1 Summary

The pharmaceutical industry is centered around small molecules and their effects. Apart from the curative effect, the absence of adverse or toxicological effects is cardinal. However, toxicity is at least as illusive as it is important. A simple definition is: ’toxicology is the science of adverse effects of chemicals on living organisms’.[1] However, this definition comprises several caveats. What is the organism? Where does therapeutic and adverse effects start and end? Even for the toxicity in simplest organisms, cytotoxicity, the mechanisms are mannifold and difficult to unravel. Hence, it remains obscure which characteristics a compound has to combine to be labelled as toxic.

One attempt to illuminate these characteristics are the novel cell-painting (CP) assays. The CP data’s roots lie in cellular fluorescence microscopy. Five fluorescent channels have been used for imaging and these channels correspond to certain cell organelles.[2] Before the images are taken and further processed, the cells are perturbed by small compounds that might affect the cellular morphology. Therefore CP contains information about each compound formatted as a unique cell-structure anomaly. Which sub-information are actually valuable within these morphological fingerprints remains elusive and therefore a significant part of this project is dedicated to exploring the CP data and their predictive capabilities comparatively. They will be compared among bioassays and to different descriptors, too. The CP data used by this project contains roughly 30 000 compounds and 1800 features.[3]

In chemistry, the structure determines the function of a compound or substance. Therefore, apart from CP, structural fingerprints are used as a benchmark to compare the information content of CP against. In this project extended-connectivity fingerprints (ECFPs) were used to encode the compounds’ structures into numerical features.

This work is concerned with morphological changes that correspond to toxicity. Thus, the CP data were combined with toxicological endpoints from a selection of PubChem assays. The selection process implemented a minimum number of active compounds, a size criterion and the occurence of toxicologically relevant targets.[4]

After the selected assays were combined with each of their descriptors, machine learning models were trained and their predicitive power was evaluated against certain metrics. The pre-

dictions can be separated into 3 cycles. In the first cycle, the CP data are used as descriptors, the second cycle used the structural fingerprints and the last cycle used a subset of both. The subsets were selected by a rigorous feature engineering process.

The evaluation of the prediction metrics illuminates which strengths and shortcomings the morphological fingerprints have compared to the structural fingerprints. It turned out that there are two groups of assays: those PubChem assays that are generally better predicted with CP features and those that have higher predictive potential when using ECFP. Additionally, it was uncovered that ECFP comprise higher specificity compared to CP data which shows higher sensitivity on the other hand. A high sensitivity means the prediction rarely mislabels a sample as negative (e.g. non-toxic) compared to the number of correctly labelled positive samples (e.g. toxic compounds.). Based on these results, CP is better suited for toxicity prediction and drug safety since the mislabelled, positive compound results in expenses or even damage to health. Furthermore, based on the fluorescent channels an enrichment measure was introduced that is calculated for the aforementioned two groups of PubChem assays. This enrichment metric indicates unusual cell organell activity. The hypothesis was that PubChem assays well predictable from CP data should have increased enrichment, which was the case for four out of five fluorescence microscopy channels.

As a final step phenotypic terms were manually generated for categorization of the different PubChem assays. These terms correspond to cellular mechanisms or morphological processes. The phenotypic terms were generated unbiasedly but are subject to imperfect knowledge and human error. A bioassay may or may not be associated with a phenotypic annotation. The phenotypic annotations that are found to be enriched for better performing PubChem assays might be able to guide the pre-selection of bioassays in future projects with CP descriptors. The enrichment analysis of phenotypic annotations detected that PubChem assays related to immune response, genotoxicity and genome regulation and cell death are best characterized by CP data.

2 Introduction

Currently, pharmacological drug development focuses on well-established biochemistry based approaches to find and optimize new drugs. However, the challenges these methods face are manifold. High costs related drug failure rates during various clinical trials and commercialization bottleneck the industry as a whole. Another important aspect is the occurrence of adverse drug reactions subjecting patients to hospitalization possibly ending up fatal. Therefore, the pharmaceutical industry is not only facing high financial risks but also humanitarian problems, that strains the trust-based relationship between the industry, physicians and patients.[5] Academia demonstrated computational tools to be employable to many challenges of the health industry like costs of drug target validation, drug safety and commercialization.[6] [7] Albeit, chemo- and bioinformatics are novel and complex disciplines recently fostered by technological advancements in high-throughput methods. Thus, ~~the~~ health industry does not yet benefit from promises like computer-aided identification of drugs and drug targets on a large scale.

