1 Introduction

Thomas Engel[1] *and Johann Gasteiger*[2]

[1] *Ludwig-Maximilians-University Munich, Department of Chemistry, Butenandtstraße 5-13, 81377 Munich, Germany*
[2] *Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstr. 25, 91052 Erlangen, Germany*

Outline

1.1 The Rationale for the Books

In 2003 we issued the book

Chemoinformatics: A Textbook
(J. Gasteiger, T. Engel, Editors, Wiley-VCH Verlag GmbH, Weinheim, Germany, ISBN 13: 978-3-527-30681-7)

which was well accepted and contributed to the development of the field of chemoinformatics. However, with the enormous progress in chemoinformatics, it is now time for an update. As we started out on this endeavor, it became rapidly clear that all the developments require presenting the field in more than a single book. We have therefore edited two volumes:

- Chemoinformatics – Basic Concept and Methods
- Applied Chemoinformatics – Achievements and Future Opportunities [1]

In this first volume, "Basic Concept and Methods," the essential foundations and methods that comprise the technology of chemoinformatics are presented.

Chemoinformatics: Basic Concepts and Methods, First Edition.
Edited by Thomas Engel and Johann Gasteiger.

The second volume, "From Methods to Applications," shows how this technology has been applied to a variety of fields such as chemistry, drug discovery, pharmacology, toxicology, agricultural, food, and material science as well as process control. The links to the second volume are referenced in the present volume by **"Applications Volume"**. The "Applications Volume" emerged from the single "Applications" chapter of the 2003 textbook. The fact that applications now merit a book of their own clearly demonstrates how enormously the field has grown. Chemoinformatics has certainly matured to a scientific discipline of its own with many applications in all areas of chemistry and in related fields.

Both volumes consist of chapters written by different authors. In order to somehow ensure that the material is not too heterogeneous, we have striven to adapt the contributions to an overall picture and inserted cross-references as mentioned above. We hope that this helps the reader to realize the interdependences of many of the methods and how they can work together in solving chemical problems.

Both volumes are conceived as textbooks for being used in teaching and self-learning of chemoinformatics. In particular, this first, "Methods Volume," is addressed to students, explaining the basic approaches and supporting this with exercises. Altogether, we wanted to present with both books a comprehensive overview of the field of chemoinformatics for students, teachers, and scientists from all areas of chemistry, from biology, informatics, and medicine.

1.2 The Objectives of Chemoinformatics

Chemistry deals with compounds and their properties and transformations. The field of chemistry has experienced an enormous development in the last two centuries, and this development has dramatically increased in the last couple of decades. On the other hand, society has become increasingly interested in the effects chemicals have on human health and on the environment. Therefore, it wants to know *a priori* about these effects and not *a posteriori*. This demands for methods that allow one to make predictions on the physical, chemical, and biological properties of compounds and to make predictions on the course and products of chemical reactions.

Quantum mechanics is a method in theoretical chemistry, but its application to many problems of high interest is too complicated to be solved or asks for computational resources that are still beyond reach. For example, this applies to the interaction of chemicals with biological systems or to the influence of reaction conditions such as time, temperature, solvent, or catalyst on chemical reactions (although quite interesting inroads have already been made).

How is it then that although the laws of chemistry are too complicated to be solved, chemists still can do their jobs and make compounds with wonderful properties that society needs and chemists run reactions from small-scale laboratory experiments to large-scale reactors in the chemical industry? The secret to success has been learning from experiments and learning from data. Chemists have done a series of experiments, have analyzed them, have looked

for common features and for those that are different, have developed models that allowed them to put these observations into a systematic ordering scheme, have made inferences and checked them with new experiments, and have then confirmed, rejected, or refined their models. This process is called *inductive learning* (Figure 1.1), a method chemists have employed from the very beginning (see Section 1.3).

