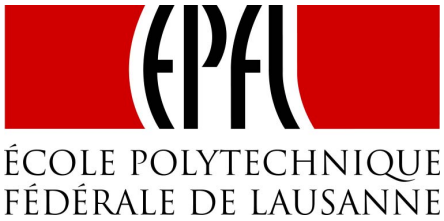
ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE SCHOOL OF LIFE SCIENCES



Master project in Bioengineering

Minor in Biocomputing

Optimizing the Image-Based Cell Profiling Pipeline

Carried out in the Imaging Plateform

at the Broad Institute of MIT and Harvard, Cambridge, MA, USA

Under the supervision of Dr. Anne E. Carpenter, Director
and Dr. Mohammad Rohban, Postdoctoral Researcher

Done by

MARIE DUC

Under the direction of

Prof. Bart Deplancke

In the Laboratory of Systems Biology and Genetics (LSGB)

EPFL

Lausanne, August 17, 2017

Abstract

Accurately capturing the effects of various small molecules (including drugs and potential drugs) on populations of cells is a critical step in image-based cell profiling, a recently developed approach which underlies a number of important applications in drug discovery and functional genomics. Cells are living beings and thus each of them will react differently to a similar environment. However, one may expect a globally similar phenotype for cells confronted with the same surrounding. In this project, we explore images of cancer cells reacting to some treatments (compounds) in a high-throughput assay. The aim is to improve the morphological profiling pipeline steps that convert cell images into a signature for each treatment. This analysis can be used as one of the first steps toward drug discovery in order to select potential new drugs for clinical tests. Thereby, we first discuss the problem of data quality and what tests can be used to assess whether the data is suitable for study. We also compare different methods to select appropriate features, to select compounds showing a strong phenotype, and to evaluate whether the profiles having the same known Mechanism of Action (MOA) are indeed more similar. We found out that an unsupervised feature selection based on SVD-entropy, combined with a similarity metric based on Jaccard distance for selecting the hits, improves compounds' overall signature specificity by 30% compared to the original pipeline. To evaluate this specificity, we defined a metric called the enrichment ratio. It uses the ground-truth MOA information in order to evaluate the global accuracy of the signatures. We additionally defined a visualization method called the "waterfall plot", that aids in understanding the morphological similarities existing for each class of drugs and can help to determine the optimal dose to be used for each drug.

Acknowledgements

First of all I really want to thank Dr. Anne Carpenter, director of the Imaging Platform. I am thankful for having the opportunity to work in a great place like the Broad Institute. She always tooks the time to guide my research and give me advice.

I would like to express my sincere gratitude to my supervisor Dr. Mohammad Hossein Rohban, without whom this project would have been impossible. He was always there to help me and guide me. He was always explaining me the theory and helping me to pursue my research.

I also want to acknowledge everybody from the Imaging Platform, who were always there to discuss and help me if needed. I was glad to see how a laboratory can act more as a team than as a sum of individuals.

Finally, I want to thank my family and friends who have always been there for me throughout all my studies and continuously encouraging me. This accomplishment would not have been possible without them. Thank you.

Marie

Contents

Chapter 1

Introduction

1.1 Overview

Drug discovery has been in free fall, even though related technologies are getting more efficient every year. The process to discover a drug is very long and expensive. It can take up to 15 years and costs around four billion dollars [12] [19]. One promising new strategy to identify potential drugs uses Cell Painting assays. A motivation towards using Cell Painting (i.e., morphological profiling) as a high-throughput assay is to decrease both the cost and the time of the first step before clinical tests. Indeed, it could reduce the costs to around tenfold [24].

High-Throughput assays allow to rapidly and automatically test a large number of compounds in an assay [2], without the need for prior knowledge on how the compounds act. The goal of high-throughput assays is to identify which compounds have useful properties in the assay, where each assay has been designed to reveal drugs that might be effective against a particular disease. However, it is complicated to optimize the experimental parameters of a compound without knowing its "mechanism of action", in other words, the precise mechanism by which the compound is affecting cells [30].

Profiling techniques such as Cell Painting can have many useful applications and it is important to consider the difference between profiling and screening [3]. Profiling aims at measuring as many properties as possible in the cells without any constraints, as compared to screening, which targets a single cell function or a single biological process and then measures a few known properties.

The idea behind morphological profiling is to treat cells with different compounds, and measure the induced phenotypic changes [21]. Using fluorescence imaging, a large number of morphological properties (such as intensity, shape, texture, size, etc) can be extracted at the cell level. Then, using image-based profiling, we can generate a signature for each compound which is a summary of single cell measurements. The signature of a compound aims to give a global description of the cells treated with the compound [1]. The general pipeline to convert cell images into signatures can be seen in Figure 1.1.

To understand this pipeline, we need to distinguish per cell analysis from profile analysis. The original measurements are made for each cell, yielding per cell level data. Each well of a plate normally contains a few hundred cells. The profiles are then obtained for each well by averaging measurements across the cells in a well. In each experiment, there are usually many replicates

for each compound. This is important in order to assess replicate correlation, i.e. whether replicate profiles are showing similar patterns or not. Then, in order to have a single signature for each treatment, we can average the profiles of a compound over the replicates.

Finally, by clustering the morphological profiles, we can determine which molecules are acting similarly on the cells, i.e. yielding a similar phenotype. These molecules can be linked with each other based on their signatures and potentially lead to the discovery of new drugs. Drugs that can induce the same phenotype in the cells likely share the same mechanism of action, although it should be noted that compounds often have unexpected off-target effects that can make such similarities less clear [1].

Because the majority of the data typically do not contain any annotation in this application (i.e. we have no information on which data should be clustered together), we use unsupervised techniques to analyze them. In unsupervised learning, the aim is to use the natural structure of the data in order to identify common patterns and to group data into clusters. Once accomplished, however, the main difficulty is to validate the clustering. There is often no conclusive way to know whether the clusters are correct, or have a real biological meaning. In some cases, however, as in this project, we do have some amount of ground truth, albeit imperfect, which we can use to test the hypothesis that the output has learned something from the data.

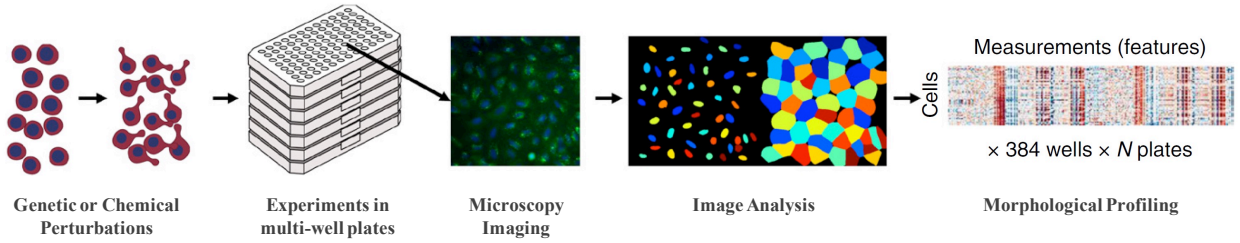


Figure 1.1: Pipeline of image-based profiling. Image obtained from [1].

1.2 Project

This project is performed at the Imaging Platform (Carpenter Lab) at the Broad Institute of Harvard and MIT. Different datasets are used to compare and determine the accuracy of the methods. Some of the raw data cannot be shared because of the intellectual property rights of the original creators, which in some cases are corporate entities, namely pharmaceutical companies.

The project is divided into many small parts, but the global aim is to improve the accuracy of the profiling analysis. At the beginning we explored the problem of the *quality of the data*, where we defined some tests that should be done prior to the analysis to be sure the data is usable. Indeed the pipeline includes a lot of processing, and with many people involved, thus increasing of the risk of problems or errors in the data. Then we tried to compare and determine the best *evaluation metric* for profiling. We also tried to find the optimal *visualization method* to evaluate the accuracy of the results. Having an unsupervised problem, we needed to find an actionable method (meaning it should be biologically relevant and useful). Having a pipeline that goes from making a profile to evaluating the similarity between the profiles, the final step was to focus on *feature selection*. We tested whether it could improve the pipeline or not and which kind of unsupervised feature selection was appropriate for our problem.

This project is separated into four chapters, including this introduction. In chapter 2, we introduce the materials and the methods. We explain the data we were working with and the computational tools we are using. We also explain the profiling pipeline and the optimization steps that we implemented. Chapter 3 presents the results of the experiments performed, followed by the last chapter 4 that discusses the outcomes of this project and what alternative analysis and additional experiments could be done in order to further improve the profiling pipeline.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Data

Multiple datasets were used to test various alternative choices in the standard profiling pipeline and to assess the significance of the results. They were created using high-throughput assays according to an established protocol called Cell Painting [1], which is based on fluorescence microscopy. Six different fluorescent dyes were used in this assay: Hoechst 33342, concanavalin A/AlexaFluor488 conjugate, SYTO14 green fluorescent nucleic acid stain, wheat germ agglutinin/AlexaFluor594 conjugate (WGA), phalloidin/AlexaFluor594 conjugate and MitoTracker Deep Red. There were five different channels imaged (DAPI, GFP, Cy3, Texas Red, Cy5) that allow the observation of eight cellular components: the nucleus, endoplasmic reticulum, nucleoli, cytoplasmic RNA, mitochondria, actin cytoskeleton, Golgi apparatus, and plasma membrane.

Each experiment contained several plates, each of them having 384 wells. Each well is exposed to a compound treatment, enabling cells to be imaged and associated with that compound. Each experiment involved either the U-2OS or A549 cell line. The former is derived from human bone osteosarcoma epithelial cells[1]. The second cell line originates from lung tissue. Specifically, they are adenocarcinomic alveolar basal epithelial cells, meaning they come from the lung cancer, which are derived from a 58 year-old Caucasian male[2]. Because the two cell lines are showing different morphologies, each type should be analyzed separately.

Each experiment involved different bioactive compounds, but the same solvent was used in all experiments: Dimethyl sulfoxide (DMSO). There were also always a number of wells that were empty (meaning that the cells are completely untreated), or that were treated only with DMSO. The latter were used to normalize the data, because they were considered as not showing any phenotype.

Besides DMSO (which was used as a negative control), some of the datasets contained positive controls. They can help to determine whether the experiment worked as expected, since the expected phenotype is known.

[1] https://www.atcc.org/Products/All/HTB-96.aspx#characteristics
[2] https://www.atcc.org/Products/All/CCL-185.aspx#characteristics

Most of the tested compounds have a specific Mechanism of Action (MOA). MOA is defined as "the process by which a molecule, such as a drug, functions to produce a pharmacological effect. A drugs mechanism of action may refer to its effects on a biological readout such as cell growth, or its interaction and modulation of its direct biomolecular target, for example a protein or nucleic acid."[3]. In other words an MOA can be seen as the target(s) of the compound along with its type of interaction, which ultimately produces an effect on the cell. Some examples of MOAs are the following: glucocorticoid receptor agonist/antagonist, acetylcholine receptor agonist/antagonist, tubulin inhibitor, calcium channel blocker, ATPase inhibitor, etc.

This information can be thought of as the ground truth, which can be used to evaluate different profiling methods. Indeed, we expected the profiles of the compounds with the same MOA to be more similar to one another. One problem is that not all compounds have known MOA, such that we often could use only a small fraction of the dataset as ground truth.

For all datasets, morphological features were extracted from images (9 different images per well) using a pipeline in CellProfiler [4]. These features measured different aspects of the cell morphology including the intensity, texture, size, shape, and radial distribution of different regions of the cell, including nuclei, cytoplasm and the cell body. More detailed information can be found in the CellProfiler manual [4].

BBBC022

The first dataset that was analyzed came from the Broad Bioimage Benchmark Collection (BBBC) [15], which contains several sets of images produced by the Carpenter laboratory and made freely available. We selected the one from Gustafsdottir which is BBBC022v1. More information about the experiment can be found in [10].

The experiment was arranged in 20 plates. U-2 OS cells were treated with around 1,600 bioactive compounds at a single dose (10 micromolar). Each compound had 4 replicates. Each such replicate was located in the same position (the same well location) across four plates. The negative control (DMSO) had 1280 replicates. Using CellProfiler, 799 features were extracted.

Target-ID

This experiment was done in two cell types (A549 and U-2 OS). There were five plates for each cell type and five replicates per compound. One particularity of these compounds was that most of them are unannotated, meaning they have an unknown mechanism of action; in fact, the goal of the experiment is to identify the mechanism of these high-value compounds which are being pursued across various disease areas at the Broad. There are 87 different compounds and 1706 features. Some of the compounds were tested in different doses, meaning there were around 300 different compound-dose combinations, or treatment conditions.

In order to make sure the assay worked correctly, positive and negative controls were also included. There were four positive controls: Fenbendazole, Parbendazole, Emetine dihydrochloride, Cycloheximide. The two first were expected to show similar phenotypes, the next two were also showing a similar phenotype but different from the previous ones. There were two

[3]https://www.nature.com/subjects/mechanism-of-action
[4]http://cellprofiler.org/manuals/current/

other controls: Adiphenine-hydrochloride and Repaglinide that are not showing any phenotypes (i.e. they are negative controls). An example of the images of these controls along with DMSO can be seen in Figure 2.1.

LINCS Cell Painting Pilot (Repurposing)

The cells in this dataset came from the A549 cell line. The assay contained 136 plates with 1762 features. 1552 FDA approved drugs were included in the experiment, which were annotated to one or a few MOAs each. There were 625 different possible MOAs in this experiment. 3,264 observations correspond to DMSO. Each compound was tested in 6 doses (ranging from 0.0041 mmoles/L to around 12 mmoles/L. Note that each compound was not necessarily tested in the same range). Each dose contained either 2, 4, or 5 replicates, because during the image processing step some replicates were removed due to poor image quality, or because no cells were detected in them, possibly due to the toxicity of the compound. Compounds having only 2 or 3 replicates involved only around 7% of all the compounds, total, indicating the loss due to known quality-control problems was not unreasonably high.

One major challenge in this dataset was that each compound had multiple doses. The question that we needed to address here was how to select the appropriate dose in which the compound shows the intended phenotype. Note that the highest possible dosage is not always the right choice, simply because high doses could easily lead to toxicity and result in uninformative and unintended phenotypes.

2.1.2 Computational Tools

The computer used for the analysis contained 8 cores (CPU) and 16 GBs of memory. The analysis could occasionally be computationally expensive, because parts of the analysis involved single cell data. In order to increase the speed, we sometimes used AWS (Amazon Web Services)[5]. They provide a service where we can use virtual computers with the configuration of our choice. The most powerful instance that we used contains 16 CPUs and 64 GB of memory. Dealing with a lot of data, they were stored on a server. To transfer and manage files between the computer to the server we used Transmit, developed by Panic[6].

In this project (source code available at https://github.com/broadinstitute/imaging_metric_comparison), we mainly used R version 3.3.3 to perform the data analysis. The IDE used in this project is RStudio version 1.0.136 [23]. The main and most useful libraries in our analysis were `dplyr` for data manipulation and `ggplot2` for data visualization. For the preprocessing of the data (as normalization, zero features variance removal or feature selection as `findCorrelation`), we used Cytominer [26].

To reduce the computation time, the code was parallelized, if possible, using two different libraries: `foreach` and `doMC` [29]. Nevertheless, in some algorithms, R was not fast enough. A solution to this problem was to implement these algorithms in C++. An R package - `Rcpp` [8] - allowed us to incorporate the code written in C++ into R scripts. A specific Rcpp library - `RcppArmadillo` [9] - was used when linear algebra operations were needed.

[5]https://aws.amazon.com/ec2/?nc2=h_m1

[6]https://www.panic.com/transmit/

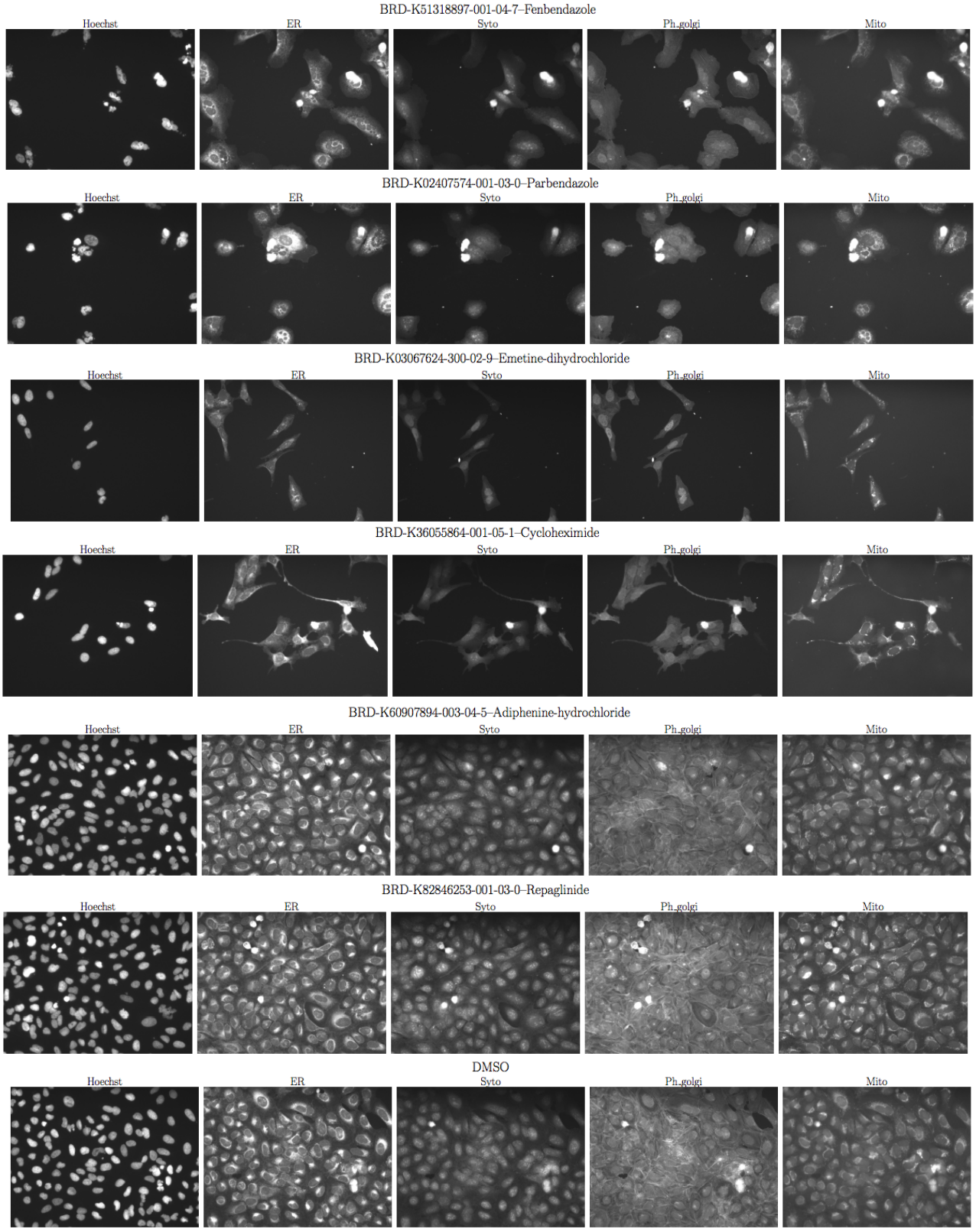


Figure 2.1: Cell images of five different channels for different controls. The first two rows correspond to the controls showing the same phenotype, the next two show the phenotype for another two controls which are expected to be different from the phenotype observed in the first pair. The last two are negative controls, showing similar phenotype to DMSO (last row). Images are acquired from BBBC022 [10].

