



universität  
wien

## MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

“Computational and experimental design and  
analysis of a compound library for high-content  
cell morphology imaging screens”

verfasst von / submitted by

Isabel Kaltenbrunner, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt /  
degree programme code as it appears  
on the student record sheet:

A 066 834

Studienrichtung lt. Studienblatt /  
degree programme as it appears  
on the student record sheet:

Masterstudium Molekulare Biologie

Betreuer / Supervisor:

Dipl.-Phys. Dr. Jörg Menche



## Acknowledgements

Zuallererst möchte ich mich bei Jörg Menche bedanken. Danke, dass du mich überhaupt in deine Gruppe aufgenommen und mir dieses Projekt zugetraut hast, obwohl ich absolut keine Ahnung vom Programmieren hatte. Danke für all deine Zeit, deine Ratschläge und vor allem den Mut, den du mir gemacht hast, wenn ich ihn gebraucht habe.

Ein besonders großes Dankeschön möchte ich außerdem an Michael Caldera richten. Du hast mir fast alles beigebracht, was zum Gelingen dieser Arbeit notwendig war und zwar mit einer unglaublichen Geduld, Hilfsbereitschaft und viel Verständnis. All deine Zeit und deine Mühe, die du mir entgegengebracht hast, weiß ich absolut zu schätzen und macht dich für mich sowohl fachlich, als auch menschlich definitiv zu einem Vorbild. Danke für alles :)

Also, special thanks to all other Menche group members for the unique and unforgettable time I had. I have never met a working group before that managed to combine smartness, craziness and kindness to such an extent as you do. Please, stay as cool as you are and take good care of my plants! :)

Vielen lieben Dank auch an meine Familie, vor allem für die finanzielle Unterstützung. Ganz besonders möchte ich mich bei meiner Mama bedanken. Danke, dass ich immer auf dich zählen kann, für dein offenes Ohr und dein Verständnis für alle Entscheidungen, die ich treffe.

Zuletzt, aber dafür umso mehr, möchte ich mich bei Christoph bedanken. Du hast mich mein ganzes Studium hindurch mit allen Mitteln unterstützt, mich ermutigt oder mir den nötigen Druck gemacht. Danke vor allem für deine Geduld und Ausdauer. Du und deine Unterstützung waren durchaus am Erfolg meiner Arbeit beteiligt und ich bin sehr froh, dich an meiner Seite zu haben.

## Abstract

Image-based cell morphology as readout provides huge potential for extracting multifaceted and complex information about biological processes that are taking place within a cell. Upon external or internal perturbations, cellular systems need to react to return to homeostasis, which often finds expression in phenotypical changes of the cell. Hence, morphological readouts are of great usefulness for identifying and investigating the impact of cellular perturbors, such as drug treatments, mutations or radiation. The overarching goal of this Master's thesis was to establish a compound library of chemical substances optimized for inducing strong morphological changes in high-throughput high-content imaging screens. We started by compiling a list of already known morphological perturbors by searching, characterizing and prioritizing compounds from the literature. We then screened more than 1400 drugs originating on the one hand from a commercial library from Sigma Aldrich and on the other hand from an in-house collection of small molecules. After incubation with these drugs, cells were stained using three different fluorescent dyes. A subsequent computational and statistical analysis on cell phenotypes was used to filter the large number of compounds to a small number of active hits that evoke a significant change in cell morphology, the so-called "CeMM Library of Morphologically active Perturbagens" (CLOMP). A following analysis of the 'CLOMP' shows its versatile properties, making it a useful resource for subsequent approaches, such as the investigation of drug effects and underlying cellular mechanisms. Moreover, its practical format allows downsizing on a single multiwell plate, which is particularly important for large-scale combination screens that address drug-drug or gene-drug interactions, in order to avoid combinatorial explosion.

## Zusammenfassung

Die Morphologie einer Zelle stellt ein sehr geeignetes Ausleseverfahren für verschiedenste bildbezogene Anwendungen dar, das es ermöglicht, vielseitige und komplexe Informationen über den biologischen Status einer Zelle zu gewinnen. Ist eine Zelle externen oder internen Perturbationen ausgesetzt, ist sie gezwungen, darauf zu reagieren, um in die zellinterne Homöostase zurückzukehren. Diese Reaktion äußert sich häufig in phänotypischen Veränderungen der Zelle. Morphologische Ausleseverfahren sind daher von großem Nutzen, wenn es darum geht, solche Störsubstanzen zu identifizieren. Das übergreifende Ziel dieser Masterarbeit war es, eine Wirkstoffsammlung zu erstellen, welche dahingehend optimiert ist, morphologische Änderungen in High-Throughput Screens hervorzurufen. Zunächst haben wir zahlreiche Substanzen aus bereits vorhandener Literatur ausgewählt, charakterisiert und gefiltert. Anschließend haben wir den Screen mit über 1400 Wirkstoffen durchgeführt, welche einerseits aus einer käuflich erwerbbarer Sammlung von Sigma-Aldrich stammen und andererseits der hausinternen Wirkstoffbibliothek entnommen wurden. Nach Inkubation mit den Substanzen wurden die Zellen mit drei unterschiedlichen Fluoreszenzfarbstoffen gefärbt und mikroskopiert. Durch die anschließende computerbasierte und statistische Analyse der Zellphänotypen konnte die große Anzahl an Wirkstoffen auf eine überschaubare Menge von Substanzen reduziert werden, welche signifikante morphologische Veränderungen hervorrufen und nun die neue Wirkstoffsammlung "CeMM Library of Morphologically active Perturbagens" (CLOMP) darstellen. Eine abschließende Charakterisierung der "CLOMP" zeigt ihre vielseitigen Eigenschaften auf, welche sie zu einer nützlichen Ressource für viele weitere Studien machen, wie beispielsweise die Untersuchung von Wirkstoff-Effekten oder zugrundeliegenden zellulären Mechanismen. Des Weiteren benötigt die CLOMP aufgrund ihres praktischen Formats lediglich eine einzige Multiwell-Platte. Dies ist vor allem für sehr umfangreiche, kombinatorische Screens von Vorteil, welche Interaktionen von Wirkstoffen mit anderen Wirkstoffen oder auch mit genetischen Faktoren behandeln und dadurch in kleinerem Umfang gehalten werden können.

# Contents

|          |                                                                                             |           |
|----------|---------------------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                                         | <b>8</b>  |
| 1.1      | Drugs – crucial tools in modern medicine . . . . .                                          | 8         |
| 1.2      | The history of drug discovery . . . . .                                                     | 8         |
| 1.3      | High-Throughput Screening . . . . .                                                         | 9         |
| 1.4      | The versatility of cell-based assays . . . . .                                              | 10        |
| 1.5      | High-Content Screening . . . . .                                                            | 11        |
| 1.6      | Common HCS-based libraries . . . . .                                                        | 12        |
| 1.7      | Aim and background of the thesis . . . . .                                                  | 12        |
| <b>2</b> | <b>Creation and characterization of a morphology-based compound library</b>                 | <b>15</b> |
| 2.1      | Browsing related publications for compiling an initial pre-selection of compounds . . . . . | 15        |
| 2.2      | The creation of the CLOUPAC compound library . . . . .                                      | 17        |
| 2.3      | A detailed characterization of the CLOUPAC . . . . .                                        | 17        |
| 2.3.1    | Collecting basic data about our compounds . . . . .                                         | 18        |
| 2.3.2    | ATC classification of the CLOUPAC . . . . .                                                 | 21        |
| 2.3.3    | Classification of CLOUPAC targets according to different databases . . . . .                | 23        |
| 2.3.4    | Outcome . . . . .                                                                           | 29        |
| <b>3</b> | <b>High-content high-throughput screen and image analysis</b>                               | <b>29</b> |
| 3.1      | Imaging screen of 1490 compounds . . . . .                                                  | 29        |
| 3.2      | Image analysis, quality control and normalization . . . . .                                 | 30        |
| <b>4</b> | <b>Acquiring an optimized morphology-associated feature set for evaluating drug effects</b> | <b>32</b> |
| 4.1      | Minimizing inter (I) and intra plate (II) effects . . . . .                                 | 32        |
| 4.2      | Calculating the effect size of features (III) . . . . .                                     | 33        |
| 4.3      | Filtering for robust replicates (IV) . . . . .                                              | 33        |
| 4.4      | Removal of redundant features . . . . .                                                     | 35        |
| 4.5      | Finding the best set of filters . . . . .                                                   | 36        |
| 4.6      | Outcome . . . . .                                                                           | 38        |
| <b>5</b> | <b>Selection of morphologically active compounds: the creation of the CLoMP</b>             | <b>39</b> |
| 5.1      | Reproducibility as a major criterion for our library . . . . .                              | 39        |
| 5.2      | Selection of strong morphological perturbors . . . . .                                      | 41        |
| 5.3      | Vector length vs. replicate correlation . . . . .                                           | 42        |

|          |                                                                                                             |           |
|----------|-------------------------------------------------------------------------------------------------------------|-----------|
| 5.4      | Removal of highly cytotoxic drugs . . . . .                                                                 | 43        |
| 5.5      | Outcome . . . . .                                                                                           | 45        |
| <b>6</b> | <b>The characterization of the CLoMP</b>                                                                    | <b>47</b> |
| 6.1      | Collecting basic data about the "CLoMP" . . . . .                                                           | 47        |
| 6.2      | Classification of CLoMP drugs according to the anatomical therapeutic chemical (ATC) code . . . . .         | 49        |
| 6.3      | Classification of CLoMP targets according to different databases .                                          | 51        |
| 6.4      | Concentration curves: the strength of the effect depends on the concentration of the applied drug . . . . . | 54        |
| 6.5      | Outcome . . . . .                                                                                           | 56        |
| <b>7</b> | <b>Summary and discussion</b>                                                                               | <b>57</b> |
| <b>8</b> | <b>Material &amp; Methods</b>                                                                               | <b>61</b> |
| 8.1      | High-throughput high-content imaging screen . . . . .                                                       | 61        |
| 8.2      | CellProfiler pipeline . . . . .                                                                             | 62        |
| 8.3      | Image quality control . . . . .                                                                             | 67        |
| 8.4      | Normalizing measurement results . . . . .                                                                   | 70        |
| 8.5      | Approach for finding the best subset of features . . . . .                                                  | 70        |
| <b>9</b> | <b>Supplementary material</b>                                                                               | <b>76</b> |
| 9.1      | Double occurring drugs . . . . .                                                                            | 76        |
| 9.2      | Classification of CLOUPAC . . . . .                                                                         | 76        |
| 9.3      | Morphology-associated features . . . . .                                                                    | 82        |
| 9.4      | Classification of CLoMP . . . . .                                                                           | 84        |

## Abbreviations

**ATC** Anatomical Therapeutic Chemical code.

**BSA** Bovine serum albumin.

**CLoMP** CeMM Library of Morphologically active Perturbagens.

**CLOUD** CeMM Library Of Unique Drugs.

**CLOUPAC** Combination of CLOUD and LOPAC<sup>®1280</sup>.

**DAPI** 6-Diamidin-2-phenylindol.

**DMEM/F12** Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12.

**DMSO** Dimethylsulfoxid.

**HCS** High Content Screening.

**HPA** The Human Protein Atlas.

**HTS** High Throughput Screening.

**LOPAC<sup>®1280</sup>** Library of Pharmacologically Active Compounds (Sigma Aldrich).

**MCF10A** Michigan Cancer Foundation-10A cells.

**PBS** Phosphate-buffered saline.

# **1 Introduction**

## **1.1 Drugs – crucial tools in modern medicine**

According to the DrugBank [1] or the ChEMBL database [2], almost 12.000 drugs exist, which found application in modern medicine. While this huge amount is to a small part made up of biologics, the vast majority, accounting for more than 80%, when we use statistics of the above-mentioned drug databases [1, 2], are described as small molecule drugs. Essentially, a small molecule drug can be described as a natural or artificially made chemical structure that is known to trigger a specific reaction within an organism. This can have toxic, addictive or therapeutic effects, meaning that it is used as a medicine for the treatment or the prevention of diseases [3, 4]. Drugs play a central and indispensable role in modern medicine that managed to achieve an extreme reduction in human mortality. In fact, chemical agents constitute the leading role in the treatment of diseases, for example antibiotics for treating bacterial infections or the usage of chemotherapy in cancer patients [5]. Nevertheless, modern drug discovery possesses a relatively short history of about 100 years.

## **1.2 The history of drug discovery**

At a time when Avogadro and Kekulé laid the foundation for chemical theory in the middle of the 19th century, no one could have anticipated that this was also the start of a new era revolutionizing the field of medicine and leading to enormous successes in drug discovery and development, as we know it today. Coal tar, which was originally a by-product of the industrialization, found use as dye in the organic chemical industry because of its high content of aromatic compounds. Paul Ehrlich, who was a medical student at this time, realized the therapeutic benefit of organic dyes, when he discovered that biological tissues could be selectively stained due to the presence of specific chemoreceptors, which was effectively the beginning of chemotherapy. Furthermore, coal tar served as the main raw material source for aromatic and aliphatic compounds, which became the basic elements of medicinal chemistry. Around the beginning of the 20th century, the field of pharmacology was initiated by Oswald Schmiedeberg. As a consequence, new interdisciplinary institutions were founded, which originally emerged from pharmacies, chemical or textile companies and which combined chemical knowledge with pharmacological studies and clinical sciences. It was the start of the now more than ever successful pharmaceutical industry [5, 6].

Within the first half of the 20th century, scientific fields, such as chemistry and biology grew slowly together, which proved to be successful for the development

of novel drugs and ultimately resulted in the creation of the new field of biochemistry [6]. Gradually, also technical advances complemented and improved the process of drug discovery, e.g. around 1912, when scientists achieved to reveal atomic structures of crystals by X-ray diffraction patterns [7]. About the same time, Joseph John Thompson developed the mass spectrometer [8]. In 1929, Alexander Fleming discovered penicillin, which was a fundamental step towards the treatment of bacterial infections [9]. Furthermore, the understanding of biological functions in conjunction with chemical compounds allowed to synthesize and improve drugs, e.g. by derivatization. Additionally, by the characterization of receptors of different organs, scientists were able to specify targets and to trigger them more precisely by drugs [6].

It was the second half of the 20th century when the young field of molecular biology has evolved to a scientific discipline having the focus on a deep understanding of diseases and their underlying genetic and chemical processes [6]. In 1953, Watson and Crick discovered the double-helix structure of DNA [10]. Over the decades that followed Watson's and Crick's discovery, many other key technologies developed towards the end of the century and found large use in the pharmaceutical industry, such as the production of recombinant proteins and the isolation of monoclonal antibodies. Scientists attempted to reveal new targets for drug intervention and to determine their exact mechanisms of action. The field of molecular biology became indispensable and contributed enormously to drug research [5, 6].

Ultimately, a completely new scientific era started shortly before the turn of the millennium, when computer processing power strongly increased. Consequently, new technologies, such as genomic sciences, DNA sequencing and especially high-throughput screens (HTS) [11] were introduced and reinvented the whole concept of drug discovery by generating huge quantities of biological data. HTS technologies lead to clear successes in productivity and proved to be of versatile use the pharmaceutical industry, as well as in academic research [12].

### **1.3 HTS: The future of modern drug discovery**

The process of finding drugs underwent a dramatic change, when scientists started to use large numbers of compounds for empirical studies instead of placing the focus on only few specific drugs. High throughput screens are dealing with whole libraries of drugs that are tested in functional assays with the aim of identifying substances that show a specific, desirable effect, referred to as 'hits' and which will further be examined in their suitability for pharmacological intervention. Typically, such screens are conducted using state of the art robotics allowing a

fast and precise conduct of multitudinous pharmacological cell-based tests at the same time [6, 11, 13]. Meanwhile, high throughput assays can be considered as a standard application within the pharmaceutical industry for identifying potential drug candidates as shown by the fact that HTS accounted for almost one third (19 out of 58 drugs) of all drugs that were approved between 1991 and 2008 [12, 14].

## **1.4 The versatility of cell-based assays**

There are two main types of assays that are routinely performed in association with HTS: biochemical and cell-based assays. Historically, biochemical assays are older and serve for the investigation of enzymatic, receptor-ligand binding, ion channel or nuclear receptor studies, as well as protein-protein interactions. Typically, those kinds of assays are specific to certain targets, which need to be purified prior to the screen. In contrast, the molecular target is in most cases not known, when cell-based assays are performed, where the aim is rather to identify physiological modulators of a specific pathway. In high-throughput format, cell-based assays can be combined with various kinds of measurements, such as cell proliferation or drug toxicity, respectively, expression of reporter genes, cell motility, activation of signaling pathways, cellular morphology and others. However, the great advantage of cell-based screens is the potential of revealing multiple intervention points of pathways, since the approach is not restricted to a very specific target as it is usually the case in biochemical screens [12, 13].

Cell-based assays have proven to be successful approaches for evaluating various scientific issues. For example, with the exception of antineoplastic or antibiotic drugs, which are usually designed for selectively harming cells, cytotoxicity is one of the major criteria for excluding drug candidates in the drug discovery process. Thus, many cell-based assays address the effect that a substance has on the proliferation or viability of cells, which eventually leads to cell death. In this context, single parameter readouts are commonly used for determining the number of viable cells remaining [15]. This was done for example in a previous combinatorial compound screen at CeMM [16] via a CellTiter-Glo<sup>®</sup> assay (Promega). This type of assay quantifies the amount of ATP present after the drug incubation for evaluating the amount of metabolically active cells. Apart from further assays based on the metabolic activity of cells [17], other readouts involve dye exclusion or DNA-condensation studies, which allow to distinguish between live and apoptotic cells via a selective staining of dead cells or the compactness of the stained nucleus, respectively [15].

Despite the simplicity and the great usefulness of viability-based assays, these kinds of readouts provide only poor information about the cellular state. It is

evident that a cell is capable of much more than living and dying. Combining HTS with microscopy offers us a wide-ranging selection of parameters to be measured, such as the localization of cells, the occupied area, size and shape of individual cells or measurements of the cellular texture, just to name a few. Imaging thus allows for the extraction of much more information and hence leads to a more detailed classification of drug effects compared to other readout methods [18, 19].

## **1.5 High content screening (HCS) - Cell morphology as a read-out**

In recent years, HCS has shown to be a powerful tool in drug discovery, enabling not only to characterize preclinical drug candidates in the pharmaceutical industry, but also for many applications in academic research, where HCS is used for studying diverse biological systems. The ‘high content’ part refers to the large amount of phenotypical data that is generated typically by automated fluorescence microscopy using fluorophores of different wavelengths [19, 20].

Several studies [21, 22, 23] of cells that were exposed to chemical stimuli led to the observation of drug-specific responses in various ways. This can be explained by the fact that drugs act as perturbations of the cellular system that change or destabilize its homeostatic state, resulting in a rearrangement of active and inactive pathways and gene expression patterns. If this transition affects the cellular morphology, we can eventually draw conclusions about a compound’s properties. Thereby, the specific area of interest has to be stained, which is achieved by the usage of specific antibodies targeting for example the cytoskeleton in case the whole cell should be visualized. In contrast, staining of the nucleus is usually done by fluorescent dyes, such as DAPI (4’,6-diamidino-2-phenylindole) [24] or Hoechst [25]. In addition, there are organic dyes available that can be applied on living cells and accumulate within specific organelles, such as the cytosol, mitochondria, the endoplasmatic reticulum, the Golgi apparatus and lysosomes [19, 26]. By the usage of different fluorophores, multiple fluorescent dyes or antibodies can be applied simultaneously and thus yield various layers of information by imaging the cells in different channels. Consequently, the phenotypic readout data allows us for example to associate cell morphologies with underlying mechanisms of action of the applied drug and thus help in the process of revealing novel drug candidates [19, 27, 28].

Numerous scientific publications use morphological perturbations of cells to uncover biological questions by screening a wide variety of substances [27, 29, 28, 23, 30] and are meanwhile based on HTS-HCS image data acquisition using hun-

dreds to thousands of compounds, followed by the extraction of quantitative measurements of the treated cells with the help of image analysis software. As mentioned above, also within the pharmaceutical industry, high-throughput screens of small molecule libraries combined with high-content imaging are meanwhile commonly used as a promising strategy for finding new drugs. The reason for this is the high amount of information that can be obtained from digital images, thus making the process of lead optimization more efficient [18]. Indeed, phenotype-based screens led to a significantly higher amount of FDA approved drugs between 1999 and 2008 than target-based screens [31]. Ultimately, the interconnection of HTS and HCS succeeded “to bridge the gap between depth and throughput of biological experiments” [19].

## **1.6 Common HCS-based libraries**

Despite the frequent usage of HCS in today’s research, most libraries used in such studies [27, 29, 23, 30], although sometimes being very large, typically have only a low percentage of morphological shaping hits. Consequently, those libraries are not well suited for systematic studies of cell morphology, since they would result in a very limited number of hits. Moreover, in most cases the used compounds are not well characterized and often lack even basic information. Hence, the usage of such libraries would mean to treat cells with a high variety of drugs without knowing how they are classified, which proteins they are targeting, if they have been approved and sometimes there is not even information about the appropriate concentration to be used.

## **1.7 Aim and background of the thesis**

The aim of this thesis was the compilation of a drug library, named the ‘CeMM Library of Morphological Perturbers’ (CLOMP), which comprises exclusively chemical compounds evoking cellular perturbations that result in statistically significant morphological changes of the treated cells. This ready-to-use library should facilitate the performance of further high-through-put high-content screens, such as drug-drug or gene-drug interaction studies, for example.

Figure 1 gives an overview of the approach we have taken to achieve this. We started by performing a high-throughput high-content imaging screen on human MCF10A cells, which included 1490 distinct compounds that were extensively characterized in advance. For the screen, each drug was used in four different concentrations in order to find the optimal amount of drug that leads to an observable effect (Figure 1 (a)). Furthermore, we were able to generate a high amount

of information by fluorescently labelling the cytoskeleton, the nucleus and the mitochondria of the cells with different fluorophores. The acquired raw image data was subsequently analyzed by using the image analysis software CellProfiler[32], which returned 638 distinct features (Figure 1 (b)) that were reduced to the most informative ones. We used multidimensional feature vectors to describe the individual drug perturbations and to efficiently validate and analyze them. Ultimately, we incorporated only those perturbors into our library that met all defined criteria, such as robust replicates, strong morphological effects and a low cytotoxicity, leading to the final establishment of the CLoMP (Figure 1 (c)). The CLoMP consists of 88 unique drugs that are fully annotated, including their ATC classification, the proteins they are targeting, and the strength of their morphological effect over increasing concentrations. Its compact format allows to accommodate the whole library on a single multiwell plate, which makes it especially easy to handle.

