UPPSALA
UNIVERSITET

UPTEC X 22030
Examensarbete 30 hp
Juli 2022

# Expanding the application of a novel proteomics tool for drug target and mechanism identification

Linnéa Yuan Andersson

Civilingenjörsprogrammet i molekylär bioteknik

UPPSALA UNIVERSITET

# Expanding the application of a novel proteomics tool for drug target and mechanism identification

Linnéa Yuan Andersson

## Abstract

In drug discovery and development, characterization of the drug targets and mechanisms of action is an essential step. ProTargetMiner is a publicly available proteome signature library of anticancer molecules and its automated bioinformatics platform can be used for drug target and mechanism deconvolution. The possibility of expanding ProTargetMiner to treatments that are non-anticancer is investigated in this project. A new proteome signature library was built for 15 versatile drugs with diverse indications, e.g. against allergies, hypertension, and depression. To comprehensively cover the proteome response to these treatments, deep expression profiling was performed in human fibroblast, breast cancer MCF7, and neuron-like SHSY5Y cells using multiplexed LC-MS/MS analysis at an optimized duration of 48h. Here, each collected proteome signature is contrasted against other signatures using OPLS-DA models to deconvolute drug targets, similar to the approach devised in the original ProTargetMiner platform. Furthermore, the drugs are further profiled by a validation technique called Proteome Integral Solubility Alteration (PISA) assay to identify the protein targets that are directly engaged by the molecules. Several known targets and mechanistic proteins are identified in the deep expression profiling experiment and are further verified by the PISA assay. Further testing and literature research could uncover novel targets for the treatments. This platform is expandable to novel drugs and provides a resource for target deconvolution of compounds in preclinical and clinical testing.

**Teknisk-naturvetenskapliga fakulteten**

**Uppsala universitet, Utgivningsort Uppsala/Visby**

Handledare: Roman A. Zubarev, Amir Ata Saei  Ämnesgranskare: Jonas Bergquist

Examinator: Siv Andersson

# Proteiners roll inom läkemedelsutveckling

Populärvetenskaplig sammanfattning

Linnéa Yuan Andersson

Proteiner kan ses som kroppens byggstenar och har en uppsjö av olika funktioner i kroppen, till exempel utgör de hormoner, enzym och bygger dessutom upp celler. Alla proteiner i en cell vid en viss tidpunkt kallas för proteomet. Genom att studera proteomet kan man dra slutsatser om vad som händer i cellen eller kroppen. Detta är den gren av biologin som kallas för proteomik.

Inom läkemedelsutveckling är det viktigt att ta reda på exakt hur ett läkemedel fungerar och vad det påverkar i kroppen. Proteomik är ett utmärkt verktyg för detta. Eftersom proteiner är involverade i viktiga processer i celler är de ofta målet för ett läkemedel. Genom att studera vilka proteiner ett läkemedel påverkar kan man även se vilka sorters processer i kroppen som påverkas, och på så sätt få mer förståelse för hur läkemedel egentligen fungerar för att t.ex. bota en sjukdom eller lindra symptom.

En metod man kan använda för studera proteomet kallas för masspektrometri (MS). Med MS kan man både identifiera vilka proteiner som finns i ett prov och kvantifiera antalet proteiner i ett prov. MS kan kvantifiera tusentals proteiner från flera olika prover och är därför väldigt användbart inom proteomik. Man kan alltså mäta proteomet i en cell efter att den blivit behandlad med ett läkemedel och sedan jämföra det med proteomet i en cell som inte blivit behandlad av ett läkemedel och se skillnaderna.

ProTargetMiner är ett bibliotek och verktyg som utvecklats av en forskargrupp på Karolinska Institutet. De studerade proteomen för flera celler då de behandlades med olika cancerläkemedel. Med hjälp av datormodeller jämförde de proteomet för ett läkemedel med proteomen för flera andra läkemedel och såg hur de olika proteinerna regleras i cellen och vilka som uttrycks i större och mindre utsträckning. De proteiner som påverkas mest undersöktes sedan med hjälp av protein databaser för att se vilka processer i cellen som de proteinerna mest sannolikt tillhör. De kunde sedan relatera de cellulära processerna tillbaka till det som redan är känt om läkemedlet, till exempel läkemedlets mekanismer eller biverkningar.

I detta projekt var målet att studera om samma metod som användes för att utveckla ProTargetMiner kan användas för andra läkemedel än cancerläkemedel. Femton olika läkemedel har undersökts i tre olika sorters celler. Dessa läkemedel inkluderar till exempel antiinflammatoriska läkemedel, läkemedel mot högt blodtryck, läkemedel för behandling av störningar i centrala nervsystemet och läkemedel mot allergi, mm.

Proteomen mättes med MS och data analyserades med hjälp av de ovannämnda datormodellerna. Flera kända målproteiner och mekanismer identifierades. Målproteinerna verifierades även med ett följdexperiment. Detta projekt visar alltså på att ProTargetMiner konceptet kan utvidgas till att även omfatta icke-cancerläkemedel. Detta koncept kan således vara värdefullt inom läkemedelsutveckling för att identifiera nya målprotein och mekanismer för läkemedel.

# Table of contents

# Abbreviations

FITExP - Functional Identification of Target by Expression Proteomics

LC-MS/MS - Liquid Chromatography coupled to tandem mass spectrometry

MOA - Mechanism of Action

MS - Mass spectrometry

MS/MS - Tandem mass spectrometry

MTX - Methotrexate

MRP - Mitochondrial Ribosomal Proteins

OPLS-DA - Orthogonal Partial Least Square-Discriminant Analysis

PISA - Proteome Integral Solubility Alteration

PCA - Principal component analysis

PC - Principal component

TPP - Thermal Proteome Profiling

TMT - Tandem Mass Tag

# 1. Introduction

In drug discovery and development, an essential but challenging task is to identify targets, cellular effects, and mechanisms of action (MOA) of compounds. Despite increasing investment in biomedical research and drug development, the approval of new drugs (new molecular entities) has not seen a similar trend, with the FDA approving around 30 new drugs annually in the past few decades, which has since slightly increased to around 40 in the past decade (Mullard 2022). However, only a small fraction of these drugs are targeting novel targets. In 2020, only 13 of the 61 novel drugs approved in the United States, European Union, and Japan had novel MOA (Avram *et al.* 2021). To facilitate the consideration of novel drug targets, an emphasis has been made on drug target identification and validation in the drug discovery pipeline (Lindsay 2003).

## 1.1 Proteomics as a tool for identifying drug targets

The proteome can be defined as the set of proteins that carry out their functions at specific times and locations in the cell (Aebersold & Mann 2016). The large-scale study of the proteome and the proteins' cellular functions is known as proteomics. Proteomics is a useful tool for addressing the challenges surrounding drug target identification as well as identifying MOA (Saei *et al.* 2019). Since proteins are common targets of drugs, the abundance of different proteins can give valuable information as to compounds' effects on the cell. Both the expression and degradation of proteins affect their abundances and can be determined uniquely by investigating the proteome.

### 1.1.1 Mass spectrometry

Mass spectrometry (MS) is a method that is widely used to study the proteome. The method has been greatly successful in proteomics due to its effectiveness in both identifying and quantifying proteins with high accuracy and sensitivity (Aebersold & Mann 2016). The most common MS workflow, called bottom-up proteomics, uses enzymatic digestion of the extracted proteins resulting in peptides which are then analyzed (Figure 1). The digestion is performed by a sequence-specific enzyme, commonly Trypsin. The peptides are then separated by chromatography and ionized. A spectrum of the peptide ions (MS1 level) is acquired. The peptide ions are then, in gas phase, fragmented in the mass spectrometer and generate a second spectrum (MS2 level). This method, generating two levels of spectra, is known as tandem mass spectrometry or MS/MS and is usually coupled with liquid chromatography (LC-MS/MS), which separates the proteins before ionization and analysis.

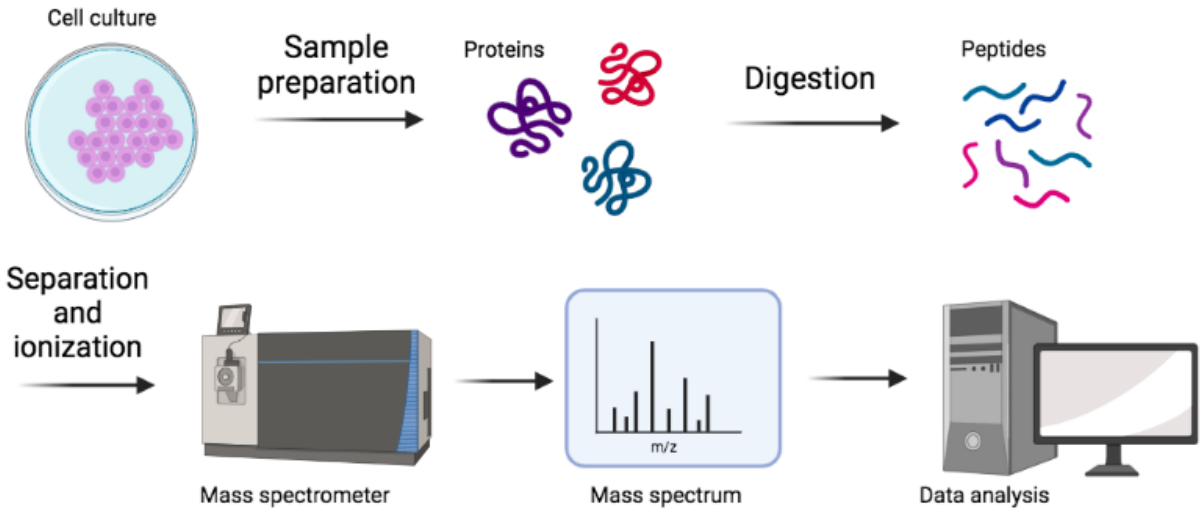


**Figure 1.** Workflow of a typical bottom-up proteomics experiment.

A mass spectrum shows the mass to charge ratio (m/z) intensities. The intensity of the signals in the MS1 level spectrum reflects the number of detected ions and can be used to quantify the peptides with quantitative proteomics software such as MaxQuant (Cox & Mann 2008, Pappireddi *et al.* 2019). The MS2 level spectrum showing the fragmented peptide ions can in turn be used to identify the amino acid sequence of the peptide as well as post-translational modifications. The masses determined in MS1 and the MS2 spectra can then be compared to theoretical spectra for known peptides in order to identify the peptides quantified in the sample (Pappireddi *et al.* 2019).

To quantify thousands of proteins from multiple samples, multiplexed proteomics can be used. Multiplexing entails that samples are labeled with isobaric tags which are distinguishable in MS2 level spectra (Pappireddi *et al.* 2019). The most common isobaric tags used are Tandem Mass Tags (TMT). In a typical workflow, the tags are used to label the peptides after digestion, after which the uniquely barcoded samples are pooled together (Figure 2). Isobaric tags have an identical total mass but a varying heavy isotope distribution in the tag. The tags contain a site that is fragmented in the MS2 spectrum, which results in reporter ions that have different masses. The tags can therefore be used to differentiate between identical peptides originating from different samples.

