Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

From data base to knowledge graph - using data in chemistry

Angiras Menon [1,3], Nenad B. Krdzavac [1,3], Markus Kraft [1,2,3]

released: August 08 2019

[1] Department of Chemical Engineering
and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom
E-mail: mk306@cam.ac.uk

[2] School of Chemical and
Biomedical Engineering
Nanyang Technological University
62 Nanyang Drive
Singapore 637459

[3] Cambridge Centre for Advanced Research and
Education in Singapore
CREATE Tower, 1 CREATE Way
Singapore 138602

Preprint No. 235

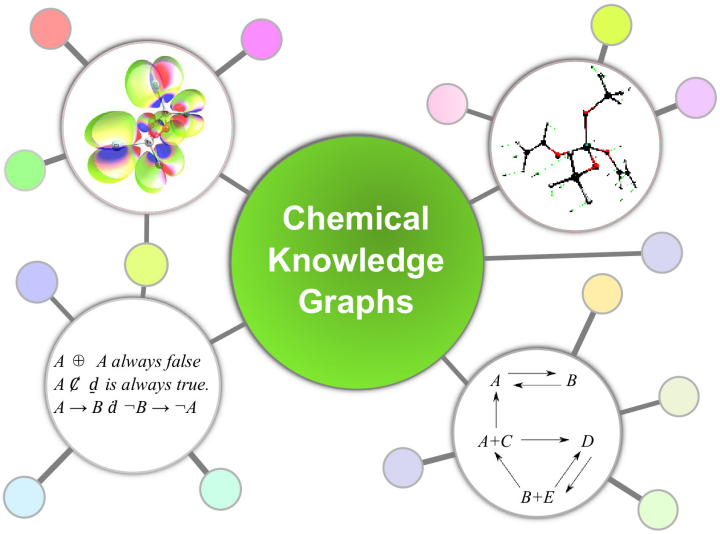
Keywords: c4e, preprint, template

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge CB3 0AS
United Kingdom

E-Mail: c4e@cam.ac.uk
World Wide Web: http://como.ceb.cam.ac.uk/

Abstract

Over the last couple of decades, the scientific community has made large efforts to process and store experimental and computational chemical data and information on the world wide web. This review summarizes several databases and ontologies available on the web for researchers to use. We also discuss briefly the categories of chemistry data that are stored, its main usage and how it can be accessed and understood in the framework of the Semantic Web.



Highlights

- A review on web based data bases used to store chemical information.
- Discussion on the role of knowledge graphs in chemical model development.

Contents

1 Introduction

As progress is being made in developing new and green chemical processes for a variety of industrial applications, an ever-growing amount of chemical information has been published and stored in databases online. This includes both experimental and computational chemical data. As a result, understanding how to store, access, and manipulate this vast amount of information is now key to further scientific progress. Increasingly, information science and mathematical methods such as data mining and graph theory are being used to guide various fields in chemistry and chemical engineering. Examples include analyzing organic reaction networks to understand and plan new synthetic routes for green chemistry [4, 9, 14, 32], and the use of process informatics to develop predictive chemical kinetics for combustion chemistry [10]. In addition, various approaches to access and generate chemical knowledge are being developed using, for example, semantic web and network analysis. Semantic web technologies like knowledge graphs offer additional functionality to represent chemical knowledge. In conjunction with semantic web services the information available in chemical databases can be retrieved and changed and allows the automation of model building [15, 24, 47]. The purpose of this review is to describe some of the main current databases available to researchers for data mining and review, as well as to discuss efforts to use ontologies as a general model for the representation of chemistry data, the improvement of the quality of these data, and the generation of resources to share consistent chemical data for a variety of purposes.

