Representation of Molecules and Molecular Systems in Data Analysis and Modeling

EEN WETENSCHAPPELIJKE PROEVE OP HET GEBIED VAN DE
NATUURWETENSCHAPPEN, WISKUNDE EN INFORMATICA

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
AAN DE RADBOUD UNIVERSITEIT NIJMEGEN
OP GEZAG VAN DE RECTOR MAGNIFICUS
PROF. MR. S.C.J.J. KORTMANN,
VOLGENS BESLUIT VAN HET COLLEGE VAN DECANEN
IN HET OPENBAAR TE VERDEDIGEN
OP WOENSDAG 2 APRIL 2008
OM 13:30 UUR PRECIES

DOOR

EGON LENNERT WILLIGHAGEN

GEBOREN OP 27 OKTOBER 1974
TE ARNHEM

Promotores

Prof. dr. L.M.C. Buydens
Prof. dr. P. Murray-Rust (University of Cambridge, United Kingdom)

Copromotor

Dr. R. Wehrens

Manuscriptcommissie

Prof. E. Vlieg
Prof. dr. J. Gasteiger (Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany)
Dr. C. Steinbeck (European Bioinformatics Institute, United Kingdom)

The work presented in this thesis was supported financially by the Netherlands Organization for Scientific Research (NWO).

ISBN 978-90-9022806-8

Contents

Chapter 1

Introduction

The topic of this thesis is representation of molecules and molecular systems. Such a representation is needed to allow analysis and manipulation of chemical structures in the computer. This is of paramount importance in areas like drug design, synthesis planning, property prediction, crystal structure engineering, structure elucidation, searching in chemical literature, exchange of chemical knowledge, and structure elucidation. Many different representations have been developed, each capturing different bits of information about the molecular system under study. Unfortunately, in many cases it is unclear which part of the information is essential for a certain application. For example, although the boiling points correlates well with the number of carbon atoms in a series of alkane homologues [1], the carbon count descriptor is not generally useful for predicting other properties, or even the same property for a more diverse set of molecules. From simple physico-chemical principles, it is clear why this is the case.

However, for more complex problems there is very little a-priori knowledge that guides us in choosing appropriate descriptors. Nevertheless, in certain areas specific habits have evolved; for example, a large part of the quantitative structure-activity and structure-property relationship (QSAR and QSPR) community routinely calculates hundreds or thousands of simple molecular descriptors, and uses various variable-selection techniques to extract the most useful ones. Unfortunately, validation of this process is almost impossible due to the small size of data sets. It would a giant leap forward if we could say beforehand, based on the characteristics of the molecular system and our aim, what descriptors would be most informative. This is currently, however, still too far-fetched. Therefore, we are forced to judge the quality of the representation on the basis on the quality of the prediction: if we are able to correctly predict properties of new compounds, then we conclude that the representations contains relevant information.

This thesis studies the role of representation in modeling properties of molecular systems of organic molecules and in the exchange of molecular information. The following paragraphs give an overview on useful representations.

1.1 Molecular Representations

The two most common methods to represent organic molecules are the (systematic) name and the 2D drawing of the molecule. They identify the molecule of interest, but cannot be used for machine processing. To prevent ambiguities, conventions describing how molecules should be named and drawn are needed. IUPAC name recommendations, and line notations such as the Wiswesser Line Notation [2] and the SMILES [3], are examples for standardized conventions for labeling molecules. In addition, these representations do not include information on the 3D conformation.

The systemic naming conventions are based on chemical graphs, which represent atoms as vertices and bonds as edges, defining the exact connectivity within the molecule. For example, IUPAC recommended names, such as 2-butanol, number attachment points based on graph theory. In combination with 3D coordinate information, many descriptors have been developed to capture particular features of the molecules and more complex systems, like reactions, crystal structures and protein-ligand complexes. For example, in reaction classification the difference in chemical graphs between reactants and products is used, and docking of ligands in the active site of proteins uses force fields to calculate binding energy, using a combination of 3D coordinates and the graph representation.

At the other end of the scale we find quantum chemical descriptors, which in detail represent the 3D molecular information. Here, atoms are represented by atomic orbitals centered on points in 3D space. The molecular bonding is represented by hybridization of atomic orbitals into molecular orbitals. The disadvantage of this method is the need to find a balance between accuracy and the required computing power. Approximations can be made to reduce the complexity of the calculations, leading to semi-empirical methods like MNDO and AM1. These methods are faster but less accurate at the same time.

Force fields provide even faster energy calculations based on 3D conformations. They use a representation of molecules where atomic coordinates are complemented by rules that approximate the energy of the system based on contributions from interactions between two, three and four atoms (bond, angle and torsion interactions). The contributions are based on physical laws where the parameters are derived from experimentally determined molecular properties. While not as accurate as quantum chemistry, it is much faster and allows to analyze much larger systems, like protein structures, crystal structures and dynamical chemical processes. The accuracy strongly depends on the parametrization of the rules that approximate the interactions. Force fields have the disadvantage that this parametrization has to be repeated for each new class of molecules and type of molecular system.

The next two sections discuss applications of graph-based representations in data analyses and in property databases, and give more details on the use of quantum chemistry as representation. The sections following these discuss the need and use of numerical representations.

1.2 Chemical Graphs

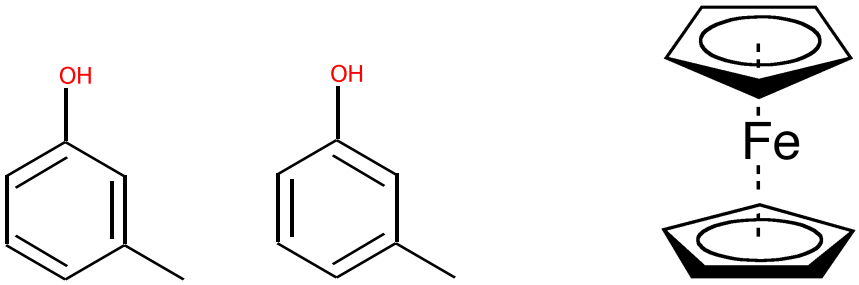
Graph-based representations are popular because they represent chemical structures in a rather intuitive way, although simplistic: molecules are atoms held together by bonds, and certain atom groups (functional groups) give rise to certain molecular properties. For example, an acid group reduces the pKa of the molecule and makes the molecule react with an amine. Searching a functional group in a molecule corresponds to finding a subgraph in the chemical graph [4, 5], when the molecule is considered a graph where atoms are vertices and bond edges.

The chemical graph also allows the use of canonization methods, such as the Morgan algorithm [6]. Using these methods, line notations can be developed which are unique for a molecule, making the look-up of molecular structures in databases much easier. The Wiswesser Line Notation is one of such notations, but nowadays the SMILES line notation is most used. However, the canonization algorithm used to generate canonical SMILES has never been published and cannot generally be used as unique molecular identifier.

The use of these line notations and the substructure searching has allowed setting up databases with molecular structures and their properties. For example, the PDB database contains crystal structures of proteins, nucleic acids and their complexes with ligands [7]. Other databases contain physical properties [8], ^{13}C and ^{1}H NMR [9], and IR spectra [8, 10]. The Chemical Abstracts Service (CAS) maintains a substance database with about 30 million chemical substances extracted from literature. At the moment this number increases by about 4000 entries each year. However, only for a fraction of these compounds more information is available in other curated databases. For example, the Cambridge Structural Database (CSD) has 400 thousand registered compounds ($\sim$1.3%) with associated crystal structures, and that number increases by only 30 thousand structures each year [11]. Moreover, the increase in information in literature is estimated at even 1 million new compounds per year from more than 700 thousand articles in chemistry-related literature [12]. In addition to these proprietary databases, open-access databases have emerged, such as PubChem [13] and ZINC [14].

These chemical graph-based databases have found many applications, such as systems for synthesis planning, where reactions are represented as changes in the molecular graph when going from reactant to the product side [15]. An example of such a tool is the Organic Chemical Simulation of Syntheses (OCSS) [16], which mimics the process of retro-synthetic synthesis planning. This has led to a number of computer-assisted synthesis design (CASD) systems, such as the LHASA system, noteworthy because it used a large knowledge base extracted from literature [17].

Another important application of chemical graphs is the use in structure generation, which fulfills a crucial role in computer-aided structure elucidation (CASE). DENDRAL is an example CASE system that elucidated molecular structures using mass spectra [18]. It derived graph constraints from the input spectrum and the molecular formula, and then generated possible structures, each of which was evaluated by comparing a predicted spectrum with the experimental one. The best spectral match was proposed as elucidated



(a) (b)

Figure 1.1: a) 2D diagrams of the two possible resonance structures of a compound with a phenyl ring. Both diagrams refer to the same compounds, but the depicted graph representations are not identical. b) 2D diagram of ferrocene, which, like all organometallic compounds, is difficult to represent with classical chemoinformatics approaches.

structure. The same approach is used for NMR-based CASE, where, in addition to the structure generation, graph theory is used to describe molecular fragments using alphanumeric codes, of which the HOSE code [19] is still widely used. Correlation of these codes with chemical shifts provides a cheap but effective method for predicting NMR spectra.

However, while the application of graph theory in chemistry has shown to be quite powerful, it is unable to reflect to full chemistry that can be found in molecules. Consider the benzene derivative diagrams shown in Figure 1.1(a). The two diagrams show non-identical graphs, but refer to the same molecular compound; the only difference is that of the resonance structures of the phenylic ring.

Organometallic compounds are excellent examples of another class of molecules that are difficult to represent using chemical graphs: they involve complex delocalized bonding systems. Ferrocene, shown in Figure 1.1(b), is an organometallic compound where two cyclopentadienyl fragments are bound to the iron. No classical two electron bonds can be drawn between the iron and any of the carbons; instead, the two six-electron π-systems of the cyclopentadienyl rings that bind to the iron. In the nineties several alternative approaches have been suggested to address this problem [20, 21, 22].

1.3 Quantum Chemistry

Quantum mechanics offers an alternative to chemical graphs as representation of molecular species. Early in the 20th century it was discovered that it can accurately describe chemical and physical properties of molecules.

Quantum chemistry takes advantage of the knowledge that electrons are not randomly distributed around the nuclei to which they are bound. Instead, their motion can be accurately described by a wave model, due to the fact that any particle both behaves as

particle as well as a wave function. Now, molecular properties, or any chemical or physical property in general, can be calculated by solving the Schrödinger equation. This leads to the exact electronic structure of the matter under study, from which any property can be calculated in arbitrary accuracy. After development of this theoretical method it was even claimed everything in chemistry was now understood; Dirac wrote [23]:

> The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.

Indeed, since the quantum mechanical description of matter is a function of all electrons and all nuclei *and* their interactions, the complexity scales rather unfortunately with the number of atoms N. Several approximations can be made to reduce the mathematical complexity, but the calculations still scale as N^4, or even N^8 for more precise calculations [24]. This means that the calculation for a molecule twice as large as ethanol, takes 16 up to 256 times as long. A small biochemically relevant molecule, e.g. nonane-4,6-dione, with only three times as many atoms, takes 81 up to 6561 times as long. Nevertheless, properties of even more complex molecules can now be calculated within hours or days. However, for current practices, like virtual screening, this is practically infinitely long. When going beyond small molecules, such as reaction mechanisms, and protein binding, the calculations become impossible.

Because neither chemical graph and quantum chemistry are practically useful for the much-needed prediction of physical, chemical and biological properties, many other representations of molecules and molecular systems have been developed. The following section discusses a class of numerical representations, which are often based on graph or quantum chemical representations, but focused to capturing molecular information relevant to the data analyzed or modeled.

1.4 Numerical Representations

One major problem is common to chemical graph and quantum chemistry representations when it comes to data analysis: their length depends on the size of the molecular system. Most statistical modeling methods, like partial least squares (PLS) [25, 26] and principal component analysis (PCA), require a fixed length representation independent of the size of the molecular system. Moreover, these methods expect that variables have the same meaning for all molecules. Additionally, many methods require the representation to be numerical, such as PLS; notable exceptions are the decision tree and random forest methods.

Many numerical representations for molecules, called molecular descriptors, have been developed [27]; examples includes descriptors which include quantum chemical features, such as the highest-occupied-molecular-orbital descriptor, or chemical graph fea-

tures, such as the fingerprint descriptor. Both of them have a fixed length and are numeric. Several programs are now available that can calculate these molecular descriptors, including Dragon, JOELib [28] and the CDK [29, 30]. The next chapter gives an overview of commonly used and recently introduced descriptors, and discusses the use of them for molecular systems with intermolecular interactions are important too.

1.5 Chemometrics

The use of these uniform-length representations has the advantage that the broad range of multivariate, statistical methods used in chemometrics can be applied. Chemometrics is traditionally described as "the application of mathematical and statistical methods to chemical measurements." [31]. Typical topics in chemometrics, therefore, are (multivariate) calibration, signal processing, experimental designs, statistics, and pattern recognition [32]. Data mining and modeling of analytical data has led to a rich field, where mathematical and statistical methods are used to analyze the chemical data. The nature of the analytical data, however, such as the high collinearity in NIR and IR spectra, has led to extensive study of multivariate regression and classification methods. These chemometrical methods turn out to have great value when used with numerical representations of molecular systems.

While chemometrics focuses on the statistical analysis of mostly multivariate chemical data, chemoinformatics generally uses the chemical graph as principal representation of molecular data. The previous section has shown that both complement each other when dealing with the understanding and prediction of properties of molecular systems (see Figure 1.2). Bridging the gap between representation of molecular structures or systems composed of molecular structures, and statistical and data mining methods, has shown to be an interesting area of research [33, 34], and standing challenges are discussed in a later section. A growing number of studies, however, use methodologies from both fields to study relationships between molecular and intermolecular information and properties of those systems. The next two sections give two illustrative examples, of the power of combining approaches, and the next chapter discusses a few applications published in recent years, referring to this type of studies as *molecular chemometrics*.

1.5.1 Example: CoMFA

Comparative Molecular Field Analysis [35] (CoMFA) is the classic example where chemoinformatics and chemometrics meet. The CoMFA method studies interactions of a molecule with its environment, often a binding site of a protein, by putting the molecules in an equidistant grid of points in three-dimensional space. At each point, the interaction energy is calculated using a hypothetical probe, for example, using the Lennard-Jones potential function and the Coulomb potential energy function. It is important to note that, because the molecules are aligned, the interaction similarities of the ligands can be compared, by means of the interaction energies of the same grid point for all molecules. It is clear that

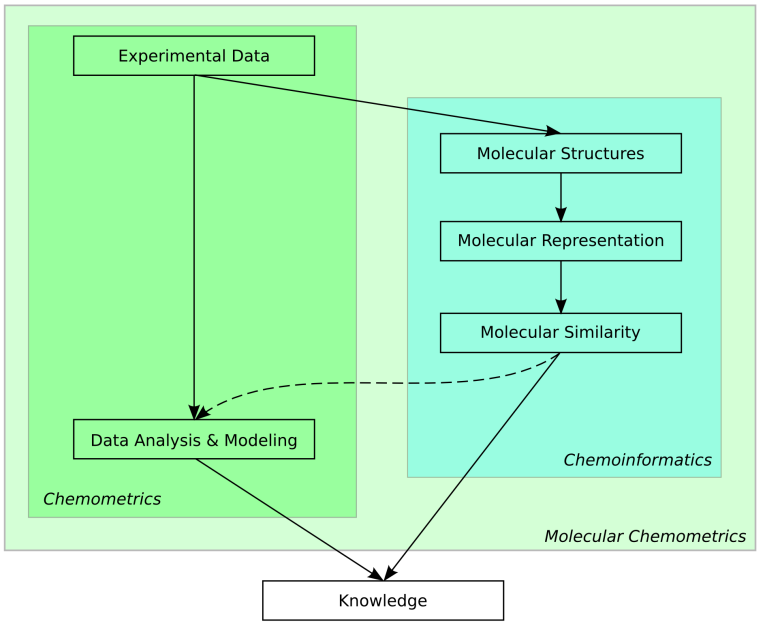


Figure 1.2: While both fields aim at increasing our knowledge about chemistry, chemometrics has traditionally focused on extracting information from analytical data (darker green area), where chemoinformatics focused on structural information of molecules (blue-ish area). Molecular chemometrics (light green area) takes approaches from both to study properties of molecules and molecular systems.

calculating the grid values involves a good deal of chemoinformatics.

Chemometrics is needed when this representation of a molecule in an environment is calibrated against the interaction of the molecule with its environment. For example, when studying ligand binding to proteins in drug design, the binding affinity is often modeled as function of this grid representation. Because of the high collinearity of interaction energies calculated for the grid points and the high variable-to-sample ratio, methods like PLS are required to correlate the matrix expansion of the grid with the activity or property.

1.5.2 Example: classification of enzyme reactions

People have been working on reaction classification since reaction databases have been set up. It was soon realized that graph-based representation of reactions, like the Ugi matrix [15], are not general enough to do database-wide reaction similarity look-up or classification. Because the Ugi matrix describes which bonds are broken, changed or formed, the size of the matrix depends on the type of reaction, and the use of this representation is restricted to reactions which involve an equal number of bonds.

Recently, however, Latino and Aires-de-Sousa have demonstrated an elegant approach to this problem [36]. To be able to compare reactions, whatever the amount of bonds are involved, a fixed-length representation is needed. Based on earlier work to numerically represent chemical bonds, they decided to use a Kohonen self-organizing map (SOM) as intermediate step to come to this fixed-length description. This is done by using a so-called MOLMAP [37], which is a trained SOM, onto which all the molecular bonds, reactant or product, are mapped. A molecule is then represented by a vector which has a length equal to the number of units of the SOM. The vector values are then determined by the specific mapping of the bonds onto the map: only the units onto which bonds map get a non-zero values, expressed in non-zero values in the corresponding vector position. This results in a sort of molecular fingerprint based on bond type information.

Similarly, this method can be used to represent reactant and product sides of a reaction. Instead of mapping just one molecule, all, for example, reactants are mapped to yield a MOLMAP for the reactant side. Likewise, a MOLMAP is created for product side. Now, the uniform length representation of the whole reaction is done in a similar way as used for the Ugi matrix: subtract the reactant side representation from the product side representation. The resulting vector is of fixed length, and represents the changes in the molecular structures during the reaction. Aires-de-Sousa and Latino showed that this approach elegantly reproduces the manually EC numbering classification of enzyme reactions [36].

1.6 Challenges

Despite the successes in molecular chemometrics in recent years, most steps in going from molecular data to chemical knowledge require human interaction: the suitability of representations and analysis needs to be explored on a problem-by-problem basis. Indeed, despite all recent efforts, the field still has several standing challenges: representation of relevant molecular information, validation of models, information loss during storage and exchange, and access to data and method implementations to improve reproducibility. This thesis focuses on the representation, storage and exchange of molecular information, discussed in the next sections.

1.6.1 Representation of Molecular Systems

New representations for molecular systems are continuously being explored to improve prediction results. It still is a challenge to find the most suitable representation for a given problem. If it does not capture the right molecular information, the modeling method will not be able to make the correlation with the property or activity of interest. The modeling method and the representation also influence each other: one representation might be suitable for a particular method, but not for others, and vice versa.

A typical example is the use of spectral similarity between, for example, NMR

spectra or crystallographic powder diffraction patterns. Such representations are peak-like; although considerable peak shifts may be the result of only minor changes at the molecular level, these may correspond to negligible changes in the property of interest. The analysis methods, however, need to deal with this property of the representation, and incorporate this prior knowledge. For peak-like spectral representations, two similarity measures have been introduced recently that take these shifts into account [38, 39]. Similarly, when modeling binding affinity the chemical graph of the molecule describes exactly which atomic groups are present in the molecule that interact with the protein. However, it does not capture any information on the actual interaction between the ligand and the protein, and a grid-like representation as used in CoMFA might be more appropriate.

Chemometric methods often provide feedback on the usefulness of (parts of) representations. For example, PLS internally determines which variables correlate best with the regressed property. These methods can also deal with non-linearities in the representation-property relationships: SVR methods use kernel functions to allow regression in a more suitable, higher-dimensional space, and various kernels, such as polynomial and RBF kernels, have been used. Specific kernels can be used to deal with particular representations, such as the Tanimoto-based kernel for chemical graph-based fingerprint representations. It is interesting to note that these kernels can be used together with other modeling methods too, such as PCA and PLS. Another often used approach to find a representation relevant to the problem at hand is variable selection, which allows the analysis method to pick descriptors or variables that provide the most accurate or predictive model.

For each new molecular system, the suitability of a chemometrical method and representation needs to be explored. The growing number of methods and approaches available from chemoinformatics and chemometrics makes validation a crucial step in this process. Interpretation of the feedback provided by the analysis methods play an important role here, in addition to statistical measures that quantify accuracy of the models. Validation is discussed in more detail in the next chapter.

1.6.2 Data Storage and Communication

Databases and small data sets are the primary source of molecular information used as input for data analysis. Accurate storage and exchange of this information is of utmost importance. Traditionally, relational database have provided uniform data, with well defined tables of information and clear relations between tables. For example, molecular databases are often defined as one table containing molecular systems as entries. Though storage and exchange of molecular information has been a well established topic in chemoinformatics for some time, the growing amount and the complexity of new data constitutes a continuing challenge. Particularly, earlier approaches did not sufficiently describe the semantics of data, leading to miscommunications, and often did not include metadata, such as details on the original source of the data, and experimental error values.

Moreover, a growing number of studies make use of information from different databases, such as proteochemometrics which models binding affinities as a function of

the molecular structure of both the ligand and the protein [40], and computer-aided structure elucidation which uses different types of spectral methods, like NMR and MS, and even physical properties like chromatographic retention time into account [41]. It is important to realize that combining information sources also means that different sources of errors have effect on the data analysis. Knowledge about the source and type of error may help during the analysis and afterward in the interpretation of results.

Exact specification of the meaning and format of data, referred to as *markup*, requires standards or common dictionaries and ontologies. Using these methods the sender of data can ensure that the receiver has the means of understanding how the data should be interpreted and may be used. For example, the markup of IR, NMR, and mass spectral information, requires explicit specification of units in which the information is expressed (absorption, intensity), and the range in which the measurement was made (e.g. 0-12 ppm). Such use of specifications still requires, however, a controlled vocabulary to be able to interpret this metadata. This is increasingly important when data sources from multiple disciplines are used.

New formats are needed that allow for this semantic markup of data and metadata. Recently, CIF has become the default format for small-molecule crystal structures. It has a well defined syntax (though not simple), and uses controlled vocabularies to add meaning to the transmitted data. More recently, the Chemical Markup Language (CML) was introduced as general transport layer for chemical data, such as molecular structures [42], molecular spectra [43] and reaction mechanisms [44]. Like CIF, CML allows the use of independent, controlled vocabularies. While this may sound more like an information-technological than a scientifical problem, but one may consider the effect of incorrectly combining data sources: the validity of scientific conclusions based on the analysis of such data is questionable. Least of all problems is that one may overlook actual patterns; more serious is the chance that false knowledge is extracted.

1.7 Selected Problems

The previous section highlights the main standing challenges in molecular chemometrics that this thesis focuses on: representation of molecular information, to aid data exchange, analysis, and modeling.

The first topic in this thesis deals with the suitability of molecular representations in property prediction; Chapter 3 describes the validation of the recently proposed use of whole NMR and IR spectra as molecular representation [45]. The chapter critically compares the use of proton and carbon NMR spectra, and compares the prediction results with models built using theoretically derived molecular descriptors calculated with Dragon. Statistical measures for model quality are used for this purpose.

While the first topic looks into representation of molecules, the second study addresses the representation of molecular systems which require incorporation of information on intermolecular interactions. Chapter 4 studies the representation of molecules in

a crystal structure environment, where the relevant information extends beyond a single molecule. Salt bridges, hydrogen patterns and other interactions define the packing pattern, and, as such, the chemical and physical properties of the crystals. Understanding the relationship between the molecular structure, interactions and crystal properties is a hot topic. For example, the prediction of likely crystal structures for a given molecular structure is yet an unsolved problem [46, 47]. One of the hurdles in this prediction process, is the definition of similarity between crystal structures.

The third topic extends on the second, and studies if numerical representations for molecular systems can be used for property prediction too, similar to QSAR and QSPR studies for single molecules. Chapter 5 describes how supervised SOMs offer an alternative method, with the added advantage that this relation can be visualized in several ways in two-dimensional plots. This in itself can be seen as a representation, similar to the application in MOLMAPs, making the information in the data easily accessible.

The three previous topics assumed the availability of accurate molecular information. To address problems originating from using more than one data source, such as different data types and error sources and types, the fourth topic deals with the markup of molecular information and metadata in a semantic-rich format: chapter 6 discusses the use of the Chemical Markup Language (CML), offering a general scheme for adding information about data units, error and other metadata, and allowing the usage of ontologies or controlled vocabularies to ensure lossless data exchange.

The fifth and last topic in this thesis deals with the (lack of) reproducibility of numerical representations. Chapter 7 discusses this issue and suggests a few methods to improve the current situation. For example, it proposes the use of a common ontology of chemoinformatics algorithms, that enables precise specification of algorithms used in a molecular chemometrical analysis; it would allow, for example, to explicitly indicate which algorithms are implemented in a particular piece of software, such as a program to calculate molecular descriptors. Additionally, reproducibility of numerical representations requires different software packages to use identical atomic properties. The chapter discusses a new and open data repository for isotope and atom type information.

The topics in this thesis only address a selection of the challenges in the field of molecular chemometrics. They show, however, how choosing a proper representation can bridge the gap between chemometrical and chemoinformatical approaches, allowing a tightly intertwined use of methods from both fields.

Bibliography

[1] H. Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69:17–20, 1947.

[2] W.J. Wiswesser. Historic development of chemical notations. *Journal of Chemical Information and Computer Sciences*, 25:258–263, 1985.

[3] D. Weininger. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

[4] L.C. Ray and R.A. Kirsch. Finding chemical records by digital computers. *Science*, 126:814–819, 1957.

[5] J. R. Ullmann. An algorithm for subgraph isomorphism. *J. ACM*, 23(1):31–42, January 1976.

[6] H. L. Morgan. The generation of a unique machine description for chemical structures - a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5:107–113, 1965.

[7] H.M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10(12):980, 2003.

[8] P. J. Linstrom and W. G. Mallard. *NIST Chemistry WebBook; NIST Standard Reference Database No. 69*. Number ; :, 2001; http:// webbook.nist.gov. Gaithersburg MD, 2001.

[9] C. Steinbeck and S. Kuhn. NMRShiftDB – compound identification and structure elucidation support through a free community-build web database. *Phytochemistry*, 65(19):2711–2717, 2004.

[10] O. Yamamoto, K. Someno, N. Wasada, J. Hiraishi, K. Hayamizu, K. Tanabe, T. Tamura, and M. Yanagisawa. An integrated spectral data-base system including ir, ms, h-1-nmr, c-13-nmr, electron-spin-resonance and raman-spectra. *Analytical Sciences*, 4(3):233–239, June 1988.

[11] F. H. Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallographica*, B58:380–388, 2002.

[12] T. Engel. Basic overview of chemoinformatics. *Journal of Chemical Information and Computer Sciences*, 46:2267–2277, 2006.

[13] Christopher P Austin, Linda S Brady, Thomas R Insel, and Francis S Collins. NIH Molecular Libraries Initiative. *Science*, 306(5699):1138–1139, Nov 2004.

[14] J.J. Irwin and B.K. Shoichet. ZINC - a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.

[15] J. Blair, J. Gasteiger, C. Gillespie, P.D. Gillespie, and I. Ugi. Representation of the constitutional and stereochemical features of chemical systems in the computer-assisted design of syntheses. *Tetrahedron*, 30:1845–1859, 1974.

[16] E. J. Corey and W. T. Wipke. Computer-Assisted Design of Complex Organic Syntheses. *Science*, 166:178–192, October 1969.

[17] E.J. Corey, W.T. Wipke, R.D. Cramer III, and W.J. Howe. Computer-aissisted synthetic analysis. facile man-machine communication of chemical structure by interactive computer graphics. *J.Am.Chem.Soc.*, 94:421–431, 1972.

[18] J. Lederberg. How dendral was conceived and born. In *Proceedings of ACM conference on History of medical informatics*, pages 5–19, New York, NY, USA, 1987. ACM Press.

[19] W. Bremser. HOSE - a novel substructure code. *Analytica Chimica Acta*, 103:355–365, 1978.

[20] A. Dietz. Yet another representation of molecular structure. *Journal of Chemical Information and Computer Sciences*, 35:787–802, 1995.

[21] E.V. Konstantinova and V.A. Skorobogatov. Molecular hypergraphs: The new representation of nonclassical molecular structures with polycentric delocalized bonds. *Journal of Chemical Information and Computer Sciences*, 35:472–478, 1995.

[22] S. Bauerschmidt and J. Gasteiger. Overcoming the limitations of a connection table description: A universal representation of chemical species. *Journal of Chemical Information and Computer Sciences*, 37(4):705–714, 1997.

[23] P.A.M. Dirac. Quantum mechanics of many-electron systems. *Proc. Roy. Soc.*, A123:714–733, 1929.

[24] J.M. Goodman. *Chemical Applications of Molecular Modelling.* Royal Society of Chemistry, 1998.

[25] P. Geladi and B. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[26] S. De Jong. SIMPLS - an alternative approach to partial least-squares regression. *Chemometrics And Intelligent Laboratory Systems*, 18(3):251–263, March 1993.

[27] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*, volume 11 of *Methods and Principles in Medicinal Chemistry.* Wiley-VCH, New York, 2000.

[28] J.K. Wegner. *Data Mining und Graph Mining auf molekularen Graphen - Cheminformatik und molekulare Kodierungen für ADME/Tox-QSAR-Analysen.* PhD thesis, Eberhard-Karls-Universität Tübingen, 2006.

[29] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 42(2):493–500, 2003.

[30] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E.L. Willighagen. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design*, 12(17):2111–2120, 2006.

[31] B. Kowalski. Chemometrics. *Anal. Chem.*, 52:112R–122R, 1980.

[32] B.K. Lavine. Chemometrics. *Analytical Chemistry*, 70(12):209–228, 1998.

[33] L.M.C. Buydens, T.H. Reijmers, M.L.M. Beckers, and R. Wehrens. Molecular datamining: a challenge for chemometrics. *Chemomet. and Intell. Lab. Syst.*, 49:121–133, 1999.

[34] R. Wehrens, R. de Gelder, G.J. Kemperman, B. Zwanenburg, and L.M.C. Buydens. Molecular challenges in modern chemometrics. *Anal. Chim. Acta*, 400:413–424, 1999.

[35] R.D. Cramer III, D.E. Patterson, and J.D. Bunce. Comparitative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carries proteins. *Journal of the American Chemical Society*, 110:5959–5967, 1988.

[36] D.A.R.S. Latino and J. Aires-de Sousa. Genome-scale classification of metabolic reactions: A chemoinformatics approach. *Angewandte Chemie International Edition in English*, 45(13):2066–2069, 2006.

[37] Qing-You Zhang and Joo Aires de Sousa. Structure-based classification of chemical reactions without assignment of reaction centers. *Journal of Chemical Information and Modeling*, 45(6):1775–1783, 2005.

[38] R. De Gelder, R. Wehrens, and J.A. Hageman. A generalized expression for the similarity spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22(3):273–289, 2001.

[39] Lorant Bodis, Alfred Ross, and Erno Pretsch. A novel spectra similarity measure. *Chemometrics and Intelligent Laboratory Systems*, 85(1):1–8, January 2007.

[40] Maris Lapinsh, Peteris Prusis, Torbjrn Lundstedt, and Jarl E S Wikberg. Proteochemometrics modeling of the interaction of amine g-protein coupled receptors with a diverse set of ligands. *Mol Pharmacol*, 61(6):1465–1475, Jun 2002.

[41] Christoph Steinbeck. Recent developments in automated structure elucidation of natural products. *Natural Product Reports*, 21(4):512–518, Aug 2004.

[42] P. Murray-Rust and H.S. Rzepa. Chemical Markup XML, and the Worldwide Web. 1. Basic Principles. *Journal of Chemical Information and Computer Sciences*, 39:928–942, 1999.

[43] S. Kuhn, P. Murray-Rust, R.J. Lancashire, H. Rzepa, T. Helmus, E.L. Willighagen, and C. Steinbeck. Chemical markup, xml, and the world wide web. 7. cmlspect, an xml vocabulary for spectral data. *J. Chem. Inf. Model.*, 2007. Accepted.

[44] Gemma L Holliday, Peter Murray-Rust, and Henry 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.

[45] R. Bursi, Y. Dao, T. Van Wijk, M. De Gooyer, E. Kellenbach, and P. Verwer. Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities. *Journal of Chemical Information and Computer Sciences*, 39:861–867, 1999.

[46] J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. Van Eijck, P. Verwer, and D. E. Williams. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica*, B56:697–714, 2000.

[47] W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scherega, B. Schweizer, M. U. Schmidt, B. P. Van Eijck, P. Verwer, and D. E. Williams. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica*, B58:647–661, 2002.

Chapter 2

Molecular Chemometrics[1]

This paper reviews literature from the past five years in the field of molecular chemometrics, which applies modeling and data analysis to molecular data. It discusses advances and standing challenges in the fields of molecular representation, similarity and diversity analysis, quantitative structure-activity and structure-property relationship modeling, and library searching.

[1]E.L. Willighagen, R. Wehrens, and L.M.C. Buydens, Crit. Rev. Anal. Chem., 2006, 36:189-198

2.1 Introduction

Molecular chemometrics is the subsection of chemometrics which applies modeling and data analysis to molecular data, such as found in diversity analysis, property or activity relationship modeling and descriptors calculation. The field includes topics as molecular representation, similarity measures for molecular data, database and diversity analysis quantitative structure activity and property relationship (QSAR/QSPR) modeling, including feature selection and model validation, and methods to automate finding and processing molecular data. We have chosen not to incorporate data of molecular mixtures, e.g. found in proteomics, image analysis, sensor data and multivariate calibration in this review, and also to exclude methods, such as quantum mechanical and force field calculations, because these do not include statistic analysis. The topic is not restricted to just data about the molecule itself, but also includes topics in which the molecule interacts with an environment. An example of this are the studies where the binding affinity is modeled by representation of both the molecule and the binding site.

The field shows a large overlap with chemoinformatics, though chemometrics tends to prefer to work with numerical data, and in chemoinformatics other fields of mathematics are strongly represented in chemoinformatics too, such as graph theory. There is also overlap with other informatics topics like data mining in general. Library searching is becoming more important again, now that more and more information is available. Increasingly, this information is available in such formats that machines can process the information, and integrate information from different sources. Extendable Markup Language (XML) applications and ontologies play a major role here, and use cases are found in the representation of molecular structures, as well as other representations of molecular information.

Problems intrinsic to this field originate from a few causes. First, the field has to deal with the huge amount of molecules in chemical space; an often cited estimation of the size of chemical space is 10^{60} unique molecules for structures with a molecular mass up to 500 atomic units [1]. Moreover, many molecular properties do not solely depend on the molecular structure itself, but may critically depend on influences from outside the molecule. For example, in binding affinity or toxicity modeling, the activities depend on the protein structure and metabolic pathways, respectively. Additionally, the discrete nature of matter at the molecular level is complicating modeling and analysis even further; we no longer deal with macroscopic properties, and simple physical laws, like the Lambert-Beer equation, get much more complicated when dealing with individual molecules.

This review discusses molecular chemometrics and touches other fields, like chemoinformatics and data mining, focusing on multivariate data analysis of molecular data. Developments reported in literature in the last five years are presented, grouped in four topics: molecular representation; chemical space, similarly and diversity; activity and property modeling; model validation; and library searching. Publications can be found in a diverse set of journals, covering many research fields including machine learning, chemometrics, analytical chemistry, bioinformatics, pharmacy and chemical information. While reviews

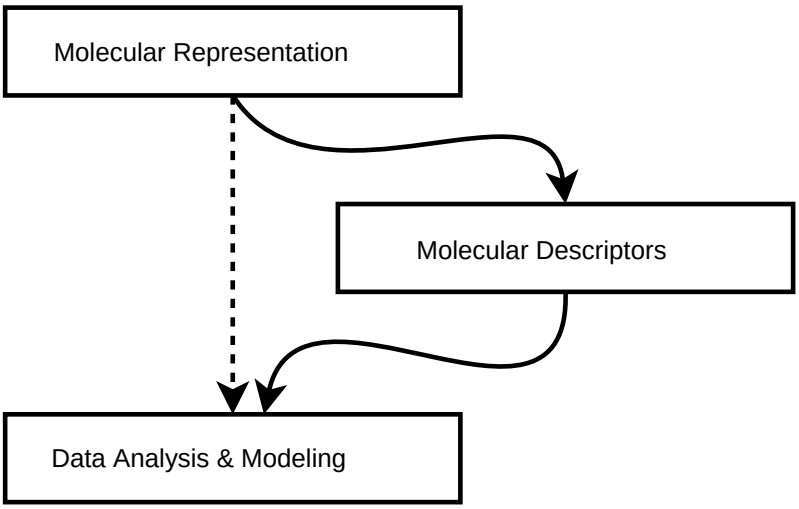


Figure 2.1: Common molecular representations, such as the quantum mechanical and the graph representation, are not well suited for direct use in statistical data analysis and modeling; instead, descriptors derived from these representations are used which match the data analysis and modeling process.

tend to be biased towards the authors preferred journals, we believe it is safe to say that the *Journal of Chemical Information and Modeling* (formerly, the Journal of Chemical Information and Computing Sciences) stands out as a source of related literature. General reading can be found in, for example, *Chemogenomics in Drug Design* edited by Kubinyi and Müller [2], *Handbook of Molecular Descriptors* by Todeschini and Consonni [3], and *Handbook of Chemoinformatics* edited by Gasteiger [4].

2.2 Molecular Representation

Central to molecular chemometrics is molecular data. The data describe chemical facts using a scientifically accepted representation. While in the eighteenth century scientists believed matter to be combinations of elements, it is now accepted that molecules are combinations of atoms held together in specific bonding patterns, governed by quantum mechanics [5]. Hence, molecular compounds are no longer identified by a pseudo molecular formula, but more detailed molecular representations are used. Molecular representations are, however, often not suitable to be use directly in data analysis and modeling; instead, descriptors derived from these representations are used which match the data analysis and modeling process. Figure 2.1 shows the relation between these representations and data analysis and modeling, and the role of descriptors in that relation.

Several basic approaches now exist to describe (small) molecular structures, each with specific function and characteristics: the first is a set of three-dimensional atomic coordinates, for example, used in crystallography. The second is the quantum mechanical

representation, where molecular structure is, in principle, a linear combination of atomic orbitals represented with three-dimensional equations, called basis functions. Deriving properties based on this representation require a lot of computation time and scales badly with a growing number of atoms. It is, therefore, not much used in molecular chemometrics; its use is outside the scope of this article. Third is the graph based representation, where atoms are nodes and bonds are edges. The abundant mathematical literature on graph theory make this representation historically successful, and even now used a lot in new research. This representation is unable to represent electron systems that cover more than two atoms, e.g. delocalization and multi atom bonds. These features are important in, for example, organometallic compounds where metals bond to electron systems, instead of atoms directly. Modifications have been proposed that allow electron systems with more than two atoms [6, 7], but application of this representation is not yet common in chemoinformatics.