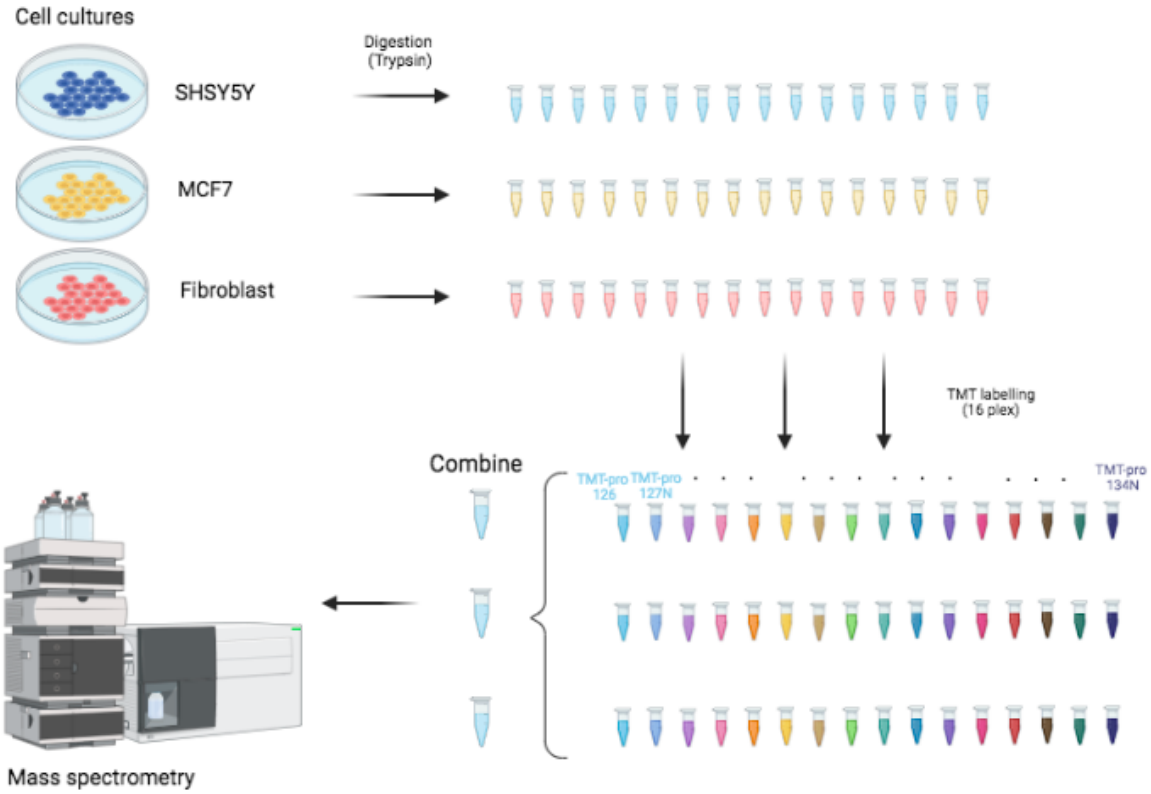


**Figure 2.** TMTpro16 multiplexing workflow for three cell lines.

### 1.1.2 Functional Identification of Target by Expression Proteomics (FITExP) and ProTargetMiner

Zubarev *et al.* previously observed that the target of the anticancer drug 5-FU, the protein TYMS, was significantly up-regulated in response to 5-FU treatment in RKO cells, particularly in late cell apoptosis. This raised the question if a target could be deduced from the proteomics data by sorting the proteins based on their regulation. However, the proteins involved in cell death were also highly regulated along with the target. In an attempt to filter out the proteins involved in cell death and highlight the target, they treated the cells with other drugs and filtered out the proteins that were reoccurring. This was unsuccessful and the target, TYMS, was still not found as a likely target among the top proteins. Following this, they added more specificity by treating two additional cell lines under the assumption that the drug should behave consistently regardless of the cell line used, while the proteins involved in cell death might be cell line specific. The added specificity successfully identified TYMS as a target.

The method developed from this observation was named Functional Identification of Target by Expression Proteomics (FITExP) (Chernobrovkin *et al.* 2015). FITExP allows for the identification of anticancer drug targets and mechanisms without the need for a chemical modification of the drug or prior knowledge of the MOA. The method also involves the

calculation of three measures (regulation, specificity and exceptionality) for every cell line, protein, and drug. The specificity parameter is obtained by normalizing the regulation for a given treatment by the average of the other treatments, which reflects the protein regulation of the compound of interest compared to the other compounds. Using specificity in combination with the other measures, proteins can be ranked along with their p-values, resulting in the identification of the drug target and MOAs.

In 2019, the FITExP concept was used to create ProTargetMiner, a publicly available proteome signature library of anticancer molecules (Saei *et al.* 2019). For ProTargetMiner, 56 compounds were profiled which resulted in an expandable library consisting of 7328 proteins and 1,307,859 protein-drug pairs. They showed that contrasting the proteome of a given anticancer drug against the proteome of many other drugs can highlight a given drug's target and mechanistic proteins. Additionally, they showed that the contrasting panel for a single cell line can be downscaled to 8 compounds and still successfully identify targets. Using this miniaturized method, 9 molecules representing the most diverse mechanisms were used to generate deep datasets in three cell lines. The combination of the three datasets revealed common targets and MOAs. Important cell-specific mechanisms could also be identified when investigating the differences between the datasets of individual cell lines. The ProTargetMiner concept involved using Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA) models, with which the proteome signature of a given compound can be contrasted against others, highlighting the likely targets of the compound under study. The automated ProTargetMiner platform can be employed for new compounds by submitting the fold changes of proteins for a given compound (and a number of replicates) in one or more specified cell lines. The output of such a query is an interactive PLS-DA model that provides the most likely drug target and shows protein rankings.

### 1.1.3 Orthogonal Partial Least Square Discriminant Analysis

OPLS-DA is a multivariate classification method that can determine discriminatory properties between two classes using supervised models (Bylesjö *et al.* 2006). In the case of ProTargetMiner where the variables are proteins, OPLS-DA determines the proteins with the largest discriminatory power between a given proteome signature and others in the library. The results can be visualized on a loading plot where the x-axis displays the predictive component which demonstrates the variations between the groups while the y-axis demonstrates the variations within the groups (Saei *et al.* 2019). For the purpose of drug target identification, only the x-axis is of importance.

Each point on the loading plot represents a protein. The proteins that are specifically up-regulated will be displayed on the right side of the plot, while the specifically down-regulated proteins will be displayed on the left side. The further the proteins are to the endpoints of the x-axis on either side of the plot, the greater the specificity and regulation of the protein in response to treatment of the contrasted drug.

R2 and Q2 values are used to characterize each OPLS-DA model. The R2 value represents the fit of the data to the model while Q2 is a measure of the model's predictive power. A model with an R2 value of 1 would perfectly describe the data and a Q2 value of 1 would indicate a perfect predictivity of the model (Saei *et al.* 2019).

### 1.1.4 Thermal Protein Profiling (TPP) & Proteome Integral Solubility Alteration (PISA)

Drugs and other cellular agents, as well as non-molecular influences such as radiation, can influence the physicochemical properties of proteins. These changes can be investigated by applying a stability- and solubility-modifying factor such as temperature changes (Gaetani *et al.* 2019). Variation in the thermal stability of proteins can be used to study ligand binding. Thermal Protein Profiling (TPP) combines Cellular Thermal Shift Assay (CETSA) with multiplexed quantitative MS for proteome-wide monitoring of drug target engagement with a given small molecule. CETSA enables the monitoring of target engagements in living cells, and in combination with quantitative MS is used in TPP to monitor the thermal stability of proteins in different states (such as under drug treatment) and identify markers of target engagement and drug efficacy (Savitski *et al.* 2014). In TPP, after incubation of living cells or cell extracts with small molecules, the proteins are incubated at different temperature points, and at each point, quantitative proteomics is used to measure relative protein abundances. The relative abundances are then used to fit a curve and calculate the specific melting temperatures ($T_m$) for each protein. For cells treated with a drug this melting curve shifts for the target proteins, and the difference in melting temperatures ($\Delta T_m$) is calculated. To increase specificity, concentration, C, is added as a second dimension. By examining the isothermal $\Delta T_m$ as a function of C, the drug concentrations needed to induce half of $\Delta T_m$ can be determined ($C_0$). Then, the proteins with the largest absolute value of $\Delta T_m$ and the lowest value of $C_0$ can indicate the most likely target of the drug. However, this approach has a low throughput and is resource-consuming. This forces researchers to perform such experiments with a limited number of replicates which can limit statistics. Also, the results partially depend on the quality of curve fitting.

In 2019, Gaetani *et al.* developed Proteome Integral Solubility Alteration (PISA) as a high throughput and resource-frugal method to overcome the previously mentioned issues. In the temperature-based approach of PISA, samples are collected from individual temperature points. However, unlike TPP, instead of labeling the trypsin digest of each sample the samples within a replicate are pooled together (Gaetani *et al.* 2019). This pooled sample is analyzed using MS. Instead of using the resulting intensities to fit a curve and extract the $T_m$ like in TPP, PISA measures the protein abundance ($S_m$) in the pooled sample and the abundance represents the area under the melting curve. If $S_m$ is the protein abundance in the untreated sample and the $S_m'$ is the protein abundance in the treated sample then $F_t(S_m, S_m')$ is the PISA T equivalence of $\Delta T_m$ in

TPP. Combining $F_t$ with the p-values of the proteins can result in a volcano plot where the best candidate targets can be identified. To add another dimension a third sample is analyzed using intermediate drug concentrations where the resulting abundance $S_m''$ represents the integral of the concentration-dependent curve. The resulting $F_t(S_m, S_m', S_m'')$ is the equivalent of TPPs $C_0$.

## 1.2 Expanding the application of ProTargetMiner

In the case of both FITExP and ProTargetMiner, the cells are treated with anticancer agents at a concentration of LC50, at which 50% of the cells die after 48h. This allows the cell states to be more relevant and comparable. However, it is also interesting to expand this tool and study whether it can be applicable for non-lethal treatments as well. Non-lethal treatments can be obtained when using lethal drugs at a lower concentration than IC50 or when normal nutrients and metabolites are used at concentrations that do not suppress cell growth. Previously, it was observed that the targets for lethal compounds were already elevated at concentrations lower than IC50, suggesting that the proposed approach could succeed. However, the non-lethal drugs cannot be compared based on a single common phenotype when using a sub-IC50 concentration. Therefore, the concentrations and treatment durations were optimized in a pilot experiment to find the concentrations that allow for comparison.

For the pilot experiment, the FITExP and ProTargetMiner methodologies were investigated to see if they could be applied to non-lethal drugs to identify drug targets in the main experiment. Furthermore, the time duration that showed the best performance for target deconvolution was investigated and optimized. The cells were treated with Methotrexate (MTX), Atorvastatin, and Celecoxib for 24h, 48h, 72h, and 96h, after which the cell proteomes were digested and multiplexed with TMT and analyzed. The TMT sets can be found in Appendix A.

For the main experiment, 15 different compounds have been chosen based on the versatility of their targets, MOA, and therapeutic indications. Among these 15 drugs, the anticancer drug MTX was used as a control, and LDC203974 was added as a novel anticancer agent for benchmarking. The drug information obtained from DrugBank (Wishart *et al.* 2018), as well as the TMT setup of both the main experiment and pilot experiment, can be found in Appendix A. The therapeutic categories of these drugs include anti-inflammatory drugs, and those for hypertension, central nervous system, allergy, hyperlipidemia, etc. The deep proteome signatures of these drugs have been acquired in human foreskin fibroblasts, neuroblastoma SH-SY5Y, and breast cancer MCF-7 cells, which represent diverse tissues.

Both for the pilot experiment and main experiments, cell viability was measured in response to the drugs. In case the drugs exerted cytotoxicity, the IC50 value was used (Appendix A), and in cases where the drug was not cytotoxic, a fixed concentration of 25 μM was used for the experiments.

This bioinformatic project aims to analyze the data of the pilot experiment in order to determine which treatment duration should be used in the main experiment. The data from the main experiment will also be analyzed to determine whether the same method as seen in ProTargetMiner can be used with non-lethal treatments to identify drug targets and MOA (Figure 3). The data from the main experiment will also be used to determine a suitable cell line for a validation experiment, performed using PISA. This data will also be analyzed to see if any potential novel targets discovered can be validated.

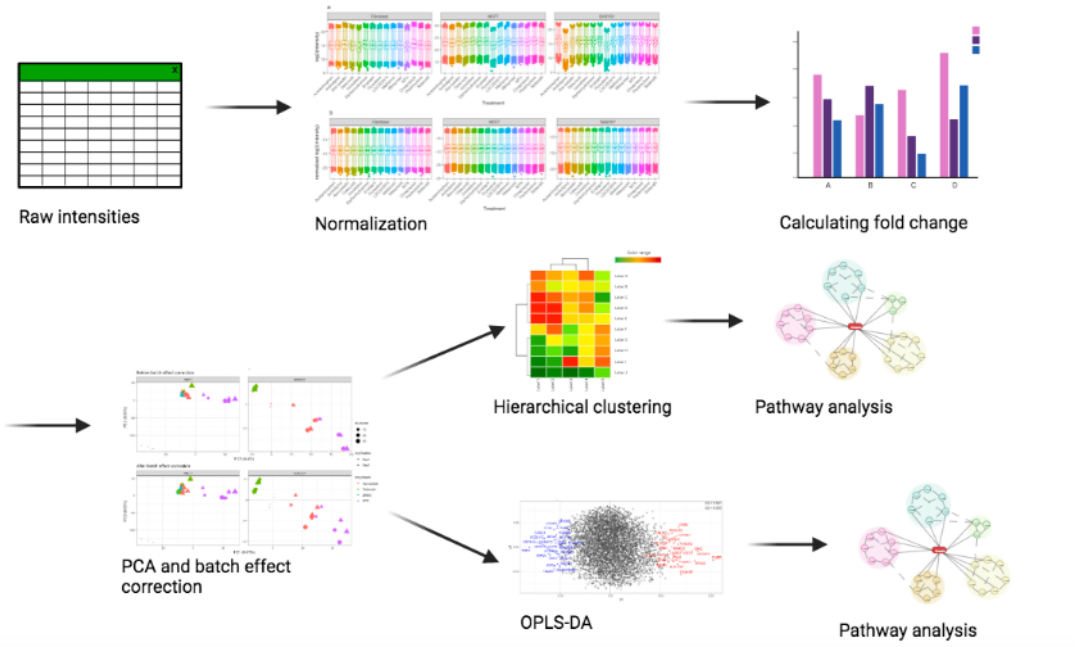


**Figure 3.** General overview of the bioinformatic workflow for analysis of the main experiment data.

# 2. Materials and methods

Raw data from the experiments were produced using LC-MS/MS. Protein quantification and identification were performed using MaxQuant. The resulting intensities for the proteins were filtered by removing contaminants and any proteins quantified with less than two peptides. The raw intensities for each protein were normalized by the total intensities of each channel (sample) and fold changes were calculated as the ratio of the intensities of the treatment and the control and then log2-transformed. PCAs were performed on the fold changes as a means of quality control. Batch effect correction was performed using the R-package Limma. P-values were calculated using two-sided Student's t-test on the log2-transformed and normalized intensities. This was done on both the pilot, main dataset and validation dataset.

