Predicting Mechanisms of Toxicity for Drug Development

A Semi-Supervised Machine Learning Approach with Metabolomic Data

Master's Thesis in Computer Science and Engineering

SÖREN RICHARD STAHLSCHMIDT

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2019

Master's thesis 2019

Predicting Mechanisms of Toxicity for Drug Development

A Semi-Supervised Machine Learning Approach with Metabolomic Data

SÖREN RICHARD STAHLSCHMIDT

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2019

Predicting Mechanisms of Toxicity for Drug Development
A Semi-Supervised Machine Learning Approach with Metabolomic Data
SÖREN RICHARD STAHLSCHMIDT

Supervisor: Alexander Schliep, PhD, CSE
Advisor: Maria Luisa Guerriero, PhD, AstraZeneca
Examiner: Devdatt Dubhashi, PhD, CSE

Master's Thesis 2019
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2019

Predicting Mechanisms of Toxicity for Drug Development
A Semi-Supervised Machine Learning Approach with Metabolomic Data
SÖREN RICHARD STAHLSCHMIDT
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

The aim of this thesis is to predict different mechanisms of toxicity from the metabolomic response of HepG2 liver cells. In order to utilize the metabolomic data the a semi-supervised machine learning approach is investigated, namely the cluster-then-label approach. The research focuses on the unsupervised part due to the centrality to this method. The dose-dependency within the data is modelled by clustering the dose-response curves according to their shape and transforming the feature space to a categorical one. This dataset is then clustered with the K-Modes algorithm. The analysis of the experimental data has shown that it is possible to distinguish toxic from non-toxic compounds on individual dose level though mechanisms can not clearly be distinguished. The proposed method is not able to clearly distinguish between toxic and non-toxic compounds or between the mechanisms of toxicity. It is hypothesized that the lack of mutually exclusive labels makes the prediction harder. Furthermore, the model could benefit from a more fine-grained dose levels in the identified range.

Keywords: Predictive Toxicology, Metabolomics, Semi-Supervised Learning, Dose-Response

Acknowledgements

I would like to extend my gratitude to my academic supervisor Dr.Alexander Schliep as well as my supervisor at AstraZeneca Dr.Maria Luisa Guerriero for their advice and suggestions throughout the project. Furthermore, I would like to thank my line manager Peter Konings for enabling an easy stay at AstraZeneca in Gothenburg. For an introduction to the biological background and the experiment generating the data I would like to thank Dr.Delyan Ivanov. Also I would like to thank Dr.Natalie Kurbatova for fruitful discussions regarding the project and Dr.Oscar Hammar for review of the probability theoretical claims. Finally, a big thank you to all colleagues from the Quantitative Biology group at AstraZeneca for making these month very enjoyable.

Sören Richard Stahlschmidt, Gothenburg, June 2019

Contents

List of Figures

List of Tables

1

Introduction

1.1 Scientific and Societal Relevance

Medicine is developed in order to improve patients' lives. Therefore, the safety of drugs is of utmost importance so that they themselves do not become a hazard to the patient's health. Knowing such effects before drugs move into the clinical trial pipeline can therefore reduce harm to animals as well as humans. From a company perspective, the attrition rate of drug development process throughout the clinical trial phases is high and costly. For instance, between 2006 and 2015 only 9.6 percent of drug development programs that were submitted to phase I of clinical trials were ultimately approved by the FDA in the United States [1]. For the therapeutic area of oncology this number drops even further to 5.1 percent. These high attrition rates have two effects. Firstly, the costs of successful drugs need to carry those costs produced by failed ones [2]. Ultimately these costs need to be transferred to the patient, the healthcare system and society at large. Thus, reducing the attrition rates in clinical trials translates into more accessible medicine for patients. Secondly, resources have been invested into every failed drug development program that could otherwise have been allocated to more promising candidate compounds. Hence, the time it takes for drugs to reach the patient is significantly increased resulting in harm to the patient. Increasing the likelihood of approval therefore is a promising way of increasing patients' health.

Beside drug efficacy, drug safety constitutes one of the major reasons why, especially in phase II of clinical trials, drug development programs fail. Toxicity of molecules is a hazard to the patient and leads to the failing of a drug candidate [2]. Chemical toxicity is the ability of a molecule to damage cells of humans, animals and other living organisms. However, toxicity can be highly idosyncratic, meaning that a drug can be highly toxic in some individuals but not in others [3]. This makes its prediction inherently difficult. Nevertheless, predicting before a clinical trial which molecules or compounds are toxic to patients bears the potential of improving the drug development process and reducing harm to trial participants. Particularly liver toxicity is damaging for patients. Methods to minimize the risk for such hazards are needed [4, 5].

1.2 Problem Definition

Traditionally, cytotoxicity has been used to measure the toxic potential of compounds. Cytotoxicity includes any form of cell death as a measurement of toxicity. However, a new approach to toxicity is to measure metabolites in order to infer toxic potential. So-called metabolomics is able to detect toxicity earlier since it becomes apparent through metabolites even before widespread cell death. Utilizing methods from artificial intelligence and machine learning to predict different liver toxicity of molecules and compounds have shown some initial promising results to detect liver toxicity with metabolomic datasets [6].

Furthermore, compounds can initiate different mechanisms of toxic actions (MeOA). The ways a compound can damage a cell can be vastly different. Understanding which MeOA is caused provides a more detailed picture of the effect on the cell and thus can aid a better drug discovery process.

Classifying compounds according to the MeOA they cause comes with challenges. The effect of a compound is highly dose dependent [6]. That means that low enough doses, even of highly toxic compounds, are non-toxic. On the other hand, compounds that at normal doses are not toxic at all, such as water, become toxic at high doses. Thus, one must take into account the effect of the compound at different doses. Along these dose levels the same reaction by the cell can occur earlier or later. Though they are shifted they constitute a similar effect.

Another challenge is that a compound can cause several MeOAs. Therefore, the compounds could cluster according to combinations of mechanisms. Additionally, labelling for compounds that are known to be toxic is sparse since identifying these is from a biological perspective a challenge in itself.

1.3 Aim and Research Questions

The aim of the research project is to investigate the potential of metabolomic data to predict mechanisms of toxic actions with machine learning methods.

More specifically, two research questions are investigated:

1. Do chemical compounds cluster in metabolomic feature space according to their mechanisms of toxic actions?
2. If such clusters exist, can a supervised machine learning algorithm be trained to predict mechanisms of unseen compounds from the metabolites they activate?

1.4 Limitations and Ethical Considerations

A limitation of the research is the gap between in vitro experiments and effects on patients. Results from in vitro experiments are not directly translatable to live organisms. For instance, the exposure of the cell to the compound is not equivalent to the dose given due to the uptake of the drug. Thus, knowledge about dose-response from in vitro experiments does not necessarily hold in vivo. Solutions to bridging the gap are based on pharmacokinetic models which estimate exposures from in vitro experiments to in vivo effects. This in vitro to in vivo extrapolation (IVIVE) however is not a trivial task [7]. This modelling is outside of the scope of this thesis.

Additionally, from a biological perspective this thesis is focused on the acute rather than the chronic exposure to the chemical compound. This differentiation is important since the same compound administered in one high dose or chronically can activate different mechanisms of toxicity [8]. Focusing on the acute exposure helps to define a clear aim for the research project.

In order to focus the research, it is limited to in vitro modelling of drug induced liver injury (DILI). Excluding injury in other organs keeps the project manageable since only data from liver cell lines are generated and analyzed. Liver toxicity is a major challenge to drug safety and thus constitutes a good use case. The liver is the organ that is metabolizing most toxins and thus has the highest chance of being adversely affected.

The measurements during the experiment are performed 24 hours after the initial exposure to the chemical compound. This gives the cell time to react to the exposure. However, observing the effects at different points in time holds the potential to gain deeper insights into the MeOA. This is suggested as a future research that builds upon this thesis.

Especially, in the medical domain high scrutiny of ethical implications of research should be made. The high impact the findings, but also the research practice itself, has on patients' lives demands high ethical standards. Therefore, the patients are the primary stakeholder from an ethical perspective [9]. The results from this research can have an impact on patients as for instance clinical trial participants. Thus, working thoroughly is of great importance. On the other hand, the research done for this thesis concerns non-patient data due to its pre-clinical trial nature. From this follows that no privacy issues with regards to patients can arise.

1.5 Related Work

Utilizing metabolic responses to predict toxicity of chemical compounds has recently received much attention in the academic literature. The main reasons are the benefits predictive models based on metabolomics offer over the traditional testing. Using metabolites as biomarkers for toxicity has the advantage that these molecules

are much earlier apparent for pathologies than widespread cell death. Additionally, metabolomic data can give insights to the mechanisms underlying toxicity (for instance [10, 11, 12] for in vivo experiments). This can be beneficial for instance to group compounds according to these and do animal testing only on representative compounds[13].

As a proof-of-principle, [13] investigated the potential of metabolomic data to predict the effect of three chemicals on the reproductive fitness of individual Daphnia, a plankton species. By applying a multiple regression to the metabolic response data of the test subjects, they demonstrated that it was possible to predict the toxic effect. Though the authors note that the findings are limited to these three chemicals tested, they showed the potential of the approach.

[6] identify different mode of toxic actions with 35 test substances by comparing the metabolomic profiles of reference substances that are known to exhibit a certain adverse outcome pathway. Using these reference compounds that can be considered labeled data to identify metabolites that are related to the different modes of actions it was possible to separate the test compounds into groups according to their modes. Again, the authors stress that their study functioned as a proof-of-concept for the utilization of metabolomic data for the prediction of toxicity and their underlying mechanics.

Building upon this research, this thesis aims to utilize the potential metabolomic data has for the prediction of mechanisms of liver toxicity.

2
Theory

Machine learning is not an end in itself and is thus always applied to a specific domain. This chapter provides the background knowledge to the domain of toxicology as well as the machine learning approach chosen in this research. This lays the foundation for the approaches described in the methods chapter.

2.1 Biochemistry of Toxicity

In order to understand the target variable, the feature space and the connection between the two, it is essential to delineate the biochemical concepts and processes forming these. Hence, in the following what it means for a compound to be toxic is explained. Additionally, the concept of metabolites is clarified. In order to define the classes for the machine learning task the MeOAs for liver toxicity are described. Finally, the connection between these mechanisms and specific metabolites are outlined.

2.1.1 Toxicity and Metabolites

Toxicity describes how much a chemical compound has adverse effects on an organism. These effects can occur on different levels, such as cell level (cytotoxicity) and organ level. In this research for instance drug induced liver injury (DILI) is measured as hepatotoxicity. When for instance foreign chemicals are metabolized the resulting biochemical consequences can cause different forms of stress on the cells which ultimately can lead to cell death. Alternatively, the adaptive responses can be activated to reduce damage inflicted upon the cell. Toxicity of a compound is dose and time dependent, meaning that the exposure of the cell must be sufficiently high and present for sufficient amount of time to result in cell death or adaptive behavior. Examples of possible cell stress pathways are oxidative stress, stress on the endoplasmic reticulum (ER) and mitochondria [5].

Oxidative stress is an imbalance between the reactive oxygen species and antioxidant defenses of the cell. While oxygen facilitates energy production in the cell it also can lead to damage or even death of the cell [14]. Thus, the imbalance induced by exogenous chemicals can be harmful. ER is an organelle involved in folding and modifying proteins in the cell. The loss of this function plays a role in a variety of diseases such as cancer, Alzheimer´s disease and bipolar disorder. Thus, the disruption of the function through compounds can lead to damages of the cell [15].

Mitochondrial stress is the disruption of the function of the mitochondria which is associated with different disease such as schizophrenia, Alzheimer's disease and hepatitis C. Oxidative stress is one cause of such disruption. However, there exist other causes such as metabolic dysregulation through for instance malnutrition that not necessarily must be drug-induced [16].

Metabolites are small molecules produced in the process of maintaining life in an organism. For instance, to provide the cell with energy the mitochondria transform adenosine diphosphate (ADP) to adenosine triphosphate (ATP). Both these molecules are metabolites of this process. Since many different processes take place within living organisms the number of metabolites is high. This is for instance apparent from the high dimensionality of the data set of this project. Measuring metabolites thus enables inferences about the processes taking place in the organism [17]. Further, measuring the concentration of metabolites gives a representation of cell health [18]. Therefore the quantitative measurement of the metabolic reaction of cells, termed metabolomics [12], to the exposure to foreign chemicals such as candidate drugs can provide insights to the toxic potential.

2.1.2 Mechanisms of Toxicity

As previously mentioned, compounds can cause cell and functional damage in different ways. The causal chains that result in these damages are termed adverse outcome pathways (AOP). These require a specified molecular initiation which is linked through different biological levels of organization to an adverse effect in the population. This means that the chain is not limited to the cell but spans further over subcellular level, cell, organ, organism and finally population. A MeOA is the complete understanding of such a chain whereas the MoOA can be considered an AOP with incomplete information [19]. The concept of AOP was introduced to clarify the uncertainty about terminology in the field of toxicology. For the purpose of this research the focus is laid upon the MeOA as a AOP with complete information.

For liver toxicity it is possible to distinguish five, not mutually exclusive mechanisms leading to DILI. In the following these mechanisms are described in detail since they constitute the classes for the classification task. Due to the in vitro nature of the experiment the immune-system-based mechanism cannot be simulated with the cell cultures that represent the model of the liver. The mechanism is nonetheless briefly described for reasons of completeness.

Toxicity on the mitochondria (mechanism 1) is the disruption of the mitochondria's function by a foreign chemical compound. This disruption can result in different outcomes such as the alteration of the mitochondrial homoeostasis. Such changing of the natural balance of the mitochondria can result for instance in oxidative stress, energy depletion and even cell death. The devastating effect this can have for patients can be observed with the antiviral therapy fialuridine which was a potential

treatment for hepatitis B. In a phase II study, 7 out of 15 patients developed acute liver failure of which 5 died and the remaining two survived through liver transplant. It was shown that fialuridine was effecting the mitochondria of HepG2 liver cells [8].

