Systems Toxicology: Beyond Animal Models

by

Alexandra Maertens

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

September, 2014

ACKNOWLEDGMENTS

The path to my Ph.D. has been longer than most and I have accumulated a correspondingly large group of people to thank.

Dr. Peter Ranslow and Dr. Elizabeth Becker both helped me get back on my feet and back into my field, and I will always be thankful they gave me the time and support necessary to finish my Ph.D.

My first advisor, Dr. Joseph Bressler, started with me when I was a scientist who could barely pipette and showed no great talent for tissue culture, yet he endured with me through thick and thin and taught me not only how to be a good scientist, but also how to be a good teacher.

Dr. James Yager was not only instrumental in helping me back on track after a personal setback, but was unfailing in his encouragement and unstinting with his time.

Dr. Vanessa sa'Rocha's calm perseverance was both an inspiration and a model, and Thomas Luechtfeld's keen coding ability kept our Integrated Testing Strategy on track.

Dr. Andre Kleensang has always been a good sparring partner and a good friend, and most importantly, he always kept the statistics honest.

I will be forever grateful to Dr. Thomas Hartung's extraordinary generosity in giving me an opportunity to finish my Ph.D. It was my great fortune to have landed in one of the most intellectually stimulating and welcoming environments possible at CAAT.

Finally, this is dedicated to my son, Mikhail Maertens, who cannot remember a time when his mother wasn't trying to finish her Ph.D.

ABSTRACT

Toxicology – much like the rest of biology – is undergoing a profound change as new technologies begin to offer a more systems oriented view of cellular physiology. For toxicology in particular, this means moving away from black-box animal models that provide limited information about mechanisms of toxicity towards the use of *in vitro* approaches which can both expedite hazard assessment while at the same time providing a more data – rich insight into toxic effects at the molecular level. One motivator of this shift is Green Toxciology, which seeks to support the Green Chemistry movement.

In order for this approach to succeed, it will require two separate but parallel efforts. The first is an Integrated Testing Strategy which seeks to use machine learning and data mining techniques to combine QSARs and *in vitro* tests in the most efficient way possible to accurately estimate hazard, which is discussed both theoretically and demonstrated practically with the example of skin sensitization. Secondly, toxicology will require new approaches that exploit the insights of network biology to look at toxic mechanisms from a systems perspective. The theoretical concept of a Pathway of Toxicity is outlined, and an example of how to extract a suggested Pathway of Toxicity is given, using a Weighted Gene Correlation Network Analysis of a small microarray study of MPTP toxicity combined with text-mining and other high-throughput data to suggest novel candidate transcription factors and proteins. In conclusion, it discusses some of the current limitations of another promising –omics technology, metabolomics.

September, 2014

ALEXANDRA MAERTENS, BS

Directed by: Professor Thomas Hartung, M.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I – Introduction: Green Toxicology as a Motivator for Systems Toxicology

> "Complexity is a term that is inversely related to the degree of understanding." —Y. Lazebnik (Lazebnik, 2004)

One common argument against the use of models in biology (and specifically, the use of computational or *in vitro* models in toxicology) is the argument of complexity—that cells (and if not cells, certainly organisms) are too complex to be captured by any abstract approach. But a system that seems complex is merely a system that is operating according to laws not yet fully understood, not laws that are impossible to learn. The counter argument, then, is that in order to understand the system, one must be able to model it—that is, one must be able to capture enough of the system to describe and predict its behavior. Given both the technological revolutions that have dramatically changed the life sciences in the last few decades, and the concomitant computational advances, complexity is no longer an acceptable excuse for using a black-box model.

While the push towards computational and *in vitro* testing comes partially from humane concerns regarding animal welfare, there is a parallel motivation for more efficient toxicology coming from the world of hazard assessment—in particular, as the pace of innovation in the chemical industry increases, there needs to be a corresponding increase in the ability of toxicology to estimate the hazard of novel chemicals (discussed more extensively in Appendix I). The field of chemistry has been undergoing a slow revolution to a more sustainable, and environmentally efficient "green" chemistry; for toxicologist to be able to keep pace, it is necessary to have a "green" toxicology. Lengthy chemical tests that require years cannot be effectively be used for front-loading toxicity testing at the beginning of the R&D process, and the current regulatory testing paradigm, which relies largely on "black-box" animal models, provides little to no information that can be useful to a chemist seeking to design a less toxic chemical replacement. For toxicologists to be able to offer some guidance to chemists seeking to design more benign alternatives, it is

necessary to specify, as completely as possible, the molecular mechanisms of toxicity.

As it stands now, the knowledge is not often available, and when it is available, it is often not accessible, as no database effectively catalogs the known molecular mechanisms of toxicity. Consider, for a moment, a chemist seeking to design an alternative to BPA – an endocrine disruptor that has been the subject of much dispute in toxicology. If she started with two of the more common databases – Chemicals of Biological Interest (CheBi) (Degtyarenko et al., 2008) or Toxin and Toxin Target (T3) database (Lim et al., 2010) she would find information that was either unspecific or poorly documented; Pubchem would have a list of molecular targets and potencies from Toxcast screens, and this (as well as the literature) would seem to indicate that the Estrogen Receptor Alpha (ER-Alpha) is the likely molecular target for BPA. On this basis, she might design an alternative – for example, BPS (bisphenol sulfate) that does not bind to the ER-Alpha receptor. Unfortunately, at least some of BPA's toxicity is thought to be mediated by bind to the Estrogen Receptor – Gamma (ER-Gamma) (Okada et al., 2008) and BPS has a higher binding affinity to ER-Gamma than BPA (Okada et al., 2008). However, this would only be clear after a copious search of the literature, and would be invisible in the Toxcast screens.

Furthermore, even if it were desirable to test every novel chemical with rigorous animal testing, the capacity is simply not available (Rovida & Hartung, 2009). The costs would not only be enormous (Bottini & Hartung, 2010) but would stifle the development of greener alternatives to known toxic compounds. Therefore, moving away from animal models towards an approach that both makes better use of computational approaches and more precisely specifies toxicity at the molecular and cellular level can both expedite hazard assessment but can also facilitate the development of chemicals that are "benign by design" (a concept discussed more extensively in Appendix I).

Currently, the most common computational approach to hazard assessment is a Quantitative Structure Active Relationship model (QSARs). While QSARs have

certainly proven their worth in some limited domains – for example, in aquatic toxicity (Voutchkova et al., 2011) and the Lipinski rules, which identifies drug-like compounds (Lipinski, 2004). However, QSARs typically only perform well when the molecular basis of toxicity is both simple and well understood. QSARs have limited usefulness for even a relatively simple toxicity mechanism such as skin sensitization, and they would almost certainly be inadequate for more complicated endpoints such as developmental neurotoxicity or endocrine disruption.

Just as QSARs have a useful domain so long as one is realistic about their limitations, at the same time, it is necessary to be realistic about the limited information an *in vitro* test can provide. No single *in vitro* test is likely to effectively replace an *in vivo* assay. Toxicity is often an emergent property of a complex system – often of multiple tissue types – and while an *in vitro* assay can perhaps mimic aspects of specific organs (e.g. skin permeability assays) or provide a read-out of a known molecular pathway (e.g. receptor binding assays), it cannot hope to effectively capture the complexity of a living system. Therefore, for computational toxicology to truly become a part of hazard assessment, it will require two separate, but parallel efforts. One, an Integrated Testing Strategy (discussed more fully in Appendix II) is needed to optimize the use of *in vitro* (and other) sources of information to predict hazard as accurately as possible while simultaneously respecting the probabilistic nature of the prediction and the possibility that a hazard estimation could be updated with additional information. Two, it is necessary for toxicology to adopt a more "systems biology" oriented approach to characterizing the molecular mechanisms that lead to adverse outcomes. In essence, this requires mapping the Human Toxome, or to put it another way, producing a model of cellular circuitry with sufficient accuracy that we can predict, with some confidence, where and how perturbations become severe enough to cause an altered phenotype.

The first goal, an ITS, is motivated by the clear need to move away from the commonly-used weight-of-evidence approach towards a more systematic methodology that uses machine learning and data-mining techniques to combine multiple sources of information (chemoinformatics, *in vitro* screening assays, and

potentially –omics technologies) in the most efficient way possible to accurately predict hazard while at the same time developing a framework that can quickly integrate new information. This is true not only because such an approach avoids the subjectivity and lack of precision endemic to a weight-of-evidence evaluation, but also because the shear explosion of newly available sources of information – e.g. from initiatives such as Toxcast (Dix et al., 2007)– produces a surfeit of data and cannot possible be processed by experts sitting around together in a room.

This abundance of data has its downside. Given the potentially large number of false positives in many high-throughput *in vitro* screens, it risks bringing the field of *in vitro* assays into disrepute. If each positive in a screening assay is misconstrued as a real hazard – even if only by consumers – it will create a perverse disincentive to avoid producing more data. Lastly, too much data can potentially result in an over-fitted model, giving an illusion of accuracy (which is in some respects worse than no knowledge at all). In other words, data is not knowledge, and as data grows in size and complexity, the task of transforming it into knowledge grows more difficult.

As an example of this, we show in Chapter 2 a practical application by demonstrating that a machine learning approach to skin sensitization benefits from pruning the data rather than using all available descriptors, and that combining chemoinformatic descriptors with *in vitro* assays outperforms a system based exclusively on descriptors of chemical structures. Lastly, we demonstrate that using domain-specific knowledge (in this case, the monotonic dose-response nature of skin sensitization) will improve the results. An obvious extension of this is that machine learning approaches to hazard prediction (and, *in vitro* tests generally) will improve in accuracy when they can be structured around existing knowledge of the mechanism of toxicity, either by taking advantage of an Adverse Outcome Pathways (AOP) – which specifies the toxic mechanisms at the organismal and population level - or, at the cellular level, a Pathway of Toxicity (PoT). A Pathway of Toxicity (discussed more extensively in Appendix III) represents "a molecular definition of the cellular processes shown to mediate adverse outcomes of toxicants".

Therefore, the second parallel development needed to make computational toxicology truly revolutionary for hazard assessment is an improved ability to efficiently extract Pathways of Toxicity from *in vitro* data – to take advantage of our newfound ability to survey and quantify subtle molecular changes at the cellular level.

Certainly, toxicology is not alone in this transformation, as a similar project is underway throughout biology as a whole, both because the limitations of the reductionist approach have become apparent and the advent of new technologies has allowed for a systems level view. Just as traffic congestion cannot be explained by the physics of automobiles or the combustion of gasoline but instead requires an understanding of both the macro level (the network of roads) and the micro-level (the cars and drivers), toxicological effects can rarely be explained in their entirety by the simple activation of one receptor or inhibition of an enzyme, but instead represent a disturbance of homeostasis within a complex system that must be appreciated at a systems-level.

Understanding such networks and pathways requires quantitative, systems-level measurements of the transcriptomic, proteomic, and metabolomics responses of a cell to a toxicological challenge. However, these "-omics" approaches come with some significant data analysis challenges.

All such –omic approaches can be noisy, and the large quantity of highly variable data creates a dilemma for data analysis – too stringent a statistical test, and one gets a handful of up- or down- regulated genes or a few obvious metabolites; too lenient, and you run the risk that of being misled by false-positives (Shi et al., 2008). Therefore, -omics approaches that depend exclusively on inferential statistics for data analysis are likely missing the very systems-level insights they promise to offer.

One way out of this dilemma is to use a pathway based approach, which both minimizes the need to correct for multiple hypothesis testing and is more robust to biological variability. Pathway based approaches can be either supervised, and depend on know annotations – e.g. using DAVID to look for enriched GO terms in a

list of genes (Dennis Jr et al., 2003), or they can be unsupervised, and attempt to reconstruct pathways or networks *de novo* based on the data, as is done with correlation-based networks (Quackenbush, 2003). The former has the disadvantage of restricting data analysis to confirm existing, known pathways. The latter can be a powerful approach to discovering novel connections, but is highly prone to spurious results and requires other data to validate any hypothesis generated. Furthermore, the networks produced from high-throughput data - often derisively referred to as "hairball diagrams" - essentially tell only a limited story of vague, putative gene interactions. In other words, it may provide an integrated view at the *genomic* level, but it provides few instead into the logic of the genetic circuitry and a somewhat limited knowledge of what is happening at the *dynamic* level.

In Chapter 3, we demonstrate the advantage of using a correlation and graph-theoretical approach for deriving a putative Pathway of Toxicity *de novo* from transcriptomic data, based on a small study of MPTP toxicity in mice. We also show that any analysis limited to known annotations may miss much that is of interest to a toxicologist; toxicological processes often a combination of physiological responses that are repurposed from inflammatory or developmental pathways, and dependence on the canonical pathways may be misleading or incomplete. Furthermore, in order to provide a model that offers more insight into the dynamics of the toxic process, we combined the transcriptomic network with an analysis of transcription factor binding sites and ChiP data to produce a rudimentary Genetic Regulatory Network (GRN).

While Chapter 3 shows both the potential insights that –omics technology can provide, as well as the disadvantages of depending exclusively on annotations for a data analysis, Chapter 4 details the problems that arise in a relatively new -omics technology - metabolomics- when the noise simply overwhelms the signal, and when annotations (and the database infrastructure that supports them) are largely inadequate for a pathway-level data analysis. Metabolomics has many of the same problems as transcriptomics, while at the same time it has both analytical and computational challenges that are unique. Some of the cautionary message of this

chapter is likely applicable to other –omics technologies, such as phosphoproteomics that are similarly immature. All –omics technologies typically come with high expectations and even a certain degree of hype – it is important to be realistic about the limitations to avoid being led astray by artifacts, but also to avoid a backlash when such technologies are inevitably shown to have pitfalls.

There are, therefore, many obstacles that remain before a comprehensive map of the Human Toxome can be said to be complete and a PoT-based toxicology can be realized. Some of them are technological, some of are computational, and some require the decidedly unglamorous but necessary work of creating databases, annotations and ontologies, all of which are necessary for the data to be more than the sum of its parts. Nonetheless, however complex the task may be, it is certainly not impossible. The number of cellular targets and metabolic pathways is finite, and thus the number of PoT should be, too.

Our understanding of the Human Toxome is, in some respects, much like cartography before the development of satellites—islands of well-described territory alongside vast oceans about which little is known; it could be said that even the extent of the unmapped territory is unknown. But the *terra incognita* should not frighten us; instead, it should beckon us towards it.

CHAPTER II – Probabilistic Hazard Assessment for Skin Sensitization Potency using Machine Learning to Design Integrated Testing Strategies

This project was performed with Tom Luechtefeld and Dr. Vanessa sa'Rocha. The author was responsible for the chemoinformatics, chemical similarity graph, developing and refining different approaches for machine learning, and writing the final draft. Tom Luechtefeld was responsible for developing and describing the machine learning approach and all coding and implementation of the algorithms. Dr. Vanessa sa'Rocha was responsible for acquiring and organizing the data as well as writing the final draft.

Abstract

Integrated Testing Strategies (ITS) aim to combine various information streams to hazard prediction. They are fueled by the increasing understanding of Adverse Outcome Pathways (AOP), i.e. mechanistic understanding and the development of tests reflecting these mechanisms. However, simple addition of further information bears the danger of adding noise and over-fitting. The problem is further amplified when potency information (dose/response) of hazard shall be estimated by these ITS.

Skin sensitization currently serves as the foster child for AOP and ITS development as legislative pressures combined with a very good mechanistic understanding of contact dermatitis, have led to test development and relatively large high-quality datasets. We curated such a dataset and combined a recursive variable selection algorithm to evaluate the information available through *in silico*, *in chemico*, and *in vitro* assays. Chemical similarity alone could not cluster chemical's sensitizing potency, and *in vitro* assays consistently ranked high in recursive feature elimination approaches. This allows for a reduction in the number of tests included in an ITS. Next we performed analysis with a Hidden Markov model that takes advantage of an intrinsic inter-relationship amongst the LLNA classes—that is, the

monotonous connection between LLNA and dose. The Dose-informed Random Forest/Hidden Markov Model was superior to the Dose-naive Random Forest model on all data sets. Although from the standpoint of balanced accuracy the improvement may seem small, this obscures the actual improvement in mis-classifications as the dose-informed Hidden Markov model had fewer "false-negatives" (i.e. extreme sensitizers as non-sensitizer) on all data sets.

Abbreviations: LLNA (Local Lymph Node Assay), HMM (Hiddem Markov Model), ITS (Integrated Testing Stretgy), AOP (Adverse Outcome Pathway)

1. Introduction

Skin sensitization, which clinically manifests in humans as allergic contact dermatitis (ACD), is an increasingly common concern among both regulators and the general population. Epidemiologic data indicate that an estimated 15-20% of the general population suffers from contact allergy (Thyssen, Johansen, & Menne, 2007). Most common are allergies to nickel, preservatives and fragrances (Peiser et al., 2012). In the particular case of fragrance allergy, prevalence estimates are ranging from 1.0-4.2% (Thyssen, Linneberg, Menne, & Johansen, 2007). Occupational contact dermatitis is particularly prevalent in the personal services industry, with an estimated prevalence of 1.2 percent in the beauty/haircare industry (Warshaw et al., 2012), as well as the petrochemical, rubber, plastic, metal and automotive industries (McDonald, Beck, Chen, & Cherry, 2006). For several decades, animal testing has been used as predictive tool to identify and characterize skin sensitizers, with the guinea pig as the initial animal of choice, which over the last 15 years has increasingly been replaced by the mouse local lymph node assay (LLNA), which has also been validated as a stand-alone (OECD). The assay uses slightly fewer animals (16 instead of 20), reduces time and suffering as it stops at the stage of lymph node swelling, and is thus considered a refinement alternative, and also provides a sensitization potency estimate, in contrast to the guinea pig assay. However, during the last few decades, there has been a growing concern about using animals for

product development and regulatory testing, especially for cosmetic products and ingredients. The drive for this change resulted first in the implementation in Europe of Cosmetic Directive (76/768/EEC), now Cosmetics Regulation (European Union, 2009), which stipulates a progressive phasing out of animal tests for the purpose of assessing the safety of cosmetics and their ingredients, and ultimately, a complete testing ban, enforced with a marketing ban with deadline in 2013. The European chemicals legislation on the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH; Regulation EU No 1907/2006) requires that animal testing for hazard assessment should be conducted only as a last resort and authorize the usage of validated *in vitro* methods. In 2007 the US National Academy of Sciences released a report called "Toxicity Testing in the 21st Century: A Vision and a Strategy" outlining a strategy for toxicity testing that would be based on human rather than animal biology and suggests moving regulatory toxicology to a more mechanistic approach requiring substantially fewer or no animals (National Research Council, 2007). Furthermore, as knowledge of the molecular key steps of skin sensitization becomes more detailed, this presents both an opportunity and a challenge to improve the availability of alternative methods.

Newer alternative methods developed for skin sensitization are based on the specific, key mechanistic steps: the chemical's ability to penetrate the skin, its capacity to bind with proteins present in the skin, as well as the recognition of this protein complex by immune cells (Adler et al., 2011). The Direct Peptide Reactivity Assay (DPRA) is the first non-animal test method formally recommended by the European Centre for the Validation of Alternative Methods (ECVAM) for skin sensitization (European Commission Joint Research Centre, 2013)—and addresses the chemical's reactivity to proteins by measuring depletion of synthetic peptides containing either cysteine or lysine (Gerberick et al., 2004; Gerberick et al., 2007). The accuracy of the DPRA for distinguishing sensitizers from non-sensitizers was 82% (sensitivity of 76%, specificity of 92%), excluding metal compounds for which the test is not applicable (Gerberick et al., 2007). More recently, ECVAM also published a recommendation indicating the usefulness of the KeratinoSens™ assay

(European Commission Joint Research Centre, 2014). The assay addresses the activation of the Keap1-Nrf2 ARE pathway in human keratinocytes (HaCaT) which is considered a major regulator of cyto-protective responses to electrophile and oxidative stress by controlling the expression of detoxification, antioxidant and stress response enzymes and proteins (Emter, Ellis, & Natsch, 2010). The accuracy was 77% based on testing of about 145 chemicals with 79% of sensitivity and 72% of specificity (Natsch et al., 2013). According to ECVAM, none of these assays can be used as stand-alone method and data should be considered in combination with other information. A similar assay using the same cell system including a combination of glutathione (GSH) depletion and gene expression known to be activated by sensitizing agents (Keap 1/Nrf 2/ARE/EpRE, ARNT/AhR/XRE and Nrf1/MTF/MRE) shown an accuracy of 84%, with a sensitivity of 81% and specificity of 92% based of 102 chemicals (McKim, Keller, & Gorski, 2010). Other assays have shown promising results to test the induction of dendritic cell (DC), which includes cell line surrogates (THP-1 a human monocytic leukemia cell line) and U937 (a human histiocytic lymphoma cell line) with DC-like characteristics for phenotypic markers of activated DC (eg. CD86 and CD54) (Ashikaga et al., 2010; Sakaguchi et al., 2006). In addition, some commercially available *in silico* models such as TIMES (Dimitrov et al., 2005) and DEREK (Sanderson & Earnshaw, 1991) have been developed based on Structure Activity Relationships (SAR).

As skin sensitization is a complex endpoint that needs more than one alternative assay to replace animal test, the open question remains on how to integrate available information for predicting the skin sensitization hazard, and more specifically how to make the best use of the cumulative information in the most efficient way possible as well as guide for future testing in such a way that the information gain is maximized and accomplished with fewest possible tests (Jaworska, Harol, Kern, & Gerberick, 2011). Recently, the use of an Integrated Testing Strategy (ITS) of batteries of *in vitro* tests combined with *in silico* models have been proposed for the replacement of LLNA (Bauch et al., 2012; Hartung, Luechtefeld, Maertens, & Kleensang, 2013; Hirota et al., 2013; Jaworska, Dancik,

Kern, Gerberick, & Natsch, 2013; Jaworska et al., 2011; Maxwell et al., 2014; McKim et al., 2010; McKim, Keller, & Gorski, 2012; Nukada, Miyazawa, Kazutoshi, Sakaguchi, & Nishiyama, 2013).

ITS provides a more formal, systematic, and quantitative approach to risk estimation (as distinct from a Weight of Evidence approach) than a fixed battery of tests. As suggested earlier, an ITS is "*an algorithm to combine (different) test result(s) and, possibly, non-test information (existing data, in silico extrapolations from existing data or modeling) to give a combined test result. They often will have interim decision points at which further building blocks may be considered*" (Hartung et al., 2013).

Since the volume of data—*in silico*, *in chemico* and *in vitro*—to be considered increases at a rapid rate and is becoming more heterogeneous in nature, there is a keen need for a new ways to combine them that offers both a robust and powerful approach to estimate hazard and support a risk decision. Likely this has to be done in a probabilistic way, where the different input parameters are combined to generate an overall probability of hazard and risk. Furthermore, understanding the effects of test substances at different doses is an essential aspect of safety testing that is not being addressed by the current proposals. We believe that an ITS based on a machine learning approach offers the best possibility to combine data for the optimal estimate of hazard given the information available. To this end, we combined a variable selection algorithm to evaluate the information available through both *in silico*, *in chemico*, and *in vitro* assays with a Hidden Markov model that takes advantage of an intrinsic inter-relationship amongst the LLNA classes—that is, the connection between LLNA and dose.

2. Material and methods

2.1 Dataset

The data set included a total of 145 distinct chemicals with *in vitro* assays from (Jaworska et al., 2013), which included the chemicals of the LLNA dataset (Gerberick et al., 2005). In addition, we obtained a subset of the original chemicals from (McKim et al., 2010) and (Natsch, Emter, & Ellis, 2009) additional *in vitro* assays. The total number of descriptors: 7, 9 and 10 *in vitro/in chemico* for datasets 1, 2 and 3, respectively, and 1666 chemoinformatic molecular descriptors available from DRAGON software. For all distinct 145 chemicals LLNA classifications were available as reference classification. Simplified molecular input line entry system (SMILES) strings were obtained via Pubchem (Bolton, Wang, Thiessen, & Bryant, 2010). DRAGON features were calculated with VCLABS E-DRAGON software (Tetko et al., 2005; Todeschini, Consonni, & Todeschini, 2009)..

The initial data set (Data Set 1, 145 distinct chemicals) was based on the work of Jaworwska (2013), which included TIMES predictions (Dimitrov et al., 2005), combined with Dragon descriptors. Data Set 1 was subdivided into smaller data sets based on additional available *in vitro* results as follows: Data Set 2 included values for ARE EC 1.5, ARE Cmax (defined as the concentration that causes as 1.5 fold increase and maximal increase in the Antioxidant Response Element induction) and Imax (the maximum fold-induction achieved) for 84 chemicals from Natsch et al. (2009), and Data Set 3 included glutathione depletion from McKim et al. (2010) for a subset of 65 chemicals. Data sets are available in Supplement 1.

	Chemicals	Descriptors	Source
Data Set 1	145	TIMES, Dragon Descriptors, keratinoSens KEC 1.5 and KEC 3.0, Cytotoxicic_ IC50, DPRACys, DPRALys, CDFree, CD86	(Jaworska et al., 2013)
Data Set 2	84	Dragon Descriptors, keratinoSens KEC 1.5 and KEC 3.0, Cytotoxicic_ IC50, DPRACys, DPRALys, CDFree, CD86, ARE EC 1.5, Imax, ARE Cmax	(Jaworska et al., 2013) (Natsch et al., 2009)
Data Set 3	65	Subset of Data Set 1 with additional Glutathione depletion data available	(Jaworska et al., 2013) (McKim et al., 2010) (Natsch et al., 2009)

Chapter II, Table 1: Overview on dataset 1 to 3 as described the section dataset.

2.3 Chemical Similarity generation

A chemical similarity map was generated by the ChemViz plug-in and Cytoscape 2.8.3 (http://www.cgl.ucsf.edu/cytoscape/chemViz/). Tanimoto distances were calculated based on SMILES strings using the Klekota and Roth fingerprint algorithm (Klekota & Roth, 2008), and any chemical with a Tanimoto similarity of greater than 0.70 was considered as link.

2.4 Random Forest

We used the scikit-learn Random Forest (Pedregosa et al., 2011) version 0.14 implementation in these analyses. Random Forest is an ensemble supervised learning model. Briefly, a Random Forest model (Breiman, 2003) is trained on a subset of all the data; during training we construct 100 random trees. Each tree is constructed via recursively splitting training data using a random selection of the available features with each permitted up to log_2 of the available features and splitting continued until the split data contains only one chemical (tree split criterion: entropy; min-samples-leaf: 1).

During each chemical prediction the class is passed down each random tree (using the feature values for that test chemical). Each tree reports the class of the chemical in the leaf node most closely matching that of the test chemical. The Random Forest then makes a prediction by picking the class most voted for (so called ensemble method).

2.5 Recursive Feature Elimination

Recursive feature elimination involves first evaluating feature importance and then eliminating low importance features. Feature importance was calculated with the scikit-learns implementation of the Breiman Random Forest Variable Importance algorithm (Breiman, 2003). This algorithm evaluates a given feature's importance in a trained model by randomly permuting all available values and recording the subsequent loss in model accuracy and the permutation that results in the greatest loss is given greater accuracy. Variable importance was normalized by dividing each feature importance value by that of the maximally important feature, which was thereby assigned a value of 1.

2.6 Dose transformation

To encode the data using LLNA classes we transformed the LLNA classification into a binary classification as follows:

Class	Low Dose	Medium Dose	High Dose
Non/weak-sensitizer	Nontoxic	Nontoxic	Nontoxic
Moderate-sensitizer	Nontoxic	Nontoxic	**Toxic**
Strong-sensitizer	Nontoxic	**Toxic**	**Toxic**

This transformation allows us to train a dose informed Random Forest that can classify chemicals combined with categories as toxic or non-toxic. Thus for a given chemical our new model can make 3 predictions:

Chemical

1-bromobutane	LLNA reference classification: non-sensitizer

Transformed LLNA classification

LLNA Low Dose	LLNA Medium Dose	LLNA High Dose
Nontoxic	Nontoxic	Nontoxic

This transformation allows us to use the predictions made by the Random Forest to build a Hidden Markov model. It should be noted that a supervised model trained with this dose transformation may very well predict/classify a chemical as follows:

Prediction Series of the Hidden Markov model

LLNA Low Dose	LLNA Medium Dose	LLNA High Dose
Nontoxic	Toxic	Nontoxic

This prediction series is concerning because our prior knowledge tells us that if a chemical is toxic at low dose it will remain toxic at higher doses; in order to avoid this we constrained the model so that a chemical that was predicted as toxic at a low dose would remain automatically be considered toxic at higher doses (see as well section 2.7 Hidden Markov Model transition probabilities).

2.7 Hidden Markov Model

A Hidden Markov model allows us to enforce proper prediction series by encoding our knowledge of allowable toxicity transformations. For example, a chemical that is toxic at low does cannot become non-toxic at higher dose. Namely that a chemical that is toxic at low dose will be as well toxic at higher doses, and that a chemical that is non-toxic at high dose must be non-toxic at lower doses.
A Hidden Markov model contains several important properties:

- **Hidden States**: These are states that cannot be directly observed. In our case a given chemical contains 6 hidden states, one for toxic or non-toxic at each of the three dose categories.
- **Transition Probabilities:** Transition probabilities tell us the probability for transitioning from one hidden state to another. Transition probabilities allow us to encode our prior knowledge about toxicity changes. By disallowing a transformation from the hidden state corresponding to low dose-toxic to the hidden state corresponding to moderate dose-non-toxic we can ensure that no prediction sequences will contain this transition.
- Empirically speaking, transition probabilities can be obtained from the data by counting how often a chemical transition from one hidden state to another takes place. Thus no special treatment is needed to encode our prior knowledge about chemical transformations since, for instance, the chemical data will contain no instances where a chemical transitions from toxic at low dose to non-toxic at higher dose occurs.
- **Emission Probabilities:** In our case, emission probabilities inform about the probability that a given hidden state will emit the prediction given by our dose-informed supervised model. This emission probability can be obtained empirically by counting how often a given prediction aligns with the given hidden state divided by the number of predictions.

The Hidden Markov Model was built using the scikit-learn HMM-module (Pedregosa et al., 2011). Transmission probabilities were built by enumeration from data and

emission probabilities by counting classifier outputs matched with actual toxicity class (10 iterations, 0.01 threshold). The trained Markov model chemical predictions were obtained using the Viterbi algorithm (Viterbi, 1967) scikit-learn implementation. For an introduction to HMM please see e.g. (Baum & Petrie, 1966).

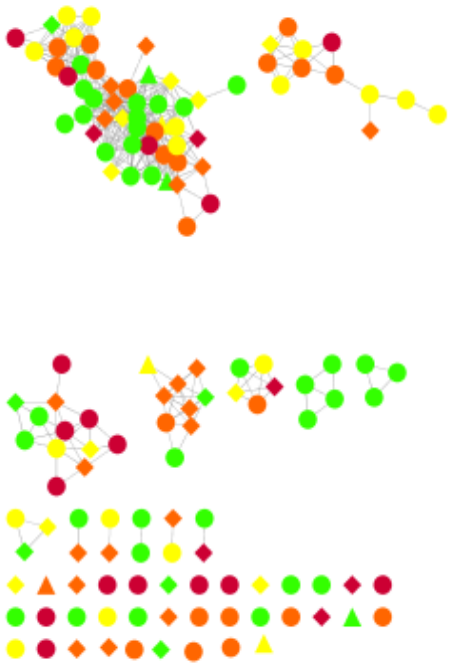
2.8 Cross Validation

In order to ensure a training data set that closely resembled the testing data set, we used 100 iterations of train/test set splits created via scikit-learn's stratified shuffle split-cross validation approach (Pedregosa et al., 2011). In testing both the dose informed and dose-naive approaches to skin-sensitization classification we allowed training on 90% of the available data and testing on the remaining unseen 10% of the data separately for each dataset, avoiding peeking by insuring that no model was trained on data it would later be tested on except unavoidably in the case of the comparison of Dataset 1 with and without the TIMES as the TIMES was trained on a number of chemicals that are included in the dataset.

3. Results

3.1 Chemical diversity of dataset

The chemical similarity map (Figure 1) indicates that many of the chemicals were highly similar compounds, but that clusters of similar chemicals did not necessarily share LLNA status—skin sensitization is therefore difficult to predict using chemicals descriptors alone. Furthermore, the data set included several chemicals (35) that were chemically dissimilar—meaning they had a Tanimoto similarity of less than .70—from all other chemicals in the data set. Interestingly, the largest cluster of similar chemicals contained several instances of chemicals from all four LLNA classes (non-weak, moderate, strong, extreme), but had only one chemical with a class error greater than 1, indicating that the model performed well in differentiating LLNA class amongst structurally similar chemicals.



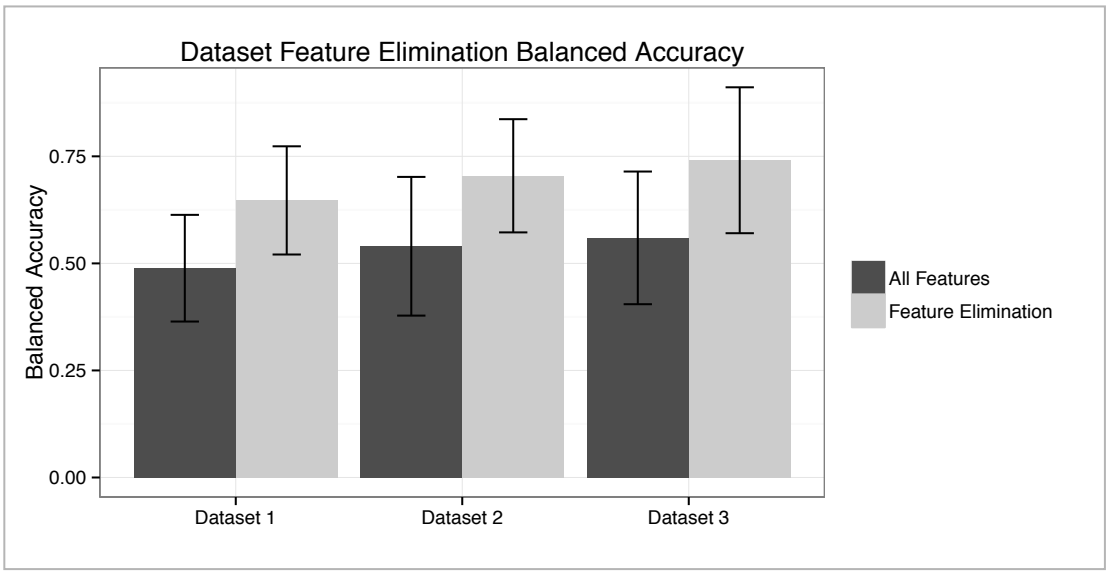
Chapter II, Figure 1: Chemical similarity map: Chemicals are colored according to LLNA status.
Figure 1, Legend: (Red = Extreme sensitizer; Orange = Strong, Yellow = Moderate, Green = Weak/Non). Chemicals with a Tanimoto index > .70 are linked; distance is proportional to Tanimoto similarity. Although there was a large cluster of highly similar chemicals, 35 chemicals had no similarity to any other chemical in the data set and an additional 8 had only one similar chemical. The difference between predicted and actual class are denoted by shape: No difference between predicted and actual are indicated by circles, one class difference by squares, and two class difference by triangles.

3.2 Feature Selection and Variable Importance

As skin sensitization is difficult to predict from chemoinformatics methods/QSARs alone, it is therefore desirable to combine *in silico* data with *in vitro* and *in chemico* assays. Feature selection methods typically improve predictive models by avoiding the over-fitting that comes from using statistically independent features in the prediction model generation (training phase), shortens computational time, and makes the model easier to understand. Recursive feature

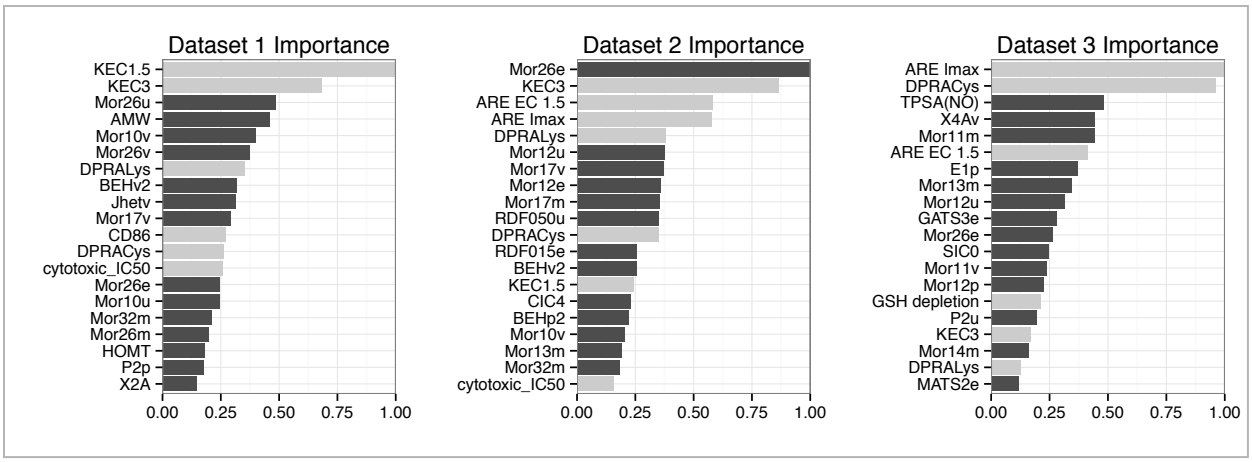
elimination can be used to trim a dataset with a large number of features—in essence, a Random Forest is trained on the dataset and the resulting features are ranked according to the Breiman feature importance test (Breiman, 2003). After ranking the dataset it is modified by the removal of the least valuable feature. The process is then repeated until the number of features in the dataset is reduced to the 20 most informative variables that were subsequently selected for building the prediction model.

Recursive feature selection indicated that the available *in vitro* tests were providing substantial information compared to the chemical descriptors alone as they were consistently ranked within the top 20 descriptors (see Figure 2). As data accumulates, recursive feature elimination will likely allow for a more informed ranking of *in vitro* assays and a better choice in terms of what test to perform next when presented with a chemical with limited available *in vitro* data, or in cases where a QSAR has predicted the potential for skin sensitization either on the basis of skin permeability or electrophilicity.

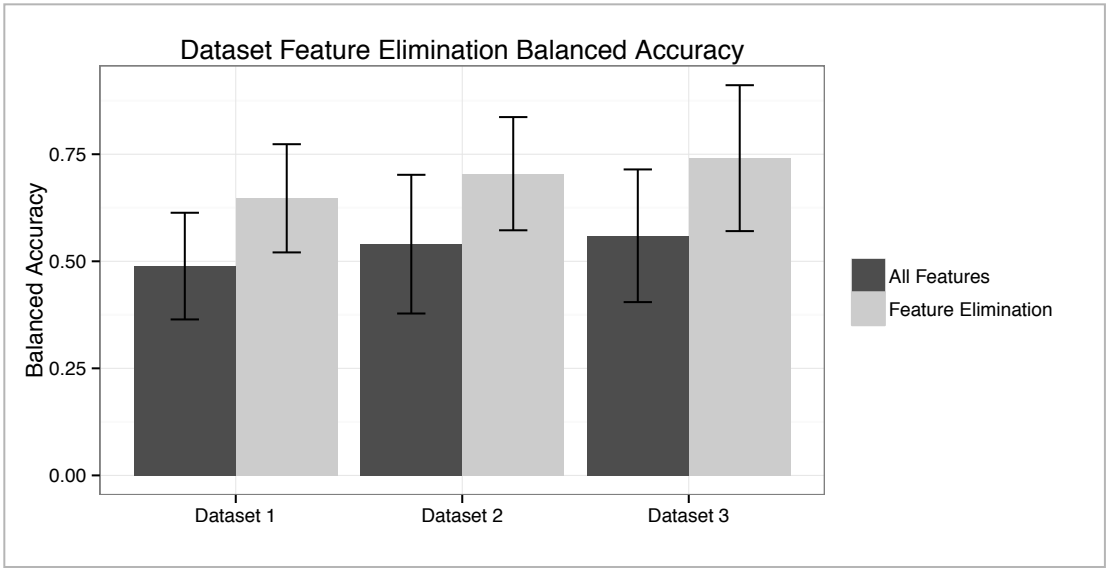


Chapter III, Figure 3: Balanced accuracy (four class problem) of data with feature selection (20 most informative features) and without.
Legend, Figure 3: Feature selection consistently improved the balanced accuracy compared to a non-balanced accuracy. Error bars indicate standard deviation of balanced accuracy estimates calculated from cross validation.

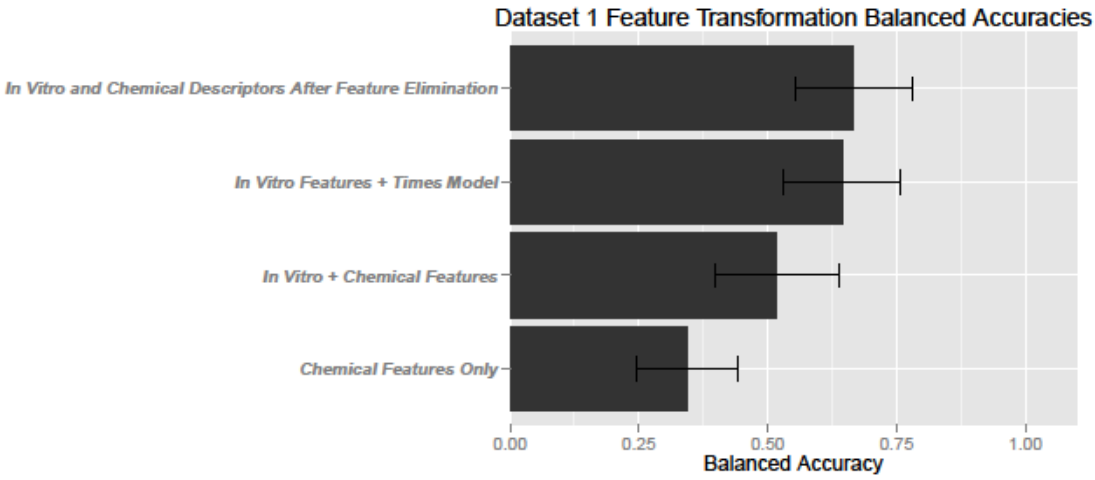
Figure 3 shows that that the balanced accuracy of data with feature (20 most informative features) and without feature selection consistently improved the balanced accuracy compared to a non-balanced accuracy for all three datasets. Chemical descriptors alone showed very poor overall accuracy for prediction. Combining *in vitro* assays with the chemical descriptors selected by the recursive feature elimination algorithm performed seemingly as well as the *in vitro* models with TIMES (see Figure 4) -but without the same restriction on applicability domain as TIMES. Furthermore, TIMES performance is likely overstated in this case by "peeking"—that is, this data set includes chemicals that were part of the TIMES training set.



Chapter III, Figure 2: Variable importance: 20 most informative features were selected by recursive feature elimination algorithm.
Legend, Figure 2: *In vitro/in chemico* assays are shown in gray and DRAGON descriptors are shown in black. *In vitro* assays consistently ranked amongst the top features selected. For more details on the DRAGON descriptors, see Supplement 1.



Chapter III, Figure 3: Balanced accuracy (four class problem) of data with feature selection (20 most informative features) and without.
Legend, Figure 3: Feature selection consistently improved the balanced accuracy compared to a non-balanced accuracy. Error bars indicate standard deviation of balanced accuracy estimates calculated from cross validation.



