A Semantically Enabled Framework for Small Molecule
Metabolic Fate Prediction: the Web as a Biochemical
Reactor

by

Leonid Chepelev

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

Doctor of Philosophy

in

Biology

Carleton University
Ottawa, Ontario

© 2011, Leonid Chepelev

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: 978-0-494-83248-6
Our file Notre référence
ISBN: 978-0-494-83248-6

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Prediction and dynamical characterization of metabolic fate of small molecules is essential to understanding chemical toxicity, activity, and mechanisms of bioaccumulation, degradation, and industrial “green” synthesis. Unfortunately, experimental methods for metabolic fate characterization are prone to missing transient metabolites, experience difficulty in adjusting to variations in metabolic networks, and are expensive in the face of the enormous chemical space that requires characterization. The quantitative structure-activity relationship-based methods employed currently to predict metabolic fate are insufficient since they often ignore the dynamics of metabolism or are limited in scope and applicability. First principle-based approaches currently require expensive manual coordination of tools and formalisms to develop a cohesive dynamical model. An integrative framework to support these efforts is vital for metabolomics research to progress efficiently while taking full advantage of the expanding biochemical information and improving computational resources.

In this work, I investigate the feasibility of constructing such a framework to integrate the computational and informational resources required for automated metabolic fate prediction of small molecules. To this end, I demonstrate a framework, based on semantic web technologies, which attempts to connect the highly fragmented world of chemical information at the levels of data representation and computational resource interoperability. This modular prototype framework is reconfigurable and adjustable for future advancements in computational chemistry and improvements in computational power. Within this framework, enzymes are distributed as semantically annotated web

services, each with logically defined molecular input and output classes. A machine client that is capable of automatically classifying small molecules strings together biochemical pathways by calling upon the appropriate distributed enzyme services and obtaining the kinetic parameters and reaction specifications for each reaction. Kinetic reaction analysis is then carried out on the resultant pathway and the toxicity of each chemical entity is assessed using semantically encoded toxicity decision trees.

Although the framework developed here does not aim to address all aspects of metabolism prediction, or to cover any appreciable portion of all human enzymes, it demonstrates the feasibility of transforming the Web into a unified, reconfigurable, distributed biochemical reactor and can open new possibilities for interdisciplinary research.

Acknowledgements

First and foremost, I would like to thank Dr. Michel Dumontier for creating a supportive, opportunity-rich environment for my research and for allowing me considerable freedom in carrying out this investigation. Truly, I shall always kindly remember the years spent on my doctoral thesis as a member of Dr. Dumontier's laboratory and working in close collaboration with Michel. I would like to thank Dr. Alain St-Amant and Dr. Maria De Rosa, the committee members who have assessed my progress and helped with constructive feedback. I would also like to thank my family for unwavering support and understanding. I would like to thank the faculty of Carleton University for providing considerable support, and the various donors of the scholarships and awards I have received throughout my studies. I am not certain whether this work would have been possible without this help. I would also like to thank the National Sciences and Engineering Research Council for their extensive funding through the Canada Graduate Scholarships program and the Michael Smith foreign studies award that has allowed me to work with some of the leaders in my field at the European Bioinformatics Institute in Cambridge, United Kingdom. I am also thankful to my lab mates for all the fun times and conversations we have had together. In particular, I would like to thank Dana Klassen for his collaboration on decision trees. Finally, I would like to thank the group of Dr. Mark Wilkinson at the University of British Columbia and the group of Dr. Christopher Baker at the University of New Brunswick for their work on some of the Semantic Web technologies which form the foundation of some parts of this work.

Preface

The vast majority of this work was carried out by me, and this thesis was written solely by me with critical input from Dr. Michel Dumontier. I would like to thank Dr. Michel Dumontier for the supervision of my work as well as valuable and insightful comments. The majority of this thesis is a result of my original contributions (which I assess at about 90%) and Dr. Dumontier's contributions in the form of suggestions and critical assessments (contributing to 10% of all work). Parts of this thesis, however, are a reflection of collaborative studies. Thus, I would like to thank Dana Klassen for his participation in the semantic decision tree project and for his work on the OWL encoding on decision trees, all reported in Section 3.3.6. For this session, I approximate that Dana Klassen has carried out 30% of the work, with me carrying out 60% and Dr. Dumontier the remaining 10%. I would also like to thank Dr. Christopher Baker's group for their collaboration in demonstrating the formalization of chemical annotation and classification with SADI services, discussed in Section 4.3.7. Finally, I would like to thank the curators at the European Biochemical Institute, and in particular, Marcus Ennis, for some of their assessment of the chemical classification outcomes for the structural hierarchy construction reported in Chapter 3.

Table of Contents

Abstract.....	ii
Acknowledgements	iv
Preface.....	v
List of Tables	xi
List of Figures.....	xii
List of Appendices.....	xv
List of Abbreviations	xvi
1 Chapter: Introduction.....	1
1.1 General Introduction.....	2
1.2 Xenobiotic Metabolism	7
1.2.1 Phase I Xenobiotic Metabolism	8
1.2.2 Phase II of Xenobiotic Metabolism.....	12
1.2.3 Non-Classical Transformations.....	14
1.3 Representation and Cataloguing of Metabolic Information.....	15
1.3.1 Pathway Map Databases.....	19
1.3.2 Enzyme Kinetics Databases	21
1.4 Metabolic Fate Prediction.....	25
1.4.1 Graph Theory and Expert-Based Approaches.....	27
1.4.2 Statistical Learning and Quantitative Structure-Activity Relationships.....	29
1.4.3 First Principles-Based Approaches	31
1.4.4 Commercial Packages for Metabolic Fate Prediction	35
1.5 Semantic Web Technologies	37
1.6 Semantic Web for Chemical Knowledge Representation.....	43
1.7 An Integrated Metabolic Fate Prediction Framework	50

2	Chapter: Chemical Entity Semantic Specification	53
2.1	Abstract.....	54
2.2	Introduction	56
2.2.1	Consensus Chemical Entity Identifiers	56
2.2.2	Common Chemical Information Representation Platform.....	58
2.2.3	Chemical Knowledge Integration.....	61
2.2.4	Overview	63
2.3	Results and Discussion	64
2.3.1	CHESS Representation Overview	64
2.3.2	Molecular Specification	66
2.3.3	Semantic Web-Enabled Cheminformatics: Chemical Searching	70
2.3.4	Chemical Descriptor Specification.....	77
2.3.5	Chemical Information Integration.....	82
2.3.6	Variable-Level Granularity Semantically Enriched Annotations and Queries....	84
2.3.7	Representing and Querying Reactive Transformations.....	89
2.3.8	Supporting Tools and Interfaces.....	94
2.4	Conclusions	94
2.5	Methods	96
2.5.1	Supporting Ontologies.....	96
2.5.2	Triplification of Chemical Information.....	97
2.5.3	Query Testing.....	97
2.5.4	Publically-Accessible Web Interface	97
2.6	Authors' Contributions	98
3	Chapter: Chemical Entity Semantic Classification	99
3.1	Abstract.....	100
3.2	Introduction	102

3.2.1	Structural Classification	102
3.2.2	Functional Classification.....	107
3.3	Results	111
3.3.1	Structure-Based Classifications.....	111
3.3.2	Automated Derivation of Formal Chemical Class Definitions	112
3.3.3	Self-Organizing Chemical Ontology.....	119
3.3.4	Chemical Classification Accuracy	125
3.3.5	Seamless Integration of Novel Chemical Classes: Enzyme Substrates.....	126
3.3.6	Functional Chemical Classification.....	126
3.4	Discussion.....	133
3.5	Methods	144
3.5.1	Training and Testing Data Sets	144
3.5.2	Identification of Consensus Molecular Fragments and Patterns	145
3.5.3	SMILES and SMARTS Pattern Canonicalization.....	147
3.5.4	Chemical Class Definitions Encoding in OWL.....	148
3.5.5	Chemical Classification Accuracy Evaluation	149
3.5.6	Data Sources and Data Preparation for Functional Decision Trees	150
3.5.7	Decision Tree Construction and Validation	150
3.5.8	Representation of Decision Trees as OWL Ontologies.....	151
3.5.9	Chemical Classification.....	152
3.6	Authors' Contributions	152
4	Chapter: Semantic Web Services for Chemistry.....	153
4.1	Abstract.....	154
4.2	Introduction	156
4.3	Results and Discussion	164
4.3.1	Exposing CDK QSAR Functionality with SADI	164

4.3.2	SADI-Enabled Format Interconversion and Software Interfacing	170
4.3.3	Lipinski Rule of Five the Semantic Way	171
4.3.4	Mechanisms for Parameter and Computational Experiment Provenance Specification.....	174
4.3.5	Integration and Repurposing of Chemical Resources	177
4.3.6	Exposing Chemical Database Resources as SADI Services	179
4.3.7	Practical Applications of SADI Services in Lipidomics	180
4.4	Conclusions	182
4.5	Methods	184
4.5.1	Supporting Service Ontologies.....	184
4.5.2	Service Creation with CDK and OpenBabel	185
4.5.3	SHARE and SADI Service Distribution.....	185
4.6	Authors' contributions.....	186
4.7	Acknowledgements.....	186
5	Chapter: A Distributed Semantic Metabolic Modeling Framework	187
5.1	Abstract.....	188
5.2	Introduction	190
5.3	Methods	194
5.3.1	Assumptions	194
5.3.2	Modeling Enzymes and Reactions as Web Services	199
5.3.3	Computing Transformation Thermodynamics	201
5.3.4	Approximating Kinetics	203
5.3.5	Toxicity Evaluation	204
5.3.6	Dynamic Model Construction	205
5.4	Results	205
5.4.1	Integrated Metabolic Fate Prediction Procedure Overview	205

5.4.2	Biochemical Transformations and Products of Benzene.....	207
5.4.3	Kinetic Model of Metabolism and Toxicological Assessment.....	210
5.5	Discussion and Conclusions	213
5.5.1	Framework Assessment.....	213
5.5.2	Predictions of Dynamic Metabolism.....	215
5.5.3	Conclusion.....	217
5.6	Authors' Contributions	218
6	Chapter: Thesis Summary and Future Directions	219
6.1	Overview	220
6.2	Alternative Architectures for MFP on the Semantic Web.....	221
6.3	Accounting for Variations in Metabolism	222
6.4	Semantic Enrichment of Pathways	223
6.5	Distribution of Physical Resources.....	224
6.6	Conclusion.....	225
Appendices.....		226
References.....		229

List of Tables

Table 1-1. Common modifications of chemical entities as part of Phase II xenobiotic metabolism.	12
Table 5-1. Selected transformations identified by the reactive web services, along with the approximated constants. Only major substrates and products shown.	206

List of Figures

Figure	Title	Page
1-1	Xenobiotic metabolism-assisted toxic activation of equilenin, an estrogen analogue.	4
1-2	An overview of absorption, distribution, metabolism, and excretion of small molecules.	8
1-3	Some of the known biochemical transformations and toxic activation of benzene as a consequence of Phase I xenobiotic metabolism.	10
1-4	An example of downstream effects of activating reactions of Phase I xenobiotic metabolism: oxidative dealkylation of amines.	12
1-5	Phase II xenobiotic metabolism does not always end with conjugation, but sometimes involves further covalent modifications, as evidenced by glutathione adducts of benzene products.	14
1-6	The process of predicting a biochemical transformation in UM-BBD PPS.	28
1-7	A reaction energy profile for a simple enzymatic reaction with Gibb's free energy of formation on the vertical.	32
1-8	Formal meaning and XML schemas.	38
1-9	RDF n-triple representation of the statement "Leo likes chemistry".	40
1-10	Architecture of the Semantic Web.	41
1-11	An ontology of plants and their properties.	42
1-12	Bio2RDF databases (outlined), within the world of open linked data.	47
1-13	Barriers preventing the streamlined knowledge integration.	50
2-1	A simplified specification of a water molecule in CML.	59
2-2	A simplified overview of the general features of the CHESS specification.	65
2-3	Principles for generating canonical identifiers for some of the many chemical entity types.	66
2-4	SPARQL query completion and average molecular size in DB100.	73
2-5	General performance trends for the two query modes relative to Open Babel matching.	74
2-6	A simplified descriptor specification, as per the CHEMINF ontology approach.	78
2-7	Mapping components of an instantiated molecule to reactive transformation definition participants can enable a range of inferences relevant to predicting reaction outcome.	89

2-8	A modification of molecular fingerprinting used for the complete description of chemical structures in terms of an exhaustive list of functional groups.	91
3-1	A simple toxicity decision tree.	107
3-2	An abstract view of the curation effort for structure-based chemical hierarchies.	110
3-3	Some of the consensus structural features identified for the class of peptides using the proposed methodology.	112
3-4	Establishing a chemical class hierarchy through direct observation of consensus structural fingerprints.	114
3-5	An example of a ‘flat’ chemical ontology with chemical class definitions, within the Protégé 4.	118
3-6	MeSH chemical class hierarchy automatically inferred using the Pellet reasoner.	119
3-7	ChEBI chemical class hierarchy automatically inferred using the Pellet reasoner.	120
3-8	The chemical class hierarchy inferred for an ontology that has resulted from the automated integration of the MeSH and ChEBI ontologies.	122
3-9	A decision tree generated from a computationally derived dataset of drug-like compounds.	125
3-10	The structure of an automatically generated OWL representation of a Rule of Five tree.	126
3-11	A simplified carcinogenic toxicity decision tree generated from a ToxCast dataset.	127
3-12	A fragment of the final, classification-friendly decision tree.	130
3-13	Relevant features of acetaminophen used in classification.	131
3-14	A novel approach to chemical hierarchy construction.	140
4-1	Conceptual overview of SADI functionality.	160
4-2	Seamless service integration into the Semantic Web with SPARQL queries over RDF-encoded resources in chemistry, as enabled by SADI.	162
4-3	Classes and methods in APIs as input/output classes and functionality-encapsulating services.	164
4-4	An overview of the prototype lipid classification and annotation framework.	179
5-1	An overview of the integrated prototype semantic MFP framework.	191
5-2	Activation free energy decomposition.	194
5-3	A sample of the correlation between the enzymatic activation free energy ΔG^* and experimentally determined reaction free energy ΔG° for the enzyme alcohol dehydrogenase.	196

5-4	An overview of the predicted metabolic fate of benzene.	207
5-5	Dynamic concentration profiles of benzene metabolites.	209
5-6	Toxicological profile of benzene metabolism, including and excluding the contribution of toxicity stemming from benzene.	210

List of Appendices

Appendix A.....	224
-----------------	-----

List of Abbreviations

API – Application Programming Interface

BioPAX – Biological Pathway Exchange Format

BRENDA – Braunschweig Enzyme Database

CDK – Chemistry Development Kit

ChEBI – Chemical Entities of Biological Interest Database and Ontology

CHEMINF – Chemical Information Ontology

CHESS – Chemical Entity Semantic Specification

CML – Chemistry Markup Language

CYP – Cytochrome P450

CYP2E1 – Cytochrome P450 2E1

DNA – Deoxyribonucleic Acid

GSH – Glutathione

HCLSIG – Health Care and Life Sciences Special Interest Group

InChI – International Chemical Identifier

IUPAC – International Union of Pure and Applied Chemistry

LODD – Linked Open Drug Data

LSO – Lipinski Service Ontology

MeSH – Medical Subject Headings

MFP – Metabolic Fate Prediction

MIRIAM – Minimum Information Required In the Annotation of Models

OWL – Web Ontology Language

QM/MM – Quantum Mechanics/Molecular Mechanics Hybrid Method

QSAR – Quantitative Structure-Activity Relationship

RDF – Resource Description Framework

REST – Representational State Transfer

SABIO-RK – System for the Analysis of Biochemical Pathways - Reaction Kinetics

SADI – Share, Analyze, Discover and Integrate Web Service Framework

SAWSDL – Semantically Annotated Web Service Description Language

SBML – Systems Biology Markup Language

SMILES – Simplified Molecular Input Line Entry Specification

SOAP – Simple Object Access Protocol

SWT – Semantic Web Technologies

W3C – World Wide Web Consortium

WSDL – Web Service Description Language

WSMO – Web Service Modeling Ontology

WSMX – Web Service Execution Environment

XML – Extensible Markup Language

1 Chapter: Introduction

1.1 General Introduction

Life is a tremendously complex phenomenon that involves the coordinated activities of innumerable chemical entities that undergo energy and matter exchanges to form a vast array of spatiotemporal chemical patterns that we recognize in our daily lives as ourselves and the living world around us. Small molecules (chemical entities with a low molecular weight) are involved in the majority of the biochemical transformations that comprise metabolism of living organisms. While some small molecules are a part of the regular dietary intake of living organisms, necessary since they bear chemical energy or structural components that can be transformed to power and develop life, other chemical entities may be unnecessary or harmful if they persist within a living organism for any appreciable period of time. In humans, such harmful molecules are eliminated through the process of xenobiotic metabolism, whereby xenochemicals, or compounds foreign to the 'normal' metabolic pathways of a human body that were either introduced as medicine, a part of food, or have inadvertently entered the body through interactions with the environment, are broken down to their most basic constituent components or are converted into forms that are readily eliminated from the organism through excretion [1-3].

This process is desirable in principle, because many xenochemicals are electrophilic and therefore reactive with the abundant cellular nucleophiles, potentially introducing a plethora of unwanted and persistent covalent modifications to the biomacromolecules of the living body [4]. Some of these modifications, like the formation of aromatic DNA adducts, may result in long-term effects on the rest of the organism, including DNA mutations and the induction of tumor growth, or cancer [5,6]. Other xenochemicals, such

as those developed by the pharmaceutical industry, are often structurally optimized for a decreased toxicity while simultaneously retaining the desirable biological activity, absorption, distribution, and elimination (ADMET) parameters [7,8]. However, the large and incompletely characterized array of chemical reactions catalyzed by human enzymatic systems involved in general metabolism, detoxification, and elimination often provides surprises for the pharmaceutical industry in the form of unexpected toxic activations of pharmaceutical agents [9,10]. This unexpected activation may, in mild cases, lead to unwanted side-effects of varying severity levels, or even to deaths and subsequent drug recalls in the worst cases [11-13].

Consider, for example, the toxic activation of extraneous estrogens and estrogen analogues, frequently prescribed in estrogen replacement therapies (Figure 1-1) [14,15]. Outwardly, it would seem that these compounds are extremely similar, bearing the same chirality and structural properties to the endogenous estrogen. However, excessive exposure to these exogenous estrogens results in their toxic activation by enzymes in the Cytochrome P450 family to the electrophilic and redox-active quinone forms through hydroxylation of the aromatic ring [15]. These quinones may then form site-specific DNA adducts, or participate in redox cycling, during which reactive oxygen species are generated and cellular energy carrying molecules are depleted. In addition to this, the catechols that result from cytochrome P450-mediated estrogen activation have been demonstrated to induce tumorigenic growth in breast epithelial cells [16]. The adverse effects of estrogen replacement therapies have led to critical re-evaluation of the risks and benefits of the practice, especially around the turn of the millennium [17,18].

Figure 1-1. Xenobiotic metabolism-assisted toxic activation of equilenin, an estrogen analogue.

The most disturbing aspect of the story of estrogen replacement therapy side effects, especially with respect to cancer induction and reactive oxygen species formation, is the fact that the redox cycling aspect of quinones was reported and well-studied by the middle of the 20th century [19]. The oversight in the identification of the connection between the adverse effects of quinones and the potential adverse effects of quinone derivatives of estrogen mimic metabolites accumulated as a result of estrogen replacement therapy at the outset of the development of these treatments and recommendation of such therapies seems somewhat unreasonable. One is left wondering, therefore, whether this case is merely an example of a recurrent pattern that is characteristic of the development cycles of drugs that have been recalled from the market. In basic terms, a biological transformation or a toxic activation of a compound may be ignored, either due to difficulties in completing an adequate theoretical analysis, or due to

the fact that the experimental systems, such as liver microsomes, employed to test the toxicity and activity of drug candidates [20,21], do not suffer from the accumulation of a normally transient but toxic metabolite that may be observed in specific sub-populations of humans endowed with an unfortunate combination of genetic variants of the key detoxification enzymes [22,23]. Various dietary and pharmacologically active supplements may also result in the adverse alteration of the metabolic fate of certain drugs, as evidenced by the inhibition of cytochrome P450 3A4 by grapefruit juice [24]. Finally, factors such as age and sex may also affect the relative distributions of the various forms of xenobiotic metabolites [25]. As a result, promising drug candidates may in mild cases have counter-indications attached to them, and in serious cases where their toxic metabolites accumulate to dangerous levels in certain circumstances and individuals, be recalled either at a late stage in drug development and testing, or after hitting the market, costing the developing companies hundreds of millions of wasted research dollars and years of regulatory approvals, not to mention the adversely affected human lives.

The underlying causes of these failures to identify potentially toxic compounds and compound combinations are numerous, but (in my opinion) may be distilled to the inadequacies in the connectivity of chemical knowledge, including knowledge stored in major biochemically relevant databases and the knowledge provided by the largely fragmented world of computational chemistry. In other words, the various knowledge barriers in the life sciences make it difficult to be aware of all chemical knowledge, let alone to employ it automatically and effectively to formulate useful and accurate predictions of metabolic products. Whether these barriers in biochemical knowledge

emerged due to disparities in chemical databases and information reporting standards across disciplines, or simply as a byproduct of historical development of the tools and formalisms supporting certain branches of biomedical chemistry, the outcome is unchanged: medicinal chemistry researchers are currently forced to employ a plethora of disconnected databases and tools in order to integrate and compute as much knowledge as possible so as to attempt the characterization and adequate prediction of the metabolic fate and resultant toxicity of candidate small molecule drugs (e.g. [26]).

One is left wishing, therefore, that a platform-, discipline-, and industry-independent framework, based on common and universal chemical principles could be developed to support such truly integrative biochemical studies. The realization of the necessity of such a platform is already coming to the world of pharmaceutical chemistry, with the first joint bodies on standards and formalisms for knowledge and resource sharing using integrative Semantic Web technologies already set up [27]. Ideally, information from journals, computations, databases, experiments, and models should be integrated on the fly to bring researchers in biochemistry an accurate response to their queries that considers all available knowledge, regardless of the originating domain or institution. Furthermore, we fully anticipate that not all knowledge necessary to address every query shall be stored in a database, while at the same time recognizing that much knowledge in biochemistry may be inferred, computed, or at the very least, approximated. Therefore, computational resources should ideally be as exposed and amenable to facile and integrative querying as is database-contained knowledge.

This thesis is concerned with the production of a prototype framework that attempts to integrate disparate computational and informational resources in biological chemistry in

order to address the problem of predicting metabolic fates of arbitrary small molecules. We postulate that the construction of such a framework is possible with the use of the integrative power of the Semantic Web technologies and formalisms to create a “self-aware” network of computational and informational resources. In other words, machine and human agents on this network would have a precise and formal definition of not only what resources are available, but also their meaning and ways in which they can be combined to address a particular logical query. We propose to distribute a network of independent services that mimic the functionalities of biochemical enzymes, thus transforming the Web into a distributed biochemical reactor, which can be readjusted for the desired accuracy/computational cost tradeoff, conditions and structural mutations in a given enzyme, or even organism under investigation. However, before we can start describing the construction and principles employed in this proposed prototype framework, we need to understand in greater detail the basics of xenobiotic metabolism, historical perspectives on experimental and theoretical metabolic fate elucidation, some basic theory on computational reaction prediction, as well as what exactly the Semantic Web is. What follows is a very brief and general overview of these complex topics.

1.2 Xenobiotic Metabolism

The process of xenobiotic metabolism is rather complex, relies on the concerted activities of multiple promiscuous enzymes coupled with non-enzymatic reactions, and involves several phases, as well as a plethora of complicating factors and interactions that make the process very difficult to predict [1-3]. The process is further complicated by the pharmacokinetics of the xenochemical under investigation, since the point of entry of a given small molecule into an organism and the relative distribution of its various forms

may impact the distribution of its metabolites (Figure 1-2), an aspect we shall declare outside the scope of this work.

Figure 1-2. An overview of absorption, distribution, metabolism, and excretion of small molecules.

1.2.1 Phase I Xenobiotic Metabolism

In the first phase of xenobiotic metabolism (Phase I), xenochemicals are modified such as to activate them for further transformations and ultimate elimination from the body. For example, benzene is a fairly hydrophobic compound that bears no overall charge, and needs to be derivatized in order to become more hydrophilic, and thus more easily eliminated through urinary excretion. Another purpose of this activation is to introduce sites that can then serve as reactive centers for the reactions of Phase II of xenobiotic metabolism that would result in benzene being derivatized with moieties that can be

specifically recognized by the transport enzymes responsible for elimination of xenochemicals into urine or feces [28].

The majority of Phase I metabolism is carried out by cytochrome P450 enzymes, a family of promiscuous catalysts so named for the absorption maximum of their heme ring cofactor when it is bound to carbon monoxide [29]. These enzymes form what is known as a cytochrome P450 system, comprised of a cytochrome P450 NADPH-dependent reductase and a cytochrome P450, among other components, the former component required to reduce the partner cytochrome P450 to enable each of its cycles. Although cytochromes P450 can catalyze a number of transformations, the most frequently considered transformation within the context of xenobiotic metabolism is hydroxylation [29]. For example, benzene can be a subject of multiple hydroxylation transformations, first in a conversion to phenol, which could also undergo hydroxylation to form a number of aromatic diols, some of which may also undergo yet another hydroxylation [28]. Although the end goal of aromatic hydroxylation in this case is the eventual elimination of benzene from the host, the interim metabolites may be equally or even more reactive than their parent compound. For instance, catechol and hydroquinone that result from phenol hydroxylation are known redox-cycling agents that exert much of their toxicity through the generation of reactive oxygen species in oxidation to their quinone forms [30], while using up cellular energy through various reduction mechanisms, such as the two-electron reduction catalyzed by the NAD(P)H Quinone Oxidoreductase 1 (NQO1) enzyme [31], for example, in a manner similar to that described for equilenin (Figure 1-3) [15]. As already discussed, quinones may also covalently bind to DNA, which may result in long-term genetic mutations in the affected cells. Another example where phase I

metabolism results in a more toxic compound is the nitrogen oxygenation of aniline compounds that results in the activation of such compounds for nucleophilic attack from DNA on the aromatic ring, permitting the formation of DNA adducts and potential long-term effects such as development of cancers [32].

Figure 1-3. Some of the known biochemical transformations and toxic activation of benzene as a consequence of Phase I xenobiotic metabolism.

Other small molecule modifications besides hydroxylation of aromatic atoms that activate xenochemicals for further metabolism and excretion include oxidative dealkylation, epoxidation, various nitrogen oxygenation reactions (by cytochromes P450 or flavin-containing monooxygenase), sulfur oxygenation reactions, reduction (e.g. of ketones,

nitro and azo compounds), and various hydrolytic reactions, among many others. Generally, as a result of these transformations, the modified compounds become more hydrophilic and often less electrophilic, which lowers their reactivity in the nucleophile-rich cellular milieu [3]. Although enzymes carrying out these reactions are fairly promiscuous, they adhere to patterns in terms of the allowed reaction transformations, and the feasibility of the transformations is dictated by steric, electronic, and thermodynamic considerations, as we shall discuss later. Therefore, the hydroxylation examples presented earlier for benzene and equilenin are not only characteristic of these compounds, but generally, of a class of molecules that contain electron-deficient organic carbon atoms. Every reactive transformation catalyzed by an enzyme can therefore be generally recorded in terms of transformations of members of one broad class of small molecules into another chemical class, with the promiscuity of a given enzyme manifested in the broadness of the definition of its substrate class. This is an important observation for the architecture of our metabolic fate prediction framework and the nature of the enzyme-mimicking web services, as we shall discuss in subsequent chapters.

Another important consideration for understanding phase I metabolism is the fact that the products of initial activation reactions do not persist in their native state, but may rather undergo further transformations that could further impact their structure and reactivity. For instance, some members of the cytochrome P450 family hydroxylate amines to form heminaminals which then undergo dissociation, completing the oxidative dealkylation process for a given amine (Figure 1-4) [33]. This is evident in metabolism of such compounds as Lysergic Acid Diethylamide (LSD) to its N-demethyl LSD form, for example [34].

Figure 1-4. An example of downstream effects of activating reactions of Phase I xenobiotic metabolism: oxidative dealkylation of amines.

1.2.2 Phase II of Xenobiotic Metabolism

In the second phase of xenobiotic metabolism, chemical entities undergo conjugation reactions with moieties that would mark a given derivatized xenochemical for immediate and active transport from the cell and into sites of excretion (primarily through kidneys). These modifications involve covalent bond formation and, with the exception of methylation (except for methylation of tertiary or aromatic nitrogen atoms), result in an overall more hydrophilic compound, further boosting its propensity to be eliminated in urine and lowering its capacity for distribution across lipid bilayers and accumulation among lipids within an organism. The enzymes that carry out these modifications transfer a part of another compound onto the eliminated molecule and are therefore termed transferases. Some of the more frequently occurring modifications, enzyme substrate classes, and the relevant transferase enzyme cofactors have been listed (Table 1) [2].

Table 1-1. Common modifications of chemical entities as part of Phase II xenobiotic metabolism.

<i>Process</i>	<i>Substrate Classes</i>	<i>Coenzyme</i>	<i>Transferase</i>
Methylation	Catechols, Alcohols, Primary Amines, Thiols	S-Adenosyl Methionine	Methyl-Transferase
Acetylation	Alcohols, Primary Amines	Acetyl Coenzyme A	Acetyl-Transferase
Sulfation	Alcohols, Primary Amines	3'-Phosphoadenosine-5'-Phosphosulfate	Sulfotransferase

Glucuronidation	Alcohols, Carboxylic Acids, Amines, Thiols, Alkanes	UDP-Glucuronic Acid	UDP-Glucuronosyl- transferase
Glycination	Carboxylic Acids	Coenzyme A, Glycine	Glycine N- Acyltransferase
Glutamination	Carboxylic Acids	Coenzyme A, Glutamine	Glutamine N- Acyltransferase
Glutathionylation	Aromatics, Epoxides	Glutathione	Glutathione S- Transferase

These transferase-catalyzed reactions may be followed by further modifications before the severely tamed xenochemical is excreted into the urine. For example, the glutathione conjugate of the various products of benzene hydroxylation may undergo subsequent modifications to lose the glycine and glutamine moieties of glutathione, with the end result of the formation of a mercapturic acid conjugate, which is excreted (Figure 1-5) [28]. A complete description of the complete spectrum of biochemical transformations that form part of this stage of metabolism is beyond the scope of this work, and has been partially compiled elsewhere [1-3]. It is important to note, however, that the metabolites leaving Phase II of xenobiotic metabolism are generally actively transported by permeability glycoprotein or other transporter proteins into the extracellular medium and leave the organism [35].

Figure 1-5. Phase II xenobiotic metabolism does not always end with conjugation, but sometimes involves further covalent modifications, as evidenced by glutathione adducts of benzene products.

1.2.3 Non-Classical Transformations

Of course, the metabolism of xenochemicals is not merely the domain of the various cytochromes, transferases, and transporters, since anything that can happen in nature will happen to a lesser or a greater extent, depending on the circumstances. A compound that is preferentially targeted to a tissue lacking the high levels of cytochromes P450 present in the smooth endoplasmic reticulum of the liver could bear a fate that is radically different to its fate in the liver, which may in turn alter its toxicity. Any enzyme with a sufficiently accommodating active site and bearing a sufficiently chemically active cofactor and permissive mechanism could, in principle, carry out further modifications on the starting compound or any of its downstream products. Indeed, understanding and adequately predicting the reactions that do not form part of the classical xenobiotic metabolism is sometimes indispensable in completely understanding the toxicity and

activity of small molecules, as we have witnessed when considering the redox cycling of quinones. Furthermore, even small molecule-enzyme interactions that do not result in productive metabolic transformations may contribute to the toxicity of a given compound. For instance, some quinones can enter the electron transport chain in the mitochondria and exert a potent toxic effect by stalling it [36]. Ideally, such interactions should be characterized and addressed, or at least flagged for further consideration by the drug designers and toxicology researchers.

1.3 Representation and Cataloging of Metabolic Information

Before we address the various approaches and modes of prediction of the many reactions of xenobiotic metabolism that we have just discussed, we have to consider the nature and state of the relevant biochemical information, as it is this knowledge that often forms the basis of metabolic fate prediction studies. Humanity has been accumulating knowledge about the nature, purpose, dynamics, and mechanisms of metabolic transformations since the discovery and characterization of enzymes in the late 19th century [37]. It comes as little surprise, then, that the largest vehicles for containing this information are not formal databases with well-defined data fields and standardized communication protocols, but rather scientific literature. The automated extraction of useful information regarding enzymatic transformations from literature has proven difficult – not surprising considering the fact that such information rests in the domain of transformations and permutations of chemical graphs and was traditionally described pictorially or verbally [38]. The terms used have often not been systematic, reproducible (e.g. through the use of common names), or applicable to the reconstruction of the chemical graphs being transformed. In order to preserve the nature of metabolic transformation patterns,

enzymatic reaction information should have been mathematically and formally presented using graph theory-derived formalisms, such as SMILES strings [39] or International Chemical Identifiers (InChI) [40]. This may be attributed to the relatively recent (with respect to enzymology) emergence of these representations and formalisms, and the lack of communication and embracement of standards across fields – the SMILES strings, for example, have originated in cheminformatics, while the field of enzymology has been traditionally at the intersection of organic chemistry and biology. Therefore, metabolic information has historically been most often depicted using diagrams, jargon, as well as non-standard descriptions and data representation formalisms, necessitating the employment of human curators for the screening and assimilation of such information. As we shall see, the majority of the currently available databases on metabolism rely on the efforts of human curators, and only a few, like BRENDA [41], are supplemented with automatically-generated information.

Only relatively recently have standards, like Systems Biology Markup Language (SBML) [42] and Biological Pathway Exchange (BioPAX) standard [43], emerged to partially address this problem. Unfortunately, while these standards do address some aspects of reaction and pathway specification, like the introduction of a formal and standard means of representing reactions, modifiers, substrates, and products, as well as the various annotations of these parts, the issue of the identity and automated integration of the various pathways represented using these standards, still stands. For example, even if a reaction is annotated to contain a rigid structure that defines that n-propanol is a substrate of a reaction in which it is converted into the product propanal, there is little means of integrating this information with another reaction that handles the conversion of

propionaldehyde into 1-propanol. A human expert may be able to identify the synonyms among these names and to recall a molecular graph structure that corresponds to these entities, but there is no such option for machines: in order for this information to be amenable to automatic integration, much more standardization, and in particular in the formal and unambiguous representation of chemical entities and their reactions has to be carried out. Although some databases now offer identifiers and structural chemical information for the participating small molecules and links to small molecule databases, this information is immediately useful only to human experts: cross-database querying from within a single interface requires extensive parsing and ‘plumbing’ to connect all the disparate databases. A standard framework to represent and interconnect this information is imperative to enable truly integrative queries across database boundaries. In addition to this, a large portion of scientific information resides within the ‘deep web’, a set of interfaces and tools distributed on the web which can compute the missing information necessary to complete a particular query [44].

Finally, the existing semantic annotations on pathways, such as those found in the BioModels database [45] of manually curated, SBML-encoded pathways, are often contained in a form that requires further processing to be useful. The BioModels semantic annotations are stored in text field as textual annotations on the biochemical reactions, and although they use standard vocabularies for annotations, require processing to be useful and useable. In addition to this, parts of biological entities are difficult to address using these representations. If there is a charge on a particular atom in a molecule, how do we represent that it is this atom, and not some other atom that is charged? If there is an important function for a residue on a protein, or a functional group on a molecule, or a

bond, how do we unambiguously specify this, such that this knowledge can then be used by machines to address our queries regarding these pathways, or to construct a consistent, detailed integrated biochemical pathway map? How can we carry out this annotation task transparently, reproducibly, and uniformly? Unfortunately, these questions have not yet been addressed adequately, and until this is done, the need for manual curation of biochemical information shall persist.

To summarize, although there has been extensive work carried out in standardizing and managing biochemical pathway information, there has been little done to improve the cross-talk between the chemistry of biological pathways and the standard annotations of reactions, pathways, and their composition. Therefore, a new representation scheme is necessary in order to address this problem, a scheme that has standard and unambiguous identifiers for all chemical entities and their components, and is extensible enough to allow for the standard and useful annotation of all biochemical entities. Furthermore, this scheme should be machine-understandable and useful for automated integration of pathway information. The institution of such a scheme can, in principle, eliminate the ambiguity in biochemical publications, and if the principal findings are reported in a machine-readable ‘abstract’ along with the human-readable one, can allow complete and automated reconstruction, integration, and cross-comparison of the reported biological information, as we shall discuss in subsequent chapters.

Despite the best efforts to institute standards for pathway and enzyme information in biochemical research communications, scientific literature is still greatly saturated with non-standard and difficult-to-integrate specifications of reactive transformations and their characterizations. Unfortunately, until this situation changes, manually curated pathway

databases shall be the primary readily accessible container of biochemical information. Biochemical pathways have two principal components: 1) the interaction graph containing all the molecular participants of a given pathway, and 2) the dynamic characterization of each reaction in terms of the kinetic form of each transformation and the kinetic parameters associated with that transformation. Let us examine the databases describing these aspects of biochemical pathways in somewhat greater detail.

