1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Finding Hidden Patterns on Cardiovascular Toxicology Problems: The Case of Doxorubicin

Rute Filipa do Carmo Pino

A thesis submitted for the degree of
Master of Biomedical Engineering
at the University of Coimbra

July 2019

Finding Hidden Patterns on Cardiovascular Toxicology Problems: The Case of Doxorubicin

Author:
Rute Filipa do Carmo Pino

Supervisors:
Prof. Dr. Nuno Lourenço
Dra. Teresa Oliveira

Dissertation presented to the Faculty of Sciences and Technology of the University of Coimbra to obtain a Master's degree in Biomedical Engineering

Coimbra, 2019

Este trabalho foi desenvolvido em colaboração com:

Faculty of Sciences and Technology - University of Coimbra

CNC - Center for Neuroscience and Cell Biology of University of Coimbra

CISUC - Center for Informatics and Systems of University of Coimbra

Abstract

Medical breakthroughs nowadays depend almost entirely on scientific research which relies in elaborating numerous hypotheses and running them through continuous and exhausting processes and experiments. These processes usually generate large amounts of data which are normally processed and treated with standard statistical methods that do not live up to the demands imposed by the technological advances that demark our era. Doxorubicin (DOX) is an antitumor anthracycline antibiotic used for treating several types of cancer, such as breast cancer, Hodgkin's disease and leukemia. Although mitochondrial disruption is an early and sensitive marker of DOX cardiotoxicity, how metabolic stress contributes to the development of cardiomyopathy still needs to be clarified. To address this problem, an experimental dataset was built at the MitoXT laboratory using a model of metabolic inhibition of perfused hearts from Saline (SAL) and DOX-treated Wistar rats to identify metabolic alterations caused by an acute DOX treatment. The hearts were removed and perfused with three different energy substrates such as glucose, Galactose plus Glutamine (GG) and Octanoate plus Malate (OM). Separately, glycolytic (Iodoacetate (IODO)) and oxidative phosphorylation (Rotenone (ROT) or Potassium Cyanide (KCN)) inhibitors were added to the distinctive metabolic perfusion buffers, aiming at detecting mitochondrial defects in the DOX-treated group. In this work we applied techniques and computational tools to organise and mine the generated data in order to expose hidden patterns. Our conclusions suggest the exclusion of the protocol concerning the hearts perfusion with OM, confirming the original analysis. Additionally, we suggest that to spare time, means and animals, future experiments could only execute the glucose perfusion protocol. We also suggested that ANT and LDH transcripts expression, absolute weight difference, and Peroxisome proliferator-activated receptor-gamma coactivator (PGC-1$alpha$) ratio are the most relevant features to be considered for this problem. Finally, we established a classifier capable of an automatic distinction between DOX- and SAL- treated groups. Thus, we not only contributed to a better understanding of how metabolic stress contributes to the development of cardiomyopathy, by selecting which parameters show greater disparity between treatments, but we also confirmed that a detailed data analysis driven by Machine Learning allows a better exploration of these biological datasets enabling new discoveries and breakthroughs in this field.

Keywords: Machine Learning, Data Analysis, Drug Toxicity, Doxorubicin

Resumo

Hoje em dia, progressos na área da medicina são quase sempre influenciados pela investigação científica, a qual se apoia na elaboração de hipóteses e processos experimentais para as validarem. Geralmente, estes processos geram grandes quantidades de dados, os quais são, posteriormente, processados e tratados com o auxílio de métodos estatísticos tradicionais, ficando, muitas vezes, aquém das expectativas impostas pelos avanços tecnológicos que marcam a nossa era. A Doxorubicina (DOX) é um fármaco antitumoral utilizado no tratamento de diversos tipos de cancro, como cancro da mama, doença de Hodgkin's e leucemia. E embora a disrupção mitocondrial seja um indicador sensível e precoce da cardiotoxicidade provoca pela DOX, subsiste ainda sobre debate a razão pela qual o stress metabólico contribui para o desenvolvimento de cardiomiopatia. Assim foi elaborado, no laboratório Mito-XT, um modelo de inibição metabólica em corações perfundidos de ratos Wistar, tratados com solução salina ou com DOX, de modo a evidenciar as alterações metabólicas causadas por este fármaco. Os corações foram removidos e perfundidos com três substratos cardíacos diferentes: glucose, galactose e glutamina, e octanoato e malato. Separadamente, foram adicionados, aos distintos tampões metabólicos da perfusão, os inibidores glicolíticos (iodoacetato), e de fosforilação oxidativa (rotenona e cianeto) com o objetivo de detectar defeitos metabólicos ocultos nos grupos tratados com o fármaco. Assim, neste estudo aplicaram-se diversas técnicas e ferramentas computacionais, incluindo algoritmos de aprendizagem automática, com o objetivo de expor padrões desconhecidos nos dados recolhidos, analisando e estruturando o dataset, de forma a estabelecer, também, um classificador capaz de distinguir automaticamente os grupos tratados com e sem DOX. As conclusões deste trabalho, verificaram a análise original dos dados, confirmando a exclusão do protocolo de perfusão correspondente ao substrato octanoato e malato, devido a não manifestar quaisquer conclusões relevantes. Adicionalmente, sugerimos que para poupar tempo, fundos e animais se deveria, apenas, implementar o protocolo relativo à perfusão com glucose. Verificámos, também, que os parâmetros mais importantes para o problema em questão são: a informação genética relativa aos transcriptos ANT e LDH, a informação proteica e a diferença dos pesos das amostras. Deste modo, este trabalho não só revelou informação importante referente à contribuição do stress metabólico para o desenvolvimento cardiomiopatia, visto que foram selecionadas as features que melhor identificam as amostras tratadas com e sem o fármaco, como também confirmou que uma análise detalhada, utilizando abordagens provenientes de Machine Learning (ML), permitem uma melhor exploração de datasets biológicos, rev-

elando novas informações que podem levar aos progressos inicialmente mencionados.

Acknowledgements

Em primeiro lugar, aos meus orientadores Dr. Nuno Lourenço e Dra. Teresa Oliveira por todo o apoio, pela paciência e pela sabedoria com que me guiaram ao longo deste projeto de investigação, muito obrigada.

Agradeço à Dra. Filipa Carvalho e ao Dr. Paulo Oliveira pelo fornecimento dos dados e o apoio imprescindível a interpreta-los.

E finalmente, agradeço também às instituições Centro de Informática e Sistemas (CISUC) da Universidade de Coimbra e ao Centro de Neurociências e Biologia Celular (CNC) da Universidade de Coimbra por acolherem o meu projeto e permitirem que se realiza-se.

Aos meus pais e ao meu irmão, a toda a família e aos amigos que me acompanharam até Coimbra e aqui continuam, muito obrigada.

A Coimbra agradeço tudo aquilo que me ensinou, todas as portas que me abriu e, sobretudo, os amigos que me deu.

Agradeço à minha segunda família e os meus amigos me acompanharam nos últimos cinco anos.

Ao João Leandro, que entrou na minha vida à menos de um ano e acredita em mim desde aí, muito obrigada.

Muito obrigada a todos!

List of Publications

R. Pino, Teresa Cunha-Oliveira, Filipa Carvalho, Rita Garcia, Ana Burgeiro, Rui A. Carvalho, Paulo J. Oliveira, Nuno Lourenço. 'S5-05 Uncovering hidden patterns in biological datasets to identify metabolic alterations caused by acute and sub-chronic DOX treatments' European Journal of Clinical Investigation, Vol 49 S1, pp 85;

Contents

List of Figures

List of Tables

List of Acronyms

ML Machine Learning

DOX Doxorubicin

SAL Saline

GG Galactose plus Glutamine

OM Octanoate plus Malate

IODO Iodoacetate

KCN Potassium Cyanide

ROT Rotenone

ANT Adenine nucleotide translocator

Hif-1*alpha* Hypoxia-inducible factor 1*alpha*

LDH Lactate dehydrogenase

TFAM Mitochondrial transcription factor A

PGC-1*alpha* Peroxisome proliferator-activated receptor-gamma coactivator

NP Nonperfused

TC Time Control

DT Decision Tree

RF Random Forest

MI Mutual Information

MA Moving Average

RFE Recursive Feature Elimination

Chapter 1

Introduction

1.1 Motivation and Objectives

We live in an era where the machines are constantly evolving, becoming more intelligent by learning from data that has been collected over the years and from our everyday routines. Thus, today there is an increasing effort for acquiring the power of processing our data through Machine Learning (ML) implementations, in order to obtain meaningful conclusions and make accurate predictions. Everyone wants to predict the future by learning the modern magic of ML algorithms and Artificial Intelligence.

Nowadays, medical breakthroughs depend almost entirely on scientific research which relies on elaborating numerous hypotheses and running them through continuous and rigorous processes and experiments. These processes usually generate large amounts of data which are normally treated with standard statistical methods that might not take into account relationships between apparently unrelated features, leading to a possible missing out on important information and hidden patterns.

In the 1960s, researchers started to investigate a drug called Doxorubicin (DOX) (also known as Adriamycin). Di Marco et al. conducted a study where they demonstrated the high therapeutic potential of this new drug. [4]

These days, DOX is often prescribed for treating several types of cancers, such as breast cancer, Hodgkin's disease and leukemia. However, it has been shown that the clinical use of this drug is limited due to its dose-dependent cardiomyopathy.[5]

Bearing in mind this limitation, several studies and experiments have been made in order to explain the possible mechanisms behind DOX's cardiomyopathy. A project was developed with the MitoXT laboratory, where and experiment was performed with the purpose of unveiling how metabolic stress contributes to the development of cardiomyopathy, using a model of metabolic inhibition in perfused hearts from Saline (SAL) and DOX-treated Wistar rats.[6]

The data obtained during these experiments was analysed solely using standard statistic methods leaving a window of opportunity to apply modern techniques, such as ML.

The main goal of this work is to use a data driven approach, in order to perform an deeper analysis of the dataset gathered, in order to confirm the previous studies conclusions [6], and possibly uncover additional information that might have been missed by the use of traditional statistical methods.

We also propose and implement a ML model capable of classifying rats that were treated with DOX or SAL, using the biological information available. In the end, our dataset investigation results confirmed the original analysis. We also concluded that to spare time, funds and animals, the protocol concerning the substrate Octanoate plus Malate (OM) hearts perfusion should be excluded, since it was proved not to add relevant information to our purpose, and only the glucose perfusion should be implemented, since it provided the most relevant information.

Moreover we concluded that the most important parameters for this investigation were the Adenine nucleotide translocator (ANT) and Lactate dehydrogenase (LDH) transcripts expression, the absolute weight difference, and the PGC-1$alpha$ ratio, due to their contribution to our models classification and clear alterations found in each treatment correlation results concerning these features. Concerning the classification phase, our models were capable of an automatic distinction of DOX- and SAL- treated groups.

Thus, we not only contributed to a better understanding of how metabolic stress influences the development of cardiomyopathy, by selecting which parameters show greater disparity between treatments, but we also confirmed that a detailed data analysis driven by ML allows a better exploration of these biological datasets, enabling new discoveries and breakthroughs in this field.

Additionally, this work will address the difficulties found in processing the data into the correct structure, capable of fitting each model's requirements. Examples of these obstacles are label incoherencies, missing values and poorly structured data. In Chapter 5, a possible systematization of the experimental data structuration process will be suggested .

1.2 Structure

This report will be divided in 7 chapters. Chapter 2 will be addressing the biologic context of the original problem from which we collected our data. It is divided in 6 sections describing all steps concerning the problem elaboration, description, the experimental protocol and the measurements taken and finally some conclusions taken from the original analysis. Chapter 3 holds the state of art revision of the computational techniques used nowadays. It is also divided in 6 sections concerning the different procedures that can be implemented. Chapter 4 deals with the experimental setup, selection and transformation the original data into our functional dataset, as well as the data cleaning process. Chapter 5 includes our results and their discussion, thus it is divided by the exploratory process results, followed by classification ones. Chapter 6 contains the lessons learned during the data selection and transformation phase, which resulted from some misstructuration and contextualization of the original data. Finally, Chapter 7 concerns our work final conclusions, and next steps to take in our implementation.

Chapter 2

Biological Context

2.1 Doxorubicin Role in Cardiotoxicity

Doxorubicin (DOX) (Figure 2.1) is an anthracycline quinone antibiotic, used for treating several types of cancers, such as breast cancer, Hodgkin's disease and leukemia. Its clinical use is, however, limited by its dose-dependent cardiomyopathy. [5]

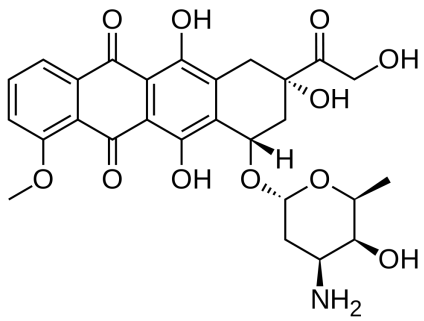


Figure 2.1: DOX chemical structure.

Since the first in vivo studies of its administration on cancerous animals, DOX promoted inhibition of neoplastic proliferation and subsequent increase of the animal survival rate.[5]

Some studies showed that the target of DOX toxicity is the cardiac tissue, which is enriched with mitochondria.[7] They also suggest that one of the reasons why the heart muscle is more susceptible to oxidant-induced damage results from the presence of low levels of antioxidant enzymes, such as catalase, in myocytes.[7]

Moreover, DOX accumulates into mitochondria over time, therefore, the selective toxicity of DOX to the heart was attributed to the selective damage to cardiac mitochondria. [5] [7]

After a single or course of therapy with DOX, acute injuries corresponding to the induced cardiac toxicity may occur. [8]

On the other hand, chronic side effects are more serious, irreversible, and involve the development of cardiomyopathy and ultimately congestive heart failure. Early-onset, chronic cardiotoxicity usually occurs within a year of treatment, persisting or even progressing after the treatment cessation, leading to chronic dilated cardiomyopathy in adult patients and to restrictive cardiomyopathy in pediatric patients. On the other hand, late-onset progressive cardiotoxicity results in ventricular dysfunction, heart failure, and arrhythmias years or even decades after chemotherapy occurred, suggesting the need for a continuous follow-up of the cardiac status of patients who received anthracyclines. [8]

Acute toxicity can affect the treatment result, since it occurs from a few minutes to a week of treatment. Whilst sub-chronic and chronic effects appear after treatment and can result not only of the compound acute toxicity, but also of the cellular adaptations the treatment effect. Thus, firstly, it should be clarified whether this compound has or has not acute toxicity.

2.2 Dataset Context and Biological protocol

Although mitochondrial disruption is an early and sensitive marker of DOX cardiotoxicity, how metabolic stress contributes to contributes to the later development of cardiomyopathy remains to be explored. To address this problem, an experimental dataset was generated using a model of metabolic inhibition in perfused hearts from Saline (SAL) and DOX-treated Wistar rats.

Concerning the protocol, for the DOX treatment model, sixteen-week-old male Wistar rats (n=46-50/group) were injected with a single dose of 20 mg/kg DOX or the equivalent volume of the vehicle solution of NaCl 0.9% (i.p), for the SAL treatment mode. Both treatments rats were sacrificed after 24 hours.[6]

Additionally, there were two groups of Wistar rats to be considered:

- Nonperfused (NP) hearts: the rats were treated, sacrificed and the hearts

were removed but not perfused;

- Time Control (TC): the rats were treated, sacrificed and the hearts were perfused only with substrate, without addition of inhibitors.

The animals were then sacrificed, the hearts were removed and perfused using a Langendorff apparatus with one of the distinct cardiac substrates:

– Glucose

– Galactose plus Glutamine (GG)

– Octanoate plus Malate (OM)

Separately, glycolytic (Iodoacetate (IODO)) and oxidative phosphorylation (Potassium Cyanide (KCN) or Rotenone (ROT)) inhibitors were added to the different metabolic perfusion buffers, aiming at exposing undercover mitochondrial defects in DOX-treated group.

	Glucose			Galactose plus Glutamine			Octanoate plus Malate		
	KCN	IODO	ROT	KCN	IODO	ROT	KCN	IODO	ROT
SAL	6	6	5	5	6	3	6	4	5
DOX	6	6	5	6	6	5	6	4	5
Total	34			31			30		

Table 2.1: Number of rats per substrate and inhibitor perfusion combination.

After perfusion, several metabolic and mitochondrial proteins were selected and semi-quantified by Western blotting.[9] mRNA levels were also obtained by RT-PCR technique.[10]

2.2.1 Injections Timeline

Figure 2.2, contains an explicative representation of the injections timeline during the hearts perfusion with both substrate and inhibitors.

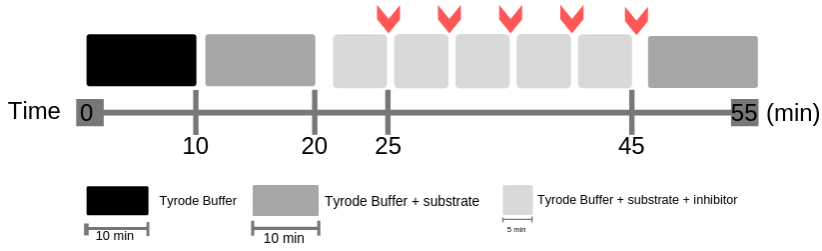


Figure 2.2: Injections timeline during perfusion.

Initially, the hearts were perfused for 10 minutes with the buffer for calibration purposes. Afterwards, the substrate was injected and the hearts were perfused again during 10 minutes. Then, 5 different injections of inhibitor, separated by 5 minutes perfusions, followed. The concentration of inhibitor increased gradually at each injection, as shown in table 2.2. Finally, the heart was again perfused with the substrate for 10 minutes more in order to evaluate its recovery from the inhibitors injections.

	1st	2nd	3rd	4th	5th	6th	7th
Cyanide (mM)	-	0.05	0.1	0.2	0.3	0.5	-
Iodoacetate (uM)	-	12.5	25	50	75	100	-
Rotenone (uM)	-	0.25	0.5	0.75	1	1.5	-

Table 2.2: Concentration of the inhibitors KCN, IODO and ROT, per injection.

At the end of each injection, the heart pressure values were measured for 30 seconds and registred.

2.2.2 Collected Data

In order to detect undercover mitochondrial defects in the DOX-treated group the following parameters were measured per animal:

- Heart and body weight;
- Tibia size;
- Heart physiology parameters: heart pressure, substrate flow, recuperation flow;

- Protein expression (Peroxisome proliferator-activated receptor-gamma coactivator (PGC-1*alpha*), Mitochondrial transcription factor A (TFAM) and Ubiquitin);
- mRNA expression (Adenine nucleotide translocator (ANT), Hypoxia-inducible factor 1*alpha* (Hif-1*alpha*) and Lactate dehydrogenase (LDH)).

2.3 Preliminary Conclusions

2.3.1 Acute Model: Glucose as substrate

In the acute model DOX treatment, hearts perfused with glucose suffered a decline in the number of heart beat and rate pressure product (RPP) when IODO was added, contrarily to the ones with ROT or KCN as inhibitors.

2.3.2 Acute Model: Galactose plus Glutamine as substrate

Perfusion with the GG substrate, inhibitor titration decreased the heart rate, despite that the decrease in the RPP was more evident adding IODO and KCN in SAL vs DOX group.

2.3.3 Acute Model: Octanoate plus Malate as substrate

Adding OM resulted in decreased heart rate an RPP in the presence of the inhibitors, working equality for each of them. However, these results weren't suficiente considering the other two protocols, so OM substrate should be excluded.

When the glycolytic and mitochondrial proteins were semi-quantified by Western blotting, alterations in the proteins involved in mitochondrial biogenesis and autophagy were observed in DOX hearts perfused with inhibitors. Looking at the conclusions, the data from the acute model appears to suggest that hearts from DOX-treated animals have improved function in the presence of metabolic inhibitors, an indication that DOX triggers metabolic adaptations

that results in a lower susceptibility to mitochondrial and glycolytic inhibition.

Chapter 3

Machine Learning

Machine Learning (ML) is about extracting knowledge from data. It is the art of guiding computers to learn without being explicitly programmed for it, in other words, autonomously.[11]

Nowadays, data is crucial for the advance of many scientific fields and, when treated correctly, has a great influence within both commercial applications and scientific studies. A more engineer-oriented point of view, states that a computer program is said to learn from experience, E, aiming at some task, T, and some performance measure, P, if its performance on T, as measured by P, improves with experience. [12]

Thus, ML algorithms can learn from input/output information. Depending on how this is implemented, they can be divided into two main types of learning: Supervised and Unsupervised.

Supervised Learning requires giving the algorithm the training to consist of pairs of input and desired output to take as example and to generalize for the desired outcome, gaining the ability to generate a human-free, accurate, output for an unprecedented input.[13]

One the other hand, for Unsupervised Learning the training set consists of unlabelled inputs, it does not require any previous knowledge of what the outcomes should be. Thus, its input information only has descriptive information (no labels). Consequently, this process is more challenging than supervised learning, since now the algorithm must learn on its own , with no examples to compare with the result.[13]

This type of learning is the preferential choice when a given dataset has dif-

ferent kinds of information and the main goal is to explore possible hidden patterns that can lead to unknown relationships between samples.

3.1 Common challenges

Usually a dataset is composed by a number of observations, called samples, and each sample is described by features. When dealing with supervised problems, datasets also include the variables allowing the samples to be distinguished/grouped and therefore classified - Labels.

Even for the simplest problems, a large amount of data is required for most ML algorithms to work accurately. Data shortage means fewer examples for our algorithm to rely on and learn from, possibly resulting in an incorrect generalization of the problem and devious classification.

However, a dataset can be filled with samples and still be unfitting for the problem. This happens to large datasets brimming with nonrepresentative training data resulting in a poor outcome of the algorithm with a compromised accuracy. As such it is important to have a training set that is representative of the problem at hand, whether we are applying supervised learning or unsupervised learning. [14]

3.2 Data Analysis Metrics

Before implementing any kind of transformation, it is useful to summarize each feature of our data into a single statement called a descriptive statistic. As the name suggests, descriptive statistics describes a particular quality of the data they summarize.

There are two general categories: the measures of central tendency and measures of spread.

Concerning data centralization:

- **Mean**: arithmetic average of a set of numbers, or distribution;

$$\mu = \frac{\sum x}{n} \tag{3.1}$$

Where n is the number of observations.

- **Median**: value in the center of the data distribution;
- **Mode**: most frequent value(s).

And concerning data distribution/spread:

- **Standard deviation**: measure of how spread numbers are;

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}} \tag{3.2}$$

Where n is the number os observations.

- **Variance**: measure of how far each value in the data set is from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1} \tag{3.3}$$

3.3 Data Visualization

Another way of analysing the data is through data visualization. A visualization has to accurately convey the data, and should not mislead or distort the viewer interpretation of what is being presented. Visualizing data involves taking data values and convert them in a systematic and logical way into the visual elements that make up the final graphic. [1] Thus, it is an important step in the exploratory process, because it allows a clearer interpretation of the data, its distribution and its tendencies, as well as facilitating comparison between features.

There are many data visualization techniques and choosing one has to take into account the type, the scale and the distribution of the data along with the purpose of the visualization. One remarkable example of visualisation tools are the Boxplots (Figure 3.1). They are simple yet informative, and they work well when plotted next to each other allowing the visualization of many distributions at once, making them easier to compare. [1]

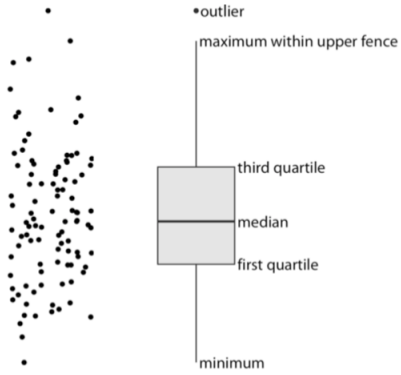


Figure 3.1: Anatomy of a boxplot - on the left there are a cloud of points and on the right the corresponding boxplot. Taken from [1]

This form of representing the data is intuitive: the line in the middle of the boxplot represents the median, and the box encloses the middle 50% of the data. The *whiskers* are represented as the vertical lines extending upwards and downwards from the box extending either to the maximum and minimum values of the data or to the maximum/minimum values that fall within 1.5 times the height of the box, whichever yields the shorter whisker. Data points that fall above or below 1.5 times the height of the box are called outliers and are usually represented as individual dots.[1]

Similar to boxplots, there are **violin plots** which can also be used to picture the data density and will accurately represent bimodal data whereas a boxplot will not.

3.4 Data Preprocessing

Preprocessing the data is the first step one has to perform, before applying any analysis. A careless disregard of this step can have a major impact on the algorithm and compromise its outcome. Thus, a good understanding of the existing preprocessing techniques will provide us with the knowledge of which is the one that best suits our solution. For example, problems with a large dataset, data filtering and data elimination should be the best approaches, as well as feature selection and reduction techniques which will help reducing the dimensionality of the data allowing the algorithm to operate faster and more efficiently.[15][16]

3.4.1 Data Cleaning

One of the first steps through preprocessing is to consider each value and its accurateness towards the context of the problem in hands.

- **Data Irregularities**

  Looking through the dataset, knowing the context of the its information, should help identify inconsistencies and illegal values within the data.

  This analysis is only straightforward if the analyst has a good background knowledge of the problem at hand. As such, this part of the analysis is usually conducted with the assistance of a domain expert, capable of providing insightful information to distinguish between an incorrect value and an abnormal value.

- **Outliers Detection**

  Outlier detection plays an essential role in the detection of abnormal observations. The removal of the faulty values should help cleanse the data and decontaminating its effect on the data set. [17]

  However, in some cases, it can have a different interpretation, and the outlier provides important information about the problem at hand. As such, the analysis of these values needs to be done with care.

  There are three fundamental approaches to the problem of outlier detection:

  – Identification of the outliers without prior knowledge of the data (analogous to unsupervised learning). This method processes the data as a static distribution, locates the most isolated points, and flags them as potential outliers. Therefore, this approach assumes that errors are separated significantly from the 'normal' data and will appear as outliers. [17]

  – Modeling both normality and abnormality which requires pre-labelled data, tagged as normal or abnormal. This approach is analogous to supervised classification since the algorithm has prior knowledge of what is normal and what is not, classifying the input accordingly. [17]

  – Modeling only normality or in a very few cases abnormality. The

algorithm knows previously what is normal but must figure out what is not.