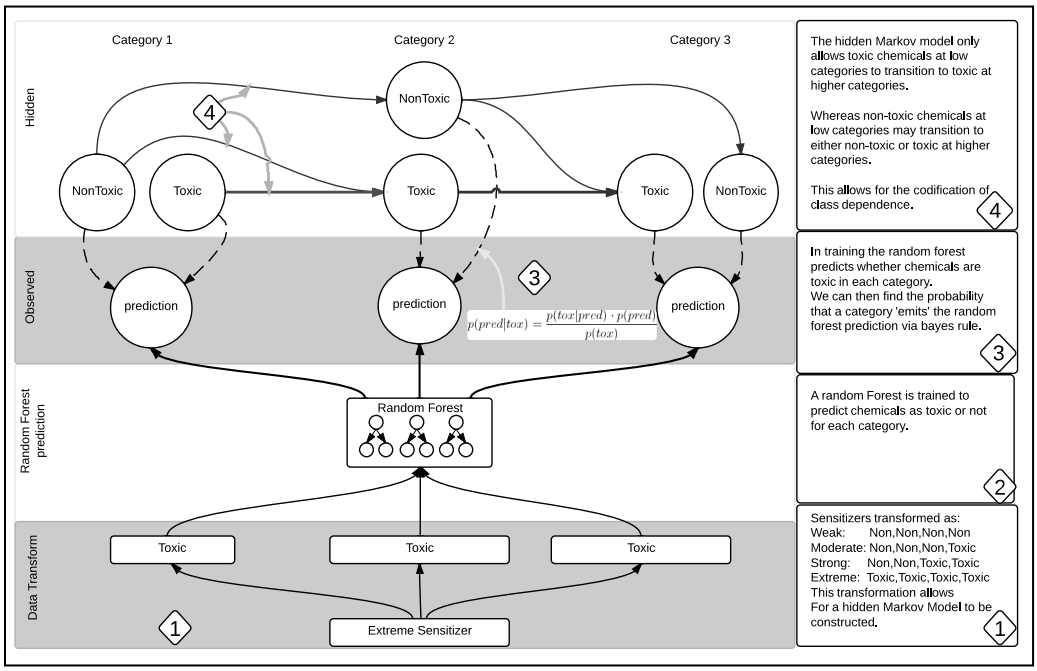
Chapter III, Figure 4: Balanced accuracies for different feature subsets of dataset 1.

3.3 Hidden Markov Model Generation and Validation

Hidden Markov Models (HMMs) are a formal methods for making probabilistic models of labeling problems; a Markov system typically has *N* discrete

states and *T* discrete time-steps; in this case, however, instead of *time* the Markov chain is based on *dose*. This required transforming our data from pairs of chemicals/LLNA class into chemical-dose pairs. In other words, each chemical was classified as toxic/non-toxic at a low dose, medium dose, or high dose (see figure 5) corresponding to LLNA which meant that the model in essence predicted a binary question—that is, whether the chemical was toxic or non-toxic at a given dose increment—instead of trying to predict a four-class problem. In principle, a model that uses this extra information—a "a dose-informed" Hidden Markov/Random Forest approach—should perform better than a "dose-naïve" Random Forest. Emission probabilities of a Hidden Markov model help us to encode and exploit this variable supervised model performance.

Figure 5 shows how HMM has been implemented to build the dose informed Hidden Markov/Random Forest approach.



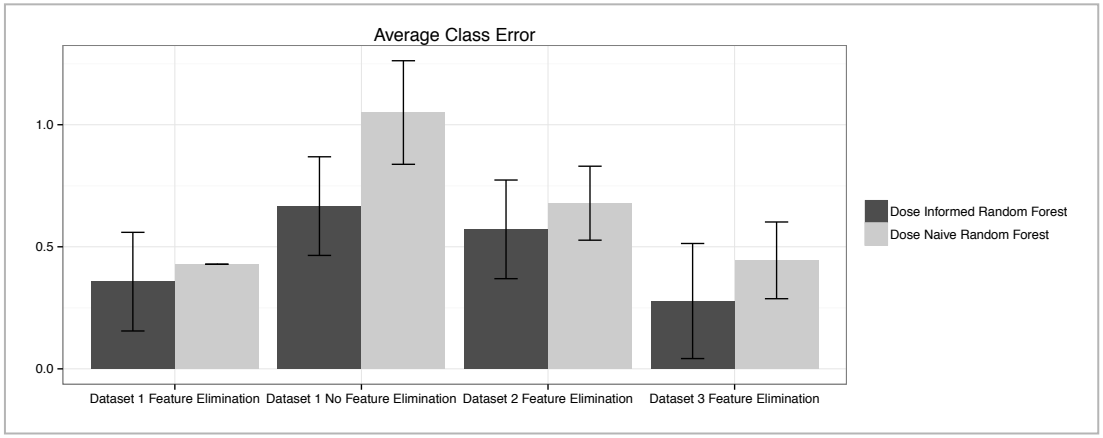
Chapter III, Figure 5: Visual description of dose transformation use in a Hidden Markov Model.

3.4 Average Class Error

With the addition of the DRAGON descriptors, the total number of features available for data sets was quite large. In this case, a comparison of both dose-informed and dose-naïve models by average class error using all available chemical descriptors from DRAGON and all *in vitro* assays performed seemingly worse than using only the 20 most informative features selected by recursive feature selection (Figure 6). Significantly, average class error for Dose-informed Random Forest/Hidden Markov Models was lower than average class error in the Dose-naive Random Forest models on all data sets. Although from the standpoint of balanced accuracy the improvement is not apparent, this obscures the actual improvement in mis-classifications.

The best performing dose-informed models (Data Set 1 and 2) had no misclassifications greater than 2 classes, i.e. no extreme sensitizers classified as non-sensitizers and no non-sensitizers classified as extreme sensitizers, indicating overall a very small rate of extreme false-negatives (an extreme sensitizer classified as having no sensitization) and no extreme false-positives (non-sensitizers classified as extreme sensitizers) in any data-set (Table 2) Interestingly, the extreme sensitizer misclassified (phthalic anhydride [CAS 85-44-9] in Data Set 3) hydrolyzes in water at pH 6.8-7.24 with half-lives of 0.5-1 min at 25 °C, forming phthalic acid and is therefore not within the applicability domain of *in vitro* assays (OECD SIDS Initial Assessment Report 2005). Phthalic acid [CAS 88-99-3] is classified on a non-sensitizer by a modification of the Maguire method and the LLNA (ECHA database on registered substances, searched on 25.07.2014), which explains the classification as non-sensitizer by our approach.

Looking at it another way, using the dose-informed Hidden Markov model had 95.8%, 92.6% and 92.1% accuracy predicting the LLNA class +/- 1 one class, versus just 90.4%, 88.6% and 90.6% balanced accuracy for the dose-naive model for dataset 1, 2 and 3, respectively.



Chapter II, Figure 6: Average Class Error and standard deviation from cross validation:

Legend, Figure 6: For all comparisons the dose-informed model gave smaller average class errors compared to the dose-naive model. Furthermore, feature selection improved the results—using all chemical descriptors significantly worsened the performance of the Random Forest.

	Dose Informed						Dose Naïve					
	LLNA Reference classification						LLNA Reference classification					
Predicted	Non	Moderate	Strong	Extreme	Sum of Predictions		Non	Moderate	Strong	Extreme	Sum of Predictions	
Data Set I												
Non	35	5	3	0	43	Balanced	36	6	5	3	50	Balanced
Moderate	6	18	5	2	31	Accuracy	2	18	5	1	26	Accuracy
Strong	1	10	25	10	46	0.65	4	8	22	9	43	0.63
Extreme	0	0	7	18	25		0	1	8	17	26	
Occurrences	42	33	40	30	145	Balanced	42	33	40	30	145	Balanced
Accuracy	0.83	0.55	0.63	0.60		Error	0.86	0.55	0.55	0.57		Error
Distance Weighted Error	0.19	0.45	0.45	0.47		0.39	0.24	0.48	0.58	0.67		0.49
Data Set II												
Non	13	3	4	0	20	Balanced	14	1	5	2	22	Balanced
Moderate	5	10	2	1	18	Accuracy	5	11	2	2	20	Accuracy
Strong	2	4	19	6	31	0.62	1	5	20	5	31	0.63
Extreme	0	0	4	10	14		0	0	2	8	10	
Occurrences	20	17	29	17	83	Balanced	20	17	29	17	83	Balanced
Accuracy	0.65	0.59	0.66	0.59		Error	0.70	0.65	0.69	0.47		Error
Distance Weighted Error	0.45	0.41	0.48	0.47		0.45	0.35	0.35	0.48	0.88		0.52
Data Set III												
Non	11	1	2	1	15	Balanced	12	1	2	1	16	Balanced
Moderate	4	9	3	1	17	Accuracy	1	9	2	1	13	Accuracy
Strong	0	2	15	4	21	0.65	2	3	16	5	26	0.66
Extreme	0	1	3	7	11		0	0	3	6	9	
Occurrences	15	13	23	13	64	Balanced	15	13	23	13	64	Balanced
Accuracy	0.73	0.69	0.65	0.54		Error	0.80	0.69	0.70	0.46		Error
Distance Weighted Error	0.27	0.38	0.43	0.69		0.44	0.33	0.31	0.39	0.77		0.45

Chapter III, Table 2: Confusion matrix of predicted chemical's sensitizing potency vs. LLNA Reference classification for datasets 1 to 3 including balanced accuracy and balanced error.

4. Discussion

Although toxicology has a handful of *in vitro* test batteries that are well-established (e.g. mutagenicity), such approaches have not kept up to date with the ability to produce high-throughput *in vitro* datasets. As *in vitro* assays grow in importance and availability, a more objective way of evaluating them becomes necessary, otherwise every positive result is a liability for the risk assessment of a given substance, and eventually this leads to an accumulation of false-positives and ultimately, inaccurate risk assessment and a lack of faith in *in vitro* method.

While skin sensitization provides a strong domain for the use of modern machine learning techniques, it also presents some challenges: since the models will be applied for regulatory purposes, we need hazard estimation models that are easily understood and visualized, which precludes black-box approaches such as Bayesian networks. However, the existing datasets have several traits that make more straightforward approaches, such as decision trees, impractical. To begin with, the datasets typically has more descriptors than samples and requires combining datasets. This means employing a methodology that is robust to both missing information, as well as highly correlated data since each dataset will likely contain redundant or overlapping data – for example, DRAGON chemical descriptors which attempt to calculate electrophilicity will likely show a high level of correlation with ARE (Antioxidant Response Element) induction. Here, we show that *in vitro* tests contribute substantial predictive information compared to the chemical descriptors alone. Furthermore, we show that given the expansion of chemical descriptors, *in vitro* tests, *in chemico* tests, etc. machine learning techniques likely require pruning the information used in a model—something that will become even more important as Toxcast and other high throughput data become available; at some point, additional data is merely adding noise or causing model over-fitting. It is always a temptation to assume that using all available data will improve accuracy; however, the reality is that more descriptors may simply be adding more noise and not offering additional information. In toxicology, we have the prominent example of the accumulation of false-positives that have made the battery of tests for mutagenicity

cumbersome (Kirkland, Aardema, Henderson, & Muller, 2005; D. J. Kirkland et al., 2005).

Furthermore, it has become increasingly evident that characterizing the dose-response relationship in *in vitro* assays is key to using them effectively in machine learning techniques. Typically, predicting LLNA is a four-class problem (predicting non-sensitizer-weak, moderate, strong and extreme), which presents a significant challenge to most machine learning techniques. While some approaches try to solve this problem by predicting sensitizer vs non-sensitizer only, this model seeks to exploit the fact that LLNA follows a monotonic dose-response curve: that is, if a chemical is a sensitizer at a low dose, it will also be a sensitizer at a high dose. By redefining the problem as predicting whether a chemical is a skin sensitizer at a given dose-increment, the prediction becomes a binary problem. From a theoretical perspective, it is clear that a Hidden Markov Model will lessen extreme mis-classifications; this is borne out in our datasets by the fairly small average class distance between predicted vs. actual for the Dose-Informed vs. Dose-Naïve. From a practical stand point, this can give users of the model some confidence that while the actual predicted class may not be accurate, a predicted non-sensitizer is unlikely to be an extreme sensitizer and vice versa. Our dose-informed Hidden Markov Model generally outperformed the dose-naïve 4-class Random Forest prediction models and minimized miss-classifications of a more than two-class distance. Furthermore, a dose-informed Hidden Markov Model could potentially be extended when used with attributes that show a dose-response curve as opposed to the single-value assays used here. This approach can likely be extended with the increasing availability of descriptors such as those from Toxcast, which capture the dose-response curve of the mechanistic steps involved in an adverse outcome.

5. Acknowledgements

The authors would like to thank Dr. Joanna Jaworska for helpful discussion and generosity with her data. Dr. James M. McKim from Iontox (Kalamazoo, USA); Dr.

Andreas Natsch from Givaudan (Dübendorf, Switzerland) for the fruitful discussions and their availability in sharing unpublished experimental data.

CHAPTER III – MPTP's Pathway of Toxicity Indicates Central Role of Transcription Factor SP1

Abstract: Deriving a Pathway of Toxicity from transcriptomic data remains a challenging task. In this paper, we explore the use of weighted gene correction network analysis (WGCNA) to extract an initial network from a small microarray study of MPTP toxicity. The resulting network was analyzed for transcription factor candidates, which were narrowed down via text-mining for relevance to the disease model, and then combined with the FANTOM4 database to generate a Genetic Regulatory Network. This analysis demonstrated that a small microarray study can capture much of the known biology of MPTP toxicity and suggests several candidates for further study. Furthermore, the analysis strongly suggests that SP1 plays a central role in co-ordinating the cellular response to MPTP toxicity.

Abbreviations: MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine), PD (Parkinson's Disease), MT (microtubules), ETC (electronic transport chain), WGCNA (Weighted Gene Correlation Network), GRN (Genetic Regulatory Network), POT (Pathway of Toxicity), TOM (Topological Overlap Metric),

1. Introduction

In order to bring toxicology into the 21st century, toxicology is undergoing a profound paradigm change: away from animal-based black-box models towards a systems toxicology approach based on higher throughput testing. The necessary mapping of the pathways of toxicity often involves using high dimensional data sets, which are traditionally analyzed by looking for a few differentially expressed genes.

However, cellular pathways leading to toxicity may involve subtle perturbations in many genes rather than drastic alterations in a few. In addition, microarrays can often be noisy and can show poor reproducibility, which only increases the difficulty of extracting meaningful, systems-level insights into biology from the data.

Here, we used an approach that derives a *de novo* network from a small data set by using a weighted approach, clusters by network topology, and uses the resulting clusters for further analysis with text-mining and other sources of high throughput data (ChIP experiments and siRNA perturbation studies), ultimately producing a more specific genetic regulatory network (GRN). Using a WGCNA approach offers, in essence, a dimensionality reduction technique that can be used to produce a more detailed genetic regulatory network based on known and predicted transcription factor interactions, and brings us a small step closer to a wiring diagram of the cell.

MPTP (methyl-4-phenyl-1,2,3,6-tetrahydropyridine) toxicity offers an excellent "proof-of-concept" for the ability to derive a Pathway of Toxicity from high-throughput data, since the broad outlines of the Pathway of Toxicity are understood. It is used widely as an animal model for a relatively data-rich disease (Parkinson's disease) (Schober, 2004), since MPTP poisoning, like Parkinson's, is highly selective for dopaminergic neurons in the *substantia nigra* and the clinical symptoms are highly similar to Parkinson's (Snyder & D'Amato, 1986).

MPTP is not itself toxic, but owing to its high lipophilicity it is able to cross the blood brain barrier, where it is metabolized in astrocytes by Monoamine Oxidase B (MOA-B) to MPP+. MPP+ is then transported selectively by the dopamine

transporter into neurons. Once inside the neuron, it is thought to exert its primary action through targeting Complex I in the mitochondria, which results in disruption of the electron transport chain (ETC). While MPTP disruption of the ETC causes a loss of ATP, it is not a critical failure of Complex I and oxidative phosphorylation that causes pathology, as MPTP typically only causes a mild decrease in ATP levels and falls short of levels required to cause significant energy depletion (Perier & Vila, 2012) and deficiency in a component of Complex I does not lead to selective dopaminergic neural death (Sterky et al., 2012). Therefore, MPTP neurodegeneration is not necessarily caused by energy depletion. More likely, a shift in energy balance is a contributing factor (Krug et al., 2014).

Another consequence of the ETC disruption is increased ROS generated by impaired mitochondria. This may in turn cause oxidative damage to Complex I, initiating a spiral of decreased mitochondrial efficiency and increased ROS. ROS can cause peroxidation of the lipids, which disrupts the normal binding of cytochrome *c* to the mitochondrial membrane and facilitates the pro-apoptotic release of cytochrome *c* to the cytosol (Perier & Vila, 2012). Mitochondria-derived ROS have also been shown to damage lysosomal membranes in MPTP-intoxicated mice, leading to an impairment of lysosomal function and defective autophagic activity (Dehay et al., 2010), including mitochondrial autophagy (Ivatt & Whitworth, 2014). In addition to proteins and lipids, MPTP-intoxicated mice also exhibit oxidative damage to nuclear and mitochondrial DNA (Hoang et al., 2009). Despite the centrality of the intracellular, mitochondrial-generated ROS, there may be other

contributors to ROS in the context of PD/MPTP toxicity —for example, astrocytes or microglia.

Another key component of MPTP toxicity is microtubule (MT) disruption. MPP+ is believed to lead to hyperphosphorylation of Microtubule Associated Protein Tau (MAPT), which leads to microtubule instability (Cappelletti, Pedrotti, Maggioni, & Maci, 2001). Depolymerization of MTs is one suggested reason for the selective vulnerability of dopaminergic (DA) neurons by toxins such as MPTP, paraquat and rotenone, as dopaminergic neurons require axonal transport of neurotransmitters to the striatum for dopamine release (Ren, Liu, Jiang, Jiang, & Feng, 2005). The traffic along the axonal length of DA neurons requires intricate coordination between MTs and the motor proteins to ensure dopamine is transported successfully through vesicle transport. Depolymerization—or, less acutely, an impairment of coordinated traffic—can lead to an impairment of neural function. Furthermore, in neurons, mitochondria are actively transported throughout the cell body; the combination of impaired mitochondrial activity and impaired transport is likely key to the toxic outcome (Sterky et al., 2012).

The final step of MPTP toxicity, apoptosis, is likely the result of several pathways that combine to produce cell death. Apoptosis is thought to be generated through a mitochondrial-initiated, BAX-dependent process. Complex I inhibition does not directly trigger mitochondrial cytochrome *c* release but instead increases the "releasable" pool of cytochrome *c* in the mitochondrial membrane—increasing the magnitude of the signal that can be released when activated by BAX (Perier & Vila, 2012).

In summary, while MPTP toxicity has an agreed-upon origin (mitochondrial disruption), there is still much to be learned about the exact Pathway of Toxicity, and the toxicity mechanism likely involves alterations of several pathways along key points (Krug et al., 2014).

2. Materials and Methods

2.1: Data: Dataset GDS2053, which represented a small study of 12 samples based on the Affymetrix Murine Genome U74A Array (normalized via RMA and RAS 5) from MPTP-treated mice, was downloaded from GEO with GEOQuery (Davis & Meltzer, 2007) and checked for outliers via the IAC function in WGCNA (Langfelder & Horvath, 2008). The top 5000 genes were filtered using the rankmeans function in WGCNA.

WGCNA uses the strength of the correlation to determine the strength of the network connection—typically, β can be chosen to fit the network to a scale-free topology $A=[aij]=[|cor(xi,xj)|^{\beta}$. Here, β was chosen as 7 based on the lowest value that produced a scale-free topology in the network. A Toplogical Overlap Metric (TOM) was calculated as described in (Yip & Horvath, 2007) and probes were clustered and assigned to modules using the "blockwisemodule" function with a signed Spearman rank correlation with β=7, and a deepsplit level of 2 (which represents a medium level of sensitivity in terms of how modules are detected), a minimum module size of 40, and clustering based on the Dynamic Treecut algorithm (Langfelder, Zhang, & Horvath, 2008). Eigengenes were calculated from each module and *p*-values calculated based on the functions in the WGCNA package (Langfelder & Horvath, 2007). The network was based on the TOM calculated from

an unsigned network with the same settings used for module detection, filtered down to only genes in the statistically significant modules and with an edge weight greater than .20.

2.2 Enrichment Analysis: Probes for each module were entered into DAVID (Dennis Jr et al., 2003) and analyzed for enrichment with a stringency set to high and all other settings at default.

2.3 Visualization: All networks were visualized in Cytoscape 3.0 (Shannon et al., 2003) with the yFiles circular layout or spring embedded bio-layout.

2.4 Genetic Regulatory Network: All probes mapped to each module were entered into MSigDB (Subramanian et al., 2005) and all the top 100 motifs and microRNA binding sites with a false discovery corrected *p*-value .01 were retrieved. Statistically significant curated gene sets were retrieved as well.

Text-mining was performed in PubMed with the name of the transcription factors and Parkinson's disease as a MeSH term and "1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine" OR "1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine" OR "mptp" and, in the cases of fewer than 5 abstracts returned, were manually inspected for relevance.

Gene symbols were entered into FANTOM4 EdgeExpressDB (Severin et al., 2009); cases of ambiguity were resolved manually. ChIP interactions, siRNA experiments, or published interactions were considered as experimental evidence. Predicted evidence was either the transcription factor binding predictions or mirRNA predictions.

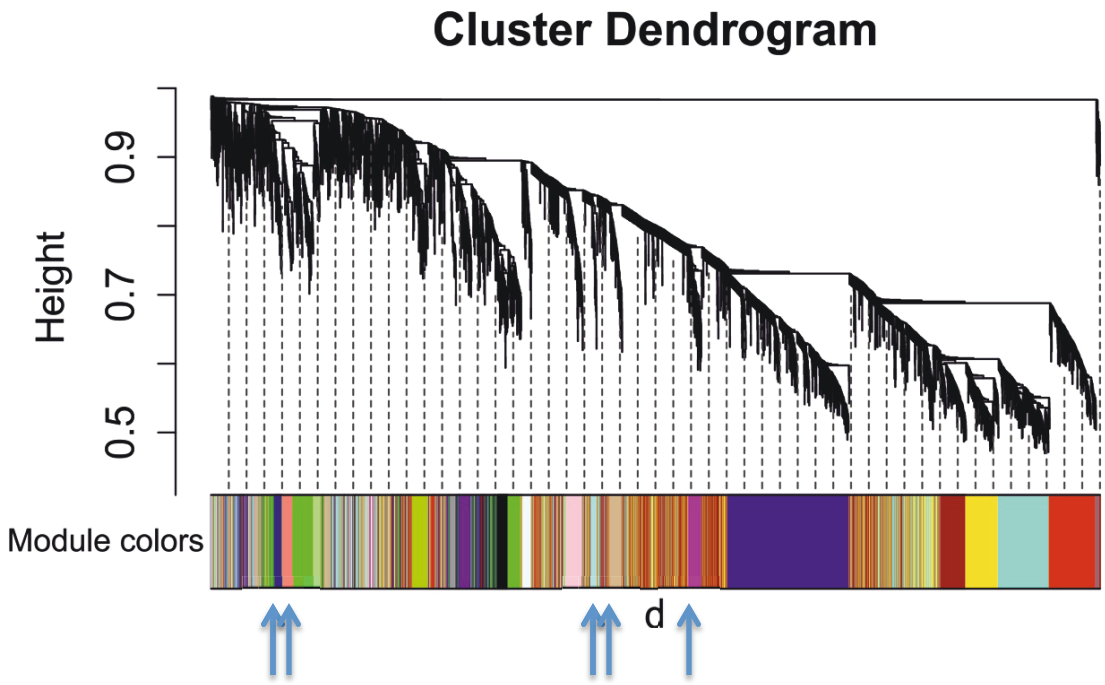
3. Results

3.1 WGCNA Clustered Probes by Function and Captured the Relevant Pathways:

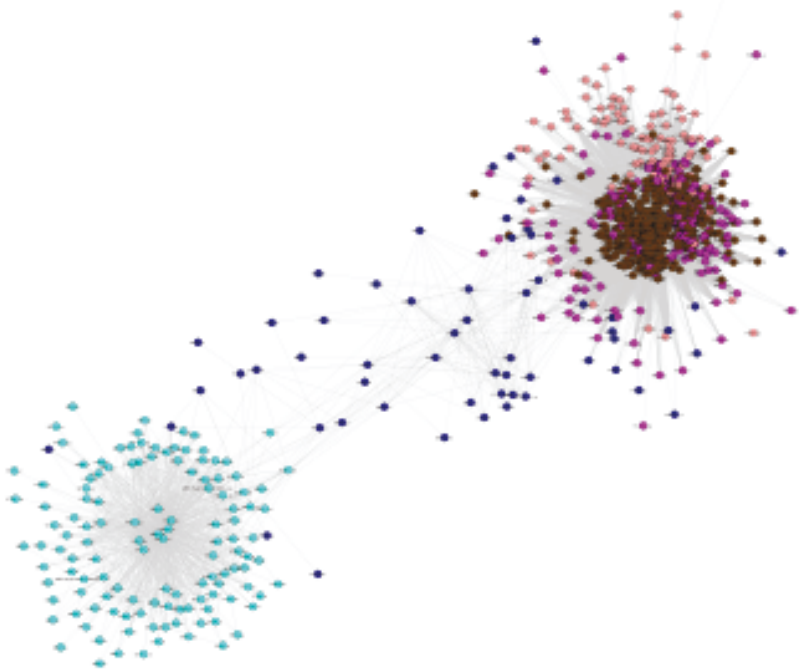
While correlation networks—often referred to as "guilt-by-association" analysis (Quackenbush, 2003)—are commonly used to derive networks *de novo* from microarray data, Weighted Gene Correlation Network Analysis (WGCNA) offers several advantages. Unweighted correlation networks typically establish a hard cut-off for a link, but WGCNA links each gene by a weight, and this network is used to derive a Topological Overlap Metric, which is most simply thought of as a measurement of gene interconnectivity. This combines the advantages of a correlation network with the insights that can be gleaned from a graph-theoretical approach; it is typically more sensitive to "weaker" connections amongst genes that may be significant, while at the same time it is somewhat more robust to noise (Langfelder and Hovarth, 2008). We chose MPTP toxicity, a commonly-used toxicity model for Parkinson's disease, and located a publicly available GEO Data Set produced from tissue isolated from the *substania nigra* of male C57BL/6J mice dosed at 10 weeks of age with a total of three doses of 30 mg/kg MPTP dosed via i.p. or saline control and killed either 24 hours or 7 days after the final dose of neurotoxin. Biological replicates were pooled and twelve arrays total were used with four arrays per group; the control group were un-dosed and sacrificed at 10 weeks. The initial data set was downloaded as RMA normalized data, filtered for the top 5000 probes by ranked mean differential expression, and used to produce the initial network which was divided into modules based on the Topological Overlap Metric as clustered by the Dynamic Tree Cut algorithm (Figure 1). The modules

were summarized as "eigengenes"—essentially, the first principal component of all genes' expression for that module, which represents an "expression signature." The eigengenes are then correlated with the phenotypic label, in this case time (Control, Day 1, and Day 7). Five modules were statistically significant (shown in figure 1) with the Midnight Blue module having the highest correlation (Table 1). Unassigned genes had no significant correlation, as would be expected. Therefore, WGCNA identified in an untargeted approach a total of 1,247 genes in five clusters that were significantly correlated to the phenotype label.

One of the underlying ideas of WGCNA is that genes with a similar function will cluster together. In order to both ensure that the clusters produced were biologically meaningful and that they captured the known biological processes involved in MPTP toxicity, we analyzed the modules using DAVID for over-represented annotations. All significant modules except one, the Midnight Blue module, were significantly enriched for terms when investigated by DAVID, and the DAVID Enrichment clusters captured the known biology of MPTP toxicity (e.g. apoptosis - Magenta module; oxidative phosphorylation/Parkinson's disease - Brown module) (Table 2). The resulting network was visualized in Cytoscape (Figure 2). One advantage of WGCNA is that it is a dimensionality reduction technique that allows for some insight into the interrelationship amongst the modules. As can be seen from the network (Figure 2), three modules (Brown, Salmon, and Magenta) were fairly tightly interconnected, while the Midnight Blue module appeared as a sparse module, which connected the Cyan module with the other three. This suggests that the Midnight Blue module may act to coordinate the distinct functions of the other three modules, which may be mediated by transcription factor TCF3 (see discussion below).



Chapter III, Figure 1: Dendrogram derived from GEODataset GDS2053, clustered by a Weighted Gene Correlation Network using Dynamic Tree Cut Algorithm. Significant modules are Midnight Blue, Salmon, Cyan, Brown, and Magenta, indicated with an arrow underneath.



Chapter III, Figure 2: Network generated by WGCNA, colored by module, using spring embedded bio-layout based on edge strength

Module	Correlation	p-Value	Genes
Magenta	0.7996246	1.80E-03	212
Salmon	0.76605824	3.67E-03	184
Brown	0.58916781	4.38E-02	560
Cyan	0.69331419	1.24E-02	177
Midnight Blue	0.94604195	3.29E-06	125
Unassigned	0.13829676	6.68E-01	68

Chapter III, Table 1: Modules Correlated with Time.
Legend, Table 1: Five of the modules produced were significantly correlated; significance is calculated via a permutation test.

BROWN		
Annotation Cluster 1	Enrichment Score: 9.1	
	GO Term	structural constituent of ribosome
	KEGG Pathway	ribosome
Annotation Cluster 2	Enrichment Score: 6.31	
	GO TERM MF	mitochondrion
	GO TERM MF	generation of precursor metabolites and energy
	KEGG Pathway	Oxidative phosphorylation
	KEGG Pathway	Parkinson's disease
SALMON		
Annotation Cluster 1	Enrichment Score: 2.37	
	GOTERM CC	vacuole
	GOTERM CC	lysosome
Annotation Cluster 2	Enrichment Score: 2.18	
	GOTERM BP	positive regulation of transcription
	GOTERM BP	positive regulation of gene expression
	GOTERM BP	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
CYAN		
Annotation Cluster 1	Enrichment Score: 3.75	
	SIR Protein Keywor	nucleotide-binding
	GO TERM MF	purine nucleotide binding
	GO TERM MF	ribonucleotide binding
Annotation Cluster 2	Enrichment Score: 2.67	
	GOTERM BP	intracellular protein transport
	GOTERM BP	cellular protein localization
	GOTERM BP	cellular macromolecule localization
	GOTERM BP	protein transport
MAGENTA		
Annotation Cluster 1	Enrichment Score: 2.71	
	GOTERM BP	regulation of protein kinase activity
	GOTERM BP	regulation of transferase activity
	GOTERM BP	regulation of phosphorus metabolic process
Annotation Cluster 2	Enrichment Score: 2.16	
	GO TERM BP	negative regulation of apoptosis
	GO TERM BP	negative regulation of programmed cell death
MIDNIGHT BLUE		
Annotation Cluster 1	Enrichment Score: 1.4	
	GO TERM - BP	hexose metabolic process
	GO TERM - BP	monosaccharide metabolic process
	GO TERM - BP	glucose metabolic process

Chapter III, Table 2: Modules annotated by DAVID.
Legend, Table 2: All modules except the Midnight Blue module were significantly enriched for annotation terms, reflecting that WGCNA had clustered probes by function and identified physiologically relevant functions, and the presence of apoptosis and the KEGG Pathway for Parkinson's disease indicates that the known physiology of MPTP toxicity was captured by the modules.

Correlation networks typically have both a high rate of false positives and provide no insight into the regulatory mechanisms. Therefore, to bring this approach closer to a mechanistically specified network and to better characterize the underlying biology, each module was analyzed for gene signatures in the Chemical and Genetic Perturbation subset of MSigDB as well as for over-represented transcription factor binding sites. Each module, except for the Magenta module, was substantially enriched for genes involved in Alzheimer's disease (Table 3), and while Alzheimer's has a different mechanism of neuronal degeneration compared to Parkinson's, this does indicate that the approach selected genes that are involve in neurodegenerative disease. Furthermore, it indicates that while the Midnight Blue module had no annotations to establish the functional significance of the cluster, the genes identified are related to neurodegeneration.

MODULE	Curated Gene Set	FDR corrected q-value
BROWN	Genes down-regulated in brain from patients with Alzheimer's	$2.05\ e^{-50}$
CYAN	Genes up-regulated in brain from patients with Alzheimer's disease.	$6.38\ e^{-8}$
MIDNIGHT BLUE	Genes up-regulated in brain from patients with Alzheimer's disease.	$6.55\ e^{-4}$

SALMON	Genes up-regulated in brain from patients with Alzheimer's disease.	2.27 e^{-3}

Chapter III, Table 3: Overlap of the modules with Alzheimer's genes, from MSigDB C2 gene sets. Alzheimer's genes are based on the Blalock dataset (Blalock et al., 2004).

3.2 Modules Were Enriched For Transcription Factors Relevant To Parkinson's Disease

Parkinson's Disease: One biological reason for correlation of gene expression is common transcription factors or microRNAs. Therefore, each module was also analyzed in MSigDB for enriched transcription factor binding sites with an FDR corrected *p*-value of less than .01. This generated a list of 114 candidate transcription factor binding sites that were enriched in the modules (of which 25 had no known transcription factor) and 23 microRNA binding sites. All modules had more than 10 predicted enriched motifs, and there was substantial overlap between enriched motifs amongst the modules. Candidate transcription factors and microRNA were text-mined for association with either Parkinson's disease or MPTP toxicity, and any transcription factor with more than two articles for Parkinson's and/or MPTP toxicity (one article for microRNA, owing to the smaller literature base) were considered relevant for building a genetic regulatory network (Table 4). This methodology found transcription factors that were well known for Parkinson's—JUN and NRF2, as well as ELK1, which had both literature evidence for Parkinson's and were in the Parkinson's Pathway in the PANTHER Database (Mi, Muruganujan, Casagrande, & Thomas, 2013). Additionally, one of the transcription

factor binding sites—SP1—had relatively few articles for Parkinson's disease, but did have binding motifs enriched in each of the modules (Table 4). SP1 was the only transcription factor with annotations for Parkinson's that was identified by MSigDB as relevant for Midnight Blue; the Cyan module had many transcription factors that were not shared with other modules, while the Brown module, in keeping with its size, had the largest number of potential transcription factors.

Transcription Factor	**Abstract for Parkinson's**	**MPTP/MPP+**	**Module**	**FDR Corrected P-Value**
JUN	4451	729	BROWN	3.44E-08
NRFR2	59	25	SALMON	6.23E-03
FOXF2	21	1	BROWN	1.69E-07
SP1	12	2	BROWN	4.25E-26
			CYAN	2.51E-06
			MAGENTA	8.08E-05
			MIDNIGHT BLUE	2.53E-03
			SALMON	7.85E-05
ATF4	12	2	BROWN	9.14E-07
TCF3	11	6	BROWN	9.20E-07
			CYAN	5.82E-04
			SALMON	2.52E-02
ELK1	3		BROWN	3.14E-15
			MAGENTA	8.08E-05
AP1	3	1	BROWN	3.44E-08
			SALMON	2.52E-02

STAT1	7	2	BROWN	3.76E-06
NRF1	6	3	BROWN	2.34E-06
			CYAN	7.59E-04
SRY	6		CYAN	4.84E-03
MIR-132	5		BROWN	3.76E-06
SREBF1	5		CYAN	7.59E-04
ATF3	4	1	CYAN	4.84E-03
MIR30C	1		CYAN	4.84E-03
SRY	6		CYAN	4.84E-03
MIR221	2		CYAN	4.95E-03
MEF2A	2		CYAN	9.52E-03
			MAGENTA	0.000809
ELK1	2		MAGENTA	8.08E-05
			BROWN	3.14E-15

Chapter III, Table 4: Candidate transcription factors associated with Parkinson's and MPTP via text-mining; *p*-values based on enrichment in MSigDB C3 gene sets.

3.3 Transcription Factors Significantly Improved The Number Of Genes That Could Be Connected In A Component: For each module, all of the genes that could be located in the FANTOM4 database were analyzed with and without the subset of transcription factors both significant for that module and identified as being relevant to Parkinson's disease to form the basis of a genetic regulatory network for that module. All modules but one—the Midnight Blue module—contained a subset of genes that were connected by experimentally verified regulatory interactions in FANTOM4 (ChIP data, siRNA, or published interactions),

indicating that the modules consisted of genes that could be connected to each other with experimental data (Table 5).

However, the "connected component"—that is to say, the largest subset of genes and proteins that were interconnected with each other—grew substantially with the addition of the predicted transcription factors as identified by MSigDB and text-mining, even when restricting to experimental evidence; the percent of the module connected by experimentally verified interactions ranged from a low of 70-80 percent for each module, and was 100 percent for the Midnight Blue module (Table 5). For each module, the transcription factor that had, by far, the highest number of interactions was SP1 and it also had substantial experimental evidence of interactions (see Fig 3, 5,6,7,8). Within the Brown module, a subnetwork centered around SP1 and JUN indicated that it not only activated JUN but was connected to several downstream components as well (see Fig. 4). Within the Midnight Blue module, even when restricted to evidence of 4 ChIP experiments, SP1 remained a significant hub (see Fig 8).

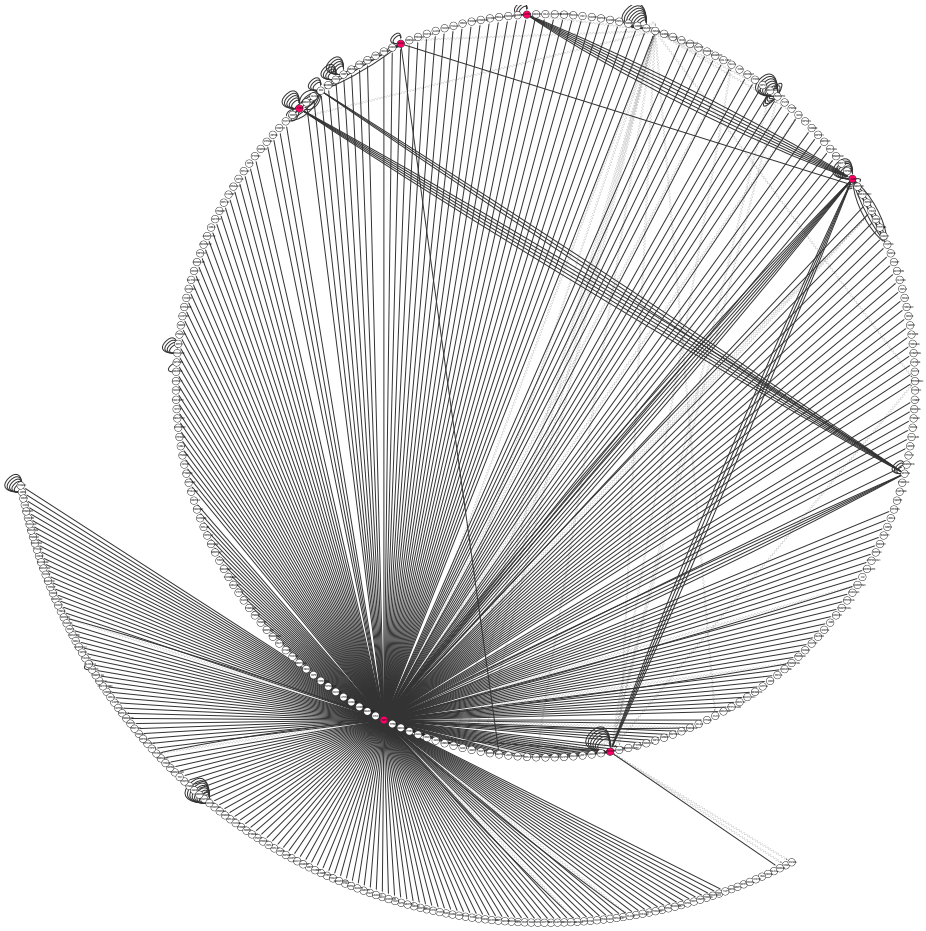
		CONNECTED COMPONENT – WITHOUT TFS			CONNECTED COMPONENT WITH TFS	
MODULE	GENES IN FANTOM	EXPERIMENTAL	PREDICTED		EXPERIMENTAL	PREDICTED
SALMON	163	14	47	SP1, NRF2, TCF3, AP1	125	132
MIDNIGHT BLUE	105	0	14	SP1	75	105
CYAN	150	16	41	SP1, NRF1, SRY,SREBF1, MIR221,MEF 2A, TCF3	121	130
BROWN	463	31	163	SP1, JUN*, FOXF2, ATF4, TCF3, ELK1, STAT1*, NRF1, MIR132	381	409
MAGENTA	106	14	26	SP1, ELK1, MEF2A	82	91

Chapter III, Table 5: Addition of Predicted Transcription Factors Substantially Increased Connected Component.
Table 5: For each module, gene symbols were entered into FANTOM4 EdgeExpressDB and a predicted regulatory network was drawn based on experimental evidence (ChIP, published interactions and siRNA experiments), with and without the addition of predicted evidence (predicted transcription factor binding and microRNA). Transcription factors were added based on evidence of significantly over-represented motifs in MSigDB and textual evidence of involvement in Parkinson's. *In the case of the Brown module, STAT1 and JUN was already in the module. "Connected component" consists of all genes that were not singletons in the predicted regulatory network.

SP1 is a ubiquitously expressed transcription factor that regulates a sweeping number of genes during development and other cellular functions. SP1 is known to play a key role in tissue differentiation; knock-out mice are embryo-lethal

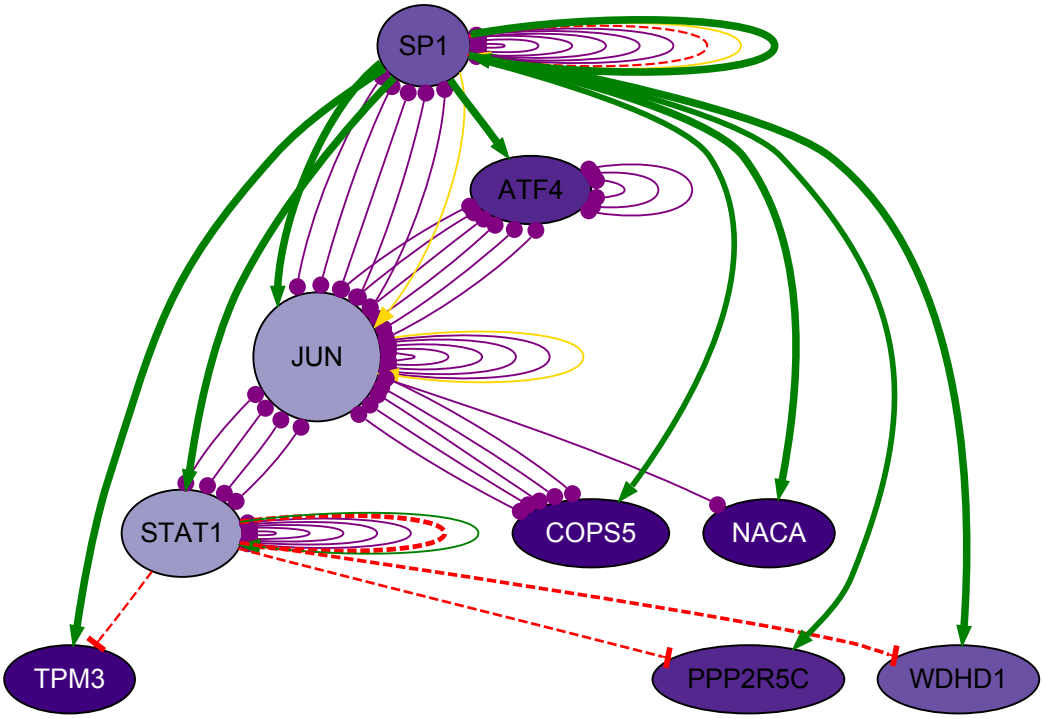
and have multiple abnormalities (OMIM). SP1 is also known to play a role in cell-cycle inhibition (Deniaud et al., 2009) and over-expression leads to apoptosis (Chuang, Wu, Lai, Chang, & Hung, 2009). Furthermore, SP1 is known to regulate the dopamine transporter (J. Wang & Bannon, 2005), and is involved in several neurodegenerative diseases (Qiu et al., 2006) (Santpere, Nieto, Puig, & Ferrer, 2006). SP1 is known to be acetylated in neurons in response to oxidative stress and works in tandem with histone deactylaces to prevent cell death (Ryu et al., 2003); acetylation is but one of many post-translational modifications that expand SP1's response repertoire.