1.3.1 Pathway Map Databases

The principal component in characterizing metabolic pathways is a pathway map that includes the transformations of the participating chemical entities. Databases of chemical information and reactions in organic chemistry trace back to as early as the 18th century [46]. However, one of the first and most authoritative efforts in standardizing enzymatic reaction information belongs to the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), which created the Enzyme Commission (EC) numbers to classify the various transformations that enzymes are capable of catalyzing [47]. In this hierarchy of chemical transformations, there are four levels to the classification, such that every enzyme is classified as A.B.C.D where A, B, C, and D, are numbers corresponding to the classes of chemical transformations an enzyme is capable of carrying out. For example, the famous alcohol dehydrogenase enzyme is classified as 1.1.1.1, which means that it belongs to oxidoreductases, acting on the CH-OH group of donors, with NAD(P)⁺ as an electron acceptor, and it acts on primary alcohols in general. The EC numbers have been a major step forward in classifying and cataloging of enzymatic information. The institution of EC number 1.1.1.1 in 1961, along with many others in the subsequent years, has prompted a more

precise redefinition of biochemical transformations, and in the process, has resulted in the accumulation of one of the first comprehensive databases of metabolic transformations, accessible today in electronic form from several sources [47-49]. Although there has originally been only a limited effort to formally specify the compounds and functional groups involved in the catalogued enzymatic reactions in a machine-accessible form that would enable chemically-aware queries, the EC system has been an important step in enzymatic information standardization.

Efforts in organizing the individual metabolic transformations characterized by NC-IUBMB have relatively recently been catalogued in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [50]. KEGG contains the transformations and interactions of various small molecules with the enzymes of various organisms, including *Homo sapiens*. The interactions catalogued in KEGG are a step forward from the previous cataloguing efforts, since each interaction is defined in terms of explicitly specified chemical entities, and this information has already been gainfully employed in numerous reaction prediction systems. Unfortunately, KEGG contains no kinetic information for the catalogued enzymatic transformations, limiting its usability to the training of chemical graph-based metabolism prediction systems only.

KEGG is not the only, but probably the best-known metabolic pathway database. Several other databases for metabolic pathways have emerged as alternatives in the past decade. The MetaCyc metabolic pathway database has been claimed by its authors to be the largest database of its kind to date [51]. The pathway tools distributed with this effort supplement the already extensive and rather well-annotated collection of reactions with predicted pathways for various organisms based on the genes identified in the sequenced

genomes of said organisms. Reactome contains a number of reactions that are not necessarily all geared towards metabolism, but the database can be exported in the SBML form, which means that it can be used more easily in SBML-encoded pathway models, although the gain in productivity is fairly marginal due to the fact that this SBML export still needs to be filtered for relevant reactions and the reactions have to be annotated with the proper kinetic forms and constants to enable modeling [52]. The emerging Rhea database at the European Bioinformatics Institute enriches its reaction definitions with Chemical Entities of Biological Interest (ChEBI) annotations and identifications, making reaction definitions less ambiguous, and opening a door to describing and querying the metabolic pathways in Rhea with concepts and terms from the ChEBI ontology [53]. Finally, the MaCiE database contains not only the enzymatic transformations, but also the mechanisms for hundreds of transformations, carefully manually reviewed from literature [54]. To summarize, there is no shortage in the variety and quantity of biochemical pathway databases, but all the databases currently available are not easily amenable to integrative, cross-domain queries due to the representational and conceptual barriers between them.

1.3.2 Enzyme Kinetics Databases

Although the focus in biochemistry has traditionally been on static maps of biological pathways, constituting a healthy portion of the undergraduate biochemistry curriculum, nature is not static, and cannot be represented in terms of a ‘flat’, two-dimensional graph of chemical entities and their connections. Life is a temporal phenomenon and requires the study of kinetics to adequately describe and understand it. Kinetic information that is so vital to understanding the bioaccumulation and toxicity of small molecules and their

metabolites has rarely been reported in a standard, unambiguous fashion. Apart from a few fundamental mathematical standardizations of reporting enzyme kinetics applicable under certain circumstances (e.g. Michaelis-Menten kinetics), there have been no well-established standards for measuring and reporting enzyme kinetics and energetics in databases. The complicating factors that prevented this from happening are numerous, and still persist in scientific literature and databases. For example, the enzymes under investigation could have diverging pH optima, the phases in which they reside when catalytically active, or they could be subject to much genetic variability, making it difficult to identify one set of representative kinetic and thermodynamic parameters for a given enzymatic transformation. The recent emergence of directed evolution as a means for the development of enzymes with kinetic parameters superior to those of the wild type (or non-mutant) enzymes [55] has further complicated the existing efforts in cataloguing kinetic information for enzymes, since characterization of the enzyme in terms of its amino acid composition has become an additional burden for efforts cataloguing data on enzyme kinetics. There also are a number of factors that are less difficult to account for, such as the species of origin of a particular enzyme – clearly, unless an enzyme is extremely well-conserved across species, its kinetic parameters can be expected to vary between species.

Cleaning up and sifting through the available kinetic data is further complicated by the fact that the EC numbers, a principal tool for characterizing and identifying a given enzyme, are assigned based on the overall reaction catalyzed by the enzyme, and not the mechanism of a given transformation or the configuration of the active site of a given enzyme. Thus, enzymes that may be radically different in composition and potency could

be classified by a single EC number and placed within a single category in a database. When kinetic information is not available for a given enzyme that is important in the kinetic modeling of a metabolic pathway in a given organism, some researchers may use kinetic data from enzymes within the same EC category. However, it is important to assure that the enzymes whose data is used are sufficiently similar in mechanism and composition to the required enzyme. This further complicates manual screening of kinetic data, as each enzyme has to be handled on a case-by-case basis. The overall process of information screening and verification would have been tolerable had there been mechanisms of unambiguously annotating enzyme and pathway information, since that would enable regular queries that could do much of the data screening for the researcher automatically. Unfortunately, the lack of such annotation and data representation standards in some databases means that the vital annotations are often reserved for the ‘comments’ fields, making the human involvement imperative even in selecting the appropriate kinetics information from a database that itself has been a result of extensive manual screening and curation.

Perhaps the biggest publicly available database of enzyme kinetic constants is the Braunschweig Enzyme Database (BRENDA) [41]. BRENDA contains kinetics data on enzyme-catalyzed reactions and their inhibitors. Collated from extensive manual and automated review of tens of thousands of scientific papers, this database is a natural extension of the NC-IUBMB transformation database, since it organizes and classifies information around the EC numbers. Literature references are given for every piece of data, which includes constants characterizing the enzymatic transformations (Michaelis-Menten and catalytic constant), as well as inhibition constants, pH and temperature

optima, and tissue location, among others. BRENDA is a prime example of a database that maintains terms and annotations that are not readily usable for automated analysis and querying. For example, some kinetic constants in BRENDA have annotations indicating experimental conditions (including temperature, pH, or enzyme mutations) under which they were derived, stored as plain text comments.

SABIO-RK [56] is a database of kinetic transformations that takes into consideration many of the limitations we have discussed here and provides annotations that rely on standard ontologies and vocabularies to simplify data filtering and use in metabolic pathway reconstruction projects. For example, the temperatures and some experimental condition annotations, as well as enzyme identity annotations (e.g. whether it is wild type or a mutant) are clearly indicated and are exposed for searching. In addition to this, SABIO-RK allows for export of database information as SBML expressions to facilitate integration into SBML-based models. This task is further helped by the fact that InChI and SMILES are provided for small molecule reaction participants and cross-database links are provided. Unlike BRENDA, SABIO-RK is not solely focussed on approximating all enzyme kinetics with the Michaelis-Menten formalism, but is rather open to a range of kinetic forms and expressions.

Having adequate information in metabolic databases is of extreme importance in the studies that are aimed at predicting enzymatic transformations of small molecules, since this information serves as a benchmark for some algorithms and a vital training guide for others. The emergence of databases (like SABIO-RK and the BioModels database) that contain well-annotated and kinetically characterized metabolic information is an encouraging sign of moving towards information standardization and the resultant

improvements in integrability and usability of biochemical information. The development of such standards as MIRIAM [57] and such formats as SBML is a further major step in this direction. However, no database or standards effort to date has fully solved the problem of automated integration of biochemically relevant information or has allowed for automated cross-domain queries to enable the full awareness of all the stored knowledge from within a single query. In other words, although data is much better represented with the recently formulated formats and standards than it used to be in such databases as BRENDA, the pipelining, connection, and integration of biochemical information is still firmly within the hands of human researchers, as the data does not naturally ‘fit’ together. For example, a cross-database hyperlink or a textual annotation still has to be processed and realized by a human expert or custom-crafted supporting software before this information can be integrated into a unified pathway map or used to address a complex query. Until a new formalism and architecture for biochemical resource integration emerges, we can expect this situation to persist.

1.4 Metabolic Fate Prediction

Metabolic fate prediction (MFP) can be historically traced to the computer-aided organic synthesis design problem, of which it is an extension. Computational retrosynthetic analysis has been under development by commercial and academic organizations since the very early days of computers and chemical informatics and played a major role in fulfilling the need of advising synthetic organic chemists of the potentially most efficient, or least costly pathways for synthesizing a particular compound [58]. The process of retrosynthetic analysis starts with the identification of the compound which needs to be synthesized. Its structure is represented in chemical graph form and matched to the

general classes of the products of any of a large collection of known general organic reactions within a pre-constructed database (hence the importance of adequate database architecture and knowledge integration). Thus, if the end product sought is an alcohol, all reactions producing alcohols are sought. The reactions that are identified as potentially producing the compound at hand are then more closely ranked based on an objective scoring function that could disqualify the reaction either on the basis of the cost of the starting substrates or on thermodynamic, structural, or historical (in terms of occurrences of specific counter-examples in the supporting database) considerations. The acceptable reactions are then processed in order to compute the starting products that could be used to synthesize the desired product. The computation of the starting products is carried out through the rearrangement of the product chemical graph according to the formally expressed rules corresponding to a given reaction, and shall be discussed at length in subsequent chapters. This process is iterated until a certain set of parameters is achieved, like all the starting compounds being available from a commercial vendor at an acceptable price, for example.

The process of metabolic fate prediction is rather similar to this, with the exception that the identification of product class matches in retrosynthetic analysis is replaced with substrate class matches. Thus, if a compound the metabolic fate of which is being predicted is classified to be an alcohol, enzymes that work on alcohols are sought, and the products of these known enzymatic transformations are computed in much the same manner as they are for retrosynthetic analysis. While the identification of *potential* reactions is fairly straightforward, the calculation of reaction feasibility and kinetic parameters in order to prune the resulting metabolic tree to solely the relevant branches is

somewhat more involved. The task of assessing reaction feasibility can be attained using three principal approaches: graph theory-based approaches, quantitative structure-activity relationship-based approaches, and with full-fledged first principles-based studies.

1.4.1 Graph Theory and Expert-Based Approaches

Graph theory-based approaches operate primarily within the domain of the molecular topology, or the two-dimensional properties of the chemical graph. Although many algorithms supporting these approaches can be traced to the early days of chemical informatics, they are still very much in use currently, both in open-source and commercial packages [58]. These approaches aim to learn the generic transformation patterns from a database of transformations or to formally encode expert knowledge on chemical and biological transformations, defining a set of reaction patterns which can then be used to infer the products of a reaction involving a particular substrate.

Figure 1-6. The process of predicting a biochemical transformation in UM-BBD PPS: given a starting compound (toluene, 1), a matching rule is identified and applied to compute a potential product (3).

One of the most primitive examples of such a system is the University of Minnesota Biocatalysis and Biodegradation Database Pathway Prediction System (UM-BBD PPS) [59], where manually encoded reaction patterns derived from literature curation efforts by a panel of biodegradation experts are used to predict the potential products of the degradation of arbitrary small molecules. The prediction system simply identifies the substrate moieties within each reaction definition and computes the products based on the matching biotransformation rule (Figure 1-6). This process is repeated for each subsequent predicted product, until either the transformation threshold is reached, or the most basic compounds and compounds commonly used in metabolism, remain. Each

biotransformation rule is also assigned likelihood, from very likely to very unlikely, based on the informal and subjective opinions of the toxicology experts who created these rules.

Biotransformation rules need not always be manually enumerated by experts, as numerous algorithms to compute these rules have emerged. For example, the researchers behind the KEGG database have recently described the implementation and application of such a system to the prediction of metabolic and bacterial xenobiotic pathways starting with a single small molecule [60]. To accomplish this, they represented the enzymatic reactions as reactive pairs comprised of a substrate and a product, and produced a mapping between the two, noting any changing molecular moieties and any moieties that stayed the same throughout the reaction. The result was a reduction of a reaction in terms of a reactive pair description, which then formed a rule that could be used to identify potential reaction substrates and compute reaction products for arbitrary small molecules. The feasibility of a given reaction was determined solely based on the graph similarity of the predicted reactions to the reactions found in the database. This general approach has also been used in several studies [61-63], and can be used to systematize and automate reaction definition generation for arbitrary enzymes, assuming training data is available.

1.4.2 Statistical Learning and Quantitative Structure-Activity Relationships

Graph theory-based transformation computation is the backbone of any metabolic fate prediction framework – this task is simply inescapable. The details of implementation of transformation prediction are in assigning the likelihood to the various biochemical transformations and computing transformation dynamics. While this can be done with heuristic approximations by only considering the topology of the input compounds or

prior (machine or human) expert knowledge, developments in statistical and *ab initio* approaches to approximating reaction kinetics form an important complement to the heuristics- or expert-based reaction prediction methods in that they provide reaction rate constants, which can then be used to construct dynamic models of metabolism.

When key enzyme kinetics information is not available, the only reasonable remedy may be an experimental characterization of the missing data. However, large-scale experiments of this nature may be costly and time-consuming. Since the more advanced first principles-based approaches are currently not feasible for a large-scale calculation of kinetics of arbitrary enzymatic reactions, Quantitative Structure-Activity Relationship (QSAR)-based models are frequently used for the prediction of some binding and kinetics constants [64]. First, a training set of structurally similar small molecules for which kinetic parameters are known are obtained and analyzed to ensure training sample homogeneity. Then, molecules within the training set are described using a feature set found to be of relevance to a particular predicted parameter. Such features may include molecular volume, polarity, overall charge, presence of key functional groups, bond dissociation energies, ionization energies, among many others. A statistical fit is then constructed between the known kinetic constants and using the values of the relevant molecular descriptors. The study is usually concluded with validation of the model on a test set of compounds that did not participate in model training. Provided that an arbitrarily defined compound is similar to the compounds in the training set, it is reasonable to expect that the model would provide transformation kinetics to its predictive capacity.

This approach continues to have widespread applications in the prediction of ligand binding kinetics, for example. The main advantage of QSAR models is the capacity to rapidly screen compounds for a desired activity. Unlike the more involved first principles-based approaches (e.g. [65]), QSAR-based approaches only require the computation or lookup of small molecule descriptors for substitution into an easily computable mathematical expression.

One final aspect of QSAR-based models that needs to be highlighted is the fact that they do not have to necessarily rely on topologically-derived molecular features. In fact, approaches within the Comparative Molecular Field Analysis (CoMFA) and 3D QSAR analysis families rely on the examination of the geometric configuration of a given molecule in order to predict its properties. Several notable examples of such studies have appeared recently and demonstrated excellent predictive capacity for selected reactions (e.g. [66,67]).

1.4.3 First Principles-Based Approaches

First principles-based approaches are the final frontier of reaction kinetics prediction. These approaches include algorithms and methods based on invariable and universal physical principles rather than statistical fits and correlations to existing kinetic information [68]. This means that unlike QSAR-based approaches, first principles-based studies are not vulnerable to errors in training sets or loss of predictive capacity due to the structural inconsistency of the assayed molecule with the training set, for example. At the same time, these approaches can be improved without training data and are increasingly accessible with increases of computational power. Computational cost has historically been the factor holding back the practical widespread application of such approaches in

large-scale metabolic model reconstruction. As new computer architectures and software algorithms become available, it is reasonable to expect these approaches to gain an increasingly widespread use.

In essence, these approaches aim to obtain the potential energy profile of the enzymatic reaction in order to compute the kinetic parameters corresponding to a given reaction. The potential energy profile is a plot of the energetic state of the reactive system with respect to the reaction coordinate, or the progress of the reaction. For a simple enzymatic reaction $E + S \rightleftharpoons ES \rightarrow EP \rightleftharpoons E + P$, where E is enzyme, S is the substrate, and P is the product, such an energy profile may appear as follows (Figure 1-7) [69].

Figure 1-7. A reaction energy profile for a simple enzymatic reaction with Gibb's free energy of formation on the vertical. Provided this is a Michaelis-Menten reaction, the constants K_M and k_{cat} can be directly computed from the energy differences shown.

The species marked with an asterisk here are referred to as the transition states in these reactions, such that ES^* is the transition state (TS) for the ligand S binding to the enzyme E, to form a stable ES complex, and EP^* is the TS for the dissociation of the product P

from the enzyme. The TS for the actual catalytic conversion carried out by the enzyme is denoted as (ES/EP)* and the difference in the Gibb's free energy of formation of this transition complex that is termed the activation energy barrier. Enzymes function by lowering this activation energy barrier and therefore accelerating the reaction.

One way to describe the kinetics of this reaction, and indeed many other enzymatic reactions, including ones that involve cofactors and coenzymes is the Michaelis-Menten (MM) formalism. MM kinetics are used widely to describe even the enzymes that are typically not expected to conform to such kinetics, and databases like BRENDA and SABIO-RK contain MM parameters for a wide range of enzymes. In MM kinetics, the enzymatic transformation is reduced as follows.

In this chemical equation, k_1 and k_{-1} are the forward and reverse reaction rate constants for substrate binding, respectively. Because the assumption in MM kinetics is that the conversion of the ES complex to the EP complex is irreversible, one forward reaction rate constant for this process, k_2 , is assigned. Among other assumptions of the MM kinetic equation are i) that the reaction is that S and P are not at their thermodynamic equilibrium; ii) ES concentration is constant because the system is in a steady-state, and; iii) that the concentration of enzyme or any cofactors is not limiting the reaction. Expressing the rate of product formation as a derivative with respect to time, the following equation, called the Michaelis-Menten equation, can be obtained.

$$\frac{dP}{dt} = v = \frac{V_{\max}[S]}{K_M + [S]}$$

Here, $V_{\max}$ is the maximal reaction velocity, and $[S]$ is the substrate concentration at a given point in time. The term K_M is the MM constant and reflects how quickly a given compound binds to its active site, the site on an enzyme where reaction occurs. This term can be computed as part of enzyme-ligand docking studies, and usually reflects to a great degree the strength and sign of electrostatic and steric interactions between the enzyme active site and the substrate.

$$K_M = \frac{[E][S]}{[ES]} = \frac{(k_{-1} + k_2)}{k_1}$$

The term $V_{\max}$ reflects the initial unbound enzyme concentration and the parameter k_2 , also known as k_{cat} , the catalytic rate constant. This parameter can be computed directly from the activation energy barrier ΔG^* as follows (k_B is the Boltzmann constant, T is temperature, R is the gas constant, and h is Planck constant).

$$V_{\max} = k_2[E_0] = [E_0] \frac{k_B T}{h} e^{\left(-\frac{\Delta G^*}{RT}\right)}$$

This means that in principle, given enough computational power and sufficiently accurate algorithms to generate a reaction energy profile, it is possible to completely characterize the kinetics of a given enzyme.

Thus, first principles-based approaches are typically concerned with the calculation of the binding free energy to approximate K_M in ligand-substrate docking studies [70], and the activation energy for the enzymatic transformations, at various levels of theory. Docking of ligands to enzymes comes in various flavors, from rapid rigid docking to full molecular dynamics studies, and is routinely carried out in pharmaceutically relevant

research with good predictive results, and has been implemented in a number of well-developed commercial packages, such as the Molecular Operating Environment [71].

First principles-based prediction of enzyme catalysis kinetics has also been addressed at length with multiple methods, from crude reaction free energy-based approximations (with heuristic group contribution, semi-empirical, and density functional methods) to full reaction transition state calculations, of which the hybrid quantum mechanical/molecular mechanical (QM/MM) approach has gained some popularity recently. Briefly, because a full QM reaction transition state calculation is currently associated with a prohibitive computational cost, only the most relevant reactive moieties of the active site, as well as the reacting molecules are treated with QM, while the rest of the protein milieu is treated with the MM approximation (Newtonian forces, Coulombic interactions). These methods have been demonstrated to provide accurate results, albeit they require careful handling by computational experts [66,67].

1.4.4 Commercial Packages for Metabolic Fate Prediction

Metabolic fate prediction is an area of interest of a number of commercial entities, and as such, has received considerable attention from chemical software vendors. There are at least a half dozen well-developed integrated software packages to enable MFP, using various levels of approximation. Let us briefly examine some of these commercial solutions and their features.

META [72] is one of the oldest commercially available packages with the specific aim of metabolic fate prediction, and operates on human-curated dictionaries of biochemical transformations for several types of organisms. This system is very similar in its

principles to the UM-BBD PPS discussed earlier in its functionality, and has been used in commercial and regulatory settings.

MetabolExpert, [73] is a commercial offering similar to META, but touting improved graphical rendering capabilities, and the possibility of graphical reaction rule definition and updateable and user-customizable reaction rule databases as its major advantages.

TIMES [74] is an expert-based system that incorporates the prediction of metabolism based on human-generated rules with toxicological predictions including skin sensitization and carcinogenicity among others. Transformation probability and feasibility are estimated based on the occurrence frequency of similar transformations within a human-curated biotransformation database. Using the kinetics data from within the reference database, the system applies statistical models to approximate the probability and kinetics of the predicted reactions. The main advantage of the system touted by its authors is the integration of toxicity prediction and MFP within one platform.

KnowItAll [75] is an integrated package containing a set of tools that can be used in conjunction to predict metabolism and toxicity, provided the necessary modules are available. The relevant modules are PredictIt Metabolism and PredictIt Toxicity, and are capable of not only MFP, but also toxicity predictions, much like TIMES. The main draw here is the integration with the other tools based on the KnowItAll platform.

Finally, **METEOR** [76] operates on very similar principles to META and MetabolExpert, and touts pathway and compound exporting capabilities, along with an extensive database of human-curated and well-described biotransformations that it operates upon.

It is worth noting that these commercial packages often tout some form of integration with other special features such as toxicity prediction, compound annotation, reaction feasibility prediction, or integration with experimental data handling packages as major advantages. Many of these packages aim to please their users by providing file import and export options into common chemical formats. However, upon closer inspection, neither these packages, nor the tools based on the principles discussed in sections 1.4.1-1.4.3, are geared to work in symphony to achieve an overall goal of metabolic fate prediction, or even to supplement the strengths of one methodology or package with the strengths of another. For example, it may be of interest to not only predict a reaction on a graph theory level, but also to search for it within scientific literature, identify the toxicity of the resultant compound, and compute the kinetics, all from within one query or interface, but using different packages that are known leaders in a given field. Instead of this, researchers in metabolic fate prediction are forced to become experts in multiple packages and approaches, and to manually piece together the metabolic puzzle (hence the highlighting of file import/export options). By what means can one unify the biochemical knowledge and all the disparate computational tools without human involvement? I believe that the Semantic Web holds the answer.

1.5 Semantic Web Technologies

While we cannot adequately cover the entirety of the Semantic Web technologies in a short introductory treatise, we shall cover some relevant aspects of the topic, and would like to refer the reader to more detailed work elsewhere [77,78].

At the turn of the last century, the idea of the Semantic Web had been proposed by Sir Tim Berners-Lee, the man credited with the birth of the internet as we know it [79].

Recognizing the potential of the Internet as an ever-expanding web of data alongside the interlinked HTML documents he proposed earlier, he also noticed that this data was typically highly insulated, and was presented such that little or no machine-understandable meaning was attached to it. As a result, this data was accessible to integration and logical analysis only upon thorough examination by human experts and extensive processing. In other words, though there was *data* on the Internet, it was of little use when one needed to obtain *knowledge*, or address questions that were never asked before or involved multiple domains of knowledge – the data on the Internet existed in the same unlinked state that documents existed before the appearance of the World Wide Web.

Berners-Lee argued that the creation of linked data would provide a solution to this problem by creating the necessary logical bridges between the disparate data points and enabling machine-based data integration and assignment of meaning to data. It was also argued that the traditional forms of data representation, even when supplemented with data structure through XML, were not sufficient for this task alone. That is, though the structure of the documents may have been apparent to a machine through XML, the logical meaning carried in the data was lost as no semantics were carried in the markup.

Figure 1-8. Though it is rather simple to identify the meaning of the information represented in XML within a given domain and given a schema (A), cross-domain integration is somewhat difficult without one (B), and reasoning is impossible as there is no machine-understandable meaning (C).

The proposed solution involved the separation of knowledge into atomic components, the minimal amount of information that would still hold meaning, called triples. These atoms of knowledge would be composed of sub-atomic knowledge particles: a subject, a predicate and an object. Instead of necessarily being values, however, these particles would be represented as *resources*, each with a unique identifier. In this manner, a linked web of data would be created where the end-users would be working with representations of *entities* and their attributes, stored in documents hosted on computers that belong to particular domains. In other words, each sub-atomic particle of knowledge would have a universal resource identifier (URI) assigned to it, would be available on the web through the hypertext transfer protocol (HTTP), and would have linked to it useful information including links to other related entities when retrieved, or *dereferenced*. Thus, one would be able to state that a particular resource has a particular property that is defined by a given vocabulary, and that another resource or a literal value is linked to the resource in question through the given property.

Figure 1-9. RDF n-triple representation of the statement “Leo likes chemistry” (A) and a SPARQL query to identify all entities (with the *x* variable) that like chemistry (B), which will bring back Leo’s URI, if queried over a triple store containing the statement in (A).

Berners-Lee’s proposal has been formalized into a set of specifications by the World Wide Web Consortium (W3C) and materialized as the Resource Description Framework (RDF) [80]. With the assistance of SPARQL Protocol and Query Language to formulate machine-interpretable queries, a beautiful orchard of integrative applications and mashups was finally able to bloom from the distributed linked data [81]. A number of efforts based on this architecture (to be discussed later), first in biology and later in chemistry have also begun appearing in recent years [82][83].

Though the act of producing RDF specifications and ongoing efforts of distributing raw data as RDF will continue to appreciate in importance as only a small fraction of data on the web is currently linked data, RDF does not constitute the fullness of the Semantic Web. Since the Semantic Web aims to add meaning to data, producing logically computable knowledge in the process, one needs to create vocabularies with term hierarchies (interchangeably referred to in this context as ontologies), and adhere to uniform specifications for ontology representation and logical querying of relationships

within the data. The problem of hierarchical vocabulary specification has been initially addressed with the RDF Schema language by introducing facilities to represent and manipulate classes of entities and corresponding class properties [84].

Figure 1-10. Architecture of the Semantic Web.

Further, the Web Ontology Language (OWL) provides a greater selection of logical constructs to express world views as ontologies – formal and explicit specifications of shared conceptualizations [85]. OWL extends RDFS vocabulary with a well understood subset of first order logic including existential and universal quantifiers, qualified cardinality restrictions, sameness, disjointness, and a number of relation attributes (transitive, reflexive, irreflexive, symmetric, antisymmetric, functional, inverse functional). In order to deal with the incomplete, dynamic, and uncertain nature of knowledge, OWL adopts the open world assumption, meaning that a statement not explicitly specified as false is unknown, rather than being automatically perceived as false. OWL also offers a number of syntactic subsets, some of which can be implemented using relational database systems while others require a more sophisticated reasoning

system. Although OWL has been largely adopted in the life sciences community, efforts are underway to promote this knowledge representation language in other domains including translational medicine, geological sciences, and retail.

By decreasing the barrier in representing fairly sophisticated knowledge, there exists a possibility that a number of ontologies may emerge to describe the same domain, thus complicating the vision of semantic web interoperability. Although we expect that not everybody will agree to a common world view, having multiple incompatible ontologies will present a problem in terms of semantic data integration and ultimately would aggravate the very problem that the Semantic Web originally sought to address. Though this may certainly emerge as a problem, mitigating efforts have been undertaken, including the establishment of so-called reference ontologies authored by domain-specific consortia [86]. Moreover, the adoption of upper-level ontologies can provide some philosophical foundation for describing domain entities while also ensuring that domain ontologies can interoperate at the most general level by providing basic types and relations.

Figure 1-11. An ontology of plants and their properties. Here, the class of fruits is defined as anything that is a plant and has sweet taste through the ‘has taste’ relationship. If banana is defined as a plant having sweet taste, it will be automatically machine-classified as a fruit (dashed line).

Practically, the ontology-defined classes and relations can be used to reason about a collection of linked data so as to potentially offer new inferences concerning the relationships of the resources within the linked data that have not necessarily been defined earlier. In fact, given adequate class and entity description, it may not even be necessary to assign class membership – it may be possible to logically infer class membership using a *reasoner*. This flexibility allows a plurality of philosophical world views, as formalized in a plurality of ontologies, to coexist while operating on the same set of RDF data. All in all, despite a number of limitations, the semantic web technologies we have described here offer boundless opportunities for data integration and analysis. Let us consider the applications of some of these technologies in greater detail.

1.6 Semantic Web for Chemical Knowledge Representation

Due to the nature of functional groups of organic chemical entities that impart to their bearers certain reproducible qualities in terms of reactivity and properties, one may argue that classification of chemical compounds and database storage are somewhat natural to chemistry. Efforts to create databases of knowledge relating to organic chemical compounds date back to the Beilstein database, founded in 1881 and cataloguing organic chemical entities that appeared in literature as far back as late 18th century [87]. More recently, the invention of computers had immediately led to the birth of cheminformatics and the development of numerous proprietary computational chemistry algorithms, software packages, and electronic chemical database systems. Unfortunately, the proprietary nature of these packages, and the disparate ways of representing chemical data for the various applications of chemistry have yielded a great number of file formats and specifications, ranging from simple line notations to complex protein structure specifications. As a testament to this, the OpenBabel project, which aims to alleviate this problem by facilitating the interconversion of the various chemical representations, now operates on over 90 chemical file formats [88]. Though there have been calls to standardize electronic information representation within the International Union of Pure and Applied Chemistry (IUPAC), it has so far only managed to arrive at a standard line notation for small molecules called International Chemical Identifier (InChI), which was adopted by major chemical databases [89]. So far, especially with the advent of high throughput chemistry and the increased accessibility of database solutions that resulted in the proliferation of uniquely designed (bio)chemical databases, the problem of unifying chemical data has only grown. It is reasonable to expect that in case the current situation persists and no common chemical information representation scheme is arrived upon to

and from which other representations may be serialized, there may be few solutions to data integration beyond producing application-specific integration programs, the number of which would have to grow as $0.5(n^2-n)$, where n is the number of representation schemes.

Henry Rzepa and Peter Murray-Rust were among the first to recognize this problem and to attempt to bring the words *semantic* and *chemistry* together. Their solution involved the introduction of the Chemical Markup Language (CML), a collection of XML-based vocabularies to act as the one language of chemistry on the web [90]. Over the past decade, CML vocabularies have been developed to address all aspects of chemistry, from compound geometries and configurations to electronic structures to spectroscopic and crystallographic data to data about reactions and mechanisms [91]. CML was also easy to develop for, since it enjoyed a wide support as an XML-based language. A supporting library, called Java Universal Molecular Browser for Objects (JUMBO), had also been created, allowing validation, manipulation, and graphical representation of information presented in CML files [92].

Despite the impressive array of features, CML was never formally accepted as a universal language of chemistry and was rarely taken seriously by the chemical community, as attested by Murray-Rust himself [93]. Further, as we have seen in the introduction to semantic web technologies, CML is a markup language, meaning that it can only be used to declare chemical information in a machine readable format, while the formal, machine-understandable meaning of data is absent; data cannot be classified with CML based on its attributes and class necessary and sufficient conditions. The named elements of CML are restricted to those defined by the schema and are not formally extensible, nor can they

be hierarchically structured as sub-types or super-types or defined as equivalent. Beyond the lack of semantics, it is apparent that data is often under-specified in CML. For example, the specification of bonds is performed through a list composed of literal values of the form “atom1 atom2” where the two atoms are bond members. While this is human-readable, it requires additional processing in the form of a string manipulation step in order to be queried. Thus, although CML has been indeed a very substantial step in the right direction, the omitted *formal* semantics may lead to a situation where CML would end up as yet another chemical file specification, albeit with greater ambitions and more extensive controlled vocabularies than most. The bottom line is that although CML is quite comprehensive in its vocabularies, the capacity to reason over the represented knowledge is non-existent. Although the power of CML and related projects has been demonstrated in a number of applications, we shall only focus on the knowledge representation and integration aspect in this work.

A few years after the original CML proposal, the realization of the usefulness of linked data and RDF-based representations was finally translated into a number of efforts, of which the CombeChem project deserves a special mention as one of the pioneering works in enabling semantic chemistry for the laboratory [94]. The aim of this project has been the development of a complete system for the capture and representation of chemical data in the laboratory with human-friendly tools based on semantic web technologies, and in particular, RDF-based representations of chemical information. The immense loss of insights into effective ways of conducting chemical experiments as well as the loss of information relating to causes of experiment failure and difficulties in retrieval of exact instrument measurement outputs (complete with units and experimental error) could have

been, in principle, prevented had the system gained a wide user base. Additionally, enhanced capabilities to identify potential causes of failed experiments or scientific fraud (in the unlikely event it should occur) were some of the indirect benefits of implementing the system. To further demonstrate the utility of the proposed framework, a semantic journal article has been produced, complete with a working example of infrastructure for semantic embedding of raw data into publications. Supporting ontologies have also been developed to address data measurements and units. The ability to conduct semantic web-enabled open science as part of the CombeChem project had helped inspire a number of Open Notebook Science efforts which managed to gain recognition in the general scientific community, despite being mostly based on pre-semantic web technologies [95].

databases across a wide range of domains is stored in as RDF triples [97]. This federated ever-expanding resource contains billions of triples of life science-related information including data from repositories like PubChem, PubMed, Protein Data Bank (PDB), Kyoto Encyclopaedia of Genes and Genomes (KEGG), and the database of Chemical Entities of Biochemical Interest (ChEBI) among many others, while dedicating significant efforts to ensuring that the represented data *fits*, or is as integrated as possible. Further, the Chem2Bio2RDF effort aims to integrate a range of linked chemical information with a particular emphasis on applications in systems chemistry [98]. Recently, World Wide Web Consortium (W3C) Health Care and Life Sciences Interest Group (HCLSIG) has been formed to formalize approaches and establish best practices for RDF archival of life science data. As a result of the Linking Open Drug Data (LODD) initiative of HCLSIG, eight million (as of 2009) RDF triples of chemical information, collected from over twenty open data repositories, have become freely available [99]. Armed with resources contained within the LODD, Bio2RDF and Chem2Bio2RDF, it is finally possible to pose complex queries that draw on the entirety of the federated linked data resource in order to address such interdisciplinary problems as the assessment of drug effects relative to its polypharmacology or the association of pathway related to drug toxicity, for example [98].

Finally, recent years have seen a growth in data integration interest on the part of major industrial players in chemical research, including Pfizer, Novartis, and GlaxoSmithKline (among multitudes of others), who aim to develop semantic web-enabled frameworks and standards with the end goal of accelerating the process of pharmaceutical drug discovery and development [27]. Indeed, with widespread community interest, backed by the

involvement of the pharmaceutical industry, the future of the Chemical Semantic Web and interdisciplinary sciences like Systems Chemistry appears to be rather exciting indeed.

1.7 An Integrated Metabolic Fate Prediction Framework

In this work, I aim to break down many of the barriers that stand in the way of knowledge integration and computational resource interoperability, and ultimately in the way of adequately addressing such problems as MFP and such fields as Systems Biology and Systems Chemistry (Figure 1-13). My aim in doing so is to gain the capacity to automatically construct integrative dynamic metabolic fate models to support the identification of the mechanisms and sources of the toxicity of arbitrary small molecules in the human body. My hope is that in the future, computational functionality could be distributed as semantically self-aggregating web services that would provide support for problems across disciplines and institutions, providing life scientists with the ability to gain new insights into chemical toxicity, and to alert them of the nature and the consequences of metabolic transformations that could be missed even with well-characterized experimental systems, as observed earlier.

Figure 1-13. Currently, a number of barriers are preventing the streamlined knowledge integration that is necessary for interdisciplinary fields like Systems Chemistry and MFP.

My objective in this work is to explore metabolic fate prediction using a first principles-based semantic biochemical reactor. To achieve this, I construct a prototype integrated framework for metabolic fate prediction based on Semantic Web technologies (Figure 5-1). I demonstrate that this integrated framework bridges the gaps between computational and database resources to provide a dynamic model of the metabolism of arbitrary small molecules that could be used to predict chemical toxicity and consequences of metabolic degradation.

In this framework, all chemical entities, from atoms to reactions are semantically specified with the Chemical Entity Semantic Specification (CHESS), an RDF-based format. All enzymes are encoded as individual Semantic Web Services within the SADI framework, with formally defined chemical classes for inputs (reflecting enzyme

behaviour in nature), and producing CHESS-encoded reaction descriptions as outputs. Because the input small molecules are fully annotated using the CHESS specification, they can be structurally classified using machine reasoners, enabling the automatic identification of acceptable enzymatic web services for their transformation. The reaction specifications produced by these enzymatic web services contain reaction products, which can feed back into the framework, and the entire procedure can be iterated until a desired goal is satisfied. The reactions, along with their kinetic annotations and the functional classifications of each small molecule product of metabolism of the starting compound can then be integrated to construct and solve a mathematical model of metabolism. Finally, the functional classifications can be used to identify the most toxic byproducts of metabolism, providing mechanistic explanations of compound toxicity.

To this end, Chapter 2 shall discuss the CHESS formalism. Chapter 3 shall address the issue of small molecule classification, both structural and functional, using OWL ontologies. In Chapter 4, I shall demonstrate the implementation and functioning of SADI web services. Chapter 5 shall provide a proof-of-principle framework of semantic web services for metabolic fate prediction. Finally, Chapter 6 shall summarize the findings of this dissertation and propose further avenues of development for the prototype semantic MFP framework.

2 Chapter: Chemical Entity Semantic Specification

This work is primarily based on the following open-access journal article.

Chepelev LL, Dumontier M. Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. *Journal of Cheminformatics*. 2011. 3(1):20.

The CHEMINF ontology used for parts of this work has been reported in the following journal article.