- **Missing Values**

  Incomplete data is a common issue regarding real world datasets. When processing unknown data, it is important to consider some factors such as:

  – human error: a value is missing because it was deleted or not registered;

  – not applicable: the feature is not valid for a given instance, e.g., it does not exist for a given instance;

  – irrelevance: the designer of the training set does not take a certain value to consideration because it is not relevant (so-called don't-care value). [16]

  When dealing with a large dataset, it is normal to reject a sample containing missing information. However, with a small number of samples it is important to find ways of keeping them so that we do not lose any information.

  Considering this problem, there are some missing values handling techniques, such as, imputation, which substitutes the missing value with another one. There are some imputation techniques described in the literature, being one of the most common the imputation using the average of the correspondent feature. Imputation is, however, only recommended when the number of missing values is very small.

3.4.2 Data Transformation

Many ML algorithms have better performance rates when all features have a close scale and similar distribution. In order to achieve these conditions, there are a vast number of methods to choose from, depending on the type of the algorithm we want to implement and the feature values we will be applying.

- **Standardization/Scaling**

  Standardize generally means changing the values so that the distribution

standard deviation from the mean equals one. Its outcomes are very close to a normal distribution, centred around $\mu = 0$, and with a deviation of $\sigma = 1$, where μ is the mean and σ is the standard deviation. This way all features should have the same magnitude. Standard scores (also called *z scores*) of the samples are calculated by:

$$z = \frac{x_i - \mu}{\sigma} \tag{3.4}$$

where x is the value to be normalized and z is the normalized value of x.

- **Min-Max scaling**

  This technique, on the other hand, normalizes the data such that all features are exactly between 0 and 1. For each feature it follows the formula:

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3.5}$$

  It should, however, be applied for the cases where the distribution is not Gaussian or when the standard deviation is very small. [3]

3.4.3 Dimensionality Reduction

Some datasets have a common issue of excessive dimension. This problem tampers with both visualization and data processing. To overcome it, there are some techniques that can be applied in order to find an approximated version of the original dataset using fewer features. Example of these techniques are feature extraction and/or feature selection methods.

Feature Reduction

The main goal behind feature reduction is to find a new subset of dimensions by combining the original ones, without losing considerable information. There are two main methods concerning this technique: Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA).[18]

Both these methods are linear transformation techniques that can be used to reduce the number of dimensions in a dataset.[19]

- **Principal Components Analysis (PCA)**

  PCA is considered an unsupervised method since it does not use the output information.[18] It can also be defined as a process for compressing a significant amount of data into something that captures the essence of the original data. Pursuing a meaningful way to reduce the dimension of the data by focusing on the differences between samples and ways to combine them.

  PCA states that there is a principal component for each dimension of the data. The first component contains the maximum variance, the second component contains the second maximum variance and so on.

  Thus, PCA tries to find a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the data varies along each dimension.[20] Hence, it will select only a subset of new features contingent on its importance for explaining the data.

- **Linear Discriminant Analysis (LDA)**

  LDA is a supervised dimensionality reduction technique used for maximizing class separability.[19] It also looks for linear combinations of input variables that best explain the data, however, whereas PCA finds components of maximum variance in a dataset, LDA focus on the components that are useful for discriminating data, maximizing the separability among known information.[19]

  So LDA projects the data on to a different space that optimizes the separation between different groups of samples, by maximizing the distance between means and the minimizing the variation (scatter) within each group.

Although PCA and LDA are the most used methods, there are other procedures that can be applied for feature extraction, such as Multidimensional Scaling (MDS), which is relatively similar to PCA, and Singular Value Decomposition (SVD).

Feature Selection

The primary objective behind feature selection is to select a relevant and infor-

mative subset of features by removing the irrelevant and redundant ones.

This process helps understanding our data, since it diminishes the effect of over dimensionality, lowering computational requirement and improving the predictor performance. [21] To achieve this result there are different methods to considerer, grouped by **filter methods**, **wrapper methods** and **embedded methods**. [22]

Filter methods

Filter methods main purpose is to rank the features according to their relevance, giving each one a score and removing the ones that are not within a certain threshold. As such, ranking methods can be considered filtering methods since they are applied before classification to filter out the less significant features.[22]

These methods include correlation coefficients, information gain, mutual information, plus more traditional statistical tests such as T-test and Chi-squared.

- **Correlation Criteria**

  Correlation coefficients measure how strong the relationship between two variables is. If two features are highly positively correlated, they have similar information, thus keeping just one will reduce the dimension of the dataset without losing important information.

  One of the simplest and most commonly used criteria of correlation is the Pearson's correlation coefficient. It shows the linear relationship between two sets of data.The correlation between variables x and y is given by:

$$r = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum (x_t - \bar{x})^2}\sqrt{\sum (y_t - \bar{y})^2}} [23] \tag{3.6}$$

  Correlation values variate between -1 and 1. Thus, a correlation coefficient near 1 has two input variables that are highly correlated, whilst a -1 value means that for every positive increase in one variable, there is a negative decrease of the other, having a strong negative linear relationship. Zero means that exists weak or non-existing relationship.

[24]

- **Mutual Information (MI)** Alternative to Correlation Criteria, MI is also an efficient filter method. MI is a measure of statistical independence

that can obtain any kind of relationship between random variables, including nonlinear relationships.

Additionally, MI is invariant under space transformations. Thus, it measures the amount of information that one variable has with another. This definition is very useful within the context of feature selection because it allows the quantification of feature subset relevance. [22]

Given two discrete random variables x1 and x2, their mutual information, I(x1,x2), is defined in terms of their probabilistic density functions (p(x1), p(x2), and p(x1,x2)): [25]

$$I(x_1, x_2) = \int \int p(x_1, x_2) log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} dx_1 dx_2 \tag{3.7}$$

Where MI is zero when x_1 and x_2 are statistically independent.

- **Kruskal Wallis**

  Kruskal Wallis is a non-parametric statistical test that considers the differences among three or more independent sampled groups on a single, non-normally distributed continuous variable. This test gives information about each feature relevancy, considering their discriminative power. [26]

  The Kruskal Wallis null hypothesis verifies if two or more variables come from the same distribution, meaning they have the same median, assuming that the shape of the distribution is the same. The H parameter measure is approximately chi-square distributed. So the probability of getting a particular value of H, if the null hypothesis is true, is the P value corresponding to a chi-square equal to H. Higher values of H suggest higher discriminative power. [27]

These methods do not incorporate the learning phase, therefore they are relatively robust against overfitting.

Wrapped methods

Wrapped methods bind the feature selection around the algorithm, using different combinations to predict the benefits of adding or removing a feature from the subset used. Hence, they search through the space of feature subsets calculating the estimated accuracy of the algorithm for each feature aiming to find the best optimization. [28]

These methods usually take higher computational power since they use learning machines as a black box to score subsets of features according to their predictive power. However, they do not incorporate knowledge about the specific structure of the classification or regression function, thus they can be combined with any learning machine.[28]

- **Recursive Feature Elimination (RFE)**

  RFE is one of the most common Wrapper methods, since it performs dimensionality reduction given an algorithm. It assigns the weights and the rankings to each input feature while training the algorithm, removing the one with the smallest weight result. This process repeats until it achieves a single feature subset.[28]

  Thus, RFE has no effect on correlation methods since the ranking criterion is computed with information about a single feature. [29]

Filters and wrappers differ mostly by the evaluation criteria. It is usually established that filters do not engage in any kind of machine learning, whereas wrappers use the performance of a learning algorithm which is trained using a given feature subset. [30]

Embedded Methods

In contrast to filter and wrapper approaches, embedded methods do not separate the learning phase from the feature selection one. They perform feature selection in the process of training the data, without splitting into training and testing sets. Thus, structure of the class of functions under consideration plays a crucial role. [31]

Some of the most common methods are Forward Selection with Least Squares and Decision Trees (see in section 3.5).

3.5 Time Series

Some datasets have observations that were taken chronologically. For example, annual birth rates datasets, daily stock prices datasets, and so on. These time-ordered sequences of observations are called Time Series. [32] Thus, a time series is a set of observations, each one being recorded at a specific time t.[33]

Usually, a time series comprises different components: a trend-cycle component, T, a seasonality component, S, and a remainder component, also called noise, R (which includes all the remaining parts).[23]

Thus, a time series decomposition can be written as:

$$y_t = T_t + S_t + R_{t_i} \tag{3.8}$$

where y is the data and t the time period.[23]

Trend:general component that exists when there is a long-term increase or decrease in the data. It does not have to be linear. [23]

Seasonal: general component that exists when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and identified frequency. [23]

Usually, a time series involves large quantities of data, since the values are being recorded during a specific period of time. Thus, if the purpose is to look for turning points in a series and interpret any changes in direction, its better to use the trend-cycle component. [23] A classical decomposition of the time series is to use a Moving Average (MA) method to estimate the trend-cycle from seasonal data and take considerations about the tendency of the time series.

Moving Average (MA) - MA of order m can be written as:

$$T_t = \frac{1}{m} \sum_{k}^{j=-k} y_{t+j} \tag{3.9}$$

with order m = 2k +1. Which represents, the estimate of the trend-cycle at time is obtained by averaging the values of the time series within k periods of t. By applying MA to the data, it eliminates some of the randomness, leaving a smooth trend-cycle component. Thus, the order, m, determines the smoothness of the trend-cycle estimate. [23]

3.6 Learning Algorithms

There are numerous ways of constructing a ML Algorithm. Depending on the task, we can use different learning approaches: supervised, semi-supervised,

and unsupervised.

Every algorithm requires some kind of input data that should be divided into training and test sets. The training set contains the input data that will be used by the learning algorithm to fit the model to the problem at hand. The testing set is used afterwards and its main goal is to evaluate the generalisation ability of the ML model, i.e., to assess the capacity of the model to work beyond the data used for training.

3.6.1 Supervised Machine Learning Algorithms

Supervised learning is one of the most commonly used types of machine learning. These algorithms are used whenever one wants to predict a certain outcome from a given input, having annotated examples of what that outcome could be. Thus, the goal is to model the relationship between the measured features and the associated labels of the data, so that it can accurately make predictions of which labels should be attributed to new sets of data. [2]

This is further subdivided into classification and regression tasks.[34]

Classification vs Regression

Classification is usually associated with prediction of discrete categories. Sometimes this is separated into binary classification, which has only two classes as input, and multiclass classification, which handles more than two classes. It is important to note that binary classification is often referred as having a positive class and a negative one. However, in this context, positive does not mean value or gain, but the object we want to classify. Contrarily, negative class refers to the class of the samples that do not represent the object in study.

On the other hand, regression tasks apply to predictions involving continuous quantities as labels. So, if there is some kind of continuity between the outputs of the algorithm performing classification, then it is considered a regression problem. [34]

Decision Trees

Decision Tree (DT) (Figure 3.3) are versatile ML algorithms capable of fitting and visualizing complex datasets. Fundamentally, they learn a hierarchy of if/else questions and answers, leading to a decision, as shown in the upcoming

figure.[2]

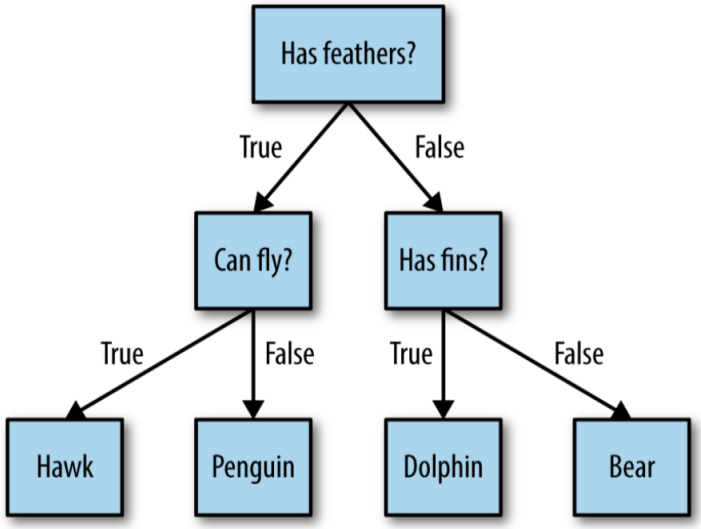


Figure 3.2: Decision Tree taken from [2].

The method implemented by decision trees begins with splitting the input dataset into subsets based on the result of testing an attribute. Afterwards, each new node of the tree is labeled with a new attribute to be tested, and its branches are labeled with their corresponding values. Thus, each new node of the tree splits the instance space into two or more sub-spaces, according to the different values of input attributes. This process ends when splitting is no longer an option or does not add any more value to the prediction (the only variable left for testing is the target).[2].

Through this process the algorithm classifies the samples by sorting them down the tree, starting in the root node, testing the attribute specified by this node, and then moving down the tree branches responding to each test until it finds the node containing the same value as the target variable, thus containing its classification. [35]

Decisions trees are great for exploratory knowledge discovery since they do not require any previous knowledge or parameter setting and can handle high dimensional data. However, they are more suited to regression problems, where the goal is to predict the value of a continuous attribute, since performing classification with few samples will result in a higher error probability. [12]

Random Forest

The Random Forest (RF) algorithm can be described as a collection of deci-

sion trees, where each tree is slightly different from the other. The main goal of applying RF is to add the random factor to the building process of each tree in order to ensure they are different from each other. This process can also be referred to as subspace sampling.[36] To each individual tree preliminary prediction probability is given. In the end, the final prediction is made considering all predictions probabilities.[2]

3.6.2 Unsupervised Machine Learning Algorithms

Unsupervised algorithms learn without any supervision. The goal is not to predict something but to discover interesting patterns between samples by exploring the distribution of the feature space.

Depending on the problem to address, one main technique to be applied is Clustering. Clustering Methods are meant to discover subgroups (or clusters) in the dataset. As an unsupervised learning class, they have no knowledge of what the output should be or the classification of the input. Thus, the main objective of these methods is to look for a way of distinguishing the data and group the samples considering their similarities. [37]

This is not a simple process, to define which samples are somewhat equal, there must be a criterion to that equality. This is often a domain-specific consideration based on prior, usually theoretical, knowledge of the input dataset. [38]

There are two best-known clustering approaches: K-means and Hierarchical Clustering. The interested reader can refer to [37].

3.7 Evaluation

After implementing the learning algorithms it is important to evaluate their performance and test their capacity to correctly classify the test data. In order to assess the classifier's performance there are some metrics and methods that can be applied to the output that can be divided into classification and regression performance methods.

Classification Performance Metrics

Confusion Matrices (CF) break the performance results up into their correctly and incorrectly predicted components for the two or more given classes. Which means, it reports how many times the classifier predicts a recurrence wrongly and how many times it predicts a nonrecurrence wrongly.[3]

It is represented as a table with four different combinations (considering a 2 classes problem) of predicted and actual values, as shown in the upcoming figure.

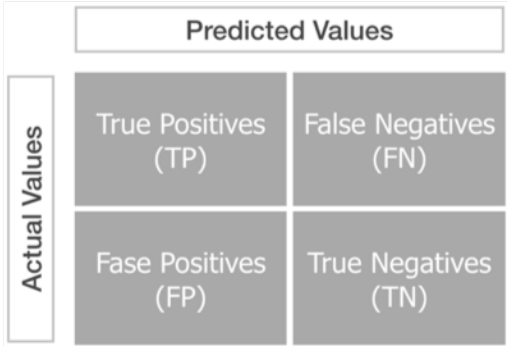


Figure 3.3: Confusion Matrix scheme.

- **Accuracy**: ratio of correct predictions, positive or negative, considering the entire predictions set.

$$Accuray = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

- **Sensitivity or Recall**: portion of correct positive predictions, considering all true positives existent. It traduces the capacity of the algorithm to predict which cases are actually positive.

$$Recall = \frac{TP}{TP + FN} \tag{3.11}$$

- **Specificity**: similar to recall, it is the fraction of true negatives predicted by the classifier, considering all true negatives, reflecting the capacity of the algorithm to predict which cases are actually negative.

$$Specificity = \frac{TN}{TN + FP} \tag{3.12}$$

- **Precision**: ratio of positive predictions that are actually true, consider-

ing all positive predictions (correct or not).

$$Precision = \frac{TP}{TP + FP} \tag{3.13}$$

- **F1-Score**: ponderation between precision and recall. A good value of F1-score suggests that the ratio of false positives and false negatives is low, which implies that there were few errors in the classification.
- **ROC Curve**: Receiver Operating Characteristic Curve (Figure 3.4).

  ROC is a graphical tool for visualizing the performance of learning algorithms. The ROC curve horizontal axis (x axis) denotes the false-positive rate FPR (1 - specificity) and the vertical axis (the y axis) denotes the true-positive rate TPR (sensitivity) of a classifier. Thus, ROC analysis shows the relationship between the sensitivity and the specificity of the classifier. [3]

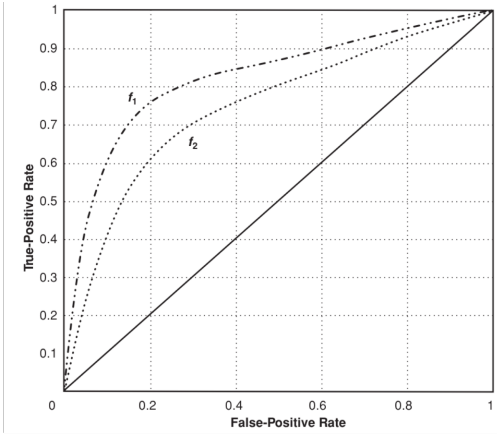


Figure 3.4: ROC Curve scheme taken from [3]

The area the total area under the ROC curve, abbreviated as AUC, representing the performance of the classifier averaged over all the possible cost ratios. [3] A good classifier performance will have an AUC value close to 1 (if equals to 1, then the classifier perfectly predicted all samples). However, a 0.5 AUC the algorithm is has the same accuracy as tossing a coin. The upcoming figures displays the ROC curve and the measures associated.

Chapter 4

Experimental Setup

4.1 Data Collection

To further study the cardiotoxic consequences of Doxorubicin (DOX), researchers at the MitoXT laboratory performed different experiments using Wistar rats treated either with Saline (SAL) or DOX injections. This project will focus on a data driven analysis of the results obtained with the acute protocol. The experiments' results were stored in different spreadsheet files which will be the aim of the following chapters analysis. Hence, it is important to understand how it was gathered and stored, before addressing how it should be processed and transformed into useful datasets.

- (Pre) Initial weight of the rat, in grams;
- (Post) Final weight of the rat, in grams;
- (Post) Tibia size of the rat, in centimeters;
- (Post) Recuperation flow, ml per minute;
- (Post) Heart weight (after its removal), in grams;
- (Post) Substrate flow, ml per minute;
- (Post) Weight of heart tissue stored in RNAlater.

Considering the rats submitted to perfusion with substrate and inhibitor, both pre- and post-weight and size measures were saved in a single spreadsheet file. Since we had three different substrates, this resulted in three different sheets,

one for each perfusion. Each sheet contains six tables aggregated in pairs by inhibitor, with the values measured for each rat.

Thus, figure 4.1, shows an example of a table concerning the measures of the rats saved in the sheet 'Galactose plus Glutamine' within the 'Weights' Excel file.

	Peso inicial (g)	Peso final (g)	Coração (g)	Comp. Tibia (cm)	Fluxo substrato (ml/min)	Fluxo recuperação (ml/min)	Coração (RNAlater)	
DOX31	378	371	1,075	3,6	21	16,5	0,51	
DOX32	-	-	-	-	-	-	-	
DOX33	377	373	1,015	3,7	8,5	4	0,408	
DOX34	409	391	1,104	4	44 fuga	32 fuga	0,483	
DOX35	393	382	1,093	3,8	20,5	11,5	0,625	
DOX36	383	375	1,051	3,7	17,5	16,5	0,561	
SAL 31	-	-	-	-	-	-	-	
SAL 32	-	-	-	-	-	-	-	
SAL 33	386	384	1,233	3,8	18,5	fuga	0,538	
SAL 34	393	389	1,223	3,7	13	3	0,502	
SAL 35	404	408	1,252	4	9	3	0,578	
SAL 36	-	-	-	-	-	-	-	

Rotenona

Figure 4.1: Spreedsheet tables of each rats pre perfusion information.

The rats that did not survive the treatment are marked with the symbol '-' as a value of each measure.

As mentioned previously in Chapter 2, during perfusion the **Heart Pressure** values of each injection last 30 seconds were registred using the Origin 8.5 software.

Each heart was submitted to perfusion with seven different substrate-inhibitor injections. These recordings were then exported to a spreadsheet, resulting in a total **ninety-five** files with seven columns each, and a total of **3761** values of heart pressure and seven additional columns with the corresponding time (from 0 to 30 seconds), as shown in figure 4.2.

Glucose		12,5uM Iodo 30seg		25uM Iodo 30seg		50uM Iodo 30seg		75uM Iodo 30seg		100uM Iodo 30seg		Glucose	
Time		Time		Time		Time		Time		Time		Time	
0	33,10	0	3,35	0	43,50	0	44,10	0	3,32	0	1,67	0	1,85
0,008	40,90	0,008	3,13	0,008	48,60	0,008	28,50	0,008	3,20	0,008	1,76	0,008	1,76
0,016	45,50	0,016	2,63	0,016	49,90	0,016	17,40	0,016	3,60	0,016	1,79	0,016	1,73
0,024	48,30	0,024	2,04	0,024	47,10	0,024	20,10	0,024	3,13	0,024	1,70	0,024	1,76
0,032	48,70	0,032	2,17	0,032	42,60	0,032	28,90	0,032	3,51	0,032	1,76	0,032	1,76
0,04	46,50	0,04	1,88	0,04	35,50	0,04	36,50	0,04	3,20	0,04	1,70	0,04	1,73
0,048	43,90	0,048	1,76	0,048	28,50	0,048	42,40	0,048	2,88	0,048	1,79	0,048	1,64
0,056	39,30	0,056	2,73	0,056	22,30	0,056	44,50	0,056	3,04	0,056	1,73	0,056	1,76
0,064	34,90	0,064	5,41	0,064	16,70	0,064	42,80	0,064	2,32	0,064	1,70	0,064	1,64
0,072	30,70	0,072	10,40	0,072	13,80	0,072	39,40	0,072	2,60	0,072	1,76	0,072	1,64
0,08	26,30	0,08	18,90	0,08	11,50	0,08	33,30	0,08	2,35	0,08	1,67	0,08	1,85
0,088	23,60	0,088	28,00	0,088	11,10	0,088	27,60	0,088	2,26	0,088	1,73	0,088	1,76
0,096	20,50	0,096	38,10	0,096	11,60	0,096	21,80	0,096	2,73	0,096	1,76	0,096	1,73
0,104	18,80	0,104	46,70	0,104	11,90	0,104	16,80	0,104	2,29	0,104	1,70	0,104	1,82
0,112	17,70	0,112	52,10	0,112	13,10	0,112	14,00	0,112	2,76	0,112	1,73	0,112	1,76
0,12	16,40	0,12	54,50	0,12	12,90	0,12	11,50	0,12	2,57	0,12	1,70	0,12	1,73
0,128	16,30	0,128	51,70	0,128	12,00	0,128	11,20	0,128	2,42	0,128	1,79	0,128	1,67
0,136	14,60	0,136	45,50	0,136	9,00	0,136	11,20	0,136	2,73	0,136	1,76	0,136	1,82
0,144	12,10	0,144	35,80	0,144	9,28	0,144	11,40	0,144	2,10	0,144	1,67	0,144	1,67
0,152	8,59	0,152	23,40	0,152	17,20	0,152	12,40	0,152	2,38	0,152	1,73	0,152	1,70
0,16	11,90	0,16	13,90	0,16	30,70	0,16	11,90	0,16	2,13	0,16	1,64	0,16	1,70
0,168	21,50	0,168	15,00	0,168	44,90	0,168	11,70	0,168	2,01	0,168	1,76	0,168	1,60
0,176	35,70	0,176	20,40	0,176	57,80	0,176	9,37	0,176	2,32	0,176	1,73	0,176	1,70
0,184	50,20	0,184	26,30	0,184	67,10	0,184	6,88	0,184	1,88	0,184	1,67	0,184	1,70
0,192	63,30	0,192	31,40	0,192	69,80	0,192	11,90	0,192	2,26	0,192	1,76	0,192	1,82
0,2	73,00	0,2	33,20	0,2	66,80	0,2	23,30	0,2	2,07	0,2	1,64	0,2	1,70
0,208	76,70	0,208	33,90	0,208	57,00	0,208	35,90	0,208	1,92	0,208	1,76	0,208	1,64
0,216	75,50	0,216	32,10	0,216	42,40	0,216	48,50	0,216	2,17	0,216	1,79	0,216	1,79
0,224	67,90	0,224	29,00	0,224	25,70	0,224	58,00	0,224	1,79	0,224	1,70	0,224	1,57
0,232	56,00	0,232	26,00	0,232	19,60	0,232	61,90	0,232	2,10	0,232	1,76	0,232	1,98
0,24	41,20	0,24	21,80	0,24	25,20	0,24	60,80	0,24	1,95	0,24	1,70	0,24	1,79
0,248	24,40	0,248	19,10	0,248	35,00	0,248	52,70	0,248	1,85	0,248	1,79	0,248	1,67
0,256	16,60	0,256	16,40	0,256	43,70	0,256	40,30	0,256	2,17	0,256	1,76	0,256	1,82
0,264	21,20	0,264	14,30	0,264	48,60	0,264	25,10	0,264	1,79	0,264	1,70	0,264	1,64
0,272	29,40	0,272	13,60	0,272	50,70	0,272	17,90	0,272	2,07	0,272	1,76	0,272	1,79

Figure 4.2: Spreedsheet table of each rats heart pressure values for each injection during perfusion.

Measures obtained after perfusion:

- **Transcript information**, of 4 pairs of samples combinations, concerning the transcripts Adenine nucleotide translocator (ANT), Hypoxia-inducible factor 1*alpha* (Hif-1*alpha*) and Lactate dehydrogenase (LDH);

- **Protein information**, of 4 pairs of samples combinations, concerning the proteins Peroxisome proliferator-activated receptor-gamma coactivator (PGC-1*alpha*) and Mitochondrial transcription factor A (TFAM).