Many chemometrical modeling methods, however, require a numerical, fixed length, vectorial representation of the molecular structure [8]; the above representations do not fulfill this requirement, and hence derived *descriptors* have been and still are being developed to bridge the gap between those representations and the mathematical modeling methods. These descriptors allow statistical modeling and analysis with, for example, classical methods like principle component analysis (PCA), partial least squares (PLS), neural networks (NN), and classification methods like linear discriminant analysis (LDA). Only very few methods, such as classification and regression trees (CART), do not require a numerical representation. Using distance-based clustering, is another example, and, for example, the distance measure based on the maximal common substructure: the more substructures two molecules have in common, the smaller the distance.

The currently used representations each have a specific field of application, and seem sufficiently adequate to solve a diverse set of chemical problems. However, all of them have limitations which restrict the applicability. With quantum mechanics on one side and the graph(-like) approaches on the other side, one might suggest there is room for an intermediate representation, allowing new types of derived derived descriptors for use in data analysis and modeling.

2.2.1 Molecular Descriptions

Todeschini published the *Handbook of Molecular Descriptors* in 2000 [3], giving a broad overview of known molecular descriptors at that time, but a universal descriptor has not been found [9], and the search for new descriptors has not stopped. Depending on the information content, descriptors are usually classified as 0D, 1D, 2D and 3D descriptors. The first category encompasses descriptors that do not take into account the molecular structure, e.g. the molecular mass and atom type counts. Where 2D descriptors are derived from the molecular connectivity, 1D descriptors can be considered a substructure list representation. The last category, 3D, additionally takes into account the three-dimensional geometry of the molecule. Recently, a fifth category has been proposed, 4D descriptors,

but several different definitions have been given. Todeshini defines the 4th dimension to describe the interaction field of the molecule [3], while others reserve this dimension to describe conformations of the molecule [10].

That insight in the three-dimensional interaction of ligand with protein cavities is important in the modeling biochemical endpoints, such as binding affinity, became apparent, and computationally feasible, in the last decade. Comparative Molecular Field analysis (CoMFA) is the primary example of this concept [11]. The CoMFA method studies molecule-environment interaction by putting the molecules in an equidistant grid of points in three-dimensional space. At each point, the interaction energy is calculated using a hypothetical probe, for example, using the Lennard-Jones potential function and the Coulomb potential energy function. It is important to note that, because the molecules are aligned, the interaction similarities of the ligands can be compared, by comparing the interaction energies of the same grid point for all molecules. Then, PLS is used to correlate the matrix expansion of the grid with the activity or property, though CoMFA is mostly applied to ligand-target binding properties [12, 13].

CoMFA requires, however, geometrical alignment of the molecules and only considers one conformation for each molecule, which is only a simplification of reality. Therefore, focus moved on to descriptors that are independent of the orientation of the molecules in its reference frame, and possibly even include information of multiple conformations. This was already acknowledged in 1997, for example, by Hopfinger who made a scheme which incorporated some ideas from CoMFA, but which also was alignment independent, and took into account multiple conformations [14]. Based on this concept Senese developed a 4D-fingerprint [15], which uses single value decomposition to transform the aforementioned representation. This vector representation still contains geometrical information about the possible conformers of the molecule.

In the last five years several new descriptors have been published. Bursi proposed the use of experimental or simulated infrared and 1D NMR spectra as molecular descriptor [16]. Most commonly used is the whole spectrum approach [16, 17, 18, 19, 20], which takes the whole spectrum as descriptor. Alternatively, the chemical shift of an atom present in all compounds can be used [21, 22]. The advantage of this method is that it explicitly focuses on information relevant to the problem; for example, when modeling chemical reactivity, one can take the chemical shift of an atom close to the reactive center. Spectrum-derived descriptors are used too, such as the accumulative differences in peak shifts of nuclei in octanol and in water to model the partition coefficient between those solvents [23]. The whole-spectrum approach was recently shown not to be suitable for all applications [24].

Not only new 3D descriptors have been introduced, but also new 2D connectivity-derived descriptors have been developed. For example, Faulon introduced the molecular signature which describes molecules by a vector of integers [25]. The length of the vector is determined by the number of unique fragments in the data set. Each integer, then, indicates the number of occurrences of the fragment in the molecule. The fragments themselves are called atomic signatures, and are line notions of the connectivity of that atom, very much like the HOSE code [26]. While substructure-based fingerprints only list

the occurrence of a number of fragments, the signature describes fragments for all atoms, and therefore, the full connectivity of the molecule. Randic used a specific counts paths of length three descriptor to model the boiling points of alcohols [27]. The include the steric hindrance around the oxygen he counted paths which included this oxygen; more path counts indicates more hindrance. Use of counts paths descriptors is not novel, but this example shows how descriptors can be customized for a specific problem.

Mansfield proposed a new class of 84 shape descriptors based on the volume distribution in three-dimensional space [28]. In itself, these descriptors are dependent on the alignment, which can be used to align molecules such that their volume distribution show the best overlap. Tuppurainen introduced the electronic eigenvalue descriptor which describes molecules by a smoothed function of the orbital eigenvalues put on an electron volt scale [29]. This descriptor is alignment independent and represents electronic substituent effects. The article shows the application in three QSAR studies for phenyl containing molecules. Stiefl developed a descriptor that represents molecules by a one-dimensional transform of a property mapped on the molecular surface [30]. The one dimensional mapping is to make the descriptor orientation independent, and shows the number of interactions between surface points, split up by distance between points, and the type of interaction, e.g. an interaction between a hydrophobic and a hydrophylic surface point.

Descriptors are a projection of the information in the molecular representation onto a lower dimension space, and, therefore, they inherently focus on a specific part of that information. Though many have theorized about this, and ideas and assumptions are present, it is generally still difficult to pick the right descriptor (set) for a given chemical problem. Should global information be taken in account, or local information, or the right mix of both? This limited understanding of how things work at a molecular level even becomes worse if the molecule starts to interact with an environment. Is only one specific conformer important, and if so, which one? Or is the system much more dynamic and should this be taken into account too, for example, when modeling the binding affinity of a ligand with a protein when this protein is also receptor at a different location, affecting the binding cavity of this ligand and thus the affinity? Much of this is unexplored territory.

2.2.2 Beyond the molecule

The use of descriptors to model molecular properties is not restricted to QSAR and QSPR. Other fields are picking up these descriptors too, e.g. in the field of proteomics, where Lapinsh and Prusis developed an extension of QSAR coined proteochemometrics [31, 32]. In this method, binding affinities are modeled on both the structure of the ligand, small peptides in this case, and the structure of the receptor. They found that including cross terms of the ligand and receptor descriptor blocks significantly improved the models, not surprising for binding affinity modeling, and stresses that binding affinity is a close interplay between both actors.

Crystallography is picking up statistical modeling methods too. For example, Habershon used a neural network to predict unit cell parameters from a powder diffraction pat-

terns [33], using a pattern recognition approach. Willighagen used a novel representation for molecular crystal structures and hierarchical clustering methods to classify experimental and simulated crystal structures [34]. Wehrens used powder diffraction patterns to classify molecular crystal structures using a self-organizing map [35].

Molecular reactions are another area where data analysis and modeling is taking off. While rule based analysis and classification of reactions has been around for quite some time (see e.g. [4]), the use of numerical representation in computational classification and modeling of reactions took off in the nineties, when Chen and Gasteiger used neural networks to classify a number of organic reaction types [36]. A year later they published a method that used a self-organizing map to classify organic reactions [37], where reactions were represented by a set of physical parameters for common atom in the reaction center on the reactant as well as the product side. The system was unable to classify reactions sets which did not have an atom in common. Zhang and Aires-de-Sousa addressed this problem by using a second self-organizing map, to map the reactant and product sides onto a fixed length representation [38]. The reaction itself can then be represented as the difference vector of the reactant and product sides. Using this approach, they were able to classify the metabolic reactions on a genome scale [39] and match this with the empirical EC numbering scheme [40].

As more and more experimental data becomes available, data analysis and modeling will become more important in a whole range of new scientific fields. The few mentioned in this section are just examples of what we can expect in the coming years.

2.3 Chemical Space, similarity and diversity

Chemical space is the term used to indicate the set of all possible connection tables, given a molecular formula [41]. The more atoms in this formula, the larger the number of possible connection tables. It is generally impossible to count the actual number of possible *isomers*, though attempts have been made to enumerate subsets of chemical space. Applications of the chemical space concept are found in many parts of molecular chemometrics, such as clustering of molecules, diversity analyzes, subset selection and structure enumeration.

Especially in structure-activity modeling it is a well-established assumption that structurally similar compounds are likely to exhibit similar properties [42]. However, biochemical activity does not just depend on the molecular structures, e.g. acknowledged in the earlier discussed proteochemometrics, and structures can bind in different ways to binding sites. The *similarity paradox* states that small changes in the structure can lead to large difference in activity [43]. The study of structural similarity is an extensive field; this review will not give a full view on this subject, and readers are recommended to read one of the comprehensive reviews available in literature, e.g. by Nikolova [43], Bender [42] or Maldonado [44]. Instead, it will highlight the function of similarity measures and show interesting developments in this field.

A similarity measure, the quantifier of similarity, is made up of two components: a

representation of the relevant molecular information, discussed earlier, and an index or coefficient suitable for this representation. Well-known similarity measures include the Euclidean distance for continuous-valued representations, and the Tanimoto coefficient for binary representations, such as fingerprints. An example which shows that a proper coefficient should be used, is the use of the weighted cross correlation when comparing crystal structures on the basis of an electronic radial distribution function [34]. The representation resembles a peak-like spectrum in which small peak shift indicates a small structural change; a Euclidean distance measure would fail to properly describe the structural similarity. The same problem is encountered when crystal structures are represented by their powder diffraction pattern [45].

An important application of similarity is diversity analysis. Especially in library design and subset selection, diversity is regarded an important feature, for similar reasons as those in *experimental design*. The goal of library design is to set up a library of molecular structures with a highest possible diversity, to achieve the largest coverage of chemical space. Another application is subset selection is which can be used to define independent test sets in activity and property modeling. Again coverage of chemical space is the goal. The analogy with experimental design is confirmed by the overlap in methods used; for example, D-optimal designs have been used for subset selection too [46]. Olsson introduced an improvement on this design, that addresses the occasional redundancy and replication [47].

Though several diversity and similarity measures have been developed in the past, the applicability all depends on the descriptors used, and the chemical problem for which they are used. For example, one can use the similarity in chemical space, but for biochemical activities this might not be the right measure. The same holds for diversity, and both will continue to evolve together with the use of new descriptors and the use in new fields of research.

2.4 Activity and Property Modeling

Although the idea of relating physical properties to molecular structures dates from the 19th century [44], the first mathematical model was developed by Wiener for boiling points of paraffins [48] only in 1947. Hansch was the first to model a biological activity, when he correlated toxicity of benzoic acids to their structures [49], seventeen years later. Modeling physical properties and biochemical activities is still a topic that receives a lot of attention. As discussed above, the search for new descriptors is still ongoing, as is the search for new modeling methods. Common methods used include multilinear regression (MLR), principle component regression (PCR), partial least squares (PLS) and neural networks (NN) for regression, and k-Nearest Neighbors (kNN), classification and regression trees (CART), linear and quadratic discriminant analysis (LDA, QDA) and soft independent modeling of class analogy (SIMCA). Regression methods are also used for supervised classification. Often, these methods are combined with feature selection methods, discussed later.

A method that has received growing attention is Support Vector Machines (SVM), originally developed by Vapnik [50]. Two types of SVM's have been developed: one that finds a hyperplane that separates two classes; and another one for regression, often referred to as SVR. While this hyperplane is linear in itself, the hyperplane can be sought in a space of higher dimension than the original data. The transformation of the data into this high-dimensional space is, and that is the elegance of the SVM method, equivalent to a formulation involving a so-called kernel function. Using such kernel functions makes SVM able to fit nonlinear behavior. Note that the use is not restricted to SVM, and can be used with partial least squares (PLS), too [51]. While most SVM applications use the radial distribution function (RBF), other kernels are available too, like the polynomial, anova [52] and Pearson IV [53] kernel. The latter is attractive as it can mimic both the RBF and the polynomial kernel. Any function that yields semi-definite kernel matrices can be used as kernel in SVM, allowing the use of chemoinformatics-specific kernels. For example, Lind et al. used a kernel based on the Tanimoto distance measure [54].

The number of articles that use SVM or SVR in QSAR and QSPR is steadily growing; this review cites a selection of the earlier studies. Serra found that SVM performed better than k-NN in classifying molecules according to their clastogenic behavior [55]. Byvatov compared SVM with neural networks in a drug/non-drug classification problem and found that SVM was slightly better [56]. Because SVM defines one hyperplane, it can only classify two classes. A common approach for dealing with more than two classes, is to make an one-against-all model for each class. For example, Ivanciuc used SVM classification for a three class odor problem [52]. Support Vector Regression (SVR) is an adaption of the SVM algorithm [57], and allows making regression models, where the hyperplane regresses through the data points. Burbidge was one of the first to apply SVR on structure-activity relationship data [58], and compared the performance of the method with C5.0 decision trees and neural networks, and found that SVM performed best. Bennett used SVR to model the retention times of proteins on an anion-exchange chromatography system [59].

Classification of molecules into two categories, can also be performed by a method called substructure mining. The method uses subgraph searching to find molecular fragments that are specific for one of the classes. An important feature of these methods is that the resulting model is easily interpreted; substructures can directly be related to the modeled end point. Since the number of possible substructures is enormous, these graph mining methods start from the data set itself and only consider substructures found. Kazius used this approach to predict mutagenicity [60], and Borgelt developed an algorithm that performs such a search to predict anti-HIV activity [61]. To reduce the number of substructures even further, only linear substructures, paths, may be considered, considerably speeding up the analysis [62].

While a lot of modeling methods have been tried in the past, it generally is still difficult to capture certain features of the data to model, including non-linearities and different modes of actions of the molecules itself. While the former can be addressed by using non-linear methods, or non-linear kernels as, for example, used in combination with PLS and SVR, the latter can be addressed by making local or sub models. Making

physically and (bio-)chemically relevant local models, explaining different modes of action, is one of the challenges we face in the next years.

2.4.1 Dimension Reduction

Calculating hundreds, if not thousands, of descriptors has become feasible with the modern computing power, and the general lack of understanding which molecular information is important, makes feature selection a continuing challenge. Feature selection, or variable selection, is a popular way to reduce the number of variables to be used in a model and is an alternative to, for example, PCA where linear combinations of variables are sought to describe the data efficiently. Feature selection has the advantage that the selected features are easier to interpret than linear combinations. Selections can be made such that the variables are orthogonal, or such that they contain most additional information content, e.g. calculated using the Shannon entropy. More importantly, the number of possible selections increases more than exponential with the number of variables to choose from.

Feature selection is, in essence, an optimization problem in which the goal is to find a subset of features, or variables, the give the best performance, e.g. for building QSAR models. Reasons to do this include model interpretability and reducing chance correlation. Classical methods include *forward selection*, in which is started with zero variables, and the one variable gets added that improves the model performance most. Likewise, in *backwards elimination* one starts with all variables, and the one variable get deleted that reduces the model performance least. To reduce ending up in local minima, the *stepwise method* can be used, which starts by forward selection, but allows elimination of earlier added variables after each addition. However, these methods often end up in local optimal.

Because feature selection is, in essence, an optimization problem, global optimization methods can be used too. For example, genetic algorithms (GAs) have been used a lot for this purpose [63]. Xu compared GAs with classical variable selection methods and found that the former performed better than the classical methods [64]. Other optimization methods used for feature selection include tabu search [65] and simulated annealing [66].

Recently, other optimization methods have been applied too, including ant colony optimization [67], and particle swarm optimization [68]. Like genetic algorithms, these methods evaluate variable subsets by making a new regression or classification model, using a prediction error measure, as discussed later. Alternatively, Byvatov used SVM to calculate the importance of features based on the support vectors, where features with a low importance were removed [69].

2.4.2 Model Validation

With a growing number of descriptors, modeling methods and feature selection methods, statistically sound performance estimation of classification or regression models is crucial.

This section discusses new insight and validation approaches from recent literature. Modeling methods are designed to make the best fit, and are not concerned with underlying physical and chemical principles, nor do they care about the so-called combinatorial explosion with a growing number of dependent variables. Consequently, overfitting is a serious risk [70]. Leave-one-out (LOO) and leave-more-out (LMO) cross-validation, also called k-fold cross-validation, have become increasingly popular. Baumann noted that while LOO performs well when selecting among a few alternatives, it yields overfitted models when used for feature selection [**?**]. In all cases, an independent hold-out test set, not used in any step of the training process, should be used to estimate the final performance of the model, though it is noted that for small data sets one loses predictive power [71]: Hawkins studied the behavior of the cross-validation q^2 and the R^2 for the independent test set, and proposed that with 100 or less compounds in the data set, only cross-validation should be used. Golbraikh further discusses the use of q^2, and argues that this statistic shows little correlation with predictive power [72].

Cross-validation is not the only available method, and others include y-randomization [73], and bootstrapping [74]. Mevik compared several prediction error estimators, amongst which LOO cross-validation, k-fold cross-validation, and three bootstrap based methods, for situations where the number of variables exceeds the number of objects [75]. Though differences are small, he recommends LOO cross-validation or the 0.632 bootstrap estimate, unless computational demand is too large, in which case LMO is a viable alternative. Several groups have worked on general guidelines for building QSAR and QSPR models, often consisting of a combination of a few performance statistics, like those mentioned here. The reader is pointed to articles by Todeschini [76], Eriksson [77], and Tropsha [78].

Statistical and machine learning modeling methods do not try to understand underlying physical and (bio-)chemical concepts; instead, they try to make the best fit between the sets of data, often molecular structures and some property. With large numbers of variables to describe molecular information, the chance to find a combination of them that correlates with the modeled activity explodes, even if the are unrelated to the physical and chemical concepts. This danger is addressed by using cross validation and test sets, and generally using data sets with more objects, which is becoming feasible with high-throughput experiments. Nevertheless, making scientifically sound and interpretable models is still an exciting challenge.

2.5 Library Searching

Lavine identified *Library Searching* as one of the key areas of chemometrics in 1998 [79], though the topic did not return in his later reviews [80, 81, 82]. Library searching is finding information in one or more libraries of data; with respect to molecular chemometrics, these libraries contain molecular information, such geometrical structures, spectra and physical and biochemical properties. Any of these can occur in literature, and sets of articles and abstracts are explicitly considered an (electronic) library in this review. To a large

extent these libraries are strictly formatted, for example, using relational databases from which extraction is easy. However, with the growing amount of diverse data produced in experimental work, a growing interest in sharing data on the Internet, and the trend towards a Semantic Web, data retrieval has become increasingly important.

Berners-Lee envisioned the Semantic Web in 2002, a future where information on the Internet is machine-readable, that is, where the information has semantic meaning [83, 84]. For example, a client program would not just be able to retrieve safety information on a molecule, but it could also give suggestions where the compound could be bought, what biological processes it is involving, how it could be synthesized, etcetera. While most of this information is already available on the web, such client software is currently not generally available. Returning to our molecule query, the following problems exist: data bases do not use one unique identifier for a particular molecule; chemical information is not stored in a well documented format; information does not have clear semantic meaning; information is not freely available [85, 86].

The last problem is slowly being addressed by a growing number of open access databases (see Table 2.1). The information available from these data bases is diverse, and includes crystal structures, biological activities and binding information, metabolic relations, NMR spectra and reaction mechanisms of enzymatic reactions.

ChemDB [87]	`http://cdb.ics.uci.edu/`
KEGG [88]	`http://www.genome.jp/kegg/`
Ligand.info [89]	`http://ligand.info/`
MACiE [90]	`http://www-mitchell.ch.cam.ac.uk/macie/`
NMRShiftDB [91]	`http://nmrshiftdb.org/`
PubChem [92]	`http://pubchem.ncbi.nlm.nih.gov/`
RCSB PDB [93]	`http://www.pdb.org/`
ZINC [94]	`http://blaster.docking.org/zinc/`

Table 2.1: Some examples of open access databases with molecular information.

It is noteworthy that although these data bases are open access, not all of them allow the content to be replicated, modified and redistributed, like in open source software, but it shows at least a new trend compared with previous decades where chemical database were mostly proprietary and expensive. Library searching is, obviously, not restricted to open access data bases, but is applicable to proprietary data bases too (an overview of those is found in [4]). A bigger challenge is the lack of a uniform access to both types of data bases. Access is not always available other than via a web interface or a custom program, making it difficult for machines to retrieve information. The use of semantic markup languages, mostly using the XML syntax, should change this. For molecular information the Chemical Markup Language [95] is receiving growing interest. For example, it is used to distribute physical properties of isotopes and elements [86], storage of reaction mechanisms [96], and enrichment of blogs and news feeds with chemical content [97].

Instead of making access to the data uniform, using web services and XML languages, one could also take another approach: trying to write a computer parser algorithm that takes unformatted documents, and to extract information from text, tables and figures. Given the huge amount of electronic journal articles in PDF format available now, this, though difficult, might proof very fruitful [98]. Townsend used this approach, and developed a system that uses regular expressions to extract information from experimental sections of articles [99]. Karthikeyan developed a system for finding chemical information on the Internet [100], also using regular expressions.

Finding information on an individual molecular structure has become easier too, with the publication of the InChI [101]. This unique molecular identifier will likely have an important function in the chemical semantic web [84]. Because most of current literature does not use such a unique index, one has to rely in IUPAC names, trivial names, and other naming schemes [98], and finding literature related to a query compound on just such names is not optimal. Singh uses both textual as well as molecular descriptors to address this problem, and defined a similarity measure between the query molecule and articles based on both pieces of information [102].

2.6 Conclusion

The research field of molecular chemometrics shows overlap with chemoinformatics, pharmaceutical studies, chemometrics and bioinformatics. Literature is scattered over a number of journals and the number of books in this area is increasing too. This review gives a view on the current trends in this field, and only a glimpse of the literature published in the past few years. The trends include the ongoing search for new ways to describe molecular structures, and, in a growing amount, molecules in some environment, for new multivariate modeling methods, and for new methods to deal with the ever growing amount of data in databases and on the Internet.

Though many of the problems have been addressed in literature, several important ones are still standing. For example, molecular chemometrics has to deal with an increasing amount of data to be analyzed and modeled. With the size of chemical space in mind, one cannot anticipate this amount to level off soon. The increase in computing power will not come close to what is needed. Consequently, more powerful searching, mining, feature selection, modeling and validation methods will become increasingly important.

Bibliography

[1] R. S. Bohacek, C. McMartin, and W. C. Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal Research Reviews*, 16(1):3–50, Jan 1996.

[2] H. Kubinyi and G Müller. *Chemogenomics in Drug Discovery*, volume 22 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, Weinheim, 2004.

[3] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*, volume 11 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, New York, 2000.

[4] J. Gasteiger, editor. *Handbook of Chemoinformatics*. Wiley-VCS, Weinheim, 2003.

[5] W.H. Brock. *The fontana history of chemistry*. Fontana Press, London, 1992.

[6] A. Dietz. Yet another representation of molecular structure. *Journal of Chemical Information and Computer Sciences*, 35:787–802, 1995.

[7] S. Bauerschmidt and J. Gasteiger. Overcoming the limitations of a connection table description: A universal representation of chemical species. *Journal of Chemical Information and Computer Sciences*, 37(4):705–714, 1997.

[8] K. Baumann. Uniform-length molecular descriptors for quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR): classication studies and similarity searching. *Trends in Analytical Chemistry*, 18(1):36–46, 1999.

[9] D.J. Livingstone. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *Journal of Chemical Information and Computer Sciences*, 40:195–209, 2000.

[10] J.S. Duca and A.J. Hopfinger. Estimation of molecular similarity based on 4D-QSAR analysis: Formalism and validation. *Journal of Chemical Information and Modeling*, 41(5):1367–1387, 2001.

[11] R.D. Cramer III, D.E. Patterson, and J.D. Bunce. Comparitative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carries proteins. *Journal of the American Chemical Society*, 110:5959–5967, 1988.

[12] K.H. Kim. List of CoMFA references, 1997. *Perspectives in Drug Discovery and Design*, 12-14(0):334–338, January 1998.

[13] K.H. Kim, G. Greco, and E. Novellino. A critical review of recent CoMFA applications. *Perspectives in Drug Discovery and Design*, 12-14(0):257–315, January 1998.

[14] A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, and C. Duraiswami. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society*, 119(43):10509–10524, 1997.

[15] Craig L Senese, J. Duca, D. Pan, A. J. Hopfinger, and Y. J. Tseng. 4D-fingerprints, universal QSAR and QSPR descriptors. *Journal of Chemical Information and Computer Sciences*, 44(5):1526–1539, 2004.

[16] R. Bursi, Y. Dao, T. Van Wijk, M. De Gooyer, E. Kellenbach, and P. Verwer. Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities. *Journal of Chemical Information and Computer Sciences*, 39:861–867, 1999.

[17] R. Begigni, L. Passerini, D.J. Livingstone, M.A. Johnson, and A. Giuliani. Infrared Spectra Information and Their Correlation with QSAR Descriptors. *Journal of Chemical Information and Computer Sciences*, 39:558–562, 1999.

[18] R.D. Beger, J.P. Freeman, J.O. Lay Jr., J.G. Wilkes, and D.W. Miller. Use of ^{13}C NMR Spectrometric Data To Produce a Predictive Model of Estrogen Receptor Binding Activity. *Journal of Chemical Information and Computer Sciences*, 41:219–224, 2001.

[19] A. Asikainen, J. Ruuskanen, and K. Tuppurainen. Spectroscopic QSAR Methods and Self-Organizing Molecular Field Analysis for Relating Molecular Structure and Estrogenic Activity. *Journal of Chemical Information and Computer Sciences*, 43:1974–1981, 2003.

[20] N.J.C. Bailey, Y. Wang, J. Sampson, W. Davis, I. Whitcombe, P.J. Hylands, S.L. Croft, and E. Holmes. Prediction of anti-plasmodial activity of *Artemisia annua* extracts: application of ^{1}H NMR spectroscopy and chemometrics. *Journal of Pharmaceutical and Biomedical Analysis*, 41:219–224, 2001.

[21] S.J. Vanderhoeven, J. Troke, G.E. Tranter, I.D. Wilson, J.K. Nicholson, and J.C. Lindon. Nuclear magnetic resonance (NMR) and quantitative structure-activity relationship (QSAR) studies on the transaceylation reactivity of model 1β-*O*-acyl glucuronides. II: QSAR modelling of the reaction using both computational and experimental NMR parameters. *Xenobiotica*, 34:889–900, 2004.

[22] P.V. Khadikar, V. Sharma, and R.G. Varma. Novel estimation of lipophilicity using ^{13}C NMR chemical shifts as molecular descriptor. *Bioorganic & Medicinal Chemistry Letters*, 15:421–425, 2005.

[23] Laura K Schnackenberg and Richard D Beger. Whole-molecule calculation of log p based on molar volume, hydrogen bonds, and simulated ^{13}C NMR spectra. *Journal of Chemical Information and Modeling*, 45(2):360–365, 2005.

[24] E.L. Willighagen, H.M.G.W. Denissen, R. Wehrens, and L.M.C. Buydens. On the use of ^{1}H and ^{13}C NMR spectra as QSPR descriptors. *Journal of Chemical Information and Modelling*, 46(2):487–494, 2006.

[25] Jean-Loup Faulon, Donald P Visco, and Ramdas S Pophale. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences*, 43(3):707–720, 2003.

[26] W. Bremser. HOSE - a novel substructure code. *Analytica Chimica Acta*, 103:355–365, 1978.

[27] M. Randi and S. C. Basak. A new descriptor for structure-property and structure-activity correlations. *Journal of Chemical Information and Computer Sciences*, 41(3):650–656, 2001.

[28] Marc L Mansfield, David G Covell, and Robert L Jernigan. A new class of molecular shape descriptors. 1. Theory and properties. *Journal of Chemical Information and Computer Sciences*, 42(2):259–273, 2002.

[29] K. Tuppurainen and J. Ruuskanen. Electronic eigenvalue (EEVA): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs). *Chemosphere*, 41(6):843–848, Sep 2000.

[30] N. Stiefl and K. Baumann. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure-Activity Relationship Technique. *Journal of Medicinal Chemistry*, 46(8):1390–1407, 2003.

[31] M. Lapinsh, P. Prusis, A. Gutcaits, T. Lundstedt, and J. E. Wikberg. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochimica Biophysica Acta*, 1525(1-2):180–190, Feb 2001.

[32] P. Prusis, R. Muceniece, P. Andersson, C. Post, T. Lundstedt, and J. E. Wikberg. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochimica Biophysica Acta*, 1544(1-2):350–357, Jan 2001.

[33] S. Habershon, E.Y. Cheung, K.D.M. Harris, and R.L Johnston. Powder diffraction indexing as a pattern recognition problem: a new approach for unit cell determination based on an artificial neural network. *Journal of Physical Chemistry A*, 108:711–716, 2004.

[34] E.L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L.M.C. Buydens. Method for the computational comparison of crystal structures. *Acta Crystallographica*, B61(1):29–36, Feb 2005.

[35] R. Wehrens, W.J. Melssen, L.M.C. Buydens, and R. De Gelder. Representing structural databases in a self-organizing map. *Acta Crystallographica*, B61:548–557, 2005.

[36] L. Chen and J. Gasteiger. Organic reactions classified by neural networks: Michael additions, friedel-crafts alkylations by alkenes, and related reactions. *Angewandte Chemie International Edition in English*, 35(7):763–765, 1996.

[37] L. Chen and J. Gasteiger. Knowledge discovery in reaction databases: Landscaping organic reactions by a self-organizing neural network. *Journal of the American Chemical Society*, 119(17):4033–4042, 1997.

[38] Qing-You Zhang and Joo Aires de Sousa. Structure-based classification of chemical reactions without assignment of reaction centers. *Journal of Chemical Information and Modeling*, 45(6):1775–1783, 2005.

[39] D.A.R.S. Latino and J. Aires-de Sousa. Genome-scale classification of metabolic reactions: A chemoinformatics approach. *Angewandte Chemie International Edition in English*, 45(13):2066–2069, 2006.

[40] K. Tipton and S. Boyce. History of the enzyme nomenclature system. *Bioinformatics*, 16(1):34–40, Jan 2000.

[41] Christopher M Dobson. Chemical space and biology. *Nature*, 432(7019):824–828, Dec 2004.

[42] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry*, 2(22):3204–3218, Nov 2004.

[43] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity - a review. *QSAR & Combinatorial Science*, 22:1006–1026, 2003.

[44] Ana G Maldonado, J. P. Doucet, Michel Petitjean, and Bo-Tao Fan. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular Diversity*, 10(1):39–79, Feb 2006.

[45] R. De Gelder, R. Wehrens, and J.A. Hageman. A generalized expression for the similarity spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22(3):273–289, 2001.

[46] Paola Gramatica, Pamela Pilutti, and Ester Papa. Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *Journal of Chemical Information and Computer Sciences*, 44(5):1794–1802, 2004.

[47] I-M. Olsson, J. Gottfries, and S. Wold. Controlling coverage of d-optimal onion designs and selections. *Journal of Chemometrics*, 18(12):548–557, 2004.

[48] H. Wiener. Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69:17–20, 1947.

[49] C. Hansch and T. Fujita. ρ-σ-π analysis - a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86:1616–1626., 1964.

[50] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[51] B. Walczak and D. L. Massart. The radial basis functions – partial least squares approach as a flexible non-linear regression technique. *Analytica Chimica Acta*, 331(3):177–185, September 1996.

[52] O. Ivanciuc. Structure-odor relationships for pyrazines with support vector machines. *Internet Electronic Journal of Molecular Design*, 1:269–284, 2002.

[53] B. Üstün, W.J. Melssen, and L.M.C. Buydens. Facilitating the application of support vector regression by using a universal pearsonnext term vii function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81:29–40, March 2006. In press.

[54] T.Maltseva P.Lind. Support vector machines for the estimation of aqueous solubility. *Journal of Chemical Information and Computer Sciences*, 43:1855–1859, 2003.

[55] P.C.Jurs J.R.Serra, E.D.Thompson. Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chemical Research in Toxicology*, 16:153–163, 2003.

[56] Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, 43(6):1882–1889, 2003.

[57] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Berlin, 1995.

[58] R. Burbidge, M. Trotter, B. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry*, 26(1):5–14, Dec 2001.

[59] M. Song, C.M. Breneman, J. Bi, N. Sukumar, B.P. Bennett, S. Cramer, and N. Tugcu. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences*, 42:1347–1357, 2002.

[60] Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, Jan 2005.

[61] Christian Borgelt, Thorsten Meinl, and Michael Berthold. MoSS: A program for molecular substructure mining. In Bart Goethals, Siegfried Nijssen, and Mohammed J. Zaki, editors, *Proceedings of OSDM 2005*, pages 6–15, 2005.

[62] Christoph Helma, Tobias Cramer, Stefan Kramer, and Luc De Raedt. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *Journal of Chemical Information and Computer Sciences*, 44(4):1402–1411, 2004.

[63] R. Leardi. Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics*, 15(7):559–569, August 2001.

[64] L. Xu and W-J. Zhang. Comparison of different methods for variable selection. *Analytica Chimica Acta*, 446:477–483, 2001.

[65] J.A. Hageman, M. Streppel, R. Wehrens, and L.M.C. Buydens. Wavelength selection with tabu search. *Journal of Chemometrics*, 17:427–437, 2003.

[66] P.C. Jurs and J.M. Sutter. *Adaption of Simulated Annealing to Chemical Optimization Problems*, volume 15 of *Data Handling in Science and Technology*, chapter Selection of molecular descriptors for quantitative structure-activity relationships. Elsevier, Amsterdam, 1995.

[67] Qi Shen, Jian-Hui Jiang, Jing-Chao Tao, Guo-Li Shen, and Ru-Qin Yu. Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. *Journal of Chemical Information and Modeling*, 45(4):1024–1029, 2005.

[68] Qi Shen, Jian-Hui Jiang, Chen-Xu Jiao, Guo-Li Shen, and Ru-Qin Yu. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *European Journal of Pharmaceutical Sciences*, 22(2-3):145–152, Jun 2004.

[69] Evgeny Byvatov and Gisbert Schneider. SVM-based feature selection for characterization of focused compound collections. *Journal of Chemical Information and Computer Sciences*, 44(3):993–999, 2004.

[70] Douglas M Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

[71] D.M. Hawkins, S.C. Basak, and D. Mills. Accessing Model Fit by Cross-Validation. *Journal of Chemical Information and Computer Sciences*, 43:579–586, 2003.

[72] A. Golbraikh and A. Tropsha. Beware of q^2! *Journal of Molecular Graphics and Modelling*, 20(4):269–276, 2002.

[73] S. Wold and L. Eriksson. *Chemometrics Methods in Molecular Design*, chapter Statistical Validation of QSAR Results, pages 309–318. VCH, Weinheim (Germany), 1995.

[74] R. Wehrens and W.E. Van der Linden. Bootstrapping principal-component regression models. *Journal of Chemometrics*, 11(2):157–171, 1997.

[75] B-H. Mevik and H.R. Cederkvist. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, 18(9):422–429, 2004.

[76] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Analytica Chimica Acta*, 515(1):199–208, 2003.

[77] Lennart Eriksson, Joanna Jaworska, Andrew P Worth, Mark T D Cronin, Robert M McDowell, and Paola Gramatica. Methods for reliability and uncertainty assessment

and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives*, 111(10):1361–1375, Aug 2003.

[78] A. Tropsha, P. Gramatica, and V. K. Gombar. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, 22(1):69–77, April 2003.

[79] B.K. Lavine. Chemometrics. *Analytical Chemistry*, 70(12):209–228, 1998.

[80] B.K. Lavine. Chemometrics. *Analytical Chemistry*, 72(12):91–98, 2000.

[81] B.K. Lavine and J. Workman. Chemometrtics. *Analytical Chemistry*, 74(12):2763–2770, 2002.

[82] B. Lavine and J.J. Workman. Chemometrics. *Analytical Chemistry*, 76(12):3365–3372, 2004.

[83] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 28–37, May 2001.

[84] Simon J Coles, Nick E Day, Peter Murray-Rust, Henry S Rzepa, and Yong Zhang. Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic & Biomolecular Chemistry*, 3(10):1832–1834, May 2005.

[85] Peter Murray-Rust, Henry S Rzepa, Simon M Tyrrell, and Yong Zhang. Representation and use of chemistry in the global electronic age. *Organic & Biomolecular Chemistry*, 2(22):3192–3203, Nov 2004.

[86] R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, R. Rzepa, S. Steinbeck, J. Wegner, and E.L. Willighagen. The blue obelisk - interoperability in chemical informatics. *Journal of Chemical Information and Modelling*, 46:991–998, 2006.

[87] Jonathan Chen, S. Joshua Swamidass, Yimeng Dou, Jocelyne Bruand, and Pierre Baldi. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, 21(22):4133–4139, Nov 2005.

[88] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, Jan 2002.

[89] Marcin von Grotthuss, Grzegorz Koczyk, Jakub Pas, Lucjan S Wyrwicz, and Leszek Rychlewski. Ligand.Info small-molecule Meta-Database. *Combinatorial Chemistry and High Throughput Screening*, 7(8):757–761, Dec 2004.

[90] Gemma L Holliday, Gail J Bartlett, Daniel E Almonacid, Noel M O'Boyle, Peter Murray-Rust, Janet M Thornton, and John B O Mitchell. MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*, 21(23):4315–4316, Dec 2005.

[91] C. Steinbeck, S. Kuhn, and S. Krause. NMRShiftDB – constructing a chemical information system with open source components. *Journal of Chemical Information and Computer Sciences*, 43(6):1733 – 1739, 2003.

[92] Christopher P Austin, Linda S Brady, Thomas R Insel, and Francis S Collins. NIH Molecular Libraries Initiative. *Science*, 306(5699):1138–1139, Nov 2004.

[93] H.M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide protein data bank. *Nature Structural Biology*, 10(12):980, 2003.

[94] J.J. Irwin and B.K. Shoichet. ZINC - a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005.

[95] P. Murray-Rust and H.S. Rzepa. Chemical Markup XML, and the Worldwide Web. 1. Basic Principles. *Journal of Chemical Information and Computer Sciences*, 39:928–942, 1999.

[96] Gemma L Holliday, Peter Murray-Rust, and Henry 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.

[97] P. Murray-Rust and H.S. Rzepa. The next big thing: From hypermedia to datuments. *J.Digital Information*, 5:248, 2004.

[98] Debra L Banville. Mining chemical structural information from the drug literature. *Drug Discovery Today*, 11(1-2):35–42, Jan 2006.

[99] Joe A Townsend, Sam E Adams, Christopher A Waudby, Vanessa K de Souza, Jonathan M Goodman, and Peter Murray-Rust. Chemical documents: machine understanding and automated information extraction. *Organic & Biomolecular Chemistry*, 2(22):3294–3300, Nov 2004.

[100] M. Karthikeyan, S. Krishnan, and A.K. Pandey. Harvesting chemical information from the internet using a distributed approach: Chemxtreme. *Journal of Chemical Information and Computer Sciences*, 46, 2006.

[101] S.E. Stein, S.R. Heller, and D. Tchekhovski. An open standard for chemical structure representation - The IUPAC Chemical Identier. In *Nimes International Chemical Information Conference Proceedings*, pages 131–143, 2003.

[102] Suresh B Singh, Richard D Hull, and Eugene M Fluder. Text Influenced Molecular Indexing (TIMI): a literature database mining approach that handles text and chemistry. *Journal of Chemical Information and Computer Sciences*, 43(3):743–752, 2003.