2 Chemical Databases

Several large chemical databases are available in the chemistry literature, providing a wealth of useful chemical information for researchers to use. The purpose of this section is to summarize some of the key features of such databases, for example, what information on chemical species they store and how this information can be queried. The world's largest freely accessible database of chemical information is PubChem [33], which stores information in three primary categories: compounds, substances, and bioactivities [20, 33]. Currently, PubChem has information on 97 million compounds, 242 million substances, and 280 million bioactivities [20, 33]. Information in PubChem can be queried by standard means, such as by text search, molecular formula, or chemical structure. For a common molecule, such as benzene, PubChem contains a variety of properties. This includes 2D and 3D structures as well as any crystal structures which can be downloaded in standard chemical formats such as JavaScript Object Notation (JSON), eXtensible Markup Language (XML) [5], or Common Interchange Format (CIF). PubChem also computes standard identifiers for the species in question, such as the IUPAC name, the canonical SMILES identifier [34, 49], or the InChI format

[17], as well as other vendor/chemical agency identifiers. These identifiers enable identification and comparison of species between databases, so are key to linking data for the same species from different sources. Essential computed and experimental chemical and physical properties for the structure are also provided by PubChem, as is any available spectral data that has been linked to the structure. PubChem also provides a large amount of information on the biological aspects of such structures, including drug information, solubility, toxicity, and biological activity, which is key data for those designing drugs or green synthesis routes.

Another major database for chemical data is Reaxys, run by Elsevier [13, 23]. Reaxys contains much of the same information as PubChem and other chemical databases, such as structure, key identifiers, physical and chemical properties, spectral data, and biological activity for various compounds. What differentiates Reaxys is its focus on providing data for developing synthetic routes. To this end, Reaxys has three key sets of information for a substance, namely preparations, reactions, and documents. Preparations displays key synthesis routes that can be used to prepare the substance in question. This includes the main reactions, reaction conditions, catalysts and any other information used in the synthesis routes. Each synthesis route also contains the source of the synthesis, which usually comes from the Journals and Patent databases that are linked to Reaxys via Elsevier. This enables the user to create a synthetic route for the substance of interest using ReaxysâĂŹ synthesis planner. Similarly, the reaction set contains the list of reactions in the Reaxys database which includes the substance the user has queried. The reactions can be filtered by structure, reagent, reaction class, solvents, catalysts, and yield among others, allowing the user to find reactions tailored to their application. Finally, the documents class lists the journal publications, patents, conference papers, and books that Reaxys has access to that are linked to the queried substance. This allows users of Reaxys to have access to both the data and source to analyze and select reactions.

Similar to Reaxys, the Chemical Abstracts Service (CAS) [1, 25] is a collection of databases containing information on organic and inorganic chemical substances. This information includes chemical structures, chemical names, and chemical reactions. Information stored in these databases is extracted from a wide range of literature such as patent records, journal publications, conference proceedings, Ph.D. theses, and web sources. The CAS Registry databases contain chemical structures, names, and experimental properties for more than 150 million molecules [1]. Building on the scope of the CAS Registry, the CASREACT database [2] contains several million single- and multi-step chemical reactions based on the molecules and the information stored in the CAS database. Much like Reaxys, this is provided to help users find reactions for their particular chemical application.

A key database for thermochemical data is the Active Thermochemical Tables (ATcT), developed by researchers at the Argonne National Laboratory [40, 41]. The principle behind the ATcT is the thermochemical network approach, which makes use of both

experimental and theoretical reaction and formation enthalpies to yield estimates for the enthalpy of formation of the species in the network. The ATcT describes thermochemistry using a graph theoretic approach, with primary vertices being the enthalpies of formation of species, secondary vertices being the reaction enthalpies, and the directed edges indicating a reaction occurring between species in the network, with the weight determined by stoichiometry. A statistical approach is then used to analyze and solve for the optimal thermochemical values that yield a self-consistent solution. Typically, this is possible because there are multiple measurements or calculations for a given formation or reaction enthalpy, providing the extra degrees of freedom necessary. This also means that the solution given by the ATcT can help to identify measurements that are potentially inconsistent with others in the network. Data computed by the ATcT can be found and queried online. Crucially, the reactions which contribute to the ATcT enthalpy of formation are displayed, as are uncertainties in the estimate of enthalpy of formation provided, making it clear which data is used and its degree of reliability.

