IMAGE INFORMATICS APPROACHES TO ADVANCE CANCER DRUG DISCOVERY

Scott J. Warchal

Doctor of Philosophy
The University of Edinburgh
2018

DECLARATION

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

– Scott Warchal, 2018

epigraph here

ACKNOWLEDGEMENTS

Acknowledgements here.

ABSTRACT

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. Putatative hits 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 disfunctional 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 subcellular 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 though 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

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

2D Two-Dimensional

3D Three-Dimensional

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 Multplicity 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

SAR Structure Activity Relationship

SMILE Simplified Molecular Input Line Entry System

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 cost of developing a new drug has steadily increased, a study by Scannel *et al.* noted the cost has approximately doubled every 9 years,[1] this observation has been dubbed "Eroom's law" in a homage to Moore's law.[i] The cost of bringing a new drug to market is now approaching £1 billion, taking 10 years from initial concept through to regulatory approval, the reasons behind this ever-increasing cost 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; for example the most effective traditional remedies have been studied and their active ingredients commercialised, natural products screened to identify the most potent bioactive molecules, many single gene disorders and eminently druggable oncogene-driven homogeneous tumours have been cured, leaving us to tackle the more complex diseases and pharmacological targets. This pessimism has led to the ever present idea that drug discovery is undergoing a productivity crisis,[2] and that the investments made in early stage research do not translate into actionable pharmacology which can be used to develop effective therapies for patients, 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 predefined 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]

[i] The 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 zebrafish[10] 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 perturbagen. Here I will focus on cell-based assays for small-molecule screening, though the same methods apply for most other assays (spheroids/organoids etc) and perturbagens (siRNA, CRISPR etc).

The standard approach to extracting numerical features from cell morphologies is through 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 morpholgical features from the raw images.[15,16]

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

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

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

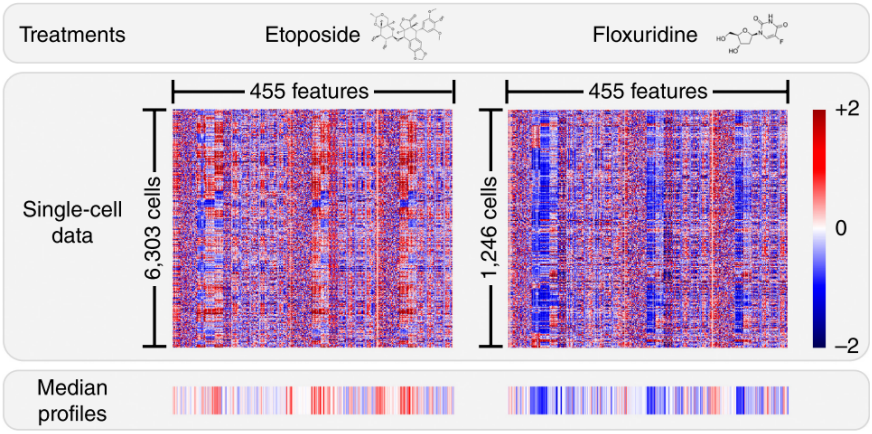


Figure 1.1: Single cell data aggregation to a median profile. Two matrices representing single cell morphology data for a treatment, with columns displaying multiple measured morphological features for each cell represented as a row. *(Figure re-used from Caicedo et al. Nat Methods, 2017)*

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 filtering[21] 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 ANOVA[23] 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 the assay has appropriate signal-to-noise attributes for testing multiple compounds. Or alternatively, a carefully designed assay in which a disease model utilising abormal patient-derived or genetially 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 technqiues.

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 trees[33] 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), these cell lines were chosen based on a number of criteria:

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

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

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.

1.4.2 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 subclasses 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 of more of these pathways is responsible for approxmiately 80-85% cases of breast cancer. The remaining 15-20% of cases are classified as triple negative (TN). Molecular subclasses 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]

1.5 Thesis structure

The following chapters focus on selected topics from my PhD, some of which has been 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 13,000 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 devloping 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 general conclusions from the my PhD and future directions.

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% CO_2.

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 μL of IncuCyte NucLight Green Lentivirus (#4624 Essen) was added to each well at an MOI of 1 with 1.5 μL 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 μg/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 successfull transduction, GFP labelled cells were maintained in 0.5 μg/mL puromycin complete DMEM.

2.2.1 Culturing cells in 96-well plates

2.2.2 Culturing cells in 384-well plates

2.3 Compound handling

2.3.1 24 compound validation set

Compounds (table 2.1) were diluted in DMSO at a stock concentration of 10 mM. Compounds plates were made in v-bottomed 96-well plates (#3363 Corning), at 1000-fold concentration in

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

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

100% DMSO by serial dilutions ranging from 10 mM to 0.3 μM in semi-log concentrations. Compounds were added to assay plates containing cells after 24 hours of incubation by first making a 1:50 dilution in media to create an intermediate plate, followed by a 1:20 dilution from intermediate plate to the assay plate, with an overall dilution of 1:1000 from the stock compound plate to the assay plate.

2.4 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 availablity of the fluorescent filter sets fitted to the microscope, reagent costs, and the scalability of the protocol when used in a large screen. Fortunately, this problem was already addressed by another group who published a protocol – named "cell painting" – for labelling 7 cellular structures, using 6 non-antibody stains imaged in the same 5 fluorescent channels available with our miscoscopy setup.[29,49]

The cell-painting protocol was initially optimised by Gustafsdottir *et al.* for use in the U2OS oesteosarcoma cell line, and briefly tested in a few other commonly used cell-lines. However, when tested on the panel of 8 breast cancer cell lines, the staining protocol was observed to induce morphological changes on certain cell lines, in the absence of compounds. It was found that changing the media, and adding the MitoTracker DeepRed stain to live MDA-MB-231 cells produced a

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

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

rounded morphology, which was not observed in the other cell lines. As any morpholgical 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 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 μL for a 96 well plate, 50 μL for a 384 well plate) and permeabilised with (50 μL 96-well, 30 μL 384-well) 0.1% Triton-X100 solution for 20 minutes at room temperature. A solution of cell painting reagents was made up in 1% bovine serum albumin (BSA) solution (see table 2.2). Cell painting solution was added to plates (30 μL 96-well, 20 μL 384-well) and left to incubate for 30 minutes at room temperature in a dark place. Plates were then washed with PBS (100 μL 96-well, 50 μL 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 equiped 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

Images were analysed using Cellprofiler v2.1.1 to extract morphological features. 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 tha dditional 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. 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

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

where μ is the mean and σ is the standard deviation.

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.

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 consisting of MCF7 breast cancer cells treated with 113 small molecules grouped into 12 mechanistic classes, these cells were then fixed, labelled and imaged in three fluorescent channels[47]. This dataset (also known as BBBC021) has become widely used as a benchmark in the field for MoA classification tasks, with multiple publications using the images to compare machine learning and data pre-processing approaches.[50,51,52,53] 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 and 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 LINCS1000[54].

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

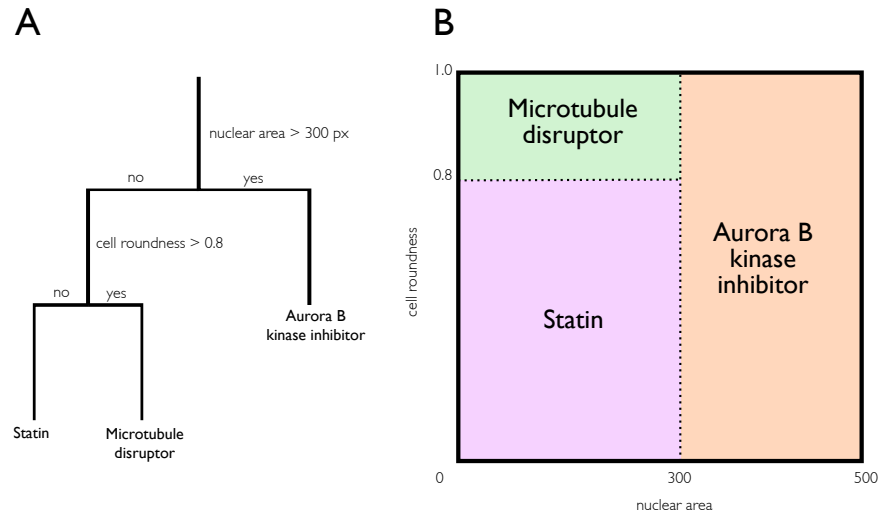


Figure 3.1: **(A)** An example of a simple mock decision tree to classify compound mechanism of action based on morphological features. **(B)** Depiction of decision space as divided by the decision tree model. Shaded areas show how new input data will be classified based on the decision rules (dotted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

the images, generating a multivariate dataset from the images, and training a classifier on these morphological features.

To develop and validate a machine learning model the dataset has to be split into training, validation and test sets. This is because overfitting is a common problem in machine learning, whereby the model is trained and accurately predicts labels on one dataset, but performs poorly when applied to new data on which is was not trained. Most 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.[55] Bagging[56] and Boosting[57] are two popular methods for constructing ensembles of decision trees. As combining the output of several decision trees is useful only if there is a disagreement among them, these two methods both attempt to solve the same problem of generating a set of correct decision tress, that still disagree with one another as much as possible on incorrect predictions.