For the transcript information, values were obtained after PCR analysis for the transcripts ANT, Hif-1*alpha* and LDH, and saved in three different files, one for each PCR plate. Within the files, there were different sheets concerning the measures of each transcript. In each sheet the values were separated by different tables regarding the substrate-inhibitor combination during perfusion. Figure 4.3, shows the tables built in the Hif-1*alpha* sheet, for the Plate 1 file. The first line of the tables shows the substrate-inhibitor combination, for example, 'G+Rotenone (ROT)' stands for the Glucose substrate combined with ROT inhibitor.

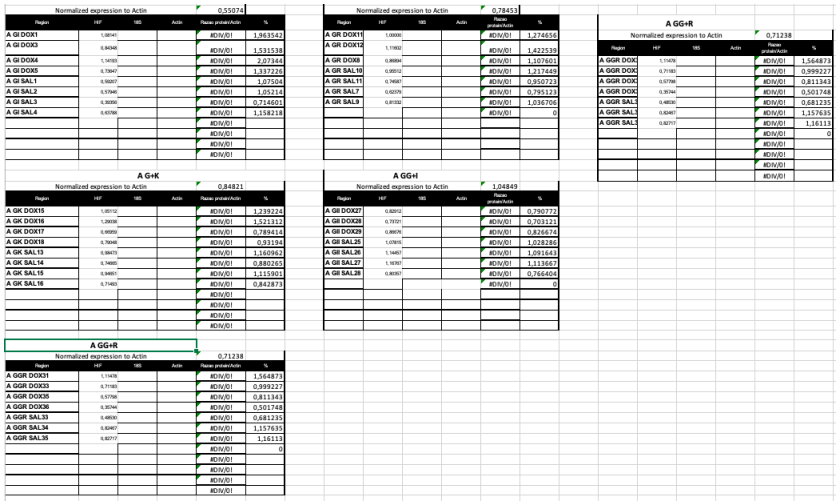
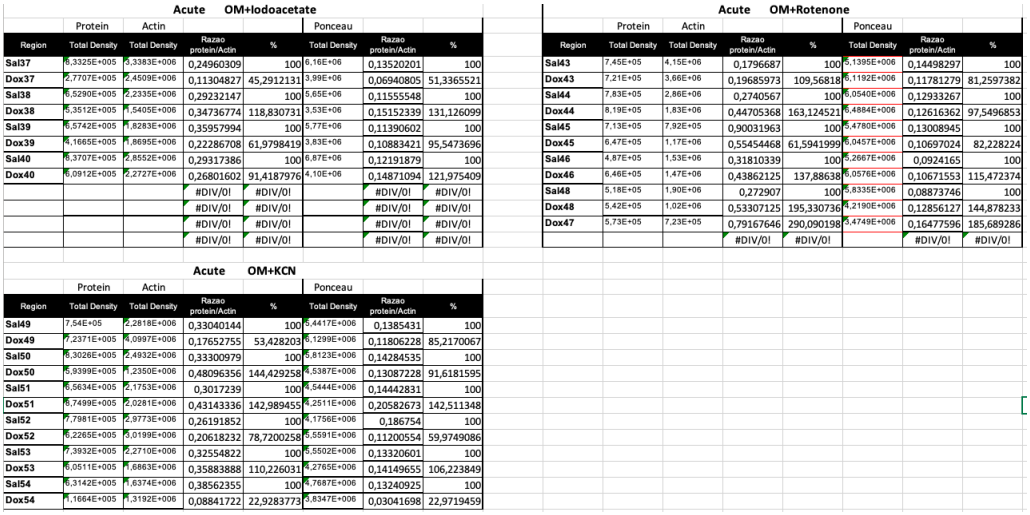


Figure 4.3: Spreedsheet tables of the transcript expression information.

Finally, the data acquired through the Western Blot technique for the proteins PGC-1*alpha* and TFAM, was stored in other three different files, one per substrate. Each file contained one PGC-1*alpha* sheet and one TFAM sheet, both with different tables for the values of each sample. Each one of these tables was grouped by inhibitor. Figure 4.4 displays the PGC-1*alpha* sheet of the Octanoate plus Malate (OM) substrate file, showing three different tables, corresponding to each three inhibitors samples.



Acute OM+Iodoacetate

	Protein	Actin			Ponceau		
Region	Total Density	Total Density	Razao protein/Actin	%	Total Density	Razao protein/Actin	%
Sal37	6,3325E+005	5,3383E+006	0,24960309	100	6,16E+06	0,13520201	100
Dox37	2,7707E+005	2,4509E+006	0,11304827	45,2912131	3,99E+06	0,06940805	51,3365521
Sal38	6,5290E+005	2,2335E+006	0,29232147	100	5,65E+06	0,11555548	100
Dox38	5,3512E+005	1,5405E+006	0,34736774	118,830731	3,53E+06	0,15152339	131,126099
Sal39	6,5742E+005	1,8283E+006	0,35957994	100	5,77E+06	0,11390602	100
Dox39	4,1665E+005	1,8695E+006	0,22286708	61,9798419	3,83E+06	0,10883421	95,5473696
Sal40	6,3707E+005	2,8552E+006	0,29317386	100	6,87E+06	0,12191879	100
Dox40	6,0912E+005	2,2727E+006	0,26801602	91,4187976	4,10E+06	0,14871094	121,975409
			#DIV/0!	#DIV/0!		#DIV/0!	#DIV/0!
			#DIV/0!	#DIV/0!		#DIV/0!	#DIV/0!
			#DIV/0!	#DIV/0!		#DIV/0!	#DIV/0!
			#DIV/0!	#DIV/0!		#DIV/0!	#DIV/0!

Acute OM+Rotenone

	Protein	Actin			Ponceau		
Region	Total Density	Total Density	Razao protein/Actin	%	Total Density	Razao protein/Actin	%
Sal43	7,45E+05	4,15E+06	0,1796687	100	5,1395E+006	0,14498297	100
Dox43	7,21E+05	3,66E+06	0,19685973	109,56818	6,1192E+006	0,11781279	81,2597382
Sal44	7,83E+05	2,86E+06	0,2740567	100	6,0540E+006	0,12933267	100
Dox44	8,19E+05	1,83E+06	0,44705368	163,124521	6,4884E+006	0,12616362	97,5496853
Sal45	7,13E+05	7,92E+05	0,90031963	100	5,4780E+006	0,13008945	100
Dox45	6,47E+05	1,17E+06	0,55454468	61,5941999	6,0457E+006	0,10697024	82,228224
Sal46	4,87E+05	1,53E+06	0,31810339	100	5,2667E+006	0,0924165	100
Dox46	6,46E+05	1,47E+06	0,43862125	137,88638	6,0576E+006	0,10671553	115,472374
Sal48	5,18E+05	1,90E+06	0,272907	100	5,8335E+006	0,08873746	100
Dox48	5,42E+05	1,02E+06	0,53307125	195,330736	6,2190E+006	0,12856127	144,878233
Dox47	5,73E+05	7,23E+05	0,79167646	290,090198	5,4749E+006	0,16477596	185,689286
			#DIV/0!	#DIV/0!		#DIV/0!	#DIV/0!

Acute OM+KCN

	Protein	Actin			Ponceau		
Region	Total Density	Total Density	Razao protein/Actin	%	Total Density	Razao protein/Actin	%
Sal49	7,54E+05	2,2818E+006	0,33040144	100	5,4417E+006	0,1385431	100
Dox49	7,2371E+005	4,0997E+006	0,17652755	53,428203	5,1299E+006	0,11806228	85,2170067
Sal50	6,3026E+005	2,4932E+006	0,33300979	100	5,8123E+006	0,14284535	100
Dox50	5,9399E+005	1,2350E+006	0,48096356	144,429258	4,5387E+006	0,13087228	91,6181595
Sal51	6,5634E+005	2,1753E+006	0,3017239	100	4,5444E+006	0,14442831	100
Dox51	6,7499E+005	2,0281E+006	0,43143336	142,989455	4,2511E+006	0,20582673	142,511348
Sal52	7,7981E+005	2,9773E+006	0,26191852	100	4,1756E+006	0,186754	100
Dox52	6,2265E+005	3,0199E+006	0,20618232	78,7200258	3,5591E+006	0,11200554	59,9749086
Sal53	7,3932E+005	2,2710E+006	0,32554822	100	5,5502E+006	0,13320601	100
Dox53	6,0511E+005	1,6863E+006	0,35883888	110,226031	4,2765E+006	0,14149655	106,223849
Sal54	6,3142E+005	1,6374E+006	0,38562355	100	4,7687E+006	0,13240925	100
Dox54	1,1664E+005	1,3192E+006	0,08841722	22,9283773	3,8347E+006	0,03041698	22,9719459

Figure 4.4: Spreedsheet tables of the protein information.

In summary, there are a total of **102 excel files**, containing the following information:

- 95 files with the heart pressure values;
- 1 file with the weight information, with 3 different sheets concerning the substrate, each with 6 tables regarding the inhibitors and the treatment;
- 3 files for the 3 different PCR plates of the transcript information, with different sheets concerning the genes, and different tables depending on the sample's perfusion combination;
- And, 3 files for the protein information, with one sheet per protein, and different tables with the correspondent samples.

Concerning the **control** rats, the same parameters were measured but in different conditions. There were two groups of control: Time Control (TC) and Nonperfused (NP), and each samples size is described in tables 4.1 and 4.2, respectively.

	NP		
	SAL	DOX	Total
Weights Information	-	-	-
Heart Pressure Values	-	-	-
Transcript Information	4	4	8
Protein Information	6	6	12

Table 4.1: Number of NP rats.

	TC							
	Glucose		Galactose + Glutamine		Octanoate + Malate		Total	
	SAL	DOX	SAL	DOX	SAL	DOX	SAL	DOX
Weights Information	4	2	2	2	2	2	8	6
Heart Pressure Values	4	6	4	5	4	6	12	17
Transcript Information	4	4	4	4	4	4	12	12
Protein Information	-	-	-	-	-	-	-	-

Table 4.2: Number of TC rats.

Concerning the TC rats, they were treated and perfused only with the substrates (four injections each) in order to understand if the hearts were capable of supporting perfusion without collapsing. Thus, there were additional 29 sheets concerning the heart perfusion measures and each sheet contains four tables, per injection similar to figure 4.2 table. Before perfusion the weights were measured and afterwards the transcript information was obtained. Both parameters were stored in files, per substrate, similar to the figures 4.1 and 4.3 respectively.

The NP rats were treated with either DOX and SAL solutions then they were sacrificed so that their protein and transcript information could be collected. Thus, there was no perfusion and therefore no heart perfusion values for these rats. The files concerning the parameters concerning this protocol are structured similar to figure 4.4 and 4.3.

4.2 Dataset Construction

In order to use the collected data, it is necessary to process and store it following a standard procedure. The final structure of the data should be simple and clear, each row should correspond to each sample, and the measures that explain the problem should match the columns. To ease the analysis categorical labels, such as **Treatment**, **Inhibitor** and **Substrate**, were transformed using the following mapping mechanism:

	Original Value	Transformation
Treatment	SAL	0
	DOX	1
Substrate	Glucose	1
	Galactose plus Glutamine (GG)	2
	OM	3
Inhibitor	Potassium Cyanide (KCN)	1
	Iodoacetate (IODO)	2
	ROT	3

Table 4.3: Treatment, **Inhibitor** and **Substrate** transformation.

To explore different data combinations and types of analysis, the data was selected and aggregated by: Weights, heart pressure, transcript information

and protein information. Concerning the aggregation by weights, there were several parameters taken from the Weight files described in the previous section and represented in figure 4.1. Advised by the experts at the MitoXT laboratory, the measurements regarding the substrate and recuperation flows were excluded due the presence of a considerable amount of missing values, and the RNALater weight feature, concerning the weight of heart tissue stored in RNAlater, was also rejected since it was not relevant for the problem in hands.

These values were used to create two different datasets. One dataset joining each sample weights information for the TCs, and another for the main dataset, which combines all features (columns) for all samples (rows), for the acute protocol, as seen in figure 4.5.

ID	Treatment	Inhibitor	Substrate	Initial Weight	Final Weight	Heart Weight	Tibia Size	
33	0	3	1	386	384	1,233	3,8	...
33	1	3	1	377	373	1,015	3,7	
...								

Figure 4.5: Structure of the TC weights dataset and the beginning of the main dataset which continues to figure 4.6.

With the help of the experts at MitoXT laboratory, four additional features were selected to be part of the exploration of possible relationships and tendencies between features. Two of the additional features were obtained by subtracting each rat's weight (equations 4.1 and 4.1):

$$WeightDifference = Finalweight - InitialWeight \tag{4.1}$$

$$AbsWeightDifference = |FinalWeight - InitialWeight| \tag{4.2}$$

Both features goal was to understand the magnitude of weight variation after each protocol and each treatment. Thus, the **Weight Difference** feature allowed us to understand if the weights decreased or increased, and was only included in the weights dataset. The **Absolute Weight Difference** feature was only included in the main dataset (figure 4.6). The other two features resulted in the division of the heart weight and the tibia size by the final weight of each rat.

The addition of these features had the purpose of understanding if the treatment had a direct or indirect influence in the weights of the rats, and if this influence happened for all protocols. By dividing the initial features by the final weight, the resulting values will represent the variation of that feature and can then be compared with the other rats results.

...	Abs Weight	Heart weight / Final weight	Tibia size / Final weight	...
	2	0,00321	0,00989	
	4	0,00272	0,00992	
	...			

Figure 4.6: Addition of features *Abs Weight*, *Heart Weight/Final Weight* and *Tibia Size/Final Weight*. Continues to table showed in figure 4.7.

Concerning the aggregation by Heart Pressure values we built several datasets following two different structures. The first structure was built in order to visualize the time series tendencies, without any statistic transformation, of the heart pressure values during perfusion with different inhibitor concentrations. Thus it combines the ID, Treatment, Inhibitor, Time, Heart Pressure and Injection number for each sample (figure 4.7).

ID	Treatment	Inhibitor	Time	Heart Pressure	Injection
33	0	3	0.000	1.6040	1
33	0	3	0.008	2.0720	1
33	0	3	0.016	4.0063	1
...					

Figure 4.7: Structure of the 3 datasets corresponding to heart pressure values timeseries. Dimensions: 816137 rows x 6 columns, for each substrate.

It is necessary to transform the data, in order to correlate the heart pressure with the remaining features. Thus, the mean, median and standard deviation were calculated for each injection value, and for all values combined, resulting in 3 additional features for each situation. These new features were then added to the dataset joining all features (table of figure 4.8).

...	Injec1_ Std	Injec1_ Mean	Injec1_ Median	...	InjecALL _Std	InjecALL _Mean	InjecALL _Median	...
	12.581	9.4608	2.4152		5.662053	2.520833	1.0424	
	2.5528	2.6037	1.2296		1.519756	1.595467	1.1048	
...								

Figure 4.8: Continuation of figure 4.6. Addition of calculated features concerning the Heart Pressure values' mean, median and std. Continuation to figure 4.9

Additionally, concerning the transcript information, aggregation the ANT, Hif-1*alpha* and LDH transcript expression values were selected and those which were from rats perfused with both substrate and inhibitor were added as features to the main dataset. The values from TC rats were joined in a new dataset: TC transcript expression information dataset. The transcript information of NP rats was also joined in a different dataset.

...	ANT_ Exp	HIF_ Exp	LDH_ Exp	...
	0.812532	0.681235	1.080604	
	0.913010	0.999227	1.15686	
...				

Figure 4.9: Continuation of figure 4.8. Addition of transcript information selected values. Ends in figure 4.10

Finally, protein information related to PGC-1*alpha* and TFAM from rats perfused with both substrate and inhibitor, completed the main dataset. The Ubiquitin protein information was not enough to be considered for the analysis, thus it was excluded. In addition, the protein values of the NP rats were joint in a new dataset: NP rats protein information dataset.

...	PGC-1alpha_ratio	TFAM_ratio
	0.111244	0.302348
	0.370401	0.200717
...		

Figure 4.10: Continuation of figure 4.9. Addition of protein information features. End of main dataset structure.

4.3 Data Cleaning

After carefully selecting and structuring the data, it is necessary to go through each component to understand the correct transformations that need to be applied to attain the best model performance and to maximize the information gain.

As mentioned in Chapter 3, data cleaning is a very important step, considering each value and its accurateness towards the problem's context. Thus, in this section we will be analysing our features possible missing values and outliers.

4.3.1 Missing values

The first phase of this process was to understand whether incomplete information exists. A general analysis during data selection showed that the protein information had very few values concerning the ubiquitin ratio. Therefore, this was not considered as a feature.

However, concerning the features that were actually selected and did not presented obvious missing information, there were other missing values that

needed to be found and treated accordingly. Thus, we performed a careful analysis to each feature.

The first analysed feature was the Heart Pressure values due to its large amount of values. Our analysis revealed that there were about 0.1% of information missing, thus we replaced it by the feature mean.

The ANT trasncript expression feature, from the TC transcript information dataset, had a missing value corresponding to a SAL-treated rat injected with GG substrate. In table 4.2, it is possible to confirm that this feature only has 4 rats. Thus, applying imputation to replace the missing value would avoid removing the sample, although it would also have great influence in our visualization of the problem. The final decision was to remove the sample from the dataset in question.

4.3.2 Outlier Detection

Outlier detection plays an essential role in detecting abnormal observations of data values. However, in a biological dataset, such as the one in hands, outlier values are frequent and may be relevant to understand the problem. Thus, they were not removed.

Chapter 5

Results and Discussion

5.1 Exploratory Analysis

In order to find possible hidden patterns in the data, an exploratory analysis was made.

Looking for new information, we applied several data visualization techniques, different descriptive analysis parameters, time series analysis, feature correlation and mutual information analysis to our datasets.

5.1.1 Descriptive analysis and Data Visualization

The first set of features to be analyzed were the measures taken before perfusion using the dataset concerning the Time Control (TC) rats information and the main dataset information.

As previously mentioned, the main goal of the TC protocol was to understand whether the hearts could survive through the entire perfusion procedure. Thus, this protocol hearts were only perfused with the substrate.

5.1.1.1 Pre perfusion

The first step was to perform a descriptive analysis of each feature using the metrics described in section 3.2 of chapter 3 and which values are shown in supplementary table A.1.

Our initial analysis of the weights for all rats showed that there are no considerable alterations between the initial weight and the final weight of the rats of each treatment, which matches the original conclusions, figure 5.1 left panel.

However, as mentioned previously, we created two additional features in order to perform a deeper exploration of the existent alterations and decided to plot the same results as before considering these variations.

In fact, the graph on the right of figure 5.1 shows that while Saline (SAL)-treated rats weight variation fell under +/- 4 grams, Doxorubicin (DOX) injected rats weights variation reached 10 or more grams.

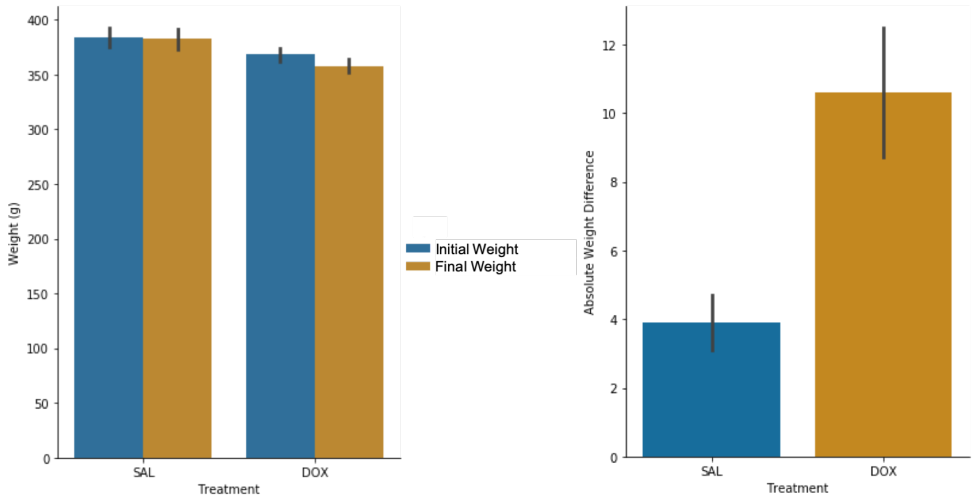


Figure 5.1: Barplot analysis of initial weight vs the final weight per treatment (left). Barplot analysis of the absolute difference of the weights (right).

A boxplot analysis of the weights difference (figure 5.2) not only confirmed that the change but it also indicates the the DOX-treated rats weight variations for most cases reveals a decrease of the weight reaching up to a minus 25 grams variation.

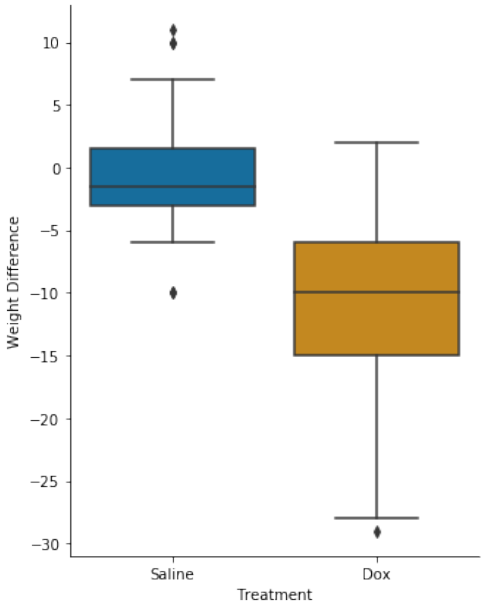


Figure 5.2: Boxplot analysis of the weight difference feature.

Additionally, figures A.1 and A.2 of Appendix A, show a boxplot analysis of *Heart weight / Final Weight*and *Tibia Size / Final Weight* results, respectively. None of the figures show noticeably differences between DOX and SAL groups.

5.1.1.2 Post perfusion: transcript information

The original analysis, done by Filipa Carvalho, suggested that in Nonperfused (NP) hearts from rats treated with DOX, Adenine nucleotide translocator (ANT) expression values were significantly reduced compared to the ones treated with SAL. This analysis also concluded that this feature presented higher values in hearts perfused with glucose and decreased for the ones perfused with Galactose plus Glutamine (GG), concerning the TC protocol, without discriminating treatments.

Regarding the addition of inhibitors, the original analysis showed that Hypoxia-inducible factor 1*alpha* (Hif-1*alpha*) expression values were higher for DOX-treated hearts perfused with glucose and Iodoacetate (IODO) inhibitor. Both GG and Octanoate plus Malate (OM) substrates combined with cyanide inhibitor had higher ANT expression values concerning DOX-treated rats. Finally, Lactate dehydrogenase (LDH) expression of DOX hearts only increased in the presence of the substrate OM combined with inhibitor Potassium Cyanide (KCN).

In order to explore these features and confirm the original conclusions, we used three datasets: the NP rats transcript information dataset; the TC rats transcript information dataset and the main dataset.

Table A.3 (see appendix) and figure 5.3 contain the results regarding the ANT expression of TC hearts and of the hearts perfused with both substrates and inhibitors. The NP hearts transcripts expression descriptive information is displayed in table A.2 of the appendix.

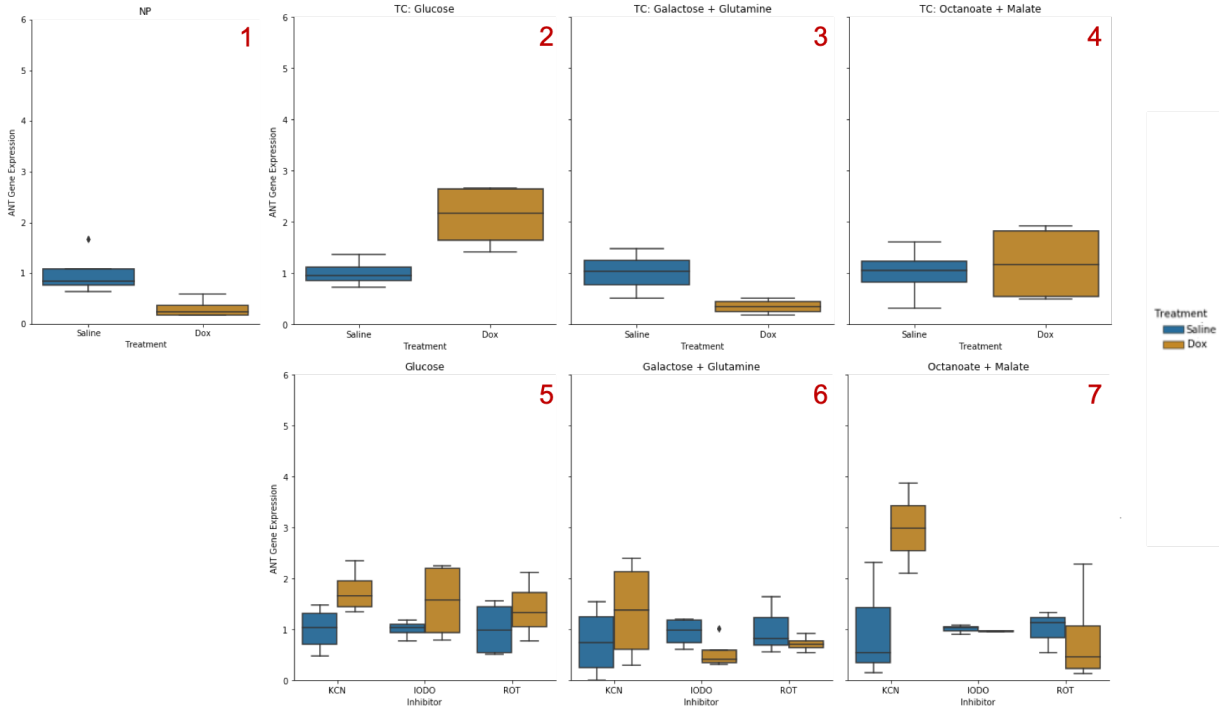


Figure 5.3: Boxplot analysis of the ANT transcript expression feature.

Regarding the NP hearts, Panel 1 shows that the SAL-treated group had higher values compared to the DOX-treated group, which had already been observed in previous studies.

