Formal Representation of Toxicology Knowledge towards Toxicity Prediction and Data Mining

By

Dana Klassen

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of:

Masters of Science Biology with Specialization in Bioinformatics

Carleton University
Ottawa, Ontario

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-87847-7

Our file Notre référence
ISBN: 978-0-494-87847-7

Abstract

We investigated the use of semantics to aid data integration and predictive model development. Toxicity data from multiple sources was converted into a Semantic Web Linked Data resource. Datasources included the U.S Environmental Protection Agency ToxCast program [1], the Comparative Toxicogenomics Database [2], and the U.S National Library of Medicine TOXNET archives [3]. The resource was used to develop predictive toxicology models represented as OWL-encoded ontologies. The use of OWL enabled automatic classification of linked data, model integration, and logical interpretation of results. A framework was developed to represent molecular mechanisms of action allowing reasoning and semantic querying answering. Finally, we investigated how this framework could be used to infer novel features for machine learning input. The completed work shows how formal knowledge representation can be used to improve integration of toxicology information and development of predictive models.

Table of Contents

List of Tables

List of Illustrations

Acknowledgements

The vast majority of this work was carried out by me and the thesis was written by me with critical input from my supervisor, Dr. Michel Dumontier. I would like to thank Dr. Dumontier for the supervision of my work as well as valuable and insightful comments. Parts of this work, however, are a result of collaboration. In particular, the work pertaining to the formalization of decision trees into OWL ontologies involved contributions from Dr. Leo Chepelev. Dr. Chepelev provided the initial chemical descriptors and data files and tested the chemical classification with sample data. This work was published as a conference proceeding in the 2011 International workshop on OWL: Experiences and Directions (OWLED 2011).

1 Introduction

1.1 Mechanisms of Action: Knowledge of Toxicity

Toxicology examines the chemical origins of adverse effects in biological systems. The scope of toxicology is broad covering investigations of dose response effects, exposure routes, biological targets, pathways, and biochemical processes. All these fields are united through interpretation based on knowledge of molecular mechanisms of action. In other words, the series of processes leading from exposure to outcome including uptake, distribution, metabolism, biological transformation, and accumulation. Knowledge of mechanism of action allow us to interpret results (e.g. dose response curves), limit scope of investigations (e.g. examine specific biological targets), and identify novel interactions [4-6]. We elucidate these mechanisms by uncovering relations between chemical and endpoints, identifying biological targets, and determining system level effects of chemical exposure.

Developing relations between chemical and endpoint provides information on how and why a chemical is toxic. Toxicity endpoints refer to observations regarding a specific disease, symptom, or sign related to toxicity. The goal is to establish a relationship between dose and effect. These relationships are specific to the test system used, in vitro, in vivo, and how the chemical is exposed such as exposure site and dosageorganism. Knowledge of endpoint relations is used to guide investigations examining chemical interaction with specific biological targets. [7]. Further statistical analysis of data can refine chemical endpoint relationships to the individual

contributions of chemical substructure [8-10]. Using chemical or substructure-endpoint relations we can determine the fate and effect of a chemical as it moves through an

Understanding how a chemical interacts with a biological target provides information to the mechanistic workings of toxicity. A chemical can interact at multiple levels of an organism, such as alter gene transcription and/or protein stability. Investigations may involve measuring the effects of individual gene transcription rate changes in response to chemical exposure. The data produced from these investigations provides information as to the targets a chemical works through to produce an outcome. We use the results of molecular toxicology investigations to refine our knowledge of chemical-endpoint relations.

For an organism, the final level mechanisms of actions are uncovered is at the level of the system. A system level of investigation seeks to understand how the entire biological system interacts due to chemical perturbation. Investigations on the system typically fall under the category of 'omics: proteomics, genomics, or metabolomics. Each seeks to understand the effect on the system at the protein, gene, and metabolite level respectively. The data from these experiments help uncover the overall system patterns reflected by chemical toxicity. These changes may be the result of individual or multiple mechanisms of action. The analysis and interpretation of systems level investigation relies on knowledge of chemical-endpoint and molecular toxicity relations.

Mechanism of action is the knowledge linking the diverse scope and purpose of toxicology investigations. This knowledge is complied from separate investigations examining a small part of the overall toxic process. Each type of investigation produces data that uncovers a facet of chemical toxicity. Any insight to be gained from these investigations depends on how we struc-

ture, store, and share information to allow examination of the complete picture. The greater the detail and integration of captured data the higher the chance of developing useful predictive patterns.

1.2 Toxicology Informatics: Supporting a new Paradigm in Toxicity Prediction

Chemicals are all around us from industrial to consumer products, food additives, and drugs. Understanding the risk these chemicals present to human health is important to protecting our society. The challenge is daunting having toxicity information for approximately 10 000 of the 60 million registered chemicals [11]. The sheer number of chemicals to be tested out paces our ability to characterize these chemicals using in vivo model systems. This is the motivation behind regulatory agencies focus on computational methods that leverage existing data to streamline chemical toxicity testing [1].

Data are the key to uncovering the potential for chemical toxicity. The analysis of data allows us to derive new facts and fill gaps in our knowledge. The challenge is making sense of the mass of data available to us and organizing it to form a complete picture of toxicity. Toxicity data are spread out across hundreds of databases with very little data available for immediate in silico use2008a; Edgar 2002; Belleau 2008; Culhane 2009; Waters 1981; WilliamsDevane 2009}-. Despite the availability of toxicology data the challenge of integration and analysis remains [1; 10; 12-15]. Progress of toxicology research depends on successful integration of current and future data.. How data are represented affects the ability to share, integrate, and interpret data during analysis [16-19]. Data management and representation have become more relevant as data moves from scientific publications to the web. The transition of data to

the internet has increased the amount of available toxicology data for analysis. Data from hundreds of sources and thousands of experiments can now be accessed for web based analysis

-{Judson 2008; Waters Toxicoinformatics is the scientific discipline concerned with providing computational support to uncover mechanisms of action. This includes the use of bioinformatic and computational tools to support the integration, development, and analysis of data from multiple levels of biological organization [10; 20]. This involves capturing results and data from toxicology investigations to quantify and improve understanding the chemical origins of toxicity [21]. Several computation systems and platforms have been developed to address the need for integration, mining, analysis, and modeling of toxicity information [1; 10; 18; 20; 22; 23]. These systems have not fully addressed the ability to integrate and share data [12; 15]. A potential solution to data integration and interoperability lies in leveraging data semantics.

The Open Toxicology Project (OpenTox) is a toxicology analysis platform utilizing core internet standards to share and access data [15]. The OpenTox project aims to make analysis of toxicology data an open and transparent process [15; 24]. The goal is to provide open web-based access to the common components, data, algorithms, models, and software used in toxicity prediction. At the core of OpenTox is the use of controlled vocabulary to represent and share data [12]. The focus on controlled vocabulary is key to providing interoperability between datasets, algorithms, and software services [12]. The OpenTox project has laid the foundation for leveraging shared vocabularies to improve data mining of toxicity data by allow meaning regarding labels/types present in datasets to be accessed [15].

1.3 Uncovering Patterns in Toxicity: Data Mining

Data mining is the process of finding patterns in data. It is about uncovering the underlying rules and trends that are inherent in data [25]. By uncovering patterns we characterize raw data into useful information by providing context and meaning. We find these rules and trends using a combination of statistics, machine learning, computer science, and information theory [26]. In the field of toxicology and risk assessment there is a need to move away from in vitro and in vivo testing [1; 27; 28]. As such, there is a move towards in silico methods, such as data mining, to characterizing potential toxicity [1; 10; 22; 28].

Predictive models can be generated based on data driven or expert based systems. In data driven systems patterns are automatically or semiautomatically extracted to build models with predictive outcome (linear models, decision trees, and bayesian networks)expert decision logic formulated by a domain expert. The patterns expressed are those learned by a domain expert to make decision regarding interpretation of data. Both these systems can take the form of rule based models. One such rule based model found in toxicology is decision trees. [29]. Data driven systems can be used to uncover chemical bioactivity relationships and experimentally derive outcomes such as phenotype. These models are validated based on their ability to correctly classify data according to the specified outcome. A second type of system are those representing human decision logic or knowledge. Expert based systems are representations of

Decision trees represent a 'divide and conquer' approach to learning and representing data patterns. Each tree is made up of nodes and paths. Nodes represent tests of specific attrib-

utes contained in the data and paths are the test result. In toxicology, decision trees were first established by Cramer and Cramer in 1976 to estimate potential toxicity [30]. Since their introduction, decision trees have been accepted as a representation for predictive models of data driven and expert based systems. Decision trees represent an interpretable model capable of handing mixed data types(e.g numerical and nominal) but are limited to categorical output [26]. However, a disadvantage with decision trees, and all predictive models, is the format we capture the model affects how and what they can be used for. The logic behind the model, rules and descriptors, are not easily shared or compared between applications due to lack of shared model standards and vocabulary. If we could represent the meaning of a model we could derive logical explanations of classification. The semantics would also enable comparison of derived models based on similarity. Knowing how models differ would allow us to identify relevant models from a collection.

1.4 Semantic Knowledge Management

1.4.1 Semantic Toxicoinformatics

Knowledge discovery is the process of uncovering useful patterns in data. It involves data contained in databases to derive implicit information. Information we use to develop models and guide scientific research. Current technology for data storage and access is outpacing the ability to derive meaningful information from data [15]. Toxicology databases inhibit the use of data to generate quantitative models due to differences in data types, object representations, and formats [12]. In toxicology this has led to a "data gap" in publicly available data hindering the prioritization of chemicals for in vitro toxicity testing [1; 10; 13-15]. As a result only frac-

tion of data are available for data mining. Recent advances have improved the ability to integrate toxicology data [31-33]. Despite these advances the ability to seamlessly integrate multiple resources is missing [12].

Research into the use of shared data standards, protocols, and representation are needed to increase integration of data and applications. Common data standards increase access to data, knowledge, and software. The Semantic Web is a collection of web standards for exposing, sharing, and connecting data. Research developing the Semantic Web provides the technologies to allow interoperability of data and data services (storage, management, software analysis tools). Our goal is to leverage shared vocabularies and annotations to improve the capture, integration, and interpretation of information relevant to predicting toxicity based on prior knowledge of mechanisms of action.

1.4.2 Hypothesis

The formal representation of toxicology knowledge will enhance our ability to predict toxicity by using linked data and inferred relationships. We will test this hypothesis by developing a predictive model of toxicity with and without inferred attributes.

1.5 What is the Semantic Web?

The World Wide Web runs on making connections. Web pages are connected through unique web addresses, called Uniform Resource Identifiers (URI), and a set of standards that allow a web browser to move from one page to the next. At the end of each web address is a web

page, a document, containing human readable information. One web address always references the same web page. Now what if we could do the same but for data with each piece of data referenced by a unique web address. Data would be able to be linked together and structured. We could reference and describe each piece of data or information. Better still, structuring data and its linkages via semantics means we could automatically traverse the web for data and uncover novel connections. The process of creating and sharing semantically encoded data are the idea behind the Semantic Web [34].

The Resource Description Framework (RDF) is the basic data model underlying the Semantic Web [34]. RDF gives us the flexibility to create web addressable statements [35]. Each statement, called a triple, is composed of three parts, the subject, the predicate, and the object. The subject denotes the topic of the statement, the predicate the traits or attributes, and the object is another resource being referenced. For instance, the statement "the protein_X binds gene_Y" can be expressed in RDF as a statement with subject ,"protein_X" predicate ,"binds" and object "gene_Y". Additional statements can be created to express that, "protein_X" is a "Protein" and "gene_Y" is a "Gene". Each part of the triple is represented by a unique URI that is web accessible (Figure 1-1).

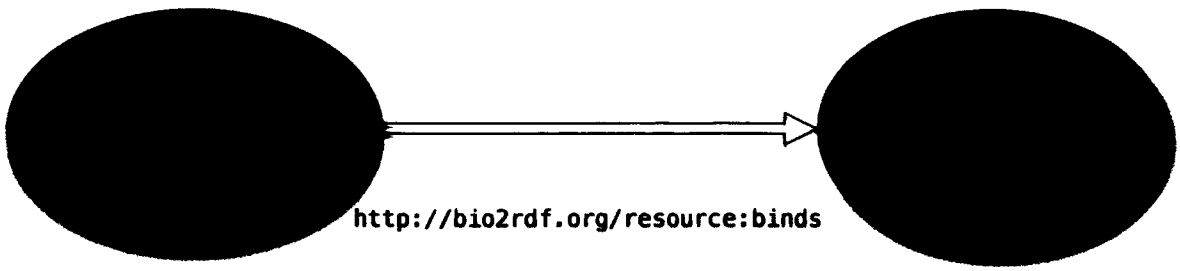

Figure 1.1: A statement represented by the RDF data model showing an example Subject Predicate Object.

Each URI used to represent a resource can be 'resolved' to retrieve a representation of the resource from the web. Exactly the same way you would enter a website address to see a website. Statements are directed, meaning a subject always refers to the object within a statement. A collection of statements forms a labeled directed graph (Figure 1-2). As more and more statements are added overlap among statements with shared syntax automatically forms new links, e.g. multiple resources that are "Genes".

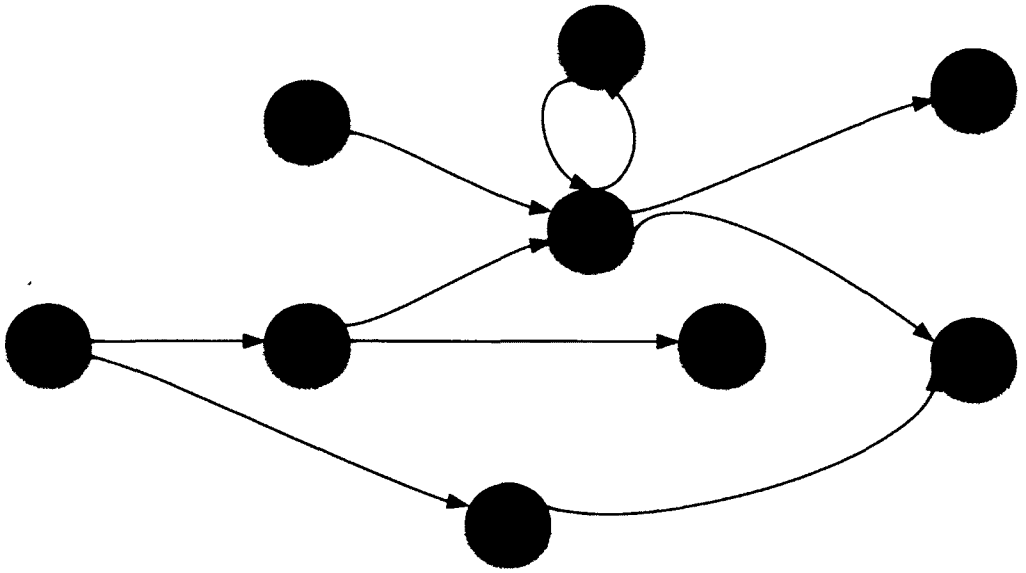

Figure 1.2 : A representation of the RDF data model showing a graph formed from RDF statements. Blue circles are resources and arrows predicates.

When talking about data it helps to talk about the entities the data is describing [36]. The RDF data model allows us to formalize descriptions about these entities. Different communities can represent the same thing differently allowing integration of data from multiple domains [36]. The semantics and interpretations are formalized in the RDF data model [35].

Figure 1-3 shows how a data measurement in text format, "56.7 g/mL" may be formalized using the RDF data model. In this example, a concentration data point has a specific value of "56.7". The example also shows how relations between a concentration and a unit, "g/ml", can be represented. Data measurements with the same units can be compared or the information about units can be used to convert measurements to other values automatically via web serv-iceslinked with other data sources that share similar concepts e.g. measurements over sev-eral experiments or from multiple patients [36-38]. Data represented in the RDF model is called Linked Data. Using RDF we can represent the semantics or meaning of linked data in an unambiguous way. . Conceptualization of data into an RDF model allows data to be

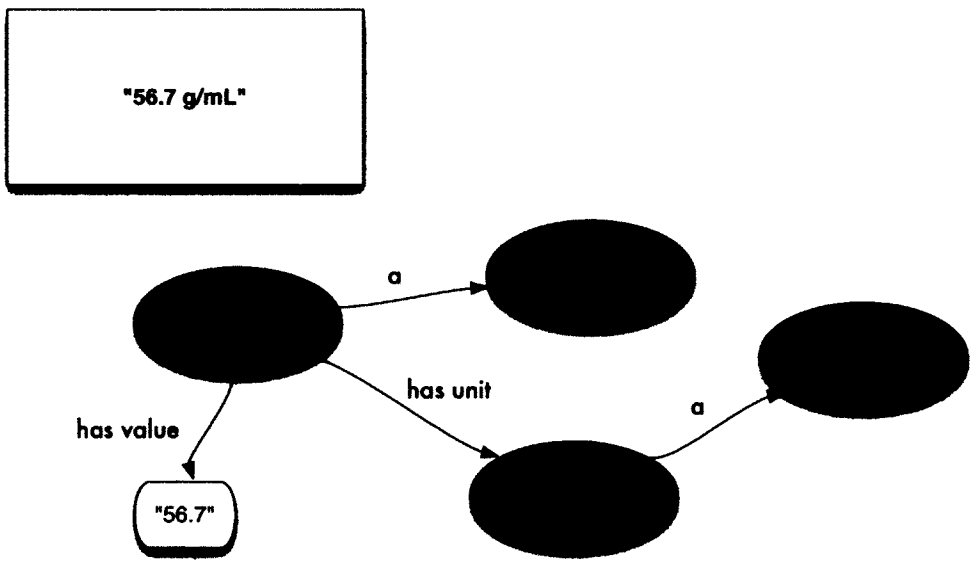

Figure 1.3 : Example formalization of a concentration measurement "56.7 g/mL" and the abstract concepts to represent the measurement in an RDF data model. Blue nodes repre-sent "types" of things, rectangles represent resources, arrows show relationships between types.

However, for more complex data manipulations a formal way involves the construction of what is called an Ontology. An Ontology is an engineering artifact which is a formal, explicit specification of a shared conceptualizationsemantics. The Web Ontology Language (OWL) is an ontology language for the Semantic Web. It is based on a family of so-called description logics, which favour a number of reasoning services including consistency checking, determining subsumption and building hierarchies, classifying instances into their most specific classes, and answering queries. [39; 40]. Ontologies formally describe entities and the relationships that hold between them. Ontologies are formalized by committing the conceptualization of the domain to a language with formal (machine understandable) These abilities allow us to check facts for satisfiability (non-contradictory statements) and if they represent more generalized versions of other relations between types, relations or individuals. Although RDF can be used to conceptualize and describe basic concepts and connect data it lacks the vocabulary to, for instance, describe relations as being transitive, symmetric, reflexive, or to specify class equivalence in terms of necessary and sufficient conditions. OWL ontologies capture the universal qualities of entities in a general sense that applies to all individuals of those types. Hence, OWL ontologies that are linked to RDF data yield a powerful knowledge base for knowledge management [41; 42].

1.6 Upper Level Ontologies for Life Sciences

When we think about the processes of taking a measurement, we might consider several aspects such as the value of the measurements relative to some unity of measurement, the time and location at which the measurement was taken, the device used for the measurement and the agent taking the measurement. While this applies intuitively to scientific measurements in

a biological laboratory, it could just as well apply to bakers that are measuring ingredients to bake a cake. Since there are commonalities across different domains for similar processes, entities and qualities, it becomes important to specify a general representation that can be leveraged by all. Upper level ontologies do just that - they specify and should logically enforce general ontology design patterns across domains.

The life sciences has been an area of active upper level ontology development. Several life sciences upper level ontologies exist such as the Basic Formal Ontology (BFO) [40], the Descriptive Ontology for Linguistic and Cognitive Engineering (DOCE) [43] and Suggested Upper Merged Ontology (SUMO) [44]. The most popular upper level ontology for life sciences ontologies is the Basic Formal Ontology (BFO) and is currently employed by the Gene Ontology (GO) and Ontology of Biomedical Investigations (OBI) [40]. However, in response to certain deficiencies in the design of the BFO the Semanticscience Integrated Ontology (SIO) [45] was developed as a competing upper level ontology for the life sciences. SIO's motivation is that BFO only admits types for which there are instances - so would exclude hypotheticals specified by some scientific theory e.g. the Higgs Boson. The Higgs Boson is a postulated sub atomic particle but has yet to be proved experimentally. It provides a means to solve problems in quantum mechanical phenomena. Since the Higgs Boson has not been shown to exist it cannot be represented and used in the BFO. SIO is the backing ontology to the Bio2RDF project which houses the majority of currently available life sciences linked data.

The choice of upper level ontology depends upon the objective of the ontology. An in depth discussion into design philosophies, philosophy of science, will not be undertaken aside from stating that the majority of upper level ontologies ascribe to a 'realist' viewpoint. The realist phi-

losophy states that entities exist independent of observation and we can represent those entities as classes in an ontology. The BFO is a set of reference ontologies restricted to experimental determined/observed instances [40]. The design was to allow community development by partioning development of domain ontologies [46].

The Semanticscience Integrated Ontology (SIO) was designed in response to deficiencies in the BFO. SIO seeks to create a single coherent and consistent ontology based on a set of defined design patterns. The result is an ontology that facilitates integration and querying across ontologies. All classes in SIO have been defined using computer amenable axiomatic expressions. The community is therefore focused more towards design pattern reuse and development. Also, SIO is the upper level ontology for the Bio2RDF project, a triple store containing 40 of the largest life sciences databases [47]. The linked data created as part of this project will be integrated with the Bio2RDF project. Using the SIO ontology for development of a knowledge base will enable access to a greater volume of linked data resources.

1.7 Syntax and Semantics

Knowledge representation languages, prepositional logic, description logic, and first order logic, all share the use of symbols to encode knowledge, and assign meaning. The use of formalized symbols to represent knowledge allows a mathematically rigorous definition and treatment. The specific use of a symbol or symbols differs from language to language resulting in different abilities in capturing and representing knowledge. However, the basic organization, or types of symbols, are the similar. Here we will be focusing on the use of Description Logics to capture knowledge.