SP1 was not present in the modules, nor was it amongst the genes differentially expressed, even with the most generous of cut-off values for significance. However, SP1 protein and mRNA levels have been shown to increase following MPP+ dosing in PC12 cells by approximately 1.5 fold, which was blocked by antioxidant treatment (Ye, Zhang, Huang, Zhu, & Chen, 2013). The lack of appearance of SP1 amongst the genes differentially expressed or in the modules may simply reflect that SP1 mRNA rises only modestly or perhaps briefly, or, alternatively, it is regulated by means other than an increase in mRNA levels, and the signal increase is therefore non-linear compared to mRNA levels (Courey, Holtzman, Jackson, & Tjian, 1989). As SP1 is constitutively expressed rather than inducible, it may also act as a preliminary sensor that initiates the cascade.

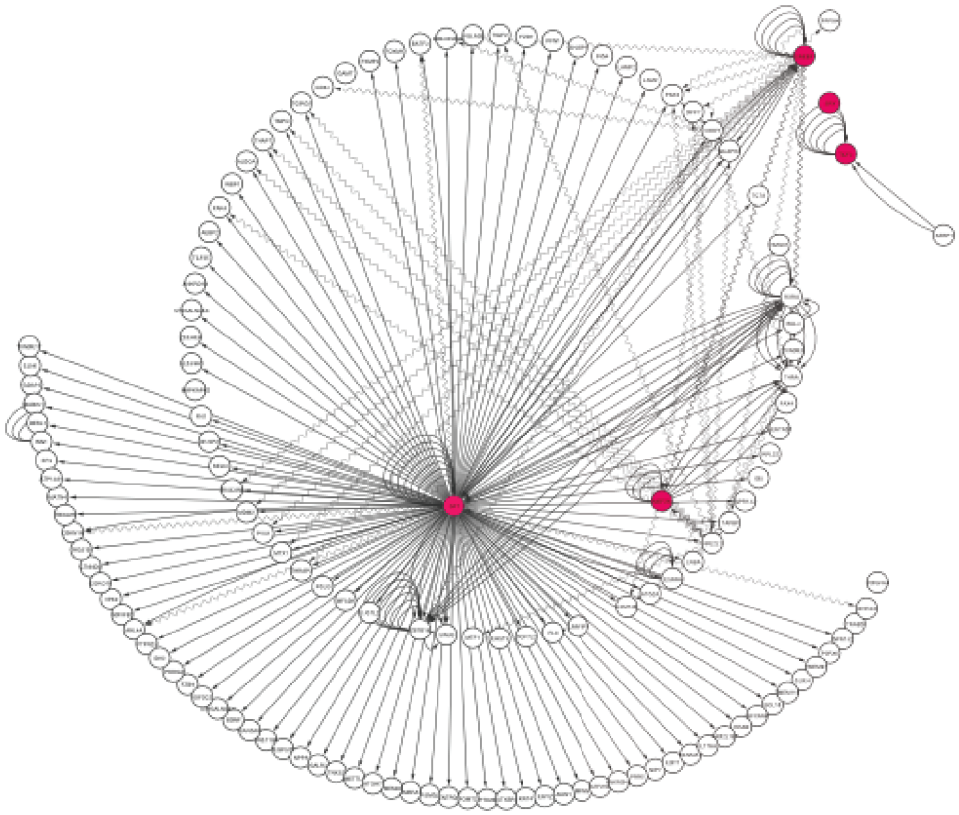


Chapter V, Figure 3: Brown Module, Identified Transcription Factors in Red.

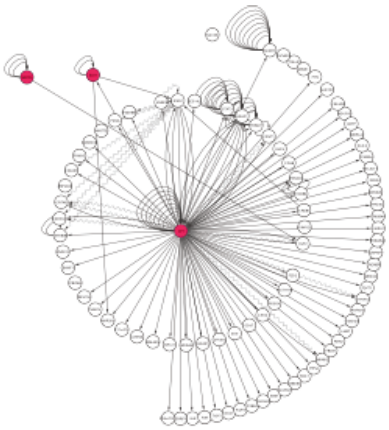
Figure 3, Legend: The Brown module formed a dense of network of regulatory interactions centered on SP1. Self-loops indicate a gene interacts with itself.



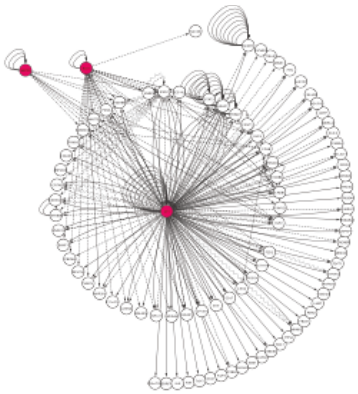
Chapter III, Figure 4: SP1, JUN, and STAT1 subnetwork from the Brown Module.
Legend: Green indicates ChIP data; red indicates perturbation experiment; yellow, published protein-DNA interactions, and purple indicates protein-protein interaction. Node size is proportional to predicted dynamics of the gene, and darker nodes indicate higher scaled expression levels. Because the FANTOM4 database gives an estimate of the dynamics of gene expression, the resulting gene regulatory network can be used as the foundation for building a dynamic model.



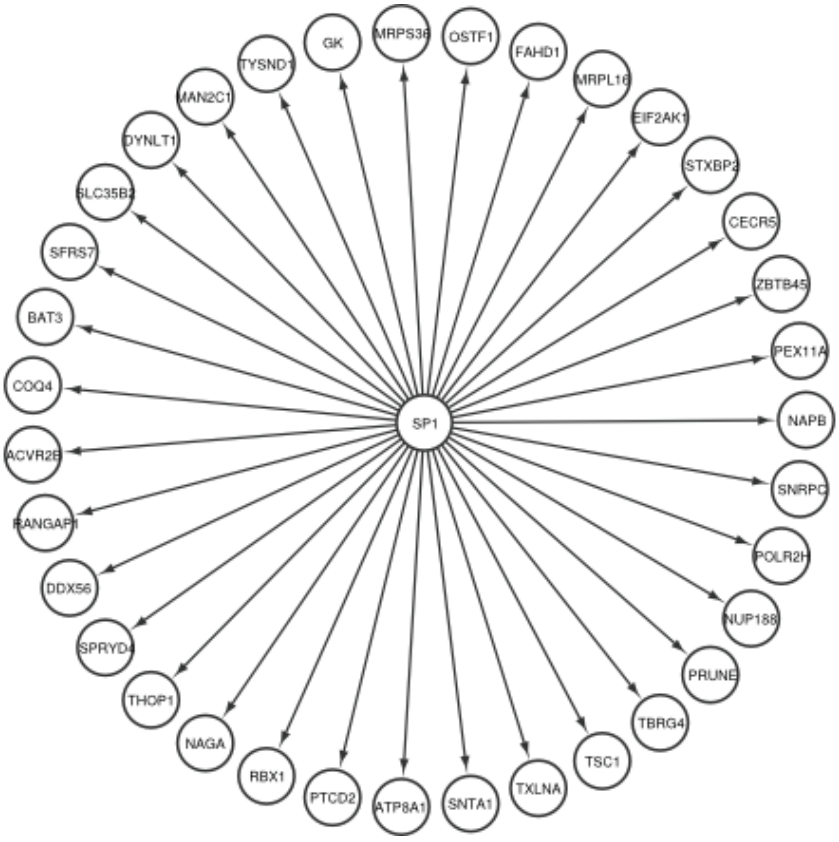
Chapter III, Figure 5: Cyan Module with TFs identified as indicated in red; SP1 is in the middle.
Legend: Edges represent experimentally verified interactions; all nodes were connected when predicted interactions were included (data not shown).



Chapter V, Figure 6: Magenta Module; Experimental



Chapter III, Figure 7: Magenta Module; Experimental and Predicted
Legend, Figure 7: Experimental Network Edges Versus Predicted; most of the edges are based on experimental data (Figure 6), the addition of transcription factor binding predictions extended the network (Figure 7), but did not substantially change the network architecture. Experimental edges are indicated with a solid line, predicted with a dash.



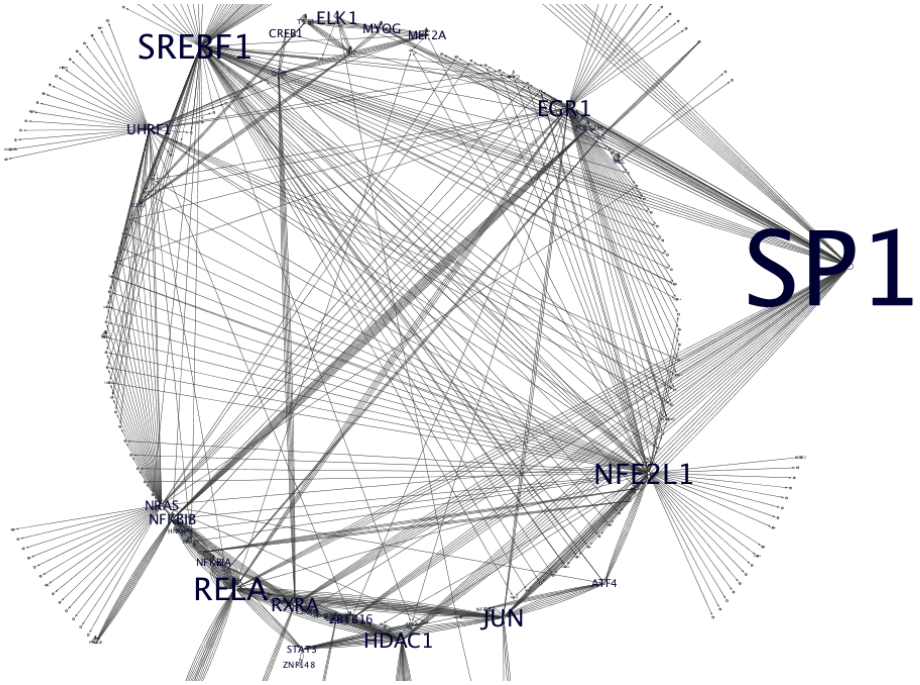
Chapter III, Figure 8: **Midnight Blue Module; SP1 interactions verified with 4 ChIP experiments.**

Genes from all modules were combined into a genetic regulatory network based on FANTOM4 interactions, as follows: (1) evidence limited to published interactions and siRNA perturbation data, (2) published/perturbation data with the addition of ChIP data, and (3) all evidence, including transcription factor binding sites. Even when restricted to published evidence, the resulting genetic regulatory network consisted of a connected component of 256 genes with several hubs (Figure 9). Including ChIP data extended it to 782 and predicted transcription factor bind sites to 830. In addition to SP1, the network hubs consist of some candidates well known for their role in Parkinson's (STAT3, JUN) but also produced other candidates that have been implicated in Parkinson's. SREBF1 has previously been identified as a risk locus for sporadic Parkinson's disease (Do et al., 2011) and in a recent RNAi screening study, it was implicated in the control of the PTEN-induced kinase 1 (PINK1)/Parkin pathways that control the autophagic destruction of mitochondria (Ivatt & Whitworth, 2014).

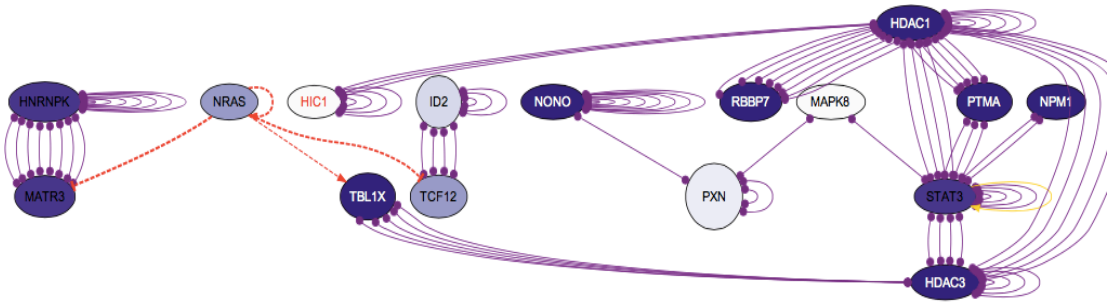
One hub identified in the reconstructed GRN from FANTOM4, HDAC1, has been implicated in cell-survival in neurotoxicity to dopaminergic neurons *in vitro* and ischemia *in vivo* (Kim et al., 2008); HDAC1 was also a hub in the WGCNA network and many of the first neighbors of HDAC1 in the FANTOM4 network were also first neighbors in the WGCNA network (Figure 10, 11). The WGCNA network also suggested a protein, LANCL1, that was connected to both HDAC1 and STAT3 (Figure 11). LANCL1 binds glutathione and is believed to play a role in neuronal

survival following oxidative insult (Zhong et al., 2012), and its connection to HDAC1 and STAT3 seems plausible.

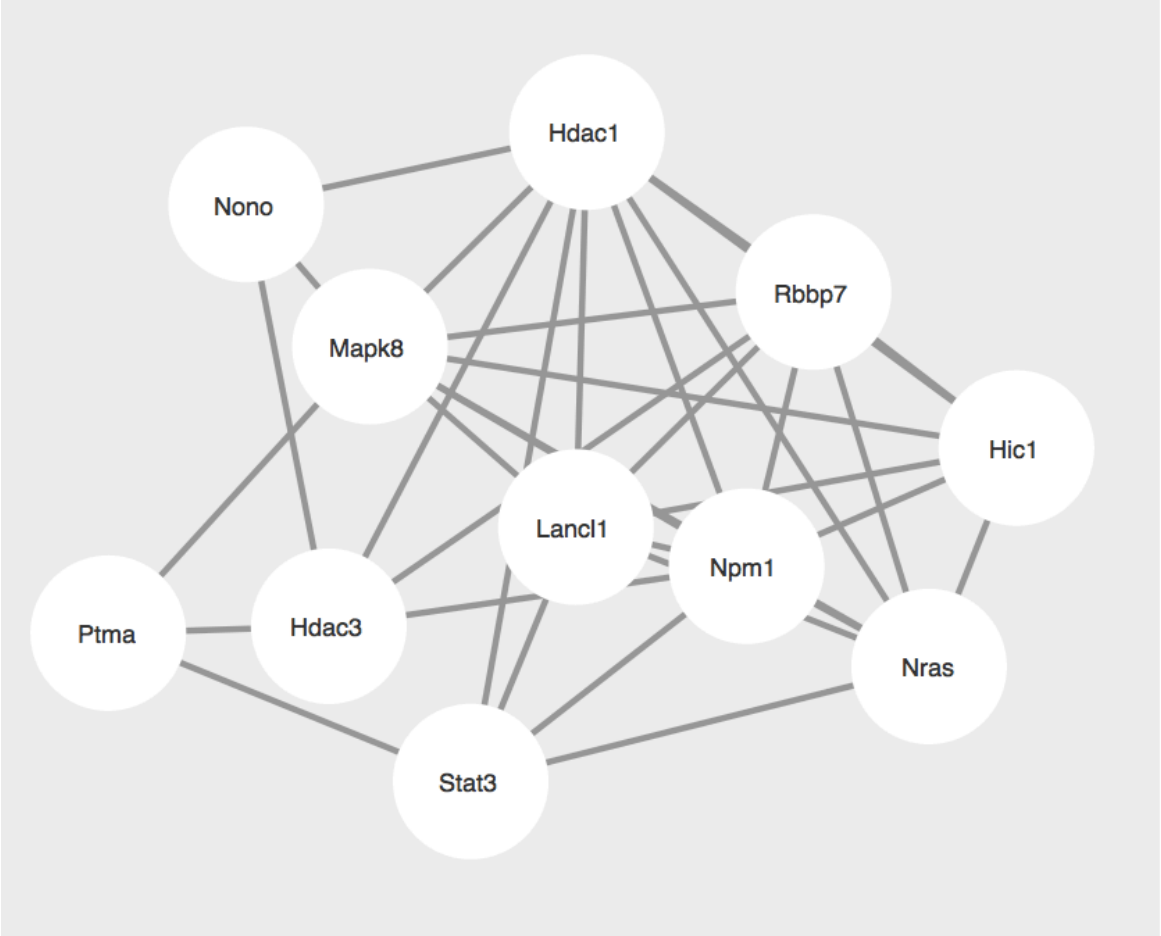
One of the smaller hubs, ZNF148, a zinc-binding transcription factor, had several predicted connections from FANTOM4 within the Midnight Blue module; of the 18 genes connected to ZNF148 in the original WGCNA network, four were also linked by predicted interactions in FANTOM4. ZNF148 (also referred to as ZPB89) is not present on any pathway in Panther or KEGG and has a relatively sparse literature base with no indication of any role in Parkinson's. However, ZNF148 is known to play a role in apoptosis (Zhang, Chen, & Lai, 2010), and would be an interesting candidate for further study. ATF4, which has recently been identified in other high-throughput studies as a key transcriptional factor in MPTP toxicity by us and others (Ye et al., 2013), (Krug et al., 2014), was also present as a small hub containing mostly protein-protein interaction connections in the network when restricted to experimentally verified interactions. Similarly, TCF3 had relatively few experimentally verified reactions and is thus relatively small in the graph; however, an expanded subnetwork that included predicted transcription factor binding sites, even when restricted to a high stringency level, would have been substantially larger. TCF3 was in the Midnight Blue module, and the Cyan, Salmon, and Brown module were all enriched for TCF3 binding motifs. This is likely a case where the relative importance of a gene is underestimated based on the lack of available experimental data in comparison to the better-studied SP1.



Chapter III, Figure 9: Genetic Regulatory Network based on published interactions.
Legend, Figure 9: Node label is proportionate to hub status as determined by edge count. Self-interactions were deleted for visual clarity.



Chapter III, Figure 10: HDAC Subnetwork from FANTOM4; single leaves collapsed for visual clarity



Chapter III, Figure 11: HDAC1 Subnetwork, WGCNA.
Legend, Figure 11: All genes from Figure 7 (and any common nearest neighbors) as seen in WGCNA. The connected component from the GRN was present in WGCNA. In addition, HDAC1 was not only directly connected to STAT3 (as it was in the FANTOM4 GRN) but was also connected through LANCL1.

One of the transcription factor binding sites that was consistently ranked by MSigDB across all modules (PAX4), had no textual evidence for involvement with Parkinson's, although it does appear to be expressed in the brain. However, there was no experimental evidence in FANTOM4 that it bound to any of the targets in the modules, the predicted targets were quite sparse, and its inclusion would not have fundamentally changed the architecture of the network. This would seem to indicate that the resultant genetic regulatory network is not just reflecting non-specific

predicted transcription factor binding motifs that are enriched in a subset of genes, but is instead constructing a genetic regulatory network that is likely enriched for biologically relevant targets.

4. Discussion

Current analysis of microarray data in toxicology does not take advantage of the data-mining and bioinformatics tools available to interpret the underlying mechanisms but remains at the level of "biomarker" or signature identification, either generating a relatively small list of genes differentially expressed using inferential statistics, or over-representation analysis which is highly dependent on pathway annotations. We chose an existing dataset, which was originally used to identify a few genes as signatures of MPTP toxicity *in vivo*, in order to explore an alternative method that would offer more insight into dynamics of gene expression compared to inferential statistics and would not be dependent on pathway annotations. WGCNA offer many advantages for analyzing microarray data: it is unsupervised, and, unlike correlation networks that are based solely on a Pearson or Spearman correlation, it preserves weak links—capturing interactions that may be small, but that may nonetheless be biologically interesting; this may be especially relevant to toxicology, as the effects may be subtle and distributed amongst many pathways.

As this represented a fairly unsophisticated approach to text-mining transcription factor candidates, it is quite probable that the proposed regulatory network is only a "10,000-ft" view; many of the transcription factors may have had textual evidence of being involved in physiological processes that are relevant to MPTP toxicity—e.g. oxidative stress or apoptosis—although extending the text-

mining in such a way would likely also have increased the false-positives, it could also fine-tune the map in some of the "neighborhoods." MPTP's Pathway of Toxicity is aided markedly by the fact it serves as an animal model for a relatively well-researched disease such as Parkinson's; depending on MPTP/MPP+ literature would have produced a much smaller subset of candidate transcription factors and no microRNAs, perhaps reflecting the relatively immature literature base from microRNAs compared to transcription factors.

Furthermore, just as the "connected component" likely contains some regulatory connections that are artifacts, the unconnected component contains both genes that are spurious correlations as well as genes that are unconnected due to lack of data about the probable regulatory mechanism. Disappointingly, neither of the two microRNAs that were identified as candidates were found to have regulatory connections; this may reflect the fact that microRNAs simply have an inadequate dataset, and it is likely that multiple microRNAs are involved but are simply invisible in this analysis. Surprisingly, one of the "unconnected" genes in the Brown module was MAO-A (Monoamine Oxidase A). Although MPTP is metabolized much more efficiently by MAO-B, MAO-A is possibly involved in dopaminergic cell death in neurons (Naoi, Maruyama, & Inaba-Hasegawa, 2012) and there is evidence that SP1 binds to the promoter of MAO-A (Zhu, Chen, & Shih, 1994).

Similarly, within the Midnight Blue module, two proteins that had relatively weak evidence of connection to SP1, AQP4 (Aquaporin-4) and TUB (Tubby protein), and were not in the final genetic regulatory network, were examined for evidence of involvement in MPTP toxicity related processes, as both have knock-out mice

models. Aquaporin-4 knock-out mice are more prone to MPTP toxicity (Fan et al., 2008) and although Aquaporin-4 may or may not be regulated by SP1, it likely does play a role in the ultimate phenotypic consequences of MPTP toxicity and perhaps Parkinson's as well. Tubby protein knock-out mice have a primary phenotype of obesity but also display neurodegeneration. There is some evidence that TUB is a regulator of microglial phagocytosis through the MerTK receptor. However, the exact nature and role of TUB in MPTP toxicity remains speculative. Nonetheless, neither the genes unconnected to the larger network nor the weaker links in the network that lack substantial experimental evidence should be discarded wholesale.

Mitochondrial disruption is a commonality for a variety of neurotoxins and neurodegenerative diseases; however, often the exact route between mitochondrial disruption and the phenotype is unclear. MPTP, like other toxins, may work primarily to disrupt the mitochondria, but the disruption likely has pleiotropic effects that differ from other toxins and disease states. Depending on annotations to reveal physiological function (or, alternatively, discarding a cluster because of lack of annotations) may miss useful information about toxic processes. In this case, the Midnight Blue module contained genes known or strongly suspected to be involved in Parkinson's or MPTP toxicity (MAPT, SYNGR1) as well as genes known to be involved in neuropathology (THOP1, which cleaves amyloid precursor protein) (Pollio et al., 2008). It also suggested novel candidates that are plausibly involved in the degenerative process (AQP4 and TUB), neither of which were on existing Parkinson's pathways (Panther, KEGG) and both of which had an inadequate literature depth on which to base enrichment analysis.

The pronounced promiscuousness of SP1 binding sites entails that many, if not most, of the predicted interactions are spurious and the experimentally verified interactions may be irrelevant within the particular context of MPTP toxicity in dopaminergic neurons. However, given the statistically significant over-representation of SP1 motifs in all the modules, the centrality of SP1 to the predicted network, the literature evidence of involvement in Parkinson's, dopamine regulation, and MPTP toxicity, and the experimental evidence of interactions with known signaling networks (such as JUN) involved in Parkinson's and MPTP, SP1 is likely necessary (though not sufficient) for MPTP toxicity and acts to integrate multiple signaling pathways in a combinatorial and complex manner. The proposed genetic regulatory network offers an advantage compared to a correlation network insofar as it offers a direction of action, an estimate of transcription factor binding site strength, multiple lines of evidence, and an estimate of the dynamics of gene expression. Therefore, it can act as scaffolding, which further experiments, both *in silico* and *in vitro,* can refine. Although this study isolated neurons, it likely benefitted from capturing the complex interplay between neurons and astrocytes, and specifically the inflammatory contribution of astrocytic processes.

This study shows that a relatively small gene array study allows for the pinpointing of mechanistic information by a combination of correlative and data-mining approaches and can suggest many plausible candidates for further study. However, any data-mining approach—especially ones that tend to generate false-positives—has to go hand-in-hand with confirmation of the (patho-)physiological sense of the distilled information. These emerging approaches for Pathway of

Toxicity identification can become even more powerful when several orthogonal omics technologies are employed and different experimental models are combined. Furthermore, a better understanding of the dynamics of the toxic process can suggest a better experimental framework to investigate the process with a greater granularity. If, as the *in vitro* data suggested, SP1 is an early sensor for oxidative stress, experiments that focused on earlier time points might offer greater insight into temporal dynamics of the initiating events. A study with time-matched controls (as opposed to this study, which did not have a vehicle-treated control at each time point) might offer greater power to see differences of treated vs normal. Lastly, using a complimentary technology – such as RNA-seq – would offer an opportunity to confirm the results with another technology, and at the same time would have the advantage of offering greater resolution for lower abundance transcripts.

The exploration of the Genetic Regulatory Network was aided by the extensive database of RNAi and ChIP-seq experiments that target SP1, and building upon these experiments by targeting other candidate transcription factors would likely identify more precisely the regulatory mechanisms involved downstream of SP1, especially if the experiments were performed in an *in vitro* neuron model.

If nothing else, this study has indicated the extant to which our knowledge of signaling networks involved in MPTP toxicity is likely limited to the downstream consequences of damage long before the initial event that perturbs homeostasis.

CHAPTER IV – Pathways of Toxicity and Metabolomics

Abbreviations: BPA (Bisphenol-A), HMDB (Human Metabolome Database), SMILES (simplified molecular-input line-entry system), QEA (Quantitative Enrcichment Analysis), ORA (Over Representation Analysis)

1. Introduction: Metabolomics—The Promise and the Pitfalls

Analyzing metabolites for altered phenotype is perhaps one of the oldest modalities employed by medicine—the tale goes that pre-modern doctors would test for *diabetes mellitus* by seeing if ants were attracted to a patient's urine, a crude but effective and accurate assay for a biomarker of disease. Nonetheless, metabolomics—defined as measuring the concentration of 'all' low molecular weight (1500 Da) molecules in a system of interest—has yet to join transcriptomics and proteomics as an essential part of systems biology.

Metabolomics represents a promising new way to approach some of the problems in the hazard assessment of endocrine disruptors. As an example, endocrine disruptors have generated a great deal of controversy regarding the adequacy of existing conventional animal assays that extrapolate from high doses to low doses and may be insensitive to subtle, long-term effects at doses relevant to human exposure; despite millions of dollars and over 5000 safety-related studies (Hengstler et al., 2011), the safety of BPA (bisphenol-A) is still relatively contentious, and this confusion is reflected at the regulatory level. While the evidence has generally indicated that the concern about BPA was based more on public perceptions of risk from poorly done studies rather than on the overall science studies (Goodman et al., 2009), it does indicate the extent to which a lack of

clarity in the data can result in a profound problem for risk assessment and regulatory toxicology, as well as the difficulty of bridging experimental science and traditional, regulatory toxicology testing (Borrell, 2010). Endocrine disruptors represent a complex physiological puzzle that would benefit from an approach that uses high-content data on a human-based tissue to explain mechanistically precisely how endocrine disruptors cause altered phenotypes (Bouhifd et al., 2014).

At the same time, metabolomics presents many challenges. The fact that metabolomics is ultimately very close to the phenotype turns out to be a double-edged sword, as it means that metabolomics is extraordinarily sensitive to slight changes in experimental parameters, and it requires a scrupulous commitment to protocol and a long-term commitment to trouble-shooting as virtually any small change—different brands of food for animals, different plastic plates in tissue culture—can introduce artifacts. Additionally, sample preparation must be kept to a minimum as every step has the potential to add artifacts.

In terms of analytical chemistry, metabolomics presents another challenge: the universe of metabolites consists of chemicals with a vast range of properties—there are approximately 2,000 polar and natural lipids, 500 class-specific metabolites, 200 redox metabolites, and 800 primary metabolites—and the different biochemical properties precludes coverage with any one platform, e.g. HPLC will have different coverage than gas chromatography, different chromatography columns and solvent gradients can have a strong , positive or negative polarity will ionize different metabolites, etc. Therefore, while untargeted metabolomics attempts to catch “all” the metabolites, the choice of platform will

likely privilege some over others. This is important to keep in mind for pathway analysis, as metabolites that are invisible to a specific platform but are heavily represented on a pathway of interest may skew the result, i.e. cells treated with estrogen may have steroid-specific pathways up-regulated, but if a technology does not adequately capture large, non-polar compounds, any impact on that pathway may be difficult to see.

Furthermore, metabolomics, unlike transcriptomics, does not produce a list of unambiguously identified "features." Instead, it depends on several intricate steps of data analysis to go from a chromatogram to a list of metabolites with concentrations, including peak alignment, deconvolution, adequate identification of ions, isotopes, and possible adduct modifications (water, sodium, or other small molecules that may be bound or lost to/from the compound and therefore reflected in the m/z), and lastly (and the one that will be the focus here), accurate metabolite identification, which is dependent not only on all of the above steps, but also on the accuracy and metabolite coverage provided by the database used for compound identification.

One critical problem for metabolomics is that knowledge of metabolic networks is still relatively incomplete, the databases still comparatively new, and the data infrastructure lacking, which presents some challenges for both metabolite identification and pathway analysis.

Currently, there are several public databases that can be used for metabolite identification: PubChem (run by NCBI), ChEBI, Metlin, and HMDB. PubChem (Y. Wang et al., 2009) is focused on acting as a repository for all chemicals and is

therefore not exclusive to metabolites, but does have the most extensive coverage of the chemical universe. ChEBI (Chemical Entities of Biological Interest) is a database of 'small' chemical compounds that are either "products of nature or synthetic products used to intervene in the processes of living organisms" (Degtyarenko et al., 2008). Metlin (Smith et al., 2005) is exclusively focused on metabolomics and is the only database to match precursor ions; Metlin has 240,493 metabolites and is the largest metabolomics database. Pubchem, Metlin, and ChEBI are extensive in terms of coverage of the chemical universe, but may match too many compounds as none of them are exclusive to humans. HMDB (Wishart et al., 2013), which has 41,828 entries, is focused exclusively on human metabolites while KEGG (Kanehisa & Goto, 2000) (which can be used both for identification and pathway-level annotations) allows filtering based on organism, but both have issues with accuracy in terms of organism specificity. For example, both Aflatoxin G (HMDB30474) and psilocin (HMDB42000) are identified as "Endogenous"; while KEGG has many pathways that are annotated to humans but involve metabolites not endogenous (e.g. neomyocin and byturison pathway, which involve bacterial synthesis of antibiotics).

Turning to the sources that focus on pathways – in other words, that try to place metabolites into known reactions, -and attempt to provide a comprehensive map (Recon/EHMNM, HumanCyc, KEGG, and Reactome), here is remarkably little overlap. These databases differ in size—from a low of 970 metabolites in Recon1 (reflecting that it is based on manual curation) to a high of 2,676 metabolites for the EHMN (which is based in part on automated annotations) (Stobbe et al., 2013). However, somewhat worryingly, there is a striking lack of agreement amongst the

databases in terms of commonality—the five main databases agree only on 402 metabolites (nine percent of the total metabolites in the different databases) and a full 3,107 of metabolites are present in only one database (See Table 1) (Stobbe et al., 2013).

This can be attributed to several reasons. One, the databases may have different levels of granularity—for example, a reaction may include all associated molecules (including "currency molecules," such as ATP and NADH) in one database, but another database may focus only on the main players. Two, various databases were started with different aims in mind and use different identifiers—the lack of database interoperability makes it exceedingly difficult to translate chemical identifiers from one database to another, because of the lack of efficient ID conversion tools, the complexity of chemical nomenclature, and the difficulty in using structural-based IDs such as InChI and SMILES for database indexing.

	Reactions	Metabolites
Union	6910	4677
Consensus	206 (3%)	402 (9%)
Majority	1015 (15%)	984 (21%)
Unique	4805 (70%)	3107 (66%)

Chapter IV, Table 1: From Consensus and Conflict Database, data taken from http://www.molgenis.org/c2cards/molgenis.do

An illustration of the different perspectives is provided below. Both estradiol and estrone were present in all databases (Tables 1 and 2), although estradiol has three different names. Both were present as a dead-end metabolite in at least one database. Looking at the reaction between estrone and estrone-sulfate indicates

agreement on the EC number (which places it in the minority of relatively well documented reactions, as only 17 percent of reactions agree on EC number in all databases), with slight variants of the sulfo-transferase—but virtually no agreement on pathway.

Metabolite	Compartment	Dead-end metabolite?	Database
Estradiol	Null	No	KEGG
estradiol-17beta	Cytosol	No	EHMN
estradiol-17beta	ER	No	EHMN
estradiol-17beta	Uncertain	No	EHMN
estradiol-17beta	Null	Yes, not consumed	HumanCyc
beta-estradiol	Cytosol	No	Reactome
Estradiol	ER	No	H. sapiens Recon 1
Estradiol	Cytosol	No	H. sapiens Recon 1
Estradiol	Null	No	KEGG
estradiol-17beta	Cytsosol	No	EHMN

Chapter IV, Table 2: Estrogen From Consensus and Conflict Database, , data taken from http://www.molgenis.org/c2cards/molgenis.do

Metabolite	Compartment	Dead-end metabolite?	Database
Estrone	Null	No	KEGG
Estrone	Uncertain	Yes, not produced	EHMN
Estrone	Cytosol	No	EHMN
Estrone	ER	No	EHMN
Estrone	Golgi	No	EHMN
Estrone	Lysosome	No	EHMN
Estrone	Null	No	HumanCyc
Estrone	Cytosol	No	Reactome
Estrone	ER	No	H. sapiens Recon 1
Estrone	Cytosol	No	H. sapiens Recon 1

Chapter IV, Table 3: Estrone, From Consensus and Conflict Database, data taken from http://www.molgenis.org/c2cards/molgenis.do

Database	Reaction	EC #	GENE	Pathway
EHMN	3'-phosphoadenylyl sulfate[uncertain] + estrone[uncertain] --> adenosine 3',5'-bisphosphate[uncertain] + estrone 3-sulfate[uncertain]	2.8.2.4	SULT1E1	Androgen and estrogen biosynthesis and metabolism
H. sapiens Recon 1	3'-phosphoadenylyl sulfate[c] + estrone[c] --> adenosine 3',5'-bisphosphate[c] + estrone 3-sulfate[c] + h+[c]	2.8.2.4	SULT1A1 or SULT1E1	Steroid Metabolism
HumanCyc	estrone + phosphoadenosine-5'-phosphosulfate |==| adenosine 3',5'-bisphosphate + estrone-sulfate + h+ [c]	2.8.2.4	SULT1E1	pathway unknown
KEGG	3'-phosphoadenylyl sulfate + estrone --> adenosine 3',5'-bisphosphate + estrone 3-sulfate	2.8.2.4	SULT1E1	Sulfur metabolism
KEGG	3'-phosphoadenylyl sulfate + estrone -- adenosine 3',5'-bisphosphate + estrone 3-sulfate	2.8.2.4	SULT1E1	Steroid hormone biosynthesis
Reactome	estrone[c] + paps[c] --> estrone 3-sulfate[c] + pap[c]	2.8.2.4	SULT1E1 or SULT2A1	Cytosolic sulfonation of small molecules

Chapter IV, Table 4: Reaction of estrone -> estrogen-sulfate, From Consensus and Conflict Database, data taken from http://www.molgenis.org/c2cards/molgenis.do

Lastly, none of the existing available networks or databases offers information about the tissue-specificity of metabolites. Many toxicants, including endocrine disruptors, exhibit tissue specific toxicity, and a cell or tissue-specific metabolic network might significantly help in establishing accurate compound identification.

A further complication is introduced by the cell system. The data generated for this study was from the Mapping the Human Toxome project (Bouhifd et al., 2014) (http://humantoxome.com), and one challenge is the variability of the system

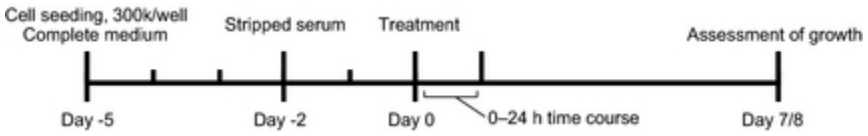
that compounds the variability of high-content results. The MCF-7 cell line (a human breast cancer epithelial derived tissue line) was originally selected as it is a fairly well-studied *in vitro* system, and has multiple published datasets with genomic, epigenomic, transcriptomic, and proteomic data widely available. However, the many different experimental settings and published results can vary significantly, even if looking at the same treatment (e.g. 17β-estradiol)—owing to different analysis and normalization methods, different experimental aims, the high rate of false positives in high-throughput data—but also, the intrinsic biological variability of the system. Nonetheless, in microarray studies, while the specific composition of gene lists from different studies shows poor overlap at a gene-by-gene level, they often coincide at the pathway level (Beltrame et al., 2009). Therefore, one expectation for metabolomics studies was that inferential statistics—looking for reproducibility at the level of fold-change for individual metabolites—might not be of much use given the noise in the analytical method and the variability in the biological system, and that a pathway approach would likely be necessary as it should in theory be less sensitive to noise, although this is assumption is predicted on a well-elucidated pathways.

In summary, metabolomics lacks the large-scale, integrated databases that have been crucial to the analysis of transcriptomic and proteomic data: specifically, it lacks databases such as Entrez and Uniprot that have established an agreed upon naming scheme, a high-level of database curation such as that provided by the NCBI to ensure accuracy, robust web services to translate identifiers, and publicly available data sets (similar to the ones provided by GEO) to allow benchmarking of

data analysis methodologies, since, much like the early years of microarrays, there are still no established methods to interpret data (Griffin, 2006). In summary, while metabolomics can theoretically measure everything, this is also a pitfall, especially if everything cannot be accurately labeled.

2. Materials and Methods

2.1 Data: This analysis is based on three different sets of experiments that were performed on MCF-7 cells. All studies used cells with a limited passage number (and in studies which focused on inter- or intra-laboratory reproducibility, passage number was harmonized as far as possible between the experiments). All cells were serum starved before dosing and charcoal-stripped media was used (See Figure 1 for Experiment Setup).



Chapter VI, Figure 1: Experiment Setup. Figure from (Bouhifd et al., 2014)

Study 1: A time and dose-response curve with 4 biological replicates at 0, 2, 4, 8, and 24 hours, and doses of 0, .001, .01, .1, and 1 nm estrogen (17-beta-estradiol) at each time point. Metabolites were extracted via methanol and HPLC-MS was performed on an Agilent QTOF 6520 with a range of 100-1100 M/z and metabolite identification performed via a recursive algorithm using Agilent Masshunter b.05. Masshunter identifies metabolites based on the amount of isotopes and ions detected and establishes a quality score that requires 70 out of 100 points to establish a true metabolite. Metabolite identification was based on

both an original library based largely on KEGG metabolites annotated as human ("KEGG Library"), and an additional, custom-built library based on HMDB and SMPD to eliminate incorrectly identified metabolites ("HMDB Library").

Studies 2 and 3: Same time, dose-response curve, and biological replicates as Study 1, and performed at Johns Hopkins and Brown University to study inter-laboratory variability. Data analysis was done using Agilent Masshunter workflow and the KEGG Library.

Studies 4 and 5: Identical studies were both performed at Johns Hopkins two weeks apart using 0 and 1 nm estrogen at 4 and 24 hours to compare intra-laboratory reproducibility, analyzed with Agilent Masshunter workflow. The same data were later analyzed using XCMS (Gowda et al., 2014) for comparison of log-fold changes between 0 and 1 nm estrogen at both time points and metabolite identification was performed based on Mummichog (Li et al., 2013).

2.2 Data Analysis: All preprocessing was done using Metaboanalyst (Xia, Mandal, Sinelnikov, Broadhurst, & Wishart, 2012); metabolites with more than 50 percent missing values were removed and missing data was imputed using k-nearest neighbors. Data was filtered using interquartile range and normalized via log transformation and Pareto scaling. MSEA (Metabolite Set Enrichment Analysis) was performed using Metaboanalyst (Xia & Wishart, 2010).

In addition to Metaboanalyst/MSEA, the following were used for metabolite enrichment analysis: IMPaLA (Kamburov, Cavill, Ebbels, Herwig, & Keun, 2011), which analyzes based on pathways from SMPD, KEGG, and REACTOME, and MBRole,

which offers the additional option of analyzing for chemical class (Chagoyen & Pazos, 2011).

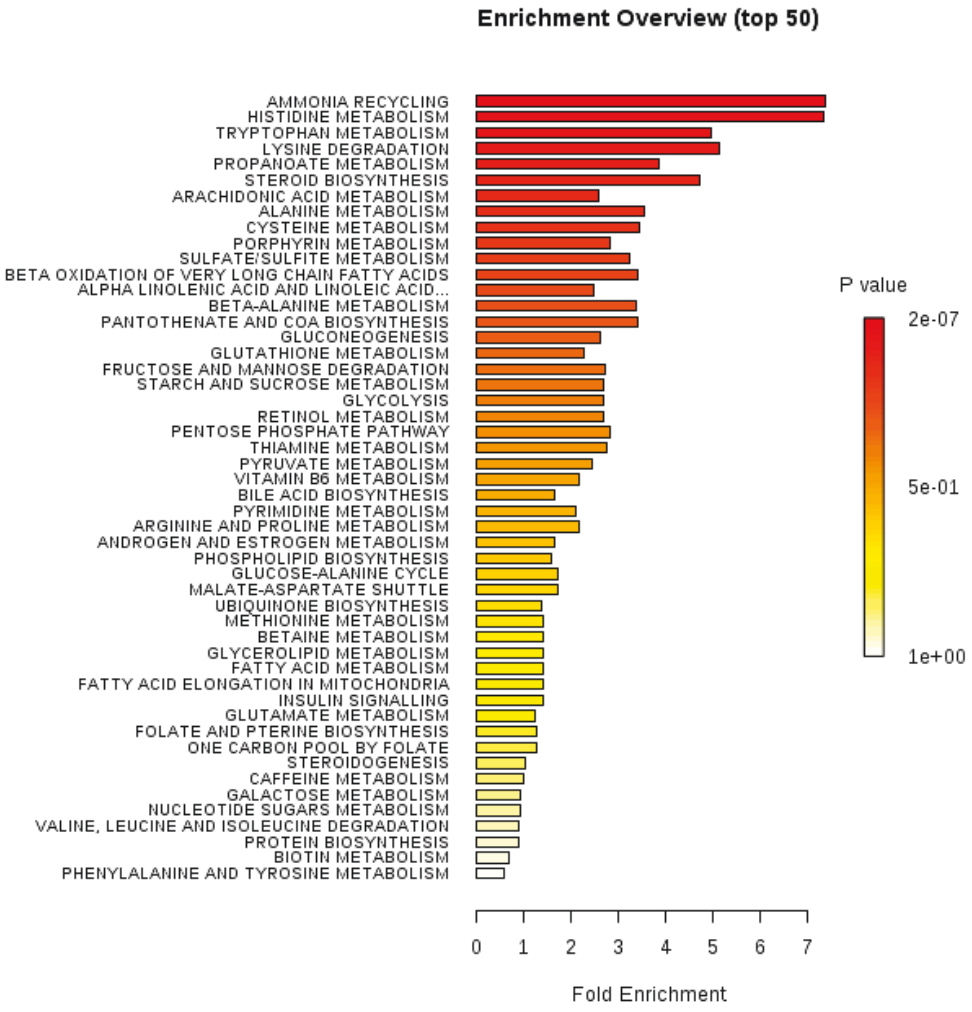
WGCNA: A weighted correlation network was built using the WGCNA package (Langfelder & Horvath, 2008) based on an unsigned Spearman rank correlation transformed via a weighting function with the weight set to 7 (for details, see Chapter Five, Materials and Methods). Clustering was done based on the Dynamic Tree Cut algorithm (Langfelder et al., 2008) with deepsplit set to 3 (based on a 1-4 scale of how sensitive the algorithm is to selecting modules; 3 was chosen to give fairly small, precise modules), a minimum module size of 10, and modules with a distance of less than .25 merged.

