Moving Towards Predictive Toxicology - A Systems Biology Approach

by

Philipp Antczak

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy (Sc, PhD)

College of Life and Environmental Sciences
School of Biosciences
The University of Birmingham
September 2011

Abstract

Human health and the environment are at risk of being exposed to a large number of hazardous chemicals each day. Unfortunately, many of these chemicals have no or little recorded toxicity information. Predictive toxicology aims to provide tools and methodologies to address this issue. In combination with systems biology approaches these can provide a powerful toolbox for understanding the impact of chemicals on biological species.

The work presented within this thesis demonstrates the effectiveness of such approaches in the context of industrial and environmentally relevant species. More specifically we focus on characterization of a general toxicity mechanism in *Rattus norvegius* and *Daphnia magna* as well as for the first time demonstrate that the transcriptional response of *D. magna* is predictive not only of chemical class but also of measured toxicity. We also show that inclusion of pathway-level information can increase biological interpretability in non-model species. Lastly, we provide evidence supporting the application of reverse engineering methodologies in the context of identifying adverse outcome pathways in *Pimephales promelas*, an environmentally relevant species.

Ultimately, our results have shown that these approaches can provide highly relevant information in model and non-model species. Further development building on these results could potentially lead to improvements in risk assessment and environmental monitoring.

Acknowledgements

First of all, I would like to dedicate this thesis to my loving wife, Kirsten Antczak, and my beautiful daughter, Lena Antczak. Their support during the PhD write-up stage has been indispensible providing me with smiles, love, care and support.

With all of my heart I would to thank my parents, Wojciech and Wioleta Antczak, for always supporting, loving and pushing me where it was needed to achieve to the best of my abilities. Their care, understanding and sacrifice to provide me with the best outlook on the future will guide me for the rest of my life. I would also like to thank my brother Lukas Antczak, who has supported me and has always been there whenever I needed him.

To the rest of my family, especially my grandparents, who have taught me many things during my life. My uncle and aunt, Krzysiek and Agnieszka Sutkowski for their support and help. My remaining uncles, aunts, cousins, and extended family members who have given me so much joy.

Professionally I would like to thank my supervisor and mentor, Dr. Francesco Falciani, for his invaluable contribution to this thesis, his guidance, patience and expertise that has given me the confidence in this field. To the head of school, Prof. Kevin Chipman, for providing his expertise towards parts of my work and to Prof. Mark Viant and Prof. Chris Vulpe for their collaborative input during my PhD.

Finally I would like to thank my friends and collegues for their professional and intellectual input in this work. Nil Turan-Jurdzinski for being a friend and colleague to discuss matters of personal and professional interest. Anna Stincone for introducing me and the rest of the group to ”but why?” and discussions of various natures. To Wazeer Varsally for the many interesting

shared inventions and captions. I would also like to thank the rest of the group, Kim Clarke, Rita Gupta, Jaanika Kronberg, Helani Munasinghe and Peter Li.

CONTENTS

List of Figures

List of Tables

Glossary

ANN	Artificial Neural Networks
AOP	Adverse Outcome Pathway
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
CLR	Context Likelihood of Relatedness
CoMFA	Comparative Molecular Field Analysis
CoMMA	Comparative Molecular Moment Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis
DAVID	Database for Annotation, Visualization and Integrated Discovery
DEREK	Deductive Estimate of Risk from Existing Knowledge
DT	Decision Trees
EPA	Environmental Protection Agency
FLU	flutamide
GA	Genetic Algorithm
GO	Gene Ontology
GRIND	Grid-Independent Descriptors

KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
MI	Mutual Information
MRNET	Multicast/Reduction Network
NOAEL	No Observable Adverse Effect Level
ODE	Ordinary Differential Equations
PCA	Principal Component Analysis
PCF	Physico-Chemical Feature
PLS	Partial least squares
QSAR	Quantitative Structure-Activity Relationship
RSM	Random Subspace Method
SA	Simulated Annealing
SVM	Support Vector Machines
WHIM	Weighted Holistic Invariant Molecular Descriptors

Scope of Thesis

Currently, the state of the art in predictive toxicology is based on a strategy combining the strengths of modern omics technologies with statistical modelling techniques borrowed from the machine learning community [1]. More specifically, a number of groups have shown that the integration of classical endpoint measurements, such as toxicity or reproduction, with these advanced computational methods provides a suitable framework for identifying relatively small subsets of molecular measurements, predictive of toxicological response. Despite their evident success, a number of challenges remained open:

1. Biomarkers based on optimized variable subsets provide relatively little knowledge of the underlying mechanism involved in the response.
2. The small number of samples available to train the classifiers can result in models that cannot be generalized.
3. Predictive models do not allow simulation of unforeseen outcomes and therefore do not provide a tool for *in silico* testing of remediation strategies.

The work described in this thesis directly address challenges 1 and 2 and set up the scene for the development of a comprehensive modelling platform mentioned in challenge 3. In summary, this work has demonstrated the following principles:

1. Biological interpretation is highly improved by integrating functional modules in the analysis pipeline, especially with non-model species
2. Hypotheses are strengthened and refined when combining results from multiple analysis approaches
3. Identification of novel adverse outcome pathways is heavily facilitated by reverse engineering of regulatory networks in the context of ecotoxicology

More specifically, we demonstrate these by utilizing three relevant species in risk assessment and environmental monitoring. Application of principles 1 and 2 to *Rattus norvegius* and *Daphnia magna* suggest that components of a general toxicity mechanism are shared between the two distant species. Furthermore, for the first time we apply machine learning methodologies to *D. magna* to identify models predictive of chemical class and measured toxicity. Lastly we characterize a novel adverse outcome pathway in *Pimephales promelas* response to flutamide, an anti-androgen. Ultimately, we believe our identified models can provide essential knowledge towards risk assessment and environmental monitoring. Consequently an extension of the work presented here has been successful in securing a Natural Environment Research Council (NERC) grant to develop multi-biomarker assays using *D. magna* for water quality assessment.

Chapter 1

Introduction and Background

1.1 Introduction to Predictive Toxicology

Predictive toxicology aims to provide quantitative tools to assess the hazard chemicals may pose to human health and the environment. In the last several decades this implied exposing large number of animals to high-doses of a stressor in mostly acute experimental setups. Uncertainty factors could then be applied to extrapolate potential adverse outcomes to humans (pharmaceutical toxicology assessment) or environmental populations (ecotoxicology). These factors account for the differences in metabolism and susceptibility to toxicity between the different species. In addition molecular biomarkers based on the understanding of the underlying mechanisms have been developed to assess potential toxicity and for use in environmental monitoring. Among the first attempts of predicting toxicity was the development of quantitative structure-activity relationship (QSAR) analysis. The purpose of these approaches was to predict biological activity, such as toxicity concentration, given only the physico-chemical features (PCFs) derived from the structure of a given chemical. Since the early 1960s technological and computational improvements have contributed to higher levels of sophistication and number of PCFs that can be calculated by current applications. Recent reports have also shown that prediction accuracy of such models can be increased when additional information from cell-based assays is used [2, 3]. In parallel the identification of molecular markers predictive of toxicity has provided a powerful approach to the identification of early toxicity events. Toxicogenomics is the

application of omics technologies to the area of toxicology. It enables simultaneous measurement of thousands of mRNAs, proteins and metabolites in single experiments. The application of bioinformatics approaches utilizing the vast information from omics datasets has proven to be highly beneficial in the identification of biomarkers [4–7]. Steiner et al [4], for example, used a supervised classification algorithm to identify patterns of gene expression profiles that classify hepatotoxic and non-hepatotoxic chemicals. Other successes in predicting toxicity using gene expression data have also been published by Bulera et al [5], Hamadeh et al [8] and others [6, 9]. Tan et al [10] used time-course gene expression profiles to successfully characterize cadmium acetate cytotoxicity and identified a set of potential biomarkers and build a hypothetical pathway that may be representative of exposure and Ellinger-Ziegelbauer et al [7] identified functional terms which were linked to genotoxic and non-genotoxic chemicals. These results, even though promising, were not able to fully provide mechanistic insights to the observed toxicity. In response to this the idea of adverse outcome pathways has been proposed [11, 12]. These utilize reverse engineering methodologies to identify the underlying regulatory network. Further application of computational biology tools, such as functional modularization or variable selection algorithms, can identify sub-networks, which are not only predictive of toxicity but can also propose a potential mechanism of action. In collaboration, Perkins et al [13] published the application of these techniques to *Pimephales promelas* (Fathead Minnow) exposed to flutamide, an androgen receptor. We identified a set of candidate pathways which lead to the formation of testable hypotheses about biological processes, biomarkers or alternative endpoints for monitoring purposes. The importance of understanding and monitoring the effects of hazardous chemicals has been well received by regulatory authorities. In Europe alone around 30,000 chemicals are produced with little or no toxicity data currently registered [14]. In response to this the European commission approved the REACH legislation in 2007, which aims to make registration mandatory for both future and existing chemicals [15]. This poses to be a great challenge to industrial and non-profit organisations alike. In this context, toxicity testing methodologies need to be improved [16–18]. A change in the toxicity paradigm has already started with EVCAM, ICCVAM and other groups developing *in vitro* tests predictive of acute

exposure, repeat dosing, or target specific *in vivo* toxicity tests [19–21]. The U.S. Epa ToxCast program [22], which integrates a number of high throughput testing approaches and computational methods, utilizes *in vitro* methodologies to prioritize chemicals for subsequent *in vivo* investigations. Such systems will provide the knowledge to transform current toxicity testing approaches to reduce the number of animals used while testing an increased number of chemicals.

1.2 Biological Systems of Relevance

In the field of toxicology animal-based tests have always been the gold standard to assess biological response. Depending on the type of toxicity (i.e. environmental hazard, hepatoxicity, cancer or reproductive/development) specific experimental designs were developed. Most of the human toxicity assessment has been performed in rodents such as *Rattus norvegius* (rat) or *Mus musculus* (Mouse). More recently, other species have been included in toxicity testing portfolios such as *Danio rerio* (zebrafish), *Xenopus laevis* (African clawed frog) and the non-model Daphnia Species (water flea). In the next sections an outline of two of these species (*Rattus norvegius* and *Daphnia magna*) is provided. These have been the bases of the work described in this thesis.

1.2.1 Studies in *Rattus Norvegicus*

For many years *Rattus norvegius* (Rat) has been one of the systems of choice in human toxicity assessment. Their high reproduction rate, known genetic backgrounds and similarity to human biology were the driving argument for toxicity testing. In most cases human toxicity was extrapolated by establishing a No Observable Adverse Effect Level (NOAEL). Due to the high reproduction rate, generation and genetic effects as a result of exposure have also been studied [23–25]. The focus on this particular species is also evident when examining the sheer volume of available omics datasets in the public domain. In particular some very large datasets, with hundreds of chemical exposures, have been published by Iconix (Pharmaceutical Industry). In particular the work done by Fielden et al in hepatocarcinogenity [26, 27] and renal tubular degeneration models [28] both of which were used to derive predictive and mechanistic mark-

ers have been of great importance. In many other cases such as the earlier described Steiner et al [4], Tan et al [10] and other publications [6, 7, 9] rodents have been the focus in toxicity testing. In the computational toxicology field the rat has been equally important in providing biological endpoints for regression models based on QSAR [29, 30].

1.2.2 *Daphnia Magna*

The field of ecotoxicology traditionally assess the effect of anthropogenic chemicals on the environment [31]. Highly susceptible are fresh water habitats especially those close to human populations [32]. Sewage and increased industrial activity are among the major causes for environmental pollution. The small waterflea *D. magna* is one of the oldest systems used in biological research [33]. It is widely geographically available and highly adaptive and sensitive to chemical stressors [34, 35]. Yet, it has only recently been added to OECD and U.S. EPA as a model organism for toxicity testing, possibly due to the lack of its genome and genetic research [33]. It is a small planktonic crustacean, which grows up to 5mm in length and has a short life cycle. *D. magna* is easy and cost-effective to cultivate and can be easily used to monitor morphological changes in response to exposure due to its transparent shell [33]. Reproduction can be mediated through cloning or egg deposition. The latter of which can be found in harbour sediments and utilized to study genetic adaptation in respect to change in environmental stressors [36]. Furthermore the Daphnia species is central to the freshwater food webs playing a crucial part in linking environmental problems to dietary components of fish and invertebrate predators. These properties contribute to its applicability as a biosensor to closely monitor the environment on an ecosystem-level. Although the *Daphnia magna* Genome has not yet been released to the public domain, the Daphnia Genome Consortium [37] is preparing the manuscript for publication. The same group has also recently made the *Daphnia pulex*, a close relative to *Daphnia magna*, genome available [38]. The study identified a large number of genes (about 30,000) which exceeds that of many other species. The authors attribute many of these to lineage-specific gene families and a large number of paralogs that have occurred due to duplication. Furthermore gene expression profiles of many of these duplicated genes were

not correlated to their counterpart suggesting a change in function of these paralogs. Other groups have also identified a phenomenon called endopolyploidy in Daphnia using flow cytometry where between 7% to 12% and 14% to 37% of neonates (< 24 hours old) and adults (> 10 days old) respectively showed tetraploid nuclei [39]. The relative close phylogenetic distance of *D. pulex* to *D. magna* may suggest that these processes could also be observed in the latter. These attributes most likely contribute to the amazing phenotypic plasticity that has been observed in the Daphnia species [40]. There are a number of publications available in the public domain that cover a wide range of acute and chronic chemical exposures measuring ecologically relevant endpoints. Poynton et al [41] identified 2 metallothioneins and a ferritin which were highlighted in their metal exposure study. Follow-up studies by the same authors showed that these can be indicative of real-life exposure [42]. Studies performed by Taylor et al [43] also showed that metabolomics on single adult daphnids or 30 neonates can provide biomarkers. They identified N-acetylspermidine as a potential novel biomarker of copper toxicity.

1.2.3 Other Commonly Used Species

In addition to the two species outlined above many other species are used for toxicity testing. In respect to human health mammals, such as mouse, rat, guinea pig, rabbit and higher primates, are being used in laboratories around the world. Each of these species has a particular field of value with mouse and rat being particularly good models for hepato and nephrotoxicity, guinea pigs for immune system and lung tissue, rabbits for ophthalmology and primates primarily for neurological and dermatological areas. In ecotoxicology, however, the areas are not as clear. Different fish species around the world are used to study the effects of chemicals in their natural environment. For example, in Europe, the stickleback (*Gasterosteus aculeatus*) and european flounder (*Platichthys esus*) have been successfully used in predictive toxicology [44–46]. For laboratory experiments the zebrafish (*Danio rerio*) is a favourite among the scientific community due to its well understood developmental behaviours, rapid embryonic development and transparent embryos, just to name a few advantages. The fathead minnow (*Pimephales promelas*) is a particular favourite in the Americas. Its wide distribution across the U.S. river-network,

robustness, animal availability and flexibility contributes towards its use in mechanistic studies and environmental monitoring. The U.S. Army ERDC, for example, have utilized this species to create a large compendium of transcriptomics data representing exposures to endocrine disruptors and explosive compounds [13]. In Chapter 5 we utilize this dataset to demonstrate the effectiveness of reverse engineering approaches in identified novel adverse outcome pathways for flutamide, a model anti-androgen. For many non-model organisms, especially in ecotoxicology, functional annotation can be a limiting factor hence hindering biological interpretation.

1.2.4 Alternatives to Animal Testing

One of the biggest challenges for the toxicological community is the reduction of *in vivo* experimental designs. This has been a growing concern as in recent years the similarity of rat to human biology has been challenged [47–49]. Approaches such as the ToxCast Program employed by the U.S. EPA are trying to address these issues by integrating *in silico*, *in vitro* and non-mammalian *in vivo* systems through high throughput assays to provide information to predict toxicity outcome in humans [22]. In their initial 2-year study (phase 1) the program characterized a total of 300 well studied chemicals (mainly pesticides). The resulting toxicity signatures are now being evaluated on their predictive ability on over 1000 compounds including consumer-end-products, food additives and drugs which have not been released to the market [50]. Most current, however, is the notion of using exclusively *in vitro* cell based systems to predict toxicity. The advantage of using such systems is the ability to perform toxicity testing in the relevant species, discarding interspecies differences, and to automate it into large-scale methods necessary for providing toxicity information for the growing number of chemicals each year. Both Heng et al [51] and also recently Laustriat et al [52] have commented on the use of pluripotent cells in drug discovery and concluded that protocols and techniques have to be standardized to create viable toxicity testing platforms. To be more generally applicable, however, these methodologies must show that the *in vitro* systems are wholly representative of *in vivo* results overcoming limitations of whole body physiology and metabolism in respect to chemical exposure.

1.3 Data Acquisition

In the last decade the development of functional genomics approaches has been indispensable in the field of biomarker identification and provided knowledge towards mechanistic models of studied systems [53]. These high density and recently also high throughput technologies provide tools that can simultaneously measure thousands of features such as mRNA, proteins or metabolites. Computational methodologies to analyse this vast amount of data equally developed in concert with the technology. Advances within the technology of microarrays and mass spectrometers have increased sensitivity and number of samples per run while constantly decreasing the cost. In particular, in transcriptomics, progress made in the chemistry and technology has decreased sample-to-data time from 5 days to just about 1.5 days while introducing multiplexing (multiple samples per experiment) and higher feature density. Independently to these developments, computational toxicology has focused on quantitative structure-activity relationship (QSAR) models. These models identify links between physico-chemical features (PCFs) and biological activity, which can be a continuous (toxicity) or categorical (toxic / non-toxic) variable.

1.3.1 Transcriptome Expression Profiling

The evolution of gene expression Profiling

Since the discovery of the DNA structure by Watson and Crick [54] research in molecular biology has focused on decrypting the code of life. Understanding this code would give answers to various functions within a living organism and provide targets for alleviating diseases and illnesses. To reach this goal many different technologies have been developed to help the scientific community. The detection of RNA molecules is only one of those technologies that have been made available and within a cell based system can give indications on disease status, protein levels and provide valuable data points for prediction purposes. Prior to the development of microarray technology, gene expression profiling was initially performed using Northern blotting technology. Introduction of reverse transcription polymerase chain reaction (RT-PCR) pro-

vided a more sensitive approach and simultaneously reduced the need for dangerous radioactive reagents. RT-PCR was quickly followed by q-PCR with increased detection sensitivity making it the gold standard of gene expression profiling. However, PCR technologies are only able to detect one mRNA molecule at one time, which hinders biomarker identification on a global scale.

Simultaneous Expression Profiling

The development of cDNA microarrays was the first attempt to measure multiple known transcripts within a single experiment. They provided a customizable and cost-effective approach to transcriptomics. More specifically, probes (in many cases large parts of nucleotide sequence representing the gene of interest) were designed prior to spotting onto glass slides using robots with fine pins or needles. Commercialization of these products resulted in improvements, gaining increased sensitivity, density and added multiplexing support allowing for multiple samples to be measured within a single experiment. Today, running large-scale transcriptomics studies using these technologies is cost and time effective allowing for identification of biomarkers, mechanistic and predictive models to help understand how biological systems work.

Differences in microarray printing technology

In the commercial sector there are three major microarray providers. Agilent, Nimblegen and Affymetrix use different printing methods and designs to build microarray slides. Traditionally microarrays were spotted or printed using presynthesized oligonucleotides. This enabled high volumes of microarrays at relatively low costs per sample. Due to the mechanisms used in this printing technology spots rarely were aligned properly adding further time consuming effort to extract feature information. Affymetrix who has pioneered the *in silico* designed microarray platform, using their GeneChip products, created high density short oligonucleotide sequence (25nt) arrays with the use of masked photo-lithography. Each transcript on these arrays is represented by a number of sequences. In addition each sequence feature was comprised of a perfect and non-perfect match. This design not only allowed to measure mRNA degradation curves and provided limited splicing knowledge but also identified unspecific binding for each

sequence making it the gold standard for high throughput gene expression profiling. A few years later Nimblegen presented its own microarray design based on maskless digital light processing using micromirrors to synthesize oligonucleotides up to 70nt at specific positions on the slide. Most innovatively however is the ink-jet technology used by Agilent to print sequences on a small piece of glass. An offshoot of Hewlett-Packard (HP), Agilent borrowed the companies well known printing technology and adapted it to simply print nucleotide bases up to 120nt with micron precision. This enables Agilent not only to be one of the fastest technologies to create arrays but also gives the customer the ability to easily design custom slides at no extra cost. This especially is useful with non-model organisms, or species for which no commercially available microarrays exist. Furthermore both Agilent and Nimblegen have introduced multiplexed slides, on which multiple samples can be run in a single experiment without the loss of feature number. Due to the different printing technologies, feature density differs greatly between the three competitors. Affymetrix, for example, can place almost 6.5 million features on its microarrays. Scanning such dense slides however requires very specific scanning equipment and hence the Affymetrix technology is bound to a series of specific devices making it one of the most expensive microarray technologies. NimbleGen and Agilent on the other hand initially competed with traditionally printed cDNA arrays and were therefore designed to use standard microarray scanners. Only recently both companies have released scanners and accompanying formats that can contain up to 1 million features per slide (Table 1.1). Their competitive price, customization and sensitivity make these a much more cost-effective choice to Affymetrix. Out of the three companies, Agilent provides one of the most comprehensive choices including a number of various formats and low cost.

A typical Sample to Data Work ow

In essence traditional microarray technology does not differ greatly from todays improved versions. Figure 1.1 shows a schematic view of gene expression profiling using modern Agilent microarrays (other technologies may differ slightly). Initially total or mRNA is extracted from the sample using phenol-chloroform or column based methods (Figure 1.1 Step 1). A primer is used to bind the desired RNA molecules within the sample and a reverse transcriptase enzyme

creates cDNA molecules based on the original sample (Figure 1.1 Step 2). In most modern labelling reactions amplification methods are used to decrease the amount of total RNA needed by employing specific enzymes in the reaction. Following the creation of the cDNA molecules a T7 RNA polymerase in the presents of either cyanine-5 (Cy5) or cyanine-3 (Cy3) bound to cytosine transcribes the double stranded cDNA molecule to give labelled antisense cRNA samples (Figure 1.1 Step 3). Most commonly in single channel experiments, where signal intensitiy is representative of mRNA concentrations, Cy3 dye is used as it is much less susceptible to ozone than Cy5. Due to the differences in wavelength emissions in these two dyes double channel experiments can be used to represent the ratio of each gene given a treated and a reference sample. The labelled sample is then hybridized to the microarray.

1.3.2 Proteomics and Metabolomics

The advancements in other omics technologies, beside transcriptomics, also highly contributed to the genome-wide molecular characterization of biological systems. Mass spectrometers (MS) used in metabolomics, proteomics and lipidomics are constantly increasing in sensitivity. Metabolomics, for example, is the study of endogenous, low molecular weight metabolites [55, 56]. Measurements can be performed from cellular to whole organism levels and its

	Affymetrix	Agilent	Nimblegen
Applications	Gene Expression, CGH, tiling arrays, CHiP chip, SNP detection	Gene Expression, CGH+SNP, DNA Methylation, ChIP chip, miRNA, CNV	Gene Expression, Sequence Capture, CGH, ChIP-chip, DNA methylation, CGS
Available Formats	Depending on Application	8x15k; 4x44k; 2x105k; 1x244k; 8x60k; 4x180k; 2x400k; 1x1M	4x72k; 1x385k; 12x135k; 3x720k*; 1x2.1M*;
Maximum Feature Size	~6.5 million	1 million	2.1 million*
Cost per Sample	~£350	From £120	From ~£110

Table 1.1: **Comparison of Modern Microarray Technologies Available.** Applications across the different Companies do not differ greatly. Cost and Formats are highly different. *not available for gene expression arrays

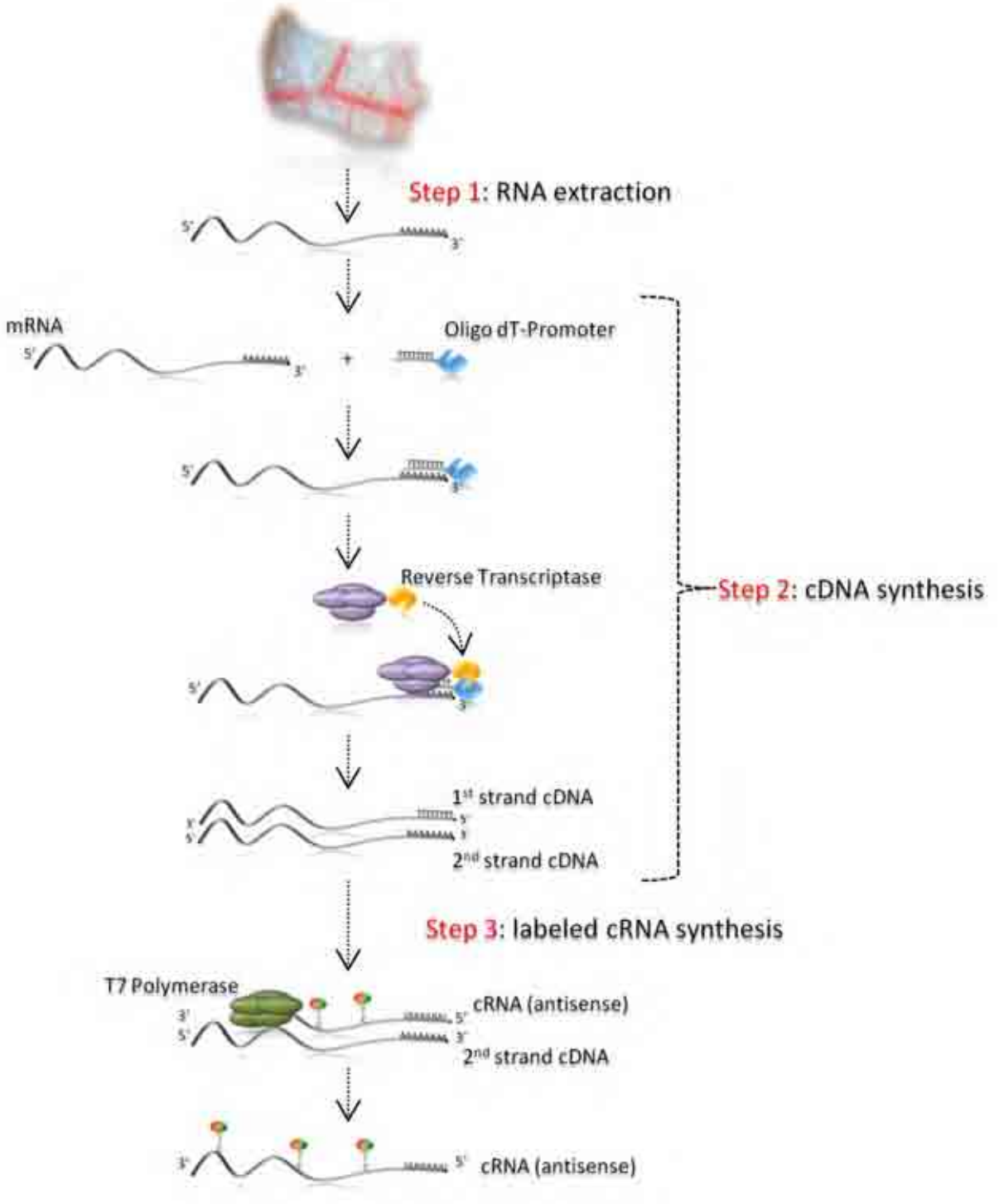


Figure 1.1: **Modern mRNA Labelling Reaction.** Modern Agilent microarrays use specific enzymes to amplify total RNA samples. Using such a techniques initial RNA concentrations can be as little as 25ng of total RNA to create enough labelled cRNA to use on the arrays.

ability to build a snapshot of compounds including lipids, sugars and amino acids can give indications about the individuals health or state of disease at a particular point in time [55, 57]. In respect to an external stressor, the metabolome often responds earlier than the transcript- or proteome [56]. Proteomics on the other hand focuses particularly on translated products of gene expression which can interact with mRNA to create further proteins or regulate their expression. Therefore the information gained by this technique could potentially be more informative of the molecular state than mRNA levels. Similarly to metabolomics proteins can be measured on various levels across the biological system. One major challenge in both of these technologies is the annotation of the various compounds that are identified. For the mass spectrometer to measure a given compound it has to be firstly ionized. A number of different adduct forms of different relative concentration can occur and in some cases a particular compound may not even ionize at all. In addition even though highly sensitive mass spectrometers are available, highly accurate mass to charge ratios alone are not sufficient information to differentiate between iso-forms, or molecules of similar weight.

1.3.3 Next Generation Sequencing

In addition to these omics technologies, advancements in next-generation sequencing in terms of speed and cost have had a great impact in the omics field. In transcriptomics, the RNA molecules of interest need to be known previous to designing the microarray. In NGS this knowledge is not needed. Here either DNA or RNA is measured by reading each base and constructing the genome (DNA-seq) or transcriptome (RNA-seq). Especially for non-model species this methodology is used to generate gene expression data or to define the transcriptome for a given species. In comparison to current transcriptomics technologies, however, RNA-seq cannot compete financially. Data generation is almost 10x more expensive per sample and requires a longer time frame until usable data is produced. The quality of the data in many cases is comparable between microarrays and sequencing [58]. A particular advantage of using NGS is the ability to characterize different splice-forms and identify even very lowly expressed genes.

1.3.4 Computing Physico-Chemical Features (PCFs)

The role of QSAR in predictive toxicology is highly important. It provides the researcher with a set of tools that enables the characterization of chemicals using a set of pre-defined descriptors. Since the early 1960s the QSAR techniques are in constant advancement and its importance in the virtual laboratory becomes more and more present. Especially in drug discovery, implemented as a high-throughput screening method, QSAR models have shown to help reduce the number of potential unwanted chemicals [59, 60]. Virtual filtering of chemicals in the pharmaceutical industry is now a standard technique to remove compounds that are predicted toxic or of poor pharmacokinetic properties but also to find compounds that are highly likely to be final candidates [61, 62]. In essence a QSAR analysis strategy is subdivided into 3 main steps:

1. Generation of Molecular Descriptors from chemical structure,
2. Filtering of descriptors to remove redundant entries,
3. Linking Descriptors to biological activity.

In most cases the extraction of descriptors from compound structure is facilitated through software on high performance computing clusters, some of which are freely available online. Standalone applications are also available but are usually catered for industry and hence not very cost-effective for academic purposes. On particular free online service is E-Dragon (hosted on `www.vcclab.org`). It is based on the DRAGON software which is able to calculate a set of molecular descriptors for a given chemical. In its newest instalment (DRAGON 6) calculates 4885 features subdivided into 29 groups. These capture a range of 2D and 3D QSAR descriptors which in general describe the topological, geometrical, electrostatic and atomic fragments of a given chemical (Table 1.2).

The large number of features however presents itself with a relatively large search space. Filtering of descriptors may be applied and this can be accomplished by a statistical, information theory, correlation based or simple criteria (i.e. number of unique numerical values across sample space) approach. Following the reduction of search parameters the remaining features are then used to link biological activity to a minimal number of descriptors. To identify feature-sub-groups variable selection algorithms, such as a Genetic Algorithm or Simulated Annealing,

	Descriptor Group	No. of Features
1	Constitutional descriptors	43
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	48
7	2D matrix-based descriptors	550
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P_VSA-like descriptors	45
11	ETA indices	23
12	Edge adjacency indices	324
13	Geometrical descriptors	38
14	3D matrix-based descriptors	90
15	3D autocorrelations	80
16	RDF descriptors	210
17	3D-MoRSE descriptors	224
18	WHIM descriptors	114
19	GETAWAY descriptors	273
20	Randic molecular profiles	41
21	Functional group counts	154
22	Atom-centred fragments	115
23	Atom-type E-state indices	170
24	CATS 2D	150
25	2D Atom Pairs	1596
26	3D Atom Pairs	36
27	Charge descriptors	15
28	Molecular properties	20
29	Drug-like indices	27

Table 1.2: **Descriptor Groups Calculated by DRAGON** The 29 descriptor groups that can be calcualted using the DRAGON software. It is available as a standalone application but it also freely available through a web-interface.

are used. Depending on whether the biological activity is a continuous or a categorical variable a regression or classification problem occurs. A regression problem involves modelling of the dependent variable against a function of descriptors whereas in classification problems the resulting model is defined by decision boundary where best separation between classes occurs [63]. In both regression and classification problems a number of linear and non-linear methodologies are available which can be easily implemented in the chosen variable selection

approach.

1.4 Omics Data Analysis

To make sense of the ever increasing number of features being measured using omics technologies analysis pipelines, computational methodologies and various other tools are needed. Initially however the challenge lies in extracting the information that has been gained by the method used. In transcriptomics, for example, this would mean analysing the image returned by the scanner and in metabolomics or proteomics processing of the peak list identified by the mass spectrometer. The second even bigger challenge is to minimize the technical variation and identify the true biological signal. Luckily during the large-scale genome initiatives in the medical field new methodologies and statistical tests have been developed and successfully applied to many areas in biology. For almost every challenge in the analysis pipeline a selection of tools exist that can give statistical or visual results that aid in the understanding of the biological system being studied. There are 6 main groups of tools available encompassing raw data preprocessing, clustering algorithms, differential expression approaches, functional analysis methods, class prediction and network inference techniques. In each category a selection of applications exist that have been developed for a particular challenge and hence creating analysis pipelines without human interaction is very difficult.

1.4.1 Raw Data Pre-processing and Normalization

In any omics technology one particular problem is the raw data processing. Devices and laboratory instruments used by the techniques create noisy data and normalization techniques are used to try and reduce this variability to reveal the true biological signal. Technologies which utilize fluorescence to represent sample features, image analysis procedures need to be applied to extract a numeric representation of that feature. Furthermore local background for each feature has to also be summarized to remove unwanted background effects such as washing artefacts or impurities. A transformation, usually log2, is then applied to the data to remove low signal intensity bias and to facilitate interpretation into fold-changes. In transcriptomics it is possible to run 2-colour microarrays in which each feature is exposed to two different samples using

different dyes and is then represented in the form of a ratio. The transformation is also applied here, however ratios are much noisier when the denominator is approaching zero. Following the transformation a normalization procedure is applied to the data. In the case of a 2-colour array, dye bias can influence the signal intensity at each feature and specialized normalization procedures are used to correct for these. A number of different normalization procedures are shown in Table 1.3. Some microarray designs contain more than one feature for a given target. This can be attributed to either the availability of multiple sequences or to the simple fact that empty spaces are available in the design. In both these cases a data summarization may be necessary. In Affymetrix, for example, each gene target contains a number of features tiled across the mRNA. In this case averaging across all given probes provides a good estimate of the expression of that gene [64, 65]. In other technologies it is important to identify whether the multiple probes are using the same or different oligonucleotide sequence. If an identical probe is present multiple times, averaging across samples, may be an option. In addition these multiple probes can be used to identify problematic areas of the array. On the other hand, if there are two probes with different sequences, one may choose to allow both or the probe with the higher signal for further analysis. The reasoning here is that features with a low expression are usually highly variable across different biological replicates. An additional filtering step, removing lowly expressed genes, may also be applied for the same reason. The resulting data file can now be subjected to a series of statistical methodologies which are outlined in the next section.

1.4.2 Identification of Differentially Expressed Genes

Often the most basic question relates to whether a set of features is differentially expressed between the classes used in the experimental design. Initially this was approached by calculating the fold change between treated and control samples. Application of an arbitrary threshold identified up and down regulated genes. This technique, however, was deemed unsuitable due to missing association to statistical significance. A number of formal statistical approaches were then developed and applied. Among the first techniques to identify differentially expressed

	Description	Availability
Scale	Most intuitive approach to simply scale the differences of medians across all samples to be 0	[66]
Loess	Local regression based methodology; usually applied to 2 colour arrays to adjust for dye bias; can adjust for print tip differences (in spotted cDNA arrays print-tip-loess)	[67]
Quantile	Assumes that the distribution of the signal between each sample is identical and corrects accordingly. This type of normalization is favoured with Affymetrix arrays.	[68]
Variable Stablizing Normalization (VSN)	Adjusts data to have equal variance for all intensities.	[69]
MAS5	A regression method used for Affymetrix data. Builds regression models for a subset of all data and adjusts each probe accordingly	[70]
rma	Each probe is background corrected, quantile normalized and then summarized.	[64]
gcrma	Improved version of rma to incorporate sequence specific probe affinities	[71]
MAANOVA	MicroArray ANalysis of VAriance, normalization is performed by fitting an ANOVA model for fixed and mixed effect models for each gene. It can be used to correct for array, dye, sample or even batch effects.	[72]

Table 1.3: **Selection of Available Microarray Normalization Procedures.**

genes was the application of the traditional t-test. It compared the averages of up to two classes and specified a p-value for each genes. A non-parametric version, the Wilcox-test, is also available. There is however a downside with the t-test as it provides very high scores for features with very low variance. This essentially biases the results towards highly reproducible genes. To compare more than two classes a generalized extension of the t-test, Analysis of Variance (ANOVA), has been widely used to identify differentially expressed genes. This particular technique compares all means across all sample groups reducing the need for performing 2 class comparisons for all possible combinations. Furthermore, ANOVA has been extended to analyse multiple factors with identical classification groups (n-way ANOVA). There are a number of other methodologies which aim to address issues such as the previously mentioned low variance in the t-test. SAM, B-statistic and samroc are among the methodologies which provide alternatives to the t-test. A comprehensive review of these methods has been published by Kim

et al [73]. With all statistical tests, a threshold is usually chosen by the user to define features that are significantly differentially expressed. This is called the probability, which is of type-1 error. This error controls the number of false positives within the statistical test. However it only assumes that one hypothesis has been tested. When many more hypothesis tests are performed the problem of multiple comparisons arises. The idea is to be able to control all the tests within a single experiment. For this reason several correction methods have been proposed [74–78]. The most commonly used correction method has been published by Benjamini and Hochberg [75] which they named the false discovery rate (FDR). This particular method was designed to capture the highest number of true positives while controlling the number of false positives. An FDR of 2%, for example, on average should yield 98% of true and 2% of false positives. Modifications of this approach have also been proposed by Storey [76] which included a bootstrap estimator for the family-wide type-1 error. SAM [79] as the only differential expression approach incorporates this Storey [76] correction and has gained a strong standing in the community. In addition SAM has been optimized for many different experimental designs, be it paired, unpaired or time-course problems.

1.4.3 Exploratory Data Analysis

Probably the most used analysis tool developed for large scale data analysis are tree-based visualization and clustering algorithms. These provide a visual representation of the similarity between genes or samples. Initially a dissimilarity matrix is created using a simple distance measure such as Euclidean distance or Pearson correlation. Next clustering algorithms attempt to identify variables which share a common trait. In some cases, these techniques are also combined with an image representative of the features signal intensity illustrated using a colour gradient heat map. In transcriptomics, the standard has been to use average linkage clustering with the Pearson correlation as a distance measure. Other clustering algorithms also include self-organising maps, k-means, multi-dimensional scaling and dimensionality reduction techniques such as principle or independent component analysis (PCA and ICA). In particular PCA has been widely used as it tries to reduce the overall dimensionality of the data by summarizing

the variance across the different samples into principle components where the first component contains most of the variance and the last the least variance from the original dataset. The first 2 or 3 components are then be represented on a 2D or even 3D plot to visually represent samples. The distance between samples can then be indicative of the difference in signal intensity across a number of features. Specific methodologies for toxicological studies have also been developed. These arose through the fact that many experiments in this field contain multiple levels of classification integrating time and/or dose-responses within several different chemicals [80]. One such approach has been published by Chou et al [81]. Their method, called EPIG (Extracting Patterns and Identifying co-expressed Genes), tries to find all the patterns within the data and categorizes them on the basis of the signal to noise ratio, signal intensity and correlation of expression profiles [81]. The integration of these allows for a much more thorough pattern discovery across the dataset. Further methodologies include semisupervised clustering approaches that integrate phenotypic data [82] and biclustering methods which partition gene expression data into cliques (subsets of samples sharing a similar expression pattern) [83, 84].

1.4.4 Machine Learning Methods for Supervised Classification

Being able to predict the outcome, class or effects as a result of an external stressor has become a major bioinformatics objective. Prediction algorithms exist in plenty different forms due to many other areas also highly interested in predicting the future. They range from very simple algorithms such as nearest centroid or k-nearest-neighbours to much more mathematically complex decision trees approaches or support vector machines. The ultimate goal in this field is to minimize the number of features needed for high prediction accuracy. To achieve this variable selection methods are used in combination with a classification algorithm to test many different sets of features and identify one possible solution representative of this goal. Schematically, the method should split the input data into a training and test-set by a predefined ratio (usually 2/3 training 1/3 test). The training set is then used to train the particular classification algorithm and is then tested against the test set. This particular setup, however, is prone to over-fitting. Over-fit models are only able to describe the training but not the test data and are therefore not

useable. To overcome this problem, error estimation procedures such as leave-one-out-cross-validation, k-folds or splits are used. As each classification problem most likely contains more than one potential good fitting model these approaches train thousands of models. A representative model is then built using the most frequently chosen features across all models. This is done by incrementally testing the top fifty features and choosing the best predictive model. Such a method can be found in the GALGO package in the statistical environment R. It uses a genetic algorithm for the variable selection approach and combines it with several in-build classification algorithms such as nearest centroid, shrunken centroid, k-nearest neighbour, linear discriminant analysis, support vector machines and even supports user built functions.

1.4.5 Functional Analysis

Through the last few years the focus on gene-level analyses has shifted towards a higher level approach. It is known that genes rarely work by themselves but work together with other genes, proteins or metabolites to complete their task. Clustering or differential expression analysis alone can therefore provide little information on the full extend of molecular change. Functional annotation tools are methods which can help scientists to identify potentially enriched clusters within the studied biological system. Several web-services are available which perform such tasks. DAVID [85] and FatiGO [86] are probably the two best known functional clustering approaches. DAVID in particular has the ability to integrate many different functional annotation databases, such as GO, KEGG, BIOCARTA, Panther and other annotation, and identifies enriched functional categories within the user submitted features. This, however, are not the only uses that functional analysis can offer. As databases such as KEGG or BIOCARTA utilize information, available from the literature, to build pathway maps, these can be considered as functional modules. In the context of KEGG specifically, pathways within their database can be considered to be conserved across a number of species. This in particular can help in biological interpretation with non-model organisms. A functional module can be defined as a collection of genes either regulated by a common factor, following a similar trend in expression or have functional similarities. To define a functional module therefore we can use these databases, such as

KEGG, network inference algorithms combined with modularization techniques or use simple clustering methods. The identified modules can then be subjected to statistical tests, prediction algorithms or linked in regression problems to other phenotypic data.

1.4.6 Network Inference

High-throughput technologies have generated vast amount of quantitative data representing the molecular state of cell-lines and tissues across many different species. In many cases these are mRNA expression level dynamics captured as a result of a perturbation to the system. It is known that proteins, which are themselves products of mRNA, have the ability to control mRNA expression levels of either themselves or other genes. It could be therefore conceivable that a statistical relationship between two potentially interacting genes can be formed albeit not directly proportional. This concept has lead to several reverse-engineering approaches and is still a matter of intense research. Some of these are designed to exploit time-course datasets while others work by analysing large compendiums of perturbation experiments. Among the first reverse-engineering effort was based on a Bayesian network approach which inferred probabilistic relationships between variables and allowed time-course, steady state and prior knowledge to be included. Correlation based and information theory approaches followed which calculate a coefficient between all the features to build the underlying regulatory network and were at first purely developed for steady state data. Some extensions of the information-based approaches are now also able to use time-course data. State space models and ODE based approaches are used to develop dynamic network models using time-course data but are only feasible with small number of genes.

1.4.7 Building Networks: Reverse Engineering and Network Inference

Fundamental to reverse engineering the underlying molecular network is the methodology used. A number of approach have been proposed over the last several years. Although transcriptomics based data has been predominantly used, these applications have been applied to other high dimensional data such as protein-protein binding strength, protein abundance, signalling (protein activation), and metabolic data [87, 88]. The development of mutual information (MI) based

analyses added an important reverse engineering technique to the already existing field. Inference algorithms such linear regression models [89–92], Bayesian networks [93, 94] or state space models [95, 96] have been previously used. Data mining or association rule mining methods have also been proposed [97], but these rely heavily on published material which is highly limited with non-model organisms. The general advantage of MI lies in the completeness of its reverse engineering attempt. Many of the previously mentioned techniques can only deal with small numbers of features. Within the MI techniques a number of methods have been proposed (for example ARACNE [98], CLR [99] or MRNET [100]), but these generally differ by the indirect edge removal approach. Generally such algorithms measure the dependency between genes, proteins, metabolites, and all relevant physiological data using a scoring function. The entropy-based MI captures a broad range of biologically relevant dependencies (positive, negative, and linear as well as nonlinear relationships) and is therefore capable of detecting more general dependencies than measures of linear correlation (i.e. Pearson or Spearman correlation). Similar to the linear correlation methods, higher values correlate with greater MI that is shared between the two variables. Due to the complicated underlying biological processes involved, causality cannot be directly inferred given a high MI score, without further validation. However examination of the resulting relationships can provide the necessary information to postulate adverse outcome pathways (AOPs). There is however one main issue with reverse engineering approach. In most cases they require a large number of data points (> 50) per node [98, 99]. However, there are instances where a smaller number of data points can lead to informative results [101]. The computational requirement for building a network using this methodology is relatively small as the estimation of the MI score is only based on two features at a time. Other methodologies such as Bayesian Networks rescore the entire network after each edge manipulation. Therefore MI based methods can deal with tens of thousands genes whereas Bayesian Networks for examples are limited to a few hundred or even less genes.

1.4.8 Identification of Adverse Outcome Pathways (AOPs) using Reverse Engineering

One application of network inference approaches is the characterization of mechanisms of action for a given environmental stressor or the underlying pathway structure of a given gene product. The process for developing and validating adverse outcome pathways can be structured into 3 distinct stages (Figure 1.2). Initially the network needs to be assembled using an appropriate methodology (Figure 1.2A). There is no limitation to what types of datasets can be included, although a pre-requisite is that the data is standardized (mean = 0, sd = 1) before attempting reconstruction. The second step is to visualize and interrogate the network (Figure 1.2B). One important aspect of network interrogation is the identification of functional modules. Modularization techniques identify highly connected sub-networks and a number of these have been proposed by other groups as well as our group during my PhD. Including phenotypic measurements can be of great advantage when trying to identify interesting modules as these should effectively cluster in the same region of the network. Additional functional information can provide information on biological processes of these networks aiding in AOP identification and biological interpretation. Lastly identifying the potential AOP and evaluating it using computational and experimental techniques is imperative before proclaiming to have a found a novel mechanism (Figure 1.2C). Computational techniques may include variable selection approaches to build predictive models of phenotypic outcome such as toxicity (classification for discrete or regression for continuous data). The resulting features could then be used to develop an assay which would predict the desired phenotypic outcome or early response to a stressor. Experimental validation of the assay using a number of independent samples and laboratories may then provide enough evidence to postulate a novel adverse outcome pathway.

1.5 Quantitative Structure-Activity Relationship Analysis (QSAR)

In the early 1960s Hansch and Fujita developed the $\rho-\sigma-\eta$ Analysis, a first attempt of linking chemical structure to a biological activity. Since then increase in level of sophistication,

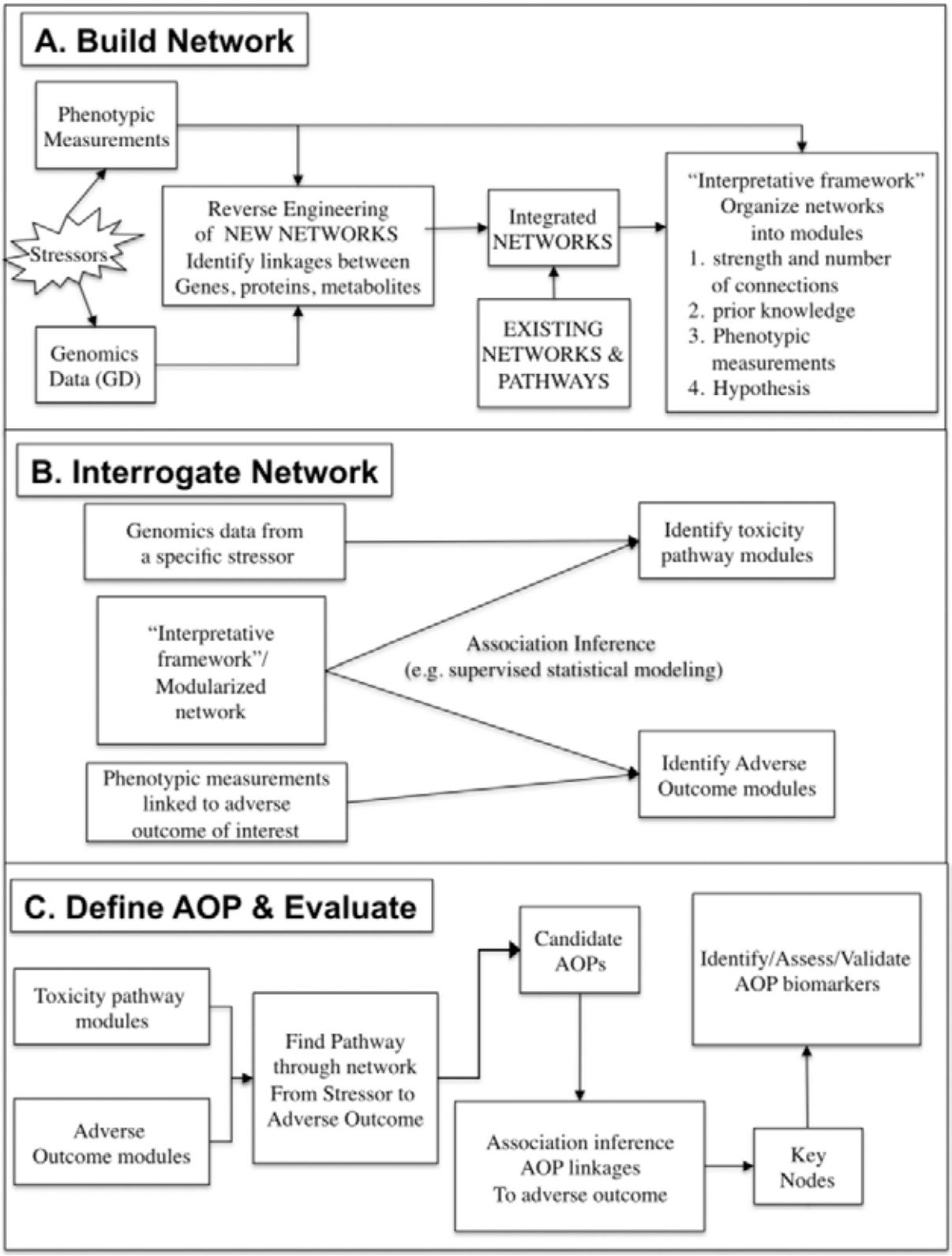


Figure 1.2: **Overall Strategy for Reverse Engineering of an Adverse Outcome Pathway (AOP).** [13]

number of chemical features and methods for identifying and correlating descriptors to biological activity have been published. A more modern term for this type of analysis is Quantitative Structure-Property Relationships (QSPR) but it is also known by Quantitative Structure-Activity Relationships (QSAR), Quantitative Structure-Toxicity Relationships (QSTR), Quantitative Proteome-Property Relationships QPPR), Quantitative Sequence-Action Model (QSAM) or Quantitative Structure-Reactivity Relationships (QSRR). The name depends on the situation it is being used for but in all cases the building of models linking biological activity to a function of chemical descriptors is performed. With the introduction of high throughput screening methods in the pharmaceutical industry, QSAR has become a vital virtual filtering method reducing overall time and cost in the development of drug-like compounds [61, 62]. Its ability to quickly and rapidly identify chemicals with potential risk factors, and narrowing of the lead-like compound library, reduces financial cost and animal suffering. Screening of hundreds of thousands of potential drug candidates is therefore greatly simplified. Focused compound selections can then be analysed using lower throughput assays [102]. These advantages however do not only apply to the pharmaceutical industry. Toxicologists, environmental protection agencies and the pharmaceutical industry use QSAR to identify potential compounds which may be hazardous to human health or the ecosystem. In many cases no toxicity information exists for new chemicals that have been manufactured or imported. In this context QSAR is used to identify and classify potentially dangerous chemicals and to help prioritize chemicals for further toxicological analyses. In the early 90s Auer et al [103] have already discussed the importance of these techniques and their potential ability to predict potential mode of action and more recently Matthews et al [104] showed that indeed QSAR models can be used to not only predict mechanisms of action but also to link chemical structure to drug-induced adverse events in humans. Furthermore the vast amount of available datasets in the public domain presents a potential supply of valuable data for building QSAR models.

1.5.1 Molecular Descriptors

As it is not possible to directly map chemical structure to biological activity, due to differences between chemistry and computers science, a set of predefined molecular descriptors has been established. A second reason for the necessity of descriptors is the fact that chemical compounds are highly diverse and most statistical data analyses require data which is uniform in feature space within each sample. There are two main broad families of descriptors currently being used in QSAR research.

2D Descriptors

This family of descriptors ranges from simple constitutional, geometrical and topological properties to electrostatic and quantum-chemical descriptors. More specifically constitutional descriptors summarize the elemental structure of a chemical such as molecular weight or counting the number of various bond strengths. These descriptors, mostly representing biochemical structures, have been used to build rule-based expert systems such as DEREK [105,106]. Identification of certain functional groups such as aromatic amines, acryl hydrazine or alkyl carbamate can highlight potentially hazardous chemicals [106]. Topological descriptors, on the other hand, are slightly more complex. Initially the chemical structure is treated in a graph format where atoms are represented by nodes and covalent bonds by edges. Calculation of indices describing the topology of this network such as the shortest path between all pairs of non-hydrogen atoms or the sum of geometric averages of edge degrees of atoms within a given length are the resulting features [107–110]. Furthermore information about valence electrons can also be included [111, 112]. Electrostatic descriptors can also provide informative features. These represent the atomic net and partial charges of a compound [113]. Molecular polarizability or areas of high negative or positive charge may also be represented using these descriptors [114, 115]. Combinations of topological descriptors and other properties such as electronic organisation or polarizability also exist. These usually represent the distribution of charge across the molecule within the eigenvalues of an atom connectivity matrix [116–118]. Geometrical descriptors, as the name suggests, relate to the spatial arrangement of a molecule. Volumetric features such as

the van der Waals areas [119, 120], gravitational indices [121] or principal moments [122] are represented.

3D Descriptors

This family of descriptors is much more complex than its conventional 2D counterpart. It is much more computationally complex to extract numerical information and in many cases is done in a stepwise fashion. More importantly is the identification of the 3D structure from experimental data or molecular mechanics. As in most cases this information does not exist, conversion tools such as CORINA (Molecular Networks GMBH, `http://www.molecular-networks.com/`) or OMEGA (Schrödinger `www.schrodinger.com`) are used to define the 3D structure from the 2D representation. Within this family there are two subfamilies which describe alignment-dependent and independent descriptors. The Comparative Molecular Field Analysis (CoMFA [123]) and the Comparative Molecular Similarity Indices (CoMSIA [124]) both belong to the alignment-dependent category and summarize the potentials of the energy fields using Coulomb and Lennard-Jones or Gaussian-type functions respectively. Within the independent subfamily the Comparative Molecular Moment Analysis (CoMMA [125]), Weighted Holistic Invariant Molecular Descriptors (WHIM [126, 127]), VolSurf [128, 129] and Grid-Independent Descriptors (GRIND [130]) are represented. These have the advantage of being invariant to molecule rotation and translation in space [63]. CoMMA features describe the mass distribution, including intertia and magnitudes of dipole moments, using the center of mass and dipole within the structure. WHIM descriptors use PCA on the atomic coordinates to summarize the variance in the structural space. Within that space several statistics are used to define directional and non-directional descriptors. In addition the contribution of each atom to the PC can be weighted by a chemical property such as mass, van der Waals volume or electrostatic properties or indices. To define hydrophobic or hydrophilic regions within a chemical the VolSurf approach is used. It virtually interacts with the molecules surface using several probes and the resulting information is then used to define the descriptors. The GRIND descriptors were specifically developed to maximize the biological properties of the compound while being alignment-independent. An important feature of these descriptors is the fact that due to the

autocorrelation transform used, the original descriptors can be regenerated and then visually represented [63].

1.5.2 Filtering of Molecular Descriptors

As described in the section above the number of methodologies and hence the number of features is very large. Automatic approaches to reducing the feature space, removing any non-contributing or highly similar descriptors, have been proposed. These can be subdivided into 4 sections, Simple Threshold Criteria (STC), Correlation-Based Methods (CBM), Information Theory Approaches (ITA) and Statistical Selection Criteria (SSC). The STC approach is the most simplest and naive method of reducing the feature space. A simple criterion, such as the number of unique categorical values across each sample can maximize the descriptor space but removing only those features which do not add any additional information as they are constant in each molecule. This type of filtering works best when used in conjunction with feature selection algorithms. CBMs dive further into this area by identifying descriptor pairs which share information. If a pair exceeds a user set coefficient threshold one feature is then randomly removed. This however can lead to differences in datasets and results. To combat this effect, prior to testing each pair, the descriptors are ranked using PCA for example [131], and the feature with the highest rank is retained. ITAs are very similar to the correlation based methods as they quantify the information content between pairs of descriptors. The same method for removing features can be used using this metric, but in addition network visualization and modularization techniques can be used to identify groups of features which share large amounts of information. Choosing a representative feature for that group can then be based on a centrally located (hub-feature) or most connected descriptor. SSC methods, such as the Fisher ratio [132], are usually used in conjunction with CBMs to rank the descriptors prior to estimating the correlation. In the case where the sample space can be separated into two classes, a simple Kolmogorov-Smirnov test [133] test can be used to compare each descriptors cumulative distribution and assign p-values based on this difference.

1.5.3 Linking Biological Activity to Chemical Structure

This is the crucial step in QSAR methodology as pioneered by Hansch et al. The desired final model should minimize the number of descriptors while at the same time be highly predictive of the observed variable. These variables can be either continuous or categorical and linear and non-linear approaches have been developed to tackle these.

Linear Models

The model build by Hansch et al in the early 1960s was based on a simple linear model to predict biological activity from different structural parameters. This type of model is still the most used today but the original methodology has been replaced with multiple linear regression (MLR) which can estimate model parameters using multiple descriptors. To identify the right parameters this approach minimizes squares of the errors between the observed and predicted variable. It has been successfully applied to various modelling problems such as predicting Caco-2 permeability [134] or predicting toxicity of nitrobenzene derivates to *T. pyriformis* [135]. Partial least squares (PLS) [136, 137] is another linear regression algorithm. Unlike multiple linear regression, which is restricted with large descriptors-to-compound ratios as over-fitting is more prominent with large number of features, PLS has been developed to overcome these short comings. The first step is to extract the latent variables from sampled factors (T) and responses (U). To build the model the extracted responses U are predicted by T and the resulting values are then used to predict the original responses. Due to its advantages to MLR, PLS has been very popular in QSAR models including a multitude of toxicity and cancer related problems [138–140]. The last modelling approach in this group is Linear Discriminant Analysis (LDA [141]). This classification algorithm was designed to use a linear transformation of the original input matrix to maximize the class difference and minimize within-class variance. Similar to the MLR method, this approach can easily lead to over-fitting when a large feature to sample ratio is present. In these cases a PCA can be utilized to reduce the number of total features summarizing the variance across feature space. Examples of the use of LDA in the public domain include compound prediction [142], antibacterial activity [143, 144] and pesticide prediction [145], and

in the latter has been shown to perform better than other algorithms especially with Friedmans extension that deals with large feature to sample ratios.

