 780–794 (2012).
- 75 Van Oostrum J, Calonder C, Rechsteiner D *et al.* Tracing pathway activities with kinase inhibitors and reverse phase protein arrays. *Proteomics Clin. Appl.* 3(4), 412–422 (2009).
- 76 Akbani R, Becker KF, Carragher N *et al.* Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: a workshop report: the RPPA (Reverse Phase Protein Array) society. *Mol. Cell. Proteomics* 13(7), 1625–1643 (2014).
- 77 Skinner HD, Mccurdy MR, Echeverria AE *et al.* Metformin use and improved response to therapy in esophageal adenocarcinoma. *Acta Oncol.* 52(5), 1002–1009 (2013).
- 78 Honjo S, Ajani JA, Scott AW *et al.* Metformin sensitizes chemotherapy by targeting cancer stem cells and the mTOR pathway in esophageal cancer. *Int. J. Oncol.* 45(2), 567–574 (2014).
- 79 Ahern TP, Pedersen L, Tarp M *et al.* Statin prescriptions and breast cancer recurrence risk: a Danish nationwide prospective cohort study. *J. Natl Cancer Inst.* 103(19), 1461–1468 (2011).
- 80 Wolfe AR, Debeb BG, Lacerda L *et al.* Simvastatin prevents triple-negative breast cancer metastasis in pre-clinical models through regulation of FOXO3a. *Breast Cancer Res. Treat.* 154(3), 495–508 (2015).
- 81 Jia L, Jin H, Zhou J *et al.* A potential anti-tumor herbal medicine, Corilagin, inhibits ovarian cancer cell growth through blocking the TGF-beta signaling pathways. *BMC Complement Altern. Med.* 13, 33 (2013).
- 82 Park ES, Rabinovsky R, Carey M *et al.* Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. *Mol. Cancer Ther.* 9(2), 257–267 (2010).
- 83 Cardnell RJ, Feng Y, Diao L *et al.* Proteomic markers of DNA repair and PI3K pathway activation predict response to the PARP inhibitor BMN 673 in small cell lung cancer. *Clin. Cancer Res.* 19(22), 6322–6328 (2013).
- 84 Taylor DL, Woo ES, Giuliano KA. Real-time molecular and cellular analysis: the new frontier of drug discovery. *Curr. Opin. Biotechnol.* 12(1), 75–81 (2001).
- 85 Gough AH, Chen N, Shun TY *et al.* Identifying and quantifying heterogeneity in high content analysis: application of heterogeneity indices to drug discovery. *PLoS ONE* 9(7), e102678 (2014).
- The development of novel analysis methods for quantifying heterogeneous responses across cell subpopulations in high-content phenotypic assays.
- 86 Graf T, Stadtfeld M. Heterogeneity of embryonic and adult stem cells. *Cell Stem Cell* 3(5), 480–483 (2008).
- 87 Young DW, Bender A, Hoyt J *et al.* Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* 4(1), 59–68 (2008).
- The intergation of chemical similarity with high-content phenotypic similarity methods to predict target mechanism.
- 88 Tanaka M, Bateman R, Rauh D *et al.* An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol.* 3(5), e128 (2005).
- 89 Caie PD, Walls RE, Ingleston-Orme A *et al.* High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol. Cancer Ther.* 9(6), 1913–1926 (2010).
- 90 Ljosa V, Caie PD, Ter Horst R *et al.* Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* 18(10), 1321–1329 (2013).
- 91 Kandaswamy C, Silva LM, Alexandre LA, Santos JM. High-content analysis of breast cancer using single-cell deep transfer learning. *J. Biomol. Screen.* 21(3), 290–297 (2016).

- 92 Smith K, Horvath P. Active learning strategies for phenotypic profiling of high-content screens. *J. Biomol. Screen.* 19(5), 685–695 (2014).
- 93 Zhang JH, Chung TD, Oldenburg KR. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* 4(2), 67–73 (1999).
- 94 Kozak K, Csucs G. Kernelized Z' factor in multiparametric screening technology. *RNA Biol.* 7(5), 615–620 (2010).
- 95 Kummel A, Gubler H, Gehin P, Beibel M, Gabriel D, Parker CN. Integration of multiple readouts into the Z' factor for assay quality assessment. *J. Biomol. Screen.* 15(1), 95–101 (2010).
- 96 Broad Institute Benchmark Collection. www.broadinstitute.org/bbbc/
- 97 Ljosa V, Sokolnicki KL, Carpenter AE. Annotated high-throughput microscopy image sets for validation. *Nat. Methods* 9(7), 637 (2012).
- 98 Pedemonte N, Tomati V, Sondo E, Galletta LJ. Influence of cell background on pharmacological rescue of mutant CFTR. *Am. J. Physiol. Cell Physiol.* 298(4), C866–C874 (2010).
- 99 Breinig M, Klein FA, Huber W, Boutros M. A chemical-genetic interaction map of small molecules using high-throughput imaging in cancer cells. *Mol. Syst. Biol.* 11(12), 846 (2015).

Development of the Theta Comparative Cell Scoring Method to Quantify Diverse Phenotypic Responses Between Distinct Cell Types

Scott J. Warchal, John C. Dawson, and Neil O. Carragher

Institute of Genetics and Molecular Medicine, Cancer Research UK Edinburgh Centre, University of Edinburgh, Edinburgh, United Kingdom.

ABSTRACT

In this article, we have developed novel data visualization tools and a Theta comparative cell scoring (TCCS) method, which supports high-throughput in vitro pharmacogenomic studies across diverse cellular phenotypes measured by multiparametric high-content analysis. The TCCS method provides a univariate descriptor of divergent compound-induced phenotypic responses between distinct cell types, which can be used for correlation with genetic, epigenetic, and proteomic datasets to support the identification of biomarkers and further elucidate drug mechanism-of-action. Application of these methods to compound profiling across high-content assays incorporating well-characterized cells representing known molecular subtypes of disease supports the development of personalized healthcare strategies without prior knowledge of a drug target. We present proof-of-principle data quantifying distinct phenotypic response between eight breast cancer cells representing four disease subclasses. Application of the TCCS method together with new advances in next-generation sequencing, induced pluripotent stem cell technology, gene editing, and high-content phenotypic screening are well placed to advance the identification of predictive biomarkers and personalized medicine approaches across a broader range of disease types and therapeutic classes.

INTRODUCTION

The treatment of complex disease in human populations is often confounded by the broad heterogeneity in the mechanisms responsible for the generation and evolution of disease-affected cells. Within an individual patient and between genetically distinct patients, such heterogeneity in disease mechanisms contributes to poor drug responses and relapses observed in the clinic.^{1,2} Sequencing of

the human genome and advances in characterizing patient disease at genetic and proteomic levels support the personalized medicine concept of treating each individual patient with the most appropriate therapy for their disease.^{3,4}

Key to the personalized medicine approach is the identification of biomarkers, which can be readily measured in patient samples to predict drug response. Many such biomarkers, for example, BRAF V600E (Melanoma/Colorectal Cancer); EGFR (Nonsmall cell lung carcinoma); and HER-2 (Breast Cancer), are associated with monitoring activation state and mutation status of known drug targets to predict response to therapy.⁵⁻⁷ Thus, the personalized medicine approach is well suited to target-directed drug discovery strategies where target pathways are clearly defined. However, such target-directed personalized medicine strategies are unsuitable for many complex diseases and drugs discovered by phenotypic drug discovery, where they are not defined by a single target or the mechanism-of-action and therapeutic targets remain to be fully elucidated.^{8,9} Thus, more unbiased approaches to the identification of biomarkers, including genetic or pathway signatures, which predict drug response are required to expand the personalized medicine concept across complex disease types and therapeutic classes.

Comparative analysis of well-characterized panels of human cell lines derived from distinct individuals has many applications in basic research, drug discovery, and personalized medicine. Genomic and transcriptional profiling of cancer cell line panels, such as the National Cancer Institute 60 human tumor cell line drug screen collection, provide a genetic context to comparison of cell function and drug sensitivity, supporting biomarker discovery and drug mechanism-of-action analysis.¹⁰

High-throughput *in vitro* pharmacogenomic studies across larger cancer cell line panels have been established and provide valuable resources, such as the Cancer Cell Line Encyclopedia (CCLE) from the Broad Institute www.broadinstitute.org/ccle/home and the Catalogue of Somatic Mutation in Cancer Cell Lines project from the Sanger Institute http://cancer.sanger.ac.uk/cell_lines, which facilitate

© Scott J. Warchal et al., 2016; Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

pharmacogenomic analysis. Such drug sensitivity profiling across genetically defined cell panels is now routinely implemented in academia and industry to identify biomarkers of response to support disease positioning and patient stratification strategies, or to further understand drug mechanism-of-action at genetic and proteomic levels.^{11,12}

To our knowledge, all examples of *in vitro* high-throughput pharmacogenomic studies carried out to date utilize either a concentration of a drug that gives half-maximal response (EC₅₀) or concentration of a drug that gives half-maximal inhibition of cell proliferation (GI₅₀) value obtained by standard cell viability assays as the primary phenotypic endpoint for correlating drug sensitivity with genomic or transcriptomic datasets. While the GI₅₀ and EC₅₀ measurements of cell viability provide the necessary univariate value for quantifying drug sensitivity across a panel of cell lines, this method has several limitations.

Accurate measurement of EC₅₀ or GI₅₀ values is dependent upon obtaining full sigmoidal dose-response curves for each drug or compound tested in the assay. Dose-response curves and thus the EC₅₀/GI₅₀ calculations are prone to fluctuation dependent upon assay conditions, including cell culture media, atmospheric conditions, cell line health and cell line batch variation, and the type of viability assay reagents used. Indeed, comparative analysis of large pharmacogenomic studies published by the Broad and Sanger institutes have resulted in reports of inconsistency between the genetic signatures of drug sensitivity assigned to drugs shared between both studies.^{13,14} Cell viability assays and EC₅₀/GI₅₀ values are also not suitable for the majority of disease models, which are not defined by a single viability endpoint, or for quantifying drug response in more complex and physiologically relevant cell assays such as three-dimensional (3D) coculture models.

High-content imaging enables the quantification of multiple phenotypic cellular endpoints with high spatial and temporal resolution supporting drug sensitivity testing across more complex *in vitro* assays including 3D and coculture models.¹⁵ Image-based phenotypic profiling combined with multiparametric analysis methods allows detailed characterization of drug mechanism-of-action and classification of phenotypic response, including identification of novel compound target associations based upon similarity of multiparametric phenotypic fingerprints with annotated reference compound sets.^{16–22}

The application of multiparametric biological profiling of compound libraries, by image-informatics and biospectra analysis methods, supports unbiased approaches to mechanism-of-action classification and identification of structure-activity relationships independent of target hypothesis.^{23–25} While multiparametric methods incorporating machine learning and artificial neural networks have steadily evolved to support phenotypic profiling across

several cell types,^{18,20,26} there are few studies that perform comparative multiparametric phenotypic analysis between distinct cell types in drug discovery. Thus, despite over 15 years of continued development in the high-content screening field, there are few reports of pharmacogenomic studies performed across the diversity of complex phenotypes that can be measured by multiparametric high-content analysis approaches. A number of challenges that must be overcome to apply high-content phenotypic profiling to pharmacogenomic or pharmacoproteomic strategies include the following: defining relevant phenotypic endpoints, which appropriately quantify drug sensitivity; quantifying diverse phenotypic response across a dose response; visualizing multiple diverse phenotypes elicited across dose response and distinct cell panels; and reducing multiparametric high-content analysis of cell phenotype to a robust univariate metric for correlating drug sensitivity with genomic or proteomic datasets.

The goals of this study were to develop a robust and scalable method for quantifying diverse multiparametric high-content phenotypes and distinct compound-induced phenotypic response across a panel of cell lines. We describe the optimization of a high-content cell-painting assay to enable analysis of a broad range of cell phenotypes across a panel of clinically relevant breast cancer subtypes. We present new methods for normalizing and displaying distinct and dose-dependent multiparametric high-content phenotypic response across multiple cell types. We introduce the development and application of the “Theta Comparative Cell Scoring” (TCCS) method for calculating distinct phenotypic response between cell types. We describe the broad utility of the TCCS method in providing a univariate metric for quantifying distinct phenotypic response between compounds tested in the same cell and for compounds tested across multiple cell types. We make available the source code to enable application of TCCS across large high-content datasets. We present proof-of-principle data from a small compound screen performed on a panel of eight breast cancer cells representing four well-characterized and clinically relevant subtypes. We demonstrate the ability of our TCCS method to cluster cell types, which have similar or distinct phenotypic response to individual compounds, to guide patient stratification hypothesis and facilitate pharmacogenomic or proteomic studies. We discuss the potential impact of this approach upon extending the application of *in vitro* pharmacogenomic and personalized medicine strategies across a wider range of disease areas and therapeutic classes.

MATERIALS AND METHODS

Cell Culture

Eight breast cancer cell lines were selected for their stratification of four well-characterized breast cancer clinical

A METHOD TO QUANTIFY PHENOTYPIC RESPONSES BETWEEN CELL TYPES

Table 1. Panel of Breast Cancer Cell Lines

Cell Line	Subclass	Mutation Status	
		PTEN ^a	PI3K ^b
MCF7	ER ^c	WT ^d	E545K
T47D	ER	WT	H1047R
MDA-MB-231	TN ^e	WT	WT
MDA-MB-157	TN	WT	WT
HCC1569	HER2 ^f	WT	WT
SKBR3	HER2	WT	WT
HCC1954	HER2	?	H1047R
KPL4	HER2	?	H1047R

^aPhosphatase and tensin homolog.

^bPhosphoinositide-3-kinase.

^cEstrogen receptor.

^dWild type.

^eTriple negative.

^fHuman epidermal growth factor receptor 2.

^gLack of consensus regarding the mutational status of those cell lines.

ER, estrogen receptor; HER2, human epidermal growth factor; PI3K, phosphoinositide 3-kinase; PTEN, phosphatase and tensin homolog; TN, triple negative; WT, wild type.

subtypes (Table 1). Authenticated cell lines were acquired from the American Type Culture Collection and carefully monitored for morphological changes to ensure authenticity. Cell lines were cultured in either Dulbecco's Modified Eagle's Medium (HCC1954, MCF7, KPL4, MDA-MB-231, MDA-MB-157, and SKBR3) or Roswell Park Memorial Institute-1640 (HCC1569 and T47D) supplemented with 10% fetal bovine serum and 2 mM L-glutamine and incubated at 37°C, 5% CO₂. Two thousand five hundred cells were seeded into each of the inner 60 wells of 96-well plates (#165305; Thermo) in 100 µL media and incubated for 24 h before compound treatment. Outer wells of plates were filled with 100 µL phosphate-buffered saline (PBS).

Compound Treatment

A panel of well-annotated compounds purchased from commercial suppliers (Table 2) were prepared as 10 mM stock solutions in dimethyl sulfoxide (DMSO). 1,000× compound plates were then created with semi-log dilutions in DMSO. Each plate contained six wells of 0.1% DMSO as a negative control and six wells of 200 nM staurosporine as a positive control. Following compound addition, cell assay plates were incubated at 37°C, in 5% CO₂ incubator for an additional 48 h before fixation, staining, and high-content imaging.

Imaging

We adapted the cell painting protocol from Gustafsdottir *et al.*²⁷ to optimize the cell staining across the eight selected breast cancer cell lines. Specific modifications to the original protocol by Gustafsdottir *et al.*²⁷ were implemented to circumvent morphological changes induced upon the MDA-MB-231 cell line, which was particularly sensitive to live cell staining. The modifications included using all stains on postfixed samples and adjusting concentrations of reagents to optimize staining across the cell lines. The following adapted cell painting protocol was therefore applied to our breast cancer cell panel.

After a 48-h incubation in the presence of compounds, an equal volume of 8% paraformaldehyde (PFA) was added to the culture media of each well resulting in a final concentration of 4% PFA fixation buffer; the plates were then incubated at room temperature for 20 min, followed by three washes in 100 µL PBS. Permeabilization was performed with the addition of 50 µL 0.1% Triton-X 100 to each well and incubation at room temperature for 20 min followed by three washes in 100 µL PBS.

The staining solution was prepared in a blocking buffer consisting of 1% bovine serum albumin in PBS (Table 3). Thirty microliters of staining solution was added to each well and incubated in darkness at room temperature for 30 min followed by three washes in 100 µL PBS, with no final aspiration. Plates were then sealed (#PCR-SQ plate max) and imaged immediately.

Plates were imaged on a Molecular Devices ImageXpress® Micro XLS, six fields of view were captured per well using a 20× objective and five filters, DAPI (387/447 nm), FITC (482/536 nm), Cy3 (531/593 nm), TxRed (562/642 nm), and Cy5 (628/692 nm). Exposure, binning, and other image settings were not altered between cell lines.

Image Analysis

Images were analyzed using CellProfiler v2.1.1¹⁹ to extract 309 features (Supplementary Table S1; Supplementary Data are available online at www.liebertpub.com/adt). Briefly, cell nuclei were segmented from the nuclei image based on intensity and shape, and used as seeds to segment cell areas in the other channels. Subcellular structures such as nucleoli and endoplasmic reticulum speckles were segmented and assigned to parent objects. From these objects, measurements such as size, shape, and spatial distribution were measured. The final CellProfiler settings applied in this study were created by iteratively adjusting the parameters and assessing the performance of cell segmentation by eye across multiple drug treatments for all cell types under evaluation, to ensure the most robust segmentation

Table 2. Compounds

Compound	Class	Sub-Class	Supplier	Cat. No.
Paclitaxel	Microtubule disrupting	Microtubule stabilizer	Sigma	T7402
Epothilone B	Microtubule disrupting	Microtubule stabilizer	Selleckchem	S1364
Colchicine	Microtubule disrupting	Microtubule destabilizer	Sigma	C9754
Nocodazole	Microtubule disrupting	Microtubule destabilizer	Sigma	M1404
Monastrol	Microtubule disrupting	Eg5 kinesin inhibitor	Sigma	M8515
ARQ621	Microtubule disrupting	Eg5 kinesin inhibitor	Selleckchem	S7355
Barasertib	Microtubule disrupting	Aurora B inhibitor	Selleckchem	S1147
ZM447439	Microtubule disrupting	Aurora B inhibitor	Selleckchem	S1103
Cytochalasin D	Actin disrupting	Actin disrupter	Sigma	C8273
Cytochalasin B	Actin disrupting	Actin disrupter	Sigma	C6762
Jasplakinolide	Actin disrupting	Actin stabilizer	Tocris	2792
Latrunculin B	Actin disrupting	Actin stabilizer	Sigma	L5288
MG132	Protein degradation	Proteasome	Selleckchem	S2619
Lactacystin	Protein degradation	Proteasome	Tocris	2267
ALLN	Protein degradation	Cysteine/calpain	Sigma	A6165
ALLM	Protein degradation	Cysteine/calpain	Sigma	A6060
Emetine	Protein synthesis	Protein synthesis	Sigma	E2375
Cycloheximide	Protein synthesis	Protein synthesis	Sigma	1810
Dasatinib	Kinase inhibitor	Src-EMT	Selleckchem	S1021
Saracatinib	Kinase inhibitor	Src-EMT	Selleckchem	S1006
Lovastatin	Statin	Statin	Sigma	PHR1285
Simvastatin	Statin	Statin	Sigma	PHR1438
Camptothecin	DNA damaging agent	Topoisomerase 1 inhibitor	Selleckchem	S1288
SN38	DNA damaging agent	Topoisomerase 1 inhibitor	Selleckchem	S4908

Src-EMT, Src kinase and Epithelial-Mesenchymal Transition inhibitor.

for each distinct cell type, and drug-induced phenotype is achieved.

Data Preprocessing

Out of focus and low-quality images were detected and removed by filtering on saturation and focus measurements provided in the CellProfiler output. Image averages of single object measurements from CellProfiler were aggregated by taking the median of each measured feature per image.

Features were normalized and standardized on a plate-by-plate basis by dividing each feature by the median DMSO response for that feature and then scaled by a z-score to have a mean of 0 and a standard deviation of 1. Feature selection was performed by calculating pair-wise correlations of features and removing one of a pair of features that have correlation greater than 0.9 and removing features with very low or zero variance, using the findCorrelation and nearZeroVar functions in the caret R package.²⁸

Quantifying Differential Morphological Responses by TCCS

Principal component analysis (PCA) was performed and the data were then centralized to the DMSO centroid. This was carried out by calculating the mean of principal component (PC) 1 and 2 for the DMSO subset of the data, and then subtracting this from the PC values. With each data point as a vector in two-dimensional (2D) space formed by the first two PCs, the norm of each vector was calculated, returning a Euclidean distance of each data point from the DMSO centroid. Then, the angles between each vector and a reference vector (0, 1) were calculated and denoted as theta (θ). The reference vector is arbitrarily set as a vector along the x-axis and enables easy comparison

between the polar coordinate histograms of the PCA biplot in Cartesian coordinates. For replicates, median values of PCs were calculated before calculating vectors; this simple approach avoids the pitfalls in calculating the mean of circular quantities—for example the arithmetic mean of 1° and 359° is 180°, despite the close proximity of the values in polar coordinates.

As any perturbations that do not produce morphological changes will be indistinguishable from negative control values, these points were found clustered within the negative

A METHOD TO QUANTIFY PHENOTYPIC RESPONSES BETWEEN CELL TYPES

Table 3. Stains and Concentrations Used in the Modified Cell-Painting Protocol

Stain	Structure Labeled	Wavelength (ex/em [nm])	Concentration	Cat. No.; Supplier
Hoechst 33342	Nuclei	387/447	2 µg/mL	#H1399; Mol. Probes
SYTO14	Nucleoli	531/593	3 µM	#S7576; Invitrogen
Phalloidin 594	F-actin	562/624	0.85 U/mL	#A12381; Invitrogen
Wheat germ agglutinin 594	Golgi and plasma membrane	562/624	8 µg/mL	#W11262; Invitrogen
Concanavalin A 488	Endoplasmic reticulum	462/520	11 µg/mL	#C11252; Invitrogen
MitoTracker DeepRed	Mitochondria	628/692	600 nM	#M22426; Invitrogen

control cloud in a scatter-plot of the first two PCs. As these compounds are centered on the origin (0, 0), the angles calculated from their vectors are uniformly distributed in all directions and meaningless as a phenotypic direction. Therefore, a minimum distance from the DMSO centroid was determined as 1 standard deviation of the vector distances from the origin, and compounds within this distance were defined as inactive in our assay and not used in further calculations. Active compounds were only included if they fell beyond this minimum limit for all the eight cell lines.

To calculate the phenotypic difference between compounds tested within the same cell line or a compound tested across different cell lines using the vector analysis described above, the absolute difference between the two theta values can be used. However, as any difference greater than 180° and approaching 360° starts to reflect morphologies becoming more similar, the absolute difference values have to be constrained between 0° and 180°; this is carried out for values greater than 180 by subtracting the value from 360, for example, 190° will become 170°. We named the method “Theta-Comparative-Cell-Scoring” to reflect the use of vectors applied to multi-parametric high-content data to quantify distinct phenotypic response between cell types.