Looking at the boxplot analysis of the TC rats' values (graphics 2, 3 and 4) showed that, similar to the original, this tendency was maintained solely for the rats perfused with the substrate GG (panel 3). As for the glucose substrate (panel 2), the DOX group showed a great increase, exceeding the one observed for SAL, which was also observed by the previous studies. The remaining substrate showed no considerable alterations between treatments.

Finally, panels 5, 6 and 7 represent the analysis of the hearts also perfused with inhibitor. Panel 5, corresponds to glucose results and shows that, independently of the inhibitor, DOX group values were higher compared to the SAL ones maintaining the TC results. Thus, the ANT expression results were not affected by glucose inhibition, which was already confirmed by the original analysis.

For the remaining substrates, the results concerning the KCN group showed a substantial increase of DOX-treated hearts values, especially in the ones of the OM substrate, compared to TC protocol, also matching the original analysis.

Thus, our ANT expression results analysis matched the original analysis.

Proceeding to analyzing the **Hif-1***alpha* results, table A.4 and figure 5.4 must

be considered.

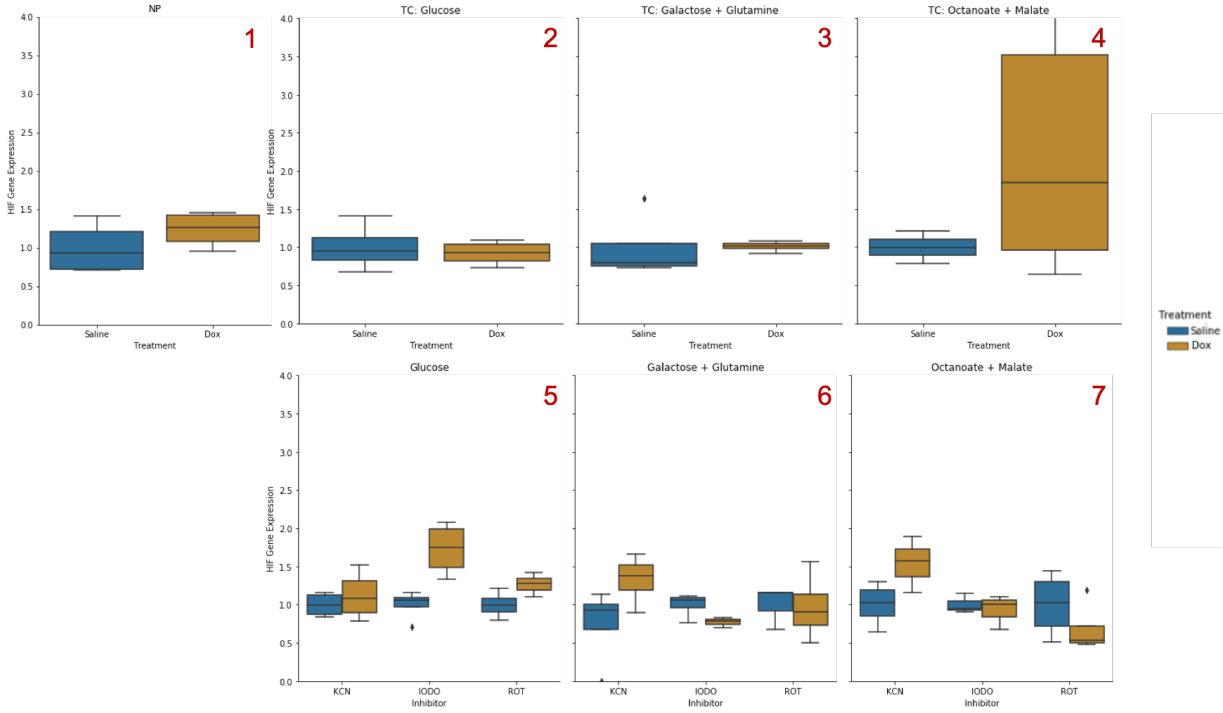


Figure 5.4: Boxplot analysis of the Hif-1*alpha* transcript expression feature. The blue boxplots correspond to SAL group and the yellow boxplots to the DOX one.

NP hearts (panel 1), showed no differences of this feature values concerning the DOX-treated hearts and compared to the SAL group. However, this is only true regarding TC hearts perfused with glucose or GG substrates, for panels 2 and 3, where the values for both treatments prevailed very similar, which matched the original analysis. One the other hand, concerning the OM substrate, the results suggest a great increase of the DOX group values (table A.4, panel 3), which was not observed by the previous studies.

Concerning the hearts also perfused with inhibitors, IODO inhibitor displayed higher values for the DOX group when perfused with glucose, surpassing the SAL group, TC and NP protocol results and confirming the original analysis, which stated that this result supported the 'idea that DOX hearts promote adaptations in glycolysis'[6].

Additionaly, the rats perfused with the remaining substrates and combined with KCN inhibitor, showed a great increase for the DOX-treated hearts for the OM+KCN substrate boxplot. Regarding the OM inhibition, the Rotenone (ROT) boxplot showed a considerable decrease in DOX-treated group, compared to all the remaining groups. Both of these analyses were not considered in the original research.

To finalize our transcript information analysis, figure 5.5 contains the boxplots

graphs regarding the **LDH expression** results and table A.5 its descriptive analysis.

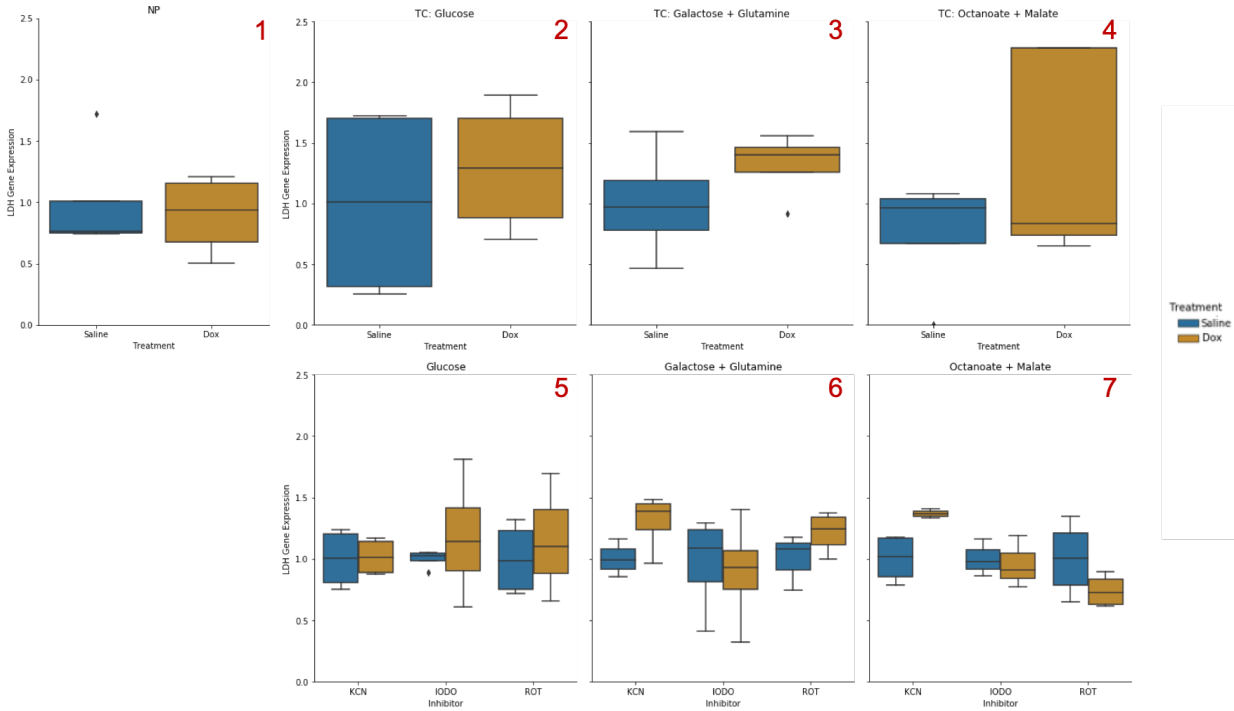


Figure 5.5: Boxplot analysis of the LDH transcript expression feature.The blue boxplots correspond to SAL group and the yellow boxplots to the DOX one.

The boxplots concerning NP hearts (panel 1) suggest that LDH expression had no alteration between treatments, although the DOX group has greater variance when compared to SAL. This effect was already observed in the original study. .

The same happened for panels 2 and 4 of the TC rats perfused with glucose and OM, respectively, but not for panel 3, regarding the GG substrate. This substrate showed an increase in the DOX treated group results, when compared to the NP panel and the SAL group. The original conclusions suggest that for GG, none of the TC results show considerable differences.

Observing all graphs, only combination OM plus KCN showed a considerable increase in the DOX hearts results. Despite both ROT and KCN inhibition with GG substrate also showing a great increase, the variance of the values is too high, compromising our conclusions. The same happened for the combination IODO+G results.

In the end, all transcript information feature values had similar responses for the same protocols and compared to the original analysis. Our results for the NP rats matched the original analysis, since ANT expression results also were substantially lower for DOX hearts, whilst LDH had similar values and Hif-1*alpha* expression had a mild increase.

Concerning the ANT feature, the original results also showed a significant increase in the DOX hearts group, regarding all combinations of the glucose perfusion, and regarding the KCN combination with the remaining two substrates.

As for the Hif-1*alpha* results, we also concluded that perfusion with the pairs G+IODO and GG+KCN or OM+KCN showed higher values for DOX treated hearts, which the previous studies stated as the support of the 'idea that DOX hearts promote adaptations in glycolysis'[6]. To conclude, we added that the OM inhibition with ROT showed an abnormal decrease in DOX results, compared to all the remaining graphs.

Finally, regarding LDH feature, the results were also similar to the original ones, since we also found an increase of the DOX hearts LDH expression for the KCN inhibition of the OM substrate, which the original analysis also concluded that this showed 'that KCN promotes susceptibility in DOX hearts including toxicity'[6].

5.1.1.3 Post perfusion: protein information

For this analysis we considered two datasets: the NP rats protein information dataset and the main dataset. Each feature was analyzed individually and afterwards these conclusions were compared to the original ones.

Figure 5.6 contains the graphs related to the Peroxisome proliferator-activated receptor-gamma coactivator (PGC-1*alpha*) ratio boxplot analysis and tables A.6 and A.7(appendix) the correspondent values.

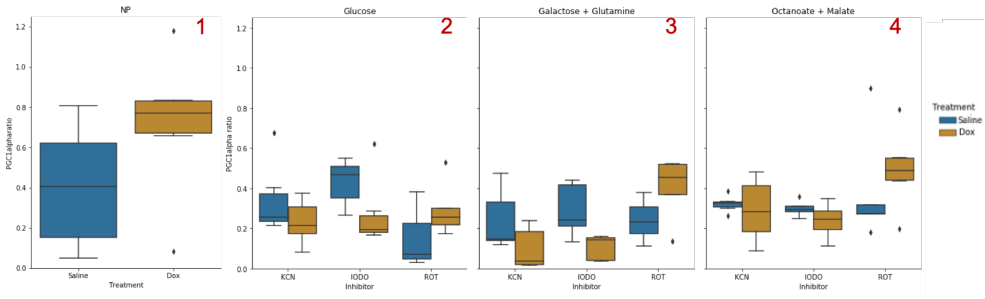


Figure 5.6: Boxplot analysis of the PGC-1*alpha* ratio feature. The blue boxplots correspond to SAL group and the yellow boxplots to the DOX one.

NP hearts of rats treated with SAL solution showed lower values when compared to the ones treated with DOX and had greater variance (panel 1). This contradicts the original analysis, which stated that, for both Mitochondrial transcription factor A (TFAM) and PGC-1*alpha*, DOX hearts results were lower.

We did in fact confirmed, however, that for DOX treatment group all perfused hearts results decreased comparing to the to the NP results. Which, the original analysis conclued as a prove for the 'down-regulation effect of the inhibitors on protein content'[6], it also made all the suggestions for both PGC-1*alpha* and TFAM ratios generalized for all inhibitors.

However our figure shows that there are some considerable differences for hearts perfused with KCN and IODO inhibitors, which had higher values for the rats treated with SAL solution (panels 2, 3 and 4), particularly concerning the substrate GG perfusion, where the DOX boxplots for these inhibitors had really low values, contrary to the NP results conclusion. Thus, we conclude that these the inhibitors which more contribute for the 'down-regulation on protein content'[6].

On the other hand, the rats whose hearts were perfused with ROT inhibitor showed an increase of the PGC-1*alpha* ratio values of DOX-treated hearts compared to the ones treated with SAL, independent of the substrate.

Finally, the TFAM ratio values boxplot analysis can be observed in figure 5.7.

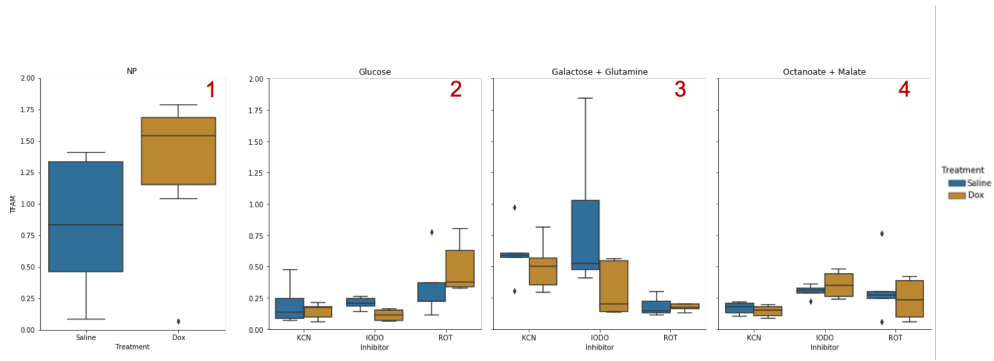


Figure 5.7: Boxplot analysis of the TFAM ratio feature. The blue boxplots correspond to SAL group and the yellow boxplots to the DOX one.

The first panel of this figure suggests that the NP hearts, whose rats were treated with saline solution, had lower values, although more separated, compared to rats treated with DOX, contradicting the original analysis.

However, the remaining graphs suggest that this feature values are lower for the perfused hearts compared with TC, which is confirmed by the descriptive analysis, A.7, suggesting the perfusion caused a great stress concerning this feature. Confirming the original analysis, which did not consider each inhibitor individual analysis.

For both substrates glucose and GG, DOX-treated hearts TFAM values decreased considerably in the presence of IODO. Besides this analysis, there were no other relevant conclusions to be taken, since the panels 2, 3 and 4 do not show any pattern concerning the substrate and inhibitors comparison.

In the end, we verified almost all original conclusions. Additionally, our analysis also suggest a great decrease of both features results concerning the perfusion protocol compared to NP protocol. Additionally, within the NP protocol we do not consider that DOX hearts results were lower than SAL ones. We also discovered that for both PGC-1*alpha* and TFAM results, IODO and KCN inhibiton showed the lowest values for the DOX group, specially IODO, contributting greatly for the 'down-regulation on protein content'. Additionally, we also consider that PGC-1*alpha* values from hearts perfused with ROT were greater for DOX hearts compared with the SAL-treated ones.

After analysing pre and post perfusion features it is time to consider the measurements taken during this procedure.

5.1.1.4 Heart Pressure Values - Time Series Analysis

In chapter 2 section 2.4, it is mentioned that at the end of each injection, the heart pressure values were measured for 30 seconds and recorded. The measurement resulted in a set of features with about 3k values per injection, per sample. Hence, in order to explore this feature, we first decided to analyse it without applying any kind of statistical transformation.

Thus, we plotted both DOX and SAL values for each substrate of the TC protocol (figure 5.8), and of the main dataset without discriminating inhibitors (figure 5.9). Lighter color lines represent the SAL values whilst darker colors represent DOX values.

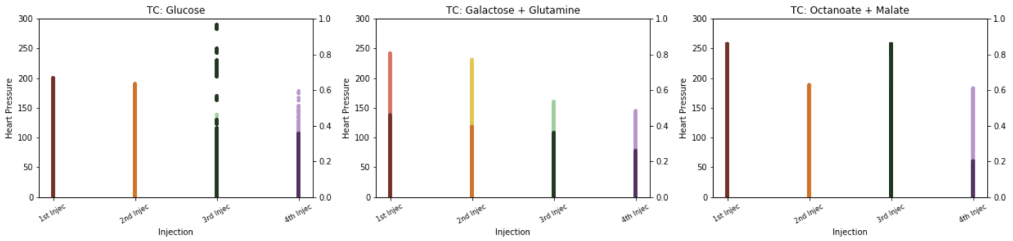


Figure 5.8: Heart Pressure values per injection, for DOX and SAL group for each substrate of the **TC** protocol.

Figure 5.9: Heart Pressure values per injection, for DOX and SAL group for each substrate of the **main dataset**.

In fact, it is possible to see differences between the treatment values. Since the main goal of the TC protocol was to understand if the hearts were capable of tolerating the injections, figure 5.8 confirms that they can. It also shows that using OM as substrate, the DOX-treated group suffered a considerable functional decline, which was already stated in the conclusions of the original thesis.

However, contrary to the GG conclusions, where the authors stated that no differences were found between groups, panel 2 of figure 5.8 shows that the DOX-treated hearts presented lower values compared to SAL ones. As for G substrate TC results, the original analysis concluded that the first injections showed higher values for the DOX group, which is also suggested in panel 1 of figure 5.8.

For the main dataset result, the OM perfusion graphs suggest that both groups values decreased considerably compared to TC outcomes of the same substrate (panel 3 figure 5.9).

However, for the remaining graphs there are no clear differences nor tendencies between treatments, protocol or injection, thus it is necessary to approach the problem differently.

Thus, the next step was to plot the same graphs of the main dataset but discriminating the inhibitors (figure 5.10) and also, in order to explore the **time series** tendencies for this measure, plot the graphs considering the time column for the x axis and not the injections.

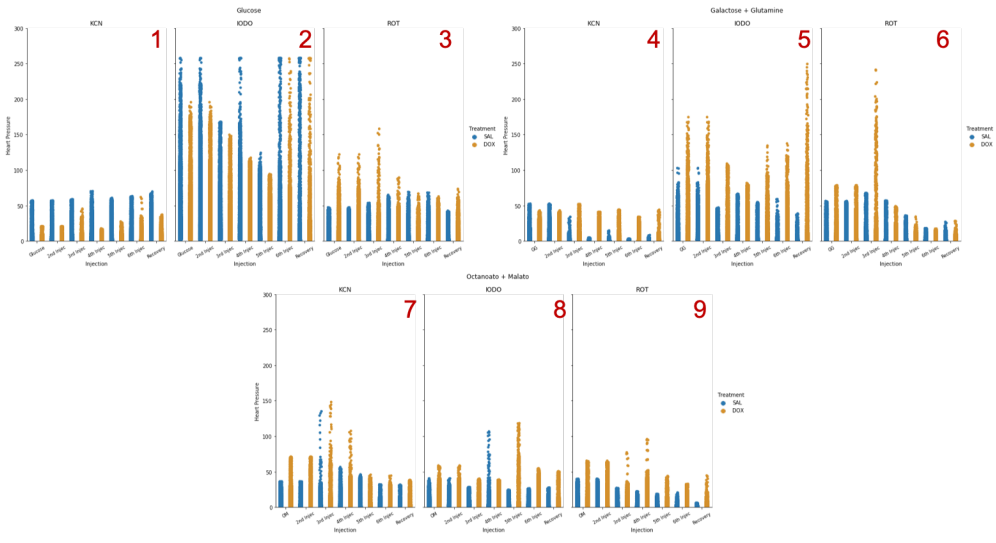


Figure 5.10: Heart Pressure values per injection, for DOX and SAL group for each substrate of the **main dataset**. The blue lines correspond to the SAL group and the yellow line to the DOX one.

In chapter 3.5, it is stated that 'a classical decomposition of the time series is to use a Moving Average (MA) method to estimate the trend-cycle'. In fact, figure

5.11 shows the same values raw, left graphic, and using MA transformation, right graphic, show that MA allows us to have a better understating of the overall time series tendencies. Thus, we applied this transformation to our data and plotted the time series for each injection, per sample pair, figures A.9 to A.58 of the attachments.

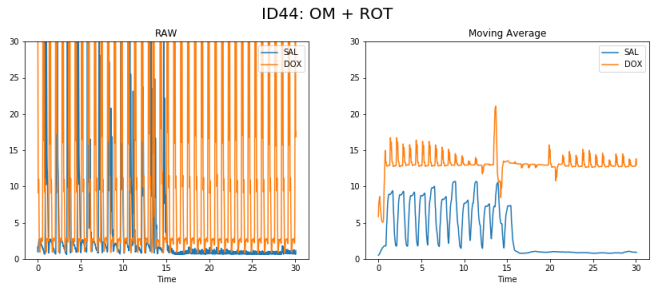


Figure 5.11: Heart Pressure time series plot for injection 1, with MA transformation(left) and without(right).

After analyzing figure 5.10, concerning the general results for each combination, and supplementary figures A.9 to A.58 concerning each pair time series tendency, some conclusions were taken in comparison to the original thesis analysis of the heart rate values taken from the same heart pressure values we are using.

Regarding the glucose perfusion, the original analysis concluded that the heart rate decreased when the IODO concentration increased, however panel 2 of figure 5.10 shows this only happens for injections 1 to 5, since the 6th injection and the recovery phase show some increase in these values for both SAL and DOX groups. This increase is explained by supplementary figures A.12 and A.14, concerning the pairs 4 and 6 of SAL and DOX hearts, perfused with glucose substrate combined with IODO inhibitor, and where both groups recovered.

As for the remaining two inhibitors, KCN and ROT, panels 1 and 3 do not exhibit a generalisation of this tendency, since the time series analysis of each ID pair showed different responses within the same combination.

As for panels 4, 5 and 6, concerning the substrate GG perfusion with KCN, IODO and ROT, respectively, the original analysis concluded that, generally, for IODO and KCN SAL group heart pressures were more affected during perfusion.

Panel 4 suggests that SAL hearts almost did not survive, whilst DOX hearts retained consist heart pressures throughout the procedure, which is supported by the individual plots, since none of the SAL group hearts for the pairs corresponding to the KCN and GG combination recovered (supplementary figures A.28 to A.32). IODO values were the ones reaching higher values for DOX-treated hearts, also confirmed by each pair plot (supplementary figures

A.33 to A.38), in which almost every DOX sample showed recovery, contrary to SAL hearts.

The last three panels, 7, 8 and 9, represent the OM perfusion with the inhibitors KCN, IODO and ROT, respectively. The original analysis concluded that none of the SAL or DOX hearts recovered for the IODO and ROT groups. Appendix figures A.44 to A.58, show that this is only accurate for the ROT results, since 2 of the 4 IODO pair plots showed recovery in DOX-treated hearts.

Looking at the results from a more generalist perspective, some conclusions could be taken. For example, the majority of samples heart pressure values being lower when injected with inhibitors compared to the TC results. Concerning the recovery graphs for the different substrates, only glucose perfused hearts showed significant recovery rates, opposite to OM substrate, for which more than half the SAL and DOX hearts did not recover, and the GG perfusion scenario where graphs show that only 2/14 SAL and 8/17 DOX hearts recovered. Finally, regarding each inhibitor's group recovery, ROT hearts were the ones which performed worse, since only 6/17 SAL and 7/17 DOX hearts recovered. As for KCN perfusion almost all hearts recovered, specially the ones injected with glucose substrate (13/13).

Overall, the analysis present great disparities between samples of the same combination, compromising any general conclusions. Thus, in order to improve the results and conclusions, an increase of the experimental sample size is required.

5.1.2 Correlation

The next step of your exploratory analysis was understanding if our features were related. Hence, we decided to do a correlation analysis, which helps us to discover linear relationships between features and their influence in the treatment. This will later help us in the development of an Machine Learning (ML) model to distinguish between DOX-treated rats and SAL-treated ones.

This analysis was divided in different phases. The first phase was to understand the features relationship with the treatment, giving information about which would help our classifier to distinguish between groups. So that then we could analyze those features relationships with the remaining ones and perhaps find new patterns.

Regarding the Glucose group, supplementary figure A.3 (panels 1 and 2) contains all features correlation analysis, including the treatment, for 4 different groups: DOX group, SAL groups, DOX-SAL group values and for groups combined. Figure 5.12, (panels 1, 2, 3) contains the ones which had the most interesting results, divided by inhibitor, and a 4th one concerning all glucose perfused data.

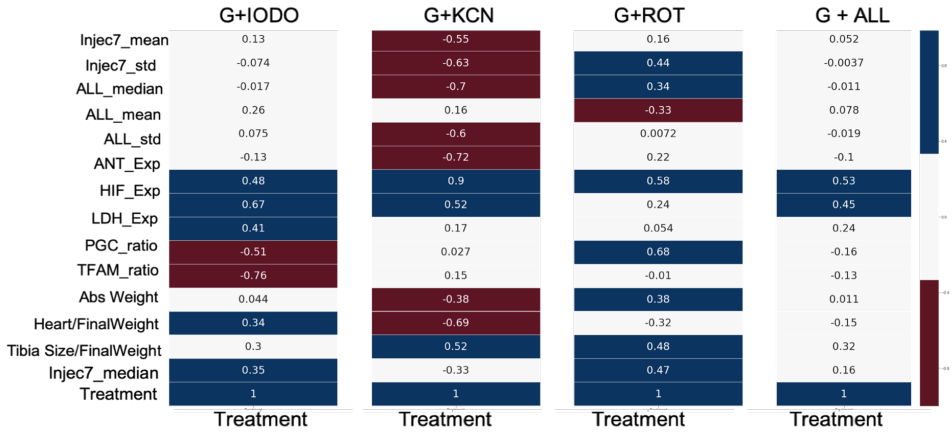


Figure 5.12: Glucose group features correlation analysis regarding the Treatment.

From the last panel of figure 5.12, one can conclude that the Hif-1*alpha* and the ANT expression features are those which will better contribute to distinguish SAL from DOX samples. For panels 1,2 and 3, the ANT feature shows a strong relationship with the Treatment, especially for the KCN inhibitor results (0.9), panel 2. As for the Hif-1*alpha* feature, the relationship is only strong for inhibitors IODO and KCN.