Non Linear Models

Understanding and interpreting non-linear models can in some cases be much more difficult than linear cases. However in biology linearity is rarely represented and hence non-linear methods usually become more accurate, in particular with large diverse datasets [63]. Classification algorithms such as the Bayes Classifier [63] or K-Nearest Neighbour (KNN) [146] have been successfully used but have been shown to perform worse than Artificial Neural Networks (ANN) [147] or Support Vector Machines (SVM) [148, 149]. ANN in particular has been developed in the context of a biological system, a neural network as the name suggests. A number of specific methodologies have been proposed in the literature but perceptron-based and radial-basis approaches were favoured. In these feed-forward methodologies, information flows from the input descriptors, through a set of layers, to the output of the network essentially predicting the biological activity. Due to their unique design these techniques have the ability to perform much better than other approaches [143, 150, 151]. SVM on the other hand uses a multiple dimension decision hyperplane to separate samples and can be extended from its linear origin to a non-linear classifier with the use of kernel functions [149]. An adaptation of the SVM core to support regression problems has also been published [152]. The main advantage of using SVM is its high threshold to over-fitting and its very small error rates. In several comparisons, such as drug-likeness prediction [153] or COX-2 inhibition [154], SVM has been shown to outperform other algorithms. Lastly Decision Trees (DT) [155, 156] are very different than all of the previously outlined approaches as they employ a logic-based systems. In such an approach a decision is taken by using a simple test criterion at each node of the tree. Once the bottom of the tree is reached the leaf provides the value that is representative of the prediction. Algorithms of this type exist in both regression and classification problems. Examples of application of decision trees to real world datasets include identification of individual amino acids in the Reverse Transcriptase pocket of HIV-1 [157] and the prediction of human hepatoxicity endpoints [158].

Ensemble Techniques

Traditionally the methods described earlier are used to build a single predictive model but more recently ensemble methods which combine several predictive models have shown to increase predictive power [159–162]. One such method is called Bagging [160], where a bootstrapping algorithm creates multiple base models, using one of the previously described methods, and uses the average of all of these models to accurately predict the outcome. Another example of ensemble techniques includes the Random Subspace Method (RSM) [161], which uses random feature subsets, and boosting. The RSM has been implemented in the Random Forest approach which is based on the decision tree algorithm [163]. In a recent comparison this has shown to achieve even higher accuracy than SVM, PLS or KNN [164, 165]. The boosting algorithm [162,166] on the other hand is specifically designed to build models with hard to predict samples by using a weighting system. Initially each sample has an identical weight but at each iteration of the classification the weights are adjusted depending on the error from the previous result regardless of the error in the next step. This allows the algorithm to create decision boundaries maximizing the difference between the classes, similar to the SVM method [167].

Variable Selection Methods in QSAR

One topic that has not been mentioned so far is the use of variable selection algorithms in QSAR. These methodologies combine a method for linking activity to chemical structure with an algorithm that efficiently searches across the feature space to identify a subset of descriptors that are highly predictive. Specifically in QSAR Genetic Algorithm (GA) [168] or Simulated Annealing (SA) [150] based approaches are used. GAs use an evolutionary process to model a population of solutions. The best of the, so called, chromosomes within the population are retained by the procedure. Multiple runs ascertain that the majority of the search space is covered. The resulting list of solutions can then be used to identify one particular model or with a specific strategy, such as a forward selection strategy, a representative model can be formed. In comparison SA approaches are rather simple as they only alter the current-best model by randomly exchanging features. With each iteration a number of changes is performed, tested,

and evaluated whether to keep the change or to discard it. The major problem with this type of approach is the fact that it can easily get stuck in a local minima. Methods for escaping this minima have been considered but in general a GA based approach can lead to highly predictive models much more efficiently.

1.6 State of the Art Predictive Toxicology

The development of toxicogenomics provided the community with a much needed high-density and high-throughput technology which allows the measuring of thousands of features for a single sample. Statistical methods exist which can deal with such large amounts of data and help in the interpretation of the results. More recently the concept of pathway-level analysis has brought forward the adverse outcome pathway approach where a change in normal pathway behaviour is identified [13,169]. And although annotation in various species, especially ones used in ecotoxicology, may be limited highly informative results leading to novel biomarker targets may be identified [41, 44, 170–172]. This particularly applies to non-model species, where a combination of homolog and local annotation can be used for pathway-level analyses. Computational toxicology also had similar advancements with QSAR predictions becoming more and more reliable [2, 3, 173]. To further develop predictive toxicology we have to combine all these advancements made in the last few years and provide the scientific community with an analysis approach that can improve the predictive power and provide potential mechanistic insights. The method developed in this thesis tackles this challenge by integrating classical QSAR methodologies with gene expression profiling datasets and applies it to two distinct species *Rattus norvegius* and *Daphnia magna*. One specific problem was the high dimensionality arising from both QSAR and microarray technologies. To reduce the search space a principal component analysis was used to summarize the gene expression data into pathways based on the KEGG database. Other dimensionality reduction techniques such as non-linear PCA (k-PCA) or independent component analysis can also be used but their effects have not been tested in this thesis. In respect to the pathway definition alternative online databases such as BIOCARTA, GO, or Panther or a more dynamic approach based on reverse engineering (ARACNE) coupled

with a functional modularization technique (FUMO, etc) may be more beneficial depending on the species employed. Utilizing the traditional QSAR techniques, biological activity (i.e. toxicity) is substituted with a pathway activity index. Due to the pathway-level design biological interpretation is highly simplified. Visualization of results indentified in both traditional and pathway-level linked models can help identify specific global functions that may be impaired as a result of exposure. Using the same methodology, given the right set and sample size of chemicals, specific mechanisms of actions for a given chemical class may also be interrogated. Identifying potential entry points for chemical toxicity either globally or for chemicals with known mechanisms of action could be highly beneficial in toxicity prediction of unknown compounds.

1.7 Concluding Remarks

There are several important issues that need to be addressed before the use of QSAR in toxicology can move to the next level. The advancements in omics technologies in the last decade have undoubtedly enriched the field of toxicology with genomics tools that are time-efficient and cost-effective. As a result large datasets have become increasingly available in the public domain that specifically tackle questions related to chemical exposure. This fast growth however has caused a few side-effects to become apparent. In particular gene annotation is proving to be extremely difficult as many non-model organisms have little or no annotation. Using the blast toolset or the KEGG Annotation Builder [174] gene homologs can be found relatively quickly but in many cases only small numbers of genes can be significantly associated. This makes biological interpretation considerably more difficult. To understand toxicity it is imperative that one deduces the adverse pathways resulting from exposure in the biological system being studied, which has shown to be highly informative [12, 13]. In the case of non-model species however the results from these reverse engineering methods are hard to interpret. Nevertheless, it gives researchers features to focus on in subsequent studies. Computational toxicology is plagued by similar annotation problems; however, the focus here has been changing to the virtual laboratory [175]. The U.S. EPA, for example, is building a virtual liver model whose

overall goal will be to accurately predict toxicity mechanisms of various compounds. Other such models are also developed by the National Biomedical Computation Resource (NBCR; `http://nbcr.sdsc.edu/`) which is building a human heart model that describes molecular interactions, diffusion, and electrostatics. Despite all these advancements in technology and resources, the field is still far from being able to accurately model whole organisms and effects of compounds on various numbers of species in the environment. On the other hand we have methodologies and technologies at hand which can greatly improve the understanding of toxicity. Integration of these techniques will prove to be indispensable to the community to undertake further research and help build on the current knowledge. The ultimate future probably lies in the development of *in vitro* systems which can replace *in vivo* experiments [16], methodologies which yield maximum biological information from minimal experimental research and technologies which are even more sensitive, cost and time-effective than the current available generation. As a stepping stone, however, the use of *in vitro* techniques to prioritize chemicals, such as in the ToxCast Program, will be indispensable in developing predictive toxicology further. *in vitro* techniques may also provide a cheap alternative to toxicity testing, essentially reducing severity of animal testing.

Chapter 2

Mapping Drug Physico-Chemical Features to Pathway Activity reveals Molecular Networks linked to Toxicity Outcome

2.1 Abstract

The identification of predictive biomarkers is at the core of modern toxicology. So far a number of approaches have been proposed. These rely on statistical inference of toxicity response from either compound features (i.e. QSAR), *in vitro* cell based assays or molecular profiling of target tissues (i.e. expression profiling). Although these approaches have already shown the potential of predictive toxicology we still do not have a systematic approach to model the interaction between chemical features, molecular networks and toxicity outcome. Here we describe a computational strategy designed to address this important need. Its application to a model of renal tubular degeneration has revealed a link between physico-chemical features and signalling components controlling cell communication pathways, which in turn are differentially modulated in response to toxic chemicals. Overall, our findings are consistent with the existence of a general toxicity mechanism operating in synergy with more specific single-target based mode of actions (MOAs) and provide a general framework for the development of an integrative approach to predictive toxicology.

2.2 Introduction

One of the most challenging tasks in toxicology is the identification of a potential toxicity via high-throughput screening, avoiding the use of animals, at an early stage in the development programme of a product such as a pharmaceutical or in the context of REACH [176]. Such screens can help to reduce attrition of products late in development and can help to prioritise existing chemicals for more complete safety assessment. In this context, the concept of quantitative structure activity relationship (QSAR) analysis was originally developed with the purpose of predicting a toxicity or pharmacological response utilizing information on the physico-chemical features (PCFs) of a chemical and the relationship to biological effects. In the last 20 years QSAR analysis has been characterized by an increasing level of sophistication as technological and computational developments have made it possible to measure or compute a higher number of chemical and physical parameters [173]. In addition, recent reports have shown that the prediction accuracy of QSAR models can be increased when additional information from cell based assays is utilized [2, 3]. Independently to these developments, the availability of functional genomics technologies facilitated the measurement of mRNA concentrations, proteins and metabolites in single experiments. This, together with the development of novel computational methods suitable for the analysis and integration of very large multilevel datasets [177], have contributed to demonstrate the usefulness of molecular fingerprinting in predicting toxicity from an early readout of the response to chemical exposure [6, 177–180]. Toxicants can in some cases be discriminated according to their mechanism of action and their target organs [4, 181]. However, there have been no successful attempts to model the interaction between a drug PCFs with genome wide molecular response to exposure and put this in context with toxicity response. It is therefore still unclear whether a true integration between traditional QSAR and functional genomics data may be possible. In this chapter we describe an analysis strategy which addresses this issue by integrating gene expression profiling measurements in the logical framework of QSAR analysis. We have applied this approach to a publicly available expression profiling dataset, representing the pre-phenotypic transcriptional response to

chemical exposure in a rat model of renal tubular degeneration which is a major toxicological response contributing to attrition during drug development [182]. Our approach has successfully linked a sub-set of PCFs to the activity of signalling pathways known from the literature to drive effector pathways differentially modulated between toxic and non-toxic chemicals. This finding suggests the existence of general toxicity mechanisms which operate in synergy with specific single-target based MOAs. The approach we have used has general validity since it can be applied to integrate different types of PCFs, molecular and phenotypic measurements to identify predictors of toxicity within a mechanistic framework for biological interpretation.

2.3 Results

2.3.1 Rational of the Approach and Data Analysis Overview

The dataset we have used in this analysis is based on a wide range of chemicals. Some of them are known to work by different mechanisms of action and have diverse chemical structures. Despite this heterogeneity it has been shown that it is possible to identify early molecular response signatures predictive of *in vivo* toxicity outcome [28]. So far, it is unclear whether these signatures represent an early convergence of the different drugs MOAs towards common toxicity pathways or whether a component of them may represent a direct interaction between the chemicals and cellular components. Here we address this question by using a multi-step computational approach. Firstly, we simplify the complexity of the transcriptional response by computing indices of overall pathway transcriptional activity (Figure 2.1, Step 1). This effectively reduces the dataset from thousands of individual gene expression profiles to 148 pathway indices. We demonstrate that toxic and non-toxic chemicals can be separated on the basis of their ability to modify pathway activity (Figure 2.1, Step 2 and 3). This proves the biological relevance of the pathway indices. We then hypothesize that the defined set of a drug PCFs may be representative of the ability of a chemical to induce changes in the homeostatic state of the target organ. In line with this hypothesis we search for statistical models based on combinations of PCFs and predictive of the transcriptional response to drug exposure (Figure 2.2, Step 1). In parallel we identified which pathways are differentially modulated between samples

treated with toxic and non-toxic drugs (Figure 2.2, Step 2). If these two pathway subsets truly represent the interaction between chemicals and underlying molecular networks we may expect that they would be part of a super-pathway. This hypothesis was addressed by mapping pathways in the two on the KEGG pathway map and testing for statistical association (Figure 2.2, Step 3).

2.3.2 Computing Indices of Molecular Pathway Activity.

The overall aim of this study was to link PCFs to drug-induced molecular responses and phenotypic outcome. A key challenge in identifying subsets of PCFs predictive of transcriptional response is the astronomical number of possible combinations of PCFs and gene subsets that need to be tested within a statistical modelling framework. In order to address this challenge we first simplified the complexity of the dataset by reducing thousands of individual gene expression profiles to a relatively small number of overall pathway activity indices. This was achieved by summarizing gene expression profiles representative of a given KEGG pathway with the first two principal components (PCs) of the gene expression matrix [183]. The choice of the number of PCs to construct the pathway activity indices was driven by the simple criteria to represent at least 80% of the variance present in the original dataset. By using this strategy we built a new dataset representing 148 KEGG Pathways (44% of the KEGG pathway database). This dataset represents 1676 out of the 7478 genes which were originally present in the processed Iconix dataset. We found that the apparent loss of gene representation was largely (77%) associated to the high frequency of non-annotated genes (i.e. function unknown or estimated by sequence homology). KEGG Pathways represented in the derived dataset are a good representation of the spectrum of functions covered by the KEGG database (See Table 2.1 and 2.2 for a detailed breakdown in the functional representation of the KEGG pathways represented in the dataset).

	KEGG Pathway	Number of Genes
rno00190	Oxidative phosphorylation	34
rno04810	Regulation of actin cytoskeleton	93

Continued on Next Page...

Table 2.1 – Continued

	KEGG Pathway	Number of Genes
rno04330	Notch signaling pathway	20
rno04080	Neuroactive ligand-receptor interaction	140
rno04010	MAPK signaling pathway	120
rno00510	N-Glycan biosynthesis	18
rno01030	Glycan structures - biosynthesis 1	40
rno05212	Pancreatic cancer	52
rno00561	Glycerolipid metabolism	21
rno00564	Glycerophospholipid metabolism	26
rno04070	Phosphatidylinositol signaling system	33
rno00330	Arginine and proline metabolism	14
rno02010	ABC transporters - General	13
rno00120	Bile acid biosynthesis	16
rno04060	Cytokine-cytokine receptor interaction	85
rno05216	Thyroid cancer	22
rno04130	SNARE interactions in vesicular transport	24
rno00980	Metabolism of xenobiotics by cytochrome P450	28
rno04610	Complement and coagulation cascades	36
rno05221	Acute myeloid leukemia	36
rno04514	Cell adhesion molecules (CAMs)	55
rno04650	Natural killer cell mediated cytotoxicity	45
rno04670	Leukocyte transendothelial migration	44
rno04120	Ubiquitin mediated proteolysis	66
rno05020	Parkinson s disease	10
rno04350	TGF-beta signaling pathway	63
rno04520	Adherens junction	42
rno04530	Tight junction	57
rno05213	Endometrial cancer	34
rno04360	Axon guidance	50
rno04510	Focal adhesion	104
rno04512	ECM-receptor interaction	42
rno05222	Small cell lung cancer	39
rno04320	Dorso-ventral axis formation	17
rno04662	B cell receptor signaling pathway	31
rno00562	Inositol phosphate metabolism	23
rno04020	Calcium signaling pathway	108
rno04310	Wnt signaling pathway	67
rno04540	Gap junction	58
rno04720	Long-term potentiation	41
rno04730	Long-term depression	46
rno04912	GnRH signaling pathway	47
rno04916	Melanogenesis	44
rno00271	Methionine metabolism	8

Continued on Next Page...

Table 2.1 – Continued

	KEGG Pathway	Number of Genes
rno00450	Selenoamino acid metabolism	6
rno00230	Purine metabolism	65
rno00670	One carbon pool by folate	8
rno05010	Alzheimer s disease	19
rno00071	Fatty acid metabolism	36
rno00592	alpha-Linolenic acid metabolism	6
rno01040	Polyunsaturated fatty acid biosynthesis	15
rno03320	PPAR signaling pathway	55
rno00860	Porphyrin and chlorophyll metabolism	14
rno00010	Glycolysis / Gluconeogenesis	31
rno00260	Glycine, serine and threonine metabolism	27
rno04630	Jak-STAT signaling pathway	61
rno04640	Hematopoietic cell lineage	38
rno00910	Nitrogen metabolism	12
rno00240	Pyrimidine metabolism	25
rno00410	beta-Alanine metabolism	10
rno01430	Cell Communication	45
rno00380	Tryptophan metabolism	18
rno05210	Colorectal cancer	50
rno05220	Chronic myeloid leukemia	48
rno00590	Arachidonic acid metabolism	24
rno00591	Linoleic acid metabolism	13
rno05219	Bladder cancer	30
rno04210	Apoptosis	47
rno04920	Adipocytokine signaling pathway	56
rno01510	Neurodegenerative Diseases	17
rno05030	Amyotrophic lateral sclerosis (ALS)	16
rno00030	Pentose phosphate pathway	15
rno00051	Fructose and mannose metabolism	18
rno00052	Galactose metabolism	13
rno04910	Insulin signaling pathway	80
rno04012	ErbB signaling pathway	51
rno04110	Cell cycle	57
rno05215	Prostate cancer	52
rno04150	mTOR signaling pathway	28
rno04370	VEGF signaling pathway	35
rno04620	Toll-like receptor signaling pathway	44
rno04660	T cell receptor signaling pathway	43
rno04664	Fc epsilon RI signaling pathway	37
rno05211	Renal cell carcinoma	54
rno05214	Glioma	39
rno05218	Melanoma	38

Continued on Next Page...

Table 2.1 – Continued

	KEGG Pathway	Number of Genes
rno05223	Non-small cell lung cancer	34
rno00350	Tyrosine metabolism	21
rno00624	1- and 2-Methylnaphthalene degradation	6
rno00641	3-Chloroacrylic acid degradation	6
rno00650	Butanoate metabolism	18
rno00040	Pentose and glucuronate interconversions	9
rno00150	Androgen and estrogen metabolism	21
rno00500	Starch and sucrose metabolism	17
rno00100	Biosynthesis of steroids	16
rno00900	Terpenoid biosynthesis	6
rno00620	Pyruvate metabolism	25
rno00640	Propanoate metabolism	16
rno05040	Huntington s disease	18
rno05050	Dentatorubropallidoluysian atrophy (DRPLA)	7
rno00521	Streptomycin biosynthesis	7
rno04930	Type II diabetes mellitus	29
rno04950	Maturity onset diabetes of the young	15
rno00360	Phenylalanine metabolism	12
rno00400	Phenylalanine, tyrosine and tryptophan biosynthesis	10
rno00401	Novobiocin biosynthesis	6
rno00950	Alkaloid biosynthesis I	7
rno03010	Ribosome	27
rno05060	Prion disease	8
rno00920	Sulfur metabolism	7
rno00020	Citrate cycle (TCA cycle)	14
rno00251	Glutamate metabolism	17
rno00252	Alanine and aspartate metabolism	18
rno04710	Circadian rhythm	6
rno03030	DNA replication	12
rno00280	Valine, leucine and isoleucine degradation	19
rno00530	Aminosugars metabolism	9
rno00272	Cysteine metabolism	16
rno03020	RNA polymerase	6
rno01031	Glycan structures - biosynthesis 2	17
rno04940	Type I diabetes mellitus	16
rno03050	Proteasome	20
rno04115	p53 signaling pathway	23
rno00970	Aminoacyl-tRNA biosynthesis	11
rno04612	Antigen processing and presentation	23
rno00602	Glycosphingolipid biosynthesis - neo-lactoseries	7
rno04340	Hedgehog signaling pathway	12
rno05217	Basal cell carcinoma	16

Continued on Next Page...

Table 2.1 – Continued

	KEGG Pathway	Number of Genes
rno00361	gamma-Hexachlorocyclohexane degradation	10
rno00740	Riboflavin metabolism	8
rno03022	Basal transcription factors	13
rno00710	Carbon fixation	15
rno00533	Keratan sulfate biosynthesis	7
rno00220	Urea cycle and metabolism of amino groups	8
rno00140	C21-Steroid hormone metabolism	7
rno00480	Glutathione metabolism	22
rno00340	Histidine metabolism	10
rno04740	Olfactory transduction	14
rno00310	Lysine degradation	13
rno00460	Cyanoamino acid metabolism	6
rno00600	Sphingolipid metabolism	9
rno04614	Renin-angiotensin system	10
rno00760	Nicotinate and nicotinamide metabolism	10
rno00532	Chondroitin sulfate biosynthesis	7
rno00960	Alkaloid biosynthesis II	6
rno00565	Ether lipid metabolism	11
rno00062	Fatty acid elongation in mitochondria	6
rno00534	Heparan sulfate biosynthesis	6

Table 2.1: **Pathways Represented by this Dataset.** This table represents the 148 Pathways (their KEGG ID) and their respective number of genes that we have identified in this dataset. These provide a good representation of the spectrum of functions covered by the KEGG database.

2.3.3 Molecular Pathway Activity in Response to Chemical Exposure is Correlated to Toxicity.

In their original paper, Fielden et al [28] demonstrated that using statistical modelling techniques it is possible to identify subsets of genes predictive of late toxicity outcome. Since our strategy is based on simplifying the complexity of the data using indices of pathway activity we first asked whether these were also effective indicators of toxicity response. We first approached this question by clustering the chemicals on the basis of their ability to modify the transcriptional activity of a given pathway. Figure 2.3A shows that the profile of pathway perturbation

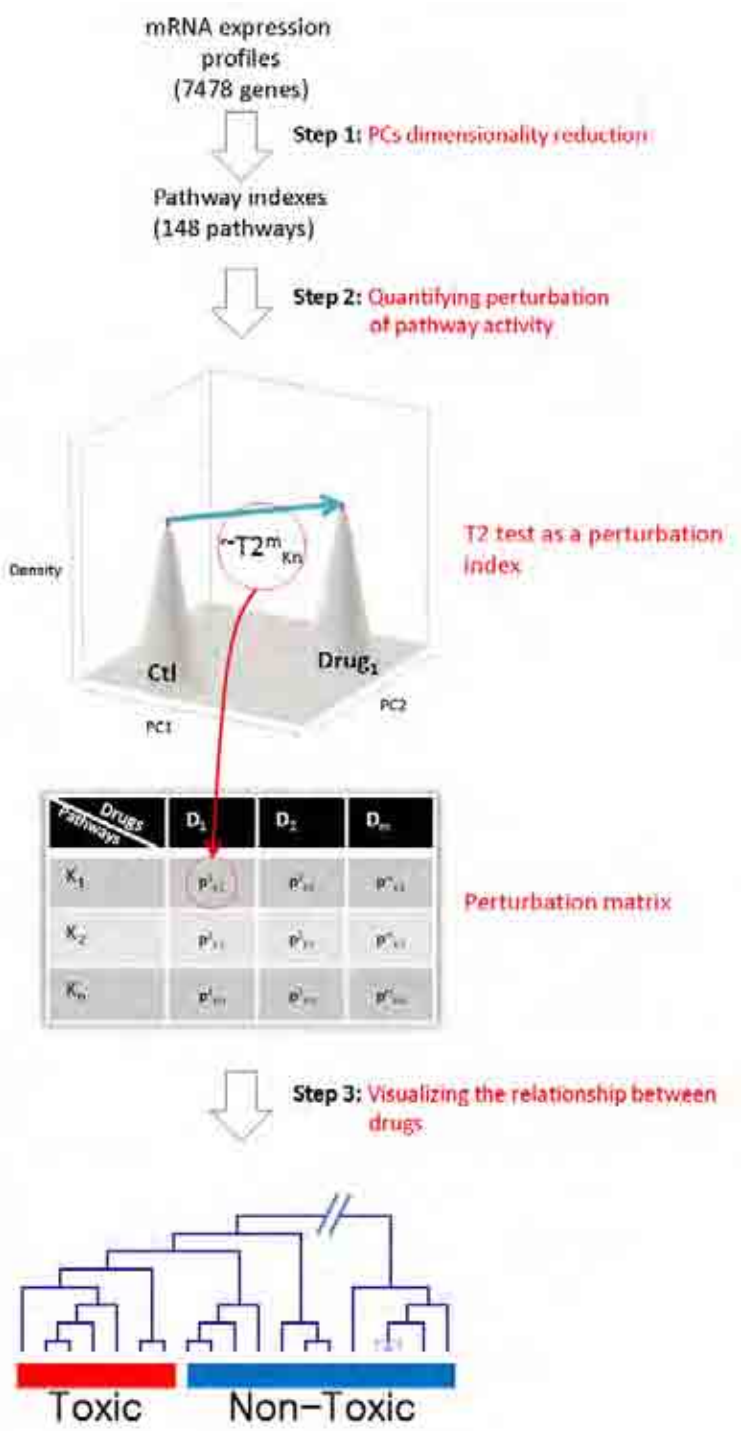


Figure 2.1: **Analysis Strategy to Compute Indices of Pathway Activity.** To compute the indices of pathway activity the first step is to summarize the gene expression profiles using PCA according to KEGG pathways. This results in 148 pathway indices summarized using two PCs. These PC can then be used as an input to a T^2 Hotelling s statistics to compute the perturbation index for a specific drug as compared to a matched control group. The third step is to visualize the relationship between the drugs with the use of a hierarchical clustering. We can then show that the dimensionality reduction in step 1 is biologically relevant to use in the subsequent analysis.

Subcategory	Percentage	Subcategory	Percentage
1.1 Carbohydrate Metabolism	80.00%	4.3 Cell Growth and Death	60.00%
1.2 Energy Metabolism	50.00%	4.4 Cell Communication	100.00%
1.3 Lipid Metabolism	76.00%	4.5 Circulatory System	0.00%
1.4 Nucleotide Metabolism	100.00%	4.6 Endocrine System	86.00%
1.5 Amino Acid Metabolism	85.00%	4.7 Immune System	69.00%
1.6 Metabolism of Other Amino Acids	44.00%	4.8 Nervous System	67.00%
1.7 Glycan Biosynthesis and Metabolism	33.00%	4.9 Sensory System	50.00%
1.8 Biosynthesis of Polyketides and Nonribosomal Peptides	0.00%	4.10 Development	100.00%
1.9 Metabolism of Cofactors and Vitamins	33.00%	4.11 Behavior	33.00%
1.10 Biosynthesis of Secondary Metabolites	17.00%	5.1 Cancers	93.00%
1.11 Xenobiotics Biodegradation and Metabolism	15.00%	5.2 Immune Disorders	0.00%
1.12 Overview	11.00%	5.3 Neurodegenerative Diseases	100.00%
2.1 Transcription	100.00%	5.4 Circulatory Diseases	0.00%
2.2 Translation	100.00%	5.5 Metabolic Disorders	100.00%
2.3 Folding, Sorting and Degradation	75.00%	5.6 Infectious Diseases	0.00%
2.4 Replication and Repair	17.00%	6.1 Chronology: Antibiotics	0.00%
3.1 Membrane Transport	25.00%	6.2 Chronology: Antineoplastics	0.00%
3.2 Signal Transduction	79.00%	6.3 Chronology: Nervous System Agents	0.00%
3.3 Signaling Molecules and Interaction	80.00%	6.4 Chronology: Other Drugs	0.00%
4.1 Transport and Catabolism	0.00%	6.5 Target Based Structure Classification	0.00%
4.2 Cell Motility	33.00%	6.6 Skeleton Based Structure Classification	0.00%

Table 2.2: **Percentage of KEGG Pathways Found within each Subcategory of the KEGG Database.**

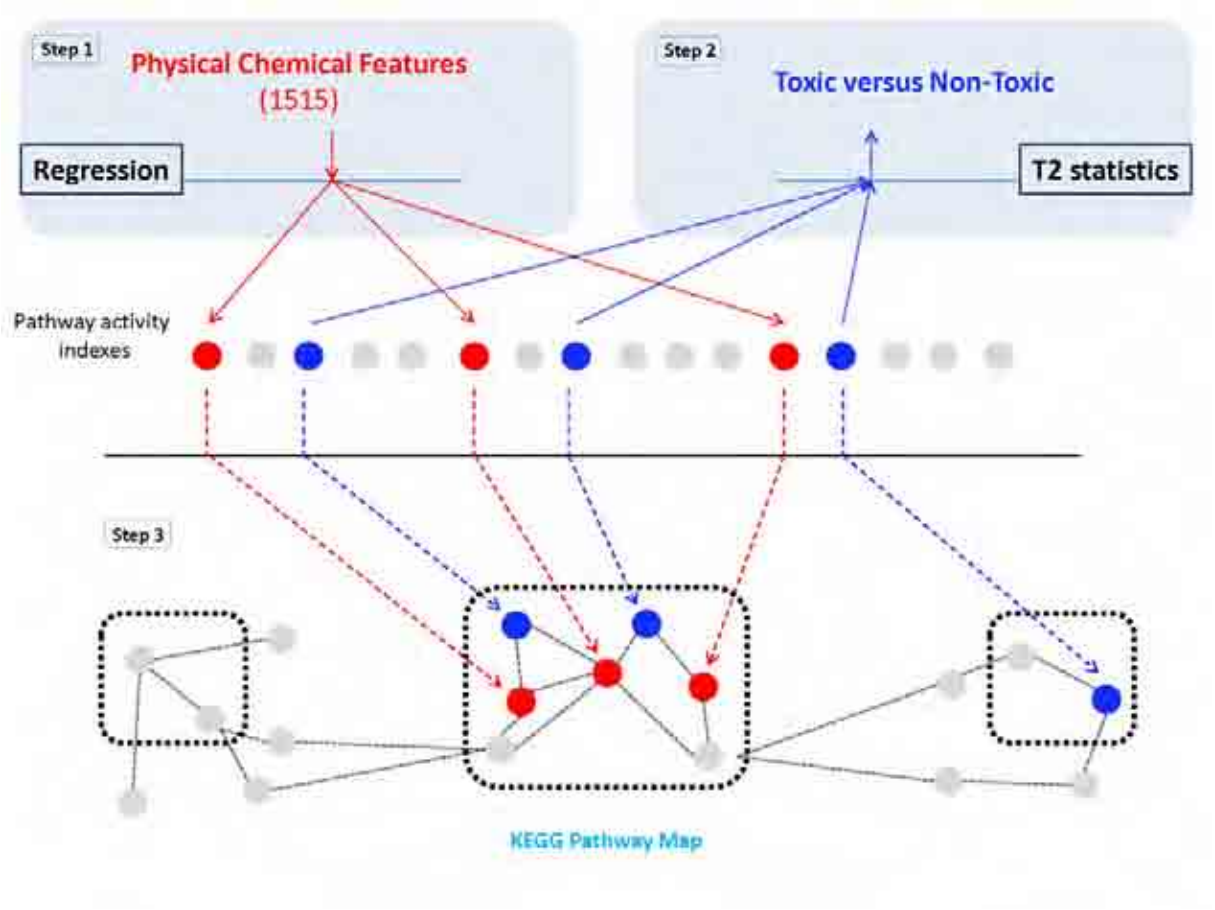


Figure 2.2: **Integrating Pathways Associated to PCFs and Toxicity.** Alongside using a regression based model to identify pathways associated to PCFs (Step 1) we also identified pathways associated to toxicity by the use of the T^2 statistics (Step 2). The resulting pathways were then mapped onto a KEGG pathway map to identify clusters of pathways associated to both PCFs and toxicity (Step 3). Finally we asked the question if the PCFs we have found to be associated to pathways are a better predictor of toxicity.

is indeed informative of chemical toxicity. In particular, cluster analysis succeeded in grouping 12 out of 15 nephrotoxic chemicals within a well-defined cluster (Figure 2.3A). Analysis of the individual PCs revealed that the second PC on its own was sufficient to reproduce clustering of toxic chemicals without significant loss of information (Figure 2.3B and 2.3C). In order to identify the molecular pathways differentially modulated in response to toxic chemicals we directly compared the index of pathway activity between samples treated either with nephrotoxic or non-nephrotoxic chemicals. This analysis identified 21 pathways which were differentially modulated (FDR<1%, Table 2.3). These can be grouped into three main functional categories: 1) metabolic pathways such as *glycerophospholipid metabolism* or *amino sugar metabolism*, 2)

pathways with a strong signalling component such as *parkinson s disease*, *phosphatidylinositol signalling* and *prostate cancer* and 3) cell communication pathways such as *cell communication* and *focal adhesion*. The KEGG pathway terms *parkinson s disease*, *prostate cancer*, *pancreatic cancer* and *renal cell carcinoma* do not specifically include the term signalling in their definition but are indeed representing primarily signalling pathways. More specifically, the pathway *parkinson s disease* represents the molecular events downstream dopamine stimulation, which is a major player in synaptic transmission and it is effectively linked to signalling pathways controlling vasoconstriction. This pathway is important for kidney physiology where dopamine release induce an increase in renal blood flow, urinary volume and excretion of sodium and potassium. This then leads to an increase in glomerular filtration rate as well as a depletion of plasma cyclic AMP [184]. The pathway *Prostate cancer* represents components of the *MAPK signalling* and *p53 signalling pathways* which are included in the response downstream of cytokine stimulation. Specific signalling pathways associated to the *Pancreatic cancer* pathway are *ErbB*, *Jak-STAT*, *VEGF*, *TGF-β*, *MAPK* and the *p53 signalling pathway*. These are not only relevant to the biology of cancer (alteration in these signalling pathways destabilize growth inhibition and promote tumour growth activity [185]) but also to kidney response to stress and regeneration [186].

2.3.4 Chemical Features are Predictive of Molecular Pathway Activity.

Having demonstrated that indices of pathway activity are representative of the biological effect of chemicals we addressed the hypothesis that a subset of PCFs may be correlated to the kidney transcriptional response to drug exposure. The statistical framework we have used to address this hypothesis (described in detail in section 2.5.5) relies on a regression model explaining the activity of a given pathway (which we remind is the first or second principal component computed from the gene expression matrix associated to a given pathway) as a linear combination of three chemical features. The model also includes interaction components to take into account potential synergistic effects between chemical descriptors. We successfully identified predictive models ($R^2 > 0.5$) for 19 of the 148 pathways represented in the Iconix dataset. It is worth

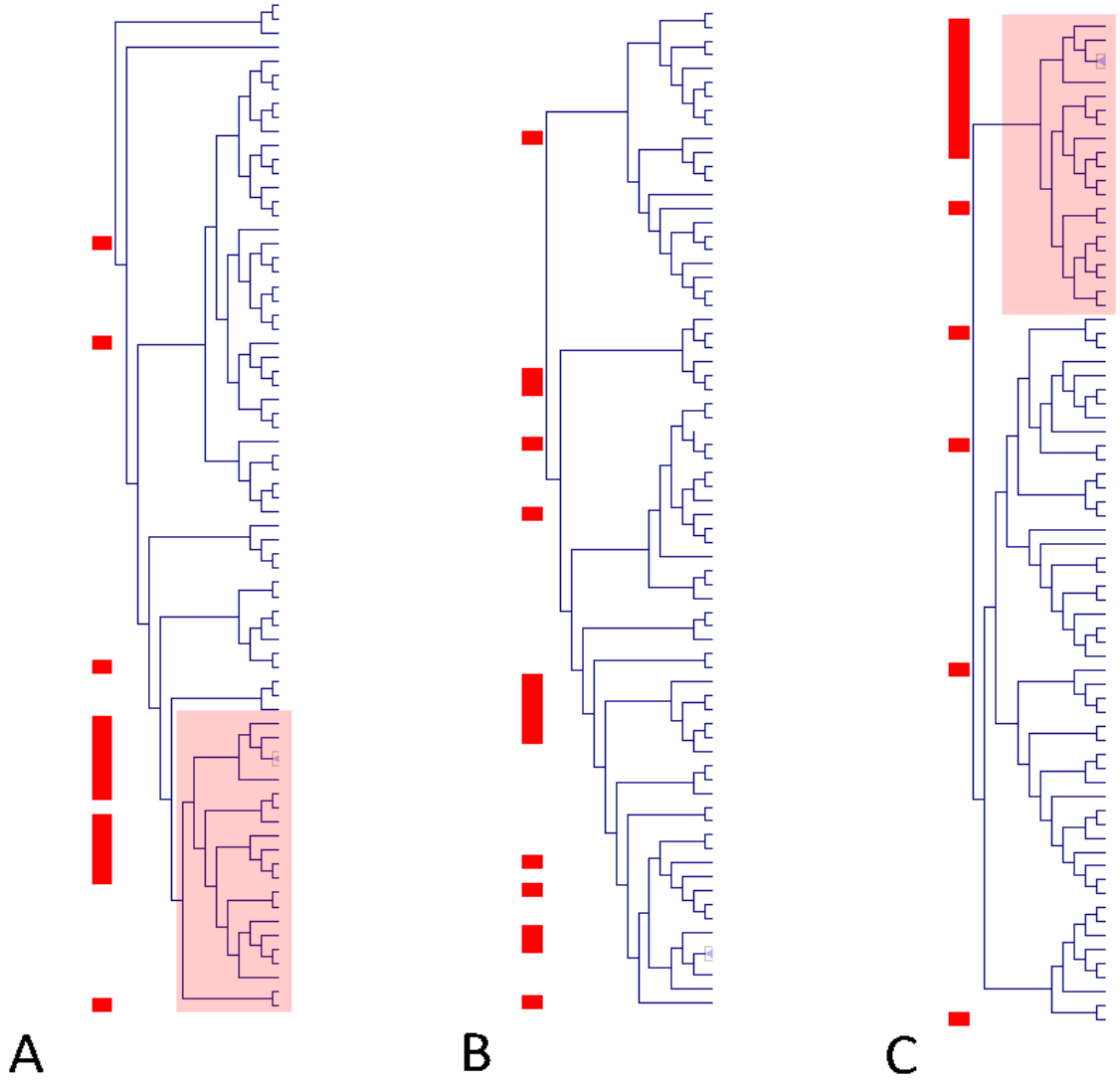


Figure 2.3: **Hierarchical Clustering of Chemicals based on Pathway Modulation Profiles.** The figure shows the clustering of chemicals on the basis of the extent of change of transcriptional activity in molecular pathways after exposure. Panel A represent the relationship between the samples when the change in pathway activity is represented simultaneously by the first and second PCs (multi-variate T^2 Hotelling test). Panels B and C represent respectively the results of cluster analysis when the change in pathway activity is estimated by the PC1 or the PC2 (univariate t-test). Notice that toxic chemicals cluster (highlighted areas on Panels A and C) on the basis of the multivariate test and that the information associated to toxicity is primarily represented by the PC2. Toxic chemicals have been highlighted using a red square on the left of each clustering.

KEGG ID	Pathway Name	Number of Genes	T^2 Score
rno05020	Parkinson s disease	10	31.71
rno00564	Glycerophospholipid metabolism	25	31.29
rno05215	Prostate cancer	52	26.66
rno02010	ABC transporters - General	13	26.35
rno04070	Phosphatidylinositol signaling system	32	22.56
rno04130	SNARE interactions in vesicular transport	24	21.54
rno00760	Nicotinate and nicotinamide metabolism	10	21.17
rno01430	Cell Communication	45	21.07
rno00530	Aminosugars metabolism	9	20.77
rno04120	Ubiquitin mediated proteolysis	65	20.40
rno05030	Amyotrophic lateral sclerosis (ALS)	16	19.64
rno03010	Ribosome	26	19.54
rno03050	Proteasome	20	19.35
rno05212	Pancreatic cancer	50	19.21
rno00230	Purine metabolism	64	16.14
rno04330	Notch signaling pathway	20	15.31
rno01031	Glycan structures - biosynthesis 2	17	14.96
rno04612	Antigen processing and presentation	23	14.96
rno04510	Focal adhesion	102	14.49
rno04320	Dorso-ventral axis formation	17	14.32
rno05211	Renal cell carcinoma	52	13.71

Table 2.3: **Pathways Perturbed by Toxic Chemicals.** This table shows 21 KEGG pathways that were found to be significantly perturbed by nephrotoxic chemicals (FDR < 1%). The number of genes in each pathway and the value of the T^2 hotelling statistics are shown respectively in the third and fourth columns.

noticing that, pathway activity indices (Figure 2.3) as well as the original gene expression data (Figure 2.5), separates toxic from non-toxic chemicals across the second PC whereas the first component is likely to represent non-specific effects (Figure 2.5). Therefore the association between PCFs and the pathway activity indices build using the second component is biologically reasonable. Among pathways associated to chemical features we observed a large number of signalling pathways as well as some metabolic pathways (i.e. *glycolysis*, *porphyrin metabolism*, *chlorophyll metabolism* and *glutathione metabolism*). Two of these pathways were also found to be associated to toxicity in the analysis described in the previous paragraph (*prostate cancer* and *cell communication*). PCFs selected in the models could be assigned to several descriptor groups. Figure 2.4 summarizes in a graph format the most frequent combination of features descriptors groups selected in the chemical feature models. A key feature of the selected mod-

els is the importance of interaction components which in most cases explain an average of 50% of the model variance (Figure 2.6). Descriptor groups pairs such as descriptors that describe patterns in the connection of specific atoms with each other (ET-State) and geometrical descriptors or descriptors of special fragments that describe a path or cycle (GSFRAG) with itself are predominantly chosen by our method. All these descriptors classes capture different types of structural information. For example, GSFRAG descriptors identify specific chemical motives such as the size of a ring, or the length of linear connections; ET-States descriptors describe patterns in the connection of specific atoms with each other and geometrical descriptors are designed to capture patterns in the overall topology of the molecule.

2.3.5 Pathways Whose Activity is Correlated to Chemical Features are Part of a Signalling System Closely Connected with Cellular Communication and Related Functions.

Regression analysis described in the previous paragraph identified 19 pathways whose activity could be predicted by a combination of chemical features (See Figure 2.7 for some examples). Because of the apparent similarity in the molecular functions represented in these pathways we reasoned that these may be closely connected within the KEGG pathway map. In order to test this hypothesis we represented the relationship between individual pathways (defined by their degree of overlap) using hierarchical clustering. In this analysis KEGG pathways which share a larger number of components are represented in close proximity in the dendrogram. The visual inspection of the dendrogram confirmed that pathways, whose overall activity can be predicted by combinations of chemical features, were grouped in a compact cluster within the KEGG map (Figure 2.8). This cluster defines a KEGG super-pathway that represents a number of signalling networks directly connected to effectors functions of direct relevance with tissue morphogenesis such as *actin remodelling* and *cell communication*. Interestingly, *cell communication* which we already mentioned to be associated to both PCFs and toxicity represents multiple signalling and effectors components of the cell to cell communication machinery. These include tight junction, gap junctions, adherence junctions, desmosomes and extracellular matrix components. The

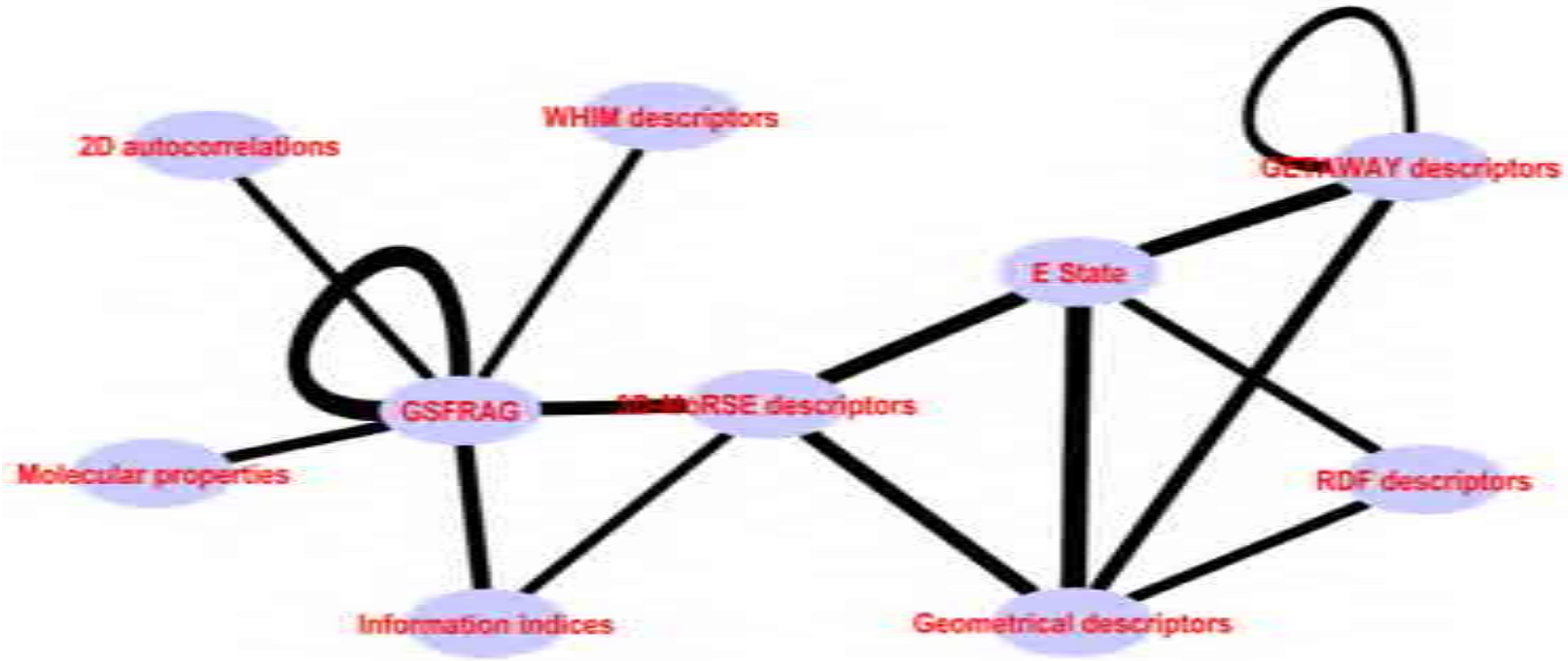
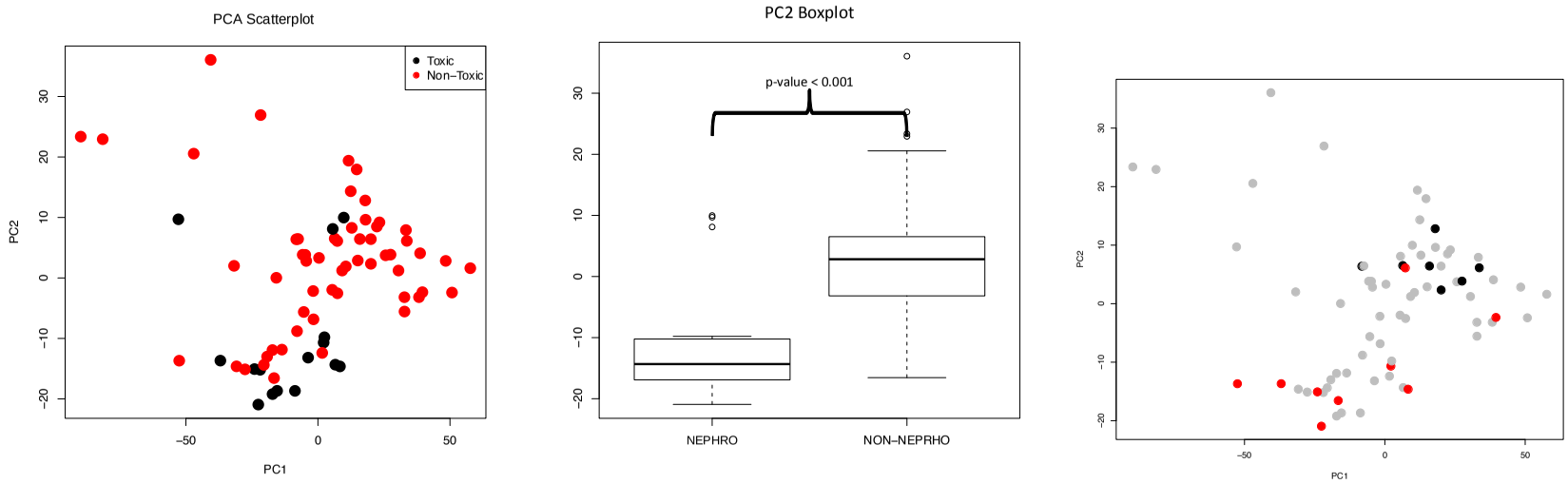


Figure 2.4: **Visual Summary of Descriptor Connections.** The network represents the number of interactions between PCFs descriptor groups computed from pooled models. The thickness of the line is proportional to the number of pathways in which PCFs of a given descriptor group are selected in an interaction component of a predictive models. The highest value edge is found between ET-State and Geometrical descriptors in which 11 out of 19 pathways were found to contain models based on features from these 2 descriptor groups.



(a) (b) (c)

Figure 2.5: **Additional Scatter Plots Representing the Chemical Space.** (a) PCA scatterplot of the chemical space using all genes clustered into KEGG Pathways. Chemicals marked black or red are nephrotoxic and non-nephrotoxic respectively. (b) Boxplot showing the separation on the second component between nephrotoxic and non-nephrotoxic chemicals. A t-test between the two sets has a p-value < 0.001. (c) Dose separation on the PCA plot, low dose chemicals are marked in red and high does chemicals in black. We observe a diagonal relationship between PC1 and PC2 separating the dose. More specifically as shown in (a) the toxic chemicals separate on the 2nd PC. This implies that part of the non-toxic dose component is summarized in PC1.

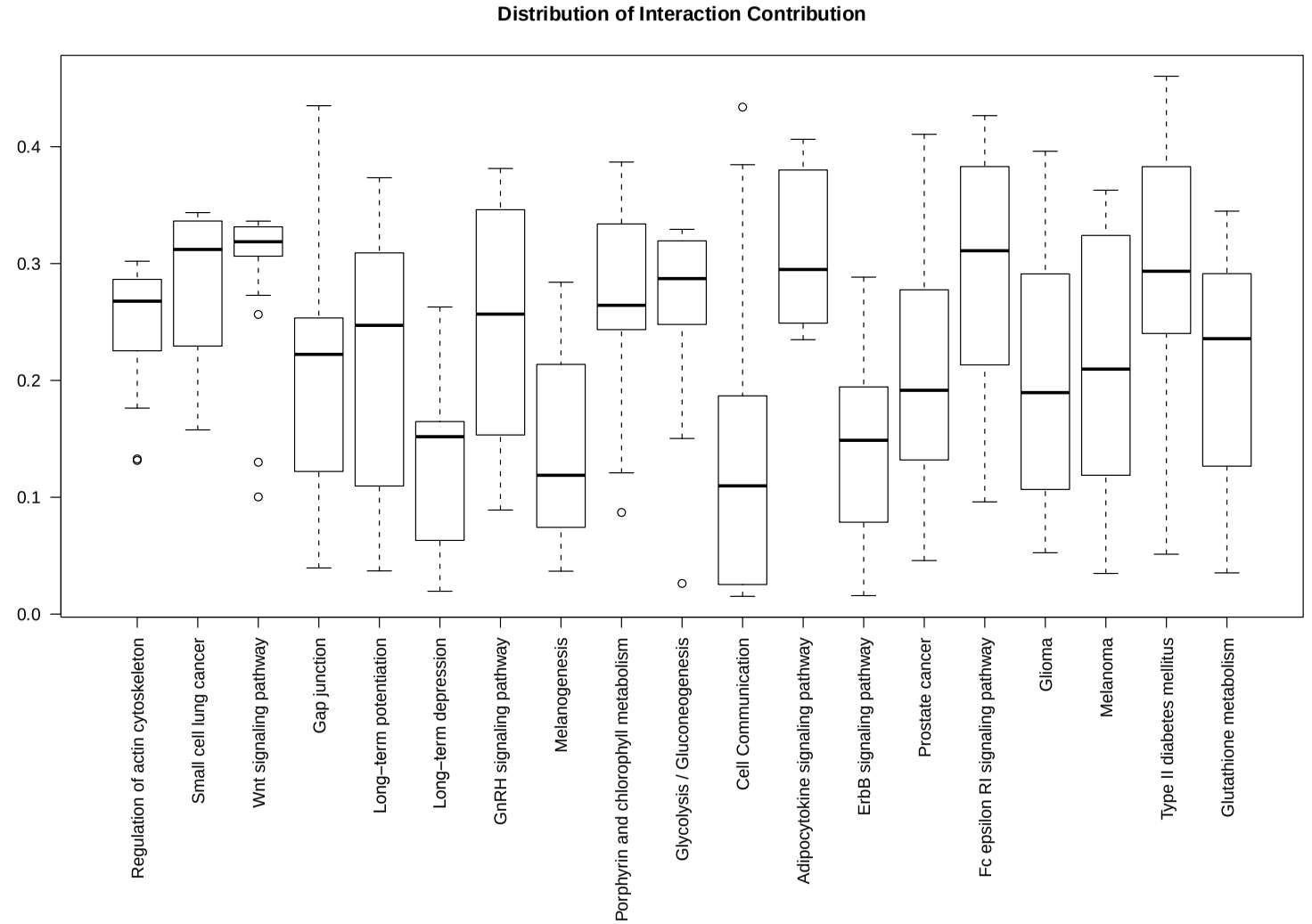


Figure 2.6: **Distribution of the Interaction Components of the 19 Pathways Associated to PCFs.**Most of the interaction components add more than 50% towards the resulting model.

cluster of interconnected KEGG pathways defined by the association with chemical physical features therefore represents a network of signalling components which directly interact with a toxicity associated core of cell communication components.

2.3.6 PCFs Correlated to Molecular Pathway Activity are Best Predictors of Chemical Induced Toxicity.

The functional link between pathways associated to PCFs and toxicity may imply that the selected PCFs may be themselves predictive of renal tubular degeneration. In order to test this hypothesis we developed multivariate statistical models predictive of toxicity selecting from PCFs associated to molecular pathway activity. We then compared these with models developed from PCFs which where uncorrelated to the pathway associated PCFs subsets. Figure 2.9A shows that features predictive of molecular response are more predictive of toxicity outcome (average accuracy of 76% versus 68%, p-value $< 10^{-3}$). In order to identify the most representative PCFs subset, we developed representative models based on the three features which were most frequently represented in the model populations. Consistent with the previous result, the model built using PCFs associated to molecular response has higher sensitivity and specificity (sensitivity 0.781, specificity 0.871) than the one build with uncorrelated PCFs. This is reflected by a clearer sample separation in the PCA plot (Figure 2.9B). Features represented in the most predictive model combine two RDF descriptors and a WHIM descriptor whereas the unselected features model contains a GSFRAG, a GETAWAY descriptor and a 2D-autocorrelation descriptor. The model based on PCFs predictive of transcriptional response shows that toxic chemicals are characterized by high polarisability (RDF020p), low electronegativity (RDF040e) and low symmetry (G2u). Although the difference in accuracy (8%) is not particularly high, the results confirm that PCFs chosen by our method have a significantly higher ability to discern toxic from non-toxic chemicals.

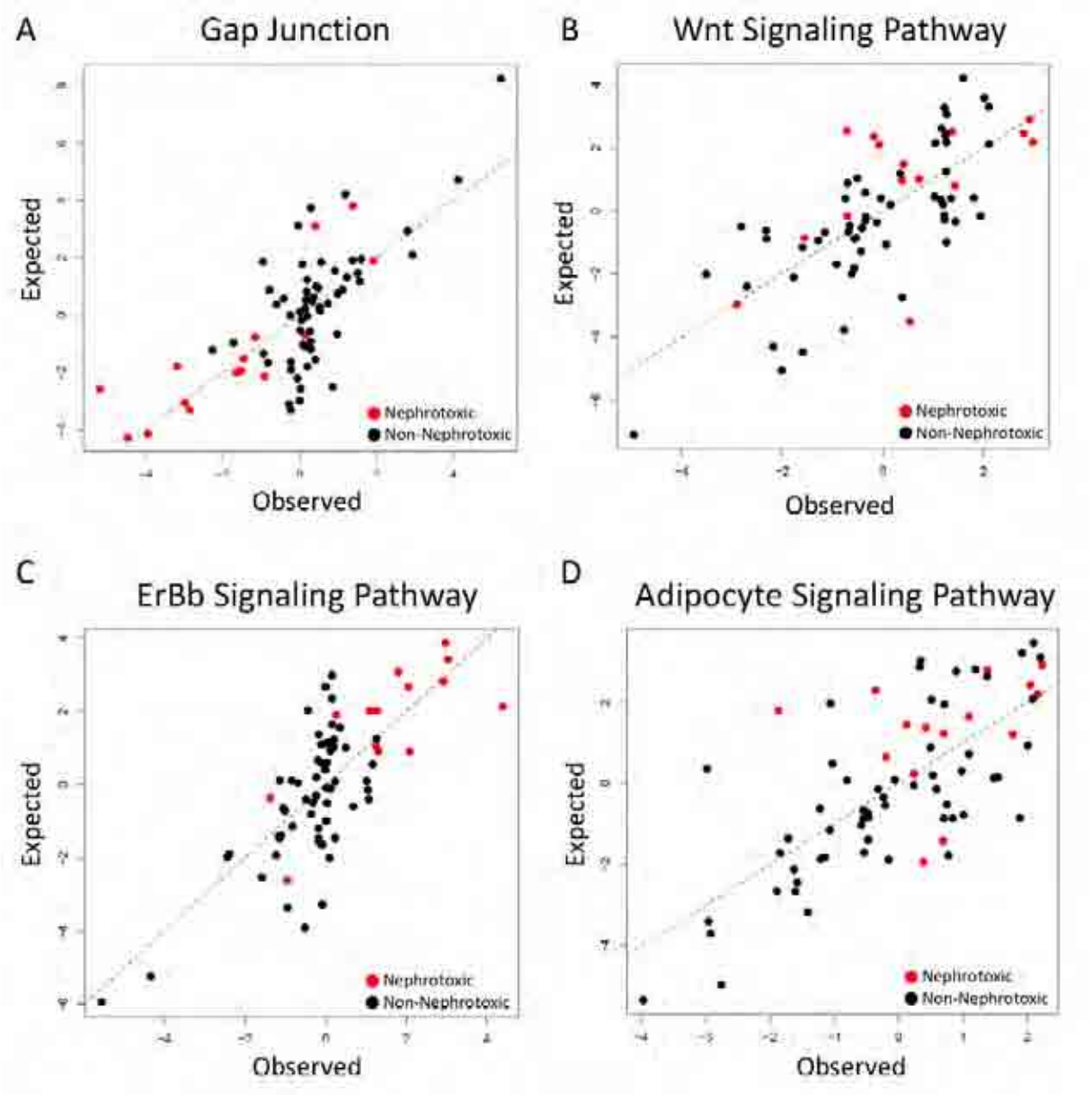


Figure 2.7: **Example Models Linking PCFs with Molecular Pathway Activity.** The figure shows the relationships between the observed (x axis) and predicted (y axis) indices of pathway activity for a number of exemplar KEGG pathways. Nephrotoxic samples are represented by red dots whereas non-nephrotoxic samples are represented by black dots. *Gap Junction* and *ErbB Signaling Pathway* contain features belonging to ET-State indices, Geometrical descriptors and RDF descriptors. The R^2 values are 0.55 and 0.57 respectively. Wnt Signaling Pathway and Adipocyte Signaling Pathway contain features belonging to GSFRAG, Information indices, Edge adjacency indices and 3D-MoRSE descriptors. The R^2 values are 0.52 and 0.51 respectively. Note that models containing a feature from E-State indices and RDF descriptors better separate nephrotoxic and non-nephrotoxic samples.