Data and Code Availability

The CellProfiler pipelines, numeric data, and R code to run the analyses and generate the figures are available at github.com/swarchal/TCCS_paper

RESULTS

High-Content Phenotypic Comparisons Between Morphologically Distinct Breast Cancer Cell Subtypes

We have modified the cell painting assay previously applied to the osteosarcoma cell line U2OS cells²⁷ to a panel of breast

cancer cell lines representing clinically relevant subtypes. Eight breast cancer cell lines representing four pairs for each of the following clinical subtypes: estrogen receptor (ER)-positive, triple negative, human epidermal growth factor receptor 2 (Her2)-positive/Phosphatase and tensin homolog (PTEN) and phosphoinositide 3-kinase (PI3K) wild type, and Her2-positive/PTEN and PI3Kmut were selected for this study (Table 1).

The modified cell painting assay was optimized to enable the CellProfiler image analysis software to segment individual cells for each well and extract features, which provide detailed morphological analysis of individual breast cancer cell phenotypes. Representative images of the eight breast cancer cells stained with the modified cell-painting protocol are displayed in three channels in Figure 1A and respective cell segmentation masks generated by CellProfiler analysis are shown in Figure 1B. As the breast cancer cell lines look inherently different from one another (Fig. 1), detecting differential phenotypic changes between them requires normalization against the negative control phenotype for each cell line. This was performed by dividing each feature by the median DMSO value for that feature on a plate-by-plate basis followed by z-scoring each feature individually for all cell lines. Normalization in this manner achieved two objectives: (1) removing any batch effects that may be present across plates and (2) normalizing all phenotypic measurements as standardized fold changes from the negative control values per cell line. PCA was then performed on the normalized dataset of all cell lines using the prcomp function in R.

Quantifying Differential Morphological Response Between Cell Lines to the Same Compound

When the first two PCs are visualized as a 2D scatter plot, low concentrations of compounds are typically found near or within the DMSO cluster. However, with increasing concentrations, the points are often seen to proceed toward a given trajectory, describing decreasing phenotypic similarity to the negative control cells with increasing compound concentration. In the case of MDA-MB-231 cells treated with Cycloheximide and Barasertib, the compounds result in trajectories with opposing directions, describing opposite morphological changes (Fig. 2). The case of Barasertib and Cycloheximide provide a proof-of-principal example in the ability of the

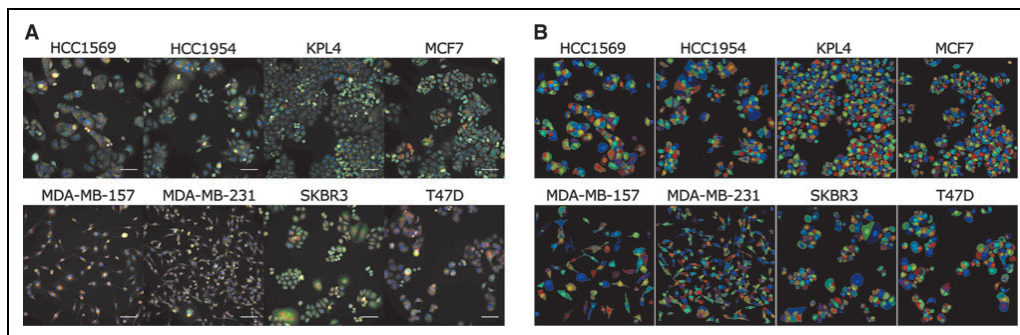


Fig. 1. Cell painting assay applied to eight distinct breast cancer cell lines. **(A)** Composite image of cell lines treated with 0.1% DMSO. Channels used: Red—MitoTracker DeepRed (mitochondria); Green—Concanavalin A (endoplasmic reticulum); Blue—Hoechst33342 (nuclei). Scale bars: 100 μm. **(B)** Image masks from CellProfiler showing nuclei and cell body segmentation. DMSO, dimethyl sulfoxide.

method described to distinguish opposing phenotypes represented by enlarged and aneuploidy nuclei characteristic of cytokinesis defects elicited by inhibitors of Aurora kinase B (Barasertib) in contrast to the condensed nuclei characteristic of the protein synthesis inhibitor (Cycloheximide).

These distinct phenotypic trajectories have been quantified as theta values against a reference vector using Equation (1), where u is the PC1, PC2 vector, and v is the reference vector of

(0, 1) (Fig. 2). A circular histogram of the theta values can then be plotted to visualize the distribution of compound induced phenotypes. The circular histogram theta plots provide an intuitive indication of a phenotypic direction produced by a specific pharmacological perturbation, as well as any change in phenotypic direction across increasing concentrations that may indicate off-target effects. Figure 3A shows a circular histogram of the data pooled from all eight cell lines treated with an eight-point half-log dose response of the Aurora B kinase inhibitor Barasertib. Using the same directional histograms, data can also be split by cell lines to directly visualize differential phenotypic response across a panel of distinct cell lines (Fig. 3B).

The difference in theta values between cell lines can then be calculated for a given compound to provide a univariate theta metric of phenotypic dissimilarity between cell types (Fig. 3C). It is possible to rank similarity and dissimilarity of each compound-induced phenotype between cells or between other compounds on a scale of 0–180° where 0 describes the most similar phenotypes and 180 the most dissimilar phenotypes. We name this method “Theta Comparative Cell Scoring” and provide the formula below:

$$\theta = \cos\left(\frac{u \cdot v}{\|u\| \|v\|}\right) \times \frac{180}{\pi} \quad (1)$$

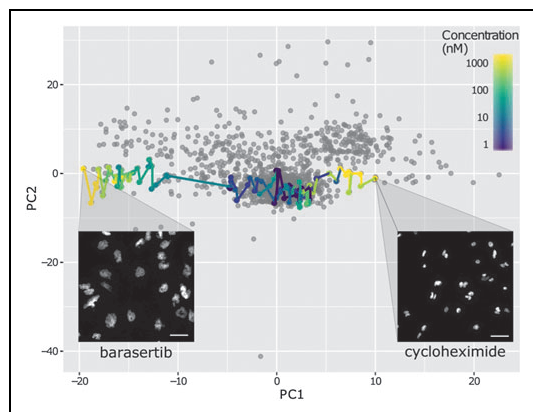


Fig. 2. Phenotypic directions in the first two PCs. Scatter plot of the first two PCs of MDA-MB-231 cells treated with a small compound library. Principal component analysis was carried out on 309 median normalized features extracted from cellular images. Barasertib and Cycloheximide compounds are colored by concentration demonstrating opposite phenotypic directions in PC space. Images show nuclei imaged with Hoechst, scale bars: 20 μm. PC, principal component.

Screening for Differential Phenotypic Response Across the Panel of Breast Cancer Subtypes

To evaluate the TCCS method for the ability to identify compounds that induce differential phenotypic responses between the breast cancer cell lines, we calculated the difference between theta values for all eight breast cancer cell lines treated with 1 μM of 24 different compounds. Compounds

A METHOD TO QUANTIFY PHENOTYPIC RESPONSES BETWEEN CELL TYPES

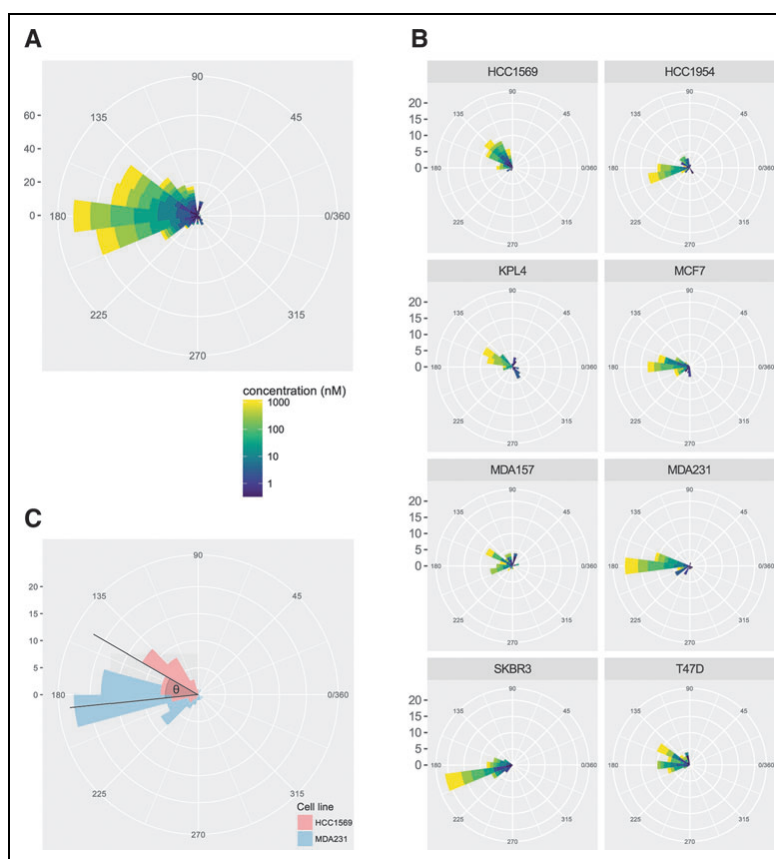


Fig. 3. Circular histograms of theta values. **(A)** Circular histogram of theta values of Barasertib calculated for all eight cell lines. **(B)** Phenotypic direction of cell lines treated with Barasertib stratified by cell line. **(C)** A diagrammatic explanation of the theta value showing the difference in theta values between HCC1569 and MDA-MB-231 cell lines treated with Barasertib.

were selected to represent 12 pairs of well-characterized mechanistic subclasses, 21 of these compounds elicited robust morphological changes in all eight cell lines.

To identify and quantify differential phenotypic responses, the difference between theta values was calculated for all pairs of cell lines, constrained to the maximum dissimilarity value of 180° and plotted as a heat map for each of the 21 compounds (Fig. 4). Compounds with high theta values indicate a differential response between pairs of cell lines for that particular compound. A representative image between KPL4 and MCF7 cells treated with 1 μM of the topoisomerase I inhibitor SN38 is

an example of a compound that induces a distinct phenotypic response between these cell types (TCCS = 179°), relative to the negative control for each cell line (Fig. 4). The majority of cell line comparisons returned low TCCS values, indicating that most of the breast cancer cell lines selected respond similarly to the compounds in our panel (Supplementary Fig. S1).

Differential Response of Breast Cancer Cell Lines Are Stratified by Molecular Subclass

To demonstrate the ability of the TCCS method to cluster high-content phenotypic response across breast cancer subtypes with a view to informing disease positioning and personalized medicine strategies, we used data from an exemplar molecular targeted therapy, the dual Src/Abl inhibitor Saracatinib (AZD0530).

To utilize the data present across multiple titrations, the mean PCs were taken across eight concentrations to create the 2D vector with which the difference between TCCS values across all pairs of cell lines is calculated. TCCS values are plotted as a heat map clustered by hierarchical clustering using Euclidean distance (Fig. 5A). This revealed that the divergent high-content phenotypic response induced by Saracatinib across the breast cancer cell panel clustered together based on their molecular subclass. Figure 5B shows images of three cell lines treated with either DMSO negative control or 1 μM Saracatinib. From Figure 5A the MDA-MB-231 cell lines are found to have responded differently to KPL4 and SKBR3 cell lines, which in turn elicited a similar response to one another. This can be seen predominantly through increased cell-cell contact in the Saracatinib-treated MDA-MB-231 cells compared to the other two cell lines, observed as an increase in normalized number of adjacent cells in MDA-MB-231 cells (Supplementary Fig. S2). Although far from

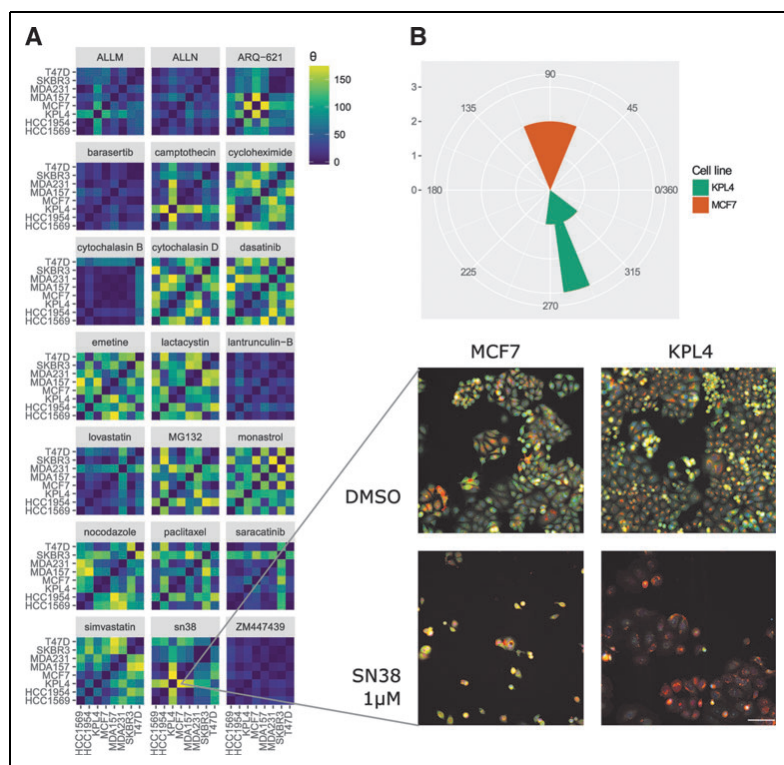


Fig. 4. Heat map of theta values between pairs of cell lines for separate compounds. **(A)** Difference in theta values calculated between pairs of cell lines treated with 21 compounds at 1 μ M concentration. Images show differential response between KPL4 and MCF7 cell lines treated with 1 μ M SN38. MCF7 cells are observed to decrease in cell area, with bright staining for the endoplasmic reticulum, whereas the KPL4 cells produce a "fried egg" morphology with large spread cells and weak endoplasmic reticulum staining. Channels used are as follows: Red—MitoTracker DeepRed (mitochondria); Green—Concanavalin A (endoplasmic reticulum); Blue—Hoechst33342 (nuclei). Scale bar: 100 μ m. **(B)** Circular histogram of theta values calculated for MCF7 and KPL4 cells treated with 1 μ M SN38.

representative of all compound responses and disease subtypes, this example does indicate the potential of high-content cell-based phenotypic screening combined with application of the TCCS method across genetically defined cell panels to provide patient stratification hypothesis for both well-characterized candidate drugs or poorly characterized active compounds identified from phenotypic screens.

DISCUSSION

The rapid evolution and convergence of new technologies, including advances in image-based high-content phenotypic

screening, induced pluripotent stem cell (iPSC) technologies, and gene editing, are well placed to advance a new era of modern phenotypic screening in more informative and disease relevant cell-based models of disease.^{15,29,30} However, a limitation of phenotypic screening is the identification of hit molecules or candidate drugs without knowledge of the target mechanism.

The lack of information on target mechanism, while not required for drug approval, impedes the design of personalized healthcare strategies to combat disease heterogeneity. Several target deconvolution strategies have been applied to compounds discovered by phenotypic screening to elucidate target mechanisms.^{31–33} However, no target deconvolution method is conclusive, and such strategies are often based upon the assumptions that a compound will only inhibit a single target and monitoring the activity and inhibition of the elucidated target will guide personalized therapy.

For the majority of compounds discovered by phenotypic screens, and for many complex human diseases where the one-drug-one-target hypothesis is unrealistic, new nontarget-centric approaches are required to understand drug mechanism-of-action and guide

personalized healthcare strategies. *In vitro* pharmacogenomic or pharmacoproteomic profiling across well-characterized cell panels, representing specific disease subtypes, exemplifies one approach for informing drug mechanism-of-action and guiding personalized healthcare strategies in the absence of target knowledge. Breast cancer is separated into four major molecular subtypes; Luminal A (ER-positive and/or progesterone receptor (PR)-positive and HER2-negative and Low Ki67); Luminal B (ER-positive and/or PR-positive and HER2-positive or HER2-negative with high Ki67); Triple negative/basal like (ER- PR- and Her2-negative); and HER2 type (ER- PR- negative and

A METHOD TO QUANTIFY PHENOTYPIC RESPONSES BETWEEN CELL TYPES

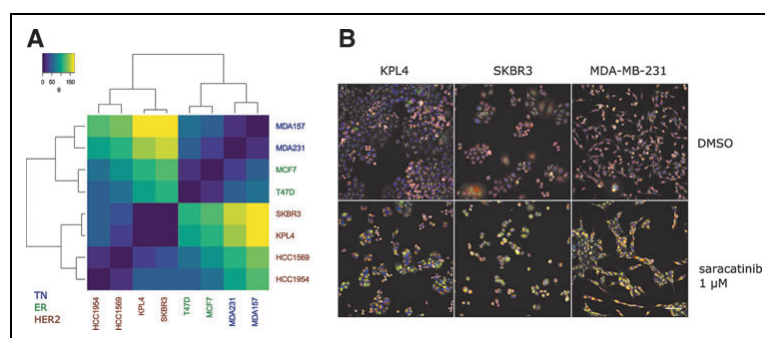


Fig. 5. Heat map and hierarchical clustering of cell lines treated with Saracatinib. **(A)** Heatmap of TCCS values calculated between all pairs of cell lines treated with Saracatinib with hierarchical clustering by complete linkage of the Euclidean distance. **(B)** Images demonstrating two similar phenotypic responses—KPL4 and SKBR3—and the dissimilar phenotypic response of MDA-MB-231 cell lines to 1 μ M Saracatinib treatment. Channels used: Red—MitoTracker DeepRed (mitochondria); Green—Concanavalin A (endoplasmic reticulum); Blue—Hoechst33342 (nuclei). Scale bar: 100 μ m. TCCS, Theta Comparative Cell Scoring.

Her2-positive). Each major molecular subtype of breast cancer can be further divided into subclasses based upon genetic mutation status and protein profiles, and the diagnosis of breast cancer subtype dictates the most appropriate personalized treatment for patients.^{34–36}

In this article, we have developed a multiparametric high-content assay, data visualization tools, and a TCCS method, which support phenotypic screening of compound libraries across genetically distinct cells representing known molecular subtypes of disease. We provide proof-of-principle data applied to eight breast cancer cells representing four disease subclasses (Table 1), demonstrating the application of the method for quantifying distinct phenotypic response between cell types and clustering of cell-associated clinical subtypes based on similar or dissimilar phenotypic response to compound treatment.

As previously discussed, several multiparametric pathway and phenotypic profiling methods have been developed to classify drug mechanism-of-action and uncover new drug-target associations, and structure activity relationships in a more holistic and unbiased manner.^{18,20–25,27} However, the majority of these methods have been applied to single cell types amenable to high-content imaging or large-scale biochemical and proteomic analysis.^{18,21–25,27} The TCCS method described in this article was developed to provide a practical method to enable comparative multiparametric phenotypic analysis across a panel of genetically distinct cell types, which provides rapid quantification and visualization of divergent compound-induced phenotypic response between cell types. An intuitive explanation of the TCCS method would be the cosine distance in

degrees of vectors in the first two PCs; this is a variation on existing methods that largely rely on correlation or Euclidean distance between compound vectors.¹⁸

The benefits of the TCCS over previous methods are as follows: (1) use of distance from the negative control to remove poorly active or inactive compounds that might produce spurious differences in correlation of cosine similarity measures; (2) The comparison of each data point to a common reference vector enables visualizations of a single metric, which depicts the relative change in phenotypic response induced by a compound (Fig. 3A).

The most critical aspect of comparing results between panels of

distinct cell lines regardless of downstream methods is during the data preprocessing stage, which requires careful normalization against the negative control values for each cell line to remove inherent differences in cell line morphology. Thus, the TCCS method represents a flexible approach with broad applicability to quantifying and visualizing distinct phenotypes induced by a panel of compounds within a single cell type and/or the response of a single compound across multiple cell types. The TCCS method removes compounds from the algorithm that are not sufficiently different from the negative control. While this increases the robustness of the calculation, it also creates the opportunity to miss compounds that possess differential sensitivity between cell lines. This limitation of the method arises where certain compounds that do not induce any morphological change in one cell line may still perturb cellular morphology in another cell line, thus any such compound would subsequently be removed from the calculation due to insufficient distance from the negative control centroid, despite eliciting a genuine differential response between cell lines. However, this limitation can be simply rectified by implementing an initial preanalytical stage of the algorithm by calculating the distance from DMSO for all compounds across all cell lines to assign either as “active” or “inactive” phenotypic responders. Differences in the activation state of all compounds across all cell lines are recorded and the active compounds then progress to TCCS analysis to quantify and visualize a distinct phenotype response across cell lines.

The TCCS method as outlined in this article utilized only the first two PCs produced from the PCA. These two variables

explain most of the variance of data in low dimensional data represented by majority of high-throughput high-content screens, which typically measure only small numbers of features.³⁷ In such high-throughput compound screens, TCCS applied to the first two PCs would be expected to provide a single value describing the difference in response across different cell lines for active compounds. The method as applied to the first two PCs in this article becomes less informative in higher dimensional data sets as more PCs are required to describe the data. As the calculation to define the angle between two vectors [Eq. (1)] uses the dot product of the two vectors, the vectors are not limited to the first two PCs, and it is entirely reasonable that they could contain any number of PCs. Therefore, an alternative option would be to implement the TCCS method on a number of PCs that satisfy a user-defined proportion of variance within the data.

Comparison of high dimensional vectors against one another rather than against a reference vector allows for direct calculation of a theta value in high dimensional space, an example workflow using the TCCS method applied to more than two PCs is provided in the online R scripts (github.com/swarchal/TCCS_paper) and is represented in the description of the TCCS workflow (Fig. 6). The TCCS method may also be applied to the normalized assay parameters rather than PCs as also demonstrated in the supplementary R workflow (github.com/swarchal/TCCS_paper). However, care should be taken to ensure that potentially uninformative parameters are

not included in such analysis to avoid introduction of unnecessary assay noise. Thus, the most optimal application of the TCCS method can be appropriately tailored to each study and nature of the underlying high-content data set.

Multiple concentrations are not often used in high-throughput cell-based screening assays, despite providing useful information to detect off-target effects and can be thought of as inherent replicates of individual compound data. A further approach to incorporate titration data into defining direction in PC space would be to fit a linear model to each compound using simple linear regression, forcing the y-intersect through 0. While this would lose information pertaining to the distance from the DMSO centroid at each concentration, it would provide information regarding goodness of fit, and data may be excluded from the TCCS analysis if they do not fit the linear model well or used to indicate compounds with off-target effects at higher concentrations. As the theta value is essentially a direction in PC space, another useful addition would be to relate theta back to the feature loadings that describe how the PCs were constructed. This would return the phenotypic features that best describe a certain direction in phenotypic space. However, PCA contains negatively weighted features and so methods such as nonnegative matrix factorization in which the feature loadings are all positive values, may be a potential avenue for this improvement.