Since, ANT and Hif-1*alpha* expression were the features which showed better relationship with the treatment, thus it is also important to understand which other features are related to them. Thus, we need to analyse figure 5.13, which contains the correlation information for the selected features concerning the SAL-treated group, first panel, and the DOX group, second panel.

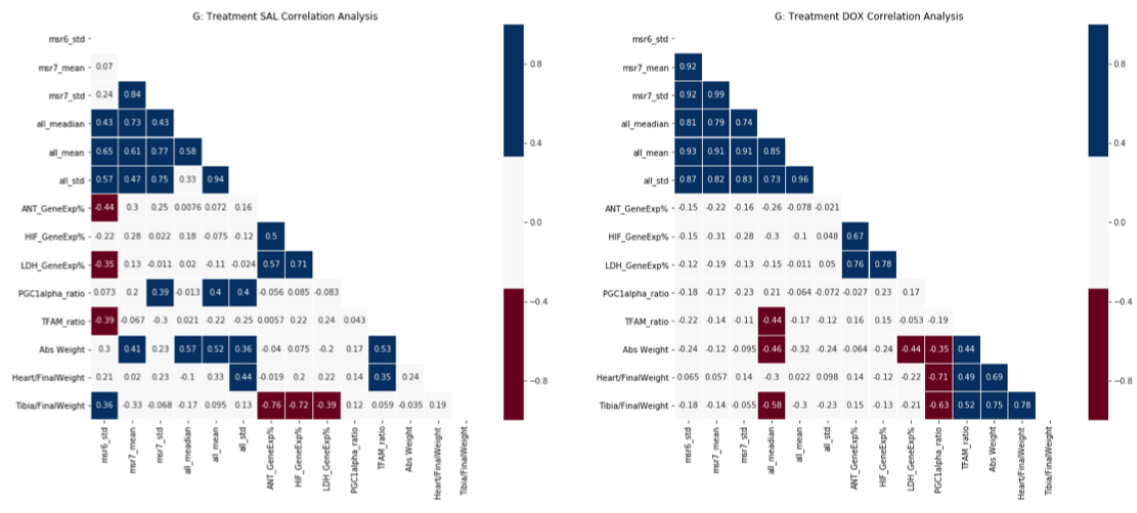


Figure 5.13: Glucose group features correlation analysis regarding the SAL and DOX treatments.

A first look at the graphs suggests that these features show different relationships between SAL and DOX groups. In panel 2, concerning the SAL group,

there is a negative correlation between all transcripts expression features and the Tibia size divided by the final weight, contrary to the results shown for the DOX group (graph 2), which shows no correlation at all.
Additionally, these graphs also show different correlations concerning the PGC-1*alpha* ratio feature with the weights features, since it showed a negative correlation in the DOX group and no correlation for the SAL group

Using the same logic as before, figure A.4 shows the results of the correlation analysis with the GG substrate, figure 5.14 the same results but selecting the most interesting ones regarding the Treatment, and figure 5.15 the correlation information concerning the SAL-treated group and the DOX group.

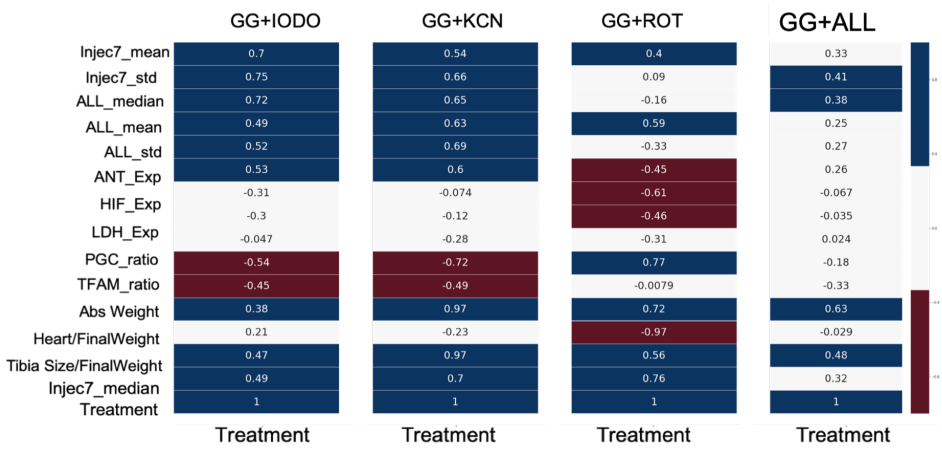


Figure 5.14: GG group features correlation analysis regarding the selected features.

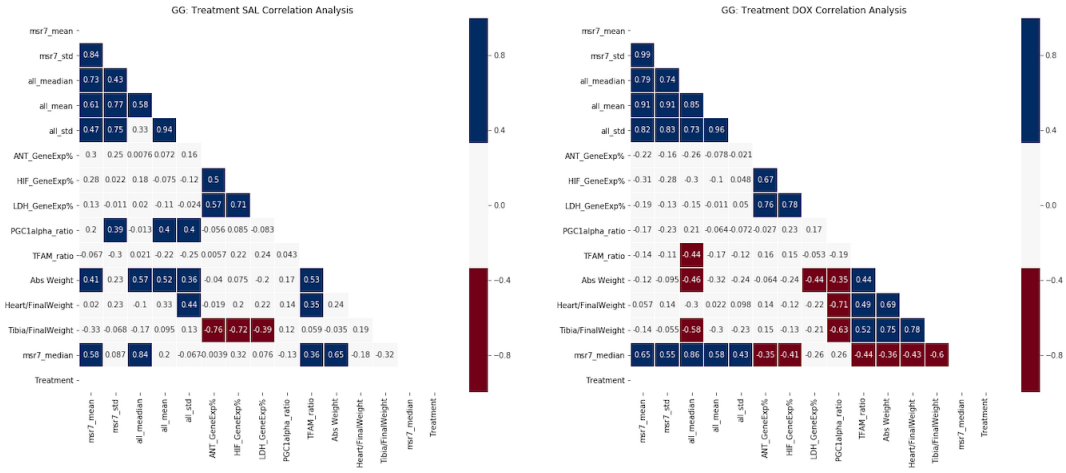


Figure 5.15: GG group features correlation analysis regarding the SAL and DOX treatments.

Analyzing panel 4 of figure 5.14, disregarding the inhibitors, we concluded that the features with considerable correlation with the Treatment are Absolute Weight Difference (0.68) and the Tibia size over the final weight feature (around 0.5).

All inhibitors show some correlation between the Treatment and the Abs Weight. For both ROT and the KCN, this correlation is considerable and almost perfect for the KCN (0.72 and around 1.00, respectively).

Concerning the remaining features, panel 3, concerning the ROT inhibitor, shows interesting contradictions to the other two inhibitors, since PGC-1*alpha* ration has a strong positive correlation (+0.77) with the treatment, whilst for IODO and KCN it has a considerable negative relationship (-0. 54 and -0.74, respectively), which also happens for the TFAM ratio feature results. Additionally, ROT's group also show a near perfect negative correlation concerning the Heart Weight division by the final weight and the Treatment, and a considerable negative correlation with both LDH and Hif-1*alpha* expression features (around -0.5 and -0.61, respectively), whilst the remaining inhibitors groups had no correlation at all.

Regarding the feature's relationships per treatment, the results showed great differences between the SAL group and the DOX group (figure 5.15). Again, if we consider the values for PGC-1*alpha*, the DOX group has negative correlation between this feature and all weights features, whilst SAL group has no correlation. For the TFAM ratio feature the opposite happens since it has a positive correlation with the Weights features, which means that TFAM could also be valuable for classifying DOX samples.

Last but not least, the OM all features correlation is presented in figure A.5, figure 5.16 the features concerning the Treatment, and figure 5.17 regarding DOX and SAL groups differences.

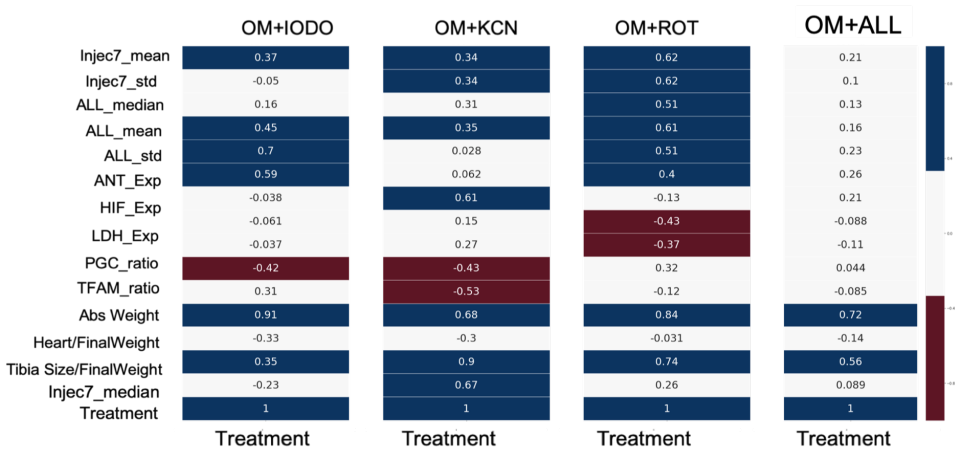


Figure 5.16: OM group features correlation analysis regarding the SAL and DOX treatments.

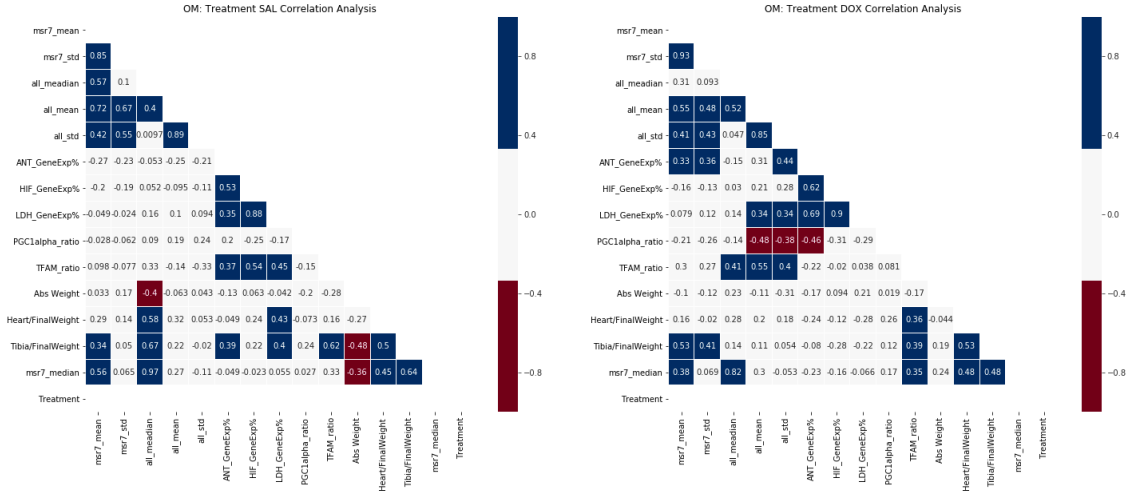


Figure 5.17: OM group features correlation analysis regarding the SAL and DOX treatments.

Figure 5.16 displays very similar results to the previously analyzed substrate. The same features, Abs Weight and Tibia size divided by the Final Weight, show considerable correlation with the treatment, panel 1. This remains true for all inhibitors, panels 1, 2 and 3.

Again, the IODO and KCN plots, panels 1 and 2 respectively, show negative relationships between PGC-1*alpha* and the Treatment, although not as strong as the GG+IODO and GG + KCN results. This inhibitors' group also show a near perfect negative correlation concerning the Heart Weight division by the final weight and the Treatment, whilst the reaming inhibitors groups had no correlation at all.

Concerning the transcripts information, the Hif-1*alpha* expression feature also shows some negative correlation for the ROT' group, and none for the remaining ones. In addition, the G+IODO panel shows considerable positive relation between ANT transcript expression and the Treatment, whilst for the remaining substrates this feature has no correlation at all with the Treatment.

As for figure 5.17, concerning the features relationships differences between treatments, the results are not very interesting compared to the remaining substrates.

Figure 5.16 considerations also did not add relevant information to our analysis, because they were very similar to the ones already made, this substrate protocol should be disregarded from the experience, saving time, money and animals that could be used to increase the remaining protocols samples size. This decision was also taken in the original thesis. Additionally, we also considere that, with more samples, we could prove that only one perfusion protocol is needed, the glucose protocol, since it is the one providing more relevant information.

5.1.3 Mutual Information

Previously we analysed how our features are linearly related by applying Pearson's correlation and plotting each graph correspondent to the differences between treatments results and each feature correlation concerning the treatment.

Then, it was necessary to consider nonlinear relationships, thus we decided to do a Mutual Information (MI) analysis. Supplementary figures A.6, A.7 and A.8 correspond to each substrate MI analysis. Each figure has 3 graphics, the first two represent the MI results for DOX group and SAL group, respectively, and remaining graphic MI results concerning all samples.

None of the graphics show relevant differences between DOX and SAL group results. And for substrates GG and OM there is also no MI between the Treatment and the remaining features.

Although, figure A.6 graphic 3, show a slight increase of MI results between the transcript information and the Treatment (around 0.5), though, in our opinion, not relevant enough.

Through this section, it has become clear that since the available features reveal different information for each treatment, it should be possible to distinguish between DOX and SAL-treated samples. Next section we will be analyzing two different classification models and their evaluation.

5.2 Classification

After a careful analysis of all features individually and their relationship with each other, we decided to analysed the discriminative capacity of the features available, by designing an ML model to distinguishing between DOX-treated samples and SAL ones.

We used two different algorithms: Decision Tree (DT) and Random Forest (RF). Our data was split into train and test datasets. To find the optimal combination of parameters, we used the GridSearchCV function, available in the python library sklearn, and applied it to our training data.

GridSearchCV stands for grid search cross validation. One of this funtion' input parameters is the estimator, where we can define the model to be implemented. Depending on the selected model, we can vary the kernel options to be tested by defining them in the param_grid. The param_grid is a list of all the model parameters we want to test. It uses K-fold cross validation in order to determine the hyper parameters values set that provide the best accuracy levels for our model.

From the GridSearchCV application, we configured our models to the following parameters.

Random Forest

- n_estimators = 10, which represents the number of trees in the forest;
- max_depth = 4, which represents the depth of each tree in the forest;
- criteria = 'entropy', where the supported criteria are 'gini' for impurity and 'entropy' for information gain. These methods are used to select which attribute would be placed at the root node or the internal node.
- max_feature = 'auto', meaning the number of features to consider when looking for the best split is the square root of our features number.

Decision Tree

- max_depth = 4, representing the depth of the tree;
- min_samples_split = 50;
- criteria = 'gini', criteria used to select which attribute would be placed at the root node or the internal node.

After choosing the optimal parameters combination for each model, we applied them to the test dataset. Figures 5.18 and 5.19 contain the confusion matrices of each algorithm outcome and tables 5.1 the results of the evaluation metrics.

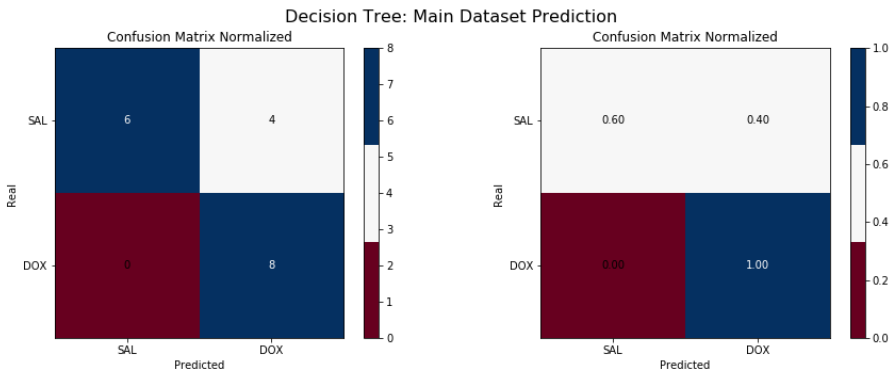


Figure 5.18: RF Confusion Matrix.

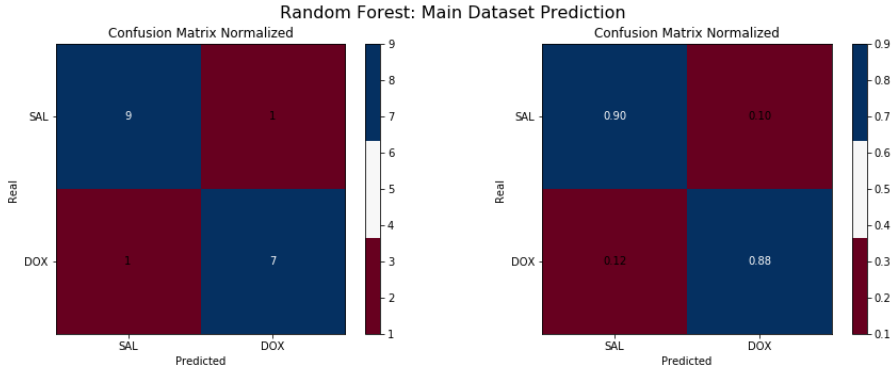


Figure 5.19: DT Confusion Matrix.

	Random Forest	Decision Tree
Accuracy	0.88	0.78
Specificity	0.90	0.60
Precision	0.875	0.66
Sensitivity	0.875	1.00
Cross Validation Mean Accuracy	75	70

Table 5.1: Classification Performance Evaluation Metrics.

Both algorithms performed well concerning the DOX samples classification, since RF only missed one sample and DT got every sample right. Concerning the SAL group, the RF had similar results, missing only one sample classification, however DT performed worse, with an 60% of SAL samples correctly classified.

A K Fold cross validation was implemented to our training data, for 30 runs, and the mean of the scores was considered in order to verify our algorithms accuracy. Thus, the RF classification mean accuracy given by this implementation is 75% and for the DT 70%.

After analyzing each feature contribution for each classifier and concerning the correlation investigation, we concluded that the most important measures to be taken were the ANT and LDH expression, the Abs Weight for their correlation with the Treatment and classification contribution, and the PGC-1*alpha* ratio due to the clear differences found between each treatment correlation results concerning this feature.

These are promising results, however the capacity of our classification methods can only be confirmed by an increase of our sample size.

Chapter 6

Lessons Learned

6.1 Context

Most times, data is created, treated and processed during the experimental phase and years later it is necessary to reuse it, for example, to confirm the original problem or to complement other experiments. However, constructing a new dataset from this data has been proved to be a hard and unpleasant process due to the lack of consistency and standardization.

Usually, researchers start their investigation with a problem, then develop a theory to solve it and finally apply a continuous trial and error process to prove it. During this process, different kinds of variables and combinations are tested resulting in numerous sets of values, which will then be treated and processed according to the researcher's necessities. To do so, the researcher has to come up with ways to store the values produced, without wasting experimental time.

Because an experiment usually has different phases of testing, the consistency of the data registration will depend on different factors, for example:

- the urge of the procedure. Procedures where the researcher has limited time to register the results will end up with sloppy measurements tables, where most variables are not logically labeled;
- need to display the data. If the investigator needs to show the data to others, it will probably be stored in a more intuitive and understandable way;
- how meticulous the researchers is. Usually, on the first phases of testing, the data is stored in concise tables and graphics. However, for experiments that take a great amount of time, the consistency will eventually disappear resulting in unstandardized tables;
- how the data is obtained. Some experiments require the use of specific equipment for recording data. Thus, the automatic exporting of this data

to an excel file will hardly match the methodology used when inserting the values directly in an excel table, formatted by the researcher.

Thus, when a different person wants to treat data from an exterior experimental setup, its format is essential to understand the connection between the variables and how the dataset should be constructed.

More important, the context of the data needs to be clear. It needs to be correctly labeled as well as it needs to be well explained, so that one can understand which measurements can be associated with which same sample and which cannot. It is essential to preserve the experimental context, in order to avoid misinterpretations and misconceptions of the data. A careless and unintuitive structuration of the data will force the investigator to spend a great deal of time in data cleaning and will most probably result in an incorrect feature selection and an unfitting construction of the dataset.

In chapter 4, we presented a review on how the data was delivered and how it was processed into functional datasets. During this phase some difficulties were found considering the problem described previously, thus the next section will be addressing these issues and what would be the good practices to help prevent them.

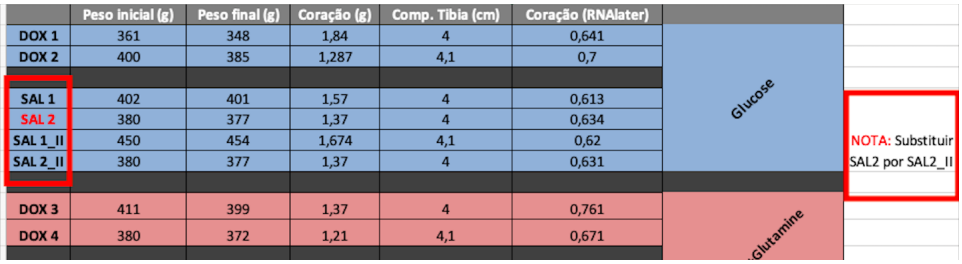
6.2 Good Practices

During data processing, two types of problems were found: problems concerning data consistency and concerning data structure.

As previously mentioned, data consistency is the most important factor to correctly construct a dataset from experimental data, so that one can understand which measurements can be associated with each sample and which cannot. Since the data used in this project came from a practical experiment, involving different trials and different kinds of data, such as timeseries, some inconsistencies were expected due to the aspects mentioned in the previous section. Thus, this section will contain an analysis of those problems and the best solutions to avoid them.

For example, as described in section 4.1, every sample had an ID such as 'SAL1', where the word corresponded to the treatment and the number corresponded to the rats pair. Thus, there were two IDs with the same number: one corresponding to the rat treated with Saline (SAL) and the other to the rat treated with Doxorubicin (DOX). However, through the analysis of the resulting spreadsheets some inconsistencies occurred.

Figure 6.1 corresponds to the weights measures spreadsheet of rats belonging to the Time Control (TC) protocol, and figure 6.2 corresponds to the weights measures of the rats which hearts were perfused with glucose and Iodoacetate (IODO).



	Peso inicial (g)	Peso final (g)	Coração (g)	Comp. Tibia (cm)	Coração (RNAlater)	
DOX 1	361	348	1,84	4	0,641	
DOX 2	400	385	1,287	4,1	0,7	
SAL 1	402	401	1,57	4	0,613	
SAL 2	380	377	1,37	4	0,634	
SAL 1_II	450	454	1,674	4,1	0,62	
SAL 2_II	380	377	1,37	4	0,631	
DOX 3	411	399	1,37	4	0,761	
DOX 4	380	372	1,21	4,1	0,671	

Figure 6.1: Original weights measures spreadsheet of rats belonging to the TC protocol.

	Peso inicial (g)	Peso final (g)	Coração (g)	Comp. Tibia (cm)	Fluxo substrato (ml/min)	Fluxo recuperação (ml/min)	Coração (RNAlater)	
DOX 1	312	312	1,1	3,8	14,5	10	0,604	
DOX 2	315	317	1,032	4,2	32	50	0,55	I
DOX 3	330	322	1,02	3,9	13,5	11,5	0,61	o
DOX 4	342	332	0,997	3,9	16	11	0,518	d
DOX 5	390	389	1,232	3,8	17	10	0,52	o
DOX 6	360	352	1,082	4,1	17,5	14	0,6	a
								c
SAL1	320	327	0,987	4	16,5	11,5	0,419	e
SAL2	368	368	1,095	4,3	11	4	0,584	t
SAL3	370	367	1,046	4,1	14,5	11,5	0,65	a
SAL4	355	352	1,135	4	14	12,5	0,6	d
SAL5	395	392	1,159	4	14	10,5	0,54	o
SAL6	330	341	1,16	3,7	19	16	0,56	

Figure 6.2: Weights measures original spreadsheet of the rats which hearts were perfused with glucose and IODO.

The highlighted area in figure 6.1, shows that there were some issues concerning the SAL1 and SAL2 IDs, since bellow them there are two similar IDs (SAL1_ll and SAL2_ll) and a note to replace SAL2 with the SAL2_ll. However, there is no instruction for the remaining IDs. Concerning the second table, figure 6.2, the same IDs of table 6.1 were found, but corresponding to a different protocol, thus to different rats. Which means that these IDs were not unique, and identifying each rat in our dataset with the given IDs could lead to different protocols aggregation and miss correspondence between features and samples.

Thus, the first lesson to learn from this analysis, is that a clear and unique ID format should be established, from the beginning of the experiment, and applied for every sample disregarding the protocol. The description of this format should accompany the data, for example, in a support text file that could easily be read by others accessing it.

In addition, each column corresponds to a different measurement, and although most column names are clear, if the researcher treating the data is not in the same field of study field, this may create some confusion. Thus, our second advice is to also include each parameter type and description in the support text file.

Regarding problems found due to data structuration, the first advice is to prefer practicality and coherence over esthetic. For instance, the previous figure 6.1 table shows some rows and columns that are merged and although it is visually more appealing and avoids repetition, when imported to a dataset using a different format the result will probably be similar to figure 6.3, which

is the importation result of table from figure 6.1 using python pandas library for data frames.

	Peso inicial (g)	Peso final (g)	Coração (g)	Comp. Tibia (cm)	Fluxo substrato (ml/min)	Fluxo recuperação (ml/min)	Coração (RNAlater)	Unnamed: 7	Unnamed: 8
DOX 1	361.0	348.0	1.840	4.0	NaN	NaN	0.641	Glucose	NaN
DOX 2	400.0	385.0	1.287	4.1	NaN	NaN	0.700	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
SAL 1	402.0	401.0	1.570	4.0	NaN	NaN	0.613	NaN	NOTA: Substituir SAL2 por SAL2_II
SAL 2	380.0	377.0	1.370	4.0	NaN	NaN	0.634	NaN	NaN

Figure 6.3: Importation result of table from figure 6.4 using python pandas library for data frames.

Thus, the simpler it is, the easier will be to adapt to other analysis and software. Moreover, for spreadsheets with multiple tables, such as shown in figure 6.4, the result will be similar but more catastrophic.