On the computational chemical database side, the largest database is the Computational Chemistry Comparison and Benchmark DataBase (CCCBDB) for thermochemical properties of species from the National Institute of Standards and Technology (NIST) [19]. Information is queried by chemical name or molecular formula. The CCCBDB stores computed information in the following main categories: energy, geometry, vibrations, electrostatics, entropy and heat capacity, and reaction. All of the computed properties are displayed for the different levels of theory at which they have been calculated, with the data split into categories based on the type of computational chemical method used. The CCCBDB also crucially has a comparison feature, where the user can compare the results of theoretical calculations to any available experimental data in NIST's databases, as well as look at the effect of different theoretical methods on calculated properties.

Other more specialized databases also exist. For example, the Alexandria library developed by van der Spoel et al. consists of molecular properties for force field development [30]. Alexandria contains molecular structures and properties for 2,704 compounds, many of which contain functional groups common to biomolecules and drugs. Alexandria contains similar information to the CCCBDB, but crucially provides more extensive multipole and polarizability calculations to guide researchers who want to develop potentials and force fields. Importantly, all properties in Alexandria are provided at the same level of theory and the Gaussian input and output files from the calculations are also given, making reproduction of the stored information significantly easier. Even more specialized databases for computational chemists exist, such as Head-Gordon and Hait's benchmark database specifically for DFT calculations on dipole moments, spanning a variety of functionals in the process [28]. The database from Simmie et al. is specifically for high-level enthalpies of formation for nitrogen based compounds [42]. The GDB-17 database specifically enumerates small organic molecules, using graph-theoretic methods to span 166 billion such molecules with the aim of guiding new drug

design [39]. Ramakrishnan et al. provide the QM9 dataset, containing DFT calculations on around 134,000 molecules for training new machine learning potentials [38]. The ANI-1 data set uniquely contains non-equilibrium DFT calculations, that is for molecules in conformers that are not their minimum energy ground state configuration [43]. ANI-1 contains around 20 million molecular conformations for 57,462 molecules taken from the GDB database. There is clearly a wide variety of chemical data, both experimental and computational, that is available to researchers in a variety of fields in chemistry. This data is ever growing, and methods to store, access, and act on this data automatically are becoming more valuable for progress to be made.

3 Ontologies for Computational Chemistry

Given the variety of chemical data available, developing a consistent framework to store and access it is crucial, even more so as the amount of data available is expanding rapidly. Further data processing will increasingly rely on automation allowing machines to interpret, integrate, share, and perform reasoning with data of various formats.

One of the early efforts in storing chemical data in a standard format was the introduction of Chemical Markup Language (CML) pioneered by Murray-Rust and coworkers [11, 26, 27, 48]. The CML format is based on XML, which is suitable for storing data of any level of complexity while providing semantic information to the data stored. CML allows the representation of complex chemical objects by employing the hierarchical tree structure of XML using chemical name tags which cover different aspects of chemistry. Over the past 20 years, CML has been developed to represent most aspects of chemistry, including CMLReact for chemical reactions [18], CMLSpec for spectral data [22], CML for crystallography [7], and CML for polymers (PML) [3] along with the standard labels and definitions for physical properties.

Building on this established format for representing chemical data, Phandungsukanan and coworkers developed a sub-domain for storing quantum chemistry calculations data based on CML, termed CompChem [37]. The main goal of CompChem was to introduce a stricter structure into CML-based documents so that software tools know exactly how to validate and process information related to computational chemistry. To this end, the semantics of data stored in the CompChem based documents is modelled based on the typical nature of computational simulations or calculations, containing information on the job type, input parameters, and output parameters that one would expect in these calculations. This enables the storage of a variety of output data from *ab initio* quantum chemistry calculations such as the results of geometry optimization, single point energy calculations, and frequency calculations, among others. The storage and access of this data was realized through a MolHub web service [37]. However, the original MolHub did not allow for semantic inter-operability between different chemistry software tools, provide an efficient query engine, or guarantee the consistency of data.