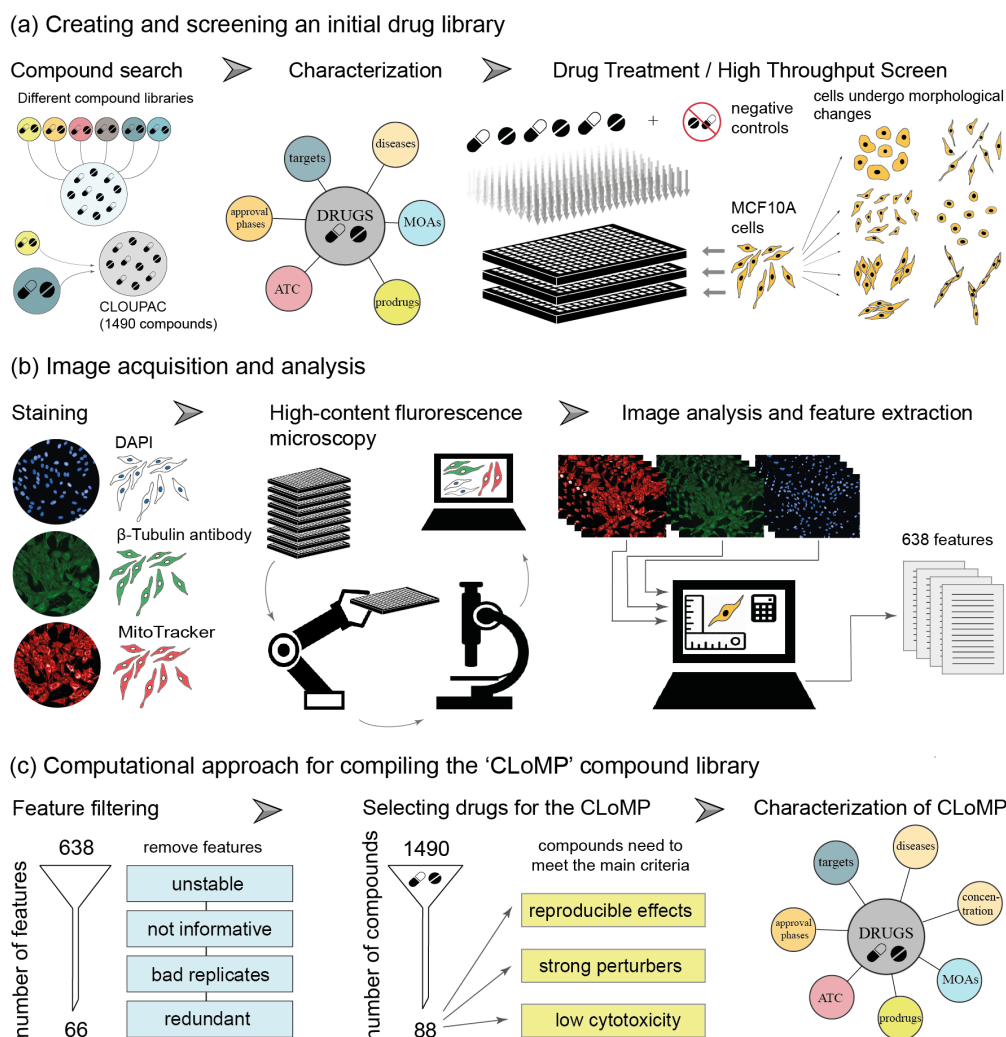


**Figure 1: Pipeline for compiling and characterizing the novel morphology-based compound library CLOMP (CeMM Library of Morphological Perturbagens).** (a) Basic approach of establishing and screening a preliminary compound library. After searching through various publications, we decided for an initial selection of 1490 compounds, which we annotated and used for treating MCF10A cells in a high-throughput imaging screen. The treatment of some drugs induced morphological changes of MCF10A cells. (b) Generation of raw imaging data by automated high-content fluorescence microscopy, followed by image analysis and feature extraction. The stained nucleus, cytoskeleton and mitochondria were imaged in three different channels, whereby large quantities of data were generated and analyzed by the software CellProfiler, returning 638 distinct features, describing the morphology of cells. (c) Extraction of suitable features for the subsequent creation of the CLOMP, resulting in 66 measurements. We then selected the final perturbors, representing the CLOMP, which meet the most important criteria and characterized them in detail.

## 2 Creation and characterization of a morphology-based compound library

Our main purpose of creating the novel compound library was to establish a specific collection of drugs that represents a preferably high diversity in terms of drug targets and mechanisms of action and simultaneously comprises a variation of agents having distinct effects on the cell morphology. Therefore, we collected an initial pool of compounds, which we subsequently characterized, screened and reduced to a final selection of drugs that showed to be morphologically active. In analogy to the existing CLOUD (CeMM Library of Unique Drugs), which is a representative CeMM-established collection of biochemically diverse and clinically relevant drugs [16], the newly compiled compound library was named the CLoMP (CeMM Library of Morphologically Active Perturbagens). In addition, we aimed to collect as much information about the CLoMP drugs as possible in order to gain insight into their basic properties as well as an overview about their targets. This characterization was mainly done via the ChEMBL database and will be described in more detail in the following section.

### 2.1 Browsing related publications for compiling an initial pre-selection of compounds

For creating the CLoMP, we first had a closer look into already existing drug libraries providing large-scale microscopy data that allowed us to select potentially interesting components evoking an obvious morphological perturbation of the cell. The publications as well as the number of hits derived from each source are depicted in Table 1.

One of our sources was the massive imaging dataset provided by the Broad Institute of Harvard and MIT that was created by using the Cell Painting protocol [29]. Over 30.000 distinct compounds were tested on two different cell lines for profiling the morphological phenotypes of the cells. The provided morphological features that were measured on these images using CellProfiler [32] allowed us by statistical validations, referring to DMSO-treated controls, to establish a ranking of all perturbations. The 55 compounds with the highest number of features changed in comparison to DMSO were considered as interesting morphological perturbagens.

Another publicly available resource by the Broad Institute[30], referred to as the L1000 compendium, provides genetic, as well as pharmacologic cell profiles from the treatment with around 20.000 small molecule compounds. By studying this

data in a similar way as in the Cell Painting protocol mentioned above, another 20 morphologically interesting perturbors were shortlisted.

A publication from Ljosa et al. [23] describes a comparative analysis of different image-based profiling methods investigating the cellular morphology upon treatment with 113 cancer-relevant cytotoxic compounds. This specifically selected collection of apparent morphological perturbors extended our drug compendium by 6 drugs.

Furthermore, we included 52 approved drugs from a database that was constructed as an encyclopedia of cellular phenotypes, the so-called 'Morphobase' [28], comprising approximately 200 "routinely used chemicals on morphology".

As another potentially interesting source for morphologically active drugs, we considered the imaging results of a chemical-genetic interaction screen. The 'Pharmacogenetic Phenome Compendium' (PGPC) was generated by screening the whole LOPAC<sup>®</sup>1280 library from Sigma-Aldrich in isogenic HCT116 cell lines, followed by nuclear and fluorescent staining [27]. Using the provided data, the perturbagens were again ranked, resulting in a list displaying several compounds that modify the morphology stronger than DMSO-treated controls. By this, we found 58 compounds that seemed relevant to us.

Finally, we selected 22 of the CLOUD drugs that clearly induced phenotypical changes in a previous screen [16].

**Table 1: List of literature sources used for compiling an initial collection of drugs that are acting on the cell morphology.** The numbers indicate the amount of compounds that we included from the respective publications. Drugs in several sources were counted only once for the final screen.

| Source                                    | Number of hits |
|-------------------------------------------|----------------|
| 30K library (Bray et al.)                 | 55             |
| L1000 compendium (Subramanian et al.)     | 20             |
| Set of cytotoxic compounds (Ljosa et al.) | 6              |
| Morphobase (Futamura et al.)              | 52             |
| PGPC (Breinig et al.)                     | 58             |
| CLOUD (Licciardello et al.)               | 22             |
| Sum                                       | 173            |

Table 1 summarizes the sources that we used for our compound search. Many of the collected drugs appeared in two or more of the above mentioned publications and were therefore only included once. In this way, we established a list of 173 drugs with prior evidence for acting on the cellular morphology.

## **2.2 The creation of the CLOUPAC compound library**

Upon closer inspection of these compounds, however, we realized that due to lacking commercial availability, higher prizes and difficulties in handling and storing single compounds, the alternative of purchasing a prepared drug library, as provided by several companies, seemed to be more advantageous for our purpose. By comparing several purchasable drug collections, the above-mentioned LOPAC<sup>®1280</sup> ('Library of Pharmacologically Active Compounds') from Sigma-Aldrich, appeared to be the most suitable one, since it was cost-efficient and there was an overlap of 114 drugs between our established list and this library, which is the highest overlap that we observed with any of the purchasable libraries.

The LOPAC<sup>®1280</sup> is described as a collection of 1280 pharmacologically active compounds and impacts a high diversity of signaling pathways and drug target classes [33]. We therefore decided to screen the whole library.

In addition, we decided to include most of the 267 CLOUD drugs into our primary drug screen, which indicated in a previous analysis [16] to induce morphological changes of the cells. Some of these CLOUD drugs were already present in the LOPAC<sup>®1280</sup>. We therefore took the union of all CLOUD and LOPAC<sup>®1280</sup> compounds. Additionally, we kept eight drugs, which have the same chemical structure, but differ in their salt form. Effectively, we used 210 CLOUD drugs and obtained in this way a final number of 1490 compounds. Table 8 (Supplementary material) lists all drugs that appear twice within the CLOUPAC library.

## **2.3 A detailed characterization of the CLOUPAC**

In order to increase the probability of finding drug candidates within the large selection of drugs, which are able of altering the cell shape, a certain diversity of properties among these compounds should be given. This helps not only to uncover morphologically active compounds, but also leads to a higher potential of resulting in a series of different cell morphologies. Such properties refer to various drug classes, mechanisms of action or drug targets. Before performing the drug screen, we therefore analyzed a variety of characteristics of the whole CLOUD-LOPAC<sup>®1280</sup> ('CLOUPAC') library with the aim of being able to make

statements about the diversity of our source material.

The usage of the ChEMBL database [2], which is maintained by the EBI (European Bioinformatics Institute) and belongs to the EMBL (European Molecular Biology Laboratory), allowed us to gather many kinds of information about our compounds via an application programming interface (API) by using the ChEMBL ID of the respective drug. The database represents a large collection of biological and chemical data including numerous bioassay measurements covering hundreds of thousands of compounds.

In addition, the database is not only accessible via a web interface, but also provides a webresource client that is written in the Python programming language and can be installed and used as a package. This allowed us to access information about various drug properties in a simple and fast way. In the following, we describe the results of this detailed characterization of our drugs.

### **2.3.1 Collecting basic data about our compounds**

Our screening library comprised the full LOPAC<sup>®1280</sup> drug collection of 1280 drugs and 210 potentially interesting CLOUD drugs. Initially, every single drug name, which was provided by both libraries, had to be translated into a ChEMBL ID by an online translation service [34]. This was a necessary step to access the ChEMBL database via the webresource client. By doing so, we found a corresponding ID for 1021 LOPAC<sup>®1280</sup> drugs and 197 CLOUD drugs, resulting in 1218 out of all 1490 compounds, meaning that for those drugs further information could be collected. For the remaining compounds we could not find a ChEMBL ID. There could be two reason for this: either these substances are rather unknown and not present in the ChEMBL database or there are alternative names of these compounds, which do not match those in ChEMBL. Therefore, we could not include them into our drug characterization.

### **Approval stages of our drugs**

We started by gathering information about the phase of approval of each single drug, which indicates the current state of development of a particular compound. In other words, a newly discovered compound has to undergo a long process of decisive tests and has to be substantially annotated with pharmacological data before being released to the market. The main part of this procedure consists of pre-clinical and clinical trials, where the compound is tested on healthy people and ultimately on patients. This encompasses phase 0, I, II and III trials. Phase 0 studies are usually carried out in patients and have an exploratory character. They serve to assess how a new drug behaves in the human body or, for example, if

it reaches the tumor in the first place. In phase I, the general safety of a drug is tested, usually on a small group of healthy people. This includes the identification of side effects and finding an appropriate dose of the treatment. If a drug reaches phase II trials, it is tested in several hundred patients, which have a certain condition that should be treated with the new drug. This helps to further evaluate the effectiveness and safety of the treatment. Finally, if the drug passes the previous tests, phase III trials are carried out in a large group of patients for collecting even more information about the effects of a treatment. In this period, the new drug will be compared with already existing or placebo treatments [35].

After this excessive and time-consuming assessment of the drug, the compound will finally be approved for clinical use by the regulatory authorities, such as the EMA (European Medicines Agency) or the FDA (U.S. Food and Drug Administration), if all necessary requirements were met. In our statistics, approved drugs will be equated with phase IV trials, which are conducted after the approval of the compound and serve as surveillance studies or for tracking risks and benefits of the treatment, respectively [35].

In sum, we found entries about the approval state for 1216 drugs in the ChEMBL database and categorized them into the different phases of clinical trials (Figure 2). Slightly more than a third of the CLOUPAC library was categorized as phase IV or ‘approved drug’. Only few drugs were in phases I-III of clinical trials. The majority was indicated as phase 0 (for see Table 9).

### **The proportion of prodrugs within the CLOUPAC**

Next, we were interested in how many prodrugs we have among our 1490 molecules. A prodrug is described as a pharmacologically inactive compound by itself that needs to be metabolized within the body to be chemically or enzymatically transformed into its active form [36]. The major intention of establishing prodrugs is to improve the physiochemical properties of a compound, which define its main characteristics: absorption, distribution, metabolism, excretion and toxicity - in basic terms the whole biopharmaceutical performance of a drug [36]. Thus, including prodrugs into our screen would in principle mean to treat some of our cells with non-active compounds since they lack the activating chemical environment or the enzymes that are needed.

The ChEMBL database declared 38 CLOUPAC drugs as prodrugs, which represent an interesting aspect in studying morphological cell shape changes, since they are not expected to have any effect on the cell phenotype. In contrast, the majority of all annotated compounds in our library, namely 421, were found to be non-prodrugs. For 1031 compounds this information was not provided by ChEMBL.

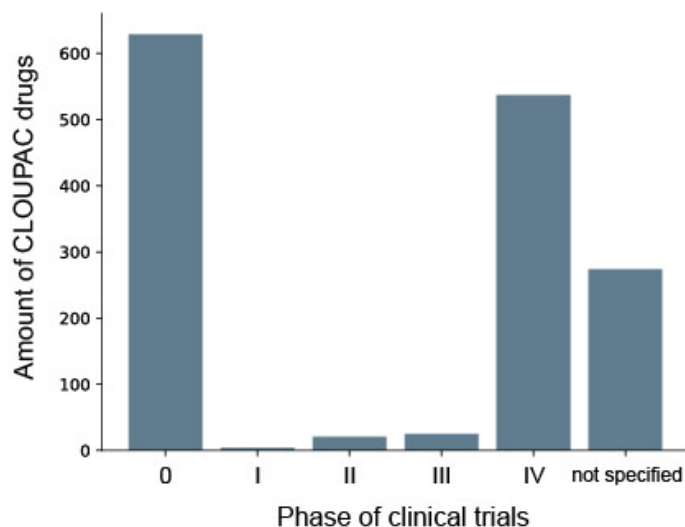


Figure 2: **Distribution of CLOUPAC drugs into phases of clinical research.** The majority of CLOUPAC drugs is either at the very beginning of clinical trials (phase 0) or already in clinical use (phase 4).

### Disease associations and mechanisms of action

Next, we searched for known drug targets, mechanisms of action (MOA) and associated diseases of our drug collection. We could find presumable targets for 506 drugs, recorded MOAs for 396 drugs and associated diseases for 422 drugs. Table 2 summarizes the different entries provided by ChEMBL and the number of drugs falling into the respective category. The target proteins of our compounds could in turn be classified and divided into different functional groups depending on the database, which will be described more precisely below.

### LOPAC<sup>®1280</sup>-provided data

The large majority of our compounds originate from the LOPAC<sup>®1280</sup> drug compilation and was equipped with some basic data about each drug that aids in evaluating the content of the library. This data includes an overview about the proportion of drugs that are targeting an enzyme. Moreover, targets were assigned to different protein classes, which will be further discussed below.

Furthermore, all drugs were divided into classes that roughly indicate the function of the drugs, such as receptors or ion channels. Thereby, we found neurotransmitter to be strongly represented, meaning that there are 468 entries for neurotransmission associated drug classes, such as ‘Adrenoreceptor’, ‘Cholinergic’, ‘Dopamine’, ‘GABA’, ‘Glutamate’, ‘Serotonin’ or a general class of ‘Neurotransmission’. Other drug classes that were ranked among the most abundant ones in-

Table 2: **Different kinds of basic information about the CLOUPAC drugs according to the ChEMBL database.** For some of the CLOUPAC drugs there was data available about whether they are acting as a prodrug or not, about their targets, their MOAs or disease indications.

| Entry in ChEMBL | #CLOUPAC drugs having this data available |
|-----------------|-------------------------------------------|
| Prodrugs        | 38                                        |
| Non-Prodrugs    | 421                                       |
| Not classified  | 1031                                      |
| Targets         | 506                                       |
| MOA             | 396                                       |
| Disease         | 422                                       |

clude ‘Adenosine’, ‘Apoptosis’, ‘Biochemistry’, ‘Gene Regulation’, ‘Histamine’ or ‘Ion channels’, just to name a few. Table 10 and 11 list all LOPAC<sup>®1280</sup>-defined drug classes and the respective number of LOPAC<sup>®1280</sup> drugs.

In some cases, the LOPAC<sup>®1280</sup> drugs were in addition equipped with some data about their selectivity, which indicates some information about the exact name of the target, its function or the associated pathway. Moreover, it is noteworthy that every single compound of the LOPAC<sup>®1280</sup> library possesses a short description of what effect it has in a biological system, meaning that for each single drug there is at least some indication of its mechanism of action, even if those drugs are neither approved, nor is there any information available in the ChEMBL database.

### 2.3.2 Classification of CLOUPAC drugs according to the Anatomical Therapeutic Chemical code

In some cases, ChEMBL provides information about the anatomical therapeutic chemical (ATC) code of a drug. This classification system was established by the World Health Organisation (WHO) Collaborating Centre for Drug Statistics Methodology as a stable classification of drugs, which is primarily related to studies about drug utilization monitoring and research. The ATC code describes a compound by five hierarchical levels. The first level comprises fourteen different letters representing the main anatomical groups. All sublevels give a more and more precise classification of the drug, including therapeutic, pharmacological and chemical subgroups. Note that one drug is not necessarily bound to a single

anatomical group, but can as well be described by various ATC classes, “if it is available in two or more strengths or routes of administration with clearly different therapeutic uses” [37].

In our validation of the CLOUPAC, we only focused on the first levels of ATC classification to get a rough overview about the compound distribution into the different categories. In doing so, the whole CLOUPAC, but as well the CLOUD and the LOPAC<sup>®1280</sup> libraries were analyzed separately in order to gain insight into how divers the individual drug collections are. We counted all occurring letters in case a drug was categorized multiple times, but we used the same letter only once, if the same first level classification appeared more often. In this way, we found that most of the used CLOUD drugs, namely 150 out of 210 (71.4%), were provided with an ATC classification, whereas only a rather small amount, 221 out of 1280 (17.3%), of the LOPAC<sup>®1280</sup> library possessed this kind of information.

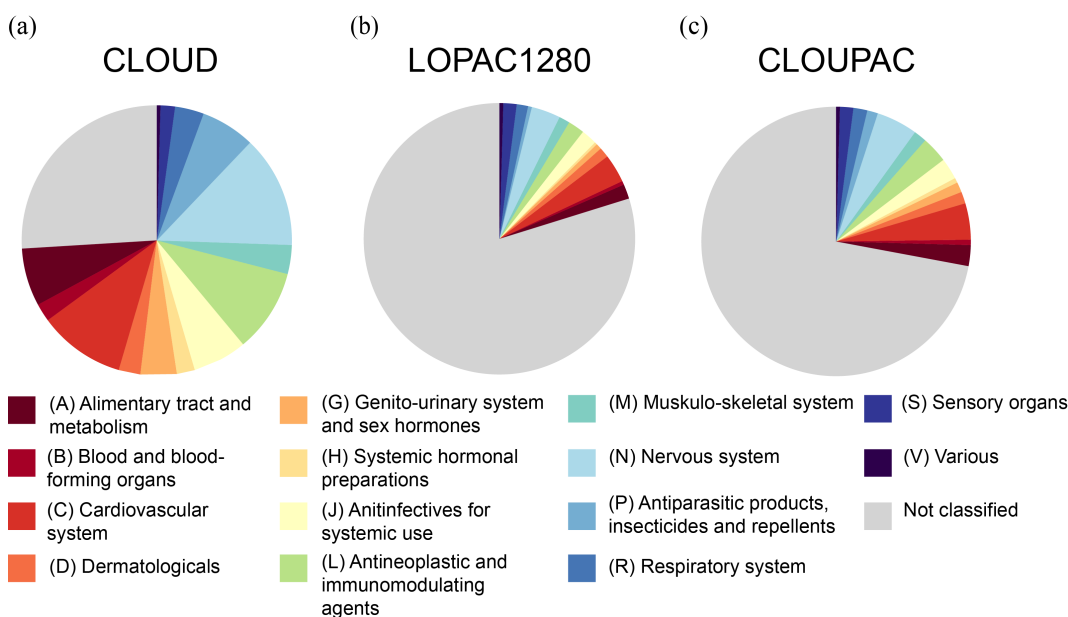


**Figure 3: Comparison of the overall ATC classification of CLOUD, LOPAC<sup>®1280</sup> and CLOUPAC drug libraries.** While the (a) CLOUD consists dominantly of well-known and divers drugs, the (b) LOPAC<sup>®1280</sup> library can be considered as a rather explorative collection of compounds. Both libraries comprise a high variety of ATC classes. In (c), both libraries were merged.

Summarizing, we can conclude that the CLOUD can be described as a collection of well-known and well annotated drugs, whereas the LOPAC<sup>®1280</sup> represents a rather explorative library, including mainly drugs with almost no data available. However, both, CLOUD and LOPAC<sup>®1280</sup>, cover all fourteen ATC classes and

thus exhibit a desirable diversity of their compounds. When both libraries were merged, or respectively, speaking of the 1490 described CLOUPAC drugs, a percentage of 24.7% ATC-classified compounds (368 of 1490) could be achieved.

Figure 3 illustrates the ratios of the ATC classes as pie charts, also indicating the ratio of classified and unclassified drugs within the libraries. For absolute numbers of compounds within the fourteen anatomical groups, see Table 12 (Supplementary material). In both libraries, and consequently also in the merged CLOUPAC collection, nervous system drugs (N) are the leading group of compounds, followed by antineoplastic and immunomodulating agents (L) in case of the CLOUD, and cardiovascular system drugs (C) in the LOPAC<sup>®1280</sup> library.

In general, all classes are relatively well represented with the exception of ‘various’ drugs, which are for example antidotes or detoxifying agents. These classes of compounds are strongly underrepresented, but since this group summarizes only those drugs that do not belong to any of the other ATC groups, they can be considered to be of minor importance.

### 2.3.3 Classification of CLOUPAC targets according to different databases

A further aspect that indicates a certain degree of diversity within the CLOUPAC library are the proteins targeted by the drugs. Usually, a drug is designed to selectively target a certain binding site of a molecule, which can range from proteins to DNA or RNA, and leads to a specific response upon activation or deactivation of this molecule and its associated pathways. However, sometimes the target is not exactly defined or there are several targets involved. Hence, the number and nature of targets that are listed in conjunction with a compound can vary, depending on the database or resource. In this section, mainly two databases will be assessed and compared in their target classification, which comprise the already introduced ChEMBL-EBI resource and the druggable proteome that is part of the Human Protein Atlas (HPA) project. Additionally, there is also some information provided by the purchased LOPAC<sup>®1280</sup>.

#### Target classification of the LOPAC<sup>®1280</sup> drugs

One type of data about the LOPAC<sup>®1280</sup> drugs provided by Sigma-Aldrich covered the information of whether the drug’s main purpose is to target an enzyme or not and revealed that 408, almost a third (31.9%), of the LOPAC<sup>®1280</sup> compounds, are altering the activity of enzymes. Most of them (92.2%) are inhibitors, while only few drugs (3.2%) act as enzyme activators. The remaining compounds were indicated to act on enzymes as ‘agonist’, ‘antagonist’, ‘co-enzyme’, ‘co-factor’, ‘metabolite’, ‘modulator’ or ‘substrate’. The other compounds not targeting an

enzyme, but any other protein or molecule were as well categorized into different actions, such as agonists, antagonists, modulators and inhibitors, but as well into binders, enhancers, regulators, releasers or toxins, only to name a few.