2.2 Methods

The first part of the methodology investigates data quality control. There are a few tests that should be performed before doing any analysis in order to ensure that the data are of high quality overall.

One of the challenges in the profiling pipeline is to find a single signature for each compound that is powerful enough to represent the sample, and also summarizes all replicates. To achieve this, we needed to define a similarity metric between the replicates and a threshold that would determine if the replicates were consistent enough. We called this process *hit selection*, as it determines compounds which showed a detectable phenotype.

After hit selection, we wanted to determine if similar profiles share the same MOA more often than expected by random chance. To do that, we applied some statistical tests and visualization methods. This part of the analysis helped to confirm the strengths and the weaknesses of the different metrics used for hit selection in order to determine which one should be used.

Being more or less confident about the hit selection, we finally investigated whether feature selection can improve the quality of the signature.

2.2.1 Data Problems and Quality Checking

Data quality control is quite challenging in biological experiments. In addition to the experimental variation in cell preparation and imaging, the data is going through multiple computational processes which might be subject to errors and thus can lead to impaired analysis. One should always double check the quality and run a few tests to avoid this scenario.

Plate layout design can help a lot assessing quality. For example to detect position effects, it might help to shuffle the compounds' location in each plate, and try to avoid putting the replicates and/or various doses of the same compound next to each other. Unfortunately the machinery used in the experiments made this impractical. Often replicates are on different plates but at the same well position.

The first easy step that we can do in the context of morphological profiling is to examine a subset of images. Although it is impractical to view the entire data set, it gives insight into the quality of the data. If positive and/or negative controls were included in the experiment, visually examining the controls aids in judging the quality more systematically.

Then, the simplest method is a visual check, plotting a measured variable (often, cell count or cell area) as a heatmap in the same spatial format as the plate; this allows easy identification of row and column effects as well as drift across multiple plates.

An additional analysis that can be done to check the quality of the data is to calculate the distance of the profiles from the feature space origin at the single well level. Since the data was normalized with respect to DMSO, we expected each compound that detectably affects cells to be as far away as possible from the origin. Looking at the Euclidean distance from the origin might give an insight on whether the compounds have produced a strong phenotype. The origin was defined as a zero vector of the size of the feature space. Then the treatments above the 95th percentile of the control distances to the origin were kept. If more than half

of the replicates were selected, we kept the compound. An illustration of the compounds that are selected can be seen in Figure 2.2. The green stars were selected even if they are sparse, because 4 out of the 6 replicates are far enough. On the other hand, the red triangles were not selected even if all the replicates are strongly similar (see following sections) because this method also cares about the phenotype strength, meaning that we will remove compounds showing a weak phenotype. When comparing with known high quality data sets, we can gain insight into whether the quality is good enough or not.

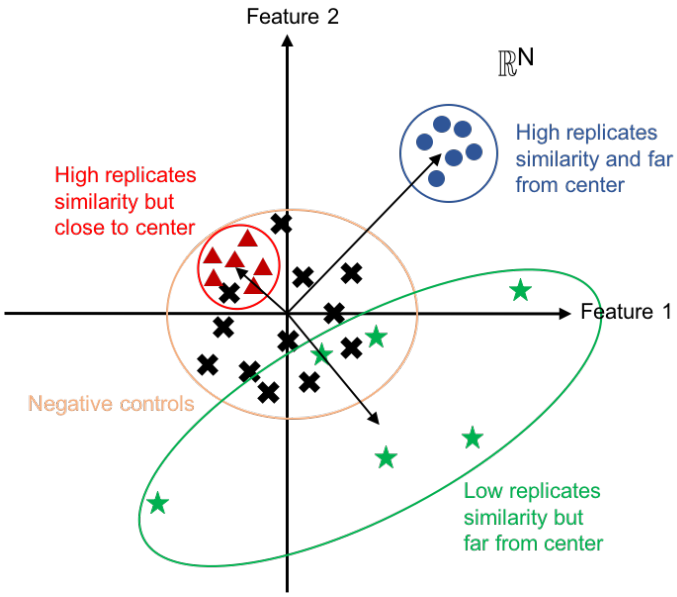


Figure 2.2: Illustration of the distance from the origin with two features. Different compounds are represented by a unique symbol with 6 replicates each. Negative controls are all close to the center and are not showing any phenotype. If more than half of the replicates for one specific compound are far from the center, the compound is selected as a hit. Red triangles are not selected while stars and circles are. Arrows show the average distance from the center.

2.2.2 Hit Selection

In order to select only compounds that have an effect on the cells, we performed hit selection. Only compounds showing a detectable and reproducible phenotype are selected.

We began by normalizing the data (z-scored) with respect to the negative controls (DMSO), because they should not show any strong phenotype. We used the median and the MAD (Median Absolute Deviation) for the normalization instead of the mean and the standard deviation, because the median is more robust to outliers [5].

Looking at the similarity between the replicates (defined in Section 2.2.3), the compounds showing a low similarity between replicates were removed because either they were not showing a phenotype different from DMSO, or the signature is not of high quality. An illustration of the definition of low versus high replicate similarity is introduced in Figure 2.3. The negative controls can be see as references; the rest of the compounds were normalized according to them and if they were all concentrated and away from the negative controls they are considered as hits. On the contrary, the red squares being spread are showing low replicate similarity.

To find the threshold over which compounds are showing a phenotype, we used a baseline

distribution defined by non-replicate similarity (see Section 2.2.4) [22].

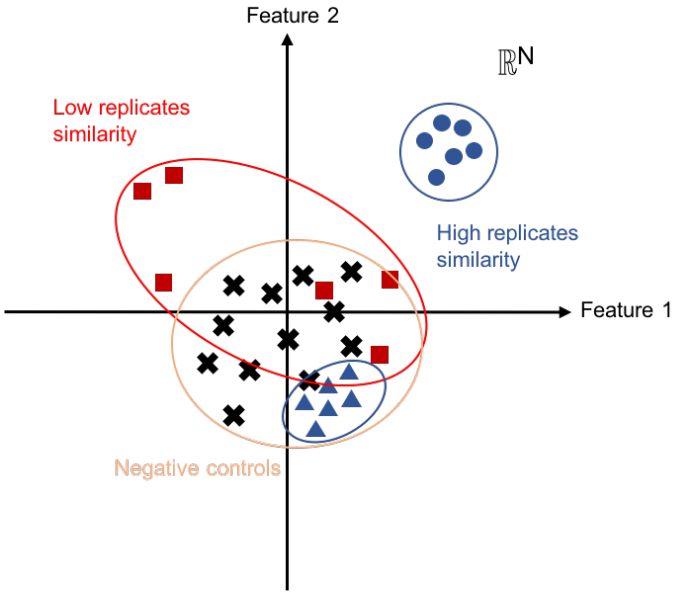


Figure 2.3: Illustration of hit selection using the feature space with the number of features N equal to 2. Different compounds are represented by a unique symbol with 6 replicates each. Negative controls are not showing any phenotype. Unlike low replicate similarity (red squares), high replicate similarity compounds (blue triangles and circles) are showing a strong phenotype and thus are hits.

2.2.3 Similarity Metric

To measure the similarity between the replicates, two different types of metrics were analyzed: the correlation and the distance metric.

Correlation Metric

Different correlation metrics were tested [25]:

- **Pearson**: The correlation metric describes a linear relationship between two variables. This works best if the two variables exhibit a linear relationship. A positive correlation means the variables are changing in the same direction. On the contrary, with a negative correlation, they are changing in opposite directions. And finally a correlation of zero means there is no linear relationship between the variables.

- **Spearman**: finds the correlation between the rank. It implicitly describes any monotonic relationship between the variables (i.e. if one increases, the other also increases and inversely). It does not assume any linearity between the variables.

- **Kendall rank**: measures ordinal association of two random variables [20].
  $\tau = \frac{(\text{number of concordant pairs}) - (\text{number of disconcordant pairs})}{n(n-1)/2}$

A perfect compound will have all the replicate correlation pairs equal to 1. Meaning that all replicates show exactly the same phenotype. But in reality, we do not expect this to happen,

because of various inter-replicate variations. We can determine the strength of the phenotype of a specific compound by finding the median of the pairwise replicate correlation.

Besides these, the cosine distance was also considered but not tested. We expected it to give similar results to the Pearson correlation, because the Pearson correlation can be seen as a normalized version of the cosine distance. Indeed mathematically we can see that:

$$cos(\theta) = \frac{< a, b >}{||a||_2 \, ||b||_2} \qquad and \qquad corr(a, b) = \frac{< a - \mu_a, b - \mu_b >}{||a - \mu_a||_2 \, ||b - \mu_b||_2}$$

are equal when $\mu_a = \mu_b = 0$. We expect this to be more or less the case here since the data is already normalized.

Distance Metric

The second way to measure similarity is by measuring the distance between the profiles. There are three metrics already implemented in R that we analyzed and another one that was implemented from scratch:

- **Euclidean**: it is the usual distance between two vectors (also named as the L2 norm).
- **Maximum**: takes the maximum distance across all the features.
- **Manhattan**: takes the absolute distance between two vectors (L1 norm).
- **Jaccard**: the Jaccard distance looks at the similarity between two sets as illustrated in Figure 2.4. For a given pair of profiles, we first sorted the normalized feature values. Then the set of top and bottom n features were obtained and compared using the Jaccard measure. 0 means that the two sets are exactly the same and 1 indicates that they are totally different.

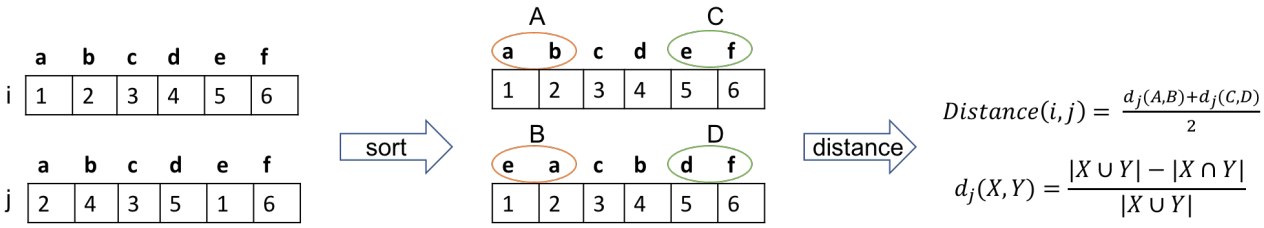


Figure 2.4: Illustration of the Jaccard distance between two profiles i and j, with n, the number of top/bottom features that are selected (here n = 2). The final distance is the mean of the Jaccard distance of the top sets (orange) and the bottom sets (green). **1.** Features are sorted based on their value. **2.** Names of n top/bottom selected features are selected. **3.** Jaccard distance of the corresponding sets are calculated. A distance of 1 indicates that i is totally different from j. **4.** Calculate the average of the Jaccard distances from top/bottom features.

The main advantage of the Jaccard distance metric is that it is independent from the features' magnitude, and it does not pay attention to the middle range features, meaning that it is more robust to noise. However, this method has an extra hyperparameter which is the number of selected top/bottom features. We set it arbitrarily to 5% of the total number of features in this project.

2.2.4 Thresholding of Poor Replicate Correlation

In order to only select the compounds that are showing a strong replicate correlation, we defined a null distribution H_0. There are two different ways of defining H_0:

1. **Median of non-replicate similarities**: looks at the similarity between random samples that come from different compounds. Calculate the median of these non-replicate samples. Repeat this step 5000 times to sample the null distribution.

2. **Replicate correlation control**: constructs a distribution based on the negative controls. This null distribution is mainly used to detect position effect. It is created by using DMSO correlation pairs of a same well.

The first null distribution — the median non-replicate similarity — was used. If position effects were suspected, we would use the null distribution based on the controls. Normally we did not expect any strong correlation between the controls, but if there were some position effects, the distribution would be shifted to the right (positive correlation).

The threshold used for hit selection is defined either as being the 5th percentile (for the replicate distance) or the 95th percentile (for the replicate correlation) of the null distribution. The percentage of compounds selected based on this threshold is called the hit ratio. This number gives insight into the sensitivity of our assay (meaning how coherent replicates are). However these metrics alone do not evaluate the specificity of compound signatures. To measure the specificity, we needed to add ground truth: the known MOA information (see Section 2.2.6).

2.2.5 Signature of a Compound

After hit selection, we averaged the replicates for each feature to obtain the compound's signatures. The main positive aspect of taking the average is that it reduces the dataset size and also removes some noise in the data. Also it should be noted that the compounds that were not stable enough across replicates were removed in the previous step, resulting in further noise removal.

Nevertheless, we also considered the case where we keep all the replicates rather than averaging them. We wondered whether having more samples could result in a better estimation of the similarity. One of the drawbacks was the increased computational cost.

2.2.6 Visualization Method

We attempted many different visualization techniques in order to find a reliable, interpretable, and informative method.

Hierarchical Agglomerative Clustering

The first method clustered the compounds using the hierarchical agglomerative clustering. The objective was to compare methods and select the one that has a higher number of compounds

sharing the same MOAs clustered together.

The first step in this clustering was to calculate the correlation between each compound pair, and find the distance defined as $1 - correlation$, needed for the Hierarchical agglomerative clustering in the function *hclust*[7] in R.

Hierarchical agglomerative clustering constructs a tree, which is also called dendrogram. It starts with each point being a separable cluster (these are the leaves of the tree), and iteratively merging them based on a specific linkage function (see below). This is repeated until one single cluster remains (root or the tree).

There are different linkage functions to decide which clusters are the closest. The main three are *single*, *complete* and *average* (see Figure 2.5 to illustrate the different linkage functions). The first one takes the minimum distance between all pairs of members in two clusters, and the two clusters with the smallest minimum distance are merged; the second one is the same except that it looks at the smallest maximum distance. The problem of single-linkage is that it tends to create long chains; on the other hand the complete-linkage creates more compact clusters. The average linkage function is a compromise between single and complete linkage. It takes the average of all possible pairwise distances from two clusters, and merges the two clusters with the minimum such average. This leads to a more stable clustering, because it is less sensitive to outliers due to averaging.

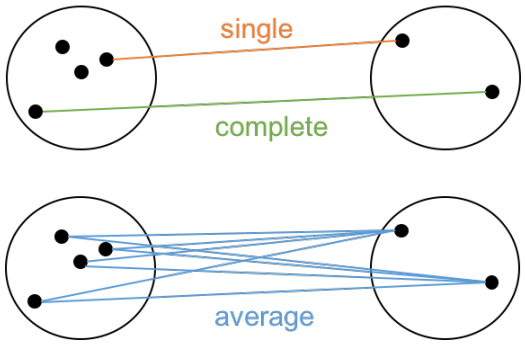


Figure 2.5: Illustration of linkage functions for the hierarchical agglomerative clustering. **Single linkage**: considers the minimum distances between all pairs of points in two clusters; **Complete linkage**: considers the maximum distance between those pairs to rank the cluster pairs to merge them. **Average linkage**: considers the average distances.

Then, in order to find the clusters, we needed to cut the tree, with the function `cutree`[8]. This function allowed us to cut the tree, based on either the number of clusters k or the maximum height h of the nodes forming the clusters. In the first attempt we used h. To set the threshold h, we looked at the stability in a heuristic way. The stability measures how much a cluster changes when changing the variable gradually. We searched for the lowest threshold with the maximal stability to estimate the threshold.

The stability is defined as follows (see Figure 2.6 as an illustration):

$$\text{Stability around } h = \frac{\#\text{ of stable clusters}}{\#\text{ of cluster in } h+}$$

The number of stable clusters is defined as the number of clusters that have not changed

[7] https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/hclust
[8] https://www.rdocumentation.org/packages/dendextend/versions/1.4.0/topics/cutree

between $h- := h - \epsilon$ and $h+ := h + \epsilon$, where the cut h is the threshold for the dissimilarity between the clusters.

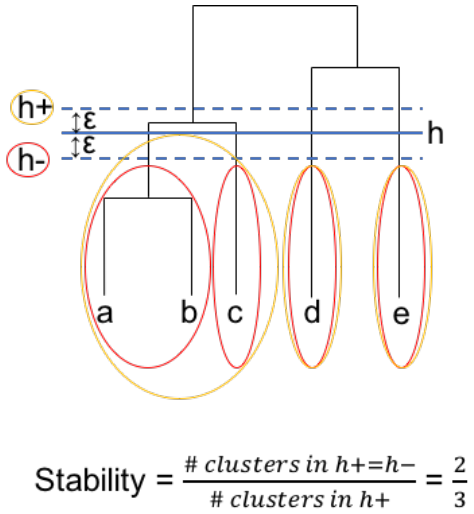


Figure 2.6: Example of the clustering stability around h for the shown dendrogram. The stability is defined as the ratio between the number of clusters that has not changed between $h-$ and $h+$ and the number of clusters in $h+$.