3. Results and Discussion

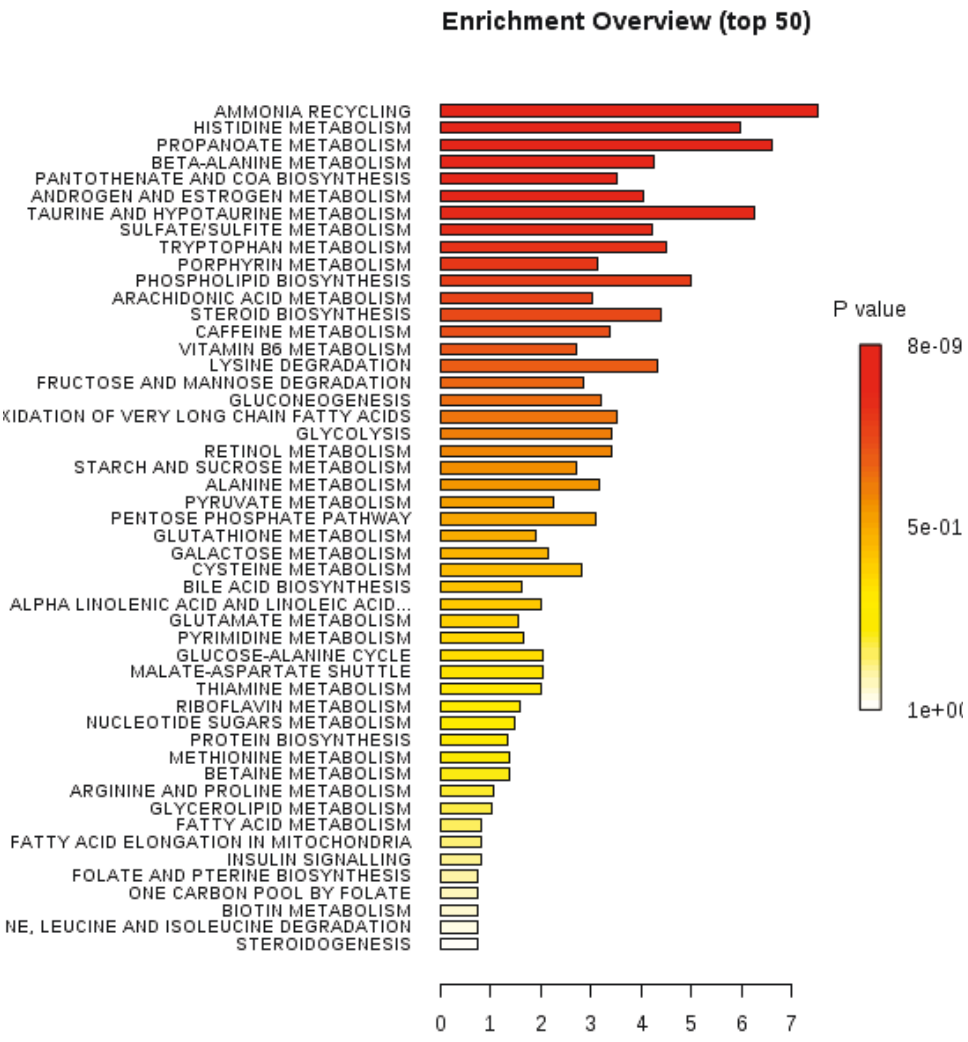
3.1 QEA COMPARED TO ORA: The Metaboanalyst platform for Metabolite Set Enrichment Analysis was used for both over-representation analysis (ORA) and Quantitative Set Enrichment Analysis (QEA) (Xia & Wishart, 2010). ORA analyzes whether a given metabolite set identified as statistically significant from an experiment is over-represented in a given pathway compared to an expected value based on the size of the pathway and assuming a hypergeometric distribution, after correcting for a false discovery rate (FDR). The other approach, QEA, is based on the "global test" algorithm (Goeman, van de Geer, de Kort, & van Houwelingen, 2004), commonly used for microarray experiments, to perform enrichment analysis directly from normalized concentration data. "Global test" was originally created to examine associations between gene sets and clinical outcomes, but it has been used extensively for microarray data and adapted for multiclass and continuous

phenotypes. It uses a generalized linear model to compute a 'Q-stat' for each gene set, using the average of the squared covariance between the expression level of the genes (or, in the case of metabolomics, concentration) and the label. MSEA includes appropriate methods to adjust for the multiple testing problems that occur during enrichment analysis (e.g. Benjamini and Hochberg FDR).

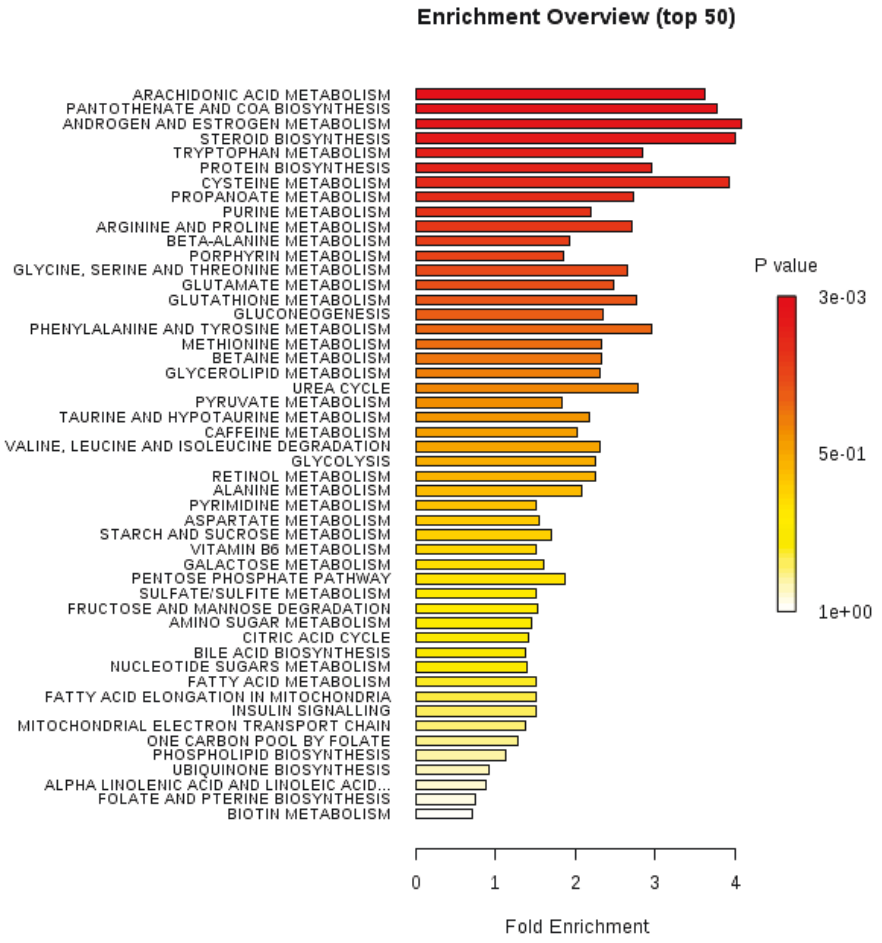
For ORA, each dose-response curve at a given time point was analyzed via one-way ANOVA and all metabolites identified as significant (*p*-value less than .05) were used for ORA. The 8-hour dose–response curve had the most significant number of pathways.



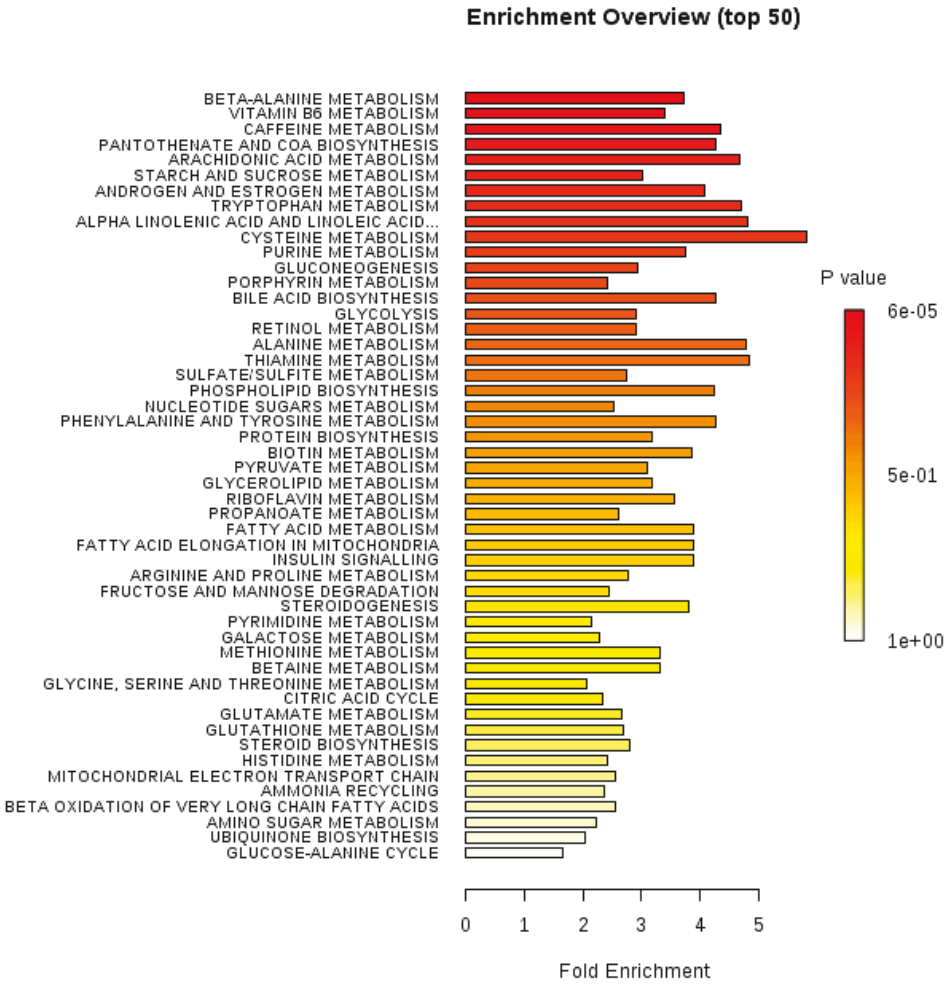
Chapter IV, Figure 2: ORA Experiment 1, 8-Hour Dose–Response Curve.
Legend, Figure 2: Metabolites identified with one-way ANOVA in the 8-hour dose–response curve as significant (adjusted *p*-value of less than .05) analyzed by Over-Representation Analysis.



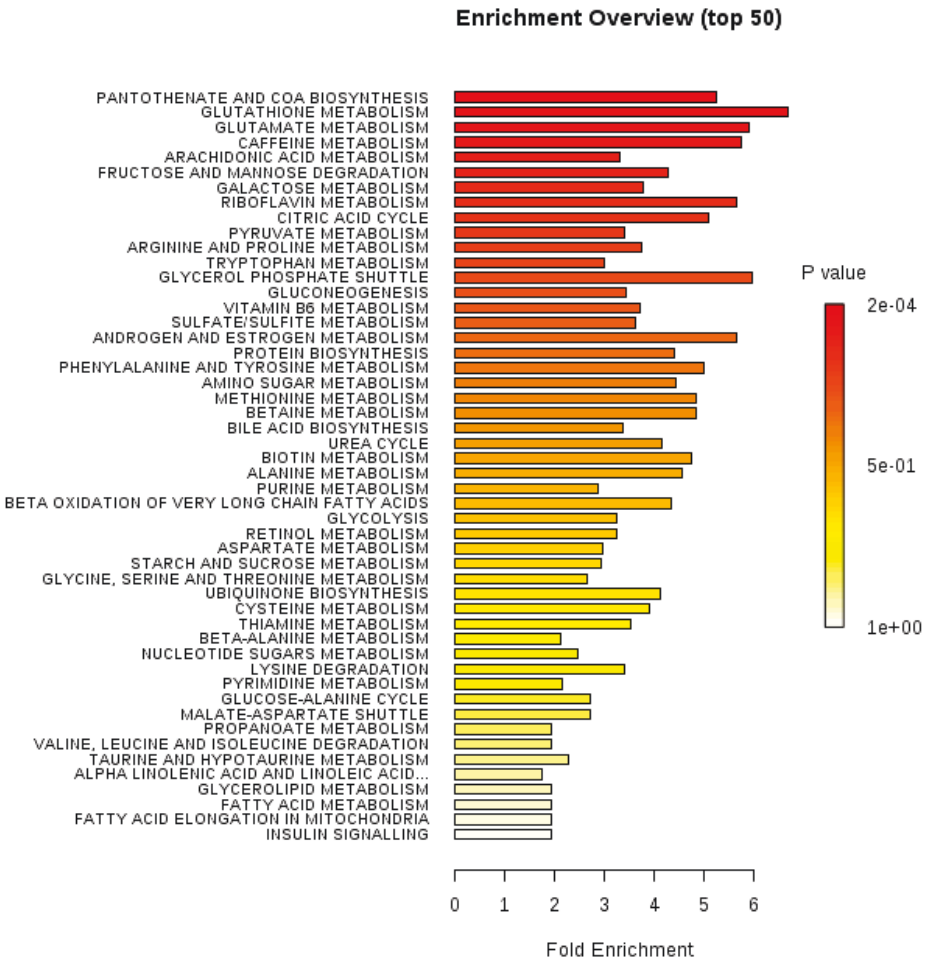
Chapter IV, Figure 3: QEA, 8-Hour Dose–Response Curve
Legend, Figure 3: Comparing ORA to QEA, the top two pathways (Ammonia Recycling and Histidine Metabolism) are the same, but the *p*-values for QEA are significantly higher than for ORA. Furthermore, QEA identified several pathways as significant that were missing from ORA.



Chapter IV, Figure 4: Experiment 1, QEA 8-Hour Dose–Response Curve



Chapter IV, Figure 5: Experiment 1, QEA, 4-Hour Dose–Response Curve



Chapter IV, Figure 6: Experiment 1, QEA 24-Hour Dose–Response Curve

	Total Cmpd	Hits	Statistic Q	Expected Q	Raw p	Holm p	FDR
PANTOTHENATE AND COA BIOSYNTHESIS	10	3	24.815	4.7476	0.000004292	0.00024465	0.00013773
GLUTATHIONE METABOLISM	10	1	31.591	4.7468	0.000004956	0.00027753	0.00013773
GLUTAMATE METABOLISM	18	2	28.022	4.748	0.000007249	0.0003987	0.00013773
CAFFEINE METABOLISM	12	4	27.266	4.7513	0.000020092	0.001085	0.00028632
ARACHIDONIC ACID METABOLISM	37	3	15.648	4.7423	0.000062942	0.0033359	0.00054516
FRUCTOSE AND MANNOSE DEGRADATION	18	2	20.337	4.7451	0.00006594	0.0034289	0.00054516

GALACTOSE METABOLISM	25	3	17.884	4.747	0.00007521	0.0038357	0.00054516
RIBOFLAVIN METABOLISM	9	2	26.779	4.7497	0.000076513	0.0038357	0.00054516
CITRIC ACID CYCLE	23	3	24.227	4.7516	0.000088921	0.0043571	0.00056316
PYRUVATE METABOLISM	20	3	16.02	4.7251	0.00011023	0.005291	0.00057284
ARGININE AND PROLINE METABOLISM	26	3	17.719	4.7436	0.00011055	0.005291	0.00057284
TRYPTOPHAN METABOLISM	34	4	14.174	4.7234	0.00012847	0.0059097	0.00061024
GLYCEROL PHOSPHATE SHUTTLE	8	1	28.356	4.7514	0.00014766	0.0066449	0.00064745
GLUCONEOGENESIS	27	2	16.178	4.7221	0.0001619	0.0071235	0.00065915
VITAMIN B6 METABOLISM	10	4	17.59	4.7417	0.00018373	0.0079004	0.00067927
SULFATE/SULFITE METABOLISM	7	3	17.202	4.7436	0.00019067	0.0080082	0.00067927
ANDROGEN AND ESTROGEN METABOLISM	17	1	26.767	4.7501	0.00024094	0.0098787	0.00080787
PROTEIN BIOSYNTHESIS	19	5	20.811	4.745	0.00026683	0.010673	0.00084496
PHENYLALANINE AND TYROSINE METABOLISM	13	1	23.606	4.7424	0.00030036	0.011714	0.00090108
AMINO SUGAR METABOLISM	15	2	21.09	4.7456	0.00040019	0.015207	0.0010645
METHIONINE METABOLISM	24	1	22.913	4.7419	0.00041087	0.015207	0.0010645
BETAINE METABOLISM	10	1	22.913	4.7419	0.00041087	0.015207	0.0010645
BILE ACID BIOSYNTHESIS	49	5	16.063	4.7502	0.00061413	0.021494	0.0014611
UREA CYCLE	20	1	19.698	4.7363	0.00061522	0.021494	0.0014611
BIOTIN METABOLISM	4	1	22.506	4.745	0.00074133	0.024464	0.0016902
ALANINE METABOLISM	6	1	21.589	4.7404	0.00090323	0.028903	0.0019802
PURINE METABOLISM	45	5	13.677	4.7404	0.0010959	0.033973	0.0023136
BETA OXIDATION OF VERY LONG CHAIN FATTY ACIDS	14	1	20.626	4.7406	0.0011527	0.034582	0.0023466
GLYCOLYSIS	21	1	15.358	4.723	0.001318	0.038221	0.0025042

RETINOL METABOLISM	18	1	15.358	4.723	0.001318	0.038221	0.0025042
ASPARTATE METABOLISM	12	3	14.097	4.7467	0.0015295	0.041297	0.0028124
STARCH AND SUCROSE METABOLISM	14	2	13.909	4.733	0.0020946	0.05446	0.003731
GLYCINE, SERINE AND THREONINE METABOLISM	26	3	12.683	4.7479	0.0025548	0.063869	0.0044128
UBIQUINONE BIOSYNTHESIS	10	1	19.619	4.7497	0.0035558	0.085338	0.0059611
CYSTEINE METABOLISM	8	1	18.486	4.7441	0.0040482	0.093108	0.0065927
THIAMINE METABOLISM	4	1	16.784	4.7428	0.0073103	0.16083	0.011575
BETA-ALANINE METABOLISM	13	5	10.116	4.7464	0.0081177	0.17047	0.012506
NUCLEOTIDE SUGARS METABOLISM	9	3	11.733	4.7395	0.0088294	0.17659	0.013244
LYSINE DEGRADATION	13	1	16.069	4.7422	0.0096781	0.18388	0.014145
PYRIMIDINE METABOLISM	36	7	10.212	4.7462	0.010089	0.18388	0.014376
GLUCOSE-ALANINE CYCLE	12	1	12.951	4.742	0.031162	0.52975	0.042291
MALATE-ASPARTATE SHUTTLE	8	1	12.951	4.742	0.031162	0.52975	0.042291

Chapter IV, Table 5: Experiment 1, QEA 24-Hour Dose–Response Curve.

On the one hand, QEA appeared to have identified pathways at each time point, indicating the presence of a dose response, but the shifts in pathways over time do not tell a consistent biological story. More worrisomely, as can be seen in Table 5, many of the most significant pathways had relatively few metabolites mapped to the pathway, and in some cases only one—meaning that a single misidentified metabolite could be significantly skewing the results, and the pathway with the largest hits (caffeine metabolism, with four metabolites) represents a

pathway not informative for this study and probably contains misidentified metabolites.

ORA and QEA have been utilized in metabolomics without necessarily taking into consideration some of the problems of metabolomic data. While microarrays present a data set where all the discrete features are labeled unambiguously at the outset and all discrete features have an assigned value, in the case of metabolomic data, due to the nature of the technology, no experiment can possibly identify a complete set of metabolites; of those identified as discrete, not all can be assigned an identity precisely. Of those assigned an identity, relatively few are annotated with pathway information. Moreover, the missing metabolites may not be missing at random, but may reflect a chemical class. As a consequence, at the same time ORA and QEA have a substantial loss of information (because of the non-mapped pathways), they also have the potential to be inaccurate due to a small number of misidentified metabolites.

Additionally, all annotation-based statistical tests are predicated on an accurate assumption of the "background"—that is, the total number of pathways and metabolites possible. Generally, the assumed background is the number of total metabolites in pathways; restricting it to the background of all metabolites identified in the experiment (945) diminishes or eliminates the statistical significance. It is difficult to know which contributes more to the error—non-random, missing data or an incorrect assumption about background size.

Furthermore, analyzing the pathways over-represented in the total number of metabolites identified in the experiment pointed to some other causes for

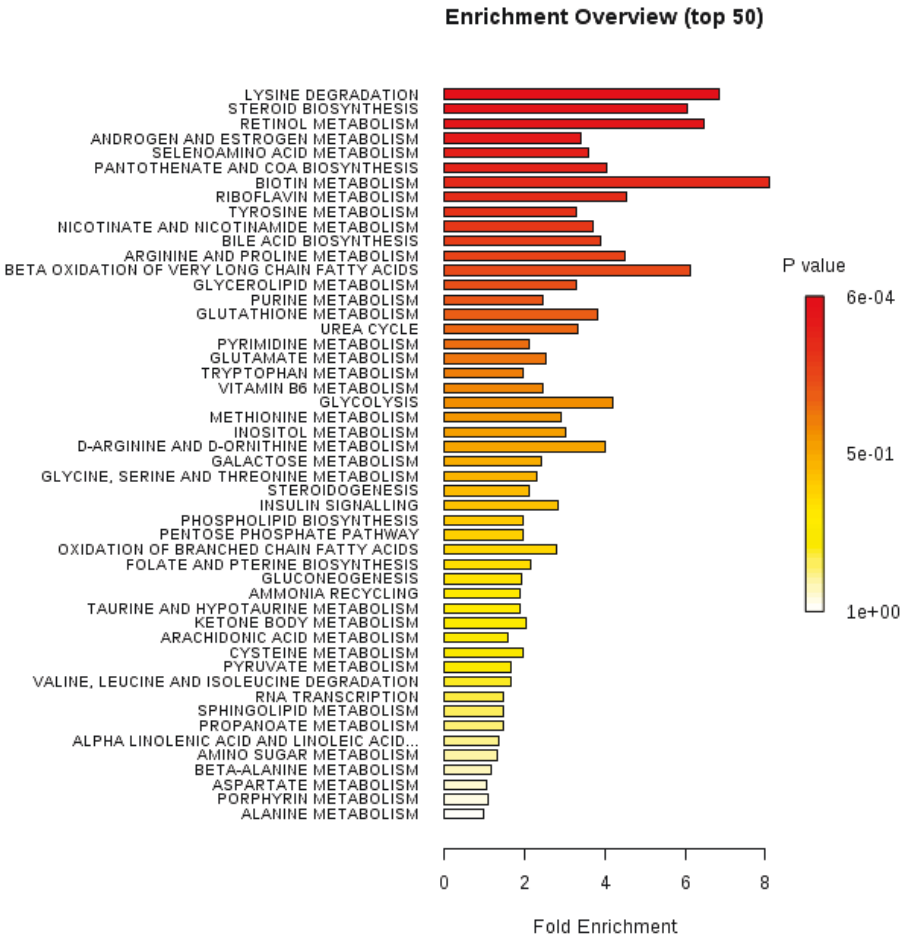
concern. One, when analyzed via IMPaLA (a platform which can combined multiple metabolite pathway databases for ORA), there appeared to be a significant pathway bias in the identification process, and of the top three pathways identified as over-represented, two pathways ("Butirosin and neomycin biosynthesis" and "Drug metabolism") were not pathways that should have been seen in this sample and likely indicate misidentified metabolites. Furthermore, of the 945 metabolites identified, only 468 were mapped via IMPaLA to pathways. A small sample taken at random of the remaining 477 non-mapped metabolites were manually checked for biological significance, and the non-mapped metabolites in the sample consisted of either plant of bacterial metabolites that were not likely candidates to be present in MCF-7 cells. A manual inspection of the 468 mapped metabolites confirmed the results of the over-representation analysis—there were several compounds (e.g. chlorophyll) that were misidentified metabolites; based on a small subsample, the error rate was estimated to be at least 10 percent. Notably, of the library used to identify the metabolites, of the 4,128, only 2,573 were mapped to pathways via IMPaLA. Given the likely error rate that this indicates in metabolite identification, this makes the results of QEA highly suspect.

Pathway	**Source**	**# Metabolites**	**Pathway (Background)**	**p-value**	**q-value**
Transport of vitamins, nucleosides, and related molecules	Reactome	25	63 (64)	1.52E-07	0.000525
Butirosin and neomycin biosynthesis - Homo sapiens (human)	KEGG	15	29 (29)	8.57E-07	0.00148
Drug metabolism - cytochrome	KEGG	29	88 (88)	1.59E-06	0.00183

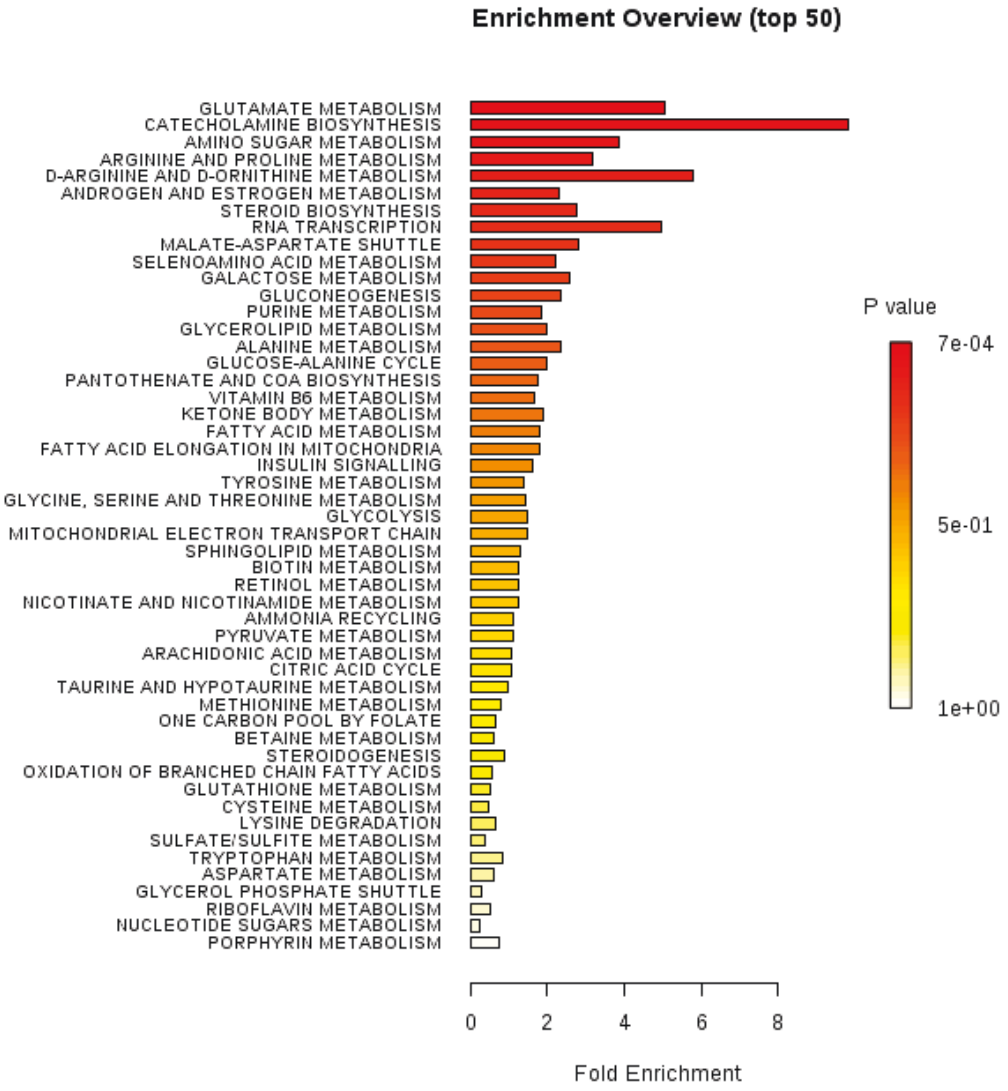
P450 - Homo sapiens (human)					

Chapter IV, Table 6: Pathways Returned by IMPaLA for All Metabolites Identified in Sample

In order to investigate the depedency of the results on the library and improve the number of human specific results, the library used to identify the metabolites, we created an alternative library based on all metabolites marked as endogenous in HMDB—2,415 metabolites total. Some obvious errors were removed (recognizable drug and pesticide names), and the library was checked for completeness against the SMPD database; any non-disease related pathways with more than 15 metabolites but only 50 percent pathway coverage were manually added to ensure the library had adequate coverage of all pathways.



Chapter IV, Figure 7: Experiment 1 – HMDB Library, 8-Hour Dose–Response Curve, QEA



Chapter IV, Figure 8: Experiment 1 – HMD Library 24-Hour Dose–Response Curve, QEA

This methodology generated different pathways than the previous library, indicating if nothing else that the results were substantially impacted by the library used for identification. The improved library also eliminated caffeine metabolism as a significant pathway, and androgen and steroid estrogen metabolism appear as pathways at both time points, indicating, if not improved accuracy, at least

improved plausibility (see Figures 7 and 8). In the pathways identified as significant, all but one pathway (biotin, which has only four metabolites for the entire pathway) had more than two metabolites in the pathways. (See Table 7). The results were therefore far less sensitive to a single misidentification.

	Total Cmpd	Hits	Statistic Q	Expected Q	Raw p	Holm p	FDR
LYSINE DEGRADATION	13	2	40.375	5.8824	9.32E-06	0.00058718	0.00031832
STEROID BIOSYNTHESIS	31	5	35.658	5.8824	1.01E-05	0.00062654	0.00031832
RETINOL METABOLISM	18	3	38.127	5.8824	8.16E-05	0.0049748	0.0017126
ANDROGEN AND ESTROGEN METABOLISM	17	6	20.182	5.8824	0.00064549	0.038729	0.010166
SELENOAMINO ACID METABOLISM	15	7	21.137	5.8824	0.00086838	0.051235	0.010942
PANTOTHENATE AND COA BIOSYNTHESIS	10	5	23.82	5.8824	0.0011924	0.069162	0.012521
BIOTIN METABOLISM	4	1	47.505	5.8824	0.0015558	0.088682	0.014002
RIBOFLAVIN METABOLISM	9	3	26.836	5.8824	0.0018759	0.10505	0.014773
TYROSINE METABOLISM	38	7	19.31	5.8824	0.0031746	0.1746	0.020541
NICOTINATE AND NICOTINAMIDE METABOLISM	13	6	21.818	5.8824	0.0032889	0.1776	0.020541
BILE ACID BIOSYNTHESIS	49	15	23.017	5.8824	0.0035864	0.19008	0.020541
ARGININE AND PROLINE METABOLISM	26	4	26.519	5.8824	0.0042354	0.22024	0.022236
BETA OXIDATION OF VERY LONG CHAIN FATTY ACIDS	14	1	36.141	5.8824	0.0083187	0.42425	0.040314
GLYCEROLIPID METABOLISM	13	5	19.304	5.8824	0.0093868	0.46934	0.04224
PURINE METABOLISM	45	12	14.601	5.8824	0.010473	0.51319	0.043988
GLUTATHIONE METABOLISM	10	3	22.619	5.8824	0.011318	0.54328	0.044566
UREA CYCLE	20	3	19.682	5.8824	0.01216	0.57152	0.045064
PYRIMIDINE METABOLISM	36	14	12.577	5.8824	0.012887	0.59279	0.045104

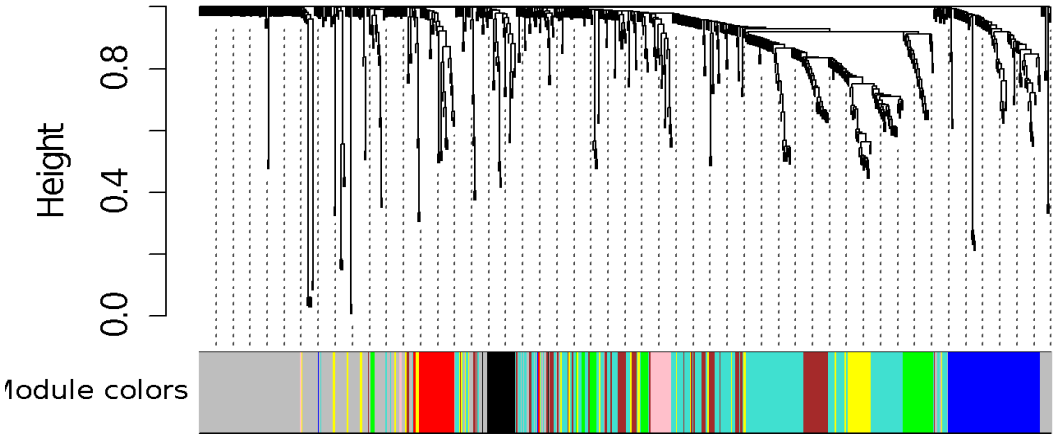
Chapter IV, Table 7: Experiment 1, 8-Hour Time Point, HMDB Library QEA

On the other hand, of the 1,025 metabolites identified in the sample, only 484 were mapped to pathways in IMPaLA—so while the improved library may have

increased the likelihood of identifying valid metabolites, it did not substantially increase the ability to use pathway-based annotations for data analysis. Furthermore, while the second library likely eliminated some misidentifications, there is no certainty that it produced more accurate metabolite identification, as that would require the verification of metabolite identity with another technology (such as MS/MS, which allows a more precise identification based on fragmentation).

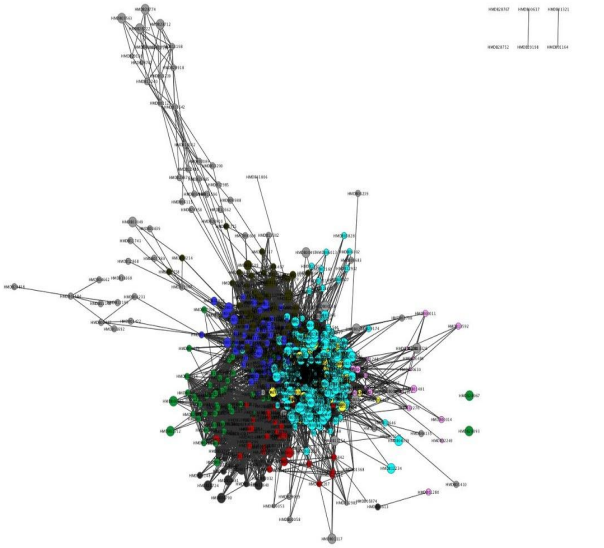
3.2 Correlation Approach: Because of the concern about annotation-dependent approaches, an unsupervised, non-annotation based approach was used to cluster the metabolites identified in experiment 1 via the HMDB library. The WGCNA package appeared to cluster the metabolites into distinct modules (Figure 9). However, the network (based on topological overlap metric) was much more dense than is typically seen with microarray data, which typically produces a network with more distinct, non-overlapping modules. This likely reflects the difficulty of clustering metabolomics data compared to microarray data as metabolites are intrinsically more correlated (Figure 10). Five of the modules were correlated with time or dose with a *p*-value of less than .01, once again indicating the presence of a dose-response. However, the module with the highest correlation for time and dose (the Red module) was not significantly over-represented for metabolites in a pathway; in fact, only two of the 20 metabolites could be mapped to a pathway, and all modules had fewer than 50 percent of the metabolites mapped to a pathway. Despite this, the Blue and Black modules did have a statistically significant over-representation, indicating that the method does appear to group

similar metabolites together. However, attempts to characterize the identity of the unmapped metabolites were not successful, as many of them were dipeptides or seemed unlikely candidates, and clustering the modules by chemical similarity did not indicate that similar chemicals were clustering together. For this application, there did not appear to be enough information for the modules to be characterized, the identity of the members verified, or the biological significance understood.



Chapter IV, Figure 9: Dendrogram based on WGCNA

Legend, Figure 9: WGCNA appeared to cluster the metabolites into distinct branches; modules indicated by color.



Chapter IV, Figure 10: Metabolites clustered by Topological Overlap Metric and colored by module.
Legend, Figure 10: Metabolites formed a fairly dense network; metabolites not assigned to a module are in grey.

Module	**R**	**p-value**	**R**	**p-value**
Red	0.377	0.0000936	0.44	0.000357
Black	0.006	0.947	0.23	0.0153
Blue	0.21	0.031	0.19	0.0455
Brown	0.18	0.006	0.24	0.00148
Pink	0.23	0.00186	0.05	0.549
Green	0.1	0.281	0.07	0.475
Turquoise	0.005	0.964	0.23	0.0177

Chapter IV, Table 8: Modules Correlated with Time and Dose

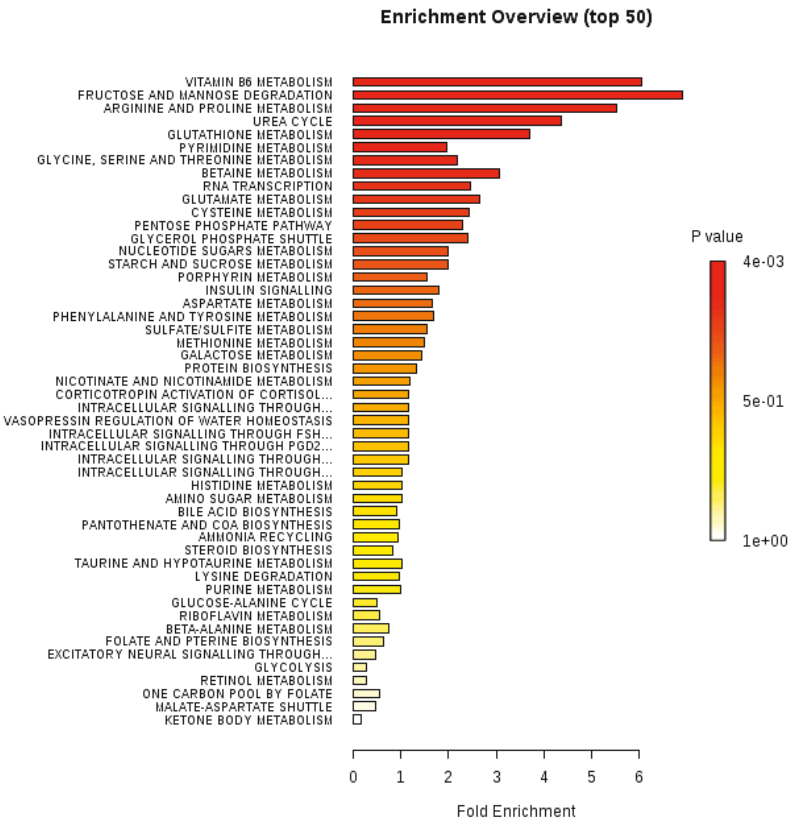
Module	**p-value**	**adjusted p-value**
Blue		
Bile Acid Biosynthesis	2.24E-06	3.39E-05
Steroidogenesis	2.24E-03	1.05E-02

Arachidonic Acid Metabolism	9.41E-04	6.59E-03
Black		
Riboflavin	2.23E-03	1.12E-02
Red		
Tauraine and Hypotaurine Metabolism	1.56E-02	1.87E-01
Brown		
Vitamin B6 Metabolism	4.35E-02	2.60E-01
Arginine and Proline Metabolism	4.35E-03	2.60E-01
Glutathione Metabolism	3.25E-02	2.60E-01

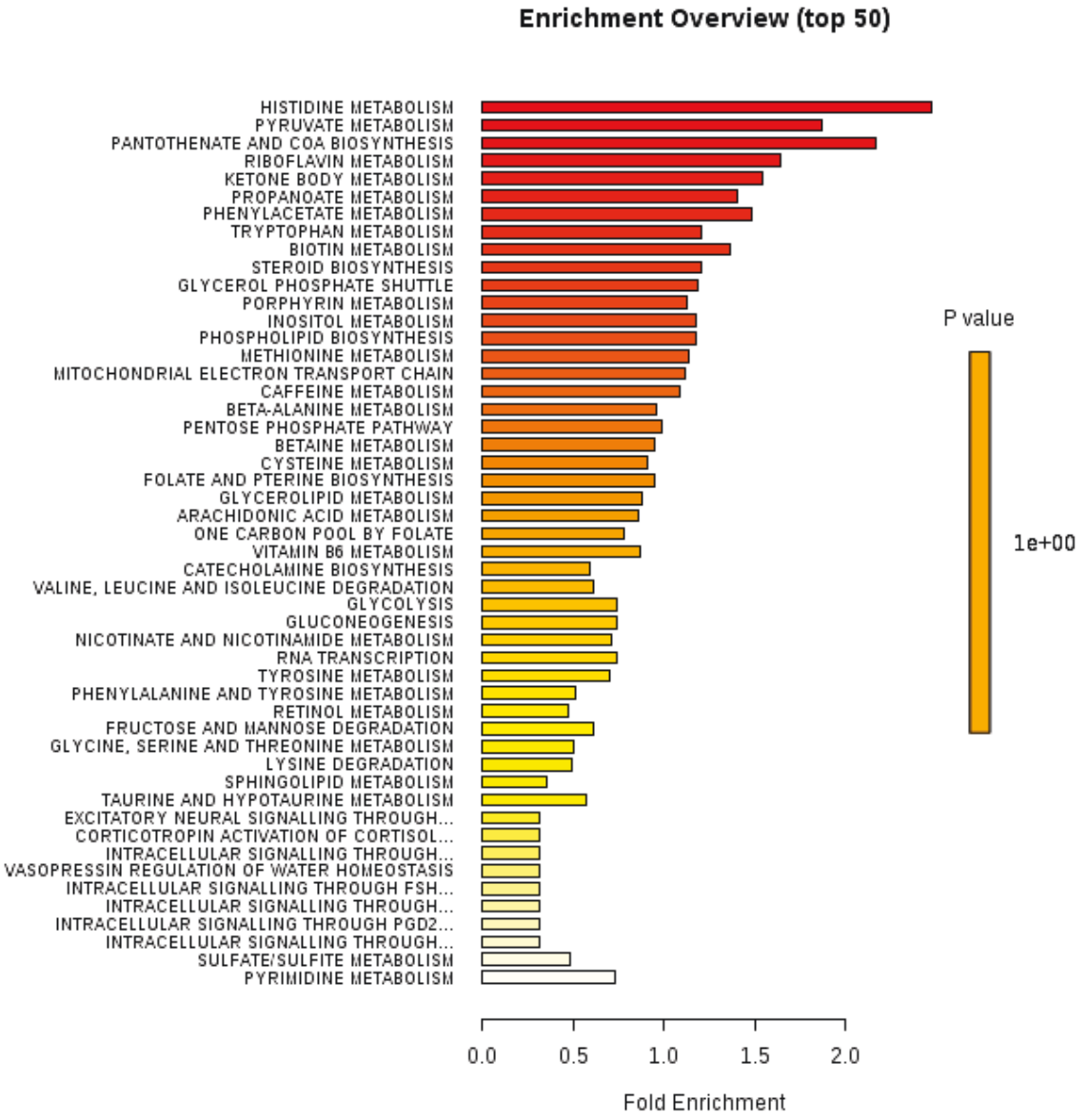
Chapter IV, Table 9: Modules annotated via MBRole

3.3 Variability and Reproducibility: Some insight into the possible source of difficulty came to light when analyzing two studies that had been done under identical conditions two weeks apart in the same laboratory (Experiments 4 and 5) and two studies done to examine inter-laboratory reproducibility (Experiments 2 and 3).