In languages related to Description logics there are two types of symbols to represent statments: logical and non-logical [48; 49]. Punctuation symbols are used to group or order symbols. Connective symbols are interpreted on the basis of logical functions, such as negation, conjunction, disjunction, for every, for some, and equality. Variable symbols are used to represent a constant.. The logical symbols have a constant meaning assigned to them regardless of the knowledge being represented [48]. Logical symbols are of three subtypes: punctuations, connectives, and variables

Non-logical symbols are dependent on the domain being represented. The non-logical symbols represent the terminology or natural language of a particular domain being represented [49]. Each non logical symbol is associated with an "arity" or the number of arguments that a particular symbol requires to have meaning. For example, hasName["John","Smith"] is a symbol with arity 2, binary, which takes a first and a last name. These symbols are of two types: atomic concepts and roles. Atomic concepts assign members to a particular concept. Atomic roles are used to make connections between concepts. The meaning of a concept or statement is a function of the interpretation of the combination of atomic concepts and roles [48].

The Description Logics syntax gives us a means to represent the knowledge of a domain [48]. We can make statements and build descriptions about 'things'. The result is a knowledge representation system containing the vocabulary of the domain being represented [48]. This system is made up of concepts, individuals, roles, and relations between concepts. Concepts and roles are defined using the logical symbols of the language to express relations between other

concepts and roles. This is called model-theoretic semantics, we interpret statements based on an abstract model or description of the world [48]. In this way increasingly complex descriptions can be created for a given concept or role.

From these descriptions we can create a knowledge resource or knowledge base. This resource is dynamic in that we can reason over it. Reasoning involves examining the formal syntax used to connect concepts, build descriptions, and define instances. We can check the descriptions for satisfiability, no contradictory statements, and subsumption, whether a concept is a more general form of another [38]. Instances of the ABox can be checked to belong to a specific concept. In other words assertions about instances are consistent. The descriptions used to build the concepts in the TBox can be used to query for groups of subclasses and instances [38].

1.8 The Web Ontology Language(OWL)

The descriptive language used to encode a knowledge base affects the kind of information to be captured and what can be done with it [38; 48; 49]. In the Semantic Web several languages exist for ontology building. We encountered RDF and RDFS and although simple and powerful they lack the formality to tackle harder knowledge representation challenges. The Semantic Web has developed the Web Ontology Language (OWL) as the language for ontology representation [38; 50].

The Web Ontology Language (OWL) is an ontology language based on Description Logics (DL) This tractability comes at the tradeoff of reasoning capabilities meaning OWL is limited to

subsumption (IS-A) reasoning. The OWL language is actually a collection of three increasingly expressive languages: OWL-lite, OWL-DL, and OWL-FULL [38; 50]. Each language designed to meet a specific purpose of knowledge representation. Together these sublanguages constitute the languages used for the construction of Ontologies in the Semantic Web. For the purposes of this research OWL-DL will be considered as the ontology language. OWL-DL was developed to maximize the expressiveness of the language but avoid the problems of intractability and decidability found in OWL-FULL . The sublanguage has the strongest links to DL allowing use of DL reasoning procedures [48; 50]. In life sciences the ability to reason over data provides the ability to access implied knowledge. For instance, the use of a chemical ontology to determine signaling molecules as the set of chemicals which bind to signaling proteins. The ability to reason over knowledge has been demonstrated over diverse sources of life sciences data [5; 41; 42].. The strong backing in DL allows OWL to leverage the research into representation and reasoning systems developed in DL for artificial intelligence systems [48]. Description logics was chosen as a basis for OWL due to its tractability, or how easily certain reasoning procedures can be performed [38; 50].

2 Semantic Integration of Toxicology Resources

2.1 Abstract

The diversity of toxicology data representation makes it challenging to integrate data. To investigate semantic integration, toxicology data from various sources was integrated into a linked data resource based on underlying data semantics. Publicly available datasources included the U.S Environmental Protection Agency ToxCast program (ToxCast), the Comparative Toxicogenomics Database (CTD), and the U.S National Library of Medicine TOXNET archives. We designed the ToxKB Ontology, formalized in OWL, to integrate resources represented in the linked data resource. The resulting toxicology linked data resource was analyzed based on the ability to discover new knowledge using semantic connections between datasets. This section demonstrates the usefulness of data semantics to overcome data integration challenges of traditional database formats. We were able to answer questions regarding toxicity for chemicals from multiple information sources. This ability was previously not possible in current data systems.

2.2 Introduction

The goal of a toxicology database is to provide access to structured data. Data are structured to provide consistent and logical access to contained information. As the number of data sources increases the ability to integrate data is hindered by lack of unified data structure to

publish with [51]. Sharing and integration of data allows us to expand the conclusions we can draw from information [19; 51]. Data integration is still a challenge in today's toxicology data systems [52]. Toxicology data are diverse spanning multiple scientific domains. This is mirrored in the scope and purpose of toxicology databases from chemical structure information for cheminformatics purposes such as QSAR analysis, to aggregated expert knowledge for use in industry regulation and risk assessment. The diversity of toxicity data present limits the ability to share and integrate data between domains. This has led to an information deficiency in toxicology limiting characterization of toxicity [1; 10; 13; 14]. By limiting access to toxicity data we limit our ability to build effective predictive models [22]. Research into better data integration is important if toxicology is to transition from in vivo animal-based toxicity testing to in vitro and computational techniques as the primary means of testing toxicity [10].

Several efforts to address the integration of toxicology data have been undertaken. One approach is to amalgamate data and the second is to employ a standard shared controlled vocabulary and data model. These approaches are not mutually exclusive. The Aggregated Computational Toxicology Resource (ACToR) is a program by the U.S Environmental Protection Agency (U.S EPA) to integrate existing toxicology information on environmental pollutants [18]. Rather than address interoperability, ACToR centralizes toxicology data into a single relational database. The underlying tables lack a domain shared standardized vocabulary or data model. This in effect restricts database access and sharing to those guidelines outlined by the U.S EPA. The Chemical Effects in Biological Systems (CEBS) database utilizes a more open approach by incorporating a vocabulary and data model based on the standard MIAME and MIAPE vocabulary [53]. CEBS uses the CEBS data dictionary [54] and CEBS SysTox Object

Model [55] to standardize representation of toxicology information. Developing a vocabulary based on a standard permits exchange and integration of information between sources using those standards. This approach treats data and data-semantics as separate entities: data and data object model. Interpretation of data is not possible in the absence of the data object model. The data semantics should be inherent and represented in the data itself [36; 51]. Semantic web technologies allow representation of data and data semantics. Doing so allows data to be automatically linked with other independent resources based on shared semantics. In this work we investigate the application of Semantic Web technologies to overcome the challenge of data integration. We show how the use of Linked Data and OWL ontologies can be used to represent both data and data semantics in a single data model. We show how formal representations of concepts, in OWL, can be used to capture meaning and relations of types within and between datasets. The goal is to create a semantic interoperable toxicology resource that can be easily integrated based on shared conceptualization and formalization. Finally, we demonstrate how formal relations between shared concepts represented using OWL ontologies can be used to integrate datasets which do not share explicit relations. Doing so allows us to simplify the analysis of the landscape of toxicology. This work is important in the development of an integrated toxicology resource to overcome the information deficiency in toxicology.

2.3 Materials and Methods

To facilitate the sharing and reusability, the creation of linked data follows a set of best practices [34; 35]. A procedure was created to convert each data format identified in a review of

publicly available toxicology data. The process was developed to facilitate the use of linked data and provide a distinct source of information from ontologies. Each database identified for conversion was examined for the presence of a database schema to provide the syntactic labels and data types a field holds. If a schema was not available a schema was developed from examination of data and identifying controlled vocabulary. A conversion process was developed to preserve the original identity of the data source and extend connections to other linked data sources. The process involved the syntactic conversion of data fields followed by the semantic conversion involving conceptualization and mapping the syntax present to the Toxicology Knowledge Base (ToxKB) ontology.

2.3.1 Data Sources and Preparation

We identified the Comparative Toxicogenomics Database (CTD) [2], Chemical Effects in Biological Systems (CEBS) [23], Distributed Structure Searchable Toxicity Database (DSSTox) [56], ToxCast [1], Genetic Toxicology Database (Gene-Tox) [57], Chemical Carcinogenesis Research Information System (CCRIS) [3], and Toxicology Literature Online(ToxLine [3] datasets for conversion to Linked Data (Table 2.1). Identified datasets were converted according to the procedure based on source dataset format.

Table 2.1: Data sources and accession for conversion to Linked Data.

CTD	Complete dataset	http://ctd.mdibl.org	2010-09-10	see accession

Table 2.1: Data sources and accession for conversion to Linked Data.

CEBS	ICONIX datafile	http://www.niehs.nih.gov/research/resources/databases/CEBS/	2010-10-30	see accession
DSSTOX CPDBAS	Carcinogenicity Potency All Species Summary datafile	http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html	2011-11-10	CPDBAS_V5a
ToxCast	Toxcast phase 1 dataset only	http://www.epa.gov/ncct/toxcast/data.html	2011-11-10	see accession
Gene-Tox	Complete dataset	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX	2010-09-10	02/11/2002
CCRIS	Complete dataset	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS	2010-09-10	08/28/2009
TOXLINE	EMIC,Archival, CIS data-files	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE	2010-09-10	05/23/2002

2.3.2 Construction of Integrated Toxicology (RDF) Network

Dataset conversion was performed using RDF libraries written in the Ruby scripting language. For each data format encountered a conversion procedure was created. All conversion procedures utilized two RDF namespaces unique to the data source. One namespace to address data from the original dataset the second namespace to reflect changes to the data structure from the original dataset. The first namespace used the URI fragment named after the datasource, the second namespace used the URI fragment created by combing the datasource name and "_resource:" string. All namespaces used the Bio2RDF host address:

"http://bio2rdf.org/", e.g. http://bio2rdf.org/toxcast: and http://bio2rdf.org/toxcast_resource:. The Bio2RDF host address was used as datasources do not provide support for Linked Data. All datafile conversion output was saved to an RDF N-TRIPLE serialized datafile. All unique URI's created used the Ruby scripting language implementation of the MD5 Hash algorithm. This algorithm creates a unique 128-bit hash value based on input information ensuring uniqueness of RDF URI's. All example RDF output from conversion procedures are presented in RDF N3 syntax notation. All software is open-sourced and available on request.

2.3.2.1 Conversion of XML to Linked Data

Extensible Markup Language(XML) data-files were converted to Linked Data using the Ruby Nokogiri XML library [58]. RDF triples were created to mirror XML hierarchical structure, the RDF subject representing a parent XML node and RDF object representing a XML child node or text. The RDF predicate of a triple connected the parent (subject) with the child(object) or text(literal). The generated RDF graph was rooted at the document level.

A URI representing the datafile was created using an MD5 hash of the entire XML datafile. Each XML record converted was connected to the datafile URI. RDF triples were created by processing a parent node for available children nodes and contained text. The RDF Subject URI fragment was created using an MD5 hash of contained XML. An RDF predicate URI fragment was created using the child node tag and prepending "has". Finally the RDF subject URI was created using an MD5 hash of contained XML or contained text. If the XML tag contained text it was attributed to the converted tag using the rdf:value predicate. Each RDF subject was

typed based on the originating XML tag. Unique identifiers referencing outside databases, e.g. CAS Registry Numbers, were assigned namespaces to the datasets they originated from.

2.3.2.2 Conversion of Flat-file to Linked Data

Flat-file data files are organized in rows and columns separated by a symbol (tab or comma). Each flat file consistsof a header containing a label to describe data within that column. Flat-file data files were converted using header information to assign type information to contained data via the rdf:type predicate.

A document level URI was created using the MD5 Hash described above of the entire document text. The same was done for each row, other than the header, in the document. These URI's would be used to reference all information contained in the document and row respectively. For each column in an individual row a URI was created using the URI of the row it was a part of, the value, and the header of the column. This was done to ensure each data point in the dataset would be unique. All URI's created used namespaces assigned to the dataset being converted. Unique identifiers referencing outside databases, i.e. CAS Registry Numbers, were assigned namespaces to the datasets they originated from.

2.3.2.3 Conversion of CEBS Data Dictionary to Linked Data Dictionary

The CEBS database uses a separate XML based controlled vocabulary, CEBS Data Dictionary (CEBS-DD), to represent a relation scheme between data fields. The conversion of the CEBS-DD to linked data extends the previous XML based conversion process to incorporate the creation of a Simple Knowledge Organization System (SKOS)dictionary. SKOS is a model and

RDF vocabulary that allows the content of controlled vocabularies to be represented as Linked Data. To convert the CEBS-DD to linked data the SKOS core vocabulary was used to represent the CEBS content scheme. The basic structure of the CEBS-DD XML document is shown below: [59] based Linked Data

```
<sift creationdate="">
        <section name="">
                <attr description="" type="" name="TERM">
                        <alias>ALIAS</alias>
                        <cv>SPECIFIC DATA LABEL</cv>
                </attr>
        </section>
</sift>
```

The document reflects the SIFT document structure provided by CEBS . Each data table is grouped into related sections mirrored in the vocabulary. Each section in the vocabulary document, denoted by a term <section> tag, contains the list of possible terms <attr> and values of those tags <cv>. Alias for terms are captured via the <alias> tag.

For conversion of the CEBS-DD each section tag was represented as a skos:Collection. Each term of that section was represented using the skis:Concept type. Concepts were related to collection through the skos:member predicate. An example XML CEBS-DD section entry is shown below:

```
 <section name="STUDY">
        <attr description="Date or time of study initiation" type="STRING" required="true"
name="START_DATE">
                <alias>STUDY_START_DATE</alias>
        <alias>STUDY START DATE</alias>
        </attr>
</section>
```

The <attr> tag entry contains xml attributes for a text description, datatype , boolean requirement, and name. Contained is also information as to the relation with other terms. In this in-

stance "START_DATE" is said to be equivalent to "STUDY_START_DATE" and "STUDY START DATE". The entry converted to RDF using the SKOS vocabulary:

```
@prefix cebs_dictionary: <http://bio2rdf.org/cebs_dictionary:> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs : <http://www.w3.org/2000/01/rdf-schema#> .
cebs_dictionary:CEBSD_2 a skos:Collection ;
        skos:prefLabel "STUDY" ;
        rdfs:label "CEBS STUDY section collection" .
cebs_dictionary:CEBSD_19 a skos:Concept ;
        rdfs:label "START_DATE [cebs_resource:CEBSD_19]";
        skos:prefLabel "START_DATE" ;
    skos:definition "Date or time of study intitiation" .
cebs_dictionary:CEBSD_20 a skos:Concept;
    rdfs:label "STUDY_START_DATE [cebs_resource:CEBSD_19]";
    skos:prefLabel "STUDY_START_DATE" ;
    skos:closeMatch cebs_dictionary:CEBSD_19 .
 cebs_dictionary:CEBSD_2 skos:member cebs_dictionary:CEBSD_20 .
 cebs_dictionary:CEBSD_21 a skos:Concept;
    rdfs:label "STUDY START DATE [cebs_resource:CEBSD_19]";
    skos:prefLabel "STUDY START DATE" ;
    skos:closeMatch cebs_dictionary:CEBSD_19 .
```

Unique URI's were created, for each entry, using an incrementing counter and appending it to the prefix "CEBSD_ ". The namespace "http://bio2rdf.org/cebs_dictionary:" was used to maintain identity of the vocabulary resource. For the example entry, 3 skos concepts: cebs_dictionary:CEBSD_19, cebs_dictionary:CEBSD_20, and cebs_dictionary:CEBSD_21, were created and assigned the according xml attributes (definition and labels) using corresponding SKOS vocabulary. Concepts were related to one another through the use of the skos:closeMatch predicate.

2.3.2.4 Conversion of SIFT Files to Linked Data

The CEBS database employs a modified Simple Investigation Formatted Text (SIFT) file format for data serializationture can be found at:

ftp://157.98.192.110/ntp-cebs/datatype/microarray/drugmatrix/SIFT_description.doc. [23]. A description of the SIFT syntax and document struc

A URI representing the datafile was created using an MD5 hash of the entire SIFT datafile. This URI would represent the root of the RDF graph. Each SIFT datafile is composed of multiple sections, each pertaining to a specific topic. Sections are composed of two parts: a header section and a tab deliminated table. For each section of a SIFT data file a unique URI was created by performing a MD5 hash of the section. Sections were typed by querying the CEBS vocabulary using the section label (see conversion of separate XML dictionary). The section was linked to the datafile using the predicate http://bio2rdf.org/cebs_resource:hasStudyPart. The header section was similarly processed and assigned to the section using the predicate http://bio2rdf.org/cebs_resource:hasMetaData.

Data tables were converted based on the procedure outlined for flat-files with modifications for typing. Each row in the data table was typed as a cebs_resource:Entry and assigned to the associated section URI using the predicate cebs_resource: hasEntry. Each data point in a row was typed based on a look up of the CEB's dictionary for the header term and assigned to the row using the cebs_resource:hasAttribute. Values for each datapoint were assigned using the rdf:value predicate.

2.3.2.5 Conversion of Structure Data Files (SDF) to Linked Data

The EPA DSSTox resource uses the SDF file format designed and described by Molecular Design Limited [60]. DSSTox created this files for the purpose of linking chemical structure data with toxicology information from outside sources. Conversion is based on the structure outlined by Dalby et al, 1992 [60].

An RDF graph root was created using an MD5 hash of the entire document. Converted records were assigned to the root via predicate http://bio2rdf.org/dsstox_resource:hasRecord. Record URI's were created using an MD5 hash of the entire record. Individual records are composed of two main components: Mol chemical structure files and data entries. Mol files contain symbols which are not URI compatible. The RUBY built-in CGI library was used to encode non-URI compliant characters. A Unique Data URI was created using a MD5 hash of the document URI, record URI, and contained data information . Each data URI was typed according to the data header provide for each data entry. Data entries were connected to the record via predicates created using the assigned namespace and URI fragment "has" + DataHeader. Unique identifiers referencing outside databases, i.e. CAS Registry Numbers, were assigned namespaces to the datasets they originated from.

2.3.2.6 Data Loading

Converted linked datasets were made accessible using the Virtuoso Open source TripleStore software available from http://www.openlinksw.com/wiki/main/Main (version 6.1.4). Software was complied and hosted on a MacBookPro using 2.2 GHz Intel Core i7 CPU using 8GB ram. Data sets were loaded into the triple store using a custom script (http://code.google.com/p/semanticscience/source/browse/trunk/lib/php/virtuoso_load.php).