To alleviate these shortcomings, a novel OntoCompChem ontology has been developed by extending the Gainesville Core (GNVC) ontology [36] while supporting the CompChem convention of CML [31]. The OntoCompChem ontology is currently populated by Gaussian quantum chemistry calculations through an updated version of the MolHub semantic web service (https://como.ceb.cam.ac.uk/resources/molhub/). The OntoCompChem knowledge graph forms part of a more general knowledge graph called the J-Park Simulator (JPS) [21]. This architecture supports semantic inter-operability between different domains and allows the use of propositional logic, formal query language, and Semantic Web tools such as the HermiT [12] reasoner to check the consistency of data within the JPS knowledge graph. More recently, the OntoKin ontology [8, 29] has been developed as a component of the JPS to represent gas phase elementary reactions, which are the building block of large reaction mechanisms found in combustion and atmospheric chemistry models. The ontology allows inference engines to detect inconsistencies in chemical mechanisms and to perform semantic queries across mechanisms stored in the JPS knowledge graph. At present, both the OntoKin and MolHub frameworks are missing an intelligent system that automatically establishes semantic inter-operability between quantum chemistry calculations and kinetic mechanisms. To achieve this goal, we are currently developing a formal framework that is based on reinforcement learning formal tools [46], modal logic [6], and a propositional logic framework with binary metric operators [45] to provide formal language support.

In addition to the JPS efforts, other semantic frameworks are currently in use. The Chemical Semantics Framework (CSF) [35] stores results of quantum chemistry calculations. The core of the CSF is the GNVC ontology which forms the knowledge component of the framework. However, the ontology does not support all of CompChem's conventions for CML features. For example, some keywords in the CML format such as geometry type are not supported. In addition, the CSF does not support semantic inter-operability between different computational chemistry tools. However, the framework allows web agents to access and, in principle, act on data stored in the CSF, representing a step towards automation of the knowledge graph. The ChEBI database stores molecular entities focused on 'small' chemical compounds, that is part of the Open Biomedical Ontologies effort. It uses the ChEBI ontology as a common model for classification of chemical compounds in the biomedical field. The ontology provides models for molecular structures such as hydrocarbons, common chemical roles for the molecules in the ontology, as well as for information pertaining to subatomic particles [16]. The ChEBI database can be explored using an advanced search interface, but semantic inter-operability and web agent access is currently not supported.

The review of ontologies for chemistry makes it clear that plenty of effort is being put towards developing methods for storing, accessing, and interpreting the available chemical data in an intelligent way. Key to the success of these efforts will be the development of standards for the publication and reporting of chemical data. By having a standard format for reporting chemical data, linking this information to a semantic

framework or ontology becomes substantially easier and less error prone. Efforts to this end include the work of the InChI consortium [17], the Allotrope Foundation's work on developing a standard data format, and the work of Cronin and coworkers on developing a chemical programming language that can be used to represent experimental organic chemistry [44]. These standards will help inspire the definition of classes in chemical ontologies. In conjunction with this, the development of tools for establishing semantic frameworks, as well as agents that can act on this data automatically, is still in process. This will eventually enable a self-consistent and ever-growing chemical knowledge graph based on ontologies and automated by web agents.

4 Summary and Outlook

In this review, we have discussed how the rapidly increasing amount of chemical information available to researchers has necessitated the development of automated methods to query, store, and share this information for a variety of applications. We have discussed some of the main databases and the usage of ontologies in the chemistry domain. Moving forward, it is hoped that more tools will be developed to provide more intelligent ways to create, update, retrieve, and maintain distributed chemical information via the Web. It is also necessary to develop tools to support more advanced community involvement, bridging data silos, and identifying "best" data for the solution of a particular problem. Eventually, the chemical knowledge graph will be fully automated and self-improving to provide, for example, new synthesis routes and more reliable chemical models built on the experimental and chemical data provided in the variety of databases online.