Finding a stable height of the tree was not straightforward. We should already notice that the higher we are in the tree, the more stable it becomes. Because there are fewer and fewer elements that will change the cluster. Indeed, at the beginning of dendrogram construction, a lot of elements were very similar, so they would rapidly merge together into the same cluster (meaning the clusters will be unstable). Instead of selecting h based on the stability, we could alternatively set a number k of clusters, where the desired cut results in that number of clusters.

Clustering Evaluation

To evaluate the resulting clustering, a two by two contingency matrix was built as shown in Table 2.1. For each pair of compounds, we counted the number of pairs that have same/different MOA within same/different clusters.

	Same MOA	Different MOA
Same cluster	a	b
Different cluster	c	d

Table 2.1: Contingency matrix for evaluation of the agglomerative clustering. For each compound pair, we counted whether it has the same MOA and the same cluster or not. Odds ratio is defined as ad/bc.

Using Fisher's exact test[9][18], we could compute the odds ratio (OR) which is defined as follows:

$$\text{Odds ratio} = \frac{\text{same MOA and cluster} \cdot \text{different MOAs and clusters}}{\text{different MOAs and same cluster} \cdot \text{same MOA and different clusters}} = \frac{\text{ad}}{\text{bc}}$$

[9]https://stat.ethz.ch/R-manual/R-devel/library/stats/html/fisher.test.html

It tested the hypothesis whether having the same MOA and being in the same cluster is linked. If this ratio was close to 1, it would suggest that the two variables are independent. An OR bigger than 1 means that if we have the same MOA, there is a higher chance to capture the same cluster. Moreover a p-value is calculated by the test, which if smaller than 0.05, means H_0 should be rejected in favor of H_1 (H_0 = the two variables are independent).

MOA vs Correlated Compound Pairs

The evaluation of a method based on clustering clearly depends on the threshold used in cutting the dendrogram. This could be an issue if the threshold is hard to establish. An alternative method is to define another contingency matrix as followed in Table 2.2.

	same MOA	different MOA
Top 5% correlated compounds pair	a	b
Rest	c	d

Table 2.2: Contingency matrix for an alternative to agglomerative clustering. For each compound pair, we count whether it has the same MOA or not either in the top 5% correlated pair or the rest. Odds ratio is defined as ad/bc.

Here we looked instead at Fisher's exact test for the top 5% correlated compound pairs. For each pair, we determined if the compounds have the same MOAs or not. This method also involved a hypeparameter that was arbitrarily set to 5%. We also examined the odds ratio for different percentages of top connections and consider that it should be set to 2% (see the Section 3.3).

Enrichment Ratio

The problem with the previously explained odds ratio was that it was not directly actionable. In other words, it was not clear to a biologist whether a particular value of odds ratio indicates strong enough enrichment to lend confidence that the predicted MOA of a compound is the correct MOA.

An alternative method was defined as follows: we calculated for how many compounds does the top 2% of other compounds matching its signature contain at least one compound with the same annotated MOA. This quality measure was more useful than the odds ratio from a practical point of view, because it told the biologist what percentage of compound-connections would likely be verified in the follow up experiment. We normalized this metric by dividing it to the same quantity but with random matches. Random matches are simply done by shuffling the compound names. Specifically we did the following:

For the top 2% of matching compounds, we divide the following two quantities:

- the number of compounds whose MOA is found among the MOAs of the top 2% best-matching compounds, divided by the total number of compounds

by

- a baseline value of the measure, where the same analysis is done 1000 times and averaged, each time with a random shuffle of the compound names.

For example an enrichment ratio of 3 means that the number of compounds with a meaningful match is 3 times higher than when the matches are made randomly.

Waterfall Plot

Instead of looking at numbers which was not always easy to interpret, we defined a new visualization technique that we called the "waterfall plot".

Here is the procedure we follow to create these plots for each MOA:

1. select the compounds that share this MOA
2. for each selected compound, calculate the correlation to all other compounds
3. sort the correlations in the decreasing order
4. create a 2D grid where the columns are the compounds sharing the same MOA and the rows are the rank of the correlation of this compound to all other compounds.
5. mark a line in this plot if the compounds share the same MOA, and nothing otherwise.
6. sort the columns based on the average rank of the marked position in an increasing order.

Compounds with the same MOA that are extremely highly correlated will show lines completely at the top of such a visualization. More commonly, when some amount of experimental noise is present, there will be instead an enrichment of lines towards the top, resulting in a shape of a waterfall in the plot. This plot could help to show if compounds sharing the same MOA are more correlated as compared to other MOAs.

These plots could also be used to determine visually which dose is optimal for a given MOA. Indeed in some experiments considered in the project, multiple doses of compounds were tested and some of them might show a stronger and more interesting phenotype than the others.

2.2.7 Unsupervised Feature Selection

Having defined methods to evaluate and interpret the results, the next step was to investigate feature selection methods. In this project, we were dealing with unsupervised feature selection. It tries to find the optimal set of features without any knowledge of data annotations. Thus it relies solely on the underlying structure of the data. It is indeed a difficult problem, because there is no guarantee that what is found is reliable, nor valid on other datasets.

In feature selection, we selected a subset of features; an alternative which we avoided, feature reduction, aims to make new non-redundant features by combining them in several ways, but suffers from loss of interpretability. For example, in PCA, we get a linear combination of features, and it may be hard to understand what aspect of the cells a principal component corresponds to. Also there are a lot of features that were redundant and highly correlated.

Combining them was not appropriate because of the feature imbalance.

Feature selection was done to ideally select only the relevant information in order to be able to have a more generalizable model and to be less subject to overfitting. It was aimed to remove redundancy in the features and help to mitigate the curse of dimensionality.

Feature selection in the context of profiling can be performed on two levels: either at the profile or single cell level. When done at the profile level, it is performed just after the normalization on the whole dataset. On the other hand, at the cell level, it is performed on the negative controls (DMSO) so that it becomes experiment independent. Larger sample size for the single cell data helps to perform feature selection more reliably. However this comes at a price: the high computational complexity due to an increase in size of the input matrix.

Filter vs Wrapper Models

There are mainly two types of feature selection: the Filter and the Wrapper approaches [6]. The Filter model is selecting a subset of features independent of any learning algorithm. For instance, it looks at the intrinsic properties (i.e. variance, entropy, correlation, etc.) of the data. It is typically faster than the Wrapper models, however it might not give good results for a particular choice of the learning algorithm. However, the Wrapper model uses a learning algorithm to select a subset of features. It results in more reliable features but has a higher computational cost. Furthermore there is the hybrid feature selection, which is a combination of both. This class of algorithms attempts to balance the computational cost and reliability of the result. However, they seem to be more applicable on data with a low number of features. Given we had a large number of features, we focused on the Filter model, to avoid the high computational cost.

FindCorrelation Method

In the standard profiling pipeline, the feature selection that is mainly used is `findCorrelation`[10] which is a filter method. It looks at the absolute pair-wise feature correlations and removes a feature from the pair that is above a threshold set to 0.9. This threshold is applied on the absolute correlation values, because anticorrelation also indicates redundancies in the data. This method has the advantage to be very efficient computationally. For d features and n observations, it has the computational complexity of $O(d^2n)$.

This feature selection is applied preferentially on the controls. There are two reasons why it is better to consider correlation on DMSO rather than all the samples for this purpose. First, it is known that the controls display a lot of heterogeneity [14]. For example, some features were not correlated in the controls but highly correlated in the treatments. Doing feature selection on all the samples might remove these features. However, because the mentioned high correlation might be specific to the compounds and hence manifests a phenotype, we may want to keep both features. The second reason was that the feature selection becomes independent of the compounds present in the experiment. This leads to a more generalizable feature selection model.

[10] https://www.rdocumentation.org/packages/caret/versions/6.0-76/topics/findCorrelation

SVD-entropy Feature Selection

In [31], the authors developed a filter method based on the SVD-entropy. This algorithm selected a feature depending on its contribution to the entropy of the singular values of the data matrix (CE), calculated on a feature leave-one-out basis. There are 3 different proposed variants, Simple ranking (SR), Forward Selection (FS) and Backward Elimination (BE).

The contribution of entropy of the feature i was calculated as follows:

$$CE_i := E(A_{n\times d}) - E(A_{n\times(d-1)}) \tag{2.1}$$

$$\begin{gathered} E(B) = -\frac{1}{\log(N)} \sum_{j=1}^{N} V_j \log(V_j), \\ V_j = \frac{s_j^2}{\sum_k s_k^2} \quad \text{and} \\ B = M^T \Lambda N \text{ is the SVD decomposition of B} \end{gathered} \tag{2.2}$$

Where A is the dataset of dimension $n \times d$, n being the number of observations and d the number of features. $E(B)$ is the entropy and V_j are the normalized eigenvalues (J_s) of B calculated by the SVD.

The entropy is a value between 0 and 1. 0 means that there is a lot of redundancy in the data and only one single eigenvector can explain it. On the other hand, 1 means that the features are independent and non-redundant. We wanted to keep the features that give the higher gain of entropy.

The paper proposeed different variations of this idea as follows:

1. Simple ranking (SR): sort the features in a decreasing order of CE and select the first m. This m was set as being all the features with a positive gain of entropy. Indeed, it means that the features are contributing to the information of the data. On the other hand, the features with more noise should be removed.

2. Forward Selection (FS): two different implementations:

   (a) FS1: accumulates features according to which set produces highest entropy. The first feature is the one giving the highest CE. Then, iteratively, it calculates the CE with this first feature and with each of the remaining features, and selects the one giving the highest entropy set. This is done until d_c features are kept.

   (b) FS2: accumulates features through the choice of the best CE out of the remaining. The first feature is the one giving the highest CE. Then, iteratively, it recalculates CE for the remaining features and select the highest. This is done until d_c features are kept.

3. Backward Elimination (BE): iteratively eliminates features with the lowest gain of entropy. In each iteration, it recalculates the CE for the remaining features. This is done until d_c features remains.

The pseudo-code for FS2 in the paper is as follows:

initialization: $\tilde{X} = \emptyset$ and $X' = X$;

while *size of* $\tilde{X} < d_c$ *(number of features to keep)* **do**

  Select the element in X' ($\forall x \notin \tilde{X}$) with the **highest** CE score;

  Remove from X', insert into $\tilde{X}$

end

$A_d \leftarrow A_{d_c}$

X is the initial matrix, A_{d_c} is the initial matrix reduced with the d_c features selected by the algorithm.

Algorithm 1: Pseudo-code of Forward Selection in method FS2.

At the single cell level, there is a problem of dimension when calculating the SVD. The singular value decomposition becomes computationally prohibitive if n (number of sample cells) $\gg d$ (number of features). In order to make it computationally feasible, we perform SVD on $A_{d\times d} = X^T X$ rather than $X_{n\times d}$.

$$SVD(X) : X = M^T \Lambda N \tag{2.3}$$

$$SVD(A) : A = N^T \Lambda M M^T \Lambda N = N^T \Lambda^2 N \tag{2.4}$$

In equation 2.4, $MM^T = I$, because the matrix M is orthonormal (i.e. orthogonal and of norm 1). By taking the square root of the eigenvalue of $SVD(X)$ we will get the same eigenvalues as in A.

The complexity to calculate the SVD of A is $O(d^2 n + d^3)$. The SVD has to be calculated once for each feature, this gives a total complexity of $O(d^4 + d^2 n)$, the last part being the matrix multiplication. For the SR algorithm, we had to calculate the SVD-entropy once for all features and we needed to sort them. It gave a complexity of $O(d^4 + d^2 n + d\ log(d))$. On the other side, the complexity for FS2 was equal to $O(d^5 - (d - d_c)^5)$, because we had to recalculate the SVD-entropy for all remaining features at each iteration.

The BE method was implemented but the complexity was too high, $O(d^5 - d_c^5)$, and thus it was computationally too expensive to run on real data. It was higher than FS2 because we normally expected d_c to be equal or lower than $d/2$.

In preliminary results, FS2 gave better results than the two other methods and thus was our initial choice for further testing. Still, we also analyzed SR for computational reasons, given its lower complexity.

2.2.8 Feature Set Comparison

To compare the sets of features that were selected during different feature selection algorithms, as a first insight, we looked at the percentage of common names in both sets. Having a percentage that is small did not mean that the results are totally different, because so many features were highly correlated. Given this, we performed another test, where we used the cosine distance between the compound-compound correlation matrices after hit selection. A cosine similarity of 1 means that both sets of features are giving exactly the same result.

We used the mentioned tests to compare whether taking more cells at the single cell level is improving the pipeline or not. We also compared two different feature selection methods, findCorrelation and SVD-entropy.

Chapter 3

Results

The first results that are presented introduce the problem of data quality. We discuss some of the data quality tests that could be used to assess whether the data is usable. This has to be the first step after the data acquisition. Being confident about the data, we next analyze the profile connectivity of the hits, which are the samples having acceptable replicate quality. Finally, we examine the problem of feature selection. The reasoning behind that is that by reducing the feature space we will avoid the curse of dimensionality and thus improve performance. The aim is to remove the features that do not contain information or the ones that are redundant, which can hide features that are underrepresented in the data.

The results were mainly obtained from the BBBC022 dataset. Some of the steps were repeated on the Repurposing dataset to have a confirmation of the results. The Repurposing dataset has different doses, so we first restricted ourselves to only the compounds having a dose of 10 mmole/L (stock concentration), being equivalent to 8.3 μmole/L (cells concentration)[1]. We selected this concentration because in the BBBC022 datasets, as can be found in the Gustafsdottir paper [10], most of the compounds were around this concentration. Finally, the Target-ID data was only analyzed for the quality control part.

3.1 Data Problem and Quality Control

One important step in all experiments was to test the data quality. Experiments consist of a long process which involves many people, both in terms of preparing the cell samples and analyzing the images. Most of the datasets that we used in this project, such as BBBC022 and the Repurposing datasets, have already been used in other projects. Therefore the quality check was already performed. However, the Target-ID data was new and thus tests were necessary to validate their quality. These tests were unfortunately not passed for this dataset. Below we explain the steps that led us to this conclusion.

First of all, having positive controls, we could look at the cell images in Figure 3.1 and compare it with the phenotypes we were expecting, as in Figure 2.1. In this new dataset, we could observe that the phenotypes in most of the wells were all visually looking similar to one another.

[1]This calculation can be made knowing that the final dilutions are the following: 1:3 from stock into a compound plate and 1:400 from compound plate to the cell plate.

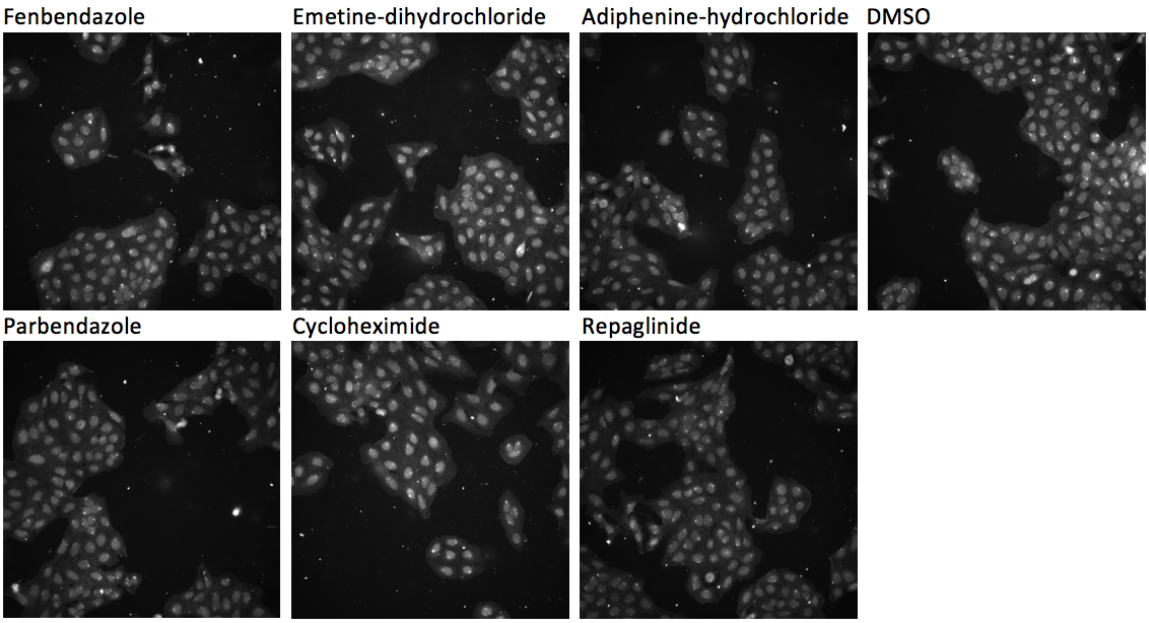
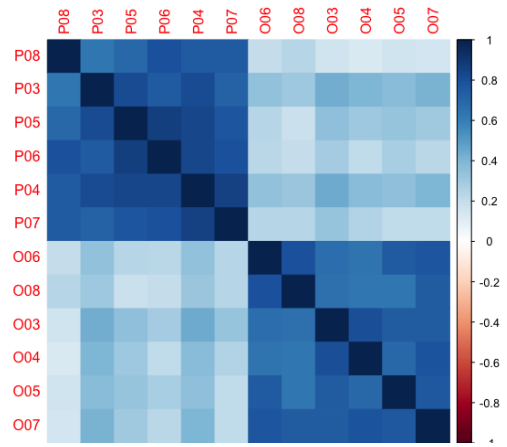


Figure 3.1: U-2 OS cells images of the different positive controls for the mitochondrial channel. Images are acquired from the Target-ID dataset.

Plotting the correlations between the controls confirmed that something was unexpected. In Figure 3.2, we can see that all the wells on the same row in the plate are correlated with one another. Moreover all the controls, irrespective of their type, are positively correlated with one another. The next step was to determine whether all the plates and all the wells were problematic or if only a subset of the data was of low quality.