2.3.2.7 Querying Integrated Toxicology (RDF) Network

The converted data sources were integrated into a single Linked Data network. Graph based queries composed in SPARQL were developed to answer toxicology questions using information contained in the integrated Toxicology (RDF) network. Each query used to following prefix header:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX genetox: <http://bio2rdf.org/genetox_resource:>
PREFIX ccris: <http://bio2rdf.org/ccris_resource:>
PREFIX toxcast: <http://bio2rdf.org/toxcast_resource:>
PREFIX ctd: <http://bio2rdf.org/ctd_resource:>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX cis:<http://bio2rdf.org/cis_resource:>
PREFIX archival:<http://bio2rdf.org/archival_resource:>
PREFIX dsstox: <http://bio2rdf.org/dsstox_resource:>
```

2.3.3 Toxicology Knowledge Base (ToxKB) Ontology Development

The ToxKB OWL ontology was constructed using the Protégé 4.1 OWL ontology editing software. The Semanticscience Integrated Ontology (http://semanticscience.org/ontology/sio.owl), SIO, was used as an upper level ontology. Classes were created based on resource types in the Toxicology (RDF) network. Each class was defined, where possible, using source data schema.

Linked Data were mapped and transformed to ToxKB ontology classes using the R2R framework (http://www4.wiwiss.fu-berlin.de/bizer/r2r/). A R2R mapping file, encoded in RDF, was constructed for each data source to specify source to target transformations. Transformed data was used to instantiate ToxKB ontology. The HermiT 1.3.5 reasoner [61] plugin for Protégé 4.1 software was used to perform reasoning procedures over knowledge base classes and instances.

2.3.3.1 Querying the OWL Toxicology Knowledge Base

The Toxicology Knowledge base (ToxKB) was instantiated using linked data converted by mapping from the source (Toxicology (RDF) network) to the ToxKB OWL ontology. Queries were asked of the knowledge base using the Protégé OWL ontology editor 4.1 software DL Query built-in plugin. Concepts and properties described in the ontology have been colored blue.

2.4 Results and Discussion

Our goal was to create a toxicology knowledge framework allowing ontology based data integration. The first phase to semantic integration was performed through conversion of toxicology resources to Linked Data. The result was a single query-able resource of toxicology information using graph based queries. Next, an OWL ontology was used to formalize concepts and relationships contained in the Linked Data. The use of an OWL ontology enabled querying information based on semantic relationships using class expression queries.

2.4.1 An Integrated Toxicology (RDF) Network

A single query-able information resource, as Linked Data, was developed from public toxicology resources. This resource contains a broad range of information including bioassay screening data, literature references, chemical structure and property, mutagenicity/carcinogenicity data, and curated chemical-gene-disease relations for 145 772 chemicals. Each converted dataset represents an independent resource of information in the form of a directed graph structure(Figure 2.1).

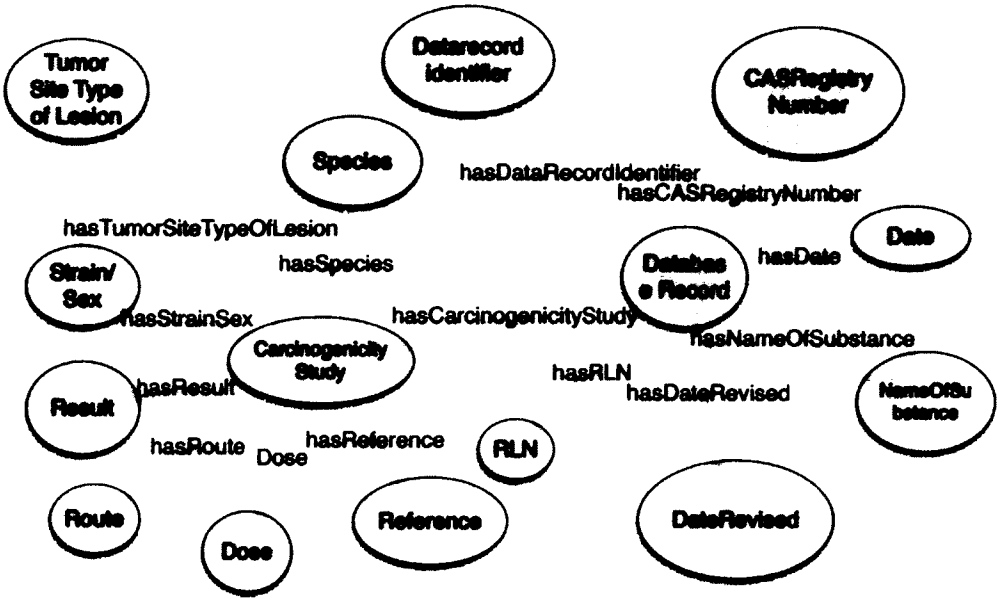

Figure 2.1: Digraph showing partial RDF directed graph structure of Chemical Carcinogenesis Research Information System (CCRIS) resource. Circles represent types of types and arrows relationships between types. Labels have been modified from source to provide human readable text.

Figure 2.1 shows information contained in the CCRIS linked data resource. Information is present regarding carcinogenicity studies including species, route, dose, and outcome of testing. The graph is difficult to understand and query due to structure of original schema. For instance, the chemical name, classification, and CAS registry number are connected separately to a database record instead of related to a chemical type associated with the experimental study(Figure 2.1).

Integration of resources was accomplished through URI's created from shared unique identifiers , e.g. Chemical Abstract Service (CAS) registry numbers. For example, the CAS registry number 72-43-5 always has the following URI <http://bio2rdf.org/cas:72-43-5> independent of the resource it was found in. The result is a query-able path between resources sharing

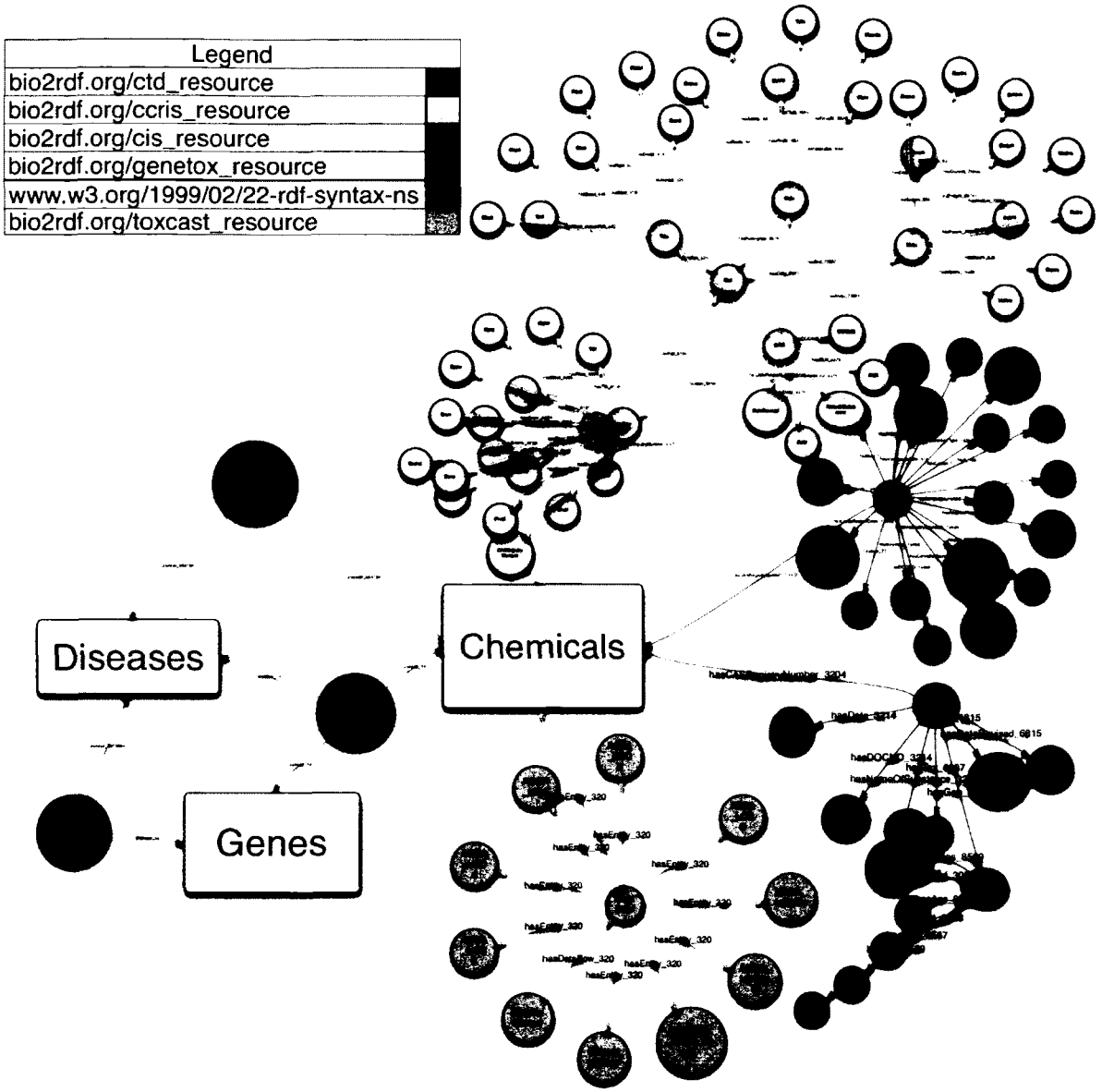

Figure 2.2 : Diagram of Linked Data graph of several converted toxicology data illustrating connections between data sets based on shared URI's. Data sets have been colour coded showing restriction of dataset to individual namespaces. Nodes represent a unique resource type. Bio2RDF, shown as rectangles, form connections to due to shared URI's allowing querying across datasets. Boxes show the types of things relating data resources.

identifiers(Figure 2.2). Shared URI's extend to other external Linked Data resources including Bio2RDF [47]. Integration was not possible with the CEBS database as it did not contain unique identifiers present in other converted databases (Figure 2.2).

Integration was hindered by use of free text and lack of controlled vocabulary. The CTD uses controlled vocabulary to describe relationships between chemicals, genes, and diseases [2]. However, this vocabulary is expressed as free text preventing querying based on relationship type. Lack of controlled vocabulary or insufficient vocabulary was also present in the Genetox and CCRIS datasets preventing querying of common result outcomes.

2.4.1.1 Querying Integrated Toxicology(RDF) Network

The Toxicology (RDF) network was queried using graph based queries expressed in the SPARQL language [62; 63]. Each query is composed of triples defining a subgraph of the Toxicology (RDF) network. Queries allow multiple variables to be specified and restricted [62; 63]. Based on these paths within and between datasets we were able to use the linked data resource to answer questions of toxicity. These questions were able to be asked simultaneously over all resources.

Query 1: How many chemicals with CAS registry numbers are there in the CTD dataset?

SPARQL Query:

```
SELECT COUNT(DISTINCT(?cas))
WHERE{ [] a ctd:Chemical ; owl:equivalentClass ?cas .}
```

The query returns the set of chemicals from the CTD dataset which also have CAS registry numbers (Table 2.2). In the CTD data set unique chemicals are identified using MeSH identifiers. CAS registry numbers were stated to be identical to the matching MeSH identifier. In the CTD dataset CAS registry numbers are considered to represent a chemical equivalent to that

denoted by MeSH identifier. However, CAS numbers in CTD were not typed and therefore cannot be queried for based on a RDF type statement. There is no way to specify only CAS registry numbers are returned. The query for CAS registry numbers allows integration with other datasets which use CAS registry numbers to identify chemicals.

Query 2: How many unique chemicals with CAS registry numbers are there in the Linked Data Network?

SPARQL Query:

```
SELECT  COUNT(DISTINCT(?cas))
WHERE{
{[] a ctd:Chemical;
   owl:equivalentClass ?cas .}
UNION{ ?cas a genetox:CASRegistryNumber .}
UNION{?cas a ccris:CASRegistryNumber.}
UNION{ ?cas a toxcast:CASRegistryNumber .}
UNION{?cas a dsstox:TestSubstance_CASRN .}
UNION{?cas a cis:CASRegistryNumber .}
UNION{ ?cas a archival:CASRegistryNumber.}}
```

This query returns the set of all chemicals identified using CAS registry numbers in the toxicology(RDF) integrated network (Table 2.2). Here the SPARQL UNION operator is used to union separate query solutions. The development of this query requires knowledge of types involved in each dataset as CAS numbers are typed differently in each dataset. Results are aggregated based on distinct CAS numbers in the final query solution.

Query 3: What is the overlap among chemicals with CAS registry numbers between the CTD and GeneTox datasets?

SPARQL Query:

```
SELECT COUNT(DISTINCT(?cas))
WHERE{
[] a ctd:Chemical ;
```

```
    owl:equivalentClass ?cas .
?cas a genetox:CASRegistryNumber .}
```

This query returns the set of chemicals that is common to both CTD and Genetox datasets. The query is expressed as a single graph with results being the intersection of datasets.

Query 4: Which chemicals are positive in a micronucleus assay in the Gene-tox Dataset?:

SPARQL Query:

```
SELECT DISTINCT(?cas_registry_number ) ?result_label ?assaytype_label
WHERE{
[] a genetox:DOC ;
   genetox:hasGen ?experimental_description;
   genetox:hasCASRegistryNumber ?cas_registry_number .
?gen genetox:hasRes ?result;
      genetox:hasAst ?assay_type .
?assay_type rdf:value ?assaytype_label .
?result rdf:value ?result_label .
FILTER(regex(?assaytype_label,"Micronucleus","i"))
FILTER(regex(?result_label,"Positive","i")))
```

This query returns the set of chemicals which have a "positive" result in a micronucleus assay from the Genetox Dataset. The query defines the graph pattern of the Genetox dataset returning all experimental results. Genetox data records are typed as "DOC", each record has an experimental description "GEN" returned via "hasGen" predicate. Assay type information was retrieved via the "hasAst" predicate. "GEN" is connected to experimental outcomes "RES" via "hasRes" predicate. Text labels for each result and assay type were retrieved via the rdf:value predicate. The SPARQL FILTER argument is used to restrict the set of chemicals to those matching the regular expressions on the variables for experiment type and result.

Query 5: Which genes are linked to chemicals that are positive in a micronucleus assay?

SPARQL Query:

```
SELECT ?gene ?label
```

```
WHERE{
[] a genetox:DOC ;
    genetox:hasGen ?gen;
    genetox:hasCASRegistryNumber ?cas .
?ctd_chemical owl:equivalentClass ?cas .
?chemgenerelation ctd:gene ?gene ;
   a ctd:ChemicalGeneRelation ;
   ctd:chemical ?ctd_chemical .
?gene rdfs:label ?label .
?gen genetox:hasRes ?res;
       genetox:hasAst ?ast .
?ast rdf:value ?ast_label .
?res rdf:value ?res_label .
FILTER(regex(?ast_label,"^Micronucleus","i"))
FILTER(regex(?res_label,"Positive","i")))
```

This is an extension of the previous query to include CTD genes with curated relationships to chemicals found "positive" in "micronucleus" test from Genetox. The query is based on the intersection of CAS registry numbers between CTD and Genetox. Here we use the CTD ChemicalGeneRelation type to retrieve related genes.

Query 6: Which diseases are linked to chemicals that are positive in a micro-nucleus assay changes?

SPARQL Query:

```
SELECT  DISTINCT(?cas) ?disease ?label
WHERE{
[] a genetox:DOC ;
    genetox:hasGen ?gen;
    genetox:hasCASRegistryNumber ?cas .
?ctd_chemical owl:equivalentClass ?cas .
?gene rdfs:label ?label .
?gen genetox:hasRes ?res;
      genetox:hasAst ?ast .
?ast rdf:value ?ast_label .
?res rdf:value ?res_label .
[] ctd:chemical ?ctd_chemical ;
    ctd:disease ?disease ;
    rdfs:label ?label .
FILTER(regex(?ast_label,"^Micronucleus","i"))
FILTER(regex(?res_label,"Positive","i")))
```

This query is similar to query # 6 but retrieves diseases associated with micronucleus positive chemicals. The query is based on the intersection between CAS registry numbers between CTD and Genetox datasets. Here the CTD ChemicalDiseaseRelation type is used to retrieve related diseases.

Query 7: What information exists to characterize methoxychlor 72-43-5 as toxic or non-toxic?

SPARQL Query:

```
SELECT DISTINCT(?genetox_experimental_uri) ?genetox_assay ?genetox_assay_result ?ccris_ex-
perimental_uri ?ccris_species ?ccris_cstu_result ?CPDB_TD50_Rat ?Genetronix_GreenScreen_Assay
?ctd_gene_interaction_uri ?ctd_gene_interaction ?ctd_action

WHERE{
OPTIONAL{
# what do we know about how this chemical affects genes?
?chem owl:equivalentClass <http://bio2rdf.org/cas:72-43-5>.
?ctd_gene_interaction_uri ctd:chemical ?chem;
    ctd:gene [] ;
    ctd:action ?ctd_action ;
    rdfs:label ?ctd_gene_interaction .
}
OPTIONAL{
#Identify the results for the ToxCast Gentronix GreenScreen Assay
[] toxcast:hasEntity <http://bio2rdf.org/cas:72-43-5>;
   toxcast:hasEntity ?gs.
 ?gs a toxcast:GreenScreen;
   rdf:value ?Genetronix_GreenScreen_Assay .
}
OPTIONAL{
 # retrieve available results from the GeneTox Database
 [] a genetox:DOC ;
   genetox:hasCASRegistryNumber <http://bio2rdf.org/cas:72-43-5>;
   genetox:hasGen ?genetox_experimental_uri .
?genetox_experimental_uri genetox:hasAst ?ast;
  genetox:hasRes ?res.
?ast rdf:value ?genetox_assay .
?res rdf:value ?genetox_assay_result .
}
OPTIONAL{
# What is the TD50 for rats? in mg/body weight kg /day
[] dsstox:hasTestSubstance_CASRN <http://bio2rdf.org/cas:62-73-7>;
```

```
  dsstox:hasTD50_Rat_mg ?td .
?td rdf:value ?CPDB_TD50_Rat .
}
OPTIONAL{
# what are the results of any carcinogenicity experiments in the CCRIS database using RAT?
[] a ccris:DOC;
  ccris:hasCASRegistryNumber <http://bio2rdf.org/cas:62-73-7>;
  ccris:hasCstu ?ccris_experimental_uri.
 ?ccris_experimental_uri ccris:hasResultc ?result;
    ccris:hasSpecc ?specc .
 ?result rdf:value ?ccris_cstu_result.
 ?specc rdf:value ?ccris_species .
 FILTER(REGEX(?ccris_species,"rat","i"))}}
```

This query retrieves a set of available annotation for methoxychlor from each specified dataset based on optional graph patterns. Querying explicitly for methoxychlor was done by stating the full URI <http://bio2rdf.org/cas:62-73-7>. The SPARQL OPTIONAL argument is used to specify optional graph patterns. Using the OPTIONAL arguments allows querying all datasets despite chemical information not being present in a dataset. The returned query solution contains all matching data from each dataset.

Currently, the only toxicology data system capable of querying multiple datasources is the Aggregated Computational Toxicology Resource (ACTOR) [64]. The Semantic Web offers key advantages over the ACTOR relational database design including the ability of data to remain federated, maintain separate data structure, encode relationships between datatypes as part of the data, and integrate new information based on shared resources (URI's) [5; 47; 52; 65].

Table 2.2 : Type and number of entities contained in the Toxicology (RDF) Resource.

Chemical	145755
Gene	19524

Table 2.2 : Type and number of entities contained in the Toxicology (RDF) Resource.

Diseases	10155
Experiments	100158

Using the toxicology (RDF) network we were able to characterize the number of entities mentioned in the resource(Table 2.2). We found overlap of chemical information across datasets was a limiting factor for characterizing toxicity (Table 2.3). The Comparative Toxicogenomics Database (CTD) contains curated information from scientific literature of chemical gene/ disease interactions. The CTD contains information for 139 252 chemicals representing the largest resource of chemical information in the linked data network. However, only 57 064 chemicals have been annotated with appropriate CAS registry numbers. The next largest dataset, CCRIS, contains 9 374 chemicals with 4 294 chemicals overlapping with the CTD (Table 2.3).

Table 2.3 : Intersection of datasets using Chemical Abstract Service Identifiers showing overlap of datasets.

GeneTox	3204	91	604	1653	1890
ToxCast		302	64	128	224
DssTox_CPDB			604	1066	999
CCRIS				9374	4294
CTD					57064
Total Chemicals with CAS registry numbers:					70548

2.4.2 Construction of Toxicology Knowledge base

The Toxicology Knowledge base(ToxKB) OWL ontology was developed to integrate resources based on shared conceptualization. An ontology allows queries based on relationships between concepts from Toxicology RDF network. Queries were expressed as class expressions. Class expressions are not as powerful as graph based queries used in the Toxicology (RDF) integrated network. Class expression queries are limited to classes, properties, and individuals listed in the ontology [42]. To facilitate integration of diverse toxicology information we choose the Semanticscience Integrated Ontology (SIO) for upper level ontology. SIO provides a basic set of extensible relations and concepts to facilitate integration of life sciences data. As the upper level ontology of Bio2RDF, SIO allows integration with the Bio2RDF network which includes 40 of the largest life science databases [47].

Instance data transformed from resources to ontology using the R2R framework(Table 2.4). The resulting linked data could be loaded into the ontology for querying instance data.

Table 2.4 : Example of one to one mapping of source GeneTox RDF types to ToxKB Classes using the R2R framework.

Gen	TOX_000189	Genetox Experimental Assay	An genetox experimental assay is a agentive processual entity designed to test a specific attribute of a study subject.

Table 2.4 : Example of one to one mapping of source GeneTox RDF types to ToxKB Classes using the R2R framework.

DOC	TOX_0202	Genetox Database Record	A genetox database record is a database record that is part of the genetox database. Each genetox record holds information regarding experimental assay information regarding a specific chemical as it relates to genotoxicity.
Spect	TOX_00087	Experimental Study Subject	An experimental test organism is an organism used as part of an experimental protocol to determine the objective of an experimental study.
Ref	TOX_000187	Genetox Reference	A reference to a scientific publication that is about some genetox experimental assay.
Rpt	TOX_000188	Genetox Panel Report	a panel report is a report evaluating the effectiveness of an experimental assay based on realizing the design objective of the experiment.
Res	TOX_000083	Experimental Outcome	An experimental outcome is the outcome of an experimental study as interpreted by the objective of the experimental study.

Mapped data integrated directly by the ToxKB ontology(Figure 2.3). Data could now be queried for based on relations between concepts described in the ToxKB.

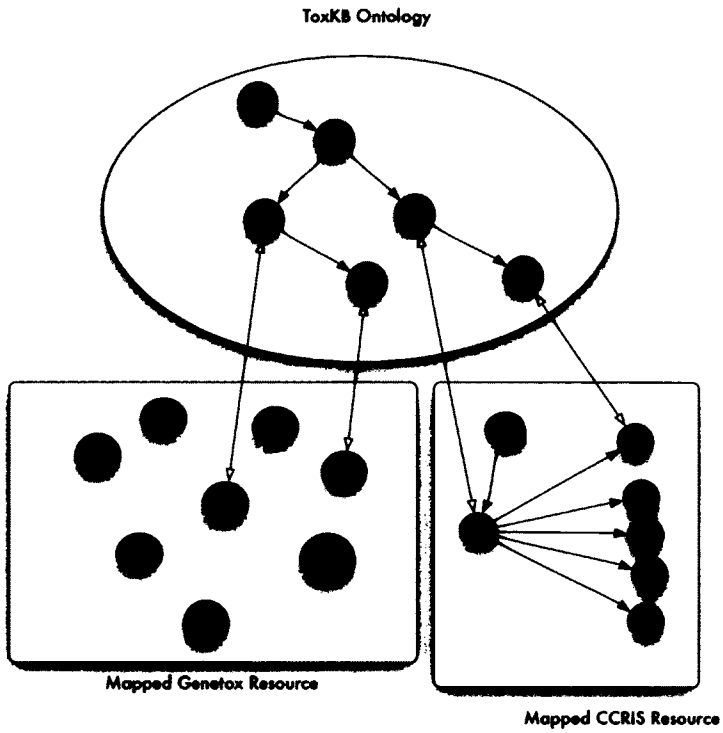

Figure 2.3: Illustration of transformed Linked Data mapped to ontology to enable querying formal relations between concepts. Nodes are colour coded to show distinction between resource sources and do not reflect namespaces.

Using semantic relationships, support based on inference could be used to derive new knowledge. Specifically the ability to state equivalences between concepts and instances. For example, a unique chemical compound may have both CAS Registry and MeSH identifiers. By stating a MeSH and CAS Registry number refer to the same chemical, information about instances from different sources can be merged. This merging of facts allows access to information not present in a single source. In the Toxicology (RDF) network querying between resources was based on the use of shared URI's. The ability to query regarding shared semantics was not developed. For instance, a search for all CAS registry numbers required knowl-

edge of what a CAS registry number was in each dataset (e.g. genetox:CASRegistryNumber or ccris:CasRegistryNumber). Each type was restricted to a namespace and no relation was defined between namespaces. Such as, using the owl:sameAs predicate to assert two concepts are identical. In the ToxKB we formally defined relations between concepts and relations from each namespace. In this way types and concepts could be mapped from multiple namespaces to an OWL class.

2.4.2.1 Querying the OWL Toxicology Knowledge Base

The Toxicology Knowledge base was queried using class expressions. We demonstrate how queries to the knowledge base can be used to retrieve information from ToxKB, such as information related by source or by defining semantic constraints.

Query 1: What chemicals are from a toxicology database?

Class Expression: molecule that 'is referred to by' some 'toxicology database'

The query returns the set of chemical individuals which have some (at least one) source toxicology database. Individuals stated as a 'CCRIS chemical' or 'ToxCast chemical' would be retrieved.

Query 2: Which positive controls are used in the CCRIS experimental studies?

Class Expression : molecule and ('has role' some 'positive control role') and 'is input in' some 'ccris experimental study'

This query returns the set of individuals which are chemicals that were used as positive controls in a CCRIS experimental study. For instance, 4-(methylnitrosamino)- 1-(3-pyridyl)-1-butanone was found to be a positive control used in a CCRIS mutagenicity study. This chemical is a potent carcinogen routinely used to induce tumors in the investigation of carcinogenicity [66].

Query 3: Which experimental assays examine effects of chemical exposure on the PPAR-alpha gene?

Class Expression: 'experimental assay' that 'has target' some 'PPAR alpha gene'

This query returns the set of individuals that are experimental assays which target the PPAR alpha gene. This query includes individuals which target at least one subclass of the PPAR alpha gene (e.g human variant). Experimental instances returned would include the TOXCAST NCGC Reporter Gene Assay PPARa Agonist assay and NCGC Reporter Gene Assay PPARg Agonist assay.

Query 4: What measurement values were generated from an in vitro experimental assay?

Class Expression: 'measurement value' that 'is output of' some ('experimental assay' that 'has quality' some 'in vitro')

This query returns the set of individuals which are measurements values of some in vitro experimental assay. Measurement values from ACEA IC50, ACEA LOC2, ACEA LOC3, and ACEA LOC4 would be retrieved.

2.5 Conclusion

We were able to successfully integrate multiple toxicology resources to answer questions of toxicity. The ability to answer similar questions is currently only possible in more complex relational data systems such as ACToR. Unlike these systems, the information in the Toxicology (RDF) network can remain federated and automatically expanded to include external sources of Linked Data. These systems separate data from data models limiting use and interpretation by outside systems. OWL ontologies represent an open data specification removing domain and application specific barriers found in today's data systems. This work represents the first step utilizing the Semantic Web as a tool for Toxicology data analysis. The challenge to characterize toxicity via data mining increases with the diversity and volume of data. The Semantic Web offers the tools necessary to capture the underlying semantics of data and provide solutions to discovering underlying knowledge patterns.

3 Representation of Decision Trees in OWL

3.1 Abstract

Development and application of in silico predictive toxicology models is hindered by the variety of published model formats. A standard representation would increase interoperability, transparency, and extensibility of predictive models, leading to increased trust in models and reproducibility of results. In this work a standard representation for predictive decision tree models using the Web Ontology Language(OWL) is investigated. The use of OWL provides a web based standard for representation capturing the semantics of predictive models. A representation scheme for multiple types of decision trees, categorical, boolean, and numerical, was developed. The use of OWL allowed decision tree model classification, integration, and querying based on stated relationships between descriptor values. OWL-encoded decision trees represent the first step towards a standard predictive toxicology framework.

3.2 Introduction

Regulatory and health agencies face a challenge in assessing potential chemical toxicity. Manufacture and use of chemicals is increasing environmental exposure to poorly characterized chemicals and by-products. There is no feasible way to assess and characterize in vivo human risk for every chemical in use [1]. In an effort to minimize risk, regulatory agencies are turning to predictive in silico methods to prioritize chemical testing for more costly in vitro/in

vivo assays [1; 28]. Current in silico hazard estimation techniques use statistical and chemical functional analysis to determine potential chemical toxicity [10; 22; 28]. These techniques use data to derive models and relationships between chemical exposure and outcome. Predictive models are a representation of the rules for toxicity. However, the problem with predictive models is the format we capture the model in affects how and what they can be used for. The lack of a standard representation and frameworks for validation hinder development and re-use of predictive toxicology models [12]. A standard model format would improve interoperability, transparency, and extensibility of predictive models. The Web Ontology Language (OWL) is the formal knowledge representation language of the Semantic Web [50]. Capturing the semantics of models should enable automated classification, integration of models and descriptors based on shared relations, and invocation of web services to identify appropriate predictive models for a given query. Representation in OWL will provide a web based standard allowing interoperability, transparency, and extensibility of predictive models. Research into encoding predictive models using OWL is largely unexplored. Decision trees represent a standard predictive model used in toxicology.

Decision trees represent an ordered set of rules used for classifying data. Since the introduction of decision trees in toxicology by Cramer and Cramer in 1976, they have become a standard technique for predicting chemical toxicity such as genotoxicity, carcinogenicity, and mutagenicity [30; 64; 67; 68]. Decision trees can be derived from data or capture expert decision logic [26]. Data driven trees are automatically generated via statistical analysis of chemical toxicity data. Several statistical algorithms exist for building decision trees including ID3 [69], C4.5 [70], CHI-Squared Automatic Interaction Detector (CHAID) [71], and Multivariate

adaptive regression splines (MARS) [72]. The most widely used statistical method for constructing decision trees is the C4.5 algorithm [70]. The C4.5 algorithm is capable of handing discrete, continuous, and missing values [64]. However, trees generated via automated algorithms can be inherently difficult for human interpretation based on the complexity of the input data [26]. Extensive knowledge of the chemicals and mechanism of action is required. Expert decision logic models are useful for summarizing testing results into a concise and human interpretable rulesets [26]. Expert decision trees have found use in predictive applications, such as Oncologic for evaluating carcinogenic potential [73]. The use and variety of decision trees in predictive toxicology offer a starting point towards investigating OWL based representation.

The purpose of this research is to investigate leveraging OWL to encode decision tree predictive models. The goal is to develop an open standards based representation for predictive models. A standard web based representation will enable integration of models and descriptors, querying by relations between descriptors, and comparison of in silico predictions.

3.3 Methods

Determining the ability of OWL ontologies to be used as predictive toxicology models, decision trees were developed using WEKA [74] with experimental and molecular features from several chemical toxicology datasets CPDB [57], CCRIS [3], ToxCast [1], IARC [75], and HMDB [76]).

Three decision trees were developed from source datasets: Lipinski Rule of Five, Boolean Feature Based Tree, and Genetox bioassay decision tree. Trees were converted to OWL ontolo-

gies and tested based on the ability to capture the original decision tree structure, infer toxicity/bioactivity class, logical equivalence between data categories, and provide explicit explanations for classification. The DT2OWL software was written in Java and is downloadable and open sourced (https://github.com/dklassen/OWLClassifier).

3.3.1 Data Sources and Preparation

The development and analysis of OWL encoded decision trees required empirical and theoretical datasets. Datasets were constructed from CPDB [56], GeneTox [57], IARC [75], and HMDB [76] resources. All datasets, except HMDB, had been previously converted to linked data. Converted linked datasets were made accessible using the Virtuoso Open source TripleStore software available from http://www.openlinksw.com/wiki/main/Main (version 6.1.4). Software was complied and hosted on a MacBookPro using 2.2 GHz Intel Core i7 CPU using 8GB ram. Data sets were loaded into virtuoso triplestore instance using a custom script (http://code.google.com/p/semanticscience/source/browse/trunk/lib/php/virtuoso_load.php).

The Lipinski Rule of Five is a set of rules for determining drug-likeness [77]. The rule of five is a popular predictive test for screening potential drug candidates [78]. The decision tree was constructed from a training set of 7000 compounds from HMDB with chemical features computed using the Chemistry Development Kit [33] based on attributes determined tests outlined by Lipinski [77]. The training set was artificially generated based on chemicals known to follow the rule. The final classification of drug-likeness was determined via procedure outlined by Lipinski [79].

A feature based boolean decision tree, IARC Classification, was constructed based on 618 chemicals from the IARC carcinogenicity classification dataset. The International Agency for Research on Carcinogenicity (IARC) is a international organization for determining classification of known chemical carcinogens. Theoretical feature values, 318, were determined using the ToxTree API [80].

The dataset from the experimental bioassay decision tree was drawn from the GeneTox and CPDB experimental datasets. The list of chemicals was calculated via a query of the intersection between GeneTox and CPDB datasets contained in the linked data resource:

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dsstox_cpdb: <http://bio2rdf.org/dsstox_resource::>
PREFIX bio2rdf: <http://bio2rdf.org/resource:>
PREFIX genetox: <http://bio2rdf.org/genetox_resource:>
SELECT DISTINCT(?cas)
FROM <genetox>
FROM <ccris>

WHERE {
      [] a genetox:Substance ;
            genetox:hasCASRegistryNumber ?cas .
      [] a dsstox_cpdb:Substance ;
            ccris:hasTestSubstaceCASRN ?cas .
      }
```