The studies initially showed no overlap when analyzed by inferential statistics at each time-point's dose response when analyzed by ANOVA. Furthermore, there was no overlap when QEA was performed on the dose–response curves of Experiments 2 and 3 (see Figures 11 and 12).



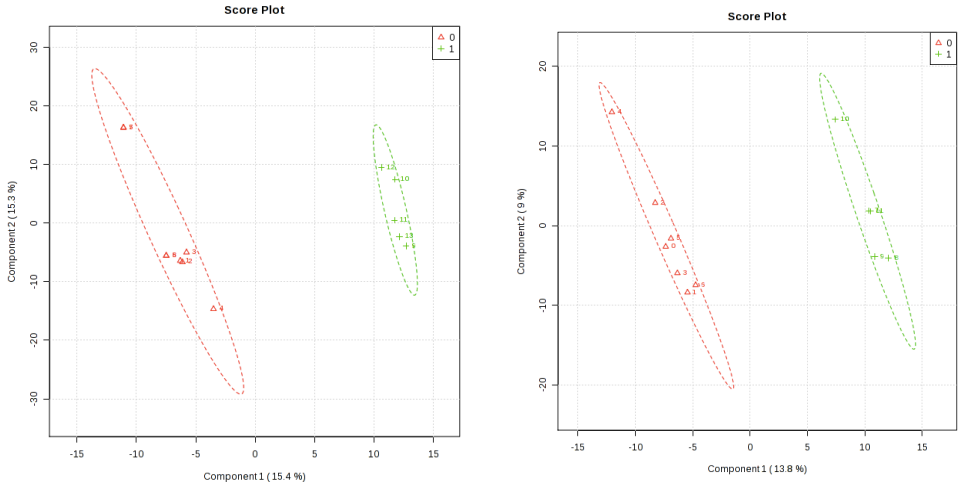
Chapter IV, Figure 11: Experiment 2 – QEA 8-Hour Time-Point



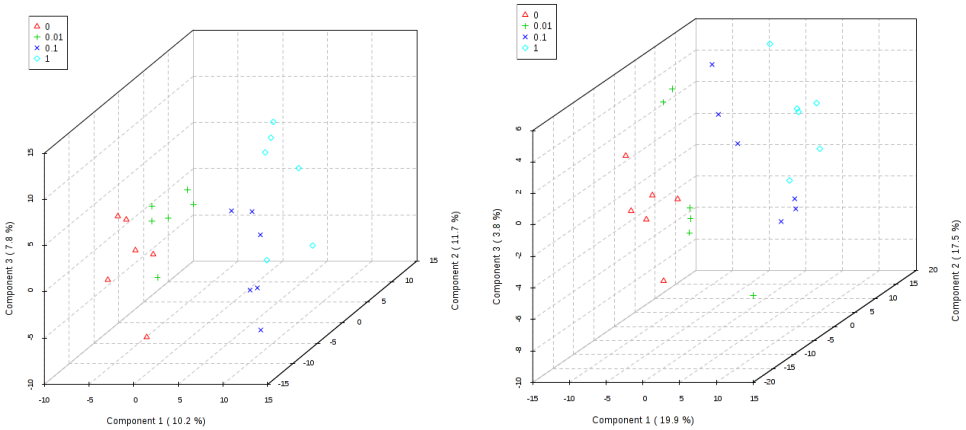
Chapter IV, Figure 12: Experiment 3, QEA 8-Hour Time-Point

To examine them further, Metaboanalyst's PLS-DA (Partial Least Squares Discriminant Analysis) was used to analyze the dose–response relationship. PLS-DA is similar to Principal Component Analysis, but unlike PCA, PLS-DA uses the class labels to maximize separation. If X is the input data from the samples and Y contains the class labels, then the PLS-DA algorithm will maximize the covariance between

the X variables and the Y labels.



Chapter VI, Figure 13: Experiments 4 and 5 (0 and 1 nm estrogen, 24 Hours) PLS, 2-D plot



Chapter IV, Figure 14: Experiments 2 and 3 (Dose–Response Curve, 24 Hours), 3-D plot

In both cases, PLS appeared to offer some separation of the samples. PLS models can be affected by systematic variation between two samples, and one caveat of PLS is that it is prone to over-fitting; the dose–response separation observed cannot be taken *ipso facto* as evidence of a robust dose–response effect.

The key finding, however, was when examining the variable importance—the metabolites that contributed to dose separation. In each case, there was minimal overlap of the top 100 metabolites that contributed to dose separation: only 35 metabolites were present in both samples; several of the top metabolites that contributed to the dose-separation were plant metabolites or drugs. In other words, despite being the same cells with the same treatment, even a fairly sensitive exploratory data technique could only find minimal indications of similarity between the two experiments.

This suggests that the failure to find consistent pathways and overlap based on inferential statistics was not due completely to biological variability, but instead that there were consistent problems with accurate metabolite identification that were skewing the results. There are several possible explanations for this, including, but not limited to: (1) peak identification may have been incorrect or incomplete, (2) ions may have been incorrectly identified, and (3) the metabolite may be incorrectly identified because of compounds with similar weight. While all three sources of error will potentially cause problems for data analysis, the first two will result in an incorrect molecular weight and incorrect concentration, while the third will result in a correct molecular weight and concentration, but simply an incorrect label. Unfortunately, attempts to match the molecular weights of some of the non-endogenous or implausible metabolites with more likely candidates in the Metlin database were not successful, and this would indicate that the problems were more than mere labels.

One possible cause of the misidentification was the recursive algorithm used in the Agilent workflow to identify compounds. To explore this possibility, we tried an alternative data analysis using XCMS, which does not identify *metabolites* but extracts a fold-change difference between *features* (m/z) for a two-class comparison—avoiding the recursive algorithm used by Masshunter. This allowed us to attempt an alternative means of metabolite identification, Mummichog (Li et al., 2013). While most metabolomics platforms attempt to assign metabolite identification based on formula, Mummichog maps all possible metabolites to the feature list and then attempts to deduce the likely correct assignments by looking for plausible biological pathways. In other words, it tries to leverage the intrinsic inter-connectedness of metabolites to correctly assign metabolite identification. Pathway significance is determined by a permutation test based on all possible pathways. In this analysis, Experiment 4 and Experiment 5 each had significant pathways identified as being differentially regulated by estradiol treatment, and there was overlap for 12 of the pathways (see Table 12).

Pathways	overlap_size	pathway_size	*p*-value	adjusted *p*-value
Glycine, serine, alanine and threonine metabolism	28	58	0.00034	0.01013
Urea cycle/amino group metabolism	24	53	0.00273	0.01033
Caffeine metabolism	8	11	0.00197	0.01052

Tryptophan metabolism	29	74	0.01293	0.01098
Ascorbate (Vitamin C) and Aldarate Metabolism	12	23	0.0085	0.01103
Histidine metabolism	13	26	0.00976	0.0111
Beta-Alanine metabolism	9	17	0.01992	0.01247
Valine, leucine and isoleucine degradation	17	41	0.02966	0.01248
Aspartate and asparagine metabolism	26	71	0.04403	0.01309
Butanoate metabolism	12	28	0.04923	0.01466
Pyruvate Metabolism	9	19	0.04459	0.01497
Pyrimidine metabolism	20	56	0.09121	0.01711
Tyrosine metabolism	32	97	0.10521	0.0173
Pentose phosphate pathway	14	37	0.09609	0.01858
Alanine and Aspartate Metabolism	10	25	0.10769	0.02131
Arginine and Proline Metabolism	15	42	0.13275	0.02206

Hexose phosphorylation	8	20	0.14328	0.02759
Nitrogen metabolism	3	4	0.06235	0.02943
Glycerophospholipid metabolism	15	46	0.23639	0.03557
Vitamin H (biotin) metabolism	3	5	0.12502	0.04508
Glutamate metabolism	5	12	0.20011	0.04566

Chapter IV, Table 10: Pathways identified as significant by Mummichog for Experiment 4, 24 Hours

Pathways	overlap_size	pathway_size	p-value	adjusted p-value
Pyruvate Metabolism	12	19	0.00026	0.00761
Valine, leucine and isoleucine degradation	18	37	0.0007	0.00764
Arginine and Proline Metabolism	18	40	0.00219	0.00776
Carnitine shuttle	12	23	0.00267	0.00788
Glycine, serine, alanine and threonine	21	57	0.01606	0.00869

metabolism				
Butanoate metabolism	12	28	0.01861	0.00926
Glycolysis and Gluconeogenesis	16	43	0.03071	0.00991
Ascorbate (Vitamin C) and Aldarate Metabolism	10	23	0.02776	0.0103
Tyrosine metabolism	30	96	0.04984	0.01062
Fructose and mannose metabolism	11	28	0.04653	0.01181
TCA cycle	10	26	0.06501	0.01387
Vitamin E metabolism	13	37	0.07577	0.01408
Beta-Alanine metabolism	7	16	0.06074	0.01494
Hexose phosphorylation	8	20	0.07682	0.01616
Phosphatidylinositol phosphate metabolism	11	32	0.11224	0.01865
Urea cycle/amino group metabolism	16	52	0.14569	0.0207
Glyoxylate and Dicarboxylate Metabolism	5	11	0.09381	0.02283
Aspartate and	20	69	0.17897	0.02369

asparagine metabolism				
Tryptophan metabolism	20	72	0.23987	0.03269
Galactose metabolism	12	40	0.21858	0.03377
Biopterin metabolism	6	17	0.19468	0.0399
Alanine and Aspartate Metabolism	8	25	0.22157	0.04019

Chapter IV, Table 11: Pathways Identified by Mummichog, Experiment 5, 24 Hours

Alanine and Aspartate Metabolism
Arginine and Proline Metabolism
Ascorbate (Vitamin C) and Aldarate Metabolism;
Aspartate and asparagine metabolism;
Beta-Alanine metabolism;
Butanoate metabolism
Glycine, serine, alanine and threonine metabolism
Hexose phosphorylation
Tryptophan metabolism
Tyrosine metabolism
Urea cycle/amino group metabolism
Valine, leucine and isoleucine degradation

Box 1: Common Pathways, Experiments 4 and 5, Mummichog

However, it is difficult to determine if the concordance between the two samples is dependent on the methodology, since Mummichog *presumes* there are pathways. This is a supervised approach, and it may simply be structuring the data rather than deriving a structure *from* the data. Although the common pathways between the two identical experiments are encouraging, and the pathways suggested are plausible, in the absence of authenticated standards, it is difficult to know if this approach is superior.

While the above outlined approach offers at least some reproducibility and plausibility, there are several drawbacks too: Mummichog cannot integrate positive and negative mode and is therefore restricted to analyzing a subset of the experimental results. Because of its methodology and the statistical stringency required when testing for significantly impacted pathways, it most likely misses smaller pathways and is disadvantaged when looking for pathways that may consist of many compounds missed by a given analytical methodology—this most likely accounts for its failure to find estrogen metabolites. Furthermore, errors in the pathway database used likely contributed to the misidentification of several compounds (e.g. benzo-a-pyrene, acetyl isoniazid, compounds with fluorine), so while the error rate is *at least* one percent – based solely on compounds that were not relevant for this biological system - the actual error rate cannot effectively be estimated. One other key disadvantage of Mummichog is that it can only be used to do a two-class comparison, as it depends on relative fold-changes to assess consistent pathway impact and therefore cannot analyze a dose–response curve or a time-course.

Going forward, it is clear that adequate peak identification and high accuracy in determining the accurate parent compound from adducts is key for an untargeted approach, as any failure in those stages will likely lead to an inaccurate molecular weight and concentration, and that cannot be easily remedied with a bioinformatics approach. One potential solution involves ProbMetab (Silva et al., 2014), an R package for Bayesian inference of identification. ProbMetab, like Mummichog, presumes metabolites in a sample are related in pathways and uses that information as part of the metabolite identification. However, unlike Mummichog, it takes advantage of several other sources of information—namely, retention time and isotope abundance—and therefore is not uniquely dependent on pathway annotations and structure. More importantly, it assigns a probability to the identification, so a measure of the underlying uncertainty is preserved. Lastly, it retains features that could not be assigned identifications, so that this information is not lost and can be mined later. Another possibility is to attempt to leverage transcriptomic data to identify pathways that are up- or down- regulated, which could be used by a pathway-based algorithm to weigh likely pathways against less likely pathways. Metabolite identification at this stage remains a data-mining puzzle of using multiple clues as to the identity of a metabolite, and assembling them in a way that respects the underlying uncertainty.

Despite the challenges, metabolomics—and specifically, the ability to do untargeted metabolomics—is key to understanding disease, drug effects, and toxicity. With every drug studied, there are always surprises at how complicated the

effects of drugs are; inhibition of an individual pathway can have unpredictable results, since all pathways are connected.

As with any "-omics" discipline, the interpretation of the data is heavily dependent on the quality of the annotations used for data analysis. Even older, more established annotations such as GO have known biases that can distort data analysis (Gene Ontology et al., 2013). Large-scale, integrated pathway databases of metabolites are relatively new (the most extensive, the Human Metabolome Database (Wishart et al., 2013), began in 2004) and are likely to have uneven coverage of metabolites. This is particularly true for toxicological applications, as the existing annotated pathways are likely canonical pathways focusing on endogenous metabolites. The HMDB includes ethanol and caffeine metabolism pathways, but has no other pathway-level annotations for other exogenous compounds. And while another database, the Toxin and Toxin-Target Database (T3DB) (Lim et al., 2010), contains 2,900 toxins and over 1,300 targets, it does not integrate this information into relevant pathways, nor does it provide an easy way to interpret the target in terms of biologic networks.

Any future metabolomics studies should start by demonstrating reproducibility—in terms of quantifying and identifying known compounds—so that analytical issues can be solved before attempting to tackle complicated biological problems. It is also essential to explore a data set via several methods (ORA, QEA, correlation analysis, and perhaps in future genome-scale network reconstruction based on parallel microarray experiments). Lastly, results that are radically dependent on the methodology should cause concern.

Currently, the most significant bottleneck is metabolite identification: while some of the challenges for metabolite identification are largely analytical, a well-characterized, biochemically complete network would significantly aid the task. In addition to reducing (if not eliminating) uncertainty in metabolite identification, it would also help go beyond pathway identification to an analysis that is systems oriented, such as flux analysis or systems control theory. As it stands, however, bioinformatics can help fill-in gaps in incomplete data, but it cannot repair extremely noisy data.

In conclusion, scientists embarking on metabolomics as a part of systems biology should remember the effort of an early pioneer: the 16th century, Italian physician Sanctorius, whose 30-year experiment (which involved meticulously weighing everything he ate and excreted) laid the foundation for the quantitative study of metabolism (Ben-Menahem, 2009). Metabolomics has always required a very long-term commitment; it is not for the scientist looking for a quick payoff.

CHAPTER V – Conclusion

"Truth is much too complicated to allow anything but approximations." —John von Neumann

Transforming toxicology from a reductionist paradigm to a more systems-based approach will require a profound change in both practical hazard assessment and the basic research that underpins our understanding of toxic mechanisms.

As outlined in Chapter II and demonstrated in Chapter III, both the proliferation of alternative assays and the increasing sophistication of chemoinformatics requires new ways to think about chemical assessments, and a move towards a more formal and quantified paradigm and away from a weight-of-evidence approach. Small improvements from the perspective of machine learning can yield large improvements for practical hazard assessment—establishing with 90 percent certainty that a chemical is not a strong sensitizer with a few *in vitro* assays can help immensely when prioritizing chemicals in the R&D chain. Data mining and machine learning approaches are no longer optional in the modern era—they are a toolkit that every discipline will need to take advantage of, and toxicology is no exception.

Secondly, toxicology must move away from simplistic mechanisms towards a more pathway and/or network-oriented approach. While a toxic process may start with interference at a discrete point in the cell—for example, inhibiting one specific enzyme—rarely does that tell the whole story. While at some level attempting to capture the complexity may seem daunting, network abstractions allow one to see

the higher-level simplicity—protein interaction networks, whether in yeast or humans, share a similar topology (Bork et al., 2004). The seemingly bewildering array of regulatory interactions in a cell, in fact, demonstrate a few simple motifs (Alon, 2007).

In Chapter V, it was shown that MPTP toxicity involves a mechanism that is known (mitochondrial disruption) but has consequences that have yet to be fully described, and suggested a methodology to look for new candidates to extend the pathway past "the usual suspects". – an approach that took advantage of the high-dimensional data in part by reducing its dimensionality to look not at individual genes but at modules of genes that were related from both a functional and regulatory perspective. In Chapter VI, the potential and pitfalls of another high-dimensional field, metabolomics, were explored.

Pathway mapping approaches have many advantages over more vague and merely predictive "signatures of toxicity" approaches or black-box animal models. However, a fully specified Pathway of Toxicity requires a very fine-grained understanding of a biological system. From a bioinformatics perspective, this means that instead of stopping at the level of an abstract connectivity map (as is typically produced from correlation networks) or a truncated, simplified pathway of a complex disease (as is common in many pathway databases), a complete "molecules to phenotype" functional characterization will be required. While this may seem daunting, small improvements in our understanding of a Pathway of Toxicity can provide for large improvements in hazard assessment.

In our machine-learning approach to skin sensitization, the modest improvement in balanced accuracy could likely be improved by adopting an approach that is structured by the known mechanisms of toxicity; previous machine learning predictive models for skin sensitization have generally approached the problem more or less blind to the known mechanistic steps. However, as skin sensitization has a well-characterized Adverse Outcome Pathway, it should be possible to build a model that takes advantage of this. The steps involved in skin sensitization—skin penetration, electrophilic activity, covalent protein binding, cytokine induction & T-cell proliferation, and the processes that determine tissue inflammation, damage and repair —can each be predicted with some degree of accuracy with either chemical descriptors or *in vitro* assays. Therefore, one obvious extension of this methodology is to adapt a Hidden Markov model that combines chemical descriptors and *in vitro* assays and essentially progresses along each critical step of the pathway—that is to say, the transition probability between one state and the next would be determined by the probability that a chemical was positive or negative for each step of the Adverse Outcome Pathway. This has the advantage of simplifying the problem—there is no point in predicting electrophilicity for compounds that will not penetrate the skin—and could serve as a proof-of-concept on how to optimally incorporate chemoinformatics, *in vitro* assays, and a known Adverse Outcome Pathway to assess the probability of hazard.

Additionally, the basic insight of the approach—to incorporate a dose-based structure for the data—could be used for the concentration–response information from the assays themselves. Currently, information from the *in vitro* assays is

summarized as an EC_{50}, but in some respect this represents a loss of information. Including a dose–response based model for the descriptors as well as the end-point should both capture more data and eliminate some of the noise (caused by cytotoxicity or non-specific responses) that is intrinsic to *in vitro* assays.

Lastly, existing skin sensitization classifications are built around LLNA testing results, and while this provides an acceptable accuracy for hazard identification, the ultimate goal is to predict hazard in humans. A recent publication (Basketter et al., 2014) has identified 131 chemicals, which can be classified with confidence into six categories, ranked 1 to 5 based on potency and class 6 as true non-sensitizers. These categories reflect more accurately the range of hazards faced occupationally, since it considers both potency and length of exposure. Compared to other data sets used for machine learning approaches, it offers the additional advantage of being balanced equally among the classes, as opposed to other data sets, which are typically weighted towards non-sensitizers. A six-class model would likely be extremely difficult for most standard machine learning approaches, but a dose-based Hidden Markov model would likely be better able to make meaningful predictions in a six-class model.

Even for a well-understood Pathway of Toxicity, neither chemical descriptors alone nor any individual *in vitro* test will work adequately as a stand-alone replacement for assessing hazard; it is necessary to combine assays that address multiple points on the Pathway of Toxicity. In instances where the Pathway of Toxicity is unknown, it requires methodologies that can better delineate critical points at a molecular level—in essence, a methodology that can locate where, within

the regulatory circuitry of the cell, a toxicant is causing a malfunction that results in an altered phenotype.

While we demonstrated that a relatively small data set can provide novel insights even for a comparatively well-characterized toxin such as MPTP, it also demonstrated the limitations of depending on existing annotations. Although ontologies and annotations have certainly proven their worth vis-à-vis microarray data, an approach overly dependent on annotations will be limited to "looking under the lamp post for your key, because that's where the light is"—it allows you to see known biology, but limits the ability to find novel connections. It will certainly be too limiting for newer technologies, such as metabolomics.

While text-mining is key to extending annotations, it typically works best when used to answer a targeted question. A question such as "Which transcription factors are involved in Parkinson's?" is not easy for text-mining to answer. On the other hand, text-mining can well prioritize a small list of candidate transcription factors likely to be involved in Parkinson's. While the approach taken here was fairly simplistic, the existence of text-mining engines such as Textpresso (Muller, Rangarajan, Teal, & Sternberg, 2008)—which both tokenizes the data and resolves synonyms, and structures the data by looking for parts of speech that indicate interactions among entities—can answer fairly specific questions more efficiently than a literature review, an approach that was used to add substantially to the number of known histone acetylation positions (Huang et al., 2009) (see Appendix). While there are a few text-mining solutions that look for associations between a short list of genes—e.g. Chilibot (Chen & Sharp, 2004)—there are no robust text-

mining solutions that can find associations between several hundred genes and suggest functional characterizations.

In the case of genes, the problem of resolving synonyms has been largely addressed. For metabolomics, only the most rudimentary solutions are available, and effectively incorporating the extensive literature-base into a useful framework for a systems biology approach will be key to moving the discipline forward.

Moreover, mere functional characterization provides little information about regulatory mechanisms. WGCNA (Weighted Gene Correlation Network Analysis) combined with other high-throughput techniques offers one methodology to make an educated guess about mechanisms—in other words, to assign arrows to the connectivity map. While this methodology appears to work well for time-course studies, it may work less well for dose–response curves, and will likely need to be adjusted slightly to capture dose-dependent effects. In particular, a time-course network will capture largely linear effects, while a dose–response tends to have more non-linear effects. Put another way, somewhere on the dose–response curve there is a threshold that triggers a key change in biology. It may be necessary, then, to use correlation networks and graph theoretical approaches in order to focus more specifically on changes in network topology. However, this may require rethinking experimental design, as determining a network at each dose point will require comparatively more microarrays than are currently used.

In order to truly take a systems level view, a “molecules to phenotype” functional characterization requires substantially more information than a correlative approach can provide, and since it is largely a method of hypothesis

generation, it requires other means—typically *in vitro* or *in vivo* studies—to confirm the regulatory mechanisms proposed. However, as bioinformatics methods generate more and more testable hypotheses, a smarter approach to exploring such proposed regulatory mechanisms is required.

One possibility is to transform the genetic regulatory networks produced by data mining into an SBML model (Systems Biology Markup Language) (Finney & Hucka, 2003). SBML is a system designed to formally describe any biological entities that are linked by interactions or processes in a machine-readable format and (through graphical interpretation) human-readable diagrams. It is sufficiently flexible to specify genetic regulatory circuits, metabolic pathways, or cell-signaling pathways and can describe a system in as much or as little detail as necessary to capture the essential features (Chaouiya et al., 2013). An additional benefit is that there are several SBML-compatible curated pathways (e.g. PANTHER Pathways) to help structure the data; therefore, SBML models often do not require starting from scratch, but usually only the far simpler task of adding the relevant information suggested by the high-throughput approach along with existing pathways.

Because SBML requires explicit, formally specified interactions, it often shows areas in proposed pathways that are poorly understood or characterized and, in some cases, conflicting. The standard diagrams employed by cell biologists to describe mechanisms in molecular biology generally involve a bunch of arrows and symbols. The arrows could mean anything—transcription, activation, phosphorylation, or merely a vague and unspecified interaction—and the symbols

could be genes, proteins, small molecules, or, worse still, vague concepts such as "oxidative stress."

Therefore, structuring proposed pathways using SBML or other standardized, controlled formats would not only help both prune and extend the network generated by "-omics" technologies, it would help make the leap from basic pathway identification and hypothesis generation to models than can be used for more complex simulations, which can both weed out false positives and point to areas where proposed regulatory mechanisms are clearly inadequate to describe the data.

Meeting this challenge will involve both larger, more integrated data sets, novel bioinformatics approaches that look for dose–response curves, and an ability to incorporate legacy data to allow for better interpretation of "-omics" results. Transcriptomics has been in use for over two decades and has well-established analysis tools and standards for best practices and documentation, and transcriptomics will likely play a key role in discerning Pathways of Toxicity. As demonstrated here, much information can be gleaned from even a small study.

Metabolomics, on the other hand, is a nascent field. Measuring the abundance of metabolites has technical challenges as well as data analysis bottlenecks due to a relatively under-developed data infrastructure and a process that is acutely dependent on a fairly complex data analysis workflow. From the bioinformatics perspective, the relatively sparse annotation data available for metabolites compared to genes will require an approach that can learn networks from data rather than depend on existing pathway maps. While correlation of co-expression

(commonly called the "guilt-by-association" method) has been a powerful method to predict gene networks *ab initio* from microarray data (Quackenbush, 2003; Stuart, Segal, Koller, & Kim, 2003) its application to metabolomics is less straightforward. While genes that show similar expression patterns likely share some level of transcriptional control, this is not the case for correlation of co-expression in metabolomics, as metabolites are intrinsically interdependent in a way that genes are not. Nonetheless, such correlations that are consistent over conditions (such as time-course analysis or different treatments) are not *necessarily* close by on a metabolic map, but do probably have some sort of linkage: they may be in chemical equilibrium, have a mass conservation relationship, are under asymmetric control, or are under the tight control of a specific gene which varies amongst the data sets (Steuer, Kurths, Fiehn, & Weckwerth, 2003). Correlative based approaches have been used to deduce novel pathways (Fukushima, Kusano, Redestig, Arita, & Saito, 2011) in previous metabolomics studies.

However, while bioinformatics approaches can solve metabolite identities if some of the network is accurately identified and quantified, it can do little to solve deeper analytical problems. One way forward for both the analytic and bioinformatics problems presented by metabolomics involves the use of stable isotope labeling. Such isotopes can be used as standards for metabolite identification, which provides a needed check on whether data analysis workflows are performing adequately. Furthermore, stable isotopes can used for fluxomics – that is, studying the reaction rates in cells – which can be considered the most direct read –out of the metabolic phenotype (Klein & Heinzle, 2012).

However, as it is, the learning curve for metabolomics is incredibly steep, and metabolomics will likely require greater maturation as a field before it becomes an equal partner to transcriptomics for Pathway of Toxicity identification and certainly has a long way to go before it becomes an everyday tool for hazard assessment.

This does, however, indicate that one key goal of any project to provide effective *in vitro* assays to discern Pathways of Toxicity will be establishing an effective dimensionality reduction of the data so that the noise from both the biological variability and technical aspects does not overwhelm the signal, and that the derived Pathways of Toxicity are not the result of over-fitting to one limited set of data and are robust when compared with existing data.

Mapping the Human Toxome will involve a degree of integration of multiple levels of molecular data with cellular responses that has not as of yet been carried out—and because, ultimately, toxicity is the study of dose, it will require a mechanistic understanding of phenotypic changes at a low-dose level and bioinformatics approaches that can tease out dose response. Theoretically, an estrogenic compound can begin affecting cells at concentrations as low as a single molecule per cell. Understanding how a Pathway of Toxicity responds to stimuli, especially at low doses typical of environmental exposures, will be a substantial bioinformatics challenge, yet it is essential for this approach to be effective

At the same time, it is necessary to be realistic about the limitations of *in vitro* approaches: cells are often misidentified, often the standard cell lines which form the foundation of much of basic science are radically genetically different—as an example, HeLa cells are radically genetically different between groups, and cells

of tumor origin may have up to 20,000 mutations (Hartung, 2013). Tissue culture cells often exist in a micro-environment that is profoundly abnormal and may lack metabolism or cell defenses (Hartung, 2007).

Finally, many, if not most, toxic processes involve interplay between tissue types. As an example, any attempt to examine MPTP toxicity that assumed astrocytes only metabolized MPTP to MPP+ may miss critical processes in astrocytes that contribute to the adverse outcome. While this may seem an intrinsic limitation for *in vitro* approaches, this is not necessarily so. More complex organ-on-a-chip *in vitro* systems offer one solution (van der Meer & van den Berg, 2012). As another, it is necessary to accept that a stand-alone replacement for an animal test is unlikely—skin sensitization is a fairly easy target compared to neurotoxicity or endocrine disruption. Any attempt to use *in vitro* assays will likely require an intelligent way to combine multiple sources of information.

Additionally, the limitations of high-throughput/high-dimensional approaches must be kept in mind. To begin with, such approaches often fail—especially when the technology is new—to conduct adequate quality assurance. Quality assurance is an essential component of science; however, in the past, much of basic science and preclinical research paid minimal attention to quality assurance and reproducibility, but this attitude must be changed, given the growth of studies that are not reproducible (Begley & Ellis, 2012) (Hartung, 2013). Both the temptation (and, owing to the complexity of interpretation, the comparative ease) of spinning high-throughput/high-dimensional data into a "good story" means that quality assurance is of critical importance to any alternative method based on such

techniques. This will be of particular importance to metabolomics, owing to the sensitivity of the technique, the ambiguity of metabolite identification, and the high probability of artifacts—the temptation will always be there for researchers to treat a fluke as a profound finding, and the only guard against this is a culture of quality assurance and reproducibility.

Lastly, all technologies offer only a narrow glimpse of the biological complexity underneath, and just as important as it is to adequately interpret the data presented, it is equally important to keep in mind what a given technology is simply incapable of seeing.

As useful as transcriptomics has been, much of cell signaling is either reflected in the phospho-proteome, the metabolome, or the complex dynamics of microRNA regulation. Each of these represent analytical and bioinformatics challenges to scale-up to the extent that they can be an equal partner to transcriptomics, but a more complete understanding of the Human Toxome will require the different perspectives. Nonetheless, a microarray study which provides a plausible genetic regulatory network is a far more efficient and informative use of animals than a study which provides only a NOAEL/LOAEL; going forward, any use of animals should aim to do so in as data-rich a way as possible, both as more humane science but also as simply better science.

As the Human Toxome is finite, it can certainly be mapped, but currently we have only a few well-characterized islands and a vast ocean of unknowns. In many instances (e.g. endocrine disruptors), the unknowns might as well be mapped "Here Be Dragons," as they become a locus onto which nebulous fears are projected. At the

same time, the finiteness of the Human Toxome will never allow us the luxury of *certainty* about toxicity mechanisms. A useful cautionary tale comes from the early history of x-ray technology. It was a commonly understood occupational hazard that individuals who worked with x-rays would often have skin burns, but the results were consistently attributed to things other than the x-ray (perhaps the chemicals used to develop it?) simply because no one could imagine light that was neither seen nor felt could possibly produce injury (Kevles, 1998). No doubt there are mechanisms of toxicity as yet undreamt of by our current philosophy.

APPENDIX I Green Toxicology

(Originally published as: Maertens, A., Anastas, N., Spencer, P. J., Stephens, M., Goldberg, A., & Hartung, T. (2013). Green Toxicology. *ALTEX*, *31*(3), 243-249.)

Food for thought…

Green Toxicology

Alexandra Maertens[1], Nicholas Anastas[3], Pamela J. Spencer[4], Martin Stephens[1], Alan Goldberg[1] and Thomas Hartung[1,2]

[1]Johns Hopkins University, Bloomberg School of Public Health, CAAT, Baltimore, USA; [2]University of Konstanz, CAAT-Europe, Germany; [3]EPA Region 1, Boston, MA; [4]Dow Chemicals, Midland, MI

Abstract

Historically, early identification and characterization of adverse effects of industrial chemicals was difficult because conventional toxicological test methods did not meet R&D needs (e.g. methods that are rapid, relatively inexpensive and amenable to small amounts of test material). The pharmaceutical industry has moved to front-loading toxicity testing, i.e. into using some *in silico*, *in vitro* and less demanding animal tests at earlier stages of product development to identify and anticipate undesirable toxicological effects and optimize product development. The Green Chemistry movement embraces similar ideas to result in less toxic products, safer processes and less waste and exposure. Going even a step further, the concept of "*benign design*" suggests ways to consider possible toxicities before the actual synthesis and to apply some structure/activity rules (SAR) and *in silico* methods. This requires not only scientific development but a change in corporate culture, where synthetic chemists work with toxicologists. An emerging discipline called *Green Toxicology* (Anastas, 2012) provides a framework for integrating the principles of toxicology into the enterprise of designing safer chemicals, thereby

minimizing potential toxicity as early in production as possible. Green toxicology's novel utility lies in driving innovation by moving safety considerations to the earliest stage in a chemical's lifecycle, i.e., to molecular design. In principle this field is no different than other sub-disciplines of toxicology that endeavor to focus the tools of toxicology on a specific area, for example, clinical, environmental or forensic toxicology. We use the same principles and tools of toxicology to evaluate an existing substance or to design a new one. The unique emphasis is in using 21st century toxicology tools as a preventative strategy to design out undesired human health and environmental effects thereby increasing the likelihood of launch of a successful, sustainable product. Starting with the formation of a steering group and a series of workshops, the *Green Toxicology* concept is currently spread internationally and refined as an iterative process.

Introduction

Over the past few decades, there has been an increase in consumer demand for less toxic, more environmentally friendly products, as well as increasing regulatory and economic pressure for more sustainable products, less wasteful manufacturing, and a switch to renewable resources as source materials—in essence, a "*Green Chemistry*" approach (Paul & John, 1998) which puts environmental and sustainable principles at the forefront of chemical design.

However, in order for Green Chemistry to flourish, there must be a parallel paradigm change in toxicology: less toxic chemicals cannot be effectively designed unless scientists have the necessary tools to quickly and accurately assess chemical hazards. Toxicology has hitherto been little concerned with developing tools to help chemists better understand toxicity and design better alternatives. The principle of "benign design" has been part of the 12 founding principles of Green Chemistry from the beginning, as principles 3 and 4 directly address this (Box 1). Other principles aim to reduce waste and use of chemicals and thus limit exposure in the environment and the workplace.

Box 1

The 12 principles of Green Chemistry . (Paul & John, 1998)

1. It is better to prevent waste than to treat or clean up waste after it is formed.
2. Synthetic methods should be designed to maximize the incorporation of all materials use in the process into the final product.
3. Wherever practicable, synthetic methodologies should be designed to use and generate substances that possess little or no toxicity to human health and the environment.
4. Chemical products should be designed to preserve efficacy of function while reducing toxicity.
5. The use of auxiliary substances (e.g. solvents, separation agents, etc.) should be made unnecessary wherever possible and innocuous when used.
6. Energy requirements should be recognized for their environmental and economic impacts and should be minimized. Synthetic methods should be conducted at ambient temperature and pressure.
7. A raw material or feedstock should be renewable rather than depleting wherever technically and economically practicable.
8. Reduce derivatives: Unnecessary derivatization (blocking group, protection/deprotection, temporary modification) should be avoided whenever possible.
9. Catalytic reagents (as selective as possible) are superior to stoichiometric reagents.
10. Chemical products should be designed so that at the end of their function they do not persist in the environment and break down into innocuous degradation products.
11. Analytical methodologies need to be further developed to allow for real-time, in-process monitoring and control prior to the formation of hazardous substances.

12. Substances and the form of a substance used in a chemical process should be chosen to minimize potential for chemical accidents, including releases, explosions, and fires.

The current industrial product development paradigm relies on time-consuming, expensive animal studies and is too slow to keep pace with technological change (Hartung, 2010b; Hartung & Rovida, 2009). For example, a typical 2-generation reproductive study costs more than $500,000, uses more than 3000 rats and takes 15 months to complete. For this reason, toxicity testing is typically reserved for the latter stages of chemical/product development after it's determined to be commercially viable. Consequently, toxic effects are identified closer to commercialization when little options for design changes exist and after significant investment of time, resources and money. Today, rapidly evolving, 21st century safety assessment methodologies have the potential to transform how companies develop and commercialize new products and chemicals

This rapid, high-throughput, high-content "*Green Toxicology*" paradigm can work in tandem with R&D by providing answers about mechanism of toxicity quickly, inexpensively, and with the small quantities of material typically available for R&D. "*Green Toxicology*" combines the *in vitro* and *in silico* tools of predictive toxicology with the principles of chemical design to develop chemicals that have negligible toxicity, and early elimination of candidates possessing undesirable traits by "failing early and failing cheaply", or to put it more positively, to enable innovation through early and inexpensive evaluation of hazard.

Consideration 1: The first principle of Green Toxicology—"Benign design"

The idea is simple: toxicologists partner with synthetic chemists to understand what chemical moiety may impart undesired hazard traits as early as feasible in product development. Toxicology is in the midst of a major transition from animal-based

methods that are slow, expensive and suffer from low-throughput, to more modern approaches utilizing cheminformatics, cell cultures, genomics and computational biology to achieve greater speed and throughput, lower cost, and ultimately, more accurate predictions of safety in humans and the environment.

For example, programs based on structure activity relationships (SAR) can be useful in guiding early selection of low hazard candidates to continue in product development. A nice illustration is the "ultimate rat carcinogen" drawn by Tennant and Ashby (Ashby & Tennant, 1991) showing the chemical features associated with mutagenicity in one theoretical molecule (Figure 1). However, when challenged to prospectively predict the outcome for 30 chemicals to be tested in the US National Toxicology Program, the authors achieved only 50-60% prediction of the carcinogenic substances and wrongly predicted 40-50% of non-carcinogens to be positive in the animal test (Benigni & Zito, 2004). This illustrates the limitations of SARs for such complex endpoints (D. Basketter et al., 2012) and

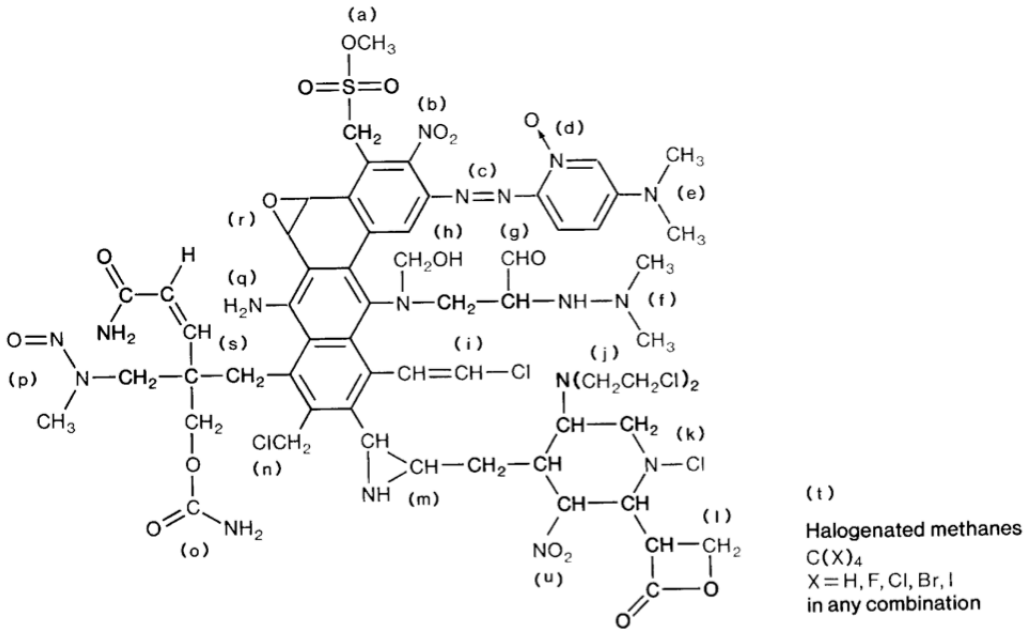


Fig. 1. Revision of the multiple electrophile model chemical upon which structural alerts in the present papers are based. The new sub-structure is labelled (u) and represents an aliphatic nitro group, as present in tetranitromethane [277]. All other substructures (a–b) are as described in detail in Ashby et al. (1989).

Chapter I, Figure 15: Ultimate Rat Carcinogen

therefore, as much as read-across and QSARs have helped to make testing more targeted and efficient, their utility lies as low cost, rapid "Tier 1" assessments of new candidate chemistries and sustainable alternatives.

It is worth pondering whether the existing QSARs will have adequate applicability for the more novel chemicals that emerge from green chemistry research—e.g. QSARs developed for industrial synthetic chemicals may not be applicable for bio-based materials. Some positive examples exist especially in the field of aquatic toxicity (Voutchkova, Ferris, Zimmerman, & Anastas, 2010; Voutchkova et al., 2011; Voutchkova, Osimitz, & Anastas, 2010) but this is arguably an easy case, where lipophilicity is key to uptake and thus hazard. However, this does not mean that helpful estimates for more complex hazards such as immunotoxicity (Hartung and Corsini, 2013), developmental neurotoxicity (Smirnova, Hogberg, Leist, & Hartung, 2014) or endocrine disruption (Juberg et al., 2013) could not be done.

Additionally, while QSARs have certainly proven their merit in the pharmaceutical industry, this success is unlikely to be repeated for industrial chemicals for a variety of reasons. Industrial chemicals may consist of polymers with a wide range of molecular weights, various impurities, left over reagents etc., while the large majority of drugs fall into a more narrow chemical space, often referred to as the Lipinsky rules (Lipinski, 2004):

- No more than 5 hydrogen bond donors (the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)
- Not more than 10 hydrogen bond acceptors (all nitrogen or oxygen atoms)
- A molecular mass less than 500 daltons
- An octanol-water partition coefficient log P not greater than 5

This is a nice example, though not for safety but efficacy, how structure considerations can help designing substances.

More fundamentally, QSARs developed for the pharmaceutical industry have a domain defined by suspected biological activity; QSARs may simply lack the accuracy necessary when the overwhelming number of chemicals are in fact lacking

toxicity, as is the case for many industrial chemicals (Box 2); the respective estimates reflect internal ECVAM analyses of the European New Chemical Database, which includes new industrial chemicals registered since 1981 under the Dangerous Substance Directive, around 2005(Hoffmann, Cole, & Hartung, 2005; Hoffmann & Hartung, 2005). Therefore, while QSARS likely will have a role to play in the development of benign alternatives, it is equally important that toxicology develop other techniques and approaches that link molecular structure with toxic outcomes in a way that can be useful to synthetic chemists.

Box 2

Most chemicals are not toxic:

90% not acutely toxic (EU New Chemical Database)

97% not skin corrosive (EU New Chemical Database)

93% not skin irritant (EU New Chemical Database)

97% not teratogenic (expert estimate, about 60% not positive in single species two-generation studies)

80-95% not carcinogenic (expert estimates, 47% not positive in rodent bioassay)

80% not eye irritating (EU New Chemical Database)

65% not skin sensitizing (EU New Chemical Database)

Consideration 2: The second principle of Green Toxicology—"Test early, produce safe"

The pharmaceutical industry has developed concepts of "*fail early, fail cheap*" as a consequence of the cost explosion in the late clinical part of development and the high failure rates observed there (Hartung, 2013; Hartung & Zurlo, 2012). For

example it was noted that in the 1990s, a large number of drugs failed because of pharmacokinetic problems, i.e. the active agent did not reach sufficient concentrations in the targeted organ in patients. Addressing this early and with human relevant methods markedly reduced this type of failure (Singh, 2006; Tsaioun & Jacewicz, 2009).