	3	4	5	6	7	8
O	Fenbendazole	Parbendazole	Emetine dihydrochloride	Cycloheximide	Adiphenine-hydrochloride	Repaglinide
P	Fenbendazole	Parbendazole	Emetine dihydrochloride	Cycloheximide	Adiphenine-hydrochloride	Repaglinide

Figure 3.2: **Top**: Correlation between the positive controls of the U-2 OS cells. **Bottom**: Platemap of the positive controls. Each color should show the same phenotype. Analysis performed on the Target-ID dataset.

To achieve this goal, we plotted the cell count in the form of a heatmap for each plate. Figure 3.3 indicates an obvious edge effect. Indeed cell count is lower at the plate edges. This pattern is usually seen when the temperature is not uniformly distributed, i.e. when it is higher in the center than in the surrounding. A lower temperature may lead to a lower growing rate,

inducing smaller cell counts [16]. Moreover one can notice that, apart from the overall patterns at the edges, the cell count in most wells is pretty uniform, with only a few wells appearing as dark "holes" in the plate. We expected more variability, because compounds should show different degrees of toxicity. However, having mostly unknown compounds in this dataset, we did not know the expected phenotypes and it is possible that all of them were similar in terms of toxicity. Lastly, we noted that for the few dark holes in the cell count pattern, their positioning was only partly consistent from replicate to replicate. The two right-most plates in Figure 3.3 showed a consistent pattern but the other three plates showed only partial similarity.

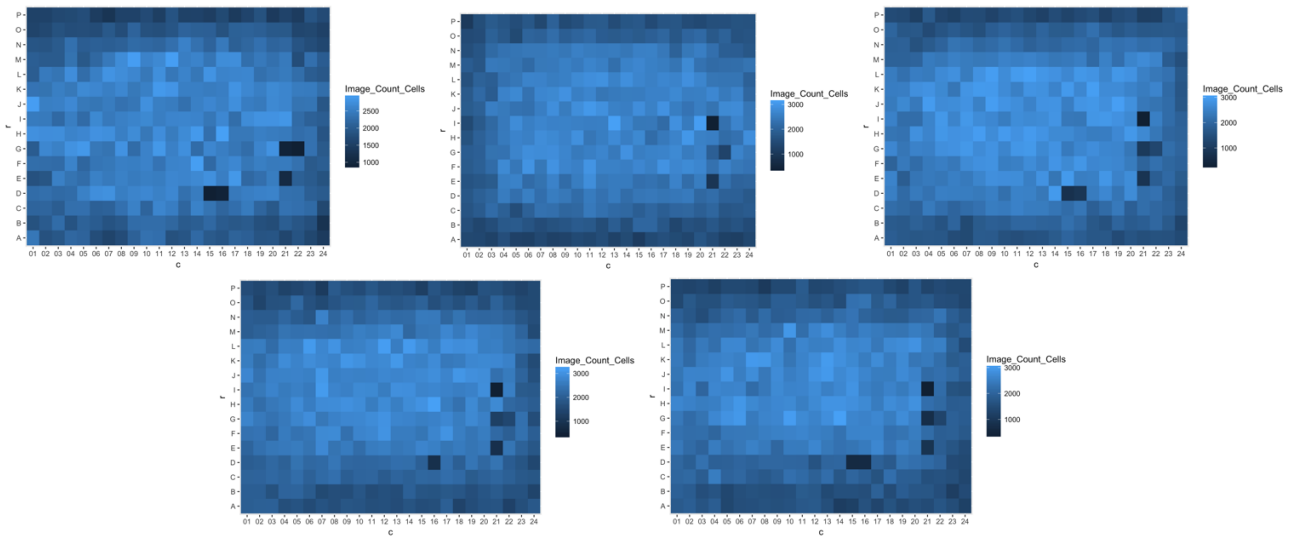


Figure 3.3: Number of cells per well for each plate in the Target-ID dataset. There are some edge effects (fewer cells in the borders compared to the center of the plate).

To look deeper into the problem of uniformity, we observed values of some features along the wells for each plates. One of those was quantifying the shape of the cells. We plotted the area of the cells in Figure 3.4 for each well in each plate. First we could again confirm the border effects. This is consistent with the cell count edge effect. Indeed, the cell area is supposed to be correlated to the cell count. We expected smaller cells when there are more cells because of space constraints. We could also observe some row-wise effects in the second plate. It was difficult to know the exact reason, but it indicates that the experiment did not go as it should have in this plate. However we could see a few patterns that are consistent between the plates. For example there were more cells in the same well across the different plates.

This experiment was relatively expensive both in terms of cost (around $10,000 just for the experiment material) and time. Although it was clear that data quality was suboptimal, we wanted to determine if part of the data could be salvaged.

We hypothesized that some compounds were not actually added to some wells in the experiment, such that many wells meant to be treated with a compound were actually untreated. To test whether we could identify treated wells, we calculated the profile Euclidean distance to the feature space origin for each sample. This distance can be thought of as a measure for the phenotype strength. It should be near zero for controls and empty wells, and much higher than zero for wells that are treated with compounds. In Figure 3.5, the distributions of controls, treatments and empty distances to the origin are plotted and observed to be very similar. Moreover, we were expecting a mean of the control values to be smaller than the mean of the treatment values which is in contradiction with what we observed. This concurs with all the previous observations, suggesting that the vast majority of treatments are behaving similar to

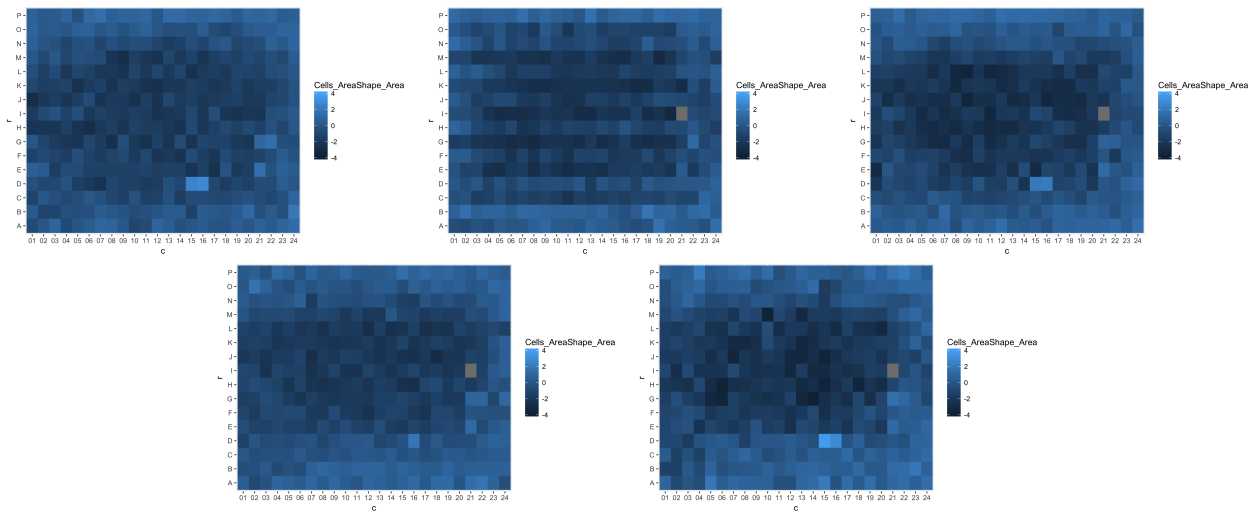


Figure 3.4: Heatmaps of the Cell Area feature for 5 plates in the Target-ID dataset. Edge-effects and a row-wise pattern indicate low data quality.

the negative controls.

To confirm the expected behavior, we repeated the same analysis for the BBBC022 dataset. Figure 3.6 shows a mean distance for the controls smaller than for the treatments. Moreover, the height of the peak is clearly higher for the controls than for the treatments, meaning that the distribution of the treatment values is broader.

In order to conclude that nothing could be saved out of this data, we looked at the percentage of treatments having a strong phenotype. A compound is defined to have a strong phenotype if its profile distance is above the 95th percentile of the DMSO distance to the origin distribution. In the Target-ID dataset, only 3% of the treatments are observed to have a strong phenotype compared to BBBC022 that has 22% compounds with that property.

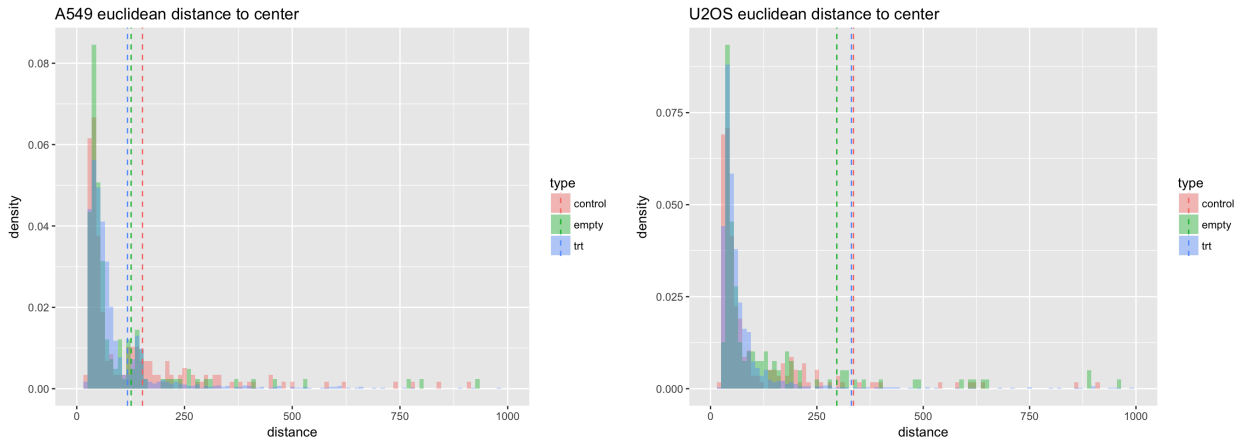


Figure 3.5: Density of the Euclidean distance to the origin (zero vector) for each sample: DMSO (red), empty wells (green) and treatment (blue). Dash lines are the mean of each distribution. Analysis performed on the Target-ID dataset for both cell lines (right: A549, left:U-2 OS).

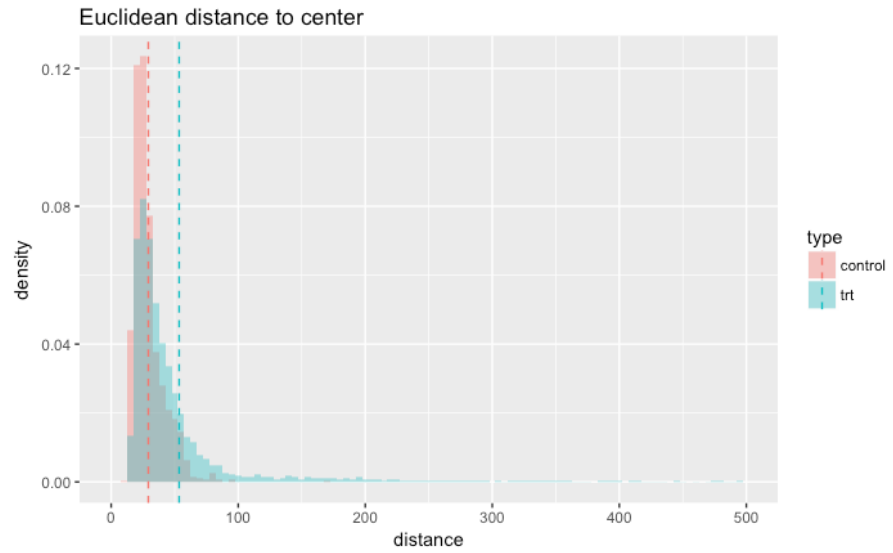


Figure 3.6: Density of the Euclidean distance to the origin (zero vector) for each sample in BBBC022 dataset: DMSO (red) and treatment (blue). Dash lines are the mean of each distribution.

3.2 Hit Selection

For datasets that have sufficient data quality, as BBBC022 and the Repurposing datasets, the next step is to select the treatments that are distinctive from negative controls, called hits. As a quality measurement, the hit ratio tells us what percentage of compounds have high replicate correlation. A high hit ratio means we have a lot of compounds that are showing a strong phenotype. We were expecting this number to be greater than 50%, because most of the compounds in the assay are supposed to show a specific phenotype.

In Figure 3.7, we can see the hit ratio for different similarity metrics (Jaccard versus Pearson) with and without feature selection (using findCorrelation). For all the methods the hit ratio is around 0.6, meaning that 60% of the compounds are hits. It is also interesting to point out that the different similarity metrics do not dramatically affect the hit ratio. Moreover the selection is quite stable, we can see that even if the null distribution is sampled based on 50 different random shufflings of the replicates, the deviation is small.

To obtain the hit ratio, we needed to set a threshold that depends on the similarity metric. If we are dealing with the correlation, we keep compounds above the 95th percentile of the null distribution. When using the distance metric, we select compounds below the 5th percentile of the null distribution. In order to see the effect of the chosen metric on the hit ratio, we compared three different metrics: the Pearson correlation, the Jaccard distance, and the Euclidean distance.

The top left Figure 3.8 shows the replicate correlation and the null distributions when Pearson's correlation was used. The vertical black line represents the threshold that identifies the hits. The density for the non-replicate correlations is centered around 0, because we did not expect any correlation between random samples. On the other hand, the density for the replicate correlations is centered around 0.5, meaning that most of the compound replicates are positively correlated.

The top right plot illustrates the same densities for the Jaccard distance. The replicate distance

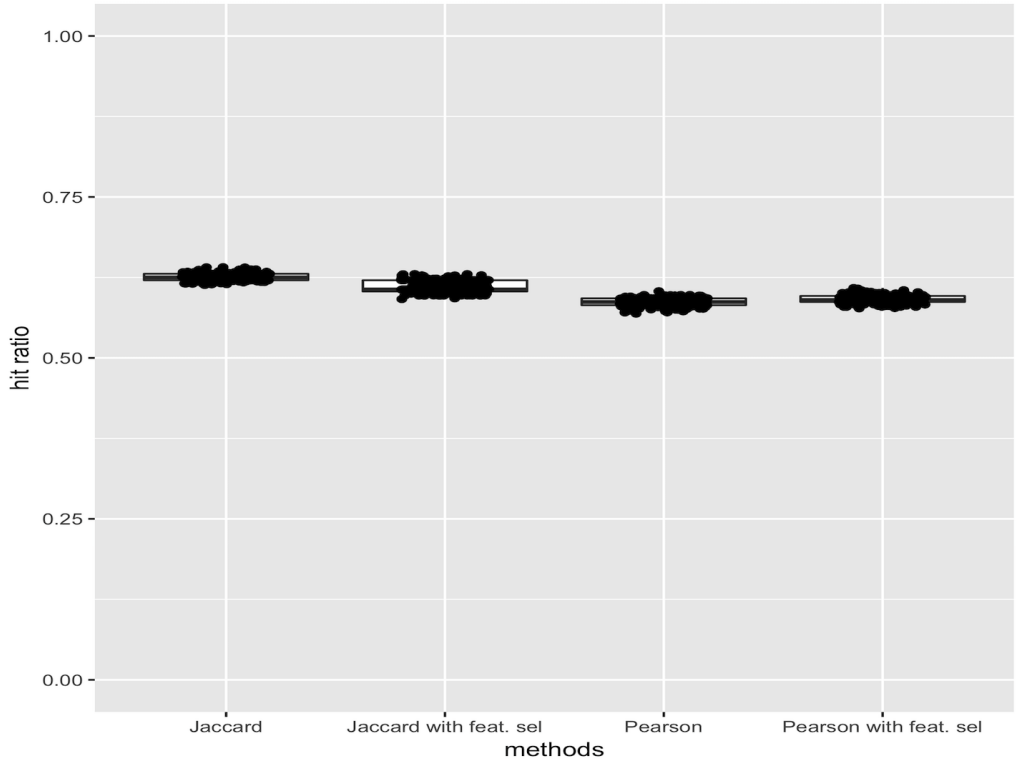


Figure 3.7: Hit ratio for different similarity metric, with or without findCorrelation features selection in BBBC022 dataset. The boxplots were made from 50 different seeds used for the random shuffling of the replicates to sample the null distribution. 5000 samples were taken in each run. Analysis performed on BBBC022 dataset.

should ideally be close to zero, which is not the case. But as expected we observe that the density of the null distribution is shifted to the right of replicate distance density.

Finally, the bottom plot shows the same densities for the Euclidean distance. Unlike the Jaccard distance, the distribution spans a larger scale, because it is not normalized between 0 and 1. We also noticed that the number of hits is lower compared to the other two previous methods.

We show the hit ratio of different metrics in Table 3.1. This preliminary analysis shows which metric is more sensitive, i.e. selects more hits. Both Pearson's correlation and the Jaccard distance yield hit ratios higher than other metrics. Therefore, we chose these two metrics for further analysis.

Pearson	Spearman	Kendall	Euclidean	Maximum	Manhattan	Jaccard
0.5806	0.5519	0.5419	0.4606	0.3612	0.4875	0.6247

Table 3.1: Hit selection ratio for different similarity metrics in BBBC022 dataset. The first two are based on correlation and the remaining are based on distance.