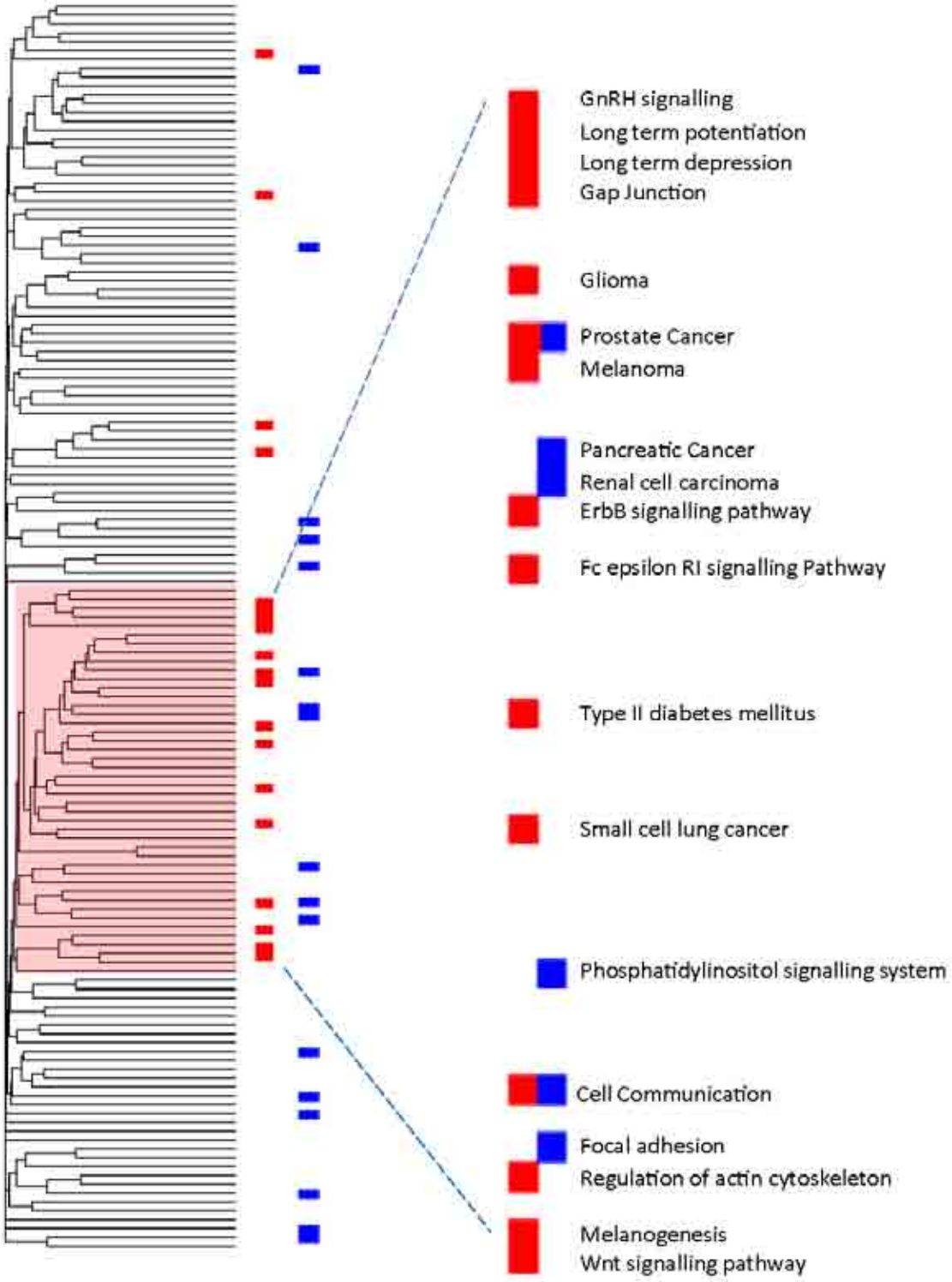


Figure 2.8: **KEGG Pathway Topology Map.** The Figure shows a dendrogram representing the degree of similarity between different KEGG pathways. Pathways marked in red are pathways that were found to be associated to chemical features (19), and pathways marked in blue have been found to be predictive of toxicity (21). Pathways whose activity is predicted by PCFs group in a tight cluster. Note that the majority of toxicity annotated pathways cluster towards the lower half of the dendrogram, close to pathways linked to PCFs.

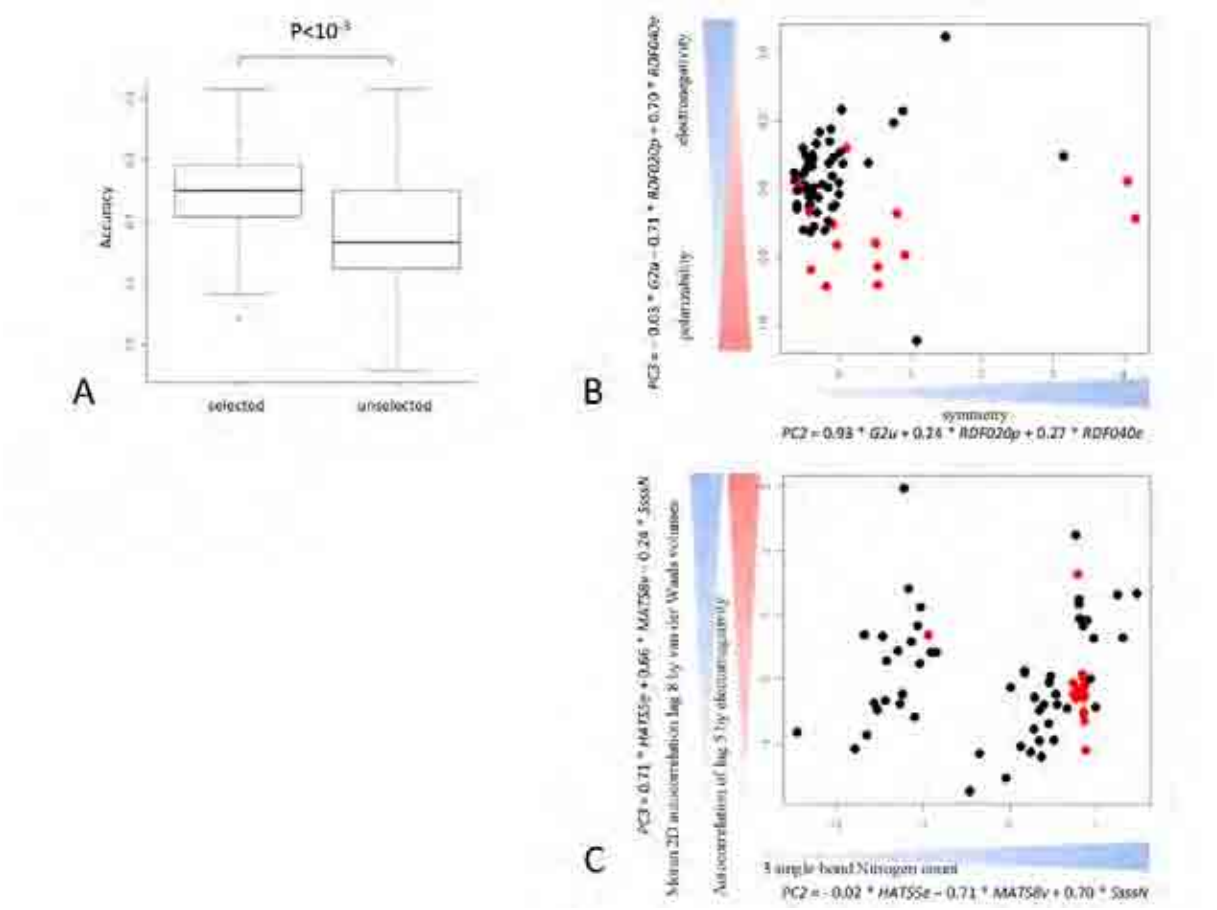


Figure 2.9: **PCFs Linked to Molecular Response are Better Predictors of Toxicity.** Panel A shows the comparison between the classification accuracy of models predictive of toxicity and developed selecting from PCFs which are predictive of molecular response and those developed using uncorrelated PCFs. Note that PCFs linked to molecular response have a higher predictivity ($p < 10^{-3}$). Panels B and C show the PCA representation of the samples using the 3 most represented features in the model populations. The information for the best separation in both instances is present in PC2 and PC3. The equations for panel B show that a high increase in symmetry and high polarizability and low electronegativity is predictive of toxicity. In the case of the unselected features panel C toxic chemicals do cluster together but are specific to containing a nitrogen with a triple single bond and a low autocorrelation.

2.4 Discussion

The most important finding of our study is the demonstration that specific combinations of chemical descriptors can be predictive of the transcriptional activity of pathways, always using the second PC, representing the molecular state of a target organ after chemical exposure. These pathways (i.e. *ErbB Signalling*, *Wnt signalling*, *Long-Term Depression*, *Long-Term potentiation* and several cancer pathways) mainly represent signalling pathways which in our model define the main domain of interaction between chemicals and cellular molecular response (Figure 2.10). We have shown that toxicity pathways with relevance to renal tubular degeneration are closely associated to this domain in the context of a KEGG pathway map. We explored close pathways by integrating the networks to establish whether, beyond the topological proximity, we could also identify a functional relationship between them. In this context we devised three interconnected pathways that could mechanistically explain the observed connection between chemical features, pathway activity and toxicity outcome. Figure 2.10A shows how a possible interaction between the *Wnt signalling pathway*, *regulation of actin cytoskeleton* (linked to PCFs) and *focal Adhesion* (predictive of toxicity outcome) could lead to a perturbation of actin cytoskeleton polymerization. More specifically, Wnt/Fz signalling activates the small GTPase Rho to control cell migration during tissue remodelling and development. This activation requires Dvl-Rho complex formation which is assisted by Daam1. From this it is clear that the integration of these topologically linked pathways represent a true series of biochemical events linking the binding of the Wnt ligand, through activation of Daam1 to the actin polymerization machinery. A plausible disturbance of mitochondrial respiration and energy balance by means of reactive oxygen species (ROS) generation is shown in Figure 2.10B. Lastly growth factor mediated modulation of the cell cycle, adhesion and cell migration through TGF-α is shown in Figure 2.10C. This pathway module results from the integration of the *ErbB signalling pathway* (linked to PCFs) and *Pancreatic Cancer* (predictive of toxicity outcome). In this case the pathway linked to toxicity is a sub-network of the *ErbB signalling pathway* which represents the specific effects on tissue remodelling via regulation of cell growth, apoptosis and differentiation.

The common feature among these hypothetical mechanisms is the association between chemical features and membrane associated cellular signalling and the large overlap between this and effectors pathways. Genes within each pathway are co-ordinately regulated across exposures suggesting that what we are modelling is not the effect of a small subset of highly regulated genes. Moreover, by mapping the direction of change between toxic and non-toxic chemicals on the KEGG pathway maps we observe that chemical exposure is associated to a coordinate overexpression of genes in signalling and effector genes (Figures 2.11 – 2.24). It is therefore not unreasonable to hypothesize that the diverse spectrum of toxic chemicals used in this study may act via a general mechanism involving interaction with cellular membranes. This hypothesis is also consistent with the finding that polarisability is a key feature of the toxic chemicals studied (Figure 2.7 and 2.9). The interaction between chemicals and cellular membranes may perturb receptor signalling inducing changes in the expression of genes encoding for signalling components and ultimately creating an unbalance in the expression of effectors pathways involved in tissue dynamics and homeostasis. The regression models we built showed that, in many cases, there is a continuum of effects influencing the molecular state of a target pathway and that, in specific pathways, (i.e. *Gap junction* and *ErbB signalling pathways*) toxicity is observed either above or below a given threshold of pathway activity (Figure 2.7). This is showing that only chemicals that can substantially perturb key signalling pathways are able to induce stress responses such as disturbance of inter-cellular communication and mitochondrial disturbances that are frequently associated with subsequent cellular toxicity [187, 188]. It is possible that the proposed mechanism may be a general unifying mode of toxicity probably secondary to a range of initial specific mechanisms and that may act in parallel to the interaction with specific molecular targets. In this context, it is known that multiple and target-specific mechanisms of action of xenobiotics are responsible for drug induced nephropathy. For example, the targets of the initial insult may be at the level of altered blood flow, glomerular injury, direct proximal tubule damage or other tubule or papillary targets [189]. Furthermore nephropathy might be a direct action of the agent on nephrons or an indirect action such as via a reduction of prostaglandin production such as with salicylic acid, or via precipitation of liver-derived alpha-

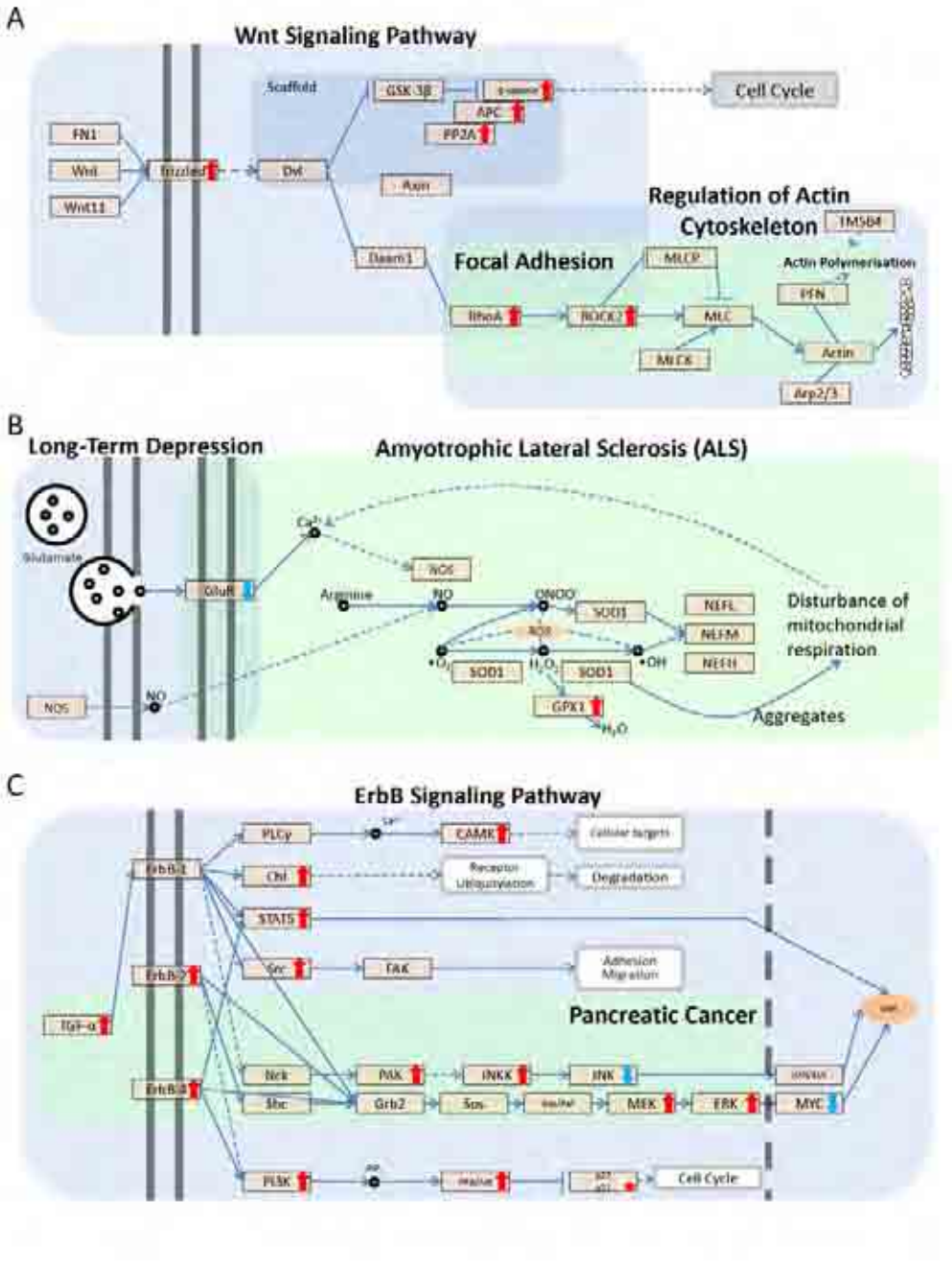


Figure 2.10: **Association between PCFs and Toxicity Associated Pathways.** The figure represents the detailed relationship between pathways associated to chemical hits and pathways associated to toxicity. Pathways with membrane component were mostly associated to chemical hits whereas pathways with downstream signalling components were mostly associated to toxicity. This figure represents three possible links between pathways associated to chemical hits (*Wnt Signaling Pathway*, *Long-Term Depression* and *ErbB Signaling Pathway*) and toxicity (*Focal Adhesion*, *ALS* and *Pancreatic Cancer*) through shared genes between the pathways. Although each link presents a mechanism of action these were only implied by the pathway associated to toxicity. Genes found to be up or down regulated have been marked with a red or a blue arrow respectively.

2-u-globulin as a result of chemical binding (e.g. d-limonene) [190]. Prominent as classes of nephrotoxic agents are halogenated hydrocarbons such as chloroform and bromobenzene and classes of therapeutic agents including nonsteroidal anti-inflammatory drugs, aminoglycosides and the anticancer agent cisplatin. These facts might suggest insurmountable difficulties in prediction of effects from structural characteristics because of a multiplicity of mechanisms. However, the focus of this paper is predominantly on agents that directly act on the tubular (principally proximal tubule) epithelial cells. Our study has shown that there are features of signalling disturbance that associate with both chemical structural parameters and also with additional molecular pathways that associate with toxicity. Integration of the datasets shows that it is possible to link structure to pathology via the two layers of analysis allowing a reconstruction of a series of pathways. The approach offers a new dimension to the existing strategies of databases that associate structure directly to known toxicity features through training (e.g. DEREK and TOPKAT [106] and the OECD Toolbox (www.oecd.org).The common signalling disturbance identified is thus hypothesised to lead to secondary effects linked to toxicity. It is the genome-wide surveillance strategy that has allowed the identification of the linkage which would not have been possible from more targeted analysis of individual mechanisms. Since the time point for the molecular changes observed is five days after exposure, it is also possible that the changes represent secondary intermediate modes of change rather than specific early mechanistic interactions. Interestingly, the modelled features associated with toxicity are not necessarily limited to nephrotoxicity. The biological implications of this work are further strengthened by the observation that chemical feature selection based on functional pathway activity leads to more predictive toxicity models (sensitivity 78.1%, specificity 87.1%). Therefore linking gene expression to chemical features identifies a sub-selection of features which are more linked to toxicity. We therefore propose that by integrating gene expression profiles with chemical feature information it may be possible to isolate a sub-group of features that are highly important in characterizing specific phenotypic effects allowing for a much better characterization of yet untested chemicals. The development of these methodologies is particularly important as large datasets representing a broader spectrum of chemicals are expected to be-

come available. An excellent example of these publicly available datasets is the ToxCast^{TM} [22] program which is currently running at the U. S. Environmental Protection Agency [191]. Several potential improvements may be necessary to make the approach fully generalizable. For example, the computation of pathway indices we have implemented is based on the use of PCs ensuring that a large percentage of variance (80% in this case) is retained. Although this is likely to work for most of the datasets, it is possible that PCA, which is based on a linear combination of variables, may not be able to capture more complex relationships with PCFs. Therefore it may be useful to consider other methods such as independent component analysis or a non-linear version of PCA. This issue is particularly important considering that in complex exposure experiments the component of variation associated to the interesting biological effect may be associated to non-specific effects of toxicity. It is therefore important that the procedure used for the construction of pathway indices has the potential to decompose these effects. However, even at the present stage of development, the broad application of the analysis strategy we have pioneered will improve our ability to identify mechanistic markers of toxicity and will help to better understanding the relationship between drug PCFs and cellular physiology.

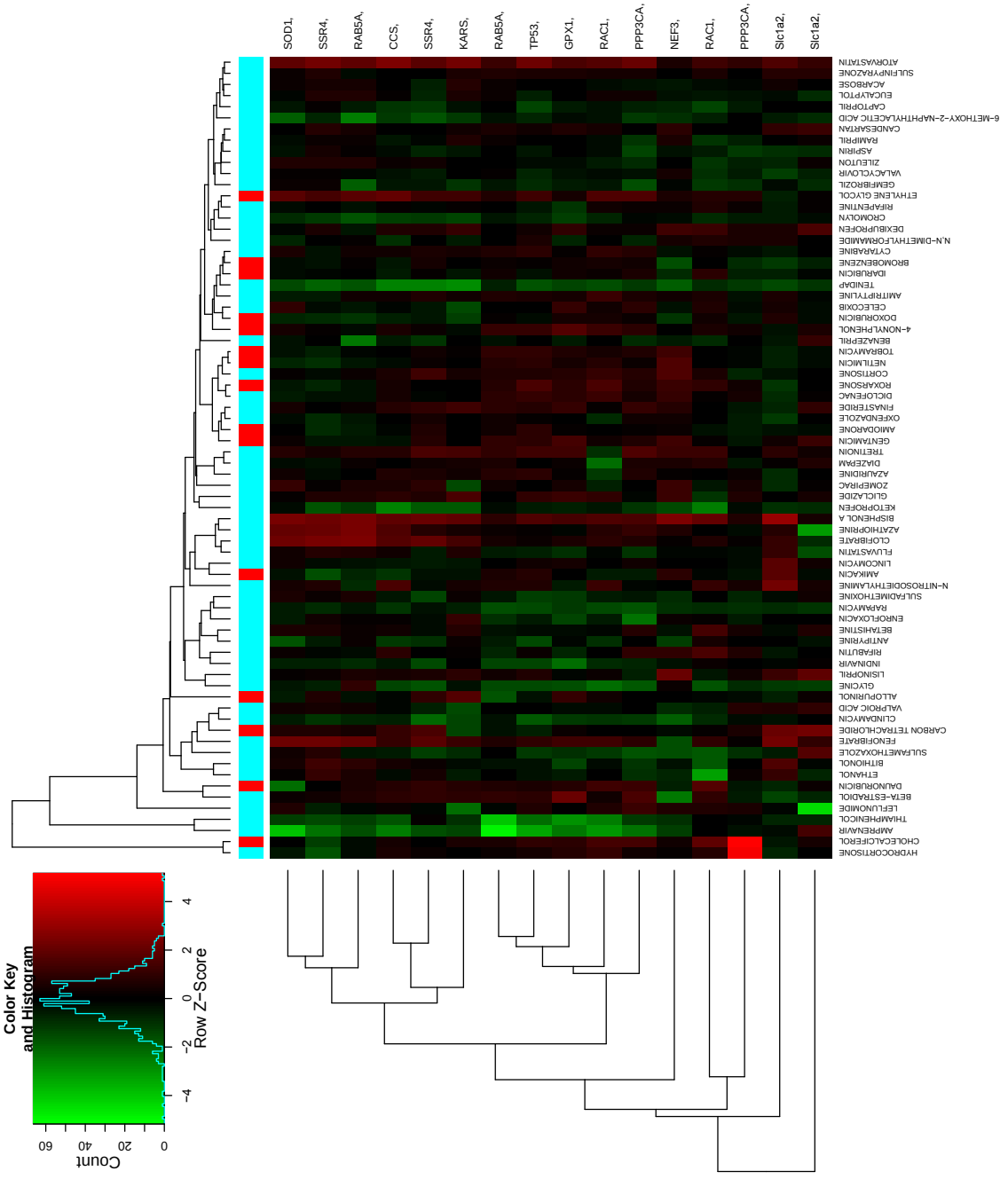


Figure 2.11: Heatmap of the Genes belonging to Amyotrophic Lateral Sclerosis (ALS).

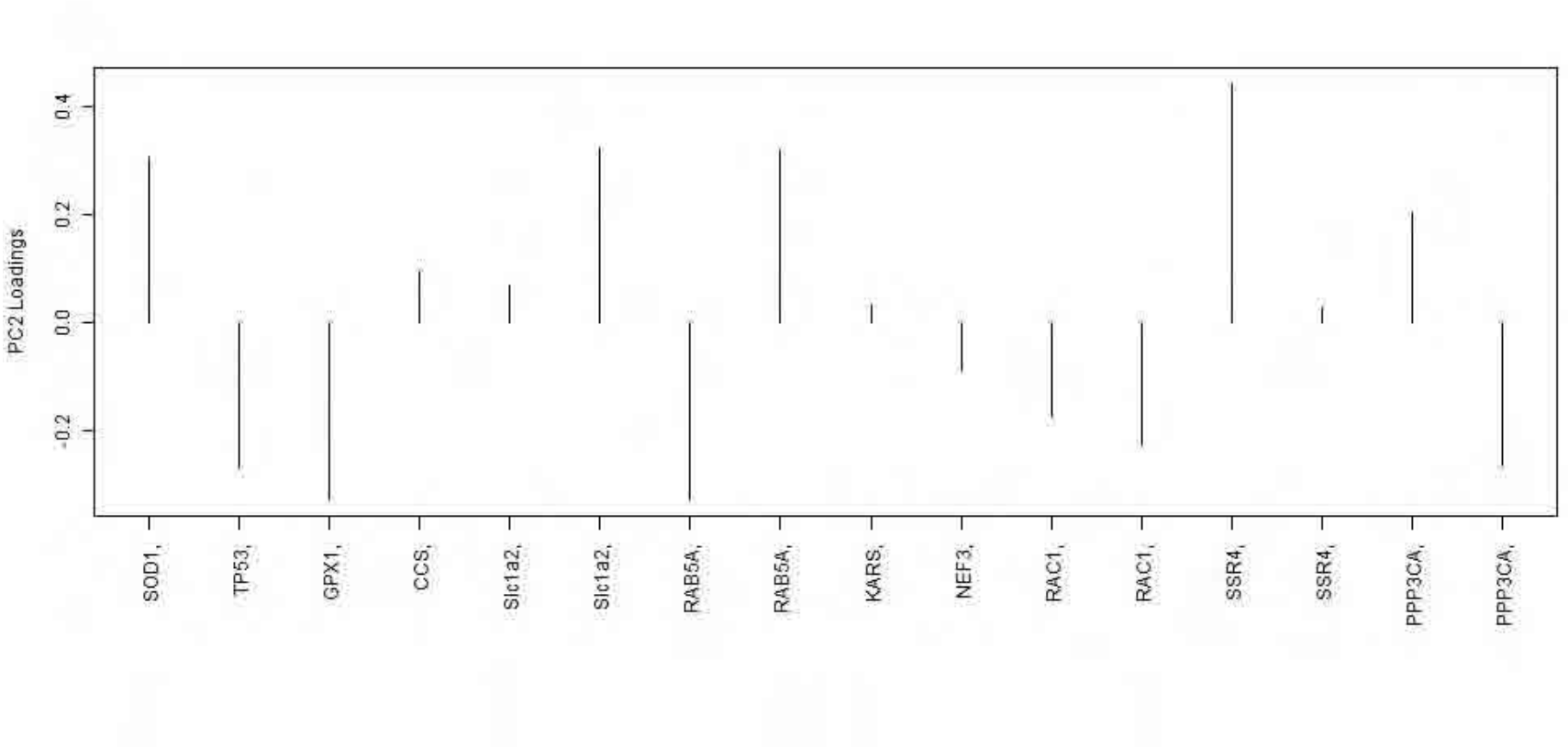


Figure 2.12: PC2 Loadings of the Genes belonging to the Amyotrophic Lateral Sclerosis (ALS).

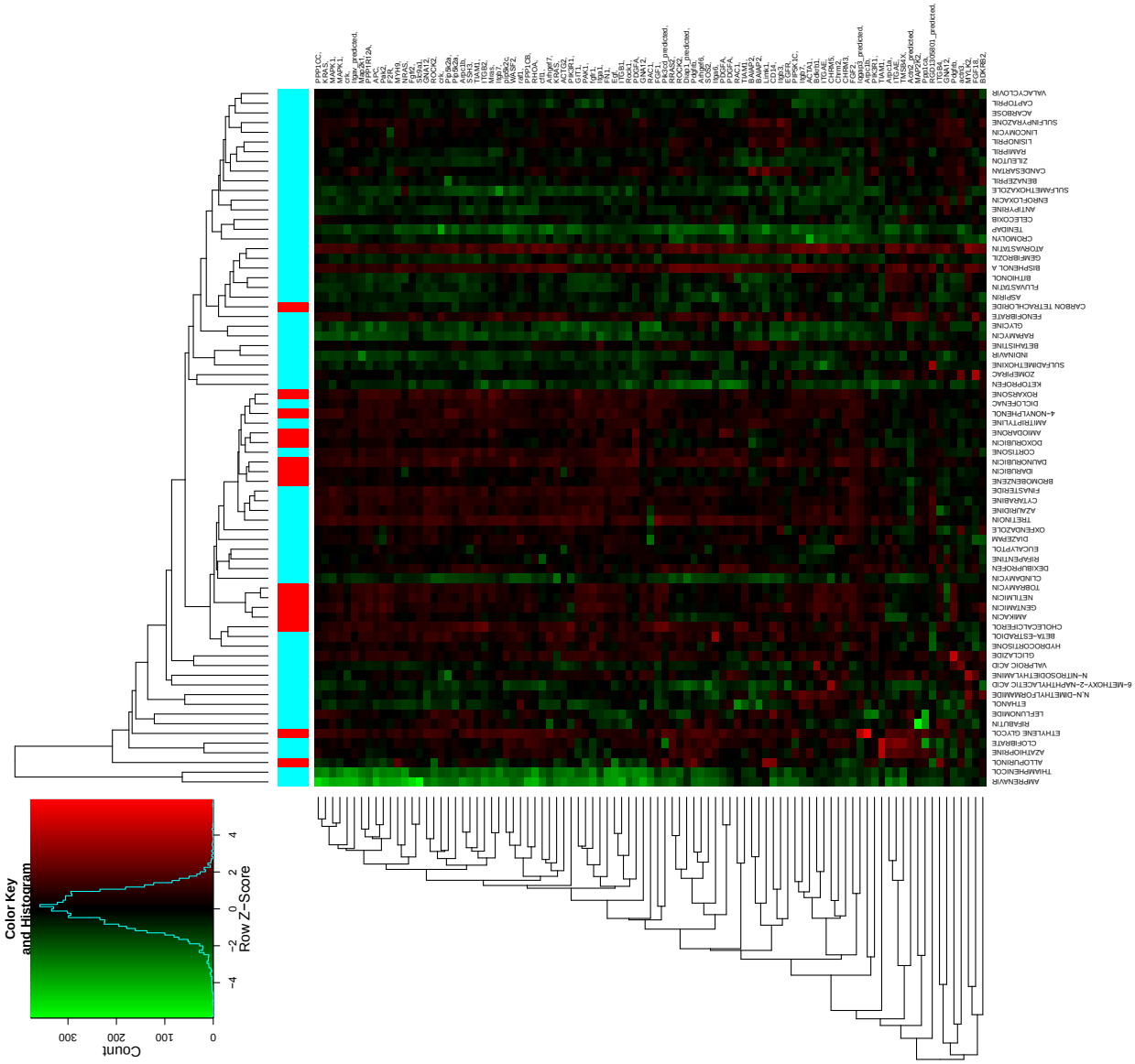


Figure 2.13: Heatmap of the Genes belonging to Regulation of Actin Cytoskeleton.

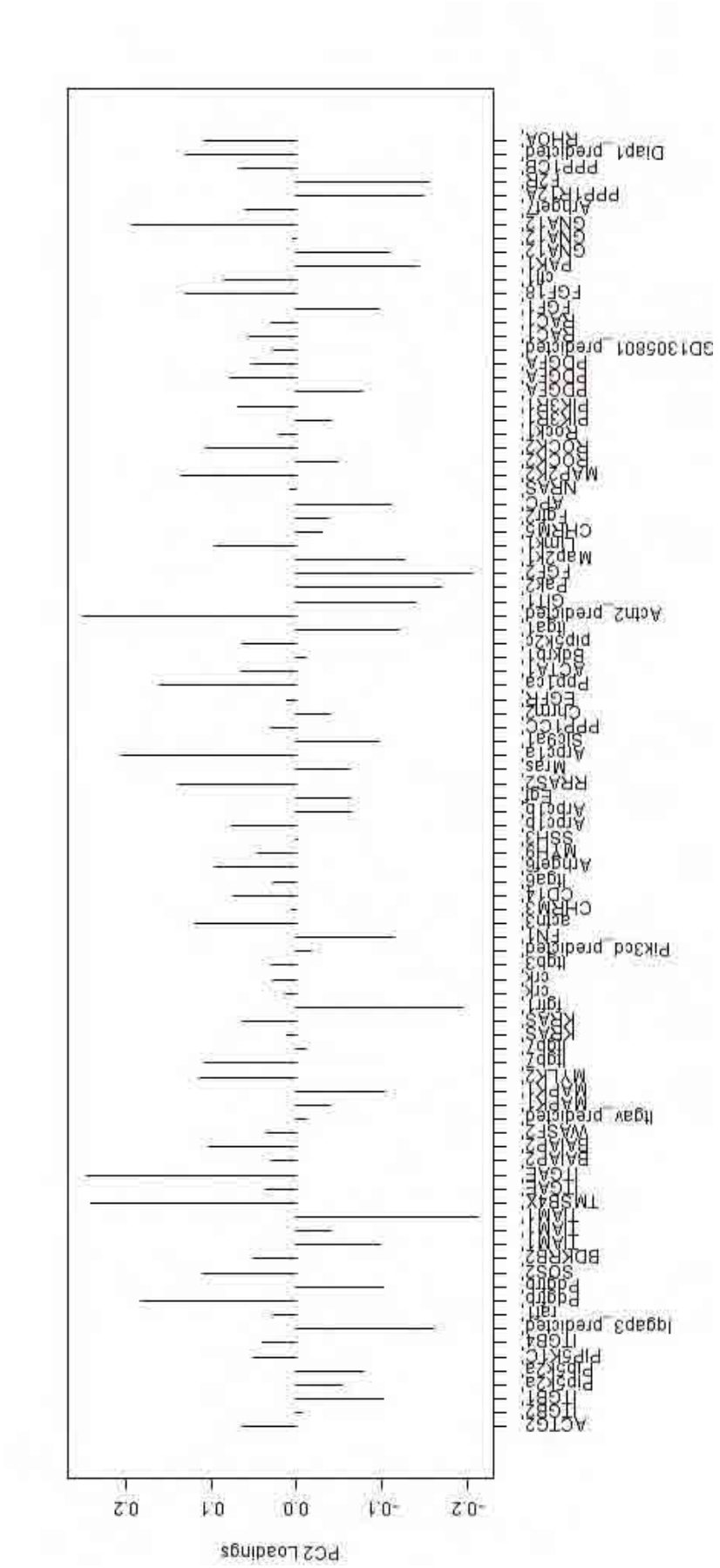


Figure 2.14: PC2 Loadings of the Genes belonging to Regulation of Actin Cytoskeleton.

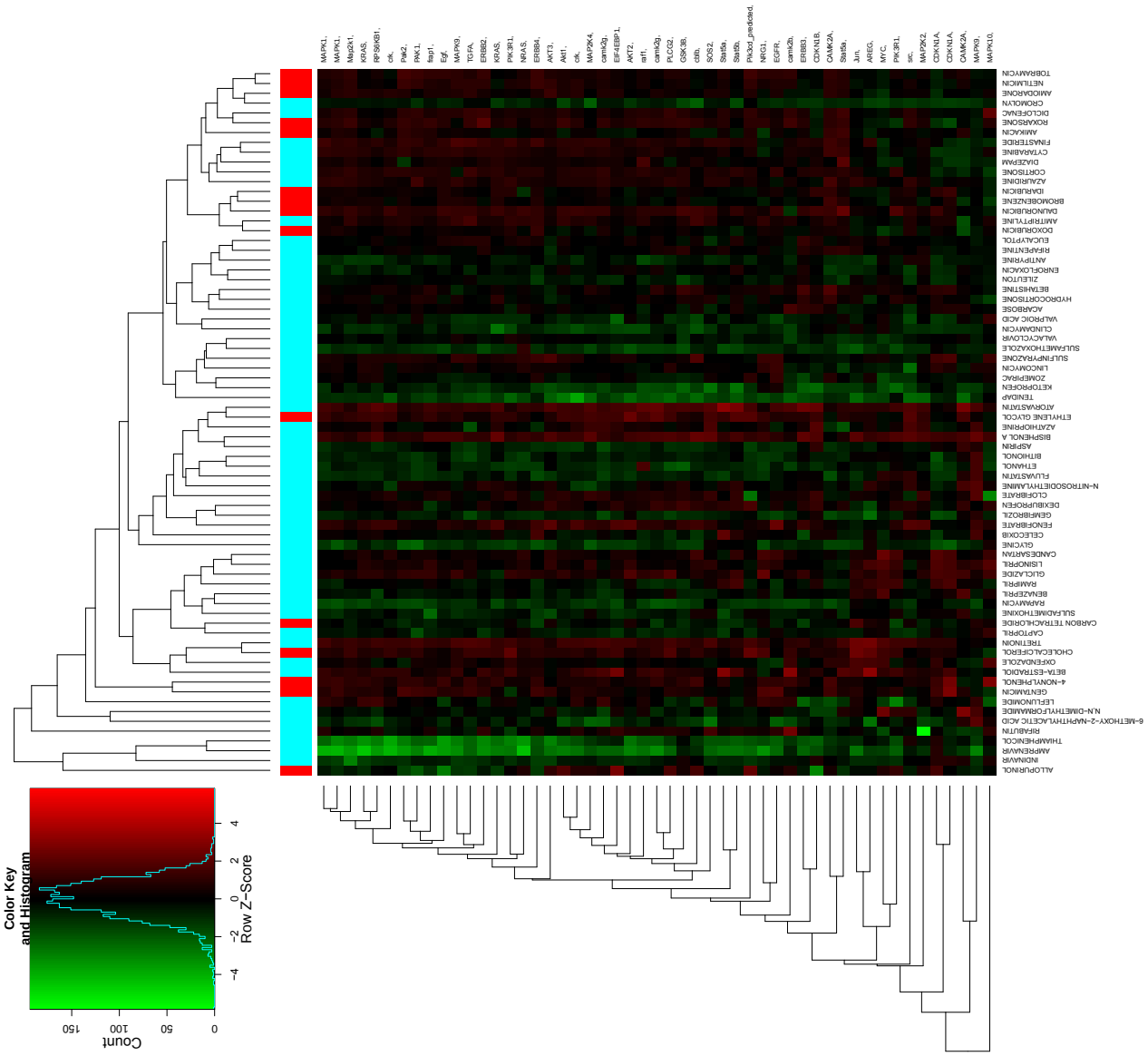


Figure 2.15: Heatmap of the Genes belonging to ErbB Signalling Pathway.

Figure 2.16: PC2 Loadings of the Genes belonging to ErbB Signalling Pathway.

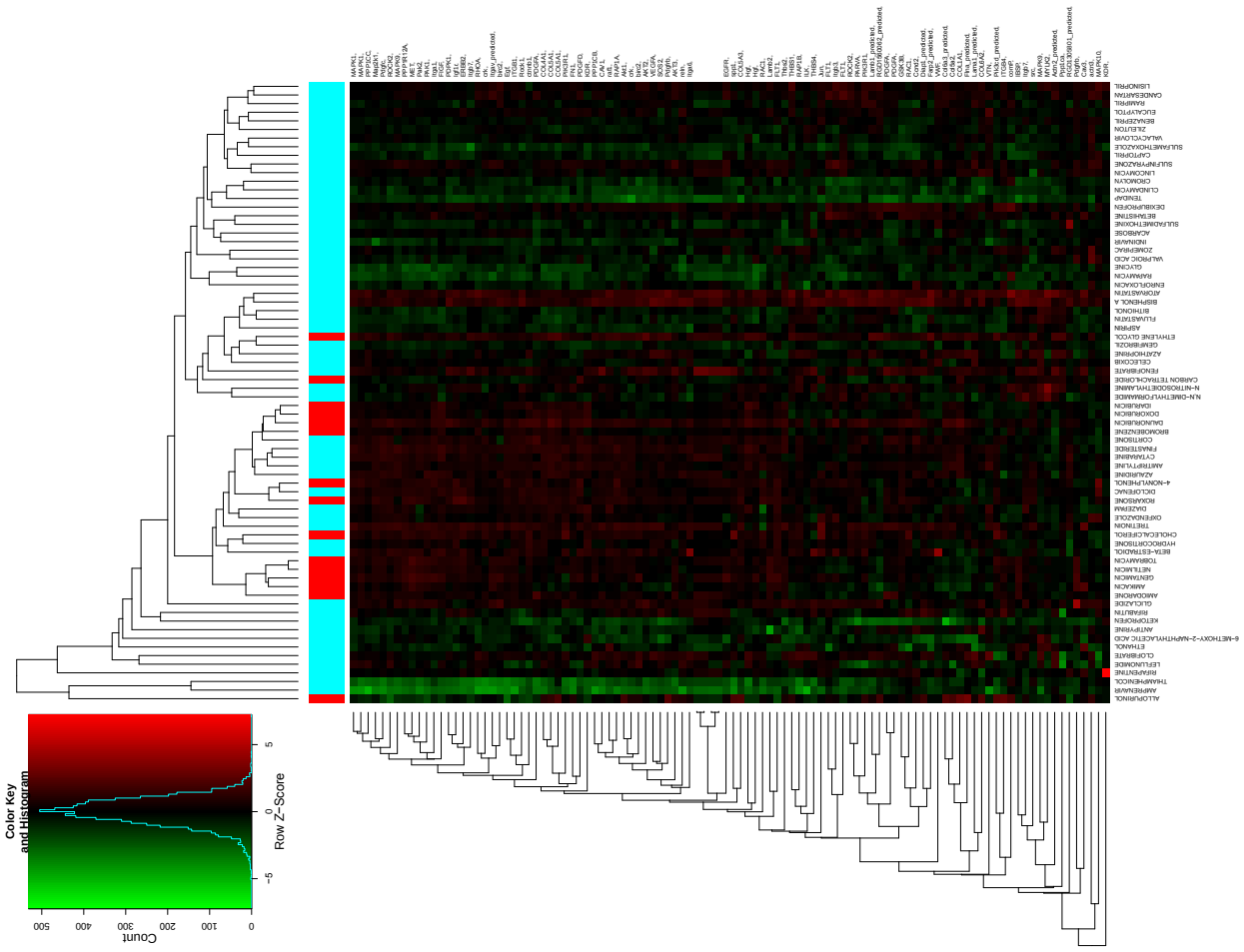


Figure 2.17: Heatmap of the Genes belonging to Focal Adhesion.

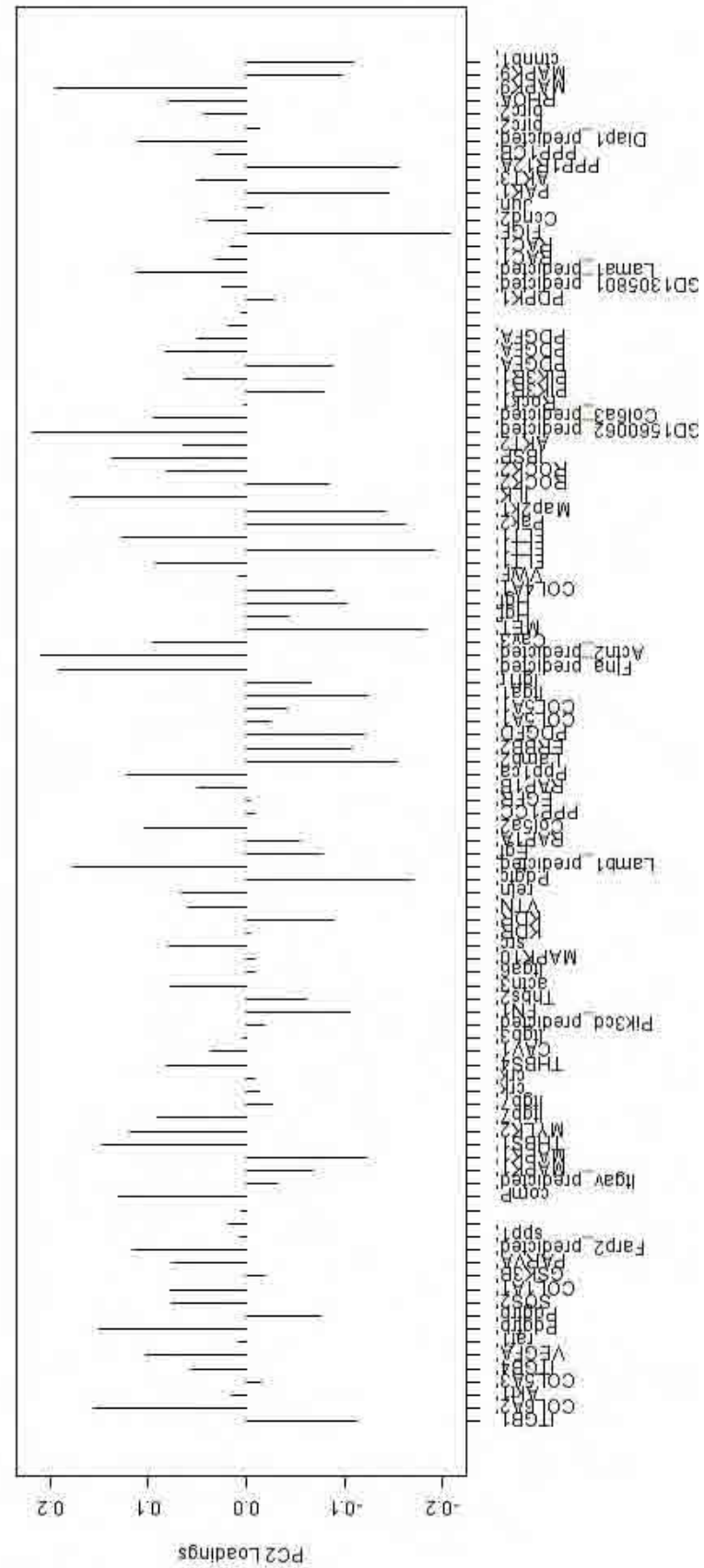


Figure 2.18: PC2 Loadings of the Genes belonging to Focal Adhesion.

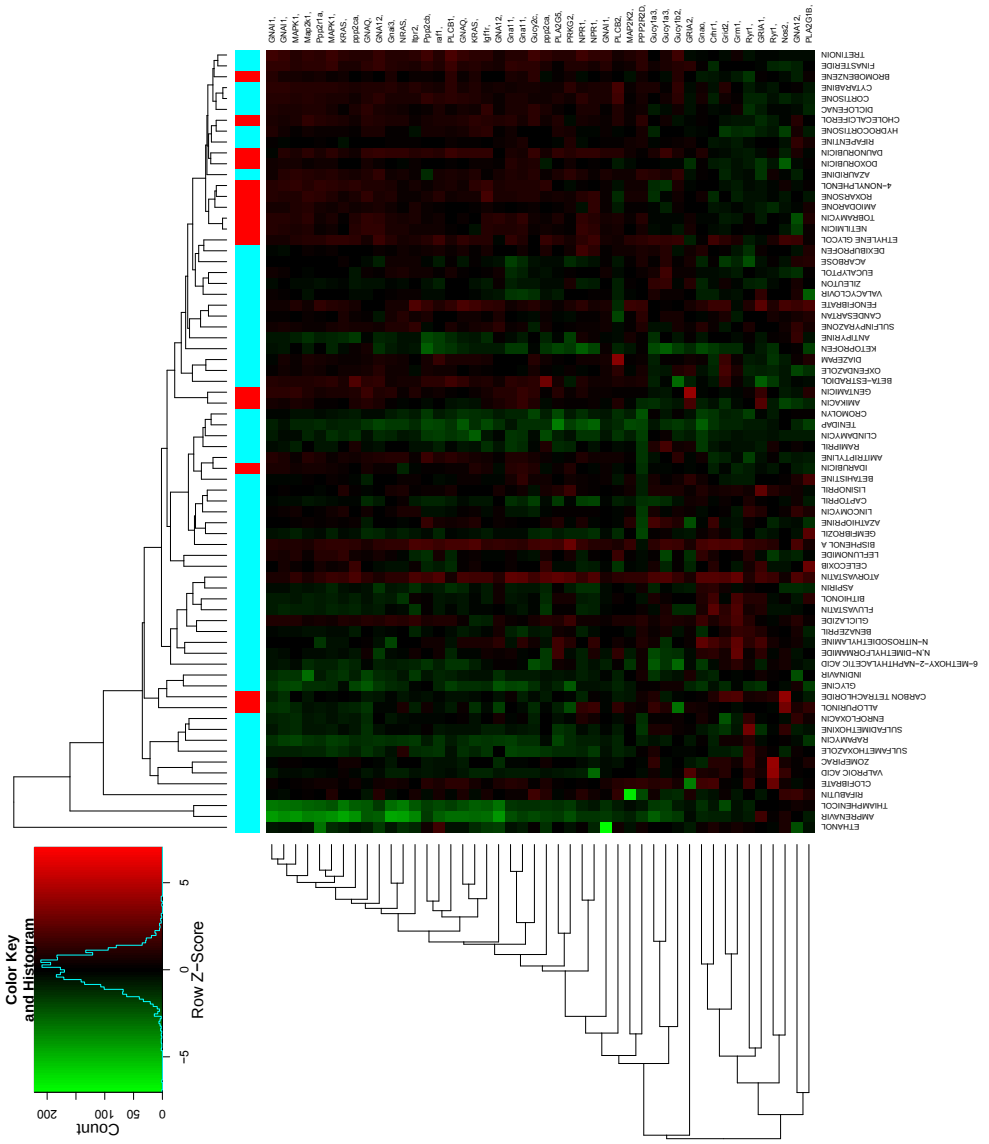


Figure 2.19: Heatmap of the Genes belonging to Long-Term Depression.

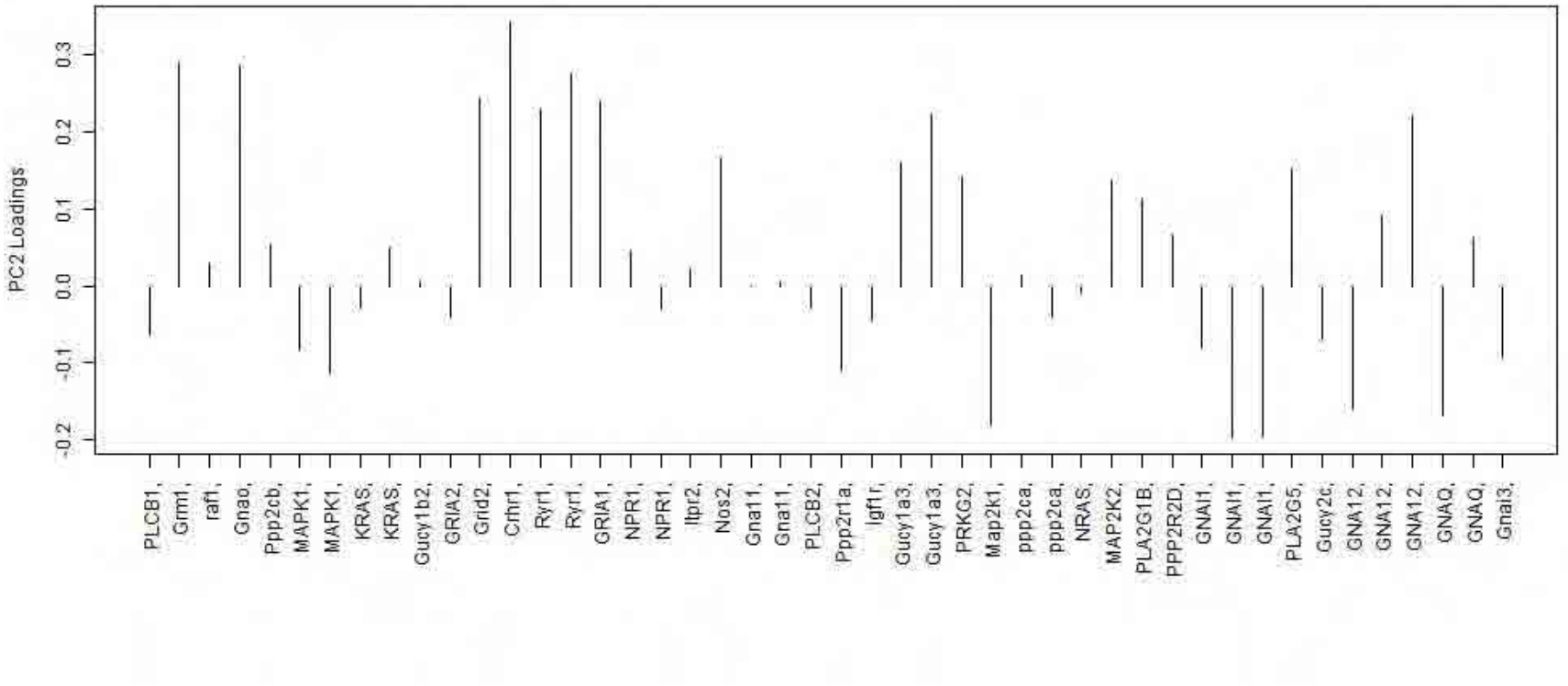


Figure 2.20: PC2 Loadings of the Genes belonging to Long-Term Depression.

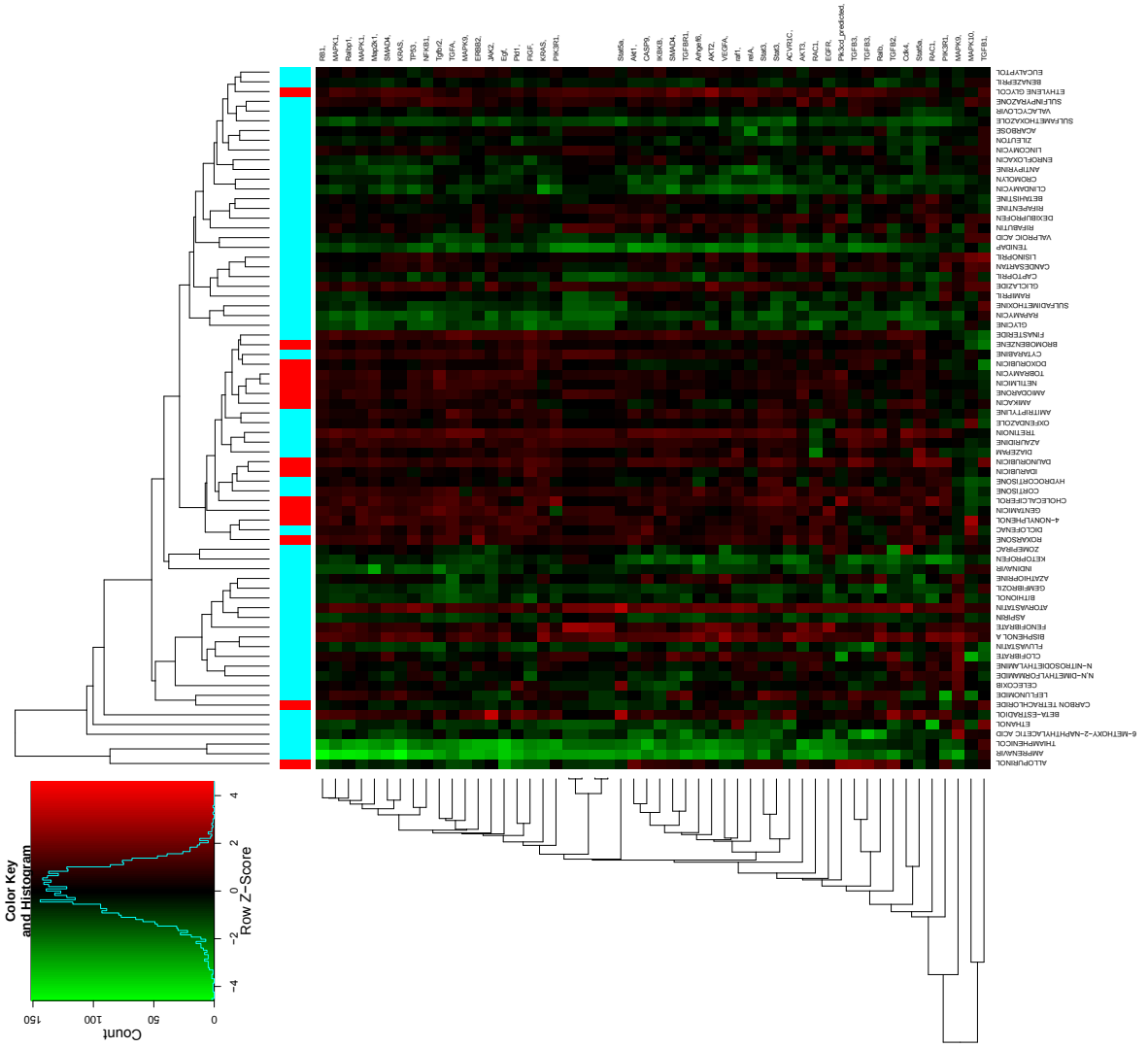


Figure 2.21: Heatmap of the Genes belonging to Pacreatic Cancer.

Figure 2.22: PC2 Loadings of the Genes belonging to Pancreatic Cancer.

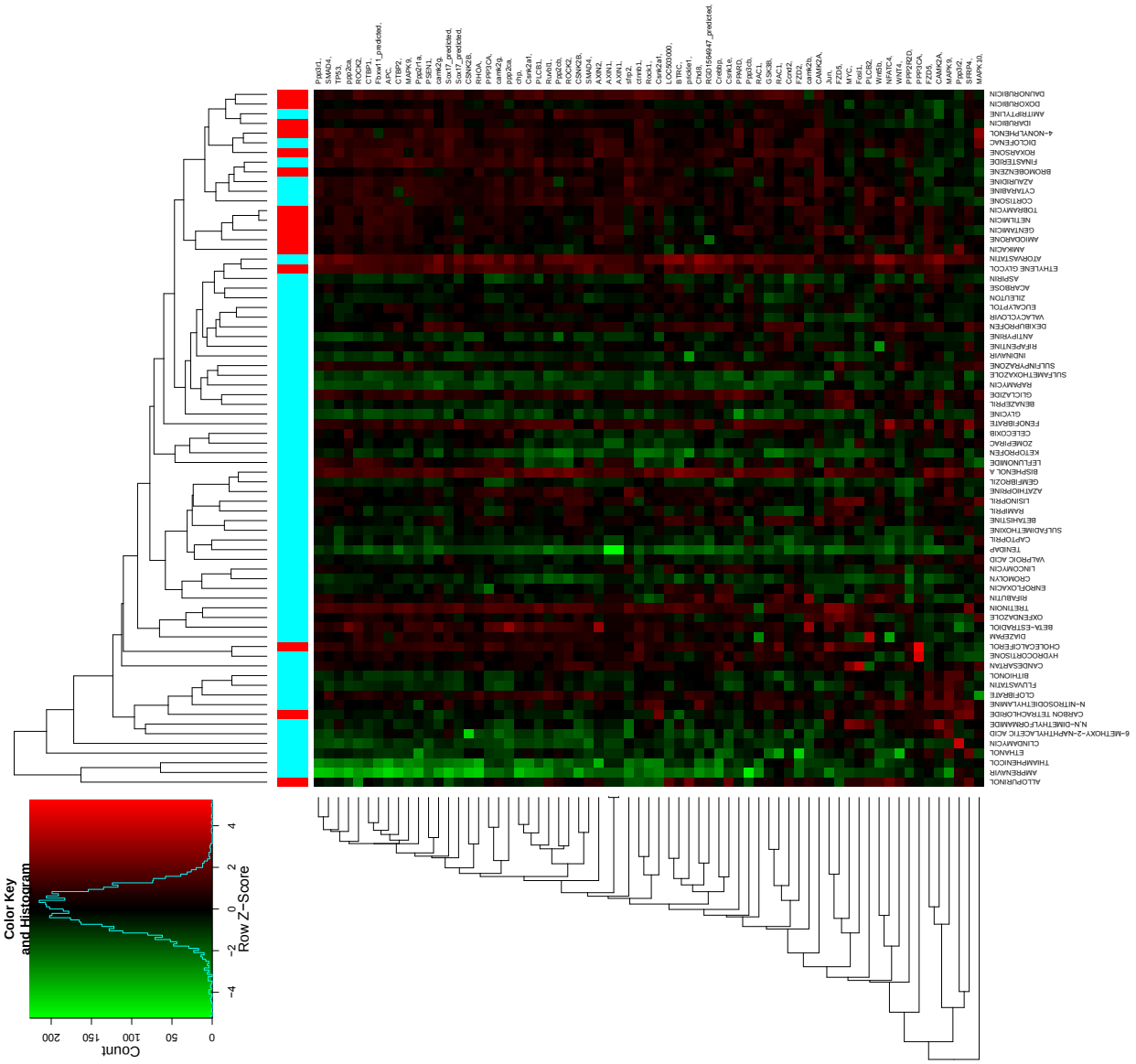


Figure 2.23: Heatmap of the Genes belonging to Wnt Signalling Pathway.