Feature attributes were determined from the assays contained in GeneTox dataset. A majority rules vote was used to calculate a single experimental value per assay type per chemical. That is a feature was positive if the majority of experimental results were positive. Ties were broken based on priority, positive having greater priority over negative results. Chemical assays that

had "no conclusion" for result values were left out. It should be noted this procedure leads to generalization of outcome across test species.

3.3.2 Decision Tree Construction and Validation

Decision trees were constructed using the J48 algorithm [25] implemented by the open sourced machine learning software WEKA [74; 81]. The J48 algorithm is an open-sourced Java implementation of the C4.5 algorithm [70]. The C4.5 is the most commonly used algorithm for developing decision trees that include mixed data type or missing values [25]. Decision trees were constructed and pruned under default conditions. A 10 fold cross-validation was applied to establish the predictive ability of generated decision trees. Overall predictive ability was not taken into account as it was not the focus of this investigation. In total 3 decision trees were created: Lipinski Rule of Five, IARC Chemical Classification, and Experimental Bioassay.

3.3.3 Representation of Decision Trees as Ontologies

Decision trees, generated using the WEKA API, were converted to OWL ontologies using the OWL APIring), and value(i.e. true/false) leading to the considered node. The expression captures the sequence of rules executed along a decision tree path (Figure 3.1). [82]. Conversion of decision tree models, represented via DOT graph notation, was done by representing each decision node as an OWL class expression. A class expression for a given node was generated via the intersection the parent node, attribute(i.e benzene

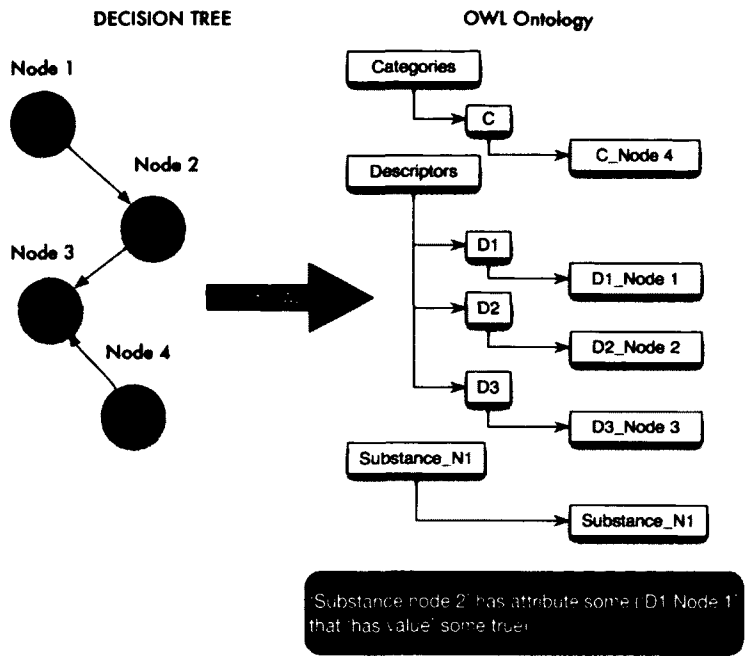

Figure 3.1 : Decision trees are represented by leaves and branches following a directed path to a terminal classification. Representation in ontology involved conversion of paths into class expression axioms which captured all unique paths in the source tree. Relationships shown by arrows in ontology represent "is-a" relationships. An example class expression axiom shows the rule between the path to node 2 from node 1.

The example decision tree in Figure 3-1 shows three decision nodes, D1, D2, and D3, with one terminal or classification nodes C. For D2 the equivalent class expression would be represented as:

> 'D2' equivalentClass 'D1 and 'has attribute' some ['attribute X' and 'has value' some TRUE]

Each class is node unique allowing identification of each path to classification. Classification classes were subclassed to the general classification via subclass axioms. Validation of this representation was shown in concurrent work [83].

3.4 Results and Discussion

For OWL classification ontologies to be of use they must be able to represent toxicity patterns mined from data. To test the representation of decision trees in OWL multiple decision trees from separate data sources were constructed. We utilized existing databases as well as linked data from the Semantic Web combining toxicity information from multiple resources. Generated decision trees were used to investigate model representation. The integration and comparison of models was demonstrated concurrently [83]. The contribution of this author was in the development of representation scheme and software for OWL-encoded decision trees.

3.4.1 Lipinski Rule of Five

The Lipinski Rule of Five is a set of rules for determining drug-likeness by considering 5 chemical attributes [70]. The simplicity and usefulness for initial drug development screening makes it an interesting challenge for representation in OWL. The Rule of Five was accurately constructed using WEKA and constructed training dataset(Figure 3-2).

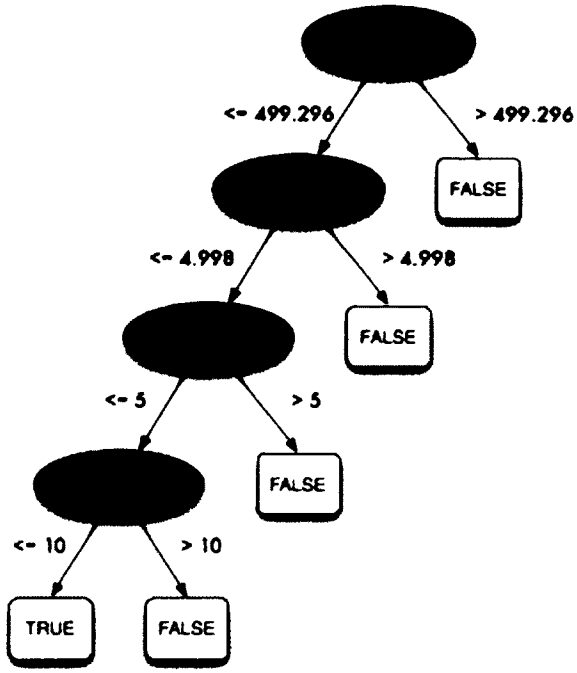

Figure 3.2: Diagram of Rule of Five decision tree. The tree was generated using a synthetic dataset 7000 compounds with final classification determined by Lipinski.

Conversion of Rule of Five to OWL presented the requirement of representing numerical comparison rules (equal too, greater than, less than). Numerical comparison rules were captured based on the example class expression:

> 'D2' equivalentClass 'D1 and 'has attribute' some ['attribute X' and
> 'has value' some float[<= 500]]

Representation involved datatype restrictions (integer, float, and double) and numerical comparisons. Representing and handling numerical comparison is vital to representing data driven QSAR based predictive models and classification of data [84-87]. Using OWL decision tree generator we were able to represent the Rule of Five (Figure 3-3).

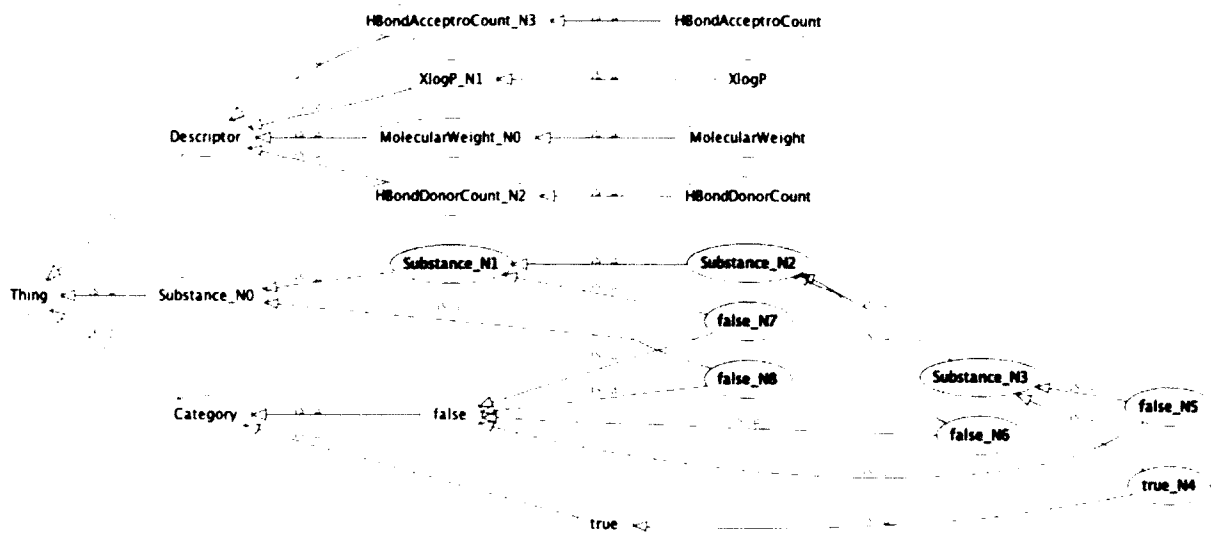

Figure 3.3: Representation of OWL encoded Lipinski Rule of Five generated using WEKA and OWL API. Each 'is-a' relationship is represented by an equivalent class axiom expression.

3.4.2 Feature-Based IARC Carcinogenicity Boolean Trees

The ToxTree API contains a collection of toxicity rules used in hazard estimation including Cramer Rules [30], Verharr [68], and Michael Acceptors [88]. Successful representation of these rules in OWL would demonstrate the application to current decision tree models. From the original 318 attributes considered a 5 rule decision tree was constructed (Figure 3-4). The node specific URI scheme implemented in the OWL decision tree generator prevented incorrect logical equivalence caused by repetition of rules. Each unique substance and attribute were made as subclasses of their generic counterpart. The pattern allowed unique paths to be represented while allowing ontology comparison of generic URI's :

Substance_N4 'subClassOf' Substance

toxicityrule_n4 'subClassOf' toxicityrule

substance_N4 'equivalentClass'

substance_N0 and 'has attribute' some (toxicityrule_n4 and 'has value' some FALSE

We were able to represent a 5 rule decision tree using OWL which was capable of classifying chemicals into the IARC chemical carcinogenicity classifications. The ability to classify linked data via OWL encoded decision trees was demonstrated in concurrent work [83].

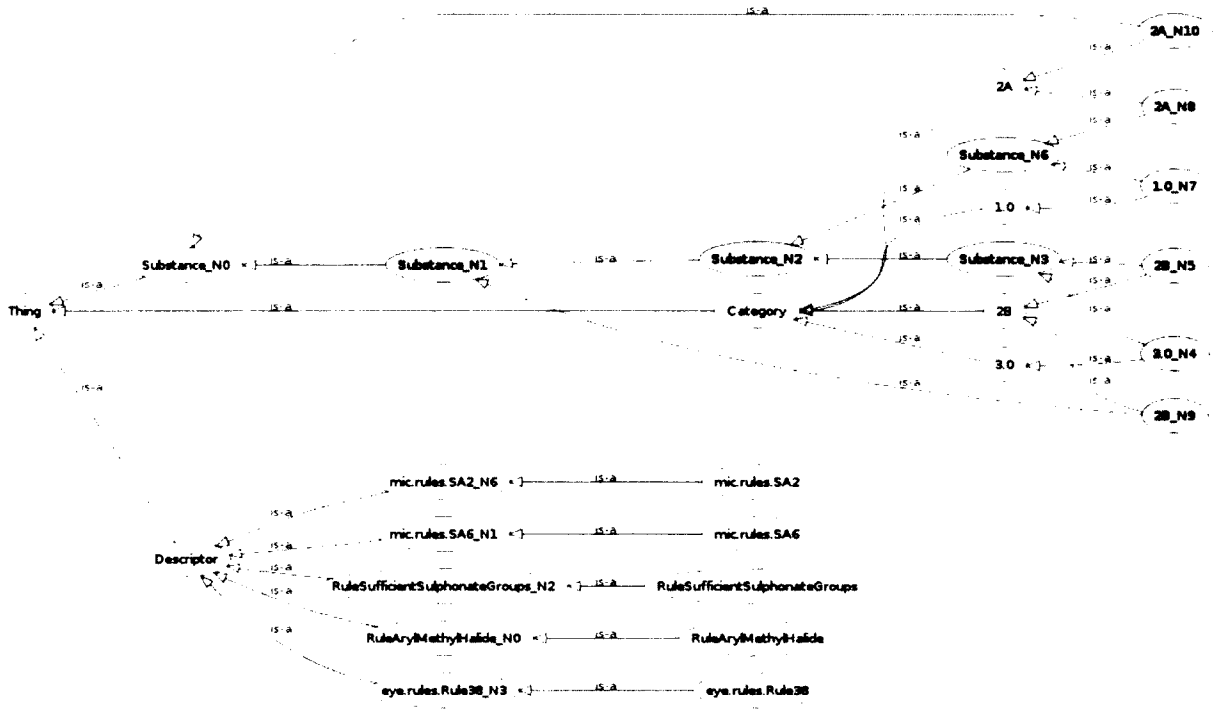

Figure 3.4: ToxTree feature based classification ontology which classifies chemicals according to IARC chemical carcinogenicity class (1, 2A, 3A, 4).

The use of rules extracted from the ToxTree API enabled the integration of decision trees and querying based on shared descriptors. The analysis of OWL-encoded decision tree integration was demonstrated from ToxTree rule set [83].

3.4.3 Experimental Bioassay Tree

An experimental bioassay decision tree was developed from experimental genotoxicity results and converted to an OWL ontology. Final classification was determined based on rat activity outcome for carcinogenicity from CPDB. However, automatic decision tree generation resulted in a single node "in vivo carcinogenicity assay", leading to classification ACTIVE/ INACTIVE. Lack of knowledge regarding the dataset and underlying relations highlighted the importance of data preprocessing prior to tree generation. Without insight into the reasoning behind the data structure it was impossible to preprocess data for decision tree building.

Data preprocessing is a necessary step for building predictive models. Access to data organization and structure is necessary to identify relevant attributes. Identifying relevant attributes means assessing the information gain towards a final classification. Lack of semantic interoperability of data sets hinders interpretation of data. The meaning underlying data needs to be captured in order to provide machine learning methods a way to guide data preprocessing [12]. The failure to develop a decision tree in this example highlights important directions to follow in the use of OWL-encode predictive models.

3.5 Conclusions and Future Applications

This work demonstrated for the first time the use and application of OWL-encoded decision trees for computational toxicology. We were able to capture data driven and expert determined decision trees. The representation scheme is extensible to numerical cut off ranges and boolean branch values. Path comparison was accomplished by creating node specific URI's which are parent classes of generalized attributes. This has the effect of allowing node specific paths to be identified while allowing classification based on generalized attributes.

Classification could be queried for based on generalized attributes [83]. The representation of OWL-encoded decision trees represents a step towards removing the barriers imposed by framework and software specific applications.

The representation of decision trees in OWL is the first step to a Semantic Web predictive toxicology framework. Model semantics have a role in logical interpretation of model classification. Using OWL encoded models allow the semantics to be leveraged for data analysis. Aside from the validation of our representation we did not investigate model generation or comparison. For example, grouping or splitting data based on shared qualities could be used to improve model accuracy. A future direction is the use of OWL encoded data relationships to aid generation of predictive models. Another direction lies in the logical comparison of OWL decision tree models. Calculating similarity of predictive models allows characterization of model diversity. If we know how models differ we can identify relevant models from a collection to a classify chemicals.