Here we discuss the rationales of using the Jaccard distance, and why we prefer it over the others. First of all, we should look at the main differences between the Pearson correlation and the rank correlation metric. Pearson correlation is more sensitive to outliers and less sensitive to noise contrary to rank based methods that are very sensitive to small noisy values (z-score close to 0). We should notice that both of these metrics have some benefits, being strong against noise or against outliers respectively. In order to combine these strengths, we investigated a

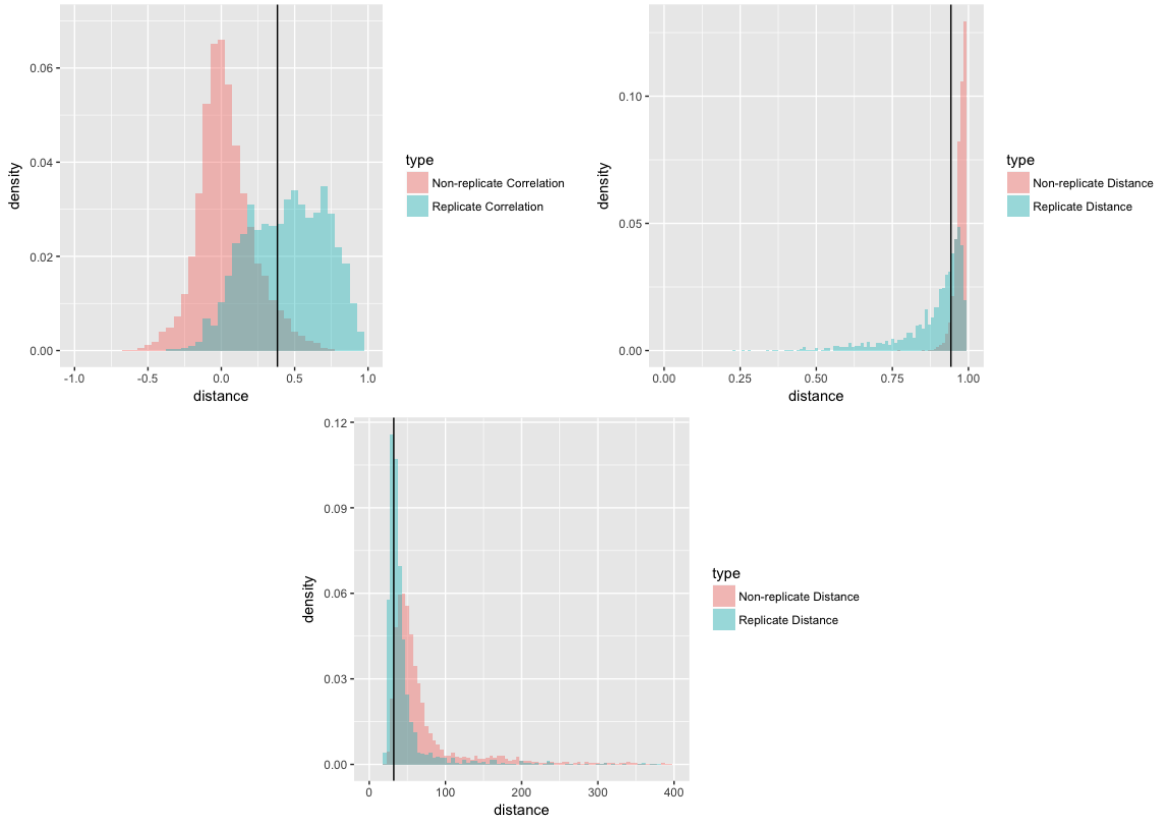


Figure 3.8: Replicate similarities analysis in the BBBC022 dataset: **Top left**: Pearson's correlation. **Top right**: the Jaccard distance. **Bottom**: the Euclidean distance. Density of the null distribution (red) (with 5,000 random non-replicate correlations/distances). The threshold (black line) marks the 95th percentile of this distribution for Pearson's correlation and the 5th percentile for the distances. The hits are the compounds with a replicate similarity (blue) higher (for the correlation) or lower (for the distances) than the threshold.

metric that strove to capture both: the Jaccard distance. Indeed, this metric only takes into consideration the extreme feature values (i.e. not considering the one with a z-score near zero) and it is looking at the sets of names, which can be seen as a kind of a ranking. Note that the Maximum distance metric is not appropriate in this application because of its sensitivity to outliers. Finally, the Manhattan and Euclidean distance yielded comparable results. But given that we wanted to select only the best, we only kept the Jaccard distance

Note that hit ratio only measures sensitivity and tells nothing about the specificity of an approach. To test the latter, we next compared profile similarity to the MOA information in the next section.

3.3 Evaluation of the Similarity Between Compound Profiles Based on MOA Information

The first evaluation method we analyzed uses a clustering algorithm. The hierarchical agglomerative clustering groups the compounds based on their morphological profile similarities. Using the MOA information, we could then evaluate these clusters, i.e. compounds sharing the same MOA should ideally be in the same cluster. In this method, we needed to set a threshold

on where to cut the clustering tree, which will then form different clusters. The objective was to obtain a small threshold with a high stability of the resulting clusters. At shallow depths we tended to get more stable clusters, because there are fewer clusters and the clusters become more distant one another, the needs in order to join them increase dramatically. We can see the stability plot of this dataset in Figure 3.9. The moving average is supposed to help to see a smoother increase towards the plateau, but we can see that the method is very unstable.

Moreover, this figure shows the dendrogram and its clusters obtained by using the threshold found in the stability plot. Dendrograms help to visualize the structure of the tree and how compounds are related but in high dimensions they may become more difficult to interpret. To evaluate the clusters and how the clusters are consistent, we used the odds ratio, given by Fisher's exact test.

		same cluster	
		True	False
same MOA	True	354	367
	False	13364	28110

		same cluster	
		True	False
same MOA	True	329	486
	False	15205	38265

Table 3.2: Contingency matrix in BBBC022 dataset for same MOA vs same cluster with hit selection using Pearson correlation (left) or Jaccard distance (right). It contains the number of compound pairs that are in each of the four categories.

The result of the Fisher's exact test is shown in Table 3.2: For Pearson's correlation hit selection, we obtain a p-value $<$ 2.2e-16 and an odds ratio of 2.0289. This means that we reject the null hypothesis. Two compounds having the same MOA have a higher chance to co-cluster. On the other hand, with the Jaccard distance metric, we obtain an odds ratio of 1.70361 with a p-value = 2.872e-13. The odds ratio here is slightly worse than for Pearson correlation. In both contingency matrices, we can observe a higher number of compounds that are in the same cluster but with different MOAs. This is expected because there are more MOAs than clusters. Moreover, some compounds might have different MOAs but expressing the same phenotype and thus are similar. On the other side, the number of compounds having the same MOA but different clusters was expected to be as small as possible. In this case, however, this number is quite high since it is comparable to the number of true positive (same MOA and same cluster category). The ideal method should have the ratio of this number and the number of true positives close to zero.

The method of stability based clustering does not seem to be reliable. Even when using a moving average, there are a lot of fluctuations and it makes the choice of maximum height quite arbitrary. Therefore, we examined the highest odds ratio at varying number of clusters. Ideally, we expected the optimal number of clusters to be at most the number of MOAs. In the BBBC022 dataset, we have around 250 compounds with strong phenotype and with one of the 100 MOAs annotated. It should be noted that some compounds can have multiple MOAs. In Figure 3.10, the maximum odds ratio is obtained for around 30 clusters for all methods except for the Jaccard distance without feature selection, which instead has a peak around 70. It is difficult at this point to determine why there is this difference.

The problem with the clustering method is that we have a hyperparameter that is hard to set: either the height of the tree or the number of clusters. We defined another evaluation method in order to avoid this issue. We picked the top 5% correlated compound pairs and checked the enrichment of this set for the same MOA pair. We expected the compounds sharing the same MOA to be more correlated, which is why we were looking at the top 5%. But this parameter might be tuned based on the prior knowledge about the dataset. In this method, we still have

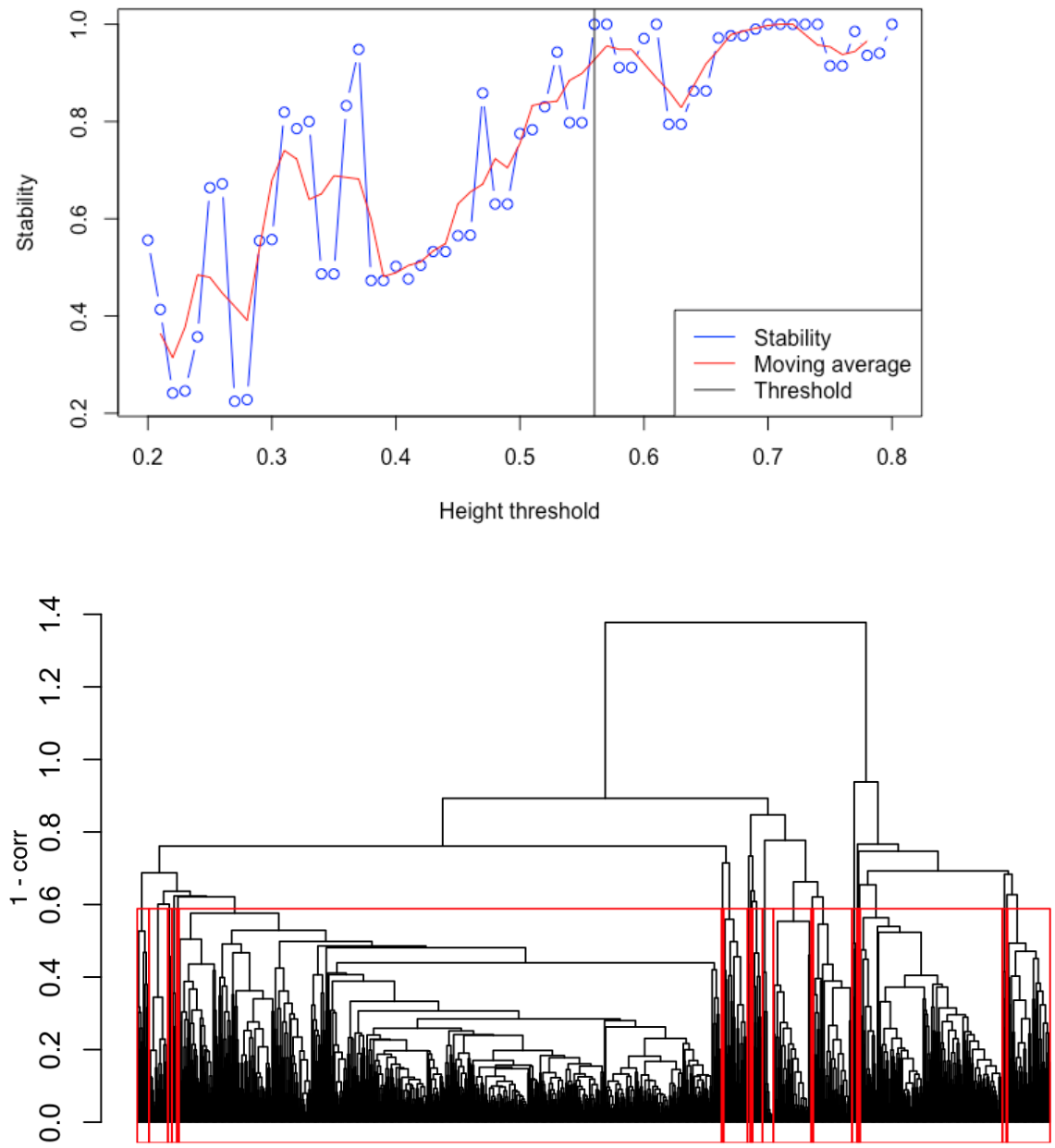


Figure 3.9: Agglomerative clustering in BBBC022 dataset. **Top**: measure of the stability as a function of the dendrogram height at which the clustering was performed. The threshold was set to the height corresponding to the first maximal peak farthest from 1. **Bottom**: The resulting dendrogram with the clusters marked by red boxes, which were obtained by cutting the dendrogram at the most stable height.

a hyperparameter but it is a lot more interpretable and hence much easier to set.

Evaluation of different metric settings based on the described method can be viewed in Table 3.3. The odds ratio with the Jaccard distance is higher than that of Pearson's correlation. Feature selection also seems to improve the results in both similarity metrics. Moreover the p-value is small ($\ll 0.05$) for all the metrics, meaning the results are significant.

Odds ratio is plotted against the percentage of selected top connections in Figure 3.11. We see that in the smaller percentage of selected connections the odds ratio is higher. However different methods do not seem to have huge differences for top connections. We are only interested in a

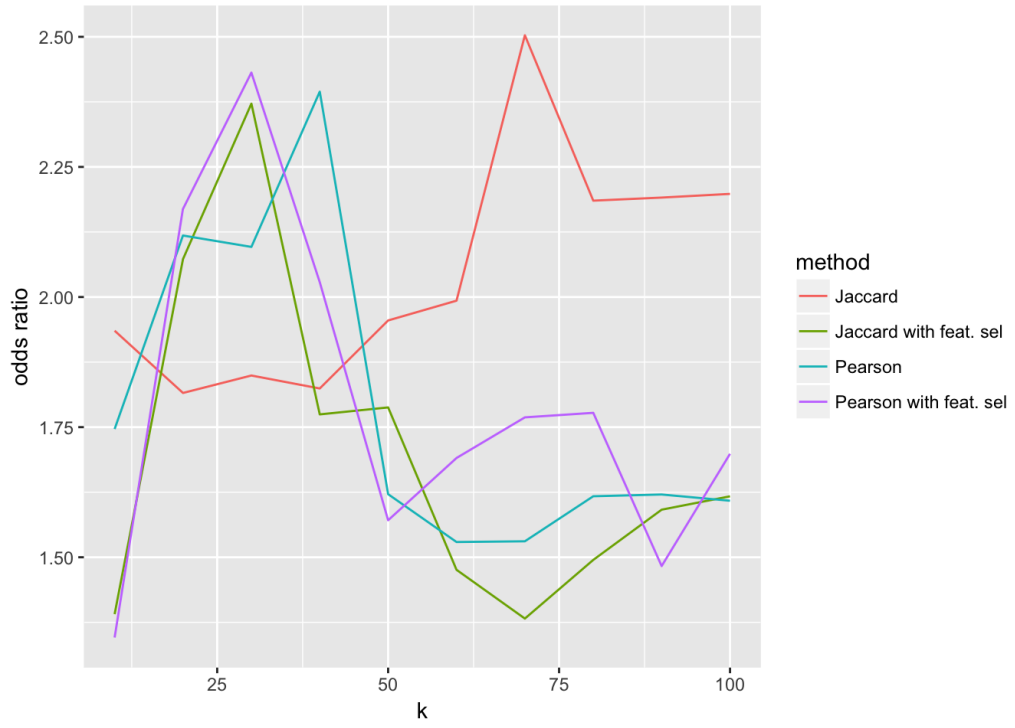


Figure 3.10: Odds ratio as a function of the number of clusters k for four different methods in the BBBC022 dataset. FindCorrelation is used as the feature selection method.

	Pearson	Pearson with feat. sel.	Jaccard	Jaccard with feat. sel
Odds ratio	2.0523	2.2943	2.3070	2.3228
(p-value)	(2.541e-09)	(7.787e-13)	(2.68e-13)	(6.555e-14)

Table 3.3: Odds ratio for different metrics for the top 5% correlated compound pairs in BBBC022 dataset. For each pair, the contingency matrix is determined by categorizing compound pairs based on their correlation and MOA annotations.

small percentage of top connections, because we are expecting the compounds with the same MOA to be highly correlated as compared to the other compounds. Moreover we can see that a plateau is reached after 10%. This means that we should consider only a small percentage of top connections, because after that the number of compounds that are "well classified" and "misclassified" are almost equal.

In order to visualize the relative ranking of compounds in a given MOA, we designed a novel plot called waterfall plot. An example of such a plot can be seen in Figure 3.12. In this 2D grid, we see blue lines when the compounds are sharing the same MOA. The rows are showing the rank of the compound-compound correlation and the columns are showing compounds ranked by increasing average rank of the marked position. We expected compounds sharing the same MOA to be highly correlated (meaning the lines should be at the top of the graph). We may lose some information in this plot, such as how well the compounds are correlated, since this plot only considers the ranking. In future iterations, the background of the plot could be shaded to indicate the correlation values themselves. On the other hand, this plot helps to understand if the compounds in an MOA yield similar profiles. For instance in the glucocorticoid receptor, we can see two groups, meaning that these two groups have the same MOA showing different phenotypes.

This plot can also help to detect outliers, which for example can be seen in Figure 3.13.

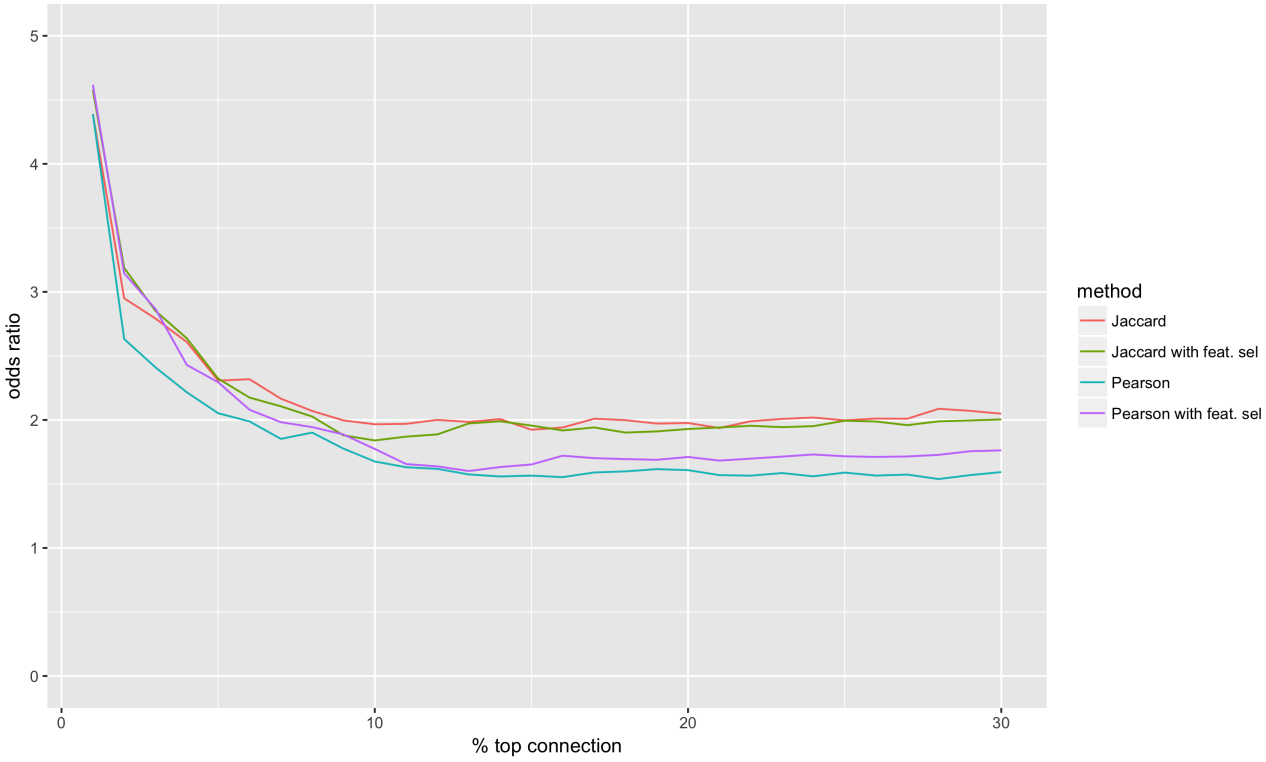


Figure 3.11: Odds ratio in function of the top $x\%$ correlated connections for different hit metrics in BBBC022. FindCorrelation is used as the feature selection method.

