Developing *in-silico* predictions of toxicity

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by CHRYSTALLA IOSIF

June/2022

Acknowledgements

Professionally I would like to thank my supervisor, Dr. Philipp Antczak, for his invaluable contribution to this study, starting from performing all the experiments and providing me with all the necessary data for my analysis, offering his guidance and expertise whenever I needed it, and his patience during the writing period. I would also like to thank my friends and colleagues, Dr Eva Caamano Gutierrez, Dr Louise Takeshita, Simon Perkins and Kim Clarke for their professional input in this work, their emotional support and the lovely time I had with them during my time in Liverpool.

Special thanks to my family, for always supporting and loving me, and for all their care, guidance and sacrifice. I also liked to thank all my friends, especially Maria Louiza Frantzi for being there for me, supporting me and helping me, especially during the PhD write-up stage.

Table of content

List of Figures

List of Tables

Abbreviation

AIC	Akaike information criterion
AO	Adverse Outcome
AOP	Adverse Outcome Pathway
BIC	Bayesian information criterion
GO	Gene Ontology
GUTS	General Unified Threshold Model
KE	Key Events
LOOCV	Leave-one-out cross-validation
MIE	Molecular Initiating Event
KEGG	Kyoto Encyclopedia of Genes and Genomes
KERs	Key Event Relationships
LASSO	least absolute shrinkage and selection operator
LC50	Lethal concentration at which 50% of the population is dead
MoA	Mode of Action
PCA	Principal Component Analysis
QSAR	Quantitative Structure-Activity Relationship models
RF	Random Forest
ROI	Region of Interest
SAM	Significance Analysis of Microarrays
STAR	Spliced Transcripts Alignment to a Reference
TMM	Trimmed mean of M-values
TPM	Transcripts per million

Developing in-silico predictions of toxicity - Chrystalla Iosif

Abstract

A large number of chemicals are released into the environment from various human activities, including the manufacturing and use of pharmaceuticals, industrial chemicals, pesticides, and insecticides. Traditional risk assessment requires the use of long-term experiments to evaluate the toxicity of each chemical on different organisms and multiple levels of biological organisation, which is time-consuming and costly. Thus, most of those chemicals are introduced into the environment before a proper risk assessment is performed, increasing the risk of releasing a potentially environmentally harmful compound. The need to reduce risk assessment time and cost, facilitated the generation of *in silico* approaches based on experimental data, computational approaches (predictive modelling, classification methods) and previous scientific knowledge. Such methods have been used to predict the toxicity of new chemicals with unknown toxicological effects, based on similarities with chemicals that have already been assessed.

Chemical structural information (molecular descriptors) that describe its physicochemical properties, such as bonds composition, chemical weight, electronegativity, lipophilicity, and chemical polarity have been associated with the mode of action of each chemical related to toxicity (MoA), and through predictive modelling to toxicity phenotypes (QSAR models). However, the high heterogeneity between chemicals and the fact that chemicals with similar structural characteristics (cis-, trans- isomers) may have different effects on organisms where only one of them is toxic or chemicals with different structures may act through similar mechanisms (endocrine disruptors), underlies the inability of such methods to describe the toxicity of each chemical fully. At the same time, signature-matching approaches have been developed that are based on the assumption that chemicals with similar gene expression signatures will cause similar biological effects and can be used in predicting the toxicity of new chemicals and identifying toxicity-related genes. This was facilitated by the development of "'omics" technologies (genomics, transcriptomics, proteomics, metabolomics and epigenomics) that allow the generation of large datasets that can be used to explore changes at the molecular level.

In this project, the ability of molecular descriptors (physicochemical information of a chemical) and mRNA sequencing data (gene expression profiles after exposure), in clustering a set of highly heterogeneous chemicals based on profile similarities were tested and compared, highlighting the difference between those two clustering methods. The large set of heterogeneous chemicals (in structural features and gene count profiles) used in this study allows for more generalizable results due to the bigger applicability domain of each method. Clustering chemicals using molecular descriptors was highly associated with changes in heart rate after chemical exposure and with the Verhaar MoA classification since both methods use chemical structural characteristics for grouping chemicals based on toxicity. However, using the differentially expressed genes between control and exposure, chemicals that significantly change zebrafish heart rate tend to cluster together, but no correlation can be seen when compared to MoA classification. These results suggest that Verhaar MoA classification is not necessarily representative of cardiotoxicity, underlying the importance of using molecular responses such as gene expression profiles in assessing chemical-induced cardiotoxicity. In an effort to connect those two datasets, molecular descriptors and gene count profile data, predictive modelling was performed to identify a set of molecular descriptors that could explain the difference between the mRNA clusters, generated using the differentially expressed genes. However, due to the high heterogeneity between the chemicals, this was not successful when the whole dataset was used, but by reducing the number of chemicals, a predictive model was generated that could predict the potential mRNA cluster of each chemical with high accuracy (R^2=0.89), thus more chemicals with high variability in structural characteristics (molecular descriptors) and gene expression profiles, can be used to generate more homogenous clusters and validate this relationship.

Molecular descriptors and gene count profiles were used as input in the predictive modelling function generated in this study, in an effort to identify a set of molecular descriptors or genes that can be used to predict changes in heart rate caused by chemical exposure. Predicting changes in zebrafish heart rate using the structural information (molecular descriptors) was not successful when the whole dataset was used (143 chemicals). The predictive function was not able to identify a set of molecular descriptors that can explain the various effects the chemicals have on the zebrafish heart rate.

On the other hand, predictive modelling identified 80 genes that can predict the effect the chemical has on zebrafish heart rate. Following those results, the chemical dataset was split into three, according to the clustering results generated using the mRNA data. Modelling with molecular descriptors (QSAR models) generated a model for two out of the three clusters, with only one common molecular descriptor, highlighting the difference between the chemicals in the separated clusters. Models based on transcriptional data did not produce comparable accuracies. The effect the chemicals in the third cluster have on zebrafish heart rate could only be predicted by gene expression profile data and not molecular descriptors. Showing the importance of both datasets in risk assessment.

Finally, based on the adverse outcome pathway (AOP) concept, the gene count data were grouped into metabolic pathways using the KEGG database, and predictive modelling identified pathways that their activity could to some extent predict changes in zebrafish heart rate, chemical concentration during exposure, and the LC50 of each chemical (concentration at which 50% of the population is dead). The pathways identified in this study to be predictive of chemical LC50 or chemical exposure concentration can be close to or represent a potential molecular initiating event (e.g. receptor binding), thus the shortest distance (lowest number of pathways) between pathways associated with chemical toxicity (molecular initiating event) and pathways related to heart-rate fold change (adverse outcome) can represent a potential AOP, with the pathways in between representing the key events. The use of large-scale genomics data and pathway analysis can be used in identifying new key events or potential adverse outcome pathways, and assist in the generation of adverse outcome pathways networks for chemical risk assessment.

Chapter 1

Introduction

1.1 Chemical risk assessment

Human activities, including the manufacturing and use of pharmaceutical and industrial compounds, have increased the abundance and diversity of such chemicals in the environment. Potentially toxic chemicals may be found in or used in the production of medical devices, pharmaceutical substances, industrial chemicals, insecticides, pesticides and packaging materials (Myatt *et al.*, 2018). Toxicity describes the various adverse effects caused by exposure to various chemicals on humans, animals, plants, and the environment. Such chemicals are persistent and can potentially be deleterious, thus increasing the need for reliable and accurate, acute toxicity testing methods (Hwang *et al.*, 2002). Chemical exposure has also been associated with heart defects (cardiotoxicity), which could lead to myocardiopathy, including arrhythmia and myocardial hypertrophy. Risk assessment of such chemicals requires the use of multiple vertebrates as model organisms, including rodents and fish (Scholz *et al.*, 2014). The exposure method, chemical dosage, frequency, and duration of exposure along with chemical properties such as absorption, distribution, metabolism, and excretion/elimination, can be used in determining the toxicological effect of a chemical on an organism (Raies *et al.*, 2016).

Danio rerio (zebrafish) is a tropical fish, native to Southeast Asia, which has been used as a model organism since the 1960s (Figure 1.1). The complete zebrafish genome sequence, available since 2013, consists of more than 26 thousand protein-coding genes. The zebrafish genome shares a lot of similarities with the human genome, where 70% of human genes have zebrafish orthologs (Howe et al., 2013), and 84% of the genes associated with human diseases are homologous to zebrafish genes. In addition, major organs and tissues, such as the heart and vascular system, are conserved among vertebrates. Zebrafish are an important model for evaluating chemical ecotoxicology and for understanding the mechanisms of development and human health (Teame *et al.*, 2019).

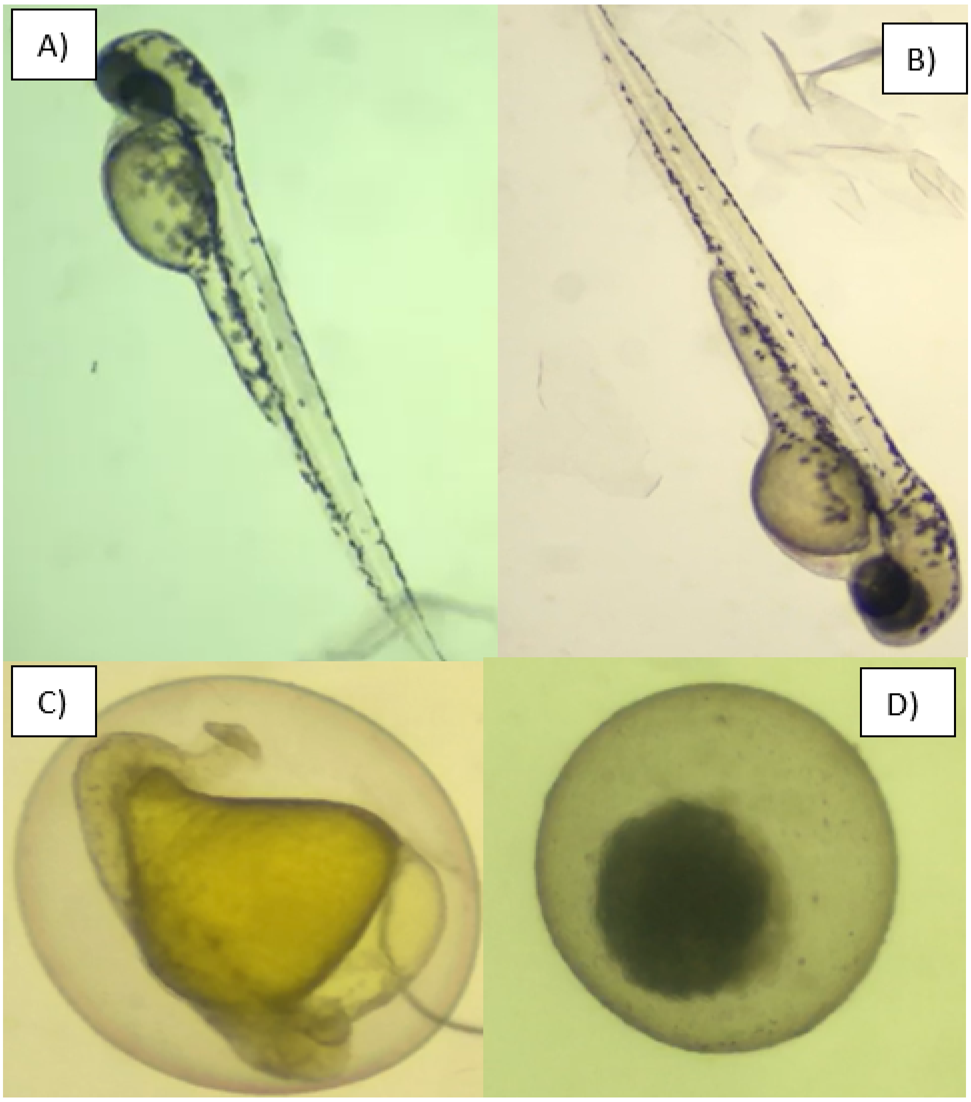


Figure 1.1: Zebrafish embryos after 72 hours of exposure. A) Zebrafish embryo under control conditions. B) Zebrafish embryo exposed to a non-toxic chemical. C) zebrafish embryo with large oedema, deformed tail and egg sac. D) A dead zebrafish due to exposure to a toxic chemical.

Zebrafish are small organisms that produce hundreds of offspring every week, grow very fast, and have lower maintenance costs compared to other experimental animals such as mice. Zebrafish embryos are transparent and develop outside the mother's body, which allows the development of internal structures to be evaluated, i.e. blood vessel defects on zebrafish embryos can be identified using a low-power microscope. This allows the evaluation of embryonic development at every stage of embryogenesis. Because of their rapid development within the first 48h after fertilisation, zebrafish's major organs such as heart and blood vessels are already completed. Within the first 24h after fertilisation anatomical structures, including somites, notochord and heart anlage, can be localised, and the body length can be measured. By 36h post fertilisation, the heartbeat is pronounced and regular (Barrionuevo *et al.*, 1999), and by 48h craniofacial features are further developed and allow the assessment of cardiac and circulatory systems. Major sensory organs and various brain and spine components can be distinguished, and an increased number of epidermal pigment cells can be seen, as well as the formation of the caudal fin and the buds of the pectoral fins (Teame *et al.*, 2019; Hellfeld *et al.*, 2020).

The low cost, small size, and rapid and external development of zebrafish embryos make them an important model organism for vertebrate development, genetics, immunity, physiology, and human diseases, like neural disorders, cancer, cardiovascular diseases, muscle disorders and DNA damage repair (Magyary, 2018; Teame *et al.*, 2019). In addition, zebrafish are widely used in environmental toxicology to explore the effect of toxic metals, endocrine disruptors and organic pollutants (Domingues *et al.*, 2010; Martínez-Sales *et al.*, 2015; Magyary, 2018).

Technology advancements have allowed the generation of large data sets, genomics, transcriptomics, proteomics, metabolomics and epigenomics (‘omics data), which require the use of bioinformatic methods for their analysis. The use of ‘omics data revealed the effect of chemicals on a molecular level, highlighting alterations in cell biochemistry and physiology, which make them a useful tool in toxicology for identifying biomarkers of adverse effects and toxicity. This will facilitate the prediction of potential adverse outcomes by evaluating whole genome alterations and replacing animal testing. ‘Omics data, such as gene expression profiles, have also been used for chemical classification and class prediction based on chemical effect, and allow identification of genes related to cellular responses and suggest mechanisms of toxicity for each compound (Sauer *et al.*, 2017).

1.2 *In silico* approaches

The large number of new chemicals that need to be assessed and the notion to move away from animal testing, reducing the cost and the experimental time, raise the need for new and more robust risk assessment methods. Computational approaches (*in silico*) are used to analyse, visualise, and predict chemical toxicity *(in silico toxicology). In silico approaches are* used in combination with toxicity tests, by generating predictive models for the potential chemical toxicity based on experimental data, structure-activity relationships and previous scientific knowledge. *In silico* approaches have been used to rapidly evaluate potential chemical effects when experimental data are unavailable (Amberg, 2013; Myatt *et al.*, 2018). Such methods are cheap, fast and high throughput; thus they have been applied to prioritise chemicals for in vitro and in vivo toxicology testing. *In silico* methods, however, are considered a “black box” by many researchers, making it harder to assess the model predictions and reliability. This can be overcome by standardisation of *in silico* tools protocols. The lack of generally accepted *in silico* protocols, including predictive modelling application, database usage and result documentation, had led to inconsistencies in the application of *in silico* approaches. Standardisation of *in silico* approaches protocols will describe the prediction process in a coherent and well-documented manner, and allow consistent performance and evaluation, ensuring reproducibility and enhancing the acceptability of the methods and their results (Myatt *et al.*, 2018).

1.2.1 Machine learning approaches

Machine learning is the process of using algorithms to identify patterns in datasets in order to fit predictive models or identify informative groupings within the dataset (Greener *et al.*, 2022). The resulting models can then be used to predict the response (e.g. toxicity endpoints) of data not used in the generation of the model (Patterson, 2017). Machine learning approaches are divided into supervised and unsupervised based on the input data. Supervised machine learning requires labelled data as input; the values to be predicted by the model are provided. On the other hand, unsupervised machine learning is used to explore patterns in unlabelled data (response is unknown), like grouping data based on similarities in expression (Greener *et al.*, 2022) (Figure 1.2).

Supervised machine learning approaches include regression and classification analysis. Regression analysis is used to predict the dependent variable of continuous data, using a set of independent variables, whereas classification analysis is used to predict defined classes. Regression analysis includes linear regression, polynomial regression, and random forest regression. Linear regression analysis is used to select one or more features (multivariable linear regression) required to explain the dependent variable (only continuous) when the relationship between them is linear. However, it should be used with caution with large datasets because it is sensitive to outliers. Logistic regression, on the other hand, is an extension of linear regression and can be used with large datasets but only with binary dependent variables, and low correlation between the independent variables. Polynomial regression is used for converting non-linear data points to polynomial features and uses a linear model for predictions (Schneider *et al.*, 2010; Myatt *et al.*, 2018).

Those models fail when the relationship between variables and outcome is not linear or when the independent variables interact with each other. This type of data requires decision tree methods that can be used for classification and regression analysis, such as random forest (RF). The popularity of RF-supervised machine learning relies on the good predictive abilities of the models, simplicity, robustness, low overfitting, and easy interpretability for both classification (categorical variable) and regression analysis (continuous variables). Features such as the evaluation of prediction accuracy, and the calculation of how each variable contributes to the model decision, make RF a strong candidate for toxicity prediction studies (Breiman, 2001; Myatt *et al.*, 2018; Degenhardt *et al.*, 2019).

Classification algorithms include support vector machines (SVM), random forest and k-nearest neighbours (k-NN). Support vector machines can be used on linear data, where the data domains can be divided linearly to separate the classes, and non-linear data needs to be transformed into a new data domain that can be divided linearly, for classification and regression. SVM generates a hyperplane that best divides a dataset into classes (Suthaharan, 2016). The k-NN algorithm is one of the oldest and simplest methods used in pattern classification. K-NN uses distance metrics, usually Euclidean distance, to calculate similarities between items based on the assumption that similar points can be found close to one another (Kilian Q. Weinberger, 2009).

Unsupervised machine learning approaches, on the other hand, include clustering and association analysis. Clustering analysis groups items based on distance measurements, thus items that are found close to each other are grouped together (Patterson, 2017). Unsupervised clustering methods include k-means and hierarchical clustering. The k-means algorithm is based on the sum-of-squares criterion, but the number of clusters needs to be defined from the beginning (Sinaga *et al.*, 2020). Hierarchical clustering analysis has been widely used in big data analysis and data mining. It can be categorised in two ways, agglomerative (bottoms-up) and divisive (top-down) (Ross et al., 2009). The agglomerative approach isolates the data points as separate groups and then they are merged together based on similarity. Euclidean distance is the most common metric used for calculating the similarities between the data points, but other measurements such as Manhattan distance have also been applied. On the other hand, in divisive clustering, the analysis starts with a single cluster that is divided based on the differences between the data. To visualise these types of clustering methods, a dendrogram is used that shows the merging or the splitting of data points at each iteration (Yogita Rani, 2013). Association analysis is used to identify the relationship between the variables in a dataset, such as dimensionality reduction methods that identify patterns or correlated variables to exclude any redundant information, and a priori algorithms that are used to identify repeated patterns (IBM Cloud Education, 2020).

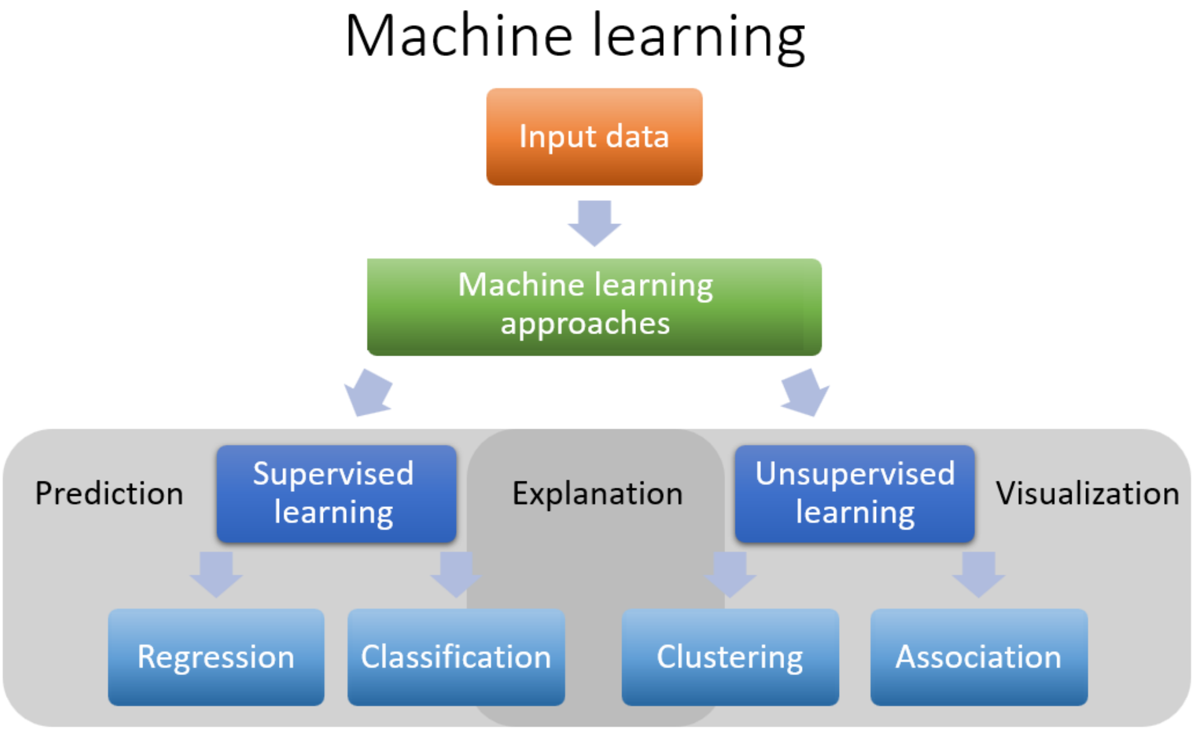


Figure 1.2: Machine learning approaches and their use in predicting biological properties. These include filling knowledge gaps, pattern identification, generation of models that can explain the observed phenotype, and data visualization. The techniques can be split into supervised and unsupervised-based approaches. Supervised classification and regression models are used for the prediction and explanation of the data. Unsupervised clustering and associating techniques are used in data explanation and visualization.

To evaluate the predictive ability of each model, the coefficient of determination, R^2, has been introduced. This value measures model performance, by comparing the observed and predicted values of the dependent variable. Thus, the value of R^2 indicates the percentage of variance in the dependent variable that can be explained by the features of the model (Schneider *et al.*, 2010). At the same time, cross-validation methods have been introduced to evaluate the predictive validity and compare models. The hold-out validation requires splitting the dataset into training and test sets. The training set is used to generate the predictive model and consists of independent variables and a dependent variable to be predicted. Once a model is generated, its efficiency is measured by its ability to predict the response of new data not included in the training step, the test dataset. In k-fold cross-validation the dataset is split into a number of sets of equal sizes (k) and the model is trained using k-1 groups, and the other group is used as a test set. The process is repeated until all groups are used as test sets. Leave-one-out cross-validation is a type of k-fold cross-validation, where the model is trained on almost all the data except for 1, and then the model is used to predict that single observation (Refaeilzadeh P *et al.*, 2009). The aim of predictive modelling is to identify patterns in the training dataset and use those patterns to predict the test set's dependent variables (Dhall *et al.*, 2020). Q^2 is the correlation coefficient calculated after cross-validation of the model (Mansouri *et al.*, 2013). The closer the R^2 and

Q^2 are to 1 the more accurate the prediction of the model is. The main concern when generating a predictive model is overfitting or underfitting. Underfitting models usually consist of a very small number of variables and fail to predict the response of both the training and test set accurately (low R^2). In contrast, complex models, with a large number of variables can predict the response of the training dataset with high accuracy, thus a very low error rate, but fail to do the same with the test dataset, this is usually an indication of overfitting (Patterson, 2017).

High throughput data suffer from high dimensionality, where the number of variables is larger than the number of samples, thus various machine learning methods have been developed for feature selection to identify only relevant data and exclude noise (Chowdhury *et al.*, 2020). Such methods aim to select the smallest number of variables required to build an accurate and reliable predictive model or identify all the variables involved in a response. The selection of appropriate variables for model generation is an important and difficult part of building a model. The selection of too few variables usually leads to a model that fails to capture the true relationship between the dependent and independent variables, underfitting. On the other hand, models with a large number of features, are more dependent on the observed data, may have reduced predictive power and fail to be generalised. Such models are usually characterised by overfitting. Simple models with fewer variables are preferred over complex models with many variables since they are easier to interpret, generalise and use (Guyon *et al.*, 1999; Chowdhury *et al.*, 2020). Variable selection aims to identify a subset of variables which can be used in predictive modelling and rank them based on their predictive power. Thus, variable selection methods provide a balance between simplicity and fit.

Lasso (least absolute shrinkage and selection operator) regression, a supervised machine learning approach, is used to reduce the complexity of the model by identifying a few features from the dataset that can be used in building the model while setting the coefficients of less contributing features to zero (or close to zero) (Mozafari *et al.*, 2020). This allows for faster variable selection by reducing the number of features. Stepwise selection methods, forward and backward, are also used to reduce the number of unnecessary variables, feature reduction. Forward selection is a stepwise approach that evaluates the contribution of a variable to the model one by one and keeps only those that help explain the dependent variable. On the other hand, backward selection is removing a variable one by one to identify the ones that are important for prediction (Schneider *et al.*, 2010; Myatt *et al.*, 2018). The various selection methods, use selection criteria for the inclusion or exclusion of features, such as the Akaike information criterion (AIC) which compares multiple models and estimates the relative information loss between them, and the Bayesian information criterion

(BIC) that penalises models also based on the number of features utilised by the model. Variable selection improves model prediction performance, reduces computation and utilisation time, facilitates data visualisation and by reducing the complexity of the model can potentially provide new insights into the underlying biological processes (Bozdogan, 2000; Lee *et al.*, 2014).

High throughput data allow a system-wide evaluation of chemical effects, identifying genes, proteins and biochemical reactions as interactive networks and linking them to biological processes and adverse effects. This allows the identification of new insights for multiple biological processes without prior knowledge, filling any knowledge gap (Garcia-Reyero *et al.*, 2011). ‘Omics data have been used in ecotoxicology to describe the underlying molecular mechanism of adverse effects (Denslow *et al.*, 2007), and identify protein-protein interactions and pathways involved in a toxic response (Martyniuk *et al.*, 2009). Large or complex datasets, such as “‘omics data", require computational approaches for analysis and pattern identification, increasing the popularity of machine-learning approaches (Greener *et al.*, 2022). Machine learning approaches generate reproducible and time-efficient pipelines, through a data-driven workflow (Wu *et al.*, 2022), and can utilise existing toxicological and chemical data, reducing the need for repeating experiments. The aim of machine learning approaches is to link chemical exposure to biological alterations, by generating predictive models and assisting in the identification of the relationship between toxins and environmental effects.

1.2.2 Quantitative Structure-Activity Relationship models

Computational and mathematical methods have been used to correlate structural properties (physicochemical characteristics) of a chemical to biological activities, Quantitative Structure-Activity Relationship models (QSAR). QSAR methods generate predictive models from molecular descriptors characterising the physicochemical properties (molecular weight, lipophilicity, absorption, distribution, excretion) electronic and topological state, and structure of a chemical (functional groups). The presence of ester, amides, hydroxyl, aldehyde, and carboxylic groups have been associated with biodegradation ability, and halogen groups, chain branching and nitro groups with chemical biodegradation time (Mansouri *et al.*, 2013). On the other hand, water-soluble chemicals are more easily degraded compared to insoluble chemicals, and heavier chemicals are harder to move into the cell (Boethling, 1996). QSAR internal validation depends on statistical measures (R^2 and Q^2) or Leave-one-out cross-validation (LOOCV), which explores the stability of the model by excluding one chemical at a time before model generation. However, the reliability of the model in predicting the toxicity profile of a new compound relies on the applicability domain of the

QSAR model, where the new chemical needs to be similar to the chemicals from the training set (Mansouri *et al.*, 2013; Carrió *et al.*, 2014; Patlewicz *et al.*, 2016). QSAR modelling has been used in various scientific disciplines, including chemistry, biology, and toxicology, in predicting biological activities, classifying chemicals and early prioritisation and exclusion of potentially harmful chemicals in drug discovery processes (Perkins *et al.*, 2003). However, the effect of some chemicals, such as endocrine disruptors, cannot be predicted solely by their structural characteristics, due to the high heterogeneity among the chemicals of such groups, chemicals with similar structural features act through different mechanisms (Perkins *et al.*, 2003; Futran Fuhrman *et al.*, 2015; Martin *et al.*, 2015).

1.2.3 The mode of action

Identifying and understanding the various toxicological mechanisms associated with chemical exposure, increases the accuracy of chemical clustering improving toxicity prediction (Kienzler *et al.*, 2017). The mode of action (MoA) of a chemical describes the series of chemical-specific events, from exposure to an observed effect, based on experimental and mechanistic data. Verhaar et al. used fish acute toxicity to split chemicals into four classes based on the MoA, baseline toxicity and narcosis (class 1), less inert chemicals (class 2), reactive chemicals (class 3) and specifically acting chemicals (class 4). They also identified structural features that can be used as rules for predicting mode of action; chemicals that do not follow those structural rules are placed in class 5 (Verhaa *et al.*, 1992; D. Villeneuve *et al.*, 2014; Kienzler *et al.*, 2017). Clustering chemicals using MoA information has been used in predictive models (QSAR) and chemical risk assessment (Yuan *et al.*, 2007; Martin *et al.*, 2015). Studies have shown that the toxicity of chemicals belonging to nonpolar (MoA class 1) and polar narcosis (MoA class 2) can be predicted using the octanol-water partition coefficient (Van Sprang *et al.*, 2013), but class 3 contains chemicals that act through multiple reactive mechanisms which increase the difficulty of predictive modelling. In addition, it has been shown that chemicals that follow the same structural rules, thus cluster together using MoA classification, might act through a different mechanism, and vice versa (Russom *et al.*, 1997; Martin *et al.*, 2015; Ellison *et al.*, 2016).

1.2.4 The Adverse Outcome Pathway

The Adverse Outcome Pathways (AOPs) framework has been proposed to help organise the existing knowledge related to toxicological effects starting from the interaction of a chemical with a biological target (molecular initiating event; MIE), to an adverse outcome (AO). At the same time, AOPs describe the various measurable/observable biological changes (key events) that occur at the molecular level and the experimentally defined relationship (key event relationships) between them in sufficient detail (Ankley *et al.*, 2010). Key events are measurable changes in the biological state compared to control, whereas key event relationships are defined by biological plausibility and allow the prediction of downstream key events based on the known upstream key events. AOPs describe the biology and the relationship between biological events and rely on the assumption that when any chemical triggers a specific molecular initiating event, it will follow the chain of defined key events if the exposure duration is sufficient, to elicit an adverse outcome (Ankley *et al.*, 2010; Sauer *et al.*, 2017) (Figure 1.3).

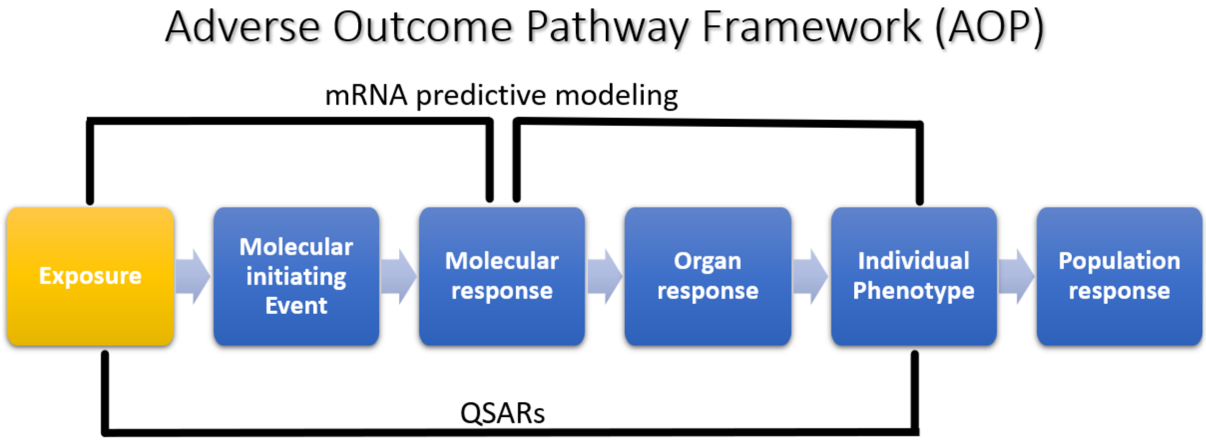


Figure 1.3: Representation of Adverse outcome pathway framework. A diagram with the key features of an adverse outcome pathway (AOP). An AOP begins with the molecular initiating event (MIE) that describes the chemical interaction with the biological target initiating a series of molecular responses leading to an AO. QSAR models have been used to directly link exposure information with phenotype. Molecular responses ('omics data) can be used as a mediator between exposure characteristics and phenotype.

There are three AOP strategies, and their use depends on the type of available information. The top-down strategy is performed when either an adverse outcome is observed but without an understanding of the underlying biological mechanisms involved, thus key events and the relationship between them, are identified using experiments or literature reviews (Ankley *et al.*, 2009), or when the MIE has been well characterised, for example through QSAR models, but the toxic effect of such interaction is not clear (L. Zhang *et al.*, 2013). The bottom-up strategy aim is the description of the multiple key events following the MIE (receptor-ligand) when the chemical concentration and exposure time are sufficient to have a toxic effect (D. L. Villeneuve *et al.*, 2014). Finally, the middle-out strategy describes the situation where a key event is characterised, but without further knowledge about the MIE or an adverse outcome, for example, transcriptional changes due to chemical exposure, but no information is provided about the MIE or the potential adverse outcome (D. L. Villeneuve *et al.*, 2014; D. Villeneuve *et al.*, 2014).

When working with AOPs it is important to remember that specific key events and key event relationships do not correspond to a single AOP, and multiple independent MIEs can be upstream connected to a single key event and a common adverse outcome, just like in biological systems where interaction and crosstalk between multiple pathways are common. This also suggests that key events and key event relationships only need to be defined once and can be reused in new AOP descriptions. However, AOPs are usually presented as a single chain of key events, to simplify the available toxicological information related to risk assessment. AOPs aim to provide more information about the underlying molecular interactions (key events and key events relationships) that link an MIE to an adverse outcome. As new tools are being developed, new key events and key event relationships are identified or measured better, thus AOPs are continuously growing (D. L. Villeneuve *et al.*, 2014). By combining the information on the various key events, the multiple AOPs can be linked into a single larger AOP network. Such AOP networks can then provide means for identifying potential new AOPs, where new MIE to adverse outcome linkages are established, and therefore support risk assessment more broadly.

Chemical classification, chemical properties and structural information (molecular descriptors) can assist in identifying or predicting chemical interactions surrounding molecular initiating events. From existing AOPs and their networks, signals can be followed through downstream key events, and likely identify adverse outcomes. QSAR models use structural information to predict the interaction between a chemical and a target biomolecule, such as receptor-ligand interaction (MIE), which is linked to an adverse outcome, and the MoA of a chemical (D. L. Villeneuve *et al.*, 2014; Martin *et al.*, 2015; Ellison *et al.*, 2016). Structural information such as absorption, distribution, metabolism, and elimination can be used to predict molecular initiating events interaction. Pharmacokinetic factors and potency of the chemical influence the duration and perturbation at the molecular initiating event (D. L. Villeneuve *et al.*, 2014). Both the MoA and AOPs frameworks describe a series of measurable biological events that are essential for the initiation and progression of toxicity, in risk assessment. However, AOPs are endpoint oriented, whereas MoAs are substance-specific. Combining traditional molecular biology and *in silico* approaches, as described earlier, such as AOPs, and their networks, can support toxicity prediction and risk assessment and thus minimise the need for extensive animal testing and reduce the cost and duration of toxicity tests (Raies *et al.*, 2016).

1.3 Scope of the thesis

Various *in silico* approaches have been developed that group chemicals based on similarities, including MoA classification using physicochemical properties, and signature-matching approaches (toxicogenomic) using gene expression profiles. These methods facilitate risk assessment and enable the move away from time-consuming and costly animal experiments. MoA classification has been very popular in risk assessment since only physicochemical information, such as molecular weight, lipophilicity, bond composition, and chemical polarity, are required. However, despite the development of numerous MoA frameworks, a large number of chemicals cannot be classified, and there is a high inconsistency between the results of those frameworks (Kienzler *et al.*, 2017; Kienzler *et al.*, 2019). On the other hand, the use of gene expression profiles for toxicity assessment, chemical clustering, and chemical prioritisation (toxicogenomic), is becoming more popular in risk assessment (Bourdon-Lacombe *et al.*, 2015). Those classification methods rely on the assumptions that the structure of a molecule is responsible for its physical, chemical and biological properties and that chemicals with similar gene expressions, will cause similar biological effects (toxicogenomic).

To test these assumptions, this project characterised a set of 143 chemicals in zebrafish larvae and then clustered these based on structural information (molecular descriptors), and gene count profiles. Two endpoints were defined by the physiological information provided in this dataset 1) heart rates and 2) chemical concentration. Clustering chemicals based on the molecular descriptors and transcriptional response was evaluated in the context of these endpoints as well as using the more traditional Verhaar MoA classification. Further effort to link these two was undertaken to examine whether gene expression clusters could be predicted by molecular descriptors.

Furthermore, this study evaluated the ability of molecular descriptors (QSAR model) and gene expression data to predict changes in zebrafish heart rates across all measured chemicals. To explore whether there are subgroups of chemicals that are more easily predicted by either of the two datasets, chemicals were grouped based on the clustering performed. According to the literature, multiple molecular descriptors have been associated with toxicity, however, QSAR models have a small applicability domain, since they are usually developed using chemicals from a single MoA class. In addition, there is limited available information on QSAR models or gene expression data being used in predicting cardiotoxicity. The model generated here will evaluate the ability of QSAR models and gene expression data to predict chemical toxicity of a highly diverse chemical dataset, and allow the identification of molecular descriptors and genes that can be used in predicting toxic effects.

While transcriptomics has been widely used to generate large sets of data that can be assessed to identify differentially expressed genes between two conditions (Anjum *et al.*, 2016; Costa-Silva *et al.*, 2017; McDermaid *et al.*, 2019), the utilisation of differentially expressed genes alone cannot describe the underlying biological mechanism since gene data usually cannot explain an entire functional trait (Ramanan *et al.*, 2012). In addition, in some cases, compounds do not significantly influence gene expression, leading to transcriptional signals dominated by noise that does not represent the effect of the chemicals on the organism. One approach to extract mechanistic information from a long list of differentially expressed genes and proteins is pathway analysis which groups genes into functional pathways through gene interactions using prior knowledge. The identification of an active pathway has more explanatory power than a set of differentially expressed genes or proteins (Glazko *et al.*, 2009; Ramanan *et al.*, 2012).

To put pathways into context, and provide an anchor for an improved understanding of the underlying biology, the AOP concept has been developed, particularly as an alternative to traditional risk assessment (Ankley *et al.*, 2010). This study makes use of this development to show that molecular and chemical data, given suitable endpoints such as heart-rate changes, LC50, or even exposure concentration, can be used to develop potential AOPs de-novo.

Compare to other studies, a highly diverse dataset was used in this work, which consists of chemicals with high variability in structural characteristics and molecular responses after exposure, which can potentially increase the applicability domain of the generated models. Chemical clustering is widely applied in risk assessment using structural characteristics, however, in this study comparing structural and molecular clustering, using a set of highly diverse chemicals, the aim is to evaluate the ability of structural clustering to group together chemicals that potentially have a similar molecular effect (gene expression profiles) on the zebrafish. Molecular descriptors and gene expression profiles have been very popular in predictive modelling, but there is limited available information on cardiotoxicity models in the literature, especially for predicting heart rate changes in zebrafish embryos. In addition, such studies use chemicals from the same MoA for the generation of predictive models, whereas in this work the dataset consists of highly diverse chemicals. Thus, this study aims to generate predictive models using molecular descriptors or gene expression profiles to predict the effect chemicals have on zebrafish heart rate, which can potentially assist in *in-silico* chemical risk assessment. Finally, the same set of chemicals and gene expression profiles are used in pathway analysis, where by organising the gene expression profiles into pathway activity information, the aim is to identify pathways that can be used to predict chemical toxicity and heart rate fold changes and arrange them into pathway networks which can potentially assist in building cardiotoxicity AOPs.

The work performed in this study is presented in the theses in the form of papers. Three manuscripts were prepared with the intention to be submitted to journals and are added to the thesis as Chapters 2,3 and 4.

Chapter 2

Disparity between structural and molecular effect clustering of chemicals

2.1 Abstract

Risk assessment of new chemicals is costly and time-consuming, thus computational approaches that can predict the toxicity of a compound based on similarities with already defined chemicals have been developed. Identifying molecular descriptors and genes associated with heart rate changes in zebrafish embryos and the increased concentration, chemicals were clustered based on similarities. Molecular descriptors such as polarity, lipophilicity, atoms composition and chemical bonds were found to be an indicator of cardiotoxicity and chemical concentration exposure. Genes involved in cell-to-cell signalling, membrane depolarization, cell death, muscle proliferation, muscle specification and morphogenesis were found to be differentially expressed during toxic chemical exposure and associated with heart rate changes in zebrafish embryos.

Classifying chemicals using the information encoded in the molecular structure (molecular descriptors) has been widely used with many advantages in risk assessment, however, clustering compounds using molecular responses (sequencing data: gene count profiles) grouped chemicals differently. The results of the two clustering methods, molecular descriptors and sequencing data, were compared to the Verhaar MoA classification and their ability to group chemicals based on their effect on zebrafish heart rate. Chemical classification based on structural information (molecular descriptors), was representative of MoA classification, where chemicals from the same MoA class tend to be grouped together. In addition, chemicals that significantly alter zebrafish heart rates are also grouped together using molecular descriptors for chemical clustering. On the other hand, grouping chemicals using gene count profiles (mRNA-seq data) was highly associated with chemicals’ effect on zebrafish heart rate, but no association can be seen with the Verhaar MoA classification.

Even when combining the structural features (molecular descriptors) and transcriptional responses (mRNA seq-data) to predict heart rate the models always tend to associate significantly with heart rate but not with Verhaar MoA classification. This suggests that one of the most used *in silico* mode of action approaches lacks the ability to represent adverse outcomes sufficiently and that more molecular approaches, such as transcriptional responses, are necessary to understand the underlying biological mechanism of chemical toxicity.

2.2 Introduction

The rapid development of the manufacturing and pharmaceutical industry has increased the abundance of numerous chemicals in the environment (Corrales *et al.*, 2015; Przybylińska *et al.*, 2016). Despite the best efforts to understand the impact of these compounds on humans and other organisms, many of them only hold limited toxicological data. Given that the number of compounds likely exceeds 100,000, it is unlikely that each compound can be tested in depth to characterise its impact on multiple organisms. As a result, computational approaches have been developed to describe chemical structural characteristics and associate them with regulatory-relevant endpoints such as lethality. Classifying chemicals based on their MoA, using structural information, has been widely used in risk assessment. Multiple toxicity mechanisms have been identified, however, most of the industrial chemicals follow either polar or non-polar narcosis and, on a few occasions, toxicity is caused by irreversible covalent bond formation (Russom *et al.*, 1997; Bearden *et al.*, 1998; Aptula *et al.*, 2006).

MoAs define the more general biological process which causes specific adversity. Several approaches have been proposed to simplify and predict these MoAs. One of the most applied classifications is defined by Verhaar (Verhaar *et al.*, 1992; Russom *et al.*, 1997). Verhaar splits chemicals into 5 classes based on structural features. Chemicals from class 1 exhibit non-polar narcosis or baseline toxicity and do not interact with specific receptors, they form non-covalent and reversible alteration at the site of action; their toxicity depends on their hydrophobicity. Class 2 chemicals are less inert, cause polar narcosis, possess strong electron-releasing substituent and aromatic structures, and form hydrogen bonds, their toxicity cannot be predicted by hydrophobicity alone. Class 3 of MoAs consists of reactive chemicals with enhanced toxicity, by forming irreversible covalent bonds with amino acid protein residues. Class 4 consists of chemicals that act by a specific mechanism, in a non-covalent manner. The chemicals that do not follow any of the structural criteria set for the previous classes are grouped in class 5 (Verhaar *et al.*, 1992).

MoA classification relies on the same premises as QSAR models, the toxicity of a new chemical can be predicted using structural information (molecular descriptors) and chemicals with already defined toxicity. QSAR models utilise multiple statistical and machine learning techniques to predict the physicochemical, biological, and environmental impact of various compounds using molecular descriptors (Kwon *et al.*, 2019). Molecular descriptors result from logical and mathematical procedures, which transform chemical information encoded within a symbolic representation of a molecule, into numbers, or the result of some standardised experiment (Todeschini *et al.*, 2009). The major types of descriptors include topological representation, connectivity of atoms, presence and nature of chemical bonds and physicochemical properties (Svetnik *et al.*, 2003; Singh *et al.*, 2013; Agatonovic-Kustrin *et al.*, 2014; Roy *et al.*, 2015). Multiple software can calculate thousands of descriptors such as Dragon 7 (Mauri *et al.*, 2006). Understanding the nature of molecular descriptors increases the interpretability of QSAR models. The classical machine learning approach for QSAR is a linear regression technique, however, not all biological properties are linear in nature. Thus non-linear techniques such as artificial neural networks and random forest have been applied (Svetnik *et al.*, 2003; Singh *et al.*, 2013; Agatonovic-Kustrin *et al.*, 2014; Roy *et al.*, 2015; Drgan *et al.*, 2016).

High variability in molecular descriptors can increase the complexity of QSAR analysis and generate models with low power and a large number of molecular descriptors. The use of chemicals from a single MoA class shows higher accuracy since they cover more similar chemical domains (Yuan *et al.*, 2007; Michielan *et al.*, 2010; Cassotti *et al.*, 2015). QSAR methods are being increasingly used in screening, testing prioritisation, hazard identification, and risk assessment, to reduce the number of experiments and testing (Cherkasov *et al.*, 2014). As MoAs are defined by their molecular interaction it stands to reason that the inclusion of molecular information can increase the reliability of structural-based predictive modelling.

To assess the large number of chemicals fast and accurately using fewer resources and experimental animals, new computational, molecular, and in vitro tools have been generated, increasing the types and amount of information available. Advances in high throughput screening techniques and genome-wide expression analysis, enable the development of signature-matching approaches that are based on the assumption that chemicals with similar gene expression signatures will cause similar biological effects and can be used in predicting the toxicity of new chemicals and identifying toxicity-related genes (Lamb *et al.*, 2006; Smalley *et al.*, 2010; Sarmah *et al.*, 2016). RNA-seq data represent the molecular state of the cell, tissue, or organism and therefore can be used as markers to uncover the mode of toxicity. Interestingly, chemical similarity profiles can be very diverse when using gene expression profiles versus structural similarities (Sirci *et al.*, 2017) as for example cis and trans isomers may have significantly different effects on the organism where only one form of the chemical is toxic (Singh *et al.*, 1988; Blisard *et al.*, 1991).

To address the lack of known MoAs and to develop better QSAR models for regulatory use, unsupervised clustering approaches can help to address these challenges using a data-driven approach. Several clustering algorithms have been proposed with varying success in the accuracy and reliability of grouping (Prasad, 2020). Clustering algorithms are divided into hierarchical and partitioning clustering. Hierarchical clustering, e.g., top to bottom, groups samples into a hierarchy of clusters and decomposes the data generating sub-clusters (Murtagh *et al.*, 2012; Murtagh *et al.*, 2017; Prasad, 2020). Partitioning clustering, the simplest and most effective method, uses various measurements (density, distribution, distance) to calculate the variation between samples. K-means, the most widely used partitioning clustering method, where each observation is placed in the cluster with the nearest mean (cluster centre), uses various distance methods to calculate similarities between data points, such as Euclidean or Manhattan distances, in order to split the dataset into a pre-defined number of clusters (Likas *et al.*, 2003; Reynolds *et al.*, 2006; Barioni *et al.*, 2014). However, K-means cannot handle outliers and non-linear data since the mean value is influenced by extreme values. Partitioning clustering is more suitable for large datasets compared to hierarchical clustering (Reynolds *et al.*, 2006). In addition, some algorithms can assign a chemical into multiple clusters (Fuzzy clustering), which is necessary when variables have a high level of correlation, but it is slower and should be used with caution with large data (Prasad, 2020; Baraldi *et al.*, 1999; Gosain *et al.*, 2016).

The most important principle in toxicology is that increasing the exposure concentration of a potentially toxic compound increases the probability of occurrence and severity of an adverse effect. Thus, the dose-response principle has been widely used in chemical risk assessment (Andersen *et al.*, 2005; Holsapple *et al.*, 2008). On the other hand, a large number of drugs were withdrawn because they were found to induce cardiotoxicity (Cai *et al.*, 2019; Ma *et al.*, 2020). Various environmental pollutants, such as heavy metals, pesticides (Georgiadis *et al.*, 2018) and multiple drugs including aspirin and terfenadine (Milan *et al.*, 2003; Zhu *et al.*, 2014), affect heart development and function (R. Li *et al.*, 2020). Heart development is a very sensitive process that can be affected by molecular, cellular and environmental factors (Sarmah *et al.*, 2016). Exposure to such toxins may lead to arrhythmia and myocardial hypertrophy, or damage to the heart muscle and other cardiac tissues. Zebrafish embryos are widely used in assessing cardiotoxicity, due to their rapid cardiac development and embryonic transparency, which allow non-invasive evaluation of heart function (Fraysse *et al.*, 2006; Scholz *et al.*, 2008; Sarmah *et al.*, 2016; Zakaria *et al.*, 2018).

Here the aim is to compare how chemical structural information (molecular descriptors) and associated molecular data cluster chemicals into specific groups and how similar these two clustering methods are. A chemical structural dataset was generated using Dragon 7 and in parallel used RNA-seq data after chemical exposure. The two datasets (molecular descriptors and gene count profiles) were used for chemical clustering and the association of structural descriptors and genes to adverse outcomes, points towards a more complex relationship between chemical structure and adverse outcomes.

2.3 Methods

2.3.1 Chemical selection and exposure

Chemical selection was governed by retrieved chemical lists from the UK, Canada, EU, and the USA high concern priority compound lists. 258 compounds were selected to ensure that a) they have previous data in other species, b) were structurally varied to other compounds, and c) were structurally singular and could be used to calculate structural features.

Zebrafish (*Danio rerio*) embryos were exposed to 258 chemicals for 72 hr using 96-well plates. For every chemical, 6 concentrations were tested (LC50, LC5, LC5/2, LC5/4, LC5/8, LC5/16) with 6 individual embryos in separated wells for each concentration; 36 zebrafish embryos in total were exposed to each chemical (6 concentrations, 6 replicates). In each well plate, two chemicals were tested along with 6 zebrafish embryos exposed to control and 6 embryos exposed to DMSO, 84 zebrafish per well-plate, and 10836 zebrafish embryos in total (129 plates,774 controls exposure, 774 DMSO exposure, 9288 chemical exposure). However, since most of the chemicals were dissolved in DMSO, DMSO was also added to all the samples and the DMSO exposure was used as a control for further analysis. Chemicals that failed to kill the zebrafish embryos even at the highest concentration (LC50) were excluded along with those that failed to generate a proper dose-response curve, resulting in 156 chemicals for further analysis. These chemicals are used in pharmaceuticals and/or industry and/or as pesticides/insecticides, some chemicals are used in multiple products such as Diethanolamine which is used in pesticides, pharmaceuticals and polishers (Table 2.1).

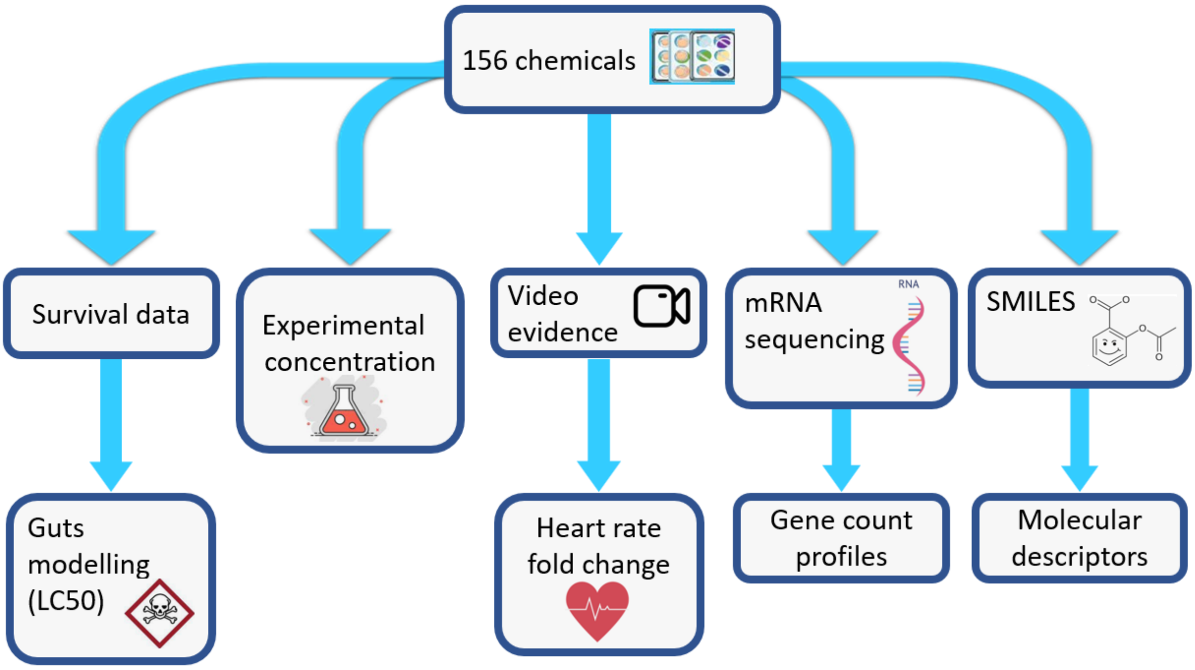


Figure 2.1: Flowchart of data acquisition. Zebrafish embryos were exposed to 156 chemicals. The number of embryonic deaths for each concentration (survival data) were used to calculate the LC50 of each chemical. Video evidence was used to calculate heart rate fold change. Zebrafish embryos underwent mRNA sequencing generating a gene count profile for each chemical exposure. A smile notation was acquired for each chemical and used to calculate molecular descriptors.

Pharmaceutical	**Industrial**	**Insecticides/Pesticides**
Acetaminophen	Acetaminophen	Warfarin
Anagrelide	Aspirin	P-Aminoazobenzene
Androstenedione	Dichlorophene	Pentachlorophenol
Anisindione	Diphenhydramine	2,2'-Dichlorodiethyl ether
Aspirin	Ibuprofen	O-Cresol
Benzamide	Melatonin	P-Chloroaniline
Bupropion	Phenacetin	Urethane
Busulfan	Progesterone	1,2,4-Trichlorobenzene
Cetirizine	Testosterone	2-Chlorophenol
Chlorpromazine	4-Hydroxybenzophenone	4-Hydroxybiphenyl
Clofibric Acid	Benzofuran	Alachlor
Clozapine	Catechol	Carbaryl
Dichlorophene	Acrylamide	Chlorobenzilate
Diclofenac	Adiponitrile	Cyanazine
Diflunisal	Benzophenone	Demeton-O
Diltiazem	Bisphenol A	Diazinon
Diphenhydramine	Dibutyl Phthalate	Dichlorvos
DL-Norepinephrine	Diisobutyl Phthalate	Diclofop-Methyl
Fenofibrate	Dodecyltrimethylammonium Chloride	Diethanolamine
Finasteride	Epichlorohydrin	Dinoseb
Fluoxetine	Ethylparaben	Diuron
Flurbiprofen	Methyl-4-Hydroxybenzoate	Ethyl Dipropylthiocarbamate
Flutamide	Nitrobenzene	Fenitrothion
Gemfibrozil	O-Anisidine	Fenthion

Pharmaceutical	Industrial	Insecticides/Pesticides
Ibuprofen	O-Dianisidine	Fipronil
Isotretinoin	O-Dinitrobenzene	Malathion
Ketoprofen	O-Phenylenediamine	Molinate
Ketotifen	O-Toluidine	O-Phenylphenol
Medetomidine	Octylamine	Pirimicarb
Melatonin	P-Aminoazobenzene	Prochloraz
Methyltestosterone	P-Cresidine	Resmethrin
Nifedipine	P-Toluidine	Tributyltin Oxide
Phenacetin	Pentachlorophenol	Fenoxaprop-Ethyl
Procarbazine	Perfluorooctanoic Acid	Fluralaner
Progesterone	Propylene Oxide	O-Tolunitrile
Propranolol HCl	Triclosan	1,8-Diamino-p-menthane
Retinoic Acid	Triphenyl Phosphate	2,4-Dichlorophenol
Rizatriptan	1,2-Diaminopropane	Acetochlor
Serotonin	1,3-Butadiene	Folpet
Tacrine	2,2'-Dichlorodiethyl ether	Benomyl
Terfenadine	2,4-Diaminotoluene	
Testosterone	2,4-Dinitroaniline	
Trenbolone	2,5-Hexanedione	
Valproate	2-Methyl-4-isothiazolin-3-one	
Warfarin	3-Ethoxy-4-hydroxybenzaldehyde	
3-Methylpyridine	4,4'-Methylenebis(2-Chloroaniline)	
4-Hydroxybenzophenone	4-Amino-2-Nitrophenol	
4-Hydroxytamoxifen	4-Chloro-o-Phenylenediamine	
Benzofuran	4-Heptylphenol	
Catechol	4-Methyl-2-Pentanone	
Cholesterol	4-Nonylphenol	
MS-222	4-Tert-Butylphenol	
Naproxen	4-Tert-Octylphenol	
Flubendazole	5-Nitro-o-Anisidine	
Estradiol	Aniline	
Caffeic Acid	Dipentyl Phthalate	
Isradipine	Ethanolamine	
Vortioxetine	N,N-Dimethylformamide	
1,2-Diaminopropane	O-Cresol	
Phenol	P-Chloroaniline	
Urethane	P-Hydroxybenzoic Acid	
2-Bromopropane	Phenol	
4-Methylimidazole	Succinic Acid	
Diethanolamine	Urethane	
4-heptyloxyphenol	1,2,4-Trichlorobenzene	
2-(1-phenylethyl)phenol	1,2-Dichlorobenzene	
	1-Hexanol	
	2,6-Dimethylaniline	
	2-Bromopropane	
	2-Chlorophenol	
	4-Hydroxybiphenyl	
	4-sec-Butylphenol	
	4-Methylimidazole	
	N,N-Dimethyl-P-Toluidine	

Pharmaceutical	Industrial	Insecticides/Pesticides
	Butyl Benzyl Phthalate	
	Michler's Ketone	
	4-Nitrobenzamide	
	Nitrapyrin	
	Diethanolamine	
	Diuron	
	O-Phenylphenol	
	Tributyltin Oxide	
	2,4-Dichlorophenol	
	Folpet	
	2-(1-phenylethyl)phenol	

Table 2.1: Chemical grouping of the 156 chemicals A) pharmaceuticals, B) Industrial and C) Insecticides/ Pesticides.

2.3.2 Video recordings

Video recordings were generated daily during those 72 hours, for every embryo used in this study. The videos generated needed to be converted to AVI format for the next step, using the FFmpeg function, an open-source audio and video converter, accessed through the command line (*FFmpeg*, 2018). FFmpeg supports multiple media formats of video and audio files and has been used for editing, video scaling, and decoding (Lei *et al.*, 2013). The resulting videos were analysed using Fiji, an open-source image processing package (Schindelin *et al.*, 2012) using the time series analyser V3 plugin (Available online: https://imagej.nih.gov/ij/plugins/time-series.html). The heart of the zebrafish is selected as the region of Interest (ROI) using the circle tool, and the “add button” is then used for choosing the ROI. Once selected, the “get average” button was used to analyse the pixel change pattern shift at the ROI and extract the heartbeat frequency (Sampurna *et al.*, 2018). This plugin is used for analysing time-lapse images and can be used to get heartbeats per minute by analysing dynamic pixel changes. Since there are 6 embryos for each condition, the mean heart rate was calculated and then compared to the control samples (DMSO) from the same well plate, to estimate the fold change in heart rate and identify the chemicals that significantly affect zebrafish heart rate.

2.3.3 Using SMILE notations to generate molecular descriptors and perform Verhaar classification

For each compound in this study, Dragon 7 software (Mauri *et al.*, 2006) was used to calculate the molecular descriptors, from smile notations identified using ChemSpider (*ChemSpider*, 2018), SigmaAldrich (*Sigma-Aldrich*, 2020) and Pubchem (PubChem, 2018). Dragon software generates 5270 descriptors that represent among others constitutional, topological and connectivity indices, ring descriptors, P-VSA-like descriptors, autocorrelation descriptors, geometrical descriptors, functional group counts and various molecular properties, such as lipophilicity (list of molecular descriptors calculated by Dragon7 (available at: http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf) (*Dragon - Talete srl*, 2018; Worachartcheewan *et al.*, 2015). These descriptors aim to quantitatively describe the physical and chemical information of a molecule. Such descriptors are used in exploring molecular structure-property relationships and similarity analysis (Sawada *et al.*, 2014). The structural information provided in the form of molecular descriptors can be very rich and can be used to showcase the similarity or dissimilarity between compounds. Before clustering, the molecular descriptors data were normalised using log transformation, and any descriptors with low variability (constant or near constant values) or those with missing values (NA), were filtered out. This resulted in a total of 2085 descriptors for further analysis (Chavan *et al.*, 2014).

To investigate how the chemicals are related to each other in terms of toxicity, MoA classification was performed using the Toxtree software (version 3.1.0), an open-source application, that classifies chemicals by applying a decision tree approach (Patlewicz *et al.*, 2008). Currently, Toxtree can perform multiple toxicity estimation schemes, such as Cramer rules, Verhaar classification, Kroes TTC decision tree and Skin irritation schemes, and can process various types of input data, including SMILES, CSV, TXT and MOL files. In this study, SMILE notations were used to assess chemical toxicity using the Verhaar classification scheme (Enoch *et al.*, 2008; Patlewicz *et al.*, 2008).

2.3.4 Generating gene count profiles from mRNA sequencing data

Zebrafish embryos were exposed to 156 chemicals, 6 concentrations, and 6 replicates each. For mRNA sequencing, the 6 replicates were grouped together as one sample resulting in 936 samples, and only five concentrations were sequenced from each chemical exposure (LC5, LC5/2, LC5/4, LC5/8, LC5/16), thus 156 samples were excluded (936-156= 780). However, for certain chemicals calculating the LC50 or the LC5 was harder than expected, due to the steepness of the dose-response curve. The inability to accurately calculate the LC50 and LC5 of each chemical and the additional equipment errors resulted in zebrafish embryos dying even at LC5 and LC5/2. Thus, for 19 chemicals, including Fenitrothion, Flubendazole, Dinoseb and Naproxen, the highest available concentration in the mRNA data was LC5/2 instead of LC5, thus these 19 samples of LC5 were excluded from the mRNA sequencing (780-19=761) (Table 2.2). In addition, for five chemicals, such as Flutamide, Clozapine and Diuron, no sequencing data were provided for either LC5 or LC5/2 since zebrafish embryos died, thus 10 samples representing the LC5 and LC5/2 of those five chemicals were excluded (761-10= 751) (Table 2.2). DMSO samples were used as controls for gene expression profiles, from 35 plates, where each sample consisted of 6 replicates from the same well plate (210 zebrafish in total). As before the 6 replicates were grouped together resulting in 35 DMSO samples for mRNA sequencing. Thus from the 936 samples from chemical exposure and 35 samples from DMSO exposure, and the exclusion of the LC50 samples (156) and the death of the zebrafish that compromise 29 samples, 786 samples were used for mRNA sequencing (936+35-156-29=786).

mRNA sequencing data - Highest available concentration	
LC5/2	***LC5/4***
Fenitrothion	Flutamide
Propylene Oxide	Butyl Benzyl Phthalate
Naproxen	Clozapine
Flubendazole	Anisindione
Dinoseb	Diuron
Fenoxaprop-Ethyl	
Gemfibrozil	
Busulfan	
Pentachlorophenol	
O-Phenylenediamine	
4,4'-Methylenebis(2-Chloroaniline)	
Benomyl	
Medetomidine	
N, N-Dimethyl-P-Toluidine	
P-Aminoazobenzene	
Bisphenol A	
Caffeic Acid	
Prochloraz	
Testosterone	

Table 2.2: Chemicals with missing mRNA sequencing profiles. For 19 chemicals the highest available concentration in mRNA-seq data was LC5/2 and for five LC5/4, instead of LC5.

These 786 samples were moved into RNALater at the end of each experiment and used for Illumina RNA sequencing. Fastq files were analysed using Cutadapt a widely used adapter trimmer (Illumina adapters:AGATCGGAAGAGCACACGTCTGAACTCCAGTCA/ AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT) (Martin, 2011), and STAR software for a two-pass alignment to generate 786 gene count profiles (Dobin *et al.*, 2013). In two-pass alignment, splice junctions are identified during the first alignment and are later used as annotations in the second pass to increase the sensitivity of the method. Compared to single-pass alignment, two-pass alignment with STAR improves splice junction annotation, provides better splice junction detection and reduces read truncation (Veeneman *et al.*, 2015). STAR alignment begins by generating a genome index using GRCz10 zebrafish reference assembly for genome reference sequences (fasta files), annotated transcripts to extract splice junctions information (GTF format), and the length of the genomic sequence was set to 99. Once the genome index is generated, it is used for the first pass alignment, along with the available fastq files. For this step, reads with inconsistent or non-canonical introns are filtered out (--outSAMstrandField intronMotif). The first alignment generates a list of splice junctions (SJ.out.tab), which is then used for generating a genome index again using the fasta file. Finally, a second alignment step is performed, using the second genome index, fastq files and the GTF file. From the second pass alignment a BAM file sorted based on coordinates (--outSAMtype BAM SortedByCoordinate) and a file containing the number of reads per gene (gene counts) (--quantMode TranscriptomeSAM GeneCounts) were generated for every sample (786 files) (Dobin, 2019).