Figure 2.24: PC2 Loadings of the Genes belonging to Wnt Signalling Pathway.

2.5 Materials and Methods

2.5.1 The Dataset.

The expression profiling dataset used in this analysis was originally developed by Iconix Biosciences [28]. It is at present the largest microarray-based analysis of chemical induced transcriptional response on a mammalian system available in a public domain. In this study, rat kidneys have been profiled five days after exposure in a 28-day repeat-dose study in male Sprague-Dawley rats. The study involved 88 chemicals, 22 of which are known to induce renal tubular degeneration at the concentrations used in this study. Details of the experimental protocol are available in the original publication. Here we report a summary for clarity of the manuscript. Rats were treated daily and sacrificed on days 5 (n = 5 rats) and 28 (n = 10 rats) for kidney histopathology evaluation. Gene expression profiles were obtained on day 5 from 3 randomly chosen rats per treatment group, before the expected appearance of the lesions. Doses were chosen so as to not cause histological or clinical evidence of renal tubular degeneration after 5 days of dosing, but to cause late-onset histological evidence of tubular degeneration as expected from the literature. The negative class of this dataset was defined based on literature knowledge of treatment effects in humans and rodents. This class included 49 non-nephrotoxic compound treatments that were administered daily for 5 or 7 days (n = 3 rats). The dose was an empirically determined maximum tolerated dose in order to ensure sufficient exposure, but not to cause overt clinical toxicity. This was defined as the dose that causes approximately a 50% decrease in body weight gain relative to controls during the course of the 5-day range finding study, and without severe clinical signs of toxicity.

2.5.2 Summarizing the Transcriptional State of Kidney using Indices of Pathway Transcriptional Activity.

In order to reduce the complexity of linking chemical descriptors to the kidney transcriptional state we have computed indices of overall pathway transcriptional activity [101, 192]. These indices were computed by mapping the 7478 genes represented in the pre-processed dataset [28]

onto KEGG pathways. In choosing the number of PCs we have used the simple criteria of selecting subsequent components to explain at least 80% of the variance. In this dataset this lead to the selection of the first two components. Using this criteria we summarized the activity of pathways including more than 5 genes (148) by computing the first two principal components (PCs) which were always able to summarize up to 89% of the variance. The advantage of using PCs is that the inter-gene correlation structure is automatically incorporated into the process of dimension reduction, so this information is not lost. Computation of the PCs has been performed using the principal component function prcomp within the software programming environment R [193].

2.5.3 Comparing Indices of Pathway Activity in Response to Chemical Exposure.

In this analysis we have compared indices of pathway activity between treated samples and matched controls (Figure 2.1, Step 2) and between toxic and non-toxic chemicals (Figure 2.2, Step 2). In both cases significantly differentially modulated pathways have been identified by a combination of dimension reduction via Principal Coordinates with Hotelling s multivariate extension of the t-test. Versions of this approach were independently developed by Kong et al. [194] and Song et al. [195], and made available in the R Bioconductor package pcot2 [195]. As mentioned in the previous paragraph, one of the advantages of using PCs is the fact that the inter-gene correlation structure is incorporated into the process of dimension reduction. The Hotelling s T^2 procedure, applied to both pathway components allows this correlation to be included in the test statistic for each pathway modules [196]. In the first case the output of the T^2 Hotelling s test has been used as an index of drug effectiveness to perturb the homeostatic state (Figure 2.1, Step 2). Indices have then been used as inputs in a hierarchical clustering procedure to compare drugs perturbation profiles. In order to identify which PC most contributes to the separation a univariate t-test has been applied to the first and second PC separately and the resulting dendrograms compared (Figure 2.3). In the second case the T^2 Hotelling s test has been used to identify pathways that are differentially modulated between toxic and non-toxic

chemicals (Figure 2.2, Step 2). The p-values obtained from this test were corrected for multiple testing using the Benjamini and Hochberg method [75]. Pathways with an FDR<1% were considered differentially active between the two experimental groups.

2.5.4 Deriving Chemical Physical Features (PCFs).

PCFs were computed using the Web-based toolset e-dragon [197]. E-dragon computes 2352 chemical descriptors by integrating several publicly available methodologies. Only features that could be computed for all chemicals in the dataset were used leading to a total of 1515 chemical physical descriptors (Dataset S1).

2.5.5 Linking Chemical Features to Pathway Activity Components.

In order to link chemical descriptors to a given pathway component we used a regression model based on the combination of three chemical descriptors, including interaction components (Equation 1). More precisely, we define:

$$PC_{i,k} = a\theta_1 + b\theta_2 + c\theta_3 + d\theta_1\theta_2 + e\theta_1\theta_3 + f\theta_2\theta_3 + g + \epsilon \tag{2.1}$$

Where $PC_{i,k}$ is the principal component i of pathway k. θ_1,θ_2 and θ_3 are three different given chemical descriptors, a, b, c, d, e, f, g are model parameters and ϵ is the noise model component. In order to select an optimal subset of chemical descriptors we have used a genetic algorithm (GA) based methodology as implemented in the R package GALGO [198]. We used this random search procedure to find an optimal sub-set of variables to maximize the model R^2 value. In this application, data where split in training (2/3 of the samples) and test (1/3 of the samples) sets. The training set was used as an input to the GA procedure to search for predictive models. The fitness function was implemented as a linear model denoted in (2.1). The fitness value for model selection was set to $R^2 > 0.5$. To estimate the R^2 accurately a 5-fold cross validation procedure was used. The GA procedure was then allowed to run for 1000 simulations. Pathways for which we could identify predictive models were considered for further analysis. This resulted in the identification of 19 pathways linked to PCFs (Table 2.4). Figure 2.7 shows

examples of models found by the GA in which the predicted values using an optimized model are plotted against the observed PC values for a given pathway.

2.5.6 Creating and Visualizing a KEGG Pathway Map.

In order to visually represent the relationship between the different KEGG pathways we computed a pathway similarity matrix based on the Jaccards Index of overlap. This is defined as the ratio between the numbers of genes shared by any two pathways (intersection) divided by the number of unique genes in the two combined pathways (union). The resulting matrix was used as an input to a hierarchical clustering procedure (average linkage). The effectiveness of the clustering procedure in representing the information described by the similarity matrix has been verified using the cophenetic function correlation fit to the input overlap matrix (r=0.9).

2.5.7 Predicting Renal Tubular Degeneration from Chemical Descriptors.

Different subsets of chemical features were used to develop multivariate predictors of chemical toxicity using a classis QSAR methodology. The first subset (92 features) was defined including descriptors represented in the models predictive of pathway activity whereas the second subset included all variables not selected in the predictive models and that were uncorrelated, an absolute pearson coefficient of less than 0.5, to PCFs from the first group (210 features). Models were developed using a maximal likelihood discriminant function coupled to a genetic algorithm for variable selection using default settings [198]. A forward selection approach was used to identify the single smallest model, with the least number of descriptors and with the highest classification accuracy [198]. Classification accuracy was estimated using a k-folds cross-validation procedure. Interpretation of the models has been performed with the help of PCA.

	Pathway	R^2	R^2 average	Feature Type			Feature 1	Feature 2	Feature 3
rno04810	Regulation of actin cytoskeleton	0.53	0.53	GETAWAY descriptors	GETAWAY descriptors	ET-state Indices	H3p	R7u	SaasC Count
rno05222	Small cell lung cancer	0.57	0.52	RDF descriptors	GSFRAG Descriptor	GETAWAY descriptors	RDF020p	p2c6A	R7p
rno04310	Wnt signaling pathway	0.57	0.53	GSFRAG Descriptor	Information indices	GSFRAG Descriptor	p5CD	SIC3	c6AC
rno04540	Gap junction	0.57	0.53	Geometrical descriptors	ET-state Indices	3D-MoRSE descriptors	DISPp	SsOH(phen)	Mor05e
rno04720	Long-term potentiation	0.57	0.53	GETAWAY descriptors	Information indices	GSFRAG Descriptor	R7v	IC3	p2c6A
rno04730	Long-term depression	0.54	0.53	RDF descriptors	ET-state Indices	Geometrical descriptors	RDF025m	Se1C3O1a	DISPp
rno04912	GnRH signaling pathway	0.55	0.52	GETAWAY descriptors	WHIM descriptors	WHIM descriptors	H3e	Am	G1s
rno04916	Melanogenesis	0.53	0.52	3D-MoRSE descriptors	Geometrical descriptors	ET-state Indices	Mor05p	DISPp	Se1C3O1a
rno00860	Porphyrin and chlorophyll metabolism	0.53	0.52	ET-state Indices	GSFRAG Descriptor	WHIM descriptors	Se1C3N1s	p3	E1e
rno00010	Glycolysis / Gluconeogenesis	0.53	0.52	GSFRAG Descriptor	GSFRAG Descriptor	WHIM descriptors	p1c6AB	p1c5ABC	G1m
rno01430	Cell Communication	0.57	0.52	Geometrical descriptors	ET-state Indices	GETAWAY descriptors	DISPp	Se1C3O1a	R7u
rno04920	Adipocytokine signaling pathway	0.53	0.51	GSFRAG Descriptor	GSFRAG Descriptor	2D autocorrelations	p4BC	p1c5AC	GATS8p
rno04012	ErbB signaling pathway	0.57	0.53	RDF descriptors	ET-state Indices	Geometrical descriptors	RDF065p	Se1C3O1a	DISPp
rno05215	Prostate cancer	0.53	0.52	RDF descriptors	GSFRAG Descriptor	GETAWAY descriptors	RDF020p	p1c6A	R7v
rno04664	Fc epsilon RI signaling pathway	0.55	0.53	3D-MoRSE descriptors	GSFRAG Descriptor	2D autocorrelations	Mor32u	p2	GATS5m
rno05214	Glioma	0.55	0.52	3D-MoRSE descriptors	Molecular properties	Topological charge indices	Mor15p	BLTD48	JGT
rno05218	Melanoma	0.59	0.51	ET-state Indices	3D-MoRSE descriptors	Geometrical descriptors	Se1C3O1a	Mor05e	DISPp
rno04930	Type II diabetes mellitus	0.57	0.53	WHIM descriptors	GSFRAG Descriptor	GSFRAG Descriptor	G1m	p1c5ABC	p1c6AC
rno00480	Glutathione metabolism	0.54	0.51	Edge adjacency indices	Information indices	3D-MoRSE descriptors	EEig15d	IC0	Mor13e

Table 2.4: **Pathways Associated to PCFs.** This table lists the 19 pathways whose activity could be predicted by combinations of PCFs. For each pathway the average R^2 value given the model population, the model with the highest R^2 value and the descriptor group features responsible for the correlation are shown.

Chapter 3

A Functional Module Based Approach to Chemical Class Prediction in *Daphnia magna*

3.1 Abstract

Daphnia magna is an accepted model organism in toxicity testing by several international agencies. Several molecular biomarkers have been discovered using hypothesis driven and omics approaches. However, statistically robust prediction systems that allow the identification of chemical contaminants from the molecular response to exposure still need to be developed. The research described in this chapter addresses this issue using a combination of transcriptomics and advanced machine learning approaches.

The models we developed successfully identifies the class of a given toxicant discriminating between endocrine disruptors, metals and a group of unrelated chemicals. We also show that models based on indices of whole pathway transcriptional activity can achieve more accurate models and facilitate biological interpretability.

3.2 Introduction

Assessing the impact of environmental exposures to chemicals released by human activity is of paramount importance. Biota within the freshwater environment are at risk from a number of pressures, including toxic substances, e.g. pesticide run-off or industrial spills, as well as from excess nutrients, e.g. from sewage, which can be released from point or diffuse sources. In response to this problem the European Union introduced the Water Framework Directive (WFD) in 2000 [199], which represents the central legislation on water quality and commits European Union member states to achieve good status of all water bodies (up to 1 km from shore) by 2015. Importantly, good status requires an assessment of both the ecological health of water bodies as well as chemical monitoring and comparison with chemical standards. Toxicity Identification and Evaluation (TIE) methods have been widely used to assess water effluents by identifying key toxicants through a series of acute toxicity tests [200]. In this context, *D. magna* has become an important biosensor in both ecology and toxicology due to their wide geographic distribution, central role in freshwater food webs, ability to adapt to a range of habitats [34] and sensitivity to anthropogenic chemicals [35]. *D. magna* has been included into toxicity testing protocols by the U.S. Environmental Protection Agency and the international Organisation of Economic Cooperation and Development (OECD) [34]. Conventionally, water quality is assessed using standard toxicity tests by exposing Daphnia neonates ($age < 24h$) [201]. The effect of the water sample is quantified by its ability to immobilize the juvenile crustaceans. Although widely used, this test is not very sensitive and importantly, does not identify the chemicals in the contaminated water samples. In order to address this challenge a number of laboratories [28, 202–205], including ours [46, 169, 206] have used a combination of omics technologies (see introduction for an overview) to perform genome-wide unbiased biomarker screenings. Several of these studies describe *D. magna* response to toxicant exposure [41, 207, 208]. For example, Poynton et al [41] used a cDNA microarray to identify biomarkers of sub-lethal exposure to copper, cadmium and zinc. This study identified genes regulated by specific metals included two metallothioneins (MT), which are already known biomarkers of metal exposure. Furthermore their work was

consistent with known mechanisms of metal toxicity and identified novel putative modes of action including zinc inhibition of chitinase activity [41]. Independently to these developments, Shaw et al also identified a set of three genes coding for the metal detoxification protein MT (in *Daphnia pulex*) whose gene structures and predicted translated sequences did not match any previously identified MTs [207]. This work showed that omics approached could lead to the identification of novel protein family memebrs by integrating expression and sequence information. Heckmann et al [208] studied the effect of ibuprofen on *D. magna* using transcriptomic and phenotypic measurements. Their results suggested a highly similar mode of action between vertebrates and invertebrates hence supporting the view that a non-model crustacean could be informative of drug toxicity effects in higher vertebrates. Although encouraging, these studies used relatively simple bioinformatics methods and did not provide truly predictive models of toxicity response. Moreover, they also did not provide robust molecular signatures that could be used to identify the chemical contaminants or at least the chemical class of the toxicant. Therefore these remain unsolved challenges in the field of ecotoxicology. In this chapter we describe the development of a predictive toxicology approach for chemical class prediction in *D. magna*. Here we demonstrate for the first time, that it is possible to predict chemical class from the transcriptional response of adult Daphnids, following exposure to sub-lethal concentrations of a given toxicant. We have also shown that predictive models based on whole-pathway activity indices provide more biologically interpretable results when compared to gene-level models. Therefore our analysis sets the scene for a broader application of computational methodologies for the development of predictive ecotoxicology.

3.3 Results

3.3.1 Analysis Overview

The aim of this study was to develop robust molecular signatures predictive of chemical class. The analysis we performed is based on a dataset representing the transcriptional response of adult Daphnids to sub-lethal exposures of 36 chemicals (Table 3.1). The analysis strategy that we employed utilizes a variable selection approach to identify multi-feature markers predictive

of the chemical class of a given toxicant. In the most comprehensive models we focused on the discrimination of chemicals in three classes. These are: 1) metals, 2) endocrine disruptors and 3) a collection of industrial chemicals, which could not be classified in either groups defined above (Table 3.1). Figure 3.1 represents the analysis strategy in a schematic format. Details for each step of the approach are reported below. Microarray data are first processed to remove

Class	Compounds
Metals	Cadmium, Nickel, Copper, Selenium, Zinc, Manganese, Arsenic, Silver, Chromium
Endocrine Disruptors	Pyriproxyfen, Ponasterone A, Methyl farnesoate, Toxaphene, Beta-estradiol, Aroclor 1242, 20-hydroxyecdysone, Methoxyclor
Industry Relevant Compounds	Methyl tert-butyl ether, Chloroform, Acrylonitrile, Bis(2-ethylhexyl)phthalate, Trichloroethylene, 2-chloroethyl vinyl ether, Atrazine, Toluene, Phenol, Phenanthrene, Dichlorobenzene, Beta-benzen hexachloride, Permethrin, Bifenthrin, λ-cyhalothrin, Diazinon, Parathion, Chloropyrifos

Table 3.1: **Chemical classification of the 36 chemicals in the dataset.** Exposures were performed at sub-lethal concentrations ($\frac{1}{10}$ of the experimentally derived LC_{50} value).

genes expressed below detection level and to eliminate duplicate probes (Figure 3.1, Step 1). Genes with a known biological function that could be mapped to KEGG pathways were then used to generate indices of pathway activity based on the first three Principal Components [209] (Figure 3.1, Step 2). These two datasets (gene expression profiles and indices of pathway activity as indicated in Figure 3.1) are the inputs of the statistical modelling procedure, which uses a combination of a genetic algorithm-based variable selection procedure and a random forest classifier [163, 198] (Figure 3.1, Step 3). The outcome of this procedure is a collection of predictive models of comparable accuracy, representing alternative solutions to the classification problem. These are condensed in two representative models, based on gene expression profiles or indices of pathway activity. At this stage, statistical models are built from a subset of the

data representing all chemicals and their accuracy is computed on a subset of data representing an independent exposure experiment. For this reason, the accuracy of the models, which in this case is larger than 95%, reflects the robustness of the classifier to biological and technical variation. The ability of these models to predict chemicals that have not been used in the model training procedure is then tested in a separate procedure (Figure 3.1, Step 4) by a leave one out cross-validation approach (LOOCV) where one chemical in every class is taken out before the model is trained (see materials and methods section 3.5.5 for details of the methodology). At this stage the prediction accuracy of gene level and pathway level models can be compared. In order to facilitate biological interpretation genes represented in the pathways identified in the models can be used as inputs to the Ingenuity Pathway Analysis software (IPA, Ingenuity Systems®, `www.ingenuity.com`) that identify potential gene to gene connections supported by mammalian literature (Figure 3.1, Step 6). This last step is of course limited by the relatively small number of genes that can be mapped between these two distant species. Therefore results have been considered indicative and used for generating hypotheses.

3.3.2 Gene-level Molecular Signatures can Discriminate Distinct Chemical Classes

In order to test whether it was possible to identify gene expression profiles predictive of chemical class we first attempted to discriminate between metals and non-metals (Table 3.2). Figure 3.2A shows a graphical representation of the increase in accuracy upon addition of the most representative genes in the model. The three most represented genes in the model population can already correctly classify 97% of the samples whereas the most predictive model (100% of accuracy) require the 14 most represented genes. These genes can also separate these two sample classes in a principle component (PC) plot (Figure 3.2B). Unfortunately, we could only identify cross species homologues for only 4 genes, within all available species in GeneBank. Of these, only one was functionally annotated (glycerolipid metabolism enzyme (brummer CG5295-PA)). The success in separating between metals and non-metals and the documented ability of *D. magna* to respond to endocrine disruptors [210] lead us to investi-

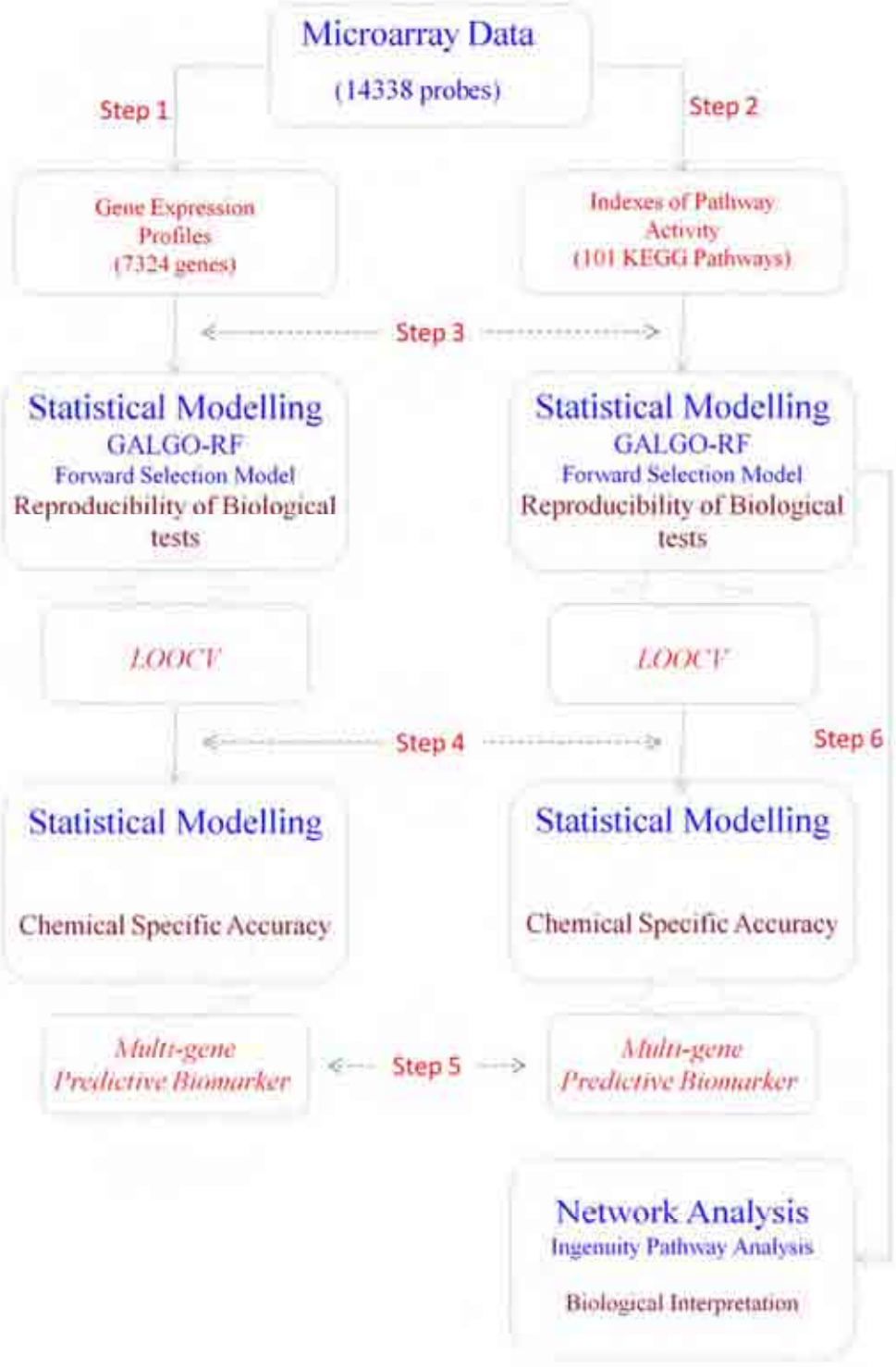


Figure 3.1: **Overview of the Analysis Strategy.**

gate whether we could further differentiate between metals and, within the non-metals class, endocrine disruptors and the group of remaining chemicals. Our procedure successfully identified a representative model based on a set of 15 genes, which had a classification accuracy of 99% (Figure 3.3A and 3.3B). Of these, only 2 genes were functionally annotated (sarcosine dehydrogenase and collagen) (Figure 3.3A).

Class	Used Categories
Metals	Metal Category
Non-metals	Endocrine Disruptors + Industry Relevant Compounds
Endocrine Disruptor	Endocrine Disruptors
Remaining Chemicals	Industry Relevant Compounds

Table 3.2: **Classification used to Predict Chemical Class.** This table provides information on the chemical classes that were used for the classification analysis. For further discrimination of the classes see Table 3.1.

3.3.3 A Pathway-level Approach to Predicting Chemical Exposure

To address the challenge of developing more biologically interpretable predictors, we converted our dataset into 303 indices of pathway activity representing 101 KEGG pathways and used this as an input to our statistical modelling procedure. Similarly to the gene-level models, we first attempted to classify between metals and non-metals. We identified a representative model, based on the activity of five pathways, which was 99.8% accurate (Figure 3.4A). These were oxidative phosphorylation, alpha-Linolenic acid metabolism, synthesis and degradation of ketone bodies, lysosome and TCA cycle. We could also develop a model discriminating between metals, endocrine disruptors and a class including a relatively heterogeneous mixture of non-metals and non-endocrine disruptor chemicals. The most predictive representative model was based on a set of 15 pathway indices (representing 14 different pathways) and was 98.5% accurate (Figure 3.5). Interestingly the top 4 pathways identified in the set of 15 are sufficient to develop a model that is 98% predictive. These are alpha-linolenic acid metabolism, ECM-receptor interactions, pyruvate metabolism and lysosome. Two of these pathways, alpha-linolenic acid metabolism and lysosme, were also included in the model discriminating between metals and non-metals (Figure 3.4). We can therefore hypothesize that the remaining two pathways have

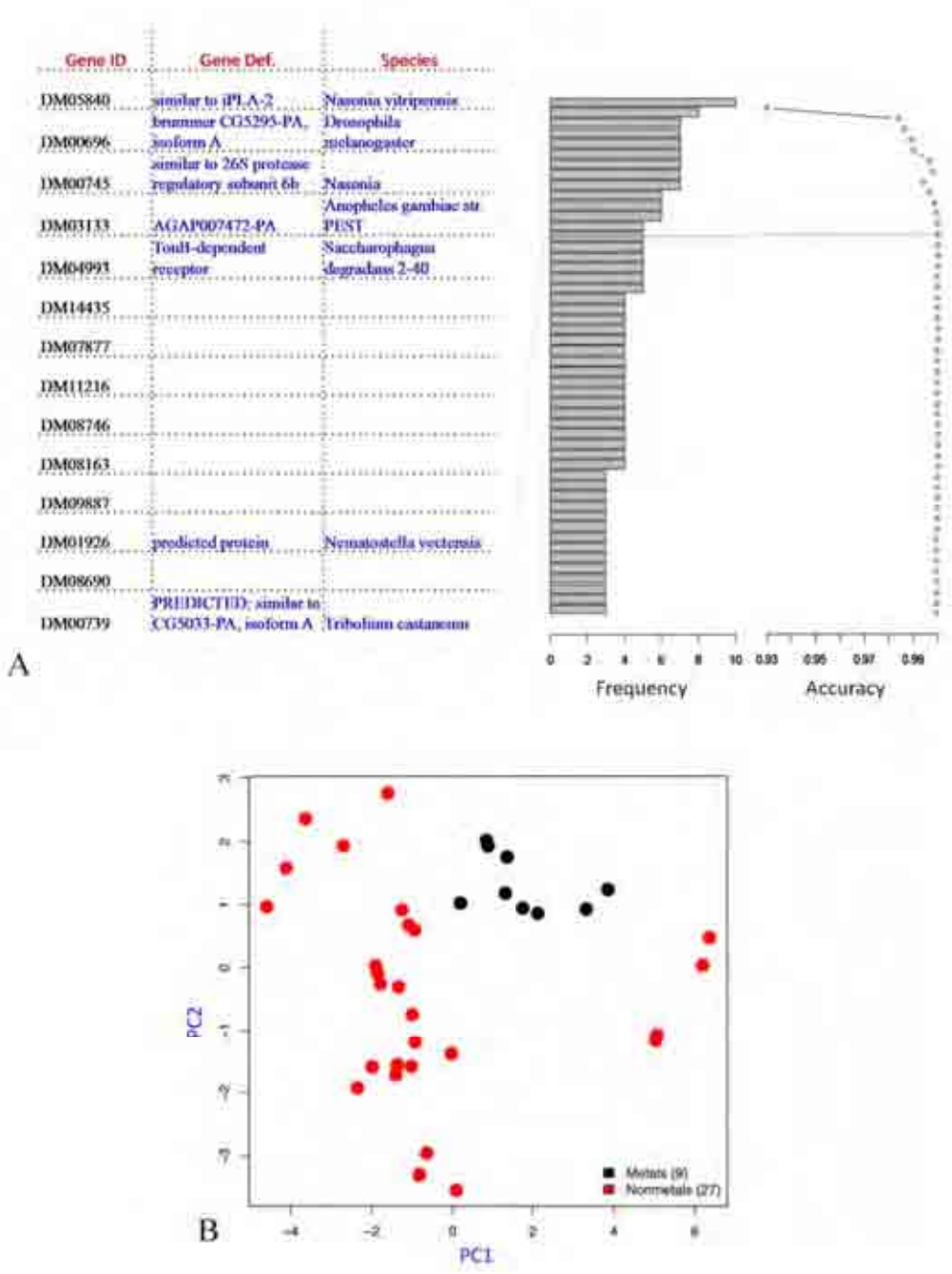


Figure 3.2: **Gene Level Model Discriminating between Metals and Non-metals.** This figure shows the result of our classification approach. Panel A represents the results from the forward selection strategy. Genes were sorted by their selection frequency (A_2). Accuracy (A_3) was calculated by incrementally adding each of the genes (top to bottom). The model was chosen to have the highest accuracy with the smallest number of components (A_3 dotted line).This resulted in 14 genes reaching an accuracy of 100%. Annotation for these genes is shown in the table (A_1). Unfortunately, most of the annotation is missing or linked to hypothetical proteins. Panel B represents these genes in a PC space. Samples marked in red and black are non-metals and metals respectively. The separation of these two classes is clearly visible.

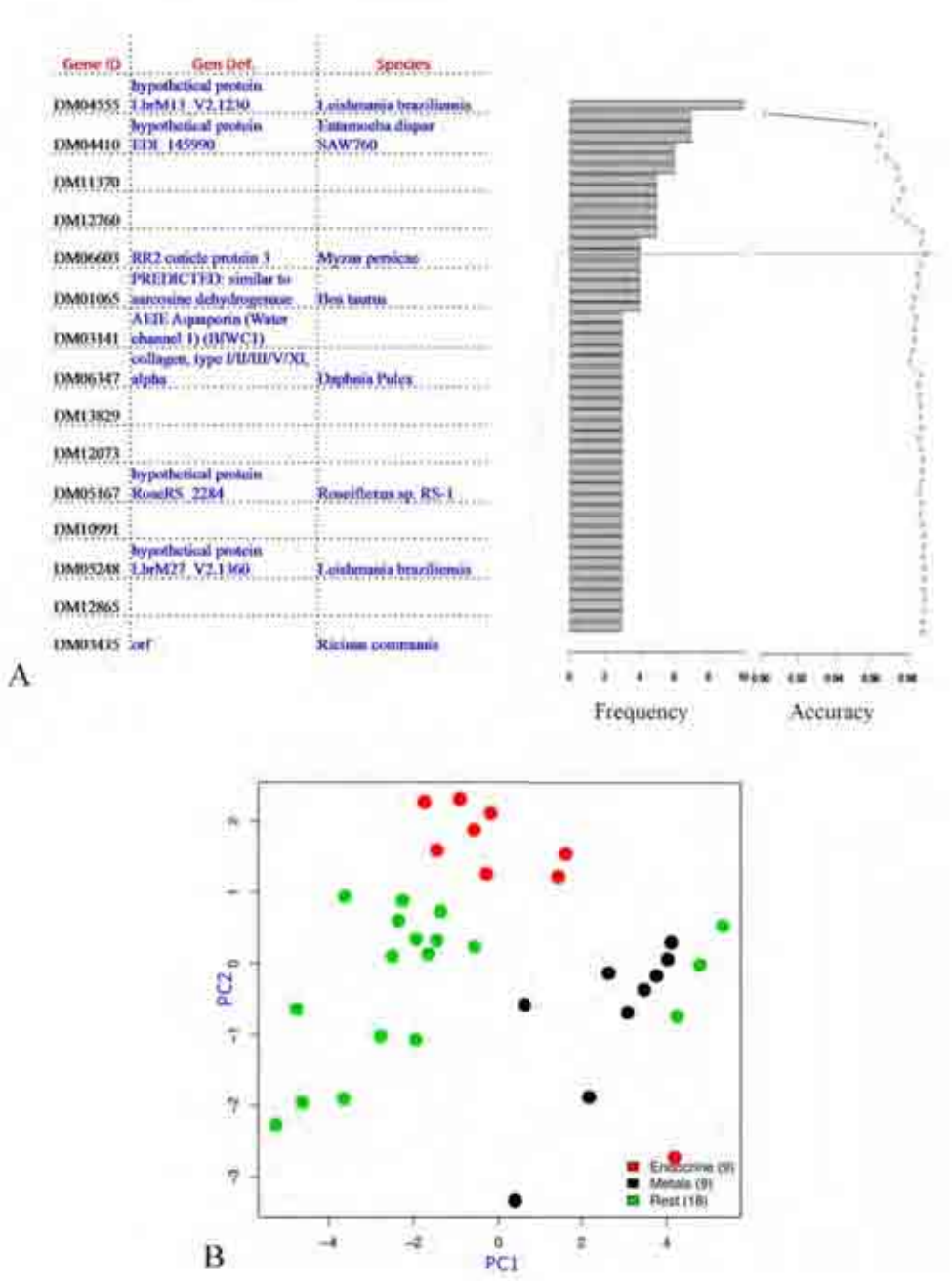


Figure 3.3: **Gene Level Model Discriminating between Metals, Endocrine Disruptors and Remaining Industry Relevant Chemicals.** Panel A represents the results from the forward selection strategy. Genes were sorted by their selection frequency (A_2). Accuracy (A_3) was calculated by incrementally adding each of the genes (top to bottom). The model was chosen to have the highest accuracy with the smallest number of components (A_3 dotted line). This resulted in 15 genes reaching an accuracy of 99%. Annotation for these genes is shown in the table (A_1). Unfortunately, most of the annotation is missing or linked to hypothetical proteins. Panel B represents these genes in a PC space. Samples marked in black, red and green are metals, endocrine and remaining chemicals respectively. Here the separation between the classes is visible with one endocrine and three industrial chemicals clustering close to the metals group.

the information required for specifying the identity of endocrine disruptors. Further work will be needed to clarify this important point.

3.3.4 Estimating Chemical-Specific Classification Accuracy

The models we have developed are effective in predicting the class of a chemical from an independent exposure experiment. Here we assessed the ability of each model to correctly classify each individual chemical in the dataset. The gene-level models discriminating between metals and non-metals were able to predict all chemicals in the correct class (Figure 3.6A, 100% accuracy). The model discriminating between metals, endocrine disruptors and other chemicals was 80% accurate failing to correctly classify copper, nickel, 20-hydroxyecdsone, pyripoxyfen and bifenthrin (Figure 3.6B). Figure 3.6C and 3.6D summarize the results of the modelling performed on the pathway-level dataset. All metals are classified correctly with the exception of manganese (Figure 3.6C). Non-metals were classified with a 95% accuracy. More specifically only 20-hydroxyecdsone was misclassified. For the classification of metals, endocrine disruptors and remaining we see a similar trend as in the gene-level model (Figure 3.6D). The endocrine disruptors have the lowest classification accuracy (77%) as compared to the other two classes (91% and 97% for metals and remaining chemicals respectively). Interestingly metals and non endocrine disruptors are classified more accurately by the pathway level model as compared to the gene-level model (86% vs. 91% for metals and 95% vs. 97% for remaining chemicals in the gene and pathway-level respectively; Figure 3.6). Similarly to the other pathway level model, manganese and 20-hydroxyecdsone are highly misclassified. In addition to these two compounds methylfarnesoate, toxaphene and acrylonitrile are only classified correctly 50% of the time. It should be noted that in 3 out of these 4 classification runs, 20-hydroxyecdsone was misclassified every time, showing that the response of *D. magna* to this chemical is highly different to other chemicals in its class.

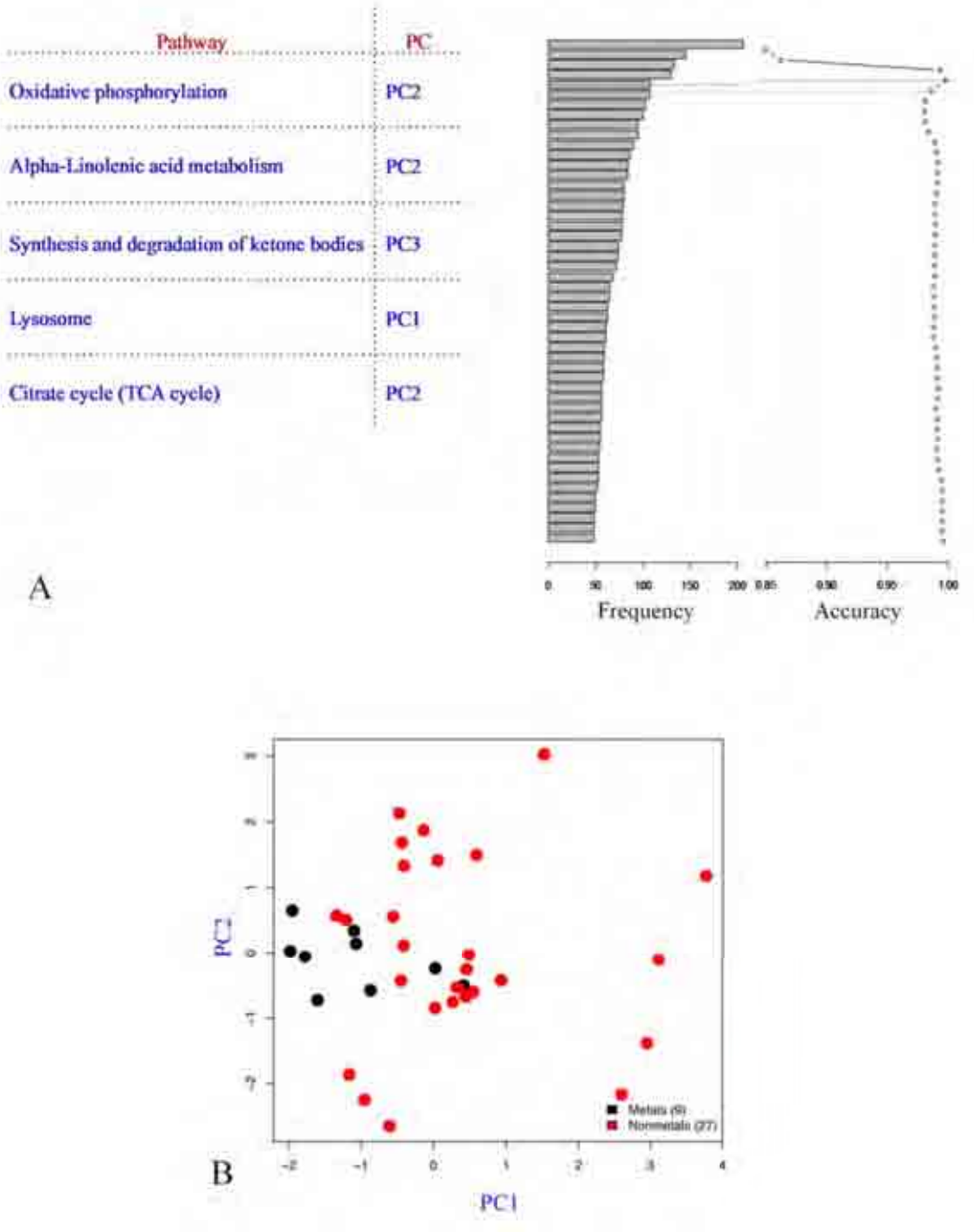


Figure 3.4: **Pathway level Model Discriminating between Metals and Non-metals.** Panel A represents the results from the forward selection strategy. Pathways are first sorted by their selection frequency (A_2). Accuracy (A_3) is then evaluated by incrementally adding the most frequent pathway components. The dotted line in A_3 shows where the model reached 99% accuracy. The 5 pathways from the resulting model are shown in A_1. Panel B shows a PCA of the indices of pathway activity. Separation between metals and non-metals is less evident than in the gene-level model.

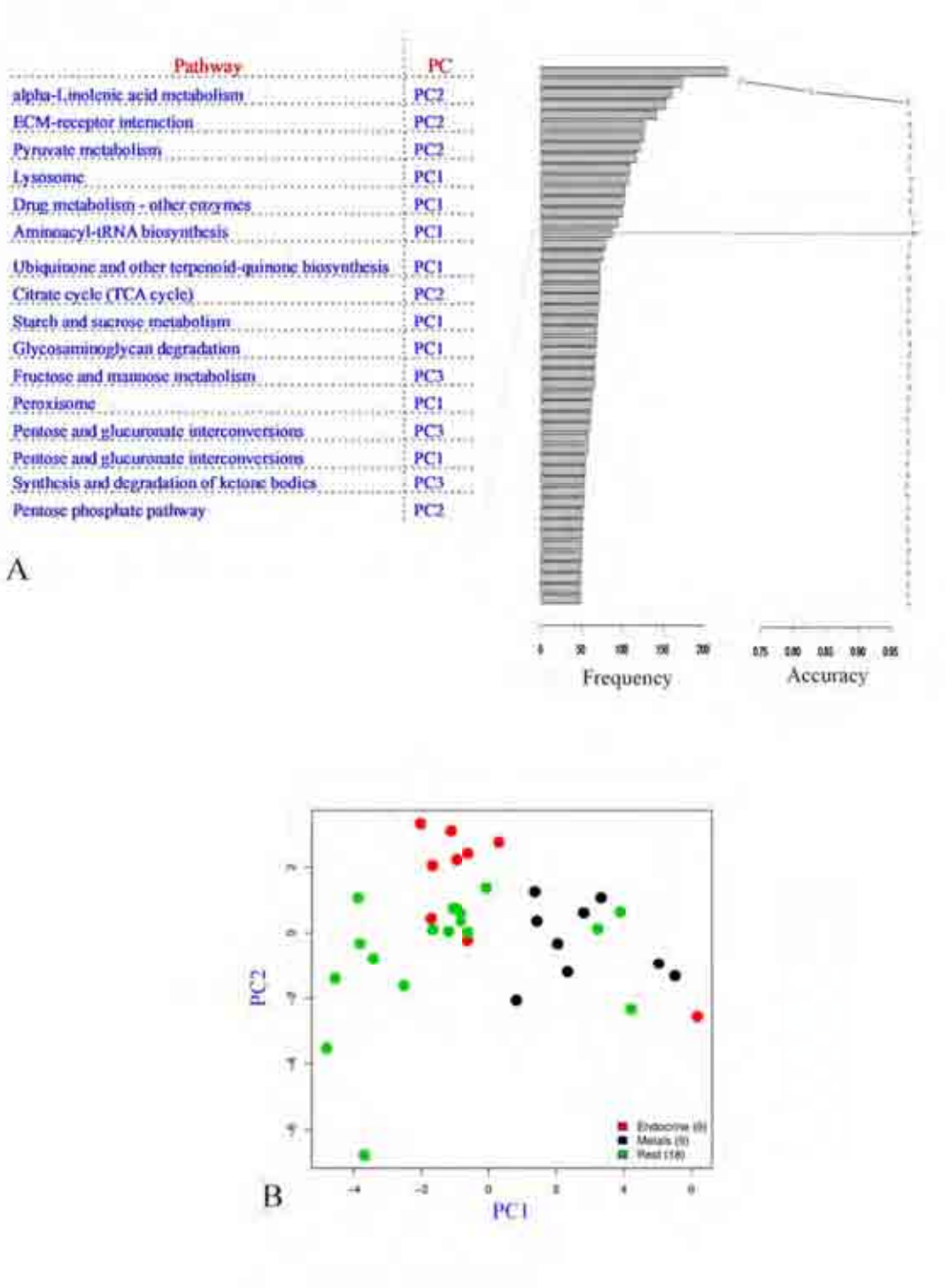


Figure 3.5: **Pathway level Model Discriminating between Metals, Endocrine Disruptors and Remaining Chemicals.** Panel A represents the results from the forward selection strategy. Pathways are first sorted by their selection frequency (A_2). Accuracy (A_3) is then evaluated by incrementally adding the most frequent pathway components. The dotted line in A_3 shows where the model reached 98.5% accuracy. The list of 16 pathways identified is shown in A_1. Note that the top 4 pathways are sufficient to reach 98% accuracy. Panel B shows a PCA of the indices of pathway activity. Samples marked in black, red and green are metals, endocrine and remaining chemicals respectively. Separation between metals, endocrine disruptors and industrial chemicals is less evident than in the gene-level model.

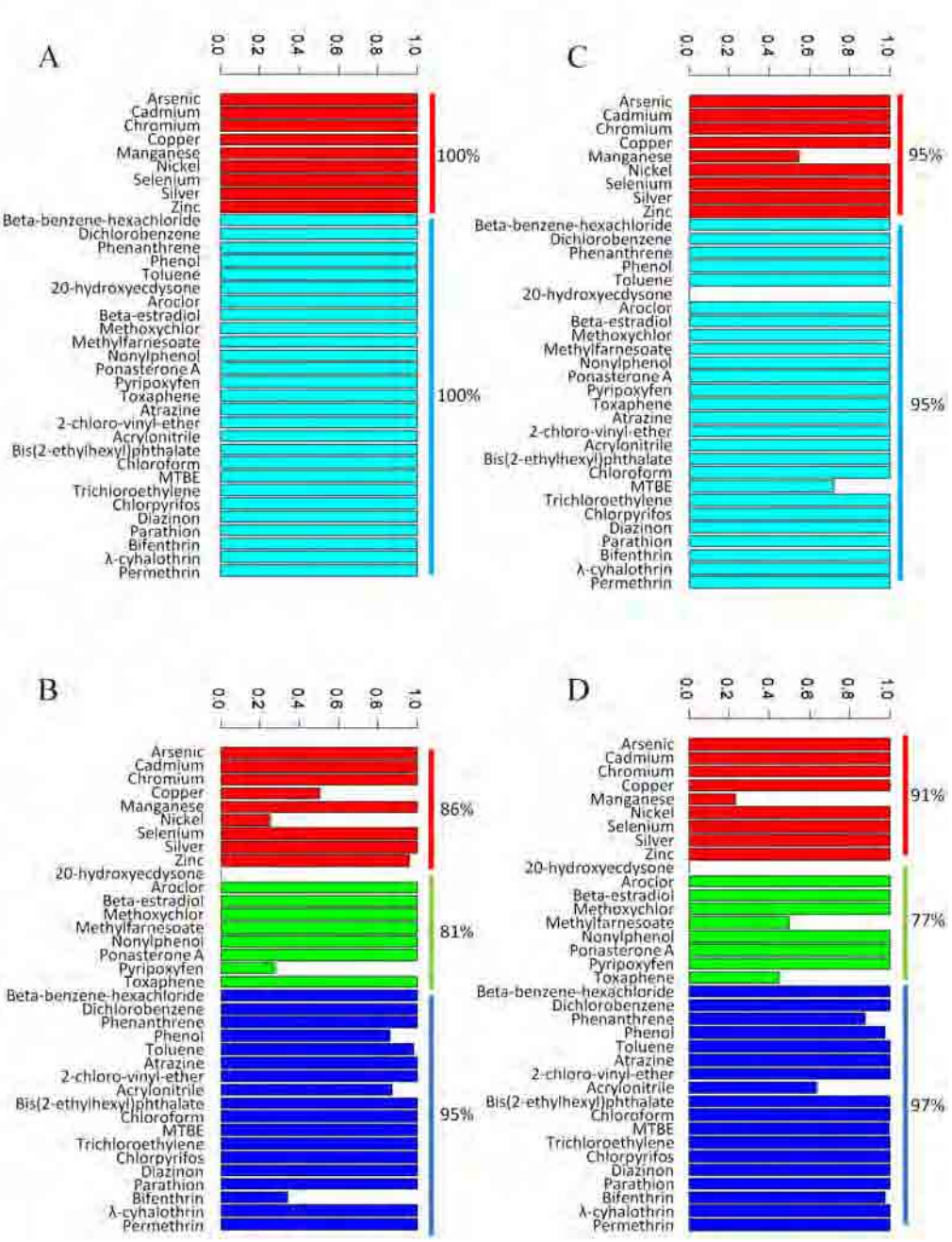


Figure 3.6: **Leave One Out Cross Validation (LOOCV) Results.** The Figure represents the prediction accuracy for each chemical, including all replicates, for the developoed models (see materials and methods section for further information). The bars are indicative of how well our models are able to predict each chemicals class. A and B represent the gene level models shown in Figures 3.2 and 3.3. The gene-level model in A shows that all chemicals are correctly classified by our model. In contrast, not all chemicals are correctly classified in panel B (for example, 20-hydroxyecdysone is always misclassified). Panels C and D are represenative for the pathway level models in Figures 3.4 and 3.5. It is interesting to note that the pathway level model in D is better at classifying metals and industry related compounds as compared to the gene level model (B). Bars across each group represent the average classification accuracy for that class. The colours correspond to, red = metals, light blue = non-metals, green = endocrine disruptors, blue = industry relevant chemicals.

3.3.5 IPA Analysis Identifies a Super-Network Representing Plausible Adverse Outcome Pathways and Integrating Energy Metabolism, beta-estradiol with TGF-β and IFN-γ signalling

In order to explore whether genes most contributing to the pathway indices (top 25%) are connected in the context of a higher order biological network, we performed an ingenuity pathway analysis for each pathway-level representative model.

From the models predictive of metals versus non-metals the ingenuity pathway analysis returned three networks, which represented the interaction between components of the oxidative phosphorylation (NADH2 dehydrogenases and ATPases) and growth factor/inflammation signals, (TGFB1, IFNG and TNF). More specifically Figure 3.7A shows the interaction between hydrogen peroxide with NADH2 dehydrogenase and lysosomal and mitochondrial ATPase components. In Figure 3.7B hydrogen peroxide is also connected to IFNG and TGFB1 which in turn are connected to genes in detoxification pathways, including CYC1, COX6C, NNMT and GLRX2, cholesterol binding/release, through VEGF to NPC1, AP1G1, AP1B1 and SMPD1 and to sugar metabolism genes including MANBA, MAN2B1, DPAGT1 and MDH2. It should be noted that several oncogenes (FOS, RAB36) and an apoptosis regulator (BID) have also been identified. The third and last network for this comparison (Figure 3.7C) mostly shows the connection between TNF and various citrate cycle genes (PDHA1, PDK3) as well as preoteases and glycoaminoglycan degradation components.

The IPA analysis of the model classifying metal, endocrine disruptors and all remaining chemicals, returned two networks (Figure 3.8). At first glance the networks in Figure 3.8A and 3.8B are highly similar to the networks we have previously discussed. However the interaction between phospholipases, TNF, glycoaminoglycan metabolism and the previously described cholesterol binding is much more evident (Figure 3.8A). In addition to these, Insulin is connected to TNF which at the centre of this network is possibly the master regulator and highly important to the response of chemical exposure. The network in Figure 3.8B goes to further characterise the role of TGFB1 and HNF4A by showing connections to IFNG, pyruvate

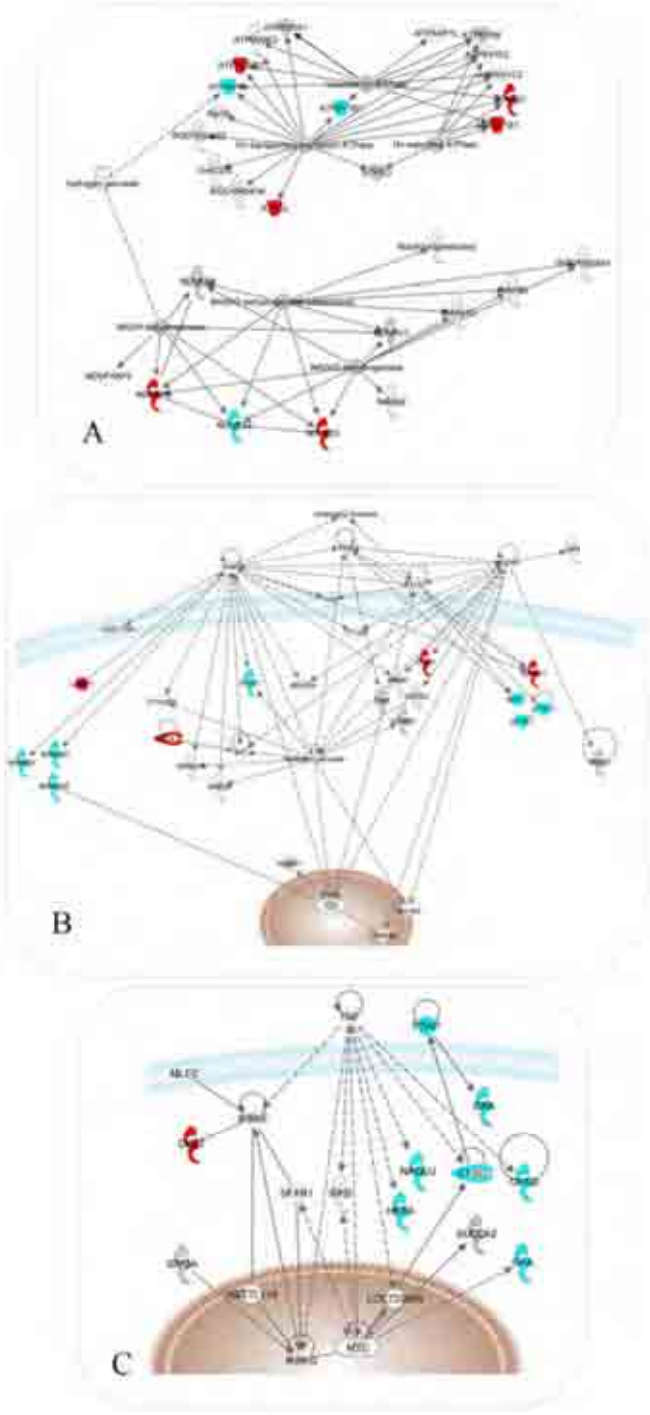


Figure 3.7: **Ingenuity Pathway Analysis Networks Build using the Pathway-level Representative Model for Metals vs. Non-metals.** Dotted and continuous lines represent indirect and direct relationships respectively. Genes coloured in red and blue are up and down regulated respectively between metals and non-metals. Grey coloured genes are part of the input list but did not show any significant change. Genes coloured in white were added by the ingenuity pathway analysis. A) shows the interaction between hydrogen peroxide with NADH2 and ATPase components, B) the interaction between hydrogen peroxide and IFNG and TGFB1 connected to detoxification, and C) the connection between TNF and the TCA-cycle.

metabolism and peptidase activity. It is interesting to note that beta-estradiol in this network could be the potential signal mediating the response.

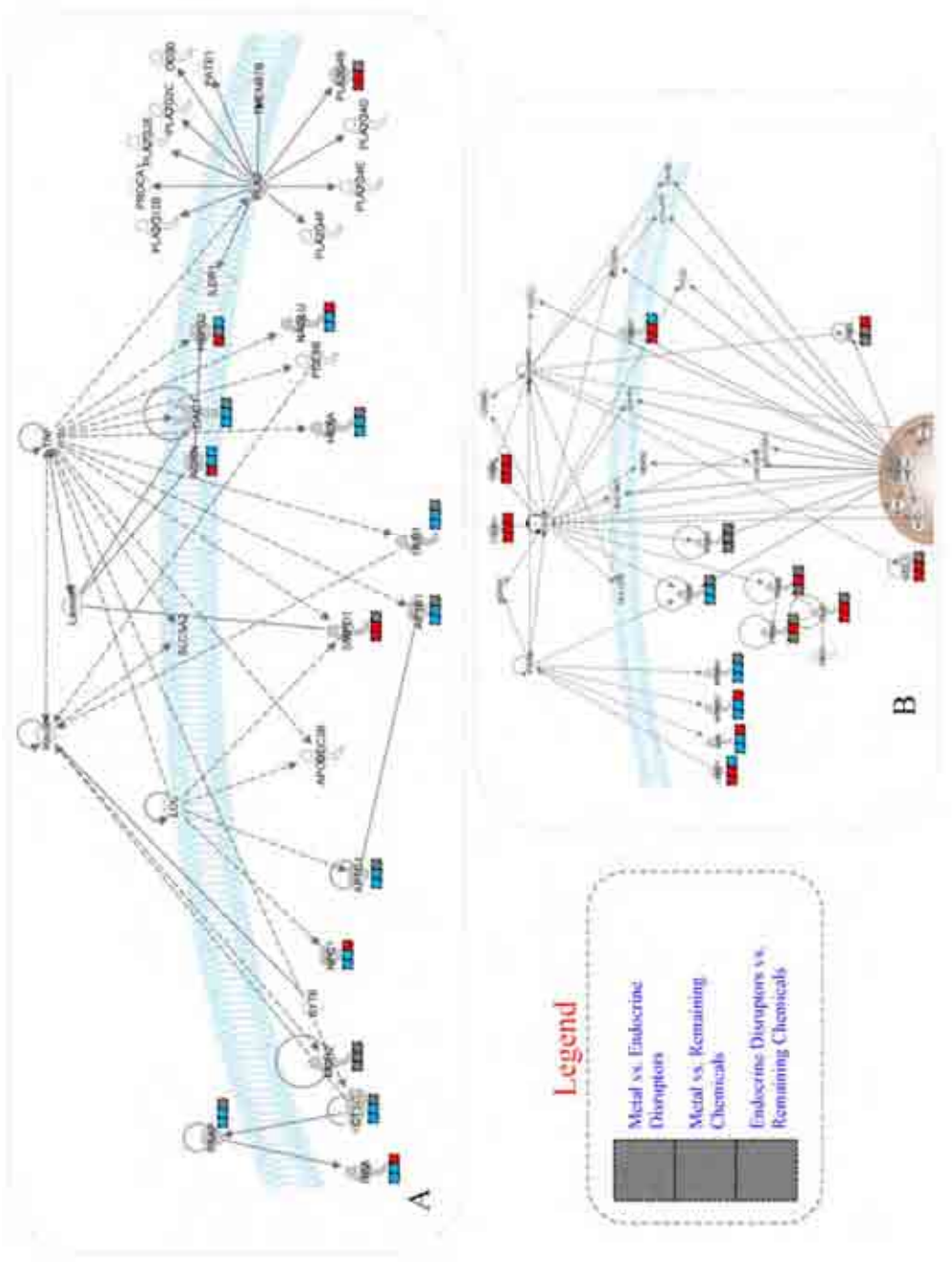


Figure 3.8: **Ingenuity Pathway Analysis Networks for the Pathway-level Representative Model of Metal vs. Endocrine Disruptors vs. All Remaining Chemicals.** Dotted and continuous lines represent indirect and direct relationships respectively. Boxes next to the genes specify if they have been up (red) or down (blue) regulated within the comparisons as denoted by the legend. Genes with white backgrounds have been added by the ingenuity analysis. A) shows the interaction between phospholipases, TNF, glycoaminoglycan metabolism and cholesterol binding. B) characterizes the role of TGFB1 and HNF4A through IFNG, pyruvate metabolism and peptidase activity.

3.4 Discussion

The most important finding of this work has been the demonstration that it is possible to predict the identity of a toxicant from molecular signatures representing exposure to sub-lethal concentrations. More broadly this is the first demonstration of the validity of a integrative approach to developing a quantitative assay for water quality assessment, informative of chemical contaminants and based on *D. magna*. Moreover we have shown that a limited annotation is not prohibitive of developing highly predictive models of chemical class. The identified pathways are consistent with the current understanding of Daphnia toxicology and in addition provide novel hypotheses on previously uncharacterized AOPs.