On the main data, the R-package Ropls was used to generate Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA) models where the proteome signature of each treatment was contrasted to those of the rest of the treatments. This resulted in specificity values for each protein, where the most and least specific protein targets were selected for further analysis.

Also on the main data, Gene ontology enrichment analyses were performed using the online tool GOrilla (Eden *et al.* 2007, Eden *et al.* 2009) where all quantified proteins were used as background. The analysis was performed on hierarchical clusters generated with the k-means algorithm (with a repeat of 100) as well as the targets identified using the OPLS-DA models. The protein targets were investigated using UniProt and existing literature.

# 3. Results

## 3.1 Pilot experiment

After preprocessing the pilot experiment data, a Principal Component Analysis (PCA) was performed before and after batch effect correction (Appendix Figure B1). The separation showed little change after batch effect removal suggesting that initially, there was no strong batch effect. There was a clear separation with the first principal component (PC) separating based on treatment. It also indicated that the batch effect was removed. However, there seemed to be three outliers in replicate two which needed further investigation.

The pilot data distribution (Appendix Figure B2) showed that replicate two in MCF7 cells with all three treatments were behaving unexpectedly. This did not seem to improve with batch effect removal and double-checking the normalization confirmed that it was not the root of the issue. Since there were only two replicates, it is difficult to conclude what the cause could be. Regardless, further analyses were performed.

In order to confirm which of the treatment durations were most suitable for the experiments, the expression of the expected targets for the treatments was investigated at each treatment duration. DHFR, which is the cognate target for MTX, showed an increased expression in MTX-treated SHSY5Y cells at all time points (Figure 4a). Additionally, the expected target for Atorvastatin, HMGCR, showed an increased expression at all time points compared to the other durations in Atorvastatin-treated MCF7 cells (Figure 4b). We concluded that the ProTargetMiner concept can be potentially generalized to non-anti-cancer treatments. Furthermore, we concluded that similar to the original FITExP and ProTargetMiner implementations, 48h would be the most suitable time of treatment based on our pilot experiment. At 48h, the regulation of drug targets is generally higher than 24h.

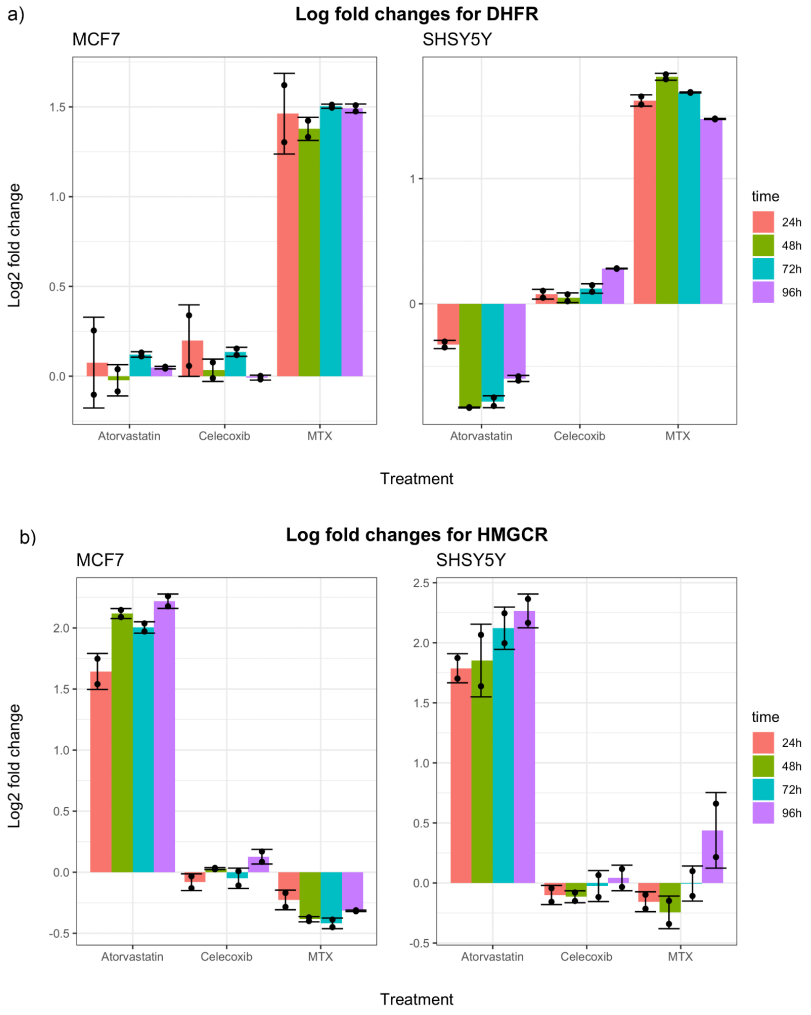


**Figure 4.** a) Regulation of DHFR in cells treated with the Atorvastatin, Celecoxib, and MTX for four different durations (24h, 48h, 72h, 96h). b) Regulation of HMGCR in cells treated with the same treatments and durations stated in a).

Furthermore, the amount of significantly expressed proteins (with a p-value < 0.05 and a log2 fold change of > 0.5 or < -0.5) was observed for each of the treatments in the two cell lines using the different durations. For instance, MCF7 cells treated with MTX for 48h (Figure 5) showed a significant increase in the number of proteins that fulfilled the conditions in comparison to a treatment duration of 24h. While 72h and 98h also showed an increase, they proved to be less practical when performing the experiments. Ultimately, 48h proved to be the most suitable duration for the main experiment.

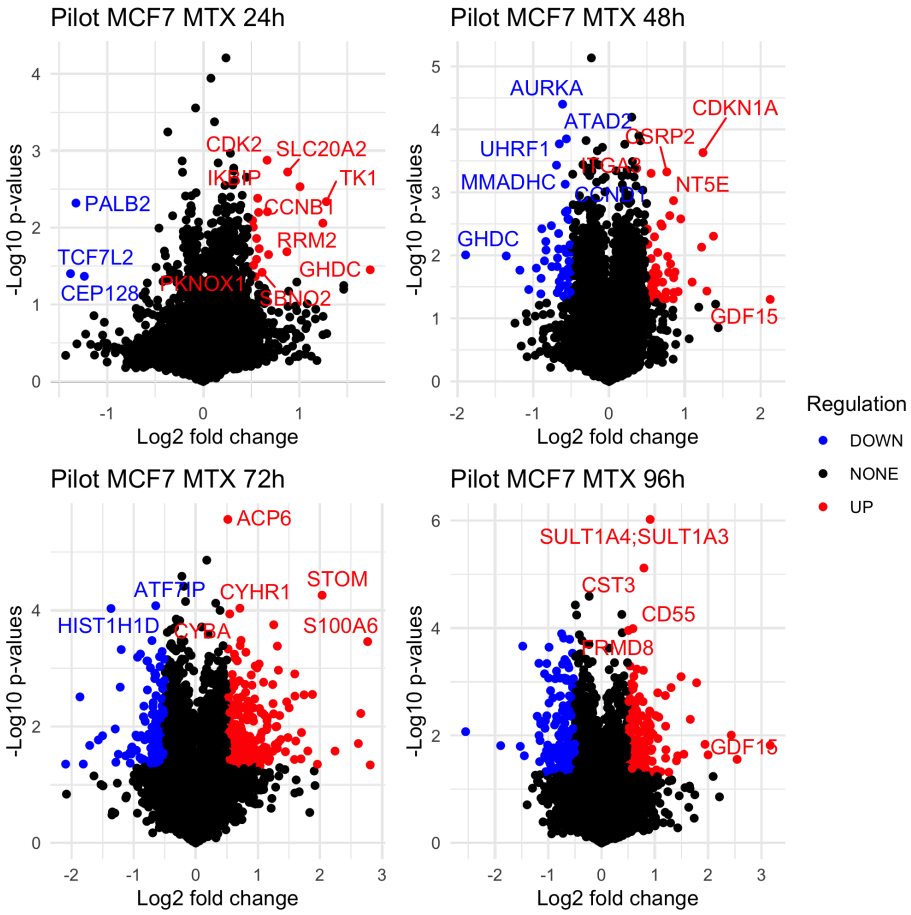


**Figure 5.** Volcano plot of cells treated with MTX for different durations. The red points were proteins that had a log2 fold change > 0.5 and a p-value < 0.05. The blue points were proteins that had a log2 fold change < -0.5 and a p-value < 0.05. The black points were the remaining quantified proteins that did not fulfill the above conditions.

## 3.2 Main experiment

### 3.2.1 Preprocessing and quality control

After the data were filtered, a normalization of the intensities resulted in a median stabilization (Appendix C). Calculation of the log2 fold changes compared to the control, and the mean log2 fold changes across each replicate showed a ranking of each drug based on their effect on the proteome (Appendix D).

After preprocessing the data, PCA was performed before and after batch effect correction, showing the absence of a significant batch effect (Appendix E). The PCA showed a separation of the treatments Fluoxetine and Amlodipine from the rest of the treatments on PC1. Overall, the PCA before and after batch effect correction did not show a significant change and there did not seem to be a significant batch effect in the data. The data was deemed viable for further analyses.

### 3.2.2 Hierarchical clustering

A heatmap of the mean log2 fold changes of each replicate was produced along with the hierarchical clustering of the proteins on the vertical axis (Figure 6a). As seen in the PCA earlier, the largest fold changes were shown for Amlodipine and Fluoxetine in SHSY5Y and Fluoxetine in MCF7 seen to the very left in the heatmap. While the clustering of the drugs on the horizontal axis is based on the cell line in which they were used.

Pathway analysis of the hierarchical clusters showed the biological processes representing each cluster and their GOrilla enrichment score. The top three biological processes of each cluster were chosen from the analysis for further investigation (Figure 6b). There were notable pathways found that can be related to the drug targets and MOA such as the cholesterol-related processes found for cluster 16 and mitochondrial translation for cluster 6.