Another potential use of TCCS method is in assay quality control (QC). For example, TCCS could be applied to the

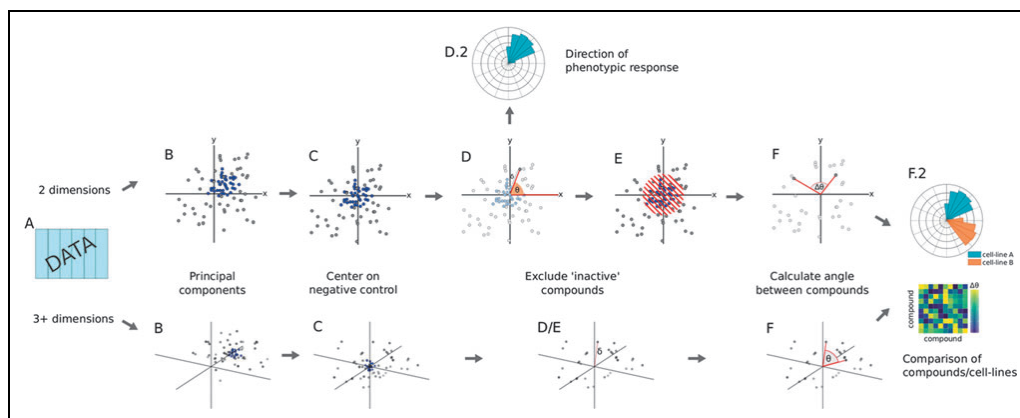


Fig. 6. TCCS workflow. (A) Normalized numerical data. (B) PC analysis, negative control values colored in blue. (C) Centering of PC values to the negative control centroid. (D) Calculation of distance from the origin to each data point, an activity cutoff is derived from the standard deviation of the distance to the negative control values. (D.2) In two-dimensional space, a directional histogram can be created by the angle of each vector against a reference vector. (E) Inactive compounds excluded based on distance from the origin. (F) Determining the angle between compounds. (F.2) Visualization or clustering of compounds based on theta values.

A METHOD TO QUANTIFY PHENOTYPIC RESPONSES BETWEEN CELL TYPES

simultaneous evaluation of two positive controls known to elicit robustly different morphologies (e.g., paclitaxel and staurosporine) along with a negative control such as DMSO to determine a theta value between the two positive controls. It would be expected that the two positive controls would have a theta value greater than a specified minimum. The variance of theta values between two positive controls per plate could therefore be used as a measure of biological assay variability during assay development and screening campaigns.

Incorporating a multiparametric QC metric that utilizes high-content analysis across two positive controls provides increased robustness and more unbiased assessment of monitoring variation in cell behavior and assay variability over current methods that use a single positive control analysis of a pre-selected parameter. Other multivariate assay QC metrics typically build on the Z' -factor using supervised machine learning techniques such as Fisher's linear discriminant analysis (LDA) to best separate the positive and negative controls.³⁸ Although more robust than single parametric analysis, a drawback of this method is that LDA is often prone to overfitting in high dimensions, which may produce overoptimistic QC values when processed to the Z' -factor calculation.

The convergence of new technologies, including next-generation sequencing, high-throughput proteomics, iPSC technology, and high-content phenotypic screening, is well placed to advance the identification of predictive biomarkers and personalized medicine approaches across a broader range of disease types and therapeutic classes.^{15,29,30,39,40}

Our study provides a broadly applicable approach for quantifying distinct phenotypic response between genetically distinct cells using high-content analysis coupled to a TCCS scoring method. The TCCS method that we describe provides a univariate metric that can be applied to any high-content assay for quantifying and visualizing a diverse phenotypic response between cell types. The TCCS metric provides a univariate score of distinct phenotypic response on a scale of 0–180° (where 0° = similar and where 180° = most dissimilar), which can be used for correlation with orthogonal genetic, epigenetic, and proteomic datasets to support the identification of biomarkers of drug phenotype and further elucidate drug mechanism-of-action at genetic and pathway levels.

ACKNOWLEDGMENTS

Cancer Research UK Edinburgh Centre studentship award to S.J.W. and Research Councils UK (RCUK) fellowship award to N.O.C.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

1. Fisher R, Pusztai L, Swanton C: Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer* 2013;108:479–485.
2. McClellan J, King MC: Genetic heterogeneity in human disease. *Cell* 2010;141:210–217.
3. Aronson SJ, Rehm HL: Building the foundation for genomics in precision medicine. *Nature* 2015;526:336–342.
4. Biankin AV, Plantadosi S, Hollingsworth SJ: Patient-centric trials for therapeutic development in precision oncology. *Nature* 2015;526:361–370.
5. Day F, Muranyi A, Singh S, et al.: A mutant BRAF V600E-specific immunohistochemical assay: correlation with molecular mutation status and clinical outcome in colorectal cancer. *Target Oncol* 2015;10:99–109.
6. Maemondo M, Inoue A, Kobayashi K, et al.: Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* 2010;362:2380–2388.
7. Ross JS, Slodkowska EA, Symmans WF, Pusztai L, Ravdin PM, Hortobagyi GN: The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. *Oncologist* 2009;14:320–368.
8. Lee JA, Berg EL: Neoclassic drug discovery: the case for lead generation using phenotypic and functional approaches. *J Biomol Screen* 2013;18:1143–1155.
9. Swinney DC: The contribution of mechanistic understanding to phenotypic screening for first-in-class medicines. *J Biomol Screen* 2013;18:1186–1192.
10. Staunton JE, Slonim DK, Coller HA, et al.: Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A* 2001;98:10787–10792.
11. Rees MG, Seashore-Ludlow B, Cheah JH, et al.: Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2016;12:109–116.
12. Cardnell RJ, Feng Y, Diao L, et al.: Proteomic markers of DNA repair and PI3K pathway activation predict response to the PARP inhibitor BMN 673 in small cell lung cancer. *Clin Cancer Res* 2013;19:6322–6328.
13. Haibe-Kains B, El-Hachem N, Birkbak NJ, et al.: Inconsistency in large pharmacogenomic studies. *Nature* 2013;504:389–393.
14. Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium: Pharmacogenomic agreement between two cancer cell line data sets. *Nature* 2015;528:84–87.
15. Bickle M: The beautiful cell: high-content screening in drug discovery. *Anal Bioanal Chem* 2010;398:219–226.
16. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ: Multidimensional drug profiling by automated microscopy. *Science* 2004;306:1194–1198.
17. Tanaka M, Bateman R, Rauh D, et al.: An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol* 2005;3:e128.
18. Ljosa V, Caie PD, Ter Horst R, et al.: Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J Biomol Screen* 2013;18:1321–1329.
19. Carpenter AE, Jones TR, Lamprecht MR, et al.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* 2006;7:R100.
20. Caie PD, Walls RE, Ingleston-Orme A, et al.: High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol Cancer Ther* 2010;9:1913–1926.
21. Reisen F, Sauty de Chalon A, Pfeifer M, Zhang X, Gabriel D, Selzer P: Linking phenotypes and modes of action through high-content screen fingerprints. *Assay Drug Dev Technol* 2015;13:415–427.
22. Reisen F, Zhang X, Gabriel D, Selzer P: Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery. *J Biomol Screen* 2013;18:1284–1297.
23. Fliri AF, Loging WT, Thadeio PF, Volkmann RA: Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J Med Chem* 2005;48:6918–6925.
24. Fliri AF, Loging WT, Thadeio PF, Volkmann RA: Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc Natl Acad Sci U S A* 2005;102:261–266.

WARCHAL, DAWSON, AND CARRAGHER

25. Kummel A, Selzer P, Siebert D, et al.: Differentiation and visualization of diverse cellular phenotypic responses in primary high-content screening. *J Biomol Screen* 2012;17:843–849.
26. Smith K, Horvath P: Active learning strategies for phenotypic profiling of high-content screens. *J Biomol Screen* 2014;19:685–695.
27. Gustafsdottir SM, Ljosa V, Sokolnicki KL, et al.: Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* 2013;8:e80999.
28. Kuhn M: Building predictive models in R using the caret package. *J Stat Software* 2008;28:1–26.
29. Yu J, Vodyanik MA, Smuga-Otto K, et al.: Induced pluripotent stem cell lines derived from human somatic cells. *Science* 2007;318:1917–1920.
30. Shalem O, Sanjana NE, Hartenian E, et al.: Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 2014;343:84–87.
31. Lee J, Bogoy M: Target deconvolution techniques in modern phenotypic profiling. *Curr Opin Chem Biol* 2013;17:118–126.
32. Rix U, Superti-Furga G: Target profiling of small molecules by chemical proteomics. *Nat Chem Biol* 2009;5:616–624.
33. Jafari R, Almqvist H, Axelsson H, et al.: The cellular thermal shift assay for evaluating drug target interactions in cells. *Nat Protoc* 2014;9:2100–2122.
34. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.
35. Nik-Zainal S, Davies H, Staaf J, et al.: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;534:47–54.
36. Kao J, Salari K, Bocanegra M, et al.: Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One* 2009;4:e6146.
37. Singh S, Carpenter AE, Genovesio A: Increasing the content of high-content screening: an overview. *J Biomol Screen* 2014;19:640–650.
38. Kummel A, Gubler H, Gehin P, Beibel M, Gabriel D, Parker CN: Integration of multiple readouts into the z' factor for assay quality assessment. *J Biomol Screen* 2010;15:95–101.
39. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: The next-generation sequencing revolution and its impact on genomics. *Cell* 2013;155:27–38.
40. Akbani R, Becker KF, Carragher N, et al.: Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: a workshop report: the RPPA (Reverse Phase Protein Array) society. *Mol Cell Proteomics* 2014;13:1625–1643.

Address correspondence to:

Neil O. Carragher, PhD
Institute of Genetics and Molecular Medicine
Cancer Research UK Edinburgh Centre
University of Edinburgh
Crewe Road South
Edinburgh, EH4 2XR
United Kingdom

E-mail: n.carragher@ed.ac.uk

Abbreviations Used

2D = two dimensional
3D = three dimensional
CCLE = Cancer Cell Line Encyclopedia
DAPI = 4,6-diamidino-2-phenylindole
DMSO = dimethyl sulfoxide
EC₅₀ = concentration of drug that produces a 50% maximal response
ER = estrogen receptor
FITC = fluorescein isothiocyanate
GI₅₀ = concentration of a drug that gives half-maximal inhibition of cell proliferation
HER2 = human epidermal growth factor receptor 2
iPSC = induced pluripotent stem cell
LDA = linear discriminant analysis
PBS = phosphate-buffered saline
PC = principal component
PCA = principal component analysis
PFA = paraformaldehyde
PI3K = phosphoinositide 3-kinase
PR = progesterone receptor
PTEN = phosphatase and tensin homolog
QC = quality control
TCCS = Theta Comparative Cell Scoring
TN = triple negative
WT = wild type

OPEN

Data-analysis strategies for image-based cell profiling





Juan C Caicedo¹, Sam Cooper², Florian Heigwer³ , Scott Warchal⁴, Peng Qiu⁵, Csaba Molnar⁶, Aliaksei S Vasilevich⁷, Joseph D Barry⁸, Harmanjit Singh Bansal⁹, Oren Kraus¹⁰, Mathias Wawer¹¹, Lassi Paavolainen¹², Markus D Herrmann¹³, Mohammad Rohban¹, Jane Hung^{1,14}, Holger Hennig¹⁵ , John Concannon¹⁶, Ian Smith¹⁷, Paul A Clemons¹¹, Shantanu Singh¹, Paul Rees^{1,18}, Peter Horvath^{6,12}, Roger G Linington¹⁹ & Anne E Carpenter¹ 

Image-based cell profiling is a high-throughput strategy for the quantification of phenotypic differences among a variety of cell populations. It paves the way to studying biological systems on a large scale by using chemical and genetic perturbations. The general workflow for this technology involves image acquisition with high-throughput microscopy systems and subsequent image processing and analysis. Here, we introduce the steps required to create high-quality image-based (i.e., morphological) profiles from a collection of microscopy images. We recommend techniques that have proven useful in each stage of the data analysis process, on the basis of the experience of 20 laboratories worldwide that are refining their image-based cell-profiling methodologies in pursuit of biological discovery. The recommended techniques cover alternatives that may suit various biological goals, experimental designs, and laboratories' preferences.

Image analysis is heavily used to quantify phenotypes of interest to biologists, especially in high-throughput experiments^{1–3}. Recent advances in automated microscopy and image analysis allow many treatment conditions to be tested in a single day, thus enabling the systematic evaluation of particular morphologies of cells. A further revolution is currently underway: images are also being used as unbiased sources of quantitative information about cell state in an approach known as image-based profiling or morphological

profiling⁴. Herein, the term morphology will be used to refer to the full spectrum of biological phenotypes that can be observed and distinguished in images, including not only metrics of shape but also intensities, staining patterns, and spatial relationships (described in 'Feature extraction').

In image-based cell profiling, hundreds of morphological features are measured from a population of cells treated with either chemical or biological perturbagens. The effects of the treatment are quantified

¹Imaging Platform, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ²Imperial College London, London, UK.

³German Cancer Research Center and Heidelberg University, Heidelberg, Germany. ⁴Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, UK. ⁵Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, Georgia, USA. ⁶Synthetic and System Biology Unit, Hungarian Academy of Sciences, Szeged, Hungary. ⁷Laboratory for Cell Biology–Inspired Tissue Engineering, MERLN Institute, Maastricht University, Maastricht, the Netherlands. ⁸Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. ⁹National Centre for Biological Sciences, Bangalore, India. ¹⁰Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada. ¹¹Chemical Biology and Therapeutics Science Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. ¹²Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. ¹³Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. ¹⁴Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. ¹⁵Department of Systems Biology & Bioinformatics, University of Rostock, Rostock, Germany. ¹⁶Department of Chemical Biology and Therapeutics, Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, USA. ¹⁷Connectivity Map Project, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. ¹⁸College of Engineering, Swansea University, Swansea, UK. ¹⁹Department of Chemistry, Simon Fraser University, Burnaby, British Columbia, Canada. Correspondence should be addressed to A.E.C. (anne@broadinstitute.org).

RECEIVED 19 MAY 2016; ACCEPTED 28 JULY 2017; PUBLISHED ONLINE 31 AUGUST 2017; DOI:10.1038/NMETH.4397

by measuring changes in those features in treated versus untreated control cells⁵. By describing a population of cells as a rich collection of measurements, termed the ‘morphological profile’, various treatment conditions can be compared to identify biologically relevant similarities for clustering samples or identifying matches or anticorrelations. This profiling strategy contrasts with image-based screening, which also involves large-scale imaging experiments but has a goal of measuring only specific predefined phenotypes and identifying outliers.

Similarly to other profiling methods that involve hundreds of measurements or more from each sample^{6,7}, the applications of image-based cell profiling are diverse and powerful. As reviewed recently^{8,9}, these applications include identifying disease-specific phenotypes, gene and allele functions, and targets or mechanisms of action of drugs.

However, the field is currently a wild frontier, including novel methods that have been proposed but not yet compared, and few methods have been used outside the laboratories in which they were developed. The scientific community would greatly benefit from sharing methods and software code at this early stage, to enable more rapid convergence on the best practices for the many steps in a typical profiling workflow (Fig. 1).

Here, we document the options at each step in the computational workflow for image-based profiling. We divide the workflow into eight main steps (Fig. 1). For each step, we describe the process, its importance, and its applicability to different experimental types and scales. We present previously published methods relevant to each step, provide guidance regarding the theoretical pros and cons for each alternative option, and refer to any prior published comparisons of methods. We do not cover the upstream steps (sample preparation and image-acquisition recommendations)^{1,2} or computational practicalities such as the necessary information-technology infrastructure to store and process images or data. The workflow’s starting point is a large set of images. The assays can be specifically designed for profiling, such as Cell Painting^{10,11}, but any image-based assays can be used, including a panel of multiple parallel image-based assays¹², or time-lapse microscopy for analyzing dynamics¹³ or even whole organisms¹⁴.

This paper is the result of a ‘hackathon’, in which the authors met to discuss and share their expertise in morphological profiling. Hands-on data-analysis challenges and the accompanying discussions helped to identify the best practices in the field and to contribute algorithms to a shared code base.

We hope to provide a valuable foundation and framework for future efforts and to lower the barrier to entry for research groups that are new to image-based profiling. The detailed workflows used by each individual laboratory contributing to this article can be found online (<https://github.com/shntnu/cytomining-hackathon-wiki/wiki/>).

Step 1: image analysis

Image analysis transforms digital images into measurements that describe the state of every single cell in an experiment. This process makes use of various algorithms to compute measurements (often called features) that can be organized in a matrix in which the rows are cells in the experiment, and the columns are extracted features.

Field-of-view illumination correction. Every image acquired by a microscope exhibits inhomogeneous illumination mainly

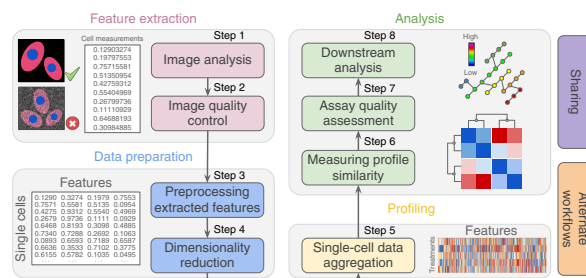


Figure 1 | Representative workflow for image-based cell profiling. Eight main steps transform images into quantitative information to support experimental conclusions.

because a nonuniform light source or optical path often yields shading around edges. This effect is often underestimated; however, intensities usually vary by 10–30%, thus corrupting accurate segmentation and intensity measurements¹⁵. Illumination correction is a process to recover the true image from a distorted one. There are three main approaches to illumination correction:

Prospective methods. These methods build correction functions from reference images, such as dark and bright images with no sample in the foreground. The approach requires careful calibration at the time of acquisition and relies on assumptions that are often inappropriate, thus yielding an incomplete correction in practice¹⁶.

Retrospective single-image methods. These methods calculate the correction model for each image individually^{17–19}. However, the result can change from image to image and thus may alter the relative intensity.

Retrospective multi-image methods. These methods build the correction function by using the images acquired in the experiment. These methods are often based on smoothing¹⁶, surface fitting²⁰, or energy-minimization models¹⁵.

Illumination correction is an important step for high-throughput quantitative profiling; the strategy of choice in most of our laboratories is a retrospective multi-image correction function. This procedure produces more robust results, particularly when separate functions are calculated for each batch of images (often with a different function for each plate and always with a different function for different imaging sessions or instruments). We recommend use of prospective and single-image methods for only qualitative experiments.

Segmentation. Typically, each cell in the image is identified and measured individually; that is, its constituent pixels are grouped to distinguish the cell from other cells and from the background. This process is called ‘segmentation’ (Fig. 2), and there are two main approaches:

Model based. The experimentalist chooses an appropriate algorithm and manually optimizes parameters on the basis of visual inspection of segmentation results. A common procedure is first to identify nuclei, as can often be done easily, and then to use the results as seeds for the identification of the cell outline. A priori

knowledge (i.e., a ‘model’) is needed, such as the objects’ expected size and shape²¹. Model-based approaches typically involve histogram-based methods, such as thresholding, edge detection, and watershed transformation²².

Machine learning. A classifier is trained to find the optimal segmentation solution by providing it with ground-truth data and manually indicating which pixels of an image belong to different classes of objects²³. This approach typically involves applying various transformations to the image to capture different patterns in the local pixel neighborhood. Segmentation is ultimately achieved by applying the trained model to new images to classify pixels accordingly.

Both approaches are used in profiling experiments. The model-based approach is most common (for example, in CellProfiler²⁴); it performs well for fluorescence microscopy images of cultured cells²². However, it requires manual parameter adjustment for each new experimental setup. Machine-learning-based segmentation (for example, in Ilastik²³) can perform better on difficult segmentation tasks, such as highly variable cell types or tissues. It does not require as much computational expertise, but it does require manual labeling of training pixels for each experimental setup and sometimes even for each batch of images. The creation of ground-truth data in the process of labeling allows for quantitative performance assessment.

Feature extraction. The phenotypic characteristics of each cell are measured in a step called feature extraction, which provides the raw data for profiling. The major types of features are:

Shape features. These features are computed on the boundaries of nuclei, cells, or other segmented compartments. These include standard size and shape metrics such as perimeter, area, and roundness^{25,26}.

Intensity-based features. These features are computed from the actual intensity values in each channel of the image on a single-cell basis, within each compartment (nucleus, cell, or other segmented compartments). These metrics include simple statistics (for example, mean intensity, and maximum intensity).

Texture features. These features quantify the regularity of intensities in images, and periodic changes can be detected by using mathematical functions such as cosines and correlation matrices. These features have been extensively used for single-cell analysis^{27–30}.

Microenvironment and context features. These features include counts and spatial relationships among cells in the field of view (on the basis of the number of and distance to cells in a neighborhood) as well as its position relative to a cell colony^{31–33}. Segmented regions are not limited to nuclei, and cells and may also include subcellular structures that can be quantified as measurements (for example, speckles within a nucleus or distances between the nucleus and individual cytoplasmic vesicles).

Whereas screening experiments typically measure one or two features of interest to quantify specific effects³⁴, cell profiling involves computing as many features as possible to select robust, concise, and biologically meaningful features to increase the

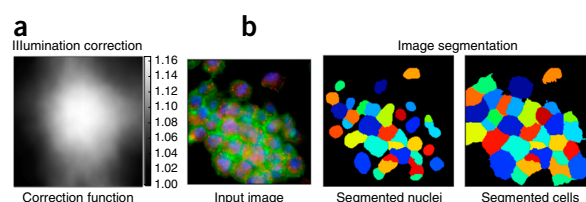


Figure 2 | Methods used for image analysis. (a) Illumination-correction function estimated with a retrospective multi-image method. Pixels in the center of the field of view are systematically brighter than pixels in the edges. (b) Image segmentation aims to classify pixels as either foreground or background, i.e. as being part of an object or not. Here, regions have been segmented with the model-based approach.

chances of detecting changes in the molecular states of cells. The most common practice is to measure hundreds or even thousands of features of many varieties; the details are typically described in the software’s documentation^{24,35,36}.

Step 2: image quality control

It is largely impossible to manually verify image quality in high-throughput experiments, so automated methods are needed to objectively flag or remove images and cells that are affected by artifacts. These methods seek to decrease the risk of contaminating the data with incorrect values.

Field-of-view quality control. Images can be corrupted by artifacts such as blurring (for example, improper autofocusing) or saturated pixels (for example, debris or aggregations that are inappropriately bright). Typically, statistical measures of image intensity are used for quality control.

Metrics can be computed to detect blurring, including the ratio of the mean and the s.d. of each image’s pixel intensities, the normalized measure of the intensity variance³⁷, and the image correlation across subregions of the image³⁸. The log–log slope of the power spectrum of pixel intensities is another effective option, because the high-frequency components of an image are lost as it becomes more blurred³⁹; this procedure has been found to be the most effective in a recent comparison for high-throughput microscopy⁴⁰. For detecting saturation artifacts, the percentage of saturated pixels has been found to be the best among all tested metrics.