New high-throughput methods and automated microscopy gave rise to the development of high-content imaging which is frequently combined with small compound perturbations inflicted on biological systems. High-content-imaging applies up to six fluorescent dyes allowing to portray up to five different compartments per cell. This method, also referred to as CP can be used to screen a wide range of compounds and capture the induced morphological changes.[8] CP assays capture compound perturbed biological systems and automatically resolve cellular characteristics that can be interpreted by computational models in context of a morphological fingerprint (or CP feature vectors).[3] The raw images from high-content imaging are processed, mostly by the software CellProfiler[2] which extracts up to 1800 numerical features per image (e.g.nucleus shape, endoplasmatic reticulum (ER) texture). The features from CP assays can be interpreted as a morphological fingerprint, unique for each compound.[8] By now, many different CP assays have been conducted and specific strategies have been developed for specific purposes. A widely used CP data set was generated by Bray *et al.*[9] The images were recorded with sixfold fluorescence staining for imaging five crucial cellular organelles (further information see section 4.4).[3] The concept of CP is visualized in figure 2.1.

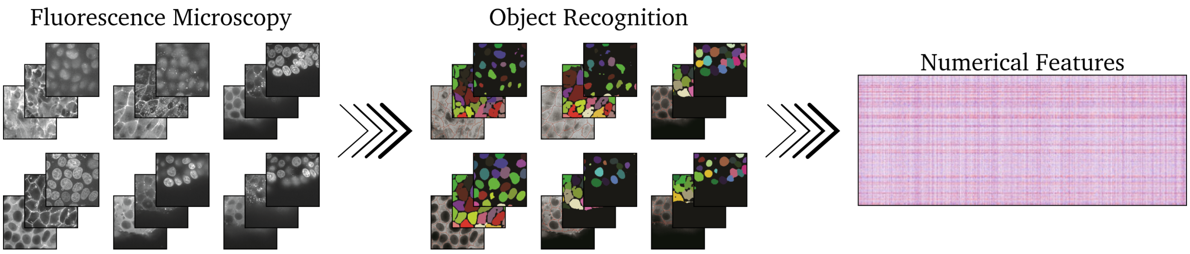

Figure 2.1: Visualization of a CP Assay. First cellular images are generated by fluorescence microscopy, then cellular objects and compartments are recognized by CellProfiler from which a large data table is generated containing the morphological fingerprint for each compound. Figure from Rohban *et al.* and Carpenter *et al.*, modified.[2,10]

Recently, Gustafsdottir *et al.*[11] used CP data to link morphological states to mechanisms of action via gene expression data. Hierarchical clustering was used to find clusters of compounds, addressing the same set of genes and therefore the same biochemical pathway. The results obtained from this CP based approach mirrored the findings in literature, which showed that CP data is directly correlated to cellular pathways responding to compound perturbations. Nassiri and McCall[12] used different CP assay data[13] and compound perturbed gene expression data in the context of machine learning methods. They used a LASSO model to predict cell morphological features against similar gene expression profiles. In-depth analysis of the results revealed strong model predictiveness among compounds that steer gene expression in the same direction, suggesting common mechanisms of action. Hence, not only can CP data be linked to mechanisms of action but also an in-depth analysis reveals a relation between compounds' mechanisms of action based on machine learning model performance. Furthermore, Rohban *et al.*[10] transduced U-2 OS cells with lentiviral particles carrying cDNA constructs for gene overexpression.[14,15] In their approach, they conducted a CP assay to annotate the overexpressed genes with morphological fingerprints (numerical CP features). After calculating the Pearson correlation between each morphological fingerprint they used hierarchical clustering resulting in 25 clusters for 110 overexpressed genes. The clusters generated from the CP data clustered genes that correspond to similar or identical pathways showing that genes can be connected using relatively inexpensive CP assays. Furthermore they predicted an unknown relationship between the Hippo- and NF-κB-pathway. Lapins and Spjuth[16] annotated compounds of the CP data from Bray *et al.*[3] and from the CMap[17] (gene expression profiles) with their mechanism of action (MoA) or target protein. The information about the compound-wise MoA and targets were obtained from the Drug Repurposing Hub or the Touchstone data base. In total they annotated 1484 compounds present in CP and CMap data with 234 MoAs or targets. As a third set of descriptors structural fingerprints were generated. For several targets and MoAs a trained random forest classifier (RFC) could present significant discriminatory power

(AUC$>$ 0.7). Furthermore, it was found that the 3 different descriptors were complementary to a certain degree, each excelling at different MoAs or targets. Lapins and Spjuth not only showed that CP data could be used to predict compounds MoA but also, that a combination with other identifiers is likely to enhance their applicability domain.