### **Target classification according to the ChEMBL database**

In ChEMBL, drug targets can be accessed via different kinds of recorded assays. This means that depending on the assay, different forms of data are provided. Confirmatory assays typically involve potency measurements over a range of concentrations. The resulting ‘active concentration’ is specified as IC<sub>50</sub>, EC<sub>50</sub>, AC<sub>50</sub>, GI<sub>50</sub> or K<sub>i</sub> values. These concentrations indicate the amount of a drug that is needed for inhibiting a defined biological process by half or the amount where half of the maximum effect is observed, respectively. By querying those assays, the associated targets could be accessed and extracted.

ChEMBL categorizes all targets into 14 functional classes of proteins, such as enzymes, ion channels, surface or structural proteins for example, and has an additional group of unclassified proteins. In order to obtain an overview about the target distribution of our compound library, we first gathered all recorded targets of each CLOUPAC drug from the ChEMBL database. This was done via IC<sub>50</sub> (inhibitory concentration) activities, which represent a typical type of measurement when testing the performance of a drug in reducing a certain enzymatic activity. Additionally, we obtained a collection of potential targets via K<sub>i</sub> (inhibitory constant) activities, a similar type of potency measurement, providing information about the binding affinity of a drug.

Depending on a chosen threshold, the number of returned targets could be reduced or extended accordingly. For convenience, we decided to use the ChEMBL-established ‘pchembl’ value and allowed only targets to be collected, if the associated IC<sub>50</sub> or K<sub>i</sub> assay exhibited a ‘pchembl’ value greater or equal to 5. This value corresponds to all above-mentioned activity measurements on a negative logarithmic scale, meaning that a ‘pchembl’ value of 5 is equivalent to an IC<sub>50</sub> or K<sub>i</sub> measurement of 10<sup>-5</sup> M or 10 μM. In short, the higher the pchembl value, the lower the applied concentration. By pursuing this approach, we collected 588 unique targets. We observed that the majority of drugs has a small number of targets, ranging from zero to ten, while only few drugs target higher numbers of proteins (Figure 4).

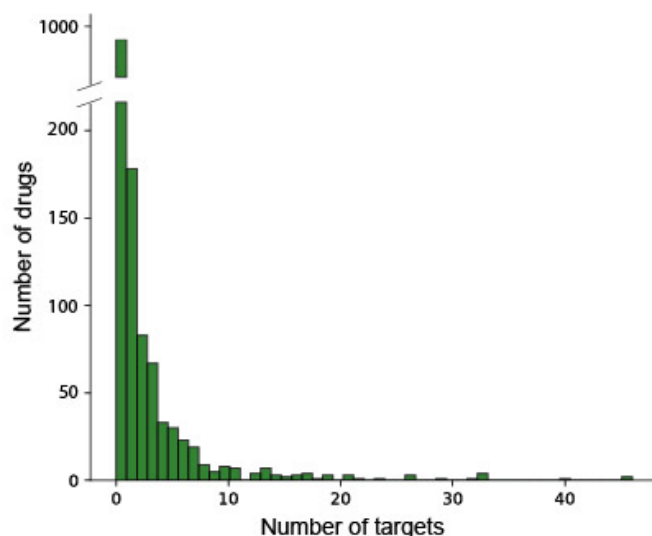


Figure 4: **Distribution of the number of targets per CLOUPAC drug identified in the ChEMBL database.** Drugs with high numbers of targets are rare, while for most drugs no or only very few targets were found.

We then categorized the collected targets into the different groups and counted the amount in each category. Subsequently, we compared these numbers to the different quantities of targets that are existing in the whole ChEMBL database. Table 13 (Supplementary material) shows the exact numbers per class of both, CLOUPAC and ChEMBL targets. If proteins appeared as target of different drugs, they were only included once.

As expected, we observed similar patterns in the abundances of target classifications when comparing CLOUPAC and overall ChEMBL targets (depicted in Figure 5). The numbers indicate that enzymes are by far the most frequent group of proteins that are targeted by drugs (60.85% of all classified proteins in ChEMBL), which is also reflected in our drug selection (58.01% of all classified proteins in CLOUPAC). Similarly, a relatively strong trend towards a high proportion of enzymes among the targets was observed by looking at the LOPAC<sup>®1280</sup>-provided data (31.9%), as described above. However, according to statistics of several publications, G-protein coupled receptors represent the most targeted protein class [38, 39, 40], which stands in contrast to our observations.

Equally, membrane receptors are in both cases the second most represented target class, directly followed by ion channels, which are still relatively well represented in both target collections. The number of transcription factors, epigenetic regula-

tors, transporters and other cytosolic proteins is in the lower range of abundances, while secreted, auxiliary transport, structural and other membrane proteins do hardly exist within the CLOUPAC library. Adhesion, surface antigens and other nuclear proteins do not occur at all in the CLOUPAC and are as well among the rarest targets in ChEMBL.

All in all, we observed resembling proportions of protein classes between the compared target collections. This lets us conclude that the CLOUPAC drug library is a suited and representative compilation of compounds targeting a variety of proteins with an emphasis on the most frequently occurring target class, namely enzymes.

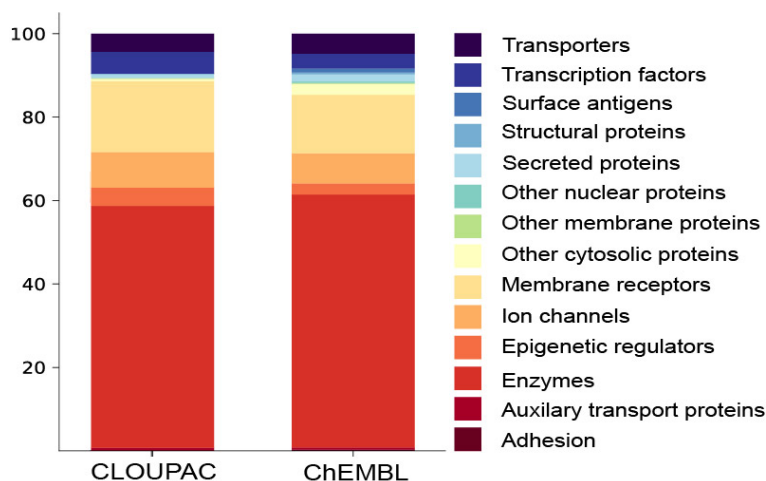


Figure 5: **Proportional comparison between targets of the CLOUPAC and the ChEMBL database.** Both drug collections show very similar proportions of protein classes: enzymes are by far the most represented group of targets, followed by membrane receptors and ion channels.

### The Human Protein Atlas (HPA) classification

Since every chemical database has its individual categorization of proteins, we considered it as useful to have a look into another database, namely the druggable proteome, which is part of the Swedish Human Protein Atlas (HPA) project, and to classify the targets of our drugs according to their classification system. In contrast to the ChEMBL database, the druggable proteome only includes 672 distinct proteins targeted by the current FDA-approved drugs and divides them only into six functional families: enzymes, transporters, voltage-gated ion channels, G-protein coupled receptors, nuclear receptors and CD (cluster of differentiation) markers.

In addition to the classification of targets into their respective functions, the HPA divides all recorded drug targets into six different locations within the cell, which mainly focus on whether the targeted protein is embedded in the membrane or not. According to this, the HPA categorization comprises ‘Integral membrane (IM)’, ‘Single pass-trans-membrane (SPTM)’, ‘IM/SPTM’, ‘Secreted’, ‘Membrane and secreted isoforms’ and ‘intracellular’ proteins. However, we questioned the usefulness of this classification, since it returns a rather superficial, little detailed and partly overlapping information about the actual location of targets. Nevertheless, we decided to include this information for an additional comparison between CLOUPAC targets and the druggable proteome in terms of the cellular occurrence of these proteins.

Separate lists for each target group were provided for downloading and gave some additional biological information as well as gene synonyms and an ensembl ID for each protein, i.e. a unique number that contains information about the species and object type of a specific gene or protein. It serves as a stable and universal code, which allows to compare objects between different data sources [41].

Because of the fact that only a small percentage of our drug library belongs to FDA approved drugs, the probability of having many matches between our targets and the druggable proteome targets was rather low. In the first place, we established a list of all 588 CLOUPAC targets that were translated to their assigned ensembl IDs. Subsequently, we were able to look for proteins from our target pool that were also present in the HPA target collection. Surprisingly, we found a match for 155 proteins. All of them were found to be assigned to a cellular location specified by the Swedish database. In contrast, 13 targets were not assigned to any functional group, whereas 39 other proteins were listed multiple times, indicating that they belong to more than one functional class.

Subsequently, we categorized all targets into the six different functional classes mentioned above. If a target had more than one function, we included all of them. The abundances of target functions within the CLOUPAC, as well as within the whole druggable proteome, are shown in Figure 6, the exact numbers are listed in Table 14 (Supplementary material). Exactly like in the LOPAC<sup>®1280</sup> and ChEMBL classifications mentioned above, enzymes represent the most prominent group in the CLOUPAC library, closely followed by G-protein coupled receptors. Since G-protein coupled receptors can be regarded as membrane receptors, the ranking of our HPA classification resembles again the pattern of the ChEMBL target classification, where membrane receptors are also the second largest group of proteins. Similarly, enzymes constitute the biggest functional group within the

druggable proteome, whereas the second most abundant proteins are in this case transporters, followed by G-protein coupled receptors.

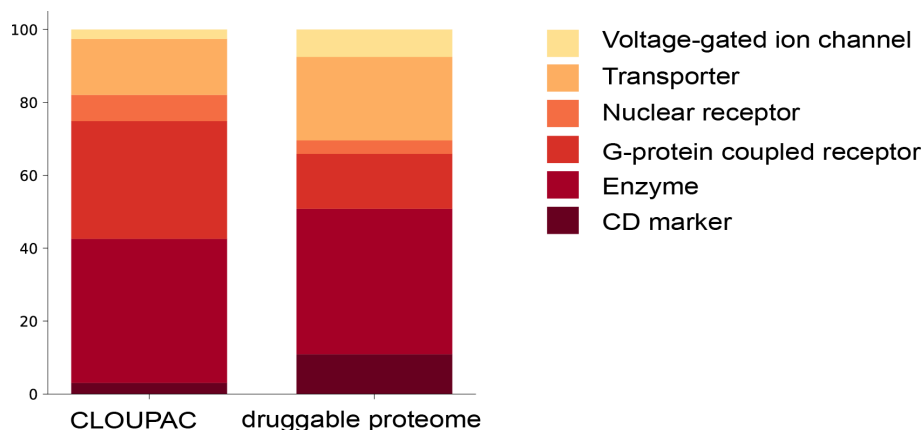


**Figure 6: Proportional comparison between CLOUPAC and HPA target functions.** In both libraries, enzymes are the most represented group of targets. While G-protein coupled receptors are present in a similar ratio as enzymes in the CLOUPAC library, transporters take the second place in the druggable proteome, whereas enzymes are rather weakly represented.

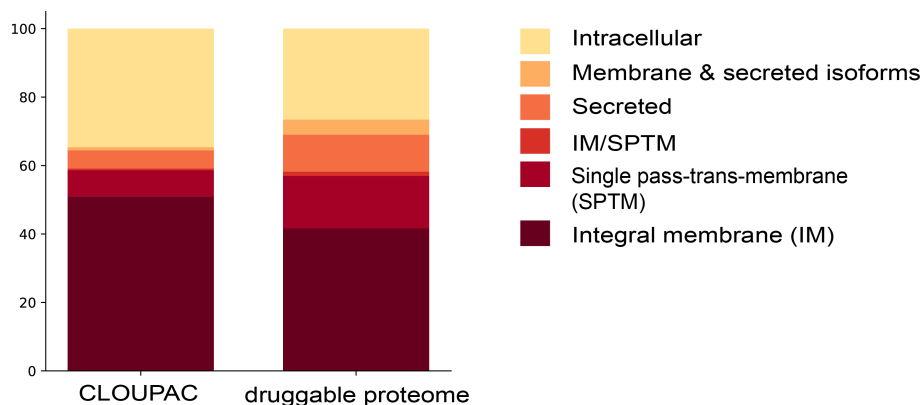


**Figure 7: Proportional comparison between CLOUPAC and HPA target locations.** In both libraries, most targets were assigned to the integral membrane, followed by intracellular proteins.

Figure 7 depicts the abundances of CLOUPAC target locations and of all annotated targets according to the HPA classification (exact numbers are listed in Table 15). Targets integrated into the membrane were the most abundant group within the CLOUPAC library, as well as within the druggable proteome. In second place, intracellular proteins were ranked among the most frequent proteins in both cases.

These results are in accordance with the ranking of the functional classes, where we found enzymes and G-protein coupled receptors to be the most abundant targets, since enzymes were in most cases classified as ‘intracellular’, whereas G-protein coupled receptors occur without exception in the membrane.

In contrast, the classes ‘Membrane & secreted isoforms’, ‘Secreted’, ‘IM/SPTM’ and ‘Single pass-trans-membrane’ are strongly underrepresented in the CLOUPAC. In the druggable proteome, ‘Secreted’ and ‘Single pass-trans-membrane’ proteins are present in larger ratios.

#### **2.3.4 Outcome**

The above-described characterization of the 1490 selected CLOUPAC drugs gave us not only insight into the general diversity and distribution of different properties that our drugs have, but simultaneously helped us to get an idea of how well described these compounds actually are or how much is known about them, respectively. Needless to say that the precisely selected FDA-approved CLOUD drugs, which already went through a selection procedure with strict criteria, are better annotated with detailed information than the LOPAC<sup>®1280</sup> library that contains many unknown or poorly annotated compounds. In fact, the vast majority of LOPAC<sup>®1280</sup> compounds does only exhibit a minimum amount of information in the ChEMBL database. Still, those compounds can potentially alter the cellular phenotype and reveal currently unknown effects on the cytoskeleton, which would make them interesting for further applications or combinatorial screens for instance and thus represent the explorative part of the CLOUPAC. After this characterization, our CLOUPAC drugs were ready for performing an automated high-content imaging screen including different concentrations.

### **3 High-content high-throughput screen and image analysis**

#### **3.1 Imaging screen of 1490 compounds**

The high-throughput imaging screen was conducted on the adherent cell line MCF10A, which is derived from a human mammary gland. Cells were treated with 1490 distinct compounds in four different concentrations (0.02, 0.2, 2.0 and 20.0  $\mu$ M) and incubated for 72 hours. Every sample was present in duplicates so that every drug was represented in 8 wells. Subsequently, nuclei, cytoskeleton and mitochondria of the cells were stained with three different fluorescent dyes and were accordingly imaged in three different channels. After image acquisition,

the raw image data was analyzed in the image analysis software ‘Cellprofiler 3.0’. A detailed description of the high-throughput high-content imaging screen as well as of the CellProfiler pipeline can be found in the ‘Material & Methods’ section.

For demonstration, Figure 8 depicts one field of view imaged in three different channels: nuclei in blue, the  $\beta$ -tubulin cytoskeleton in green and mitochondria in red, as well as the merged image of all three channels.

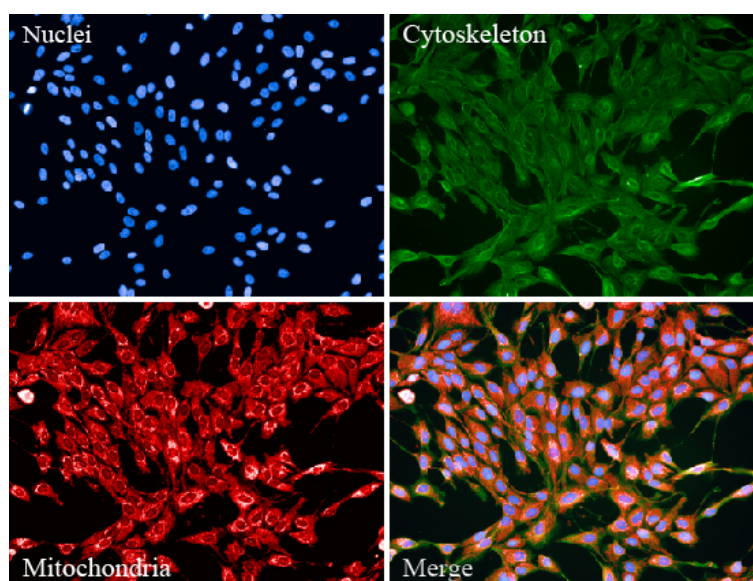


Figure 8: **Immunofluorescent staining of MCF10A cells.** Images were acquired by an automated microscope in three different channels, visualizing the stained nuclei in blue, the  $\beta$ -tubulin cytoskeleton in green and mitochondria in red. The images were additionally overlaid for depicting whole cells.

## 3.2 Image analysis, quality control and normalization

### Image analysis using CellProfiler 3.0

The large amount of raw data that has been generated after the screen served as input material for the computational processing using the open-source imaging analysis software ‘CellProfiler 3.0’, which was established by the group of Anne Carpenter from the Broad Institute. The program was specifically designed for identifying phenotypes automatically from a large number of images and measuring many different kinds of properties of the respective objects of interest, involving various shape parameters, intensities or position coordinates, for example. Additionally, it provides the opportunity to specifically adjust the analysis to the objects to be measured, which refers to cells, nuclei and mitochondria in our case,

by determining the order of processing steps.

### **Building a CellProfiler pipeline**

The analysis is based on a selection of modules, such as an illumination correction, quality assessments and ultimately various kinds of measurements that will be performed on the objects. Thereby, the basic approach is a preceding identification of primary objects (nuclei), which are later used as a reference for secondary object recognition (cells). Reason for that is the fact that nuclei are in general more uniform, well separated from each other and easily identifiable from the background in contrast to stained cell bodies. Secondary objects can consequently be recognized more easily by the software.

The modules can be freely selected to a certain degree and arranged in a variable order with the aim of establishing a pipeline for an optimal automated object identification, which can continuously be tested and optimized by the usage of a small set of images. Finally, before running the pipeline on a large scale, the composition of the desired measurements to be performed has to be defined, which involved predominantly morphology-associated statistics in case of our analysis. The exact design of our ‘CellProfiler’ pipeline is provided in the ‘Materials & Methods’ section 8.2.

### **Quality measurements and ‘features’ are returned as output data after CellProfiler analysis**

After the analysis, numerous data files were returned by CellProfiler as output, comprising all measurement results per image or per object, which are referred to as “features”. A feature corresponds to a single type of measurement that was performed on a cell or nucleus, resulting in a single value per object. For each feature, an output file was created, which facilitated the statistical evaluation and comparison of features.

Having established and tested the full CellProfiler pipeline on a small set of images, we started the analysis for the entire raw data. Various data tables were created after the pipeline was passed through. As a result, 51 quality measurements were returned, which were used for distinguishing useable images from those that exhibit illumination artifacts (see ‘Materials & Methods’ 8.3) and could thus not be used. Subsequently, the remaining values had to undergo two necessary normalization steps for removing plate effects on the one hand and for bringing them into a normalized range of numbers between zero and one on the other hand (see ‘Materials & Methods’ 8.4).

Moreover, 638 distinct features were returned per object and per well. These

features constitute the most valuable aspect of our image analysis. As not all of them were suitable for our project or did not contain usable information, we reduced them strongly to only retain those measurements that met all of the criteria that we defined. These include small variability within DMSO-treated samples, strong morphological effects, well correlating replicates and a low redundancy.

## **4 Acquiring an optimized morphology-associated feature set for evaluating drug effects**

In sum, there were 638 features returned by CellProfiler on both, a single cell and a per-well level. Since we were not interested in a comparative single cell analysis, but rather in the overall morphological effect that one drug has on the cells, we exclusively worked with the averaged data across the whole cell population per well.

The extracted features included a variety of cell shape measurements, as well as texture and intensity statistics for all three channels. In our approach, we considered the set of all features as a vector in a multidimensional space, representing not only a specific direction of drug effect, but also its magnitude. By summing up all feature values per compound, we created unique drug vectors, which we used for evaluating the individual impact that a drug has on the cells.

We used this vector-based approach also for reducing the number of features, since not all of them contained useful information. Our aim was to establish a condensed set of informative, reliable and non-redundant measurements. The following steps describe the criteria that we defined to obtain this subset. Features that did not match the requirements were considered as ineligible and were removed.

### **4.1 Minimizing inter (I) and intra plate (II) effects**

One important criterion that a feature had to fulfil was a certain consistency within the negative controls of DMSO treated samples. Hence, the fluctuation within the DMSO controls of the same plate, as well as between all plates was determined for each feature separately via the coefficient of variation (CV). In other words, standard deviation ( $\sigma$ ) and mean ( $\mu$ ) were extracted from the DMSO values per feature. This was either done per plate (intra plate effects) or for all plates of the same batch (inter plate effects). The ratio of both values ( $c_v = \frac{\sigma}{\mu}$ ) gives information about the dispersion of the distribution.

Small coefficients would indicate a small extent of variability, while in contrast, high coefficients imply a stronger variation of values. However, a feature exhibiting high intra or inter plate fluctuations is not suitable as a negative control, since a statistical comparison with drug treated samples would in such cases not be meaningful, nor informative. We considered 0.2 as a proper threshold for both, intra and inter plate effects that should not be exceeded.

## 4.2 Calculating the effect size of features (III)

As a next filtering step, we selected only those features, which showed a certain extent of effect between drug-treated and untreated cells, meaning that we could potentially gain valuable information from this measurement. For quantifying the magnitude of an effect, the values of all drug-treated samples as well as of all negative controls were collected separately per feature. The Cohen's D effect size (equation 1) was then determined by calculating the mean difference between both distributions. Resulting values of 0.2 or smaller indicate in general a small effect, whereas values above 0.8 are regarded as strong effects [42]. In order to select features that definitely represent some effect, we defined the criterion for the effect size to be greater than three.

$$d = \frac{\mu_1 - \mu_2}{s} \quad (1)$$

where:

$\mu_1$  = mean of drug-treated samples

$\mu_2$  = mean of negative controls (DMSO)

$s$  = pooled standard deviation for two independent samples [43]

## 4.3 Filtering for robust replicates (IV)

A feature can be regarded as suitable and reliable, if two replicates result generally in similar values. An informative statistical approach to validate this is the Pearson correlation coefficient, which measures the linear correlation between both replicates and returns values close to 1 in case of a strong positive correlation and values close to -1 in case of a strong negative correlation, while a coefficient around 0 indicates no linear relationship.

The Pearson correlation over all replicates was calculated for each feature individually. Thereby, we observed that there was a relatively weak overall correlation

with an average coefficient of 0.28 for batch 1 and 0.31 for batch 2. As a consequence, in order to avoid rejecting too many features, we chose a reasonably low threshold of 0.2 for the Pearson correlation, which had to be exceeded.

Figure 9 depicts four examples for the distribution of replicate values. An overall Pearson correlation higher than our selected threshold of 0.2 can be observed particularly in features relating to shape measurements of the cells, such as the area or the minimum Feret diameter, which indicates a good reproducibility of replicates. In contrast, granularity features for instance, which measure the texture of an image, score in some cases badly. However, this does not necessarily imply that the replicates differ strongly with regards to the cell shape, but can be due to varying cell distributions or densities. Poor correlations were as well observed for features counting the number of objects, such as cells, nuclei, mitochondria or children objects, which are touching or located within parent objects.