Compounds can also be toxic by forming chemically reactive metabolites (CRM) (mechanism 2). During drug absorption the parent drug can be metabolized to a hydrophilic, i.e. water-soluble, entity. This process can also result in CRMs. These metabolites constitute the molecular initiation of the AOP. The metabolites react with macromolecules such as proteins, thereby disrupting the normal function of the cell. Whereas some CRMs react locally inside the cell others can diffuse in surrounding cells and cause damage in a larger area. CRMs are one of mechanism that make acetaminophen toxic. This chemical is involved in many acute liver failures since at therapeutic doses the drug is safe but at higher doses exhibits toxic behaviors [8, 20].

Lysosomal impairment (mechanism 3) is another mechanism by which compounds can be toxic. The lysosomes are organelles in the cell that break down different extra- and intracellular molecules. The lysosomes are involved in autophagy which is the process of breaking down intracellular molecules. The disruption of this process can lead to diseases such as cancer, neurodegeneration and diabetes [21]. Moreover, the lysosome is involved in endocytosis which is the breakdown of extracellular molecules such as drugs. Certain drugs can accumulate in the lysosome. The accumulation of fats and the drug results in the imbalance of charges in parts of the cells which ultimately causes dysfunction of the lysosome. The retention of fats such as phospholipids (steatosis) [22] in the liver can cause the formation of excess tissue (fibrosis) which quickly leads to the disruption of normal liver function through excessive amounts of scar tissue (cirrhosis) [8]. In the labelling for the dataset in this project the mechanism is focused on the damaging through phospholipids.

A fourth MeOA is the accumulation of bile acid in the liver termed cholestasis (mechanism 4). Bile is a fluid that is produced by the liver and assists the small intestines with the digestion of, for instance, fats. Therefore, bile acid is moved by different transporters from liver cells to the small intestines. Drugs that cause cholestasis affect these transporters which inhibits the secretion of bile. For instance, drugs can inhibit the bile salt export pump (BSEP) (Padda, Sanchez, Akhtar, Boyer, 2011). This results in the bile acid in the liver not flowing into the small intestine. The accumulation of acid damages the liver. Bosentan is an example of a chemical where such BSEP inhibition can be observed [23, 8].

Finally, drugs can cause toxicity through immune mechanisms (mechanism 5). Immune responses to compounds and compound-induced autoimmune reactions can cause acute liver injury (Dragovic et al., 2016). This mechanism cannot be seen in the current in vitro setup as hepatocellular cell lines are cultured as monocultures in the absence of immune cells. Thus, this mechanism is not detectable in the dataset.

The four first mechanisms constitute the target variables. Classifying which mechanisms are initiated by a compound can be of importance determining the safety of

drug candidates before they are tested in animals and patients. Thereby harm to animals can be reduced and patient safety increased. As these pathways are taking place in the cells, different metabolites can be observed. Reversely, the mechanism can be inferred from observing metabolites.

2.1.3 Mechanisms and Metabolites

Though the connection between different biomarkers and specific mechanisms of DILI is not fully explored, the association between some metabolites and specific mechanisms have been shown [24, 23]. Thus, this leads to these metabolites informing the annotation after clustering and at the same time justifies why other metabolites are kept in the dataset. The metabolites that are not known to be associated with any mechanism are used as the space for exploratory biomarker development as suggested by [4]. Associations found through this exploratory method can in future research be investigated further by subject matter experts on a case by case bases. In the following the known association between specific metabolites and the four mechanisms outlined above is described.

For toxicity on the mitochondria (mechanism 1) the change in ATP, ADP and adenosine monophosphate (AMP) should be possible to observe. These metabolites are essential for energy production. When foreign compounds damage the mitochondria, the energy-carrying metabolite ATP should be depleted. There should be an increase of ADP and AMP since they are not transformed into ATP. This follows logically from the function of mitochondria in the cell. Thus, observing such an increase can indicate that a compound is initiating mechanism 1.

Detoxification done in the liver generally happens in three phases. Phase I and II metabolize the foreign compound through enzymes whereas phase III is the transport and elimination from the body. However, in phase I the compound can be metabolized to more reactive and toxic metabolites which are detoxified in phase II. A foreign compound can be metabolized to for instance N-acetyl-p-benzoquinone imine (NAPQI) in the case of APAP which then is detoxified by glutathione (GSH). If an overdose of APAP enters the cell GSH is depleted and NAPQI causes damage to the cell. Since GSH has this function for many toxins it can be considered a biomarker for CRM (Mechanism 2). The depletion of GSH can activate the increase in the metabolite ophthalmic acid (OA). Thus, OA can be considered another metabolite indicating the activation of mechanism 2. 5-Oxoproline (5-OP) is an intermediate metabolite in the GSH biosynthesis and is more directly linked to GSH and can be used as an indication for CRM. There is some indication that the amino acids cysteine, glutamine and glycine who are precursors of GSH are elevated through hepatotoxic compounds. At least in plasma and urine elevated levels could be shown [23].

The third mechanism which causes lysosomal impairment (mechanism 3) is associated with an increase in lipids which are free fatty acids, phospholipids and lipid metabolites. As described above, fatty acids accumulate in the liver and

form molecules with the foreign compounds. This increase should be visible in the metabolomic profile of the compound [8]. Furthermore, the accumulation of lipids would result in increased utilization of fatty acids to produce energy in the cells in a process called —oxidation leading to the formation of 2-hydroxy and 2-keto fatty acid derivatives and acyl-carnitine metabolites.

Cholestatic injury concerns the transport of bile acid from the liver. Due to the inhibition of the secretion of bile acids from the liver increased levels thereof are to be expected if a chemical compound is initiating mechanism 4 [23]. Though some of the bile acids are produced in the gut and are not modelled with current in vitro cell cultures, many bile acids are produced by liver cells.

2.2 Semi-Supervised Learning

In the realm of machine learning there exists the distinction between supervised and unsupervised learning. In order to make classifications, supervised approaches learn from fully labelled datasets. These datasets are required to represent the underlying structure of the true distribution of the target variable in order to enable sufficiently good classifications. However, a major challenge is that the annotation of data usually has to be done by human expert annotators or specialized machines. In the context of this project that means that subject matter experts have to review the current toxicology literature to annotate chemical compounds with the known mechanisms of toxicity. These procedures take significant amount of time, are costly and are generally incomplete. Therefore, sufficiently large labelled datasets are rare. Unsupervised learning however finds patterns in unlabeled data which is more readily available. Usually this involves assigning similar observations into the same group. Such clustering does not require human input. The main drawback is that the clusters do not straight forward explicate to what category they correspond [25].

Semi-supervised learning (SSL) utilizes a labelled dataset in combination with an unlabeled dataset to develop better classifiers. Thus, SSL refers to semi-supervised classification in contrast to semi-supervised clustering since the final goal is the assigning of a category to an instance. Furthermore, one has to distinguish between transductive and inductive learners. The former is the learning that cannot handle unseen data. The latter in contrast is the approach pursued in this thesis. The aim of the learning is the training of a model that can correctly classify unseen instances. The major drawback of SSL approaches is that assumptions about the problem structure, as for instance the assumption that the decision boundru should not g through dense areas, are made by the different methods. If these assumptions are actually violated the methods perform poorly [26].

A rationale for including unlabeled observations is provided by [27]. As a simple example, say data is generated from a mixture model consisting of n normally distributed components. This model can be expressed as

$$f(x|\Theta) = \sum_{j=1}^{n} \alpha_j f(x|\theta_j). \tag{2.1}$$

Where α_i represents the coefficient for the respective component, functioning as a weighting of the different components. Thus, $\sum_{j=1}^{n} \alpha_j = 1$. Additionally, $\Theta = \{\theta_i\}$ represents the model parameters. This model characterizes how observations x are drawn from the distribution with such parameters. The class membership y_i can be considered a random variable. The class membership is dependent on the feature vector x_i and the mixture component g_i, i.e. $P(y_i|g_i, x_i)$. In other words, the class membership is determined by the characteristics of an observation (x_i) and the densities of the different Gaussian distributions g_i. Assuming for simplicity a binary classification, through the maximum a posteriori (MAP) criterion the following model describes the optimization objective for the classification task,

$$h(x) = argmax_{c \in \{Y,N\}} \sum_{j=1}^{n} P(y_i = c|g_i = j, x_i) \times P(g_i = j|x_i). \tag{2.2}$$

Here the first factor in the sum represents the probability of the class being c given the component and the observation. The second factor is the probability that the observation x_i is generated from component g_i. More specifically,

$$P(g_i = j|x_i) = \frac{\alpha_j f(x_i|\Theta)}{\sum_{k=1}^{n} \alpha_k f(x_i|\Theta_k)} \tag{2.3}$$

shows that the prior belief that x_i is generated by g_i can be viewed as the frequency it is generated by g_i over the frequency that the observation is generated by any component. The crucial explanation for why SSL can result in improved classification is that this probability is not dependent on a class label y_i. The unlabeled data enables the estimation of the different distributions of the components. If an unlabeled observation has a higher probability to be generated by a component that dominantly generates positive classes, it is also more likely to be a positive class. This is connected to the cluster assumption underlying SSL. It is assumed that observations that belong to the same cluster also have the same label. Similar assumptions are made in supervised learning [27, 26]

Having established the theoretical justification for why including unlabeled data in classification can result in better accuracy, it is of interest to formalize how datasets for SSL look like in practice. A data set for semi-supervised classification can be split into two parts,

$$S = \{(x_i, y_i)|x_i \in \mathbb{R}^d, y_i \in Y, 1 \leq i \leq m\}. \tag{2.4}$$

The set S consists of observations for which labels exists. These instances inform the first factor in (2.2).

$$U = \{x_i|x_i \in \mathbb{R}^d, 1 \leq i \leq M\} \tag{2.5}$$

The observations in U are unlabeled and can possibly assist in inferring the distribution of the population [28]. Though the two subsets are distinct, they are sampled from the same underlying distribution.

3
Methods

3.1 Research Design

The research is designed to address the question of whether chemical compounds cluster in metabolomic feature space according to their MeOAs. To answer this question (1) the compound-dose combinations are considered as individual observations and (2) the dose levels are summarized in curve shapes to predict MeOAs. By analyzing the data with these different approaches, insights into the relationship between toxicity and metabolites can be gained. This can facilitate the prediction of drug safety before clinical trials as well as inform further experimentation to explore the role metabolomics can play in predictive toxicology.

3.1.1 Individual Dose Levels

First, the compounds are considered on individual dose levels, i.e. each compound-dose combination is considered as an observation. The aim of the approach is to investigate whether on individual dose level toxic and non-toxic effects can be observed. As mentioned above, even toxic compounds are non-toxic in low doses. Distinguishing if and when compounds become toxic can be observed on the individual levels. The distinction can be made by comparing the metabolomic response to that of control substances. Dimethyl sulfoxide (DMSO) is used as a control for the lack of toxicity. Chlorpromazine at 316µM is utilized to kill the cells, thus representing control substance for toxicity.

The compounds are specifically investigated at low concentration where all compounds should cluster with DMSO. At the highest concentration toxic compounds should cluster with the "Death" control. The differentiation between the MeOA should be visible in-between these concentrations since the cells are still alive and the metabolizing of the compounds should be observable. As a side-effect the dose-respone nature of the data can be confirmed. Thus, the compounds are plotted at selected dose level.

Following the semi-supervised approach described previously, the identified clusters are labeled and a supervised machine learning algorithm is trained. This algorithm is then evaluated on a held out test set where toxicity is known. By using the accuracy score of the held out data set a fair evaluation is possible. Due to the focus

on the unsupervised part of the semi-supervised approach the supervised part is not implemented in this thesis.

3.1.2 Dose-Response Analysis

From the theoretical expectation of dose-response dependencies and empirical confirmation by [6] follows that there is a potential to extract more information from the dose-response relationship than from considering compounds at individual dose levels. As the dose is increased, toxic compounds might show different dose-response curves even though the final outcome (cytotoxicity) is the same. Incorporating the metabolomic response from all dose levels into the analysis thus could lead to better insights into the MeOAs.

3.1.2.1 Independence Assumption

The most straight forward way of including all dose levels is to ignore the dose all together and do the SSL on all observations. However, this violates the independence of observations assumption fundamental to statistical inference. Intuitively, when sampling data from the population any two observations should not have an influence on each other. Let $S = \{x_1, ...x_n\}$ be observations produced by an experiment. If the probability of observing x_i has no influence on observing x_j, where $x_i, x_j \in S$, both observations are independent.

More formally, let event $E_i \in S$ be an observation from an experiment. If the observations are independent that means,

$$P(E_1, E_2, ..., E_n) = P(E_1)P(E_2)...P(E_n) \tag{3.1}$$

where $P(E_i)$ is the probability distribution of observing that event [29]. In other words, observing a set of values from an experiment is the multiplication of the individual probabilities. Thus, it is feasible to estimate these.

Translating this to the experiment at hand, the observations would be independent if the occurrence of a metabolomic profile of one observation would not influence the values seen in another observation. However, due to the dose dependency this is not the case. Let x and y be two compounds measured at 9 dose levels, thus resulting in the metabolomic feature matrices $X = \{x_1, ..., x_9\}$ and $Y = \{y_1, ..., y_9\}$, where $x_i and y_i$ are vectors of length number of metabolites. Because the values in X are all generated from compound x, the vector x_1 will be more similar to $\forall x \in X$ than to $\forall y \in Y$. This violates the independence assumption which in turn can result in biased estimations.
When dependency between the observation exists the estimation needs to take these into account. (3.1) becomes

$$P(E_1, E_2, ..., E_n) = P(E_n|E_1, ..., E_{n-1})P(E_{n-1}|E_1, ..., E_{n-2})...P(E_2|E_1)P(E_1) \tag{3.2}$$

Failing to model the dependencies in (3.2) results in biased estimations since it does not approximate the true distribution generating the data. In regards to the experiment, not accounting for the systemic similarity between the observations in X, namely all observations being generated by the same compound, will result in biased estimations and thus in poor generalization.

3.1.2.2 Modelling of Dose-Dependency

To circumvent this violation of the independence assumption, a method that models the dependencies by summarizing the dose-response values of each compound is proposed. This method is based on knowing how the dependencies between observations, e.g. $P(E_n|E_1, ..., E_{n-1})$, relate to each other. Taking the known dependencies into account can lead to independent observations and enable further analysis.