Janna Hastings, Leonid L Chepelev, Egon Willighagen, Nico Adams, Christoph Steinbeck, Michel Dumontier. The Chemical Information Ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLOS One*. 2011. *In Press*.

2.1 Abstract

Over the past several centuries, chemistry has permeated virtually every facet of human lifestyle, enriching fields as diverse as medicine, agriculture, manufacturing, warfare, and electronics, among numerous others. Unfortunately, application-specific, incompatible chemical information formats and representation strategies have emerged as a result of such diverse adoption of chemistry. Although a number of efforts have been dedicated to unifying the computational representation of chemical information, disparities between the various chemical databases still persist and stand in the way of cross-domain, interdisciplinary investigations. Through a common syntax and formal semantics, Semantic Web technology offers the ability to accurately represent, integrate, reason about and query across diverse chemical information.

Here we specify and implement the Chemical Entity Semantic Specification (CHESS) for the representation of polyatomic chemical entities, their substructures, bonds, atoms, and reactions using Semantic Web technologies. CHESS provides means to capture aspects of their corresponding chemical descriptors, connectivity, functional composition, and geometric structure while specifying mechanisms for data provenance. We demonstrate that using our readily extensible specification, it is possible to efficiently integrate multiple disparate chemical data sources, while retaining appropriate correspondence of chemical descriptors, with very little additional effort. We demonstrate the impact of some of our representational decisions on the performance of chemically-aware knowledgebase searching and rudimentary reaction candidate selection. Finally, we provide access to the tools necessary to carry out chemical entity encoding in CHESS, along with a sample knowledgebase.

By harnessing the power of Semantic Web technologies with CHESS, it is possible to provide a means of facile cross-domain chemical knowledge integration with full preservation of data correspondence and provenance. Our representation builds on existing cheminformatics technologies and, by the virtue of RDF specification, remains flexible and amenable to application- and domain-specific annotations without compromising chemical data integration. We conclude that the adoption of a consistent and semantically-enabled chemical specification is imperative for surviving the coming chemical data deluge and supporting systems science research.

2.2 Introduction

The importance of cataloguing and adequately representing chemical information has been realized fairly early in the development of chemistry and related sciences. From the dawn of the era of organic synthesis, thousands of chemical entities, reactions, and experimental outcomes were catalogued and stored in a human-readable form, some dating to as early as the eighteenth century when the understanding of molecular reactivity and chemical structure was nowhere near its current level [1]. During the relatively long history of the development of chemical information archiving technologies, a large number of persistent redundancies and factors complicating chemical knowledge federation have been introduced. It may be argued, however, that these problems may be reduced to three major categories, some of which have been only recently partially addressed: i) a lack of consensus canonical identifiers of all chemical entities, including reactions and macromolecules, as well as their constituents, ii) absence of a single common flexible representation to satisfy the needs of most sub-disciplines of chemistry, and iii) a lack of a consensus chemical database structure or schema. We argue that the bulk of present-day complications in integrating chemical information can be traced to these three problems and that until an information representation that addresses these issues is introduced, truly integrative chemical research shall be a complicated and costly endeavour.

2.2.1 Consensus Chemical Entity Identifiers

Many modern chemical databases rely on internal chemical entity identifiers which are usually created in a sequential manner, producing an index whose value increases for every new chemical entity in the database [2-5]. Unfortunately, such indexing systems

require manual or semi-automated cross-database matching, resulting in difficulties of federating chemical data. Attempts to create reproducible and canonical molecular identifiers that bear molecular graph information could be dated to the introduction of computers to the field of chemistry in the 20th century. The fact that molecules could be represented as graphs and features of these graphs could be used to arrive at a shorthand depiction of molecular structure, had swiftly led to the creation of the first fragment-based line notation in chemistry: the Wiswesser Line Notation (WLN) [6]. In this notation, molecular fragments were abbreviated and recorded with a limited character set to reconstitute various molecular parts and their connectivity. Unfortunately, no efficient way to create a *canonical* molecular representation for this line notation existed, meaning that a given molecule could be referred to by multiple different WLN strings in different chemical databases. This shortcoming was overcome with the introduction of the SMILES notation [7], which explicitly represented chemical molecules as graphs with atoms being nodes and bonds edges, along with an efficient algorithm to create a canonical, reproducible SMILES string representation of a given molecule. Unfortunately, multiple algorithms for SMILES canonicalization have been devised over the years, leading to software-, and therefore, database-specific canonical molecular SMILES representations. Finally, The International Chemical Identifier (InChI) notation has addressed this issue by providing algorithms and software to enable consistent canonical representation of chemical structures, but has unfortunately not yet addressed the efficient canonical representation of many other chemical entities, such as reactions and macromolecules [8]. More recently, due to the unwieldy nature of InChI for many larger molecules as well as web search engine complications, InChI keys have been

introduced, producing a 25-character hash based on the elements of chemical graph structure [8]. Although InChI keys cannot be used to reconstitute chemical structure without lookup tables, their use has enabled cross-database chemical searches using common web search engines. It is not unreasonable to believe that a universal adoption of the IUPAC standard InChI keys in the role of database indexes could potentially facilitate knowledge federation immensely.

2.2.2 Common Chemical Information Representation Platform

Simple line notations have been useful as chemical structure identifiers and bearers of information necessary for the vast majority of cheminformatics tasks, such as chemical database searching or basic reactive transformation outcome predictions. However, these notations could not address the needs of structural, biological, and computational chemists, among others. For this purpose, myriads of chemical file formats incorporating elements of discipline-specific controlled annotations and geometric molecular configuration, have been developed over the past half century. One of the most popular formats to address this need has been the Structure-Data File (SDF) [9], which combines molecular structural and atomic connectivity information with data annotations. Unfortunately, these annotations may often be confusing or contradictory, as they commonly bear no units, data source information, or specific references to the moieties or molecular entities to which the annotations correspond. For instance, an octanol-water partition coefficient annotation may be specified as follows.

```
> <logP> .  
1.3856
```


While it may be a straightforward annotation to the creators of a given database, and while it may be more or less easily interpreted by a human agent, it bears no information with respect to corresponding units, algorithms, or parameters used in generating this value. Furthermore, if two different databases containing SDF data for partition coefficients for the same molecule were to be integrated, this integration would require human interpretation, specialized parsers, and if a relational database is used, a special field to store this information in order to enable queries over it. This task is convoluted by the limited availability of annotation specification or outright lack thereof, prompting many cheminformatics applications to re-evaluate descriptor values in a given study.

Thus, an ideal representation would be able to refer to every chemical entity and its part unambiguously and to capture information in a controlled, reproducible, and machine-understandable way to enable machine reasoning and to facilitate data integration. To address this, numerous XML-based chemical representation schemes have been created (e.g. [10-12]), enabling highly detailed reaction modeling and chemical representations but unfortunately they have neither been widely embraced by the chemical community, nor have they allowed for seamless machine-mediated information integration. One XML-based representation, the Chemical Markup Language [13], backed by a controlled vocabulary, has been rather successful in specifying most aspects of chemistry, from small molecules and their connectivity to polymers and crystal structures [14].

```

<?xml version="1.0"?>
<molecule xmlns="http://www.xml-cml.org/schema">
  <atomArray>
    <atom id="a1" elementType="O"/>
    <atom id="a2" elementType="H"/>
    <atom id="a3" elementType="H"/>
  </atomArray>
  <bondArray>
    <bond atomRefs2="a1 a2" order="1"/>
    <bond atomRefs2="a1 a3" order="1"/>
  </bondArray>
</molecule>

```

Figure 2-1. A simplified specification of a water molecule in CML.

Unfortunately, while most elements of this specification can be parsed out using one of the many XML libraries, certain elements do not render themselves to facile interpretation. Consider the sample CML specification of a water molecule (Figure 2-1). In order to identify the member atoms in a given bond, it is necessary to carry out string processing as an intermediate step. Further, while many of the elements of CML are defined in a controlled vocabulary, the lack of explicit, consistent, and formal axiomatization of the involved concepts gives rise to difficulties in inferring connections between chemical concepts where no such connections are stated explicitly, something that is possible in formal ontology-backed RDF-based information specifications. Although CML specifications have been increasingly evolving to incorporate elements of the Semantic Web, the lack of widespread adoption of the format, and the limited availability of large-scale CML-based chemical knowledge repositories, have somewhat limited CML-assisted federation of the world of chemical data. Furthermore, the implementation of coverage of additional chemical concepts in most chemical representations requires a formal, rigorous representation specification, complicating the

incorporation of data represented using domain-specific representation extensions. We believe that an ideal chemical representation would require no specialized wrapper or interpreter, would be generic such as to allow for facile and conflict-free extensions, would be based on a formal ontology, and would be encoded in a machine-*understandable* (bearing formal, computable meaning, as opposed to simply machine-readable, as in CML) manner and therefore facilitates automated reasoning and data integration.

2.2.3 Chemical Knowledge Integration

The final point of contention in the world of chemical information archiving is a universal open architecture for chemical databases. Chemical data currently exists in a large collection of application- or institution-specific databases that offer little in the way of an integrative searching approach. In fact, many of these databases expect the end user to rely solely on the information that they provide in their research. In the world where cross-discipline borders are increasingly disappearing, such philosophy should have no place or foothold. It should be possible to seamlessly query for, say, the physical properties of a given chemical entity, as well as for its metabolic fates and toxicity data, and ordering information, all from a single interface. Though a number of databases, such as PubChem [3] and ChemSpider [4] currently offer database cross-links to a number of relevant information providers, the perusal and integration of this information still requires human or human-assisted procedures. Furthermore, until an explicit mapping to a given data source is introduced within a database interface, this data source is inaccessible or difficult to access, making the data practically non-existent to the users of these databases. This situation is complicated by the fact that many data repositories and

web services that could potentially generate the required data require unique interfaces and access methods, leaving an immense amount of potentially useful information inaccessible.

With the advent of the Semantic Web, a number of these issues have been addressed. With the concept of linked data and the Resource Description Framework (RDF)-based knowledge representations, a new way of modeling, querying and distributing data became available [15]. With RDF, knowledge is represented in terms of subject-predicate-object triples, where each member of a given triple may be a dereferenceable Universal Resource Identifier (URI) for a particular concept or entity. Thus, what was traditionally referred to as a database, could now be considered a knowledge base, as the initially inert data points were given a machine-understandable meaning through reference to formally specified concepts in supporting ontologies. Thus, two entries from two different knowledge bases could be inferred to relate to the same concept even if no such inference had been explicitly stated, through machine reasoning over axioms in the supporting ontologies. Furthermore, truly integrative queries that could draw on the entirety of the linked data web have now become a reality.

A number of efforts [16-19] have already been successful in the integration of a large portion of chemical information into the linked open data cloud, demonstrating the utility of doing so with successfully fulfilled integrative queries. Such integrative efforts address a number of issues, from the representation of small molecules and their adverse effects to explicit specification of multiple facets of macromolecular structure and interactions. In fact, the chemically-relevant data cloud constitutes a major portion of the entirety of the linked data available on the web [20]. Many of these efforts provide facile means and

tools, such as our chemical information RDFization plugin for Open Babel, to represent and distribute any arbitrary chemical information on the Semantic Web [21]. Although these efforts provide a means of integrating chemical information and of breaking data out of domain or institutional data silos and exposing it for common searches, the triplified data often bears, to a greater or a lesser extent, the same problem as those found in the original databases. For instance, multiple redundant entries for a given molecule, based on database-specific indices may exist without explicit equivalence assertions. Parthood relationships may be missing from such knowledge-bases altogether, and chemical descriptors may be assigned to entities without reference to generating software, experimental conditions, parameters, or data sources. Finally, many of these serializations implement specifications that preclude facile extension of asserted knowledge. For example, a model where an octanol-water partition coefficient is expressed through a predicate and a value, as in '*ethanol hasPartitionCoefficient 1.5*' is not as readily amenable to specification of the conditions under which this value had been generated as its counterpart where the descriptor is given a URI and is fully annotatable with the required information.

2.2.4 Overview

To rectify the aforementioned chemical information integration problems, we propose CHESS, an RDF-based chemical information specification that is backed by the CHEMINF ontology [22]. Here, we shall only focus on selected aspects of semantic chemical information encoding with CHESS, emphasizing principles and consequences rather than specification details. Thus, we shall explore the representation of molecules and all of their constituents with the exception of electrons, representation of chemical

descriptors, the consequences of our representation in terms of efficiency of chemical database searches *without specialized cheminformatics plugins*, and finally, we shall briefly cover reaction representation and implications of our representation on reaction candidate selection. By no means do we claim that the representation specification presented here is complete, but would like to rather refer the reader elsewhere for a more detailed and rigorous explanation and implementation examples [23].

2.3 Results and Discussion

2.3.1 CHESS Representation Overview

The underlying principle in CHESS is to minimize the amount of context-specific rules and regulations, while maximizing the coverage of information represented with the given set of rules. We have also followed an expanded set of principles and requirements in formulating CHESS specification in order to ensure its suitability as a universal chemical exchange language on the Semantic Web, as follows.

1. The most important requirement for CHESS as a universal chemical information framework is the ability to identify and represent chemical entities in a database-, software-, and discipline-independent fashion. For this purpose, we recruit InChI keys and canonical atom numbering arising from the InChI canonicalization algorithm. This also means that atoms, bonds, and functional groups within a molecule have unique and consistent identifiers. Furthermore, all other (physical and informational) entities, such as descriptors, reactions, and macromolecules should also have canonical representations from which consistent identifiers could be obtained.
2. The flexibility and extensibility to support interfacing with and capture of information from a majority of the existing file formats and databases. In previous work, our group

has demonstrated OWL serialization of approximately 100 different structure formats with an Open Babel plugin [21]. We continue this trend by providing means to triplify topological and structural information encoded by SMILES, InChI, and SDF chemical formats, as well as by providing an extensible model for specifying chemical properties and descriptors.

3. CHESS should be descriptive enough to provide means for capturing data at various levels of granularity, from atoms to substances, as well as heterogeneity of data from molecular orbitals to chemical reactions. The information thus captured should preserve explicit correspondence to the circumstances of its creation and the other related data points. That is, all the positional descriptors for the atoms in a particular molecule should preserve a correspondence to each other, reconstituting a single conformer, as well as to the parameters and conditions under which they were observed or computed.

4. CHESS should be supportive of Semantic Web Technology-based implementations of basic cheminformatics tasks pertinent to useful analysis of chemical information, such as Semantic Web-assisted drug discovery, chemical similarity searching, or reactive pattern matching.

5. Concepts used in CHESS must be backed by a formal ontology in order to facilitate reasoner-mediated integration of chemical information. For the purposes of our specification, we have chosen the CHEMINF ontology.

Thus, the overall specification for CHESS is quite simple, and involves only three broad categories of players: i) chemical entities, comprised of reactions, complexes, molecules, functional groups, bonds, and atoms (but extensible to e.g. electrons, macromolecular assemblies or even subatomic particles), ii) chemical descriptors which could be entity

variant or invariant or could be complex and contain multiple CHEMINF ontology-typed descriptors, each with appropriate value, uncertainty, unit, and chemical configuration annotations, and iii) chemical configurations themselves, which reflect upon the sum of relevant conditions under which data has been derived, as well as the sources of data (Figure 2-2).

Figure 2-2. A simplified overview of the general features of the CHESS specification. Please note that the chemical configuration is a reflection of the sum of the conditions that may change the value of a given descriptor, as well as the data source.

2.3.2 Molecular Specification

Let us consider in greater detail the methodology of CHESS specification generation on the example of an ethanol molecule. First of all, it is necessary to decide which chemical entities are of interest in a given study. Here, we shall focus on the molecule itself as well

as the connectivity of its constituents, including functional groups, bonds, and atoms. In order to respect the first principle of CHESS, it is imperative to generate unique canonical identifiers for each of these entities, according to a set of simple rules that will consistently result in reproducible, database-independent identifiers, some of which we outline here (Figure 2-3). Please note that although we use our own base URI in our work (http://semanticscience.org/resource/CHESS_), we would like to invite the wider chemical community to initiate a discussion on adopting a single, standard base URI for semantic chemistry, to which all entities will be assigned.

Molecule	Standard InChI Key
Functional Group Instance	Containing Entity InChI Key, FG, Hash: (Canonical Specification, Ordered List of Atoms)
Bond	Containing Entity InChI Key, B, Canonical Specification, Ordered List of Atom Indices
Atom	Containing Entity InChI Key, A, Canonical Index, Symbol
Reaction	R, Hash: Ordered List of Reactant, Catalyst, and Product InChI Keys, Stoichiometric Coefficients

Figure 2-3. Principles for generating canonical identifiers for some of the many chemical entity types. Please note that these identifiers are for instances of chemical entities rather than classes of chemical entities (e.g. all oxygen atoms or all C-O bonds) and necessarily involve the canonical identifier of their containing entity, molecule in this case.

In order to enable reasoning and inference over this chemical information, each represented entity also has to be explicitly assigned to a class that is defined within a

supporting ontology. This is important to enable querying over broad general concept topics, such as the identification of all instances of oxygen atoms in a given database, or the weakest bond of a particular type in a given molecule, for example. For this purpose, we draw on the concepts present in the Chemical Entities of Biological Interest (CHEBI) ontology and the Semanticscience Integrated Ontology (SIO) [24] to assign general classes to the appropriate chemical entities. For instance, the ethanol molecule may be assigned to a general class of molecular entities (CHEBI:23367), or if the correspondence is present, to a more specific class of molecules, such as that of primary alcohols (CHEBI:15734). Functional groups or molecular substructures may also be assigned to a general class within the CHEBI ontology (e.g. CHEBI:33249), bonds to an appropriate SIO class (SIO_011118), atoms to their appropriate types in CHEBI (e.g. CHEBI:25805 for oxygen), and reactions to the *chemical reaction* class in SIO (SIO_010345). The end-user is not limited to the pre-defined classes in the SIO or CHEBI ontologies. Because these classes are fully extensible, it is possible to define a more specific class for each of the chemical entities presented here. For instance, one may extend SIO's *covalent chemical bond* (SIO_011118) to create a subclass corresponding to carbon-oxygen single bonds, or extend the broad class of functional groups or molecular substructures in CHEBI to correspond to a class of substructures that satisfy or exactly match a general pattern, such as CCO, as we shall see later.

Though functional group or substructure specification is optional, as is that of any component not relevant to the chemical information represented, in order to demonstrate chemical database searching and reaction candidate matching in this study, we have automatically generated a set of unique atom-centric molecular sub-graphs consisting of

heavy atoms and containing first, second, and/or third neighbours of each heavy atom in a given molecule. For example, the oxygen-centered fragmentation products of *n*-propanol of connectivity 1, 2, and 3 are the hydrogen-suppressed graphs CO, CCO, and CCCO, respectively. These fragments are given unique and reproducible identifiers computed from the fragment chemical graph structure and the canonical indices of the member heavy atoms, as well as their molecule of origin. It must be noted that the CHES specification itself is not limited to this automatically generated set of fragments, but rather we are using this set in order to achieve further goals of enabling chemical similarity searching. Customized fragment or functional group annotations, just like annotations of any other type, may be added to the triple store containing the chemical entities under investigation at any time.

To complete the semantic description of the molecular skeleton, mereological relationships between the various sub-components of a given entity have to be asserted. These relationships are captured with *has proper part* (SIO_000053). Based on this information, the complete chemical graph can be reconstituted, and our chemical entity under investigation is ready for further annotation or querying. Here, we provide examples of the specification for each entity discussed (Listing 1).

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix sio: <http://semanticscience.org/resource/SIO_>
@prefix chess: < http://semanticscience.org/resource/CHESS_>
@prefix chebi: <http://purl.org/obo/owl/CHEBI#>

#Specify ethanol as a CHEBI molecule using its InChI key
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N
rdf:type chebi:CHEBI_23367.

#Specify ethanol's alcohol group as a CHEBI organic group instance.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-
FGd787664f213a8ffded68d4d945d012b0cfaf7aa4
rdf:type chebi:CHEBI_33247.

#Specify ethanol's C-O bond as an instance of SIO's single bond.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-BCO23
rdf:type sio:010498.

#Specify ethanol's oxygen atom as an instance of CHEBI's oxygen atom.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-AO3
rdf:type chebi:CHEBI_25805.

#Assert proper part between the oxygen atom and the molecule.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N
sio:000053 chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-AO3.

```

Listing 2-1. RDF/N3 CHESS representation of ethanol and its constituent parts.

2.3.3 Semantic Web-Enabled Cheminformatics: Chemical Searching

Since the specification we have so far described can be used to reconstitute the molecular graph, we can now venture to study the most optimal approaches to enabling some of the most common tasks in cheminformatics. Here, we shall initially focus on the classical task of chemical database searching by chemical similarity. Although it is possible to invoke plugins or intermediary specialized software to enable rapid database searching, we argue that unlike many other information representations and formats, CHESS allows us to fully represent the chemical graph and should therefore be readily amenable to graph manipulation and similarity searching. Furthermore, we believe that, within the limit of providing an expressive enough specification of chemical entities under investigation, the efficiency of querying a knowledgebase created using a given knowledge representation is a good indicator of the efficiency of the representation itself.

Chemical similarity searching is a complex topic that lies at the heart of cheminformatics and can be carried out in a wide variety of ways to address a number of problems. Because our specification provides us a complete chemical graph description, we have chosen to first attempt a Semantic Web-native solution for the substructure matching problem, using description logic-safe rules [25] and SPARQL query language-based queries [26] on the molecular structure. As a benchmark, we have used representative subsections of the LIPIDMAPS database [27] of lipids and their structures of sizes 10, 100, and 10000 molecules in partitions DB10, DB100, and DB10000, respectively. As query graphs, we have used a series of linear carbon chains, from ethyl to pentyl, cyclopentene, and a number of lipid-related functional groups, including glycerol, sterol, a fatty acyl moiety, a sphingolipid moiety, and a prenol lipid moiety.

While carrying out our tests, we have become aware of the complexity of modeling bonds as explicit entities. While this specification allowed for facile annotation of bonds with properties and descriptors, it resulted in significant search performance hits, forcing us to reconsider elements of our specification. As a result, we have created a test set where, apart from specifying explicit bonds, we also linked bonded atom instances with the appropriate bidirectional relationships that corresponded to bond type (single, double, triple, or aromatic). This improved performance considerably and allowed us to carry out our tests. Herein we find another demonstration of the versatility of semantically enabled information representations: we have been able to extend and amend the information in our knowledge repository without much additional effort or adverse effects on the knowledge repository. So long as our specification is consistent with the formal axioms

underpinning the concepts in a supporting ontology, there is no barrier preventing us from extending our specification indefinitely.

One may search for molecules containing a particular sub-graph using the explicit specification of the structure of the sought molecular sub-graph as a SPARQL query (Listing 2). It must be noted that answering this query in a typical SPARQL engine involves the exhaustive examination of all the candidate molecules that may potentially contain a collection of atoms that satisfy the laid out bonding criteria. Unfortunately, since the SPARQL query engines currently available have not been explicitly optimized for chemical searching needs, they often lack many of the mechanisms developed over the past decades to accelerate the solution of this problem, and resemble the brute force approach to graph matching more closely.

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix sio: <http://semanticscience.org/resource/SIO_>
prefix chess: <http://semanticscience.org/resource/CHESS_>
prefix chebi: <http://purl.org/obo/owl/CHEBI#>

SELECT distinct ?m
WHERE {
?m rdf:type      chebi:25367. #molecule
?a0 rdf:type     chebi:27594. #carbon atom
?a1 rdf:type     chebi:27594. #carbon atom
?m sio:0000053 ?a0.      #has proper part
?m sio:0000053 ?a1.      #has proper part
?a0 chess:'has single bond with' ?a1.
FILTER(?a0 != ?a1).
}
```

Listing 2-2. An automatically generated (based on graphical user input) SPARQL query to identify all molecules containing an ethyl subgraph, altered for improved readability. Note the use of ‘has single bond with’ direct atom relationship to improve query performance.

Description logic-safe rules may be used to reason about instances of an OWL-DL ontology where the description follows a graph-like pattern instead of the general ‘tree-

like' expression. In this way, one can combine classical DL-reasoning with graph-like descriptions to classify molecules as more specific kinds of compounds, provided that they satisfy certain sub-graph conditions. For instance, an ethyl compound would be a kind of a molecule that contains an ethyl group with two carbon atoms linked together by a single bond. Overlooking for the sake of simplicity of the immediate example the other necessary conditions, such as the lack of branching or membership in a ring, this may be represented as the following rule (Listing 3).

```
'molecule'(?m1), 'carbon atom'(?a0), 'carbon atom'(?a1),  
'is part of'(?a0, ?m1), 'is part of'(?a1, ?m1),  
'single bond'(?a0, ?a1), 'different from'(?a0, ?a1)  
-> 'ethyl containing molecule'(?m1)
```

Listing 2-3. An automatically generated (based on graphical user input) dl-safe rule for identifying an ethyl group-containing molecule, altered for improved readability.

Both, the SPARQL querying, and the rule-based reasoning, completed in the allocated time of 120 seconds on the simple carbon chain-based queries. However, the iterative identification of compounds with more complex substructures, such as those relevant in the classification of lipids, fails with rule reasoning using DB10 for the majority of molecules in the database. In all cases, we observed either memory exhaustion (despite having allocated 5GB of memory), or premature termination. Note that the time limit was imposed after observing that even if Pellet was allowed to reason over an unconstrained amount of time, this would only delay termination due to memory exhaustion. SPARQL-based query answering, on the other hand, succeeds in identifying molecules containing many fatty acyl patterns, but fails between 20-40% of the time in DB10, and between 60-70% of the time in DB100 for prenol lipid, sphingolipid and sterol lipid SPARQL queries. Generally, SPARQL query completion exhibits a dramatic drop across all test

cases with increasing number of atoms in the database molecules searched at a given time (Figure 2-4). In addition to this discouraging result, the increase of search pattern size or complexity had an even more profound effect on query completion, from exceeding the completion time limit to SPARQL query engine-triggered query termination due to the query computational load exceeding any reasonable expectations, of 11 years, for example (results not shown).

Figure 2-4. SPARQL query completion and average molecular size in DB100.

In terms of performance for queries that were successful, both DL-safe rule reasoning and SPARQL-based querying are 1-4 orders of magnitude slower than using Open Babel, and DL-safe rules are 1-2 orders of magnitude slower than SPARQL queries. By comparing the number of atoms in the query structure to the completion time we observe that Babel

performance is linear, but the performance of searching with both, SPARQL and DL-safe rules, appears exponential or parabolic (Figure 2-5).

Figure 2-5. General performance trends for the two query modes relative to Open Babel matching. Note that the presence of two points at six query atoms for SPARQL queries is due to alternate structures: cyclical structures are more time-consuming than linear ones.

Given these discouraging results, even for an extremely small data set, it was clear that an alternative approach or a redefinition of the problem to a more manageable one was in order. As we have discussed, by incorporating functional group/sub-structure information at the time of creating the molecular specification, or by adding it to the existing representation stores, we can redefine our searching problem. By doing so, it becomes possible to use the same device as the one used in fingerprint-based searches: in order to

identify database molecules that are similar to the query, it is simply necessary to rank them in terms of the number of substructures that belong to the same class as those in the query. The resultant query exhibits relatively rapid completion times that depend weakly on the complexity of the queried structure and linearly on database size (Listing 4). Overall, the completion time of this query for DB10000 is well within the stipulated time limit for all the queries attempted (results not shown). Although this is encouraging, many commercially and scientifically important chemical databases enumerate several orders more molecular entities, casting a shadow over the applicability of this approach to large-scale applications, until a more detailed study of search performance demonstrates otherwise.

```
PREFIX sio: <http://semanticscience.org/resource/SIO_>
PREFIX chess: <http://semanticscience.org/resource/CHESS_>

select distinct ?x count(?y) where {
  ?x rdf:type chebi:25367. #molecule
  ?x sio:0000053 ?y.      #has proper part
  FILTER (?y = chess:FG1 || ?y = chess:FG2 )
}
ORDER BY (100/(count(?y))) LIMIT 100
```

Listing 2-4. The general form of the SPARQL query to identify molecules similar to a queried molecule, which can be obtained from graphical user input. Query amended for readability.

Although we have not tested the more efficient tools currently available, and although the performance of SPARQL query engines [28] and machine reasoners [29] continue to improve, we can see that chemical similarity searches that rely solely on existing Semantic Web tools are possible, but may be problematic for very large chemical knowledge bases. Certainly, we believe that optimization of Semantic Web-based chemical searching solutions with the latest and the most efficient tools, as well as the application of existing tools to very large (over a million molecules) stores of chemical

information, warrants further elaboration in an expanded, separate study. Other, significantly more rapid searching solutions that draw on both, the Semantic Web technologies, and existing methodologies, are also available. For example, it is possible to create custom SPARQL functions that encapsulate specialized code to either carry out pairwise similarity comparisons between a query and database molecules by reconstructing and comparing chemical graphs from InChI, SMILES, or SMARTS annotations. It is also possible to envision specialized functions to carry out simple Tanimoto comparison of query and database molecule fingerprint strings. Finally, semantically-enabled web services may be used to carry out such searches [30]. However, the purpose of this excursion has not been the immediate creation of a solution that could outperform specialized and streamlined code carrying out optimized sub-graph detection or similarity calculations on in-memory stores of molecular fingerprint strings [31]. Rather, the fact that it is possible to attempt parser- and specialized tool-free analysis and integration of chemical data, demonstrates the potential, power, and versatility of investigations afforded by adopting a semantic specification of chemical entities.

2.3.4 Chemical Descriptor Specification

Having explored in detail semantic specification of various chemical entities and their parts, let us turn our attention to their annotation. Chemical annotations may be classified into two broad types: those that are dependent solely on the composition and the nature of a given entity (e.g. standard InChI strings or heavy atom count), and those that capture empirically or theoretically derived data that varies depending on the circumstances of the chemical entity or data observation (e.g. solubility, computed logP). In CHESS, the

two cases are specified through a common approach stipulated in the CHEMINF ontology with one major difference: non-constant descriptors are assigned to a *chemical configuration* that reflects upon the circumstances of descriptor creation and the circumstances of the entity to which these descriptors refer. For instance, the calculated free energy of formation of a molecule in gas phase depends not only upon the computational package employed to derive the value and parameters like temperature or the level of theory used, but also on the geometric configuration of the molecule.

Unlike chemical entities, it is not absolutely crucial that descriptors have canonical and reproducible names, as they are rarely used as focal points around which other annotations are integrated. That is, although it is important to ensure that descriptor identifiers do not clash, thus overwriting or contradicting any data previously assigned to a given descriptor, the precise form of descriptor identifier and its canonical nature are of secondary importance. In this work, invariant descriptors receive unique and canonical identifiers, based only on a hash of essential descriptor components: value, uncertainty (if any), and units. Variable descriptors, on the other hand, derive their identifiers from these parameters in addition to the canonical identifier of their corresponding chemical configuration, itself derived by hashing values of its annotations, presented in lexicographical order. Multi-component descriptors that include other descriptors, such as the non-invariant positional descriptor for atoms, receive an identifier that is a hash of the canonical identifiers of constituent descriptors in lexicographical order instead of a direct hash on the value, uncertainty, and unit annotations. In both cases, the resultant hash is appended to the identifier of the entity to which these descriptors refer in order to obtain the final identifier.

The scheme for specifying descriptors is quite simple (Figure 2-6). Descriptors may contain other descriptors and may have an unlimited number of annotations, such as units and uncertainty, but must have a value assigned. Furthermore, CHESS follows the CHEMINF approach to modeling computationally-derived descriptor provenance. Thus, a chemical descriptor is an information content entity that is a specified output of an algorithm (e.g. Mannhold logP algorithm [32]), and an output of a parameterized or non-parameterized software execution process. This process has to be annotated with the identity of the software agent employed, which may be further annotated with e.g. software version, and any parameters (formally defined in an ontology) used in carrying out the calculation. Experimentally-derived descriptors follow a similar scheme, but refer to experimental observations and processes.

Figure 2-6. A simplified descriptor specification, as per the CHEMINF ontology approach.

In addition to this specification, descriptors are also annotated with a chemical configuration. This concept is useful not only as a nodal point for storing source and other provenance information, but also in uniting the descriptors that have been derived under a uniform set of conditions and for the system under investigation that is in a given,

well-defined state. Because chemical configurations are also, in part, specific to a geometric configuration of a given entity, it also allows a non-confounded aggregation of data, e.g. in representing atomic coordinates for multiple conformations of the same molecular entity or investigating the thermodynamic properties of molecules that may have different electronic configurations (e.g. singlet or triplet oxygen) or molecules that may be exposed to different temperatures. This approach also does away with data retrieval and integration complexities arising from heterogeneously derived information attached to a single chemical entity in certain databases, manifested in the requirement for extensive involvement of a human expert to identify a set of descriptors suitable for a particular comparison. So long as chemical information is properly annotated, descriptors specified with our approach (Listing 5) are readily amenable to facile querying and retrieval (Listing 6), significantly reducing the workload on an individual researcher. Alternatively, querying the knowledgebase for all descriptors from different databases that refer to the same geometric configuration or experimental conditions, is also possible.

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix sio: <http://semanticscience.org/resource/SIO_>
@prefix chess: <http://semanticscience.org/resource/CHESS_>
@prefix chebi: <http://purl.org/obo/owl/CHEBI#>

#Specify an oxygen atom in ethanol.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N rdf:type chebi:CHEBI_23367.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-A3O rdf:type chebi:CHEBI_25805.
chess:LFQSCWFLJHTTHZ-UHFFFAOYSA-N sio:SIO_000053 :LFQSCWFLJHTTHZ-
UHFFFAOYSA-N-AO3.

#Three-dimensional coordinate specification.
:DCCC rdf:type sio:3D Cartesian coordinate.
:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-A3O sio:'has attribute' :DCCC.
:DXXX rdf:type sio:'x Cartesian coordinate'.
:DCCC sio:'has direct part' :DXXX.
:DYYY rdf:type sio:'y Cartesian coordinate'.
:DCCC sio:'has direct part' :DYYY.
:DZZZ rdf:type sio:'z Cartesian coordinate'.
:DCCC sio:'has direct part' :DZZZ.

#Define the x Cartesian coordinate.
:DXXX sio:'has value' "1.55" .
:DXXX sio:'has unit' <http://purl.org/obo/owl/UO#UO_0000019> .

#Define a chemical configuration.
:CCXXX rdf:type sio:000659.
:CCXXX sio:'has provider' "http://pubchem.ncbi.nlm.nih.gov" .

#Link the chemical configuration to the 3D Cartesian coordinate.
:DCCC sio:has attribute :CCXXX.

```

Listing 2-5. Sample specification of variable descriptors with reference to a chemical configuration, amended for readability.

```

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix sio: <http://semanticscience.org/resource/SIO_>
prefix : <http://semanticscience.org/resource/CHESS_>
prefix chebi: <http://purl.org/obo/owl/CHEBI#>

select ?a, ?x, ?y, ?z where {
?d    rdf:type          sio:'3D Cartesian coordinate'.
?d    sio:'has provider' "http://pubchem.ncbi.nlm.nih.gov" .
?a    rdf:type          chebi:atom.
?a    sio:'has attribute' ?d.
:LFGSCWFLJHTTHZ-UHFFFAOYSA-N sio:'has proper part' ?a.
?d    sio:'has direct part' ?xx.
?xx   sio:'has value'    ?x.
?d    sio:'has direct part' ?yy.
?yy   sio:'has value'    ?y.
?d    sio:'has direct part' ?zz.
?zz   sio:'has value'    ?z.
}

```

Listing 2-6. Sample query of variable descriptors that would retrieve coordinate information for all atoms of ethanol that originate from PubChem, amended for readability.

2.3.5 Chemical Information Integration

Having described at length the representation of chemical information and various entities, let us consider the practical effects of decisions taken in CHESS on information integration. Although we have generated and successfully integrated moderately-sized subsections of various publically accessible chemical databases, the overall effect of our specification can be illustrated on the example of a limited set of molecules present in multiple databases or having descriptors created by different computational procedures. To demonstrate the facile cross-database and cross-study information integration afforded by consistent canonical entity identifiers, consider two instances of the same compound, antidepressant melitracene, in two different databases, PubChem and ChEMBL. Although these entries are cross-linked in their respective databases, the comparison and integration of information regarding this entity involves a procedure that requires a degree of human involvement, especially if this entry has to be further cross-linked to

another entry in any number of other databases, each having unique data fields or approaches to data presentation and representation.

With CHESS, it is possible to independently encode the information in each repository in separate RDF graphs. However, because chemical entity URIs for melitracene are the same in all cases (http://semanticscience.org/resource/CHESS_GWWLWDURRGNSRS-UHFFFAOYSA-N), all of these graphs effectively collapse into a single graph, with a SPARQL query to retrieve information relevant to melitracene capable of seamlessly drawing on the entirety of the chemical knowledge, without regards to originating database- or software-specific integration barriers. At the same time, this allows us to address issues relating to data correspondence, as bonds with the same URI are assured to be the same entity. This eliminates the need for substructure matching or other intermediate steps in carrying out cross-database comparisons. This ability may be especially useful in cases where, for example, multiple computational experiments are performed and the computed bond lengths need to be compared to the experimentally observed bond lengths or to the results of other computational experiments.

As a simple demonstration of this capability, we have generated a small set of 90 CHESS-encoded antidepressants containing selected information from two databases, as well as three different computational packages. The descriptors we have represented in this knowledgebase were of relevance to satisfying Lipinski's Rule of Five [33], allowing us to potentially pool all of the available information in these disparate databases and computational experiments to address the question of whether the compounds in our knowledgebase were, in fact, drug-like. For each source of information, a number of descriptors were intentionally left out to demonstrate the assurance of information

complementarity and preservation of information correspondence with the CHESS representation. To test this, we have created a custom class with a formal definition corresponding to Lipinski's Rule of Five and reasoned over it using the Pellet reasoner plugin (version 1.4) [34] in Protégé software [35] (version 4.0, build 115). As this class was correctly populated with 88 instances of chemical entities satisfying the Rule of Five, it became apparent that consistent molecular identifiers permitted the effective collapse of the multiple knowledge sources to provide the information necessary in order to fulfil the classification and identify drug-like chemical entities. Furthermore, the correspondence of the molecular descriptors specified in our knowledgebase was adequately preserved.