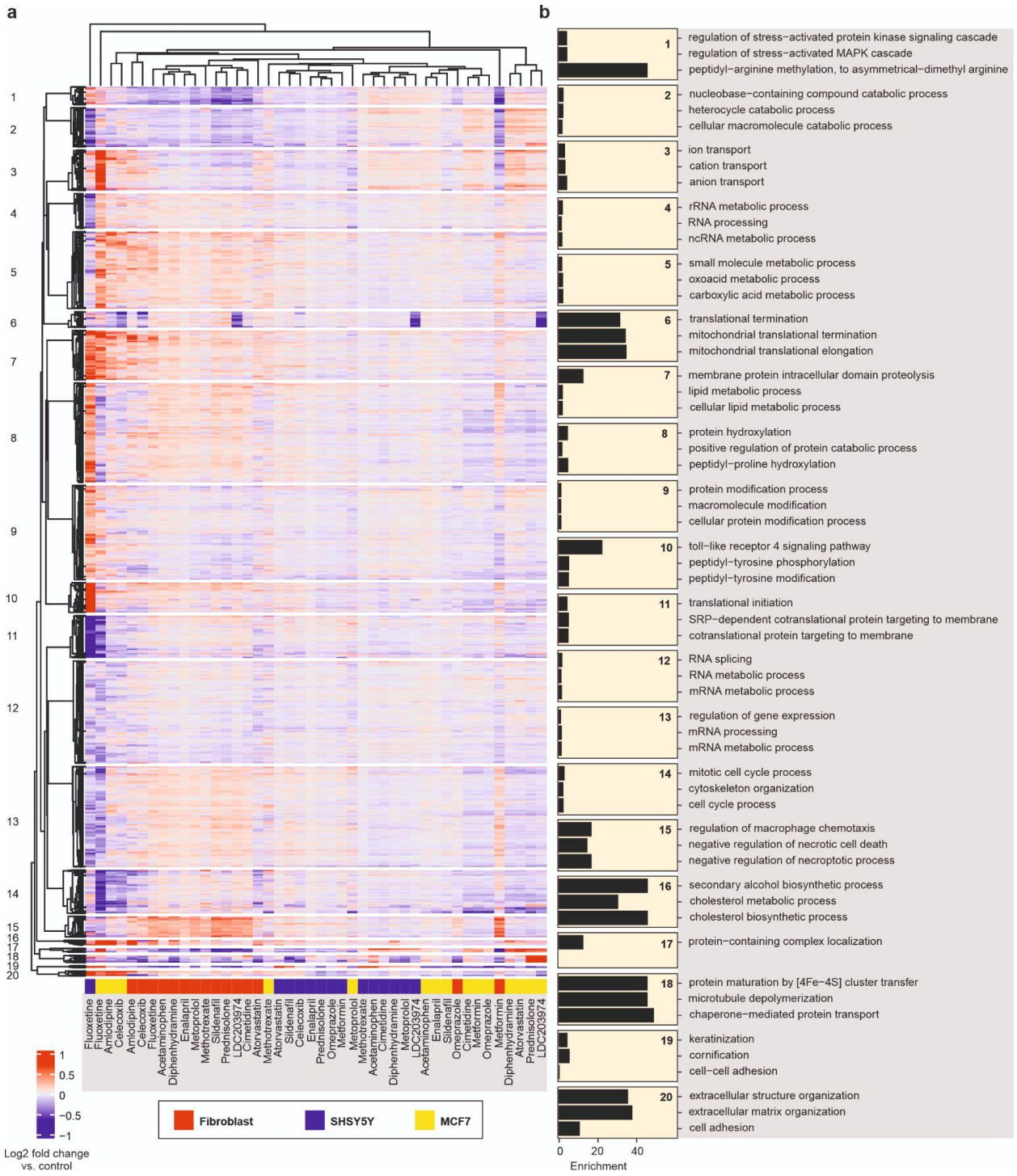


**Figure 6.** a) Hierarchical clustering of the main data. Performed on the mean log2 fold changes in each cell line. b) The results of a GO enrichment analysis on the hierarchical clusters. The top 3 biological processes were selected from the results of each cluster. All cluster enrichment scores were scaled up 10x except for cluster 16.

### 3.2.3 Orthogonal Partial Least Square Discriminant Analysis (OPLS-DA)

Targets for each of the treatments were also investigated using OPLS-DA models (Figure 7a-e). These OPLS-DA models were built on the main experiment dataset.

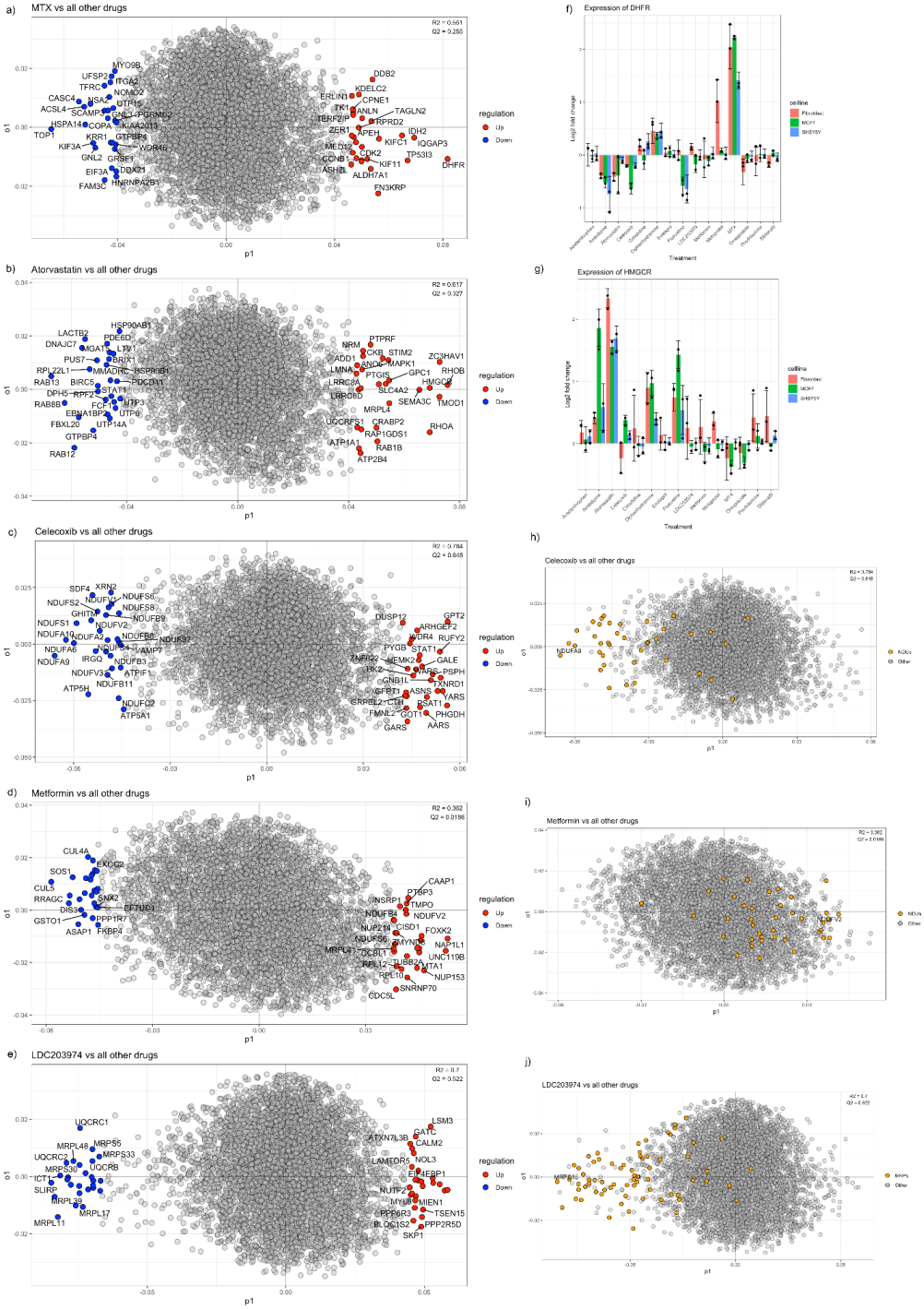


**Figure 7.** a-e) OPLS-DA models of five of the treatments (MTX, Atorvastatin, Celecoxib, Metformin, LDC203974). f-g) Log2 fold change of DHFR and HMGCR for each treatment vs control. h-j) OPLS-DA models for Celecoxib, Metformin, and LDC203974 only show specific protein groups labeled (NDU and MRP proteins).

The OPLS-DA models always show the compound of interest on the right-hand side, while the rest of the proteome signatures are on the left-hand side. So the proteins that are least specific to the treatment (therefore downregulated in response to the treatment) are to the very left in the plot, and the upregulated proteins are to the very right. The model for MTX showed a clear upregulation of the expected target DHFR (Figure 7a). When looking at the specific expression for DHFR it was also clear that treatment with MTX resulted in a significant upregulation in all cell lines compared to other treatments (Figure 7f).

The model for Atorvastatin also showed an upregulation of the expected target, HMGCR (Figure 7b). Plotting the expression of HMGCR specifically also revealed that in comparison to other treatments, Atorvastatin resulted in a significant upregulation of HMGCR (Figure 7g). However, it was also interesting to see that Amlodipine also resulted in an upregulation of HMGCR in the MCF7 cell line.

In the case of Celecoxib, many NADH:Ubiquinone Oxidoreductase (NDU) proteins were shown as down-regulated in the OPLS-DA model (Figure 7c). Metformin on the other hand showed an upregulation of NDU proteins (Figure 7d), while LDC203974 showed downregulation of Mitochondrial Ribosomal Proteins (MRPs), which is in line with its effect on mitochondria (Figure 7f) (Bonekamp *et al.* 2020).

## 3.2.4 Pathway analysis

A pathway analysis showed the top three biological processes for the top 21 up-regulated and top 21 down-regulated proteins of each treatments OPLS-DA model built on the main experiment dataset (Figure 8). Many of these pathways are in line with the known biological effects of these molecules. For example, Amlodipine shows an enrichment in the cholesterol metabolic process while LDC203974 shows an enrichment in mitochondrial translation.


**Figure 8.** Top 3 biological processes for the top 21 up and down-regulated proteins of each treatment from the main experiment dataset.

## 3.2.5 Merging the MCF7 data with data from ProTargetMiner

The MCF7 data from the main experiment was merged with the MCF7 data from the 9 anticancer drugs used in ProTargetMiner, to investigate if the inclusion of more drugs can add to

the specificity in target selection. For example, in the Prednisolone OPLSA-DA model from the merged data, we found MTPN or myotrophin as a down-regulated protein. This protein might be involved in the anti-inflammatory effects of Prednisolone, as MTPN regulates NF-kappa-B transcription factor activity. Building an OPLS-DA model contrasting Prednisolone against the 14 non-lethal drugs in the current study and the 9 anticancer drugs in ProTargetMiner, resulted in higher rankings for this mechanistic protein as shown in Figure 9.

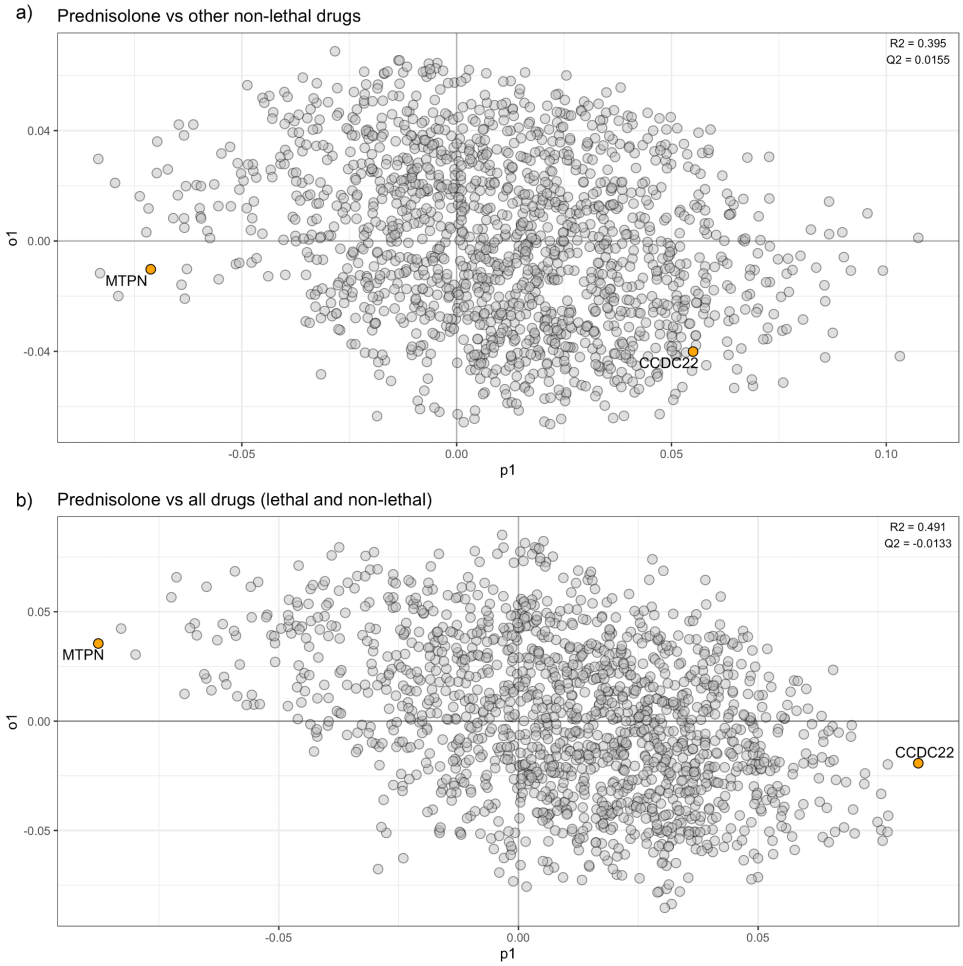


**Figure 9.** Prednisolone OPLS-DA models based on the combined dataset and lethal treatment dataset where prednisolone is on the right-hand side and the other proteome signatures are on the left-hand side.