The read count data needed to be normalised before use. The TPMCalculator package (version 0..4) (Vera Alvarez *et al.*, 2019) that calculates transcripts per million (TPM) has been widely used in comparing multiple samples from several experiments since it allows normalisation by transcript gene length.TPMCalculator quantifies mRNA abundance using the same GTF file, used in STAR alignment, and the BAM file generated through STAR. In addition, in this study, the smaller size allowed for an intron set to 90 to exclude small introns (-c 90) and only the properly paired reads were included (-p). Following TPM normalisation, TMM was also performed, a trimmed mean of M-values normalisation method, that is based on the assumption that the majority of genes are not differentially expressed, part of the EdgeR package (version 3.28.1) (Robinson *et al.*, 2010). This function was performed using the gene count profiles (normalised by TPM) and by defining the normalisation method. This double normalisation results in gene expression values that represent the changes in expression relative to a core set of genes (Monaco *et al.*, 2019). The normalisation step is important as it allows us to remove systematic technical effects from raw counts and adjust the gene counts, generating a gene expression profile for every sample.

2.3.5 Chemical screening

2.3.5.1 Batch effect

A source of variation in large-scale experiments is batch effects, defined as the presence of variation in data due to non-biological factors and not scientific variables, such as variations in laboratory conditions, experiments conducted in separated days and the presence of multiple technicians (Leek *et al.*, 2010; Reese *et al.*, 2013). Few methods have been developed for removing such samples, as batch effects can lead to incorrect correlations. Principal component analysis (PCA) is an unsupervised learning approach, that transforms the observations in a dataset into new observations which are uncorrelated with each other and account for decreasing proportions of the total variance of the original variables, with the first principal component explaining most of the variation (Stefatos *et al.*, 2007; Yang *et al.*, 2008; Holmes *et al.*, 2011). PCA has been widely used for visualising batch effects by plotting the first two principal components against each other (Reese *et al.*, 2013). PCA was performed on the RNA-seq data, in order to identify potential batch effects, resulting in six chemicals being removed before further analysis, Naproxen, Acetochlor, Estradiol, Folpet, Isradipine and Vortioxetine (Figure 2.2). The resulting dataset consists of 150 chemicals.

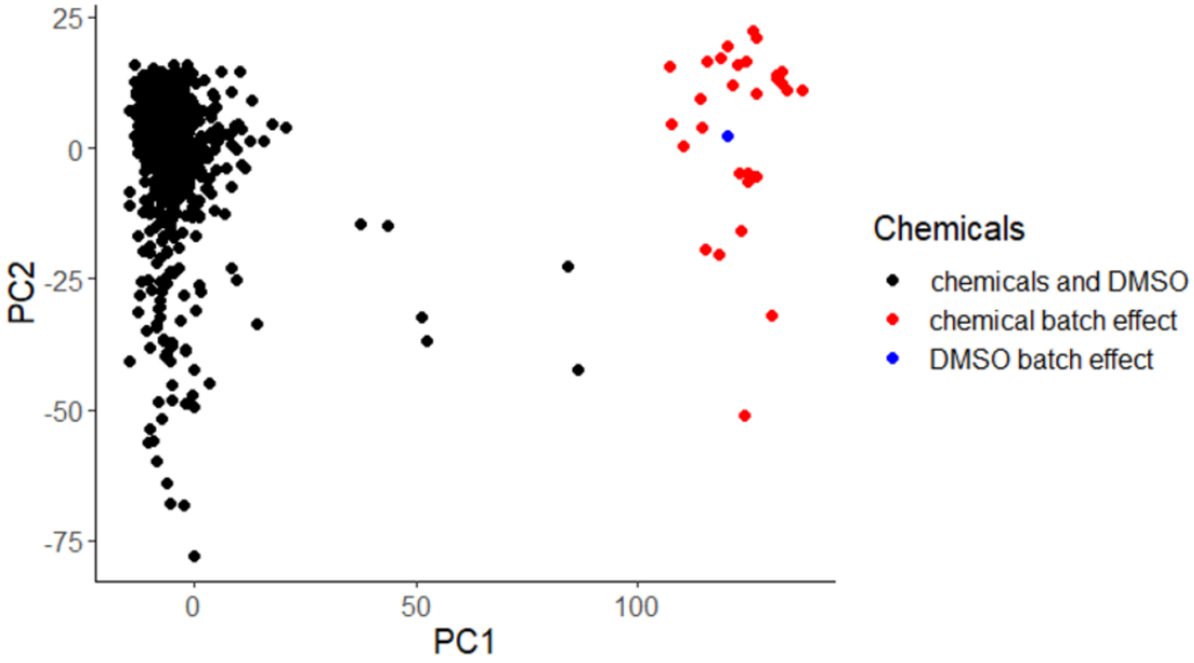


Figure 2.2: PCA plot for identification and visualisation of batch effect. PCA analysis was performed using the gene count profiles of the 156 chemicals in order to identify any batch effect among the data. The red colour represents chemicals that were excluded due to batch effect (five concentrations each). Black colour represents the rest of the available chemicals and blue is the corresponding DMSO, whose variation cannot be explained by biological factors, batch effect.

2.3.5.2 Survival data

Survival data, a record of the number of embryos that died during the experiment at each concentration, were used for calculating the concentration at which 50% of the population is dead (LC50). General Unified Threshold model of Survival (GUTS) modelling uses a combination of toxicokinetic-toxicodynamic (TKTD) models to predict the probabilities of survival. The external concentrations are translated to internal concentrations and associated with the likely damage which triggers death. This, therefore, ensures that the concentrations leading to 50% of death (LC50) are directly based on the likely internal concentrations. GUTS modelling consists of four modules, internal concentration (toxicokinetic model) and damage, hazard rate and threshold distribution (toxicodynamic models) (Figure 2.3) (Jager *et al.*, 2011). GUTS modelling offers stochastic dead (SD) and individual tolerance (IT) methods, based on different assumptions for those TKTD models. IT approach is based on the assumption that a percentage of the population will survive indefinitely, implying that individuals have different sensitivity levels, thus the first pulse is killing off the most sensitive individuals and subsequent pulses will have less and less effect on the population (Jager *et al.*, 2011). On the other hand, the SD approach assumes that all individuals are identical, and mortality is a stochastic process where the individuals that die are not more sensitive, and any exposure that causes some mortality can eventually affect the whole population given enough time. GUTS modelling uses all the available data to estimate model parameters and does not require constant exposure concentration (Ashauer *et al.*, 2010).

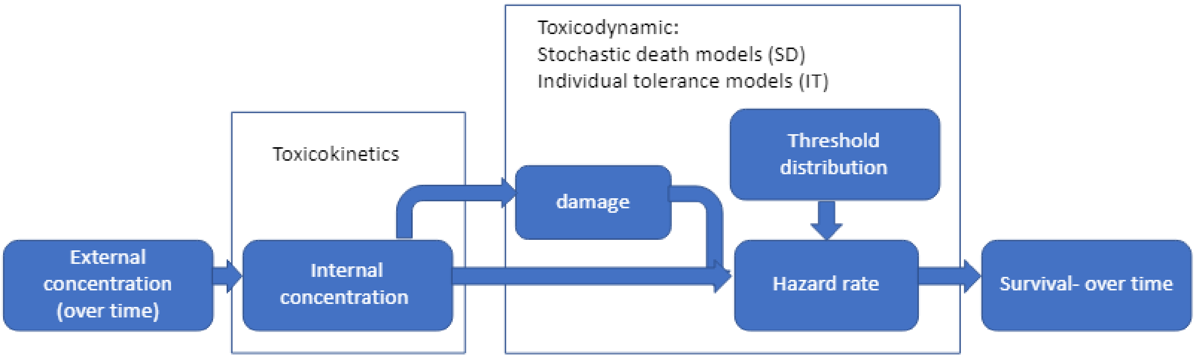


Figure 2.3: Workflow of GUTS modelling. GUTS consist of toxicokinetics and toxicodynamic models and allow for both Stochastic death models (SD) and Individual Tolerance models (IT). GUTS modelling begins by calculating the internal concentration and then by identifying the damage and hazard rate, it calculates the estimated probability of survival over time.

The openGUTS, a free and open-source software, is based on likelihood, and through the use of a genetic algorithm, and likelihood profiling, it explores the parameter space to find the best-fitting parameters. The reduced models for stochastic death (SD) are implemented through the openGuts application, (Jager, 2019) (Figure 2.4). Damage as a function of time is calculated using the recovery rate (k_d), exposure concentration (C_w) and scaled damage (D_w) that is proportional to the damage level (equation 1) (Jager, 2020).

$dD_w/dt = k_d (C_w - D_w)$ with $D_w (0) = 0$ (1)

The hazard rate due to chemical stress (h_c), which is the probability to die, represents the continuous changes in probability over time. When the scaled damage (D_w) is lower than the threshold (m_w), the probability of an individual dying is zero, but when damage is greater than the threshold, then the h_c is increasing linearly with proportionality constant b_w (the killing rate) (equation 2) (Jager, 2020). H_c can be used to calculate the survival probability (S_c) (equation 3).

$h_c = b_w \max (0, D_w - m_w)$ (2)

$S_c = exp (- \int^{t} h_c) (\tau)d\tau$ (3)

Finally, openGuts takes into consideration background mortality (h_b), death caused by random events such as handling, and not due to exposure (Jager, 2020). Thus, when the survival probability is calculated, the background survival probability is also taken into account (equation 4).

$S= S_c \times exp (-h_b t)$ (4)

The log K_{ow} value was calculated by openGuts and was compared to Pubchem (PubChem, 2020) information before generating the dose-response curve and calculating the LC50 of the compound. However, the LC50 could not be calculated for two chemicals, N-N-dimethyl-p-toluidine and MS-222. After excluding those chemicals 148 chemicals were used for further analysis.

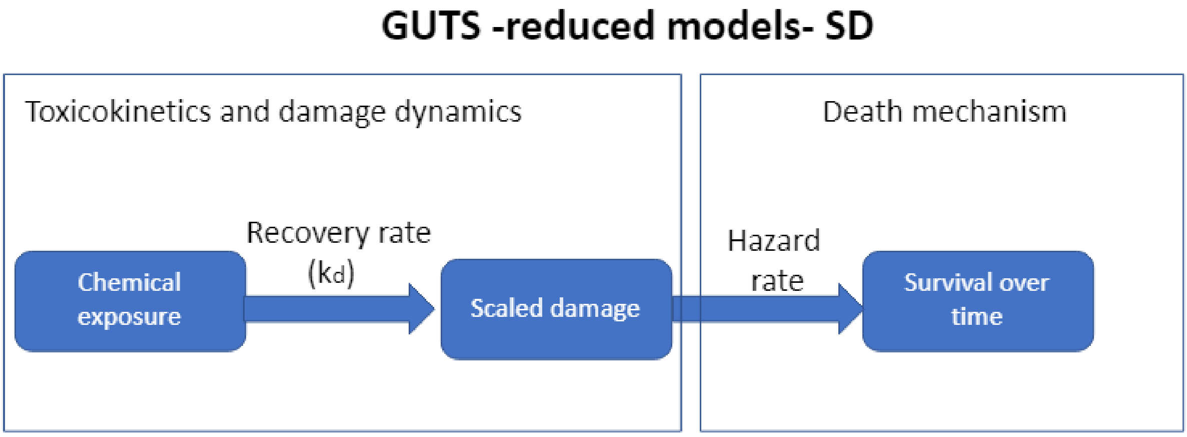


Figure 2.4: Workflow of openGUTS software. OpenGUTS uses the reduced versions of models performed by GUTS. In this study, only the stochastic death approach is shown. From chemical exposure, the recovery rate and scale damage can be calculated and using the hazard rate the survival over time can be estimated.

2.3.5.3 Heart rate data

As mentioned above, for 19 chemicals LC5/2 was the highest available concentration in mRNA sequencing data instead of LC5. In most cases, heart-rate fold change at LC5/2 was representative of the effect the chemical has on zebrafish embryos, i.e. heart rate changed significantly (or not) after exposure to both LC5 and LC5/2. However, five chemicals, Caffeic Acid, Naproxen, Michler's Ketone, Flubendazole and Benomyl, were removed as at the concentration of LC5/2 they cause no significant changes in heart rate but when the concentration is increased to LC5, zebrafish heart rate was altered significantly. The same rationale was applied to the chemical that the highest available concentration in the mRNA data was LC5/4. Butyl-benzyl-phthalate was removed as the highest available concentration was LC5/4 which was again not representative of the effect the chemical had when LC5 was used. The final dataset consists of 143 chemicals. It is important to note that among the 143 chemicals used in this study, some significantly increase heart rate and some significantly reduce heart rate, thus heart rate fold change range from 0.23 to 1.137, where the control (DMSO) had a heart rate fold change equal to 1.

2.3.6 Selection of features for classification

To ensure that as little noise as possible is added to downstream analyses Significance Analysis of Microarrays (SAM) function was applied, from the samr R package (version 3.0) (Tusher *et al.*, 2001; Monaco *et al.*, 2019). SAM applies a t-test for each independent variable to identify whether the pattern of the variable is significantly changing based on changes in experimental conditions (such as increasing chemical concentrations). A score for each independent variable is calculated based on its correlation to the dependent variable, by identifying changes relative to a standard deviation of measurements for that variable. The relative difference is compared to the distribution of relative differences through random permutations (Larsson *et al.*, 2005). When the score is higher than the adjustable threshold the variable is significant. But due to the likelihood of some variable being falsely identified as significant, the false discovery rate (FDR) is calculated by permuting over the repeated measurements (Tusher *et al.*, 2001).

Four datasets were processed by SAM. The first one consists of 2085 molecular descriptors for each chemical (143 chemicals), and the heart-rate fold changes for the highest available concentration (from mRNA-seq data) for each chemical as the dependent variable (124 chemicals- LC5, 15 chemicals- LC5/2, 4 chemicals- LC5/4) (Heart rate fold change ~ molecular descriptors +error). The second dataset consisted again of 2085 molecular descriptors for each chemical, but the experimental concentration (log format) of the highest mRNA-seq data available, was used as the dependent variable (124 chemicals- LC5, 15 chemicals- LC5/2, 4 chemicals- LC5/4) (log(experimental concentration)~ molecular descriptors +error). It is important to note that with SAM analysis (differential expression analysis) that is performed using the actual experimental concentration, the smaller the concentration the more toxic a chemical is. Thus, SAM analysis using those two datasets aims to identify molecular descriptors (structural features) whose values change when the zebrafish heart rate is altered or based on chemical toxicity.

The other two datasets consist of 726 gene count profiles with 31954 genes each, 692 genes profiled from chemical exposure, and 34 from DMSO exposure. For 124 chemicals five concentrations were available (620 profiles), for 15 chemicals four concentrations were available (60 profiles), and for four chemicals three concentrations were available (12 profiles), a total of 692 gene count profiles. The dependent variable of the third dataset was the heart-rate fold change calculated from the six replicates of each concentration and compared to the DMSO heart rate (Heart rate fold change ~ gene expression +error). On the other hand, the dependent variable of the fourth dataset consists of the values “0” for DMSO samples, “1” for all the LC5/16 concentrations, “2” for all the LC5/8 concentrations, “3” for all the LC5/4 concentrations, “4” for all the LC5/2 concentrations and “5” for all the LC5 concentrations (chemical concentration (0 to 5) ~ gene expression +error). The last dataset is used to identify genes whose expression is altered as the chemical concentration is increasing thus chemical effect (toxicity) is greater.

The SAM function was performed with 1000 permutations (nperms=1000) used for estimating the FDR, the nature of the data was specified (resp.type="Quantitative"), and the FDR cutoff for output in the significant genes table was set to 0.1 (fdr.output=0.1). For each run, variables found to be differentially expressed were identified. The genes identified by SAM to be differentially expressed when chemicals affect heart rate or chemical concentration is increased (3791 genes) were processed using the R function gost from the gprofiler2 R package (version 0.2.1) (Kolberg *et al.*, 2020). Gprofiler2 performs over-representation analysis to identify significantly enriched biological functions and pathways from multiple sources such as Gene Ontology (GO) (Ashburner *et al.*, 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2019), Reactome (REAC) (Fabregat *et al.*, 2018), and miRNA targets mostly based on Ensembl databases (Yates *et al.*, 2020). The gost function was performed using FDR as the algorithm for correcting for multiple testing (correction_method = "fdr") with a threshold set to 0.1 (user_threshold = 0.1), specifying the organisms used (organism = "drerio") and generating a list of only the statistically significant results (significant = TRUE) among all the genes of the given organism (domain_scope = c("known")). GO terms indicate the function of the genes in biological processes, molecular functions, and cellular components, thus a gene is usually represented by more than one GO term. Gene enrichment analysis was performed to reveal the set of biological processes (GO: BP) and KEGG pathways that the identified genes are found to be involved in. In addition, the miRNAs identified by gprofiler2 indicate that certain miRNAs may have been present and regulating the genes identified. As the sequencing approach does not identify miRNAs per se the terms shown can be considered analogous to GO Terms.

2.3.7 Clustering algorithm

The variables identified by differential expression analysis (SAM) were used for clustering the 143 chemicals. Two datasets were used, one with all the molecular descriptors identified to be associated with changes in chemical concentration, or the effect the chemical has on zebrafish heart rate, and one with all the genes found to be differentially expressed when chemical exposure affects zebrafish heart rate, or chemical concentration is increased. Both datasets contain quantitative data.

The k-means clustering algorithm is used in this study through the cmeans function from the e1071 R package (version 1.7.6) (Monaco *et al.*, 2019), which allows for hard (chemicals belong to only one cluster) and fuzzy clustering (chemicals belong to multiple clusters) (Bezdek *et al.*, 1984). This algorithm assigns membership to each data point (chemical) corresponding to each cluster centre based on the distance between the cluster centre and the data point. The sums of squares (euclidean) or the absolute values (manhattan) are applied for calculating the distance between observations and the cluster centre. However, as in the traditional k-means method, using the c-means function requires the number of expected clusters to be defined. The Elbow method was used in this study, to determine the number of clusters in a dataset by applying the k-means algorithm using various numbers of clusters and calculating the Sum Square Error (euclidean distance) (Thorndike, 1953; Marutho *et al.*, 2018). The explained variation is plotted as a function of the number of clusters. The first cluster explains a lot of variance, but the marginal gain drops giving an angle in the graph as the number of clusters is increased (Anand, 2017). The elbow method was performed using 1 to 7 clusters. From the plots generated (1-clustering using molecular descriptors, 2-clustering using gene count profiles), the suggested number of clusters using the data from this study is three and four.

For performing the c-means function the maximum number of iterations was set to 10000 (iter.max= 10000), the euclidean distance method (dist="euclidean") was chosen and the degree of fuzzification was set to 1.5 (m=1.5). Clustering was performed with three clusters for both input datasets (molecular descriptors, gene count profiles) for further analysis. The output file of the c-means function includes a matrix with the membership values of each chemical to the clusters, which is used for fuzzy clustering. Chemicals were assigned to one or more clusters when the membership coefficient for each cluster was greater than 0.333 (since there are three clusters in total). Fuzzy clustering can deal more effectively with outliers and data points that are found between the cluster centres and allow the identification of potential overlap between clusters.

2.3.8 Predicting mRNA (gene counts) clustering using molecular descriptors

In order to test whether the molecular descriptors can be used to predict the mRNA (gene count profiles) clustering, an R function was built. The dataset used as an input into that function, consists of 2088 molecular descriptors, as the independent variables, and the mRNA (gene counts) hard clustering of the 143 chemicals, as the dependent variable (values "1", "2", "3", representing the 3 clusters). The first step of the predictive function is splitting the dataset into test and training datasets, using the createDataPartition function part of the Caret R package (version 6.0.86) (Kuhn, 2020). This function splits the dataset randomly but aims to preserve the overall class distribution. In this study, the dataset is split 300 times into training and test sets (times = 300), where 63.2% of the chemicals are part of the training dataset (p= 0.632). One of the training datasets is selected randomly and used in downstream analysis.

To reduce the dimensionality of the dataset, by reducing the number of molecular descriptors or genes, and computational time, LASSO was applied as implemented in the stabpath function in the c060 R package (version 0.2.8) (Sill *et al.*, 2014). The stabpath function identifies features related to the dependent variable while setting the coefficients of less contributing features to zero (or close to zero) (Mozafari *et al.*, 2020). It begins by splitting the training dataset again by 63.2% (size=0.632) 1000 times, subsamples (steps= 1000), and since the dependent variable is categorical, multinomial analysis was performed (family="multinomial"). In addition, to increase the consistency of estimations, the weakness parameter is used that indicates the amount of additional randomization. In this study, for each subsample, the features are reweighted by random weight uniformly sampled (weakness= 1).

The features selected through the stabpath function were used as input to forward selection that was performed in combination with ranger (RF). The features selected in the previous step were ranked by their LASSO coefficient and then sequentially added to the RF model (one by one) and tested against unused data, the 300 data splits performed in the first step. The ranger function, part of the ranger R package (version 0.12.1) (Wright *et al.*, 2017), was applied for performing classification RF (classification= TRUE). RF calculates numerous decision trees, where each tree generates a class prediction and the class with the most votes is identified as the prediction of the model. The decision trees method is sensitive to the input data, thus RF randomly performs row and feature sampling for each tree, ensuring that each tree is different (bagging- bootstrap aggregation). Bootstrapping, a resampling technique is a statistical technique that describes the use of multiple small data samples and the generation of an average estimate from all the small data samples, in order to improve the robustness of the predicted model (Efron, 1979).

Each model was then used to predict the dependent variable, through the predict function part of the stat R package (version 3.6.1) and then the cor function from the stat R package, is applied to measure the correlation coefficient value between the dependent variable and the predicted values. The R^2 of the model is calculated by raising the correlation coefficient to the power of two. R^2 is a measurement of how much variation of the dependent variable is explained by the independent variables. Finally, for the last part of the function, the model that on average performs best across all 300 splits was identified by selecting the model with the higher mean R^2. Once the best model is identified, LOOCV was performed to evaluate the model, by training the model on almost all the data except one, and then using the model to predict the one observation left behind.

The train function, part of the caret R package, that fits predictive models over multiple tuning parameters, performed LOOCV (trControl = trainControl(method = "LOOCV")) using ranger (method= "ranger") for classification data (classification =TRUE). For classification analysis, "gini" and "extratrees" split rules are used (splitrule=c("gini"," extratrees")). The input dataset consists of the dependent variable, that is the mRNA hard clustering results, and the molecular descriptors identified as part of the model.

In order to measure the validity of the selected model, the dependent variable was randomly permuted, using the sample function in R, and used as an input for the predictive modelling function where the mean R^2 of the best model across the 300 datasets was recorded. This was repeated 1000 times and the resulting R^2 values were used to calculate the p-value of the selected model using the ecdf function part of the stat package in R (empirical cumulative distribution function) (1- ecdf(R^2 from LOOCV)). The p-value is an indication of the probability of the generated model occurring by chance, thus, a high p-value indicates that the model is not representative of the actual data, and it can be generated even with randomised data.

To further evaluate the performance of the generated model, accuracy, precision and recall values are calculated that are based on the confusion matrix. The confusion matrix is a table that consists of the comparison between the observed and predicted classification, identifying false negatives, and false positives. The correctly classified elements are located on the main diagonal (from top left to bottom right). The accuracy of the model is calculated by dividing the number of correctly classified instances per class by the total number of instances, measuring the ability of the model to correctly predict the class of the entire dataset. On the other hand, precision and recall metrics are calculated for each class independently that depend on the number of false positive and false negative results. The precision of the model is calculated, by dividing the number of correctly classified predictions for each class by the total number of positively predicted units (true and false positives), thus as the number of false positives is increase the precision is reduced. The recall is calculated by dividing the number of correctly classified instances per class by the number of instances per class (true positive and false negative), indicating the degree of reliability of the model, the higher the number of false negatives the smaller the recall value (M *et al.*, 2015; Grandini *et al.*, 2020). The workflow for the predictive modelling function and model evaluation is shown in Figure 2.5.

The package ggplot2 (version 3.3.5) was used for data visualization (Wickham, 2016). To evaluate the importance of each variable to the final model the varImpPlot function was applied, from the random forest R package (version 4.7-1.1) (Liaw et al., 2002) that generates the mean decrease in the accuracy plot and the mean decrease in Gini coefficient plot. The mean decrease in accuracy plot shows how much accuracy is lost when a variable is removed from the model, and the Gini plot measures how each variable contributes to the homogeneity of random forest nodes and leaves.

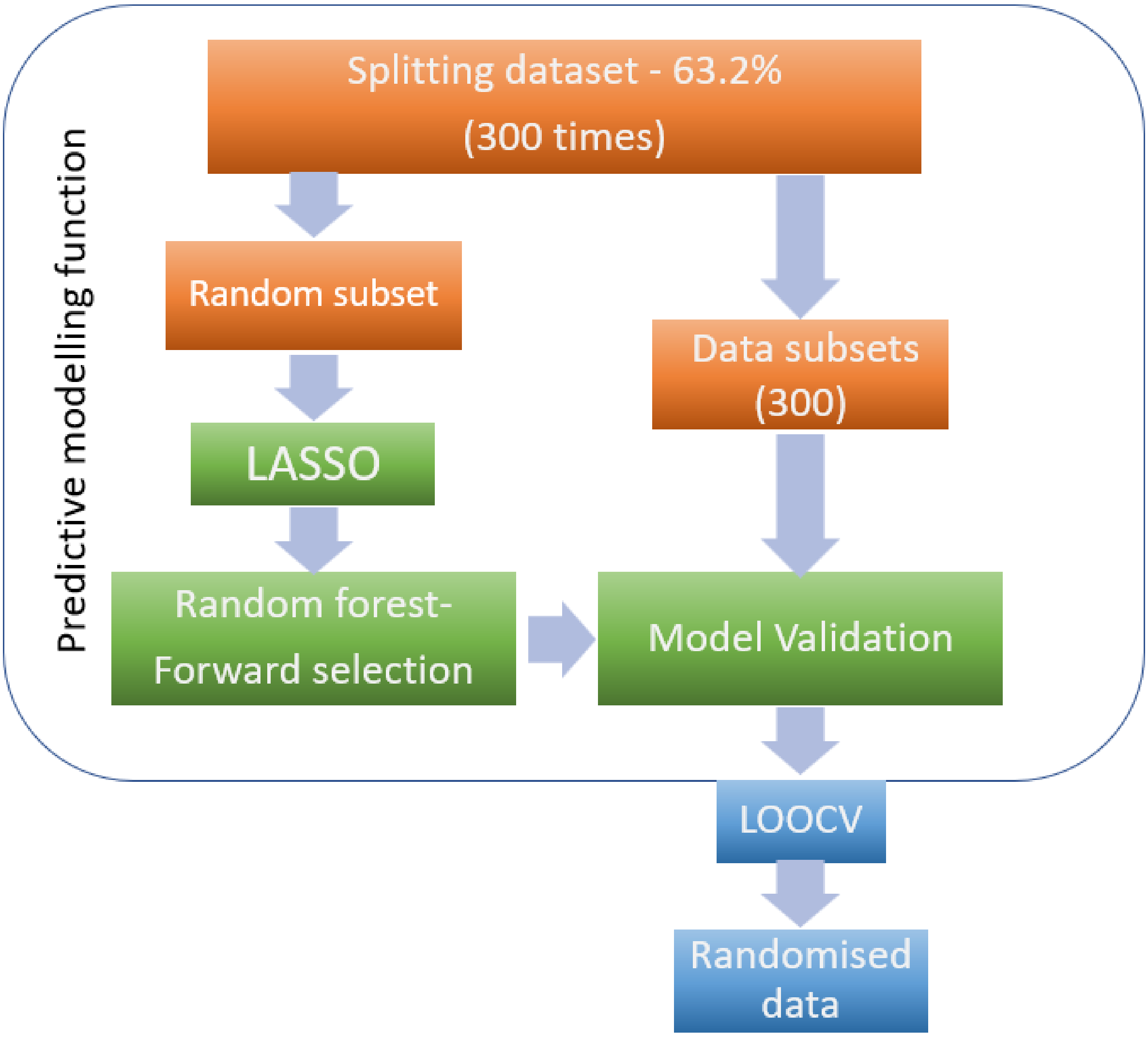


Figure 2.5:Generating and evaluating a predictive model. A predictive modelling function was developed that begins by splitting the dataset into training and test sets. LASSO stability path is then used to identify variables that can be used in predicting the dependent variable and using RF, a variable is added each time, and the resulting model was evaluated using the splits generated in the first step. The model that performs best across all splits is selected and used in LOOCV. The predictive modelling function was repeated 1000 times but the dependent variable is randomised before running the function. Using the ecdf function the probability of the model being representative of the effect is evaluated by calculating the p-value of the model.

2.4 Results

2.4.1 Majority of compounds affect zebrafish heart rate significantly

Most of the chemicals used in this study (66%- 94 out of the 143 chemicals) significantly upregulate or downregulate the heart rate of zebrafish embryos compared to the control. MoA classification is one of the most widely used methods for toxicity evaluation, thus the MoA of each chemical was identified in an effort to define the MoA classes of the chemicals that can significantly alter zebrafish heart rate. MoA classification using Toxtree revealed the high variability of the data in structural features, with chemicals that were classified in all 5 classes of Verhaar MoA (Table 2.3). Looking into the potential relationship between MoA and the ability of a chemical to significantly alter heart rate in zebrafish embryos, it can be seen that there is no particular enrichment since all five classes of MoA are represented in that list (Table 2.3) (Fisher's Exact Test p-value= 0.785). These results show no relationship between the Verhaar MoA of a chemical with its ability to alter zebrafish embryos' heart rates.

	MoA 1	**MoA 2**	**MoA 3**	**MoA 4**	**MoA 5**
Significant heart-rate changes	14	4	6	9	61
Not significant heart-rate changes	10	3	3	6	27

Table 2.3: The Verhaar MoA of the chemicals split based on their effect on heart rate, showing no correlation between toxic chemicals that significantly alter the heart rate of zebrafish, and the Verhaar MoA of the chemicals (Fisher's Exact Test p-value= 0.785).

2.4.2 Clustering chemicals using molecular descriptors reveals distinct feature sets for changes in heart rate and chemical concentration

After filtering, 2085 molecular descriptors were selected for differential analysis using SAM. SAM was performed twice, using the heart rate fold change (significantly upregulated and downregulated after chemical exposure) and experimental chemical concentrations (the lower the LC5 of a chemical the more toxic a chemical is) of the highest mRNA-seq profiles available concentration, for 124 chemicals the LC5, for 15 chemicals the LC5/2 and for four chemicals the LC5/4, as the dependent variable. This analysis aims to identify structural features (i.e. molecular descriptors) that contribute to the ability of a chemical to alter heart rate in zebrafish embryos or has the potential to kill them (toxicity).

A total of 1188 molecular descriptors were identified by differential expression analysis (SAM) to be significantly associated with heart rate changes, and 1240 with chemical concentration (toxicity). Comparing those two lists, 1031 molecular descriptors were found to be associated with both chemical toxicity and the ability of a chemical to affect heart rate. This list contained some basic features related to the composition of the compound, including molecular weight (MW, SpMax), the number of multiple atoms (nAT, nH, nSK, C%, N%, nN, nC), the number of bonds and their nature (nBT, nBO, nAB, SpPosA), distance (VE1_L, VE1_X), spatial autocorrelation (MATS) and topological descriptors (SpDiam, SM06). Certain functional groups such as phenol/carboxyl OH (O-057), primary alcohols (nOHp) and the number of aromatic carbons (nCar), also seem to be associated with the ability of a chemical to alter zebrafish heart rate and with chemical toxicity. In addition, those two SAM analyses have identified various descriptors related to the presence and structure of rings in a chemical (nCIC, TRS, Rperim, nR06, nR07, nBnz, NRS). Finally, descriptors relating the polarity of the compounds such as unipolarity (UNIP), the octanol-water partition coefficient (MLOGP2), polarity number (Pol, P_VSA, SpPosA, ATSC6p, ATS7p), the sum of atomic polarizabilities (Sp), ionisation potential (Si, ATS2i, ATSC4i, Wi_B(i)) and electronegativity (ATS3e, GATS2e), were also important in associating with the effect chemicals have on cardiac function of zebrafish embryos, and chemical toxicity.

A total of 209 molecular descriptors were identified by differential expression analysis (SAM) to be associated only with chemical concentration (chemical toxicity) including the presence of certain chemical features such as ethers (nROR), number of primary amines (nRNH2), hydrogen atoms attached to alpha-Carbon (H-051), the number of Pyrroles (nPyrroles), the number of aromatic hydroxyls (nArOH), number of nitriles (nRCN), and the number of nonaromatic conjugated Carbons (nCconj). In addition, the percentage of hydrogen atoms and the number of oxygen atoms (H%, nO), the number of double and rotatable bonds (nDB, RBN), the number of 5-membered rings (nR05) and the hydrophilic factor (Hy) were found to be associated only with chemical toxicity (Figure 2.6).

On the other hand, 157 molecular descriptors were identified by differential expression analysis (SAM) to be associated only with changes in heart rate. Some of those descriptors are the percentage of halogen atoms (X%), the number of triple bonds (nTB), aromatic ratio (ARR), the number of 10-membered rings (nR10), ring fusion density (RFD), ring bridge count (Rbrid) and the ratio of multiple path count (PCR). The chemicals that tend to dysregulate zebrafish heart rate are characterised by the presence and the number of tertiary amines (nRNR2), the number of sulphides (nRSR) and the number of secondary and tertiary alcohols (nOHs, nOHt) (Figure 2.6). The two lists were combined to represent the largest possible list of molecular descriptors associated with chemical concentration (toxicity) and the changes in heart rate fold change, with 1397 molecular descriptors in total to be used for clustering.

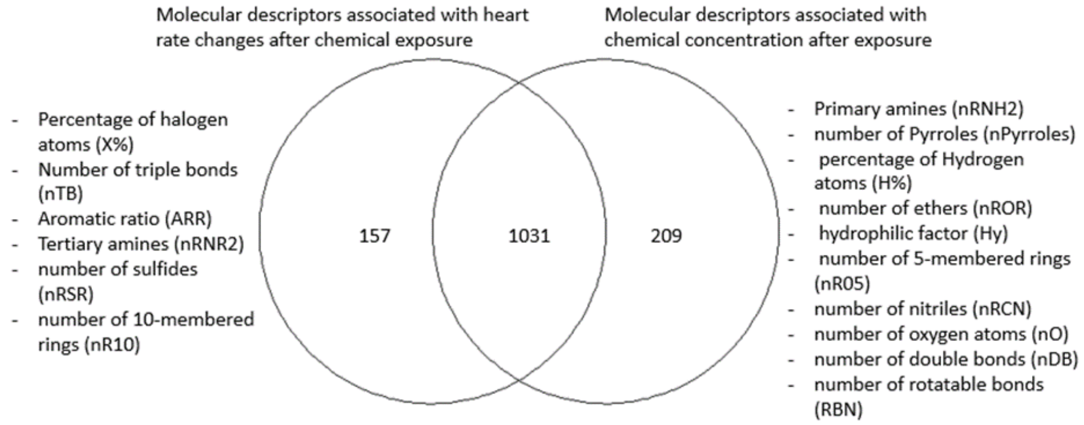

Molecular descriptors associated with heart rate changes after chemical exposure
Molecular descriptors associated with chemical concentration after exposure
- Percentage of halogen atoms (X%)
- Number of triple bonds (nTB)
- Aromatic ratio (ARR)
- Tertiary amines (nRNR2)
- number of sulfides (nRSR)
- number of 10-membered rings (nR10)
157
1031
209
- Primary amines (nRNH2)
- number of Pyrroles (nPyrroles)
- percentage of Hydrogen atoms (H%)
- number of ethers (nROR)
- hydrophilic factor (Hy)
- number of 5-membered rings (nR05)
- number of nitriles (nRCN)
- number of oxygen atoms (nO)
- number of double bonds (nDB)
- number of rotatable bonds (RBN)

Figure 2.6: Molecular descriptors identified by differential expression analysis (SAM) to be associated with heart rate and toxicity. Comparing the molecular descriptors selected by SAM analysis using changes in heart rate and chemical concentration exposure.

The c-means function was performed for chemical clustering with 1.5 degree of fuzzification, the molecular descriptors were identified by differential expression analysis (1397 molecular descriptors) and by setting the expected number of clusters to three. This analysis split the chemicals into 42, 52, and 49 chemicals (Table 2.4). Examining each cluster shows that chemicals that significantly change heart rate due to exposure are found in all clusters (Table 2.5), however, using Fisher's Exact Test, a significant association was found between the chemical effect on heart rate and the molecular descriptors clustering, indicating that chemicals that influence heart rate were more likely to be grouped together (Fisher's Exact Test p-value 0.017). Clustering using the molecular descriptors was compared to Verhaar MoA classification, showing that molecular descriptors clustering also represents MoA classification strongly with Fisher's Exact Test p-value of 0.007, as expected, since both classification methods use the structural characteristics of a chemical, for chemical grouping.

The c-means function also allows for fuzzy clustering, where chemicals can be part of more than one cluster, based on the degree of fuzzification to explore any potential overlap between the clusters and identify chemicals that are found in between clusters and share structural features with chemicals from more than one cluster. Fuzzy clustering is not based on distance like in traditional k-means function but is based on probability score or likelihood (Rehman *et al.*, 2019). The chemicals in this study were divided into three clusters, consisting of 46, 60, and 60 chemicals (Table 2.4), with low overlap (Figure 2.7) indicating that chemicals from different clusters have different molecular descriptors profiles (structural characteristics). Chemicals that significantly influence heart rate in zebrafish embryos were found in all fuzzy clusters (Table 2.5). Fisher's Exact Test was performed and revealed an association between molecular descriptors' fuzzy clustering and chemical effects on zebrafish embryos' heart rate (p-value= 0.008) and between the fuzzy clustering and the

MoA classification of chemicals (p-value = 0.009) (Table 2.6), indicating that chemicals from the same MoA class and chemicals that significantly affect zebrafish heart rate tend to cluster together.

	Molecular descriptors clustering Number of chemicals		**mRNA Sequencing clustering Number of chemicals**	
	Hard	**Fuzzy**	**Hard**	**Fuzzy**
Cluster 1	42	46	43	34
Cluster2	52	60	34	109
Cluster 3	49	60	66	/

Table 2.4: Clustering of the 143 chemicals using molecular descriptors and mRNA sequencing, hard and fuzzy clustering.

Molecular descriptor Hard Clustering			
	Cluster 1 (n=42)	**Cluster 2 (n=52)**	**Cluster 3 (n=49)**
Significant heart-rate changes	20	39	35
Not significant heart-rate changes	22	13	14
Molecular descriptor Fuzzy Clustering			
	Cluster 1 (n=46)	**Cluster 2 (n=60)**	**Cluster 3 (n=60)**
Significant heart-rate changes	24	47	46
Not significant heart-rate changes	22	13	14

Table 2.5: The distribution of chemicals that significantly affect heart rate among the clusters generated using molecular descriptors for hard clustering (Fisher's Exact Test p-value=0.017) and fuzzy clustering (Fisher's Exact Test p-value=0.008).

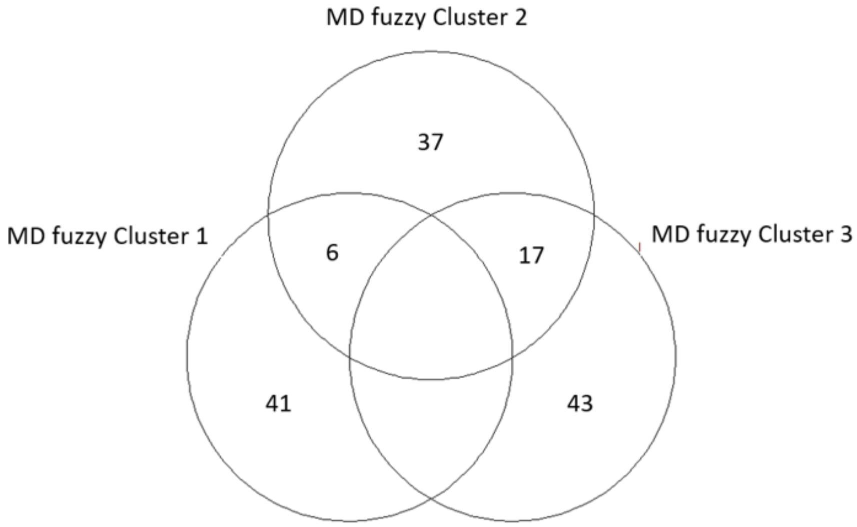


Figure 2.7: Comparing the Fuzzy clustering results generated using molecular descriptors. Chemicals were grouped into three clusters with low overlap.

Hard Clustering using Molecular descriptors					
	MoA 1	**MoA 2**	**MoA 3**	**MoA 4**	**MoA 5**
Cluster 1 (n=42)	9	7	2	2	22
Cluster 2 (n=52)	5	0	3	7	37
Cluster 3 (n=49)	10	0	4	6	29
Fuzzy Clustering using Molecular descriptors					
	MoA 1	**MoA 2**	**MoA 3**	**MoA 4**	**MoA 5**
Cluster 1 (n=46)	10	7	2	2	25
Cluster 2 (n=60)	6	1	3	9	41
Cluster 3 (n=60)	12	0	5	8	35

Table 2.6: The Verhaar MoA distribution of the chemicals in each molecular descriptor cluster generated from hard clustering (Fisher's Exact Test p-value=0.007) and fuzzy clustering (Fisher's Exact Test p-value=0.009), both showing a strong correlation between mRNA clustering and Verhaar MoA classification.

2.4.3 Clustering chemicals using mRNA sequencing data highlights the strong association with heart rate changes

The mRNA sequencing data generated after exposure of zebrafish embryos to 143 chemicals at multiple concentrations, were used to generate gene count profiles that consist of 31,953 genes each. To identify gene expression changes associated with increasing chemical concentration (chemical toxicity) or heart rate defects, a quantitative association analysis was performed using differential expression analysis (SAM).

Differential expression analysis identified a set of 3780 genes associated with increases in chemical concentration, 930 of those genes were upregulated and 2850 were downregulated as chemical concentration was increased. Functional enrichment analysis was performed to identify gene ontology terms that are significantly associated with the selected genes. The targets of six microRNAs were identified to be upregulated, thus the expression of the equivalent miRNA is reduced (Table 2.7). The miR-126a-3p (p-value=0.045) is involved in various toxicant exposures and its downregulation contributes to cardiac dysfunction (Shen *et al.*, 2019; Balasubramanian *et al.*, 2020). The miR-155 (p-value=0.045), is a key regulator for cell homeostasis and regulates hematopoietic lineage differentiation, immunity and inflammation (Cao *et al.*, 2016; Liu *et al.*, 2021). The miR-216b (p-value=0.045) and miR-499-5p (p-value=0.045) are involved in chemical stress, myocardial autophagy and apoptosis (Ahkin Chin Tai *et al.*, 2020; Wang *et al.*, 2021). The miR-30a-5p (p-value=0.06) is involved in muscle cell types specification and differentiation and downregulation contributes to endoplasmic reticulum stress in cardiac muscle and vascular smooth muscle cells (Ketley *et al.*, 2013; Chen *et al.*, 2014). Finally, the miR-145-5p (p-value=0.06) regulates smooth muscle cell differentiation and cardiac specification during heart development, found to be downregulated in coronary artery disease (Zhao *et al.*, 2015; Vacante *et al.*, 2019). The genes found to be upregulated when the chemical concentration is increased were associated with 103 significant biological properties, amino sugar metabolism, cytoskeleton-dependent intracellular transport, regulation of microtubule polymerization or depolymerization, cell death, nervous system and regulation of DNA-templated transcription (p-value range 0.02-0.093) (Table 2.7, Supplementary materials Table S.1). Finally, one KEGG GO term was selected, thiamine metabolism (p-value=0.015), which is involved in vitamin b1 biosynthesis that plays an important role in changing carbohydrates into energy.

Using the downregulated genes identified by differential expression analysis (SAM) one miRNA was identified by GO analysis, miR-1 (p-value=0.093), indicating that since the targets of this miRNA are downregulated the miRNA is overexpressed. miR-1 is a muscle-specific cardiac miRNA and upregulation of miR-1 has been associated with various cardiac conditions (Ai *et al.*, 2012; Ahkin Chin Tai and Freeman, 2020). Finally, the downregulated genes were involved in two KEGG pathways, MAPK signalling pathway (p-value= 0.008) involved in cell proliferation, differentiation and migration, and cellular senescence pathway (p-value=0.066) responsible for irreversible cellular arrest (KEGG PATHWAY Database).

Chemical concentration- Genes upregulated		
GO term name	**P value**	**GO term ID**
miRNA-126a-3p	0.045	dre-miR-126a-3p
miRNA-155	0.045	dre-miR-155
miRNA-216b	0.045	dre-miR-216b
miRNA-499-5p	0.045	dre-miR-499-5p
miRNA-30a-5p	0.06	dre-miR-30a-5p
miRNA-145-5p	0.06	dre-miR-145-5p
Thiamine metabolism	0.015	KEGG:00730
Morphogenesis	0.044-0.093	Biological Processes
Cell membrane organisation	0.02-0.093	Biological Processes
Nervous system	0.05-0.09	Biological Processes
Microtubule	0.045-0.086	Biological Processes
Cell death	0.07-0.093	Biological Processes
Chemical concentration- Genes downregulated		
MAPK signalling pathway	0.008	KEGG:04010
Cellular senescence	0.066	KEGG:04218
miRNA-1	0.093	dre-miR-1
Heart-rate fold changes		
Name	**P value**	**GO term ID**
miRNA- 206-3p	0.07	dre-miR-206-3p
miRNA-126a-3p	0.01	dre-miR-126a-3p
miRNA-155	0.01	dre-miR-155
miRNA-216b	0.01	dre-miR-216b
miRNA-499-5p	0.01	dre-miR-499-5p
miRNA-430a-3p	0.01	dre-miR-430a-3p

Table 2.7: Functional enrichment analysis of the genes selected by differential expression analysis (SAM), to be differentially expressed based on chemical toxicity (chemical concentration) and heart-rate changes.

When establishing the association to heart rate changes a set of 1890 genes were identified to be differentially expressed. Functional enrichment analysis was performed and identified six miRNAs, miR-206-3p (p-value=0.07), known to be expressed in skeletal muscles and regulates muscle proliferation and differentiation (Lin *et al.*, 2017),miR-126a-3p (p-value=0.01), involved in circulatory system development and is expressed in the cardiovascular system (Khanaghaei *et al.*, 2016) and miR-155 (p-value=0.01), associated with cardiovascular diseases (Faraoni *et al.*, 2009). The miR-216b (p-value=0.01), has been associated with cardiomyocyte proliferation (Faraoni *et al.*, 2009; Ahkin Chin Tai *et al.*, 2020), miR-499-5p (p-value=0.01), is involved in cardiac differentiation (Shieh *et al.*, 2011; Garreta *et al.*, 2017) and finally, miR-430a-3p (p-value=0.01) that is important in embryonic heart morphogenesis (Li *et al.*, 2017) (Table 2.7).

Comparing the two gene lists obtained (heart rate fold change and increasing concentration), 11 genes were only associated with changes in the heart rate of zebrafish embryos, related to protein targeting and localization to the mitochondrion, apoptotic signalling, MAPK, TCR and AKT signalling and innate immune system (p-value range= 0.01 - 0.098). On the other hand, 1901 genes were associated only with chemical exposure concentration related to organ development, cell-cell signalling, wnt, Foxo and VEGF signalling, cell cycle and fatty acid elongation (p-value range= 2.8e-11 - 0.08).

Combining the two lists generated by differential expression analysis (SAM), to get the list of unique genes that are associated with both heart-rate effects on zebrafish and chemical exposure, 3791 genes were identified and used for clustering. Using the Elbow method, the dataset was split into three clusters consisting of 43, 34 and 66 chemicals (Table 2.4). Checking the distribution of chemicals that do not significantly influence heart rate in the clusters after mRNA clustering, it can be seen that most of them are grouped in cluster 3 (Table 2.8). Fisher's Exact Test verified the association between heart rate changes and mRNA clustering with a p-value of 6.084e-08. On the other hand, the mRNA clustering does not seem to be associated with Verhaar MoA classification, Fisher's Exact Test p-value = 0.706 (Table 2.9).

Fuzzy clustering resulted in the generation of two clusters, the first one consisting of the same 34 chemicals as the mRNA cluster 2 hard clustering and the second cluster, consisting of 109 chemicals, the combination of hard mRNA cluster 1 (43 chemicals) and cluster 3 (66 chemicals) (Table 2.4), indicating that the chemicals from hard mRNA clustering cluster 1 and 3 share similar gene count profiles. Most of the chemicals from fuzzy cluster 1, 32 out of 34 chemicals, significantly affect the heart rate in zebrafish embryos (Table 2.8). As expected mRNA fuzzy clustering is highly associated with the effect of the chemical on the zebrafish embryo heart rate (Table 2.8) (Fisher's Exact Test p-value = 2.363e-05) and not associated with the Verhaar MoA classification (Table 2.9) (Fisher's Exact Test p-value = 0.453).

Hard Clustering with mRNA sequencing data			
	Cluster 1 (n=43)	**Cluster 2 (n=34)**	**Cluster 3 (n=66)**
Significant heart-rate changes	34	32	28
Not significant heart-rate changes	9	2	38
Fuzzy Clustering with mRNA sequencing data			
	Cluster 1 (n=34)	**Cluster 2 (n=109)**	
Significant heart-rate changes	32	62	
Not significant heart-rate changes	2	47	

Table 2.8: The distribution of chemicals that significantly affect heart rate among the clusters generated using mRNA sequencing data. Both clustering methods, hard and fuzzy, are shown to cluster chemicals based on heart rate changes, hard clustering Fisher's Exact Test p-value=6.084e-08 and fuzzy clustering p-value=2.363e-05.

Hard Clustering with mRNA sequencing data					
	MoA 1	**MoA 2**	**MoA 3**	**MoA 4**	**MoA 5**
Cluster 1 (n=43)	7	2	1	7	26
Cluster 2 (n=34)	5	2	3	1	23
Cluster 3 (n=66)	12	3	5	7	39
Fuzzy Clustering with mRNA sequencing data					
	MoA 1	**MoA 2**	**MoA 3**	**MoA 4**	**MoA 5**
Cluster 1 (n=34)	5	2	3	1	23
Cluster 2 (n=109)	19	5	6	14	65

Table 2.9: The Verhaar MoA of the chemicals from each cluster generated from mRNA clustering. Hard clustering (Fisher's Exact Test p-value= 0.706) and fuzzy clustering (Fisher's Exact Test p-value=0.453) show no correlation between the clusters and the chemical MoA.

2.4.4 Combining molecular descriptors and gene expression significantly improves association with heart rate and chemical concentration

Cluster membership of chemicals clustered using molecular descriptors and mRNA data differ significantly (Table 2.10). Based on Fisher's Exact Test comparing the two clustering methods (molecular descriptors versus mRNA-seq data clustering) it is clear that the distribution of values is fairly random and that no single enrichment of chemicals has been observed. For the hard and fuzzy clustering, this is highlighted by Fisher's Exact Test p-value of 0.106 and 0.183 respectively. In addition, clustering chemicals using molecular descriptors represent both heart-rate changes and the Verhaar MoA of the chemicals, but clustering chemicals using only mRNA information, although it does represent heart-rate changes very well, there is no association between the mRNA clustering of the chemicals and their MoA.

Hard Clustering				
		Molecular descriptors		
		Cluster 1 (n=42)	**Cluster 2 (n=52)**	**Cluster 3 (n=49)**
mRNA clustering	**Cluster 1 (n=43)**	7	2	1
	Cluster 2 (n=34)	5	2	3
	Cluster 3 (n=66)	12	3	5
Fuzzy Clustering				
		Molecular descriptors		
		Cluster 1 (n=46)	**Cluster 2 (n=60)**	**Cluster 3 (n=60)**
mRNA clustering	**Cluster 1 (n=109)**	5	2	3
	Cluster 2 (n=34)	19	5	6

Table 2.10: Comparing cluster composition between molecular descriptors and mRNA sequencing data clustering, highlighting the difference between the two methods, hard clustering (Fisher's Exact Test p-value=0.106) and fuzzy clustering (Fisher's Exact Test p-value=0.183).

Thus, in order to explore those differences, the molecular descriptors and genes identified to be altered under the two conditions explored in this study, chemical concentration (toxicity) and heart rate fold change caused by chemical exposure, were combined. The new dataset consists of 5188 variables (3791 genes and 1397 molecular descriptors). The same method was followed here, where the elbow method was used along with the c-means function to cluster the 143 chemicals into three classes. Clustering chemicals using the new, combined dataset, resulted in three clusters with 53, 39 and 51 chemicals each (Table 2.11). Fuzzy clustering on the other hand also resulted in three clusters, clusters 1 and 2 containing 90 chemicals each, 88 of them shared between the clusters, and cluster 3, consisting of 52 chemicals, from which only one (o-Phenylenediamine) is shared with cluster 2 (Table 2.11).

	Data combined clustering- number of chemicals per cluster	
	Hard	**Fuzzy**
Cluster 1	53	90
Cluster 2	39	90
Cluster 3	51	52

Table 2.11: Clustering chemicals using a combination of molecular descriptors and mRNA information used to explain chemical toxicity and heart-rate effects on zebrafish organisms.

		Molecular descriptors clustering		
		Cluster 1 (n=42)	**Cluster 2 (n=52)**	**Cluster 3 (n=49)**
Data combined clustering	**Cluster 1 (n=53)**	0	18	35
	Cluster 2 (n=39)	5	20	14
	Cluster 3 (n=51)	37	14	0
		mRNA clustering		
		Cluster 1 (n=43)	**Cluster 2 (n=34)**	**Cluster 3 (n=66)**
Data combined clustering	**Cluster 1 (n=53)**	24	0	29
	Cluster 2 (n=39)	5	34	0
	Cluster 3 (n=51)	14	0	37

Table 2.12: Comparing the combined clustering with the molecular descriptors clustering (Fisher's Exact Test p-value <2.2e-16) and mRNA clustering (Fisher's Exact Test p-value=2.2e-16) both showing a high correlation between the clustering methods.

Comparing the combined datasets' clustering results, with the clustering methods using only molecular descriptors or mRNA-seq data, a high correlation can be seen, with both of them, as expected (Fisher's Exact Test p-value <2.2e-16) (Table 2.12). Looking further into the combined data clustering, a high correlation was observed between the clusters and the effect chemicals have on zebrafish heart rate (Fisher's Exact Test p-value= 3.45e-06) (Table 2.13) but no correlation with the Verhaar MoA classification of chemicals (Fisher's Exact Test p-value= 0.222) (Table 2.14). The inability to find any correlation between MoA classification and any clustering performed in this study using the mRNA-seq data, using only mRNA seq-data or combined with molecular descriptors, highlighting the difference between structural profiles and gene expression clustering and the variation in the information provided by those two types of data (molecular descriptors, gene count profiles).

Data combined clustering			
	Cluster 1 (n=53)	**Cluster 2 (n=39)**	**Cluster 3 (n=51)**
Significant heart-rate changes	32	37	25
Not significant heart-rate changes	21	2	26

Table 2.13: The distribution of chemicals that significantly affect heart rate among the clusters generated using the combined dataset. This clustering method is highly representative of the effect toxic chemicals have on zebrafish embryos (Fisher's Exact Test p-value= 3.45e-06).

Data combined clustering					
	MoA 1	**MoA 2**	**MoA 3**	**MoA 4**	**MoA 5**
Cluster 1 (n=53)	10	0	3	9	31
Cluster 2 (n=39)	5	2	4	3	25
Cluster 3 (n=51)	9	5	2	3	32

Table 2.14: The Verhaar MoA of the chemicals from each cluster generated using the combined dataset, showing no correlation between them (Fisher's Exact Test p-value=0.222).

2.4.5 Prior clustering of chemicals based on mRNA improves prediction using structural features

As the mRNA data were able to distinguish and associate much more closely with the effect on heart rate compared to the structural features themselves, the ability of structural features to predict the classification generated by the mRNA data was evaluated. The final model, based on the LASSO stability selection approach, achieved a mean R^2 of 0.11 and used 66 molecular descriptors (Model 1 - Figure 2.8). Along with the low R^2, the accuracy of the model was 0.53 (table 2.15), with the precision of prediction to be equal to 0.53 for class 1, 0.31 for class 2 and 0.57 for class 3, and the recall 0.4 for class 1, 0.15 for class 2 and 0.82 for class 3 (Figure 2.8). The low predictive ability of this model seems to suggest that predicting the mRNA clustering using molecular descriptors is extremely challenging and that more data is required to define more homogenous chemical groupings which will likely result in an improved effect prediction.

To identify the cause of such a weak model, and generate a more reliable one, the chemicals with the highest overlap between the mRNA and structural clusters were selected. Comparing molecular descriptors clustering and the combined clustering, 92 chemicals were selected (Figure 2.9). The modelling function described in the methods was used with 2012 molecular descriptors (after removing molecular descriptors with NA values of those with no variations across the samples) to predict mRNA classification. The resulting model (Model 2) had an R^2 of 0.22 and consisted of 98 descriptors (Table 2.15). Along with the low R^2, the accuracy of the model was 0.70 (table 2.15), with the precision of prediction to be equal to 0.6 for class 1, 0.73 for class 2 and 0.75 for class 3, and the recall 0.5 for class 1, 0.53 for class 2 and 0.89 for class 3 (Figure 2.9).

On the other hand, when only the chemicals with the highest overlap between mRNA clustering and combined dataset clustering were used (molecular descriptors, mRNA), 95 chemicals were selected (Figure 2.10). The model generated (Model 3) resulted in 72 molecular descriptors out of the 2022 descriptors (after removing molecular descriptors with NA values of those with no variations across the samples from the 5270 molecular descriptors) and a mean R^2 across the 300 splits of 0.6 (p-value=0) (Table 2.15). The low p-value is an indication that this model has a very low probability to be generated when the dataset is randomised, thus the model is representative of the data provided and the phenomenon under study. Along with the low R^2, the accuracy of the model was 0.72 (Table 2.15), with the precision of prediction to be equal to 0.75 for class 1, 0.61 for class 2 and 0.82 for class 3, and the recall 0.625 for class 1, 0.65 for class 2 and 0.865 for class 3 (Figure 2.10).

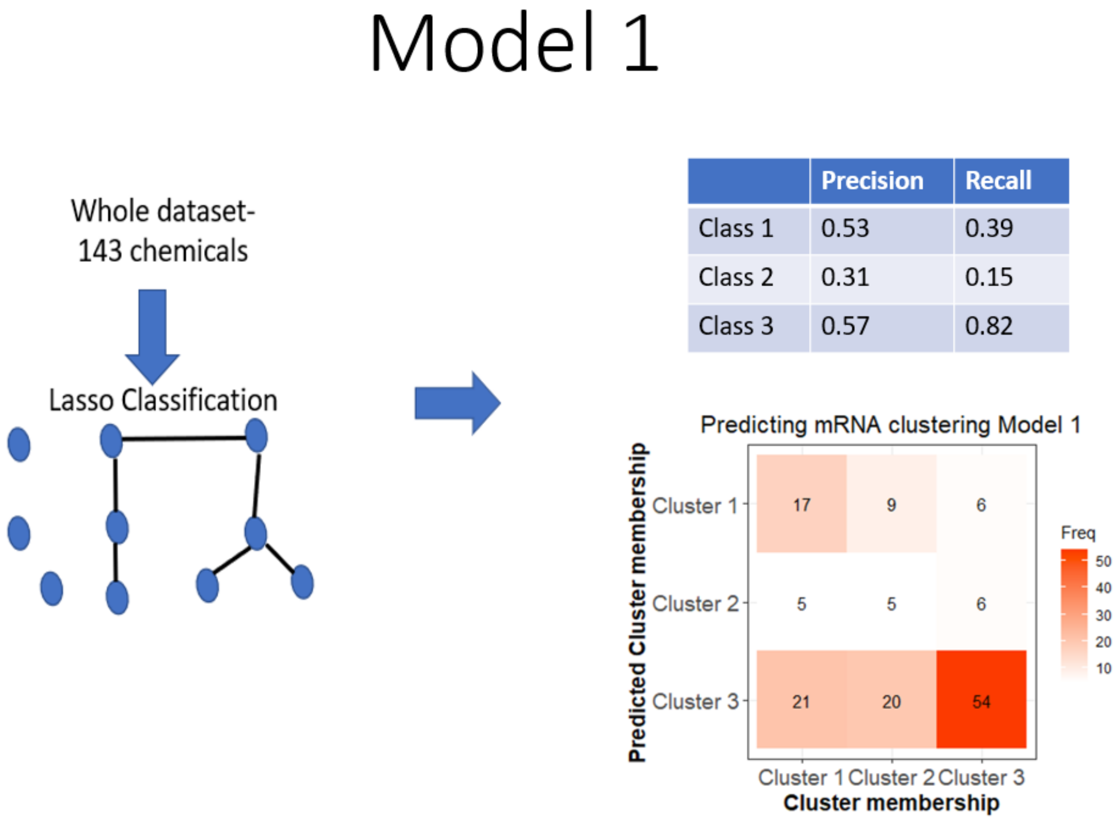
The selected molecular descriptors were evaluated based on their predictive power using the mean decrease accuracy plot, which expresses how much accuracy is lost by excluding each variable, and the mean decrease in Gini coefficient plot, which measures how each variable contributes to the homogeneity of the nodes and leaves. The higher those values are, the higher the importance of the variable to the model. The most important variables based on the mean decrease accuracy plot and mean accuracy Gini coefficient were identified (figure 2.10). Six 2D autocorrelation descriptors were selected, Geary autocorrelation of lag 8 (the topological distance between pairs of atoms) weighted by van der Waals volume (GATS8v), by ionization potential (GATS8i) or by mass (GATS8m) (Geary, 1954), Centred Broto-Moreau autocorrelation of lag 8 weighted by polarizability (ATSC8p) and by ionization potential (ATSC8i) (Moreau *et al.*, 1980) and the mean topological charge index of order 6 (JGI6) (Galvez *et al.*, 1994). These spatial autocorrelation descriptors provide information on how the considered property (weight) is distributed along the topological structure (topological distance 8). CATS2D, atom-type autocorrelation descriptors, are based on the potential pharmacophore points (PPP), in this case, hydrogen-bond donor lipophilic properties (CATS2D_04_DL) (Schneider *et al.*, 1999).

One 2D atom pair descriptor was identified, that describes the frequency of carbon and oxygen atoms at topological distance 3 (F03[C-O]), where topological distance is defined as the number of bonds of the shortest path between two atoms. Extended topochemical atom (ETA) eta x shape index (Eta_sh_x), calculated from molecular composition information and two-dimensional representations, provides information about the branched distribution of atoms (Pal *et al.*, 1989). A 2D matrix-based descriptor was identified, the average Wiener-like index from the Barysz matrix weighted by polarizability (WiA_Dz(p)). The barysz distance matrix (Dz) is a weighted distance matrix accounting for the presence of heteroatoms and multiple bonds (Todeschini *et al.*, 2008). Finally, the smallest eigenvalue n. 5 of the Burden matrix weighted by polarizability (SpMin5_Bh(p)), a Burden eigenvalues descriptor was selected, which is calculated from the Burden matrix, where the diagonal elements are atomic properties (polarizability), and the off-diagonal elements represent the pairs of bonded atoms (Burden, 1989)

Following those results, selecting only the chemicals that highly overlap between the mRNA and molecular descriptors clustering, resulted in 58 chemicals (Figure 2.11). Predictive modelling identified 10 out of 2031 molecular descriptors (after removing molecular descriptors with NA values of those with no variations across the samples from the 5270 molecular descriptors) that were selected to predict the mRNA clustering results (Model 4), with an R² of 0.93 (p-value=0, low probability of the model to be generated with random data) (Figure 2.11). Along with the low R², the accuracy of the model was 0.95 (table 2.15), with the precision of prediction being equal to 0.94 for class 1, 0.87 for class 2 and 1 for class 3, and the recall 0.91 for class 1, 0.93 for class 2 and 1 for class 3 (Figure 2.11). The importance of each descriptor is evaluated using the decrease accuracy plot and the mean decrease in the Gini coefficient plot. Two of those descriptors were similar to the ones selected for model 3 but weighted by different properties, Geary autocorrelation of lag 4 weighted by Sanderson electronegativity (GATS4e) and the smallest eigenvalue n. 8 of Burden matrix weighted by ionization potential (SpMin8_Bh(i)). One descriptor related to the drug-like ability of the chemicals was selected, Ghose-Viswanadhan-Wendoloski antidepressant-like index at 80% (Depressant-80) (Ghose, Viswanadhan and Wendoloski, 1999). Three descriptors were selected related to the edge adjacency indices, that provide information about bonds, eigenvalue n. 3 from edge adjacency matrix weighted by dipole moment (Eig03_EA(dm)), eigenvalue n. 13 from augmented edge adjacency mat. weighted by bond order (Eig13_AEA(bo)) and eigenvalue n. 14 from augmented edge adjacency mat. weighted by bond order (Eig14_AEA(bo)) (Laskar, 1969). Finally, the number of the 5-membered rings (nR05), the presence or absence of single carbon-carbon bonds at topological distance 6 (B06[C-C]), the presence or absence of oxygen sulfur bonds at topological distance 2 (B02[O-S]) and the presence or absence of nitrogen and chlorine bonds at topological distance 3 (B03[N-Cl]) were found to be essential for predicting the chemical clustering.