Acknowledgements

AM acknowledges Johnson Matthey for financial support. The authors also acknowledge the financial support of the Singapore National Research Foundation (NRF) through the Campus for Research Excellence and Technological Enterprise (CREATE) program. MK gratefully acknowledges the support of the Alexander von Humboldt foundation.

References

[1] Cas registry database, 2019. URL https://www.cas.org/support/documentation/cas-databases. Accessed May 23rd, 2019.

[2] Casreact - cas chemical reactions database, 2019. URL https://www.cas.org/support/documentation/reactions. Accessed May 23rd, 2019.

[3] N. Adams, J. Winter, P. Murray-Rust, and H. S. Rzepa. Chemical markup, xml and the world-wide web. 8. polymer markup language. *Journal of chemical information and modeling*, 48(11):2118–2128, 2008.

[4] K. J. M. Bishop, R. Klajn, and B. A. Grzybowski. The core and most useful molecules in organic chemistry. *Angewandte Chemie International Edition*, 45(32):5348–5354, 2006. doi:10.1002/anie.200600881. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200600881.

[5] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau. Extensible markup language (xml) 1.0, 2000.

[6] A. V. Chagrov and M. Zakharyaschev. *Modal Logic*, volume 35 of *Oxford logic guides*. Oxford University Press, 1997. ISBN 978-0-19-853779-3. URL https://dblp.org/rec/bib/books/daglib/0030819.

[7] N. Day, J. Downing, S. Adams, N. England, and P. Murray-Rust. Crystaleye. *URL http://wwmm. ch. cam. ac. uk/crystaleye/. Online*, 2008.

[8] F. Farazi, N. B. Krdzavac, J. Akroyd, S. Mosbach, A. Menon, D. Nurkowski, and M. Kraft. Linking reaction mechanisms and quantum chemistry: An ontological approach. 2019. Submitted for publication.

[9] M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, and B. A. Grzybowski. Architecture and evolution of organic chemistry. *Angewandte Chemie International Edition*, 44(44):7263–7269, 2005. doi:10.1002/anie.200502272. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200502272.

[10] M. Frenklach. Transforming data into knowledgeâĂŤprocess informatics for combustion chemistry. *Proceedings of the Combustion Institute*, 31(1):125 – 140, 2007. ISSN 1540-7489. doi:https://doi.org/10.1016/j.proci.2006.08.121. URL http://www.sciencedirect.com/science/article/pii/S1540748906003841.

[11] G. V. Gkoutos, P. Murray-Rust, H. S. Rzepa, and M. Wright. Chemical markup, xml, and the world-wide web. 3. toward a signed semantic chemical web of trust. *Journal of Chemical Information and Computer Sciences*, 41(5):1124–1130, 2001. doi:10.1021/ci000406v. URL https://doi.org/10.1021/ci000406v. PMID: 11604013.

[12] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, and Z. Wang. HermiT: An OWL 2 reasoner. *J. Autom. Reasoning*, 53(3):245–269, 2014. doi:10.1007/s10817-014-9305-1.

[13] J. Goodman. Computer software review: Reaxys, 2009.

[14] B. A. Grzybowski, K. J. Bishop, B. Kowalczyk, and C. E. Wilmer. The'wired'universe of organic chemistry. *Nature Chemistry*, 1(1):31, 2009.

[15] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, and M. Dumontier. The chemical information ontology: Provenance and disambiguation for chemical data on the biological semantic web. *PLOS ONE*, 6(10):1–13, 10 2011. doi:10.1371/journal.pone.0025513. URL https://doi.org/10.1371/journal.pone.0025513.

[16] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck. Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214–9, 2016. ISSN 0305-1048. doi:10.1093/nar/gkv1031.

[17] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23, 2015.

[18] G. L. Holliday, P. Murray-Rust, and H. S. Rzepa. Chemical markup, xml, and the world wide web. 6. cmlreact, an xml vocabulary for chemical reactions. *Journal of chemical information and modeling*, 46(1):145–157, 2006.