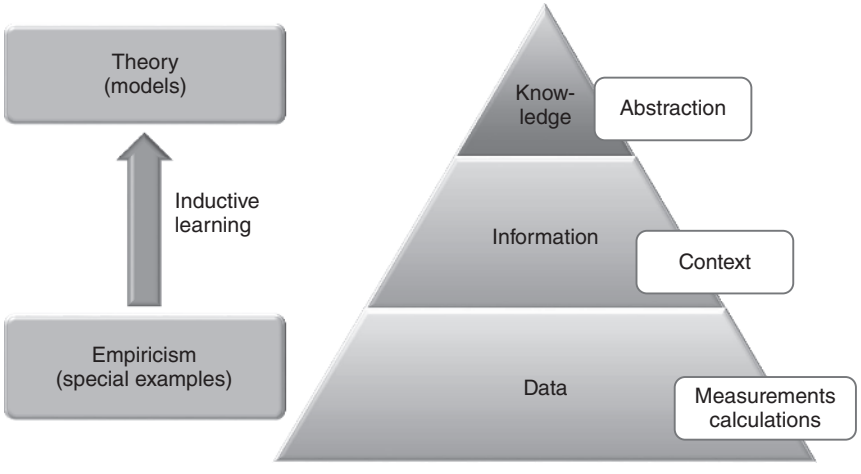

Figure 1.1 Inductive learning.

In this manner, the laws and rules of nature and of compounds and their reactions were learned. Thus, enough knowledge was accumulated to launch an entire industry, the chemical industry, that produces a cornucopia of chemicals having a wide range of properties that allow us to maintain our present standard of living. This process of inductive learning is still not over; we are still far away from understanding and predicting all chemical phenomena. This is most vividly illustrated by our poor knowledge of undesired side effects of compounds, such as toxicity. We still have to strive to increase our knowledge of chemistry.

This is where chemoinformatics comes in!

Typical challenges where chemoinformatics methods might assist are for the three fundamental questions of a chemist:

1. **What structure do I need for a certain property, be it a drug, a paint, or a glue?**
   This is the domain of establishing *structure–property relationship* (SPR) or *structure–activity relationship* (SAR) or even finding such relationships on a quantitative basis (QSPR or QSAR).
2. **How can I synthesize the compound that should have the desired property?**
   This is the domain of *synthesis design* and the *planning of chemical reactions*.

3. **What is the structure of the compound that was obtained in my reaction?** This is the domain of *structure elucidation*, which, in most part, utilizes information from a battery of spectra (infrared, NMR, and mass spectra).

An additional difficulty soon became apparent: the amount of chemical information was dramatically increasing. It became clear that managing this huge quantity of information could only be handled with electronic means, by storing the information in databases. Chemoinformatics methods have been developed to assist in the process of inductive learning, in supporting chemists in solving their three fundamental questions, and in storing chemical information in databases.

In this book, we want to build on the long history of applying informatics methods to chemical problems and pay tribute to the scientists who have started out decades ago to develop this interdisciplinary field. For this field we prefer the broad and general definition:

> *"Chemoinformatics is the application of informatics methods to solve chemical problems."*

This broad definition was the basis for the conception of this textbook.

The term has different spellings: chemoinformatics and cheminformatics. The varying use of both terms seems to indicate a geographical (or perhaps cultural) divide, with "cheminformatics" mainly used in the United States and "chemoinformatics" originating and more widely used in Europe and the rest of the world.

1.3 Learning in Chemoinformatics

In the endeavor to deepen an understanding of chemistry, many experiments have been performed, and many data have been accumulated. The chapter on databases will give a vivid picture of the enormous amounts of data that have been determined and made accessible. The task is then to derive knowledge from these data by inductive learning. In this context we have to define the terms data, information, and knowledge, and we do so in a widely accepted manner:

- *Data:* Any observation provides data. These could be the result of a physical measurement, a yes/no answer whether a reaction occurs or not, or the determination of a biological activity.
- *Information:* If data are put into context with other data, we call it information. The measurement of the biological activity of a compound gains in value if we also know the molecular structure of that compound.
- *Knowledge:* Obtaining knowledge needs some level of abstraction. Many pieces of information are ordered in the framework of a model; rules are derived from a sequence of observations; predictions can be made by analogy.