3.4.1 Metabolic Imbalance Characterized Response of *D. magna* to Metal Exposure

The top 2 most representative pathways, in the model discriminating between metals and non-metals, are oxidative phosphorylation and alpha-linolenic acid metabolism. Abnormal expression of genes in these pathways can lead to higher production of reactive oxygen species [211–213]. More specifically enzymes within the alpha-linolenic acid metabolism are part of the peroxisome which is the main site within the cell where generation and removal of oxygen free radicals occurs [211]. In fact enzymes part of this pathway, including acyl-CoA oxidase, D-bifunctional protein (DBP), Sterol carrier protein X (SCPx) and 2-methylacyl-CoA racemase (AMCR), have been associated with a peroxisomal beta-oxidation enzymes deficiency in humans [214–216]. Further investigation of the pathways contributing to the overall accuracy, lysosome related genes were highly important in combination with the other components to reach the high accuracy rate. Publications dating back to the late 60s and early 70s have shown that heavy metals accumulate in lysosomes and in some cases such as mercury or cadmium have the ability to inhibit specific enzyme activities [217, 218]. A more recent publication has also shown that lysosomal activity is linked to reduced chemical toxicity in mice [219]. This raises the interesting hypothesis that the transcriptional regulation of genes encoding for lyso-

somal components (LAMAN, AP1G1, CTSL1, NAGLU, SMPD, GUSB, HEXA, MANBA, NPC, GBA, GLB1, AP1B1, TCIRG1 and CTSD) may be an adaptive mechanism to metal exposure. The remaining two pathways in this model, synthesis and degradation of ketone bodies and TCA cycle, have also been associated to heavy metal exposure [220–224]. More specifically hydroxymethylglutaryl-CoA synthase (HMGCS1) which creates HMG-CoA ((S)-3-hydroxy-3-methylglutaryl-CoA) from either Acetyl-CoA or Acetoacetyl-CoA has been reported to be strongly down-regulated as a result of heavy metal exposure in human HepG2 cells [220]. Within the TCA cycle several of the metals have the ability to inhibit specific steps. Arsenic, for example, has been shown to interfere with pyruvate dehydrogenase (PDH) essentially preventing oxidisation of pyruvate to acetyl-CoA [221–224]. Moreover, cadmium has been shown to interfere with the electron transport chain, directly downstream of the TCA cycle, inhibiting complexes II (succinate:ubiquinone oxidoreductase) and III (ubiquinol:cytochrome c oxidoreductase) [225]. The previously described mechanisms therefore show that a metabolic imbalance caused by the combination of energy metabolism inhibition, aggregation of metals in lysosomes and the creation of reactive oxygen species by perturbing oxidative phophorylation and peroxisomes may hypothetically be the mechanism of action to the response of *D. magna* to metal exposure.

3.4.2 Endocrine Disrupting Chemicals may Act through Pyruvate Metabolism and Extracellular Matrix Receptor Interaction Pathways

The model discriminating metals and endocrine disruptors included the KEGG pathways extracellular matrix (ECM) receptor interaction and pyruvate metabolism. This may be consistent with the observation that displacement of pancreatic tissue by ECM is thought to underlie many major human endocrine diseases [226]. There is evidence that this may be mediated by epidermal growth factor receptor (EGFR) signalling but the specific molecular mechanisms have yet to be identified [227]. Pyruvate metabolism pathway, which is closely related to the TCA cycle, may, in addition to the effects described for the citric cycle, also be linked to endocrine perturbation [228]. An example of this connection has been shown in the exposure of two crus-

tacean species, *D. magna* and *Acartia tonsa*, to 3,4-dichloroaniline. Although this particular chemical is not represented within our dataset, it is also classed as an endocrine disruptor and is included in the EU and US EPA endocrine disruptor monitoring programs [199, 229]. Its effect on *D. magna* includes inhibition of pyruvate kinase (PK) and malate dehydrogenase (MDH), and significant reduction in fecundity [230]. This was also noted by Andersen et al, suggesting that 3,4-dichloroaniline has some specific effects on metabolic enzymes that influence growth in *A. tonsa* [228]. Although this may be a unique scenario, the inclusion of this specific pathway within that specific model may indicate that other known endocrine disruptors may have a similar effect on *D. magna*.

3.4.3 Future Directions

The ultimate goal of this approach was to show that we can indeed predict the nature of chemical contamination using an assay based on relatively simple measurements in *D. magna*. .Further developments may include the development of a qPCR-based assay based on the relatively small number of genes included in the pathways identified by the modelling procedure. In addition, the approach should be extended to a larger number of chemicals and include a broader range of doses and time points.

3.5 Materials and Methods

3.5.1 Exposures and LC_{50} measurements

The laboratory experiments, including exposures, LC_{50} measurements and transcriptomics analyses have been performed by various group members in Chris Vulpes laboratory at the University of California, Berkley, USA. The data has been made available to us prior publication as part of a collaboration to apply systems biology approaches in the field of ecotoxicology. Genetically homogenous *D. magna* were cultured in COMBO media [231] at 23.5°C in a Percival environmental chamber according standard protocols [232, 233]. Chemical exposures were performed using ~40 adult (16 – 18 days) *D. magna* placed in 2L of COMBO media for 24h. LC_{50} values were identified using standard protocols as described in [41]. Sub-lethal concentrations

of chemicals at $\frac{1}{10}$ of LC_{50} were added to the culture. Along side the exposure, zero concentration exposures to each solvent have also been performed. Harvested D.magna were directly ground in liquid nitrogen followed by a standard RNA isolation using Trizol (Invitrogen, Carlsbad, CA).

3.5.2 Expression Profiling

Custom designed microarrays (Agilent Microarray ID: 023710) were ordered from Agilent and three biological replicates of each chemical were hybridized to the array using standard Agilent Low Input Quick-Amp Protocol v.6.0 (Agilent Technologies). The slides were then scanned on a GenePix 4000B Scanner and data was extracted using GenePix Pro 6.0 Software. Microarray data for each chemical were then loess normalized using the suitable control solvent microarrays as reference samples [67]. Out of the 14338 probes available on the array 7324 genes were expressed at a significantly higher intensity as compared to the background.

3.5.3 Computing Indices of Pathway Activity

Since the *Daphnia pulex* genome is completed and all coding regions have been already annotated on the KEGG database [174] we first mapped (by protein blast) the *D. magna* genes represented in our array on the *D. pulex* complete genome. At the time of analysis the KEGG database for *D. pulex* had 3846 genes assigned to KEGG pathway terms. To make sure that pathway indices were representative of a significant fraction of genes within a pathway we only considered pathways where 5 genes or more were available in the input dataset. This reduced the list of 7324 expressed genes to 1671 KEGG annotated genes, representing 101 distinct pathways. These were representative of the broad spectrum of functions represented in this database. Indices of pathway activity were then computed as the first three principal components (PC) of the gene expression profiles representative of each KEGG pathways using the prcomp function within the statistical environment R [234]. These represented at least 70% of variance present in the data. This procedure generated a new derived dataset with 303 pathway components and 144 samples. This was previously demonstrated to be an effective strategy to improve biological interpretability and reduce the computational space [209].

3.5.4 Statistical Modelling Procedure

To identify gene and pathway-level models we employed a genetic algorithm-based variable selection strategy coupled with the classification algorithm Random Forest [163, 235], as implemented in the GALGO package [198] developed in the statistical environment R [234]. This procedure integrates an efficient multivariate variable selection procedure designed to optimize small subset of predictive variables and an advanced classification algorithm that minimize the possibility of overtraining with an in-built out-of-bag cross validation procedure [163]. The modelling procedure was initialized using the default settings in GALGO [198] with a model size of 10. The classification accuracy was estimated as follows: Data was first split into training and test datasets, representing respectively $\frac{2}{3}$ and $\frac{1}{3}$ of the original data. Both training and test sets represent all chemicals but included independent biological exposures. In order to avoid overtraining models were trained within a second level split ($\frac{2}{3}$ training, $\frac{1}{3}$ test) of the training data. Up to 1000 independent models were collected and a representative model developed using a forward selection procedure, as described in [198]. This approach ranks the model variables (genes or pathway activity indices) on the basis of the frequency they appear in the population. The top 50 most frequent variables are then incrementally tested, by adding each variable one by one starting with the most frequent. The model with the smaller number of variables and the higher accuracy is then selected as the final representative model (Figure 3.1, Step 5).

3.5.5 Computing Chemical-Specific Classification Accuracy

Although the main aim of this project was to assess whether it was possible to develop predictors of chemical class within a defined subset of chemicals we also wanted to assess whether the models would be able to predict chemicals that have not been used to train the model itself. We therefore implemented a leave one cross validation (LOOCV) procedure where each chemical in every class is removed to generate a training set. The model is then tested on the chemicals taken out and the predictions matched against the known identity of the chemical. The procedure is then repeated for every combination of chemicals.

3.5.6 Ingenuity Pathway Analysis

Using the Ingenuity Pathway Analysis (IPA, Ingenuity Systems®, www.ingenuity.com) software we performed biological interpretation and identification of biological networks defined by genes represented in the predictive KEGG pathways. These were identified by choosing the top 20% most contributing genes to the particular identified PC of a given pathway. Once the gene lists were uploaded into the application, each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. These genes, called focus genes, were overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of these focus genes were then algorithmically generated based on their connectivity according to the following procedure implemented in the IPA software application. The specificity of connection for each focus gene was calculated by the percentage of its connection to other focus genes. The initiation and the growth of pathways proceed from the gene with the highest specificity of connections. Each network had a maximum of 35 genes for easier interpretation and visual inspection. Pathways of highly interconnected genes were identified by statistical likelihood. Networks with a Score greater than 20 and containing more than 60% of focus genes were selected for biological interpretation.

Chapter 4

A Pathway-based Approach to Predictive Toxicology in the Crustacean *Daphnia magna*

4.1 Abstract

Identifying the response of an organism as a result to chemical exposure is vital in understanding the impact of human activities on the environment. Freshwater species are among the most endangered due to industrial spills, sewage or pesticide run offs. Unfortunately, often, the impact of environmental pollution on the physiology of an organism and the identity of the toxicants are hard to establish. The work described in this chapter is based on the underlying hypothesis that the molecular response of a given organism when exposed to a complex mixture of chemicals is predictive of post hoc toxicity. An extension of this principle is that even an early response to sub-lethal exposure may be predictive of potential toxicity at higher doses. In order to test this hypothesis, we exposed *Daphnia magna* adults to sub-lethal doses (10% of the measured LC_{50}) of chemicals of environmental relevance. Here we demonstrate that components of the whole-organism transcriptional response to exposure are predictive of toxicity outcome (measured as LC_{50}). In depth analysis of the pathways represented in these predictive signatures are consistent with a perturbation at the level of signalling pathways leading to changes in the

expression of genes encoding for enzymes involved in amino acid metabolism. Furthermore we propose a general toxicity mechanism, which we hypothesise to be conserved across different species.

4.2 Introduction

The U.S. EPA reported that in 2005 alone 3.8 billion pounds of toxic chemical waste was released into the environment [236]. In most cases the effects of these complex set of exposures are only visible after they have caused severe developmental, generational and physiological damage to the organisms in the environment. This makes remediation a very difficult task. Exposure effects have been characterized either with conventional physiological endpoints or with more advanced molecular techniques. Data on exposed fish populations showed that ambient levels of known endocrine disruptors causes sexual disruption and experiments in developing rodents showed a concerning change in reproductive organ development [237–241]. These and other groups of chemicals are not only dangerous to species residing in fresh and salt water habitats but also to the human and animal populations who regularly feed on them. Most fish, such as the European eel, accumulate xenobiotics over their lifetime and so endanger their predators health [242, 243]. Other dangers, particularly to humans, may include exposure to toxins through drinking water. Although there are strict regulations on water quality assessment, the employed techniques are not very sensitive or indicative of contamination source. In most cases, information about mechanisms of action or toxicity is missing for a large number of chemicals. In this context, legislations such as REACH, have been adopted to address this challenge. Characterizing the toxicity profile of the more than 30,000 chemicals, which are produced by Europe alone, is however an impossible task. The development of easy to implement quantitative and sensitive essay for predicting toxicity effects from the response of key reference/biosensor species is a potential way forward. Other approaches, such as quantitative structure relationship (QSAR) analyses aim at predicting toxicity outcome from the analysis of compound physico-chemical features (PCFs). Although effective in some applications [104, 145, 244], these methods have failed to provide a comprehensive solution to the

problem. We have shown (Chapter 2 and [209]) that QSAR models can benefit from integrating data representative of the molecular response to exposure to produce more accurate predictive models [2,3]. In this chapter we report the results of a pilot study that supports the use of *Daphnia magna* as a biosensor for predicting potential toxicity effects at whole organism level. We demonstrate that the transcriptional response to sub-lethal doses of chemicals is predictive of toxicity outcome (measured LC_{50} value). Using an approach developed in chapter 2, we show that pathways whose activity is predictive of toxicity are also linked to specific compound PCFs. Interestingly our results suggested the existence of a general toxicity mechanism, similar to the one described in chapter 2, raising the possibility that these mechanisms may be conserved across philogenetically divergent species.

4.3 Results

4.3.1 Rational of the Approach and Data Analysis Overview

The work described here is based on the application of statistical modelling integrating molecular response data (mRNA expression) and compound physical chemical features to develop predictors of toxicity outcome (measured as LC_{50}).Similarly to what was described in the previous chapter, our strategy is based on indices of overall pathway activity that enable the development of more biologically interpretable statistical models (Figure 4.1, Step 1). Using this technique we can effectively reduce the complexity of a 7324 gene data set into 303 KEGG Pathway indices, representing a total of 101 pathways. A regression modelling approach can then be employed to identify pathway indices (these can represent individual pathways or a combination of these) predictive of LC_{50} values (Figure 4.1, Step 2). The degree of gene level overlap between each of the KEGG pathways can be used to visualize the complexity of the Daphnia whole genome pathway map either in the form of a dendrogram (using hierarchical clustering) or as a graph (using an edge weighted layout procedure as implemented in cytoscape [245]) (see Figure 4.1, Step 3 for an overview and Figures 4.4 and 4.7 for specific examples). These visual representations can provide initial hypotheses based on the significantly associated pathways (Figure 4.1 Step 4a). A QSAR approach (chapter 3) is used to identify potential pathways tar-

geted by chemicals (Figure 4.1 Step 4b). The formulated hypothesis can then be experimentally tested (Figure 4.1 Step 5). In this chapter we provide a proposal for experimental validation in the discussion.

4.3.2 The Transcriptional Profile of *Daphnia magna* Exposed to Sub-lethal Chemicals Concentrations is Predictive of Toxicity

We first asked whether the transcriptional response registered after 24hrs of exposure to sub-lethal toxicants concentrations is predictive of toxicity outcome. By using the approach described in Step 2 of Figure 4.1 we could prove that this was indeed the case. The most predictive model (Figure 4.2B, R^2=0.64) was based on 7 genes. Since interaction components in the model were significantly contributing to its accuracy (Figure 4.2A) we concluded that there was a strong evidence for gene-gene synergic effects. The gene NDUFB8 was the only single-gene component to significantly contribute to the model prediction. All other model components represented the interaction between gene pairs (A4GALT -RPL18, A4GALT-ITGA8 and SC4MOL-ITGA8) (Figure 4.2A). Pathway association and KEGG gene names are shown in Table 4.1.

	Pathway	Gene Name
DM14949	Oxidative phosphorylation	NDUFB8 - NADH dehydrogenase (ubiquinone) 1 beta subcomplex 8
DM11923	Glycosphingolipid biosynthesis globo series	A4GALT - lactosylceramide 4-alpha-galactosyltransferase
DM04328	Steroid biosynthesis	C5orf4/SC4MOL - methylsterol monooxygenase
DM04121	Ribosome	RPL18 - large subunit ribosomal protein L18e
DM06070	Tyrosine metabolism, Jak-STAT signaling pathway	TPO - thyroid peroxidase
DM09549	ECM-receptor interaction	ITGA8 - integrin alpha 8
DM06985	Nicotinate and nicotinamide metabolism	NNMT - nicotinamide N-methyltransferase

Table 4.1: **Genes Predictive of Measured Toxicity (LC_{50}).** The table represents the genes identified in our model to be predictive of measured toxicity. Association to KEGG pathways and their human gene names, derived from the KEGG orthology database are shown.

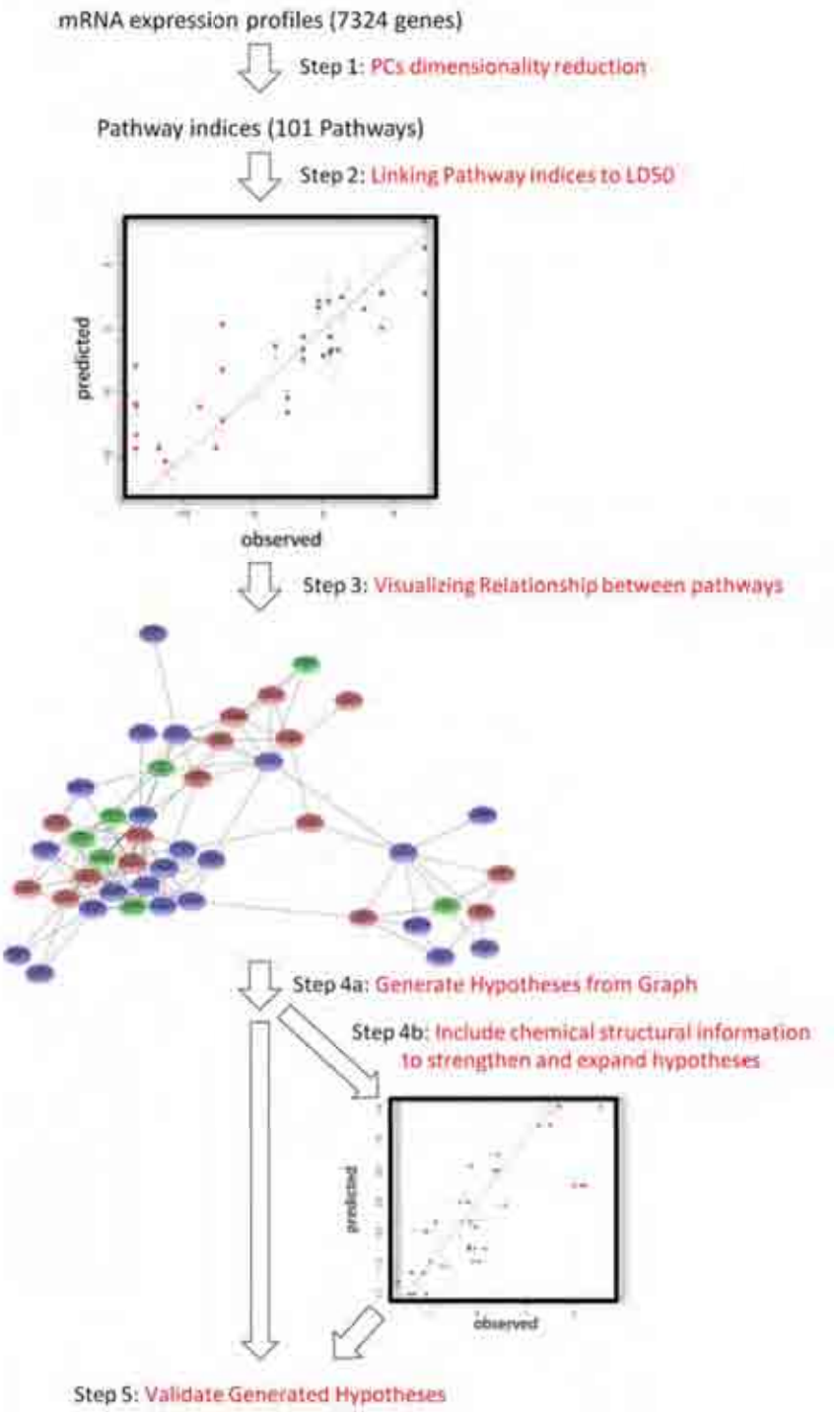


Figure 4.1: **Analysis Strategy Overview.** Step 1: We reduce the dimension of the dataset by using the KEGG database to create pathway indices. Step 2: We link these indices to the LC_{50} value. Step 3: Visualizing relationship between pathways with the use of Jaccards Index and a Force Driven layout or hierarchical clustering. Step 4a: Generate Hypotheses from the graph, identify clusters of pathways closely related. Step 4b: We can get additional information using the structural information of the chemicals. Step 5: We validate the generated hypotheses.

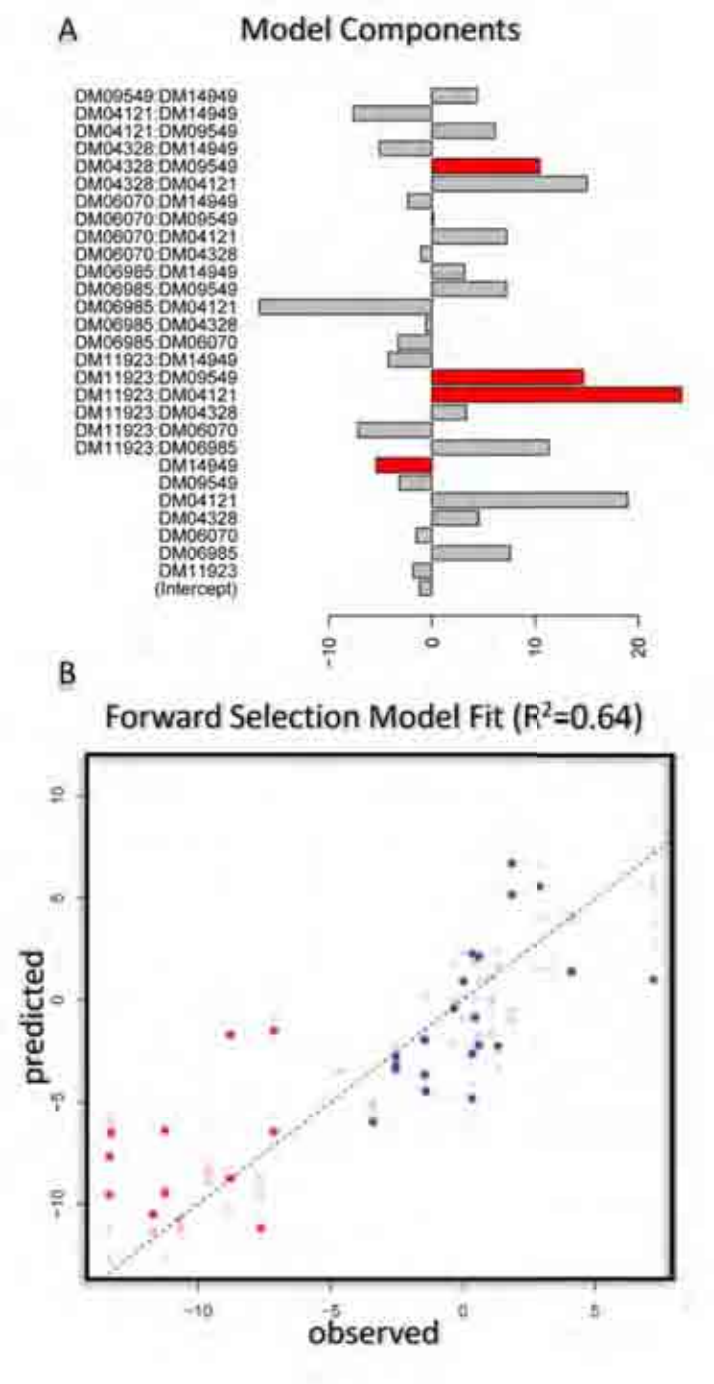


Figure 4.2: **Model Linking Gene Expression Data to Toxicity Outcome (LC_{50}).** The figure shows detailed information on the model we have identified. Panel A provides a graphical representation of the strength of the coefficients (x axis) for each model component (y axis). Significant coefficients marked in red ($p-value < 0.05$. Panel B shows the relationship between the measured LC_{50} values (x axis) and the predicted values by the model (y axis). Chemicals that we identified as high and low toxic have been coloured in red and blue respectively. Transparent points represent the replicates which were used in training the model whereas opaque points represent the test set.

4.3.3 Computing and Validating Indices of Molecular Pathway Activity

We have already shown (chapters 2 and 3) that reducing the complexity of a gene expression profiling dataset by using indices of pathway activity can lead to more predictive and biologically meaningful models. We therefore decided to use this approach to analyse our *Daphnia magna* dataset. Once the indices are computed it is important to assess whether they still contain biologically relevant information. We have addressed this question by assessing the ability of each chemical to perturb pathways indices in respect to its suitable control (T2-hotelling test). We then clustered the calculated statistics, showing that seven out of the nine highly toxic chemicals (77%) cluster within a defined group (Figure 4.3).

4.3.4 The Transcriptional Activity of Some Pathways is Predictive of Toxicity Outcome

Having shown that indexes of pathway activity were able to discriminate between different levels of toxicity, we set to develop statistical regression models linking the three PCs based pathway activity indexes to the measured LC_{50} values. Details of the regression modelling approach are described in detail in the material and methods section of this chapter (Equation 4.2). We could identify 31 out of a total 101 predictive pathways (Table 4.2). These represent three major functional clusters:

1. Signalling pathways such as Wnt, Notch and TGFβ signalling,
2. Lipid metabolism representing both sphingolipid and glycosphingolipid biosynthesis,
3. Amino Acid metabolism such as histidine and tyrosine metabolism (Figure 4.4).

4.3.5 Pathway-based Models are Predictive of Toxicity Outcome

Having demonstrated that indices of pathway activity retain toxicity information and that these can be significantly associated to toxicity, we asked the question whether we can optimize prediction accuracy by developing models that integrate information on the activity of multiple pathways. We not only discovered that this is possible (Figure 4.5) but that the model accuracy is higher than the gene-level model (Figure 4.5B, R^2= 0.7). In comparison to the gene-level model

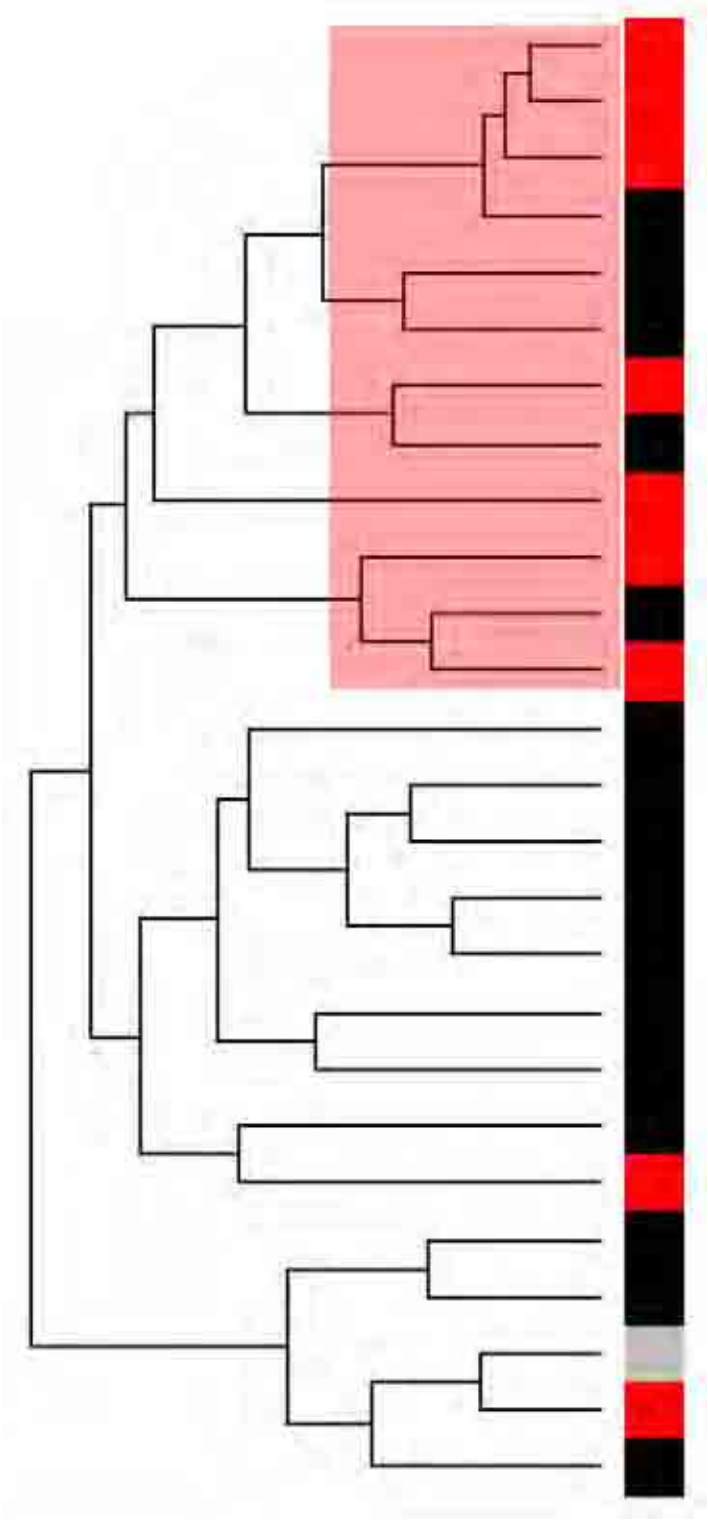


Figure 4.3: **Clustering Analysis of Indices of Pathway Activity.** The figure shows the results of a cluster analysis of individual chemical ability to perturb pathway activity. Chemicals marked in black are low toxic, red highly toxic, and grey is Bis(2-ethylhexyl)phthalate which we found to be in between these two classifications (Figure 4.10). Note that the analysis shows a clear separation between highly toxic and less toxic chemicals.

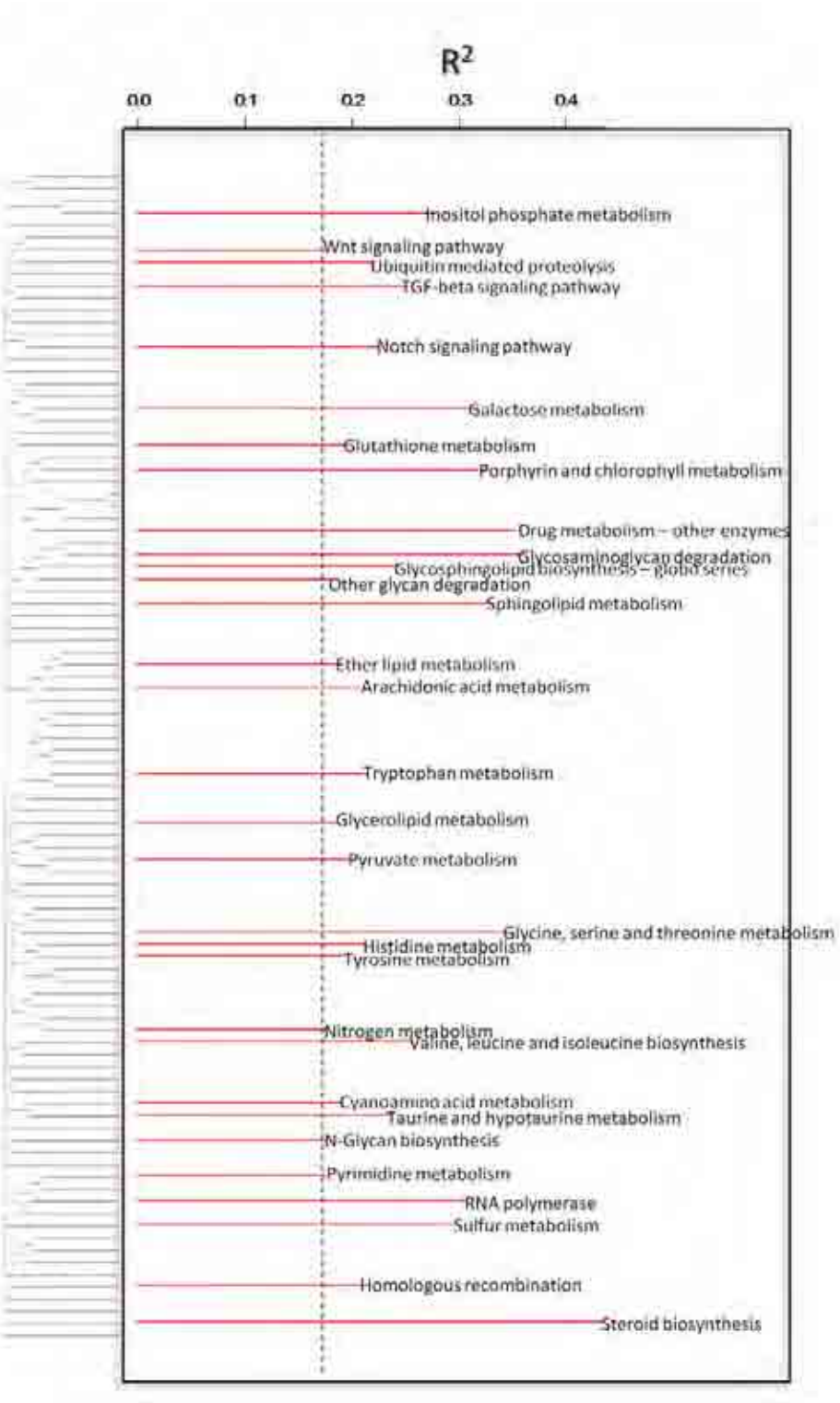


Figure 4.4: **Hierarchical Cluster Representation of the Pathways Identified in the Dataset.** This figure highlights in a graph format the pathways predicitve of toxicity. The pathway order is defined by the dendogram (y axis) which is based on the Jaccard s index of overlap (see Materials and Methods section for more detail). The predictive power (R^2) of the significant pathways is represented on the x axis. The dotted line refers to the cutoff ($FDR < 1\%$) at which significance was assessed.

	Toxicity R^2
Inositol phosphate metabolism	0.27
Wnt signaling pathway	0.17
Ubiquitin mediated proteolysis	0.22
TGF-beta signaling pathway	0.25
Notch signaling pathway	0.22
Galactose metabolism	0.31
Glutathione metabolism	0.19
Porphyrin and chlorophyll metabolism	0.32
Drug metabolism - other enzymes	0.35
Glycosaminoglycan degradation	0.36
Glycosphingolipid biosynthesis - globo series	0.24
Other glycan degradation	0.18
Sphingolipid metabolism	0.33
Ether lipid metabolism	0.19
Arachidonic acid metabolism	0.21
Tryptophan metabolism	0.21
Glycerolipid metabolism	0.18
Pyruvate metabolism	0.20
Glycine, serine and threonine metabolism	0.34
Histidine metabolism	0.21
Tyrosine metabolism	0.19
Nitrogen metabolism	0.17
Valine, leucine and isoleucine biosynthesis	0.25
Cyanoamino acid metabolism	0.19
Taurine and hypotaurine metabolism	0.23
N-Glycan biosynthesis	0.17
Pyrimidine metabolism	0.17
RNA polymerase	0.30
Sulfur metabolism	0.29
Homologous recombination	0.20
Steroid biosynthesis	0.42

Table 4.2: **Pathways Predictive of Toxicity Outcome.** The table lists the predictive power (R^2) of the 31 KEGG pathways whose activity is predictive of toxicity outcome. The order was defined by identifying the gene level overlap between the different pathways as in Figure 4.4.

we noticed that a larger number of individual pathway components significantly contributed to the model prediction. This is consistent with the fact that the indexes we developed already integrate information from multiple genes. The representative model we developed included 5 different pathway components. These were:

1. PC2 of taurine and hypotaurine metabolism

2. PC3 of aminoacyl-tRNA biosynthesis
3. PC3 of nitrogen metabolism
4. PC2 of steroid biosynthesis
5. PC2 of limonen and pinene degradation.

4.3.6 Linking Compound PCFs to Pathway Activity

We have shown that the molecular pathways most contributing to the toxicity prediction represent amino acid and lipid metabolism. We reasoned that similarly to the rat kidney toxicity model described in chapter 3, these may be somehow connected to KEGG pathways linked to PCFs. In order to test this hypothesis, we developed regression models linking PCFs to the transcriptional activity of a given pathway. We successfully identified predictive models ($R^2 > 0.7$) for 32 unique pathways including different combinations of PCs (a detailed overview of these models can be found in Table 4.7). Among pathways associated to chemical features we observed signalling pathways (e.g. phosphatidylinositol signalling system, mTOR signalling and Wnt signalling pathway), several amino acid metabolism (e.g. histidine, glycine, serine and threonine and taurine and hypotaurine metabolism) and drug metabolism (e.g. glutathione and cytochrome P450 metabolism) pathways. Some examples of pathways linked to PCFs are shown in Figure 4.6. Most interestingly PC2 of taurine and hypotaurine pathway (Figure 4.6D) has the ability to separate high and low toxic chemicals and was previously identified in the pathway-level model. We then asked the question which PCFs were among the most selected across all identified models (Table 4.3). Within the top 4 features we find MATS6e, T(O..Cl), BEHm2 and E1p (Table 4.3). Interestingly electronegativity and polarizability are among the most selected features.

4.3.7 A Network Linking PCFs, Transcriptional Response to Exposure and Toxicity Outcome

In order to assess whether KEGG pathways associated to PCFs and/or toxicity outcome are linked within the overall KEGG pathway system we develop a graph-based representation of their relationship (Figure 4.7). The visual inspection of the resulting graph shows that indeed

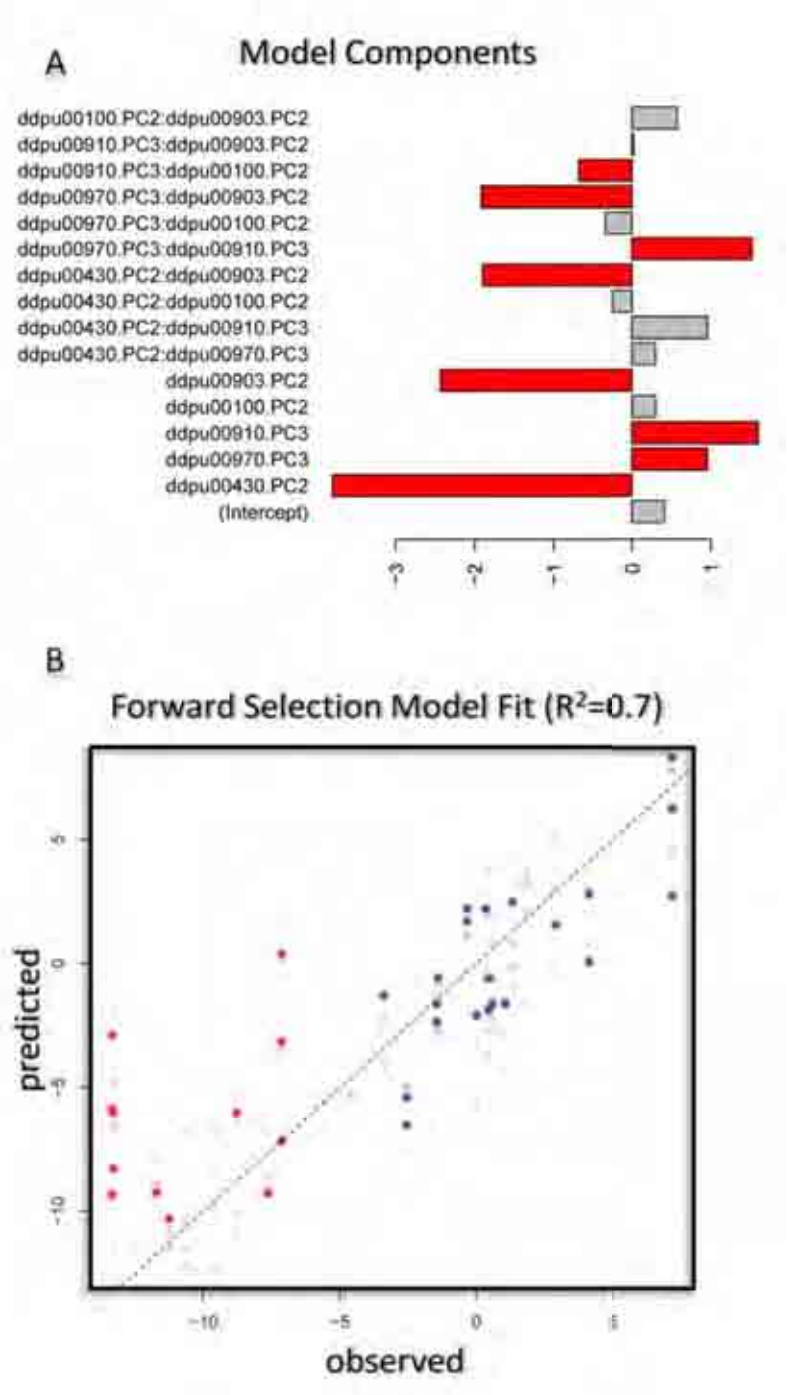


Figure 4.5: **Model Linking Indices of Pathway Activity to Toxicity.** The figure describes the results of the statistical modelling approach. We have identified 5 pathway and their interaction components with an increased goodness of fit ($R^2 = 0.7$). Panel A shows how strong each coefficient contributes to the final model. Interestingly a larger number (8) of statistically significant components as in the gene level model is observed (p-value < 0.05). Panel B shows the experimentally measured LC_{50} values (x axis) versus the ones predicted by the model (y axis). Transparent points are samples that were used in the training set where as opaque points are from the test set. It is interesting to note that the taurine and hypotaurine metabolism pathway is the highest contributing factor in this model.

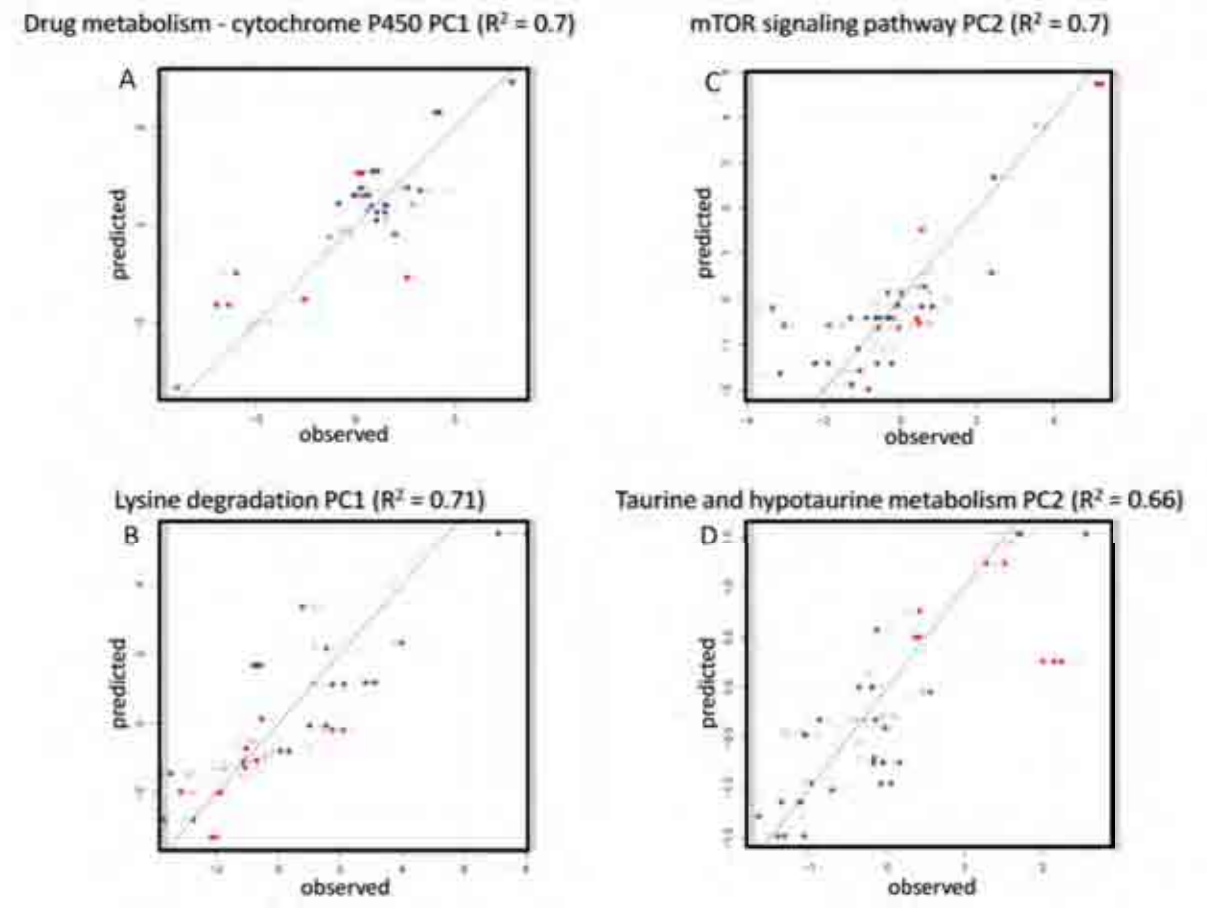


Figure 4.6: **Examples of the Models Found using the Genetic Algorithm Approach.** The figure represents scatter plots of identified models linking PCFs to indices of pathway activity. The x axis represents the calculated PC values for each pathway where as the y axis shows the predicted pathway activity. Panels A and B were derived from pathways significant in PC1 (drug metabolism - cytochrome P450 and lysine degradation respectively. Panels C and D show models derived from PC2 (mTOR signalling pathway and taurine and hypotaurine metabolism). Chemicals which were defined as high and low toxic are represented in red and blue respectively. Transparent points represent the replicates which were used in the training set and opaque points were used in the test set. It should be noted that the taurine and hypotaurine pathway is able to separate high (red) and low (blue) toxic chemicals.

Descriptor	Frequency	Description
MATS6e	250	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)
T(O..Cl)	199	sum of topological distances between O..Cl (2D Atom Pairs)
BEHm2	191	highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses (BCUT descriptors)
E1p	187	1st component accessibility directional WHIM index / weighted by polarizability (WHIM descriptors)
RDF155u	186	Radial Distribution Function - 155 / unweighted (RDF descriptors)
Mor24m	180	signal 24 / weighted by mass (3D-MoRSE descriptors)
H-052	138	H attached to C0(sp3) with 1X attached to next C (Atom-centred fragments)
RDF155e	124	Radial Distribution Function - 155 / weighted by Sanderson electronegativity (RDF descriptors)
GATS4v	108	Geary autocorrelation of lag 4 weighted by van der Waals volume (2D autocorrelations)
E1v	93	1st component accessibility directional WHIM index / weighted by van der Waals volume (WHIM descriptors)

Table 4.3: **Top 10 PCFs Selected by our Approach.** This table shows the specific descriptors which have been most frequently chosen within models ($R^2 > 0.7$) of significantly associated pathways (Table 4.7). Description of the feature is given in column 3 with the descriptor groups in parantheses.

these pathways are linked and that three main functional clusters can be defined. These are:

1. amino Aaid Metabolism
2. signalling pathways
3. lipid metabolism.

Within each of the groups we identified the pathways associated to toxicity (red), PCFs (blue) and pathways that associated to both (green). Interestingly the pathways linked to both toxicity and PCFs are concentrated in the amino acid metabolism cluster.

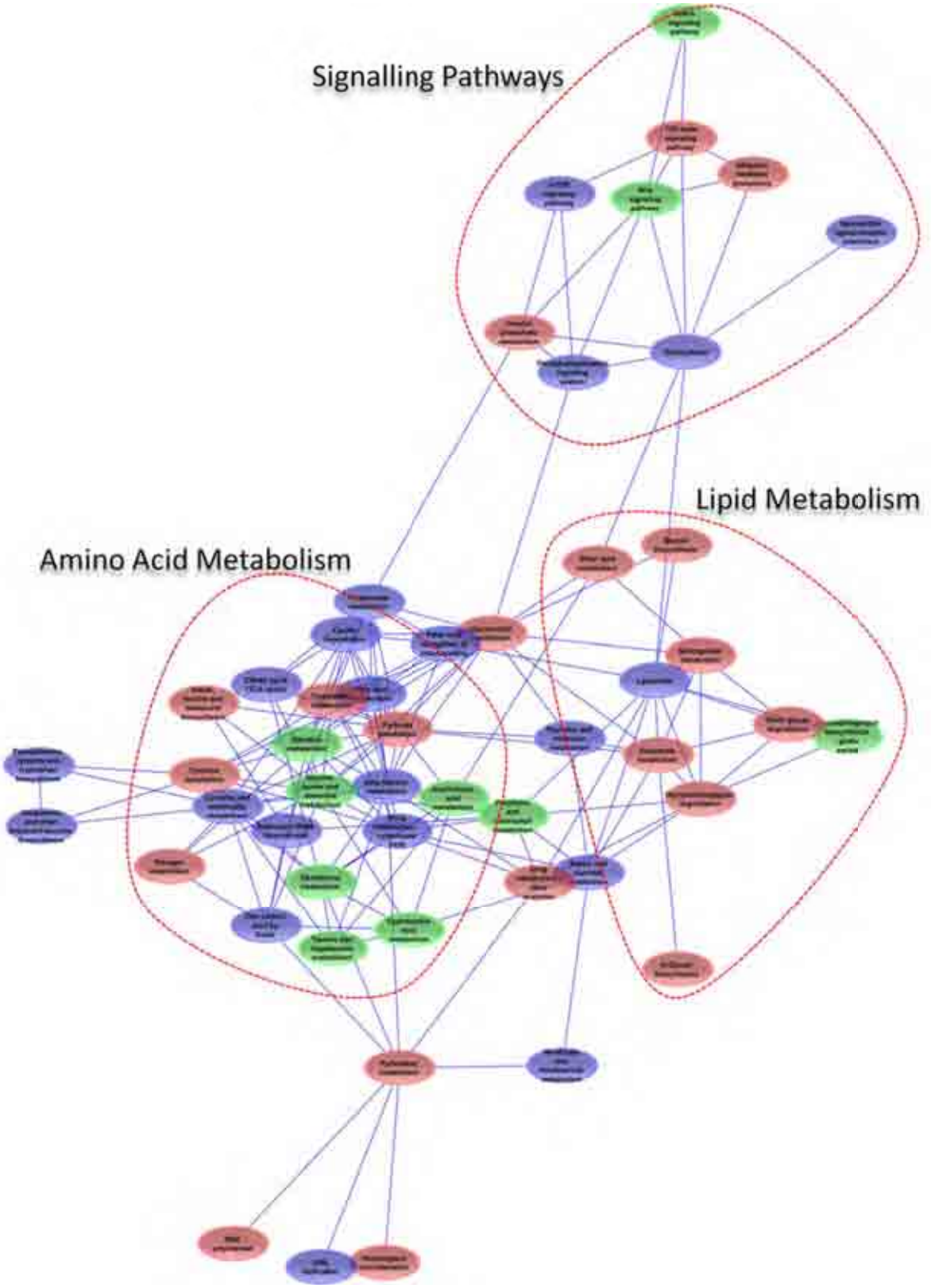


Figure 4.7: **Pathway Representation of the Identified Associations.** Undirected network of the pathways that were identified to be significantly associated to either toxicity (red), chemical structure (blue) or both (green). The distance between nodes is approximately correlated to the number of genes shared by two different KEGG pathways.

4.3.8 Increased Expression of Genes Within the Amino Acid Metabolism and Signalling Pathways is a Landmark of Toxicity Response

Having identified an amino acid metabolism and signalling cluster within Figure 4.7 we set out to characterize each pathways gene-level response in relation to toxicity. We expect that this may give an indication on the mechanism linked to the toxicity response. Within the amino acid metabolism cluster we focused on pathways which were associated to both PCFs and toxicity (glycine, serine and threonine, taurine and hypotaurine, glutathione, cyanoamino acid and porphoryin and chlorophyll metabolism, Figure 4.8). We show that the genes involved in amino acid conversion and metabolism are mainly up-regulated in response to highly toxic chemicals. Moreover it shows that the production of cytochrome C in the porphyrin and chrolophyll metabolism and conversion of NADP+ to NADPH is down regulated. This leads to the hypothesis that amino acids play an important role in *D. magna* toxicity response. Within the signalling pathways cluster we focused on WNT signalling and two additional related pathways, Notch and phosphatidylinositol signalling (Figure 4.9). The Wnt signalling pathway presents several possibilities by which exposure to chemicals could lead to a molecular toxicity response. Interactions between Frizzled and Dsh can lead to perturbations in cytoskeleton or induce ubiquitin mediated proteolysis, in addition Frizzled can activate PLC, which is highly interconnected with the phosphatidylinosotiol (PI) signalling pathway. Downstream pathways of PI signalling include inositol phosphate metabolism, which we found to be associated to toxicity, leading to the hypothesis that calcium may play a vital role in chemical toxicity.

4.4 Discussion

We have shown, for the first time that it is possible to identify gene and pathway-level models predictive of toxicity outcome in *D. magna*. The models we have developed, contributed to the identification of novel potential mechanisms involved in whole organism toxicity of a large set of heterogeneous chemicals.

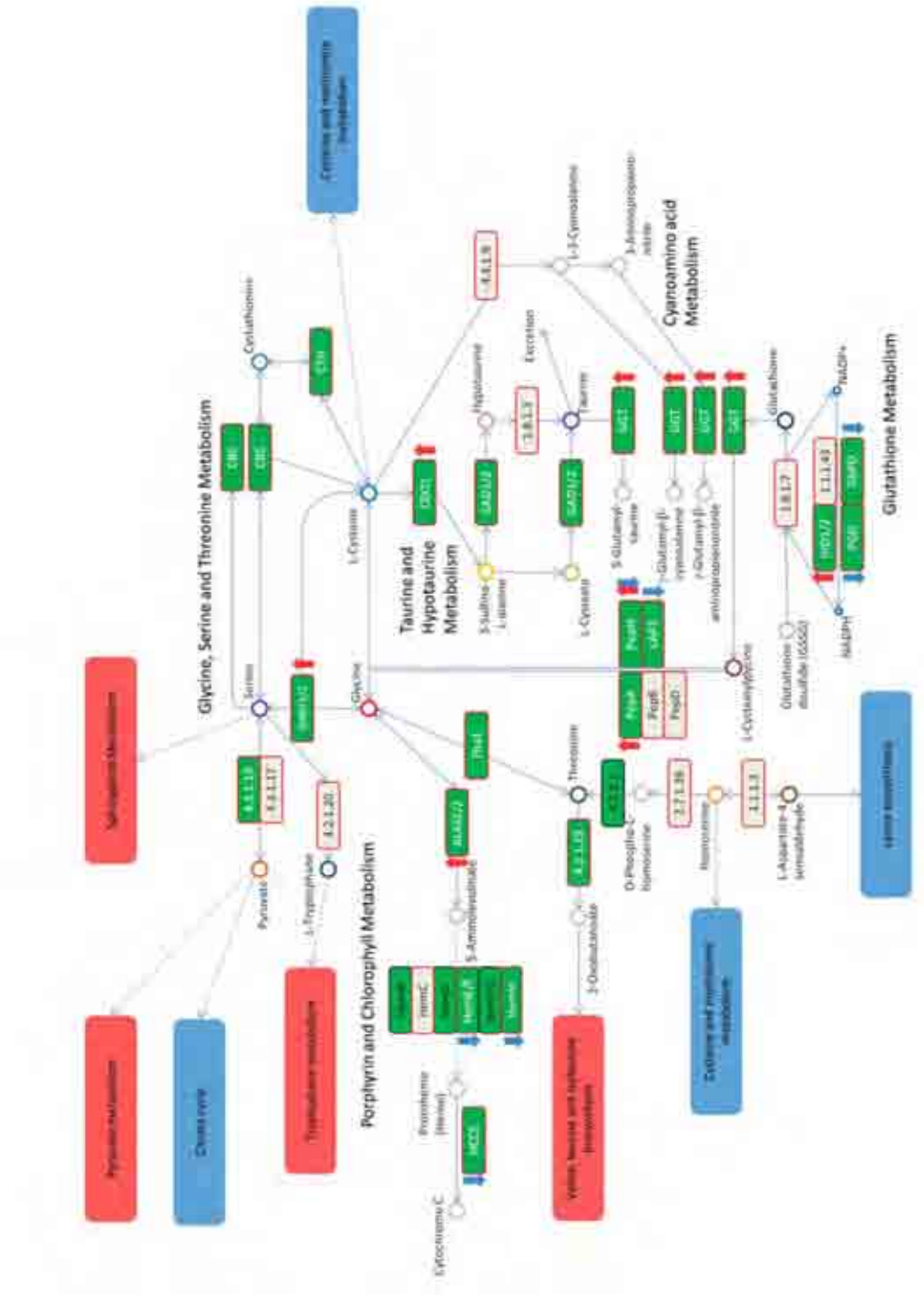


Figure 4.8: **A Cartoon Representation of the Amino Acid Metabolism Pathways we Identified using our Approach.** Central to the remaining pathways is the Taurine and hypotaurine metabolism pathway. Pathways in boxes marked in red and blue are associated to toxicity and chemical structure respectively. Genes marked in red and blue are up and down-regulated as compared to high vs. low toxic chemicals.

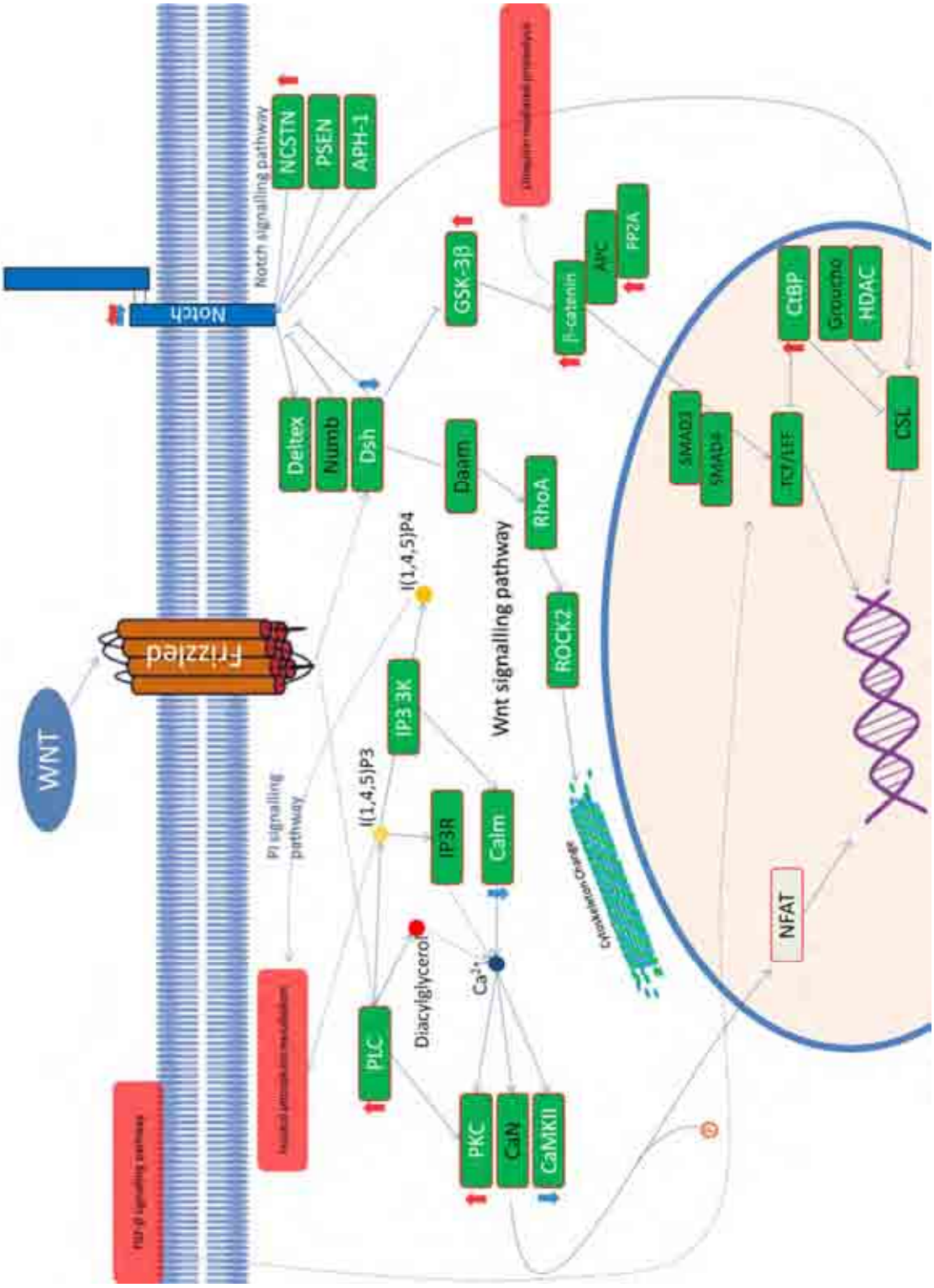


Figure 4.9: **Cartoon Representation of the Signalling Pathways we found to be Associated to Toxicity and Chemical Structure.** Pathways marked in red and blue are associated to toxicity and chemical structure respectively. Genes marked in red or blue are up and down-regulated as compared to high vs. low toxic chemicals.