```
'chemical entity' and 'has attribute'  
  some ('mass descriptor' that 'has value' some double[<500.0])  
and 'has attribute'  
  some ('hydrogen bond acceptor count' that 'has value' some int[<=10])  
and 'has attribute'  
  some ('hydrogen bond donor count' that 'has value' some int[<=5])  
and 'has attribute'  
  some ('logP descriptor' that 'has value' some double[>-5.0,<5.0])
```

Listing 7. The formal axiomatic definition of a class of chemical entities that satisfy Lipinski's Rule of Five, using CHEMINF concepts (specified using the Manchester OWL syntax).

2.3.6 Variable-Level Granularity Semantically Enriched Annotations and Queries

We have established that CHESS addresses the problem of seamless data integration across multiple sources of information by demonstrating a query that reproduces a common Rule of Five filter, on an integrated data set from three different sources. However, this kind of filtering can currently be readily carried out for most chemical databases through their search engine interfaces provided by the suppliers of such databases, such as PubChem. In the absence of such exposed search interfaces, however (e.g. when there is no option to restrict search results based on molecular mass), users of

these chemical information repositories are faced with the task of manually parsing or calculating the chemical information that is needed for answering their research question. In addition to this, annotations of existing entities with new information in smaller studies are often lost due to the database barriers discussed at length in this work, or the practicality of publishing ‘smaller’ scientific data as an accessible database that is open to querying. Finally, the level of annotation granularity allowed in a given database may be insufficient for a particular application. That is, while data on individual atoms and bonds certainly exists in PubChem, it is impossible to refer to, retrieve, or annotate these individual entities, at least not in a manner that would immediately meaningfully connect the annotations generated as a part of a given study with a given PubChem entity and would allow these annotations to be discovered and queried.

In contrast, CHESS is flexible and extensible: annotations and information represented in CHESS is assured to be searchable, no matter what information the database vendor feels like exposing. So long as there is chemical information represented in CHESS, it is subject to logical queries and semantic integration. Furthermore, all chemical entities, such as atoms, bonds, and molecules, can be fully annotated using a range of vocabularies, and these annotations can be linked directly to the entities being annotated, even if the original entity and its annotations reside in different RDF graphs. With CHESS, it is now possible to facilitate open publishing of scientific information and assure that the precious scientific knowledge is preserved, no matter how small a study has been carried out.

To demonstrate the benefits of this approach, let us consider a practical case that is impossible to address with currently existing chemical databases. For this, let us examine

phenolic antioxidants, which constitute an important class of molecules that are used in industrial processes and as nutritional supplements to help alleviate the damaging effects of free radicals, such as lipid peroxidation in oils. It has been shown that Bond Dissociation Enthalpies (BDEs) of the phenolic O-H bonds can be used as excellent predictors of the antioxidant potency of a phenolic compound. When designing phenolic antioxidants for biological systems (e.g. humans), one has to be mindful that the BDE of the phenolic O-H bond has to be higher than that of the weakest O-H bond in ascorbate (67 kcal/mol) to allow for biological antioxidant recycling, but lower than that of the O-H bond of α -tocopherol (78 kcal/mol) to surpass the potency of existing physiological antioxidant defences [41]. While the accurate annotation and searching of bond-level information is currently impossible in major chemical information repositories, we have developed a demonstrative set of chemical entities with O-H BDE information annotation (available from our companion website [23]), including the computational method, software, and some of the parameters used to compute this information for ethanol (Listing 8).

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix sio: <http://semanticscience.org/resource/SIO_>
@prefix : <http://semanticscience.org/resource/CHESS_>
@prefix chebi: <http://purl.org/obo/owl/CHEBI#>
@prefix cheminf: <http://semanticscience.org/resource/CHEMINF_>

#Specify an OH bond in ethanol.
:LFQSCWFLJHTTHZ-UHFFFAOYSA-N rdf:type chebi:CHEBI_23367.
:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-BOH7 rdf:type sio:SIO_010498.
:LFQSCWFLJHTTHZ-UHFFFAOYSA-N sio:SIO_000053 :LFQSCWFLJHTTHZ-UHFFFAOYSA-
N-BOH7.

#BDE Annotation Specification.
:DBDE rdf:type cheminf:000252. #BDE Descriptor.
:LFQSCWFLJHTTHZ-UHFFFAOYSA-N-B)H7 sio:'has attribute' :DBDE.
:DBDE sio:'has value' "104.5" .
:DBDE sio:'has unit' "kcal/mol" .

#BDE Provenance Specification.
:DBDE cheminf:'is specified output of' :AM1 .
:AM1 rdf:type cheminf:000144 .
:DBDE cheminf:'is output of' :ExecutionXYZ .
#Parameterized software execution.
:ExecutionXYZ rdf:type cheminf:000147.
:ExecutionXYZ sio:'has attribute' :TempX #Temperature parameter.
:TempX sio:'has value' "298" .
:TempX sio:'has unit' "<http://purl.org/obo/owl/UO#UO_0000012">

#Software Package Used
:ExecutionXYZ sio:'has participant' :MOPACV7 .
:MOPACV7 rdf:type sio:'software application' .
:MOPACV7 sio:'has attribute' :VMOPAC7 .
:VMOPAC7 rdf:type sio:'software version' .
:VMOPAC7 sio:'has value' "7.1.11" .

```

Listing 2-8. A representative portion of the CHESS specification of a ethanol, its constituent OH bond, and the parameters used in BDE calculation for this bond. Please note that this representation was simplified and modified to improve readability.

A knowledgebase with such information could have been published and shared as an outcome of any of a number of studies on this subject, with annotations attached directly to existing chemical entities in one of the major chemical databases. Instead, this information was sealed away in a series of PDF and HTML documents, accessible only to those with the time and resources to locate and read them. In order to retrieve *phenolic*

antioxidants with potential applications in biological systems, one could combine the molecular structural features (the presence of a phenol group) with thermochemical annotation information on the OH bond. Furthermore, since computationally derived thermochemical parameters are best compared when the method, software, and parameters used to derive them are uniform, we may include these requirements in our query to retrieve uniform and useable information. Such an integrative query will provide all the potentially potent novel phenolic antioxidants from a chemical knowledgebase (Listing 9). Please note that such integrative, variable-granularity, and flexible queries are impossible for the major conventional chemical information repositories. Truly, the imagination of CHESS users is the limit for the expressivity and the semantic enrichment of the represented chemical information.

```

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix sio: <http://semanticscience.org/resource/SIO_>
prefix : <http://semanticscience.org/resource/CHES_>
prefix chebi: <http://purl.org/obo/owl/CHEBI#>

select ?m, ?value where {
?m   rdf:type          chebi:CHEBI_23367.    #A molecule
?m   sio:'has proper part' ?f .              #that has a moiety
?f   rdf:type          :FGPhenol            #which is phenol.

?m   sio:'has proper part' ?b .              #This molecule
?b   rdf:type sio:010498.                    #also has a single bond
?b   sio:'has proper part' ?o.               #which has atoms O and H.
?o   rdf:type chebi:25805.
?b   sio:'has proper part' ?h.
?h   rdf:type chebi:49637.

?b   sio:'has attribute' ?d.                 #This bond has
?d   rdf:type cheminf:000252.                #a BDE descriptor
?d   cheminf:'is specified output of' :AM1.  #computed with AM1
?d   sio:'has unit'    "kcal/mol".           #in units of kcal/mol
?d   sio:'has value'   ?value .              #and with a value

FILTER(?value > 67 && ?value < 78)          #between 78 and 67.
}

```

Listing 2-9. Sample query that may be carried out across a single or multiple SPARQL endpoints to retrieve all potentially promising phenolic antioxidants from an annotated compound collection, amended for readability.

2.3.7 Representing and Querying Reactive Transformations

As a final demonstration of usability of CHES- encoded chemical information, let us consider reaction representation and querying. As mentioned earlier, reactions in CHES are considered to be chemical entities, whose identifiers come from their participants and stoichiometry. Though reactions may or may not take place under a certain range of circumstances, we model them as abstract processes that exist without regards to their likelihood of occurring. Because as long as a specified reaction respects universal principles, such as the conservation of matter and energy, any reaction may happen at any time (some are more likely than others), we maintain that the sole features relevant to

unique reaction specification are the chemical identities and stoichiometries of the participating chemical entities (Figure 2-3).

In general terms, there are two types of reactive transformations specified: those involving generic chemical moieties and their transformations and those that involve precisely defined chemical entities. Both cases may be represented in CHESS in a uniform and consistent fashion. The only difference between the two cases is the specification of the entities involved, which in the former case are of the functional group/substructure type, and in the latter are instantiated molecular entities. In either case, the most important feature that decides whether a molecule will undergo reactive conversion is the presence of a characteristic functional group within the substrate molecule. That is, reactions can be viewed as transformations between the various isolated regions, or functional groups, with the rest of the molecule attenuating or increasing reactivity. The presence of the requisite functional group through explicit mapping of constituent atoms and bonds makes it possible to infer the role of a given molecule and its components in a reactive process (Figure 2-7).

Figure 2-7. Mapping components of an instantiated molecule to reactive transformation definition participants can enable a range of inferences relevant to predicting reaction outcome.

The amount of information regarding a reactive transformation that could be inferred is entirely dependent upon the amount of information present in the reference knowledgebase and in the reaction specification RDF graph. For a demonstrative example, let us consider a transformation of primary alcohols to aldehydes, which could be catalyzed by an inorganic catalyst or by an enzyme. Since generic functional group transformations are involved, reaction specification would involve functional group types and atom types, rather than corresponding instances. Substrates are linked to the reaction using *has input* (SIO_000230), products using *has product* (SIO_000312) and catalysts using *has agent* (SIO_000139). Further, to maintain a record of the correspondence of every transforming entity, the transformations would be specified with *transforms into* (SIO_000655), which can operate upon whole molecules, functional groups, and atoms (Listing 10).

```
@prefix      : <http://semanticscience.org/resource/CHESS_>
@prefix  sio: <http://semanticscience.org/resource/SIO_>
@prefix  chebi: < http://purl.org/obo/owl/CHEBI#>

#Reaction inputs, outputs, and agents.
:RXXX      rdf:type                sio:'chemical reaction'.
:RXXX      sio:'has input'         :FGXXX.
:RXXX      sio:'has input'         :FGXXXA3O.
:RXXX      sio:'has product'       :FGYYY.
:RXXX      sio:'has product'       :FGYYYA3O.
:RXXX      sio:'has agent'         :KJFCCLURYALNSL-UHFFFAOYSA-N. #Ru/Al2O3

#Transformations of functional groups.
:FGXXX      rdf:type                chebi:'organic group'.
:FGYYY      rdf:type                chebi:'organic group'.
:FGXXX      sio:'transforms into'   :FGYYY.

#Transformations of atom types in these groups.
:FGXXXA3O    rdf:type                chebi:'oxygen atom'.
:FGXXX      sio:'has proper part'   :FGXXXA3O.
:FGYYYA3O    rdf:type                chebi:'oxygen atom'.
:FGYYY      sio:'has proper part'   :FGYYYA3O.
:FGXXXA3O    sio:'transforms into'   :FGYYYA3O.
```

Listing 2-10. Sample reaction definition of CCO functional group transformed into CC=O functional group with CHESS in N3-turtle, amended for readability.

To enable reaction matching, it is necessary to also obtain the atom and functional group typing information for the chemical entities in the CHESS knowledgebase. For this purpose, one may use an extension of the described atom-centric fingerprinting procedure by typing atom and functional group instances to the generic classes of atoms and functional groups they instantiate. Thus, instead of merely asserting membership of group LFQSCWFLJHTTHZ-UHFFFAOYSA-NFGXYZ in the ethanol molecule (where to save space, XYZ is the appropriate hash), it would be typed to the FGXXX functional group. A similar procedure has to be carried out for atoms if atom-centric searches need to be enabled, e.g. for tracing the precise passing of atoms between molecules, and checking the metabolic history of every single atom (Figure 2-8). As mentioned earlier, this process may be carried on at any point, and the knowledgebase may be amended with the required information.

Figure 2-8. A modification of molecular fingerprinting used for the complete description of chemical structures in terms of an exhaustive list of functional groups. Molecular fragments resulting from fingerprinting (G1, G2, G3) may be stored and treated as descriptive functional groups, along with user-submitted ones.

The resultant specification of reactions and chemical entities enables simple, yet powerful SPARQL-based queries, e.g. to find the potential reactions of a given molecule and identify the atoms and functional group instances involved in these reactions (Listing 11).

```
prefix      : <http://semanticscience.org/resource/CHESS_>
prefix  sio: <http://semanticscience.org/ontology/sio.owl#>
prefix  chebi: < http://purl.org/obo/owl/CHEBI#>

select * where {
?reaction rdf:type sio:'chemical reaction' .
?reaction sio:'has input' ?FGAtom .
?FGAtom  rdf:type chebi:'atom' .
?MAAtom  rdf:type ?FGAtom .
?MAAtom  sio:'is part of' :LFQSCWFLJHTTHZ-UHFFFAOYSA-N
}
```

Listing 2-11. A query that will return all the reactions that ethanol may potentially be involved in, amended for readability.

2.3.8 Supporting Tools and Interfaces

An evolving project site with supporting information, graphical user interface-backed tools, sample RDF specifications and data sets (including a sample set of ~17000 ChEBI compounds), as well as further information, is available [23]. The implementation of CHESS provided on the website has a scope limited to compounds that can be well-represented with current version of InChI. Therefore, we currently exclude polymers with repeating units of arbitrary length and Markush structures from the scope of CHESS, for example. On the other hand, organometallic compounds and metals that can be represented with InChI are included within our scope. While there is no theoretical limit to the size of a compound that can in principle be represented using CHESS, our sample implementation is practically limited (by the available computational resources on the server) to lower molecular weight compounds.

2.4 Conclusions

Unfortunately, many of the large chemical databases currently do not possess the means of chemical data integration and federation. Either for historical reasons or for efficiency improvement, a large number of these databases have been purpose-built for capturing data within a particular domain, and without much consideration of trans-domain knowledge aggregation. This further complicates the task of database integration and poses as an obstacle to productive chemical research. Fortunately however, the Semantic Web provides an excellent opportunity for significantly simplifying this problem with the appropriate data representation and sufficiently advanced data conversion tools.

In this work, our principal goal has been to present a novel chemical representation formalism that draws on the Semantic Web principles. We have attempted to make a

compelling case for a universal semantic specification of chemical entities in cheminformatics by demonstrating the power, integrative capacity, and the flexibility of representation afforded by fully embracing Semantic Web technologies. By adopting consistent and canonical identifiers for every aspect of chemical entities identified here, we have demonstrated facile cross-domain chemical knowledge integration while preserving correct data correspondence and explicit data provenance information. Furthermore, we have demonstrated the power of CHESS in enabling integrative chemical research that draws on the entirety of chemical information available on the Web. While we do not believe that any specification can natively address outright errors in databases, CHESS representations allow us to explicitly and formally define the meaning of our data and enable machine agents to automatically reason over this data, checking it for consistency and completeness. This, in turn, enables a more accurate scientific discourse and a more reproducible and transparent way of doing science.

We have also demonstrated mechanisms by which chemical configuration-specific information may be encoded without loss of inter-configuration information aggregation and without introducing intra-configuration information mixing that is sometimes an unfortunate occurrence in traditional databases. For example, atomic coordinate information and heat of formation data may exist for multiple conformers of a single molecule, with every conformer annotated with the appropriate data. In principle, this chemical configuration concept allows one to aggregate information about the various electronic states of a given molecule when CHESS will be extended to include the explicit specification of electrons and related concepts. While CHESS does not aim to make statements with respect to the preferred chemical configuration of a given

compound under a given set of conditions, CHESS allows the unambiguous and explicit identification of precise chemical configurations for the purpose of advancing and facilitating interdisciplinary scientific discourse.

We believe that outsourcing of chemical information integration to machine agents is of increasing importance as the rapidly growing collection of diverse chemical information already available on the web is overwhelming human integrative capacity. If no steps are taken to create, standardize, and adopt a set of consistent standard Semantic Web chemical information exchange ontologies and representation formalisms soon, we are risking missing yet another opportunity to truly federate the chemical web and trigger a transition to a new era of chemical research. Truly, the infinite chemical space is full of mysteries, marvels, and opportunities – and we believe that it is only through the concerted and unified efforts of researchers in all fields of science, enabled by Semantic Web technologies, that we may hope to one day chart it.

More importantly to this dissertation, we would like to conclude that by allowing the representation and annotation of every aspect of chemical entities and reactions, CHESS answers the representational needs of our prototype metabolic modeling framework. As such, CHESS forms the representational backbone of the entire framework.

2.5 Methods

2.5.1 Supporting Ontologies

CHEMINF is a collaboratively developed Web Ontology Language (OWL) [36] ontology for representing chemical information and descriptors, freely available to the broad cheminformatics community. Semanticscience Integrated Ontology (SIO) is an general ontology that provides over a 150 object relations and over 900 classes of various

entities, including physical, processual, and informational ones. There is only one data property in the ontology, ‘has value’, with the other relations describing aspects of mereology, spatial positioning, temporal ordering, qualities, attributes, representations, participation, and agency. In addition to these ontologies, we use the CHEBI ontology for chemical entities and concepts.

2.5.2 Triplification of Chemical Information

The encoding of chemical information into its CHEMA form was carried out with software we developed (sample source code available on the companion website [23]), based on the Jena API [37] and the Chemistry Development Kit [38]. A sample dataset of 90 antidepressants was obtained from a mesh-based keyword search on ‘antidepressant’ over the PubChem database. Chemical descriptors were computed using CDK and the Open Babel Java API [39]. Unique identifiers were obtained by hashing pertinent chemical data as a 40-letter SHA-1 hash using the Java Security API.

2.5.3 Query Testing

In order to test our chemical similarity queries, we used OpenLink Virtuoso version 6.01.3126 SPARQL endpoint [40]. To test DL Rule-based queries, we used the Pellet reasoner version 2.2 [42], running on a single CPU core. The machine used for these tests had dual Intel Xeon CPU at 2.9GHz, with 32GB of RAM.

2.5.4 Publically-Accessible Web Interface

We have created a publically-accessible web interface that generates CHEMA RDF descriptions of chemical entities using the methodology described in section 2.3.2 for the triplification of chemical information and the JME Molecular Editor for the graphical input of chemical structures. The interface is available at

<http://semanticscience.org/projects/chess/tools.php> . A working knowledge of SMILES strings is highly recommended for users of this interface.

2.6 Authors' Contributions

LLC implemented supporting tools, generated triple stores, and ran tests. LLC and MD created the supporting ontologies. MD provided guidance and ran initial tests.

3 Chapter: Chemical Entity Semantic Classification

Parts of this work dealing with functional classification of small molecules are based on the following conference paper.

Chepelev LL, Klassen D, Dumontier M. Hazard Estimation and Method Comparison with OWL-Encoded Toxicity Decision Trees. OWL Experiences and Directions. June 5-6, 2011. San Francisco, the United States of America.

3.1 Abstract

The advent of high-throughput experimentation in biochemistry has led to the generation of vast amounts of chemical data, necessitating the development of novel analysis, characterization, and cataloguing techniques and tools. Recently, a movement to publically release such data has advanced biochemical structure-activity relationship research, while providing new challenges, the biggest being the curation, annotation, and classification of this information to facilitate useful biochemical pattern analysis. Unfortunately, the human resources currently employed by the organizations supporting these efforts (e.g. ChEBI) are expanding linearly, while new useful scientific information is being released in a seemingly exponential fashion. Compounding this, currently existing chemical classification and annotation systems are not amenable to automated classification, formal and transparent chemical class definition axiomatization, facile class redefinition, or novel class integration, thus further limiting chemical ontology growth by necessitating human involvement in curation. Clearly, there is a need for the automation of this process, especially for novel chemical entities of biological interest. To address this, we present a formal framework based on Semantic Web technologies for the automatic development of chemical ontologies which can be used for automated classification of novel entities. We demonstrate the automatic self-assembly of a structure-based chemical ontology based on 60 MeSH and 40 ChEBI chemical classes. This ontology is then used to classify 200 compounds with an accuracy of 92.7%. We extend these structure-based classes with molecular feature information and demonstrate the utility of our framework for classification of functionally relevant chemicals. Finally, we discuss an iterative approach that we envision for future biochemical ontology

development. We anticipate that the proposed methodology shall ease the burden of chemical data annotators and dramatically increase their productivity. We anticipate that the use of formal logic in our proposed framework will make chemical classification criteria more transparent to humans and machines alike and will thus facilitate predictive and integrative bioactivity model development.

3.2 Introduction

Over the hundreds of years of biochemical research, humanity has encountered myriads of chemical entities with countless combinations of functional groups that imparted upon their bearers distinct reactivities and properties. Grouping these entities into structure- and property-based classes within a larger chemical ontology (formal logical specification of a chemical hierarchy conceptualization) may enable the recapitulation of or improvements in the prediction of their biological functionality and chemical reactivity patterns, thus providing indispensable assistance in understanding the molecular nature of metabolism, toxicity, and bioactivity. In addition to this, the assignment of chemical entities to a given class within a chemical ontology or hierarchy may facilitate the inference of the potential anticipated roles and properties of these entities.

3.2.1 Structural Classification

Despite the numerous efforts to develop automated chemical classification and ontology construction approaches, this process has so far practically remained firmly within the hands of human curators, as exemplified by perhaps the largest chemical ontology constructed to date – the Chemical Entities of Biological Interest (ChEBI) database and ontology [1]. As more chemical data becomes readily available from large-scale experimentation, ChEBI and numerous similar repositories of biological and chemical information (e.g. [2,3]) are increasingly coming under strain due to lack of human curators, the overwhelming wealth and diversity of the new chemical information, and the boundlessness of the chemical space. Clearly, a means to automate and facilitate these classification efforts is urgently needed if we desire to understand, expose, and integrate the entirety of the available chemical information.

Manual chemical classification efforts frequently involve the tedious work of multiple human experts who carefully peruse the relevant chemical literature to gain competence in assigning chemical entities to a range of classes [1]. Although the reliance of biochemical activity-based classification efforts on the predictions or informed estimates of human experts has been largely relegated to the informal or speculative domain of scientific communications thanks to the development of numerous quantitative structure-activity relationship (QSAR) elucidation approaches, chemical hierarchy construction is still very much a human-dominated domain. Chemical hierarchies in common use, such as ChEBI and the Medical Subject Headings (MeSH) organic chemistry hierarchy [4], are structure-based but often have the capacity to incorporate chemical classes that are based on functional chemical attributes, such as those bearing a certain pharmaceutical activity (e.g. ChEBI antibiotics). Although structure- and function-based classification schemes involve somewhat different approaches for construction, the basic accepted SAR dogma is that chemical structure and chemical function are inseparably linked [5]. Therefore, structure-based classification efforts can be useful in initiating the construction of predictive models of small molecule bioactivity.

The correlation and overlap between chemical topology-based and biological function-based chemical hierarchies is expected since the vast majority of biological activity is enacted through interactions and transformations that are specific to a particular set of structural features of small molecules. For instance, the majority of enzymes operate on a very limited set of chemical structures: carbonyl reductases, operate on carbonyl and alcohol functional groups [6] while bacterial transpeptidases bind to and operate upon structures that resemble the peptide bond [7], leading to the potency of beta lactam-based

drugs in the inhibition of their operation. While there may be more than one 2D chemical structure that results in an interaction with a given enzyme (due to its 3D configuration), we claim that there is a finite set of 2D chemical skeletons that result in a bioactive interaction, and that given a sufficiently large dataset of compounds known to be bioactive, it is possible to characterize a class of functionally active compounds as a collection of classes with well-defined, consistent structural features. Concerns for the predictive capacity of such classifications, like the distinction between stereoisomers and a range of chemical properties such as molecular volume that may result in disqualifying an otherwise well-suited 2D skeleton from a bioactivity-inducing interaction can be laid to rest by explicitly incorporating these features into class definition. In this work, we demonstrate that given the logical expressivity afforded by the Web Ontology Language (OWL), we can automatically create formal logical definitions for a range of topology and property-based chemical classes which can then self-assemble into a chemical ontology.

While manual literature searching and experimental data analysis is often involved in the classification of chemical entities into functional classes [1], structure-based hierarchy construction is much more well-defined and easily amenable to automation. In constructing structure-based chemical classes, a modern curator would familiarize themselves with any existing informal class definitions, as well as any known chemical class members. Based on this expertise, the curator would then define a class, create a short textual or pictorial description of this class, and add child/parent relationships to the other classes, as well as including several representative entities. In essence, this approach requires the resulting ontology to consist of a formalization of expert domain

knowledge in the area, while providing no means for this knowledge to be captured in a machine-understandable format. The result of this is a lack of facile re-use of the work of human curators in classification and annotation of novel entities and establishing child/parent relationships for newly added classes when the ontology grows or is modified. Unfortunately, the majority of chemical ontologies, including ChEBI, MeSH (chemical hierarchy), and LIPIDMAPS [2], suffer from this, resulting in artificial discipline-specific barriers to research, since none of these can be easily or automatically integrated. As the chemical ontologies grow in size and the range of relationships covered, manual ontology development becomes increasingly difficult due to the multiplying number of consistency checks that have to be performed for each new class and relationship added.

To date, automated chemical classification has mostly belonged to the domain of QSAR studies which involve the characterization of a training set of chemical entities with respect to a particular set of computable or measurable molecular features, as well as the property of interest, such as biochemical activity or biomedical potency [6]. The features that are used for training are often selected such that the cost of their acquisition is significantly lower than that of the modeled property. This feature set is then used to construct a predictive model using one of many available mathematical procedures, including the creation of decision trees, multidimensional parametric fits, or three-dimensional pharmacophore models, to name a few. The resultant models are then verified for their capacity to adequately classify molecules into a range of classes of interest using a test set of compounds not used in model creation. While the resultant models may achieve a great degree of accuracy, the vast majority of such models are not

readily amenable to human understanding or automated incorporation into a chemical hierarchy. In order to fully appreciate the structural and physicochemical molecular features of relevance to a particular functional or structural classification, post-processing of the resultant mathematical models is often required. Compounding this problem is the fact that chemical class definitions derived using approaches such as support-vector machine training are often inherently probabilistic – that is, no class membership requirement is formally stated as a necessary or sufficient condition. In order to derive a truly self-organizing chemical ontology, a deterministic classification in terms of well-defined features and their combinations is necessary.

While function-based chemical hierarchy construction has not been adequately addressed, methods to generate structure-based chemical ontologies have been studied in the past to some extent. For example in [8], a very early attempt at formalizing chemical classes and their relationships for chemical inference has been carried out, but never practically implemented for any large-scale classification exercise. In [9], a chemical ontology has been manually constructed with classes formally defined as containing molecules that possessed a particular functional group or a set of functional groups. While this work has been successfully applied for practical chemical functionality analysis through chemical semantic similarity identification enrichment, the basic ontology construction still followed a manual process and was limited to a set of pre-defined functional groups, insufficient for characterizing a number of more complex chemical classes. Furthermore, the lack of a formal, accessible framework for chemical classification (in the form of OWL) has resulted in limited logical chemical class axiom expressivity, therefore limiting the applicability of the approach for characterizing some chemical classes (e.g.

those requiring cardinality restrictions, such as dienes). Finally, the work of [10] has resulted in a structure-based chemical ontology that employed combinations of chemical substructures of increasing complexity to produce increasingly specialized chemical classes, with the aim of improving the structural analysis and pattern identification in sets of biologically active compounds.

Unfortunately, a framework or approach for fully automated construction of chemical ontologies has never been achieved. In this work, we propose a radically novel approach for structure-based chemical classification automation by abstracting and automating the majority of chemical curation to construct machine- and human-understandable, formally axiomatized chemical class definitions in OWL. Further, we use these automatically generated chemical classes to enable machine reasoning over the ontology and demonstrate ontology self-organization, as well as classification of individual chemical entities represented using the Chemical Entity Semantic Specification formalism. Finally, we demonstrate that certain functionally-relevant classes can be defined unambiguously using our approach, and integrated into the overall chemical ontology seamlessly.

3.2.2 Functional Classification

Our industrialized society relies on millions of diverse chemical entities in applications as broad as energy production, combating disease, and manufacturing. As novel chemicals are developed and as industrial processes evolve, we become heavily exposed to an increasingly diverse pool of environmental pollutants and their poorly characterized by-products. The resource commitment necessary to fully characterize the toxicity of even a single chemical entity experimentally is very substantial. Since the pool of chemicals in need of routine toxicity screening by organizations such as environmental protection

agencies and pharmaceutical companies is practically infinite and the resources for this task are often scarce, alternative means of toxicity screening are often applied to prioritize compound screening or alert chemical researchers to the potential adverse effects of their molecule of interest, especially in the early stages of compound development.

Such predictive *in silico* approaches may be broadly characterized into two major categories: data-driven systems and expert systems [11]. Data-driven systems involve the generation of mathematical models (regression, neural network, or any other method) to correlate computed or observed physicochemical molecular properties to their experimentally obtained functional characteristics, such as toxicity, binding affinity for a given enzyme, or biological activity of a given type. The result of data-driven systems are QSAR that are specific to the class of compounds represented in the training set and are often difficult to logically interpret or integrate even for a human operator.

Expert-based systems, on the other hand, strive to capture the knowledge of human toxicology experts into machine-readable models with the aim of automating chemical classification and chemical information analysis. Expert-based systems can take a number of forms, among which rule-based and decision tree-based systems are quite prominent. Rule-based systems rely on the formulation of a number of independent rules that can be integrated to construct a logical conclusion about the toxicity or activity of a given compound. Decision tree-based systems involve the sequential execution of a series of logical tests, with each branch point of the tree containing a logical test, and each leading either to a final classification, or a deferral to further tests (Fig. 1).

Figure 3-1. A simple toxicity decision tree: at each branching point, a rule is evaluated, and based on the outcome of this rule, either a final activity decision is made, or judgment is deferred to another node.

Since their introduction four decades ago [12], decision tree-based toxicity and activity prediction systems have gained acceptance by academics and industrial researchers alike, finding applications in predicting molecular properties such as mutagenicity, toxicity, and skin sensitization among others [13]. Furthermore, automated objective methods have appeared to emulate the work of human experts by creating decision trees in which rules and tree structures are drawn based on the analysis of empirical toxicity data [14]. Aside from simplifying and automating classification efforts, and unlike data-driven toxicology prediction systems, decision tree-encoded expert knowledge is understandable to humans and machines alike. Unfortunately, the potential of OWL ontologies to formally capture and enact such expert-based decisions in chemical toxicology and many other fields has not yet been fully realized. Consequently, the decision frameworks and the supporting databases for making such decisions are still largely fragmented along discipline, software, and institutional divides. Since biological and chemical information is

increasingly standardized and integrated into the Semantic Web through initiatives like Bio2RDF [15], we find ourselves at the point where OWL-based formalization of expert rule bases and decision trees, combined with ready access to vast amounts of linked data can yield unprecedented, tangible benefits in integrated bioactivity and toxicity prediction and predictive method comparison and integration.

In this work, we demonstrate the automated generation of biologically relevant decision trees and their subsequent representation as OWL ontologies. We show how the OWL ontologies can be used for toxicologically relevant classification over RDF-based linked data and discuss the potential for the application of OWL-based decision trees on large RDF chemical knowledgebases. We believe that this work is an important initial development in the formalization, standardization, and integration of computational toxicology resources and predictive classification methods.

3.3 Results

3.3.1 Structure-Based Classifications

We have reduced the activity of a human curator to several steps, namely a) the identification of high-confidence chemical class members, b) identification of consensus chemical patterns that have to be present in a given compound to qualify for membership in a given class, c) creation of formal chemical class definition axioms, and d) chemical classification and class relationship assignments based on the acquired expertise (Figure 3-2). The ontology that we aim to generate with the framework we have developed is primarily based on, but not limited to, chemical structure. We operate on the premise that chemical class sub-specialization is enabled through chemical structure extension. We claim that for each class, a collection of functional groups exist that are necessary for a molecule to be considered a member of this class. This forms the basis for the self-organization of our automatically generated chemical ontologies.

Figure 3-2. An abstract view of the curation effort for structure-based chemical hierarchies.

3.3.2 Automated Derivation of Formal Chemical Class Definitions

Given a set of chemical entities known to be members of a given chemical class, our framework is capable of abstracting and generalizing a set of consensus molecular sub-graphs and features that are present in all representative members of this class (see Methods). In identifying such features, we consider not only the chemical graph that is formed by the overlap of several smallest class representatives, but also the chemical sub-graphs formed by the fragmentation of class members, and atomic features, such as ring membership, aromaticity, and connectivity. Therefore, we generate a set of both, canonicalized SMILES and canonicalized SMARTS [16] patterns (See Methods) that are present in a given molecule. Canonicalization of SMILES and SMARTS patterns is

important because the fragmentation of chemical structures and their generalized forms often yields chemical patterns that are equivalent, but can be written down as valid SMILES or SMARTS in multiple ways. Canonicalization of consensus patterns ensures that each pattern has exactly one definition, thus allowing for facile integration and comparison of consensus fingerprints from multiple class definition generation procedures.

The defining features, or chemical structural motifs, that are present in every member of a given class are deemed to be necessary for class membership and are termed *consensus chemical features* and their collection may be viewed to constitute a *consensus chemical fingerprint* for a given chemical class. The consensus chemical features that contain but are not contained in other consensus chemical features are termed *principal characteristic substructures*. These substructures can be used directly by curators for quick validation of the automatically generated class definition. For instance, for the ChEBI Peptide class (CHEBI:16670), one would expect the principal characteristic substructure to constitute the amino acid backbone (Figure 3-3), which is indeed the case.

Fragment	Contained	Principal Characteristic Sub-Graph A
C	y
C=O	y	
CC	y	
CC(=O)N	y	
CC(=O)O	y	
CC=O	y	
CCN	y	
CCO	y	
CN	y	
CNC	y	
CNC(=O)C	n	
CNC=O	y	
CNCC	y	
CO	y	
N	y	
NC=O	y
NCC(=O)O	n	
NCC=O	y	
NCCO	y	
O	y	
OC=O	y	
		Principal Characteristic Sub-Graph B

Figure 3-3. Some of the consensus structural features identified for the class of peptides using the proposed methodology. Note that in the two consensus fragments that are not contained in other consensus fragments, the basic peptide backbone is conserved, even for proline, a structurally exotic amino acid, since ‘N’ in SMARTS does not necessarily mean an NH₃ in practice.

In the process of determining principal characteristic substructures, we have identified certain classes for which our algorithm could not converge on an expected substructure set. Surprisingly, this was observed for carboxylic acid esters, where only the pattern associated with a carboxylic acid, but without extension to create an ester was found. Upon closer investigation, we have discovered that these difficulties in convergence were exclusively due to errors in the manual curation of input data. In the case of carboxylic acid esters, the failure of convergence was due to the inclusion of a molecule containing RC(=O)ONR pattern instead of the RC(=O)OCR pattern in the training set. Technically,

carboxylic acid esters conform to a pattern $RC(=O)OR'$, where both, R and R' are aryls or alkyls, meaning that the inclusion of the nitrogen-linked ester derivative was erroneous. This was a very fortunate development, since this identified one possible method to verify the consistency and correctness of human curators.

Our second premise is that for one class to be considered a child class of another chemical class, it has to contain all the consensus structural features of the parent class, and some additional features specific to it only (most likely, but not necessarily, principal characteristic substructures). Therefore, while the principal characteristic substructures will be useful in establishing the unique structural identity of a given class, the rest of consensus structural features are useful in establishing the class hierarchy (Figure 3-4). This basic premise applies to both, simple consensus structural fingerprints, and the formal logical definitions of chemical classes and forms the foundation of the self-organizing nature of the chemical ontology generated by our framework.

Figure 3-4. Establishing a chemical class hierarchy through direct observation of consensus structural fingerprints is similar to logical hierarchy construction. It is possible to infer parent/child relationships based solely on the consensus features of a class. Formal logic affords much more powerful expressions.

The next step in self-organizing chemical ontology generation is the formalization of chemical class definitions based on the consensus structural fingerprints. For this, we generate and maintain a single OWL ontology that is constantly updated with the unique consensus chemical substructures that are identified as a result of class analysis, using previously reported principles of unique, canonical, and invariable URIs for each functional group class (Chapter 2). The functional groups defined in this ontology are then used in converting the consensus structural fingerprints. For instance, a class of 'Organic Alcohols' is found to contain the consensus structural features expressed in SMARTS notation as [#6] (any carbon atom, aromatic or not), [#8] (any oxygen atom), [#6][#8] (carbon and oxygen in a single bond), and [#6][OX2H] (carbon connected to an

OH group). This can be translated into the following equivalent class statement in OWL using the Manchester DL query syntax, which states that an instance of the class of organic alcohols is exactly defined by being an instance of a molecular entity and having at least one (keyword: *some*) instance of each of the four functional groups:

'Organic Alcohols'

EquivalentTo

'Molecular entity'

and 'has part' some '[#6]'

and 'has part' some '[#8]'

and 'has part' some '[#6][#8]'

and 'has part' some '[#6][OX2H]'

Clearly, this specification is human-readable (for humans trained to recognize SMARTS patterns), and can be further enriched by adding annotations on the functional groups involved, including graphical representations or textual descriptions for the various functional groups, in order to facilitate chemical class definition understanding.

Further, if one wishes to handle cases where more expressive logical statements are necessary, this is always an option. Consider, for example, the class of Glycols (also sometimes referred to as Diols), which require the presence of a minimum of two alcohol functional groups in a given molecule. Please note that we state that diols contain at least two alcohol functional groups, since in our approach, triols should be identified as simply a subclass of diols. The cardinality restriction to at least two alcohol functional groups allows for this. Thus, diols can be defined as follows. Please also note that cardinality is

not currently screened for by the framework proposed here and all the cardinality restrictions in our automatically generated ontologies were added manually.