Figure 1.1 illustrates this hierarchy in going from data through information to knowledge.

In the case of chemoinformatics, this process of abstraction will mostly be performed to gain knowledge about the properties of compounds. Physical, chemical, or biological data of compounds will be associated with each other or with data on the structure of a compound. These pieces of information will then be analyzed by inductive learning methods to obtain a model that allows one to make predictions.

1.4 Outline of the Book

The general outline of the book by and large follows the outline of the 2003 textbook. However, some new chapters and sections were added and new methods were included.

The following four chapters (Chapters 2–5) introduce the basic concepts for representing the entities in the computer that chemistry deals with, chemical structures, chemical reactions, and data on properties. This then leads in Chapter 6 to a presentation of databases on chemical information and in Chapter 7 how to find this information in these databases. It is clear that without the use of databases, modern chemical research cannot be performed anymore. The construction of databases on chemical information is one of the major achievements of chemoinformatics, a gift by chemoinformatics that was given into the hands of scientists to perform their research more efficiently.

However, many data on compounds are not known and therefore cannot be found in databases. Thus, one has to strive to predict the needed data. Several types of data can directly be calculated by deductive methods, be it by empirical methods, molecular mechanics, or quantum mechanics as shown in Chapter 8. For most types of data, however, the relationships between chemical structure and the data are too complicated to be directly calculable. This is where inductive learning methods come in. Chapters 9–12 present the methods required for building models that can predict data on properties of compounds.

Chapter 13 presents an overview of bioinformatics methods.

Finally, in Chapter 14 we dare to make predictions on where the field of chemoinformatics will venture in the future.

After this overview on the strategy built into the outline of the book, we briefly present the individual chapters:

Chapter 2, "Principles of Molecular Representation," introduces some chemical concepts and shows some important methods for the representation of chemical structures. The way how this structural information can be represented by computer methods is shown with various methods in "Computer Processing of Chemical Structure Information" (Chapter 3).

Concepts and methods for the representation of chemical reactions are collected in Chapter 4, "Representation of Chemical Reactions."

Chapter 5, "The Data," presents the different types of data that are met in chemistry and how they can be stored in various file formats. It illustrates the

data acquisition pathway and discusses data complexity. Chapter 6, "Databases and Data Sources in Chemistry," presents an overview of the various types of databases for storing chemical information.

The various forms of structure search methods, full structure, substructure, superstructure, and similarity searches are presented in "Searching Chemical Structures" (Chapter 7). This chapter also indicates methods for searching for biological sequences of proteins and nucleic acids.

Chapter 8, "Computational Chemistry," shows that quite a few properties of chemical compounds can be calculated explicitly. These deductive methods start with the listing of simple *empirical methods* (Section 8.1). Then, various *molecular mechanics* approaches (Section 8.2) are presented, and the presentation continues with *molecular dynamics* calculations (Section 8.3). Then, methods from *quantum chemistry* are explained starting from the simple Hückel molecular orbital (MO) method through semiempirical methods to *ab initio* and density functional theory calculations (Section 8.4).

Chapter 9, "Modeling and Prediction of Properties (QSPR/QSAR)," provides an outline of the methodology for predicting properties of chemical compounds by inductive learning methods. These methods are usually subsumed as quantitative structure–property or structure–activity relationships (QSPR/QSAR). This chapter gives an outline of the steps involved in establishing such relationships that are presented in more detail in Chapters 10–12.

Chapter 10, "Calculation of Structure Descriptors," presents methods for extracting descriptors from chemical structures for use in the qualitative classification of quantitative modeling methods. They can be represented with various degrees of sophistication. In fact, the structure descriptors form a hierarchy starting from global descriptors to 1D, 2D, and 3D representation of chemical structures or the representation of molecular surface properties. Also, briefly, methods are given for the representation of chemical compounds whose molecular structure is not known.