We recommend computing various measures that represent a variety of artifacts that might occur in an experiment to increase the chance of artifact identification. Then, with data-analysis tools, these measurements can be reviewed to identify acceptable quality-control thresholds for each measure⁴⁰. It is also possible to use supervised machine-learning algorithms to identify problematic images^{41,42}, but these algorithms require example annotations and classifier training and validation, and thus may require more effort and introduce a risk of overfitting.

Cell-level quality control. Outlier cells may exhibit highly unusual phenotypes but may also result from errors in sample preparation, imaging, image processing, or image segmentation. Errors include incorrectly segmented cells, partly visible cells at image edges, out-of-focus cells, and staining artifacts. Although errors

are best decreased through careful techniques and protocols, there are several strategies for detecting outlier cells:

Model-free outlier detection. This strategy includes methods to define normal limits by using statistics. Data points represented with a single variable (for example, distance values or single features) can be analyzed with univariate statistical tools, including the 3- or 5-s.d. rules, Winsorizing, and the adjusted box-plot rule⁴³. Robust statistics based on estimators such as the median and the median absolute deviation⁴⁴ can also be used and extended to multivariate situations⁴⁵. Additional multivariate methods include principal component analysis (PCA) and Mahalanobis-based outlier detection⁴⁶.

Model-based outlier detection. This strategy involves training a model of normal samples to aid in detecting outlier cells⁴⁷. For instance, if a linear regression among features is suitable, outliers can be detected as data points with a large residual that does not follow the general trend⁴⁸. Alternately, a supervised-machine-learning classifier can be trained by providing examples of outliers^{49–51}.

After they are detected, outlier cells can be removed, or when the number of outliers in the sample is too high, the entire sample can be examined manually or omitted from analysis^{47,52}. Importantly, cell-outlier detection should be performed at the whole-population level; that is, it should not be separately configured per well, per replicate, or per plate. Extreme caution is recommended, to avoid removing data points that represent cells and samples with interesting phenotypes^{53,54}. Samples can be composed of various subpopulations of cells, and outlier-detection methods may incorrectly assume normality or homogenous populations (Fig. 3). For this reason, most laboratories skip outlier detection at the level of individual cells, other than to check for segmentation problems.

Step 3: preprocessing extracted features

Preparing extracted cell features for further analysis is a delicate step that can enhance the observation of useful patterns or can corrupt the information and lead to incorrect conclusions.

Missing values. Feature-extraction software may yield non-finite symbols (such as NaN and INF) representing incomputable values. In general, use of these symbols is preferred to assigning a numerical value that could be interpreted as having a phenotypic meaning. The presence of non-finite symbols poses challenges to applying statistics or machine-learning algorithms. There are three alternate solutions for handling missing values:

Removing cells. If a small proportion of cells have missing values, excluding them can be considered. However, those cells may indicate a valid and relevant phenotype, a possibility that should be assessed carefully (described in ‘Cell-level quality control’).

Removing features. If a large proportion of cells have a missing value for a particular feature, they might be removed on the grounds that the feature is insufficiently informative. Again, this removal should be assessed carefully for its effect on unexpected cell phenotypes.

Applying imputation. If the proportion of cells with missing values for certain features is relatively small, several statistical rules may

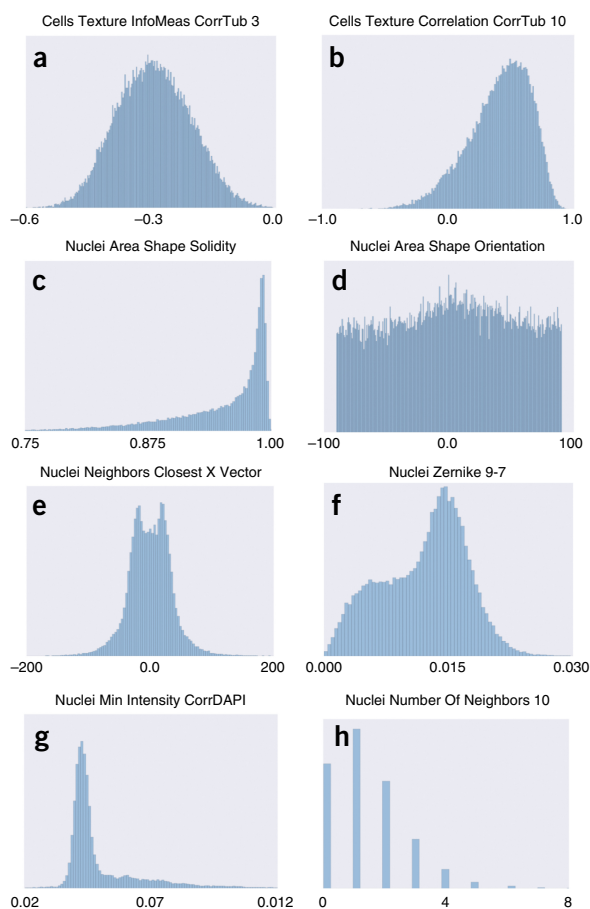


Figure 3 | Diversity of feature distributions in morphological profiling. (a–h) Morphological features display various types of distributions, including normal (a), skewed (b,c), uniform (d), multimodal (e–g), and even discrete distributions (h). The ranges in which features are represented also vary considerably. These histograms were obtained with feature values from a sample of 10,000 cells in the BBBC021 data set¹⁰⁸. The names of features correspond to conventions used in the CellProfiler software. The x axes show feature values (in different units), and the y axes show frequencies (cell counts).

be applied to complete these values. The use of zeros or the mean value is common in general statistical analysis but should not be the default option for single-cell profiling. If too many values are artificially added to the data matrix, the downstream analysis may be affected or biased by false data.

Deciding how to proceed with missing values is primarily dependent on experimental evaluations and empirical observations. Removing cells or features is more common than applying imputation. However, there is no single rule that applies in all cases, and the best practice is to collect convincing evidence supporting these decisions, especially with the use of quality measures and replicate analysis (described in ‘Downstream analysis’).

Plate-layout-effect correction. High-throughput assays use multiwell plates, which are subject to edge effects and gradient artifacts. Concerns regarding spatial effects across each plate are

not unique to imaging and have been widely discussed in both the microarray-normalization and high-throughput-screening literature^{44,55–58}. They can be decreased to some degree at the sample-preparation step⁵⁹.

We recommend checking for plate effects to determine whether any artifacts are present within plates or across multiple batches. The simplest method is a visual check, through plotting a measured variable (often cell count or cell area) as a heat map in the same spatial format as the plate; this procedure allows for easy identification of row and column effects as well as drift across multiple plates.

We recommend using a two-way median polish to correct for positional effects. This procedure involves iterative median smoothing of rows and columns to remove positional effects, then dividing each well value by the plate median absolute deviation to generate a *B* score⁶⁰. However, this procedure cannot be used on nonrandom plate layouts such as compound titration series or controls placed along an entire row or column⁵⁴. Other approaches include 2D polynomial regression and running averages, both of which correct spatial biases by using local smoothing⁶¹. Notably, image-based profiling is often sufficiently sensitive to distinguish among different well positions containing the same sample. Thus, to mitigate these positional effects, samples should be placed in random locations with respect to the plate layout. However, because such scrambling of positions is rarely practical, researchers must take special care to interpret results carefully and to consider the effects that plate-layout effects might have on the biological conclusions.

Batch-effect correction. Batch effects are subgroups of measurements that result from undesired technical variation (for example, changes in laboratory conditions, sample manipulation, or instrument calibration) rather than constituting a meaningful biological signal (Fig. 4). Batch effects pose a major challenge to high-throughput methodologies, and correction is an important preliminary step; if undetected, batch effects can lead to misinterpretation and false conclusions⁶².

We recommend identifying batch effects by inspecting correlations among profiles (described in ‘Single-cell data aggregation’). Specifically, by plotting heat maps of the correlation between all pairs of wells within an experiment, sorted by experimental repeat, batch effects can be identified as patterns of high correlation corresponding to technical artifacts (Fig. 4a). As a quantitative check, within-plate correlations should be in the same range as across-plate correlations.

When correction is needed, standardization and quantile normalization, as discussed in ‘Feature transformation and normalization’, can be applied within plates rather than to the entire screen⁶³. This procedure should be performed only if samples are relatively randomly distributed across plates. Canonical correlation analysis can also be used to transform data to maximize the similarity between technical replicates across experiments^{64,65}. Nonetheless, care should be taken to ensure that batch effects have been correctly decreased without false amplification of other sources of noise.

Feature transformation and normalization. Morphological profiles include features that display varying shapes of statistical distributions⁶⁶. It is therefore essential to transform feature values

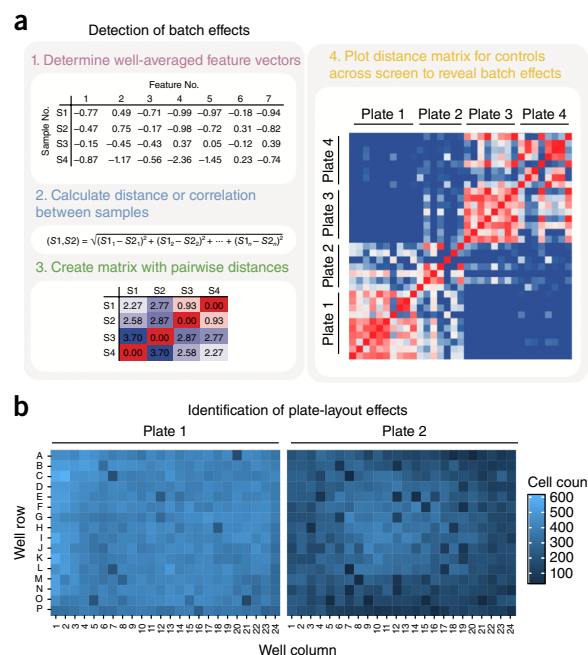


Figure 4 | Example diagnostic plots for detecting batch effects and plate-layout effects. **(a)** Process of detecting batch effects. The largest matrix on the right shows how plates 1 and 2 are more correlated to each other than to plates 3 and 4, and vice versa. This pattern suggests that plates 1 and 2, as well as 3 and 4, were prepared in batches that have noticeable differences in their experimental conditions. **(b)** Two plate layouts illustrating the cell count in each well. The visualization allows for identification of plate-layout effects, such as unfavorable edge conditions. Plate 1 shows that cells can grow normally in any well, whereas plate 2 shows markedly lower cell counts at the edges, thus indicating the presence of experimental artifacts.

with simple mathematical operations, such that the values are approximately normally distributed and mean centered and have comparable s.d. Normal distributions make it easier to work with numeric values from a mathematical, statistical, and computational point of view. We highlight three key steps in this process:

Distribution testing. The need for transforming feature values can be evaluated for each feature on the basis of diagnostic measures and plots (Fig. 3). Graphical methods such as histograms, cumulative distribution curves, and quantile–quantile plots allow for visual identification of features that deviate from symmetric distributions. Analytical tests can also be used, including the Kolmogorov–Smirnov (KS) test and the Kullback–Leibler divergence, both of which aim to compute ratios of deviation from normality.

Logarithmic transformations. These transformations are often used to obtain approximate normal distributions for features that have highly skewed values or require range correction^{67,68}. Transformations include the generalized logarithmic function⁶⁸ and other adaptations that use shrinkage terms to avoid problems with nonpositive and near-zero feature values^{69,70}, as well as the Box–Cox transformation⁶⁷.

Relative normalization. This procedure consists of computing statistics (for example, median and median absolute deviation) in one population of samples, and then centering and scaling the rest with respect to that population. Ideally, features are normalized across an entire screen in which batch effects are absent; however, normalization within plates is generally performed to correct for batch effects (described in ‘Batch-effect correction’). When choosing the normalizing population, we suggest the use of control samples (assuming that they are present in sufficient quantity), because the presence of dramatic phenotypes may confound results. This procedure is good practice regardless of the normalization being performed within plates or across the screen. Alternately, all samples on a plate can be used as the normalizing population when negative controls are unavailable, too few, or unsuitable for some reason, and when samples on each plate are expected to not be enriched in dramatic phenotypes.

We recommend applying normalization across all features. Normalization can be applied even if features are not transformed, and it is preferable to remove biases while simultaneously fixing range issues. *z*-score normalization is the most commonly used procedure in our laboratories. Normalization also aligns the range of different features, thus decreasing the effects of unbalanced scales when computing similarities (described in ‘Measuring profile similarity’) or applying analysis algorithms (described in ‘Downstream analysis’). It is advisable to compare several transformation and normalization methods, because their performance can vary significantly among assays⁷¹.

Step 4: dimensionality reduction

At this point in the workflow, it can be useful to ask which of the measured features provide the most value in answering the biological question being studied.

Dimensionality reduction aims to filter less informative features and/or merge related features in the morphological profiles, given that morphological features calculated for profiling are often relatively redundant. The resulting compact representation is computationally more tractable, and it additionally avoids overrepresentation of similar features, that is, having a subgroup of features that measure similar or redundant properties of cells. Redundant features can diminish the signals of other more complementary features that are underrepresented, thus confounding downstream analysis.

Feature selection. Feature selection reduces dimensionality by discarding individual features while leaving the remainder in their original format (and thus retaining their interpretability). Options include:

Finding correlated features. One feature is selected from a subgroup that is known to be correlated. For instance, some texture features are highly correlated; thus, not all of them are needed, because they may represent the same underlying biological property. The feature–feature correlation matrix is computed, and pairs with a correlation exceeding a given threshold are identified iteratively. At each step, the feature with the largest mean absolute correlation with the rest of the features is removed.

Filtering on the basis of replicate correlation. Features that provide the highest additional information content^{69,70} on the basis of

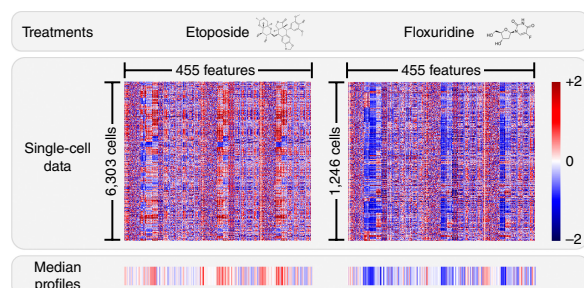


Figure 5 | Single-cell data aggregation. The feature matrices of two treatments show the measurements of their cell populations in the experiment. These measurements have been collapsed into median profiles that show very distinct signatures corresponding to two selected compounds: etoposide and floxuridine.

replicate correlation are iteratively selected as follows. An initial set of features is selected, and each of the remaining features is regressed on the selected set. The resulting residual data vector represents the additional information not already present in the selected features. The correlation of this residual vector across replicates is used to quantify information content. As a separate step, features with low replicate correlation are often excluded from analysis because they are too noisy^{69,72}.

Minimum redundancy–maximum relevance. A subset of features can have high replicate correlation without contributing substantially new information. To prevent selecting redundant features, minimum redundancy–maximum relevance⁷³ adds a constraint based on mutual information to the selection algorithm. The resulting selected features have high replicate correlation while preserving a diverse set of measurements⁷⁴.

Support-vector-machine-based recursive-feature elimination. A support vector machine is trained to implicitly weigh useful features in a classification task. Then, the features with the lowest weight are iteratively removed until the accuracy of the classification task begins to decline⁷⁵. In profiling applications, it may be desirable to select the features that best separate the treatments from the negative controls^{76,77}; the selected features would then be those that maximally differentiate phenotypes.

No previous studies have compared these options. Most groups use the filter method based on replicate correlation^{69,70,72}, and some add more powerful algorithms despite the computational cost. A combination of methods could be used, especially in tandem with the replicate-correlation strategy. There are other methodologies that may be useful, such as rescaling features in correlated groups such that their sum is one or selecting the features that contribute to most of the variance in the first two principal components.

Linear transformation. Methods of linear transformation seek lower-dimensional subspaces of higher-dimensional data that maintain information content. Linear transformation can be performed on single-cell profiles and aggregated sample-level profiles. Unlike feature selection, transformations can combine individual features, thus making the resulting features more powerful and information rich but potentially impeding their interpretability.

Linear transformation across all samples in the experiment is often needed for downstream analysis, to avoid overrepresentation of related features. Options used in morphological profiling are:

PCA. This procedure maximizes variance in successive orthogonal dimensions. PCA has been shown to outperform other dimensionality-reduction methods, such as random-forest selection for discriminating small-molecule-inhibitor effects⁷⁸, and independent component analysis and statistical image moments (Zernike/Fourier) for separating cell lines and preserving cell morphology after reconstruction from a lower-dimensional space⁷⁹.

Factor analysis and linear discriminant analysis. Factor analysis, which is closely related to PCA, finds nonorthogonal combinations of features representing frequent patterns in the data⁸⁰. Linear discriminant analysis finds a projection that maximizes the separation between positive and negative controls⁸¹. Both procedures have been successfully used in morphological profiling.

Among our laboratories, and in data science more generally, PCA is the most commonly used choice. Its simplicity and ability to retain a large amount of information in fewer dimensions probably explains its popularity. One comparative analysis using image-based profiling data has shown that factor analysis, compared with some alternate transformations, can identify a compact set of dimensions and improve downstream analysis results⁷⁷.

Step 5: single-cell data aggregation

Profiles are data representations that describe the morphological state of an individual cell or a population of cells. Population-level (also called image-level or well-level) representations are obtained by aggregating the measurements of single cells into a single vector to summarize the typical features of the population, so that populations can be compared (Fig. 5).

Simple aggregations. There are three simple and commonly used strategies for creating aggregated population-level profiles from all individual cell profiles in the sample:

Mean profile. Assuming a normal distribution of features, a profile built from the means of each feature for all cells in the population can provide a useful summary. This method has been used for compound classification^{77,82}. The profile length is sometimes doubled by also computing the s.d. of each feature.

Median profile. Taking the median for each feature over all the cells in a sample (and optionally the median absolute deviation) can be more robust to non-normal distributions and can mitigate the effects of outliers. If outliers are artifacts or errors, this procedure is useful, but the median may misrepresent populations with rare phenotypes by considering them as undesired outliers.

KS profile. This profile compares the probability distribution of a feature in a sample with respect to negative controls by using the KS nonparametric statistical test⁸³. The resulting profile is the collection of KS statistics for the features, which reveal how different the sample is with respect to the control.

There are other tests that may perform well but have not been evaluated for morphological profiling. Such tests include the

Anderson–Darling statistic and the Mann–Whitney *U* test. Other aggregation strategies can be designed by using bootstrap estimators previously used for phenotype classification⁸⁴.

The median profile has been found to have better performance than other profiling strategies in two different studies^{16,77} and is the preferred choice in most of our laboratories. One choice that varies among groups is whether to construct profiles at the level of images, fields of view, wells, or replicates. One could, for example, calculate a mean profile across all cells in a given replicate (regardless of the image or well) or instead calculate means for each image individually and then calculate means across images to create the replicate-level profile.

Subpopulation identification and aggregation. In most image-based cell-profiling workflows, it is implicitly assumed that ensemble averages of single-cell measurements reflect the dominant biological mechanism influenced by the treatment condition. However, subpopulations of cells are known to exhibit different phenotypes even within the same well^{85,86}. Classifying populations of single cells on the basis of their shape^{87–90}, cell-cycle phase^{13,88,91}, or signaling state⁹² can aid in interpretation and visualization of cell-profiling data⁹³. Cellular heterogeneity poses practical challenges for effective measurement methods that account for this variability.

Making use of subpopulations usually involves three key steps:

Subpopulation identification. Cells are clustered according to their morphological phenotypes, by using single-cell profiles (from controls or from the whole experiment). Clustering can be supervised, wherein reference phenotypes are selected^{94–96}, or unsupervised, as in *k*-means clustering^{90,97} and Gaussian mixture model fitting⁹².

Classification. Single-cell data points from all treatment conditions are then assigned to one of the subpopulations identified in the previous step. This assignment can be done by using a feature-evaluation rule, such as proximity, similarity, or feature weighting. This step is necessary because subpopulation identification is typically performed only on a subset of cells.

Aggregation. For each treatment condition, vectors are calculated and yield the number (or fraction) of cells within each subpopulation. Thus, the dimensionality of these vectors is the number of identified subpopulations.

An unproven hypothesis in the field is that profiles based on identification of phenotypically coherent subpopulations of cells should improve the accuracy of profiling, given the prevalence of heterogeneity and the existence of small subpopulations that might be ignored in mean or median profiling. In fact, to date, subpopulation-based profiling has not improved separation of treatment conditions^{77,98}. Nonetheless, defining subpopulations can assist in inferring biological meaning, by identifying over- and underrepresented subpopulations of cells under a given treatment condition⁹⁹ and by improving understanding of the dynamics of how cells transition between different phenotypes^{98,100}.

Step 6: measuring profile similarity

A key component of downstream analysis is the definition of a metric to compare treatments or experimental conditions. Similarity metrics reveal connections among morphological profiles.

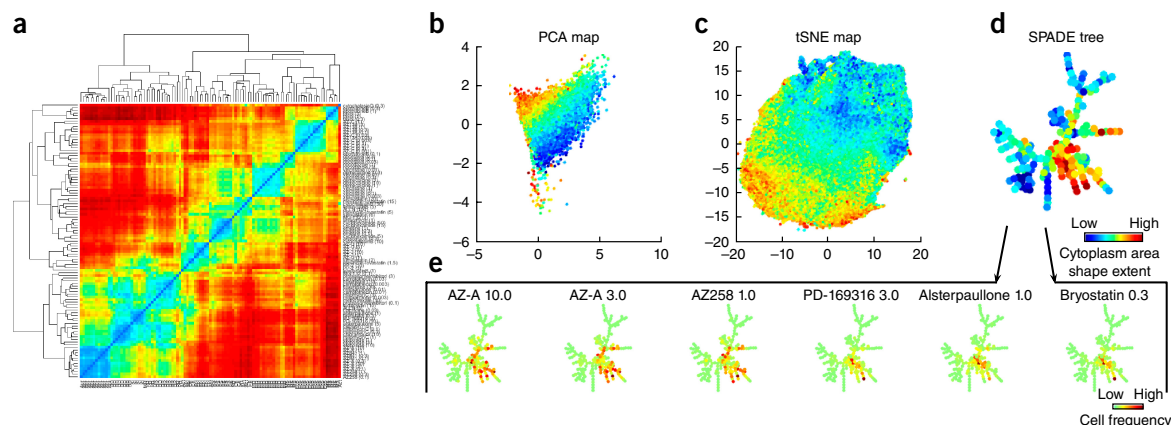


Figure 6 | Visualizations for downstream analysis. The source data are from 148,649 cells from the BBBC021 data set¹⁰⁸. **(a)** A heat map of the distance matrix shows the correlations between all pairs of samples, by using sample-level data (described in 'Measuring profile similarity'). **(b–d)** The cellular heterogeneity landscape can be visualized from single-cell data by using PCA **(b)**, tSNE scatter plots **(c)** or a SPADE tree **(d)**. In these examples, single-cell data points are colored according to a single-cell shape feature 'cytoplasm area shape extent' (red, high; blue, low). **(e)** A separate visualization for each treatment can assist in interpreting phenotypic changes induced by sample treatments. A constant SPADE tree is shown, and treatment-induced shifts in the number of cells in each 'node' of the tree are shown by the color scale depicted. The first three treatments are known to have a similar functional effect (Aurora kinase inhibition), and they exhibit similar cell distributions on the SPADE tree. The remaining three treatments are known to induce protein degradation, inducing cell distributions that differ from the first three.