[19] R. Johnson III. Cccbdb computational chemistry comparison and benchmark database. *NIST Standard Reference Database Number*, 101, 1999.

[20] S. Kim, P. A. Thiessen, T. Cheng, B. Yu, B. A. Shoemaker, J. Wang, E. E. Bolton, Y. Wang, and S. H. Bryant. Literature information in pubchem: associations between pubchem records and scientific articles. *Journal of cheminformatics*, 8(1): 32, 2016.

[21] M. Kraft and S. Mosbach. The future of computational modelling in reaction engineering. *Philos. Trans. R. Soc., A*, 368(1924):3633–3644, 2010. doi:10.1098/rsta.2010.0124.

[22] S. Kuhn, T. Helmus, R. J. Lancashire, P. Murray-Rust, H. S. Rzepa, C. Steinbeck, and E. L. Willighagen. Chemical markup, xml, and the world wide web. 7. cmlspect, an xml vocabulary for spectral data. *Journal of chemical information and modeling*, 47(6):2015–2034, 2007.

[23] A. J. Lawson. The making of reaxys-towards unobstructed access to relevant chemistry information. *The Future of the History of Chemical Information*, 1164: 127–48, 2014.

[24] M. F. Lopez, A. Gomez-Perez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems and their Applications*, 14(1):37–46, Jan 1999. ISSN 1094-7167. doi:10.1109/5254.747904.

[25] K. J. Meloche, J. Mears, and R. J. Schenck. *Intriguing Records in CAS Databases*, chapter 2, pages 21–40. doi:10.1021/bk-2013-1153.ch002.

[26] P. Murray-Rust and H. S. Rzepa. Chemical markup, xml, and the worldwide web. 1. basic principles. *Journal of Chemical Information and Computer Sciences*, 39(6):928–942, 1999. doi:10.1021/ci990052b. URL https://doi.org/10.1021/ci990052b.

[27] P. Murray-Rust and H. S. Rzepa. Chemical markup, xml, and the world wide web. 4. cml schema. *Journal of Chemical Information and Computer Sciences*, 43(3): 757–772, 2003. doi:10.1021/ci0256541. URL https://doi.org/10.1021/ci0256541. PMID: 12767134.

[28] *D. Hait and M. Head-Gordon. How accurate is density functional theory at predicting dipole moments? an assessment using a new database of 200 benchmark values. *Journal of chemical theory and computation*, 14(4):1969–1981, 2018.
The authors provide 200 benchmark dipole moments calculated using coupled cluster theory. This study then develops a hierarchy of density functionals for accurately predicting dipole moments, crucial to the development of intermolecular potentials.

[29] * F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, and M. Kraft. OntoKin: An ontology for chemical kinetic reaction mechanisms, 2019. Submitted for publication.
The authors develop an ontology capable of storing data from chemical kinetics and chemical reaction mechanisms by using OWL and formal reasoning tools. The new ontology's use is demonstrated by querying and browsing different mechanism as well as modelling the atmospheric dispersion of pollutants formed in an internal combustion engine.

[30] *M.M. Ghahremanpour, P. J. Van Maaren, and D. Van Der Spoel. The alexandria library, a quantum-chemical database of molecular properties for force field development. *Scientific data*, 5:180062, 2018.
The authors provide an open source database of quantum chemistry calculations for 2704 compounds. This establishes a key training set for the development of empirical forcefields for a variety of molecules and applications.

[31] *N. Krdzavac, S. Mosbach, D. Nurkowski, P. Buerger, J. Akroyd, J. Martin, A. Menon, and M. Kraft. An ontology and semantic web service for quantum chemistry calculations. *Journal of chemical information and modeling*, 59(7): 3154–3165, 2019.
The authors develop the OntoCompChem ontology by extending the Gainesville Core (GNVC) ontology and establish semantic interoperability between different tools used in quantum chemistry and thermochemistry calculations. The new ontology's use is demonstrated by querying the results from quantum chemistry calculations and using these to perform thermodynamic data calculations for the species of interest.