Inductive methods for establishing a correlation between chemical compounds and their properties are the theme of Chapter 11 that is allocated to three sections: Section 11.1, "Methods for Multivariate Data Analysis," lists methods for a basic statistical analysis of data and for data preprocessing such as centering, scaling, and normalization and methods for the definition of distance and similarity of data. Then, methods for quantitative modeling (calibration) are indicated followed by methods for the classification of objects. Section 11.2, "Artificial Neural Networks," presents methods that model the information processing in the brain both for unsupervised and for supervised learning as well as for classification and for modeling tasks. Recently the term "deep learning" has appeared in many areas of economics and science, including chemoinformatics. Section 11.3 puts the methods subsumed under "deep learning" into perspective with more established methods.

Chapter 12, "QSPR/QSAR Revisited," explains how data, structure descriptors, and data analysis methods are best utilized to obtain good prediction results. Section 12.1, "Best Practices of QSAR Modeling," emphasizes how care has to be taken at each step to build a successful and useful QSAR model. "The Data Science of QSAR Modeling" (Section 12.2) presents methods for data collection,

data curation, data integration, and data structuring and shows how QSAR models can be used as decision support systems. Knowing that chemoinformatics and bioinformatics methods are increasingly used jointly to solve difficult problems in drug discovery and in understanding biological systems, we have included a chapter on *bioinformatics* (Chapter 13). This chapter lists various sequence databases of proteins and nucleic acids and indicates the methods for searching in these sequence databases and for the interpretation of the search results. It further presents methods for the characterization of protein families and for homology modeling.

The book concludes with Chapter 14, "Future Directions," giving a personal view of the further development of the field of chemoinformatics.

1.5 The Scope of the Book

This book is conceived as a textbook to be used in teaching and self-learning of chemoinformatics. We wanted to present a comprehensive overview of the field of chemoinformatics for students, teachers, and scientists from all areas of chemistry, biology, informatics, and medicine. Those interested in a more in-depth presentation and analysis of some of the topics of this book are referred to an accompanying set of four volumes,

Handbook of Chemoinformatics: From Data to Knowledge

which was printed in 2003 and is also available on the Web since 2008 (*http://onlinelibrary.wiley.com/book/10.1002/9783527618279*).

These additional volumes present many of the topics of the present book in more detail by leading experts in the various fields.

Clearly, some of the topics touched in this "Methods Volume" would deserve entire books of their own. This is particularly true for the methods discussed in Chapters 8 and 11. The subjects

- Molecular mechanics
- Quantum mechanics
- Free energy relationships
- Chemometrics
- Neural networks
- Fuzzy logic
- Genetic algorithms
- Expert systems

are presented in a variety of books that are cited in the references of the corresponding chapters of this volume. Here, we can only present the major foundations, methods, and uses of these subjects as deemed necessary. Readers interested in a more in-depth presentation of these topics are referred to these references. Furthermore, we emphasize here the concepts and contents of chemistry in chemoinformatics. We had to refrain from introducing those parts that would deal more with the informatics part, such as

- Theory of algorithms
- Programming languages
- Database management systems

This is not to say that we deem these topics not to be important. On the contrary, we think that those interested in chemoinformatics should strive to obtain a basic knowledge in these subjects. Nevertheless, presenting those aspects of informatics here would have gone beyond the scope of this book on chemoinformatics.

1.6 Teaching Chemoinformatics

Chemoinformatics has matured to a scientific discipline that will – and in some cases did already – change the way we perceive chemistry. The chemical industry and, in particular, pharmaceutical industry are in high need of chemoinformatics specialists. Thus, this field has to be taught in academia, both in specialized courses on chemoinformatics and by integrating chemoinformatics into regular chemistry curricula. In particular, this first "*Methods Volume*" is addressed to students, explaining the basic approaches and supporting this with exercises. In addition, parallel to this book, Alexandre Varnek has published a very valuable collection of tutorials in Chemoinformatics [2].

References

[1] Engel, T. and Gasteiger, J. (eds) (2017) *Applied Chemoinformatics – Achievements and Future Opportunities*, Wiley-VCH Verlag GmbH, Weinheim, 648pp.

[2] Varnek, A. (ed.) (2017) *Tutorials in Chemoinformatics*, J. Wiley & Sons Ltd., 482pp.