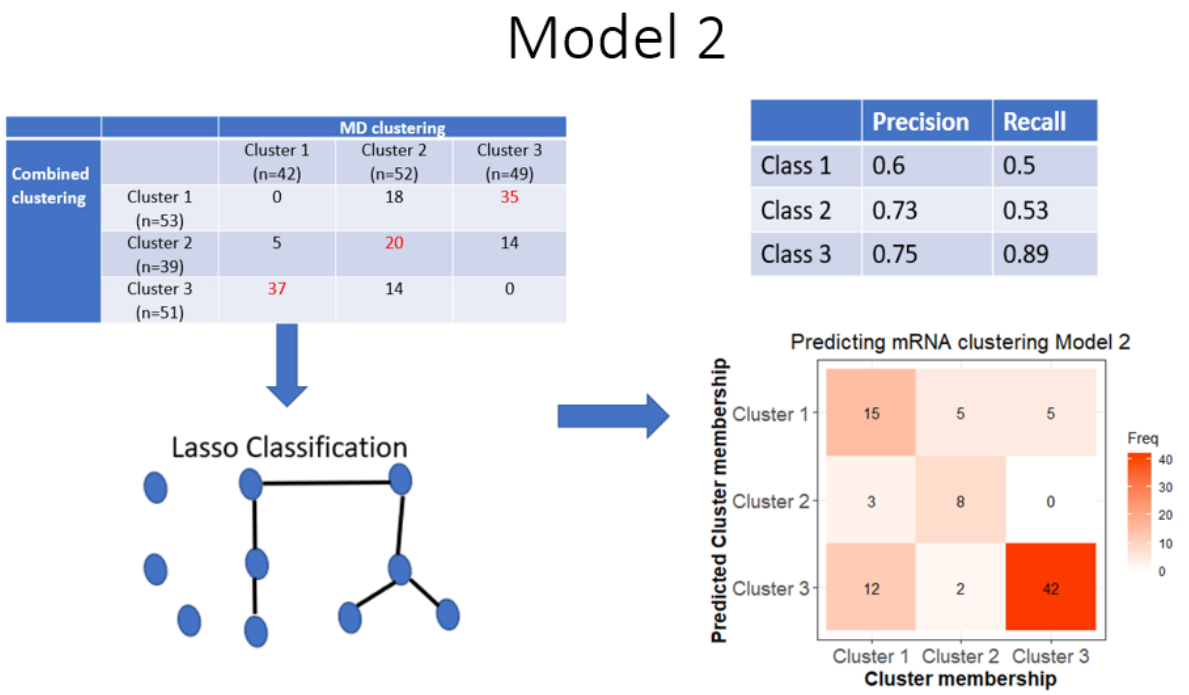
	R^2	Accuracy
Model 1 (n=143)	0.11	0.53
Model 2 (n=92)	0.22	0.71
Model 3 (n=95)	0.6	0.72
Model 4 (n=58)	0.93	0.95

Table 2.15: The calculated R^2 and accuracy of the four models generated to predict the mRNA-seq hard clustering using molecular descriptors. Each model consists of a different number of chemicals.



	Precision	Recall
Class 1	0.53	0.39
Class 2	0.31	0.15
Class 3	0.57	0.82

Figure 2.8: Graphical representation of Model 1 generation and evaluation. 143 chemicals represented by molecular descriptors were used as input for the predictive modelling function, to predict mRNA hard clustering. The predicted and c-means function clustering membership were plotted and the precision and recall ability of the predictive model were provided.

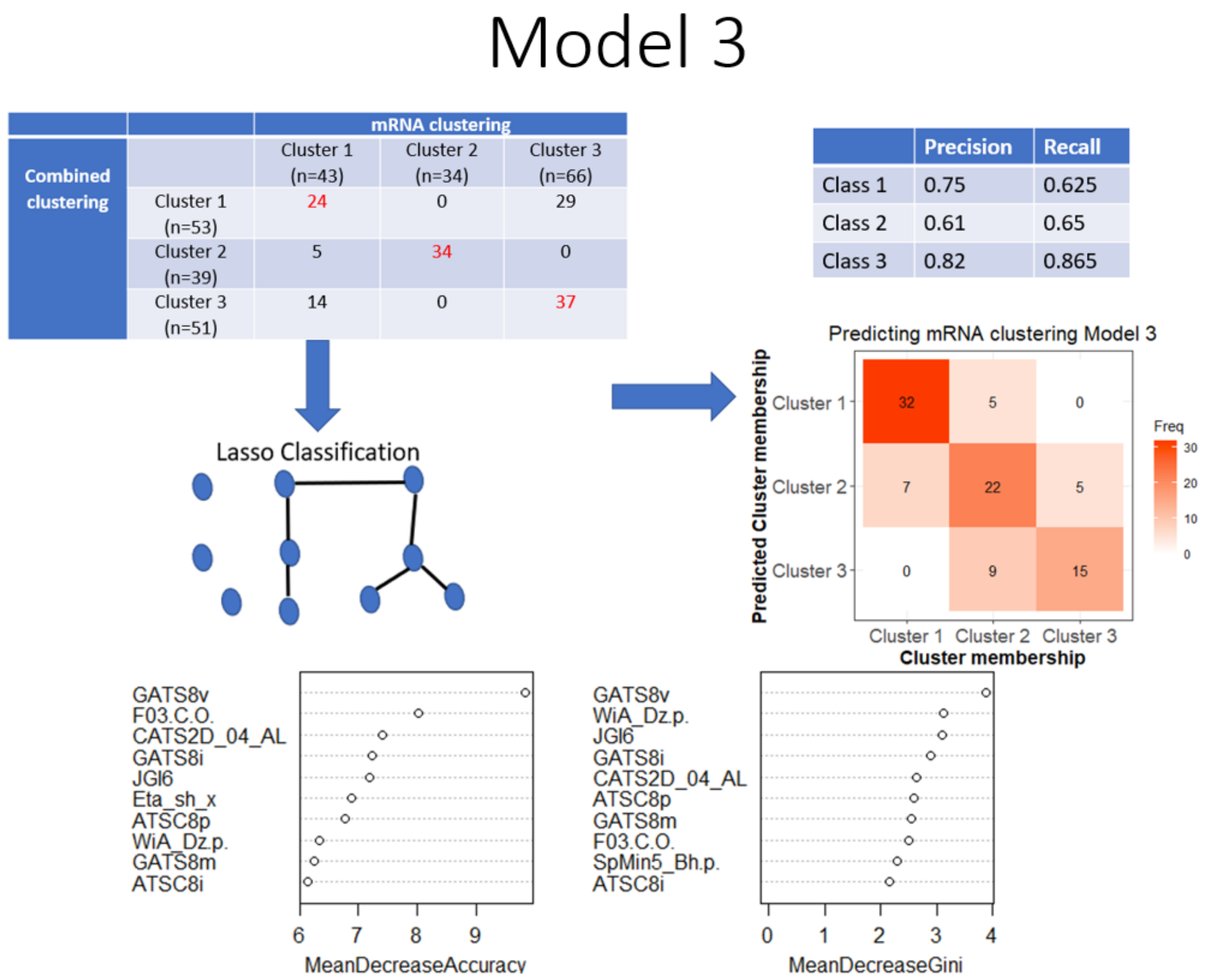


Model 2

		MD clustering		
		Cluster 1 (n=42)	Cluster 2 (n=52)	Cluster 3 (n=49)
Combined clustering	Cluster 1 (n=53)	0	18	35
	Cluster 2 (n=39)	5	20	14
	Cluster 3 (n=51)	37	14	0

	Precision	Recall
Class 1	0.6	0.5
Class 2	0.73	0.53
Class 3	0.75	0.89

Figure 2.9: Graphical representation of Model 2 generation and evaluation. 92 chemicals represented by molecular descriptors were used as input for the predictive modelling function, to predict mRNA hard clustering. The predicted and c-means function clustering membership were plotted and the precision and recall ability of the predictive model were provided.

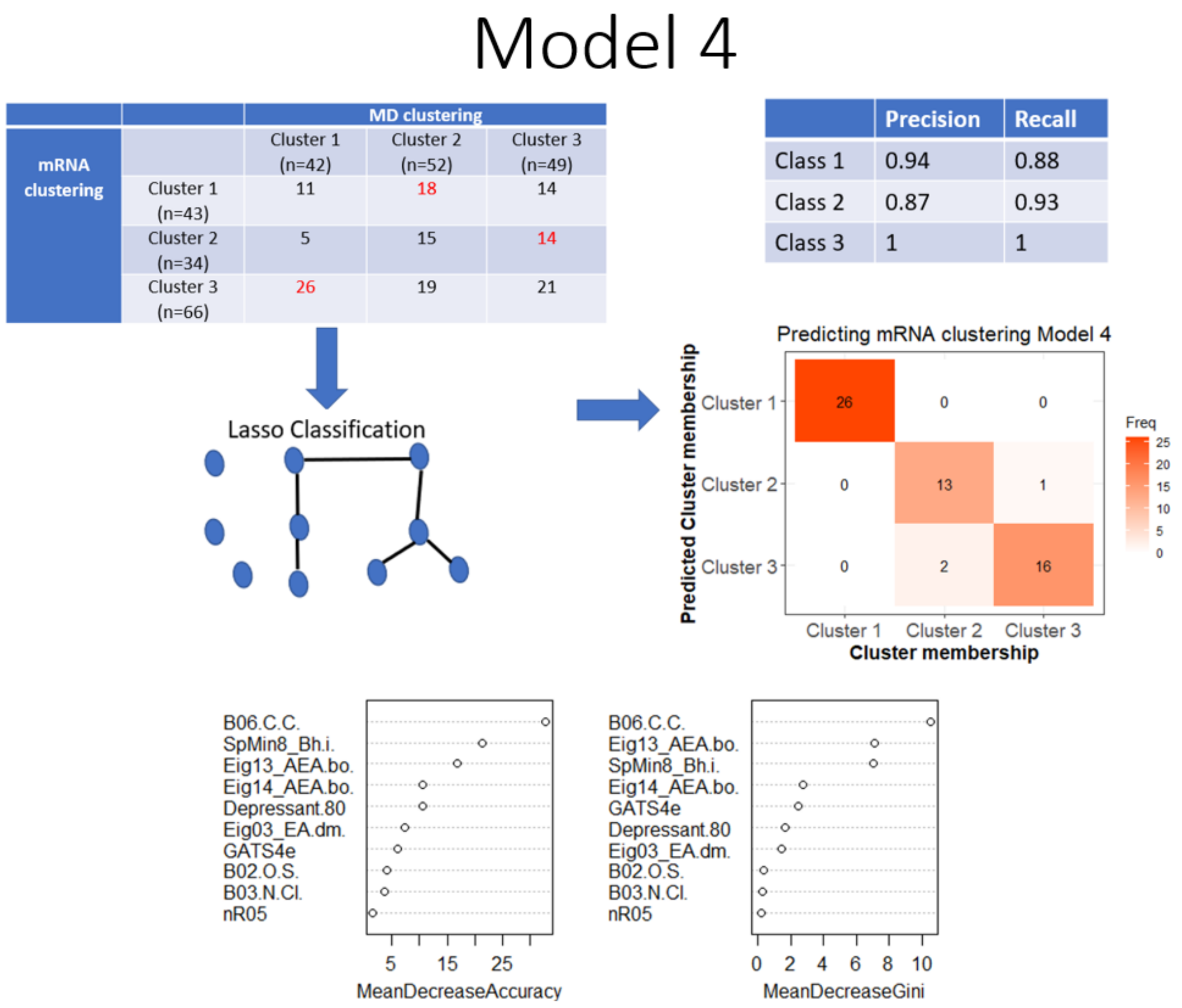


Model 3

		mRNA clustering		
		Cluster 1 (n=43)	Cluster 2 (n=34)	Cluster 3 (n=66)
Combined clustering	Cluster 1 (n=53)	24	0	29
	Cluster 2 (n=39)	5	34	0
	Cluster 3 (n=51)	14	0	37

	Precision	Recall
Class 1	0.75	0.625
Class 2	0.61	0.65
Class 3	0.82	0.865

Figure 2.10: Graphical representation of Model 3 generation and evaluation. 95 chemicals represented by molecular descriptors were used as input for the predictive modelling function, to predict mRNA hard clustering. The predicted and c-means function clustering membership were plotted and the precision and recall ability of the predictive model were provided. The mean decrease accuracy and mean decrease Gini plots show the molecular descriptors that contribute the most to the predictive power of the model.



Model 4

		MD clustering		
		Cluster 1 (n=42)	Cluster 2 (n=52)	Cluster 3 (n=49)
mRNA clustering	Cluster 1 (n=43)	11	18	14
	Cluster 2 (n=34)	5	15	14
	Cluster 3 (n=66)	26	19	21

	Precision	Recall
Class 1	0.94	0.88
Class 2	0.87	0.93
Class 3	1	1

Figure 2.11: Graphical representation of Model 4 generation and evaluation. 58 chemicals represented by molecular descriptors were used as input for the predictive modelling function, to predict mRNA hard clustering. The predicted and c-means function clustering membership were plotted and the precision and recall ability of the predictive model were provided. The mean decrease accuracy and mean decrease Gini plots show the molecular descriptors that contribute the most to the predictive power of the model.

2.5 Discussion

2.5.1 Structural characteristics associated with chemical toxicity

More than 1000 descriptors were identified to be associated with chemical toxicity or the ability of the chemical to affect zebrafish heart rate, where the majority of them are associated with both toxicity and cardiotoxicity. The popularity of QSAR methods helped identify various structural features associated with chemical toxicity, however, the available information on the association between molecular descriptors and cardiotoxicity is limited. The structure of the chemical has been associated with chemical-induced toxicity, where different structural features are associated with different MoA, based on the assumption that the structural characteristics of a chemical can be an indication of its biological properties.

QSAR zebrafish toxicity studies have identified a set of molecular descriptors that are associated with chemical toxicity, including lipophilicity, polarizability, electronegativity, ionization potential, branching, presence of rings, number and nature of bonds, chemical atom composition, and molecular weight (Estrada, 1996; Ghorbanzadeh *et al.*, 2016; Lavado *et al.*, 2020). Lipophilicity values, such as AlogP and MLOGP, have been identified by QSARs models to be associated with toxicity, since lipophilic chemicals can cross biological membranes easier (Verhaar *et al.*, 1992; Vaes *et al.*, 1998; Klüver *et al.*, 2019). The toxicity of inert chemicals can be explained very well by the lipophilicity of the compound alone (Verhaar classification MoA class 1). Chemical toxicity is also affected by polarization, described by molecular descriptors such as P_VSA_e_2, SpMax4_Bh(p) and SpPosA_B(p) which are found to be higher for toxic compounds. Polarizability describes the ability of a chemical to interact with endogenous molecules that are important in the developmental regulation of zebrafish embryos (Lavado *et al.*, 2020). Electronegativity, the ability of an atom to attract a pair of electrons in a chemical bond, is used to describe the polarity of the bond (ATS4e, GATS2e) (Wong *et al.*, 2014; Przybyłek, 2020). Ionization potential is defined as the amount of energy required to remove one electron from an atom or a molecule and describes the reactivity of the molecule and is used associated with the ability of a chemical to be absorbed by the organism (SpMin4_Bhi, ATS4i, ATS2i) (Schindler, 2016; Przybyłek, 2020; Wang *et al.*, 2022). Chemical toxicity is influenced by molecular size since larger chemicals are absorbed slower by the organism and highly branched chemicals are less toxic compared to straight chain isomers since branched chemicals are less membrane soluble, due to decreased hydrophobicity (Zhang *et al.*, 2006; Ghorbanzadeh *et al.*, 2016). Functional groups such as hydroxyls, characterise low-toxicity chemicals since they provide high solubility that affects the excretion rate and the ability of a chemical to cross biological membranes and accumulate (Gadaleta *et al.*, 2019). To our knowledge, QSAR models that link cardiotoxicity to structural features are very limited but descriptors such as lipophilicity and molecular weight were used to predict ion current changes involved in cardiac action potential generation (Wiśniowska *et al.*, 2015). Most of the chemicals used in this study, from all five Verhaar MoA classes, significantly alter zebrafish heart rates after exposure, raising the question of whether structural characteristics (molecular descriptors) provide enough information to explain the ability of a chemical to also cause cardiotoxicity.

Differential expression analysis revealed a large set of molecular descriptors that contribute to the toxic effect of a chemical and the ability of a chemical to affect zebrafish heart rate. Most of them have been previously described in the literature, however, differential expression analysis can assist in identifying more structural features related to biological properties and help in generating more reliable models for toxicity and cardiotoxicity prediction.

2.5.2 Gene Ontology terms associated with chemical toxicity

Differential expression analysis (SAM) also revealed a set of genes whose expression is altered when toxic chemicals affect the heart rate of zebrafish embryos. Enrichment analysis of these genes primarily identified associations with microRNAs used to regulate fundamental cellular functions, developmental processes and vascular integrity. Four miRNAs were found to be associated with both conditions, where the expression of miR-126a, miR-216, miR-155, and miR-499 are altered. MiR-126a is expressed in the cardiovascular system and is required to maintain vessel integrity during zebrafish vascular development. This miRNA regulates cell migration, reorganisation of the cytoskeleton, capillary network stability and cell survival in zebrafish (Fish *et al.*, 2008; Zou *et al.*, 2011). MiR-126 is only expressed in the endothelial cell lineage and endothelial cell lines regulate vascular endothelial growth factor-dependent PI3 kinase and MAP kinase signalling by directly targeting PI3KR2 and SPRED1 in zebrafish (Fish *et al.*, 2008) and regulates oxidative stress and inflammation (Helmy *et al.*, 2020).

MiR-155, a multifunctional microRNA, is involved in hematopoietic lineage differentiation, apoptosis, immunity, inflammation, viral infections, and vascular remodelling. MiRNA-155 target genes associated with DNA damage-repairing process and found to regulate the MAPK signalling pathway which is involved in cellular stress response, apoptosis, and inflammation responses in zebrafish (Hu *et al.*, 2019). Exposure to toxins such as polystyrene causes cellular damage that increases the generation of reactive oxygen species, causing an upregulation of miR-155 in humans (Ng *et al.*, 2011; Grogg *et al.*, 2016). Various studies have associated lower expression levels of miR-155 with acute coronary syndromes (Cao *et al.*, 2016), but at the same time it is highly expressed in patients with acute myocardial infarction (Matsumoto *et al.*, 2012; Xie *et al.*, 2014), coronary heart disease (Zhang *et al.*, 2019) or congestive heart failure (Cao *et al.*, 2016), and miR-155 knockout improves cardiac remodelling (He *et al.*, 2016). These contradictory results indicate the complexity of the underlying mechanisms involved.

MiR-499 is involved in cardiac and muscle growth and it is overly expressed in developing hearts (van Rooij *et al.*, 2009). MiR-499 is highly expressed in the myocardium and is important in heart development, function and pathology (Sluijter *et al.*, 2010; Wilson *et al.*, 2010; Fu *et al.*, 2011) and is associated with human heart diseases (Chistiakov *et al.*, 2016) It is controlling the expression of the β-myosin heavy chain, enhancing myocardial oxygen metabolism and tolerance under normal conditions (Wan *et al.*, 2018). MiR-499 protects cardiomyocytes from stress-induced apoptosis and has several regulatory factors as targets that inhibit mitochondrial cell apoptosis (Li *et al.*, 2016; Wan *et al.*, 2018). MiRNA-216 has been associated with loss of vascular integrity, haemorrhage during zebrafish development, and coronary artery disease in humans (Wang *et al.*, 2017). MiR-216b, miR-155 and miR-499 downregulation increase cyb561d2 expression in zebrafish, which is important for electron transfer and cell defence and chemical stress (Ahkin Chin Tai *et al.*, 2020).

In addition, among the genes dysregulated after toxic chemical exposure, two more miRNAs were identified after differential expression analysis. MiR-30 has an important role in zebrafish embryonic development by regulating muscle phenotype. This miRNA controls Hedgehog (Hh) signalling during zebrafish embryonic development (Ketley *et al.*, 2013). Studies have identified the role of Hh signalling in muscle specification controlling the switch from fast-twitch fibres to slow-muscle fibres (Blagden *et al.*, 1997). MiR-145 is expressed in vascular smooth muscle cells and its level increases during embryogenesis, especially in the heart (Zeng *et al.*, 2012). MiR-145 expression alterations cause delayed onset of heartbeat pericardial oedema and unlooped heart (Zeng *et al.*, 2009; Zeng *et al.*, 2012) and control cell death regulation through regulation of apoptosis (J. Li *et al.*, 2020; Zhao *et al.*, 2020) and cell proliferation, differentiation and organ development, including heart formation, by regulating the expression of sox9a and sox9b (Yokoi *et al.*, 2009; Lin *et al.*, 2021). Various biological properties were identified to be explanatory of the toxic effects a chemical has on zebrafish. The genes selected were found to be involved in the determination of heart left/right asymmetry, anatomical structure development, involved in morphogenesis, neuron differentiation and development, which are essential for zebrafish embryogenesis. Microtubule stability is important in maintaining cell shape, intracellular transport, cell motility and division (Díaz-Martín *et al.*, 2021), and proper protein degradation is essential for the maintenance of normal cell homeostasis but exposure to toxins alters those processes.

One microRNA and two KEGG pathways were found to be associated with the downregulated genes after toxic chemical exposure. MiR-1, a muscle miRNA, promotes embryonic muscle gene expression (Chen *et al.*,2006; Mishima *et al.*, 2009), and is essential in cardiac and skeletal muscle development and disease in zebrafish (Wang *et al.*, 2019). During cardiac development, miR-1 has been reported to control the balance between cell proliferation and differentiation, it promotes myogenic differentiation and is involved in cell cycle regulation and migration processes and also inhibits cell apoptosis in zebrafish (Zhao *et al.*, 2005; Lu *et al.*, 2014). MiR-1 enhances angiogenesis during muscle regeneration by silencing SARS proteins (Nakasa *et al.*, 2010; Stahlhut *et al.*, 2012), thus knockdown of miR-1 increases SARS protein abundance, which represses VEGFA expression to affect zebrafish embryonic angiogenesis (Lin *et al.*, 2013). MAPK signalling and cellular senescence pathways are found to be downregulated as the chemical concentration increases, causing toxicity. The MAPK signalling pathway is important for the transduction of various extracellular signals to the nucleus, regulating macrophage activity and angiogenesis. Cellular senescence is caused by DNA damage and oxidative stress and protects damaged cells from proliferating. During embryonic development, senescence promotes morphogenesis through cell turnover, tissue remodelling and growth (Da Silva-Álvarez *et al.*, 2020).

On the other hand, two miRNAs were identified to be associated only with changes in zebrafish heart rate. MiR-430 is mainly functional during zebrafish embryonic development and was found to regulate developmental pathways for cell movement, germ layer specification, and axis patterning organ progenitor formation allowing embryonic body plan formation. Altered expression of this miRNA is associated with developmental delay and disturbed cardiovascular and neural systems (Liu *et al.*, 2020). MicroRNA-206-3p targets, that regulate muscle proliferation and differentiation in zebrafish embryos and mice (Kim *et al.*, 2006; Chen *et al.*, 2010; Goljanek-Whysall *et al.*, 2011; Lin *et al.*, 2017). MiR-206 is essential for gastrulation and is one of the most abundant miRNAs during zebrafish embryogenesis and is regulating the proliferation and differentiation of muscle fibroblasts during somitogenesis (Chen *et al.*, 2005). Alteration of miR-206 expression results in severe cell migration defects during zebrafish embryonic development (Liu *et al.*, 2012; Lin *et al.*, 2017).

2.5.3 Clustering chemicals using molecular descriptors and gene count profiles information

The results of this study support the idea that clustering chemicals using molecular descriptors is very different compared to clustering using mRNA sequencing data since chemicals with similar structures have different transcriptional responses, and vice-versa (Sirci *et al.*, 2017). As expected, molecular descriptors clustering was highly associated with the Verhaar MoA of the chemicals, as both methods use structural characteristics for classification. However, molecular descriptor clustering also successfully grouped chemicals based on the heart-rate effect and concentration changes in chemical exposure, underlying the importance of structural classification. Clustering chemicals using mRNA data was found to be useful in clustering chemicals based on their effect on zebrafish heart rate, with higher accuracy compared to molecular descriptors clustering. Still, there was no association between mRNA clustering and MoA classification. These results suggest that Verhaar MoA classification is not necessarily representative of the ability of a chemical to alter zebrafish embryos' heart rate, indicating that chemicals with similar structural characteristics may have different gene expression profiles and consequently affect the organisms through different mechanisms.

2.5.4 Predictive modelling connects molecular descriptors to mRNA clustering

The effort to connect molecular descriptors to mRNA clustering results was not successful when the whole dataset was used. However, after selecting chemicals for Model 3 or Model 4 two relatively accurate models were generated. These models show that molecular descriptors have the potential to predict gene profile clustering and the use of more chemicals, with high variability (in structural characteristics and gene count profiles), can validate this relationship.

Multiple molecular descriptors were selected by the two predictive models (Models 3 and 4). Model 3 molecular descriptors were related to molecular size (GATS8m, GATS8v) (Baldim *et al.*, 2017; Moussa *et al.*, 2021), ionization potential (GATS8i, ATSC8i) (Guan *et al.*, 2018; Bittremieux *et al.*, 2022), polarizability (ATSC8p, WiA_Dz(p), SpMin5_Bh(p)) (Gaudêncio *et al.*, 2022), lipophilicity (CATS2D_04_DL), molecular branching (Eta_sh_x) (Carnesecchi *et al.*, 2020), charge transfer between a pair of atoms (JGI6) (Doucet *et al.*, 2018) and the frequency of carbon and oxygen bonds (F03[C-O]) (Elsayad *et al.*, 2020). These types of molecular descriptors have been identified previously by various studies to be associated with chemical toxicity or used for clustering chemicals based on structural features.

On the other hand, the molecular descriptors identified by model 4, were related to chemical electronegativity (GATS4e) (Carnesecchi *et al.*, 2020), ionisation potential (SpMin8_Bh(i)), the drug-like ability of the chemicals (Depressant-80) (Solimeo *et al.*, 2012) dipole moment (Eig03_EA(dm)) (Zhang *et al.*, 2015), bond order (Eig13_AEA(bo), Eig14_AEA(bo)) (Watkins *et al.*, 2016), the number of the rings (nR05) and the presence or absence of various bonds (B06[C-C], B02[O-S], B03[N-Cl]) (Lavado *et al.*, 2022).

Molecular size, lipophilicity, polarizability, electronegativity and ionization potential of a chemical influence its absorption, distribution, metabolism and excretion (Gleeson, 2008; Waring, 2009; Yang *et al.*, 2012; Gajewicz-Skretna *et al.*, 2021). Lipophilicity is also associated with molecular branching and the presence of rings; chemical lipophilicity is decreased when the molecular branching is high, and ring count correlates positively with lipophilicity (Ritchie *et al.*, 2009; Yang *et al.*, 2012; Gajewicz-Skretna *et al.*, 2021). Dipole moment can be used to predict toxic potency since it is related to the binding affinity of the molecule (Forrest *et al.*, 2014). Chemical drug-like properties derived from the analysis of various physicochemical properties such as lipophilicity, molecular size, molar refractivity, and the number of atoms (Anuta *et al.*, 2014; Tsantili-Kakoulidou *et al.*, 2021). Bond order, described the number of bonds between atoms indicating the stability of the bond, whereas more stable substances require more time to break down, leading to bioaccumulation within the organism, carbon-carbon bonds are strong and stable whereas oxygen-sulphur bonds are weak (Savoca *et al.*, 2021).

The adverse outcome pathway (AOP) framework was developed in order to organise all available information from multiple levels of biological organisation related to risk assessment. AOP is a conceptual construct that portrays existing knowledge concerning the linkage between a direct MIE, the interaction between the chemical and its biological target at the molecular level, and an AO via key events, biological events as a response to MIE, at a biological level of organisation relevant to risk assessment (Ankley *et al.*, 2010). *In silico* screening of new chemicals against known MIE or key events of AOPs allows the classification of chemicals based on their biological activation profile and prediction of AOs (Leist *et al.*, 2017).

At the core of the QSAR and AOP fields is the predictive ability towards a relevant, mainly regulatory, endpoint. However, conceptually these two approaches differ significantly. While QSAR focuses on the assumption that chemical structural information must be associated with its impact on an organism, AOPs argue that one, or more, MIE causes a cascade of molecular perturbations that lead to an AO, indicating the importance of transcriptomics data.

Predicting molecular responses clustering using molecular descriptors will reveal potential connections between MIE and KE and make chemical classification and AO prediction easier, and more accurate. Using molecular descriptors to predict gene count profiles will allow the prediction of possible MIE and KE *in silico*, reducing the number of experiments while taking into consideration the molecular effect of a chemical.

2.5.5 Conclusion

Differential expression analysis can be used to identify structural features that contribute to the ability of chemicals to cause toxicity or cardiotoxicity, uncover important details related to the structural features-toxicity relationship, or uncover potential mechanisms of toxicity by identifying differentially expressed genes. Identifying the genes associated with an AO (death, heart-rate changes), allow the identification of the underlying events occurring at a molecular level which can represent key events in an AOP. Clustering chemicals using structural information has been widely used, however, clustering using molecular responses (gene count profiles) is gaining a lot of attention since those two clustering methods cluster chemicals differently. In addition, the fact that structurally similar compounds (cis- and trans isomers) may have different toxic effects on the organism, or chemicals with different structural characteristics that act through the same mechanisms, increases the popularity of clustering using molecular responses. Structural information can be used as an indicator of the toxicity effect of a chemical. Still, molecular information provides more details about the mechanism of toxicity and the combination of those will provide more information about the underlying biological processes of toxicity.

Molecular descriptors are used in the generation of QSAR models that have been widely used for predicting chemical toxicity, but their power decreases when they are used on data with chemicals from multiple MoAs and for MoAs 4 and 5. However, the results of this study show that chemicals that affect the heart rate of zebrafish belong to all 5 Verhaar MoAs, suggesting that chemicals that cause cardiotoxicity are characterized by various structural features, increasing the difficulty of predictive model generation.

In this study, the ability of molecular descriptors to predict molecular response clustering (mRNA-seq data) was evaluated. The results suggest that molecular descriptors have the potential to predict molecular response clustering and the use of more chemicals, with high variability (in structural characteristics and gene count profiles), can validate this relationship. The ability to predict molecular information such as MIE and key events without the need for experiments, will allow integration into existing AOPs and increase the accuracy of *in silico* risk assessment.

Chapter 3

Predicting changes in heart rate in the model species Danio rerio

3.1 Abstract

Heart development and function are very sensitive processes, and susceptible to environmental toxins. Chemicals that disrupt vascular remodelling, cardiomyocyte proliferation and cardiac differentiation, ATP production, cell death, and various signalling pathways, have a significant impact on the heart-rate fold change of zebrafish embryos. In this study, the ability of structural features (molecular descriptors) or gene expression profiles to explain the variation in heart rates was evaluated using predictive modelling with random forest. Molecular descriptors fail to predict changes in heart rate when the dataset is characterised by high variability, chemicals with various structural features (and modes of action) or different gene expression profiles. On the other hand, predictive modelling identified multiple genes associated with signal transduction, cell death and cardiac action potential that were found to be predictive of the heart-rate fold change of zebrafish embryos when 143 chemicals, characterized by high variability (diverse structural and gene expression profiles) were used.

Clustering chemicals based on molecular responses (gene count profiles) reduces the variability of the dataset since only chemicals with similar gene count profiles are grouped together, generating three clusters. For chemicals from clusters 1 and 2 structural information including lipophilicity, electrochemical characteristics and the presence and nature of bonds were found to be predictive of the chemical effect on zebrafish heart rate. Finally for chemicals from cluster 3, only mRNA information was able to predict the impact on the heart rate using genes associated with ATP production, signalling pathways and nervous system development. The results suggest that structural information and molecular responses (gene count profiles) are complementary, and when used together can assist chemical risk assessment.

3.2 Introduction

Environmental pollution continues to be a major threat to public health. Exposure to environmental contaminants and toxins during embryonic development can affect heart development and function in various organisms, defined as cardiotoxicity (Onakpoya *et al.*, 2016; Georgiadis *et al.*, 2018). Cardiotoxicity, the damage of the heart muscle and other cardiac tissues or disruption of the electrophysiology of the heart due to toxic compounds, can lead to inadequate pumping of blood through the body, or cardiac muscle dysfunction (Basak *et al.*, 1991). Heart development is a very sensitive process, susceptible to environmentally toxic compounds, and molecular, and cellular factors. However, chemical assessment in mammals is time-consuming and relatively expensive (R. Li *et al.*, 2020). Zebrafish (*Danio rerio),* a small tropical fish native to Southeast Asia, can be used as an alternative animal model for toxicological screening, due to its simple maintenance, low cost, and fast growth. Their small size allows the use of well plates and the simultaneous evaluation of multiple individual organisms (Dai *et al.*, 2014; Nasrallah *et al.*, 2018) and their highly permeable skin allows small molecules to be added directly into the water and absorbed by the fish (Milan *et al.*, 2003; McLeish *et al.*, 2010; Martin *et al.*, 2019). Their rapid and external development of zebrafish embryos and their transparency allows for easier and relatively straightforward phenotypic assessment.

The zebrafish genome has been fully sequenced and shares high similarity with the human genome, where 70% of human genes have zebrafish orthologs (Howe *et al.*, 2013), and a large number of genes (84%) and regulatory networks that have been associated with human diseases have also been identified in zebrafish. Thus zebrafish have been used as model organisms for studying various human diseases (Poon *et al.*, 2013). In addition, Zebrafish and humans respond similarly to cardiotoxic compounds (Zhu *et al.*, 2014), and show similar electrocardiogram (ECG) recordings (Hodgson *et al.*, 2018).

Zebrafish have been used as a model organism in developmental biology and molecular genetics providing important information about the molecular regulation of vertebrate cardiac development due to their high conservation among vertebrates (Miura *et al.*, 2011; Staudt *et al.*, 2012; Chen, 2013). The stark similarities between zebrafish and human cardiogenesis, the low cost, rapid cardiovascular development and the ability of the zebrafish to survive without a fully functional cardiovascular system increase zebrafish popularity in cardiovascular research (Nishimura *et al.*, 2016). In addition, the transparency of the embryo allows for non-invasive heart measurements that can be used to assess stage-specific exposure effects on zebrafish embryos, which facilitates the understanding of functions involved in cardiac development (Sarmah *et al.*, 2016; Caballero *et al.*, 2018).

The rapid increase of anthropogenic chemicals in the environment, from industrial and pharmaceutical advances, make *in silico* approaches that can evaluate new chemicals based on similarities with previously defined compounds more and more essential (Kausar *et al.*, 2018). Quantitative Structure-Activity Relationship (QSAR) models have been used to predict the physicochemical properties, and biological and environmental impact of a chemical, using molecular descriptors. Molecular descriptors are the representation of physicochemical properties, such as topological representation, connectivity of atoms, the presence and nature of chemical bonds and chemical lipophilicity and polarizability, into numeric values either as a result of standardised experiments or as a result of logical and mathematical procedures (Todeschini *et al.*, 2009; Roy *et al.*, 2015). QSARs reduce the number of experiments, by relating the structure and physicochemical properties of a chemical with biological activities, selecting only a number of compounds for in vivo experiments (Neves *et al.*, 2018) in risk assessment (Melnikov *et al.*, 2016), in medicinal chemistry, in academy, industry and government institutions (Cherkasov *et al.*, 2014).

On the other hand, signature-matching approaches have evolved, that are based on the assumption that chemicals with similar gene expression signatures, will cause similar biological effects, thus can be used in predicting the toxicity of new chemicals and identifying toxic-related genes (Lamb *et al.*, 2006; Smalley *et al.*, 2010; Sarmah *et al.*, 2016). The popularity of those methods was facilitated by the advances in microarray technologies, and subsequently, RNA-seq, where variations between RNA profiles enabled the understanding of the toxic effect at a system level. The transcriptome represents the response to numerous cellular signals, including that of exogenous compounds. With the latest technologies even low abundant RNA species can be measured in a given sample providing one of the most complete representations of the molecular state of a sample.

The numerous approaches in which an organism might respond to a given signal, however, provide a challenge when trying to establish regulatory-relevant response profiles for given exposure scenarios. The adverse outcome pathway (AOP) framework was developed to organise all this available information from multiple levels of a biological organisation with a focus on exogenous perturbation. An AOP portrays existing knowledge of the linkage between an MIE and an AO via key events and the respective key event relationships to improve risk assessment (Ankley *et al.*, 2010). AOPs are chemically agnostic and can be represented by multiple levels of biological organisation. Their modular properties enable the linkage and interaction between AOPs to form AOP networks which can be utilised to identify key events, which are central to several AOs and MIEs. AOPs are represented as unidirectional chains of events from MIE to AO, but this is rarely true in biology. The AOP concept is a simplified version of the response which usually overlooks the presence of positive/negative feedback loops and multiple pathways that connect to the AOP. Despite these shortcomings, AOPs are extremely useful in regulatory applications as *in silico* screening of new chemicals against known MIE or key events of AOPs allows the classification of chemicals based on their biological activation profile and adds information to that of structural similarity (Leist *et al.*, 2017).

The AOP concept allows the integration of molecular mechanisms into the field of regulatory toxicology, identifies uncertainties and research priorities (Ankley *et al.*, 2010) and supports decision-making in hazard identification and risk assessment by chemical prioritisation or exclusion early in development (Leist *et al.*, 2017). The development of AOPs is partly based on scientific literature and partly on newly generated data, however the lack of reproducibility (Hartung *et al.*, 2013), the presence of contradictory data and the lack of transparency (Leist *et al.*, 2010; Leist *et al.*, 2012), increase the difficulty of those process. AOPs reflect the current state of knowledge, thus they can continue to evolve as new information becomes available.

Both QSAR and AOPs use the similarities between the chemicals to predict an AO. QSARs are relying only on the available structural characteristics of a chemical, whereas AOPs use all the available information that allows the identification of key events that lead to an AO. Chemical similarity profiles can be very different when using gene expression profiles versus structural similarities (Sirci *et al.*, 2017). QSAR models have been widely used for predicting toxicity when the dataset consists of highly similar chemicals, especially chemicals that act through polar and non-polar narcosis. However, data with high variability, in structure and mode of action, or the use of reactive chemicals and chemicals that act through specific mechanisms, decreases the accuracy of the QSAR model.

In this study, the aim is to identify if molecular descriptors (QSAR) or mRNA data (genes) can be used to generate a reliable and accurate model that can predict the toxicity of a highly diverse set of chemicals (in structural features and gene count profiles). In addition, since the classification of data reduces the variability within the dataset, by grouping similar chemicals together, clustering based on gene count profiles (chapter 2) can be used to evaluate the strength of QSAR compared to mRNA data in predicting toxicity.

3.3 Methods

3.3.1 Building heart rate datasets and chemical clustering

The chemicals used in this study were the same as in the Chapter 2 analysis, 143 chemicals. Zebrafish embryos were exposed to various chemicals (6 concentrations- LC50, LC5, LC5/2, LC5/4, LC5/8, LC5/16, 6 individual embryos in separated wells per concentration), and their heart rate was estimated using Fiji with the time series analyser V3 plugin (Schindelin *et al.*, 2012). DMSO exposure was used as a control. The structural characteristics in the form of molecular descriptors were calculated for each chemical using Dragon 7 software (Mauri *et al.*, 2006) and mRNA sequencing provides us with a gene count profile for each chemical as described in chapter 2. Two datasets were generated with 143 chemicals each, with the heart-rate fold change as the dependent variable for both, the molecular descriptors as the independent variables for the first dataset and the gene count profiles for the second dataset.

The high variability of the data, chemicals with various structural features and gene count profiles, increases the difficulty of generating robust and reliable models. One way to overcome this complication is grouping the dataset into smaller clusters based on similarities in gene count profiles. The hard chemical clustering using mRNA-seq data, described in Chapter 2, was used to generate three smaller (43,34 and 66 chemicals each) datasets.

3.3.2 Identifying differentially expressed genes associated with chemical impact on zebrafish heart rate

The dataset used in this study, consists mostly of chemicals that significantly change the heart rate of zebrafish embryos compared to the controls. To explore the effect of those chemicals on the zebrafish at the gene level, differential expression analysis (SAM) was performed (Tusher *et al.*, 2001). In an effort to identify differentially expressed genes between exposure to chemicals that do not affect zebrafish heart rate, including controls, (dependent variable = “0”), and those that cause bradycardia or tachycardia (dependent variable value= “1”), chemicals were classified into those two categories. Gene count profiles were used as the independent variable. The SAM function was performed with 1000 permutations (nperms=1000) used for estimating the FDR, the nature of the data was specified (resp.type= "Two class unpaired"), and the FDR cutoff for output in significant genes table was set to 0.1 (fdr.output=0.1).

The genes identified by SAM to be differentially expressed when chemicals affect heart rate were processed using the R function gost from the gprofiler2 package for functional enrichment analysis (version 0.2.1) (Kolberg *et al.*, 2020). The gost function was performed using FDR as the algorithm for correcting for multiple testing (correction_method = "fdr") with a threshold set to 0.1 (user_threshold = 0.1), specifying the organisms used (organism = "drerio") and generating a list of only the statistically significant results (significant = TRUE) among all the genes of the given organism (domain_scope = c("known")). Gene enrichment analysis reveals a set of biological processes (GO: BP), and KEGG and Reactome (REAC) pathways that the identified genes are found to be involved in, and also identifies the targets of multiple miRNAs within that list (miRNA).

3.3.3 Predicting heart rate fold change using molecular descriptors and mRNA-seq data

Eight datasets in total were used as input to the predictive function described in Chapter 2. Four datasets consist of gene count profiles as the independent variables and the heart rate fold change as the dependent variable. These four datasets differ in the number of chemicals, the first one consists of all 143 chemicals used in this study, and the other 3 represent the clusters with 43, 34 and 66 chemicals each. The other four datasets, based on molecular descriptors, followed the same rationale with the number of chemicals per dataset (143,43,34,66 chemicals each) and the heart-rate fold change caused by exposure as the dependent variable.

The same approach (predictive modelling function) as in chapter 2 was applied for the generation of the predictive model but instead of classification, regression analysis was used, since the dependent variable is quantitative (heart-rate fold change), compared to the classification data (clusters) used in chapter 2. Thus, the “family” parameter of the LASSO function was set to “gaussian” instead of "multinomial", the “classification” parameter for random forest analysis was set to “FALSE” instead of “TRUE” and "variance" and "extratrees” were used as splitrules for LOOCV.

The package ggplot2 (version 3.3.5) was used for visualising the results of the models (Wickham, 2016). To evaluate the importance of each variable to the model, the mean square error (%IncMSE) and the contribution of each variable to the homogeneity of random forest nodes and leaves were calculated for every variable, the higher those values the higher the importance of that variable to the model (varImpPlot function, random forest R package version 4.7-1.1) (Liaw et al., 2002). The genes identified by predictive modelling were then processed using the gost function from the gprofiler2 R package (version 0.2.1) for functional enrichment analysis.

3.4 Results

3.4.1 Differential expression and functional enrichment analysis to identify biological properties involved in cardiotoxicity

The heart rate fold change was estimated for the 143 chemicals used in this study, at multiple concentrations, and as it can be seen in Figure 3.1, increasing chemical concentration is associated with bigger effects on heart rate as expected. Chemicals such as clozapine, prochloraz, terfenadine, tacrine and chlorpromazine cause bradycardia in zebrafish embryos and have a greater effect on heart rate than expected based on their concentration (Figure 3.1). These five chemicals cause bradycardia in zebrafish but based on the mRNA clustering from Chapter 2 their gene count profiles are different, as terfenadine was grouped in cluster 1 whereas the rest, clozapine, prochloraz, tacrine and chlorpromazine were grouped in cluster 2. This is an indication that various biological mechanisms are involved in heart development and function.

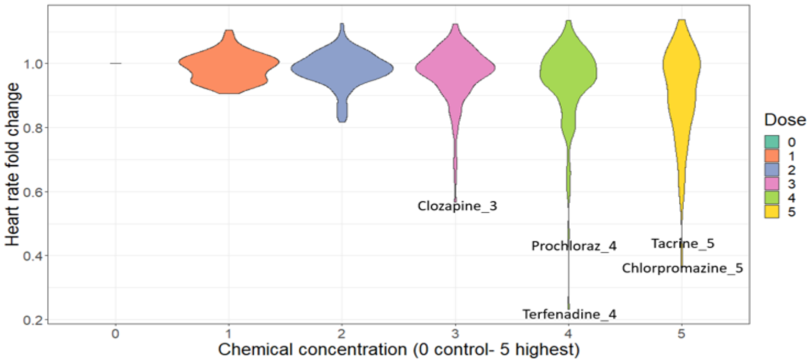


Figure 3.1: Zebrafish were exposed to multiple chemical concentrations and their heart-rate fold change was calculated. As the concentration increases (dose) the effect chemicals have on heart rate also increases. Some outliers can be identified.

Most of the chemicals used in this study (66%) significantly affect the heart rate of zebrafish embryos. Differential expression analysis (SAM) was performed in an effort to identify genes that are upregulated or downregulated after exposure to chemicals that significantly affect zebrafish heart rate. From the 31954 genes used for differential expression analysis, 7938 genes were identified to be upregulated when the heart rate is significantly influenced by chemical exposure and 6977 genes were downregulated. Compared to the results from chapter 2, more genes were identified to be upregulated or downregulated, however, not all of the genes identified in chapter 2 were selected by the differential expression analysis

(SAM) using the two-class unpaired method. This analysis allows the selection of genes that are differentially expressed when the chemical has a significant impact on zebrafish heart rate.

Functional enrichment analysis identified multiple biological properties, KEGG and Reactome pathways that are significantly associated with the upregulated genes (Table 3.1, Supplementary materials Table S.2). Biological functions and pathways associated with organ development (p-value= 9.2e-12 - 8.9E-02), heart development and function (p-value= 9e-05-0.09), circulatory system development and negative regulation of hematopoietic stem cell differentiation (p-value=0.005-0.09), were identified to be significantly associated with heart rate changes caused by chemical exposure. Nervous and immune system development and function (p-value= 2e-08-0.09) and mitochondria function and ATP production (p-value= 0.002-0.097) were identified as expected to be significantly associated with cardiotoxicity (bradycardia or tachycardia). Ion transmembrane transport (p-value= 2.3e-05- 0.09) and cell-cell communication (p-value= 6e-04 -0.08), describe the movement of ions such as potassium and sodium and other small molecules across biological membranes regulating cardiac function and contraction (Fountoulaki *et al.*, 2015; Grandi *et al.*, 2017). In addition, the genes found to be upregulated when the chemical caused cardiotoxicity, were involved in various signalling pathways including Wnt, PPAR, FoxO, mTOR, MAPK, ErbB and VEGF signalling pathways (p-value=2e-07 -0.096), since signal transduction facilitates the response to extracellular stimuli to ensure proper heart function (Wheeler-Jones, 2005). Finally, among the genes found to be upregulated, pathways involved in the negative regulation of DNA transcription and RNA biosynthesis, metabolic processes splicing and degradation, were identified (p-value= 2.2e-04 - 0.9).

On the other hand, the downregulated genes were involved in pathways associated with embryonic development (cell and organ development) (p-value=3.3e-05 - 9.3e-02) and heart and muscle development, muscle contraction and ATP production (p-value= 0.001-0.09) as expected. Genes involved in cell death (p-value= 6.8e-05 - 0.09), nervous system development and neuroactive ligand-receptor interaction (p-value= 0.0005 - 0.086) and immune system responses (p-value= 2.9e-0.5 - 8.1e-02) were selected. In addition DNA replication, transcription and repair, generation of mature mRNA, ribosomes and post-translational modifications pathways were identified after functional enrichment analysis (p-value= 7.6e-05 - 0.095). Genes involved in cell-cell communication and transmembrane transport of ions and molecules (p-value= 5.4E-09 - 0.09) were found to be downregulated. Finally, genes were involved in multiple signalling pathways associated with various processes including development (Wnt, VEGF, VEGFA-VEGFR2, MAPK), immune response (Toll-like, Interleukins, TRC), cell death (TGF-beta, ErbB, FoxO) and gene expression (mTOR, FoxO, p53) (p-value=4.4e-05 - 0.09) were identified by differential expression analysis (Table 3.1, Supplementary materias Table S.2).

Heart-rate fold changes -Genes upregulated	
GO term	**P value**
Organ development	9.2E-12 - 8.9E-02
Heart development and function	0.04-0.08
Circulatory system	0.005-0.09
Mitochondria	0.002-0.097
Neuronal system	7E-05 - 0.86
Immune system	2E-08 - 0.09
Transmembrane ion transport	2.3E-05 - 0.09
Cell-cell communication	0.0006-0.08
Signalling pathways	2E-07 - 0.096
DNA and RNA	2.2E-04- 0.09
Heart-rate fold changes -Genes downregulated	
Embryonic development	3.3E-05 - 9.3E-02
Cardiac function and ATP production	0.001-0.09
Cell death	6.8E-05 - 0.09
Nervous system	0.0005-0.086
Immune responses	2.9E-05 - 8.1E-02
DNA and RNA	7.6E-05 - 0.095
Cell-cell communication	5.4E-09 - 0.09
Signalling pathways	4.4E-05 - 0.09

Table 3.1: Functional enrichment analysis of the upregulated and downregulated genes associated with significant heart rate fold change.

3.4.2 Prior clustering of chemicals by molecular effects, significantly improves heart-rate prediction using molecular descriptors

The relationship between toxicity and concentration suggests that chemical structure plays a crucial role in predicting physiological outcomes resulting from chemical exposure. To test the ability of the chemical structure to predict heart rate the stability path feature selection approach was used for the whole dataset (143 chemicals). The model reached an R^2 of 0.19 with an adjusted p-value of 0.006, where the p-value indicates the probability of the selected model being generated with randomized data, thus the smaller the p-value the more reliable the model is, however, the low R^2 indicates the low accuracy of this model, the inability of the model to accurately predict the heart rate fold change caused by chemical exposure (Table 3.2, Figure 3.2). A set of 80 descriptors was selected for the generation of this model and were related to multiple functional groups that are used in pharmaceuticals or industrial products, including pyridines, aromatic primary amides, and amines, however, this model can not predict the experimental heart rate changes caused by chemical exposure in this study, especially the effect of triclosan, tacrine, terfenadine, propylene oxide, prochloraz and chlorpromazine (Figure 3.2).

	Molecular descriptors				**mRNA sequencing data**			
	Number of variables	**R^2**	**Validate (p-value)**	**FDR**	**Number of variables**	**R^2**	**Validate (p-value)**	**FDR**
143 chemicals	80	0.19	0.002	0.006	80	0.68	0	0
Cluster 1	35	0.39	0.054	0.072	28	0.58	0.19	0.25
Cluster 2	11	0.63	0.003	0.006	29	0.6	0.35	0.35
Cluster 3	6	0.17	0.219	0.219	21	0.64	0.03	0.05

Table 3.2: Generating predictive models for heart-rate fold change in zebrafish using molecular descriptors and mRNA sequencing data. Several models were generated using the whole dataset (143 chemicals) and the 3 clusters separately. For each model, the number of variables selected, the R^2, the p-value from validation by randomisation of the dependent variable and the adjusted p-value (FDR), were recorded.

The low predictive ability of the model raises the question of whether chemical heterogeneity plays a crucial role in heart-rate prediction using structural information since QSAR models usually fail to generate accurate models when chemicals from different MoA classes are used. It stands to reason that selecting chemicals with similar structural characteristics for predictive modelling can generate a more accurate and reliable model as these should contain aspects of structural similarity that carry the effect on heart rate and associated toxicity. Therefore, the mRNA clustering generated in Chapter 2 was applied to divide the chemicals into three groups (43, 34, 66).

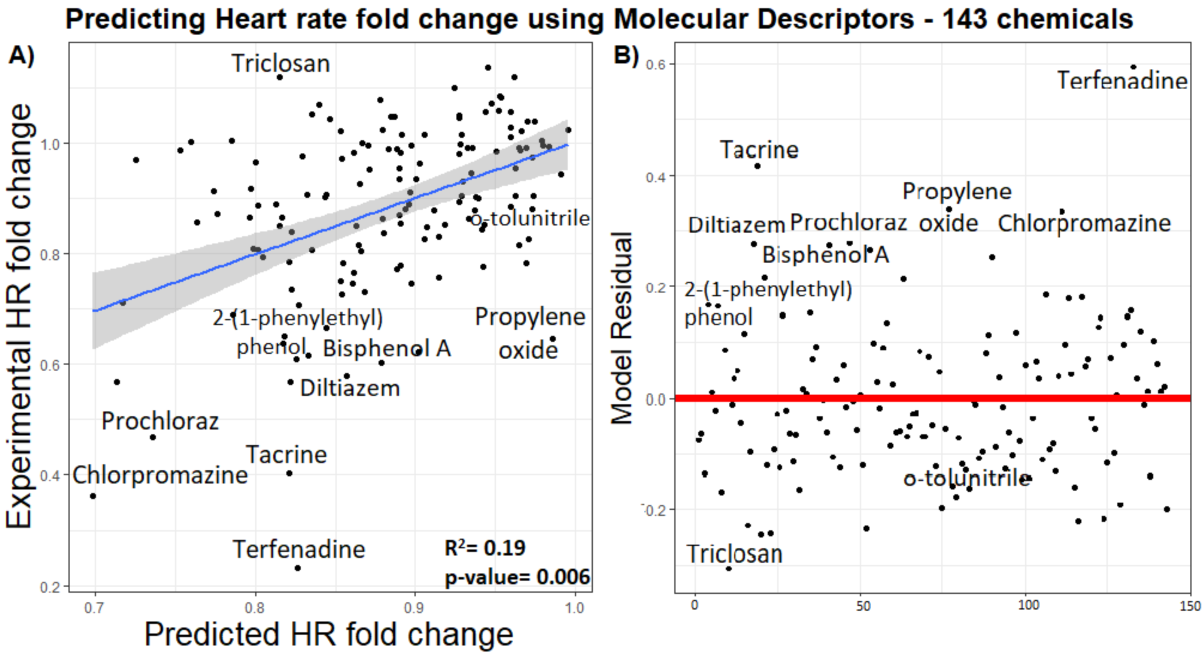


Figure 3.2: Predicting heart rate fold change using molecular descriptors and 143 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change, $R^{\wedge 2}$=0.19. B) Model Residuals calculated by the predicted heart rate fold change values, using molecular descriptor model generated for the whole dataset (143 chemicals).

The model generated using chemicals from mRNA cluster 1 (43 chemicals) consisted of 35 molecular descriptors and had a mean $R^{\wedge 2}$ of 0.387, and an adjusted p-value of 0.072, an indication that the model is representative of the data provided, thus more reliable (Figure 3.3, Table 3.2). The predictive accuracy of this model (using chemicals from cluster 1) is higher compared to the model generated using a set of 143 chemicals, but still generally weak. Descriptors such as the number tertiary amines (aliphatic) (nRNR2), along with the presence or absence of N-O and C-N bonds (B05[N-O], F05[N-O], B02[N-O], B03[N-O], (F01[C-N], B05[C-N], B01[C-N], B02[C-N]), the charge transfer between a pair of atoms (JGI4), molecular mass (ATSC2m, GATS1m), ionization potential (MATS4i) and molecular electronegativity (rGes) were found among the most important variables for the generation of the model. In addition, the amount of van der Waals surface area having potential pharmacophore points from atoms belonging to cycles (P_VSA_ppp_cyc) and two molecular descriptors related to the number of atom-centred fragments C-006 (CH2RX) and C-026 (R- - CX- -R) were identified, where “R” is defined as any group linked through carbon, “X” is any electronegative atom (O, N, S, P, Se, halogens) and “- -” represents an aromatic bond. Despite the improved predictive capacity the model failed to accurately predict the effect of terfenadine and bisphenol A on zebrafish embryo heart rate, however, this model was able to predict the effect of triclosan more accurately compared to the model generated using the 143 chemicals (Figure 3.3).

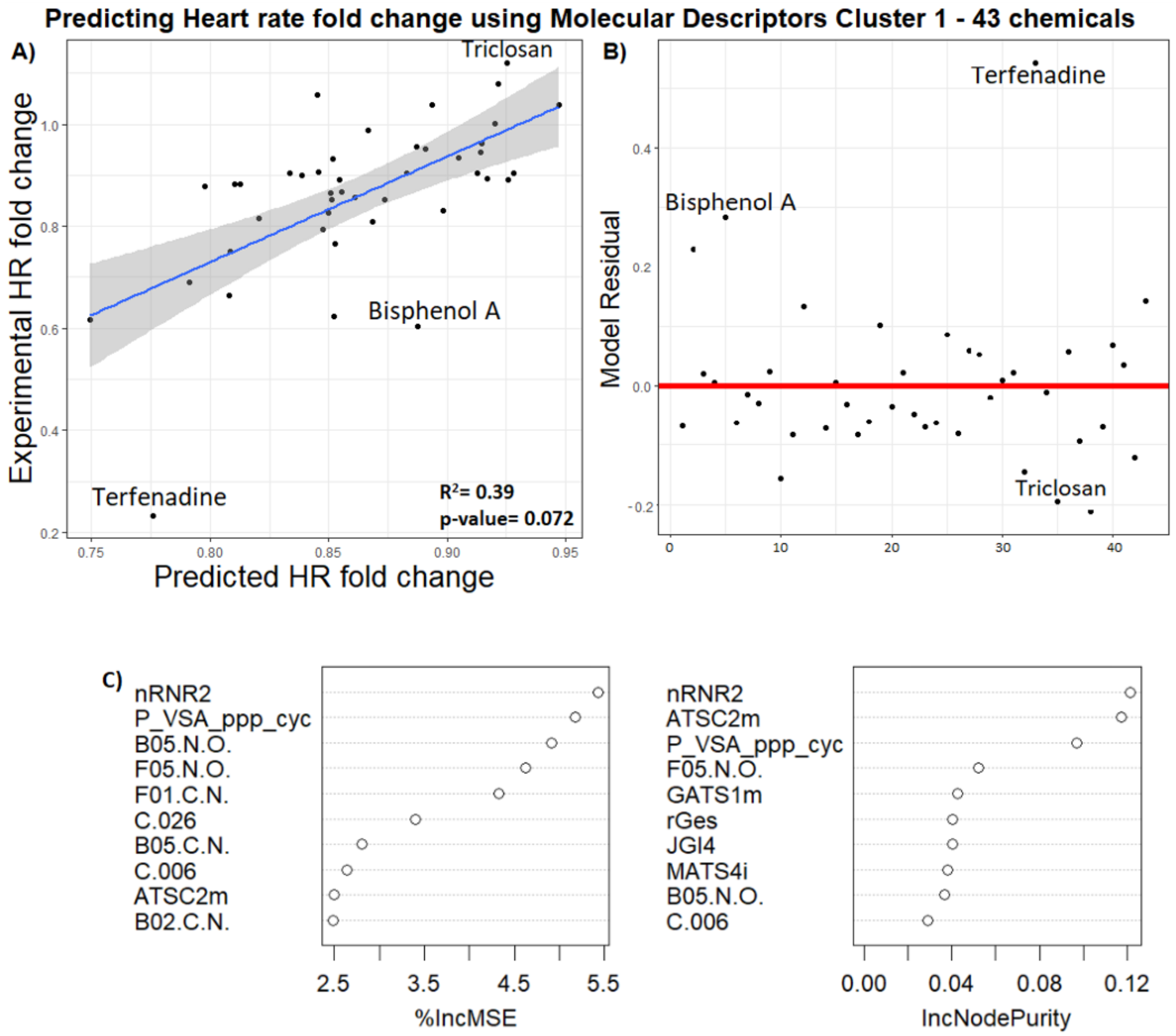


Figure 3.3: Predicting heart rate fold change using molecular descriptors and cluster 1 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change (43 chemicals, R^2= 0. 39. B) Model Residuals calculated by the predicted heart rate fold change values. C) The mean square error (%IncMSE) and the contribution of each variable to the homogeneity of random forest nodes and leaves (IncNodePurity) of the 10 most important variables.

For predicting the heart rate fold change of the chemicals that belong to cluster 2, which consists of 34 chemicals, 11 molecular descriptors were selected and had an R^2 of 0.634, with an adjusted p-value of 0.006, the smaller the p-value the more reliable the model is (Figure 3.4, Table 3.2). While the number of chemicals is low and overfitting may occur, the bootstrap-based approach used can mitigate some of these effects by using multiple subsamples of the original dataset to evaluate the performance of the model. A set of 11 molecular descriptors were identified including the presence or absence of C-Cl and N-Cl bonds (B08[C-Cl], B08[N-Cl]), the number of triple bonds (nTB), the number of aromatic tertiary amines (nArNR2), ring fusion density (RFD), molecular descriptors related to molecular mass (MATS1m), ionization potential (MATS4i), polarizability (MATS7p) and the charge transfer between a pair of atoms (JGI8). Finally, atom-centred fragments C-008 (CHR2X) and C-040 (R-C(=X)-X / R-C#X / X=C=X) were selected, where “R” is any group linked through carbon, “X” is any electronegative atom, “#” represents a triple bond and “=” represents a double bond. Despite the relatively high predictive ability of the model (with R^2 of 0.63), Figure 3.4 shows that predicting the effect of Tacrine, a chemical that causes bradycardia in zebrafish was harder compared to the rest of the chemicals from cluster 2. On the other hand, this model was able to better predict the effect of chlorpromazine, prochloraz and propylene oxide compared to the model generated using the 143 chemicals.

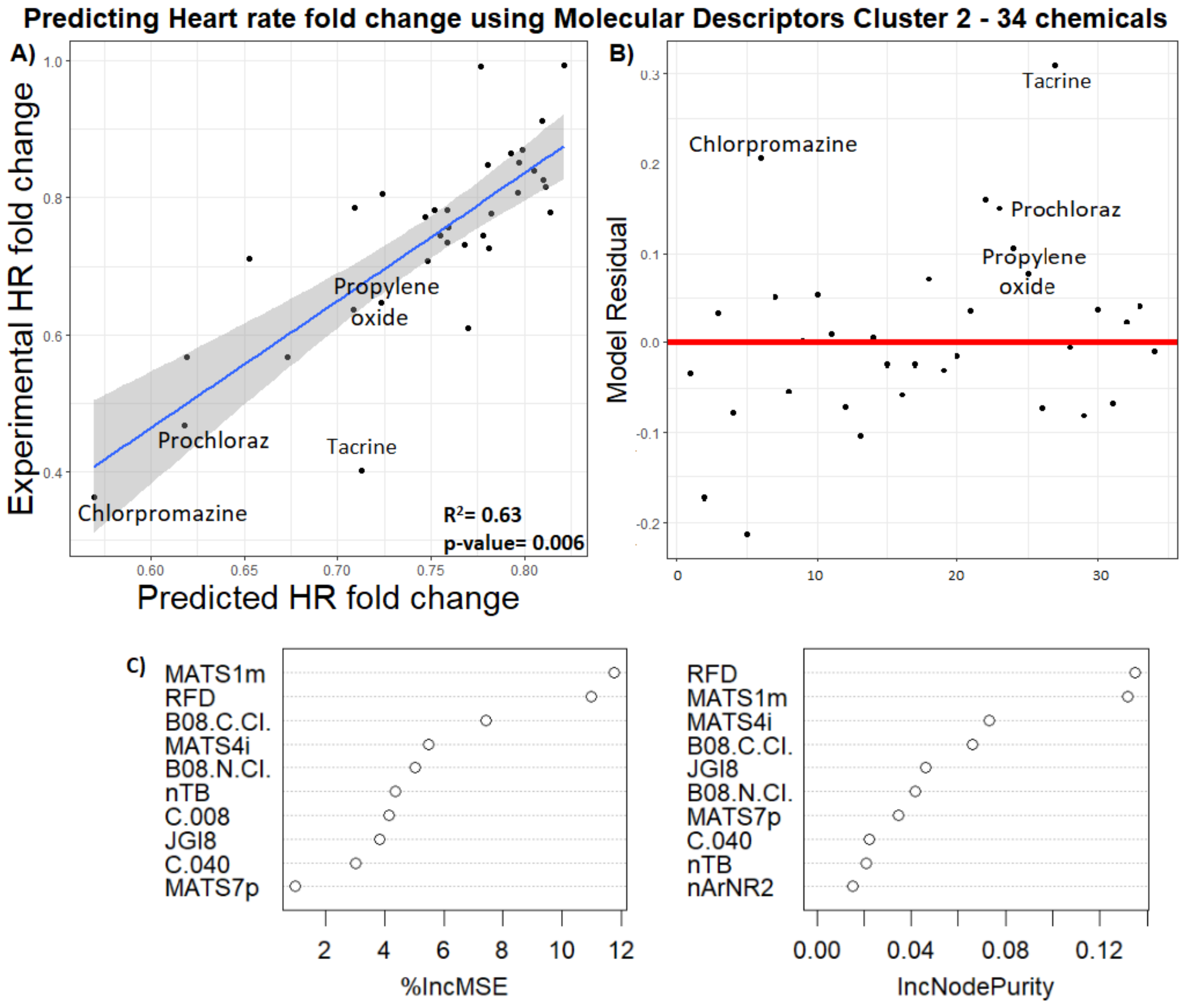


Figure 3.4: Predicting heart rate fold change using molecular descriptors and cluster 2 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change (34 chemicals, R^2= 0. 634). B) Model Residuals calculated by the predicted values of heart rate fold change.C) The mean square error (%IncMSE) and the contribution of each variable to the homogeneity of random forest nodes and leaves (IncNodePurity) of the 10 most important variables.