4 Mechanistic Toxicology Ontology (MechTox)

4.1 Abstract

The volume and diversity of toxicology data is a challenge for integration and analysis. Mechanisms of action represent the underlying process knowledge connecting chemical exposure to phenotypic outcome. In order to discover trends among chemicals that share similar mechanisms, this information needs to be machine accessible. A broad range of mechanisms of action were surveyed and formalized into the Mechanistic Toxicology (MechTox) Ontology. This ontology formalizes domain vocabulary for mechanisms of action, chemical participants, and phenotypic outcomes. We demonstrate the ability to represent complex mechanisms with varying degrees of granularity and certainty. This work sets the stage for more sophisticated research towards discovering novel links between chemical exposure and phenotypic outcomes.

4.2 Introduction

Interpretation of toxicology data is done in view of mechanistic toxicology knowledge [6; 8; 22; 28; 89]. Understanding how a group of chemicals elicit an effect can be used to characterize unknown chemicals. We use this knowledge to ground predictive models, provide context for model application, and provide common dialogue between model developers, toxicologists, and regulators [9]. Mechanistic toxicology knowledge covers information related to bioactivity, biological targets, chemical substructures, and experimental outcomes to give an overview of tox-

icity [64]. It represents the molecular basis of how a chemical elicits an effect; The mechanism of action [64]. Currently, mechanistic toxicology knowledge must be pieced together, from scientific journals, by an expert. The volume of knowledge makes this task nearly impossible increasing the likelihood of overlooking important data. The format of this knowledge is not open to computer analysis. We cannot efficiently leverage what is known to analyze data in context, eliminate false positives, and check new knowledge against a body of evidence.

The field of knowledge representation is the study of symbolic representation of knowledge, explicit statements about what we know, to enable reasoning and inference capabilities. Represented knowledge can be queried, manipulated, and checked for consistency against an entire body of represented knowledge. We can check facts for satisfiability (non-contradictory statements) and if they represent more generalized versions of other facts (subsumption). Capturing mechanism of action knowledge in a computer interpretable format would allow us to check whether our mechanistic knowledge is consistent. The goal of this chapter is to investigate the ability to capture and formally represent chemical mechanism of action knowledge. Knowledge representation differs from application hardcoded expert decision logic used in current systems.

The process of experimental toxicology produces knowledge. New facts are uncovered and used to build a picture of the overall toxic process. This process is an open ended task, the value of current facts change in light of future experimentation. As new facts are generated they must be compared to what is currently known and verified. The open ended nature of experimental toxicology is not captured and utilized in traditional databases or applications for

answering questions about toxicity. Current applications use expert knowledge in the form of hardcoded decision logic rules. The goal was to incorporate SAR based analysis and expert decision logic to determine possible toxicity. Several expert based rule systems have been developed and used successfully to guide experimental testing [73; 90]. However, the hardcoded format of this expert information is limited by:

1. No separation between application logic and encoded knowledge
2. Behaviour of the system is not entirely dependent on represented knowledge
3. Knowledge is not encoded in a format which is computer interpretable

Knowledge based systems differ from expert based rule systems by separating application logic from knowledge. Separated knowledge can be shared independently of application, reused, and repurposed. The abilities of the system are determined by examining represented knowledge. This property of knowledge based systems over hardcoded systems allows dealing with open-ended tasks or questions. A knowledge representation system can be told facts and adjust its behaviour accordingly. For example, we may know that a specific chemical binds to and inhibits a protein. However, that value of that knowledge does not become apparent until we know the function or role that protein plays. If the system is told the function or role of a protein we can infer that chemical binding may affect protein function. Finally, we can state explicitly the reason for a particular behaviour. All explicit statements of behaviour are both human and machine interpretable.

The goal of this work is to develop a representation of mechanisms of action, formalized in OWL, capable of representing molecular mechanisms of toxicity. The MechTox Ontology is a result of manual curation of current knowledge regarding mechanisms of action. Components from existing ontologies were used or extended to capture domain specific knowledge of mechanistic toxicology. This framework can be used to check new knowledge in light of what is currently known, validate facts, and infer connections between chemicals and mechanisms of action.

4.3 Methods

4.3.1 Development of Example Mechanism of Actions

The creation of a mechanistic toxicology ontology required a base of information to draw from. This information was used to detail the type, kind, and representational requirements of contained knowledge. Example mechanisms of action were compiled from scientific literature. The following section details a section of completed examples used for the knowledge representation of mechanistic toxicology. Each example is made up of a text base description of a mechanism of action of a specific chemical followed by the computer encoded description using Manchester Style Syntax [91].

4.3.2 Construction of Mechanistic Toxicology Ontology

The Mechanistic Toxicology Knowledge Base (MechToxKB) ontology was developed using the Protégé 4.1 OWL ontology editing software. The Semanticscience Integrated Ontology(http://semanticscience.org/ontology/sio.owl), SIO, was used as an upper level on-

tology. An initial ontology was constructed using vocabulary extracted from the set of example mechanisms of action. These terms were then matched to existing ontologies using the Bio-Portal [http://bioportal.bioontology.org/] RESTful API for ontology vocabulary searching. The HermiT 1.3.5 reasoner [61] plugin for Protégé was used to reason over knowledge base classes and instances.

Queries were expressed in the Manchester Style Syntax [91] and asked of the knowledge base using the Protégé OWL ontology editor software 4.1 DL Query built-in plugin. Concepts and properties described in the ontology have been colored blue. Queries are documented as part of example descriptions (see results section - mechanisms of action).

4.4 Results and Discussion

4.4.1 Mechanisms of Action

We collected a broad set of 22 chemical mechanisms/modes of action arising from chemical exposure. These mechanisms were obtained by a review of scientific literature and categorized based on those outlined by Boelsterli [64]. The 22 mechanisms can be categorized into 5 major categories below:

1. QSAR-like substructure-based mechanisms with known bio activities or phenotypes.
2. Mechanisms focusing on simple reactions , e.g. inactivation of specific enzyme.
3. Mechanisms that involve a chain of events from stimuli to response (e.g. cell signaling pathway)
4. Species-variant, organ variant and universal mechanisms of action.

5. Mechanisms (or signatures) that involve chemical perturbation of a system (chemical genomics / toxicogenomics) that leads to increased/decreased gene expression, and for which the gene products are known to participate in certain pathways and regulate certain processes in particular organisms.

Table 4.1 lists the mechanisms collected, which include the mechanism of DNA point mutation formation by Benzo[a]pyrene (BaP) and mechanism of AHR-mediated toxicity induced by TCDD (2,3,7,8-Tetrachloro-p-dibenzodioxin). Both chemicals are carcinogens, BaP, acts through a direct acting DNA mechanism while TCDD binds to AHR nuclear receptor altering signaling and gene expression.

Table 4.1 : Chemical mechanisms of action

Qsar-like substructure-based mechanisms with known bio activities or phenotypes	1	Mechanisms of CDK inhibition via 3,2' Bisindole compounds
Simple Reaction Mechanism	2	Mechanism of Aconite Inhibition by Fluroacetate
	3	Mechanism of Isoform selective based inhibition of CYP p450 by 8-methylxanthine furafylline
	4	Mechanism of Mitochondrial Complex II suicide inhibition by 3-nitropropinonic acid
	5	Mechanism of DNA polymerase alpha inhibition by aphidicolin
	6	Mechanism of Oxidative Phosphorylation Uncoupling by Carbonyl Cyanide M-Chlorophenylhydrazone (CCCP)
	7	Mechanism of Microtubule destabilization by Z-3,4,5'-trimethoxystilbene
Multi-Step Reaction Mechanism	8	Mechanism of AHR-mediated toxicity induced by TCDD

Table 4.1 : Chemical mechanisms of action

	9	Mechanism of DNA Point Mutation Formation by Benzo[a]pyrene
	10	Mechanism of Vitamin K Antagonism by Dicoumarol
	11	Mechanism of Dopaminergic Neuron injury by MPTP
	12	Mechanism of Nephrotoxicity by Mercury (Hg2+)
Species-variant, organ variant, and universal mechanisms of action	13	Mechanism of Cephaloridine Nephrotoxicity (CER-induced nephrotoxicity)
	14	Mechanism of Sodium/Potassium ATPase inhibition in cardiac cells by digitoxin
	15	Mechanism of Selective Blockage of Voltage Gated Na+ channels by tetrodotoxin
System Perturbation Mechanism	16	Mechanism of CDK inhibition by Indirubin-3' monoxide
	17	Mechanism of Hsp90 Inhibition by 17-allylamino-17-demethoxygeldanamycin(geldanamycin)
	18	Mechanism of mTOR inhibition by Rapamycin
	19	Mechanism of Calcineurin Inhibition by Cyclosporin A
	20	Mechanism of Estrogen Receptor Antagonism by Raloxifene
	21	Mechanism of HMG-CoA Reductase Inhibition by Atorvastatin
	22	Mechanism of Retinoic Acid Receptor (RAR) agonist bexarotene

4.4.2 Construction of Ontology

We first annotated the descriptions of mechanisms of action against 268 existing vocabularies using the NCBO annotated service [92]. For each vocabulary terms were matched to terms in the MechTox ontology in a process called mapping. 1033 mappings were obtained for 209 terms matching to 112 ontologies were identified based on syntactic similarity. After manual curation, we approved 813 mappings largely to NCI Thesaurus (9.7%), CRISP Thesaurus (6.4%), Medical Subject Headings (6.1%), SNOWMED Clinical Terms (6.0%), and Chemical Entities of Biological Interest (3.3%). From the initial set of terms 11.3% had no mappings to other ontologies in NCBO.

4.4.3 Knowledge Representation Requirements

The expressive requirements are the basic tools to express the knowledge represented in the examples. They represent the types of relations which can be expressed and reasoned over a particular language. Identifying the limitations of a particular language, from the requirements , will help determine how to best to model the knowledge (Table 4.2). Each mechanism of action was examined for two types of representational requirements (Table 5.2). The first were biochemical requirements or how to accurately represent what was taking place biochemically in the mechanism of action. The second was expressive, the expressive requirements are those pertaining to the representation language that would enable achieving the biochemical representational requirements. So in order to accurately model some biochemical requirement all the expressive requirements must be met.

To illustrate this process, mechanism 2: aconitase inhibition by fluoroacetate will be used. This use case describes the enzymatic conversion of fluoroacetate to 4-hydroxy-aconitate which binds and inhibits aconitase:

> "Fluoroacetate, an analog of acetate, is coupled to CoA to form fluoroacetyl-CoA by acetate thiokinase [93]. Fluoroacetyl-CoA is converted to yield 2-fluorocitrate through a condensation reaction with oxaloacetate. The (-)-erythro diastereomer ((2R,3R)-2-fluorocitrate) of 2-fluorocitrate is then converted to fluoro-cis-aconitate, which is followed by addition of hydroxide and with loss of a fluoride atom to form 4-hydroxy-trans-aconitate (HTn), which binds tightly, but not covalently to the enzyme [46; 64]. Binding of 4-hydroxy-trans-aconitate results in inhibition of aconitase enzyme followed by inhibition of the citric acid cycle [46]."

The goal is to represent the knowledge contained in these descriptions so that we may ask questions such as:

- "What proteins are involved in this mechanism?"
- "What derivative of fluoroacetate inhibits aconitase?"
- "What does inhibition of aconitase result in?"

By looking at the description above several biochemical processes can be identified, e.g enzyme catalyzed reactions and molecular complex formations. It can be seen that some process exists involving conjugation of fluoroacetate to coenzyme-a, a process catalyzed by acetate thiokinase. Therefore, there is a requirement to be able represent this biochemical process of converting fluoroacetate to fluoroacetyl-CoA. Other biochemical requirements present are those related to the entities involved in biochemical processes such as chemical derivatives and molecular complexes. This mechanism of action involves the transformation of fluoroacetate to several intermediates before aconitase is inhibited by 4-hydroxy-cis-aconitate. Accu-

rately representing these intermediates is necessary to capturing the mechanism. Keeping track of chemical derivatives requires us to be able to relate chemicals through time and process. In the example, fluoroacetate would have fluoroacetyl-CoA, 2-fluorocitrate, fluoro-cis-aconitate, and 4-hydroxy-cis-aconitate as derivatives. These chemicals are all related by being derived from fluoroacetate. This expressive requirement is called 'transitivity'. By processing each use case a list of requirements was identified (Table 4.2). For each requirement, independent of mechanism of action, a design pattern was developed to represent the information. The development of design patterns allows a consistent specification of encoding future information.

Table 4.2: Knowledge representational requirements determined from examination of mechanistic toxicology knowledge base examples.

Biochemical	Molecular complexes
	Association and Dissociation of Molecular Complexes
	Chemical similarity/variation and Metabolic Derivatives
	Biochemical catalysis
	Biochemical process rate change
	Biochemical transport
	Regulation of function
	Function or role-based classification
	Temporal ordering
Expressive	Existential restrictions
	Qualified cardinality restrictions
	Equivalence
	Transitive relations

Table 4.2: Knowledge representational requirements determined from examination of mechanistic toxicology knowledge base examples.

	Anti-symmetrical relations
	Symmetrical relations
	Mereology

4.4.4 Design Patterns: Expressive Requirements

Symmetry

Symmetrical relations are those which given a set X, for every member a and b in X it holds true for a is related to b and b is related to a, $\forall a,b \in X$, $aRb \Rightarrow bRa$.

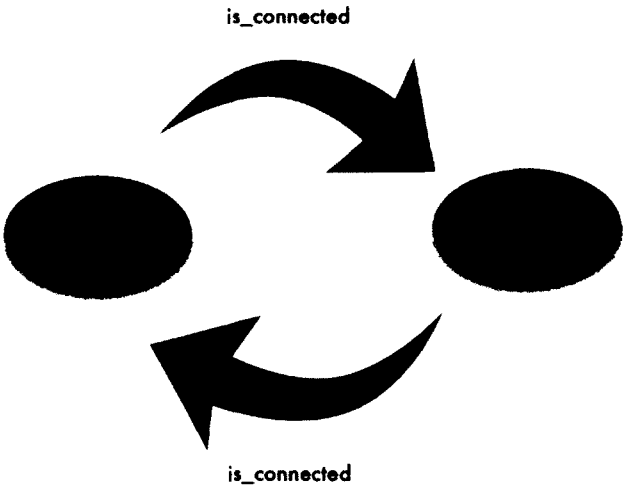

Figure 4.1: Example of symmetrical relations 'is_connected' relating two classes a and b.

These relations are used to express relations that occur where directionality, intent, is not required. When 4-hydroxy-cis aconitate forms a complex with aconitase it is stated to be 'is weakly interacting with'. The relation 'is weakly interacting with' is defined in SIO as describing a

non-covalent interaction. This is a symmetrical relation that allows us to specify aconitase is also weakly interacting, through non-covalent forces, with some 4-hydroxy-cis aconitate.

Equivalence

A more useful function is the ability to infer equivalence. Equivalence means statements or concepts refer to the same entity. OWL allows equivalence to be stated directly at the class, property, and instance levels through OWL:equivalentClass, OWL:equivalentProperty, and OWL:sameAs relations respectively. Class equivalence can be inferred through the construction of equivalent class axioms, or descriptions expressed using vocabulary in the ontology. The equivalent class axiom means that the attributes specified in the class expression are not only necessary, but sufficient. Thus, with automated reasoning, it becomes possible to find subclasses because they satisfy the equivalence axiom. When two classes are subclasses of one another, they are said to be equivalent.

Transitivity

Transitivity is a relational quality which means a relation, P, is transitive if for instance pairs (x,y) and (y,z) related by P the pair (x,z) can be inferred (If $\forall x,x=y, y=z$, then $x=z$). For mechanisms of action, it is important to consider chemicals as inputs across many reaction processes, e.g. chemicals in a metabolic pathway. Transitive relations allow us to represent a chemical and its derivates that are formed. Another important use for transitive relations is specifying components or parts of a thing. We can infer that if a chemical is connected to a protein subunit that it is connected to the protein. In SIO transitive relations are used in the design pattern of a 'transitive closure' to deal with derivatives and parts of things.

5.4.5 Design Patterns: Biochemical Requirements

An ontology design pattern is an example framework created using classes and relations to enable answering of specific types of questions [94-97]. The pattern represents an example model for a context or situation. Every time we want to represent a context, e.g. chemical binding, we use a specific pattern. Doing so ensures consistent knowledge representation. The following section deals with design patterns created to meet the biochemical requirements identified from the list of examples. Design patterns will be expressed using the Manchester Style Syntax [91] and use relations directly or extended from SIO (http://semanticscience.googlecode.com/svn/trunk/ontology/sio.owl).

Dispositions, Roles, Capabilities, and Functions

In order to accurately capture mechanisms, we must be able to specify the ability and behaviour of participants, including the conditions under which they are realized. SIO provides the vocabulary to specify roles, capabilities, dispositions, and functions. Roles are realizable entities that describe the behaviors, rights, and obligations that may be realized under circumstances. For instance, a chemical that is used as a reference in an experimental assay can be said to have a control role (positive or negative). Capabilities are realizable entities whose basis lies in one or more parts or qualities and reflects the possibility of an entity to act in a specified way under certain conditions or in response to a certain stimulus (trigger), normally towards some entity. For instance, a signaling molecule has the capability to bind to a receptor protein based on certain structural qualities. Dispositions are capabilities which have a tendency to be exhibited under certain conditions or in response to a certain stimulus (trigger). For instance, solu-

bility is a disposition of a chemical compound to dissolve when placed in fluid. Functions are capabilities that simultaneously satisfies some agentive design or natural selection. For instance, the function of the heart is to pump blood satisfying the purpose of the circulatory system.

In mechanism 2 [Table 4.1], 4-cis-hydroxy aconitate forms a complex with aconitase and inhibits its ability to form isocitrate from citrate. We can represent this as class expressions for the two entities , 4-cis hydroxy aconitate and aconite, involved in the formation of a complex:

```
'4-cis hydroxy aconitate' subClassOf
molecule and 'has disposition' some ('to inhibit' and
    'is realized in' some 'negative regulation of aco-
    nitase by 4-hydroxy-cis- aconitate')
```

The above class expression states 4-cis hydroxy aconitate is some molecule with the disposition to inhibit. The disposition is only realized during an inhibition process between 4-cis hydroxy aconitate and aconite.

```
aconitase subClassOf
protein and 'has disposition' some ('to be inhibited'
    and 'is realized in' some 'negative regulation of
    aconitase by 4-hydroxy-cis- aconitate')
```

Here we state the reverse that aconitase is a protein that has the disposition to be inhibited.

```
'biochemical inhibition of aconitase by 4-hydroxy-cis-
    aconitate' subClassOf
'negative regulation of aconitase' and realizes some
    ('to inhibit' that 'is disposition of' some '4-
    hydroxy-cis aconitate' and 'in relation to' some
    aconitase) and realizes some ('to be inhibited'
    that 'is disposition of' some aconitase and 'in re-
    lation to' some '4-hydroxy-cis aconitate)
```

Finally, we realize the dispositions for 4-cis hydroxy aconitate and aconitase through there participation in the process of negative regulation of aconitase. Both dispositions are stated to be

in relation to the entity that realizes them. In the example above two dispositions are represented, "to inhibit" and "to be inhibited". Now we can query for entities which have the disposition to inhibit aconitase:

```
Chemical and 'has disposition' some ('to be inhibited'
    and 'is realized in' some ('process' that 'has
    agent' some aconitase'))
```

The answer to this query would be "4-cis hydroxy aconitate" and any other chemicals that satisfy those conditions.

Chemical similarity/variation and Metabolic Derivatives

Toxicity may arise not only from a chemical an organism was exposed to, but also from its metabolic by-products. To answer questions about how a chemical is related to the overall toxic process it is necessary to capture their involvement in the overall toxic mechanism. Chemicals can be related through shared structure, function, or process.

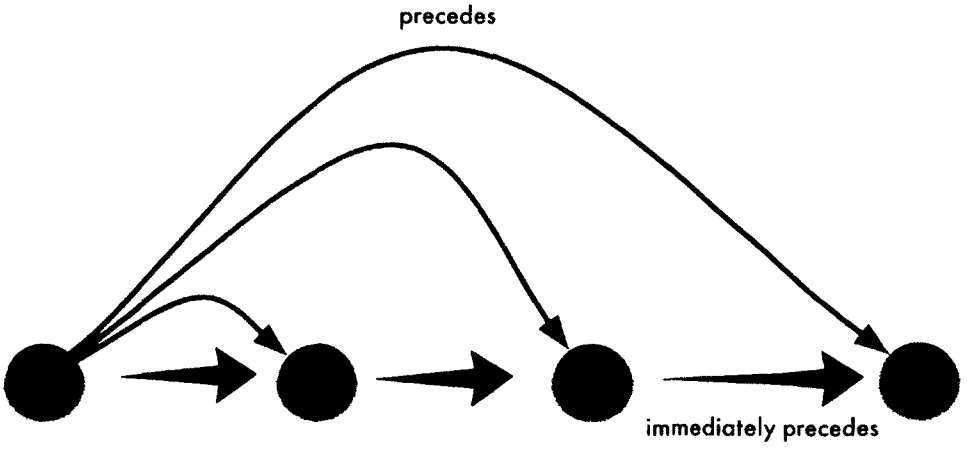

Figure 4.2: Example of how a transitive closure works. Red arrows indicate non-transitive relations between resources. Blue arrows indicate transitive relations.

Chemical derivatives are chemicals which are related through some biological process. At its simplest it is the binary relation between an input, or target chemical, and the output, or product, chemical of a process. The relations used to relate two chemicals that are, respectively, the input and output of a single process is 'is immediately derives from' and 'immediately derives into'. These relations are binary and non-transitive. The use of these two relations allows an ordering of a set of chemicals in a mechanism. However, since the relations are non-transitive it is not possible to query for all derivates of a chemical. Allowing querying for the complete set of chemicals is done using a 'transitive closure' (Figure 4.2). The binary, non-transitive relations 'immediately derives into' and 'is immediately derives from' are subclasses of 'derives into' and derives from' respectively. These parent classes are defined as transitive relations enabling querying from any chemical to another where a sequence path of processes exists.

OWL based ontologies operate on the basis of subsumption, or is-a relationships between classes. This principle allows us to represent classes as increasingly defined or specific children of parents. Therefore to represent similarities and variants, mechanism of action, biological target, or chemical substructure, parent classes can be defined to represent the parts that are shared in common with child classes. Specifying child classes was done through the creation of subclass axioms. This process resulted in all superclasses being inferred as opposed to stated.