4.4.1 Amino Acid Transporters Provide the Link Between Signalling Pathways and Amino Acid Metabolism

We have shown that the transcriptional regulation of genes involved in amino acid metabolism is central to toxicity prediction. We have also identified a link between signalling pathways, such as TGF-β and the expression of these genes. There is experimental evidence, which provide a precise mechanism that may explain these statistical relationships. Perturbation in Wnt signalling is known to affect amino acid transport through the PI pathway [246,247]. Moreover, TGF-β also has this ability, via direct regulation of system A amino acid transporter 2 (SAT2) in vascular smooth muscle cells [248, 249]. TGF-β can also modulate amino acid uptake in myofibroblasts [250] and stimulate glycolysis in NRK-49F cells [251]. Other publications also point towards a link between intracellular calcium concentrations and toxicity [252–255]. Our results are therefore consistent with these findings.

4.4.2 Amino Acid Metabolism and Whole Organism Toxicity

In our results one pathway, taurine and hypotaurine, is particularly interesting as it was associated to both toxicity and chemical descriptors. Taurine is the only amino acid that is not used to form proteins and is one of the most abundant free amino acids (FAA) in crustaceans [256]. In addition taurine can be used to estimate the concentration of FAA in the system [257]. *D. magna* however does not seem to be only marine species whose amino acid pool is altered as a result of chemical exposure. Graney and Giesy showed that long-term exposure to pentachlorophenol (PCP) significantly reduced free amino acid reserves within 5 days of exposure even at the lowest tested exposure in *Gammarus pseudolimnaeus Bous eld* [258]. Williams et al demonstrated that exposure to polycyclic aromatic hydrocarbons (PAHs) alters concentrations of taurine, malonate, glutamate, and alanine in three-spined sticklebacks (Gasterosteus aculeatus) while observed gene expression changes related to bile acid biosynthesis, steroid metabolism, and endocrine function [44]. Furthermore, Katsiadaki et al then also showed changes in amino acid concentrations in the same species exposed to ethinylestradiol [45].

4.4.3 Blocking of Amino Acid Transporters may Reduce Toxic Effects

Our data and previous publications demonstrated that initial interaction of chemicals with the cell membrane may cause an imbalance within signalling pathways leading to a change in amino acid transporters [248, 249, 259], reducing the overall amino acid pool [44, 45, 258] and concluding in toxicity. A similar connection between stimulation of receptors and amino acid transport through accumulation of intracellular calcium has already been identified by Turski et al [252, 260]. More specifically they administered MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine), an excitatory amino acid antagonists, to reduce the effects from neurotoxicity [260]. Hence the observed toxicity of a chemical may depend on its ability to perturb the cell membrane and subsequently the discussed pathways. It may therefore be possible to identify a specific set of biomarkers, based on a combination of gene expression and amino acid concentration levels, to accurately predict sub-lethal exposures to toxic chemicals even in the environment.

4.4.4 Electro-Potential Features Associate to Identified Pathways

The identification of a super-pathway representing a link between signalling events and amino acid metabolism that is both correlated to PCFs and predictive of toxicity is suggestive of a broad functional response involving events localized on the cell membrane (signalling) and in the cytoplasm (metabolism). This highlighted electronegativity and polarizability among the most predictive features of molecular response. It is interesting to note that the rat model of renal tubular degeneration we describe in Chapter 2 [209] was also based on a similar set of PCFs. The models were also similar at the functional level in respect to the signalling component but appear to diverge in the downstream effector pathways (Actin remodelling and increased ROS production versus amino acid metabolism). We cannot rule out the possibility that the failure to identify a link between chemical exposure and changes in the expression of amino acid metabolism genes in the rat model may be a consequence of the different time frame in monitoring the molecular response or the different toxicity endpoint. In any case, we speculate our models may capture a general mechanism of response, possibly centred on membrane effects

and shared by these two very different toxicity models.

4.5 Conclusion

The key question, which is still remaining, is whether our approach has identified mechanisms which apply outside of the controlled laboratory environment. Can we utilize this knowledge to build models predictive of exposure in real life scenarios entailing complex mixtures, varying environmental properties and genetic diversity? We believe that with further experiments and application of our computational methodology may indeed address some of these issues. For a more detailed discussion see Chapter 6.

4.6 Material and Methods

4.6.1 The Experimental System

The work described in this paper is based on a model of chemical toxicity in *Daphna magna*, developed in Prof. Chris Vulpe s laboratory at University of California, Berkley, USA [41]. Briefly *D. magna* were exposed to sub-lethal ($\frac{1}{10}LC_{50}$) concentrations of 26 industry relevant chemicals including endocrine disruptors, benzenes, pyrethroids, organophosphates, herbicide and other unclassified chemicals (Table 4.4). Control exposures using the solvents used were also performed. The mRNA was extracted by first grinding harvested *D. magna* in liquid nitrogen using a pestle and mortar and then using Trizol according to standard methods (Invitrogen, Carlsbad, CA). Agilent custom microarrays (AMAID: 023710) were used for transcriptomics measurements. Data were loess normalized using suitable solvent controls and lowly expressed genes (twice the raw average background) were removed before further analysis.

4.6.2 Annotation of *Daphnia magna* Microarrays

The annotation of the *Daphnia magna* array, particularly in reference to the KEGG pathway database, was crucial to our approach. The challenge was facilitated by the availability of the complete *Daphnia pulex* genome sequence, which is already partly mapped in the KEGG pathway database. More precisely, we used a protein blast to identify homologues of the 7324

Classification	Compound
Endocrine Disruptors	20-hydroxyecdysone
	Beta-estradiol
	Methoxychlor
	Methylfarnesoate
	4-nonylphenol
	Ponasterone A
	Pyripoxyfen
	Toxaphene
Organophosphates	Chlorpyrifos
	Diazinon
	Parathion
Benzene Derivatives	Beta benzene hexachloride
	Dichlorobenzene
	Phenanthrene
	Phenol
	Toluene
Pyrethroids	Bifenthrin
	λ-Cyhalothrin
	Permethrin
Herbicide	Atrazine
Industry Relevant Chemicals	2-chloro-vinyl-ether
	Acrylonitrile
	Bis(2-ethylhexyl)phthalate
	Chloroform
	MTBE
	Trichloroethylene

Table 4.4: **Chemicals and their Classes Represented within this Dataset.** This table shows the chemicals and their relative classes represented in this dataset.

genes represented in the *Daphnia magna* array in *Daphnia pulex*. We identified 4958 homologues using the default blast settings (blastx v2.2.21 `blast.ncbi.nlm.nih.gov`). Further filtering to include only marginally significant ($e < 0.1$) genes reduced this list to 1869. Out those we were able to map 1686 genes to 116 *Daphnia pulex* KEGG pathways. Pathways representing fewer than 5 genes in the array were removed resulting in 101 pathways representing 1671 genes. These represented an unbiased sample of the 371 possible KEGG pathways in the original database (Table 4.5).

Table 4.5

High Level Pathway Names	Identified Pathways (in percent)	Total No of Pathways in KEGG
Metabolism	0%	3
Carbohydrate Metabolism	93.33%	15
Energy Metabolism	37.5%	8
Lipid Metabolism	70.59%	17
Nucleotide Metabolism	100%	2
Amino Acid Metabolism	92.31%	13
Metabolism of Other Amino Acids	55.56%	9
Glycan Biosynthesis and Metabolism	40%	15
Metabolism of Cofactors and Vitamins	58.33%	12
Metabolism of Terpenoids and Polyketides	10%	20
Biosynthesis of Other Secondary Metabolites	4.76%	21
Xenobiotics Biodegradation and Metabolism	15%	20
Overview of biosynthetic pathway	0%	9
Transcription	100%	3
Translation	40%	5
Folding, Sorting and Degradation	71.43%	7
Replication and Repair	71.43%	7
Membrane Transport	33.33%	3
Signal Transduction	46.67%	15
Signaling Molecules and Interaction	50%	4
Transport and Catabolism	80%	5
Cell Motility	0%	3
Cell Growth and Death	0%	7
Cell Communication	0%	4
Immune System	0%	15
Endocrine System	14.29%	7
Circulatory System	0%	2
Digestive System	0%	9
Excretory System	0%	5
Nervous System	0%	8
Sensory System	0%	4
Development	33.33%	3
Environmental Adaptation	0%	4
Cancers	0%	15
Immune Diseases	0%	7
Neurodegenerative Diseases	0%	5
Cardiovascular Diseases	0%	4
Endocrine and Metabolic Diseases	0%	3
Infectious Diseases	0%	22
Chronology: Antibiotics	0%	8
Chronology: Antineoplastics	0%	5

Continued on Next Page...

Table 4.5 – Continued

High Level Pathway Names	Identified Pathways (in percent)	Total No of Pathways in KEGG
Chronology: Nervous System Agents	0%	9
Chronology: Other Drugs	0%	9
Target Based Classification: G Protein-Coupled Receptors	0%	10
Target Based Classification: Nuclear Receptors	0%	4
Target Based Classification: Ion Channels	0%	5
Target Based Classification: Transporters	0%	2
Target Based Classification: Enzymes	0%	4
Structure Based Classification	0%	5
Skeleton Based Classification	0%	7

Table 4.5: **Distribution of Identified Pathways in Relation to the Full KEGG Database.** This table shows the percentage of KEGG pathways in relation to their higher level annotation. Column 2 shows the percentage of identified pathways from our approach and column 3 shows the total number of pathways available for each top level in the KEGG database. Note that many of these pathways may not be directly related to crustaceans but apply generally with a slight bias to human physiology.

4.6.3 Summarizing the Transcriptional State of Adult *D. magna* by using Indices of Pathway Transcriptional Activity

In order to reduce the complexity of the transcriptional state of *Daphnia magna* transcriptomics, we computed indices of overall transcriptional pathway activity. PCA was used to summarize each pathways transcriptional activity and the number of PCs was chosen as to explain at least 70% of the variance. The advantage of using PCs is that the inter-gene correlation structure is automatically incorporated into the process of dimension reduction, so this information is not lost. Computation of the PCs has been performed using the principal component function prcomp within the statistical programming environment R [234].

As mentioned above, we first mapped all the expressed probes (7324) on the *Daphnia magna* microarray to the *Daphnia pulex* genome by protein blast. Out of these 5255 probes were returned with significant matches. Furthermore we then mapped these to the available genes in the KEGG database and found that 1686 genes were represented by 117 KEGG pathways. We

discarded any pathway for which we could map fewer than 5 genes resulting in a total of 101 pathways and 1671 genes. The resulting dataset represented 87% of the up to date *D. pulex* annotation in KEGG but only represented 27% of the possible total number of reference KEGG Pathways. The relatively low coverage of the KEGG database in *D. pulex* may be due to the fact that the genome sequence is still in the process of being annotated.

4.6.4 Predicting Toxicity (LC_{50}) by Gene Expression Profiling

To predict toxicity from gene expression profiling we devised a regression model which either takes 3 gene expression profiles or indices of pathway activity (here indicated by θ_{1-3}) and their interactions into account. More precisely, we define:

$$LC_{50} = a\theta_1 + b\theta_2 + c\theta_3 + d\theta_1\theta_2 + e\theta_1\theta_3 + f\theta_2\theta_3 + g + \epsilon \tag{4.1}$$

Where a, b, c, d, e, f, g are model parameters and ϵ is the noise model component. To optimally select genes or indices within our regression model we used a genetic algorithm as implemented in the R package GALGO. Initially the data is split by the algorithm into a training and a test set. The training set is primarily used to train 1000 optimized models using a k-fold cross-validation procedure. Identified models are then further validated using the test set. In our case we identified a single representative model using a forward selection strategy [198]. This approach initially sorts the individual features by their frequency and then incrementally tests the top 50 most selected genes or indices. The combination of features with the highest accuracy is then labelled as our representative model. Figures 4.2 and 4.5 show the results of the approach for the gene and pathway level analysis respectively

4.6.5 Deriving Chemical Physical Features (PCFs).

PCFs were computed using the Web-based toolset E-dragon [197]. E-dragon computes 2352 chemical descriptors by integrating several publicly available methodologies. Only features that could be computed for all chemicals in the dataset were used leading to a total of 1260 chemical physical descriptors.

4.6.6 Linking Chemical Features to Pathway Activity Components.

In order to link chemical descriptors to a given pathway component we used a regression model highly similar to the one proposed earlier (Equation 4.1). We replace the LC_{50} within the equation with the pathway activity ($PC_{i,k}$, where $PC_{i,k}$ is the principal component i of pathway k) and θ_{1-3} now refers to three chemical descriptors including their interaction component. The remaining model components stay the same. The resulting equation is shown in Equation 4.2.

$$PC_{i,k} = a\theta_1 + b\theta_2 + c\theta_3 + d\theta_1\theta_2 + e\theta_1\theta_3 + f\theta_2\theta_3 + g + \epsilon \tag{4.2}$$

As in developing the regression models to predict toxicity outcome, here we identified variable subsets by using a multivariate variable selection procedure based on a genetic algorithm, as implemented in the R package GALGO. Here the data is split again into a training ($\frac{2}{3}$) and a test ($\frac{1}{3}$) set. 300 models are selected on the training set aiming to reach the fitness goal $R^2 > 0.7$. To estimate the R^2 accurately a 5-fold cross validation procedure was used. Pathways for which we could identify predictive models were considered for further analysis. This resulted in the identification of 32 pathways linked to PCFs (Table 4.7 for further details). Figure 4.6 shows examples of models found by the GA in which the predicted values using an optimized model are plotted against the observed PC values for a given pathway.

Further we identified the PCFs, which were selected most frequently, for each significantly associated pathway by our approach. We then chose to represent the top 3 descriptors (Table 4.6).

Table 4.6

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Citrate cycle (TCA cycle) (PC1)	E1p (46)	E1v (35)	BEHm1 (26)	1st component accessibility directional WHIM index / weighted by polarizability (WHIM descriptors)	1st component accessibility directional WHIM index / weighted by van der Waals volume (WHIM descriptors)	highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses (BCUT descriptors)
Citrate cycle (TCA cycle) (PC3)	JGI7 (28)	TPSA(NO) (19)	SEigp (13)	mean topological charge index of order 7 (2D autocorrelations)	topological polar surface area using N,O polar contributions (Molecular propterties)	Eigenvalue sum from polarizability weighted distance matrix (topological descriptors)
Fructose and mannose metabolism (PC2)	HATS7m (31)	T(O..Cl) (21)	R6p+ (19)	leverage-weighted autocorrelation of lag 7 / weighted by mass (GETAWAY descriptors)	sum of topological distances between O..Cl (2D Atom Pairs)	R maximal autocorrelation of lag 6 / weighted by polarizability (GETAWAY descriptors)
Fatty acid elongation in mitochondria (PC1)	BEHm2 (48)	GATS4v (38)	RDF150p (30)	highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses (BCUT descriptors)	Geary autocorrelation of lag 4 weighted by van der Waals volume (2D autocorrelations)	Radial Distribution Function - 150 / weighted by polarizability (RDF descriptors)
Fatty acid metabolism (PC1)	E1p (38)	BEHm2 (30)	RDF155u (30)	1st component accessibility directional WHIM index / weighted by polarizability (WHIM descriptors)	highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses (BCUT descriptors)	Radial Distribution Function - 155 / unweighted (RDF descriptors)
Ubiquinone and other terpenoid-quinone biosynthesis (PC2)	R1e+(24)	MATS1e (22)	GATS1e (20)	R maximal autocorrelation of lag 1 / weighted by Sanderson electronegativity (GETAWAY descriptors)	Moran autocorrelation of lag 1 weighted by Sanderson electronegativity (2D autocorrelations)	Geary autocorrelation of lag 1 weighted by Sanderson electronegativity (2D autocorrelations)

Continued on Next Page...

Table 4.6 – Continued

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Glycine, serine and threonine metabolism (PC2)	RDF145v (38)	RDF150p (30)	RDF150v (28)	Radial Distribution Function - 145 / weighted by van der Waals volume (RDF descriptors)	Radial Distribution Function - 150 / weighted by polarizability (RDF descriptors)	Radial Distribution Function - 150 / weighted by van der Waals volume (RDF descriptors)
Cysteine and methionine metabolism (PC2)	DISPp (21)	RDF040m (16)	Mor06m (10)	displacement value / weighted by polarizability (geometrical descriptors)	Radial Distribution Function - 040 / weighted by mass (RDF descriptors)	signal 06 / weighted by mass (3D-MoRSE descriptors)
Lysine degradation (PC1)	E1m (31)	Dm (30)	RDF040m (24)	1st component accessibility directional WHIM index / weighted by mass (WHIM descriptors)	D total accessibility index / weighted by mass (WHIM descriptors)	Radial Distribution Function - 040 / weighted by mass (RDF descriptors)
Histidine metabolism (PC1)	Mor22m (33)	H-052 (29)	GATS3v (26)	signal 22 / weighted by mass (3D-MoRSE descriptors)	H attached to C0(sp3) with 1X attached to next C (Atom-centred fragments)	Geary autocorrelation of lag 3 weighted by van der Waals volume (2D autocorrelations)
Phenylalanine, tyrosine and tryptophan biosynthesis (PC1)	RDF155e (66)	E1p (56)	RDF155u (54)	Radial Distribution Function - 155 / weighted by Sanderson electronegativity (RDF descriptors)	1st component accessibility directional WHIM index / weighted by polarizability (WHIM descriptors)	Radial Distribution Function - 155 / unweighted (RDF descriptors)
beta-Alanine metabolism (PC1)	RDF155e (58)	RDF155u (57)	HATS1m (38)	Radial Distribution Function - 155 / weighted by Sanderson electronegativity (RDF descriptors)	Radial Distribution Function - 155 / unweighted (RDF descriptors)	leverage-weighted autocorrelation of lag 1 / weighted by mass (GETAWAY descriptors)

Continued on Next Page...

Table 4.6 – Continued

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Taurine and hypotaurine metabolism (PC2)	Mor32p (50)	Mor32e (46)	Mor32u (39)	signal 32 / weighted by polarizability (3D-MoRSE descriptors)	signal 32 / weighted by Sanderson electronegativity (3D-MoRSE descriptors)	signal 32 / unweighted (3D-MoRSE descriptors)
Cyanoamino acid metabolism (PC2)	Mor24m (55)	MATS6e (53)	GATS6e (14)	signal 24 / weighted by mass (3D-MoRSE descriptors)	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)	Geary autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)
Cyanoamino acid metabolism (PC3)	RTe+ (22)	Mor21e (16)	R1e+ (16)	R maximal index / weighted by Sanderson electronegativity (GETAWAY descriptors)	signal 21 / weighted by Sanderson electronegativity (3D-MoRSE descriptors)	R maximal autocorrelation of lag 1 / weighted by Sanderson electronegativity (GETAWAY descriptors)
Glutathione metabolism (PC1)	BEHm2 (39)	E1v (30)	E1p (25)	highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses (BCUT descriptors)	1st component accessibility directional WHIM index / weighted by van der Waals volume (WHIM descriptors)	1st component accessibility directional WHIM index / weighted by polarizability (WHIM descriptors)
Glutathione metabolism (PC2)	MATS6e (25)	E1p (22)	GATS3v (18)	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)	1st component accessibility directional WHIM index / weighted by polarizability (WHIM descriptors)	Geary autocorrelation of lag 3 weighted by van der Waals volume (2D autocorrelations)
Starch and sucrose metabolism (PC3)	T(O..Cl) (93)	G(O..Cl) (58)	GATS6e (34)	sum of topological distances between O..Cl (2D Atom Pairs)	sum of geometrical distances between O..Cl 3D Atom Pairs	Geary autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)

Continued on Next Page...

Table 4.6 – Continued

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Arachidonic acid metabolism (PC3)	MATS1e (20)	HATS7p (17)	R7p+ (17)	Moran autocorrelation of lag 1 weighted by Sanderson electronegativity (2D autocorrelations)	leverage-weighted autocorrelation of lag 7 / weighted by polarizability (GETAWAY descriptors)	R maximal autocorrelation of lag 7 / weighted by polarizability (GETAWAY descriptors)
Glycosphingolipid biosynthesis - globo series (PC3)	GATS4v (34)	G(O..Cl) (20)	T(O..Cl) (14)	Geary autocorrelation of lag 4 weighted by van der Waals volume (2D autocorrelations)	sum of geometrical distances between O..Cl 3D Atom Pairs	sum of topological distances between O..Cl (2D Atom Pairs)
Propanoate metabolism (PC1)	DISPp (33)	E1v (28)	RDF155u (19)	displacement value / weighted by polarizability (geometrical descriptors)	1st component accessibility directional WHIM index / weighted by van der Waals volume (WHIM descriptors)	Radial Distribution Function - 155 / unweighted (RDF descriptors)
One carbon pool by folate (PC2)	H-052 (33)	RDF090m (28)	TPSA(NO) (22)	H attached to C0(sp3) with 1X attached to next C (Atom-centred fragments)	Radial Distribution Function - 090 / weighted by mass (RDF descriptors)	topological polar surface area using N,O polar contributions (Molecular propterties)
Nicotinate and nicotinamide metabolism (PC3)	Mp (58)	Mv (27)	TIC4 (23)	mean atomic polarizability (scaled on Carbon atom) (Constitutional indices)	mean atomic van der Waals volume (scaled on Carbon atom) (Constitutional indices)	Total Information Content index (neighborhood symmetry of 4-order) (Information indices)
Porphyrin and chlorophyll metabolism (PC2)	R6m+ (50)	T(Cl..Cl) (29)	R7v+ (27)	R maximal autocorrelation of lag 6 / weighted by mass (GETAWAY descriptors)	sum of topological distances between Cl..Cl (2D Atom Pairs)	R maximal autocorrelation of lag 7 / weighted by van der Waals volume (GETAWAY descriptors)

Continued on Next Page...

Table 4.6 – Continued

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Porphyrin and chlorophyll metabolism (PC3)	MATS3e (40)	HATS7m (15)	Mor09e (10)	Moran autocorrelation of lag 3 weighted by Sanderson electronegativity (2D autocorrelations)	leverage-weighted autocorrelation of lag 7 / weighted by mass (GETAWAY descriptors)	signal 09 / weighted by Sanderson electronegativity (3D-MoRSE descriptors)
Aminoacyl-tRNA biosynthesis (PC3)	Mor32e (28)	RDF155u (26)	Mor32u (26)	signal 32 / weighted by Sanderson electronegativity (3D-MoRSE descriptors)	Radial Distribution Function - 155 / unweighted (RDF descriptors)	signal 32 / unweighted (3D-MoRSE descriptors)
Drug metabolism - cytochrome P450 (PC1)	MATS6e (26)	GATS8m (22)	MATS8e (15)	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)	Geary autocorrelation of lag 8 weighted by mass (2D autocorrelations)	Moran autocorrelation of lag 8 weighted by Sanderson electronegativity (2D autocorrelations)
ABC transporters (PC1)	E1e (40)	MATS4v (31)	G2u (24)	1st component accessibility directional WHIM index / weighted by Sanderson electronegativity (WHIM descriptors)	Moran autocorrelation of lag 4 weighted by van der Waals volume (2D autocorrelations)	2nd component symmetry directional WHIM index / unweighted (WHIM descriptors)
DNA replication (PC1)	H-052 (44)	RDF130m (34)	MATS1v (33)	H attached to C0(sp3) with 1X attached to next C (Atom-centred fragments)	Radial Distribution Function - 130 / weighted by mass (RDF descriptors)	Moran autocorrelation of lag 1 weighted by van der Waals volume (2D autocorrelations)
DNA replication (PC3)	MATS6e (50)	GATS4v (36)	MATS8e (20)	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)	Geary autocorrelation of lag 4 weighted by van der Waals volume (2D autocorrelations)	Moran autocorrelation of lag 8 weighted by Sanderson electronegativity (2D autocorrelations)

Continued on Next Page...

Table 4.6 – Continued

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Protein export (PC3)	RDF095p (33)	RDF095v (30)	RDF095e (28)	Radial Distribution Function - 095 / weighted by polarizability (RDF descriptors)	Radial Distribution Function - 095 / weighted by van der Waals volume (RDF descriptors)	Radial Distribution Function - 095 / weighted by Sanderson electronegativity (RDF descriptors)
Phosphatidylinositol signaling system (PC1)	BEHm2 (74)	RDF130m (26)	RDF150v (19)	highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses (BCUT descriptors)	Radial Distribution Function - 130 / weighted by mass (RDF descriptors)	Radial Distribution Function - 150 / weighted by van der Waals volume (RDF descriptors)
Neuroactive ligand-receptor interaction (PC3)	FDI (73)	H-052 (32)	Mor08u (26)	folding degree index (geometrical descriptors)	H attached to C0(sp3) with 1X attached to next C (Atom-centred fragments)	signal 08 / unweighted (3D-MoRSE descriptors)
Lysosome (PC2)	Mor24m (47)	MATS6e (43)	T(O..Cl) (37)	signal 24 / weighted by mass (3D-MoRSE descriptors)	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)	sum of topological distances between O..Cl (2D Atom Pairs)
Endocytosis (PC3)	Mor24m (62)	MATS6e (53)	MATS8e (50)	signal 24 / weighted by mass (3D-MoRSE descriptors)	Moran autocorrelation of lag 6 weighted by Sanderson electronegativity (2D autocorrelations)	Moran autocorrelation of lag 8 weighted by Sanderson electronegativity (2D autocorrelations)
mTOR signaling pathway (PC2)	MATS6p (22)	Mor24m (7)	GATS3e (6)	Moran autocorrelation of lag 6 weighted by polarizability (2D autocorrelations)	signal 24 / weighted by mass (3D-MoRSE descriptors)	Geary autocorrelation of lag 3 weighted by Sanderson electronegativity (2D autocorrelations)
Wnt signaling pathway (PC2)	Mor24v (44)	Mor24p (42)	T(O..Cl) (34)	signal 24 / weighted by van der Waals volume (3D-MoRSE descriptors)	signal 24 / weighted by polarizability (3D-MoRSE descriptors)	sum of topological distances between O..Cl (2D Atom Pairs)

Continued on Next Page...

Table 4.6 – Continued

Pathway	PCF1	PCF2	PCF3	Description PCF1	Description PCF2	Decription PCF3
Notch signaling pathway (PC2)	Mor24m (9)	Mor28v (8)	Mor24p (7)	signal 24 / weighted by mass (3D-MoRSE descriptors)	signal 28 / weighted by van der Waals volume (3D-MoRSE descriptors)	signal 24 / weighted by polarizability (3D-MoRSE descriptors)

Table 4.6: **PCFs most Frequently Selected by our Procedure.** For each pathway which was significantly associated to PCFs ($R^2 > 0.7$) we identified the top 3 most frequently selected descriptors. Numbers in parantheses within the features is the frequency. Descriptor groups of each PCF is given in parentheses in the descriptions.

	PC1	PC2	PC3
mTOR signaling pathway	0.45	0.70	0.54
Phosphatidylinositol signaling system	0.76	0.08	0.67
Wnt signaling pathway	0.60	0.81	0.59
Endocytosis	0.67	0.36	0.74
Notch signaling pathway	0.46	0.77	0.35
Neuroactive ligand-receptor interaction	0.60	0.69	0.70
Fructose and mannose metabolism	0.66	0.71	0.55
Starch and sucrose metabolism	0.63	0.55	0.73
Glutathione metabolism	0.73	0.76	0.64
Porphyrin and chlorophyll metabolism	0.62	0.72	0.72
Drug metabolism - cytochrome P450	0.71	0.56	0.55
Glycosphingolipid biosynthesis - globo series	0.57	0.59	0.75
Lysosome	0.64	0.78	0.66
Arachidonic acid metabolism	0.64	0.37	0.73
beta-Alanine metabolism	0.75	0.45	0.50
Propanoate metabolism	0.73	0.49	0.40
Lysine degradation	0.74	0.55	0.50
Fatty acid metabolism	0.78	0.68	0.40
Fatty acid elongation in mitochondria	0.73	0.40	0.12
Citrate cycle (TCA cycle)	0.71	0.57	0.77
Cysteine and methionine metabolism	0.63	0.72	0.42
Glycine, serine and threonine metabolism	0.61	0.74	0.66
Histidine metabolism	0.74	0.66	0.61
Phenylalanine, tyrosine and tryptophan biosynthesis	0.75	0.32	0.41
Aminoacyl-tRNA biosynthesis	0.67	0.61	0.70
One carbon pool by folate	0.66	0.72	0.22
Cyanoamino acid metabolism	0.58	0.73	0.76
Taurine and hypotaurine metabolism	0.52	0.70	0.58
Ubiquinone and other terpenoid-quinone biosynthesis	0.60	0.73	0.53
Nicotinate and nicotinamide metabolism	0.62	0.66	0.73
DNA replication	0.71	0.60	0.73
ABC transporters	0.70	0.37	0.51

Table 4.7: **32 Pathways Linked to PCFs.** The table lists Pathways whose median R^2 across all identified models was > 0.7. Here we show the 32 pathways for which at least one of the PCs reached this goal. For a list of all pathways refer to Table 4.8

4.6.7 Developing a KEGG Pathway Map

In order to visually represent the relationship between the different KEGG pathways we computed a pathway similarity matrix based on the Jaccards index of overlap. This is defined as the ratio between the numbers of genes shared by any two pathways (intersection) divided by

the number of unique genes in the two combined pathways (union). The resulting matrix was used as an input of either a hierarchical clustering procedure (average linkage) (Figure 4.4) or a graph visualization tool Cytoscape [245] using a force driven layout (Figure 4.7). The effectiveness of the clustering procedure in representing the information described by the similarity matrix has been verified using the cophenetic function correlation fit to the input overlap matrix (r=-0.9). Specific KEGG pathways of interest were represented in a cartoon format indicating the direction of change between low and high toxicity chemicals (Figure 4.8 and 4.9). These were defined by applying arbitrary thresholds on the distribution of LC_{50} values. The chemical Bis(2-tehylhexyl)phthalate was eliminated in this categorization due to an intermediate toxicity (Figure 4.10).

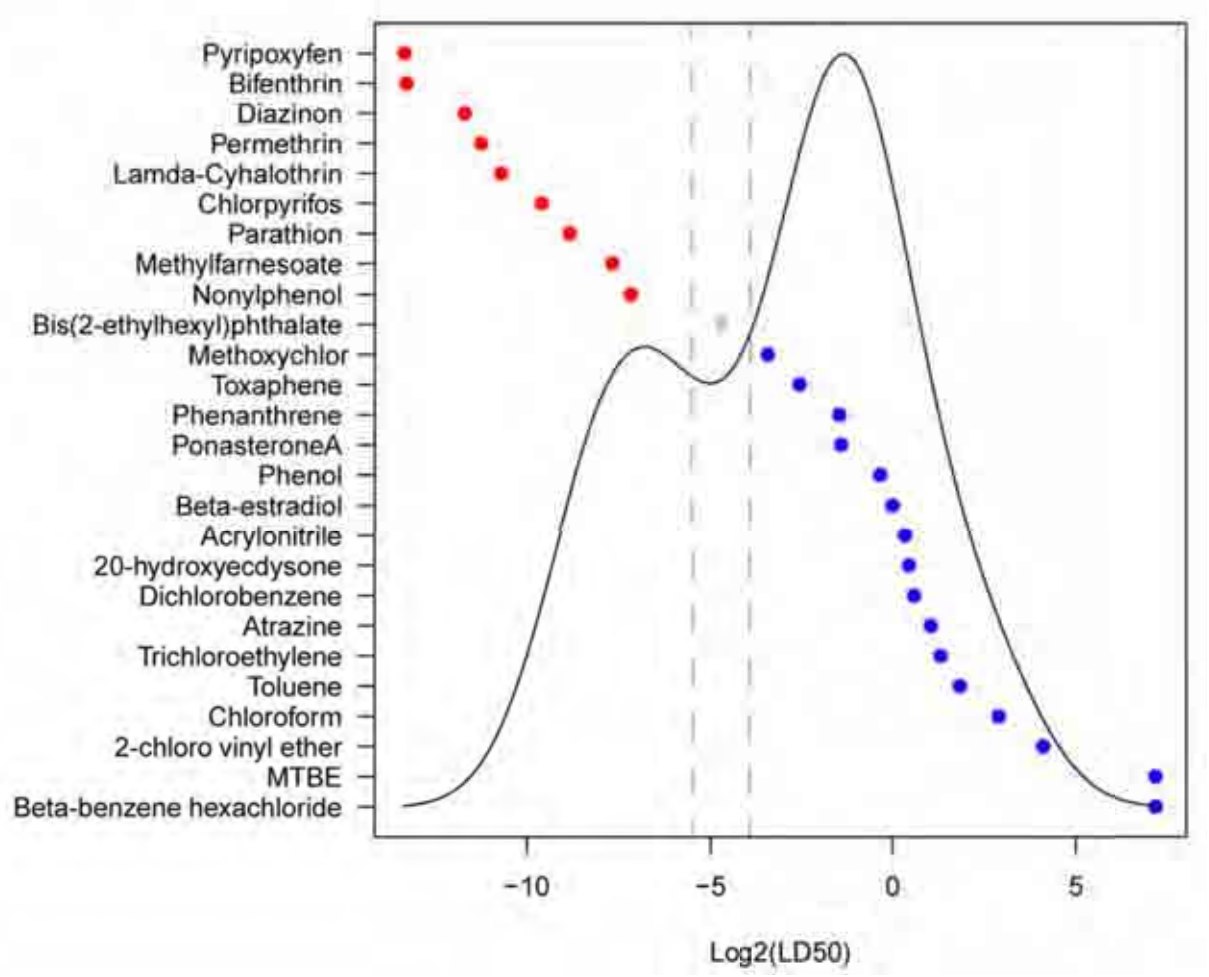


Figure 4.10: **A Graphical Representation of the Toxicity Values Across the Chemical Space.** The figure represents the distribution of log LC_{50} values across all chemicals in the dataset. We used this distribution to define lower and higher toxic chemicals. Higher LC_{50} values (x axis) correspond to a lower toxicity. Regions of high and low toxicity were defined on the basis of the bimodal distribution. Bis(2-ethylhexyl)phthalate was designated outside of both domains as it falls directly in between this classfication.

Table 4.8

	Tox	PCFs (PC1)	PCFs (PC2)	PCFs (PC3)
mTOR signaling pathway	0.02	0.45	0.70	0.54
Regulation of autophagy	0.04	0.61	0.65	0.61
Phosphatidylinositol signaling system	0.13	0.76	0.08	0.67
Inositol phosphate metabolism	0.27	0.60	0.66	0.59
Ribosome	0.08	0.56	0.57	0.54
Hedgehog signaling pathway	0.11	0.62	0.25	0.51
Wnt signaling pathway	0.17	0.60	0.81	0.59
Ubiquitin mediated proteolysis	0.22	0.39	0.42	0.62
Progesterone-mediated oocyte maturation	0.15	0.65	0.56	0.37
TGF-beta signaling pathway	0.25	0.54	0.64	0.69
ECM-receptor interaction	0.13	0.44	0.51	0.51
Endocytosis	0.07	0.67	0.36	0.74
Spliceosome	0.15	0.60	0.70	0.62
Dorso-ventral axis formation	0.14	0.14	0.64	0.66
Notch signaling pathway	0.22	0.46	0.77	0.35
Neuroactive ligand-receptor interaction	0.08	0.60	0.69	0.70
RNA degradation	0.13	0.45	0.36	0.37
Fructose and mannose metabolism	0.15	0.66	0.71	0.55
Amino sugar and nucleotide sugar metabolism	0.10	0.56	0.62	0.60
Galactose metabolism	0.31	0.43	0.59	0.70
Starch and sucrose metabolism	0.07	0.63	0.55	0.73
Pentose phosphate pathway	0.07	0.43	0.64	0.57
Glutathione metabolism	0.19	0.73	0.76	0.64
Pentose and glucuronate interconversions	0.11	0.47	0.56	0.62
Porphyrin and chlorophyll metabolism	0.32	0.62	0.72	0.72
Ascorbate and aldarate metabolism	0.00	0.69	0.69	0.63
Drug metabolism - cytochrome P450	0.11	0.71	0.56	0.55
Metabolism of xenobiotics by cytochrome P450	0.03	0.37	0.53	0.59
Retinol metabolism	0.11	0.63	0.57	0.50
Drug metabolism - other enzymes	0.35	0.63	0.61	0.59
Glycosphingolipid biosynthesis - ganglio series	0.15	0.55	0.61	0.56
Glycosaminoglycan degradation	0.36	0.58	0.67	0.40
Glycosphingolipid biosynthesis - globo series	0.24	0.57	0.59	0.75
Other glycan degradation	0.18	0.58	0.64	0.61
Lysosome	0.15	0.64	0.78	0.66
Sphingolipid metabolism	0.33	0.52	0.49	0.67
Biosynthesis of unsaturated fatty acids	0.03	0.55	0.39	0.68

Continued on Next Page...

Table 4.8 – Continued

	Tox	PCFs (PC1)	PCFs (PC2)	PCFs (PC3)
Peroxisome	0.14	0.14	0.43	0.32
Caffeine metabolism	0.08	0.51	0.67	0.63
alpha-Linolenic acid metabolism	0.12	0.59	0.54	0.37
Ether lipid metabolism	0.19	0.67	0.41	0.65
Linoleic acid metabolism	0.10	0.70	0.43	0.41
Arachidonic acid metabolism	0.21	0.64	0.37	0.73
Glycerophospholipid metabolism	0.12	0.55	0.57	0.46
Limonene and pinene degradation	0.10	0.43	0.61	0.54
beta-Alanine metabolism	0.00	0.75	0.45	0.50
Butanoate metabolism	0.06	0.68	0.62	0.63
Propanoate metabolism	0.03	0.73	0.49	0.40
Valine, leucine and isoleucine degradation	0.00	0.46	0.68	0.37
Tryptophan metabolism	0.21	0.53	0.61	0.64
Lysine degradation	0.08	0.74	0.55	0.50
Fatty acid metabolism	0.09	0.78	0.68	0.40
Fatty acid elongation in mitochondria	0.05	0.73	0.40	0.12
Glycerolipid metabolism	0.18	0.56	0.61	0.56
Glyoxylate and dicarboxylate metabolism	0.17	0.47	0.60	0.61
Citrate cycle (TCA cycle)	0.09	0.71	0.57	0.77
Pyruvate metabolism	0.20	0.48	0.63	0.69
Glycolysis / Gluconeogenesis	0.09	0.46	0.53	0.69
Synthesis and degradation of ketone bodies	0.02	0.49	0.54	0.49
Terpenoid backbone biosynthesis	0.08	0.57	0.66	0.57
Selenoamino acid metabolism	0.15	0.66	0.65	0.49
Cysteine and methionine metabolism	0.15	0.63	0.72	0.42
Glycine, serine and threonine metabolism	0.34	0.61	0.74	0.66
Histidine metabolism	0.21	0.74	0.66	0.61
Tyrosine metabolism	0.19	0.65	0.64	0.61
Phenylalanine, tyrosine and tryptophan biosynthesis	0.11	0.75	0.32	0.41
Phenylalanine metabolism	0.04	0.66	0.58	0.66
Arginine and proline metabolism	0.16	0.65	0.69	0.66
Alanine, aspartate and glutamate metabolism	0.12	0.64	0.51	0.59
D-Glutamine and D-glutamate metabolism	0.00	0.49	0.57	0.56
Nitrogen metabolism	0.17	0.66	0.67	0.64
Valine, leucine and isoleucine biosynthesis	0.25	0.66	0.41	0.57

Continued on Next Page...

Table 4.8 – Continued

	Tox	PCFs (PC1)	PCFs (PC2)	PCFs (PC3)
Aminoacyl-tRNA biosynthesis	0.08	0.67	0.61	0.70
Pantothenate and CoA biosynthesis	0.17	0.38	0.59	0.53
One carbon pool by folate	0.04	0.66	0.72	0.22
Folate biosynthesis	0.04	0.52	0.52	0.45
Cyanoamino acid metabolism	0.19	0.58	0.73	0.76
Taurine and hypotaurine metabolism	0.23	0.52	0.70	0.58
Ubiquinone and other terpenoid-quinone biosynthesis	0.04	0.60	0.73	0.53
N-Glycan biosynthesis	0.17	0.61	0.50	0.65
Glycosphingolipid biosynthesis - lacto and neolacto series	0.13	0.55	0.54	0.65
Fatty acid biosynthesis	0.00	0.63	0.43	0.40
Pyrimidine metabolism	0.17	0.63	0.31	0.67
Purine metabolism	0.05	0.60	0.66	0.45
RNA polymerase	0.30	0.70	0.61	0.56
Nicotinate and nicotinamide metabolism	0.13	0.62	0.66	0.73
Sulfur metabolism	0.29	0.47	0.57	0.07
Nucleotide excision repair	0.02	0.62	0.65	0.01
Mismatch repair	0.06	0.57	0.63	0.49
DNA replication	0.06	0.71	0.60	0.73
Base excision repair	0.06	0.64	0.52	0.23
Homologous recombination	0.20	0.57	0.44	0.57
Oxidative phosphorylation	0.15	0.37	0.62	0.38
Jak-STAT signaling pathway	0.11	0.68	0.63	0.38
Steroid biosynthesis	0.42	0.63	0.59	0.67
ABC transporters	0.16	0.70	0.37	0.51
O-Glycan biosynthesis	0.15	0.59	0.59	0.27
SNARE interactions in vesicular transport	0.05	0.60	0.58	0.48
Proteasome	0.04	0.69	0.42	0.58
Basal transcription factors	0.04	0.45	0.52	0.62
Protein export	0.01	0.50	0.58	0.73

Table 4.8: **All R^2 Values of Pathways Associated to Toxicity and PCFs.** Here the results for all pathways are shown. First column represents the results for the association to toxicity. The threshold for 1% FDR is 0.1713. Columns 2 – 4 are representative of the association to PCFs of PC1, PC2 and PC3 respectively. Here the median R^2 across all models is shown. The threshold for association to PCFs was chosen to be $R^2 > 0.7$.

Chapter 5

Application of Reverse Engineering in Ecotoxicology

5.1 Abstract

The mechanisms of action of many chemicals of environmental concern are either unknown or incompletely characterized. Omics technologies have provided a high-throughput unbiased approach to address this issue. Statistical modelling approaches, identifying groups of features predictive of toxicity outcome have shown to provide informative results, especially when pathway knowledge is incorporated. This approach, however, is limited to the gene to gene interactions represented by the functional annotation. In this context, reverse engineering methodologies have provided means of inferring the underlying regulatory network without prior knowledge. Identification of functional modules linked to physiological outcome can then aid in characterizing adverse outcome pathways. Here we demonstrate the application of these approaches to a large compendium of fathead minnow (FHM, *Pimephales promelas*) microarrays developed by the U.S. Army Engineering Research and Development Center. We focus our efforts on flutamide (FLU), an anti-androgen, to further characterize its effect. We applied a well validated reverse engineering methodology, ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks), to infer the underlying regulatory network and identified highly interconnected sub-networks. Our results provide evidence that FLU acts through a novel

androgen-receptor independent pathway in the fathead minnow.

5.2 Introduction

Environmental protection agencies worldwide are presented with a huge challenge as a result of human innovation. Large numbers of chemicals are being released into the environment through chemical spills, sewage, industrial waste or agricultural run-offs such as excess nutrients or pesticides. Many underlying mechanisms for toxicologically relevant compounds have been characterized in the literature. There are, however, several instances where the underlying mechanisms for specific responses are mediated through currently unknown or incompletely characterized toxicity pathways. In chapters 3 and 4 we have shown that machine learning techniques and pathway level analyses provide very powerful tools for identifying adverse outcome pathways. However, these approaches do not implicitly infer the structure of the underlying regulatory networks. In this context, network inference approaches, originally developed and validated in other biological systems have the potential to address this issue [261–264].

This chapter will focus on demonstrating the potential of these approaches in the context of a very relevant ecotoxicological system. This is the response of the fathead minnow (FHM) to FLU, a model anti-androgen compound. This pharmaceutical is currently used to treat prostate cancer as it specifically competes with testosterone for binding the androgen receptor (AR) in the prostate gland, essentially reducing cancer cell growth [265]. In fish, FLU is assumed to act via the same mechanism of action [266–268]. More specifically, exposure to FLU in early development of guppies resulted in demasculinization [266], an anti-androgen response also observed in mammalian studies [269]. Reduced fecundity, a decrease in mature oocytes in females and degeneration of spermatocytes in addition to necrosis have also been observed in FHM [268]. However the specific binding of FLU to fish AR(s) has revealed conflicting results [270]. Ankley et al [271] suggested that the reason for these conflicting results was in part due to the design of the experiment where binding of FLU was measured but its metabolite, 2-hydroxy-flutamide, was neglected. This metabolite has been shown to have a strong affinity to both fish and mammalian AR(s) [271].

To address whether the transcriptional response to FLU is consistent with an AR mediated mechanism in FHMs, Garcia-Reyero et al [202] compared the response of fish treated with FLU to fish treated with the model androgen 17β-trenbolone (TB). If FLU effects are mediated by the canonical mechanism, exposure to these two compounds should result in opposite effects in respect to gene expression. Interestingly a direct comparison on the gene level identified only 70 genes that were regulated reciprocally between the two exposures. The authors also performed a QPCR analysis on specific genes involved in steroidogenesis. The remaining 1351 differentially expressed genes did not follow the paradigm. Overall this work suggested that in addition to its anti-androgen activity, FLU might act via a still uncharacterized yet important mechanism.

To address this challenge we utilized a network inference approach based on an information theoretical approach [98, 272]. The methodology was applied to, a large compendium of over 800 microarrays representing the transcriptional response of FHM ovaries to exposure of endocrine disruptors (Table 5.1). This dataset, provided by the U.S. Army Engineering Research and Development Centre is ideally suited for reverse engineering [273].

This proof of concept study proved the effectiveness of the approach and provided evidence of a novel AR independent pathway, linked to ovary development and perturbed by FLU exposure in FHM ovaries.

5.3 Results

5.3.1 A Compendium of Gene Expression Profiling Experiments Representing *in vivo* and *in vitro* Response of the Fathead Minnow Ovary to Endocrine Disruptors

The work described in this chapter relies on a large compendium of microarray and hormone measurements in the FHM, made available by Dr. Ed Perkins (U.S. Army Engineer Research and Development Center). Here we describe this dataset in some detail to facilitate the understanding of our modelling effort. This compendium includes transcriptomics data generated in

Chemical experiment	*in-vivo* / *in-vitro*	Time	Concentration	conditions no controls	conditions + controls	total Arrays #
Fadrozole	*in-vivo*	30min, 1, 2, 4, 6 hr	5, 50 uM	10	15	60
Fadrozole	*in-vivo*	6, 12, 24 hr	50 um	3	6	24
Fadrozole	*in-vitro*	0, 1, 2, 3, 4, 6, 8, 10, 12 hr	50 uM	8	16	64
Fadrozole	*in-vivo*	exposed (1, 2, 4, 8d), recovery (1, 2 ,4, 8d)	3,30 uM	16	24	182
Flutamide	*in-vivo*	1, 2, 4, 8, 12h	500 ug/l	5	10	39
Ketoconazole	*in-vitro*	2, 4, 6, 8, 10, 12 hr	0.5 uM, 5 pools of 5	6	12	52
Ketoconazole	*in-vitro*	15, 30, 45, 60, 75, 90, 105, 120, 135, 150min	0.5 uM	10	20	84
Ketoconazole	*in-vivo*	6, 12, 24 hr	0.5 uM	3	6	24
Stages	*in-vivo*	NA	Pre-vitellogenic, Vitellogenic, Mature ovary, Ovulated eggs, Atretic	5	NA	23
Prochloraz	*in-vitro*	2, 4, 6, 8, 10, 12 hr	2.5 uM, 5 pools of 5	6	12	53
Prochloraz	*in-vivo*	6, 12, 24 hr	2.5 uM	3	6	22
RDX	*in-vivo*	1, 21d	5 mg/L	2	4	22
TNT	*in-vivo*	30min, 1, 2, 4, 6, 24 hr	5 mg/L	6	12	60
Trenbolone	*in-vivo*	exposed (1, 2, 4, 8d), recovery (1, 2 ,4, 8d)	low, high dose	16	3	162
Total				99	146	871

Table 5.1: **Overview of the FHM Dataset.**

14 separate FHM experiments involving 7 different chemical stressors. All experiments were continuous flow-through exposures. Chemicals were delivered in UV-treated, 0.4um filtered Lake Superior water (LSW) without the use of carrier solvents. Nine experiments were acute

time course studies involving a minimum of three different exposure durations, each less than 24h. Generally a single concentration per chemical was used with the exception of fadrazole. Only female FHMs were exposed with a minimum of 8 fish per treatment group with a minimum of 2 exposure replicates were performed. Fish were added after the desired concentration has been verified and replicates were staggered to ensure collection within minutes of the intended time point. In addition to these, two exposure and recovery time-course experiments were performed. Both experiments were run for 8 days followed by a recovery phase in clean water of similar time length using two different concentrations. The remaining two experiments followed a slightly different approach. One of these specifically focused on profiling the different ovary stages based on histological examination [203] while the other investigated a 21 day reproduction effect to RDX (cyclotrimethylenetrinitramine).

5.3.2 Estimation of Gene to Gene Connections: A Comparison of Different MI Estimation Methods Using minet

The reverse engineering approach we have chosen is a modification of the well-validated ARACNE methodology. Each gene-to-gene connection is inferred by using a measure of variable dependency called mutual information (MI) (see methods section for a formal definition). There are a number of MI estimation methods available, which differ by the way data is processed (discretization) and in the MI estimation procedure itself to reverse engineer the underlying regulatory network. Several of these are implemented in the minet package available in R [234]. A detailed description is given in the material and methods section. Therefore we first set out to identify which of the combinations of data discretization and MI estimation may be more appropriate for the analysis of this dataset. For each estimation procedure linked to a discretization method (Shrink entropy EqualWidth (SEW), Shrink entroy EqualFreq (SEF), Empirical EqualWidth (EEW), Miller-Madow EqualWidth (MMW) and Schurmann-Grassberger EqualWidth (SGW)) we plotted pearson correlation against the calculated MI values to be able to visually inspect the results of the different estimators. In order to select the method of choice we used the criteria that the diagnostic plot should best match the theoretical relationship between pearson correlation (R) and MI (M) (Equation 5.1 and Figure 5.1 [274, 275]):

$$M(X, Y) = -log\sqrt{1 - R(X, Y)^2} \tag{5.1}$$

We were able to disregard three of the five tested approaches (Figure 5.9). More specifically SEF, EEW and SGW all inflated MI values where no linear relationship was identified. SEF in particular inflated all MI values across the sample space. The remaining two approaches SEW and MMW both produced comparable results. Referring to the package documentation; although the default estimation routine is the empirical estimator the authors mention that the Miller-Madow Corrected Estimator reduces bias generated by the nave approach. This led us to choose the MMW approach for network inference. Calculation of MI values generated from a randomized dataset enabled the calculation of an FDR threshold (Figure 5.2). In order to extract

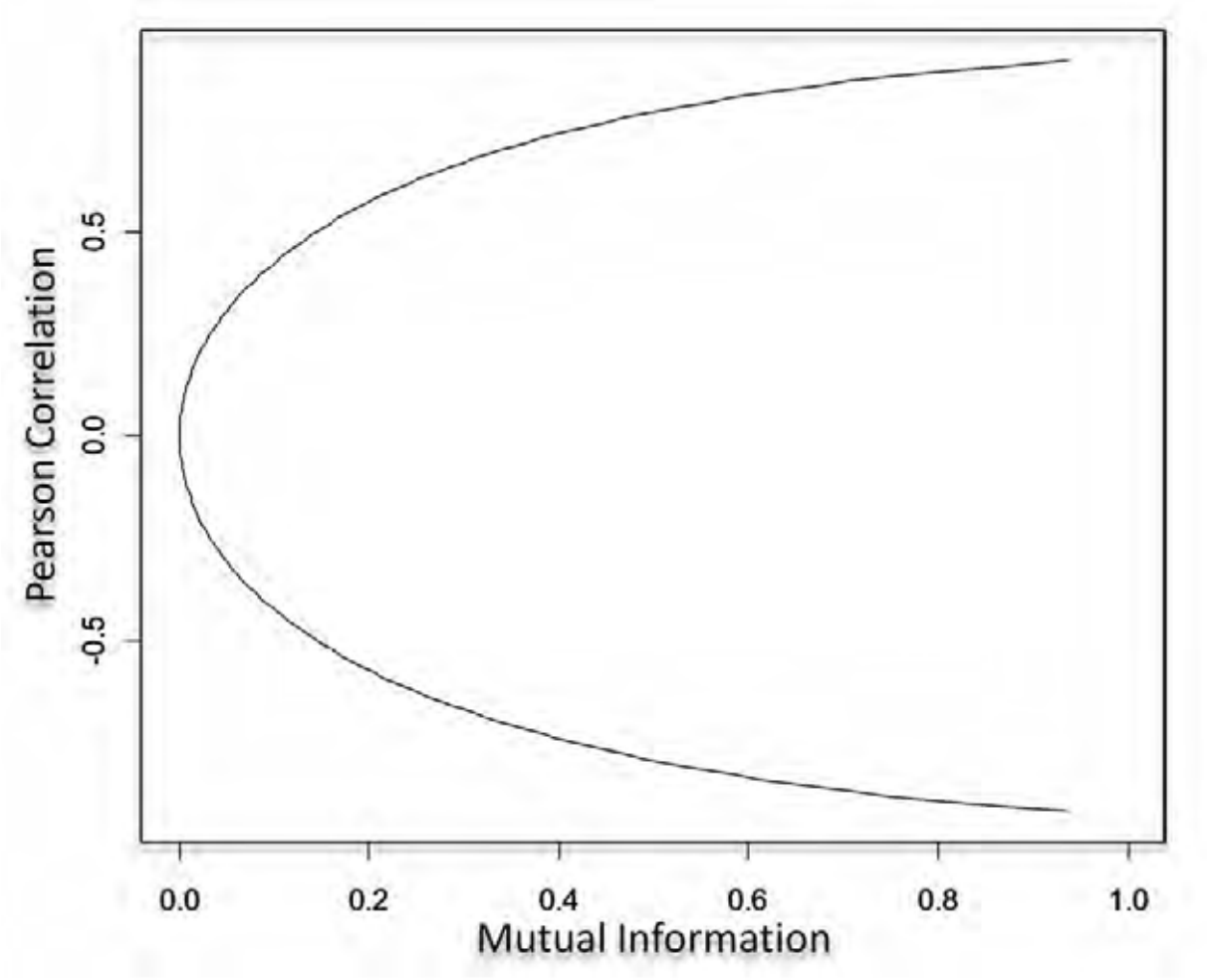


Figure 5.1: **Theoretical Distribution between Mutual Information and Pearson Correlation.** When the distribution of the Pearson correlation (y axis) in a given dataset is Gaussian like, the relationship between Pearson Correlation and Mutual Information (x axis) should follow this theoretical graph (Equation 5.1, [274, 275]).

only the most significant connections we applied a conservative threshold, which exceeded the calculated 1% FDR value (Figure 5.2).

5.3.3 Inference of a Regulatory Network Representing the Receptor Neighbourhood in Fathead Minnow

Having identified the appropriate network inference approach we set out to develop a gene to gene interaction network representative of FHM ovaries. In order to maximize biological interpretability we focus on reconstructing the network as a union of the neighbourhood of biologically important hubs. We choose to focus on genes that code for receptors and in-

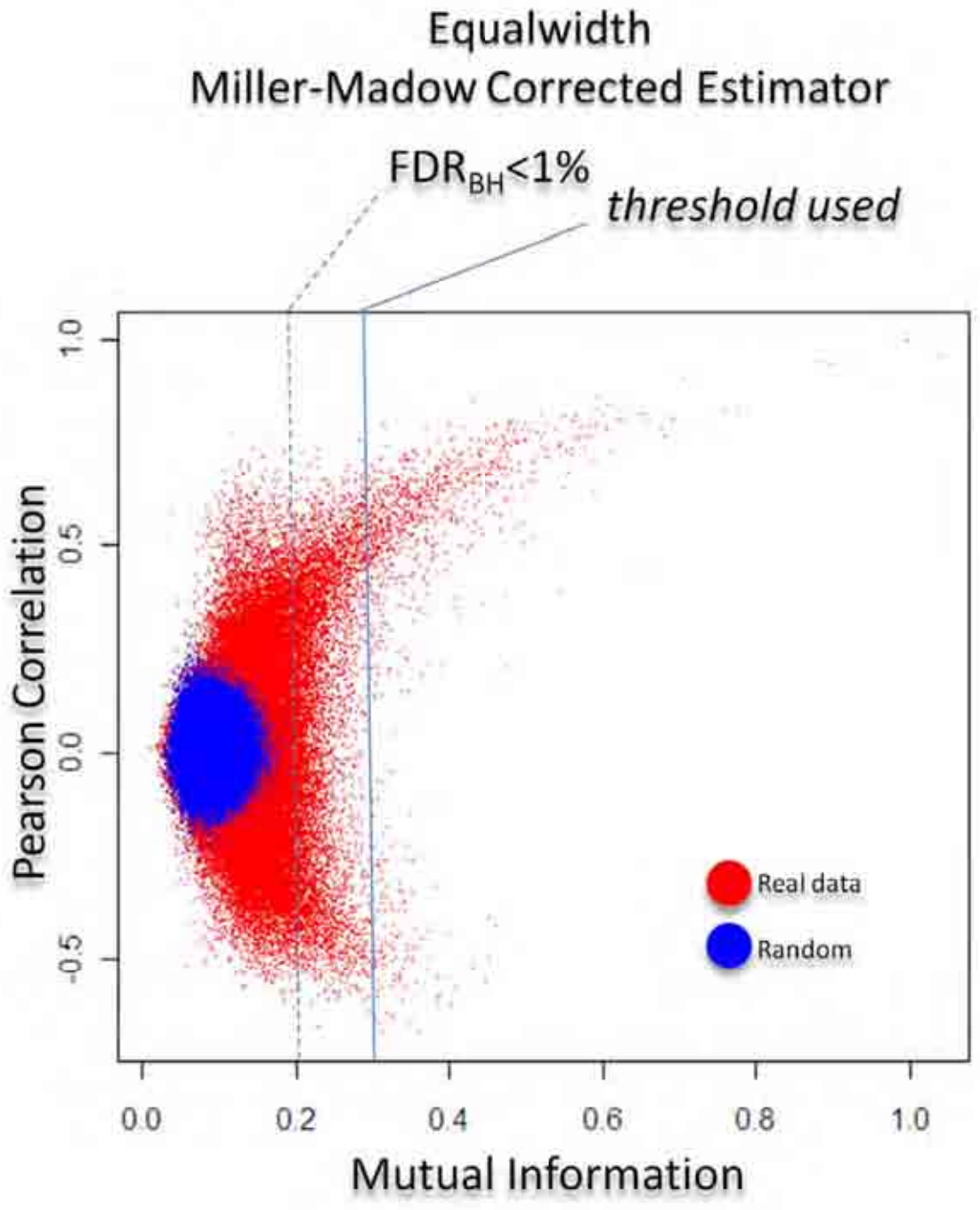


Figure 5.2: **Distribution of MI values in Relation to Pearson Correlation for this Dataset.** The MI values from the Miller-Madow Corrected Estimator best fit the theoretical distribution. From the distribution of Random MI (blue dots) values a FDR threshold can be calculated. The final threshold that was used for building the network is shown in the figure by a continuous line and exceeds the calculated 1% FDR threshold.

cluded the physiological measurements (see methods section for a detailed description of the procedure). The resulting network was then visualized within cytoscape. We then searched for naturally occurring highly connected network modules using the network modularization procedure MCODE [276]. The procedure identified 6 modules within our network linked to specific biological process of potential relevance for ovary biology (Figure 5.3A). We were able to functionally annotate four of these modules (Figure 5.3A). Furthermore we highlighted the genes directly connected to testosterone (Figure 5.3B, yellow), serum vitellogenin (Figure 5.3B, blue) and the gene receptors (Figure 5.3B, red). A more detailed analysis of the neighbourhood of testosterone (defined by connections with $MI > 0.47$, Figure 5.4), revealed a number of relationships supported by the literature. The direct interaction between the oestrogen receptor and testosterone [277] and the link between testosterone levels and aryl hydrocarbon (or dioxin) receptor (AhR) [278] are the two most relevant. Furthermore, opioid receptors have also been linked to pain tolerance based on testosterone and oestrogen levels in male and female rats [279]. The fact that our results reflect known regulatory networks in respect to serum testosterone validates our approach and provides confidence in further analysis of the rest of the inferred network.