Similarity-metric calculation. With a suitable metric, the similarities among a collection of treatment conditions can facilitate downstream analysis and allow for direct visualization of data structure, for example in distance heat maps (**Fig. 6a**). Image-based cell-profiling studies use three types of metrics:

Distance measures. These measures involve calculating how far apart two points are in the high-dimensional feature space. Those used in morphological profiling include Euclidean^{72,83}, Mahalanobis¹⁰¹, and Manhattan distances. Distance measures are very useful to quantify the difference in magnitude between profiles, because they aggregate the lengths of feature variations regardless of directionality. This procedure is useful to compute estimates of phenotypic strength of treatments with respect to controls.

Similarity measures. These measures involve computing a statistical estimate of the likelihood of a relation between two profiles. Statistics used in morphological profiling include Pearson's correlation¹⁰², Spearman's rank correlation¹⁰³, Kendall's rank correlation⁷⁸, and cosine similarity⁷⁷. Similarity measures quantify the proximity between profiles, because they detect deviations from one sample to another regardless of the absolute magnitude. This procedure is useful in finding relations and groups of samples that share common properties.

Learned similarity measures. These measures involve training machine-learning models that weight features differently according to prior knowledge about samples. The model can be a classifier that systematically identifies differences between two samples by using cross-validation¹⁰⁴ or by determining transformations of features that lead to maximal enrichment of groups of related samples⁸⁹. These strategies can highlight patterns that are not discriminated by regular metrics and that usually require more computational power to be calculated.

The performance of distance and similarity metrics relies on the quality of selected features (described in 'Feature selection'). High-dimensional feature profiles are often prone to the drawback of dimensionality, which consists of a decreasing ability of metrics to discern differences between vectors when the dimensionality increases. Dimensionality reduction can mitigate this effect (described in 'Linear transformations'). However, the choice of the metric can also be crucial, because good metrics better exploit the structure of the available features.

A comparison of metrics on one particular imaging data set has demonstrated that rank correlations (Spearman's and Kendall's) perform best for multiple untransformed feature vectors, whereas Euclidean and Manhattan distances are best for calculating z-prime factor values between positive and negative controls⁷⁸. A comparison of metrics in gene expression data sets has suggested that Pearson's correlation performs best when features are ratios, whereas Euclidean distance is best on other distributions¹⁰⁵.

The consensus from our laboratories is that selecting an optimal metric is probably specific to feature-space dimensionality and distributions that result from prior steps in the pipeline. For a typical pipeline, Pearson's correlation generally appears to be a good choice. Notably, indexes measuring clustering quality¹⁰⁶, for example the Davies–Bouldin Index, silhouette statistic, and receiver operating characteristic–area under the curve can aid in metric choice^{78,98}.

Concentration-effect handling. In experiments involving chemical perturbations, multiple concentrations are usually tested. Generally, researchers are interested in identifying phenotypic similarities among compounds even if those similarities occur at different doses. The following strategies are used to compute dose-independent similarity metrics:

Titration-invariant similarity score. First, the titration series of a compound is built by computing the similarity score between

each dose and negative controls. Then, the set of scores is sorted by increasing dose and is split into subseries by using a window of certain size (for instance, windows of three doses). Two compounds are compared by computing the correlation between their subwindows, and only the maximum value is retained⁸³.

Maximum correlation. For a set of n doses for each compound, the $N \times N$ correlation matrix is computed between all pairs of concentrations, and the maximum value is used as the dose-independent similarity score⁷².

The use of the maximum correlation is practical when a small number of concentrations are being tested. Depending on the experimental design, multiple concentrations can be treated differently. For instance, doses that do not yield a profile distinct from those of negative controls can be omitted, and the remaining doses can be combined to yield a single profile for the compound. Alternatively, if all concentrations are expected to have a phenotype, an entire compound can be left out of the analysis when its doses do not cluster together consistently¹⁰⁷. In addition, high doses can be removed if they are observed to be too toxic according to certain criteria, such as a minimum cell count^{102,107}.

Step 7: assay quality assessment

Assessing quality for morphological profiling assays can be challenging: basing the assessment on a few positive controls is not reassuring, but there are rarely a large number of controls available, nor are there other sources of ground truth. Every measured profile combines a mixture of the signal relating to the perturbation together with unintended effects such as batch effects and biological noise. Tuning the sample-preparation technique, choosing cell lines or incubation times, and choosing among alternatives within the computational pipeline all benefit from use of a quantitative indicator of whether the assay is better or worse as a result of particular design choices. Options include:

Comparison to ground truth. If the expected similarities between pairs of biological treatments are known, they can be used to validate predicted values. For instance, different concentrations of the same compound are expected to cluster together, and computed similarities should reflect that clustering. Similarly, if a subset of biological treatments is known to fall into particular classes, classification accuracy can be an appropriate metric⁷⁷. However, it is challenging to obtain ground-truth annotations at a large scale. To our knowledge, the only publicly available image data set with a large number of class annotations is for human MCF7 breast cancer cells (in this case, various classes of compound ‘mechanisms of action’)¹⁰⁸. Importantly, for proper evaluation of this data set, one complete compound set, including all concentrations, should be left out of training. A common mistake is to leave out a single dose of a single compound, inappropriately leaving the remaining doses of the same compound available to the classifier for training. Additional benchmarks beyond this data set are greatly needed.

Replicate reproducibility. This is typically measured as the similarity among the profiles of replicate pairs of the same biological treatment, which should be significantly higher than the similarity to profiles of other experimental conditions (controls and/or other biological treatments). This procedure requires at least two replicates of the experiment, a condition usually met for modern

morphological profiling experiments. To assess significance, similarity scores are compared with a suitable null distribution. A null distribution is usually built with pairs of samples that are not expected to be highly correlated, and it mainly depends on the hypothesis being tested. For instance, the use of all pairs of biological treatments can provide a diverse null distribution for measuring replicate correlation, and a null formed by random pairs of control samples can be compared against controls grouped by well location to reveal position effects. A P value can be computed nonparametrically by evaluating the probability of random pairs having greater similarity than a particular replicate pair.

Effect size. The difference between positive and negative controls, also known as the effect size, can be used as a measure of quality. This measure can be computed with a wide variety of statistical formulations, including univariate and multivariate methods, and also by assuming parametric and nonparametric models^{109,110}. The disadvantage of this approach is that maximizing effect size alone may cause a bias toward detecting only those phenotypes that distinguish the control while ignoring other phenotypes.

Exploratory approaches. Several methods have not been tested but might prove useful. Clustering can be used to ascertain the overall structure of relationships among samples in the experiment: a pipeline that produces substructures or many distinct clusters is likely to be preferable over one in which the distances between all pairs of samples are similar. The cumulative variance of the principal components is a metric not yet applied to morphological profiling experiments. Highly diverse signals from different biological treatments should require more components to explain a predefined fraction of variance (for example, 99%).

Currently, replicate reproducibility is the most commonly used method, given that ground truth is rarely available. Specifically, methods are often optimized to maximize the percentage of replicates that are reproducible relative to a null (under suitable cross validation). Using a null comprising pairwise correlations between different treatments is safer than using a null comprising correlations between treatments and negative controls; in the latter case, it is possible to optimize the assay to distinguish samples from negative controls while diminishing important differences among samples.

Step 8: downstream analysis

Downstream analysis is the process of interpreting and validating patterns in the morphological profiles. The most important readouts are the similarities and relationships among the experimental conditions tested. Visualization of the relationships and the use of machine learning can help to uncover biologically meaningful structures and connections among various treated samples. Most laboratories use a combination of these strategies; generally, unsupervised clustering is a good starting point for exploring the data. From there, the goals of the study strongly influence the combination of approaches used.

Clustering. Finding clusters is one of the most effective ways of extracting meaningful relationships from morphological profiles. Clustering algorithms can be used for identifying new associations

among treatments as well as validating known connections and ruling out batch effects. There are several ways of clustering a data set. Hierarchical clustering, the most widely adopted strategy, is used to identify groups of highly correlated experimental conditions⁸⁷ and to identify treatments with unexpected positive or negative connections⁹⁹. Although it is not discussed in detail here, examining relationships among features rather than among samples can yield useful biological insights: for example, the amount of mitochondrial material in cells is generally proportional to cell size, thus revealing stereotyped control of these parameters, but certain chemical perturbants can disrupt this relationship¹¹¹.

Hierarchical clustering is computed by using a similarity matrix that contains the similarity values for all pairs of samples (described in ‘Measuring profile similarity’). This similarity matrix can be visualized as a heat map to reveal patterns in the data for several or up to hundreds of samples. The heat maps’ rows and columns are typically sorted by using the hierarchical structure discovered by the clustering algorithm. This hierarchical structure is known as a dendrogram, which links samples together according to their proximity in the feature space, and is usually visualized together with the heat map to highlight negative and positive correlations in the data (Fig. 6a). Bootstrapping has been used to evaluate the statistical significance of the results obtained with hierarchical clustering, as well as other probabilistic algorithms used in the analysis of single-cell populations³². Resampling methods can generally be used to estimate variance, error bars, or other statistical properties of the data and can aid in making more accurate predictions and interpretations.

Visualization of high-dimensional data. Visualizations are useful to reveal the distribution and grouping of high-dimensional data points by using a 2D (and sometimes 3D) map layout that approximates their positions in the feature space. The relationships among points are implicitly encoded in how close together or far apart they are in the visualization. This method can be used to study cell heterogeneity by using single-cell data points, or sample relations by using aggregated profiles. Single-cell data are usually downsampled for practical reasons: to decrease data size and identify rare cell types^{112,113}. The following are the most common approaches for data visualization:

Data projections. A projection of the data matrix is displayed in a 2D (or 3D) scatter plot that approximates the geometry of the original point cloud. The most common methods include PCA (Fig. 6b), Isomap¹¹⁴, *t*-distributed stochastic neighbor embedding (tSNE)¹¹⁵ (Fig. 6c), and viSNE¹¹⁶.

Hierarchical visualizations. Plots are used to find structures in the data and reveal relationships between samples (Fig. 6d,e). The most commonly used choices are spanning-tree progression analysis of density-normalized events (SPADE)^{113,117} and minimum spanning trees¹¹⁸, which allow for relationships among hierarchical groups of single cells or samples to be identified by using branches that may represent phenotypes or treatments.

In many cases, data points in a visualization are colored on the basis of positive controls or otherwise known labels in the data, a common practice in analysis of single-cell flow cytometry data^{116,119,120}. The color code can also illustrate other information

in the data set, such as cell phenotypes, compound doses, values of measured features, or treatment conditions (Fig. 6e). Visualizations can be more effective if they are interactive, thereby allowing researchers to create and test hypotheses *ad hoc*. Software packages such as Shiny, GGobi, iPlots in R, Bokeh in Python, and D3.js in Javascript provide interactive plotting capacities, most of which can also be deployed in server-client environments for dissemination to the public.

Classification. Classification rules can be useful for transferring labels from annotated samples to unknown data points, for example, classifying the mechanism of action of new compounds in a chemical library. As such, classification strategies require prior knowledge in the form of annotations for at least some of the data points in the collection. Given examples of data points that belong to different classes of interest, supervised classification algorithms learn a rule that computes the probability of each unknown data point falling into one of the classes.

It is relatively uncommon to have a large number of annotated samples in morphological profiling, because most experiments are designed to be exploratory. However, when this information is available, a classification strategy can provide informative insights into the treatments. The most commonly used classification rule in morphological profiling experiments is the nearest-neighbors algorithm, which finds the closest data points in the collection of annotated samples and recommends a label for the new sample. For instance, this algorithm has been used for classifying the mechanism of action in a compound library⁷⁷. Other supervised prediction models can also be used to learn relations between morphological features and biological activity assays, such as Bayesian matrix factorization, neural networks, and random forests¹²¹.

The classification performance is validated in a holdout test using precision, recall, and accuracy measures. It is absolutely critical for confidence in these metrics that the holdout test set not overlap with any data points in the training set. The most recommended practice is to use samples treated in a different experimental batch to create the holdout test set (other ground-truth recommendations are described in ‘Assay quality assessment’).

Sharing

Both authors and the scientific community benefit from sharing code and data¹²². Numerous tools currently exist that address the steps outlined in this paper (Box 1); these tools can be useful both for beginners to experiment with and learn from and for experts to integrate into pipelines and build upon. Although data must be kept confidential for sensitive patient material, intellectual-property concerns are generally not the major issue with sharing; the primary hurdle in the process is usually the often substantial time and effort required of the authors. We do not consider code or data labeled ‘available upon request’ to qualify as being openly shared, given the poor efficacy statistics^{123,124}. We therefore recommend the following options to make code and data available publicly online.

Code sharing. Options for sharing code include:

Step-by-step narrative. For software with only a graphical user interface, a detailed walkthrough of each step of the workflow can be provided; however, this option is suboptimal.

BOX 1 SOFTWARE TOOLS

A large range of software tools and libraries currently exist that seek to address the steps outlined in this paper. For each step, the alternatives are usually several software packages or programming languages that require either parameterization or coding.

Tools for image-analysis software have been previously reviewed¹⁵⁰, and the variety in functionalities and platforms can fit a diverse range of workflows. Some of the open-source alternatives include CellProfiler²⁴ and EBIImage³⁵, whereas Columbus and MetaXpress are commercial solutions.

After collection of features or measurements with image-analysis software, the next steps in the workflow may require a combination of tools and programming languages. Statistical

packages such as R have proven to be very useful for single-cell data analysis, including *cytominer*, which is specific to morphological profiling. Other programming languages such as Python, Matlab and shell scripts can be used to process data with specific algorithms, including machine learning, data transformation, or simple data filtering and selection.

Each step may require specialized methods or may be solved with off-the-shelf implementations. The field is constantly changing, and the next breakthroughs in theory and practice may require new tools not yet available. In either case, the practice of sharing code is highly valued, to ensure rapid implementation of techniques, optimization of pipelines, and reproducibility of the results by others.

Online code repository. The code should preferably be publicly hosted rather than being provided on a university website or as journal supplemental files. The options range from repositories such as Github and BitBucket to tools such as Jupyter notebooks and knitr documents¹²⁵, which allow for reproducible reports containing code, documentation, and figures to be shared within a single document.

Packaging. Researchers can capture and share the computational environment used to create the results, such as providing virtual machines or Docker containers. Doing so ensures that all code, dependencies and data are available in a single container^{126,127}, which is convenient for the user and also protects against changes in software libraries and dependencies.

Data sharing. In image-based cell profiling, publicly available data are valuable not only for reproducing results but also for identifying completely new biological findings. Options include:

Sharing processed data only. Sharing only processed data (for example, extracted features) has been common, often through supplemental data files available via the journal or via a general-purpose data repository such as Dryad (<http://datadryad.org/>).

Sharing images and data online. Few raw-image sets have been made available online, primarily because of the large size of the image data (tens of gigabytes for each 384-well plate) and therefore the high cost of maintaining the data on public servers. However, recent initiatives are decreasing this cost for authors, including the Image Data Resource (IDR; <https://idr-demo.openmicroscopy.org/>)¹²⁸, which accepts cellular images at the scale of high-throughput image profiling experiments. Generally, smaller sets of annotated images for testing image analysis methods are available in the Broad Bioimage Benchmark Collection (<https://data.broadinstitute.org/bbbc/>)¹⁰⁸ and the Cell Image Library (<http://www.cellimagelibrary.org/>). Some resources, such as IDR, support using an ontology for describing phenotypes¹²⁹. Before these public resources became available, some laboratories provided the data through their institutional servers^{13,32,52,89,103,130,131}. Tools such as OMERO¹³² and open-BIS¹³³ have been used to create project-specific portals for easy

online data exploration^{32,52,130}, but bulk download of very large data sets can remain challenging.

We strongly encourage sharing of both data and images online, given how rapidly feature-extraction methods are changing, particularly via deep-learning methods (described in 'Alternate workflows').

Alternate workflows

The data-processing workflow and recommendations presented in this paper have evolved as a result of years of efforts in different laboratories. They have been robustly used in various studies and have proven to be successful in making biological discoveries^{8,9}. However, the field is eager to adapt as the computer-vision and machine-learning communities make progress in designing new algorithms for processing image data. Some of our laboratories are already exploring alternate workflows, such as those described below.

Segmentation-free classical-feature extraction. Instead of identifying single cells that are measured and characterized, this strategy computes classical features from whole field-of-view images or from discrete tiles within images. Examples of these include PhenoRipper^{134,135} and WND-Charm/CP-CHARM^{136–138}.

Deep-learning feature extraction. Deep learning techniques have recently and dramatically come to dominate the state-of-the-art performance in various computer vision tasks¹³⁹. The most relevant model for image analysis is currently the convolutional neural network (CNN), which learns to extract useful features directly from raw pixel data by using multiple nonlinear transformations, in contrast to the classical features described in 'Feature extraction'. This model has been used for segmentation and classification of biomedical images^{140,141}, for phenotype discovery in single-cell images from imaging flow cytometry¹⁴², and more recently for deep-learning approaches for morphological profiling: morphological profiling^{143,144}. The following are the most relevant deep-learning approaches for morphological profiling:

Learning features from raw pixels. This approach has been used for problems in which phenotypes of interest are predefined, and a set of categorized examples is needed to train the network. This approach has been successfully used for protein-localization

problems^{145–147} and mechanism-of-action prediction¹⁴⁴. Input images can be single cells^{146,147} or full fields of view^{144,145}.

Transferring learned features from other domains. Using a CNN trained on a large data set for other tasks different from the original is known as transfer learning. CNNs pretrained with natural images have been evaluated as feature extractors for full image profiling of compounds; its accuracy matches the results of classical features without requiring segmentation or training¹⁴³. The preprocessing steps described in 'Field-of-view quality control' and 'Field-of-view illumination correction' are still likely to be necessary for obtaining improved results. If there are few annotations available for phenotype-classification tasks, transfer learning can also be used to improve performance¹⁴⁶.

Learning transformations of classical features. feature transformations similar to those described in 'Linear transformations' can be obtained with a technique known as the autoencoder. Deep autoencoders have been evaluated for high-content morphology data, thus suggesting that they may potentially have better performance for downstream analysis according to homogeneity of clusters¹⁴⁸. Another study has evaluated deep autoencoders for profiling and has also obtained competitive performance¹⁴⁹.

Using full images results in a loss of single-cell resolution but offers several advantages: the avoidance of the segmentation step eliminates the sometimes tedious manual tuning of segmentation and feature extraction algorithms, saves computation time, avoids segmentation errors, and may better capture visual patterns resulting from multiple cells. Using single-cell images explicitly captures heterogeneity and may offer improved accuracy with less training.

Although segmentation-free classical-feature extraction can be helpful for quality control, we generally consider it to be incapable of accomplishing most profiling tasks. Deep-learning techniques, although not yet proven to be more powerful than the standard workflow, are nonetheless very promising. We are actively pursuing optimized workflows based on deep learning and are gaining an understanding of how these techniques can be adapted for improving the computation and interpretation of useful image features.

We caution that it is possible to obtain excellent results on a ground-truth data set with a method that fails in realistic-use cases. This phenomenon may be especially true for machine-learning-based methods with millions of internal parameters and again reinforces the need for new and disparate sets of ground-truth data in the field.

Conclusions

It is an exciting time for the field of image-based cell profiling, as methods are rapidly evolving and applications leading to major biological discoveries are beginning to be published. We see the collection and sharing of large biologically interesting image sets, the organizing of benchmark ground-truth data sets, and the testing of new methods to be the major areas in which effort is currently most needed.

In future work, as a community, we aim to build shared codebases, namely toolboxes of algorithms in R and Python. The beginnings of this effort can be found online (<https://github.com/>

[CellProfiler/cytominer/](#)), and we welcome additional contributors as well as participants in the cytominer hackathon, which will be held annually. A shared codebase will facilitate the development and dissemination of novel methods and the comparison of alternative methods, particularly as additional ground-truth data become publicly available.

Data availability

This work did not analyze new data. The plots and figures presented in the manuscript were obtained by processing the BBBC021 image collection, which is publicly available in <https://data.broadinstitute.org/bbbc/BBBC021/>.

ACKNOWLEDGMENTS

The cytominer hackathon 2016 was supported in part by a grant from the National Institutes of Health BD2K program (U54 GM114833). Work on this paper was supported in part by NSF CAREER DBI 1148823 (to A.E.C.).

AUTHOR CONTRIBUTIONS

All authors contributed to writing the manuscript and editing the text.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

1. Boutros, M., Heigwer, F. & Lauffer, C. Microscopy-based high-content screening. *Cell* **163**, 1314–1325 (2015).
2. Mattiazzi Usaj, M. *et al.* High-content screening for quantitative cell biology. *Trends Cell Biol.* **26**, 598–611 (2016).
3. Fetz, V., Prochnow, H., Brönstrup, M. & Sasse, F. Target identification by image analysis. *Nat. Prod. Rep.* **33**, 655–667 (2016).
4. Pennisi, E. 'Cell painting' highlights responses to drugs and toxins. *Science* **352**, 877–878 (2016).
5. Grys, B.T. *et al.* Machine learning and computer vision approaches for phenotypic profiling. *J. Cell Biol.* **216**, 65–71 (2017).
6. Feng, Y., Mitchison, T.J., Bender, A., Young, D.W. & Tallarico, J.A. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discov.* **8**, 567–578 (2009).
7. Mader, C.C., Subramanian, A. & Bittker, J. Multidimensional profile based screening: understanding biology through cellular response signatures. in *High Throughput Screening Methods: Evolution and Refinement* (eds. Bittker, J.A. & Ross, N.T.) 214–238 (RSC Publishing, 2016).
8. Caicedo, J.C., Singh, S. & Carpenter, A.E. Applications in image-based profiling of perturbations. *Curr. Opin. Biotechnol.* **39**, 134–142 (2016).
9. Bougen-Zhukov, N., Loh, S.Y., Lee, H.K. & Loo, L.-H. Large-scale image-based screening and profiling of cellular phenotypes. *Cytometry A* **91**, 115–125 (2017).
10. Gustafsdottir, S.M. *et al.* Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* **8**, e80999 (2013).
11. Bray, M.-A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
12. Kang, J. *et al.* Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. *Nat. Biotechnol.* **34**, 70–77 (2016).