The dependency stems from some observation being produced by the same compound at different dose levels. These doses are monotonically increasing and thus can be ordered. Thereby the values form a dose response curve with the interpretation of the cells' reaction to being exposed to increasing concentrations of the same compound. The shape of that curve includes the information over all dose levels. Hence, it is possible to summarize the information by determining the shape. Knowing that the dependency exits between observations in X and Y but not between them enables condensing for example X from $\{x_1, ..., x_9\}$ to a vector x_{DR} which holds categorical values for each curve shape.

The shapes of the curves are determined inductively, by clustering them and finding representative curves for each cluster. After grouping by compound and ordering by concentration, for each compound there exist a $m \times n$ matrix M where each row is a metabolite and the column a dose level. Each row forms a dose-response curve. A matrix A can be formed by adding the different compound matrices together such that A becomes a $l \times m \times n$ matrix where l is the number of compounds. The curves of all compound-metabolite combinations are then clustered with the K-Means clustering algorithm.

The same shape can occur at different dose levels, meaning that the same shape is simply shifted. In this case the biological interpretation is similar which means they should be clustered together. Thus, the identification of typical curves can potentially benefit from aligning the curves. This is done by using dynamic time warping (DTW) (as applied in [30] to K-Means++ clustering). This technique does not compare dose by dose but finds the most appropriate dose to compare the values with. Thus, shifts of the curves do not distort the assignment to a cluster. This distance function is implemented in addition to Euclidean distance.

To arrive at a single value that summarizes the response of a specific metabolite to a compound at different dose levels, the curves in the matrix are replaced by the cluster membership. As previously mentioned, this reduces the dimensionality of A to $l \times m$ by condensing X to x_{DR} and avoids the violation of the independence

assumption.

Subsequently, A is used to cluster the compounds according to the metabolomic responses they cause in liver cells. Since the feature space consists of categorical variables the K-Modes clustering algorithm is appropriate. Ultimately, a supervised algorithm is trained to predict unseen compounds MeOAs.

3.2 Dataset

In the following, the experiment generating the data and the preprocessing steps taken are described. The conditions under which the data was collected have essential influence on the external and internal validity of the findings. The experimental conditions specify how the data is generated. Therefore, they limit the generalizability to a larger population of conditions. The preprocessing steps specify how the data has to be transformed in order to arrive at the results. Ideally the preprocessing helps in making the data approximating better the underlying patterns which constitute ground truth by removing for instance noise and effects resulting from the experiment itself such as batch effects.

3.2.1 Experimental Design

The data is generated by in vitro experiments on HepG2 liver cells. The cell cultures are exposed to different chemical compounds and their reactions in form of metabolites is measured. These measurements are done with Acoustic Mist Ionization Mass Spectrometry (AMI-MS) [31]. In this project an untargeted metabolomics approach is chosen. This means that all known and unknown metabolites are captured by the AMI-MS. In contrast, targeted metablomics focuses predefined metabolites [32]. This done due to the exploratory objective of the research. The technology and approach enable high throughput which results in a rich data set of at the moment 160 chemical compounds at 9 different concentrations with 3000 features detected. Of these 3000 metabolites around 1000 could be identified and named as known metabolites. The metabolite concentration in the cells is measured 24h after the start of the exposure so that enough time has past for the cell to react. The data is generated in three different experiments performed in June, July and August 2018 and referred to accordingly in the rest of the report. The June data set itself consists of two different experimental runs. The July dataset constitutes the largest one.

The compounds are drugs, some of which are toxic. These toxicity levels are known since they have failed clinical trials or have been retracted from the market. For most compounds therefore exist drug induced liver injury (DILI) severity level annotation which categorizes the compound crudely in different levels of toxicity. However, for annotation for the different MeOAs subject matter experts have to investigate the scientific literature to label the compounds. Additionally, researching in what way a specific compound is toxic is in itself difficult and thus labelling is sparse.

As mentioned previously, the concentration, or dose, to which the cell is exposed is determinant in the adverse effects. This raises the challenge of determining which concentrations to include in the experiment. Additionally, including several observations of the same chemical at different concentrations violates the independence assumption underlying traditional statistical methods. Different methods rooted in the experimental design can be found in the literature. For instance, [11] and [10] include the concentration that inhibits 50% of the cell growth of the experimental subjects. [13] use 5% to 10% of the neonatal LC_{50} dose for a 21 day chronic exposure study. The neonatal LC_{50} is the lethal concentration at which 50% of newly born (in their case <24h old) die. Both these approaches are not applicable for the study at hand. The former is not suitable because the experiment in performed on in vitro cell cultures and thus the cells are not growing. The latter investigates chronic rather than acute toxicity. Ramirez et al.[6] performed range finding experiments a priori to the main experiment. They determined a high and low dose by considering the reduction in protein after 48h. They compared these two to a control substance for one chemical bezafribrate. Using repeated measurements, on a PCA plot the control, low dose and high dose all clustered separate from one another. This suggests dose-response dependency. Analysis with high and low doses with a subset of their observations supported the dose dependency argument. Ultimately, the authors decide to include only the observations at the highest dose in their analysis. They argue that at this level the toxic effects and the MeOA are most clearly observable.

To circumvent this challenge, this research proposes to take a more holistic view on the dose response behavior. Therefore, it is proposed to take into account the entire dose-response curve as input to the machine learning model by summarizing its information into one value. The results gained from this investigation can contribute to the scientific discussion on dose-response data.

3.2.2 Preprocessing

As is common with dose-response data, the concentrations of the compounds are log transformed in order to get a linear interpretation of the data due to the exponential increase in the original concentration. Additionally, the detected concentrations of the metabolites are log transformed for outliers to have less influence and due to the skewness of the data from the AMI-MS hence approximating a normal distribution more closely.

In the feature space there exist missing values. Most of these stem from the concentration of that metabolite being below the limit of detection by the AMI-MS. Different data sets are created with different cutoffs at which features are excluded from the data set. The cut offs are chosen at 10%, 20% and 40%. For features below the thresholds, two different imputation methods are applied. One method is based on the rationale that missing values are caused by undetectably small concentrations. Therefore, the minimum value of that feature is taken and divided by two.

The second method is based on the k-nearest neighbors with non-missing values. These values are averaged to impute the missing value.

Since compounds are tested in different experiments and on different plates of the AMI-MS, batch effects are introduced. This means that systematic bias is introduced by the conditions of the experiment [33]. Therefore, normalization methods have to take this into account. A common way of normalizing metabolomic data is to include an internal standard of which the concentration is known in each sample. Thus, all metabolites can be expressed in relation to this standard. However, the metabolites to be analyzed can have an influence on the internal standard if they have overlapping chromatographical peaks which is termed cross-contribution. A reliable normalization algorithm termed cross-contribution compensating multiple standard normalization (CCMN) that takes such cross-contributions into account is thus applied to the dataset [34].

Once the scalar values of the metabolites for each composite of compound and log-concentration are grouped into curves they are smoothed by applying a moving average and standardized to a mean of zero and unit variance. The former is done to decrease the influence of outliers on the curve. The latter is done to make the curves comparable with each other.

3.2.3 Batch Effect Correction

As mentioned above, the data set is generated by three experiments which can introduce batch effects. Even though the experimental conditions (cell type, solvent etc.) were kept the same there are conditions not under the control of the experimenter. Since these conditions introduce systemic bias to the data generated the data can show pattern not due to biological phenomena but because of the conditions. When integrating several data sets from different experiments into one global data set the difference in bias can mask the true biological signal. Therefore, it is important to control for such batch effects and if needed correct for them [35].

In order to correct for batch effects, the mutual nearest neighbors algorithm developed by [36] is applied to the three data sets. The algorithm has the advantage that not all observations in both datasets need to be the same. It suffices that there exists an intersection of some size. This is the case in the datasets in this thesis, thus making it a good candidate for batch correction.

3.3 Dimensionality Reduction

One of the major challenges that high-dimensional data brings with it is the difficulty of visualizing the observation so that the researcher can inspect. For visualization purposes, dimensionality reduction methods are needed to plot data in two dimensional space and at the same time preserve as much information as possible [37].

Principal Component Analysis (PCA) for continuous and Multiple Correspondence Analysis for categorical feature space are presented in the following.

3.3.1 Principal Component Analysis

PCA's aim is to reduce the dimensionalty by preserving the maximum amount of variability. This is done by finding new, uncorrelated variables, so-called Principal Components (PCs), that are linear functions of the original features. More specifically, PCA finds vectors in the original feature space in which the most variation can be found. This means that the first PC is a vector in the direction in which when projecting the observations onto that vector the variation is maximized. The second PC is an orthogonal vector that captures the second most variation from the data [37]. For a dataset with n features, or dimensions, n PCs would capture all variation and thereby information from the data. However, since it is known which PCs capture the most variation the k first PCs, where $k \leq n$, can be used to reduce the feature space by only losing small amounts of information.

To find the vector that maximizes the variation it is possible to multiply a random vector by the co-variance matrix repeatedly. It has been shown that this turns the random vector into the direction with the greatest variation as well as extends or reduces its length. The vectors that do not change directions when multiplied by the co-variance matrix are termed *eigenvectors*. Formally, $\Sigma e = \lambda e$ where Σ is the co-variance matrix and λ is the scaling of the *eigenvector*, termed *eigenvalue*. The principle component is the *eigenvector* with the largest *eigenvalue* [37].

The first PCs lend themselves for visualization since they capture large amounts of information from the data and can be plotted in two-dimensional space. It is important though to show how much variation is captured by these PCs since if all dimensions have similar variation the captured variation by the first PCs could be low, thereby not reflecting the structure of the original dataset well. A distinct advantage over other visualization techniques is the simple interpretation of PCA. Distance between observations in the reduced space reflect distance in the original space. Other methods such as t-distributed stochastic neighbor embedding (t-SNE) do not have such straight forward interpretation and the shape can be heavily influenced by hyperparameters such as perplexity in the case of t-SNE.

3.3.2 Multiple Correspondence Analysis

In order to visualize the feature space a multiple correspondence analysis is applied (MCA). This similar to principal component analysis (PCA) though it is appropriate to apply to categorical variables such as in the data set at hand. As mentioned above, the similarity measurement of Euclidean distance is not appropriate when analyzing categorical variables. MCA solves this by creating an indicator matrix M expressing the values as binary indicators, 1 for a category being present and 0 otherwise. With M the eigenvector problem is solved to find components that

summarize the data. Thus, MCA can be considered as an extension of PCA for categorical variables. However, since M inflates the number of variables by having binary columns for each category for each variable the explained variance is too low. Thus, this is corrected for as a final step. Plotting the resulting components enable a similar interpretation between rows as PCA [38].

3.4 Clustering

As mentioned above, clustering is essential in the approach chosen to predict MeOAs from metabolomic data. Thus, in the following K-Means and K-Modes, the main clustering algorithms used in this project, are described in depth. In addition to the traditional implementation with Euclidean distance, it is investigated whether the extraction of typical curves benefits from applying the K-Means algorithm with Dynamic Time Warping instead. For the clustering of the compounds the K-Modes algorithm is also applied due to its appropriateness to categorical feature space [39].

3.4.1 K-Means Clustering

The ultimate goal of a clustering algorithm is to assign observations $X = \{x_1, x_2, \ldots, x_n\}$ to clusters such that members of a cluster are more similar to each other than to those of other clusters. The K-Means algorithm assigns the observations to k different clusters by minimizing the within-cluster variation (i.e. the within-cluster sum of squares(WCSS)). Formally expressed,

Minimize

$$\sum_{j=1}^{k} \sum_{i=1}^{n} w_{i,j} ||x_i - c_j||_2 \tag{3.3}$$

subject to

$$\sum_{j=1}^{k} w_{i,j} = 1, \quad 1 \leq i \leq n \tag{3.4}$$

and

$$w_{i,j} \in \{0, 1\}, \quad 1 \leq i \leq n. \tag{3.5}$$

Where c_j represents the mean or centroid of the jth cluster. $w_{i,j}$ determines whether observation x_i is member of the jth cluster by taking the value of 1 if so and 0 otherwise. (3.2) ensures that each observation can only be member of one cluster [39].

This minimization problem is solved in an iterative manner in a two-step process. Before starting the iterative process, centroids $C = \{c_1, c_2, ..., c_k\}$ where $|C| = k$ are, for example randomly, initialized. The goal is to divide X into partitions $S = \{s_1, s_2, ..., s_k\}$ with corresponding centroids C such that the WCSS is minimized. To do so in the first step the Euclidean distance from an observation to all centroids in C is calculated and it becomes a member of the partition s whose centroid is closest. This is done for all observations in X. As a second step, all

centroids in C are recalculated as the mean of the corresponding members of S. For instance, c_1 is recalculated as the mean of s_1, c_2 as the mean of s_2 and so on. As these steps are repeated, the algorithm converges[39]. The value of k is a hyperparameter that needs to be chosen and different methods exist [40]. In the context of this research project, particular during the curve extraction, the number of curves is also influenced by the biological interpretation of the shapes.

It is important to note that there are drawbacks to the K-Means algorithm. Due to the optimization problem (3.1) to (3.3) being non-convex the algorithm tends to converge on a local minimum. Therefore, the algorithm is usually initialized several times with different centroids and the best result in terms of WCSS is chosen. Furthermore, the clusters in S are convex. Thus, K-Means is unlikely to find clusters that have different shapes [39].

3.4.2 Dynamic Time Warping

A common problem with sequential data is that the discrete points at which the observations are collected can vary even if the pattern, i.e. the dependency between the consecutive observations, is similar. An intuitive example from speech recognition is the same word spoken at different speeds. The shape of the sound waves is similar. However, calculating the Euclidean distance between them would be relatively large due to the difference in speed. Considering the problem at hand, when extracting curves, a cell might react similar to two different compounds however just delayed. That is, the dose-response curve is shifted either to the left or right. If this is the case, applying the clustering algorithm not with Euclidean distance but a more appropriate measure could be beneficial.