Diols

EquivalentTo

'Molecular entity'

and 'has part' some '[#6]'

and 'has part' some '[#8]'

and 'has part' some '[#6][#8]'

and 'has part' min 2 '[#6][OX2H]'

Attributes and molecular descriptors can also be handled with such definitions. For instance, we may find out that a given enzyme catalyzes the oxidation of organic alcohols that are no heavier than 500 Daltons. This mixing of chemical sub-graphs and physicochemical attributes is also possible, using the principles and concepts defined in the CHEMINF ontology.

'Reactive Diols'

EquivalentTo

'Molecular entity'

and 'has part' some '[#6]'

and 'has part' some '[#8]'

and 'has part' some '[#6][#8]'

and 'has part' min 2 '[#6][OX2H]'

and 'has attribute' some ('molecular weight' and 'has value' [= < 500] and 'has unit some 'Dalton')

Using the dataset of diols annotated with their computed molecular weight with the concepts from the CHEMINF ontology, we were able to correctly identify that all 14 diols in the ChEBI training set for this class were indeed ‘Reactive Diols’ as per the logical definition above.

3.3.3 Self-Organizing Chemical Ontology

In order to test the integrative capacity and accuracy of our chemical classification framework, we have manually generated a dataset containing 40 ChEBI and 60 MeSH chemical classes, for a total of 100 classes with at least 7 and at most 20 representatives in each chemical class. To demonstrate the self-organizing chemical ontology creation from a specific source, we have created a stand-alone self-organizing ontology for each, the ChEBI and the MeSH data sets, containing the formal class definitions derived from consensus chemical fingerprints of each class (Figure 3-5). Upon application of the Pellet reasoner to classify these ontologies, we observed self-organization of the formally axiomatized chemical classes into chemical ontologies with a hierarchy (Figures 3-6 and 3-7). Each child-parent inference was screened by a curator to identify that the automatically generated ontologies were consistent (from a structural viewpoint) for both datasets.

Figure 3-6. MeSH chemical class hierarchy automatically inferred using the Pellet reasoner. Please note that all ‘is-a’ relationships except those to ‘molecularentity’ are inferred automatically.

Figure 3-7. ChEBI chemical class hierarchy automatically inferred using the Pellet reasoner. Please note that all 'is-a' relationships except those to 'molecularentity' are inferred automatically.

Because both ontologies were built using one central functional group ontology and the CHEMINF ontology for chemical descriptors, we were then able to trivially import the ChEBI ontology into the MeSH ontology within Protégé 4.1 by specifying direct ChEBI ontology import within the MeSH ontology, and to apply the Pellet machine reasoner to the resultant ontology to reconstitute a unified chemical ontology with 100 chemical classes (Figure 3-8). Thus, for the first time, we were able to automatically and trivially integrate and re-use the manual curation efforts from multiple data sources in order to construct a single seamless chemical ontology. Manual analysis of the class relationship inferences was carried out to ensure that the inferred chemical hierarchy was consistent and coherent from the viewpoint of chemical structure.

3.3.4 Chemical Classification Accuracy

For a demonstration of chemical classification, we have randomly selected a set of 200 chemical entities and employed the integrated ontology in order to classify each individual compound. Because each chemical entity could be classified into multiple classes, the number of class membership inferences (452) exceeded the number of tested chemical entities. Within this dataset, 91% of chemical entities received inferences that exactly matched their annotation within the test set and additional 8.5% were matched to classes consistent with their test set-assigned class. For example, if a compound was a member of the alcohols in the test set, but actually contained two OH groups, this compound was correctly identified as belonging to diols or glycols, without the explicit specification of its membership among the alcohols, even though this membership is implied.

Overall, 92.7% of the entity inferences were correct. Most of the observed errors were in the classes that necessitated the definition of negations. For example, a carboxylic acid is a structural derivative of an alcohol, but is strictly speaking not one. From a structural perspective, classification of a carboxylic acid as an alcohol is not incorrect. However, from a classical organic chemistry viewpoint, this is an incorrect classification, and has been noted as such in this work. In addition to this, we have observed that identifying incorrect inferences in some cases led to facile identification of adjustment of the training sets to automatically generate a more accurate chemical entity description or suggested manual refinements to definitions (e.g. glycols, diols, and alkadienes with cardinality restrictions). The entire classification process for 200 compounds took 34 minutes and 10

seconds, running on a Core i7-870 machine with 6GB of RAM, executed serially, while the computational time to classify non-populated ontologies was negligible.

3.3.5 Seamless Integration of Novel Chemical Classes: Enzyme Substrates

In order to demonstrate the flexibility and facile amenability of our ontology to the incorporation of novel chemical classes, and especially chemical classes of biological relevance, we have created a stand-alone class of chemical entities that are substrates to yeast alcohol dehydrogenase, using information in the BRENDA database. Because this data was experimentally derived and the reactions indicated were confirmed to occur in nature, no human curation (beyond eliminating redundant entries) was necessary for the training set composed of 23 unique chemical entities. Upon the repetition of the exercise described above with this chemical class, we have, unsurprisingly, discovered that the resultant class of yeast alcohol dehydrogenase substrates was a subclass of alcohols, with the added requirement of possessing an oxygen atom connected to one hydrogen atom and a non-aromatic carbon atom.

3.3.6 Functional Chemical Classification

3.3.6.1 OWL-Based Decision Trees: Rule of Five

The first task that we addressed with our automated OWL ontology decision tree generator was the construction of simple ontologies where the classification rules involved the evaluation of numerical values associated with various molecular descriptors. This is a fairly common mode of preliminary screening of large compound datasets in initial stages of cheminformatics analysis. The decision tree generated by Weka [17] using computed data reproduced the Rule of Five criteria (Figure 3-9).

Figure 3-9. A decision tree generated from a computationally derived dataset of drug-like compounds. Drug-like compound classification is indicated as *true*. Correctly classified molecule counts are given in brackets. No classification was incorrect.

There was little surprise that the Rule of Five criteria (used as an example, not a practical application) which we imposed in the computationally derived dataset were perfectly returned to us after data-based decision tree construction in Weka. However, this demonstrated to us that, given a sufficient amount of data with low levels of noise, one could successfully derive meaningful and useful numerical cutoff-based decision trees which could subsequently be converted to predictive ontologies.

In order to carry out the conversion, we have followed the scheme indicated in Section 2.3 to obtain a set of substance classes that followed numerical cutoff rules, such as the following.

Substance_N1:

Substance_N0 and has_attribute some

(MolecularWeight and has_value some double[<= "500"^^double])

As a result of applying our generator, we have obtained an ontology that perfectly captured the Rule of Five decision tree (Figure 3-10).

Figure 3-10. The structure of an automatically generated OWL representation of a Rule of Five tree (Figure 3-9).

3.3.6.2 OWL-Based Decision Trees: Large-Scale Boolean Feature-Based Trees

Unfortunately, biological information is often subject to extensive variation, whether due to noise in experimental conditions or the abundance of the variable parameters that may

differ even within a single laboratory and experiment. Compounding this is the limited experimental data availability to characterize most forms of biological activity, especially for experiments that are not high-throughput at inception. As a result, the real-world data is rarely as neatly classifiable as in the decision tree above. However, our primary concern in this work has been the proof of principle for the utility of OWL-based decision trees. To this end we have been able to generate a number of useable trees with the full 318-feature set (not shown due to complexity), as well as the more presentation-friendly limited feature sets (Figure 3-11).

Figure 3-11. A simplified carcinogenic toxicity decision tree generated from a ToxCast dataset, using a restricted set of chemical features for ease of presentation. Note the repetition of some rules at multiple decision tree nodes. The path taken to classify acetaminophen, as detected with the explanation functionality of Protégé, has been highlighted with double arrows.

Upon closer examination of such increasingly complex decision trees, we have identified several unanticipated classification challenges. The greatest surprise has come from the identification of the logical equivalence of several branches within some of the generated trees. While that was considered completely plausible at the level of the individual nodes, the subsequent identification of the logical equivalence of the final toxicity and bioactivity classifications upon the application of reasoners to our generated ontologies has led to some concerns over the validity and applicability of our approach. Clearly, the equivalence of the class of toxic compounds to the non-toxic compounds is not an anticipated or desirable effect for an ontology used to replace the existing classification systems. Further, in order to make the decision tree more transparent, we needed a way to trace the logical path taken to activity classification leaves, while still preserving broad activity classification capacity.

After careful consideration of the logical explanation of the equivalence of these practically distinct classes, we identified the cause of the problem to lie in the repetition of rules within a single decision tree and the lack of the distinction between the nodes that executed rules in a particular order. As such, it was quite possible to arrive at a situation where, having ignored the context of the rest of the tree, the classifier technically correctly assigned class equivalence between the toxic and non-toxic compounds simply because parts of the paths taken to these classifications were similar, while the other parts were not mutually exclusive.

To rectify this problem, we have recognized that node-specific classification rule tracking had to be implemented. Thus, we amended our generator to include a local set of node-

specific classification features within a given ontology. This translated into alterations to substance classifications, as follows.

Substance_N6:

Substance_N0 and has_attribute some

(RuleToxicFunctionalGroups_N0 and has_value value false)

Note that what used to be the *RuleToxicFunctionalGroups* descriptor became the *RuleToxicFunctionalGroups_N0* descriptor. This amendment was effective in solving our misclassification problem from a computational viewpoint. However, the introduction of ontology-specific descriptors would negate our ability to integrate and compare the different ontologies, as well as to draw on existing repositories of chemical entities annotated with the general standard descriptors and features. To rectify ontology comparison deficiency, we have created versions of our decision tree ontologies where node-specific rules were explicitly defined as subclasses of their generic counterparts. Similarly, node-specific activity leaves were introduced to enable tracing classification paths. Thus, although we had to artificially distinguish activity categories and rules, we were still able to query for the compounds falling into the general activity classes, as well as to trace classification paths, important in e.g. automated toxicity tree comparison.

3.3.6.3 Chemical Entity Classification

While the above amendment permitted comparison between multiple ontologies and still avoided erroneous class equivalence conclusions, it did not address drawing on existing data repositories, as there is no direct inference that if a general rule bears a particular value, there exists an instance of its subclass that bears the same value. The first intuitive

suggestion to clear this task is to modify our generator to also create ontologies where the general rules were specified as subclasses of the node-specific rules. This has allowed us to automatically make the necessary inference to import data from existing chemical knowledge repositories in RDF.

However, upon carrying out the classification within such ontologies, we have been unpleasantly surprised to find out that due to the introduced equivalences at the data level, some of our instances were capable of adopting both, active and inactive classifications. In order to rectify this problem, we defined node-specific final classifications (e.g. *active_N3*) which were declared to be subclasses of the general final classes (e.g. *active* and *inactive* in this context) (Figure 3-12).

Figure 3-12. A fragment of the final, classification-friendly decision tree.

Classification was successfully carried out by querying whether a given instance belonged to one of these general classes. Using thus constructed decision tree-based ontologies, we encountered no problems classifying numerous RDF-encoded molecules bearing the requisite information. A sample OWL model is available [18].

As an example, consider the case of acetaminophen, a known non-carcinogen. Its attributes (Figure 3-13) were imported from its CHESS representation and correctly

classified as *inactive* according to the decision tree presented earlier. This classification was also reproduced using numerical trees (omitted for brevity).

Structure	Acyl	Rule Alcohol	Rule	Rule	Rule	Rule	Acyl	Rule	Rule	Toxic	SN2 Rule
	Transita		Aldehydes	Aromatic	Aromatic	Methyl	Tertiary	Functional			
	Rule		Aldehyde	Aldehyde	Aldehyde	Halide	Amine	Group
TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Figure 3-13. Relevant features of acetaminophen used in classification.

Further, unlike the traditional statistical classification systems, which are often essentially black boxes, our semantic decision tree-based approach has allowed us *to automatically trace the exact route taken to classifying acetaminophen as a non-carcinogen*, using the explanation feature of Protégé [19]. The fact that we created artificially distinct activity classes in our trees did not prevent us from querying for chemical activity in terms of general categories.

3.4 Discussion

Here, we have presented a radically streamlined framework and approach to chemical ontology construction based on explicit formal semantic specification of chemical classes and rooted in chemical graph analysis. By eliminating the manual assertion of ancestry relationships between the various chemical classes and instead emphasizing the adequate curation of the limited (as small as five entities) training set of the starting chemical class representatives and the automatically generated logical class definitions, we have greatly simplified the work of human curators, thus improving their productivity and making the annotation and classification of the rapidly multiplying chemical information manageable.

One may argue that curators may also easily define chemical patterns in order to define a given chemical class instead of starting with identifying a diverse set of compounds that

would then be used to automatically derive the relevant patterns. While this is indeed true, automated consensus chemical structure identification has numerous benefits, including the exhaustive identification of primitive structural features that are useful in constructing class hierarchies, as well as in the inherently objective identification of the structural patterns of interest. Manual supplementation of class definitions is always possible, as we have shown with class customization, and is sometimes required, in cases where our algorithms cannot identify a consensus pattern despite the best efforts of the curators to include only valid structures in the training sets.

For example, classes that include much variability within the key structural elements may not be adequately characterized by our algorithm as it stands currently. For example, the IUPAC-recommended interpretation of the term ‘ester’ [20] includes compounds that conform to $C(=O)OC$ (i.e. carboxylic esters) and $C(=S)OC$ patterns, among others. If entities representing both of these patterns were grouped into a single training set, a single descriptive pattern would not be identified with our current algorithm, since there is variation on the central atom. In this case, a human curator’s intervention may be necessary to include a pattern such as $C(=[O,S])OC$ that describes both cases.

Although this is a difficult problem to address, it is not unavoidable with future improvements to the consensus chemical pattern identification algorithms. For example, had we not considered the atomic ring membership property in the generation of our consensus chemical patterns, we would not have been able to adequately model the class of cyclic organic compounds. One approach around this would be to manually define a feature $[#6;R]$ and screen for this feature to identify whether it is a consensus chemical feature for a given class. Another, much more labor-intensive approach would be to

identify the cyclic organic compound sub-classes and explicitly enumerate them in the definition of this class, as follows.

'Cyclic Organic Compounds'

EquivalentTo 'Molecular Entity' and 'has part' some Cycle

'Three-Membered Cycle' subClassOf Cycle

'Five-Membered Cycle' subClassOf Cycle

Fortunately, our algorithm currently identifies and screens for the presence of a number of ring features and abstracts chemical structures so as to include considerations for atoms' hydrogenation state (such as OH versus O, CH₃ versus CH₂), ring membership, aromaticity, general type (such as [#6] is any carbon atom), as well as combinations of these attributes. Thus, the feature [#6;R] is automatically identified by our algorithm by examining the chemical entities in the training set. In addition to this, we screen for 396 pre-defined patterns designed to capture the nuanced chemical class definitions, such as the general ester feature discussed above, and we screen each class with all the previously-generated consensus patterns in the other class definitions. This allows us to handle the definition of many more classes than just those with automatically generated consensus substructures, including, for example, bicyclic molecules, which are composed of two fused rings and are characterized, among other automatically identified features, with the manually pre-defined [R2] SMARTS pattern, which defines any atom contained within two rings simultaneously.

While some cases can be handled automatically in principle, for some classes of compounds such as Nanotubes (CHEBI:50796) or Polycyclic Cages (CHEBI:33640), manual curation is unavoidable, as we do not yet have the capacity to adequately

characterize the requisite consensus chemical feature information in terms of SMILES, SMARTS, or properties, but only some very complex logical expressions or expressions in some molecular query languages. This may potentially be handled automatically in the future by screening for the requisite higher-order chemical graph features and patterns using seamlessly integrated semantic web services with the SADI framework [28], but is not handled by us at present.

Another aspect that we would like to address in the future is the automated identification of consensus chemical feature values in order to investigate the handling of more biologically relevant chemical classes, as demonstrated for the ‘Reactive Diols’ class. While we have already demonstrated the learning and formal logical encoding of biological activity and toxicity decision trees, we have not yet connected that effort to structure-based classification. In addition to this, classes of biologically active compounds are rarely as well defined as they are for the alcohol dehydrogenase substrates. For instance, structurally distinct compounds may sometimes engage in a number of productive binding or interaction modes with the same enzyme. While such annotation is not our primary focus, we claim that each individual binding mode should be amenable to topology-based and consensus feature-based logical characterization as a sub-class of the broader enzyme substrate class. Failing that, as we have demonstrated, the existing chemical graph-based class definitions can always be easily augmented with various chemical property and geometric characterization restrictions manually.

Unfortunately, certain aspects of chemical class definition formulation are difficult to address in principle. Of particular concern is the negation of chemical statements in some class definitions. For instance, some people may argue that fatty acids constitute a class

of compounds that should not contain any rings, among other features. This negative relationship is rather difficult to capture with the proposed methodology, and in fact, the majority of cases where classification of small molecules was incorrect were due to the lack of automatically generated negations in the class definitions. The difficulty here is not posed so much by the identification of the features that are consistently absent from every member of a given class, as it is by the identification of which of the numerous features identified as absent from every member of a class should indeed be absent. Of course, given a diverse enough training set for a given class, many of the chemical features would be eliminated from the forbidden list. However, one may never be entirely certain that the groups still on the list should indeed be forbidden in a given class, until all the possible class members have been included in the training set.

The final aspect where human curation still has an edge is human-created, arbitrary or otherwise necessarily diverse groupings, such as is the case for vitamins and many natural products, or terms that refer to broad functional roles rather than chemical structure. Certainly, this is still very much a domain of subjective human classification where humans are irreplaceable, and will likely never be wholly replaceable by automated and objective chemical classification approaches. Thus, human involvement is necessary in our proposed mode of chemical ontology construction and refinement, but it is also more subtle, nuanced, and much more productive than it has ever been in any of the ongoing chemical ontology construction efforts.

Having discussed the limitations and caveats of the proposed framework as it currently stands, let us examine more closely what it enables. By providing chemical class definitions in a logical and objectively consistent manner, we have created a means by

which chemical hierarchies can be seamlessly integrated and new, previously uncharacterized classes can be added without necessitating a manual review of the whole ontology structure. The process of annotation of chemical entities has been greatly simplified and reduced to merely chemically fingerprinting a molecule and importing its CHESS-encoded fingerprint for facile automated classification with machine reasoning agents, all of which can be done on-the-fly as the chemical entities are loaded into a given ontology.

Finally, there is no further reason for compartmentalization and erection of artificial barriers between the ontologies derived by distinct chemical classification efforts. In principle, all chemical hierarchies built using concepts within common or integrable ontologies can be fused into a single, consistent, and self-organized ontology, where chemical class equivalences and relationships are automatically inferred and reconciled. We have demonstrated this with the automatic integration of chemical classifications derived from multiple, independently-curated data sets, one using the MeSH-annotated PubChem molecules, and one that relied on ChEBI small molecule annotations. Previously, classification efforts in either manually created hierarchy stressed the manual establishment of explicit parent/child relationships between the various chemical classes, as well as potentially formulating a short description regarding what exactly went into deciding whether certain compounds should be annotated as members of a given class. In this regard, ChEBI is, beyond comparison, superior to the MeSH classification, because unlike MeSH it includes a programmatically examinable, formal ontology of chemical classes that can be queried for small molecule class membership. In addition to this, the

ChEBI ontology understandably includes many more terms and specific classes of biological interest.

Unfortunately, neither of the two chemical ontologies currently has means for facile integration with each other, nor facile class extension and inclusion of previously uncharacterized chemical classes. For both ontologies, the precise location of novel class addition is performed using subjective human assessment and examination of the possible child and parent classes. When a decision is made for the location of the class, no formal justification is provided for the decision, meaning that the location of the class is subject to errors and interpretation. On the other hand, in the formally specified ontologies that we generate by screening the data that is already available with PubChem and ChEBI, every subclass inference is explainable logically and amenable to facile correction through adjustments in formal chemical class definitions rather than an unexplainable reshuffling of the class relationships in the current ontologies. Furthermore, the effort of adding a novel class into the ontology is limited to defining the class itself, rather than analyzing all the other existing classes for their potential role as parent classes or subclasses of a given novel class. This is the principal reason for the ease of the integration of the MeSH and ChEBI ontologies. To summarize, the principal benefits of our proposed approach are i) the ease of chemical class integration into existing hierarchies, ii) facilitated chemical class definition and characterization, and iii) logically explainable and therefore easily correctable chemical class hierarchy assignments. Furthermore, to generate the hierarchies presented in this work, we have essentially re-used the previous work done by annotators on the ChEBI and MeSH projects already, with slight screening

modifications of the existing data sets. This gives hope that the existing chemical hierarchies can be somewhat easily converted into their logically encoded forms.

The automatically generated ontology that we present here also removes much subjectivity and ambiguity from chemical class assignments and annotation of individual small molecules. It is no longer necessary to manually adjust the annotation of every small molecule that needs to be classified. This is of critical importance as the ChEBI effort transitions to the classifications of hundreds of thousands compounds present in the ChEMBL database, and the PubChem database experiences constant exponential growth of millions of new, partially characterized or completely uncharacterized compounds. The classification procedure for the novel small molecules presented here is highly accurate, and when each class is screened using a testing set, classification inconsistencies are highly visible and easily correctable. For example, the classes of alkadienes and diols necessitated cardinality restrictions in their formal definition to account for the fact that these classes required a given molecule to contain two alkene or two alcohol groups, respectively. Upon screening these classes with a test set of known positives and known negatives, the misclassification of known negative compounds containing only one of these groups has alerted us to the omission of cardinality restriction in the definition of these classes. While we anticipate adding automatic cardinality restriction identification in future work, this feature is currently not implemented and appropriate class definitions had to be manually corrected. Thus, awareness of the formal definition of these classes has alerted us to an omission in our work. On the other hand, there is no capacity or provision for explaining chemical classifications and annotations in the major chemical ontology building efforts, which

means that any errors in chemical entity annotation assignment have to be manually identified by the users of these hierarchies, discussed by curators, and then manually corrected, in a further time-consuming endeavour. We shall reiterate: in the proposed framework, only class definitions need to ever be adjusted, while chemical entity classification is automatically inferred, and can be generated at will, for any arbitrary compound.

We must note that the high degree of accuracy of automated small molecule classification is not surprising since we were operating on structurally-derived classes and we have performed a nearly exhaustive characterization of chemical structure to identify unambiguous, deterministic class definitions. This is in sharp contrast with the inherently probabilistic, ‘black box’ training approaches based on artificial neural networks or support vector machines, where features of chemical entities in a given training set are incorporated into (often difficult to interpret) mathematical expressions which are expected to be correct only a fraction of time, and due to the statistical nature of these approaches may involve chemical features that may have little to do with the proper classification of a given compound into a given class, in reality. Our task is very much simplified, as we do not concern ourselves with chemical features that only appear in a fraction of compounds within a given class, but rather admit to class definition only the consensus features.

Based on our own experiences in constructing the chemical hierarchies presented here, we envision a new iterative approach to chemical ontology development (Figure 3-14). In this scheme, curators are primarily concerned with the assimilation of a representative set of chemical entities for each chemical class based on whether the desired principal

chemical substructures for a given class have been identified. Manual adjustment of some class definitions and adjustment of other definitions to explicitly include disparate classes that may contain a set of well-defined, but disparate and non-overlapping sub-classes may also be performed. However, explicit specification of class membership or chemical ontology hierarchy is possible but not recommended in this approach to maintain the self-organizing and objective nature of the chemical hierarchy.

Figure 3-14. A novel approach to chemical hierarchy construction, maintaining human involvement and supplemented with automated cheminformatics algorithms developed in this work.

Finally, several words have to be said about the functional classification of chemical entities presented here. In this work, we have demonstrated for the first time the automated construction and practical application of OWL-encoded decision trees in chemical toxicology. The OWL ontologies that we generate can capture numerical cutoff-based rules, as well as Boolean-based rules, and can be used to represent both automatically and expert-generated decision trees. Using our approach, decision trees that

form the basis for predictive chemical toxicology classification and are either manually (expert-based systems) or algorithmically (data-based systems) generated can be routinely converted to OWL ontologies. Due to the explicit and formal specification of concepts within these ontologies, toxicity and bioactivity classes can be exposed for comparison and logical integration. In addition to this, these ontologies can also be easily applied to classify chemical entities in the rapidly growing knowledgebases of RDF-encoded chemical information. As our single-molecule classification performance data suggests, large sets of chemical entities could be efficiently classified. In replacing framework-, software-, and domain-specific classification engines with standard OWL ontologies, we allow for the chemical toxicology efforts to break free of their respective boundaries and support their current shift towards the Semantic Web technologies. As this shift occurs, we are confident that the work we present here will play an important role in informing future efforts in integrating and analyzing the future Chemical Semantic Web to support open, transparent, and reproducible chemical toxicology research.

This work marks a first step towards an OWL-based predictive toxicology framework that is currently under development. In this framework, ontologies capture the decision tree-based toxicology and bioactivity mathematical models are generated on the fly from linked open data. These ontology-specified models will subsequently be accessible for further automated classification of large collections of semantically represented chemical entities. Preliminary results point to the possibility of logically comparing formalized decision trees of multiple types so as to provide explanations for [19] and to identify points of equivalence of toxicity and bioactivity classes. Finally, the capture of classification statistics presents an interesting avenue to explore probabilistic reasoning

[21] using description logics which would be well suited for toxicity prediction within a set of confidence intervals.

With the rising of systems sciences to prominence, we believe that chemical annotation and classification frameworks, such as the one presented here, shall increasingly rise in importance. The already impressive collection of chemical data on the Semantic Web and its fruitful applications in chemical research suggest that semantically enabled chemical classification systems are of immediate applicability. Finally, the ability to formally specify and exchange chemical class definitions can open new doors to collaboration in chemical and biological research, driving new innovations and discoveries.

3.5 Methods

3.5.1 Training and Testing Data Sets

For our proof of principle work, we have identified 60 MeSH classes and 40 ChEBI classes that were subjectively identified as potentially related by a curator and which would thus be expected to either have equivalent logical definitions or identified as sub- or super-classes of each other. For MeSH-defined chemical classes, we have used PubChem to identify all compounds previously annotated with a given MeSH chemical class term. The resultant set of chemical structures was carefully manually screened by a curator to assure that the MeSH term has been correctly assigned in the training data and that the representative class instance set adequately covers the full spectrum of compounds included in the class. For ChEBI classes, we have used the ChEBI API to select only the high-confidence (three-star in the ChEBI star system) annotated chemical entities that were identified as instances of a given ChEBI chemical class. In order to assure complete adherence to the ChEBI class definition in the class instances, a curator

was assigned to manually screen these instances with the same criteria as those for the MeSH data set. This resulted in 766 chemical entities manually screened for the MeSH data set and 606 chemical entities manually screened for the ChEBI data set. The screening process involved not only the querying of the appropriate databases for members of a given class, but also consultation with the existing literature on the definition of each class, as class definitions in ChEBI, and especially in MeSH were often lacking in detail.

3.5.2 Identification of Consensus Molecular Fragments and Patterns

For the identification of consensus chemical substructures, structural patterns, and molecular skeletons, we have created custom software that relied upon the Chemistry Development Kit and OpenBabel. The software we have created is freely accessible upon request. For each class, the set of representative chemical entities is partitioned into a training set which contains at least five least complex representatives (in terms of the number of atoms and bonds) of the investigated class (manually assuring a complete coverage of the class), and a test set which is used to assess the accuracy of formal class definition. The assumption that allowed selecting the least complex class representatives for training is that the smallest class members have to contain all the consensus structural patterns that the rest of the class contains. This allows us to save on computational time.

A chemical substructure, pattern, or molecular skeleton is considered to be a consensus one for a given class if it is found to be present in all members of the training set. In order to identify chemical features to consider as candidates for consensus features, we employ a hybrid approach involving limited molecular fragmentation, subgraph isomorphism detection, predefined chemical feature detection, and the fragmentation of chemical

entities with atoms that are annotated for full hydrogenation state (indicating a terminal atom in a chain, e.g. CH₃ or OH), ring membership, aromaticity, general atom type, as well as binary combinations of these attributes including aromatic and fully hydrogenated atoms, ring and fully hydrogenated atoms, and general atom types and fully hydrogenated atoms.

To identify smaller consensus substructures, we have carried out only a limited molecular fragmentation since we were limited by the apparent and highly perceptible combinatorial explosion of possible fragmentation choices on large and ring-containing molecules. We have therefore limited the fragmentation to exhaustively consider all the possible fragmentation patterns resulting from cutting all possible combinations of up to four bonds in every molecule. If the fragmented compound was the native SMILES-encoded structure, each fragment was then canonicalized using OpenBabel and only the unique fragments were collected. For SMARTS-transformed compounds, the procedure was somewhat more involved (see below).

To identify principal characteristic substructures for each class, we have employed Chemistry Development Kit in order to compute the maximal common substructures in binary combinations of the five smallest class members. Again, the selection of a limited number of the smallest class members has been made in order to eliminate unnecessary computational expense. The principal characteristic substructure for a given class has to be present in every member of a given class, including the smallest members. This also means that one class member in the curated set should be selected such that it contains the smallest set of features while still qualifying as a class member. For example for the class of alcohols, this could be methanol.

The resultant collection of maximal common substructures was then used to derive the maximal common substructures of the maximal common substructures. That is, even if the smallest class members did not contain the one smallest member that is still representative of a given class, this procedure would have identified the principal characteristic substructure for a given class. Consider, for example, the class of Benzenes, which is composed of benzene and its derivatives. The principal characteristic substructure is the benzene ring. However, if we included among the smallest substructures only three different products of addition to the benzene ring, the ring itself would have been still identified as a result of this procedure. We accounted for all three smallest structures involving consensus features (such as if all were aniline-based) by also fragmenting the maximal common substructure thus identified using the fragmentation approach described above.

Finally, each member of the training set was screened for the presence of each feature identified using fragmentation and maximal common substructure analysis, along with 396 manually pre-defined features, obtained from a number of collections. The positive responses were tallied and the fraction of compounds that contained every given substructure or structural pattern was then computed. If this fraction was below 1.0, the fragment or pattern was discarded from the consensus fragment list. If, on the other hand, this fraction was 1.0, a given fragment or pattern was identified as a consensus structural pattern or fragment.

3.5.3 SMILES and SMARTS Pattern Canonicalization

Canonicalization is important since, among other things, it helps prevent repetition and unnecessary screening of the chemical entities for features multiple times. The SMILES

patterns obtained using the above approaches were canonicalized using the OpenBabel software. For SMARTS patterns, the situation was complicated by the fact that, to the best of the author's knowledge, there is no single piece of software that can canonicalize SMARTS patterns. To accomplish this, we have used the fact that atomic mass has a bearing on the ordering of atoms within a canonical SMILES representation, as generated in OpenBabel. Thus, we have represented each atomic feature as a separate isotope type (e.g. a ring carbon is ^{14}C , while an aromatic carbon is ^{15}C) and used such SMILES strings in order to determine a canonical atom ordering using the normal SMILES canonicalization procedure in OpenBabel. After obtaining the canonical isotopic SMILES string for a given fragment, we converted it using a lookup table into a SMARTS pattern specification. This gave us a robust canonical form for every SMARTS pattern we generated.

3.5.4 Chemical Class Definitions Encoding in OWL

In order to translate the consensus chemical fingerprints to OWL-based axioms, we have used the OWL Java API, version 3.2.3 [22]. We have set up a central ontology of functional groups and structural features, which is referenced by every ontology that contains chemical class definitions generated using the proposed methodology. This allows for facile chemical hierarchy integration from multiple distinct efforts. In the functional group ontology, all the functional groups are merely declared with a canonical URI, which is generated using reproducible SHA1 40-character hashes of the canonical functional group SMILES or SMARTS pattern. This ontology contains no hierarchy, but only the instances of functional groups and their labels (SMILES or SMARTS).

The ontologies containing chemical class definitions are generated by iterating over every consensus feature and adding class restrictions to the definition, through the ‘has attribute’ object property defined in Semanticscience Integrated Ontology, and using the ‘some values from’ restriction on this property. Cardinality restrictions were added manually to the existing ontologies using the Protégé 4.1 ontology editor, only for a very limited set of classes that required these restrictions.

3.5.5 Chemical Classification Accuracy Evaluation

The accuracy of chemical class hierarchy was assessed by applying the Pellet reasoner to each generated ontology, and then recording all the pairwise child-parent relationships within the resultant automatically generated hierarchies. The relationships that conformed to the hierarchies from which they were derived, as well as relationships that were acceptable according to the judgement of a curator were tallied up, and this score divided by the total number of the inferred relationships was reported as classification accuracy.

In order to classify chemical compounds using the formally axiomatized chemical class definitions, we represented 200 randomly selected chemical compounds from the MeSH data set (previously removed from training) using the CHESS formalism, with the Jena Java API version 2.6.2. Each compound was screened to detect the presence of each of the chemical features defined in the central functional group ontology. Unique occurrences of functional groups within each chemical entity were individually instantiated and mereological relationships between these functional group instances and their containing chemical entity instance were defined. Each instantiated entity was identified explicitly as distinct to allow classification into cardinality restriction-containing classes. For the chemical entities that were used to demonstrate chemical class

definition extension, we have also generated CHESS-encoded molecular weight descriptors. After this instance data in the RDF/XML form was imported into the ontology received from the integration of the MeSH- and ChEBI-derived ontologies, the Pellet reasoner was used again to realize the ontology with instances. Each chemical class inference was then manually screened by a curator and scored for exact matches to the test set data, correct classifications, and incorrect classifications.

3.5.6 Data Sources and Data Preparation for Functional Decision Trees

Our analysis made use of empirically and theoretically derived datasets. A carcinogenic toxicity dataset, from which 1400 chemical entities were selected, was obtained from the ToxCast database [23]. These compounds were either active or inactive with respect to single cell mutagenicity. Then, 318 non-redundant features for each molecule were computed using the ToxTree API version 2.1.0 [24] to determine a Boolean value for each feature: true for feature presence and false for absence. These features corresponded to rules at decision tree branch points: true if satisfied, and false if not.

Features for the Rule of Five training set, consisting of 7000 compounds selected from HMDB [25], were computed using the Chemistry Development Kit, and the drug-likeness attribute was derived using the logical tests outlined by Lipinski [26]. Software and data are available upon request.

3.5.7 Decision Tree Construction and Validation

Weka was used to construct and validate binary decision trees using the experimental and computed feature information. Decision trees were constructed using the J48 algorithm [27]. We applied ten-fold cross-validation to derive a set of statistical measures of tree predictive ability. Though these statistical measures are not directly relevant for this

work, they have been included as annotations on resultant OWL-encoded decision trees for completeness. For the purposes of discussion in this work, we generated five decision trees: Lipinski Rule of Five, modified Lipinski Rule of Five, as well as trees resulting from different partitions of the ToxCast datasets.

3.5.8 Representation of Decision Trees as OWL Ontologies

OWL ontologies were constructed using the OWL API from the decision tree graphs represented with the DOT graph description language. Each decision node is represented as being equivalent to a class expression involving the parent decision node intersected with a restriction on the attribute value (true/false) that the parent node represents (e.g. contains an alcohol moiety). For example, given three substances (A, B and C), where A is the parent substance and B and C are defined with respect to the exact value of the parent feature X, and given Substance classes, ‘has attribute’ object property, and ‘has value’ functional datatype property, the equivalent class expressions corresponding to Substance B and Substance C are:

‘Substance B’ EquivalentTo

‘Substance A’ and ‘has attribute’ some (‘Attribute X’ and ‘has value’ true)

‘Substance C’ EquivalentTo

‘Substance A’ and ‘has attribute’ some (‘Attribute X’ and ‘has value’ false)

EquivalentClass axioms were added to terminal nodes corresponding to the final classification, e.g. toxic or non-toxic. This enabled us to reflect both the structure of the decision tree and the formal axioms leading to the classification of a given chemical entity into a given biological functional class. We did not include covering axioms (e.g. A can have the disjoint subclasses B or C) because we would like to avoid

inconsistencies in some manually created trees where multiple classification outcomes may be possible and the most hazardous classification outcome is selected.

3.5.9 Chemical Classification

Molecular entities were instantiated using conventions set out by the Chemical Entity Semantic Specification and the Chemical Information Ontology. These entities annotated with chemical feature data were classified using Pellet through the OWL API into the predicted toxicity classes using our automatically generated OWL-based decision trees.

3.6 Authors' Contributions

This study has been jointly conceived by me and Michel Dumontier. I have created the software, ran the tests, and generated the chemical ontologies. I have worked in collaboration with the European Bioinformatics Institute's Chemical Entities of Biological Interest team to screen and validate the hierarchy inferences. Thus, one team member (a ChEBI curator), Marcus Ennis, has identified inconsistencies in the final auto-generated hierarchy, which helped me identify programming bugs. Finally, Dana Klassen (a fellow student) has been instrumental in the implementation of conversion of decision trees generated by Weka into the OWL format, as described in section 3.5.8.

4 Chapter: Semantic Web Services for Chemistry

This chapter is principally based on the following open-access journal article.

Chepelev LL, Dumontier M. Semantic Web integration of Cheminformatics resources with the SADI framework. *Journal of Cheminformatics*. 2011. 16;3:16.

The following open-access article has been used for subsection 4.3.7.

Chepelev LL, Riazanov A, Kouznetsov A, Low HS, Dumontier M, Baker CJO. Prototype Semantic Infrastructure for Automated Small Molecule Classification and Annotation in Lipidomics. *BMC Bioinformatics*. 2011. 12:303.