13. Neumann, B. *et al.* Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).
14. Hasson, S.A. & Ingles, J. Innovation in academic chemical screening: filling the gaps in chemical biology. *Curr. Opin. Chem. Biol.* **17**, 329–338 (2013).
15. Smith, K. *et al.* CIDRE: an illumination-correction method for optical microscopy. *Nat. Methods* **12**, 404–406 (2015).
16. Singh, S., Bray, M.-A., Jones, T.R. & Carpenter, A.E. Pipeline for illumination correction of images for high-throughput microscopy. *J. Microsc.* **256**, 231–236 (2014).
17. Likar, B., Maintz, J.B., Viergever, M.A. & Pernus, F. Retrospective shading correction based on entropy minimization. *J. Microsc.* **197**, 285–295 (2000).
18. Lévesque, M.P. & Lelièvre, M. Evaluation of the iterative method for image background removal in astronomical images. (TN 2007-344) (DRDC Valcartier, 2008).
19. Babaloukas, G., Tentolouris, N., Liatis, S., Sklavounou, A. & Perrea, D. Evaluation of three methods for retrospective correction of vignetting on medical microscopy images utilizing two open source software tools. *J. Microsc.* **244**, 320–324 (2011).
20. Can, A. *et al.* Multi-modal imaging of histological tissue sections. in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 288–291 (2008).
21. Molnar, C. *et al.* Accurate morphology preserving segmentation of overlapping cells based on active contours. *Sci. Rep.* **6**, 32412 (2016).
22. Stoeger, T., Battich, N., Herrmann, M.D., Yakimovich, Y. & Pelkmans, L. Computer vision for image-based transcriptomics. *Methods* **85**, 44–53 (2015).
23. Sommer, C., Straehle, C., Köthe, U. & Hamprecht, F.A. Ilastik: interactive learning and segmentation toolkit. in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 230–233 (2011).
24. Carpenter, A.E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
25. Rodenacker, K. & Bengtsson, E. A feature set for cytometry on digitized microscopic images. *Anal. Cell. Pathol.* **25**, 1–36 (2003).
26. Wählby, C. *Algorithms for applied digital image cytometry* PhD thesis. Uppsala University (2003).
27. Haralick, R.M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).
28. Turner, M.R. Texture discrimination by Gabor functions. *Biol. Cybern.* **55**, 71–82 (1986).
29. Bolland, M.V., Markey, M.K. & Murphy, R.F. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* **33**, 366–375 (1998).
30. Coelho, L.P. *et al.* Determining the subcellular location of new proteins from microscope images using local features. *Bioinformatics* **29**, 2343–2349 (2013).
31. Snijder, B. *et al.* Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461**, 520–523 (2009).
32. Snijder, B. *et al.* Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol. Syst. Biol.* **8**, 579 (2012).
33. Sero, J.E. *et al.* Cell shape and the microenvironment regulate nuclear translocation of NF- κ B in breast epithelial and tumor cells. *Mol. Syst. Biol.* **11**, 790 (2015).
34. Singh, S., Carpenter, A.E. & Genovesio, A. Increasing the content of high-content screening: an overview. *J. Biomol. Screen.* **19**, 640–650 (2014).
35. Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBIImage: an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).
36. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
37. Groen, F.C., Young, I.T. & Ligthart, G. A comparison of different focus functions for use in autofocus algorithms. *Cytometry* **6**, 81–91 (1985).
38. Haralick, R.M. Statistical and structural approaches to texture. *Proc. IEEE* **67**, 786–804 (1979).
39. Field, D.J. & Brady, N. Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. *Vision Res.* **37**, 3367–3383 (1997).
40. Bray, M.-A., Fraser, A.N., Hasaka, T.P. & Carpenter, A.E. Workflow and metrics for image quality control in large-scale high-content screens. *J. Biomol. Screen.* **17**, 266–274 (2012).
41. Goode, A. *et al.* Distributed online anomaly detection in high-content screening. in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 249–252 (2008).
42. Lou, X., Fiaschi, L., Koethe, U. & Hamprecht, F.A. Quality classification of microscopic imagery with weakly supervised learning. in *Machine Learning in Medical Imaging* (eds. Wang, F., Shen, D., Yan, P. & Suzuki, K.) 176–183 (Springer Berlin Heidelberg, 2012).
43. Barnett, V. & Lewis, T. *Outliers in statistical data* (Wiley, 1994).
44. Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J. & Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167–175 (2006).
45. Liberali, P., Snijder, B. & Pelkmans, L. Single-cell and multivariate approaches in genetic perturbation screens. *Nat. Rev. Genet.* **16**, 18–32 (2015).
46. Prastawa, M., Bullitt, E., Ho, S. & Gerig, G. A brain tumor segmentation framework based on outlier detection. *Med. Image Anal.* **8**, 275–283 (2004).
47. Hulsman, M. *et al.* Analysis of high-throughput screening reveals the effect of surface topographies on cellular morphology. *Acta Biomater.* **15**, 29–38 (2015).
48. Rousseeuw, P.J. & Leroy, A.M. *Robust Regression and Outlier Detection* (Wiley, 2005).
49. Rämö, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25**, 3028–3030 (2009).
50. Horvath, P., Wild, T., Kutay, U. & Csucs, G. Machine learning improves the precision and robustness of high-content screens: using nonlinear multiparametric methods to analyze screening results. *J. Biomol. Screen.* **16**, 1059–1067 (2011).
51. Dao, D. *et al.* CellProfiler Analyst: interactive data exploration, analysis and classification of large biological image sets. *Bioinformatics* **32**, 3210–3212 (2016).
52. Liberali, P., Snijder, B. & Pelkmans, L. A hierarchical map of regulatory genetic interactions in membrane trafficking. *Cell* **157**, 1473–1487 (2014).
53. Zhu, Y., Hernandez, L.M., Mueller, P., Dong, Y. & Forman, M.R. Data acquisition and preprocessing in studies on humans: what is not taught in statistics classes? *Am. Stat.* **67**, 235–241 (2013).
54. Mpindi, J.-P. *et al.* Impact of normalization methods on high-throughput screening data with high hit rates and drug testing with dose-response data. *Bioinformatics* **31**, 3815–3821 (2015).
55. Kluger, Y., Yu, H., Qian, J. & Gerstein, M. Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics* **4**, 49 (2003).
56. Yu, H. *et al.* Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool. *Nucleic Acids Res.* **35**, e8 (2007).
57. Makarenkov, V. *et al.* An efficient method for the detection and elimination of systematic error in high-throughput screening. *Bioinformatics* **23**, 1648–1657 (2007).
58. Homouz, D., Chen, G. & Kudlicki, A.S. Correcting positional correlations in Affymetrix genome chips. *Sci. Rep.* **5**, 9078 (2015).
59. Lundholt, B.K., Scudder, K.M. & Pagliaro, L. A simple technique for reducing edge effect in cell-based assays. *J. Biomol. Screen.* **8**, 566–570 (2003).
60. Brideau, C., Gunter, B., Pikounis, B. & Liaw, A. Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **8**, 634–647 (2003).
61. Reisen, F. *et al.* Linking phenotypes and modes of action through high-content screen fingerprints. *Assay Drug Dev. Technol.* **13**, 415–427 (2015).
62. Leek, J.T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
63. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
64. Vaisipour, S. Detecting, correcting, and preventing the batch effects in multi-site data, with a focus on gene expression microarrays. PhD thesis University of Alberta (2014).
65. Stein, C.K. *et al.* Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. *BMC Bioinformatics* **16**, 63 (2015).

66. Haney, S.A. Rapid assessment and visualization of normality in high-content and other cell-level data and its impact on the interpretation of experimental results. *J. Biomol. Screen.* **19**, 672–684 (2014).
67. Durbin, B.P., Hardin, J.S., Hawkins, D.M. & Rocke, D.M. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* **18** (Suppl. 1), S105–S110 (2002).
68. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** (Suppl. 1), S96–S104 (2002).
69. Laufer, C., Fischer, B., Billmann, M., Huber, W. & Boutros, M. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat. Methods* **10**, 427–431 (2013).
70. Fischer, B. *et al.* A map of directional genetic interactions in a metazoan cell. *eLife* **4**, e05464 (2015).
71. Birmingham, A. *et al.* Statistical methods for analysis of high-throughput RNA interference screens. *Nat. Methods* **6**, 569–575 (2009).
72. Woehrmann, M.H. *et al.* Large-scale cytological profiling for functional analysis of bioactive compounds. *Mol. Biosyst.* **9**, 2604–2617 (2013).
73. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **3**, 185–205 (2005).
74. Ng, A.Y.J. *et al.* A cell profiling framework for modeling drug responses from HCS imaging. *J. Biomol. Screen.* **15**, 858–868 (2010).
75. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
76. Loo, L.-H., Wu, L.F. & Altschuler, S.J. Image-based multivariate profiling of drug responses from single cells. *Nat. Methods* **4**, 445–453 (2007).
77. Ljosa, V. *et al.* Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* **18**, 1321–1329 (2013).
78. Reisen, F., Zhang, X., Gabriel, D. & Selzer, P. Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery. *J. Biomol. Screen.* **18**, 1284–1297 (2013).
79. Pincus, Z. & Theriot, J.A. Comparison of quantitative methods for cell-shape analysis. *J. Microsc.* **227**, 140–156 (2007).
80. Young, D.W. *et al.* Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **4**, 59–68 (2008).
81. Kümmel, A. *et al.* Integration of multiple readouts into the Z' factor for assay quality assessment. *J. Biomol. Screen.* **15**, 95–101 (2010).
82. Adams, C.L. *et al.* Compound classification using image-based cellular phenotypes. *Methods Enzymol.* **414**, 440–468 (2006).
83. Perlman, Z.E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
84. Candia, J. *et al.* From cellular characteristics to disease diagnosis: uncovering phenotypes with supercells. *PLoS Comput. Biol.* **9**, e1003215 (2013).
85. Altschuler, S.J. & Wu, L.F. Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559–563 (2010).
86. Snijder, B. & Pelkmans, L. Origins of regulated cell-to-cell variability. *Nat. Rev. Mol. Cell Biol.* **12**, 119–125 (2011).
87. Bakal, C., Aach, J., Church, G. & Perrimon, N. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* **316**, 1753–1756 (2007).
88. Jones, T.R. *et al.* CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics* **9**, 482 (2008).
89. Fuchs, F. *et al.* Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol. Syst. Biol.* **6**, 370 (2010).
90. Sailem, H., Bousgouni, V., Cooper, S. & Bakal, C. Cross-talk between Rho and Rac GTPases drives deterministic exploration of cellular shape space and morphological heterogeneity. *Open Biol.* **4**, 130132 (2014).
91. Mukherji, M. *et al.* Genome-wide functional analysis of human cell-cycle regulators. *Proc. Natl. Acad. Sci. USA* **103**, 14819–14824 (2006).
92. Singh, D.K. *et al.* Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol. Syst. Biol.* **6**, 369 (2010).
93. Sailem, H.Z., Cooper, S. & Bakal, C. Visualizing quantitative microscopy data: History and challenges. *Crit. Rev. Biochem. Mol. Biol.* **51**, 96–101 (2016).
94. Kiger, A.A. *et al.* A functional genomic analysis of cell morphology using RNA interference. *J. Biol.* **2**, 27 (2003).
95. Yin, Z. *et al.* Online phenotype discovery in high-content RNAi screens using gap statistics. in *Proc. Int. Symposium on Computational Models of Life Sciences* Vol. 952 (eds. Pham, T.D. & Zhou, X.), 86–95 (AIP Publishing, 2007).
96. Jones, T.R. *et al.* Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proc. Natl. Acad. Sci. USA* **106**, 1826–1831 (2009).
97. Volz, H.C. *et al.* Single-cell phenotyping of human induced pluripotent stem cells by high-throughput imaging. Preprint at <http://www.biorxiv.org/content/early/2015/09/16/026955/> (2015).
98. Cooper, S., Sadok, A., Bousgouni, V. & Bakal, C. Apolar and polar transitions drive the conversion between amoeboid and mesenchymal shapes in melanoma cells. *Mol. Biol. Cell* **26**, 4163–4170 (2015).
99. Rohban, M.H. *et al.* Systematic morphological profiling of human gene and allele function via Cell Painting. *eLife* **6**, e24060 (2017).
100. Gordonov, S. *et al.* Time series modeling of live-cell shape dynamics for image-based phenotypic profiling. *Integr. Biol.* **8**, 73–90 (2016).
101. Caie, P.D. *et al.* High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol. Cancer Ther.* **9**, 1913–1926 (2010).
102. Schulze, C.J. *et al.* “Function-first” lead discovery: mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **20**, 285–295 (2013).
103. Singh, S. *et al.* Morphological profiles of RNAi-induced gene knockdown are highly reproducible but dominated by seed effects. *PLoS One* **10**, e0131370 (2015).
104. Zhang, X. & Boutros, M. A novel phenotypic dissimilarity method for image-based high-throughput screens. *BMC Bioinformatics* **14**, 336 (2013).
105. Gibbons, F.D. & Roth, F.P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* **12**, 1574–1581 (2002).
106. Rendón, E., Abundez, I. & Arizmendi, A. Internal versus external cluster validation indexes. *Int. J. Computers Communications* **5**, 27–34 (2011).
107. Vial, M.-L. *et al.* A grand challenge. 2. Phenotypic profiling of a natural product library on Parkinson’s patient-derived cells. *J. Nat. Prod.* **79**, 1982–1989 (2016).
108. Ljosa, V., Sokolnicki, K.L. & Carpenter, A.E. Annotated high-throughput microscopy image sets for validation. *Nat. Methods* **9**, 637 (2012).
109. Hutz, J.E. *et al.* The multidimensional perturbation value. *J. Biomol. Screen.* **18**, 367–377 (2013).
110. Rajwa, B. Effect-size measures as descriptors of assay quality in high-content screening: a brief review of some available methodologies. *Assay Drug Dev. Technol.* **15**, 15–29 (2017).
111. Kitami, T. *et al.* A chemical screen probing the relationship between mitochondrial content and cell size. *PLoS One* **7**, e33755 (2012).
112. Zare, H., Shoostari, P., Gupta, A. & Brinkman, R.R. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* **11**, 403 (2010).
113. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
114. Tenenbaum, J.B., de Silva, V. & Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
115. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
116. Amir, A.D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
117. Anchang, B. *et al.* Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat. Protoc.* **11**, 1264–1279 (2016).
118. Qiu, P., Gentles, A.J. & Plevritis, S.K. Discovering biological progression underlying microarray samples. *PLoS Comput. Biol.* **7**, e1001123 (2011).
119. Bendall, S.C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
120. Haghverdi, L., Büttner, F. & Theis, F.J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
121. Simm, J. *et al.* Repurposed high-throughput images enable biological activity prediction for drug discovery. Preprint at <http://www.biorxiv.org/content/early/2017/03/30/108399/> (2017).

122. Carpenter, A.E., Kametsky, L. & Eliceiri, K.W. A call for bioimaging software usability. *Nat. Methods* **9**, 666–670 (2012).
123. Ince, D.C., Hatton, L. & Graham-Cumming, J. The case for open computer programs. *Nature* **482**, 485–488 (2012).
124. Collberg, C., Proebsting, T. & Warren, A.M. *Repeatability and Benefaction in Computer Systems Research* (Technical Report 14-04) (University of Arizona, 2015).
125. Shen, H. Interactive notebooks: sharing the code. *Nature* **515**, 151–152 (2014).
126. Boettiger, C. An introduction to Docker for reproducible research. *Oper. Syst. Rev.* **49**, 71–79 (2015).
127. Beaulieu-Jones, B.K. & Greene, C.S. Reproducibility of computational workflows is automated using continuous analysis. *Nat. Biotechnol.* **35**, 342–346 (2017).
128. Williams, E. *et al.* Image Data Resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).
129. Jupp, S. *et al.* The cellular microscopy phenotype ontology. *J. Biomed. Semantics* **7**, 28 (2016).
130. Breinig, M., Klein, F.A., Huber, W. & Boutros, M. A chemical-genetic interaction map of small molecules using high-throughput imaging in cancer cells. *Mol. Syst. Biol.* **11**, 846 (2015).
131. Badertscher, L. *et al.* Genome-wide RNAi Screening identifies protein modules required for 40S subunit synthesis in human cells. *Cell Rep.* **13**, 2879–2891 (2015).
132. Allan, C. *et al.* OMERO: flexible, model-driven data management for experimental biology. *Nat. Methods* **9**, 245–253 (2012).
133. Bauch, A. *et al.* openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics* **12**, 468 (2011).
134. Rajaram, S., Pavie, B., Wu, L.F. & Altschuler, S.J. PhenoRipper: software for rapidly profiling microscopy images. *Nat. Methods* **9**, 635–637 (2012).
135. Pavie, B. *et al.* Rapid analysis and exploration of fluorescence microscopy images. *J. Vis. Exp.* **e51280** (2014).
136. Shamir, L. *et al.* Wndchrm: an open source utility for biological image analysis. *Source Code Biol. Med.* **3**, 13 (2008).
137. Orlov, N. *et al.* WND-CHARM: multi-purpose image classification using compound image transforms. *Pattern Recognit. Lett.* **29**, 1684–1693 (2008).
138. Uhlmann, V., Singh, S. & Carpenter, A.E. CP-CHARM: segmentation-free image classification made accessible. *BMC Bioinformatics* **17**, 51 (2016).
139. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
140. Kraus, O.Z. & Frey, B.J. Computer vision for high content screening. *Crit. Rev. Biochem. Mol. Biol.* **51**, 102–109 (2016).
141. Van Valen, D.A. *et al.* Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS Comput. Biol.* **12**, e1005177 (2016).
142. Eulenberg, P., Koehler, N., Blasi, T., Filby, A. & Carpenter, A.E. Deep learning for imaging flow cytometry: cell cycle analysis of Jurkat cells. Preprint at <http://www.biorxiv.org/content/early/2016/10/17/081364/> (2016).
143. Pawlowski, N., Caicedo, J.C., Singh, S., Carpenter, A.E. & Storkey, A. Automating morphological profiling with generic deep convolutional networks. Preprint at <http://www.biorxiv.org/content/early/2016/11/02/085118/> (2016).
144. Godínez, W.J., Hossain, I., Lazic, S.E., Davies, J.W. & Zhang, X. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics* (2017).
145. Kraus, O.Z., Ba, J.L. & Frey, B.J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**, i52–i59 (2016).
146. Kraus, O.Z. *et al.* Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.* **13**, 924 (2017).
147. Pärnamaa, T. & Parts, L. Accurate classification of protein subcellular localization from high throughput microscopy images using deep learning. *G3 (Bethesda)* **7**, 1385–1392 (2017).
148. Zamparo, L. & Zhang, Z. Deep autoencoders for dimensionality reduction of high-content screening data. Preprint at <https://arxiv.org/abs/1501.01348/> (2015).
149. Kandaswamy, C., Silva, L.M., Alexandre, L.A. & Santos, J.M. High-content analysis of breast cancer using single-cell deep transfer learning. *J. Biomol. Screen.* **21**, 252–259 (2016).
150. Eliceiri, K.W. *et al.* Biological imaging software tools. *Nat. Methods* **9**, 697–710 (2012).



Chapter 13

High-Dimensional Profiling: The Theta Comparative Cell Scoring Method

Scott J. Warchal, John C. Dawson, and Neil O. Carragher

Abstract

Principal component analysis enables dimensional reduction of multivariate datasets that are typical in high-content screening. A common analysis utilizing principal components is a distance measurement between a pertubagen—such as small-molecule treatment or shRNA knockdown—and a negative control. This method works well to identify active pertubagens, though it cannot discern between distinct phenotypic responses. Here, we describe an extension of the principal component analysis approach to multivariate high-content screening data to enable quantification of differences in direction in principal component space. The theta comparative cell scoring method can identify and quantify differential phenotypic responses between panels of cell lines to small-molecule treatment to support in vitro pharmacogenomics and drug mechanism-of-action studies.

Key words Phenotypic screening, High-content analysis, Cell-based profiling

1 Introduction

Phenotypic screening allows the identification of treatments that modify a disease model without prior knowledge of the molecular target. This re-emerging method can generate hypotheses for the etiology behind poorly understood diseases, in addition to the discovery of potential therapeutics that act through novel biological mechanisms [1].

One form of phenotypic screening is high-content image-based screening which uses multiple measurements to create a detailed multivariate profile of a pertubagen. This can make screens less biased to preominated target or therapeutic class hypothesis and also create a phenotypic fingerprint to inform mechanism of action [2–5].

A distinct phenotypic response between cell types which represent the broad heterogeneity of human disease and/or more defined clinical subtypes can highlight differences in cellular signaling, metabolic, and biochemical transporter mechanisms that

explain the variation of drug efficacy between patients often observed in the clinic. Correlation of distinct phenotypic response and drug sensitivity across genetically distinct cell types with genomic, transcriptomic, and proteomic data can help elucidate compound mechanism of action and identify molecular biomarkers which predict drug sensitivity and clinical outcomes [6, 7]. We can also use phenotypic similarity between different perturbagens to infer mechanistic similarities. One such example is that small molecules which elicit similar cellular phenotypes are likely to have similar mechanisms of action [8]. Phenotypes can also be used to model disease biology where the underlying signaling pathways and molecular targets associated with disease progression are lacking or poorly understood [9].

In order to quantify complex phenotypes, high-content screening generates multivariate datasets in which multiple phenotypic measurements are taken from a single cell or image. These datasets are usually subjected to some form of dimensionality reduction technique in order to make analysis more manageable. A common dimensional reduction method is principal component analysis, which creates new features (principal components) through orthogonal linear combinations of the original features in order to maximize variation. As principal components are ranked in order of variation, a subset of the principal components can be taken as a replacement for the original feature measurements—with the aim of reducing the number of variables while still retaining as much information as possible. This approach helps visualize complex multivariate data points by plotting them in 2D or 3D principal component space [10, 11].

A simple method used to identify active perturbagens in multivariate datasets is a distance measurement such as Euclidean or Mahalanobis distance between the perturbagen and the negative control in principal component space. This can be used to create a threshold distance to separate the active from inactive, as well as rank perturbations on phenotypic activity [11]. However, this distance metric cannot readily discern between different active phenotypes. Two perturbations may produce very different phenotypes and coordinates in principal component space, and yet have similar distances from the negative control.

In order to discern between perturbations such as these we need a measure of directionality. The idea behind the theta comparative cell scoring (TCCS) method is that different directions in phenotypic space indicate different phenotypes. Therefore measuring the angle between perturbagen-induced phenotypes can be used as a phenotypic similarity score independent of potency. This is very similar to the use of cosine similarity, though the TCCS method centers measurements on the negative control and removes inactive perturbagens that may otherwise produce inaccurate measures of directionality.