Chapter 3

On the use of ^{1}H and ^{13}C 1D NMR spectra as QSPR descriptors[1]

Recently, 1D NMR and IR spectra have been proposed as descriptors containing 3D information. And, as such, said to be suitable for making QSAR and QSPR models where 3D molecular geometries matter, for example in binding affinities. This article presents a study on the predictive power of 1D NMR spectra-based QSPR models using simulated proton and carbon 1D NMR spectra. It shows that the spectra-based models are outperformed by models based on theoretical molecular descriptors, and that spectra-based models are not easy to interpret. We therefore conclude that the use of such NMR spectra offers no added value.

[1] E.L. Willighagen, H.M.G.W. Denissen, R. Wehrens, L.M.C. Buydens, J. Chem. Inf. Comp. Sci., 2006, 46:487-494

3.1 Introduction

After several decades, methodological research on quantitative structure activity/property relationship (QSAR and QSPR) modeling still receives much attention [1]. Focus has been both on new modeling methods, e.g. support vector regression [2], as well as on describing the molecular structures. Even though many theoretical molecular descriptors have been developed in the past to represent molecular structures in mathematical models, new descriptors are being introduced every day. While some descriptors are more useful in some applications, no general descriptor type is available that can be used for all QSAR/QSPR studies.

Descriptors capture certain features of the molecular structure and are often categorized into descriptor classes according to the information they represent [3]. The first class of descriptors, including the Wiener index and the Kier shape descriptors, represents topological properties of a molecule. These only describe the connectivity and not the geometry. The second class represents descriptors which describe geometrical properties and contains descriptors like WHIM descriptors and solvent-accessible surface areas. Such descriptors are often named 3D descriptors, while the former are 2D descriptors. The third class of descriptors contains the electronic descriptors, describing the electronic features of the molecules. Examples include the HOMO and LUMO energies, and electronegativity. The fourth and last class of descriptors contains features derived chemical formula, like atom counts.

While such a classification is somewhat artificial, the notion that a descriptor may represent geometrical information instead of just topological information is important. If the modeled activity is highly depending on the 3D geometry of the molecule, which is, for example, the case with binding affinities, the descriptors need to represent geometrical features of the molecules. When the 3D geometry is relatively unimportant, for example in the case of solubility, then such features need not to be present in the descriptor set in order to obtain predictive models.

Recently, IR and 1D NMR spectra have been proposed as 3D molecular descriptors [4] in QSAR modeling. Both spectra types show unique spectra for different compounds. Moreover, these spectra depend on the 3D geometry of the molecules, which can, for example, be seen with the low temperature NMR spectrum of cyclohexane where the axial and equatorial hydrogens show different chemical shifts. Additionally, the through space spin-spin coupling in proton NMR is used as a restriction in elucidating 3D protein structures. From these examples it can be concluded that spectra indeed contain 3D information, but unlike grid-based representations, such as CoMFA [5], spectra do not require molecular alignment prior to analysis, simplifying the model building considerably. It is questionable, though, whether this 3D information is useful and relevant for modeling the activities or properties.

QSAR and QSPR models correlate molecular structures with a measured activity or property using numerical descriptors, attempting to capture the relation between the chemical and physical information in the descriptors with that activity. When modeling

water solubilities or partition coefficients, the model will focus on descriptors describing features that have a high influence, positively or negatively, on the activity. Consequently, when using NMR spectra as descriptors, the modeling method will find shift areas which correlate with the activity. For example, if ^{1}H NMR are used as descriptor, the peak shift areas where phenyl protons are found, are expected to negatively correlate with the water solubility and positively with the octanol/water partition coefficient.

^{1}H NMR and ^{13}C NMR spectra have been used in several QSAR and QSPR studies [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Three different methods have been used in those studies to include NMR descriptors, though other approaches can be considered too. Most used is the whole spectrum approach [4, 6, 7, 8, 9, 10, 11, 12, 13, 14]. As explained in the previous paragraph, shift areas will then correlate with the modeled activity. Optionally, specific features of the spectrum can be selected, for example, a few areas where relevant information is found.

A second method that uses NMR spectra, uses the chemical shift of an atom common to all compounds [15, 17]. The advantage of this method is that it explicitly focuses on information relevant to the problem; for example, when modeling chemical reactivity, one can take the chemical shift of an atom close to the reactive center. Obviously, this method is restricted to homologous compound series, and that peaks need to be assigned, restricting its general use.

A third method that uses NMR spectral information, is specific for modeling the logarithm of the partition coefficient between octanol and water (logP) [16]. In this research advantage was taken from the fact that compounds have different NMR spectra in the two solvents. By summing the differences in chemical shifts for the atoms have in the two solvents, an estimate is made of the solvent effects on the whole molecule. This difference was used to model the activity, though the influence on the predicted activity is rather small, if significant.

Generally, small data sets were used in these QSAR and QSPR studies, in many cases without independent test sets, making it hard to study the true predictive power of the constructed models. The current article studies the potential of the proposed use of simulated ^{1}H NMR and ^{13}C NMR spectra as molecular descriptor and compares it with theoretically calculated molecular descriptors, derived from a symbolic representation of the molecules, in this case the connection table. Three data sets are used, of which three contain a diverse set of more than 100 compounds, and have physical properties as end points. For these data sets any possible 3D information in the descriptor is unlikely to be important. Results for a fourth data set with binding affinities as end point (used in the original NMR-QSAR article, [4]), for which such 3D information would be important, has been left out because modeling the activity was unsuccessful with any descriptor used. For all data sets an independent test set is used to be able to estimate the true predictive power. As in most relevant literature, we used full spectra: it does not require peak assignment, nor one or more atoms to be common to all compounds. Other approaches used in literature did not show clear advantages over the full spectrum approach, and are not further considered in this paper.

3.2 Experimental

3.2.1 Methods

1D NMR spectra have been simulated with ACD's ^{1}H Predictor and ^{13}C Predictor version 7.0. Proton NMR spectra were scaled to a resolution of 0.05 ppm per data point in the range of 0 to 11 ppm using custom scripts, resulting in 220 variables. Likewise, carbon NMR spectra were scaled to a resolution of 1 ppm in the range of 1 to 220 ppm, also giving 220 variables.

Theoretical molecular descriptors are calculated with Dragon 5, though alternatives are abundant including open-source variants like JOELib and the CDK [18, 19]. Binary and constant descriptors are removed, resulting in about 1200 to 1300 descriptors, depending on the data set, from which 220 descriptors were randomly selected, to give a descriptor set with the same number of variables are the NMR sets. We used models based on these descriptors for benchmarking only, because it was not our goal to make optimal models based on these descriptors. Therefore, we explicitly did not do feature-selection on these descriptors, as is usually done. Columns were autoscaled in order to make each descriptor equally important.

The Dragon 5 program defines 20 different descriptor classes. Replicate random selections for the data sets at least 18 of all descriptor classes represented (not shown). This indicates that the used subset of 220 descriptors has a high diversity in information content, including constitutional, topological, connectivity, geometrical descriptors and many others, covering molecular properties that correlate with dipole moment, weight and hydrophobicity. For completeness, the random selections for the three data sets used to calculated the presented results, are found in the Supporting Information.

The amount of information in the X matrices for the descriptor sets is first studied by investigating the mathematical ranks of those matrices. The maximum rank equals the lower value of the rows and columns of the matrix. A matrix rank lower than this maximum indicates correlation in the matrix in either the columns or the rows. By comparing ranks for the different descriptor types, the differences can only be caused by correlation between columns.

Partial Least Squares (PLS) [20] was then used to make mathematical models that relate the molecular descriptor set (X matrix, consisting of either spectra or theoretical descriptors) with the activity (Y vector). To pick the number of latent variables for the model, we used the root mean square error (RMSE) of leave-one-out cross validation (LOO-CV). This is done using an automatic procedure that picks the lowest number of LV's that has a cross validation error is lower than one standard deviation above the absolute minimum in that error [21]. This might not be the optimal decision, as choosing the best number of LV's is a difficult problem, but at least it is conservative and consistent.

To validate the performance of the different types of descriptors, several statistics are monitored that describe the differences in predicted and real activity: the, in QSAR/QSPR

research commonly used, R^2 and Q^2 [22] and the root mean square error of cross validation (RMSECV) and of prediction (RMSEP). The RMSEP is used to get a independent estimate of predictive power of the model for unknowns. For each data set, five random divisions in training and test sets have been used to get an estimate on the errors on these statistics due to these divisions.

The RMSE values for the models are compared with a no-information limit which is calculated from the activities for a data set. It considers a QSPR model where the predicted activity is the mean activity for all compounds in the data set, i.e. $y_{pred} = \bar{y}$. Obviously, the RMSE of a truly predictive PLS models should be significantly lower than this limit.

Calculations have been performed in the statistical program R 2.1.0 [23] on a dual AMD64 processor system running the 64 bits Debian GNU/Linux 3.1 (sarge) operating system. The pls.pcr package was used for building the PLS models [24].

3.2.2 Data

This article presents the results of three data sets. These data sets were used to compare the power of NMR spectra in QSPR modeling to theoretical molecular descriptors. The first data set, called WS, contains 431 compounds with aqueous solubilities. This set is a subset of a published test set that was selected on diversity [25]. Models were trained with 400 compounds, and the remaining 31 compounds were used as test set. The second data set, called BP, contains 269 heteroatom-containing compounds excluding nitrogen compounds (data set II from [26]) with associated boiling points. Eight compounds from the original data set lacking any hydrogens were removed. A test set with 42 compounds was used, while training models was done with 227 compounds. The third data set, called LogP, contains 154 compounds with associated log P values [16], the partition coefficients between octanol and water. Models were trained with 120 compounds, and the remaining 34 compounds were used as test set. Activities and InChI's for these three data sets can be found in the Supporting Information.

3.3 Results

3.3.1 Data rank

The median ranks of the training X matrices for the five random training/test set divisions are shown in Table 3.1. For all data sets, the rank for the Dragon descriptor set was found to be equal or close to the minimum of the number of rows and columns of the matrix. For proton NMR the rank was lower, and for carbon NMR the rank was lowest. This indicates that carbon NMR shows most correlation. Dragon descriptors show the least correlation of all three descriptor types. Less correlation does not directly mean better PLS models, though. That the ranks for the NMR spectra are lower than the maximum is

	^{1}H NMR	^{13}C NMR	Dragon	Limit
WS	198	195	219	220
BP	163	157	219	220
LogP	120	117	119	120

Table 3.1: The median ranks of five randomly chosen training sets for descriptor types for the three data sets. The limit is the maximal rank possible for that descriptor type and data set. Clearly, Dragon-based descriptor matrices are always of nearly full rank, which indicates a high amount of uncorrelated information.

not surprising. Spectra normally have shift areas where no peaks are found. Those matrix columns have zero intensity for all compounds, and are obviously correlated.

3.3.2 Predictivity

The RMSECV plots to select the number of latent variables for the three descriptor types typically look like those for the LogP data set shown in Figure 3.1. Error plots for carbon NMR and Dragon models show that the RMSECV drops with the first few number of latent variables, after which it stabilizes and then increases. This can be explained by assuming that the first few LV's add information to the model, after which the model starts to be overtrained. For Dragon-based models, typically 6 or 7 LV's are chosen and for carbon NMR-based models typically 3 or 4 LV's are chosen. The error plots for proton NMR look different: the error rises from the first or second LV on. For this descriptor type, only one or two LV's are chosen.

The performance of the models is studied using several statistics. Five replicate training/test set divisions are used to allow comparing calculated statistics; a small improvement in one of the statistics might not indicate an significant improvement of the model. Taking into account the errors on the statistics is important when picking one model over another.

When looking at the R^2 and Q^2 values (see Figure 3.2) for the WS, BP and LogP data sets, it is apparent that it was not possible to create acceptable models based on proton NMR, as shown by the low R^2 and Q^2 values. While carbon NMR based models perform reasonably, they are still outperformed by the Dragon-based models which have higher R^2 and Q^2 statistics. Only the Dragon-based models have statistics approaching the optimal value of 1.0. It is also clear that the error due to the choice of the training/test set division is much smaller than the differences between the three descriptor sets. This strengthens our conclusion, that the differences between the descriptor sets are significant.

These results are confirmed by the RMSECV values and the independent RMSEP's for the independent test sets as shown in Figure 3.3. The RMSE values show that proton NMR in general does not show a prediction performance significantly better than the no-information limit provided by the $y_{pred} = \bar{y}$ model. Also, in agreement with the R^2 and Q^2 statistics, is the observation that carbon NMR performs reasonably, but is outperformed

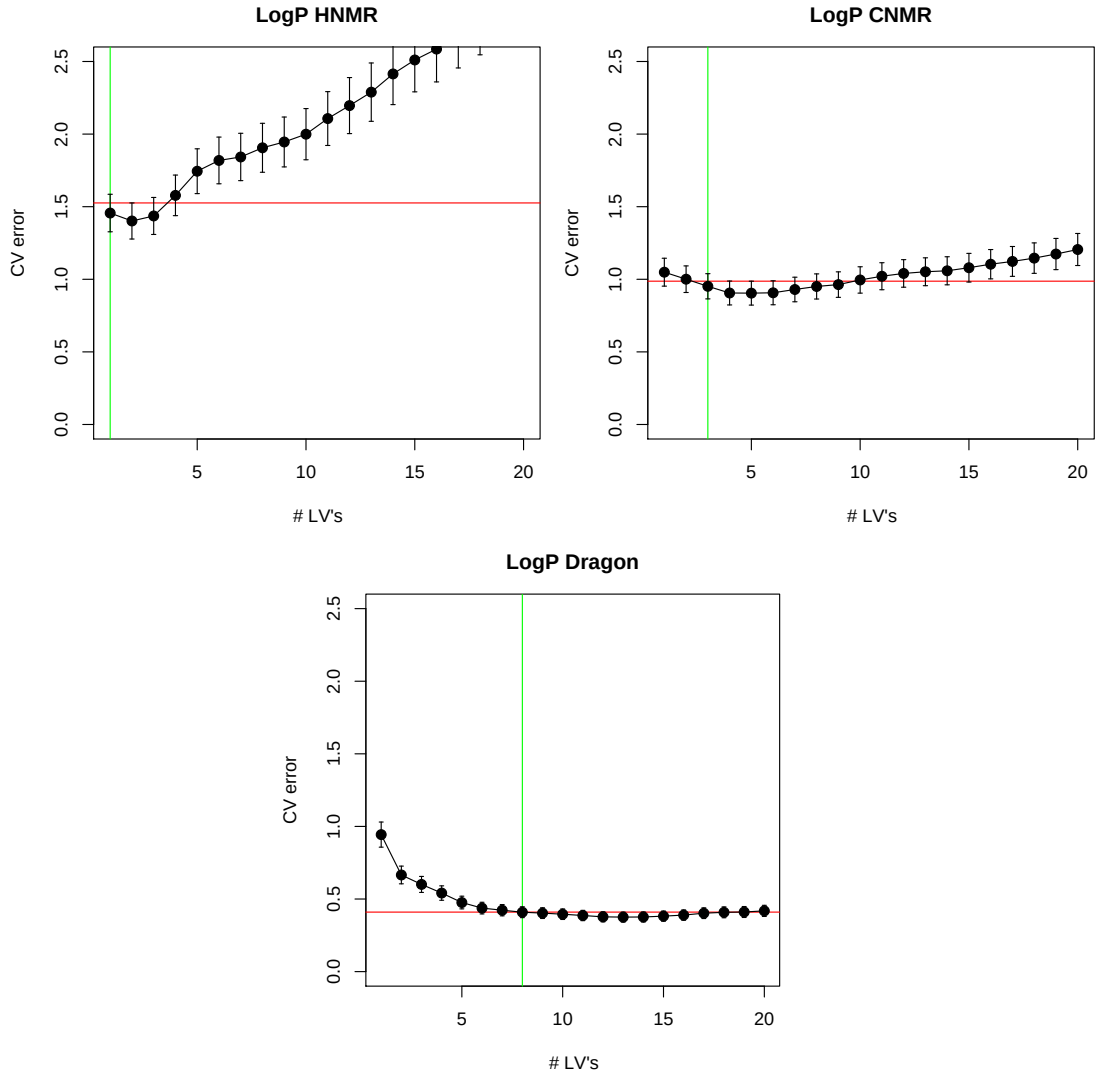


Figure 3.1: The chosen number of latent variables is based on the LOO-CV error. The LogP plots for the three data sets are representative for the other data sets. The red line indicates one standard deviation above the absolute minimum in the LOO-CV error used to choose the number of latent variables for the PLS model, which is indicated by the green line.

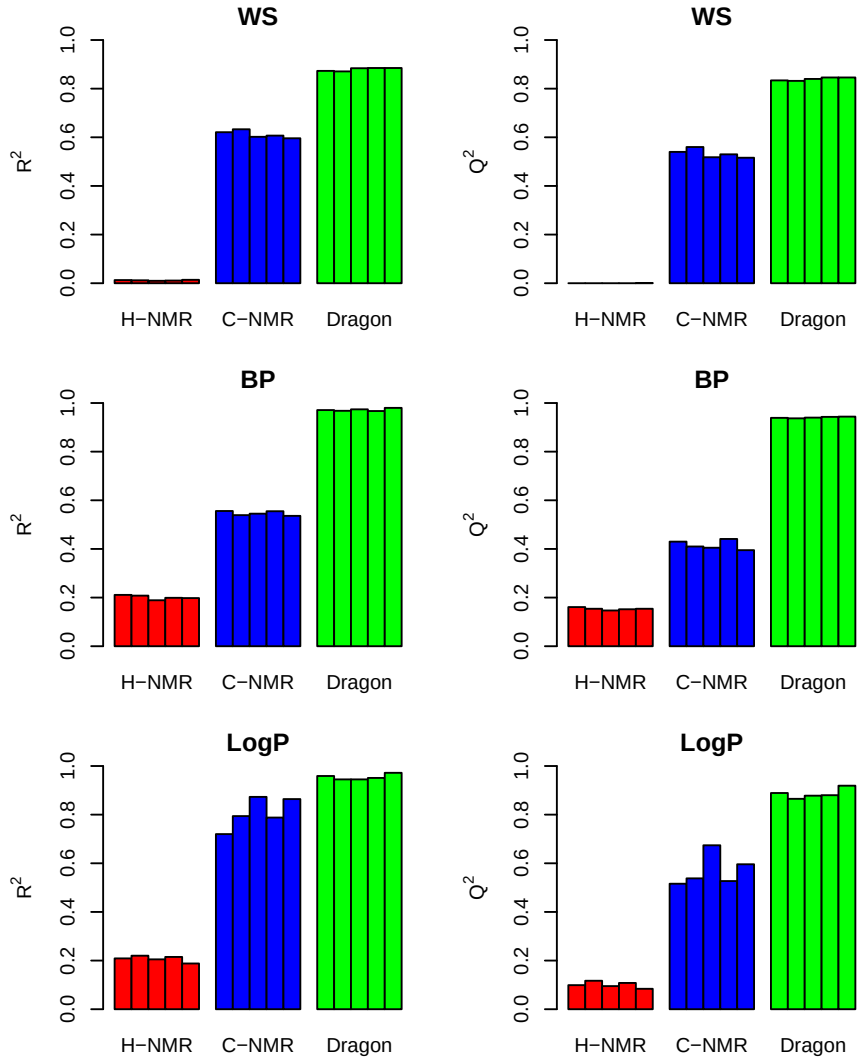


Figure 3.2: The internal performance statistics R^2 and Q^2 for the three data sets, each with five random training/test set divisions. In all cases the Dragon-based descriptors clearly perform best.

by the Dragon-based models, which clearly have lower prediction errors.

In addition to looking at numerical prediction error differences, one can also look at $y_{measured} - y_{predicted}$ plots. For the WS, BP and LogP data sets, the three plots for the different descriptor sets look similar to those for the LogP data set shown in Figure 3.4. The recall, i.e. the prediction of the training samples, is plotted with black open circles, and the test set predictions are drawn with red dots. These plots confirm that proton NMR-based models do not improve significantly on the $y_{pred} = \bar{y}$ model. The plot for carbon NMR shows regression around the $y_{pred} = y_{measured}$ line, but the regression is clearly better for the Dragon-based models. The results in Figure 3.4 are based on one random test set, but are representative for other training/test set divisions.

3.3.3 Model interpretation

In addition to looking at the predictive power of the models, the explanatory nature of the models is often informative too. In PLS this is done by looking at the regression vectors. In NMR one would expect shift ranges with high positive coefficients, where peaks occur characteristic for molecular fragments, positively affecting the activity; and ranges with high negative coefficients for groups which negatively affect the activity.

Such shift ranges are found for carbon NMR, as shown in Figure 3.5 for the LogP data set. Chemical shift ranges where peaks are to be expected for molecular fragments with electron withdrawing atoms, like C-O and C=O, have a negative influence on the calculated property. Additionally, ranges where hydrophobic groups, like CHx and C=C, are found, show positive coefficients. The regression coefficients do not seem to provide information beyond the observed influence of these molecular atom groups. The blue lines indicate $\pm$ standard deviation for the five random training/test set divisions, and show that the patterns are found for all five replicates.

Proton NMR also seems to show some pattern. Clearly, the area between 3 and 4 ppm has positive contributions. In this area, shifts are expected for protons connected to carbons that bond with heteroatoms, like oxygen and nitrogen, indicating a positive effect of polar groups. This contradicts the interpretation of the PLS coefficients for the carbon NMR models. Moreover, the coefficients are three orders of magnitude smaller than those for the carbon NMR models. Even though the regression vector seems to contain information, proton NMR spectra are not predictive.

The regression vector of the Dragon-based model was sorted in ascending order to allow easier interpretation of the significances of the coefficients. As it is not the intention to produce the best possible models based on theoretical molecular descriptors, we will not discuss which individual descriptors had high (positive or negative) coefficients. We do note that for all build models, the 20 descriptors with highest coefficients represent at least 8 different descriptor classes, with an average of 10. It is important to note that all descriptors with high coefficients show this for all five training/test set divisions. From these results, we conclude that by randomly picking 220 descriptors from the larger set, predictive models can be constructed. We anticipate that by carefully selecting descriptors

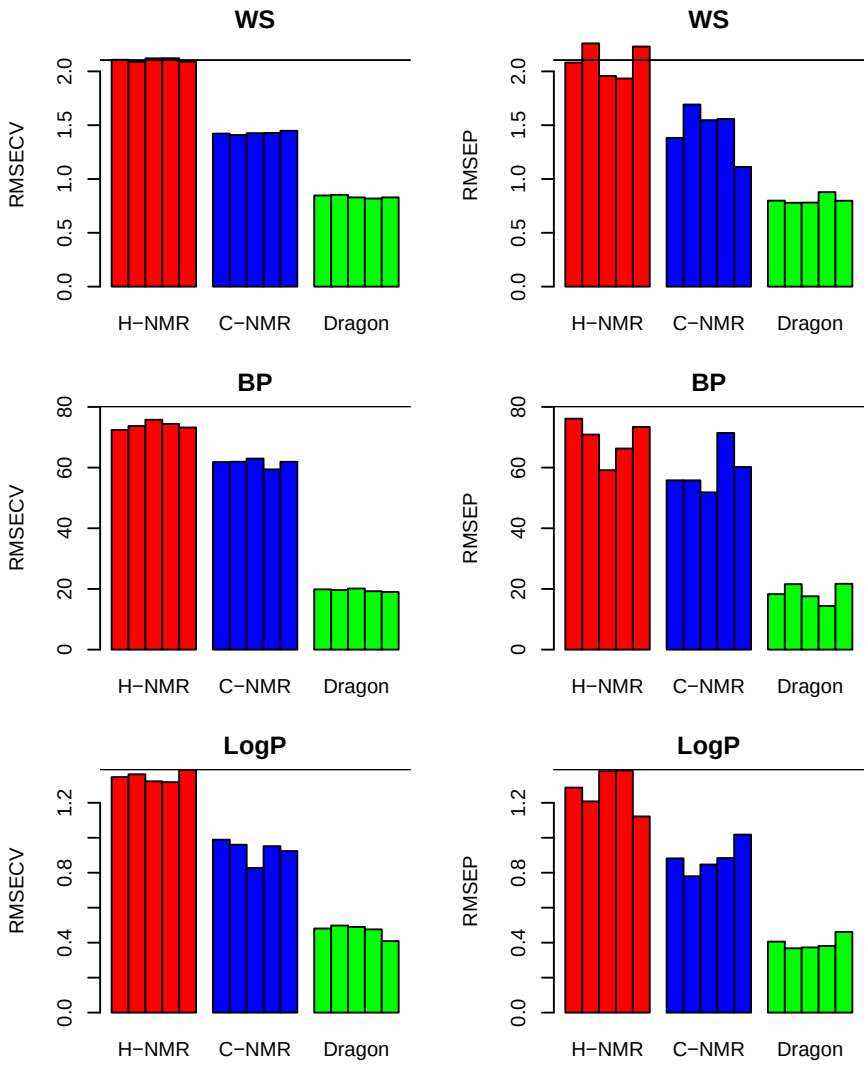


Figure 3.3: The cross-validation and test set performance statistics RMSECV and RMSEP for the three data sets, each with the same five test sets as in Figure 3.2. The horizontal line indicates the no-information limit defined by the $y_{pred} = \bar{y}$ model.

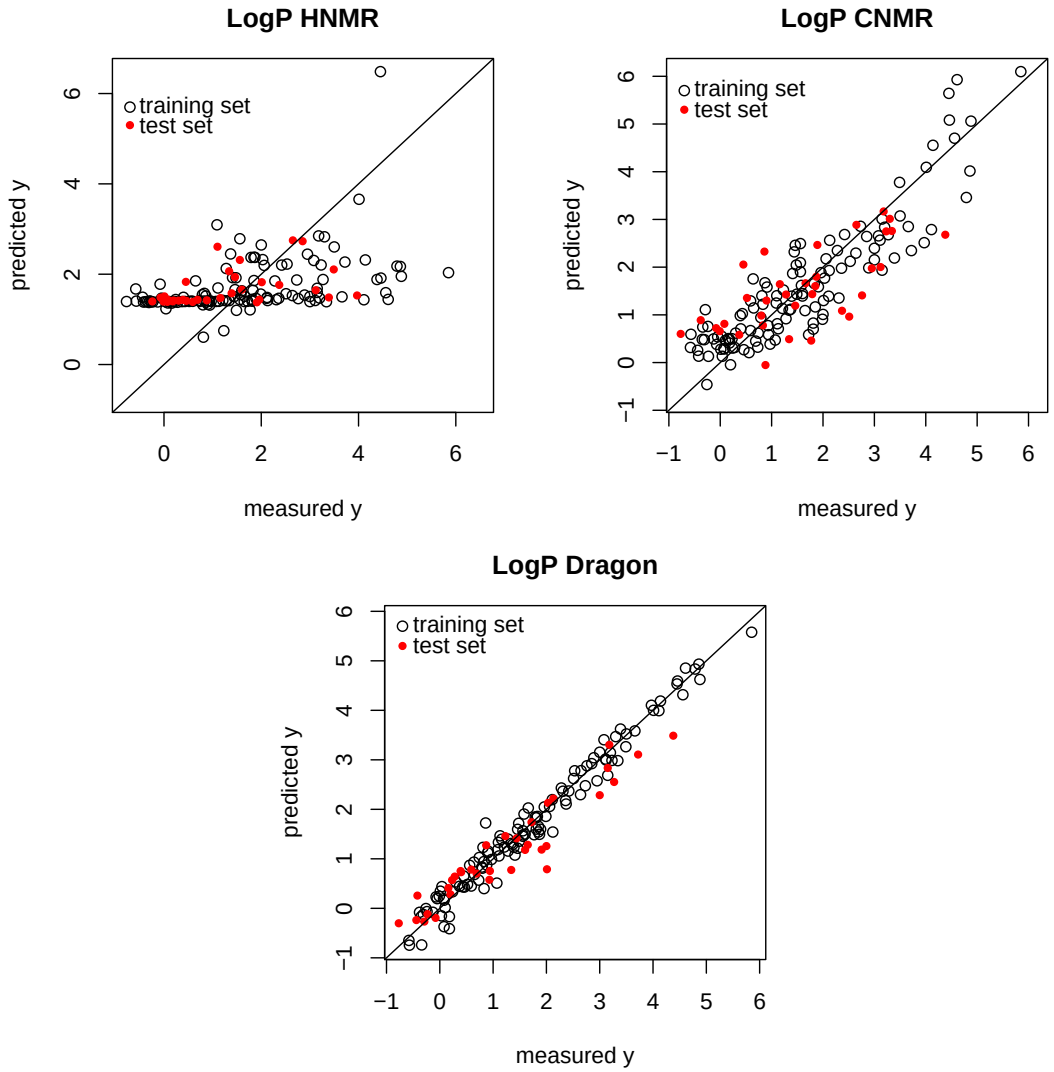


Figure 3.4: The $y_{measured} - y_{predicted}$ plots for the three descriptor types, proton and carbon NMR and Dragon, for the LogP data set. These plots show that Dragon-based models outperform the NMR-based models: the predicted activities are much closer to the expected values, indicated by the $x = y$ line. These figures are based on one random test set and are typical for other training/test set divisions.

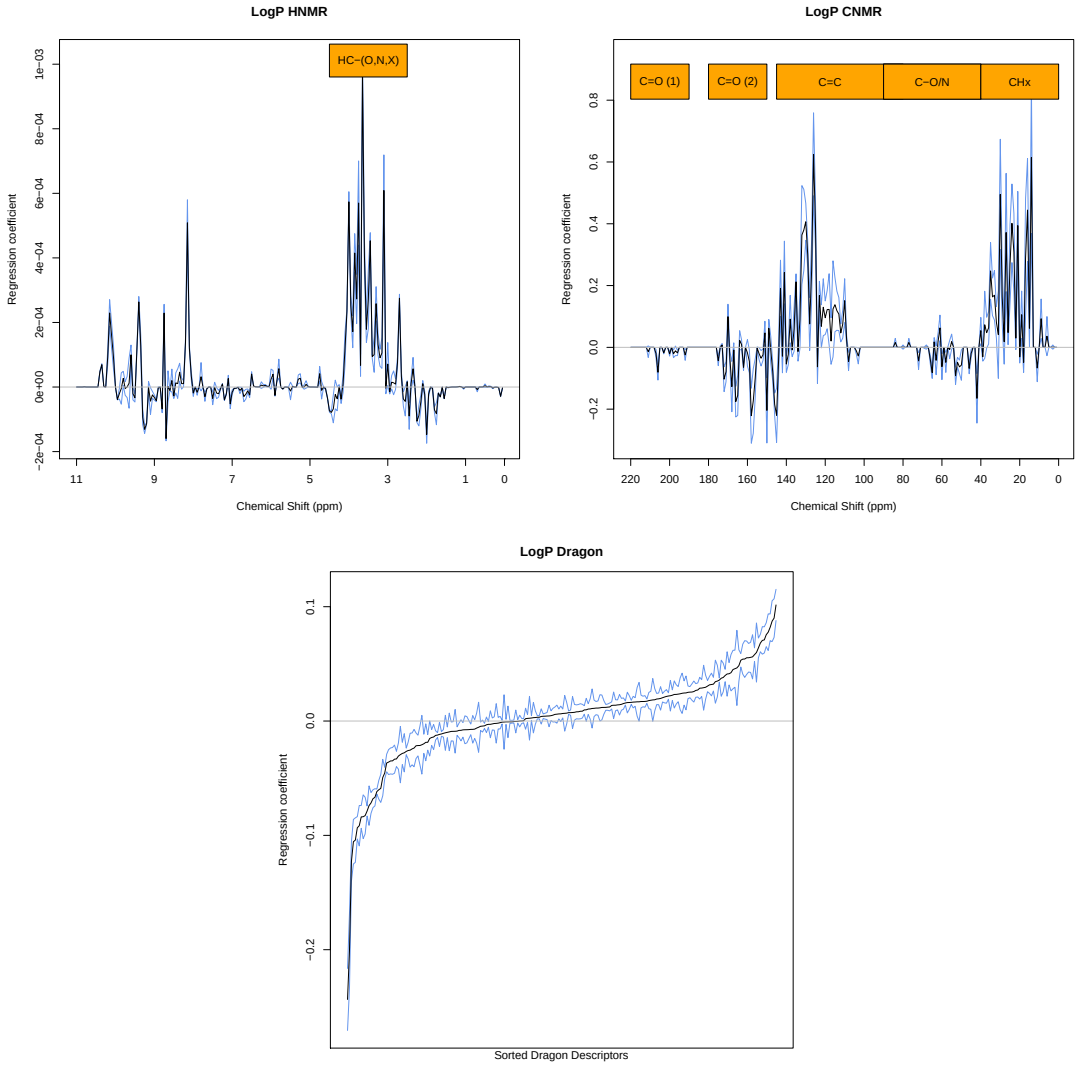


Figure 3.5: The mean PLS coefficients for the three types of spectra calculated from the five replicates, including ± standard deviation (blue lines). The coefficient vector for the Dragon descriptor set was sorted by size to show more clearly the significance of the most negative and most positive Dragon coefficients.

		^{13}C NMR	Dragon	Reference	
WS	R^2	0.61	0.88	*0.92* *	[25]
	RMSEP	1.46	0.81	0.59	
BP	R^2	0.55	0.97	0.99	[26]
	RMSEP	59.0	18.7	7.14	
LogP	R^2	0.81	0.95	0.88	[16]
	RMSEP	0.88	0.40	†	

Table 3.2: The R^2 values for the carbon NMR-based and the Dragon-based PLS models. The the right column are the published R^2 values as reference. *: The reference value for the WS data set is for a larger data set. †: No test set was used.

even better predictive models can be built.

3.4 Discussion

Important features of a QSAR or QSPR model are its predictive ability and the interpretability. The latter feature is an important tool to help scientists understand the influences of molecular features on the modeled activities. In such cases, the statistical fit is important, and one can focus on training set statistics [27]. An increasingly important application of QSAR and QSPR modeling, however, is virtual screening. For such applications, the predictive power of the model is more important, and just the statistical fit is not enough to characterize the model; an independent test set is then obligatory to estimate the models predictive power. We feel, however, that the use of an independent test set should in both cases be used. It ensures that observed influences of molecular features on the activities are true cause-effect relationships instead of just random correlation.

Spectral areas in NMR spectra are indicative for molecular features, but do not offer much information on the important molecular features. This makes the NMR based models not optimal for explanatory purposes. Moreover, the results indicate that the predictive power of models based on proton and carbon NMR spectra is not sufficient when compared to models based on theoretical molecular descriptors. For the WS, BP and LogP data sets, the R^2, Q^2 statistics and RMSE errors for the Dragon-based models were all favorable as compared to the NMR-based models. The results even indicate that proton NMR-based models do not improve on the null hypothesis model $y_{pred} = \bar{y}$. One possible reason for the inability of PLS to make spectra-based models, might be that PLS is a linear regression method unable to model non-linear problem well. Unpublished results using support vector machines, classification and regression trees, and wavelength selection, showed not to improve the predictive power of the models.

Comparing the means of the R^2 and RMSEP statistics for the five training/test set divisions with literature values (see Table 3.2), shows that spectra-based models are inferior to Dragon-based models and models published in literature. The fact that the statistics for

the Dragon-based models are comparable with statistics reported in literature, indicates that PLS in itself is a proper regression method for these data sets.

Although the use of full NMR spectra for proton and carbon nuclei does not give satisfactory results, NMR spectra in general might still be useful. For example, the combination of NMR spectra types has been suggested to improve models [4], though improvement is not apparent from literature and our own experiments. Moreover, data fusion of two spectra types is not trivial, and includes scaling issues. Additionally, NMR spectra of other nuclei, e.g. nitrogen and phosphorus, might be used, but these nuclei are more rarely found in organic compounds, and would restrict the applicability of the models, even if they decrease prediction errors. Other approaches are the combination NMR spectra with theoretical descriptors, where scaling issues occur again, and the use of spectra derived descriptors, such as the number of chemical shifts or the total sum of shift values. Finally, 2D and 3D methods might provide additional structural information that allows better modeling of the activities. Though interesting, such spectra types are, however, beyond the scope of the current QSAR/QSPR literature that uses NMR spectra, and will not be further discussed in this article.

3.5 Conclusions

The predictive powers of the PLS model for the three data sets indicate that proton NMR is not suitable for building QSPR models: the predictive power, as measured by the RMSECV and RMSEP, is never better than the $y_{pred} = \bar{y}$ model, as is clearly visible from the typical $y_{measured} - y_{predicted}$ plot of the LogP data set.

Carbon NMR-based models, however, do give acceptable QSPR models as was shown by the prediction errors. Moreover, the regression vectors correlate with areas of relevant molecular fragments, as was exemplified for the LogP data set. However, it was noted, that the regression vectors only indicate a few broad chemical shift ranges and do not indicate in detail which molecular features are interesting for modeling the activities.

Importantly, the predictive power of the carbon NMR-based spectra is less than basic Dragon-based models. We did not interpret Dragon descriptors which were found to be important for the models, but did notice that the training/test set division did not effect the importance of those descriptors. From the fact that Dragon performs better than spectra-based models, and that NMR-based models do not offer much information about important molecular features, we conclude that NMR spectra should not be considered first choice when making predictive models in general, and that proton NMR should probably not be used at all.

Bibliography

[1] D.J. Livingstone. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *Journal of Chemical Information and Computer Sciences*, 40:195–209, 2000.

[2] N. Christianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[3] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors*, volume 11 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, New York, 2000.

[4] R. Bursi, Y. Dao, T. Van Wijk, M. De Gooyer, E. Kellenbach, and P. Verwer. Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities. *Journal of Chemical Information and Computer Sciences*, 39:861–867, 1999.

[5] R.D. Cramer III, D.E. Patterson, and J.D. Bunce. Comparitative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carries proteins. *Journal of the American Chemical Society*, 110:5959–5967, 1988.

[6] R. Begigni, L. Passerini, D.J. Livingstone, M.A. Johnson, and A. Giuliani. Infrared Spectra Information and Their Correlation with QSAR Descriptors. *Journal of Chemical Information and Computer Sciences*, 39:558–562, 1999.

[7] R Begigni, A. Giuliani, and L. Passerini. Infrared Spectra as Chemical Descriptors for QSAR Models. *Journal of Chemical Information and Computer Sciences*, 41:727–730, 2001.

[8] R.D. Beger, J.P. Freeman, J.O. Lay Jr., J.G. Wilkes, and D.W. Miller. Use of ^{13}C NMR Spectrometric Data To Produce a Predictive Model of Estrogen Receptor Binding Activity. *Journal of Chemical Information and Computer Sciences*, 41:219–224, 2001.

[9] R.D. Beger and J.G. Wilkes. Developing ^{13}C NMR quantitative spectrometric data-activity relationship (QSADR) models of steroid binding to the corticosteroid binding globulin. *Journal of Computer-Aided Molecular Design*, 15:659–669, 2001.

[10] R.D. Beger and J.G. Wilkes. Models of Polychlorinated Dibenzodioxins, Dibenzofurans and Biphenyls Binding Affinity to the Aryl Hydrocarbon Recepter Developed Using ^{13}C NMR Data. *Journal of Chemical Information and Computer Sciences*, 41:1322–1329, 2001.

[11] R.D. Beger, D.A. Buzatu, J.G. Wilkes, and J.O. Lay Jr. Comparative Structural Connectivity Spectra Analysis (CoSCOSA) Models of Steroid Binding to the Corticosteroid Binding Globulin. *Journal of Chemical Information and Computer Sciences*, 42:1123–1131, 2002.

[12] R.D. Beger, D.A. Buzatu, and J.G. Wilkes. Combining NMR spectral and structural data to form models of polychlorinated dibenzodioxins, dibenzofurans and biphenyls binding to AhR. *Journal of Computer-Aided Molecular Design*, 16:727–740, 2002.