# 3.3 Validation experiment

## 3.3.1 Choice of cell line

In order to determine which cell line should be used in the validation experiment, some criteria were set such as the number of quantified proteins in the given cell line as well as the number of proteins with differential regulation vs. control: log2 fold change > 0.5 or <-0.5 and a p-value <

0.05. The number of significantly up and down-regulated proteins for each treatment was used to generate a stacked bar plot (Figure 10a), and a box plot showing the median and spread of the number of significant proteins was also generated (Figure 10b). MCF7 had the highest median number of significant proteins, which made it a better candidate for the follow-up experiment. As shown previously, MCF7 also showed a higher expression specifically for targets such as DHFR when treated with MTX and HMGCR when treated with Atorvastatin and Amlodipine. As such, MCF7 was selected as opposed to Fibroblasts and SHSY5Y for the validation experiments.

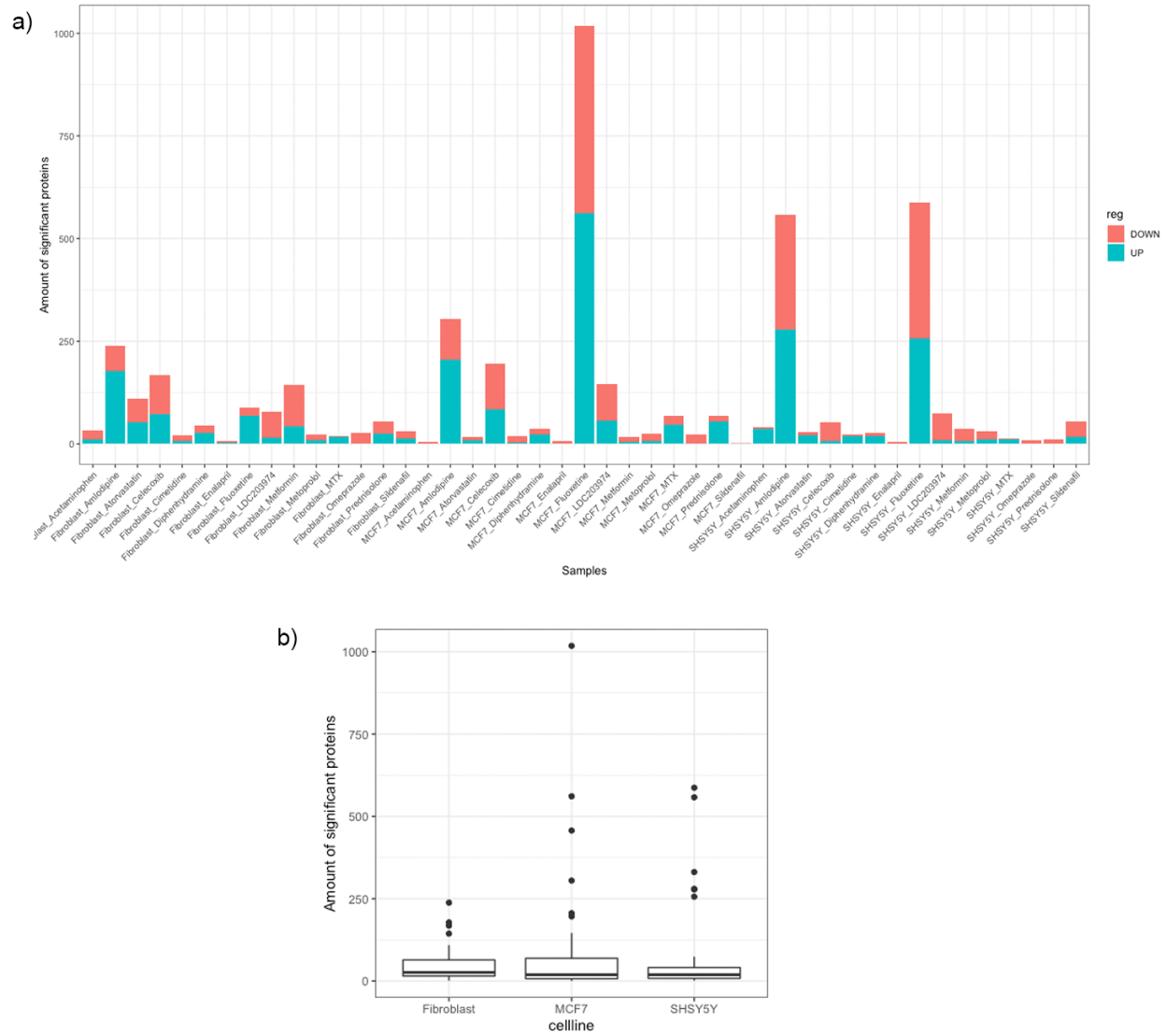


**Figure 10.** a) Stacked bar plot showing the number of significant proteins from the main dataset (for each cell line and treatment) that fulfill the two conditions: log2 fold change > 0.5 or < -0.5 and a p-value. < 0.05 b) Box plot for each cell line that shows the distribution of the number of significant proteins for each treatment.

## 3.3.2 Significant proteins

The calculated p-values and log2 fold changes of the validation experiment data were filtered based on protein IDs with the MCF7 data from the main experiment so that the dataset only contained the proteins found in both datasets. This was then used to create volcano plots for each treatment (Figure 11).

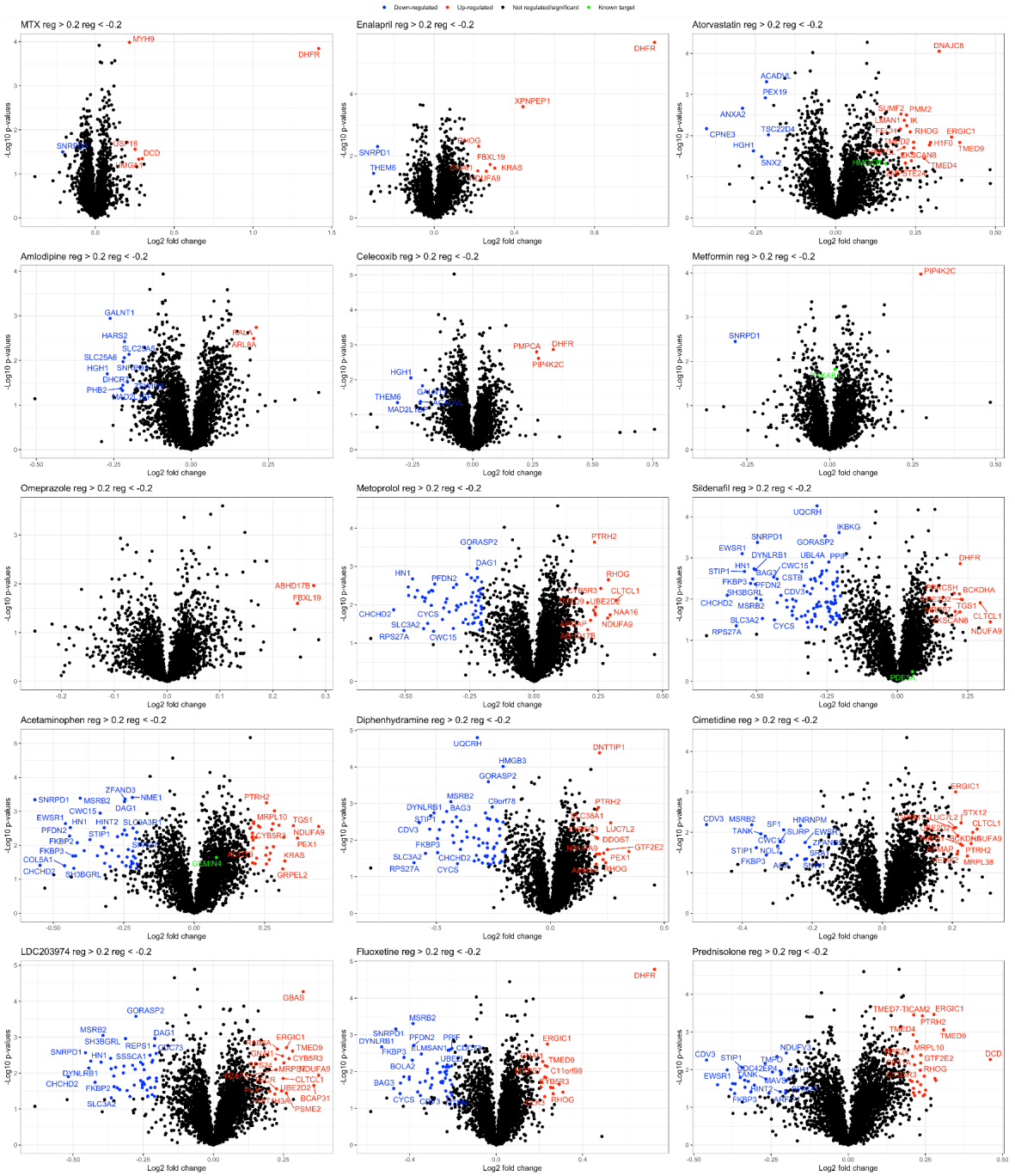


**Figure 11.** Volcano plots from the validation experiment data. The proteins shown in red are significant and up-regulated; proteins in blue are significant and down-regulated; proteins shown in green are known targets of the drug and are labeled even if they don't fulfill the conditions of being significant.

The resulting volcano plots successfully showed DHFR as an up-regulated target of MTX. HMGCR was also slightly up-regulated for Atorvastatin and passed the significance threshold ($p < 0.05$).

### 3.3.3 PISA vs Expression

The expression data (log2 fold changes) of MCF7 cells generated in the main experiment was plotted against the log2 fold changes from the PISA validation experiment to see which potential targets were found in both data sets (Figure 12).

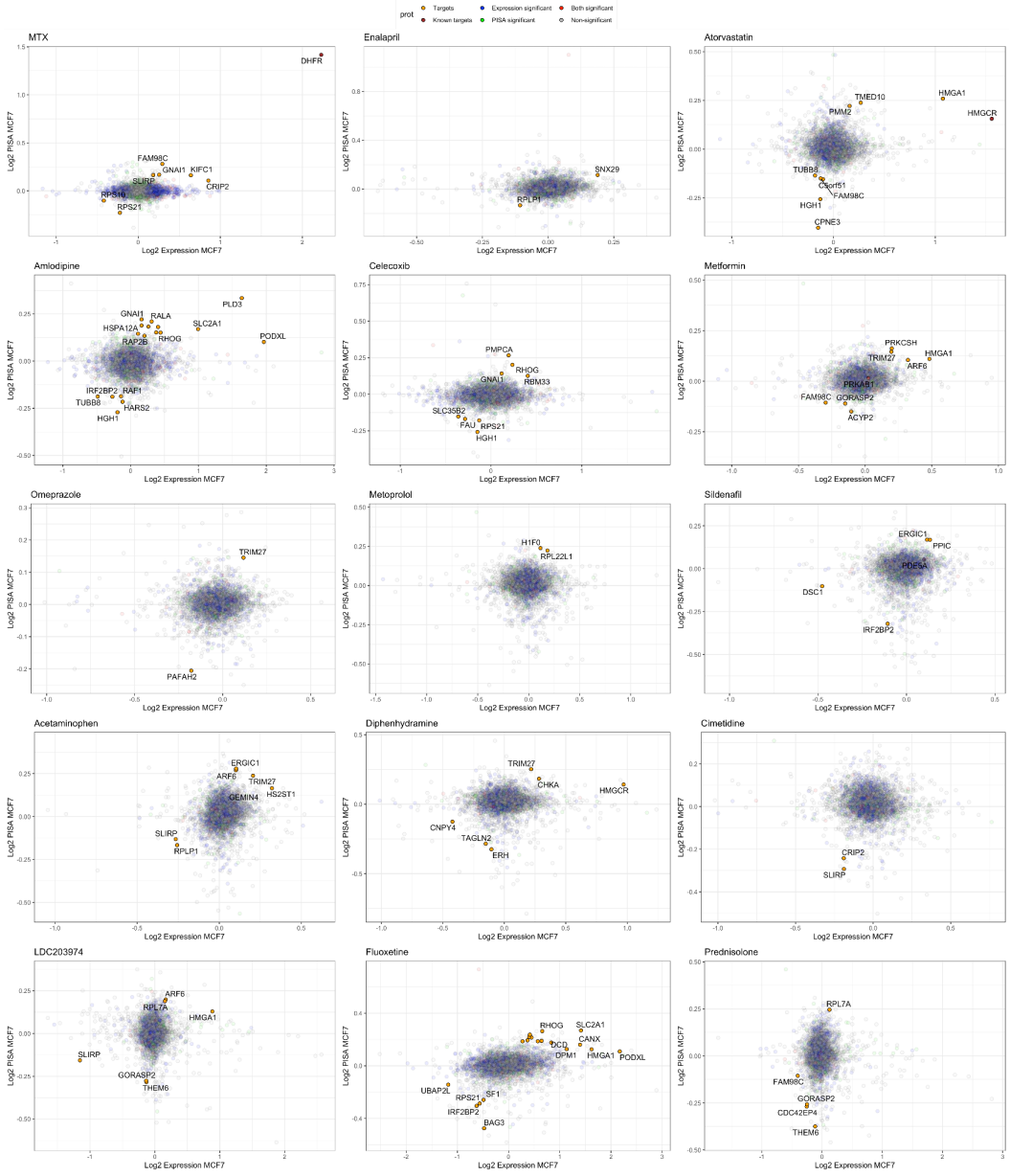


**Figure 12.** Scatterplots showing protein regulation in expression profiling vs. Log2 fold changes of proteins in the PISA validation experiments in MCF-7 cells. Blue points are proteins that are statistically

significant (p-value < 0.05) in the expression data; green points are statistically significant proteins in the PISA data; red points are statistically significant in both datasets; orange points are the top 1% log fold change of proteins that have a fold change that is not zero in both datasets; crimson points are expected targets of the drugs.