Lastly, the number of descriptors required to generate a model to predict heart-rate changes using chemicals from mRNA hard cluster 3 (66 chemicals) highlighted 6 MDs with an R^2 =0.17, and an adjusted p-value of 0.219 (Table 3.2, Figure 3.5). Indicating the inability of structural information to predict the effect these chemicals have on zebrafish heart rate. Incidentally, this is also the cluster where over 50% of chemicals have no or very little effect on the zebrafish embryo, suggesting that these compounds act through a different, non-cardiotoxicity-related pathway.

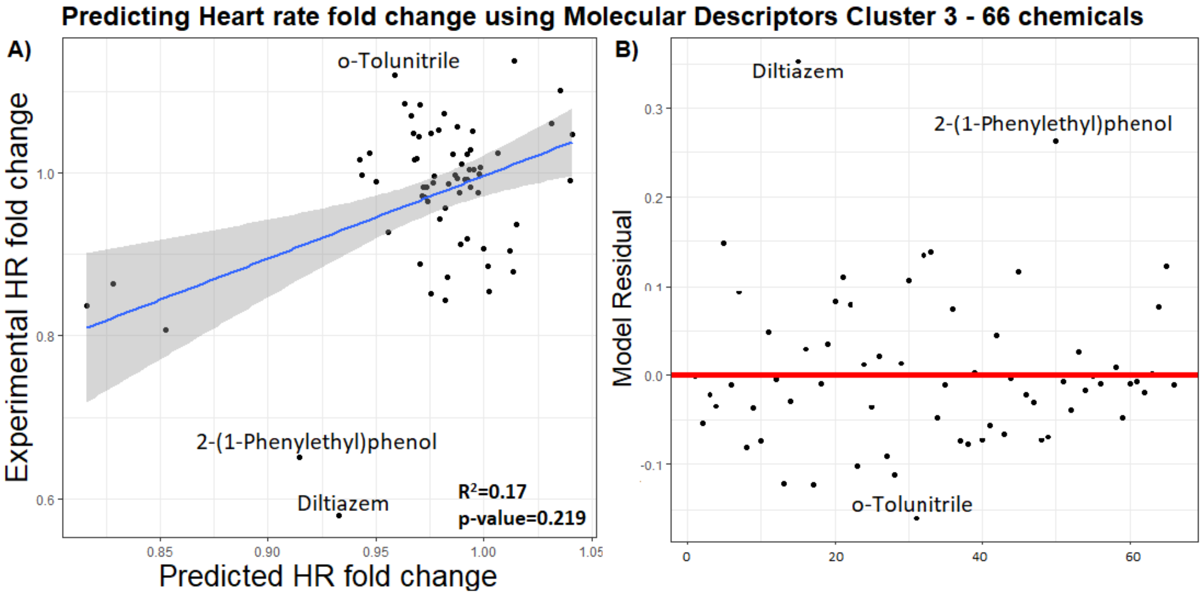


Figure 3.5: Predicting heart rate fold change using molecular descriptors and cluster 3 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change (66 chemicals), R^2= 0. 174 B) Model Residuals calculated by the predicted values of heart rate fold change.

Comparing the molecular descriptors selected by the various models a minimum overlap was observed (Figure 3.6, Table 3.3). Three of the molecular descriptors selected by the whole dataset model (143 chemicals), were also selected by the cluster 1 model (ChiA_X,nArCONH2, B08[Cl-Cl]), two molecular descriptors (nArNR2, B08[N-Cl]) were also selected by cluster 2 model, and only one (CATS2D_06_AN) was shared with cluster 3 model. The model generated using chemicals only from cluster 1 shares only one molecular descriptor (MATS4i) with the cluster 2 model and one (JGI4) with the cluster 3 model. No common molecular descriptors were identified between cluster 2 and cluster 3 models. This very low overlap between the molecular descriptors selected by the three models generated using the three clusters highlights the difference between the chemicals of each cluster in structural characteristics. The R^2 values generated for each training data set indicate that each model has the same rate of success when multiple datasets were used with only a few outliers, showing that model performance was consistent when subsamples of each dataset were used (Figure 3.7).

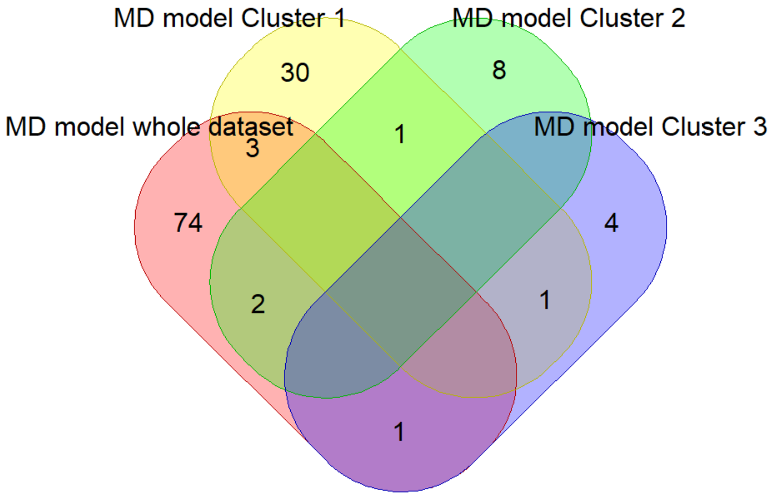


Figure 3.6: Comparing the molecular descriptors selected by the 4 molecular descriptors models. Optimisation of the models resulted in very different models with little to no overlap between the models.

Shared Molecular Descriptors	143 chemicals	Cluster 1 (n=43)	Cluster 2 (n=34)	Cluster 3 (n=66)
ChiA_X	X	X		
nArCONH2	X	X		
B08[Cl-Cl]	X	X		
nArNR2	X		X	
B08[N-Cl]	X		X	
CATS2D_06_AN	X			X
MATS4i		X	X	
JGI4		X		X

Table 3.3: Molecular descriptors shared by the various models generated using multiple datasets, whole dataset (143 chemicals), mRNA clusters 1,2 and 3.

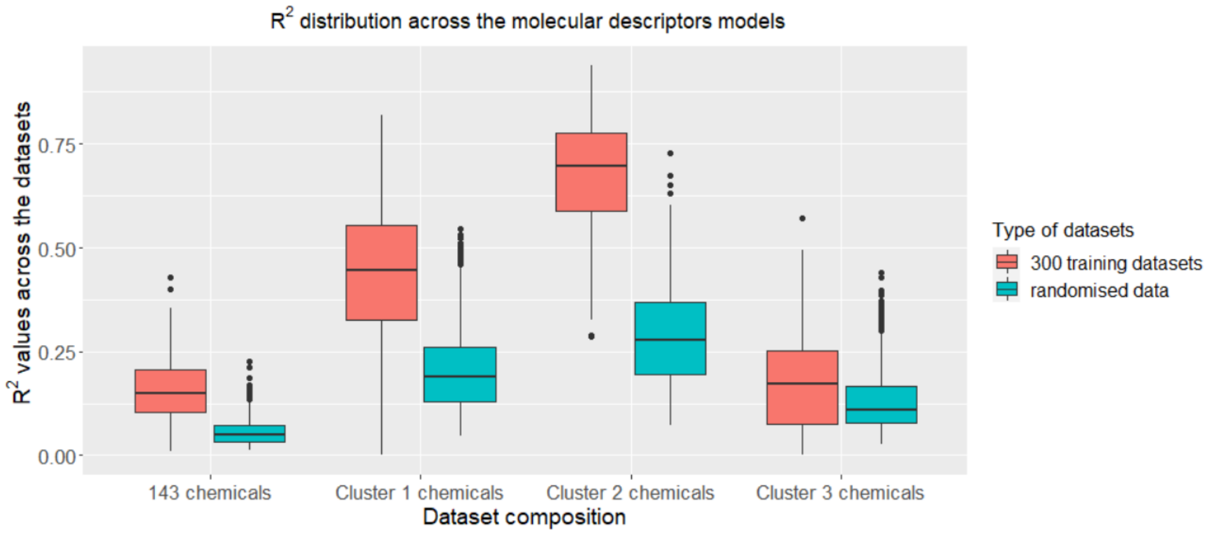


Figure 3.7: The R^2 values generated using the molecular descriptor models over the 300 training datasets (colour=red) and the results of the validation with the randomised data (colour= green), over the different datasets, 143 chemicals, cluster 1, cluster 2 and cluster 3 chemicals.

3.4.3 Molecular response is highly predictive of heart rate

As the molecular state (gene expression profile) is much closer to the phenotypic outcome, heart rate should be easier to be predicted both across the whole dataset and within the defined effect clusters from chapter 2. Therefore, in order to predict the heart-rate fold change of zebrafish embryos for all 143 chemicals, 80 genes were needed with R^2 of 0.683 (adjusted p-value = 0), indicating a relatively accurate and reliable model (Table 3.2, Figure 3.8). This already highlights the significant improvement in heart-rate prediction. The 80 genes selected were associated significantly with 11 biological processes, including cardiac jelly development, chemical stimulus detection, sensory perception of smell, immune response regulation and cholecystokinin signalling pathway (p-value=0.096) (Table 3.4). A total of 11 genes were identified to contribute the most towards the accuracy of the model, including natriuretic peptide A (nppa) involved in cardiac jelly development, Cholecystokinin A receptor (cckar) involved in muscle contraction, internexin neuronal intermediate filament protein alpha a (inaa) and TIAM Rac1 associated GEF 2a (tiam2a) involved in neurons morphogenesis and axogenesis, defensin beta-like 3 (defbl3), FYN binding protein b (fybb), si:ch211-170i2.2 and Janus kinase 3 (jak3), that are involved in innate and adaptive immune response, lens intrinsic membrane protein 2.2 (lim 2.2) involved in eye lens development, and gap junction protein alpha 13.2 (gja13.2) and si:ch211-203k16.3 that are involved in cell adhesion (*ZFIN The Zebrafish Information Network*, 2020, *National Center for Biotechnology Information*, 2020). Despite the high R^2, this model failed to predict the effect of Terfenadine, Chlorpromazine, Tacrine and Prochloraz on zebrafish heart rate (Figure 3.8). These chemicals were also identified as highly toxic previously, suggesting that exposure to these chemicals influences heart development and function through different mechanisms compared to the rest.

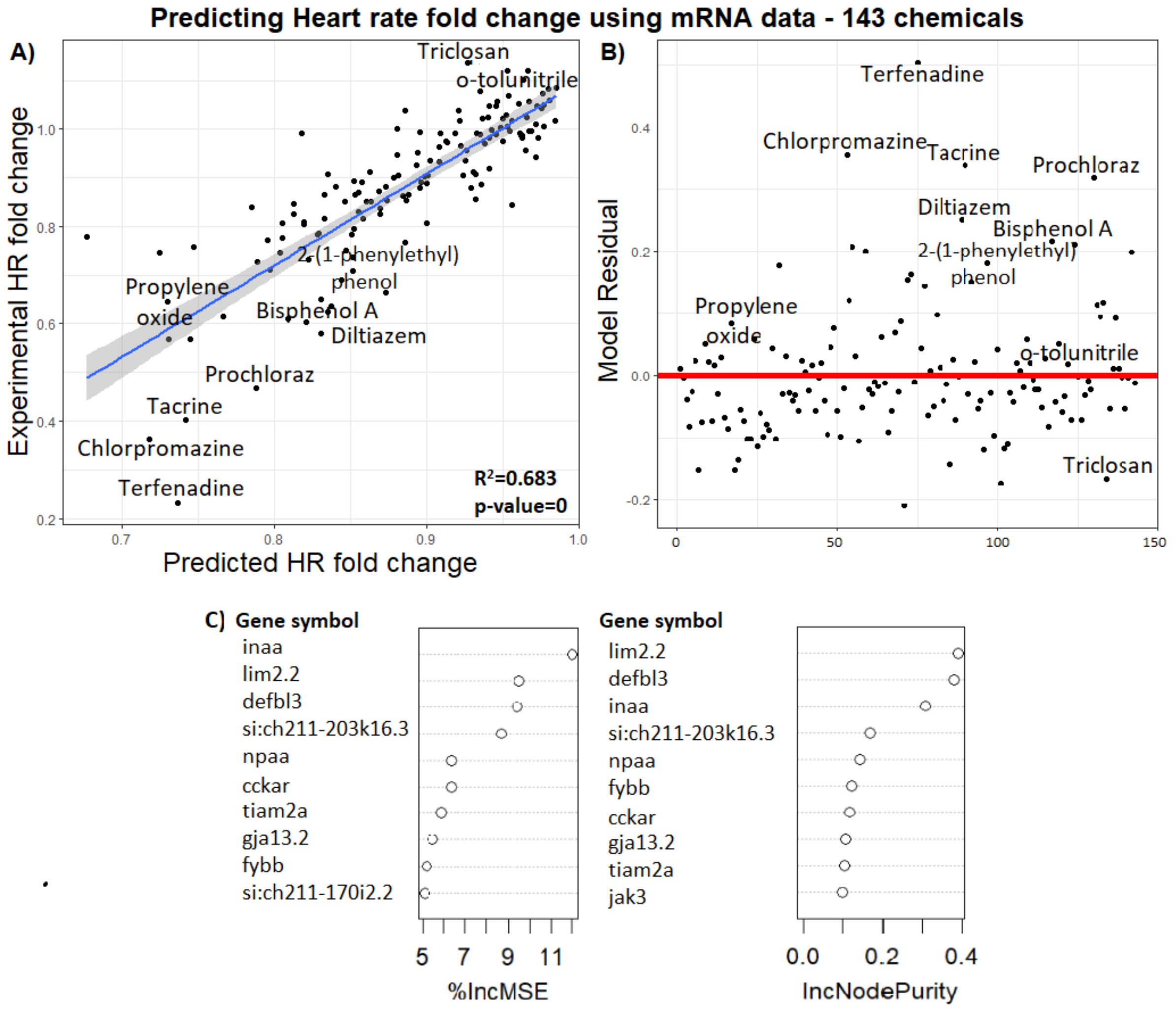


Figure 3.8: Predicting heart rate fold change using mRNA data and 143 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change, R^2= 0. 683. B) Model Residuals calculated by the predicted values of heart rate fold change. C) The mean square error (%IncMSE) and the contribution of each variable to the homogeneity of random forest nodes and leaves (IncNodePurity) of the 10 most important variables.

In addition, the chemicals within each cluster were used in predictive modelling using mRNA data. Cluster 1 resulted in a model with 28 genes and R^2 of 0.58, however, after bootstrap validation the adjusted p-value calculated was 0.250, indicating that the prediction has near a random probability of occurring (Table 3.2, Figure 3.9). The genes selected by this model are associated with neuroactive ligand-receptor interaction, SNARE interactions in vesicular transport, secretin family receptors, GPCR ligand binding, vasopressin-like receptors, Tie2 Signalling, signalling by GPCR and RET signalling (Table 3.5). In addition, this model also failed to predict the effect of terfenadine on zebrafish heart rate (Figure 3.9). The model generated using the 143 chemicals could better predict the chemical effect on zebrafish embryos’ heart rate and is more reliable (adjusted p-value=0) compared to the model generated using only chemicals from cluster 1.

Functional enrichment - 143 chemicals modelling		
GO term name	**p-value**	**GO term ID**
Cholecystokinin signalling pathway	0.09	GO:0038188
Detection of chemical stimulus involved in sensory perception of smell	0.09	GO:0050911
Sensory perception of smell	0.09	GO:0007608
Detection of chemical stimulus involved in sensory perception	0.09	GO:0050907
Germinal center formation	0.09	GO:0002467
Detection of chemical stimulus	0.09	GO:0009593
Regulation of immune response	0.09	GO:0050776
Cardiac jelly development	0.09	GO:1905072
Cell communication	0.09	GO:0007154

Table 3.4: Functional enrichment analysis performed using the genes selected to be predictive of heart-rate changes for the whole dataset.

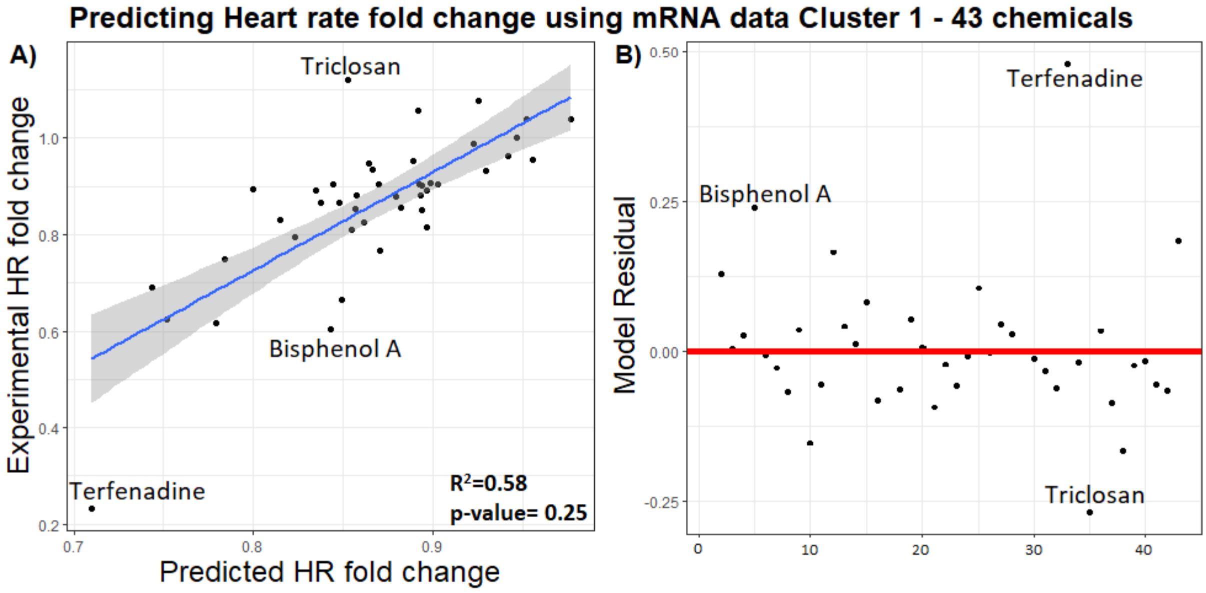


Figure 3.9: Predicting heart rate fold change using mRNA data and cluster 1 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change, (43 chemicals, R^2= 0. 58. B) Model Residuals calculated by the predicted values of heart rate fold change.

Functional enrichment-Cluster 1 modelling		
GO term name	**P value**	**GO term ID**
Neuroactive ligand-receptor interaction	0.04	KEGG:04080
SNARE interactions in vesicular transport	0.04	KEGG:04130
Class B/2 (Secretin family receptors)	0.027	REAC:R-DRE-373080
GPCR ligand binding	0.04	REAC:R-DRE-500792
Vasopressin-like receptor	0.04	REAC:R-DRE-388479
Tie2 signalling	0.07	REAC:R-DRE-210993
Signaling by GPCR	0.08	REAC:R-DRE-372790
RET signalling	0.099	REAC:R-DRE-8853659

Table 3.5: Functional enrichment analysis performed using the genes selected to be predictive of heart-rate changes when only chemicals from cluster 1 were used.

For the prediction of heart-rate fold change caused by the 34 chemicals from cluster 2, 29 genes were selected, with a mean R^2 of 0.60, and an adjusted p-value of 0.353, also indicating that this model can be generated even when using randomised data, showing that it is unsuitable for predicting heart rate (Table 3.2, Figure 3.10). The 29 genes selected were associated with two KEGG pathways and 26 Reactome pathways, including sphingolipid metabolism (KEGG pathway: p value=0.04), homologous recombination (KEGG and Reactome pathways: p-value= 0.04-0.07) and potassium and calcium ion channels (Reactome pathways: p-value=0.05-0.09) (Table 3.6). Despite the high R^2 of this model still fails to accurately predict the effect of chlorpromazine and tacrine on zebrafish embryos' heart rates (Figure 3.10). The model generated using the 143 chemicals could better predict the chemical effect on zebrafish heart rate and is more reliable (adjusted p-value=0) compared to the model generated using only chemicals from cluster 2.

Functional enrichment-Cluster 2 modelling		
GO term name	**P value**	**GO term ID**
Sphingolipid metabolism	0.075	KEGG:00600
Homologous recombination	0.075	KEGG:03440
Homologous DNA Pairing and Strand Exchange	0.072	REAC:R-DRE-5693579
Homology Directed Repair	0.091	REAC:R-DRE-5693538
HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA)	0.091	REAC:R-DRE-5693567
HDR through Homologous Recombination (HRR)	0.086	REAC:R-DRE-5685942
Presynaptic phase of homologous DNA pairing and strand exchange	0.072	REAC:R-DRE-5693616
Resolution of D-loop Structures through Holliday Junction Intermediates	0.072	REAC:R-DRE-5693568
Resolution of D-Loop Structures	0.072	REAC:R-DRE-5693537
VxPx cargo-targeting to cilium	0.072	REAC:R-DRE-5620916
Activation of G protein gated Potassium channels	0.072	REAC:R-DRE-1296041
Activation of GABAB receptors	0.072	REAC:R-DRE-991365
Classical Kir channels	0.072	REAC:R-DRE-1296053
G protein gated Potassium channels	0.072	REAC:R-DRE-1296059
GABA B receptor activation	0.072	REAC:R-DRE-977444
Phase 4 - resting membrane potential	0.072	REAC:R-DRE-5576886
Synthesis of PA	0.072	REAC:R-DRE-1483166
Inwardly rectifying K+ channels	0.072	REAC:R-DRE-1296065
Inhibition of voltage gated Ca2+ channels via Gbeta/gamma subunits	0.072	REAC:R-DRE-997272
Insulin processing	0.072	REAC:R-DRE-264876
GABA receptor activation	0.086	REAC:R-DRE-977443
Cargo trafficking to the periciliary membrane	0.086	REAC:R-DRE-5620920

Table 3.6: Functional enrichment analysis performed using the genes selected to be predictive of heart-rate changes when only chemicals from cluster 2 were used.

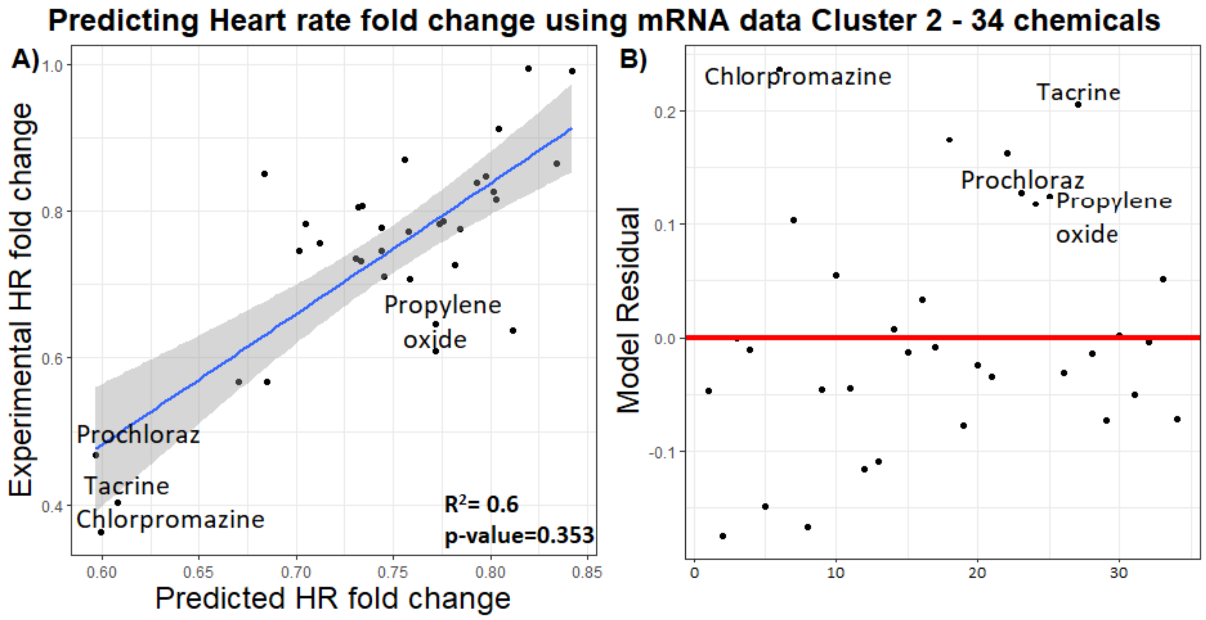


Figure 3.10: Predicting heart rate fold change using mRNA data and cluster 2 chemicals. A): Plotting the predicted and the experimentally calculated heart rate fold change, (34 chemicals, R^2= 0.6. B) Model Residuals calculated by the predicted values of Heart rate fold change.

Lastly, for cluster 3 (66 chemicals), the model generated consists of 21 genes, has an R^2 of 0.64 and an adjusted p-value of 0.05, the high R^2 indicates a relatively accurate model and the low adjusted p-value the high reliability of the model (Table 3.2, Figure 3.11). The 21 genes were associated with various biological processes, KEGG and Reactome pathways, involved in mitochondrial translation and ATP production (p-value= 0.27-0.099), regulation of development and cell growth (p-value= 0.034-0.099), synthesis of ribonucleic acids, DNA synthesis and repair (p-value= 0.017-0.099) and in immune responses (p-value=0.06) (Table 3.7, Supplementary materials Table S.3). This model failed to predict the effect of Diltiazem and 2,1-Phenylethyl phenol on heart rate (Figure 3.11). The model generated using the 143 chemicals once again could better predict the chemical effect on zebrafish heart rate and is more reliable (adjusted p-value=0) compared to the model generated using only chemicals from cluster 3.

Looking through the genes selected by the various models, again there is very little overlap (Figure 3.12). Two genes (ENSDARG00000092668 -F1R128, ENSDARG00000093868-si:dkey-11o15.6p) were shared between the full data set and the model generated using only the chemicals from cluster 1 (Table 3.8). These genes are associated with the regulation of immune system processes and defence responses. One microRNA (ENSDARG00000104494) was found in the models generated from the 143 chemical dataset and the chemicals from mRNA cluster 2. Finally, one gene (ENSDARG00000052900-zgc:153642) was selected for both models generated by using mRNA cluster 2 and 3 chemicals, associated with ion binding and guanyl ribonucleotide binding (Table 3.8). The models generated using the whole dataset (143 chemicals) and the chemicals from cluster 3 (66 chemicals) do not have any genes in common. In addition, the model generated using cluster 1, did not share any genes with the models generated using chemicals from cluster 2 or cluster 3, highlighting the difference between the chemicals in the separated clusters. A closer look through the R^2 calculated for each training dataset, the model generated for the whole dataset, had more consistent results among the various training sets compared to the models generated for the three clusters (Figure 3.13). In addition, all models seem to behave similarly when multiple training datasets were used, with very few outliers.

Functional enrichment-Cluster 3 modelling	
GO term	**p-value**
Mitochondria and ATP production	0.025-0.099
Regulation of development and cell growth	0.034-0.099
Ribonucleic acid synthesis **DNA transcription** **DNA repair**	0.017-0.099
Immune response	0.06

Table 3.7: Functional enrichment analysis performed using the genes selected to be predictive of heart-rate changes when only chemicals from cluster 3 were used.

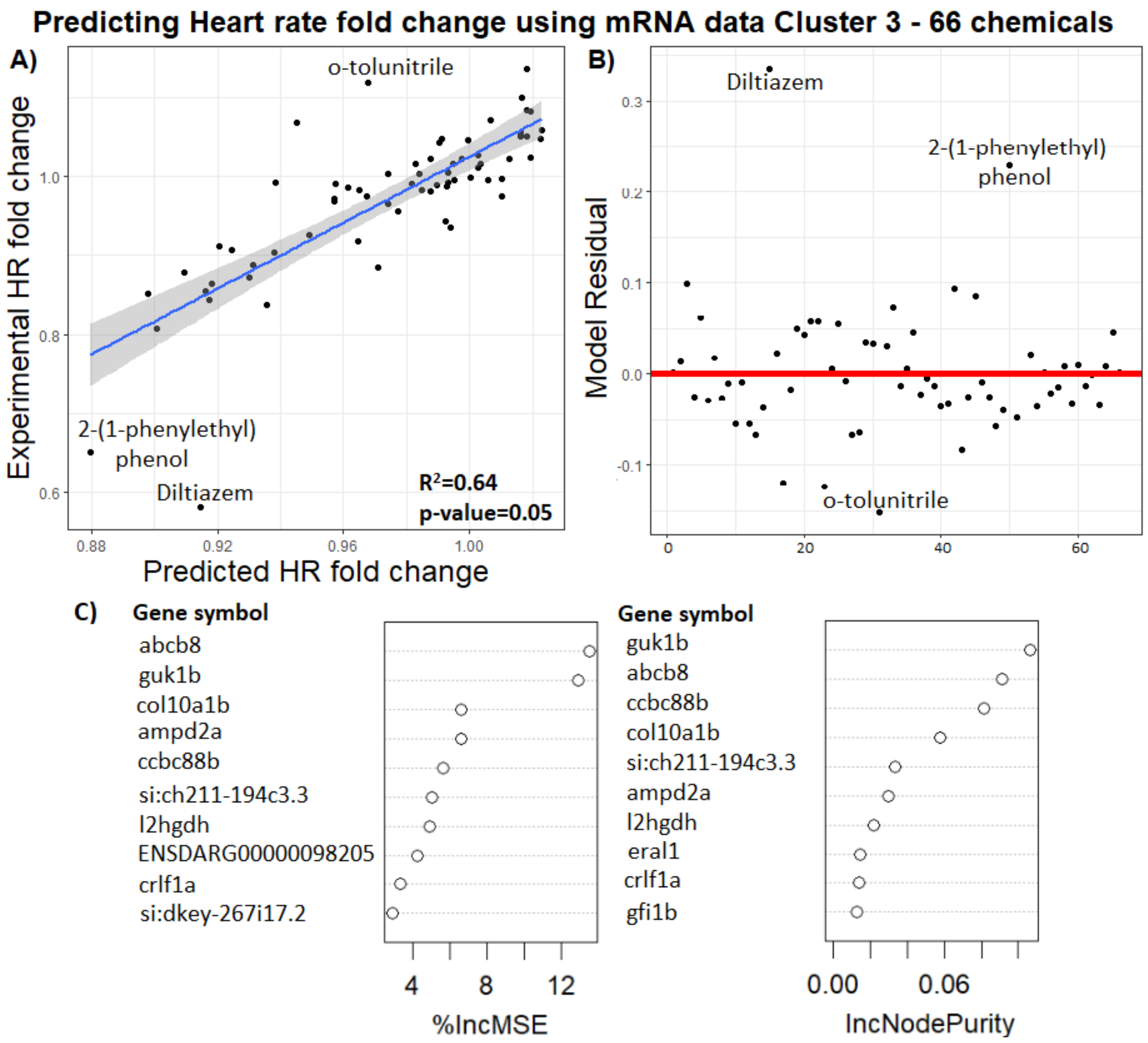


Figure 3.11: Predicting heart rate fold change using mRNA data and cluster 3 chemicals. A) Plotting the predicted and the experimentally calculated heart rate fold change, (66 chemicals), R^2=0. 6401664, B) Model Residuals calculated by the predicted values of Heart rate fold change. C) The mean square error (%IncMSE) and the contribution of each variable to the homogeneity of random forest nodes and leaves (IncNodePurity) of the 10 most important variables.

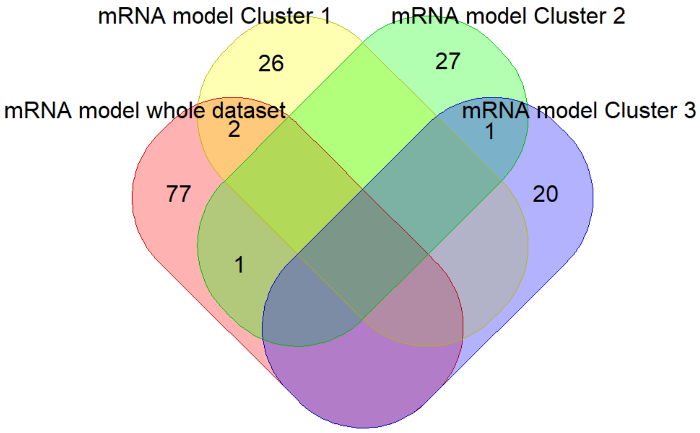


Figure 3.12: Comparing the genes selected by the 4 mRNA models. Optimisation of the models resulted in very different models with little to no overlap between the models.

Shared Genes	143 chemicals	Cluster 1 (n=43)	Cluster 2 (n=34)	Cluster 3 (n=66)
F1R128	X	X		
si:dkey-11o15.6p	X	X		
ENSDARG0000010449	X		X	
zgc:153642			X	X

Table 3.8: Genes shared by the models generated using the whole dataset (143 chemicals), the chemicals from cluster 1, cluster 2 and cluster 3.

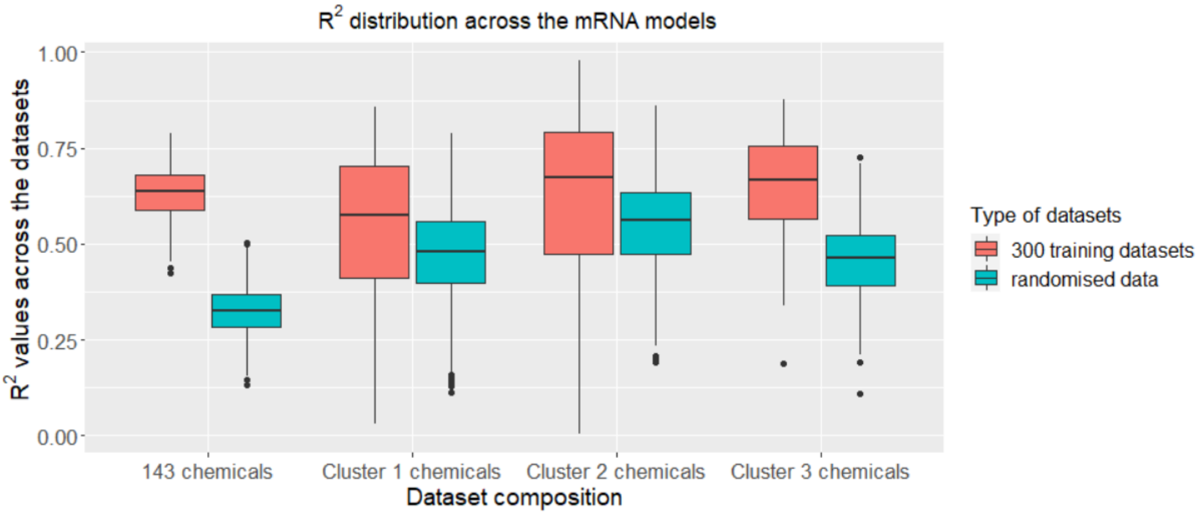


Figure 3.13: The R^2 values generated using the mRNA models over the 300 training datasets (colour= red), and the results of the validation with the randomised data (colour=green), over the different datasets, 143 chemicals, cluster 1, cluster 2 and cluster 3 chemicals.

3.5 Discussion

3.5.1 Chemicals that cause bradycardia

Zebrafish heart development is highly controlled by cellular differentiation, migration, proliferation and apoptosis, thus toxic chemicals that alter those processes can lead to cardiovascular diseases, bradycardia, and deformities. Chemicals, such as clozapine, prochloraz, terfenadine, chlorpromazine and tacrine, cause severe bradycardia in the zebrafish embryos after exposure. Clozapine, an antipsychotic drug for treating schizophrenia, has been found to cause cardiotoxicity in zebrafish embryos by increasing oxidative stress and causing up-regulation of inflammatory cytokines. In addition, morphological abnormalities have been identified including oedema, incomplete heart looping and bradycardia in zebrafish (Abdel-Wahab *et al.*, 2015; Zhang *et al.*, 2021).

Prochloraz, a widely used fungicide, has been associated with teratogenesis and induces multiple embryonic developmental anomalies, like spine deformation, slower heart rate, oedemas, and even hatching failure of zebrafish embryos (Domingues *et al.*, 2013). Terfenadine is a withdrawn antihistamine that increases the interval between heart contracting and relaxing (inducing QT interval prolongation), causing cardiac arrhythmias. However, stroke blood volume seems to be increased due to increased filling time, resulting in stable cardiac output, compensating for bradycardia in zebrafish. Zebrafish embryo terfenadine exposure was also found to be associated with pericardial oedema and ventricle collapse (Maciag *et al.*, 2022). Chlorpromazine is an antipsychotic drug that has been found to suppress the current of potassium channels (hERG) causing QT prolongation (He *et al.*, 2021; Li, Tang and Li, 2021). Finally, tacrine has been used to treat Alzheimer’s symptoms, by blocking the breakdown of acetylcholine neurotransmitters, increasing their abundance. Such inhibitors cause the accumulation of acetylcholine in zebrafish embryos causing bradycardia (Lin *et al.*, 2007).

3.5.2 Genes associated with significant heart rate fold change

Differential expression analysis revealed a set of genes found to be up or down-regulated after exposure to chemicals that significantly affect the heart rate of zebrafish embryos. The upregulated genes were found to be involved in various embryonic developmental processes, such as multiple organ development and assembly and localization of cellular components, as toxic exposure can cause morphological abnormalities, developmental delay and death (Yang *et al.*, 2009; Bambino *et al.*, 2017; Hellfeld *et al.*, 2020).

Processes related to heart development and function, including muscle, circulatory system and nervous system developmental processes and function were found to be altered after exposure to toxic chemicals, as expected. Cardiac development and function are very sensitive processes to external and internal stimuli during embryonic development (Sarmah *et al.*, 2016). Heart contraction is controlled by the autonomic nervous system and utilises a large number of ATP, thus defects in muscle tissue formation, mitochondrial respiration or nervous system development and function have been associated with cardiotoxicity in zebrafish (Dubińska-Magiera *et al.*, 2016; Fedele *et al.*, 2020). In addition, cross-talk between the heart and immune system, through hormones, neurotransmitters and cytokines, where dysregulation of the immune system and inflammatory pathways leads to arrhythmias and heart failure (Dal Lin *et al.*, 2019). Finally, the genes found to be upregulated are involved in cell-cell communication which describes the ability of the cell to receive, process and transmits signals with the environment and itself, in order to respond and adapt to changes, controlling cell death and survival, cell proliferation and cellular differentiation (Uings *et al.*, 2000).

On the other hand, the genes identified to be downregulated by differential expression analysis (SAM) were also associated with multiple KEGG and Reactome pathways. Pathways involved in embryonic development, cardiovascular development and function, nervous system development, immune responses and cell-cell communication, were found to be significantly altered after chemical exposure. In addition, pathways involved in cellular senescence and apoptosis were also found to be influenced by toxic exposure.

Cardiac development is coordinated by cellular, molecular, and environmental factors (Olson, 2006). Biochemical signalling, components of extracellular matrix and cell-cell communication (Bornhorst *et al.*, 2019), cardiomyocyte contractility (Auman *et al.*, 2007) and intracardiac hemodynamic flow (Hove *et al.*, 2003; Radisic *et al.*, 2004) control the heart development of zebrafish. G-protein receptors perceive multiple extracellular signals and transduce them intracellularly. Electric currents can influence cell morphology and cardiac organogenesis (Chi *et al.*, 2010), and *in vivo* studies have shown that cardiomyocytes are realigning based on conduction directionality (Hove *et al.*, 2003; Radisic *et al.*, 2004). Disrupted cardiac conduction can lead to changes in the intracellular calcium gradient, which causes the redistribution of integrins of N-cadherin, specifically expressed in the myocardium, which could lead to loss of cell-cell contact between cardiomyocytes and alter cell shape and overall cardiac morphogenesis (Chi *et al.*, 2010). Cell surface receptors can be redistributed in the cell membrane by electric fields, including N-cadherin which mediates calcium-dependent homophilic cell-cell contact, after physiologic depolarization of synapses (Tanaka *et al.*, 2000). Various drugs have been found to block the hERG channels that control the electrical activity of the heart by mediating the repolarization currents (potassium ions) in the cardiac action potential that helped coordinate the heart's beating. Blockage of hERG channels, lead to delayed repolarization and predisposition to lethal arrhythmia (Simpson *et al.*, 2020). Exposure to toxic chemicals tends to increase reactive oxygen species (ROS), nitric oxide synthase (NOS) activity and expression of cytokines. Autophagy is the engulfment, degradation, and recycling of dysfunctional or damaged cellular components is important for cardiac development, and cardiomyocyte differentiation (Zhang *et al.*, 2012; Lee *et al.*, 2014). Apoptosis during embryogenesis is involved in cell differentiation of heart myocardium and endocardial cushions through cell-cell interactions (Pyati *et al.*, 2007), thus changes in apoptosis may lead to vascular remodelling (Poelmann *et al.*, 2005).

The large number of genes identified to be altered after toxic exposure are involved in various biological processes and pathways, highlighting the complexity of cardiac development and function, and the various mechanisms through which chemical exposure can lead to cardiotoxicity, toxic chemicals can alter zebrafish heart rate through various mechanisms (chemical MoA).

3.5.3 Gene expression profiles for predicting Heart rate fold change of a highly diverse dataset

Predictive modelling indicates that structural information (molecular descriptors- QSAR models) cannot accurately predict the heart-rate fold change of a set of highly diverse chemicals (in structural characteristics) as efficiently as the mRNA profiles. Qsar models are usually performed using chemicals with similar structural characteristics, from a single MoA, in order to generate accurate and reliable models, as the use of reactive chemicals and chemicals that act through specific mechanisms decrease the accuracy of the QSAR model (Yuan *et al.*, 2007; Michielan *et al.*, 2010; Cassotti *et al.*, 2015). The mRNA model generated when a total of 143 chemicals were used, consists of 80 genes that are associated with multiple biological processes including, cardiac jelly development (nppa gene), immune responses (defbl3, fybb, si:ch211-170i2.2 genes), cholecystokinin signalling pathway (cckar gene), detection to chemical stimulus (gja13.2, si:ch11-203k16.3) and sensory perception of smell. Despite the relatively high R^2, of the model, failed to predict the full effect of terfenadine, chlorpromazine, tacrine, and prochloraz, chemicals identified to cause bradycardia in zebrafish as described above.

The genes selected by this model were also associated with cardiac jelly development, the extracellular matrix (ECM) separating the myocardium from the endocardium (Stankunas *et al.*, 2008; Lockhart *et al.*, 2011) and is involved in heart morphogenesis (Taber, 1998; Segert *et al.*, 2018; Männer *et al.*, 2019). Cardiac jelly components are involved in cell shape, migration, proliferation and differentiation, through interactions with growth factors, controlling cell behaviour, cell-to-cell communication and gene expression (Luxán *et al.*, 2016; Silva *et al.*, 2020). The cholecystokinin pathway is involved in muscle contraction and regulated blood pressure (Lovick, 2009; Mikulášková *et al.*, 2016; Dong *et al.*, 2017). The immune system is activated after cardiac tissue injury or stress, caused by toxic chemical exposure, driving acute inflammatory response and regenerative response (Epelman *et al.*, 2015; Carrillo-Salinas *et al.*, 2019). Innate immune cells migrate to the affected site and release mediators such as reactive oxygen species and proteases to remove factors responsible for heart damage. In addition, injured cardiomyocytes release pro-inflammatory cytokines triggering the adaptive immune system (Monda *et al.*, 2020; He *et al.*, 2022). Chemical stimuli detection and sensory perception of smell describe the process by which a chemical stimulus is received by cell and converted into a molecular signal (EMBL-EBI, 2021). In addition, the expression of inaa and tiam2a which are involved in neurons morphogenesis and axogenesis are found to be highly associated with heart rate changes caused by chemical exposure, highlighting the importance of the nervous system in controlling cardiac contractility (Gordan *et al.*, 2015).

The 143 chemicals used in this study are characterised by high variability in structural characteristics, where chemicals with similar structural features, such as functional groups, number and nature of bonds, lipophilicity, electronegativity, and polarizability, may act through different mechanisms and have a diverse effect on the zebrafish embryos. Utilising additional information from gene expression profiles after chemical exposure allows the identification of biological processes and pathways that are responsible for the chemical effect.

3.5.4 Molecular descriptors for predicting Heart rate fold change of cluster 1 and 2 chemicals

Clustering chemicals reduces the variability of the dataset since chemicals with similar gene expression profiles are grouped together, thus the reduction of the dataset (cluster 1= 43 chemicals, cluster 2=34 chemicals) is expected to improve the accuracy and reliability of the predictive models. However, that was not true for the models generated using the gene expression profiles for cluster 1 (R^2 =0.58, p-value=0.25) and cluster 2 chemicals (R^2 =0.6, adjusted p-value=0.353), where both models despite the high R^2 were not representative of the data provided (adjusted p-value). On the other hand, molecular descriptors generated two models that could to some extent predict the phenotypic effect of the chemicals, cluster 1 R^2=0.39, cluster 2 R^2=0.63. The model generated from cluster 1 chemicals identified 35 descriptors that their combination can partially predict heart-rate fold change caused due to exposure to toxins.

A set of 14 descriptors were identified as the ones that contribute the most towards an accurate and reliable prediction associated with molecular mass, electronegativity, ionization potential and presence of tertiary aliphatic amines. Toxicity of various metals has been associated with electrochemical characteristics, such as ionisation potential (MATS4i), and electronegativity (rGes, C-006, C-026, electronegative atoms) (Walker, Enache and Dearden, 2003; Gajewicz-Skretna *et al.*, 2021), based on the assumption that chemical reactivity and toxicity are proportionally related, the most reactive chemical will be the most toxic (Fan *et al.*, 2018). Ionisation potential is the lowest energy needed to remove an electron from a chemical system and has been used to measure the capability of a molecule to lose an electron and is related to electronegativity (Walker *et al.*, 2003). JGI4 is a topological descriptor that provides a measure of the charge transfer between pairs of atoms and the total charge transfer in the molecule (Galvez *et al.*, 1994). Two descriptors were selected that are associated with molecular mass (ATSC2m, GATS1m), where molecular size influences the retention of compounds (Ogadimma *et al.*, 2016).

In addition, descriptors associated with chemical lipophilicity/hydrophobicity (log P) have been widely used to predict the toxicity of inert chemicals. Hydrophobicity is the ability to accumulate organic substances in water, and lipophilicity determines the intermolecular relationships between an organic substance and a solvent, describing the bioavailability, permeability, and toxicity of a drug (Kujawski *et al.*, 2012). LogP can be used for drug distribution estimation and provide information about the ability of a chemical to passively diffuse across biological membranes (Kujawski *et al.*, 2012). The results of this study indicate the importance of P_VSA_LogP_5, which defines the amount of van der Waals surface weighted by logP. Three descriptors were selected that have been associated with the intrinsic state of a chemical (P_VSA_s_3, MATS5s and GATS6s) that describes the availability of the valence electrons for intermolecular interactions that influence the half-life of a chemical (Kujawski *et al.*, 2012; Liu *et al.*, 2017). The intrinsic state has been associated with the atomic contribution to partition coefficient, molecular refractivity, and atomic partial charge (Labute, 2000; Lavado *et al.*, 2020).

Molecular polarizability (VE1sign_B(p)) enhances the non-covalent interaction during molecular transport through the membrane, thus high polarizability indicates high lipophilicity hence associated with acute aquatic toxicity (Gajewicz-Skretna *et al.*, 2021). Chemical toxicity has also been related to chemical size, the number and nature of bonds and the presence of various functional groups (Gajewicz-Skretna *et al.*, 2021). The results indicate that the presence and frequency of nitrogen-oxygen, carbon-nitrogen, oxygen-oxygen and chlorine-chlorine bonds, the number of aromatic primary amides (nArCONH2), are important for predicting the chemical effect on heart-rate fold change. CATS2D descriptors are based on the number of bonds between anions, cations, hydrogen bond acceptors, hydrogen bond donors, and hydrophobic (lipophilic) atoms within the chemical (CATS2D_06_DA, CATS2D_02_AP) (Chang *et al.*, 2013).

The 11 molecular descriptors (R^2= 0.63) identified to be predictive of the heart-rate effect of cluster 2 chemicals, are also related to chemical size (MATS1m),) and the presence of multiple functional groups number of aromatic tertiary amines (nArNR2), carbon-chlorine bonds, nitrogen- chlorine bond. As described above ionisation potential (MATS4i), molecular polarizability (MATS7p) and the charge transfer between pairs of atoms, the total charge-transfer on the molecule (JGI8) and the presence of electronegative atoms (C-040, C-008).

Lastly, the presence of triple bonds can be used as an indicator of the reactivity of the chemical as triple bonds are less stable than single and double bonds (nTB), and the ring fusion density (RFD) that describes the electrophilic nature of the molecule, as it describes the electron mobility of a compound (Mukherjee *et al.*, 2022). The accuracy of this model is relatively high, however, once again the effect of tacrine and chlorpromazine chemicals was not accurately predicted. The effect of those chemicals was better predicted by this model compared to the model generated using the 143 chemicals and the gene expression profiles, suggesting that structural information (molecular descriptors) can be useful in predicting cardiotoxicity when chemicals with similar profiles are used.

As it can be seen from the nature of the molecular descriptors identified by the predictive modelling to be associated with chemical toxicity, molecular mass, ionisation potential, electronegativity, polarizability and lipophilicity are the most important properties for predicting cardiotoxicity in zebrafish embryos. These properties are related to the ability of a chemical to cross the cell membrane and accumulate within the organism and the reactivity of the chemicals that have been widely associated with chemical toxicity.

3.5.5 Gene expression profiles for Predicting Heart rate fold change of cluster 3 chemicals

Finally, the effect of cluster 3 chemicals on zebrafish embryos' heart rate could be predicted only by the mRNA data (R^2=0.64, adjusted p-value= 0.05) and not by the molecular descriptors (R^2=0.17, adjusted p-value= 0.219). Despite the high R^2 of the model generated using the gene expression profiles, and the reduction in dataset size, the model failed to predict the effect of diltiazem, 2-(1-phenylethyl)phenol and o-tolunitrile more accurately than the model generated using the 143 chemicals. This model consists of 21 genes associated with various biological processes, including mitochondria function and ATP production, various developmental processes, signalling pathways and immune system function.

The primary function of mitochondria is energy production and supply by generating ATP through oxidative phosphorylation, to ensure muscle contraction, metabolism, and ion homeostasis (Chen *et al.*, 2010; Nguyen *et al.*, 2019). In addition, mitochondria are also involved in cell death regulation, apoptosis, and necrotic cell death, by responding to stress signals including growth factors, DNA damage, and oxidative stress (Gustafsson *et al.*, 2008). The genes selected by this model are involved in mitochondrial function and energy production, including genes such as eral1 which is active in the mitochondrial matrix and enables rRNA binding activity, ribosomal small subunit binding activity and ribosome biogenesis. The gene abcb8 is part of the mitochondrial ATP-gated potassium channel complex and the 2hgdh gene enables 2-hydroxyglutarate dehydrogenase activity in mitochondria and is involved in energy production. Some of the genes identified are related to carnitine metabolism, used in the transfer of long-chain fatty acids inside the mitochondrial for oxidation in the myocardium and other muscle tissues (Waber *et al.*, 1982; Fu *et al.*, 2013), leading to the generation of adenosine triphosphate (ATP) (Park *et al.*, 2021). Thus, carnitine deficiency may lead to impaired ATP production, reduced ketogenesis, and lipid accumulation in the cytosol (Fu *et al.*, 2013). Heart failure has been associated with defective carnitine transport and fatty acid oxidation in mitochondria (Zhou *et al.*, 2018).

The results of this study indicate that disruption of purine synthesis and metabolism (guk1b, ampd2a), involved in development (Wang, 2016), can be predictive of the effect a chemical has on zebrafish heart rate. Purines are used as metabolic signals, they provide energy and control cell growth (Fumagalli *et al.*, 2017). Nucleotide balance is important for DNA and RNA integrity in replicating cells, and imbalance can induce base substitutions, frameshift mutations, delay of replication form progression and DNA replication and increase the frequency of fragile sites (Weinberg *et al.*, 1981; Copeland *et al.*, 2014; Nogueira *et al.*, 2014; Fasullo *et al.*, 2015). In addition, inhibition of purine synthesis can inhibit cell proliferation through p53-cell cycle arrest (G_0/G_1), which can lead to cytotoxicity (Linke *et al.*, 1996; Quéméneur *et al.*, 2003; Desler *et al.*, 2010). Cytosolic and mitochondrial dNTP pools are sensitive to oxidative stress (Wang, 2016). Oxidative stress results in DNA damage in mitochondria, where nucleotide imbalance can lead to mitochondrial depletion (Fasullo *et al.*, 2015).

Cell communication and signalling pathways are involved in the cellular response to the environment, through cell growth and division, differentiation, migration and apoptosis. Genes associated with extracellular matrix organization (col10a1b) and cytoskeleton-dependent intracellular transport (ccdc88b) are among the genes used by the model. Signalling pathways such as the tyrosine phosphorylation of STAT protein and the JAK/STAT pathway (crlf1a) have been found to be modified after exposure to chemicals that can cause altered heart rate changes in zebrafish embryos. STAT pathways integrate inputs from multiple signalling pathways, and responses to multiple extracellular ligands, such as cytokines and growth factors (Decker, 1999; Levy, 1999; Mui, 1999; Yeh *et al.*, 1999; Imada *et al.*, 2000; Hou *et al.*, 2002; O'Shea *et al.*, 2002). Stat proteins are involved in animal development, growth, cell proliferation, differentiation, survival, immune response, hematopoiesis and migration in normal tissues (Darnell, 1997; Hirano*et al.*, 2000; Yamashita *et al.*, 2002; Liu *et al.*, 2017). In addition, the transcriptional repressor gfi1b is involved in hematopoiesis, cell proliferation, and apoptosis and influences cell fate decisions (Moore *et al.*, 2018).

3.5.6 Conclusion

Cardiac development and function are very sensitive processes and exposure to toxic chemicals during embryonic development may lead to cardiotoxicity in zebrafish embryos. Differential expression analysis revealed that the expression of genes involved in organ development, cardiovascular system development and function, nervous and immune system function, cell-cell communication, and cell death are significantly altered after exposure to chemicals that significantly alter zebrafish embryo's heart rate. These results highlight the various mechanisms through which chemicals can cause cardiotoxicity.

In this study, during the evaluation of the ability of structural features (molecular descriptors) and gene expression profiles to be predictive of heart rate fold change in zebrafish embryos, we found that the use of chemicals with high variability (in structural features and molecular responses) increases the complexity of the models. Thus, as expected, the model generated using molecular descriptors (QSAR) failed to accurately predict the effect of the 143 chemicals, as usually, such models are MoA specific. On the other hand, using the gene expression profiles a model with relatively high accuracy (R^2=0.683) and reliability (adjusted p-value= 0) was generated. The accuracy of this model suggests that gene expression data can potentially be used in risk assessment for assessing chemical-induced cardiotoxicity, of chemicals from multiple MoA classes with a single model.

Clustering chemicals, into three clusters, based on gene count profiles reduces the variability of the data since chemicals with similar profiles are grouped together and is expected to improve model performance, especially of the QSAR models. However, even after clustering only one out of the three models generated using molecular descriptors was accurate and reliable (R^2= 0.63, adjusted p-value= 0.006, cluster 2). In addition, the models generated using gene expression data for the three clusters were less accurate compared to the full dataset model (Cluster 1: R^2= 0.58, Cluster 2: R^2=0.6, Cluster 3: R^2=0.64), and only one of them can be characterised as reliable (Cluster 1: adjusted p-value = 0.25, Cluster 2: adjusted p-value =0.35, Cluster 3: adjusted p-value =0.05). Clustering chemicals in this study did not improve the performance of the predictive models. However, the number of chemicals used in this study is relatively small (43, 34, 66), thus increasing the number of chemicals may improve the predictive power of those models.

Chapter 4

Evaluating the basal toxicity mechanisms in Danio rerio

4.1 Abstract

Exposure to various toxins influences gene expression and physiology of zebrafish embryos. Sequencing approaches generate high-dimensional datasets containing a breadth of molecular functions and biological processes. To improve interpretability, dimensional reduction techniques, such as PCA, have been widely used. In this study, molecular pathways were represented by principal components and used as inputs to predictive modelling approaches to identify potential linkages between exposure, molecular response and chemical toxicity, or chemical effect on zebrafish heart rate. This effectively generates AOPs based on a high-level understanding of the underlying biology. When the data are characterised by high heterogeneity (highly diverse chemicals in gene count profiles), no pathway activity was able to predict the calculated LC50 or the chemical concentration during exposure. On the other hand, most of the pathways described in this study were found to be predictive of heart-rate fold change, indicating that a variety of mechanisms are involved in proper heart development and function. Splitting chemicals based on gene count profiles revealed a set of pathways whose activity was found to be predictive of LC50 and chemical concentration during exposure.

This highlighted several links between exposure and AOs where pathways involved in transport and catabolism, replication and repair, cell growth and death, signalling pathways, sucrose metabolism, and vascular smooth muscle contraction were predictive of chemical LC50 or the experimental concentration, associated with chemical toxicity. Pathways involved in amino acid and nucleotide metabolism, signalling pathways, cellular growth and death, endocrine-related pathways, cell motility, and immune system pathways were associated with heart rate fold change. By considering shared genes between the pathways a network was established highlighting the distances between the pathways associated with chemicals and pathways associated with outcomes. From here several examples have been extracted which highlighted potential new adverse outcome pathways. Incidentally, the

analysis also showed that the selected pathways were connected closer than expected by random chance.

4.2 Introduction

Zebrafish (Danio rerio) is a small tropical fish, with rapid development, a short reproduction cycle and a large number of offspring. They are very easy to maintain in the laboratory due to their small size and ex-utero embryonic development, reducing housing space and husbandry costs. Zebrafish embryos are transparent, enabling fast evaluation of the developmental process and identification of genes related to development (Kimmel *et al.*, 1995). Exposure of Zebrafish to contaminants affects gene expression, physiology and behaviour, due to zebrafish's high sensitivity to environmental changes (Dai *et al.*, 2014). Identifying the toxicity profile of a new chemical requires long experiments and a large number of experimental animals. The ability to predict and understand toxicity mechanisms will consequently reduce the need for animal testing, cost, and duration for chemical assessment.

Signature-matching approaches are based on the assumption that compounds with similar gene expression profiles will have the same effect in a biological system. This allows the comparison of compound-induced gene expression profiles to identify the toxicity of a new compound using chemicals with known toxicity profiles (Alexander-Dann *et al.*, 2018). However, such lists do not provide information related to the underlying biological mechanism, since gene data usually cannot explain an entire functional trait. In addition, in some cases, compounds do not significantly influence gene expression, leading to transcriptional signals dominated by noise that does not represent the effect of the chemicals on the organism (Ramanan *et al.*, 2012).

High throughput data suffer from high dimensionality since usually, the number of variables is greater than the number of samples. These characteristics may lead to model overfitting and false correlation that will affect the accuracy and the computational complexity of modelling (Clarke *et al.*, 2008). These limitations raise the need for a method that simplifies the analysis by dimensionality reduction. Such strategies aim to reduce storage space and computation time, remove noise, and redundant and correlated features. It also allows easier visualisation of data and pattern identification. Variable reduction methods lower the accuracy of the data due to some information loss, but the simplicity offered by the smaller datasets allows easier visualisation and exploration of the data and increases the speed of the analysis.

There are two main types of dimensionality reduction methods, feature selection, selecting a subset of features as an input for further analysis, and feature extraction where using all the available variables a new feature space is generated (such as principal component analysis, PCA). Feature selection, such as grouping the list of genes into clusters of functionally related gene sets (pathways) has the potential to cope with such limitations. When using RNA-sequencing data, grouping genes into pathways reduces dataset dimensionality and allows the differentiation between response and noisy data (Alexander-Dann *et al.*, 2018) and improves the interpretation of the results and hypothesis generation (García-Campos *et al.*, 2015).

Pathway analysis is based on existing biological knowledge, quantitative data such as RNA-seq, statistical testing, mathematical analyses, and computational algorithms; they are used to connect ‘omics data and existing knowledge (Antczak *et al.*, 2015; García-Campos *et al.*, 2015). The use of functionally derived gene sets can reveal larger effects compared to gene-based analysis (Wang *et al.*, 2010; Zhong *et al.*, 2010). Pathway analysis has been used for the identification of the biological role of candidate genes and facilitating the understanding of the underlying mechanism (García-Campos *et al.*, 2015). However, the lack of guidelines increases variability between pathway studies (Ramanan *et al.*, 2012).

To assist with risk assessment, the AOP concept has been proposed, which organises existing knowledge between biological changes (key events) from an MIE to an AO. An AOP begins with the interaction of a chemical with a biological target (MIE), which is followed by cellular responses, leading to an organ response, individual phenotype, and population response (AO) (Ankley *et al.*,2010). An AOP is considered active when data shows that all the key events that lead to an AO are covered (Serra *et al.*, 2020). However, collecting experimental data to prove the activation of all identified key events is time-consuming, thus *in silico* approaches can be used to predict the downstream or upstream key events, and possible key events relationships. The activation of molecular mechanisms can be predictive of a phenotypic effect, thus using toxicogenomic data can provide the necessary information spanning through the multiple key events in an AOP (Vinken, 2019). AOPs reflect the current state of knowledge, thus they can continue to evolve as new information becomes available. Computational methods have been applied in identifying AOPs, by linking transcriptomics and structural information to an AO; thus structural characteristics can be used to predict pathway activity which can be used to predict adverse outcomes (Antczak *et al.*,2015).

In this study, the aim is to utilise pathway-based information to uncover the mechanism of how chemicals impact the molecular state of zebrafish embryos. Over 140 chemicals were used to derive pathway-level information and aimed to establish an AOP-based network that can be used to predict AOs such as heart rate fold changes.

4.3 Methods

4.3.1 Selecting transcriptional profiles for pathway analysis

The same 143 chemicals used for the analysis in chapters 2 and 3 were also used in this chapter for further analysis. In addition, the 143 chemicals were grouped together into three clusters (43, 34 and 66 chemicals) based on the gene expression profiles as described in chapter 2 (mRNA hard clustering), thus a total of 4 datasets were used for further analysis in this chapter. The gene expression profiles of the highest available concentration from the mRNA-seq data, for each chemical (124 chemicals- LC5, 15 chemicals- LC5/2, 4 chemicals- LC5/4, as described in chapter 2), were used for pathway analysis. The calculated heart rate fold change and the experimental concentration (log format) (124 chemicals- LC5, 15 chemicals- LC5/2, 4 chemicals- LC5/4) were used as the dependent variable in the predictive modelling.

4.3.2 Modelling of toxicity using the General Unified Threshold model and dose-response curves

To derive a more regulatory applicable model the General Unified Threshold Model (GUTS) was used to define the concentration at which 50% of the population is dead (LC50) of each chemical (Jager *et al.*, 2011). The openGUTS, a free and open-source software, calculate the LC50 using toxicokinetic-toxicodynamic models to estimate survival. The LC50 is negatively correlated with toxicity, where low LC50 values indicate that chemical exposure will kill zebrafish even at low concentrations since the chemical is toxic. The more traditional and popular method for calculating the LC50 (toxicity) in the public domain of chemicals is by using dose-response models like the log-logistic model. In this study the drc package in R (version 3.0.1) was used, combining the four-parameter log-logistic function (LL.4) with the drm function for fitting function for dose-response analysis (Ritz *et al.*, 2015). The GUTS and the dose-response model predictions were compared. The LC50 of each chemical calculated by GUTS were also used as the dependent variable in predictive modelling.

4.3.3 Data dimensionality reduction and predictive modelling

To establish pathway representations of the gene expression data the Kyoto Encyclopaedia of Genes and Genomes (KEGG) was used. The KEGG database currently contains more than 500 pathways and covers 756 eukaryotes, 7011 bacteria and 389 archaea. The gene expression data were generated using the Ensembl id for each gene and consisted of more than 30 thousand genes. However, to be able to split the data based on the KEGG pathway, the biomaRT R package (version 2.42.1) was used to convert the ensemble id into the Entrez gene id (NCBI gene id) using the function getBM. The NCBI gene id was then used on the KEGG website (*KEGG Mapper – Convert ID*, 2021) to get the equivalent KEGG id for each gene. The keggGet function from the R package KEGGREST (version 1.26.1) was used to map the genes to the available pathways, and only the pathways represented in the dataset by more than five genes were selected for further analysis. The genes used in this study were found to be involved in 160 pathways with at least 5 genes per pathway. Each of the 4 datasets was then split into these pathways. To summarise the expression of each pathway principal component analysis (PCA), an unsupervised learning method was then performed. Here the gene expression data are linearly transformed into new, uncorrelated observations (principal components -PC) that account for decreasing proportion of the total variance in gene expression, the first PC explains most of the variation. PCA was performed in R using the function prcomp (from the stats package version 3.6.1), which uses a singular value decomposition of the covariance and correlation between variables (Jolliffe *et al.*, 2016), using the parameters center=TRUE, indicating that the variables should be shifter to be zero centred, and scale=TRUE, indicating that variables should be scaled to have unit variance before the analysis. After the PCA only the PCs that cumulatively explain up to 80% of the variance between the gene expression profiles were selected to represent pathway activity, in each of the 640 datasets.

Finally, the predictive capability of each pathway towards heart rate fold change, chemical exposure concentration, and the LC50 calculated by GUTS was estimated using the same predictive modelling function described in chapter 2. The predictive function split the dataset into training and test sets using the Caret R package (Kuhn, 2020), and LASSO regression-stability selection to select and sort features based on importance (parameter family= "gaussian" (Sill *et al.*, 2014). The selected features were added sequentially to the model (forward selection) and using random forest regression (ranger) (Wright *et al.*, 2017) (classification =FALSE), the model that on average performed best across all the splits, was selected. For model validation, the same method was repeated 1000 times, while randomising the dependent variable every time, and the p-value (the probability of the model to be generated even with random data), was calculated using the empirical cumulative distribution function.