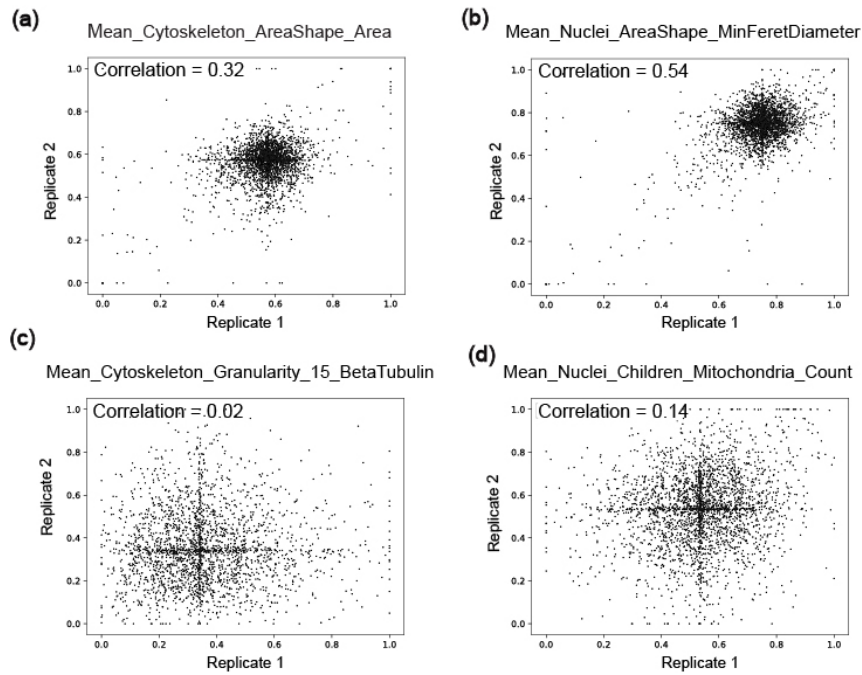


Figure 9: **Visualization of replicate correlations for four different features.** (a) Comparing all measurements of the feature “Mean Cytoskeleton AreaShape Area” resulted in a Pearson correlation of 0.32 and thus exceeds the threshold (which is 0.2). (b) “Mean Nuclei AreaShape MinFeretDiameter” shows an even higher correlation of 0.54, while the features (c) ‘Mean Cytoskeleton Granularity 15 BetaTubulin’ and (d) ‘Correlation Mean Nuclei Children Mitochondria Count’ show an extremely weak Pearson correlation and were excluded from the further analysis.

#### 4.4 Removal of redundant features by creating a correlation network (V)

At this point, useful features could already be selected via the application of four different kinds of filtering criteria that were mainly based on statistical validations of their reliability and their informative content. Nevertheless, many of the remaining features exhibited a certain degree of redundancy with others. In those cases, we considered it as sufficient to pick only one feature from a pool of very similar measurements. In order to define which features can be regarded as redundant, a correlation network was constructed by using again the Pearson correlation. For that, all DMSO values were collected per feature for a subsequent pairwise correlation. If this correlation was higher than 0.6, an edge was drawn between these two features. In other words, the network consists of features as nodes and they were connected, if they were similar or redundant, respectively.

In a following step, the network was fragmented, meaning that nodes (features) were stepwise removed from the network until there were only nodes left that were not connected to any other. All single nodes were kept. If only two nodes were connected to each other, the one having a greater effect size was selected. In this way, the remaining pool of features could be reduced to a non-redundant set of measurements.

Table 3: **Summary of statistical approaches and applied thresholds.** The different filtering methods were applied with the aim of reducing the original amount of features to a meaningful set of reliable, informative and non-redundant measurements.

| Filtering method                          | Applied threshold |
|-------------------------------------------|-------------------|
| I. Intra plate effects                    | $< 0.2$           |
| II. Inter plate effects                   | $< 0.2$           |
| III. Effect size                          | $> 3.0$           |
| IV. Pearson correlation of replicates     | $> 0.2$           |
| V. Pairwise Pearson correlation (network) | $< 0.6$           |

Table 3 lists the above-described filtering steps and the respective applied thresholds. Still, we were not convinced, whether all established criteria served the purpose of retaining the best suited set of features. The following section explains in detail how we proceeded to find the optimal combination of filters to be applied.

## 4.5 Finding the best set of filters by comparing cosine similarities of drug vector pairs

After we defined several criteria for minimizing the original amount of 638 features to an appropriate subset, an empirical validation whether these methods are actually reasonable was still missing. In order to not only test, which of these steps were useful, but as well in which constellation, we applied different combinations of the described filtering methods. As a result, different amounts and compositions of features were left, depending on the combination of applied filters, which were, according to the Roman numbers of Table 3, the following: (A) all filters, (B) no filters, (C) IV, (D) I-III, V, (E) I-IV, (F) IV, V, (G) III, V, (H) V and (I) I, II, V.

Subsequently, we summarized all drug vectors, consisting of the respective feature subsets, into three different groups: ‘Replicates’, ‘Same Drug’ and ‘All Drugs’. Each group comprised the entirety of cosine similarities between the respective drug vectors, whereas a high average similarity is expected within the group of replicates. Contrary, ‘All drugs’ were expected to show a wide dispersion. For the detailed approach see ‘Materials & Methods’ 8.5.

The dispersion of cosine similarities of the three groups was visualized by box-plots. Figure 10 shows the results after using either all of the filters, only certain subsets of filters or none at all. A significant difference between the distinct subgroups is clearly observable in all cases.

As expected, filtering features according to the correlation of replicates (Filtering method V) is useful to increase the overall cosine similarity within the group of replicates. In addition, it is shown that cosine similarities are widely scattered within the group of all drugs, meaning that there are various kinds of effects evoked by the compounds, while this wide range was strongly reduced by comparing only replicates. This observation indicates that there are a lot of specific perturbations occurring within the treated cells.

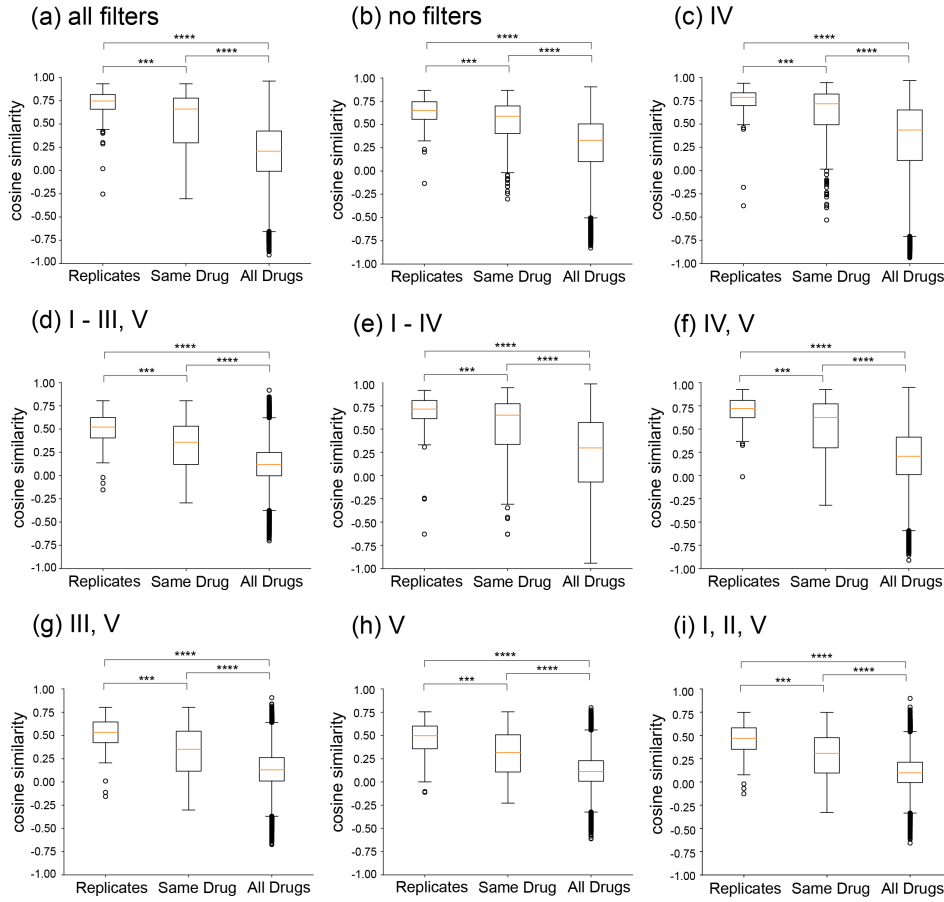


**Figure 10: Boxplots showing that different combinations of applied filtering methods result in all cases in highly significant differences of cosine similarities between the individual drug subgroups.** Roman numerals indicate, which filtering methods were applied: **(a)** all filters **(b)** no filters **(c)** Pearson correlation of replicates **(d)** Intra plate effects, Inter plate effects, Effect size **(e)** Intra plate effects, Inter plate effects, Effect size, Pearson correlation of replicates **(f)** Pearson Correlation of replicates, Pairwise Pearson correlation (network) **(g)** Effect size, Pairwise Pearson correlation (network) **(h)** Pairwise Pearson correlation (network) **(i)** Intra plate effects, Inter plate effects, Pairwise Pearson correlation (network). (Only results of batch 1 are shown.)

For determining, which feature collection to use, we decided to judge by the maximum separation between the group ‘Replicates’ and the group ‘All drugs’, since this difference represents the discriminability between specific and unspecific effects. This was done by calculating the p-values between these two distributions by using the Mann-Whitney U test. This was done for both batches separately. Table 4 lists the three filtering subsets, where we observed the highest p-values in each batch.

Table 4: **List of the three filtering subsets, where the highest significant difference of cosine similarities was observed between the drug subgroups ‘Replicates’ and ‘All Drugs’.** The ranking of filter subsets slightly differs between batch 1 and 2. However, in both cases, ‘all filters’ resulted in one of the most significant approaches.

| Batch | Subset of filters | p-value between distributions |
|-------|-------------------|-------------------------------|
| 1     | (F) IV + V        | $2.78 * 10^{-99}$             |
|       | (A) all filters   | $1.24 * 10^{-97}$             |
|       | (G) III + V       | $1.93 * 10^{-80}$             |
| 2     | (A) all filters   | $1.42 * 10^{-89}$             |
|       | (F) IV + V        | $3.17 * 10^{-85}$             |
|       | (E) I-IV          | $3.36 * 10^{-79}$             |

The results revealed that the application of all filters indeed resulted in the highest significance after averaging both batches, even though there was not a great difference to the p-values of the other filtering subsets.

## 4.6 Outcome

Based on the above analysis, we decided to reduce the pool of features by applying all described filtering steps that are listed in table 3. This resulted in a final set of 66 features, which met all the required criteria: low intra and inter plate effects, high effect sizes and a strong overall correlation of replicates. In addition, there is only a low degree of redundancy present among the remaining features. A list of the remaining 66 features is provided in Supplementary material (Table 16 and 17).

For the following approach, a 66-dimensional feature vector was created for each drug that allowed us to establish a condensed library of compounds exhibiting crucial properties, such as a good reproducibility and a strong morphological perturbation of the cell, but at the same time a rather low cytotoxicity. The following section describes the stepwise approach of identifying and collecting the most promising perturbors.

## 5 Selection of morphologically active compounds: the creation of the CLoMP

Since our main goal of the CLoMP was to design a library of specific morphological perturbors, one particularly important step was to find a way to validate the strength of the morphological effect of a drug, which is ought to be preferably high. Equally important was a certain reproducibility or correlation among both replicates as well as finding the appropriate concentration of the drug, which is inducing a morphological change without killing the cell.

The previous creation of feature vectors for each drug allowed us to filter for these criteria. According to that, we could select compounds by strong effects, which are equivalent to high vector lengths. Equally, we could easily identify the similarity of replicates by calculating the cosine of the angle between the two vectors, related to the approach described in section 5.3. The following section is now dealing with the process of filtering the 1490 chemical compounds. At the beginning of the procedure, we merged the feature vectors of both batches into the same file and, unless otherwise stated, processed them together.

### 5.1 Reproducibility as a major criterion for our library

If cells are treated with a drug for several times at the same and additionally at different concentrations, the morphological response of the cell is in general expected to be the same in each case, albeit to different extents. Based on that, we decided to reject those drugs from our collection, which did not show a reasonable consistency over both replicates. In order to do so, we defined as a criterion for a drug to be accepted, a cosine distance of both replicate vectors smaller than 0.4 in at least one of the four concentrations. The cosine distance is defined as  $1 - \text{cosine similarity}$ , implying a possible range of values between 0 (equal vectors) and 2 (opposed vectors), meaning that using a cosine similarity of 0.6 as a minimum threshold is effectively the same.

In the course of this, we deliberately decided on using the cosine distance instead of the Euclidean distance, which is in simple terms the measure of the straight line between two points, which we do not regard as sufficiently reliable. For example, in the case of a drug that evokes hardly any morphological effect and will thus result in two short and DMSO-resemblant feature vectors, the Euclidean distance would be extremely small. This would indicate strongly correlating replicates, even if both vectors are pointing into opposite directions. In contrast, two vectors that have an extremely similar direction, but are of different lengths, would result

in a high Euclidean distance and hence be rejected, even if the morphological outcome is almost the same. The cosine distance is thus the more reasonable option for deciding on suitable drugs, since it takes only the angle between both vectors, i.e., the similarity of their directions, into account.

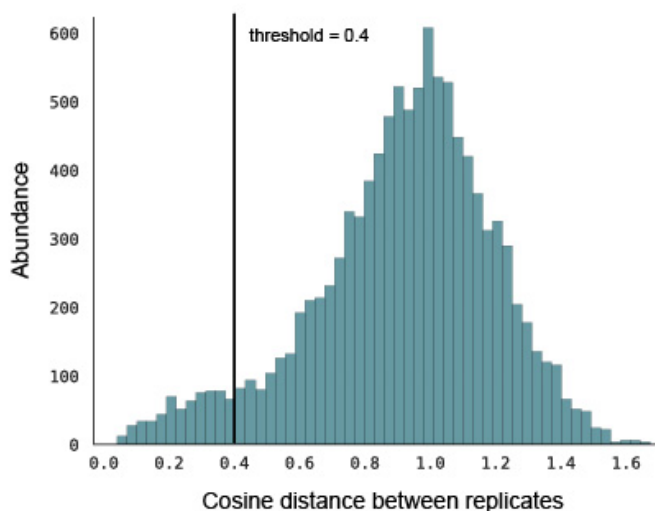


Figure 11: **Distribution of cosine distances between all pairs of drug replicates.** The smaller the angle between both replicate vectors, the more consistent is the effect of the drug. Vectors of drug replicates need to exhibit cosine distances smaller or equal to 0.4 in order to qualify for the CLoMP.

Figure 11 depicts the distribution of cosine distances between all pairs of replicates. The values are generally normally distributed with a peak around 1 (corresponding to an angle between the vectors of  $90^\circ$ ), likely reflecting stochastical fluctuations. However, the distribution shows a significant left tail of exceedingly similar replicates. This indicates a collection of drugs having a specific and reproducible effect on the cell morphology. The visual inspection also serves to explain the choice of the threshold set at 0.4, which leads to the retention of only those replicates causing a distinct and directed phenotype.

The application of this first approach in the drug searching process resulted in a strong decrease of the original number of drugs, namely 274 distinct compounds (or 626, if different concentrations were counted as different perturbations). This represents 18.4% of the original quantity of drugs.

## 5.2 Selection of strong morphological perturburbers

In the following, we proceeded with the previously selected drugs and further reduced them by searching for the most effective morphological perturburbers. This was necessary, since the impact of the drug on the cellular morphology is a crucial and probably even the most essential criterion for a drug to be part of the final library. Since drugs were represented as vectors in a multidimensional space, defined by 66 different feature values, the actual length of a vector could easily be calculated to serve as an indicator for the strength of an effect.

Thereby, the entireties of DMSO and drug vectors were collected separately, since based on the dispersion of the DMSO vector lengths, it was possible to make an estimate about what is noise and what can be considered as actual drug-induced change. In order to set an appropriate threshold for demarcating strong effects from noise, the 98th percentile of all DMSO values was calculated, which resulted in a minimal vector length of 1.23, and used as a cut-off, meaning that the highest 2% of DMSO values were regarded as outliers. All drug vectors exceeding the cut-off were considered as sufficiently strong.

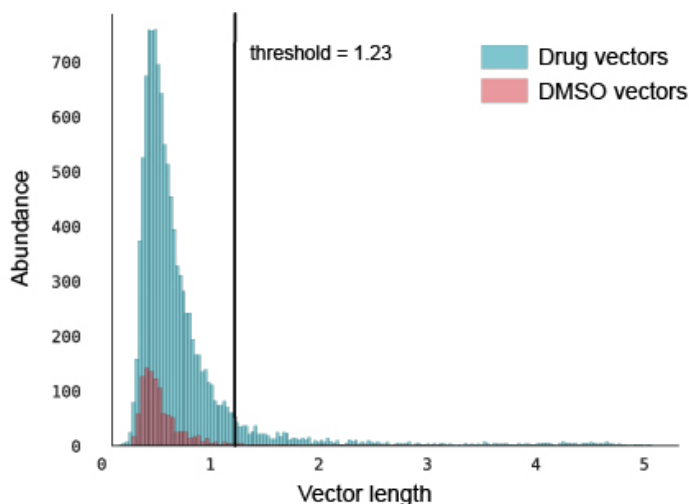


Figure 12: **Distribution of DMSO and drug vector lengths.** The length of the drug vector can be equated with the strength of the drug effect. For determining a suitable threshold for a minimal vector length, we used the 98th percentile of DMSO values. Hence, drugs are considered as sufficiently effective, if they exhibit a vector length of at least 1.23.

In this manner, the previously diminished pool of drugs was again reduced. Figure

12 visualizes both, the general distribution of all drug vector lengths, as well as the distribution of DMSO vector lengths, which has been used to find a suitable threshold value. This approach led to a remaining amount of 205 unique drugs, (or 226, if different concentrations of the same drug were taken into account). These are still 13.76% of the initial number of drugs.

### 5.3 The relationship between the correlation of replicates and the vector length

As already elucidated, weak drug effects result in short and unspecific feature vectors, whereas a strong effect leads to a long and specific feature vector. Consequently, we assumed that a highly effective compound would usually generate two specific replicate vectors of the same kind. Short vectors are in contrast generally associated with a high extent of noise. Hence, it can be concluded that they show only a low accordance of duplicates.

In order to illustrate the relationship between the strength of the effect and the replicate correlation, the average vector length of two replicates was calculated and plotted against the cosine similarity between those vectors. Figure 13 depicts that exceedingly long vectors in fact comprise a certain tendency of remarkably well correlating replicates, meaning that strong effects seem to correlate with similar effects. In comparison, short vectors show a broad variety of cosine similarities, indicating unspecific effects and noise.

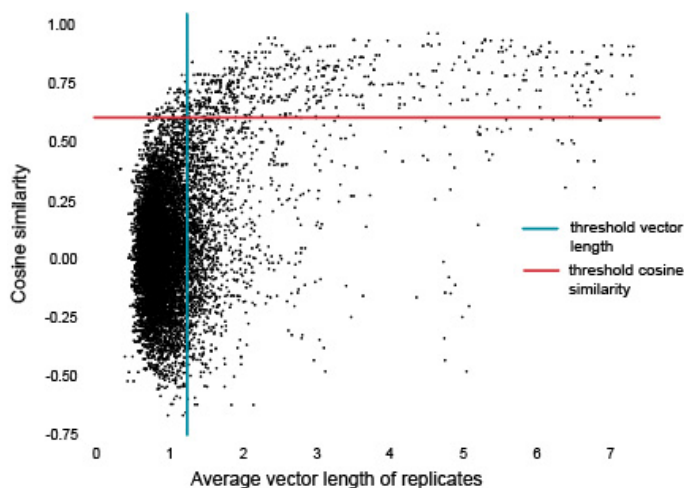


Figure 13: **Relationship between the replicate correlation and the strength of the drug effect.** Drugs exhibiting a high average vector length generally show a better correlation than drugs with short vectors.

The applied thresholds (as described in sections 5.1 and 5.2), shown as horizontal and vertical lines, make it possible to collect only those vector pairs representing a consistent and strong phenotype at the same time.

## **5.4 Removal of highly cytotoxic drugs**

As explained above, various classes of drugs are existing, which are highly diverse in their chemical structure and are involved in numerous different biological processes. As a consequence, chemical compounds differ strongly in their activity and thus also in their toxic impact on cells. Therefore, it was important to exclude those drugs from our library, which are killing too many cells, since this would lead to wrong statistical assumptions. Also, if a compound is too harmful, morphological changes due to cell damage or apoptosis cannot be excluded and would potentially result in misleading interpretations.

In order to prevent this kind of biases, we decided to implement another criterion that a drug needs to meet for being qualified as a suitable perturber, which is a minimal number of cells being still present after the drug treatment. In general, the cell count largely depends on the seeding density at the very beginning of the screen. However, as described above, the whole screen was divided into two batches that were handled separately with a few days in between, leading to a clearly observable difference in the cell count distribution between both batches. For this reason, the threshold was determined individually for each batch, which was done by collecting all negative control (DMSO) cell counts per well and subtracting two standard deviations from the mean.

The resulting number represents the minimal requirement of cells surviving the drug treatment. In case of batch 1, these were 374 cells per well, while in batch 2 it was a lower number of only 192 cells. Figures 14 and 15 illustrate the distribution of cell counts per well for both batches individually, depicting the span of DMSO values compared to drug values and additionally showing the applied thresholds for the minimal number of remaining cells.

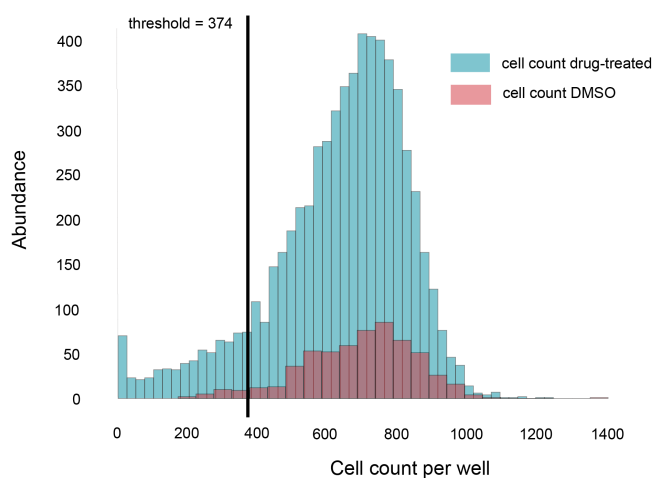


**Figure 14: Cell count distributions of drug-treated and negative controls of batch 1.** The threshold was calculated by subtracting two standard deviations from the DMSO cell count mean. Only drugs leaving more than 374 cells per well after the treatment were included in the CLoMP.

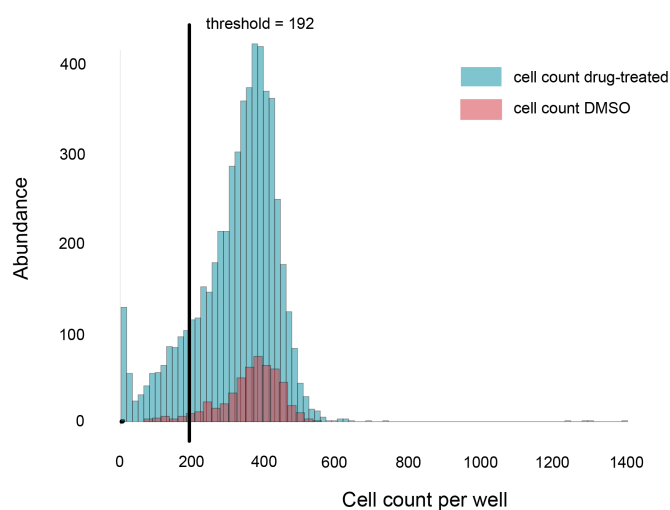


**Figure 15: Cell count distributions of drug-treated and negative controls of batch 2.** The threshold was calculated by subtracting two standard deviations from the DMSO cell count mean. Only drugs leaving more than 192 cells per well after the treatment were included in the CLoMP.

After filtering the remaining drugs according to this criterion, the resultant compound lists of both batches were merged again to one list for the subsequent steps.