The validation experiment data merged with the MCF7 data from the main experiment showed that DHFR is a significantly upregulated target in both datasets in response to MTX. HMGCR was also seen as an upregulated target of Atorvastatin. The validation experiment also showed that GEMIN4, found among the top proteins in the OPLS-DA model of Acetaminophen, was also stabilized in the PISA experiment, although the fold change was not as high.

# 4. Discussion

## 4.1 Main experiment

The OPLS-DA models successfully identified the expected targets for MTX and Atorvastatin. Though MTX is an anticancer and lethal drug, it was included in this study as a quality control since the expected target, DHFR, is already known. The OPLS-DA model and bar dot plot (Figure 7a & 8f) successfully identified DHFR as the target of MTX assuring the quality of the analysis.

The expected target of Atorvastatin (a non-lethal drug), HMGCR, was identified as a target of Atorvastatin. Atorvastatin is a drug used to reduce the risk of cardiovascular disease and lower lipid levels (Drugbank). Atorvastatin works by competitively inhibiting HMGCR which causes higher expression of LDL receptors in the liver and successively increases the catabolism of plasma LDL which lowers the concentration of cholesterol in plasma (O'Leary *et al.* 2016). The OPLS-DA model showed HMGCR to be one of the most specific up-regulated proteins in response to Atorvastatin (Figure 7b). Furthermore, plotting the expression of HMGCR for each treatment vs the control showed that HMGCR was significantly up-regulated when treated with Atorvastatin (Figure 7g).

However, HMGCR was also significantly up-regulated in response to treatment with Amlodipine in MCF7 cells (Figure 7g). Amlodipine is a treatment for high blood pressure and coronary artery disease which is known to also lower cholesterol (Salehi *et al.* 2012). Therefore it could also have an effect on HMGCR. The effect, in this study, even being comparable even to a specified drug such as Atorvastatin is a noteworthy observation.

In the hierarchical clustering, Cluster 16 showed larger fold changes in response to treatment with Amlodipine and Atorvastatin (Figure 6a). Further inspection of the pathway analysis performed on the clusters shows that the pathways associated with the proteins in Cluster 16 are

from the cholesterol metabolic pathways (Figure 6b). Hence the hierarchical clusters also confirm that Atorvastatin and Amlodipine have an effect on the cholesterol metabolic process.

OPLS-DA models also identified protein groups related to the processes that are affected in response to some of the drugs. For instance, in the model for Celecoxib, NDU proteins are down-regulated (Figure 7h). NDU proteins are a part of the mitochondrial membrane respiratory process. Celecoxib is a treatment for inflammation that is known to inhibit the cyclooxygenases COX-1 and COX-2 (DrugBank). It has been known to suppress mitochondrial function and inhibition of mitochondrial oxygen consumption (Tatematsu *et al.* 2018, Pritchard *et al.* 2018). This would explain the downregulation of NDU proteins. Likewise, the pathway analysis of the top-ranking down-regulated proteins from the OPLS-DA model showed that they were primarily involved in NADH dehydrogenase complex assembly and other mitochondrial processes (Figure 8).

On the other hand, the OPLS-DA model for Metformin shows an up-regulation of NDU proteins (Figure 7i). Metformin is a treatment for diabetes and polycystic ovary syndrome and is known to accumulate in the mitochondria and inhibit the activity of mitochondrial complex 1 activity (DrugBank). This could be related to the regulation of NDU proteins shown in the OPLS-DA model.

LDC203974 is an anticancer drug that is known to affect mitochondrial translation and mitochondrial RNA polymerases. Since the mechanism is known, similar to MTX, it was also included as quality control in this project. The majority of the up-regulated proteins in the OPLS-DA model for LDC203974 are Mitochondrial Ribosomal Proteins (MRPs) (Figure 7j). These proteins are a part of the mitochondrial ribosome involved in mitochondrial translation. As seen in the pathway analysis of the OPLS-DA model for LDC203974, the down-regulated proteins are involved in mitochondrial translation (Figure 8). This is in line with the activity of the compound against mitochondria (Bonekamp *et al.* 2020). This was also found in the enrichment analysis of both the hierarchical clustering and OPLS-DA.

## 4.2 Merged dataset

While the models have effectively identified known targets for several different treatments, it also raised the question of if a larger dataset containing both lethal and non-lethal treatments such as the MCF7 data combined with the ProTargetMiner data can more effectively identify the targets. In the case of Prednisolone, the OPLS-DA model based on the merged data showed better rankings for mechanistic proteins compared to individual datasets. Prednisolone is a glucocorticoid and these are known to inhibit the transcription factor NF-Kappa B which regulates multiple aspects of immune functions and is a key mediator of immune responses (DrugBank).

In the dataset containing only the merged lethal and non-lethal data, the highest-ranking protein for Prednisolone is Myotrophin (MTPN) which promotes the dimerization of NF-Kappa B subunits and regulates the NF-Kappa B transcription factor activity (UniProt). This is shown to be down-regulated in the OPLS-DA model which could be explained by Prednisolone inhibiting NF-Kappa B.

The second top-ranking protein in the merged dataset is Coiled-coil domain-containing protein 22 (CCDC22) which is involved in the regulation of NF-Kappa B signaling and may be involved in the downregulation of NF-Kappa B activity (UniProt). This was shown to be up-regulated in the OPLS-DA model and could also be explained by Prednisolone's inhibitory effect on NF-Kappa B.

In the dataset with non-lethal data only, MTPN ranked 44$^{th}$ and CCDC22 ranked 147$^{th}$, while in the merged dataset, they ranked 1$^{st}$ and 2$^{nd}$, respectively (Figure 9). This shows that the targets ranked higher for Prednisolone in the combined dataset than the dataset of only non-lethal drugs.

## 4.3 Validation experiment

DHFR was successfully validated as a target for MTX in both the PISA volcano plots (Figure 11) and in the PISA vs. Expression plots (Figure 12). DHFR is a known target for MTX and this validation assures the quality of the analysis. HMGCR was also validated as a target for Atorvastatin. HMGCR and DHFR were significant and upregulated in both the PISA and main experiment datasets. GEMIN4 which was identified as a likely target for Acetaminophen using OPLS-DA was also upregulated in both data sets, but its fold changes were not remarkable. GEMIN4 is a part of the SMN complex which catalyzes the assembly of small nuclear ribonucleoproteins and is involved in rRNA processing and splicing of pre-mRNAs (UniProt). Since it was detected in both experiments it could be an interesting novel target for further investigation, as this protein is currently not associated with Acetaminophen in any literature.

# 5. Future outlook

This project can be further developed into an online tool, similar to ProTargetMiner. It could either be incorporated into ProTargetMiner or be developed into an independent library with a similar function. Since the combined dataset showed a higher ranking of relevant targets in the case of Prednisolone, it could also be interesting to build upon this dataset and further improve the target deconvolution through the OPLS-DA models. However, the addition of too many compounds is not advised, as if two compounds affect the same target, the specificity can be lost to some extent. The targets identified could also be investigated further with experiments or a deeper literature study to evaluate if they could be related to the drugs' MOAs.

# 6. Conclusion

To conclude, this project investigated whether the concepts used in FITExP and ProTargetMiner could be expanded to non-lethal datasets. In several cases, the models were able to identify targets and mechanistic proteins that can be related to the non-lethal drugs' mechanisms. Several targets were also validated using PISA. This shows that the expansion of ProTargetMiner can be done to successfully identify targets and MOAs for non-lethal drugs.

# 7. Acknowledgments

I would like to thank my supervisors Roman A. Zubarev and Amir Ata Saei for this wonderful opportunity and interesting project. A special thanks to Amir for the daily guidance and support. I want to thank my subject reviewer Jonas Bergquist for reviewing my work and making sure I was always on the right track. I would also like to thank the course coordinator Lena Henriksson for always being able to help with course-related issues and cheering us students on. A final thanks to my student opponent Albin Lundin, for reviewing my work and working with me throughout the project, and always being available as a sounding board.

# References


Aebersold R, Mann M. 2016. Mass-spectrometric exploration of proteome structure and function. Nature 537: 347–355.

Avram S, Halip L, Curpan R, Oprea TI. 2021. Novel drug targets in 2020. Nature Reviews Drug Discovery 20: 333–333.

Bonekamp NA, Peter B, Hillen HS, Felser A, Bergbrede T, Choidas A, Horn M, Unger A, Di Lucrezia R, Atanassov I, Li X, Koch U, Menninger S, Boros J, Habenberger P, Giavalisco P, Cramer P, Denzel MS, Nussbaumer P, Klebl B, Falkenberg M, Gustafsson CM, Larsson N-G. 2020. Small-molecule inhibitors of human mitochondrial DNA transcription. Nature 588: 712–716.

Bylesjö M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. 2006. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. Journal of Chemometrics 20: 341–351.

Chernobrovkin A, Marin-Vicente C, Visa N, Zubarev RA. 2015. Functional Identification of Target by Expression Proteomics (FITExP) reveals protein targets and highlights mechanisms of action of small molecule drugs. Scientific Reports 5: 11176.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology 26: 1367–1372.

Eden E, Lipson D, Yogev S, Yakhini Z. 2007. Discovering Motifs in Ranked Lists of DNA Sequences. PLOS Computational Biology 3: e39.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48.

Gaetani M, Sabatier P, Saei AA, Beusch CM, Yang Z, Lundström SL, Zubarev RA. 2019. Proteome Integral Solubility Alteration: A High-Throughput Proteomics Assay for Target Deconvolution. Journal of Proteome Research 18: 4027–4037.

Lindsay MA. 2003. Target discovery. Nature Reviews Drug Discovery 2: 831–838.

Mullard A. 2022. 2021 FDA approvals. Nature Reviews Drug Discovery 21: 83–88.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44: D733-745.

Pappireddi N, Martin L, Wühr M. 2019. A Review on Quantitative Multiplexed Proteomics. ChemBioChem 20: 1210–1224.

Pritchard R, Rodríguez-Enríquez S, Pacheco-Velázquez SC, Bortnik V, Moreno-Sánchez R, Ralph S. 2018. Celecoxib inhibits mitochondrial O2 consumption, promoting ROS dependent death of murine and human metastatic cancer cells via the apoptotic signalling pathway. Biochemical Pharmacology 154: 318–334.

Saei AA, Beusch CM, Chernobrovkin A, Sabatier P, Zhang B, Tokat ÜG, Stergiou E, Gaetani M, Végvári Á, Zubarev RA. 2019. ProTargetMiner as a proteome signature library of

anticancer molecules for functional discovery. Nature Communications 10: 5715.
Salehi I, Mohammadi M, Mirzaei F, Soufi F. 2012. Amlodipine attenuates oxidative stress in the heart and blood of high-cholesterol diet rabbits. Cardiovascular Journal of Africa 23: 18–22.
Savitski MM, Reinhard FBM, Franken H, Werner T, Savitski MF, Eberhard D, Molina DM, Jafari R, Dovega RB, Klaeger S, Kuster B, Nordlund P, Bantscheff M, Drewes G. 2014. Tracking cancer drugs in living cells by thermal profiling of the proteome. Science 346: 1255784.
Tatematsu Y, Fujita H, Hayashi H, Yamamoto A, Tabata A, Nagamune H, Ohkura K. 2018. Effects of the Nonsteroidal Anti-inflammatory Drug Celecoxib on Mitochondrial Function. Biological and Pharmaceutical Bulletin 41: 319–325.
Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research 46: D1074–D1082.