Decision tree methods work best with multivariate tabular data, with well defined features describing each observation, this is in contrast to image data which consists of 2D arrays of pixel intensities. Therefore, in order to train such a model, cellular morphology needs to be quantified by measuring cellular features. This is a common task with multiple software packages avail-

able, which follow two main steps: (1) Segment objects from the background. Objects may be sub-cellular structures or whole-cell masks (2) Measure various attributes from the object, this is typically based on size, shape and intensity. Cellprofiler[58] was chosen primarily due to the high configurability and the permissive license enabling large-scale distributed processing on compute clusters in order to reduce the image analysis time. The images captured on the ImageXpress were analysed using Cellprofiler, quantifying approximately 400 morphological features. The datasets produced by the Cellprofiler analysis contained morphological measurements on an individual cell level. Although we can train a model on single cell data we are not interesting in classifying morphologies of single cells, but rather classifying an image or a collection of images that represent a compound treatment, this therefore allows several approaches to structuring the training data:

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

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 training[60] 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

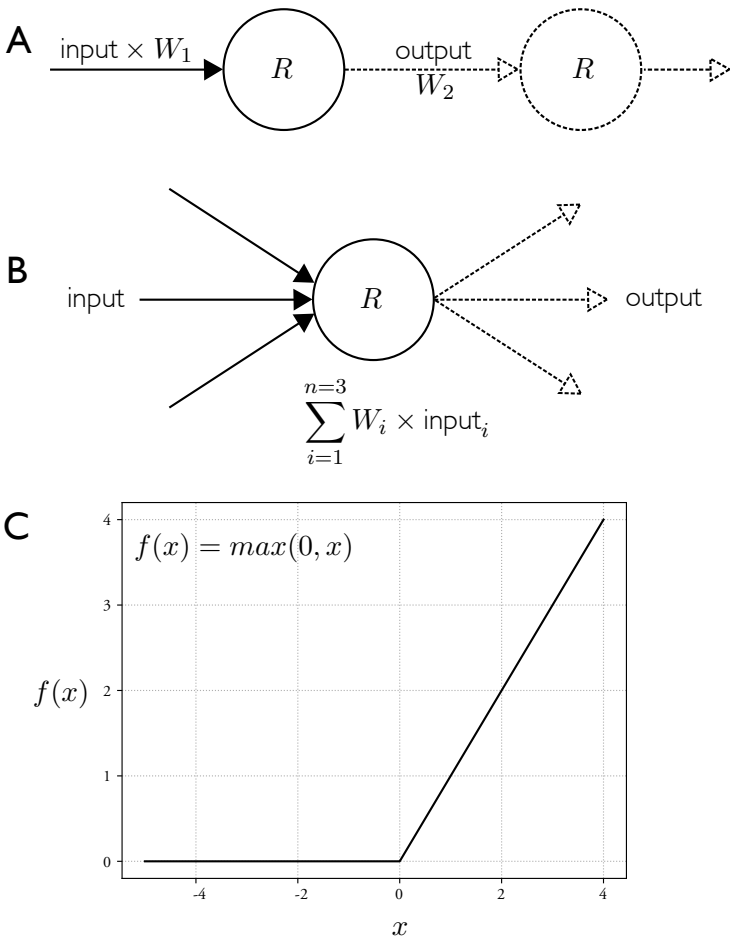


Figure 3.2: **(A)** A representation of a single connected neuron in an ANN, the input value to the neuron is multiplied by the weight (W_1), before being passed through the activation function R, the output of which is then multiplied by W_2, and passed as the input to the next neuron. **(B)** A neuron 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 when directed to all connected neurons in the next layer. Where W_i is the weight of input_i. **(C)** A common activation function also known as a rectifier, in this example a rectified linear unit (ReLU), in the inputs (x) is transformed and passed as output. So $f(x)$ can be viewed as the output for a given value of x.

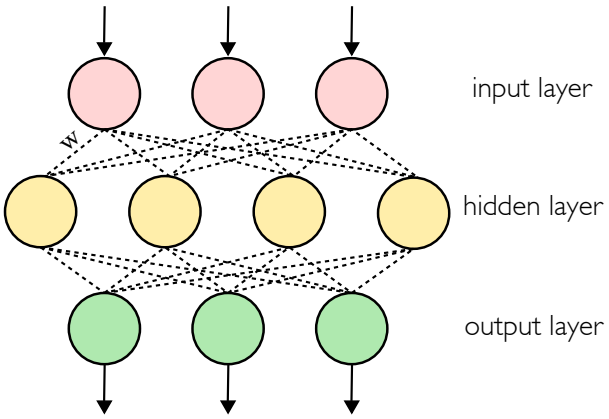


Figure 3.3 Representation of a simple 3-layer ANN with a single fully connected hidden layer, three input neurons and three output neurons. W denotes a weighted connection between an input neuron and a hidden-layer neuron, with all connections between neurons having an associated adjustable weight. A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

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 ANNs by starting with many randomly initialised kernels, and updating the kernel values through training in order to best highlight features which prove useful for accurately predicting classes. Using a single convolutional layer highlights simple features in an image such as edges and speckles, by combining several convolutional layers more complex features are highlighted through combinations of these simple features. These convolved images are then flattened into a one-dimensional vector which is used as an input in a fully connected ANN such as that depicted in figure 3.3.

3.1.4 Chapter aims

The aims of this chapter are to assess how well machine learning models to 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 consist of 24 annotated compounds with well defined MoAs.

3.2 Results

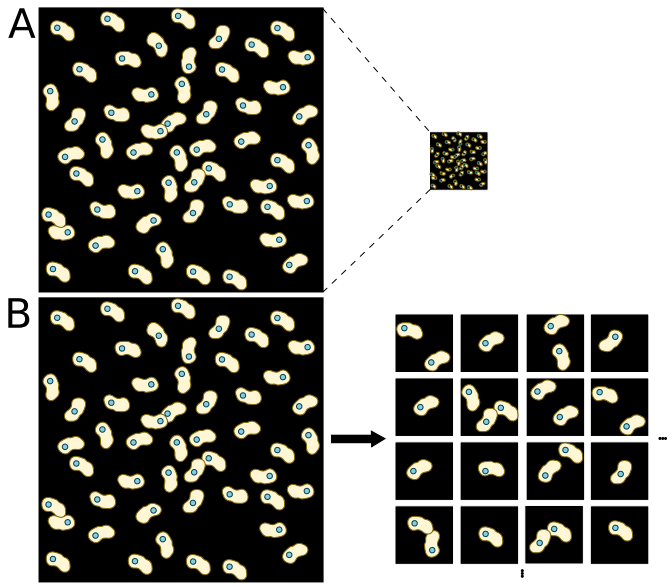


Figure 3.4: Two options for adapting large microscope images to work with the smaller input size of typical CNNs. **(A)** Full-sized images are downsized to the desired dimensions via bi-linear or bi-cubic interpolation. **(B)** Images are chopped into smaller sub-images, cell detection can be carried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

3.2.1 CNN predictions are improved using sub-images

The images generated by the ImageXpress microscope with zero binning 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 small images, one option is to downscale the entire image using bi-linear or bi-cubic interpolation, while a second option is to chop the original image up into smaller sub-images (figure 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 was found 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.

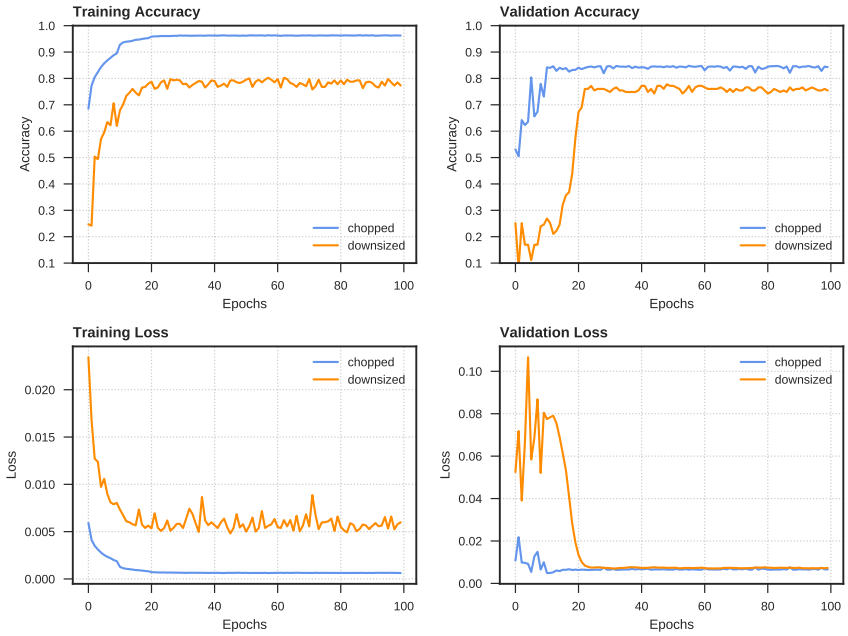


Figure 3.5: Comparison of training ResNet18 model on chopped sub-images vs downsized images from the MDA-MB-231 cell line. Chopped images were 300×300 crops centred on nuclei. Whole images were 2160×2160 images downsized to 300×300 pixels.

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 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 it should be noted that 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 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 μm^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.

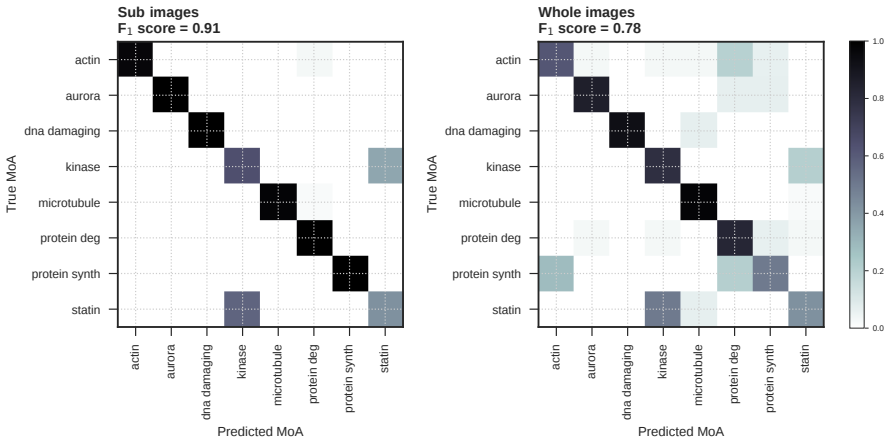


Figure 3.6: Confusion matrices comparing sub-image and whole image classification accuracy on 8 mechanistic classes of compounds.