[32] *P.M Jacob and A. Lapkin. Statistics of the network of organic chemistry. *React. Chem. Eng.*, 3:102–118, 2018. doi:10.1039/C7RE00129K. URL http://dx.doi.org/10.1039/C7RE00129K.
Using graph-theoretic methods, the authors analyze the structure of a network of organic reactions built on chemical data mined from Reaxys. The authors show that on average most molecules can be synthesized within six steps from any other molecule, in what is the first study on such a large network.

[33] *S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, 10 2018. ISSN 0305-1048. doi:10.1093/nar/gky1033.
The authors summarize the information available in PubChem, the world's largest open source chemical database. The authors have also expanded PubChem to include spectral information, links to scientific articles, as well as biological properties for food and agricultural chemicals.

[34] N. M. O'Boyle. Towards a universal smiles representation-a standard method to generate canonical smiles based on the inchi. *Journal of cheminformatics*, 4(1): 22, 2012.

[35] N. S. Ostlund and M. Sopek. Applying the semantic web to computational chemistry. In A. Paschke, editor, *Proceedings of the 6th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2013)*, Edinburgh, United Kingdom, 2013. URL http://ceur-ws.org/Vol-1114/Poster_Ostlund.pdf/. Accessed February 7th, 2019.

[36] N. S. Ostlund and M. Sopek. GNVC: Gainesville core ontology - standard for publishing results of computational chemistry, ver. 0.7, 2015. URL http://ontologies.makolab.com/gc/gc07.owl. Accessed October 24th, 2018.

[37] W. Phadungsukanan, M. Kraft, J. A. Townsend, and P. Murray-Rust. The semantics of chemical markup language (cml) for computational chemistry: Compchem. *Journal of cheminformatics*, 4(1):15, 2012.

[38] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1:140022, 2014.

[39] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

[40] B. Ruscic, R. E. Pinzon, M. L. Morton, G. von Laszevski, S. J. Bittner, S. G. Nijsure, K. A. Amin, M. Minkoff, and A. F. Wagner. Introduction to active thermochemical tables: Several âĂIJkeyâĂİ enthalpies of formation revisited. *The Journal of Physical Chemistry A*, 108(45):9979–9997, 2004.

[41] B. Ruscic, R. E. Pinzon, G. Von Laszewski, D. Kodeboyina, A. Burcat, D. Leahy, D. Montoy, and A. F. Wagner. Active thermochemical tables: thermochemistry for the 21st century. In *Journal of Physics: Conference Series*, volume 16, page 561. IOP Publishing, 2005.

[42] J. M. Simmie. A database of formation enthalpies of nitrogen species by compound methods (cbs-qb3, cbs-apno, g3, g4). *The Journal of Physical Chemistry A*, 119(42):10511–10526, 2015.

[43] J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific data*, 4:170193, 2017.

[44] S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 363(6423):eaav2211, 2019.

[45] N. Stojanovic, N. Ikodinovic, and R. Djordjevic. A propositional logic with binary metric operators. *Journal of Applied Logics - IfCoLog Journal of Logics and their Applications*, 5(8):1605–1622, 2018.

[46] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge-Massachusetts, London -England, 2nd edition, 2018. ISBN 9780262039246.

[47] K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris, and D. C. De Roure. Bringing chemical data onto the semantic web. *Journal of Chemical Information and Modeling*, 46(3):939–952, 2006. doi:10.1021/ci050378m. URL https://doi.org/10.1021/ci050378m. PMID: 16711712.

[48] J. A. Townsend and P. Murray-Rust. Cmllite: a design philosophy for cml. *Journal of Cheminformatics*, 3(1):39, Oct 2011. ISSN 1758-2946. doi:10.1186/1758-2946-3-39. URL https://doi.org/10.1186/1758-2946-3-39.

[49] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Citation index