This approach can also be adapted to the front-loading of toxicity testing of industrial chemicals. In the short term, predictive safety assessment offers a way to enrich the R&D pipeline for chemicals that are most likely to clear challenging regulatory hurdles. Because predictive methods focus on the root causes of toxicity at the cellular and molecular levels, they also generate new knowledge to inform the design of safer and more sustainable products. Traditional toxicity tests total several million dollars for a product to go to the market. These studies also take a lot of time, in some cases taking years to complete, e.g., the rat cancer bioassay entails two years of treatment plus time for planning, histopathology and reporting. And often, at the end of this process, the results are equivocal and may be of questionable relevance to humans. If the results are positive, such bioassays typically provide no mechanistic information for the synthetic chemist to design a less toxic alternative. Clearly, under the pressure of "*time to market*" and the running clock of the patents and competitive economic pressures, these are not the best tools for early decision taking.

Front-loading thus requires screening level tests that are both less costly and much faster, and a movement to a smarter approach that begins with *in silico* screening to predict possible targets and progresses to targeted *in vitro* tests that can examine suspected Pathways of Toxicity (Hartung and McBride, 2011; Kleensang et al., 2014). For those candidates that do move on to whole animal tests, a smarter testing approach might allow for reduced reliance on high-dose testing that causes gross pathological change as an indication of toxicity and focuses more precisely on the molecular initiating event at doses that can meaningfully be related to possible human exposures.

Another advantage of front-loading toxicity in the R&D process would be to reduce cases of "*out of the frying pan, into the fire*"—in other words, often replacements that are promoted as alternatives to known "*bad actors*" turn out to be not necessarily less toxic, but in fact simply have less data. This was the case with flame retardants (Lakind & Birnbaum, 2010). This creates a somewhat perverse incentive *not* to gather toxicity data, which is compounded by the fact that consumer preference can be markedly influenced by the results of toxicity tests that are taken out of context. More rigorous toxicity testing as an essential part of the R&D process would likely produce a more rational selection of benign replacements.

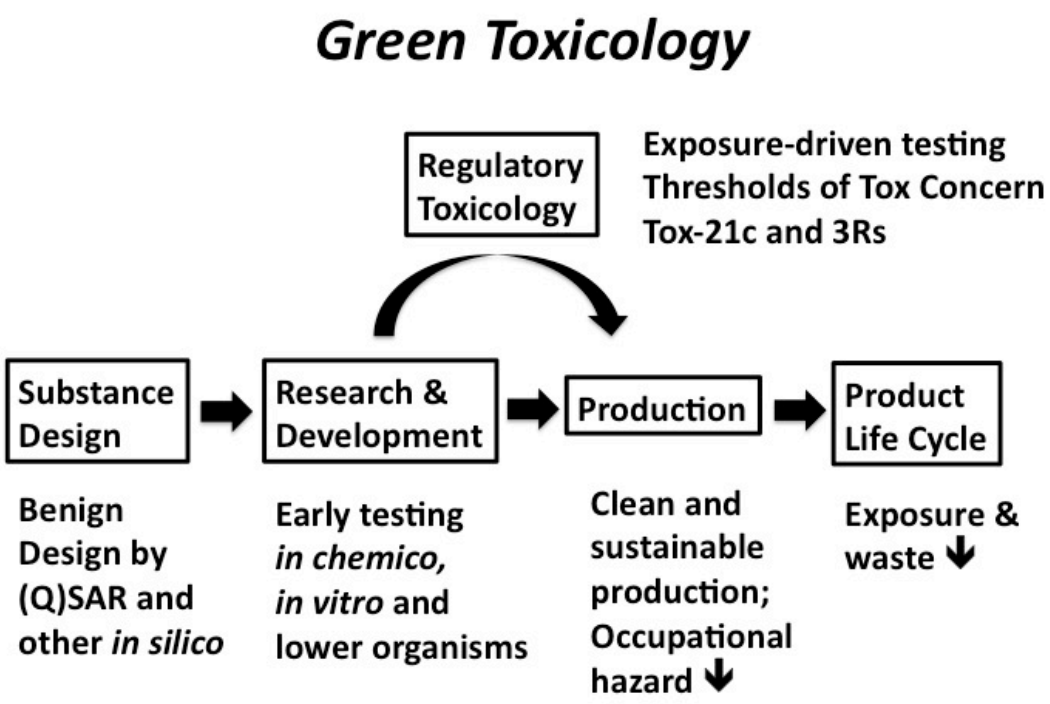
Consideration 3: The third principle of Green Toxicology—"Avoid exposure and thus testing needs"

Traditionally, toxicologists are trained to think in terms of molecules and not in terms of the production processes behind them. However, within the many steps involved in the production of industrial chemicals, there are often small alterations that can achieve significant reductions in terms of exposure and therefore minimize risk—e.g. toxicity may reside in a catalytic agent that can be eliminated with alternative routes of synthesis. For many polymers, the final product has a sufficiently large molecular weight so as to preclude bioavailability, and any hazard is likely because of residual monomer. Consequently, small changes in the efficiency of the reaction or the purification step can drastically reduce the hazard while conserving resources. Similarly, a change to "*one-pot synthesis*" (meaning that all reactions take place in the same vessel) can decrease the number of exposed workers. In this respect, the goals of Green Toxicology dovetail with the Green Chemistry goal of improved efficiency and emphasize the importance of close collaboration between the chemist and the toxicologist. Together their measures directly affect occupational health and then via reduced exposure also influence risk assessment and management. Such scenarios are ripe for exposure-driven testing strategies, which can result in reduced testing demands. Reduced exposure also

makes it more likely that Thresholds of Toxicological Concern (TTC) (R Kroes, Kleiner, & Renwick, 2005), (Munro, Renwick, & Danielewska-Nikiel, 2008) are not exceeded, an example of a formalized exposure-driven assessment. The idea is simple: we can assess how much of the known toxicants is necessary to exert a certain effect (which will give a distribution of doses) and then define a point of departure dose. With an appropriate safety factor, it is unlikely that an unknown agent will exert this toxicity. The concept has been pioneered for food (Robert Kroes et al., 2004) and cosmetics (Blackburn et al., 2005; R Kroes et al., 2007) but also adapted to pharmaceuticals, especially for genotoxic impurities. Noteworthy, the World Health Organization is currently reviewing the approach[1]. TTCs have first been used for carcinogens, but the concept was also adapted for example to reproductive toxicity testing (Van Ravenzwaay, Dammann, Buesen, & Schneider, 2011). We argued in a similar way in the context of immunotoxicity (Hartung & Corsini, 2013), showing that clinical immunosuppressants require mg/kg quantities to exert their effects and that this could be used for establishing TTC for substances not optimized for this purpose.

In conclusion, Green Chemistry—by reducing exposure and thus testing demands—has more to offer to toxicology, safety testing and risk management than just Benign Design and Early Testing (figure 2).

[1] http://www.who.int/foodsafety/chem/ttc/en/



Chapter I, Figure 16: Green Toxicology Paradigm

Consideration 4: The fourth principle of Green Toxicology—"Make testing sustainable"

Animal testing is not just costly in terms of time and money, but is inefficient with regards to resources, requiring energy and producing a great deal of biohazard waste. Consequently, we argue that the long-term use of animals is fundamentally not sustainable. It has been estimated that regulatory requirements in Europe require approximately 300 animals to test new chemical compounds up to 10,000 for a pesticide (A. Bottini & Hartung, 2010). Notably, the 10.000 animals per pesticide do not include abandoned products. Before REACH in Europe 90 to 140 thousand animals were used for chemical testing of roughly 200 new chemicals per year, but this does not include testing outside of Europe. In the US, discrepancies in testing demands between are even stronger between different products with 7 out of 8 new industrial chemicals having no toxicity data at pre-marketing notification under the Toxic Substance Control Act (Hartung, 2010) and similar requests of more

than 30 animal tests for pesticides. With 40,000 animals tested for safety per new substance entering the market and 350,000 for R&D (factoring in the animals used for candidate drugs that do not succeed), the pharmaceutical sector still has very high animal use despite the impressive reductions in recent years. This is not only unsustainable but may impose an economic barrier that is prohibitive for niche chemicals that may have limited profitability. A smarter, *in vitro* testing strategy can reduce the use of resources for testing by better prioritization and more efficient screening-level tests. Longer term we hope agencies will find greater application of predictive methods to address some requirements of their programs.

Consideration 5: "Early testing can use methods not yet mature for regulating".

Regulation tends to take a precautionary approach that is oriented towards minimizing mistakes rather than optimizing the cost/benefit analysis. This makes it profoundly difficult to change traditional approaches. Furthermore, a traditional validation study takes about one decade. Consequently, a validated test is "frozen in time," and it is simply impossible for regulatory mechanisms to keep up-to-date with the current rate of change in science (Hartung, 2007;(Leist, Hasiwa, Daneshian, & Hartung, 2012).

Frontloading toxicity at the research and development stage, however, allows a more flexible approach. Prioritization of substances as lead for development can be based on methods, which still have some margins of error. Early testing allows the use of methods not yet validated. *In silico* and *in vitro* tests that are individually too inaccurate for regulatory purposes will likely have a useful place in an integrated testing strategy (Hartung et al., 2013). Such strategies allow uncertainty in results and seek to combine data from multiple tests in a flexible manner that maximizes predictive power while also providing an estimate of the uncertainty in the data. This helps to build capacity and capability to perform these assays also for later regulatory use, if validated and accepted. In the meantime, these front-loaded methods will be generating data, and thereby facilitating an assessment of the

predictive value of these methods and thus contributing to the validation and acceptance process.

This opens up also a role for new risk assessments based on toxicity "pathways" (cell/molecular level changes) and data-driven uncertainty factors (e.g., intra-human variability factors based on genetic analysis). It will take tremendous time to base regulatory testing on pathways of toxicity (PoT), as the respective database first would need to be sufficiently comprehensive and validated (Hartung and McBride, 2011). However, with each and every PoT identified the respective assays can be included in integrated testing strategies. A pathway-based approach can also allow for more precise understanding of individual variation in response to toxicity as well as susceptible populations by illuminating more precisely the differences in PoT. Similarly, default safety and assessment factors might be replaced by knowledge on intra-species and inter-individual differences in PoT.

Consideration 6: Green Toxicology as a Driver of 21st Century Toxicology

Biology has been transformed over the last decade from a reductionist and largely qualitative science to a more quantitative approach that requires systems-level thinking, large-scale data analysis, and multi-scale modeling. Although certain areas of toxicology (such as PBPK modeling) have long embraced mathematical models and certain elements of systems-level thinking, the insights gained from systems biology have not generally been reflected in regulatory toxicology or hazard assessment. Furthermore, the field of toxicology is only beginning to assemble the type of large-scale data sets that have been transformative for molecular biology. As the green toxicology paradigm of high-throughput, -omics based approaches for screening many compounds gathers data, this can act as a driver towards transforming toxicology from a reductionist approach based on "feet-up/feet-down" assays (i.e. the LD_{50}) towards an approach that uses the insights of systems biology, computational modeling, and exploratory data mining to locate the mechanism of toxicity in perturbed networks. Green toxicology can serve as a bridge between 21st century toxicology methods and the development of safer, sustainable products.

This paradigm shift and transformation is necessarily a slow and long process as safety of workers and consumers is at stake. This delay makes regulatory science less attractive for academic research and even more for the commercialization of test methods. If companies have to wait a decade for the regulatory acceptance of a test with unclear prospects for the validation phase, the return of investment is rather unlikely. Early non-regulatory testing creates an immediate market for new test methods. It therefore liberates the market forces necessary to standardize and disseminate tests also internationally (A. A. Bottini, Amcoff, & Hartung, 2007), Bottini and Hartung, 2008).

Consideration 7: The Green Toxicology Program

Following on the initial success of our Green Toxicology Day in November of 2013 and its forerunner at University of Connecticut in December 2012[2], a follow-up series of webinars is planned. In addition, a proposal for a session at the 2015 SOT meeting has been accepted. Information sessions at the GlobalChem conference and the ACS Green Chemistry Conference are planned. A multi-day "Green Toxicology Workshop" is planned for the spring of 2015 in Washington State. Curricula for students especially of synthetic chemistry—who typically are given minimal training in toxicology—are a further goal. Development of dedicated scientific articles and a textbook on "Green Toxicology" as well as a compendium of Design Rules for Reduced Hazard aimed at synthetic chemists ('the green toolbox"), will be significant products of the effort. Furthermore, the CAAT policy program will inform policy makers about the opportunities of a Green Toxicology approach. Key to this outreach will be bringing together two communities—toxicologists and chemists—that have long worked in parallel but have heretofore rarely worked collaboratively.

[2] http://caat.jhsph.edu/programs/workshops/greenTox.html

Conclusions

Alternative methods in toxicology increasingly represent themselves as enabling technologies, i.e. they can do more than optimize and replace current regulatory testing: The pharmaceutical field has for a while been taking advantage of front-loading of testing and mechanistic understanding for early determination of possible toxic liabilities. The chemical industry has started to embrace similar concepts in the Green Chemistry movement. A Green Toxicology is emerging, which uses structure-activity relationships for the design of less harmful substances, tests early in the development process to prioritize less dangerous chemicals and reduces exposures thus reducing risk and testing demands. These approaches promise to create opportunities for the development and use of alternative test methods and support a transition to sustainable chemistry.

APPENDIX II Integrated Testing Strategy

(Originally publishes as: Hartung, T., Luechtefeld, T., Maertens, A., & Kleensang, A. (2013). Food for Thought… Integrated Testing Strategies for Safety Assessments. *Altex*, *30*(1), 3)

"Playing safe is probably the most
unsafe thing in the world.

You cannot stand still.

You must go forward"

Robert Collier (1885-1950)

Food for thought…
Integrated Testing Strategies for Safety Assessments

Thomas Hartung[1,2], Tom Luechtefeld[1], Alexandra Maertens[1] and Andre Kleensang[1]

[1]Johns Hopkins University, Bloomberg School of Public Health, CAAT, Baltimore, USA; [2]University of Konstanz, CAAT-Europe, Germany

Abstract

Despite the fact that toxicology uses many stand-alone tests, very often a systematic combination of several information sources is required: Examples include, when not all possible outcomes of interest (e.g. modes of action), classes of test substances (applicability domains) or severity classes of effect are covered in a single test;

furthermore, sometimes the positive test result is rare (low prevalence leading to excessive false-positive results) or the gold standard test is too costly / uses too many animals creating a need for prioritization by screening. Similarly, tests are combined when the human predictivity of a single test is not satisfying or existing data and evidences from various tests shall be integrated. Increasingly, also kinetic information shall be integrated to make an *in vivo* extrapolation from *in vitro* data.

The solution to these problems is Integrated Testing Strategies (ITS). They have been discussed for more than a decade and some attempts have been made in test guidance for regulations. But despite their obvious potential to revamp regulatory toxicology, we still have little guidance on the composition, validation and adaptation of ITS for different purposes. Similarly to approaches of Weight of Evidence and Evidence-based Toxicology, different pieces of evidence and test data need to be weighed and combined. ITS represent also the logical way of combining pathway-based tests as suggested in Toxicology for the 21st Century. Here, the state of the art of ITS is described and suggestions as to definition, systematic combination and quality assurance of ITS are made.

Introduction

Replacing a test on a living organism with a cellular, chemico-analytical or computational approach is obviously reductionistic. Sometimes this might work well, e.g. when an extreme pH is a clear indication of corrosivity. However, in general it is quite naïve to expect a single, system to substitute for all mechanisms, the entire applicability domain (substance classes) and degrees of severity. Still toxicology has long neglected this when requesting a replacement to substitute one by one the traditional animal test. We might even extend this to say it is similarly naïve to address an entire human health effect with a single animal experiment using inbred, young rodents… The only way to approximate human relevance is to mimic the complexity and responsiveness of the organ situation and model the respective kinetics, i.e. what the human-on-a-chip approach targets (Hartung and Zurlo, 2012). Everything else requires making use of several information sources if not compromising the coverage of the test. Genotoxicity is a nice example, where patches have continuously been added to cover the various mechanisms. However, here the simplest possible strategy, i.e. a battery of tests, where every positive result is considered a liability, causes problem. We have seen where the inevitable accumulation of false-positives leads (Kirkland et al., 2005), ultimately undermining the credibility of *in vitro* approaches.

The solution is the "intelligent" or "integrated" use of several information sources in a testing strategy (ITS). There is a lot of confusion around this term and even more, how to design, validate and use ITS.

This article aims to elaborate on these aspects with examples and outline the prospects of ITS in toxicology. It thereby expands the thoughts elaborated for the introduction to the roadmap for animal-free systemic toxicity testing (Basketter et al., 2012). The underlying problems and the approach is actually not unique to toxicology. The most evident similarity is to diagnostic testing strategies in clinical medicine, where similarly several sources of information are used for differential

diagnosis; we have discussed earlier these similarities (Hoffmann and Hartung, 2005).

Consideration 1: The two origins of ITS in safety assessments

When do we need a test and when do we need a testing strategy? We need more than one test, if:

- not all possible outcomes of interest (e.g. modes of action) are covered in a single test
- not all classes of test substances are covered (applicability domains)
- not all severity classes of effect are covered
- when the positive test result is rare (low prevalence) and the number of false-positive results becomes excessive (Hoffmann and Hartung, 2005)
- when the gold standard test is too costly or uses too many animals and substances need to be prioritized
- when the accuracy (human predictivity) is not satisfying and predictivity can be improved
- existing data and evidences from various tests shall be integrated
- kinetic information shall be integrated to make an *in vivo* extrapolation from *in vitro* data (Basketter et al., 2012)

All together, it is difficult to imagine a case, where we should not apply a testing strategy. It is astonishing how long we have still pursued *"one test suits all"* solutions in toxicology. If at all, a restricted usefulness (applicability domain) was stated, but it was only with the discussion on Integrated Testing of *in vitro, in silico* and toxicokinetics (adsorption, distribution, metabolism, excretion, i.e. ADME) information that such integration was attempted. Bas Blaauboer and colleagues was for long spearheading this (Blaauboer, 2010; Blaauboer, Barratt, & Houston, 1999; DeJongh, Nordin-Andersson, Ploeger, & Forsby, 1999; Forsby & Blaauboer, 2007). The first ITS were accepted as OECD test guidelines in 2002 for eye and skin irritation (OECD TG 404, 2002; OECD TG 405, 2002). A major driving force was then

the emerging REACH legislation, which sought to make use of all available information for registration of chemicals (especially existing chemicals) in order to limit costs and animal use, prompted the call for Intelligent TS (Ahlers, Stock, & Werschkun, 2008; Anon, 2005; Combes & Balls, 2011; Gabbert & Benighaus, 2012; Leist et al., 2012; Schaafsma, Kroese, Tielemans, Van de Sandt, & Van Leeuwen, 2009; van Leeuwen, Patlewicz, & Worth, 2007; Vonk et al., 2009)..

The two differ to some extent as the REACH-ITS include also *in vivo* data and are somewhat restricted to the tools prescribed in legislation. This excludes largely the 21st century methodologies (van Vliet, 2011) i.e. omics, high-throughput and high-content imaging techniques, which are not mentioned in the legislative text. The very narrow interpretation of the legislative text in administrating REACH does not encourage such additional approaches. This represents a tremendous opportunity lost and some more flexibility and "*learning on the road*" would benefit one of the largest investments in consumer safety ever attempted.

Astonishingly, despite these prospects and billions of Euros spent for REACH the literature on ITS for safety assessments is still poor and little progress toward consensus and guidance have been made. For example, two In Vitro Testing Industrial Platform workshops were summarized stating (De Wever et al., 2012): *"As yet, there is great dispute among experts on how to represent ITS for classification, labelling or risk assessments of chemicals, and whether or not to focus on the whole chemical domain or on a specific application. The absence of accepted Weight of Evidence (WoE) tools allowing for objective judgements was identified as an important issue blocking any significant progress in the area."* Similarly, the ECVAM/EPAA workshop concluded (Kinsner-Ovaskainen et al., 2012): *"Despite the fact that some useful insights and preliminary conclusions could be extracted from the dynamic discussions at the workshop, regretfully, true consensus could not be reached on all aspects."*

We have earlier commissioned a whiter paper on ITS (Jaworska & Hoffmann, 2010) in the context of our transatlantic think tank for toxicology (t^4) and a 2010

conference on *21st century validation for 21st century tools*. It similarly concluded: *"Although a pressing concern, the topic of ITS has drawn mostly general reviews, broad concepts, and the expression of a clear need for more research on ITS (Benfenati, Gini, Hoffmann, & Luttik, 2010; J. Hengstler et al., 2006; Worth et al., 2007). Published research in the field remains scarce (Gubbels-van Hal et al., 2005; Hoffmann et al., 2008; Jaworska, Gabbert, & Aldenberg, 2010)."*

Noteworthy, testing strategies from pharmaceutical industry do not help a lot. They try to identify an active compound (the future drug) out of thousands of substances, without regard to what they miss—but this approach is unacceptable in a safety ITS. Pharmacology screening also typically starts with a target, i.e. a mode of action, while toxicological assessments need to be open to various mechanisms, some as yet uncharacterized, until we have a comprehensive list of relevant pathways of toxicity (Hartung and McBride, 2011).

Due to its origin from alternative methods and REACH, ITS discussions are much more predominant in Europe (Hartung, 2010b). However, they resonate in principle very strongly with the US approach of toxicity testing in the 21st century (Tox-21c) (Hartung, 2009). The latter suggests moving regulatory toxicology to mechanisms (the pathways of toxicity, PoT). This means breaking the hazard down to its modes of action and combining it with chemico-physical properties (including QSAR) and PBPK models. This implies in similar ways that different pieces of evidence and tests are strategically combined.

Consideration 2: The need for a definition of ITS

The currently best reference for definitions of terminology is provided by OECD guidance document 34 on validation (OECD, 2005). An extract of the most relevant definitions is given in box 1. Notably, (integrated) test strategy is not defined.

Following a series of ECVAM internal meetings, an ECVAM/EPAA workshop was held to address this (Kinsner-Ovaskainen et al., 2009) and came up with a working

definition: *"As previously defined within the literature, an ITS is essentially an information-gathering and generating strategy, which in itself does not have to provide means of using the information to address a specific regulatory question. However, it is generally assumed that some decision criteria will be applied to the information obtained, in order to reach a regulatory conclusion. Normally, the totality of information would be used in a weight-of-evidence (WoE) approach."* WoE had been addressed in an earlier ECVAM workshop *(Balls et al., 2006)*: *"Weight of evidence (WoE) is a phrase used to describe the type of consideration made in a situation where there is uncertainty, and which is used to ascertain whether the evidence or information supporting one side of a cause or argument is greater than that supporting the other side."* It is of critical importance to understand that WoE and ITS are two different things though they combine the same types of information! In WoE there is no formal integration, usually no strategy and often no testing. WoE is much more a *"poly-pragmatic shortcut"* to come to a preliminary decision, where there is no or limited certainty. As proponents of evidence-based toxicology (EBT) (Hoffman and Hartung, 2006), we have to admit that the term EBT further contributes to this confusion (Hartung, 2009b). However, there is obvious cross-talk between these approaches, when for example the quality scoring of studies developed for EBT (Schneider et al., 2009) helps to filter their use in WoE and ITS approaches.

The following definition was put forward by the ECVAM/EPAA workshops (Kinsner-Ovaskainen et al., 2009): ***"In the context of safety assessment, an Integrated Testing Strategy is a methodology which integrates information for toxicological evaluation from more than one source, thus facilitating decision-making. This should be achieved whilst taking into consideration the principles of the Three Rs (reduction, refinement and replacement)"***. In line with the proposal put forward in the 2007 OECD Workshop on Integrated Approaches to Testing and Assessment, they reiterated, "*a good ITS should be structured, transparent and hypothesis driven"* (OECD, 2008).

Jaworska and Hoffmann (Jaworska and Hoffmann, 2010) defined ITS somewhat differently: "*In narrative terms,* ***ITS can be described as combinations of test batteries covering relevant mechanistic steps and organized in a logical, hypothesis-driven decision scheme, which is required to make efficient use of generated data and to gain a comprehensive information basis for making decisions regarding hazard or risk****. We approach ITS from a system analysis perspective and understand them as decision support tools that synthesize information in a cumulative manner and that guide testing in such a way that information gain in a testing sequence is maximized. This definition clearly separates ITS from tiered approaches in two ways. First, tiered approaches consider only the information generated in the last step for a decision as, for example, in current regulated sequential testing strategy for skin irritation (OECD TG 405, 2002) or the recently proposed* in vitro *testing strategy for eye irritation* (Laurie Scott et al., 2010)*. Secondly, in tiered testing strategies the sequence of tests is prescribed, albeit loosely, based on average biological relevance and is left to expert judgment. In contrast, our definition enables an integrated and systematic approach to guide testing such that the sequence is not necessarily prescribed ahead of time but is tailored to the chemical-specific situation. Depending on the already available information on a specific chemical the sequence might be adapted and optimized for meeting specific information targets."*

It might be useful to start from the scratch with our definitions to get around some glitches.

- The leading principle should be that a test gives one result, and it does not matter how many endpoints (measurements) the test requires. Figure 1 shows these different scenarios. A **test / assay** thus consists of a test system (biological *in vivo* or *in vitro* model) and a Standard Operation Protocol (SOP) including endpoint(s) to measure, reference substance(s), data interpretation procedure (a way to express the result), information on reproducibility / uncertainty, applicability domain / information on limitations and favorably performance standards. Note, that tests can include

multiple test systems and/or multiple endpoints as long as they lead to one result.

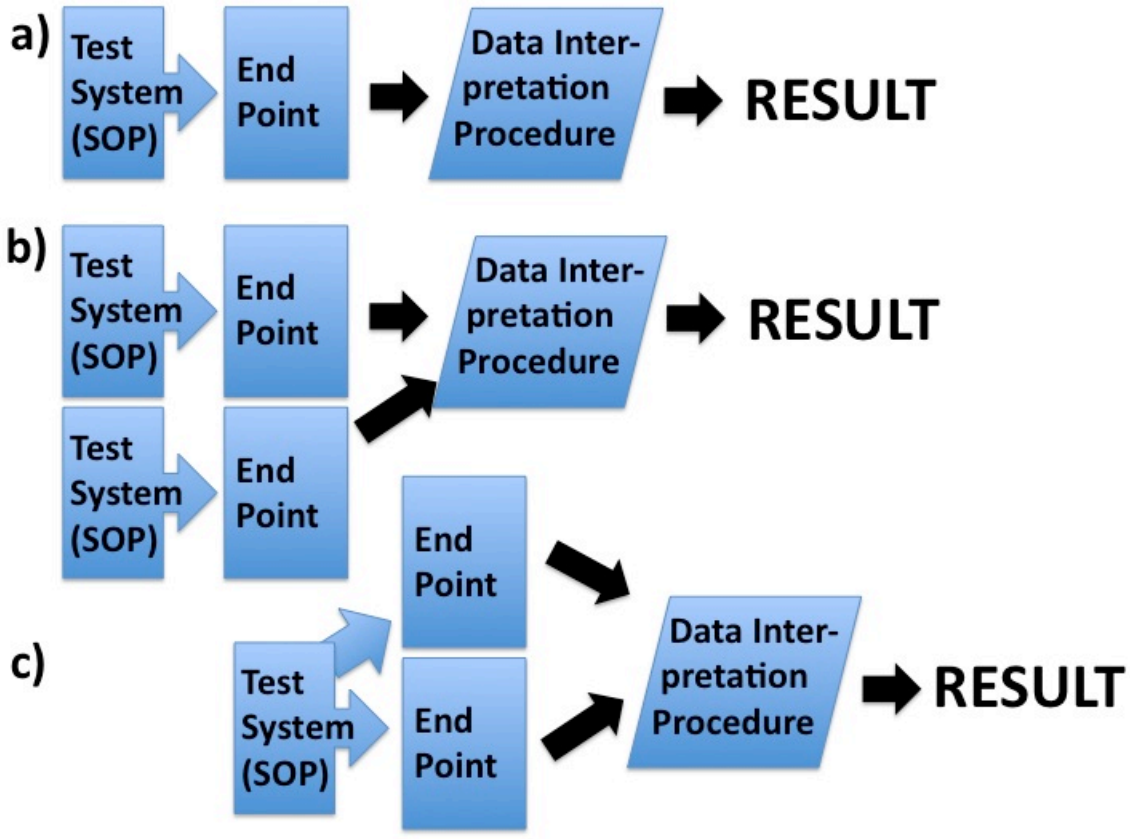


Figure 1: Three prototypic tests

Three prototypic tests, i.e. (a) a simple test with one endpoint, (b) two test systems giving a joint result and (c) multiple endpoints (including omics and other high-content analysis)

- An **integrated test strategy** is an algorithm to combine (different) test result(s) and possibly non-test information (existing data, *in silico* extrapolations from existing data or modeling) to give a combined test result. They often will have interim decision points on which further building blocks to consider.
- A **battery of tests** is a group of tests, which complement each other but are not integrated in a strategy. A classical example is the genotoxicity testing battery.

- **Tiered testing** describes the simplest ITS, where a sequence of tests is defined without formal integration of results.
- A **probabilistic TS** describes an ITS, where the different building blocks change the probability for a test result.
- **Validation** of a test or an ITS requires a **prediction model** (a way to translate it to the point of reference) and the **point of reference** itself, which can be correlative on the basis of results or mechanistic.

Some of these aspects are shown in Figure 2.

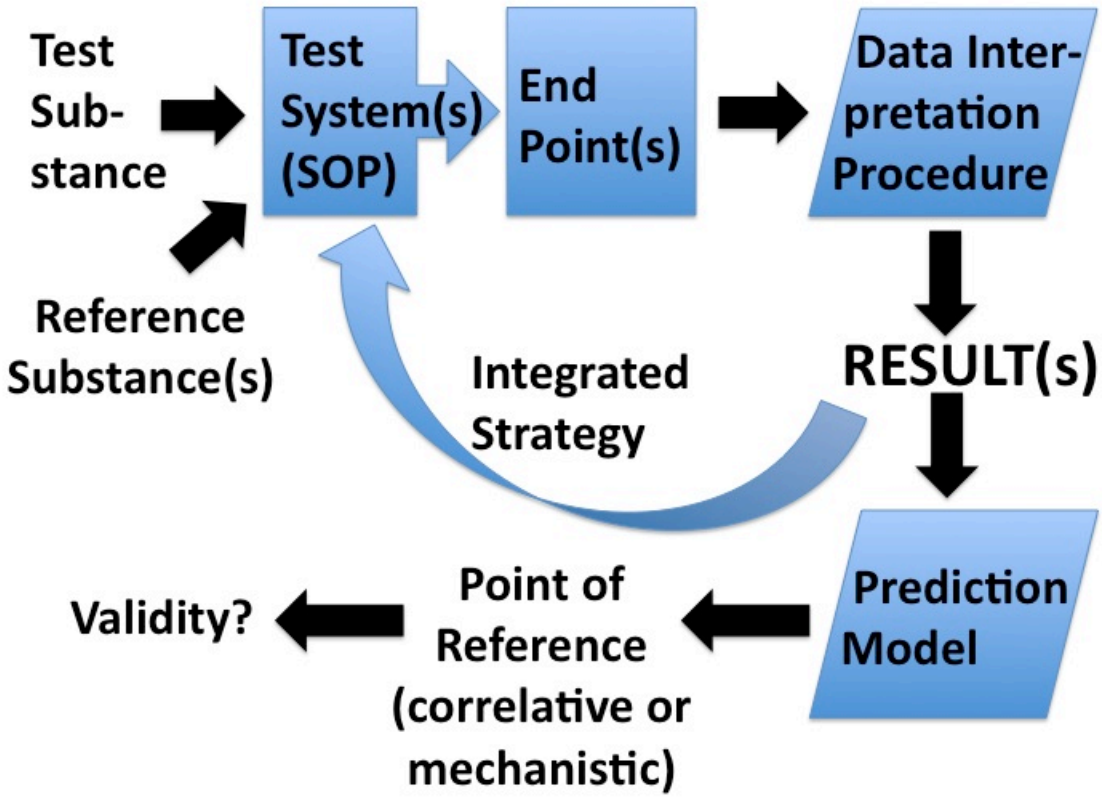


Figure 2: Components of a test (strategy) and its traditional (correlative) or mechanistic validation

Consideration 3: Composition of ITS – no GOBSATT!

The ITS in use to date is based on consensus processes often called *"weight of evidence"* (WoE) approaches. Such *"Good old boys sitting around the table"* (GOBSATT) is not really a way forward to compose ITS. The complexity of data and

the multiplicity of performance aspects to consider (costs, animal use, time, predictivity etc.) (Gabbert & Benighaus, 2012; Nordberg, Rudén, & Hansson, 2008)call for simulation based on test data. Shortcomings of existing ITS were recently analyzed in detail by (Jaworska et al., 2010): *"Though both current ITS and WoE approaches are undoubtedly useful tools for systemizing chemical hazard and risk assessment, they lack a consistent methodological basis for making inferences based on existing information, for coupling existing information with new data from different sources, and for analyzing test results within and across testing stages in order to meet target information requirements"* and in more detail in (Jaworska & Hoffmann, 2010): *"The use of flow charts as the ITS' underlying structure may lead to inconsistent decisions. There is no guidance on how to conduct consistent and transparent inference about the information target, taking into account all relevant evidence and its interdependence. Moreover, there is no guidance, other than purely expert-driven, regarding the choice of the subsequent tests that would maximize information gain."* A pioneering example of ITS evaluation, focused on skin irritation, has been provided by Hoffmann et al. (Hoffmann et al., 2008). They compiled a database of 100 chemicals. A number of strategies, both animal-free and inclusive of animal data were constructed and subsequently evaluated considering predictive capacities, severity of misclassifications and testing costs. Noteworthy, the different ITS to be compared where "hand-made", i.e. based on scientific reasoning and intuition, but not any construction principles. They correctly conclude: *"To promote ITS, further guidance on construction and multi-parameter evaluation need to be developed."* Similarly, the ECVAM/EPAA workshop only stated needs (Kinsner-Ovaskainen et al., 2009): *"So far, there is also a lack of scientific knowledge and guidance on how to develop an ITS, and in particular, on how to combine the different building blocks for an efficient and effective decision-making process. Several aspects should be taken into account in this regard, including:*

— the extent of flexibility in combining the ITS components;

— the optimal combination of ITS components (including the minimal number of components and/or combinations that have a desired predictive capacity);

— the applicability domain of single components and the whole ITS; and

— the efficiency of the ITS (cost, time, technical difficulties)"

Using this "wish list" as guidance some aspects shall be discussed.

***Extent of flexibility in combining the ITS components**:* This is a key dilemma—any validation "puts tests into stone" and "freezes them in time" (Hartung, 2007). An ITS is, however, so much larger than individual tests that there is even more reasons for change (technical advances, limitations of individual ITS components for the given substance to study, availability of all tests in a given setting etc.). What is needed here is a measure of similarity of tests and performance standards. The latter concept was introduced in the modular approach to validation (Hartung et al., 2004) and is now broadly used for the new validations. It defines what criteria a "me-too" development (a term borrowed from pharmaceutical industry, where a competitor follows the innovative, pioneering work of another company introducing a compound with the same work principle) has to fulfill to be considered equivalent to the original one. The idea is that this meant to avoid undertaking again a full-blown validation ring trial with its enormous resources. There is some difference in interpretation, whether this still needs to be multi-laboratory exercise to establish also inter-laboratory reproducibility and transferability. Noteworthy, this requires demonstrating the similarity of tests, for which we have no real guidance. However, it also implies that any superiority of the new test compared to the originally validated one cannot be shown. For ITS components, in the same way similarity and performance criteria need to be established to allow exchange for something different without a complete reevaluation of the ITS. This can first be based on the scientific relevance and the PoT covered as argued earlier (Hartung, 2010b). This means that two assays, which cover the same mechanism can substitute for each other. Alternatively, it can be based on correlation of results. Two assays, which agree (concordance) to a sufficient degree, can be considered similar. We might call these two options *"mechanistic similarity"* and *"correlative similarity".*

***The optimal combination of ITS components**:* The typical combination of building blocks so far is following a Boolean logic, i.e., the logical combinations are AND, OR and NOT. Table 1 gives the different examples for combining two tests with dichotomous (plus/minus) outcome with such logic and the consequences for the joint applicability domain and the validation need. Noteworthy, in most cases the validation of the building blocks will suffice, but the joint applicability domain will just be the overlap of the two tests' applicability domains. This is a simple application of set theory. Only if the two tests measure the same but for different substances / substance severity classes, the logical combination OR results in the combined applicability domain. If the result requires that both tests are positive, e.g. when a screening tests and a confirmatory test are combined, it is necessary to validate the overall ITS outcome.

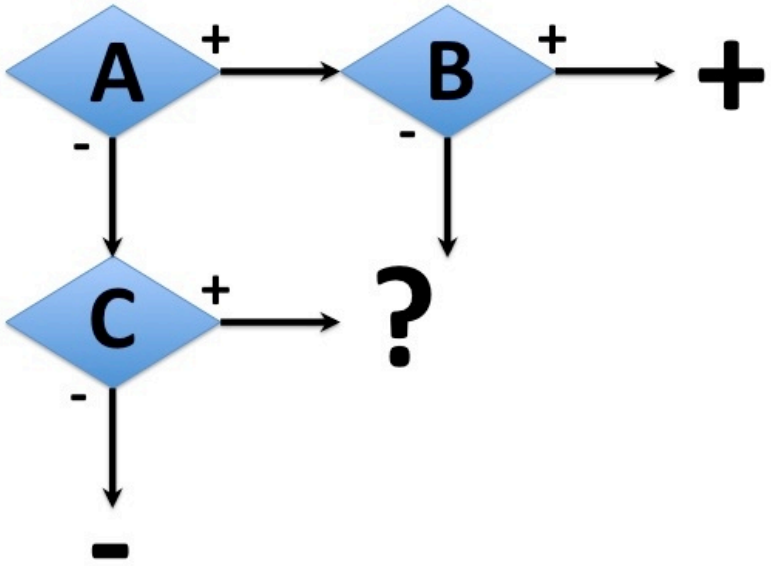
The principal opportunities in combining tests into the best ITS lie, however, in interim decision points (Figure 3 is showing a simple example, where the positive or negative outcome is confirmed). Here, the consequences for the joint applicability domain are more complex and typically only the overall outcome can be validated. The other opportunity is combining tests not with Boolean logic but with fuzzy / probabilistic logic. This means that the result is not dichotomous (toxic or not) but a probability or score is assigned. We could say that a value in-between 0 (non-toxic) and 1 (toxic) is assigned. Such combinations will typically only allow use in the overlapping applicability domains. It also implies that only the overall ITS can be validated. The challenge lies here mostly in the point of reference, which normally needs to be graded and not dichotomous as well.

The advantages of a probabilistic approach were recently summarized by Jaworska and Hoffmann (2010): "*Further, probabilistic methods are based on fundamental principles of logic and rationality. In rational reasoning every piece of evidence is consistently valued, assessed and, coherently used in combination with other pieces of evidence. While knowledge- and rule-based systems, as manifested in current testing strategy schemes, typically model the expert's way of reasoning, probabilistic systems describe dependencies between pieces of evidence (towards an information target)*

within the domain of interest. This ensures the objectivity of the knowledge representation. Probabilistic methods allow for consistent reasoning when handling conflicting data, incomplete evidence, and heterogeneous pieces of evidence."

Test combinations and consequences for applicability domain and validation needs			
Logic	**Example**	**Joint Applicability Domain**	**Validation Need**
Boolean			
A AND B	**Screening plus confirmatory test**	**Overlap**	**Total ITS**
A OR B	**Different Mode of Action**	**Overlap**	**Building Blocks**
	Different Applicabilty Domain or Severity Grades	**Combined**	**Building Blocks**
A NOT B	**Exclusion of a property (such as cytotoxicity)**	**Overlap**	**Total ITS**
IF A positive: B **IF A negative: C** **See Figure 1**	**Decision points, here confirmation of result in a second test**	**Combined overlap A/B and overlap A/C**	**Total ITS**
Fuzzy / Probabilistic			
p(A, B) **i.e. probability as function of A and B**	**Combined change of probability, e.g. priority score**	**Overlap**	**Building Blocks**

Appendix II, Table 12: Boolean versus Fuzzy Logic



Chapter II, Figure 3: Illustration of a simple decision tree, where outcomes of test A are confirmed by different second tests B or C

The applicability domain of single components and the whole ITS*:* Simple logic shows as discussed above that in most instances an ITS can only be applied where all building blocks applied to a substance allow so. Only if the combination serves exactly the purpose of expanding the applicability domain (by combining two tests with OR) the picture changes. However, this implies that essentially the same thing is measured (i.e. similarity of tests); if tests differ in applicability domain and what they measure, a hierarchy needs to be established first. This is one of the key arguments for flexibility of ITS as we need to exchange building blocks for others to meet the applicability domain for a given substance.

The efficiency of the ITS*:* Typically, here resources such as cost and labor are referred to. However, animal use and suffering is outside of this equation. It is a societal decision how to value the replacement of an animal test. In the EU legislation, the term *"reasonably available"* is used to mandate the use of an alternative (Hartung, 2010a). This leaves room for interpretation but there are

certainly limits: How much more costly can an alternative method be to be reasonably available? And the cost/benefit calculation needs to include also societal acceptability. However, this is missing the point: Especially for safety assessments it centers in the end around predicting human health and environmental effects. What are the costs of a test versus the risk of a scandal? However, if we only attempt to be as good as the animal test, this argument has no leverage. We thus need to advance to human relevance if we really want to impact. This is difficult on the level of correlation, because we typically do not have the human data for a statistically sufficient number of substances. However, we have more and more the mechanisms relevant to human health effects. Thus, the efficacy to cover relevant mechanisms for human health and environmental effects is becoming increasingly important. I have called this *"mechanistic validation"* (Hartung, 2007). This does require that we establish causality for a given mechanism to create a health or environmental effect. The classical frameworks of the Koch-Dale (Dale, 1929) and Bradford Hill (Hill, 1965) principles for assessing evidence of causation come to mind first. Dale translated the Koch postulates for a pathogen to cause a certain disease to a mediator (at the time histamine and neurotransitter) of a physiological effect. We have recently applied this to systematically evaluate the nature of the Gram-positive bacterial endotoxin (Hartung, 2012). We can similarly translate to a PoT being responsible for the manifestation of an adverse cellular outcome of substance X:

- Evidence for presence of the PoT in affected cells
- Perturbation / activation of the PoT leads to or amplifies the adverse outcome
- Hindering PoT perturbation / activation diminishes manifestation of the adverse outcome
- Blocking the PoT once perturbed / activated PoT diminishes manifestation of the adverse outcome

Please note that the current debate whether a PoT represents a chemico-biological interaction impacting on the biological system or the perturbed normal physiology is reflected in using both terminologies.

Similarly, the Bradford-Hill criteria can be applied:

- Strength: The stronger an association between cause and effect the more likely a causal interpretation, but a small association does not mean that there is not a causal effect.
- Consistency: Consistent findings of different persons in different places with different samples increase the causal role of a factor and its effect.
- Specificity: The more specific an association is between factor and effect, the bigger the probability of a causal relationship.
- Temporality: The effect has to occur after the cause.
- Biological gradient: Greater exposure should lead to greater incidence of the effect with the exception that it can also be inverse, meaning greater exposure leads to lower incidence of the effect.
- Plausibility: A possible mechanism between factor and effect increases the causal relationship, with the limitation that knowledge of the mechanism is limited by best available current knowledge.
- Coherence: A coherence between epidemiological and laboratory findings leads to an increase in the likelihood of this effect. However, the lack of laboratory evidence cannot nullify the epidemiological effect on the associations.
- Experiment: Similar factors that lead to similar effects increase the causal relationship of factor and effect.

Most recently, a new approach to causation was proposed originating from ecological modeling (Sugihara et al., 2012; Marshall, 2012). Whether this offers an avenue for systematically testing causality in large datasets from omics and/or high-throughput testing needs to be explored. It might represent an alternative to choosing meaningful biomarkers (Blaauboer et al., 2012), being always limited to the current state of knowledge.