Molecular Complex

A molecular complex is any entity made up of at least two chemical entities interacting non-covalently. Molecular complexes are ubiquitous from signal proteins binding chemical signals to enzymes binding substrates. A complex's behaviour and function is due to the parts making up the whole. Representing a molecular complex, its parts, order, and capabilities is necessary to capturing the toxic process. The following design patterns were developed to represent molecular complexes:

```
'example chemical' subClassOf 'chemical entity'

'example protein' subClassOf 'protein'

'example complex' subClassOf 'molecular complex'
   and 'has component part' some 'example chemical'
    and 'has component part' some 'example protein'
```

Here was have a molecular complex made up of a chemical and protein. We have no knowledge of how the components are connected. This could be a complex formed in an enzymatic reaction with a natural substrate or exogenous chemical. The parts making up the whole determine the identity and involvement in downstream events. The parts are specified using the SIO 'has component part' which is a relation between a whole and its part where the part is intrinsic to the identity of the whole. We can extend the above pattern to capture what we know of how the parts are connected:

```
'example chemical' subClassOf 'chemical entity'

'example protein' subClassOf 'protein'

'example chemical substructure' subClassOf 'chemical substructure'

'example binding site' subClassOf 'protein binding domain'

'example complex' subClassOf 'molecular complex'
   and 'has component part' some
```

```
      ('example chemical' and 'has component part' some
         ('example chemical substructure' and 'is weakly inter-
acting with' some 'example protein')))
    and 'has component part' some
    ('example protein' and 'has component part' some
        ('example binding site' that 'is weakly interacting
with' some 'example chemical')))
```

Now we have captured the knowledge of how each component is related to the whole. The relationship between parts has been specified using 'is weakly interacting with' to identify a non-covalent interaction. Addition relations have been specialized to capture specific interaction forces (covalent, non-covalent, Van-der-waals). We can further extend the pattern using relations to specify the spatial relation between components such as the containment of a chemical in a binding pocket of a protein catalytic subunit. SIO has a set of relations for handling relative spatial location including containment.

Association and Dissociation of Molecular Complexes

Molecular complexes are not static, they are constantly changing, forming, and separating. We need to be able to capture the separate processes of molecular complex formation and disassociation. Each pattern needs to specify the participants and how they are involved. Below is the pattern for capturing the process of molecular complex formation:

```
'example molecule' subClassOf 'molecule'
  and 'has disposition' some ('to be part of a group' and 'is
  realized in'     some 'example complex formation')

'example protein' subClassOf 'protein'
  and 'has disposition' some ('to be part of a group' and 'is
  realized in'     some 'example complex formation')

'example complex' subClassOf 'molecular complex'
```

```
    that 'has component part' some 'example molecule'
    that 'has component part' some 'example protein'

'example complex formation' subClassOf 'molecular complex forma-
    tion'
    and 'has participant' some 'example molecule'
    and 'has participant' some 'example protein'
    and 'has product' some 'example complex'
    and 'realizes' some ('to be part of a group' and 'is disposi-
    tion of' some 'example molecule
    and 'in relation to' some 'example protein')
    and 'realizes' some ('to be part of a group' and 'is disposi-
    tion of' some 'example protein'
    and 'in relation to' some 'example molecule')
```

The pattern describes the participants of the complex formation. In SIO 'has participant' is a relation between a processual entity and another entity which is involved in but unchanged due to the process. The design pattern for molecular complex and molecular complex formation are connected. The process captures participants and capabilities, while the complex captures knowledge of parts.

Temporal Ordering

A mechanism of action is a series of events. Each event occurs in a linear sequence to produce an outcome. Querying how the toxic process unfolds required a set of relations to define event order. Relations would need to specify if processes overlap (one occurs immediately after another) or are distinct in time. The pattern to order temporal events, in SIO, is similar to that used to capture chemical derivates by constructing a transitive closure. Two sets of relations, one transitive and one not are defined with domain and range 'process'. The transitive relations, 'precedes' and 'is preceded by' are defined in SIO to specify two related processes which do not overlap temporally. The relations 'immediately precedes' and 'is immediately pro-

ceeded by' were subclasses of the previous relations to represent processes that overlap temporally. These relations are non-transitive. A design pattern using the relations is seen below:

```
'example process 1' subClassOf 'Processual Entity'
    and 'immediately precedes' some 'example process 2'

'example process 2' subClassOf 'Processual Entity'
    and 'is immediately preceded by' some 'example process 1'
    and 'immediately precedes' some 'example process 3'

'example process 3' subClassOf 'Processual Entity'
    and 'is preceded by' some 'example process 1'
    and 'is preceded by' some 'example process 2'
```

Using the design pattern we can query for all events that come before or after a process. Furthermore, we can specify we are only looking for patterns that occur immediately after or before an event. In toxicology, not all events happen in a linear sequence or we may not know how the exact chain of events that leads to an outcome. For instance, in biochemical inhibition of a protein we may know a chemical binds to a protein and immediately inhibits all processes that protein is involved it. On the other hand a chemical binding to a signaling protein may elicit a downstream effect that results in some increased gene expression. All we know is increased gene expression is preceded by upstream chemical binding. We can capture cause and event without knowing the full details.

Chemical Transport

Toxicity is not just local but a whole body phenomena. A toxin has to move from exposure site to endpoint. Understanding how a chemical is transported is important to understanding overall toxicity. As a chemical is transported it may be modified and converted to more toxic

metabolites. Those modifications may convey the potential for toxicity that is specific to an organ, tissue, or cell type. Representing the movement of chemicals through a system, organ, or membrane required a multi process design pattern. SIO provides several relations for specifying the temporal components of a process such as start and end points:

```
'example transport process' subClassOf 'transport process'
    and 'has component start' some
          ('process start' and 'has participant' some 'example
    chemical' and 'is located in' some 'example spatiotemporal
    region 1')
    and 'precedes' some ('example protein transport' that 'has
    agent' some    'example protein' and 'is located in' some
    'example membrane')
    and 'precedes' some ('process end' and 'has participant'
    some 'example  chemical' and 'is located in' some 'example
    spatiotemporal region 2')
```

The example pattern shows transport of a chemical across a membrane via a transport protein. Separating the transport process into separate events we can query based on chemical start and end points, transport type (active or passive), and proteins involved.

Process Rate Change

Chemicals often exert a toxic influence by altering the rate of a biological process. If the catalytic rate of an enzyme is modified it will affect all downstream biological processes. A process, in SIO, is a physical entity that exists only in time and made up only of other process parts. The process is defined by having a definite start and end point in time and only have physical participants. A toxin that affects a process rate does so relative to some "normal" rate. The "normal" rate is defined as a relative measure as it is not determined from an individual measure or experimental collection of results. This allows us to define a class to assign an effect given

chemical exposure. The design pattern below was designed to capture knowledge of rate change that lead to different outcomes:

```
'process rate' subClassOf 'processual property'

'example decreased process' subClassOf ' example normal process'
    and 'has quality' some ('Process rate' and 'is lesser than'
    some ('process rate' that 'is  quality of' some ' example
    normal process
```

The design pattern above is a relative comparison of process with no measurement values considered. The development of the design pattern required the creation of comparison relations in SIO , "is lesser than" and "is greater than".

Regulation of Function

Biochemical inhibition, in SIO, is a molecular interaction process that decreases the catalytic rate of a target enzyme. Enzyme inhibition can be through direct physical interaction with a catalytic or allosteric site. Often these interactions involve specific amino acid side chains located on the enzyme. The result of these processes is the modification of all processes the target is involved in (i.e. a process rate change). The design pattern created is an extensible pattern to support various states of knowledge:

```
'example inhibition' subClassOf 'biochemical inhibition'
    and 'has target' some 'example enzyme'
    and 'has participant' some 'example chemical'
    and 'results in' some 'decreased enzymatic conversion of A
    to B'
```

In the first example pattern only the participants and outcome are known. We know nothing of the physical interaction between chemical and enzyme. The pattern can be extended to capture the molecular interaction between the chemical and enzyme as follows:

```
'example complex' subClassOf 'molecular complex'
    and 'has component part' some 'example enzyme'
    and 'has component part' some 'example chemical'

'example inhibition' subClassOf 'biochemical inhibition'
    and 'has target' some ('example enzyme' and 'is component
    part of' some 'example complex')
    and 'has ' some ('example chemical' and 'is component part
    of' some 'example complex'
    and realizes some ('to inhibit' that 'is disposition of'
    some 'example chemical' and 'in relation to' some 'example
    enzyme')
    and realizes some ('to be inhibited' that 'is disposition
    of' some 'example enzyme and 'in relation to' some 'example
    chemical'
    and 'results in' some 'decreased enzymatic conversion of A
    to B'
```

By including the molecular complex design pattern the specific interaction between the chemical and enzyme can be captured and queried.

Biochemical catalysis

Biochemical regulation is a process that alters the frequency, rate, or extent of a downstream biochemical process. The distinction with regulation of catalytic activity is it does not involve an enzyme and reflects a modification of a downstream process. It is a regulation of process compared to a regulation of capability. An example of a negative regulation is the negative regulation of tumorigenesis, the process of tumor formation. The design pattern is specified below:

```
'example negative regulation' subClassOf 'negative regulation'
    and 'has agent'  some 'example chemical'
    and 'has target' some 'example protein'
    and 'results in' some  'example result'
```

The 'has agent' relation is used to specify an entity that is involved in a process but not affected. The 'has target' relation is used to specified a participating entity who undergoes a change process or transformation.

Function or Role based Classification

Knowledge of chemical targets and substrates is important to characterizing mechanism of action. Specifying the target and product of processes involved in mechanisms of action allows development of a signature which can be used for data mining. A target is some entity that undergoes a change or transformation as being part of a process. Targets are specified using the relations , "is target in" and "is target of" . A target is a chemical that undergoes a change to produce a product. Products are specified in much the same way using , "is product in" and "is product of" relations. The design pattern allows querying from both chemical and process:

```
'example target' subClassOf 'chemical entity'
    and 'is target in' some 'example process'

'example product' subClassOf 'chemical entity'
    and 'is product of' some 'example process'

'example process' subClassOf 'process'
    and 'has target' some 'example target'
    and 'has product' some 'example product'
```

Knowing the targets and products of processes we can using that knowledge to infer target and substrate membership. If a chemical is a target in a process we can classify it as a target.

The specific type of target can be refined based on the processes the chemical is involved in. We can the query between types of targets such as cytochrome p450 targets or nuclear receptor targets. Inferring target and substrate type is done using equivalent class axioms:

```
'target' subClassOf 'molecule'
    and 'is target in' some 'process'

'substrate' subClassOf 'molecule'
    and 'is product of' some 'process'
```

The same pattern can be used to allow querying for chemical based on class of toxin. A chemical can be classified based on how a chemical interacts within a mechanism.
By using equivalent class axioms we state the exact requirements for membership to a particular class. Membership is inferred rather than directly stated. The advantage is the hierarchy of targets and products is built from the descriptions and reflect the captured processes. The descriptions, used to build the hierarchy, can be used to investigate common linkages between chemicals and provide explicit explanations of toxicity.

```
'CYP inhibitor' equivalentClassOf 'molecule'
    that 'is agent in' some ('process' that 'has participant'
    some 'CYP protein' and realizes some ('to be inhibited' that
    'is disposition of' some CYP protein)
```

The design pattern above details the axioms used to classify chemicals that are 'Cytochrome P450 inhibitors'. The pattern makes use of chemical dispositions and participation in processes that involve inhibition of cytochrome p450 proteins.

4.4.6 Querying MechTox Ontology

The ability to meet the expressive requirements is determined by the description language. OWL was found to meet all the expressive requirements outlined in the example mechanisms of action analysis (Table 4.2). Table 4.3 lists example class based queries and expressive and biochemical requirements. The ability to query and retrieve specific classes is validation of constructed design patterns used to build the ontology.

Query 1. Which mechanisms involve inhibition of aconitase?

Class Expression: 'mechanism of action' and 'has component part' some ('biochemical inhibition' that realizes some ('to be inhibited' and 'is disposition of' some 'aconitase'

This query returns the set of individuals that have some (at least one) biochemical inhibition process where the specific capability of aconitase to be inhibited is realized.

Query 2. Which chemicals are derivatives of Benzo[a]pyrene?

Class Expression: molecule that 'derives from' some 'benzo[a]pyrene'

This query returns the set of individuals which are molecules and have derived from benzo[a]pyrene. The query makes use of the transitive relations to identify all molecules related to benzo[a]pyrene. The result of this query includes molecule classes of Benzo[a]pyrene-7,8-epoxide, trans-benzo[a]pyrene-7,8-dihydrodiol, and Benzo[a]pyrene-7,8-dihydrodiol 9,10-epoxide.

Query 3. Which chemicals decrease the rate of formation of mevalonic acid?

Class expression: molecule that 'has disposition' some ('to decrease rate of formation' and 'is realized in' some (process that 'has product' some 'mevalonic acid'))

This query returns the set of individuals which are molecules and have the disposition to decrease the rate of formation of a process. This disposition is only realized in a process that involves the production of mevalonic acid.

Query 4. What protein complexes contain Hsp90 and Aryl Hydrocarbon receptor?

Class expression: 'molecular complex' that ('has component part' some Hsp90) and 'has component part' some 'aryl hydrocarbon receptor')

This query is a conjunctive query returning the set of individuals that are molecular complex and have both an Hsp90 and Aryl hydrocarbon receptor as component parts.

Query 5. What chemicals bind to a human retinoic acid receptor?

Class expression: 'molecule' that 'is weakly interacting with' some ('binding site' that 'is component part of' some ('retinoic acid receptor' that 'is part of' some human))

This query returns the set of chemicals stated to be non-covalently bound to some binding site of a retinoic acid receptor belonging to a human.

Query 6. Which chemicals are HMG-CoA reductase inhibitors?

Class expression: 'inhibitor and 'has disposition' some ('to inhibit' and ('is realized in' some ('biochemical inhibition' and (realizes some ('to be inhibited' and ('is disposition of' some 'HMG-Coa reductase'))))))

This query returns the set of chemicals having the disposition 'to inhibit' is realized in a biochemical inhibition process where HMG-CoA reductase is inhibited. This class expression is an example of an equivalent class expression of the class "HMG-CoA Reductase Inhibitors". This is an example of functional or role based classification. It states the necessary and sufficient conditions for a molecular entity to be classified as a HMG-CoA reductase inhibitor.

Table 4.3: Example queries posed to MechTox Ontology along with class expressions. Each query is associated with the query features that are used.

(1) Which mechanisms involve the inhibition of aconitase?	mechanism of action' that 'has component part' some ('biochemical inhibition' that realizes some ('to be inhibited' and 'is disposition of' some 'aconitase'	mereology, dispositions
(2) Which chemicals are derivatives of Benzo[a] pyrene?	molecule that 'derives from' some 'benzo [a]pyrene'	transitivity
(3) Which chemicals decrease the rate of formation of mevalonic acid?	Molecule that 'has disposition' some ('to decrease rate of formation' and 'is realized in' some (process that 'has product' some 'mevalonic acid'))	dispositions, biochemical processes
(4) What protein complexes contain Hsp90 and Aryl Hydrocarbon receptor?	molecular complex' that ('has component part' some Hsp90) and 'has component part' some 'aryl hydrocarbon receptor')	mereology, molecular complex
(5) Which chemicals bind to the retinoic acid receptor?	'chemical entity' that 'is weakly interacting with' some ('binding site' that 'is component part of' some ('retinoic acid receptor' that 'is part of' some human))	dispositions, molecular complex, mereology
(6) Which chemicals are HMG-CoA reductase inhibitors?	inhibitor and 'has disposition' some ('to inhibit' and ('is realized in' some ('biochemical inhibition' and (realizes some ('to be inhibited' and ('is disposition of' some 'HMG-CoA reductase'))))))	functional or role based classification, dispositions, regulation of function.

4.5 Example Mechanisms of Action

Below is a collection of 5 representative example mechanisms of action. Examples contain a the context describing the general use of the chemical and mechanism of action description.

4.5.1 Mechanism of Aconitase Inhibition by Fluroacetate
Example of mechanistic inhibition

Context

Sodium fluroacetate is a metabolic poison found it over 40 plants in Australia, Brazil, and Africa. Symptoms of fluroacetate poisoning begins after 30 min from exposure and include nausea, vomiting, stomach pain, sweating and confusion.

Mechanism

Fluroacetate, an analog of acetate , is coupled to CoA to form fluoroacetyl-CoA by acetate thiokinase [93]. Fluoroacetyl-CoA is converted to yield 2-fluorocitrate through a condensation reaction with oxaloacetate. The (-)-erythro diastereomer ((2R,3R)-2-fluorocitrate) of 2-fluorocitrate is then converted to fluoro-cis-aconitate, which is followed by addition of hydroxide and with loss of a fluoride atom to form 4-hydroxy-trans-aconitate (HTn), which binds tightly, but not covalently to the enzyme [98]. Binding of 4-hydroxy-trans-aconitate results in inhibition of aconitase enzyme followed by inhibition of the citric acid cycle.

4.5.2 Mechanism of DNA Point Mutation Formation by Benzo[a]pyrene
Example of direct acting DNA modification

Context

Benzo[a]pyrene (BaP) is a 5 member ring polycyclic aromatic hydrocarbon that is present in diesel fumes, cooked food, cigarettes, and smoke. Metabolites of BaP have been found to be mutagenic and carcinogenic. BaP itself has been classified as a group 1 human carcinogen by the International Agency for Research on Cancer (IARC).

Mechanism

Benzo[a]pyrene is metabolically activated through the 'bay region dihydrodiol epoxides pathway [99]. This pathway involves three enzyme-mediated reactions involving CYP p450, epoxide hydrolase(EH), and a second CYP p450 [99]. Metabolic activation begins with epoxidation of the 7-8 carbon double bond of Benzo[a]pyrene (BaP) to form Benzo[a]pyrene-7,8-epoxide. The epoxide is hydrolyzed, by epoxide hydrolase, to form trans-benzo[a]pyrene-7,8-dihydrodiol. The final stage of the pathway involves a CYP p450 enzyme catalyzed epoxidation to generate benzo[a]pyrene-7,8-dihydrodiol 9,10-epoxide (specifically 8alpha-dihydroxy-9alpha,10alpha-epoxy-7,8,9,10-tetrahydrobenzene[a]pyrene) [100]. Benzo[a]pyrene-7,8-dihydrodiol 9,10-epoxide is forms n2-deoxyguanosine adducts with DNA which leads to point mutations and mutagenesis [101]. Point mutations have a direct link with the neoplastic initiation phase of carcinogenesis.

4.5.3 Mechanism of Cephaloridine Nephrotoxicity (CER-induced nephrotoxicity) Example of organ selective toxicity

Context

The kidneys are the primary route of excretion for xenobiotics, as such they are a large target for organ selective toxicity. Cephalosporins belong to the class of beta-lactame antibiotics first isolated from Cephalosporium acremonium. Cephalosporin is a nephrotoxin due to its selective accumulation in the proximal tubular epithelia of the kidney. Another Cephalosporin, Cephaloridine, is not a nephrotoxin attributed to a missing pyridine group. Selective accumulation is the function of the organic anion transporters(OAT) which are expressed in high levels specifically in the kidney. OAT transporters show a broad substrate specificity and transport many chemically unrelated compounds. The criteria to be transported by OAT are possession of a hydrophobic moiety, the ability to form hydrogen bonds, and the presence of ionic or partial electrical charges.

Mechanism

The mechanism is based on the concentrating uptake of the drug into the proximal tubular epithelial in the kidney cortex. Cephaloridine, a zwitterion at physiologic conditions, is transported across the basolateral membrane via the organic anion transporter, OAT1, into proximal tubular cells. However, Cephaloridine does not move as easily across the apical membrane. This is the result of a lower affinity for Cephaloridine by OAT4 leading to increase in concentration. Contrast with anionic Cephalothin molecule which is missing a pyridine group and can move from proximal tubule cells across apical membrane more efficiently (Km 0.20 vs 3.63 for OAT 4 Cephaloridine/Cephalothin) [64; 102]

. The differences in affinity have been to Cephalothin missing a pyridine group making it anionic at physiological conditions. Nephrotoxicity induced by Cephaloridine is characterized by acute proximal tubular necrosis [102]. The nephrotoxic effects are associated with increased production of free radical species and depletion of glutathione resulting in oxidative stress [103].

4.5.4 Mechanism of Dopaminergic Neuron injury by MPTP
Example of cellular transport and selective accumulation

Context

1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine is a neurotoxin resulting in Parkinson like symptoms. The effect is caused through targeting of the dopaminergic neurons in the substantia nigra.

Mechanism

1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) is a lipophilic compound capable of crossing the blood brain barrier and absorbed by glial cells. In astrocytes, MPTP is converted by monoamine oxidase to MPDP, which then auto oxidizes to MPP+ [64]. MPP+ accumulates in dopaminergic neurons due to selective uptake by the Dopamine transporter (DAT) resulting in cell death [64]. MPP+ toxicity due to ability to bind to and inhibit complex 1 activity in mitochondria [64]. There is a high concentration of dopaminergic neurons in the substantia nigra, cell death causes Parkinson's like symptoms.

4.5.5 Mechanism of AHR-mediated toxicity induced by TCDD (2,3,7,8-Tetrachloro-p-dibenzodioxin)

Example of nuclear receptor mediated toxicity

Context

2,3,7,8-Tetrachloro-p-dibenzodioxin (TCDD) is a member of the chemical dioxin family and classified as group 1 human carcinogen by IARC. TCDD is a non-genotoxic carcinogen whose primary mechanism involves acting through binding to the Aryl Hydrocarbon receptor(AhR) [104]. Binding of the AhR receptor results in altered gene expression.

Mechanism

The ligand free Aryl hydrocarbon Receptor (AhR) is located in the cytosol as part of the Aryl hydrocarbon Receptor complex. This complex is composed to two Hsp90 chaperones bound to AhR PAS domains, one prostaglandin E synthase 3 protein, one immunophilin-like protein hepatitis B virus X-associated protein 2 (XAP2) bound to AhR nuclear localization sequence, and one AhR-activated 9 (ARA9) [105]. TCDD binding to the AhR ligand binding domain causes XAP2 to dissociate exposing the nuclear localization sequence located in the bHLH region [106]. The TCDD-ArH complex is then localized to the nucleus where the two Hsp90 proteins dissociate from the PAS domains allowing binding of ArNT [107-109]. The TCDD-AhR-AnRT complex binds to dioxin-responsive-elements (AhR responsive elements) containing the core 5'-GCGTG-3'[36] and consensus sequence 5'-T/GNGCGTGA/CG/CA-3' [107].