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 architectures in the literature. 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 (ResNet18[61]) 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 AlexNet[62] 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.

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

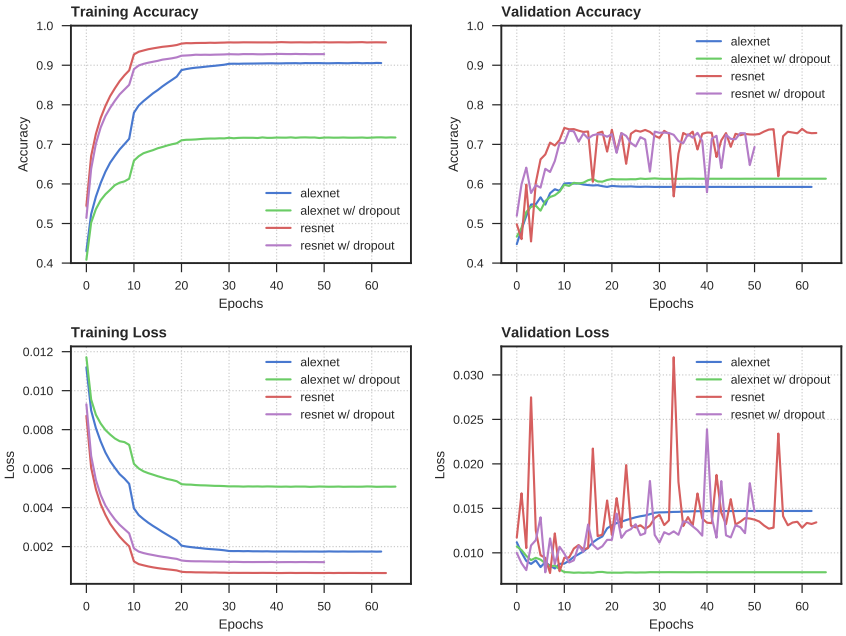


Figure 3.7: Comparison of CNN architectures. A comparison of AlexNet and ResNet18 architectures with and without dropout training and predicting on 5 channel 244×244 pixel 5 channel images of all eight cell lines. Loss was calculated using cross entropy on 8 mechanistic classes of compounds.

over-fitting 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].

[i] Although 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.

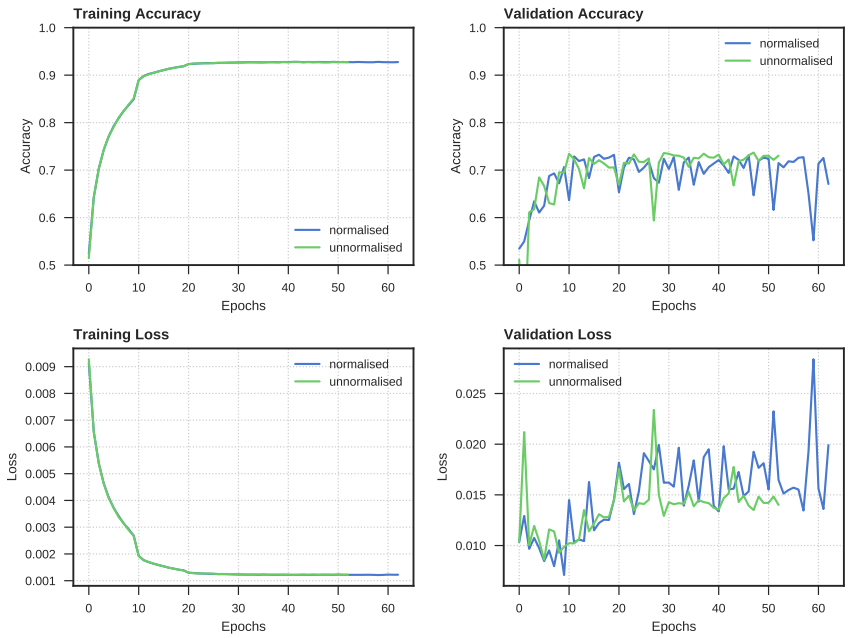


Figure 3.8: Effect of image intensity normalisation on CNN training. ResNet18 models training and predicting on eight pooled cell-lines with and without standardising image intensities per image per channel.

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

3.2.4 Single cell/Image aggregates improve classification accuracies with decision trees.

TODO, delete as appopriate when results come in.

[ii] The 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.

3.2.5 Principal component analysis does (not) improve classification accuracy with decision trees.

TODO, delete as appopriate when results come in.

3.2.6 CNN and random forest show equivalent performance at predicting MoA on a single cell-line

3.2.7 Additional data from more cell lines does necessarily improve model performance

3.2.8 On the transferrability of classifiers applied to unseen cell lines

3.3 Discussion

TODO.

3.4 Methods

3.5 Dataset

The imaging dataset used in this work is the same as in chapter 4. For each compound data were used for three concentrations (0.1 μM, 0.3 μM, 1.0 μM).

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 $\hat{y}_i$:

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

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):

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{3.2}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{3.3}$$

the F_1 score can be calculated as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{3.4}$$

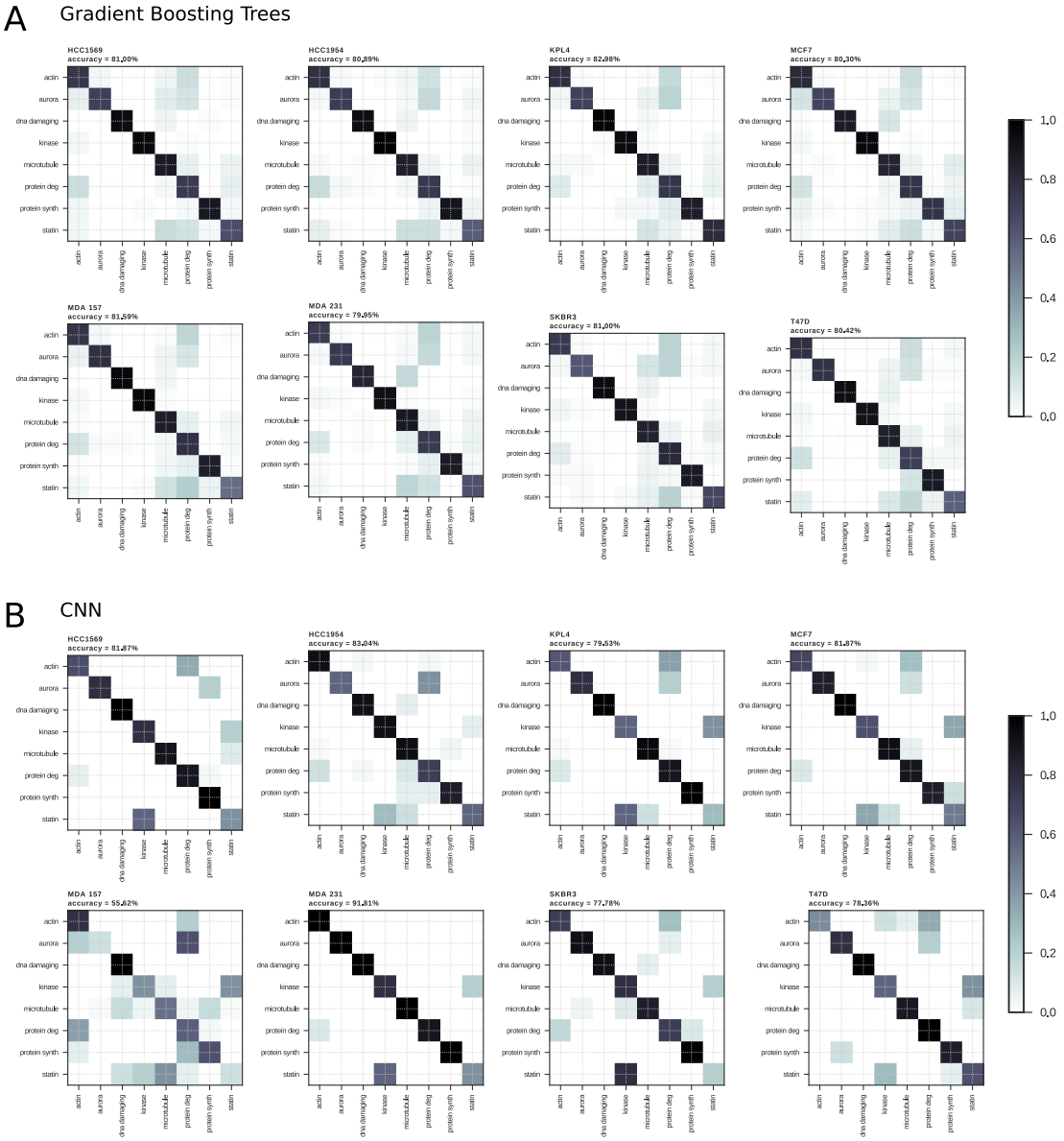


Figure 3.9: Comparison of ensemble based tree classifier and CNN at predicting compound MoA when trained on tested on an individual cell-line. **(A)** Gradient Boosting tree classifier. **(B)** ResNet18 CNN classifier. Accuracy measured as the Jaccard similarity score.