[13] A. Asikainen, J. Ruuskanen, and K. Tuppurainen. Spectroscopic QSAR Methods and Self-Organizing Molecular Field Analysis for Relating Molecular Structure and Estrogenic Activity. *Journal of Chemical Information and Computer Sciences*, 43:1974–1981, 2003.

[14] N.J.C. Bailey, Y. Wang, J. Sampson, W. Davis, I. Whitcombe, P.J. Hylands, S.L. Croft, and E. Holmes. Prediction of anti-plasmodial activity of *Artemisia annua* extracts: application of ^{1}H NMR spectroscopy and chemometrics. *Journal of Pharmaceutical and Biomedical Analysis*, 41:219–224, 2001.

[15] S.J. Vanderhoeven, J. Troke, G.E. Tranter, I.D. Wilson, J.K. Nicholson, and J.C. Lindon. Nuclear magnetic resonance (NMR) and quantitative structure-activity relationship (QSAR) studies on the transaceylation reactivity of model 1β-*O*-acyl glucuronides. II: QSAR modelling of the reaction using both computational and experimental NMR parameters. *Xenobiotica*, 34:889–900, 2004.

[16] Laura K Schnackenberg and Richard D Beger. Whole-molecule calculation of log p based on molar volume, hydrogen bonds, and simulated ^{13}C NMR spectra. *Journal of Chemical Information and Modeling*, 45(2):360–365, 2005.

[17] P.V. Khadikar, V. Sharma, and R.G. Varma. Novel estimation of lipophilicity using ^{13}C NMR chemical shifts as molecular descriptor. *Bioorganic & Medicinal Chemistry Letters*, 15:421–425, 2005.

[18] J. Wegner. JOELib. http://joelib.sourceforge.net/, 2005.

[19] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 42(2):493–500, 2003.

[20] P. Geladi and B. Kowalski. Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer-Verlag, Heidelberg, 2001.

[22] A. Golbraikh and A. Tropsha. Beware of q^2! *Journal of Molecular Graphics and Modelling*, 20(4):269–276, 2002.

[23] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[24] R. Wehrens. PLS and PCR functions: pls.pcr. http://cran.r-mirror.de/src/contrib/Descriptions/pls.pcr.html, 2005.

[25] A. Yan and J. Gasteiger. Prediction of Aqueous Solubility of Organic Compounds based on a 3D Structure Representation. *Journal of Chemical Information and Computer Sciences*, 43:429–434, 2003.

[26] E.S. Goll and P.C. Jurs. Prediction of the Normal Boiling Points ot Organic Compounds from Molecular Structures with a Computational Neural Network Model. *Journal of Chemical Information and Computer Sciences*, 39:974–983, 1999.

[27] M.T.D. Cronin and T.W. Schultz. Pitfalls in QSAR. *Journal of Molecular Structure*, 622:39–51, 2003.

Chapter 4

A Method for the Computational Comparison of Crystal Structures[1]

A new method to assess crystal structure similarity is described. A similarity measure is important in classification and clustering problems in which the crystal structures are the source of information. Classification is particularly important for the understanding of properties of crystals, while clustering can be used as a data reduction step in polymorph prediction. The method described uses a radial distribution function that combines atomic coordinates with partial atomic charges. The descriptor is validated using experimental data from a classification study of clathrate structures of cephalosporins, and data from a polymorph prediction run. In both cases, excellent results were obtained.

[1]E.L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L.M.C. Buydens, Acta.Cryst. B, 61:29-36, 2005

Introduction

Comparing crystal structures is important in both classification and clustering problems. Classification is important for the understanding of the relation between physical properties and the underlying structure of materials. The specific packing of molecules in a crystal directly influences the physical properties of compounds. As an example, in crystal engineering crystal packings are classified according to intermolecular interactions [1, 2, 3, 4, 5]. A second application of the similarity measure is in the clustering stage of ab initio crystal structure prediction [6, 7, 8]. In this process, hundreds or thousands of different hypothetical crystal packings for the same molecule, called polymorphs, are generated. They need to be clustered to get representative subsets for which analysis and geometry optimization is feasible.

For clustering and classification of crystal structures, two things are needed: a properly defined descriptor and a similarity function applied to this descriptor. In literature, several requirements for both the descriptor of crystal structures and the similarity function have been described [9, 10, 11]. The most obvious requirement for a descriptor-similarity combination is that more dissimilar crystal structures result in larger dissimilarity values. Although this seems trivial, several well-known descriptors do not generally satisfy this requirement [9, 10, 12, 11]. Many descriptors require a choice of origin, or some other setting. Among such descriptors is the combination of unit cell parameters and fractional coordinates. A descriptor based on reduced unit cell parameters can vary significantly with only minor lattice distortions [13, 14]. While it is in some cases possible to adapt the similarity function to deal with such instabilities, we believe that this issue should be addressed by using a proper descriptor.

Recently, powder diffraction patterns have been used to compare crystal structures of both simulated and experimental structures [15, 3]. This descriptor does not suffer from the problems mentioned above, and has an interpretable physical meaning. A potential disadvantage is that it is not always unique under certain conditions [16].

The current article investigates a new direct space descriptor for comparing crystal structures. It is based on a radial distribution function and includes the electronic properties of the atoms. Section two will introduce the descriptor in detail and will introduce the dissimilarity measure used to express the dissimilarities between structures using this descriptor.

Validation of the descriptor and the dissimilarity measure is done in two ways; first, by comparing calculated dissimilarity values with empirical values, and, secondly, by comparing a clustering created from the calculated dissimilarities with an empirical clustering. Empirical dissimilarity values, however, are normally not known on a continuous scale, but are expressed on a binary scale (identical or not) or are described textually using visual inspection. To our knowledge, there is no data set available from literature in which the dissimilarities between a set of crystal structures are known on a continuous scale, which is needed for a quantitative validation of the descriptor and its dissimilarity measure. Section three describes the two data sets for which empirical dissimilarity values and the

clustering or classification are obtained. These are used to validate the application of the descriptor and dissimilarity measure.

Section four describes experimental details and the fifth section discusses the calculated dissimilarity values and clusterings for the two data sets.

4.1 The Descriptor

To be able to compare crystal structures a descriptor is needed that represents the structure in mathematical form, and a dissimilarity measure that expresses the differences between two crystal structures using the descriptor. The resulting dissimilarity values can then be used to cluster or classify the crystal structures by grouping together structures which have a low dissimilarity between them.

Crystal structures can be uniquely represented by a radial distribution function (RDF) describing the distribution of neighboring atoms around a central atom. Each neighboring atom gives rise to a peak in the function. RDFs are independent of cell choice, and can be physically interpreted. RDFs have been used to describe molecules with the goal to simulate infra red spectra [17, 18], and have been used in the form of a radial distribution matrix for crystals [16]. In the latter application, each row in the distribution matrix is a RDF describing the interatomic distances for one atom type pair. As such, the descriptor does not differentiate between, e.g., hydroxyl and carbonyl oxygens.

In our approach, the RDF is adapted to include more specific information about the atoms. To do so, the RDF is weighted by the electrostatic interactions. To indicate the inclusion of electrostatic information in the descriptor, we will refer to this as the electronic radial distribution function, or R_eDF. The reason for including electrostatics is the assumption that these play a major role in crystal packing [19, 2, 20]. By including partial atomic charges, the R_eDF focuses on atom groups with large partial charges, in particular functional groups, and differentiates between attractive interactions, between oppositely charged atoms, and repulsive interactions.

An atomic R_eDF describes the distribution of coulombic interactions of one atom with surrounding atoms; the R_eDF for the crystal structure is obtained by summing all atomic R_eDFs of all N atoms in the asymmetric unit:

$$R_eDF(r) = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{q_i q_j}{N \cdot r_{i,j}} \delta(r - r_{i,j}) \tag{4.1}$$

where M is the number of neighboring atoms within a radius r, q_i and q_j are partial atomic charges of the atoms i and j, and δ places the electrostatic interaction at the right distance by its definition $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. The function is scaled for the number of atoms in the asymmetric unit, N.

The R_eDF in Eq. 4.1 is a continuous function and is implemented as a discrete

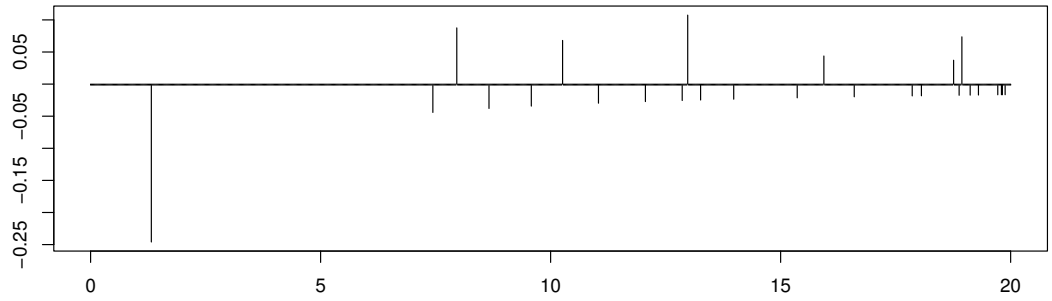


Figure 4.1: Example R_eDF of an artificial crystal structure with a positively and a negatively charged atom ($a = 7.97$, $b = 10.26$, $c = 18.77$, $\alpha = \beta = \gamma = 90^o$).

function with S intervals of size b, hereafter called bins:

$$R_eDF(s) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\frac{q_i q_j}{N \cdot r_{i,j}} D((s+\frac{1}{2})b - r_{i,j})\right) \tag{4.2}$$

where s is the bin index and $s = 0..S$, $r_{i,j}$ is the distance between the two atoms i, j, q_i and q_j are partial atomic charges and D is

$$D(x) = \begin{cases} 1 & \text{if } |x| < \frac{1}{2}b \\ 0 & \text{if } |x| \geq \frac{1}{2}b \end{cases} \tag{4.3}$$

Figure 4.1 shows the R_eDF for an artificial crystal with two atoms in the unit cell, a positively and a negatively charged one ($a = 7.97$, $b = 10.26$, $c = 18.77$, $\alpha = \beta = \gamma = 90^o$). The first, negative peak is the interaction between the two atoms at exactly the bonding distance. The other negative peaks are also peaks between two oppositely charged atoms. The overall decrease in intensities is caused by the $\frac{1}{r}$ term in the R_eDF equation. The first positive peak is related to the translation along the a axis, i.e. $\pm\vec{a}$, and the second peak to the translation along the b axis. The third peak is the translation in the direction $a \pm b$; for this orthogonal structure there are twice as many contributions to this peak as for the first two positive peaks, resulting in the higher intensity.

The R_eDFs of four experimental crystal structures, described in a later section, are given in Figures 4.2 and 4.3. They show a few distinct high intensity peaks and many smaller peaks. The locations of these peaks are specific for the crystal packing: Figure 4.2(a) and (b) show the R_eDFs of two cephalosporin structures from the same class, while (c) shows the R_eDF for a different packing.

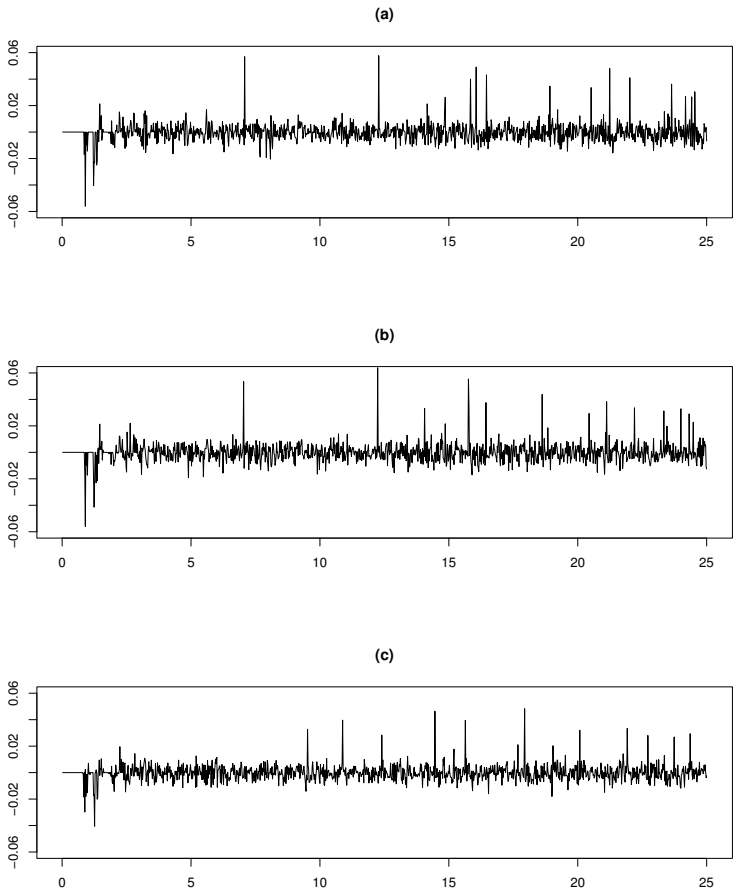


Figure 4.2: Example R_eDFs of cephalosporins a) A9, b) A10 from the same class A, and c) N19 from a different class N.

Figure 4.3(a) shows the function for a simulated estrone crystal structure; a similar pattern can be observed. Figure 4.3(b) shows the effect of cutting away peaks with intensities lower than some threshold. It was found that the cut off value must be around 20 percent of the highest peak. Cutting away the smaller peaks emphasizes the major features in the R_eDF and leads to better discrimination.

Because of the nature of the R_eDF, one can expect positive contributions at those distances which match the translational symmetry in the crystal. However, since such contributions can be canceled out by other, negative contributions, they do not always show up in the R_eDF. Moreover, peaks not related to translational symmetry are especially interesting, because they provide information additional to periodicity.

Figure 4.4 shows the R_eDF for cephalosporin structure A1 (top) and the locations

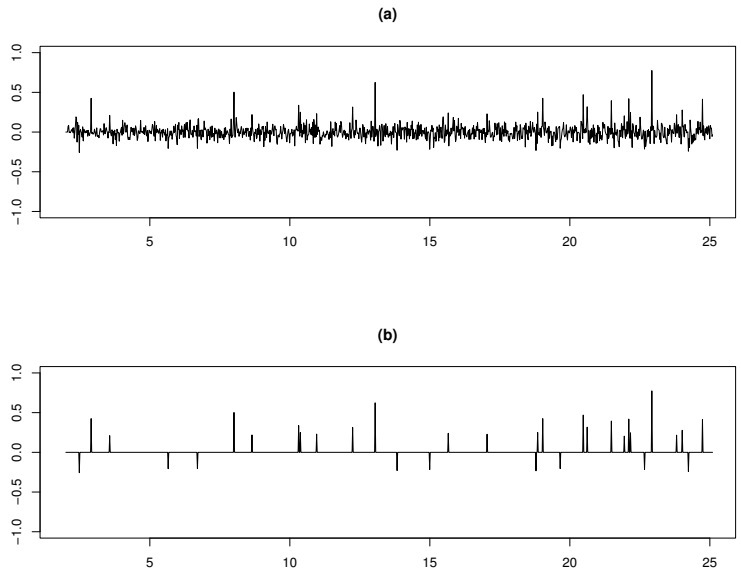


Figure 4.3: Example R_eDF of one of the simulated estrone structures shown in a), and the effect of cutting away of peaks below 20 percent of the intensity of the highest peak in b).

of peaks caused by the translational symmetry. Clearly, a significant number of peaks are not caused by translational symmetry and contain additional structural information. Each peak consists of many contributing atom pairs resulting in a netto positive (repulsive) or negative (attractive) peak in the function.

Dissimilarities between crystal structures are represented by the difference between the two corresponding R_eDFs. For this, a weighted cross correlation (WCC) is used [3] which is applied to the high intensity peaks of the R_eDF.

4.2 Data

Two data sets are used in this article to show the application of the descriptor. The first data set contains experimental crystal structures of inclusion complexes of cephalosporins. These twenty structures are classified into seven classes, but there is no knowledge about the similarity between structures other than belonging or not belonging to the same class. To our knowledge, there is no data set available from literature in which the dissimilarities between all crystal structures are known on a continuous scale, which would be ideal for validation of the proposed descriptor and its dissimilarity measure. The second data set contains simulated polymorphs of estrone, for which detailed information is available

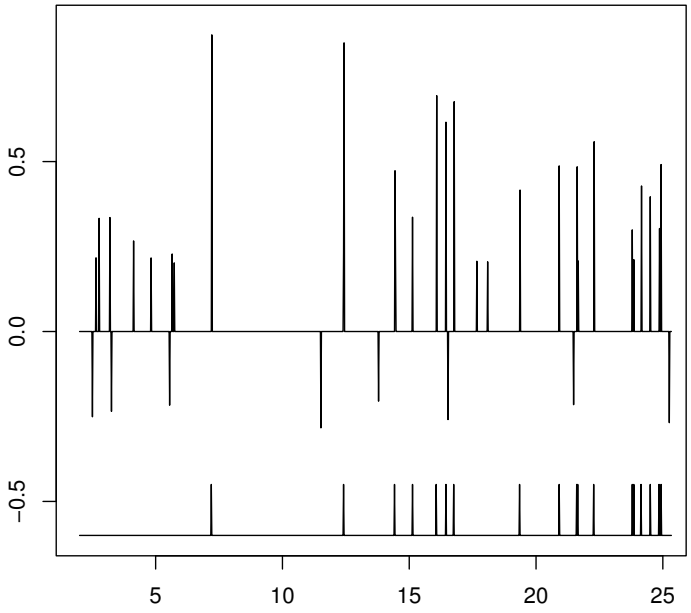


Figure 4.4: This figure shows that the nature of the peaks in the R_eDFs is not only describing the translation symmetry of the crystal structure: the top function is the R_eDF of cephalosporin A1 after applying the peak selection. The bottom black line shows the locations originating from translation symmetry.

about the dissimilarities between the structures, as is explained below. The 48 structures in this data set are classified into 25 classes, based on visual inspection as described below.

4.2.1 Cephalosporin data set

The cephalosporin data set consists of twenty clathrate structures of cephalosporins [21, 3]. The twenty compounds were classified into seven isomorphic classes based on their crystal form: A,B,C,D,E,F and N. Class A has ten structures, all in the C2 space group. Class B has four structures in the $P2_12_12_1$ space group. Classes C, D, E and F all have one structure, and have space groups $P2_1$, C2, P1, and $P2_1$ respectively. Class N has two structures both have the $P2_1$ space group. A brief overview of the unit cell parameters of this data set is given in Table 4.1. Further details on these structures can be found in Ref. [21, 3].

For a set of twenty crystal structures, there are 190 unique pairs of structures ($\frac{1}{2} * n *$

Table 4.1: Unit cell parameters of the cephalosporin data set, grouped into seven clusters (A, B, C, D, E, F, and N).

cluster	a	b	c	α	β	γ
A	23.47	7.12	14.93	90.0	108.27	90.00
	23.42	6.97	15.00	90.0	110.41	90.00
	23.46	7.12	14.89	90.0	108.57	90.00
	23.41	7.11	14.81	90.0	108.15	90.00
	23.39	7.20	14.76	90.0	108.58	90.00
	23.02	7.15	14.55	90.0	104.64	90.00
	23.40	7.06	14.92	90.0	109.80	90.00
	23.43	7.11	14.88	90.0	108.19	90.00
	23.49	7.08	14.85	90.0	108.95	90.00
	23.45	7.03	14.84	90.0	110.55	90.00
B	7.11	21.72	30.96	90.0	90.00	90.00
	7.00	20.99	30.69	90.0	90.00	90.00
	7.11	21.86	32.31	90.0	90.00	90.00
	7.09	21.27	31.00	90.0	90.00	90.00
C	14.92	7.38	20.50	90.0	105.77	90.00
D	23.56	7.13	18.69	90.0	109.38	90.00
E	7.07	10.70	14.23	87.15	79.00	89.74
F	15.40	7.30	23.57	90.00	99.35	90.00
N	10.87	9.51	12.39	90.00	98.70	90.00
	10.91	9.41	12.20	90.00	98.53	90.00

$(n-1) = \frac{1}{2} * 20 * 19$). The dissimilarity associated with each pair is unknown. However, it is known whether the pair is a within-cluster or a between-cluster pair, i.e. the dissimilarity of a pair of structures from the same class is marked as within-cluster, and for a pair of structures that do not belong to the same structure class it is marked as between-cluster.

4.2.2 Estrone data set

The second data set consists of 48 simulated crystal structures of the estrone steroid, which has three known naturally occurring polymorphs (CSD refcodes ESTRON10, ESTRON11, ESTRON12) [22]. Two thousand polymorphic structures were generated using the Polymorph Predictor module in Cerius^2 [6, 23]. The method used by this program consists of a generation step where random crystal structures are generated. After removal of duplicates, the remaining 1278 structures were minimized in energy using a force field. For this data set, the estrone molecule was kept rigid and the $P2_12_12_1$ space group symmetry was imposed during the initial generation. The energy minimization was done with the DREIDING-2.21 force field [24] using Ewald summation to calculate the van der Waals and

Coulomb interactions. Electrostatic potential (ESP) derived atomic charges for estrone were calculated using Gaussian94 [25] with the HF/6-31G* basis set.

From the 1278 structures, a set of 48 structures were selected in the low energy region which represent crystal structures that might be found in nature. The densities of these simulated structures are in the range of $[1.043, 1.173]$ g/cm^3, while the experimental structures have densities around 1.2 g/cm^3. It is common for predicted crystal structures to have deviating densities, due to the force field used. The energies are in a range of 5.03 kcal/mol.

To classify the crystal structures, the 1128 pairwise comparisons between the 48 estrone structures ($\frac{1}{2} * 48 * 47$) were manually grouped into three dissimilarity classes by visual inspection. The classification of the pairwise dissimilarities was done by trying to overlap the crystal structures. However, an attempt has been made to qualify the differences in terms of packing parameters. These properties were taken into account during the clustering: cell parameters, placement in the cell, and orientation in the cell (see Table 4.3). The cell parameters were compared and show big differences (for -), small differences (for +), or hardly any differences (for ++). The placement in the cell is compared visually: ++ means that the four molecules in the unit cell can be placed on top of each other perfectly within 0.01 Å, + means they fit well, and - means that they cannot be aligned simultaneously at all. Similarly, for the rotations around the various axes, ++ means that the molecules in the two structures have an identical orientation, + means a rotation up to about 10°. Larger rotations do not occur in the data sets, as the actual molecular packing becomes different then. The number of dissimilarity classes is chosen to reflect the number of visually distinguishable dissimilarity types in the above analysis.

The first dissimilarity class is called *identical*, as the structures are *visually* identical. The second class is called *similar* and consists of pairs of crystal structures that show small displacements or small rotations of the molecules in the unit cell, but the location of the molecules in the cell and the cell parameters itself are similar. The third class is called *dissimilar* and consists of all dissimilarities not classified in the other two classes. No further distinction between dissimilarities can be made in this class. Note that the first two classes have far less structure pairs than the *dissimilar* class, which reflects the diversity of the data set.

Based on the visually determined dissimilarities, identical and similar crystal structures were grouped, leading to 25 true classes, labeled A to Y. Table 4.2 shows the members of each class. The diversity of unit cell axes between the structures is apparent from this table. The similarity within classes is mostly clear, for example in class A.

An additional analysis has been done to qualify the similarity of structures within classes: for all structures, the hydrogen bonding pattern was determined as described by two variables. Because estrone has only one hydrogen bond donor, and only one acceptor, the bonding pattern can only exist in the form of chains. Thus, the axis along which the chain is directed is given, as well as the form of the chain: linear, or zigzagged. In all cases the structure pairs with *identical* and *similar*, similarity values show an identical

scheme of hydrogen bond chains. The hydrogen bonding patterns are given in Table 4.2 and support the clustering found by visual analysis of the structures.

Table 4.2: An overview of the estrone dataset showing the lengths a, b, c of the orthogonal unit cell axes of the 48 structures, and the direction (a, b or c direction) and form of the hydrogen bond chain (linear of zigzagged).

cluster	a	b	c	direction of H-bond chain	form of chain
A	7.063	11.530	19.481	c	zigzagged
	7.971	10.262	18.772	c	zigzagged
	8.427	10.958	17.286	c	zigzagged
B	7.742	9.110	23.163	c	linear
	7.658	9.188	22.419	c	linear
	7.691	8.865	23.262	c	linear
	7.706	8.910	24.038	c	linear
C	6.457	12.421	19.679	c	zigzagged
	6.678	13.305	18.966	c	zigzagged
D	5.946	12.940	20.499	c	zigzagged
	6.332	13.037	19.066	c	zigzagged
E	8.687	10.067	18.082	b	linear
	9.381	9.432	18.147	b	linear
F	8.742	13.276	13.617	b	linear
	9.649	12.309	13.279	c	linear
G	7.456	14.441	15.324	b	zigzagged
	8.281	13.521	15.177	c	zigzagged
	9.025	11.533	15.931	c	zigzagged
H	8.507	10.087	18.943	b	linear
	9.331	9.410	17.887	b	linear
I	6.903	9.589	23.719	c	zigzagged
J	7.980	10.539	18.293	b	linear
K	9.868	12.127	13.063	c	linear
L	7.969	10.597	18.254	b	linear
	7.969	10.597	18.255	b	linear
M	7.968	13.259	14.687	b	linear
	7.968	13.259	14.688	b	linear
N	7.581	10.387	19.439	b	linear
	7.581	10.387	19.439	b	linear
O	9.306	9.445	18.111	b	linear
	9.306	9.445	18.111	b	linear
P	7.733	9.526	21.196	b	zigzagged
	7.733	9.526	21.196	b	zigzagged
Q	7.500	12.300	17.088	c	linear

	7.500	12.300	17.088	c	linear
R	8.560	13.268	14.186	c	linear
	8.560	13.268	14.186	c	linear
S	7.829	13.975	15.743	b	zigzagged
	7.829	13.975	15.743	b	zigzagged
T	7.135	10.876	20.431	c	zigzagged
	7.442	10.043	22.177	c	zigzagged
U	9.183	13.104	13.198	c	linear
	9.750	12.673	13.044	c	linear
V	7.235	11.743	19.066	c	zigzagged
	7.293	10.763	20.544	c	zigzagged
W	7.772	9.123	23.078	c	zigzagged
X	7.302	13.266	16.788	b	linear
Y	9.228	13.127	13.254	b	linear

4.3 Experimental

For both data sets the R_eDF was used with a bin size of 0.02 Å and in a domain of [2,25] Å. The bin size was chosen such that high intensity peaks showed clearly. Below 2 Å there is mostly intramolecular information, which does not describe crystal packing and is therefore not included in the chosen domain. The distance up to which the R_eDF is calculated, 25 Å, is found to be the smallest distance containing enough informative peaks, and is used for both data sets. When calculating the dissimilarities between the R_eDFs with the WCC measure, a triangle is used of 0.6 Å, which is about half a bond length. Much larger and much smaller values showed worse clustering results.

The descriptor is validated for both data sets, by grouping all dissimilarities calculated with the descriptor into the dissimilarity classes, as defined earlier. The median, minimal and maximal dissimilarity values for the classes can be compared and ideally show distinct classes. The larger the overlap between two dissimilarity classes, the worse the descriptor. The better the trend in the calculated dissimilarity values, the better the descriptor.

In addition to this, the calculated dissimilarities are used to cluster the crystal structures into a dendrogram using hierarchical average linkage clustering. The dendrogram can be cut at a height yielding a certain number of clusters. Cutting at a small height will give many clusters, while cutting at a large height will give only a few clusters. The height at which the dendrogram is cut is chosen to give that number of clusters that matches the number of classes defined for that data set.

Finally, the simulated estrone structures are matched against the experimentally determined ESTRON10 structure to find the structure with the same packing. This is done by calculating the R_eDF for the experimental and simulated structures and calculating

the dissimilarity between ESTRON10 and all of the simulated structures. The structure with the smallest dissimilarity to ESTRON10 is identified to have the same packing.

The simulated structures are not matched against the ESTRON11 polymorph which also has $P2_12_12_1$ symmetry, because the hydroxyl group in ESTRON11 points in a different directions than in the simulated structure, leading to a different packing. Neither was it matched against ESTRON12 which has a different space group symmetry. Both experimental structures do not have a corresponding structure in the simulated data set.

Calculation of R_eDF descriptions for crystal structures and dissimilarity measures is implemented in C++. Clustering of structures based on the dissimilarities matrix is done in R [26] with the average linkage method. Calculations were performed on both Solaris and GNU/Linux systems.

4.4 Results

4.4.1 Dissimilarity Classes

The descriptor is validated by using it to calculate the dissimilarity values between all pairs of crystal structures. The dissimilarity values calculated for the cephalosporin data set are shown as box plots in Figure 4.5, where the within-cluster and between-cluster groupings are based on the known classification. As desired, the two medians show a rise going from the *within-cluster* class to the *between-cluster* class. There is, however, a slight overlap between the two dissimilarity classes. The calculated dissimilarities on the basis of powder diffraction patterns [3] are shown in Figure 4.6 and show the same increase for the median and overlap, though the separation of the classes is better with the R_eDF descriptor.

The results for the estrone data set are plotted as box plots in Figure 4.7. The calculated dissimilarities are an order of magnitude larger than those for the cephalosporin set. This is caused by the higher intensities of the peaks in the estrone R_eDFs. The medians in the plot show a gradual rise going from the *identical* class to the *dissimilar* class. This is what one would expect, but the figure shows that the two most dissimilar classes are not fully separated. The *identical* class is completely separated from the other two dissimilarity classes.

4.4.2 Dendrograms and Partitionings

The dendrogram determined for the cephalosporin data set with the new descriptor using average linkage is given in Figure 4.8. Given a properly chosen height, it predicts the true classes without errors. Partitioning the dendrogram into seven clusters was done by cutting the tree at a height of 0.4 (horizontal line).

The use of the R_eDF descriptor for the experimental data set was compared with the

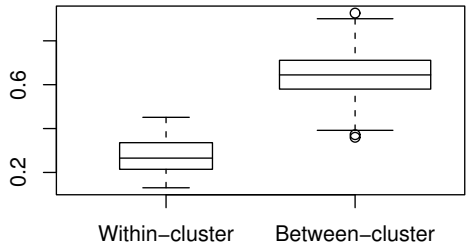


Figure 4.5: Box plot for dissimilarities between the two defined dissimilarity classes (within-cluster and between-cluster) calculated for the cephalosporin structures with the R_eDF.

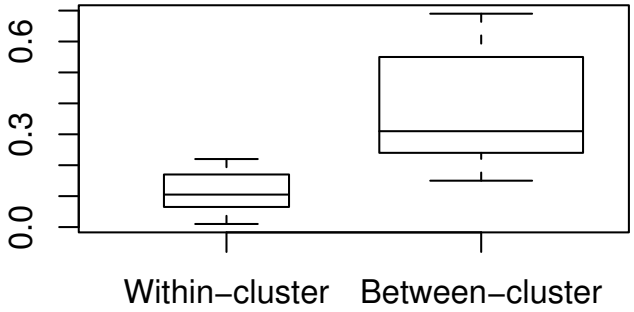


Figure 4.6: Box plot for dissimilarities between the two defined dissimilarity classes (within-cluster and between-cluster) calculated for the cephalosporin structures using powder diffraction data (see Ref. [3]).

dendrogram determined on the basis of powder diffraction patterns (see Figure 5d in [3]). The latter shows a clustering which is essentially correct, but the dendrogram based on the R_eDF gives a better discrimination of the separate groups.

The dendrogram for the 48 crystal structures of estrone was calculated with av-

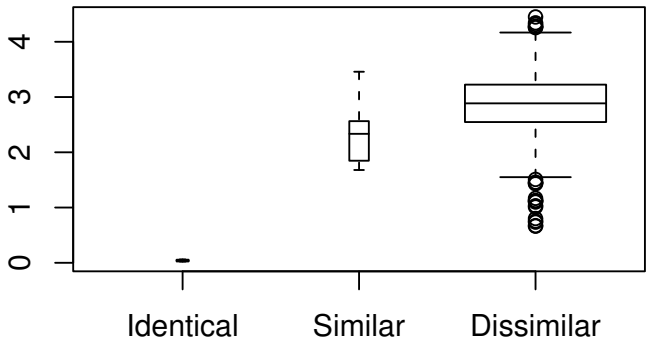


Figure 4.7: Box plot for dissimilarities between the estrone crystal structures grouped by the three dissimilarity classes as defined in Table 4.3 calculated with the R_eDF. The widths of the boxes are proportional to the number of objects in that class. The circles in this plot indicate dissimilarities that fall outside the fourth quantile of the distribution.

erage linkage from the R_eDF-generated dissimilarities and is given in Figure 4.9. The dendrogram shows that the crystal structures which are known to have a dissimilarity in the *identical* class (clusters L-S), are correctly grouped together. The structures from cluster B, with dissimilarities in the *similar* class are grouped together, but cluster A, also with dissimilarities in the *similar* class, is scattered over the right hand side of the dendrogram. This reflects the fact that the dissimilarities for the two dissimilarity classes have an overlap (see Figure 4.7). A partitioning with 25 clusters is generated from the dendrogram by cutting at a height of 0.45 (horizontal line).

4.4.3 Matching ESTRON10

In case of the simulated estrone structure, it is interesting to know if the method is able to tell which simulated structure matches an experimental structure. This has been done for ESTRON10, and the results are given in Figure 4.10. The R_eDF for ESTRON10 is calculated in the same way as done for the simulated structures, and the dissimilarity measure is able to identify structures 6 and 1 having the same packing. Structures 6 and 1 both belong to cluster A with a dissimilarity between them in the *similar* class.

The large dissimilarity between the simulated structures and ESTRON10 is due to the fact that the set of simulated structures is the result of a molecular mechanics

Table 4.3: Visual criteria used to classify the 1128 dissimilarities between the 48 crystal crystals. The qualifier ++ means *almost identical*, + *similar*, and - means *dissimilar*. See text for a more specified definition.

Dissimilarity Class	Number of Pairs	Unit Cell Parameters	Placement in Cell	Orientation in Cell
Identical	8	++	++	++
Similar	21	+	+	+
Dissimilar	1099	-	-	-

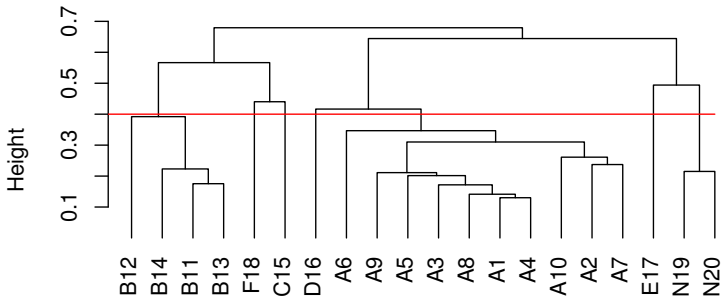


Figure 4.8: Dendrogram for the cephalosporin data set calculated with the optimized descriptor for the twenty structures with average linkage. The seven structure classes that are compared with the known classes (A,B,C,D,E,F,N) were determined by cutting the dendrogram at a height of 0.4.

optimization. Force field artifacts lead to longer unit cell axes than experimentally found; therefore, the y-scales of Figure 4.10 and Figure 4.7 are not directly comparable. The important thing here is that the order of dissimilarities is correct.

which explains why the dissimilarities found are larger than those in the dissimilar class in Table 4.7. It also makes comparing the dissimilarities of ESTRON10 versus 6 and 1 with the dissimilarity of ESTRON10 versus the third most ESTRON10 like compound less intuitively; the small differences in those three values do not necessarily indicate that the third structure has almost the same packing as ESTRON10 as structure 6 does.

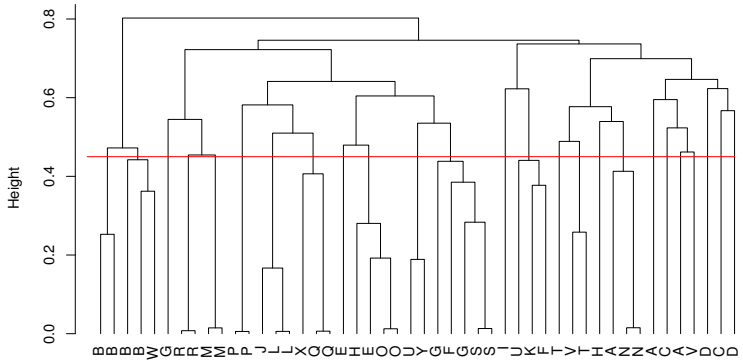


Figure 4.9: Dendrogram of the 48 structures in the estrone data set clustered with average linkage using the dissimilarity values calculated with the optimized descriptor. The 25 clusters that are compared with the validation set were determined by cutting the dendrogram at a height of 0.44 (horizontal line). Object labels are taken from Table 4.2.

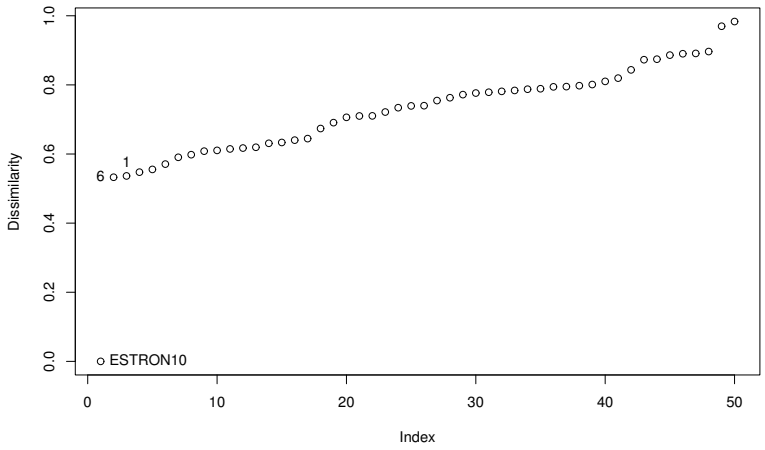


Figure 4.10: Dissimilarities between the experimentally found polymorph (ESTRON10) and all simulated estrone structures (1-48). Structures 6 and 1 have the same packing as the ESTRON10, and are identified with the new descriptor.

4.5 Conclusions

This article presents a new computational method to compare crystal structures. It is conceptually easy and contains only a few parameters to tune; within broad ranges, the exact values of these hardly influence the results. The method is, therefore, very general. It correctly shows increasing dissimilarity values when going from identical crystal structures to similar, and finally to dissimilar structures. It is difficult to order dissimilar structures in a meaningful way, and therefore, the main use of the descriptor is twofold: to gather similar structures from a large set, and to recognize the most similar structure from a set of candidate structures. Both have numerous and important applications.

Bibliography

[1] J. Perlstein, K. Steppe, S. Vaday, and E. M. N. Ndip. Molecular self-assemblies. 5. analysis of the vector properties of hydrogen bonding in crystal engeneering. *Journal of the American Chemical Society*, 118:8433–8443, 1996.

[2] B. Moulton and M. J. Zaworotko. From molecules to crystal engineering: Supramolecular isomerism and polymorphism in network solids. *Chemical Reviews*, 101:1629–1658, 2001.

[3] R. De Gelder, R. Wehrens, and J.A. Hageman. A generalized expression for the similarity spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22(3):273–289, 2001.

[4] M. D. Hollingsworth. Crystal engineering: from structure to function. *Science*, 295:2410–2413, 2002.

[5] G. Ilyushin, N. Blatov, and Y. Zakutin. Crystal chemistry of orthosilicates and their analogs: the classification by topological types of suprapolyhedral structural units. *Acta Crystallographica*, B58:948–964, 2002.