4.1 Abstract

The diversity and the largely independent nature of chemical research efforts over the past half century are, most likely, the major contributors to the current poor state of chemical computational resource and database interoperability. While open software for chemical format interconversion and database entry cross-linking have partially addressed database interoperability, computational resource integration is hindered by the great diversity of software interfaces, languages, access methods, and platforms, among others. This has, in turn, translated into limited reproducibility of computational experiments and the need for application-specific computational workflow construction and semi-automated enactment by human experts, especially where emerging interdisciplinary fields, such as systems chemistry, are pursued. Fortunately, the advent of the Semantic Web, and the very recent introduction of RESTful Semantic Web Services (SWS) may present an opportunity to integrate all of the existing computational and database resources in chemistry into a machine-understandable, unified system that draws on the entirety of the Semantic Web.

We have created a prototype framework of Semantic Automated Discovery and Integration (SADI) framework SWS that exposes the QSAR descriptor functionality of the Chemistry Development Kit. Since each of these services has formal ontology-defined input and output classes, and each service consumes and produces RDF graphs, clients can automatically reason about the services and available reference information necessary to complete a given overall computational task specified through a simple SPARQL query. We demonstrate this capability by carrying out QSAR analysis backed by a simple formal ontology to determine whether a given molecule is drug-like. Further,

we discuss parameter-based control over the execution of SADI SWS. Finally, we demonstrate the value of computational resource envelopment as SADI services through service reuse and ease of integration of computational functionality into formal ontologies.

We conclude that envelopment of chemical computational resources as SADI SWS facilitates interdisciplinary research by enabling the definition of computational problems in terms of ontologies and formal logical statements instead of cumbersome and application-specific tasks and workflows.

4.2 Introduction

The introduction and subsequent widespread availability of computers in the latter half of the 20th century has had an enormous impact on chemistry and related sciences. A wide range of problems which could only be addressed by tedious manual or semi-automated computation a few decades prior suddenly became readily accessible with computers. The explosion of the diversity of the various software packages addressing every aspect of chemistry that followed can only be compared, in relative terms, to the Cambrian explosion in species diversity. Myriads of file formats, programming languages, platforms, operating systems, programming paradigms, distribution models, and access methods have been employed in hundreds of largely-independent projects, each vying for widespread adoption and often offering a unique set of functionalities and features to target a specific subdomain or application of chemistry. Consequently, computational life scientists are now obliged to spend considerable efforts on software package integration to make any progress in their daily investigations.

This problem has been especially acute for interdisciplinary studies, perhaps rising in relevance and importance with the relatively recent rise of Systems Science to prominence. For instance, to build a simple ordinary differential equation-based model of a system of partially enzyme-catalysed reactions, one may need to generate the three-dimensional structures of involved molecules, compute their energies of formation and solvation, approximate pKa values, evaluate enzyme interactions, predict kinetics, and finally solve kinetic equations, all in different software packages, which might be located on different operating systems or have unique execution procedures known only to the experts of these packages. Even practitioners of narrower specialities are not spared the

wrath of software integration, albeit on a smaller scale. Although tools to interconvert the output and input files for many of these software components have been developed [1,2], and although a number of chemical packages offer access to their functionalities through programming interfaces [e.g. 3,4,5], one is left wishing that researchers in chemistry-related fields could still do more science rather than pipelining.

As scientific publishing accelerates and high-throughput experimentation platforms become increasingly pervasive, the problem of integration of the disparate computational, literature, and experimental resources is transformed from that of removing a daily nuisance to that of finding a solution without which science cannot move forward effectively. With the introduction of computational web services for life sciences [e.g. 6, 7], a step in the direction of addressing this problem has been made. With web services, tasks can be posted directly to computational resources with job execution instructions that conform to service-specific schemas, usually defined with a standard specification, most prominently the Web Service Definition Language (WSDL) [8]. Given sufficient knowledge of the service schemas, it is technically possible to automate workflow construction and provide seamless integration of web service components to fulfil a greater overarching task. In practice, however, the lack of shared and consistent schema elements with formal semantics has severely limited this integrative potential due to difficulties of automatically integrating service schemas themselves.

The next step of the evolution of web services was reached with the adoption of the Semantic Web and the corresponding development of Semantic Web Technologies to enable not only machine-understandable knowledge representation, but also the exposition of this knowledge and underlying concepts to automated, formal logic-based

reasoning. Given a collection of knowledge triples that utilize types and relations from a formal ontology, it has finally become possible to automatically classify, integrate, and interconnect entities and concepts, much like a human expert would. To truly capitalize on this potential, simple XML-based approaches in resource specification and annotation would have to make way for Resource Description Framework (RDF) [9] and Web Ontology Language (OWL) [10] to enable automated integration of static knowledge resources with computationally generated information and provide results for cross-domain queries. This ability is indispensable in the life sciences domain to address interdisciplinary problems in toxicology or metabolism, for example. For such problems, it is not only often the case that no single database contains all the information necessary to build a working model or formulate trustworthy predictions, but it is also true that much database information is fragmented and often incomplete. Some of this missing information could be computed to fill in the gaps preventing integrative model construction, but relevant computational resources, many of which are web services, remain inaccessible to a single query method, partly due to the aforementioned integration issues. Although large collections of chemical data have recently become represented in RDF and exposed to SPARQL querying [11,12,13], seamless and facile integration of computational resource output to enable query completion has been difficult to attain with currently existing technologies.

Early solutions proposed for automated service integration in life sciences often drew on elements of Semantic Web Technologies. With Semantic Annotations for WSDL and XML Schema (SAWSDL), it has become possible to annotate WSDL documents with terms from formal ontologies, in a process termed 'lifting' to enable a greater degree of

resource integration than with simple XML service specifications [14]. Services thus annotated would then be more readily available to integration from a central service registry. However, tight integration into the Semantic Web that would allow natural formal reasoning over such services has not yet been adequately addressed, with SAWSDL and WSDL services often requiring adaptors for service integration. The Web Service Modelling Ontology (WSMO), on the other hand, aims to construct and support a complex framework, implemented in Web Service Modelling eXecution environment (WSMX), whereby web services and all aspects of their behaviour are formally semantically represented [15].

Within the life sciences web service domain, numerous practical solutions to web service integration have been proposed, but are too numerous to completely discuss here [16]. Often, these solutions focused on the construction of a common domain ontology, vocabulary, or registry to improve service annotation and discovery in a given domain [e.g. 17,18]. More recently, service frameworks that relied on more general ontologies for service annotation and input/output specification have also become available [19,20]. Further, frameworks integrating REST service discovery and ontology-assisted workflow composition have also appeared recently [21]. Implicit in these approaches has been the need to adopt and adhere to common service annotation, input and output specifications, or common domain ontologies. This has meant that although some of these SWS initiatives have enjoyed considerable success in enabling service interoperability and supporting facile manual workflow composition, truly seamless and automated service integration into the Semantic Web was not reached, as it was inhibited by semantic service platforms themselves.

This situation has changed with the recent introduction of the SADI framework [22]. Web services created with SADI consume and produce RDF graphs, operating on instances of input and output classes formally defined in supporting service OWL ontologies. The input class of a SADI service subsumes the output class, as these services are stateless, atomic, and annotative. That is, each service carries out a single primitive function and annotates an instance of the input class with information through a particular predicate. SADI services are also REST-like, in that there is only a standard basic set of HTTP verbs that they may respond to, namely GET and POST. A GET operation on a given service returns its semantic description, while a POST of a well-formed RDF graph to the service initiates service execution and returns the same RDF graph with the annotations created by the service. If a SADI service is computationally-intensive, standard asynchronous execution mechanisms are available.

Unlike its aforementioned predecessors, SADI service specification is extremely simple as it neither imposes nor invents a central schema, ontology, or message structure, using standard web components instead. Because of the formal logical definition of the input class, output class, and the introduced predicate, SADI services can be tightly integrated into the Semantic Web and very naturally reasoned about by a machine client. This combination of simplicity of specification and power of formal reasoning allows SADI web services to be seamlessly integrated into SPARQL queries with simple machine reasoning clients, as if the data that they can potentially generate was already available in an RDF triple store. Essentially, this allows the various web services and resources to be treated like a set of LEGO blocks, and naturally fitted together to accomplish a single goal, without human intervention.

One such prototype client, Semantic Health And Research Environment (SHARE), operates on SPARQL queries and is capable of reasoning about the desired overall query goal and chaining services and information together such as to reach this goal in the least computationally expensive way [23]. SHARE draws on a central freely accessible service registry that contains information about service input, output, and predicate types to carry out this automated workflow construction. Thus, in order to pose a query through SHARE, a human agent has to be aware of service specification details in the registry to be able to create a well-formed query. For this, the user needs to acquaint themselves with the input and output classes operated upon and annotations created by the collection of SADI services available in a given SADI instance by perusing the service registry [24]. Thus, one only needs to identify an input class that contains the information that is already available as a starting point, as well as the service-specific predicates corresponding to the annotation that is desired (Figure 4-1).

Figure 4-1. Conceptual overview of SADI functionality: thanks to their semantic specification, web services can naturally fit together through a machine agent to complete a given integrative query.

In order to maximize service interoperability, it is of course recommended to reuse concepts and classes as much as possible by adhering to upper-level domain ontologies, but this is not a requirement in SADI and concepts can be manually mapped to those appearing in supporting service ontologies, if required. By the virtue of allowing external formal ontologies to be referenced in SPARQL queries executed by SHARE, this approach provides support for discourse in science while disambiguating and explicitly highlighting the points of disagreement. For example, a small molecule according to one researcher may be one that is no heavier than 500 Daltons, while another may insist that number to be 750 Daltons. With OWL, both of these viewpoints may be represented with an explicit specification which can then be used to classify a set of molecules automatically using the same SADI services, and filter down into further assertions and reasoning seamlessly. Further, multiple researchers may construct ontologies that may

model molecules differently. So long as a formal logical mapping can be either inferred or directly made to concepts used by the service ontologies, this difference can be accommodated and the construction of computational workflows may be initiated. In other words, SADI enables liberation from conformity in existing consortia-generated ontologies and facilitates discourse and disagreement which drive science forward, at the cost of making the end user aware of the ontologies used by the existing SADI services.

SADI, with supporting machine reasoning clients, also enables the conversion of ontologies into workflows. As we have seen, a formal definition of a small molecule as having a molecular mass descriptor within a particular range in an external ontology will trigger the execution of an appropriate service through a SPARQL query posted to SHARE, if no such data is already available, and the input and output classes the service operates upon are consistent with the external ontology. Therefore, integrative workflows (in e.g. molecular classification) can be constructed just by redefining the problem in terms of classifying a given entity into a particular formal ontology-defined class or a set of classes of interest in a given study (Figure 4-2). Furthermore, because computational tasks are explicitly specified and service invocation is controlled, SADI allows for a greater reproducibility and interoperability of computational analysis.

- ▶ **submolecule**
- ▶ **molecule**
 - ◊ **organic molecule**
 - ◊ **substrate**
 - ◊ **product**
 - ◊ **target**
 - ◊ **ligand**
 - ▶ **molecular regulator**
 - ▶ **signal transducer**
 - ▶ **ionomer**
 - ▶ **small molecule**
 - ◊ **oligonucleotide molecule**
 - ◊ **ribonucleoside molecule**
 - ◊ **ribonucleoside molecule**
 - ◊ **lipid nucleoside molecule**
 - ◊ **nucleoside molecule**
- ▶ **weakly connected entity**
- ◊ **zwitterion**
- ◊ **functional substance**
- ◊ **antigen**
- ▶ **ion**
- ▶ **material aggregate**
- ▶ **material boundary**
- ◊ **spatiotemporal region**

Query results

52200 600

In this work, we shall describe and discuss the exposition of a range of computational chemical resources as SADI services and their resultant amenability to seamless automated semantic integration through ontologies, SPARQL queries, and with graphical interfaces.

4.3.1 Exposing CDK QSAR Functionality with SADI

164

services may envelop more than one basic method in a given API or software. One limitation that arises as a consequence of the annotative nature of SADI services is the requirement to support transformative functionalities by differentiating the input and output, even if both are of the same class in the software package. For example, a method that removes hydrogen atoms from a given input molecule specified by a SMILES string and returns a SMILES string (or void), would have to be converted into a SADI service that operates upon an input class that consists of molecules that have a SMILES string specified and produces output typed to a class that consists of molecules that are annotated with a SMILES string as well as a hydrogen-free SMILES string. It must be noted that SADI services do not have to return all of the information present in the specification of the annotated entity that is a member of the input class. For a member of the output class of a service whose input is a molecule that has a SMILES string, it is possible to introduce only the new service-generated annotations on the existing input entity (present at a dereferenceable URI) in the output. Thus, in our hydrogen depletion example, only a reference to the input entity and the new annotation need to be included in the output, without the need to include all of the inherited entity attributes.

Figure 4-3. In principle, classes and methods in APIs can, with some adjustments, be converted to input/output classes and functionality-encapsulating services. Note that since SADI services are annotative, the input class subsumes the output class which merely contains the extra annotation computed by the service.

In this study, we have exposed a portion of QSAR descriptor calculating functionality, to calculate descriptors starting from a SMILES string molecular specification, as implemented in the Chemistry Development Kit. Because we are primarily concerned with demonstrating the envelopment of functionalities of CDK in this work, we have chosen to distribute the calculation of every available descriptor as a separate web service, even if such services relied on multiple methods and classes in CDK (Table 1).

The second step in service creation is formal description of the input and output classes in a service ontology. For our whole set of QSAR descriptor services, we have created a single service ontology [25], extended from the CHEMINF ontology [26] that contains concepts relevant to formal specification of chemical information in general and descriptor information in particular. The reuse of concepts and relations from widely

accepted higher-level ontologies in CHEMINF translates into greater integration of service input and output classes into cross-domain queries.

For each descriptor calculating service, the input class is a smilesmolecule which is formally defined as the following.

```
molecule and 'has attribute' some
    ('SMILES descriptor' and 'has value' some string)
```

This input specification assures that the service will receive and operate upon an entity of the type molecule that has a SMILES descriptor and that this descriptor has a string value which the service can parse, transform into a molecular graph, and for which it can carry out descriptor calculations with a given API, in this case CDK. Note that the terms molecule, has attribute, and has value are reused from upper-level ontologies, meaning that concepts introduced in third-party ontologies constructed using the same upper-level ontologies will be much easier to integrate with than if we were to invent our own terms. The input is an RDF-XML graph submitted to the service through a simple HTTP POST to the service URL (Listing 1).

Listing 1. A fragment of the RDF input graph for the CDK descriptor services, in N3 form.

```
@prefix ss: <http://semanticscience.org/> .
@prefix sio: <http://semanticscience.org/resource/> .
ss:Ethanol rdf:type sio:SIO_011125 .
ss:Ethanol sio:SIO_000008 ss:EthanolSmilesDescriptor .
ss:EthanolSmilesDescriptor rdf:type sio:CHEMINF_000018 ; sio:SIO_000300
"OCC" .
```

Because the input class has to subsume the output class, the output entity has to have all the features of the input entity, but service-computed annotations should decorate the entity in the output. For instance, the definition of the output class *bondcountsmilesmolecule* for a bond count descriptor calculating service [27] is as follows.

```
smilesmolecule and 'has attribute' some
    ('bond count' and 'has value' some int)
```

This class definition specifies that in the output, a given *smilesmolecule* instance will be annotated with a bond count descriptor which would have an integer value (Listing 2).

Listing 2. A fragment of the RDF output graph produced by the CDK bond count calculator service, converted to N3 RDF form.

```
@prefix :
<http://semanticscience.org/sad1/ontology/lipinskiserviceontology.owl#>.
@prefix ss: <http://semanticscience.org/resource/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
ss:Ethanol a :bondcountsmilesmolecule, :smilesmolecule .
ss:Ethanol sio:SIO_000008 ss:EthanolSmilesDescriptor .
ss:EthanolSmilesDescriptor a sio:CHEMINF_000018; sio:SIO_000300 "OCC" .
ss:Ethanol sio:SIO_000008 ss:EthanolBondCount .
ss:EthanolBondCount a sio:CHEMINF_000233 .
ss:EthanolBondCount sio:SIO_000300 "3"^^<http://www.w3.org/2001/XMLSchema#int>
.
```

The service OWL ontology, containing these input and output class specifications, along with the relevant predicate (has attribute) specifications, has to be made distributed such as to assure that these resources have dereferenceable URIs and that the ontology itself is

readily available for machine reasoning agents. In order to expose services for invocation and automated workflow composition with the SHARE client, one needs to also register the service on the central SADI registry. Descriptor information can then be obtained by submitting SPARQL queries that are no different from queries over triple stores that are already populated with RDF knowledge, to the SHARE client. In essence, we are seamlessly querying all the data at our disposal, even the knowledge that does not yet exist, but can be generated. For example, to determine the number of hydrogen bond donors in a given molecule, one may submit the following query to the SHARE client (Listing 3).

Listing 3. A sample SPARQL query to determine the value (specified by the ?value variable) of the hydrogen bond donor count descriptor for a molecule specified in the given (lipinski_test) RDF graph, submitted to a SHARE client [28].

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ss: <http://semanticscience.org/resource/>
PREFIX lso:
<http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl#>
select ?s ?value
FROM <http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl>
FROM <http://semanticscience.org/sadi/ontology/lipinski_test.rdf>
where {
    ?s rdf:type lso:smilesmolecule .    #S is a molecule with a SMILES descriptor.
    ?s lso:hasChemicalDescriptor ?attr .    #S has a chemical descriptor attr.
    ?attr rdf:type ss:CHEMINF_000244 .    #Attr is a CHEMINF H bond donor descriptor.
    ?attr ss:SIO_000300 ?value .          #Attr must have some value.
}
```

This query returns the identity of the input molecule (whose SMILES descriptor is specified in the `lipinski_test` RDF graph), along with the value of its corresponding logP descriptor. In the background, SHARE reasons about the available services based on the information requested in the SPARQL query as well as the information already available in the input graph, and automatically matches the appropriate service or services to the request. Thanks to the formal reasoning carried out by the SHARE client using the service-specific ontologies and any other ontologies referenced in the query, it is possible to infer the services needed to carry out a particular task even if the request does not use concepts identical to those found in the service definition. For example, if a predicate *hatChemischeDeskriptor* can be inferred to be equivalent to *hasChemicalDescriptor* through its formal axiomatic definition, the same set of services shall be called to fulfill queries using either predicate. Thus, the integration of SADI services into the Semantic Web by means of integration into SPARQL queries with SHARE is seamless and requires no additional programming on the part of the life science researcher.

4.3.2 SADI-Enabled Format Interconversion and Software Interfacing

Though ubiquitous in chemical databases, SMILES strings do not address every chemical entity specification need. For example, one may be interested in the three-dimensional configuration of a given molecule directly, or in a more standard and canonical way of representing chemical graph structure with InChI strings [29]. Conversely, the SMILES string needed for our services to operate may not be present, but an InChI descriptor may be available instead. Finally, disparate services may operate on different formats, such that one may produce a molecule specified with an InChI string, while another may need to consume a SMILES string. Clearly, the ability to interconvert a wide range of chemical

formats and representations is essential for wrapping and integrating into a single workflow the functionality of entire software packages that have no exposed programming interfaces, but are accessible for command-line interface scripting.

As a means of demonstrating the format conversion capacity as well as integration of multiple disparate software packages with SADI, we have created an Open Babel (version 2.3.0) based format conversion service to convert InChI strings to SMILES strings [30]. The implementation of this functionality and SPARQL querying for resultant data is virtually identical to that of other descriptor computing services, and is readily accessible, either through the SHARE client or through a direct POST of an RDF graph containing an instance of the *inchimolecule* class (specified below) as input.

```
molecule and 'has attribute' some
    ('InChI descriptor' and 'has value' some string)
```

The resultant output, by virtue of classifying into the *smilesmolecule* class, since it now contains the SMILES string representation of the queried molecule, can subsequently be consumed by all of the QSAR descriptor computing services. A collection of services to convert file formats can therefore be envisioned in order to connect multiple chemical calculation packages together, on the fly.

4.3.3 Lipinski Rule of Five the Semantic Way

The simplicity of SADI architecture allows for natural computational resource integration into the Semantic Web, as demonstrated by seamless service invocation through simple SPARQL queries. The automated computational workflow construction that can be achieved thanks to this tight resource integration can be demonstrated by carrying out simple Lipinski Rule of Five analysis [31]. This well-known rule postulates that drug-like

compounds can be most often characterized as having a molecular mass of less than 500 Daltons, fewer than 5 hydrogen bond donors, fewer than 10 hydrogen bond acceptors, and a logP value between -5 and 5. The definition of a Lipinski-consistent molecule lends itself quite easily to formal representation using concepts from the CHEMINF ontology, as follows.

```
smilesmolecule
    and 'hasChemicalDescriptor' some ('mass descriptor' and 'has value' some
double[<= 500.0])
    and 'hasChemicalDescriptor' some ('hydrogen bond donor count' and 'has
value' some int[< 5])
    and 'hasChemicalDescriptor' some ('hydrogen bond acceptor count' and 'has
value' some int[< 10])
    and 'hasChemicalDescriptor' some ('logP descriptor' and 'has value' some
double[< 5.0, > -5.0])
```

In this formal definition of a drug-like molecule, each statement linking the input SMILES molecule to a particular descriptor conforms to the output class and annotating predicate specification of a corresponding descriptor calculator service. This means that when a SPARQL query is posed to a SHARE client to determine whether a given instance of the smilesmolecule class is drug-like, the client will be capable of identifying and executing the four services necessary for the completion of this query (Listing 4).

Listing 4. A SPARQL query to determine whether a molecule (in lipinski_test RDF graph) is drug-like. If it is the case, the URI corresponding to the matching molecule will be returned.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lso:
<http://semanticscience.org/sad1/ontology/lipinskiserviceontology.owl#>
select ?s
FROM <http://semanticscience.org/sad1/ontology/lipinskiserviceontology.owl>
FROM <http://semanticscience.org/sad1/ontology/lipinski_test.rdf>
where {
    ?s rdf:type lso:lipinskismilesmolecule .
}
```

The overall effect of this is that in the absence of necessary existing data, SHARE creates a web service workflow to complement the information already available, based solely on the formal definition of Lipinski drug-like molecules in a reference ontology. Not only does this lead to improvements in computational workflow reproducibility and concept disambiguation, but it also allows for straightforward means of concept reassessment from within a common framework during the course of scientific discourse. For example, the Lipinski Rule of Five has been extensively discussed, assessed and revised since its introduction [32]. By expressing their alternative definitions of drug-like compounds within the same framework as that of the original Rule of Five, it may have been possible to reduce ambiguity and inconsistencies in results stemming from inadvertent inconsistencies of data sources or precise methods of computational resource invocation. Further, the integration of service-computed data with more involved analysis (e.g. statistical regressions) as well as operations on entire data sets are possible, and demonstrations of this have been discussed at length elsewhere [22].

4.3.4 Mechanisms for Parameter and Computational Experiment Provenance

Specification

A number of algorithms and software packages, which may be wrapped as SADI services, require the specification of one or more parameters. The abundance of algorithm implementations and implementation-specific parameters, coupled with their under-reporting in scientific literature may often result in irreproducibility of computational experiments or discrepancies in research findings and conclusions. SADI services can specify parameters defined in an OWL ontology to control the execution of a specific computational algorithm, or to select a given algorithm from a set of equivalent algorithms which would otherwise be logically equivalent and called either together or at random, depending on the preferences and settings employed by the end user. The service description (using the GET) will display which type corresponds to the parameter class. A user wanting to specify the parameter must do so by adding its explicit description to the input RDF graph. Additionally, the provenance for the data item obtained by running the service is preserved by annotating the output as being the product of a parameterized data transformation. Besides the parameters used, this approach also allows us to explicitly specify the software (and its version), as well as the agent (who executed it). For instance, using CHEMINF concepts, one may construct the following simplified generic output class.

```

molecule that
  'has attribute' some ('descriptor' and 'is output of' some (
    'parameterized data transformation'
    and 'has agent' some 'software'
    and 'has input' some 'smiles'
    and 'has parameter' some (
      'parameter'
      and 'has value' some double)
    )
  )

```

An instance of parameterized data transformation may be placed, along with the input, into the input RDF graph and referred to in the SPARQL query to execute service computational functionality according to explicit, precise, and reproducible specifications. To demonstrate parametric execution capacity, we have created a prototype service to compute a scaled octanol-water partition coefficient value [33]. For some compounds, it may be necessary to apply correction factors to arrive at more accurate predicted logP values. Our service computes a logP value which is multiplied by the value of the scaling factor parameter specified in the input as follows.

Listing 5. A simplified input to the parameterized logP calculating service, converted to N3.

```

@prefix ss: <http://semanticscience.org/> .
@prefix sio: <http://semanticscience.org/resource/> .
ss:parameterX rdf:type sio:SIO_000144 ; sio:SIO_000300 "1.05" .
ss:Ethanol rdf:type sio:SIO_011125 .
ss:Ethanol sio:SIO_000008 ss:EthanolSmilesDescriptor .
ss:EthanolSmilesDescriptor rdf:type sio:CHEMINF_000018 ; sio:SIO_000300 "OCC" .

```

If no parameter is specified, the service has an internally-specified default parameter to fall back on. In both cases, the value of the parameter is reported in the output, and the parameter itself is linked to the process executed in order to obtain the value of the descriptor (Listing 6). Because the output of a parameterized service preserves this provenance information explicitly on the descriptor this service generates, it is then possible to query over only descriptors generated using a particular set of parameters, or with a given software package. This is useful when addressing the construction of toxicological models using data derived from multiple disparate data sources or across chemical entity databases. Finally, this preserved provenance information makes our calculation fully and unambiguously reproducible.

Listing 6. A simplified output of the parameterized logP calculating service, converted to N3.

```
@prefix :
<http://semanticscience.org/sad1/ontology/lipinskiserviceontology.owl#> .
@prefix ss: <http://semanticscience.org/resource/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
ss:Ethanol a :bondcountsmilesmolecule, :smilesmolecule .
ss:Ethanol sio:SIO_000008 ss:EthanolSmilesDescriptor .
ss:EthanolSmilesDescriptor a sio:CHEMINF_000018; sio:SIO_000300 "OCC" .
ss:Ethanol sio:SIO_000008 ss:EthanolBondCount .
ss:EthanolParamlogP a sio:CHEMINF_000251 .
ss:EthanolParamlogP sio:SIO_000300
"1.727"^^<http://www.w3.org/2001/XMLSchema#double> .
ss:EthanolParamlogP sio:SIO_000232 ss:PDTOCCLOGP .
ss:PDTOCCLOGP :hasParameter ss:parameterX .
ss:parameterX rdf:type sio:SIO_000144 .
ss:parameterX sio:SIO_000300 "1.05"^^<http://www.w3.org/2001/XMLSchema#double>
.
```

4.3.5 Integration and Repurposing of Chemical Resources

Service interoperability, even within a single framework, relies on the compatibility of service inputs and outputs. With SADI, formal definition of input and output classes in supporting service ontologies, especially if these ontologies draw on common upper-level concepts, facilitates service integration by enabling class equivalence inference. However, if one service produces output in terms of a molecule that has a SMILES descriptor for example, no conceivable web service framework will magically enable that output to be directly consumed by a service that demands three-dimensional molecular structure specified. In these cases, intermediary services have to be made available to bridge the gap. For example, the existing SADI service to retrieve the KEGG pathways a given drug is involved in, based on a molecule's KEGG Drug identifier [34], would have to be connected to *smilesmolecule* instance-generating services through a KEGG Drug identifier matching service.

If service input/output classes are logically equivalent or compatible, however, no such pipelining services are required to repurpose services for uses not originally anticipated. Consider, for example a functional group annotation service created by us to assist in lipid annotation and classification [35]. Given an instance of a *smilesmolecule*, this service enumerates functional group instances (from a predefined collection) occurring in the input molecule through an upper-level ontology “*has proper part*” predicate and produces a semantic equivalent of a chemical fingerprint in the *annotatedsmilesmolecule* output class. Although this information was originally used to classify molecules into

various lipid classes, we may repurpose it for defining a customized class of chemical compounds: drug-like alkynes, as follows.

```
lipinskismilesmolecule and hasProperPart some Alkyl_Group
```

In order to invoke service execution, one needs to submit a SPARQL query, much like that for the Lipinski Rule of Five use case, to SHARE. The SHARE client will then be capable of inferring not only the necessary services to invoke in order to classify an input molecule into the *lipinskismilesmolecule* class, but would also call on the functional group annotator service to obtain *hasProperPart* annotations and complete the reasoning. Thus, it is possible to build up increasingly complex queries ad infinitum and let the machine reasoning clients take care of the invocation and orchestration of the web services necessary to obtain the information needed to address the query.

It is also easy to imagine a service that enumerates pharmacologically active functional groups working in conjunction with QSAR descriptor computing services to logically select compounds that are predicted to be drug-like and non-toxic, out of a large collection of combinatorially-generated chemical entities. Furthermore, thanks to the ready integration and repurposing of SADI services, it is also possible to combine QSAR descriptors with molecular pharmacological activity data to obtain a formally defined QSAR model as an output of a model creator service that could wrap existing QSAR software or mathematical scripts. Finally, it is worth stressing that due to the simplicity of SADI services, they are not precluded from working with other services, or be described and accessed through other web service frameworks.

4.3.6 Exposing Chemical Database Resources as SADI Services

Web services are not solely limited to tasks involving computational capacities, but can be linked to a range of processes, including those carried out by experimental or industrial platforms. In certain cases, it is advantageous to use web services to encapsulate relational database lookup and the conversion of resultant information to RDF. Although large corpora of RDF data derived from the numerous publically accessible chemical databases have been exposed for querying through SPARQL endpoints [11,13], and although this has a proven potential in facilitating cross-domain querying in chemistry, there are multiple reasons web service-based lookups may be preferable. For example, because the major data providers do not directly publish their information in RDF there may sometimes be a delay in the conversion and incorporation of new data by the RDF triple store providers. Further, not all of the desirable information may be available in the RDF triple stores, or information might be available in a form that makes it difficult or awkward to map to one's own service ontologies, for example.

To preserve the atomic nature of SADI web services and allow for maximal flexibility in workflow construction, it is preferable to encapsulate the lookup of each index-value pair type as a separate web service in a manner identical to that of creating a CDK QSAR service. Here, the input class definition would have to require specification of the index used to look up the database and the service output class would contain entities annotated with the value or values retrieved from the database. Because this task has been demonstrated and implemented elsewhere, we shall limit our discussion of implementation to what is already stated. One point that we would like to observe is that

encapsulation of database lookup functionality as SADI services allows seamless integration of chemical database resources into SPARQL queries even in the absence of corresponding RDF data. In the end, both computational and experimental resources will be available in addressing a given SPARQL query.

4.3.7 Practical Applications of SADI Services in Lipidomics

The development of high-throughput experimentation has led to astronomical growth in biologically relevant lipids and lipid derivatives identified, screened, and deposited in numerous online databases. Unfortunately, efforts to annotate, classify, and analyze these chemical entities have largely remained in the hands of human curators using manual or semi-automated protocols, leaving many novel entities unclassified. Since chemical function is often closely linked to structure, accurate structure-based classification and annotation of chemical entities is imperative to understanding their functionality.

As part of a separate exploratory work conducted in collaboration with the laboratory of Dr. Christopher J.O. Baker, we have investigated the utility of semantic web technologies in automated chemical classification and annotation of lipids. Our prototype framework consisted of two components: a Lipid Eicosanoid Ontology (LEO) and a set of federated web services that operated upon it (Figure 4-4).

Figure 4-4. An overview of the prototype lipid classification and annotation framework.

The formal lipid ontology we used drew on the lipid hierarchy in the LIPID MAPS database, as well as literature-derived knowledge. The federated semantic web services that operated upon this ontology were deployed within the SADI framework. Structure-based lipid classification was enacted by two core services. Firstly, a structural annotation service detected and enumerated relevant functional groups for a specified chemical structure. A second service reasoned over lipid ontology class descriptions using the attributes obtained from the annotation service and identified the appropriate lipid classification. We extended the utility of these core services by combining them with additional SADI services that retrieved associations between lipids and proteins and identified publications related to the inferred lipid types. We analyzed the performance of

SADI-enabled eicosanoid classification relative to the LIPID MAPS classification and obtained a classification accuracy of 96%.

Thus, our prototype framework was capable of accurate automated classification of lipids and facile integration of lipid class information with additional data obtained with SADI web services. The potential of low level programming-free integration of external web services through the SADI framework offers an opportunity for development of powerful novel applications in lipidomics. We concluded that semantic web technologies can provide an accurate and versatile means of classification and annotation of lipids.

4.4 Conclusions

Chemistry is indeed an immense and rapidly growing discipline with a wealth of disparate computational and database resources which are currently largely isolated and inaccessible to truly integrative queries across the entirety of the chemical (deep) web. Thus, we believe that there is an urgent need of exposing chemical resources in a manner that would be conducive to supporting a more productive way of carrying out chemical research. In this work, we have attempted to address this issue by demonstrating what we believe to be the future of chemical service distribution and chemical resource integration into the rapidly expanding Semantic Web. Using our set of SADI services to envelop the CDK QSAR-relevant descriptor functionality to decide whether a molecule was drug-like, we have demonstrated a highly integrative behavior afforded by the simplicity of the formal semantic service specification of the SADI framework. In the future, the widespread adoption of explicit formal specification of computational tasks afforded by Semantic Web technologies may lead to an improved reproducibility and reduced ambiguity of chemical research.

With the Lipinski Rule of Five example of SHARE-assisted automated workflow construction, we have demonstrated the kinds of powerful and natural queries that could be accessible in cheminformatics research if all of the functionalities of CDK were distributed as SADI services. Although the complexity of queries amenable to SHARE automated reasoning is somewhat limited to the capacity of the supporting formal reasoning software and computational resources of the host machine, we believe that with time, this limitation shall diminish to the point of vanishing, as existing reasoners are improved and new ones become available. Engineering limitations aside, provided an ontology of common tasks and a set of adequately specified services, researchers in the future would, in principle, only need to specify their end goal or the kind of information they seek, potentially with natural language queries, and obtain it without having to be well-versed with computational tools, programming, or pipelining. At the same time, parameter-based service control would enable advanced users to express service execution specifics. Intermediate users or those wishing to specify parameters manually or string together SADI services alongside the many other cheminformatics and bioinformatics services would also be able to do this through graphical programming in the Taverna web service interface, using the Taverna SADI plugin [36]. This approach could be applied to computational queries, both big and small, because the SADI framework specifically addresses synchronous and asynchronous service execution modes. This paves the way to integration of more than just database and computational resources into scientific queries, but also potentially to automation of experimentation platforms, similar to the platform deployed for the robot scientist.

Finally, distribution of resources with SADI may act as a form of insurance against computational resources being lost into oblivion as a result of changes in platform popularity or difficulties in porting computational resources across platforms, since SADI services expose a standard, platform-independent interface. Distributing computational capacity as SADI web services in the cloud may become an attractive possibility in the future. In our future work, we intend to significantly expand our collection of web services to envelop all of chemical functionality of CDK, as well as openly accessible cheminformatics and computational packages, potentially in the cloud.

We believe that the amount of knowledge created or creatable in chemistry and related fields on a daily basis has far exceeded the potential of a single human to analyze and integrate information efficiently. In order for chemistry to progress and in order for us to handle these massive and exponentially growing amounts of data, the greater chemistry and life sciences communities have to start exploiting the power of the Semantic Web and deferring some reasoning to machine agents. We believe that the SADI web services, the semantic resource envelopment, and the seamless machine reasoning they enable constitute the first step on our journey to a way of practicing science that transcends disciplines, knows no barriers, and encompasses all human knowledge without taxing the beholder with menial and irrelevant tasks: a self-aware science.

4.5 Methods

4.5.1 Supporting Service Ontologies

We have developed a formal OWL ontology, Lipinski Service Ontology (LSO) to capture the formal definition of service input and output classes, as well as the predicate with which a given service carries out annotations. LSO is a derivative of the CHEMINF

ontology for representing chemical information and chemical descriptors, and relies on an upper level ontology, Semantic Science Integrated Ontology (SIO) [37]. Within LSO, we have defined a single input class for all the services, smilesmolecule, and a large and growing set of output classes to correspond to the output of each service individually.

4.5.2 Service Creation with CDK and OpenBabel

We implemented descriptor calculating functionality based on classes implementing the IMolecularDescriptor interface of CDK, version 1.3.0. Where a descriptor calculation returned multiple results, we created a separate service for each of the results within the descriptor vector thus returned, in order to preserve the atomic nature of SADI services. For cases where a three-dimensional molecular configuration was necessary in order to compute a particular descriptor, we employed the ModelBuilder3D class of CDK. For the InChI-to-SMILES service demonstrating the wrapping of programmatically inaccessible computational capacity distribution, we employed Java system calls to Open Babel [38] (version 2.3.0) from within the SADI service. Although we are aware of the Open Babel API, we have chosen to access the compiled Babel binary from the command line as a means of demonstrating that numerous other command-line tools may be semantically exposed in a similar fashion.

4.5.3 SHARE and SADI Service Distribution

The SHARE client and SADI skeleton for generating services are freely available for download and development. We distributed our services as Java servlets, using the Jetty servlet container. We then registered our SADI services to the central service registry and queried them on the freely accessible public SHARE interface with the queries provided in text. Functionality of SADI web services that are registered in either the central or a

local service registry can also be employed in manually created workflows in Taverna through the SADI Taverna plugin.

4.6 Authors' contributions

LLC wrote the paper and created the demonstration services. LLC and MD created the supporting service ontologies. MD provided guidance.

4.7 Acknowledgements

This research was funded in part by NSERC CGS for LLC and the CANARIE NEP-2 Program for the C-BRASS project. I would like to thank Dr. Mark Wilkinson and Luke McCarthy for helpful discussions on SADI.

5 Chapter: A Distributed Semantic Metabolic Modeling Framework

5.1 Abstract

Semantic Web technologies have so far been shown to be useful in the unification of biochemical information on the level of chemical entities and reactions. We have also seen the application of Semantic Web technologies to the automated classification and annotation of small molecules, as well as computational characterization of these small molecules with precisely-regulated, reconfigurable, and machine-integrable semantic web services. With these elements, I have laid the foundation for the creation of a unified Semantic Web-based framework for metabolic fate prediction and biochemical pathway analysis of arbitrary small molecules.