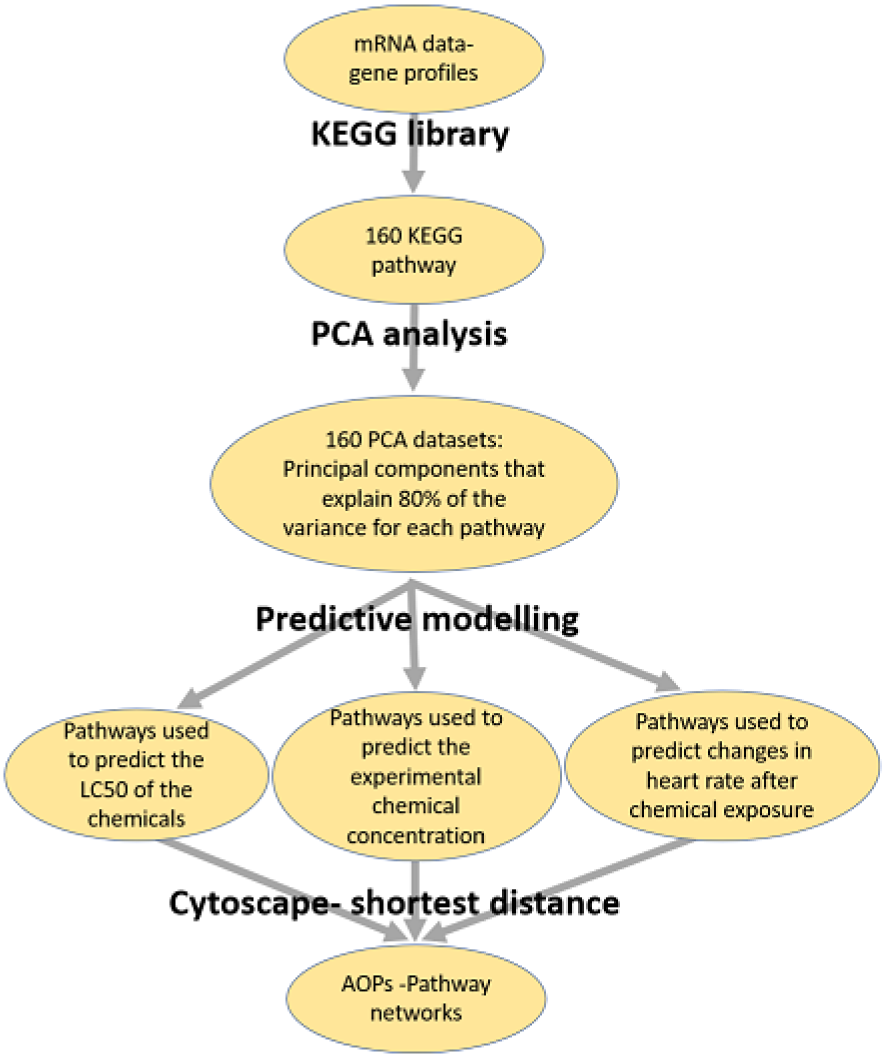

mRNA data- gene profiles
KEGG library
160 KEGG pathway
PCA analysis
160 PCA datasets: Principal components that explain 80% of the variance for each pathway
Predictive modelling
Pathways used to predict the LC50 of the chemicals
Pathways used to predict the experimental chemical concentration
Pathways used to predict changes in heart rate after chemical exposure
Cytoscape- shortest distance
AOPs -Pathway networks

Figure 4.1: Representation of the workflow related to the generation of putative AOPs using mRNA data. Using the KEGG library, the gene profiles were split into 160 pathways. PCA analysis was performed on the gene profiles for each pathway. For predictive modelling, the principal components that cumulatively explain up to 80% of the variance were used as the independent variables in predictive modelling for each pathway. The pathways activity that can be used to predict chemicals LC50, experimental chemical concentration and changes in heart rate fold change were identified and used in the construction of AOP networks using the number of shared genes between pathways and Cytoscape software.

The predictive modelling function was applied 1920 times (160 X4 X3). In this study 4 datasets were used that consisted of different numbers of chemicals, the larger one consisted of 143 chemicals, and the other 3 (43,34,66 chemicals) are the result of the chemical clustering based on gene expression profiles, describe in chapter 2. Each of those 4 datasets consists of more than 30 thousand genes, thus after pathway analysis, each dataset was split into 160 smaller datasets (4X160 = 640). Three dependent variables were used, the calculated heart rate fold change (described in chapter 2), the experimental chemical concentration after log transformation (124 chemicals- LC5, 15 chemicals- LC5/2, 4 chemicals- LC5/4), and the calculated LC50 using GUTS (4 X 640=1920).

To evaluate how the pathways identified to be predictive of the three dependent variables (heart rate fold change, experimental chemical concentration in log format, and chemicals LC50) are connected, a network was developed between all the available pathways. The connections between the selected pathways were defined by the presence of shared genes based on the KEGG database, using the jaccardSets function, part of the R package bayesbio (version 1.0.0) and the data downloaded from the KEGG database that describes the genes involved in each pathway. The Jaccard index was calculated for any two pathways, which describes the number of intersections (i.e. shared genes between each pair of pathways) divided by the total number of elements of the two pathways. The Jaccard index was then used to develop a network of pathways. To explore how the pathways that can be used to predict chemical-induced heart rate changes with the pathways that can predict chemical concentration (experimental exposure- log format) or chemical LC50, the shortest distance was calculated between those pathways using the R package igraph (version 1.2.6) (Csardi et al., 2006). The resulting linkages were extracted and visualised using Cytoscape, an open-source software network visualisation tool (Shannon *et al.*, 2003), and the KEGGScape extension (Nishida *et al.*, 2014).

4.4 Results

4.4.1 GUTS Modelling of toxicity

Modelling toxicity from observational data can be achieved through more traditional dose-response models or by utilising the toxicokinetics and toxicodynamics (TKTD) inspired modelling approaches. Here the GUTS model has established itself as an important regulatory approach to modelling endpoint survival (Jager *et al.*, 2011). The external concentrations are translated to internal concentrations and associated with the likely damage which triggers death. This, therefore, ensures that the concentrations leading to 50% of death (LC50) are directly based on the likely internal concentrations. At the same time dose-response modelling, a log-logistic model, one of the most popular methods for predicting the LC50 of each chemical was applied. The results show that the outcome of those two models is similar with very few outliers (R^2= 0.89) (Figure 4.2).

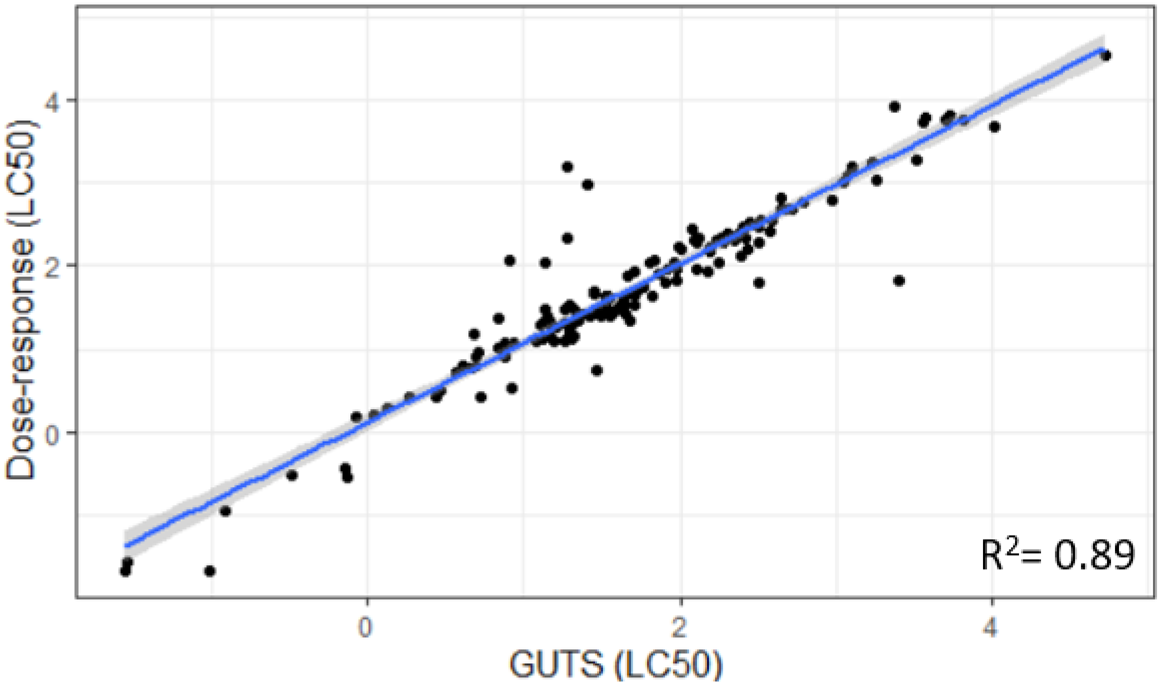


Figure 4.2: Comparing the estimated LC50 of each chemical using dose-response log-logistic modelling and GUTS (R^2= 0.89).

4.4.2 Predicting LC50 of chemicals using pathway activity

Based on the gene expression data, pathway activity information can be obtained using pathways originating from the KEGG database. Using pathway activity information (160 pathways) in the form of principal components as the independent variables, the LC50 (GUTS) of each chemical was predicted. When the whole dataset was used, no pathway information was able to predict the LC50 accurately (highest R^2 = 0.12).

The dataset was then split based on mRNA (gene count profiles) clustering described in chapter 2. The first cluster consists of 43 chemicals, and predictive modelling identified two pathways that can to some extent used to predict the LC50s of these chemicals, lysosomes activity ($R^{\wedge 2}$ = 0.342) and VEGF signalling pathways ($R^{\wedge 2}$ = 0.202) (Table 4.1). Lysosomes are involved in intracellular macromolecule degradation, such as lipids and damaged organelles, where the end products can be used for energy or as building blocks for other macromolecules (Cooper, 2000; Pei *et al.*, 2021). On the other hand, the VEGF signalling pathway is associated with vasculogenesis and angiogenesis, by controlling the expression of genes involved in vascular permeability and promotion of cell migration, proliferation and survival (Apte *et al.*, 2019).

To predict the LC50 of the 34 chemicals from mRNA cluster 2, 10 pathways were selected (Table 4.1). Some of these pathways are associated with lipid metabolism, glycerophospholipid metabolism (R^2 = 0.33), where glycerophospholipids are the main structural components of biological membranes and are involved in signal induction and transport (Hishikawa *et al.*, 2014) and steroid hormone biosynthesis (R^2 = 0.28), that are important in growth, development, sexual differentiation and reproduction (Adhya *et al.*, 2018). The non-homologous end-joining pathway (R^2 = 0.3), responsible for repairing double-strand breaks in DNA throughout the cell cycle were identified by the model (Chang *et al.*, 2017). Pathways involved in environmental information processing such as the ECM-receptor interaction pathway (R^2 = 0.260) that regulates cell behaviour and cell shape, growth, survival and differentiation (Lukashev *et al.*, 1998), Adherens junctions (R^2 = 0.266), cell-cell adhesion complexes involves in tissue architecture maintaining under external stress and contribute to embryogenesis and tissue homeostasis (Harris *et al.*, 2010) and Gap junctions (R^2 = 0.259) that are clusters of intercellular channels that allow ion and small molecules diffusion between adjacent cells (Goodenough *et al.*, 2009) can be used to predict chemical LC50. Cellular senescence (R^2 = 0.304), the cell-cycle arrest that prevents the proliferation of damaged cells (Huang *et al.*, 2022), the Intestinal immune network for IgA production (R^2 = 0.250), a noninflammatory immunoglobulin antibody used in the defence against microorganisms (Gutzeit *et al.*, 2014), and GnRH signalling pathway (R^2 = 0.274) that regulates the production and release of the gonadotropins hormones (LH, FSH), controlling the reproductive system, and are involved in cell proliferation inhibition (Kraus *et al.*, 2001). Finally, the last pathway whose activity is predictive of the LC50 of the chemical is the Herpes simplex virus 1 infection (R^2 = 0.252) which links to apoptosis and multiple signalling components (*KEGG PATHWAY: dre05168*, 2021). For the chemicals within mRNA cluster 3, only one pathway was selected, starch and sucrose metabolism (R^2 = 0.237), which play an important role in development and stress response by generating sugars for growth and essential compounds synthesis and signals that regulate microRNA expression, transcription factors and genes involved in defence signalling (Ruan, 2014) (Table 4.1).

Predicting LC50 of mRNA cluster 1 chemicals		
KEGG Pathways	**R^2**	**P value**
Lysosome	0.34	0.0005
VEGF signalling pathway	0.2	0.02
Predicting LC50 of mRNA cluster 2 chemicals		
Glycerophospholipid metabolism	0.33	0.008
Steroid hormone biosynthesis	0.28	0.02
Non-homologous end-joining pathway	0.3	0.006
ECM- receptor interaction pathway	0.26	0.032
Cellular senescence	0.3	0.013
Adherens junction	0.266	0.028
Gap junction	0.26	0.032
Intestinal immune network for IgA production	0.25	0.04
GnRH signalling pathway	0.27	0.02
Herpes simplex virus 1 infection	0.25	0.037
Predicting LC50 of mRNA cluster 3 chemicals		
Starch and sucrose metabolism	0.24	0.0004

Table 4.1: A list of pathways whose activity is predictive of GUTS-LC50 of each chemical using three datasets, chemicals form mRNA cluster 1, 2 and 3 respectively.

4.4.3 Predicting chemical concentration using pathway activity

After predicting the estimated LC50 of each chemical, the experimental concentration (log) used in this study was predicted (124 chemicals- LC5, 15 chemicals- LC5/2, 4 chemicals- LC5/4). As with the LC50 data, using the whole dataset no pathways were identified to be predictive of the experimental concentration. Splitting the data into mRNA clusters increases the predictive power of the generated models. Two pathways were identified to be predictive of the chemical concentration when 43 chemicals were used (mRNA cluster 1), lysosome (R^2=0. 282), also identified to be predictive of LC50 and selenocompound metabolism (R^2=0. 205), that are essential immunonutrients as they have anti-inflammatory and antioxidant properties and are essential components for multiple enzymes activities (Hariharan *et al.*, 2020) (Table 4.2).

Predicting experimental concentration (log) of mRNA cluster 1 chemical		
KEGG Pathways	**R^2**	**P value**
Lysosome	0.28	0.004
Selenocompound metabolism	0.2	0.037
Predicting experimental concentration (log) of mRNA cluster 2 chemicals		
Amino sugar and nucleotide sugar metabolism	0.3	0.02
Glycolysis/Gluconeogenesis	0.3	0.025
Steroid biosynthesis	0.31	0.012
Fatty acid elongation	0.3	0.017
Alpha-Linolenic acid metabolism	0.26	0.04
Various types of N-glycan biosynthesis	0.29	0.03
Pantothenate and CoA biosynthesis	0.27	0.05
Metabolism of xenobiotics by cytochrome	0.27	0.03
Ribosome biogenesis in eukaryote	0.3	0.025
Non-homologous end-joining	0.3	0.008
Apeling signalling pathway	0.31	0.016
Cellular senescence	0.34	0.008
Vascular smooth muscle contraction	0.26	0.04
Predicting experimental concentration (log) of mRNA cluster 3 chemicals		
Vascular smooth muscle contraction	0.26	0.0005

Table 4.2: A list of pathways whose activity is predictive of experimental concentration (log) of each chemical using three datasets, chemicals form mRNA cluster 1, 2 and 3 respectively.

The models generated with chemicals from mRNA cluster 2, indicate that the activity of 13 pathways is predictive of the chemical concentration (Table 4.2). Three pathways were selected that are associated with DNA repair, translation and post-translational modifications, Ribosome biogenesis in eukaryotes (R^2 = 0.303), non-homologous end-joining (R^2 = 0.299) and post-translational modification N-glycan biosynthesis (R^2 = 0.289) (Toustou *et al.*, 2022). Pathways involved in carbohydrate metabolism, amino sugar and nucleotide sugar metabolism (R^2 = 0.289) and glycolysis/gluconeogenesis pathways (R^2 = 0.285) (Brosnan, 1999), lipid metabolism, including steroid biosynthesis (R^2 = 0.315), fatty acid elongation ($R^2 = 0.304$) and alpha-Linolenic acid metabolism ($R^2 = 0.257$), and Pantothenate and CoA biosynthesis ($R^2 = 0.267$) (cofactor metabolism) (Leonardi *et al.*, 2007) ensure a constant supply of energy through the generation of ATP. Pathways involved in the immune system such as the metabolism of xenobiotics by cytochrome P450 enzyme ($R^2 = 0.274$) through iron oxidation, including drug metabolism (Stavropoulou *et al.*, 2018), and the Apelin signalling pathway ($R^2 = 0.309$) which is involved in angiogenesis, heart muscle contractility, energy metabolism and homeostasis (Chapman *et al.*, 2014), were identified to be predictive of chemical concentration. Finally, cellular senescence ($R^2 = 0.339$) and the vascular smooth muscle contraction pathway ($R^2 = 0.267$) which is controlled by the increase of the intracellular calcium ions (Ca^{2+}) and upon contraction, the vascular smooth muscle is shortened, decreasing the diameter of blood vessels regulating the blood flow and pressure (Ets *et al.*, 2016), was found to some extent predict the experimental concentration. Predictive modelling with the datasets that consist of only the chemicals from cluster 3 indicated that only one pathway can predict the chemical concentration, Vascular smooth muscle contraction ($R^2 = 0.257$) (Table 4.2).

4.4.4 Predicting Heart rate fold change of chemical using pathway activity

Out of the 160 Kegg pathways, 137 were found to be predictive of the heart-rate fold change after chemical exposure across the whole dataset. In an effort to select the most important pathway, only the ones with R^2 higher than 0.35 are described here (41 pathways). As it was expected, the activity of pathways involved in cardiac muscle and vascular smooth muscle contraction ($R^2 = 0.37\text{-}0.4$) and pathways involved in carbohydrate metabolism, lipid metabolism, purine metabolism, and amino acid metabolism, were found to be predictive of heart-rate changes. These pathways ensure a constant supply of energy through the generation of ATP and proper cardiovascular function. Predictive modelling identified the importance of the immune system in cardiovascular function, where pathways involved in recognizing and responding to pathogenic microorganisms and non-self-components were found to some extent predictive of changes in heart rate after chemical exposure ($R^2 = 0.36\text{-}0.4$). Two pathways were also identified by the predictive modelling function, associated with cell death, cellular senescence ($R^2 = 0.42$) and apoptosis ($R^2 = 0.41$), that are activated by cellular stress. Pathways involved in the transmission of regulatory signals between the extracellular matrix and an interacting cell ($R^2 = 0.35\text{-}0.39$)(Table 4.3).

Predicting heart rate of 143 chemicals	
KEGG pathways	**R^2 range**
Cardiovascular muscle contraction	0.37-0.4
Metabolism	0.35-0.43
Immune system	0.36-0.4
Cell death	0.41-0.42
Cellular community	0.35-0.4
Predicting heart rate of mRNA cluster 1 chemicals	
Metabolism	0.21-0.32
Immune system	0.21-0.31
Environmental information processing	0.24-0.26
Cellular processes	0.2-0.23
Infectious disease	0.23-0.27
Predicting heart rate of mRNA cluster 2 chemicals	
Metabolism	0.25-0.375
MAPK signalling	0.38
Cellular processes	0.26-0.264
Organismal system	0.3-0.34
Predicting heart rate of mRNA cluster 3 chemicals	
Metabolism	0.2-0.29
Environmental information processing	0.22-0.31
Cellular processes	0.2-0.27
Organismal system	0.22-0.36
Infectious disease	0.22

Table 4.3: A list of pathways whose activity is predictive of the heart-rate fold change of each chemical using four datasets, chemicals from the whole dataset, mRNA cluster 1, 2 and 3 respectively.

While it is possible to predict heart rate fold change across the whole dataset, for direct comparison with the toxicity-based prediction performed earlier, clustered data were also used in predicting heart rate fold change. The heart rate fold change of the chemicals from mRNA cluster 1, could be predicted by the activity of 20 pathways (Table 4.3). Ten of the pathways selected are associated with metabolism, lipid and Amino acid metabolism, glycan biosynthesis and metabolism, the metabolism of cofactors and vitamins (R^2 = 0.21-0.32) and three pathways involved in the immune system (R^2 =0.21-0.31). Two of the pathways were related to environmental information processing, cytokine-cytokine receptor interaction, and cell adhesion molecules (R^2 =0.24-0.26). Three of the pathways are involved in cellular processes, autophagy, necroptosis and regulation of actin cytoskeleton (R^2 =0.2-0.23). Finally, the activities of pathways associated with viral and bacterial infectious diseases were also able to predict heart-rate fold change (R^2 =0.23-0.27).

The heart-rate fold change of the chemical form cluster 2, could be predicted by 18 pathways (Table 4.3). 13 of those pathways are associated with lipid, nucleotide, amino acid, cofactors and vitamin metabolism, and glycan biosynthesis and metabolism (R^2 = 0.25-0.375). Environmental information processing (MAPK pathway) (R^2 = 0.38), p53 signalling (R^2 = 0.26), regulation of the actin cytoskeleton (R^2 = 0.26), the intestinal immune network for IgA production (R^2 = 0.3) and the adipocytokine signalling pathway (R^2 = 0.34), were selected after predictive modelling.

The mRNA cluster 3 which consists of 66 chemicals, could be predicted by 32 pathways (Table 4.3). 15 pathways were associated with carbohydrate, energy, lipid, nucleotide and amino acid metabolism and glycan biosynthesis and metabolism (R^2 = 0.20-0.29). Five pathways are associated with environmental information processing, signal transduction and signalling molecules and interaction (R^2 = 0.22-0.31). Seven pathways were associated with cellular processes, transport and catabolism, cell growth and death, and cellular community-eukaryotes (R^2 = 0.20-0.27). Five pathways were associated with organismal systems, the immune system, endocrine system and circulatory system (R^2 = 0.21-0.36). Finally, the salmonella infection pathway activity was also able to some extent predict heart-rate fold change (R^2 = 0.22).

4.4.5 Shortest distance between pathways

After identifying the pathways that are associated with the chemical LC50 (calculated by GUTS), experimental concentration and heart-rate fold change, the distance between the pathways was investigated. KEGG pathways represent a molecular system, where all the pathways are connected with each other within a larger metabolism-inspired network. First, the shortest paths between all pathways (the path that connects the pathways in question and consists of the least number of pathways) used in this study were identified. This showed that the median distance between any two pathways is four, highlighting that although biological systems are closely related to each other, many smaller clusters exist that drive the response (Figure 4.3).

The pathways associated with LC50 or chemical concentration could be considered to be closer to the potential MIE, thus the distance between those pathways from those that are associated with an adverse outcome (heart rate fold change) highlights the distance the signal must travel to achieve its effect. As it is expected the average distance between LC50 or chemical concentration (MIE) and heart rate (AO), was smaller than the average distance observed between the whole dataset, where the median is reduced to 2, suggesting that only 2 connections were necessary to direct the signal from MIE to AO (Figure 4.3).

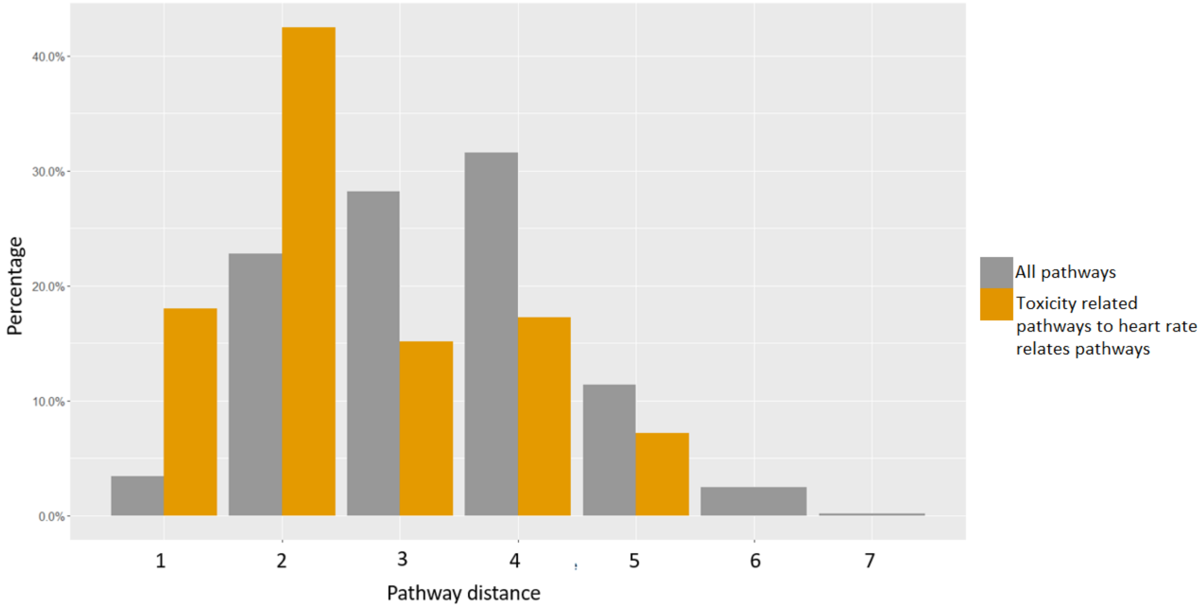


Figure 4.3: The distribution of the shortest distance between all the pathways in the dataset (grey) and the distribution of the shortest path between the pathways that can predict chemical concentration (LC50 or experimental concentration) and heart-rate fold change (orange).

To explore how the pathways that are found to be predictive of GUTS calculated LC50 and the chemical concentration used in this study, are connected, the shortest distance between them is also calculated. Figure 4.4 shows clearly that the pathways that are found close to the MIE (pathways predictive of chemical LC50 and the concentration of the chemical) share a significant number of genes as the pathways are directly connected (1 step) or there is only 1 pathway in between (2 steps). This highlights that although this data is based on a highly heterogeneous set of compounds the effect or the MIE that drive the response seem limited to a few higher-order biological functions. This is in line with the concept that pathways close to extracellular regions are more likely to contain MIEs than any other available pathway, such as various signalling pathways, Including the VEGF signalling pathway, apelin signalling pathway, extracellular matrix receptor interaction pathways and cellular community.

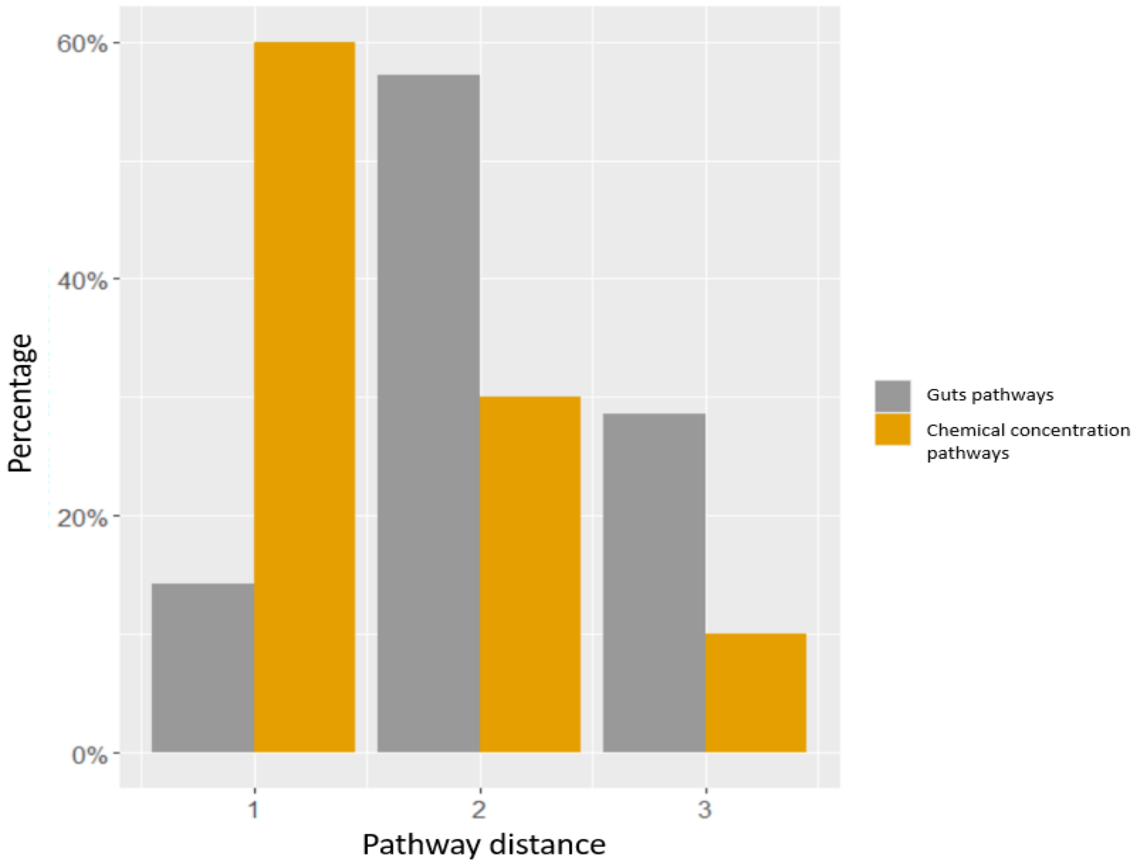


Figure 4.4: The distribution of the shortest path between pathways that can predict Guts- LC50 (grey) and chemical concentration (orange).

4.4.6 Pathway network

To identify how the pathways found to be predictive of LC50, experimental concentration, and heart rate fold change interact with each other, a network was developed that highlights the shortest distances identified in the previous step. Cluster-specific networks were developed to highlight the potentially different mechanisms represented by each.

Cluster 1 chemicals reveal the importance of lysosome activity in predicting LC50 and chemical concentration. Downstream of the lysosome pathway, eight pathways were identified, whose activity was predictive of change in heart rate. Four of these were linked directly to Lysosome while the other four used a single proxy pathway. These pathways represented lipid metabolism, biosynthesis of factors, and signalling pathways. Additionally, the approach identified salmonella infection to be predictive of heart rate. This pathway contains several of the other pathways identified to be directly associated with the lysosome pathway such as the NOD-like receptor signalling pathway. On the other hand, both VEGF signalling pathway activity that has been associated with LC50 and selenocompound metabolism pathway activity associated with chemical concentration were directly linked to metabolic pathways activity predictive of heart rate changes in zebrafish embryos. The metabolic pathways are further associated with amino acid metabolism, cell death, signalling pathways and immune system activation (Herpes simplex virus 1 infection) (Figure 4.5).

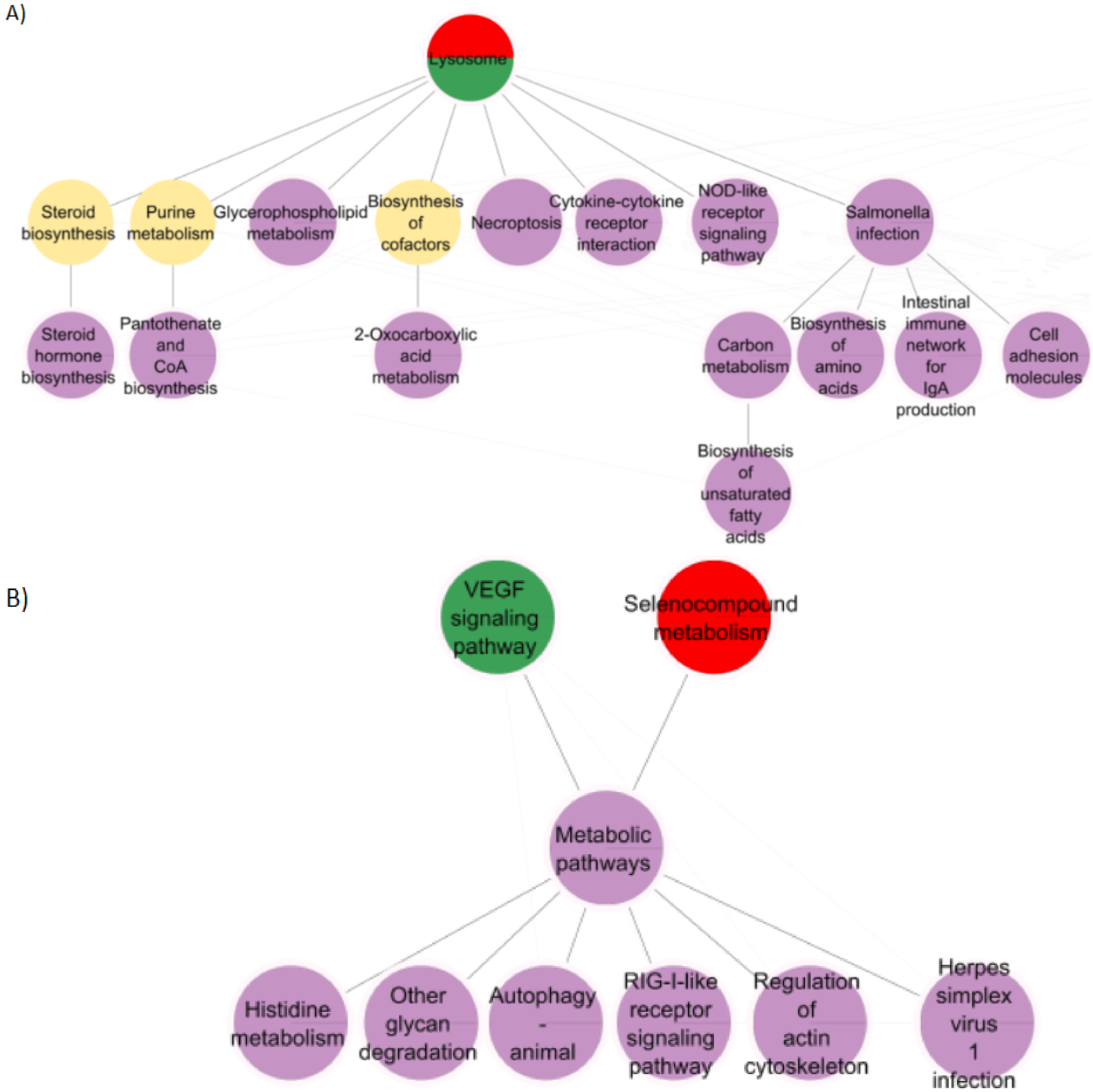
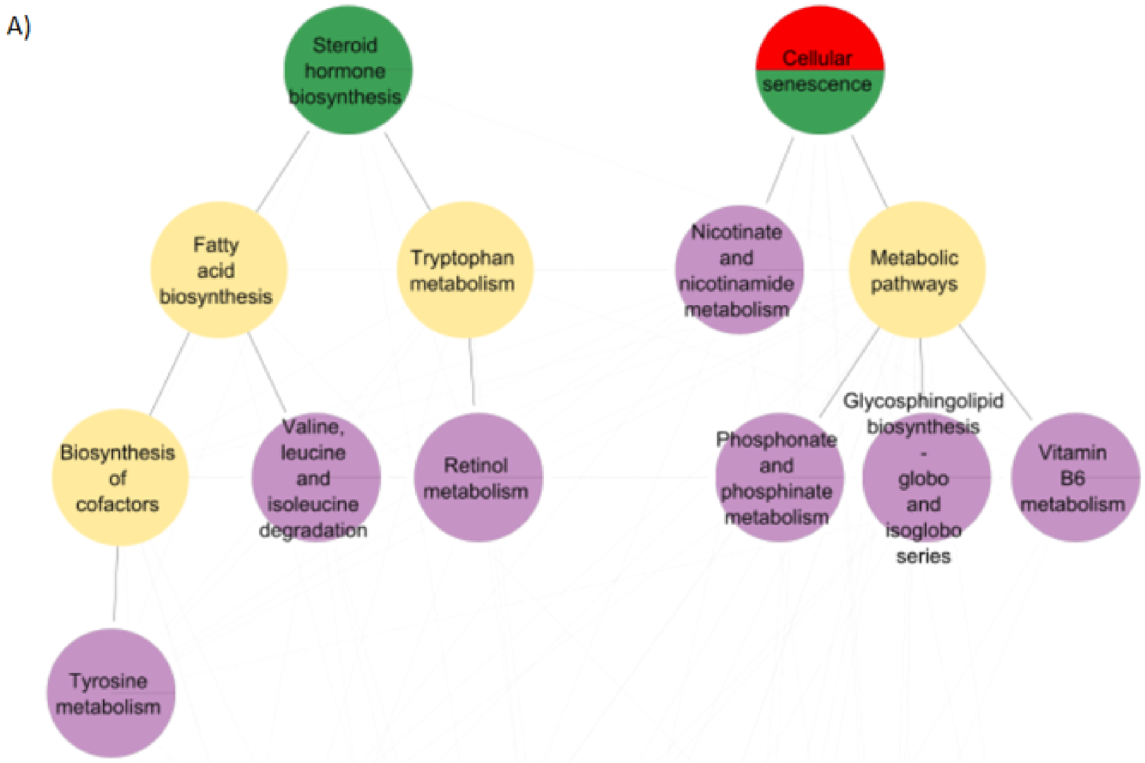
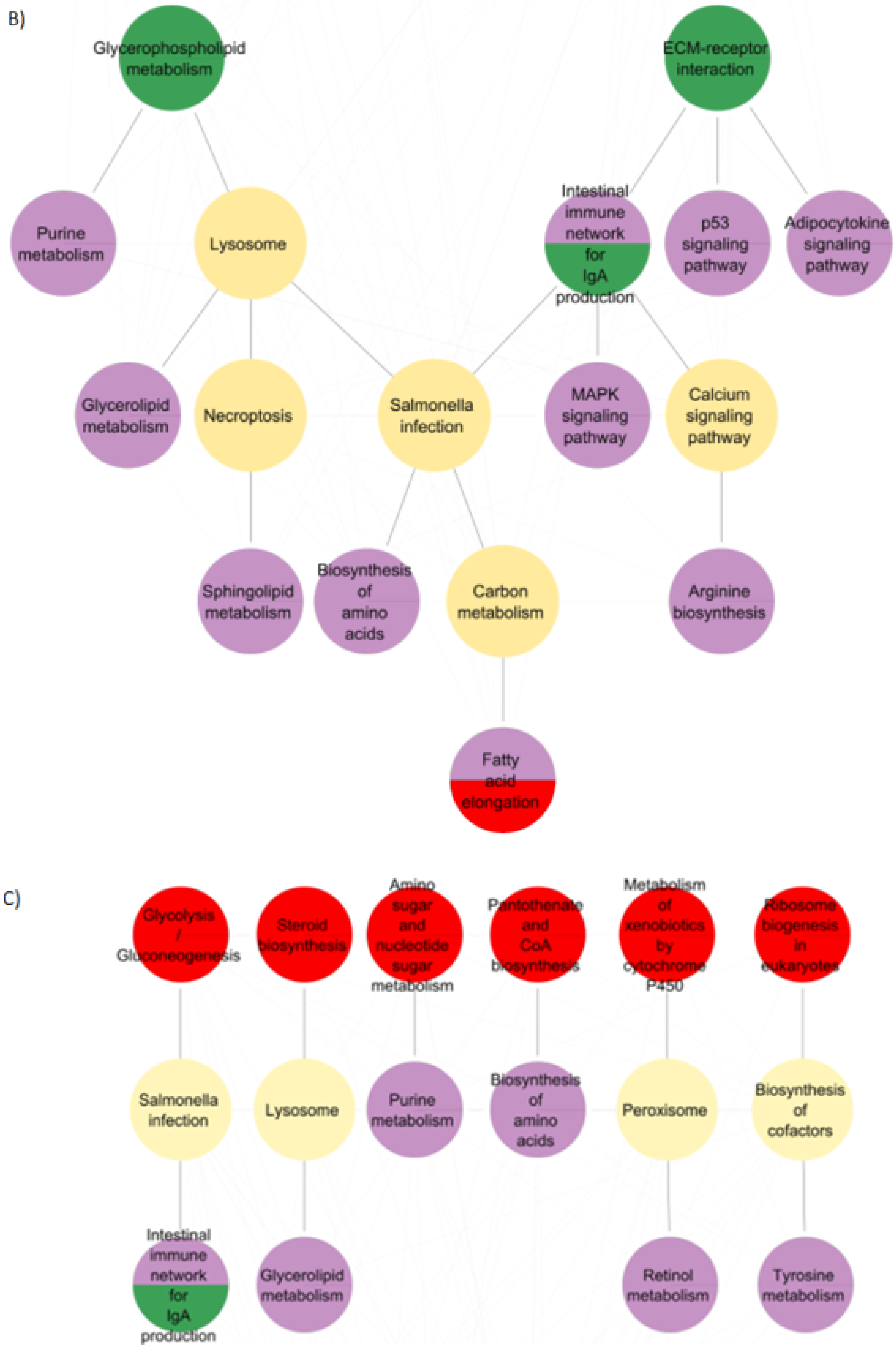

A)
Lysosome
Steroid biosynthesis
Purine metabolism
Glycerophospholipid metabolism
Biosynthesis of cofactors
Necroptosis
Cytokine-cytokine receptor interaction
NOD-like receptor signaling pathway
Salmonella infection
Steroid hormone biosynthesis
Pantothenate and CoA biosynthesis
2-Oxocarboxylic acid metabolism
Carbon metabolism
Biosynthesis of amino acids
Intestinal immune network for IgA production
Cell adhesion molecules
Biosynthesis of unsaturated fatty acids
B)
VEGF signaling pathway
Selenocompound metabolism
Metabolic pathways
Histidine metabolism
Other glycan degradation
Autophagy - animal
RIG-I-like receptor signaling pathway
Regulation of actin cytoskeleton
Herpes simplex virus 1 infection

Figure 4.5: Putative AOP for cluster 1 chemicals. Pathway network constructed using cluster 1 chemicals, illustrating how the pathways whose activity can be used to predict LC50 (green) and experimental concentration (red) are linked to pathways that can be used to predict heart-rate fold change (purple), using the shortest distance. A) How pathways predictive of heart rate fold change are connected to the Lysosome pathway and B) to the VEGF signalling pathway and selenocompound metabolism.

From cluster 2 chemicals, five pathways were found to be predictive of the chemical LC50, steroid biosynthesis, cellular senescence, glycerophospholipid metabolism, ECM-receptor interaction and intestinal immune network for IgA production (Figure 4.6). The steroid biosynthesis pathway is not directly linked to pathways predictive of heart-rate effect but is associated with amino acid and vitamin metabolism, through lipid metabolism and cofactors biosynthesis. The cellular senescence pathway is directly linked to the nicotinamide vitamin metabolism pathway, which the activity was found to be predictive of zebrafish heart rate. In addition, through metabolic pathways, cellular senescence pathway activity was found to be associated with glycan biosynthesis, amino acid and vitamin B6 metabolism, a pathway predictive of cardiotoxicity. Alteration in glycerophospholipid metabolism pathway activity was found to be directly linked to purine metabolism. Furthermore, four more pathways that have been predictive of heart rate, are associated with lipid metabolism and amino acid biosynthesis, were found downstream of glycerophospholipid metabolism but with one to three pathways in between. The ECM-receptor interaction pathway is linked downstream to three pathways, related to the immune system, cell death and endocrine system, where their activity is predictive of heart rate. In addition, the production of IgA is also directly linked to the MAPK signalling pathway and through calcium signalling, to arginine biosynthesis (Figure 4.6). The activity of 11 pathways was found to be predictive of chemical concentration and 18 with heart rate fold change. Seven out of the 11 pathways were directly linked to heart-rate-related pathways, and the rest have a single pathway in between (Figure 4.6). Comparing the pathway results from cluster 2 analysis, cellular senesce pathway activity was predictive of both LC50 and chemical concentration. At the same time, the IgA production pathway active was associated with LC50 and heart rate, and fatty acid elongation with chemical concentration and heart rate.


A)
Steroid hormone biosynthesis
Cellular senescence
Fatty acid biosynthesis
Tryptophan metabolism
Nicotinate and nicotinamide metabolism
Metabolic pathways
Biosynthesis of cofactors
Valine, leucine and isoleucine degradation
Retinol metabolism
Phosphonate and phosphinate metabolism
Glycosphingolipid biosynthesis - globo and isoglobo series
Vitamin B6 metabolism
Tyrosine metabolism


B)
Glycerophospholipid metabolism
ECM-receptor interaction
Purine metabolism
Lysosome
Intestinal immune network for IgA production
p53 signaling pathway
Adipocytokine signaling pathway
Glycerolipid metabolism
Necroptosis
Salmonella infection
MAPK signaling pathway
Calcium signaling pathway
Sphingolipid metabolism
Biosynthesis of amino acids
Carbon metabolism
Arginine biosynthesis
Fatty acid elongation
C)
Glycolysis / Gluconeogenesis
Steroid biosynthesis
Amino sugar and nucleotide sugar metabolism
Pantothenate and CoA biosynthesis
Metabolism of xenobiotics by cytochrome P450
Ribosome biogenesis in eukaryotes
Salmonella infection
Lysosome
Purine metabolism
Biosynthesis of amino acids
Peroxisome
Biosynthesis of cofactors
Intestinal immune network for IgA production
Glycerolipid metabolism
Retinol metabolism
Tyrosine metabolism

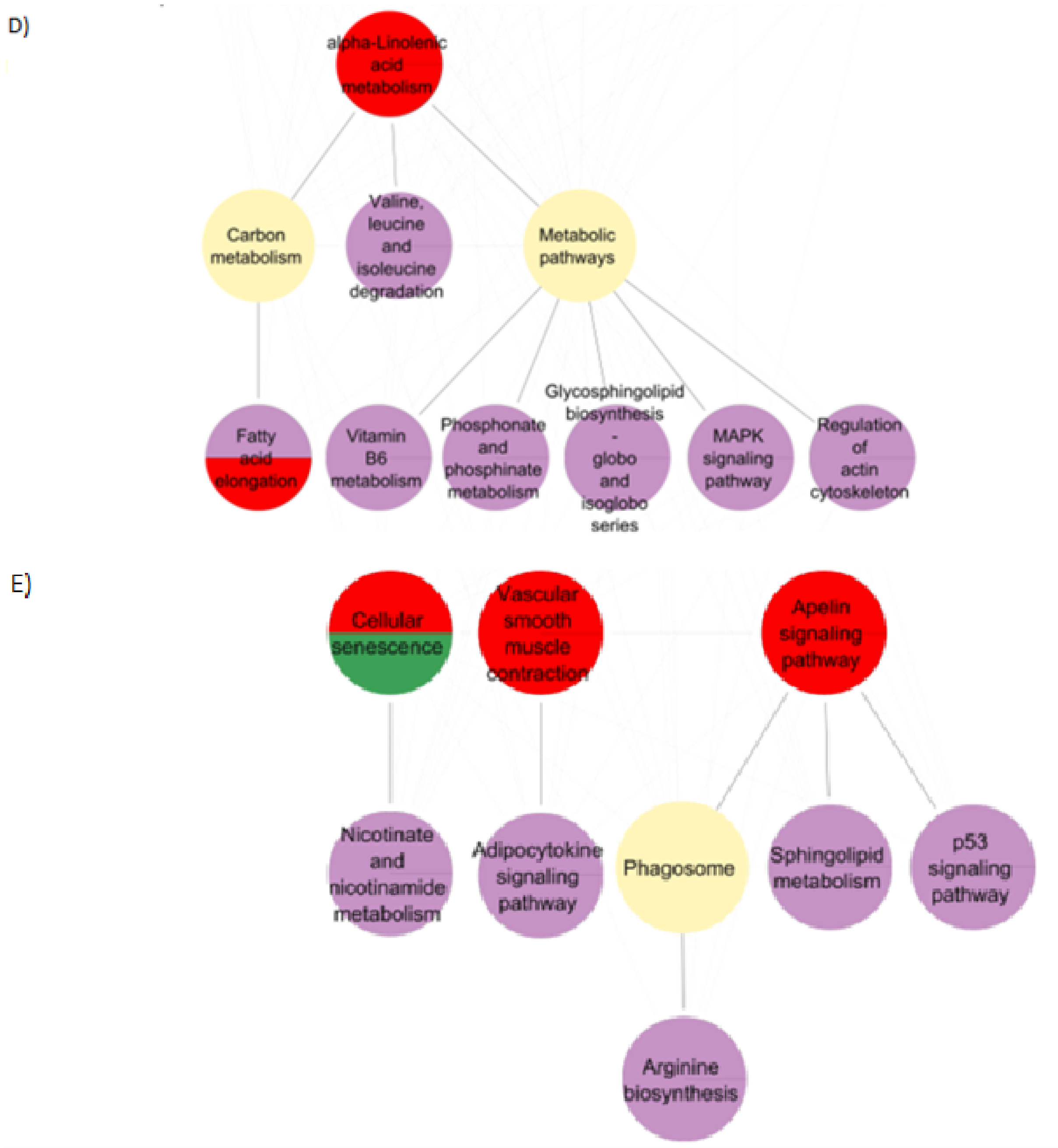
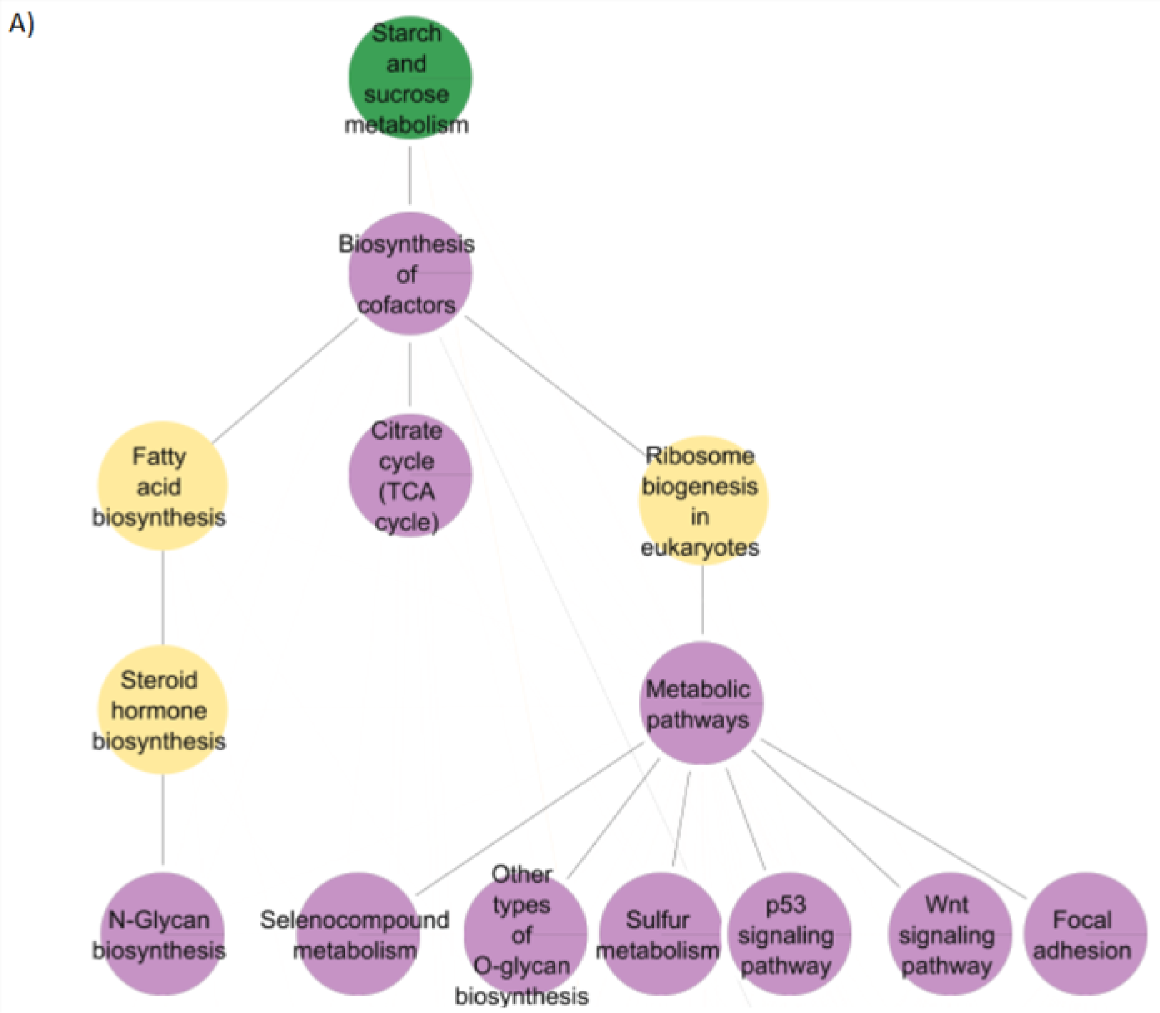

D)
alpha-Linolenic acid metabolism
Carbon metabolism
Valine, leucine and isoleucine degradation
Metabolic pathways
Fatty acid elongation
Vitamin B6 metabolism
Phosphonate and phosphinate metabolism
Glycosphingolipid biosynthesis - globo and isoglobo series
MAPK signaling pathway
Regulation of actin cytoskeleton
E)
Cellular senescence
Vascular smooth muscle contraction
Apelin signaling pathway
Nicotinate and nicotinamide metabolism
Adipocytokine signaling pathway
Phagosome
Sphingolipid metabolism
p53 signaling pathway
Arginine biosynthesis

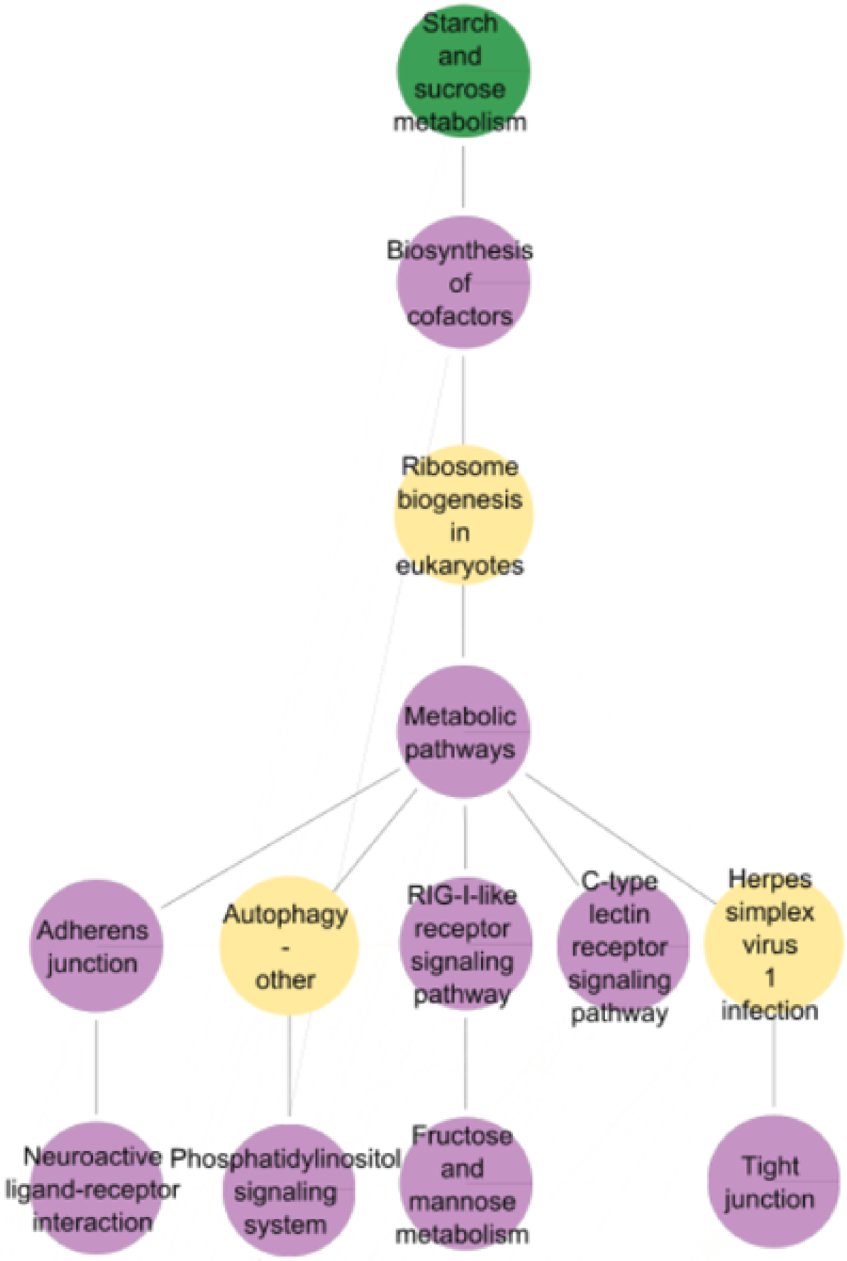
Figure 4.6: Pathway network constructed using cluster 2 chemicals, illustrating how the pathways whose activity can be used to predict LC50 (green) or chemical concentration (red) are linked to pathways that can be used to predict heart-rate fold change (purple) using the shortest distance. A) How pathways predictive of heart rate fold change are connected to the steroid hormone biosynthesis and cellular senescence pathways (green) and B) to glycerophospholipid metabolism, ECM-receptor interaction and the intestinal immune network for IgA production pathways (green). C) How pathways predictive of heart rate fold change are connected to glycolysis/gluconeogenesis, steroid biosynthesis, amino sugar and nucleotide sugar metabolism, pantothenate and CoA biosynthesis, metabolism of xenobiotics by cytochrome P450, and ribosome biogenesis in eukaryotes pathways (red), D) alpha-linolenic acid metabolism, fatty acid elongation pathways (red) and E) cellular senescence, vascular smooth muscle concentration and apelin signalling pathways (red).

Finally, looking through the pathway results from cluster 3 chemicals, only one pathway was identified to be predictive of LC50, starch and sucrose metabolism, which is linked to several pathways associated with heart rate changes through the cofactors biosynthesis pathway (Figure 4.7). The cofactors biosynthesis pathway is linked to 16 pathways related to heart rate changes, either directly (citrate cycle), or with one to three pathways in between, such as energy metabolism, cell death, signal transduction, cellular community and immune response. Lysosomes activity, on the other hand, is linked to 15 pathways that have been associated with heart rate, either directly (nucleotide and carbohydrate metabolism, salmonella infection – cell death, immune system, signalling pathway), or with one to three pathways in between, such as vascular smooth muscle contraction and endocrine system pathways. When looking into the pathways that were predictive of chemical concentration, it can be seen that only one was selected. It can also be used to predict zebrafish heart rate fold change and vascular smooth muscle contraction pathway (Figure 4.7). This pathway is directly linked downward with the adipocytokine signalling pathway, predictive of heart rate, which in turn is directly or indirectly (one to three pathways in between) associated with 23 pathways whose activity was found to be predictive of heart rate (cell death, metabolic pathways, energy metabolism).