[6] P. Verwer and F. J. J. Leusen. *Computer Simulation to Predict Possible Crystal Polymorphs*, volume 12 of *Reviews in Computational Chemistry*, chapter 7. Wiley-VCH, New York, 1998.

[7] J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. Van Eijck, P. Verwer, and D. E. Williams. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica*, B56:697–714, 2000.

[8] W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scherega, B. Schweizer, M. U. Schmidt, B. P. Van Eijck, P. Verwer, and D. E. Williams. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica*, B58:647–661, 2002.

[9] A. V. Dzyabchenko. Method of crystal-structure similarity searching. *Acta Crystallographica*, B50:414–425, 1994.

[10] L. C. Andrews and H. J. Bernstein. Bravais lattice invariants. *Acta Crystallographica*, A51:413–416, 1995.

[11] A. Kálmán and L. Fábián. Volumetric measure of isostructurality. *Acta Crystallographica*, B55:1099–1108, 1999.

[12] B. P. Van Eijk and J. Kroon. Fast clustering of equivalent structures in crystal structure prediction. *Journal of Computational Chemistry*, 18:1036–1042, 1997.

[13] L. C. Andrews, H. J. Bernstein, and G. A. Pelletier. A perturbation stable cell comparison technique. *Acta Cryst.*, A36:248–252, 1980.

[14] L. C. Andrews and H. J. Bernstein. Lattices and reduced cells as points in 6-space and selection of bravais lattice type by projections. *Acta Crystallographica*, A51:1009, 1988.

[15] H. R. Karfunkel, B. Rohde, F. J. J. Leusen, R. J. Gdanitz, and G. Rihs. Continous similarity measure between nonoverlapping x-ray powder diagrams of different crystal modifications. *Journal of Computational Chemistry*, 14:1125–1135, 1993.

[16] H. Karfunkel, H. Wilts, Z. Hao, A. Iqbal, J. Mizuguchi, and Z. Wu. Local similarity in organic crystals and the non-uniqueness of x-ray powder patterns. *Acta Crystallographica*, B55:1075–1089, 1999.

[17] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, and V. Steinhauer. Chemical information in 3d space. *Journal of Chemical Information and Computer Sciences*, 36:1030–1037, 1996.

[18] M. C. Hemmer, V. Steinhauer, and J. Gasteiger. Deriving the 3d structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy*, 19:151–164, 1999.

[19] L. Pauling and M. Delbrück. The nature of the intermolecular forces operative in biological processes. *Science*, 92:77–79, 1940.

[20] G. R. Desiraju. Supramolecular synthons in crystal engineering - a new organic synthesis. *Angewandte Chemie International Edition in English*, 34:2311–2327, 1995.

[21] G. J. Kemperman, R. De Gelder, F. J. Dommerholt, P. C. Raemakers-Franken, A. J. H. Klunder, and B. Zwanenburg. Induced fit phenomena in clathrate structures in cephalosporins. *Journal of the Chemical Society, Perkin Transactions 2*, 7:1425–1429, 2000.

[22] B. Busetta, C. Courseille, and M. Hospital. Structures cristallines et moléculaires de trois formes polymorphes de l'oestrone. *Acta Crystallographica*, B29:298–313, 1973.

[23] Molecular Simulations Inc. *Cerius*^{2} *User Guide*, chapter 7. Molecular Simulations Inc., San Diego, 1997.

[24] S. L. Mayo, B. D. Olafson, and W. A. Goddard III. Dreiding: A generic force field. *Journal of Physical Chemistry*, 94:8897–8909, 1990.

[25] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople. *Gaussian*. Gaussian, Inc., Pittsburg, PA, U.S.A., 2001.

[26] R. Gentleman and R. Ihaka. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

Chapter 5

Supervised self-organizing maps in crystal structure prediction[1]

This article shows the use of supervised self-organizing maps (SOMs) to explore large numbers of, experimental or simulated, crystal structures and to visualize structure-property relations. The examples show how powder diffraction patterns together with one or more structural properties, such as cell volume, space group and lattice energy, are used to determine the positions of the crystal structures in the maps. The weighted cross-correlation criterion is used as the similarity measure for the diffraction patterns. The results show that supervised SOMs offer a better and more interpretable mapping than unsupervised SOMs, which makes exploration of large sets of structures easier and allows for the classification and prediction of properties. Combining diffraction pattern and lattice energy similarity using a SOM outperforms the separate use of those properties and offers a powerful tool for subset selection in polymorph prediction.

[1]E.L. Willighagen, R. Wehrens, W. Melssen, R. de Gelder, and L.M.C. Buydens, Cryst.Growth & Design, 2007, 7:1738-1745

5.1 Introduction

To explore large databases of crystal structures, self-organizing maps (SOMs) have been introduced recently as a method to map structures, represented by their powder diffraction patterns, onto a two-dimensional grid [1]. This method provides a visualization of the similarities of structures and may show grouping of patterns that cannot easily be found otherwise. Applications include providing an overview of structural diversity of the crystal structures in a database and selection of archetypical structures. To allow prediction of certain crystal properties, for example the values of unit cell parameters, SOMs can be extended to allow supervised learning, and several approaches have been suggested. For example, a combination of a SOM with linear vector quantification has been suggested [2]. However, this does not allow self-organization of the extra property: it imposes a predefined topological structure of the property during training by penalizing the mapping of an object onto a unit when the property does not match. This has the drawback that the topological structure of the property must be given beforehand. For example, when archetypical structure classes are used as the extra property, each map unit is assigned to one of those classes before training is started.

Two supervised SOM methods that do allow self-organization of the property of interest have recently been proposed for this purpose: the XY-fused (XYF) and Bi-Directional Kohonen SOMs [3]. These methods are capable of mapping the topological structures of several properties simultaneously. For example, it allows training of maps using powder diffraction patterns and other crystal properties at the same time. Supervised SOMs allow new ways to map crystal structures by incorporating extra crystallographic information, and allow for the prediction of physical properties. Possible crystallographic properties of interest are space group information, cell volume, and lattice energy. With space group information, the map will be trained in a supervised way to ensure that structures with the same space group are grouped together. As such, a supervised SOM orders crystal structures depending not just on the descriptions of their diffraction patterns, but on a target property too. This has the effect that during training, structures with similar powder patterns but with different space group assignments, get pushed away from each other in the map, while they would be grouped together by an unsupervised SOM. Similarly, when unit cell volume is used as the property, crystal structures with a similar powder diffraction pattern, but with different cell volumes, will map onto different regions in the supervised map, while onto the same unit for the unsupervised approach.

This paper introduces the use of supervised self-organizing maps in crystal structure prediction and describes the method used for this in the next section. The experimental section gives details on the data used, how crystal structures and structure properties are represented and compared, and how the SOMs are trained. The fourth section presents applications of SOMs in structure prediction. It is shown how the supervised SOMs can be used and that combining different types of structural information, such as powder diffraction patterns, space group, structural class or lattice energy, has advantages as compared to unsupervised SOMs. The first example shows the improved classification of structure classes by incorporating unit cell volume information, while the second example

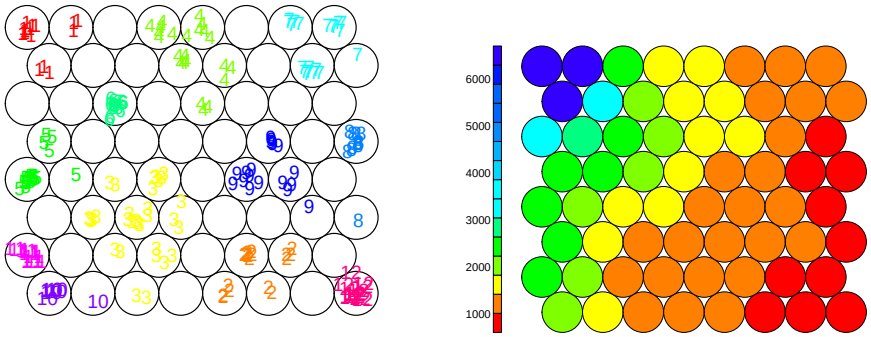


Figure 5.1: Plot showing the Xmap with the mapped structures, colored by class (left) and the Ymap with unit cell volumes, colored by range. The colors in the two plots are unrelated. The class volumes are given in Table 5.1.

Table 5.1: Mean volumes (in Å^3) for the twelve classes shown in Fig. 5.1.

class	volume	class	volume
12	641.93	4	1694.39
8	663.55	10	2033.90
2	1014.68	11	2340.04
9	1150.99	5	2510.71
7	1197.78	6	2568.88
3	1455.53	1	6564.35

shows how the supervised maps can be used for prediction of crystallographic properties. It also demonstrates that different property types can be combined into one single map. The last example shows that supervised SOMs are a viable tool for subset selection in polymorph predictions.

5.2 Supervised self-organizing maps

Unsupervised SOMs provide a non-linear mapping method where the map consists of a two-dimensional hexagonal or rectangular grid of units. Each unit is associated with a codebook vector, or weight vector, of equal length as the input vector. For example, a map trained with diffraction patterns will have weight vectors that resemble a diffraction

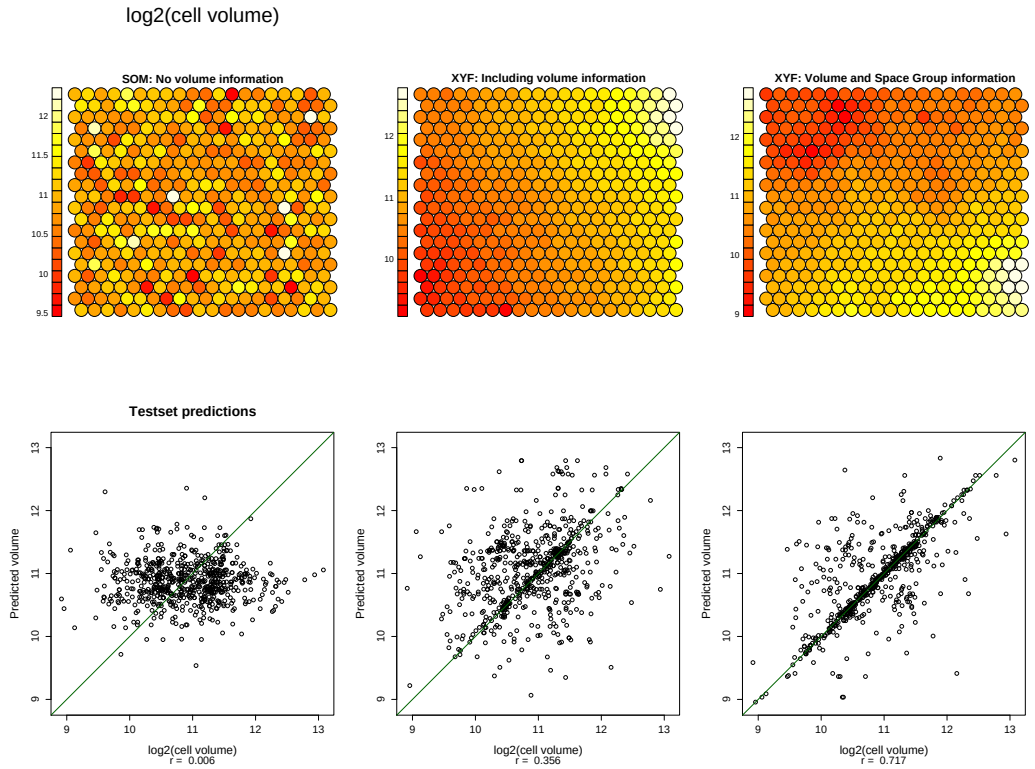


Figure 5.2: Ymaps showing volumes (top) of an unsupervised SOM (left) and two supervised XYF maps, trained with volume alone (middle) and volume as well as space group (right). The bottom plots show the volume prediction for a random test set (seed 1).

pattern. The training of an unsupervised SOM is performed by repeatedly feeding input vectors, representing the training objects, to the map and updating the winning unit in such a way that it more closely resembles this training object. In addition, the immediate neighborhood of the winning unit is updated too. Both the size of the neighborhood and the size of the applied changes to the units are decreased during the training. At the end of the training, only minor adjustments are made which are only applied to the winning unit. The units of a supervised SOM are not only represented by an input vector describing the training objects (X space), but also by a vector describing the properties of interest (Y space). In XYF SOMs, the winning unit is determined by calculating the distance in the fused XY space, instead of the distance in only X space, as is done in unsupervised training [3].

During training, the vectors for the X and Y spaces are both updated, resulting in a map that not only learns about the relations between objects in X space (Xmap), but also learns about the spatial relationship of the objects in Y space (Ymap). For example, applying XYF maps to a data set with structures in two space groups, will result in two distinct areas in the Ymap, one for each space group.

Table 5.2: Correlation coefficients for three seeds for an unsupervised SOM, a XYF map trained with volume information only, and a XYF map trained with space group and volume information.

	Seed 1	Seed 2	Seed 3
SOM	0.01	-0.04	0.01
XYF (volume only)	0.36	0.41	0.41
XYF (class and volume)	0.72	0.28	0.68

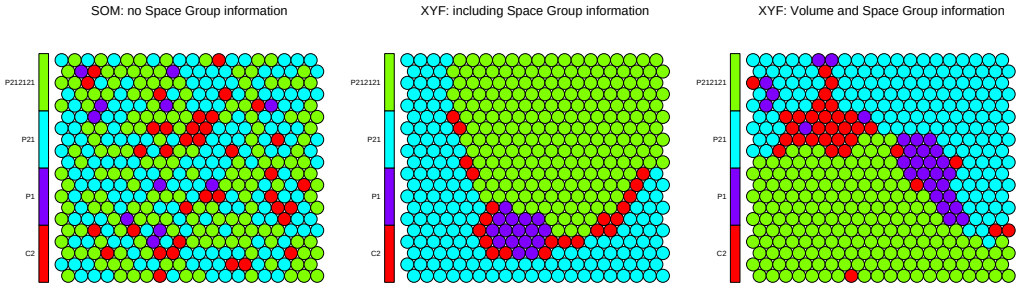


Figure 5.3: Ymap showing space group classification of an unsupervised SOM (left) and two supervised XYF maps, trained with volume alone (middle) and volume and space group (right).

This additional organization of the self-organizing maps allows the study of diversity of crystal structure properties and structural diversity in one single map, and to visualize relations between those two. Trained maps allow prediction of a target property, as represented in Y space of the SOM (Ymap), and as such, offers an alternative to other supervised learning methods like artificial neural networks [4]. For example, when a SOM is trained with diffraction patterns and volume, mapping a new structure onto the map using its diffraction pattern alone will indicate units having Y vectors that can be used for prediction of the volume of the new structure.

5.3 Experimental

5.3.1 Data

This articles uses three data sets to show the applications of supervised training as compared with unsupervised SOMs, and the effects on the organization of the map caused

Table 5.3: The space group prediction results for three seeds for an unsupervised SOM, a XYF map trained with space group information only, and a XYF map trained with space group and volume information.

	Seed 1	Seed 2	Seed 3
SOM	43%	43%	24%
XYF (class only)	87%	86%	85%
XYF (class and volume)	79%	46%	66%

by the supervised training. The first data set is a set of 205 crystal structures, created by searching structures similar to twelve quite different seed structures proposed in [1], effectively creating twelve clusters. The property of interest for this data set is the unit cell volume. The second data set contains 2303 steroid crystal structures, created by searching the CSD for molecules that have a sterane skeleton. Volumes of these structures range from around 600 to around 7000 $Å^3$. The majority of the structures belong to four space groups: $P2_12_12_1$ (978 structures), $P2_1$ (843), $C2$ (98), and $P1$ (93). Both space group and unit cell volume can be used as additional variables in Y space. The last data set contains 1954 simulated crystal structures of acetic acid, created with a polymorph prediction run in space group $P1$. The structures are minimized with respect to their lattice energy and this energy is used as the Y property.

5.3.2 Representation in X and Y space

Comparing crystal structures needs sophisticated methods that require a unique and practical representation of the structures [5, 6, 7, 8, 9]. These, in turn, can also be used for training self-organizing maps, as well as, amongst others, computational clustering of polymorphs [10, 11, 12], and classification of crystal packings [13, 14, 15, 16, 17]. Recently, two new methods have been proposed: a method that uses a radial distribution function incorporating electronic features to describe packing patterns of molecular crystal structures [8], and a method that uses powder diffraction patterns to represent crystal structures [15].

The latter has shown useful for self-organizing maps [1], and is used in this article, too. The first two data sets consist of diffraction patterns with 2θ angles up to 25 degrees, with a sampling rate of 0.05 degrees. Values below one degree are not taken into account since no features are present. A pattern therefore consists of 481 intensity values (counts). The Cu-Kα_1 wavelength is used for calculation of the powder diffraction patterns. These settings lead to a crystal structure description with a resolution of approximately 3.6 Å. The acetic acid data set consists of diffraction patterns with 2θ angles from 0 up to 30 degrees with 501 intensity values, using the same sampling rate. For this data set the resolution is approximately 3.1 Å. Other choices are possible, and exact values do not

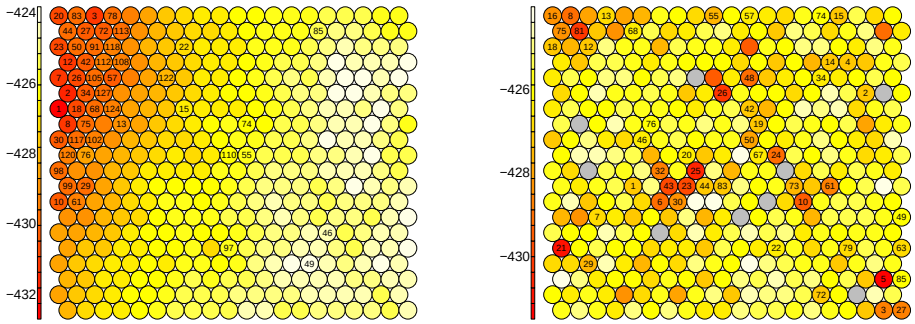


Figure 5.4: Trained XYF map (left) and unsupervised SOM (right) with a selected subset of 50 structures. Numbering is according to increasing lattice energy, with structure 1 having the lowest energy. The lattice energy for the units of the unsupervised map are the means of the lattice energies of the structures mapped onto each unit.

seem critical. Intensity counts are scaled by taking square roots, analogous to the IsoQuest program [18] The largest intensity is then set to 100 units.

Space group information is represented in Y space by one column for each space group, for which the value is 1 if the object belongs to this class, and 0 otherwise. Consequently, the number of columns equals the number of space groups. Similarly, structural classes are represented in the same way. Both classifications are, in the remainder of this paper, referred to as *classes*, not to be mistaken with crystal classes. Continuous variables are represented by one column in Y space and can be scaled. For example, a logarithmic scale and mean centering is applied to unit cell volume. The logarithmic scaling is useful as a difference of 100 A^3 for small volumes is more important than for large volumes. The lattice energies calculated in the polymorph prediction data set are mean centered.

5.3.3 Similarity calculations

Because the X and Y spaces are different in nature, crystal structure similarity is calculated for both spaces separately; the distances in X and Y spaces are then rescaled so that the largest value in each space equals one. Finally, a weighted sum is taken to obtain the overall similarity. Equal weights are employed for the X and Y similarities in this paper. The measure used in this article for the powder diffraction patterns in X space, is the weighted cross-correlation (WCC) [15, 1, 19]. When calculating the WCC criterion, a triangle width of 1.0° (20 data points) was used. Triangles that are too narrow ignore the neighborhood of the features, while too broad triangles lead to uniformly high similarity

Table 5.4: The top twenty selected representative units (first column), for the supervised XYF map (see Figure 5.4), and their lowest energy structures (second column) are given, together with a lists of all structures mapped onto that unit (third column).

unit	lowest energy structure	all structures mapped on that unit
261	1	1, 6
281	2	2, 4, 14
383	3	3, 5, 11, 17
301	7	7, 21
241	8	8, 9, 16
141	10	10, 73, 104, 106, 114, 263, 1525
321	12	12, 19
244	13	13, 165, 173, 232, 632
268	15	15, 37, 166, 271, 336, 455, 485, 524, 549
262	18	18, 39, 71
381	20	20, 25, 28, 31, 33, 36, 38, 59
348	22	22, 380, 407, 415
341	23	23, 24, 40, 54, 77, 81, 359
302	26	26
362	27	27, 35, 41, 52, 53, 62, 64, 111, 150, 385, 973, 1014
162	29	29, 96, 136, 567
221	30	30, 32, 43, 45, 69, 84, 101, 115
282	34	34, 63
322	42	42, 47, 48, 51, 58, 65, 66, 82
361	44	44, 128, 182

values without enough discriminatory power. A triangle of 1.0° was found to give generally good results [15, 1]. The similarity used for the Y space is the Euclidean distance for continuous and mixed-type variables, and the Tanimoto distance for class variables.

5.3.4 SOM training

In addition to unsupervised SOMs, an XYF map has only one extra parameter: the weight determines the contribution of the X and Y spaces in the calculation of the overall distance. The map size, map topology, the learning parameter α and the neighborhood function apply to both types of SOMs. The map size scales with the amount of detail visualized: more units allow training of more distinct features. Commonly, the number of map units is at least twice the number of classes, and less than the number of objects [20]. For the steroid and acetic acid data sets we, therefore, used 20x20 maps, and for the 12-class data set we used a 8x8 map.

The settings for the other parameters are identical to those used in earlier work on SOMs for powder diffraction patterns [1]: hexagonal networks are used in which the

Table 5.5: The top 20 selected representative units (first column), for the unsupervised map (see Figure 5.4), and their lowest energy structures (second column) are given, together with a lists of all structures mapped onto that unit (third column).

unit	lowest energy structure	all structures mapped on that unit
165	1	1, 791, 1328, 1662
298	2	2, 472, 533, 644, 741, 1289
19	3	3, 11, 17, 113, 639, 1244, 1263
337	4	4, 421, 586, 781, 965, 1379
59	5	5
147	6	6, 101, 118, 120, 129, 131, 149, 159, 178, 232, 632, 858
123	7	7, 452, 1229, 1655
382	8	8, 9, 39, 71, 845, 1176
155	10	10, 114, 234, 257
343	12	12, 252, 295, 316, 318, 419, 1158, 1422, 1630
384	13	13, 213, 359, 809, 811, 987, 1058, 1507
336	14	14, 1163, 1327
397	15	15, 37, 271, 453, 683, 1321, 1366, 1748
381	16	16, 69, 84, 102, 138, 140, 168, 172, 203, 221, 448, 491, 498, 1185, 1296, 1306
341	18	18, 160, 517, 959, 1055, 1334
252	19	19, 307, 792, 836, 870
208	20	20, 924, 972, 988, 1423
81	21	21
93	22	22, 664, 925, 1293, 1563
168	23	23, 40, 121, 133, 182

distances of a unit to all six direct neighbors are equal. The maps are not toroidal, and thus have edges, and units at these edges have fewer neighbors than the units in the center. The number of training events used in this paper is 200 times the number of patterns in the training set. With each event, the winning unit, and the units in the neighborhood, are updated with a weighted average of the unit weight vector and the new pattern, in both X and Y space. The weighting of the average depends on the learning parameter α and the neighborhood. During the training, the α decreases linearly from 0.05 to 0.01, and the neighborhood decreases exponentially from covering two-third of the map, to only the winning unit after one third of the training phase has passed.

5.3.5 Software

All methods are implemented in R [21], and available from the R package "wccsom" (version 1.2.0) [1, 22]. The package uses C code for the time critical calculations of the WCC criterion. The package provides methods to train XYF maps using the WCC to

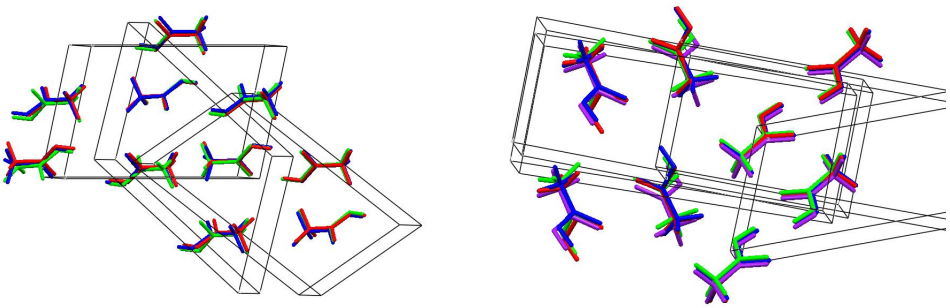


Figure 5.5: Structures 2 (red), 4 (green) and 14 (blue) from unit 281 (left) and structures 3 (red), 5 (green), 11 (blue) and 17 (purple) from unit 383 (right). Unit cell details for the first two structures are given in Table 5.6.

calculate the distances between diffraction patterns in X space. It also provides various plotting methods to visualize Xmap and Ymaps.

5.4 Applications

The applications in this section show that supervised training allows maps to have organization in X space as well as in Y space, offering new possibilities to analyze large sets of crystal structures, and prediction of crystal properties based on the powder diffraction patterns. The examples of the XY-fused SOMs demonstrate possible applications of these supervised maps in crystallography and crystal structure prediction. The method is not restricted to these applications, and other properties can easily be used to train maps, such as packing patterns, hydrogen bond networks or other crystal properties.

5.4.1 Unit cell volume in Y space

The first example shows the prediction of the unit cell volume from the powder diffraction pattern, and the effect of taking into account this volume as Y space variable when training the XYF map. A hexagonal XYF map of 8x8 units was trained for the set of 205 crystal structures grouped into 12 clusters in 200 iterations, using the Euclidean distance measure on the log-scaled and mean-centered volumes.

Figure 5.1 shows a trained XYF map on which the training set was mapped after training (left), and the Y weight vector values, i.e. the unit cell volumes (right). This Ymap shows areas with high volumes (>6000 Å) for the units in the top left corner, while showing low volumes (<1000 Å) in the bottom right corner. Indeed, when the 205 structures are mapped on the trained map, by assignment based on similarity in X space only, it is clearly visible that these clusters are now ordered according to their volume (see

Table 5.6: Crystal data for the acetic acid polymorphs pairs 1 and 6, 2 and 4, and 3 and 5.

	1	6	2	4	3	5
space group	P1	P1	P1	P1	P1	P1
a, Å	5.49	4.45	4.501	4.454	4.086	4.086
b, Å	6.06	6.75	7.238	6.747	5.547	5.473
c, Å	9.09	11.63	10.666	11.628	14.112	14.245
α, deg	90.0	104.1	101.3	104.1	90.0	87.6
β, deg	90.0	112.5	115.0	112.5	100.5	77.4
γ, deg	90.0	90.0	90.0	90.0	90.0	85.8
lattice energy	-108.5	-107.9	-108.2	-107.9	-107.9	-107.9
density, g cm^{-3}	1.320	1.298	1.297	1.282	1.293	1.287

Table 5.1): the cluster of structures with the lowest mean volume, cluster 12, maps onto a unit in the bottom right corner of the SOM, next to cluster 8 which has a low volume too. The cluster with the highest volume, cluster 1, maps onto units in the top left corner. Class 1 and 12 will never be mapped next to each other, because the supervised training imposed a large distance between them, reflecting the large difference in volumes of the two classes. Noteworthy is the non-linear behavior of volumes in the Y weight vectors: the mapped volumes show a sharp decrease just outside the top left corner.

As was noted earlier, structures can be mapped onto the map using similarity in X space only, by calculating WCC criteria for all units. Therefore, the method can be used to predict Y values. For example, the volume can be predicted for the mapped crystal structures by taking the winning unit's Y value, and compare it with the real volume. Now, for this set, volume prediction is not really challenging and is merely illustrative and proof-of-concept: the correlation coefficient between predicted and true volumes is 0.98, with a mean WCC of 0.986 and a lowest WCC of 0.879.

5.4.2 Adding space group information in Y space

Training an XYF map is not restricted to just one property; multiple properties can be represented in Y space. The next example shows a XYF map trained using 2012 structures in the $P2_12_12_1$, $P2_1$, $C2$ and $P1$ space groups from the steroid data set with 2303 structures. The space group classification information in Y space is represented by a set of four numbers indicating class likeliness (0.0 for unlikely, 1.0 for likely). For example, (0.0, 1.0, 0.0, 0.0) indicates that the structure has the $P2_1$ space group. A fifth number is used for the mean-centered logarithm of the volume. Similarity in Y space is measured with the Euclidean distance.

Figure 5.2 shows the Ymaps for three SOMs: the left most is an unsupervised SOM,

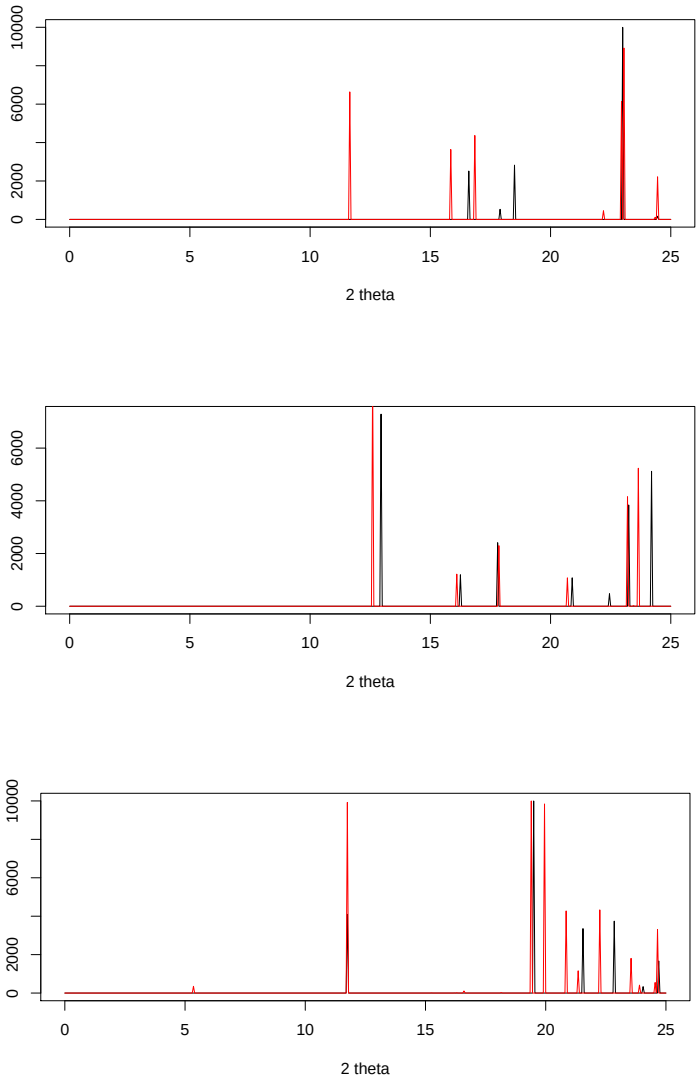


Figure 5.6: Powder diffraction patterns of structure 1 and 6 (top, WCC=0.83), 2 and 4 (middle, WCC=0.70) and 3 and 5 (bottom, WCC=0.82). Details for the four structures are given in Table 5.6.

where the volume associated with a unit equals the mean volume of the training structures mapped onto that unit. The right two maps are XYF maps, trained with only volume (middle) and volume and space group (right). The top Ymaps show that the unsupervised SOM does not have any spatial structure regarding the volume, which the XYF maps were specifically trained for. Prediction of unit cell volume can be performed for all three maps

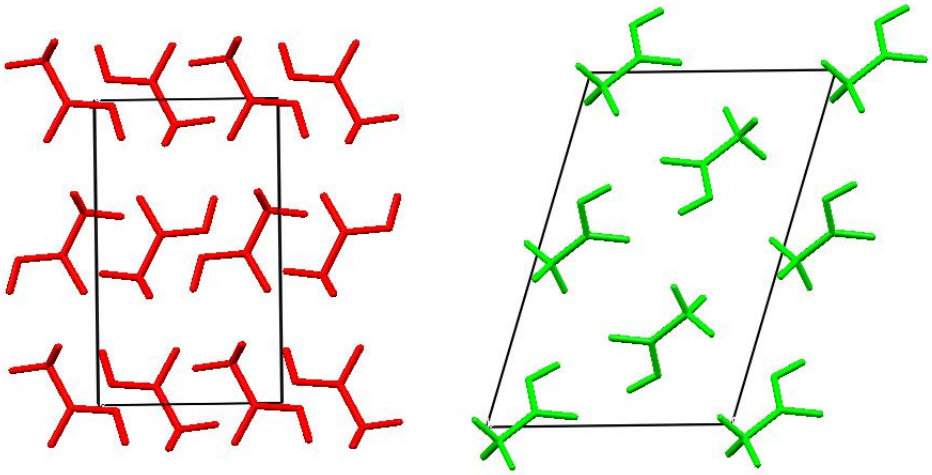


Figure 5.7: Structures 1 and 6 from unit 281, which are both mapped onto unit 261, but are structurally different.

by mapping test structures onto the map, and assigning as prediction volume the volume in the Ymap of the unit cell onto which the structure is mapped. The bottom plots in Figure 5.2 show the predicted versus the true volumes for a random test set for one of the three replicate maps trained starting with different seeds, and Table 5.2 gives the correlation coefficients. The unsupervised map is not able to predict unit cell volume at all, as is clear from the zero correlation coefficient. The XYF map trained with volume, however, is able to predict, to some extent, the volume, reflected by the positive, though low, correlation. Including space group information during training of the map further increases the predictive power, an occasionally badly trained map excluded (seed 2).

Using the same approach, space group information can be predicted. Figure 5.3 shows the Ymap with the space groups associated with the units of the trained maps. For the unsupervised map, space groups were assigned by majority vote. Two XYF maps were trained: one with only space group information, and one with space group and volume information. Again, the unsupervised map does not show spatial organization of the Ymap, which is present in the supervised maps. When volume is taken into account during the training, the spatial organization of the space groups is somewhat distorted, caused by the dual organizational restrictions imposed by the space groups and volumes. Table 5.3 shows the prediction results for one test set for three replicates using different seeds for each of the three map types. The results are best for a XYF map trained with only space group information. Unlike the fact that including space group improves results

when predicting unit cell volume, the reverse does not seem to apply: including unit cell volume when predicting space groups deteriorates the prediction.

Note that the SOM and XYF (class and volume) mappings in Table 5.2 and 5.3 are the same; there is no need to retrain since prediction for both cell volumes and space groups can be performed. The mapping corresponding to seed 2 clearly corresponds to a situation where the training is stuck in a local optimum: both cell volumes and space group predictions are bad, but still better than unsupervised prediction.

5.4.3 Analyzing simulated polymorphs

An interesting application of supervised self-organizing maps is the analysis of simulated structures generated in polymorph prediction. Polymorph predictions often generate hundreds, if not thousands, of structures, from which a subset of structures needs to be selected which is likely to contain real polymorphs occurring in nature. These experimentally determined polymorphs usually appear somewhere in the list of predicted low energy structures, and selecting a few stable structures based on the predicted lattice energy alone has not shown sufficiently reliable yet [23, 24], while a recent test at an international conference has shown that crystallographic intuition based on visual inspection as a complement to computational methods has doubtful reliability too [25]. Moreover, visual inspection of all structure pairs is practically impossible. Energy-density plots have been used for subset selection, but density is not always considered to be a useful property to base selection on. Here, we propose the use of XYF maps as alternative for selecting structures from large sets of predicted structures, with diffraction patterns and energy as the basic properties for non-linear 2D mapping. The setup could be extended to make use of additional properties, such as H-bonding patterns, though that has not been explored in this work.

This example shows an XYF map trained for the set of 1954 simulated acetic acid polymorphs, trained with the diffraction patterns in X space and the calculated lattice energy in Y space. Again, the WCC criterion and the Euclidean distances are used to calculate similarities during training. Low energy structures are likely candidates of naturally occurring crystal structures, and the low energy areas on the map are, therefore, the areas of interest. Selecting crystal structures could be performed by taking one structure for each unit in the low energy area of the map, e.g. the one with the lowest energy. This way structural diversity is achieved in the selected subset. However, because we want to make sure we represent all low energy structures, we take the following approach. Disregarding their location on the map, 50 units are selected that contain the low energy structures. Mapping the 50 lowest energy structures results in fewer occupied units, because many of those structures have similar powder diffraction patterns, and will end up in the same unit. Therefore, an increasing amount of lowest energy structures is mapped onto the map, starting with 50 structures, until exactly 50 units are occupied. For example, an XYF map requires, about, the 275 lowest energy structures to be mapped to get 50 distinct units occupied, while an unsupervised maps requires around 200 lowest energy structures.

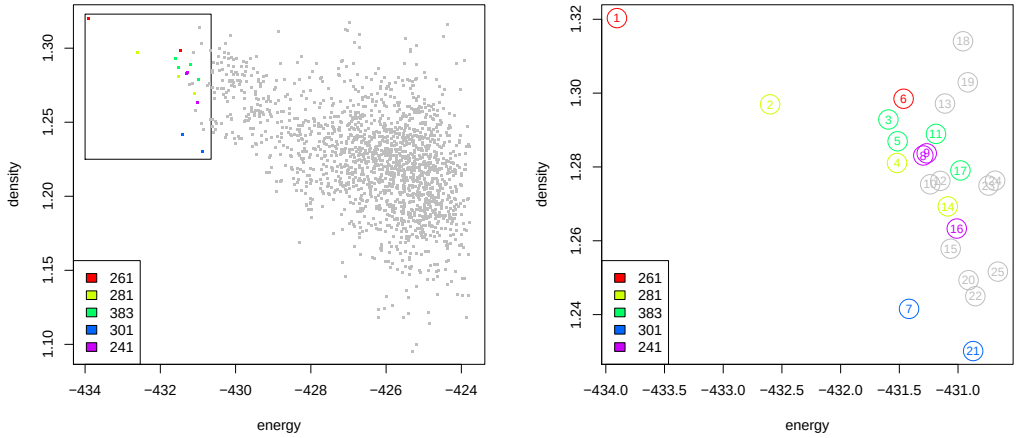


Figure 5.8: Energy-density plots for the complete acetic acid data set (left) and the subset of structures mapped onto the five lowest energy units of the XYF (261, 281, 383, 301 and 241) colored by unit (right).

Figure 5.4 shows the Ymap with the energy of a XYF map, and an unsupervised map trained for the same data. The energies given for the unsupervised map are calculated by taking the mean of the energies of the structures that map into that unit after training. The XYF map has a low energy region in the top left corner of the map, whereas the unsupervised map does not have organization with respect to the energy at all. Likewise, the low energy structures are mapped together onto the supervised map, but not when mapped onto the unsupervised map. The structures are numbered by increasing predicted lattice energy, where structure 1 has the lowest energy. Tables 5.4 and 5.5 list all structures mapped onto the twenty selected units with the lowest energy structures. For example, four structures are mapped onto unit 383 of the supervised map (Table 5.4), of which structure 3 has the lowest energy.

Effectively, including the energy as property has the effect that the supervised map encourages low energy structures to map onto the same area of the map. As a result, structures may be structurally more different while still ending up in the same or a neighboring unit during and after training of the SOM. For example, structures 2 and 4 are mapped onto the same XYF map unit, while mapped onto different units of the unsupervised map, though not that far apart. Being mapped onto the same unit, it is easier to detect their relative similarity, which is more difficult when one has to analyze the content of one unit *and* all neighboring units. While structure 2 and 4 are only one unit apart, detecting their similarity would have required to analyze 18 units instead of one.