Griseofulvin is an outlier compared to the other compounds in the Tubulin inhibitor MOA because it is the only one not highly correlated with others. The remainder of same MOA compounds are ranked to be in top 2% of connections.

This visualization technique could also be a way to find which dose of a drug is the most effective. In Figure 3.14, we see that the dose around 3.33 mmole/L is the most effective one. Indeed, the correlation between compounds is highly concentrated at the top of the plot. Moreover it is interesting to notice that some compounds, for example KW-2449, are not correlated to other compounds of the same MOA for low doses, but at high doses it becomes correlated to them. Finally, we can also see that the lowest and the highest doses are less consistent than the middle range ones.

However not all MOAs are showing the expected patterns. Some of them seem to have an entirely random pattern. Figure 3.15 gives an example of such case. This could mean that this particular MOA is not yielding a strong phenotype. It is also possible that we do not have any pattern here because different compounds are not showing similar phenotypes at the same dose even if they are sharing the same MOA.

Finally, the last approach for evaluating different methods based on the MOA information was the enrichment ratio. This method is the one that we selected as being more reliable than the odds ratio. According to the waterfall plots we selected the top 2% correlated connections. In this method, we compared all the different feature selection methods (see Section 3.4).

But before exploring the feature selection, using the enrichment ratio, we compared the two following methods to define a signature: taking the average or keeping all the replicates separately. The enrichment ratio for these two methods is outlined in Table 3.4. Although the results of this preliminary analysis were better without averaging across the replicates, because

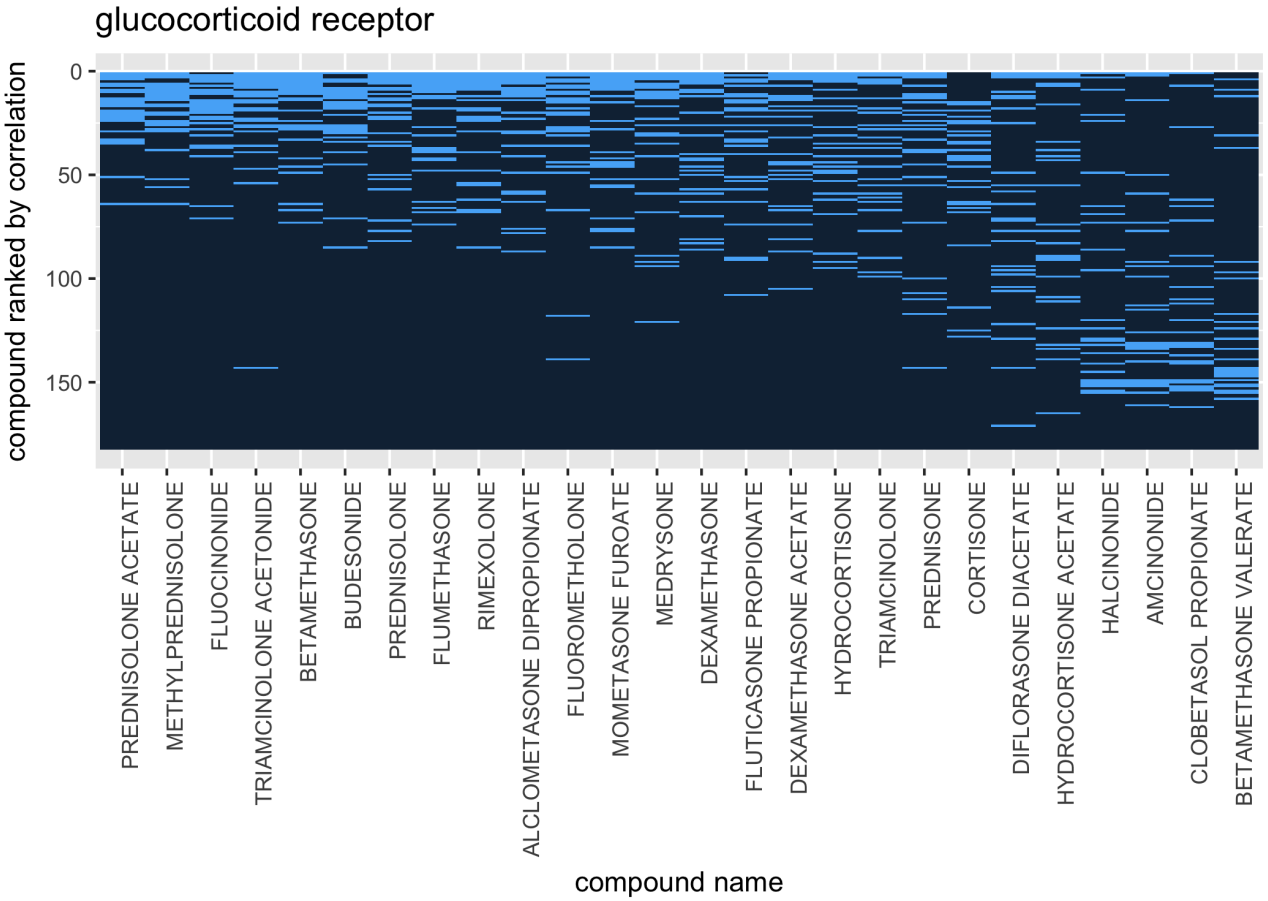


Figure 3.12: The Waterfall plot for the Glucocorticoid receptor agonist/antagonist MOA in BBBC022 data. A blue line means the compound at that correlation rank shared the same MOA as the compound noted at the bottom of the column. For each column the rows are ordered based on the correlation to the corresponding compound. **Rows**: compounds sorted based on the correlation to compounds in each column. **Columns**: compounds ranked by increasing average rank of the marked positions.

of the computational time and the higher complexity, we decided to continue working with averaged profiles. However we should note that we might lose some useful information and this could be pursued in the future.

Enrichment ratio	Average profile	Individual profile
Jaccard distance	1.8206	1.9256
Pearson's correlation	1.8785	2.0681

Table 3.4: Enrichment ratio for different settings of the metrics in BBBC022 dataset (Jaccard distance and Pearson's correlation) with averaging of the replicate profiles or not.

At this point we explored the findCorrelation feature selection with two different similarity metrics using different visualization techniques. But it is quite difficult to assess which method is really better than the other. Indeed, they all seem to be comparable. However, we expected the feature selection to play an important role on the quality of the signature. For this reason, we then explored a new feature selection method.

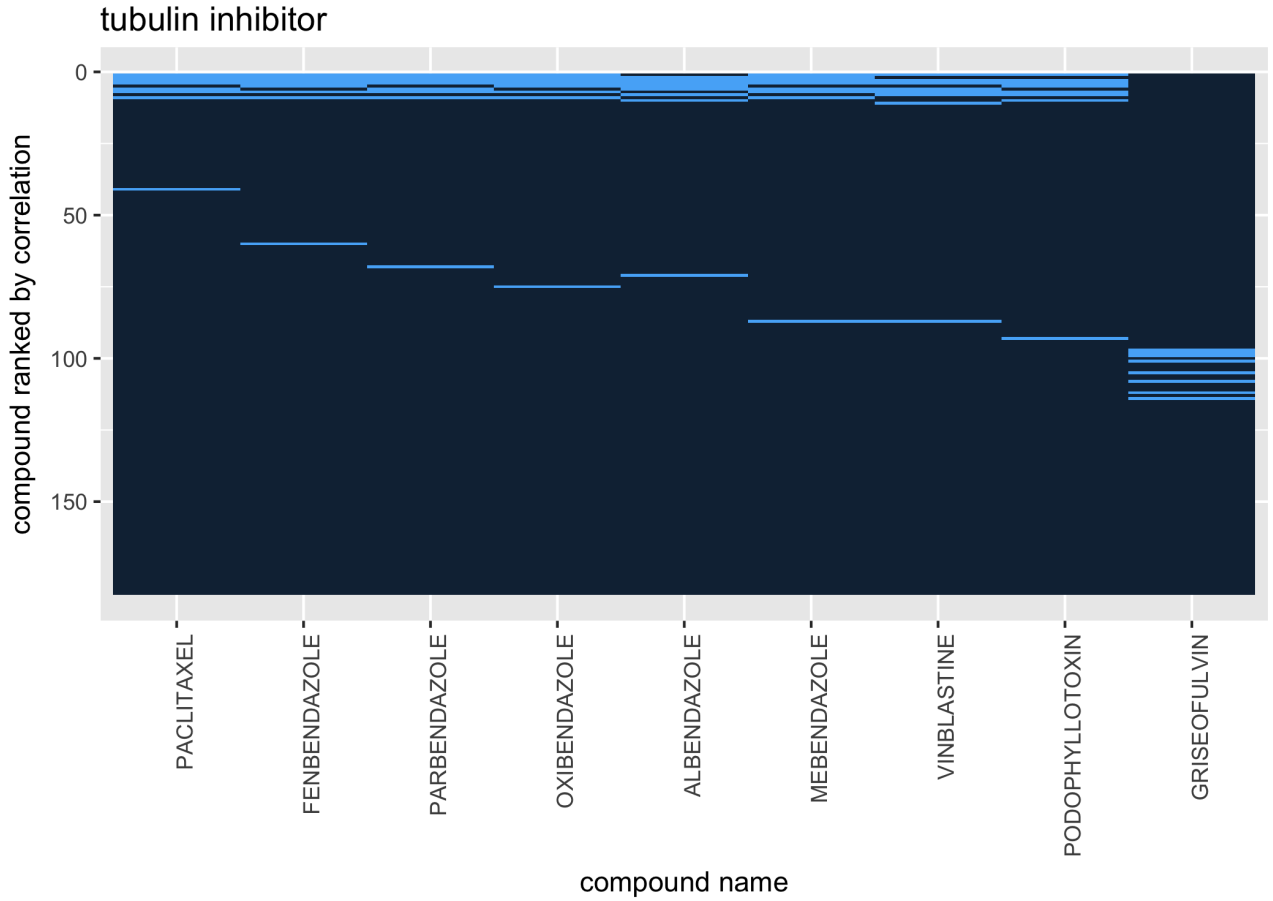


Figure 3.13: The Waterfall plot for the Tubulin inhibitor MOA in BBBC022 data. A blue line means the compound at that correlation rank shared the same MOA as the compound noted at the bottom of the column. For each column the rows are ordered based on the correlation to the corresponding compound. **Rows**: compounds sorted based on the correlation to compounds in each column. **Columns**: compounds ranked by increasing average rank of the marked positions.

3.4 Unsupervised Feature Selection

In this part of the analysis, we wanted to answer the question whether feature selection could improve the quality of the signature or at least not hinder it. Obtaining a comparable statistical performance with fewer features is defined as an improvement because in such case we use limited data to obtain the same result. Here, we compared various feature selection methods, and considered the random feature selection as a baseline. In order to make a more reliable conclusion, we performed the analysis on both BBBC022 and the Repurposing datasets.

The standard profiling pipeline uses findCorrelation at the profile level. We compared this method with the same at the single cell level as well as the method of SVD-entropy mentioned in Section 2.2.7. We initially tested the SVD-entropy using a variant of the method called FS2, at the profile and the cell levels. But eventually the SR algorithm variant was also tested at the single cell level because the FS2 algorithm was too computationally expensive. Indeed, the Repurposing dataset has twice as many features as BBBC022, thus taking too long for the FS2 algorithm to run.

We first discuss the results we obtained in the BBBC022 datasets. In the top Figure 3.16, the

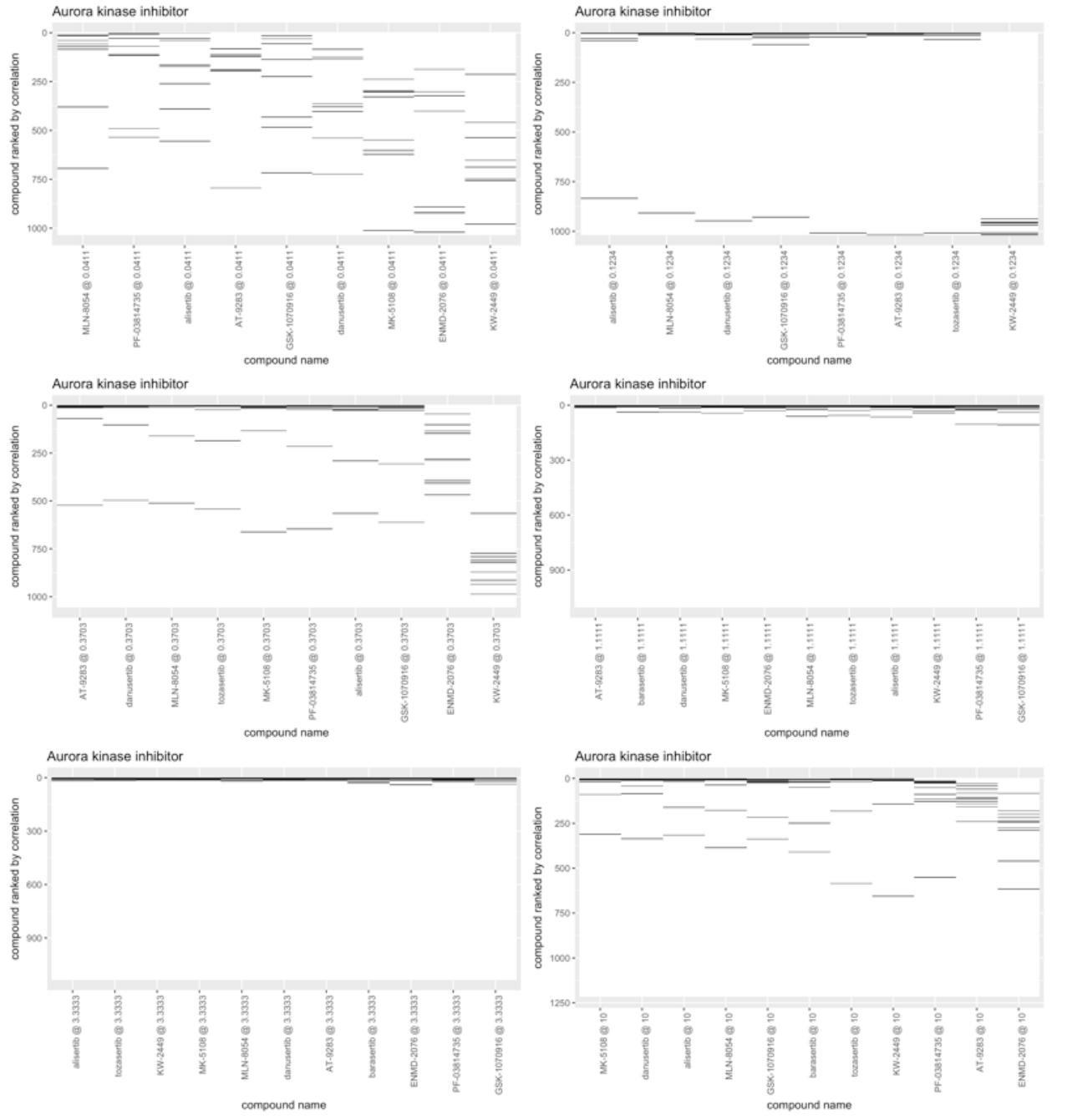


Figure 3.14: The Waterfall plots for the Aurora Kinase Inhibitors MOA at six different doses in Repurposing dataset from lowest (top left) to highest (bottom right). A black line means the compound at that correlation rank shared the same MOA as the compound noted at the bottom of the column. For each column the rows are ordered based on the correlation to the corresponding compound. **Rows**: compounds sorted based on the correlation to compounds in each column. **Columns**: compounds ranked by increasing average rank of the marked positions.

enrichment ratio is shown as a function of number of features selected for different methods. This was performed for both Pearson's correlation and the Jaccard distance metrics. The Jaccard distance results in an improved enrichment ratio compared to Pearson's correlation across different feature selection methods. Moreover, we can see that findCorrelation at the profile level (which is the standard practice) is reducing the enrichment ratio compared to taking the whole feature set. We also realized that it gives a worse result compared to random