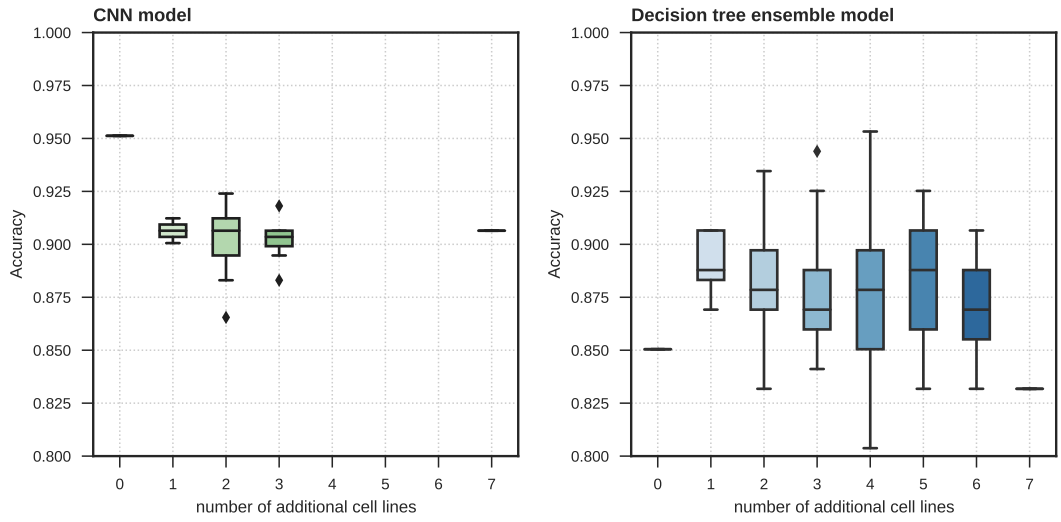


Figure 3.10: The effect of using additional cell-lines during model model training. Models accuracy when tested on a with-held proportion of MDA-MB-231 data. Box-plots show accuracy when tested on different combinations of additional cell-lines.

3.5.2 Ensemble of decision trees

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

3.5.3 Convolutional neural networks

All code related to neural networks was written in pytorch v0.3 for python 3.5, and all ANN models were trained on nvidia K80 GPUs.

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.

Architecture

Training parameters

Image intensities were standardised on an individual image and channel basis by taking each image in the form of an array [width $\times$ height $\times$ 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 64 images. 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

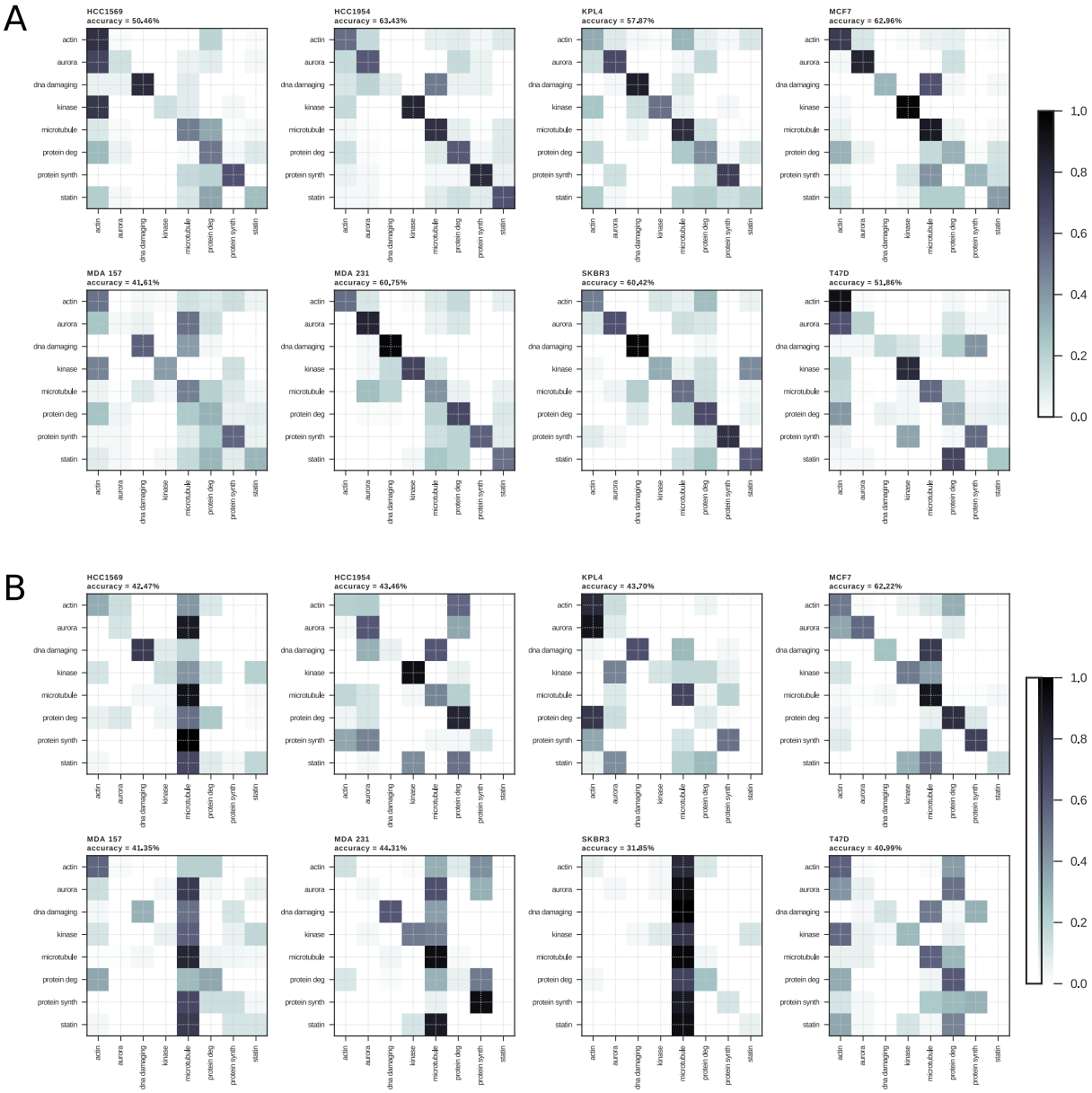


Figure 3.11: Confusion matrices of classifiers applied to unseen cell-lines. Models were trained on 7 out of the eight cell-lines and tested on the with-held cell-line (named above confusion matrix).

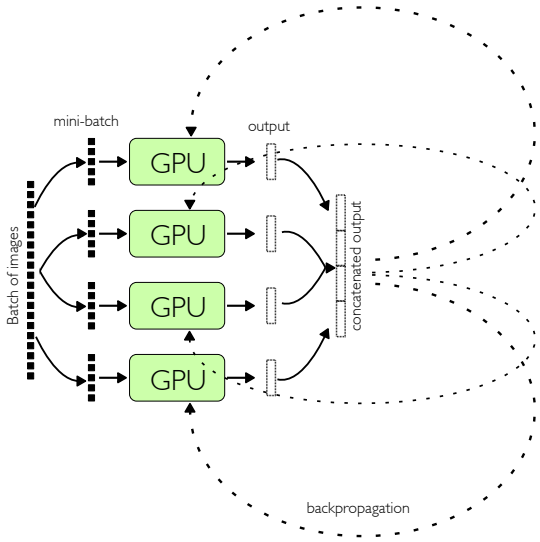


Figure 3.12: Increased training speed by data parallelism. Models are replicated across an array of GPUs, the input batch is split evenly among the devices, with each device processing a portion in parallel. During backpropagation the updated weights for all replicas are averaged and models weights are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

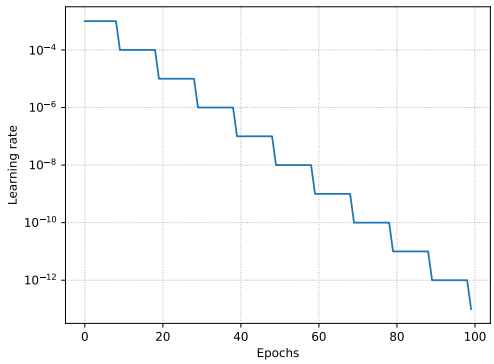


Figure 3.13: Learning rate and decay for training CNN models, initialised at $1e^{-3}$ and reduced 10-fold every 10 epochs.

10 epochs (figure 3.13). Decay was used to aid gradient descent and model convergence. Models were trained for 100 epochs or 48 hours, whichever was reached first, with model checkpoints every epoch there was an increase maximum validation accuracy. The optimiser used was ADAM[63] with the categorical cross entropy loss function.

Image preparation

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.[64] 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 EC_{50} 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.

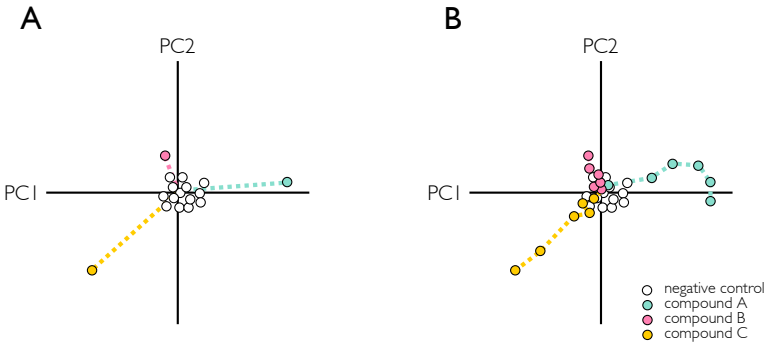


Figure 4.1: Diagram illustrating measuring magnitude of compound response by distance from the negative control centroid in principal component space. **(A)** Phenotypic distance to three different compounds. Compound A and B show phenotypic activity as they are distanced from the negative control cluster, 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 C 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

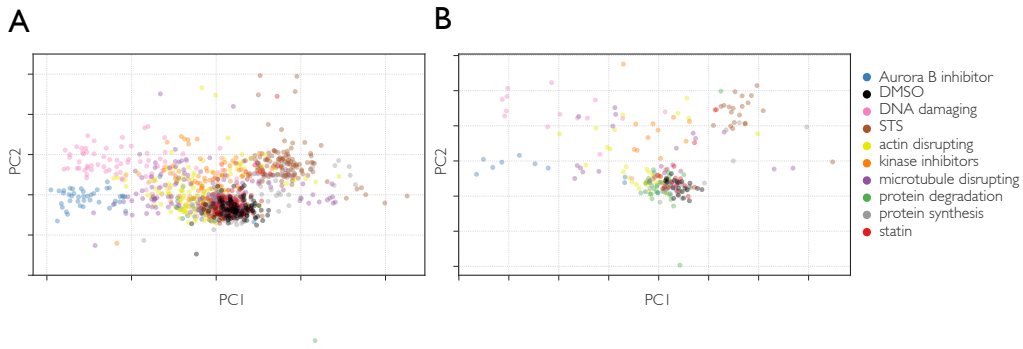


Figure 4.2: MoA clustering of compounds based on PCA of their morphological features. Principal components calculated from morphological features of 24 compounds grouped into 8 mechanistic classes. **(A)** Principal components calculated from an image average of individual cell measurements. **(B)**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