# Appendix A - TMT-sets of the pilot and main experiments.

**Table 1.** The setup of the pilot dataset for one cell line with TMT16 labeling. These experiments were performed with two replicates for two cell lines (MCF-7 and SH-SY5Y) resulting in a total of four TMT sets.

| TMT | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rep 1 | DMSO-24h | DMSO-48h | DMSO-72h | DMSO-96h | MTX-24h | MTX-48h | MTX-72h | MTX-96h | Atorvas-24h | Atorvas-48h | Atorvas-72h | Atorvas-96h | Celecoxib-24h | Celecoxib-48h | Celecoxib-72h | Celecoxib-96h |
| Rep 2 | DMSO-24h | DMSO-48h | DMSO-72h | DMSO-96h | MTX-24h | MTX-48h | MTX-72h | MTX-96h | Atorvas-24h | Atorvas-48h | Atorvas-72h | Atorvas-96h | Celecoxib-24h | Celecoxib-48h | Celecoxib-72h | Celecoxib-96h |

**Table 2**. The setup of the main dataset with TMT16 labeling and a list of the treatments, indications, and expected MOA. These experiments were performed with three replicates for three cell lines resulting in 9 TMT sets.

| TMT | Drug | Category | Target (drugbank) | Mechanism | Indication |
|---|---|---|---|---|---|
| **1** | DMSO | - | - | - | - |
| **2** | MTX | Anticancer drug | DHFR | Targets the folate pathway | Cancer |
| **3** | Enalapril | ACE inhibitors | ACE | Decrease the formation of angiotensin II | Hypertension |
| **4** | Atorvastatin | HMG-CoA reductase inhibitors (statins) | HMGCR | Inhibition of cholesterol synthesis | Hyperlipidaemia |
| **5** | Amlodipine | Calcium channel blocker | Calcium channels including e.g. CACNA1C | Inhibits calcium uptake and muscle contraction | High blood pressure and coronary artery disease |
| **6** | Celecoxib | NSAID | COX1 and COX2 | Inhibition of prostaglandin production | Inflammation |
| **7** | Metformin | Anti-diabetes | PRKAB1, ETFDH and GPD1 | Not completely understood | Diabetes and polycystic ovary syndrome. |
| **8** | Omeprazole | Proton pump inhibitor | ATP4A | Inhibits proton pump | Antacid |
| **9** | Metoprolol | Beta Blocker | ADRB1 and ADRB2 | Beta blocker | Hypertension and angina |
| **10** | Sildenafil | Sex enhancer | PDE5A | blocking PDE5A, enzyme that promotes cGMP breakdown, which regulates blood flow in the penis | Erection problems |
| **11** | Acetaminophen | Cold medication | Unknown | Unknown | Pain and fever |

| | | | | | |
|---|---|---|---|---|---|
| **12** | Diphenhydramine | Antihistamine | HRH1, CHRM2 | Inhibition of histamine receptor 1 | Allergy |
| **13** | Cimetidine | Antihistamine | HRH2 | Inhibition of histamine receptor 2 | Antacid |
| **14** | LDC203974 | Anticancer drug | POLRMT | Inhibition of mitochondrial RNA polymerase | Cancer |
| **15** | Fluoxetine | Antidepressant, selective serotonin reuptake inhibitor (SSRI) | SLC6A4, HTR2C, CHRNA2, CHRNA3, CHRNB4 CKS1B and KCNH2 | Inhibiting serotonin re-uptake in the synapse | Depression, etc. |
| **16** | Prednisolone | Steroid | NR3C1 | Inhibits the glucocorticoid receptor | Inflammation and immunity |

**Table 3.** IC50 of non-anticancer drugs in 3 different cell lines (in μM);100 means that the drugs did not affect the viability up till 100 μM. Labeling scheme is given. For drugs not affecting cell viability at 100 μM, 25 μM was used in main experiments. At higher doses, the drugs can be unspecific.

| **TMT** | **Drug** | **IC50_MCF7** | **IC50_Sushi** | **IC50_fibroblasts** |
|---|---|---|---|---|
| **1** | DMSO | - | - | - |
| **2** | MTX | 100 | 0.05 | 50 |
| **3** | Enalapril | 100 | 100 | 100 |
| **4** | Atorvastatin | 1 | 2.5 | 100 |
| **5** | Amlodipine | 7.5 | 7.5 | 10 |
| **6** | Celecoxib | 50 | 25 | 50 |
| **7** | Metformin | 100 | 100 | 100 |
| **8** | Omeprazole | 100 | 100 | 100 |
| **9** | Metoprolol | 100 | 0.05 | 100 |
| **10** | Sildenafil | 100 | 100 | 100 |
| **11** | Acetaminophen | 100 | 100 | 100 |
| **12** | Diphenhydramine | 100 | 100 | 100 |
| **13** | Cimetidine | 0.05 | 100 | 100 |
| **14** | LDC203974 | 100 | 100 | 100 |
| **15** | Fluoxetine | 25 | 37.5 | 10 |
| **16** | Prednisolone | 100 | 100 | 100 |

# Appendix B - Pilot data distribution and PCA

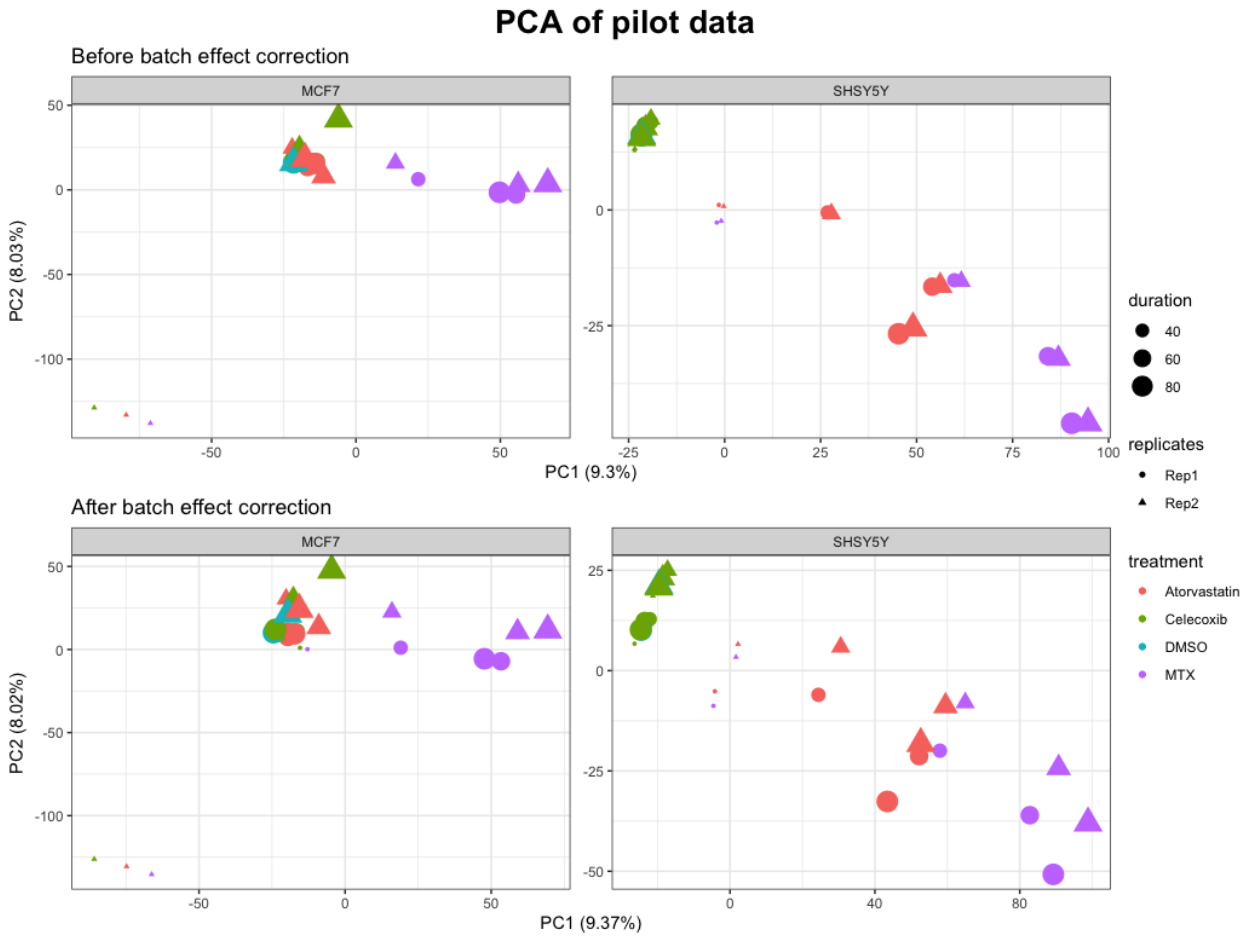


**Figure B1.** Principal component analysis of the pilot data before and after batch effect correction showing a reduced batch effect and separation of treatments.

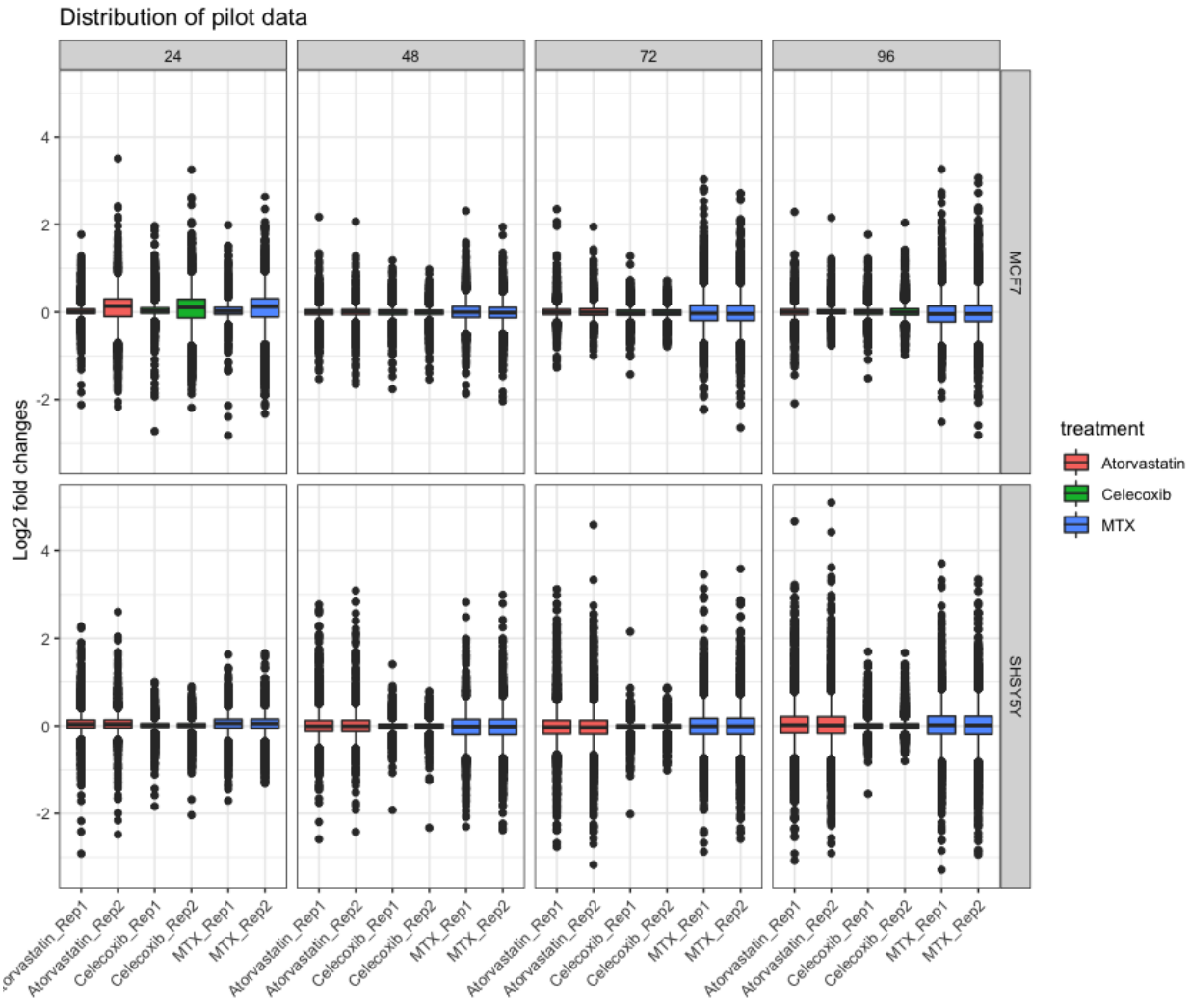


**Figure B2.** Barplot showing the spread of the pilot data in each replicate for each treatment, cell line, and duration.