5.3.4 The Transcriptional Response to Flutamide is Largely Independent of Testosterone Activity

Having developed a network model and mapped the genes linked to testosterone we set to address the key question whether FLU activity could be related to a non-testosterone dependent mechanism. We reasoned that if this original hypothesis would be correct genes differentially regulated by FLU would map in a different region of the network, far from the testosterone target genes. We discovered that the 179, differentially expressed genes (67 up, 112 down-regulated) were only marginally linked to the testosterone neighbourhood (Figure 5.5). In fact only 12 genes were in common between the two gene lists (566 genes linked to testosterone). Moreover, these genes clustered in close proximity to the light-blue module shown in Figure 5.3A, which was enriched in Gene Ontology terms for translation, inflammatory response

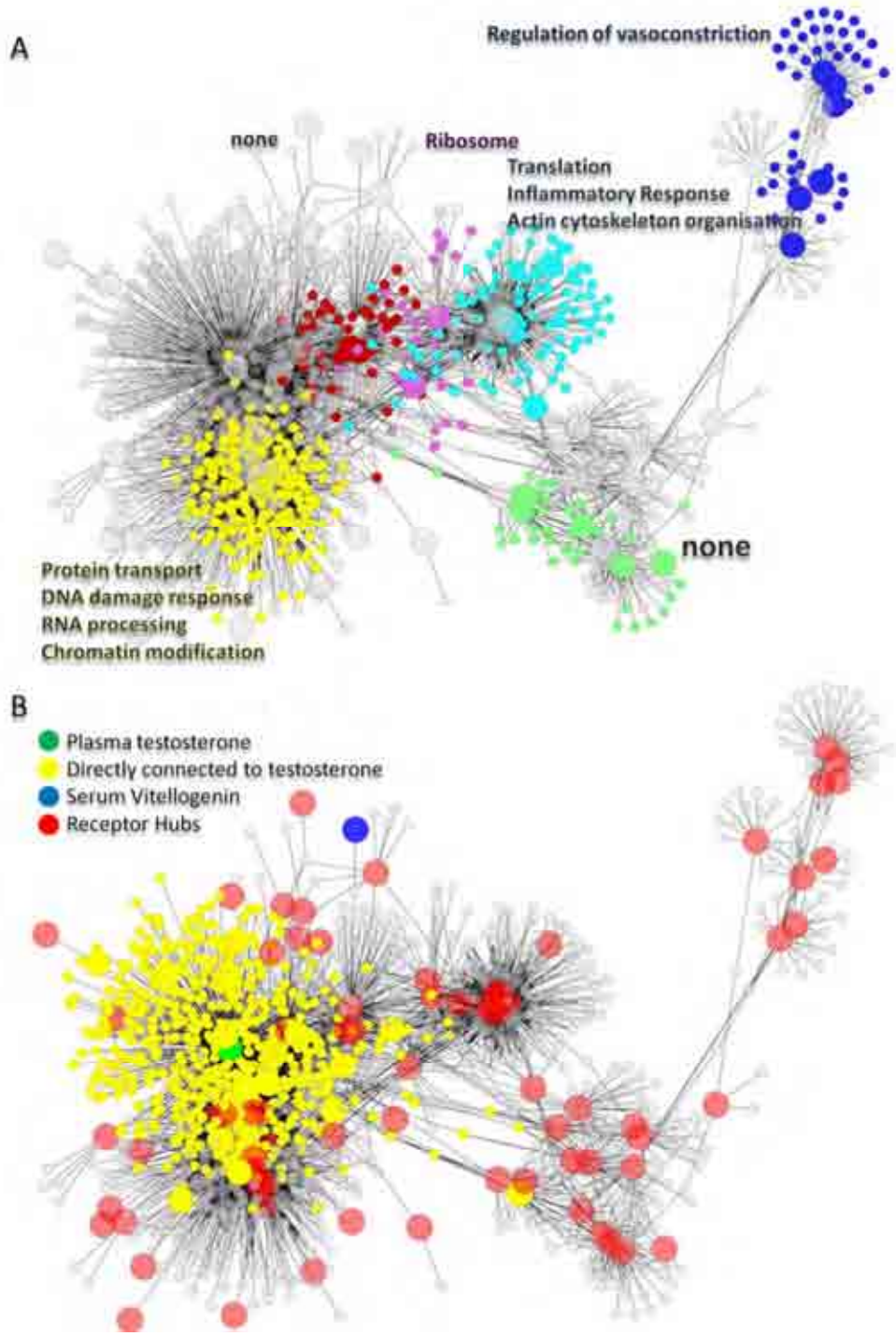


Figure 5.3: **The Network Derived using the Algorithm for the Reconstruction of Accurate Cellular Networks.** A) Colour overlay of six modules present in the overall network. Cyan and yellow modules were significantly enriched for Gene Ontrology biological processes. B) Illustration that hub genes tend to be receptor genes (large red nodes) distributed throughout the network. A large cluster of genes (yellow nodes) are linked to serum testosterone (green node), whereas serum vitellogenin (blue) has few linked nodes.

and actin cytoskeleton organisation. This provided further evidence that FLU does not act directly through a testosterone mediated pathway.

5.3.5 Molecular Networks Involved in Ovary Development in Fathead Minnow

Our work has shown that impact of FLU is not directly associated to testosterone, which begs the question what biological process FLU actually interferes with. In order to address this question we wondered which component, of our network, was enriched in genes regulated during ovary development and atresia. We therefore first identified 71 genes differentially expressed in during ovary development (Figure 5.6, Table 5.2). We then asked the question where in our inferred network these genes would map (Figure 5.7).

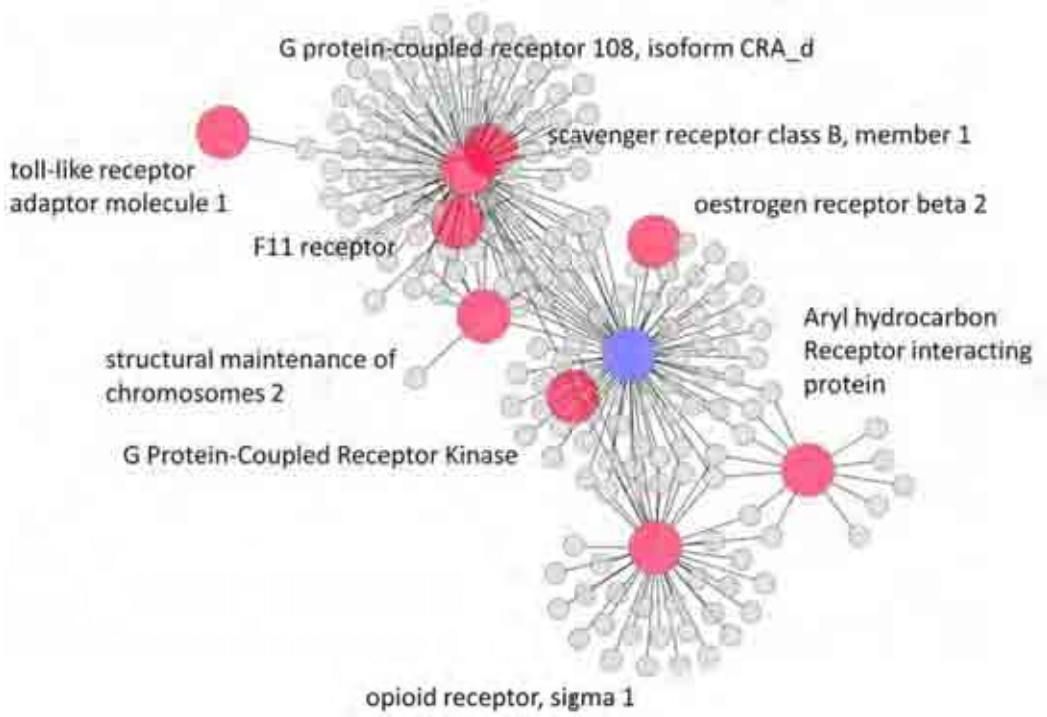


Figure 5.4: **Identifying the Neighbourhood of Testosterone.** Sub-network Representing the Relationship between Oestrogen Receptor, Aryl Hydrocarbon (dioxin) Receptor and Opioid Receptor Expression.

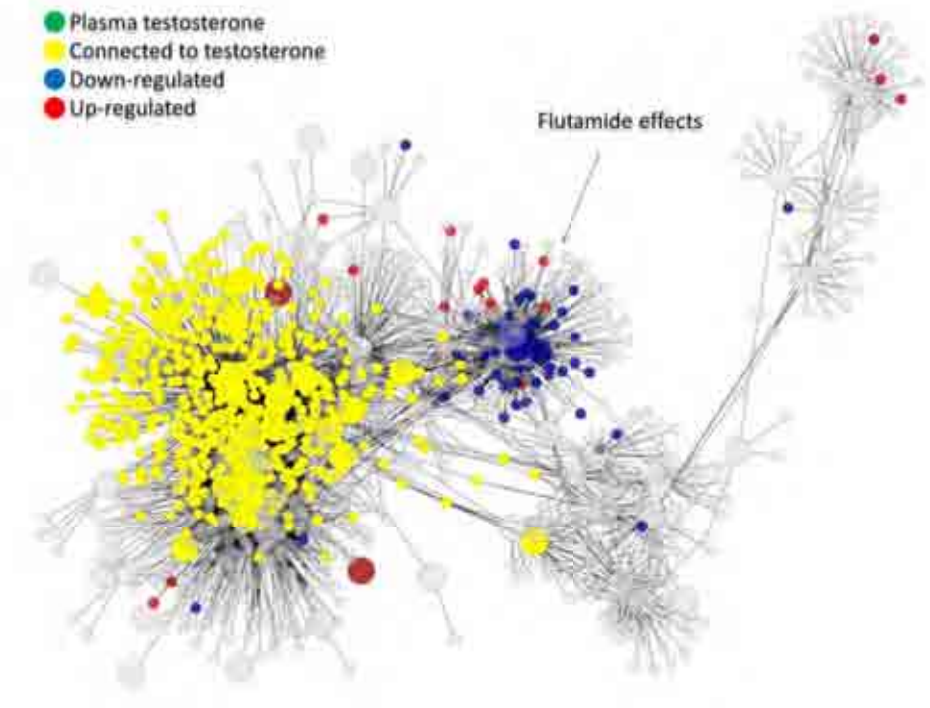


Figure 5.5: **Mapping the Transcriptional Response to Flutamide Exposure.** Genes modulated by FLU exposure were mapped onto the overall ARACNE network. The genes cluster in proximity of module 4 (light-blue in Figure 5.3A) enriched in the GO terms anatomical structure development, cell motility, and inflammatory response ($ES > 1.5$).

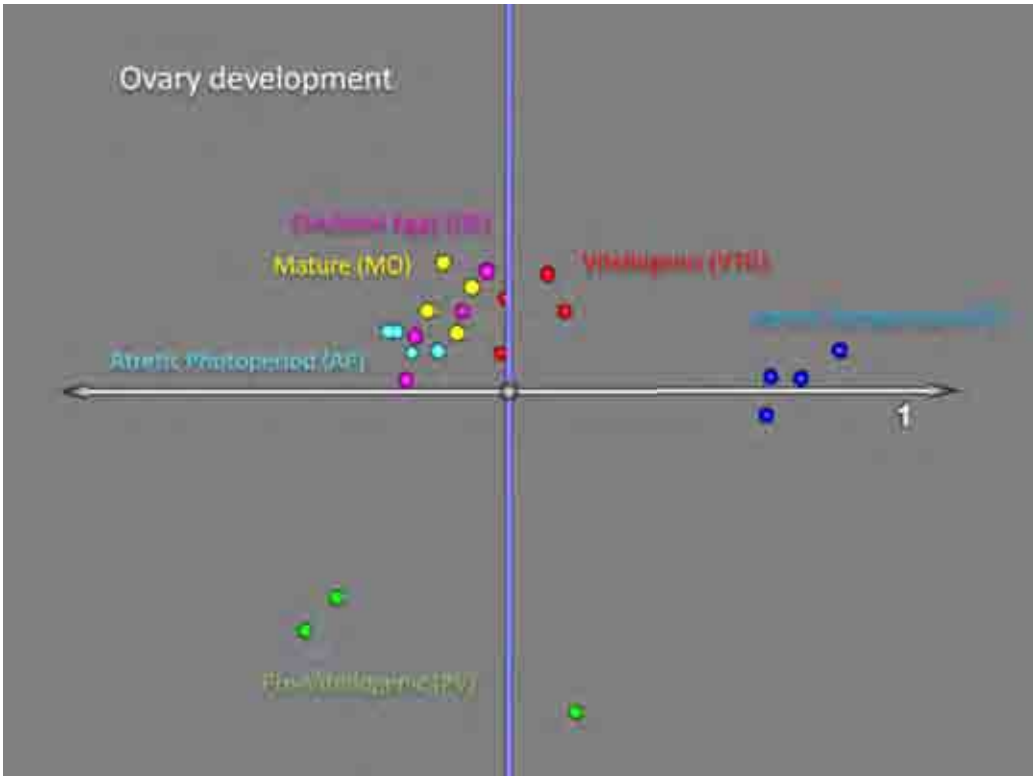


Figure 5.6: **PCA of the Different Ovary Stages in FHM.** At 10% FDR only 71 genes were identified differentially expressed in at least one of the stages. The AT and PV stages are the most diverse in this dataset. Initially PV stages must adapt to the VTG stage, here signified by a change in PC2. From the VTG stage there are two options, mature or follow the artretic pathway. Artresion can occur at any stage during maturation and hence gene expression changes up to OE (VTG, MO and AP) are only minimal. Lastly conclusion of the artretic cycle requires a large change in transcription, here signified by the change in PC1.

Table 5.2

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AF_102714	LOC571170	NP_001030165	NM_001034993	Danio rerio Indian hedgehog homolog a (ihha), mRNA	NP_000184
UF_Ppr_AF_107602		BAA92179	AB031424	Cyprinus carpio mRNA for CD45, complete cds	-
UF_Ppr_AF_110174		YP_001304555	BC083534	Danio rerio amyloid beta (A4) precursor protein b, mRNA (cDNA clone MGC:92771 IMAGE:7086881), complete cds	-
UF_Ppr_AF_108239	smyhc1	ABW87635	XR_029314	PREDICTED: Danio rerio similar to slow myosin heavy chain 1 (LOC100001366), mRNA	NP_000248
UF_Ppr_AF_113874	fn1b	AAU14809	AY725818	Danio rerio fibronectin 1b (fn1b) mRNA, complete cds	NP_997643
UF_Ppr_AF_115536	LOC565021	CAP08005	BR000041	TPA: TPA_inf: Danio rerio SN4TDR gene for 4SNc-Tudor domain protein, complete cds	-
UF_Ppr_AF_112364	nr2f1	NP_571255	BC056574	Danio rerio nuclear receptor subfamily 2, group F, member 1, mRNA (cDNA clone MGC:65769 IMAGE:6800668), complete cds	NP_005645
UF_Ppr_AF_114711		NP_001070182	BC115104	Danio rerio hypothetical protein LOC553426, mRNA (cDNA clone IMAGE:7434433), partial cds	NP_067544
UF_Ppr_AF_111699		EDL17202	AB034198	Carassius auratus mRNA for lamin B2, complete cds	-
UF_Ppr_AF_107084		AAP45037	NM_001024109	Danio rerio protocadherin 1 gamma 9 (pcdh1g9), mRNA	-

Continued on Next Page...

Table 5.2 – Continued

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AF_101189	myhz2	NP_694514	AB231798	Cyprinus carpio MYH emb1 mRNA for myosin heavy chain embryonic type 1, complete cds	NP_060003
UF_Ppr_AF_109393	igf1	AAT02176	AY533140	Pimephales promelas insulin-like growth factor-I mRNA, complete cds	NP_000609
UF_Ppr_AF_108530	atp1b1b	NP_571746	BC071293	Danio rerio ATPase, Na+/K+ transporting, beta 1b polypeptide, mRNA (cDNA clone MGC:86581 IMAGE:6896300), complete cds	NP_001668
UF_Ppr_AF_111501	dhcr7	NP_958487	NM_201330	Danio rerio 7-dehydrocholesterol reductase (dhcr7), mRNA	NP_001351
UF_Ppr_AF_113974	zgc:109896	NP_001018353	XR_029478	PREDICTED: Danio rerio similar to collapsin response mediator protein 2 (LOC798555), mRNA	NP_001377
UF_Ppr_AF_112335	LOC563708	NP_001076305	NM_001082836	Danio rerio integrin, beta 5 (itgb5), mRNA	NP_002204
UF_Ppr_AF_116141	wu:fc25c04	AAH80223	DQ317971	Danio rerio follistatin-like 2 mRNA, complete cds	NP_009016
UF_Ppr_AF_106736	prdm1	NP_955809	DQ851841	Danio rerio PR domain containing 1c (prdm1c) mRNA, partial cds	NP_001189
UF_Ppr_AF_105852	zgc:113032	NP_001038671	NM_001045206	Danio rerio si:dkeyp-119b4.5 (si:dkeyp-119b4.5), mRNA	NP_056344
UF_Ppr_AF_112702		Q04956	BX649389	Zebrafish DNA sequence from clone DKEY-28D3 in linkage group 1, complete sequence	-
UF_Ppr_AF_111504	dlst	AAH65943	BC065943	Danio rerio dihydrolipoamide S-succinyltransferase, mRNA (cDNA clone MGC:77238 IMAGE:6963253), complete cds	NP_001924

Continued on Next Page...

Table 5.2 – Continued

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AF_104873	ela2	AAH42328	BC042328	Danio rerio elastase 2, mRNA (cDNA clone IMAGE:3817357), partial cds	NP_254275
UF_Ppr_AF_109618	paf1l	NP_001019619	NM_001024448	Danio rerio myosin, heavy polypeptide 11, smooth muscle (myh11), mRNA	-
UF_Ppr_AF_117146	LOC562957	AAI33924	XM_686324	PREDICTED: Danio rerio hypothetical LOC562957 (LOC562957), mRNA	NP_077304
UF_Ppr_AF_112068	nqo1	NP_991105	NM_205542	Danio rerio NAD(P)H dehydrogenase, quinone 1 (nqo1), mRNA	NP_000894
UF_Ppr_AF_102176		AAI42762	BC142761	Danio rerio sulfotransferase family 2, cytosolic sulfotransferase 2, mRNA (cDNA clone MGC:165385 IMAGE:8156282), complete cds	NP_814444
UF_Ppr_AF_114968		AAI41835	CP000647	Klebsiella pneumoniae subsp. pneumoniae MGH 78578, complete sequence	NP_001034795
UF_Ppr_AF_106128	vox	AAH92695	AC144824	Danio rerio clone CH211-172M22, complete sequence	-
UF_Ppr_AF_100438	nol5a	AAT68132	BC075769	Danio rerio nucleolar protein 5A, mRNA (cDNA clone IMAGE:6900431), partial cds	NP_006383
UF_Ppr_AF_109131	zgc:113111	NP_001013469	AL954132	Zebrafish DNA sequence from clone DKEY-32N7, complete sequence	-
UF_Ppr_AM_119089		AAR22965	NM_213362	Danio rerio epsin 1 (epn1), mRNA	-
UF_Ppr_AF_114278	zgc:103600	NP_001103338	XM_688236	PREDICTED: Danio rerio hypothetical LOC564918 (LOC564918), mRNA	-
UF_Ppr_AF_102006	zgc:101812	NP_001004619	NM_001004619	Danio rerio serine/threonine/tyrosine interacting-like 1 (styxl1), mRNA	NP_057170

Continued on Next Page...

Table 5.2 – Continued

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AF_101031	LOC555286	BAA34528	XR_030049	PREDICTED: Danio rerio similar to Thymus high mobility group box protein TOX (LOC569666), mRNA	NP_055544
UF_Ppr_AF_109600	flot1b	NP_958864	NM_201456	Danio rerio flotillin 1b (flot1b), mRNA	NP_005794
UF_Ppr_AF_100277	derl1	NP_998609	NM_213444	Danio rerio Der1-like domain family, member 1 (derl1), mRNA	NP_077271
UF_Ppr_AF_112628	lcp1	NP_571395	NM_131320	Danio rerio lymphocyte cytosolic plastin 1 (lcp1), mRNA	NP_002289
UF_Ppr_AF_107119	dlg7	NP_001004592	NM_001004592	Danio rerio discs, large homolog 7 (Drosophila) (dlg7), mRNA	-
UF_Ppr_AF_100861	ube1	NP_998227	AB035495	Carassius auratus mRNA for ubiquitin-activating enzyme E1, complete cds	NP_003325
UF_Ppr_AF_118166	gna12l	AAR25616	BC133071	Danio rerio guanine nucleotide binding protein (G protein) alpha 12, mRNA (cDNA clone MGC:158144 IMAGE:6963705), complete cds	NP_031379
UF_Ppr_AF_115320	zgc:56589	NP_957241	XM_001331559	PREDICTED: Danio rerio hypothetical protein LOC791759 (LOC791759), mRNA	NP_002632
UF_Ppr_AF_100723	FBXO9	NP_956012	BC076528	Danio rerio F-box protein 9, mRNA (cDNA clone MGC:92017 IMAGE:7043937), complete cds	NP_258441
UF_Ppr_AF_113631		ZP_01834647	BX072578	Zebrafish DNA sequence from clone CH211-286A10 in linkage group 5, complete sequence	-

Continued on Next Page...

Table 5.2 – Continued

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AM_119860		BAC05810	BC139644	Danio rerio zgc:136268, mRNA (cDNA clone MGC:162844 IMAGE:2640944), complete cds	-
UF_Ppr_AF_104808	zgc:63716	NP_956141	NM_199847	Danio rerio FXYD domain containing ion transport regulator 6 (fxyd6), mRNA	NP_071286
UF_Ppr_AF_107159	atp1a1b	P25489	BC085663	Danio rerio ATPase, Na+/K+ transporting, alpha 1b polypeptide, mRNA (cDNA clone MGC:92351 IMAGE:7055852), complete cds	NP_000692
UF_Ppr_AF_101905	slc9a3r2	AAH64290	BC064290	Danio rerio solute carrier family 9 (sodium/hydrogen exchanger), isoform 3 regulatory factor 2, mRNA (cDNA clone MGC:77629 IMAGE:6996791), complete cds	NP_004243
UF_Ppr_AF_118999	sb:cb36	NP_001093576	NM_001100106	Danio rerio Fc receptor, IgE, high affinity I, gamma polypeptide (fcer1g), mRNA	NP_004097
UF_Ppr_AF_102886	zgc:114107	NP_001025433	NM_001030262	Danio rerio zgc:110411 (zgc:110411), mRNA	NP_001113
UF_Ppr_AF_101389	cplx2	NP_001002459	NM_001002459	Danio rerio complexin 2 (cplx2), mRNA	NP_006642
UF_Ppr_AF_109406		YP_001253881	CR478288	Zebrafish DNA sequence from clone DKEY-118K5 in linkage group 15, complete sequence	-
UF_Ppr_AF_112259		YP_356790	NM_131465	Danio rerio l-isoaspartyl protein carboxyl methyltransferase (pcmt), mRNA	-
UF_Ppr_AF_104866	zgc:100973	NP_001002740	NM_001002740	Danio rerio elongation factor-2 kinase (eef2k), mRNA	NP_037434
UF_Ppr_AF_102888	csenl	NP_957113	NM_200819	Danio rerio Kv channel interacting protein 3, calsenilin, like (kcnip3l), mRNA	NP_038462

Continued on Next Page...

Table 5.2 – Continued

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AF_116319		AAI52512	NM_001083050	Danio rerio family with sequence similarity 82, member C (fam82c), mRNA	NP_060615
UF_Ppr_AF_101077	LOC567275	ABD67515	DQ411318	Cyprinus carpio microsomal glutathione S-transferase 3 mRNA, complete cds	NP_004519
UF_Ppr_AF_117736		NP_519492	CT943666	Zebrafish DNA sequence from clone CH73-83C16 in linkage group 18, complete sequence	-
UF_Ppr_AF_112920		NP_001095148	NM_001101678	Danio rerio lysosomal-associated protein transmembrane 4 alpha (LOC100003844), mRNA	NP_055528
UF_Ppr_AF_112481	zgc:86749	NP_999945	XM_001331734	PREDICTED: Danio rerio hypothetical protein LOC791970 (LOC791970), mRNA	NP_006414
UF_Ppr_AF_103711	gclm	NP_956139	NM_199845	Danio rerio glutamate-cysteine ligase, modifier subunit (gclm), mRNA	NP_002052
UF_Ppr_AF_113419	foxd3	AAH95603	BC095603	Danio rerio forkhead box D3, mRNA (cDNA clone MGC:111934 IMAGE:7432677), complete cds	NP_036315
UF_Ppr_AF_105534	Trit1	AAI35063	NM_001044774	Danio rerio si:ch211-194e15.1 (si:ch211-194e15.1), mRNA	NP_060116
UF_Ppr_AF_102915	LOC560753	NP_705954	BC152271	Danio rerio triosephosphate isomerase 1b, mRNA (cDNA clone MGC:174772 IMAGE:7176806), complete cds	NP_000356
UF_Ppr_AF_118851	etr1	EDL38737	XM_688633	PREDICTED: Danio rerio similar to Ribonucleoprotein (LOC565354), mRNA	NP_009116

Continued on Next Page...

Table 5.2 – Continued

Probe ID	Gene Symbol	NR Accession	NT Accession	NT Description	Human Homolog RefSeq
UF_Ppr_AF_108780		EAY57363	BX908726	Zebrafish DNA sequence from clone DKEY-74I2 in linkage group 22 Contains the 5 end of the gene for a novel protein similar to vertebrate aryl hydrocarbon receptor family, a novel gene, the ahr2 gene for aryl hydrocarbon receptor 2 and a CpG island, compl	-
UF_Ppr_AF_101551		CAL51952	XM_001331694	PREDICTED: Danio rerio hypothetical protein LOC791615 (LOC791615), mRNA	-
UF_Ppr_AF_107108	cxcl12a	NP_840092	AJ627274	Cyprinus carpio mRNA for stromal cell-derived factor 1a precursor (cxcl12a gene)	NP_001029058
UF_Ppr_AF_104795	zgc:110080	NP_001025309	CR376839	Zebrafish DNA sequence from clone DKEY-75G22 in linkage group 2, complete sequence	-
UF_Ppr_AF_100275		NP_998535	NM_213370	Danio rerio cox4 neighbor (cox4nb), mRNA	NP_006058
UF_Ppr_AF_110335	ppp1cb	ABC94584	EF540902	Carassius auratus protein serine/threonine phosphatase-1 catalytic subunit beta isoform mRNA, complete cds	NP_002700
UF_Ppr_AF_109037	LOC560753	ABN80450	AY825430	Lepidomeda aliciae isolate tc14 TPI-B gene, partial sequence	NP_000356

Table 5.2: **Genes Differentially Expressed during Ovary Development.** This table shows the 71 genes differentially expressed during ovary development. Annotation has been acquired by identifying cross species homologs. Human homolog protein refseq IDs are shown in the last column.

5.3.6 Flutamide Target Genes Map in Proximity of a Network Module Linked to Ovary Development

The 71 genes differentially expressed in ovary development all localized in the same area of the network identified by the genes differentially expressed as a result of FLU exposure (Figure 5.7). We therefore hypothesised that FLU has the potential to interfere with ovary development via a non testosterone mediated pathway. This is consistent with the observation by Villeneuve et al [203] showing that ovary development is impaired in the FHM. When assessing the overlap between these genes and the FLU regulated genes we find that 18% of the ovary genes (13 genes) are shared. These functionally represent plasma membrane (8 genes), wound healing (4 genes) and cell motility genes (4). To specify the exact mechanism, by which FLU perturbs ovary development, will require a number of further experiments. However, this study has shown great potential in the investigation and identification of AOPs using network inference within a systems biology framework in the field of ecotoxicology.

5.3.7 Ingenuity Analysis of the Flutamide Associated Sub-Network

To further characterize relationship between genes regulated by FLU and formulate hypothesis on the mechanisms linking FLU exposure and ovary development we performed an ingenuity analysis on the genes in common between the two gene lists. This resulted in one network, which represents the interaction between insulin-like growth factor 1 (IGF1), collagen, fibronectin (FN1) and hedgehog (SHH) signalling with a number of plasma membrane proteins involved in solute transport and response to inflammatory signals (Figure 5.8).

5.4 Discussion

5.4.1 A Mechanism for Flutamide AR-independent Toxicity?

Interference of the growth hormone/IGF system has been previously shown to be mediated by $17\alpha - ethinylestradiol$ (EE2) in tilapia (*O. niloticus*) [280]. The authors suggested that perturbation of reproductive functions and growth may be mediated by cross talk between sex steroids

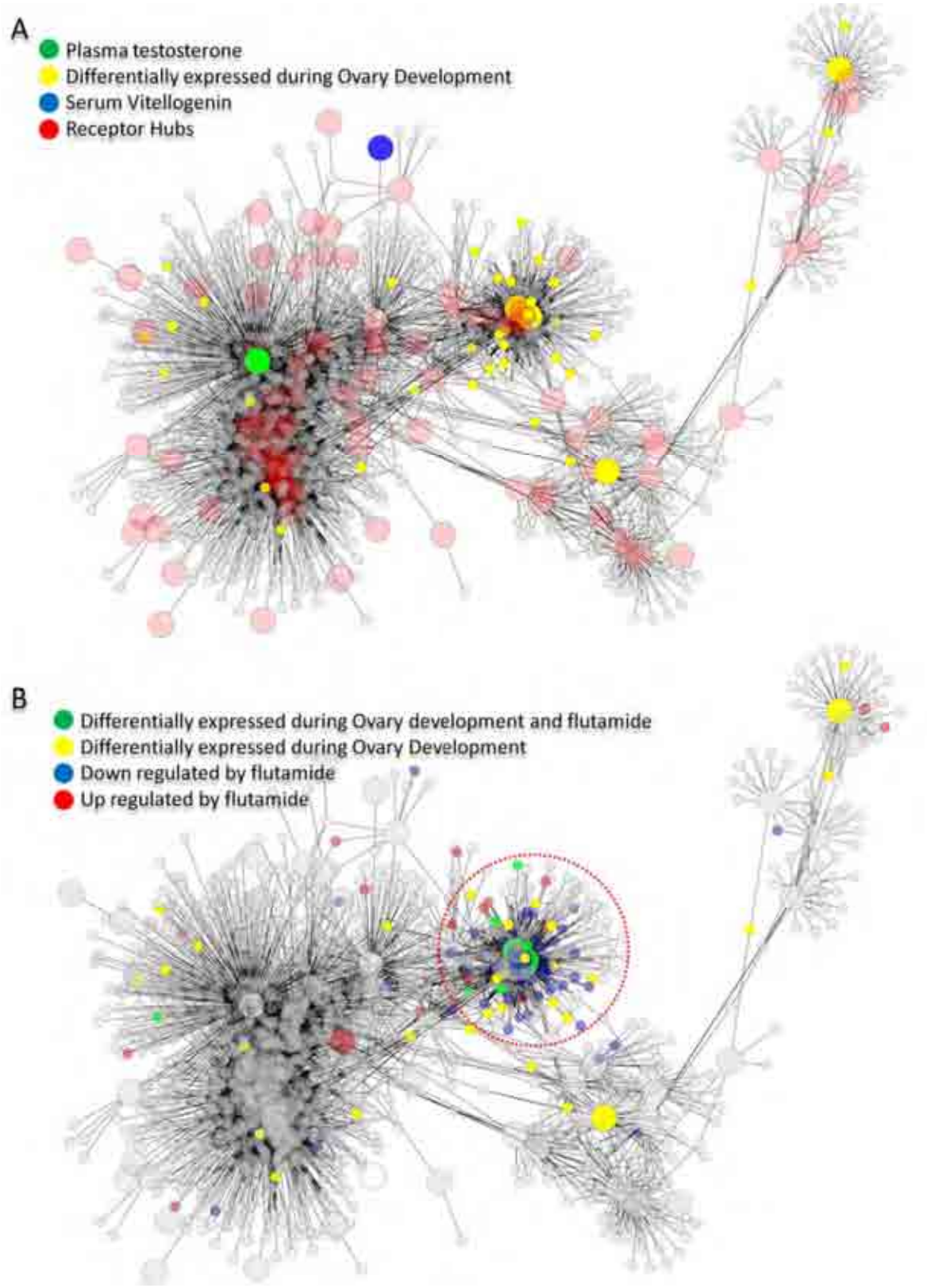


Figure 5.7: **Mapping of Differentially Expressed Genes during Ovary Development and in Response to Flutamide.** A) represents where the identified genes, differentially expressed during ovary development, localize in our inferred network. B) Addition of the genes differentially expressed as a result of FLU exposure reveal a sub-network enriched in genes from both lists.

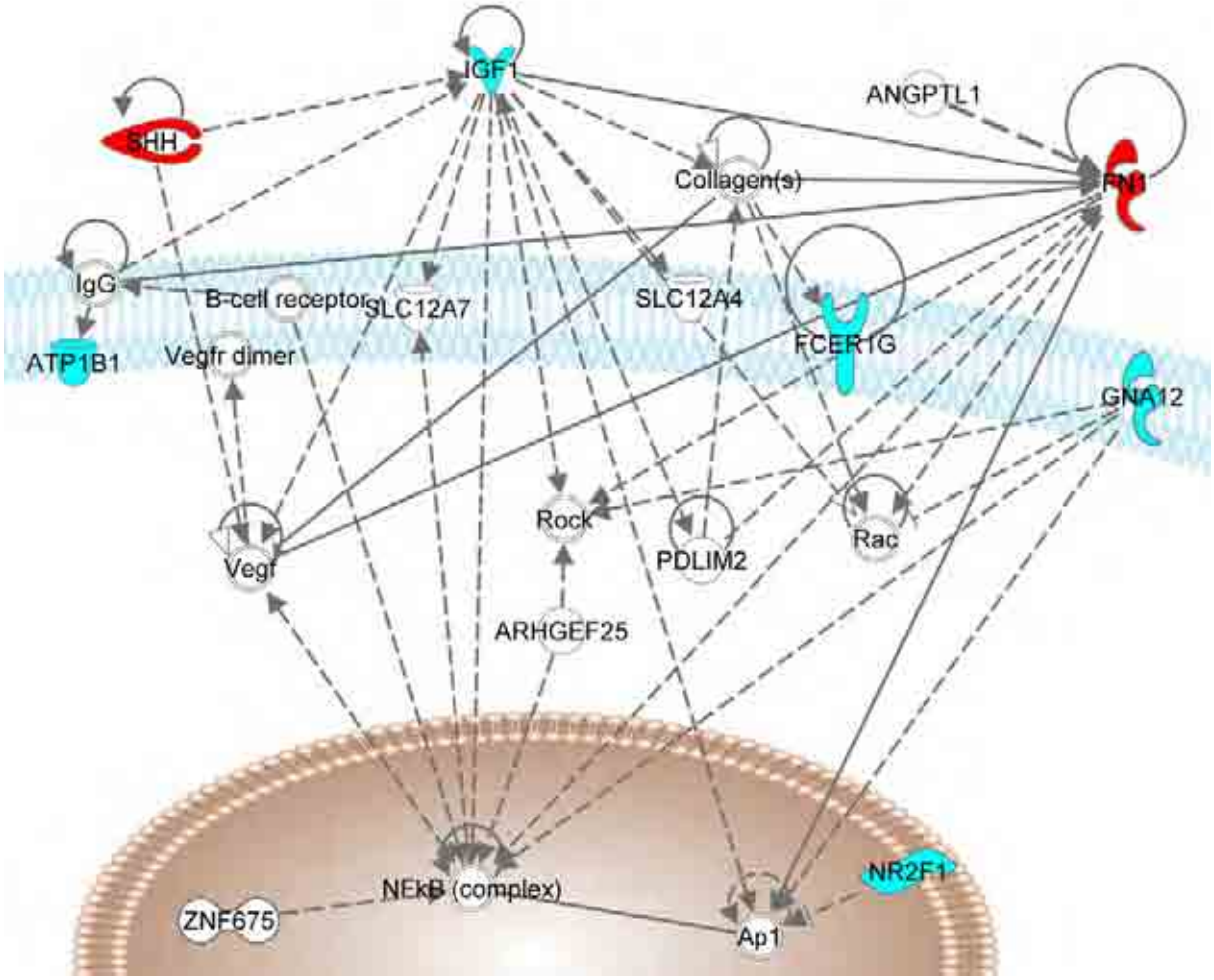


Figure 5.8: **Ingenuity Pathway Analysis of Genes Overlapping between Ovary Development and in Response to Flutamide.** Genes marked in red and blue are up and down regulated in response to FLU. Direct and indirect relationships are represented by a continuous and dotted line respectively. Genes with white backgrounds have been added by the ingenuity pathway analysis algorithm.

and the IGF-1 system rather than as a direct result [280]. SHH on the other hand has shown alteration of bone patterning in zebrafish [281] and is associated to various aspects of empryonic and adult organ development [282–284]. We therefore hypothesise that the effect of FLU is directed towards a perturbation effect of IGF-1 signalling pathways including overexpression of developmental genes such as hedgehogs causing changes in the endocrine system and the observed reproductive effects. Further, more specific experiments will be needed to specify the exact mechanism by which FLU acts on FHMs.

5.4.2 Further Developments

Several potential improvements of our approach may be necessary to allow it to be generally applicably in the ecotoxicological community. Kernel based MI estimation, such as in the original ARACNE implementation [98], is generally favoured to discretization based methods. Although, as samples size increase, different MI estimation methods should converge, choosing the correct methodology is imperative to the success and accuracy of the reverse engineering approach. The accuracy of MI estimation is however not the only issue in network inference. One important challenge is the elimination of indirect connections. ARACNE and CLR use different strategies to achieve this. A systematic comparison may be needed to limit the number of false positives.

To improve biological interpretation we focused our efforts to only explore receptors, phenotypic measurements and their neighbourhoods to identify novel AOPs. In the case of a well characterized anti-androgen this may be favourable [13, 169]; however, with less defined compounds, a more complete approach may need to be pursued. There are also a number of modularization techniques available which may be better suited for identifying functional modules within a regulatory network [101, 285–287]. In many cases, these are more sophisticated and in need of much higher computational effort making them exclusive for the use by computational biologists, unlike the method described in this chapter.

An additional challenge when working with a non-model species such as FHM and the relatively large dataset is the biological variability introduced as a result of fish acquisition. Although

these fish were cultured on an on-site facility at the US EPA Mid-Continent Ecology Division in Duluth, MN, USA, the genetic variability and subsequent transcriptional differences can provide analytical difficulties.

Nevertheless, our approach was successful in identifying known regulatory components regarding testosterone biology and identified novel adverse outcome pathways as a result of FLU exposure.

5.5 Materials and Methods

5.5.1 Comparing MI Estimation Approaches using minet

Mutual information quantifies the information between any given pair of discrete variables by calculating the dependencies between them. More specifically it is defined as

$$I(x, y) = S(x) + S(y)S(x - y), \tag{5.2}$$

where each component is the entropy of an arbitrary variable. For continuous data, such as gene expression data, the entropy is infinite. To solve this, $S(x)$ can be replaced with the differential entropy, which averages the log-probability density rather than the log-mass [98] for mutual information estimation. This in combination with a Gaussian Kernel estimator provides one of the gold standard techniques in reverse engineering [273].

Discretization

In this application we have the classic definition of MI, which involves discretizing the data prior to estimating the MI value. More specifically if a continuous random variable is defined by the interval $[a, b]$, it can be discretized by sorting the data into a pre-defined number of bins. There are a number of methodologies which address this issue such as Equal Width, Global Equal Width or Equal Frequency. Both equal width methods are based on discretizing the data into sub-intervals of equal size. Where the non-global approach considers discretization of each variable separately, the global methodology discretizes based on the minimum and maximum value across the whole dataset. On the other hand the equal frequency method tries to partition

the bins in such as way that for each variable in the datasets each bin has the same number of data points.

MI Estimation

Estimation of the mutual information is also available in a number of methodologies. Within the minet package the empirical, Miller-Madow corrected, Shrink entropy or Schurmann-Grassberger estimator are implemented. More specifically the Miller-Madow correction is an extension of the empirical, or nave, estimator. The use of a logarithmic function within the empirical estimator creates an asymptotic bias which can be easily corrected by adjusting the empirical entropy. Miller-Madow is often favoured to the nave estimator as it does not add to the computational cost and reduces bias without changing variance. The Shrink entropy estimator on the other hand was developed as a combination of two different estimators whose advantages lied in low variance and low bias. Particularly with small sample sizes the shrink estimator should provide improved MI estimation. Lastly the Schurmann-Grassberger estimator is based on the Dirichlet distribution which can be used to estimate the entropy of a discrete random variable. Its origins lie in Bayesian statistcs where it is used as the conjugate prior of the multinomial distribution [288].

Network Inference Methodology

Within the minet package three network inference methodologies have been implemented. These include ARACNE, CLR, and MRNET. ARACNE [98] is based on the data processing inequality which essentially reduces the weakest edge within each triplet given a selected threshold. CLR [99] on the other hand derives a z-score related to the empirical distribution of MI values. Finally MRNET [100] infers the network with the use of the maximum relevance/minimum redundancy (MRMR) feature selection method [289, 290]. Fundamentally each gene is defined as a target for which a variable selection procedure is performed [272].

Comparison of MI Estimation and Discretization Methodologies

We generated MI values using five different approaches (Shrink estimator with equal frequency (SEF), Shrink estimator with equal width (SEW), Empirical estimator with equal width (EEW),

Miller-Madow Corrected estimator with equal width (MMW) and Schurmann-Grassberger estimator with equal width (SEW)). We then plotted each pair value against its pearson correlation value (Figure 5.9). Each MI estimation and discretization procedure provides a varying result. The empirical as well as the Schurmann-Grassberger estimator both show an increase in MI score for gene pairs with no linear relationship. In contrast, equal frequency discretization, inflates all MI scores across the whole dataset. The Shrink Entropy and Miller-Madow Corrected Estimator linked to Equal Width result in a comparable shape across the MI space. We chose to use the MMW approach guided by the fact that the default mi estimator in the minet package is the empirical estimator and that Miller-Madow corrects its bias.

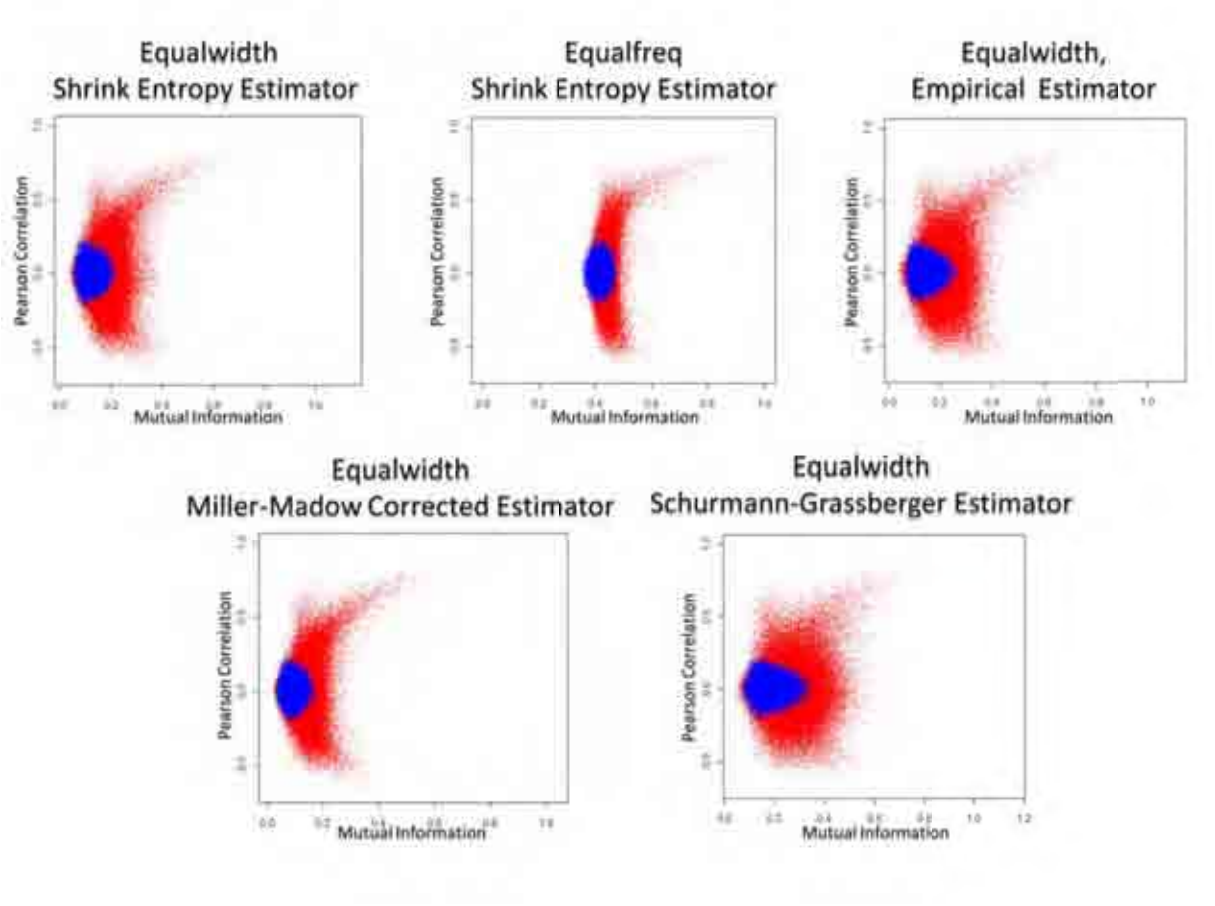


Figure 5.9: **Comparison of Various MI Estimation Techniques Coupled to Discretization Approaches.** MI values computed from the original dataset are represented in red and MI values derived from a randomized dataset in blue. Equal Freqency discretization increases MI in both the random and original dataset. Both the empirical and Schurmann-Grassberger Estimators inflate the MI value with gene pairs with low linear relationships.

5.5.2 Building and Visualizing the Identified Network

In order to develop molecular networks we combined the 868 gene expression arrays and normalized these using a median centred linked to quantile normalization. In addition we added information on plasma testosterone and vitellogenin levels. We then standardized the dataset (mean = 0; sd = 1) using the QuantPsyc R package. We decided to focus on the phenotypic and receptor genes represented in this dataset (Table 5.1). We hypothesised that these and their neighbourhoods will play an important role in the context of the response of FHM to FLU. For each of the receptors we build a network, using our chosen approach (MMW), comprised of all potential connections to all other genes and measurements. Statistically significant connections were chosen by calculating an FDR threshold based on MI values derived from a randomized dataset ($MI > 0.3$). Individual networks were then merged and visualized using the software application cytoscape. A force-driven layout was used to change the topology of the graph to represent the strength of association by the value of MI of each tested pair. Genes with a strong relationship will therefore appear in close proximity of each other.

Identifying and Annotating Functional Modules

In order to identify molecular subnetworks we utilized the cytoscape plugin MCODE which searches for highly interconnected regions. To functionally annotate the identified modules we used human homologs of the FHM genes as an input to the web-service DAVID. Functional terms were chosen to have an enrichment score (ES) of >= 1.5. The enrichment score is the overall enrichment given all the members of its class. It is calculated by taking the geometric mean of the p-values and is therefore representative of a significant enrichment. In this case an enrichment score of 1.5 equates a $p - value < 0.03$.

5.5.3 Identifying Differentially Expressed Genes

To identify genes differentially expressed in respect to FLU a one way ANOVA was performed on the subset of FHM data containing the specific FLU exposure in TMEV [291, 292]. Resulting p-values were corrected by Benjamini & Hochberg [75] false discovery rate correction

(FDR_{BH}). Genes differentially expressed at 10% FDR were identified and mapped onto the inferred network (Table x). During ovary development a significance analysis of microarrays (as implemented in TMEV) was applied to identify genes differentially expressed. At 20% FDR 71 genes were differentially expressed (Table 5.2). The resulting list of genes was then mapped onto the identified network.

Chapter 6

The Future of Predictive Toxicology: A Systems Biology Perspective

Animal models have been the focus in toxicity testing for many years. In industry this has provided a stable base for risk assessment in human health. Traditionally, toxicological assessment is performed by exposing laboratory animals to relatively high doses of chemicals and acceptable concentrations in other species, including humans, were inferred by applying uncertainty and/or safety factors. These factors were meant to represent interspecies differences in sensitivity, metabolism or physiology, and include deviation in environmental variables such as food intake, air or water properties or general stress of the system (may it be through predators or additional environmental factors). Such experiments, however, are costly and in many cases do not provide the relevant information to support human safety assessment [16]. In recent years a growing interest in alternative approaches in the scientific and industrial community has been fuelled by societys interest in animal rights, governmental pressure, financial concerns and advancements of research tools [17]. In this context, predictive toxicology provides a combination of omics technologies and statistical modelling techniques within a framework for identifying biomarkers relevant of the studied characteristics.

In 2007 the U.S National Research council (NRC) provided a new paradigm that would achieve to cover a broad range of chemicals, reduce time and cost, develop robust methodologies and subsequently reduce the need for animal testing [293]. This would encompass a number of

computational, molecular and toxicological fields in a collaborative effort to integrate their expertise. The key to success is the understanding of the biology involved as a result of chemical exposure. This would lead to the identification of adverse outcome pathways, which provide a mechanistic basis to exposure.

These systems biology approaches are not novel in other fields of biology. Several studies have been published where similar methodologies were successfully applied to microbiology [206, 294] or in clinical areas [169, 295, 296]. The pharmaceutical industry is already implementing some of the in-silico [297] and *in vitro* [298] ideas to efficiently identify drugs but has yet to include a systems biology approach [299, 300].

6.1 Open Challenges in Predictive Toxicology

The focus of predictive toxicology in recent years lay in biomarker discovery, identifying groups of features predictive of phenotypic endpoints. These studies, however, did not provide much mechanistic insight, such as toxicity or recovery mechanism. It is fundamental for risk assessment or environmental monitoring to understand the underlying biological effect. Furthermore this should enable simulation of chemical exposure and provide more accurate extrapolation to other species by including species specific knowledge of physiology and metabolism.

In the strategy proposed by the NRC, high-throughput *in vitro* assays, ideally in the species of interest (removing the need for extrapolation), would be carried out and subsequent analysis using computational methodologies provide extensive knowledge on toxicity pathways [16]. Due to the low cost of these, a broad range of concentrations, ranging from very low (lower than environmentally relevant) to high (toxic), could be characterized. This application, however, may provide additional challenges, especially in industry. Although, a number of *in vitro* models have been developed to provide suitable alternatives to animal testing [301, 302] very few of these have been accepted for regulatory purposes [302, 303]. This reflects the enormous challenge of representing whole organism physiology, metabolism, and the plethora of cell types in a single *in vitro* experimental system. Nevertheless, progress is being made with the inclusion of pluripotent stem cells with the ability to differentiate into a number of different cell types and

even form 3D structures [304].

Most importantly, for these techniques to become generally applicable, in risk assessment and environmental monitoring, the identified mechanisms, biomarkers or adverse outcome pathways have to be translated into a field based monitoring program. Predicting real life scenarios, while reducing the number of animals and achieving at least comparable prediction accuracy, is therefore the ultimate aim of these developments in the pharmaceutical and ecotoxicology area. Assays resulting from this innovation should conform to be: of low economic impact, easy, minimize the use of expensive instruments and be broadly applicable.

6.2 Current Applications in Systems Toxicology

A number of groups in academia and health agencies have shown that systems biology approaches can be highly informative and relevant within the context of predictive toxicology. Most notably are the large scale applications by the U.S. EPA including a number of programs within the computational toxicology framework CompTox [305], which includes ExpoCast [306], ToxCast [22], Tox21 [307] and virtual liver and embryo projects [308, 309]. ToxCast, in particular, was a direct response to the NRC report mentioned previously and combines high throughput screening and computational methodologies to prioritize chemicals for further toxicological evaluation [22]. Tox21 on the other hand combines several U.S. federal agencies to develop models for risk assessment. The ultimate goal of this project is to provide activity profiles for a large number of chemicals which are predictive of *in vivo* toxicity [307]. Smaller projects have also shown that characterization of specific areas, such as neuro or hepatotoxicity, can be analysed in a systems biology framework [296, 310, 311]. Slikker et al [311] used such an approach to build models representative of ketamine exposure in developing rodents, specifically focusing on neurotoxicity. Craig et al [310] integrated several omics analyses to characterize Methapyrilene induced hepatotoxicity in rats. In ecotoxicology such application has only been a very recent development. Large scale efforts include datasets created by the U.S. Army ERDC to investigate the impacts of endocrine disruptors on ovary tissue in fathead minnow [203, 271]. A proof of concept study applying a systems biology approach to

identify adverse outcome pathways using a preliminary dataset has been demonstrated in this thesis (Chapter 5, [13]). Other groups such as De Wit et al [312] combined transcriptomics and proteomics to identify the impact of tetrabromobisphenol-A on zebrafish livers. Previously noted publications, such as Williams et al [44] and Katsiadaki et al [45], also fall within this category. Many of these publications utilized non-model species, which can impede biological interpretation. Among the best examples for a systems biology approach in the field of toxicology employing non-model species has been published by Bundy et al [170]. The authors focused their efforts on the earthworm *Lumbricus rubellus* Hoffmeister and characterized the chronic effect (70 day exposure) to sub-lethal levels of copper. Integration of transcriptomic, metabolomic and phenotypic endpoints clearly showed changes in oxidative phosphorylation and carbohydrate use [170]. This shows that species heterogeneity is a particular issue in ecotoxicology. The animals sampled from the environment may react according to their genetic make-up, showing different responses relating to their geographical location or previous exposure. On the other hand, heterogeneity may provide the biological variation needed to develop a generalized assay, which would be applicable across a number of environments. Working with environmental samples can also bring additional problems. Finding samples, capturing and retrieving these back to the laboratory without causing too much external influence (stress response, change in environment etc.) can be difficult.

6.3 Predictive Toxicology in the 21^{st} Century

The very recent application of systems biology approaches in the field of ecotoxicology tells us that despite the difficulties of working with non-model species useful information on adverse outcome pathways can be obtained. Within this thesis we have provided further evidence to strengthen this approach by integrating mainly transcriptomics and QSAR analyses (Chapter 2, 4), developing predictive models of toxicity, as measured by LC_{50} (Chapter 4), or chemical class (Chapter 3) all within integrating pathway information to facilitate biological interpretation. We discussed the relevance of the results in context of the environmental stressors and provided speculations on potential general mechanisms which may be shared between the rat

and waterflea. In most cases our results provided important links between already existing knowledge and showed that pathway-level analyses with non-model species is one way to succeed in identifying the impact of chemicals on biological species. Finally, we demonstrated, within a proof of concept study, the advantages of reverse engineering in identifying novel adverse outcome pathways (Chapter 5). This secondary approach is particularly advantageous when large datasets are available. The take-home-message from this work is that to succeed in preventing ecological disasters, identifying exogenous compounds in the environment and chemical risk assessment, a more forward approach to toxicity testing is needed. This encompasses utilizing the systems biology toolbox, integrating different technologies, including model and non-model species alike and most importantly derive mechanistic biomarkers which are relevant to the phenotypic outcome.

6.4 Translating Mechanistic Biomarkers into Safety Assessment

The ultimate goal of the work presented, is the development of predictive assays which can be applied to real life scenarios. As mentioned earlier, these need to possess a number of characteristics to become successful. Most importantly, industrial standards need to be met to become a viable alternative. In most cases this means validating the developed assay across a number of laboratories with a large selection of chemicals and then showing that inter-laboratory variation is at a minimum while preserving high sensitivity and specificity. In addition to these requirements, the assay should be easily applied (ideally even by untrained staff), be economical in production and acquisition, and be easily implemented into current protocols. While the above features should be maintained a closely as possible it should be noted that the perfect should not be the enemy of the good. In fact rapid advances are already made in implementing stem-cell biology [313] and computational modelling of cellular response pathways [314, 315] into viable systems. This and the recent development of easily scalable high data content assays for cellular responses [316] show the field is progressing fast towards an improved toxicity testing approach. Within this PhD, we also ventured into developing a stem cell based predic-

tivity toxicology approach. In our initial analysis we compared the transcriptional response of fibroblasts derived from three different tissues (Skin, Bone Marrow, Synovium) in the context of a abnormal wound healing response in rheumatoid and osteoarthritis. Furthermore, we are also endeavouring towards an ecotoxicology approach to water quality assessment (see next section).

6.5 Translating Systems Ecotoxicology into Environmental Monitoring

To further expand our work in *D. magna* we have applied for a Natural Environment Research Council (NERC) grant to focus on the development of a systems biology approach to water quality assessment. This proposal has been funded with me as a named candidate and is set to commence in the beginning of 2012. In the following sections I describe the overall strategy behind this study.

6.5.1 Overall Aim and Objective

The aim of this project is to pioneer a combination of advanced compoutational modelling techniques to identify adverse outcome pathways and evaluate their potential as a predictive tool in toxicity assessment. The projects involves characterizing the response of *D. magna* to exposure of a number of environmentally and industry relevant chemicals utilizing transcriptomics, metabolomics, lipidomics and several physiological endpoints. These are the overall project objectives:

1. Acquire datasets representative of the molecular and physiological responses of *D. magna* to priority substances within the Water Framework Directive.
2. Develop computational models representing adverse outcome pathways that link physiology to molecular responses to chemical exposure.
3. In collaboration with the Environmental Agency, validate the predictive power of the molecular biomarkers identified in objective 2 for their ability to predict complex exposure patterns from chemically defined mixtures and environmentally sampled waters with

different degrees of contamination.