In this work, I demonstrate a prototype framework consisting of just a few semantic web services that mimic the behavior of enzymes. As analogues of natural enzymes, these *in silico* enzymes operate on semantically-defined classes of chemical entities, and produce an RDF-encoded document containing the description of the nature and dynamics of a given transformation. Just as demonstrated with the SADI web services presented in the previous chapter, these enzymatic web services are interoperable, widely distributable to the Web, and can be adjusted to carry out computations of various levels of complexity and cost, given future improvements in software or hardware. I show that the process of identifying the appropriate web services for a particular classified small molecule can be iterated to produce a full kinetic description of a metabolic pathway. The resultant pathway description can then be used to construct and solve an ordinary differential equation-based model of the metabolic pathway resulting from the metabolic fate prediction study, allowing us for the first time to observe the dynamics of toxicological effect of small molecule metabolism. Finally, I demonstrate the application of

toxicological classification of arbitrary chemical entities using the semantically-encoded decision trees we have seen earlier to produce an overall assessment of the toxicity of the starting small molecule for which metabolic fate is being predicted. Unlike the existing approaches, this will allow researchers to dynamically assess the relative importance and the actual impact of small molecule modifications on chemical toxicity – an important asset, especially in the pharmaceutical industry.

I conclude that Semantic Web technologies can be used to create a metabolic fate prediction framework that overcomes many of the barriers currently present in metabolomics research, and can thus one day lead to improved predictions of chemical toxicity and activity.

5.2 Introduction

Despite the fact that there has been a large array of algorithms and approaches to enable the various components of MFP, as discussed in Chapter 1, there has been no effort to incorporate the strengths of each approach into a single pathway prediction system. The lack of interoperability of databases relevant to MFP, such as the biotransformation pattern databases, further complicates the task of researchers wishing to undertake comprehensive MFP studies. Without layers of software pipelining, scientists are needlessly occupied with manual knowledge and computation integration for metabolic pathway model construction. More importantly, crucial information, such as biotransformation patterns contained in databases that are not immediately accessible to a given researcher, or existing toxicological information buried behind an arcane database interface, could be overlooked. Although a number of web service architectures and data distribution models have appeared over the past decade to help alleviate this problem (see Section 4.2), none have quite managed the automated web service and database integration demonstrated by the SHARE client within the SADI web service framework. Before we could begin addressing MFP with SADI, we needed to establish Semantic Web-based solutions for overcoming barriers to the unification of chemical information and resources on the representational, conceptual, and computational levels. With CHESS, we were able to represent a wide range of chemical entities from atoms to reactions and their precise composition and structure. In addition to this, thanks to the use of CHEMINF, we were able to develop a standardized way of annotating chemical entities with information relevant to their subsequent classification and use in further investigations such as QSAR studies and reaction feasibility calculations relevant to

MFP, for example. Conceptual barriers to integration of chemical information were addressed by automatically generating machine-understandable, unambiguous definitions of classes of small molecules, be they structural or functional, within chemical OWL ontologies. Finally, computational interoperability barriers were erased in this work with the semantic distribution of computational functionality with the SADI framework and machine reasoning and web service orchestration by the SHARE client.

With these components in place, it is now possible to start tackling the problem of integrative MFP for arbitrary small molecules. One may recall that many known enzymes operate on a well-defined class of substrates and catalyze a very well-defined set of general biological transformation patterns. Thus, it is possible to define enzymatic inputs semantically, using the same methodology as that developed in Chapter 3 for chemical classification. The ability to formally define enzyme substrate classes translates into the ability to use such formal definitions in semantic web service input descriptions, and this logically leads to the possibility of construction of semantic web services that mimic the behavior of enzymes, turning the World Wide Web into an *in silico* biochemical reactor.

Furthermore, since all enzymes obey the same fundamental laws of physics that govern the universe, it is also, in principle, possible to employ first principles-based approximations to compute the catalytic and binding constants for each enzymatic reaction. In theory, given a sufficiently advanced computer, it is possible to accurately compute the reaction energy profile and obtain reaction kinetics. Practically, such calculations may either be impossible due to missing enzyme structure and lack of ability to obtain a reasonable homology model, or are infeasible due to prohibitive computational cost for all but the very simplest of enzymatic systems. In such cases,

approximations are necessary in order to compute the relevant information reasonably quickly.

One may identify two extremes in integrating the computational resources needed to carry out MFP on the semantic web: a process-centric approach, and an enzyme-centric approach. In the process-centric approach, the procedure of MFP can be subdivided into the basic functionalities which can then be distributed individually as semantic web services. For instance, the identification of candidate reactions and transformation of the input molecule according to the transformation rules stored within a knowledgebase could be handled by a 'reactor' service, calculations of reaction thermodynamics, kinetics, and feasibility could be handled by a 'computational chemistry' service, and the approximation of the enzyme-ligand binding constants could be handled by a 'docking' service, for example. This approach could potentially produce broadly reusable, universal services, and while being a major step forward in terms of enabling computational resource interoperability, it would be highly reminiscent of the current practices in MFP prediction. The second approach would rely on the construction of thousands of individual enzyme-mimicking services that would each have a chemical ontology-defined class of input molecules, and each producing a fully characterized CHESS-encoded reaction description. Each service would therefore contain the reactor, thermodynamic calculation, and kinetics approximation functionalities embedded within. While the enzyme-centric approach highlights the potential of semantic resource integration in somewhat more accessible terms, the process-centric approach provides broadly applicable and reusable services. We combine the two approaches to distribute the 'reactor' functionality with enzyme and reaction mimicking services and the assessment

of thermodynamics and kinetics with process-centric services to arrive at an integrated framework for metabolic fate prediction (Figure 5-1).

Figure 5-1. An overview of the integrated prototype semantic MFP framework.

In this work, we demonstrate the semantic distribution and subsequent concerted action of these services for MFP of benzene, with a very limited set of both enzymatic and non-catalyzed transformations. Resulting from this is a dynamic model of benzene metabolism which allows us to supplement the information resulting from the assessment

of the toxicity of all of the chemical intermediates in this metabolic pathway with their bioaccumulation information. We argue that this helps pinpoint the metabolites with the greatest impact on small molecule toxicity, and therefore to identify potential molecular sites that could be modified to avoid a particular metabolic route. Finally, we discuss how the integrated MFP framework presented here can be easily extended to include better predictive models as well as a more comprehensive compendium of biological and chemical transformations, and could pave the way towards more accurate and more facile model construction in the future.

5.3 Methods

5.3.1 Assumptions

Before progressing much further with the description of the MFP framework, we need to define the scope and assumptions for this study. The aim of this study is to demonstrate semantic service interoperability in support of the MFP task. We leave the selection, optimization, and streamlining of the precise computational methods employed in the individual components of the framework for the broader scientific community and researchers specializing in the relevant fields of computational chemistry.

Therefore, we make a vast range of simplifying assumptions that allow us to demonstrate the integrative use of computational chemistry and cheminformatics web services in addressing MFP. The first assumption is that absorption and distribution plays no role in the accurate prediction of a compound's metabolic fate. Thus, both, the metabolites and the enzymes exist in a well-stirred aqueous solution - a gross oversimplification of the human body. Secondly, because the development of rapid and accurate algorithms for the prediction of arbitrary reaction kinetics is outside of the scope of this study, we simplify

this task by making a series of assumptions that help us approximate reaction kinetics. It must be noted, however, that the methods we employ are still more flexible and accurate than the methods employed in a number of metabolic pathway reconstruction studies, and many integrated MFP packages, most of which rely on human expert-created rules to determine reaction feasibility. Thus, we assume that all enzymes in this study obey irreversible Michaelis-Menten kinetics.

To predict reaction kinetics, it would have been ideal to have rapid reaction transition state energy calculations. However, because we are not aware of any package that carries this out within reasonable time (seconds), we use an approximation that is inspired by the Evans-Polanyi principle [1]. This principle states that for homolytic atom transfer reactions of molecules that lie within a single family, the activation energy barrier is proportional to the overall reaction enthalpy change, formalized as follows.

$$E_a = E_0 + \alpha\Delta H$$

Here, E_a is the activation energy for a given reaction, E_0 is the activation energy of the reference reaction, α is a constant, and ΔH is the enthalpy change for a given reaction. In our reaction kinetics approximation, we assume that the activation Gibbs' free energy, necessary in calculating the Michaelis-Menten k_{cat} parameter (as shown in Section 1.4.3), can be approximated in a similar fashion. This is not entirely unreasonable because reactions for a given enzyme all belong to the same family, and substrates fitting within the enzyme cavity and containing active, transformable moieties, can undergo a given reaction with a promiscuous enough enzyme, which the Phase I and II enzymes and many other metabolic enzymes generally are. In our approximation, we depart somewhat from

the form, but not the spirit of the Evans-Polanyi principle, and use the following formula for enzymes that have sufficient information in the training set.

$$\Delta G^* = \Delta G_0^* + \alpha \Delta G_r$$

In this equation, ΔG^* is the activation free energy for the evaluated reaction, ΔG_0^* is the difference between the activation free energy and the overall reaction free energy change for a reference reaction, and ΔG_r is the overall reaction free energy for the evaluated reaction. The factor α is a constant parameter. In other words, we claim that it is possible to decouple the reaction free energy from the activation free energy, and to use the difference to predict activation free energies for arbitrary reactions within a given family (Figure 5-2).

Figure 5-2. Activation free energy could be decomposed into a reaction free energy, and the term that describes the activation free energy of the reverse reaction.

When sufficiently homogeneous data is used, it is possible to appreciate that this relationship is not wholly unrealistic for certain promiscuous enzymes, such as alcohol dehydrogenase (Figure 5-3), albeit more studies need to be undertaken to demonstrate unambiguously that it is realistic, as a seven-point correlation for one enzyme is barely a good indicator of universality. Nonetheless, this is our assumption. By taking into account the fact that k_{cat} is essentially the reaction rate constant for the transformation of substrates to products with an enzyme catalyst, we can compute the reaction free energy for enzymes that have a k_{cat} value reported, using the following equation (k_B is the Boltzmann constant, T is temperature, h is Planck's constant, and R is the gas constant) [2].

$$k_{cat} = \frac{k_B T}{h} e^{-\frac{\Delta G^*}{RT}}$$

Figure 5-3. A sample of the correlation between the enzymatic activation free energy ΔG^* and experimentally determined reaction free energy ΔG^0 for the enzyme alcohol dehydrogenase (E.C. 1.1.1.1), data from BRENDA and [3].

For cases where no such correlations can be drawn due to limited data available, we shall assume that α is 1, and simply add the ΔG^{**} term obtained for the reference reaction to the computed reaction free energy for the studied reaction. We shall also assume that gas-phase free energies can be used for these calculations, in order to save computational time. Finally, we also assume that no enzyme is faster than acetylcholine esterase, and we set this ceiling at $1.2 \times 10^4 \text{ s}^{-1}$ [4].

To approximate binding parameters, we assume that K_M can be approximated by taking into account substrate similarity – a guiding principle in QSAR studies, further discussed below. Finally, we shall assume that once a compound undergoes a Phase II modification

and is marked for excretion, it disappears from the system, and that happens with irreversible mass-action kinetics with an arbitrary uniform rate constant for all substrates. We make no claims with respect to the accuracy of these approximations, and use these approximating assumptions solely in order to demonstrate the distribution and integration of computational chemistry functionality packages, as placeholders for much better suited approximations and computational approaches, which can be swapped in for ours within the framework we propose at any point in the future, harnessing the integrative potential of our framework.

5.3.2 Modeling Enzymes and Reactions as Web Services

To model the enzymatic reactions as web services, we first derived the input chemical class definitions for each enzymatic reaction, based on a training set of at least five substrates for each reaction. We then used the approach reported in Chapter 3 to identify the formal structure-based chemical class definitions for each substrate class and to represent these substrate classes in terms of OWL axioms. Each compound encountered in the process of MFP was subjected to the structural classification process outlined in Chapter 3 into the newly derived reaction-specific classes. These classes were specified as the input classes for each reactive web service. Where available, automatically derived reaction rules from the KEGG database were converted into reactive SMARTS definitions, and where such rules were unavailable, they were generated manually based on supporting literature. Graph theoretic biotransformation computation was carried out using these definitions by interfacing with the RDKit software, version 2011.06.1, Python 2.6 edition [5]. In cases where multiple possible products were identified, only unique products were selected and multiple reactions were produced in the output of the reactive

web service. Enzymatic reactions were also annotated with the enzyme's UniProt ID [6], the native (experimentally observed) substrate, as well as the K_M and k_{cat} parameters for this substrate. Non-enzymatic reactions were annotated with a reference activation Gibbs free energy and an alpha parameter. These annotations were necessary for further processing by our kinetics approximation services, but could be wholly ignored by web services using only the enzyme and substrate structures for full-fledged reaction energy profile calculations, for example (Listing 5-1).

Listing 5-1. A simplified output of a reactive web service.

```
@prefix : <http://semanticscience.org/sadi/ontology/enzymeontology.owl#>.
@prefix ss: <http://semanticscience.org/resource/> .
@prefix sio: <http://semanticscience.org/ontology/sio.owl#> .
ss:MoleculeX a :smilesmolecule .
ss:MoleculeX sio:'is input in' ss:ReactionX .
ss:ReactionX a sio:'organic reaction' .
ss:ReactionX sio:'has product' ss:MoleculeY .
ss:MoleculeY a :smilesmolecule .
ss:MoleculeY sio:'has attribute' ss:SmilesDescriptorY .
ss:SmilesDescriptorY a sio:'SMILES descriptor' .
ss:SmilesDescriptorY sio:'has value' "SmilesString" .
```

These services were distributed in a local SADI service registry. To show automated web service invocation, they were queried with a local SHARE client, with SPARQL input that contained the functional group annotations for each input compound. The share client inferred class membership of the input molecule and assigned it to transformations with the appropriate reactive services (Listing 5-2).