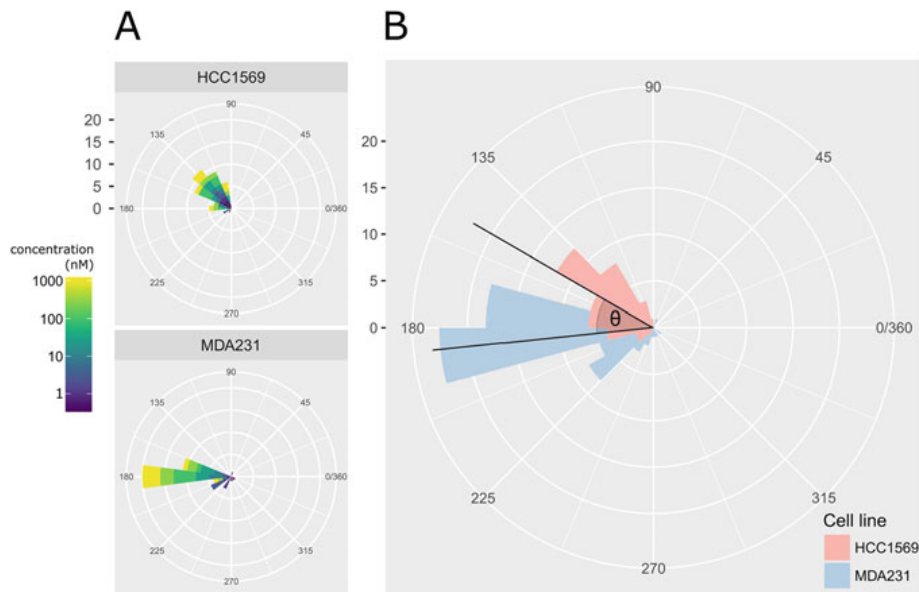


Fig. 1 Circular histograms showing the similar phenotypic direction of HCC1569 and MDA-MB-231 (MDA231) breast cancer cell lines treated with the aurora kinase inhibitor barasertib. (a) Theta values calculated from the first two principal components against a reference vector for both HCC1569 and MDA-MB-231 cell lines treated with barasertib at multiple concentrations. (b) Depiction of the θ value when calculated between a pair of cell lines representing the difference in phenotypic response

The idea of directionality can also be used to produce intuitive and quantitative figures such as circular histograms depicting the direction in phenotypic space or the difference in theta values between two perturbations or samples (Fig. 1).

2 Materials

1. Optical-bottom imaging plates (96- or 384-well).
2. Cell culture medium.
3. Trypsin.
4. Perturbagen Library.
5. Paraformaldehyde (PFA).
6. Triton X-100.
7. Wheat-germ agglutinin 594 (WGA), diluted in dH₂O.
8. SYTO14 green fluorescent nucleic acid stain.
9. Microtiter plate seals.
10. Aluminum foil.
11. Cell painting stock solution: 10 mg/mL Hoechst 33342, 1 mg/mL concanavalin A (diluted in 0.1 M NaHCO₃), 200 U/mL phalloidin-594 (diluted in methanol), 1 mg/mL WGA, 1 mM MitoTracker DeepRed.

12. Blocking buffer: 1% Bovine serum albumin (BSA) in PBS (w/v).
13. Cell painting working solution: 2 $\mu\text{g/mL}$ Hoechst 33342, 11 $\mu\text{g/mL}$ concanavalin A, 3 μM SYTO14, 2.5 U/mL phalloidin-594, 0.25 $\mu\text{g/mL}$ WGA, 600 nM MitoTracker DeepRed.

3 Methods

3.1 Cell Seeding

Preliminary studies are required to determine the optimal number of cells to seed per well (*see Note 1*). This number is dependent on the characteristics of the cell line(s) and the well area in a given plate. Approximate values are provided in Table 1.

1. Using a sub-confluent population of cells, detach the cells by short-term incubation with trypsin and suspend to the desired concentration in cell culture medium.
2. Seed the cells into each well of an optical bottom microtiter 96- or 384-well plate. Make sure that the cells do not settle in the stock of cell suspension by frequently agitating the stock of cell suspension.
3. Incubate the plates containing cells for 24 h.

3.2 Compound Addition

1. Make up stock compound plates in DMSO at 1000 \times the final concentration.
2. Make an intermediate plate by diluting stock compound plate 1:50 in cell culture medium.
3. Remove cell plates from the incubator and transfer a volume from the intermediate plate to the cell plate in a 1:20 dilution.
4. Return cell plates to the humidified, 37 $^{\circ}\text{C}$, 5% CO_2 incubator for an additional 48 h.

3.3 Fluorescent Labeling

3.3.1 Fixation

1. Make a solution of 8% paraformaldehyde (PFA) in phosphate-buffered saline (PBS).
2. Add an equal volume of PFA to each well, and incubate at room temperature for 30 min.
3. Wash wells three times with 50 μL of PBS.

Table 1
Approximate cell seeding densities for different plates

Plate	Cells/well	Volume/well (μL)
96	2000–3000	100
384	750–1500	50

- 3.3.2 Permeabilization**
1. Add 30 μL of 0.1% Triton-X100 solution in PBS to each well, and incubate for 20 min at room temperature.
 2. Wash wells three times with 50 μL of PBS.
- 3.3.3 Cell Labeling**
- Cell labeling protocol adapted from the cell painting protocol [12, 13].
1. Protect the staining solution from light sources by wrapping in aluminum foil.
 2. Add 30 μL of cell painting solution and incubate in a dark place at room temperature for 30 min.
 3. Wash plate three times with 50 μL of PBS. Do not aspirate the final volume.
 4. Seal the plates. If the plates are not imaged immediately, then store them at 4 °C in the dark or wrapped in aluminum foil.
- 3.4 Imaging**
1. Set up the microscope to image five channels at 20 \times magnification. *See* Table 2 for suggested filter settings.
 2. Image multiple sites per well; we recommend a minimum of four.
 3. Adjust the focus and exposure settings (*see* **Note 2**). These settings should be kept constant between batches and comparable experiments as intensity measurements are a function of exposure time.
- 3.5 Image Analysis**
- The following image analysis instructions use CellProfiler [14] nomenclature, though other image analysis software packages may be used to achieve similar results.
1. Extract metadata from either the image or the file path; record the date, plate number, plate name, well, site, and channel for each image.

Table 2
Cell painting reagents and suggested filters

Stain name	Filter name	Filter wavelength	
		Excitation (nm)	Emission (nm)
Hoechst 33342	DAPI	377 \pm 40	447 \pm 60
Con A	FITC	482 \pm 35	536 \pm 40
SYTO14	Cy3	531 \pm 40	594 \pm 40
Phalloidin & WGA	TxRed	562 \pm 40	624 \pm 40
MitoTracker DeepRed	Cy5	628 \pm 40	692 \pm 40

2. Add in external metadata from a .csv file such as compound labels or concentrations and match via plate name and well name/position.
3. Assign each image to a channel name using the extracted channel metadata.
4. Segment the nucleus using `IdentifyPrimaryObjects`.
5. Segment the cell body/cytoplasm using the nucleus object as a seed in the phalloidin/WGA channel with the `IdentifySecondaryObjects` module.
6. Measure image quality in the DAPI channel using `MeasureImageQuality`. Out-of-focus images and any debris can usually be detected in the DAPI channel. Image quality can also be measured in all the channels though the `MeasureImageQuality` module although this will increase analysis time.
7. Measure object size and shape of both the nucleus and cell body with `MeasureObjectSizeShape`.
8. Measure intensity of the nucleus in the DAPI channel and intensity of the cell body in the other four channels using `MeasureObjectIntensity`.
9. Measure texture in the channels for Golgi apparatus and actin staining (WGA channel) in the cell body objects, and the DAPI channel for the nuclei objects using `MeasureTexture`.
10. Measure object neighbors for both nuclei and cell bodies with `MeasureObjectNeighbors`.
11. Export measurement data as .csv files or to a database, excluding any feature measurements that may not be relevant such as object number or object x - y position.

3.6 Data Analysis

1. Check the data produced by the CellProfiler analysis for any missing rows or columns; these need to be removed as appropriate (*see* **Note 3**).
2. Using the `ImageQuality` measurements produced by CellProfiler, identify any images that may be out of focus or contain debris and after visually checking the images remove the data relating to that image if necessary (*see* **Note 4**).
3. If the data is at the object-level, i.e., measurements per cell, then aggregate this to a well median, so each measurement describes the median measurement per feature per well.
4. Remove non-informative features (any measurement columns that are not metadata) such as those with zero or very low variance.
5. Remove redundant features, such as one of a pair of features that are very highly correlated with each other. This can be performed by calculating a correlation matrix of the feature

dataset and finding groups of features that have Spearman's correlations greater than 0.95, and then removing all but one of these features from the dataset.

6. Normalize the data to the negative control values on each plate. This is performed by subtracting the median of the negative control for each feature, per plate (*see Note 5*).
7. Scale the features. For each feature: subtract the feature mean from each individual value, and then divide by the standard deviation of the feature. This standardizes the features to have a mean of zero and unit variance. This is done otherwise features with large values/small units—such as object area which is measured in pixels—will skew the subsequent statistical methods.
8. Calculate the principal components of the feature data and determine the number of principal components needed to account for a proportion of the variance in the dataset, typically 80–90% (*see Note 6*).
9. Remove those principal components that fall outside of this subset.
10. Calculate the negative control medoid, which is the median value for each feature of the negative controls.
11. Adjust the principal component values so that the negative control medoid is centered on the origin (*see Note 7*).
12. Calculate the l1-norm (AKA city-block or Manhattan distance) from the negative control medoid to each data point in principal component space.
13. Calculate the l1-norm of each negative control point from the origin and calculate a distance threshold as 2 standard deviations of these negative control distances. Any compound that has a distance less than this threshold from the medoid of the negative control can be labeled as inactive.
14. Once the inactive compounds have been removed, perturbation similarity can be determined by the angle between perturbation vectors (θ). In two dimensions—using the first two principal components—this can be visualized on a scatter plot. The θ value can be calculated in any number of dimensions, although visualization becomes more difficult. The similarity angle can be calculated by the cosine similarity converted to degrees (*see Eq. (1)*). Note that 180° is the value of maximum dissimilarity, where two perturbagens having completely different directions in phenotypic space, with values greater than 180° becoming increasingly similar as they approach 360° . Therefore θ values are constrained between 0 and 360 by subtracting from 360 any value greater than 180, i.e., $\theta > 180 \rightarrow \theta := 360 - \theta$:

$$\theta = \cos^{-1}\left(\frac{u \cdot v}{\|u\| \|v\|}\right) \times \frac{180}{\pi} \quad (1)$$

where u and v are the vectors of principal components for each compound.

15. If two principal components capture a large proportion of the variance in the dataset then a visualization can be made by calculating θ for every perturbagen against a common reference vector, and then plotting a circular histogram of the θ values (*see Note 8*).
16. Identify cell line pairs treated with the same compound that have significantly different theta values (*see Note 9*), indicating a distinct phenotypic response between cell lines to a compound treatment.
17. *See Notes 10 and 11* for additional troubleshooting steps.

4 Notes

1. Too few cells will provide fewer replicates and may run the risk of having no cells contained in an image if a perturbagen reduces the cell number. Seeding too many cells can mean cells do not form a single monolayer which makes image analysis more difficult. We advise seeding the number of cells to result in approximately 60–70% confluence.
2. After setting the focus for the first channel (DAPI/Hoechst), all additional channel's focus settings are based on these measurements. Therefore adjusting the focus settings for the first channel will also affect all of the other channels, so it is advised to set this first and check a few different wells to ensure that the settings are robust.
3. It is recommended to remove columns containing large amounts of missing numbers. This can often be caused by missing metadata in certain samples, or some features that remain constant between samples—such as Euler number—that may be transformed to missing data entries after scaling or aggregation. Once columns of largely missing data have been removed, rows containing missing values can be removed. Without first removing the missing data columns it is often possible to erroneously remove the entire or large proportions of the dataset when using missing rows as the first step.
4. Out-of-focus images can be detected using ImageQuality_PowerLogLogSlope measurements in the nuclei channel. Images with very low values are likely to be out of focus [15]. Debris such as dust or fibers typically show up in the nuclei channel,

and can be detected by identifying images with a large percentage of saturated pixels.

5. Normalizing to the negative control is a useful step in any plate-based screen to remove any batch effects between plates that may influence the results. It is especially important when comparing effects between cell lines as this converts the values to changes from the negative control for that particular plate; as we expected to have a single cell line per plate this also removes any inherent phenotypic differences between the cell lines, and allows the compound-induced changes to be comparable.
6. The number of principal components required to capture a specified proportion of the variance in the data can be calculated in *R* (assuming that data is numeric feature data), to calculate the value for 80% of the variance:

```
threshold <- 0.8
pca_output <- prcomp(data)
pc_variance <- pca_output$sdev^2
cumulative_proportion_variance <- cumsum(pc_variance) / sum(pc_variance)
n_components <- min(which(cumulative_proportion_variance >= threshold))
```

7. To center the principal component data so that the medoid of the negative control lies on the origin, find the medoid for the negative control values, which is the median value for each feature column for the negative control values; find how much this differs from the origin for each feature; shift all values for each feature by this difference, e.g., in *R*:

```
centre_control <- function(df, feature_cols, compd_col, neg_control = "DMSO") {
  # 1. the median value for the DMSO values for each measured feature
  mediodid <- apply(df[df[, compd_col] == neg_control, feature_cols], 2, median)
  # 2. calculate the difference from the origin for each mediodid position
  delta <- 0 - mediodid
  # 3. iterate along columns and adjust to centre the DMSO data
  for (i in seq_along(feature_cols)) {
    feature <- feature_cols[i]
    df[, feature] <- df[, feature] + d[i]
  }
  return(df)
}
```

8. Creating circular histograms: If the principal component vector only contains information regarding two principal components, then we can calculate a θ value for each perturbagen against a common reference such as (0, 1). This generates a θ value for each perturbagen which can be plotted as a histogram. Without constraining them, the θ values are ranged between

0 and 360. As either end of this range is equivalent to the x -axis of this histogram can be wrapped round into a circle which can be used to visualize the phenotypic direction induced by a perturbagen (Fig. 1).

9. To identify distinct phenotypic responses between cell lines treated with a perturbagen, a theta value has to be calculated for each pair of cell lines per perturbagen. Cell lines that elicit a similar response to a given perturbagen will produce a low θ value, indicating that they produce similar phenotypic trajectories, whereas a θ value approaching 180 indicates opposite phenotypic directions. In our experience a histogram of all measured θ values produces a log-normal distribution, indicating that most perturbagens produce similar phenotypic response between cell lines.
10. Image analysis can take considerable time for large numbers of images. We recommended using either a computing cluster or a cloud computing service to process many images in parallel.
11. Large .csv files can also cause problems. If files exceed several GBs we recommend users switch to a database format such as SQLite.

Acknowledgments

This work was supported by a Cancer Research UK Ph.D. Studentship award to the Cancer Research UK Edinburgh Centre.

References

1. Swinney DC, Anthony J (2011) How were new medicines discovered? *Nat Rev Drug Discov* 10:507–519
2. Ljosa V, Caie PD, Ter Horst R, Sokolnicki KL, Jenkins EL, Daya S et al (2013) Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J Biomol Screen* 18:1321–1329
3. Singh S, Carpenter AE, Genovesio A (2014) Increasing the content of high-content screening: an overview. *J Biomol Screen* 19:640–650
4. Reisen F, Sauty de Chalon A, Pfeifer M, Zhang X, Gabriel D, Selzer P (2015) Linking phenotypes and modes of action through high-content screen fingerprints. *Assay Drug Dev Technol* 13:150810081821009
5. Kümmel A, Selzer P, Siebert D, Schmidt I, Reinhardt J, Götte M et al (2012) Differentiation and visualization of diverse cellular phenotypic responses in primary high-content screening. *J Biomol Screen* 17:843–849
6. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW et al (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483:570–575
7. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S et al (2012) The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483:603–607
8. Perlman Z, Slack M, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ (2004) Multidimensional drug profiling by automated microscopy. *Science* 306:1194–1199
9. Vincent F, Loria P, Pregel M, Stanton R, Kitching L, Nocka K et al (2015) Developing predictive assays: the phenotypic screening “rule of 3”. *Sci Transl Med* 7:293ps15
10. Tanaka M, Bateman R, Rauh D, Vaisberg E, Ramachandani S, Zhang C et al (2005) An unbiased cell morphology-based screen for

- new, biologically active small molecules. *PLoS Biol* 3:0764–0776
11. Caie PD, Walls RE, Ingleston-Orme A, Daya S, Houslay T, Eagle R et al (2010) High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol Canc Ther* 9:1913–1926
 12. Gustafsdottir SM, Ljosa V, Sokolnicki KL, Wilson JA, Walpita D, Kemp MM et al (2013) Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* 8:e80999
 13. Warchal SJ, Dawson JC, Carragher NO (2016) Development of the theta comparative cell scoring method to quantify diverse phenotypic responses between distinct cell types. *Assay Drug Dev Technol* 14:395–406
 14. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O et al (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* 7:R100
 15. Bray M-A, Fraser AN, Hasaka TP, Carpenter AE (2012) Workflow and metrics for image quality control in large-scale high-content screens. *J Biomol Screen* 17:266–274

CellProfiler features

Cells_AreaShape_Compactness	Cells_AreaShape_Eccentricity	Cells_AreaShape_MaxFeretDiameter
Cells_AreaShape_Orientation	Cells_AreaShape_Perimeter	Cells_AreaShape_Solidity
Cells_AreaShape_Zernike_1_1	Cells_AreaShape_Zernike_2_0	Cells_AreaShape_Zernike_2_2
Cells_AreaShape_Zernike_3_1	Cells_AreaShape_Zernike_4_0	Cells_AreaShape_Zernike_4_2
Cells_Correlation_Correlation_W2_W3	Cells_Correlation_Costes_W2_W3	Cells_Correlation_Costes_W3_W2
Cells_Correlation_K_W2_W3	Cells_Correlation_K_W3_W2	Cells_Correlation_Manders_W2_W3
Cells_Correlation_Manders_W3_W2	Cells_Correlation_Overlap_W2_W3	Cells_Correlation_RWC_W3_W2
Cells_Granularity_10_W4	Cells_Granularity_10_W5	Cells_Granularity_11_W4
Cells_Granularity_11_W5	Cells_Granularity_12_W4	Cells_Granularity_12_W5
Cells_Granularity_13_W4	Cells_Granularity_13_W5	Cells_Granularity_14_W4
Cells_Granularity_14_W5	Cells_Granularity_15_W4	Cells_Granularity_15_W5
Cells_Granularity_16_W4	Cells_Granularity_16_W5	Cells_Granularity_1_W4
Cells_Granularity_1_W5	Cells_Granularity_2_W4	Cells_Granularity_2_W5
Cells_Granularity_3_W4	Cells_Granularity_3_W5	Cells_Granularity_4_W4
Cells_Granularity_4_W5	Cells_Granularity_5_W4	Cells_Granularity_5_W5
Cells_Granularity_6_W4	Cells_Granularity_6_W5	Cells_Granularity_7_W5
Cells_Granularity_8_W4	Cells_Granularity_8_W5	Cells_Granularity_9_W4
Cells_Granularity_9_W5	Cells_Intensity_IntegratedIntensityEdge_W2	Cells_Intensity_IntegratedIntensityEdge_W3
Cells_Intensity_IntegratedIntensityEdge_W4	Cells_Intensity_IntegratedIntensityEdge_W5	Cells_Intensity_IntegratedIntensity_W3
Cells_Intensity_IntegratedIntensity_W5	Cells_Intensity_MADIntensity_W4	Cells_Intensity_MassDisplacement_W2
Cells_Intensity_MassDisplacement_W4	Cells_Intensity_MassDisplacement_W5	Cells_Intensity_MaxIntensity_W3
Cells_Intensity_MaxIntensity_W4	Cells_Intensity_MeanIntensityEdge_W2	Cells_Intensity_MeanIntensityEdge_W3
Cells_Intensity_MeanIntensityEdge_W4	Cells_Intensity_MeanIntensityEdge_W5	Cells_Intensity_MedianIntensity_W4
Cells_Intensity_MedianIntensity_W5	Cells_Intensity_MinIntensity_W2	Cells_Intensity_MinIntensity_W3
Cells_Intensity_MinIntensity_W4	Cells_Intensity_MinIntensity_W5	Cells_Intensity_StdIntensityEdge_W2
Cells_Intensity_StdIntensityEdge_W4	Cells_Intensity_StdIntensityEdge_W5	Cells_Intensity_StdIntensity_W3
Cells_Intensity_StdIntensity_W4	Cells_Intensity_StdIntensity_W5	Cells_Neighbors_PercentTouching_2
Cells_RadialDistribution_FracAtD_W2_2of4	Cells_RadialDistribution_FracAtD_W3_2of4	Cells_RadialDistribution_FracAtD_W3_3of4
Cells_RadialDistribution_FracAtD_W3_4of4	Cells_RadialDistribution_FracAtD_W4_2of4	Cells_RadialDistribution_FracAtD_W5_1of4
Cells_RadialDistribution_FracAtD_W5_2of4	Cells_RadialDistribution_FracAtD_W5_3of4	Cells_RadialDistribution_MeanFrac_W2_2of4
Cells_RadialDistribution_MeanFrac_W3_2of4	Cells_RadialDistribution_MeanFrac_W4_2of4	Cells_RadialDistribution_MeanFrac_W4_3of4
Cells_RadialDistribution_MeanFrac_W4_4of4	Cells_RadialDistribution_MeanFrac_W5_1of4	Cells_RadialDistribution_MeanFrac_W5_2of4
Cells_RadialDistribution_MeanFrac_W5_3of4	Cells_RadialDistribution_RadialCV_W2_1of4	Cells_RadialDistribution_RadialCV_W2_2of4
Cells_RadialDistribution_RadialCV_W2_3of4	Cells_RadialDistribution_RadialCV_W3_1of4	Cells_RadialDistribution_RadialCV_W3_2of4
Cells_RadialDistribution_RadialCV_W3_3of4	Cells_RadialDistribution_RadialCV_W3_4of4	Cells_RadialDistribution_RadialCV_W4_1of4
Cells_RadialDistribution_RadialCV_W4_2of4	Cells_RadialDistribution_RadialCV_W4_3of4	Cells_RadialDistribution_RadialCV_W4_4of4
Cells_RadialDistribution_RadialCV_W5_1of4	Cells_RadialDistribution_RadialCV_W5_2of4	Cells_RadialDistribution_RadialCV_W5_3of4
Cells_RadialDistribution_RadialCV_W5_4of4	Cells_RadialDistribution_ZernikeMagnitude_W2_2_2	Cells_RadialDistribution_ZernikeMagnitude_W2_3_1
Cells_RadialDistribution_ZernikeMagnitude_W2_4_0	Cells_RadialDistribution_ZernikeMagnitude_W2_4_2	Cells_RadialDistribution_ZernikeMagnitude_W2_5_1
Cells_RadialDistribution_ZernikeMagnitude_W3_3_1	Cells_RadialDistribution_ZernikeMagnitude_W4_1_1	Cells_RadialDistribution_ZernikeMagnitude_W4_3_1
Cells_RadialDistribution_ZernikeMagnitude_W4_4_0	Cells_RadialDistribution_ZernikeMagnitude_W5_2_2	Cells_RadialDistribution_ZernikeMagnitude_W5_3_1
Cells_RadialDistribution_ZernikeMagnitude_W5_4_0	Cells_RadialDistribution_ZernikeMagnitude_W5_4_2	Cells_RadialDistribution_ZernikeMagnitude_W5_5_1
Cells_RadialDistribution_ZernikeMagnitude_W5_6_0	Cells_Texture_AngularSecondMoment_W2_3_135	Cells_Texture_AngularSecondMoment_W3_3_135
Cells_Texture_AngularSecondMoment_W4_3_135	Cells_Texture_AngularSecondMoment_W5_3_135	Cells_Texture_Contrast_W2_3_135
Cells_Texture_Contrast_W3_3_135	Cells_Texture_Contrast_W4_3_135	Cells_Texture_Contrast_W5_3_135
Cells_Texture_Correlation_W2_3_135	Cells_Texture_Correlation_W3_3_135	Cells_Texture_Correlation_W4_3_135
Cells_Texture_Correlation_W5_3_135	Cells_Texture_DifferenceEntropy_W5_3_45	Cells_Texture_Entropy_W2_3_135
Cells_Texture_Entropy_W3_3_135	Cells_Texture_Entropy_W4_3_135	Cells_Texture_Entropy_W5_3_135
Cells_Texture_Gabor_W2_3	Cells_Texture_Gabor_W3_3	Cells_Texture_Gabor_W4_3
Cells_Texture_Gabor_W5_3	Cells_Texture_InfoMeas1_W2_3_135	Cells_Texture_InfoMeas1_W3_3_135
Cells_Texture_InfoMeas1_W4_3_135	Cells_Texture_InfoMeas1_W5_3_135	Cells_Texture_InfoMeas2_W2_3_135
Cells_Texture_InfoMeas2_W3_3_135	Cells_Texture_InfoMeas2_W4_3_135	Cells_Texture_InfoMeas2_W5_3_45
Cells_Texture_InverseDifferenceMoment_W2_3_135	Cells_Texture_InverseDifferenceMoment_W3_3_135	Cells_Texture_InverseDifferenceMoment_W4_3_135
Cells_Texture_InverseDifferenceMoment_W5_3_135	Cells_Texture_SumAverage_W2_3_135	Cells_Texture_SumAverage_W3_3_135
Cells_Texture_SumAverage_W4_3_135	Cells_Texture_SumAverage_W5_3_135	Cells_Texture_SumVariance_W2_3_135
Cells_Texture_SumVariance_W3_3_135	Cells_Texture_SumVariance_W4_3_135	Cells_Texture_SumVariance_W5_3_135
Correlation_Correlation_W2_W3	Correlation_Manders_W3_W2	Correlation_RWC_W2_W3
Correlation_RWC_W3_W2	Correlation_Slope_W2_W3	Count_Nuclei
Granularity_10_W4	Granularity_10_W5	Granularity_11_W1
Granularity_11_W4	Granularity_11_W5	Granularity_12_W1
Granularity_12_W4	Granularity_12_W5	Granularity_13_W1
Granularity_13_W4	Granularity_13_W5	Granularity_14_W1
Granularity_14_W4	Granularity_14_W5	Granularity_15_W1
Granularity_15_W4	Granularity_15_W5	Granularity_16_W1
Granularity_16_W4	Granularity_16_W5	Granularity_1_W1
Granularity_1_W5	Granularity_2_W1	Granularity_2_W4
Granularity_2_W5	Granularity_3_W1	Granularity_3_W4
Granularity_4_W1	Granularity_4_W4	Granularity_4_W5
Granularity_5_W1	Granularity_6_W1	Granularity_6_W4
Granularity_6_W5	Granularity_7_W1	Granularity_7_W5
Granularity_8_W4	Granularity_8_W5	Granularity_9_W1