By using Dynamic Time Warping (DTW) it is possible to align curves that are shifted. DTW minimizes the distance between to curves not on a point by point bases, as is done when applying Euclidean distance. Rather, for two curves A and B a $n \times m$ matrix is created where one curve is captured by the horizontal and the other by the vertical axis. Each cell represents the distance from a_i to b_j. The warping path W that aligns the two curves can then be calculated by minimizing the distance between a_i and b_j[41].

To visualize the different shapes that are grouped in a cluster the centroid is not appropriate since it only reflects the mean of the different shifted curves. It is expected that the centroid will have less of a clear shape as when applying K-Means with Euclidean distance. Therefore, a K-Means algorithm with Euclidean distance is applied to each cluster from the K-Means with DTW. The centroids of the resulting clusters reflect the different shifts of the curve shape and thus can be visualized.

3.4.3 K-Modes Clustering

The K-Means algorithm is suitable for the first step in the approach chosen for predicting MeOAs, namely the curve extraction. However, for clustering the compounds

with categorical values K-Means becomes less appropriate due to the dissimilarity measure used and the way cluster centroids are defined. For categorical variables the Euclidean distance between two values is meaningless and thus cannot function as an appropriate dissimilarity measure. Similarly, calculating the mean of categorical variables has no proper interpretation. Therefore, [39] proposed the K-Modes algorithm to cluster observations in categorical feature space.

The K-Modes algorithm has the same steps as the K-Means algorithm. However, to solve the issue of the dissimilarity measure a simple matching is used instead of Euclidean distance. Between two observations X, Y with m features the dissimilarity is calculated by comparing the observations with corresponding features and counting how often they agree and disagree. The fewer times the observations disagree the more similar they are [39]. Formally expressed,

$$d_1(X, Y) = \sum_{j}^{m} \delta(x_j, y_j) \tag{3.6}$$

where

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j). \end{cases} \tag{3.7}$$

Instead of the mean as the centroid the mode of the partition is chosen. The mode is the combination of categorical values for the m features that minimizes the disagreement between itself and all the cluster members. Thus, let X be the set of all observations and $S \subset X$ be the partition for one cluster with n observations. The m dimensional vector Q is the mode of S if it minimizes,

$$D(S, Q) = \sum_{i=1}^{n} d_1(S_i, Q) \tag{3.8}$$

Note that $Q \in S$ is not a requirement.
From these changes from the K-Means algorithm the new cost function (the equivalent to (3.1)) is,

Minimize,

$$\sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{i,l} \delta(x_{i,l}, q_{l,j}) \tag{3.9}$$

subject to (3.2) and (3.3)

By minimizing the cost function the within-cluster similarity is maximized and the observations in X grouped so that similar observations belong to the same group.

The K-Modes algorithm is used to cluster the chemical compounds. Those clusters in turn are used to infer cluster labels according to the MeOAs.

4

Results

The results are reported for (1) the analysis of one dataset, the July experiment, and (2) the batch corrected large dataset separately. In doing so the advantages of a "clean" dataset, (1), and those of a large but potentially biased dataset, (2), can be compared. Each analysis follows the two approaches, individual dose levels and dose-response analysis, outlined in the research design. Since the results from the unsupervised part of the semi-supervised learning (SSL) approach do not suffice for a good labelling of the observations no supervised algorithm is trained and presented here. It is important to note that the ultimate goal of the unsupervised part is to inform a classification algorithm in order to make prediction. Thus, the results of this thesis are to be considered in the wider context of the SSL paradigm.

4.1 Single Experiment Analysis

For the analysis of a single dataset, the July experiment has been chosen due to its larger number of compounds included. Also the June experiment is composed of two different experiments introducing batch effects (see Figure A.2 in the appendix). The August experiment has significantly fewer compounds (see Figure A.3 in the appendix).

4.1.1 Individual Dose Levels

The differentiation between toxic and non-toxic compounds can most clearly be seen when inspecting the metabolomic profile at the highest concentration (HC). Figure 4.1 shows the July dataset reduced with PCA to two dimensions where the first PC captures 39% of variation and the second 10%. The compounds are plotted with control substances DMSO, "Death" control, MMP[1] as control for lipogenesis as well as compounds that are known to be toxic (active)[2].

[1] Tamoxifen at 150 µM known to be toxic [42]

[2] Tamoxifen [42], Chlorpromazine [43], Fluoxetine [44], cccp [45]

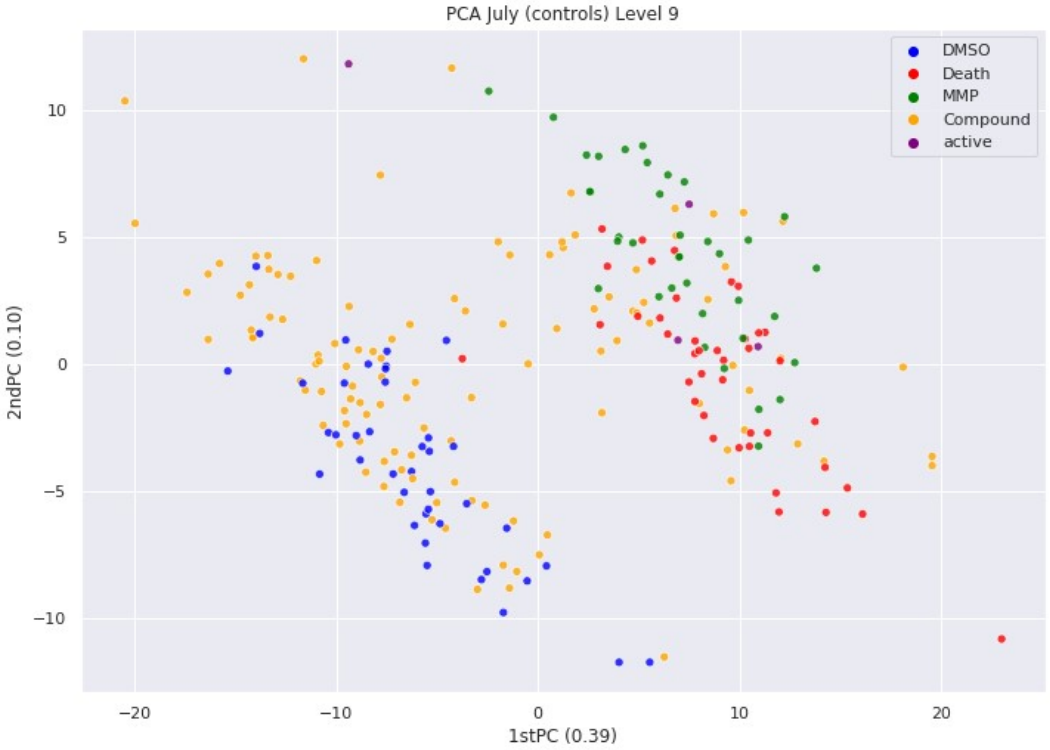
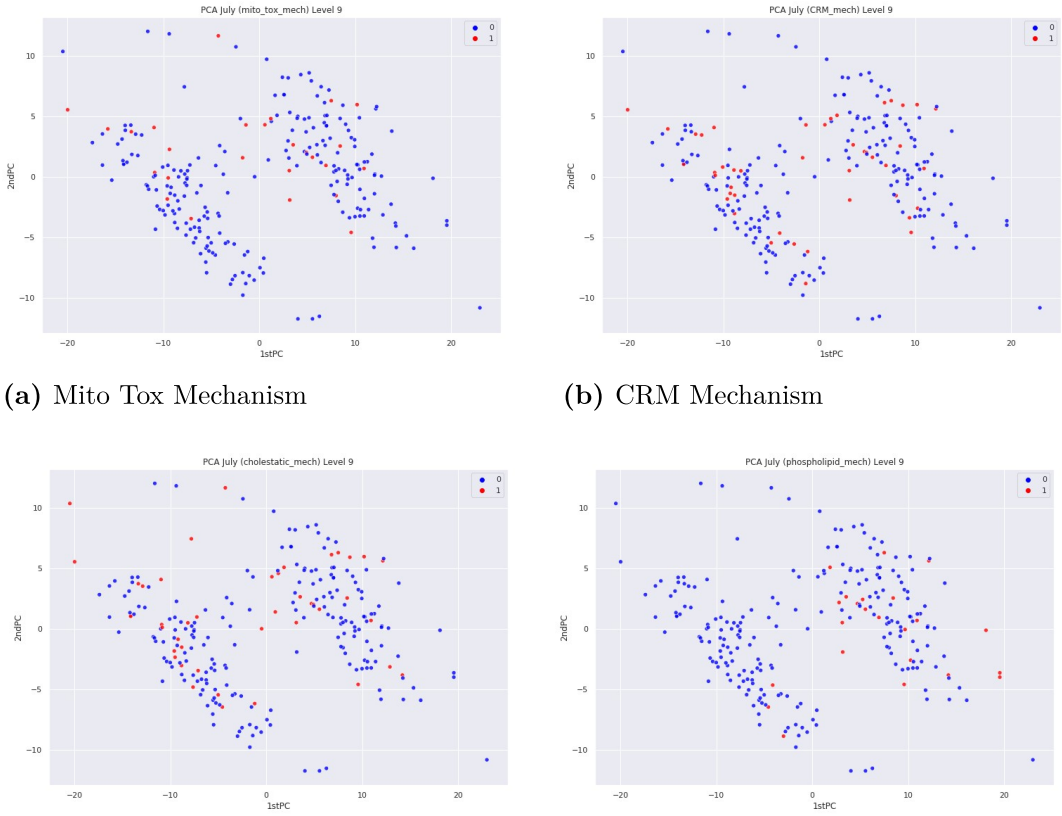


Figure 4.1: Control Substances at HC for July

As can be seen in Figure 4.1 two clusters can be identified in the data. The left cluster is dominated by DMSO indicating that compounds in this cluster have no toxic effect on the cell. The right cluster is dominated by the "Death" control substance as well as the MMP control. This indicates that compounds in this cluster have toxic effects on the cells. This is confirmed by the four active compounds which are part of the right cluster. Also five inactive compounds are part of the left cluster (see Figure A.2). Similar clusters can be seen in the June and August dataset (Figure A.2 and Figure A.3). The low density areas between the two clusters indicate a clear distinction between toxic and non-toxic compounds. Figure A.4 in the appendix shows that compounds move from the non-toxic to the toxic cluster starting from the seventh to ninth dose level. This supports the expectation of dose-dependency and that metabolomic data has the potential to show at least toxic effects in general.



(a) Mito Tox Mechanism

(b) CRM Mechanism

(c) Cholestatic Mechanism

(d) Phospholipids Mechanism

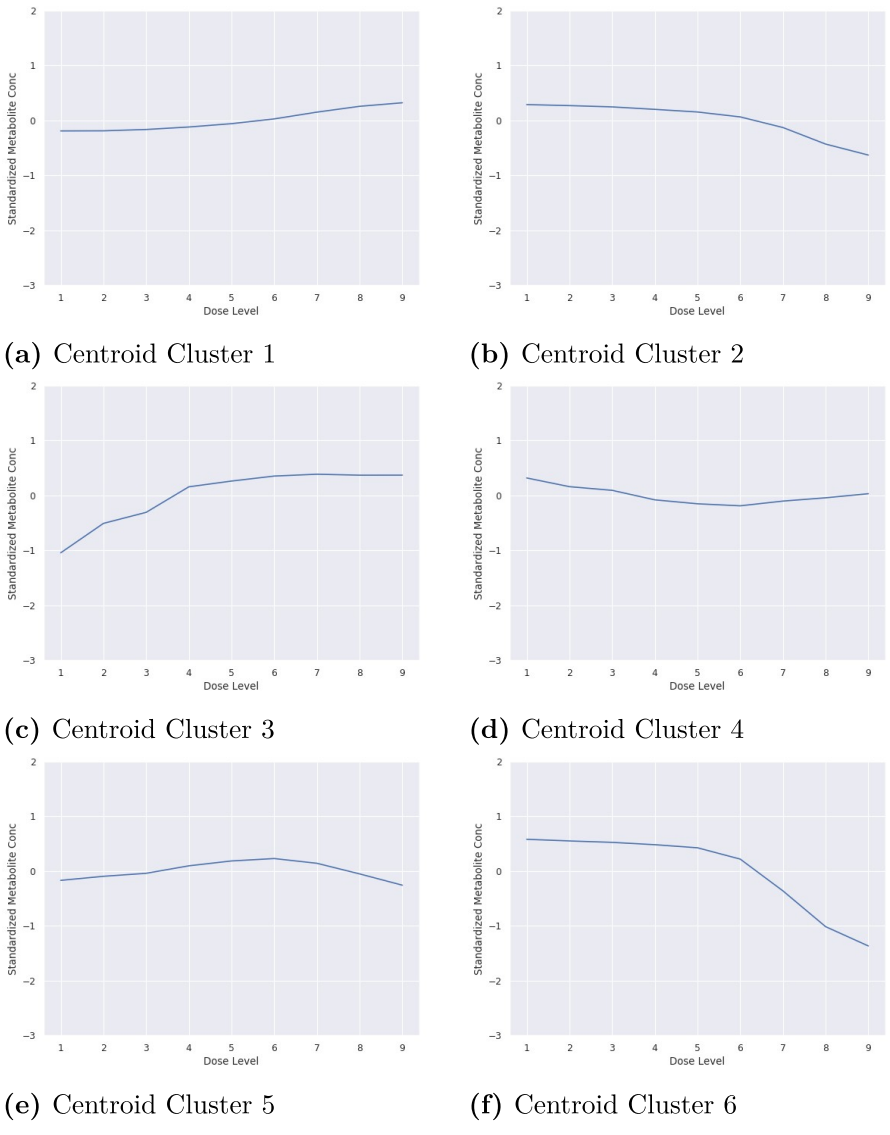
Figure 4.2: Different MeOAs at HC for July Dataset

Figure 4.2 shows the MeOAs plotted at the HC for the July dataset. It can be seen that not all compounds that are known to be toxic in human appear to be member of the right cluster which is associated with toxic controls. Also, the compounds that are known to cause toxicity through the respective MeOA and are part of the toxic cluster do not separate as clusters within the larger cluster. Thus, in this metabolomic feature space MeOAs are not obviously distinguishable. This in turn calls for methods that take into account more information than one dose level.