4.5 Conclusions and Future Directions

In this work, we demonstrated the use of OWL encoded mechanism of action knowledge to answer questions of toxicity. We were able to capture diverse toxicity process knowledge in a consistent and extensible framework. Toxicity information can now be leveraged to classify toxins according to mechanism of action. Formal descriptions were leveraged to increase interoperability with other OWL-ontology encoded information.

Knowledge of molecular mechanism of action allows interpretation of experimental toxicology data [6]. Using the MechTox ontology it is possible to capture, query, and interrogate toxicology data based on mechanism of action knowledge. The use of Semantic Web to represent knowledge allows better use of current knowledge for open ended question answering. These systems can check the validity of data and adapt to new information. The value of knowledge is increased by being interpretable by human and computer. Capturing prior knowledge of mechanistic toxicology exposes it for use in relational learning techniques [110-112]. A novel direction lies in utilizing ontologies to aid machine learning techniques for model induction.

5 Inference of Novel Features for Data Mining: Application of the MechTox Ontology

5.1 Abstract

A common issue with machine learning methods is feature construction and aggregation [66]. Simplifying complex toxicological endpoint data requires construction of relational or aggregated features [29]. In this work, we present a proof of concept on the use of ontologies to aid construction of features for machine learning. Mechanism of action knowledge, encoded in OWL, was used to infer a chemical derivative feature based on a genotoxic endpoint. Use of this feature resulted in improved accuracy and reduced complexity of generated classification rules. This application of the MechTox ontology demonstrates the potential of ontologies to improve machine learning techniques.

5.2 Introduction

As the diversity of data increases so does the number of available features for data mining [113]. A feature is an attribute of data which can be used to generalize or develop patterns in data [66]. In toxicology chemical substructure represents features which can be associated with toxic potential (e.g. heterocyclic structures). Making sense of and simplifying complex toxicology data requires the construction of relational and aggregate features [29; 113]. Tech-

niques exist to do this automatically [114-116]. However, these techniques do not make use of prior knowledge of data and underlying process. Ontologies represent prior knowledge which can enhance the process of learning relations in data [111; 112; 117]. Unlike relational databases, Ontologies contain the necessary semantic information to guide feature construction [112; 118]. The Mechanistic Toxicology (MechTox) Ontology was designed to capture mechanism of action knowledge. Mechanism of action knowledge captures the underlying process connecting chemical exposure to expressed phenotype. This information represents patterns in toxicity which can be leveraged to generalize and aggregate data features for model development.

In this work, we show a proof of concept for how relational features can be developed to generalize model development. A single feature based on knowledge of genotoxic metabolites was created to provide additional information for classifying potential genotoxins. This feature was constructed based on inferring relations between input chemical and known genotoxins. The identity of genotoxins was based on stated knowledge of chemical derivatives contained in the MechTox ontology.

5.3 Methods

An artificial dataset was created containing structural features used in genotoxicity prediction [67; 80]. Values for chemical features were determined using the ToxTree predictive software and Cramer and Benigni/Bossa rulesets [119]. The Octanol/Water partition coefficient (Log Kow) was predicted by the US Environmental Production Agency's EPI SUITE [120]. Evaluation

of constructed features was done by comparing decision trees constructed using the J48 algorithm [25] implemented by the open sourced machine learning software WEKA [74; 81]. Trees were evaluated based overall accuracy and ability of generated feature to simplify model rulesets. Accuracy measures for example datasets were generated using a 2-fold cross validation.

The Mechanistic Toxicology (MechTox) ontology was used to develop a chemical derivative feature based on mechanism of action. The HermiT 1.3.5 reasoner was used to reason over and classify concepts contained in the MechTox Ontology. Class expression queries were constructed to determine the chemical derivative feature value for a given chemical in the artificial dataset. Feature value was TRUE if a known genotoxic derivative with mechanism existed and FALSE otherwise. Queries were asked of the ontology using the DL query built-in expression tester.

5.4 Results and Discussion

To investigate how an ontology can be used to construct relational features an example training set was created. The training set was used to demonstrate how knowledge contained in the MechTox Ontology could be leveraged to simplify toxicology data. A single feature based on knowledge of chemical derivatives was created based on this knowledge.

5.4.1 Chemical Derivative Relational Feature

Degradation of xenobiotics may lead to the production of reactive intermediates [121-125]. Identifying reactive intermediates is useful for assessing potential risk of a chemical related to endpoint [126]. The MechTox ontology contains information regarding metabolism of chemicals related to genotoxic endpoints. Table 5.1 shows an example list of structural features relevant to potential carcinogenicity/genotoxicity [67; 80].

Table 5.1: Inference of a chemical relational feature based on being a known metabolic precursor of a chemical ACTIVE for the endpoint Genotoxicity. Inferred relation is shown highlighted in red.

Chemical	Log Kow	Aromatic?	Heterocyclic	Open Chain?	Lactone or cyclic diester		Genotoxic
benzo[a]pyrene	7.82	TRUE	FALSE	FALSE	FALSE		ACTIVE
2-Methylimidazole	4.01	FALSE	TRUE	FALSE	FALSE		INACTIVE
Geranyl pyrophosphate	2.42	FALSE	FALSE	TRUE	FALSE		INACTIVE
Artemisinin	9.60	FALSE	TRUE	FALSE	TRUE		ACTIVE
Estradiol	3.94	TRUE	FALSE	FALSE	FALSE		ACTIVE
naphthalene	3.30	TRUE	FALSE	FALSE	FALSE		ACTIVE
Benzene	2.13	TRUE	FALSE	FALSE	FALSE		ACTIVE
1,3-butadiene	1.99	FALSE	FALSE	TRUE	FALSE		ACTIVE
lucidinprimeveroside	1.17	TRUE	FALSE	FALSE	FALSE		ACTIVE
hexane	3.90	FALSE	FALSE	TRUE	FALSE		INACTIVE
5-methylchrysene	6.07	TRUE	FALSE	FALSE	FALSE		ACTIVE
1,3-propane sultone	-0.28	FALSE	TRUE	FALSE	FALSE		ACTIVE
Tetrachloroethylene	2.97	FALSE	FALSE	FALSE	FALSE		INACTIVE

Querying the MechTox ontology we were able to infer an additional feature related to metabolites and input chemical(Table 5.1). The value of this feature was determined based on me-

tabolites being known genotoxins. The inclusion of the relational feature, reactive metabolite, simplified decision tree construction while retaining model accuracy (Table 5.2). The inferred feature had greater predictive value to the composite rule of "Aromaticity?" and "Log Kow"(Table 5.2).

Table 5.2: Results of feature input for decision tree building. Values are weighted averages for ACTIVE/INACTIVE classes.

Without Inferred	0.642	0.615	0.625	61.5	38.5	aromatic? = True: Active (5.0) aromatic? = FALSE | LOGKow <= 2.13: ACTIVE(3.0) | LOGKow > 2.13: INACTIVE (5.0/1.0)
With Inferred Feature	0.747	0.692	0.704	69.2	30.8	Reactive_metabolite =TRUE: ACTIVE(10/1) Reactive_metabolite = FALSE:INACTIVE(3.0)

The advantage of the rule based on "reactive_metabolite" lies in the ability to simplify the data to a single feature. Determining the value of this feature in a real world example is challenging. Xenobiotic metabolism plays an important role in determining toxicity [126-128]. However, inferring the relation to a reactive intermediate did not take into account the complex kinetics of xenobiotic metabolism.

This value of the feature was determined based on the class expression query:

'chemical' that 'derives into' some genotoxin

The query returns the set of chemicals inferred to be derivatives of some genotoxin. To answer this query, two inferences were required based on information contained in MechTox.

First, which chemicals are equivalent to genotoxin. Second, which chemicals derive into a genotoxin.

The class genotoxin was defined using class expressions to entail the set of chemicals involved in some process resulting in DNA damage. For instance, Benzo[a]pyrene-7,8-dihyrodiol-9,10-epoxide is the ultimate genotoxin derived from Benzo[a]pyrene (BaP) involved in N2-

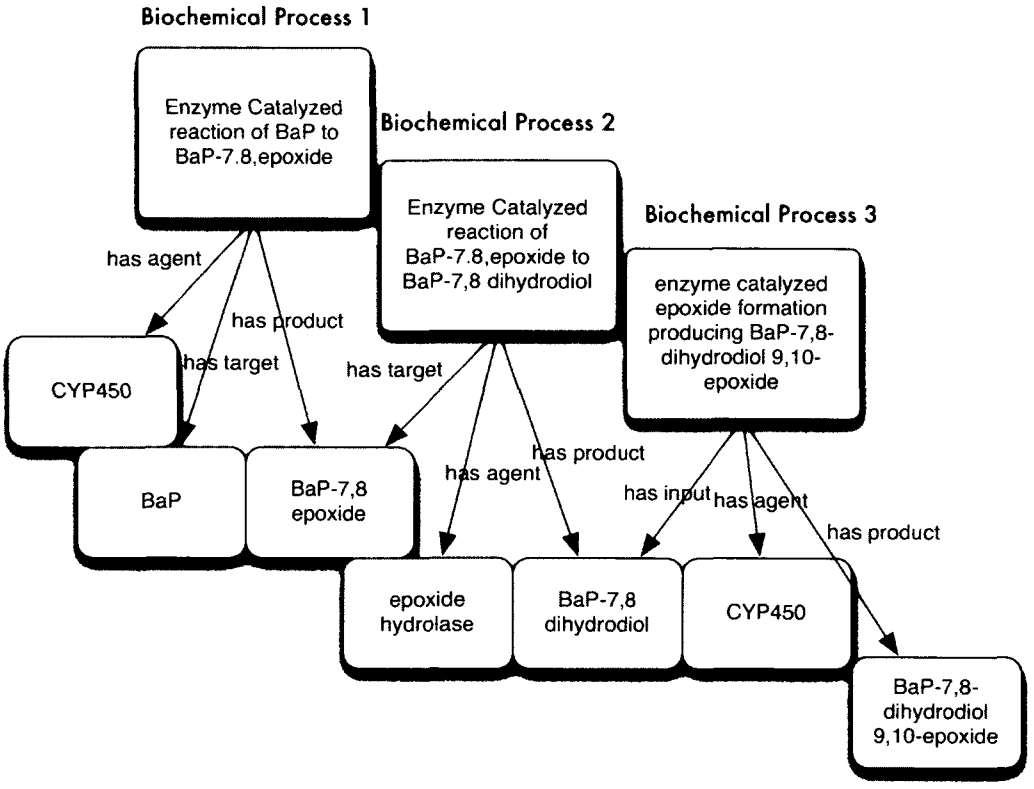

Figure 5.1: Diagram showing relationship between Benzo[a]pyrene and the genotoxin Benzo[a]pyrene–7,8-dihydrodiol-9,10-epoxide as stated based on involvement in metabolic processes.

deoxyguanosine DNA adduct formation [101]. Capturing the metabolism of BaP to Benzo[a]pyrene 7,8-diol-9,10-epoxide, via diol epoxide pathway, in OWL allowed us to infer the connection between BaP and the ultimate genotoxin BaP-7,8-dihydrodiol-9,10-epoxide(Figure 5.1).

Figure 5.1 shows each participant involved in biological processes responsible for converting BaP to BaP-7,8-hydrodiol-9,10-epoxide. The agent is the enzyme responsible for the transformation, the target is the entity undergoing a change, and the product is the result of the change. The targets and products are connected via their involvement in the separate processes. A transitive closure is defined using the relations, "immediately derives into' and 'derives into', to infer all chemicals related through biochemical process(Figure 5.2).

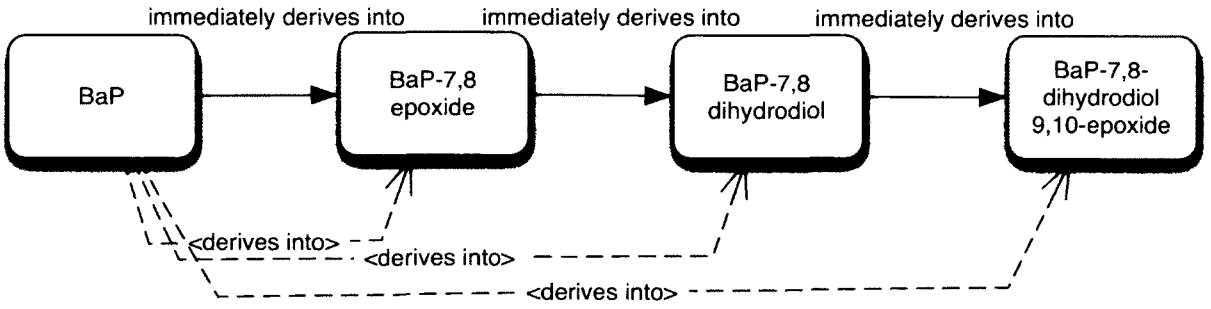

Figure 5.2: Illustration identifying all chemicals which derive into BaP-7,8-dihydrodiol-9,10-epoxide. Explicit statements in MechTox are shown as solid lines. Inferred relations are shown as dashed lines.

Although the potential for inferring novel features was demonstrated the actual practicability was not. The example data-set is biased due to size and selection of chemicals based on known reactive metabolites. Also, no consideration was taken for the complex toxicokinetic and toxicodynamic processes involved. For instance, the fact a chemical may not be genotoxic despite having reactive intermediates. Therefore, the results can only be accepted on the basis of potential given further development of the MechTox ontology. Examining the feasibility would require a larger chemical dataset and metabolic information would need to be encoded for

each chemical present. This would involve knowledge of enzyme kinetics and metabolic fate parameters.

We examined development of a single feature type, chemical derivative, however many semantic relations are possible. These include features based on enrichment of gene and protein function (e.g all differentially expressed genes share a function) and examination of biochemical pathways (e.g. all target proteins belong to lipid metabolism pathways).

5.5 Conclusion

In this work, we demonstrated the use of mechanism of action knowledge encoded in OWL to infer a novel relational feature for model induction. This feature was based on explicit knowledge of mechanism of action encoded in the MechTox ontology. Inferring novel relational features has the potential to improve accuracy of predictive models. The use of mechanism of action provides a context for developing new features for investigating toxicity. However, to fully encompass chemical exposure, ontologies need to be developed to represent the complex ADME (absorption, distribution, metabolism, and excretion) properties of toxicity.

The application of the MechTox ontology demonstrates a novel approach for ontology use in development of predictive models. A future direction lies in extending this approach to include relational learning techniques, such as statistical relational learning (SLR) and inductive logic programming(ILP) [111; 112; 129; 130]. Work has been done on the use of description logic based ontologies as input for inductive logic programming [112]. This represents the first step

towards a learning framework that uses prior information to identify new relations in complex toxicity data.

Conclusion

This body of work represents the application of semantics to improving the integration and use of toxicology data to answer questions of toxicity. I developed a toxicology knowledge base using Semantic Web technologies to demonstrate the integration of multiple, previously separate, data sources. In collaboration with Leo Chepelev, I created a representation of a toxicity decision tree as an OWL ontology in order to facilitate the classification of linked data, and enable answering of questions across ontologies and data resources. It is now possible to develop and run predictive models from the Linked Data network. We created the Mechanistic Toxicology Ontology to formally represent mechanistic toxicology knowledge so that it could be used for answering questions, and could be used in enhancing the feature vectors in development of predictive models of toxicity. We successfully demonstrated the ability of background knowledge, encoded in bio-ontologies, to generate novel descriptors which improved predictive ability. While our framework is promising, more work is required to discover novel patterns in curated datasets of toxicity.

Currently, the ability to leverage ontologies as background knowledge in bioinformatics is not fully realized [112]. An interesting avenue lies in using prior background knowledge to improve statistical relational learning techniques [111; 112; 129; 130]. In this scenario, patterns in the ontology would be used to provide assumptions reducing complexity of a computational problem. Another lies in capturing the probabilities involved to model the toxicodynamics and better reflect the concentration/amount which elicit a phenotype. This would involve exploration of probabilistic reasoning [131] in description logics to refine modeling of mechanism of

action to include statistical confidence measures. Taken together, this work forms a foundation for more explorations into the formalization and use of toxicology knowledge for predictive toxicology.

References

[1] Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, and Kavlock RJ. "The Toxcast Program for Prioritizing Toxicity Testing of Environmental Chemicals." Toxicological Sciences 95.1 (2007): 5.

[2] Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, and Mattingly CJ. "Comparative Toxicogenomics Database: A Knowledgebase and Discovery Tool for Chemical-Gene-Disease Networks." Nucleic Acids Research 37.Database issue (2009): D786.

[3] Fonger GC, Stroup D, Thomas PL, and Wexler P. "TOXNET: A Computerized Collection of Toxicological and Environmental Health Information." Toxicology and Industrial health 16.1 (2000): 4-6.

[4] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, et al. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease." Science 313.5795 (September 29, 2006): 1929-35.

[5] Qu XA, Gudivada RC, Jegga AG, Neumann EK, and Aronow BJ. "Inferring Novel Disease Indications for Known Drugs by Semantically Linking Drug Action and Disease Mechanism Relationships." BMC bioinformatics 10 Suppl 5 (2009): S4.

[6] Waters MD, Jackson M, and Lea I. "Characterizing and Predicting Carcinogenicity and Mode of Action Using Conventional and Toxicogenomics Methods." Mutation Research/Reviews in Mutation Research 705.3 (2010): 184-200.

[7] Hodgson E, Levi PE.A Textbook of Modern Toxicology. N.p.: John Wiley & Sons, Inc., 2004.

[8] Benigni R, and Bossa C. "Predictivity and Reliability of QSAR Models: The Case of Mutagens and Carcinogens." Toxicology Mechanisms and Methods 18.2-3 (2008): 137-147.

[9] Benigni R, and Bossa C. "Structure Alerts for Carcinogenicity, and the Salmonella Assay System: A Novel Insight Through the Chemical Relational Databases Technology." Mutation Research/Reviews in Mutation Research 659.3 (2008): 248 - 261.

[10] Kavlock RJ, Ankley G, Blancato J, Breen M, Conolly R, Dix D, Houck K, Hubal E, Judson R, Rabinowitz J, et al. "Computational Toxicology–A State of the Science Mini Review." Toxicological Sciences 103.1 (2008): 14.

[11] Americal Chemical Society, American Chemical Society. "CAS Database Content at a Glance." [CAS Database Content at a Glance]. List of chemical classes and types in the CAS registry database. Chemical Abstract Service, December 1, 2010. http://www.cas.org/expertise/cascontent/ataglance/index.html. Web. 1 Dec. 2010.

<http://www.cas.org/expertise/cascontent/ataglance/index.html>.December 1, 2010

[12] Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, et al. "Collaborative Development of Predictive Toxicology Applications." Journal of Cheminformatics 2.1 (2010): 1-29. <http://dx.doi.org/10.1186/1758-2946-2-7>.

[13] Allanou R, Hansen BG, and van der Bilt Y. "Public Availability of Data on EU High Production Volume Chemicals, European Commission, Joint Research Centre." Institute for Health and Consumer Protection, Chemical Bureau, I-21020 Ispra (VA), Italy, EUR 18996 (1999).

[14] Andersen ME, and Krewski D. "Toxicity Testing in the 21St Century: Bringing the Vision to Life." Toxicological Sciences 107.2 (2009): 324.

[15] Girschick T, Buchwald F, Hardy B, and Kramer S. "Opentox: A Distributed REST Approach to Predictive Toxicology." Third-Generation Data Mining: Towards Service-Oriented Knowledge Discovery SoKD'10 (2010).

[16] Boyle J. "Gene-Expression Omnibus Integration and Clustering Tools in Seqexpress." Bioinformatics 21.10 (2005): 2550-2551.

[17] Williams-Devane CR, Wolf MA, and Richard AM. "Toward a Public Toxicogenomics Capability for Supporting Predictive Toxicology: Survey of Current Resources and Chemical Indexing of Experiments in GEO and Arrayexpress." Toxicological Sciences 109.2 (2009): 358-371. <http://toxsci.oxfordjournals.org/content/109/2/358.abstract>.

[18] Judson R, Richard A, Dix D, Houck K, Elloumi F, Martin M, Cathey T, Transue TR, Spencer R, and Wolf M. "Actor—Aggregated Computational Toxicology Resource." Toxicology and Applied Pharmacology 233.1 (2008): 7-13.

[19] Yang C, Hasselgren CH, Boyer S, Arvidson K, Aveston S, Dierkes P, Benigni R, Benz RD, Contrera J, Kruhlak NL, et al. "Understanding Genetic Toxicity Through Data Mining: The Process of Building Knowledge by Integrating Multiple Genetic Toxicity Databases." Toxicology mechanisms and methods 18.2-3 (2008): 277-295.

[20] Nigsch F, Macaluso NJM, Mitchell JBO, and Zmuidinavicius D. "Computational Toxicology: An Overview of the Sources of Data and of Modelling Methods." Expert Opin. Drug Metab. Toxicol. 5.1 (2009): 1-14.

[21] Waters MD, Fostel JM, Wetmore BA, and Merrick BA. "Toxicogenomics and Systems Toxicology: Aims and Prospects." Nature Review Genetics 5.12 (2004): 936-48.

[22] BENFENATI E, BENIGNI R, DEMARINI DM, HELMA C, KIRKLAND D, MARTIN TM, MAZZATORTA P, OUDRAOGO-ARRAS G, RICHARD AM, SCHILTER B, et al. "Predictive Models for Carcinogenicity and Mutagenicity: Frameworks, State-Of-The-Art, and Perspectives." Journal of Environmental Science and Health, Part C: Environmental Carcinogenesis and Ecotoxicology Reviews 27.2 (2009): 57-.

[23] Waters M, Stasiewicz S, Alex Merrick B, Tomer K, Bushel P, Paules R, Stegman N, Nehls G, Yost KJ, Johnson CH, et al. "CEBS Chemical Effects in Biological Systems: A Public Data Repository Integrating Study Design and Toxicity Data with Microarray and Proteomics Data." Nucleic acids research 36.Database issue (2008): D892.

[24] Willighagen E, Jeliazkova N, Hardy B, Grafstrom R, and Spjuth O. "Computational Toxicology Using the Opentox Application Programming Interface and Bioclipse." BMC research notes 4.1 (2011): 487.