At this point, we counted 138 drugs reaching the minimal amount of cells, including some that were present in different concentrations and 90 drugs, if each compound was only counted once. This corresponds to 6% of the original amount of drugs. Still, we did not have a collection of unique drugs indicating the optimal concentration to use, which represents the last approach in filtering our compilation of morphological perturbers. Thereby, we included an additional criterion, which allowed only drugs to be accepted, if they exhibited at least one replicate for each concentration, meaning that all four concentrations are represented and were not rejected due to bad image quality. This step led to the removal of only two drugs, whereby 88 unique drugs (5.9 % of the CLOUPAC) still remained, or 135 respectively, if different concentrations were included. This selection of drugs represents the final library ‘CLOMP’.

The last task in the procedure of establishing a list of suitable morphological perturbers was the selection of the appropriate concentration for each compound, since within the 135 remaining drugs, several of those appeared in more than one concentration. Therefore, we simply decided to choose the concentration, at which we observed the highest drug vector length and thus directly correlates with the strongest morphological effect that a drug has on the cell. By this, we generated a final list of 88 compounds in their optimal concentrations to use for potential subsequent screens. A complete list of all CLOMP drugs and their respective concentrations can be found in Supplementary Table 22.

## 5.5 Outcome

With 88 compounds, the CLOMP represents a relatively small fraction of the originally compiled CLOUPAC library (5.9%). However, we can be confident that the remaining compounds are highly suitable for morphology-based studies, as they passed a variety of strict thresholds. In this context, the application of representative feature vectors proved to be an efficient method for evaluating correlating effects between drug replicates, but also for estimating the strength of a morphological effect evoked by a drug. Both criteria were crucial for an appropriate drug filtering process. We additionally excluded drugs that apparently have too toxic effects on the cells by referring to the remaining numbers of cells after the drug treatment. In this way, we achieved to establish the CLOMP, consisting of 88 compounds with three major characteristics: having highly reproducible replicates, a strong morphological effect on cells and exhibiting low cytotoxicity.

A manual examination of images from cells treated with CLOMP drugs confirmed that the remaining compounds are capable of inducing distinct morphological changes in MCF10A cells. Figure 16 shows some examples of observed phe-

notypes after drug treatment, including not only changes in cellular size, but as well an increased formation of filopodia or different types of cell clustering. This association of morphologies to specific drugs allows investigating fundamental biological challenges, such as understanding drug-drug interactions, for example. More specifically, when a new phenotype emerges from a combined application of two drugs, which individually did not induce changes in cell morphology, it would imply an interaction between the drugs, which in turn induces the activation of structure-related pathways. Similarly, two compounds that previously led to a characteristic morphology can potentially reveal particular effects that one drug has on the other, when administered in combination. In this case, deviations from the expected phenotype could uncover a potential synergistic or antagonistic interference of both compounds. In the same way, genetical perturbations can be investigated and combined with chemical perturbations. In this regard, the CLoMP provides the basis for numerous follow-up studies based on any perturbations that manifest themselves in the cellular morphology.

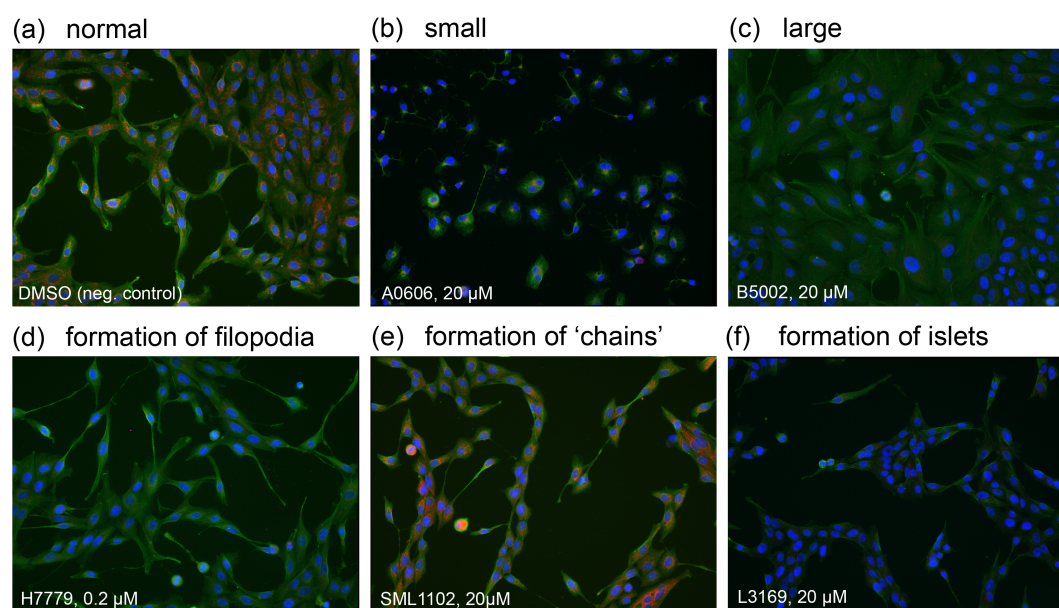


**Figure 16: Examples of drug-induced phenotypes observed in MCF10A cells.** (a) Untreated control. (b) Treatment with Aurothioglucose (A0606) results in cells with a very small and round phenotype. (c) The treatment with 5-Bromo-2'-deoxyuridine (B5002) induces a large phenotype and a high density of cells. (d) Retinoic acid p-hydroxyanilide (H7779) treatment appears to lead to an increased formation of filopodia. (e) Upon treatment with Darifenacin hydrobromide (SML1102), we observed chain-like structures of cells, while in case of (f) LP 12 hydrochloride hydrate (L3169) cells tend to be smaller and form isolated groups.

## 6 The characterization of the CLoMP

Now that our 1490 CLOUPAC drugs were screened, imaged and evaluated with regards to their morphological effect on cells by numerous aspects, they could be downsized to a small collection of 88 non-redundant morphological perturbors that showed to have a significant impact on the cellular phenotype. Referring to the CeMM in-house library ‘CLOUD’, we named the new drug collection ‘CLoMP’, standing for ‘CeMM Library of Morphological Perturbors’. The CLoMP consists of 88 drugs, whereas 69 drugs originate from the LOPAC<sup>®1280</sup> library, which are 4.63% of the original Sigma-Aldrich library, while 19 drugs were part of the CLOUD, accounting for 9,05%.

In order to provide us with insights into the properties of our newly compiled ‘CLoMP’, we again characterized the remaining compounds in the same way as we did with the CLOUPAC before performing the screen. Taken together, 72 of 88 drugs possessed ChEMBL IDs, meaning that 16 compounds remained uncharacterized. From the 19 CLOUD drugs, there is one drug without a ChEMBL identifier, see Table 5 for an accurate overview.

Table 5: **Summary of CLoMP drugs.** The majority of CLoMP compounds originates from the LOPAC<sup>®1280</sup> library. In addition, the greatest part of the CLoMP is characterized, meaning that for most of the drugs, a ChEMBL ID was found.

| Source                 | characterized | uncharacterized | Total |
|------------------------|---------------|-----------------|-------|
| LOPAC <sup>®1280</sup> | 54            | 15              | 69    |
| CLOUD                  | 18            | 1               | 19    |
| Total                  | 72            | 16              | 88    |

The following section elaborates on the main characteristics of our compounds, comprising some general data, the ATC classification and target classification of our compounds. In addition, we looked at the morphological behaviour of the cells over the range of the four applied concentrations.

### 6.1 Collecting basic data about the ”CLoMP”

In order to gain a first impression of the basic properties of the CLoMP, we again started with taking a look into the approval state of the final 88 perturbors. Thereby, the proportion of approved drugs is particularly of interest since it allows us to make statements about whether our drugs, which underwent strict filtering

processes, contain primarily well-known and extensively studied compounds or if we encounter rather unknown compounds.

Most of our 88 compounds were found to possess information about their approval state, whereas there was no data available about the approval state of 19 drugs (Table 18). Among them were the 16 above-mentioned drugs that did not have a ChEMBL ID at all. Apart from that, the overall results correspond to the proportions of the initial CLOUPAC and are visualized in Figure 17. There are 33 FDA-approved drugs, which account for more than a third of the CLoMP. The second most abundant approval phase of the CLoMP drugs is phase 0, the lowest state that a drug can have in the conduct of clinical trials. Hardly any compounds were found for approval phases between both extrema, meaning that the classified compounds could be divided into two separate clusters: well-known and hardly known drugs.

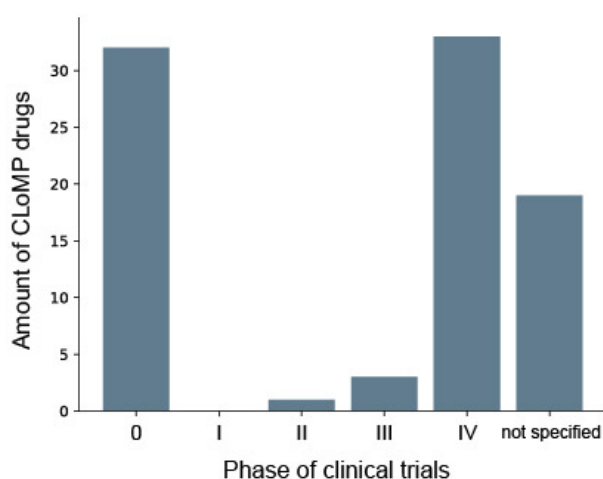


Figure 17: **Distribution of CLoMP drugs into phases of clinical research.** Similar to the CLOUPAC, phases 0 and 4 are the most represented ones, even though the number of phase 4 drugs have proportionally increased in the CLoMP.

Further information was gathered about the CLoMP drugs by counting the number of prodrugs and looking for targets, mechanisms of action and associated diseases. The numbers of drugs falling into the respective category or having these kinds of data available are summarized in Table 6.

Within the whole CLoMP, there was only one drug found to act as a pro-drug, while 24 were declared to be non-pro-drugs. The remaining drugs did not contain

this information on ChEMBL. Targets were found for 35 drugs, which is more than a third of the library and resembles the ratio of the CLOUPAC. Mechanisms of action are known for 17 compounds, which constitutes a smaller proportion than in the CLOUPAC. Furthermore, there were 27 drugs with indications for medical use, which is again a similar percentage as in the CLOUPAC.

Table 6: **Different kinds of basic information about the CLoMP drugs according to the ChEMBL database.** As for the CLOUPAC, we collected available data about the CLoMP compounds pertaining to prodrugs, targets, MOAs or diseases.

| Entry in ChEMBL | #CLOUPAC drugs having this data available |
|-----------------|-------------------------------------------|
| Prodrugs        | 1                                         |
| Non-Prodrugs    | 24                                        |
| Not classified  | 63                                        |
| Targets         | 35                                        |
| MOA             | 17                                        |
| Disease         | 27                                        |

All in all, we can conclude from this primary characterization that the CLoMP is essentially a representative selection of the initially compiled and screened CLOUPAC.

## 6.2 Classification of CLoMP drugs according to the anatomical therapeutic chemical (ATC) code

The classification of drugs into their respective target location is an especially interesting aspect, since this part of the characterization gives an indication of how divers our final library of morphological perturbors is, at least for the well-known compounds. Therefore, we made again use of the above-described ATC system for categorizing our 88 drugs into different anatomical or pharmacological groups. Taken together, the ChEMBL database provided entries containing this information for 21 drugs, whereas two of them possess two different first-level classifications. The absolute frequencies are shown in table 19 (Supplementary material). Figure 18 illustrates additionally the relative abundances of ATC classes as a pie chart.

In contrast to the CLOUPAC, which covers the full spectrum of ATC groups, there are only 8 out of 14 ATC classes represented in the CLoMP. Drugs targeting the nervous system are still the most abundant ones within the CLoMP, followed by antineoplastic and immunomodulating agents (L). While the previously strongly represented cardio vascular system (C) drugs declined strongly, antiparasitic products, insecticides and repellents (P) take over a higher proportion in the CLoMP than in the CLOUPAC.

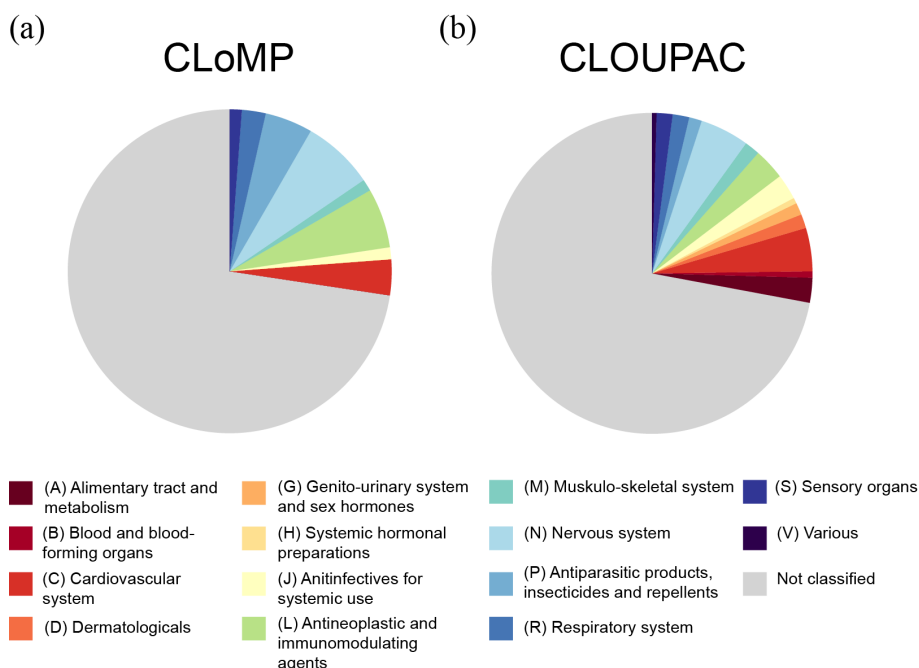


Figure 18: **Comparison of the overall ATC classification of CLOUD, LOPAC<sup>®1280</sup> and CLOUPAC drug libraries.** While the CLOUD (a) consists dominantly of well-known and divers drugs, the LOPAC<sup>®1280</sup> library (b) can rather be considered as an explorative collection of compounds. Both libraries comprise a high variety of ATC classes. In (c), both libraries were merged.

Categories A (alimentary tract and metabolism), B (blood and blood forming organs), D (dermatologicals), G (genito urinary system and sex hormones), H (systemic hormonal preparations, excl. sex hormones and insulins) and V (various) are not present anymore in the CLoMP.

However, we note that the large majority of CLoMP drugs could not be classified into any ATC group, which makes it difficult to obtain an accurate representation of ATC classifications of the CLoMP.

### 6.3 Classification of CLoMP targets according to different databases

The classification of drugs into distinct anatomical target areas (ATC classes) provides a general overview about the underlying diversity of those compounds in terms of their anatomical location and their pharmacological effect.

Furthermore, the targets of our remaining, morphologically active drugs were gathered from different databases and analyzed, as we already did it in the characterization of the CLOUPAC. Initially, we used again the LOPAC<sup>®1280</sup>-provided data for those drugs that originate from the purchased library and which were equipped with the information about whether they target an enzyme or not. Additionally, we collected again all available drugs from the ChEMBL database via bioassays and classified them into the 14 different target classes. Lastly, we tried to match all found ChEMBL targets to the druggable proteome compiled in the course of the Human Protein Atlas Project.

#### LOPAC<sup>®1280</sup>-provided data

With reference to the information about the LOPAC<sup>®1280</sup> supplied by Sigma, we identified 20 out of 69 LOPAC<sup>®1280</sup> drugs that target an enzyme.

#### ChEMBL classification

For a comparative analysis between CLOUPAC and CLoMP all available targets were again collected via IC50 and Ki bioassay activities in ChEMBL, if the required criterium of a pchembl value greater than 5 was met, meaning that the concentration of the tested compound must not exceed the concentration of  $10^{-5}$  M or 10  $\mu$ M in their half-maximal effect, respectively. By this, we obtained 124 target ChEMBL IDs. The collected targets were subsequently divided into the 14 functional target groups as defined by ChEMBL. For a direct comparison between CLoMP, CLOUPAC and ChEMBL, the abundances of all three libraries are shown in Table 13 and depicted in Figure 19.

At a first glance at the target abundances of the CLoMP, two protein classes are again clearly standing out: enzymes and membrane receptors. To a certain degree, this observation was expected, since both groups already showed an exceedingly high abundance in the CLOUPAC as well as in the ChEMBL database. Moreover, ion channels and transporter proteins are present in a proportional amount when compared to the CLOUPAC and to ChEMBL. There are few targets that were classified as transcription factors and there is at least one epigenetic regulator, one other cytosolic protein and one secreted protein (Table 13). All other groups are not represented in the pool of CLoMP targets, but were already extremely rare in the CLOUPAC and in ChEMBL. Again, we can conclude that the CLoMP is

a miniature version of the merged CLOUD-LOPAC<sup>®1280</sup> compilation that is still targeting mainly enzymes and membrane receptors.

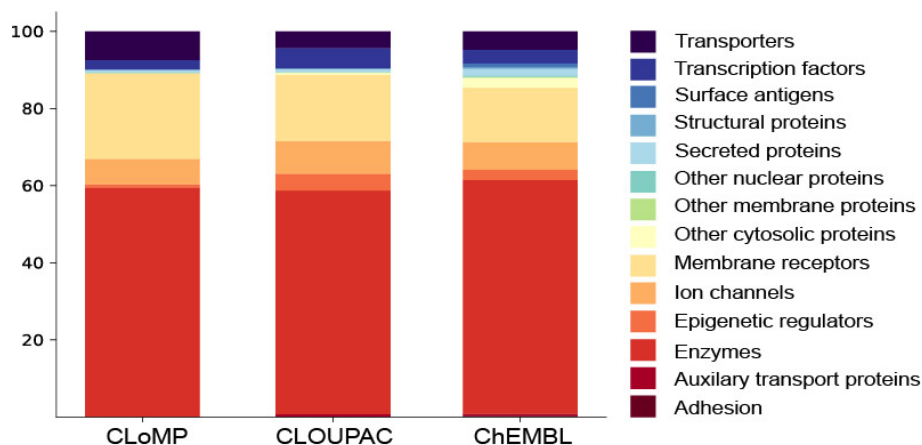


Figure 19: **Proportional comparison between target groups of the CLoMP, the CLOUPAC and the ChEMBL database.** Similar to the other libraries, enzymes are by far the most represented group in the CLoMP, followed by membrane receptors. Overall, the relative proportions are very similar in all three libraries.

### HPA classification

Next, we were interested into the categorization of targets according to the Human Protein Atlas classification system. Therefore, we collected again all available ensemble IDs of the 124 targets, which was necessary for matching them to the druggable proteome and the associated information about the function and the location of targets. An overlap with the druggable proteome was found for 70 targets.

Previously, it was shown that there is a high prevalence of enzymes and G-protein coupled receptors in the CLOUPAC. A very similar ranking could be observed for the abundances of CLoMP targets. However, other than in the CLOUPAC, the greatest part of proteins in the CLoMP is made up by G-protein coupled receptors, while enzymes are taking the second place. All other groups are rather weakly represented, whereas transporters are still the most frequent proteins among the rare ones in both cases. Figure 20 depicts the abundances of CLoMP and CLOUPAC targets among the different functional groups (exact numbers are listed in Table 20). Note that a target can be assigned to multiple functions.

Equally, the position of each target within the cell was recorded and classified by matching the targets to the druggable proteome. As shown in Figure 21, this approach returned again a similar distribution of CLoMP targets into the differ-

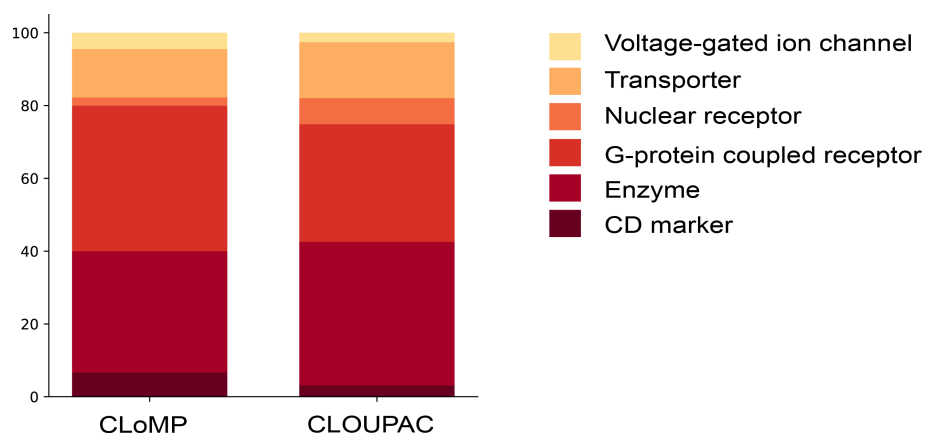


Figure 20: **Proportional comparison of protein functions between the CLoMP and CLOUPAC library.** In contrast to the ChEMBL database and the CLOUPAC, G-protein coupled receptors are the most abundant group of proteins within the CLoMP, enzymes only the second largest group.

ent classes as we observed it in the CLOUPAC, meaning that most proteins were found to be located in the membrane, followed by intracellular proteins. In contrast to the CLOUPAC, there were no proteins classified as ‘IM/SPTM’ or ‘secreted’ within our reduced set of targets (Table 21).

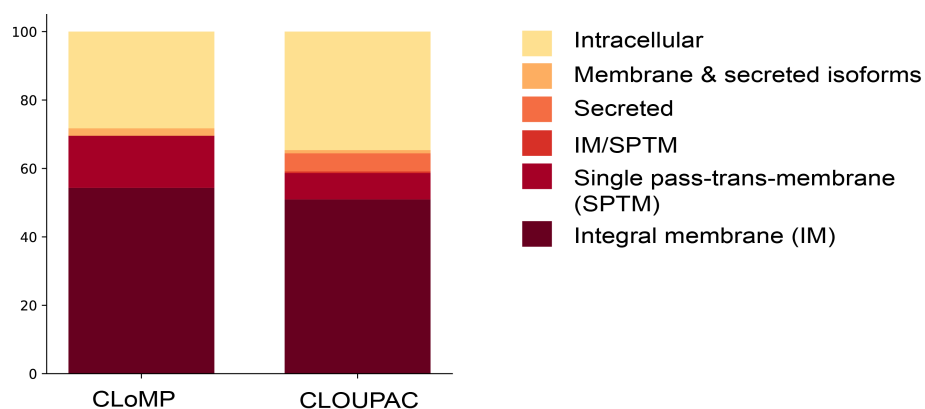


Figure 21: **Proportional comparison between target locations of the CLoMP and the CLOUPAC library.** In both drug collections, integral membrane proteins were found to be the most abundant group, followed by intracellular proteins.

## 6.4 Concentration curves: the strength of the effect depends on the concentration of the applied drug

The compilation of a novel drug library implicates not only finding suitable drugs, but as well the important question of what is the optimal compound concentration evoking a morphological change of the cell shape, but is simultaneously not harming or killing the cell. Since the collection of drugs was diverse and, other than in the CLOUD library, many compounds were not well described in terms of their pharmacologically effective amount, we decided to screen every agent in four different concentrations spanning several orders of magnitude, namely 0.02, 0.2, 2.0 and 20.0  $\mu\text{M}$ . This allowed us to observe the phenotypical change of the cells over increasing drug amounts by plotting the average duplicate vector lengths against the four concentration points and identifying the maximum of the curve, which corresponds to the strongest morphological effect.

Thereby, we discovered that the remaining drugs resulted in various types of concentration curves, such as ascending, declining or rather flat curves. In most cases, there was only one out of four concentrations exceeding the threshold and thus qualifying for the final CLoMP library, which was in most cases, but not in all, the highest concentration. Tables 22 and 23 (Supplementary material) list all 88 CLoMP drugs and the respective concentration, where we observed the strongest phenotypical effect.