As a more pragmatic approach, DeWever et al. (De Wever et al., 2012) suggested key elements of an ITS:

"(1) Exposure modelling to achieve fast prioritisation of chemicals for testing, as well as the tests which are most relevant for the purpose. Physiologically based pharmacokinetic modelling (PBPK) should be employed to determine internal doses in blood and tissue concentrations of chemicals and metabolites that result from the administered doses. Normally, in such PBPK models, default values are used. However, the inclusion of values or results form in vitro *data on metabolism or exposure may contribute to a more robust out-come of such modelling systems.*

(2) Data gathering, sharing and read-across for testing a class of chemicals expected to have a similar toxicity profile as the class of chemicals providing the data. In vitro *results can be used to demonstrate differences or similarities in potency across a category or to investigate differences or similarities in bioavailability across a category (e.g. data from skin penetration or intestinal uptake).*

(3) A battery of tests to collect a broad spectrum of data focussing on different mechanisms and mode of actions. For instance changes in gene expression, signalling pathway alterations could be used to predict toxic events which are meaningful for the compound under investigation.

(4) Applicability of the individual tests and the ITS itself has to be assured. The acceptance of a new method depends on whether it can be easily transferred from the developer to other labs, whether it requires sophisticated equipment and models, or if intellectual property issues and the costs involved are important. In addition, an accurate description of the compounds that can and cannot be tested is essential in this context.

(5) Flexibility allowing for adjustment of the ITS to the target molecule, exposure regime or application.

(6) Human-specific methods should be prioritised whenever possible to avoid species differences and to eliminate 'low dose' extrapolation. Thus, the in vitro *methods of choice are based upon human tissues, human tissue slices or human primary cells and cell lines for* in vitro *testing. If* in vivo *studies be unavoidable, transgenic animals should be the preferred choice if available. If not, comparative genomics (animal*

versus human) and computational models of kinetics and dynamics in animals and humans may help to overcome species differences."

This "shopping list" extends ITS from hazard identification to exposure considerations and the inclusion of existing data beyond *de novo* testing (including some quite questionable approaches of read-across and forming of chemical classes, for which no guidance and quality assurance is yet available). It similarly calls for flexibility, a key difference to current guidance document from ECHA or OECD. Compared to REACH it calls for human predictivity and mode-of-action information in the sense of *Toxicity Testing for the 21st Century*. Similarly, an earlier report also based on an IVTP symposium, to which the author contributed, made further recommendations more along a concept based on pathways of toxicity (Berg et al., 2011): *"When selecting the battery of* in vitro *and* in silico *methods addressing key steps in the relevant biological pathways (the building blocks of the ITS) it is important to employ standardized and internationally accepted tests. Each block should be producing data that are reliable, robust and relevant (the alternative 3R elements) for assessing the specific aspect (e.g. biological pathway) it is supposed to address. If they comply with these elements they can be used in an ITS."*

An important additional consideration was made by Hoffmann et al. (Hoffmann et al., 2008): "*Furthermore, the study underlined the need for databases of chemicals with testing information to facilitate the construction of practical testing strategies. Such databases must comprise a good spread of chemicals and test data in order that the applicability of approaches may be effectively evaluated. Therefore, the (non-) availability of data is a caveat at the start of any ITS construction. Whilst* in silico *and* in vitro *data may be readily generated,* in vivo *data of sufficient quality are often difficult to obtain."* This comes back again to both the need for data-sharing (D. Basketter et al., 2012) and the construction of a point of reference for validation exercises(Hoffmann et al., 2008).

The most comprehensive framework for ITS composition so far was produced by Jaworska and Hoffmann as a t[4] commissioned white-paper see (Jaworska & Hoffmann, 2010) also (Jaworska et al., 2010)**:**

"ITS should be:

a) ***Transparent and consistent***

– As a new and complex development, key to ITS, as to any methodology, is the property that they are comprehensible to the maximum extent possible. In addition to ensuring credibility and acceptance, this may ultimately attract the interest needed to gather the necessary momentum required for their development. The only way to achieve this is a fundamental transparency.

– Consistency is of similar importance. While difficult to achieve for weight of evidence approaches, a well-defined and transparent ITS can and should, when fed with the same, potentially even conflicting and/or incomplete information, always (re-)produce the same results, irrespective of who, when, where, and how it is applied. In case of inconsistent results, reasons should be identified and used to further optimize the ITS consistency.

– In particular, transparency and consistency are of utmost importance in the handling of variability and uncertainty. While transparency could be achieved qualitatively, e.g. by appropriate documentation of how variability and uncertainty were considered, consistency in this regard may only be achievable when handled quantitatively.

b) ***Rational***

– Rationality of ITS is essential to ensure that information is fully exploited and used in an optimized way. Furthermore, generation of new information, usually by testing, needs to be rational in the sense that it is focused on providing the most informative evidence in an efficient way.

c) ***Hypothesis-driven***

– ITS should be driven by a hypothesis, which will usually be closely linked to the information target of the ITS, a concept detailed below. In this way the efficiency of an ITS can be ensured, as a hypothesis-driven approach offers the flexibility to adjust the hypothesis whenever new information is obtained or generated.

… Having defined and described the framework of ITS, we propose to fill it with the following five elements:

1. Information target identification;

2. Systematic exploration of knowledge;

3. Choice of relevant inputs;

4. Methodology to evidence synthesis;

5. Methodology to guide testing"

The reader is referred to the original article (Jaworska and Hoffmann, 2010) and its implementation for skin sensitization (Jaworska et al., 2011).

Consideration 4: Guidance from testing strategies in clinical diagnostics

We have earlier stressed the principal similarities of a diagnostic and a toxicological test strategy (Hoffmann and Hartung, 2005). In both cases, different sources of information have to be combined to come to an overall result. Vecchio pointed out already in 1966 the problem of single tests in unselected populations (Vecchio, 1966) leading to unbearable false-positive rates. Systematic reviews of an evidence-based toxicology (EBT) approach (Hoffman and Hartung, 2006; Hartung 2009b) and meta-analysis could serve the evaluation and quality assurance of toxicological tests. The frameworks for evaluation of clinical diagnostic tests are well developed (Deeks, 2001) (Devillé et al., 2002; Leeflang et al., 2008) and led to the Cochrane Handbook for Diagnostic Test Accuracy Reviews (Anon, 2011). Deville et al. (Devillé et al., 2002) give very concise guidance how to evaluate diagnostic methods. This is closely linked to efforts to improve reporting on diagnostic tests; a set of minimal

reporting standards for diagnostic research has been proposed: Standards for Reporting of Diagnostic Accuracy statement (STARD) [http://www.consort-statement.org/]. We have argued earlier that this represents an interesting approach to complement or substitute for traditional method validation (Hartung, 2010b). Deeks et al. (Deeks, 2001) summarize their experience as follows [with translation to toxicology inserted in brackets]: *"Systematic reviews of studies of diagnostic* [hazard assessment] *accuracy differ from other systematic reviews in the assessment of study quality and the statistical methods used to combine results. Important aspects of study quality include the selection of a clinically relevant cohort* [relevant test set of substances], *the consistent use of a single good reference standard* [reference data], *and the blinding of results of experimental and reference tests. The choice of statistical method for pooling results depends on the summary statistic and sources of heterogeneity, notably variation in diagnostic thresholds* [thresholds of adversity]. *Sensitivities, specificities, and likelihood ratios may be combined directly if study results are reasonably homogeneous. When a threshold effect exists, study results may be best summarised as a summary receiver operating characteristic curve, which is difficult to interpret and apply to practice."*

Interestingly, Schunemann et al. (Holger J Schünemann, 2008) developed GRADE for grading quality of evidence and strength of recommendations for diagnostic tests and strategies. This framework uses *"patient-important outcomes"* as measures, in addition to test accuracy. A less invasive test can be better for a patient even if it does not give the same certainty. Similarly, we might frame our choices by aspects such as throughput, costs or animal use.

Consideration 5: The many faces of (I)TS for safety assessments

As defined earlier, any systematic combination of different (test) results represents a testing strategy. It does not really matter of these results already exist, are estimated from structures or related substances, measured by chemico-physical methods or stem from testing in a biological system or from human observations

and studies. Jaworska et al. (Jaworska et al., 2010)and Basketter et al. (D. Basketter et al., 2012) list many of the more recently proposed ITS. One of the authors (THA) had the privilege to coordinate from the side of the European Commission the ITS development within the guidance for REACH implementation for industry, which formed the basis for current ECHA guidance (http://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment). Classical examples, some of them commonly used without the label ITS, in toxicology are:

Test battery of genotoxicity assays: Several assays (3-6) depending on the field of use (Hartung, 2008) are carried out and typically any positive result is taken as an alert. They are often combined with further mutagenicity testing *in vivo* (Hartung, 2010a). The latter is necessary to reduce the tremendous rate of false-positive classifications of the battery as discussed earlier (Basketter et al., 2012). Interestingly, Aldenberg and Jaworska (Aldenberg & Jaworska, 2010) applied a Bayesian network to the dataset assembled by Kirkland et al. showing the potential of a probabilistic network to analyze such datasets.

ITS for eye and skin irritation: As already mentioned, these were the first areas to introduce internationally accepted ITS, though relatively simple, e.g. suggesting a pH test before progressing to corrosivity testing. The rich data available from six International validation studies, eight retrospective assessments and three recently completed validation studies of new tests (Adler et al., 2011; Zuang, Eskes, Griesinger, & Hartung, 2007) makes it an ideal test case for ITS development. For ocular toxicity, since 2002 the OECD TG 405 has provided an ITS approach for eye irritation and corrosion. In spite of this TG, the Office of Pesticide Programs (OPPs) of the US EPA, requested the development of an *in vitro* eye irritation strategy to register anti-microbial cleaning products. The Institute for In-Vitro Sciences in collaboration with industry partners developed such an ITS of three *in vitro* approaches, which was then accepted by regulators (De Wever et al., 2012). ITS development has very much advanced around this test case (McNamee et al., 2009; L. Scott et al., 2010).

For skin irritation, we already referred to the work by Hoffmann et al. (Hoffmann et al., 2008), which was based on an evaluation of the prevalence of this hazard among new chemicals (HOFFMANN et al., 2005). The study showed the potential of simulations to guide ITS construction.

Embryonic Stem Cell test (EST) – an ITS? The EST (Marx-Stoelting et al., 2009; Seiler & Spielmann, 2011; Spielmann et al., 2006) is an interesting test case for our definition of an ITS. It consists of two test systems (mouse embryonic stem cells and 3T3 fibroblasts) and two endpoints (cell differentiation into beating cardiomyocytes and cytotoxicity in both cell systems). The result (embryotoxicity), however, is only deduced from all this information. According to the suggested definition of tests and ITS, therefore, this represents a test and not an ITS. Noteworthy, the EST formed a key element of the ITS developed at the end of the Integrated Project ReProTect (Hareng, Pellizzer, Bremer, Schwarz, & Hartung, 2005); a final feasibility study showed the tremendous potential of this approach (Schenk et al., 2010).

Skin sensitization: The area has been subject to intense work over the last decade, which resulted in about 20 test systems. As outlined in the roadmap process (Basketter et al., 2012), the area now requires the creation of an ITS. It seems that only the gridlock of the political decision process on the 2013 deadline, which includes skin sensitization as an endpoint, hinders the finalization of this important work. Since at the same time this represents a critical endpoint for REACH (notably all chemicals under REACH require at the moment a local lymph node assay for skin sensitization), such delays are hardly acceptable. It is very important that BASF has pushed the area by submitting their ITS (Mehling et al., 2012) for ECVAM evaluation already. Pioneering work to develop a Bayesian ITS for this hazard was referred to earlier (Jaworska et al., 2011).

In silico *ITS*: There are also attempts to combine only various *in silico* (QSAR) approaches. We have discussed some of the limitations of the *in silico* approaches in isolation earlier (Hartung and Hoffmann, 2009). Since they are referred to in REACH as *"non-testing methods"* they might actually be called *"Integrated Non-Testing*

Strategies" (INTS). An example for bioaccumulation, already earlier proposed to suit ITS (De Wolf et al., 2007; Ahlers et al., 2008), was reported recently (Fernández et al., 2012) showing improved prediction by combining several QSAR.

Consideration 6: Validation of ITS

Concepts for the validation of ITS are only emerging. The ECVAM/EPAA workshop (Kinsner-Ovaskainen et al., 2009) noted only: *"There is a need to further discuss and to develop the ITS validation principles. A balance in the requirements for validation of the individual ITS components versus the requirements for the validation of a whole ITS should be considered."* Later in the text, the only statement made was: *"It was concluded that a formal validation should not be required, unless the strategy could serve as full replacement of an* in vivo *study used for regulatory purposes."* The workshop stated that for screening, hazard classification & labeling and risk assessment neither a formal validation of the ITS components nor the entire ITS is required. We would kindly disagree, as validation is certainly desirable also for other uses, but should be tailored to the use scenario and the available resources. The follow-up workshop (Kinsner-Ovaskainen et al., 2012) did not go much further with regard to recommendations for validation: *" Firstly, it was agreed that the validation of a partial replacement test method (for application as part of a testing strategy) should be differentiated from the validation of an* in vitro *test method for application as a stand-alone replacement. It was also agreed that any partial replacement test method should not be any less robust, reliable or mechanistically relevant than stand-alone replacement methods. However, an evaluation of predictive capacity (as defined by its accuracy when predicting the toxicological effects observed* in vivo*) of each of these test methods would not necessarily be as important when placed in a testing strategy, as long as the predictive capacity of the whole testing strategy could be demonstrated. This is especially the case for test methods for which the relevant prediction relates to the impact of the tested chemical on the biological pathway of interest (i.e. biological relevance). The extent to which (or indeed how) this*

biological relevance of test methods could, and should, be validated, if reference data (a 'gold standard') were not available, remained unclear.

Consequently, a recommendation of the workshop was for ECVAM to consider how the current modular approach to validation could be pragmatically adapted for application to test methods, which are only used in the context of a testing strategy, with a view to making them acceptable for regulatory purposes.

Secondly, it was agreed that ITS allowing for flexible and ad hoc approaches cannot be validated, whereas the validation of clearly defined ITS would be feasible. However, even then, current formal validation procedures might not be applicable, due to practical limitations (including the number of chemicals needed, cost, time, etc).

Thirdly, concerning the added value of a formal validation of testing strategies, the views of the group members differed strongly, and a variety of perspectives were discussed, clearly indicating the need for further informed debate. Consequently, the workshop recommended the use of EPAA as a forum for industry to share case studies demonstrating where, and how, in vitro *and/or integrated testing strategies have been successfully applied for safety decision-making purposes. Based on these case studies, a pragmatic way to evaluate the suitability of partial replacement test methods could be discussed, with a view to establishing conditions for regulatory acceptance and to reflect on the cost/benefit of formal validation, i.e. the confirmation of scientific validity of a strategy by a validation body and in line with generally accepted validation principles, as provided in OECD Guidance Document 34 (OECD, 2005).*

Finally, the group agreed that test method developers should be encouraged to develop and submit to ECVAM, not only tests designed as full replacements of animal methods, but also partial replacements in the context of a testing strategy."

Somewhat going further, De Wever et al. 2012 noted: *"In some cases, the assessment of predictive capacity of a single building block may not be as important, as long as the predictive capacity of the whole testing strategy is demonstrated. However, … the*

predictive capacity of each single element of an ITS and that of the ITS as a whole needs to be evaluated".

Berg et al. go further challenging validation need and suggesting a more hands-on approach to gain experiences (Berg et al., 2011): *"Does it make sense to validate a strategy that builds upon tests for hazard identification which change over time, but is to be used for risk assessment? One needs to incorporate new thinking into risk assessment. Regulators are receptive to new technologies but concrete data are needed to support their use. Data documentation should be comprehensive, traceable and make it possible for other investigators to retrieve information as well as reliably repeat the studies in question regardless of whether the original work was performed to GLP standards."*

What is the problem? If we follow the traditional approach of correlating results, we need a good coverage of each branch of the ITS with suitable reference substances to establish correct classification. However, even for these very simple stand-alone tests we are often limited by the low number of available well characterized reference compounds and how much testing we can afford. However, such an approach would only be valid for static ITS anyway and would lose all the flexibility of exchanging building blocks. The opportunity lies in the earlier suggested *"mechanistic validation"*. If we can agree that a certain building block covers a certain relevant mechanism, we might relax our validation requirements and also accept as equivalent another test covering the same mechanism. This does not blunt the need for reproducibility assessments, but a few pertinent toxicants relevant to humans should suffice to show that we at least identify the liabilities of the past. The second way forward is to stop making any test a "game-changer": If we accept that each and every test only changes probabilities of hazard, we can relax and fine-tune the weight added with each piece of evidence "on the road". It appears that such probabilistic hazard assessment also should be ideally compatible with probabilistic PBPK modeling and probabilistic exposure modeling (Van Der Voet & Slob, 2007). This is the tremendous opportunity of probabilistic hazard and risk assessment (Thompson & Graham, 1996) (Hartung et al., 2012a).

Consideration 7: Challenges ahead

Regulatory acceptance: A key recommendation from the ECVAM/EPAA workshop (Kinsner-Ovaskainen et al., 2009) was: *"It is necessary to initiate, as early as possible, a dialogue with regulators and to include them in the development of the principles for the construction and validation of ITS."* An earlier OECD workshop in 2008 (OECD, 2008) made some first steps and posed some of the most challenging questions:

- *how these tools and methods can be used in an integrated approach to fulfil the regulatory endpoint, independent of current legislative requirements;*
- *how the results gathered using these tools and methods can be transparently documented; and*
- *how the degree of confidence of using them can be communicated throughout the decision making process.*

With impressive crowd-sourcing of about 60 nominated experts and three case studies, a number of conclusions were reached:

- *There is limited acceptability for use of structural alerts to identify effects. Acceptability can be improved by confirming the mode of action (e.g.* in vitro *testing,* in vivo *information from an analogue or category).*
- *There is a higher acceptability for positive (Q)SAR results compared to negative (Q)SAR results (except for aquatic toxicity).*
- *The communication on how the decision to accept or reject a (Q)SAR result can be based on the applicability domain of a (Q)SAR model and/or the lack of transparency of the (Q)SAR model.*
- *The acceptability of a (Q)SAR result can be improved by confirming the mechanism/mode of action of a chemical and using a (Q)SAR model applicable for that specific mechanism/mode of action.*
- *Read-across from analogues can be used for priority setting, classification & labelling and risk assessment.*

- *The combination of analogue information and (Q)SAR results for both target chemical and analogue can be used for classification & labelling and risk assessment for acute aquatic toxicity if the target chemical and the analogue share the same mode of action and if the target chemical and analogue are in the applicability domain of the QSAR.*
- *Confidence in read-across from a single analogue improves if it can be demonstrated that the analogue is likely to be more toxic than the target chemical or if it can be demonstrated that the target chemical and the analogue have similar metabolisation pathways.*
- *Confidence in read-across improves if experimental data is available on structural analogues "bracketing" the target substance. The confidence is increased with an increased number of "good" analogues that provide concordant data.*
- *Lower quality data on a target chemical can be used for classification & labelling and risk assessment if it confirms an overall trend over analogues and target.*
- *Confidence is reduced in cases where robust study summaries for analogues are incomplete or inadequate.*
- *It is difficult to judge analogues with missing functional groups compared to the target; good analogues have no functional group compared to the target and when choosing analogues, other information on similarity than functional groups is requested.*

Taken together, these conclusions address more a WoE approach and the use of non-testing information than actually ITS. They still present important information on the comfort zone of regulators and how to handle such information for inclusion into ITS. Noteworthy, the questions of documentation and expressing confidence were not tackled.

Flexibility by determining the Most Valuable (next) Test: A key problem is to break out of the rigid test guideline principles of the past. ITS must not be forced

into a scheme with yearlong debate of expert consensus and committees. Too often, technological changes to components, difficulties with availability and applicability of building blocks and case-by-case adaptations for the given test sample will be necessary. For example, the integration of existing data, obviously at the beginning of an ITS, already creates a very different starting point. Chemico-physical, structural properties (including read-across or chemical category assignments) and prevalence will also change the probability of risk even before the first tests are applied. In order to maintain the desired flexibility in applying an ITS, at each moment the MVT (most valuable test) to follow needs to be determined. Such an approach should have the following features:

1. Assess finally the probability of toxicity from the different test results.

2. Determine most valuable next test given from previous test results and other information.

3. Have a measure of model stability (e.g. confidence intervals) and robustness.

Assessing the probability of toxicity for given tests can be done by a machine learning tools. Generative models work best for providing the values needed to find a most valuable test given prior tests. One simple generative model would predict probability of toxicity using a discriminative model (e.g. Random Forest), and test probability via a generative model (e.g. Naive Bayes). A classifier for determining risk of chemical toxicity must have the following traits:

- Outputs unbiased and consistent probability estimates for toxicity (e.g. by cross-validation).
- Outputs probability estimates even when missing certain results (both Random Forests and Naive Bayes can handle missing values).
- Reliable and stable results based on cross-validation measures.

The MVT identification based on previous tests is not a direct consequence of building a toxicity probability estimator. To find MVTs we need a generative model capable of determining test probabilities. One simple and effective way to determine

the MVT is via the same method that decision trees use, i.e. an iterative process of determining, which tests gives the most 'information' on the endpoint. Information gain can be calculated given a generative model. To determine the test, which gives the most information, we can find the test that yields the greatest reduction in Shannon entropy. This is basically a measure that quantifies information as a function of the probability of different values for a test and the impact those values have on the endpoint category (toxic vs. non-toxic). The mathematical formula is:

$$H(T) = -\sum_{i=0}^{n} p(T_i) * p(toxicity|T_i) * log(p(toxicity|T_i))$$

Where T is the test in question and *p*(Ti) signifies the probability of a test taking on one of its values (enumerated by *i*). To determine the most valuable test we need not only the toxicity classifier, but we need probability estimates for every test as a function of all other tests. To determine these transition probabilities we need to discretize every test into the *n* buckets shown in the above equation.
We can expect that users applying this model would want to determine probabilities of toxicity for their test item within some risk threshold in the fewest number of test steps or minimizing the costs. When we start testing for toxicity we may want to check on the current level of risk before deciding on more testing. For example we might decide to stop testing if a test item has less than 10% chance of being toxic or a greater than 90% chance. Finding MVTs from a generative model has an advantage over directly using decision trees. Unfortunately, decision trees cannot handle sparse data effectively. The amount of data needed to determine *n* tests increases exponentially with the number of tests. By calculating MVTs on top of a generative model we can leverage a simple calculation from a complex model that is not as heavily constrained by data size.

Combining the ITS concept with Tox-21c: As discussed above, Tox-21c relies on breaking risk assessment down in many components. These need to be put together again in a way to allow decision-taking, ultimately envisioned as Systems Toxicology

by simulation (Hartung et al., 2012). Before this, an ITS-like integration and possibly a probabilistic condensation of evidences into a probability of risk are the logical approaches. However, there are special challenges: Most importantly, the technologies promoted by Tox-21c, at this stage mainly omics and high-throughput (Hattis, 2009), are very different to the information sources promoted in the European ITS discussion. However, we see how the ITS discussion is crossing the Atlantic—for example, in the context of endocrine disruptor testing (Willett et al., 2011). They are so data-rich that from the beginning that a data-mining approach is necessary, which means that the weighing of evidence is left to the computer. Not all regulators are comfortable with this.

Our own research is approaching this for metabolomics (Bouhifd et al., in press) using endocrine disruption as test case might illustrate some of the challenges of the high-throughput, systems biology methods and "-omics" technologies. Metabolomics—defined as measuring the concentration of 'all' low molecular weight (1500 Da) molecules in a system of interest—is the closest "-omics" technology to the phenotype and represents the upstream consequences of whatever changes are observed in proteomic or transcriptomic studies. Small changes in the concentration of a protein, which might be undetectable at the level of transcriptomics or proteomics, can result in large changes in the concentrations of metabolites, and changes which are often invisible at one level of analysis (i.e. co-factor regulation of an enzyme), are more likely to be apparent in a metabolic profile. By taking a global view, metabolomics provides clues to the systemic response to a challenge from a toxin and does so in a way that provides both mechanistic information and candidates for biomarkers (Griffin, 2006; Robertson et al., 2011). In other words, metabolomics offers both the possibility of seeing the high-level pattern of altered biological pathways while drilling down for relevant mechanistic details.

Metabolomics produces many of the same challenges as other high-content methods—namely, how to integrate the surfeit of data into a meaningful framework—but at the same, it has some unique challenges. In particular,

metabolomics lacks the large-scale, integrated databases that have been crucial to the analysis of transcriptomic and proteomic data. Like the early years of microarrays, there are still no established methods to interpret data. Exploring data sets via several methods (Sugimoto, Kawakami, Robert, Soga, & Tomita, 2012) (ORA, QEA, correlation analysis, and genome-scale network reconstruction) will hopefully provide some guidance for future toxicological applications for metabolomics and help better understand the puzzle as well as develop and provide new perspectives on how to integrate several '-omics' technologies. At some level, metabolomics remains at this stage one of hypothesis generation and potentially, biomarker discovery, and as such will be dependent on validation by other means.

One critical problem for metabolomics is that while a more-or-less complete "parts list" and wiring diagrams exist for genomic and proteomic networks, knowledge of metabolic networks is still relatively incomplete. Currently, there are three non-tissue specific genome-scale human metabolic networks: Recon 1 (Rolfsson, Palsson, & Thiele, 2011) (Rolfsson et al., 2011), the Edinburgh Human Metabolic Network (EHMN) (Ma et al., 2007), and HumanCyc (Romero et al., 2004). These reconstructions are "first-drafts": in addition to genes and proteins of unknown function as well as "dead-end" or "orphaned" metabolites which are not associated with specific anabolic or catabolic pathways. Furthermore, the networks are not tissue-specific. Many toxicants, including endocrine disruptors, exhibit tissue-specific toxicity, and a cell or tissue-specific metabolic network (Hao, Ma, Zhao, & Goryanin, 2012) should provide a more accurate model of pathology than a generic, global human metabolic network. Longer-term, a well-characterized, biochemically complete network will help make the leap from pathway identification to a parameterized model than can be used for more complex simulations such as metabolic control analysis, flux analysis, and systems control theory to understand the wiring diagram that allows the cell to maintain homeostasis and where, within that wiring diagram, there are vulnerabilities.

Steering the new developments: At this stage, no strategic planning and coordination for the challenge of ITS implementation exists. This was noticed in

most of the meetings do far, e.g., (Berg et al., 2011): *"…there was a clear call from the audience for a credible leadership with the capacity to assure alignment of ongoing activities and initiation of concerted actions, e.g. a global human toxicology project."* The Human Toxicology Project Consortium (http://htpconsortium.wordpress.com) is one of the advocates for such steering (Seidle and Stephens, 2009). There is still quite a road to go (Hartung, 2009a). While we aim to establish some type of coordinating center in the US at Johns Hopkins (working title PoToMaC—Pathway of Toxicity Mapping Center), no such effort is yet in place in Europe. We have been suggesting the creation of a *European Safety Sciences Institute* (ESSI) in our policy program, but this discussion is only starting. It is, however, evident that we need such structures for developing the new toxicological toolbox and a global collaboration of regulators of the different sectors to finally revamp regulatory safety assessments.

Acknowledgement

The support by NIH transformative research grant "Mapping the Human Toxome by Systems Toxicology" (RO1ES020750) and FDA grant "DNTox-21c Identification of pathways of developmental neurotoxicity for high throughput testing by metabolomics" (U01FD004230) is gratefully appreciated.

Box 1

RELEVANT DEFINITIONS FROM OECD SERIES ON TESTING AND ASSESSMENT
OECD SERIES ON TESTING AND ASSESSMENT Number 34

GUIDANCE DOCUMENT ON THE VALIDATION AND INTERNATIONAL ACCEPTANCE OF NEW OR UPDATED TEST METHODS FOR HAZARD ASSESSMENT

Adjunct test: Test that provides data that add to or help interpret the results of other tests and provide information useful for the risk assessment process

Assay: Uses interchangeably with Test.

Data interpretation procedure (DIP): An interpretation procedure used to determine how well the results from the test predict or model the biological effect of interest. See Prediction Model.

Decision Criteria: The criteria in a test method protocol that describe how the test method results are used for decisions on classification or other effects measured or predicted by the test method.

Definitive test: A test that is considered to generate sufficient data to determine the specific hazard or lack of hazard of the substance without the need for further testing, and which may therefore be used to make decisions pertaining to hazard or safety of the substance.

Hierarchical (tiered) testing approach: An approach where a series of tests to measure or elucidate a particular effect are used in an ordered sequence. In a typical hierarchical testing approach, one or a few tests are initially used; the results from these tests determine which (if any) subsequent tests are to be used. For a particular chemical, a weigh-of-evidence decision regarding hazard could be made at any stage (tier) in the testing strategy, in which case there would be no need to proceed to subsequent tiers.

In silico models: Approaches for the assessment of chemicals based on the use computer-based estimations or simulations. Examples include structure-activity relationships (SAR), quantitative structure-activity relationships (QSARs), and expert systems.

(Q)SARs (Quantitative Structure-Activity Relationships): Theoretical models for making predictions of physicochemical properties, environmental fate parameters, or biological effects (including toxic effects in environmental and mammalian species). They can be divided into two major types, QSARs and SARs. QSARs are quantitative models yielding a continuous or categorical result while SARs are qualitative relationships in the form of structural alerts that incorporate molecular substructures or fragments related to the presence or absence of activity.

A screen/screening test is often a rapid, simple test method conducted for the purpose of classifying substances into a general category of hazard. The results of a screening test generally are used for preliminary decision making in the context of a testing strategy (i.e., to assess the need for additional and more definitive tests). Screening tests often have a truncated response range in that positive results may be considered adequate to determine if a substance is in the highest category of a hazard classification system without the need for further testing, but are not usually adequate without additional information/tests to make decisions pertaining to lower levels of hazard or safety of the substance

Test (or assay): An experimental system used to obtain information on the adverse effects of a substance. Used interchangeably with assay.

Test battery: A series of tests usually performed at the same time or in close sequence. Each test within the battery is designed to complement the other tests and generally to measure a different component of a multi-factorial toxic effect. Also called base set or minimum data set in ecotoxicological testing.

Test method: A process or procedure used to obtain information on the characteristics of a substance or agent. Toxicological test methods generate information regarding the ability of a substance or agent to produce a specified biological effect under specified conditions. Used interchangeably with “test” and “assay”.

APPENDIX III Pathways of Toxicity Workshop Report

(*Originally published as*: Kleensang, A., Maertens, A., Rosenberg, M., Fitzpatrick, S., Lamb, J., Auerbach, S., ... & Hartung, T. (2014). t4 workshop report: pathways of toxicity. *Altex*, *31*(1), 53)

Pathways of Toxicity

Andre Kleensang[1,,], Alexandra Maertens[1,,], Michael Rosenberg[2], Suzanne Fitzpatrick[3], Justin Lamb[4], Scott Auerbach[5], Richard Brennan[6], Kevin M. Crofton[7], Ben Gordon[8], Albert J. Fornace Jr.[9], Kevin Gaido[3], David Gerhold[10], Robin Haw[11], Adriano Henney[12], Avi Ma'ayan[13], Mary McBride[2], Stefano Monti[14], Michael F. Ochs [15], Akhilesh Pandey[16], Roded Sharan[17], Rob Stierum[18], Stuart Tugendreich[19], Catherine Willett[20], Clemens Wittwehr[21], Jianguo Xia[22], Geoffrey W. Patton [23], Kirk Arvidson[23], Mounir Bouhifd[1], Helena T. Hogberg[1], Thomas Luechtefeld[1], Lena Smirnova[1], Liang Zhao[1], Yeyejide Adeleye[24], Minoru Kanehisa[25], Paul Carmichael[24], Melvin E. Andersen[26], and Thomas Hartung[1]

[1]Johns Hopkins University, Bloomberg School of Public Health, Center for Alternatives to Animal Testing, Baltimore, MD, USA; [2]Agilent Technologies, Inc., Santa Clara, CA, USA; [3]US Food and Drug Administration, Center for Food Safety & Applied Nutrition, College Park, MD, USA; [4]Genometry Inc, Cambridge, MA, USA; [5]Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA; [6]Thomson Reuters Inc., Carlsbad, CA, USA; [7]U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, Research Triangle Park, NC, USA; [8]Dept of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; [9]Dept. of Biochemistry and Molecular & Cellular Biology, and Lombardi Comprehensive Cancer Center, Georgetown

[,,] authors contributed equally

University, Washington, DC, USA; [10]The National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, MD, USA; [11]Community Outreach Manager Reactome, Ontario Institute for Cancer Research Toronto, Canada; [12]The German Virtual Liver Network, University of Heidelberg, Heidelberg, Germany; [13] Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [14]Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA; [15]Johns Hopkins University, School of Medicine, Department of Oncology, Baltimore, MD, USA; [16]Institute of Genetic Medicine and Departments of Biological Chemistry, Oncology and Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [17]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel; [18]TNO Healthy Living, Microbiology & Systems Biology, Zeist, The Netherlands; [19]Ingenuity Systems, Inc., Redwood City, CA, USA; [20]The Humane Society of the United States, Washington, DC, USA; [21]European Commission, Joint Research Centre, Systems Toxicology Unit, Ispra, Italy; [22]Department of Microbiology and Immunology, University of British Columbia, Vancouver, Canada; [23]U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, Office of Food Additive Safety, College Park, MD, USA; [24]Unilever, Safety & Environmental Assurance Centre, Colworth Science Park, Sharnbrook, Bedfordshire, UK; [25]Institute for Chemical Research, Kyoto University, Uji, Kyoto, Japan; [26]The Hamner Institutes for Health Sciences, Research Triangle Park, NC, USA; [27]CAAT-Europe, University of Konstanz, Germany

There is a goal,
but no way;
what we call a way is
hesitation.

Franz Kafka
(Kafka, 1931, p 230)

Summary

Despite wide-spread consensus on the need to transform toxicology and risk assessment in order to keep pace with technological and computational changes that have revolutionized the life sciences, there remains much work to be done to achieve the vision of toxicology based on a mechanistic foundation. To this end, a workshop was organized to explore one key aspect of this transformation—the development of Pathways of Toxicity as a key tool for hazard identification based on systems biology. Several issues were discussed in depth in the workshop:

The first was the challenge of formally defining the concept of a Pathway of Toxicity (PoT), as distinct from, but complementary to, other toxicological pathway concepts such as mode of action (MoA). The workshop came up with a preliminary definition of PoT as "A molecular definition of cellular processes shown to mediate adverse outcomes of toxicants". It is further recognized that normal physiological pathways exist that maintain homeostasis and these, sufficiently perturbed, can become PoT. Second, the workshop sought to define the adequate public and commercial resources for PoT information, including data, visualization, analyses, tools, and use-cases, as

well as the kinds of efforts that will be necessary to enable the creation of such a resource.

Third, the workshop explored ways in which systems biology approaches could inform pathway annotation, and which resources are needed and available that can provide relevant PoT information to the diverse user communities.

1. Introduction

The "Toxicology in the 21st Century" (Tox-21c) movement, initiated with the 2007 NRC report (Krewski et al., 2010; National Research Council, 2007), has stirred the toxicological community (T. Hartung, 2008; Hartung, 2009, 2011; Hartung & Leist, 2008) and initiated a far-reaching discussion about current practices in risk assessment and possible avenues for advancement. A critical overview of the extensive dialog that ensued after the publication of the report has been summarized by Andersen and Krewski (Andersen & Krewski, 2010). Within a few years the discussion has moved from whether the field of toxicology should change to discussions on how and when to do so—from the call for a Human Toxicology Project (Seidle & Stephens, 2009; http://www.humantoxicologyproject.org) to the ongoing programs by the US federal agencies (R. S. Judson et al., 2010; Knudsen et al., 2011) and the redefinition of the EPA toxicity-testing paradigm (Firestone, Kavlock, Zenick, Kramer, & Testing, 2010).

The United States Food and Drug Administration (FDA) has recently embraced this strategy (Hamburg, 2011):

> *"We must bring 21st century approaches to 21st century products and problems. Toxicology is a prime example. Most of the toxicology tools used for regulatory assessment rely on high-dose animal studies and default extrapolation procedures and have remained relatively unchanged for decades, despite the scientific revolutions of the past half-century. We need better predictive models to identify concerns earlier in the product development process to reduce time and costs. We also need to modernize the tools used to assess emerging concerns about potential risks from food and other product exposures. … With an advanced field of regulatory science, new tools, including functional genomics, proteomics, metabolomics, high-throughput screening, and systems biology, can replace*

current toxicology assays with tests that incorporate the mechanistic underpinnings of disease and of underlying toxic side effects. This should allow the development, validation, and qualification of preclinical and clinical models that accelerate the evaluation of toxicities during drug development. … Ultimately, investments in regulatory science can lead to a new era of progress and safety. Because such investments will promote not only public health but also the economy, job creation, and global economic competitiveness, they have major implications for the nation's future."

We could not summarize it better.

The key proposal of Tox-21c is straightforward: we have to base regulatory toxicology (for environmental chemicals, because this was the mandate of the National Academy of Sciences panel) on mechanism and mode of action. The term "toxicity pathways" was coined in the NRC report and later the term "*Pathway of Toxicity*" (PoT) was created by Hartung and colleagues (Hartung, 2009, 2011). OECD uses *adverse outcome pathway* in the context of their QSAR Toolbox and ecotoxicology (Ankley et al., 2006) and recently published a proposal for a template, and guidance on developing and assessing the completeness of adverse outcome pathways as a draft document (OECD, 2012). This is in line with the science of toxicology moving toward a more complete mechanistic understanding. There have already been some tentative efforts to identify and describe PoT. One component of the Tox-21 alliance formed by US EPA (ToxCast), the NIEHS (within the National Toxicology Program), NIH Chemical Genomics Center (the high-throughput testing program) and FDA (the Critical Path Initiative), is focused on use of HTS data to facilitate and test PoT[3].

The limitations of the existing paradigm are well known. Hazard assessment based on animal testing has limited throughput achieved at a high cost; if traditional tests are applied to the backlog of existing chemicals of concern for which there is limited safety data, the costs would be enormous and, even if that were not an obstacle, the capacity is simply not there (see e.g. Hartung & Rovida, 2009; Rovida &

[3] http://epa.gov/ncct/Tox21

Hartung, 2009; Seok et al., 2013). Furthermore, while the continued or expanded use of animal testing has become more and more objectionable to the general public, as well as to many in the toxicology community, there is at the same time a public mandate to perform more thorough hazard assessment and testing for industrial chemicals (e.g., European REACH legislation), not to mention the demands of the drug and consumer industry. New types of products—such as nanomaterials—that will likely play a large role in our economic future require a more sophisticated hazard assessment paradigm (Hartung, 2010b). The necessary practice of high-dose to low-dose extrapolation is both imprecise and often results in an overly cautious approach.

To foster the ideas of the NRC report, in Oct 2012 the Center for Alternatives to Animal Testing supported by the Doerenkamp-Zbinden Foundation, Zurich, Switzerland, and Unilever held a workshop on "Pathways of Toxicity" that discussed the concept of PoT as well as defining the necessary associated tools, standards, and core competencies. The three-day workshop brought together a diverse group of more than 30 front-line researchers and experts from academia (e.g., Universities in Boston, Alberta, Tel-Aviv and Johns Hopkins University in Baltimore), independent research institutes (TNO Netherlands and The Hamner Institutes for Health Sciences), industry (e.g., Agilent and Unilever), non-governmental organizations (e.g., The Humane Society of the US), systems biology/toxicology content and tool providers (e.g., KEGG, Thomson Reuters, WikiPathways, Reactome, Ingenuity Systems, Genometry), and the regulatory professionals that employ toxicology studies and data analysis tools to protect public health (e.g., NIH & NIEHS, US EPA, US FDA, European Commission). This report presents the conclusions and perspectives from that conference. We outline the possible benefits of mapping PoT, clarify the meaning and definition of PoT, complemented by a thorough discussion of the usefulness and validation of a public PoT database. Finally, we discuss the future challenges and directions including the idea of the creation of a PoT consortium.

2. What are the benefits of mapping PoT?

Toxicology, like the rest of biology, is undergoing a shift from a reductionist approach to a more system-oriented view that takes advantage of the newer, high-content and high-throughput technologies (van Vliet, 2011). The opportunity to move away from the limited mechanistic information provided by traditional animal tests to a pathway-based approach that provides detailed, specific mechanistic understanding at a cellular level, predictive for target organ toxicities in a causal (ideally dose dependent) manner, presents both challenges and opportunities (Hartung & McBride, 2011; Hartung et al., 2012b). As part of this challenge, the production of a comprehensive list of all PoT—that is, the "Human Toxome"—would be of great benefit. This concept is based on the assumption that the number of PoT is finite, and that, once mapped, toxicology can move towards more certainty while sharply reducing and eventually eliminating the need for animal testing [see also section IV].

Pathway-based approaches for toxicity testing require different methods for extrapolations. With animal testing, an expensive, two-year animal assay may establish, for example, that a 6 ppm dose exposure concentration is a point-of-departure for specific adverse responses. For non-cancer effects, this in-life point-of-departure would be divided by various uncertainty factors to arrive at a "safe" dose. Linear low-dose modeling would be used with carcinogens to estimate a dose associated with some level of risk (e.g., 1/100,000 or 1/1,000,000). With a PoT approach, the point-of-departure will arise from observations in the *in vitro* test batteries that provide greater multi-dose concentration response curves. These *in vitro* PoDs will be adjusted using *in vitro-in vivo* extrapolation (Rotroff et al., 2010; Wetmore et al., 2012) and there will be a need for computational pathway models (Bhattacharya, Zhang, Carmichael, Boekelheide, & Andersen, 2011) to derive proposed "safe doses" depending on characteristics of the pathway architecture. These pathway approaches will link dose and dynamics—especially at low doses—

and will show a clear causal linkage between initiating event and adverse outcome that should be useful both for setting safe doses as well as identifying biomarkers.

Lastly, it is necessary to move toxicology away from an approach that extrapolates from rodents, and instead uses a human-tissue based approach; this necessitates by definition understanding toxicological mechanisms at the cellular and pathway level, jointly with *in vitro* to *in vivo* extrapolations of dose levels. Ultimately, a pathway-based approach that uses human tissue informed by a deeper mechanistic understanding of toxicity, as well as mechanistic understanding of human disease decreases uncertainty in decision-making.

As an example of the existing problems that face regulators when testing a substance with current approaches, consider the dilemma posed by negative results: there is always the possibility that a different dosing scheme, different species or other experimental variation might yield very different results. The uncertainty only increases when we consider that animals might have a defense mechanism not present in humans or in sensitive populations, like newborns, who, for instance, lack a functional blood-brain barrier for chemicals. Conventionally however, we assume that with some additional measures (high dose, species selection, more than one species, structural alerts, etc.) we can overcome this problem. However, a more definitive answer could be given if we had a complete list of human relevant PoT and a corresponding validated test battery. Then we could, for the first time, be reasonably confident that a substance does not trigger/perturb relevant PoT. Similarly, we can establish concentrations of substances (*in vitro* no-effect levels—$NOEL_{in\ vit}$) where no PoT is triggered. It is important to note that the triggering of a PoT does not necessarily indicate harm, but a *potential* for harm.