Listing 5-2. SHARE client SPARQL query to identify the appropriate reactive web services for a given substrate, defined in moleculeDefinition.rdf.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX eo: <http://semanticscience.org/sad1/ontology/enzymeontology.owl#>
PREFIX sio: <http://semanticscience.org/ontology/sio.owl#>
SELECT ?substrate ?reaction
FROM <http://semanticscience.org/sad1/ontology/enzymeontology.owl>
FROM <http://semanticscience.org/sad1/ontology/moleculeDefinition.rdf>
where {
    ?substrate rdf:type eo:smilesmolecule . #Identify the substrate.
    ?substrate sio:'is input in' ?reaction . #Identify appropriate reaction.
    ?reaction rdf:type sio:'chemical reaction'. #That is a chemical reaction.
}
```

5.3.3 Computing Transformation Thermodynamics

To predict the thermodynamic feasibility of a given reaction, we have developed two generic services: one for enzymatic transformations, and one for non-enzymatic ones. The non-enzymatic transformation feasibility service takes instance of SIO 'organic reaction' class with complete reaction definitions as input, as well as the parameters necessary for our approximation and produces thermodynamic and kinetic annotations on them in the output. In terms of thermodynamics feasibility calculations, both service exhibit the same behavior.

Reaction substrates and products are identified by parsing the input RDF with the Jena API to obtain the SMILES strings corresponding to each substrate and reactant. Using the Chemistry Development Kit, we then generate initial three-dimensional coordinates with a Molecular Mechanics Force Field approximation. These coordinates are then used to

prepare input to standalone MOPAC 7.1 software [7], which optimizes geometries and computes free energies after its scripted invocation on the input. The output is parsed, and free energy for the reaction is computed by subtracting the free energy of formation of substrates from the free energy of formation of products.

Finally, to demonstrate the extensibility of our framework for future improvements in algorithms and hardware, a service to evaluate reaction thermodynamics based on Gaussian 03 software was also created. This service takes a reaction specification as input and prepares it in much the same way as described above. The difference is the use of Gaussian 03, functionality instead of MOPAC 7.1 functionality. In this service, we use a modification of an earlier described method for computing reaction free energies. Molecular geometries are optimized with the AM1 method, vibrational frequencies are calculated at the AM1 level as well, in gas-phase, and at a temperature of 298K. Free energies of solvation are calculated in Gaussian 03 [8], with the CPCM method and water solvent (relative static permittivity of 78.3553), using Density Functional Theory with the Becke, three-parameter, Lee-Yang-Parr (B3LYP) exchange-correlation functional, with the 6-31G(d) basis set. The overall aqueous free energies of formation of each molecule were calculated by summing the free energy of formation, and a correction factor for the transfer from gas-phase to solution phase was added.

All these services are parameterized services and use a parameter of type '*algorithm to calculate a chemical descriptor*' (CHEMINF 000144) from within the CHEMINF ontology to precisely control the algorithm used for free energy calculation, as described in the section 4.3.4 in greater detail.

5.3.4 Approximating Kinetics

Reaction kinetics approximations for the non-enzymatic reactions are produced similarly to the procedure of producing kinetics to determine the k_{cat} parameter, using the principles described in Section 5.3.1, except for the fact that we are obtaining the reaction rate constant for a mass action kinetics expression. That is, for an irreversible reaction that converts A to B, mass action kinetics state that the rate of B formation is as follows.

$$\frac{dB}{dt} = kA$$

It is this k parameter that we approximate for non-enzymatic reactions.

The K_M parameter is estimated by computing the Tanimoto similarity score for the similarity of the prospective substrate with the native substrate, using the Chemistry Development Kit Java API version 1.2, according to the following equation.

$$Tanimoto = \frac{AB}{A + B - AB}$$

Here, A is the number of chemical features in molecule A, B is the number of chemical features in molecule B, and AB is the number of common features in the two molecules.

As stated earlier, the following equation describes the K_M coefficient.

$$K_M = \frac{[E][S]}{[ES]} = \frac{(k_{-1} + k_2)}{k_1}$$

Assuming that k_2 is negligible with respect to k_{-1} , which is one of the basic assumptions of Michaelis-Menten kinetics, and taking in consideration the thermodynamics of the enzyme-substrate binding reaction, we can obtain the following expression.

$$K_M = \frac{[E][S]}{[ES]} = \frac{k_{-1}}{k_1} = e^{\left(\frac{\Delta G_{diss}}{RT}\right)}$$

Here, ΔG_{diss} is the enzyme-substrate complex dissociation free energy, R is the gas constant, and T is the temperature. To obtain the K_M value for a given arbitrary compound, we obtain the product of the value of ΔG_{diss} and the Tanimoto similarity coefficient and substitute the resultant value in the equation above. In other words, we claim that, because all the compounds entering a particular biotransformation are already within a single family, they are likely similar enough to permit the scaling of their free energies according to their precise similarity, and that we would not often observe chemical substitutions that could entirely make this assumption invalid.

Finally, all elimination reactions have been arbitrarily set as mass action irreversible reactions, with a rate constant of 0.1 s^{-1} . All reference data was obtained from BRENDA and [3].

5.3.5 Toxicity Evaluation

Toxicity of the individual compounds was evaluated using a toxicity approximation service that uses the ToxTree API, version 2.1.0, to enumerate chemical features of each compound and to represent these features in CHESS. Each compound is then evaluated using an OWL-encoded toxicity decision tree based on the Cramer decision tree into one of three categories: high, moderate, and low toxicity compounds. Due to difficulties of encoding the Cramer decision tree itself (associated with tree complexity) using the procedure described in Chapter 3, the decision tree used here was created by computing the Cramer toxicity classes for 30000 randomly chosen organic compounds with a mass

less than 300Da from the PubChem database and training a decision tree using this information as described in Section 3.5.

5.3.6 Dynamic Model Construction

All the reactions resulting from reactive web service transformations, as well as the annotations introduced by the kinetics approximation services were dynamically collected into a single RDF document while iterating through the reaction prediction stages, and converted into an SBML file, which was then used to simulate the predicted pathway dynamically with the COPASI Java API, version 4.6 [9]. The starting concentrations for all chemical species are zero mM, except for the starting compound, which the starting concentration is 10 mM. All enzymes are assumed to have a constant concentration of 0.1mM.

In order to determine the overall toxicity levels present throughout the duration of small molecule metabolism, we assign a factor to each toxicity category, 0 for low toxicity, 0.5 for moderate toxicity, and 1 for high toxicity, and use the sum of the products of these factors and their corresponding species at a given time point in order to report the overall toxicological impact of all of the compound's metabolites at a given time point.

5.4 Results

5.4.1 Integrated Metabolic Fate Prediction Procedure Overview

The MFP procedure with the proposed framework is semi-automated and scripted, but the appropriate services for biochemical transformation identification and reaction parameter calculation are located automatically from within the SHARE client. Thus, the procedure begins with the semantic description of the starting compound. The compound is drawn by the user, its SMILES string is identified, and it is encoded in CHESS. In

addition, CHESS descriptions are added to identify and enumerate the functional groups that the starting compound contains, from the list of functional groups in the functional group ontology specific to the previously developed enzyme substrate classes. This results in a molecule that can be classified into one of the substrate classes (or not), and automatically matched to these services for transformation by the SHARE client, using the query shown in methods. The results of the reactive web services are parsed to enumerate the reactions in the output of these services, and the products contained in those reactions, if any. The procedure is repeated for each product of the transformation, until no further transformations are possible. For example, in the case of a cytochrome P450 2E1 (CYP2E1) service, the input class is specified as consisting of molecules that have at most two alcohol functional groups, thus preventing complete hydroxylation of the benzene ring. Further, the transformation is defined such that two adjacent unsubstituted aromatic carbon atoms are converted to an unsubstituted carbon atom and a hydroxylated carbon atom. Thus, we do not see a hydroxylation of resorcinol to pyrogallol, for example.

After the substrate iteration procedure completes, the thermodynamic and kinetic characterization of each individual reaction begins. In order to select the appropriate service, two things are ensured. Firstly, the annotation from each reactive web service that identifies whether the reaction is enzymatic or non-enzymatic. This allows the SHARE client to select the appropriate thermodynamics and kinetics calculation web service. Secondly, each reaction, with substrates, products, and the relevant annotations produced by the reactive web services is further annotated with the appropriate method parameter, as described in Section 5.3.3. When this parameter references the AM1

algorithm, the less computationally expensive (~ 1 s/molecule) web services are selected, while a reference to the B3LYP-based calculation triggers the more expensive (~ 4 min/molecule) web service.

The output of these services is collected in a single RDF document describing every reaction, and is then converted to SBML. This allows us to then solve the system of kinetically annotated ordinary differential equations with the COPASI software and to obtain a dynamic profile of metabolism. In addition to this, a final characterization of all small molecule participants with respect to their toxicity as a result of their annotation with Cramer decision tree-based descriptors and machine classification with the Pellet reasoner on an OWL-encoded decision tree allows us to identify the toxicity of each individual compound and assess the dynamic toxicological profile of the integrated metabolic pathway.

5.4.2 Biochemical Transformations and Products of Benzene

The reactions we have chosen to model in this study with semantic web services were the hydroxylation of aromatic moieties including benzene by CYP2E1, the auto-oxidation of hydroquinones and catechols (actually a rolled-up and simplified auto-oxidative pathway), the two-electron reduction of quinones by enzymes (we are aware of non-enzymatic two-electron reduction of o-quinones, but assume Quinone Reductase, QR, does it), and the elimination of all the hydroxylated species lacking a ring carbonyl through Phase II metabolism. The transformations identified are shown (Table 5-1, Figure 5-4).

Table 5-1. Selected transformations identified by the reactive web services, along with the approximated constants. Only major substrates and products shown.

Substrate	Product	Reaction	Rate Constant or k_{cat} (s^{-1})	K_M (mM)
Benzene	Phenol	CYP2E1	10.4	227.4
Phenol	Catechol	CYP2E1	1.2×10^4	40.4
Phenol	Hydroquinone	CYP2E1	1.2×10^4	40.4
Catechol	Hydroxyquinol	CYP2E1	13.2	47.6
Hydroquinone	Hydroxyquinol	CYP2E1	32.8	31.6
Phenol	Resorcinol	CYP2E1	1.2×10^4	40.4
Resorcinol	Hydroxyquinol	CYP2E1	35.6	47.6
Resorcinol	Phloroglucinol	CYP2E1	4.6×10^2	47.6
Catechol	Pyrogallol	CYP2E1	24.5	47.6
Phenol	-	Phase II	0.1	-
Catechol	-	Phase II	0.1	-
Hydroquinone	-	Phase II	0.1	-
Resorcinol	-	Phase II	0.1	-
Hydroxyquinol	-	Phase II	0.1	-
Phloroglucinol	-	Phase II	0.1	-
Pyrogallol	-	Phase II	0.1	-
Hydroquinone	p-Quinone	Auto-oxidation	26.7	-
Catechol	o-Quinone	Auto-oxidation	0.5	-
Hydroxyquinol	2-Hydroxy p-Quinone	Auto-oxidation	30.3	-
Hydroxyquinol	4-Hydroxy o-Quinone	Auto-oxidation	0.6	-
Pyrogallol	3-Hydroxy o-Quinone	Auto-oxidation	2.2	-
3-Hydroxy o-Quinone	Pyrogallol	QR	2.0×10^3	346.2
2-Hydroxy p-Quinone	Hydroxyquinol	QR	2.4×10^2	346.2
4-Hydroxy o-Quinone	Hydroxyquinol	QR	1.1×10^4	381.3
p-Quinone	Hydroquinone	QR	2.7×10^2	391.7
o-Quinone	Catechol	QR	1.2×10^4	146.4

Figure 5-4. An overview of the predicted metabolic fate of benzene. Compounds classified for some level of toxicity are denoted in bold. Highly toxic compounds included benzene, o-Quinone, 4- and 3-Hydroxy o-Quinone. Moderately toxic compounds included p-Quinone and 2-Hydroxy p-Quinone.

5.4.3 Kinetic Model of Metabolism and Toxicological Assessment

The solved kinetic model of metabolism has allowed us to observe the accumulation and distribution of every product of benzene metabolism over time (Figure 5-5). In addition to this, combined with the toxicological information on the individual products of benzene metabolism, this has allowed us to observe the overall toxicity profile of the process, including the overall impact of the accumulation of the toxic byproducts of benzene metabolism, as well as the trends in the toxicological consequences of this predicted metabolic pathway (Figure 5-6).

Figure 5-5. Dynamic concentration profiles of benzene metabolites. All vertical axes are in the units of concentration (mM), and all horizontal axes denote time (fraction of metabolism completion).

Figure 5-6. Toxicological profile of benzene metabolism, including and excluding the contribution of toxicity stemming from benzene. Toxicity has been computed according to Section 5.3.6.

5.5 Discussion and Conclusions

In this chapter, we have harnessed the power of the semantic web technology-based components developed in the earlier chapters in order to construct a freely extensible, reconfigurable, and integrated framework for metabolic fate prediction. In this framework, the semantic annotation of the computational components works in harmony with the semantic annotations on the biochemical entities to form a workflow of LEGO-like components which produces a dynamic metabolic pathway and a mechanistic explanation of chemical toxicity.

5.5.1 Framework Assessment

The metabolic fate of benzene has been our primary means for validating the functionality of our framework. With just four types of enzymatic and non-enzymatic transformations, we have been able to predict the majority of the transformations described for benzene earlier, and discussed at length in Section 1.2. In addition to this, we have identified transformations that were not described earlier, but that are not necessarily incorrect until proof of the contrary can be provided. Overall, the behavior we have observed for the biochemical transformation part of our framework is unremarkable, in that it exactly mirrors the typical behavior of the monolithic software discussed in Section 1.4.4 – which is exactly what we have set out to demonstrate.

Throughout the process of predicting the metabolic fate, the SHARE machine client querying the services for their capability to carry out a biochemical transformation on a particular enzyme was, in fact, going through a logical process that was very similar to that of a human expert carrying out an analogous task. The machine examined the relevant structural features of the molecule at hand, logically inferred its chemical

classification, and identified the enzymes that operated on molecules of a given class. This is the reason that, for example, we have not observed products of benzene metabolism with more than three alcohol functional groups - the formal definition of the input class for the CYP2E1-mimicking service allowed for compounds with a maximum of two hydroxyl groups.

This procedure is extensible to all accessible services, provided they are registered within a central service registry, no matter which physical machine they reside upon. In this sense, the World Wide Web has just become a biochemical reactor. Unlike the previously described commercial packages for MFP, however, the framework we propose here harnesses the power of any computational package, method, and database that has been distributed as a SADI web service. In principle, it is even possible to distribute certain functionalities of the existing commercial MFP packages in order to supplement and support parts of this process. In this respect, the framework we have proposed can potentially surpass the predictive ability of any commercial framework currently available. Furthermore, the individual components of our framework can also be re-used for procedures and workflows not envisioned or pursued in this study.

On the positive side, a number of barriers preventing integrative research and comprehensive model construction have been struck down with the introduction of reactive semantic web services. In addition to this, the formalization of the specification and annotation transformation rules, identification and characterization of reactions, kinetic constants, algorithms, and biochemical entities paves the way for a more clearly defined, more reproducible, and more machine-computable science.

On the other hand, one may be wary of the erosion of too many barriers – a web as a biochemical reactor may be an interesting idea, but what happens when the reactor needs virtual walls and partitions to model e.g. human metabolism, as opposed to the entire possible metabolic space of all life? Thankfully, the annotation of web services and their parametric execution can help in such partitioning, as demonstrated with the selection of a DFT method over a semi-empirical method with the modification of a single parameter. In fact, service parameters could be used to control for multiple factors, including species of origin of a given enzyme, enzyme's subcellular location, pH of the medium, or the precise enzyme sequence and structure to be used. The importance of parameters is felt especially acutely when considering the fact that many rule-based systems and statistically-based models for predicting the nature and feasibility of biochemical transformations are geared towards the statistical majority of cases, while it is the minority that may have severe adverse effects to a given compound. Parametric service control enables the examination of the impact of such variations.

5.5.2 Predictions of Dynamic Metabolism

Metabolism is a phenomenon that occurs over time, with the dynamic distribution of metabolites playing an important role in the toxicity, activity, and adverse effects of the metabolized compound. Because pathway dynamics are often complex and non-linear, representations of metabolism that merely enumerate the possible biochemical transformations and rank them according to feasibility, are of very limited use. Such representations are only suitable for alerting researchers of the potential formation of a given compound, without any indication of whether that compound would persist or

would be quickly scavenged. Thus, the prediction of reaction and pathway dynamics is a crucial step in forming an adequate toxicological model of the starting chemical entity.

Unfortunately, common experimental methods in metabolomics are sometimes prone to missing metabolites that are either unexpected or transient. A theoretical prediction of a reaction that has not been previously observed, or is unexpected in a given system, but can be characterized thermodynamically and kinetically using first-principles approaches, could help with the detection of such compounds. The predicted reaction may proceed relatively slowly or could be followed by very rapid reactions that degrade its product before it is accumulated to any appreciable levels. However, under certain conditions or in certain situations, such as pH changes, substrate depletion, or subtle genetic variations in key metabolic enzymes, a given reaction or its products may become more prominent and could potentially exert adverse effects. Thus, a dynamic MFP is indispensable to understanding chemical metabolism and toxicity.

In this work, we have used a range of simplifying assumptions to model the kinetics of benzene degradation. Thanks to this, we have been able to identify p-quinone and 2-hydroxy p-quinone as the major contributors of the toxicity of benzene. Interestingly, these compounds have not been identified by the Cramer scheme as the most toxic chemical entities, yet they were clearly the most important contributors to the overall toxicity (in our assessment) due to their appreciable accumulation throughout the metabolic pathway. Clearly, benzene itself was one of the most toxic compounds, and due to the relatively slow conversion to phenol, the most abundant compound throughout the duration of the simulation. Thus, the overall effect of the pathway has been clearly to lower the overall concentration of the toxic metabolites. However, the process of

xenobiotic metabolism sometimes involves toxic activation of the starting compound, and so dynamic toxicity profiles for many other innocuous – looking compounds, like the earlier discussed estrogen analogues, could be an indispensable aid in the prediction of the toxicological effect of these chemical entities. Thus, our framework also provides a means to obtain valuable mechanistic insights into the toxicity and activity of small molecules.

While the development of an accurate dynamic characterization of the metabolic pathways of arbitrary small molecules has not been the principal focus of this study, we feel that this is a goal we have to strive towards. We are acutely aware of the fact that the large number of simplifying assumptions adopted in our approach to the approximation of reaction kinetics translates into a pathway with dynamic descriptors that can barely be trusted, we expect that the ongoing development of novel algorithms for pathway prediction, as well as the ease of envelopment of the existing tools, as demonstrated with the distribution of the functionality of the Chemistry Development Kit, MOPAC, Gaussian 03, and RDKit, shall pave the way to increasingly accurate dynamic assessment of small molecule metabolic fates with improved computational tools running on faster hardware.

5.5.3 Conclusion

With the prototype framework presented here, we have demonstrated that metabolic fate prediction can be undertaken using a first principles-based semantic biochemical reactor. This framework employs the representational framework developed with CHESS in order to automatically classify chemical compounds into semantically defined structural and functional classes, and draws on a collection of self-organizing semantic web services to

carry out an array of reactions on the appropriate compounds and to predict their dynamic distribution. With this framework, it has become possible to mechanistically trace the origins of toxicity of small molecules, and to suggest possible routes for small molecule modification to lower this toxicity. Thanks to the readily extensible nature of the framework, it can be further developed in the future to produce increasingly accurate toxicological predictions, and to provide increasingly rich information on the interactions and effects of the produced compounds. Scientific curiosity and imagination is the limit!

5.6 Authors' Contributions

The study has been conceived and carried out by me, under the supervision of Dr. Michel Dumontier.

6 Chapter: Thesis Summary and Future Directions

6.1 Overview

In this work, I have demonstrated the feasibility of metabolic fate prediction using a first principles-based, semantically distributed biochemical reactor. In Chapter 2, I have proposed, implemented, and characterized a new formalism for the representation, annotation, and query of biochemical knowledge, the Chemical Entity Semantic Specification. For the first time, a framework has been developed that could adequately address the problems of canonical naming of chemical entities and their sub-components and support the facile integration of chemical information across databases. In the third chapter, I have formalized the expert knowledge in the two major areas of chemical classification: structural chemical hierarchy construction and functional chemical classification. With automated structural hierarchy construction, I have demonstrated for the first time a self-assembling and unambiguously defined hierarchy of chemical entities that was trained on the existing annotated chemical information. By training and representing toxicological decision trees as formal Web Ontology Language-encoded ontologies, I have demonstrated a way to classify chemical entities into functional classes and to obtain a formal explanation of these classifications. I then demonstrated the application of both classification approaches by classifying a set of CHESSE-encoded chemical entities and demonstrating high classification accuracy. In the fourth chapter, I have provided a means to uniformly distribute computational chemistry functionality, and to unify multiple software packages such that they could be automatically constructed into a single workflow, using the SADI framework. Finally in Chapter 5, I have demonstrated that semantic web services, supported by semantic chemical representation and classification, can mimic enzymatic behavior, thus enabling metabolic fate prediction

with the Web acting as an *in silico* biochemical reactor. The framework I have proposed here is very extensible to account for future developments in hardware and software, and can be readily evolved in several promising directions.

6.2 Alternative Architectures for MFP on the Semantic Web

While the architecture of a prototype semantic MFP framework presented here has been successfully applied to the automated construction of the metabolic pathway of benzene, there are many other approaches and architectures that could be suitable for this task. It would therefore be of interest to identify the limitations and benefits of a wide range of such architectures. For example, one of the principal potential limitations on the sustainable growth of the *in silico* enzymatic reactome may be its sheer size and the difficulty of querying the reactome for the appropriate reaction. In this case, a procedure-centric approach could have been applied to the identification of the biochemical transformations feasible for a particular compound, by examining a distributed reaction ontology or reaction specification RDF knowledgebase. Reactor growth would involve the extension of the ontology or the addition of triples describing novel reactions to the knowledgebase, and novel reactions would be identified through a single lookup of these sources and transformation creation by a single reactor service. The limitation in this approach is the assumption that all enzymes and reactions could be modeled using a single approach within a downstream thermodynamics service, meaning that all the nuances of reactive transformations, such as cooperativity, non-Michaelis-Menten kinetics, or feedback inhibition, could all be adequately described and processed by this service. While this may be possible to achieve in the future with first principles

approaches, no single software package or computational resource has achieved this task to date.

Fully reaction-centric approaches, on the other hand, could involve the construction of a large number of fully autonomous reactive web services that would be capable of providing reaction nature, kinetics, and thermodynamics – all from a single two-dimensional description of the input molecule. The strength of this architecture is in its versatility and provision for heterogeneity, in that many different modeling paradigms could be incorporated into a single framework. For instance, one web service may carry out a docking calculation to approximate the Michaelis-Menten constant, while another service may rely on a QSAR model earlier developed for this task. Provided that the user is not concerned with maintaining homogeneity in computational methods, this could be a powerful method combining the strengths of many approaches to construct a single model. If the user is concerned about calculation homogeneity, however, the selection of appropriate services for a given model hinges on their accurate annotation upon creation.

6.3 Accounting for Variations in Metabolism

While the ‘static’ metabolic fate of a given compound may vary little in various individuals and experimental conditions, due to its non-linear and complex nature, a compound’s dynamic fate could potentially be a subject to dramatic variations due to subtle variations on the level of tissue, individual, or the cell. For example, not all tissues have equal enzyme distributions, not all cells are subjected to the same extracellular pH, and not all individuals have catalytic sites that are identical. At the same time, this information is often available and could be incorporated into dynamic metabolic fate prediction models. For instance, relative enzyme distributions in various organs and

tissues is available within the GeneAtlas database [1], and polymorphisms of individual enzymes could be modeled using many homology modeling tools, by starting from a previously characterized enzyme crystal structure in the Protein Data Bank [2].

In the future, we anticipate that this information could become available by semantic cross-linking of the various necessary biological databases. In recent times, we have seen the process of semantic envelopment of biochemical information from various databases occur on a large scale, with billions of triples of biochemical information already accessible. Further integration of this information with our semantic MFP framework to produce organ- or individual-specific predictions of toxicity could prove to be an important asset in mechanistically explaining the toxicity of various drug candidates and other small molecules.

6.4 Semantic Enrichment of Pathways

Unlike the situation in many commercial packages for MFP, the pathway information we generate is readily accessible for further analysis and characterization. Thus, the lifecycle of a metabolic fate prediction does not need to end with a characterized pathway, but can be extended at any point in time with the application of additional web services and integrative queries. As we have demonstrated in our work on lipid classification and the annotation of the classified lipids with literature information, semantically encoded biochemical pathway information can be further enriched at any time.

The recent concerted efforts from the government, academia, and industry to release vast amounts of information detailing the toxicity, activity, adverse effects, and known drug-drug interactions, as evidenced by the PubChem and ChEMBL databases, among others, provides us with ample opportunities to extend our predictions. Not only can we *predict*

toxicological information, we could potentially expose the information that we *know* already – no small feat, considering the woeful state of informational resource fragmentation in the life sciences. Certainly, this is an avenue of research we would like to pursue further.

6.5 Distribution of Physical Resources

An intriguing new development in the world of automated scientific analysis has appeared recently: a fully automated and autonomous system mimicking the logical process involved in biochemical experimentation, has reported the first scientific discoveries [3]. Robot scientists, as they have come to be called, are currently capable of reproducing much of the systematized effort that was previously reserved solely to the human scientists. A typical robot scientist is capable of formulating hypotheses and testing them using a minimal set of experiments necessary to support them. While the logical process happens in the virtual domain, the outcomes of this logical process are manifested as a planned experimental workflow and executed in the physical world using digitized and automated instruments.

The recent development of hypothesis testing frameworks using semantic web technologies [4], coupled with an increasingly growing collection of automated and robotic scientific instruments and the capacity of putting said instruments under the control of software components that could reside within semantic web services, opens a wide array of new research possibilities. For instance, compounds could be assayed for a given property on demand to fill in gaps of information necessary for metabolic model construction, entire experimental workflows could be constructed using a single query, and manufacturing processes could be automated to produce custom goods on demand.

Finally, health sciences could use semantic web services to supplement disease diagnosis and patient background information and automatically assay biological samples to confidently infer the disease state of a given individual. Truly, we are only now beginning to feel the potential of semantic web technologies.

6.6 Conclusion

In conclusion, we have established a novel approach to predicting metabolic fates of arbitrary small molecules with the proposed framework. For the first time, we demonstrate that it is possible to produce a self-organizing, distributed biochemical reactor on the Web, and to use this reactor to gain valuable new insights into the nature and dynamics of cellular metabolism. We believe that a wider acceptance of the semantic web services in computational biology, as demonstrated by our MFP framework, can revolutionize and unify chemical and biological sciences, enabling cross-barrier collaboration, and earning us valuable new insights into the nature of life.

Appendices

Appendix A

A highly abbreviated representative list of the descriptor-generating semantic web services created for the study presented in Chapter 4.

Descriptor	Explanation
AlogP	Atomic calculation-based octanol-water partition coefficient.
Aromatic Atom Count	Number of aromatic atoms in a given molecule.
Atom Count	Number of atoms in a molecule.
Atomic Polarizability	Sum of atomic polarizabilities of all atoms in a given molecule.
Bond Count	Number of bonds in a molecule.
Eccentric Connectivity Index	A topological molecular descriptor that reflects on atom connectivity and distance.
Fractional Polar Surface Area	Total partially positively charged molecular surface area divided by the total molecular surface area.
Hydrogen Bond Acceptor Count	Number of atoms that can act as hydrogen bond acceptors.
Hydrogen Bond Donor Count	Number of atoms acting as hydrogen bond donors.
Ionization Potential	Propensity of a given molecule to lose an electron.
Largest Chain	The length of the longest chain of heavy atoms in a molecule
Largest PI System	The number of atoms in the largest conjugated pi-bond system.

Maximal Length/Breadth Ratio	Descriptor of molecular shape describing the ratio of a molecule's length to breadth in the region where it is highest.
Minimal Length/Breadth Ratio	Descriptor of molecular shape describing the ratio of a molecule's length to breadth in the region where it is lowest.
MlogP	Mannhold algorithm-based octanol-water partition coefficient.
Molar Refractivity	Total polarizability of a mole of a given molecule.
Molecular Complexity	A descriptor reflecting on the complexity of a given molecule in terms of quantities of heteroatoms and their connectivity.
Molecular Formula	A molecular formula that captures the types and counts of atoms present in a molecule.
Molecular Mass	The mass of a given molecule, in Daltons.
Petitjean Geometric Shape Index	A descriptor that reflects on the shape of the molecular connectivity graph and factors in distance information.
Petitjean Number	An index characterizing molecular graph topology.
Petitjean Topological Shape Index	A descriptor that reflects on the topological shape of the molecular graph.
Relative Hydrophobic Surface Area	The fraction of the overall molecular surface area that is hydrophobic.
Topological Polar Surface Area	Total molecular surface area that has a non-zero partial charge.
Total Hydrophilic Surface Area	Sum of solvent-accessible surfaces of partially charged atoms.

Total Hydrophobic Surface Area	Sum of solvent-accessible surfaces of nonpolar atoms.
Total Partially Negative Surface Area	Total surface area of a molecule that has a partially negative charge.
Total Partially Positive Surface Area	Total molecular area that has a partially positive charge.
Vertex Adjacency Magnitude	A descriptor reflecting upon the number of bonds between heavy atoms in a given molecule.
Wiener Polarity Number	A descriptor that reflects on atomic connectivity and molecular topology.
XlogP	Group contribution-based octanol-water partition coefficient.
Zagreb Index	The sum of squares of all heavy atom degrees.

References

Chapter 1

1. Coleman MD. Human drug metabolism: an introduction. John Wiley and Sons, New York. 2005.
2. Pearson PG, Wienkers LC, Eds. Handbook of drug metabolism, 2 Ed. Informa Healthcare. 2009.
3. Gibson GG, Skett P. Introduction to drug metabolism, 3 Ed, Nelson Thornes. 2001.
4. Nakamura Y, Miyoshi N. Electrophiles in foods: the current status of isothiocyanates and their chemical biology. *Biosci Biotechnol Biochem*. 2010. 74(2):242-55.
5. Wu Q, Dohnal V, Huang L, Kuča K, Wang X, Chen G, Yuan Z. Metabolic pathways of ochratoxin A. *Curr Drug Metab*. 2011. 12(1):1-10.
6. Mesárošová M, Valovičová Z, Srančíková A, Krajčovičová Z, Milcová A, Sokolová R, Schmuczerová J, Topinka J, Gábelová A. The role of human cytochrome P4503A4 in biotransformation of tissue-specific derivatives of 7H-dibenzo[c,g]carbazole. *Toxicol Appl Pharmacol*. 2011. Jul 21. [Epub ahead of print]
7. Tsaïoun K, Kates SA. ADMET for Medicinal Chemists: A Practical Guide. Wiley, New York. 2011.
8. van de Waterbeemd H, Testa B. Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability. Wiley-VCH, New York. 2008.
9. Nelson SD. Metabolic activation and drug toxicity. *J Med Chem*. 1982. 25 (7):753–765.
10. Benigni R, Bossa C. Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. *Chem Rev*. 2011. 111:2507–2536.

11. O'Connell TM, Watkins PB. The application of metabonomics to predict drug-induced liver injury. *Clin Pharmacol Ther.* 2010. 88(3):394-9.
12. Hsiao CJ, Younis H, Boelsterli UA. Trovafloxacin, a fluoroquinolone antibiotic with hepatotoxic potential, causes mitochondrial peroxynitrite stress in a mouse model of underlying mitochondrial dysfunction. *Chem Biol Interact.* 2010. 188(1):204-13.
13. Tujios S, Fontana RJ. Mechanisms of drug-induced liver injury: from bedside to bench. *Nat Rev Gastroenterol Hepatol.* 2011. 8(4):202-11.
14. Cavalieri E, Chakravarti D, Guttenplan J, Hart E, Ingle J, Jankowiak R, Muti P, Rogan E, Russo J, Santen R, Sutter T. Catechol estrogen quinones as initiators of breast and other human cancers: implications for biomarkers of susceptibility and cancer prevention. *Biochim Biophys Acta.* 2006. 1766(1):63-78.
15. Bolton JL, Thatcher GR. Potential mechanisms of estrogen quinone carcinogenesis. *Chem Res Toxicol.* 2008. 21(1):93-101.
16. Hui Y, Yasuda S, Liu MY, Wu YY, Liu MC. On the sulfation and methylation of catecholestrogens in human mammary epithelial cells and breast cancer cells. *Biol Pharm Bull.* 2008. 31(4):769-73.
17. Brinton RD. Investigative models for determining hormone therapy-induced outcomes in brain: evidence in support of a healthy cell bias of estrogen action. *Ann N Y Acad Sci.* 2005. 1052:57-74.
18. Henderson VW. Action of estrogens in the aging brain: dementia and cognitive aging. *Biochim Biophys Acta.* 2010. 1800(10):1077-83.

19. Trebst A. The Role of Benzoquinones in the Electron Transport System. Proc Royal Soc London. Series B, Biological Sciences, A Discussion on Photosynthesis. 157(968):355-366
20. Youdim KA, Saunders KC. A review of LC-MS techniques and high-throughput approaches used to investigate drug metabolism by cytochrome P450s. J Chromatogr B Analyt Technol Biomed Life Sci. 2010. 878(17-18):1326-36.
21. Xu R, Manuel M, Cramlett J, Kassel DB. A high throughput metabolic stability screening workflow with automated assessment of data quality in pharmaceutical industry. J Chromatogr A. 2010. 1217(10):1616-25.
22. Zhou SF, Liu JP, Chowbay B. Polymorphism of human cytochrome P450 enzymes and its clinical impact. Drug Metab Rev. 2009. 41(2):89-295.
23. Meyer U. Pharmacogenetics and adverse drug reactions. Lancet. 2000. 356:1667-1671
24. Valenzuela B, Rebollo J, Pérez T, Brugarolas A, Pérez-Ruixo JJ. Effect of grapefruit juice on the pharmacokinetics of docetaxel in cancer patients: a case report. Br J Clin Pharmacol. 2011. [Epub ahead of print]
25. Montgomery PR, Berger LG, Mitenko PA, Sitar DS. Salicylate metabolism: effects of age and sex in adults. Clin Pharmacol Ther. 1986. 39(5):571-6.
26. Zhang J, Shan Y, Pan X, Wang C, Xu W, He L. Molecular Docking, 3D-QSAR studies and in-silico ADME prediction of p-aminosalicylic acid derivatives as neuraminidase inhibitors. Chem Biol Drug Des. 2011. [Epub ahead of print]
27. Pistoia Alliance. [<http://www.pistoiaalliance.org/>]

28. Snyder R, Hedli CC. An overview of benzene metabolism. *Environ Health Perspect.* 1996. 104 S6:1165-71.
29. Costas Ioannides, Ed. *Cytochromes P450: role in the metabolism and toxicity of drugs and other xenobiotics*, Royal Society of Chemistry, London. 2008.
30. Wright JS, Shadnia H. Computational modeling of substituent effects on phenol toxicity. *Chem Res Toxicol.* 2008. 21(7):1426-31.
31. Dinkova-Kostova AT, Talalay P. NAD(P)H:quinone acceptor oxidoreductase 1 (NQO1), a multifunctional antioxidant enzyme and exceptionally versatile cytoprotector. *Arch Biochem Biophys.* 2010. 501(1):116-23.
32. Skipper PL, Trudel LJ, Kensler TW, Groopman JD, Egner PA, Liberman RG, Wogan GN, Tannenbaum SR. DNA adduct formation by 2,6-dimethyl-, 3,5-dimethyl-, and 3-ethylaniline in vivo in mice. *Chem Res Toxicol.* 2006. 19(8):1086-90.
33. Nouri-Nigjeh E, Permentier HP, Bischoff R, Bruins AP. Lidocaine oxidation by electrogenerated reactive oxygen species in the light of oxidative drug metabolism. *Anal Chem.* 2010. 82(18):7625-33.
34. Karch SB, Ed. *Drug abuse handbook*. CRC Press. 1998.
35. Pérez-Tomás R. Multidrug resistance: retrospect and prospects in anti-cancer drug treatment. *Curr Med Chem.* 2006;13(16):1859-76.
36. Gautam AH, Zeevalk GD. Characterization of reduced and oxidized dopamine and 3,4-dihydrophenylacetic acid, on brain mitochondrial electron transport chain activities. *Biochim Biophys Acta.* 2011. 1807(7):819-28.

37. Buchner E. Alkoholische Gärung ohne Hefezellen (Vorläufige Mitteilung). *Berichte der Deutschen Chemischen Gesellschaft*. 1897. 30:117–124.
38. Kolluru B, Hawizy L, Murray-Rust P, Tsujii J, Ananiadou S. Using workflows to explore and optimise named entity recognition for chemistry. *PLoS One*. 2011. 6(5):e20181.
39. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model*. 1988. 28:31.
40. McNaught A. The IUPAC International Chemical Identifier:InChI. *Chemistry International (IUPAC)*. 2006. 28(6).
41. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res*. 2002. 30(1):47-9.
42. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novère N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J; SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*. 2003. 19(4):524-31.
43. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-

- Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, van Iersel M, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le Novère N, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, Bader GD. The BioPAX community standard for pathway data sharing. *Nat Biotechnol.* 2010. 28(9):935-42.
44. Toby Segaran, Jeff Hammerbacher, Eds. *Beautiful Data: The Stories Behind Elegant Data Solutions*. O'Reilly Media, Inc. 2009. pp. 133-147.
45. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI, Snoep JL, Hucka M, Le Novère N, Laibe C. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol.* 2010. 4:92.
46. Reaxys Database. [<https://www.reaxys.com/info/>]
47. International Union of Biochemistry and Molecular Biology Nomenclature Recommendations. [<http://www.chem.qmul.ac.uk/iubmb/>]

48. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 2003. 31 (13): 3784–8.
49. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R. IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* 2004. 32:D434–7.
50. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 2004. 32:D277-80.
51. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 2008. 36:D623-31.
52. Haw R, Hermjakob H, D'Eustachio P, Stein L. Reactome Pathway Analysis to Enrich Biological Discovery in Proteomics Datasets. *Proteomics.* 2011. [Epub ahead of print]
53. Rhea Database. [<http://www.ebi.ac.uk/rhea/>]
54. Holliday GL, Bartlett GJ, Almonacid DE, O'Boyle NM, Murray-Rust P, Thornton JM, Mitchell JB. Bioinformatics. MACiE: a database of enzyme reaction mechanisms. 2005. 21(23):4315-6.
55. Blagodatski A, Katanaev VL. Technologies of directed protein evolution in vivo. *Cell Mol Life Sci.* 2011. 68(7):1207-14.
56. Li P, Dada JO, Jameson D, Spasic I, Swainston N, Carroll K, Dunn W, Khan F, Malys N, Messiha HL, Simeonidis E, Weichart D, Winder C, Wishart J, Broomhead DS,

- Goble CA, Gaskell SJ, Kell DB, Westerhoff HV, Mendes P, Paton NW. Systematic integration of experimental data and models in systems biology. *BMC Bioinformatics*. 2010. 11:582.
57. Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol*. 2005. (12):1509-15.
58. Wipke T, Howe J, Eds. *Computer-Assisted Organic Synthesis*. American Chemical Society. 1977.
59. Ellis LB, Gao J, Fenner K, Wackett LP. The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res*. 2008. 36:W427-32.
60. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, Kanehisa M. PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res*. 2010. 38:W138-43.
61. Faust K, Croes D, van Helden J. Metabolic pathfinding using RPAIR annotation. *J Mol Biol*. 2009. 388(2):390-414.
62. Hou BK, Ellis LB, Wackett LP. Encoding microbial metabolic logic: predicting biodegradation. *J Ind Microbiol Biotechnol*. 2004. 31(6):261-72.
63. McDonald AG, Tipton KF, Boyce S. Tracing metabolic pathways from enzyme data. *Biochim Biophys Acta*. 2009. 1794(9):1364-71.
64. Romualdo Benigni, Ed. *Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens*. CRC Press. 2003.

65. Bathelt CM, Mulholland AJ, Harvey JN. QM/MM modeling of benzene hydroxylation in human cytochrome P450 2C9. *J Phys Chem A*. 2008. 112(50):13149-56.
66. Zhou Z, Madura JD. CoMFA 3D-QSAR analysis of HIV-1 RT nonnucleoside inhibitors, TIBO derivatives based on docking conformation and alignment. *J Chem Inf Comput Sci*. 2004 Nov-Dec;44(6):2167-78.
67. Ravichandran V, Prashantha Kumar BR, Sankar S, Agrawal RK. Predicting anti-HIV activity of 1,3,4-thiazolidinone derivatives: 3D-QSAR approach. *Eur J Med Chem*. 2009. 44(3):1180-7.
68. Kamerlin SC, Haranczyk M, Warshel A. *J Phys Chem B*. Progress in ab initio QM/MM free-energy simulations of electrostatic energies in proteins: accelerated QM/MM studies of pKa, redox reactions and solvation free energies. 2009. 113(5):1253-72.
69. Segel, IH. *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*. Wiley-Interscience. New York. 1993.
70. Rester U. Dock around the Clock – Current Status of Small Molecule Docking and Scoring. *QSAR & Combinatorial Science*. 2006. 25(7): 605–615.
71. MOE: Molecular Operating Environment. [<http://www.chemcomp.com/software.htm>]
72. Talafous J, Sayre LM, Mieyal JJ, Klopman G. META. 2. A dictionary model of mammalian xenobiotic metabolism. *J Chem Inf Comput Sci*. 1994. 34(6):1326-33.
73. Darvas F, Marokházi S, Kormos P, Kulkarni G, Kalász H, Papp Á. Metabolexpert: Its Use in Metabolism Research and in Combinatorial Chemistry, in *Drug Metabolism*,

Databases and High-Throughput Testing During Drug Design and Development, Ed. by P.W. Erhardt. IUPAC Press. 1999 pp. 237-270.

74. Roberts DW, Patlewicz G, Dimitrov SD, Low LK, Aptula AO, Kern PS, Dimitrova GD, Comber MIH, Phillips RD, Niemelä J, Madsen C, Wedebye EB, Bailey PT, Mekenyan OG. TIMES-SS—A Mechanistic Evaluation of an External Validation Study Using Reaction Chemistry Principles. *Chem. Res. Toxicol.* 2007. 20(9):1321-1330.

75. KnowItAll Academic Edition. [<http://www.knowitall.com/academic/>]

76. Testa B, Balmat AL, Long A, Judson P. Predicting drug metabolism--an evaluation of the expert system METEOR. *Chem Biodivers.* 2005. 2(7):872-85.

77. Brachman R, Levesque H. Knowledge Representation and Reasoning. Morgan Kaufmann Publishers. New York. 2004.

78. Allemang D, Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann Publishers. New York. 2008.

79. Berners-Lee T. Design Issues for the World Wide Web.

[<http://www.w3.org/DesignIssues/>]

80. World Wide Web Consortium RDF Specifications.

[http://www.w3.org/standards/techs/rdf#w3c_all]

81. Prud'hommeaux E, Seaborne A, Eds. SPARQL Query language for RDF.

[<http://www.w3.org/TR/rdf-sparql-query/>]

82. Casher O, Rzepa HS. SemanticEye: A Semantic Web Application to Rationalize and Enhance Chemical Electronic Publishing. *J Chem Inf Model.* 2006. 46(6):2396–2411.

83. Downing J, Harvey MJ, Morgan PB, Murray-Rust P, Rzepa HS, Stewart DC, Tonge AP, Townsend JA. SPECTRa-T: machine-based data extraction and semantic searching of chemistry e-theses. *J Chem Inf Model*. 2010. 50(2):251-61.
84. McBride B, Ed. RDF Vocabulary Description Language 1.0: RDF Schema. [<http://www.w3.org/TR/rdf-schema/>]
85. McGuinness DL, van Harmelen F, Eds. OWL Web Ontology Language Overview. [<http://www.w3.org/TR/owl-features/>]
86. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007; 25(11):1251-5.
87. Crossfire Database Suites. [<http://www.info.crossfiredatabases.com/>]
88. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner JK, Willighagen E. The Blue Obelisk - Interoperability in Chemical Informatics. *J Chem Inf Model*. 2006; 46(3):991-998.
89. McNaught, A. The IUPAC International Chemical Identifier: InChI. *Chemistry International (IUPAC)*. 2006; 28(6).
90. Murray-Rust P, Rzepa HS. Chemical Markup, XML, and the World Wide Web. 1. Basic Principles. *J Chem Inf Comput Sci*. 1999; 39:928-942.

91. Holliday GL, Murray-Rust P, Rzepa HS. Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *J Chem Inf Model.* 2006; 46(1):145–157.
92. Zhang Y, Murray-Rust P, Dove M, Glen R, Rzepa H, Townsend J, Tyrell S, Wakelin J, Willighagen E. JUMBO: An XML infrastructure for eScience. UK e-Science All Hands Meeting, 2004.
93. Murray Rust P. The Semantic Chemical Web GoogleInChI and Other Mashups. GoogleTechTalks. <http://video.google.com/videoplay?docid=5871900250712259806> (accessed May 1, 2010).
94. Hughes G, Mills H, De Roure D, Frey JG, Moreau L, Schraefel MC, Smith G, Zaluska E. The semantic smart laboratory: a system for supporting the chemical eScientist. *Org Biomol Chem.* 2004; 2(22):3284-93.
95. Sanderson K, Neylon C. Data on display. *Nature.* 2008; 455(7211):273.
96. Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems.* 2010; 6.
97. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform.* 2008; 41(5):706-16.
98. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics.* 2010; 11:255.

99. Jentzsch A, Zhao J, Hassanzadeh O, Cheung K, Samwald M, Andersson B. Linking Open Drug Data. I-Semantics Conference, Graz, Austria, September 204, 2009.

Chapter 2

1. Crossfire Database Suite [<http://www.info.crossfiredatabases.com/>]
2. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: Chemical Entities of Biological Interest: an update. Nucl Acids Res 2010, 38:D249-D254.
3. Li Q, Cheng T, Wang Y, Bryant SH: PubChem as a public resource for drug discovery. Drug Discov Today 2010, 15:1052-1057.
4. ChemSpider Database of Chemical Structures and Property Predictions [<http://www.chemspider.com/>]
5. ChEMBL Database [<https://www.ebi.ac.uk/chembl/>]
6. Wiswesser WJ: How the WLN began in 1949 and how it might be in 1999. J Chem Inf Comput Sci 1982, 22:88-93.
7. Weininger D: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988, 28:31-36.
8. McNaught A: The IUPAC International Chemical Identifier: InChI - A New Standard for Molecular Informatics. Chem International 2006, 28:12-15.
9. Connectivity Table File Formats [<http://www.symyx.com/downloads/public/ctfile/ctfile.jsp>]

10. Fujita S: XyM Markup Language (XyMML) for Electronic Communication of Chemical Documents Containing Structural Formulas and Reaction Schemes. *J Chem Inf Comput Sci* 1999, 39:915-927.
11. Sankar P, Aghila G: Ontology Aided Modeling of Organic Reaction Mechanisms with Flexible and Fragment Based XML Markup Procedures. *J Chem Inf Model* 2007, 47:1747-1762.
12. Sankar P, Alain K, Aghila G: Model Tool to Describe Chemical Structures in XML Format Utilizing Structural Fragments and Chemical Ontology. *J Chem Inf Model* 2010, 50:755-770.
13. Murray-Rust P, Rzepa HS: Chemical Markup, XML, and the World Wide Web. 1. Basic Principles. *J Chem Inf Comput Sci* 1999, 39:928-942.
14. Adams N, Winter J, Murray-Rust P, Rzepa HS: Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language. *J Chem Inf Model* 2008, 48:2118-2128.
15. Resource Description Framework Specification [<http://www.w3.org/RDF/>]
16. Hughes G, Mills H, De Roure D, Frey JG, Moreau L, Schraefel MC, Smith G, Zaluska E: The semantic smart laboratory: a system for supporting the chemical eScientist. *Org Biomol Chem* 2004, 2:3284-3293.
17. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008, 41:706-716.

18. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 2010, 11:255-267.
19. Linking Open Drug Data Project [<http://esw.w3.org/HCLSIG/LODD>]
20. Linking Open Data Initiative
[<http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>]
21. Konyk M, De Leon A, Dumontier M: Chemical Knowledge for the Semantic Web. In: *Data Integration in the Life Sciences Conference*; June 25-27, 2008; Evry, France. Springer 2008, 169-176.
22. CHEMINF Ontology
[<http://semanticchemistry.googlecode.com/svn/trunk/ontology/cheminf.owl>]
23. Chemical Entity Semantic Specification [<http://semanticscience.org/projects/chess/>]
24. Semanticscience Integrated Ontology [<http://semanticscience.org>]
25. Semantic Web Rule Language [<http://www.w3.org/Submission/SWRL/>]
26. SPARQL Query Language for RDF [<http://www.w3.org/TR/rdf-sparql-query/>]
27. Fahy E, Sud M, Cotter D, Subramaniam S: LIPID MAPS online tools for lipid research. *Nucleic Acids Res* 2007, 35:W606-612.
28. BigOWLIM High-Performance Semantic Repository
[<http://www.ontotext.com/owlim/big/>]
29. FaCT++ Reasoner [<http://owl.man.ac.uk/factplusplus/>]

30. Wilkinson MD, Vandervalk BP, McCarthy EL: SADI SemanticWeb Services – 'cause you can't always GET what you want! In: Asia-Pacific Services Computing Conference; December 7-11, 2009; Singapore. IEEE Press 2009, 13–18.
31. Rahman SA, Bashton M, Holliday GL, Schrader R, Thornton JM: Small Molecule Subgraph Detector (SMSD) toolkit. *J Cheminformatics* 2009, 1:12.
32. Mannhold R, Petrauskas A. Substructure versus Whole-molecule Approaches for Calculating Log P. *QSAR & Combinatorial Science* 2003, 22:466–475.
33. Lipinski CA: Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today: Technologies* 2004, 1:337-341.
34. Pellet OWL 2 Reasoner [<http://clarkparsia.com/pellet/>]
35. Protégé Ontology Editor [<http://protege.stanford.edu/>]
36. Hussain HH, Babic G, Durst T, Wright JS, Flueraru M, Chichirau A, Chepelev LL: Development of novel antioxidants: design, synthesis, and reactivity. *J Org Chem* 2003, 68:7023-7032.
37. Web Ontology Language Specification [<http://www.w3.org/TR/owl2-overview/>]
38. Jena Semantic Web Framework [<http://jena.sourceforge.net/>]
39. Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen EL: The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J Chem Inf Comp Sci* 2003, 43:493-500.
40. Open Babel Chemistry Toolbox [http://openbabel.org/wiki/Main_Page]
41. OpenLink Virtuoso [<http://virtuoso.openlinksw.com/>]

42. Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y: Pellet: A practical OWL-DL reasoner. *Web Semantics* 2007, 5:51-53.

Chapter 3

1. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 36, D344–D350.
2. *Nucleic Acids Res.* 2007 Jan;35(Database issue):D527-32. Epub 2006 Nov 10. LMSD: LIPID MAPS structure database. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH Jr, Murphy RC, Raetz CR, Russell DW, Subramaniam S.
3. ChEMBL Database. [<https://www.ebi.ac.uk/chembl/db>]
4. Medical Subject Headings [<http://www.nlm.nih.gov/mesh/>]
5. Benigni R, Bossa C. Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. *Chem Rev*, 2011, 111:2507–2536.
6. *Curr Drug Metab.* 2010 Oct;11(8):639-58. Human carbonyl reductases. Malátková P, Maser E, Wsól V.
7. *Am J Respir Cell Mol Biol.* 2009 Nov;41(5):509-15. Epub 2009 Aug 14. Redox regulation of gamma-glutamyl transpeptidase. Zhang H, Forman HJ.
8. Gordon JE. Chemical inference. 3. formalization of the language of relational chemistry: ontology and algebra. *J. Chem. Inf. Comput. Sci.*, 1988, 28 (2), pp 100–115.
9. *FEBS Lett.* 2005 Aug 29;579(21):4685-91. CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. Feldman HJ, Dumontier M, Ling S, Haider N, Hogue CW.
10. M. D. Prasanna, J. Vondrasek, A. Wlodawer, H. Rodriguez, and T. N. Bhat. Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds. *Proteins*, vol. 63, pp. 907-17, Jun 1 2006.

11. Helma, C.: In Silico predictive toxicology: The state-of-the-art and strategies to predict human health effects. *Curr. Opin. Drug Discov. Devel.* 8, 27-31 (2005)
12. Cramer, G.M., Ford, R.A., Hall, R.L.: Estimation of Toxic Hazard - A Decision Tree Approach. *J. Cosmet. Toxicol.* 16, 255-276 (1978)
13. Kroes, R., et al.: Structure based thresholds of toxicological concern. *Food Chem. Toxicol.* 42, 65–83 (2004)
14. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann (1999)
15. Belleau, F. et al.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 41, 706-716 (2008)
16. SMARTS Theory Manual.
[<http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>]
17. Ian H. Witten; Eibe Frank (2005). "Data Mining: Practical machine learning tools and techniques, 2nd Edition". Morgan Kaufmann, San Francisco.
18. Semtox Project Page. {<http://semanticscience.org/projects/semtox/>}
19. Horridge, M., Parsia, B., Sattler, U.: Laconic and precise justifications in OWL. In: *Proc. of ISWC-08, LNCS*. 5318, 323-338 (2008)
20. International Union of Pure and Applied Chemists - Ester Definition.
[<http://goldbook.iupac.org/E02219.html>]
21. Klinov, P.: Pronto: A Non-monotonic Probabilistic Description Logic Reasoner. In: *The Semantic Web: Research and Applications, LNCS*. 5021, 826-830 (2008)
22. Horridge, M., Bechhofer, S.: The OWL API. *OWLED 2009, 6th OWL Experiences and Directions Workshop*. Chantilly, Virginia, USA. (2009)

23. Carcinogenic Potency Database. [http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html]
24. Patlewicz, G. et al.: An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. SAR QSAR Environ. Res. 19, 495-524 (2008)
25. Wishart, D.S., et al.: HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 37, D603-D610 (2009)
26. Lipinski, C.A. et al.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. Adv. Drug. Del. Rev. 46, 3–26 (2001)
27. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1999)
28. SADI Framework. [<http://sadiframework.org>]

Chapter 4

1. Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen EL: The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. J Chem Inf Comp Sci 2003, 43:493-500.
2. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner JK, Willighagen E: The Blue Obelisk - Interoperability in Chemical Informatics. J Chem Inf Model 2006, 46:991-998.
3. Benfenati E, Benigni R, Demarini DM, Helma C, Kirkland D, Martin TM, Mazzatorta P, Ouédraogo-Arras G, Richard AM, Schilter B, Schoonen WG, Snyder RD, Yang C: Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the-art, and perspectives. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 2009, 27:57-90.

4. JOELib Java-based cheminformatics library [<http://www.ra.cs.uni-tuebingen.de/software/joelib/>]
5. RDKit Cheminformatics Package [<http://rdkit.org/>]
6. DiBernardo M, Pottinger R, Wilkinson M: Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework. *J Biomed Inform* 2008, 41:837-47.
7. Kuhn T, Willighagen EL, Zielesny A, Steinbeck C: CDK-Taverna: an open workflow environment for cheminformatics. *BMC Bioinformatics* 2010, 11:159.
8. Web Service Description Language Specification [<http://www.w3.org/TR/wsdl>]
9. Resource Description Framework Specification [<http://www.w3.org/RDF/>]
10. Web Ontology Language Specification [<http://www.w3.org/TR/owl2-overview/>]
11. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008, 41:706-16.
12. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 2010, 11:255.
13. Linking Open Drug Data Project [<http://esw.w3.org/HCLSIG/LODD>]
14. Semantic Annotations for WSDL and XML Schema Specification [<http://www.w3.org/TR/sawSDL/>]
15. Vitvar T, Mocan A, Kerrigan M, Zaremba M, Zaremba M, Moran M, Cimpian E, Haselwanter T, Fensel D: Semantically-enabled Service Oriented Architecture: Concepts, Technology and Application. *J Service Oriented Comp Applications* 2007, 1:129-154.

16. Neerincx PB, Leunissen JA: Evolution of web services in bioinformatics. *Brief Bioinform* 2005, 6:178-88.
17. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A: TAMBIS:transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000, 16:184–185.
18. Lord P, Alper P, Wroe C, Goble C: Feta: A Light-Weight Architecture for User Oriented Semantic Service Discovery. *ESWC 2005, LNCS 3532*, pp. 17–31, 2005.
19. Wilkinson MD, Links M: BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics* 2002, 3:331–41.
20. Gessler DDG, Schiltz GS, May GD, Avraham S, Town CD, Grant D, Nelson RT: SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. *BMC Bioinformatics* 2009, 10:309.
21. Kietz J, Serban F, Bernstein A, Fischer S: Data Mining Workflow Templates for Intelligent Discovery Assistance and Auto-Experimentation. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*; September 24, 2010.
22. Wilkinson MD, Vandervalk BP, McCarthy EL: SADI SemanticWeb Services – 'cause you can't always GET what you want! In: *Asia-Pacific Services Computing Conference*; December 7-11, 2009; Singapore. *IEEE Press* 2009, 13–18.
23. Vandervalk BP, McCarthy EL, Wilkinson MD: SHARE: A Semantic Web Query Engine for Bioinformatics. *Lecture Notes in Computer Science* 2009, 5926:367-369.
24. SADI Service Registry [<http://sadiframework.org/registry/services/>]

25. Lipinski Service Ontology

[<http://semanticscience.org/sadi/ontology/lipinskiserviceontology.owl>]

26. CHEMINF Ontology

[<http://semanticchemistry.googlecode.com/svn/trunk/ontology/cheminf.owl>]

27. Bond Count Descriptor Service

[<http://s6.semanticscience.org:8080/bondcount/bondcount>]

28. SHARE Web Interface [<http://dev.biordf.net/cardioSHARE/>]

29. McNaught A: The IUPAC International Chemical Identifier: InChI — A New Standard for Molecular Informatics. *Chem International* 2006, 28:12-15.

30. Service for converting InChI to SMILES

[<http://s7.semanticscience.org:9090/wyrm/inchi2smilesbabel>]

31. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev* 1997, 23:3–25.

32. Lipinski CA: Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today: Technologies* 2004, 1:337-341.

33. Parameterized LogP Calculator Service

[<http://s7.semanticscience.org:9090/parameters/paramlogp>]

34. KEGG Pathway Association Retrieval Service [<http://dev.biordf.net/~kawas/cgi-bin/getKeggPathwaysByKeggDrug>]

35. Functional Group Annotation Service

[<http://s7.semanticscience.org:9090/computesas/fgannotate>]

36. SADI Taverna Plugin [http://www.taverna.org.uk/documentation/taverna-plugins/taverna-2-x-plugins/#sadi_plugin]
37. SemanticScience Integrated Ontology [<http://semanticscience.org>]
38. Open Babel Open Source Chemistry Toolbox [http://openbabel.org/wiki/Main_Page]

Chapter 5

1. Roy S, Goedecker S, Hellmann V. Bell-Evans-Polanyi principle for molecular dynamics trajectories and its implications for global optimization. *Phys. Rev. E* 77:67-72 (2008)
2. Segel IH. *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*. Wiley-Interscience. New York.(1993)
3. Goldberg RN, Tewari YB, Bell D, et al. Thermodynamics of enzyme-catalyzed reactions: Part 1. Oxidoreductases. *J. Phys. Chem. Ref. Data* 22:515-82 (1993)
4. Wilson IB, Harrison MA. Turnover Number of Acetylcholinesterase. *J Biol Chem.* 236:2292-2295 (1961)
5. RDKit. [<http://www.rdkit.org>]
6. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38:D142-8 (2010)
7. James J. P. Stewart. MOPAC: A semiempirical molecular orbital program. *J Comp Aided Mol Design* 4:1-103 (1990)
8. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Montgomery, Jr., J. A., Vreven, T., Kudin, K. N., Burant, J. C., Millam, J. M., Iyengar, S. S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani,

- G., Rega, N., Petersson, G. A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J. E., Hratchian, H. P., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Ayala, P. Y., Morokuma, K., Voth, G. A., Salvador, P., Dannenberg, J. J., Zakrzewski, V. G., Dapprich, S., Daniels, A. D., Strain, M. C., Farkas, O., Malick, D. K., Rabuck, A. D., Raghavachari, K., Foresman, J. B., Ortiz, J. V., Cui, Q., Baboul, A. G., Clifford, S., Cioslowski, J., Stefanov, B. B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Martin, R. L., Fox, D. J., Keith, T., Al-Laham, M. A., Peng, C. Y., Nanayakkara, A., Challacombe, M., Gill, P. M. W., Johnson, B., Chen, W., Wong, M. W., Gonzalez, C., and Pople, J. A. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford CT, 2004.
9. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U. COPASI — a COMplex PATHway Simulator. *Bioinformatics*. 22:3067-74 (2006)

Chapter 6

1. GeneAtlas. [<http://www.genatlas.org>]
2. Protein Data Bank. [<http://www.pdb.org>]
3. Sparkes A, Aubrey W, Byrne E, Clare A, Khan MN, Liakata M, Markham M, Rowland J, Soldatova LN, Whelan KE, Young M, King RD. Towards Robot Scientists for autonomous scientific discovery. *Autom Exp.* 2010. 2:1.
4. Callahan A, Dumontier M, Shah NH. HyQue: evaluating hypotheses using Semantic Web technologies. *J Biomed Semantics.* 2011. Suppl 2:S3.