4.1.2 Dose Response Analysis

As shown above, analyzing compounds in metabolomic feature space can distinguish well between toxic and non-toxic compounds. However, considering the different mechanisms a distinction is not possible. Thus, in the following the results of the dose-response analysis is presented.

Firstly, the typical curves identified with K-Means clustering are presented for $k = 6$ as well as the cluster size. This is done to show the biological interpretation of the curves. Secondly, the MCA plots of the categorical feature space is presented, color coded according to MeOA. Lastly, the output of K-Modes clustering algorithm is presented in form of a table with the distribution of MeOAs.



(a) Centroid Cluster 1

(b) Centroid Cluster 2

(c) Centroid Cluster 3

(d) Centroid Cluster 4

(e) Centroid Cluster 5

(f) Centroid Cluster 6

Figure 4.3: Typical Curve of July Dataset

Figure 4.3 shows the centroids as a typical curve for each cluster generated by the K-Means algorithm. As described previously, the centroid is the point from which the distance to all observations is measured. It also represents the mean value of the cluster. Thus, it is a representation of the shape of the clusters. Since the mean is sensitive to outliers the analysis has also been done with the K-Median algorithm which uses the median instead of the mean.[3] The centroids in this case showed approximately the same shape and thus it is inferred that the results of the K-Means algorithm are not influenced significantly by any outliers and hence are reliable.

The curves that were used for clustering were standardized to a mean of zero by subtracting the mean of the curve to make them comparable with one another. The curves were not standardized to unit variance in order to observe, for instance, flat curves. Furthermore, the curves were smoothed by applying a moving average such that outliers have less influence and the noise level gets reduced. $k = 6$ was chosen inductively by comparing whether an increase of k by one would yield a significantly new curve shape.

The six typical curves presented in Figure 4.3 all have biological interpretations in regards to the metabolomic reaction by the cell. Cluster 1 groups curves that change insignificantly over the dose levels and thus can be considered flat curves. As Table 4.1 shows, this is one of the most common types of curve. This is biologically sound since most metabolites are not connected to toxicity. This is supported by the high number of curves in cluster 4 and 5 whose centroid also show fairly low activity. Cluster 6 shows a significant drop at dose level 6. This coincides with the observed changes in metabolomic profile of compounds (see Figure A.4). A possible interpretation would be the shutting down of processes within the cell or that molecules that are used to metabolize the foreign compound are used up. Centroid of cluster 3 shows the opposite reaction. An interpretation could be the increase in metabolites that are byproducts of the metabolizing of a compound as the cell gets under duress. Cluster 2 groups those reactions that show a less strong decline than cluster 6. In conclusion, all cluster shapes have a sound biological interpretation and cover a vast range of reactions by the cell, thus supporting the choice of $k = 6$.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Custer 6
#Members	5782	3672	2114	5927	5609	943

Table 4.1: Cluster Sizes for July DR-Analysis

[3] Results not shown

The categorical feature space can be clustered with the K-Modes algorithm. Ideally, the resulting clusters correspond, or at least are dominated by, one MeOA. However, due to the multiple MeOAs a compound can cause the cluster might be shared by two compounds. Nevertheless, some structure should be observable.

Mechanism	C-1 p	C-1 n	C-2 p	C-2 n	C-3 p	C-3 n	C-4 p	C-4 n
MitoTox	7	1	9	1	3	0	7	1
CRM	9	1	15	1	8	2	8	2
Phospho.	3	10	12	8	0	11	8	6
Cholest.	10	0	15	0	8	0	10	0

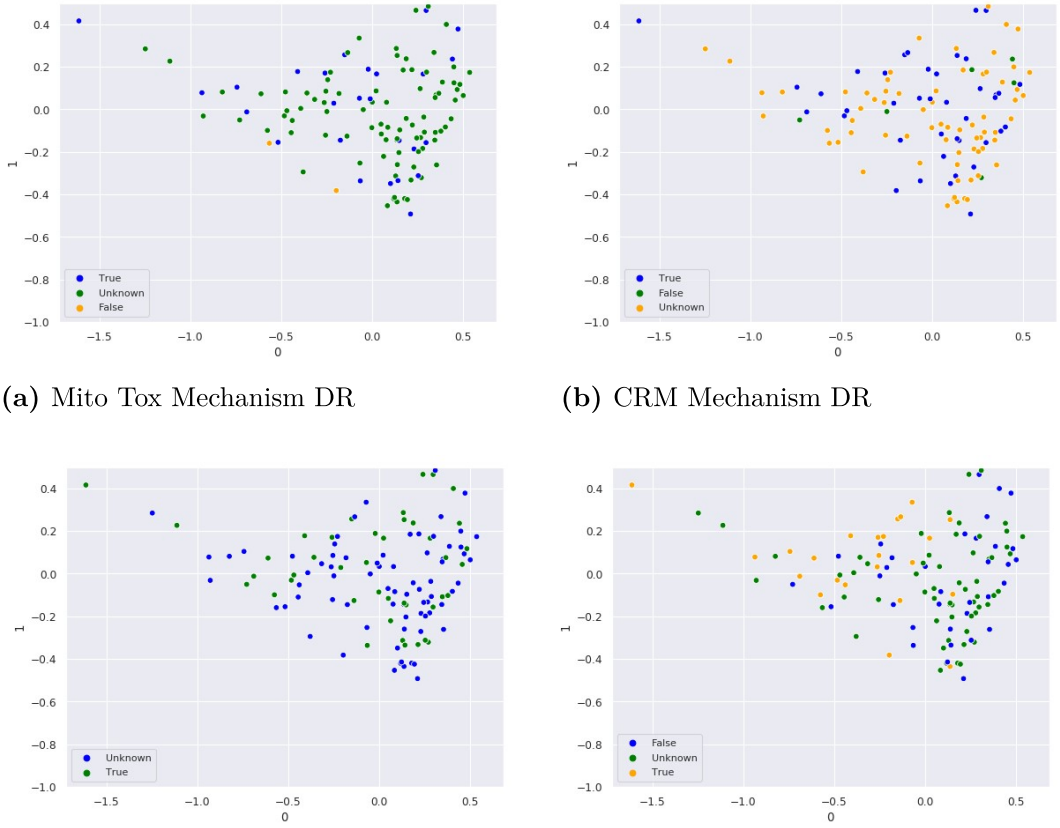
Table 4.2: Distribution of MeOAs in K-Modes Clusters (p = positive, n = negative)

From the distribution of MeOAs in the clusters it can be seen that mitochondrial impairment (MitoTox), chemically reactive metabolites (CRM) and the cholestatic MeOA have a fairly similar distribution over the four clusters. Considering the MeOA of lysosomal impairment, in form of phospholipidosis, one can identify that cluster one and three are dominated by the negative category and cluster two and four by the positive class. This could give some indication that they cluster in space.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
#Members	31	37	27	24

Table 4.3: Cluster Sizes for July K-Modes

Figure 4.4 shows the compounds in the metabolomic feature space after transforming the features into curve shapes thereby including the information over all dose levels in one analysis. The feature space is reduced by applying a MCA. No features selection has been done. In the data no clear clusters can be identified. Only the MeOA phospholipids shows that compounds that exhibit the mechanism cluster somewhat together. This is in line with the results from K-Modes clustering which identified two clusters that were dominated by "False" category for phospholidosis and two that tended to include dominantly "True" values. Moreover, the other different MeOAs do not cluster together.



(a) Mito Tox Mechanism DR

(b) CRM Mechanism DR

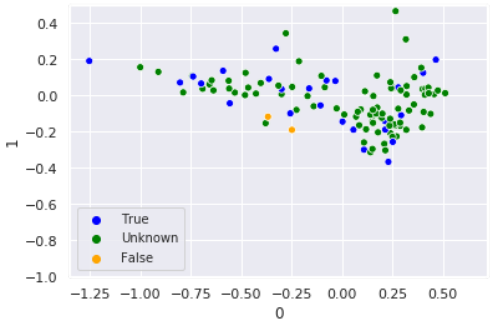
(c) Cholestatic Mechanism DR

(d) Phospholiqids Mechanism DR

Figure 4.4: Different MeOAs Dose-Response Analysis July

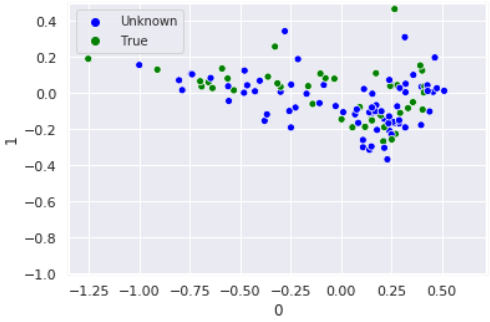
Due to the fixed dose range in the experiment the same reaction can occur at different dose levels which can be captured by using DTW instead of Euclidean distance as the similarity measurement between curves. Applying K-Means with DTW to the data shows that some curves are shifted but are correctly grouped in the same cluster (see Figure A.5). Cluster 1, 2 and 4 show the same curve shape at different dose levels. This shows that applying DTW to curve data can help grouping curves that express biologically similar phenomena together.

Coding the observations according to these clusters results in similar MCA plots as with Euclidean distance though the variance is decreased in the y-dimension and slightly increased in the x-dimension. Also, here can be seen that only the cholestatic MeOA shows some structure although again this could be due to uneven labeling of the observations. Many unknowns can be found to the right side of the plot. Ultimately, it is difficult to determine to what extent the application of DTW has improved the results. It is important to note that K-Means with DTW is computationally significantly more expensive which could be a challenge when including more dose levels.

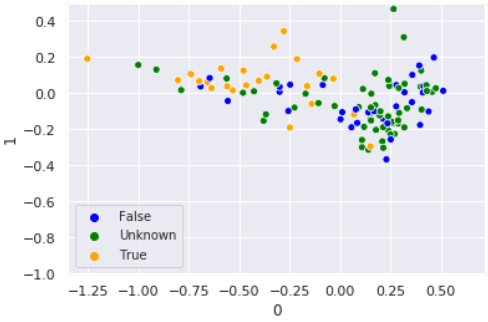


(a) Mito Tox Mechanism DTW July

(b) CRM Mechanism DTW July



(c) Cholestatic Mechanism DTW July



(d) Phospholiqids Mechanism DTW July

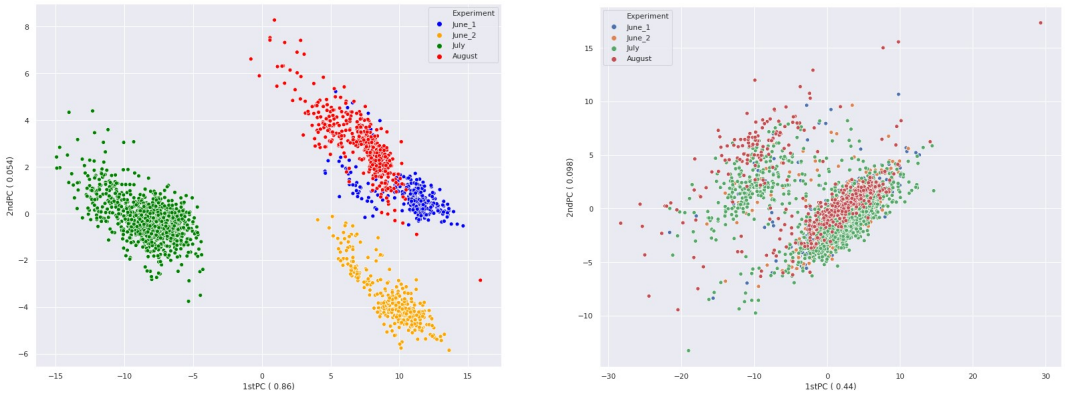
Figure 4.5: Different MeOAs Dynamic Time Warping July

4.2 Combined Experiment Analysis

The analysis can potentially benefit from including more observations and thereby increasing the chance of representing the underlying distribution generating the data. Thus, the following chapter presents the results from the merging of data from four different experiments. The number of compounds included in the analysis is significantly increased.

However, as Figure 4.6a shows, batch effects in the datasets can be found. This means that systemic influences from the actual experimental run affect all observations in each dataset. The clusters form according to the experiments thereby denying the chance to identify patterns rooted in biological effects such as toxicity.

To correct for the systemic effects in the different datasets the MNN algorithm with 20 mutual nearest neighbors has been applied to the data. Figure 4.6b shows the resulting dataset with batch effects corrected. The parameter of k has been identified inductively by plotting the batch corrected datasets for different k.



(a) Batch Effects Between Experiments **(b)** Batch Effect Removed (MNN k = 20)

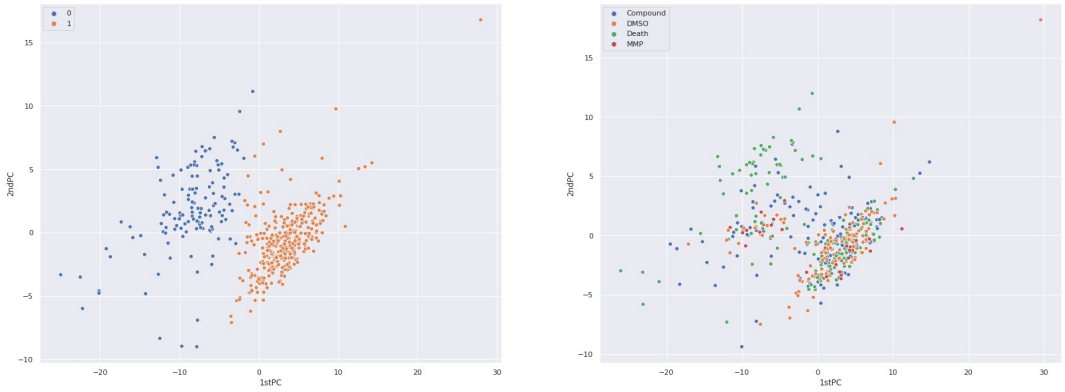
Figure 4.6: Batch Effect Correction with MNN

Though the algorithm was able to correct for the individual bias to a large extent still some pattern can be identified. This can be seen in the August and July dataset where the former is located on the right side of the latter. This can be an artifact of the correction. Also, the August dataset has lower variation which is the case in the uncorrected and corrected representation of the datasets. The difference in variation can be due to batch effects and thus to a failing of correcting it. However, an alternative hypothesis is that the smaller variation stems from the compounds included in the experiment. In this case the variation would have a biological cause and should not be corrected for. It is difficult to find out definitively which explanation is correct. Therefore, the analysis proceeds with the output of the algorithm, presented in Figure 4.6b, knowing that bias due to batch effects could influence the final results.