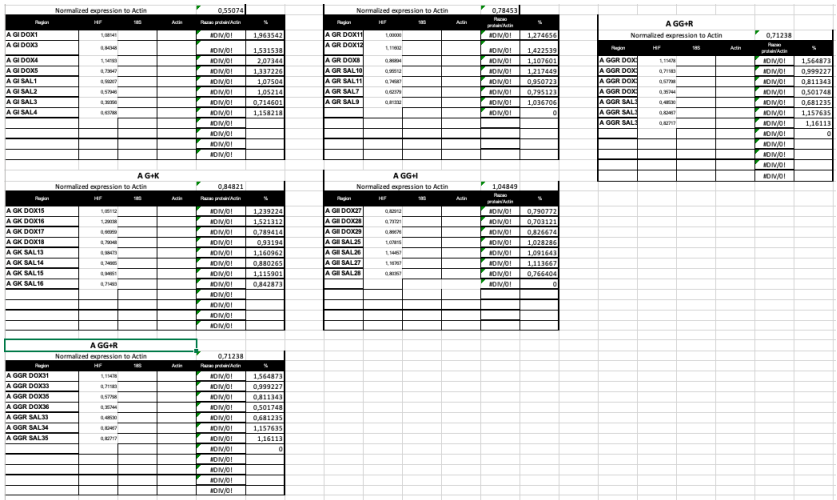


Figure 6.4: Spreedsheet tables of the transcript expression information. Also present in figure 4.3.

Whereas for figure 6.3 we just had to drop the unwanted columns, and each column value would correspond to the same rat (thus, to the same row). For figure 6.4, the same row contains information about at least two different rats, thus the most likely solution would be to separate the tables by hand into different spreadsheets and then import them into the program.

Again, the best structuration solution would be the simplest possible, without loosing logic. The ideal scenario is to think of every program as just columns and rows, without merging tools, and that each row can only contain information about the same observation, so that the columns correspond to the features of the experiment, meaning the measurements, and the rows correspond to each sample.

Finally, after deciding on how to best structure the data. It is very important to standardize the chosen format. For instance, considering figure 6.3, even if the program importation is not clear and one has to build a code to fix it, if the structure is the same for all spreadsheets this repairing code could be applied over and over until the dataset is complete.

In the end, these advices resulted from the difficulties found through the original experimental data spreadsheets processing to functional data frames. Their application should spare time for both researcher and the person who collected the data, since they would not have to explain and confirm each step of the transformation process.

Chapter 7

Conclusion and Future Work

The main goal of this project was to expose possible hidden patterns concerning an experiment whose purpose was to unveil the clinical of Doxorubicin (DOX)'s cardiotoxicity using a model of metabolic inhibition in perfused hearts from Saline (SAL)- and DOX-treated Wistar rats.

From the data collected in this process we had to build and structure a functional dataset capable of being analyzed by different computational tools, including Machine Learning (ML) Algorithms.

The goal was then to not only find new conclusions but also to implement an algorithm capable of distinguishing rats treated with DOX from rats treated with SAL solutions.

During the dataset construction a series of problems emerged concerning the original data structuration and format. Thus, chapter 6. addressed these issues and aims at providing directions to what we think are good practices to help prevent them.

Concerning the dataset analysis, most of our conclusions confirmed the results of the original work. We have also shown that the perfusion with glucose had the most interesting results and the one with Octanoate plus Malate (OM) added no relevant information to our purpose. Thus, our analysis indicates that, to spare time, means and animals, we could only execute one perfusion protocol, which is the glucose protocol.

Completly new from the previous study, we implemented a correlation analysis and ML algorithms to better understand our features relationships and to distinguishing rats treated with DOX from rats treated with SAL solutions.

Concerning our problem' features relationships, during correlation analysis we concluded that the most important parameters for this investigation were the Adenine nucleotide translocator (ANT) and Lactate dehydrogenase (LDH) transcripts expression for their correlation with the treatment and their contribution to the classification algorithms confirmed by their strong correlation with the treatment; the absolute weight difference, which not only showed

great contribution to the algorithms classification, but also presented considerable differences between SAL treated rats results and DOX ones; and finally Peroxisome proliferator-activated receptor-gamma coactivator (PGC-1$alpha$) ratio due to the clear alterations found in each treatment correlation results concerning this feature.

Then, we implemented two different ML algorithms: Decision Trees and Random Forests. Our data was split into train and test datasets and using the train one, a grid search was implemented in order to optimize the algorithms' input parameters. To this dataset an additional cross validation analysis was made in order to help us assess their performance. After optimization, the models were applied to the test set and both their performance was evaluated.

Both our algorithms were able to classify correctly the DOX group, with a sensitivity of 88% for the RF model and 100% for the DT model. However, concerning the SAL group classification DT specificity fell to 60%. In the end, RF model performed considerably better with an accuracy of 88%, whereas DT model accuracy was 78%.

It is safe to say that we confirmed that not only DOX' treatment affects the rat's metabolic system, but also that it is possible to have an automatic model capable of distinguishing both treatment groups. Thus, a detailed data analysis driven by ML allows a better exploration of this biological datasets enabling new discoveries and breakthrough in this field.

This work was featured in one of the most prestigious european conferences on clinical investigation, in which it was orally presented and an abstract on Uncovering hidden patterns in biological datasets to identify metabolic alterations caused by acute and sub-chronic DOX treatments was published. For more information, the reader should acess
https://onlinelibrary.wiley.com/doi/10.1111/eci.13108

Despite our algorithms' classification which was capable of distinguish DOX-treated rats from the SAL ones, these results are limited by our small sample size. Thus, part of this project future work will be to find more examples to complement our dataset.

As mentioned in section 2.3, concerning the biological context, there were two general protocols implemented in Filipa's PhD thesis: the acute protocol, which was the basis of our work, and a sub-chronic protocol. Hence, the next phase of this project is to implement a similar analysis but with the sub-chronic that and hopefully prove that it can complement our dataset and help train our model.

This work was funded by: PTDC/BTM-SAL/29297/2017,POCI-01-0145-FEDER-029297.

Bibliography

[1] C. O. Wilke, *Fundamentals of Data Visualization.* O'REILLY, 2019.

[2] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python.* O'REILLY, 2017.

[3] N. Japkowic and M. Shah, *Evaluating Learning Algorithms.* Cambridge University Press, 2011.

[4] F. Arcamone, *Doxorubicin Anticancer Antibiotics.* Academic Press, 1981.

[5] K. Chandran, D. Aggarwal, R. Q. Migrino, J. Joseph, D. McAllister, E. A. Konorer, W. E. Antholine, J. Zielonka, S. Srinivasan, N. G. Avadhani, and B. Kalyanaraman, "Doxorubicin inactivates myocardial cytochrome c oxidase in rats: Cardioprotection by mito-q," *Biophysical Journa*, 2009.

[6] F. S. Carvalho, *Clarification of the Mitochondrial Role in the Cardiotoxicity of Doxorubicin Using a Whole Heart Perfusion System - Impact of Different Doxorubicin Treatment Regimens Clarification of the Mitochondrial Role in the Cardiotoxicity of Doxorubicin Using a Whole Heart Perfusion System - Impact of Different Doxorubicin Treatment Regimens.* PhD thesis, University of Coimbra, 2014.

[7] K. J. A. Davies and J. Doroshow, "Redox cyclingof anthracyclines by cardiac mitochondria," *The Journal of Biological Chemistry*, 1986.

[8] F. S. Carvalho, R. Burgeiro, Ana Burgeiro Garcia, A. J. Moreno, R. A. Carvalho, and P. J. Oliveira, "Doxorubicin-induced cardiotoxicity: From bioenergetic failure and cell death to cardiomyopathy," *Wiley Online Library*, 2013.

[9] B. T. Kurien and R. H. Scofield, *Western Blotting: An Introduction*, vol. 1312. Humana Press, New York, NY, 2015.

[10] F. H. Stephenson, *Calculations for Molecular Biology and Biotechnology.* Elsevier, 2016.

[11] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, 1959.

[12] T. M. Mitchell, *Machine Learning.* McGraw-Hill, Inc., 1997.

[13] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, 2018.

[14] J. Brownlee, *Master Machine Learning Algorithms.* 2016.

[15] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Elsevier Science B.V.*, 1997.

[16] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *International Journal of Computer Science*, vol. 1, 2006.

[17] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, 2004.

[18] *Introduction to Machine Learning.* The MIT Press, Alpaydin, Ethem.

[19] S. Raschka and V. Mirjalili, *Python Machine Learning.* Packt Publishing, 2017.

[20] C. R. Shalizi, *Advanced Data Analysis from an Elementary Point of View.* Packt Publishing Ltd, 2017.

[21] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, 2014.

[22] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Springer-Verlag London*, 2013.

[23] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice.* OTexts, 2nd ed., 2016.

[24] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 09, pp. 1226–1238, 2002.

[25] P. A. Estévez, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, 2009.

[26] P. E. Mckight and J. Najab, "Kruskal-wallis test," *John Wiley & Sons, Inc.*, 2009.

[27] M. J. H., *Handbook of Biological Statistics.* Sparky House Publishing, 2008.

[28] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh, and (Eds.), *Feature Extraction Foundations and Applications.* Springer, 2006.

[29] I. Guyon, J. Weston, and B. Stephen, "Gene selection for cancer classification using support vector machines," *Kluwer Academic Publishers*, 2002.

[30] I. Guyon and A. Elisseeff, *An Introduction to Feature Extraction.* Springer, Berlin, Heidelberg, 2006.

[31] T. Navin Lal, O. Chapelle, J. Weston, and A. Elisseeff, *Embedded Methods.* Springer.

[32] *Time Series Analysis Univariated and Multivariated Methods.* Greg Tobin, Wei, William W. S.

[33] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer, 2002.

[34] J. VanderPlas, *Pyhton Data Science Handbook*. O'REILLY, 2017.

[35] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*. World Scientific, 2nd ed., 2014.

[36] P. Flach, *Machine Learning The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

[37] C. C. Aggarwal and C. Reddy, *Data Clustering: Algorithms and Applications*. Taylor & Francis Group, LLC, 2014.

[38] J. Lin and D. Demner-Fushman, "Semantic clustering of answers to clinical questions," *AMIA Annu Symp Proc. 2007*, pp. 458–462, 20017.

Appendix A

		Mean		Median		STD		RANGE	
		SAL	DOX	SAL	DOX	SAL	DOX	SAL	DOX
Weight Difference	G	-1.412	-5.059	-3.0	-6.0	4.542	4.038	[-6 ; 11]	[-12 ; 2]
	GG	-2.142	-13.588	-2.5	-13.0	4.975	7.567	[-10 ; 10]	[-29 ; -2]
	OM	1.667	-12.875	1.0	-13.0	4.220	6.249	[-6 ; 10]	[-24 ; 1]
Heart Weight / Final Weight ($x10^{-3}$)	G	3.1	3.2	3.1	3.2	0.3	0.2	[2.60 ; 4.70]	[2.90 ; 3.70]
	GG	3.0	3.0	3.0	3.0	3.0	3.0	[2.60 ; 3.49]	[2.70 ; 3.41]
	OM	2.89	2.82	2.91	2.80	0.19	0.24	[2.50 ; 3.35]	[2.48 ; 3.21]
Tibia Size / Final Weight $x10^{-3}$	G	11.10	10.85	10.85	11.59	0.65	0.92	[10.20 ; 12.89]	[9.77 ; 13.25]
	GG	10.00	10.63	9.99	10.59	0.43	0.68	[9.23 ; 10.77]	[9.65 ; 11.71]
	OM	9.92	10.75	10.00	10.61	0.67	0.59	[8.89 ; 10.89]	[9.97 ; 12.38]

Table A.1: Descriptive analysis of the features *Weight Difference*, *Heart Weight/ Final Weight* and *Tibia Size/Final Weight*.

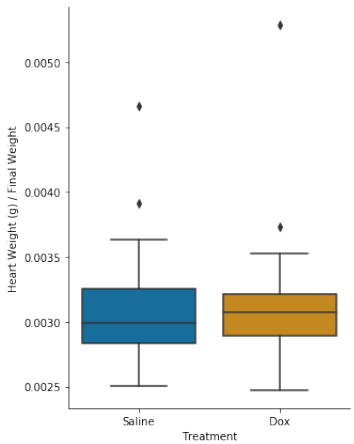


Figure A.1: Boxplot Analysis of *Heart Weight/FinalWeight* feature.

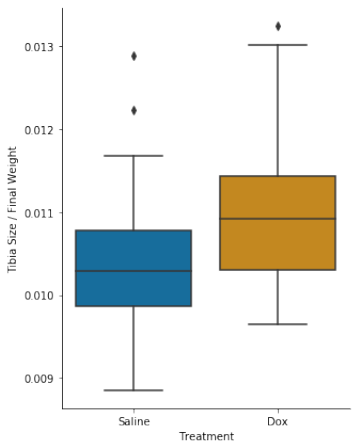


Figure A.2: Boxplot Analysis of *Tibia Size/Final Weight* feature.

	NP							
	Mean		Median		STD		Range	
	SAL	DOX	SAL	DOX	SAL	DOX	SAL	DOX
ANT Exp Expression	1.00	0.302	0.847	0.229	0.458	0.198	[0.637 ; 1.669]	[0.166 ; 0.585]
HIF Exp Expression	1.00	1.234	0.936	1.267	0.342	0.237	[0.710 ; 1.418]	[0.952 ; 1.452]
LDH Exp Expression	1.00	0.896	0.765	0.937	0.482	0.336	[0.746 ; 1.723]	[0.503 ; 1.208]

Table A.2: Descriptive analysis of the NP dataset transcript information.

		ANT Transcript Expression								
		Glucose			Galactose plus Glutamine			Octanoate plus Malate		
		Mean	Meadian	Std	Mean	Meadian	Std	Mean	Median	Std
KCN	SAL	1.000	1.024	0.451	0.750	0.732	0.714	1.000	0.536	1.157
	DOX	1.744	1.650	0.444	1.359	1.377	1.015	2.980	2.980	1.256
IODO	SAL	1.000	1.031	0.173	0.939	0.978	0.292	1.000	1.025	0.088
	DOX	1.547	1.581	0.771	0.534	0.405	0.324	0.957	0.957	0.013
ROT	SAL	1.000	0.553	0.974	1.00	0.813	0.570	1.000	1.129	0.408
	DOX	1.401	1.323	0.679	0.713	0.698	0.154	0.831	0.453	0.994
Total	SAL	1.000	1.031	0.383	0.887	0.813	0.506	1.000	1.025	0.615
	DOX	1.579	1.474	0.592	0.868	0.689	0.672	2.056	0.964	2.498
TC	SAL	1.000	0.957	0.274	0.750	0.766	0.638	1.00	1.047	0.535
	DOX	2.103	0.638	2.171	0.342	0.339	0.149	1.185	1.167	0.769

Table A.3: Descriptive analysis of the ANT expression feature values for TC and main datasets.

		HIF Transcript Expression								
		Glucose			Galactose plus Glutamine			Octanoate plus Malate		
		Mean	Meadian	Std	Mean	Meadian	Std	Mean	Median	Std
KCN	SAL	1.000	0.998	0.162	0.750	0.929	0.511	1.000	1.032	0.287
	DOX	1.120	1.086	0.327	1.327	1.375	0.328	1.540	1.571	0.369
IODO	SAL	1.000	1.064	0.196	1.000	1.060	0.160	1.000	0.949	0.124
	DOX	1.726	1.748	0.349	0.774	0.791	0.063	0.990	1.008	0.227
ROT	SAL	1.000	0.994	0.176	1.00	1.158	0.276	1.000	1.023	0.427
	DOX	1.401	1.275	0.679	0.969	0.905	0.447	0.689	0.538	0.339
Total	SAL	1.000	1.044	0.161	0.909	1.028	0.342	1.000	0.950	0.288
	DOX	1.381	1.337	0.390	1.046	0.893	0.386	1.015	1.057	0.471
TC	SAL	1.000	0.955	0.310	0.994	0.801	0.433	0.750	0.530	0.891
	DOX	0.923	0.934	2.171	1.009	1.020	0.067	2.639	1.844	2.541

Table A.4: Descriptive analysis of the Hif-1*alpha* expression feature values for TC and main datasets.

		LDH Transcript Expression								
		Glucose			Galactose plus Glutamine			Octanoate plus Malate		
		Mean	Meadian	Std	Mean	Meadian	Std	Mean	Median	Std
KCN	SAL	1.000	1.010	0.251	1.000	0.993	0.133	1.000	1.019	0.200
	DOX	1.016	1.010	0.153	1.305	1.386	0.237	1.370	1.365	0.039
IODO	SAL	1.000	1.007	0.075	0.967	1.085	0.399	1.000	0.976	0.152
	DOX	1.175	1.141	0.508	0.893	0.926	0.443	0.956	0.907	0.214
ROT	SAL	1.000	0.981	0.306	1.000	1.081	0.228	1.000	1.002	0.315
	DOX	1.152	1.103	0.520	1.212	1.241	0.171	0.739	0.722	0.135
Total	SAL	1.000	1.027	0.210	0.988	1.048	0.252	1.000	0.976	0.215
	DOX	1.111	1.103	0.380	1.137	1.241	0.332	0.993	0.900	0.305
TC	SAL	1.000	1.013	0.821	0.999	0.970	0.466	0.750	0.959	0.530
	DOX	1.295	1.292	0.565	1.320	1.403	0.280	2.185	0.836	2.825

Table A.5: Descriptive analysis of the LDH expression feature values for TC and main datasets.

	Mean		Median		STD		Range	
	SAL	DOX	SAL	DOX	SAL	DOX	SAL	DOX
PGC1alpha Ratio	0.404	0.715	0.406	0.771	0.306	0.360	[0.048; 0.806]	[0.082 ; 1.179]
TFAM Ratio	0.833	1.283	0.834	1.543	0.554	0.650	[0.083 ; 1.407]	[0.070 ; 1.787]

Table A.6: Descriptive analysis of the TC dataset protein information.

		Mean		Median		STD		RANGE	
		SAL	DOX	SAL	DOX	SAL	DOX	SAL	DOX
PGC1alpha Ratio	G	0.328	0.274	0.266	0.142	0.195	0.142	[0.030 ; 0.678]	[0.168 ; 0.621]
	GG	0.241	0.183	0.221	0.452	0.145	0.180	[0.00 ; 0.476]	[0.00 ; 0.523]
	OM	0.339	0.353	0.302	0.353	0.163	0.189	[0.180 ; 0.900]	[0.088 ; 0.792]
TFAM Ratio	G	0.251	0.205	0.224	0.160	0.187	0.162	[0.069; 0.776]	[0.059; 0.627]
	GG	0.562	0.322	0.499	0.202	0.471	0.221	[0.000 ; 1.843]	[0.000 ; 0.813]
	OM	0.258	0.233	0.223	0.186	0.164	0.139	[0.062 ; 0.765]	[0.059; 0.483]

Table A.7: Descriptive analysis of the protein information features for the main dataset.

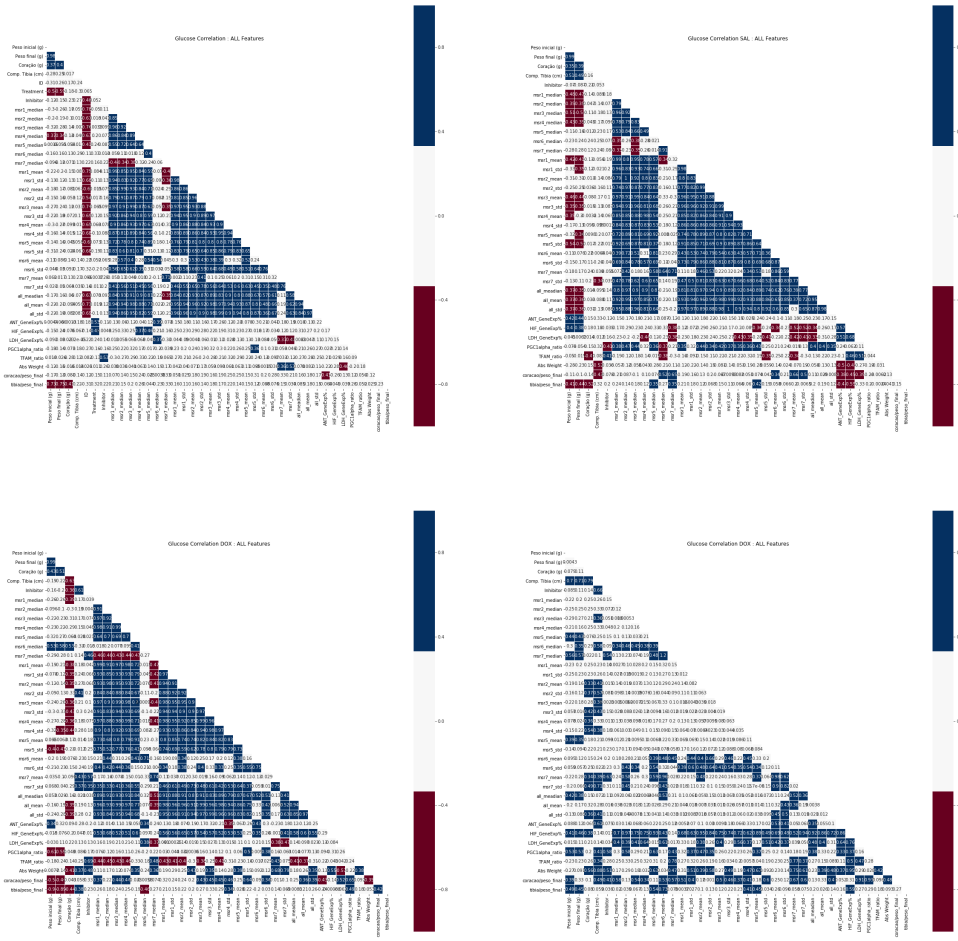


Figure A.3: Glucose group all features correlation analysis.

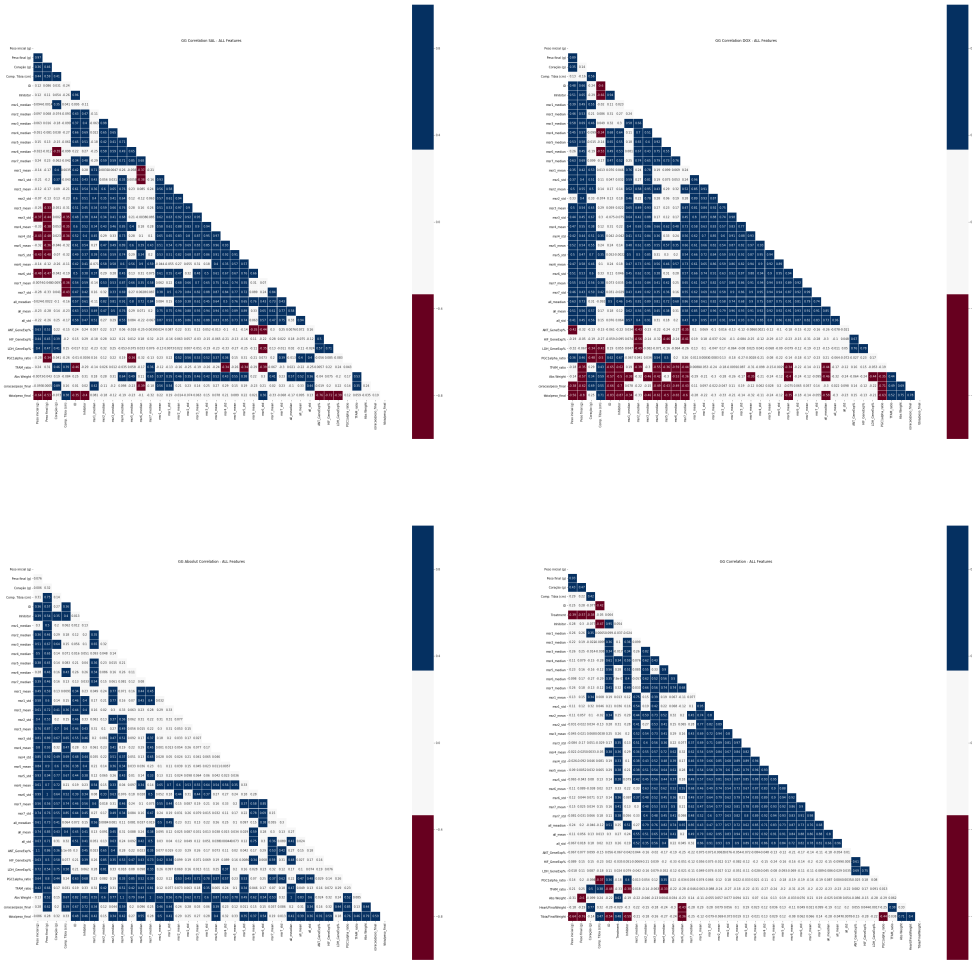


Figure A.4: GG group all features correlation analysis.