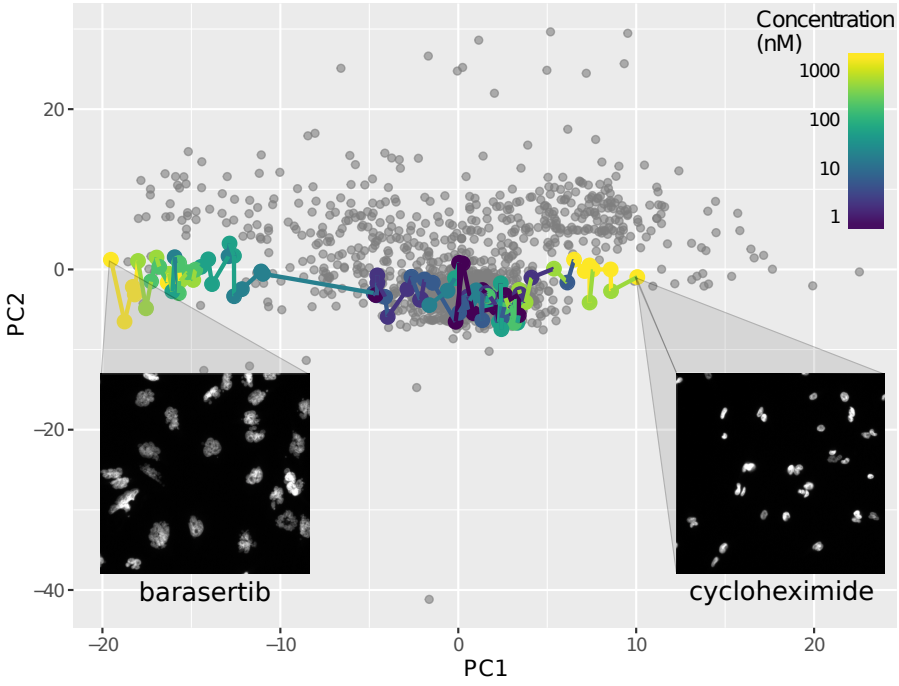


Figure 4.3: Principal components of Cellprofiler features calculated from a 24 compound high content screen in MDA-MB-231 cells. Barasertib (left) and cycloheximide (right) titrations are highlighted to show two active compounds with distinct phenotypes heading in different directions in phenotypic space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

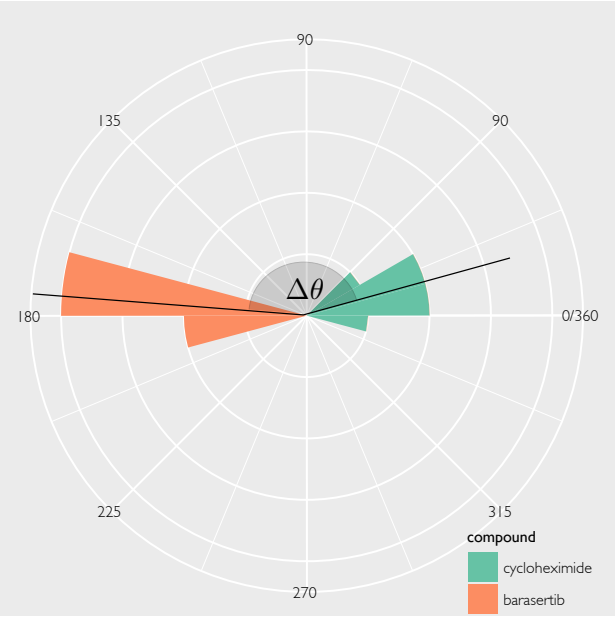


Figure 4.4: Visualisation of $\Delta\theta$ to quantify the difference in phenotypic direction between two compounds. Histograms in polar co-ordinates show the θ values of treatments against a fixed reference vector, with $\Delta\theta$ calculated as the difference between the average θ (black lines) 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 position 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 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 μM, 0.3 μM, 1 μM). 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

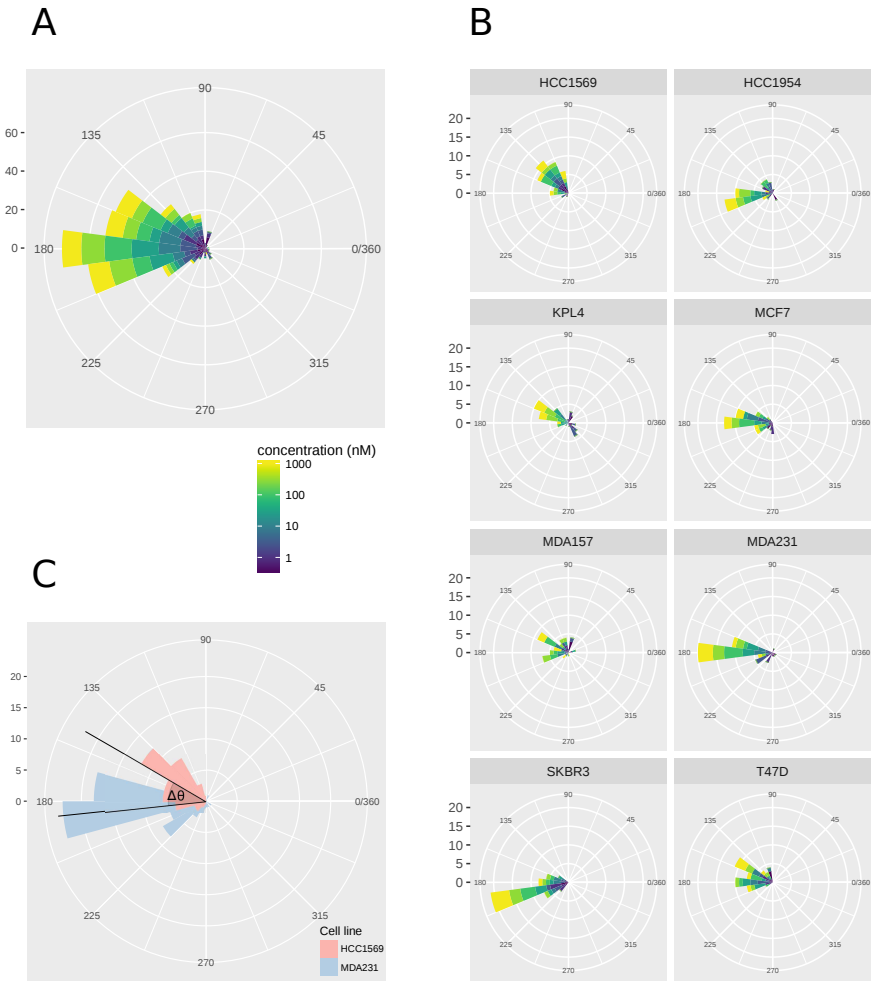


Figure 4.5: Visualisation of $\Delta\theta$ to quantify the difference in phenotypic response between cell lines when treated with barasertib. **(A)** Circular histogram of θ values of barasertib calculated for eight cell lines. **(B)** Phenotypic direction of cell lines treated with barasertib stratified 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.

of the calculated $\Delta\theta$ 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.

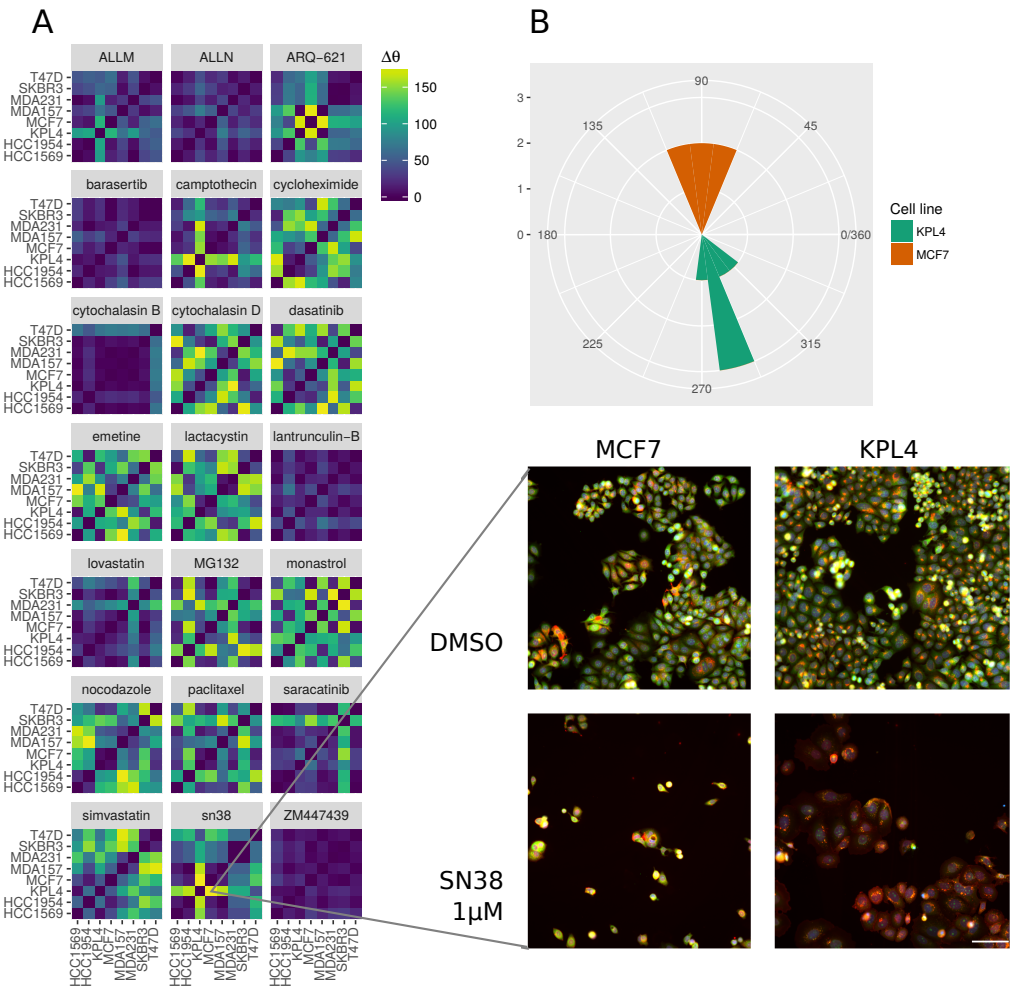


Figure 4.6: Heatmap of $\Delta\theta$ values between pairs of cell lines for separate compounds. **(A)** $\Delta\theta$ calculated between pairs of cell lines treated with 21 compounds at (0.1 μM, 0.3 μM, 1 μM). Images show differential response between KPL4 and MCF7 cell lines treated with 1 μM SN38. 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 μm. **(B)** Histogram of θ values calculated for MCF7 and KPL4 cells treated with 1 μM SN38. $\Delta\theta = 179°$