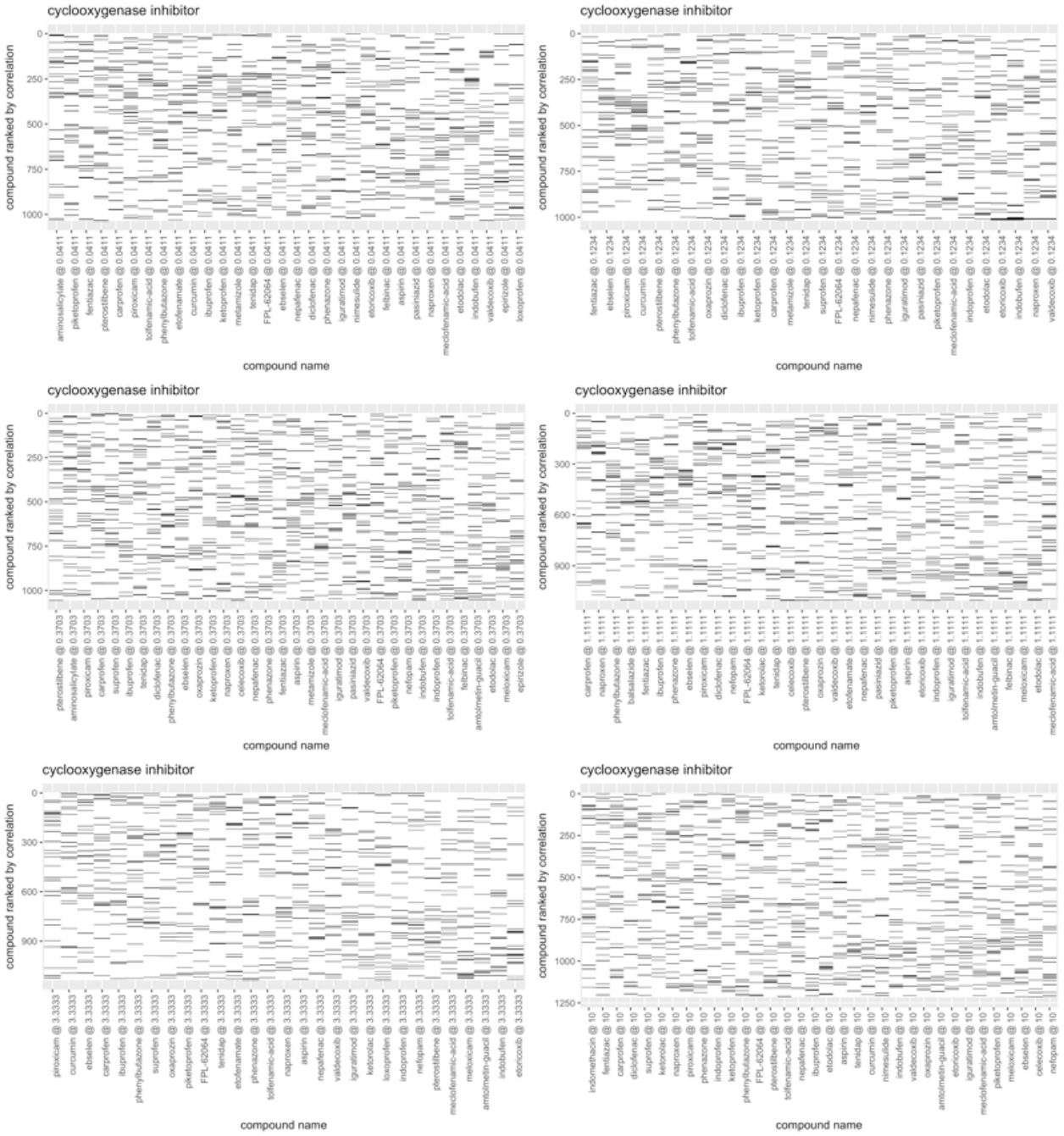


Figure 3.15: The Waterfall plots for the Cyclooxygenase Inhibitor MOA at six different doses in Repurposing dataset from lowest (top left) to highest (bottom right). A black line means the compound at that correlation rank shared the same MOA as the compound noted at the bottom of the column. For each column the rows are ordered based on the correlation to the corresponding compound. **Rows**: compounds sorted based on the correlation to compounds in each column. **Columns**: compounds ranked by increasing average rank of the marked positions.

selection, meaning the features that are selected are not optimal. Another observation we made is that using findCorrelation at the single cell level along with Jaccard distance improves the result, and according to this plot seems to be the way to go.

We compared feature selection based on SVD-entropy (denoted as SVD), in the bottom Figure 3.16, against other methods at single cell and profile levels. Note that with a very low

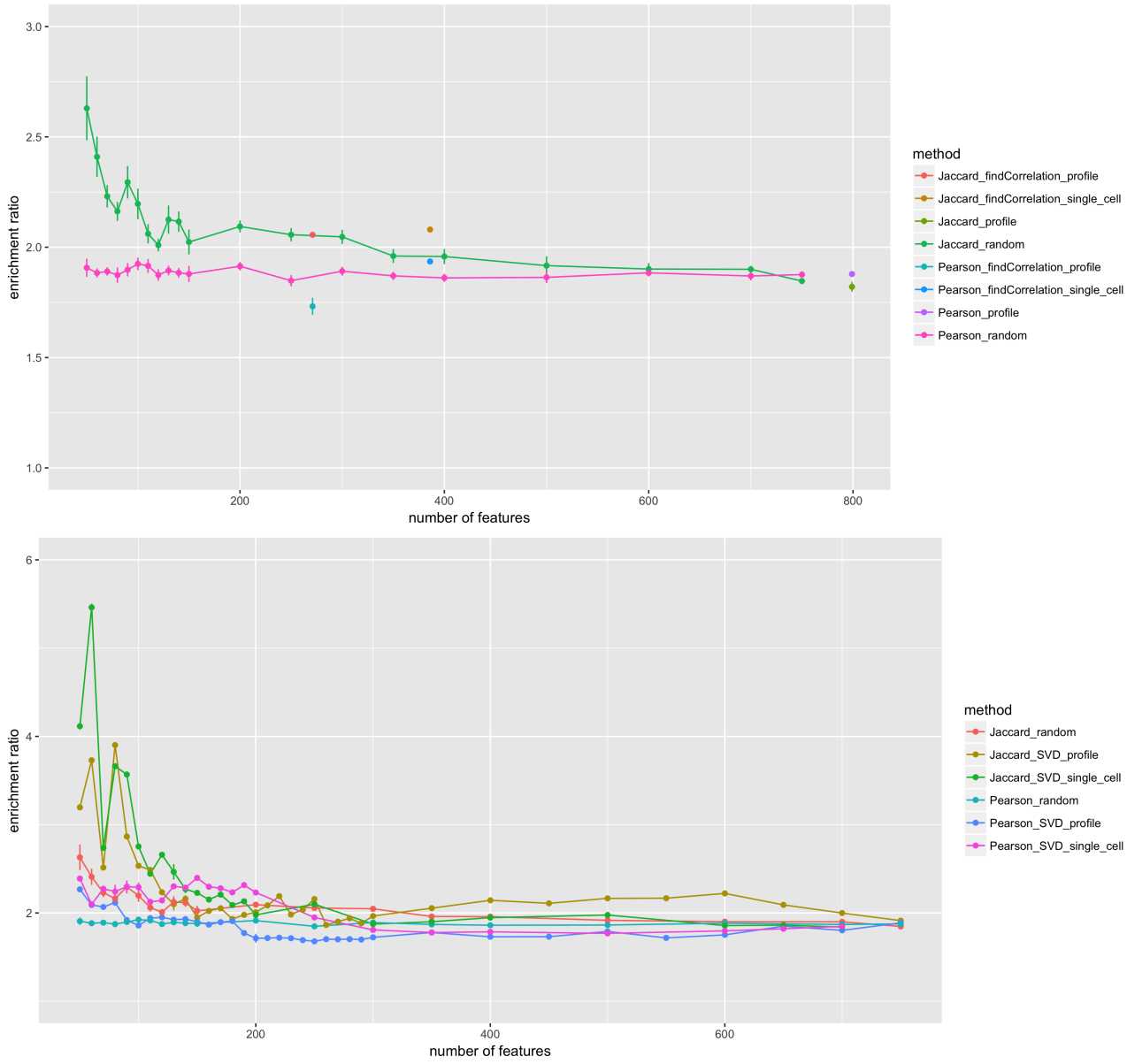


Figure 3.16: Enrichment ratio as a function of number of selected features in the BBBC022 dataset for the following methods: **Top**: no feature selection, findCorrelation at the single cell and profile levels, random feature selection. **Bottom**: SVD-entropy using FS2 on single cells and on profile levels, random feature selection. Evaluation were made with both Pearson's correlation and the Jaccard distance. The error bars are calculated based on 5 different simulations.

number of features, the enrichment ratio is highly increased. This stems from the fact that we have fewer hits, and the ones selected are probably the ones with a stronger phenotype. Moreover, we also observe the instability of the Jaccard metric for a low number of features. For example, if our dataset only has 50 features, selecting 5% of them, we will have 3 features in each set which makes the hit selection unstable. However, after around 200 features we can see a stability of all the methods. The Jaccard metric with SVD seems slightly better than random.

We replicated this analysis on the Repurposing dataset, expecting more promising results given the existence of more MOA annotations. However, before doing that, we wanted to determine whether the number of cells used at the single cell level influences the method of feature

selection. The hypothesis is that more cells will provide more information but also more noise. Also more cells means a larger matrix and longer time to run the analysis. The comparison between feature selection using 10,000 and 200,000 cells is shown in Figure 3.17. We can obviously see that, at least on this dataset, increasing the number of cells from 10,000 to 200,000 does not improve the results. However, in the future it would be interesting to look at the minimum number of cells needed to reach the same result.

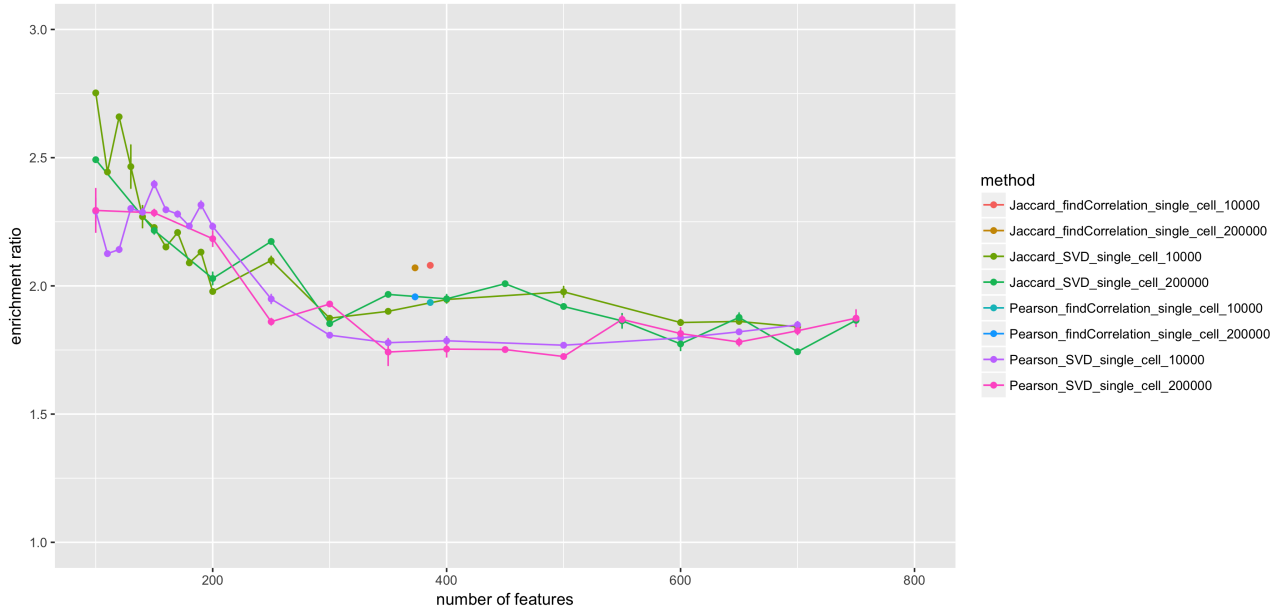


Figure 3.17: Enrichment ratio as a function of number of selected features in BBBC022 dataset for the following methods: SVD-entropy using the FS2 algorithm at the single cell level with 10,000 cells compared to 200,000 cells. Evaluation with both Pearson's correlation and the Jaccard distance. The error bars are calculated based on 5 different simulations.

One final analysis was done on BBBC022 dataset. We wanted to determine whether the sets of features that are selected with these two methods are capturing the same information. Specifically, we reported the number of common features in two methods, as well as the cosine similarity between the resulting compound correlation matrices.

First, we compared findCorrelation at the cell level, with a different number of initial cells (10,000 vs 200,000). The first observation was that only around 50% of the features were sharing the same name (186 out of 373). This should not be a problem, because some features are highly correlated with different names, and the method of findCorrelation is removing highly correlated features randomly. Moreover we obtain a cosine distance of resulting correlation matrices to be 0.9987.

Then, we compared findCorrelation and SVD-entropy (for 300 features in order to have a comparable dataset size) feature selection methods. We obtained that 248 out of the 386 features were in common. However, we obtained a cosine distance of 0.9682. This difference might explain the improvement in the odds ratio. Indeed, the features that are not in common, contrary to the previous case, might contain totally different information.

In order to validate the methods that were selected until now, we evaluated the two feature selection methods on the Repurposing dataset. First, we selected only the profiles of compounds at 10 mmoles/L stock concentration (1531 compounds). Figure 3.18 shows the enrichment ratio as a function of number of features that are selected in two methods. Eventually, all the methods are reaching the same point, around 1.8. First, we observe that random feature selection, which

is used as a baseline, decrease gradually with the number of features. This is likely due to the curse of dimensionality. As before, findCorrelation performed better at the cell level compared to the profile level. However, surprisingly it seems to perform a bit better with Pearson's correlation than the Jaccard distance.

The most interesting observation is that performance improves dramatically and consistently for the SVD-entropy feature selection (with SR). As said before, we could not perform the feature selection with FS2 because of high computational cost. Indeed, using this approach we get an improvement of enrichment ratio from 1.8 to around 2.4. Using the threshold of taking the features that have a positive gain of entropy, we should select 373 features. At some point we can see a big dropout of the curve. We tried to determine the cause of this drop, but it appears there were just a few features that are "harmful" to the data, meaning they come with a lot of noise. But it seems that the algorithm is capable of removing them up to almost 90% of features selected. We also noticed that 10 out of 20 of these "noisy" features are linked to granularity metrics. One explanation for this might be non-optimal scale parameters for the granularity features for this particular dataset/cell line.

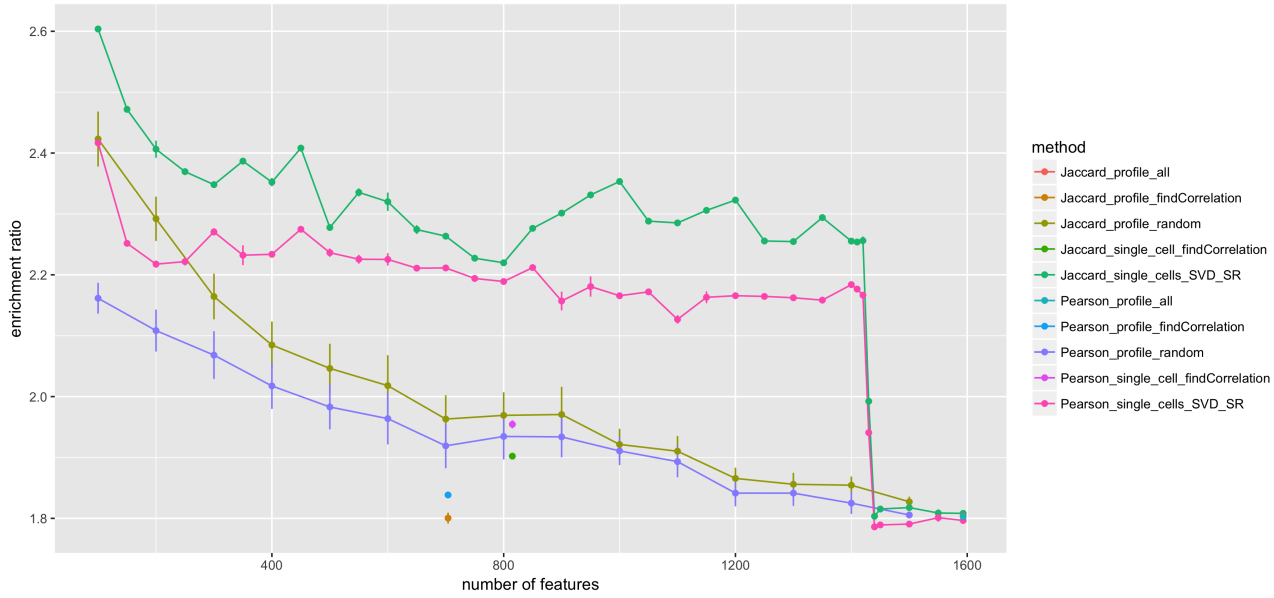


Figure 3.18: Enrichment ratio as a function of number of selected features in the Repurposing dataset at 10 mmoles/L dose for the following methods: findCorrelation on single cells and profiles, random feature selection and SVD-entropy using SR algorithm on single cells. These methods were evaluated based on both Pearson's correlation and the Jaccard distance. The error bars are calculated based on 5 different simulations.

We also tried to run the FS2 algorithm after applying findCorrelation, reducing the number of features to a size comparable to BBBC022. However, the results are worse compared to random feature selection. Indeed, when using findCorrelation a lot of information is lost and cannot be recovered (see Figure 3.19).

In order to confirm these results, we plotted the SVD-entropy SR feature selection method for each dose in Figure 3.20. Firstly, the SVD-entropy method combined with the Jaccard hit selection always gives the higher enrichment ratio for all doses. Secondly, we observe a dose response. With a higher dose, we get a higher enrichment ratio, meaning the phenotypes become stronger. Overall, this new method of combining the Jaccard distance with the SR SVD-entropy feature selection appears to be substantially and consistently improving the signatures compared to random feature selection or retaining all the features.