This nicely shows the effect of including the lattice energy during training. While the diffraction pattern alone is not enough to clearly group together similar packing patterns,

the combination makes this more apparent. For example, structures 2, 4 and 14 (see Figure 5.5 left), are all mapped onto unit 281 and show the same packing. However, their powder diffraction patterns show a lower similarity, as expressed by the WCC measure, than the patterns of structures 1 and 6 (see Figure 5.6), which have a different packing (see Figure 5.7). The lattice energies could not have been used to highlight the similarity between structures 2, 4 and 14, either. While other units show this perfect alignment of crystal packing too, such as units 383 (see Figure 5.5), 301 and 241 (not shown), unit 261, with structures 1 and 6, shows that this is not always the case. While mapped onto the same unit, the packing of structures 1 and 6 is quite different (see Table 5.6). It is noteworthy that in the unsupervised map, structures 1 and 6 are mapped only one unit apart, too. To address false positives, other crystal structure properties could be included, such as hydrogen bonding patterns, during training of the SOM, to further enhance organization.

To visualize the subset selection, Figure 5.8 shows energy-density plots for all structures in the data set (left) and the 25 lowest energy structures (right). The colored structures are colored according to the units they are mapped onto (261, 281, 383, 301 and 241), while the gray structures are mapped onto other units. Units 281, 383, 301 and 241 showed perfect overlap of crystal structures, which is not apparent from the energy-density plot, with structures 8 and 9 as a possible exception. The Gray structures in the right plot are clustered into seven distinct units, as can bee seen in Table 5.4.

5.5 Conclusions

Mapping sets of crystal structures onto self-organizing feature maps has been shown to have many applications. This article shows how this approach can be extended to create maps of which the topological structure not only depends on the powder diffraction data, but on other properties of interest, such as cell volume, space group, and lattice energy, or a combination of both, too. These supervised maps not only give a better mapping, they can also be used to predict properties trained in the Ymap, using similarity in the Xmap alone, and for subset selection in polymorph prediction.

The applications show that unit cell volume, space group classification and lattice energies can be used in Y space as dependent properties. The map trained for the data set with 205 structures and twelve structural classes shows that these classes will organize on the map according to their mean volumes when trained with the unit cell volume in Y space, adding interpretability to the map.

The steroid data set was used to demonstrate the possibility of including different property types in Y space, such as binary class information and continuous variables such as unit cell volume and lattice energy. The Ymaps of the space group classification and unit cell volume show a shared but complementary topology, where space groups are roughly separated in one dimension and the unit cell volume in another. The P_1 and $C2$ space group classes are broken up into separate areas due to the imposed organization on

volume.

The third application shows the use of a XY-fused SOM in selecting a structurally diverse set of low lattice energy structures from a polymorph prediction set. The similarities in X space ensure that different units have structural diversity, while the similarities in Y space impose a topological structure that results in low and high energy regions. The example shows that the supervised map, using powder diffraction pattern and lattice energy, gives much better clustering of crystal structures with similar packing patterns. As such, we believe that this application offers a flexible and viable alternative to the often used energy-density diagrams for subset selection in polymorph predictions.

Concluding, the XY-fused SOM provides a new tool to visualize and analyze large set of crystal structures via powder diffraction pattern data, and a new method to predict crystal structure properties from their diffraction patterns.

Bibliography

[1] R. Wehrens, W.J. Melssen, L.M.C. Buydens, and R. De Gelder. Representing structural databases in a self-organizing map. *Acta Crystallographica*, B61:548–557, 2005.

[2] T. Kohonen. *Self-Organizing Maps*. Number 30 in Springer Series in Information Sciences. Springer, Berlin, 3 edition, 2001.

[3] W.J. Melssen, R. Wehrens, and L.M.C. Buydens. Supervised Kohonen networks for classification problems. *Chemom. Intell. Lab. Syst.*, 2006. In press.

[4] S. Habershon, E.Y. Cheung, K.D.M. Harris, and R.L Johnston. Powder diffraction indexing as a pattern recognition problem: a new approach for unit cell determinination based on an artificial neural network. *Journal of Physical Chemistry A*, 108:711–716, 2004.

[5] A. V. Dzyabchenko. Method of crystal-structure similarity searching. *Acta Crystallographica*, B50:414–425, 1994.

[6] L. C. Andrews and H. J. Bernstein. Bravais lattice invariants. *Acta Crystallographica*, A51:413–416, 1995.

[7] A. Kálmán and L. Fábián. Volumetric measure of isostructurality. *Acta Crystallographica*, B55:1099–1108, 1999.

[8] E.L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L.M.C. Buydens. Method for the computational comparison of crystal structures. *Acta Crystallographica*, B61(1):29–36, Feb 2005.

[9] R. Hundt, J.C. Schön, and M. Jansen. Cmpz - an algorithm for the efficient comparison of periodic structures. *J.Appl.Cryst.*, 39:6–16, 2006.

[10] P. Verwer and F. J. J. Leusen. *Computer Simulation to Predict Possible Crystal Polymorphs*, volume 12 of *Reviews in Computational Chemistry*, chapter 7. Wiley-VCH, New York, 1998.

[11] J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. Van Eijck, P. Verwer, and D. E. Williams. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica*, B56:697–714, 2000.

[12] W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scherega, B. Schweizer, M. U. Schmidt, B. P. Van Eijck, P. Verwer, and D. E. Williams. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica*, B58:647–661, 2002.

[13] J. Perlstein, K. Steppe, S. Vaday, and E. M. N. Ndip. Molecular self-assemblies. 5. analysis of the vector properties of hydrogen bonding in crystal engeneering. *Journal of the American Chemical Society*, 118:8433–8443, 1996.

[14] B. Moulton and M. J. Zaworotko. From molecules to crystal engineering: Supramolecular isomerism and polymorphism in network solids. *Chemical Reviews*, 101:1629–1658, 2001.

[15] R. De Gelder, R. Wehrens, and J.A. Hageman. A generalized expression for the similarity spectra: application to powder diffraction pattern classification. *Journal of Computational Chemistry*, 22(3):273–289, 2001.

[16] M. D. Hollingsworth. Crystal engineering: from structure to function. *Science*, 295:2410–2413, 2002.

[17] G. Ilyushin, N. Blatov, and Y. Zakutin. Crystal chemistry of orthosilicates and their analogs: the classification by topological types of suprapolyhedral structural units. *Acta Crystallographica*, B58:948–964, 2002.

[18] R. De Gelder and J.M.M. Smits. SYSTER and ISOQUEST: How good and unique are your data and structure? *Acta Crystallographica*, A60:s78, 2004.

[19] J. Van de Streek and S. Motherwell. Searching the cambridge structural database for polymorphs. *Acta Crystallographica*, B61:504–510, 2005.

[20] W.J. Melssen, J.R.M. Smits, L.M.C. Buydens, and G. Kateman. Using artificial neural networks for solving chemical problems. part ii. kohonen self-organizing feature maps and hopfield networks. *Chemom. Intell. Lab. Syst.*, 23:267–291, 1994.

[21] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[22] R. Wehrens. wccsom - SOM networks for comparing patterns with peak shifts. http://cran.r-project.org/, 2006.

[23] G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, and P. Verwer. A third blind test of crystal structure prediction. *Acta Crystallographica*, B61:511–527, October 2005.

[24] P.G. Karamertzanis and S.L. Price. Energy minimization of crystal structures containing flexible molecules. *Journal of Chemical Theory and Computation*, 2006. Web Release.

[25] G.M. Day and W.D.S. Motherwell. An experiment in crystal structure prediction by popular vote. *Cryst. Growth Des.*, 6(9):1985–1990, 2006.

Chapter 6

Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators[1]

Examples of the use of the RSS 1.0 (RDF Site Summary) specification together with CML (Chemical Markup Language) to create a metadata based alerting service termed CMLRSS for molecular content are presented. CMLRSS can be viewed either using generic software or with modular opensource chemical viewers and editors enhanced with CMLRSS modules. We discuss the more automated use of CMLRSS as a component of a World Wide Molecular Matrix of semantically rich chemical information.

[1]P. Murray-Rust, H.S. Rzepa, M.J. Williamson, and E.L. Willighagen, J.Chem.Inf.Comput.Sci., 2004, 44(2):462-469

6.1 Introduction

There is increasing recognition that the World Wide Web has vast untapped potential as an infrastructure for structured data interchange rather than just being a medium for delivering documents. This recognition underpins the Semantic Web [1], Berners Lees vision of its evolutionary future. Its construction will involve developing mechanisms which precisely and predictably associate data with descriptions of its meaning, context, and validity (whether it is fit for purpose). XML is now universally recognized as providing syntactic architectures for achieving this. XML is itself a specification for creating families and subfamilies of more specific markup languages. The best known of these is XHTML, which evolved from the original requirements of the Web to create documents which could be rendered readable for humans via the Web browser. In fact, XML was designed to serve an even more fundamental role for specifying data and data structures. Via a formalism known as namespacing, several XML languages can in turn be combined to create a compound document, and these components can be transformed into other appropriate forms by invoking other XML-based tools known as stylesheets. These can be appropriate either for presentation to a human for reading or for further processing (transformation) by empowered software agents according to defined algorithms. In recognition of the dual purposes that XML can serve, we have coined the term datuments [2, 3] to describe these compound information objects.

As the structure of datuments and the number of components they may contain grows more complex and the datuments themselves become larger (possibly very much larger), methods for achieving higher order organization and aggregation become required. Metadata (data about data) provides a mechanism for providing concise descriptions of the type of content expected in the datument, enabling high level decisions about further processing or filtering to be made. What is required is a more finely grained elaboration of the MIME approach we used to achieve appropriate postprocessing of discrete data files on the first generation Web [4].

At this stage, it is worthwhile noting an early experiment of ours in creating a complex environment of documents, chemical data, metadata, and processes applied to the collection, using the Web technologies available in 1995. The ECTOC electronic conferences [5] were designed to investigate innovative electronic metaphors for the conventional but expensive and time-consuming physical meetings which the scientific communities have evolved over many decades to promote cross fertilization of ideas among humans. Each of the four ECTOC conferences held during the period 1995-1998 contained about 100 posters and articles. These were an intertwingled mixture [6] of bit-mapped images, chemical data expressed in a variety of formats [4], discussion forums and lists of titles, with associated provenance of authors, comments by participants, and clear time stamps. Part of this experiment was an attempt to create navigational aids to this diverse but inter-related information collection which would help participants to identify chemical subject matter of interest to them. This would in turn help identify similarities in this material which would promote serendipitous chemical discovery. While conventional navigational aids were presented, (tables of contents, subthemes, indices) we also introduced

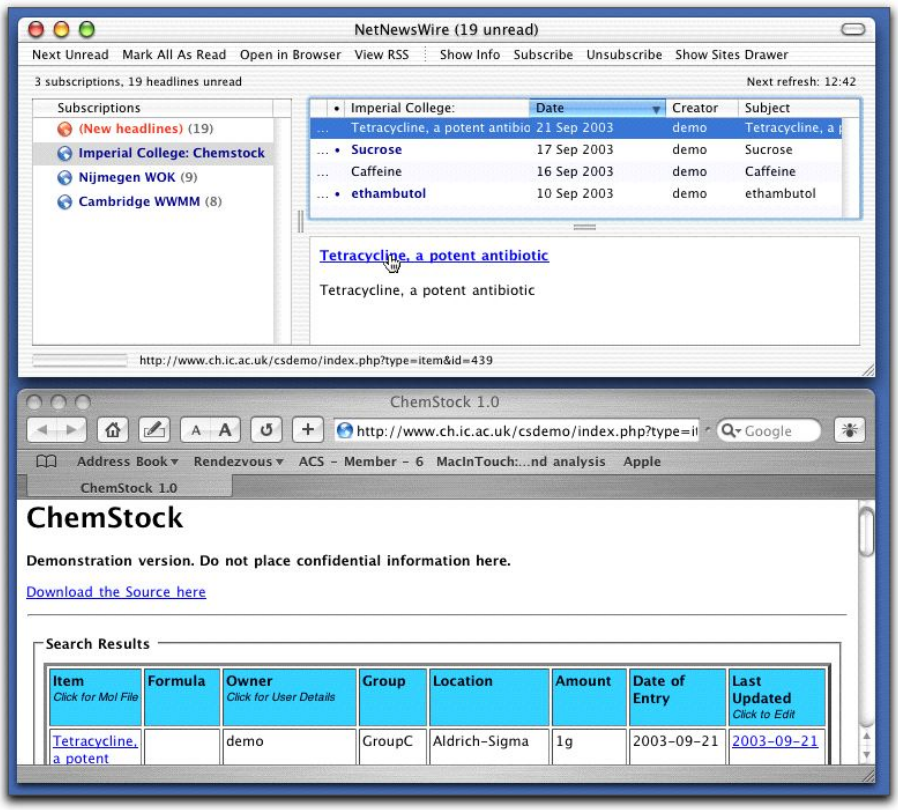


Figure 6.1: A generic RSS Viewer illustrating the RSS feed from `http://www.ch.ic.ac.uk/csdemo/feed.php` (top) and expansion of one item within a browser window (bottom).

a novel metadata based mechanism using the Meta Content Framework (MCF) which had been developed by a small group within Apple Computer. MCF was used to provide metadescriptions of the various conference components and was presented to the human as a nonlinear visual navigation map of the conferences containing links between related components of the conferences based on this metadata. The MCF-based map and software to view it was included on the subsequent ECHET96 CDROM archives, although its use was not developed further at the time. A particular limitation was that chemical information such as "how many molecules are described in this article" still had to be (slowly) organized and then discovered by the human editor or reader.

MCF itself underwent a number of evolutions after being abandoned by Apple in 1996, including adoption by Netscape for use in their own information portals under the name RSS. The ideas espoused by MCF were also adopted by the W3C for their Resource Description Framework (RDF), itself seen as an integral part of the Semantic Web noted in our introduction. These various concepts, along with a decision to recast the syntax into XML, merged around the year 2000 with the specification of a protocol

now known as RSS (RDF Site Summary) 1.0. The background history to this evolution has been recently summarized elsewhere [7], and this latter article also provides a concise description of various more formal metadata schemas that can be incorporated into RSS. These include the standard Dublin Core (DC) schema and PRISM [8] which provides an XML metadata vocabulary specifically for journal publishing.

With RSS now cast as an XML language, for the first time it becomes possible to consider how an entire collection of data, metadata, and information could be constructed using XML components (something not possible at the time of the ECTOC conferences) and which in turn could make use of the increasing array of standard (often opensource) software tools which have become available for processing XML. In an earlier article where we first introduced the ideas behind RSS [9], we concluded by alluding to the prospects of such unification in the specific area of chemistry. In the present article, we provide explicit examples of the use of RSS to provide metainformation about three diverse chemical sites, including mechanisms for molecule discovery largely absent in conventional Web pages. We also show how the use of XML throughout greatly facilitates the development of authoring applications which make use of these concepts via reuse of standard components and tools.

6.2 Implementations of RSS for chemical data sources

We have previously described [9] the structure and use of a basic RSS document, noting how XML namespaces [10] allowed explicit chemical information and metadata to be added. Here we elaborate upon the topic of namespaces which we had introduced in the earlier article and then illustrate this usage via three deliberately diverse examples of how these concepts can be added to repositories of chemical data.

6.2.1 Namespaces and RSS 1.0

Namespaces are central to modern XML but not always widely deployed in some domains, including chemistry. Large documents (e.g. journal articles, regulatory submissions, patents, books, etc.) may contain material from many disciplines and be created by many authors. Moreover material may be copied or transcluded from other sources. It is unrealistic to expect a globally controlled vocabulary, and namespaces allow authors to create local information components and merge them without name collisions. Thus chemistry and XML both use the vocabulary "element" which would collide unless disambiguated.

Namespaces use URIs [10] for disambiguation. The creator of a namespace devises a unique URI, usually based on their domain name to provide uniqueness. Thus many XHTML documents start with the following syntax *<html xmlns="http://www.w3.org/-1999/xhtml">*.

This states that, by default, all elements and attributes in the document belong to the "http://www.w3.org/1999/xhtml" namespace, conventionally referred to as the

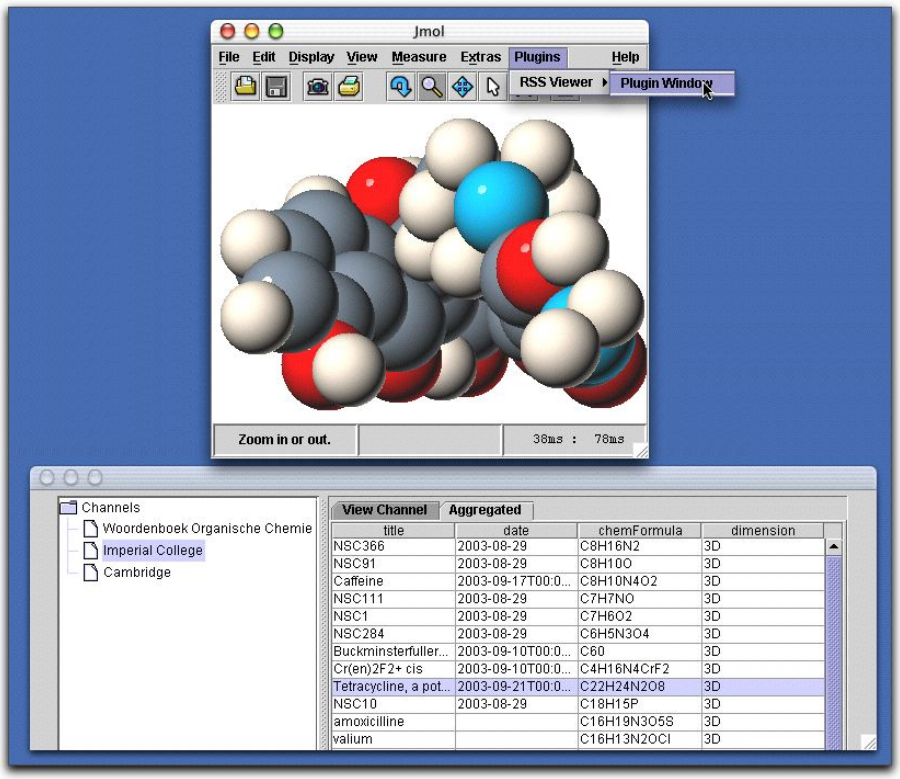


Figure 6.2: The Jmol 3D Molecule viewer showing the RSS plugin window. Three CMLRSS channels are shown sorted by date, with the Jmol window showing the most recent item. The formula is computed from the CML molecular information, and the dimension indicates what type of coordinates are available (1D, 2D, 3D, 2D+3D, fractional etc).

"XHTML V1.0" namespace. The use of the HTTP protocol for namespaces is an unhappy and confusing syntax. It is not required and could be replaced by a URN (a naming, rather than addressing convention). We dispose of the following common myths:

- "You have to be connected to the web to use namespaces". In fact no XML tool should try to resolve these as addresses and if it does it is an error.
- "There is a web page with something useful at the URL address". It is not an address, and it is only coincidental if there is a page to which it resolves.

In practice many namespace designers do put some form of specification or help pages at the "URI address", but there is no consistency in content or syntax.

The namespace specification should be read carefully by designers of multinamespace documents (including RSS), but the following simple guide and example illustrating

namespace use in an RSS 1.0 document [11] (Scheme 1) is sufficient for this article.

- The document may (but need not) start with an XML declaration (line 1).
- The document may (but need not) have a DTD reference: *<!DOCTYPE foo SYSTEM "http://foo.org/ bar.dtd">*. In practice it is difficult to construct DTDs for multinamespace documents, and they are not normally DTDvalidatable, so our approach does not use them. It is possible to create schemas which describe and validate, but content models and attributes are complex in RDF and schema-based validation is not always cost-effective [12].
- Processing instructions (line 2) are not under namespace control. PIs are hints or instructions to processing software but do not affect the content of the document. The PI target ("xml-stylesheet") is used in user-agents (browsers) which support the W3C style guidelines. It means "if you are stylesheet aware, and if you support CSS stylesheets ("type" pseudoattribute) then retrieve the stylesheet at URL ("href") and apply it". In this example, the stylesheet specified ensures that if the RSS feed is displayed in a browser, the result is at least readable.
- There can be any number of namespaces in an XML document. For RSS there will normally be at least RSS, RDF, and DC. Most human-readable news feeds also include XHTML. In the example (Scheme 1) there are the following:

```
"http://www.w3.org/1999/02/22-rdf-syntax-ns#"
"http://purl.org/rss/1.0/"
"http://usefulinc.com/rss/manifest/"
"http://purl.org/dc/elements/1.1/"
"http://www.xml-cml.org/schema/cml2/core"
```

- Namespaces do not need to be declared at the start of the document unless they are required in the rootElement. They can also be declared several times. All namespaces (except the default namespace, which in this example corresponds to the schema for RSS 1.0 itself) are associated with a prefix. This prefix is arbitrary and is required to be unique only within the document. The prefixes are determined by the xmlns pseudoattribute mechanism. Thus, xmlns:dc="http://purl.org/dc/-elements/1.1/" associates the prefix "dc" with the namespace "http://purl.org/-dc/elements/1.1/". Any element (or attribute) whose name starts with dc: has a namespace URI of xmlns:dc="http://purl.org/dc/elements/1.1/". Thus the element <dc:date>2003-0916T00:00:00-00:00</dc:date> consists of the pair localname+namespaceURI of "date"+"http://purl.org/dc/elements/1.1/". This is what is passed to an application program, and the actual prefix used is irrelevant. We emphasize this through the following example: <dCore:subject xmlns:dCore="http://-purl.org/dc/elements/1.1/">Caffeine</dCore:subject> and <dc:subject xmlns:dc=-"http://purl.org/dc/elements/1.1/">Caffeine</dc:subject> are semantically identical.

Listing 6.1: RSS Code Generated for a CMLRSS Aggregation Feed (Shown Truncated).

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet href="http://www.w3.org/2000/08/w3c-synd/style.css"
                 type="text/css"?>
<rdf:RDF
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns="http://purl.org/rss/1.0/"
         xmlns:mn="http://usefulinc.com/rss/manifest/"
         xmlns:dc="http://purl.org/dc/elements/1.1/"
         xmlns:cml="http://www.xml-cml.org/schema/cml2/core">
  <channel rdf:about="http://">
    <title>Chemstock</title>
    <link>http://www.ch.ic.ac.uk/csdemo</link>
    <description>A chemical database based upon MySQL and PHP</description>
    <dc:publisher>Chemstock</dc:publisher>
    <dc:creator>rzepa@imperial.ac.uk</dc:creator>
    <image rdf:resource="http://www.ch.ic.ac.uk/logo.gif" />
    <items><rdf:Seq>
      <rdf:li rdf:resource="http://www.ch.ic.ac.uk/csdemo/?type=item&id=437" />
    </rdf:Seq></items>
  </channel>
  <image rdf:about="http://www.ch.ic.ac.uk/logo.gif">
    <title>ChemStock, Imperial College London</title>
    <url>http://www.ch.ic.ac.uk/logo.gif</url>
    <link>http://www.ch.ic.ac.uk/csdemo</link>
    <dc:description>A chemical database based upon MySQL and PHP</dc:description>
  </image>
  <item rdf:about="http://www.ch.ic.ac.uk/csdemo/index.php?type=item&id=437">
    <link>http://www.ch.ic.ac.uk/csdemo/index.php?type=item&id=437</link>
    <title>Caffeine</title>
    <description>Caffeine</description>
    <dc:subject>Caffeine</dc:subject>
    <dc:date>2003-09-16T00:00:00-00:00</dc:date>
    <dc:creator>demo</dc:creator>
    <cml:molecule xmlns:cml="http://www.xml-cml.org/schema/cml2/core"
       title="CAFFEINE">
      <cml:metadataList title="generated automatically from Openbabel">
        <cml:metadata name="dc:creator" content="OpenBabel version 1-100.1"/>
        <cml:metadata name="dc:description"
                      content="Conversion of legacy filetype to CML"/>
        <cml:metadata name="dc:content"/>
        <cml:metadata name="dc:rights" content="unknown"/>
        <cml:metadata name="dc:type" content="chemistry"/>
        <cml:metadata name="dc:contributor" content="unknown"/>
        <cml:metadata name="dc:creator" content="Openbabel V1-100.1"/>
        <cml:metadata name="dc:date" content="Wed Sep 17 15:03:57 bst 2003"/>
        <cml:metadata name="cmlm:structure" content="yes"/>
      </cml:metadataList>
      <cml:atomArray atomID="a1 a2 a3 a4 etc"
        elementType="C C N C O N C O N N C C C C H H H H H H H H H H"
        formalCharge="0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0"
        x3="0.310000 -0.355000 -1.534000  etc"
        y3="0.025000 -0.629000 -0.232000 etc"
        z3="0.266000 -0.740000 -1.149000 etc"/>
      <cml:bondArray atomRefs1="a1 a1 a1 a2 a2 a3 a3 a4 a4 etc"
        atomRefs2="a2 a7 a9 a3 a10 a4 etc"
        order="2 1 1 1 1 1 2 etc"/>
    </cml:molecule>
  </item>
</rdf:RDF>
```

6.2.2 Example 1. The ChemStock System

ChemStock, described fully elsewhere [13], is a simple data capture system designed using OpenSource software to maintain an inventory of molecules and selected properties. It is designed to answer questions such as "what are the latest chemical additions to the inventory?" or "what is the owner/location of a specific item?". At its heart is a user interface written using the PHP scripting system and populated by querying a MySQL database. Part of the data capture process involves contributors supplying molecular coordinate and atom connection table descriptors of a collection of molecules. Currently, two different formats for contributors to upload this information are implemented, the MDL Molfile format (which is not XML-compliant) and CML (Chemical Markup Language, which is XML based) [14]. Because the Molfile cannot be used within the XML-based RSS format, it is necessary to preprocess this legacy format into CML using the OpenBabel converter [15]. As XML conformance in the chemical community increases, the need for this particular step, itself susceptible to potential data loss, will eliminate. The CML expression of the molecule is then stored in the MySQL database, along with other information such as the creation/modification date, the name of the author, and a textbased descriptor (i.e. name) of the substance. Other molecular properties could also be computed at this stage if needed, most prominently the InChI unique molecular identifier [16].

To create a CMLRSS feed, the MySQL database must then be queried to retrieve the information and to format it according to the RSS 1.0 specification [11] using appropriate PHP tools [17]. The resulting RSS document is shown in Scheme 1.

The subscription URL takes the form `http://www.ch.ic.ac.uk/csdemo/feed.php?num=5`. This default query retrieves the last five entries added by users to the ChemStock database, including any CML components, along with metadata appropriate for the DC schema such as the author, date, and description. An example of the RSS generated is shown in Scheme 1. If used within a generic RSS news reader (which does not support the CML namespace) the results take the form shown in Figure 1. Note particularly that the CML components are not displayed (since no handler for these is present or has been specified) but that the Dublin Core (DC) fields are displayed, and these can be used to sort the aggregated display. Selecting the link associated with any individual entry will display the ChemStock page.

6.2.3 Example 2. The Dutch Dictionary on Organic Chemistry

The Dutch Dictionary on Organic Chemistry (WOC, "Woordenboek Organische Chemie" in Dutch) is an 8-year old Web site about organic chemistry and mainly in Dutch [18]. It contains descriptions of terminology, named reactions, and compounds. The 10 most recently changed items in the dictionary have been made available as a CMLRSS feed at `http://www.woc.sci.kun.nl/cgi-bin/rssfeed.rss`. The content is similar to that of the ChemStock RSS feed, including CML metadata for molecular content. Though not available at this moment, it is planned that the named reactions will be available using

the CMLReact namespace.

6.2.4 Example 3. The World Wide Molecular Matrix

The molecular matrix [19] is a bold and innovative attempt to create a global open repository of molecular information and associated properties using a grid-based peerto-peer model for collaboration and dissemination. The adoption of XML syntax throughout ensures that the diverse molecular information held in the matrix can be aggregated in a fully extensible and interoperable manner and that it is exposed to other chemical and nonchemical disciplines that may wish to access it in a semantically rich manner. Another pervasive concept is that of adding value to existing information ("accretion"). The Cambridge node in the WWMM for example can process molecular information contributed by users and add to it via e.g. a full MOPAC based [20] quantum mechanical computation of selected molecular properties. Other nodes on such a grid would compute other properties. The matrix gains content from the contributing user, the latter gains a valuable property calculation, and the community gains from patterns that may emerge from the aggregation of this on a large scale. Within such an environment therefore, it becomes valuable to readily identify new entries, or newly computed properties for existing entries, and to filter and sort these according to specified criteria. CMLRSS provides a mechanism for achieving this. The RSS feed `http://wwmm.ch.cam.ac.uk/Bob/rss` contains the appropriate chemical metadata for selected entries to the WWMM which could form the basis for further interactions with the matrix.

although clearly a much wider range of computed properties could be included either via modular functionality of the program itself or a call to an appropriate Web service. If several CMLRSS feeds are defined in the Jmol or JChemPaint properties file, the aggregated molecule entries can sorted by the various fields such as title, date, or formula. An example of filtering the content by atom type is shown in Figure 6.4. More complex calls to the CDK toolkit could in principle provide other sorting mechanisms, such as by e.g. chemical substructure. If the molecule is associated with a published journal article, then appropriate PRISM-based metadata can link to this information.

6.3 Chemical postprocessing and aggregation of RSS metadata

RSS functionality is not limited to generic viewers but can also be incorporated into chemical application software. This has been done for the opensource programs Jmol [21] and JChemPaint [22] via a plugin module interface written as part of the CDK (Chemistry Development Kit) [23, 24]. This functionalized RSS reader is then rendered capable of parsing the XML and extracting both the DC and e.g. the CML namespaced components for display. If atom coordinates (of various dimensionality) are present in the CMLRSS feed, these are extracted and displayed within the Jmol (Figure 2) or JChemPaint (Figure

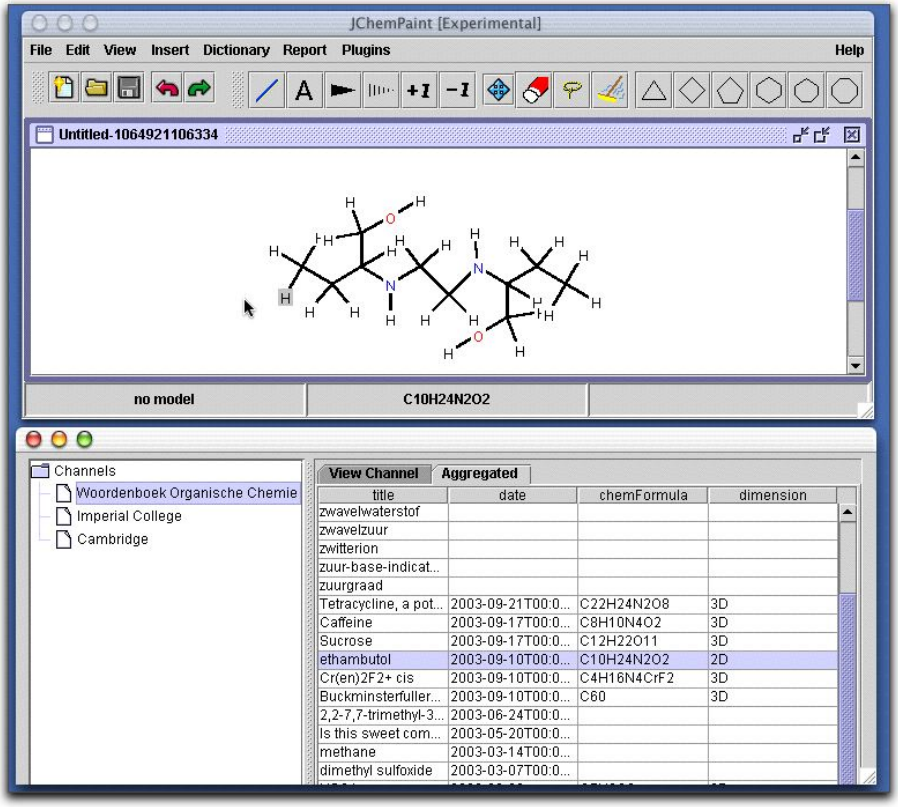


Figure 6.3: The JChemPaint 2D Molecule viewer/editor showing the RSS plugin window.

3) window when the item is selected. Additionally, a call is made to the CDK toolkit [24] to compute (in this example) the molecular formula for display,

6.4 Discussion and Conclusions

Scientific research generates large amounts of potentially valuable data. Most existing models for handling and disseminating such data adopt a variety of (often quite inadequate) approaches [3, 25]:

- The data are discarded at the completion of a project or archived on paper in a box file stored in a cupboard. The mere existence of such data is often forgotten.
- The data are converted (by scanning or other means) to a PDF (Acrobat) file and submitted as Supporting Information along with the associated scientific publication. This material may be made available on a publishers Web site but possibly only for a limited period. It is unlikely to be indexed in any manner by the publisher and

is therefore unlikely to be retrieved by any logical search procedures. Reuse of such data is only possible if a human (or good OCR system) rekeys it.

- Data are saved in original (non XML) formats and made available via a publishers (or authors) Web site in association with the article. It is potentially reusable by others if the particular formats (and any variations) are documented on the site or the prospective (human) user can "reverse engineer" the probable syntax and context and identify any software capable of handling it. These data too are unlikely to be indexed or searchable. Its existence is less likely to be advertised, and there may be many issues in its reuse, such as unambiguous knowledge of the particular scientific units used and ambiguity or multiple meanings for any terms describing the data.

- Some forms of data, particularly those resulting from X-ray crystallography, may be submitted by the publisher to an agency or specialist for added-value processing such as validation and deposition into a formal database for subsequent searching and retrieval. Most such data are currently not "open" and hence available without a commercial license. 5. Ultimately, only a small proportion of scientific data will reach recognized definitive repositories such as Chemical Abstracts or Beilstein; here again its reuse is restricted to those who can afford it, and again proprietory software may be needed to process it.

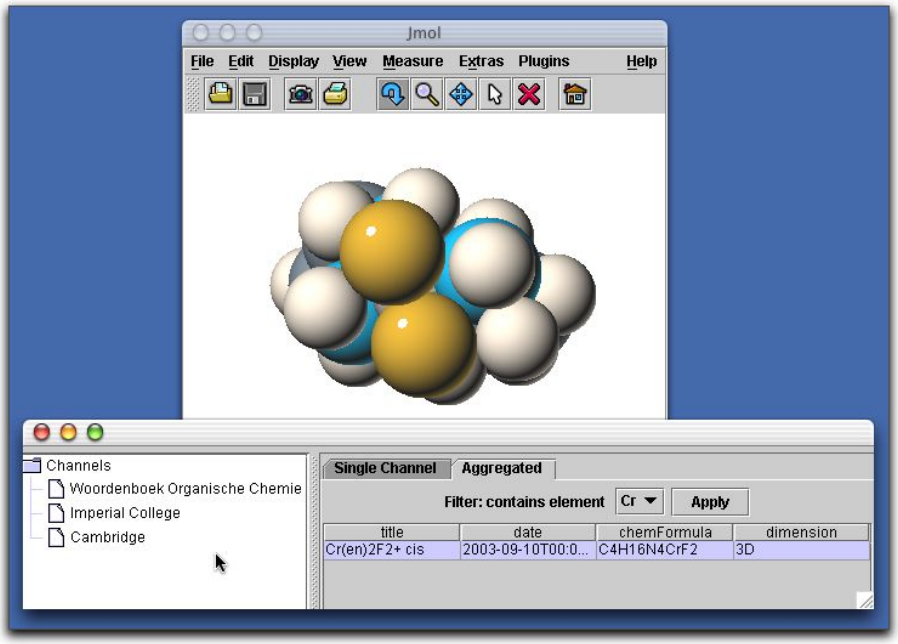


Figure 6.4: The Jmol 3D Molecule viewer showing the RSS plugin window with selection by atom type (in this example the element Cr).

It is quite likely that metadata about any stage in the above processes is also likely to be missing. There are no mechanisms for even identifying the existence of any data in categories 1 and 2, while metadata in category 3 may be restricted to information such as the date of deposition, the authors, and just possibly a hint (recognized only by a human) that it may include more specific information such as molecular connectivity information or molecular coordinates and their dimensionality (summarized as 1D, 2D, 3D, 5D (2D+3D), 3D/fractional, etc.). The provenance of the data may also be uncertain or inferred only by implicit association with (separately located) journal articles. The situation is not quite so dire in category 4, but even here it has been estimated that less than half of all determined crystal structures are actually deposited in any retrievable form. In addition, access to metadata (i.e. the existence of a molecular structure) is again restricted to those who have purchased access licenses and are using the dedicated software so provided. These data cannot easily be reconciled with other data about perhaps the same molecule (or indeed distinguished from data for an isomeric molecule) or with data resident in journal articles, etc. Some of the missing connections between the data and its provenance probably exist in collections in category 5 but again in a proprietary manner. Thus Chemical Abstracts and Beilstein can be seen as competitors and hence would have no (perceived) business case for providing mutual metadata about each others holdings. Each of these databases holds extensive metadata about chemical substances, including for example enumeration of property lists (plists) for each substance. These plists however may not always overlap or inter-relate, and hence aggregation of the data is not possible.

Against this background, we introduce CMLRSS as an XML-based RSS carrier for chemical metadata. The three implementations described above allow the discrete capture of several types of metadata. Simple information such as date, author, text descriptions, etc. can be contained with the base DC schema. More finely grained chemical metadata (one could of course argue this to be an oxymoron!) is carried using the CMLCore schema. This enables capture of information such as the type of molecular coordinate available and also allows derived metadata such as a molecular formula to be algorithmically computed on the fly. The CMLRSS feeds described above include sufficient molecular information to allow humans (or software) to decide if the data is appropriate for the purpose they had in mind, including the possibility of filtering/sorting the information not just by e.g. molecular formula but also by substructure content or unique (InChI) identifier

The adoption of a unifying XML-based syntax has other particular benefits. We have illustrated this by writing a CMLRSS parser using standard XML-compliant components which can be easily incorporated into a (2D) chemical editor (e.g. JChemPaint) and a (3D) molecular viewer (Jmol). This enables the user of either tool to subscribe to the appropriate CMLRSS feeds and hence to sort, select, and reuse the molecular data. The selection could be either by a human according to their own perceptions, or by software tools alone, but prearmed with particular chemical criteria. The current RSS plugin allows for filtering out news items that do not contain a specific element. By enfranchising chemical data sources of the types outlined above (particularly categories 1-4) with CMLRSS feeds, many of the problematic issues discussed above can be reduced if not entirely eliminated. This raises the intriguing prospect of what the role of aggregators such

as 4 and 5 should indeed be. A case could indeed be made that the greater availability of interchangeable data and metadata from such sources, routinely accessible from standard chemical software, would greatly improve their business models. Certainly this prospect would move the chemical community a great deal closer to the realization of the chemical semantic Web.

Bibliography

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 28–37, May 2001.

[2] P. Murray-Rust and H.S. Rzepa. Scientific publications in XML - towards a global knowledge base. *Data Sci.*, 1:84–98, 2002.

[3] P. Murray-Rust and H.S. Rzepa. The next big thing: From hypermedia to datuments. *J.Digital Information*, 5:248, 2004.

[4] H.S. Rzepa, P. Murray-Rust, and B.J. Whitaker. The application of chemical multipurpose internet mail extensions (Chemical MIME) internet standards to electronic mail and world wide web information exchange. *J.Chem.Inf.Comput.Sci.*, 38:976–982, 1998.

[5] R.S. Rzepa, J. Snyder, and C. Leach, editors. *Proceedings of the Electronic Conference on Trends in Heterocyclic Chemistry (ECHET96).* The Royal Society of Chemistry, 1997. ISBN 0-85404-894-4, CD ROM Version. See also `http://www.ch.ic.ac.uk/ectoc/`.

[6] The term intertwingling was coined for this process by Ted Nelson in his own far sighted vision encapsulated in the Xanadu project, now re-incarnated as Zig-Zag: `http://xanadu.com/zigzag/`.

[7] T. Hammond. Why choose RSS 1.0? `http://www.xml.com/pub/a/2003/07/23/rssone.html`, July 2003.

[8] PRISM: Publishing Requirements for Industry Standard Metadata. `http://www.prismstandard.org/`.

[9] P. Murray-Rust and H.S. Rzepa. Towards the chemical semantic web. An introduction to RSS. *Internet J. Chem.*, 6:article 4, 2003.

[10] T. Bray, D. Hollander, A. Layman, and R. Tobin. Namespaces in XML 1.0 (Second Edition). `http://www.w3.org/TR/REC-xml-names/`, August 2006.

[11] An XHTML profile for RDF Site Summaries. `http://www.w3.org/2000/08/w3c-synd/`, 2000.

[12] For an RSS validation service, see `http://feedvalidator.org/`.

[13] H.S. Rzepa and M.J. Williamson. Chemstock: A web-based chemical inventory system built from opensource software components. *Internet J. Chem.*, page article 6, 2002.

[14] Peter Murray-Rust and Henry 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.

[15] The OpenBabel Chemical File Format Conversion Package. `http://openbabel.sourceforge.net/`, 2005.

[16] S.E. Stein, S.R. Heller, and D. Tchekhovski. An open standard for chemical structure representation - The IUPAC Chemical Identier. In *Nimes International Chemical Information Conference Proceedings*, pages 131–143, 2003.

[17] For information about RSSWriter see `http://usefulinc.com/rss/rsswriter/`.

[18] E.L. Willighagen, G. Josten, and M. Fleuren. Woordenboek Organische Chemie. `http://www.woc.science.ru.nl/`, 1995.

[19] World-Wide Molecular Matrix (WWMM). `http://wwmm.ch.cam.ac.uk/`.

[20] J.J.P. Stewart. MOPAC: a semiempirical molecular orbital program. *J. Comp. Aided Mol. Design*, 4:1–105, 1990. See also `http://www.cachesoftware.com/mopac/`.

[21] D. Gezelter, B. Smidth, E.L. Willighagen, M. Howard, and Hanson. B. The Jmol 3D Molecular Visualization Software. `http://www.jmol.org/`, January 2005.

[22] S. Krause, E.L. Willighagen, and C. Steinbeck. JChemPaint - using the collaborative forces of the internet to develop a free editor for 2D chemical structures. *Molecules*, 5:93–98, 2000.

[23] E.L. Willighagen. Processing CML conventions in java. *Internet Journal of Chemistry*, 4, 2001.

[24] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 42(2):493–500, 2003.

[25] H.S. Rzepa and P. Murray-Rust. A new publishing paradigm: STM articles as part of the semantic web. *Learned Publishing*, 14:177, 2001.

Chapter 7

The Blue Obelisk-interoperability in chemical informatics.[1]

The Blue Obelisk Movement (`http://www.blueobelisk.org/`) is the name used by a diverse Internet group promoting re-usable chemistry via open source software development, consistent and complimentary chemoinformatics research, open data, and open standards has formed. We outline recent examples of cooperation in the Blue Obelisk group: a shared dictionary of algorithms and implementations in chemoinformatics algorithms drawing from our various software projects, a shared repository of chemoinformatics data including elemental properties, atomic radii, isotopes, atom typing rules, etc., and web services for platform-independent use of chemoinformatics programs.

[1] R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen, J.Chem.Inf.Model., 2006, 46(3):991–998

7.1 Introduction

While the past 20 or 30 years of development in chemoinformatics has created a plethora of published software systems and algorithms for solving chemical problems, little effort has been spent in providing the community with open components and data, to be reused and improved by communal efforts. Bioinformatics, with its much younger history, adopted the principles taught by success stories of the open source movement in general, and Linux in particular, from the very beginning. Recent years, however, have seen the emergence of open tools and databases also in chemical informatics [1, 2, 3, 4]. These draw on the existing ideas of independent peer review and scientific collaboration, mixed with "open source" software development paradigms. Community involvement, including assessments, suggestions, critiques, and rapid evolution is a core component of these efforts. The benefits of open source software have been discussed in great detail by Eric Raymond in his seminal paper "The Cathedral and the Bazaar" and following works [5]. The Open Source Initiative OSI summarizes: "Open source promotes software reliability and quality by supporting independent peer review and rapid evolution of source code. To be OSI certified, the software must be distributed under a license that guarantees the right to read, redistribute, modify, and use the software freely" [6].

In the beginning, most scientific software *was* free. It was so difficult to port that scientists didn't bother about licenses – one was delighted if someone else could get it working on another machine. But the 1980's saw the value of chemical informatics and the need to "productize" it. Much of this was meritorious, as it brought informatics into the classroom and the research lab, and helped pay for some chemistry research, but it also had hidden costs, which we are now facing today. In particular, costs include non-interoperability and centralized control of informatics.

Now, several open chemistry and chemoinformatics projects (Table 7.1) have pooled forces to enhance interoperability between these tools in a movement we call "The Blue Obelisk" (BO). The name originates from an informal meeting place in San Diego, California, during the American Chemical Society 2005 Spring National Meeting (see Figure 7.1) and was coined by one of the authors. Since contributors to the component projects live around the world, few had met in person – instead collaborating and meeting via the Internet.

We identify three core areas for the Blue Obelisk Movement:

- Open Source. One can use other people's code without further permission, including changing it for one's own use and distributing it again.

- Open Standards. One can find visible community mechanisms for protocols and communicating information. The mechanisms for creating and maintaining these standards cover a wide spectrum of human organizations, including various degrees of consent. We have been heavily influenced by the mantra of the IETF (Internet Engineering Task Force): "rough consensus and running code."

Figure 7.1: Where it all began. The Blue Obelisk in San Diego, California, at the 2005 American Chemical Society meeting.

- Open Data. One can obtain all data in the public domain when wanted and re-use it for whatever purpose. This is an under-used term which we are resurrecting. It is independent of "Open Access" and has relevance to "closed access" as well.

As outlined above, these areas are independent of the concept of "open access" to read publications freely. Instead, the three points focus on access to the scientific data, algorithms, and implementations themselves, rather than the formatted manuscript. In particular, we believe that these concepts strongly continue the spirit of communal peer review and reproducibility at the heart of modern scientific research.

It is well known in software development that 80 percent of the costs are caused by maintaining software and not by the initial implementation [7]. This holds both for the in-house development in pharmaceutical companies and the development for commercial chemoinformatics suppliers. Besides judging software by its standardized functional quality, it can be also compared on the basis of its long term stability and interoperability. Openly standardized algorithms and chemical information can help to reduce the maintenance costs, because developers can reuse available modules or test their tools against open source software and open data. This reduces the risk for both the "buy" and "build" strategies for software implementation. We agree with De Lano [8], that the try-before-buy paradigm for open source software does not necessarily require open standards. Open specifications for standard algorithms like kekulization [9], chirality coding [10], and atom

Project	URL	Principal Authors
CML, JUMBO [12]	http://cml.sf.net/	PMR, RZ
JChemPaint [13]	http://jchempaint.sf.net/	CS, EW
Jmol	http://jmol.sf.net/	MH, EW
NMRShiftDB [4]	http://www.nmrshiftdb.org/	CS
JOElib	http://joelib.sf.net/	JW
Kalzium	http://edu.kde.org/kalzium/	Carsten Niehaus
Octet	http://octet.sf.net/	Rich Apodaca
Open Babel	http://openbabel.sf.net/	GH
QSAR	http://qsar.sf.net/	EW, RG, CS, JW
The Chemistry Development Kit [3]	http://cdk.sf.net/	EW, CS
WWMM	http://wwmm.sf.net/	PMR

Table 7.1: Current Blue Obelisk Projects

typing [11], however, are indispensable in academic chemoinformatics research to build better, more stable, and more reproducible chemical information systems.

In this contribution, we outline several examples for how the Blue Obelisk projects address this need: a shared dictionary of algorithms and implementations in chemoinformatics algorithms drawing from our various software projects and a shared repository of chemoinformatics data including elemental properties, atomic radii, isotopes, atom typing rules, a set of web-based chemoinformatics services, and the process of providing open algorithms and data. All of these projects were developed with continual community involvement, an open standardization process, and provide open data to key chemoinformatics processes. Anyone can take part, we welcome those in commercial organizations, academia, government, etc., and contributions come as code, compilations of data and molecules, testing, and more.

7.2 The Importance of Open Specifications for Algorithms and Data

The World Wide Web as it is used today is a collection of linked HTML pages and other data formats. Whenever there is chemical or other scientific knowledge or data published via this mechanism, it is often difficult or impossible to discover, because it lacks the semantics that would help machines – the only practical way to harvest information "from the Internet" – to identify and classify it. Recognizing this lack, Tim Berners-Lee introduced the concept he termed the "Semantic Web." The Semantic Web is a mesh of information linked up in such a way as to be easily processable by machines, on a global scale. One can think of it as being an efficient way of representing data on the World Wide Web, or as a globally linked database. An analogy of the Semantic Web, projected onto the currently heavily researched idea of creating global networks of computational resources, so-called Grids, are the Semantic Grids. A Semantic Web, and even more a

Semantic Grid, are predicated on the supply of information and services without requiring the user to know the details of *how* the resource was obtained. The "users," who may be humans or robots, request precise services but should be unconcerned exactly how or where they originate. For example the calculation of a molecular property might depend on a precise method, but should not, in principle, depend on the actual program used, its version, the operating system and the machine involved.

We note that many chemical calculations are described in an imprecise manner. For example "molecular weight" is an imprecise term and the result of an algorithm returning this cannot be regarded as precise. The IUPAC Gold Book [14] describes

> *relative molecular mass*, M_r: Ratio of the mass of a molecule to the unified atomic mass unit. Sometimes called the molecular weight or relative molar mass.
>
> *relative molar mass*: Molar mass divided by 1 g mol^{-1} (the latter is sometimes called the standard molar mass).

and

> *unified atomic mass unit*: Non-SI unit of mass (equal to the atomic mass constant), defined as one twelfth of the mass of a carbon-12 atom in its ground state and used to express masses of atomic particles, $u \approx 1.6605402(10) \times 10^{-27}$ kg.

but "molar mass" does not occur as a term. These appear to refer to the mass of a single molecule, not to the properties of a bulk sample. However atomic masses include the concept of "average" as in:

> *relative atomic mass (atomic weight)*, A_r: The ratio of the average mass of the atom to the unified atomic mass unit. See also standard atomic weight .

There are at least two algorithms that could be used to obtain the "molecular mass":

- sum the average masses of all the atoms in the molecule (the normal "molecular weight")
- sum the precise masses of the most frequent isotopes in the molecule (giving the "high resolution molecular mass"). Even this latter is imprecise as in mass spectroscopy it relates to ions, and presumably the mass of the ionizing electron(s) should be accounted for.

Moreover, the actual values of atomic weights varies between program systems. We have frequently observed variations in molecular weights between different authorities – often at the second decimal place.

Current practice does not constrain any of this. Many chemoinformatics and computational chemistry papers use data resources which are not available to reviewers and readers, and algorithms which are not portable or distributed. It is a matter of trust rather than verification whether such work is accepted by the community. We believe it is essential that computational chemistry is able to provide the basic scientific tenet of reproducibility – if a scientist repeats the work in an article they should be able to duplicate the result. This is simple in principle: computers should run reliably and if the same data are given to the same algorithm identical results should be obtained. However it is surprisingly difficult to assert that the "same" method is being used. Wirth [15] observed: that "Data Structures + Algorithms = Programs". We can amend this to "known validated data resources" + "known validated algorithms" = "validated web resources."

There is relatively little practice of public validation of data resources and certification of algorithms in the field of chemistry, but without this, a global chemical semantic web is difficult to implement. This article explores the basis for such interoperability and outlines a working proof-of-concept. We hope that in the long term appropriate bodies such as IUPAC and other learned societies might come to oversee this practice; until then the Blue Obelisk can be seen as an informal, neutral mechanism to which those interested in open semantics can contribute.

An interoperable chemical approach requires at least the following communally agreed components in its architecture (in no particular order):

- terminology
- datatyping
- extensible data structures
- conformance specification and tools
- links and references
- namespaces
- metadata for provenance and discoverability

Syntactic support for all of these is provided by Chemical Markup Language [16] and other XML namespaces (XHTML, MathML, etc.). This article is largely concerned with how the semantic containers for terminology, data and algorithms are populated. There is also an important need for machine-enforceable behavior, which may also benefit from inheritance mechanisms but is not discussed in this work.

Our design and practice is heavily influenced by the practice and specifications from the International Union of Crystallography (IUCr). For the last three decades the IUCr, through its Data Commission and other bodies, has actively developed communal practice for the interchange of data. One of us (PM-R) has been associated with the COMCIFs project for a decade. The Crystallographic Information File (CIF) is the latest design

of the IUCr's semantically rich data structures and fully described in this Journal and the recently published Volume G of the *Int. Tab.* The primary approach is through *dictionaries*, each of which can describe a subdomain (e.g., core, macromolecules, powder diffraction, publications, etc.) Any valid crystallographic data *must* conform to one or more dictionaries. The dictionaries are similarly constrained by a dictionary definition language (DDL) which is also recursively conformant.

The groundbreaking DDL and CIF specifications are the major vehicle for publications of crystallographic information, both textual and numeric. The community has developed software for validation and processing, though the full power of the DDL is only recently becoming realized. DDL and CIF predated XML by a decade and are almost isomorphic to XMLSchema (XSD) and XML in their architecture. CIF dictionaries traditionally describe the human-readable meaning of a term, together with its structure and constraints (cardinality, lexical form, numeric range, enumerations, etc.).

This architecture can reasonably be considered an ontology for the hard sciences. Since the semantics of crystallography have been well understood for many decades, much of the ontology, including the algorithms, can be "hard-coded." More recently, through the dREL specification, the IUCr has started to add machine enforceable semantics into their dictionaries Listing 7.1 shows a typical CIF dictionary entry using the starDDL approach (courtesy Prof S.R. Hall and Dr. N. Spadaccini). This specification is being actively considered by the IUCr's COMCIFs committee.

Much of this example is self-explanatory. *description.text* (within ; ... ;) is the human-readable meaning, where there are references to other dictionary items. *_type.container, _type.value* and *_units.code* correspond to *<scalar dataType="float" units="daltons">* in CML. The *enumeration.range* term describes a non-negative integer (e.g., *xsd:nonNegativeInteger* in XML Schema.) The main enhancement is the machine-readable semantics in the *method.* loop_*. In this loop, a piece of code, based on Python and extended in the dRel language describes the precise algorithm for the evaluation of the atomic mass of the cell. It defines a mass, initially zero and a list of *atom_types* in the data object (the CIF). The *atom_types* have sub-fields *number_in_cell* (provided by the author) and *atomic_mass* (from a lookup table provided by IUCr). The sum of the atomic masses of all atoms is returned *_cell.atomic_mass*, the id of the dictionary entry.

These dictionaries are now compilable and executable in a proof-of-concept system [17]. They are powerful enough to allow the complete calculation of many crystallographic quantities (e.g., structure factors from atomic sites and form factors). The code can be run directly as Python, in Java through Jython and compiled into other languages through the JJTree compiler compiler.

This type of approach has great benefits for chemistry. Many of the BO algorithms (e.g., hundreds of JUMBO [18] methods) are sufficiently simple to be documented as machine-enforceable semantics. The dictionary approach enforces communal semantics for objects (e.g., through Octet) – e.g., a molecule contains atoms and bonds which can provide dRel-like iterators.

Listing 7.1: An example of a CIF dictionary entry