4.2.1 Individual Dose Levels

As has been shown with the individual dataset in section 4.1.1, by inspecting the compounds at the highest dose level it can be possible to identify clusters that show a distinction between toxic and non-toxic compounds. The controls for "Death" and for inactivity, DMSO, can provide information for such distinction.

Figure 4.7 shows that there are two clusters that are separated by a low density area between them.[4] The right cluster has low and the left cluster high variation. In contrast to the individual dataset from July, the control substances do not clearly correspond with the clusters. However, it seems that the upper part of the left cluster is dominated by "Death" control and the right cluster by DMSO. The lack of clear distinction between the controls could be a result of the merging of different datasets since there might still be distortions remaining.



(a) Batch Corrected K-Means **(b)** Batch Corrected HC

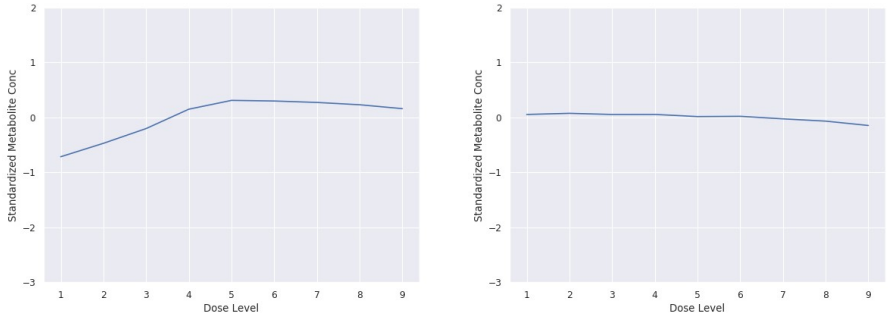
Figure 4.7: Batch Corrected at HC

As with the individual dataset, the batch corrected dataset could benefit from taking into account the metabolomic path a compound takes. Hence, the next section presents the results from the dose-response analysis.

4.2.2 Dose-Response Analysis

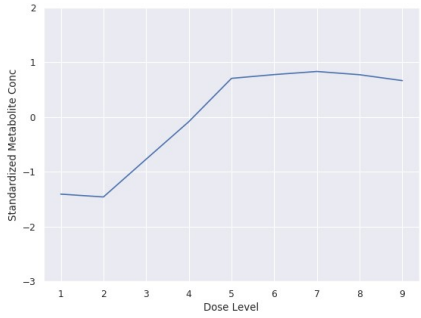
The following presents the dose-response analysis of the batch corrected dataset including the experiments from June, July and August. First the identified typical curves are presented as well as their number of occurrences within the dataset. The compounds are then visualized in metabolomic feature space according to curve shape by inspecting the MCA plots. Thereafter, the clustering with K-Modes is shown. Due to the insufficient results from the unsupervised methods a supervised algorithm is not applied and thus no results shown.

[4] Silhouette Score of 0.351 which indicates some clustering

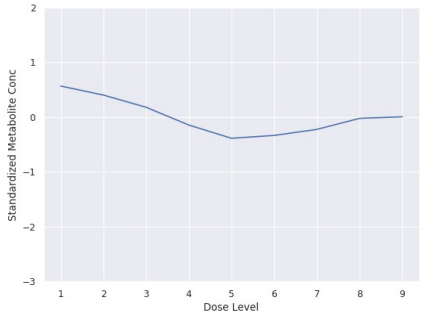


(a) Centroid Cluster 1

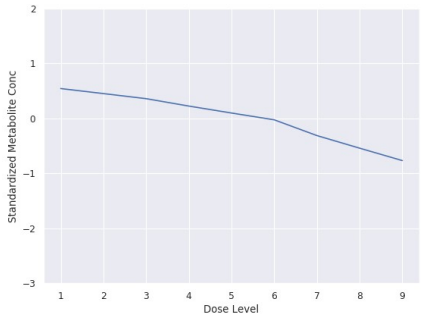
(b) Centroid Cluster 2



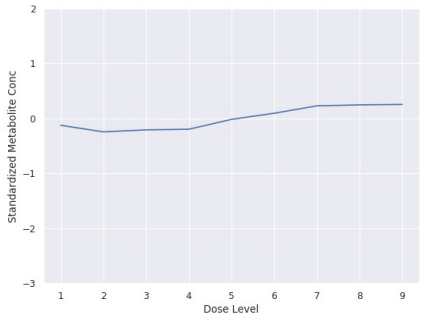
(c) Centroid Cluster 3



(d) Centroid Cluster 4



(e) Centroid Cluster 5



(f) Centroid Cluster 6

Figure 4.8: Typical Curve of Batch Corrected Dataset

The typical curves shown in Figure 4.8, as in the previous analysis on the single dataset, correspond to a biological interpretation as a reaction of a cell to a foreign compound. Centroid of cluster 1 corresponds to a cell reacting to a compound by metabolizing it but this increase in this levels off. Cluster 2 clearly groups the flat curves which represent the inactivity of that metabolite. This could be due to the metabolite not being involved in the metabolizing of a compound or that the compound is not toxic and thus not requiring such metabolizing. Similarly cluster 6 shows low activity. As in the individual dataset these clusters are dominating the dataset as can be seen in Table 4.4 which is expected. Cluster 3 shows a strong reaction to the increased dose that levels off whereas cluster 4 shows a strong decrease that becomes steady. Cluster 5 shows a constant decrease. In conclusion, the different identified curves represent different biochemical reactions by the cells and thus should be suitable for further analysis.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Custer 6
#Members	2042	5154	533	1387	1041	3367

Table 4.4: Cluster Sizes for Batch Corrected DR-Analysis

As can be seen in Table 4.5 , the positive and negative categories are evenly spread between the clusters when considering the different cluster sizes. There is no clusters that clearly groups the positive category and one for the negative category for any of the MeOAs.

Mechanism	C-1 p	C-1 n	C-2 p	C-2 n	C-3 p	C-3 n	C-4 p	C-4 n
MitoTox	7	1	16	1	1	1	6	0
CRM	12	1	23	4	1	0	8	1
Phospho.	8	7	21	15	1	2	11	13
Cholest.	12	0	25	0	3	0	11	0

Table 4.5: Distribution of MeOAs in K-Modes Clusters (p = positive, n = negative)

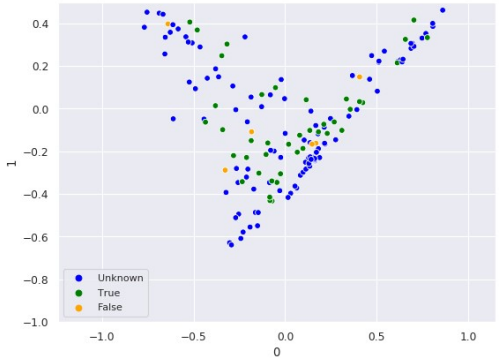
Between the clusters there is a strong imbalance with cluster two having the highest number of members with 83 compounds and cluster three only 13. Figure 4.9 can offer a possible explanation. The compounds in the middle are fairly dense and at the upper left and right corner are more sparsely distributed. Therefore, K-Modes clusters the dense area together in one cluster and the sparser ones into the other clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
#Members	27	83	13	38

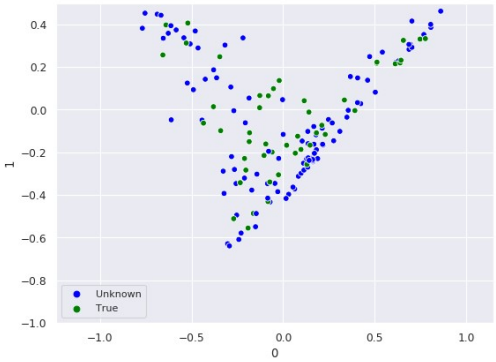
Table 4.6: Cluster Sizes for BC K-Modes

After coding the metabolomic reactions of the cell to the compounds according to their curve shapes and reducing the feature space to two dimensions with MCA, the plots in Figure 4.9 show the different compounds and the MeOAs they cause. Inspecting the plots no clear clusters can be identified. Also, the different compounds do not cluster together within the point cloud according to the MeOAs.

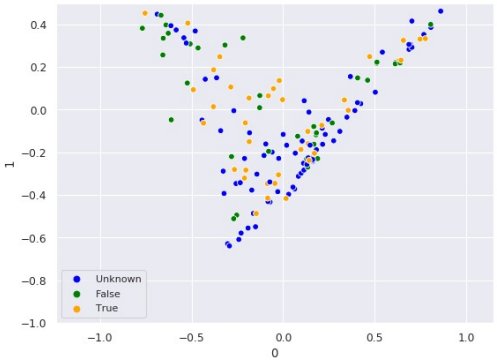
(a) Mito Tox Mechanism DR BC



(b) CRM Mechanism DR BC



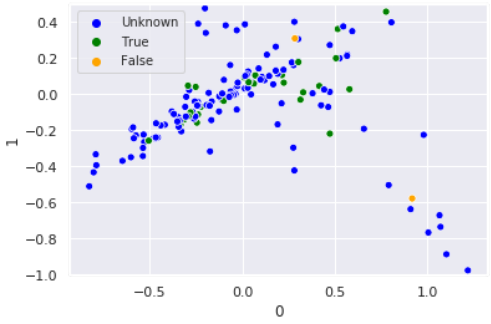
(c) Cholestatic Mechanism DR BC



(d) Phospholiqids Mechanism DR BC

Figure 4.9: Different MeOAs Dose-Response Analysis Batch Corrected (BC)

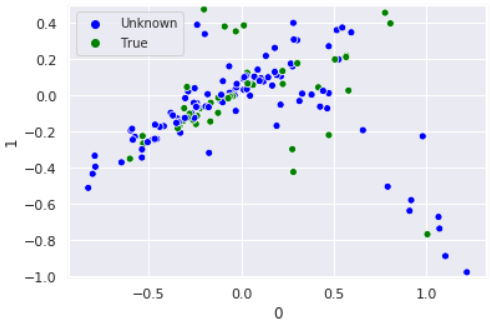
Considering the MCA plots of the K-Means applied with DTW, one can see that the curve alignment did not result in more clearly separated clusters of points. Figure A.6 (see Appendix) shows that the alignment had some effect for instance in the case of cluster one where a shifted U-shape can be observed in two of the curves. However, adjusting for the shifts in curve did not result in a better distinction between the MeOAs.



(a) Mito Tox Mechanism DTW BC

(b) CRM Mechanism DTW BC

Cholestatic.png



(c) Cholestatic Mechanism DTW BC

(d) Phospholiqids Mechanism DTW BC

Figure 4.10: Different MeOAs Dynamic Time Warping BC

Overall, the results show that the methodology to identify typical curves worked well since a variety of different curves could be identified. The second step of the dose-response analysis however did not result in positive results. Neither the individual nor on the batch corrected dataset did result in clusters that could be utilized for labelling unknown observations. This hinders building a good supervised model to predict MeOAs of unseen compounds. This impediment is further analyzed in the discussion section.

5
Conclusion

5.1 Discussion

The analysis of observations on high concentration from an individual experiment has shown that toxic effects can be identified from metabolomic data. This could not definitively be confirmed with the batch corrected dataset. Yet, this could stem from distortions not corrected for by the batch correction algorithm. Combining the strong evidence from the individual dataset and the weaker evidence from the batch corrected dataset one can conclude that toxicity can be identified on individual dose levels in metabolomic feature space.

However, it turned out to be difficult to distinguish between the different MeOAs on individual dose levels as well as when taking into account all dose levels. This can stem from different causes. (1) Dead cells could show a common metabolomic profile regardless by which MeOA they died. (2) The range of doses tested are not fine enough to identify meaningful paths to cell death. (3) The MeOAs not being mutually exclusive could distort the clustering.

In case (1) is correct, the dose-response analysis should be able to provide better insights since it is taking into account the metabolomic path the cell has taken in the process of dying. The analysis of the July dataset has shown that at dose levels 7 to 9 the compounds are moving from the non-toxic cluster to the toxic one. Thus, the process of dying is not a discrete event in terms of metabolomic response by the cell. Rather the process can be observed as the dose is increased until reaching a lethal level. Since a toxic effect can only be seen in the highest three concentrations, including more dose levels between level 7 and 9 could be providing more insights by representing the metabolomic reaction by the cell more fine grained (2). Approaches to this could be to use dose levels six to nine from the dataset analyzed in this thesis or by performing range finding experiments as done in [6].

Additionally, the analysis could benefit by using the information from the metabolomic reaction over increasing dose levels in different ways. During the dose-response analysis it has been shown that biologically sound shapes in the data have been found. However, when coding the shapes into categorical variables some information is lost. As an example, consider three curves a, b, c where a is increasing with the concentration, b is strongly decreasing as the dose increases and c is slightly decreasing as the dose increases. It stands to reason that curve b and c are describing a more

similar reaction by the cell than a and b and a and c. However, when transforming these variables into categorical variables this information about the relation between the shapes is lost since no notion of distance between categorical variables exists. Therefore, research building upon the findings presented in this thesis could develop such methodology that can model these relationships by quantifying the differences between curve shapes. Thereby such potential information loss could be avoided.

As another impediment to finding meaningful clusters in the data is the lack of mutually exclusive labelling of the compounds (3). Since a compound can be toxic by causing multiple MeOAs a cluster for the different combinations of MeOAs should form. A compound that causes mechanisms m_1 and m_2 should have a different metabolomic profile than one that causes m_1 and m_3. This makes finding clusters with clear low density areas between them could be more difficult since the two compounds share parts of their metabolomic profile due to m_1.