[25] Witten IH, Frank E, Trigg L, Hall M, Holmes G, and Cunningham SJ. "Weka: Practical Machine Learning Tools and Techniques with Java Implementations." Computing and Mathematical Science Papers 99 (1999): 192-196.

[26] Witten IH, Frank E.Data Mining: Practical Machine Learning Tools and Techniques. N.p.: Morgan Kaufmann, 2005.

[27] Hengstler JG, Foth H, Kahl R, Kramer PJ, Lilienblum W, Schulz T, and Schweinfurth H. "The REACH Concept and Its Impact on Toxicological Sciences." Toxicology 220.2-3 (2006): 232-239.

[28] Kavlock RJ, Austin CP, and Tice RR. "Toxicity Testing in the 21St Century: Implications for Human Health Risk Assessment." Risk Analysis 29.4 (2009): 485-487.

[29] Friedman N, Getoor L, Koller D, and Pfeffer A. "Learning Probabilistic Relational Models." International Joint Conference on Artificial Intelligence 16 (1999): 1300-1309.

[30] Cramer GM, Ford RA, and Hall RL. "Estimation of Toxic Hazard–A Decision Tree Approach." Food and Cosmetics Toxicology 16.3 (1976): 255 - 276.

[31] "Dsstox." United States Environmental Protection Agency. Web <http://www.epa.gov/ncct/dsstox/index.html>.August 10, 2011

[32] "Toxml : Leadscope - Chemoinformatics Platform for Drug Discovery." Leadscope. Web <http://www.leadscope.com/toxml.php>.November 30, 2011

[33] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, and Willighagen E. "The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo-And Bioinformatics." Journal of chemical information and computer sciences 43.2 (2003): 493-500.

[34] Berners-Lee T, and Hendler J. "Scientific Publishing on the Semantic Web." Nature 410 (2001): 1023-1024.

[35] Lassila O, and Swick RR. "Resource Description Framework (RDF) Model and Syntax." World Wide Web Consortium, http://www. w3. org/TR/WD-rdf-syntax

[36] Wang X, Gorlitsky R, and Almeida JS. "From XML to RDF: How Semantic Web Technologies Will Change the Design of 'Omic' Standards." Nat Biotech 23.9 (September, 2005): 1099-1103.

[37] Wilkinson MD, Vandervalk B, and McCarthy L. "SADI Semantic Web Services--Cause You Can't Always GET What You Want!." Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific (2009): 13-18.

[38] Horrocks I, Patel-Schneider PF, and Van Harmelen F. "From SHIQ and RDF to OWL: The Making of a Web Ontology Language." Web semantics: science, services and agents on the World Wide Web 1.1 (2003): 7-26.

[39] Gruber TR. "A Translation Approach to Portable Ontology Specifications." Knowledge acquisition 5.2 (1993): 199-220.

[40] Smith B, and Grenon P. "Basic Formal Ontology (Bfo)." Institute for Formal Ontology and Medical Information Science (IFOMIS) 2007 (2008).

[41] Dumontier M, and Villanueva-Rosales N. "Towards Pharmacogenomics Knowledge Discovery with the Semantic Web." Briefings in Bioinformatics 10.2 (2009): 153.

[42] Villanueva-Rosales N, and Dumontier M. "Yowl: An Ontology-Driven Knowledge Base for Yeast Biologists." Journal of Biomedical Informatics 41.5 (October, 2008): 779-89.

[43] Masolo C, Borgo S, Gangemi A, Guarino N, Oltramari A, and Schneider L. "DOLCE: A Descriptive Ontology for Linguistic and Cognitive Engineering." WonderWeb Project, Deliverable D 17.

[44] Pease A, Niles I, and Li J. "The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and Its Applications." Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web 28 (2002).

[45] Dumontier, Michel. "Semanticscience - Scientific Knowledge Discovery." Web <http://code.google.com/p/semanticscience/>.September 12, 2011

[46] Lauble H, Kennedy MC, Emptage MH, Beinert H, and Stout CD. "The Reaction of Fluorocitrate with Aconitase and the Crystal Structure of the Enzyme-Inhibitor Complex." Proc Natl Acad Sci U S A 93.24 (November 26, 1996): 13699-703.

[47] Belleau F, Nolin M-A, Tourigny N, Rigault P, and Morissette J. "Bio2Rdf: Towards a Mashup to Build Bioinformatics Knowledge Systems." Journal of Biomedical Informatics 41.5 (2008): 706 - 716.

[48] Baader F. The Description Logic Handbook: Theory, Implementation, and Applications. N.p.: Cambridge Univ Pr, 2003.

[49] Brachman RJ, Levesque HJ.Knowledge Representation and Reasoning. N.p.: Morgan Kaufmann Pub, 2004.

[50] "OWL Web Ontology Language Overview." Ed. Deborah McGuinness and Frank Van Harmelen. W3C, 2004. Web <http://www.w3.org/TR/owl-features/>.November 30, 2011

[51] Slater T, Bouton C, and Huang ES. "Beyond Data Integration." Drug Discovery Today 13.13-14 (2008): 584 - 589.

[52] Hardy B. "The Opentox Predictive Toxicology Framework." Toxicology Letters 189 (2009): 260-260.

[53] Xirasagar H, Gustafson S, Merrick BA, Tomer KB, Stasiewicz S, Chan D, Sumner S, Xiao N, and Waters MD. "CEBS Object Model for Systems Biology Data, Sysbio-Om." Bioinformatics 20.13 (2008): 2004-2015.

[54] Fostel J, Choi D, Zwickl C, Morrison N, Rashid A, Hasan A, Bao W, Richard A, Tong W, and Bushel PR. "Chemical Effects in Biological Systems—Data Dictionary (CEBS-DD): A Compendium of Terms for the Capture and Integration of Biological Study Design Description, Conventional Phenotypes, and 'Omics Data." Toxicological Sciences 88.2 (2005): 585.

[55] Xirasagar S, Gustafson SF, Huang CC, Pan Q, Fostel J, Boyer P, Merrick BA, Tomer KB, Chan DD, and Yost KJ. "Chemical Effects in Biological Systems (CEBS) Object Model for Toxicology Data, Systox-Om: Design and Application." Bioinformatics 22.7 (2006): 874.

[56] Williams-DeVane CLR, Wolf MA, and Richard AM. "Dsstox Chemical-Index Files for Exposure-Related Experiments in Arrayexpress and Gene Expression Omnibus: Enabling Toxico-Chemogenomics Data Linkages." Bioinformatics 25.5 (2009): 692.

[57] Waters MD, and Auletta A. "The GENE-TOX Program: Genetic Activity Evaluation." Journal of Chemical Information and Computer Sciences 21.1 (February 1, 1981): 35-38. <http://dx.doi.org/10.1021/ci00029a007>.

[58] Patterson, Aaron, Mike Dalessio, Charles Nutter, Sergio Arbeo, Patrick Mahoney, and Yoko Harada. "Nokogiri." Web <http://nokogiri.org/>.October 11, 2011

[59] Miles A, Bechhofer S, Miles A, and Bechhofer S. "SKOS Simple Knowledge Organization System Reference." W3C Recommendation (2008).

[60] Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, and Laufer J. "Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited." Journal of Chemical Information and Computer Sciences 32.3 (1992): 244-255.

[61] Shearer R, Motik B, and Horrocks I. "Hermit: A Highly-Efficient Owl Reasoner." Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008) (2006).

[62] "SPARQL Query Language for RDF." Ed. Eric Prud'hommeaux and Andy Seaborne. W3C. Web <http://www.w3.org/TR/rdf-sparql-query/>.December 1, 2011

[63] Pérez J, Arenas M, and Gutierrez C. "Semantics and Complexity of SPARQL." The Semantic Web - ISWC 2006 4273 (2006): 30-43. <http://dx.doi.org/10.1007/11926078_3>.

[64] Boelsterli UA. Mechanistic Toxicology : The Molecular Basis of How Chemicals Disrupt Biological Targets. London ; New York: Taylor & Francis, 2003.

[65] Berners-Lee T, Hendler J, Lassila O, and others. "The Semantic Web." Scientific american 284.5 [2001]: 28-37.

[66] Perlich C, and Provost F. "Aggregation-Based Feature Invention and Relational Concept Classes." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining [2003]: 167-176.

[67] Benigni R, Bossa C, Jeliazkova N, Netzeva TI, and Worth AP. "The Benigni/Bossa Rulebase for Mutagenicity and Carcinogenicity-A Module of Toxtree." EUR 23241 [2008].

[68] Verhaar HJM, van Leeuwen CJ, and Hermens JLM. "Classifying Environmental Pollutants." Chemosphere 25.4 [1992]: 471-491.

[69] Quinlan JR. "Induction of Decision Trees." Machine learning 1.1 [1986]: 81-106.

[70] Quinlan JR. "Combining Instance-Based and Model-Based Learning." Proceedings of the Tenth International Conference on Machine Learning. N.p.: Morgan Kaufmann, 1993. 236-243.

[71] Kass GV. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." Applied statistics [1980]: 119-127.

[72] Friedman JH. "Multivariate Adaptive Regression Splines." The annals of statistics 19.1 [1991]: 1-67.

[73] Woo YT, Lai DY."Oncologic: A Mechanism-Based Expert System for Predicting the Carcinogenic Potential of Chemicals." Predictive Toxicology. N.p.: Taylor and Francis, Boca Raton, FL, 2005. 385-413.

[74] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten IH. "The WEKA Data Mining Software: An Update." ACM SIGKDD Explorations Newsletter 11.1 [2009]: 10-18.

[75] "IARC Monographs." International Agency for Research on Cancer. Web <http://www.iarc.fr/en/research-groups/sec2/index.php>.August 10, 2011

[76] Wishart, Knox, and Guo. "HMDB: A Knowledgebase for the Human Metabolome." Nucleic acids research 37 [2009]: D603-610. <http://www.hmdb.ca/>.

[77] Lipinski CA, Lombardo F, Dominy BW, and Feeney PJ. "Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings." Advanced drug delivery reviews 23.1-3 [1997]: 3-25.

[78] Lipinski CA. "Lead-And Drug-Like Compounds: The Rule-Of-Five Revolution." Drug Discovery Today: Technologies 1.4 [2004]: 337-341.

[79] Lipinski CA, Lombardo F, Dominy BW, and Feeney PJ. "Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development." Adv Drug Deliv Rev 23.1 (1997): 3-25.

[80] Patlewicz G, Jeliazkova N, Safford RJ, Worth AP, and Aleksiev B. "An Evaluation of the Implementation of the Cramer Classification Scheme in the Toxtree Software." SAR and QSAR in Environmental Research, 19 5.6 (2008): 495-524.

[81] Wexler P. "TOXNET: An Evolving Web Resource for Toxicology and Environmental Health Information." Toxicology 157.1-2 (2001): 3 - 10. <http://www.sciencedirect.com/science/article/pii/S0300483X00003371>.

[82] Horridge M, and Bechhofer S. "The Owl Api: A Java Api for Working with Owl 2 Ontologies." Proc. OWLED 2009 Workshop on OWL: Experiences and Directions. CEUR Workshop Proceedings 529.

[83] Chepelev, Klassen, and Dumontier. "Hazard Estimation and Method Comparison with

[84] Owl-Encoded Toxicity Decision Trees." Proc. of 8th International Workshop on OWL: Experiences and Directions (OWLED2011) 796 (June 6, 2011). <CEUR-WS.org/Vol-796>.

[85] Benigni R. Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens. Boca Raton, Fla.: CRC Press, 2003.

[86] Carlsson L, Helgee EA, and Boyer S. "Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data." Journal of Chemical Information and Modeling 49.11 (November 23, 2009): 2551-2558.

[87] Perkins R, Fang H, Tong W, and Welsh WJ. "Quantitative Structure-Activity Relationship Methods: Perspectives on Drug Discovery and Toxicology." Environmental Toxicology and Chemistry 22.8 (2003): 1666-1679.

[88] Venkatapathy R, Wang CY, Bruce RM, and Moudgal C. "Development of Quantitative Structure-Activity Relationship (QSAR) Models to Predict the Carcinogenic Potency of Chemicals: I. Alternative Toxicity Measures As An Estimator of Carcinogenic Potency." Toxicology and Applied Pharmacology 234.2 (2009): 209 - 221.

[89] Schultz TW, Yarbrough JW, Hunter RS, and Aptula AO. "Verification of the Structural Alerts for Michael Acceptors." Chemical research in toxicology 20.9 (2007): 1359-1363.

[90] Benigni R, Bossa C, and Worth A. "Structural Analysis and Predictive Value of the Rodent in Vivo Micronucleus Assay Results." Mutagenesis 25.4 (2010): 335-341.

[91] Sanderson DM, and Earnshaw CG. "Computer Prediction of Possible Toxic Action From Chemical Structure; The DEREK System." Human & Experimental Toxicology 10.4 (July 7, 1991): 261-273.

[92] Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, and Wang HH. "The Manchester Owl Syntax." OWL: Experiences and Directions (2006): 10-11.

[93] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, and Chute CG. "Bioportal: Ontologies and Integrated Data Resources at the Click of a Mouse." Nucleic acids research 37.suppl 2 (2009): W170.

[94] Voet D, Voet JG.Biochemistry. 3, illustrated ed. New York: J. Wiley & Sons, 2004.

[95] Aranguren ME, Antezana E, Kuiper M, and Stevens R. "Ontology Design Patterns for Bio-Ontologies: A Case Study on the Cell Cycle Ontology." BMC bioinformatics 9.Suppl 5 (2008): S1.

[96] Egaña M, Rector A, Stevens R, and Antezana E. "Applying Ontology Design Patterns in Bio-Ontologies." Knowledge Engineering: Practice and Patterns (2008): 7-16.

[97] Gangemi A. "Ontology Design Patterns for Semantic Web Content." The Semantic Web–ISWC 2005 (2005): 262-276.

[98] Presutti V, and Gangemi A. "Content Ontology Design Patterns As Practical Building Blocks for Web Ontologies." Conceptual Modeling-ER 2008 (2008): 128-141.

[99] Lauble H, Kennedy MC, Emptage MH, Beinert H, and Stout CD. "The Reaction of Fluorocitrate with Aconitase and the Crystal Structure of the Enzyme-Inhibitor Complex." Proc Natl Acad Sci U S A 93.24 (November 26, 1996): 13699-703.

[100]Xue W, and Warshawsky D. "Metabolic Activation of Polycyclic and Heterocyclic Aromatic Hydrocarbons and DNA Damage: A Review." Toxicology and Applied Pharmacology 206.1 (2005): 73 - 93.

[101]Jernström B, and Gräslund A. "Covalent Binding of Benzo[A]Pyrene 7,8-Dihydrodiol 9,10-Epoxides to DNA: Molecular Structures, Induced Mutations and Biological Consequences." Biophysical Chemistry 49.3 (1994): 185 - 199.

[102]Jernström B, and Gräslund A. "Covalent Binding of Benzo[A]Pyrene 7,8-Dihydrodiol 9,10-Epoxides to DNA: Molecular Structures, Induced Mutations and Biological Consequences." Biophysical Chemistry 49.3 (1994): 185 - 199.

[103]Takeda M, Tojo A, Sekine T, Hosoyamada M, Kanai Y, Endou H, Takeda M, Tojo A, Sekine T, Hosoyamada M, et al. "Role of Organic Anion Transporter 1 (OAT1) in Cephaloridine (Cer)-Induced Nephrotoxicity." Kidney International 56.6 (December 12, 1999): 2128.

[104]Rokushima M, Fujisawa K, Furukawa N, Itoh F, Yanagimoto T, Fukushima R, Araki A, Okada M, Torii M, and Kato I. "Transcriptomic Analysis of Nephrotoxicity Induced by Cephaloridine, a Representative Cephalosporin Antibiotic." Chemical research in toxicology 21.6 (2008): 1186-1196.

[105]Boverhof DR, Burgoon LD, Tashiro C, Chittim B, Harkema JR, Jump DB, and Zacharewski TR. "Temporal and Dose-Dependent Hepatic Gene Expression Patterns in

Mice Provide New Insights Into Tcdd-Mediated Hepatotoxicity." Toxicological Sciences 85.2 (2005): 1048.

[106]Carver LA, and Bradfield CA. "Ligand-Dependent Interaction of the Aryl Hydrocarbon Receptor with a Novel Immunophilin Homolog in Vivo." Journal of Biological Chemistry 272.17 (1997): 11452.

[107]Ikuta T, Eguchi H, Tachibana T, Yoneda Y, and Kawajiri K. "Nuclear Localization and Export Signals of the Human Aryl Hydrocarbon Receptor." Journal of Biological Chemistry 273.5 (1998): 2895.

[108]Probst MR, Reisz-Porszasz S, Agbunag RV, Ong MS, and Hankinson O. "Role of the Aryl Hydrocarbon Receptor Nuclear Translocator Protein in Aryl Hydrocarbon (Dioxin) Receptor Action." Molecular pharmacology 44.3 (1993): 511.

[109]Carver LA, and Bradfield CA. "Ligand-Dependent Interaction of the Aryl Hydrocarbon Receptor with a Novel Immunophilin Homolog in Vivo." Journal of Biological Chemistry 272.17 (1997): 11452.

[110]Pandini A, Soshilov AA, Song Y, Zhao J, Bonati L, and Denison MS. "Detection of the TCDD Binding-Fingerprint Within the Ah Receptor Ligand Binding Domain by Structurally Driven Mutagenesis and Functional Analysis." Biochemistry 48.25 (2009): 5972-5983.

[111]Mansingh G, Osei-Bryson K-M, and Reichgelt H. "Using Ontologies to Facilitate Post-Processing of Association Rules by Domain Experts." Information Science 181.3 (February, 2011): 419-434. <http://dx.doi.org/10.1016/j.ins.2010.09.027>.

[112]Rettinger A, Nickles M, and Tresp V. "Statistical Relational Learning with Formal Ontologies." Machine Learning and Knowledge Discovery in Databases 5782 (2009): 286-301.

[113]Lisi F, and Esposito F. "On Ontologies As Prior Conceptual Knowledge in Inductive Logic Programming." Knowledge Discovery Enhanced with Semantic and Social Information 220 (2009): 3-17. <http://dx.doi.org/10.1007/978-3-642-01891-6_1>.

[114]Getoor L, and Taskar B. "Statistical Relational Learning." MIT Press, Cambridge MA 2 (2007): 6.

[115]Markovitch S, and Rosenstein D. "Feature Generation Using General Constructor Functions." Machine Learning 49.1 (2002): 59-98.

[116]Guo H, Jack LB, and Nandi AK. "Feature Generation Using Genetic Programming with Application to Fault Classification." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 35.1 (2005): 89-99.

[117]Popescul A, and Ungar LH. "16 Feature Generation and Selection in Multi-Relational Statistical Learning." Introduction to statistical relational learning (2007): 453.

[118]Armengol E, and Plaza E. "Lazy Learning for Predictive Toxicology Based on a Chemical Ontology." Artificial intelligence methods and tools for systems biology (2004): 1-18.

[119]Bard JBL, and Rhee SY. "Ontologies in Biology: Design, Applications and Future Challenges." Nature Reviews Genetics 5.3 (2004): 213-222.

[120]Devillers J, and Mombelli E. "Evaluation of the OECD QSAR Application Toolbox and Toxtree for Estimating the Mutagenicity of Chemicals. Part 2. **A-B** Unsaturated Aliphatic Aldehydes." SAR and QSAR in Environmental Research 21.7-8 (2010): 771-783.

[121]"Epa/Oppt/Exposure Assessment Tools and Models/Estimation Program Interface (EPI) Suite Version 3.12 (August 17, 2004)." Web <http://www.epa.gov/opptintr/exposure/pubs/episuite.htm>.December 13, 2011

[122]Walsh JS, and Miwa GT. "Bioactivation of Drugs: Risk and Drug Design." Annual review of pharmacology and toxicology 51 (2011): 145-167.

[123]Teffera Y, Choquette D, Liu J, Colletti AE, Hollis LS, Lin MHJ, and Zhao Z. "Bioactivation of Isothiazoles: Minimizing the Risk of Potential Toxicity in Drug Discovery." Chemical research in toxicology (2010).

[124]Reese M, Sakatis M, Ambroso J, Harrell A, Yang E, Chen L, Taylor M, Baines I, Zhu L, and Ayrton A. "An Integrated Reactive Metabolite Evaluation Approach to Assess and Reduce Safety Risk During Drug Discovery and Development." Chemico-Biological Interactions 192.1 (2011): 60-64.

[125]Park BK, Laverty H, Srivastava A, Antoine DJ, Naisbitt D, and Williams DP. "Drug Bioactivation and Protein Adduct Formation in the Pathogenesis of Drug-Induced Toxicity." Chemico-Biological Interactions 192.1-2 (2010): 30-36.

[126]Guengerich F. "Cytochrome P450S and Other Enzymes in Drug Metabolism and Toxicity." The AAPS Journal 8.1 (2006): E101-E111. <http://dx.doi.org/10.1208/aapsj080112>.

[127]Bugrim A, Nikolskaya T, and Nikolsky Y. "Early Prediction of Drug Metabolism and Toxicity: Systems Biology Approach and Modeling." Drug discovery today 9.3 (2004): 127-135.

[128]Sun H, and Scott DO. "Structure-Based Drug Metabolism Predictions for Drug Design." Chemical Biology & Drug Design 75.1 (2010): 3-17. <http://dx.doi.org/10.1111/j.1747-0285.2009.00899.x>.

[129]Dobo KL, Obach RS, Luffer-Atlas D, and Bercu JP. "A Strategy for the Risk Assessment of Human Genotoxic Metabolites." Chem Res Toxicol 22.2 (January 27, 2009): 348-356. <http://dx.doi.org/10.1021/tx8004339>.

[130]Fanizzi N, d'Amato C, and Esposito F. "Statistical Learning for Inductive Query Answering on OWL Ontologies." The Semantic Web - ISWC 2008 5318 (2008): 195-212. <http://dx.doi.org/10.1007/978-3-540-88564-1_13>.

[131]Lehmann J. "DI-Learner: Learning Concepts in Description Logics." Journal of Machine Learning Research 10 (December, 2009): 2639-2642. <http://dl.acm.org/citation.cfm?id=1577069.1755874>.

(132)Klinov P. "Pronto: A Non-Monotonic Probabilistic Description Logic Reasoner." The Semantic Web: Research and Applications (2008): 822-826.

(133)