# Appendix C - Effect of normalization on main experiment data

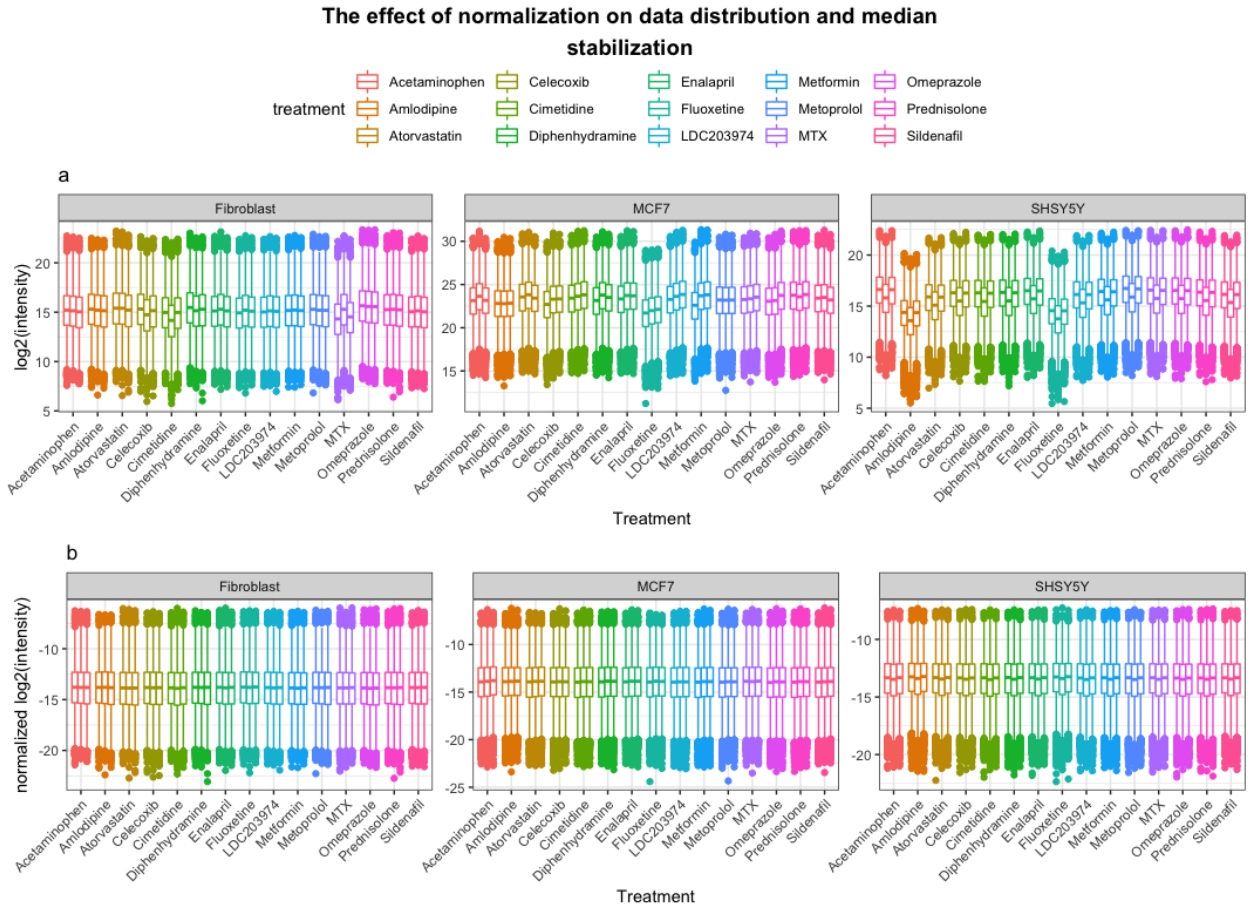


**Figure C**. Effect of normalization on the distribution of the main experiment data for each treatment in each cell line. a) before normalization using total intensities. b) after normalization using total intensities.

# Appendix D - Ranking of each treatment based on the total effect on the proteome

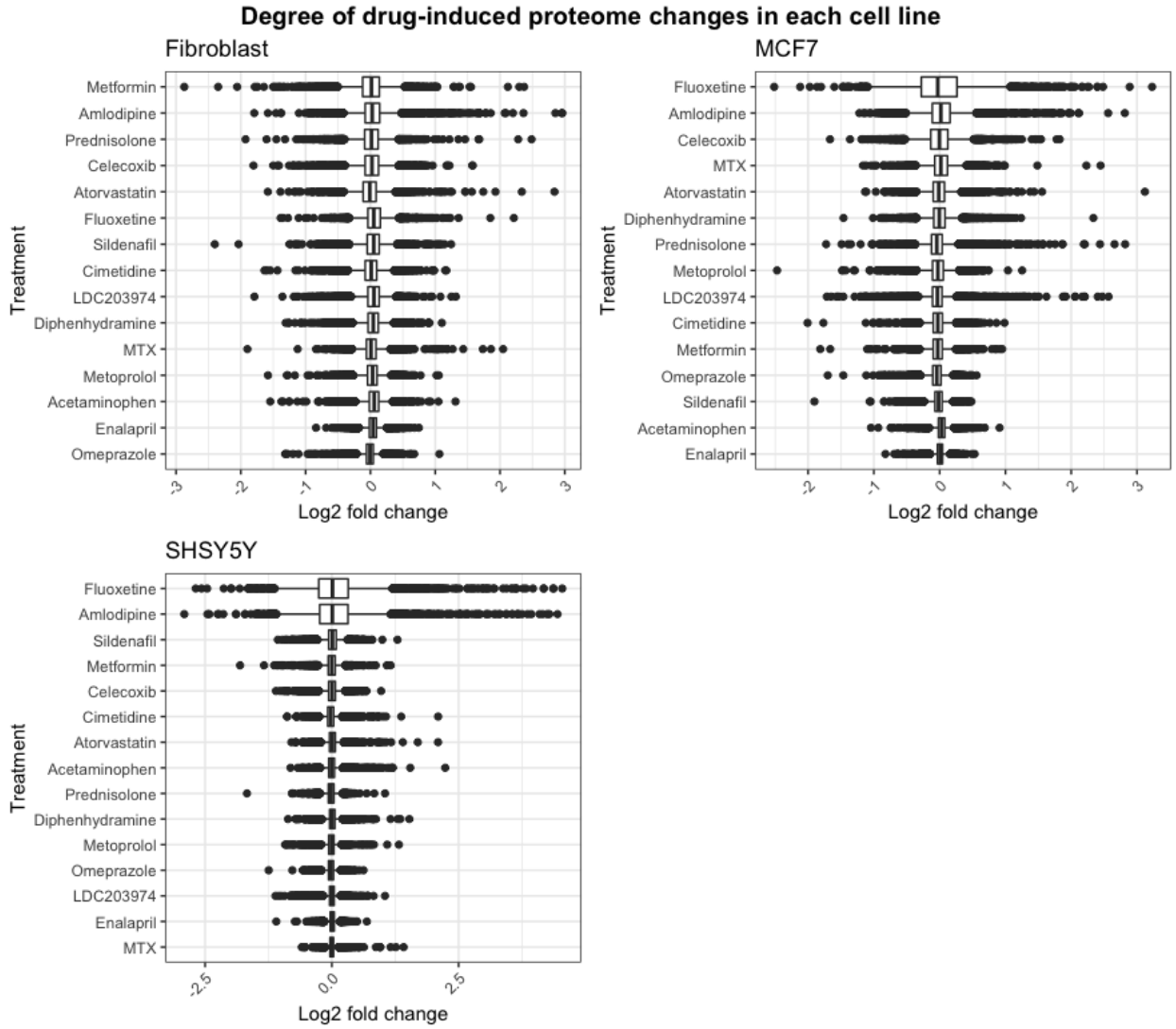


**Figure D.** Data distribution and rankings of the treatment's effect on the proteome.

# Appendix E - Main experiment data PCAs before and after batch effect correction

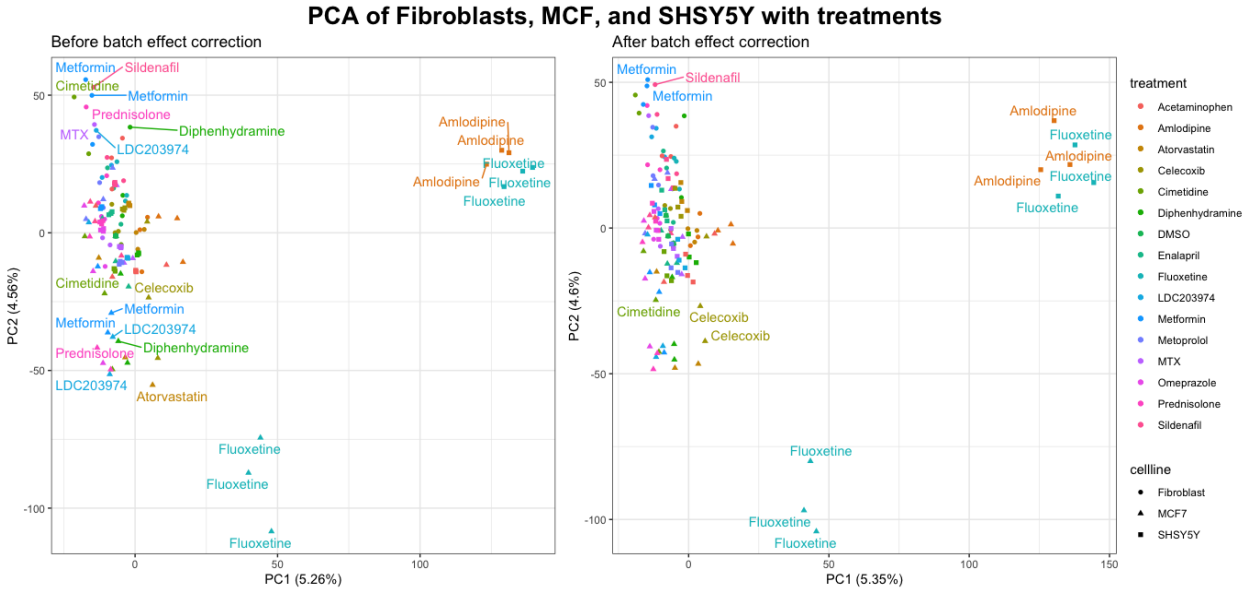


**Figure E1.** PCA plots of the main experiment data with all treatments and cell lines before and after batch effect correction.

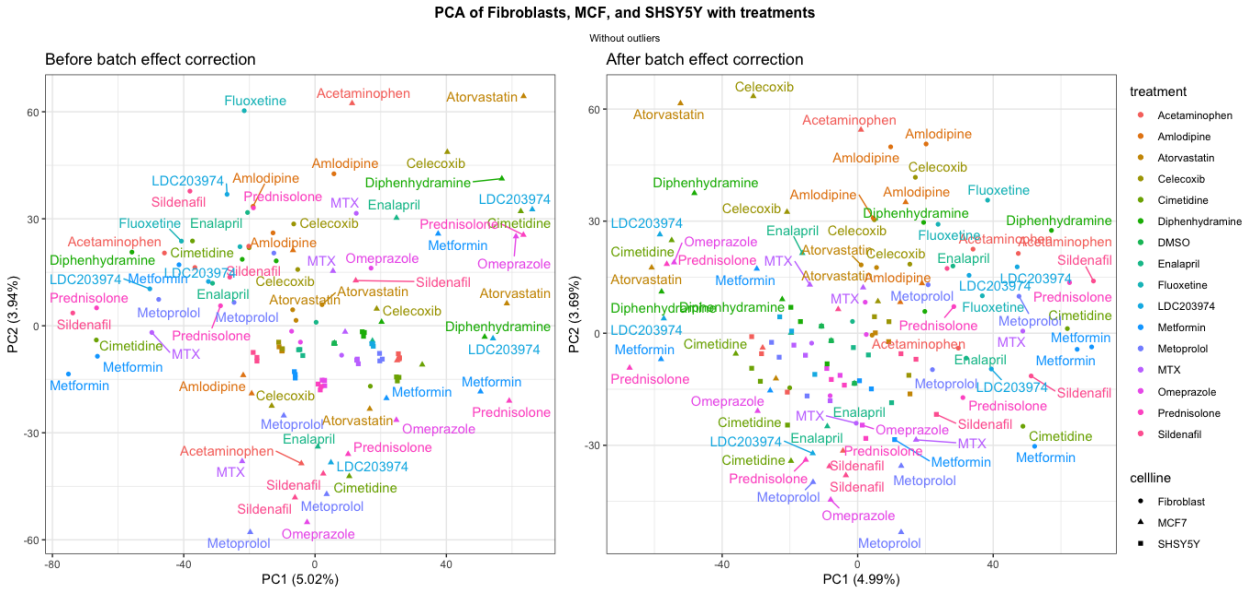


**Figure E2.** PCA plots of the main experiment data (with Amlodipine and Fluoxetine in SHSY5Y and Fluoxetine in MCF7 removed to clarify the spread of the other treatments).