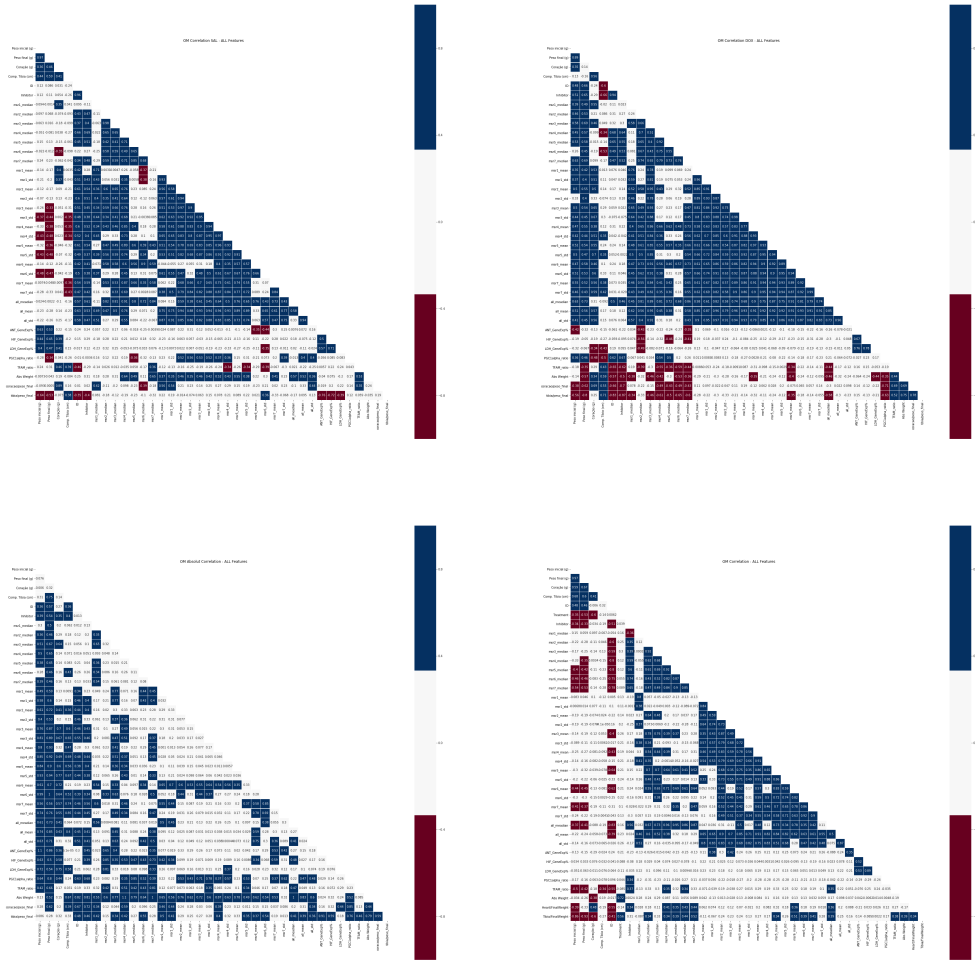


Figure A.5: OM group all features correlation analysis.

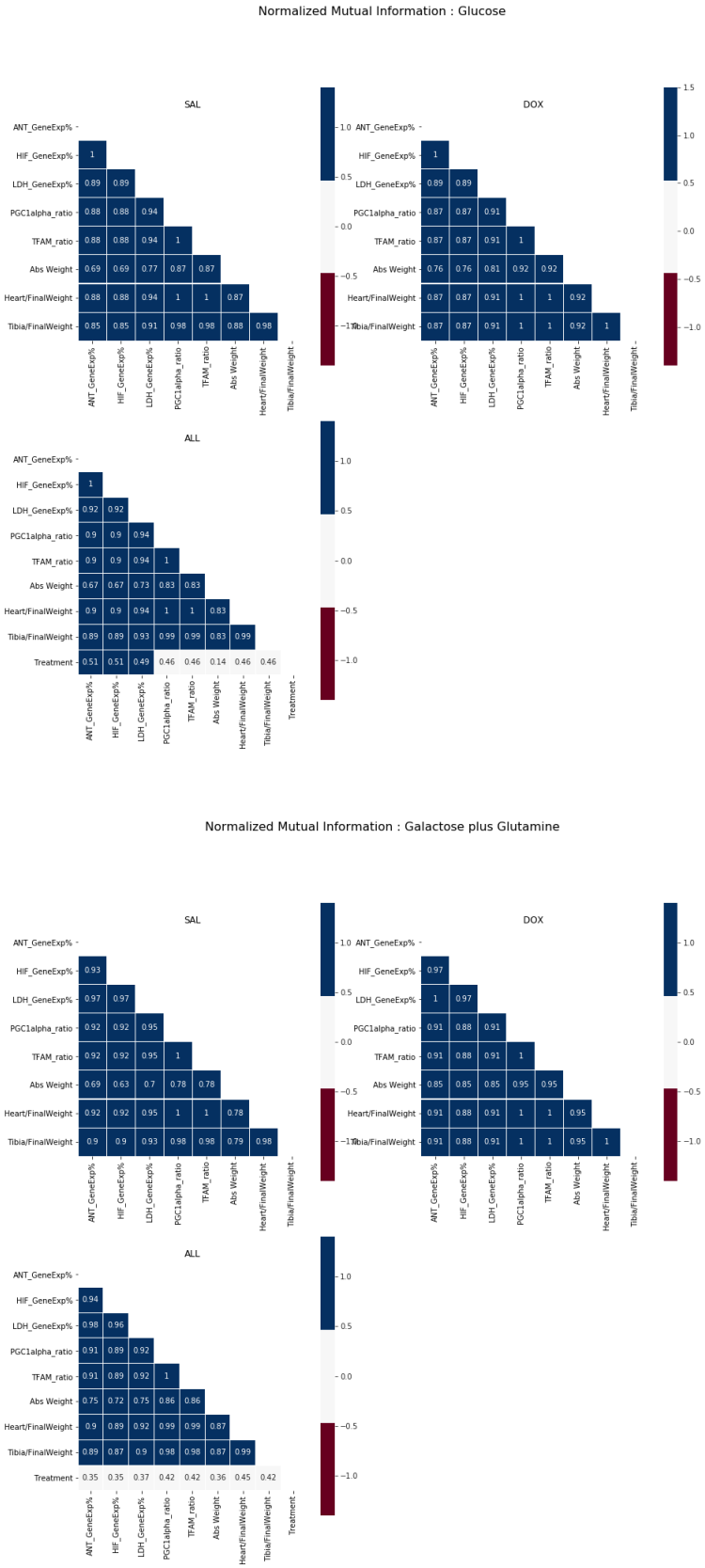


Figure A.7: GG group features mutual information analysis.

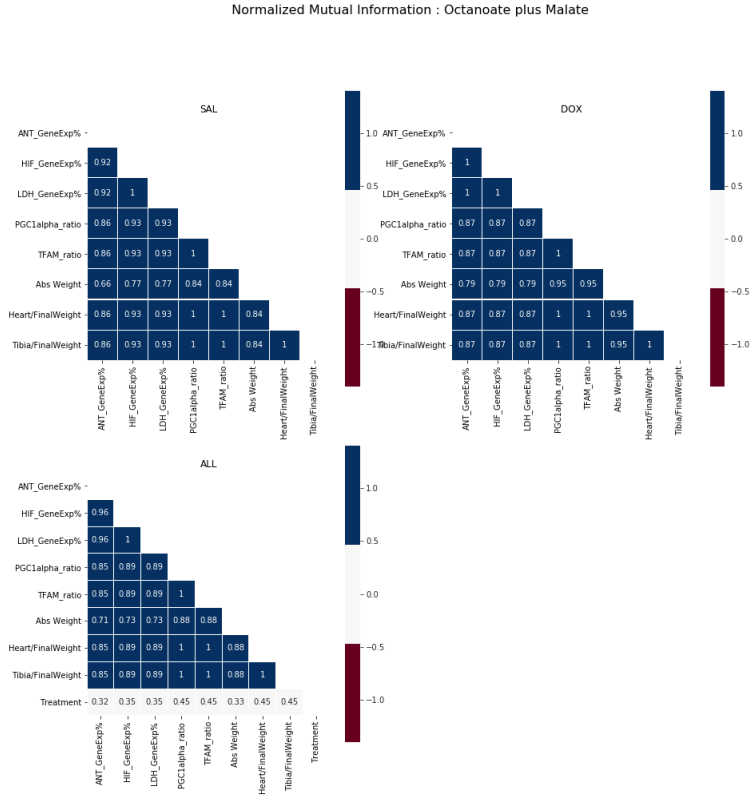


Figure A.8: OM group features mutual information analysis.

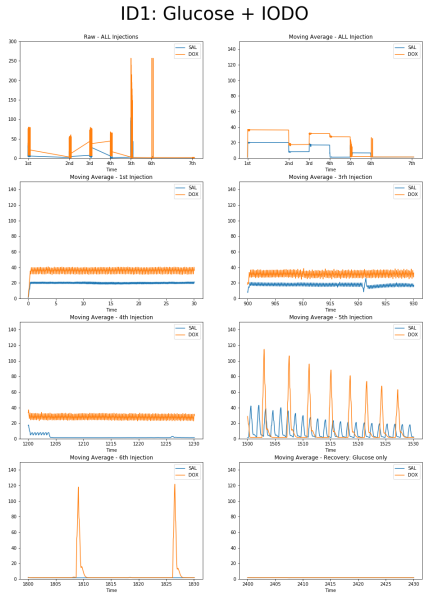


Figure A.9: G+IODO Time series ID1.

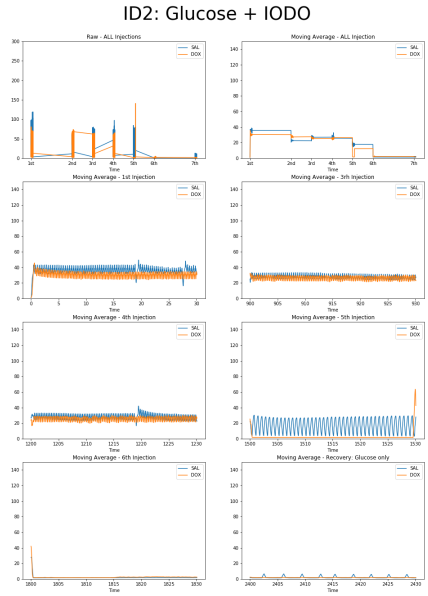


Figure A.10: G+IODO Time series ID2.

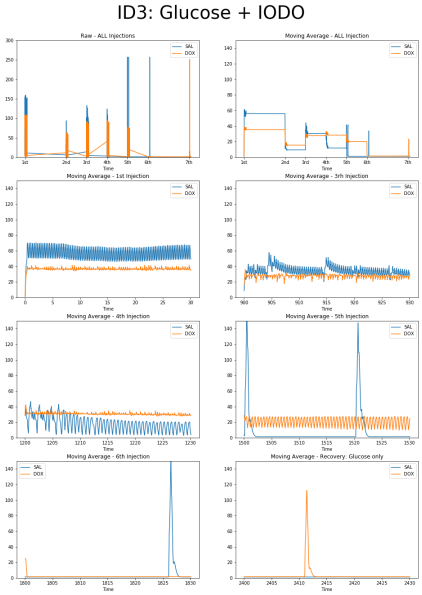


Figure A.11: G+IODO Time series ID3.

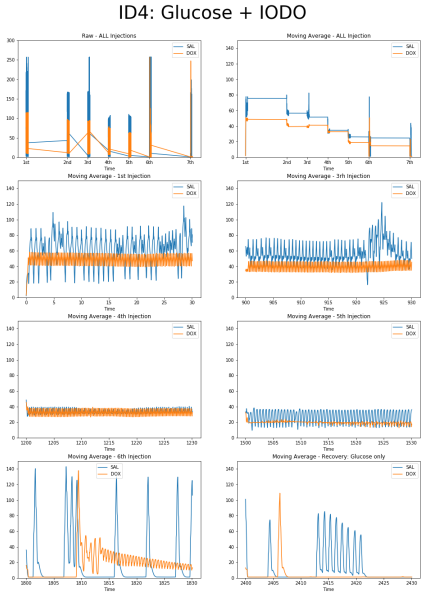


Figure A.12: G+IODO Time series ID4.

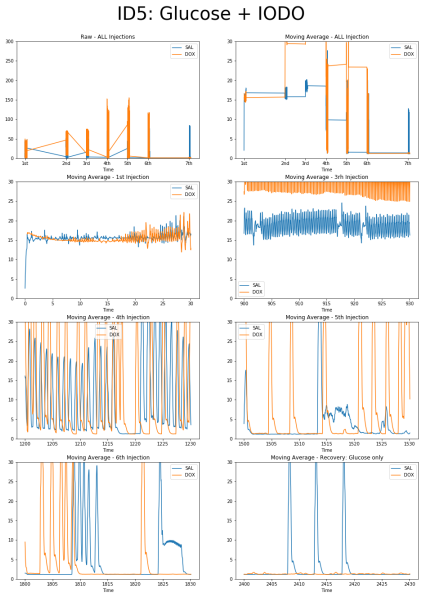


Figure A.13: G+IODO Time series ID5.

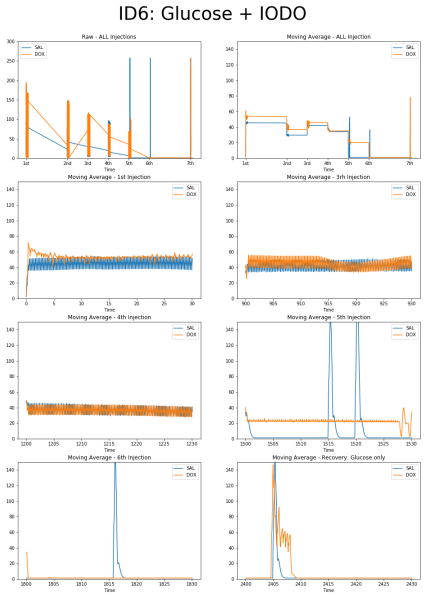


Figure A.14: G+IODO Time series ID6.

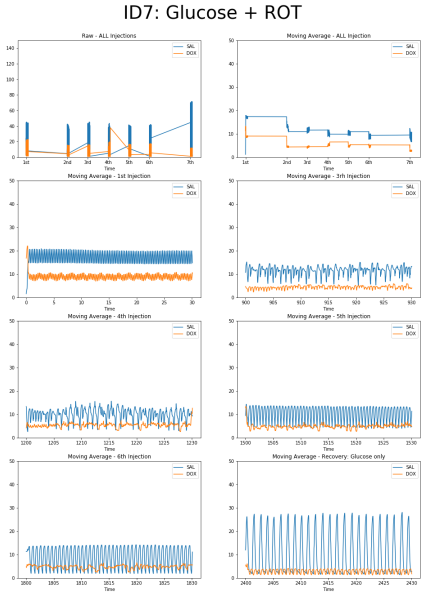


Figure A.15: G+ROT Time series ID7.

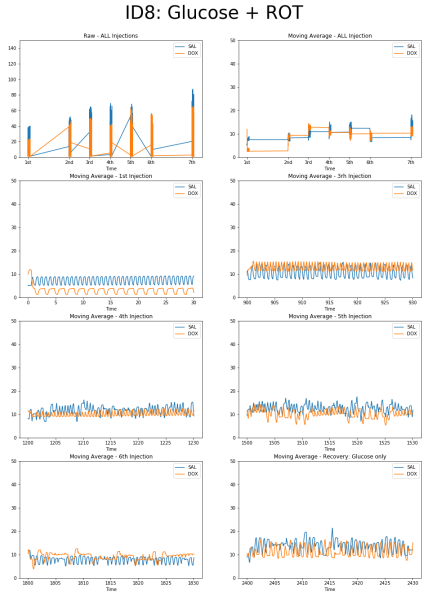


Figure A.16: G+ROT Time series ID8.

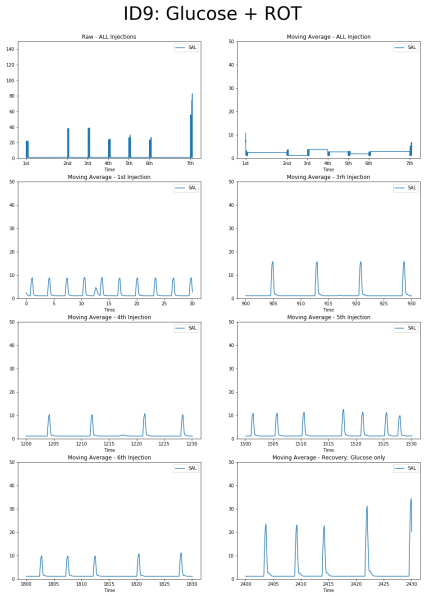


Figure A.17: G+ROT Time series ID9.

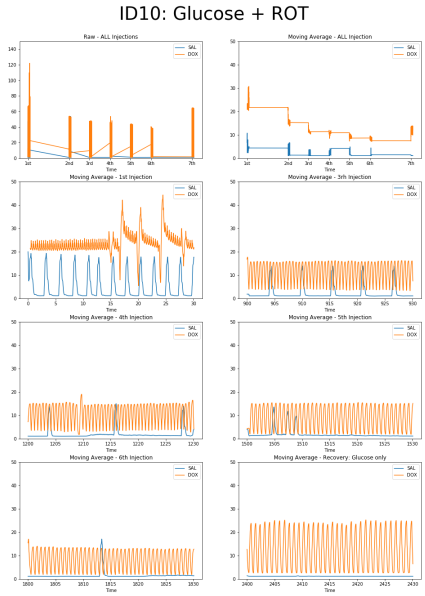


Figure A.18: G+ROT Time series ID10.

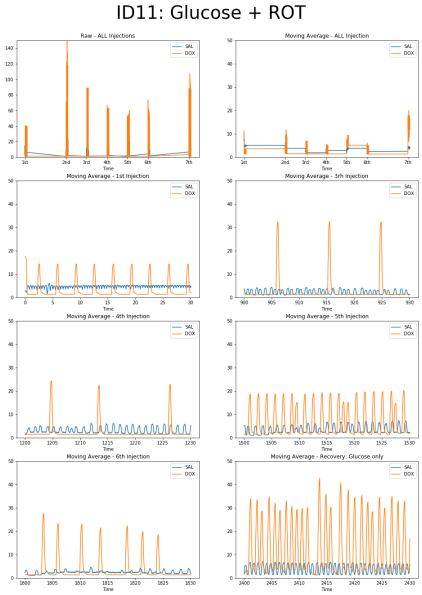


Figure A.19: G+ROT Time series ID11.

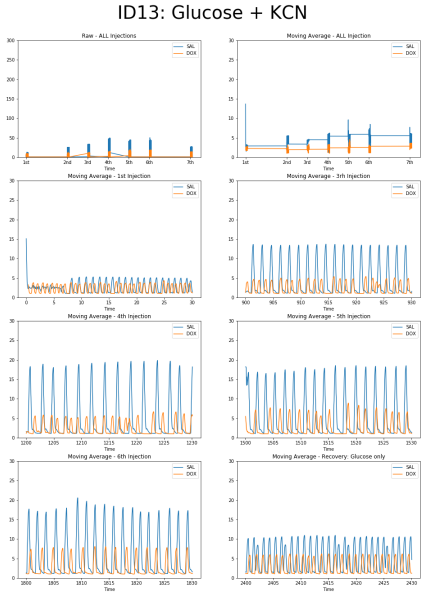


Figure A.20: G+KCN Time series ID13.

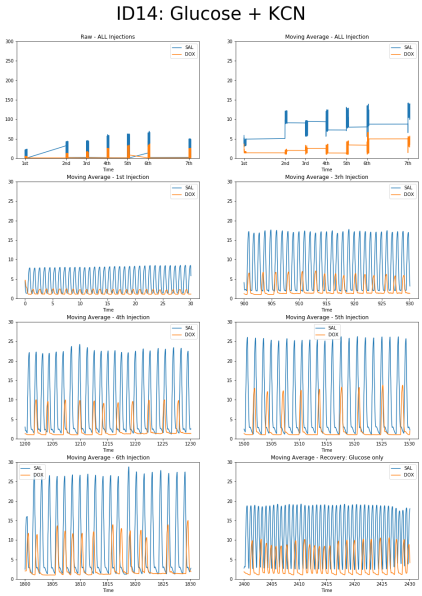


Figure A.21: G+KCN Time series ID14.

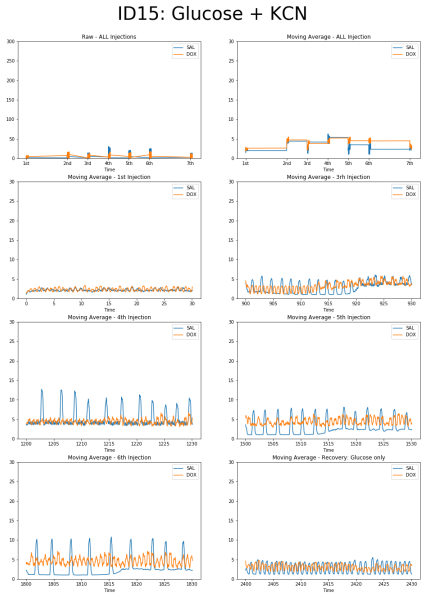


Figure A.22: G+KCN Time series ID15.

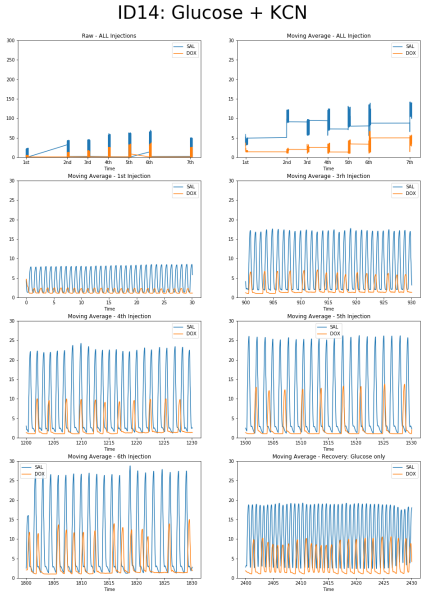


Figure A.23: G+KCN Time series ID14.

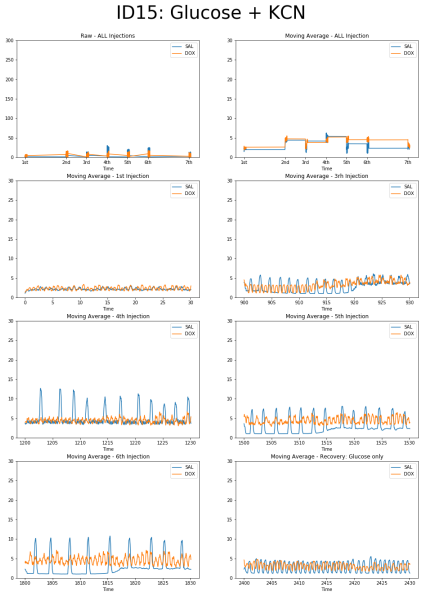


Figure A.24: G+KCN Time series ID15.

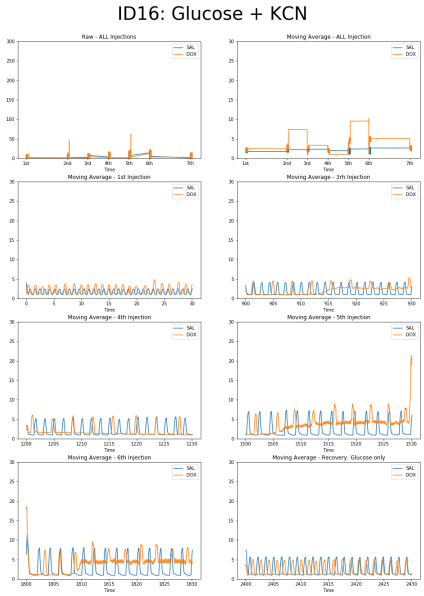


Figure A.25: G+KCN Time series ID16.

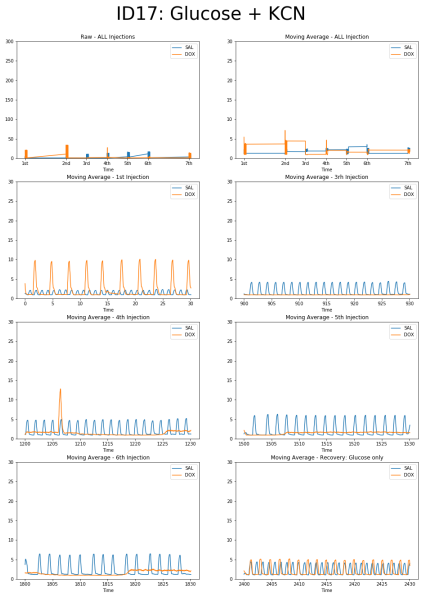


Figure A.26: G+KCN Time series ID18.

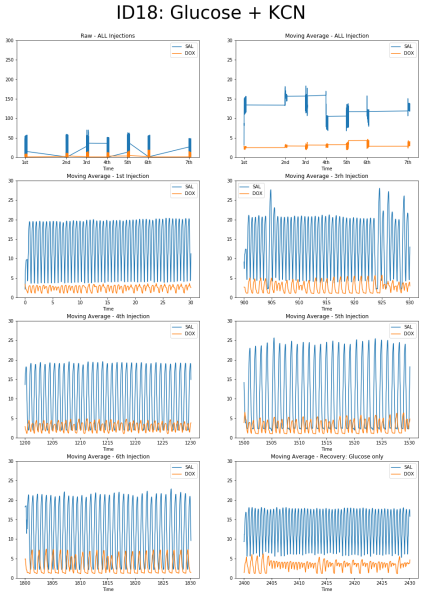


Figure A.27: G+KCN Time series ID18.

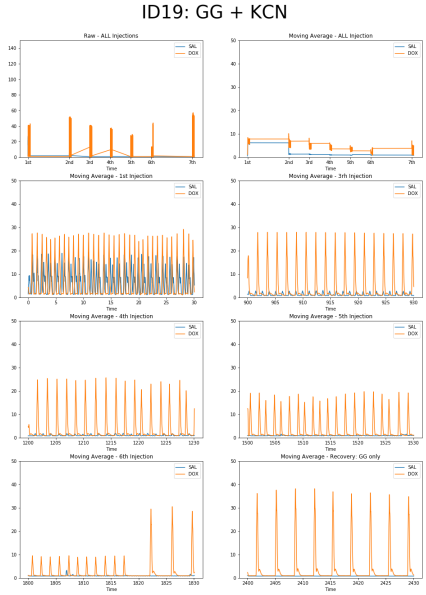


Figure A.28: GG+KCN Time series ID19.

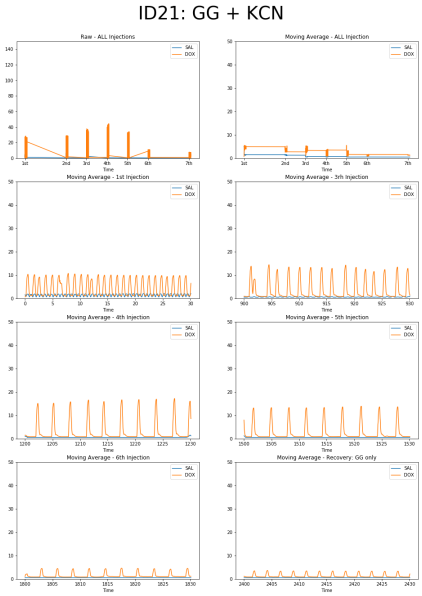


Figure A.29: GG+KCN Time series ID21.