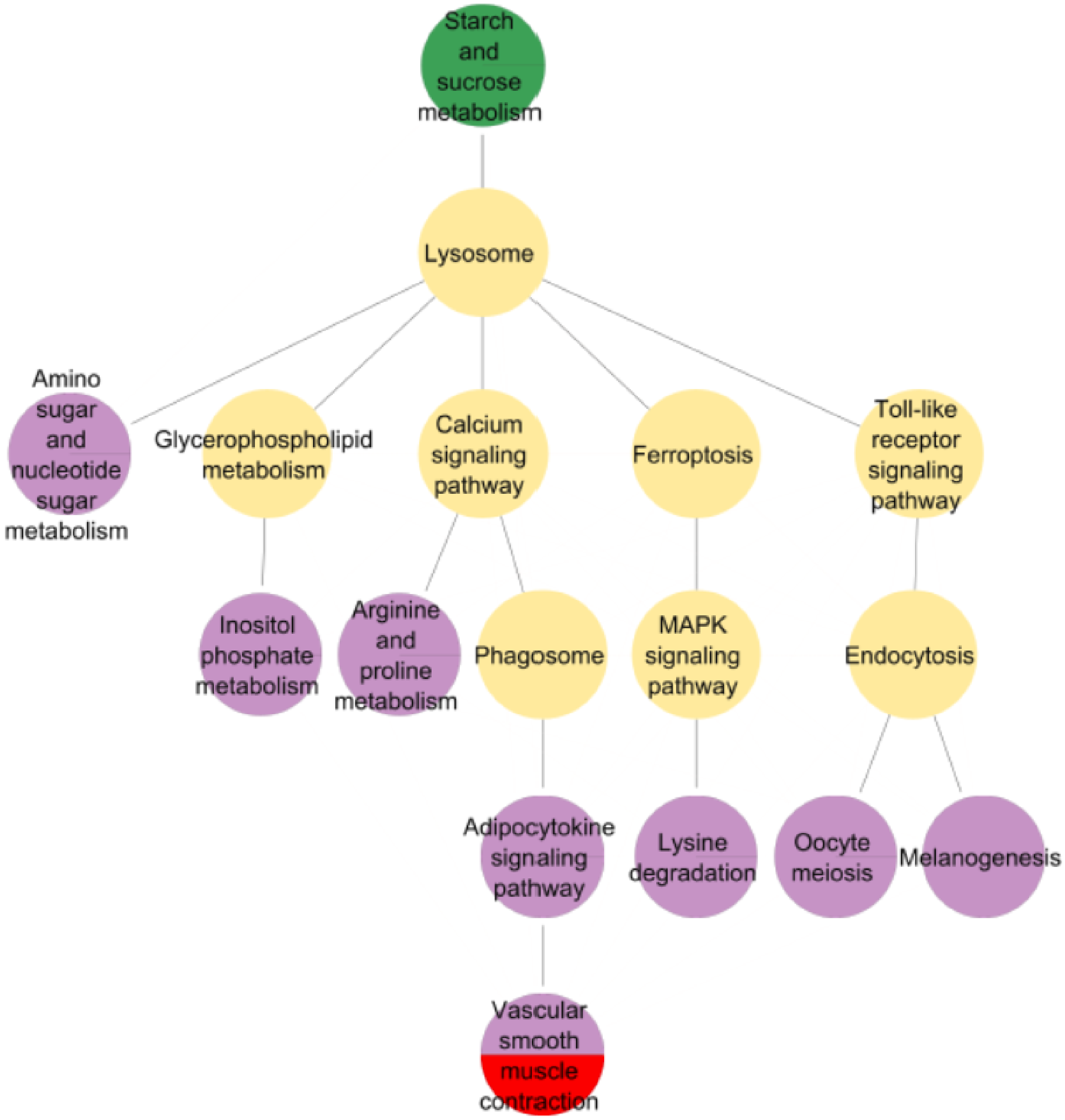
A)


Starch and sucrose metabolism
Biosynthesis of cofactors
Fatty acid biosynthesis
Citrate cycle (TCA cycle)
Ribosome biogenesis in eukaryotes
Steroid hormone biosynthesis
Metabolic pathways
N-Glycan biosynthesis
Selenocompound metabolism
Other types of O-glycan biosynthesis
Sulfur metabolism
p53 signaling pathway
Wnt signaling pathway
Focal adhesion

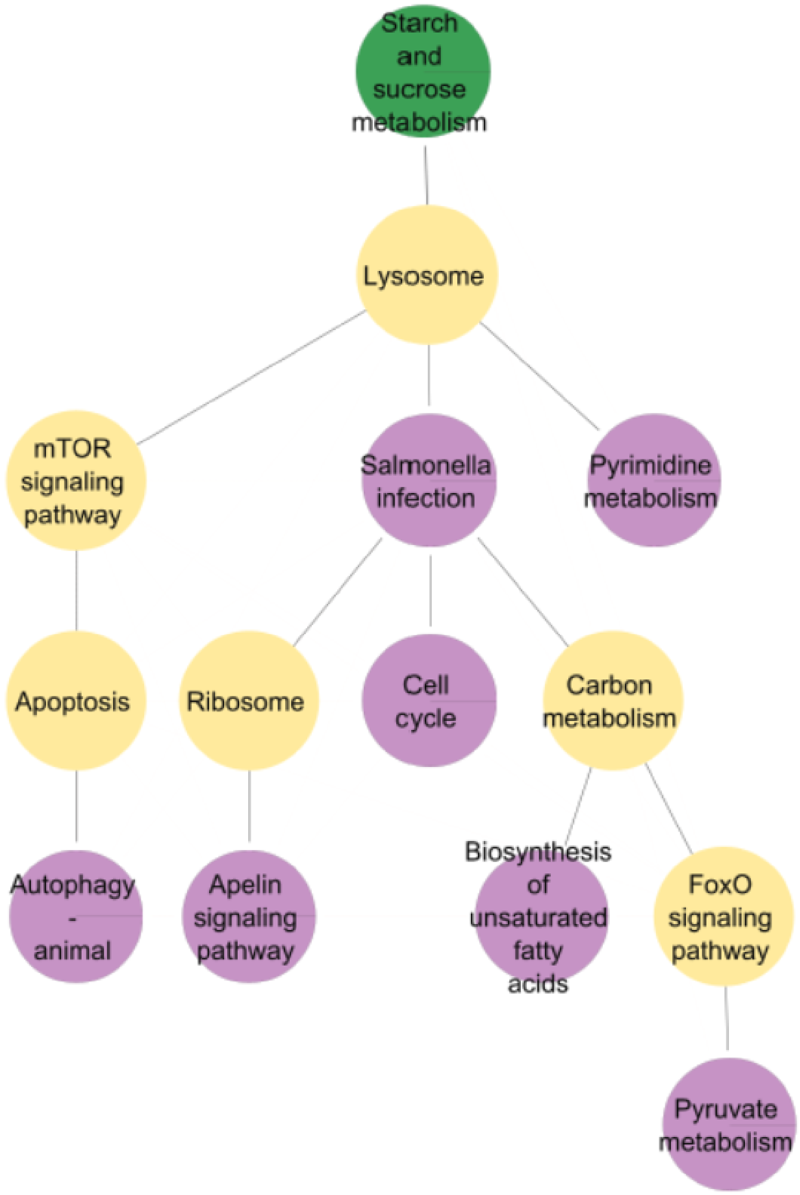
B)


Starch and sucrose metabolism
Biosynthesis of cofactors
Ribosome biogenesis in eukaryotes
Metabolic pathways
Adherens junction
Autophagy - other
RIG-I-like receptor signaling pathway
C-type lectin receptor signaling pathway
Herpes simplex virus 1 infection
Neuroactive ligand-receptor interaction
Phosphatidylinositol signaling system
Fructose and mannose metabolism
Tight junction

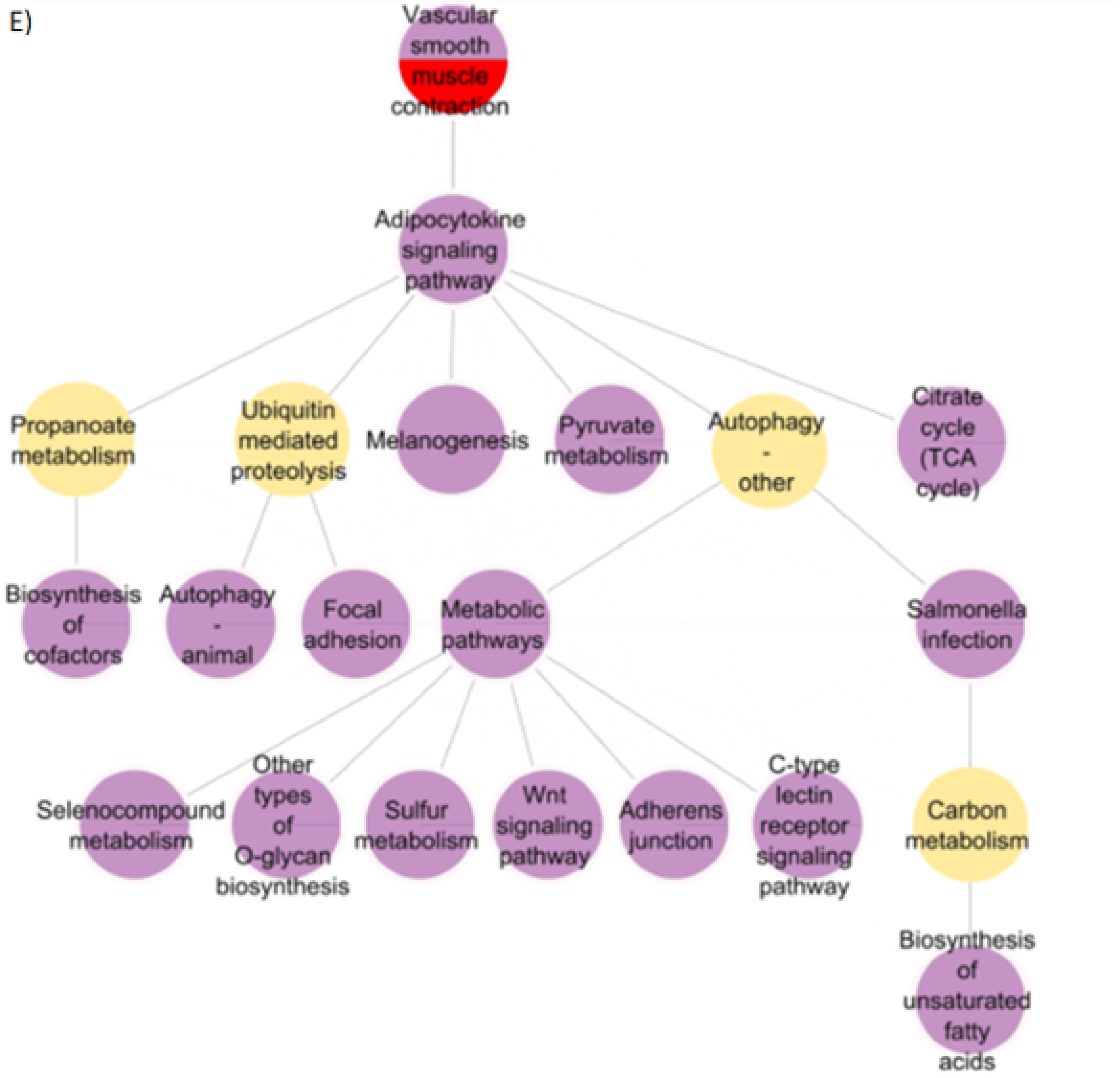
C)


Starch and sucrose metabolism
Lysosome
Amino sugar and nucleotide sugar metabolism
Glycerophospholipid metabolism
Calcium signaling pathway
Ferroptosis
Toll-like receptor signaling pathway
Inositol phosphate metabolism
Arginine and proline metabolism
Phagosome
MAPK signaling pathway
Endocytosis
Adipocytokine signaling pathway
Lysine degradation
Oocyte meiosis
Melanogenesis
Vascular smooth muscle contraction

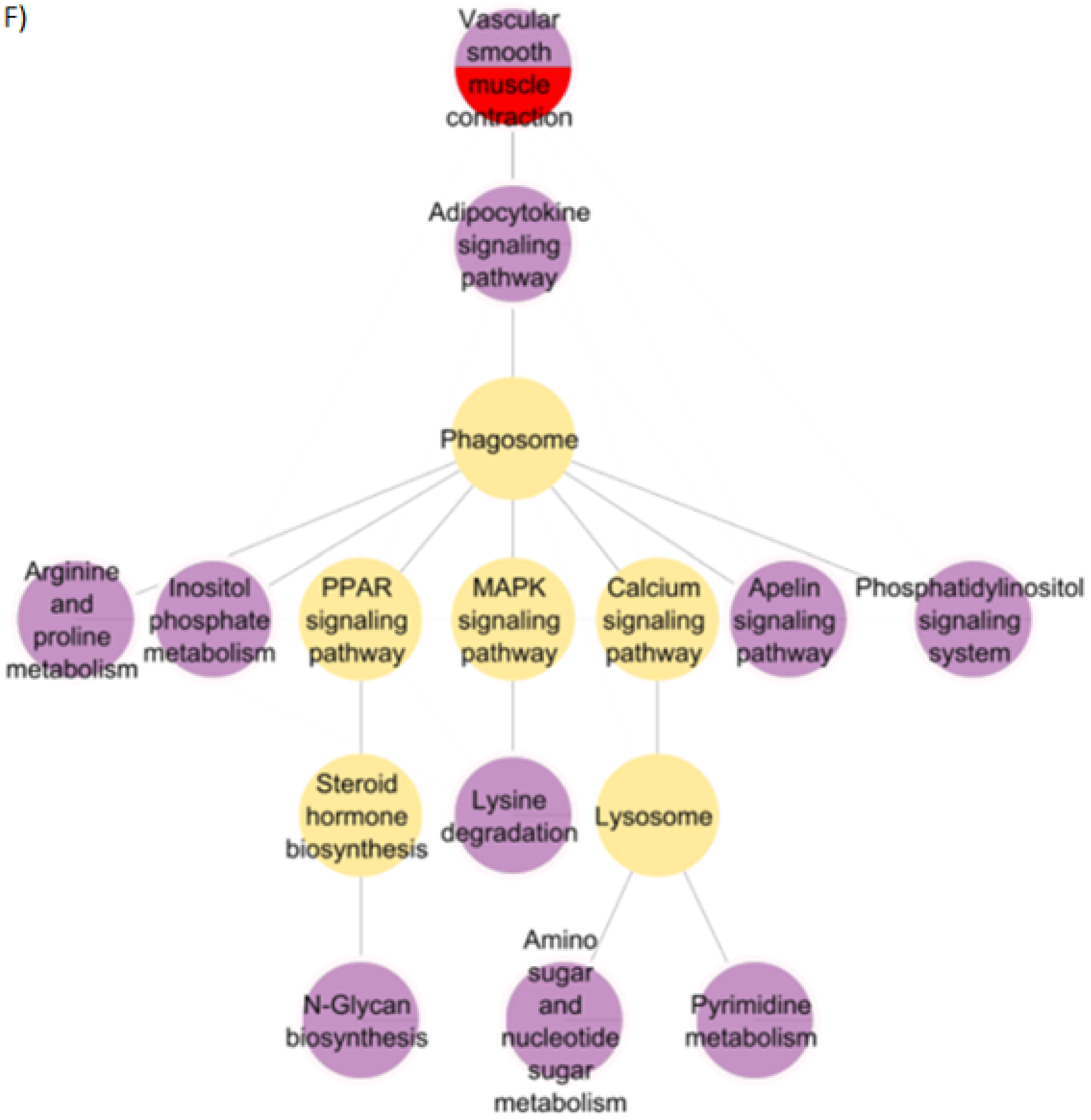
D)


Starch and sucrose metabolism
Lysosome
mTOR signaling pathway
Salmonella infection
Pyrimidine metabolism
Apoptosis
Ribosome
Cell cycle
Carbon metabolism
Autophagy - animal
Apelin signaling pathway
Biosynthesis of unsaturated fatty acids
FoxO signaling pathway
Pyruvate metabolism

E)


Vascular smooth muscle contraction
Adipocytokine signaling pathway
Propanoate metabolism
Ubiquitin mediated proteolysis
Melanogenesis
Pyruvate metabolism
Autophagy - other
Citrate cycle (TCA cycle)
Biosynthesis of cofactors
Autophagy - animal
Focal adhesion
Metabolic pathways
Salmonella infection
Selenocompound metabolism
Other types of O-glycan biosynthesis
Sulfur metabolism
Wnt signaling pathway
Adherens junction
C-type lectin receptor signaling pathway
Carbon metabolism
Biosynthesis of unsaturated fatty acids

F)



Vascular smooth muscle contraction
Adipocytokine signaling pathway
Phagosome
Arginine and proline metabolism
Inositol phosphate metabolism
PPAR signaling pathway
MAPK signaling pathway
Calcium signaling pathway
Apelin signaling pathway
Phosphatidylinositol signaling system
Steroid hormone biosynthesis
Lysine degradation
Lysosome
N-Glycan biosynthesis
Amino sugar and nucleotide sugar metabolism
Pyrimidine metabolism

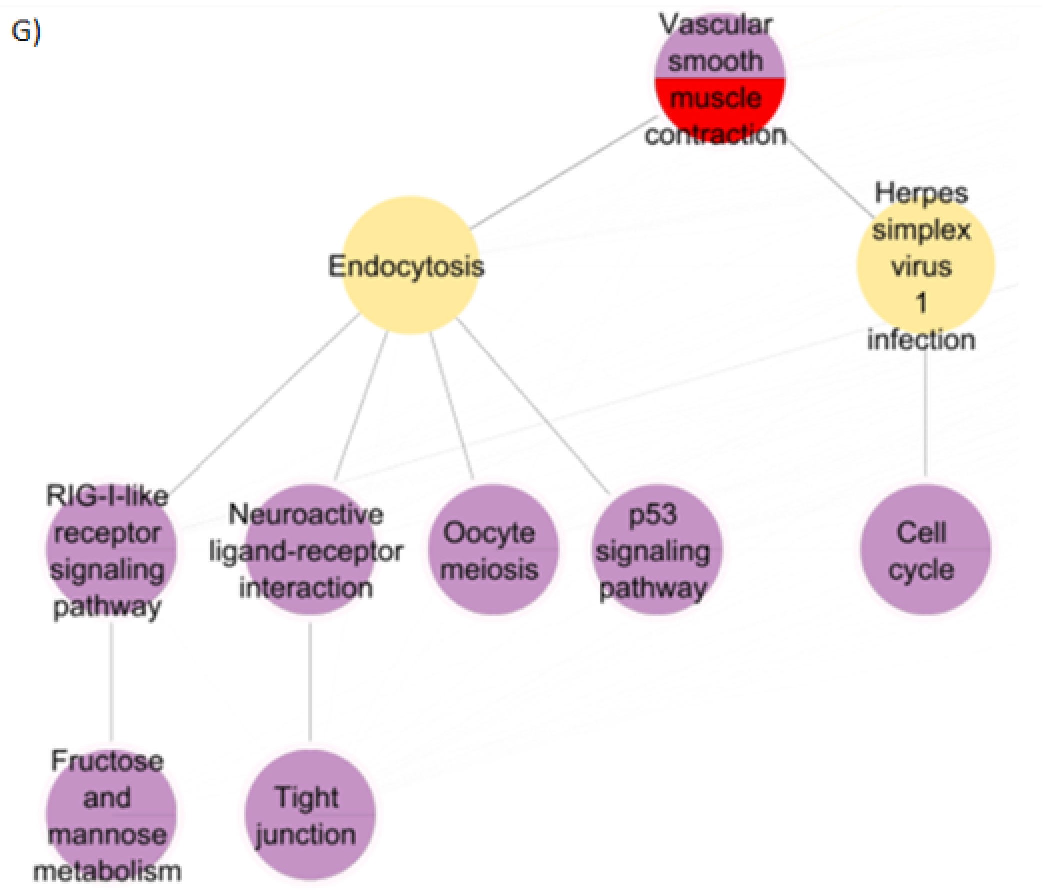

G)
Vascular smooth muscle contraction
Endocytosis
Herpes simplex virus 1 infection
RIG-I-like receptor signaling pathway
Neuroactive ligand-receptor interaction
Oocyte meiosis
p53 signaling pathway
Cell cycle
Fructose and mannose metabolism
Tight junction

Figure 4.7: Pathway network constructed using cluster 3 chemicals, illustrating how the pathways whose activity can be used to predict LC50 (green) or chemical concentration (red) are linked to pathways that can be used to predict heart-rate fold change (purple) using the shortest distance. A and B) How pathways predictive of heart rate fold change are connected to starch and sucrose metabolism pathway (green) through the biosynthesis of cofactors pathway and C and D) how pathways predictive of heart rate fold change are connected to starch and sucrose metabolism pathway (green) through the lysosome pathway. E and F) How pathways predictive of heart rate fold change are connected to vascular smooth muscle contraction pathway (red) through the adipocytokine signalling pathway and G) how pathways predictive of heart rate fold change are connected to vascular smooth muscle contraction pathway (red) through endocytosis pathway and herpes simplex virus 1 infection pathway.

4.5 Discussion

4.5.1 Pathway activity for predicting heart-rate fold change using the whole dataset and mRNA clustering

The aim of this study was to identify the link between pathway activity and heart-rate fold change, experimental chemical concentration and chemical LC50. When the whole dataset was used, out of the 160 KEGG pathways, 137 pathway activities were identified to be predictive of heart-rate changes in zebrafish embryos showing the wide spectrum of pathways that are involved in heart development and function, and the high sensitivity of cardiogenesis processes. This also highlights the diverse nature of the chemicals used in this study, as chemicals can cause cardiotoxicity through multiple mechanisms. The pathways identified are involved in energy production and cardiac muscle contraction, nucleic acid and protein integrity, and environmental information processing. Reducing the dimensionality of the dataset by clustering chemicals based on gene expression profiles, resulted in 20 pathways being selected to be predictive of the effect chemicals from cluster 1 have on zebrafish heart rate, 18 pathways for cluster 2 chemicals, and 32 pathways for cluster 3 chemicals. Clustering based on gene expression, chemicals are grouped based on their effect on the molecular level, thus reducing the diversity among the gene expression profiles.

Pathways that are related to cardiac muscle contraction and blood flow, including the cardiac muscle contraction pathway, vascular smooth muscle contraction pathway and adrenergic signalling in cardiomyocytes have been predictive of the effect chemicals have on zebrafish heart rate. Most of the pathways selected were related to energy production. Heart muscles have a constant need for energy, thus production and turnover of ATP are essential for cardiac contractility. ATP is mainly generated through fatty acid oxidation and glucose metabolism (Goldberg *et al.*, 2012; Tran *et al.*, 2019). Various pathways relating to DNA, RNA and protein integrity, including nucleotide metabolism, replication and repair, folding, sorting and degradation, have been selected showing the importance of DNA-RNA synthesis and repair in proper embryonic development, through cell growth and repair (Diehl *et al.*,2021). Protein function depends on the proper three-dimensional structures, the ability of the protein to reach the area of action and in case of an error the degradation of such proteins, to avoid their accumulation. Amino acid metabolism is critical in nutrient metabolism, protein synthesis and immune responses. For example, arginine is involved in nutrient metabolism, stimulating insulin release, is involved in nonspecific immune response, regulates energy homeostasis (AMPK), protein synthesis (TOR signalling) and regulates endocrine and metabolic systems (Wang *et al.*, 2021). Branched-chain amino acids are also involved in mTOR signalling, have a stimulatory effect on insulin secretion and have an inhibitory effect on muscle proteolysis (Holeček, 2018). Aromatic amino acids are broken down or converted into neurotransmitters (Holeček, 2018; Parthasarathy *et al.*, 2018).

Cardiac development and function also rely on signal transductions, signalling molecules and interactions. Signalling pathways such as MAPK are involved in proliferation, differentiation, metabolism, survival, and apoptosis. ErbB is a tyrosine kinase receptor that binds to extracellular growth factors such as neuregulin-1 (Nrg-1) and plays an important role in cardiovascular development by regulating tissue organisation during development and maintaining cardiac function. Dysregulation of the calcium signalling pathway has been involved in cardiotoxicity since it is responsible for the excitation-contraction coupling of the heart (Salgado-Almario *et al.*, 2020). MTOR signalling pathway regulates multiple biological properties, including lipid metabolism, autophagy, protein synthesis, ribosome biogenesis and proteostasis in cardiomyocytes (Bu *et al.*, 2021). The extracellular matrix (EMC) provides structural support and plays a crucial role in cardiac homeostasis by force transmission and transducing key signals to cardiomyocytes, vascular cells, and interstitial cells. Changes in the biochemistry of EMC have been associated with the expansion of the cardiac interstitium (Frangogiannis, 2019). WNT and Notch signalling pathways control cardiac developmental asymmetry and regulate blood vessel stability in zebrafish (Blankesteijn, 2020).

In addition, the endocrine system is responsible for hormonal signalling pathways to control and coordinate metabolism, energy release, reproduction growth and development. Among the chemicals selected for this study, some are characterised as endocrine disruptors that have been associated with cardiovascular health by hormone hype- or hypofunction. Endocrine disruptors cause insulin production and function defects and have been linked to type-2 diabetes, carbohydrate, and lipid metabolic disorders (Toyoshima *et al.*, 2008; Kirkley *et al.*, 2014). The PPAR signalling pathway is important for heart function since it is involved in lipid metabolism and modulates energy production by the breakdown of lipids (Den Broeder *et al.*, 2015), and gonadotropin-releasing hormone (GnRH) for the regulation of the reproductive axis and neuron migration (Onuma *et al.*, 2011). In addition, the adipocytokine signalling pathway regulates energy balance and participates in inflammation, coagulation, and fibrinolysis.

Reducing the diversity of the dataset using mRNA clustering, the number of pathways that can be used to predict the effect a chemical has on zebrafish heart rate was reduced. This decrease in the number of pathways selected was expected since mRNA clustering groups chemicals based on their molecular effect, thus only chemicals with similar gene count profiles were used, reducing the spectrum of the effect they have on zebrafish. Some pathways were common among all three clusters, such as lipids, amino acids and nucleotide metabolism, metabolism of cofactors and vitamins, cell death, signalling and energy production. Regulation of the actin cytoskeleton is important for maintaining the cell structure and shape and is involved in cell migration, polarity, intracellular or extracellular trafficking, cell-cell interaction, and cell division (Balta *et al.*, 2020). Autophagy and necroptosis are forms of cell death in drug-induced cardiotoxicity. Autophagy maintains intracellular metabolic homeostasis by removing unwanted or damaged cellular components, and necroptosis cell death occurs after exposure to extreme physical or chemical insults (Ma *et al.*, 2020).

The chemicals from (hard) mRNA cluster 1 (chapter 2) apart from the pathways described above, cell adhesion molecules pathway is also associated with heart-rate fold changes. Cell adhesion molecules are proteins on the cell surface and are involved in various biological processes such as embryogenesis and the development of neuronal tissues. Using the mRNA cluster 2 chemicals the adipocytokine signalling pathway, responsible for leptin production, regulation of energy intake and metabolic rate, and adiponectin, involved in skeletal muscle fatty acid oxidation and glucose uptake, was also predictive of heart-rate fold change after exposure. Finally for the chemicals from mRNA cluster 3, the activity of pathways selected after predictive modelling were related to vascular smooth muscle contraction, cell adhesion, and the control of the cell cycle.

4.5.2 Pathway activity for predictive toxicity using the whole dataset and mRNA clustering

On the other hand, no KEGG pathways were found to be predictive of the experimental concentration or the LC50 of the chemical, showing the difficulty of predicting toxicity from highly diverse datasets. The mRNA (hard) clustering was applied from chapter 2 in an effort to reduce the diversity of the dataset. When only chemicals from mRNA cluster 1 were used, one pathway was identified to be predictive of both toxicity phenotypes (LC50/ chemical concentration), lysosome activity. On the other hand, selenocompound metabolism was only identified to be predictive of the experimental concentration, and VEGF signalling for LC50 predictions. Lysosomal destabilisation can be used as an indicator of chemical stress since organic and inorganic chemicals tend to accumulate in those organelles and damage the lysosomes (Hwang *et al.*, 2002). Lysosome cell organelles contain digestive enzymes that break down excess or worn-out cell parts and macromolecules and promote cell death by regulating apoptosis in case of cell damage (KEGG Database). Chemical toxicity has been linked to oxidative stress by free radicals that can cause lipid peroxidation, disruption of cell membrane and nucleic acid oxidation followed by cell damage. Antioxidant enzymes such as selenocompounds, deactivate free radicals by reducing their energy, increasing their stability to minimise cellular damage (Amjad *et al.*, 2020). The VEGF signalling pathway in zebrafish is involved in the formation and growth of blood vessels, by regulating gene expression, and vascular permeability and promoting cell migration, proliferation, and survival (Bussmann *et al.*, 2008).

The LC50 of the chemicals from mRNA cluster 2, was predicted by 10 pathways. Some of the pathways found to be predictive of chemicals LC50 were described earlier, lipid metabolism, cell-cell interactions and signalling including extracellular matrix interactions, GnRH signalling pathway. In addition, pathways that are involved in steroid biosynthesis, DNA repair and cellular senescence were identified. Environmental chemicals are found to disrupt the endocrine system of zebrafish. Steroid hormones play a key role in sex determination, reproduction, growth, and development alterations. Disruption of cortisol production has been linked to growth impairment, pericardial oedema, vascular system defects and altered somitogenesis, in zebrafish embryos (Tokarz *et al.*, 2013).

Non-homologous end-joining is responsible for double-strand breaks in DNA repair. Double strand breaks are introduced either by endogenous sources, such as reactive oxygen species and replication errors, or exogenous sources, such as toxic chemicals. When those breaks are not repaired can lead to cellular senescence (Davis *et al.*, 2013). The 13 pathways selected to be associated with the experimental chemical concentration are involved in lipid metabolism, carbohydrate metabolism, ATP production, vascular smooth muscle contraction, CoA biosynthesis, DNA repair and ribosomes biogenesis, cellular senescence, steroid biosynthesis, biodegradation of chemicals foreign to the animal, such as drugs and pesticides. The activity of the apelin signalling pathway, which plays an important role in angiogenesis, cardiovascular function, cell proliferation and energy metabolism regulation is also altered under toxic conditions (Helker *et al.*, 2020).

Finally, only one pathway activity was predictive of chemical LC50, starch and sucrose metabolism, which is involved in development and stress response by providing sugars necessary for growth and the synthesis of important compounds such as proteins, used as signals regulating the expression of microRNA and transcription factors among others (Ruan, 2014). One pathway was also selected to be predictive of the experimental concentration, the vascular smooth muscle contraction.

4.5.3 Pathway networks

The pathways selected by the various models in this study were turned into networks connecting the pathways related to toxicity (chemical LC50, experimental concentration) with the pathways associated with the chemical effect on heart rate. The results show that the distance between toxicity-related pathways and heart rate-related pathways was shorter compared to the rest of the pathways, indicating that those phenotypes are closely related. In addition, the results suggest that pathways found to be associated with LC50 and chemical concentration are found closer to an MIE (receptor binding, cellular membrane changes in fluidity and transport), and are closely related to pathways associated with an adverse outcome such as changes in heart rate.

4.5.4 Conclusion

Pathway analysis and more specifically the use of the KEGG pathway database allows the identification of specific pathways associated with chemical-induced toxicity instead of a set of genes, which provide information related to the underlying biological mechanism involved in chemical-induced toxicity. As it can be seen, pathways identified in this study to be predictive of chemical LC50 or chemical exposure concentration can be close to or represent a potential MIE (e.g. receptor binding), and the pathways found to be predictive of heart rate fold change can represent an AO. Shared genes between the pathways can be used to generate pathway networks, linking pathways based on shared genes, which allows for the Identification of the shortest path, i.e. the path with the least number of pathways that connects two pathways (e.g. MIE to AO). This can potentially assist in generating cardiotoxicity AOPs, where a pathway identified to be predictive of chemical toxicity (LC50 or experimental concentration) represents an MIE, a pathway identified to be predictive of heart rate fold change represents an AO and all the pathways in between (identified through the shortest path) represent the key events. The results suggest that the use of large-scale genomics data and pathway analysis can be used in identifying new key events or potential AOPs and assist in the generation of AOPs networks for chemical risk assessment.

Chapter 5

General discussion

5.1 AOP framework can assist risk assessment

Exposure to environmental toxins during embryonic development can be dangerous. Cardiac development is a highly sensitive process and is regulated by molecular cellular and environmental factors, thus exposure to toxic chemicals may lead to cardiotoxicity, including changes in heart rate, and damage to the myocardium (Sarmah *et al.*, 2016). Risk assessment is used to evaluate the harmful effect chemicals may have on an organism, by hazard identification (whether a chemical is harmful (e.g. MoA)), dose-response assessment (mathematical relationship between exposure and toxic effect), exposure assessment (frequency, duration and levels of chemical exposure required for an adverse outcome), and risk characterization (define how the chemical should be used by combining all the information collected) (Kang *et al.*, 2018). Mathematical and statistical models (QSAR) have been introduced in chemical risk assessment, identifying biochemical and physiological factors identified through in vivo and in vitro experiments such as chemical absorption, distribution, metabolism, and excretion (Zvinavashe *et al.*, 2008).

Zebrafish embryos have been widely used in risk assessment and cardiotoxicity evaluation, due to the high conservation of cardiovascular physiology and electrical properties between vertebrates. In addition, zebrafish heart rates are similar to humans (Zhang *et al.*, 2011; Chen, 2013). Various toxins, including acetylcholinesterase inhibitors, organic pollutants, and β-adrenergic receptors, reduce heart rate and cause cardiac oedema in zebrafish embryos, through multiple mechanisms, including oxidative stress, iron overload, and DNA and mitochondrial damage (Lymperopoulos *et al.*, 2013; Y. Zhang *et al.*, 2013; Watson *et al.*, 2014; Hoeger *et al.*, 2020). Such chemicals, including terfenadine, tacrine, chlorpromazine and prochloraz, are highly toxic and exposure may lead to bradycardia even at relatively low concentrations.

The AOP concept organizes the existing knowledge related to the various biological events that lead to an adverse effect, connecting an MIE and an AO through the identification of KE that occur at a biological level. AOPs have been used in risk assessment for predicting an AO caused due to chemical exposure, reducing animal testing, experimental cost, and time (Ankley *et al.*, 2010).

Advances in ‘omics technologies and system-level data analysis facilitate the identification of molecular responses (MIE, KEs) associated with chemical exposure, through gene and pathway analysis, and provide more information about the molecular toxicity mechanisms and the molecular effects (Antczak *et al.*, 2015; Brockmeier *et al.*, 2017). Pathway analysis and the identification of pathways whose activity is altered due to exposure, uncover the mechanisms involved in chemical toxicity and can be used to describe the molecular responses after toxic exposure. The results of this study also suggest that pathways activity analysis can assist in generating potential AOPs by uncovering the pathways that are involved in the response and the potential connection between them. Pathway analysis and predictive modelling allow the identification of pathways, whose activity is altered after chemical exposure and can be used to predict chemical toxicity (LC50/chemical concentration). Such pathways are found to be close to or represent a potential MIE, for example, receptor binding and extracellular matrix activity. On the other hand, alteration of heart rate in zebrafish is an indication of the ability of a chemical to cause cardiotoxicity after exposure, thus pathways whose activity can be used to predict heart rate fold change can be used as potential AOs. Pathway analysis and shared genes between pathways are used to generate pathway networks, indicating the path that a signal needs to follow in order to generate a response. The shortest path (least number of pathways) between pathways associated with toxicity (LC50/experimental chemical concentration) and pathways predictive of heart-rate fold change can potentially represent an AOP, with the pathways in between representing the KEs. Thus, utilising pathway information and predictive modelling may assist in the generation of cardiotoxicity AOPs, reducing extensive animal experiments and providing insight into the underlying biological mechanisms involved in the response to exposure to toxic chemicals.

Two putative AOPs were generated and shown here, based on the data generated in this thesis. AOP 1 was generated using the chemicals from cluster 2, and AOP 2 from cluster 3 chemicals and they were both selected based on the predictive power of the individual models. These models are examples of the potential use of pathway analysis in generating cardiotoxicity AOP networks using pathways predictive of chemical toxicity, related to MIE, and pathways predictive of heart rate fold change, AO.

5.1.1 Potential cardiotoxicity AOP example 1

The first proposed cardiotoxicity AOP described in this study, provides a link between the ECM-receptor interaction pathway, the intestinal immune network for IgA production and the MAPK signalling pathway and is generated using the chemicals from mRNA cluster 2 (Figure 5.1). ECM-receptor interaction pathway, which is found to be predictive of toxicity (chemical L50), represented in this study by 76 genes, is related to environmental information processing. Several drugs have been found to influence ECM metabolism and regulate ECM composition and organisation, such as cytokine inhibitors, glucocorticoids, ACE inhibitors and calcium channel blockers (Järveläinen *et al.*, 2009). Multiple diseases have been associated with ECM structure alterations, disturbances in metabolism, and dysregulation in ECM-cell signalling (Ricard-Blum *et al.*, 2016; Sainio *et al.*, 2020). ECM components are involved in tissue and organ morphogenesis and the maintenance of cell and tissue structure and function. ECM macromolecules such as collagens, provide structural support (Gordon *et al.*, 2010), elastin and fibrillin are responsible for tissue elasticity (Czirok *et al.*, 2006), proteoglycans and hyaluronan are involved in the formation of pericellular matrix and tissue homeostasis (Knudson *et al.*, 1991; Smith *et al.*, 2019), and glycoproteins such as fibronectins are important in multiple embryogenic processes (Rozario *et al.*, 2009), and laminins modulate cell adhesion, differentiation, and migration (Patarroyo *et al.*, 2002).

ECM molecules such as growth factors, cytokines, chemokines, matrix-degrading enzymes, and their inhibitors (Sainio *et al.*, 2020) are involved in cell signalling regulation. ECM-cell interactions are mediated by transmembrane molecules, such as integrins (adhesion receptors) (Kechagia *et al.*, 2019), discoidin domain receptor (DDR) family for collagens (Johnson *et al.*, 1993), proteoglycans and other cell-surface- associated components, mediate cell signalling by ECM macromolecules. These interactions control cellular activities including adhesion, migration, differentiation, proliferation, and apoptosis.

ECM- receptor interaction pathway was found to be directly linked to the intestinal immune network for the IgA production pathway, which is predictive of both LC50 and heart rate fold change and consists of 32 genes. These two pathways share two genes, integrin beta and integrin alpha 4 proteins, that are involved in cell adhesion and transmission of signals to the cytoplasm. The intestinal immune network for the IgA production pathway consists of multiple genes associated with cytokine-cytokine receptor interaction, growth factors and integrins, which are important for cell signalling.

The last part of this potential AOP is the Mitogen-activated protein kinase (MAPK) signalling pathway, predictive of heart-rate fold change, that is represented by 310 genes. The intestinal immune network for the IgA production pathway and MAPK pathway share four genes associated with TGF-β, which is involved in apoptosis control and angiogenesis (Prud’homme, 2007), ADP-ribosylation factor 2b enables GTP binding and is involved in endocytic recycling and intracellular protein transport and Mitogen-activated protein kinase

kinase kinase 14a, a serine/theonine protein-kinase that stimulated NF-kB activity inducing the expression of pro-inflammatory genes.

MAPKs form complex signalling networks and are activated by a variety of stimuli following a canonical cascade activation and are involved in proliferation, differentiation, metabolism, survival, and apoptosis (Kyriakis *et al.*, 2001; Bogoyevitch *et al.*, 2006; Rincón *et al.*, 2009; Rose *et al.*, 2010). Extracellular signal-regulated kinases (ERK1/2, ERK5), c-Jun N-terminal kinase (JNK) and p38 are the most popular MAPK pathways. ERK1/2 pathway is activated by many hormones, G proteins, growth factors and insulin (Kyriakis *et al.*, 2001; Goldsmith *et al.*, 2007; Katz *et al.*, 2007; Raman *et al.*, 2007). This pathway is involved in multiple biological processes, including cell cycle progression, proliferation, cytokinesis, senescence, migration, GAP junction formation, actin and microtubule networks and cell adhesion (Ramos, 2008). JNK and p38 are stress-activated MAPKs and are activated by growth factors, G protein-coupled receptors (Goldsmith *et al.*, 2007; Katz *et al.*, 2007), but also by physiological stressors such as oxidative stress, hyperosmolarity, cellular stress osmotic shock, infection, cytokines, DNA damage, and ER stress (Kyriakis *et al.*, 2001; Raman *et al.*, 2007). They are both involved in multiple biological processes, including cell proliferation, differentiation, apoptosis, cell survival and cytokine production (Kyriakis *et al.*, 2001; Bogoyevitch *et al.*, 2006; Rincón *et al.*, 2009; Rose *et al.*, 2010). Finally, the ERK5/BMK pathway is important in growth, stress signalling (Hayashi *et al.*, 2004; Hayashi *et al.*, 2004), vascular formation (Hayashi *et al.*, 2004; Hayashi *et al.*, 2004), cell survival, differentiation, proliferation, and growth (Nishimoto *et al.*, 2006; Wang *et al.*, 2006). This pathway is activated by growth factors, such as VEGF and nerve growth factors, and stress stimuli like oxidative stress (Hayashi *et al.*, 2004).

Various studies on heart formation have identified multiple signalling pathways and transcription factors associated with MAPK pathways, including Hedgehog, bone morphogenic protein (BMP), FGF and Wnt-JKNK (Dunwoodie, 2007). FGFs (growth factors) and their receptors are expressed throughout development in the epicardium, endocardium, and myocardium (Sugi *et al.*, 2003; Lavine *et al.*, 2005). They are important in cardiogenic induction, and morphogenesis, and are involved in various cellular processes during development (Böttcher *et al.*, 2005). Wnts on the other hand is involved in many developmental processes including cell polarity (Nelson *et al.*, 2004) and its activation promoted cardiac cell fate induction and inhibition (Zhou *et al.*, 2007; Cohen *et al.*, 2008). VEGF (growth factor) promotes cardiomyocyte differentiation through the ERK pathway (Chen *et al.*, 2006). All four MAPK pathways described here have been associated with cardiovascular diseases, cancer, and diabetes (Ramos, 2008; Rose *et al.*, 2010). Cardiac hypertrophy is a common response to external stressors including mechanical overload and oxidative stress (Rose *et al.*, 2010).

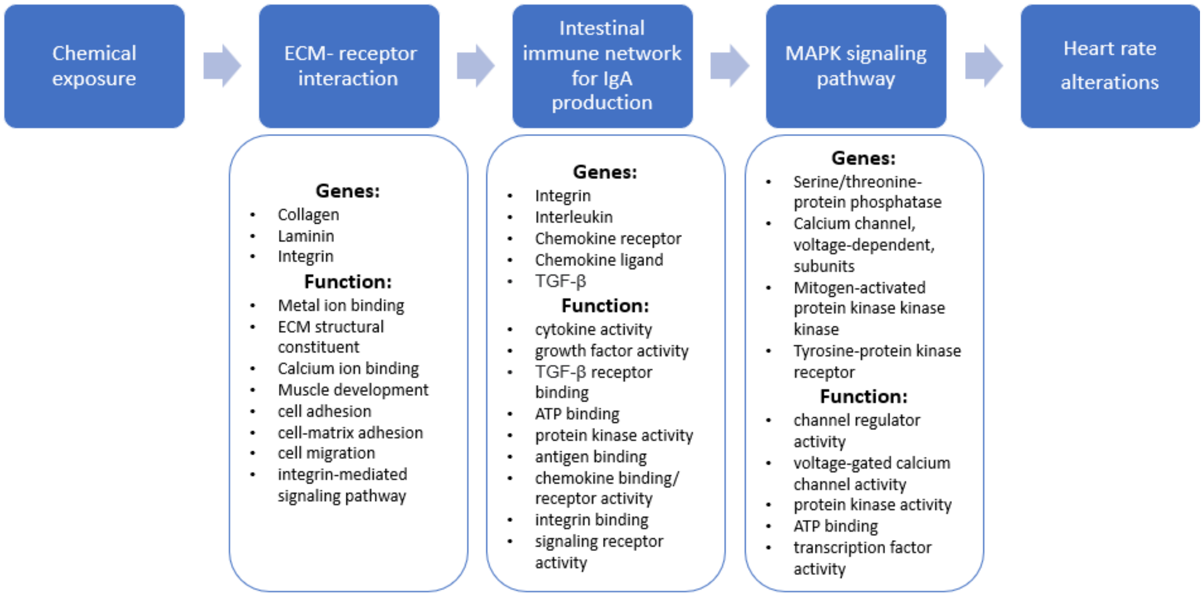


Figure 5.1: Potential AOP example 1 representation, generated from cluster 2 chemicals. An adverse outcome pathway, starting from ECM-receptor interaction, that is downstream linked to the Intestinal immune network for IgA production, is linked to the MAPK signalling pathway. Some of the genes and functions related to each pathway are also provided.

5.1.2 Potential cardiotoxicity AOP example 2

The second potential cardiotoxicity AOP constructed here using chemicals from mRNA cluster 3, highlights the link between starch and sucrose metabolism, biosynthesis of cofactors and citrate cycle (TCA cycle) (Figure 5.2). The first pathway identified here is starch and sucrose metabolism, which is predictive of toxicity (LC50) and consists of 36 genes. This list of genes consists mostly of enzymes that are involved in carbohydrate metabolism and degradation, glycogen synthesis, and glycolysis. Glycolysis is the breakdown of glucose into pyruvate that is used further down for the generation of acetyl-CoA, an essential cofactor (Shi *et al.*, 2015). This pathway is directly linked to the biosynthesis of cofactors, which is predictive of heart-rate fold change and is represented in this study by 187 genes. Starch and sucrose metabolism and biosynthesis of cofactors pathway share two genes encoding for UDP-glucose pyrophosphorylase uridylyltransferase, which is involved in glycogenesis and cell wall synthesis.

Cofactors generated through the biosynthesis cofactor pathway, including coenzyme A, NAD+, TPP and FAD are used in the next step of the AOP, the citrate cycle (TCA cycle), which has been found to be predictive of heart rate fold change in zebrafish embryos and is represented by 35 genes. The last two pathways of this cardiotoxicity AOP share four genes.

Two of those genes encode for transforming growth factor beta (1a, 1b), a cytokine that controls multiple processes during development through the activation of several signalling pathways including MAPK pathways (Chaudhury *et al.*, 2009). The other two genes are encoding ADP-ribosylation factor 2b, which enables GTP binding activity, and mitogen-activated protein kinase kinase kinase 14a, which enables MAP kinase activity. TCA cycle utilises various molecules, such as acetyl CoA, to generate GTP, ATP, NADH2 and FADH. The reducing equivalents (NADH2 and FADH) are then required for the electron transport chain for oxidative phosphorylation in the mitochondria to produce ATP. The heart demands high levels of ATP to maintain myocardial contraction and ion homeostasis, thus inadequate ATP production in myocardium caused by Krebs cycle regulation dysfunction, or NADH supply and activity of electron transport chain alterations leads to energy deprivation and potential apoptotic cell death (Giordano, 2005; Sheeran *et al.*, 2006; Doenst *et al.*, 2013).

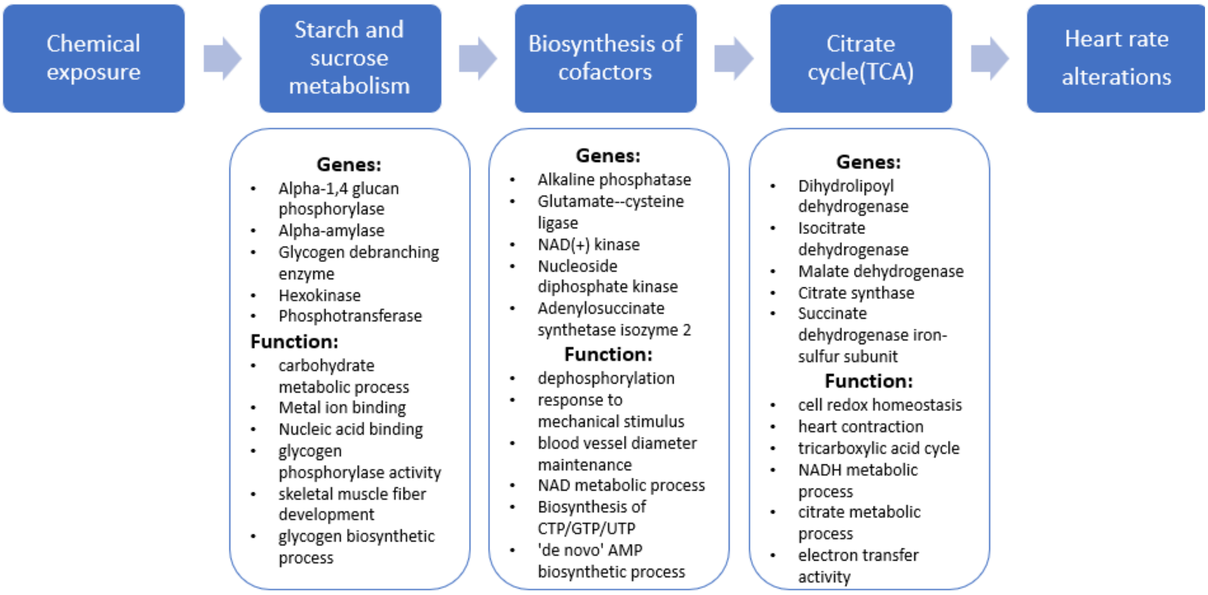

Chemical exposure
Starch and sucrose metabolism
Biosynthesis of cofactors
Citrate cycle(TCA)
Heart rate alterations
Genes:
• Alpha-1,4 glucan phosphorylase
• Alpha-amylase
• Glycogen debranching enzyme
• Hexokinase
• Phosphotransferase
Function:
• carbohydrate metabolic process
• Metal ion binding
• Nucleic acid binding
• glycogen phosphorylase activity
• skeletal muscle fiber development
• glycogen biosynthetic process
Genes:
• Alkaline phosphatase
• Glutamate--cysteine ligase
• NAD(+) kinase
• Nucleoside diphosphate kinase
• Adenylosuccinate synthetase isozyme 2
Function:
• dephosphorylation
• response to mechanical stimulus
• blood vessel diameter maintenance
• NAD metabolic process
• Biosynthesis of CTP/GTP/UTP
• 'de novo' AMP biosynthetic process
Genes:
• Dihydrolipoyl dehydrogenase
• Isocitrate dehydrogenase
• Malate dehydrogenase
• Citrate synthase
• Succinate dehydrogenase iron-sulfur subunit
Function:
• cell redox homeostasis
• heart contraction
• tricarboxylic acid cycle
• NADH metabolic process
• citrate metabolic process
• electron transfer activity

Figure 5.2: Potential AOP example 2 representation generated from cluster 3 chemicals. An adverse outcome pathway, starting from starch and sucrose metabolism, that is downstream linked to the biosynthesis of cofactors, is associated with the citrate cycle. Some of the genes and functions related to each pathway are also provided.

5.2 Chemical classification increases prediction ability of pathway analysis

Pathway analysis and predictive modelling of 143 chemicals, with high variability between gene expression profiles, failed to generate a model that can be predictive of chemical toxicity (LC50, experimental chemical concentration). On the other hand, utilising the 143 chemicals predictive modelling identified multiple models that to some extent can be predictive of heart rate fold changes after chemical exposure, showing that different chemicals can potentially alter zebrafish embryo's heart rate through multiple mechanisms and that a large number of biological processes are involved in cardiac development and function. However, pathway analysis indicates that cardiotoxicity (changes in zebrafish heart rate) and chemicals toxicity (LC50 or chemical concentration) can be, to some extent, predicted through pathway activity when the chemicals involved in predictive modelling share similar gene profiles (mRNA clustering from chapter 2). A popular method for predicting chemical toxicity is identifying their MoA based on structural features, such as Verhaar classification. However, the results of this study suggest that MoA classification is not representative of the effect chemicals have on heart rate in this study; chemicals that significantly alter the heart rate of zebrafish embryos are grouped in all Verhaar MoA classes.

The models generated using pathway activity (chapter 4) are relatively weak, especially after chemical clustering. The results of this study suggest that pathway activity can potentially be useful in predicting toxicity and heart rate fold change, and the use of more chemicals can increase the accuracy and predictive ability of those models.

5.3 Structural characteristics associated with chemical exposure and cardiotoxicity

The results of this study and the various structural relationship models (QSAR) published, have identified multiple molecular descriptors that are related to chemical toxicity, including chemical lipophilicity, molecular polarity, branching, bond nature, functional groups (amines) and molecular weight (Mansouri *et al.*, 2013; Ghorbanzadeh *et al.*, 2016; Lavado *et al.*, 2020). Lipophilic chemicals pass through the cell membrane easily, accumulate in the tissues and reach their target of toxicity (Verhaar *et al.*, 1992; Vaes *et al.*, 1998; Klüver *et al.*, 2019). QSAR studies have shown that the toxicity of inert chemicals can be predicted using only lipophilicity descriptors, but the toxicity of reactive chemicals is determined by chemical polarizability, the ability to interact with cellular molecules including nucleic acids and proteins (Lavado *et al.*, 2020). In addition, topological descriptors that represent the connections between adjacent atom pairs and provide information related to branching have been associated with developmental toxicity (Estrada, 1996; Mansouri *et al.*, 2013).

On the other hand, only a small number of studies have been looking into molecular descriptors associated with cardiotoxicity in zebrafish, showing that descriptors associated with lipophilicity and molecular weight can be used to predict ion currents changes involved in cardiac action potential generation (Wiśniowska *et al.*, 2015). In this study, similar descriptors were identified to be associated with heart rate fold change after exposure. However, in some cases, chemicals with similar structural characteristics act through different mechanisms, where only one of them may be toxic, thus gene and pathway analysis can be used to provide more information about the molecular mechanisms involved in chemical toxicity or cardiotoxicity and improve the associated AOP.

5.4 Gene ontology of genes related to chemical toxicity and cardiotoxicity

At the same time, high throughput sequencing has identified multiple genes, biological properties and pathways related to chemical toxicity and cardiotoxicity. The expression of multiple micro-RNA that are involved in regulating cellular functions and development processes were identified to be associated with chemical toxicity and cardiotoxicity. Most of the miRNAs identified in this study were both associated with increased chemical toxicity and cardiotoxicity, miR-126a, miR-216 and miR-155, associated with vascular integrity and development (Fish *et al.*, 2008; Cao *et al.*, 2016) and miR-499 that is involved in cardiac and muscle growth (Sluijter *et al.*, 2010; Wilson *et al.*, 2010; Fu *et al.*, 2011; Chistiakov *et al.*,2016).

Multiple biological processes, KEGG pathways and more miRNAs have been identified to be altered with increased chemical concentration (lower toxicity). MiR-30 regulates muscle phenotype, by controlling the Hedgehog signalling pathway (Ketley *et al.*, 2013) MiR-145 is expressed in vascular smooth muscle cells and controls cell death through regulation of apoptosis, cell proliferation, differentiation, and organ development (Yokoi *et al.*, 2009; J. Li *et al.*, 2020; Zhao *et al.*, 2020; Lin *et al.*, 2021) and miR-1 promotes embryonic muscle gene expression (Chen *et al.*, 2006; Mishima *et al.*, 2009). Finally, genes that are involved in heart left/right asymmetry determination, neuron differentiation and development, microtubule stability (Díaz-Martín *et al.*, 2021), protein degradation MAPK signalling, and cellular senescence pathways (Da Silva-Álvarez *et al.*, 2020) were also identified to be altered during chemical exposure. Cardiotoxicity in zebrafish was also mediated by miR- 206-3p, involved in muscle proliferation and differentiation (Kim *et al.*, 2006; Chen *et al.*, 2010; Goljanek-Whysall *et al.*, 2011; Lin *et al.*, 2017), miR-430 that regulates developmental pathways for cell movement (Liu *et al.*, 2020) and phosphorylation of AKT targets that are involved in muscle growth development was also associated with cardiotoxicity (Brunet *et al.*, 1999; Manning *et al.*, 2007).

5.5 Predictive modelling, Genes vs Pathway analysis

The dataset used in this study consists of highly diverse chemicals, but the experimental heart rate fold change could be predicted using molecular responses (genes and pathway analysis). Predictive modelling identified 80 genes related to cell communication, cardiac jelly development (Stankunas *et al.*, 2008; Lockhart *et al.*, 2011; Segert *et al.*, 2018), signalling pathways (Olson, 2006; Chi *et al.*, 2010), cardiomyocyte contractility, myocardium development (Radisic *et al.*, 2004; Auman *et al.*, 2007; Apaydin *et al.*, 2020), intracardiac hemodynamic flow (Hove *et al.*, 2003), regulation of immune response (Dong *et al.*, 2018; Qiu *et al.*, 2020) and cell death (Poelmann *et al.*, 2005; Pyati *et al.*, 2007; Zhang *et al.*, 2012; Lee *et al.*, 2014)(R^2=0.68). On the other hand, pathway analysis revealed that most of the available pathways can be used to predict heart rate fold change indicating that for proper heart development and function, a variety of genes and mechanisms are involved from multiple biological levels.

Clustering chemicals using mRNA profiles, reduce the diversity within the dataset, by grouping chemicals into three clusters. Predictive modelling using gene count profiles failed to generate accurate and reliable models for predicting heart rate in two out of the three clusters, (mRNA cluster). However, gene data were enough to predict the heart rate effect of chemicals from mRNA cluster 3, generating a model with 21 genes (R^2=0.64) associated with energy production by fatty acid oxidation (Waber *et al.*, 1982; Fu *et al.*,2013; Park *et al.*, 2021), acetylcholinesterase inhibitors that are associated with bradycardia (Watson *et al.*, 2014; Koenig *et al.*, 2016; Altenhofen *et al.*, 2019) and improper muscle development (McCollum *et al.*, 2011), various signalling pathways (O'Shea *et al.*, 2002; Yamashita *et al.*, 2002; Liu *et al.*, 2017) and nervous system development (Fedele *et al.*, 2020). In contrast, predictive modelling based on indices of pathway activity identified multiple pathways able to predict heart rate changes in all three clusters.

Generally, however, the gene-count based model across all chemicals outperformed the pathway-based cluster models. The results suggest that cardiotoxicity is associated with a wide spectrum of biological responses and interactions, thus reducing the dimensionality of the data into KEGG pathways leads to some gene (information) loss that is shared across all chemicals.

5.6 Application of pathway analysis in risk assessment

The large number of uncharacterised chemicals and the effort to move away from long and costly experiments, increase the need of utilising ‘omics and *in silico* approaches for chemical risk assessment. QSAR models have been widely used in risk assessment, however as it can be seen from the results of this study structural models although predictive of heart rate fold change, a prior classification is required reducing the applicability domain of the model. These results show that structural information is useful, but sometimes fails to be predictive of cardiotoxicity, since in some cases chemicals with similar structural features might act through different mechanisms, and vice versa (Russom *et al.*, 1997; Martin *et al.*, 2015; Ellison *et al.*, 2016), or the presence of cis and trans isomers were only one of them is toxic (Singh *et al.*, 1988; Blisard *et al.*, 1991).

The utilisation of ‘omics data can be used to overcome those limitations. ‘Omics data have been used to explore the molecular level changes and underlying cell biochemistry and physiology alterations. In this study, a small dataset was used (143 chemicals), characterised by high variability in chemical structure and gene expression profiles after chemical exposure, but predictive models were generated using both gene expression data and pathway analysis. A highly accurate and reliable model for the prediction of chemical-induced changes in zebrafish heart rate was generated using the gene expression profiles, but the models generated for predicting heart rate fold change using pathway activity were relatively weak, but this can be attributed to the small number of chemicals and the high variability between the gene expression profiles which increase the complexity of predictive modelling.

‘Omics data and pathway analysis can be used to identify potential key events and the relationship between them, through pathway networks, and potentially assist in the generation of AOPs. In addition, large-scale ‘omics data can be used in generating AOP networks that consist of two or more AOPs that share one or more key events, MIE or AO and offers a more realistic representation of the biological interaction underlying toxic exposure. Analysis of AOP network intersections can reveal unknown or unexpected biological connections and provide more information about the biological mechanism underlying chemical toxicity (Knapen *et al.*, 2018).

Reference

Abdel-Wahab, B.A. and Metwally, M.E. (2015) ‘Clozapine-Induced Cardiotoxicity: Role of Oxidative Stress, Tumour Necrosis Factor Alpha and NF-κβ’, *Cardiovascular toxicology*, 15(4), pp. 355–365.

Adhya, D., Annuario E., Lancaster M.A., Price J., Baron-Cohen S. and Srivastava D.P. (2018) ‘Understanding the role of steroids in typical and atypical brain development: Advantages of using a “brain in a dish” approach’, *Journal of neuroendocrinology*, 30(2). Available at: https://doi.org/10.1111/jne.12547.

Agatonovic-Kustrin, S., Morton, D.W. and Razic, S. (2014) ‘In silico modelling of pesticide aquatic toxicity’, *Combinatorial chemistry & high throughput screening*, 17(9), pp. 808–818.

Ahkin Chin Tai, J.K. and Freeman, J.L. (2020) ‘Zebrafish as an integrative vertebrate model to identify miRNA mechanisms regulating toxicity’, *Toxicology reports*, 7, pp. 559–570.

Ai, J., Zhang, R., Gao, X., Niu, H.F., Wang, N., Xu, Y., Li, Y., Ma, N., Sun, L.H., Pan, Z.W., Li, W.M. and Yang, B.F. (2012) ‘Overexpression of microRNA-1 impairs cardiac contractile function by damaging sarcomere assembly’, *Cardiovascular research*, 95(3), pp. 385–393.

Alexander-Dann, B., Pruteanu, L.L., Oerton, E., Sharma, N., Berindan-Neagoe, I., Modos, D. and Bender, A. (2018) ‘Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data’, *Molecular omics*, 14(4), pp. 218–236.

Altenhofen, S., Nabinger, D.D., Bitencourt, P.E.R. and Bonan, C.D. (2019) ‘Dichlorvos alters morphology and behavior in zebrafish (Danio rerio) larvae’, *Environmental pollution*, 245, pp. 1117–1123.

Amberg, A. (2013) ‘In Silico Methods’, in *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1273–1296.

Amjad, S., Rahman, M.S. and Pang, M.-G. (2020) ‘Role of Antioxidants in Alleviating Bisphenol A Toxicity’, *Biomolecules*, 10(8). Available at: https://doi.org/10.3390/biom10081105.

Anand, S. (2017) *Finding Optimal Number of Clusters*, *R-bloggers*. Available at: https://www.r-bloggers.com/2017/02/finding-optimal-number-of-clusters/ (Accessed: 7 November 2021).

Andersen, M.E., Thomas, R.S., Gaido, K.W. and Connolly, R.B.(2005) ‘Dose-response modeling in reproductive toxicology in the systems biology era’, *Reproductive toxicology*, 19(3), pp. 327–337.

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A. and Rai, A. (2016) ‘Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach’, *Journal of computational biology: a journal of computational molecular cell biology*, 23(4), pp. 239–247.

Ankley, G.T., Bencic, D.C., Breen, M.S., Colette, T.W., Connolly, R.B., Denslow, N.D., Edwards, S.W., Ekman, D.R., Garcia-Reyero, N., Jensen, M.J., Lazorchak, J.M., Martinovic, D., Miller, D.H., Perkins, E.J., Orlando, E.F., Vileneuve, D.L., Wang, R.L. and Watanabe, K.H. (2009) ‘Endocrine disrupting chemicals in fish: developing exposure indicators and predictive models of effects based on mechanism of action’, *Aquatic toxicology*, 92(3), pp. 168–178.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W., Johnson, R.D., Mountm D.R., Nichols, J.W., Russom, C.L., Schmieder, P.K., Serraro, J.A., Tietge, J.E. and Villeneuve, D.L (2010) ‘Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment’, *Environmental toxicology and chemistry / SETAC*, 29(3), pp. 730–741.

Antczak, P., White, T.A., Giri, A., Michelangeli, F., Viant, M.R., Cronin, M.T.D., Vulpe, C. and Falciani, F. (2015) ‘Systems Biology Approach Reveals a Calcium-Dependent Mechanism for Basal Toxicity in Daphnia magna’, *Environmental science & technology*, 49(18), pp. 11132–11140.

Anuta, V, Nitulescu, G.M., Dinu-Pîrvu, C.E. and Olaru, O.T. (2014) ‘Biopharmaceutical profiling of new antitumor pyrazole derivatives’, *Molecules*, 19(10), pp. 16381–16401.

Apaydin, D.C., Jaramillo, P.A.M., Corradi, L., Cosco, F., Rathjen, F.G., Kammertoens, T., Filosa, A. and Sawamiphak, S (2020) ‘Early-Life Stress Regulates Cardiac Development through an IL-4-Glucocorticoid Signaling Balance’, *Cell reports*, 33(7), p. 108404.

Apte, R.S., Chen, D.S. and Ferrara, N. (2019) ‘VEGF in Signaling and Disease: Beyond Discovery and Development’, *Cell*, 176(6), pp. 1248–1264.

Aptula, A.O. and Roberts, D.W. (2006) ‘Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity’, *Chemical research in toxicology*, 19(8), pp. 1097–1105.

Ashauer, R., Hintermeister, A., Caravatti, I., Kretschmann, A. and Escher, B. (2010) ‘Toxicokinetic and toxicodynamic modeling explains carry-over toxicity from exposure to diazinon by slow organism recovery’, *Environmental science & technology*, 44(10), pp. 3963–3971.

Ashburner, M. Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G.(2000) ‘Gene ontology: tool for the unification of biology. The Gene Ontology Consortium’, *Nature genetics*, 25(1), pp. 25–29.

Auman, H.J., Coleman, H., Riley, H.E., Olale, F., Tsai, H.J. and Yelon, D. (2007) ‘Functional modulation of cardiac form through regionally confined cell shape changes’, *PLoS biology*, 5(3), p. e53.

Balasubramanian, S., Gunasekaran, K., Sasidharan, S., Jeyamanickavel Mathan, V. and Perumal, E. (2020) ‘MicroRNAs and Xenobiotic Toxicity: An Overview’, *Toxicology reports*, 7, pp. 583–595.

Baldim, J.L., Alcântara, B.G.V., Domingos, O.S., Soares, M.G., Caldas, I.S., Novaes, R.D., Oliveira, T.B., Lago, J.H.G. and Chagas-Paula, D.A.(2017) ‘The Correlation between Chemical Structures and Antioxidant, Prooxidant, and Antitrypanosomatid Properties of Flavonoids’, *Oxidative medicine and cellular longevity*, 2017, p. 3789856.

Balta, E., Kramer, J. and Samstag, Y. (2020) ‘Redox Regulation of the Actin Cytoskeleton in Cell Migration and Adhesion: On the Way to a Spatiotemporal View’, *Frontiers in cell and developmental biology*, 8, p. 618261.

Bambino, K. and Chu, J. (2017) ‘Zebrafish in Toxicology and Environmental Health’, *Current topics in developmental biology*, 124, pp. 331–367.

Baraldi, A. and Blonda, P. (1999) ‘A survey of fuzzy clustering algorithms for pattern recognition. I’, *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*, 29(6), pp. 778–785.

Barioni, M.C.N., Razente, H., Marcelino, A.M.R., Traina, A.J.M. and Traina Jr, C.(2014) ‘Open issues for partitioning clustering methods: an overview’, *Wiley interdisciplinary reviews. Data mining and knowledge discovery*, 4(3), pp. 161–177.

Barrionuevo, W.R. and Burggren, W.W. (1999) ‘O_2consumption and heart rate in developing zebrafish (*Danio rerio*): influence of temperature and ambient O_2’, *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, pp. R505–R513. Available at: https://doi.org/10.1152/ajpregu.1999.276.2.r505.

Basak, S.C., Niemi, G.J. and Veith, G.D. (1991) ‘Predicting properties of molecules using graph invariants’, *Journal of mathematical chemistry*, 7(1), pp. 243–272.

Bearden, A.P. and Schultz, T.W. (1998) ‘Comparison of *Tetrahymena* and *Pimephales* Toxicity Based on Mechanism of Action’, *SAR and QSAR in Environmental Research*, pp. 127–153. Available at: https://doi.org/10.1080/10629369808039153.

Behra, M., Cousin, X., Bertrand, C., Vonesch, J.L., Biellmann, D., Chatonnet, A. and Strahle, U. (2002) ‘Acetylcholinesterase is required for neuronal and muscular development in the zebrafish embryo’, *Nature neuroscience*, 5(2), pp. 111–118.

Bezdek, J.C., Ehrlich, R. and Full, W. (1984) ‘FCM: The fuzzy c-means clustering algorithm’, *Computers & geosciences*, 10(2-3), pp. 191–203.

Binukumar, B.K., Bal, A., Kandimalla, R.J.L. and Gill, K.D. (2010) ‘Nigrostriatal neuronal death following chronic dichlorvos exposure: crosstalk between mitochondrial impairments, α synuclein aggregation, oxidative damage and behavioral changes’, *Molecular brain*, 3, p. 35.

Bittremieux, W., Advani, R.S., Jarmusch, A.K., Aguirre, S., Lu, A., Dorrestein, P.C. and Tsunoda, S.M. (2022) ‘Physicochemical properties determining drug detection in skin’, *Clinical and translational science*, 15(3), pp. 761–770.

Blagden, C.S., Currie, P.D., Ingham, P.W. and Hughes, S.M. (1997) ‘Notochord induction of zebrafish slow muscle mediated by Sonic hedgehog’, *Genes & development*, 11(17), pp. 2163–2175.

Blankesteijn, W.M. (2020) ‘Interventions in WNT Signaling to Induce Cardiomyocyte

Proliferation: Crosstalk with Other Pathways', *Molecular pharmacology*, 97(2), pp. 90–101.

Blisard, K.S., Harrington, D.A., Long, D.A. and Jackson, J.E (1991) 'Relative lack of toxicity of transplatin compared with cisplatin in rodents', *Journal of comparative pathology*, 105(4), pp. 367–375.

Boethling, R.S. (1996) 'Designing Biodegradable Chemicals', in *ACS Symposium Series*. Washington, DC: American Chemical Society (ACS symposium series. American Chemical Society), pp. 156–171.

Bogoyevitch, M.A. and Kobe, B. (2006) 'Uses for JNK: the many and varied substrates of the c-Jun N-terminal kinases', *Microbiology and molecular biology reviews: MMBR*, 70(4), pp. 1061–1095.

Bornhorst, D., Xia, P., Nakajima, H., Dingare, C., Herzog, W., Lecaudey, V., Mochizuki, N., Heisenberg, C.P., Yelon, D. and Abdelilah-Seyfried, S. (2019) 'Biomechanical signaling within the developing zebrafish heart attunes endocardial growth to myocardial chamber dimensions', *Nature communications*, 10(1), p. 4113.

Böttcher, R.T. and Niehrs, C. (2005) 'Fibroblast growth factor signaling during early vertebrate development', *Endocrine reviews*, 26(1), pp. 63–77.

Bourdon-Lacombe, J.A., Moffat, I.D., Deveau, M., Husain, M., Auerbach, S., Krewski, D., Thomas, R.S., Bushel, P.R., Williams, A. and Yauk, C.L.,(2015) 'Technical guide for applications of gene expression profiling in human health risk assessment of environmental chemicals', *Regulatory toxicology and pharmacology: RTP*, 72(2), pp. 292–309.

Bozdogan, H. (2000) 'Akaike's Information Criterion and Recent Developments in Information Complexity', *Journal of Mathematical Psychology*, pp. 62–91. Available at: https://doi.org/10.1006/jmps.1999.1277.

Breiman, L. (2001) *Machine learning*, 45(1), pp. 5–32.

Brockmeier, E.K., Hodges, G., Hutchinson, T.H., Butler, E., Hecker, M., Tollefsen, K.E., Garcia-Reyero, N., Kille, P., Becker, D., Chipman, K., Colbourne, J., Collette, T.W., Cossins, A., Cronin, M., Graystock, P., Gutsell, S., Knapen, D., Katsiadaki, I., Lange, A., Marshall, S., Owen, S.F., Perkins, E.J., Plaistow, S., Schroeder, A., Taylor, D., Viant, M., Anklet, G., and Falciani, F (2017) 'The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment', *Toxicological sciences: an official journal of the Society of Toxicology*, 158(2), pp. 252–262.

Brosnan, J.T. (1999) 'Comments on metabolic needs for glucose and the role of gluconeogenesis', *European journal of clinical nutrition*, 53 Suppl 1, pp. S107–11.

Brunet, A., Bonni, A., Zigmond, M.J., Lin, M.Z., Juo, P., Hu, L.S., Anderson, M.J., Arden, K.C., Blenis, J. and Greenberg, M.E. (1999) 'Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor', *Cell*, 96(6), pp. 857–868.

Bu, H., Ding, Y., Li, J., Zhu, P., Shih, Y.H., Wang, M., Zhang, Y., Lin, X. and Xu, X. (2021) 'Inhibition of mTOR or MAPK ameliorates vmhcl/myh7 cardiomyopathy in zebrafish', *JCI insight*, 6(24). Available at: https://doi.org/10.1172/jci.insight.154215.

Burden, F.R. (1989) 'Molecular identification number for substructure searches', *Journal of chemical information and modeling*, 29(3), pp. 225–227.

