1 Introduction

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

[1] *Ludwig-Maximilians-University Munich, Department of Chemistry, Butenandtstr. 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 [1]
- Applied Chemoinformatics – Achievements and Future Opportunities

In the first volume, "Basic Concept and Methods," the essential foundations and methods that comprise the technology of chemoinformatics are presented. The links to this first volume are referenced in the present volume by "**Methods Volume.**"

The *Application Volume* – tagged as shortcut in the first 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

Applied Chemoinformatics: Achievements and Future Opportunities, First Edition.
Edited by Thomas Engel and Johann Gasteiger.

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, the 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 Development of the Field

We are happy to see – and demonstrate within this book – that chemoinformatics has ventured into and found applications in many fields other than drug discovery. Drug discovery is still the most important area of application of chemoinformatics methods (Chapter 6), but we expect that the other fields of applications will continue to grow.

Some comments on terminology are appropriate. The varying use of the terms *chemoinformatics* and *cheminformatics* 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.

Chemometrics is a field that originated in the early 1970s and is mainly associated with analytical chemistry, as is shown in Chapter 9. The two fields, chemometrics and chemoinformatics, have borrowed heavily from each other and use many of the same methods. As chemoinformatics has a much broader focus, it is fair to consider the data analysis methods in chemometrics as a part of chemoinformatics.

Molecular informatics has appeared as a term, but this is certainly too narrow a name to cover all potential chemoinformatics applications, as many tasks in chemistry deal with compounds and materials that are not limited to or cannot be identified with single molecular structures. The applications in materials science, presented in Chapter 12, are a case in point, as are the applications in process control illustrated in Chapter 13.

More fortunate developments have brought chemoinformatics closer together with other disciplines. The borders between chemoinformatics and *bioinformatics* are becoming quite transparent. Some important and challenging tasks in drug design, and cosmetics development, and in trying to understand living systems need approaches from both fields. We have taken account of this with Section 4.3 on biochemical pathways; Section 6.2 on drugs, targets, and diseases; and Chapter 11 on computational approaches in cosmetics products discovery.

In a similar manner, many tasks faced in chemistry and related fields are tackled by methods of both chemoinformatics and *computational chemistry*, thus blurring the borders between the two disciplines. We indicate the combined utilization of chemoinformatics and computational chemistry in Chapter 5 on structure–spectra correlations and *computer-assisted structure elucidation* (CASE), in Chapter 7 on computational approaches in agricultural research, and in Chapter 11 on computational approaches in cosmetics products discovery. Furthermore, the basic concepts of *computational chemistry* are described in Chapter 8 of the *Methods Volume* [1].

It is exciting to see that the use of computers in chemistry, by methods both of chemoinformatics and of computational chemistry, has gained more widespread recognition. This culminated in the awarding of the *Nobel Prize in Chemistry in 2013* to Martin Karplus, Michael Levitt, and Arieh Warshel. The Swedish Academy of Sciences motivated its decision by stating:

> "Today the computer is just as important a tool for chemists as the test tube."
> (https://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/press.html)

1.3 The Basis of Chemoinformatics and the Diversity of Applications

We were fortunate to gain many excellent scientists to demonstrate the various applications of chemoinformatics in chemistry and related fields by writing chapters for this book, but it should be emphasized from the very beginning that these applications are, by no means, all of the established or possible applications: the field is still expanding and growing.

1.3.1 Databases

Probably the most important achievement of chemoinformatics is the building of databases on chemical information. The development of chemoinformatics technologies made it possible to store and retrieve chemical information in a variety of databases. Interaction with these databases can be made by the international language of the chemist that is graphical in nature: structure diagrams and reaction equations. Databases are so routinely used in chemical research that it seems to have been forgotten that their construction was only possible by the work and developments by chemists, mathematicians, and computer scientists in the decades from 1960 to 1990. The building of databases on chemical information is presented in Chapter 6 of the *Methods Volume* [1]. In the present volume the exploitation of these databases will be shown in a variety of applications presented in the chapters.

Without these databases an overview of chemical information, which has grown enormously in recent decades, could not be maintained anymore. It is fair to say that modern chemical research can nowadays not be done without consulting databases on chemical information.

1.3.2 Fundamental Questions of a Chemist

In structuring the presentation of the applications of chemoinformatics methods, we were motivated by the fundamental questions of a chemist and how they can be supported by chemoinformatics methods (see Figure 1.1).

1.3.2.1 Prediction of Properties

In his Norris Award Lecture of 1968, George S. Hammond said:

> "The most fundamental and lasting objective of synthesis is not production of new compounds but *production of properties*."

With this in mind, the first fundamental task of a chemist is to relate the desired property, be it a drug, a pesticide, a paint, or an antiaging compound, with a chemical structure. This is the domain of *structure–property relationships* (SPR) or *structure–activity relationships* (SAR), or even finding such relationships on a quantitative basis by quantitative structure- property relationships (QSPR) or quantitative structure–activity relationships QSAR.