- **Box 1: How should a PoT database be designed?**

- Main target should be regulatory context and quality of the
- database.
- PoT should be grounded at the molecular and cellular level
- It should include:

    structured and hierarchal vocabulary of adverse events, MoA and pathway description,

    spatial and temporal effects,

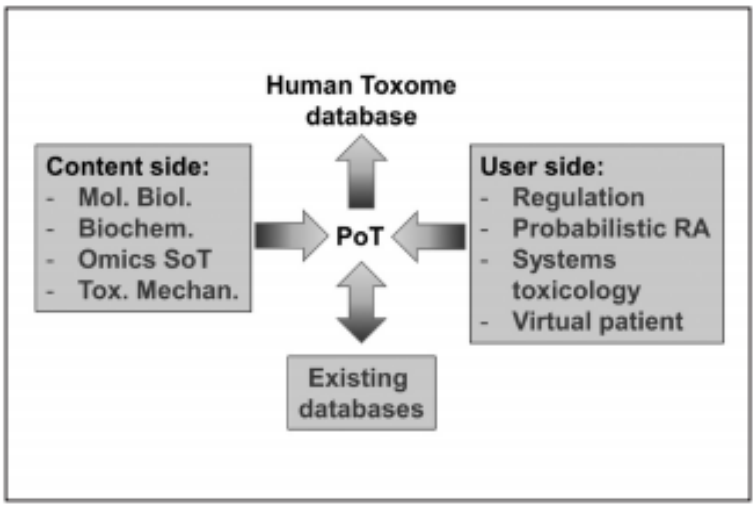
    dose-response: thresholds of adversity are essential,

    links to evidence and raw data needed

    both machine and human readable,

    quality assurance/validation summaries based on evidence-based toxicology principles, and

    interfaces with other databases (e.g., WikiPathways) for import/export



Chapter IV, Figure 1: Possible Structure of a PoT Database

3. What gap could a PoT database fill that is not met by existing databases?

As -omics technologies have increasingly added to our knowledge of biology, there has been a proliferation of pathway oriented databases such as KEGG, WikiPathways, Reactome etc., so the question might be asked, is there really a need for another pathway database?

Participants identified several needs unmet by currently available resources [see also Figure 1 and Box 1]:

Firstly, existing databases do not focus on toxicology-related pathways—any approach that uses "off-the-shelf" pathway annotations such as KEGG focuses on highly conserved pathways and may be missing much that is of interest to toxicologists.

Secondly, many toxicology related databases, such as the T3 (Toxin and Toxin Targets) database (Lim et al., 2010), have extensive documentation on various toxins and their biological targets, but the information is not available in a manner that facilitates a systems-based approach. For example, informative descriptions are

often provided as free-text which is not machine readable, and does not use a structured vocabulary or ontology that describes mechanism and targets. While ontologies exist for certain outcomes (for example, the commercial CCNet's ToxWiz ontology for histopathology[4]), no comprehensive, agreed upon, open-access ontology currently exists for toxicology especially at the molecular level, although a few are in development (for example, eTox, Cases, Pastor, & Sanz, 2013)[5]. An ontology is defined as a "formal, explicit specification of a shared conceptualization" (Gruber, 1993, p 199-200). An ontology provides both a shared controlled vocabulary—a collection of precisely defined terms—and an explicit model of the properties and relations between those terms (Hardy et al., 2012a, 2012b). Although ontologies may seem somewhat academic, most people use them everyday—whether as a library card catalog or the more specialized ontologies, such as GO (Gene Ontology), SNOMED-CT (Systemized Nomenclature of Medicine – Clinical Terms) or MeSH (Medical Subject Headings, US National Library of Medicine). Although toxicology has, in recent years, seen a vast increase in the availability of databases (e.g., ToxRefDB, Chemical Effects in Biological Systems, Comparative Toxicogenomics Database) and datasets (e.g., ToxCast), the lack of commonly agreed upon ontology and structured vocabulary has held back both data-sharing and data-mining. One key to transforming data into knowledge is the use of an ontology to provide structure and access to the data. Fortunately the toxicology community need not start from scratch but can build on existing ontologies such as SNOMED, MeDRA (the Medical Dictionary for Regulatory Activities), ChEBI (for chemicals) and GO.

Thirdly, existing databases do not "connect the head to the tail"—that is to say, they are not comprehensive from initiating event to adverse outcome. Lastly, one of the concerns unique to toxicology (and specifically, regulatory toxicology) is having certainty with respect to negative results; absence of evidence is not the same as evidence of absence, and the database user must be able to distinguish "no effect" from "no evidence". Unlike databases such as KEGG that focus on

[4] ToxWiz, http://toxwiz.com/ (accessed 12 June 2013)
[5] eTOX, http://www.etoxproject.eu/ (accessed 12 June 2013)

comprehensive coverage of biological processes, a PoT database does not have to offer global coverage. Instead, it can focus on relevant pathways that are both curated and quality-controlled for the specific needs of the regulatory community and toxicology researchers. Having a strong emphasis on quality-control does not preclude acting as a more general repository of data useful for data-mining—a PoT can be low confidence, but depending on the consequences of the decision for the regulator (or the interest of a researcher) could still offer useful information. Ultimately, it is hoped that a PoT database will function both as a data repository for the research community and a knowledge-base that regulators can rely upon.

Participants agreed as well that, ideally, the database should be constructed to allow easy answers to inquiries that might come from researchers—(e.g., What nodes with a signaling network are suspected of being involved in endocrine disruption?) as well as from regulatory scientists looking to de-risk chemicals early in the R&D process (e.g., What nodes in a PoT have assays?). And lastly, it should be able to answer the question, "What nodes are important for regulatory purposes?"

4. What is a Pathway of Toxicity; how many PoT are there and is the number finite?

After extensive discussion, the workshop participants came up with a formal definition of a Pathway of Toxicity:

A *Pathway of Toxicity* is
a *molecular definition*
of the *cellular processes*
shown *to mediate adverse outcomes*
of toxicants.

This definition focuses our attention on understanding thoroughly the molecular mechanisms of toxicity while maintaining the emphasis on the cellular context. PoT are relevant to regulators, if and only if, we can define necessary and sufficient pathways for adverse outcomes and establish their relevance by evaluating the scientific evidence. Evidence-Based Toxicology (EBT) could serve as a framework to establish the tools necessary for validating the PoT [see also section V] (Hartung, 2010a).

It is important to keep in mind that a linear pathway is an artificial construct—all pathways are abstracted from a broader, global cellular network and therefore are, at some level, an oversimplification (Figure 2 and for an overview see e.g. Kholodenko, Yaffe, & Kolch, 2012). Nonetheless, the complexity of a network is both difficult to represent on a map and distracts from focusing on key-events. Nonetheless, it may be necessary not to think of the pathways as sharply and precisely delineated from the broader cellular network, but rather to keep in mind that a pathway representation may always be a "warm, fuzzy, cloud": that is to say, warm since the answer is close but not necessarily exact; fuzzy, since the membership of components in a pathway is graded; and a cloud, since the boundaries are not sharply defined.

There will be several challenges to refining the definition of PoT into a useful working definition—how does one choose where a pathway ends? How does a pathway-based approach refine our understanding of a dose-response dependency? Toxicological processes are both spatially and temporally dynamic—how will this be represented in a pathway-based approach?

There are other questions that will need to be addressed as evidence accumulates: are PoT perturbations of known physiological pathways? For example, proliferation is a normal process—when does one re-label it as a Pathway of Toxicity? Is it possible that certain PoT are novel pathways active only in the presence of a toxicant? Are there any PoT that are distinct pathways altogether? How many PoT can we expect to find? "132" Mel Andersen, one of the proponents of

Tox-21c and workshop organizer, often answers adding, after a pause, "As a toxicologist/risk assessor, I am accustomed to false accuracy."

At this moment, any such questions about the number and nature of PoT is a pure speculation and will have to wait for more experimental evidence. Nonetheless, the number of cellular targets and metabolic pathways is finite, and thus the number of PoT should thus be, too. Evolution cannot have left too many vulnerable points given the number of xenobiotics we are exposed to, and the astonishingly large number of healthy years we enjoy on average. We see the enormous redundancy and buffering provided via biological networks when you consider the surprising number of viable homozygous knockout mice, which often have only subtle phenotypic changes, despite lacking an entire gene. The recent finding that each human individual is null for both alleles of in excess of twenty genes, also attests to the genomes' redundancy (MacArthur et al., 2012).

One unique challenge for the PoT database will be the requirement not only to represent the PoT or their network but also the kinetics and cellular or tissue location of these events, as a PoT represents a spatio-temporal event in the cell. In this respect, it may be necessary to extend the definition of PoT to include a more quantitative model, similar to those discussed in Uri Alon's Introduction to Systems Biology (Uri Alon, 2007). From this perspective, a pathway represents not just a link between a series of nodes but instead might be thought of as a wiring diagram with components such as positive and negative feedback loops, along with quantitative information about inputs, thresholds, and outputs.

5. How to identify and validate a PoT?

Most importantly, toxicology is not alone in identifying pathways—all the life sciences are on the same quest under the label of systems biology. It is the logical next step stemming from the advent of high-content technologies (-omics), attempting to create order by identifying the underlying pathways. Therefore, we

will not have to reinvent the wheel as pathway mapping, visualization and database tools are increasingly developed in other areas of the life sciences [e.g., Cytoscape (Cline et al., 2007), PathVisio (van Iersel et al., 2008), iPath (Letunic, Yamada, Kanehisa, & Bork, 2008), CellDesigner (Funahashi, Morohashi, Kitano, & Tanimura, 2003), VANTED (Junker, Klukas, & Schreiber, 2006), IPA from Ingenuity Systems, Agilent Genespring, or MetaCore from Thomson Reuters].

As an example for primary data analysis, identification of statistically significant signatures and mapping cross-technology datasets on known pathways, the Human Toxome Consortium—which initiated this PoT workshop—is largely relying on Agilent GeneSpring software. GeneSpring is a comprehensive package that combines advanced bioinformatics tools for analysis of gene expression microarrays, NGS, LC/MS and GC/MS data with unique ability to conduct joint analysis in the context of curated or customized pathways. At the time of this writing, GeneSpring supports WikiPathways, Biocyc, Ingenuity and Metacore content, KEGG will become available later this year. Besides data normalization, QC, clustering, and statistical analyses of their primary gene expression and metabolite abundance data users can perform pathway enrichment computations that leverage multiple data types and seamlessly explore and co-analyze the results overlaid on pathway diagrams in the Pathway Architect module. Additional analysis and visualization methods tailored to specific needs of PoT projects, such as multi-omics correlation tools, will be developed soon in collaboration with members of the NIH transformative research project on “Mapping the Human Toxome by Systems Toxicology” (http://humantoxome.com).

WikiPathways (Kelder et al., 2012; Pico et al., 2008) facilitates the contribution and maintenance of pathway information by the biology community. It is an open, collaborative platform dedicated to online pathway curation. WikiPathways thus complements ongoing efforts, such as KEGG, and Reactome (see next paragraph). Building on the same MediaWiki software that powers Wikipedia, custom graphical pathway editing tool and integrated databases are included covering major small-(bio)molecule systems. The web-based format of

WikiPathways reduces the barrier for biologists (e.g., toxicologists) to participate in pathway curation. More importantly, the open, public approach of WikiPathways allows for wider participation by the entire toxicological community. This approach also shifts the bulk of peer review, editorial curation, and maintenance to the toxicological community, and as such can represent content for more peer-reviewed efforts such as Reactome or the creation of a PoT database. Efforts to use WikiPathway content/tools in the context of *in vitro* toxicology, specifically to address the use of human disease mechanisms *in silico* in the interpretation of *in vitro* toxicological data have started under the Assuring Safety Without Animal testing (ASAT) initiative for allergic contact dermatitis, hepatocellular cancer and soon to be extended with models for cholestasis.

Reactome, another valuable resource, is a freely accessible, open-source, curated and peer-reviewed biological knowledgebase of human bioreactions, pathways and processes, which serves as a platform for pathway visualization and analysis of complex experimental data sets (Croft et al., 2010). A recent extension of the Reactome data model permits the capture of normal biological pathway behavior and predicts its response to a stress like a mutational change in a protein's function or the presence of a novel small molecule in the environment, in a comprehensive and internally consistent format (Milacic et al., 2012). The Reactome data model allows for annotation of small molecules, toxicological agents, and their specific mode of action. Pathway data visualization is facilitated by the Reactome Pathway Browser, a Systems Biology Graphical Notation (SBGN)-based interface (Le Novere et al., 2009), which exploits the Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) web services (Aranda et al., 2011) to overlay molecular interaction data from external interaction databases. Overlaying interaction data from ChEMBL or Drugbank (Gaulton et al., 2012; Knox et al., 2011) databases of bioactive drug-like compounds provides an opportunity to identify protein variant-drug interactions, identify novel small molecule targets, off-target effects, or pharmaceuticals that can perturb or moderate reactions or pathways of toxicity. Reactome also provides the Functional Interaction (FI) network plug-in for

Cytoscape, which can identify gene network patterns related to diseases, including cancer (Wu & Stein, 2012). Future expansion of the Reactome pathway database and the FI network with interactions based upon Pathways of Toxicity should significantly improve coverage, enrich the functional annotations supported, and enhance the functionality of the pathway and network analyses.

MetaCore™ from Thomson Reuters (formerly GeneGo) is a commercial systems biology platform for network and pathway analysis. MetaCore includes a large manually-curated database of molecular interactions (protein-protein, compound-protein, enzyme-reaction, reaction-substrate, miRNA, etc.), and tools to flexibly reconstruct and analyze biological networks. MetaCore also contains over 800 Canonical Pathway Maps—interactive visual representations of precise molecular pathways for well-characterized and annotated biological, metabolic, disease and toxicological processes. At this time, 260 of these maps, covering a wide range of pathways relevant to toxicological and disease processes have been made freely available at http://pathwaymaps.com.

However, many of these existing pathway and network mapping tools are more suited to hypothesis generation and do not provide the necessary precision and reproducibility for predicting full dose-dependent *in vivo* toxicity in man that will be required for PoT to become a useful tool for regulators. Validating PoT will likely require a sustained, coordinated effort to generate the necessary datasets to benchmark and provide context to the scoring of PoT.

Furthermore, we will need to develop tools which are suitable for looking at systems toxicology with the aim of validating them for regulatory purposes. As part of this effort, an evidence-based toxicology collaboration (EBTC, http://www.ebtox.com) has been established, which promises to generate a partnership between agency representatives, individuals from the corporate sector, and those promoting the paradigm shift in toxicology (Zurlo, 2011). Evidence-based toxicology uses concepts learned from evidence-based medicine, mechanistic/molecular toxicology, biostatistics, and validation to bring the necessary consistency and objectivity to the process. Moreover, evidence-based

toxicology can help concisely summarize existing evidence on a specific topic so that experts and non-experts can use an EBT-assessed PoT database for decision-making in a regulatory context. Noteworthy, EBT has embarked on developing the validation concepts for 21st century tools (Hartung, 2010a; Hartung, Hoffmann, & Stephens, 2013; R. Judson et al., 2013).

6. Future challenges and directions; creation of a PoT consortium

There are many obstacles that remain before a comprehensive, PoT-based toxicology can be realized. Some of them are technological. While transcriptomics is a mature technology, metabolomics is just beginning to contribute to systems-toxicology (Bouhifd et al., 2014), and some technologies—such as phosphoproteomics—remain in their infancy (van Vliet, 2011). Furthermore, even though gene and protein networks are relatively complete for humans (Tang, Zhong, & Xie, 2012; Taylor & Wrana, 2012), such "hairball" networks tell only a limited story—it is difficult to extract complete concise pathways or to take into account dose, and spatial and temporal effects. In particular, causality with respect to predicting target organ specificity needs to be addressed (Hartung, Hoffmann, et al., 2013). It will be necessary to analyze new methodologies for determining dose-response with high-throughput, high-content data and a PoT-based approach. It may be necessary then, to bootstrap our way from what we know to what we don't know in an iterative process. The workshop participants agreed, however, that we do not need to know every detail of a pathway to use it in the context of a PoT, but we need to establish fit-for-purpose principles.

Depending on the specific PoT, it may also be necessary to address the question of what types of data will be included and how the data will be integrated. Combining datasets of transcriptomics, metabolomics and other -omics still represents a challenge, although some progress in the application of systems biology approaches to such cross-domain data integration in toxicology has already

been made (e.g. Xu et al., 2008). Integrating biomarker and epidemiology data will require new ways to turn the surfeit of existing data into useful information.

Other challenges will involve a dedicated process of consensus building in the toxicology community to develop a useful ontology and structured vocabulary to facilitate sharing information. And lastly, it will require new tools and concepts within the risk assessment community as toxicology moves away from older paradigms into a more probabilistic approach (Hartung, Luechtefeld, et al., 2013; Hartung et al., 2012a).

The creation of a PoT database will make it necessary to form and coordinate a larger consortium and linking it to the development of the necessary concepts. Central steering needs to be established, incorporating the ideas of opinion leaders and the needs of stakeholders, especially regulators who ultimately have to accept the changes derived from novel approaches (Hartung, 2009). Regulators, therefore, need a seat at the table to provide input into the processes from the very beginning. The governance of such a consortium effort needs to be established, as does the quality assurance (validation), comparison to the current approaches, and possible transition. CAAT with its partners is at the moment trying to form such a consortium to define and set up a public resource for PoT information.

The vision represented here takes advantage of new innovations afforded by our rapidly evolving understanding of systems biology and a host of molecular, informational, and computational tools. Toxicity testing today is much like cartography before the development of satellites—islands of well-described territory alongside vast oceans about which little is known; it could be said that even the extent of the unmapped territory is unknown. A mapped Human Toxome available in a PoT-database would provide the necessary perspective to bring toxicology into the 21st century.

Freeman Dyson (Princeton), and his 1995 book, *The Scientist as Rebel* said: "*The great advances in science usually result from new tools rather than from new doctrines*" (Dyson, 2006, p 805). The map of the Human Toxome available in a PoT database promises to be such a new tool.

Acknowledgements

This CAAT workshop on Pathways of Toxicity was made possible through support from Unilever and the extensive discussions and experiences from the NIH transformative research project on “Mapping the Human Toxome by Systems Toxicology” (R01ES020750) and FDA grant “DNTox-21c Identification of pathways of developmental neurotoxicity for high throughput testing by metabolomics” (U01FD004230).

REFERENCES:

Adler, S., Basketter, D., Creton, S., Pelkonen, O., van Benthem, J., Zuang, V., . . . Zaldivar, J. M. (2011). Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. *Arch Toxicol, 85*(5), 367-485

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet, 8*(6), 450-461

Ashikaga, T., Sakaguchi, H., Sono, S., Kosaka, N., Ishikawa, M., Nukada, Y., . . . Itagaki, H. (2010). A comparative evaluation of in vitro skin sensitisation tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). *Altern Lab Anim, 38*(4), 275-284

Basketter, D. A., Alepee, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., . . . Templier, M. (2014). Categorization of chemicals according to their relative human skin sensitizing potency. *Dermatitis, 25*(1), 11-21

Bauch, C., Kolle, S. N., Ramirez, T., Eltze, T., Fabian, E., Mehling, A., . . . Landsiedel, R. (2012). Putting the parts together: combining in vitro methods to test for skin sensitizing potentials. *Regul Toxicol Pharmacol, 63*(3), 489-504

Baum, L. E., & Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics, 37*(6), 1554-&

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533

Beltrame, L., Rizzetto, L., Paola, R., Rocca-Serra, P., Gambineri, L., Battaglia, C., & Cavalieri, D. (2009). Using pathway signatures as means of identifying similarities among microarray experiments. *PLoS One, 4*(1), e4128

Ben-Menahem, A. (2009). *Historical encyclopedia of natural and mathematical sciences* (1st ed.). New York: Springer.

Bolton, E. E., Wang, Y. L., Thiessen, P. A., & Bryant, S. H. (2010). PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry, Vol 4, 4*, 217-241

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., & Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr Opin Struct Biol, 14*(3), 292-299

Borrell, B. (2010). Toxicology: The big test for bisphenol A. *Nature, 464*(7292), 1122-1124

Bottini, A., & Hartung, T. (2010). The economics of animal testing. *Altex, 27*, 67-77

Bouhifd, M., Hogberg, H. T., Kleensang, A., Maertens, A., Zhao, L., & Hartung, T. (2014). Mapping the human toxome by systems toxicology. *Basic Clin Pharmacol Toxicol, 115*(1), 24-31

Breiman, L. (2003). Random Forests. *Machine Learning, 45*, 5-32

Caberoy, N. B., Alvarado, G., & Li, W. (2012). Tubby regulates microglial phagocytosis through MerTK. *Journal of neuroimmunology, 252*(1), 40-48

Cappelletti, G., Pedrotti, B., Maggioni, M. G., & Maci, R. (2001). Microtubule assembly is directly affected by MPP(+)in vitro. *Cell Biol Int, 25*(10), 981-984

Chagoyen, M., & Pazos, F. (2011). MBRole: enrichment analysis of metabolomic data. *Bioinformatics, 27*(5), 730-731

Chaouiya, C., Berenguier, D., Keating, S. M., Naldi, A., van Iersel, M. P., Rodriguez, N., . . . Helikar, T. (2013). SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst Biol, 7*, 135

Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics, 5*, 147

Chuang, J. Y., Wu, C. H., Lai, M. D., Chang, W. C., & Hung, J. J. (2009). Overexpression of Sp1 leads to p53 dependent apoptosis in cancer cells. *International Journal of Cancer, 125*(9), 2066-2076

Courey, A. J., Holtzman, D. A., Jackson, S. P., & Tjian, R. (1989). Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. *Cell, 59*(5), 827-836

Davis, S., & Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England), 23*(14), 1846-1847

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., . . . Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res, 36*(Database issue), D344-350

Dehay, B., Bove, J., Rodriguez-Muela, N., Perier, C., Recasens, A., Boya, P., & Vila, M. (2010). Pathogenic lysosomal depletion in Parkinson's disease. *J Neurosci, 30*(37), 12535-12544

Deniaud, E., Baguet, J. l., Chalard, R., Blanquier, B., Brinza, L., Meunier, J., . . . Wierinckx, A. (2009). Overexpression of transcription factor Sp1 leads to gene expression perturbations and cell cycle inhibition. *PLoS One, 4*(9), e7035

Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol, 4*(5), P3

Dimitrov, S. D., Low, L. K., Patlewicz, G. Y., Kern, P. S., Dimitrova, G. D., Comber, M. H., . . . Mekenyan, O. G. (2005). Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. *Int J Toxicol, 24*(4), 189-204

Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., & Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences, 95*(1), 5-12

Do, C. B., Tung, J. Y., Dorfman, E., Kiefer, A. K., Drabant, E. M., Francke, U., . . . Langston, J. W. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS genetics, 7*(6), e1002141

Emter, R., Ellis, G., & Natsch, A. (2010). Performance of a novel keratinocyte-based reporter cell line to screen skin sensitizers in vitro. *Toxicol Appl Pharmacol, 245*(3), 281-290

European Commission Joint Research Centre. (2013). EUR 26383 - EURL ECVAM Recommendation on the Direct Peptide Reactivity Assay (DPRA) for Skin Sensitisation Testing. Luxembourg: Publications Office of the European Union.

European Commission Joint Research Centre. (2014). EUR 26427 - EURL ECVAM Recommendation on the KeratinoSensTM Assay for Skin Sensitisation Testing. Luxembourg: Publications Office of the European Union.

European Union. (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on Cosmetic Products. *Official Journal of the European Union, L 342*, 59-209

Fan, Y., Kong, H., Shi, X., Sun, X., Ding, J., Wu, J., & Hu, G. (2008). Hypersensitivity of aquaporin 4-deficient mice to 1-methyl-4-phenyl-1, 2, 3, 6-tetrahydropyrindine and astrocytic modulation. *Neurobiology of aging, 29*(8), 1226-1236

Finney, A., & Hucka, M. (2003). Systems biology markup language: Level 2 and beyond. *Biochem Soc Trans, 31*(Pt 6), 1472-1473

Fukushima, A., Kusano, M., Redestig, H., Arita, M., & Saito, K. (2011). Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Syst Biol, 5*, 1

Gene Ontology, C., Blake, J. A., Dolan, M., Drabkin, H., Hill, D. P., Li, N., . . . Westerfield, M. (2013). Gene Ontology annotations and resources. *Nucleic Acids Res, 41*(Database issue), D530-535

Gerberick, G. F., Ryan, C. A., Kern, P. S., Dearman, R. J., Kimber, I., Patlewicz, G. Y., & Basketter, D. A. (2004). A chemical dataset for evaluation of alternative approaches to skin-sensitization testing. *Contact Dermatitis, 50*(5), 274-288

Gerberick, G. F., Ryan, C. A., Kern, P. S., Schlatter, H., Dearman, R. J., Kimber, I., . . . Basketter, D. A. (2005). Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. *Dermatitis, 16*(4), 157-202

Gerberick, G. F., Vassallo, J. D., Foertsch, L. M., Price, B. B., Chaney, J. G., & Lepoittevin, J. P. (2007). Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. *Toxicol Sci, 97*(2), 417-427

Goeman, J. J., van de Geer, S. A., de Kort, F., & van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics, 20*(1), 93-99

Goodman, J. E., Witorsch, R. J., McConnell, E. E., Sipes, I. G., Slayton, T. M., Yu, C. J., . . . Rhomberg, L. R. (2009). Weight-of-evidence evaluation of reproductive and developmental effects of low doses of bisphenol A. *Crit Rev Toxicol, 39*(1), 1-75

Gowda, H., Ivanisevic, J., Johnson, C. H., Kurczy, M. E., Benton, H. P., Rinehart, D., . . . Siuzdak, G. (2014). Interactive XCMS Online: Simplifying Advanced

Metabolomic Data Processing and Subsequent Statistical Analyses. *Anal Chem, 86*(14), 6931-6939
Griffin, J. L. (2006). The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philos Trans R Soc Lond B Biol Sci, 361*(1465), 147-161
Hartung, T. (2007). Food for thought... on cell culture. *Altex, 24*(3), 143-152
Hartung, T. (2013). Look back in anger - what clinical studies tell us about preclinical work. *Altex, 30*(3), 275-291
Hartung, T., Luechtefeld, T., Maertens, A., & Kleensang, A. (2013). Integrated testing strategies for safety assessments. *Altex, 30*(1), 3-18
Hengstler, J. G., Foth, H., Gebel, T., Kramer, P. J., Lilienblum, W., Schweinfurth, H., . . . Gundert-Remy, U. (2011). Critical evaluation of key evidence on the human health hazards of exposure to bisphenol A. *Crit Rev Toxicol, 41*(4), 263-291
Hirota, M., Kouzuki, H., Ashikaga, T., Sono, S., Tsujita, K., Sasa, H., & Aiba, S. (2013). Artificial neural network analysis of data from multiple in vitro assays for prediction of skin sensitization potency of chemicals. *Toxicol In Vitro, 27*(4), 1233-1246
Hoang, T., Choi, D. K., Nagai, M., Wu, D. C., Nagata, T., Prou, D., . . . Przedborski, S. (2009). Neuronal NOS and cyclooxygenase-2 contribute to DNA damage in a mouse model of Parkinson disease. *Free Radic Biol Med, 47*(7), 1049-1056
Huang, H., Maertens, A. M., Hyland, E. M., Dai, J., Norris, A., Boeke, J. D., & Bader, J. S. (2009). HistoneHits: a database for histone mutations and their phenotypes. *Genome Res, 19*(4), 674-681
Ivatt, R., & Whitworth, A. J. (2014). SREBF1 links lipogenesis to mitophagy and sporadic Parkinson's disease. *Autophagy, 10*(8), 34-33
Jaworska, J., Dancik, Y., Kern, P., Gerberick, F., & Natsch, A. (2013). Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. *J Appl Toxicol*
Jaworska, J., Harol, A., Kern, P. S., & Gerberick, G. F. (2011). Integrating non-animal test information into an adaptive testing strategy - skin sensitization proof of concept case. *Altex, 28*(3), 211-225
Kamburov, A., Cavill, R., Ebbels, T. M., Herwig, R., & Keun, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics, 27*(20), 2917-2918
Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30
Kevles, B. (1998). *Naked to the bone : medical imaging in the twentieth century*. Reading, Mass.: Addison-Wesley.
Kim, D., Frank, C. L., Dobbin, M. M., Tsunemoto, R. K., Tu, W., Peng, P. L., . . . Tsai, L. H. (2008). Deregulation of HDAC1 by p25/Cdk5 in neurotoxicity. *Neuron, 60*(5), 803-817
Kirkland, D., Aardema, M., Henderson, L., & Muller, L. (2005). Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens I. Sensitivity, specificity and relative predictivity. *Mutat Res, 584*(1-2), 1-256

Kirkland, D. J., Henderson, L., Marzin, D., Muller, L., Parry, J. M., Speit, G., . . . Williams, G. M. (2005). Testing strategies in mutagenicity and genetic toxicology: an appraisal of the guidelines of the European Scientific Committee for Cosmetics and Non-Food Products for the evaluation of hair dyes. *Mutat Res, 588*(2), 88-105

Klein, S., & Heinzle, E. (2012). Isotope labeling experiments in metabolomics and fluxomics. *Wiley Interdiscip Rev Syst Biol Med, 4*(3), 261-272

Klekota, J., & Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics, 24*(21), 2518-2525

Krug, A. K., Gutbier, S., Zhao, L., Poltl, D., Kullmann, C., Ivanova, V., . . . Leist, M. (2014). Transcriptional and metabolic adaptation of human neurons to the mitochondrial toxicant MPP(+). *Cell death & disease, 5*, e1222

Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol, 1*, 54

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*, 559

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics, 24*(5), 719-720

Lazebnik, Y. (2004). Can a biologist fix a radio? -- Or, what I learned while studying apoptosis, (Cancer Cell. 2002 Sep;2(3):179-82). *Biochemistry (Mosc), 69*(12), 1403-1406

Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., . . . Pulendran, B. (2013). Predicting network activity from high throughput metabolomics. *PLoS Comput Biol, 9*(7), e1003123

Lim, E., Pon, A., Djoumbou, Y., Knox, C., Shrivastava, S., Guo, A. C., . . . Wishart, D. S. (2010). T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res, 38*(Database issue), D781-786

Lipinski, C. A. (2004). Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies, 1*(4), 337-341

Maxwell, G., MacKay, C., Cubberley, R., Davies, M., Gellatly, N., Glavin, S., . . . Summerfield, V. (2014). Applying the skin sensitisation adverse outcome pathway (AOP) to quantitative risk assessment. *Toxicol In Vitro, 28*(1), 8-12

McDonald, J. C., Beck, M. H., Chen, Y., & Cherry, N. M. (2006). Incidence by occupation and industry of work-related skin diseases in the United Kingdom, 1996-2001. *Occup Med (Lond), 56*(6), 398-405

McKim, J. M., Jr., Keller, D. J., 3rd, & Gorski, J. R. (2010). A new in vitro method for identifying chemical sensitizers combining peptide binding with ARE/EpRE-mediated gene expression in human skin cells. *Cutan Ocul Toxicol, 29*(3), 171-192

McKim, J. M., Jr., Keller, D. J., 3rd, & Gorski, J. R. (2012). An in vitro method for detecting chemical sensitization using human reconstructed skin models and its applicability to cosmetic, pharmaceutical, and medical device safety testing. *Cutan Ocul Toxicol, 31*(4), 292-305

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc, 8*(8), 1551-1566

Muller, H. M., Rangarajan, A., Teal, T. K., & Sternberg, P. W. (2008). Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics, 6*(3), 195-204

Naoi, M., Maruyama, W., & Inaba-Hasegawa, K. (2012). Type A and B monoamine oxidase in age-related neurodegenerative disorders: their distinct roles in neuronal death and survival. *Current topics in medicinal chemistry, 12*(20), 2177-2188

National Research Council. (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: National Academies Press.

Natsch, A., Emter, R., & Ellis, G. (2009). Filling the concept with data: integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. *Toxicol Sci, 107*(1), 106-121

Natsch, A., Ryan, C. A., Foertsch, L., Emter, R., Jaworska, J., Gerberick, F., & Kern, P. (2013). A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J Appl Toxicol*

Nukada, Y., Miyazawa, M., Kazutoshi, S., Sakaguchi, H., & Nishiyama, N. (2013). Data integration of non-animal tests for the development of a test battery to predict the skin sensitizing potential and potency of chemicals. *Toxicol In Vitro, 27*(2), 609-618

OECD. (1992). *Test No. 406: Skin Sensitisation*: OECD Publishing.

OECD. (2010). *Test No. 429: Skin Sensitisation*: OECD Publishing.

Okada, H., Tokunaga, T., Liu, X., Takayanagi, S., Matsushima, A., & Shimohigashi, Y. (2008). Direct evidence revealing structural elements essential for the high binding ability of bisphenol A to human estrogen-related receptor-gamma. *Environmental Health Perspectives, 116*(1), 32-38

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830

Peiser, M., Tralau, T., Heidler, J., Api, A. M., Arts, J. H., Basketter, D. A., . . . Luch, A. (2012). Allergic contact dermatitis: epidemiology, molecular mechanisms, in vitro methods and regulatory aspects. Current knowledge assembled at an international workshop at BfR, Germany. *Cell Mol Life Sci, 69*(5), 763-781

Perier, C., & Vila, M. (2012). Mitochondrial biology and Parkinson's disease. *Cold Spring Harbor perspectives in medicine, 2*(2), a009332

Pollio, G., Hoozemans, J. J. M., Andersen, C. A., Roncarati, R., Rosi, M. C., van Haastert, E. S., . . . Fiorentini, A. (2008). Increased expression of the oligopeptidase THOP1 is a neuroprotective response to AÎ² toxicity. *Neurobiology of disease, 31*(1), 145-158

Qiu, Z., Norflus, F., Singh, B., Swindell, M. K., Buzescu, R., Bejarano, M., . . . Hersch, S. M. (2006). Sp1 is up-regulated in cellular and transgenic models of

Huntington disease, and its reduction is neuroprotective. *J Biol Chem, 281*(24), 16672-16680
Quackenbush, J. (2003). Genomics. Microarrays--guilt by association. *Science, 302*(5643), 240-241
Ren, Y., Liu, W., Jiang, H., Jiang, Q., & Feng, J. (2005). Selective vulnerability of dopaminergic neurons to microtubule depolymerization. *J Biol Chem, 280*(40), 34105-34112
Rovida, C., & Hartung, T. (2009). Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals—a report by the transatlantic think tank for toxicology (t (4)). *Altex, 26*(3), 187-208
Ryu, H., Lee, J., Olofsson, B. A., Mwidau, A., Dedeoglu, A., Escudero, M., . . . Ratan, R. R. (2003). Histone deacetylase inhibitors prevent oxidative neuronal death independent of expanded polyglutamine repeats via an Sp1-dependent pathway. *Proceedings of the National Academy of Sciences of the United States of America, 100*(7), 4281-4286
Sakaguchi, H., Ashikaga, T., Miyazawa, M., Yoshida, Y., Ito, Y., Yoneyama, K., . . . Suzuki, H. (2006). Development of an in vitro skin sensitization test using human cell lines; human Cell Line Activation Test (h-CLAT). II. An inter-laboratory study of the h-CLAT. *Toxicol In Vitro, 20*(5), 774-784
Sanderson, D. M., & Earnshaw, C. G. (1991). Computer prediction of possible toxic action from chemical structure; the DEREK system. *Hum Exp Toxicol, 10*(4), 261-273
Santpere, G., Nieto, M., Puig, B., & Ferrer, I. (2006). Abnormal Sp1 transcription factor expression in Alzheimer disease and tauopathies. *Neuroscience letters, 397*(1), 30-34
Schober, A. (2004). Classic toxin-induced animal models of Parkinson's disease: 6-OHDA and MPTP. *Cell Tissue Res, 318*(1), 215-224
Severin, J., Waterhouse, A. M., Kawaji, H., Lassmann, T., van Nimwegen, E., Balwierz, P. J., . . . Hayashizaki, Y. (2009). FANTOM4 EdgeExpressDB: an integrated database of promoters, genes, microRNAs, expression dynamics and regulatory interactions. *Genome Biol, 10*(4), R39
Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504
Shi, L., Jones, W. D., Jensen, R. V., Harris, S. C., Perkins, R. G., Goodsaid, F. M., . . . Fang, H. (2008). The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics, 9*(Suppl 9), S10
Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., . . . Vencio, R. Z. (2014). ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics, 30*(9), 1336-1337
Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., . . . Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. *Ther Drug Monit, 27*(6), 747-751

Snyder, S. H., & D'Amato, R. J. (1986). MPTP A neurotoxin relevant to the pathophysiology of Parkinson's disease: The 1985 George C. Cotzias Lecture. *Neurology, 36*(2), 250-250

Sterky, F. H., Hoffman, A. F., Milenkovic, D., Bao, B., Paganelli, A., Edgar, D., . . . Larsson, N. G. (2012). Altered dopamine metabolism and increased vulnerability to MPTP in mice with partial deficiency of mitochondrial complex I in dopamine neurons. *Human molecular genetics, 21*(5), 1078-1089

Steuer, R., Kurths, J., Fiehn, O., & Weckwerth, W. (2003). Interpreting correlations in metabolomic networks. *Biochem Soc Trans, 31*(Pt 6), 1476-1478

Stobbe, M. D., Swertz, M. A., Thiele, I., Rengaw, T., van Kampen, A. H., & Moerland, P. D. (2013). Consensus and conflict cards for metabolic pathway databases. *BMC Syst Biol, 7*, 50

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science, 302*(5643), 249-255

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., . . . Prokopenko, V. V. (2005). Virtual computational chemistry laboratory--design and description. *J Comput Aided Mol Des, 19*(6), 453-463

Thyssen, J. P., Johansen, J. D., & Menne, T. (2007). Contact allergy epidemics and their controls. *Contact Dermatitis, 56*(4), 185-195

Thyssen, J. P., Linneberg, A., Menne, T., & Johansen, J. D. (2007). The epidemiology of contact allergy in the general population--prevalence and main findings. *Contact Dermatitis, 57*(5), 287-299

Todeschini, R., Consonni, V., & Todeschini, R. (2009). *Molecular descriptors for chemoinformatics* (2nd, rev. and enl. ed.). Weinheim: Wiley-VCH.

van der Meer, A. D., & van den Berg, A. (2012). Organs-on-chips: breaking the in vitro impasse. *Integr Biol (Camb), 4*(5), 461-470

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *Ieee Transactions on Information Theory, 13*(2), 260-+

Voutchkova, A. M., Kostal, J., Steinfeld, J. B., Emerson, J. W., Brooks, B. W., Anastas, P., & Zimmerman, J. B. (2011). Towards rational molecular design: derivation of property guidelines for reduced acute aquatic toxicity. *Green Chemistry, 13*(9), 2373-2379

Wang, J., & Bannon, M. J. (2005). Sp1 and Sp3 activate transcription of the human dopamine transporter gene. *J Neurochem, 93*(2), 474-482

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res, 37*(Web Server issue), W623-633

Warshaw, E. M., Wang, M. Z., Mathias, C. G., Maibach, H. I., Belsito, D. V., Zug, K. A., . . . Sasseville, D. (2012). Occupational contact dermatitis in hairdressers/cosmetologists: retrospective analysis of north american contact dermatitis group data, 1994 to 2010. *Dermatitis, 23*(6), 258-268

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., . . . Scalbert, A. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res, 41*(Database issue), D801-807

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012). MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. *Nucleic Acids Res, 40*(Web Server issue), W127-133

Xia, J., & Wishart, D. S. (2010). MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res, 38*(Web Server issue), W71-77

Ye, Q., Zhang, X., Huang, B., Zhu, Y., & Chen, X. (2013). Astaxanthin suppresses MPP - induced oxidative damage in PC12 cells through a Sp1/NR1 signaling pathway. *Marine drugs, 11*(4), 1019-1034

Yip, A. M., & Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics, 8*, 22

Zhang, C. Z., Chen, G. G., & Lai, P. (2010). Transcription factor ZBP-89 in cancer growth and apoptosis. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1806*(1), 36-41

Zhong, W. X., Wang, Y. B., Peng, L., Ge, X. Z., Zhang, J., Liu, S. S., . . . Luo, J. H. (2012). Lanthionine synthetase C-like protein 1 interacts with and inhibits cystathionine beta-synthase: a target for neuronal antioxidant defense. *J Biol Chem, 287*(41), 34189-34201

Zhu, Q. S., Chen, K., & Shih, J. C. (1994). Bidirectional promoter of human monoamine oxidase A (MAO A) controlled by transcription factor Sp1. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 14*(12), 7393-7403

CURRICULUM VITAE:

Education

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Department of Environmental Health Sciences, Division of Toxicology

Ph.D.

University of California, San Diego, San Diego, CA

Post-graduate Certificate in Data Mining (April 2010)

Excelsior College, Albany, NY (2001)

B.S., English and Sociology

Research Experience

- Developed an assay for aluminum-induced iron toxicity in astrocytes
- Curated and designed a histone post-translational modification database using Pymol to visualize structural information and Textpresso for text-mining
- Analyzing transcriptomic and metabolomic data for the Mapping the Human Toxome Project
- Using weighted correlation networks to characterize dose-response relationships using microarray data
- Using machine learning and chemoinformatics to predict skin sensitization

Work Experience

Consortium For Environmental Research, 2010-Current. Toxicologist

- Researching and compiling dossiers on human health effects in support of EPA submissions
- Developing QSARs to characterize chemicals with unknown toxicity
- Comprehensive analysis of chemical reagents to rank chemicals for health and safety

Teaching Experience

Pennsylvania Institute of Technology, Philadelphia, PA. 2011-2012. Adjunct Faculty

- Courses taught: Introduction to Biology (Lecture and Lab), Introduction to Chemistry (Lecture and Lab)

University of Maryland University College, Adelphi, MD. 2002-2006. Teaching Assistant

- Courses TA'd: Introduction to Bioinformatics, Regulatory Issues in Biotechnology, Societal Issues in Biotechnology

Posters:

"Using Weighted Gene Correlation Network Analysis to Derive Networks from Microarrays: MPTP at day 1 and 7 post-lesion compared to controls " Maertens A, Kleensang A, Hartung, T. Developmental Neurotoxicity Conference 2014.

"Probabilistic Hazard Assessment for Skin Sensitization Potency using Machine Learning to Design Integrated Testing Strategies.". Luechtefeld, T, Maertens, A, Kleensang A, Hartung, T. Sarocha, V. Society of Toxicology 2014.

"Using Weighted Gene Correlation Network Analysis for Microarray Meta-Analysis" Maertens A, Kleensang A, Hartung T. International Systems Toxicology Conference 2013.

Publications:

Bouhifd, M., Hogberg, H. T., Kleensang, A., Maertens, A., Zhao, L., & Hartung, T. (2014). Mapping the human toxome by systems toxicology. *Basic & Clinical Pharmacology & Toxicology.*

Bressler, J. P., Olivi, L., Cheong, J. H., Kim, Y., Maerten, A., & Bannon, D. (2007). Metal transporters in intestine and brain: Their involvement in metal-associated neurotoxicities. *Human & Experimental Toxicology, 26*(3), 221-229.

Hartung, T., Luechtefeld, T., Maertens, A., & Kleensang, A. (2013). Food for thought. . . Integrated testing strategies for safety assessments. *Altex, 30*(1), 3.

Huang, H., Maertens, A., Hyland, E. M., Dai, J., Norris, A., Boeke, J. D., et al. (2009). HistoneHits: A database for histone mutations and their phenotypes. *Genome Research, 19*(4), 674-681.

Kim, Y., Olivi, L., Cheong, J. H., Maertens, A., & Bressler, J. P. (2007). Aluminum stimulates uptake of non-transferrin bound iron and transferrin bound iron in human glial cells. *Toxicology and Applied Pharmacology, 220*(3), 349-356.

Kleensang, A., Maertens, A., Rosenberg, M., Fitzpatrick, S., Lamb, J., Auerbach, S., ... & Hartung, T. (2014). t4 workshop report: Pathways of Toxicity. *ALTEX*, *31*(1), 53.

Maertens, A., Anastas, N., Spencer, P. J., Stephens, M., Goldberg, A., & Hartung, T. (2013). Green Toxicology. *ALTEX*, *31*(3), 243-249.