Granularity_9_W4	Granularity_9_W5	Nuclei_AreaShape_Compactness
Nuclei_AreaShape_Eccentricity	Nuclei_AreaShape_FormFactor	Nuclei_AreaShape_MajorAxisLength
Nuclei_AreaShape_Orientation	Nuclei_AreaShape_Solidity	Nuclei_AreaShape_Zernike_1_1
Nuclei_AreaShape_Zernike_2_0	Nuclei_AreaShape_Zernike_2_2	Nuclei_AreaShape_Zernike_3_1
Nuclei_AreaShape_Zernike_3_3	Nuclei_AreaShape_Zernike_4_0	Nuclei_AreaShape_Zernike_4_4
Nuclei_Granularity_10_W1	Nuclei_Granularity_11_W1	Nuclei_Granularity_12_W1
Nuclei_Granularity_13_W1	Nuclei_Granularity_14_W1	Nuclei_Granularity_15_W1
Nuclei_Granularity_16_W1	Nuclei_Granularity_1_W1	Nuclei_Granularity_2_W1
Nuclei_Granularity_3_W1	Nuclei_Granularity_4_W1	Nuclei_Granularity_5_W1
Nuclei_Granularity_6_W1	Nuclei_Granularity_7_W1	Nuclei_Granularity_8_W1
Nuclei_Granularity_9_W1	Nuclei_Intensity_IntegratedIntensityEdge_W1	Nuclei_Intensity_IntegratedIntensity_W1
Nuclei_Intensity_LowerQuartileIntensity_W1	Nuclei_Intensity_MassDisplacement_W1	Nuclei_Intensity_MaxIntensityEdge_W1
Nuclei_Intensity_MaxIntensity_W1	Nuclei_Intensity_MeanIntensityEdge_W1	Nuclei_Intensity_MinIntensity_W1
Nuclei_Intensity_StdIntensityEdge_W1	Nuclei_Texture_AngularSecondMoment_W1_3_135	Nuclei_Texture_Contrast_W1_3_135
Nuclei_Texture_Correlation_W1_3_135	Nuclei_Texture_Entropy_W1_3_135	Nuclei_Texture_Gabor_W1_3
Nuclei_Texture_InfoMeas1_W1_3_135	Nuclei_Texture_SumAverage_W1_3_135	Nuclei_Texture_SumEntropy_W1_3_135
Nuclei_Texture_SumVariance_W1_3_135	Cells_AreaShape_MaximumRadius	Cells_AreaShape_MeanRadius
Cells_AreaShape_MedianRadius	Cells_AreaShape_MinFerretDiameter	Cells_AreaShape_MinorAxisLength
Cells_Intensity_IntegratedIntensity_W4	Cells_AreaShape_Area	Cells_AreaShape_Zernike_0_0
Cells_AreaShape_Extent	Cells_AreaShape_FormFactor	Cells_AreaShape_MinFerretDiameter
Cells_AreaShape_MinorAxisLength	Cells_AreaShape_MajorAxisLength	Cells_Correlation_RWC_W2_W3
Cells_Granularity_7_W4	Cells_Intensity_IntegratedIntensity_W4	Cells_Intensity_IntegratedIntensity_W2
Cells_Intensity_MedianIntensity_W2	Cells_Intensity_LowerQuartileIntensity_W2	Cells_Intensity_MedianIntensity_W3
Cells_Intensity_LowerQuartileIntensity_W3	Cells_Intensity_LowerQuartileIntensity_W4	Cells_Intensity_LowerQuartileIntensity_W5
Cells_Intensity_MaxIntensity_W2	Cells_Intensity_StdIntensity_W2	Cells_Intensity_UpperQuartileIntensity_W2
Cells_RadialDistribution_ZernikeMagnitude_W2_2_0	Cells_Intensity_MADIntensity_W2	Cells_Intensity_MADIntensity_W3
Cells_Intensity_UpperQuartileIntensity_W5	Cells_RadialDistribution_ZernikeMagnitude_W5_2_0	Cells_Intensity_MADIntensity_W5
Cells_Intensity_MassDisplacement_W3	Cells_Intensity_MaxIntensityEdge_W2	Cells_Intensity_StdIntensityEdge_W3
Cells_Intensity_MaxIntensityEdge_W3	Cells_Intensity_MaxIntensityEdge_W4	Cells_Intensity_MaxIntensityEdge_W5
Cells_Intensity_MaxIntensity_W5	Cells_Intensity_UpperQuartileIntensity_W2	Cells_RadialDistribution_ZernikeMagnitude_W2_0_0
Cells_RadialDistribution_ZernikeMagnitude_W2_2_0	Cells_Intensity_MeanIntensity_W2	Cells_Intensity_UpperQuartileIntensity_W3
Cells_RadialDistribution_ZernikeMagnitude_W3_0_0	Cells_RadialDistribution_ZernikeMagnitude_W3_1_1	Cells_RadialDistribution_ZernikeMagnitude_W3_2_0
Cells_Intensity_MeanIntensity_W3	Cells_Intensity_UpperQuartileIntensity_W4	Cells_RadialDistribution_ZernikeMagnitude_W4_0_0
Cells_RadialDistribution_ZernikeMagnitude_W4_2_0	Cells_Intensity_MeanIntensity_W4	Cells_Intensity_UpperQuartileIntensity_W5
Cells_RadialDistribution_ZernikeMagnitude_W5_0_0	Cells_RadialDistribution_ZernikeMagnitude_W5_2_0	Cells_Intensity_MeanIntensity_W5
Cells_Intensity_MinIntensityEdge_W2	Cells_Intensity_MinIntensityEdge_W3	Cells_Intensity_MinIntensityEdge_W4
Cells_Intensity_MinIntensityEdge_W5	Cells_Neighbors_NumberOfNeighbors_2	Cells_RadialDistribution_FracAtD_W4_1of4
Cells_RadialDistribution_FracAtD_W2_1of4	Cells_RadialDistribution_FracAtD_W4_3of4	Cells_RadialDistribution_FracAtD_W2_3of4
Cells_RadialDistribution_FracAtD_W4_4of4	Cells_RadialDistribution_FracAtD_W5_4of4	Cells_RadialDistribution_MeanFrac_W2_4of4
Cells_RadialDistribution_FracAtD_W2_4of4	Cells_RadialDistribution_FracAtD_W3_1of4	Cells_RadialDistribution_MeanFrac_W2_1of4
Cells_RadialDistribution_MeanFrac_W2_3of4	Cells_RadialDistribution_MeanFrac_W3_1of4	Cells_RadialDistribution_MeanFrac_W3_3of4
Cells_RadialDistribution_MeanFrac_W5_4of4	Cells_RadialDistribution_MeanFrac_W3_4of4	Cells_RadialDistribution_MeanFrac_W4_1of4
Cells_RadialDistribution_RadialCV_W2_4of4	Cells_RadialDistribution_ZernikeMagnitude_W2_1_1	Cells_RadialDistribution_ZernikeMagnitude_W5_1_1
Cells_Texture_AngularSecondMoment_W2_3_45	Cells_Texture_AngularSecondMoment_W2_3_90	Cells_Texture_AngularSecondMoment_W2_3_0
Cells_Texture_AngularSecondMoment_W3_3_45	Cells_Texture_AngularSecondMoment_W3_3_90	Cells_Texture_AngularSecondMoment_W3_3_0
Cells_Texture_AngularSecondMoment_W4_3_45	Cells_Texture_AngularSecondMoment_W4_3_90	Cells_Texture_AngularSecondMoment_W4_3_0
Cells_Texture_AngularSecondMoment_W5_3_45	Cells_Texture_AngularSecondMoment_W5_3_90	Cells_Texture_AngularSecondMoment_W5_3_0
Cells_Texture_Contrast_W2_3_45	Cells_Texture_Contrast_W2_3_90	Cells_Texture_DifferenceEntropy_W2_3_0
Cells_Texture_DifferenceEntropy_W2_3_135	Cells_Texture_DifferenceEntropy_W2_3_45	Cells_Texture_DifferenceEntropy_W2_3_90
Cells_Texture_DifferenceVariance_W2_3_0	Cells_Texture_DifferenceVariance_W2_3_135	Cells_Texture_DifferenceVariance_W2_3_45
Cells_Texture_DifferenceVariance_W2_3_90	Cells_Texture_Contrast_W2_3_0	Cells_Texture_Contrast_W3_3_45
Cells_Texture_Contrast_W3_3_90	Cells_Texture_DifferenceEntropy_W3_3_0	Cells_Texture_DifferenceEntropy_W3_3_135
Cells_Texture_DifferenceEntropy_W3_3_45	Cells_Texture_DifferenceVariance_W3_3_0	Cells_Texture_DifferenceVariance_W3_3_135
Cells_Texture_DifferenceVariance_W3_3_45	Cells_Texture_DifferenceVariance_W3_3_90	Cells_Texture_Contrast_W3_3_0
Cells_Texture_Contrast_W4_3_45	Cells_Texture_Contrast_W4_3_90	Cells_Texture_DifferenceEntropy_W4_3_0
Cells_Texture_DifferenceEntropy_W4_3_135	Cells_Texture_DifferenceEntropy_W4_3_45	Cells_Texture_DifferenceEntropy_W4_3_90
Cells_Texture_DifferenceVariance_W4_3_0	Cells_Texture_DifferenceVariance_W4_3_135	Cells_Texture_DifferenceVariance_W4_3_45
Cells_Texture_DifferenceVariance_W4_3_90	Cells_Texture_Contrast_W4_3_0	Cells_Texture_Contrast_W5_3_45
Cells_Texture_Contrast_W5_3_90	Cells_Texture_DifferenceEntropy_W5_3_0	Cells_Texture_DifferenceVariance_W5_3_0
Cells_Texture_DifferenceVariance_W5_3_135	Cells_Texture_DifferenceVariance_W5_3_45	Cells_Texture_DifferenceVariance_W5_3_90
Cells_Texture_Contrast_W5_3_0	Cells_Texture_Correlation_W2_3_45	Cells_Texture_Correlation_W2_3_90

Cells_Texture_Correlation_W2_3_0	Cells_Texture_Correlation_W3_3_45	Cells_Texture_Correlation_W3_3_90
Cells_Texture_Correlation_W3_3_0	Cells_Texture_Correlation_W4_3_45	Cells_Texture_Correlation_W4_3_90
Cells_Texture_Correlation_W4_3_0	Cells_Texture_Correlation_W5_3_45	Cells_Texture_Correlation_W5_3_90
Cells_Texture_Correlation_W5_3_0	Cells_Texture_DifferenceEntropy_W3_3_90	Cells_Texture_DifferenceEntropy_W5_3_90
Cells_Texture_DifferenceEntropy_W5_3_135	Cells_Texture_Entropy_W2_3_45	Cells_Texture_Entropy_W2_3_90
Cells_Texture_SumEntropy_W2_3_0	Cells_Texture_SumEntropy_W2_3_135	Cells_Texture_SumEntropy_W2_3_45
Cells_Texture_SumEntropy_W2_3_90	Cells_Texture_Entropy_W2_3_0	Cells_Texture_Entropy_W3_3_45
Cells_Texture_Entropy_W3_3_90	Cells_Texture_SumEntropy_W3_3_0	Cells_Texture_SumEntropy_W3_3_135
Cells_Texture_SumEntropy_W3_3_45	Cells_Texture_SumEntropy_W3_3_90	Cells_Texture_Entropy_W3_3_0
Cells_Texture_Entropy_W4_3_45	Cells_Texture_Entropy_W4_3_90	Cells_Texture_SumEntropy_W4_3_0
Cells_Texture_SumEntropy_W4_3_135	Cells_Texture_SumEntropy_W4_3_45	Cells_Texture_SumEntropy_W4_3_90
Cells_Texture_Entropy_W4_3_0	Cells_Texture_Entropy_W5_3_45	Cells_Texture_Entropy_W5_3_90
Cells_Texture_SumEntropy_W5_3_0	Cells_Texture_SumEntropy_W5_3_135	Cells_Texture_SumEntropy_W5_3_45
Cells_Texture_SumEntropy_W5_3_90	Cells_Texture_Entropy_W5_3_0	Cells_Texture_InfoMeas1_W2_3_45
Cells_Texture_InfoMeas1_W2_3_90	Cells_Texture_InfoMeas1_W2_3_0	Cells_Texture_InfoMeas1_W3_3_45
Cells_Texture_InfoMeas1_W3_3_90	Cells_Texture_InfoMeas1_W3_3_0	Cells_Texture_InfoMeas1_W4_3_45
Cells_Texture_InfoMeas1_W4_3_90	Cells_Texture_InfoMeas1_W4_3_0	Cells_Texture_InfoMeas1_W5_3_45
Cells_Texture_InfoMeas1_W5_3_90	Cells_Texture_InfoMeas1_W5_3_0	Cells_Texture_InfoMeas2_W2_3_45
Cells_Texture_InfoMeas2_W2_3_90	Cells_Texture_InfoMeas2_W5_3_0	Cells_Texture_InfoMeas2_W2_3_0
Cells_Texture_InfoMeas2_W3_3_45	Cells_Texture_InfoMeas2_W3_3_90	Cells_Texture_InfoMeas2_W3_3_0
Cells_Texture_InfoMeas2_W4_3_45	Cells_Texture_InfoMeas2_W4_3_90	Cells_Texture_InfoMeas2_W4_3_0
Cells_Texture_InfoMeas2_W5_3_90	Cells_Texture_InfoMeas2_W5_3_135	Cells_Texture_InverseDifferenceMoment_W2_3_45
Cells_Texture_InverseDifferenceMoment_W2_3_90	Cells_Texture_InverseDifferenceMoment_W2_3_0	Cells_Texture_InverseDifferenceMoment_W3_3_45
Cells_Texture_InverseDifferenceMoment_W3_3_90	Cells_Texture_InverseDifferenceMoment_W3_3_0	Cells_Texture_InverseDifferenceMoment_W4_3_45
Cells_Texture_InverseDifferenceMoment_W4_3_90	Cells_Texture_InverseDifferenceMoment_W4_3_0	Cells_Texture_InverseDifferenceMoment_W5_3_45
Cells_Texture_InverseDifferenceMoment_W5_3_90	Cells_Texture_InverseDifferenceMoment_W5_3_0	Cells_Texture_SumAverage_W2_3_45
Cells_Texture_SumAverage_W2_3_90	Cells_Texture_SumAverage_W2_3_0	Cells_Texture_SumAverage_W3_3_45
Cells_Texture_SumAverage_W3_3_90	Cells_Texture_SumAverage_W3_3_0	Cells_Texture_SumAverage_W4_3_45
Cells_Texture_SumAverage_W4_3_90	Cells_Texture_SumAverage_W4_3_0	Cells_Texture_SumAverage_W5_3_45
Cells_Texture_SumAverage_W5_3_90	Cells_Texture_SumAverage_W5_3_0	Cells_Texture_SumVariance_W2_3_45
Cells_Texture_SumVariance_W2_3_90	Cells_Texture_Variance_W2_3_0	Cells_Texture_Variance_W2_3_135
Cells_Texture_Variance_W2_3_45	Cells_Texture_Variance_W2_3_90	Cells_Texture_SumVariance_W2_3_0
Cells_Texture_SumVariance_W3_3_45	Cells_Texture_SumVariance_W3_3_90	Cells_Texture_Variance_W3_3_0
Cells_Texture_Variance_W3_3_135	Cells_Texture_Variance_W3_3_45	Cells_Texture_Variance_W3_3_90
Cells_Texture_SumVariance_W3_3_0	Cells_Texture_SumVariance_W4_3_45	Cells_Texture_SumVariance_W4_3_90
Cells_Texture_Variance_W4_3_0	Cells_Texture_Variance_W4_3_135	Cells_Texture_Variance_W4_3_45
Cells_Texture_Variance_W4_3_90	Cells_Texture_SumVariance_W4_3_0	Cells_Texture_SumVariance_W5_3_45
Cells_Texture_SumVariance_W5_3_90	Cells_Texture_Variance_W5_3_0	Cells_Texture_Variance_W5_3_135
Cells_Texture_Variance_W5_3_45	Cells_Texture_Variance_W5_3_90	Cells_Texture_SumVariance_W5_3_0
Correlation_Manders_W2_W3	Count_Cells	Granularity_10_W1
Granularity_1_W4	Granularity_3_W5	Granularity_5_W4
Granularity_5_W5	Granularity_7_W4	Granularity_8_W1
Nuclei_AreaShape_MaxFeretDiameter	Nuclei_AreaShape_MaximumRadius	Nuclei_AreaShape_MeanRadius
Nuclei_AreaShape_MedianRadius	Nuclei_AreaShape_MinFeretDiameter	Nuclei_AreaShape_MinorAxisLength
Nuclei_AreaShape_Perimeter	Nuclei_Texture_InverseDifferenceMoment_W1_3_0	Nuclei_Texture_InverseDifferenceMoment_W1_3_135
Nuclei_Texture_InverseDifferenceMoment_W1_3_45	Nuclei_Texture_InverseDifferenceMoment_W1_3_90	Nuclei_AreaShape_Area
Nuclei_AreaShape_Extent	Nuclei_AreaShape_Zernike_0_0	Nuclei_Intensity_MeanIntensity_W1
Nuclei_Intensity_StdIntensity_W1	Nuclei_Intensity_UpperQuartileIntensity_W1	Nuclei_Intensity_MADIntensity_W1
Nuclei_Intensity_MedianIntensity_W1	Nuclei_Intensity_MinIntensityEdge_W1	Nuclei_Texture_AngularSecondMoment_W1_3_45
Nuclei_Texture_AngularSecondMoment_W1_3_90	Nuclei_Texture_AngularSecondMoment_W1_3_0	Nuclei_Texture_Contrast_W1_3_45
Nuclei_Texture_Contrast_W1_3_90	Nuclei_Texture_DifferenceEntropy_W1_3_0	Nuclei_Texture_DifferenceEntropy_W1_3_135
Nuclei_Texture_DifferenceEntropy_W1_3_45	Nuclei_Texture_DifferenceEntropy_W1_3_90	Nuclei_Texture_DifferenceVariance_W1_3_0
Nuclei_Texture_DifferenceVariance_W1_3_135	Nuclei_Texture_DifferenceVariance_W1_3_45	Nuclei_Texture_DifferenceVariance_W1_3_90
Nuclei_Texture_Contrast_W1_3_0	Nuclei_Texture_Correlation_W1_3_45	Nuclei_Texture_Correlation_W1_3_90
Nuclei_Texture_InfoMeas2_W1_3_0	Nuclei_Texture_InfoMeas2_W1_3_135	Nuclei_Texture_InfoMeas2_W1_3_45
Nuclei_Texture_InfoMeas2_W1_3_90	Nuclei_Texture_Correlation_W1_3_0	Nuclei_Texture_Entropy_W1_3_45
Nuclei_Texture_Entropy_W1_3_90	Nuclei_Texture_Entropy_W1_3_0	Nuclei_Texture_InfoMeas1_W1_3_45
Nuclei_Texture_InfoMeas1_W1_3_90	Nuclei_Texture_InfoMeas1_W1_3_0	Nuclei_Texture_SumAverage_W1_3_45
Nuclei_Texture_SumAverage_W1_3_90	Nuclei_Texture_SumAverage_W1_3_0	Nuclei_Texture_SumEntropy_W1_3_45
Nuclei_Texture_SumEntropy_W1_3_90	Nuclei_Texture_SumEntropy_W1_3_0	Nuclei_Texture_SumVariance_W1_3_45
Nuclei_Texture_SumVariance_W1_3_90	Nuclei_Texture_Variance_W1_3_0	Nuclei_Texture_Variance_W1_3_135
Nuclei_Texture_Variance_W1_3_45	Nuclei_Texture_Variance_W1_3_90	Nuclei_Texture_SumVariance_W1_3_0