chem-bla-ics

TODO: April 2nd, defend my PhD work

Egon Willighagen

Published March 1, 2008

Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/87t55-whn79

Keywords
Cheminf, Chemometrics, Phd

In 4.5 weeks, on Wednesday April 2 (13:30 precisely, Aula, Comeniuslaan 2, Nijmegen) I will publicly defend my PhD work performed in the Analytical Chemistry group of Prof. Lutgarde Buydens at the Radboud University Nijmegen:

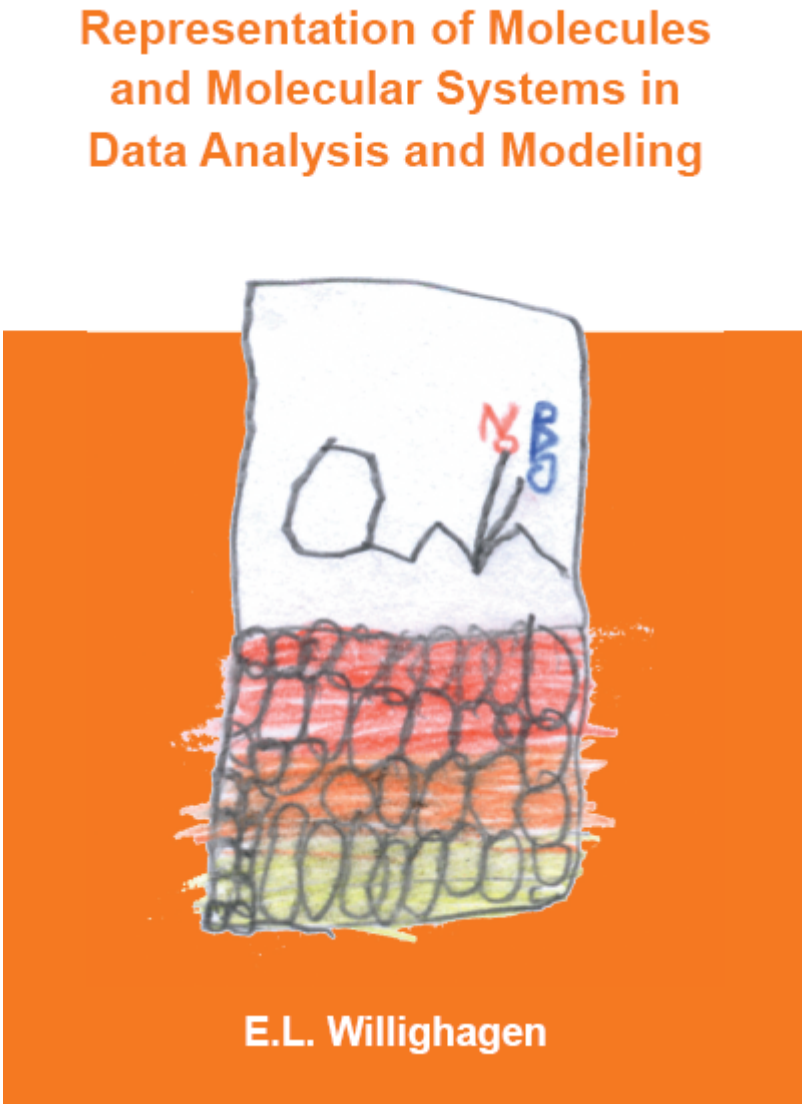

Table of Contents

Chapters 2, 3, 4, and 5 are first author papers, while for chapters 6 and 7 I am just co-author.

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.