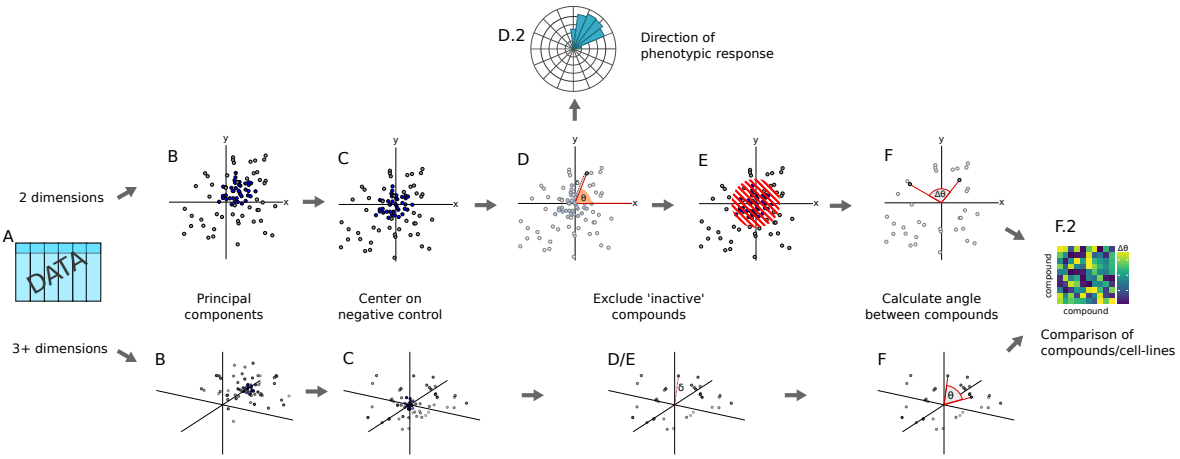


Figure 4.7: Theta comparative cell scoring (TCCS) workflow. **(A)** Normalised and standardised numerical data. **(B)** Principal component analysis, negative control values coloured in blue. **(C)** Centering of principal component values to the negative control centroid. **(D)** Calculation of distance from the origin to each data point, an activity cutoff is derived fro 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:

```
threshold = 0.8
pca_output = prcomp(data)
pc_variance = pca_output$stdev^2
cumulative_prop_variance = cumsum(pc_variance) / sum(pc_variance)
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 columns 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:

```
mediods = apply(data[data[, "compound_name"] == "DMSO"], 2, median)
delta_m = 0 - mediods # δm
for (i in seq_along(feature_cols)) {
    feature = feature_cols[i]
    # feature_column_i := feature_column_i + δm_i
    data[, feature] = data[, feature] + delta_m[i]
}
```

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 μM 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.

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

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:

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

5 LARGE COMPOUND SCREEN ACROSS 8 BREAST CANCER CELL LINES

5.1 Introduction

5.1.1 subsection

5.2 Results

5.2.1 Hit selection

5.2.2 Serotonin related compounds

5.2.3 Spheroids

5.2.4 RPPA

5.3 Methods

5.3.1 Identifying hits

5.3.2 Spheroids

Creating spheroids

Spheroids were created by seeding approxmiately 10,000 GFP-expressing cells per well in 50 μL of media into each well of a 96-well 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 μL per well was added to the existing cell suspension, for a final Matrigel concentraion of 2% and 1% DRAQ7. Plates were then centrifuged for 10 minutes at 1000X G and 4° with break 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 μL from the source plate to an intermediate plate containing 150 μL of media. From the intermediate plate 5 μL were transferred to the spheroid assay plate containing 100 μL 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 μm intervals for a total range of 350 μm. 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.

5.3.3 Western blotting for SERT and TPH1

Protein extraction

Protein extraction from cells. Cells were grown for protein extraction in 6-well plates and lysed when roughly 80% confluent. RIPA buffer was made from 25 mM Tris (pH 7.5), 150 mM NaCl, 0.1% SDS, 1% Triton-X100, 0.5% deoxycholate and phosphatase inhibitors (cOmplete ULTRA #05892970001 Roche), and placed on ice. Media was removed from the 6-well plates with a pipette, and wells were washed with cold PBS before addition of 300 μL of cold RIPA lysis solution. Cells were incubated in lysis buffer for 2-3 minutes on ice and scraped with a cell-scraper. The lysis buffer/cell-lysate was then transferred to a 1.5 mL eppendorf tube and incubated on ice for 10 minutes. Cell lysate solutions were centrifuged at 13,000x G for 10 minutes at 4 °C and the supernatent transferred to new 1.5 mL eppendorf tubes.

Protein extraction from mouse brain. Two grain of rice sized pieces were cut from different regions of a frozen FVB adult mouse brain and separately homogenised (MP lysing matrix type D 1.4mm sphere) for 20 seconds at 4 M/s (MP FastPrep-24). Once homogenised, 500 μL of RIPA lysis buffer was added to each tube and placed on a rotator for 15 minutes at 4°C. The homogenising tubes and lysis solution were then briefly centrifuged to settle the beads and the lysates extracted with a pipette into pre-chilled 1.5 mL eppendorf tubes which were centrifuged at 13,000x G for 10 minutes at 4°C and the supernatents were transferred to new 1.5 mL tubes.

Standardising protein concentrations. Protein concentrations were measured using Precision-Red protein assay reagent (#ADV02-A Cytoskeleton, Inc.). For each sample, 10 μL of lysate was added to a plastic cuvette followed by 1 mL of Precision-Red. The absorbance was measured after 1 minute incubation at room temperature with a spectrometer at 500 nm and compared to a blank of 1 mL Precision-Red and no lysate. Protein concentration in mg/mL was determined as 100 times the corrected absorbance value at 600 nm, as per the suppliers instructions. Sample protein concentrations were all made to 2.5 mg/mL by diluting in RIPA lysis buffer. Samples were then prepared for Western blot by mixing 75 μL of sample with 25 μL of 4x loading buffer (200 mM Tris, 400 mM DTT, 8% SDS, 0.4% bromophenol blue, 40% glycerol) denaturing at 90°C for 10 minutes.

Western blotting

Gel eletrophoresis and transfer. Using pre-cast 4-15% gradient gels (#4568085 Bio-Rad) and SDS-page running buffer (20 mM Tris, 195 mM Glycine, 0.1% SDS, pH 8.5), 7.5 μL of seeBlue Plus2 ladder (#LC5925 Invitrogen) and 20 μL of sample (50 μg) were added to separate lanes, and the gel was run at 200V constant voltage for 30 minutes. Proteins were transferred to PVDF (polyvinylidene difluoride) membranes (P0.45 Aversham hybond) pre-wetted in in MeOH sandwiched with blotting paper at 400 mA constant ampage for 60 minutes in transfer buffer (SDS-page running buffer with 20% MeOH) with ice packs.

Antibody staining and visualising. Following transfer, membranes were blocked in 20 mL of 1% casein blocking solution (#11921681001 Roche) for 10 minutes on a rocking plate. Antibody solutions were made up in the same blocking buffer, TPH1 (#LS-C117936 LSBio) was diluted 1:500, SERT (#AMT-004 Amalone) was diluted 1:500. Membranes were placed in separate 50 mL Falcon tubes with the protein side facing inwards in 7 mL of antibody solution and placed on rollers overnight at 4°C. The membranes were then removed from the antibody solution was washed 3 times with 10 mL of TBS-T (50 mM Tris, 150 mM NaCl, 0.1% Tween 20, pH 7.5) for 5 minutes. The secondary antibody (HRP-linked anti-rabbit IGg #70742 CST) was diluted 1:5000 in blocking buffer and used to wash the membranes for 1 hour at room temperature. Membranes were then washed 3 times with 10 mL TBS-T for 5 minutes followed by 3 washes with 10 mL TBS (50 mM Tris, 150 mM NaCl, pH 7.5) for 5 minutes. Imaging of the membrane was carried out using BM chemiluminescence blotting substrate (#11500619001 Roche) following the manufacturers instructions.

TODO:housekeeper control, imaging

5.3.4 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 μL 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 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 μL 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 5.1). Samples were diluted 1:20 in PBS by adding 2.5 μL of sample to 47.5 μL of PBS and mixed with a vortex. 10 μL of diluted samples and standard were added to each well of a flat-bottomed 96-well plate, followed by 240 μL 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 (μL)	PBS (μL)	Lysis Buffer (μL)
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 5.1: Volumes for the BSA standard curve. Lysis buffer was CLB1, the same as used for the sample preparation.

6 CHEMINFORMATICS AND HIGH-CONTENT IMAGING

6.1 Introduction

6.1.1 Cheminformatics

The term "cheminformatics" was first coined in 1998[65] 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.[66] 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[67]. Since then the QSAR field has advanced to include many more parameters and in 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 endevour, 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 a 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 covering 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.[68] 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".[69]

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.

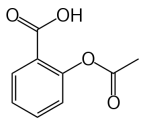
Relating changes to a compound's structure to biological activity is relatively straightforward if compound activity can be represented as a single variable such as binding affinity or EC_{50}, applying quantitative SAR (QSAR) to multiparametric data such as that found with high-content imaging is not 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 sterochemistry (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.[70] However, these naive methods suffer from a number of drawbacks, mainly stemming from the ambiquity 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)[71] which are abstract representations of molecules in the form of fixed-length binary arrays – generated from local patterns in the molecule such as the iden-

A



B