6.5.2 A new Vision for Biomarker Discovery

The current 1st generation use of biomarkers in environmental monitoring has been limited. The difficulty is that the current set of biomarkers, such as induction of CYP1A, vitellogenin, metallothionein, is limited to specific types of exposures and lacks a strong link between exposure and biological effect. We propose that moving away from single biomarkers and employing a battery of indicators, related to AOPs, can overcome this. Even relatively simple transcriptomic profiles have been able to distinguish between the biological MOAs of chemicals (e.g. genotoxic versus non-genotoxic carcinogens [7]), and similarly for metabolomics studies of toxicant MOAs in *D. magna* [171]. In the clinical setting such profiles are allowing improved diagnosis and prognosis; e.g. Mammaprint that offers a DNA microarray-based *in vitro* analysis to predict the likelihood of tumour recurrence [317]. Added to these successes, we have shown in a species of ecotoxicology relevance that it is possible to successfully identify sub-networks whose activity was predictive of environmental exposure and linked to organism health indicators (see Chapters 3–5). Thus we are now poised to be able to develop 2nd generation biomarkers, derived from omics and computational studies, that will be linked to health, non-biased in discovery, mechanistically based and applicable to simple targeted assays. Any successful environmental biomarker discovery program needs to be developed in close interaction with regulatory authorities. We have been engaged in extensive knowledge transfer activities to integrate such strategies into the thinking and ultimately the practice of regulators. In close interaction with the EA we will provide proof of concept studies, exposing *D. magna* to waters of known contaminants with known environmental impact. We wish to apply the AOP determined in this project to targeted assays based on quantitative PCRs and targeted metabolite measurements. Ultimately these targeted assays can be converted to high-throughput and thus be readily employed in environmental monitoring. The ultimate assays will concur with the 5Rs of Reproducibility, Representative, Responsive, Robust and Relevant that are key requirements of monitoring techniques [318].

Appendix A

Publications

Below I present you the list of publications to which I contributed to during my PhD. I have highlighted publications in red where my involvement played a major role towards the publication.

A.1 In Preparation

A.1.1 First Author Papers

1. Antczak P*, Filer A*, Parsonage G, Leguault H, OToole M, Pearson M, Thomas AMC, Scheel-Toellner D, Raza K, ONeill L, Salmon M, Buckley CD, Falciani F. Characterization of the Serum Response Programme in Rheumatoid and Osteoarthritis (Advanced Manuscript write-up status) (* Authors contributed equally)

   Author Contribution: This paper reports the results of a project I initiated at the beginning of my PhD, before I shifted to predictive toxicology. The aim of the project was to identify molecular signatures associated to wound healing and predictive of disease. My role in this publication was to conceive the analysis strategy and defines the questions. I then performed all analysis and wrote the paper with input from collaborators.

2. P. Antczak, J. Hun, L. Scandlan , M. Viant, C. Vulpe, F. Falciani A Pathway-based Approach to Predictive Toxicology in the crustacean Daphnia Magna (Advanced Manuscript write-up status)

3. P. Antczak, T. White, C. Vulpe, F. Falciani Towards a Systems Biology approach to chemical class prediction in Daphnia Magna (Advanced Manuscript write-up status)

   Author Contribution: The two papers above are the result of an on-going collaboration with Chris Vulpe who shared a large transcriptomics dataset derived from D. magna exposures with us to perform a rigorous bioinformatics analysis. The aim of this dataset was to identify whether D. magna transcriptional response is predictive of chemical exposure and whether such a system would be more effective in water quality assessment. I therefore performed an in-depth analysis of this data using a number of techniques which resulted in these 2 interesting stories (Chapters 3 and 4 of my thesis). I then wrote the papers and revised it with the help of collaborators.

A.2 Published

A.2.1 First Author Papers

4. P. Antczak, F. Ortega, J.K. Chipman, F. Falciani. Mapping Drug Physico-Chemical Features to Pathway Activity Reveals Molecular Networks Linked to Toxicity Outcome. PloS one 5(8): 580-588. 2010.

   Author Contribution: Fernando Ortega initially developed the proof of concept study for this publication. I have taken his ideas and expanded these and applied this to a dataset representing renal tubular degeneration as a result of exposure. I packaged the paper and developed it for publication with the help of JK. Chipman and F Falciani. This paper is part of my thesis and is discussed in chapter 2.

5. P. Antczak*, F. Soulet*, W.W. Kilarski*, J. Herbert, R. Bicknell, F. Falciani, A. Bikfalvi. Gene signatures in wound tissue as evidenced by molecular profiling in the chick embryo model. BMC Genomics 11(1): 495. 2010. (* Authors contributed equally)

   Author Contribution: This particular project was at the very beginning of my PhD, before I worked on predictive toxicology and the wound healing project described earlier. In fact

the RA/OA project was initiated as a result from the analysis that I performed on a set of data acquired by F. Soulet and W. Kilarski in A. Bikfalvi group. I performed the bioinformatics analysis to identify functions which provided insight into the wound healing model of the CAM in chicken embryos. This particular system is devout of immunocompetent cells and so provides an interesting insight into wound healing by chemokines only.

A.2.2 Work in Collaboration

6. H. Lin, J. Halsall, P. Antczak, L. ONeill, F. Falciani, B. Turner. Up-regulated expression of X-linked genes in mouse embryonic stem cells is consistent with Ohnos hypothesis. Nature Genetics (In-Press)

   Author Contribution: Prof. Bryan Turner approached F.Falciani and me to help in performing a robust statistical analysis of the data that was generated in Prof Turners group. I analysed the data to show that genes present on the x-chromosome are up-regulated to be consistent with Ohnos hypothesis.

7. V. Trevino, M. G. Tadesse, M. Vannucci, P. Antczak, S. Durant, F. Al-Shahrour, J. Dopazo, M. J. Campbell and F. Falciani. Analysis of Normal-Tumour Tissue Interaction in Solid Tumours: Prediction of Prostate Cancer Features from the Molecular Profile of Adjacent Normal Cells. PloS one 6(3): e16492 2011

   Author Contribution: In this study, we needed to validate the up-regulation of pro-metastatic chemokines CX3CL1 and CCL20 as a result of IL-1 induction. S. Durant, our technician, and I provided support by performing the experiment to validate this hypothesis. I analysed the results and provided F. Falciani with the necessary figures.

8. E. J Perkins, J.K. Chipman, S. Edwards, T. Habib, F. Falciani, R. Taylor, G. Van Aggelen, C. Vulpe, P. Antczak, A. Loguinov. Reverse Engineering Adverse Outcome Pathways. Environmental Toxicology and Chemistry 2010

   Author Contribution: Data generated by the US EPA was provided to show the effective-

ness of reverse engineering of adverse outcome pathways in ecotoxicology. I provided computational support in normalization and analysis of the data. The full analysis was not published in this review, but is shown in Chapter 5 of my thesis.

9. N.A. Burton, M.D. Johnson, P. Antczak, A. Robinson and P.A. Lund. Novel aspects of the acid response network of E. coli K-12 are revealed by a study of transcriptional dynamics. Journal of Molecular Biology 2010.
   Author Contribution: P.A. Lund approached me to provide support in analysing some of the data created by N.A. Burton. I produced several clustering figures representing the dynamics of several promoters in related mutants (mainly Figure 7c in the paper).

10. K. Sameith, P. Antczak, E. Marston, N. Turan, D. Maier, T. Stankovic, F. Falciani. Functional modules integrating essential cellular functions are predictive of the response of leukaemia cells to DNA damage. Bioinformatics 24(22): 2602-2607. Nov 2008.
    Author Contribution: This is my first paper for which I provided some support in the analysis of the data. I also reformatted and corrected the manuscript where indicated by the reviewers.

11. R. Gupta, A. Stincone, P. Antczak, S. Durant, R. Bicknell, A. Bikfalvi, F. Falciani. A Computational Framework for Gene Regulatory Network Inference that Combines Multiple Methods and Datasets. BMC Systems Biology 5(1): 52 2011
    Author Contribution: The method itself was implemented and validated by R. Gupta. I contributed heavily to the processing and analysis of the experimental data before this new method was applied.

12. A. Stincone, N. Daudi, A. Rahman, P. Antczak, I. Henderson, J. Cole, M. Johnson, P. Lund, F. Falciani. A systems biology approach sheds new light on Escherichia coli acid resistance. Nucl. Acids Res. 39 (17), 7512-7528.
    Author Contribution: I helped A. Stincone with the analysis and interpretation of the data.

13. Mura M, Swain RK, Zhuang X, Vorschmitt H, Reynolds G, Durant S, Beesley JF, Herbert JM, Sheldon H, Andre M, Sanderson S, Glen K, Luu NT, McGettrick HM, Antczak P, Falciani F, Nash GB, Nagy ZS, Bicknell R. Identification and angiogenic role of the novel tumor endothelial marker CLEC14A. Oncogene 2011 doi: 10.1038/onc.2011.233.
Author Contribution: I provided bioinformatics support during the analysis of this project. I also provided experimental support for S. Durant who was running some of the microarrays.

14. S Moro, JK Chipman, P Antczak, N Turan, W Dekant, F Falciani, A Mally. Identification and pathway mapping of furan target proteins reveal mitochondrial energy production and redox regulation as critical targets of furan toxicity Toxicological Sciences 2012
Author Contribution: To help with the identification of specific groups of proteins identified by the authors approach, I analysed their proteins by their domains and grouped them accordingly using a modified EASE score.

List of References

[1] Kotsiantis S, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. 2007;160:3.

[2] Merlot C. Computational toxicology-a tool for early safety evaluation. Drug discovery today. 2009;.

[3] Vayer P, Arrault A, Lesur B, Bertrand M, Walther B. Chemoinformatics and virtual screening of molecules for therapeutic use. Médecine sciences: M/S. 2009;25(10):871.

[4] Steiner G, Suter L, Boess F, Gasser R, de Vera MC, Albertini S, et al. Discriminating different classes of toxicants by transcript profiling. Environmental health perspectives. 2004;112(12):1236.

[5] Bulera SJ, Eddy SM, Ferguson E, Jatkoe TA, Reindel JF, Bleavins MR, et al. RNA expression in the early characterization of hepatotoxicants in Wistar rats by high-density DNA microarrays. Hepatology. 2001;33(5):1239–1258.

[6] Thomas RS, Rank DR, Penn SG, Zastrow GM, Hayes KR, Pande K, et al. Identification of toxicologically predictive gene sets using cDNA microarrays. Molecular Pharmacology. 2001;60(6):1189.

[7] Ellinger-Ziegelbauer H, Stuart B, Wahle B, Bomann W, Ahr HJ. Comparison of the expression profiles induced by genotoxic and nongenotoxic carcinogens in rat liver. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2005;575(1-2):61–84.

[8] Hamadeh HK, Bushel PR, Jayadev S, DiSorbo O, Bennett L, Li L, et al. Prediction of compound signature using high density gene expression profiling. Toxicological Sciences. 2002;67(2):232.

[9] Waring JF, Jolly RA, Ciurlionis R, Lum PY, Praestgaard JT, Morfitt DC, et al. Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. Toxicology and applied pharmacology. 2001;175(1):28–42.

[10] Tan Y, Shi L, Hussain SM, Xu J, Tong W, Frazier JM, et al. Integrating time-course microarray gene expression profiles with cytotoxicity for identification of biomarkers in primary rat hepatocytes exposed to cadmium. Bioinformatics. 2005;22(1):77.

[11] Schultz T. Adverse outcome pathways: A way of linking chemical structure to in vivo toxicological hazards. 2010;.

[12] Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, et al. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry. 2010;29(3):730–741.

[13] Perkins EJ, Chipman JK, Edwards S, Habib T, Falciani F, Taylor R, et al. Reverse engineering adverse outcome pathways. Environmental Toxicology and Chemistry. 2011;.

[14] Abbott A. Animal testing: More than a cosmetic change. Nature. 2005;438(7065):144–146.

[15] Parliament E. Regulation (EC) No. 1907/2006 of the European parliament and the council of December 18, 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No. 793/93 and Commission Regulation (EC) No. 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC, Off. J. Eur. Commun. 396 (2006) 1. Off J Eur Commun. 2006;396(1).

[16] Andersen ME, Krewski D. Toxicity testing in the 21st century: bringing the vision to life. Toxicological sciences. 2009;107(2):324.

[17] Kimber I, Humphris C, Westmoreland C, Alepee N, Negro GD, Manou I. Computational chemistry, systems biology and toxicology. Harnessing the chemistry of life: revolutionizing toxicology. A commentary. Journal of Applied Toxicology. 2011;31(3):206–209. Available from: `http://dx.doi.org/10.1002/jat.1666`.

[18] Garcia-Reyero N, Perkins EJ. Systems biology: Leading the revolution in ecotoxicology. Environmental Toxicology and Chemistry. 2011;.

[19] Clemedson C, Blaauboer B, Castell J, Prieto P, Risteli L, Vericat JA, et al. AcuteToxOptimation and pre-validation of an in vitro test strategy for predicting human acute toxicity. ALTEX. 2006;23:254–258.

[20] Prieto P, Baird AW, Blaauboer BJ, Ripoll JVC, Corvi R, Dekant W, et al. The assessment of repeated dose toxicity in vitro: a proposed approach. ATLA-NOTTINGHAM-. 2006;34(3):315.

[21] Spielmann H, Müller L, Averbeck D, Balls M, Brendler-Schwaab S, Castell JV, et al. The second ECVAM workshop on phototoxicity testing. The report and recommendations of ECVAM workshop 42. Altern Lab Anim. 2000;28:777–814.

[22] Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicological Sciences. 2007;95(1):5.

[23] Kimmel CA, Grant LD, Sloan CS, Gladen BC. Chronic low-level lead toxicity in the rat: I. Maternal toxicity and perinatal effects. Toxicology and applied pharmacology. 1980;56(1):28–41.

[24] McGivern RF, Sokol RZ, Berman NG. Prenatal lead exposure in the rat during the third week of gestation: long-term behavioral, physiological, and anatomical effects associated with reproduction. Toxicology and applied pharmacology. 1991;110(2):206–215.

[25] MacGregor J. Mutagenicity studies of flavonoids in vivo and in vitro. Toxicol Appl Pharmacol. 1979;48:A47.

[26] Fielden MR, Nie A, McMillian M, Elangbam CS, Trela BA, Yang Y, et al. Interlaboratory evaluation of genomic signatures for predicting carcinogenicity in the rat. Toxicological sciences. 2008;103(1):28.

[27] Fielden MR, Brennan R, Gollub J. A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. Toxicological sciences. 2007;99(1):90.

[28] Fielden MR, Eynon BP, Natsoulis G, Jarnagin K, Banas D, Kolaja KL. A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity. Toxicologic pathology. 2005;33(6):675.

[29] Lessigiarska I, Worth AP, Netzeva TI, Dearden JC, Cronin MTD. Quantitative structure-activity-activity and quantitative structure-activity investigations of human and rodent toxicity. Chemosphere. 2006;65(10):1878–1887.

[30] Venkatapathy R, Moudgal C, Bruce R. Assessment of the oral rat chronic lowest observed adverse effect level model in TOPKAT, a QSAR software package for toxicity prediction. Journal of chemical information and computer sciences. 2004;44(5):1623–1629.

[31] Chapman PM. Integrating toxicology and ecology: putting the. Marine Pollution Bulletin. 2002;44(1):7–15.

[32] Diffuse Pollution and the Water Framework Directive; 2006. Available from: http://www.environment-agency.gov.uk/static/documents/Research/briefing2006_diffuse_1622466.pdf. Accessed 28.09.2011.

[33] Ebert D. A Genome for the Environment. Science. 2011;331(6017):539.

[34] Hogstrand C, Kille P. Comparative toxicogenomics. vol. 2. Elsevier Science; 2008.

[35] Dodson SI, Hanazato T. Commentary on effects of anthropogenic and natural organic chemicals on development, swimming behavior, and reproduction of Daphnia, a key member of aquatic ecosystems. Environmental health perspectives. 1995;103(Suppl 4):7.

[36] Decaestecker E, Gaba S, Raeymaekers JAM, Stoks R, Van Kerckhoven L, Ebert D, et al. Host–parasite Red Queendynamics archived in pond sediment. Nature. 2007;450(7171):870–873.

[37] Daphnia Genome Consortium;. Available from: https://wiki.cgb.indiana.edu/display/DGC/Home. Accessed 28.09.2011.

[38] Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, et al. The ecoresponsive genome of Daphnia pulex. Science. 2011;331(6017):555.

[39] Korpelainen H, Ketola M, Hietala J. Somatic polyploidy examined by flow cytometry in Daphnia. Journal of plankton research. 1997;19(12):2031.

[40] Laforsch C, Ngwa W, Grill W, Tollrian R. An acoustic microscopy technique reveals hidden morphological defenses in Daphnia. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(45):15911.

[41] Poynton HC, Varshavsky JR, Chang B, Cavigiolio G, Chan S, Holman PS, et al. Daphnia magna ecotoxicogenomics provides mechanistic insights into metal toxicity. Environmental science & technology. 2007;41(3):1044–1050.

[42] Poynton HC, Zuzow R, Loguinov AV, Perkins EJ, Vulpe CD. Gene expression profiling in Daphnia magna, part II: validation of a copper specific gene expression signature with effluent from two copper mines in California. Environmental science & technology. 2008;42(16):6257–6263.

[43] Taylor NS, Weber RJM, Southam AD, Payne TG, Hrydziuszko O, Arvanitis TN, et al. A new approach to toxicity testing in Daphnia magna: application of high throughput FT-ICR mass spectrometry metabolomics. Metabolomics. 2009;5(1):44–58.

[44] Williams TD, Wu H, Santos EM, Ball J, Katsiadaki I, Brown MM, et al. Hepatic transcriptomic and metabolomic responses in the stickleback (Gasterosteus aculeatus) exposed to environmentally relevant concentrations of dibenzanthracene. Environmental science & technology. 2009;43(16):6341–6348.

[45] Katsiadaki I, Williams TD, Ball JS, Bean TP, Sanders MB, Wu H, et al. Hepatic transcriptomic and metabolomic responses in the Stickleback (Gasterosteus aculeatus) exposed to ethinyl-estradiol. Aquatic Toxicology. 2010;97(3):174–187.

[46] Falciani F, Diab A, Sabine V, Williams T, Ortega F, George S, et al. Hepatic transcriptomic profiles of European flounder (Platichthys flesus) from field sites and computational approaches to predict site from stress gene responses following exposure to model toxicants. Aquatic Toxicology. 2008;90(2):92–101.

[47] Lauer B, Tuschl G, Kling M, Mueller SO. Species-specific toxicity of diclofenac and troglitazone in primary human and rat hepatocytes. Chemico-biological interactions. 2009;179(1):17–24.

[48] Kotokorpi P, Ellis E, Parini P, Nilsson LM, Strom S, Steffensen KR, et al. Physiological differences between human and rat primary hepatocytes in response to liver X receptor activation by 3-[3-[N-(2-chloro-3-trifluoromethylbenzyl)-(2, 2-diphenylethyl) amino] propyloxy] phenylacetic acid hydrochloride (GW3965). Molecular pharmacology. 2007;72(4):947.

[49] Brock BJ, Waterman MR. Biochemical differences between rat and human cytochrome P450c17 support the different steroidogenic needs of these two species. Biochemistry. 1999;38(5):1598–1606.

[50] EPA US. EPA ToxCast Program;. Available from: `http://www.epa.gov/ncct/toxcast/`. Accessed 28.09.2011.

[51] Heng BC, Richards M, Shu Y, Gribbon P. Induced pluripotent stem cells: a new tool for toxicology screening? Archives of toxicology. 2009;83(7):641–644.

[52] Laustriat D, Gide J, Peschanski M. Human pluripotent stem cells in drug discovery and predictive toxicology. Biochem Soc Trans. 2010;38(4):1051–7.

[53] Wolkenhauer O. Systems biology: The reincarnation of systems theory applied in biology? Briefings in Bioinformatics. 2001;2(3):258.

[54] Watson JD, Crick FHC. Molecular structure of nucleic acids. Nature. 1953;171(4356):737–738.

[55] Lin CY, Viant MR, Tjeerdema RS. Metabolomics: Methodologies and applications in the environmental sciences. Journal of Pesticide Science. 2006;31(3):245–251.

[56] Viant MR. Metabolomics of aquatic organisms: the new omics on the block. Introducing genomics, proteomics and metabolomics in marine ecology. 2007;332:301–306.

[57] Schmidt CW. Metabolomics: what s happening downstream of DNA. Environmental Health Perspectives. 2004;112(7):A410.

[58] Willenbrock H, Salomon J, Søkilde R, Barken KB, Hansen TN, Nielsen FC, et al. Quantitative miRNA expression analysis: Comparing microarrays with next-generation sequencing. Rna. 2009;15(11):2028.

[59] Hodgson J. ADMET-turning chemicals into drugs. Nature Biotechnology. 2001;19(8):722–726.

[60] van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? Nature Reviews Drug Discovery. 2003;2(3):192–204.

[61] Proudfoot JR. Drugs, leads, and drug-likeness: an analysis of some recently launched drugs. Bioorganic & medicinal chemistry letters. 2002;12(12):1647–1650.

[62] Roche O, Schneider P, Zuegge J, Guba W, Kansy M, Alanine A, et al. Development of a virtual screening method for identification of frequent hitters in compound libraries. Journal of medicinal chemistry. 2002;45(1):137–142.

[63] Dudek AZ, Arodz T, Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Combinatorial Chemistry &# 38; High Throughput Screening. 2006;9(3):213–228.

[64] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research. 2003;31(4):e15.

[65] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249.

[66] Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31(4):265–273.

[67] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association. 1979;p. 829–836.

[68] Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185.

[69] Huber W, Von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002;18(suppl 1):S96.

[70] Affymetrix Mircoarray Suite User Guide - Statistical Algorithms Description Document, Santa Clara, CA;.

[71] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association. 2004;99(468):909–917.

[72] Wu H, Kerr M, Cui X, Churchill G. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. The analysis of gene expression data. 2003;p. 313–341.

[73] Kim SY, Lee JW, Sohn IS. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Statistical methods in medical research. 2006;15(1):3.

[74] Walsh B. Multiple comparisons: Bonferroni corrections and false discovery rates. Lecture Notes (EEB 581, 14 May 2004), Department of Ecology and Evolutionary Biology, University of Arizona, online at http://nitro biosci arizona edu/courses/EEB581-2004/handouts/Multiple pdf (the 14th of August, 2006). 2004;.

[75] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;p. 289–300.

[76] Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B, Statistical Methodology. 2002;p. 479–498.

[77] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;p. 65–70.

[78] Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988;75(2):383.

[79] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 2001;98(9):5116.

[80] Afshari CA, Hamadeh HK, Bushel PR. The Evolution of Bioinformatics in Toxicology: Advancing Toxicogenomics. Toxicological Sciences. 2011;120(suppl 1):S225.

[81] Chou J, Zhou T, Kaufmann W, Paules R, Bushel P. Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. BMC bioinformatics. 2007;8(1):427.

[82] Bushel P, Heinloth A, Li J, Huang L, Chou J, Boorman G, et al. Blood gene expression signatures predict exposure levels. Proceedings of the National Academy of Sciences. 2007;104(46):18211.

[83] Cheng Y, Church GM. Biclustering of expression data. In: Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology. vol. 8; 2000. p. 93.

[84] Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006;22(9):1122.

[85] Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4(5):P3.

[86] Al-Shahrour F, D´az-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004;20(4):578.

[87] Çak r T, Hendriks MMWB, Westerhuis JA, Smilde AK. Metabolic network discovery through reverse engineering of metabolome data. Metabolomics. 2009;5(3):318–329.

[88] González-D´az H, González-D´az Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics. 2008;8(4):750–778.

[89] de la Fuente A, Brazhnik P, Mendes P. Linking the genes: inferring quantitative gene networks from microarray data. Trends in genetics. 2002;18(8):395–398.

[90] Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome biology. 2006;7(5):R36.

[91] Van Someren E, Vaes B, Steegenga W, Sijbers A, Dechering K, Reinders M. Least absolute regression network analysis of the murine osteoblast differentiation network. Bioinformatics. 2005;22(4):477.

[92] Bussemaker HJ, Foat BC, Ward LD. Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annual review of biophysics and biomolecular structure. 2007;36(1):329.

[93] Peer D. Bayesian network analysis of signaling networks: a primer. Sci STKE. 2005;2005(281):l4.

[94] Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics. 2004;20(18):3594.

[95] Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, et al. Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics. 2004;20(9):1361.

[96] Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. Bioinformatics. 2008;24(7):932.

[97] Kamber M, Han J. Data mining: Concepts and techniques. Morgan Kaufmann Publishers; 2001.

[98] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics. 2006;7(Suppl 1):S7.

[99] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology. 2007;5(1):e8.

[100] Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology. 2007;2007:8–8.

[101] Sameith K, Antczak P, Marston E, Turan N, Maier D, Stankovic T, et al. Functional Modules integrating essential cellular functions are predictive of the response of leukaemia cells to DNA damage. Bioinformatics. 2008;24(22):2602.

[102] Bajorath J. Integration of virtual and high-throughput screening. Nature Reviews Drug Discovery. 2002;1(11):882–894.

[103] Auer CM, Nabholz JV, Baetcke KP. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. Environmental Health Perspectives. 1990;87:183.

[104] Matthews EJ, Kruhlak NL, Daniel Benz R, Sabaté DA, Marchant CA, Contrera JF. Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part C: Use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities. Regulatory toxicology and pharmacology. 2009;54(1):43–65.

[105] Sanderson D, Earnshaw C. Computer prediction of possible toxic action from chemical structure; the DEREK system. Human & experimental toxicology. 1991;10(4):261.

[106] Cariello NF, Wilson JD, Britt BH, Wedd DJ, Burlinson B, Gombar V. Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. Mutagenesis. 2002;17(4):321.

[107] Wiener H. Structural determination of paraffin boiling points. Journal of the American Chemical Society. 1947;69(1):17–20.

[108] Randic M. Characterization of molecular branching. Journal of the American Chemical Society. 1975;97(23):6609–6615.

[109] Balaban AT. Highly discriminating distance-based topological index. Chemical Physics Letters. 1982;89(5):399–404.

[110] Schultz HP. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. Journal of Chemical Information and Computer Sciences. 1989;29(3):227–228.

[111] Kier LB, Hall LH. Derivation and significance of valence molecular connectivity. Journal of Pharmaceutical Sciences. 1981;70(6):583–589.

[112] Gálvez J, Garcia R, Salabert M, Soler R. Charge indexes. New topological descriptors. Journal of Chemical Information and Computer Sciences. 1994;34(3):520–525.

[113] Mülliken R. Electronic Population Analysis on LCAO [Single Bond] MO Molecular Wave Functions. I. J Chem Phys. 1955;23:1833.

[114] Stanton DT, Egolf LM, Jurs PC, Hicks MG. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. Journal of chemical information and computer sciences. 1992;32(4):306–316.

[115] Cammarata A. An Apparent Correlation between the in Vitro Activity of Chloramphenicol Analogs and Electronic Polarizability1. Journal of Medicinal Chemistry. 1967;10(4):525–527.

[116] Pearlman RS, Smith K. Novel software tools for chemical diversity. Perspectives in Drug Discovery and Design. 1998;9:339–353.

[117] Stanton DT. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. Journal of chemical information and computer sciences. 1999;39(1):11–20.

[118] Burden FR. Molecular identification number for substructure searches. Journal of Chemical Information and Computer Sciences. 1989;29(3):225–227.

[119] Labute P. A widely applicable set of descriptors. Journal of Molecular Graphics and Modelling. 2000;18(4-5):464–477.

[120] Higo JI, Gō N. Algorithm for rapid calculation of excluded volume of large molecules. Journal of Computational Chemistry. 1989;10(3):376–379.

[121] Katritzky AR, Mu L, Lobanov VS, Karelson M. Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. The Journal of Physical Chemistry. 1996;100(24):10400–10407.

[122] Rohrbaugh RH, Jurs PC. Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. Analytica Chimica Acta. 1987;199:99–109.

[123] Cramer III RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. Journal of the American Chemical Society. 1988;110(18):5959–5967.

[124] Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. Journal of Medicinal Chemistry. 1994;37(24):4130–4146.

[125] Silverman B, Platt DE. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. Journal of medicinal chemistry. 1996;39(11):2129–2140.

[126] Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. Theory. Journal of chemometrics. 1994;8(4):263–272.

[127] Gramatica P, Navas N, Todeschini R. 3D-modelling and prediction by WHIM descriptors. Part. 1998;9:53–63.

[128] Cruciani G, Crivori P, Carrupt PA, Testa B. Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. Journal of Molecular Structure: THEOCHEM. 2000;503(1-2):17–30.

[129] Crivori P, Cruciani G, Carrupt PA, Testa B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. Journal of medicinal chemistry. 2000;43(11):2204–2216.

[130] Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. Journal of medicinal chemistry. 2000;43(17):3233–3243.

[131] Verdu-Andres J, Massart DL. Comparison of prediction-and correlation-based methods to select the best subset of principal components for principal component regression and detect outlying objects. Applied spectroscopy. 1998;52(11):1425–1434.

[132] Lin TH, Li HT, , Tsai KC. Implementing the Fisher s Discriminant Ratio in ak-Means Clustering Algorithm for Feature Selection and Data Set Trimming. Journal of chemical information and computer sciences. 2004;44(1):76–87.

[133] Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association. 1951;p. 68–78.

[134] Hou T, Zhang W, Xia K, Qiao X, Xu X. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. Journal of chemical information and computer sciences. 2004;44(5):1585–1600.

[135] Cronin M, Gregory B, Schultz TW. Quantitative structure-activity analyses of nitrobenzene toxicity to Tetrahymena pyriformis. Chemical research in toxicology. 1998;11(8):902–908.

[136] Wold S, Ruhe A, Wold H, Dunn III W. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing. 1984;5:735.

[137] Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems. 2001;58(2):109–130.

[138] Sun H. A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. Journal of chemical information and computer sciences. 2004;44(2):748–757.

[139] Adenot M, Lahana R. Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates. Journal of chemical information and computer sciences. 2004;44(1):239–248.

[140] Feng J, Lurati L, Ouyang H, Robinson T, Wang Y, Yuan S, et al. Predictive toxicology: benchmarking molecular descriptors and statistical methods. Journal of chemical information and computer sciences. 2003;43(5):1463–1470.

[141] Fisher RA. The use of multiple measures in taxonomic problems. Ann Eugenics. 1936;7(179-188):12–56.

[142] Guha R, Jurs PC. Determining the validity of a QSAR model-A classification approach. Journal of chemical information and modeling. 2005;45(1):65–73.

[143] Murcia-Soler M, Pérez-Giménez F, Garc´a-March FJ, Salabert-Salvador MT, D´az-Villanueva W, Castro-Bleda MJ, et al. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. Journal of chemical information and computer sciences. 2004;44(3):1031–1041.

[144] Molina E, D´az HG, González MP, Rodr´guez E, Uriarte E. Designing antibacterial compounds through a topological substructural approach. Journal of chemical information and computer sciences. 2004;44(2):515–521.

[145] Mazzatorta P, Benfenati E, Lorenzini P, Vighi M. QSAR in ecotoxicity: an overview of modern classification techniques. Journal of chemical information and computer sciences. 2004;44(1):105–112.

[146] Cover T, Hart P. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on. 1967;13(1):21–27.

[147] Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer. 1996;29(3):31–44.

[148] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

[149] Burges CJC. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998;2(2):121–167.

[150] Kirkpatrick S, Gelatt Jr C, Vecchi M. Optimization by simulated annealing, Readings in computer vision: issues, problems, principles, and paradigms. Morgan Kaufmann Publishers Inc., San Francisco, CA; 1987.

[151] Tino P, Nabney IT, Williams BS, Lösel J, Sun Y. Nonlinear prediction of quantitative structure-activity relationships. Journal of chemical information and computer sciences. 2004;44(5):1647–1653.

[152] Alex JS, Schoelkopf B. A tutorial on support vector regression. Statistics and Computing. 2004;14(2):199–222.

[153] Müller KR, Rätsch G, Sonnenburg S, Mika S, Grimm M, Heinrich N. Classifying drug-likeness with kernel-based learning methods. Journal of chemical information and modeling. 2005;45(2):249–253.

[154] Yao X, Panaye A, Doucet J, Zhang R, Chen H, Liu M, et al. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. Journal of chemical information and computer sciences. 2004;44(4):1257–1266.

[155] Quinlan JR. Induction of decision trees. Machine learning. 1986;1(1):81–106.

[156] Gelfand SB, Ravishankar C, Delp EJ. An iterative growing and pruning algorithm for classification tree design. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1991;p. 163–174.

[157] Daszykowski M, Walczak B, Xu QS, Daeyaert F, de Jonge M, Heeres J, et al. Classification and Regression Trees Studies of HIV Reverse Transcriptase Inhibitors. Journal of chemical information and computer sciences. 2004;44(2):716–726.

[158] DeLisle RK, Dixon SL. Induction of decision trees via evolutionary programming. Journal of chemical information and computer sciences. 2004;44(3):862–870.

[159] Meir R, Rätsch G. An introduction to boosting and leveraging. Advanced lectures on machine learning. 2003;p. 118–183.

[160] Breiman L. Bagging predictors. Machine learning. 1996;24(2):123–140.

[161] Ho TK. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1998;20(8):832–844.

[162] Freund Y, Schapire R. A decision-theoretic generalization of online learning. Comput. System Sci. 1997;55(1):119–139.

[163] Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

[164] Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. Journal of Chemical Information and Modeling. 2005;45(3):786–799.

[165] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences. 2003;43(6):1947–1958.

[166] Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence. 1999;14(771-780):1612.

[167] Schapire RE, Freund Y, Bartlett P, Lee WS. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of statistics. 1998;p. 1651–1686.

[168] Michalewicz Z. Genetic algorithms + data structures. Springer; 1996.

[169] Turan N, Kalko S, Stincone A, Clarke K, Sabah A, Howlett K, et al. A Systems Biology Approach Identifies Molecular Networks Defining Skeletal Muscle Abnormalities in Chronic Obstructive Pulmonary Disease. PLoS Computational Biology. 2011;7(9):e1002129.

[170] Bundy J, Sidhu J, Rana F, Spurgeon D, Svendsen C, Wren J, et al. Systems toxicology approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm Lumbricus rubellus. BMC biology. 2008;6(1):25.

[171] Taylor NS, Weber RJM, White TA, Viant MR. Discriminating between different acute chemical toxicities via changes in the daphnid metabolome. Toxicological Sciences. 2010;118(1):307.

[172] Owen JR, Morris CA, Nicolaus B, Harwood JL, Kille P. Induction of expression of a 14-3-3 gene in response to copper exposure in the marine alga, Fucus vesiculosus. Ecotoxicology. 2011;p. 1–15.

[173] Vedani A, Dobler M, Lill MA. The challenge of predicting drug toxicity in silico. Basic and Clinical Pharmacology and Toxicology. 2006;99(3):195.

[174] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28(1):27.

[175] Kavlock RJ, Ankley G, Blancato J, Breen M, Conolly R, Dix D, et al. Computational toxicologya state of the science mini review. Toxicological sciences. 2008;103(1):14.

[176] Mayr LM, Bojanic D. Novel trends in high-throughput screening. Current Opinion in Pharmacology. 2009;9(5):580–588.

[177] Conesa A, Bro R, Garc´a-Garc´a F, Prats JM, Götz S, Kjeldahl K, et al. Direct functional assessment of the composite phenotype through multivariate projection strategies. Genomics. 2008;92(6):373–383.

[178] Amin RP, Hamadeh HK, Bushel PR, Bennett L, Afshari CA, Paules RS. Genomic interrogation of mechanism (s) underlying cellular responses to toxicants. Toxicology. 2002;181:555–563.

[179] Waring JF, Gum R, Morfitt D, Jolly RA, Ciurlionis R, Heindel M, et al. Identifying toxic mechanisms using DNA microarrays: evidence that an experimental inhibitor of cell adhesion molecule expression signals through the aryl hydrocarbon nuclear receptor. Toxicology. 2002;181:537–550.

[180] Lobenhofer E, Auman JT, Blackshear P, Boorman G, Bushel P, Cunningham M, et al. Gene expression response in target organ and whole blood varies as a function of target organ injury phenotype. Genome biology. 2008;9(6):R100.

[181] Bushel PR, Hamadeh HK, Bennett L, Green J, Ableson A, Misener S, et al. Computational selection of distinct class-and subclass-specific gene expression signatures. Journal of biomedical informatics. 2002;35(3):160–170.

[182] Dieterle F, Marrer E, Suzuki E, Grenet O, Cordier A, Vonderscher J. Monitoring kidney safety in drug development: emerging technologies and their implications. Current opinion in drug discovery & development. 2008;11(1):60.

[183] Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3/4):325–338.

[184] Schoeppe W. Effects of dopamine on kidney function. Proceedings of the Royal Society of Medicine. 1977;70(Suppl 2):36.

[185] Freeman JW, DeArmond D, Lake M, Huang W, Venkatasubbarao K, Zhao S. Alterations of cell signaling pathways in pancreatic cancer. Front Biosci. 2004;9:1889–98.

[186] Geng H, Lan R, Wang G, Siddiqi AR, Naski MC, Brooks AI, et al. Inhibition of Autoregulated TGF beta Signaling Simultaneously Enhances Proliferation and Differentiation of Kidney Epithelium and Promotes Repair Following Renal Ischemia. American Journal of Pathology. 2009;174(4):1291.

[187] Chipman JK, Mally A, Edwards GO. Disruption of gap junctions in toxicity and carcinogenicity. Toxicological Sciences. 2003;71(2):146.

[188] Pereira CV, Moreira AC, Pereira SP, Machado NG, Carvalho FS, Sardao VA, et al. Investigating Drug-induced Mitochondrial Toxicity: A Biosensor to Increase Drug Safety? Current Drug Safety. 2009;4(1):34–54.

[189] Klaassen CD, Amdur MO, et al. Casarett and Doull s toxicology: the basic science of poisons. Journal of Occupational and Environmental Medicine. 1993;35(1):76.

[190] Doi AM, Hill G, Seely J, Hailey JR, Kissling G, Bucher JR. α2u-Globulin Nephropathy and Renal Tumors in National Toxicology Program Studies. Toxicologic pathology. 2007;35:533–540.

[191] Judson R, Richard A, Dix DJ, Houck K, Martin M, Kavlock R, et al. The toxicity data landscape for environmental chemicals. Environmental Health Perspectives. 2009;117(5):685.

[192] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. NIH Public Access; 2000. p. 455.

[193] Team RDC. R: A language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria. 2005;.

[194] Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics. 2006;22(19):2373.

[195] Song S, Black M. Principal Coordinates and Hotelling s T^2 method.; 2006. Available from: `http://bioconductor.org/packages/bioc/html/pcot2.html`. (accessed 2010).

[196] Song S, Black MA. Microarray-based gene set analysis: a comparison of current methods. BMC bioinformatics. 2008;9(1):502.

[197] Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, et al. Virtual computational chemistry laboratory–design and description. Journal of computer-aided molecular design. 2005;19(6):453–463.

[198] Trevino V, Falciani F. GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics. 2006;22(9):1154.

[199] Directive WF. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities. 2000;22(12):2000.

[200] Ankley G, Hockett J, Mount D, Mount D. Early evolution of the toxicity identification evaluation process: contributions from the United States environmental protection agency effluent testing program. Effect-Directed Analysis of Complex Environmental Contamination. 2011;p. 1–18.

[201] Agency UE. The direct toxicity assessment of aqueous environmental samples using the juvenile *Daphnia magna* immobilisation test (2007);.

[202] Garcia-Reyero N, Villeneuve DL, Kroll KJ, Liu L, Orlando EF, Watanabe KH, et al. Expression signatures for a model androgen and antiandrogen in the fathead minnow (Pimephales promelas) ovary. Environmental science & technology. 2009;43(7):2614–2619.

[203] Villeneuve DL, Garcia-Reyero N, Martinovic D, Cavallin JE, Mueller ND, Wehmas LC, et al. Influence of ovarian stage on transcript profiles in fathead minnow (Pimephales promelas) ovary tissue. Aquatic Toxicology. 2010;98(4):354–366.

[204] Suter L, Schroeder S, Meyer K, Gautier JC, Amberg A, Wendt M, et al. EU Framework 6 Project: Predictive Toxicology (PredTox)-Overview and Outcome. Toxicology and applied pharmacology. 2010;.

[205] Kulkarni K, Larsen P, Linninger AA. Assessing chronic liver toxicity based on relative gene expression data. Journal of theoretical biology. 2008;254(2):308–318.

[206] Stincone A, Daudi N, Rahman AS, Antczak P, Henderson I, Cole J, et al. A systems biology approach sheds new light on Escherichia coli acid resistance. Nucleic Acids Research. 2011;.

[207] Shaw J, Colbourne J, Davey J, Glaholt S, Hampton T, Chen C, et al. Gene response profiles for Daphnia pulex exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins. BMC genomics. 2007;8(1):477.

[208] Heckmann LH, Sibly RM, Connon R, Hooper HL, Hutchinson TH, Maund SJ, et al. Systems biology meets stress ecology: linking molecular and organismal stress responses in Daphnia magna. Genome biology. 2008;9(2):R40.

[209] Antczak P, Ortega F, Chipman JK, Falciani F. Mapping Drug Physico-Chemical Features to Pathway Activity Reveals Molecular Networks Linked to Toxicity Outcome. PloS one. 2010;5(8):e12385.

[210] Olmstead AW, LeBlanc GA. Effects of endocrine-active chemicals on the development of sex characteristics of Daphnia magna. Environmental toxicology and chemistry. 2000;19(8):2107–2113.

[211] Dansen TB, Wirtz KWA. The peroxisome in oxidative stress. IUBMB life. 2001;51(4):223–230.

[212] Chen Q, Vazquez EJ, Moghaddas S, Hoppel CL, Lesnefsky EJ. Production of reactive oxygen species by mitochondria. Journal of Biological Chemistry. 2003;278(38):36027.

[213] Thannickal VJ, Fanburg BL. Reactive oxygen species in cell signaling. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2000;279(6):L1005.

[214] Shimozawa N. Molecular and clinical aspects of peroxisomal diseases. Journal of inherited metabolic disease. 2007;30(2):193–197.

[215] Funato M, Shimozawa N, Nagase T, Takemoto Y, Suzuki Y, Imamura Y, et al. Aberrant peroxisome morphology in peroxisomal beta-oxidation enzyme deficiencies. Brain and Development. 2006;28(5):287–292.

[216] Wanders R, Ferdinandusse S, Brites P, Kemp S. Peroxisomes, lipid metabolism and lipotoxicity. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids. 2010;1801(3):272–280.

[217] Fowler BA. General subcellular effects of lead, mercury, cadmium, and arsenic. Environmental health perspectives. 1978;22:37.

[218] Chvapil M, Ryan JN, Zukoski C. The effect of zinc and other metals on the stability of lysosomes. Experimental Biology and Medicine. 1972;140(2):642.

[219] Ndolo RA, Forrest ML, Krise JP. The role of lysosomes in limiting drug toxicity in mice. Journal of Pharmacology and Experimental Therapeutics. 2010;333(1):120.

[220] Kawata K, Yokoo H, Shimazaki R, Okabe S. Classification of heavy-metal toxicity by human DNA microarray analysis. Environmental science & technology. 2007;41(10):3769–3774.

[221] Hughes MF. Arsenic toxicity and potential mechanisms of action. Toxicology letters. 2002;133(1):1–16.

[222] Hu Y, Su L, Snow ET. Arsenic toxicity is enzyme specific and its affects on ligation are not caused by the direct inhibition of DNA repair enzymes1. Mutation Research/DNA Repair. 1998;408(3):203–218.

[223] Peters R. Biochemistry of some toxic agents. I. Present state of knowledge of biochemical lesions induced by trivalent arsenical poisoning. Bulletin of the Johns Hopkins Hospital. 1955;97(1):1.

[224] Szinicz L, Forth W. Effect of As2O3 on gluconeogenesis. Archives of toxicology. 1988;61(6):444.

[225] Wang Y, Fang J, Leonard SS, Krishna Rao KM. Cadmium inhibits the electron transfer chain and induces reactive oxygen species. Free Radical Biology and Medicine. 2004;36(11):1434–1443.

[226] Kim SH, Turnbull J, Guimond S. Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. Journal of Endocrinology. 2011;209(2):139.

[227] Blaine SA, Ray KC, Branch KM, Robinson PS, Whitehead RH, Means AL. Epidermal growth factor receptor regulates pancreatic fibrosis. American Journal of Physiology-Gastrointestinal and Liver Physiology. 2009;297(3):G434.

[228] Andersen HR, Wollenberger L, Halling-Sørensen B, Kusk KO. Development of copepod nauplii to copepoditesa parameter for chronic toxicity including endocrine disruption. Environmental toxicology and chemistry. 2001;20(12):2821–2829.

[229] Duncan BJ. The Environmental Protection Agencys (EPA) Endocrine Disruptor Screening Program (EDSP). Focus On. 2010;49:13.

[230] Morgado J, Soares A. Activity of pyruvate kinase and malate dehydrogenase in Daphnia magna under 3, 4-dichloroaniline stress. Archives of Environmental Contamination and Toxicology. 1995;29(1):94–96.

[231] Kilham SS, Kreeger DA, Lynn SG, Goulden CE, Herrera L. COMBO: a defined freshwater culture medium for algae and zooplankton. Hydrobiologia. 1998;377(1):147–159.

[232] Weber CI, Environmental Monitoring Systems Laboratory (Cincinnati O, Agency USEP. Methods for measuring the acute toxicity of effluents and receiving waters to freshwater and marine organisms. Environmental Monitoring Systems Laboratory, Office of Research and Development, US Environmental Protection Agency; 1991.

[233] Lewis PA, Environmental Monitoring Systems Laboratory (Cincinnati O, of Research USEPAO, Development. Short-term methods for estimating the chronic toxicity of effluents and receiving waters to freshwater organisms. US Environmental Protection Agency, Environmental Monitoring Systems Laboratory; 1994.

[234] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2010. ISBN 3-900051-07-0. Available from: `http://www.R-project.org/`.

[235] Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002;2(3):18–22.

[236] EPA US. What Are the Trends in Chemicals Used on the Land and Their Effects on Human Health and the Environment? US EPA 2008 Report on the Environment. 2008;(4.5):29–41.

[237] Jobling S, Nolan M, Tyler CR, Brighty G, Sumpter JP. Widespread sexual disruption in wild fish. Environmental Science & Technology. 1998;32(17):2498–2506.

[238] Iguchi T, Watanabe H, Katsu Y. Developmental effects of estrogenic agents on mice, fish, and frogs: a mini-review. Hormones and behavior. 2001;40(2):248–251.

[239] Iguchi T, Sato T. Endocrine disruption and developmental abnormalities of female reproduction. American Zoologist. 2000;40(3):402.

[240] Shioda T, Wakabayashi M. Effect of certain chemicals on the reproduction of medaka (Oryzias latipes). Chemosphere. 2000;40(3):239–243.

[241] Suzuki A, Sugihara A, Uchida K, Sato T, Ohta Y, Katsu Y, et al. Developmental effects of perinatal exposure to bisphenol-A and diethylstilbestrol on reproductive organs in female mice. Reproductive Toxicology. 2002;16(2):107–116.

[242] Geeraerts C, Belpaire C. The effects of contaminants in European eel: a review. Ecotoxicology. 2010;19(2):239–266.

[243] Van der Oost R, Beyer J, Vermeulen NPE. Fish bioaccumulation and biomarkers in environmental risk assessment: a review. Environmental Toxicology and Pharmacology. 2003;13(2):57–149.

[244] Baurin N, Mozziconacci JC, Arnoult E, Chavatte P, Marot C, Morin-Allory L. 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. Journal of chemical information and computer sciences. 2004;44(1):276–285.

[245] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13(11):2498.

[246] Tsakiridis T, McDowell H, Walker T, Downes C, Hundal H, Vranic M, et al. Multiple roles of phosphatidylinositol 3-kinase in regulation of glucose transport, amino acid transport, and glucose transporters in L6 skeletal muscle cells. Endocrinology. 1995;136(10):4315.

[247] Slusarski DC, Corces VG, Moon RT. Interaction of Wnt and a Frizzled homologue triggers G-protein-linked phosphatidylinositol signalling. Nature. 1997;390(6658):410–413.

[248] Ensenat D, Hassan S, Reyna SV, Schafer AI, Durante W. Transforming growth factor-beta 1 stimulates vascular smooth muscle cell L-proline transport by inducing system A amino acid transporter 2 (SAT2) gene expression. Biochemical Journal. 2001;360(Pt 2):507.

[249] Durante W, Liao L, Reyna SV, Peyton KJ, Schafer AI. Transforming Growth Factor-beta 1 Stimulates L-Arginine Transport and Metabolism in Vascular Smooth Muscle Cells: Role in Polyamine and Collagen Synthesis. Circulation. 2001;103(8):1121.

[250] Subramanian M, Kuang PP, Wei L, Rishikof DC, Liu H, Goldstein RH. Modulation of amino acid uptake by TGF-β in lung myofibroblasts. Journal of cellular biochemistry. 2006;99(1):71–78.

[251] Boerner P, Resnick RJ, Racker E. Stimulation of glycolysis and amino acid uptake in NRK-49F cells by transforming growth factor beta and epidermal growth factor. Proceedings of the National Academy of Sciences. 1985;82(5):1350.

[252] Nicotera P, Bellomo G, Orrenius S. Calcium-mediated mechanisms in chemically induced cell death. Annual review of pharmacology and toxicology. 1992;32(1):449–470.

[253] Komulainen H, Bondy SC. Increased free intrasynaptosomal Ca2+ by neurotoxic organometals: distinctive mechanisms. Toxicology and applied pharmacology. 1987;88(1):77–86.

[254] Bondy S, Komulainen H. Intracellular calcium as an index of neurotoxic damage. Toxicology. 1988;49(1):35–41.

[255] Orrenius S, McCabe Jr MJ, Nicotera P. Ca2+-dependent mechanisms of cytotoxicity and programmed cell death. Toxicology letters. 1992;64:357–364.

[256] Claybrook D. Nitrogen metabolism. In, The Biology of Crustacea, DE Bliss (Ed), Vol. 5, Internal anatomy and physiological regulation, LH Mantel. New York: Academic Press; 1983.

[257] Ventura M. Linking biochemical and elemental composition in freshwater and marine crustacean zooplankton. Marine Ecology Progress Series. 2007;327:233–246.

[258] Graney RL, Giesy JP, et al. Effects of long-term exposure to pentachlorophenol on the free amino acid pool and energy reserves of the freshwater amphipod Gammarus pseudolimnaeus Bousfield (Crustacea, Amphipoda). Ecotoxicology and environmental safety. 1986;12(3):233–251.

[259] Pritchard JB. Toxic substances and cell membrane function. In: Federation proceedings. vol. 38; 1979. p. 2220.

[260] Turski L, Bressler K, Rettig KJ, Löschmann PA, Wachtel H. Protection of substantia nigra from MPP+ neurotoxicity by N-methyl-D-aspartate antagonists. 1991;.

[261] Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D. How to infer gene networks from expression profiles. Molecular systems biology. 2007;3(1).

[262] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences. 2007;104(6):1777.

[263] Ramsey SA, Klemm SL, Zak DE, Kennedy KA, Thorsson V, Li B, et al. Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS computational biology. 2008;4(3):e1000021.

[264] Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: Data integration in dynamic models–A review. Biosystems. 2009;96(1):86–103.

[265] Singh SM, Gauthier S, Labrie F. Androgen receptor antagonists (antiandrogens) structure-activity relationships. Current medicinal chemistry. 2000;7(2):211–247.

[266] Bayley M, Junge M, Baatrup E. Exposure of juvenile guppies to three antiandrogens causes demasculinization and a reduced sperm count in adult males. Aquatic toxicology. 2002;56(4):227–239.

[267] Hagino S, Kagoshima M, Ashida S. Effects of ethinylestradiol, diethylstilbestrol, 4-t-pentylphenol, 17β-estradiol, methyltestosterone and flutamide on sex reversal in S-rR strain medaka (Oryzias latipes). Environ Sci. 2001;8(1):75–87.

[268] Jensen KM, Kahl MD, Makynen EA, Korte JJ, Leino RL, Butterworth BC, et al. Characterization of responses to the antiandrogen flutamide in a short-term reproduction assay with the fathead minnow. Aquatic toxicology. 2004;70(2):99–110.

[269] Gray Jr LE, Wolf C, Lambright C, Mann P, Price M, Cooper RL, et al. Administration of potentially antiandrogenic pesticides (procymidone, linuron, iprodione, chlozolinate, p, p -DDE, and ketoconazole) and toxic substances (dibutyl-and diethylhexyl phthalate, PCB 169, and ethane dimethane sulphonate) during sexual differentiation produces diverse profiles of reproductive malformations in the male rat. Toxicology and industrial health. 1999;15(1-2):94.

[270] Wells K, Van Der Kraak G. Differential binding of endogenous steroids and chemicals to androgen receptors in rainbow trout and goldfish. Environmental toxicology and chemistry. 2000;19(8):2059–2065.

[271] Ankley GT, Defoe DL, Kahl MD, Jensen KM, Makynen EA, Miracle A, et al. Evaluation of the model anti-androgen flutamide for assessing the mechanistic basis of responses to an androgen in the fathead minnow (Pimephales promelas). Environmental science & technology. 2004;38(23):6322–6327.

[272] Meyer P, Lafitte F, Bontempi G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. BMC bioinformatics. 2008;9(1):461.

[273] Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A. Reverse engineering cellular networks. Nat Protoc. 2006;1(2):662–671.

[274] Daub C, Steuer R, Selbig J, Kloska S. Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. BMC bioinformatics. 2004;5(1):118.

[275] Steuer R, Daub C, Selbig J, Kurths J. Measuring distances between variables by mutual information. Innovations in Classification, Data Science, and Information Systems. 2005;p. 81–90.

[276] Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics. 2003;4(1):2.

[277] Maggiolini M, Donzé O, Jeannin E, Ando S, Picard D. Adrenal androgens stimulate the proliferation of breast cancer cells as direct activators of estrogen receptor α. Cancer research. 1999;59(19):4864.

[278] Ohtake F, Takeyama K, Matsumoto T, Kitagawa H, Yamamoto Y, Nohara K, et al. Modulation of oestrogen receptor signalling by association with the activated dioxin receptor. Nature. 2003;423(6939):545–550.

[279] Claiborne J, Nag S, Mokha SS. Activation of opioid receptor like-1 receptor in the spinal cord produces sex-specific antinociception in the rat: estrogen attenuates antinociception in the female, whereas testosterone is required for the expression of antinociception in the male. The Journal of neuroscience. 2006;26(50):13048.

[280] Shved N, Berishvili G, Baroiller JF, Segner H, Reinecke M. Environmentally Relevant Concentrations of 17α-Ethinylestradiol (EE2) Interfere With the Growth Hormone (GH)/Insulin-Like Growth Factor (IGF)-I System in Developing Bony Fish. Toxicological sciences. 2008;106(1):93.

[281] Quint E, Smith A, Avaron F, Laforest L, Miles J, Gaffield W, et al. Bone patterning is altered in the regenerating zebrafish caudal fin after ectopic expression of sonic hedgehog and bmp2b or exposure to cyclopamine. Proceedings of the National Academy of Sciences. 2002;99(13):8713.

[282] Chuang PT, Kornberg TB. On the range of hedgehog signaling. Current opinion in genetics & development. 2000;10(5):515–522.

[283] Johnson R, Tabin C. The long and short of hedgehog signaling. Cell. 1995;81:313–316.

[284] Varjosalo M, Taipale J. Hedgehog: functions and mechanisms. Genes & development. 2008;22(18):2454.

[285] Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Systems Biology. 2007;1(1):8.

[286] Rivera C, Vakil R, Bader J. NeMo: network module identification in Cytoscape. BMC bioinformatics. 2010;11(Suppl 1):S61.

[287] Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH. Modular organization of protein interaction networks. Bioinformatics. 2007;23(2):207.

[288] Nemenman I, Bialek W, Van Steveninck RDR. Entropy and information in neural spike trains: Progress on the sampling problem. Physical Review E. 2004;69(5):056111.

[289] Tourassi GD, Frederick ED, Markey MK, Floyd Jr CE. Application of the mutual information criterion for feature selection in computer-aided diagnosis. Medical Physics. 2001;28:2394.

[290] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence. 2005;p. 1226–1238.

[291] Saeed A, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003;34(2):374.

[292] Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, et al. [9] TM4 Microarray Software Suite. Methods in enzymology. 2006;411:134–193.

[293] on Toxicity Testing NRCUC, of Environmental Agents A. Toxicity testing in the 21st century: A vision and a strategy. Natl Academy Pr; 2007.

[294] Nielsen J, Vidal M. Systems biology of microorganisms Editorial overview. Current Opinion in Microbiology. 2010;13:1–2.

[295] Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Molecular systems biology. 2008;4(1).

[296] Villoslada P, Steinman L, Baranzini SE. Systems biology and its application to the understanding of neurological diseases. Annals of neurology. 2009;65(2):124–139.

[297] Valerio Jr LG. In silico toxicology for the pharmaceutical sciences. Toxicology and applied pharmacology. 2009;241(3):356–370.

[298] Houck KA, Kavlock RJ. Understanding mechanisms of toxicity: Insights from drug discovery research. Toxicology and applied pharmacology. 2008;227(2):163–178.

[299] Pujol A, Mosca R, Farrés J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends in pharmacological sciences. 2010;31(3):115–123.

[300] Abbott A. Pharmaceutical futures: a fiendish puzzle. Nature. 2008 Oct;455:1164–1167.

[301] Fentem J, Archer G, Balls M, Botham P, Curren R, Earl L, et al. The ECVAM international validation study on in vitro tests for skin corrosivity. 2. Results and evaluation by the Management Team. Toxicology in vitro. 1998;12(4):483–524.

[302] Vinardell M, Mitjans M. Alternative methods for eye and skin irritation tests: an overview. Journal of pharmaceutical sciences. 2008;97(1):46–59.

[303] Lynch AM, Wilcox P. Review of the performance of the 3T3 NRU in vitro phototoxicity assay in the pharmaceutical industry. Experimental and Toxicologic Pathology. 2010;.

[304] Nirmalanandhan VS, Sittampalam GS. Stem cells in drug discovery, tissue engineering, and regenerative medicine: emerging opportunities and challenges. Journal of Biomolecular Screening. 2009;14(7):755.

[305] EPA US. EPA CompTox Program;. Available from: `http://www.epa.gov/ncct/`. Accessed 28.09.2011.

[306] EPA US. EPA ExpoCast Program;. Available from: `http://www.epa.gov/ncct/expocast/`. Accessed 28.09.2011.

[307] Schmidt CW. TOX 21: new dimensions of toxicity testing. Environmental health perspectives. 2009;117(8):A348.

[308] EPA US. EPA Virtual Liver Program;. Available from: http://www.epa.gov/ncct/virtual_liver/. Accessed 28.09.2011.

[309] EPA US. EPA Virtual Embryo Program;. Available from: http://www.epa.gov/ncct/v-Embryo/. Accessed 28.09.2011.

[310] Craig A, Sidaway J, Holmes E, Orton T, Jackson D, Rowlinson R, et al. Systems toxicology: integrated genomic, proteomic and metabonomic analysis of methapyrilene induced hepatotoxicity in the rat. Journal of proteome research. 2006;5(7):1586–1601.

[311] Slikker Jr W, Paule MG, Wright LKM, Patterson TA, Wang C. Systems biology approaches for toxicology. Journal of applied toxicology. 2007;27(3):201–217.

[312] De Wit M, Keil D, Remmerie N, Ven K, Brandhof EJ, Knapen D, et al. Molecular targets of TBBPA in zebrafish analysed through integration of genomic and proteomic approaches. Chemosphere. 2008;74(1):96–105.

[313] De Coppi P, Bartsch G, Siddiqui MM, Xu T, Santos CC, Perin L, et al. Isolation of amniotic stem cell lines with potential for therapy. Nature biotechnology. 2007;25(1):100–106.

[314] Alon U. Network motifs: theory and experimental approaches. Nature Reviews Genetics. 2007;8(6):450–461.

[315] Muzzey D, van Oudenaarden A. When it comes to decisions, myeloid progenitors crave positive feedback. Cell. 2006;126(4):650–652.

[316] Inglese J, Johnson RL, Simeonov A, Xia M, Zheng W, Austin CP, et al. High-throughput screening assays for the identification of chemical probes. Nature chemical biology. 2007;3(8):466–479.

[317] Agendia. MammaPrint; 2007. Available from: http://www.agendia.com/. (accessed 2010).

[318] SP H. In Handbook of ecotoxicology. Wiley-Blackwell; 1997.