Simm *et al.*[8] studied a CP assay specifically designed for glucocorticoid receptor (GCR) nuclear translocation. After treating H4 brain neuroglioma cells with 524 371 compounds, hydrocortisone is added to stimulate GCR translocation. Next, the treated cells were stained, imaged and processed analogous to the work of Gustafsdottir *et al.*[11] The 524 371 compounds were not only annotated with morphological information, but also with target activity information from 600 biochemical assays. Noticeably, most compounds were covered in few assays only, amounting in a fill rate of 1.6 %. From this sparse activity matrix they built a machine learning (ML) model with CP data as side information to predict all labels within the activity matrix. They evaluated the discriminatory power of their model and 34 bioassays (out of 600) showed high predictivity (AUC>0.9). One of the assays was part of an ongoing discovery project. Within this assay the highest ranking 342 compounds, by matrix factorization, were experimentally tested. 141 (41.2%) of these resulted in submicromolar hits which means a 60-fold enrichment over the initial high-throughput-screening (HTS). Another assay with AUC greater 0.9 was part of an ongoing drug discovery project and could achieve a 250-fold hit enrichment over the initial HTS in an analogous way. Their work presented CP data as highly informative descriptors, that might be repurposed for prediction of sparse activity matrices. Additionally, they demonstrated their potential in ongoing drug discovery projects.[18]

Data science in computational biochemistry has great potential, however, there are caveats that need to be aware of when working with CP and drug safety. The first one is imbalanced data. An assay testing compounds on a potential target will always feature less actives than inactives. Chawla *et al.*[19] propsed a technique that mitigates this effect called synthetic minority oversampling technique (SMOTE). This technique allows to generate synthetic samples that fit into the distribution of the real data points. Another problem is the high dimensionality of CP data ~~which is an intrinsic problem connected to~~ its richness in information. Only a comparably small number of features contains most of the information necessary to predict a given target, which is why feature engineering is usually conducted in CP studies. A CP data set that contains too many features is bound to overfit the data and give overoptimistic predictions on the test set without generalizing particularly well. The aforementioned project of Rohban *et al.*[10] tackled this problem by ~~conducting a~~ principal component analysis (PCA). They reduced the number of features from 2769 to 158 that comprise most of the variance.

In this explorative project, the CP data set of Bray *et al.*[3] is used for the prediction of PubChem assays. The PubChem assays are selected based on their relation to targets presumably contributing to cytotoxicity.[4] The results are compared against structural fingerprints of the small molecules and the performance metrics are analyzed extensively. From this analysis, conclusions can be drawn, whether and when to use CP data. Whether structural fingerprints add complementary information and which cellular processes are corresponding to excelling predicitve performance of CP descriptors. Insights gained from this project might be able to guide future decision making when it comes to prediction of biological endpoints.

3 Scientific Aim

The goal of this project is to generate heuristics that simplify working with CP data. Generally, CP can be used to predict ~~compound wise~~ biochemical readouts. However, which types of readouts work well and why remains elusive. Therefore, this work aims at understanding the results obtained from RFC prediction and to link the results to cellular mechanisms and the concept of cytotoxicity.

Conceptually, this means finding bioassays whose endpoints are related to toxicity for annotation of the CP descriptors. Since this is a comparative approach, structural fingerprints (ECFPs) are used as another set of descriptors. Both annotated data sets are inputted into a ML model and the discriminatory power of this model is evaluated using generic statistical metrics. Another model is trained using the combined set of descriptors which is evaluated analogously.

The detailed procedure starts by preprocessing of the CP raw image data into a machine learning ready data frame. Next, the bioassays are selected from PubChem,[20] based on size and their relation to cytotoxicity. Every bioassay data frame is combined with the CP data frame to obtain annotated sets containing inputs as well as targets for prediction. As a comparison, structural descriptors are generated for all data sets using sklearn funtionalities.[21] To this end, the annotations in all data sets are highly imbalanced, comprising mostly samples labelled as inactive. ~~Herein,~~ this problem is tackled by applying SMOTE to the data sets in combination with undersampling of the majority label (inactive). Two RFC models are trained on each data set and their predictive power is evaluated. Furthermore, to examine if shortcomings of one descriptor can be mitigated by the other, both descriptors are combined and another model is trained and evaluated. Since there is no apparent way to select features manually, statistical methods can be used to meaningfully conduct features selection. PCA, minimal-redundancy-maximal-relevance criterion (MRMR) and random forest feature importance are applied to score and select features from each set of descriptors for merging. Eventually, another RFC model is trained and evaluated using the joint descriptors. Based on the prediction evaluation and feature selection process a rigorous analysis is conducted aiming to explain the results with respect to cellular morphology and cytotoxicity. The focus lies in detecting and analyzing prediction similarities and dissimilaries between either descriptor sets, bioassays or groups characterized by their performance. Thereby, patterns can be detected that transform into heuristics which facilitate the application of CP data to ML problems.