Chapter 2 presents the methodology of the QSPR and QSAR approach for establishing models that allow the prediction of properties. Quite a few of the steps in this approach have been outlined in more detail in the *Methods Volume* [1] in Chapters 9–12. However, because of the importance of this approach in many areas of chemistry and for the prediction of a wide variety of properties,

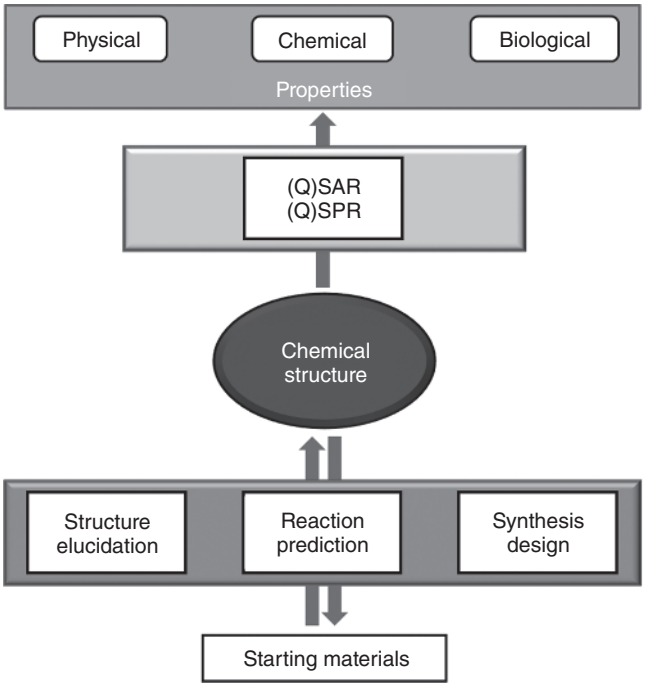

Figure 1.1 Fundamental questions of a chemist and the chemoinformatics methods that can be used in providing support for solving these tasks.

we have decided to outline the QSPR/QSAR approach as an entry to the various chapters in this *Application Volume*.

Many such QSPR and QSAR models for the prediction of physical, chemical, or biological properties have been established. Chapter 3 on the prediction of physicochemical properties presents various approaches to the prediction of some important physical and chemical properties. The prediction of some more properties relevant for drug discovery is indicated in Section 6.9 on the prediction of absorption, distribution, metabolism, and excretion(ADME) properties. The effects of chemicals on living species and on the environment are of much concern to society. The evaluation of these effects is indicated in Chapter 8 on chemoinformatics in modern regulatory science. The prediction of properties of importance in other fields of chemistry is presented in Chapter 7 on computational approaches to agricultural research and in Chapter 9 on chemometrics in analytical chemistry.

Some more recent areas of applications are given in Chapter 10 on chemoinformatics in food chemistry, in Chapter 11 on computational approaches in cosmetics products discovery, and in Chapter 12 on chemoinformatics in material science.

1.3.2.2 Chemical Reaction Prediction and Synthesis Design

Once a chemist has an idea which chemical structure will carry the desired property, he or she is faced with the task of synthesizing this compound and first coming up with a plan for performing the synthesis. This is the domain of *synthesis design* and *planning chemical reactions*. Section 4.2 on chemical reactions and synthesis design addresses these tasks, reflecting on work on computer-assisted synthesis design (CASD) that can be considered as one of the roots of chemoinformatics. Particularly interesting types of reactions are those that keep living organisms alive and thriving. These are discussed in Section 4.3 on biochemical pathways. This section also shows how chemoinformatics and bioinformatics can work together to obtain interesting insights such as discovering the essential pathways of diseases.

1.3.2.3 Structure Elucidation

As our knowledge on the driving forces of chemical reactions is still incomplete and many factors influence the outcome of a reaction, we must first verify whether the reaction that was performed produced the desired product. This is usually established by measuring various spectra to deduce the structure of the reaction product. This requires analysis of the relationships between the spectral data and chemical structure. At an early stage, in the late 1960s, methods for CASE were developed and can be considered as another root of chemoinformatics. In fact, the DENDRAL project at Stanford University is widely considered to be the first application of artificial intelligence to chemical problems. Chapter 5 on structure–spectra correlations and CASE presents the most recent developments in this field.