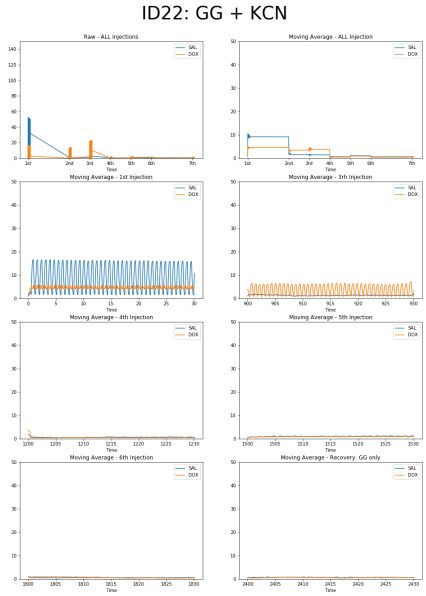


Figure A.30: GG+KCN Time series ID22.

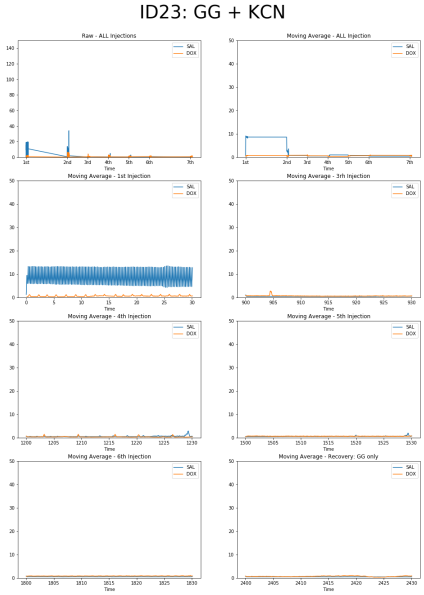


Figure A.31: GG+KCN Time series ID23.

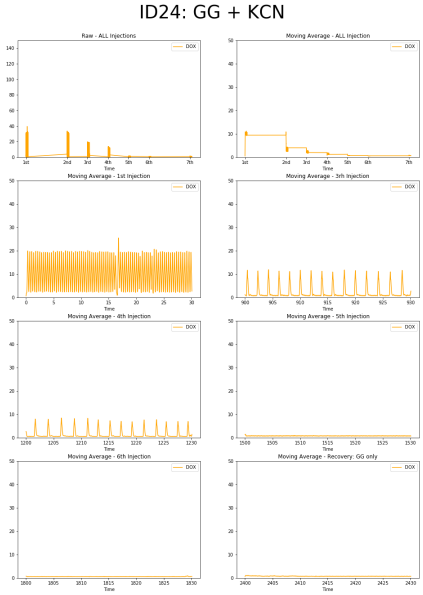


Figure A.32: GG+KCN Time series ID24.

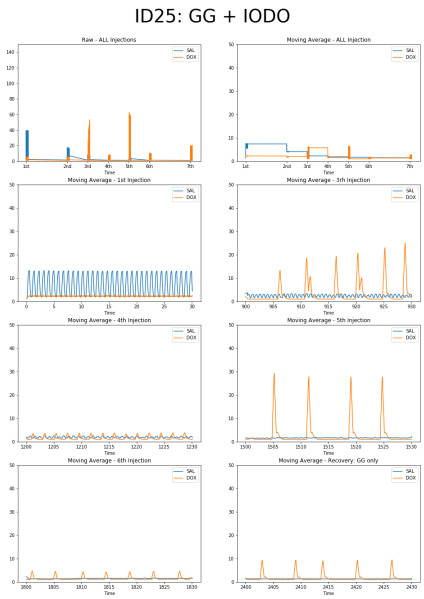


Figure A.33: GG+IODO Time series ID25.

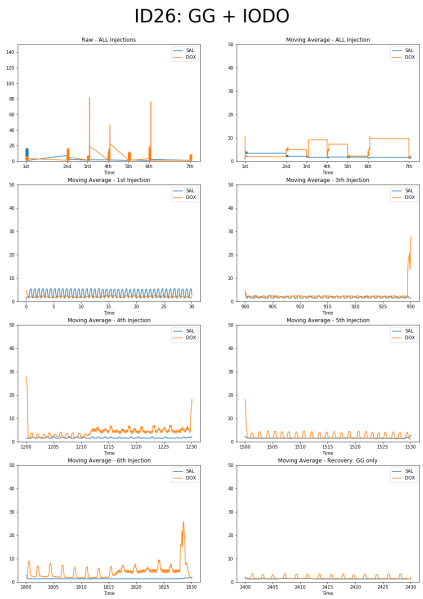


Figure A.34: GG+IODO Time series ID26.

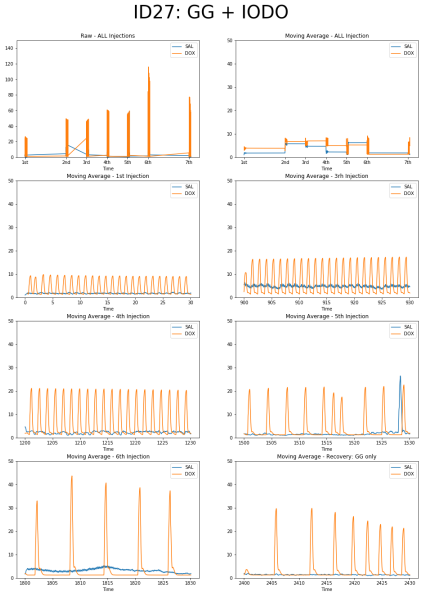


Figure A.35: GG+IODO Time series ID27.

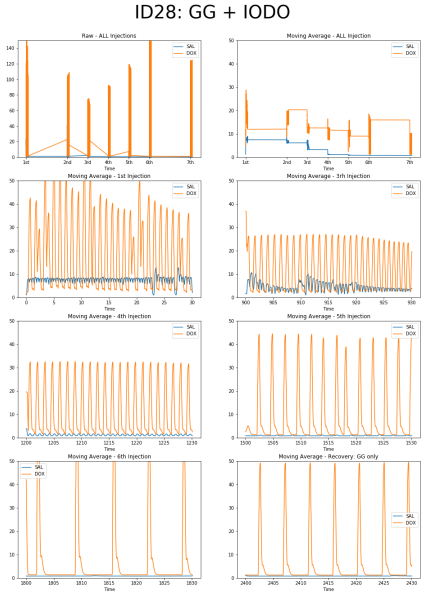


Figure A.36: GG+IODO Time series ID28.

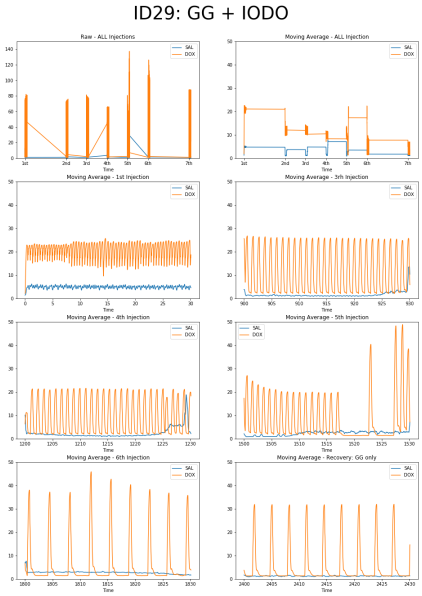


Figure A.37: GG+IODO Time series ID29.

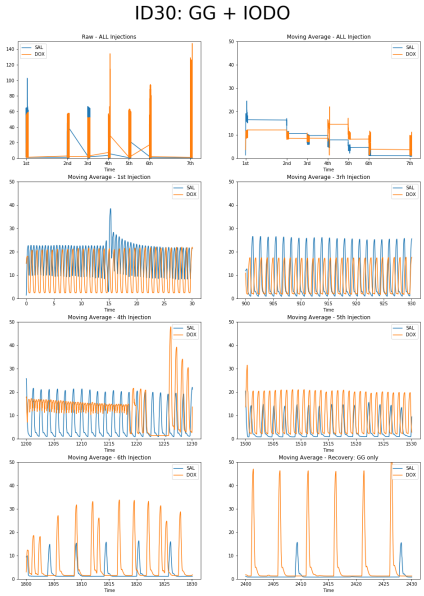


Figure A.38: GG+IODO Time series ID30.

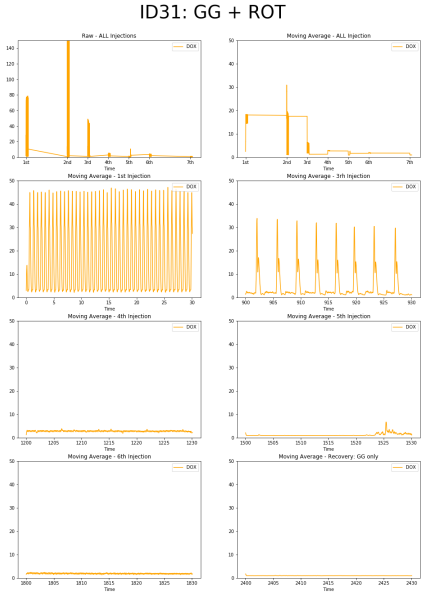


Figure A.39: GG+ROT Time series ID31.

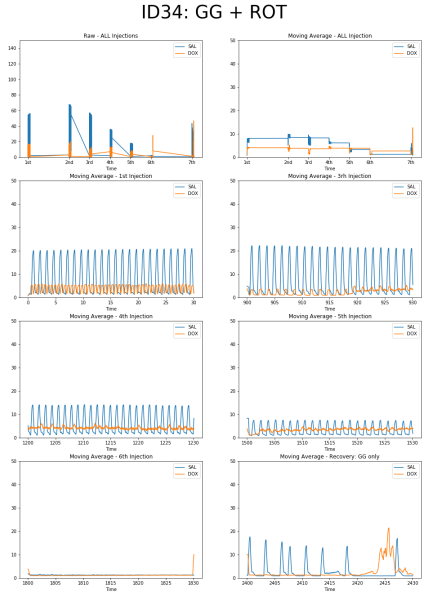


Figure A.40: GG+ROT Time series ID34.

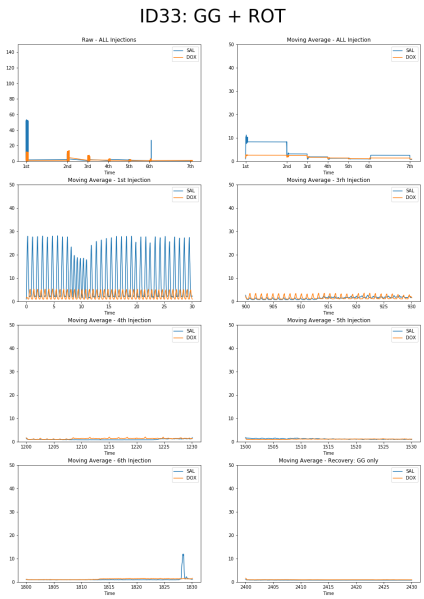


Figure A.41: GG+ROT Time series ID33.

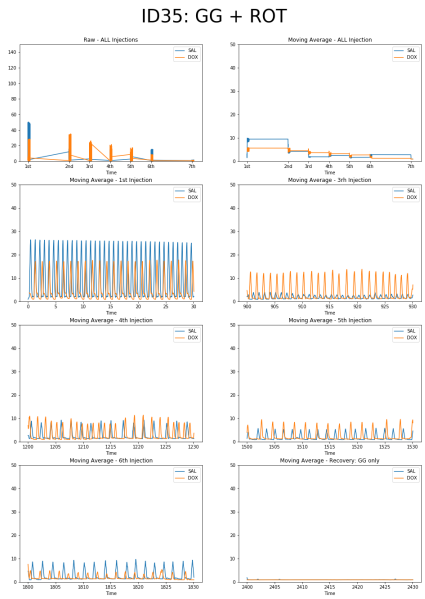


Figure A.42: GG+IODO Time series ID35.

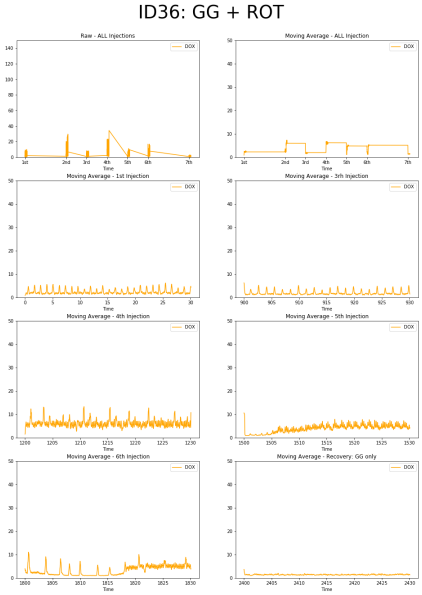


Figure A.43: GG+ROT Time series ID36.

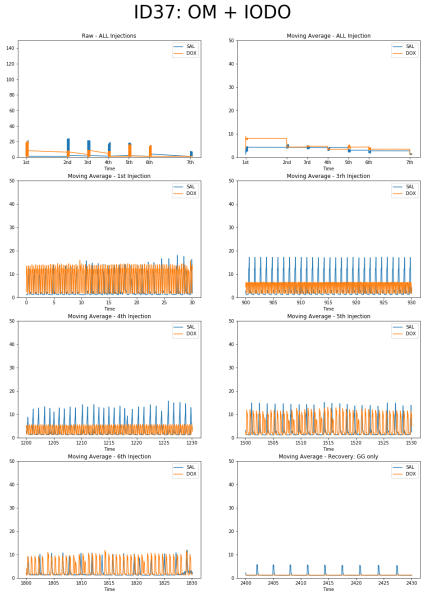


Figure A.44: OM+IODO Time series ID37.

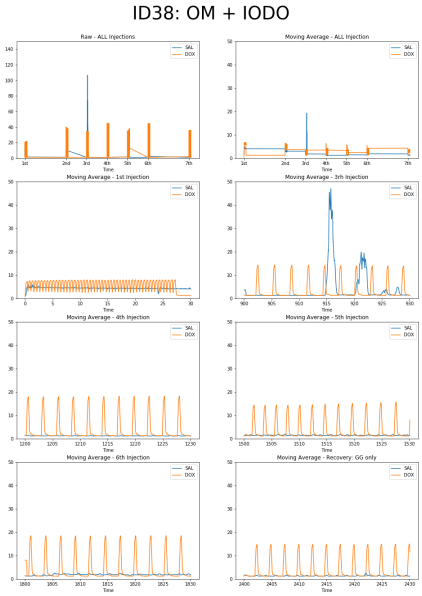


Figure A.45: OM+IODO Time series ID38.

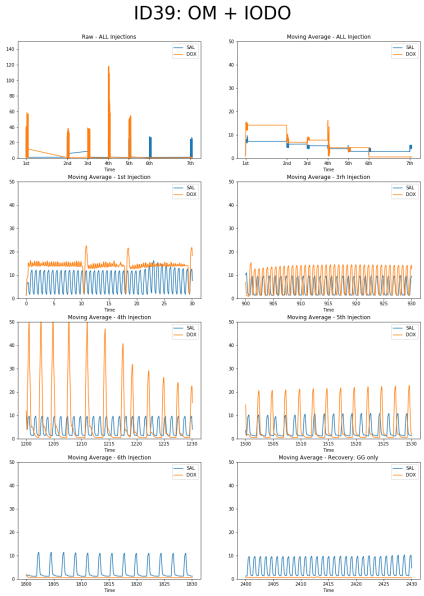


Figure A.46: OM+IODO Time series ID39.

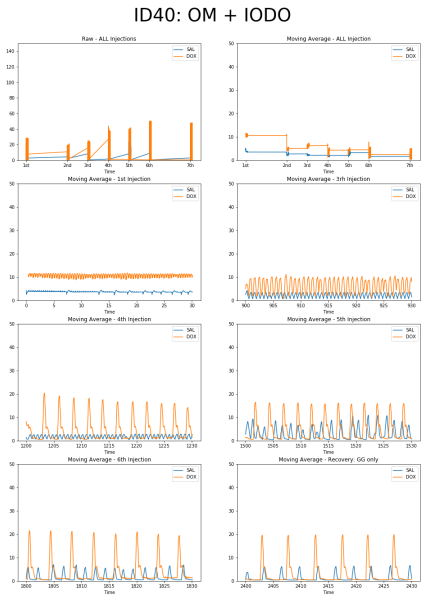


Figure A.47: OM+IODO Time series ID40.

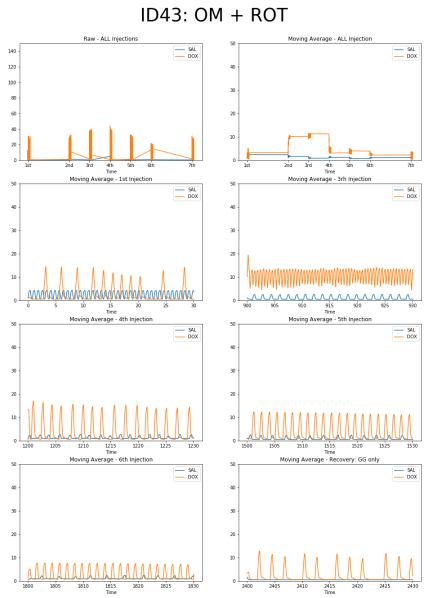


Figure A.48: OM+ROT Time series ID43.

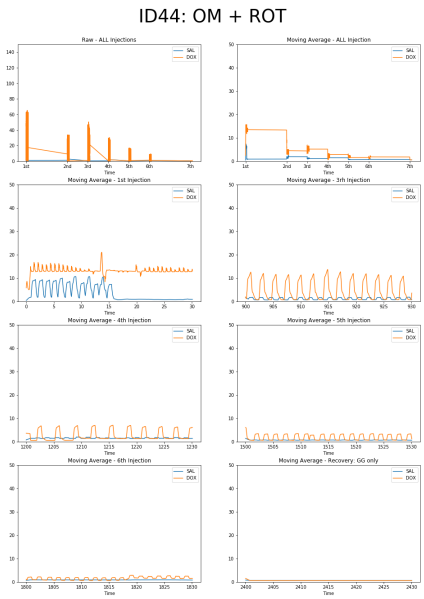


Figure A.49: OM+ROT Time series ID44.

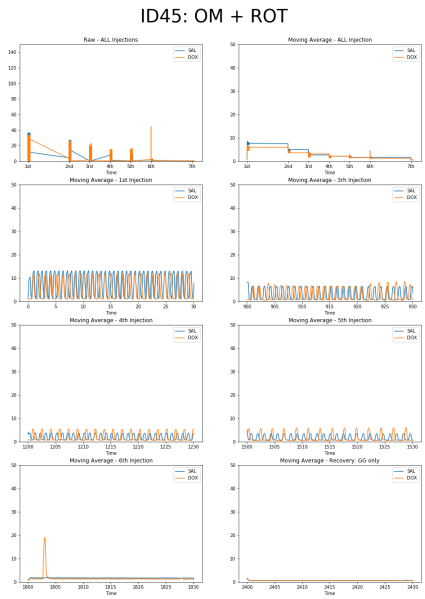


Figure A.50: OM+ROT Time series ID45.

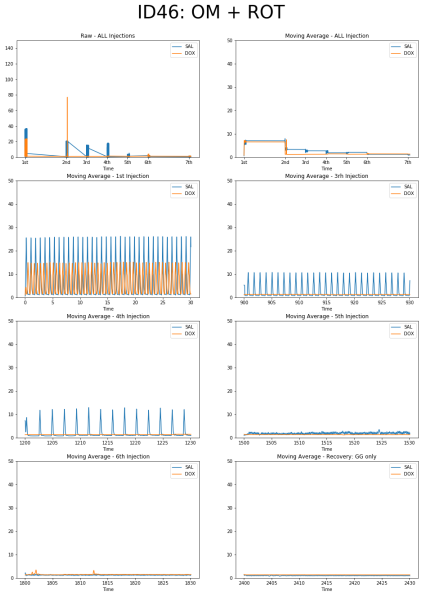


Figure A.51: OM+ROT Time series ID46.

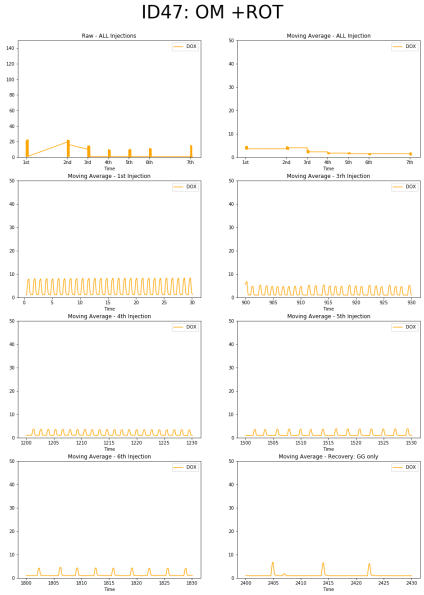


Figure A.52: OM+ROT Time series ID47.

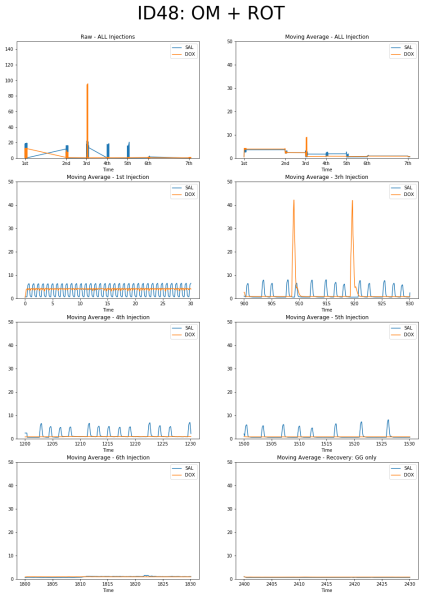


Figure A.53: OM+ROT Time series ID48.

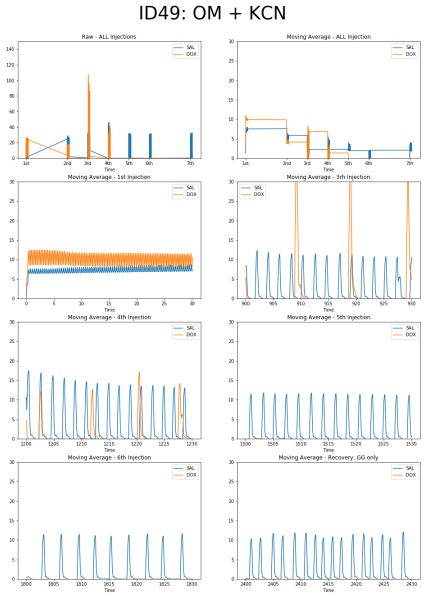


Figure A.54: OM+KCN Time series ID49.

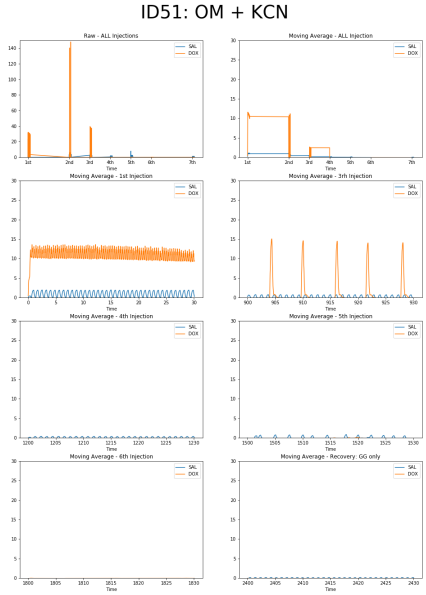


Figure A.55: OM+KCN Time series ID51.

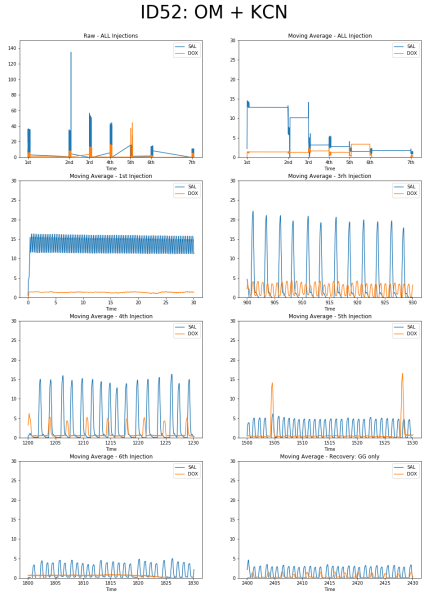


Figure A.56: OM+KCN Time series ID52.

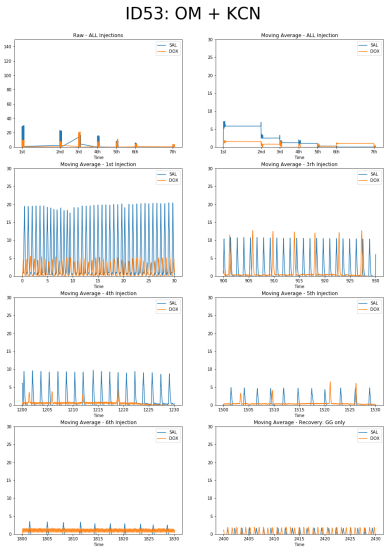


Figure A.57: OM+KCN Time series ID53.

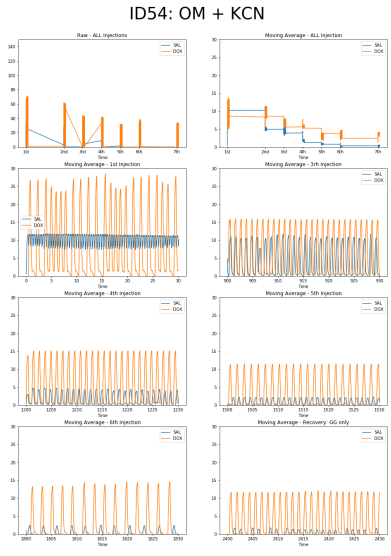


Figure A.58: OM+KCN Time series ID54.