5.2 Conclusion

The aim of this thesis is to predict MeOAs of chemical compounds that are drug candidates in order to assess the probability of safety before clinical trials are initiated. To do so the potential of metabomolics as a source of information for such prediction is assessed. Since labelling is generally sparse in this domain, as is the case in this thesis, a semi-supervised machine learning approach is taken to utilize a larger number of compounds. The thesis investigates an approach that takes into account the information over several dose levels, thereby capturing information about the reaction of cells to different concentrations. Since in the cluster-then-label approach the unsupervised part is essential for good performance of classification algorithms the finding of clusters in the data that correspond to MeOAs has been focused on.

The analysis of the individual observations at different dose levels has shown that toxicity clusters in metabolomic feature space. This could be clearly shown when applying the analysis on individual datasets and it is less clearly shown when merging data from different experiments and correcting for batch effects. At higher dose levels compounds could be distinguished gradually from the non-toxic control. Thus, it can be concluded that metabolomic responses can show toxic effects, that AMI-MS experiments are suitable for capturing this information and that dose-response effects hold information that can be utilized for predicting toxicity.

In this thesis, a method is proposed that codes the metabolomic response by cells exposed to different compounds according to their dose-response curves in order to create a transformed feature space that can take into account not only single dose levels but to add the dimension of dose levels. This is done by K-Means to cluster the curves and using K-Modes to cluster the transformed feature space.

Results from the method show that the compounds do not clearly cluster according to their MeOAs in the transformed feature space. This is also the case when

applying K-Means with DTW to align curves with the same shape. As reasons for these results it is hypothesized that this could be due to the lack of mutual exclusivity between the MeOAs. This would increase the expected number of clusters and hence require more data and make distinctions more difficult. Moreover, it is expected that a finer grained increase in dose levels could benefit the analysis since this could refine the curves extracted from the dose-dependency.

Bibliography

[1] David W. Thomas, Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay, "Clinical Development Success Rates," *BioMedTracker*, vol. June, no. June, 2016.

[2] A. B. Haberman, "Overcoming phase II attrition problems," *Genetic engineering & biotechnology News*, vol. 29, no. 14, pp. 64–67, 2009.

[3] R. Powers, "NMR metabolomics and drug discovery," *Magnetic Resonance in Chemistry*, vol. 2009, no. May, pp. S2–S11, 2009.

[4] C. Goldring, A. Norris, N. Kitteringham, M. D. Aleo, D. J. Antoine, J. Heslop, B. A. Howell, M. Ingelman-sundberg, R. Kia, L. Kamalian, S. Koerber, J.-c. Martinou, A. Mercer, J. Moggs, D. J. Naisbitt, C. Powell, J. Sidaway, R. Sison-young, J. Snoeys, B. V. D. Water, P. B. Watkins, R. J. Weaver, A. Wolf, F. Zhang, and B. K. Park, "Mechanism-Based Markers of Drug-Induced Liver Injury to Improve the Physiological Relevance and Predictivity of In Vitro Models," vol. 1, no. 3, pp. 175–186, 2015.

[5] N. Noureddin and N. Kaplowitz, "Overview of Mechanisms of Drug-Induced Liver Injury (DILI) and Key Challenges in DILI Research," in *Drug-Induced Liver Toxicity* (M. Chen and Y. Will, eds.), ch. Overview o, pp. 3–21, Humana Press, 2018.

[6] T. Ramirez, A. Strigun, A. Verlohner, H. A. Huener, E. Peter, M. Herold, N. Bordag, W. Mellert, T. Walk, M. Spitzer, X. Jiang, S. Sperber, T. Hofmann, T. Hartung, H. Kamp, and B. van Ravenzwaay, "Prediction of liver toxicity and mode of action using metabolomics in vitro in HepG2 cells," *Archives of Toxicology*, vol. 92, no. 2, pp. 893–906, 2018.

[7] M. Yoon, J. L. Campbell, M. E. Andersen, and H. J. Clewell, "Quantitative in vitro to in vivo extrapolation of cell-based toxicity assay results," *Critical Reviews in Toxicology*, vol. 42, no. 8, pp. 633–652, 2012.

[8] S. Dragovic, N. P. Vermeulen, H. H. Gerets, P. G. Hewitt, M. Ingelman-Sundberg, B. K. Park, S. Juhila, J. Snoeys, and R. J. Weaver, "Evidence-based selection of training compounds for use in the mechanism-based integrated prediction of drug-induced liver injury in man," *Archives of Toxicology*, vol. 90, no. 12, pp. 2979–3003, 2016.

[9] H. Kong and S. West, "WMA DECLARATION OF HELSINKI – ETHICAL PRINCIPLES FOR Scienti c Requirements and Research Protocols," no. June 1964, pp. 29–32, 2013.

[10] R. Baptista, D. M. Fazakerley, M. Beckmann, L. Baillie, and L. A. J. Mur, "Untargeted metabolomics reveals a new mode of action of pretomanid (PA-824)," *Scientific Reports*, no. March, pp. 1–7, 2018.

[11] S. Halouska, R. J. Fenton, R. G. Barletta, and R. Powers, "Predicting the in vivo Mechanism of Action for Drug Leads using NMR Metabolomics," *ACS Chem Biol*, vol. 7, no. 1, pp. 166–171, 2012.

[12] S. V. Vulimiri, A. Berger, and B. Sonawane, "The potential of metabolomic approaches for investigating mode(s) of action of xenobiotics: Case study with carbon tetrachloride," *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*, vol. 722, no. 2, pp. 147–153, 2011.

[13] N. S. T. Taylor, A. Gavin, and M. R. Viant, "Metabolomics Discovers Early-Response Metabolic Biomarkers that Can Predict Chronic Reproductive Fitness in Individual Daphnia magna Nadine," *Metabolites*, vol. 8, no. 42, 2018.

[14] G. J. Burton, E. Jauniaux, and F. Medicine, "Oxidative stress," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 25, no. 3, pp. 287–299, 2011.

[15] N. Chaudhari, P. Talwar, A. Parimisetty, and C. Lefebvre, "A molecular web : endoplasmic reticulum stress , inflammation , and oxidative stress," vol. 8, no. July, pp. 1–15, 2014.

[16] J. Neustadt and S. R. Pieczenik, "Medication-induced mitochondrial damage and disease," *Mol.Nutr.FoodRes*, vol. 52, pp. 780–788, 2008.

[17] A. Sinha, X. Lu, L. Wu, D. Tan, Y. Li, J. Chen, and R. Jain, "Voltammetric sensing of biomolecules at carbon based electrode interfaces : A review," *Trends in Analytical Chemistry*, vol. 98, pp. 174–189, 2018.

[18] M. Cuperlovic-Culf, "Machine learning methods for analysis of metabolic data and metabolic pathway modeling," *Metabolites*, vol. 8, no. 1, 2018.

[19] G. T. Ankley, R. S. Bennett, R. J. Erickson, D. J. Hoff, M. W. Hornung, R. D. Johnson, D. R. Mount, J. W. Nichols, C. L. Russom, P. K. Schmieder, J. A. Serrrano, J. E. Tietge, and D. L. Villeneuve, "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment," *Environmental Toxicology and Chemistry*, vol. 29, no. 3, pp. 730–741, 2010.

[20] L. Cui, S. Yoon, R. F. Schinazi, and J.-p. Sommadossi, "Cellular and Molecular Events Leading to Mitochondrial Toxicity Liver Cells," vol. 95, no. February, pp. 555–563, 1995.

[21] D. Glick, S. Barth, and K. F. Macleod, "Autophagy : cellular and molecular mechanisms," *Journal of Pathology*, vol. 221, no. 1, pp. 3–12, 2010.

[22] J. Shayman and A. Abe, "DRUG INDUCED PHOSPHOLIPIDOSIS: AN ACQUIRED LYSOSOMAL STORAGE DISORDER," *Biochim Biophys Acta*, vol. 1831, no. 3, pp. 602–611, 2013.

[23] X. Yang, L. K. Schnackenberg, Q. Shi, and W. F. Salminen, "Hepatic toxicity biomarkers," *Biomarkers in Toxicology*, pp. 241–259, 2014.

[24] M. Robles-díaz, I. Medina-caliz, C. Stephens, R. Andrade, and M. I. Lucena, "Biomarkers in DILI : One More Step Forward," *Frontiers in Pharmacology*, vol. 7, no. August, 2016.

[25] H. Gan, N. Sang, R. Huang, X. Tong, and Z. Dan, "Neurocomputing Using clustering analysis to improve semi-supervised classification," *Neurocomputing*, vol. 101, pp. 290–298, 2013.

[26] X. Zhu, *Semi-supervised Learning Literature Survey*. University of Wisconsin Madison, 2005.

[27] Z.-h. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. August 2017, pp. 44–53, 2018.

[28] F. Schwenker and E. Trentin, "Pattern classification and clustering : A review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.

[29] G. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford University Press, 2001.

[30] G. F. Zhang, Yuan & Hepner, "The dynamic-time-warping-based k-means++ clustering and its application in phenoregion delineation," *International Journal of Remote Sensing*, vol. 38, no. 6, pp. 1720–1736, 2017.

[31] I. Sinclair, M. Bachman, D. Addison, M. Rohman, D. C. Murray, G. Davies, E. Mouchet, M. E. Tonge, R. G. Stearns, L. Ghislain, S. S. Datwani, L. Majlof, E. Hall, G. R. Jones, E. Hoyes, J. Olechno, R. N. Ellson, P. E. Barran, S. D. Pringle, M. R. Morris, and J. Wing, "Acoustic Mist Ionization Platform for Direct and Contactless Ultrahigh-Throughput Mass Spectrometry Analysis of Liquid Samples," *Anal.Chem.*, vol. 91, pp. 3790 – 3794, 2019.

[32] C. H. Johnson, J. Ivanisevic, and G. Siuzdak, "Metabolomics: beyond biomarkers and towards mechanisms," *Molecular Cell Biology*, vol. Advanced Online Publication, pp. 1–9, 2016.

[33] A. M. D. Livera, M. Sysi-aho, L. Jacob, J. A. Gagnon-bartsch, S. Castillo, J. A. Simpson, T. P. Speed, B. Unit, Z. B. Oy, and B. Division, "HHS Public Access," *Analytical Chemistry*, vol. 87, no. 7, pp. 3606–3615, 2016.

[34] H. Redestig, A. Fukushima, H. Stenlund, T. Moritz, and M. Arita, "Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data," *Analytical Chemistry*, vol. 81, no. 19, pp. 7974–7980, 2009.

[35] W. Wen, B. Goh, W. Wang, and L. Wong, "Why Batch Effects Matter in Omics Data , and How to Avoid Them," *Trends in Biotechnology*, vol. 35, no. 6, pp. 498–507, 2017.

[36] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Europe PMC Funders Group Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbours," *Nat Biotechnol.*, vol. 36, no. 5, pp. 421–427, 2018.

[37] I. T. Jolliffe, J. Cadima, and J. Cadima, "Principal component analysis : a review and recent developments," *Phil. Trans. R. Soc A*, vol. 374, no. 20150202, 2016.

[38] D. Abdi, Hervé Valentin, "Multiple correspondence analysis." `https://www.utdallas.edu/~herve/Abdi-MCA2007-pretty.pdf`, 2007. Accessed: 2019-04-11.

[39] Z. Huang, "Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 304, no. 2, pp. 283–304, 1998.

[40] T. M. Kodinariya and P. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.

[41] D. J. Bemdt and J. Clifford, "Using Dynamic Time Warping to FindPatterns in Time Series," *AAAI Technical Report WS-94-03.*, vol. 03, pp. 359–370, 1994.

[42] G. Yang, S. Nowsheen, K. Aziz, and A. G. Georgakilas, "Pharmacology & Therapeutics Toxicity and adverse effects of Tamoxifen and other anti-estrogen drugs," *Pharmacology and Therapeutics*, vol. 139, no. 3, pp. 392–404, 2013.

[43] L. Radenovic and G. Kartelija, "Effect of Chlorpromazine on Human and Murine Intracellular Carboxylesterases," *Biochemistry*, vol. 69, no. 4, pp. 381–386, 2004.

[44] I. Inkielewicz-stêpniak, "Impact of fluoxetine on liver damage in rats," *Pharmacol Rep.*, vol. 63, no. 2, pp. 441–447, 2011.

[45] Y. Sun, S. Eun, and H. Chul, "PGAM5 regulates PINK1 / Parkin-mediated mitophagy via DRP1 in CCCP- induced mitochondrial dysfunction," *Toxicology Letters*, vol. 284, no. August 2017, pp. 120–128, 2018.