1.3.3 Drug Discovery

Drug discovery is certainly the most prominent field for the application of chemoinformatics methods. All major drug companies have divisions of

chemoinformatics in various guises, and drug companies are the largest employers of chemoinformatics specialists. Furthermore, all drugs developed in the last few years have benefited from the use of chemoinformatics methods in one way or other. Because of its importance, the chapter on drug discovery is the most extensive one in this book. The introductory material found in Section 6.1 on drug discovery provides the framework for a series of the more specialized Sections 6.2–6.13. In addition, it also mentions methods and applications that are not represented in these sections. These 12 specialized sections present a variety of topics such as the use of bioinformatics methods on drugs, targets, and diseases in Section 6.2, the exploitation of natural products for drug discovery in Section 6.3, and in Section 6.4 on chemoinformatics in Chinese herbal medicines. Section 6.3 also shows applications in natural products research that have other focuses beyond drug discovery. In Section 6.5, the efforts made at the US National Institutes of Health are described in providing the general public with data on biological screening results on chemical compounds by PubChem. The PubChem database is an enormous resource for academic and industrial researchers and promises to heavily increase the insights made by scientists working in drug discovery.

The following three sections present methods for analyzing the three-dimensional structure of drug candidates and of proteins as targets for a biological response: Section 6.6 on pharmacophore perception and analysis, Section 6.7 on prediction and analysis of active sites, and Section 6.8 on structure-based virtual screening.

The further development of many drug candidates has to be discontinued in the preclinical phase because of adverse effects such as poor bioavailability, unfavorable pharmacokinetic effects, and undesirable metabolic stability. Many models have been developed for the prediction of the corresponding ADME properties. These are presented in Section 6.9 on the prediction of ADME properties and in Section 6.10 on the prediction of xenobiotic metabolism. The Computer-Aided Drug Design (CADD) Group at the National Cancer Institute has collected and developed a series of chemoinformatics methods for assisting in the drug design process and is making them generally available as detailed in Section 6.11. Many different information resources, such as those found in printed media and various databases, offer interesting information that can be harnessed for drug discovery. The use of these resources is discussed in Section 6.12 on the exploration of new data sources. The last section in Section 6.13 on drug design, status, and future of drug design, offers the personal views of an active academic researcher on where the drug discovery process will develop. His ideas are supported with recent developments from his research group.

1.3.4 Additional Fields of Application

It is gratifying to observe that, as we had hoped in the chapter on future directions in the 2003 textbook, various new fields have benefited from the use of chemoinformatics methods. A number of these will be shown in Chapters 7–13.

The *agrochemical industry* is employing the very same methods used in drug discovery for the development of new plant protection products. These include ligand-based and structure-based methods and are discussed in Chapter 7 on computational approaches in agricultural research.

The effects of chemicals on *human health* and their impact on the *environment* have become of increasing concern to society. At the same time, there is much interest in society to reduce, or even phase out, the use of animals for testing for toxicity of chemicals. For this purpose, computer models for the prediction of toxicity, environmental effects, and bioaccumulation have gained much interest for the registration of new and old chemicals, as indicated in Chapter 8 on chemoinformatics in modern regulatory science.

The analysis of data from *analytical chemistry* by statistical and other mathematical methods, subsumed under the term chemometrics, has a long history. Chapter 9 on chemometrics in analytical chemistry presents the methods used in this field and some typical applications (see also *Methods Volume*, Chapter 11).

It is interesting to see that rather novel areas of applications have been developed in recent years. For example, this is true for *food chemistry* as shown in Chapter 10 on chemoinformatics in food chemistry.

With the pressure on reducing, or even phasing out, the testing of chemicals with animals, the cosmetics industry has started to employ chemoinformatics and other computational chemistry methods. This has meant that the development of new *cosmetics products* is made more efficient, as presented in Chapter 11 on computational approaches in cosmetics products discovery.

One of the potentially largest and most promising new fields of application of chemoinformatics is *materials science*. Here, new ways for representing the objects of study have to be developed, since for many of the materials, no molecular structure can be given. The challenge of properly representing materials in the computer is of crucial influence on the success of any study. This will become clear in Chapter 12 on chemoinformatics in material science.

Many processes in the chemical industry are influenced by a multitude of factors and are governed by these factors in a nonlinear fashion. Indeed, in many cases no explicit mathematical relationship can be specified. In these situations, chemoinformatics methods can be helpful in *controlling* processes, mainly by measuring various control factors through sensors, and using those data for modeling a process. This is shown with a few examples in Chapter 13 on process control and soft sensors.

We hope that the many varied applications in these chapters show the importance of chemoinformatics in supporting chemical research and development in many fields, but it should be emphasized that by far, not all of the problems have been solved. Many challenging problems remain, and chemoinformatics will further evolve in parallel with chemical and biological sciences to provide deeper insights into these and related sciences. Thus, chemoinformatics is in very active development and provides many opportunities for future students. Chapter 14 on future directions collects some ideas on the further development of the field.

Reference

[1] Engel, T. and Gasteiger, J. (eds) (2017) *Chemoinformatics – Basic Concepts and Methods*, Wiley-VCH, Weinheim, 600 pp.