Based on this observation, we were striving for a method for dividing the drugs into different clusters according to the pattern of their concentration curves. This would allow us to make statements about the presence of different types of morphological responses over increasing drug concentrations. To this end, we generated a heatmap that categorized the compounds into groups of drugs with resembling curves. This was done by correlating all pairs of concentration curves to each other and converting the linkage data into a distance matrix, as shown in Figure 22. As a result, an  $n \times n$  table was created, where clusters (dark squares) were formed along the diagonal indicating the distance between each pair of drugs by colors ranging from dark blue, for a very low distance (= high correlation), to light orange for a high distance (= low correlation) between two compounds.

In addition, the dendrogram depicts the hierarchical arrangement of the clusters produced by the above described distance matrix and allows to extract drugs that belong to the same cluster by determining a cut-off, which can be thought of a horizontal line cutting the tree at a certain distance. Depending on the height of the cut-off, varying amounts of clusters are formed and can subsequently be analysed or compared to each other. We chose a cut-off of 1.1531, whereby our CLoMP

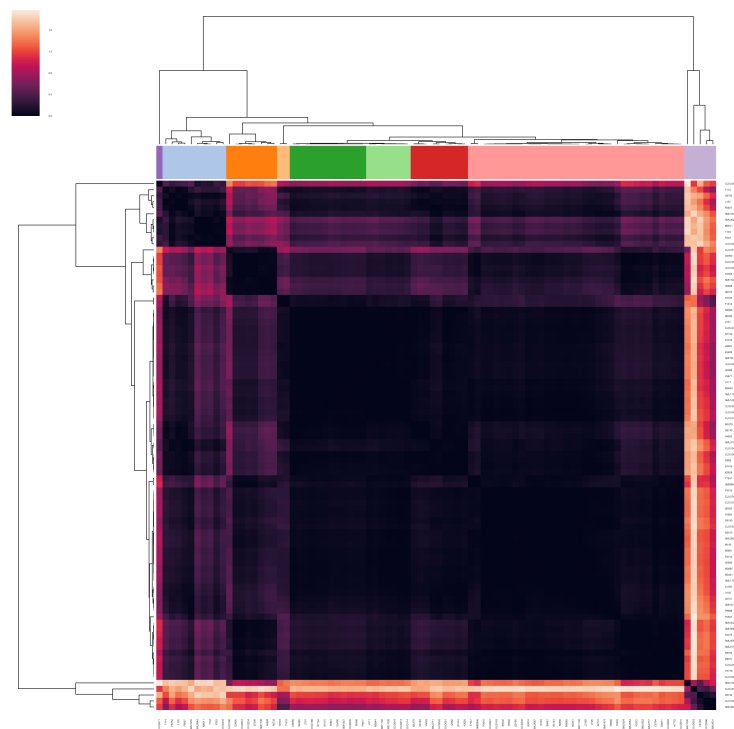


Figure 22: **Distance matrix based on pairwise correlations of concentration values between all pairs of drugs.** By correlating the four concentration values between all drugs pairwise, nine distinct clusters formed along the diagonal of the matrix that include drugs, whose concentration curves have similar shapes.

drugs were divided into nine distinct clusters, depending on the shape of their concentration curve. For a clearer visual understanding, the clusters are highlighted by different column colors in Figure 22.

In order to visually inspect the accuracy of the clustering, but as well to categorize the drugs according to different groups of concentration curves, we overlaid the plots of the individual clusters. As shown in Figure 23, we saw a clear resemblance of concentration curves within the same clusters and can thus assume that the clustering returned by the distance matrix as well as our selected threshold are reasonable.

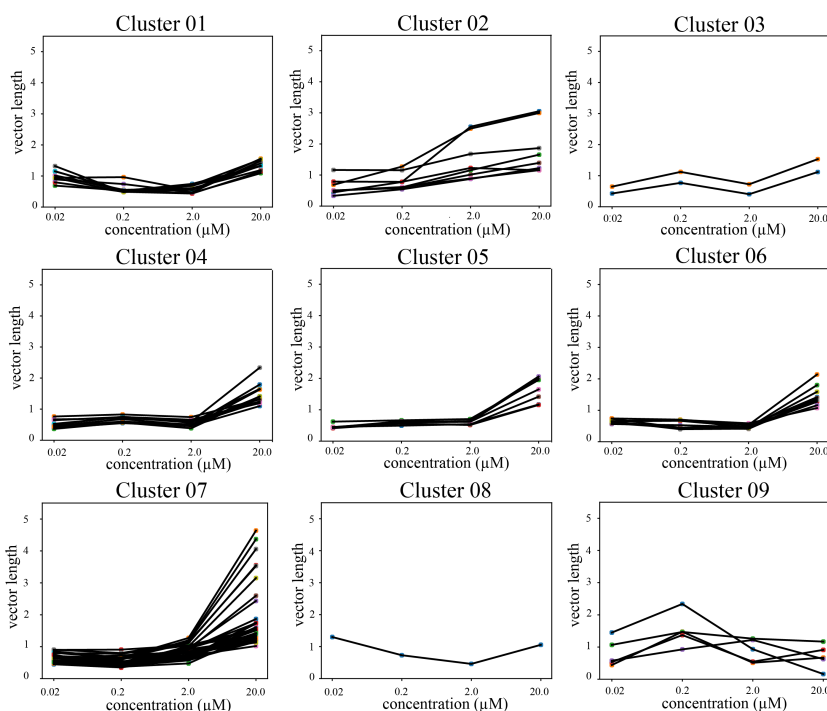


Figure 23: **Overlaid curves of the different clusters returned by the distance matrix.** The clustering allows to categorize drugs into different kinds of dose dependencies, meaning that each cluster shows a specific kind of concentration curve, which varies from rather flat curves to ascending or descending curves.

## 6.5 Outcome

Summarizing, the CLoMP consists of 88 distinct chemical compounds, whereof the majority originates from the LOPAC<sup>®1280</sup> library. However, 9.05% of the original CLOUD drugs, but only 4.63% of the original LOPAC<sup>®1280</sup> have passed the filtering process, meaning that CLOUD drugs are in relative terms stronger represented than LOPAC<sup>®1280</sup> drugs. The drugs can be split into a group of well annotated or approved compounds and a group of poorly characterized or hardly known compounds. There is only one prodrug among all CLoMP drugs, which is 'Proguanil'. Targets were found for approximately half of all compounds equipped with a ChEMBL ID. For many drugs there was as well information available about their MOA or disease indications.

Furthermore, the proportions of ATC classes has not dramatically changed in the CLoMP, if we compare it to the CLOUPAC, even though not all classes are represented anymore. However, we can only make statements about a small number of classified drugs, since the majority of CLoMP drugs does not have an ATC

classification. Within the group of annotated drugs, nervous system drugs are represented the most in both, the CLOUPAC and the CLoMP library. The proportion of antineoplastic drugs is slightly higher in the CLoMP than in the CLOUPAC. Additionally, there is a surprisingly high amount of antiparasitic agents among the classified CLoMP drugs, which are in contrast rather poorly represented in the CLOUPAC.

Moreover, we observed a strong similarity between CLOUPAC and CLoMP with regards to the relative proportions of target classes, meaning that most targeted proteins were found to be enzymes, followed by membrane receptors and ion channels. Very resembling proportions were as well observed by analyzing all existing targets within the ChEMBL database. By classifying the targets according to the druggable proteome of the Human Protein Atlas database, we noticed a similarly high amount of enzymes and G-protein coupled receptors in the CLoMP. Accordingly, most targets of the CLoMP were classified as integral membrane proteins, constituting the common cellular location of G-protein coupled receptors. Referring to drug concentrations, the large majority of drugs showed the strongest morphological effect when applied in the highest concentration of 20  $\mu$ M. In addition, we were able to cluster the CLoMP compounds according to the strength of their morphological effect over four concentrations. This resulted in nine clusters of drugs, each exhibiting a typical pattern of concentration dependency.

Ultimately, we clearly observed specific morphological phenotypes upon treating MCF10A cells with CLoMP compounds, ranging from small to large phenotypes, an increased amount of filopodia and different kinds of cell clustering, for example. With that, we achieved to establish a compound library that indeed showed to induce specific and significant morphological changes in cells, making it directly applicable for further morphology-based screens addressing various kinds of biological questions.

## **7 Summary and discussion**

Here, we described a strategic approach for designing the novel drug library ‘CLoMP’ that is based on compounds that significantly induce morphological changes in MCF10A cells. By carefully investigating previous publications dealing with morphologically active compounds, we obtained an initial collection of potentially interesting drugs. For reasons of simplicity and cost-effectiveness, we decided to purchase the LOPAC<sup>®1280</sup> compound library from Sigma-Aldrich and combined it with the in-house compound collection ‘CLOUD’, both being

enriched for morphologically active drugs. The resulting ‘CLOUPAC’ consisted of 1490 distinct small molecules, which we annotated extensively with available information of different databases, such as ATC classes, phases of clinical trials or targets. Subsequently, we performed a high-throughput/high-content imaging screen of the CLOUPAC, which we optimized in preceding test runs. By carefully considering the seeding number of cells, the drug concentrations, the duration of the treatment and a precise schedule for staining, our final protocol complies with optimal conditions for inducing and visualizing morphological changes of the cells and allowed to screen the whole CLOUPAC within only a few days. The images were acquired via microscopy in three different fluorescent channels, visualizing the cytoskeleton, the nucleus and the mitochondria of the cells. We processed the imaging data with the help of the image analysis software ‘CellProfiler’ by designing a specifically adapted pipeline of analysis modules, outputting 638 different features. We then elaborated a statistical and vector-based approach to reduce this amount to 66 meaningful measurements. By this procedure, we removed strongly fluctuating, non-informative or very redundant features and could additionally increase the speed for subsequent analysis steps. We further used feature vectors for validating the effect of the CLOUPAC compounds and their reproducibility as well as the most effective drug concentration by various statistical tests. In this way, we obtained a final set of compounds that consists of 88 distinct chemicals, which proved to dynamically change the cellular phenotype in very distinct and specific ways. In analogy to the CLOUD, we named the novel drug collection the ‘CeMM Library of Morphological Perturbagens’, or short ‘CLOMP’. We analyzed the remaining compounds once again in order to gain a broad insight into the main characteristics of the CLOMP. This library is now perfectly suitable and ready to use for various kinds of further studies that use cell morphology as readout, such as drug-drug, gene-drug or mechanism of action studies.

The CLOMP, as we designed it, consists to almost 80% of LOPAC<sup>®1280</sup> drugs and to slightly more than 20% of CLOUD drugs, which is a ratio that was within our expectations. With 72 out of 88 compounds that possess a ChEMBL ID, we can say that the CLOMP contains mainly annotated compounds. Around half of the characterized drugs were found to be approved, such as the antineoplastic agents ‘Imatinib’, ‘Sunitinib’ or ‘Pazopanib’ or the nervous system drugs ‘Citalopram’ and ‘Tomoxetine’, for example. The majority of the other half of annotated drugs is in phase 0 of clinical trials. Furthermore, ‘Proguanil’, the only active pro-drug compound that was identified, is commonly used as chemoprophylaxis and treatment for malaria by specifically inhibiting the parasitic dihydrofolate reductase (DHFR) [44].

Furthermore, the CLoMP exhibits a lower diversity in terms of ATC classes than the CLOUPAC, but still the compounds are classified into eight different groups - nervous system and antineoplastic drugs at the top. Reason for the reduction in ATC diversity is probably a drastic decrease of compounds in general, since the CLoMP constitutes only 6 percent of the CLOUPAC. Moreover, since we screened for morphologically active compounds on a specific cell line derived from a mammary gland, drugs targeting the alimentary tract, the blood or the skin, for example, were not expected to show significant effects. From this point of view, having 8 out of 14 ATC classes, the CLoMP still covers a broad spectrum of drug diversity.

A closer analysis of the CLoMP targets revealed that the ratios of the different protein classes clearly reflect the ratios observed in the CLOUPAC and in the ChEMBL database. In particular, enzymes and membrane receptors, which represent the two biggest groups of proteins, are present in similar proportions. This stands in contrast with the results of the target classification according to the drug-gable proteome from the Human Protein Atlas (HPA) database, where enzymes were found to be only the second most abundant group of targets, while G-protein coupled receptors were ranked first. Appropriately, the main target location was found to be the membrane. Reason for this difference in the classification of targets is with high probability the varying compositions of the databases.

Moreover, by clustering the drugs according to the effect curves that resulted from treating the cells with four increasing drug concentrations, we identified different types of dose dependencies and divided them into nine individual groups. The visual representation of the overlaid concentration curves facilitates finding the appropriate drug concentration for the individual clusters in order to obtain the strongest morphological effect. In most clusters, we observed a positive correlation between the morphological activity and the drug amount, meaning that the effect of a drug was increasing with increasing concentration, either in a consistent manner or only at the highest dose. Another cluster, however, indicates that the morphological activity of the drug decreases at a certain level of drug amount. Reason for that could be a high cytotoxicity of the drugs leading to cell death or some kind of drug inactivation, such as oxidation or evaporation. In few cases, we also observed slightly irregular curves, fluctuating at low levels of drug activity, which do not seem to have a clear maximum effect. Nevertheless, the clustering allows obtaining a clear image of how drugs affect cell morphology over several magnitudes, resulting in dividable patterns of concentration dependencies.

In general, we find that cell morphology can be considered as an extremely informative and powerful readout method. The cellular phenotype comprises mul-

tiparametric layers of information by responding in highly individual ways to chemical stimuli, which allows extracting large quantities of treatment-specific data. Indeed, we found various morphology classes evoked by the CLoMP drugs, such as small or large cells, increased numbers of filopodia or the formation of chain- or islet-like structures of cells. Correlations of phenotypes of well-studied drugs and poorly annotated compounds enables thus a variety of investigations, since it would e.g. raise the question of whether these drugs are related in their MOAs or acting within the same pathway. Additionally, the rich repertoire of morphologies can be useful to study the impact of two or more perturbations on each other, revealing specific drug-drug or even gene-drug interactions.

Moreover, given the nature of the CLoMP only containing morphologically active drugs, makes this library much more suitable for morphology-based screens, whereas other libraries would have more than 90 percent of the compounds showing no effect at all. We expect that the CLoMP would lower this number significantly in future studies. Additionally, we characterized and annotated our compound library in various ways. All this information can again help in downstream analyses, e.g. to associate drugs to yet unknown targets or MOAs, as well as to find underlying patterns based on morphologies.

Furthermore, we want to emphasize the great benefit of our analysis pipeline, which is based on a multidimensional threshold for filtering the compounds. By adjusting the individual thresholds, the CLoMP is able to grow or shrink in its size, depending on the individual requirements and the nature of the screen. Since our group aims to investigate in particular drug-drug or gene-drug interactions, the size of the library to be screened should in the best case not exceed the extent of a single 384- or even 96-well plate in order to avoid an enormous expansion of pair-wise combinations.

In conclusion, we consider the CLoMP as a versatile and divers drug library that represents a diligently selected collection of chemical compounds making it applicable for various kinds of morphology-based studies. Moreover, it cannot only be used as a look-up-table for finding morphological perturbors, but can additionally be extended and reduced for individual purposes. Since many CLoMP drugs are well annotated, specific morphologies could eventually be linked to particular drug targets or MOAs in downstream analyses and thus help in revealing underlying patterns, which allow to assign cell phenotypes to certain kinds of chemical perturbations. Altogether, we believe that the CLoMP constitutes a valuable and promising compound source in numerous ways.

## 8 Material & Methods

### 8.1 High-throughput high-content imaging screen

The high-throughput high-content imaging screen was conducted on the adherent, epithelial, non-tumorigenic cell line MCF10A, which is derived from a human mammary gland. All washing and dispensing steps were done by automated devices. The selected 1490 CLOUPAC drugs were transferred to 384-well plates by an acoustic liquid handler in the following concentrations: 0.02, 0.2, 2.0 and 20.0  $\mu$ M. Additionally, every sample was present in duplicates so that every drug was represented in 8 wells. After transferring the compounds to the plates, the MCF10A cells were seeded on top of the compounds at a density of approximately 1000 cells per well and incubated with the drugs for 72 hours at 37°C and 5% CO<sub>2</sub> in DMEM/F12 medium plus ingredients (see Table 7). Due to better practical handling, the plates were divided into two batches.

After the three-day treatment, the medium was aspirated and the cells were directly stained with a 1  $\mu$ M live-stain for mitochondria (MitoTracker, ThermoFisher). After 40 minutes, the cells were washed with 1xPBS and fixed with 3.7% formaldehyde. Subsequently, the cells were washed again with 1xPBS and blocked with 3% BSA in 1xPBS + 0.1% Triton-X for at least 30 minutes. The medium was aspirated and the primary antibody for  $\beta$ -tubulin (1:500, ThermoFisher) in 3% BSA, 0.1% Triton-X in PBS was added. The plates were incubated over night at 4°C. Cells were washed again with 1xPBS and the secondary antibody (anti-mouse, 1:1000, ThermoFisher) as well as 6-Diamidin-2-phenylindol (DAPI, ThermoFisher) were added (1:5000) and the cells were incubated for at least 1 hour in the dark. The cells were washed in a last step with 1xPBS and 50 $\mu$ L 1xPBS were added to each well.

We included 32 negative (DMSO) controls per plate. For image acquisition, four fields of view were captured per well in 20x resolution in three different channels (blue - nuclei, green - cytoskeleton, red - mitochondria) with the Operetta High-Content Imaging System (PerkinElmer). The raw images were directly used for image analysis using CellProfiler 3.0.

Table 7: **DMEM/F12 medium composition.**

| Ingredient     | Final concentration |
|----------------|---------------------|
| DMEM/F12       |                     |
| Horse serum    | 5%                  |
| Human EGF      | 20 ng/mL            |
| Hydrocortisone | 500 ng/mL           |
| Cholera Toxin  | 100 ng/mL           |
| Insulin        | 10 ug/mL            |
| Pen/Strep      | 1x                  |

## 8.2 CellProfiler pipeline

The CellProfiler 3.0 software is capable of distinguishing and identifying individual objects by itself, which is already working remarkably well on the raw image data. Still, there was a lot of improvement potential left for making the whole process working more robustly and reliably. All following descriptions of the modules relate to explanations from the CellProfiler website ([www.cellprofiler.org](http://www.cellprofiler.org)).

As mentioned above, the images were recorded in two separated batches, whereas the image acquisition parameters, such as focus and exposure time, had to be adapted individually for each batch. As a consequence, we considered it appropriate to adjust as well our CellProfiler pipeline to each batch of images, even if there were only minor differences. If not explicitly mentioned, the settings were the same for both batches.

### **LoadData**

At first, an input data file in CSV format was uploaded to CellProfiler that contained a list of all image file names to be processed. In addition, the image intensities were rescaled so that the camera's highest possible intensity value was rescaled to 1.

### **FlagImage**

By the usage of this module, images with extremely high or low values can be flagged and skipped in the analysis, if they do not meet a certain criterion. For batch 1, we defined 1.0 as minimum and maximum value. For batch 2, we used 2.0.

### **GreyToColor**

As soon as the data to analyze has been determined, the color scheme had to be selected, since all images were obtained only in black and white from the fluorescence microscope. The displayed colors were matched to the staining color of the appropriate cell compartment, meaning that nuclei will be shown in blue, the  $\beta$ -tubulin cytoskeleton in green and the mitochondria in red. The relative weight for each color was equally set to 1.0.

### **MeasureImageQuality**

This module serves as a quality control for a later distinction between good and bad images, meaning that upon usage of this module on the raw image data, several features of an image will be measured. The returned values would indicate different image aberrations, such as blurriness and unwanted intensity or saturation effects, which would be very likely to disturb the object recognition process. In terms of our analysis, the calculation of blur, intensity and saturation metrics were activated. The spatial scales for blur measurement were set to 30 and 75 and are explained more precisely in section 8.3 ‘Image quality control’.

### **CorrectIlluminationCalculate (primary objects - nuclei) (only for batch 1)**

As already explained above, primary object recognition is always the preceding step for secondary object recognition. For this reason, processing the blue channel images, which refer to DAPI-stained nuclei, was the initial step in our analysis pipeline. By calculating an illumination function, uneven exposure levels could be corrected by applying the associated module ‘CorrectIlluminationApply’ afterwards. For our purpose, we used a regular calculation of the illumination function, while rescaling the pixel intensities to equal or greater than 1. The function was calculated for each image individually and smoothed the resulting image by applying the ‘Fit Polynomial’ method.

### **CorrectIlluminationApply (primary objects - nuclei) (only for batch 1)**

By applying the previously calculated illumination function by division, uneven illumination, lighting or shading within images could be corrected.

### **RescaleIntensity (primary objects - nuclei)**

All blue-channel images were rescaled by using this module, meaning that the full intensity range of an image was scaled to the range between zero and one. We chose “Maximum for each image” as method of calculating the maximum intensity in batch 1 and chose a lower intensity limit of 0.008 for the input image. In case of batch 2, we decided to use a custom range between 0.008 to 0.6.

### **Smooth (primary objects - nuclei)**

Smoothing our images led to a more uniform pattern of primary objects, which helped to further reduce incorrect object recognition. For this, we used the Gaussian Filter and defined 4 pixels as a typical artifact diameter in batch 1, while increasing it to 5 pixels in batch 2.

### **IdentifyPrimaryObjects**

Running several modules that improved the distinguishability between nuclei and background, the first basic module for image analysis was applied, which was able to identify our primary objects to a satisfactory extent. We defined the expected diameter of the imaged nuclei to be between 20 and 70 pixels and discarded all objects exceeding this size. We chose a 'Global' threshold strategy by applying the 'Otsu' method with thresholding in two classes (foreground and background). Furthermore, we used a threshold smoothing scale of 1.3488 and a threshold correction factor of 0.8. Lower and upper bounds on the threshold were defined as 0.05 and 0.4 for batch 1, while setting them to 0.02 and 0.4 in batch 2. Clumped objects were distinguished by shape, which was the same method as for drawing the dividing line between them. The size of the smoothing filter, as well as the minimum allowed distance between local maxima were calculated automatically by using lower-resolution images. In addition, holes within the identified objects were filled after thresholding and declumping. We did not erase objects when they occurred in large numbers.

### **CorrectIlluminationCalculate (secondary objects - cells)**

#### **(only for batch 2)**

For images of batch 2, an illumination correction was additionally performed for secondary objects (cells). The illumination function was calculated for each image individually. As smoothing method, we chose 'Gaussian Filter' with an object diameter of 150 pixels.

### **CorrectIlluminationApply (secondary objects - cells)**

#### **(only for batch 2)**

We applied the previously calculated illumination function by division.

### **RescaleIntensity (secondary objects - cells)**

After the pipeline for primary object identification was optimized, the next set of images could be processed, namely all green-channel images referring to the fluorescently stained  $\beta$ -tubulin cytoskeleton. Other than the DAPI-images, we did not find a necessity for an illumination correction. The first processing step of secondary objects was thus the rescaling of pixel intensities. For that, we defined

a minimum of 0.005 (batch 1) or 0.0025 (batch 2) and a maximum of 1.0 for the input images, which were scaled to numbers between zero and one in the output images.

#### **Smooth (secondary objects - cells)**

For smoothing secondary objects, we chose the Median Filter with an artifact diameter of 2 pixels for batch 1, while in batch 2 we defined 3 pixels as typical artifact size to be blurred.

#### **IdentifySecondaryObjects**

By using the primary objects as input, secondary objects were identified by the 'Watershed - Image' method. We used a global threshold strategy with two-class 'Otsu' thresholding. We set the threshold smoothing scale to 2 and the threshold correction factor to 0.8. Same as in the primary object identification, holes within the objects were filled. Objects were discarded when they touched the border of the image.

#### **IdentifyTertiaryObjects**

As an additional necessary step, tertiary objects were identified, which relate to the cytoplasm. This was done by simply subtracting the primary objects (nuclei) from the secondary objects (cells). Thereby, we used the option to shrink the smaller objects before the subtraction.

#### **Smooth (secondary objects - mitochondria)**

Another smoothing module was applied for the third kind of objects, referring to mitochondria imaged in the red channel. We used the 'Median Filter' with a typical artifact diameter of 4.