A

Appendix 1

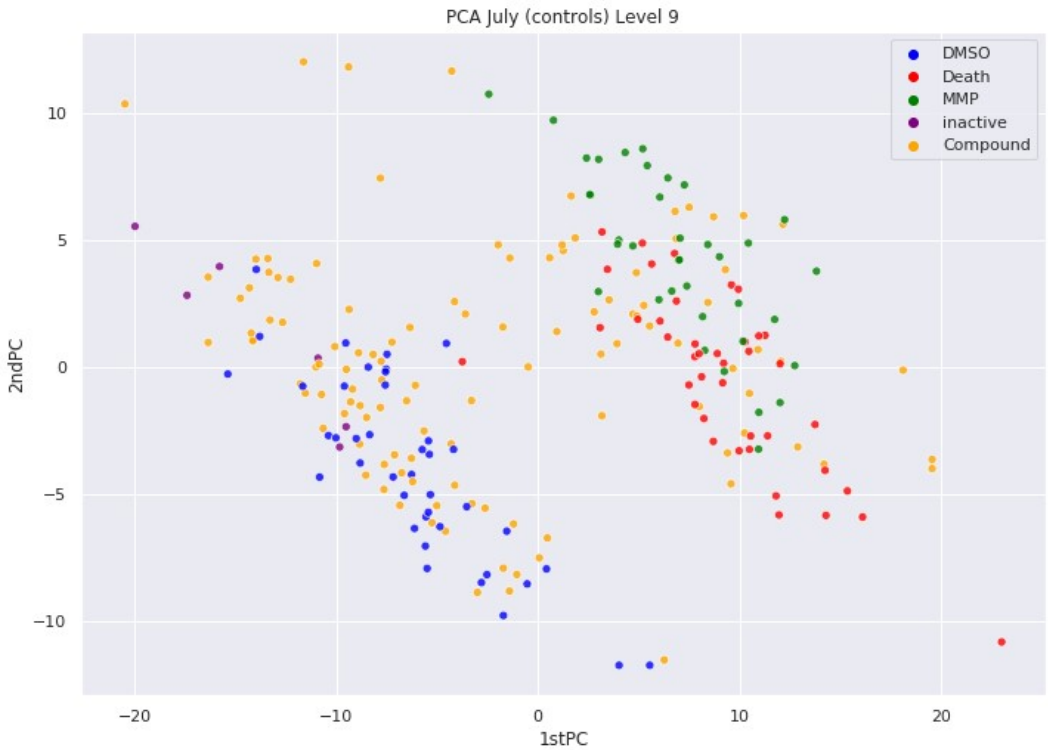


Figure A.1: Control Substances at HC for July (inactive)

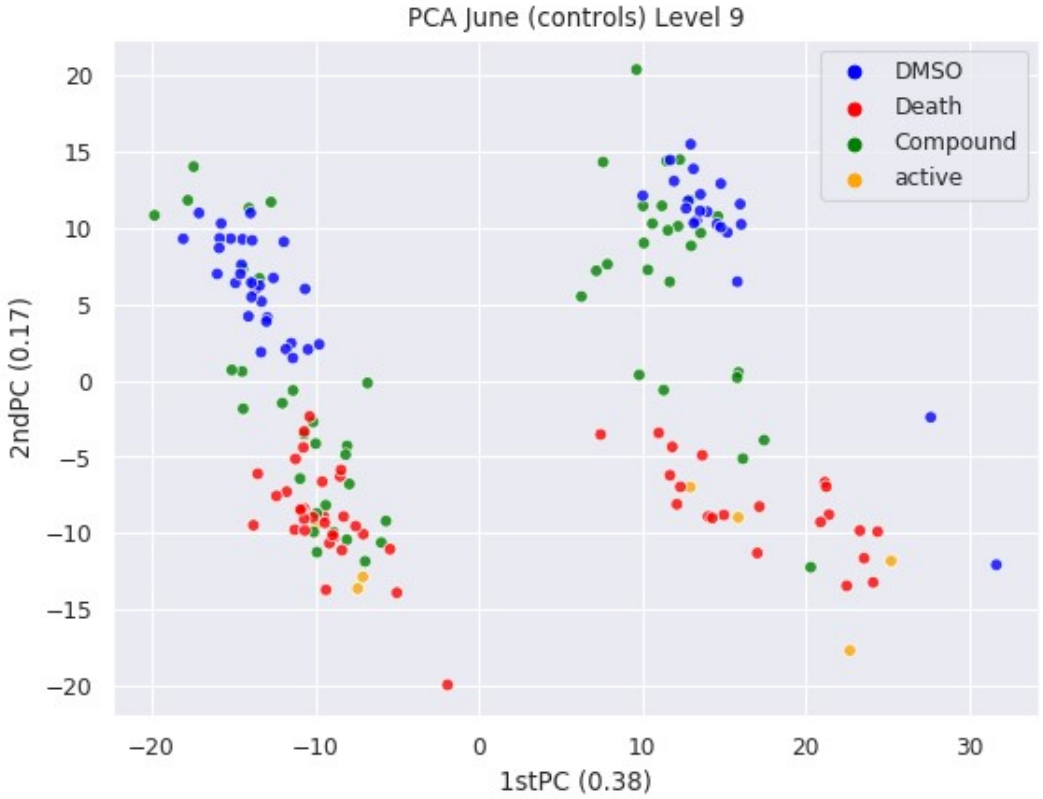


Figure A.2: Control Substances at HC for June

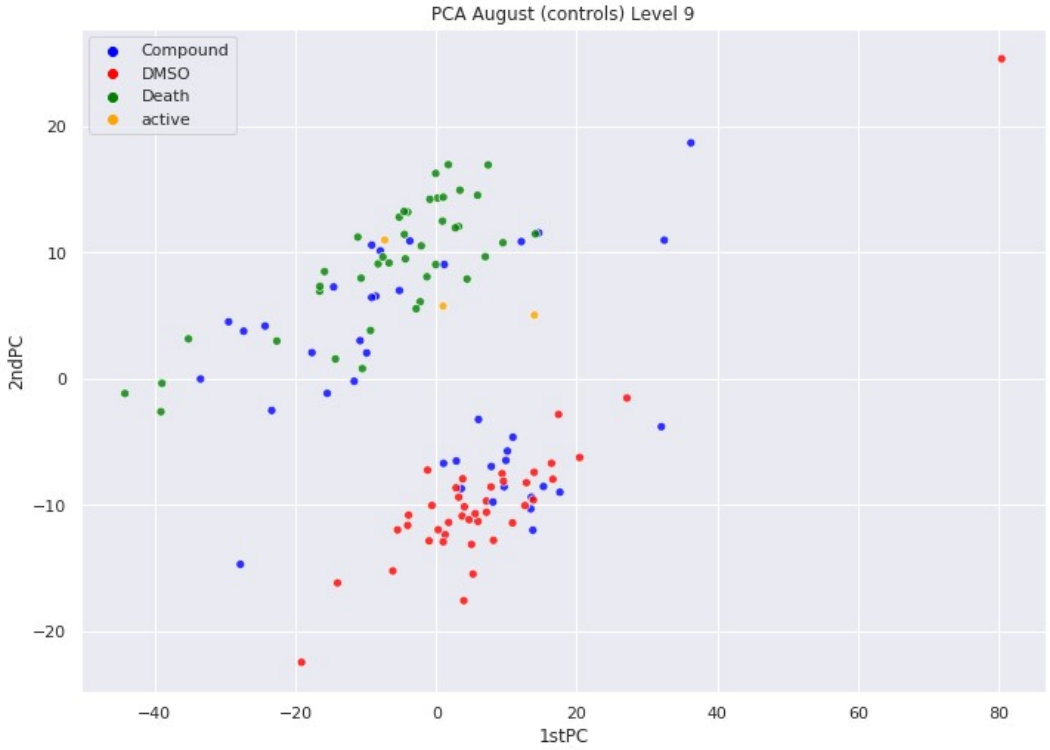
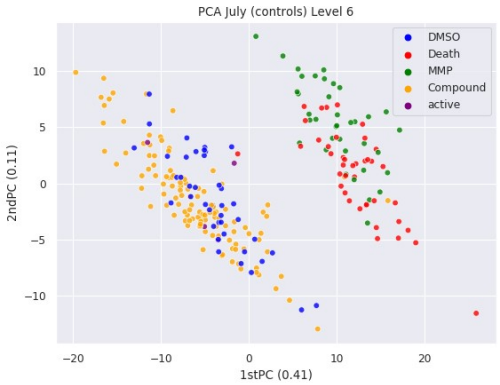
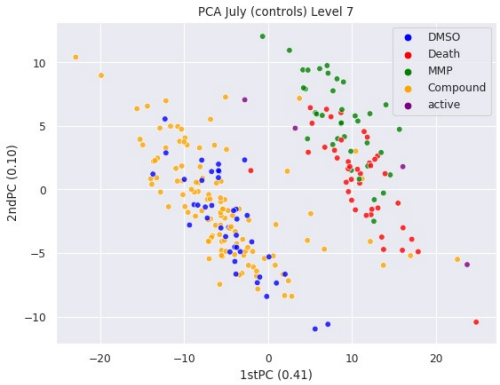


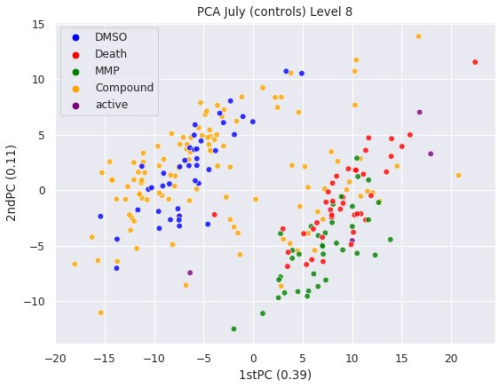
Figure A.3: Control Substances at HC for August



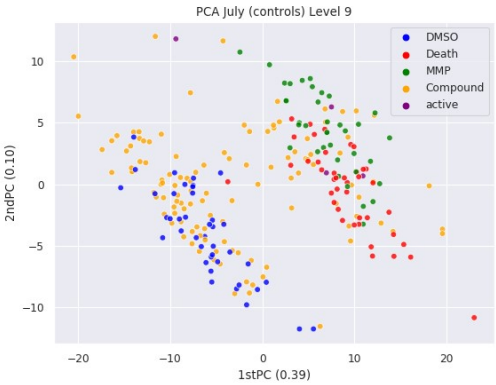
(a) Controls at Dose Level 6



(b) Controls at Dose Level 7



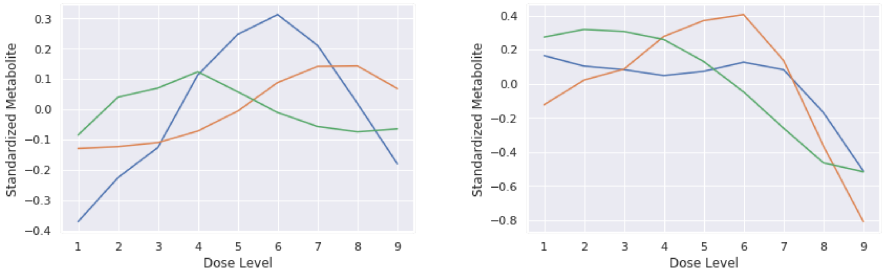
(c) Controls at Dose Level 8



(d) Controls at Dose Level 9

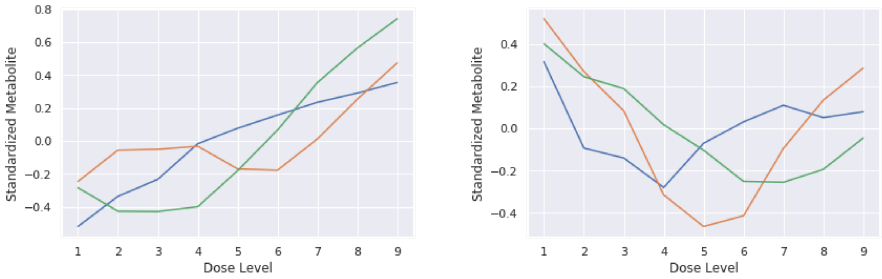
Figure A.4: July Different Dose Levels

A.1 Dynamic Time Warping



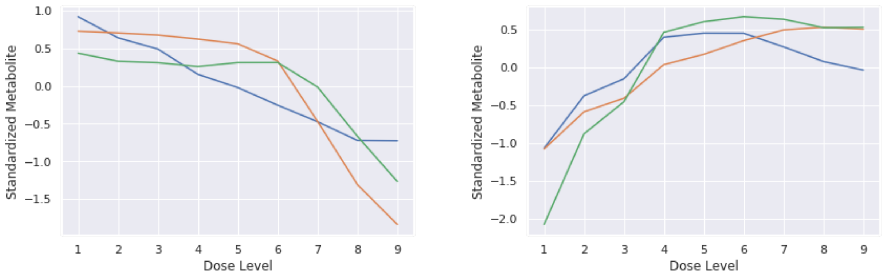
(a) Curves in Cluster 1

(b) Curves in Cluster 2



(c) Curves in Cluster 3

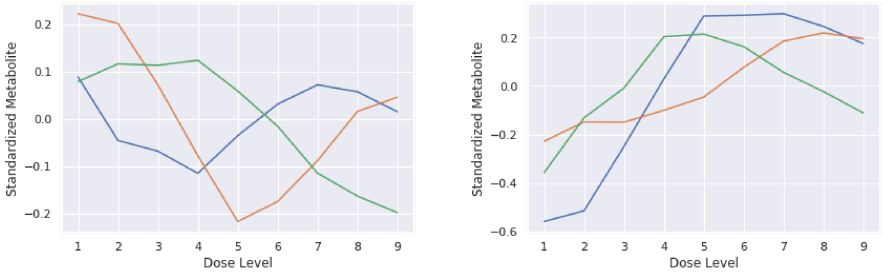
(d) Curves in Cluster 4



(e) Curves in Cluster 5

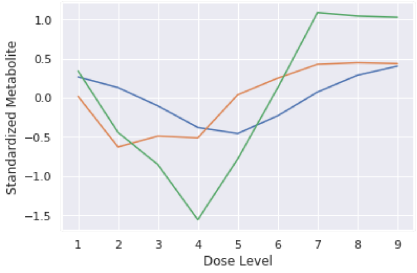
(f) Curves in Cluster 6

Figure A.5: Curve Shapes within DTW Clusters July

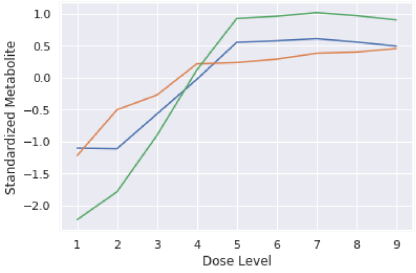


(a) Curves in Cluster 1

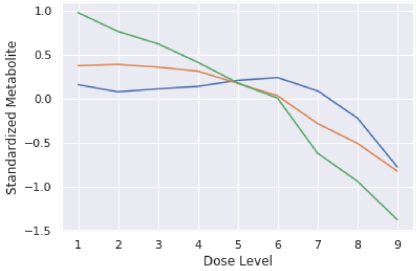
(b) Curves in Cluster 2



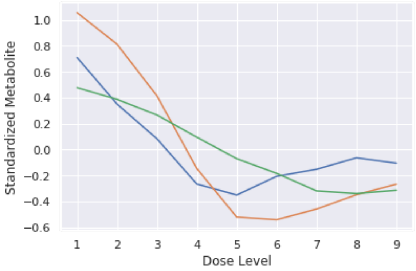
(c) Curves in Cluster 3



(d) Curves in Cluster 4



(e) Curves in Cluster 5



(f) Curves in Cluster 6

Figure A.6: Curve Shapes within DTW Clusters BC