```
save_cell.atomic_mass
    _definition.id              '_cell.atomic_mass'
    _definition.update           2000-11-03
    _description.text
;
     Atomic mass of the contents of the unit cell. This is calculated
     from the atom sites present in the ATOM_TYPE list, rather than
     the ATOM_SITE lists of atoms in the refined model.}
;
    _description.compact        'CellAtomicMass'
    _name.category_id            cell
    _name.attribute_id           atomic_mass
    _type.container              Single
    _type.value                  Real
    _enumeration.range           0.:
    _units.code                  daltons
    loop_
    _method.class
    _method.expression
    EVALUATION
;
    mass = 0.
    Loop t as atom_type  {
      mass += t.number_in_cell * t.atomic_mass
    }
    _cell.atomic_mass = mass
;
    save_
```

There may be concerns about using a procedural language rather than a functional one (e.g., Scheme or LISP). We believe that the approach above is easily implemented and can run in a wide range of environments. It has the benefit of synergy with code and systems developed in crystallography.

Note that the approach also contains precise identification of, and therefore retrieval of, algorithms. Thus *_cell.atomic_mass.EVALUATION* is a precise pointer to a defined algorithm. The BO approach is informed by this architecture, though the precise syntax and semantics use XML-based approaches rather than CIF.

7.3 The Blue Obelisk Dictionary

The Blue Obelisk Chemoinformatics Dictionary is our effort of defining a standard set of chemoinformatics algorithms [19]. If a software project implements one of these algorithms, they can refer to this dictionary. By using unique identifiers, the dictionary allows using Web search engines, like Google.com, to find implementations for an algorithm in the dictionary. A similar dictionary has been developed for QSAR descriptors previously [20].

7.3.1 The Dictionary

The dictionary uses the following technologies: Scientific, Technical and Medical Markup Language (STMML, `http://www.xml-cml.org/stmml/`) was used as a general container and Mathematical Markup Language (MathML) is used to contain mathematical formula. Likewise, Scalable Vector Graphics (SVG) could be used to add graphics to the dictionary, though this is currently not used. References are contained in BibTeXML, an extended markup language for managing bibliographies. The full source of the latest XML source for the dictionary can be retrieved from Ref. [21].

The XML document is accompanied by a XML Schema document that encompasses the used XML languages. This allows XML aware editors to syntactically validate the document and filter out syntax errors in either of the three XML languages.

Each entry in the dictionary has an associated identifier (id) which is unique throughout the XML document. Using XML namespace technologies a world wide unique identifier can be composed that uniquely points to the entry in the dictionary. For example, by defining a namespace *http://qsar.sourceforge.net/dicts/blue-obelisk* with a related prefix *blue-obelisk*, one can uniquely point to an entry describing a Kabsch algorithm to align two molecules (id=alignmentKabsch) [22], within this namespace by referring to *blue-obelisk:alignmentKabsch*. Listing 7.2 is an example of an entry currently used in the Blue Obelisk dictionaries.

In this example an entry is defined for an algorithm that finds the smallest set of smallest rings, given a molecular graph. BibTeXML is used using the *bibtex* namespace prefix, to cite the article in which the algorithm was described. The entry has a bit of

Listing 7.2: An example of an XML dictionary entry

```
<entry id="findSmallestSetOfSmallestRings_Berger"
       term="Find Smallest Set of Smallest Rings (Berger Algorithm)">
  <annotation>
    <documentation>
      <metadata name="dc:contributor" content="elw"/>
      <metadata name="dc:date" content="2005-06-22"/>
    </documentation>
  </annotation>
  <definition>
    Algorithm to find the smallest set of smallest rings starting with a
    molecular graph <bibtex:cite ref="BGdV04a"/>.
  </definition>
  <metadataList dictRef="blue-obelisk-metadata:isClassifiedAs">
    <metadata dictRef="blue-obelisk-metadata:category"
              content="blue-obelisk-metadata:graph"/>
  </metadataList>
  <annotation>
    <documentation title="bibliography">
      <bibtex:entry id="BGdV04a">
        <bibtex:article>
          <bibtex:author>
            Berger, F. and Gritzmann, P. and De Vries, S.
          </bibtex:author>
          <bibtex:title>
            Minimum cycle bases for network graphs
          </bibtex:title>
          <bibtex:journal>Algorithmica</bibtex:journal>
          <bibtex:year>2004</bibtex:year>
          <bibtex:number>1</bibtex:number>
          <bibtex:pages>51-62</bibtex:pages>
        </bibtex:article>
      </bibtex:entry>
    </documentation>
  </annotation>
  <relatedEntry type="blue-obelisk-metadata:instanceOf"
                href="findSmallestSetOfSmallestRings"/>
</entry>
```

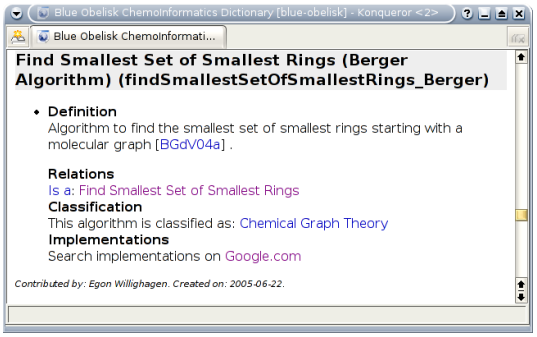


Figure 7.2: Screen shot of the XHTML output of the Blue Obelisk Chemoinformatics Dictionary showing the "Search implementations on Google.com" feature.

meta content using the Dublin Core standard, for which the namespace uses the prefix *dc*. Additionally, a classification is made (into the area of graph theory), and a related entry is mentioned.

XSLT (Extensible Stylesheet Language Transformation) is used to transform the XML source code into an XHTML document which can be displayed by a MathML aware web browser, like Mozilla Firefox.

7.3.2 Finding Implementations

The Blue Obelisk movement agreed on using the same namespace prefix, i.e. *blue-obelisk*, allowing web pages for specific software projects to cite entries in the dictionary. Links from those pages currently must be made explicitly, but having the citations on those pages allows web search engine to easily find software projects that implement a specific algorithm. The XHTML web page generated from the XML source of the dictionary, contains, for each entry, a link to Google.com that shows available implementations of that algorithm (see Figure 7.2). This setup provides a powerful tool to find software that implements published algorithms.

At the time of writing, CDK and Jmol each provide a web page that cites and links to individual Blue Obelisk Chemoinformatics Dictionary entries. The CDK page can be found at [23] and the Jmol page can be found at [24]. The Open Babel project has also included links to the dictionary in its developer documentation and is in the process of producing a complete index of entries as a separate webpage. All projects are continuing to add entries to the dictionary for common algorithms.

property type	property	sources
physical properties	isotope abundances	
	isotope masses	[31]
	atomic masses	[32]
	ionization energies	
chemical properties	affinities radii	[33]
	electronegativities	
	element densities	
discovery	year of discovery	
	name and etymology	
other	atom type definitions	
	2D and 3D coloring schemes	

Table 7.2: The current content of the data repository, with a few of the used sources.

7.4 The Blue Obelisk Repository

Since many chemoinformatics projects rely on accurate atomic and molecular data such as atomic masses, isotopes, electronegativities, van der Waals radii, covalent radii and so on, we have initiated a repository of a standard set of chemoinformatics data, building on the processes involved in the dictionary mentioned above.

Conventional standards bodies, such as IUPAC, have established a variety of published data, particularly on isotopes, atomic masses, elemental abundances, element symbols and names and so on. Many chemoinformatics algorithms, however, rely on other data which may not have a clear-cut definition. For example, there is no obvious way to specify a van der Waals radius – not all elements are perfectly spherical and multiple definitions exist including those taken from crystal structures, gas-phase measurements, and molecular mechanics force fields [25, 26, 27, 28, 29].

To address these issues, the Blue Obelisk Movement has established the Blue Obelisk Data Repository [30]. Software can use and refer to this repository when it needs standardized data for a wide range of chemical properties and other facts, of which an overview is given in Table 7.2. It is anticipated that over the next year the repository will considerably increase in the amount of available data.

The repository uses CML and dictionaries to allow explicit markup of data types, units, and the experimental errors, as well as metadata like bibliographic sources, creation dates and indications of authority. An example entry in the Blue Obelisk Data Repository is presented in Listing 7.3, and lists properties for hydrogen. For example, it states that the ionization energy is 13.5984 eV, and that the mass is 1.00794 a.m.u. It does not explicitly state which mass is meant, but refers for the definition to the Blue Obelisk Dictionary (see Section 7.3).

Listing 7.3: An example of a Blue Obelisk Data Repository entry