#### **IdentifySecondaryObjects**

Similar to cell recognition, mitochondria were identified via the 'Watershed - Image' method and a global threshold strategy. We used the minimum cross entropy for thresholding, while setting the threshold smoothing scale to 2.0 and the threshold correction factor to 0.8. Lower and upper bounds were set to 0.0 and 1.0. Holes within the identified objects were filled and objects touching the border of the image were discarded.

At this point, the established CellProfiler pipeline included, in our view, the most important modules and was considered as suitable for the analysis. The hereafter described modules served now to measure a variety of features that involve area, intensity, shape or other parameters. Those measurements, which are the most relevant to us are the ones containing all kinds of morphological data. Therefore,

the module ‘MeasureObjectSizeShape’ represented the probably most essential element in our analysis. However, we included as well other kinds of features to measure, such as intensities or adjacent objects.

### **MeasureObjectOverlap**

This module measures the overlap of different identified areas. It was inserted in case there is an interest in comparing the area occupied by mitochondria to the cytoplasmatic area calculated by the subtraction of the nuclei from the cells. For this purpose, we selected the calculated cytoskeletal area as the ground truth and tested for an overlap with the area taken up by mitochondria.

### **MeasureGranularity**

The next module that we incorporated into our pipeline covered a range of size measurements on the texture of the image. Thereby, a variety of structure elements of increasing sizes is tried to match the image texture, resulting in a spectrum of measurements indicating how well they fit. We did this for all three channels by using the same settings: we used 0.25 as subsampling factor for granularity measurements, as well as for background reduction. We set the radius of the structuring element to 10 and the range of the granular spectrum to 16.

### **MeasureObjectIntensity**

By applying this module, intensity features for each identified object were returned. We used these measurements for nuclei and cells.

### **MeasureObjectIntensityDistribution**

This module uses previously identified objects and measures their spatial intensity distribution by defining the center of the object and creating radial bins towards the cell boundary, while measuring the intensity within each bin. We did this for cells and mitochondria and used the associated nuclei as centers. Therefore, we selected the objects to be radially scaled into 4 bins. Intensity Zernike moments were calculated while saving the magnitude and phase information. The maximum radial moment was set to 9.

### **MeasureObjectNeighbours**

Another type of measurement involves the count of neighbours that each object has. We used this module only for cells and chose the method “Adjacent” for determining neighbours, meaning that objects will only be counted as neighbouring if the two objects have adjacent boundary pixels. We did not make use of the option for retaining images showing the number of neighbours or the percentage of touching pixels.

### **MeasureObjectSizeShape**

Since we were mainly interested in the morphological changes of cells upon drug treatment, this module was for our analysis definitely the most important one. It is able to measure a variety of area and shape features of each identified object. We applied the module for all three kinds of objects and also included the Zernike shape features.

### **MeasureTexture**

By the application of this module, textures within a whole image were measured in order to evaluate its roughness or smoothness. We selected all green and all red channel objects and images to be measured and set the scale of texture to be measured to 3 pixel units.

### **ExportToDatabase**

With this module, the data was directly deposited into a MySQL database after the analysis. Thereby, we also included the calculation of per-image mean values for each type of object measurement and selected the option to export object relationships produced by the module ‘MeasureObjectNeighbours’. We decided to create a single object table in the database that comprises all object measurements and is easier to handle for the downstream data analysis.

### **CreateBatchFiles**

The last module was needed to produce files that are submitted to a cluster for processing batches of images in parallel, which increases the speed of the analysis process.

## **8.3 Image quality control**

The image analysis software CellProfiler provides the module “MeasureImageQuality” that is performing and collecting measurements on the acquired raw images for each single field of view, which can be used to find and remove images with various aberrations, such as fluorescent particles, highly saturated or blurry, defocused areas. The main reason for the need of excluding those “bad images” from the further analysis is that such artifacts lead to an incorrect separation of foreground objects and background. As a consequence, errors are likely to arise in the nuclei or cell segmentation process, resulting in wrong cell counts, erroneous feature or intensity measurements and many other issues. For those reasons, the recognition and removal of affected images before the analysis procedure is essential.

In sum, 51 quality control measurements are performed when this module is applied, which can be categorized into blur, saturation, intensity and threshold metrics. By having a more detailed look at the value distributions of the different measurements over all images, we found that in some cases the values were not normally distributed, but rather resulted in only very few different numbers, which did not allow to choose a proper threshold for separation, so they were excluded.

All normally distributed metrics were used to set a border between “good” and “bad” images, respectively to determine the percentile that yields the best possible partition, assuming that extremely high measurements correlate with flawed images.

In this way, the remaining metrics were tested individually, but also in various combinations, resulting in the empirical conclusion that the blur metric “correlation” appeared to be the best method to filter our images. What it does is to measure the intensity distribution across a chosen spatial scale of the image. If the image is blurred, neighboring pixels have a high correlation and produce therefore high values. In order to capture different kinds or sizes of blur, one small scale of 30 pixels and one bigger scale of 75 pixels were selected. In contrast, the other methods were not found to perform well in identifying blurry or over-saturated images or did only in one or two channels, but failed in the others. However, according to those correlation values and by trying various percentiles, the images were labelled as either “good” or “bad”.

Subsequently, parts of the images of both sets were inspected manually to validate how efficiently the selection worked. Since technical differences during the microscopy process could not be excluded, both batches were handled individually. Ultimately, the 96.5th percentile was selected for batch 1 and the 97.0th percentile for batch 2. To be more precise, we reviewed the 120 first images of the two prior established sets and evaluated them subjectively to count all true positive and negative images as well as all false positive and negative images. In this way, the confusion matrices below were compiled, which show the exact counts of the images of both batches.

For statistically measuring the accuracy of this method, sensitivity and specificity were determined, which was done for each batch separately. These measurements are interesting, since they give some indication of both, the proportion of “bad” images that were effectively identified as such (sensitivity) and equally the amount of “good” images that were correctly classified (specificity). This yielded similar results for both batches with a sensitivity of 0.84 for batch 1 and 0.85 for batch 2, and a specificity of 0.86 for batch 1 and 0.83 for batch 2.

### Batch 1

|                      |          | Subjective evaluation |             | Total |
|----------------------|----------|-----------------------|-------------|-------|
|                      |          | Positive              | Negative    |       |
| Quality measurements | Positive | $TP = 104$            | $FP = 16^*$ | 120   |
|                      | Negative | $FN = 20$             | $TN = 100$  | 120   |
| Total                |          | 124                   | 116         | 240   |

### Batch 2

|                      |          | Subjective evaluation |             | Total |
|----------------------|----------|-----------------------|-------------|-------|
|                      |          | Positive              | Negative    |       |
| Quality measurements | Positive | $TP = 99$             | $FP = 21^*$ | 120   |
|                      | Negative | $FN = 17$             | $TN = 103$  | 120   |
| Total                |          | 116                   | 124         | 240   |

TP (true positives) = “good” image correctly classified as “good”; TN (true negatives) = “bad” image correctly classified as “bad”; FP (false positives) = “bad” image not recognized as “bad”; FN (false negatives) = “good” image classified as “bad”.

11 of the 16 FPs (\*) of batch 1 as well as 11 of the 21 FPs (\*) of batch 2 were not classified as “bad”, even though they exhibited illumination effects. Anyway, the artifacts were in these cases relatively weak and would most likely not have disturbed the segmentation process. Thus, if these images were counted as TPs, the specificity for batch 1 would rise to 0.96, the sensitivity to 0.93. For batch 2, the specificity would rise to 0.92, the sensitivity to 0.95.

This approach allowed us to create a list of all acceptable fields of view. However, it was considered as reasonable to additionally judge the quality of an image according to the number of its fields rated as tolerable. In other words, if there was one out of four fields per image that was considered as inadequate, only this field has been removed, while the other three were kept. In contrast, if two or more foci of the same image exceeded the threshold, the whole image, meaning all four fields, has been removed, since it could be assumed that contaminations or other artifacts were spreading as well over the remaining fields. In this way, 685 wells, corresponding to 5.27%, were completely lost, whereas 29 of them were DMSO wells. The remaining 13.008 wells were still represented with at least three fields per well. Interestingly, there was one drug, ‘Gemcitabine hydrochloride’, where all eight wells were considered as “bad” images and were consequently removed. Hence, this compound was the only one out of all 1490 drugs that was excluded

from the analysis. The remaining imaging data was further processed.

## **8.4 Normalizing measurement results**

### **Plate normalization (Tukey's median polish)**

A commonly observed phenomenon when growing cells in multiwell plates is the influence of the position of the well on the final outcome, which should be taken into account. This means, beside the actual interesting effect, which is the impact of the applied drug on the cell, there are generally two more factors contributing to the measured result and those are the row and the column effect. One way for revealing the real drug effect is the usage of the median polish by Tukey, which allows the dissection of the variable into its different effects via an iterative process. Thereby, row and column medians are removed from the original value repeatedly until both medians approach zero.

As described above, the values of a number of wells were completely removed, since the corresponding images were declared as blurry or artifactual. However, in order to circumvent the fact that due to this, some of the wells remain empty and disturb the normalization process, the average DMSO value of the whole plate was used for these wells. All subsequent steps were performed using the normalized values.

### **Normalizing values between 0 and 1**

Furthermore, there was the necessity for standardizing the values for each feature within the range of zero and one. This allowed an adequate and more informative comparison of the features, since depending on the respective feature the range of values, meaning the minimum and maximum, differ greatly. Without normalizing the values, the following feature analysis would not be reasonable. For normalization, the highest and the lowest values per feature were identified with neglecting all outlier values below the 0.5th and above the 95.5th percentile.

The next step was to set the minimal value to zero and the maximal value to one, while accordingly assigning each value in between a normalized value between zero and one. All outlier values below the minimum were set to zero and those above the maximum were set to one. All following steps were performed using the filtered and normalized data.

## **8.5 Approach for finding the best subset of features**

We applied different combinations of filters (see 5.5) in order to reduce the initial amount of 638 features to a small set of features that was able to provide reliable

and significant information about morphological changes of the cells after drug treatment.

Out of the resulting features, multidimensional vectors were created for each drug, where each feature was representing one dimension. By means of the cosine similarity, which measures the cosine of an angle, the proximity between all pairs of vectors was calculated in a subsequent step. The smaller the angle, the more similar are both vectors to each other in terms of their morphological properties. This information allowed us now to differentiate between more and less convenient feature compositions. As an explanation, we divided all drug vectors into different groups: ‘Replicates’, which comprise only replicate vectors, ‘Same Drug’, which includes all available vectors of the same drug independent of the concentration and the collection ‘All Drugs’. From a suitable set of features we would expect generally high cosine similarities between replicates and moderately high similarities among samples of the same drug, but in different concentrations (‘Same Drug’), while the entirety of all drug vectors (‘All Drugs’) is not expected to show a significant proximity. Clear differences between the three groups would suggest that the applied measurements are reproducible and that the selected set of features is able to distinguish between different kinds of perturbations. As a matter of noise reduction, we filtered additionally for long vectors, meaning that only vector lengths higher than the mean of all drug vectors plus one standard deviation were accepted.

## References

- [1] Drugbank statistics. <https://www.drugbank.ca/stats/>. Accessed: 2019-04.
- [2] ChEML database. <https://www.ebi.ac.uk/chembl/g/#browse/drugs>. Accessed: 2019-04.
- [3] Cambridge University Press. *Cambridge Advanced Learner's Dictionary & Thesaurus*. Cambridge University Press, 2008.
- [4] Cambridge University Press. *Cambridge Business English Dictionary*. Cambridge University Press.
- [5] Juergen Drews. In *Quest of Tomorrow's Medicine*, pages 51–68. Springer-Verlag, New York, 1999.
- [6] Jürgen Drews. Drug discovery: a historical perspective. *science*, 287(5460):1960–1964, 2000.
- [7] Michael Eckert. Max von laue and the discovery of x-ray diffraction in 1912. *Annalen der Physik*, 524(5):A83–A85, 2012.
- [8] Jennifer Griffiths. A brief history of mass spectrometry. *Anal. Chem*, 80(15):5678–83, 2008.
- [9] Alexander Fleming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. *Bulletin of the World Health Organization*, 79:780–790, 2001.
- [10] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.
- [11] James Inglese and Douglas S Auld. High throughput screening (hts) techniques: applications in chemical biology. *Wiley Encyclopedia of Chemical Biology*, pages 1–15, 2007.
- [12] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren VS Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, et al. Impact of high-throughput screening in biomedical research. *Nature reviews Drug discovery*, 10(3):188, 2011.
- [13] Elisa Michelini, Luca Cevenini, Laura Mezzanotte, Andrea Coppa, and Aldo Roda. Cell-based assays: fuelling drug discovery. *Analytical and bioanalytical chemistry*, 398(1):227–238, 2010.

- [14] Emanuele Perola. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *Journal of medicinal chemistry*, 53(7):2986–2997, 2010.
- [15] Christina Nicole Ramirez, Christophe Antczak, and Hakim Djaballah. Cell viability assessment: toward content-rich platforms. *Expert opinion on drug discovery*, 5(3):223–233, 2010.
- [16] Marco P Licciardello, Anna Ringler, Patrick Markt, Freya Klepsch, Charles-Hugues Lardeau, Sara Sdelci, Erika Schirghuber, André C Müller, Michael Caldera, Anja Wagner, et al. A combinatorial screen of the cloud uncovers a synergy targeting the androgen receptor. *Nature chemical biology*, 13(7):771, 2017.
- [17] Terry L Riss, Richard A Moravec, Andrew L Niles, Sarah Duellman, Hélène A Benink, Tracy J Worzella, and Lisa Minor. Cell viability assays. 2016.
- [18] Kenneth A Giuliano, Robbin L DeBiasio, R Terry Dunlay, Albert Gough, Joanne M Volosky, Joseph Zock, George N Pavlakis, and D Lansing Taylor. High-content screening: a new approach to easing key bottlenecks in the drug discovery process. *Journal of Biomolecular Screening*, 2(4):249–259, 1997.
- [19] Fabian Zanella, James B Lorens, and Wolfgang Link. High content screening: seeing is believing. *Trends in biotechnology*, 28(5):237–245, 2010.
- [20] Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-based high-content screening. *Cell*, 163(6):1314–1325, 2015.
- [21] Joseph X Zhou, Zerrin Isik, Caide Xiao, Irit Rubin, Stuart A Kauffman, Michael Schroeder, and Sui Huang. Systematic drug perturbations on cancer cells reveal diverse exit paths from proliferative state. *Oncotarget*, 7(7):7415, 2016.
- [22] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.
- [23] Vebjorn Ljosa, Peter D Caie, Rob Ter Horst, Katherine L Sokolnicki, Emma L Jenkins, Sandeep Daya, Mark E Roberts, Thouis R Jones, Shantanu Singh, Auguste Genovesio, et al. Comparison of methods for image-based

- profiling of cellular morphological responses to small-molecule treatment. *Journal of biomolecular screening*, 18(10):1321–1329, 2013.
- [24] Jan Kapuscinski. Dapi: a dna-specific fluorescent probe. *Biotechnic & Histochemistry*, 70(5):220–233, 1995.
  - [25] Jonas Bucevicius, Grazvydas Lukinavicius, and Ruta Gerasimaite. The use of hoechst dyes for dna staining and beyond. *Chemosensors*, 6:18, 04 2018.
  - [26] Michael J Ignatius and Jeffrey T Hung. Physiological indicators of cell function. In *High Content Screening*, pages 233–244. Springer, 2007.
  - [27] Marco Breinig, Felix A Klein, Wolfgang Huber, and Michael Boutros. A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells. *Molecular systems biology*, 11(12):846, 2015.
  - [28] Yushi Futamura, Makoto Kawatani, Sayaka Kazami, Kenichi Tanaka, Makoto Muroi, Takeshi Shimizu, Koji Tomita, Nobumoto Watanabe, and Hiroyuki Osada. Morphobase, an encyclopedic cell morphology database, and its use for drug target identification. *Chemistry & biology*, 19(12):1620–1630, 2012.
  - [29] 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, 2016.
  - [30] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.
  - [31] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507, 2011.
  - [32] Lee Kamentsky, Thouis R Jones, Adam Fraser, Mark-Anthony Bray, David J Logan, Katherine L Madden, Vebjorn Ljosa, Curtis Rueden, Kevin W Elieci, and Anne E Carpenter. Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. *Bioinformatics*, 27(8):1179–1180, 2011.
  - [33] Lopac1280 – the library of pharmacologically active compounds. <https://www.sigmaaldrich.com/life-science/cell-biology/>

- bioactive-small-molecules/lopac1280-navigator.html. Accessed: 2019-03.
- [34] Cts – the chemical translation service. <http://cts.fiehnlab.ucdavis.edu/>. Accessed: 2019-03.
- [35] Clinical trials explained. <https://clinicaltrialsexplained.com/what-is-a-clinical-trial/>. Accessed: 2019-04.
- [36] Jarkko Rautio, Nicholas A Meanwell, Li Di, and Michael J Hageman. The expanding role of prodrugs in contemporary drug design and development. *Nature Reviews Drug Discovery*, 2018.
- [37] Who collaboration centre for drug statistics methodology. [https://www.whocc.no/atc/structure\\_and\\_principles/](https://www.whocc.no/atc/structure_and_principles/). Accessed: 2019-03-30.
- [38] Alexander S Hauser, Misty M Attwood, Mathias Rask-Andersen, Helgi B Schiöth, and David E Gloriam. Trends in gpcr drug discovery: new agents, targets and indications. *Nature reviews Drug discovery*, 16(12):829, 2017.
- [39] Suhas Vasaikar, Pooja Bhatia, Partap Bhatia, and Koon Chu Yaiw. Complementary approaches to existing target based drug discovery for identifying novel drug targets. *Biomedicines*, 4(4):27, 2016.
- [40] Michael S Kinch, Denton Hoyer, Eric Patridge, and Mark Plummer. Target selection for fda-approved medicines. *Drug discovery today*, 20(7):784–789, 2015.
- [41] Ensembl stable ids. [https://www.ensembl.org/info/genome/stable\\_ids/index.html](https://www.ensembl.org/info/genome/stable_ids/index.html). Accessed: 2019-04.
- [42] Shlomo S Sawilowsky. New effect size rules of thumb. 2009.
- [43] J Cohen. Statistical power analysis for the behavioral sciences. routledge. *New York, NY*, 1988.
- [44] Andrea K Boggild, Monica E Parise, Linda S Lewis, and Kevin C Kain. Atovaquone-proguanil: report from the cdc expert meeting on malaria chemoprophylaxis (ii). *The American journal of tropical medicine and hygiene*, 76(2):208–223, 2007.

## 9 Supplementary material

### 9.1 Double occuring drugs

Table 8: List of drugs appearing in both, CLOUD and LOPAC<sup>®1280</sup> library and were kept twice.

| LOPAC drug name                                 | LOPAC ID | ChEMBL ID     | CLOUD drug name   | CLOUD ID | ChEMBL ID     |
|-------------------------------------------------|----------|---------------|-------------------|----------|---------------|
| Argatroban monohydrate                          | A0487    | -             | Argatroban        | CLOUD008 | CHEMBL1166    |
| Fluvoxamine maleate                             | F2802    | CHEMBL1409    | Fluvoxamine       | CLOUD033 | CHEMBL814     |
| Eptifibatide acetate                            | SM11042  | -             | Eptifibatide      | CLOUD090 | CHEMBL1174    |
| Tetramisole hydrochloride                       | T1512    | CHEMBL554891  | Levamisole        | CLOUD097 | CHEMBL1454    |
| Meclofenamic Acid Sodium                        | M4531    | CHEMBL876     | Meclofenamic Acid | CLOUD166 | CHEMBL509     |
| L-alpha-Methyl-p-tyrosine                       | M8131    | CHEMBL1200862 | Metyrosine        | CLOUD202 | CHEMBL1200862 |
| DL-alpha-Difluoromethyl-ornithine hydrochloride | D193     | CHEMBL536971  | Eflornithine      | CLOUD207 | CHEMBL830     |
| Quinine sulfate                                 | Q1250    | CHEMBL170     | Quinine           | CLOUD232 | CHEMBL170     |

### 9.2 Classification of CLOUPAC

Table 9: Classification of CLOUPAC drugs into the four phases of clinical research.

| Phase of approval | Number of drugs |
|-------------------|-----------------|
| IV (approved)     | 537             |
| III               | 25              |
| II                | 21              |
| I                 | 4               |
| 0                 | 629             |
| not specified     | 274             |
| Total             | 1490            |

**Table 10: Classification of the LOPAC<sup>®</sup>1280 drugs into various functional groups (I).**  
Sigma-Aldrich classifies the LOPAC<sup>®</sup>1280 drugs into 94 different classes. Drug classes marked with \* are involved into some neurotransmitting actions.

| Drug class                | #drugs | Drug class                             | #drugs |
|---------------------------|--------|----------------------------------------|--------|
| Actin                     | 1      | Cytokines, Growth Factors and Hormones | 21     |
| Adenosine                 | 34     | Cytoskeleton and ECM                   | 9      |
| Adrenoceptor*             | 84     | DNA                                    | 8      |
| Alzheimer's disease       | 1      | DNA Metabolism                         | 10     |
| Androgen                  | 1      | DNA Repair                             | 3      |
| Angiogenesis              | 3      | DNA-RNA transcription regulators       | 1      |
| Angiotensin               | 2      | Dopamine*                              | 82     |
| Anhydrases                | 1      | Endosytosis                            | 1      |
| Antibiotic                | 29     | Endothelin                             | 1      |
| Anticoagulants            | 2      | Enzyme inhibitors                      | 1      |
| Anticonvulsant            | 9      | Epigenetics                            | 1      |
| Antifungal                | 4      | Estrogen                               | 1      |
| Antioxidants              | 2      | Extracellular Matrix                   | 1      |
| Antiviral                 | 4      | GABA*                                  | 32     |
| Apoptosis                 | 34     | Gene Regulation                        | 33     |
| Arachidonic acid cascade  | 1      | Glucocorticoid                         | 1      |
| Benzodiazepine            | 2      | Glutamate*                             | 68     |
| Bioactive Small Molecules | 1      | Glycine                                | 1      |
| Biochemistry              | 39     | G protein                              | 3      |
| Ca <sup>2+</sup> channel  | 26     | Heat shock proteins                    | 1      |
| Cation channel            | 1      | Histamine                              | 34     |
| Cell Biology              | 4      | Hormone                                | 28     |
| Cell Cycle                | 18     | Imidazoline                            | 8      |
| Cell Signaling (CS)       | 26     | Immune system                          | 17     |
| CS and Neuroscience       | 15     | Inflammation                           | 2      |
| Cell Stress               | 21     | Intracellular Ca <sup>2+</sup>         | 4      |
| Cholecystokinin           | 3      | Ion channels/pumps                     | 32     |
| Cholinergic*              | 73     | K <sup>+</sup> channel                 | 23     |
| Cl <sup>-</sup> channel   | 3      | Kinase inhibitors                      | 7      |
| Cyclic Nucleotides        | 25     | Kinase/Phosphatase inhibitor           | 9      |
| Cytokines                 | 7      | Leukotriene                            | 4      |

Table 11: **Classification of the LOPAC<sup>®</sup>1280 drugs into various functional groups (II).**

| Drug class                   | #LOPAC drugs | Drug class            | #LOPAC drugs |
|------------------------------|--------------|-----------------------|--------------|
| Lipids                       | 21           | Phosphorylation       | 67           |
| Melatonin                    | 4            | Phosphatase inhibitor | 1            |
| Multi drug resistance*       | 12           | PPAR                  | 1            |
| Muscarinic                   | 1            | Prostaglandin         | 19           |
| Na <sup>+</sup> channel      | 18           | Protease              | 1            |
| Neurobiology                 | 8            | Purinoceptor          | 1            |
| Neurotransmission*           | 62           | Serotonin*            | 67           |
| Nitric oxide                 | 26           | Sigma receptor        | 1            |
| Nootropic                    | 3            | Somatostatin          | 1            |
| Opioid                       | 16           | Sphingolipid          | 4            |
| Orexin                       | 1            | Tachykinin            | 5            |
| Oxidative stress             | 1            | Transcription         | 12           |
| P2 receptor                  | 9            | Tyrosine kinase       | 5            |
| Phosphodiesterase inhibitors | 2            | Vanilloid*            | 4            |
| Phospholipase inhibitor      | 1            | Not classified        | 2            |