Bussmann, J.Lawson, N., Zon, L. and Schulte-Merker, S. (2008) ‘Zebrafish VEGF receptors: a guideline to nomenclature’, *PLoS genetics*, 4(5), p. e1000064.

Caballero, M.V. and Candiracci, M. (2018) ‘Zebrafish as Toxicological model for screening and recapitulate human diseases’, *Journal of unexplored medical data*, 3(2), p. 4.

Cai, C.,Guo, P., Zhou, Y., Zhou, J., Wang. Q., Zhang, F., Fang, J. and Cheng, F. (2019) ‘Deep learning-based prediction of drug-induced cardiotoxicity’, *Journal of chemical information and modeling*, 59(3), pp. 1073–1084.

Cao, R.Y., Li, Q., Miao, Y., Zhang, Y., Yuan, W., Fan, L., Liu, G., Mi, Q. and Yang, J. (2016) ‘The Emerging Role of MicroRNA-155 in Cardiovascular Diseases’, *BioMed research international*, 2016, p. 9869208.

Carnesecchi, E., Toma, C., Roncaglioni, A., Kramer, N., Benfenati, E. and Dorne, J.L.C.M. (2020) ‘Integrating QSAR models predicting acute contact toxicity and mode of action profiling in honey bees (A. mellifera): Data curation using open source databases, performance testing and validation’, *The Science of the total environment*, 735, p. 139243.

Carrillo-Salinas, F.J., Ngwenyama, N., Anastasiou, M., Kaur, K. and Alcaide, P. (2019) ‘Heart Inflammation: Immune Cell Roles and Roads to the Heart’, *The American journal of pathology*, 189(8), pp. 1482–1494.

Carrió, P., Pinto, M., Ecker, G., Sanz, F. and Pastor, M. (2014) ‘Applicability Domain ANalysis (ADAN): a robust method for assessing the reliability of drug property predictions’, *Journal of chemical information and modeling*, 54(5), pp. 1500–1511.

Cassotti, M., Ballabio, D., Todeschini, R. and Consonni, V.(2015) ‘A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas)’, *SAR and QSAR in environmental research*, 26(3), pp. 217–243.

Chang, C.-Y., Hsu, M.T., Esposito, E.X. and Tseng, Y.J. (2013) ‘Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods’, *Journal of chemical information and modeling*, 53(4), pp. 958–971.

Chang, H.H.Y., Pannunzio, N.R., Adachi, N. and Lieber, M.R. (2017) ‘Non-homologous DNA end joining and alternative pathways to double-strand break repair’, *Nature reviews. Molecular cell biology*, 18(8), pp. 495–506.

Chapman, N.A., Dupré, D.J. and Rainey, J.K. (2014) ‘The apelin receptor: physiology, pathology, cell signalling, and ligand modulation of a peptide-activated class A GPCR’, *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 92(6), pp. 431–440.

Chaudhury, A. and Howe, P.H. (2009) ‘The tale of transforming growth factor-beta (TGFbeta) signaling: a soigné enigma’, *IUBMB life*, 61(10), pp. 929–939.

Chavan, S., Nicholls, I.A., Karlsson, B.C., Rosengren, A.M., Ballabio, D., Consonni, V. and Todeschini, R. (2014) ‘Towards Global QSAR Model Building for Acute Toxicity: Munro Database Case Study’, *International journal of molecular sciences*, 15(10), p. 18162.

ChemSpider (2018). Available at: http://www.chemspider.com/ (Accessed: 7 February 2018).

Chen, J. (2013) ‘Impaired cardiovascular function caused by different stressors elicits a common pathological and transcriptional response in zebrafish embryos’, *Zebrafish*, 10(3), pp. 389–400.

Chen, J.F., Mandel, E.M., Thomson, J.M., Wu, Q., Callus, T.E., Hammond, S.M., Conlon, F.L. and Wang, D.Z. (2006) ‘The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation’, *Nature genetics*, 38(2), pp. 228–233.

Chen, J.F., Mandel, E.M., Thomson, J.M., Wu, Q., Callus, T.E., Hammond, S.M., Conlon, F.L. and Wang, D.Z. (2010) ‘microRNA-1 and microRNA-206 regulate skeletal muscle satellite cell proliferation and differentiation by repressing Pax7’, *The Journal of cell biology*, 190(5), pp. 867–879.

Chen, L. and Knowlton, A.A. (2010) ‘Mitochondria and heart failure: new insights into an energetic problem’, *Minerva cardioangiologica*, 58(2), pp. 213–229.

Chen, M., Ma, G., Yue, Y., Wei, Y., Li, Q., Tong, Z., Zhang, L., Miao, G. and Zhang, J. (2014) ‘Downregulation of the miR-30 family microRNAs contributes to endoplasmic reticulum stress in cardiac muscle and vascular smooth muscle cells’, *International journal of cardiology*, 173(1), pp. 65–73.

Chen, P.Y., Manninga, H., Slanchev, K., Chien, M., Russo, J.J., Ju, J., Sheridan, R., John, B., Marks, D.S., Giadatzis, D., Sander, C., Zavolan, M. and Tuschi, T. (2005) ‘The developmental miRNA profiles of zebrafish as determined by small RNA cloning’, *Genes & development*, 19(11), pp. 1288–1293.

Chen, Y., Amende, I., Hampton, T.G., Yang, Y., Ke, Q., Min, J.Y., Xiao, Y.F. and Morgan, J.P. (2006) ‘Vascular endothelial growth factor promotes cardiomyocyte differentiation of embryonic stem cells’, *American journal of physiology. Heart and circulatory physiology*, 291(4), pp. H1653–8.

Cherkasov, A.,Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuzmin, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A. and Tropsha, A. (2014) ‘QSAR modeling: where have you been? Where are you going to?’, *Journal of medicinal chemistry*, 57(12), pp. 4977–5010.

Chi, N.C., Bussen, M., Brand-Arzamendi, K., Ding, C., Olgin, J.E., Shaw, R.M., Martin, G.R. and Stainier, D.Y.R. (2010) ‘Cardiac conduction is required to preserve cardiac chamber morphology’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(33), pp. 14662–14667.

Chistiakov, D.A., Orekhov, A.N. and Bobryshev, Y.V. (2016) ‘Cardiac-specific miRNA in cardiogenesis, heart function, and cardiac pathology (with focus on myocardial infarction)’, *Journal of molecular and cellular cardiology*, 94, pp. 107–121.

Chowdhury, M.Z.I. and Turin, T.C. (2020) ‘Variable selection strategies and its importance in clinical prediction modelling’, *Family medicine and community health*, 8(1), p. e000262.

Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A. and Wang, Y.

(2008) ‘The properties of high-dimensional data spaces: implications for exploring gene and protein expression data’, *Nature reviews. Cancer*, 8(1), pp. 37–49.

Cohen, E.D., Tian, Y. and Morrisey, E.E. (2008) ‘Wnt signaling: an essential regulator of cardiovascular differentiation, morphogenesis and progenitor self-renewal’, *Development*, 135(5), pp. 789–798.

Colombo, A., Orsi, F. and Bonfanti, P. (2005) ‘Exposure to the organophosphorus pesticide chlorpyrifos inhibits acetylcholinesterase activity and affects muscular integrity in Xenopus laevis larvae’, *Chemosphere*, 61(11), pp. 1665–1671.

Cooper, G.M. (2000) ‘Lysosomes’, in *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates.

Copeland, W.C. and Longley, M.J. (2014) ‘Mitochondrial genome maintenance in health and disease’, *DNA repair*, 19, pp. 190–198.

Corrales, J., Kristofco, L.A., Steele, W.B., Yates, B.S., Breed, C.S., Williams, E.S. and Brooks, B.W. (2015) ‘Global Assessment of Bisphenol A in the Environment: Review and Analysis of Its Occurrence and Bioaccumulation’, *Dose-response: a publication of International Hormesis Society*, 13(3), p. 1559325815598308.

Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017) ‘RNA-Seq differential expression analysis: An extended review and a software tool’, *PloS one*, 12(12), p. e0190152.

Czirok, A., Zach, J., Kozel, B.A., Mecham, R.P., Davis, E.C. and Rongish, B.J. (2006) ‘Elastic fiber macro-assembly is a hierarchical, cell motion-mediated process’, *Journal of cellular physiology*, 207(1), pp. 97–106.

Dai, Y.J., Jia, Y.F., Chen, N., Bian, W.P., Li, Q.K., Ma, Y.B., Chen, Y.L. and Pei, D.S. (2014) ‘Zebrafish as a model system to study toxicology’, *Environmental toxicology and chemistry / SETAC*, 33(1), pp. 11–17.

Dal Lin, C., Tona, F. and Osto, E. (2019) ‘The crosstalk between the cardiovascular and the immune system’, *Vascular biology (Bristol, England)*, 1(1), pp. H83–H88.

Darnell, J.E., Jr (1997) ‘STATs and gene regulation’, *Science*, 277(5332), pp. 1630–1635.

Da Silva-Álvarez, S., Guerra-Varela, J., Sobrido-Cameán, D., Quelle, A., Barreiro-Iglesias, A., Sanchez, L. and Collado, M. (2020) ‘Cell senescence contributes to tissue regeneration in zebrafish’, *Aging cell*, 19(1), p. e13052.

Davis, A.J. and Chen, D.J. (2013) ‘DNA double strand break repair via non-homologous end-joining’, *Translational cancer research*, 2(3), pp. 130–143.

Decker, T. (1999) ‘Introduction: STATs as essential intracellular mediators of cytokine responses’, *Cellular and molecular life sciences: CMLS*, 55(12), pp. 1505–1508.

Degenhardt, F., Seifert, S. and Szymczak, S. (2019) ‘Evaluation of variable selection methods for random forests and omics data sets’, *Briefings in bioinformatics*, 20(2), pp. 492–503.

Den Broeder, M.J., Kopylova, V.A., Kamminga, L.M. and Legler, J. (2015) ‘Zebrafish as a Model to Study the Role of Peroxisome Proliferating-Activated Receptors in

Adipogenesis and Obesity', *PPAR research*, 2015, p. 358029.

Denslow, N.D., Garcia-Reyero, N. and Barber, D.S. (2007) 'Fish "n" chips: the use of microarrays for aquatic toxicology', *Molecular bioSystems*, 3(3), pp. 172–177.

Desler, C., Lykke, A. and Rasmussen, L.J. (2010) 'The effect of mitochondrial dysfunction on cytosolic nucleotide metabolism', *Journal of nucleic acids*, 2010. Available at: https://doi.org/10.4061/2010/701518.

Dhall, D., Kaur, R. and Juneja, M. (2020) 'Machine learning: A review of the algorithms and its applications', in *Lecture Notes in Electrical Engineering*. Cham: Springer International Publishing (Lecture notes in electrical engineering), pp. 47–63.

Díaz-Martín, R.D., Valencia-Hernández, J.D., Betancourt-Lozano, M. and Yáñez-Rivera, B. (2021) 'Changes in microtubule stability in zebrafish () embryos after glyphosate exposure', *Heliyon*, 7(1), p. e06027.

Diehl, F.F., Miettinen, T.P., Elbashir R, Nabel, C.S., Manalis, Lewis, C.A. and Heiden, M.G.V (2021) 'Nucleotide imbalance decouples cell growth from cell proliferation', *bioRxiv*. Available at: https://doi.org/10.1101/2021.12.06.471399.

Dobin, A. (2019) *STAR manual 2.7.0a*. Available at: https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf (Accessed: 2019).

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, N. and Gingeras, T.R. (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics* , 29(1), pp. 15–21.

Doenst, T., Nguyen, T.D. and Abel, E.D. (2013) 'Cardiac metabolism in heart failure: implications beyond ATP production', *Circulation research*, 113(6), pp. 709–724.

Domingues, I., Oliveira, R., Lourenço, J., Grisolia, C.K., Mendo, S. and Soares, A.M.V.M. (2010) 'Biomarkers as a tool to assess effects of chromium (VI): comparison of responses in zebrafish early life stages and adults', *Comparative biochemistry and physiology. Toxicology & pharmacology: CBP*, 152(3), pp. 338–345.

Domingues, I. Oliveira, R., Musso, C., Cardoso, M., Soares, A.M.V.M. and Loureiro, S. (2013) 'Prochloraz effects on biomarkers activity in zebrafish early life stages and adults', *Environmental toxicology*, 28(3), pp. 155–163.

Dong, X., Wang, C., Zhang, J., Wang, S., Li, H., Kang, Y., Tian, S. and Fu, L. (2017) 'Cholecystokinin Expression in the Development of Postinfarction Heart Failure', *Cellular physiology and biochemistry: international journal of experimental cellular physiology, biochemistry, and pharmacology*, 43(6), pp. 2479–2488.

Dong, X., Zhang, Z., Meng, S., Pan, C., Yang, M., Wu, X., Yang, L. and Xu, H. (2018) 'Parental exposure to bisphenol A and its analogs influences zebrafish offspring immunity', *The Science of the total environment*, 610-611, pp. 291–297.

Doucet, J.P. and Doucet Panaye, A. (2018) 'Quantitative structure activity relationships for carboxamides and related compounds active on Aedes aegypti adult females', *Vector biology journal*, 03(01). Available at: https://doi.org/10.4172/2473-4810.1000127.

Dragon - Talete srl (2018). Available at: http://www.talete.mi.it/products/dragon_description.htm.

Drgan, V., Župerl, Š., Vračko, M., Como, F. and Novič, M. (2016) 'Robust modelling of acute toxicity towards fathead minnow (Pimephales promelas) using counter-propagation artificial neural networks and genetic algorithm', *SAR and QSAR in environmental research*, 27(7), pp. 501–519.

Dubińska-Magiera, M., Daczewska, M., Lewicka, A., Migocka-Patrzałek, M., Niedbalska-Tarnowska, J. and Jagla, K. (2016) 'Zebrafish: A Model for the Study of Toxicants Affecting Muscle Development and Function', *International journal of molecular sciences*, 17(11). Available at: https://doi.org/10.3390/ijms17111941.

Dunwoodie, S.L. (2007) 'Combinatorial signaling in the heart orchestrates cardiac induction, lineage specification and chamber formation', *Seminars in cell & developmental biology*, 18(1), pp. 54–66.

Du, S.J., Devoto, S.H., Westerfield, M. and Moon, R.T. (1997) 'Positive and negative regulation of muscle cell identity by members of the hedgehog and TGF-beta gene families', *The Journal of cell biology*, 139(1), pp. 145–156.

Efron, B. (1979) 'Bootstrap methods: Another look at the jackknife', *Annals of statistics*, 7(1), pp. 1–26.

Ellison, C.M., Piechota, P., Madden, J.C., Enoch, S.J. and Cronin Mark, T.D. (2016) 'Adverse Outcome Pathway (AOP) Informed Modeling of Aquatic Toxicology: QSARs, Read-Across, and Interspecies Verification of Modes of Action', *Environmental science & technology*, 50(7), pp. 3995–4007.

Elsayad, A.M., Nassef, AM., Al-Dhaifallah, M. and Elsayad, K.A. (2020) 'Classification of Biodegradable Substances Using Balanced Random Trees and Boosted C5.0 Decision Trees', *International journal of environmental research and public health*, 17(24). Available at: https://doi.org/10.3390/ijerph17249322.

EMBL-EBI (2021) *QuickGO*. Available at: https://www.ebi.ac.uk/QuickGO/ (Accessed: 19 February 2021).

Enoch, S.J., Hewitt, M., Cronin, M.T., Azam, S. and Madden, J.C. (2008) 'Classification of chemicals according to mechanism of aquatic toxicity: an evaluation of the implementation of the Verhaar scheme in Toxtree', *Chemosphere*, 73(3), pp. 243–248.

Epelman, S., Liu, P.P. and Mann, D.L. (2015) 'Role of innate and adaptive immune mechanisms in cardiac injury and repair', *Nature reviews. Immunology*, 15(2), pp. 117–129.

Estrada, E. (1996) 'Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes', *Journal of chemical information and computer sciences*, 36(4), pp. 844–849.

Ets, H.K., Seow, C.Y. and Moreland, R.S. (2016) 'Sustained Contraction in Vascular Smooth Muscle by Activation of L-type Ca2 Channels Does Not Involve Ca2 Sensitization or Caldesmon', *Frontiers in Pharmacology*. Available at: https://doi.org/10.3389/fphar.2016.00516.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. and D'Eustachio, P. (2018) 'The Reactome Pathway Knowledgebase', *Nucleic acids research*, 46(D1), pp. D649–D655.

Fan, T. , Sun, G., Zhao, L., Cui, X. and Zhong, R. (2018) 'QSAR and Classification Study on Prediction of Acute Oral Toxicity of -Nitroso Compounds', *International journal of molecular sciences*, 19(10). Available at: https://doi.org/10.3390/ijms19103015.

Faraoni, I., Antonetti, F.R., Cardone, J. and Bonmassar, E (2009) 'miR-155 gene: a typical multifunctional microRNA', *Biochimica et biophysica acta*, 1792(6), pp. 497–505.

Fasullo, M. and Endres, L. (2015) 'Nucleotide salvage deficiencies, DNA damage and neurodegeneration', *International journal of molecular sciences*, 16(5), pp. 9431–9449.

Fedele, L. and Brand, T. (2020) 'The Intrinsic Cardiac Nervous System and Its Role in Cardiac Pacemaking and Conduction', *Journal of cardiovascular development and disease*, 7(4). Available at: https://doi.org/10.3390/jcdd7040054.

FFmpeg (2018). Available at: http://ffmpeg.org/ (Accessed: 23 November 2018).

Fish, J.E., Santoro, M.M., Morton, S.U., Yu, S., Yeh, R.F., Wythe, J.D., Ivey, K.N., Bruneau, B.G., Stainier D.Y.R. and Srivastava, D. (2008) 'miR-126 regulates angiogenic signaling and vascular integrity', *Developmental cell*, 15(2), pp. 272–284.

Forrest, J., Bazylewski1, P., Bauer1, R., Hong2, S., Kim, C.Y., Giesy. J.P., Khim2, J.S. and Chang, G.S. (2014) 'A comprehensive model for chemical bioavailability and toxicity of organic chemicals based on first principles', *Frontiers in Marine Science*. Available at: https://doi.org/10.3389/fmars.2014.00031.

Fountoulaki, K., Dagres, N. and Iliodromitis, E.K. (2015) 'Cellular Communications in the Heart', *Cardiac failure review*, 1(2), pp. 64–68.

Frangogiannis, N.G. (2019) 'The Extracellular Matrix in Ischemic and Nonischemic Heart Failure', *Circulation research*, 125(1), pp. 117–146.

Fraysse, B., Mons, R. and Garric, J. (2006) 'Development of a zebrafish 4-day embryo-larval bioassay to assess toxicity of chemicals', *Ecotoxicology and environmental safety*, 63(2), pp. 253–267.

Fu, J.-D., Rushing, S.N., Lieu, D.K., Chan, C.W., Kong, C.-W., Geng, L., Wilson, K.D., Chiamvimonvat, N., Boheler, K.R., Wu, J.C., Keller, G., Hajjar, R.J. and Li, R.A. (2011) 'Distinct roles of microRNA-1 and -499 in ventricular specification and functional maturation of human embryonic stem cell-derived cardiomyocytes', *PloS one*, 6(11), p. e27417.

Fu, L., Huang, M. and Chen, S. (2013) 'Primary carnitine deficiency and cardiomyopathy', *Korean circulation journal*, 43(12), pp. 785–792.

Fumagalli, M., Lecca, D., Abbracchio, M.P. and Ceruti, S. (2017) 'Pathophysiological role of purines and pyrimidines in neurodevelopment: Unveiling new pharmacological approaches to congenital brain diseases', *Frontiers in pharmacology*, 8, p. 941.

Futran Fuhrman, V., Tal, A. and Arnon, S. (2015) ‘Why endocrine disrupting chemicals (EDCs) challenge traditional risk assessment and how to respond’, *Journal of hazardous materials*, 286, pp. 589–611.

Gadaleta, D., Vuković, K., Toma, C., Lavado, G.J., Karmaus, A.L., Mansouri, K., Kleinstreuer, N.C., Benfenati, E. and Roncaglioni, A. (2019) ‘SAR and QSAR modeling of a large collection of LD rat acute oral toxicity data’, *Journal of cheminformatics*, 11(1), p. 58.

Gajewicz-Skretna, A., Gromelski, M., Wyrzykowska, E., Furuhama, A., Yamamoto, H. and Suzuki, N. (2021) ‘Aquatic toxicity (Pre)screening strategy for structurally diverse chemicals: global or local classification tree models?’, *Ecotoxicology and environmental safety*, 208, p. 111738.

Galvez, J., Garcia, R., Salabert, M.T. and Soler, R. (1994) ‘Charge indexes. New topological descriptors’, *Journal of chemical information and computer sciences*, 34(3), pp. 520–525.

García-Campos, M.A., Espinal-Enríquez, J. and Hernández-Lemus, E. (2015) ‘Pathway Analysis: State of the Art’, *Frontiers in physiology*, 6, p. 383.

Garcia-Reyero, N. and Perkins, E.J. (2011) ‘Systems biology: leading the revolution in ecotoxicology’, *Environmental toxicology and chemistry / SETAC*, 30(2), pp. 265–273.

Garreta, E., Prado, P., Izpisua Belmonte, J.C. and Montserrat, N. (2017) ‘Non-coding microRNAs for cardiac regeneration: Exploring novel alternatives to induce heart healing’, *Non-coding RNA research*, 2(2), pp. 93–99.

Gaudêncio, S.P. and Pereira, F. (2022) ‘Predicting Antifouling Activity and Acetylcholinesterase Inhibition of Marine-Derived Compounds Using a Computer-Aided Drug Design Approach’, *Marine drugs*, 20(2). Available at: https://doi.org/10.3390/md20020129.

Geary, R.C. (1954) ‘The contiguity ratio and statistical mapping’, *The incorporated statistician*, 5(3), p. 115.

Georgiadis, N., Tsarouhas, K., Tsitsimpikou, C., Vardavas, A., Rezaee, R., Germanakis, I., Tsatsakis, A., Stagos, D. and Kouretas, D. (2018) ‘Pesticides and cardiotoxicity. Where do we stand?’, *Toxicology and applied pharmacology*, 353, pp. 1–14.

Ghorbanzadeh, M., Zhang, J. and Andersson, P.L. (2016) ‘Binary classification model to predict developmental toxicity of industrial chemicals in zebrafish’, *Journal of chemometrics*, 30(6), pp. 298–307.

Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J. (1999) ‘A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases’, *Journal of combinatorial chemistry*, 1(1), pp. 55–68.

Giordano, F.J. (2005) ‘Oxygen, oxidative stress, hypoxia, and heart failure’, *The Journal of clinical investigation*, 115(3), pp. 500–508.

Glazko, G.V. and Emmert-Streib, F. (2009) ‘Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets’, *Bioinformatics*, 25(18), pp. 2348–2354.

Gleeson, M.P. (2008) ‘Generation of a set of simple, interpretable ADMET rules of thumb’, *Journal of medicinal chemistry*, 51(4), pp. 817–834.

Goldberg, I.J., Trent, C.M. and Schulze, P.C. (2012) ‘Lipid metabolism and toxicity in the heart’, *Cell metabolism*, 15(6), pp. 805–812.

Goldsmith, Z.G. and Dhanasekaran, D.N. (2007) ‘G protein regulation of MAPK networks’, *Oncogene*, 26(22), pp. 3122–3142.

Goljanek-Whysall, K., Sweetman, D., Abu-Elmagd, M., Chapnik, E., Dalmay, T., Hornstein, E. and Münsterberg, A. (2011) ‘MicroRNA regulation of the paired-box transcription factor Pax3 confers robustness to developmental timing of myogenesis’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(29), pp. 11936–11941.

Goodenough, D.A. and Paul, D.L. (2009) ‘Gap junctions’, *Cold Spring Harbor perspectives in biology*, 1(1), p. a002576.

Gordan, R., Gwathmey, J.K. and Xie, L.H. (2015) ‘Autonomic and endocrine control of cardiovascular function’, *World journal of cardiology*, 7(4), pp. 204–214.

Gordon, M.K. and Hahn, R.A. (2010) ‘Collagens’, *Cell and tissue research*, 339(1), pp. 247–257.

Gosain, A. and Dahiya, S. (2016) ‘Performance analysis of various fuzzy clustering algorithms: A review’, *Procedia computer science*, 79, pp. 100–111.

Grandi, E., Sanguinetti, M.C., Bartos, D.C., Bers, D.M., Chen-Izu, Y., Chiamvimonvat, N., Colecraft, H.M., Delisle, B.P., Heijman, J., Navedo, M.F., Noskov, S., Proenza, C., Vandenberg, J.I. and Yarov-Yarovoy, V. (2017) ‘Potassium channels in the heart: structure, function and regulation’, *The Journal of physiology*, 595(7), pp. 2209–2228.

Grandini, M., Bagli, E. and Visani, G. (2020) ‘Metrics for multi-class classification: An overview’. Available at: https://doi.org/10.48550/ARXIV.2008.05756.

Greener, J.G., Kandathil, S.M., Moffat, L. and Jones, D.T. (2022) ‘A guide to machine learning for biologists’, *Nature reviews. Molecular cell biology*, 23(1), pp. 40–55.

Grogg, M.W., Braydich-Stolle, L.K., Maurer-Gardner, E.I., Hill, N.T., Sakaram, S., Kadakia, M.P. and Hussain, S.M. (2016) ‘Modulation of miRNA-155 alters manganese nanoparticle-induced inflammatory response’, *Toxicology research*, 5(6), pp. 1733–1743.

Guan, D., Fan, K., Spence, I. and Matthews, S. (2018) ‘QSAR ligand dataset for modelling mutagenicity, genotoxicity, and rodent carcinogenicity’, *Data in brief*, 17, pp. 876–884.

Gustafsson, A.B. and Gottlieb, R.A. (2008) ‘Heart mitochondria: gates of life and death’, *Cardiovascular research*, 77(2), pp. 334–343.

Gutzeit, C., Magri, G. and Cerutti, A. (2014) ‘Intestinal IgA production and its role in host-microbe interaction’, *Immunological reviews*, 260(1), pp. 76–85.

Guyon, X. and Yao, J.-F. (1999) ‘On the Underfitting and Overfitting Sets of Models

Chosen by Order Selection Criteria', *Journal of Multivariate Analysis*, pp. 221–249. Available at: https://doi.org/10.1006/jmva.1999.1828.

Hariharan, S. and Dharmaraj, S. (2020) 'Selenium and selenoproteins: it's role in regulation of inflammation', *Inflammopharmacology*, 28(3), pp. 667–695.

Harris, T.J.C. and Tepass, U. (2010) 'Adherens junctions: from molecules to morphogenesis', *Nature reviews. Molecular cell biology*, 11(7), pp. 502–514.

Hartung, T., Hoffmann, S. and Stephens, M. (2013) 'Mechanistic validation', *ALTEX*, 30(2), pp. 119–130.

Hayashi, M., Kim, S.W., Imanaka-Yoshida, K., Yoshida, T., Abel, E.D., Eliceiri, B., Yang, Y., Ulevitch, R.J. and Lee, J.D. (2004) 'Targeted deletion of BMK1/ERK5 in adult mice perturbs vascular integrity and leads to endothelial failure', *The Journal of clinical investigation*, 113(8), pp. 1138–1148.

Hayashi, M. and Lee, J.-D. (2004) 'Role of the BMK1/ERK5 signaling pathway: lessons from knockout mice', *Journal of molecular medicine*, 82(12), pp. 800–808.

He, B., Quan, L.P., Cai, C.Y., Yu, D.Y., Yan, W., Wei, Q.J., Zhang, Z., Huang, X.N. and Liu, L. (2022) 'Dysregulation and imbalance of innate and adaptive immunity are involved in the cardiomyopathy progression', *Frontiers in cardiovascular medicine*, 9, p. 973279.

He, L., Yu, Y., Wei, Y., Huang, J., Shen, Y. and Li, H. (2021) 'Characteristics and Spectrum of Cardiotoxicity Induced by Various Antipsychotics: A Real-World Study From 2015 to 2020 Based on FAERS', *Frontiers in pharmacology*, 12, p. 815151.

Helker, C.S., Eberlein, J., Wilhelm, K., Sugino, T., Malchow, J., Schuermann, A., Baumeister, S., Kwon, H.B., Maischein, H.M., Potente, M., Herzog, W. and Stainier, D.Y.R. (2020) 'Apelin signaling drives vascular endothelial cells toward a pro-angiogenic state', *eLife*, 9. Available at: https://doi.org/10.7554/eLife.55589.

Hellfeld, R. von, Brotzmann, K., Baumann, L., Strecker, R. and Braunbeck, T. (2020) 'Adverse effects in the fish embryo acute toxicity (FET) test: a catalogue of unspecific morphological changes versus more specific effects in zebrafish (Danio rerio) embryos', *Environmental Sciences Europe*. Available at: https://doi.org/10.1186/s12302-020-00398-3.

Helmy, H.S., Senousy, M.A., El-Sahar, A.E., Sayed, R.H., Saad, M.A. and Elbaz, E.M. (2020) 'Aberrations of miR-126-3p, miR-181a and sirtuin1 network mediate Di-(2-ethylhexyl) phthalate-induced testicular damage in rats: The protective role of hesperidin', *Toxicology*, 433-434, p. 152406.

He, W., Huang, H., Xie, Q., Wang, Z., Fan, Y., Kong, B., Huang, D. and Xiao, Y. (2016) 'MiR-155 Knockout in Fibroblasts Improves Cardiac Remodeling by Targeting Tumor Protein p53-Inducible Nuclear Protein 1', *Journal of cardiovascular pharmacology and therapeutics*, 21(4), pp. 423–435.

Hirano, T., Ishihara, K. and Hibi, M. (2000) 'Roles of STAT3 in mediating the cell growth, differentiation and survival signals relayed through the IL-6 family of cytokine receptors', *Oncogene*, 19(21), pp. 2548–2556.

Hishikawa, D., Hashidate, T., Shimizu, T. and Shindou, H. (2014) 'Diversity and function of membrane glycerophospholipids generated by the remodeling pathway in mammalian cells', *Journal of lipid research*, 55(5), pp. 799–807.

Hodgson, P., Ireland, J. and Grunow, B. (2018) 'Fish, the better model in human heart research? Zebrafish Heart aggregates as a 3D spontaneously cardiomyogenic in vitro model system', *Progress in biophysics and molecular biology*, 138, pp. 132–141.

Hoeger, C.W., Turissini, C. and Asnani, A. (2020) 'Doxorubicin cardiotoxicity: Pathophysiology updates', *Current treatment options in cardiovascular medicine*, 22(11). Available at: https://doi.org/10.1007/s11936-020-00842-w.

Holeček, M. (2018) 'Branched-chain amino acids in health and disease: metabolism, alterations in blood plasma, and as supplements', *Nutrition & metabolism*, 15, p. 33.

Holmes, S., Alekseyenko, A., Timme, A., Nelson, T., Pasricha, P.J. and Spormann, A. (2011) 'VISUALIZATION AND STATISTICAL COMPARISONS OF MICROBIAL COMMUNITIES USING R PACKAGES ON PHYLOCHIP DATA', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 142.

Holsapple, M.P. and Wallace, K.B. (2008) 'Dose response considerations in risk assessment--an overview of recent ILSI activities', *Toxicology letters*, 180(2), pp. 85–92.

Hou, S.X., Zheng, Z., Chen, X. and Perrimon, N. (2002) 'The Jak/STAT pathway in model organisms: emerging roles in cell movement', *Developmental cell*, 3(6), pp. 765–778.

Hove, J.R., Köster, R.W., Forouhar, A.S., Acevedo-Bolton, G., Fraser, S.E. and Gharib, M. (2003) 'Intracardiac fluid forces are an essential epigenetic factor for embryonic cardiogenesis', *Nature*, 421(6919), pp. 172–177.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch G.J., White, S., Chow, W., Kilian, B. *et al.* (2013) 'The zebrafish reference genome sequence and its relationship to the human genome', *Nature*, 496(7446), pp. 498–503.

Hsieh, D.J.-Y. and Liao, C.-F. (2002) 'Zebrafish M2 muscarinic acetylcholine receptor: cloning, pharmacological characterization, expression patterns and roles in embryonic bradycardia', *British journal of pharmacology*, 137(6), pp. 782–792.

Huang, W., Hickson, L.J., Eirin, A., Kirkland, J.L. and Lerman, L.O. (2022) 'Cellular senescence: the good, the bad and the unknown', *Nature reviews. Nephrology*, 18(10), pp. 611–627.

Hu, M., Jovanović, B. and Palić, D. (2019) 'In silico prediction of MicroRNA role in regulation of Zebrafish (Danio rerio) responses to nanoparticle exposure', *Toxicology in vitro: an international journal published in association with BIBRA*, 60, pp. 187–202.

Hwang, H.-M., Wade, T.L. and Sericano, J.L. (2002) 'Relationship between lysosomal membrane destabilization and chemical body burden in eastern oysters (Crassostrea virginica) from Galveston Bay, Texas, USA', *Environmental toxicology and chemistry / SETAC*, 21(6), pp. 1268–1271.

IBM Cloud Education (2020) *Unsupervised Learning*, *IBM*. Available at:

https://www.ibm.com/cloud/learn/unsupervised-learning (Accessed: 10 January 2022).

Imada, K. and Leonard, W.J. (2000) ‘The Jak-STAT pathway’, *Molecular immunology*, 37(1-2), pp. 1–11.

Jager, T., Albert, C., Preuss, T.G. and Ashauer, R. (2011) ‘General unified threshold model of survival--a toxicokinetic-toxicodynamic framework for ecotoxicology’, *Environmental science & technology*, 45(7), pp. 2529–2540.

Jager, T. (2019) *openGUTS*. Available at: http://openguts.info/about.html (Accessed: 2019).

Jager, T. (2020) *Interpretation of output of the openGUTS software*, *open guts*. Available at: http://openguts.info/downloads/openguts_interpret.pdf (Accessed: 2019).

Järveläinen, H., Sainio, A., Koulu, M., Wight, T.N. and Penttinen, R. (2009) ‘Extracellular matrix molecules: potential targets in pharmacotherapy’, *Pharmacological reviews*, 61(2), pp. 198–223.

Johnson, J.D., Edman, J.C. and Rutter, W.J. (1993) ‘A receptor tyrosine kinase found in breast carcinoma cells has an extracellular discoidin I-like domain’, *Proceedings of the National Academy of Sciences of the United States of America*, 90(12), pp. 5677–5681.

Jolliffe, I.T. and Cadima, J. (2016) ‘Principal component analysis: a review and recent developments’, *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065), p. 20150202.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. and Tanabe, M. (2019) ‘New approach for understanding genome variations in KEGG’, *Nucleic acids research*, 47(D1), pp. D590–D595.

Kang, D.S., Yang, J.H., Kim, H.S., Koo, B.K., Lee, C.M., Ahn, Y.S., Jung, J.H. and Seo, Y.R. (2018) ‘Application of the Adverse Outcome Pathway Framework to Risk Assessment for Predicting Carcinogenicity of Chemicals’, *Journal of cancer prevention*, 23(3), pp. 126–133.

Katz, M., Amit, I. and Yarden, Y. (2007) ‘Regulation of MAPKs by growth factors and receptor tyrosine kinases’, *Biochimica et biophysica acta*, 1773(8), pp. 1161–1176.

Kausar, S. and Falcao, A.O. (2018) ‘An automated framework for QSAR model building’, *Journal of cheminformatics*, 10(1). Available at: https://doi.org/10.1186/s13321-017-0256-5.

Kechagia, J.Z., Ivaska, J. and Roca-Cusachs, P. (2019) ‘Integrins as biomechanical sensors of the microenvironment’, *Nature reviews. Molecular cell biology*, 20(8), pp. 457–473.

KEGG Mapper – Convert ID (2021). Available at: https://www.genome.jp/kegg/tool/conv_id.html (Accessed: 5 January 2021).

KEGG PATHWAY Database. https://www.genome.jp/kegg/pathway.html. Accessed May 12, 2022.

KEGG PATHWAY: dre05168 (2021). Available at: https://www.kegg.jp/entry/dre05168 (Accessed: 11 January 2021).

Ketley, A., Warren, A., Holmes, E., Gering, M., Aboobaker, A.A. and Brook, J.D. (2013) ‘The miR-30 microRNA family targets smoothened to regulate hedgehog signalling in zebrafish early muscle development’, *PloS one*, 8(6), p. E65170.

Khanaghaei, M., Tourkianvalashani, F., Hekmatimoghaddam, S., Ghasemi, N., Rahaie, M., Khorramshahi, V., Sheikhpour, A., Heydari, Z. and Pourrajab, F. (2016) ‘Circulating miR-126 and miR-499 reflect progression of cardiovascular disease; correlations with uric acid and ejection fraction’, *Heart international*, 11(1), pp. E1–e9.

Kienzler, A., Barron, M.G., Belanger, S.E., Beasley, A. and Embry, M.R. (2017) ‘Mode of Action (MOA) Assignment Classifications for Ecotoxicology: An Evaluation of Approaches’, *Environmental science & technology*, 51(17), pp. 10203–10211.

Kienzler A., Connors, K.A., Bonnell, M., Barron, M.G., Beasley, A., Inglis, C.G., Norberg-King, T.J., Martin, T., Sanderson, H., Vallotton, N., Wilson, P. and Embry, M.R. (2019) ‘Mode of Action Classifications in the EnviroTox Database: Development and Implementation of a Consensus MOA Classification’, *Environmental toxicology and chemistry / SETAC*, 38(10), pp. 2294–2304.

Kilian Q. and Weinberger, L.K.S. (2009) ‘Distance Metric Learning for Large Margin Nearest Neighbor Classification’, *Journal of machine learning research: JMLR*, 10, pp. 207–244.

Kim, H.K., Lee, Y.S., Sivaprasad, U., Malhotra, A. and Dutta, A. (2006) ‘Muscle-specific microRNA miR-206 promotes muscle differentiation’, *The Journal of cell biology*, 174(5), pp. 677–687.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B. and Schilling, T.F. (1995) ‘Stages of embryonic development of the zebrafish’, *Developmental dynamics: an official publication of the American Association of Anatomists*, 203(3), pp. 253–310.

Kirkley, A.G. and Sargis, R.M. (2014) ‘Environmental endocrine disruption of energy metabolism and cardiovascular risk’, *Current diabetes reports*, 14(6), p. 494.

Klüver, N., Bittermann, K. and Escher, B.I. (2019) ‘QSAR for baseline toxicity and classification of specific modes of action of ionizable organic chemicals in the zebrafish embryo toxicity test’, *Aquatic toxicology*, 207, pp. 110–119.

Knapen, D., Angrish, M.M., Fortin, M.C., Katsiadaki, I., Leonard, M., Margiotta-Casaluci, L., Munn, S., O'Brien, J.M., Pollesch, N., Smith, L.C., Zhang, X. and Villeneuve, D.L. (2018) ‘Adverse outcome pathway networks I: Development and applications’, *Environmental toxicology and chemistry / SETAC*, 37(6), pp. 1723–1733.

Knudson, W. and Knudson, C.B. (1991) ‘Assembly of a chondrocyte-like pericellular matrix on non-chondrogenic cells. Role of the cell surface hyaluronan receptors in the assembly of a pericellular matrix’, *Journal of cell science*, 99 (Pt 2), pp. 227–235.

Koenig, J.A., Dao, T.L., Kan, R.K. and Shih, T.M. (2016) ‘Zebrafish as a model for acetylcholinesterase-inhibiting organophosphorus agent exposure and oxime reactivation', *Annals of the New York Academy of Sciences*, 1374(1), pp. 68–77.

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. and Peterson, H. (2020) 'gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler', *F1000Research*, 9. Available at: https://doi.org/10.12688/f1000research.24956.2.

Kraus, S., Naor, Z. and Seger, R. (2001) 'Intracellular signaling pathways mediated by the gonadotropin-releasing hormone (GnRH) receptor', *Archives of medical research*, 32(6), pp. 499–509.

Kuhn M. (2020) Classification and Regression Training [R package caret version 6.0-90]. October 2021. https://cran.r-project.org/package=caret.

Kujawski, J., Bernard, M.K., Janusz, A. and Kuźma, W. (2012) 'Prediction of log P: ALOGPS application in medicinal chemistry education', *Journal of chemical education*, 89(1), pp. 64–67.

Kujawski, J., Popielarska, H., Myka, A., Drabińska, B. and Bernard, M.K. (2012) 'The log P parameter as a molecular descriptor in the computer-aided drug design – an overview', *Computational Methods in Science and Technology*, 18(2), pp. 81–88.

Kwon, S., Bae, H., Jo, J. and Yoon, S. (2019) 'Comprehensive ensemble in QSAR prediction for drug discovery', *BMC bioinformatics*, 20(1), p. 521.

Kyriakis, J.M. and Avruch, J. (2001) 'Mammalian mitogen-activated protein kinase signal transduction pathways activated by stress and inflammation', *Physiological reviews*, 81(2), pp. 807–869.

Labute, P. (2000) 'A widely applicable set of descriptors', *Journal of molecular graphics & modelling*, 18(4-5), pp. 464–477.

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., and Golub, T.R. (2006) 'The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease', *Science*, 313(5795), pp. 1929–1935.

Larsson, O., Wahlestedt, C. and Timmons, J.A. (2005) 'Considerations when using the significance analysis of microarrays (SAM) algorithm', *BMC bioinformatics*, 6, p. 129.

Laskar, R. (1969) 'Eigenvalues of the adjacency matrix of cubic lattice graphs', *Pacific Journal of Mathematics*, 29(3), pp. 623–629.

Lavado, G.J., Gadaleta, D., Toma, C., Golbamaki, A., Toropov, A.A., Toropova, A.P., Marzo, M., Baderna, D., Arning, J. and Benfenati, E. (2020) 'Zebrafish AC modelling: (Q)SAR models to predict developmental toxicity in zebrafish embryo', *Ecotoxicology and environmental safety*, 202, p. 110936.

Lavado, G.J., Baderna, D., Carnesecchi, E., Toropova, A.P., Toropov, A.A., Dorne, J.L.C.M. and Benfenati, E. (2022) 'QSAR models for soil ecotoxicity: Development and validation of models to predict reproductive toxicity of organic chemicals in the collembola Folsomia candida', *Journal of hazardous materials*, 423(Pt B), p. 127236.

Lavine, K.J., Yu, K., White, A.C., Zhang, X., Smith, C., Partanen, J. and Ornitz, D.M.

(2005) ‘Endocardial and epicardial derived FGF signals regulate myocardial proliferation and differentiation in vivo’, *Developmental cell*, 8(1), pp. 85–95.

Lee, E., Koo, Y., Ng, A., Wei, Y., Luby-Phelps, K., Juraszek, A., Xavier, R.J., Cleaver, O., Levine, B. and Amatruda, J.F. (2014) ‘Autophagy is essential for cardiac morphogenesis during vertebrate development’, *Autophagy*, 10(4), pp. 572–587.

Lee, E.R., Noh, H. and Park, B.U. (2014) ‘Model selection via Bayesian information criterion for quantile regression models’, *Journal of the American Statistical Association*, 109(505), pp. 216–229.

Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K. and Irizarry, R.A. (2010) ‘Tackling the widespread and critical impact of batch effects in high-throughput data’, *Nature reviews. Genetics*, 11(10), pp. 733–739.

Leist, M. Hasiwa, N., Daneshian, M. and Hartung, T. (2012) ‘Validation and quality control of replacement alternatives – current status and future challenges’, *Toxicology research*, 1(1), pp. 8–22.

Leist, M., Ghallab, A., Graepel, R., Marchan, R., Hassan, R., Bennekou, S.H., Limonciel, A., Vinken, M,, Schildknecht, S., Waldmann, T., Danen, E., van Ravenzwaay, B., Kamp, H., Gardner, I., Godoy, P., Bois, F.Y., Braeuning, A., Reif, R., Oesch, F., Drasdo, D. *et al.* (2017) ‘Adverse outcome pathways: opportunities, limitations and open questions’, *Archives of toxicology*, 91(11), pp. 3477–3505.

Leist, M., Efremova, L. and Karreman, C. (2010) ‘Food for thought ... considerations and guidelines for basic test method descriptions in toxicology’, *ALTEX*, 27(4), pp. 309–317.

Lei, X., Jiang, X. and Wang, C. (2013) ‘Design and Implementation of a Real-Time Video Stream Analysis System Based on FFMPEG’, *2013 Fourth World Congress on Software Engineering* [Preprint]. Available at: https://doi.org/10.1109/wcse.2013.38.

Leonardi, R. and Jackowski, S. (2007) ‘Biosynthesis of Pantothenic Acid and Coenzyme A’, *EcoSal Plus*, 2(2). Available at: https://doi.org/10.1128/ecosalplus.3.6.3.4.

Levy, D.E. (1999) ‘Physiological significance of STAT proteins: investigations through gene disruption in vivo’, *Cellular and molecular life sciences: CMLS*, 55(12), pp. 1559–1567.

Li, J., Wu, L., Pei, M. and Zhang, Y. (2020) ‘YTHDF2, a protein repressed by miR-145, regulates proliferation, apoptosis, and migration in ovarian cancer cells’, *Journal of ovarian research*, 13(1), p. 111.

Liaw, A. and Wiener, M. (2002) Classification and Regression by Random forest. R News, 2, 18-22. http://CRAN.R-project.org/doc/Rnews/

Likas, A., Vlassis, N. and J. Verbeek, J. (2003) ‘The global k-means clustering algorithm’, *Pattern recognition*, 36(2), pp. 451–461.

Lin, C.C., Hui, M.N.Y. and Cheng, S.H. (2007) ‘Toxicity and cardiac effects of carbaryl in early developing zebrafish (Danio rerio) embryos’, *Toxicology and applied pharmacology*, 222(2), pp. 159–168.

Lin, C.Y., Lee, H.C., Fu, C.Y., Ding, Y.Y., Chen, J.S., Lee, M.H., Huang, W.J. and Tsai, H.J. (2013) ‘MiR-1 and miR-206 target different genes to have opposing roles during angiogenesis in zebrafish embryos’, *Nature communications*, 4, p. 2829.

Lin, C.Y., He, J.Y., Zeng, C.W., Loo, M.R., Chang, W.Y., Zhang, P.H. and Tsai, H.J. (2017) ‘modulates an Rtn4a/Cxcr4a/Thbs3a axis in newly forming somites to maintain and stabilize the somite boundary formation of zebrafish embryos’, *Open biology*, 7(7). Available at: https://doi.org/10.1098/rsob.170009.

Linke, S.P., Clarkin, K.C., Di Leonardo, A., Tsou, A. and Wahl, G.M. (1996) ‘A reversible, p53-dependent G0/G1 cell cycle arrest induced by ribonucleotide depletion in the absence of detectable DNA damage’, *Genes & development*, 10(8), pp. 934–947.

Lin, Q., He, Y., Gui J.F. and Mei, J. (2021) ‘Sox9a, not sox9b is required for normal cartilage development in zebrafish’, *Aquaculture and fisheries*, 6(3), pp. 254–259.

Li, R., Zupanic, A., Talikka, M., Belcastro, V., Madan, S., Dörpinghaus, J., Berg, C.V., Szostak, J., Martin, F., Peitsch, M.C. and Hoeng, J. (2020) ‘Systems Toxicology Approach for Testing Chemical Cardiotoxicity in Larval Zebrafish’, *Chemical research in toxicology*, 33(10), pp. 2550–2564.

Liu, H., Wei, M., Yang, X., Yin, C. and He, X. (2017) ‘Development of TLSER model and QSAR model for predicting partition coefficients of hydrophobic organic chemicals between low density polyethylene film and water’, *The Science of the total environment*, 574, pp. 1371–1378.

Liu, T., Wang, J., Subedi, K., Yi, Q., Zhou, L. and Mi, Q.S. (2021) ‘MicroRNA-155 Regulates MAIT1 and MAIT17 Cell Differentiation’, *Frontiers in cell and developmental biology*, 9, p. 670531.

Liu, X, Ning, G., Meng, A. and Wang, Q. (2012) ‘MicroRNA-206 regulates cell movements during zebrafish gastrulation by targeting prickle1a and regulating c-Jun N-terminal kinase 2 phosphorylation’, *Molecular and cellular biology*, 32(14), pp. 2934–2942.

Liu, Y., Zhu, Z., Ho, I.H.T., Shi, Y., Li, J., Wang, X., Chan, M.T.V. and Cheng, C.H.K. (2020) ‘Genetic Deletion of Disrupts Maternal-Zygotic Transition and Embryonic Body Plan’, *Frontiers in genetics*, 11, p. 853.

Liu, Y., Sepich, D.S. and Solnica-Krezel, L. (2017) ‘Stat3/Cdc25a-dependent cell proliferation promotes embryonic axis extension during zebrafish gastrulation’, *PLoS genetics*, 13(2), p. e1006564.

Li, W.M., Chan, C.M., Miller, A.L. and Lee, C.H. (2017) ‘Dual functional roles of molecular beacon as a MicroRNA detector and inhibitor’, *The Journal of biological chemistry*, 292(9), pp. 3568–3580.

Li, X.-Q., Tang, X.-R. and Li, L.-L. (2021) ‘Antipsychotics cardiotoxicity: What’s known and what's next’, *World journal of psychiatry*, 11(10), pp. 736–753.

Li, Y., Lu, J., Bao, X., Wang, X., Wu, J., Li, X. and Hong, W. (2016) ‘MiR-499-5p protects cardiomyocytes against ischaemic injury via anti-apoptosis by targeting PDCD4’, *Oncotarget*, 7(24), pp. 35607–35617.

Lockhart, M., Wirrig, E., Phelps, A. and Wessels, A. (2011) ‘Extracellular matrix and heart development', *Birth defects research. Part A, Clinical and molecular teratology*, 91(6), pp. 535–550.

Lovick, T.A. (2009) 'CCK as a modulator of cardiovascular function', *Journal of chemical neuroanatomy*, 38(3), pp. 176–184.

Lu, J.W., Liao, C.Y., Yang, W.Y., Lin, Y.M., Jin, S.L., Wang, H.D. and Yuh, C.H. (2014) 'Overexpression of endothelin 1 triggers hepatocarcinogenesis in zebrafish and promotes cell proliferation and migration through the AKT pathway', *PloS one*, 9(1), p. E85318.

Lukashev, M.E. and Werb, Z.(1998) 'ECM signalling: orchestrating cell behaviour and misbehaviour', *Trends in cell biology*, 8(11), pp. 437–441.

Luxán, G., D'Amato, G., MacGrogan, D. and de la Pompa, J.L. (2016) 'Endocardial Notch Signaling in Cardiac Development and Disease', *Circulation research*, 118(1), pp. E1–e18.

Lymperopoulos, A., Rengo, G. and Koch, W.J. (2013) 'Adrenergic nervous system in heart failure: pathophysiology and therapy', *Circulation research*, 113(6), pp. 739–753.

Maciag, M., Wnorowski, A., Mierzejewska, M. and Plazinska, A. (2022) 'Pharmacological assessment of zebrafish-based cardiotoxicity models', *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 148, p. 112695.

Magyary, I. (2018) 'Recent advances and future trends in zebrafish bioassays for aquatic ecotoxicology', *Ecocycles*, 4(2), pp. 12–18.

Männer, J. and Yelbuz, T.M. (2019) 'Functional Morphology of the Cardiac Jelly in the Tubular Heart of Vertebrate Embryos', *Journal of cardiovascular development and disease*, 6(1). Available at: https://doi.org/10.3390/jcdd6010012.

Manning, B.D. and Cantley, L.C. (2007) 'AKT/PKB signaling: navigating downstream', *Cell*, 129(7), pp. 1261–1274.

Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R. and Consonni, V. (2013) 'Quantitative structure-activity relationship models for ready biodegradability of chemicals', *Journal of chemical information and modeling*, 53(4), pp. 867–878.

Martínez-Sales, M., García-Ximénez, F. and Espinós, F.J. (2015) 'Zebrafish as a possible bioindicator of organic pollutants with effects on reproduction in drinking waters', *Journal of environmental sciences*, 33, pp. 254–260.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10.

Martin, T.M., Young, D.M., Lilavois, C.R. and Barron, M.G. (2015) 'Comparison of global and mode of action-based models for aquatic toxicity', *SAR and QSAR in environmental research*, 26(3), pp. 245–262.

Martin, W.K., Tennant, A.H., Conolly, R.B., Prince, K., Stevens, J.S., DeMarini, D.M., Martin, B.L., Thompson, L.C., Gilmour, M.I., Cascio, W.E., Hays, M.D., Hazari, M.S., Padilla, S. and Farraj, A.K. (2019) 'High-Throughput Video Processing of Heart Rate Responses in Multiple Wild-type Embryonic Zebrafish per Imaging Field', *Scientific reports*, 9(1), p. 145.

Martyniuk, C.J. and Denslow, N.D. (2009) ‘Towards functional genomics in fish using quantitative proteomics’, *General and comparative endocrinology*, 164(2-3), pp. 135–141.

Marutho, D., Handaka, S.H., Wijaya, E. and Muljono (2018) ‘The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News’, *2018 International Seminar on Application for Technology of Information and Communication* [Preprint]. Available at: https://doi.org/10.1109/isemantic.2018.8549751.

Matsumoto, S., Sakata, Y., Nakatani, D., Suna, S., Mizuno, H., Shimizu, M., Usami, M., Sasaki, T., Sato, H., Kawahara, Y., Hamasaki, T., Nanto, S., Hori, M. and Komuro, I. (2012) ‘A subset of circulating microRNAs are predictive for cardiac death after discharge for acute myocardial infarction’, *Biochemical and biophysical research communications*, 427(2), pp. 280–284.

Mauri, A., Consonni, V., Pavan, M. and Todeschini, R. (2006) ‘DRAGON SOFTWARE: AN EASY APPROACH TO MOLECULAR DESCRIPTOR CALCULATIONS’. Available at: https://www.semanticscholar.org/paper/DRAGON-SOFTWARE%3A-AN-EASY-APPROACH-TO-MOLECULAR-Mauri-Consonni/ae36d8ff9c925f34ba31d0ec9769f278f1327596 (Accessed: 27 May 2021).

Ma, W., Wei, S., Zhang, B. and Li, W. (2020) ‘Molecular Mechanisms of Cardiomyocyte Death in Drug-Induced Cardiotoxicity’, *Frontiers in cell and developmental biology*, 8, p. 434.

McCollum, C.W., Ducharme, N.A., Bondesson, M. and Gustafsson, J.A. (2011) ‘Developmental toxicity screening in zebrafish’, *Birth defects research. Part C, Embryo today: reviews*, 93(2), pp. 67–114.

McDermaid, A., Monier, B., Zhao, J., Liu, B. and Ma, Q. (2019) ‘Interpretation of differential gene expression results of RNA-seq data: review and integration’, *Briefings in bioinformatics*, 20(6), pp. 2044–2054.

McLeish, J.A., Chico, T.J., Taylor, H.B., Tucker, C., Donaldson, K. and Brown, S.B. (2010) ‘Skin exposure to micro- and nano-particles can cause haemostasis in zebrafish larvae’, *Thrombosis and haemostasis*, 103(4), pp. 797–807.

Kostal, J., Voutchkova-Kostal A., Zimmerman, J.B. and Anastas, P.T. (2016) ‘Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals’, *Green chemistry: an international journal and green chemistry resource: GC*, 18(16), pp. 4432–4445.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2021) ‘e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.R package version 1.7-9 from R-Forge.

M, H. and M.n, S. (2015) ‘A review on evaluation metrics for data classification evaluations’, *International Journal of Data Mining & Knowledge Management Process*, 5(2), pp. 01–11.

Michielan, L., Pireddu, L., Floris, M. and Moro, S. (2010) ‘Support Vector Machine (SVM) as Alternative Tool to Assign Acute Aquatic Toxicity Warning Labels to Chemicals’, *Molecular informatics*, 29(1-2), pp. 51–64.

Mikulášková, B., Maletínská, L., Zicha, J. and Kuneš, J. (2016) ‘’, *Molecular and cellular endocrinology*, 436, pp. 78–92.

Milan, D.J., Peterson, T.A., Ruskin, J.N., Peterson, R.T. and MacRae, C.A. (2003) ‘Drugs that induce repolarization abnormalities cause bradycardia in zebrafish’, *Circulation*, 107(10), pp. 1355–1358.

Misgeld, T., Burgess, R.W., Lewis, R.M., Cunningham, J.M., Lichtman, J.W. and Sanes, J.R. (2002) ‘Roles of neurotransmitter in synapse formation’, *Neuron*, 36(4), pp. 635–648.

Mishima, Y., Abreu-Goodger, C., Staton, A.A., Stahlhut, C., Shou, C., Cheng, C., Gerstein, M., Enright, A.J. and Giraldez, A.J. (2009) ‘Zebrafish miR-1 and miR-133 shape muscle gene expression and regulate sarcomeric actin organization’, *Genes & development*, 23(5), pp. 619–632.

Miura, G.I. and Yelon, D. (2011) ‘A guide to analysis of cardiac phenotypes in the zebrafish embryo’, *Methods in cell biology*, 101, pp. 161–180.

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., Zippelius, A., Pedro de Magalhães, J. and Larbi, A. (2019) ‘RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types’, *Cell reports*, 26(6), pp. 1627–1640.e7.

Monda, E., Palmiero, G., Rubino, M., Verrillo, F., Amodio, F., Di Fraia, F., Pacileo, R., Fimiani, F., Esposito, A., Cirillo, A., Fusco, A., Moscarella, E., Frisso, G., Russo, M.G., Pacileo, G., Calabrò, P., Scudiero, O., Caiazza, M. and Limongelli, G. (2020) ‘Molecular Basis of Inflammation in the Pathogenesis of Cardiomyopathies’, *International journal of molecular sciences*, 21(18). Available at: https://doi.org/10.3390/ijms21186462.

Moore, C., Richens, J.L., Hough, Y., Ucanok, D., Malla, S., Sang, F., Chen, Y., Elworthy, S., Wilkinson, R.N. and Gering, M. (2018) ‘Gfi1aa and Gfi1b set the pace for primitive erythroblast differentiation from hemangioblasts in the zebrafish embryo’, *Blood advances*, 2(20), pp. 2589–2606.

Moreau, G. and Broto, P. (1980) ‘THE AUTOCORRELATION OF A TOPOLOGICAL STRUCTURE: A NEW MOLECULAR DESCRIPTOR’. Available at: https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=PASCAL8040406871 (Accessed: 21 November 2021).

Moussa, N., Hassan, A. and Gharaghani, S. (2021) ‘Pharmacophore model, docking, QSAR, and molecular dynamics simulation studies of substituted cyclic imides and herbal medicines as COX-2 inhibitors’, *Heliyon*, 7(4), p. e06605.

Mozafari, Z., Arab Chamjangali, M. and Arashi, M. (2020) ‘Combination of least absolute shrinkage and selection operator with Bayesian Regularization artificial neural network (LASSO-BR-ANN) for QSAR studies using functional group and molecular docking mixed descriptors’, *Chemometrics and Intelligent Laboratory Systems*, 200(103998), p. 103998.

Mui, A.L. (1999) ‘The role of STATs in proliferation, differentiation, and apoptosis’, *Cellular and molecular life sciences: CMLS*, 55(12), pp. 1547–1558.

Mukherjee, R.K., Kumar, V. and Roy, K. (2022) ‘Ecotoxicological QSTR and QSTTR

Modeling for the Prediction of Acute Oral Toxicity of Pesticides against Multiple Avian Species', *Environmental science & technology*, 56(1), pp. 335–348.

Murtagh, F. and Contreras, P. (2012) 'Algorithms for hierarchical clustering: an overview', *Wiley interdisciplinary reviews. Data mining and knowledge discovery*, 2(1), pp. 86–97.

Murtagh, F. and Contreras, P. (2017) 'Algorithms for hierarchical clustering: an overview, II', *Wiley interdisciplinary reviews. Data mining and knowledge discovery*, 7(6), p. e1219.

Myatt, G.J., Ahlberg, E., Akahori, Y., Allen, D., Amberg, A., Anger, L.T., Aptula, A., Auerbach, S., Beilke, L., Bellion, P., Benigni, R., Bercu, J., Booth, E.D., Bower, D. and Brigo, A. (2018) 'In silico toxicology protocols', *Regulatory toxicology and pharmacology: RTP*, 96, pp. 1–17.

Nakasa, T., Ishikawa, M., Shi, M., Shibuya, H., Adachi, N. and Ochi, M. (2010) 'Acceleration of muscle regeneration by local injection of muscle-specific microRNAs in rat skeletal muscle injury model', *Journal of cellular and molecular medicine*, 14(10), pp. 2495–2505.

Nasrallah, G.K, Zhang, Y., Zagho, M.M., Ismail, H.M., Al-Khalaf, A.A., Prieto, R.M., Albinali, K.E., Elzatahry, A.A. and Deng, Y. (2018) 'A systematic investigation of the bio-toxicity of core-shell magnetic mesoporous silica microspheres using zebrafish model', *Microporous and mesoporous materials: the official journal of the International Zeolite Association*, 265, pp. 195–201.

National Center for Biotechnology Information (2020). Available at: https://www.ncbi.nlm.nih.gov/ (Accessed: 19 December 2020).

Nelson, W.J. and Nusse, R. (2004) 'Convergence of Wnt, beta-catenin, and cadherin pathways', *Science*, 303(5663), pp. 1483–1487.

Neves, B.J., Braga, R.C., Melo-Filho, C.C., Moreira-Filho, J.T., Muratov, E.N. and Andrade, A.H. (2018) 'QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery', *Frontiers in pharmacology*, 9, p. 1275.

Ng, C.T., Dheen, S.T., Yip, W.C., Ong, C.N., Bay, B.H. and Lanry Yung, L.Y. (2011) 'The induction of epigenetic regulation of PROS1 gene in lung fibroblasts by gold nanoparticles and implications for potential lung injury', *Biomaterials*, 32(30), pp. 7609–7615.

Nguyen, B.Y., Ruiz-Velasco, A., Bui, T., Collins, L., Wang, X. and Liu, W. (2019) 'Mitochondrial function in the heart: the insight into mechanisms and therapeutic potentials', *British journal of pharmacology*, 176(22), pp. 4302–4318.

Nishida, K., Ono, K., Kanaya, S. and Takahashi, K. (2014) 'KEGGscape: a Cytoscape app for pathway data integration', *F1000Research*, 3, p. 144.

Nishimoto, S. and Nishida, E. (2006) 'MAPK signalling: ERK5 versus ERK1/2', *EMBO reports*, 7(8), pp. 782–786.

Nishimura, Y., Inoue, A., Sasagawa, S., Koiwa, J., Kawaguchi, K., Kawase, R., Maruyama, T., Kim, S. and Tanaka, T. (2016) 'Using zebrafish in systems toxicology for developmental toxicity testing', *Congenital anomalies*, 56(1), pp. 18–27.

Nogueira, C., Almeida, L.S., Nesti, C., Pezzini, I., Videira, A., Vilarinho, L. and Santorelli, F.M. (2014) ‘Syndromes associated with mitochondrial DNA depletion’, *Italian journal of pediatrics*, 40, p. 34.

Ogadimma, A.I. and Adamu, U. (2016) ‘Quantitative structure activity relationship analysis of selected chalcone derivatives as Mycobacterium tuberculosis inhibitors’, *OAlib*, 03(03), pp. 1–13.

Olson, E.N. (2006) ‘Gene regulatory networks in the evolution and development of the heart’, *Science*, 313(5795), pp. 1922–1927.

Onakpoya, I.J., Heneghan, C.J. and Aronson, J.K. (2016) ‘Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature’, *BMC medicine*, 14, p. 10.

Onuma, T.A., Ding, Y., Abraham, E., Zohar, Y., Ando, H. and Duan, C. (2011) ‘Regulation of temporal and spatial organization of newborn GnRH neurons by IGF signaling in zebrafish’, *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 31(33), pp. 11814–11824.

O’Shea, J.J., Gadina, M. and Schreiber, R.D. (2002) ‘Cytokine signaling in 2002’, *Cell*, 109(2), pp. S121–S131.

Pal, D.K., Sengupta, C. and De, A.U. (1989) ‘Introduction of a novel topochemical index and exploitation of group connectivity concept to achieve predictability in QSAR and RDD’, *INDIAN* [Preprint].

Park, D.D., Gahr, B.M., Krause, J., Rottbauer, W., Zeller, T. and Just, S. (2021) ‘Long-Chain Acyl-Carnitines Interfere with Mitochondrial ATP Production Leading to Cardiac Dysfunction in Zebrafish’, *International journal of molecular sciences*, 22(16). Available at: https://doi.org/10.3390/ijms22168468.

Parthasarathy, A., Cross, P.J., Dobson, R.C.J., Adams, L.E., Savka, M.A. and Hudson, A.O. (2018) ‘A Three-Ring Circus: Metabolism of the Three Proteogenic Aromatic Amino Acids and Their Role in the Health of Plants and Animals’, *Frontiers in molecular biosciences*, 5, p. 29.

Patarroyo, M., Tryggvason, K. and Virtanen, I. (2002) ‘Laminin isoforms in tumor invasion, angiogenesis and metastasis’, *Seminars in cancer biology*, 12(3), pp. 197–207.

Patlewicz, G., Jeliazkova, N., Safford, R.J., Worth, A.P. and Aleksiev, B. (2008) ‘An evaluation of the implementation of the Cramer classification scheme in the Toxtree software’, *SAR and QSAR in environmental research*, 19(5-6), pp. 495–524.

Patlewicz, G., Worth, A.P. and Ball, N. (2016) ‘Validation of Computational Methods’, *Advances in experimental medicine and biology*, 856, pp. 165–187.

Patterson, J., and Gibson, A., (2017) *Deep Learning: A Practitioner’s Approach.* O’Reilly Media, Inc., Sebastopol.

Pei, J., Wang, G., Feng, L., Zhang, J., Jiang, T., Sun, Q. and Ouyang, L. (2021) ‘Targeting Lysosomal Degradation Pathways: New Strategies and Techniques for Drug Discovery’, *Journal of medicinal chemistry*, 64(7), pp. 3493–3507.