```
CC(=O)OC1=CC=CC=C1C(=O)O
```

C

```
InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
```

D

```
13 13  0  0  0  0            999 V2000
   -0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1433    0.4126    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    1.6501    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.0002    1.6500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    2.0625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4290    1.6500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    2.8875    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
 2  1  1  0  0  0  0
 3  1  2  0  0  0  0
 4  7  1  0  0  0  0
 5  3  1  0  0  0  0
 4  5  2  0  0  0  0
10  5  1  0  0  0  0
 6  4  1  0  0  0  0
 2  6  2  0  0  0  0
 8  7  1  0  0  0  0
 9  7  2  0  0  0  0
11 10  1  0  0  0  0
12 11  1  0  0  0  0
13 11  2  0  0  0  0
M  END
```

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.

tity 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 distance[i] (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.[72] Recently a number of studies have leveraged advances in the machine learning field to generate alternative chemical fingerprints using neural networks.[73,74,75,76] 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

[i] Whilst 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 consine 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.*[77] incorportated 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 morphological screen and selected a subset of compounds which produced a diverse range of bioactivities. 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.[78]

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 Lipinksi'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.

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.

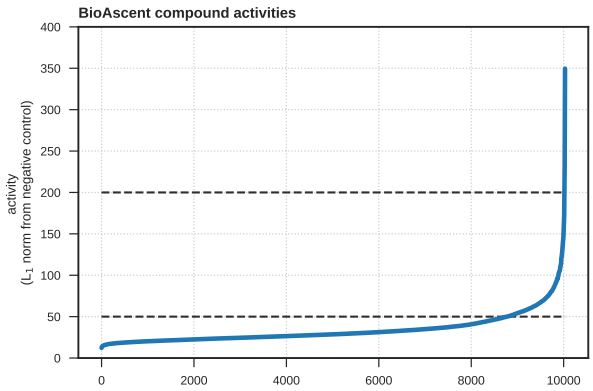


Figure 6.2: Selection of active BioAscent compounds based on the l_1 norm distance from the DMSO negative control centroid in PCA space. Lower and upper bounds of the selected compounds are indicated by dashed lines. In total 1244 compounds were selected.

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 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 cytoxic 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 algorithm[79] 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 idenfified by hierarchical clustering they appeared to match up with the tSNE 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 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 agglomative 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 phenotypically

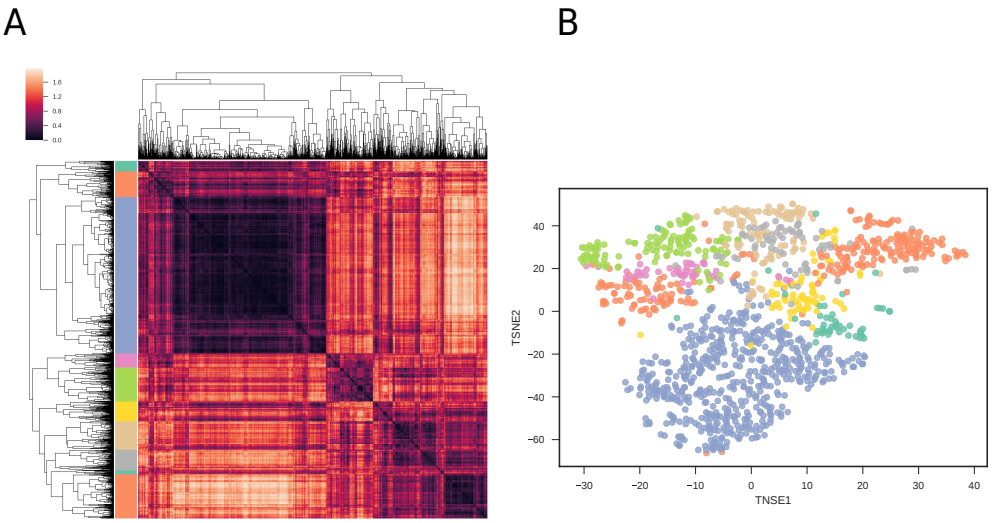


Figure 6.3: Morphological clustering of active compounds within the BioAscent library. **(A)** Hierarchical clustering of the 1244 active BioAscent compounds based on a distance matrix of principal components. Clusters formed by cutting the produced dendrogram. **(B)** Unsupervised t-SNE clustering of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

active molecules. The use of more novel compound fingerprinting techniques such as USRCAT[72] and autoencoded features[80] 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[81]. 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 activiting 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

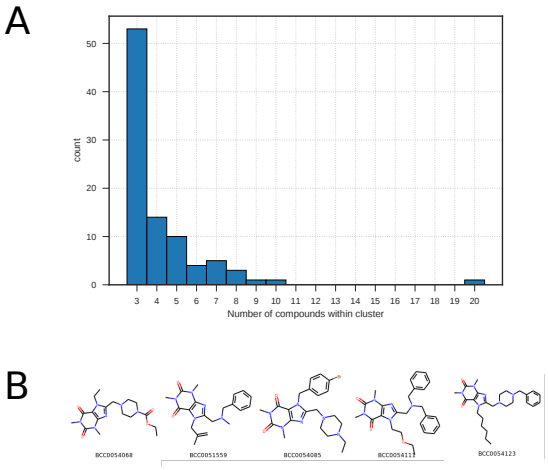


Figure 6.4: **(A)** Histogram of number of compounds within structurally similar clusters, with most clusters only containing 3 molecules. **(B)** An example of one of the structurally similar clusters as found with the Butina clustering algorithm.

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 two 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[82], 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 availble with the BioAscent library is through querying publically 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/IC_{50} 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/IC_{50} values. This returns a list of Uniprot accession codes which were used with interpro[83] 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

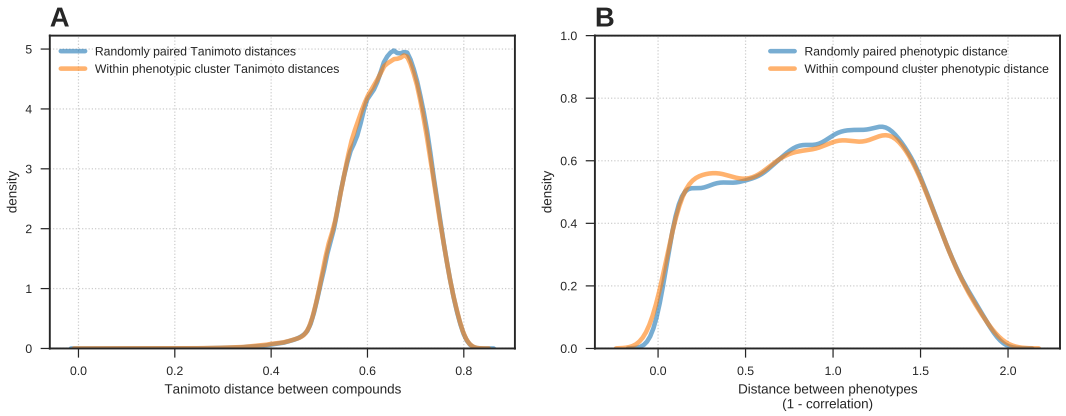


Figure 6.5: (A) Tanimoto distance between compounds from within phenotypic clusters (as found in fig. 6.3) and between randomly paired active compounds. ($p = 1.81 \times 10^{-15}$, $D = 0.011$, 2-sample Kolmogorov-Smirnov test) **(B)** Phenotypic distance between compounds from within structurally similar 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.[69] One way to incoporate 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 undersireable 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),[84] 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

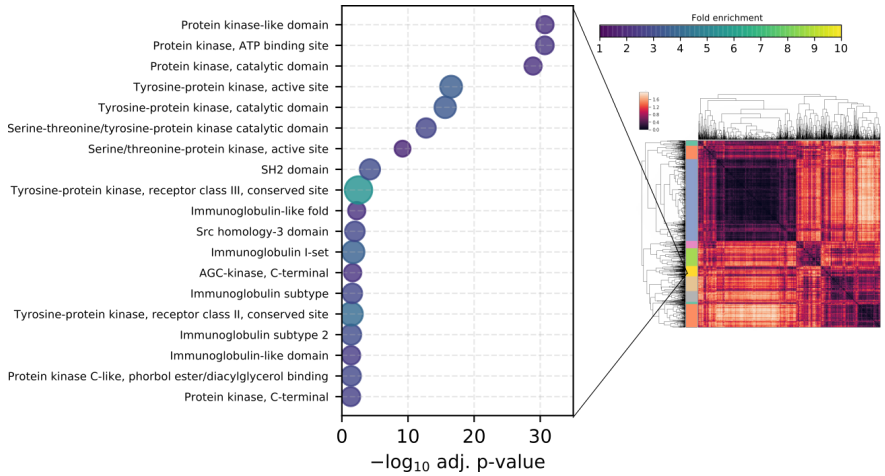


Figure 6.6: Enriched interpro targets found within a phenotypic cluster of the BioAscent library when compared to a background of all active BioAscent compounds.

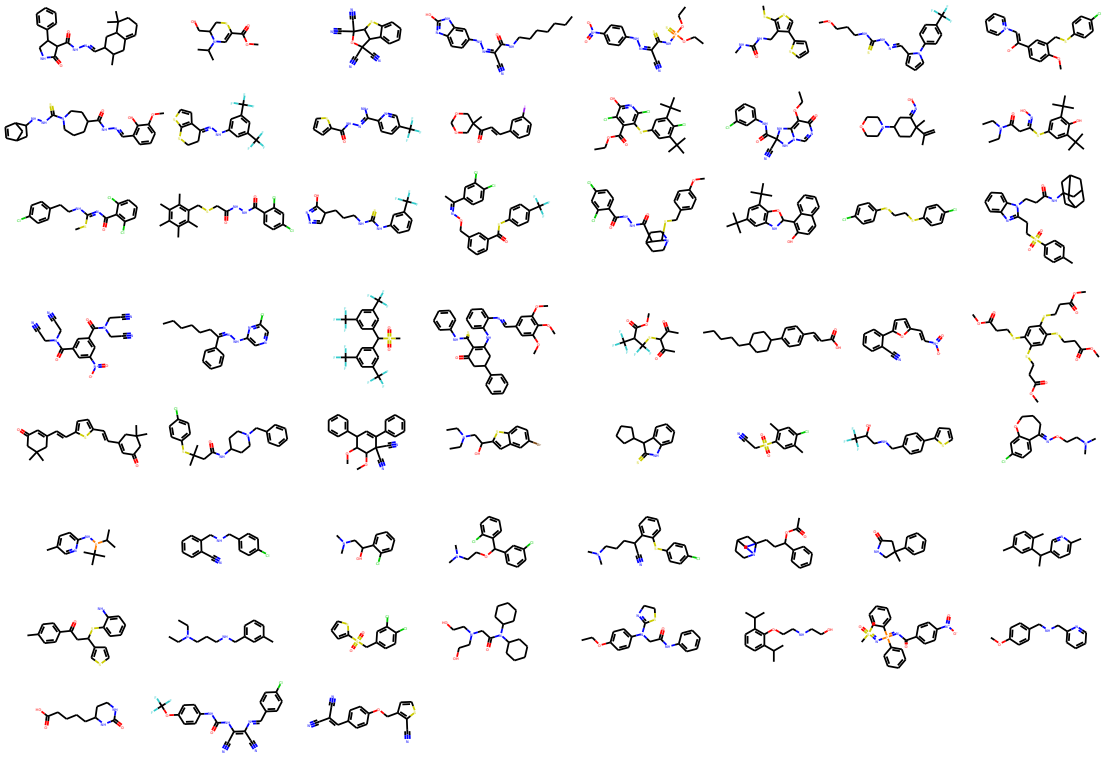


Figure 6.7: 59 phenotypically active BioAscent compounds with no close structural analogues in the ChEMBL database.

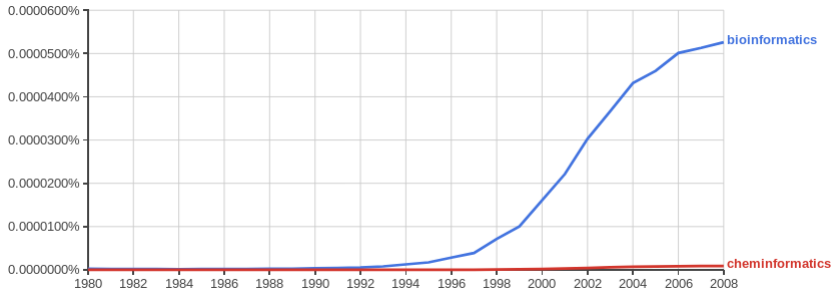


Figure 6.8: Popularity of the terms 'bioinformatics' and 'cheminformatics' in the published literature as found using Google's Ngram viewer between 1980 and 2010. y-axis represents the cumulative percentage of literature containing the term, x-axis represents the year.

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 suprising 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 structually 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 structually similar clusters as found with the Butina algorithm were more phenotypically similar (figure 6.5), despite statisical 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 biding of a small molecule to a target receptor, althought 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 contructed from many hetergeneous 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 heterogenenous 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.[69] 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 where structurally distinct from any listed in the ChEMBL records (figure 6.7). There is also the possibiliy 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 exlusion 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 dataset [ii], based on the work published by Gomez-Bombarelli *et al.*[80] 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 Cellprofiler.

[ii] www.github.com/cxhernandez/molencoder