Table 12: **Classification of CLOUD, LOPAC<sup>®1280</sup> and CLOUPAC drugs into the 14 ATC classes.**

| ATC first letter code | Group description                                               | CLOUD drugs | LOPAC <sup>®1280</sup> drugs | CLOUPAC drugs |
|-----------------------|-----------------------------------------------------------------|-------------|------------------------------|---------------|
| A                     | Alimentary tract and metabolism                                 | 16          | 23                           | 39            |
| B                     | Blood and blood forming organs                                  | 5           | 6                            | 10            |
| C                     | Cardiovascular system                                           | 24          | 45                           | 68            |
| D                     | Dermatologicals                                                 | 6           | 16                           | 22            |
| G                     | Genito urinary system and sex hormones                          | 10          | 9                            | 19            |
| H                     | Systemic hormonal preparations, excl. sex hormones and insulins | 5           | 4                            | 9             |
| J                     | Antiinfectives for systemic use                                 | 15          | 25                           | 40            |
| L                     | Antineoplastics and immunomodulating agents                     | 23          | 26                           | 49            |
| M                     | Musculo-skeletal system                                         | 8           | 17                           | 24            |
| N                     | Nervous system                                                  | 31          | 46                           | 77            |
| P                     | Antiparasitic products, insecticides and repellents             | 15          | 6                            | 20            |
| R                     | Nervous system                                                  | 8           | 18                           | 26            |
| S                     | Sensory organs                                                  | 4           | 21                           | 25            |
| V                     | Various                                                         | 1           | 6                            | 7             |
| Not classified        |                                                                 | 60          | 1059                         | 1122          |
| Total                 |                                                                 | 210         | 1280                         | 1490          |

Table 13: **Classification of targets into 14 functional groups according to the ChEMBL database.**

| Protein class               | #targets CLoMP | #targets CLOUPAC | #targets ChEMBL |
|-----------------------------|----------------|------------------|-----------------|
| Adhesion                    | 0              | 0                | 21              |
| Auxiliary transport protein | 0              | 4                | 35              |
| Enzyme                      | 72             | 333              | 4380            |
| Epigenetic regulator        | 1              | 25               | 191             |
| Ion channel                 | 8              | 49               | 523             |
| Membrane receptor           | 27             | 98               | 1014            |
| Other cytosolic protein     | 0              | 3                | 186             |
| Other membrane protein      | 0              | 1                | 20              |
| Other nuclear protein       | 0              | 0                | 25              |
| Secreted protein            | 1              | 5                | 120             |
| Structural protein          | 0              | 1                | 36              |
| Surface antigen             | 0              | 0                | 68              |
| Transcription factor        | 3              | 30               | 256             |
| Transporter                 | 9              | 25               | 345             |
| Unclassified proteins       | 3              | 14               | 1019            |
| Total                       | 124            | 588              |                 |

Table 14: **Classification of CLOUPAC and 'druggable proteome' targets into 6 functional groups according to the HPA database.**

| Protein class              | Number of targets of CLOUPAC | Number of targets in druggable proteome |
|----------------------------|------------------------------|-----------------------------------------|
| CD marker                  | 6                            | 71                                      |
| Enzyme                     | 77                           | 260                                     |
| G-protein coupled receptor | 63                           | 98                                      |
| Nuclear receptor           | 14                           | 24                                      |
| Transporter                | 30                           | 149                                     |
| Voltage-gated ion channel  | 5                            | 49                                      |
| Not classified             | 421                          | 120                                     |
| Total                      | 616                          | 771                                     |

**Table 15: Classification of CLOUPAC and 'druggable proteome' targets into 6 locations according to the HPA database.**

| Protein class                     | Number of targets of CLOUPAC | Number of targets in druggable proteome |
|-----------------------------------|------------------------------|-----------------------------------------|
| Integral membrane (IM)            | 106                          | 321                                     |
| Single pass-trans-membrane (SPTM) | 16                           | 118                                     |
| IM/SPTM                           | 1                            | 10                                      |
| Secreted                          | 11                           | 83                                      |
| Membrane & secreted isoforms      | 2                            | 34                                      |
| Intracellular                     | 72                           | 205                                     |
| Not classified                    | 408                          | -                                       |
| Total                             | 616                          | 771                                     |

### 9.3 Morphology-associated features

Table 16: List of remaining CellProfiler features (I)

Mean\_Cytoskeleton\_AreaShape\_Extent  
Mean\_Cytoskeleton\_AreaShape\_MedianRadius  
Mean\_Cytoskeleton\_AreaShape\_Zernike\_1\_1  
Mean\_Cytoskeleton\_AreaShape\_Zernike\_6\_4  
Mean\_Cytoskeleton\_AreaShape\_Zernike\_7\_7  
Mean\_Cytoskeleton\_AreaShape\_Zernike\_8\_6  
Mean\_Cytoskeleton\_Intensity\_IntegratedIntensityEdge\_Mitotracker  
Mean\_Cytoskeleton\_Intensity\_IntegratedIntensity\_Mitotracker  
Mean\_Cytoskeleton\_Intensity\_MassDisplacement\_Mitotracker  
Mean\_Cytoskeleton\_Intensity\_MeanIntensityEdge\_Mitotracker  
Mean\_Cytoskeleton\_Intensity\_MedianIntensity\_Mitotracker  
Mean\_Cytoskeleton\_Neighbors\_AngleBetweenNeighbors\_Adjacent  
Mean\_Cytoskeleton\_Neighbors\_PercentTouching\_Adjacent  
Mean\_Cytoskeleton\_RadialDistribution\_FracAtD\_BetaTubulin\_3of4  
Mean\_Cytoskeleton\_RadialDistribution\_MeanFrac\_Mitotracker\_1of4  
Mean\_Cytoskeleton\_RadialDistribution\_ZernikeMagnitud\_Mttrckr\_3\_3  
Mean\_Cytoskeleton\_RadialDistribution\_ZernikePhase\_Mitotracker\_6\_0  
Mean\_Cytoskeleton\_Texture\_AngularSecondMoment\_BetaTubulin\_3\_00  
Mean\_Cytoskeleton\_Texture\_InfoMeas1\_BetaTubulin\_3\_01  
Mean\_Cytoskeleton\_Texture\_InfoMeas1\_Mitotracker\_3\_00  
Mean\_Cytoskeleton\_Texture\_InverseDifferenceMoment\_Mitotrckr\_3\_03  
Mean\_Cytoskeleton\_Texture\_SumVariance\_BetaTubulin\_3\_01  
Mean\_Mitochondria\_AreaShape\_Extent  
Mean\_Mitochondria\_AreaShape\_MajorAxisLength  
Mean\_Mitochondria\_AreaShape\_MinFeretDiameter  
Mean\_Mitochondria\_AreaShape\_Solidity  
Mean\_Mitochondria\_AreaShape\_Zernike\_3\_1  
Mean\_Mitochondria\_AreaShape\_Zernike\_4\_2  
Mean\_Mitochondria\_AreaShape\_Zernike\_4\_4  
Mean\_Mitochondria\_AreaShape\_Zernike\_5\_3  
Mean\_Mitochondria\_AreaShape\_Zernike\_6\_4  
Mean\_Mitochondria\_AreaShape\_Zernike\_6\_6  
Mean\_Mitochondria\_AreaShape\_Zernike\_7\_5  
Mean\_Mitochondria\_AreaShape\_Zernike\_8\_6  
Mean\_Mitochondria\_AreaShape\_Zernike\_8\_8  
Mean\_Mitochondria\_AreaShape\_Zernike\_9\_5

**Table 17: List of remaining CellProfiler features (II)**

Mean\_Mitochondria\_AreaShape\_Zernike\_9\_7  
 Mean\_Mitochondria\_Granularity\_1\_Mitotracker  
 Mean\_Mitochondria\_Granularity\_2\_Mitotracker  
 Mean\_Mitochondria\_RadialDistribution\_FracAtD\_BetaTubulin\_1of4  
 Mean\_Mitochondria\_RadialDistribution\_FracAtD\_BetaTubulin\_3of4  
 Mean\_Mitochondria\_RadialDistribution\_MeanFrac\_BetaTubulin\_1of4  
 Mean\_Mitochondria\_RadialDistribution\_MeanFrac\_Mitotracker\_1of4  
 Mean\_Mitochondria\_RadialDistribution\_RadialCV\_Mitotracker\_4of4  
 Mean\_Mitochondria\_RadialDistribution\_ZernikeMagnitud\_Mttrckr\_6\_0  
 Mean\_Mitochondria\_RadialDistribution\_ZernikeMagnitud\_Mttrckr\_9\_9  
 Mean\_Mitochondria\_RadialDistribution\_ZernikePhase\_BetaTubulin\_6\_0  
 Mean\_Mitochondria\_RadialDistribution\_ZernikePhase\_Mitotracker\_8\_0  
 Mean\_Nuclei\_AreaShape\_Compactness  
 Mean\_Nuclei\_AreaShape\_Extent  
 Mean\_Nuclei\_AreaShape\_MajorAxisLength  
 Mean\_Nuclei\_AreaShape\_Solidity  
 Mean\_Nuclei\_AreaShape\_Zernike\_2\_0  
 Mean\_Nuclei\_AreaShape\_Zernike\_6\_0  
 Mean\_Nuclei\_AreaShape\_Zernike\_6\_2  
 Mean\_Nuclei\_AreaShape\_Zernike\_6\_4  
 Mean\_Nuclei\_AreaShape\_Zernike\_8\_2  
 Mean\_Nuclei\_AreaShape\_Zernike\_8\_6  
 Mean\_Nuclei\_Granularity\_1\_DAPI  
 Mean\_Nuclei\_Granularity\_2\_DAPI  
 Mean\_Nuclei\_Granularity\_3\_DAPI  
 Mean\_Nuclei\_Granularity\_4\_DAPI  
 Mean\_Nuclei\_Granularity\_6\_DAPI  
 Mean\_Nuclei\_Intensity\_IntegratedIntensity\_DAPI  
 Mean\_Nuclei\_Intensity\_MassDisplacement\_DAPI  
 Mean\_Nuclei\_Intensity\_StdIntensityEdge\_DAPI

## 9.4 Classification of CLoMP

Table 18: Classification of CLoMP drugs into the four phases of clinical research.

| Phase of approval | Number of drugs |
|-------------------|-----------------|
| IV (approved)     | 33              |
| III               | 3               |
| II                | 1               |
| I                 | 0               |
| 0                 | 32              |
| not specified     | 19              |
| Total             | 88              |

Table 19: Classification of CLOUPAC and CLoMP drugs into the 14 ATC classes.

| ATC first letter code | Group description                                               | CLOUPAC drugs | CLoMP drugs |
|-----------------------|-----------------------------------------------------------------|---------------|-------------|
| A                     | Alimentary tract and metabolism                                 | 39            | 0           |
| B                     | Blood and blood forming organs                                  | 10            | 0           |
| C                     | Cardiovascular system                                           | 68            | 3           |
| D                     | Dermatologicals                                                 | 22            | 0           |
| G                     | Genito urinary system and sex hormones                          | 19            | 0           |
| H                     | Systemic hormonal preparations, excl. sex hormones and insulins | 9             | 0           |
| J                     | Antiinfectives for systemic use                                 | 40            | 1           |
| L                     | Antineoplastics and immunomodulating agents                     | 49            | 5           |
| M                     | Musculo-skeletal system                                         | 24            | 1           |
| N                     | Nervous system                                                  | 77            | 6           |
| P                     | Antiparasitic products, insecticides and repellents             | 20            | 4           |
| R                     | Nervous system                                                  | 26            | 2           |
| S                     | Sensory organs                                                  | 25            | 1           |
| V                     | Various                                                         | 7             | 0           |
| Not classified        |                                                                 | 1122          | 67          |
| Total                 |                                                                 | 1490          | 88          |

**Table 20: Classification of CLoMP and CLOUPAC targets into 6 functional groups according to the HPA database.**

| Protein class              | Number of CLoMP targets | Number of CLOUPAC targets |
|----------------------------|-------------------------|---------------------------|
| CD marker                  | 3                       | 6                         |
| Enzyme                     | 15                      | 77                        |
| G-protein coupled receptor | 18                      | 63                        |
| Nuclear receptor           | 1                       | 14                        |
| Transporter                | 6                       | 30                        |
| Voltage-gated ion channel  | 2                       | 5                         |
| Not classified             | 86                      | 421                       |
| Total                      | 131                     | 616                       |

**Table 21: Classification of CLoMP and CLOUPAC targets into six locations according to the HPA database.**

| Protein class                     | Number of CLoMP targets | Number of CLOUPAC targets |
|-----------------------------------|-------------------------|---------------------------|
| Integral membrane (IM)            | 25                      | 106                       |
| Single pass-trans-membrane (SPTM) | 7                       | 16                        |
| IM/SPTM                           | 0                       | 1                         |
| Secreted                          | 0                       | 11                        |
| Membrane & secreted isoforms      | 1                       | 2                         |
| Intracellular                     | 13                      | 72                        |
| Not classified                    | 85                      | 408                       |
| Total                             | 131                     | 616                       |

Table 22: List of all 88 CLoMP drugs and their most effective concentrations (I).

| Drug name | concentration ( $\mu$ M) | Drug name | concentration ( $\mu$ M) |
|-----------|--------------------------|-----------|--------------------------|
| A0606     | 20.0                     | D9071     | 20.0                     |
| A2731     | 20.0                     | D9190     | 20.0                     |
| A6733     | 20.0                     | E5406     | 20.0                     |
| A9861     | 20.0                     | F114      | 20.0                     |
| B5002     | 20.0                     | F8927     | 20.0                     |
| B8385     | 20.0                     | F9677     | 20.0                     |
| CLOUD010  | 20.0                     | H5752     | 20.0                     |
| CLOUD013  | 20.0                     | H7779     | 20.0                     |
| CLOUD018  | 20.0                     | I2892     | 20.0                     |
| CLOUD036  | 20.0                     | K1015     | 20.0                     |
| CLOUD042  | 20.0                     | K2628     | 20.0                     |
| CLOUD047  | 20.0                     | K3375     | 20.0                     |
| CLOUD053  | 0.2                      | K4144     | 20.0                     |
| CLOUD066  | 0.2                      | L135      | 20.0                     |
| CLOUD069  | 20.0                     | L3169     | 20.0                     |
| CLOUD071  | 0.02                     | M149      | 20.0                     |
| CLOUD081  | 20.0                     | M5379     | 20.0                     |
| CLOUD091  | 20.0                     | M5441     | 20.0                     |
| CLOUD114  | 20.0                     | M5644     | 20.0                     |
| CLOUD163  | 20.0                     | M5685     | 20.0                     |
| CLOUD196  | 2.0                      | M9511     | 20.0                     |
| CLOUD198  | 20.0                     | N3510     | 20.0                     |
| CLOUD204  | 20.0                     | P203      | 20.0                     |
| CLOUD232  | 20.0                     | P3510     | 20.0                     |
| CLOUD234  | 20.0                     | P8139     | 0.2                      |
| D1414     | 20.0                     | P8688     | 20.0                     |
| D3900     | 20.0                     | R7150     | 20.0                     |
| D5794     | 20.0                     | S0568     | 20.0                     |
| D6140     | 20.0                     | S5890     | 20.0                     |
| D8696     | 20.0                     | S7395     | 20.0                     |

Table 23: **List of all 88 CLoMP drugs and their most effective concentrations (II).**

| Drug name | concentration ( $\mu$ M) | Drug name | concentration ( $\mu$ M) |
|-----------|--------------------------|-----------|--------------------------|
| SML0524   | 20.0                     | SML1506   | 2.0                      |
| SML0552   | 20.0                     | SML1547   | 20.0                     |
| SML0601   | 0.2                      | SML1639   | 20.0                     |
| SML0644   | 20.0                     | T104      | 20.0                     |
| SML0667   | 20.0                     | T1516     | 20.0                     |
| SML0683   | 20.0                     | T6950     | 20.0                     |
| SML0711   | 20.0                     | T7947     | 20.0                     |
| SML0777   | 20.0                     | T8703     | 20.0                     |
| SML0846   | 20.0                     | U105      | 20.0                     |
| SML1026   | 20.0                     | U111      | 20.0                     |
| SML1102   | 20.0                     | V4629     | 20.0                     |
| SML1108   | 20.0                     | Y0503     | 20.0                     |
| SML1138   | 20.0                     | Z101      | 20.0                     |
| SML1474   | 20.0                     |           |                          |

Table 24: **List of drugs within the nine distinct clusters extracted from the distance matrix.**

| Cluster | Drugs                                                                                                                                                                                                                                                                    |
|---------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1       | CLOUD204, F114, F8927, H5752, L135, M9511, P203, SML0644, SML1639, T104                                                                                                                                                                                                  |
| 2       | A0606, A6733, CLOUD047, CLOUD196, CLOUD234, D3900, S7395, SMI1108                                                                                                                                                                                                        |
| 3       | T1516, T8703                                                                                                                                                                                                                                                             |
| 4       | A9861, B8385, CLOUD036, CLOUD198, D8696, E5406, F9677, K1015, R7150, S0568, SMI1547, Z101                                                                                                                                                                                |
| 5       | CLOUD013, CLOUD018, CLOUD114, M5644, SML1026, SML1102, U111                                                                                                                                                                                                              |
| 6       | LOUD042, CLOUD091, D1414, D6140, I2892, K2628, M5379, SML0711, V4629                                                                                                                                                                                                     |
| 7       | A2731, B5002, B5681, CLOUD010, CLOUD069, CLOUD081, CLOUD163, CLOUD232, D5794, D9071, D9190, H7779, K3375, K4144, L3169, M149, M5441, M5685, N3510, P3510, P8688, S5890, SML0524, SML0552, SML0667, SML0683, SML0777, SML1138, M10846, SMI1474, T6950, T7947, U105, Y0503 |
| 8       | CLOUD071                                                                                                                                                                                                                                                                 |
| 9       | LOUD053, CLOUD066, P8139, SML0601, SML1506                                                                                                                                                                                                                               |

## List of Tables

|    |                                                                                                                                    |    |
|----|------------------------------------------------------------------------------------------------------------------------------------|----|
| 1  | Literature sources for an initial library compilation. . . . .                                                                     | 16 |
| 2  | Basic details about CLOUPAC drugs found in ChEMBL. . . . .                                                                         | 21 |
| 3  | List of filtering methods and applied thresholds. . . . .                                                                          | 35 |
| 4  | List of the three most efficient filtering subsets. . . . .                                                                        | 38 |
| 5  | Summary of CLoMP drugs. . . . .                                                                                                    | 47 |
| 6  | Basic details about CLOUPAC drugs found in ChEMBL. . . . .                                                                         | 49 |
| 7  | DMEM/F12 medium composition. . . . .                                                                                               | 62 |
| 8  | List of drugs appearing in both, CLOUD and LOPAC <sup>®1280</sup> library<br>and were kept twice. . . . .                          | 76 |
| 9  | Classification of CLOUPAC drugs into the four phases of clinical<br>research. . . . .                                              | 76 |
| 10 | Classification of the LOPAC <sup>®1280</sup> drugs into 94 different classes (I). . . . .                                          | 77 |
| 11 | Classification of the LOPAC <sup>®1280</sup> drugs into 94 different classes<br>(II). . . . .                                      | 78 |
| 12 | Classification of CLOUD, LOPAC <sup>®1280</sup> and CLOUPAC drugs into<br>the 14 ATC classes. . . . .                              | 79 |
| 13 | Classification of targets into 14 functional groups according to the<br>ChEMBL database. . . . .                                   | 80 |
| 14 | Classification of CLOUPAC and 'druggable proteome' targets ac-<br>cording to the HPA database into six functional groups. . . . .  | 80 |
| 15 | Classification of CLOUPAC and 'druggable proteome' targets ac-<br>cording to the HPA database into six cellular locations. . . . . | 81 |
| 16 | List of remaining CellProfiler features (I) . . . . .                                                                              | 82 |
| 17 | List of remaining CellProfiler features (II) . . . . .                                                                             | 83 |
| 18 | Classification of CLoMP drugs into the four phases of clinical<br>research. . . . .                                                | 84 |
| 19 | Classification of CLOUPAC and CLoMP drugs into the 14 ATC<br>classes. . . . .                                                      | 84 |
| 20 | Classification of CLoMP and CLOUPAC into six functional groups<br>according to the HPA database. . . . .                           | 85 |
| 21 | Classification of CLoMP and CLOUPAC targets into six locations<br>according to the HPA database. . . . .                           | 85 |
| 22 | List of all 88 CLoMP drugs and their most effective concentra-<br>tions (I). . . . .                                               | 86 |
| 23 | List of all 88 CLoMP drugs and their most effective concentra-<br>tions (II). . . . .                                              | 87 |
| 24 | List of drugs within the nine distinct clusters extracted from the<br>distance matrix. . . . .                                     | 87 |

## List of Figures

|    |                                                                                                            |    |
|----|------------------------------------------------------------------------------------------------------------|----|
| 1  | Pipeline for compiling and characterizing the novel morphology-based compound library CLoMP. . . . .       | 14 |
| 2  | Distribution of CLOUPAC drugs into clinical trial phases. . . . .                                          | 20 |
| 3  | Comparison of CLOUD, LOPAC <sup>®1280</sup> and CLOUPAC ATC classifications . . . . .                      | 22 |
| 4  | Distribution of the number of targets per CLOUPAC drug. . . . .                                            | 25 |
| 5  | Proportional comparison between CLOUPAC and ChEMBL target classifications. . . . .                         | 26 |
| 6  | Proportional comparison between CLOUPAC and HPA target functions. . . . .                                  | 28 |
| 7  | Proportional comparison between CLOUPAC and HPA target locations. . . . .                                  | 28 |
| 8  | Immunofluorescent staining of MCF10A cells. . . . .                                                        | 30 |
| 9  | Examples for the correlation of replicates for four different features. . . . .                            | 34 |
| 10 | Boxplots showing cosine similarities of drug vectors of different drug subgroups. . . . .                  | 37 |
| 11 | Distribution of cosine distances between replicate pairs. . . . .                                          | 40 |
| 12 | Distribution of DMSO and drug vector lengths. . . . .                                                      | 41 |
| 13 | Relationship between the replicate correlation and the strength of the drug effect. . . . .                | 42 |
| 14 | Cell count distributions of drug-treated samples and negative controls of batch 1. . . . .                 | 44 |
| 15 | Cell count distributions of drug-treated samples and negative controls of batch 2. . . . .                 | 44 |
| 16 | Examples for drug-induced phenotypes in MCF10A cells. . . . .                                              | 46 |
| 17 | Distribution of CLoMP drugs into clinical trial phases . . . . .                                           | 48 |
| 18 | Comparison of CLOUD, LOPAC <sup>®1280</sup> and CLOUPAC ATC classifications . . . . .                      | 50 |
| 19 | Proportional comparison between target groups of the CLoMP, the CLOUPAC and the ChEMBL database. . . . .   | 52 |
| 20 | Proportional comparison between CLoMP and CLOUPAC target functions. . . . .                                | 53 |
| 21 | Proportional comparison between CLOUPAC and HPA target locations. . . . .                                  | 53 |
| 22 | Distance matrix based on pairwise correlations of concentration values between all pairs of drugs. . . . . | 55 |
| 23 | Overlaid curves of the different clusters returned by the distance matrix. . . . .                         | 56 |