Perkins, R., Fang, H., Tong, W. and Welsh, W.J. (2003) ‘Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology', *Environmental toxicology and chemistry / SETAC*, 22(8), pp. 1666–1679.

Poelmann, R.E. and Gittenberger-de Groot, A.C. (2005) 'Apoptosis as an instrument in cardiovascular development', *Birth defects research. Part C, Embryo today: reviews*, 75(4), pp. 305–313.

Poon, K.L. and Brand, T. (2013) 'The zebrafish model system in cardiovascular research: A tiny fish with mighty prospects', *Global cardiology science & practice*, 2013(1), pp. 9–28.

Prasad, S. (2020) *Types of Clustering Algorithms in Machine Learning With Examples*, *Blogs & Updates on Data Science, Business Analytics, AI Machine Learning*. Available at: https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/ (Accessed: 27 May 2021).

Prud'homme, G.J. (2007) 'Pathobiology of transforming growth factor beta in cancer, fibrosis and immunologic disease, and therapeutic considerations', *Laboratory investigation; a journal of technical methods and pathology*, 87(11), pp. 1077–1091.

Przybyłek, M. (2020) 'Application 2D Descriptors and Artificial Neural Networks for Beta-Glucosidase Inhibitors Screening', *Molecules*, 25(24). Available at: https://doi.org/10.3390/molecules25245942.

Przybylińska, P.A. and Wyszkowski, M. (2016) 'Environmental contamination with phthalates and its impact on living organisms', *Ecological Chemistry and Engineering S*, pp. 347–356. Available at: https://doi.org/10.1515/eces-2016-0024.

PubChem (2018) *PubChem*. Available at: https://pubchem.ncbi.nlm.nih.gov/ (Accessed: 23 January 2018).

Pyati, U.J., Look, A.T. and Hammerschmidt, M. (2007) 'Zebrafish as a powerful vertebrate model system for in vivo studies of cell death', *Seminars in cancer biology*, 17(2), pp. 154–165.

Qiu, W., Chen, B., Greer, J.B., Magnuson, J.T., Xiong, Y., Zhong, H., Andrzejczyk, N.E., Zheng, C. and Schlenk, D. (2020) 'Transcriptomic Responses of Bisphenol S Predict Involvement of Immune Function in the Cardiotoxicity of Early Life-Stage Zebrafish ()', *Environmental science & technology*, 54(5), pp. 2869–2877.

Quéméneur, L., Gerland, L.M., Flacher, M., Ffrench, M., Revillard, J.P. and Genestier, L. (2003) 'Differential control of cell cycle, proliferation, and survival of primary T lymphocytes by purine and pyrimidine nucleotides', *Journal of immunology*, 170(10), pp. 4986–4995.

Radisic, M., Park, H., Shing, H., Consi, T., Schoen, F.J., Langer, R., Freed, L.E. and Vunjak-Novakovic, G. (2004) 'Functional assembly of engineered myocardium by electrical stimulation of cardiac myocytes cultured on scaffolds', *Proceedings of the National Academy of Sciences of the United States of America*, 101(52), pp. 18129–18134.

Raies, A.B. and Bajic, V.B. (2016) 'In silico toxicology: computational methods for the prediction of chemical toxicity', *Wiley interdisciplinary reviews. Computational molecular science*, 6(2), pp. 147–172.

Ramanan, V.K., Shen, L., Moore, J.H. and Saykin, A.J. (2012) ‘Pathway analysis of genomic data: concepts, methods, and prospects for future development’, *Trends in genetics: TIG*, 28(7), pp. 323–332.

Raman, M., Chen, W. and Cobb, M.H. (2007) ‘Differential regulation and properties of MAPKs’, *Oncogene*, 26(22), pp. 3100–3112.

Ramos, J.W. (2008) ‘The regulation of extracellular signal-regulated kinase (ERK) in mammalian cells’, *The international journal of biochemistry & cell biology*, 40(12), pp. 2707–2719.

Reese, S.E., Archer, K.J., Therneau, T.M., Atkinson, E.J., Vachon, C.M., de Andrade, M., Kocher, J.P. and Eckel-Passow, J.E. (2013) ‘A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis’, *Bioinformatics*, 29(22), p. 2877.

Refaeilzadeh, P., Tang L. and Liu H. (2009) ‘Cross-Validation’, in *Encyclopedia of Database Systems*. Boston, MA: Springer US, pp. 532–538.

Rehman, T.U., Mahmud, S., Chang, Y.K., Jin, J. and Shin, J. (2019) ‘Current and future applications of statistical machine learning algorithms for agricultural machine vision systems’, *Computers and Electronics in Agriculture*, 156, pp. 585–605.

Reynolds, A.P., Richards, G., de la Iglesia, B. and Rayward-Smith V.J. (2006) ‘Clustering rules: A comparison of partitioning and hierarchical clustering algorithms’, *Journal of Mathematical Modelling and Algorithms*, 5(4), pp. 475–504.

Ricard-Blum, S. and Vallet, S.D. (2016) ‘Matricryptins Network with Matricellular Receptors at the Surface of Endothelial and Tumor Cells’, *Frontiers in pharmacology*, 7, p. 11.

Rincón, M. and Davis, R.J. (2009) ‘Regulation of the immune response by stress-activated protein kinases’, *Immunological reviews*, 228(1), pp. 212–224.

Ritchie, T.J. and Macdonald, S.J.F. (2009) ‘The impact of aromatic ring count on compound developability--are too many aromatic rings a liability in drug design?’, *Drug discovery today*, 14(21-22), pp. 1011–1020.

Ritz, C., Baty, F., Streibig, J.C. and Gerhard, D. (2015) ‘Dose-Response Analysis Using R’, *PLOS ONE*, p. e0146021. Available at: https://doi.org/10.1371/journal.pone.0146021.

Robinson, M.D. and Oshlack, A. (2010) ‘A scaling normalization method for differential expression analysis of RNA-seq data’, *Genome biology*, 11(3), p. R25.

van Rooij, E., Quiat, D., Johnson, B.A., Sutherland, L.B., Qi, X., Richardson, J.A., Kelm, R.J. Jr and Olson, E.N. (2009) ‘A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance’, *Developmental cell*, 17(5), pp. 662–673.

Rose, B.A., Force, T. and Wang, Y. (2010) ‘Mitogen-activated protein kinase signaling in the heart: angels versus demons in a heart-breaking tale’, *Physiological reviews*, 90(4), pp. 1507–1546.

Roy, K., Kar, S. and Das, R.N. (2015) ‘Chemical Information and Descriptors’, in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and*

Risk Assessment. Elsevier, pp. 47–80.

Rozario, T., Dzamba, B., Weber, G.F., Davidson, L.A. and DeSimone, D.W. (2009) ‘The physical state of fibronectin matrix differentially regulates morphogenetic movements in vivo’, *Developmental biology*, 327(2), pp. 386–398.

Ruan, Y.-L. (2014) ‘Sucrose metabolism: gateway to diverse carbon use and sugar signaling’, *Annual review of plant biology*, 65, pp. 33–67.

Bradbury, S.P., Broderius, S.J., Hammermeister D.E. and Drummond, R.A. (1997) ‘Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas)’, *Environmental toxicology and chemistry / SETAC*, 16(5), pp. 948–967.

Sainio, A. and Järveläinen, H. (2020) ‘Extracellular matrix-cell interactions: Focus on therapeutic applications’, *Cellular signalling*, 66, p. 109487.

Salgado-Almario, J., Vicente, M., Vincent, P., Domingo, B. and Llopis, J. (2020) ‘Mapping calcium dynamics in the heart of zebrafish embryos with ratiometric genetically encoded calcium indicators’, *International journal of molecular sciences*, 21(18), p. 6610.

Sampurna, B., Audira, G., Juniardi, S., Lai, Y.H. and Hsiao, C.D. (2018) ‘A simple ImageJ-based method to measure cardiac rhythm in zebrafish embryos’, *Inventions*, 3(2), p. 21.

Sarmah, S. and Marrs, J.A. (2016) ‘Zebrafish as a Vertebrate Model System to Evaluate Effects of Environmental Toxicants on Cardiac Development and Function’, *International journal of molecular sciences*, 17(12). Available at: https://doi.org/10.3390/ijms17122123.

Sauer, U.G., Deferme, L., Gribaldo, L., Hackermüller, J., Tralau, T., van Ravenzwaay, B., Yauk, C., Poole, A., Tong, W. and Gant, T.W. (2017) ‘The challenge of the application of ’omics technologies in chemicals risk assessment: Background and outlook’, *Regulatory toxicology and pharmacology: RTP*, 91 Suppl 1, pp. S14–S26.

Savoca, D. and Pace, A. (2021) ‘Bioaccumulation, Biodistribution, Toxicology and Biomonitoring of Organofluorine Compounds in Aquatic Organisms’, *International journal of molecular sciences*, 22(12). Available at: https://doi.org/10.3390/ijms22126276.

Sawada, R., Kotera, M. and Yamanishi, Y. (2014) ‘Benchmarking a Wide Range of Chemical Descriptors for Drug-Target Interaction Prediction Using a Chemogenomic Approach’, *Molecular informatics*, 33(11-12), pp. 719–731.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P. and Cardona, A. (2012) ‘Fiji: an open-source platform for biological-image analysis’, *Nature methods*, 9(7), pp. 676–682.

Schindler, M. (2016) ‘A QSAR for the prediction of rate constants for the reaction of VOCs with nitrate radicals’, *Chemosphere*, 154, pp. 23–33.

Schneider, A., Hommel, G. and Blettner, M. (2010) ‘Linear regression analysis: part 14 of a series on evaluation of scientific publications’, *Deutsches Arzteblatt* *international*, 107(44), pp. 776–782.

Schneider, G., Neidhart, W., Giller, T. and Schmid, G. (1999) '"Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening', *Angewandte Chemie International Edition*, pp. 2894–2896. Available at: https://doi.org/10.1002/(sici)1521-3773(19991004)38:19<2894::aid-anie2894>3.0.co;2-f.

Scholz, S., Fischer, S., Gündel, U., Küster, E., Luckenbach, T. and Voelker, D. (2008) 'The zebrafish embryo model in environmental risk assessment--applications beyond acute toxicity testing', *Environmental science and pollution research international*, 15(5), pp. 394–404.

Scholz, S., Ortmann, J., Klüver, N. and Léonard, M. (2014) 'Extensive review of fish embryo acute toxicities for the prediction of GHS acute systemic toxicity categories', *Regulatory toxicology and pharmacology: RTP*, 69(3), pp. 572–579.

Segert, J., Schneider, I., Berger, I.M., Rottbauer, W. and Just, S. (2018) 'Mediator complex subunit Med12 regulates cardiac jelly development and AV valve formation in zebrafish', *Progress in biophysics and molecular biology*, 138, pp. 20–31.

Serra, A., Fratello, M., Cattelani, L., Liampa, I., Melagraki, G., Kohonen, P., Nymark, P., Federico, A., Kinaret, P.A.S., Jagiello, K., Ha, M.K., Choi, J.S., Sanabria, N., Gulumian, M., Puzyn, T., Yoon, T.H., Sarimveis, H., Grafström, R., Afantitis, A. and Greco, D. (2020) 'Transcriptomics in Toxicogenomics, Part III: Data Modelling for Risk Assessment', *Nanomaterials (Basel, Switzerland)*, 10(4). Available at: https://doi.org/10.3390/nano10040708.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome research*, 13(11), pp. 2498–2504.

Sheeran, F.L. and Pepe, S. (2006) 'Energy deficiency in the failing heart: linking increased reactive oxygen species and disruption of oxidative phosphorylation rate', *Biochimica et biophysica acta*, 1757(5-6), pp. 543–552.

Shen, L., Li, C., Zhang, H., Qiu, S., Fu, T. and Xu, Y. (2019) 'Downregulation of miR-146a Contributes to Cardiac Dysfunction Induced by the Tyrosine Kinase Inhibitor Sunitinib', *Frontiers in pharmacology*, 10, p. 914.

Shieh, J.T., Huang, Y., Gilmore, J. and Srivastava, D. (2011) 'Elevated miR-499 levels blunt the cardiac stress response', *PloS one*, 6(5), p. e19481.

Shi, L. and Tu, B.P. (2015) 'Acetyl-CoA and the regulation of metabolism: mechanisms and consequences', *Current opinion in cell biology*, 33, pp. 125–131.

Sigma-Aldrich (2020). Available at: http://www.sigmaaldrich.com/sigma-aldrich/home.html (Accessed: 2020).

Sill, M., Hielscher, T., Zucknick, M., Becker, N. *and* Zucknick, M. (2014) 'C060: Extended inference with lasso and elastic-net regularized cox and generalized linear models', *Journal of statistical software*, 62(5). Available at: https://doi.org/10.18637/jss.v062.i05.

Silva, A.C., Pereira, C., Fonseca, A.C.R.G., Pinto-do-Ó, P. *and* Nascimento, D.S.

(2020) ‘Bearing My Heart: The Role of Extracellular Matrix on Cardiac Development, Homeostasis, and Injury Response’, *Frontiers in cell and developmental biology*, 8, p. 621644.

Simpson, K.E., Venkateshappa, R., Pang, Z.K., Faizi, S., Tibbits, G.F.*and* Claydon, T.W. (2020) ‘Utility of Zebrafish Models of Acquired and Inherited Long QT Syndrome’, *Frontiers in physiology*, 11, p. 624129.

Sinaga, K.P. and Yang, M.-S. (2020) ‘Unsupervised K-means clustering algorithm’, *IEEE access: practical innovations, open solutions*, 8, pp. 80716–80727.

Singh, G. and Koropatnick, J. (1988) ‘Differential toxicity of cis and trans isomers of dichlorodiammineplatinum’, *Journal of biochemical toxicology*, 3, pp. 223–233.

Singh, K.P., Gupta, S. and Rai, P. (2013) ‘Predicting acute aquatic toxicity of structurally diverse chemicals in fish using artificial intelligence approaches’, *Ecotoxicology and environmental safety*, 95, pp. 221–233.

Sirci, F., Napolitano, F., Pisonero-Vaquero, S., Carrella, D., Medina, D.L.*and* di Bernardo, D. (2017) ‘Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses’, *NPJ systems biology and applications*, 3, p. 23.

Sluijter, J.P., van Mil, A., van Vliet, P., Metz, C.H., Liu, J., Doevendans, P.A.*and* Goumans, M.J. (2010) ‘MicroRNA-1 and -499 regulate differentiation and proliferation in human-derived cardiomyocyte progenitor cells’, *Arteriosclerosis, thrombosis, and vascular biology*, 30(4), pp. 859–868.

Smalley, J.L., Gant, T.W. and Zhang, S.-D. (2010) ‘Application of connectivity mapping in predictive toxicology based on gene-expression similarity’, *Toxicology*, 268(3), pp. 143–146.

Smith, S.M. and Melrose, J. (2019) ‘Type XI collagen-perlecan-HS interactions stabilise the pericellular matrix of annulus fibrosus cells and chondrocytes providing matrix stabilisation and homeostasis’, *Journal of molecular histology*, 50(3), pp. 285–294.

Solimeo, R., Zhang, J., Kim, M., Sedykh, A.*and* Zhu, H. (2012) ‘Predicting chemical ocular toxicity using a combinatorial QSAR approach’, *Chemical research in toxicology*, 25(12), pp. 2763–2769.

Stahlhut, C., Suárez, Y., Lu, J., Mishima, Y.*and* Giraldez, A.J. (2012) ‘miR-1 and miR-206 regulate angiogenesis by modulating VegfA expression in zebrafish’, *Development*, 139(23), pp. 4356–4364.

Stankunas, K., Hang, C.T., Tsun, Z.Y., Chen, H., Lee, N.V., Wu, J.I., Shang, C., Bayle, J.H., Shou, W., Iruela-Arispe, M.L.*and* Chang, C.P. (2008) ‘Endocardial Brg1 represses ADAMTS1 to maintain the microenvironment for myocardial morphogenesis’, *Developmental cell*, 14(2), pp. 298–311.

Staudt, D. and Stainier, D. (2012) ‘Uncovering the molecular and cellular mechanisms of heart development using the zebrafish’, *Annual review of genetics*, 46, pp. 397–418.

Stavropoulou, E., Pircalabioru, G.G. and Bezirtzoglou, E. (2018) ‘The Role of Cytochromes P450 in Infection’, *Frontiers in immunology*, 9, p. 89.

Stefatos, G. and Hamza, A.B. (2007) ‘Cluster pca for outliers detection in high-dimensional data’, in *2007 IEEE International Conference on Systems, Man and Cybernetics*. *2007 IEEE International Conference on Systems, Man and Cybernetics*, IEEE. Available at: https://doi.org/10.1109/icsmc.2007.4414244.

Sugi, Y., Ito, N., Szebenyi, G., Myers, K., Fallon, J.F., Mikawa, T.*and* Markwald, R.R. (2003) ‘Fibroblast growth factor (FGF)-4 can induce proliferation of cardiac cushion mesenchymal cells during early valve leaflet formation’, *Developmental biology*, 258(2), pp. 252–263.

Suthaharan, S. (2016) ‘Support Vector Machine’, in *Machine Learning Models and Algorithms for Big Data Classification*. Boston, MA: Springer US, pp. 207–235.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P.*and* Feuston, B.P. (2003) ‘Random forest: a classification and regression tool for compound classification and QSAR modeling’, *Journal of chemical information and computer sciences*, 43(6), pp. 1947–1958.

Taber, L.A. (1998) ‘Mechanical aspects of cardiac development’, *Progress in biophysics and molecular biology*, 69(2-3), pp. 237–255.

Tanaka, H., Shan, W., Phillips, G.R., Arndt, K., Bozdagi, O., Shapiro, L., Huntley, G.W., Benson, D.L. *and* Colman, D.R. (2000) ‘Molecular modification of N-cadherin in response to synaptic activity’, *Neuron*, 25(1), pp. 93–107.

Teame, T., Zhang, Z., Ran, C., Zhang, H., Yang, Y., Ding, Q., Xie, M., Gao, C., Ye, Y., Duan, M. *and* Zhou, Z. (2019) ‘The use of zebrafish () as biomedical models’, *Animal frontiers: the review magazine of animal agriculture*, 9(3), pp. 68–77.

Thorndike, R.L. (1953) ‘Who belongs in the family?’, *Psychometrika*, 18(4), pp. 267–276.

Todeschini, R. and Consonni, V. (2008) *Handbook of Molecular Descriptors*. John Wiley & Sons.

Todeschini, R. and Consonni, V. (2009) *Molecular descriptors for chemoinformatics: Volume I: Alphabetical listing / volume II: Appendices, references 2 volume set* [PDF]. 2nd edn. Edited by R. Todeschini and V. Consonni. Weinheim, Germany: Wiley-VCH Verlag (Methods and Principles in Medicinal Chemistry).

Tokarz, J., Möller, G., de Angelis, M.H. *and* Adamski, J. (2013) ‘Zebrafish and steroids: what do we know and what do we need to know?’, *The Journal of steroid biochemistry and molecular biology*, 137, pp. 165–173.

Toustou, C., Walet-Balieu, M.L., Kiefer-Meyer, M.C., Houdou, M., Lerouge, P., Foulquier, F. *and* Bardor, M. (2022) ‘Towards understanding the extensive diversity of protein N-glycan structures in eukaryotes’, *Biological reviews of the Cambridge Philosophical Society*, 97(2), pp. 732–748.

Toyoshima, Y., Monson, C., Duan, C., Wu, Y., Gao, C., Yakar, S., Sadler, K.C. *and* LeRoith, D. (2008) ‘The role of insulin receptor signaling in zebrafish embryogenesis’, *Endocrinology*, 149(12), pp. 5996–6005.

Tran, D.H. and Wang, Z.V. (2019) ‘Glucose Metabolism in Cardiac Hypertrophy and Heart Failure’, *Journal of the American Heart Association*, 8(12), p. e012673.

Tsantili-Kakoulidou, A. and Demopoulos, V.J. (2021) ‘Drug-like Properties and Fraction Lipophilicity Index as a combined metric’, *ADMET & DMPK*, 9(3), pp. 177–190.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) ‘Significance analysis of microarrays applied to the ionizing radiation response’, *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), pp. 5116–5121.

Uings, I.J. and Farrow, S.N. (2000) ‘Cell receptors and cell signalling’, *Molecular pathology: MP*, 53(6), pp. 295–299.

Vacante, F., Denby, L., Sluimer, J.C. *and* Baker, A.H. (2019) ‘The function of miR-143, miR-145 and the MiR-143 host gene in cardiovascular development and disease’, *Vascular pharmacology*, 112, pp. 24–30.

Vaes, W.H.J., Ramos, E.U., Verhaar, H.J.M. *and* Hermens, J.L.M. (1998) ‘Acute toxicity of nonpolar versus polar narcosis: Is there a difference?’, *Environmental toxicology and chemistry / SETAC*, 17(7), pp. 1380–1384.

Claeys L., Iaccino, F., Janssen, C.R., Van Sprang, P. and Verdonck, F. (2013) ‘Development and validation of a quantitative structure-activity relationship for chronic narcosis to fish’, *Environmental toxicology and chemistry / SETAC*, 32(10), pp. 2217–2225.

Veeneman, B.A., Shukla, S., Dhanasekaran, S.M., Chinnaiyan, A.M. and Nesvizhskii, A.I. (2015) ‘Two-pass alignment improves novel splice junction quantification’, *Bioinformatics* , 32(1), pp. 43–49.

Vera Alvarez, R., Pongor, L.S., Mariño-Ramírez, L. and Landsman, D. (2019) ‘TPMCalculator: one-step software to quantify mRNA abundance of genomic features’, *Bioinformatics*, 35(11), pp. 1960–1962.

Verhaar, H.J.M., van Leeuwen, C.J. and Hermens, J.L.M. (1992) ‘Classifying environmental pollutants’, *Chemosphere*, 25(4), pp. 471–491.

Villeneuve, D., Volz, D.C., Embry, M.R., Ankley, G.T., Belanger, S.E., Léonard, M., Schirmer, K., Tanguay, R., Truong, L. and Wehmas, L. (2014) ‘Investigating alternatives to the fish early-life stage test: a strategy for discovering and annotating adverse outcome pathways for early fish development’, *Environmental toxicology and chemistry / SETAC*, 33(1), pp. 158–169.

Villeneuve, D.L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T.H., LaLone, C.A., Landesmann, B., Lettieri, T., Munn, S., Nepelska, M., Ottinger, M.A., Vergauwen, L. and Whelan, M. (2014) ‘Adverse outcome pathway (AOP) development I: strategies and principles’, *Toxicological sciences: an official journal of the Society of Toxicology*, 142(2), pp. 312–320.

Vinken, M. (2019) ‘Omics-based input and output in the development and use of adverse outcome pathways’, *Current opinion in toxicology*, 18, pp. 8–12.

Waber, L.J., Valle, D., Neill, C., DiMauro, S. and Shug, A.(1982) ‘Carnitine deficiency presenting as familial cardiomyopathy: A treatable defect in carnitine transport’, *The Journal of pediatrics*, 101(5), pp. 700–705.

Walker, J.D., Enache, M. and Dearden, J.C. (2003) ‘Quantitative cationic-activity relationships for predicting toxicity of metals', *Environmental toxicology and chemistry / SETAC*, 22(8), pp. 1916–1935.

Wang, D., Weng, Y., Guo, S., Qin, W., Ni, J., Yu, L., Zhang, Y., Zhao, Q., Ben, J. and Ma, J. (2019) 'microRNA-1 Regulates NCC Migration and Differentiation by Targeting', *International journal of biological sciences*, 15(12), pp. 2538–2547.

Wang, J., Hao, D., Zeng, L., Zhang, Q. and Huang, W. (2021) 'Neuropeptide Y mediates cardiac hypertrophy through microRNA-216b/FoxO4 signaling pathway', *International journal of medical sciences*, 18(1), pp. 18–28.

Wang, K., Li, M. and Hakonarson, H. (2010) 'Analysing biological pathways in genome-wide association studies', *Nature reviews. Genetics*, 11(12), pp. 843–854.

Wang, L. (2016) 'Mitochondrial purine and pyrimidine metabolism and beyond', *Nucleosides, nucleotides & nucleic acids*, 35(10-12), pp. 578–594.

Wang, Q., Xu, Z. and Ai, Q. (2021) 'Arginine metabolism and its functions in growth, nutrient utilization, and immunonutrition of fish', *Animal nutrition (Zhongguo xu mu shou yi xue hui)*, 7(3), pp. 716–727.

Wang, S., Zhang, X., Gui, B., Xu, X., Su, L., Zhao, Y.H. and Martyniuk, C.J. (2022) 'Comparison of modes of action between fish, cell and mitochondrial toxicity based on toxicity correlation, excess toxicity and QSAR for class-based compounds', *Toxicology*, 470, p. 153155.

Wang, X., Lian, Y., Wen, X., Guo, J., Wang, Z., Jiang, S. and Hu, Y. (2017) 'Expression of miR-126 and its potential function in coronary artery disease', *African health sciences*, 17(2), pp. 474–480.

Wang, X. and Tournier, C. (2006) 'Regulation of cellular functions by the ERK5 signalling pathway', *Cellular signalling*, 18(6), pp. 753–760.

Wan, Q., Xu, T., Ding, W., Zhang, X., Ji, X., Yu, T., Yu, W., Lin, Z. and Wang, J. (2018) 'miR-499-5p Attenuates Mitochondrial Fission and Cell Apoptosis via p21 in Doxorubicin Cardiotoxicity', *Frontiers in genetics*, 9, p. 734.

Waring, M.J. (2009) 'Defining optimum lipophilicity and molecular weight ranges for drug candidates-Molecular weight dependent lower logD limits based on permeability', *Bioorganic & medicinal chemistry letters*, 19(10), pp. 2844–2851.

Watkins, M., Sizochenko, N., Rasulev, B. and Leszczynski, J. (2016) 'Estimation of melting points of large set of persistent organic pollutants utilizing QSPR approach', *Journal of molecular modeling*, 22(3), p. 55.

Watson, F.L., Schmidt, H., Turman, Z.K., Hole, N., Garcia, H., Gregg, J., Tilghman, J. and Fradinger, E.A. (2014) 'Organophosphate pesticides induce morphological abnormalities and decrease locomotor activity and heart rate in Danio rerio and Xenopus laevis', *Environmental toxicology and chemistry / SETAC*, 33(6), pp. 1337–1345.

Weinberg, G., Ullman, B. and Martin, D.W., Jr (1981) 'Mutator phenotypes in mammalian cell mutants with distinct biochemical defects and abnormal deoxyribonucleoside triphosphate pools', *Proceedings of the National Academy of Sciences of the United States of America*, 78(4), pp. 2447–2451.

Wheeler-Jones, C.P.D. (2005) ‘Cell signalling in the cardiovascular system: an overview’, *Heart*, 91(10), pp. 1366–1374.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wilson, K.D., Hu, S., Venkatasubrahmanyam, S., Fu, J.D., Sun, N., Abilez, O.J., Baugh, J.J., Jia, F., Ghosh, Z., Li, R.A., Butte, A.J. and Wu, J.C. (2010) ‘Dynamic microRNA expression programs during cardiac differentiation of human embryonic stem cells: role for miR-499’, *Circulation. Cardiovascular genetics*, 3(5), pp. 426–435.

Wiśniowska, B., Mendyk, A., Szlęk, J., Kołaczkowski, M. and Polak, S. (2015) ‘Enhanced QSAR models for drug-triggered inhibition of the main cardiac ion currents’, *Journal of applied toxicology: JAT*, 35(9), pp. 1030–1039.

Wong, K.Y., Mercader, A.G., Saavedra, L.M., Honarparvar, B., Romanelli, G.P. and Duchowicz, P.R. (2014) ‘QSAR analysis on tacrine-related acetylcholinesterase inhibitors’, *Journal of biomedical science*, 21(1), p. 84.

Worachartcheewan, A., Nantasenamat, C., Isarankura-Na-Ayudhya, C. and Prachayasittikul, C. (2015) ‘Probing the origins of anticancer activity of chrysin derivatives’, *Medicinal Chemistry Research*, pp. 1884–1892. Available at: https://doi.org/10.1007/s00044-014-1260-1.

Wright, M.N. and Ziegler, A. (2017) ‘Ranger: A fast implementation of random forests for high dimensional data in C++ and R’, *Journal of statistical software*, 77(1). Available at: https://doi.org/10.18637/jss.v077.i01.

Wu, X., Zhou, Q., Mu, L. and Hu, X. (2022) ‘Machine learning in the identification, prediction and exploration of environmental toxicology: Challenges and perspectives’, *Journal of hazardous materials*, 438, p. 129487.

Xie, W., Li, P., Wang, Z., Chen, J., Lin, Z., Liang, X. and Mo, Y. (2014) ‘Rosuvastatin may reduce the incidence of cardiovascular events in patients with acute coronary syndromes receiving percutaneous coronary intervention by suppressing miR-155/SHIP-1 signaling pathway’, *Cardiovascular therapeutics*, 32(6), pp. 276–282.

Yamashita, S., Miyagi, C., Carmany-Rampey, A., Shimizu, T., Fujii, R., Schier, A.F. and Hirano, T. (2002) ‘Stat3 controls cell movements during zebrafish gastrulation’, *Developmental cell*, 2(3), pp. 363–375.

Yang, H., Harrington, C.A., Vartanian, K., Coldren, C.D., Hall, R. and Churchill, G.A. (2008) ‘Randomization in Laboratory Procedure Is Key to Obtaining Reproducible Microarray Results’, *PloS one*, 3(11). Available at: https://doi.org/10.1371/journal.pone.0003724.

Yang, L., Ho, N.Y., Alshut, R., Legradi, J., Weiss, C., Reischl, M., Mikut, R., Liebel, U., Müller, F. and Strähle, U. (2009) ‘Zebrafish embryos as models for embryotoxic and teratological effects of chemicals’, *Reproductive toxicology*, 28(2), pp. 245–253.

Yang, Y., Engkvist, O., Llinàs, A. and Chen, H. (2012) ‘Beyond size, ionization state, and lipophilicity: influence of molecular topology on absorption, distribution, metabolism, excretion, and toxicity for druglike compounds’, *Journal of medicinal chemistry*, 55(8), pp. 3667–3677.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J.C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C.G., Gil, L., Grego, T., *et al.* (2020) 'Ensembl 2020', *Nucleic acids research*, 48(D1), pp. D682–D688.

Yeh, T.C. and Pellegrini, S. (1999) 'The Janus kinase family of protein tyrosine kinases and their role in signaling', *Cellular and molecular life sciences: CMLS*, 55(12), pp. 1523–1534.

Yogita Rani, D.H.R. (2013) 'A Study of Hierarchical Clustering Algorithm', *International Journal of Information and Computation Technology.*, 3, pp. 1225–1232.

Yokoi, H., Yan, Y.L., Miller, M.R., BreMiller, R.A., Catchen, J.M., Johnson, E.A. and Postlethwait, J.H. (2009) 'Expression profiling of zebrafish sox9 mutants reveals that Sox9 is required for retinal differentiation', *Developmental biology*, 329(1), pp. 1–15.

Yuan, H., Wang, Y.-Y. and Cheng, Y.-Y. (2007) 'Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow', *Journal of molecular graphics & modelling*, 26(1), pp. 327–335.

Zakaria, Z.Z., Benslimane, F.M., Nasrallah, G.K., Shurbaji, S., Younes, N.N., Mraiche, F., Da'as, S.I. and Yalcin, H.C. (2018) 'Using Zebrafish for Investigating the Molecular Mechanisms of Drug-Induced Cardiotoxicity', *BioMed research international*, 2018, p. 1642684.

Zeng, L., Carter, A.D. and Childs, S.J. (2009) 'miR-145 directs intestinal maturation in zebrafish', *Proceedings of the National Academy of Sciences of the United States of America*, 106(42), pp. 17793–17798.

Zeng, L. and Childs, S.J. (2012) 'The smooth muscle microRNA miR-145 regulates gut epithelial development via a paracrine mechanism', *Developmental biology*, 367(2), pp. 178–186.

ZFIN The Zebrafish Information Network (20200). Available at: https://zfin.org/ (Accessed: 19 December 2020).

Zhang, B., Li, B., Qin, F., Bai, F., Sun, C. and Liu, Q. (2019) 'Expression of serum microRNA-155 and its clinical importance in patients with heart failure after myocardial infarction', *The Journal of international medical research*, 47(12), pp. 6294–6302.

Zhang, F., Han, L., Wang, J., Shu, M., Liu, K., Zhang, Y., Hsiao, C., Tian, Q. and He, Q. (2021) 'Clozapine Induced Developmental and Cardiac Toxicity on Zebrafish Embryos by Elevating Oxidative Stress', *Cardiovascular toxicology*, 21(5), pp. 399–409.

Zhang, J., Liu, J., Huang, Y., Chang, J.Y., Liu, L., McKeehan, W.L., Martin, J.F. and Wang, F. (2012) 'FRS2α-mediated FGF signals suppress premature differentiation of cardiac stem cells through regulating autophagy activity', *Circulation research*, 110(4), pp. e29–39.

Zhang, L., Sedykh, A., Tripathi, A., Zhu, H., Afantitis, A., Mouchlis, V.D., Melagraki, G., Rusyn, I. and Tropsha, A. (2013) 'Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches', *Toxicology and applied pharmacology*, 272(1), pp. 67–76.

Zhang, P.C., Llach, A., Sheng, X.Y., Hove-Madsen, L. and Tibbits, G.F. (2011) 'Calcium handling in zebrafish ventricular myocytes', *American journal of physiology. Regulatory, integrative and comparative physiology*, 300(1), pp. R56–66.

Zhang, S., Golbraikh, A. and Tropsha, A. (2006) 'Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces', *Journal of medicinal chemistry*, 49(9), pp. 2713–2724.

Zhang, Y., Huang, L., Wang, C., Gao, D. and Zuo, Z. (2013) 'Phenanthrene exposure produces cardiac defects during embryo development of zebrafish (Danio rerio) through activation of MMP-9', *Chemosphere*, 93(6), pp. 1168–1175.

Zhang, Y.H., Xia, Z.N., Yan, L. and Liu, S.S. (2015) 'Prediction of placental barrier permeability: a model based on partial least squares variable selection procedure', *Molecules*, 20(5), pp. 8270–8286.

Zhao, J., Zhou, K., Ma, L. and Zhang, H. (2020) 'MicroRNA-145 overexpression inhibits neuroblastoma tumorigenesis in vitro and in vivo', *Bioengineered*, 11(1), pp. 219–228.

Zhao, W., Zhao, S.-P. and Zhao, Y.-H. (2015) 'MicroRNA-143/-145 in Cardiovascular Diseases', *BioMed research international*, 2015, p. 531740.

Zhao, Y., Samal, E. and Srivastava, D. (2005) 'Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis', *Nature*, 436(7048), pp. 214–220.

Zhong, H., Yang, X., Kaplan, L.M., Molony, C. and Schadt, E.E. (2010) 'Integrating pathway analysis and genetics of gene expression for genome-wide association studies', *American journal of human genetics*, 86(4), pp. 581–591.

Zhou, B. and Tian, R. (2018) 'Mitochondrial dysfunction in pathophysiology of heart failure', *The Journal of clinical investigation*, 128(9), pp. 3716–3726.

Zhou, W., Lin, L., Majumdar, A., Li, X., Zhang, X., Liu, W., Etheridge, L., Shi, Y., Martin, J., Van de Ven, W., Kaartinen, V., Wynshaw-Boris, A., McMahon, A.P., Rosenfeld, M.G. and Evans, S.M. (2007) 'Modulation of morphogenesis by noncanonical Wnt signaling requires ATF/CREB family-mediated transcriptional activation of TGFbeta2', *Nature genetics*, 39(10), pp. 1225–1234.

Zhu, J.J., Xu, Y.Q., He, J.H., Yu, H.P., Huang, C.J., Gao, J.M., Dong, Q.X., Xuan, Y.X. and Li, C.Q. (2014) 'Human cardiotoxic drugs delivered by soaking and microinjection induce cardiovascular toxicity in zebrafish', *Journal of applied toxicology: JAT*, 34(2), pp. 139–148.

Zou, J., Li, W.Q., Li, Q., Li, X.Q., Zhang, J.T., Liu, G.Q., Chen, J., Qiu, X.X., Tian, F.J., Wang, Z.Z., Zhu, N., Qin, Y.W., Shen, B., Liu, T.X. and Jing, Q. (2011) 'Two functional microRNA-126s repress a novel target gene p21-activated kinase 1 to regulate vascular integrity in zebrafish', *Circulation research*, 108(2), pp. 201–209.

Zvinavashe, E., van den Berg, H., Soffers, A.E., Vervoort, J., Freidig, A., Murk, A.J. and Rietjens, I.M.(2008) 'QSAR models for predicting in vivo aquatic toxicity of chlorinated alkanes to fish’, *Chemical research in toxicology*, 21(3), pp. 739–745.

Supplementary material

Chemical concentration- Genes upregulated		
GO term name	**P-value**	**GO term ID**
Regulation of cellular metabolic process	0.02	GO:0031323
Cell projection organization	0.02	GO:0030030
Plasma membrane bounded cell projection organization	0.02	GO:0120036
Regulation of primary metabolic process	0.02	GO:0080090
Regulation of nitrogen compound metabolic process	0.02	GO:0051171
Microtubule-based transport	0.04	GO:0099111
Metencephalon development	0.04	GO:0022037
Cerebellum development	0.05	GO:0021549
Biological_process	0.05	GO:0008150
Nervous system development	0.05	GO:0007399
Purine nucleoside monophosphate metabolic process	0.05	GO:0009126
Regulation of macromolecule biosynthetic process	0.05	GO:0010556
Cell differentiation in hindbrain	0.05	GO:0021533
Purine ribonucleoside monophosphate metabolic process	0.05	GO:0009167
Cellular process	0.05	GO:0009987
Cerebellar cortex formation	0.05	GO:0021697
Cell projection assembly	0.05	GO:0030031
Developmental growth involved in morphogenesis	0.05	GO:0060560
Regulation of cellular process	0.05	GO:0050794
Regulation of nucleobase-containing compound metabolic process	0.05	GO:0019219
Hindbrain development	0.05	GO:0030902
Regulation of cellular biosynthetic process	0.05	GO:0031326
Regulation of biosynthetic process	0.05	GO:0009889
Cellular metabolic process	0.06	GO:0044237

Regulation of DNA-templated transcription, elongation	0.06	GO:0032784
Primary metabolic process	0.07	GO:0044238
Axis elongation	0.07	GO:0003401
Regulation of RNA metabolic process	0.07	GO:0051252
Regulation of RNA biosynthetic process	0.07	GO:2001141
Regulation of nucleic acid-templated transcription	0.07	GO:1903506
Cilium organization	0.07	GO:0044782
Regulation of transcription, DNA-templated	0.07	GO:0006355
Macromolecule biosynthetic process	0.07	GO:0009059
Biosynthetic process	0.07	GO:0009058
Microtubule-based process	0.07	GO:0007017
Organic substance biosynthetic process	0.07	GO:1901576
Neuron differentiation	0.07	GO:0030182
Nucleobase-containing compound biosynthetic process	0.07	GO:0034654
Developmental process	0.07	GO:0032502
Ribonucleoside diphosphate biosynthetic process	0.07	GO:0009188
ADP biosynthetic process	0.07	GO:0006172
Purine nucleoside diphosphate biosynthetic process	0.07	GO:0009136
Negative regulation of transcription by RNA polymerase II	0.07	GO:0000122
Ribonucleoside monophosphate metabolic process	0.07	GO:0009161
RNA biosynthetic process	0.07	GO:0032774
Neurogenesis	0.07	GO:0022008
Transcription, DNA-templated	0.07	GO:0006351
Regulation of macromolecule metabolic process	0.07	GO:0060255
Determination of bilateral symmetry	0.07	GO:0009855
Cerebellar cortex morphogenesis	0.07	GO:0021696
Heterocycle biosynthetic process	0.07	GO:0018130
Cerebellar cortex development	0.07	GO:0021695
Nucleic acid-templated transcription	0.07	GO:0097659

Specification of symmetry	0.07	GO:0009799
Generation of neurons	0.07	GO:0048699
Purine ribonucleoside diphosphate biosynthetic process	0.07	GO:0009180
Aromatic compound biosynthetic process	0.07	GO:0019438
Transport along microtubule	0.07	GO:0010970
Positive regulation of nitrogen compound metabolic process	0.07	GO:0051173
Selective autophagy	0.07	GO:0061912
Cellular biosynthetic process	0.07	GO:0044249
Cellular nitrogen compound biosynthetic process	0.07	GO:0044271
Determination of heart left/right asymmetry	0.08	GO:0061371
Anatomical structure development	0.08	GO:0048856
Organic cyclic compound biosynthetic process	0.08	GO:1901362
Regulation of cyclin-dependent protein serine/threonine kinase activity	0.08	GO:0000079
Protein deubiquitination	0.08	GO:0016579
Convergent extension involved in axis elongation	0.08	GO:0060028
Regulation of cyclin-dependent protein kinase activity	0.08	GO:1904029
Engulfment of apoptotic cell	0.08	GO:0043652
Intraciliary transport	0.08	GO:0042073
Microtubule-based movement	0.08	GO:0007018
Regulation of metabolic process	0.08	GO:0019222
Negative regulation of macromolecule biosynthetic process	0.08	GO:0010558
Positive regulation of macromolecule metabolic process	0.08	GO:0010604
Cell development	0.08	GO:0048468
Brain development	0.09	GO:0007420
Nucleoside diphosphate biosynthetic process	0.09	GO:0009133
Determination of left/right symmetry	0.09	GO:0007368
Regulation of protein kinase A signaling	0.09	GO:0010738
Cerebellar granular layer development	0.09	GO:0021681

Cerebellar granular layer morphogenesis	0.09	GO:0021683
Cerebellar granular layer formation	0.09	GO:0021684
Cerebellar granule cell differentiation	0.09	GO:0021707
Amino sugar metabolic process	0.09	GO:0006040
Plasma membrane bounded cell projection assembly	0.09	GO:0120031
Positive regulation of cell growth	0.09	GO:0030307
Neuron projection development	0.09	GO:0031175
Positive regulation of nucleobase-containing compound metabolic process	0.09	GO:0045935
Regulation of microtubule polymerization or depolymerization	0.09	GO:0031110
Cytoskeleton-dependent intracellular transport	0.09	GO:0030705
Positive regulation of cellular process	0.09	GO:0048522
Positive regulation of cellular metabolic process	0.09	GO:0031325
Neuron development	0.09	GO:0048666
Cellular catabolic process	0.09	GO:0044248
Cerebellum morphogenesis	0.09	GO:0021587
Negative regulation of cellular biosynthetic process	0.09	GO:0031327
Multicellular organism development	0.09	GO:0007275
Negative regulation of biosynthetic process	0.09	GO:0009890
Cilium assembly	0.09	GO:0060271
Head development	0.09	GO:0060322
Catabolic process	0.09	GO:0009056
Ubiquitin-dependent protein catabolic process	0.09	GO:0006511

Table S.1: Functional enrichment analysis of the genes selected by differential expression analysis (SAM), to be upregulated when chemical concentration is increased.

Heart-rate fold changes -Genes upregulated		
GO term name	**P-value**	**GO term ID**
Developmental process	9.16E-12	GO:0008150
Anatomical structure development	3.55E-11	GO:0048856
Multicellular organism development	1.67E-10	GO:0007275
System development	1.29E-09	GO:0048731
Animal organ development	1.73E-08	GO:0048513
Immune System	2.05E-08	REAC:R-DRE-168256
Signal Transduction	2.07E-07	REAC:R-DRE-162582
Cell differentiation	6.22E-06	GO:0030154
Cellular developmental process	6.61E-06	GO:0048869
Cation transport	2.27E-05	GO:0006812
Anatomical structure morphogenesis	4.35E-05	GO:0009653
Nervous system development	6.97E-05	GO:0007399
Transport of small molecules	7.35E-05	REAC:R-DRE-382551
Metal ion transport	8.19E-05	GO:0030001
Extracellular matrix organization	8.78E-05	REAC:R-DRE-1474244
Cell development	0.0001	GO:0048468
PPAR signaling pathway	0.0001	KEGG:03320
Inorganic ion transmembrane transport	0.0001	GO:0098660
Ion transport	0.0002	GO:0006811
Regulation of RNA metabolic process	0.0002	GO:0051252
Inorganic cation transmembrane transport	0.0002	GO:0098662
Regulation of developmental process	0.0004	GO:0050793
Cell projection organization	0.0004	GO:0030030
Wnt signaling pathway	0.0006	KEGG:04310
FoxO signaling pathway	0.0006	KEGG:04068
Cell adhesion molecules	0.0006	KEGG:04514
Localization	0.0010	GO:0051179

Plasma membrane bounded cell projection organization	0.0012	GO:0120036
Neurogenesis	0.0014	GO:0022008
Ion transmembrane transport	0.0014	GO:0034220
RNA biosynthetic process	0.0014	GO:0032774
Regulation of nucleic acid-templated transcription	0.0017	GO:1903506
Regulation of DNA-templated transcription	0.0017	GO:0006355
Regulation of RNA biosynthetic process	0.0017	GO:2001141
DNA-templated transcription	0.0019	GO:0006351
Nucleic acid-templated transcription	0.0019	GO:0097659
Inorganic ion homeostasis	0.0025	GO:0098771
mTOR signaling pathway	0.0025	KEGG:04150
Fatty acid metabolism	0.0025	KEGG:01212
Cellular ion homeostasis	0.0026	GO:0006873
Ion homeostasis	0.0036	GO:0050801
Homeostatic process	0.0040	GO:0042592
Transport	0.004	GO:0006810
Cation transmembrane transport	0.004	GO:0098655
Generation of neurons	0.004	GO:0048699
Establishment of localization	0.004	GO:0051234
Cellular cation homeostasis	0.005	GO:0030003
Adaptive Immune System	0.005	REAC:R-DRE-1280218
Circulatory system development	0.005	GO:0072359
Phosphatidylinositol signaling system	0.006	KEGG:04070
Calcium signaling pathway	0.006	KEGG:04020
Insulin signaling pathway	0.006	KEGG:04910
Regulation of actin cytoskeleton	0.006	KEGG:04810
Purine metabolism	0.006	KEGG:00230
Hindbrain development	0.006	GO:0030902

Cation homeostasis	0.006	GO:0055080
RNA Polymerase II Transcription	0.007	REAC:R-DRE-73857
Post-translational protein modification	0.007	REAC:R-DRE-597592
Potassium ion transmembrane transport	0.007	GO:0071805
Innate Immune System	0.008	REAC:R-DRE-168249
Protein modification process	0.009	GO:0036211
Transmembrane transport	0.009	GO:0055085
Regulation of multicellular organismal development	0.009	GO:2000026
SLC-mediated transmembrane transport	0.009	REAC:R-DRE-425407
Membrane Trafficking	0.01	REAC:R-DRE-199991
Tissue development	0.01	GO:0009888
Neuron differentiation	0.01	GO:0030182
Adipocytokine signaling pathway	0.011	KEGG:04920
Adherens junction	0.011	KEGG:04520
ECM-receptor interaction	0.011	KEGG:04512
Cell morphogenesis	0.011	GO:0000902
Gene expression (Transcription)	0.011	REAC:R-DRE-74160
Chemical homeostasis	0.015	GO:0048878
Macromolecule modification	0.015	GO:0043412
Cytokine Signaling in Immune system	0.015	REAC:R-DRE-1280215
Regulation of gene expression	0.016	GO:0010468
Embryo development	0.016	GO:0009790
Neuron development	0.018	GO:0048666
Stem cell differentiation	0.025	GO:0048863
Cellular chemical homeostasis	0.025	GO:0055082
Cell cycle	0.026	KEGG:04110
Signaling by Receptor Tyrosine Kinases	0.027	REAC:R-DRE-9006934
Somite development	0.027	GO:0061053
Apelin signaling pathway	0.029	KEGG:04371

MAPK signaling pathway	0.03	KEGG:04010
Brain development	0.036	GO:0007420
Cardiac conduction	0.038	REAC:R-DRE-5576891
Cellular component morphogenesis	0.038	GO:0032989
Class I MHC mediated antigen processing & presentation	0.038	REAC:R-DRE-983169
Muscle contraction	0.039	REAC:R-DRE-397014
Degradation of the extracellular matrix	0.039	REAC:R-DRE-1474228
Hematopoietic stem cell differentiation	0.039	GO:0060218
Transcription by RNA polymerase II	0.039	GO:0006366
Negative regulation of RNA metabolic process	0.039	GO:0051253
Animal organ morphogenesis	0.04	GO:0009887
Cell morphogenesis involved in differentiation	0.042	GO:0000904
Hematopoietic progenitor cell differentiation	0.042	GO:0002244
Head development	0.043	GO:0060322
Notch signaling pathway	0.044	KEGG:04330
Homologous recombination	0.044	KEGG:03440
Base excision repair	0.044	KEGG:03410
Heart process	0.045	GO:0003015
Cell adhesion	0.045	GO:0007155
Neuron projection development	0.046	GO:0031175
Sprouting angiogenesis	0.046	GO:0002040
Negative regulation of DNA-templated transcription	0.046	GO:0045892
Negative regulation of RNA biosynthetic process	0.046	GO:1902679
Negative regulation of nucleic acid-templated Transcription	0.046	GO:1903507
Metal ion homeostasis	0.046	GO:0055065
Platelet activation, signaling and aggregation	0.048	REAC:R-DRE-76002
RNA degradation	0.049	KEGG:03018
Ion channel transport	0.05	REAC:R-DRE-983712

Vesicle-mediated transport	0.05	REAC:R-DRE-5653656
Cell communication	0.051	GO:0007154
Cellular metal ion homeostasis	0.053	GO:0006875
Negative regulation of hematopoietic progenitor cell differentiation	0.053	GO:1901533
Regulation of transcription by RNA polymerase II	0.053	GO:0006357
Neuron recognition	0.053	GO:0008038
NOD-like receptor signaling pathway	0.054	KEGG:04621
Protein processing in endoplasmic reticulum	0.054	KEGG:04141
Axon development	0.056	GO:0061564
Cardiac muscle contraction	0.058	KEGG:04260
Potassium ion transport	0.059	GO:0006813
Cell junction organization	0.059	GO:0034330
Cell junction organization	0.059	GO:0034330
Ferroptosis	0.061	KEGG:04216
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	0.062	REAC:R-DRE-72662
Toll-like receptor signaling pathway	0.062	KEGG:04620
Embryo development ending in birth or egg hatching	0.062	GO:0009792
Chordate embryonic development	0.063	GO:0043009
Vasculogenesis	0.064	GO:0001570
Developmental growth involved in morphogenesis	0.064	GO:0060560
Transport to the Golgi and subsequent modification	0.064	REAC:R-DRE-948021
ER to Golgi Anterograde Transport	0.064	REAC:R-DRE-199977
Signal transduction	0.065	GO:0007165
Axonogenesis	0.065	GO:0007409
Signaling by Hedgehog	0.065	REAC:R-DRE-5358351
Mesenchymal cell differentiation	0.066	GO:0048762
Cellular homeostasis	0.066	GO:0019725

Voltage gated Potassium channels	0.066	REAC:R-DRE-1296072
ErbB signaling pathway	0.066	KEGG:04012
Myeloid cell homeostasis	0.066	GO:0002262
Canonical Wnt signaling pathway	0.067	GO:0060070
T cell proliferation	0.067	GO:0042098
Cellular monovalent inorganic cation homeostasis	0.069	GO:0030004
Neuronal System	0.069	REAC:R-DRE-112316
C-type lectin receptor signaling pathway	0.07	KEGG:04625
Regulation of T cell activation	0.073	GO:0042129
Regulation of ion transmembrane transport	0.077	GO:0034765
TGF-beta signaling pathway	0.077	GO:0019219
Mesenchyme development	0.08	GO:0060485
Regulation of nervous system development	0.08	GO:0051960
Regulation of ion transport	0.08	GO:0043269
Heart contraction	0.08	GO:0060047
Immune system development	0.08	GO:0002520
Cell-cell signaling by wnt	0.08	GO:0198738
Wnt signaling pathway	0.08	GO:0016055
Regulation of lymphocyte proliferation	0.08	GO:0050670
Cell projection assembly	0.08	GO:0030031
Negative regulation of cell-cell adhesion	0.08	GO:0022408
Signaling by GPCR	0.081	REAC:R-DRE-372790
Presynaptic phase of homologous DNA pairing and strand exchange	0.081	REAC:R-DRE-5693616
Nervous system development	0.081	REAC:R-DRE-9675108
G alpha (i) signalling events	0.081	REAC:R-DRE-418594
Homologous DNA Pairing and Strand Exchange	0.081	REAC:R-DRE-5693579
Receptor-type tyrosine-protein phosphatases	0.081	REAC:R-DRE-388844
Ribosomal scanning and start codon recognition	0.081	REAC:R-DRE-72702

Regulation of RNA splicing	0.083	GO:0043484
Erythrocyte differentiation	0.083	GO:0030218
Cell junction assembly	0.083	GO:0034329
Heart development	0.085	GO:0007507
Neuron projection guidance	0.086	GO:0097485
Stabilization of membrane potential	0.087	GO:0030322
Regulation of transmembrane transport	0.087	GO:0034762
Anatomical structure formation involved in morphogenesis	0.087	GO:0048646
Regulation of T cell proliferation	0.087	GO:0042129
IMP biosynthetic process	0.087	GO:0006188
IMP metabolic process	0.087	GO:0046040
GPCR downstream signalling	0.088	REAC:R-DRE-388396
YAP1- and WWTR1 (TAZ)-stimulated gene expression	0.088	REAC:R-DRE-2032785
Regulation of TP53 Activity	0.088	REAC:R-DRE-5633007
Sensory system development	0.089	GO:0048880
Hemopoiesis	0.089	GO:0030097
Positive regulation of immune effector process	0.089	GO:0002699
Metencephalon development	0.089	GO:0022037
Ubiquitin mediated proteolysis	0.092	KEGG:04120
Erythrocyte homeostasis	0.092	GO:0034101
Sodium ion transport	0.094	GO:0006814
Cell projection morphogenesis	0.094	GO:0048858
Organelle assembly	0.094	GO:0070925
VEGF signaling pathway	0.096	KEGG:04370
Mitophagy - animal	0.096	KEGG:04137
Citrate cycle (TCA cycle)	0.098	KEGG:00020
Heart-rate fold changes -Genes downregulated		
Endocytosis	5.42E-09	KEGG:04144

Animal organ development	3.26E-05	GO:0048513
Anatomical structure morphogenesis	3.52E-05	GO:0009653
MAPK signaling pathway	4.42E-05	KEGG:04010
Apoptotic cleavage of cellular proteins	6.76E-05	REAC:R-DRE-111465
Metabolism of RNA	7.62E-05	REAC:R-DRE-8953854
Metabolism of proteins	0.0003	REAC:R-DRE-392499
Adrenergic signaling in cardiomyocytes	0.0005	KEGG:04261
Neuroactive ligand-receptor interaction	0.001	KEGG:04080
Nervous system development	0.001	GO:0007399
Regulation of actin cytoskeleton	0.001	KEGG:04810
Tissue development	0.001	GO:0009888
FoxO signaling pathway	0.001	KEGG:04068
Post-translational protein modification	0.001	REAC:R-DRE-597592
Developmental Biology	0.001	REAC:R-DRE-1266738
Processing of Capped Intron-Containing Pre-mRNA	0.001	REAC:R-DRE-72203
Innate Immune System	0.002	REAC:R-DRE-168249
p53 signaling pathway	0.002	KEGG:04115
Cytokine-cytokine receptor interaction	0.002	KEGG:04060
Apoptosis	0.002	REAC:R-DRE-109581
Insulin signaling pathway	0.002	KEGG:04910
Programmed Cell Death	0.002	REAC:R-DRE-5357801
Cellular senescence	0.002	KEGG:04218
Gene expression (Transcription)	0.003	REAC:R-DRE-74160
Apelin signaling pathway	0.003	KEGG:04371
Calcium signaling pathway	0.003	KEGG:04020
Oocyte meiosis	0.003	KEGG:04114
Membrane Trafficking	0.003	REAC:R-DRE-199991
Epithelium development	0.004	GO:0060429
ErbB signaling pathway	0.004	KEGG:04012

Protein processing in endoplasmic reticulum	0.004	KEGG:04141
mRNA Splicing	0.004	REAC:R-DRE-72172
Cell development	0.005	GO:0048468
Spliceosome	0.005	KEGG:03040
Apoptotic execution phase	0.006	REAC:R-DRE-75153
mRNA Splicing - Major Pathway	0.006	REAC:R-DRE-72163
Transport	0.006	GO:0006810
Regulation of protein metabolic process	0.006	GO:0051246
Cytokine Signaling in Immune system	0.007	REAC:R-DRE-1280215
Gap junction	0.008	KEGG:04540
Vascular smooth muscle contraction	0.008	KEGG:04270
Regulation of cell junction assembly	0.008	GO:1901888
Regulation of synapse assembly	0.009	GO:0051963
Embryo development	0.012	GO:0009790
Lysosome	0.014	KEGG:04142
Neuron differentiation	0.014	GO:0030182
Ribosome	0.015	KEGG:03010
Apoptosis	0.017	KEGG:04210
Circulatory system development	0.018	GO:0072359
Pyruvate metabolism	0.018	KEGG:00620
Central nervous system development	0.018	GO:0007417
RNA Polymerase II Transcription	0.018	REAC:R-DRE-73857
Vesicle localization	0.018	GO:0051648
Response to stimulus	0.02	GO:0050896
Germ cell development	0.02	GO:0007281
Neurogenesis	0.02	GO:0022008
Tube development	0.02	GO:0035295
Actin filament-based transport	0.02	GO:0099515
Heart development	0.02	GO:0007507

Nervous system development	0.021	REAC:R-DRE-9675108
Axon guidance	0.021	REAC:R-DRE-422475
Generation of neurons	0.021	GO:0048699
Cardiac muscle contraction	0.023	KEGG:04260
Adherens junction	0.023	KEGG:04520
Transmembrane transport	0.024	GO:0055085
GnRH signaling pathway	0.026	KEGG:04912
Tight junction	0.026	KEGG:04530
TGF-beta signaling pathway	0.026	KEGG:04350
Biosynthesis of amino acids	0.026	KEGG:01230
Signaling by Receptor Tyrosine Kinases	0.027	REAC:R-DRE-9006934
Fatty acid metabolism	0.027	KEGG:01212
Vasopressin-like receptors	0.028	REAC:R-DRE-388479
Neutrophil degranulation	0.028	REAC:R-DRE-6798695
Organelle organization	0.031	GO:0006996
Valine, leucine and isoleucine degradation	0.035	KEGG:00280
Mismatch repair	0.035	KEGG:03430
Regulation of cellular component organization	0.038	GO:0051128
Lysine degradation	0.04	KEGG:00310
Focal adhesion	0.042	KEGG:04510
Nucleotide Excision Repair	0.042	REAC:R-DRE-5696398
Neuronal System	0.043	REAC:R-DRE-112316
Signaling by VEGF	0.044	REAC:R-DRE-194138
Nucleotide excision repair	0.046	KEGG:03420
Organelle localization	0.053	GO:0051640
Signal transduction	0.053	GO:0007165
Regulation of protein modification process	0.053	GO:0031399
Ion transmembrane transport	0.053	GO:0034220
Phosphatidylinositol signaling system	0.054	KEGG:04070

Plasma membrane bounded cell projection organization	0.056	GO:0120036
Ribosome biogenesis in eukaryotes	0.057	KEGG:03008
Toll-like receptor signaling pathway	0.057	KEGG:04620
VEGFA-VEGFR2 Pathway	0.058	REAC:R-DRE-4420097
Signaling by Interleukins	0.058	REAC:R-DRE-449147
GPCR downstream signalling	0.058	REAC:R-DRE-388396
DNA replication	0.059	KEGG:03030
Adaptive Immune System	0.059	REAC:R-DRE-1280218
Death Receptor Signalling	0.059	REAC:R-DRE-73887
Post-translational protein phosphorylation	0.066	REAC:R-DRE-8957275
mTOR signaling pathway	0.067	KEGG:04150
Adipocytokine signaling pathway	0.067	KEGG:04920
TCR signaling	0.069	REAC:R-DRE-202403
Autophagy	0.069	REAC:R-DRE-9612973
Small GTPase mediated signal transduction	0.07	GO:0007264
VEGF signaling pathway	0.073	KEGG:04370
PPAR signaling pathway	0.073	KEGG:03320
Pantothenate and CoA biosynthesis	0.074	KEGG:00770
Fatty acid elongation	0.074	KEGG:00062
Fatty acid degradation	0.074	KEGG:00071
Cell projection organization	0.074	GO:0030030
Morphogenesis of an epithelium	0.074	GO:0002009
Cellular component morphogenesis	0.076	GO:0032989
Base excision repair	0.078	KEGG:03410
Cellular localization	0.079	GO:0051641
Cation transmembrane transport	0.079	GO:0098655
Striated Muscle Contraction	0.079	REAC:R-DRE-390522
MHC class II antigen presentation	0.081	REAC:R-DRE-2132295

Signaling by GPCR	0.081	REAC:R-DRE-372790
FLT3 Signaling	0.082	REAC:R-DRE-9607240
Axonogenesis	0.086	GO:0007409
ncRNA metabolic process	0.086	GO:0034660
Tissue morphogenesis	0.09	GO:0048729
Wnt signaling pathway	0.09	KEGG:04310
mRNA surveillance pathway	0.09	KEGG:03015
Cell adhesion molecules	0.09	KEGG:04514
Peroxisome	0.09	KEGG:04146
Establishment of organelle localization	0.093	GO:0051656
Golgi vesicle transport	0.095	GO:0048193
Phagosome	0.095	KEGG:04145

Table S.2: Functional enrichment analysis of the upregulated and downregulated genes associated with significant heart rate fold change.

GO term name	P-value	GO term ID
Purine metabolism	0.017	KEGG:00230
Butanoate metabolism	0.025	KEGG:00650
Homologous recombination	0.026	KEGG:03440
ABC transporters	0.027	KEGG:02010
Glycerolipid metabolism	0.034	KEGG:00561
PPAR signaling pathway	0.04	KEGG:03320
Peroxisome	0.04	KEGG:04146
Interconversion of 2-oxoglutarate and 2-hydroxyglutarate	0.047	REAC:R-DRE-880009
IL-6-type cytokine receptor ligand interactions	0.06	REAC:R-DRE-6788467
Interleukin-6 family signaling	0.06	REAC:R-DRE-6783589
Triglyceride biosynthesis	0.06	REAC:R-DRE-75109
Triglyceride metabolism	0.065	REAC:R-DRE-8979227
Resolution of D-Loop Structures	0.065	REAC:R-DRE-5693537
Metalloprotease DUBs	0.065	REAC:R-DRE-5689901
Collagen chain trimerization	0.065	REAC:R-DRE-8948216
Resolution of D-loop Structures through Holliday Junction Intermediates	0.065	REAC:R-DRE-5693568
Processing of DNA double-strand break ends	0.068	REAC:R-DRE-5693607
Presynaptic phase of homologous DNA pairing and strand exchange	0.068	REAC:R-DRE-5693616
HDR through Single Strand Annealing (SSA)	0.068	REAC:R-DRE-5685938
Homologous DNA Pairing and Strand Exchange	0.069	REAC:R-DRE-5693579
Collagen biosynthesis and modifying enzymes	0.072	REAC:R-DRE-1650814
DNA Double Strand Break Response	0.072	REAC:R-DRE-5693606
Recruitment and ATM-mediated phosphorylation of repair and signaling proteins at DNA double strand breaks	0.072	REAC:R-DRE-5693565
Pyruvate metabolism and Citric Acid (TCA) cycle	0.072	REAC:R-DRE-71406
Collagen formation	0.077	REAC:R-DRE-1474290

G2/M DNA damage checkpoint	0.077	REAC:R-DRE-69473
HDR through Homologous Recombination (HRR)	0.077	REAC:R-DRE-5685942
HDR through Homologous Recombination (HRR) or Single Strand Annealing (SSA)	0.077	REAC:R-DRE-5693567
Homology Directed Repair	0.077	REAC:R-DRE-5693538
UCH proteinases	0.077	REAC:R-DRE-5689603
Regulation of TP53 Activity through Phosphorylation	0.077	REAC:R-DRE-6804756
Mitochondrial translation termination	0.08	REAC:R-DRE-5419276
Mitochondrial translation elongation	0.08	REAC:R-DRE-5389840
Mitochondrial translation	0.08	REAC:R-DRE-5368287
Positive regulation of tyrosine phosphorylation of STAT protein	0.099	GO:0042531
IMP biosynthetic process	0.099	GO:0006188
Alditol phosphate metabolic process	0.099	GO:0052646
Glycerol metabolic process	0.099	GO:0006071
Purine ribonucleotide salvage	0.099	GO:0106380
Glycerol-3-phosphate metabolic process	0.099	GO:0006072
Amino-acid betaine metabolic process	0.099	GO:0006577
IMP metabolic process	0.099	GO:0046040
Carnitine metabolic process	0.099	GO:0009437
Carnitine metabolic process, CoA-linked	0.099	GO:0019254
Purine-containing compound salvage	0.099	GO:0043101
Alditol metabolic process	0.099	GO:0019400
AMP metabolic process	0.099	GO:0046033
Alditol catabolic process	0.099	GO:0019405
Glycerol catabolic process	0.099	GO:0019563
Organophosphate metabolic process	0.099	GO:0019637
Purine nucleotide salvage	0.099	GO:0032261
IMP salvage	0.099	GO:0032264
Tyrosine phosphorylation of STAT protein	0.099	GO:0007260

Regulation of tyrosine phosphorylation of STAT protein	0.099	GO:0042509
Glycerol-3-phosphate biosynthetic process	0.099	GO:0046167

Figure S.3: All the GO terms associated with the genes identified by predictive modeling using cluster 3 chemicals.