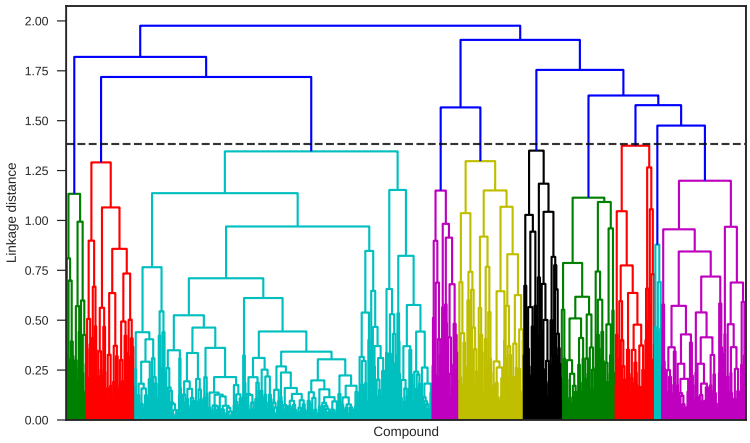


Figure 6.9: Dendrogram thresholding to determine the number of phenotypic clusters in the active BioAscent compounds. Dashed line indicates cutoff of 70% of the maximum linkage distance, resulting in 10 clusters.

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/IC_{50} value less than 1 μM. 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's 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.[85] 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 interprot 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 DISCUSSION AND CONCLUSION

7.1 Section name

BIBLIOGRAPHY

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

[2] Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. "The productivity crisis in pharmaceutical R&D." *Nature reviews. Drug discovery* 10.6 (2011), pp. 428–38.

[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.Database issue (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. arXiv: 15334406.

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

[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* 8.11 (2016), pp. 1133–1144.

[9] Y. Tabata, N. Murai, T. Sasaki, S. Taniguchi, S. Suzuki, K. Yamazaki, and M. 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, Swetlana Friedel, Benjamin Kilian, Ming Chen, Thomas Altmann, and Christian Klukas. "Dissecting the Phenotypic Components of Crop Plant Growth and Drought Responses Based on High-Throughput Image Analysis". *The Plant Cell Online* 26.12 (2014), pp. 4636–4655.

[12] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". *IEEE Transactions on systems, man and cybernetics* 20.1 (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, K Ullrich, and Fred a Hamprecht. "ILASTIK : Interactive learning and segmentation toolkit". *Eighth IEEE International Symposium on Biomedical Imaging (ISBI)* 1 (2011), pp. 230–233.

[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. 635–7.

[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. arXiv: NIHMS150003.

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

[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: Identifying Density-Based Local Outliers". *Proceedings of the 2000 Acm Sigmod International Conference on Management of Data* (2000), pp. 1–12.

[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 Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

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

[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. arXiv: 15334406.

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

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

[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), p. 150810081821009.

[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. 507–19.

[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 Supp". *Nature* 483.7391 (2012), pp. 603–7. arXiv: NIHMS150003.

[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. arXiv: NIHMS150003.

[42] Wu Lin, Smythe Anne, Stinson Sherman, Mullendore Leslie, Anne Monks, Dominic Scudiero, Kenneth Paull, Kuotsoukis Antonis, Lawrence Rubinstein, Michael Boyd, and Robert Shoemaker. "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, 2016, pp. 3029–3034.

[43] R H Shoemaker. "The NCI60 human tumour cell line anticancer drug screen". *Nature Rev.* 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. arXiv: 15334406.

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

[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 Methods* 11.9 (2016), pp. 1757–1774.

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

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

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

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

[54] Qiaonan Duan, Corey Flynn, Mario Niepel, Marc Hafner, Jeremy L Muhlich, F Fernandez, Andrew D Rouillard, Christopher M Tan, Edward Y Chen, R Golub, Peter K Sorger, Aravind Subramanian, and Avi Ma. "LINCS Canvas Browser : interactive web app to query , browse and interrogate LINCS L1000 gene expression signatures". *Nucleic acids research* 42.W1 (2014), W449–W460.

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

[56] Leo Breiman. "Bagging predictors". *Machine Learning* 24.2 (1996), pp. 123–140.

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

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

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

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

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

[63] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization" (2014), pp. 1–15. arXiv: 1412.6980.

[64] 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.

[65] Frank K. Brown. "Chemoinformatics: What is it and How does it Impact Drug Discovery." *Annual Reports in Medicinal Chemistry* 33.C (1998), pp. 375–384.

[66] Louis C Ray and Russell A Kirsch. "Finding Chemical Records by Digital Computers". *Science* 126.3278 (1957), pp. 814–819.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] David Rogers and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of chemical information and modeling* 50.5 (2010), pp. 742–54.

[72] Adrian M. Schreyer and Tom Blundell. "USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints". *Journal of Cheminformatics* 4.11 (2012), p. 1.

[73] 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.

[74] 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.

[75] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. "Drug Similarity Integration Through Attentive Multi-view Graph Auto-Encoders". *ArXiv* (2018). arXiv: 1804.10850.

[76] Shengchao Liu, Thevaa Chandereng, and Yingyu Liang. "N-Gram Graph, A Novel Molecule Representation". *ArXiv* (2018). arXiv: 1806.09206.

[77] 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.

[78] 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.

[79] Laurens Van Der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[80] 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" (2016), pp. 1–28. arXiv: 1610.02415.

[81] 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.

[82] N Mantel. "The detection of disease clustering and a generalized regression approach." *Cancer research* 27.2 (1967), pp. 209–20. arXiv: arXiv:1011.1669v3.

[83] 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.

[84] 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.

[85] 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. arXiv: 9411012 [chao-dyn].