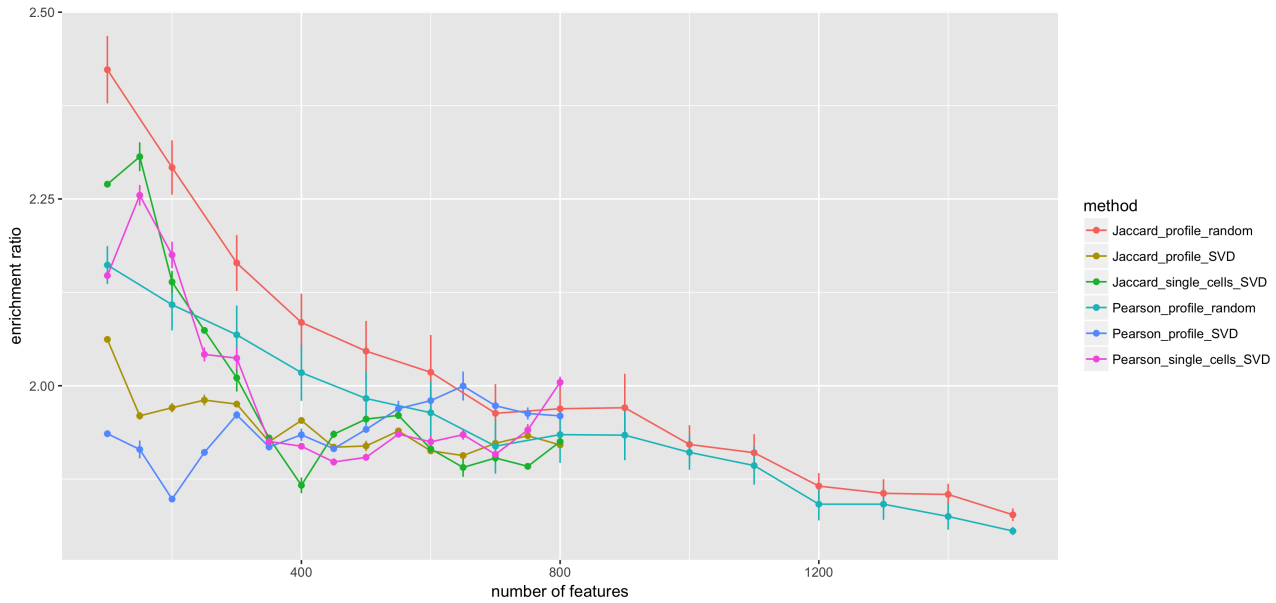


Figure 3.19: Enrichment ratio as a function of number of selected features in the Repurposing dataset at 10 mmoles/L dose for the following methods: random feature selection, findCorrelation at the cell level followed by SVD-entropy using FS2 for both cell and profile levels. These methods were evaluated based on both Pearson's correlation and the Jaccard distance. The error bars are calculated based on 5 different simulations.

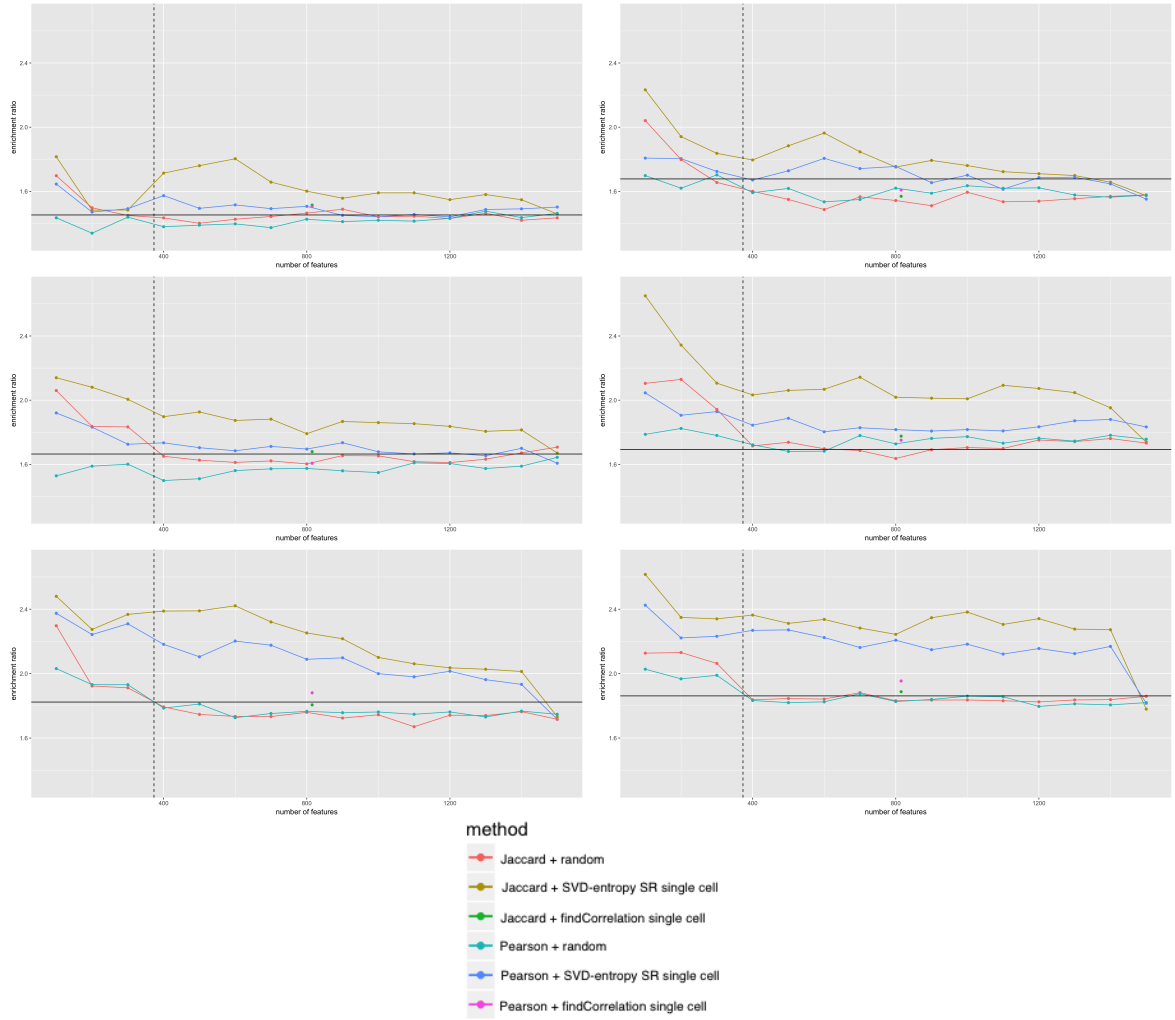


Figure 3.20: Enrichment ratio as a function of number of selected features in the Repurposng dataset for the following methods: random feature selection, findCorrelation and SVD-entropy using SR algorithm on single cells. These methods are evaluated based on both Pearson's correlation and the Jaccard distance. The analysis is performed on all doses: dose 1 top-left to dose 6 bottom-right: **Dose 1**: around 0.04 mmoles/L. **Dose 2**: around 0.1 mmoles/L. **Dose 3**: around 0.3 mmoles/L. **Dose 4**: around 1 mmoles/L. **Dose 5**: around 3 mmoles/L. **Dose 6**: around 10 mmoles/L. Legend: **horizontal black**: Pearson with FindCorrelation feature selection at the profile level (current pipeline). **Dashed black**: threshold at which the features contribute to the SVD entropy positively, which in this case is 373.

Chapter 4

Discussion

In this project, we first explored the problem of data quality. This process is very important and not straightforward. The Target-ID dataset is a good example of data that cannot be recovered. After a few tests on its quality we concluded that there was an error during the experiment and that nothing could be deduced from it. It is thus important to check the quality of the data to avoid a meaningless analysis.

The step of selecting hits can be seen as a quality selection. Indeed, only compounds that have reproducible signatures will be selected (all replicates are similar enough to one another). Without any feature selection, we could conclude that both Pearson's correlation and the Jaccard distance yield comparable results. At this point, it appeared as though the step of hit selection could not really be optimized. It was hypothesized that in order to see differences between the various metrics, we should first optimize the previous step in the pipeline which is the feature selection step.

SVD-entropy feature selection showed substantial and consistent improvement versus other (or no) feature selection methods. Using SVD-entropy, we then found the Jaccard distance metric seems to improve the enrichment ratio compared to Pearson's correlation metric. Reducing the feature space, thus removing noise and redundancy in the data, gives the metrics more ability to act depending on their properties. The Jaccard distance metric has the advantage of being less sensitive to outliers, because we are looking at the feature names and not their value. However, one of the drawbacks is that we need to set the size of features chosen as a subset. We found it convenient to set it as a percentage of feature set size, which we set to 5%. This quantity could be optimized in future work, which we kept constant in this study.

Moreover, if we wanted to have a more conservative hit selection, in order to select only compounds that are showing a phenotype different than negative controls, we could add a step using the profile Euclidean distance to the feature space origin for each sample (as explained in Section 2.2.1). We did not apply this method, partly to avoid the computational increase, but mainly because it may be too strict and we might also be interested in phenotypes that are weaker. Still, this step may be a useful quality check.

We should note here that although the Repurposing dataset has FDA approved drugs that generally should impact cells in some way, they do not necessarily have any target in the cancer cell line (A549) that we are using, or the cellular effect may not be detectable with the stains chosen. This could result in a compound with no signal that should be removed during the hit selection step. We also saw some examples of drugs that do not show any consistency in the

waterfall plot, where lines were randomly distributed.

One important aspect of all datasets that we analyzed is that they all have replicates for each compound. This gives us an insight on the intra-variability of each compound. Having replicates will reduce the number of false negatives (hits not detected as hits) and false positives because of less variability in the data, as explained in [17].

Having a profile for each replicate, we compared averaging the profiles to a single signature or taking all of them separately. It was difficult to hypothesize whether averaging would remove noise, or on the other hand induce a significant loss of information. We did not deeply explore this issue, but after a preliminary analysis, it seems that by reducing the data we are losing some important information. However, having a matrix that includes individual replicates, the computation time of the analysis is also vastly increased. At this early stage of the analysis, we concluded that the method is not worth the computational cost, but this might be further investigated.

After analyzing how to select the hits, we looked at the specificity of the profiles. We first clustered the compounds using an agglomerative clustering [13] [28]. We did not use k-means clustering, because k-means is more appropriate for circular shapes since it makes the assumption that the clusters are circular and it is also very sensitive to outliers. On the other hand, the agglomerative clustering is a non-parametric method; it is not very sensitive to outliers since they will just be singletons. But this method is computationally more expensive and it becomes impractical for large dataset. We also decided not to continue the evaluation based on the agglomerative clustering because it introduces hyperparameters which are difficult to set.

We considered two other methods to evaluate the signatures: one method looking at the top 2% of all pair-wise connections (leading to the odds ratio) and another one that considers the top 2% connections for each compound (leading to the enrichment ratio). We conclude that the enrichment ratio is more appropriate, because it considers all the compounds equally, even the ones with weaker connections. It is also easier to interpret from a practical point of view.

In the whole project, we used the MOA information as an annotation for each compound. Unfortunately these annotations are incomplete, particularly in the BBBC022 dataset. We lost a lot of compounds, because a lot of them are unannotated. Most of the project was done using the BBBC022 dataset, but at some point we realized that this dataset might not be the best for this analysis, not only because of the poor annotations but also because the number of features is lower than what the latest version of the pipeline gives. Also, the experimental part, the microscope and also the preprocessing pipeline have progressed since the data was created. For these reasons and also because we wanted to validate the results obtained so far, we shifted our analysis to the Repurposing dataset.

This dataset has the particularity of having six doses for each drug. In [17], the authors are mentioning the problem of the potency differences depending on the compounds being studied. It is difficult to determine what the optimal dose for each compound is. We have only compared compounds tested at the same dose, as if we had six different smaller datasets. But this is not optimal, and we should also look at the interactions between compounds at different doses.

We tried to find some possibilities to do that. We could look at the correlation for two compounds between all the doses and take the maximum of them. The issue here is that toxic phenotypes will be strongly correlated. We could consider taking the mean or the median, however this will probably not find the optimal dose. A better idea is to take the more extreme

of the 75% quantile and 25% quantile, because this removes the strongly correlated compounds, where the correlation is caused by toxicity or other irrelevant phenotypes. We did not have the opportunity to test any of these methods, and the problem of how to select the right dose is definitely not solved. However, we were able to have an overview on which dose is globally more relevant for each MOA using the waterfall plots.

Being able to evaluate whether the signatures are consistent with the MOA annotation, we finally tried to optimize the feature selection part of the pipeline. SVD-entropy showed better performance compared to findCorrelation algorithm. The advantage of this method is that it is looking at the contribution to the entropy for each feature, meaning that it considers the data as a whole, with all the feature interactions in deciding to include a feature or not. FindCorrelation is a greedy algorithm, and it has a complexity of $O(d^2n)$, where d is the number of features ans n is the number of observations. On the other side, SVD-entropy with simple ranking has a complexity of $O(d^4 + d^2n)$. The enrichment ratio using the SVD-entropy is 30% better but at a cost of a higher complexity.

There are also other feature selection methods that could be tested in future work. Some filter methods are based on spectral graph theory [32]. As an example [11] is an unsupervised feature selection based on the Laplacian score. It considers the locality preserving power: Two data points probably belong to the same class if they are close to each other. The method constructs a nearest neighbor graph and assesses how similar features are. We also investigated a hybrid technique (see Section 2.2.7) aiming at selecting the features by clustering using the Laplacian score [27]. The advantage of a hybrid method is that it is a compromise between efficiency and effectiveness. In this paper, the authors are using the Laplacian score as the filter part, in order to rank the features, followed by a wrapper part where they are using the Calinski–Harabasz index. This index measures the similarity within each cluster normalized by between cluster similarities. The ideal clustering will have a high cohesion inside a cluster and a high separability between the clusters. However, the efficiency of the wrapper part seems to be still too expensive.

For computational reasons we did not explore any wrapper feature selection method, but they may give better results and be worth analyzing. For example in [7], the authors present a method called FSSEM (Feature subset selection using expectation-maximization clustering). It aims at combining feature selection and clustering at the same time. The search algorithm iteratively adds one feature to the set of features and evaluates whether this feature increases a given criterion. On the other side, the clustering uses Expectation-Maximization clustering, with the underlying assumption that the clusters have arose by a Gaussian Mixture Models (GMM).

Many other alternatives not mentioned here could be tested. For example we could explore some deep learning techniques to extract and select some optimal sets of features. But the method that we used in this project, SVD-entropy feature selection, has the advantage of being independent of any learning algorithm and completely unsupervised. It is only based on the patterns found in the data.

To conclude, we should note that high-throughput experiments are used at the early stages of drug discovery. There are some benefits and some caveats. On the one hand, it is very fast, cheap and we can deal with many compounds at the same time in order to filter the most relevant ones. On the other hand, there are some parameters that we have to tune which might not be optimal for all compounds. We might filter relevant compounds that are just optimal at a different time point or concentration. But in the end, the results obtained by the profiling

pipeline, even if we do not necessarily have the perfect experimental settings, are helpful in reducing the number of drug possibilities, and can lead to promising follow-up experiments.

We tried to optimize the profiling pipeline as much as possible. We finally succeed in improving the enrichment ratio by around 30%. By tuning the parameters and through testing other datasets we could even further increase this improvement. However, we should realize that this pipeline is very complex. There are many steps and probably each of them could be improved. But we have to start somewhere and by using the SVD-entropy feature selection and the similarity metric based on the Jaccard distance, we have taken a step towards a better pipeline that will give a better signature, and thus a higher probability to discover new drugs.

Bibliography

[1] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016.

[2] Jonathan J Burbaum, Thomas DY Chung, Gregory L Kirk, James Inglese, and Daniel Chelsky. High-throughput assay, March 2 1999. US Patent 5,876,946.

[3] Juan C Caicedo, Shantanu Singh, and Anne E Carpenter. Applications in image-based profiling of perturbations. *Current opinion in biotechnology*, 39:134–142, 2016.

[4] 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, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

[5] Namjin Chung, Xiaohua Douglas Zhang, Anthony Kreamer, Louis Locco, Pei-Fen Kuan, Steven Bartz, Peter S Linsley, Marc Ferrer, and Berta Strulovici. Median absolute deviation to improve hit selection for genome-scale rnai screens. *Journal of biomolecular screening*, 13(2):149–158, 2008.

[6] Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of machine learning research*, 5(Aug):845–889, 2004.

[7] Jennifer G. Dy, Carla E. Brodley, Avi Kak, Lynn S. Broderick, and Alex M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE transactions on pattern analysis and machine intelligence*, 25(3):373–378, 2003.

[8] Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[9] Dirk Eddelbuettel and Conrad Sanderson. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, March 2014.

[10] 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, et al. Multiplex cytological profiling assay to measure diverse cellular states. *PloS one*, 8(12):e80999, 2013.

[11] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Advances*

in neural information processing systems, pages 507–514, 2006.

[12] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[13] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[14] Thouis R Jones, Anne E Carpenter, Michael R Lamprecht, Jason Moffat, Serena J Silver, Jennifer K Grenier, Adam B Castoreno, Ulrike S Eggert, David E Root, Polina Golland, et al. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences*, 106(6):1826–1831, 2009.

[15] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nat Methods*, 9(7):637, 2012.

[16] Betina Kerstin Lundholt, Kurt M Scudder, and Len Pagliaro. A simple technique for reducing edge effect in cell-based assays. *Journal of biomolecular screening*, 8(5):566–570, 2003.

[17] Nathalie Malo, James A Hanley, Sonia Cerquozzi, Jerry Pelletier, and Robert Nadon. Statistical practice in high-throughput screening data analysis. *Nature biotechnology*, 24(2):167, 2006.

[18] John H McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

[19] Serge Mignani, Scot Huber, Helena Tomas, Joao Rodrigues, and Jean-Pierre Majoral. Why and how have drug discovery strategies in pharma changed? what are the new mindsets? *Drug discovery today*, 21(2):239–249, 2016.

[20] F.P. Miller, A.F. Vandome, and M.B. John. *Kendall Tau Rank Correlation Coefficient*. VDM Publishing, 2010.

[21] Elizabeth Pennisi. Cell painting highlights responses to drugs and toxins. *Science*, 352(6288):877–878, 2016.

[22] Mohammad Hossein Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S Boehm, and Anne E Carpenter. Systematic morphological profiling of human gene and allele function via cell painting. *eLife*, 6:e24060, 2017.

[23] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2016.

[24] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191–200, 2012.

[25] A.K. Sharma. *Text Book Of Correlations And Regression*. DPH mathematics series. Discovery Publishing House, 2005.

[26] Shantanu Singh. Cytominer. https://github.com/cytomining/cytominer, 2017.

[27] Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. A new hybrid filter–wrapper feature selection method for clustering based on ranking. *Neurocomputing*, 214:866–880, 2016.

[28] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[29] Steve Weston. *Getting Started with doMC and foreach*. Vignette, CRAN, 2015.

[30] Mojca Mattiazzi Usaj, Erin B Styles, Adrian J Verster, Helena Friesen, Charles Boone, and Brenda J Andrews. High-content screening for quantitative cell biology. *Trends in cell biology*, 26(8):598–611, 2016.

[31] Roy Varshavsky, Assaf Gottlieb, Michal Linial, and David Horn. Novel unsupervised feature filtering of biological data. *Bioinformatics*, 22(14):e507–e513, 2006.

[32] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.