```
<elementType id="H">
  <scalar dataType="xsd:Integer" dictRef="bo:atomicNumber">1</scalar>
  <label dictRef="bo:symbol">H</label>
  <label dictRef="bo:name" xml:lang="en">Hydrogen</label>
  <scalar dataType="xsd:float" dictRef="bo:mass"
    unit="boUnits:amu" errorValue="7">1.00794</scalar>
  <scalar dataType="xsd:float" dictRef="bo:exactMass"
    unit="boUnits:amu">1.007825032</scalar>
  <scalar dataType="xsd:float" dictRef="bo:ionization"
    unit="boUnits:electronVolt">13.5984</scalar>
  <scalar dataType="xsd:float" dictRef="bo:electronAffinity"
    unit="boUnits:electronVolt" errorValue="3">0.75420375</scalar>
  <scalar dataType="xsd:float" dictRef="bo:electronegativityPauling"
    unit="boUnits:paulingScaleUnit">2.20</scalar>
  <scalar dataType="xsd:String" dictRef="bo:nameOrigin"
    xml:lang="en">Greek 'hydro' and 'gennao' for 'forms water'</scalar
  <scalar dataType="xsd:float" dictRef="bo:radiusCovalent"
    unit="boUnits:angstrom">0.37</scalar>
  <scalar dataType="xsd:float" dictRef="bo:radiusVDW"
    unit="boUnits:angstrom">1.2</scalar>
  <array title="color" dictRef="bo:elementColor"
    size="3" dataType="xsd:float">1.00 1.00 1.00</array>
  <scalar dataType="xsd:float" dictRef="bo:boilingpoint"
    unit="boUnits:kelvin">20.28</scalar>
  <scalar dataType="xsd:float" dictRef="bo:meltingpoint"
    unit="boUnits:kelvin">13.81</scalar>
  <scalar dataType="xsd:String"
    dictRef="bo:periodTableBlock">s</scalar>
  <scalar dataType="xsd:date"
    dictRef="bo:discoveryDate">1766</scalar>
  <scalar dataType="xsd:string"
    dictRef="bo:discoverers">C. Cavendish</scalar>
  <scalar dataType="xsd:int"
    dictRef="bo:period">1</scalar>
  <scalar dataType="xsd:int"
    dictRef="bo:acidicbehaviour">1</scalar>
  <scalar dataType="xsd:int"
    dictRef="bo:group">1</scalar>
  <scalar dataType="xsd:String"
    dictRef="bo:electronicConfiguration">1s1</scalar>
  <scalar dataType="xsd:String"
    dictRef="bo:family">Non-Metal</scalar>
</elementType>
```

7.5 Web Services

The preceding material has described how chemoinformatics data can be managed and accessed in a collaborative manner. Another aspect of collaboration is the use of distributed functionality. That is, the use of function implementations that are not necessarily on the local machine. An example of this type of approach is the use of web services. Though Web based applications are ubiquitous, they are generally full fledged applications that are monolithic in nature. The term web services refers to functionality that can be accessed over the Internet in a programmatic manner. In the context of chemoinformatics, this means that a programmer can access functions, that for example calculate binary fingerprints, over the Internet without having to understand what language the underlying function is written in or whether the function is up to date. Of course, this implies that the calling mechanism for the given function is well-defined and that the maintainer has kept it up-to-date. This approach is useful on a smaller scale, say at the organizational level. The advantage of having web based services implies that updates and modifications can be made on a single server, rather than requiring updates on individual machines.

We have used the CDK to provide web services for molecular similarity and descriptor calculations, available at `http://blue.chem.psu.edu/~rajarshi/code/java/cdkws.html`. Access to these services can be programmatic (using the SOAP [34] protocol) or by a web based interface which simply calls the service and presents the results. Since the algorithms are well documented and the calling mechanism is well defined, the service provides a relatively transparent method to obtain chemoinformatics functionality in a distributed manner.

The downside of web service functionality is that the user does not have control. This can be a problem if the service is not documented but at the same time it can be an advantage in that it relieves the user of the maintenance of yet another library. Furthermore, with the advent of Open Source and Open Data, a user is free to investigate the inner workings of a web service if he so wishes. This would allow the user to ensure that the a web service does indeed do what it advertises. Once again, this depends on the fact that the maintainer of the web service actually assigns an open license to the web service (in terms of access as well as code). Clearly, increased usage of web services is dependent on the transparency of the service. That is, a user must be able to ensure that a web service does indeed do what it says and should be able to rely on the provider of the service. We believe that the open principles underlying the Blue Obelisk movement are conducive to the development of transparent web services which provide easy access to a variety of functionality in a distributed manner.

7.6 Social Aspects

It has been mentioned previously that the Blue Obelisk movement is a communal effort. Given the three goals of the movement it is obvious why such an endeavour must be

a community effort rather than that of an individual. In this sense, the Blue Obelisk movement characterizes the nature of Open Source development in general and serves as an example of how this mode of development can be applied to problems in the field of chemical algorithms, standards and data. A striking feature of the Blue Obelisk movement is the wide variety of contributors to the individual projects that make up the movement. Contributors range from full professors to graduate students to commercial employees. The contributions themselves range from things as large as entire programs or frameworks to things as small as small amounts of data (e.g., to the data repository) or bug reports. However, it should be understood that though a bug report may appear to be a minor contribution compared to a whole framework, each contribution plays a vital role in the communal development and peer review of these projects.

At the same time it is important to realize that Open Source efforts represented by the Blue Obelisk movement do not always involve renumeration. Thus in many cases, the contributors work on the respective projects in their spare time. This leads to the situation where some areas in a project do not get as much attention as others, simply because it has not caught the attention of a contributor or due to lack of expertise amongst the contributors. In many cases, contributions to these projects are the result of a developer having "an itch" that needed to be "scratched." Thus compared to commercial projects, it may appear that the projects represented by the Blue Obelisk movement lack in certain areas. Given the open nature of these projects, it is a simple matter for anybody with the interest and expertise to contribute to such an area, thus filling the gap.

The above discussion paints a picture of many people contributing whatever they feel like. Naturally this would lead one to think of a chaotic development process. How is all this managed? This is an important question as the contributors to the Blue Obelisk projects are located all over the world. Furthermore most projects are large enough that a single person cannot always manage the contributions from a large user community.

The fundamental mechanism for distributed communal development are mailing lists, i.e., via email. Mailing lists are the mode by which the majority of decisions are made the community for a given project, both in terms of use and development. Decisions are made by consensus, although sometimes the "benevolent dictator" model of development is followed. Mailing lists also serve as archives of discussion, in addition to the use of traditional web pages and collaborative web pages (Wiki) for the development of documentation. A more real time mode of communication is the use of Internet Relay Chat (IRC) which allows multiple people to "convene" in a virtual room and communicate in real time. In general this is restricted to text, but current Instant Messaging (IM) services allow for the use of both audio and video based communication. This type of interaction is very fruitful, as contributors can discuss current problems and decisions in real time as they are working on the projects themselves.

These methods represent approaches to communication between the contributors. But how are the contributions (such as code or documents) themselves managed? Once again this is a very important question as multiple people will be working on a program or document and manually managing individual contributions does not scale for projects of

even moderate size. The workhorses for managing actual contributions are version control systems such as CVS or Subversion. These allow multiple contributors to submit changes to a program file or a document to a centrally located repository. If multiple contributors make changes to the same document, the system allows them to intelligently merge the resultant conflicts. These systems also allow developers to track changes and essentially view the "history" of a project. Workflows and web services can also be used in the development process and the utility of such types of applications have been mentioned previously.

Many of the Blue Obelisk projects make use of services provided by Sourceforge.net which is a community effort to provide Open Source projects with a set of tools and functionality for efficient code maintenance and communication. The site supports a number of features such as CVS, mailing lists, bug trackers and so on, all of which are freely available to Open Source projects.

Clearly, current Internet based technology allows for easy and efficient management of contributions to the various Blue Obelisk projects from contributors located all over the world. In a sentence, the Blue Obelisk movement is an example of the use of Open Source technology and methods to customize tools and social practices for the development of chemical information services.

7.7 Conclusion

We have described a communal effort to realize interoperability in chemical informatics, which we call the Blue Obelisk movement, named after the first meeting place of our community. The BO movement currently consists of more than ten open source and open data projects all related to chemoinformatics. We identify concepts and algorithms, codify them in a collaborative dictionary and link them to concrete implementations in Blue Obelisk projects and beyond to make those machine-searchable. We have started a public repository of chemical data of general interest, including data for chemical elements and isotopes, (boiling points, colors, electron affinities, masses, covalent radii, etc.), definitions of atom types, and more. All the data is augmented with documentation, citations of origin and bibliography. We are working on a system of web services to provide access to chemoinformatics functionality without the knowledge of the details of the individual implementation and without the need to master the installation and programming interface of yet another chemoinformatics library. We emphasize that this work in progress, which due to its emphasis on interoperability has a value beyond that of open source and open data efforts. While standardization efforts in chemistry have a long history, modern computing and data processing, the Internet and the World Wide Web have for the first time created the possibility of effortlessly searchable and reusable data and computer programs. Thus this article adresses the "old guard" of developers to contribute their wisdom and their work. The result can be the survival of a work of a lifetime which otherwise might not survive the emeritation or the next sale of the company. This article is also addressed to newcomers to adopt the ideas of open data and software from the very beginning. We

welcome those in commercial organizations. What is prized are contributions that help support the communal vision (e.g., Raymond [35]) Our approach is not incompatible with commercial systems, though the preservation of authorship moral rights is taken very seriously.

Bibliography

[1] The OpenBabel Chemical File Format Conversion Package. `http://openbabel.sourceforge.net/`, 2005.

[2] JOELib - a java based computational chemistry package. http://joelib.sourceforge.net/, Aug 2005.

[3] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 42(2):493–500, 2003.

[4] C. Steinbeck, S. Kuhn, and S. Krause. NMRShiftDB – constructing a chemical information system with open source components. *Journal of Chemical Information and Computer Sciences*, 43(6):1733 – 1739, 2003.

[5] Eric S. Raymond. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary.* O'Reilly and Associates, Sebastopol, California, 1999.

[6] The Open Source Initiative (OSI). `http://www.opensource.org/`, Aug 2002.

[7] D. C. Weaver. Build vs. Buy vs. Both. *Pharmaceutical Discovery*, pages 42–43, February 2005.

[8] W. L. DeLano. The case for open-source software in drug discovery. *Drug Discovery Today*, 10:213–217, 2005.

[9] A. Miličević, S. Nikolić, and N. Trinajstić. Coding and Ordering Kekulé Structures. *J. Chem. Inf. Comput. Sci.*, 44:415–421, 2004.

[10] J. Aires-de Sousa, J. Gasteiger, I. Gutman, and D. Vidović. Chirality Codes and Molecular Structure. *J. Chem. Inf. Comput. Sci.*, 44:831–836, 2004.

[11] P. Labute. On the Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.*, 45:215–221, 2005.

[12] P. Murray-Rust and H. S. Rzepa. Chemical markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *Journal of Chemical Information and Computer Sciences*, 41(5):1113–1123, 2001.

[13] S. Krause, E.L. Willighagen, and C. Steinbeck. JChemPaint - using the collaborative forces of the internet to develop a free editor for 2D chemical structures. *Molecules*, 5:93–98, 2000.

[14] A. D. McNaught and A. Wilkinson. Blackwell Science, Inc., Malden, MA, USA, 2nd edition, 1997.

[15] N. Wirth. *Algorithms + Data Structures = Programs.* Prentice Hall, Upper Saddle River, NJ, 1976.

[16] P. Murray-Rust and H.S. Rzepa. Chemical Markup XML, and the Worldwide Web. 1. Basic Principles. *Journal of Chemical Information and Computer Sciences*, 39:928–942, 1999.

[17] S. Hall, N. Spadaccini, D. du Boulay, and I. Castleden. Semantics for Scientific Data: Smart Dictionaries as Ontologies. `http://www.biomedchem.uwa.edu.au/our_people/homepages/hall/swwspaper`.

[18] Y. Zhang, P. Murray-Rust, M. Dove, R.C. Glen, H.S. Rzepa, J.A. Townsend, S. Tyrell, J. Wakelin, and E.L. Willighagen. JUMBO - an XML infrastructure for eScience. *Proceedings of UK e-Science All Hands Meeting 2004*, pages 930–933, 2004.

[19] C. Hoppe, P. Murray-Rust, C. Steinbeck, and E.L. Willighagen. Blue Obelisk ChemoInformatics Dictionary. `http://qsar.sourceforge.net/dicts/blue-obelisk/index.xhtml`.

[20] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E.L. Willighagen. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design*, 12(17):2111–2120, 2006.

[21] The Blue Obelisk Dictionary of Algorithms. `http://cvs.sourceforge.net/viewcvs.py/qsar/bo-dicts/blue-obelisk.xml?view=markup`.

[22] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, A32:922–923, 1976.

[23] CDK Reference to the The Blue Obelisk Dictionary of Algorithms. `http://wiki.jmol.org/JmolBlueObelisk`.

[24] Jmol Reference to the The Blue Obelisk Dictionary of Algorithms. `http://almost.cubic.uni-koeln.de/cdk/cdk_top/docu/dictrefindex/`.

[25] Y. V. Zefirov. Van der waals radii and current problems of their application. *Rus. J. Inorg. Chem.*, 46:568–572, 2001.

[26] S. S. Batsanov. Van der waals radii of elements. *Inorg. Mater.*, 37:871–885, 2001.

[27] N. L. Allinger, X. F. Zhou, and J. Bergsma. Molecular mechanics parameters. *Theochem. J. Mol. Struct.*, 118:69–83, 1994.

[28] A. Bondi. van der waals volumes and radii. *J. Phys. Chem.*, 68:441–451, 1964.

[29] L. Pauling. Cornell University, Ithaca, NY, USA, 3rd edition, 1960.

[30] G.R. Hutchison, P. Murray-Rust, C. Steinbeck, and E.L. Willighagen. Blue Obelisk Data Repository. `http://bodr.sf.net/`.

[31] A.H. Wapstra, G. Audi, and C. Thibault. The AME2003 atomic mass evaluation (I). Evaluation of input data, adjustment procedures. *Nuclear Physics*, A729:129, 2003.

[32] R.D. Loss. Atomic weights of the elements 2001 (IUPAC Technical Report). *Pure Appl. Chem.*, 75:1107–1122, 2003.

[33] T. Andersen, H.K. Haugen, and H. Hotop. Atomic weights of the elements 2001 (IUPAC Technical Report). *Journal of Physical and Chemical Reference Data*, 28:1511–1533, 1999.

[34] Simple Object Access Protocol. `http://www.w3.org/TR/soap/`.

[35] "Homesteading the Noosphere", article at Wikipedia. `http://en.wikipedia.org/wiki/Homesteading_the_Noosphere`.

Chapter 8

Discussion and Outlook

The main requirements for successful data analysis in molecular chemometrics are that results are properly validated and give the user new insights in the chemistry behind the data. A prerequisite to this is the appropriate formulation of the problem. This calls for suitable machine representations of the molecular systems, and statistical methods that are able to give meaningful feedback. For example, QSAR studies require molecular descriptors that contain information relevant to the biological activity, and modeling methods that indicate which molecular features are important in predicting the activity. This chapter discusses the principal problems in molecular chemometrics, and how they might be addressed.

8.1 Information Content

The information content of a representation determines what problems can be studied: while the chemical graph is suitable for exploring the molecular diversity of a set of potential drugs, 3D geometrical information is required to model and predict their binding affinities with the biological target. While such expert knowledge suggests a certain representation, for each application the question should be raised which molecular representation contains the most relevant information. Additionally, it must be studied if the representation allows the modeling method to use the information, which is not necessarily the case.

For example, it was recently proposed to use NMR and IR spectra to describe molecules in QSAR studies. These spectra contain information on the molecular connectivity and 3D geometries, and describe features relevant to physical properties like solubility and boiling point. PLS models have shown, however, unable to predict various properties from the ^{1}H NMR spectra of the molecules. It can be concluded that containing relevant information is not a sufficient requirement for yielding optimal models: the modeling method is not able to use the information experts use to deduce structural and geometrical properties of the molecule.

Capturing relevant information is indeed not straightforward, which becomes even more complicated when not just molecules are involved, but complex molecular systems such as organic crystal structures and enzymatic reactions. For multimolecular systems, the representation not only needs to describe molecular features, but also the interactions that define how the molecules aggregate. For example, while (some) packing features for molecular crystal structures can be represented by properties of the unit cell, this representation does not allow to quantify similarity: a small structural change can lead to large differences in the unit cell parameters. The relevant information is present in the description, but the analysis method is unable to use it.

Consequently, although this classic representation of crystal structures is suitable for crystallographic databases, it is inappropriate for statistical analysis. Instead, powder diffraction patterns or RDF-based descriptors can be used. These describe the same molecular information, only in a different form. The advantage of the RDF over the diffraction pattern is that it may be tuned to capture additional information relevant to intermolecular interactions. For example, a RDF might be defined with more focus on hydrogen bonding, potentially leading to better clustering according to expert knowledge. The diffraction pattern in itself, however, can directly be compared to experimental data and is, as such, easier to interpret.

8.2 Representation Characteristics

The representation of the molecular system strongly determines the outcome of the analysis. For example, PLS requires a fixed-length and numerical representation of the objects and properties of interest. Chemoinformatics has provided a plethora of numerical descriptors that fulfill this need. They may be derived from 3D geometry, chemical graph and otherwise, and represent certain molecular features and allow to describe more complex molecular systems.

Even though the number of available representations is huge, each has specific characteristics that may make statistical analysis challenging. For example, collinearity and a high variable-to-object ratio are two often encountered problems. Methods like MLR are known to become unstable when the representation contains a high number of collinear variables. With a growing number of variables, the chance of overfitting increases too, though internal validation can address this problem.

Specific statistical methods may have additional requirements. For example, kNN estimates the property of a new object by averaging those of the k nearest neighbors, and needs an appropriate similarity measure. A similarity is not always easily defined and strongly depends on characteristics of the representation. Peak-like spectra are interesting examples, where shifts need to be approached differently from one application to another. When searching for the molecular structure in a database of NMR spectra using a query spectrum, one should realize the peak shifts do not contain relevant information and should be neglected. This can be achieved by using a binning scheme or a more sophisticated

alignment procedure, after which a normal difference measure can be used.

On the other hand, in other applications these peak shifts can be used to express similarity. An example of this is the use of the weighted cross-correlation for comparing crystal structures, both for powder diffraction patterns and R_eDFs. Hierarchical clustering algorithms and self-organizing maps (SOMs) can use similarities to find groups of structures. The representation also affects how the similarity changes when the molecular system changes; it is preferred here that the similarity shows a smooth transition when going from highly similar systems, to more dissimilar systems. This smoothness, or lack thereof, affects how well statistical methods are able to reproduce expectations.

8.3 Validation

Chemoinformatics provides many methods to represent molecules and molecular systems, while chemometrics offers a wide range of statistical methods to process the resulting data. Statistical methods are even used to define better representations, such as MOLMAPs for representing enzymatic reactions, as explained in the introduction. Because the choice of both representation and modeling methods depends strongly on the combination of the two, validation is a very important tool to ensure that conclusions are scientifically sound.

Validation can be split up into two complementary types: one is statistical in nature and expresses model and analysis quality in numerical estimates, such as the root mean squared error of prediction; the other is expert validation, where the analysis is (visually) examined and the results are compared against expert knowledge. Both have advantages and disadvantages, which are discussed now.

Of these two methods, statistical validation provides the means to compare alternative models in an objective manner. The big advantage here is that the use of such statistics allows automated model building, as it does not require time-consuming human intervention: if the quality parameter is favorable for one model, then that one is chosen. This is used in modeling methods for parameter optimization, such as choosing the number the latent variable in PLS modeling, and selecting kernel parameters in kernel methods, such as SVM. However, the error margins on these statistics make it difficult to prefer one model over another if the difference is only small. For example, in QSAR studies the squared correlation coefficient between predicted and real activity is used; this statistic has a range of 0 to 1, and the error margin for small data sets (less than 100 molecules) might go up to 0.05. Consequently, one is unable to distinguish models with $R^2 = 0.85$ and $R^2 = 0.90$.

For situations like these, visual validation becomes important: where statistical validation is no longer able to make decisions, expert knowledge is needed as complement. While prior knowledge should be incorporated into the process as early as possible, capturing a sufficiently large and relevant knowledge base has shown to be difficult. Therefore, expert validation of models remains critical. This kind of validation can be as simple as plotting predicted versus real activities in QSAR studies, or plotting a clustering of ob-

jects using PCA or non-linear mapping methods. It is often also useful to visualize model features itself; for example, the regression vector in PLS indicates which variables are important for the prediction. This, for example, can be used with ^{13}C NMR to indicate which chemical shift ranges correlate with the modeled property.

SOMs are interesting as tools to give visual feedback. Objects can be mapped onto units of the map which visualizes their grouping into clusters. This can be matched against expert knowledge: objects mapped onto the same unit show a high similarity, while objects mapped onto neighboring units show similarity of lower but significant degree. The map units themselves have associated weight vectors which resemble the representations of the objects mapped onto that unit. These vectors can be explored and indicate which features in the representation are typical for the objects that map onto that unit. Supervised SOMs have additional possibilities and allow to visualize the relation between the cell volumes for crystal structure classes, by creating a SOM where the map units are colored according to the cell volume and where the objects are mapped according to their unit assignment and color-coded by the classification scheme.

An interesting feature of the SOM is not explored in this thesis: while the prediction of a property of interest is principally based on the property value associated with the winning unit, the map really returns an array of more (and less) likely property values. Consider mapping a crystal structure on one of the maps trained in Chapter 5 which has unit cell volume as property of interest: the structure will have similarities with all units; the unit with the highest similarity, the winning unit, has the most likely cell volume. However, a second unit may have an almost equally high similarity, but with a rather different cell volume. The map does not exclude the second option, and only an expert may be able to decide on the difference. Indeed, for the crystal structures maps it was observed that volumes were predicted which were twice or half to true volume, caused by occurrence of structurally similar crystal structures in the training data set with a different number of molecules in the unit cell.

This feature can be translated to other applications too, such as QSAR. The method may now be able distinguish two different classes of molecules, which show a high structural similarity, but different binding affinities too. Such situations can occur when the molecules exhibit different modes of action. In contrast to methods like PCR and PLS, a supervised SOM would learn to distinguish between both molecular classes, using the similarity based on both descriptor and property space. Like conventional screening methods, the SOM would still be able to predict the binding affinity, but now predict two likely affinities instead of one; the similarity of the test molecule with the units provide a likeliness for the predicted affinities. Although there is no way to decide which affinity is the true one, it is anticipated that virtual screening results may improve by making advantage of this feature.

8.4 Reproducibility

While the statistical and visual expert validation discussed above provides important ways to verify the quality of models, more is needed when validating the scientific conclusions that are drawn from the analysis. For example, the scope of a new representation or new method can only be explored if other scientists are able to apply it to their data too. This requires that both the representation and data analysis are reproducible. However, this is not straightforward and often not possible in current molecular chemoinformatics

Molecular descriptors are good examples where this reproducibility is not easily achieved. These theoretical descriptors are derived from principal molecular representations, like chemical graphs and 3D geometries, but often also include information on atomic features, like partial charges and atomic weights. It is not always well defined how these descriptors should be calculated, resulting in a situation where these descriptors cannot be recalculated using the same software package with the same version but using different libraries or platforms. Recalculation of those descriptors is, however, the only way a published QSAR and QSPR model can be validated independently.

Reproducibility demands several things from the chemoinformatics and chemometrics involved: primarily, clear descriptions of algorithm are needed to allow alternative implementations; and, identical data should be used in those algorithms to ensure identical numerical output. For example, the molecular weight descriptor requires the algorithm to specify which atomic weights are used, e.g. the atomic weight of the most abundant isotope or the weighted atomic mass according to natural occurrence of the isotopes, and what the values for those atomic/isotopic properties are.

Such reproducibility in descriptor calculation can, for example, be achieved by using ontologies or dictionaries that describes the exact algorithms and open standards of molecular properties like isotopic weights. Communal efforts are undertaken which aim to realize such interoperability in chemical informatics. One is called the Blue Obelisk Movement. Within this project common concepts and algorithms in the field of chemoinformatics are identified, and codified into a collaborative dictionary. Implementations can link to this dictionary to specify which algorithm is used. The project also provides a repository of standard atomic properties, to increase the reproducibility between implementations of descriptors which rely on such data. A partial solution to the reproducibility problem is the use of open source, open data and open standards, where the implementation is the exact specification. It is nevertheless acknowledged that the use of these concepts requires considerable change in the way we currently do science. The Blue Obelisk promotes adoption of these ideas.

8.5 Data Storage and Communication

Additionally, accurate storage and communication of molecular data is important when aiming at reproducible data analysis: changing the data set, such as the removal of a single

object, can significantly change the results of an analysis. The complexity and interaction between the representation and the analysis method is again an important factor.

One source of problems is the lack of data exchange in an information-rich manner. Listing the SMILES of the molecules in a QSAR data set is not sufficient: it does not contain information on the conformation used in the analysis, nor does it indicate which protonation state was used. Both are important when calculating binding affinities. The loss of information may be countered by assuming likely values, but this makes reproduction of the predictive models difficult. A second problem is the format itself. For example, exchanging NMR spectra as images in Word or PDF documents makes it rather difficult to extract the data needed to verify if the spectrum supports the suggested molecular structure. This complicates, for example, dereplication, where the spectrum is matched against a database to determine if the spectrum is already associated with a molecular structures.

Such processes would be possible if semantic document formats would have been used. Method that do allow semantically-rich distribution of data provide the tools to automate validation. For example, explicitly defining measurement units and conditions for NMR or IR spectra make it easier to validate the experimental results against expert knowledge. Additionally, adding more semantics also allows automated aggregation of data from different sources. The Chemical Markup Language (CML) is such an language aimed at molecular data, and supports a wide range of chemical data, including 3D molecular geometries and crystal structures, as well as experimental data such as 1D and 2D MS spectra.

A realistic application of this technology is to distill new patterns. For example, consider a crystallographic database, such as CrystalEye, which publishes new entries using CMLRSS. A preliminary statistical analysis of the database may result in normal bond lengths and geometries. These normal conditions can then be used to validate entries in the CMLRSS feed; an entry which has geometrical properties outside the range of expected values can be highlighted as *interesting* in a derived CMLRSS feed, suggesting manual inspection. Moreover, the feed may also be used to update the so-far assumed normal conditions, therefore covering a more diverse range of crystal structures.

This approach has many applications. For example, supervised SOMs for crystal structures could automatically be retrained with new types of crystal structures, if suitable. And QSAR, QSPR and virtual screening methods can automatically be validated against new data, highlighting new chemistry. These integrated data flows, where molecular data is streamed and processed in an automatic way, allow scientists to detect, and react sooner on, yet unexplained chemistry.

8.6 Outlook

The research shows how effective the combination of chemoinformatics and chemometrics can be in data analysis and data mining of chemical data. Several problems, typical for

either field, are encountered and addressed. The work extends and improves on earlier work, and offers a number of new methods to analyze chemical data. Whether it is the whole data aggregation, fusion, analysis and modeling process, model validation and reproducibility, or addressing complex molecular information, important steps have been made. To indicate what can be expected in the next decade, in light of the research described in this manuscript, the following sections will illustrate some new applications.

8.6.1 Crystal Engineering

The supervised self-organizing map was shown to have favorable data mining features, allowing visualization of relations between crystal structure and one or more crystal properties. Currently, clustering crystal structures according to packing patterns is an important step in understanding how molecular features give rise to properties of the crystal. Hydrogen bonding patterns have been used on several occasions, but often require manual examination. Predicting hydrogen bond patterns using a computational method allows the processing of much larger data sets, as the expert validation is now guided by the machine-learned patterns. This approach also allows the visualization of relations between this bonding pattern and other molecular and/or crystal properties. This provides a tool to assess hypotheses on much larger sets of crystal structures.

8.6.2 Data Fusion

New technologies like CMLRSS and standardization efforts like the Blue Obelisk will increase interoperability and automation of data processing, paving the way for online library searching, data mining and analysis. For example, it is anticipated that molecules and their associated properties will automatically be extracted from new publications, stored into databases, and analyzed against normal statistics on those properties. If different from those normal conditions, it may indicate either new chemistry, or false chemistry. Because these methods can be applied to the reviewing process too, as already done partially with publications of crystal structures, an improved scientific dissemination is expected. The RSC's Project Prospect already exemplifies this trend.

8.7 Conclusion

It is clear that molecular chemometrics shows a strong interaction between representation of objects and the methods used for data analysis. The representation does not just need to capture relevant information, it must also be compatible with the statistical methods used to analyze the data. The need for fixed-length representations is discussed, as well as the importance of understanding the data characteristics when choosing a similarity measure. Validation provides an important tool in determining the quality of the analysis. Statistical validation must be complemented with expert knowledge, putting emphasis on (visual) feedback.

List of Abbreviations

AM1 Austin Model 1

BO Blue Obelisk

CART Classification And Regression Trees

CAS Chemical Abstracts Service

CASD Computer-Assisted Synthesis Design

CASE Computer-Aided Structure Elucidation

CSD Cambridge Structural Database

CDK The Chemistry Development Kit

CD-ROM Compact Disc Read-Only Memory

CIF Crystallographic Information File

CML Chemical Markup Language

CMLRSS CML-enabled RSS

COMCIFS COmmittee for the Maintenance of the CIF Standard

CoMFA Comparative Molecular Field Analysis

CSS Cascading Style Sheets

CV Cross-Validation

CVS Concurrent Versioning System

DC Dublin Core

DDL Dictionary Definition Language

DENDRAL DENDRitic ALgorithm

DTD Document Type Definition

EC number Enzyme Commission numbering scheme

ECHET96 Electronic Conference on HETerocyclic chemistry 1996

ECTOC Electronic Conferences on Trends in Organic Chemistry

ESP ElectroStatic Potential

GA Genetic Algorithm

GNU GNU's Not Unix

HF Hatree-Fock

HIV Human Immunodeficiency Virus

HOMO Highest Occupied Molecular Orbital

HOSE Hierarchical Organisation of Spherical Environments

HTML HyperText Markup Language

HTTP HyperText Transfer Protocol

IETF Internet Engineering Task Force

IM Instant Messaging

InChI IUPAC International Chemical Identifier

IR InfraRed

IRC Internet Relay Chat

IUCr International Union of Crystallography

IUPAC International Union of Pure and Applied Chemistry

KEGG Kyoto Encyclopedia of Genes and Genomes

kNN k-Nearest Neighbors

LDA Linear Discriminant Analysis

LHASA Logic and Heuristics Applied to Synthetic Analysis

LISP LISt Processing

LMO Leave-more-out

LOO Leave-one-out

LUMO Lowest Unoccupied Molecular Orbital

LV Latent Variable

MACiE Mechanism, Annotation and Classification in Enzymes

MathML Mathematical Markup Language (MathML)

MCF Meta Content Framework

MIME Multipurpose Internet Mail Extensions

MLR MultiLinear Regression

MNDO Modified Neglect of Differential Overlap

MOLMAP MOLecular Map of Atom-level Properties

MOPAC Molecular Orbital PACkage

MS Mass Spectroscopy

NIR Near InfraRed

NMR Nuclear Magnetic Resonance

NN Neural Network

OCR Optical Character Recognition

OCSS Organic Chemical Simulation of Syntheses

OSI Open Source Initiative

PCA Principal Component Analysis

PDB RCSB Protein Data Bank

PDF Portable Document Format

PHP PHP: Hypertext Preprocessor

PI Processing Instruction

pKa The logarithm of the acid dissociation constant K_a

PLS Partial Least Squares

ppm parts per million

PRISM Publishing Requirements for Industry Standard Metadata

QDA Quadratic Distriminant Analysis

QSAR Quantitative Structure-Activity Relationship

QSPR Quantitative Atructure-Property Relationship

RBF Radial Basis Function

RCSB Research Collaboratory for Structural Bioinformatics

RDF 1) Radial Distribution Function 2) Resource Description Framework

RMSE Root Mean Square Error

RMSECV Root Mean Square Error of Cross Validation

RMSEP Root Mean Square Error of Prediction

RSC Royal Society of Chemistry

RSS RDF Site Summary (based on Resource Description Framework)

SI Système International

SIMCA Soft Independent Modeling of Class Analogy

SMILES Simplified Molecular Input Line Entry Specification

SOAP Simple Object Access Protocol

SOM Self-Organizing Map

SVG Scalable Vector Graphics

SVM Support Vector Machine

SVR Support Vector Regression

URI Uniform Resource Identifier

URL Uniform Resource Locator

URN Uniform Resource Name

W3C WWW Consortium

WCC Weighted Cross Correlation

WHIM Weighted Holistic Invariant Molecular (descriptor)

WLN Wiswesser Line Notation

WOC Woordenboek Organische Chemie (= Dictionary of Organic Chemistry)

WWMM World Wide Molecular Matrix

WWW World Wide Web

XHTML Extensible HyperText Markup Language

XML Extensible Markup Language

XSD XML Schema Definition

XSLT Extensible Stylesheet Language Transformation

XYF XY-Fusion

ZINC ZINC Is Not Commercial

Summary

Chemometrics and chemoinformatics play important roles in the analysis and modeling of molecular data. In particular, in understanding and prediction of properties of molecules and molecular systems. Both chemometrics and chemoinformatics apply statistics, machine learning and informatics methodologies to chemical questions, though originating from a different background. Where chemometrics had its origins in the extraction of information from chemical experiments, chemoinformatics had roots in the representation of chemical data for storage in databases. The technological advances in chemistry and biochemistry in the past decades have led, however, to a flood of data and new questions, and the data analysis and modeling have become more complex. The standing challenge in data analysis and data exchange, is how to represent the molecular features relevant to the problem at hand. This representation of molecular information is the topic of this thesis.

Chapter 1 introduces the field of data analysis and modeling of molecular data and describes the aforementioned importance of representation of relevant features. It discusses different approaches to molecular representation, such as line notations, chemical graphs, and quantum chemical models. Each of these have limitations when used in data analysis and modeling. Numerical representations are then introduced, which allow the application of statistical and mathematical modeling approaches. These numerical representations are commonly derived from chemical graph and quantum chemical representations. CoMFA and the classification of enzyme reactions are examples were the choice of molecular representation as well as the analysis method are important.

The term *molecular chemometrics* is coined in Chapter 2 for the field that applies statistical modeling methods to molecular structure. It reviews the advances made in this field in recent years. New numerical descriptors for molecules are discussed, as well as approaches to represent molecules in more complex systems like crystal structures and reactions. Molecular descriptors are used in similarity and diversity analysis. The applications of new methods for structure-activity and structure-property modeling, and dimension reduction are described. An overview of recent approaches in model validation show new insights and approaches to estimate the performance of classification and regression models. The last section of this chapter lists new databases and introduces new methods that improve the extracting of chemical data from database and repositories. Semantic markup languages improve the exchange of data, and new methods have been

introduced to extract molecular properties from text documents.

Chapter 3 studies the in literature proposed use of 1D ^{13}C and ^{1}H NMR spectra as molecular descriptor. These spectra are known to describe features relevant to physical properties like solubility and boiling point. The NMR representation is studied for the predictive powers of its PLS models for three structure-property data sets. The results indicate that proton NMR is not suitable for building QSPR models in combination with PLS. Carbon NMR-based models, however, do give reasonable QSPR models, and the regression vectors for the carbon NMR data, correlate with spectral regions relevant to molecular fragments. Nevertheless, the predictive power of the carbon NMR-based spectra is still less than models based on common molecular descriptors. It is concluded that NMR spectra should not be considered first choice when making predictive models in general, and that proton NMR should probably not be used at all.

A computational method to calculate similarities between crystal structures based on a new representation is introduced in Chapter 4. While a reference method is perfectly able to identify structures with high similarity, it fails to recognize the different similarities between two similar structures and two completely different structures. This makes it very difficult for clustering algorithm to organize small clusters of identical and highly similar structures into larger clusters. The new representation of crystal structures introduced in this chapter shows a much smoother transition in similarity values when crystal structures go from identical, via similar, and finally to dissimilar structures. Clustering a set of simulated polymorphic structures of estrone, and classification of a set of experimental cephalosporin structures reproduce expected clustering and classification.

Chapter 5 uses supervised self-organizing maps to cluster crystal structures represented by their powder diffraction pattern and one or more properties. The topological structure of the resulting maps not only depends on the similarity of the diffraction data, but also on the properties of interest, such as cell volume, space group, and lattice energy. This approach is used to analyze and visualize large sets of crystal structures, and the results show that these supervised maps not only give a better mapping, they can also be used to predict crystal properties based on the diffraction patterns, and for subset selection in polymorph prediction. The two applications in crystallography show that suitable representations and similarity measures that allow data analysis and modeling of molecular crystal data are now available. Both approaches are flexible enough to open up a new field of research; especially combinations with other classification schemes for crystal structures, such as those based on hydrogen bonding patterns, come to mind.

Chapter 6 introduces and discusses a method that allows information rich distribution of molecular data between machines, such as measuring devices and computers. Existing approaches often imply not or badly documented semantics which may lead to information loss. CMLRSS is proposed and combines two existing web standards: Rich Site Summaries (RSS), also known as RDF Site Summaries, and the Chemical Markup Language (CML). Here, RSS is used as transport layer, while CML is used to contain the chemical information. CML supports a wide range of chemical data, including molecular (crystal) structures, reaction schemes, and experimental data such as NMR spectra. It is

shown that this semantic representation allows automated dissemination of chemical data, and is increasingly used to exchange data between web resources.

Chapter 7 describes a communal effort to realize interoperability in chemical informatics, which is called the Blue Obelisk movement. This movement currently consists of more than ten smaller and larger, open source and open data projects all related to chemoinformatics and chemistry in general. To increase the reproducibility of molecular representations, this chapter introduces a collaborative dictionary of chemoinformatics algorithms, and a public repository of chemical data of general interest, including data for chemical elements and isotopes, (boiling points, colors, electron affinities, masses, covalent radii, etc.), definitions of atom types, and more. The availability of a standard set of atomic properties, open source algorithms and open data (for example via CMLRSS feeds), it is much easier to reproduce and validate published results in molecular chemometrics. Results from Chapter 3 show that such ability is no luxury.

The last chapter summarizes the efforts in this thesis and how they address the challenges in molecular chemometrics. This thesis shows the strong interaction between representation and the methods used for data analysis: molecular representation need to capture relevant information and be compatible with the statistical methods used to analyze the data. The chapters review molecular representations and put focus on model validation using statistics, visualization methods, and standardization approaches.

Samenvatting

Chemometrie en chemoinformatica hebben belangrijke rollen bij de analyse en het modelleren van moleculaire data. Met name bij het begrijpen en voorspellen van eigenschappen van moleculen en moleculaire systemen. Zowel chemometrie als ook chemoinformatica gebruiken statistiek, wiskundige en informatietechnologisch methoden om chemische vraagstukken op te lossen; maar beide hebben een verschillende achtergrond. Chemometrie, bijvoorbeeld, komt voort uit de tak van wetenschap die informatie haalt uit natchemische experimenten, terwijl chemoinformatica begonnen is als wetenschap zie zich bezig houdt met het representeren van chemische eigenschappen en structuren in databases.

Echter, de technologische ontwikkelingen in de chemie en biochemie in de afgelopen twintig tot dertig jaar hebben geleid tot een tsunami aan nieuwe meetmethoden die elke steeds meer data geeft en daarmee ook nieuwe vragen, terwijl ook de analyse van die data steeds complexer wordt. Het is nog steeds een uitdaging in de data analyse en data uitwisseling om de voor een probleem meest relevante informatie accuraat te weer te geven. De representatie van moleculaire informatie is het onderwerp van deze studie.

Hoofdstuk 1 introduceert het wetenschapsveld van data analyse en het maken van modellen voor het beschrijven van moleculaire data. Het beschrijft het belang van een goede representatie van relevante aspecten en geeft een overzicht van de verschillende methoden om moleculaire structuren te representeren: de lijnnotatie, chemische graven, en de quantumchemische modellen. Elke kent zijn specifieke beperkingen bij het gebruik in data analyse en modelleren. Numerieke representaties worden daarna geïntroduceerd als benadering die het mogelijk maakt veelgebruikte statistische en wiskundige methoden te gebruiken met het beschrijven van moleculen en moleculaire systemen. Deze representaties zijn normaliter afgeleid van de chemische graaf en quantumchemische representatie. Het maken van een juiste keuze in de representatie bepaalt het succes van de analyse zoals gellustreerd met de toepassingen CoMFA en de classificatie van enzymereacties.

De term *moleculaire chemometrie* wordt in hoofdstuk 2 geïntroduceerd als naam voor het wetenschapsveld dat statistische methoden toepast op moleculaire structuren. Het geeft een overzicht van de ontwikkelingen in dit veld in de afgelopen jaren. Nieuwe numerieke descriptoren voor moleculen worden beschreven, maar ook hoe moleculen in een kristal of andere complexe systemen beschreven kan worden. Deze moleculaire descriptoren worden gebruikt in similariteits- en diversiteitsstudies, en vinden hun toepassing in het modelleren van structuur-activiteits en structuur-eigenschaps relaties en dimensie

reductie. Het hoofdstuk geeft ook een overzicht in recente ontwikkelingen op het gebied van modelvalidatie, zoals nieuwe inzichten en methoden om de voorspellingskracht van regressie en classificatiemodellen te beschrijven.

Moleculaire descriptoren worden gebruik bij de vergelijken van moleculen, bijvoorbeeld om hun gelijkenis te bepalen, of de diversiteit binnen een set van moleculen te bepalen. Nieuwe toepassingen van descriptoren binnen het modelleren van structuur-activiteits- en structuur-eigenschapsrelaties worden besproken. Een overzicht van recente benaderingen voor het valideren van classificatie- en regressiemodelen wordt ook besproken. Het laatste stuk van dit hoofdstuk geeft een lijst van nieuwe databases en introduceert methoden die het halen van informatie uit databases en opslagmedia verbeteren. Semantische opmaaktalen maken het uitwisselen van chemische informatie eenvoudiger, terwijl nieuwe methoden het makkelijker maken om gegevens te extraheren uit bestaande tekstdocumenten.

Hoofdstuk 3 beschrijft een analyse van het in de literatuur voorgestelde gebruik van 1D ^{13}C en ^{1}H NMR-spectra als moleculaire descriptor. Het is bekend dat deze spectra moleculaire eigenschappen beschrijven die van invloed zijn of fysische eigenschappen zoals oplosbaarheid en kookpunt. De NMR-representatie wordt getest of deze gebruikt kan worden in voorspellende PLS-modellen voor drie structuur-eigenschap datasets. De resultaten geven echter aan dat proton NMR geen geschikte descriptor is. Koolstof NMR-modellen geven redelijke QSPR-modellen geven, waarbij de regressievector zelfs relevante moleculaire fragmenten aanwijst die de fysische eigenschap benvloeden. Maar zelfs deze vorm van NMR-descriptor geeft slechtere modellen dan eenvoudig te berekenen numerieke descriptoren afgeleid van de moleculaire structuur. De conclusie is dan ook dat koolstof NMR-spectra niet de eerste keus moeten zijn bij het maken van voorspellende structuur-eigenschap modellen, en dat proton NMR helemaal niet gebruikt moet worden.

Een wiskundige methode gebaseerd op een nieuwe representatie van kristalstructuren wordt in Hoofdstuk 4 geïntroduceerd om de similariteit tussen twee kristalstructuren te berekenen. Een oudere methode is als referentie gebruikt die wel in staat was identieke maar niet op elkaar lijkende kristalstructuren te herkennen. Deze laatste eigenschap is belangrijk bij het clusteren van kristalstructuren in kleine groepen van kristalstructuren met een gelijke pakkingspatroon. De nieuwe representatie voor kristalstructuren die in dit hoofdstuk beschreven is, kan wel gelijkende structuren herkennen. Het clusteren van met de computer gesimuleerde kristalpolymorfen van estron en de classificatie van een set experimenteel gemeten cefalosporine kristalstructuren laten zien dat deze methode in staat is om juiste clusters te vinden.

Hoofdstuk 5 beschrijft het gebruik van gestuurd-getrainde zelf-organiserende kaarten voor het in kaart brengen van kristalstructuren die zijn gerepresenteerd via hun poederdiffractiepatronen en een of meer eigenschappen. De topologische structuur van de kaarten zijn afhankelijk van zowel de representatie als ook van de getrainde eigenschappen, zoals celinhoud, ruimtegroep en de energie van het kristalrooster. Met deze kaarten is het mogelijk grote sets van kristalstructuren te analyseren. De resultaten laten zien dat de kristalstructuren zich beter verdelen over de kaart en dat ze ook gebruikt kunnen

worden om eigenschappen te voorspellen uitgaande van het diffractie patroon en om een representatieve subset van kristalstructuren te selecteren. De twee toepassingen laten zien dat het gebruik van een goede representatie en similariteitsmaat het mogelijk maakt om grote hoeveelheden kristalstructuren te analyseren.

Hoofdstuk 6 introduceert en bespreekt een methode die het mogelijk maakt om data op een informatie-rijke manier tussen twee apparaten te verzenden, zoals tussen een meetapparaat en een computer. Bestaande technieken vereisen vaak kennis van niet of slecht gedocumenteerde details over het formaat, dat tot informatieverlies kan leiden. De CMLRSS-techniek is beschreven en combineert twee bestaande web-standaarden: Rich Site Summaries (RSS), ook bekend als RDF Site Summaries, en de Chemical Markup Language (CML). RSS wordt in CMLRSS gebruikt als transportlaag, terwijl CML gebruikt wordt als drager van de chemische informatie. CML ondersteund veel soorten chemische data, waaronder kristalstructuren, reactievergelijkingen en -schemas, en experimentele data zoals NMR-spectra. De resultaten laten zien dat deze semantische representatie het mogelijk maakt geautomatiseerd chemische data te verspreiden.

Het zevende hoofdstuk introduceert de gezamenlijke inzet van een internationale groep wetenschappers om interoperabiliteit binnen informatische chemie te realiseren. Het noemt zichzelf de *Blue Obelisk*. Deze beweging omvat op dit moment meer dan tien kleinere en grotere open broncode en open data projecten op het gebied van chemoinformatica en chemie in het algemeen. Om de reproduceerbaarheid van moleculaire representaties te vergroten, is een woordenboek van chemoinformatica algoritmes en een database van chemische en fysische informatie opgesteld, die in dit hoofdstuk beschreven worden. De database bevat informatie over de chemische elementen en hun isotopen, zoals kookpunten, electronaffiniteiten, massa's, en radii. De beschikbaarheid van deze standaardlijst van atomaire eigenschappen en het gebruik daarvan door meerdere programma's, maakt het reproduceren en valideren van gepubliceerde resultaten in de moleculaire chemometrie eenvoudiger. De conclusies uit hoofdstuk 3 laten zien dat dit geen overbodige luxe is.

Het laatste hoofdstuk vat de resultaten die in dit proefschrift beschreven staan samen, en laat zien hoe deze gebruikt kunnen worden bij nieuwe uitdagingen in de moleculaire chemie. Dit werk laat de sterke interactie tussen representatie en analysemethode zien: de representatie moet relevante informatie bevatten, maar de methode moet tevens in staat zijn deze informatie uit de representatie te halen om het effectief te kunnen gebruiken. De hoofdstukken beschrijven voorbeelden van deze interactie en gebruikt modelvalidatie op basis van statistische en visuele methoden als middel om een goede interactie tussen representatie en analysemethode zichtbaar te maken.

Curriculum Vitae

Egon Lennert Willighagen werd geboren op 27 oktober 1974 in Arnhem. Na het behalen van zijn VWO diploma, begin hij in 1993 aan de studie Scheikunde aan de Katholieke Universiteit Nijmegen, met als specialisatie Informatische Chemie. Nevenafstudeerrichting was de synthese and characterisatie van gemini amfifielen in de groep van prof. Nolte. Hoofdafstudeerrichting was de representatie van moleculaire kristallen in de Analytische Chemie vakgroep van prof. Buydens. In 2001 werd hij aangesteld als junioronderzoeker bij dezelfde groep. Tijdens de promotie volgde hij een leerstage aan Cambridge University (prof. Murray-Rust) en het European Bioinformatics Institute (prof. Thornton). Dit proefschrift beschrijft het onderzoek dat tijdens deze promotie is uitgevoerd.

Publication List

- E.L. Willighagen, N.M. O'Boyle, H.Gopalakrishnan, D.Jiao, R.Guha, C.Steinbeck, D.J. Wild, *Userscripts for the Life Sciences*, BMC Bioinformatics, 8:487, **2007**, doi:10.1186/1471-2105-8-487

- S. Kuhn, P. Murray-Rust, R.J. Lancashire, H. Rzepa, T. Helmus, E.L. Willighagen, C. Steinbeck, *Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data*, J. Chem. Inf. Model., 47:2015-2034, **2007**, doi:10.1021/ci600531a

- E.L. Willighagen, R. Wehrens, W. Melssen, R. de Gelder, and L.M.C. Buydens, *Supervised self-organizing maps in crystal structure prediction*, Crystal Growth & Design, 7:1738-1745, **2007**, doi:10.1021/cg060872y

- O. Spjuth, T. Helmus, E.L. Willighagen, S. Kuhn, M. Eklund, J. Wagener, P. Murray-Rust, C. Steinbeck, J.E.S. Wikberg, *Bioclipse: An open source workbench for chemo- and bioinformatics*, BMC Bioinformatics, 8:59, **2007**, doi:10.1186/1471-2105-8-59

- E.L. Willighagen, R. Wehrens, and L.M.C. Buydens, Molecular Chemometrics, Crit. Rev. Anal. Chem., 36:189-198, **2006**, doi:10.1080/10408340600969601

- R. Guha, P. Murray-Rust, M. Howard, H. Rzepa, G.R. Hutchison, C. Steinbeck, J. Wegner, E.L. Willighagen, *The Blue Obelisk - Interoperability in Chemical Informatics*, J. Chem. Inf. Model., 46:991-998, **2006**, doi:10.1021/ci050400b

- E.L. Willighagen, H.M.G.W. Denissen, R. Wehrens, and L.M.C. Buydens, *On the use of 1H and 13C NMR spectra as QSPR descriptors*, J. Chem. Inf. Model., 46:487-494, **2006**, doi:10.1021/ci050282s

- C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E.L. Willighagen, *Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics*, Curr. Pharm. Design, 12(17):2111-2120, **2006**, doi:10.2174/138161206777585274

- R. Wehrens, E.L. Willighagen, *Mapping databases of X-ray powder patterns*, RNews, 6(3):24-28, **2006**

- E.L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder, and L.M.C. Buydens, *Method for the Computational Comparison of Crystal Structures*, Acta.Cryst., B61:29-36, **2005**, doi:10.1107/S0108768104028344

- P. Murray-Rust, H.S. Rzepa, M.J. Williamson, E.L. Willighagen, *Chemical Markup, XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators*, J.Chem.Inf.Comp.Sci., 44, 462-9, **2004**, doi:10.1021/ci034244p

- C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E.L. Willighagen, *The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics*, J.Chem.Inf.Comp.Sci., 43:493-500, **2003**, doi:10.1021/ci025584y

- P.J.J.A. Buijnsters, C.L. Garcia Rodriguez, E.L. Willighagen, N.A.J.M. Sommerdijk, A. Kremer, P. Camilleri, M.C. Feiters, R.J.M. Nolte, B. Zwanenburg, *Cationic Gemini Surfactants Based on Tartaric Acid: Synthesis, Aggregation, Monolayer Behaviour, and Interaction with DNA*, European Journal of Organic Chemistry, 1397-1406, **2002**, doi:10.1002/1099-0690(200204)2002:8<1397::AID-EJOC1397>3.0.CO;2-6

- E.L. Willighagen, *Processing CML Conventions in Java*, Internet Journal of Chemistry, 4, **2001**

- S. Krause, E.L. Willighagen, and C. Steinbeck, *JChemPaint - Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures*, Molecules, 5:93-98, **2000**

Dankwoord

This section is a special and tricky one to write. It is so easy to forget to express gratitude (*Dankwoord*) to those who deserve it: those who actively participated in getting this thesis written and printed on paper; those who made the PhD period enjoyable; and, all who contributed in other ways. Thank you. Now, it is expected that this section provides a long list of names. So, here goes.

First of all, I would like to thank Lutgarde Buydens and Ron Wehrens for their guidance during the research that led to, and the writing of this thesis. I have learned a lot; about the importance of validation in particular. I always enjoyed work meetings the three of us had, even though those tended to result in disturbances in the fabric of my space-time continuum, both before and after the meeting. Keeps you on edge, so to say. These meetings always provide new insights, simply by not having been around in the field as long as Lutgarde and Ron have been. Anyway, a discussion about what *kernel PLS* is, is always interesting. The concept of *prove me wrong* I will propagate during my further life.

Secondly, I warmly thank everyone with whom I have worked together on scientific problems, and with whom I ended up so fellow authors of research papers. This includes René de Gelder and Paul Verwer who taught me about the fine details of crystallography and polymorph prediction. And Willem Melssen who developed the XYF method and with whom I had pleasant discussions on supervised learning. I would also like to thank Harm Denissen who studied during his M.Sc. the role of NMR spectra as molecular descriptor, which formed the foundation for one of the chapters in this thesis. Henry Rzepa came up with the idea of CMLRSS when I visited him in London where we discussed the Chemical Markup Language. Peter Murray-Rust invited me for a project in cooperation with Janet Thornton. I warmly thank the both of them for giving me the opportunity to work at the Unilever Center in Cambridge and at the European Bioinformatics Institute in Hinxton, for tools to support the development of a reaction database. I include Gemma Holliday and Gail Bartlett; we regularly discussed the information rich markup needed for describing enzymatic reactions. Rajarshi Guha, Geoff Hutchinson, Jörg Wegner and Christoph Steinbeck I thank for their interesting discussions on Open Data, Open Source and Open Standard going back to even four years before the start of my PhD studies. These discussions lead to the paper on the Blue Obelisk, and several papers outside the scope of thesis.

I also want to mention my fellow PhD students and the other people in the Analytical Chemistry of Lutgarde, with whom I have an overlapping working period. This includes Theo, Han, Philip, Arjan, Uwe, Jos, Bülent, Jorn, Patrick, Velitchka, Simon, and of course Geert and Brigitte, and the students who I guided in their literature studies. YY and Jürgen should certainly not be forgotten, with whom worked in Cambridge on the World Wide Molecular Matrix.

Finally, last but not least, I thank Karin, Lars and Fien, who forgave me the late hours that required me to finish this thesis. The representation of molecules on the frontside of this book was designed by Lars, something I'm quite uncapable of myself. Karin helped me by proofreading most of the content, helping with the layout of the cover, and getting this thesis printed. Lars and Karin, therefore, actively participated in the completion.

To all others who feel left out, please take comfort in the fact that a good deal of this thesis is freely available as Open Data or Open Source. Think of it as free, as in free beer :)